
RESEARCH CENTER
**Grenoble - Rhône-Alpes**

FIELD

Activity Report 2018

# Section New Results

Edition: 2019-03-07

**AGORA Project-Team**

# 7. New Results

## 7.1. Wireless network deployment

*Participants : Walid Bechkit, Amjed Belkhiri, Jad Oueis, Hervé Rivano, Razvan Stanica, Fabrice Valois*

### *7.1.1. UAVs positionning*

Mobile base stations mounted on unmanned aerial vehicles (UAVs) provide viable wireless coverage solutions in challenging landscapes and conditions, where cellular/WiFi infrastructure is unavailable. Operating multiple such airborne base stations, to ensure reliable user connectivity, demands intelligent control of UAV movements, as poor signal strength and user outage can be catastrophic to mission critical scenarios. In [17], we propose a deep reinforcement learning based solution to tackle the challenges of base stations mobility control. We design an Asynchronous Advantage Actor-Critic (A3C) algorithm that employs a custom reward function, which incorporates SINR and outage events information, and seeks to provide mobile user coverage with the highest possible signal quality. Preliminary results reveal that our solution converges after 4 × 10 5 steps of training, after which it outperforms a benchmark gradient-based alternative, as we attain 5dB higher median SINR during an entire test mission of 10,000 steps.

### *7.1.2. Network functions placement*

Emerging mobile network architectures (e.g., aerial networks, disaster relief networks) are disrupting the classical careful planning and deployment of mobile networks by requiring specic self-deployment strategies. Such networks, referred to as self-deployable, are formed by interconnected rapidly deployable base stations that have no dedicated backhaul connection towards a traditional core network. Instead, an entity providing essential core network functionalities is co-located with one of the base stations. In [5], we tackle the problem of placing this core network entity within a self-deployable mobile network, i.e., we determine with which of the base stations it must be co-located. We propose a novel centrality metric, the ow centrality, which measures a node capacity of receiving the total amount of ows in the network. We show that in order to maximize the amount of exchanged trac between the base stations and the core network entity, under certain capacity and load distribution constraints, the latter should be co-located with the base station having the maximum ow centrality. We rst compare our proposed metric to other state of the art centralities. Then, we highlight the signicant trac loss occurring when the core network entity is not placed on the node with the maximum ow centrality, which could reach 55

### *7.1.3. Mobile edge computing orchestration*

Orchestrating network and computing resources in Mobile Edge Computing (MEC) is an important item in the networking research agenda. In [12], we propose a novel algorithmic approach to solve the problem of dynamically assigning base stations to MEC facilities, while taking into consideration multiple time-periods, and computing load switching and access latency costs. In particular, leveraging on an existing state of the art on mobile data analytics, we propose a methodology to integrate arbitrary time-period aggregation methods into a network optimization framework. We notably apply simple consecutive time period aggregation and agglomerative hierarchical clustering. Even if the aggregation and optimization methods represent techniques which are different in nature, and whose aim is partially overlapping, we show that they can be integrated in an efficient way. By simulation on real mobile cellular datasets, we show that, thanks to the clustering, we can scale with the number of time-periods considered, that our approach largely outperforms the case without time-period aggregations in terms of MEC access latency, and at which extent the use of clustering and time aggregation affects computing time and solution quality.

### 7.1.4. On User Mobility in Dynamic Cloud Radio Access Networks

The development of virtualization techniques enables an architectural shift in mobile networks, where resource allocation, or even signal processing, become software functions hosted in a data center. The centralization of computing resources and the dynamic mapping between baseband processing units (BBUs) and remote antennas (RRHs) provide an increased flexibility to mobile operators, with important reductions of operational costs. Most research efforts on Cloud Radio Access Networks (CRAN) consider indeed an operator perspective and network-side performance indicators. The impact of such new paradigms on user experience has been instead overlooked. In [20], we shift the viewpoint, and show that the dynamic assignment of computing resources enabled by CRAN generates a new class of mobile terminal handover that can impair user quality of service. We then propose an algorithm that mitigates the problem, by optimizing the mapping between BBUs and RRHs on a time-varying graph representation of the system. Furthermore, we show that a practical online BBU-RRH mapping algorithm achieves results similar to an oracle-based scheme with perfect knowledge of future traffic demand. We test our algorithms with two large-scale real-world datasets, where the total number of handovers, compared with the current architectures, is reduced by more than 20%. Moreover, if a small tolerance to dropped calls is allowed, 30% less handovers can be obtained.

### 7.1.5. Wireless sensor network deployment for environmental monitoring

Air pollution has major negative effects on both human health and environment. Thus, air quality monitoring is a main issue in our days. In [9], we focus on the use of mobile WSN to generate high spatio-temporal resolution air quality maps. We address the sensors' online redeployment problem and we propose three redeployment models allowing to assess, with high precision, the air pollution concentrations. Unlike most of existing movement assisted deployment strategies based on network generic characteristics such as coverage and connectivity, our approaches take into account air pollution properties and dispersion models to offer an efficient air quality estimation. First, we introduce our proposition of an optimal integer linear program based on air pollution dispersion characteristics to minimize estimation errors. Then, we propose a local iterative integer linear programming model and a heuristic technique that offer a lower execution time with acceptable estimation quality. We evaluate our models in terms of execution time and estimation quality using a real data set of Lyon City in France. Finally, we compare our models' performances to existing generic redeployment strategies. Results show that our algorithms outperform the existing generic solutions while reducing the maximum estimation error up to 3 times.

## 7.2. Wireless data collection

*Participants : Walid Bechkit, Ahmed Boubrima, Alexis Duque, Abdoul-Aziz Mbacke, Hervé Rivano, Razvan Stanica, Yosra Zguira*

### 7.2.1. RFID paradigm

While RFID technology is gaining increased attention from industrial community deploying different RFID-based applications, it still suffers from reading collisions. As such, many proposals were made by the scientific community to try and alleviate that issue using different techniques either centralized or distributed, mono-channel or multi-channels, TDMA or CSMA. However, the wide range of solutions and their diversity make it hard to have a clear and fair overview of the different works. In [4], we propose a survey of the most relevant and recent known state-of-the- art anti-collision for RFID protocols. It provides a classification and performance evaluation taking into consideration different criteria as well as a guide to choose the best protocol for given applications depending on their constraints or requirements but also in regard to their deployment environments.

### 7.2.2. Anti-collision and routing protocol for RFID

In the midst of Internet of Things development, a first requirement was tracking and identification of those mentioned "things" which could be done thanks to Radio Frequency Identification. However, since then, the development of RFID allowed a new range of applications among which is remote sensing of environmental values. While RFID can be seen as a more efficient solution than traditional Wireless Sensor Networks,

two main issues remain: first reading collisions and second proficient data gathering solution. In [18], we examine the implementation of two applications: for industrial IoT and for smart cities, respectively. Both applications, in regards to their requirements and configuration, challenge the operation of a RFID sensing solution combined with a dynamic wireless data gathering over multi-hops. They require the use of both mobile and fixed readers to cover the extent of deployment area and a quick retrieval of tag information. We propose a distributed cross-layer solution for improving the efficiency of the RFID system in terms of collision and throughput but also its proficiency in terms of tag information routing towards one or multiple sinks. Simulation results show that we can achieve high level of throughput while maintaining a low level of collision and a fairness of reader medium access above 95% in situations where readers can be fix and mobile, while tag information is routed with a data rate of 97% at worst and reliable delays for considered applications.

### *7.2.3. Routing priority information in RFID*

Long being used for identification purposes, a new set of applications is now available thanks to the development of RFID technology. One of which is remote sensing of environmental values using passive RFID tags. This leap forward allowed a more energy efficient and cheaper solution for applications like logistics or urban infrastructure monitoring. Nevertheless, serious issues raised with the use of RFID: (i) reading collisions and (ii) gathering of tag information. Indeed, tags information retrieved by readers have to be transmitted towards a base station through a multi-hop scheme which can interfere with neighboring readers activity. In [19], we propose cross-layer solutions meant for both scheduling of readers' activity to avoid collisions, and a multi-hop routing towards base stations, to gather read tag data. This routing is performed with a data priority aware mechanism allowing end-to-end delay reduction of urgent data packets delivery up to 13% faster compared to standard ones. Using fuzzy logic, we combine several observed metrics to reduce the load of forwarding nodes and improve latency as well as data rate. We validate our proposal running simulations on industrial and urban scenarios.

### *7.2.4. Data collection in DTN networks*

Intelligent Transport Systems (ITS) are an essential part of the global world. They play a substantial role for facing many issues such as traffic jams, high accident rates, unhealthy lifestyles, air pollution, etc. Public bike sharing system is one part of ITS and can be used to collect data from mobiles devices. In this paper, we propose an efficient, *Internet of Bikes*, IoB-DTN routing protocol based on data aggregation which applies the Delay Tolerant Network (DTN) paradigm to Internet of Things (IoT) applications running data collection on urban bike sharing system based sensor network. In [6], we propose and evaluate three variants of IoB-DTN: IoB based on spatial aggregation (IoB-SA), IoB based on temporal aggregation (IoB-TA) and IoB based on spatiotemporal aggregation (IoB-STA). The simulation results show that the three variants offer the best performances regarding several metrics, comparing to IoB-DTN without aggregation and the low-power long-range technology, LoRa type. In an urban application, the choice of the type of which variant of IoB should be used depends on the sensed values.

### *7.2.5. Data sensing in Internet of Bikes*

Following the trend of the Internet of Thing, public transport systems are seen as an efficient bearer of mobile devices to generate and collect data in urban environments. Bicycle sharing system is one part of the city's larger transport system. In [23], we study the *Internet of Bikes* IoB-DTN protocol which applies the Delay Tolerant Network (DTN) paradigm to the Internet of Things (IoT) applications running on urban bike sharing system based sensor network. We evaluate the performances of the protocol with respect to the transmission power. Performances are measured in terms of delivery rate, delivery delay, throughput and energy cost. We also compare the multi-hop IoB-DTN protocol to a low-power wide-area network (LPWAN) technology. LPWAN have been designed to provide cost-effective wide area connectivity for small throughput IoT applications: multiyear lifetime and multi-kilometer range for battery-operated mobile devices. This work aims at providing network designers and managers insights on the most relevant technology for their urban applications that could run on bike sharing systems. To the best of our knowledge, this work is the first to provide a detailed performance comparison between multi-hop and long range DTN-like protocol being applied to mobile network IoT devices running a data collection applications in an urban environment.

### 7.2.6. Reducing IoT traffic through data aggregation mechanisms

Intelligent Transport Systems (ITS) are an essential part of the global world. They play a substantial role for facing many issues such as traffic jams, high accident rates, unhealthy lifestyles, air pollution, etc. Public bike sharing system is one part of ITS and can be used to collect data from mobiles devices. In this paper, we propose an efficient, " Internet of Bikes " , IoB-DTN routing protocol based on data aggregation which applies the Delay Tolerant Network (DTN) paradigm to Internet of Things (IoT) applications running data collection on urban bike sharing system based sensor network. In [6], we propose and evaluate three variants of IoB-DTN: IoB based on spatial aggregation (IoB-SA), IoB based on temporal aggregation (IoB-TA) and IoB based on spatiotemporal aggregation (IoB-STA). The simulation results show that the three variants offer the best performances regarding several metrics, comparing to IoB-DTN without aggregation and the low-power long-range technology, LoRa type. In an urban application, the choice of the type of which variant of IoB should be used depends on the sensed values.

### 7.2.7. Environmental modeling

Wireless sensor networks (WSN) are widely used in environmental applications where the aim is to sense a physical parameter such as temperature, humidity, air pollution, etc. Most existing WSN-based environmental monitoring systems use data interpolation based on sensor measurements in order to construct the spatiotemporal field of physical parameters. However, these fields can be also approximated using physical models which simulate the dynamics of physical phenomena. In [11], we focus on the use of wireless sensor networks for the aim of correcting the physical model errors rather than interpolating sensor measurements. We tackle the activity scheduling problem and design an optimization model and a heuristic algorithm in order to select the sensor nodes that should be turned off to extend the lifetime of the network. Our approach is based on data assimilation which allows us to use both measurements and the physical model outputs in the estimation of the spatiotemporal field. We evaluate our approach in the context of air pollution monitoring while using a dataset from the Lyon city, France and considering the characteristics of a monitoring system developed in our lab. We analyze the impact of the nodes' characteristics on the network lifetime and derive guidelines on the optimal scheduling of air pollution sensors.

### 7.2.8. Multi-robot routing for evolving missions

In [22], we propose Dynamic Multi Robot-Routing (DMRR), as a continuous adaptation of the multi-robot target allocation process (MRTA) to new discovered targets. There are few works addressing dynamic target allocation. Existing methods are lacking the continuous integration of new targets, handling its progressive effects, but also lacking dynamic support (e.g. parallel allocations, participation of new robots). This work proposes a framework for dynamically adapting the existing robot missions to new discovered targets. Missions accumulate targets continuously, so the case of a saturation bound for the mission costs is also considered. Dynamic saturation-based auctioning (DSAT) is proposed for allocating targets, providing lower time complexities (due to parallelism in allocation). Comparison is made with algorithms ranging from greedy to auction-based methods with provable sub-optimality. The algorithms are tested on exhaustive sets of inputs, with random configurations of targets (for DMRR with and without a mission saturation bound). The results for DSAT show that it outperforms state-of-the-art methods, like standard sequential single-item auctioning (SSI) or SSI with regret clearing.

### 7.2.9. Measuring information using VLC

The use of visible light for bidirectional communication between regular smartphones and the small LEDs integrated in most consumer electronics nowadays raises new challenges. In [13], we enhance the state of the art with an efficient image processing algorithm to accurately detect the LEDs and decode their signal in real time. We propose an efficient decoding algorithm, which can detect the LED position, process and decode the signal on average in 18.4 ms, for each frame, on a Nexus 5 unrooted smartphone. Thus, this implementation is convenient for low latency indoor localization or real-time transmission with a moving receiver. Also, as the ROI detection is the most complex step of the algorithm, scenarios with several transmitters can be envisaged, enabling MIMO-like transmissions. We also present smart mechanisms and protocols to build a robust flash-to-LED communications channel using off- the-shelf smartphones and small LEDs. Our experimental evaluation

shows a throughput of 30 bit/s, which is suitable for feedback, wake-up or even some limited communication purposes. We believe that such bidirectional VLC communication system will be a great opportunity for smart and connected consumer electronic products, providing bidirectional smartphone- to-device communication at lower cost.

**AIRSEA Project-Team**

# 6. New Results

## 6.1. Modeling for Oceanic and Atmospheric flows

### *6.1.1. Coupling Methods for Oceanic and Atmospheric Models*

**Participants:** Eric Blayo, Florian Lemarié, Sophie Thery.

Coupling methods routinely used in regional and global climate models do not provide the exact solution to the ocean-atmosphere problem, but an approximation of one [72]. For the last few years we have been actively working on the analysis of ocean-atmosphere coupling both in terms of its continuous and numerical formulation.Our activities over the last few years can be divided into four general topics

1. *Stability and consistency analysis of existing coupling methods*: in [72] we showed that the usual methods used in the context of ocean-atmosphere coupling are prone to splitting errors because they correspond to only one iteration of an iterative process without reaching convergence. Moreover, those methods have an additional condition for the coupling to be stable even if unconditionally stable time stepping algorithms are used. This last remark was further studied recently in [47] and it turned out to be a major source of instability in atmosphere-snow coupling.
2. *Study of physics-dynamics coupling*: during the PhD-thesis of Charles Pelletier (funded by Inria and defended on Feb. 15, 2018, [2]) the scope was on including the formulation of physical parameterizations in the theoretical analysis of the coupling, in particular the parameterization schemes to compute air-sea fluxes [79]. To do so, a metamodel representative of the behavior of the full parameterization but with a continuous form easier to manipulate has been derived thanks to a sensitivity analysis. This metamodel is more adequate to conduct the mathematical analysis of the coupling while being physically satisfactory [80]. In parallel we have contributed to a general review gathering the main international specialists on the topic [64]. More recently we have started to work specifically on the discretization methods for the parameterization of planetary boundary layers in climate models [27] which takes the form of an nonstationary nonlinear parabolic equation. The objective is to derive a discretization for which we could prove nonlinear stability criteria and show robustness to large variations in parabolic Courant number while being consistent with our knowledge of the underlying physical principles (e.g. the Monin-Obukhov theory in the surface layer).
3. *Design of a coupled single column model*: in order to focus on specific problems of ocean-atmosphere coupling, a work on simplified equation sets has been started. The aim is to implement a one-dimensional (in the vertical direction) coupled model with physical parameterizations representative of those used in realistic models. Thanks to this simplified coupled model the objective is to develop a benchmark suite for coupled models evaluation. Last year the single column oceanic and atmospheric components have been developed and coupled during the PhD-thesis of Rémi Pellerej (defended on Mar. 26, 2018) and in the framework of the SIMBAD project [17]. A publication describing this model is currently in preparation for the Geoscientific Model Development journal.
4. *Analysis of air-sea-wave interactions in realistic high-resolution realistic simulations*: part of our activity has been in collaboration with atmosphericists and physical oceanographers to study the impact on some modeling assumptions (e.g. [73]) in large-scale realistic ocean-atmosphere coupled simulations [14]. Moreover, within the ALBATROS project, we have contributed to the development of a 2-way coupling between an ocean global circulation model (NEMO) with a surface wave model (WW3). Such coupling is not straightforward to implement since it requires modifications of the governing equations, boundary conditions and subgrid scale closures in the oceanic model. A paper is currently under review in Geoscientific Model Development journal on that topic.

5. *Efficient coupling methods*: we have been developing coupling approaches for several years, based on so-called Schwarz algorithms. In particular, we addressed the development of efficient interface conditions for multi-physics problems representative of air-sea coupling [28] (paper in preparation). This work is done in the framework of S. Théry PhD (started in fall 2017).

These topics are addressed through strong collaborations between the applied mathematicians and the climate community (Meteo-France, Ifremer, LMD, and LOCEAN). Indeed, Our work on ocean-atmosphere coupling has steadily matured over the last few years and has reached a point where it triggered interest from the climate community. Through the funding of the COCOA ANR project (started in January 2017, PI: E. Blayo), Airsea team members play a major role in the structuration of a multi-disciplinary scientific community working on ocean-atmosphere coupling spanning a broad range from mathematical theory to practical implementations in climate models. An expected outcome of this project should be the design of a benchmark suite of idealized coupled test cases representative of known issues in coupled models. Such idealized test cases should motivate further collaborations at an international level.

### *6.1.2. Numerical Schemes for Ocean Modelling*

**Participants:** Eric Blayo, Matthieu Brachet, Laurent Debreu, Emilie Duval, Nicholas Kevlahan, Florian Lemarié, Christopher Eldred, Farshid Nazari.

The increase of model resolution naturally leads to the representation of a wider energy spectrum. As a result, in recent years, the understanding of oceanic submesoscale dynamics has significantly improved. However, dissipation in submesoscale models remains dominated by numerical constraints rather than physical ones. Effective resolution is limited by the numerical dissipation range, which is a function of the model numerical filters (assuming that dispersive numerical modes are efficiently removed). A review paper on coastal ocean models has been written with German colleagues and has been published in Ocean Modelling ([11]).
F. Lemarié and L. Debreu (with H. Burchard, K. Klingbeil and J. Sainte-Marie) have organized the international COMMODORE workshop on numerical methods for oceanic models (Paris, Sept. 17-19, 2018). https://commodore2018.sciencesconf.org/, see [12] for a summary of the scientific discussions

With the increase of resolution, the hydrostatic assumption becomes less valid and the AIRSEA group also works on the development of non-hydrostatic ocean models. The treatment of non-hydrostatic incompressible flows leads to a 3D elliptic system for pressure that can be ill conditioned in particular with non geopotential vertical coordinates. That is why we flavor the use of the non-hydrostatic compressible equations that removes the need for a 3D resolution at the price of reincluding acoustic waves [24].
In addition, Emilie Duval started her PhD in September 2018 on the coupling between the hydrostatic incompressible and non-hydrostatic compressible equations.
The team is involved in the HEAT (Highly Efficient ATmospheric Modelling) ANR project. This project aims at developing a new atmospheric dynamical core (DYNAMICO) discretized on an icosahedral grid. This project is in collaboration with Ecole Polytechnique, Meteo-France, LMD, LSCE and CERFACS. This year we worked on dispersion analysis of compatible Galerkin schemes for shallow water model both in 1D ([5]) and 2D ([39]). In addition, we worked on the discrete formulation of the thermal rotating shallow water equations. This formulation, based on quasi-Hamiltonian discretizations methods, allows for the first time total mass, buoyancy and energy conservation to machine precision ([4]).

### *6.1.3. Data assimilation for coupled models*

In the context of operational meteorology and oceanography, forecast skills heavily rely on proper combination of model prediction and available observations via data assimilation techniques. Historically, numerical weather prediction is made separately for the ocean and the atmosphere in an uncoupled way. However, in recent years, fully coupled ocean-atmosphere models are increasingly used in operational centers to improve the reliability of seasonal forecasts and tropical cyclones predictions. For coupled problems, the use of separated data assimilation schemes in each medium is not satisfactory since the result of such assimilation process is generally inconsistent across the interface, thus leading to unacceptable artefacts. Hence, there is a strong need for adapting existing data assimilation techniques to the coupled framework. As part of our ERACLIM2 contribution, R. Pellerej started a PhD on that topic late 2014 and defended it early 2018

[1]. Three general data assimilation algorithms, based on variational data assimilation techniques, have been developed and applied to a single column coupled model. The dynamical equations of the considered problem are coupled using an iterative Schwarz domain decomposition method. The aim is to properly take into account the coupling in the assimilation process in order to obtain a coupled solution close to the observations while satisfying the physical conditions across the air-sea interface. Results shows significant improvement compared to the usual approach on this simple system. The aforementioned system has been coded within the OOPS framework (Object Oriented Prediction System) in order to ease the transfer to more complex/realistic models.

Finally, CASIS, a new collaborative project with Mercator Océan has started late 2017 in order to extend developments to iterative Kalman smoother data assimilation scheme, in the framework of a coupled ocean-atmospheric boundary layer context.

### *6.1.4. Optimal control of grids and schemes for ocean model.*

**Participants:** Laurent Debreu, Eugene Kazantsev.

In [33], variational data assimilation technique is applied to a simple bidimensional wave equation that simulates propagation of internal gravity waves in the ocean in order to control grids and numerical schemes. Grid steps of the vertical grid, Brunt-Vaisala frequency and approximation of the horizontal derivative were used as control parameters either separately or in the joint control. Obtained results show that optimized parameters may partially compensate errors committed by numerical scheme due to insufficient grid resolution.

Optimal vertical grid steps and coefficients in horizontal derivative approximation found in the variational control procedure allow us to get the model solution that is rather close to the solution of the reference model. The error in the wave velocity on the coarse grid is mostly compensated in experiments with joint control of parameters while the error in the wave amplitude occurs to be more difficult to correct.

However, optimal grid steps and discretization schemes may be in a disagreement with requirements of other model physics and additional analysis of obtained optimized parameters from the point of view of they agreement with the model is necessary.

### *6.1.5. Nonhydrostatic Modeling*

**Participants:** Eric Blayo, Laurent Debreu, Emilie Duval.

In the context of the French initiative CROCO (Coastal and Regional Ocean COmmunity model, https://www.croco-ocean.org) for the development of a new oceanic modeling system, Emilie Duval started a PhD (Oct. 2018) focused on the design of methods to couple local nonhydrostatic models to larger scale hydrostatic ones. Such a coupling is quite delicate from a mathematical point of view, due to the different nature of hydrostatic and nonhydrostatic equations (where the vertical velocity is either a diagnostic or a prognostic variable).

## 6.2. Model reduction / multiscale algorithms

### *6.2.1. Model Order Reduction*

**Participants:** Mohamed Reda El Amri, Youssef Marzouk, Maëlle Nodet, Clémentine Prieur, Alessio Spantini, Olivier Zahm.

Another point developed in the team for sensitivity analysis is model reduction. To be more precise regarding model reduction, the aim is to reduce the number of unknown variables (to be computed by the model), using a well chosen basis. Instead of discretizing the model over a huge grid (with millions of points), the state vector of the model is projected on the subspace spanned by this basis (of a far lesser dimension). The choice of the basis is of course crucial and implies the success or failure of the reduced model. Various model reduction methods offer various choices of basis functions. A well-known method is called “proper orthogonal decomposition" or “principal component analysis". More recent and sophisticated methods also exist and may be studied, depending on the needs raised by the theoretical study. Model reduction is a natural way to overcome difficulties due to huge computational times due to discretizations on fine grids. In [68], the

authors present a reduced basis offline/online procedure for viscous Burgers initial boundary value problem, enabling efficient approximate computation of the solutions of this equation for parametrized viscosity and initial and boundary value data. This procedure comes with a fast-evaluated rigorous error bound certifying the approximation procedure. The numerical experiments in the paper show significant computational savings, as well as efficiency of the error bound.

When a metamodel is used (for example reduced basis metamodel, but also kriging, regression, ...) for estimating sensitivity indices by Monte Carlo type estimation, a twofold error appears: a sampling error and a metamodel error. Deriving confidence intervals taking into account these two sources of uncertainties is of great interest. We obtained results particularly well fitted for reduced basis metamodels [69]. In [66], the authors provide asymptotic confidence intervals in the double limit where the sample size goes to infinity and the metamodel converges to the true model. These results were also adapted to problems related to more general models such as Shallow-Water equations, in the context of the control of an open channel [70].

When considering parameter-dependent PDE, it happens that the quantity of interest is not the PDE's solution but a linear functional of it. In [67], we have proposed a probabilistic error bound for the reduced output of interest (goal-oriented error bound). By probabilistic we mean that this bound may be violated with small probability. The bound is efficiently and explicitly computable, and we show on different examples that this error bound is sharper than existing ones.

A collaboration has been started with Christophe Prieur (Gipsa-Lab) on the very challenging issue of sensitivity of a controlled system to its control parameters [70]. In [71], we propose a generalization of the probabilistic goal-oriented error estimation in [67] to parameter-dependent nonlinear problems. One aims at applying such results in the previous context of sensitivity of a controlled system.

More recently, in the context of the Inria associate team UNQUESTIONABLE, we have extended the focus of the axis on model order reduction. Our objectives are to understand the kinds of low-dimensional structure that may be present in important geophysical models; and to exploit this low-dimensional structure in order to extend Bayesian approaches to high-dimensional inverse problems, such as those encountered in geophysical applications. Our recent and future efforts are/will be concerned with parameter space dimension reduction techniques, low- rank structures in geophysical models and transport maps tools for probability measure approximation. At the moment, scientific progress has been achieved in different directions, as detailed below: A first paper [45] has been submitted on gradient- based dimension reduction of vector-valued functions. Multivariate functions encountered in high-dimensional uncertainty quantification problems often vary most strongly along a few dominant directions in the input parameter space. In this work, we propose a gradient-based method for detecting these directions and using them to construct ridge approximations of such functions, in the case where the functions are vector-valued. The methodology consists of minimizing an upper bound on the approximation error, obtained by subspace Poincaré inequalities. We have provided a thorough mathematical analysis in the case where the parameter space is equipped with a Gaussian probability measure. A second work [46] has been submitted, which proposes a dimension reduction technique for Bayesian inverse problems with nonlinear forward operators, non-Gaussian priors, and non-Gaussian observation noise. In this work, the likelihood function is approximated by a ridge function, i.e., a map which depends non-trivially only on a few linear combinations of the parameters. The ridge approximation is built by minimizing an upper bound on the Kullback-Leibler divergence between the posterior distribution and its approximation. This bound, obtained via logarithmic Sobolev inequalities, allows one to certify the error of the posterior approximation. A sample-based approximation of the upper bound is also proposed. In the framework of the PhD thesis of Reda El Amri, a work on data-driven stochastic inversion via functional quantization was submitted. In this paper [36], a new methodology is proposed for solving stochastic inversion problems through computer experiments, the stochasticity being driven by functional random variables. Main tools are a new greedy algorithm for functional quantization, and the adaptation of Stepwise Uncertainty Reduction techniques.

## 6.3. Dealing with uncertainties

### 6.3.1. Sensitivity Analysis

**Participants:** Elise Arnaud, Eric Blayo, Laurent Gilquin, Maria Belén Heredia, François-Xavier Le Dimet, Clémentine Prieur, Laurence Viry.

#### 6.3.1.1. Scientific context

Forecasting geophysical systems require complex models, which sometimes need to be coupled, and which make use of data assimilation. The objective of this project is, for a given output of such a system, to identify the most influential parameters, and to evaluate the effect of uncertainty in input parameters on model output. Existing stochastic tools are not well suited for high dimension problems (in particular time-dependent problems), while deterministic tools are fully applicable but only provide limited information. So the challenge is to gather expertise on one hand on numerical approximation and control of Partial Differential Equations, and on the other hand on stochastic methods for sensitivity analysis, in order to develop and design innovative stochastic solutions to study high dimension models and to propose new hybrid approaches combining the stochastic and deterministic methods.

### 6.3.2. Extensions of the replication method for the estimation of Sobol' indices

**Participants:** Elise Arnaud, Eric Blayo, Laurent Gilquin, Alexandre Janon, Clémentine Prieur.

Sensitivity analysis studies how the uncertainty on an output of a mathematical model can be attributed to sources of uncertainty among the inputs. Global sensitivity analysis of complex and expensive mathematical models is a common practice to identify influent inputs and detect the potential interactions between them. Among the large number of available approaches, the variance-based method introduced by Sobol' allows to calculate sensitivity indices called Sobol' indices. Each index gives an estimation of the influence of an individual input or a group of inputs. These indices give an estimation of how the output uncertainty can be apportioned to the uncertainty in the inputs. One can distinguish first-order indices that estimate the main effect from each input or group of inputs from higher-order indices that estimate the corresponding order of interactions between inputs. This estimation procedure requires a significant number of model runs, number that has a polynomial growth rate with respect to the input space dimension. This cost can be prohibitive for time consuming models and only a few number of runs is not enough to retrieve accurate informations about the model inputs.

The use of replicated designs to estimate first-order Sobol' indices has the major advantage of reducing drastically the estimation cost as the number of runs n becomes independent of the input space dimension. The generalization to closed second-order Sobol' indices relies on the replication of randomized orthogonal arrays. However, if the input space is not properly explored, that is if n is too small, the Sobol' indices estimates may not be accurate enough. Gaining in efficiency and assessing the estimate precision still remains an issue, all the more important when one is dealing with limited computational budget.

We designed an approach to render the replication method iterative, enabling the required number of evaluations to be controlled. With this approach, more accurate Sobol' estimates are obtained while recycling previous sets of model evaluations. Its main characteristic is to rely on iterative construction of stratified designs, latin hypercubes and orthogonal arrays [61]

In [7] a new strategy to estimate the full set of first-order and second-order Sobol' indices with only two replicated designs based on orthogonal arrays of strength two. Such a procedure increases the precision of the estimation for a given computation budget. A bootstrap procedure for producing confidence intervals, that are compared to asymptotic ones in the case of first-order indices, is also proposed.

The replicated designs strategy for global sensitivity analysis was also implemented in the applied framework of marine biogeochemical modeling, making use of distributed computing environments [43].

### 6.3.3. Sensitivity analysis with dependent inputs

An important challenge for stochastic sensitivity analysis is to develop methodologies which work for dependent inputs. For the moment, there does not exist conclusive results in that direction. Our aim is to define an analogue of Hoeffding decomposition [65] in the case where input parameters are correlated. Clémentine

Prieur supervised Gaëlle Chastaing's PhD thesis on the topic (defended in September 2013) [53]. We obtained first results [54], deriving a general functional ANOVA for dependent inputs, allowing defining new variance based sensitivity indices for correlated inputs. We then adapted various algorithms for the estimation of these new indices. These algorithms make the assumption that among the potential interactions, only few are significant. Two papers have been recently accepted [52], [55]. We also considered the estimation of groups Sobol' indices, with a procedure based on replicated designs [63]. These indices provide information at the level of groups, and not at a finer level, but their interpretation is still rigorous.

Céline Helbert and Clémentine Prieur supervised the PhD thesis of Simon Nanty (funded by CEA Cadarache, and defended in October, 2015). The subject of the thesis is the analysis of uncertainties for numerical codes with temporal and spatio-temporal input variables, with application to safety and impact calculation studies. This study implied functional dependent inputs. A first step was the modeling of these inputs [75]. The whole methodology proposed during the PhD is presented in [76].

More recently, the Shapley value, from econometrics, was proposed as an alternative to quantify the importance of random input variables to a function. Owen [77] derived Shapley value importance for independent inputs and showed that it is bracketed between two different Sobol' indices. Song et al. [82] recently advocated the use of Shapley value for the case of dependent inputs. In a very recent work [78], in collaboration with Art Owen (Standford's University), we show that Shapley value removes the conceptual problems of functional ANOVA for dependent inputs. We do this with some simple examples where Shapley value leads to intuitively reasonable nearly closed form values. We also investigated further the properties of Shapley effects in [41].

### *6.3.4. Global sensitivity analysis for parametrized stochastic differential equations*

**Participant:** Clémentine Prieur.

Many models are stochastic in nature, and some of them may be driven by parametrized stochastic differential equations. It is important for applications to propose a strategy to perform global sensitivity analysis for such models, in presence of uncertainties on the parameters. In collaboration with Pierre Etoré (DATA department in Grenoble), Clémentine Prieur proposed an approach based on Feynman-Kac formulas [40].

### *6.3.5. Parameter control in presence of uncertainties: robust estimation of bottom friction*

**Participants:** Victor Trappler, Elise Arnaud, Laurent Debreu, Arthur Vidard.

Many physical phenomena are modelled numerically in order to better understand and/or to predict their behaviour. However, some complex and small scale phenomena can not be fully represented in the models. The introduction of ad-hoc correcting terms, can represent these unresolved processes, but they need to be properly estimated.

A good example of this type of problem is the estimation of bottom friction parameters of the ocean floor. This is important because it affects the general circulation. This is particularly the case in coastal areas, especially for its influence on wave breaking. Because of its strong spatial disparity, it is impossible to estimate the bottom friction by direct observation, so it requires to do so indirectly by observing its effects on surface movement. This task is further complicated by the presence of uncertainty in certain other characteristics linking the bottom and the surface (eg boundary conditions). The techniques currently used to adjust these settings are very basic and do not take into account these uncertainties, thereby increasing the error in this estimate.

Classical methods of parameter estimation usually imply the minimisation of an objective function, that measures the error between some observations and the results obtained by a numerical model. In the presence of uncertainties, the minimisation is not straightforward, as the output of the model depends on those uncontrolled inputs and on the control parameter as well. That is why we will aim at minimising the objective function, to get an estimation of the control parameter that is robust to the uncertainties.

The definition of robustness differs depending of the context in which it is used. In this work, two different notions of robustness are considered: robustness by minimising the mean and variance, and robustness based on the distribution of the minimisers of the function. This information on the location of the minimisers is not a novel idea, as it had been applied as a criterion in sequential Bayesian optimisation. However, the constraint of

optimality is here relaxed to define a new estimate. To evaluate this estimation, a toy model of a coastal area has been implemented. The control parameter is the bottom friction, upon which classical methods of estimation are applied in a simulation-estimation experiment. The model is then modified to include uncertainties on the boundary conditions in order to apply robust control methods.

### 6.3.6. Development of a data assimilation method for the calibration and continuous update of wind turbines digital twins

**Participants:** Adrien Hirvoas, Elise Arnaud, Clémentine Prieur, Arthur Vidard.

In the context of the energy transition, wind power generation is developing rapidly in France and worldwide. Research and innovation on wind resource characterisation, turbin control, coupled mechanical modelling of wind systems or technological development of offshore wind turbines floaters are current research topics.

In particular, the monitoring and the maintenance of wind turbine is becoming a major issue. Current solutions do not take full advantage of the large amount of data provided by sensors placed on modern wind turbines in production. These data could be advantageously used in order to refine the predictions of production, the life of the structure, the control strategies and the planning of maintenance. In this context, it is interesting to optimally combine production data and numerical models in order to obtain highly reliable models of wind turbines. This process is of interest to many industrial and academic groups and is known in many fields of the industry, including the wind industry, as "digital twin".

The objective of Adrien Hirvoas's PhD work is to develop of data assimilation methodology to build the "digital twin" of an onshore wind turbine. Based on measurements, the data assimilation should allow to reduce the uncertainties of the physical parameters of the numerical model developed during the design phase to obtain a highly reliable model. Various ensemble data assimilation approches are currently under consideration to address the problem.

This work in done in collaboration with IFPEN.

### 6.3.7. Non-Parametric Estimation for Kinetic Diffusions

**Participants:** Clémentine Prieur, Jose Raphael Leon Ramos.

This research is the subject of a collaboration with Chile and Uruguay. More precisely, we started working with Venezuela. Due to the crisis in Venezuela, our main collaborator on that topic moved to Uruguay.

We are focusing our attention on models derived from the linear Fokker-Planck equation. From a probabilistic viewpoint, these models have received particular attention in recent years, since they are a basic example for hypercoercivity. In fact, even though completely degenerated, these models are hypoelliptic and still verify some properties of coercivity, in a broad sense of the word. Such models often appear in the fields of mechanics, finance and even biology. For such models we believe it appropriate to build statistical non-parametric estimation tools. Initial results have been obtained for the estimation of invariant density, in conditions guaranteeing its existence and unicity [48] and when only partial observational data are available. A paper on the non parametric estimation of the drift has been accepted recently [49] (see Samson et al., 2012, for results for parametric models). As far as the estimation of the diffusion term is concerned, a paper has been accepted [49], in collaboration with J.R. Leon (Montevideo, Uruguay) and P. Cattiaux (Toulouse). Recursive estimators have been also proposed by the same authors in [50], also recently accepted. In a recent collaboration with Adeline Samson from the statistics department in the Lab, we considered adaptive estimation, that is we proposed a data-driven procedure for the choice of the bandwidth parameters.

In [51], we focused on damping Hamiltonian systems under the so-called fluctuation-dissipation condition. Idea in that paper were re-used with applications to neuroscience in [74].

Note that Professor Jose R. Leon (Caracas, Venezuela, Montevideo, Uruguay) was funded by an international Inria Chair, allowing to collaborate further on parameter estimation.

We recently proposed a paper on the use of the Euler scheme for inference purposes, considering reflected diffusions.This paper could be extended to the hypoelliptic framework.

We also have a collaboration with Karine Bertin (Valparaiso, Chile), Nicolas Klutchnikoff (Université Rennes) and Jose R. León (Montevideo, Uruguay) funded by a MATHAMSUD project (2016-2017) and by the LIA/CNRS (2018). We are interested in new adaptive estimators for invariant densities on bounded domains [32], and would like to extend that results to hypo-elliptic diffusions.

### *6.3.8. Multivariate Risk Indicators*

**Participants:** Clémentine Prieur, Patricia Tencaliec.

Studying risks in a spatio-temporal context is a very broad field of research and one that lies at the heart of current concerns at a number of levels (hydrological risk, nuclear risk, financial risk etc.). Stochastic tools for risk analysis must be able to provide a means of determining both the intensity and probability of occurrence of damaging events such as e.g. extreme floods, earthquakes or avalanches. It is important to be able to develop effective methodologies to prevent natural hazards, including e.g. the construction of barrages.

Different risk measures have been proposed in the one-dimensional framework . The most classical ones are the return level (equivalent to the Value at Risk in finance), or the mean excess function (equivalent to the Conditional Tail Expectation CTE). However, most of the time there are multiple risk factors, whose dependence structure has to be taken into account when designing suitable risk estimators. Relatively recent regulation (such as Basel II for banks or Solvency II for insurance) has been a strong driver for the development of realistic spatio-temporal dependence models, as well as for the development of multivariate risk measurements that effectively account for these dependencies.

We refer to [56] for a review of recent extensions of the notion of return level to the multivariate framework. In the context of environmental risk, [81] proposed a generalization of the concept of return period in dimension greater than or equal to two. Michele et al. proposed in a recent study [57] to take into account the duration and not only the intensity of an event for designing what they call the dynamic return period. However, few studies address the issues of statistical inference in the multivariate context. In [58], [60], we proposed non parametric estimators of a multivariate extension of the CTE. As might be expected, the properties of these estimators deteriorate when considering extreme risk levels. In collaboration with Elena Di Bernardino (CNAM, Paris), Clémentine Prieur is working on the extrapolation of the above results to extreme risk levels [35]. This paper has now been accepted for publication.

Elena Di Bernardino, Véronique Maume-Deschamps (Univ. Lyon 1) and Clémentine Prieur also derived an estimator for bivariate tail [59]. The study of tail behavior is of great importance to assess risk.

With Anne-Catherine Favre (LTHE, Grenoble), Clémentine Prieur supervised the PhD thesis of Patricia Tencaliec. We are working on risk assessment, concerning flood data for the Durance drainage basin (France). The PhD thesis started in October 2013 and was defended in February 2017. A first paper on data reconstruction has been accepted [83]. It was a necessary step as the initial series contained many missing data. A second paper is in revision, considering the modeling of precipitation amount with semi-parametric models, modeling both the bulk of the distribution and the tails, but avoiding the arbitrary choice of a threshold. We work in collaboration with Philippe Naveau (LSCE, Paris).

## 6.4. Assimilation of Images

**Participants:** Elise Arnaud, François-Xavier Le Dimet, Maëlle Nodet, Arthur Vidard, Long Li.

### *6.4.1. Direct assimilation of image sequences*

At the present time the observation of Earth from space is done by more than thirty satellites. These platforms provide two kinds of observational information:

- Eulerian information as radiance measurements: the radiative properties of the earth and its fluid envelops. These data can be plugged into numerical models by solving some inverse problems.
- Lagrangian information: the movement of fronts and vortices give information on the dynamics of the fluid. Presently this information is scarcely used in meteorology by following small cumulus clouds and using them as Lagrangian tracers, but the selection of these clouds must be done by hand and the altitude of the selected clouds must be known. This is done by using the temperature of the top of the cloud.

Our current developments are targeted at the use of « Level Sets » methods to describe the evolution of the images. The advantage of this approach is that it permits, thanks to the level sets function, to consider the images as a state variable of the problem. We have derived an Optimality System including the level sets of the images. This approach is being applied to the tracking of oceanic oil spills in the framework of a Long Li's Phd in co-supervision with

A collaborative project started with C. Lauvernet (IRSTEA) in order to make use of our image assimilation strategies on the control of pesticide transfer.

### *6.4.2. Optimal transport for image assimilation*

We investigate the use of optimal transport based distances for data assimilation, and in particular for assimilating dense data such as images. The PhD thesis of N. Feyeux studied the impact of using the Wasserstein distance in place of the classical Euclidean distance (pixel to pixel comparison). In a simplified one dimensional framework, we showed that the Wasserstein distance is indeed promising. Data assimilation experiments with the Shallow Water model have been performed and confirm the interest of the Wasserstein distance. Results have been presented at conferences and seminars and a paper has been published at NPG [6].

## 6.5. Land Use and Transport Models Calibration

**Participants:** Thomas Capelle, Laurent Gilquin, Clémentine Prieur, Arthur Vidard, Peter Sturm, Elise Arnaud.

Given the complexity of modern urban areas, designing sustainable policies calls for more than sheer expert knowledge. This is especially true of transport or land use policies, because of the strong interplay between the land use and the transportation systems. Land use and transport integrated (LUTI) modelling offers invaluable analysis tools for planners working on transportation and urban projects. Yet, very few local authorities in charge of planning make use of these strategic models. The explanation lies first in the difficulty to calibrate these models, second in the lack of confidence in their results, which itself stems from the absence of any well-defined validation procedure. Our expertise in such matters will probably be valuable for improving the reliability of these models. To that purpose we participated to the building up of the ANR project CITiES led by the STEEP EPI. This project started early 2013 and two PhD about sensitivity analysis and calibration were launched late 2013. Laurent Gilquin defended his PhD in October 2016 [62] and Thomas Capelle defended his in April 2017 and published his latest results in [3].

**ARIC Project-Team**

# 7. New Results

## 7.1. Efficient approximation methods

### *7.1.1. A High Throughput Polynomial and Rational Function Approximations Evaluator*

In [21] we present an automatic method for the evaluation of functions via polynomial or rational approximations and its hardware implementation, on FPGAs. These approximations are evaluated using Ercegovac's iterative E-method adapted for FPGA implementation. The polynomial and rational function coefficients are optimized such that they satisfy the constraints of the E-method. We present several examples of practical interest; in each case a resource-efficient approximation is proposed and comparisons are made with alternative approaches.

### *7.1.2. Continued fractions in power series fields*

In [4], we explicitly describe a noteworthy transcendental continued fraction in the field of power series over $\mathbb{Q}$, having irrationality measure equal to 3. This continued fraction is a generating function of a particular sequence in the set $\{1, 2\}$. The origin of this sequence, whose study was initiated in a recent paper, is to be found in another continued fraction, in the field of power series over $\mathbb{F}_3$, which satisfies a simple algebraic equation of degree 4, introduced thirty years ago by D. Robbins.

### *7.1.3. A Lattice Basis Reduction Approach for the Design of Finite Wordlength FIR Filters*

Many applications of finite impulse response (FIR) digital filters impose strict format constraints for the filter coefficients. Such requirements increase the complexity of determining optimal designs for the problem at hand. In [6], we introduce a fast and efficient method, based on the computation of good nodes for polynomial interpolation and Euclidean lattice basis reduction. Experiments show that it returns quasi-optimal finite wordlength FIR filters; compared to previous approaches it also scales remarkably well (length 125 filters are treated in $< 9$s). It also proves useful for accelerating the determination of optimal finite wordlength FIR filters.

### *7.1.4. Validated and numerically efficient Chebyshev spectral methods for linear ordinary differential equations*

In [7], we develop a validated numerics method for the solution of linear ordinary differential equations (LODEs). A wide range of algorithms (i.e., Runge-Kutta, collocation, spectral methods) exist for numerically computing approximations of the solutions. Most of these come with proofs of asymptotic convergence, but usually, provided error bounds are non-constructive. However, in some domains like critical systems and computer-aided mathematical proofs, one needs validated effective error bounds. We focus on both the theoretical and practical complexity analysis of a so-called *a posteriori* quasi-Newton validation method, which mainly relies on a fixed-point argument of a contracting map. Specifically, given a polynomial approximation, obtained by some numerical algorithm and expressed in Chebyshev basis, our algorithm efficiently computes an accurate and rigorous error bound. For this, we study theoretical properties like compactness, convergence, invertibility of associated linear integral operators and their truncations in a suitable coefficient space of Chebyshev series. Then, we analyze the almost-banded matrix structure of these operators, which allows for very efficient numerical algorithms for both numerical solutions of LODEs and rigorous computation of the approximation error. Finally, several representative examples show the advantages of our algorithms as well as their theoretical and practical limits.

### 7.1.5. Validated semi-analytical transition matrices for linearized relative spacecraft dynamics via Chebyshev series appproximations

In [14], we provide an efficient generic algorithm to compute validated approximations of transition matrices of linear time-variant systems using Chebyshev expansions, and apply it to two different examples of relative motion of satellites (spacecraft rendezvous with Tschauner-Hempel equations and geostationary station keeping with J2 perturbation in the linearized Orange model).

### 7.1.6. A Newton-like Validation Method for Chebyshev Approximate Solutions of Linear Ordinary Differential Systems

In [22], we provide a new framework for *a posteriori* validation of vector-valued problems with componentwise tight error enclosures, and use it to design a symbolic-numeric Newton-like validation algorithm for Chebyshev approximate solutions of coupled systems of linear ordinary differential equations. More precisely, given a coupled differential system with polynomial coefficients over a compact interval (or continuous coefficients rigorously approximated by polynomials) and componentwise polynomial approximate solutions in Chebyshev basis, the algorithm outputs componentwise rigorous upper bounds for the approximation errors, with respect to the uniform norm over the interval under consideration.

A complexity analysis shows that the number of arithmetic operations needed by this algorithm (in floating-point or interval arithmetics) is proportional to the approximation degree when the differential equation is considered fixed. Finally, we illustrate the efficiency of this fully automated validation method on an example of a coupled Airy-like system.

### 7.1.7. Fuel-optimal impulsive fixed-time trajectories in the linearized circular restricted 3-body-problem

In [41], the problem of fixed-time fuel-optimal trajectories with high-thrust propulsion in the vicinity of a Lagrange point is tackled via the linear version of the primer vector theory. More precisely, the proximity to a Lagrange point i.e. any equilibrium point-stable or not-in the circular restricted three-body problem allows for a linearization of the dynamics. Furthermore, it is assumed that the spacecraft has ungimbaled thrusters, leading to a formulation of the cost function with the 1-norm for space coordinates, even though a generalization exists for steerable thrust and the 2-norm. In this context, the primer vector theory gives necessary and sufficient optimality conditions for admissible solutions to two-value boundary problems. Similarly to the case of rendezvous in the restricted two-body problem, the in-plane and out-of-plane trajectories being uncoupled, they can be treated independently. As a matter of fact, the out-of-plane dynamics is simple enough for the optimal control problem to be solved analytically via this indirect approach. As for the in-plane dynamics, the primer vector solution of the so-called primal problem is derived by solving a hierarchy of linear programs, as proposed recently for the aforementioned rendezvous. The optimal thrusting strategy is then numerically obtained from the necessary and sufficient conditions. Finally, in-plane and out-of-plane control laws are combined to form the complete 3-D fuel-optimal solution. Results are compared to the direct approach that consists in working on a discrete set of times in order to perform optimization in finite dimension. Examples are provided near various Lagrange points in the Sun-Earth and Earth-Moon systems, hinting at the extensive span of possible applications of this technique in station-keeping as well as mission analysis, for instance when connecting manifolds to achieve escape or capture.

## 7.2. Floating-point and Validated Numerics

### 7.2.1. Optimal bounds on relative errors of floating-point operations

Rounding error analyses of numerical algorithms are most often carried out via repeated applications of the so-called standard models of floating-point arithmetic. Given a round-to-nearest function $fl$ and barring underflow and overflow, such models bound the relative errors $E_1(t) = |t - fl(t)|/|t|$ and $E_2(t) = |t - fl(t)|/|fl(t)|$ by the unit roundoff $u$. In [10] we investigate the possibility and the usefulness of refining these bounds, both in the case of an arbitrary real $t$ and in the case where $t$ is the exact result of an

arithmetic operation on some floating-point numbers. We show that $E_1(t)$ and $E_2(t)$ are optimally bounded by $u/(1+u)$ and $u$, respectively, when $t$ is real or, under mild assumptions on the base and the precision, when $t = x \pm y$ or $t = xy$ with $x, y$ two floating-point numbers. We prove that while this remains true for division in base $\beta > 2$, smaller, attainable bounds can be derived for both division in base $\beta = 2$ and square root. This set of optimal bounds is then applied to the rounding error analysis of various numerical algorithms: in all cases, we obtain significantly shorter proofs of the best-known error bounds for such algorithms, and/or improvements on these bounds themselves.

### 7.2.2. On various ways to split a floating-point number

In [32] we review several ways to split a floating-point number, that is, to decompose it into the exact sum of two floating-point numbers of smaller precision. All the methods considered here involve only a few IEEE floating-point operations, with rounding to nearest and including possibly the fused multiply-add (FMA). Applications range from the implementation of integer functions such as `round` and `floor` to the computation of suitable scaling factors aimed, for example, at avoiding spurious underflows and overflows when implementing functions such as the hypotenuse.

### 7.2.3. Algorithms for triple-word arithmetic

Triple-word arithmetic consists in representing high-precision numbers as the unevaluated sum of three floating-point numbers. In [45], we introduce and analyze various algorithms for manipulating triple-word numbers. Our new algorithms are faster than what one would obtain by just using the usual floating-point expansion algorithms in the special case of expansions of length 3, for a comparable accuracy.

### 7.2.4. Error analysis of some operations involved in the Fast Fourier Transform

In [44], we are interested in obtaining error bounds for the classical FFT algorithm in floating-point arithmetic, for the 2-norm as well as for the infinity norm. For that purpose we also give some results on the relative error of the complex multiplication by a root of unity, and on the largest value that can take the real or imaginary part of one term of the FFT of a vector $x$, assuming that all terms of $x$ have real and imaginary parts less than some value $b$.

## 7.3. Lattices: algorithms and cryptology

### 7.3.1. Reduction of orthogonal lattice bases

As a typical application, the LLL lattice basis reduction algorithm is applied to bases of the orthogonal lattice of a given integer matrix, via reducing lattice bases of a special type. With such bases in input, we have proposed in [26] a new technique for bounding from above the number of iterations required by the LLL algorithm. The main technical ingredient is a variant of the classical LLL potential, which could prove useful to understand the behavior of LLL for other families of input bases.

### 7.3.2. Lattice-Based Zero-Knowledge Arguments for Integer Relations

The paper [36] provides lattice-based protocols allowing to prove relations among committed integers. While the most general zero-knowledge proof techniques can handle arithmetic circuits in the lattice setting, adapting them to prove statements over the integers is non-trivial, at least if we want to handle exponentially large integers while working with a polynomial-size modulus $q$. For a polynomial $L$, the paper provides zero-knowledge arguments allowing a prover to convince a verifier that committed $L$-bit bitstrings $x$, $y$ and $z$ are the binary representations of integers $X$, $Y$ and $Z$ satisfying $Z = X + Y$ over $\mathbb{Z}$. The complexity of the new arguments is only linear in $L$. Using them, the paper constructs arguments allowing to prove inequalities $X < Z$ among committed integers, as well as arguments showing that a committed $X$ belongs to a public interval $[\alpha, \beta]$, where $\alpha$ and $\beta$ can be arbitrarily large. The new range arguments have logarithmic cost (i.e., linear in $L$) in the maximal range magnitude. Using these tools, the paper obtains zero-knowledge arguments showing that a committed element $X$ does not belong to a public set $S$ using $O(n \cdot \log |S|)$ bits of communication, where $n$ is the security parameter. The paper finally gives a protocol allowing to argue

that committed $L$-bit integers $X$, $Y$ and $Z$ satisfy multiplicative relations $Z = XY$ over the integers, with communication cost subquadratic in $L$. To this end, the paper uses its new protocol for integer addition to prove the correct recursive execution of Karatsuba's multiplication algorithm. The security of the new protocols relies on standard lattice assumptions with polynomial modulus and polynomial approximation factor.

### 7.3.3. Logarithmic-Size Ring Signatures With Tight Security from the DDH Assumption

Ring signatures make it possible for a signer to anonymously and, yet, convincingly leak a secret by signing a message while concealing his identity within a flexibly chosen ring of users. Unlike group signatures, they do not involve any setup phase or tracing authority. Despite a lot of research efforts in more than 15 years, most of their realizations require linear-size signatures in the cardinality of the ring. In the random oracle model, two recent constructions decreased the signature length to be only logarithmic in the number N of ring members. On the downside, their suffer from rather loose reductions incurred by the use of the Forking Lemma. This paper considers the problem of proving them tightly secure without affecting their space efficiency. Surprisingly, existing techniques for proving tight security in ordinary signature schemes do not trivially extend to the ring signature setting. The paper [37] overcomes these difficulties by combining the Groth-Kohlweiss $\Sigma$-protocol (Eurocrypt'15) with dual-mode encryption schemes. The main result is a fully tight construction based on the Decision Diffie-Hellman assumption in the random oracle model. By full tightness, we mean that the reduction's advantage is as large as the adversary's, up to a constant factor.

### 7.3.4. Adaptively Secure Distributed PRFs from LWE

In distributed pseudorandom functions (DPRFs), a PRF secret key $SK$ is secret shared among $N$ servers so that each server can locally compute a partial evaluation of the PRF on some input $X$. A combiner that collects $t$ partial evaluations can then reconstruct the evaluation $F(SK, X)$ of the PRF under the initial secret key. So far, all non-interactive constructions in the standard model are based on lattice assumptions. One caveat is that they are only known to be secure in the static corruption setting, where the adversary chooses the servers to corrupt at the very beginning of the game, before any evaluation query. The paper [38] constructs the first fully non-interactive adaptively secure DPRF in the standard model. The construction is proved secure under the LWE assumption against adversaries that may adaptively decide which servers they want to corrupt. The new construction is also extended in order to achieve robustness against malicious adversaries.

### 7.3.5. Unbounded ABE via Bilinear Entropy Expansion, Revisited

This paper [24] presents simpler and improved constructions of unbounded attribute-based encryption (ABE) schemes with constant-size public parameters under static assumptions in bilinear groups. Concretely, we obtain: a simple and adaptively secure unbounded ABE scheme in composite-order groups, improving upon a previous construction of Lewko and Waters (Eurocrypt'11) which only achieves selective security; an improved adaptively secure unbounded ABE scheme based on the k-linear assumption in prime-order groups with shorter ciphertexts and secret keys than those of Okamoto and Takashima (Asiacrypt'12); the first adaptively secure unbounded ABE scheme for arithmetic branching programs under static assumptions. At the core of all of these constructions is a "bilinear entropy expansion" lemma that allows us to generate any polynomial amount of entropy starting from constant-size public parameters; the entropy can then be used to transform existing adaptively secure "bounded" ABE schemes into unbounded ones.

### 7.3.6. Improved Anonymous Broadcast Encryptions: Tight Security and Shorter Ciphertext

This paper [35] investigates anonymous broadcast encryptions (ANOBE) in which a ciphertext hides not only the message but also the target recipients associated with it. Following Libert et al.'s generic construction [PKC, 2012], we propose two concrete ANOBE schemes with tight reduction and better space efficiency.

- The IND-CCA security and anonymity of our two ANOBE schemes can be tightly reduced to standard k-Linear assumption (and the existence of other primitives). For a broadcast system with n users, Libert et al.'s security analysis suffers from $\mathcal{O}(n^3)$ loss while our security loss is constant.
- Our first ANOBE supports fast decryption and has a shorter ciphertext than the fast-decryption version of Libert et al.'s concrete ANOBE. Our second ANOBE is adapted from the first one.

We sacrifice the fast decryption feature and achieve shorter ciphertexts than Libert et al.'s concrete ANOBE with the help of bilinear groups. Technically, we start from an instantiation of Libert et al.'s generic ANOBE [PKC, 2012], but we work out all our proofs from scratch instead of relying on their generic security result. This intuitively allows our optimizations in the concrete setting.

### *7.3.7. Compact IBBE and Fuzzy IBE from Simple Assumptions*

This paper [29] proposes new constructions for identity-based broadcast encryption (IBBE) and fuzzy identity-based encryption (FIBE) in composite-order groups equipped with a bilinear pairing. Our starting point is the IBBE scheme of Delerablée (Asiacrypt 2007) and the FIBE scheme of Herranz et al. (PKC 2010) proven secure under parameterized assumptions called generalized decisional bilinear Diffie-Hellman (GDDHE) and augmented multi-sequence of exponents Diffie-Hellman (aMSE-DDH) respectively. The two schemes are described in the prime-order pairing group. We transform the schemes into the setting of (symmetric) composite-order groups and prove security from two static assumptions (subgroup decision). The Déjà $Q$ framework of Chase et al. (Asiacrypt 2016) is known to cover a large class of parameterized assumptions (dubbed "Uber assumption"), that is, these assumptions, when defined in asymmetric composite-order groups, are implied by subgroup decision assumptions in the underlying composite-order groups. We argue that the GDDHE and aMSE-DDH assumptions are not covered by the Déjà $Q$ uber assumption framework. We therefore work out direct security reductions for the two schemes based on subgroup decision assumptions. Furthermore, our proofs involve novel extensions of Déjà $Q$ techniques of Wee (TCC 2016-A) and Chase et al. Our constructions have constant-size ciphertexts. The IBBE has constant-size keys as well and achieves a stronger security guarantee as compared to Delerablée's IBBE, thus making it the first compact IBBE known to be selectively secure without random oracles under simple assumptions. The fuzzy IBE scheme is the first to simultaneously feature constant-size ciphertexts and security under standard assumptions.

### *7.3.8. Improved Inner-product Encryption with Adaptive Security and Full Attribute-hiding*

This paper [25] proposes two IPE schemes achieving both adaptive security and full attribute-hiding in the prime-order bilinear group, which improve upon the unique existing result satisfying both features from Okamoto and Takashima [Eurocrypt'12] in terms of efficiency.

- Our first IPE scheme is based on the standard $k$-Lin assumption and has shorter master public key and shorter secret keys than Okamoto and Takashima's IPE under weaker $DLIN$=2-lin assumption.
- Our second IPE scheme is adapted from the first one; the security is based on the XDLIN assumption (as Okamoto and Takashima's IPE) but now it also enjoys shorter ciphertexts.

Technically, instead of starting from composite-order IPE and applying existing transformation, we start from an IPE scheme in a very restricted setting but already in the prime-order group, and then gradually upgrade it to our full-fledged IPE scheme. This method allows us to integrate Chen et al.'s framework [Eurocrypt'15] with recent new techniques [TCC'17, Eurocrypt'18] in an optimized way.

### *7.3.9. Improved Security Proofs in Lattice-Based Cryptography: Using the Rényi Divergence Rather than the Statistical Distance*

The Rényi divergence is a measure of closeness of two probability distributions. In this paper [5], we show that it can often be used as an alternative to the statistical distance in security proofs for lattice-based cryptography. Using the Rényi divergence is particularly suited for security proofs of primitives in which the attacker is required to solve a search problem (e.g., forging a signature). We show that it may also be used in the case of distinguishing problems (e.g., semantic security of encryption schemes), when they enjoy a public sampleability property. The techniques lead to security proofs for schemes with smaller parameters, and sometimes to simpler security proofs than the existing ones.

### *7.3.10. CRYSTALS-Dilithium: A Lattice-Based Digital Signature Scheme*

This paper [8] presents Dilithium, a lattice-based signature scheme that is part of the CRYSTALS (Cryptographic Suite for Algebraic Lattices) package that will be submitted to the NIST call for post-quantum standards. The scheme is designed to be simple to securely implement against side-channel attacks and to have

comparable efficiency to the currently best lattice-based signature schemes. Our implementation results show that Dilithium is competitive with lattice schemes of the same security level and outperforms digital signature schemes based on other post-quantum assumptions.

### *7.3.11. On the asymptotic complexity of solving LWE*

In this paper [9], we provide for the first time an asymptotic comparison of all known algorithms for the search version of the Learning with Errors (LWE) problem. This includes an analysis of several lattice-based approaches as well as the combinatorial BKW algorithm. Our analysis of the lattice-based approaches defines a general framework, in which the algorithms of Babai, Lindner–Peikert and several pruning strategies appear as special cases. We show that within this framework, all lattice algorithms achieve the same asymptotic complexity. For the BKW algorithm, we present a refined analysis for the case of only a polynomial number of samples via amplification, which allows for a fair comparison with lattice-based approaches. Somewhat surprisingly, such a small number of samples does not make the asymptotic complexity significantly inferior, but only affects the constant in the exponent. As the main result we obtain that both, lattice-based techniques and BKW with a polynomial number of samples, achieve running time $2^{O(n)}$ for $n$-dimensional LWE, where we make the constant hidden in the big-$O$ notion explicit as a simple and easy to handle function of all LWE-parameters. In the lattice case this function also depends on the time to compute a BKZ lattice basis with block size $\Theta(n)$. Thus, from a theoretical perspective our analysis reveals how LWE 's complexity changes as a function of the LWE-parameters, and from a practical perspective our analysis is a useful tool to choose LWE-parameters resistant to all currently known attacks.

### *7.3.12. Measuring, Simulating and Exploiting the Head Concavity Phenomenon in BKZ*

The Blockwise-Korkine-Zolotarev (BKZ) lattice reduction algorithm is central in cryptanalysis, in particular for lattice-based cryptography. A precise understanding of its practical behavior in terms of run-time and output quality is necessary for parameter selection in cryptographic design. As the provable worst-case bounds poorly reflect the practical behavior, cryptanalysts rely instead on the heuristic BKZ simulator of Chen and Nguyen (Asiacrypt'11). It fits better with practical experiments, but not entirely. In particular, it over-estimates the norm of the first few vectors in the output basis. Put differently, BKZ performs better than its Chen-Nguyen simulation.

In this article [15], we first report experiments providing more insight on this shorter-than-expected phenomenon. We then propose a refined BKZ simulator by taking the distribution of short vectors in random lattices into consideration. We report experiments suggesting that this refined simulator more accurately predicts the concrete behavior of BKZ. Furthermore, we design a new BKZ variant that exploits the shorter-than-expected phenomenon. For the same cost assigned to the underlying SVP-solver, the new BKZ variant produces bases of better quality. We further illustrate its potential impact by testing it on the SVP-120 instance of the Darmstadt lattice challenge.

### *7.3.13. CRYSTALS - Kyber: A CCA-Secure Module-Lattice-Based KEM*

Rapid advances in quantum computing, together with the announcement by the National Institute of Standards and Technology (NIST) to define new standards for digital signature, encryption, and key-establishment protocols, have created significant interest in post-quantum cryptographic schemes. This paper [17] introduces Kyber (part of CRYSTALS - Cryptographic Suite for Algebraic Lattices - a package submitted to NIST post-quantum standardization effort in November 2017), a portfolio of post-quantum cryptographic primitives built around a key-encapsulation mechanism (KEM), based on hardness assumptions over module lattices. Our KEM is most naturally seen as a successor to the NEWHOPE KEM (Usenix 2016). In particular, the key and ciphertext sizes of our new construction are about half the size, the KEM offers CCA instead of only passive security, the security is based on a more general (and flexible) lattice problem, and our optimized implementation results in essentially the same running time as the aforementioned scheme. We first introduce a CPA-secure public-key encryption scheme, apply a variant of the Fujisaki-Okamoto transform to create a CCA-secure KEM, and eventually construct, in a black-box manner, CCA-secure encryption, key exchange,

and authenticated-key-exchange schemes. The security of our primitives is based on the hardness of Module-LWE in the classical and quantum random oracle models, and our concrete parameters conservatively target more than 128 bits of postquantum security.

### *7.3.14. Learning with Errors and Extrapolated Dihedral Cosets*

The hardness of the learning with errors (LWE) problem is one of the most fruitful resources of modern cryptography. In particular, it is one of the most prominent candidates for secure post-quantum cryptography. Understanding its quantum complexity is therefore an important goal. In this paper [20], we show that under quantum polynomial time reductions, LWE is equivalent to a relaxed version of the dihedral coset problem (DCP), which we call extrapolated DCP (eDCP). The extent of extrapolation varies with the LWE noise rate. By considering different extents of extrapolation, our result generalizes Regev's famous proof that if DCP is in BQP (quantum poly-time) then so is LWE (FOCS'02). We also discuss a connection between eDCP and Childs and Van Dam's algorithm for generalized hidden shift problems (SODA'07). Our result implies that a BQP solution for LWE might not require the full power of solving DCP, but rather only a solution for its relaxed version, eDCP, which could be easier.

### *7.3.15. Pairing-friendly twisted Hessian curves*

This paper [27] presents efficient formulas to compute Miller doubling and Miller addition utilizing degree-3 twists on curves with j-invariant 0 written in Hessian form. We give the formulas for both odd and even embedding degrees and for pairings on both $G_1 \times G_2$ and $G_2 \times G_1$. We propose the use of embedding degrees 15 and 21 for 128-bit and 192-bit security respectively in light of the NFS attacks and their variants. We give a comprehensive comparison with other curve models; our formulas give the fastest known pairing computation for embedding degrees 15, 21, and 24.

### *7.3.16. On the Statistical Leak of the GGH13 Multilinear Mapand some Variants*

At EUROCRYPT 2013, Garg, Gentry and Halevi proposed a candidate construction (later referred as GGH13) of cryptographic multilinear map (MMap). Despite weaknesses uncovered by Hu and Jia (EUROCRYPT 2016), this candidate is still used for designing obfuscators. The naive version of the GGH13 scheme was deemed susceptible to averaging attacks, i.e., it could suffer from a statistical leak (yet no precise attack was described). A variant was therefore devised, but it remains heuristic. Recently, to obtain MMaps with low noise and modulus, two variants of this countermeasure were developed by Döttling et al. (EPRINT:2016/599). In this work [28], we propose a systematic study of this statistical leak for all these GGH13 variants. In particular, we confirm the weakness of the naive version of GGH13. We also show that, among the two variants proposed by Döttling et al., the so-called conservative method is not so effective: it leaks the same value as the unprotected method. Luckily, the leak is more noisy than in the unprotected method, making the straightforward attack unsuccessful. Additionally, we note that all the other methods also leak values correlated with secrets. As a conclusion, we propose yet another countermeasure, for which this leak is made unrelated to all secrets. On our way, we also make explicit and tighten the hidden exponents in the size of the parameters, as an effort to assess and improve the efficiency of MMaps.

### *7.3.17. Higher dimensional sieving for the number field sieve algorithms*

Since 2016 and the introduction of the exTNFS (extended tower number field sieve) algorithm, the security of cryptosystems based on nonprime finite fields, mainly the pairing- and torus-based ones, is being reassessed. The feasibility of the relation collection, a crucial step of the NFS variants, is especially investigated. It usually involves polynomials of degree 1, i.e., a search space of dimension 2. However, exTNFS uses bivariate polynomials of at least four coefficients. If sieving in dimension 2 is well described in the literature, sieving in higher dimensions has received significantly less attention. In this work [30], we describe and analyze three different generic algorithms to sieve in any dimension for the NFS algorithms. Our implementation shows the practicability of dimension-4 sieving, but the hardness of dimension-6 sieving.

### 7.3.18. Speed-Ups and Time-Memory Trade-Offs for Tuple Lattice Sieving

In this work [31], we study speed-ups and time–space trade-offs for solving the shortest vector problem (SVP) on Euclidean lattices based on tuple lattice sieving. Our results extend and improve upon previous work of Bai–Laarhoven–Stehlé [ANTS'16] and Herold–Kirshanova [PKC'17], with better complexities for arbitrary tuple sizes and offering tunable time–memory tradeoffs. The trade-offs we obtain stem from the generalization and combination of two algorithmic techniques: the configuration framework introduced by Herold–Kirshanova, and the spherical locality-sensitive filters of Becker–Ducas–Gama–Laarhoven [SODA'16]. When the available memory scales quasi-linearly with the list size, we show that with triple sieving we can solve SVP in dimension $n$ in time $2^{0.3588n+o(n)}$ and space $2^{0.1887n+o(n)}$, improving upon the previous best triple sieve time complexity of $2^{0.3717n+o(n)}$ of Herold–Kirshanova. Using more memory we obtain better asymptotic time complexities. For instance, we obtain a triple sieve requiring only $2^{0.3300n+o(n)}$ time and $2^{0.2075n+o(n)}$ memory to solve SVP in dimension $n$. This improves upon the best double Gauss sieve of Becker–Ducas–Gama–Laarhoven, which runs in $2^{0.3685n+o(n)}$ time when using the same amount of space.

### 7.3.19. Improved Quantum Information Set Decoding

In this paper [34], we present quantum information set decoding (ISD) algorithms for binary linear codes. First, we refine the analysis of the quantum walk based algorithms proposed by Kachigar and Tillich (PQCrypto'17). This refinement allows us to improve the running time of quantum decoding in the leading order term: for an n-dimensional binary linear code the complexity of May-Meurer-Thomae ISD algorithm (Asiacrypt'11) drops down from $2^{0.05904n+o(n)}$ to $2^{0.05806n+o(n)}$. Similar improvement is achieved for our quantum version of Becker-JeuxMay-Meurer (Eurocrypt'12) decoding algorithm. Second, we translate May-Ozerov Near Neighbour technique (Eurocrypt'15) to an 'updateand-query' language more common in a similarity search literature. This re-interpretation allows us to combine Near Neighbour search with the quantum walk framework and use both techniques to improve a quantum version of Dumer's ISD algorithm: the running time goes down from $2^{0.059962n+o(n)}$ to $2^{0.059450+o(n)}$.

### 7.3.20. Quantum Attacks against Indistinguishablility Obfuscators Proved Secure in the Weak Multilinear Map Model

We present in [39] a quantum polynomial time attack against the GMMSSZ branching program obfuscator of Garg et al. (TCC'16), when instantiated with the GGH13 multilinear map of Garg et al. (EUROCRYPT'13). This candidate obfuscator was proved secure in the weak multilinear map model introduced by Miles et al. (CRYPTO'16). Our attack uses the short principal ideal solver of Cramer et al. (EUROCRYPT'16), to recover a secret element of the GGH13 multilinear map in quantum polynomial time. We then use this secret element to mount a (classical) polynomial time mixed-input attack against the GMMSSZ obfuscator. The main result of this article can hence be seen as a classical reduction from the security of the GMMSSZ obfuscator to the short principal ideal problem (the quantum setting is then only used to solve this problem in polynomial time). As an additional contribution, we explain how the same ideas can be adapted to mount a quantum polynomial time attack against the DGGMM obfuscator of Döttling et al. (ePrint 2016), which was also proved secure in the weak multilinear map model.

### 7.3.21. On the Ring-LWE and Polynomial-LWE Problems

The Ring Learning With Errors problem (RLWE) comes in various forms. Vanilla RLWE is the decision dual-RLWE variant, consisting in distinguishing from uniform a distribution depending on a secret belonging to the dual $O_K^\vee$ of the ring of integers $O_K$ of a specified number field $K$. In primal-RLWE, the secret instead belongs to $O_K$. Both decision dual-RLWE and primal-RLWE enjoy search counterparts. Also widely used is (search/decision) Polynomial Learning With Errors (PLWE), which is not defined using a ring of integers $O_K$ of a number field $K$ but a polynomial ring $Z[x]/f$ for a monic irreducible $f \in Z[x]$. We show that there exist reductions between all of these six problems that incur limited parameter losses. More precisely: we prove that the (decision/search) dual to primal reduction from Lyubashevsky et al. [EUROCRYPT 2010] and Peikert [SCN 2016] can be implemented with a small error rate growth for all rings (the resulting reduction is nonuniform polynomial time); we extend it to polynomial-time reductions between (decision/search) primal

RLWE and PLWE that work for a family of polynomials $f$ that is exponentially large as a function of $\deg(f)$ (the resulting reduction is also non-uniform polynomial time); and we exploit the recent technique from Peikert et al. [STOC 2017] to obtain a search to decision reduction for RLWE for arbitrary number fields. The reductions incur error rate increases that depend on intrinsic quantities related to $K$ and $f$.

### 7.3.22. Non-Trivial Witness Encryption and Null-iO from Standard Assumptions

A *witness encryption (WE)* scheme can take any NP statement as a public-key and use it to encrypt a message. If the statement is true then it is possible to decrypt the message given a corresponding witness, but if the statement is false then the message is computationally hidden. Ideally, the encryption procedure should run in polynomial time, but it is also meaningful to define a weaker notion, which we call *non-trivially exponentially efficient* WE (XWE), where the encryption run-time is only required to be much smaller than the trivial $2^m$ bound for NP relations with witness size $m$. In [19], we show how to construct such XWE schemes for all of NP with encryption run-time $2^{m/2}$ under the sub-exponential learning with errors (LWE) assumption. For NP relations that can be verified in $\mathsf{NC}^1$ (e.g., SAT) we can also construct such XWE schemes under the sub-exponential Decisional Bilinear Diffie-Hellman (DBDH) assumption. Although we find the result surprising, it follows via a very simple connection to *attribute-based encryption*.

We also show how to upgrade the above results to get non-trivially exponentially efficient *indistinguishability obfuscation for null circuits (niO)*, which guarantees that the obfuscations of any two circuits that always output 0 are indistinguishable. In particular, under the LWE assumptions we get a XniO scheme where the obfuscation time is $2^{n/2}$ for all circuits with input size $n$. It is known that in the case of indistinguishability obfuscation (iO) for all circuits, non-trivially efficient XiO schemes imply fully efficient iO schemes (Lin et al., PKC 2016) but it remains as a fascinating open problem whether any such connection exists for WE or niO.

Lastly, we explore a potential approach toward constructing fully efficient WE and niO schemes via multi-input ABE.

### 7.3.23. Function-Revealing Encryption

Multi-input functional encryption is a paradigm that allows an authorized user to compute a certain function—and nothing more—over multiple plaintexts given only their encryption. The particular case of two-input functional encryption has very exciting applications, including comparing the relative order of two plaintexts from their encrypted form (order-revealing encryption).

While being extensively studied, multi-input functional encryption is not ready for a practical deployment, mainly for two reasons. First, known constructions rely on heavy cryptographic tools such as multilinear maps. Second, their security is still very uncertain, as revealed by recent devastating attacks.

In [33], we investigate a simpler approach towards obtaining practical schemes for functions of particular interest. We introduce the notion of function-revealing encryption, a generalization of order-revealing encryption to any multi-input function as well as a relaxation of multi-input functional encryption. We then propose a simple construction of order-revealing encryption based on function-revealing encryption for simple functions, namely orthogonality testing and intersection cardinality. Our main result is an efficient order-revealing encryption scheme with limited leakage based on the standard DLin assumption.

### 7.3.24. Exploring Crypto Dark Matter: New Simple PRF Candidates and Their Applications

Pseudorandom functions (PRFs) are one of the fundamental building blocks in cryptography. We explore a new space of plausible PRF candidates that are obtained by mixing linear functions over different small moduli. Our candidates are motivated by the goals of maximizing simplicity and minimizing complexity measures that are relevant to cryptographic applications such as secure multiparty computation.

In [16], we present several concrete new PRF candidates that follow the above approach. Our main candidate is a *weak* PRF candidate (whose conjectured pseudorandomness only holds for uniformly random inputs) that first applies a secret mod-2 linear mapping to the input, and then a public mod-3 linear mapping to the result. This candidate can be implemented by depth-2 $\mathsf{ACC}^0$ circuits. We also put forward a similar depth-3 *strong*

PRF candidate. Finally, we present a different weak PRF candidate that can be viewed as a deterministic variant of "Learning Parity with Noise" (LPN) where the noise is obtained via a mod-3 inner product of the input and the key.

The advantage of our approach is twofold. On the theoretical side, the simplicity of our candidates enables us to draw natural connections between their hardness and questions in complexity theory or learning theory (e.g., learnability of depth-2 $ACC^0$ circuits and width-3 branching programs, interpolation and property testing for sparse polynomials, and natural proof barriers for showing super-linear circuit lower bounds). On the applied side, the "piecewise-linear" structure of our candidates lends itself nicely to applications in secure multiparty computation (MPC). Using our PRF candidates, we construct protocols for distributed PRF evaluation that achieve better round complexity and/or communication complexity (often both) compared to protocols obtained by combining standard MPC protocols with PRFs like AES, LowMC, or Rasta (the latter two are specialized MPC-friendly PRFs). Our advantage over competing approaches is maximized in the setting of MPC with an honest majority, or alternatively, MPC with preprocessing.

Finally, we introduce a new primitive we call an *encoded-input PRF*, which can be viewed as an interpolation between weak PRFs and standard (strong) PRFs. As we demonstrate, an encoded-input PRF can often be used as a drop-in replacement for a strong PRF, combining the efficiency benefits of weak PRFs and the security benefits of strong PRFs. We conclude by showing that our main weak PRF candidate can plausibly be boosted to an encoded-input PRF by leveraging error-correcting codes.

### *7.3.25. Related-Key Security for Pseudorandom Functions Beyond the Linear Barrier*

Related-key attacks (RKAs) concern the security of cryptographic primitives in the situation where the key can be manipulated by the adversary. In the RKA setting, the adversary's power is expressed through the class of related-key deriving (RKD) functions which the adversary is restricted to using when modifying keys. Bellare and Kohno (Eurocrypt 2003) first formalised RKAs and pin-pointed the foundational problem of constructing RKA-secure pseudorandom functions (RKA-PRFs). To date there are few constructions for RKA-PRFs under standard assumptions, and it is a major open problem to construct RKA-PRFs for larger classes of RKD functions. We make significant progress on this problem. In [3], we first show how to repair the Bellare-Cash framework for constructing RKA-PRFs and extend it to handle the more challenging case of classes of RKD functions that contain claws. We apply this extension to show that a variant of the NaorReingold function already considered by Bellare and Cash is an RKA-PRF for a class of affine RKD functions under the DDH assumption, albeit with an exponential-time security reduction. We then develop a second extension of the Bellare-Cash framework, and use it to show that the same Naor-Reingold variant is actually an RKA-PRF for a class of degree d polynomial RKD functions under the stronger decisional d-Diffie-Hellman inversion assumption. As a significant technical contribution, our proof of this result avoids the exponential-time security reduction that was inherent in the work of Bellare and Cash and in our first result.

### *7.3.26. Practical Fully Secure Unrestricted Inner Product Functional Encryption modulo $p$*

In [23], we provide adaptively secure functional encryption schemes for the inner product functionality which are both efficient and allow for the evaluation of unbounded inner products modulo a prime p. Our constructions rely on new natural cryptographic assumptions in a cyclic group containing a subgroup where the discrete logarithm (DL) problem is easy which extend Castagnos and Laguillaumie's assumption (RSA 2015) of a DDH group with an easy DL subgroup. Instantiating our generic construction using class groups of imaginary quadratic fields gives rise to the most efficient functional encryption for inner products modulo an arbitrary large prime p. One of our schemes outperforms the DCR variant of Agrawal et al.'s protocols in terms of size of keys and ciphertexts by factors varying between 2 and 20 for a 112-bit security.

## 7.4. Algebraic computing and high-performance kernels

### *7.4.1. Generalized Hermite Reduction, Creative Telescoping and Definite Integration of D-Finite Functions*

Hermite reduction is a classical algorithmic tool in symbolic integration. It is used to decompose a given rational function as a sum of a function with simple poles and the derivative of another rational function. In [18], we extend Hermite reduction to arbitrary linear differential operators instead of the pure derivative, and develop efficient algorithms for this reduction. We then apply the generalized Hermite reduction to the computation of linear operators satisfied by single definite integrals of D-finite functions of several continuous or discrete parameters. The resulting algorithm is a generalization of reduction-based methods for creative telescoping.

### 7.4.2. Hermite-Padé approximant bases

In [46] we design fast algorithms for the computation of approximant bases in shifted Popov normal form. For $\mathsf{K}$ a commutative field, let $F$ be a matrix in $\mathsf{K}[x]^{m\times n}$ (truncated power series) and $\vec{d}$ be a degree vector, the problem is to compute a basis $P \in \mathsf{K}[x]^{m\times m}$ of the $\mathsf{K}[x]$-module of the relations $p \in \mathsf{K}[x]^{1\times m}$ such that $p(x)\cdot F(x) \equiv 0 \mod x^{\vec{d}}$. We obtain improved complexity bounds for handling arbitrary (possibly highly unbalanced) vectors $\vec{d}$. We also improve upon previously known algorithms for computing $P$ in normalized shifted form for an arbitrary shift. Our approach combines a recent divide and conquer strategy which reduces the general case to the case where information on the output degree is available, and partial linearizations of the involved matrices.

### 7.4.3. Resultant of bivariate polynomials

We have proposed in [42] an algorithm for computing the resultant of two generic bivariate polynomials over a field $\mathsf{K}$. For such $p$ and $q$ in $\mathsf{K}[x,y]$ both of degree $d$ in $x$ and $n$ in $y$, the algorithm computes the resultant with respect to $y$ using $(n^{2-1/\omega}d)^{1+o(1)}$ arithmetic operations, where $\omega$ is the exponent of matrix multiplication. Previous algorithms from the early 1970's required time $(n^2d)^{1+o(1)}$. We have also described some extensions of the approach to the computation of generic Gröbner bases and of characteristic polynomials of generic structured matrices and in univariate quotient algebras.

### 7.4.4. Recursive Combinatorial Structures: Enumeration, Probabilistic Analysis and Random Generation

The probabilistic behaviour of many data-structures, like series-parallel graphs used as a running example is this tutorial [13], can be analysed very precisely, thanks to a set of high-level tools provided by Analytic Combinatorics, as described in the book by Flajolet and Sedgewick. In this framework, recursive combinatorial definitions lead to generating function equations from which efficient algorithms can be designed for enumeration, random generation and, to some extent, asymptotic analysis. With a focus on random generation, this tutorial given at STACS first covers the basics of Analytic Combinatorics and then describes the idea of Boltzmann sampling and its realisation. The tutorial addresses a broad TCS audience and no particular pre-knowledge on analytic combinatorics is expected.

### 7.4.5. Linear Differential Equations as a Data-Structure

A lot of information concerning solutions of linear differential equations can be computed directly from the equation. It is therefore natural to consider these equations as a data-structure, from which mathematical properties can be computed. A variety of algorithms has thus been designed in recent years that do not aim at "solving", but at computing with this representation. Many of these results are surveyed in [11].

# AVALON Project-Team

# 7. New Results

## 7.1. Energy Efficiency in HPC and Large Scale Distributed Systems

**Participants:** Laurent Lefèvre, Dorra Boughzala, Christian Perez, Issam Raïs, Mathilde Boutigny.

### *7.1.1. Building and Exploiting the Table of Leverages in Large Scale HPC Systems*

Large scale distributed systems and supercomputers consume huge amounts of energy. To address this issue, an heterogeneous set of capabilities and techniques that we call leverages exist to modify power and energy consumption in large scale systems. This includes hardware related leverages (such as Dynamic Voltage and Frequency Scaling), middleware (such as scheduling policies) and application (such as the precision of computation) energy leverages. Discovering such leverages, benchmarking and orchestrating them, remains a real challenge for most of the users. We have formally defined energy leverages, and we proposed a solution to automatically build the table of leverages associated with a large set of independent computing resources. We have shown that the construction of the table can be parallelized at very large scale with a set of independent nodes in order to reduce its execution time while maintaining precision of observed knowledge [22], [25].

### *7.1.2. Automatic Energy Efficient HPC Programming: A Case Study*

Energy consumption is one of the major challenges of modern datacenters and supercomputers. By applying Green Programming techniques, developers have to iteratively implement and test new versions of their software, thus evaluating the impact of each code version on their energy, power and performance objectives. This approach is manual and can be long, challenging and complicated, especially for High Performance Computing applications. In [24], we formally introduces the definition of the Code Version Variability (CVV) leverage and present a first approach to automate Green Programming (*i.e.,* CVV usage) by studying the specific use-case of an HPC stencil-based numerical code, used in production. This approach is based on the automatic generation of code versions thanks to a Domain Specific Language (DSL), and on the automatic choice of code version through a set of actors. Moreover, a real case study is introduced and evaluated though a set of benchmarks to show that several trade-offs are introduced by CVV 1. Finally, different kinds of production scenarios are evaluated through simulation to illustrate possible benefits of applying various actors on top of the CVV automation.

### *7.1.3. Performance and Energy Analysis of OpenMP Runtime Systems with Dense Linear Algebra Algorithms*

In the article [9], we analyze performance and energy consumption of five OpenMP runtime systems over a non-uniform memory access (NUMA) platform. We also selected three CPU-level optimizations or techniques to evaluate their impact on the runtime systems: processors features Turbo Boost and C-States, and CPU Dynamic Voltage and Frequency Scaling through Linux CPUFreq governors. We present an experimental study to characterize OpenMP runtime systems on the three main kernels in dense linear algebra algorithms (Cholesky, LU, and QR) in terms of performance and energy consumption. Our experimental results suggest that OpenMP runtime systems can be considered as a new energy leverage, and Turbo Boost, as well as C-States, impacted significantly performance and energy. CPUFreq governors had more impact with Turbo Boost disabled, since both optimizations reduced performance due to CPU thermal limits. An LU factorization with concurrent-write extension from libKOMP achieved up to 63% of performance gain and 29% of energy decrease over original PLASMA algorithm using GNU C compiler (GCC) libGOMP runtime.

### *7.1.4. Energy Simulation of GPU based Infrastructures*

Through the IPL Hac-Specis and the PhD of Dorra Boughzala we begin to explore the modeling and calibrating of energy consumption of GPU architectures. We use the SimGrid simulation framework for the integration and validation on large scale systems.

## 7.2. HPC Component Models and Runtimes

**Participants:** Thierry Gautier, Christian Perez, Jérôme Richard.

### *7.2.1. On the Impact of OpenMP Task Granularity*

Tasks are a good support for composition. During the development of a high-level component model for HPC, we have experimented to manage parallelism from components using OpenMP tasks. Since version 4-0, the standard proposes a model with dependent tasks that seems very attractive because it enables the description of dependencies between tasks generated by different components without breaking maintainability constraints such as separation of concerns. In [20], we present our feedback on using OpenMP in our context. We discover that our main issues are a too coarse task granularity for our expected performance on classical OpenMP runtimes, and a harmful task throttling heuristic counter-productive for our applications. We present a completion time breakdown of task management in the Intel OpenMP runtime and propose extensions evaluated on a testbed application coming from the Gysela application in plasma physics.

### *7.2.2. Building and Auto-Tuning Computing Kernels: Experimenting with BOAST and StarPU in the GYSELA Code*

Modeling turbulent transport is a major goal in order to predict confinement performance in a tokamak plasma. The gyrokinetic framework considers a computational domain in five dimensions to look at kinetic issues in a plasma; this leads to huge computational needs. Therefore, optimization of the code is an especially important aspect, especially since coprocessors and complex manycore architectures are foreseen as building blocks for Exascale systems. This project [6] aims to evaluate the applicability of two auto-tuning approaches with the BOAST and StarPU tools on the gysela code in order to circumvent performance portability issues. A specific computation intensive kernel is considered in order to evaluate the benefit of these methods. StarPU enables to match the performance and even sometimes outperform the hand-optimized version of the code while leaving scheduling choices to an automated process. BOAST on the other hand reveals to be well suited to get a gain in terms of execution time on four architectures. Speedups in-between 1.9 and 5.7 are obtained on a cornerstone computation intensive kernel.

## 7.3. Modeling and Simulation of Parallel Applications and Distributed Infrastructures

**Participants:** Eddy Caron, Zeina Houmani, Frédéric Suter.

### *7.3.1. SMPI Courseware: Teaching Distributed-Memory Computing with MPI in Simulation*

It is typical in High Performance Computing (HPC) courses to give students access to HPC platforms so that they can benefit from hands-on learning opportunities. Using such platforms, however, comes with logistical and pedagogical challenges. For instance, a logistical challenge is that access to representative platforms must be granted to students, which can be difficult for some institutions or course modalities; and a pedagogical challenge is that hands-on learning opportunities are constrained by the configurations of these platforms. A way to address these challenges is to instead simulate program executions on arbitrary HPC platform configurations. In [15] we focus on simulation in the specific context of distributed-memory computing and MPI programming education. While using simulation in this context has been explored in previous works, our approach offers two crucial advantages. First, students write standard MPI programs and can both debug and analyze the performance of their programs in simulation mode. Second, large-scale executions can be simulated in short amounts of time on a single standard laptop computer. This is possible thanks to SMPI, an MPI simulator provided as part of SimGrid. After detailing the challenges involved when using HPC platforms for HPC education and providing background information about SMPI, we present SMPI Courseware. SMPI Courseware is a set of in-simulation assignments that can be incorporated into HPC courses to provide students with hands-on experience for distributed-memory computing and MPI programming learning objectives. We describe some these assignments, highlighting how simulation with SMPI enhances the student learning experience.

### 7.3.2. Evaluation through Realistic Simulations of File Replication Strategies for Large Heterogeneous Distributed Systems

File replication is widely used to reduce file transfer times and improve data availability in large distributed systems. Replication techniques are often evaluated through simulations, however, most simulation platform models are oversimplified, which questions the applicability of the findings to real systems. In [17], we investigate how platform models influence the performance of file replication strategies on large heterogeneous distributed systems, based on common existing techniques such as prestaging and dynamic replication. The novelty of our study resides in our evaluation using a realistic simulator. We consider two platform models: a simple hierarchical model and a detailed model built from execution traces. Our results show that conclusions depend on the modeling of the platform and its capacity to capture the characteristics of the targeted production infrastructure. We also derive recommendations for the implementation of an optimized data management strategy in a scientific gateway for medical image analysis.

### 7.3.3. WRENCH: Workflow Management System Simulation Workbench

Scientific workflows are used routinely in numerous scientific domains, and Workflow Management Systems (WMSs) have been developed to orchestrate and optimize workflow executions on distributed platforms. WMSs are complex software systems that interact with complex software infrastructures. Most WMS research and development activities rely on empirical experiments conducted with full-fledged software stacks on actual hardware platforms. Such experiments, however, are limited to hardware and software infrastructures at hand and can be labor- and/or time-intensive. As a result, relying solely on real-world experiments impedes WMS research and development. An alternative is to conduct experiments in simulation.

In [16] we presented WRENCH, a WMS simulation framework, whose objectives are (i) accurate and scalable simulations; and (ii) easy simulation software development. WRENCH achieves its first objective by building on the SimGrid framework. While SimGrid is recognized for the accuracy and scalability of its simulation models, it only provides low-level simulation abstractions and thus large software development efforts are required when implementing simulators of complex systems. WRENCH thus achieves its second objective by providing high- level and directly re-usable simulation abstractions on top of SimGrid. After describing and giving rationales for WRENCH's software architecture and APIs, we present a case study in which we apply WRENCH to simulate the Pegasus production WMS. We report on ease of implementation, simulation accuracy, and simulation scalability so as to determine to which extent WRENCH achieves its two above objectives. We also draw both qualitative and quantitative comparisons with a previously proposed workflow simulator.

### 7.3.4. A Microservices Architectures for Data-Driven Service Discovery

Usual microservices discovery mechanisms are normally based on a specific user need *(Goal-based approaches)*. However, in today's evolving architectures, users need to discover what features they can take advantage of before looking for the available microservices. In collaboration with RDI2 (Rutgers University) we developped a data-driven microservices architecture that allows users to discover, from specific objects, the features that can be exerted on these objects as well as all the microservices dedicated to them [28]. This architecture, based on the main components of the usual microservices architectures, adopts a particular communication strategy between clients and registry to achieve the goal. This article contains a representation of a microservice data model and a P2P model that transforms our architecture into a robust and scalable system. Also, we designed a prototype to validate our approach using Istio library.

## 7.4. Cloud Resource Management

**Participants:** Eddy Caron, Hadrien Croubois, Jad Darrous, Christian Perez.

### 7.4.1. Nitro: Network-Aware Virtual Machine Image Management in Geo-Distributed Clouds

Recently, most large cloud providers, like Amazon and Microsoft, replicate their Virtual Machine Images (VMIs) on multiple geographically distributed data centers to offer fast service provisioning. Provisioning a service may require to transfer a VMI over the wide-area network (WAN) and therefore is dictated by the distribution of VMIs and the network bandwidth in-between sites. Nevertheless, existing methods to facilitate VMI management (*i.e.,* retrieving VMIs) overlook network heterogeneity in geo-distributed clouds. In [19], we design, implement and evaluate Nitro, a novel VMI management system that helps to minimize the transfer time of VMIs over a heterogeneous WAN. To achieve this goal, Nitro incorporates two complementary features. First, it makes use of deduplication to reduce the amount of data which will be transferred due to the high similarities within an image and in-between images. Second, Nitro is equipped with a network-aware data transfer strategy to effectively exploit links with high bandwidth when acquiring data and thus expedites the provisioning time. Experimental results show that our network-aware data transfer strategy offers the optimal solution when acquiring VMIs while introducing minimal overhead. Moreover, Nitro outperforms state-of-the-art VMI storage systems (*e.g.,* OpenStack Swift) by up to 77%.

### 7.4.2. Toward an Autonomic Engine for Scientific Workflows and Elastic Cloud Infrastructure

The constant development of scientific and industrial computation infrastructures requires the concurrent development of scheduling and deployment mechanisms to manage such infrastructures. Throughout the last decade, the emergence of the Cloud paradigm raised many hopes, but achieving full platform autonomicity is still an ongoing challenge. We built a workflow engine that integrated the logic needed to manage workflow execution and Cloud deployment on its own. More precisely, we focus on Cloud solutions with a dedicated Data as a Service (DaaS) data management component. Our objective was to automate the execution of workflows submitted by many users on elastic Cloud resources. This contribution proposes a modular middleware infrastructure and details the implementation of the underlying modules:

- A workflow clustering algorithm that optimises data locality in the context of DaaS-centered communications;
- A dynamic scheduler that executes clustered workflows on Cloud resources;
- A deployment manager that handles the allocation and deallocation of Cloud resources according to the workload characteristics and users' requirements.

All these modules have been implemented in a simulator to analyse their behaviour and measure their effectiveness when running both synthetic and real scientific workflows. We also implemented these modules in the Diet middleware to give it new features and prove the versatility of this approach. Simulation running the WASABI workflow (waves analysis based inference, a framework for the reconstruction of gene regulatory networks) showed that our approach can decrease the deployment cost by up to 44% while meeting the required deadlines [13].

### 7.4.3. Madeus: A Formal Deployment Model

Distributed software architecture is composed of multiple interacting modules, or components. Deploying such software consists in installing them on a given infrastructure and leading them to a functional state. However, since each module has its own life cycle and might have various dependencies with other modules, deploying such software is a very tedious task, particularly on massively distributed and heterogeneous infrastructures. To address this problem, many solutions have been designed to automate the deployment process. In [18], we introduce Madeus, a component-based deployment model for complex distributed software. Madeus accurately describes the life cycle of each component by a Petri net structure, and is able to finely express the dependencies between components. The overall dependency graph it produces is then used to reduce deployment time by parallelizing deployment actions. While this increases the precision and performance of the model, it also increases its complexity. For this reason, the operational semantics need to be clearly defined to prove results such as the termination of a deployment. In this paper, we formally describe the operational semantics of Madeus, and show how it can be used in a use-case: the deployment of a real and large distributed software (*i.e.,* , OpenStack).

In [18], we have proposed an extension based on component behavioral interfaces to the Aeolus component model to better separate the concerns of component users (*e.g.,* application architect) from component developers.

## 7.5. Data Stream Processing on Edge Computing

**Participants:** Eddy Caron, Felipe Rodrigo de Souza, Marcos Dias de Assunção, Laurent Lefèvre, Alexandre Da Silva Veith.

### *7.5.1. Latency-Aware Placement of Data Stream Analytics on Edge Computing*

The interest in processing data events under stringent time constraints as they arrive has led to the emergence of architecture and engines for data stream processing. Edge computing, initially designed to minimize the latency of content delivered to mobile devices, can be used for executing certain stream processing operations. Moving opera- tors from cloud to edge, however, is challenging as operator-placement decisions must consider the application requirements and the network capabilities. We introduce strategies to create placement configurations for data stream processing applications whose operator topologies follow series parallel graphs[35]. We consider the operator characteristics and requirements to improve the response time of such appli- cations. Results show that our strategies can improve the response time in up to 50% for application graphs comprising multiple forks and joins while transferring less data and better using the resources.

### *7.5.2. Estimating Throughput of Stream Processing Applications in FoG Computing*

Recent trends exploit decentralized infrastructures (e.g.. Fog computing) to deploy DSP (Data Stream Processing) applications and leverage the computational power. Fog computing overlaps some features of Cloud computing and includes others, for instance, location awareness. The operator placement problem consists of determining, within a set of distributed computing resources, the computing resources that should host and execute each operator of the DSP application, with the goal of optimizing QoS requirements of the application. The QoS requirements of the application refer to processing time, costs, throughput, etc. We propose a model to estimate the application throughput at each layer of Fog computing (Devices, Edge and Cloud) by considering a given placement solution. The estimated throughput provides a useful insight to determine the amount of physical resources to meet the QoS requirements. The model allows to identify the application bottleneck, when facing data rate variations, and provides information to self-scale in or out the DSP application.

**BEAGLE Project-Team**

# 7. New Results

## 7.1. Dopamine interacts with endocannabinoids to regulate spike timing dependent plasticity

participants: H. Berry, I. Prokin

Dopamine modulates striatal synaptic plasticity, a key substrate for action selection and procedural learning. Thus, characterizing the repertoire of activity-dependent plasticity in striatum and its dependence on dopamine is of crucial importance. In collaboration with L. Venance Lab (CIRB, Collège de France) we recently unraveled a striatal spike-timing-dependent long-term potentiation (tLTP) mediated by endocannabinoids (eCBs) and induced with few spikes ( 5-15). Whether this eCB-tLTP interacts with the dopaminergic system remains to be investigated. We found that eCB-tLTP is impaired in a rodent model of Parkinson's disease and rescued by L-DOPA. Dopamine controls eCB-tLTP via dopamine type-2 receptors (D2R) located presynaptically in cortical terminals. Dopamine-endocannabinoid interactions via D2R are required for the emergence of tLTP in response to few coincident pre-and post-synaptic spikes and control eCB-plasticity by modulating the long-term potentiation (LTP)/depression (LTD) thresholds. While usually considered as a depressing synaptic function, our results show that eCBs in the presence of dopamine constitute a versatile system underlying bidirectional plasticity implicated in basal ganglia pathophysiology. These results have been published in Nature Communications [23]

## 7.2. Estimating the robustness of spike timing dependent plasticity to timing jitter

participants: H. Berry, I. Prokin

In Hebbian plasticity, neural circuits adjust their synaptic weights depending on patterned firing. Spike-timing-dependent plasticity (STDP), a synaptic Hebbian learning rule, relies on the order and timing of the paired activities in pre- and postsynaptic neurons. Classically, in ex vivo experiments, STDP is assessed with deterministic (constant) spike timings and time intervals between successive pairings, thus exhibiting a regularity that differs from biological variability. Hence, STDP emergence from noisy inputs as occurring in in vivo-like firing remains unresolved. In collaboration with the laboratories of L. Venance (CIRB, Collège de France) and A. De Kerchove d'Exaerde (Univ. Libre Bruxelles), we used noisy STDP pairings where the spike timing and/or interval between pairings were jittered. We explored with electrophysiology and mathematical modeling, the impact of jitter on three forms of STDP at corticostriatal synapses: NMDAR-LTP, endocannabinoid-LTD and endocannabinoid-LTP. We found that NMDAR-LTP was highly fragile to jitter, whereas endocannabinoid-plasticity appeared more resistant. When the frequency or number of pairings was increased, NMDAR-LTP became more robust and could be expressed despite strong jittering. Our results identify endocannabinoid-plasticity as a robust form of STDP, whereas the sensitivity to jitter of NMDAR-LTP varies with activity frequency. This provides new insights into the mechanisms at play during the different phases of learning and memory and the emergence of Hebbian plasticity in in vivo-like activity. These results have been published in Scientific Reports [14]

## 7.3. A new method to monitor gap junctional comunication in astrocytes

participants: H. Berry

Intercellular communication through gap junction channels plays a key role in cellular homeostasis and in synchronizing physiological functions, a feature that is modified in number of pathological situations. In the brain, astrocytes are the cell population that expresses the highest amount of gap junction proteins, named connexins. Several techniques have been used to assess the level of gap junctional communication in astrocytes, but so far they remain very difficult to apply in adult brain tissue. Using specific loading of astrocytes with sulforhodamine 101, we adapted in collaboration with C. Giaume's laboratory (CIRB, Collège de France) the gap-FRAP (Fluorescence Recovery After Photobleaching) to acute hippocampal slices from 9 month-old adult mice. We show that gap junctional communication monitored in astrocytes with this technique was inhibited either by pharmacological treatment with a gap junctional blocker or in mice lacking the two main astroglial connexins, while a partial inhibition was measured when only one connexin was knocked-out. We validate this approach using a mathematical model of sulforhodamine 101 diffusion in an elementary astroglial network and a quantitative analysis of the exponential fits to the fluorescence recovery curves. Consequently, we consider that the adaptation of the gap-FRAP technique to acute brain slices from adult mice provides an easy going and valuable approach that allows overpassing this age-dependent obstacle and will facilitate the investigation of gap junctional communication in adult healthy or pathological brain. These results have been published in J. Neuroscience Methods [24].

## 7.4. Kir4.1 upregulation in astrocytes of the lateral habenula is involved in depression

participants: H. Berry, A. Foncelle

Enhanced bursting activity of neurons in the lateral habenula (LHb) is essential in driving depression-like behaviours, but the cause of this increase has been unknown. In collaboration with H. Hu's laboratory (Zhejiang University, China), using a high-throughput quantitative proteomic screen, we show that an astroglial potassium channel (Kir4.1) is upregulated in the LHb in rat models of depression. Kir4.1 in the LHb shows a distinct pattern of expression on astrocytic membrane processes that wrap tightly around the neuronal soma. Electrophysiology and modelling data show that the level of Kir4.1 on astrocytes tightly regulates the degree of membrane hyperpolarization and the amount of bursting activity of LHb neurons. Astrocyte-specific gain and loss of Kir4.1 in the LHb bidirectionally regulates neuronal bursting and depression-like symptoms. Together, these results show that a glia–neuron interaction at the perisomatic space of LHb is involved in setting the neuronal firing mode in models of a major psychiatric disease. Kir4.1 in the LHb might have potential as a target for treating clinical depression. These results have been published in Nature [15] and were commented in the "News and views" section of the journal: Howe WM and Kenny PJ (2018). Burst firing sets the stage for depression.

## 7.5. The evolutionary complexity ratchet

participants: G Beslon, V Liard, D Parsons, Jonathan Rouzaud-Cornabas

Using the *in silico* experimental evolution platform Aevol, we evolved populations of digital organisms in conditions where a simple functional structure is best.

Strikingly, we observed that in a large fraction of the simulations, organisms evolved a complex functional structure and that their complexity increased during evolution despite being a lot less fit than simple organisms in other populations. However, when submitted to a harsh mutational pressure, we observed that a significant proportion of complex individuals ended up with a simple functional structure.

Our results suggest the existence of a complexity ratchet that is powered by epistasis and that cannot be beaten by selection. They also show that this ratchet can be overthrown by robustness because of the strong constraints it imposes on the coding capacity of the genome.

This result has been published in the International conference ALife in Tokyo (July 2018) where it received the best paper award [28]

## 7.6. Weight-based search to find clusters around medians in subspaces

participants: C Rigotti, G Beslon

There exist several clustering paradigms, leading to different techniques that are complementary in the analyst toolbox, each having its own merits and interests. Among these techniques, the K-medians approach is recognized as being robust to noise and outliers, and is an important optimization task with many different applications (e.g., facility location). In the context of subspace clustering, several paradigms have been investigated (e.g., centroid-based, cell-based), while the median-based approach has received less attention. Moreover, using standard subspace clustering outputs (e.g., centroids, medoids) there is no straightforward procedure to compute the cluster membership that optimizes the dispersion around medians. We advocated for the use of median-based subspace clustering as a complementary tool. Indeed, we showed that such an approach exhibits satisfactory quality clusters when compared to well-established paradigms, while medians have still their own interests depending on the user application (robustness to noise/outliers and location optimality). We showed that a weight-based hill climbing algorithm using a stochastic local exploration step can be sufficient to produce the clusters.

This research has been published in the proceedings of the ACM-SAC conference (Pau, March 2018) where it received the best paper award [26].

## 7.7. The surprising creativity of digital evolution

participants: C Knibbe, G Beslon

Natural evolution is a creative fount of complex adaptations that often surprise the scientists who discover them. However, the creativity of evolution is not limited to the natural world; artificial organisms evolving in computational environments are also able to elicit a similar degree of surprise and wonder from the researchers studying them. The process of evolution has proven to be an algorithmic process that transcends the substrate to which it is applied. Indeed, most digital evolution researchers can relate anecdotes highlighting how common it is for their algorithms to creatively subvert their expectations or intentions, expose unrecognized bugs in their code, produce unexpectedly potent adaptations, or engage in behaviors and outcomes uncannily convergent with ones found in nature. Such stories routinely reveal surprise and creativity by evolution in these digital worlds, but they rarely fit into the standard scientific narrative and are thus often treated as obstacles to be overcome rather than results that warrant publication in their own right. Bugs are fixed, experiments are refocused, one-off surprises are collapsed 1into a single data point. The stories themselves are traded among researchers through oral tradition, but that mode of information transmission is lossy, inefficient and error-prone. Moreover, the very fact that these stories tend to be confined to practitioners means that many natural scientists do not recognize how lifelike digital organisms are and how natural their evolution can be. We actively participated to a crowd-sourced research in which evolutionary computation researchers providing first-hand reports of such cases, and thus functions as a written, fact-checked collection of entertaining and important stories.

## 7.8. HPC support for Aevol

participants: Jonathan Rouzaud-Cornabas, David Parsons, Guillaume Beslon

During the year, we had three internships that focus around HPC. The three of them were founded through the Federation Informatique de Lyon (FIL FR2000) and were common between the Inria Beagle team (LIRIS) and the Inria Avalon team (LIP).

The first one (Lukas Schmidt - M2) was working on component-based software engineering and HPC with Aevol as use-case. The goal was to see if and how the COMET [1] task-based parallel component model (and its implementation Halley) can fit the parallelization requirement of Aevol. An extension of the model was proposed to support hierarchical data structure and a prototype implementation has been done. In the future, we will work on the formalization of the extension and an efficient implementation on it. The goal is to ease the development and replacement of core components of the Aevol software (e.g. be able to easily replace the 2-base DNA code by a 4-base one).

The second internship (Valentin Huguet - M2) was evolving around Aevol and how to ease the distribution of the computation. To do so, an extension of the DIET software [2] was proposed and a fully functional webboard was implemented. We have a first prototype that support the execution of a large set of distributed computing resources and the control of its execution through a webboard. Moreover, basic visualization of the simulation results can be done through the same webboard. A following internship (starting Feb. 2019) will continue the work. The goal is to support workflow composed of multiple execution of Aevol and its pre/post treatments to automate the execution of large campaign that are done manually at the moment.

The goal of the third internship (Nathan Payre - L3) was to propose a prototype of a bitset for Aevol and its efficient implementation on modern hardware (Intel Skylake and Intel Xeon Phi). Indeed, the current implementation of Aevol DNA (2 base) uses a char type (8bit) to store a bit value (0 or 1). Accordingly, working at the bitset level could save up to 8 time memory space and speed up the computation (as Aevol is memory bound, reducing the memory transfer by 8 could dramatically speed up the global execution). Moreover, modern processors have vectorization extension that are perfectly fitting our requirements (we could process 512bit per cycle with AVX512 extension). During the internship, the bitset and the different operation we use in Aevol model (e.g. Hamming distance) were formalized and implemented. The preliminary results show a speed up of 140x on these operations. A full evaluation on the impact of the performance of Aevol and how different modern processor react to such implementation will be done in the future.

Last, a part of the Beagle team (Guillaume Beslon, David Parsons, Jonathan Rouzaud-Cornabas) were selected and participate to the EuroHack 2018 GPU Programming Hackathon in Lugano (Switzerland) organized by CSCS (Swiss National Supercomputing Centre) and NVidia. The goal was to port Aevol to modern GPU and thus to the CUDA programming language. In order to be able to do so in a week, we propose a mini-application (mini-Aevol) of Aevol [3] that is representative of the computation and memory pattern of the full Aevol. This prototype will be reuse in our collaboration with team focusing on HPC research. At the end of the week, we had a full implementation of mini-Aevol on GPU. New core algorithms of Aevol have been proposed to support massively parallel processors such as GPU. The prototype will be transfer to the full Aevol code in the future to be able to support GPU. It is worth noting that this mini-apps is also used in teaching context (INSA Lyon - Computer Science M2) to learn how to parallelize and optimize code with OpenMP and CUDA.

[1] Olivier Aumage, Julien Bigot, Hélène Coullon, Christian Pérez, Jérôme Richard. Combining Both a Component Model and a Task-based Model for HPC Applications: a Feasibility Study on GYSELA. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)., May 2017, Madrid, Spain.

[2] https://graal.ens-lyon.fr/diet/

[3] https://gitlab.inria.fr/rouzaudc/mini-aevol

[4] https://github.com/fomics/EuroHack18/blob/master/final/beagle_aevol.pdf

## 7.9. Exploring the evolution of chromatin-associated proteins

participants: A Crombach

Eukaryotic gene regulation depends strongly on chromatin state. High-throughput studies in the fruit fly Drosophila melanogaster have shown that instead of the canonical two types of chromatin, hetero- and euchromatin, one can subdivide chromatin into five states. These states are each characterized by a unique combination of chromatin-associated proteins (CAPs). We were interested in the evolution of CAPs and studied them by means of phylogenomic methods. We found three evolutionary trends. One type of heterochomatin, called GREEN, is specific to centromeres and some of its proteins are found to be under a Red Queen type evolution, where they rapidly accumulate amino acid changes. The second type of heterochromatin, BLUE, is tightly linked to Polycomb Group proteins. These proteins are important regulators in developmental processes and our findings confirm their origin in multicellular organisms. Finally, the two euchromatic types, YELLOW and RED, have strong lineage-specific characteristics. Their origins seem to date back to the start of eukaryotic life.

## 7.10. Evolutionary interplay of genome content and 3D spatial structure

participants: A Crombach
Genomes are hierarchically folded, which involves transposable elements (TEs). The most prominently observed folding domains are conserved between cell types and across species, yet their building blocks, TEs, are powerful mutagens. This paradox raises the question why we observe evolutionary stable folding domains. Using in silico evolution of polymer genomes, the aim is to elucidate the interplay between mutations and folding structure. We have built the software and are in the process of generating data. First results indicate that due to accessibility in 3D (some parts of the genome are more tightly compacted than others), a positive feedback is created between (1) where mutations happen, (2) how genome content is changed, and (3) how genomes fold in 3D.

## 7.11. Network inference for mammalian cortex development

participants: A Crombach
The mammalian cortex divides into two major regions, neocortex (NCx) and the structurally simpler allocortex. Whereas NCx is well-characterized, the allocortex is much less studied. Its best known region is the olfactory (piriform, PCx) cortex. The regions have a laminar structure, with distinct neuronal cell types in each of the layers: NCx has 6 layers, PCx has 3. The differentiation of precursor cells into various neuronal cell types determines to which layer these cells will migrate. This process is mostly studied in NCx and depends on the activity of 10–20 developmental genes. In PCx the same genes are used, yet they appear in other combinations and may indicate diverse target layers, sometimes violating rules-of-thumb derived from NCx. Current understanding is rather incomplete with respect to how cortical neurons are specified. We propose that, despite apparent contradictions, a single gene network can explain the development of distinct cortical regions.

In collaboration with Dr. A. Fleischmann at Brown University (USA), we are measuring the expression of genes involved in neurodevelopment at cellular resolution using light-sheet microscopy. These data will form the basis for the inference of a regulatory network describing neuronal differentiation in NCx and PCx. Inference is done by fitting mathematical models of gene regulation to the data using global optimization methods. Currently, we are processing the image data. Moreover, single cell RNA sequencing will allow the study of the temporal dynamics of the expression of these genes and many others. We are completing an in-depth statistical analysis of the resulting genome-wide expression data.

## 7.12. Gene transfers can date the tree of life

participants: E Tannier
Biodiversity has always been predominantly microbial, and the scarcity of fossils from bacteria, archaea and microbial eukaryotes has prevented a comprehensive dating of the tree of life. Here, we show that patterns of lateral gene transfer deduced from an analysis of modern genomes encode a novel and abundant source of information about the temporal coexistence of lineages throughout the history of life. We use state-of-the-art species tree-aware phylogenetic methods to reconstruct the history of thousands of gene families and demonstrate that dates implied by gene transfers are consistent with estimates from relaxed molecular clocks in Bacteria, Archaea and Eukarya. We present the order of speciations according to lateral gene transfer data calibrated to geological time for three datasets comprising 40 genomes for Cyanobacteria, 60 genomes for Archaea and 60 genomes for Fungi. An inspection of discrepancies between transfers and clocks and a comparison with mammalian fossils show that gene transfer in microbes is potentially as informative for dating the tree of life as the geological record in macroorganisms. [16]

## 7.13. The devil in the details of evolvability

Participants: E Tannier, P Biller, V Liard, G Beslon

The theory of Evolvability consists in studying the evolution of living organisms as a computational learning process. It defines the possibilities of a population under Darwinian selection, to evolve in a certain direction, in a reasonable amount of time. While its robustness to certain parameters has been theoretically assessed, this theory has not been experimentally tested. We use a standard *in silico* experimental evolution tool to compare some predictions of the theory and the behavior of digital populations designed to resemble biological organisms. We obtain that the evolvability of monotone conjunctions under the uniform distribution of environmental conditions, presented as a major result of the theory, is not reproduced by the experiments. We show that this is due to different mutation algorithms, by a proof of exponential expectation time to target under theoretical conditions closer to the experiments. We examine into detail the choices of mutation algorithms. In the Evolvability theory it is any Turing machine, while much more restricted in the experimental design. This definition allows a wider range of conditions and in a certain way is conform to biological reality, where mutators evolve and can be selected. However it also allows, if it is misused, for the inclusion of oracles that are incompatible with the principles of a Darwinian evolution. Unfortunately these oracles are extensively used in the current evolvability proofs.

**CASH Team**

# 7. New Results

## 7.1. Monoparametric Tiling of Polyhedral Programs

**Participant:** Christophe Alias.

Tiling is a crucial program transformation, adjusting the ops-to-bytes balance of codes to improve locality. Like parallelism, it can be applied at multiple levels. Allowing tile sizes to be symbolic parameters at compile time has many benefits, including efficient autotuning, and run-time adaptability to system variations. For polyhedral programs, parametric tiling in its full generality is known to be non-linear, breaking the mathematical closure properties of the polyhedral model. Most compilation tools therefore either perform fixed size tiling, or apply parametric tiling in only the final, code generation step.

We introduce monoparametric tiling, a restricted parametric tiling transformation. We show that, despite being parametric, it retains the closure properties of the polyhedral model. We first prove that applying monoparametric partitioning (i) to a polyhedron yields a union of polyhedra, and (ii) to an affine function produces a piecewise-affine function. We then use these properties to show how to tile an entire polyhedral program. Our monoparametric tiling is general enough to handle tiles with arbitrary tile shapes that can tesselate the iteration space (e.g., hexagonal, trapezoidal, etc). This enables a wide range of polyhedral analyses and transformations to be applied.

This is a joint work with Guillaume Iooss (Inria Parkas) and Sanjay Rajopadhye (Colorado State University).

This work is under submission [12].

## 7.2. Improving Communication Patterns in Polyhedral Process Networks

**Participant:** Christophe Alias.

Process networks are a natural intermediate representation for HLS and more generally automatic parallelization. Compiler optimizations for parallelism and data locality restructure deeply the execution order of the processes, hence the read/write patterns in communication channels. This breaks most FIFO channels, which have to be implemented with addressable buffers. Expensive hardware is required to enforce synchronizations, which often results in dramatic performance loss. In this paper, we present an algorithm to partition the communications so that most FIFO channels can be recovered after a loop tiling, a key optimization for parallelism and data locality. Experimental results show a drastic improvement of FIFO detection for regular kernels at the cost of a few additional storage. As a bonus, the storage can even be reduced in some cases.

This work has been published in the HIP3ES workshop [1]

## 7.3. FIFO Recovery by Depth-Partitioning is Complete on Data-aware Process Networks

**Participant:** Christophe Alias.

In this paper, we build on our algorithm for FIFO recovery based on depth partitioning. We describe a class of process networks where the algorithm can recover all the FIFO channels. We point out the limitations of the algorithm outside of that class. Experimental results confirm the completeness of the algorithm on the class and reveal good performance outside of the class.

This work is under submission [9]

## 7.4. Parallel code generation of synchronous programs for a many-core architecture

**Participant:** Matthieu Moy.

Embedded systems tend to require more and more computational power. Many-core architectures are good candidates since they offer power and are considered more time predictable than classical multi-cores. Data-flow Synchronous languages such as Lustre or Scade are widely used for avionic critical software. Programs are described by networks of computational nodes. Implementation of such programs on a many-core architecture must ensure a bounded response time and preserve the functional behavior by taking interference into account. We consider the top-level node of a Lustre application as a software architecture description where each sub-node corresponds to a potential parallel task. Given a mapping (tasks to cores), we automatically generate code suitable for the targeted many-core architecture. This minimizes memory interferences and allows usage of a framework to compute the Worst-Case Response Time.

This is a joint work with Amaury Graillat, Pascal Raymond (IMAG) and Benoît Dupont de Dinechin (Kalray).

This work has been published at the DATE conference [6].

## 7.5. Estimation of the Impact of Architectural and Software Design Choices on Dynamic Allocation of Heterogeneous Memories

**Participant:** Matthieu Moy.

Reducing energy consumption is a key challenge to the realization of the Internet of Things. While emerging memory technologies may offer power reduction, they come with major drawbacks such as high latency or limited endurance. As a result, system designers tend to juxtapose several memory technologies on the same chip. This paper studies the interactions between dynamic memory allocation and architectural choices regarding this heterogeneity. We provide cycle accurate simulations of embedded platforms with various memory technologies and we show that different dynamic allocation strategies have a major impact on performance. We demonstrate that interesting performance gains can be achieved even for a low fraction of heap objects in fast memory, but only with a clever data placement strategy between memory banks.

This is a joint work with Tristan Delizy, Kevin Marquet, Tanguy Risset, Guillaume Salagnac (Inria Socrate) and Stéphane Gros (eVaderis).

This work has been published at the French Compas workshop [8] and the RSP Symposium [2].

## 7.6. Dataflow-explicit futures

**Participant:** Ludovic Henrio.

A future is a place-holder for a value being computed, and we generally say that a future is resolved when the associated value is computed. In existing languages futures are either implicit, if there is no syntactic or typing distinction between futures and non-future values, or explicit when futures are typed by a parametric type and dedicated functions exist for manipulating futures. We defined a new form of future, named data-flow explicit futures [38], with specific typing rules that do not use classical parametric types. The new futures allow at the same time code reuse and the possibility for recursive functions to return futures like with implicit futures, and let the programmer declare which values are futures and where synchronisation occurs, like with explicit futures. We prove that the obtained programming model is as expressive as implicit futures but exhibits a different behaviour compared to explicit futures. The current status of this work is the following:

- A paper showing formally the difference between implicit and explicit futures is under submission
- We are working with collaborators from University of Uppsala and University of Oslo on the design of programming constructs mixing implicit and dataflow-explicit futures
- Amaury Maillé will do his internship in the Cash team (advised by Matthieu Moy and Ludovic Henrio), working on an implementation of dataflow-explicit futures and further experiments with the model.

## 7.7. Locally abstract globally concrete semantics

**Participant:** Ludovic Henrio.

This research direction aims at designing a new way to write semantics for concurrent languages. The objective is to design semantics in a compositional way, where each primitive has a local behavior, and to adopt a style much closer to verification frameworks so that the design of an automatic verifier for the language is easier. The local semantics is expressed in a symbolic and abstract way, a global semantics gathers the abstract local traces and concretizes them. We have a reliable basis for the semantics of a simple language (a concurrent while language) and for a complex one (ABS), but the exact semantics and the methodology for writing it is still under development, we expect to submit a journal article during 2019 on the subject.

This is a joint with Reiner Hähnle (TU Darmstadt), Einar Broch Johnsen, Crystal Chang Din, Lizeth Tapia Tarifa (Univ Oslo), Ka I Pun (Univ Oslo and Univ of applied science).

## 7.8. Memory consistency for heterogeneous systems

**Participant:** Ludovic Henrio.

Together with Christoph Kessler (Linköping University), we worked on the formalization of the cache coherency mechanism used in the VectorPU library developed at Linköping University. Running a program on disjoint memory spaces requires to address memory consistency issues and to perform transfers so that the program always accesses the right data. Several approaches exist to ensure the consistency of the memory accessed, we are interested here in the verification of a declarative approach where each component of a computation is annotated with an access mode declaring which part of the memory is read or written by the component. The programming framework uses the component annotations to guarantee the validity of the memory accesses. This is the mechanism used in VectorPU, a C++ library for programming CPU-GPU heterogeneous systems and this article proves the correctness of the software cache-coherence mechanism used in the library. Beyond the scope of VectorPU, this article can be considered as a simple and effective formalisation of memory consistency mechanisms based on the explicit declaration of the effect of each component on each memory space. This year, we have the following new results:

- provided a formalization showing the correctness of VectorPU approach (published in 4PAD 2018, a symposium affiliated to HPCS).
- extended the work to support the manipulation of overlapping array (submitted as an extended version of the 4PAD paper)

We now plan to extend the work with support for concurrency.

## 7.9. PNets: Parametrized networks of automata

**Participant:** Ludovic Henrio.

pNets (parameterised networks of synchronised automata) are semantic objects for defining the semantics of composition operators and parallel systems. We have used pNets for the behavioral specification and verification of distributed components, and proved that open pNets (i.e. pNets with holes) were a good formalism to reason on operators and parameterized systems. This year, we have the following new results:

- A weak bisimulation theory for open pNets is under development (a strong isimulation had already been defined in the past) and its properties are being proven, especially in terms of compositionality. This work is realized with Eric Madelaine (Inria Sophia-Antipolis) and Rabéa Ameur Boulifa (Telecom ParisTech).
- A translation from BIP model to open pNets is being formalized and encoded, this work is done in collaboration with Simon Bliudze (Inria Lille).

These works are under progress and should be continued in 2019.

## 7.10. Decidability results on the verification of phaser programs

**Participant:** Ludovic Henrio.

Together with Ahmed Rezine and Zeinab Ganjei (Linköping University) we investigated the possibility to analyze programs with phasers (a construct for synchronizing processes that generalizes locks, barrier, and publish-subscribe patterns). They work with signal and wait messages from the processes (comparing the number of wait and signal received to synchronize the processes). We proved that in many conditions, if the number of phasers or processes cannot be bounded, or if the difference between the number of signal and the number of wait signal is unbounded, then many reachability problems are undecidable. We also proposed fragments where these problems become decidable, and proposed an analysis algorithm in these cases. The results are currently under review in a conference.

## 7.11. Practicing Domain-Specific Languages: From Code to Models

**Participant:** Laure Gonnord.

Together with Sebastien Mosser, we proposed a new Domain- Specific Language course at the graduate level whose objectives is to reconciliate concepts coming from Language Design as well as Modeling domains. We illustrate the course using the reactive systems application domain, which prevents us to fall back in a toy example pitfall. This paper describes the nine stages used to guide students through a journey starting at low-level C code to end with the usage of a language design workbench. This course was given as a graduate course available at Université Côte d'Azur (8 weeks, engineering-oriented) and École Normale Supérieure de Lyon (13 weeks, research-oriented).

The results have been published in a national software engineering conference [4] and the Models Educator Symposium [5].

## 7.12. Polyhedral Dataflow Programming: a Case Study

**Participant:** Laure Gonnord.

With Lionel Morel and Romain Fontaine (Insa Lyon), we have studied the benefits of jointly using polyhedral compilation with dataflow languages. We have proposed to expend the parallelization of dataflow programs by taking into account the parallelism exposed by loop nests describing the internal behavior of the program's agents. This approach is validated through the development of a prototype toolchain based on an extended version of the SigmaC language. We demonstrated the benefit of this approach and the potentiality of further improvements on several case studies.

The results have been published in the Sbac-PAD conference on High Performance computing [3].

## 7.13. Semantic Array Dataflow Analysis

**Participants:** Laure Gonnord, Paul Iannetta.

Together with Lionel Morel (Insa/CEA) and Tomofumi Yuki (Inria, Rennes), we revisited the polyhedral model's key analysis, dependency analysis. The semantic formulation we propose allows a new definition of the notion of dependency and the computation of the dependency set. As a side effect, we propose a general algorithm to compute *an over-approximation of* the dependency set of general imperative programs.

We argue that this new formalization will later allow for a new vision of the polyhedral model in terms of semantics, which will help us fully characterize its expressivity and applicability. We also believe that abstract semantics will be the key for designing an approximate abstract model in order to enhance the applicability of the polyhedral model.

The results is published in a research report [11].

## 7.14. Static Analysis Of Binary Code With Memory Indirections Using Polyhedra

**Participant:** Laure Gonnord.

Together with Clement Ballabriga, Julien Forget, Giuseppe Lipari, and Jordy Ruiz (University of Lille), we proposed a new abstract domain for static analysis of binary code. Our motivation stems from the need to improve the precision of the estimation of the Worst-Case Execution Time (WCET) of safety-critical real-time code. WCET estimation requires computing information such as upper bounds on the number of loop iterations, unfeasible execution paths, etc. These estimations are usually performed on binary code, mainly to avoid making assumptions on how the compiler works. Our abstract domain, based on polyhedra and on two mapping functions that associate polyhedra variables with registers and memory, targets the precise computation of such information. We prove the correctness of the method, and demonstrate its effectiveness on benchmarks and examples from typical embedded code.

The results have been accepted to VMCAI'19 on Model Checking and Abstract Interpretation [7].

## 7.15. Polyhedral Value Analysis as Fast Abstract Interpretation

**Participant:** Laure Gonnord.

Together with Tobias Grosser, (ETH Zurich, Switzerland), Siddhart Bhat, (IIIT Hydrabad, India), Marcin Copik (ETH Zurich, Switzerland), Sven Verdoolaege (Polly Labs, Belgium) and Torsten Hoefler (ETH Zurich, Switzerland), we tried to bridge the gap between the well founded classical abstract interpretation techniques and their usage in production compilers.

We formulate the polyhedral value analysis (a classical algorithm in production compilers like LLVM, scalar evolution based on Presburger set as abstract interpretation), present a set of fast join operators, and show that aggressively falling back to top (rather than continuing with approximations) results in a scalable analysis. By formally describing the required analysis, we provide the necessary theoretical foundations for analysing large program systems with hundred thousands of loops and complex control flow structures at a precision high enough to cater for high-precision users such as polyhedral optimization frameworks, at a compile-time cost comparable with just compiling the application.

The paper is under redaction process.

**Chroma Project-Team**

# 7. New Results

## 7.1. Bayesian Perception

**Participants:** Christian Laugier, Lukas Rummelhard, Jean-Alix David, Thomas Genevois, Jerome Lussereau, Nicolas Turro [SED], Jean-François Cuniberto [SED].

### *7.1.1. Conditional Monte Carlo Dense Occupancy Tracker (CMCDOT) Framework*

**Participants:** Lukas Rummelhard, Jerome Lussereau, Jean-Alix David, Thomas Genevois, Christian Laugier, Nicolas Turro [SED].

Recognized as one of the core technologies developed within the team over the years (see related sections in previous activity report of Chroma, and previously e-Motion reports), the CMCDOT framework is a generic Bayesian Perception framework, designed to estimate a dense representation of dynamic environments [83] and the associated risks of collision [85], by fusing and filtering multi-sensor data. This whole perception system has been developed, implemented and tested on embedded devices, incorporating over time new key modules [84]. In 2018, this framework, and the corresponding software, has continued to be the core of many important industrial partnerships and academic contributions [17] [18] [16] [15] [45] [47], and to be the subject of important developments, both in terms of research and engineering. Some of those recent evolutions are detailed below.

- CMCDOT evolutions : important developments in the CMCDOT, in terms of calculation methods and fundamental equations, were introduced and tested this year. These developments could lead, in the coming months, to the proposal of a new patent, then to academic publications. These changes introduced, among other evolutions, a much higher update frequency, greater flexibility in the management of transitions between states (and therefore a better system reactivity), as well as the management of a high variability in sensor frequencies (for each sensor over time, and in the set of sensors). The technical documents describing those developments are currently being redacted, and will be described in the next annual report.
- Multi-sensor integration in the Ground Estimator : the module of dynamic estimation of the shape of the ground and data segmentation, based solely on the sensor point clouds (no prior map information), the first step of data interpretation in CMCDOT framework, has been developed since 2016, patented and published in 2017. The corresponding software, until this year, could not take into account more than one sensor. In case of multiple sensors, several different modules were to be launched, their respective occupancy grids then fused, not only increasing the global computation use, but also preventing each sensor from benefiting from the ground models generated by the others. This point was corrected this year, by introducing the management of multiple input sensors, unifying the ground estimation in a single model, thus leading to improved performance, both in terms of calculation and results.
- Velocity display : in the CMCDOT framework, velocity of every element of the scene is inferred at a cell level, without object segmentation. This low-level velocity estimation is one of the most original and important aspects of the method, and should be displayed accordingly. A velocity display module, displaying for each occupied cell of the grid the average of the estimated velocity, generating colors depending on the intensity and the orientation, has been developed, see Fig. 5 .
- Software optimization : the whole CMCDOT framework has been developed on GPUs (implementations in C++/Cuda), an important focus of the engineering has always been, and continued to be in 2018, on the optimization of the software and methods to be embedded on low energy consumption embedded boards (now Nvidia Jetson TX2).

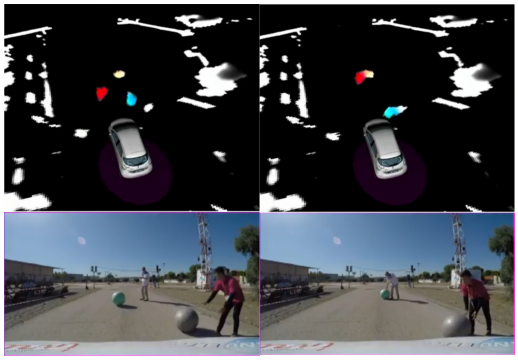


*Figure 5. Image from the Velocity Display module : in every occupied cell of the grid, the average velocity is represented by a color code, the hue being based on the orientation, the saturation on its norm. A static cell is white, a cell moving in the same direction as the vehicle is red, in the opposite direction in blue. In the grid can be seen the moving balloons, the pedestrians being static.*

- IROS 2018 Autonomous Driving event : https://hal.inria.fr/medihal-01963296v1 As already mentioned in the highlights of the year, the experimental Zoe platform, funded by IRT Nanoelec, has participated at IROS2018 in the Autonomous Vehicle Demonstrations, a full day of demonstration of autonomous vehicle capacities from various research centers. During this successful event, it has been presented and demonstrated on live conditions the effectiveness of the embedded CMCDOT framework, in connection with the newly developed control and decision making systems.

### *7.1.2. Simulation based validation*

**Participants:** Thomas Genevois, Lukas Rummelhard, Nicolas Turro [SED], Christian Laugier, Anshul Paigwar, Alessandro Renzaglia.

Since 2017, we are working to address the concept of *simulation based validation* in the scope of the EU Enable-S3 project, with the objective of searching for novel approaches, methods, tools and experimental methodology for validating BOF-based algorithms. For that purpose, we have collaborated with the Inria Tamis team (Rennes) and with Renault for developing the simulation platform that is used in the test platform. The simulation of both the sensors and the driving environment are based on the Gazebo simulator. A simulation of the prototype car and its sensors has also been realized, meaning that the same implementation of *CMCDOT* can handle both real data and simulated data. The test management component that generates random simulated scenarios has also been developed. Output of *CMCDOT* computed from the simulated scenarios are recorded by *ROS* and analyzed through the Statistical Model Checker *(SMC)* developed by the Inria Tamis team. In [41], we presented the first results of this work, where a decision-making approach for intersection crossing (see Section 7.2.3 ) has been analyzed. In particular new KPIs expressed as Bounded Linear Temporal Logic (BLTL) formula have been defined. Temporal formulas allow a finer formulation of KPIs by taking into account the evolution of the metrics during time. A further work in this direction will be done in the next months to provide new results on the validation of the perception algorithm, namely for the velocity estimation and collision risk assessment. For this part, we are also exploring the advantages and potentiality of a new open-source vehicle simulator (Carla), which would allow considering more realistic scenarios with respect to Gazebo. This work on simulation-based validation will be continued in 2019.

Previously, in 2017, CHROMA has developed a model of the Renault Zoe demonstrator within the simulation framework Gazebo. In 2018, we have improved it to keep it up-to-date after several evolutions of the actual demonstrator. Namely, the drivers of the simulated lidars and the control law have been updated. Thus the model now provides the outputs corresponding to a simulated Inertial Measurement Unit.

### 7.1.3. Control and navigation

**Participants:** Thomas Genevois, Lukas Rummelhard, Jerome Lussereau, Jean-Alix David, Christian Laugier, Nicolas Turro [SED], Rabbia Asghar.

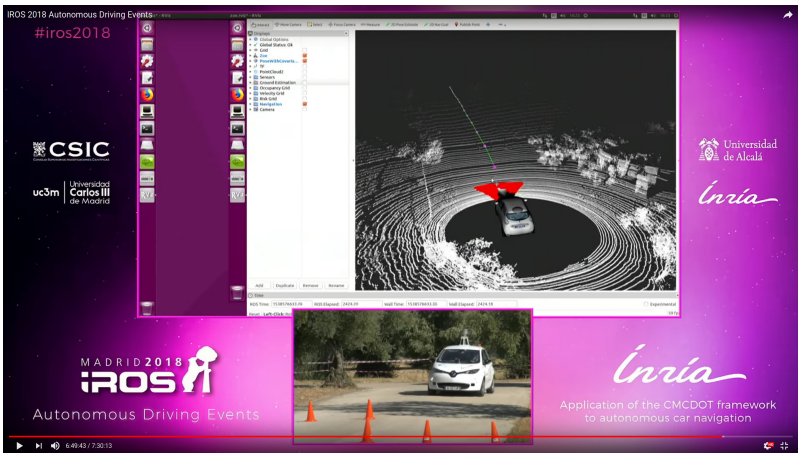



*Figure 6. Image taken from the live diffusion of the Autonomous Vehicles event at IROS2018. The demonstrator Renault Zoe is about to go through an obstacle course*

In 2018, we have updated the Renault Zoe demonstrator in collaboration with the LS2N (Laboratoire des Sciences Numérique de Nantes). The control codes have been transferred to the micro-controllers of the car for a faster and more precise control. An electric signal has been added to identify when the driver acts on the manual controls of the car. Finally the control law of the vehicle has been modified in order to consider a command in acceleration. These modifications allowed us to improve the software we use to control the vehicle. We have improved our implementation of DWA (Dynamic Window Approach) local planner in order to handle acceleration commands. This local planner has also been modified to take in account maxima of lateral acceleration and to integrate a path following module in its cost function. Thanks to this, the new version of this program provides a smooth command for a combination of path following and obstacle avoidance with the demonstrator Renault Zoe. This has been showed at the Autonomous Vehicle Demonstration event at IROS2018, Madrid, Figure 6 [46].

We have also experimented a driving assistant for autonomous obstacle avoidance. We showed that it is possible on the Renault Zoe demonstrator to let a driver drive manually the car and then, when a collision risk is identified, to take over the control with the autonomous drive and perform an avoidance maneuver. A simple ADAS [0] system has been developed for this purpose. In addition, we have developed on the Renault Zoe demonstrator, a localization system which merges the data of wheel speed, accelerometer, gyrometer, magnetometer and GPS into a position estimation. This relies on an Extended Kalman Filter. This will probably be extended later to consider the localization with respect to roads identified on a map.

Finally a Dijkstra Algorithm have been tested in simulation to define a global navigation path allowing management of waypoints to give to the DWA planner for local navigation.

## 7.2. Situation Awareness & Decision-making

**Participants:** Christian Laugier, Olivier Simonin, Jilles Dibangoye, David Sierra-Gonzalez, Mathieu Barbier, Victor Romero-Cano [Universidad Autónoma de Occidente, Cali, Colombia], Ozgur Erkent, Christian Wolf.

[0] Advanced Driving Assistance System

### 7.2.1. Dense & Robust outdoor perception for autonomous vehicles

**Participants:** Christian Laugier, Victor Romero-Cano, Özgür Erkent, Christian Wolf.

Robust perception plays a crucial role in the development of autonomous vehicles. While perception in normal and constant environmental conditions has reached a plateau, robustly perceiving changing and challenging environments has become an active research topic, particularly due to the safety concerns raised by the introduction of autonomous vehicles to public streets. Solving the robustness issue in road and urban perception applications is the first challenge. Then, it is also mandatory to develop an appropriate framework for extracting relevant semantic information. Our approach is to reason about vision-based data and the output of our grid-based multi-sensors perception approach (see previous section).

The work presented in this section has partly been done in 2017 and completed in 2018, in the scope of our collaboration with Toyota Motor Europe (TME). The main objective was to develop a framework for integrate the outcomes of the deep learning methods with a well-established area, occupancy grids obtained with a Bayesian filtering method in the grid space.

In this work, we are interested in 2D egocentric representations. We propose a method, which estimates an occupancy grid containing detailed semantic information. The semantic characteristics include classes like *road, car, pedestrian, sidewalk, building, vegetation, etc.*. To this end, we leverage and fuse information from multiple sensors including Lidar, odometry and monocular RGB video. To benefit from the respective advantages of the two different methodologies, we propose a hybrid approach leveraging i) the high-capacity of deep neural networks as well as ii) Bayesian filtering, which is able to model uncertainty in a unique way.

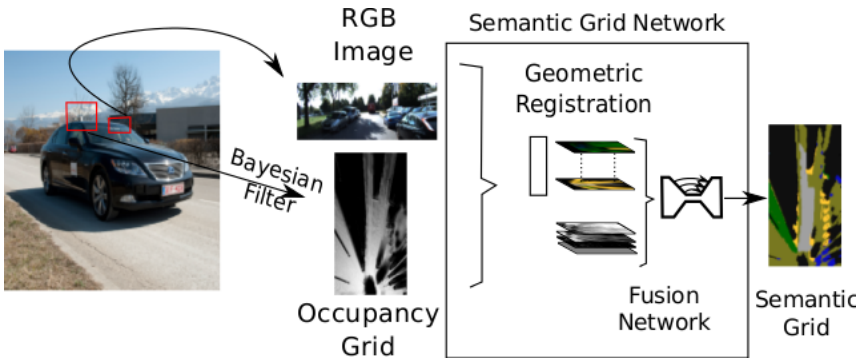



*Figure 7. The Semantic Grid framework.*

In the system depicted by Figure 7 , Bayesian particle filtering processes the Lidar data as well as odometry information from the vehicle's motion in order to robustly estimate an egocentric bird's eye view in the form of an occupancy grid. This grid contains a 360° view of the environment around the car and integrates information from the observation history through temporal filtering; however, it does not include fine-grained semantic classes.

Deep Learning is used for two different tasks in our work. Firstly, a deep network performs semantic segmentation of monocular RGB images. This network has been pre-trained on large scale datasets for image classification and fine-tuned on the vehicle datasets. Secondly, a deep network fuses the occupancy grid with the segmented image of the projective view in order to estimate the semantic grid. Since the occupancy grid is dense, the semantic grid is also expected to be dense. We pay particular attention to correctly model the transformation from the egocentric projective view of the RGB image to the bird's eye view of the occupancy grid as input to the neural network. This work was filed for a patent [98] and published in [28], [14].

**Novel approach: Semantic Grid Estimation with a Hybrid Bayesian and Deep Neural Network Approach.**

Current and future work in the scope of our collaboration with TME, aims at constructing *Semantic Occupancy Grids*. We propose a hybrid approach, which combines the advantages of Bayesian filtering and deep neural networks. Bayesian filtering provides robust temporal/geometrical filtering and integration and allows for modelling of uncertainty. RGB information and deep neural networks provide knowledge about the semantic class labels like *sideway* vs *road*. The fusion process is fully learned and due to dense structure of occupancy grid, we can construct a dense semantic grid even if we have a sparse point cloud.

### *7.2.2. Towards Human-Like Motion Prediction and Decision-Making in Highway Scenarios*

**Participants:** David Sierra González, Victor Romero-Cano, Özgür Erkent, Jilles Dibangoye, Christian Laugier.

The objective is to develop human-like motion prediction and decision-making algorithms to enable automated driving in highways. This research work is done in the scope of the Inria-Toyota long-term cooperation on Autonomous Driving and of the PhD thesis work of David Sierra González.

Previous work from our team has shown the predictive potential of driver behavioral models learned from demonstrations using Inverse Reinforcement Learning (IRL) [87] [88]. Unfortunately, these models are hard to learn from real-world driving data due to the inability of traditional IRL algorithms to handle continuous state spaces and dynamic environments. To facilitate this task, we have proposed in 2018 an approximated IRL algorithm for driver behavior modeling that successfully scales to continuous spaces with moving obstacles, by leveraging a spatio-temporal trajectory planner [35]. The proposed algorithm was validated using real-world data gathered with an instrumented vehicle. As an example, Figure 8 shows the similarity between the trajectory obtained using a driver model learned with the proposed method and that of a real human driver in a highway overtake scenario. Current efforts are directed towards integrating the learned behavioral models and the predictive models developed in the scope of this project into a decision-making framework for highways. David Sierra González will defend his PhD thesis in March 2019.

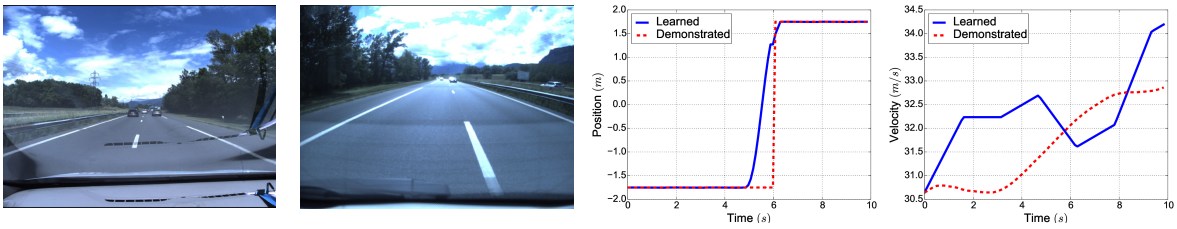



*Figure 8.*

*Comparison of the trajectory obtained with a driver model learned from demonstrated driving data using the method proposed in [35] and that of a human driver for a typical highway overtake scenario*
*a. Front view at $t = 5.0$ b. Back view at $t = 5.0$ c. Position prediction d. Velocity prediction*

### *7.2.3. Decision-making for safe road intersection crossing*

**Participants:** Mathieu Barbier, Christian Laugier, Olivier Simonin.

Road intersections are probably the most complex segment in a road network. Most major accidents occur at intersections, mainly caused by human errors due to failures in fully understanding the encountered situations. Indeed, as drivers approach a road intersection, they must assess the situation and quickly adapt their behaviour accordingly. When this task is performed by a computer, the available information is partial and uncertain.

Any decision requires the system to use this information as well as taking into account the behaviour of other drivers to avoid collisions. However, metrics such as collision rate can remain low in an interactive environment because of other driver's actions. Consequently, evaluation metrics must depend on other driving aspects.

In this framework, we developed a decision-making mechanism and designed metrics to evaluate such a system at road intersection crossing [22]. For the former, a Partially Observable Markov Decision Process (POMDP) is used to model the system with respect to uncertainties in the behaviour of other drivers. For the latter, different key performance indicators are defined to evaluate the resulting behaviour of the system in different configurations and scenarios. The approach has been demonstrated within an automotive grade simulator.

Current work aims at increasing the complexity of the scenario, to include pedestrians and more vehicles, and improving the model used for the dynamics of the vehicle and the observation of the physical state to get closer to real world scenarios.

This work has been carried out in the framework of the PhD thesis of Mathieu Barbier, which will be defended in the first trimester of 2019.

## 7.3. Robust state estimation (Sensor fusion)

This research is the follow up of Agostino Martinelli's investigations carried out during the last five years, which are in the framework of the visual and inertial sensor fusion problem and the unknown input observability problem.

### *7.3.1. Visual-inertial structure from motion*

**Participants:** Agostino Martinelli, Alexander Oliva, Alessandro Renzaglia.

During this year, we have obtained the full analytic solution of the cooperative visual inertial sensor fusion problem in the case of two agents, starting from the closed-form solution obtained in the last years (this latter solution will be published on the journal of Autonomous Robots [76]). Additionally, we also validated this solution with real experiments and in particular we showed that the analytic solution significantly outperforms our previous closed-form solution in [76]. The new analytic solution has been accepted for publication by the IEEE Robotics and Automation Letters [13].

Specifically, we obtained the analytic solution of the problem by first proving that, this sensor fusion problem, is equivalent to a simple polynomial equations system that consists of several linear equations and three polynomial equations of second degree. The analytic solution of this polynomial equations system was easily obtained by using an algebraic method (developed by Bernard Mourrain, the leader of AROMATH at Inria Sophia Antipolis). The power of the analytic solution is twofold. From one side, it allows us to determine the relative state between the agents (i.e., relative position, speed and orientation) without the need of an initialization. From another side, it provides fundamental insights into all the theoretical aspects of the problem. During this year, we focused on the first issue. Our next objective is to exploit the analytic solution to obtain basic structural properties of the problem.

### *7.3.2. Unknown Input Observability*

**Participant:** Agostino Martinelli.

The Unknown Input Observability problem (UIO) in the nonlinear case was an open problem since the sixties years, when it was solved only in the linear case. In the last five years, I have obtained its general analytic solution. The mathematics apparatus necessary to obtain this solution is very sophisticated and is based on Ricci calculus, borrowed from theoretical physics. On the other hand, this mathematics can be avoided in the case of driftless systems and characterized by a single unknown input.

All the results (i.e., in the general case that also accounts for a drift and more than one unknown input) are fully described in a book available on ArXiv (arXiv:1704.03252).

During this year, my effort was devoted to make the analytic derivation of the solution palatable for a large audience (in particular, without knowledge of Ricci calculus). Hence, I focused on the simple case of a single unknown input and without drift. This solution has been published on a full paper on the IEEE Transaction on Automatic Control [75].

Regarding the general case available on ArXiv (arXiv:1704.03252), I was invited by the SIAM to write a book, palatable for a large audience. The scope of writing this book, is to present to the control theory and information theory communities a very powerful mathematics framework borrowed from theoretical physics. This could provide the possibility of revisiting many aspects of the control and information theory and bring new fundamental results, open new research domains etc. In this sense the book could be the kick-off of a new season of research in control and information theory. This will be the objective of the next years.

## 7.4. Motion-planning in human-populated environment

We study new motion planning algorithms to allow robots/vehicles to navigate in human populated environment, and to predict human motions. Since 2016, we investigate several directions exploiting vision sensors : prediction of pedestrian behaviors in urban environments (extended GHMM), mapping of human flows (statistical learning), and learning task-based motion planning (RL+Deep-Learning) . These works are presented here after.

### *7.4.1. Urban Behavioral Modeling*

**Participants:** Pavan Vasishta, Anne Spalanzani, Dominique Vaufreydaz.

The objective of modeling urban behavior is to predict the trajectories of pedestrians in towns and around car or platoons (PhD work of P. Vasishta). In 2017 we proposed to model pedestrian behaviour in urban scenes by combining the principles of urban planning and the sociological concept of Natural Vision. This model assumes that the environment perceived by pedestrians is composed of multiple potential fields that influence their behaviour. These fields are derived from static scene elements like side-walks, cross-walks, buildings, shops entrances and dynamic obstacles like cars and buses for instance. This work was published in [95], [94]. In 2018, an extension to the Growing Hidden Markov Model (GHMM) method has been proposed to model behavior of pedestrian without observed data or with very few of them. This is achieved by building on existing work using potential cost maps and the principle of Natural Vision. As a consequence, the proposed model is able to predict pedestrian positions more precisely over a longer horizon compared to the state of the art. The method is tested over legal and illegal behavior of pedestrians, having trained the model with sparse observations and partial trajectories. The method, with no training data (see. Fig. 9 .a), is compared against a trained state of the art model. It is observed that the proposed method is robust even in new, previously unseen areas. This work was published in [36] and won the **best student paper** of the conference.

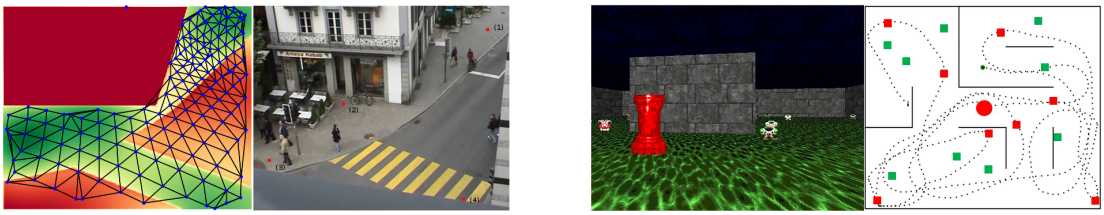


*Figure 9.*

*a. Prior Topological Map of the dataset from the Traffic Anomaly Dataset : first figure shows the generated potential cost map and second figure the "Prior Topology" of the image from scene.*
*b. Illustration of learning task-based motion planning.*

### 7.4.2. Learning task-based motion planning

**Participants:** Christian Wolf, Jilles Dibangoye, Laetitia Matignon, Olivier Simonin, Edward Beeching.

Our goal is the automatic learning of robot navigation in human populated environments based on specific tasks and from visual input. The robot automatically navigates in the environment in order to solve a specific problem, which can be posed explicitly and be encoded in the algorithm (e.g. recognize the current activities of all the actors in this environment) or which can be given in an encoded form as additional input. Addressing these problems requires competences in computer vision, machine learning, and robotics (navigation and paths planning).

We started this work in the end of 2017, following the arrival of C. Wolf, through combinations of reinforcement learning and deep learning. The underlying scientific challenge here is to automatic learn representations which allow the agent to solve multiple sub problems require for the task. In particular, the robot needs to learn a metric representation (a map) of its environment based from a sequence of ego-centric observations. Secondly, to solve the problem, it needs to create a representation which encodes the history of ego-centric observations which are relevant to the recognition problem. Both representations need to be connected, in order for the robot to learn to navigate to solve the problem. Learning these representations from limited information is a challenging goal. This is the subject of the PhD thesis of Edward Beeching who started on October 2018, see illustration Fig. 9 .b.

### 7.4.3. Human-flows modeling and social robots

**Participants:** Jacques Saraydaryan, Fabrice Jumel, Olivier Simonin, Benoit Renault, Laetitia Matignon, Christian Wolf.

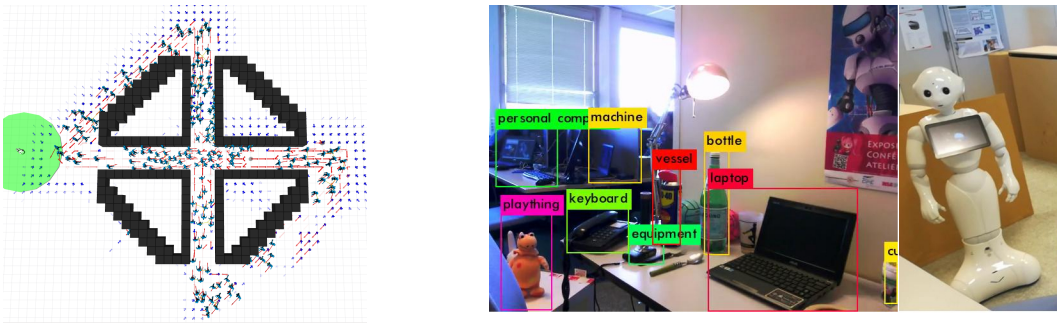



*Figure 10.*

*(a) Flow-grid mapping in a cross-corridor where 200 moving pedestrians turns*
*(b) Object detection with Pepper based on vision/deep learning techniques.*

In order to deal with robot/humanoid navigation in complex and populated environments such as homes, we investigate since 2 years several research avenues :

- Mapping humans flows. We defined a statistical learning approach (ie. a counting-based grid model) exploiting only data from robots embedded sensors. See illustration in Fig. 10 .a and publication [66].
- Path-planning in human flows. We revisited the A* path-planning cost function under the hypothesis of the knowledge of a flow grid. See publication [66].
- In 2018 we started to study NAMO problems (Navigation Among Movable Obstacles) by considering populated environments and muti-robot cooperation. After his Master thesis on this subject, Benoit Renault started a PhD in Chroma focusing on the extension of NAMO algorithms to such dynamic environments.

- RoboCup competition. In the context of the **RoboCup** international competition, we created the 'LyonTech' team, joining members from Chroma (INSA/CPE/UCBL). We investigated several humanoid tasks in home environments with our Pepper robot : social aware architecture, decision making and navigation, deep-learning based human and object detection (see Fig. 10 .b), human-robot interaction. In July 2018, we participated for the first time to the RoboCup and reaching the 5th rank of the SSL league (Pepper@home). We also published our social-aware architecture to the RoboCup Conference [31]. In October 2018, we qualified for the next final phase of RoboCup SSL (Pepper) to be organized on July 2019, in Sydney.

## 7.5. Decision Making in Multi-Robot Systems

### *7.5.1. Multi-robot planning in dynamic environments*

*7.5.1.1. Global-local optimization in autonomous multi-vehicles systems*

**Participants:** Guillaume Bono, Jilles Dibangoye, Laetitia Matignon, Olivier Simonin, Florian Peyreron [VOLVO Group, Lyon].

This work is part of the PhD. thesis in progress of Guillaume Bono, with the VOLVO Group, in the context of the INSA-VOLVO Chair. The goal of this project is to plan and learn at both global and local levels how to act when facing a vehicle routing problem (VRP). We started with a state-of-the-art paper on vehicle routing problems as it currently stands in the literature [53]. We were surprise to notice that few attention has been devoted to deep reinforcement learning approaches to solving VRP instances. Hence, we investigated our own deep reinforcement learning approach that can help one vehicle to learn how to generalize strategies from solved instances of travelling salesman problems (an instance of VRPs) to unsolved ones. The difficulty of this problem lies in the fact that its Markov decision process' formulation is intractable, i.e., the number of states grows doubly exponentially with the number of cities to be visited by the salesman. To gain in scalability, we build inspiration on a recent work by DeepMind, which suggests using pointer-net, i.e., a novel deep neural network architecture, to address learning problems in which entries are sequences (here cities to be visited) and output are also sequences (here order in which cities should be visited). Preliminary results are encouraging and we are extending this work to the multi-agent setting.

*7.5.1.2. Multi-Robot Routing (MRR) for evolving missions*

**Participants:** Mihai Popescu, Olivier Simonin, Anne Spalanzani, Fabrice Valois [INSA/Inria, Agora team].

After considering Multi-Robot Patrolling of known targets in 2016 [81], we generalized to MRR (multi-robot routing) and to DMRR (Dynamic MRR) in the work of the PhD of M. Popescu. Target allocation problems have been frequently treated in contexts such as multi-robot rescue operations, exploration, or patrolling, being often formalized as multi-robot routing problems. There are few works addressing dynamic target allocation, such as allocation of previously unknown targets. We recently developed different solutions to variants of this problem :

- MRR : Multi-robot routing has been the main testbed in the domain of multi-robot task allocation, where decentralized solutions consist in auction-based methods. Our work addresses the MRR problem and proposes MRR with saturation constraints (MRR-Sat), where the cost of each robot treating its allocated targets cannot exceed a bound (called saturation). We provided a NP-Complete proof for the problem of MRR-Sat. Then, we proposed a new auction-based algorithm for MRR-Sat and MRR, which combines ideas of parallel allocations with target-oriented heuristics. An empirical analysis of the experimental results shows that the proposed algorithm outperforms state-of-the art methods, obtaining not only better team costs, but also a much lower running time. Results are submitted to RSS'2019 conference.
- DMRR : we defined the Dynamic-MRR problem as the continuous adaptation of the ongoing robot missions to new targets. We proposed a framework for dynamically adapting the existent robot missions to new discovered targets. Dynamic saturation-based auctioning (DSAT) is proposed for adapting the execution of robots to the new targets. Comparison was made with algorithms ranging

from greedy to auction-based methods with provable sub-optimality. The results for DSAT shows it outperforms state-of-the-art methods, like standard SSI or SSI with regret clearing, especially in optimizing the target allocation w.r.t. the target coverage in time and the robot resource usage (e.g. minimizing the worst mission cost). First results have been published in [34].

- Synchronization : When patrolling targets along bounded cycles, robots have to meet periodically to exchange information, data (e.g. results of their tasks). Data will finally reach a delivery point (e.g. the base station). Hence, patrolling cycles sometimes have common points (rendezvous points), where the information needs to be exchanged between different cycles (robots). We investigated this problem by defining the following first solutions : random-wait, speed adaptation (first-multiple), primality of periods, greedy interval overlapping. We developed a simulator, allowing experiments that show the approaches have different performances and robustness. This work will be submitted to IROS'2019 conference.
- PHC DRONEM [0] : We started a collaboration in 2017 with the team of Prof. Gabriela Czibula from Babes-Bolyai University in Cluj-Napoca, Romania. The DRONEM project focuses on optimization and online adaptation of the multi-cycle patrolling with machine learning (RL) techniques in order to deal with the arrival of new targets in the environment.

*7.5.1.3. Middleware for open multi-robot systems*

**Participants:** Stefan Chitic, Julien Ponge [INSA/CITI, Dynamid], Olivier Simonin.

Multi-robots systems (MRS) require dedicated software tools and models to face the complexity of their design and deployment. In the context of the PhD work of Stefan Chitic, we addressed service self-discovery and property proofs in an ad-hoc network formed by a fleet of robots. This led us to propose a robotic middleware, SDfR, that is able to provide service discovery, see [54]. In 2017, we defined a tool-chain based on timed automata, called ROSMDB, that offers a framework to formalize and implement multi-robot behaviors and to check some (temporal) properties (both offline and online). Stefan Chtic defended his Phd thesis on March 2018 [11].

### *7.5.2. Multi-robot Coverage and Mapping*

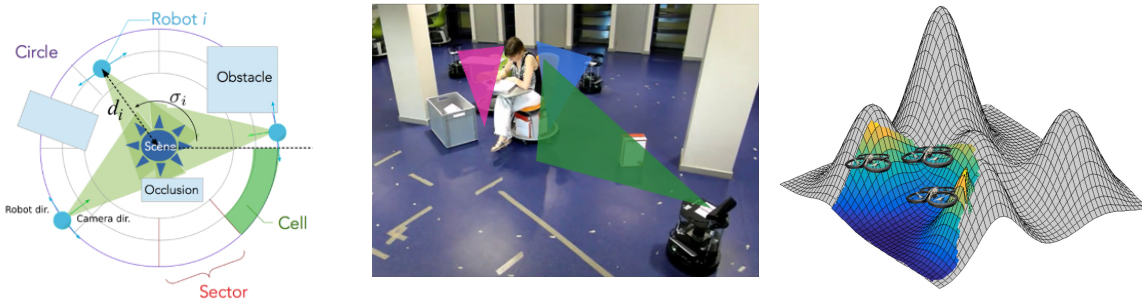



*Figure 11. (a) Concentric navigation model and (b) its experimental setup. (c) Illustration of the local search method for multi-UAV coverage.*

*7.5.2.1. Human scenes observation*

**Participants:** Laetitia Matignon, Olivier Simonin, Stephane d'Alu, Christian Wolf.

[0] Hubert Curien Partnership

Solving complex tasks with a fleet of robots requires to develop generic strategies that can decide in real time (or time-bounded) efficient and cooperative actions. This is particularly challenging in complex real environments. To this end, we explore anytime algorihms and adaptive/learning techniques.

The "CROME" and "COMODYS" [0] projects [0] are motivated by the exploration of the joint-observation of complex (dynamic) scenes by a fleet of mobile robots. In our current work, the considered scenes are defined as a sequence of activities, performed by a person in a same place. Then, mobile robots have to cooperate to find a spatial configuration around the scene that maximizes the joint observation of the human pose skeleton. It is assumed that the robots can communicate but have no map of the environment and no external localisation.

To attack the problem, we proposed an original concentric navigation model allowing to keep easily each robot camera towards the scene (see fig. 11 .a). This model is combined with an incremental mapping of the environment and exploration guided by meta-heuristics in order to limit the complexity of the exploration state space. Results have been published in AAMAS'2018 [32]. An extended version has been submitted to the Journal JAAMAS.

For experiment with multi-robot systems, we defined an hybrid metric-topological mapping. Robots individually build a map that is updated cooperatively by exchanging only high-level data, thereby reducing the communication payload. We combined the on-line distributed multi-robot decision with this hybrid mapping. These modules has been evaluated on our platform composed of several Turtlebots2, see fig. 11 .b. This robotic architecture has been presented in [77] (ECMR). A Demo has been done in AAMAS'2018 international conference [33].

*7.5.2.2. Multi-UAV Visual Coverage of Partially Known 3D Surfaces*

**Participants:** Alessandro Renzaglia, Olivier Simonin, Jilles Dibangoye, Vincent Le Doze.

It has been largely proved that the use of Unmanned Aerial Vehicles (UAVs) is an efficient and safe way to deploy visual sensor networks in complex environments. In this context, a widely studied problem is the cooperative coverage of a given environment. In a typical scenario, a team of UAVs is called to achieve the mission without a perfect knowledge on the environment and needs to generate the trajectories on-line, based only on the information acquired during the mission through noisy measurements. For this reason, guaranteeing a global optimal solution of the problem is usually impossible. Furthermore, the presence of several constraints on the motion (collision avoidance, dynamics, etc.) as well as from limited energy and computational capabilities, makes this problem particularly challenging.

Depending on the sensing capabilities of the team (number of UAVs, range of on-board sensor, etc.) and the dimension of the environment to cover, different formulations of this problem can be considered. We firstly approached the deployment problem, where the goal is to find the optimal static UAVs configuration from which the visibility of a given region is maximized. A suitable way to tackle this problem is to adopt derivative-free optimization methods based on numerical approximations of the objective function. In 2012, Renzaglia et al. [82] proposed an approach based on a stochastic optimization algorithm to obtain a solution for arbitrary, initially unknown 3D terrains (see fig. 11 .c). However, adopting this kind of approaches, the final configuration can be strongly dependent on the initial positions and the system can get stuck in local optima very far from the global solution. We identified that a way to overcome this problem can be found in initializing the optimization with a suitable starting configuration. An a priori partial knowledge on the environment is a fundamental source of information to exploit to this end. The main contribution of our work is thus to add another layer to the optimization scheme in order to exploit this information. This step, based on the concept of Centroidal Voronoi Tessellation, will then play the role of initialization for the on-line, measurement-based local optimizer. The resulting method, taking advantages of the complementary properties of geometric and stochastic optimization, significantly improves the result of the previous approach and notably reduces the probability of a far-to-optimal final configuration. Moreover, the number of iterations necessary for the convergence of the on-line algorithm is also reduced. This work led to a paper submitted to AAMAS 2019 [0], currently under review. The development of a realistic simulation environment based on

[0] COoperative Multi-robot Observation of DYnamic human poSes

[0] Funded by a LIRIS transversal project in 2016-2017 and a FIL project in 2017-2019 (led by L. Matignon)

Gazebo is an important on-going activity in Chroma and will allow us to further test the approach and to prepare the implementation of this algorithm on the real robotic platform of the team.

We are currently also investigating the dynamic version of this problem, where the information is collected along the trajectories and the environment reconstruction is obtained from the fusion of the total visual data.

### *7.5.3. Sequential decision-making*

This research is the follow up of a group led by Jilles S. Dibangoye carried out during the last three years, which include foundations of sequential decision making by a group of cooperative or competitive robots or more generally artificial agents. To this end, we explore combinatorial, convex optimization and reinforcement learning methods.

*7.5.3.1. Optimally solving cooperative and competitive games as continuous Markov decision processes*

**Participants:** Jilles S. Dibangoye, Olivier Buffet [Inria Nancy], Vincent Thomas [Inria Nancy], Christopher Amato [Univ. New Hampshire], François Charpillet [Inria Nancy, Larsen team].

Our major findings this year include:

1. (Theoretical) – As an extension of [58] in the cooperative case [44], we characterize the optimal solution of partially observable stochastic games.
2. (Theoretical) – We further exhibit new underlying structures of the optimal solution for both cooperative and non-cooperative settings.
3. (Algorithmic) – We extend a non-trivial procedure in [27] for computing such optimal solutions when only an incomplete knowledge about the model is available.

This work proposes a novel theory and algorithms to optimally solving a two-person zero-sum POSGs (zs-POSGs). That is, a general framework for modeling and solving two-person zero-sum games (zs-Games) with imperfect information. Our theory builds upon a proof that the original problem is reducible to a zs-Game—but now with perfect information. In this form, we show that the dynamic programming theory applies. In particular, we extended Bellman equations [50] for zs-POSGs, and coined them maximin (resp. minimax) equations. Even more importantly, we demonstrated Von Neumann & Morgenstern's minimax theorem [99] [100] holds in zs-POSGs. We further proved that value functions—solutions of maximin (resp. minimax) equations—yield special structures. More specifically, the maximin value functions are convex whereas the minimax value functions are concave. Even more surprisingly, we prove that for a fixed strategy, the optimal value function is linear. Together these findings allow us to extend planning and learning techniques from simpler settings to zs-POSGs. To cope with high-dimensional settings, we also investigated low-dimensional (possibly non-convex) representations of the approximations of the optimal value function. In that direction, we extended algorithms that apply for convex value functions to lipschitz value functions [27].

*7.5.3.2. Learning to act in (continuous) decentralized partially observable Markov decision process*

**Participants:** Jilles S. Dibangoye, Olivier Buffet [Inria Nancy].

During the last year, we investigated deep and standard reinforcement learning for solving decentralized partially observable Markov decision processes. Our preliminary results include:

1. (Theoretical) Proofs that the optimal value function is linear in the occupancy-state space, the set of all possible distributions over hidden states and histories.
2. (Algorithmic) Value-based and policy-based (deep) reinforcement learning for common-payoff partially observable stochastic games.

[0] A. Renzaglia, J. Dibangoye, V. Le Doze and O. Simonin, "Multi-UAV Visual Coverage of Partially Known 3D Surfaces: Voronoi-based Initialization to Improve Local Optimizers", International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2019, *under review*.

This work addresses a long-standing open problem of Multi-Agent Reinforcement Learning (MARL) in decentralized stochastic control. MARL previously applied to finite decentralized decision making with a focus on team reinforcement learning methods, which at best lead to local optima. In this research, we build on our recent approach [44], which converts the original problem into a continuous-state Markov decision process, allowing knowledge transfer from one setting to the other. In particular, we introduce the first optimal reinforcement learning method for finite cooperative, decentralized stochastic control domains. We achieve significant scalability gains by allowing the latter to feed deep neural networks. Experiments show our approach can learn to act optimally in many finite decentralized stochastic control problems from the literature [43], [26].

*7.5.3.3. Study of policy-gradient methods for decentralized stochastic control*

**Participants:** Guillaume Bono, Jilles S. Dibangoye, Laëtitia Matignon, Olivier Simonin, Florian Peyreron [VOLVO Group, Lyon].

This work is part of the Ph.D. thesis in progress of Guillaume Bono, with VOLVO Group, in the context of the INSA-VOLVO Chair. The work aims at investigating an attractive family of reinforcement learning methods, namely policy-gradient and more generally actor-critic methods for solving decentralized partially observable Markov decision processes. Our preliminary results include:

1. (Theoretical) Proofs of the policy-gradient theorems for both total- and discounted-reward criteria in decentralized stochastic control.
2. (Algorithmic) (deep) actor-critic reinforcement learning methods for centralized and decentralized stochastic control.

Reinforcement Learning (RL) for decentralized partially observable Markov decision processes (Dec-POMDPs) is lagging behind the spectacular breakthroughs of single-agent RL. That is because assumptions that hold in single-agent settings are often obsolete in decentralized multi-agent systems. To tackle this issue, we investigate the foundations of policy gradient methods within the centralized training for decentralized control (CTDC) paradigm. In this paradigm, learning can be accomplished in a centralized manner while execution can still be independent. Using this insight, we establish policy gradient theorem and compatible function approximations for decentralized multi-agent systems. Resulting actor-critic methods preserve the decentralized control at the execution phase, but can also estimate the policy gradient from collective experiences guided by a centralized critic at the training phase. Experiments demonstrate our policy gradient methods compare favorably against standard RL techniques in benchmarks from the literature [42], [23]. Guillaume Bono also designed a simulator for urban logistic reinforcement learning, namely SULFR [39].

*7.5.3.4. Towards efficient algorithms for two-echelon vehicle routing problems*

**Participants:** Mohamad Hobballah, Jilles S. Dibangoye, Olivier Simonin, Elie Garcia [VOLVO Group, Lyon], Florian Peyreron [VOLVO Group, Lyon].

During the last year, Mohamad Hobballah (post-doc INSA VOLVO Chair) investigated efficient meta-heuristics for solving two-echelon vehicle routing problems (2E-VRPs) along with realistic logistic constraints. Algorithms for this problem are of interest in many real-world applications. Our short-term application targets goods delivery by a fleet of autonomous vehicles from a depot to the clients through an urban consolidation center using bikers. Preliminary results include:

1. (Methodological) Design of a novel meta-heuristic based on differential evolution algorithm [56] and iterative local search [97]. The former permits us to avoid being attracted by poor local optima whereas the latter performs the local solution improvement.
2. (Empirical) Empirical results on standard benchmarks available at http://www.vrp-rep.org/datasets.html show state-of-the-art performances on most VRP, MDVRP and 2E-VRP instances.

**CONVECS Project-Team**

# 6. New Results

## 6.1. New Formal Languages and their Implementations

### *6.1.1. LOTOS and LNT Specification Languages*

**Participants:** Hubert Garavel, Frédéric Lang, Wendelin Serwe.

LNT [5] [36] is a next-generation formal description language for asynchronous concurrent systems. The design of LNT at CONVECS is the continuation of the efforts undertaken in the 80s to define sound languages for concurrency theory and, indeed, LNT is derived from the ISO standards LOTOS (1989) and E-LOTOS (2001). In a nutshell, LNT attempts to combine the best features of imperative programming languages, functional languages, and value-passing process calculi.

LNT is not a frozen language: its definition started in 2005, as part of an industrial project. Since 2010, LNT has been systematically used by CONVECS for numerous case studies (many of which being industrial applications — see § 6.5 ). LNT is also used as a back-end by other research teams who implement various languages by translation to LNT. It is taught in university courses, e.g., at University Grenoble Alpes and ENSIMAG, where it is positively accepted by students and industry engineers. Based on the feedback acquired by CONVECS, LNT is continuously improved.

In 2018, the CADP tools that translate LNT to LOTOS have been enhanced in various ways. In the warning and error messages emitted by LNT2LOTOS, line numbers have been made more precise. In addition to a bug fix, the LNT_DEPEND tool, which computes dependencies between LNT modules has been entirely rewritten and made much faster. Also, the LNT language has been simplified by removing "!external" pragmas for constructors, as "!external" pragmas for types are sufficient.

We also continued improving the TRAIAN compiler for the LOTOS NT language (a predecessor of LNT), which is used for the construction of most CADP compilers and translators.

In February 2018, we released version 2.9 of TRAIAN. We scrutinized the source code of TRAIAN, deleting all parts of code corresponding to those features of the LOTOS NT language that were either not fully implemented or seldom used in practice. This reduced the source code of TRAIAN by 40% and the binaries by 50%. External LOTOS NT functions are now allowed to return a non-void result. Support for 64-bit macOS executables was added. A few bugs have been fixed and the reference manual of TRAIAN was entirely revised.

The main limitation of TRAIAN 2.x is that it is a 20-year-old compiler that is increasingly difficult to maintain. It consists in a large collection of attribute grammars and is built using the FNC-2 compiler generation system, which is no longer supported. For this reason, TRAIAN only exists in 32-bit version, and sometimes hits the 3–4 GB RAM limit when dealing with large compiler specifications, such as those of LNT2LOTOS or EVALUATOR 5.

For this reason, we undertook a complete rewrite of TRAIAN to get rid of FNC-2. Two main design decisions behind TRAIAN 3.0 are the following: (i) it supports (most of) the LOTOS NT language currently accepted by TRAIAN 2.9, but also extensions belonging to LNT, so as to allow a future migration from LOTOS NT to LNT; and (ii) TRAIAN 3.0 is currently written in LOTOS NT and compiled using TRAIAN 2.9, but should be ultimately capable of bootstrapping itself.

So far, a lexer and parser for LOTOS NT have been developed using the SYNTAX compiler-generation system [0] developed at Inria Paris. This work triggered an in-depth reexamination of the programming interfaces offered by SYNTAX and led to enhancements of these interfaces (see § 6.1.6 ).

[0] http://syntax.gforge.inria.fr

The abstract syntax tree of LOTOS NT, and the library of predefined LOTOS NT types and functions have been redesigned; previously specified as FNC-2 attribute grammars, they are now themselves written in LOTOS NT, so as to allow bootstrap, using the current version of TRAIAN to build the next one. The construction of the abstract syntax tree has also been completed. Finally, we set several non-regression test bases gathered all available programs written in LOTOS NT.

### *6.1.2. NUPN*

**Participant:** Hubert Garavel.

Nested-Unit Petri Nets (NUPNs) is an upward-compatible extension of P/T nets, which are enriched with structural information on their concurrent structure. Such additional information can easily be produced when NUPNs are generated from higher-level specifications (e.g., process calculi); quite often, such information allows logarithmic reductions in the number of bits required to represent states, thus enabling verification tools to perform better. The principles of NUPNs are exposed in [39] and its PNML representation is described here [0].

The NUPN model has been adopted by the Model Checking Contest and the Rigorous Examination of Reactive Systems challenge. It has been so far implemented in thirteen different tools developed in four countries.

In 2018, a journal article (to appear in 2019) has been written to formalize the complete theory of NUPNs. The CAESAR.BDD tool for NUPNs has been extended with twelve new options. A new tool named NUPN_INFO has been added to CADP to perform three normalizing transformations of NUPNs.

### *6.1.3. MCL and XTL Property Specification Languages*

**Participants:** Hubert Garavel, Radu Mateescu.

CADP provides two different languages, named MCL and XTL, for expressing data-handling temporal properties of concurrent systems. MCL is an extension of alternation-free modal $\mu$-calculus with data values, programming language constructs, generalized regular formulas on transition sequences, and fairness operators. XTL is a functional-like programming language interpreted on Labeled Transition Systems, enabling the definition of temporal operators by computing their interpretation using fixed point iterations over sets of states and transitions.

In 2018, we enhanced these languages and their associated tools as follows:

- The MCL v4 language was enhanced with a new operator "**loop**" on regular formulas over transition sequences. This general iteration operator parameterized by data variables is able to characterize complex (recursively definable) sequences in an LTS. Two auxiliary regular operators "**continue**" and "**exit**" carrying data values were also introduced to express the repetition and the termination of a loop regular formula, respectively. These operators are particularly useful for specifying transition sequences having a particular cumulated cost (e.g., number of transitions, sum of weights associated to actions, etc.) in the context of probabilistic verification (see § 6.3.2 ).
- The MCL v3 language was modified and aligned on MCL v4 by removing syntactic differences that existed between both languages concerning the infinite repetition operator ("@") and the respective precedences of the concatenation (".") and choice ("|") operators in regular formulas. MCL v3 has also been enriched with the option operator ("?") on regular formulas already present in MCL v4.
- Consequently, the two versions of MCL_EXPAND for MCL v3 and MCL v4 have been unified in one single tool, which is now invoked by both EVALUATOR 3 and EVALUATOR 4. The corresponding manual pages have been simplified accordingly, with the introduction of two overarching manual pages ("mcl" and "evaluator"). In addition to five bug fixes, the memory footprint of MCL_EXPAND has been reduced. The error messages displayed by MCL_EXPAND, EVALUATOR 3, and EVALUATOR 4 have been improved in terms of accuracy and explanatory contents.

[0] http://mcc.lip6.fr/nupn.php

- In addition to four bug fixes, the XTL model checker now performs consistency checks on the C identifiers specified by the pragmas "!implementedby", "!comparedby", "!enumeratedby", and "!printedby".
- Two new options were added to the EVALUATOR and XTL model checkers: "-depend", which displays the libraries transitively included in an MCL or XTL file, and "-source", which is used by SVL to display correct file names and line numbers for MCL or XTL formulas embedded in SVL scenarios.

### *6.1.4. Translation of Term Rewrite Systems*

**Participant:** Hubert Garavel.

We pursued the development undertaken in 2015 of a software platform for systematically comparing the performance of rewrite engines and pattern-matching implementations in algebraic specification and functional programming languages. Our platform reuses the benchmarks of the three Rewrite Engine Competitions (2006, 2009, and 2010). Such benchmarks are term-rewrite systems expressed in a simple formalism named REC, for which we developed automated translators that convert REC benchmarks into many languages, among which AProVE, Clean, Haskell, LNT, LOTOS, Maude, mCRL, MLTON, OCAML, Opal, Rascal, Scala, SML-NJ, Stratego/XT, and Tom.

In 2018, we corrected and/or enhanced several of the existing REC translators and finalized experiments. The results of this study have been presented during an invited talk at WRLA'2018 (*12th International Workshop on Rewriting Logic and its Applications*) and an article [15] was published in the WRLA post-proceedings.

### *6.1.5. Formal Modeling and Analysis of BPMN*

**Participant:** Gwen Salaün.

A business process is a set of structured activities that provide a certain service or product. Business processes can be modeled using the BPMN standard, and several industrial platforms have been developed for supporting their design, modeling, and simulation.

In collaboration with Francisco Durán and Camilo Rocha (University of Málaga, Spain), we proposed a rewriting logic executable specification of BPMN with time and extended with probabilities. Duration times and delays for tasks and flows can be specified as stochastic expressions, while probabilities are associated to various forms of branching behavior in gateways. These quantities enable discrete-event simulation and automatic stochastic verification of properties such as expected processing time, expected synchronization time at merge gateways, and domain-specific quantitative assertions. The mechanization of the stochastic analysis tasks is done with Maude's statistical model checker PVeStA. These results led to a publication in an international journal [10].

We also worked on an extension of BPMN with data, which is convenient for describing real-world processes involving complex behavior and data descriptions. By considering this level of expressiveness due to the new features, challenging questions arise regarding the choice of the semantic framework for specifying such an extension of BPMN, as well as how to carry out the symbolic simulation, validation, and assess the correctness of the process models. These issues were addressed first by providing a symbolic executable rewriting logic semantics of BPMN using the rewriting modulo SMT framework, where the execution is driven by rewriting modulo axioms and by querying SMT decision procedures for data conditions. Second, reachability properties, such as deadlock freedom and detection of unreachable states with data exhibiting certain values, can be specified and automatically checked with the help of Maude, thanks to its support for rewriting modulo SMT. These results led to a publication in an international conference [21].

### *6.1.6. Other Language Developments*

**Participants:** Hubert Garavel, Frédéric Lang, Wendelin Serwe.

The ability to compile and verify formal specifications with complex, user-defined operations and data structures is a key feature of the CADP toolbox since its very origins.

In 2018, we enhanced the SYNTAX compiler generator [0] in various ways: (i) The "string manager" has been generalized to allow several symbol tables to be handled simultaneously; (ii) The "source manager" has been extended with new relocation primitives that enable the caller to specify alternative file names and line numbers for the source file being parsed; for instance, this is typically useful for implementing the "#line" pragma of the C preprocessor; this mechanism has been extended to transparently handle multiple relocations (triggered by the lexer) while recognizing the right-hand side of a syntax rule in the grammar; (iii) The "include manager" has been modified to store file names in a distinct symbol table than the table of identifiers, and to provide the list of all files transitively included from the principal module; (iv) Finally, the main programming interface of SYNTAX has been extended with new primitives, so that at present only 5 calls (rather than 9–13 calls, formerly) are required to launch a compiler written using SYNTAX.

All the CADP compilers have been modified to take advantage of the improvements of the SYNTAX library.

Also, a master student started to study an automated translation from Event-B to LNT. He reviewed the syntax and semantics of Event-B and proposed a pencil-paper translation of most Event-B operators. He applied it to a small example consisting of a bank system, where accounts can be created and closed, and money can be deposited or withdrawn. This was a preliminary work that did not lead to a full implementation, due to lack of time. However, this work is a solid basis for a later implementation.

## 6.2. Parallel and Distributed Verification

### *6.2.1. Distributed State Space Manipulation*

**Participant:** Wendelin Serwe.

For distributed verification, CADP provides the PBG format, which implements the theoretical concept of *Partitioned LTS* [44] and provides a unified access to an LTS distributed over a set of remote machines.

In 2018, we improved the usability of distributed state space manipulation tools. In particular:

- A memory shortage error that occurs on a computing node now triggers a distributed termination of the computation, producing proper error messages in the log file of that node.
- A similar naming scheme for log files produced by computing nodes was enforced for all distributed verification tools, which prevents interferences between different invocations of the tools.

### *6.2.2. Debugging of Concurrent Systems using Counterexample Analysis*

**Participants:** Gianluca Barbon, Gwen Salaün.

Model checking is an established technique for automatically verifying that a model satisfies a given temporal property. When the model violates the property, the model checker returns a counterexample, which is a sequence of actions leading to a state where the property is not satisfied. Understanding this counterexample for debugging the specification is a complicated task for several reasons: (i) the counterexample can contain hundreds of actions, (ii) the debugging task is mostly achieved manually, (iii) the counterexample does not explicitly highlight the source of the bug that is hidden in the model, (iv) the most relevant actions are not highlighted in the counterexample, and (v) the counterexample does not give a global view of the problem.

We proposed an approach that improves the usability of model checking by simplifying the comprehension of counterexamples. Our solution aims at keeping only actions in counterexamples that are relevant for debugging purposes. This is achieved by detecting in the models some specific choices between transitions leading to a correct behaviour or falling into an erroneous part of the model. These choices, which we call "neighbourhoods", provide key information for understanding the bug behind the counterexample. To extract such choices, we proposed a first method for debugging the counterexamples of safety property violations. To do so, it builds a new model from the original one containing all the counterexamples, and then compares the two models to identify neighbourhoods.

[0] http://syntax.gforge.inria.fr

In 2018, we proposed a different method for debugging the counterexamples of liveness property violations. Given a liveness property, it extends the model with prefix and suffix information w.r.t. that property. This enriched model is then analysed to identify neighbourhoods. A set of abstraction techniques we developed exploit the enriched model annotated with neighbourhoods to extract relevant actions from counterexamples, which makes their comprehension easier. This work led to a publication in an international conference [16].

Both approaches are fully automated by a tool we implemented and that has been validated on real-world case studies from various application areas. We extended the methodology and tool with 3D visualization techniques to visualize the erroneous part of the model with a specific focus on neighbourhoods, in order to have a global view of the bug behaviour. This work led to a publication to appear in an international conference.

A detailed description of the proposed methodology is available in G. Barbon's PhD thesis [8].

## 6.3. Timed, Probabilistic, and Stochastic Extensions

### *6.3.1. Tools for Probabilistic and Stochastic Systems*

**Participants:** Hubert Garavel, Frédéric Lang.

Formal models and tools dealing with quantitative aspects (such as time, probabilities, and other continuous physical quantities) have become unavoidable for a proper study and computer-aided verification of functional and non-functional properties of cyber-physical systems. The wealth of such formal models is sometimes referred to as a quantitative "zoo" [48].

The CADP toolbox already implements some of these probabilistic/stochastic models, namely DTMCs and CTMCs (*Discrete-Time* and *Continuous-Time Markov Chains*), and IMCs (*Interactive Markov Chains*) [50]. Our long-term goal is to increase the capability and flexibility of the CADP tools, so as to support other quantitative models more easily.

In 2018, BCG_STEADY and BCG_TRANSIENT were enhanced along the following lines:

- They were extended to handle single-state Markov chains and to properly compute state solution vectors and transition throughputs on such models.
- Their command-line options were simplified and warnings are emitted when the input Markov chain contains no stochastic transition.
- A problem which caused correct Markov chains to be rejected was corrected. This problem was due to floating point conversion and rounding errors.
- A confusion between state numbers and matrix indices was fixed in the output and error messages.
- Models containing probabilistic self-loops are now rejected, as was already the case of longer circuits of probabilistic transitions, as both represent similar "timelock" situations.

### *6.3.2. On-the-fly Model Checking for Extended Regular Probabilistic Operators*

**Participant:** Radu Mateescu.

Specifying and verifying quantitative properties of concurrent systems requires expressive and user-friendly property languages combining temporal, data-handling, and quantitative aspects. In collaboration with José Ignacio Requeno (Univ. Zaragoza, Spain), we undertook the quantitative analysis of concurrent systems modeled as PTSs (*Probabilistic Transition Systems*), whose actions contain data values and probabilities. We proposed a new regular probabilistic operator that extends naturally the Until operators of PCTL (*Probabilistic Computation Tree Logic*) [47], by specifying the probability measure of a path characterized by a generalized regular formula involving arbitrary computations on data values. We integrated the regular probabilistic operator into MCL, we devised an associated on-the-fly model checking method based on a combined local resolution of linear and Boolean equation systems, and we implemented the method in a prototype extension of the EVALUATOR model checker.

In 2018, we continued improving and using the extended model checker as follows:

- The model checker now determinizes the dataless regular formulas contained in regular probabilistic operators, ensuring automatically that the linear equation systems produced by the verification of these operators have a unique solution.
- For nondeterministic data-handling regular formulas contained in regular probabilistic operators, the model checker now produces a warning message informing the user that the determinization has to be done manually.
- We carried out further experiments to analyze the quantitative behaviour of the Bounded Retransmission Protocol, namely the variation of the probability of transmission failure w.r.t. the total number of retransmissions attempts.

A paper describing the probabilistic extension of MCL and of the on-the-fly model checker was published in an international journal [13].

## 6.4. Component-Based Architectures for On-the-Fly Verification

### *6.4.1. Compositional Verification*

**Participants:** Hubert Garavel, Frédéric Lang.

The CADP toolbox contains various tools dedicated to compositional verification, among which EXP.OPEN, BCG_MIN, BCG_CMP, and SVL play a central role. EXP.OPEN explores on the fly the graph corresponding to a network of communicating automata (represented as a set of BCG files). BCG_MIN and BCG_CMP respectively minimize and compare behavior graphs modulo strong or branching bisimulation and their stochastic extensions. SVL (*Script Verification Language*) is both a high-level language for expressing complex verification scenarios and a compiler dedicated to this language.

In 2018, we improved these tools along the following lines:

- SVL now invokes EVALUATOR 3, EVALUATOR 4, and XTL with their new "-source" option, so that error and warning messages regarding temporal logic formulas now display line numbers in the SVL file itself, rather than in the temporary files generated to contain the temporal logic formulas, making it easier for users to modify incorrect MCL and XTL formulas contained in SVL files.
- SVL has been modified so that both EVALUATOR 3 and EVALUATOR 4 can now be used to compute "deadlock" and "livelock" statements.
- SVL does not require anymore that every "property" statement contains at least one verification statement, namely "comparison", "verify", "deadlock", "livelock", or a shell-line command with an "expected" clause.
- In addition to a bug fix, the EXP.OPEN tool was enhanced with a new option "-depend", displaying both the list of EXP files included (directly or transitively) in the input EXP file, and the list of automata, hide, rename, and cut files used (directly or transitively) in the input EXP file.

A paper containing both a tutorial and a survey on compositional verification was published in an international conference [14].

### *6.4.2. On-the-Fly Test Generation*

**Participants:** Lina Marsso, Radu Mateescu, Wendelin Serwe.

The CADP toolbox provides support for conformance test case generation by means of the TGV tool. Given a formal specification of a system and a test purpose described as an input-output LTS (IOLTS), TGV automatically generates test cases, which assess using black box testing techniques the conformance of a system under test w.r.t. the formal specification. A test purpose describes the goal states to be reached by the test and enables one to indicate parts of the specification that should be ignored during the testing process. TGV does not generate test cases completely on the fly (i.e., *online*), because it first generates the complete test graph (CTG) and then traverses it backwards to produce controllable test cases.

To address these limitations, we developed the prototype tool TESTOR [0] to extract test cases completely on the fly. TESTOR presents several advantages w.r.t. TGV: (i) it has a more modular architecture, based on generic graph transformation components taken from the OPEN/CAESAR libraries ($\tau$-compression, $\tau$-confluence, $\tau$-closure, determinization, resolution of Boolean equation systems); (ii) it is capable of extracting a test case completely on the fly, by exploiting the diagnostic generation features of the Boolean equation system resolution algorithms; (iii) it enables a more flexible expression of test purposes, taking advantage of the multiway rendezvous, a primitive to express communication and synchronization among a set of distributed processes.

In 2018, we improved TESTOR and TGV as follows:

- TESTOR has been ported to the Windows operating system.
- TESTOR can now be directly connected (by means of Unix pipes) to a system under test (SUT), executing the test case, rather than generating an abstract test-case that has to be connected to the SUT.
- We revised the architecture of TESTOR, so that the interface for the user is more similar to the one of TGV. This enables a user to easily switch between both tools.
- Taking advantage of the similar interfaces, we merged the non-regression test bases of TESTOR and TGV.
- We also fixed a bug and added a new option "-self" to TGV, reducing the number of warning messages.

These activities led to a new version 3.0 of TESTOR and two publications in international conferences [24], [18].

### *6.4.3. Other Component Developments*

**Participants:** Pierre Bouvier, Hubert Garavel, Frédéric Lang, Radu Mateescu, Wendelin Serwe.

In 2018, several components of CADP have been improved as follows:

- The CADP toolbox now contains a new tool named SCRUTATOR for pruning Labeled Transition Systems on the fly.
- The OPEN/CAESAR environment was enriched with a new SOLVE_2 library for solving linear equation systems on the fly.
- Two manual pages ("bes" and "seq") have been added, which provide standalone definitions of CADP's BES format for Boolean Equation Systems and SEQ format for execution traces. The OPEN/CAESAR manual pages have been enhanced to give full prototypes for function parameters.
- The CADP toolbox has been ported to Solaris 11 and to SunOS 5.11 OpenIndiana "Hipster". CADP has also been ported to macOS 10.14 "Mojave" and a 64-bit version of CADP is now available for macOS.
- We also designed new C functions for handling path names in order to replace the traditional POSIX primitives basename(), dirname(), and realpath(), which suffer from limitations and ambiguities.

## 6.5. Real-Life Applications and Case Studies

### *6.5.1. Autonomous Resilience of Distributed IoT Applications in a Fog Environment*

**Participants:** Umar Ozeer, Gwen Salaün.

Fog computing provides computing, storage and communication resources (and devices) at the edge of the network, near the physical world (PW). These end-devices nearing the physical world can have interesting properties such as short delays, responsiveness, optimized communications and privacy, which are especially appealing to IoT (Internet of Things) applications. However, IoT devices in the fog have low stability and are prone to failures.

[0] http://convecs.inria.fr/software/testor

In the framework of the collaboration with Orange Labs (see § 7.1.1 ), we are working on the key challenge of providing reliable services. This may be critical in this context since the non-containment of failures may impact the physical world. For instance, the failure of a smoke detector or a lamp in a smart home for elderly/medicated people may be hazardous. The design of such resilience solutions is complex due to the specificities of the environment, i.e., (i) dynamic infrastructure, where entities join and leave without synchronization; (ii) high heterogeneity in terms of functions, communication models, network, processing and storage capabilities; and (iii) cyber-physical interactions, which introduce non-deterministic and physical world's space and time dependent events.

In 2018, our work focused on proposing an end-to-end resilience approach for stateful IoT applications in the fog taking into account the three specificities mentioned above. The resilience protocol is functionally divided into four phases: (i) state-saving; (ii) monitoring and failure detection; (iii) failure notification and reconfiguration; and (iv) decision and recovery. The protocol implements a combination of different state-saving techniques based on rules and policies to cope with the heterogeneous nature of the environment and recover from failures in a consistent way, including PW-consistency. This work led to a publication in an international conference [25].

To illustrate our protocol at work, we mounted a smart home testbed with objects that can be found in real-life smart homes to test our solution. Our resilience approach was also implemented as a framework and deployed onto the testbed. The empirical results showed that multiple failures are recovered in an acceptable time in regard to end users. This work led to a publication to appear in an international conference.

### *6.5.2. Verified Composition and Deployment of IoT Applications*

**Participants:** Radu Mateescu, Ajay Muroor Nadumane, Gwen Salaün.

The Internet of Things (IoT) is an interconnection of physical devices and software entities that can communicate and perform meaningful tasks largely without human intervention. The design and development of IoT applications is an interesting problem as these applications are typically dynamic, distributed and, more importantly, heterogeneous in nature.

In the framework of the collaboration with Nokia Bell Labs (see § 7.1.2 ), we proposed to build and deploy reliable IoT applications using a series of steps: (i) IoT objects and compositions are described using an interface-based behavioural model; (ii) the correctness of the composition is ensured by checking a behavioural compatibility notion that we proposed for IoT systems; and (iii) finally, a deployment plan respecting the dependencies between the objects is generated to facilitate automated deployment and execution of the application.

Regarding implementation, behavioural models and composition are specified in LNT and we take advantage of the CADP toolbox to perform compatibility checks. The deployment is automated using the Majord'Home platform developed by Nokia Bell Labs. The entire implementation is packaged as a Web tool available for end-users. This work led to a publication to appear in an international conference.

### *6.5.3. Memory Protection Unit*

**Participants:** Hubert Garavel, Radu Mateescu, Wendelin Serwe.

Asynchronous circuits have key advantages in terms of low energy consumption, robustness, and security. However, the absence of a global clock makes the design prone to deadlock, livelock, synchronization, and resource-sharing errors. Formal verification is thus essential for designing such circuits, but it is not widespread enough, as many hardware designers are not familiar with it and few verification tools can cope with asynchrony on complex designs. In the framework of the SECURIOT-2 project (see § 8.2.2.1 ), we are interested in the rigorous design of asynchronous circuits used in the secure elements for IoT devices developed in the project.

In collaboration with Aymane Bouzafour and Marc Renaudin (Tiempo Secure), we suggested an extension of Tiempo's industrial design flow for asynchronous circuits, based upon the standard Hardware Description Language SystemVerilog (SV), with the formal verification capabilities provided by CADP. This was achieved by translating SV descriptions into LNT, expressing correctness properties in MCL, and verifying them using the EVALUATOR model checker of CADP. It turned out that the constructs of SV and LNT are in close correspondence, and that the synthesizable SV subset can be entirely translated into LNT. The MCL language was also shown adequate for expressing all property patterns relevant for asynchronous circuits.

The practicality of the approach was demonstrated on an asynchronous circuit (4000 lines of SV) implementing a memory protection unit (MPU). The MPU block exhibits a high degree of internal concurrency, comprising 660 parallel execution flows and 250 internal communication channels. The corresponding state space was generated compositionally, by identifying a suitable minimization and composition strategy described in SVL (the largest intermediate state space had more than 116 million states and 862 million transitions). A set of 184 MCL properties were successfully verified on the state space, expressing the correct initialization of the MPU configuration registers, the mutual exclusion of read and write operations on registers, the correct responses to stimuli, and the security requirements related to the many access-control policies enforced by the MPU. This work led to a publication in an international conference [17].

### *6.5.4. TLS 1.3 Handshake Protocol*

**Participants:** Lina Marsso, Radu Mateescu.

Security services are extensively used in fields like online banking, e-government, online shopping, etc. To ensure a secure communication between peers in terms of authenticity, privacy, and data integrity, cryptographic protocols are applied to regulate the data transfer. These protocols provide a standardized set of rules and methods for the interaction between peers. The Transport Layer Security (TLS) is a widely used security protocol, encompassing a set of rules for the communication between clients and servers, and relying on public-key cryptography to ensure integrity of exchanged data. However, despite multiple prevention measurements, several vulnerabilities (such as Heartbleed and DROWN), have been discovered recently. Therefore, testing the implementations of security protocols is still a crucial issue.

In the framework of the RIDINGS PHC project (see § 8.3.1 ), we are interested in testing protocols and distributed systems. In collaboration with Josip Bozic and Franz Wotawa (TU Graz, Austria), we undertook the formal modelling of the draft TLS 1.3 handshake protocol [0]. Taking as input the informal description of TLS 1.3 in the draft standard, we developed a formal model (1293 lines of LNT) specifying the handshake messages and client-server interactions. As far as we are aware, this is the first formal model of the draft TLS 1.3 handshake.

We used our LNT model for conformance testing with the OpenSSL version 1.0.1e implementation of the TLS protocol [0]. We defined three test purposes specifying requirements from the draft TLS 1.3 handshake, and applied the newly developed TESTOR tool (see § 6.4.2 ) to generate the test cases from the LNT model and each test purpose. The execution of these test cases on the OpenSSL implementation spotted a discrepancy of the server's response to a client certificate request w.r.t. the draft TLS 1.3 standard. This work led to a publication in an international workshop [18].

### *6.5.5. Message Authenticator Algorithm*

**Participants:** Hubert Garavel, Lina Marsso.

The Message Authenticator Algorithm (MAA) is one of the first cryptographic functions for computing a Message Authentication Code. Between 1987 and 2001, the MAA was adopted in international standards (ISO 8730 and ISO 8731-2) to ensure the authenticity and integrity of banking transactions. The MAA also played a role in the history of formal methods, as the National Physical Laboratory (NPL, United Kingdom) developed, in the early 90s, three formal, yet non-executable, specifications of the MAA in VDM, Z, and LOTOS abstract data types.

[0] https://tools.ietf.org/html/draft-ietf-tls-tls13-24
[0] https://www.openssl.org/

In 2018, we examined how the new generation of formal methods can cope with the MAA case study. We specified the MAA in both LOTOS and LNT and checked these specifications using the CADP tools. The C code generated by the CADP compilers was executed w.r.t. a set of reference MAA test vectors, as well as supplementary test vectors devised to improve the coverage of byte permutations and message segmentation. This enabled us to detect and correct several errors in the reference test vectors given in the ISO 8730 and ISO 8731-2 standards. This work led to a publication in an international workshop [22].

### *6.5.6. Other Case Studies*

**Participants:** Hubert Garavel, Frédéric Lang, Lina Marsso, Radu Mateescu, Wendelin Serwe.

Based on the work described above, the demo examples of the CADP toolbox have been enriched. Two new demo examples have been added: demo_06 (Transport Layer Security v1.3 handshake protocol specified in LNT), and demo_11 (a hardware block implementing a Dynamic Task Dispatcher). The demo_12 (Message Authenticator Algorithm) is now documented in a publication [22]. The demo_17 (distributed leader election protocol) has been converted from LOTOS to LNT. Finally, most existing demo examples have been updated to reflect the evolution of the MCL v3 and SVL languages.

**CORSE Project-Team**

# 7. New Results

## 7.1. Profiling Feedback based Optimizations and Performance Debugging

**Participants:** Fabrice Rastello, Diogo Sampaio, Fabian Gruber, Christophe Guillon [STMicroelectronics], Antoine Moynault [STMicroelectronics], Changwan Hong [OSU, USA], Aravind Sukumaran-Rajam [OSU, USA], Jinsung Kim [OSU, USA], Prashant Singh Rawat [OSU, USA], Sriram Krishnamoorthy [PNNL, USA], Louis-Noël Pouchet [CSU, USA], P. Sadayappan [OSU, USA].

Profiling feedback is an important technique used by developers for performance debugging, where it is usually used to pinpoint performance bottlenecks and also to find optimization opportunities. Our contributions in this area are twofold: (1) we developed a new technique that combines abstract simulation and sensitive analysis that allows to pinpoint performance bottleneck; (2) we developed a new technique to build a polyhedral representation out of an execution trace that allows to provide feedback on possible missed transformations.

### *7.1.1. Compiler Optimization for GPUs Using Bottleneck Analysis*

Optimizing compilers generally use highly simplified performance models due to the significant challenges in developing accurate analytical performance models for complex computer systems. In this work, we develop an alternate approach to performance modeling using abstract execution of GPU kernel binaries. We use the performance model to predict the bottleneck resource for a given kernel's execution through differential analysis by performing multiple abstract executions with varying machine parameters. The bottleneck analysis is then used to develop an automated search through a configuration space of different grid reshaping, thread/block coarsening, and loop unrolling factors. Experimental results using a number of benchmarks from the Parboil/Rodinia/SHOC suites demonstrate the effectiveness of the approach. The bottleneck analysis is also shown to be useful in assisting high-level domain-specific code generators for GPUs.

This work is the fruit of the collaboration 9.4.1.1 with OSU. It has been presented at the ACM/SIGPLAN conference on Programming Language Design and Implementation, PLDI 2018.

### *7.1.2. Data-Flow/Dependence Profiling for Structured Transformations*

Profiling feedback is an important technique used by developers for performance debugging, where it is usually used to pinpoint performance bottlenecks and also to find optimization opportunities. Assessing the validity and potential benefit of a program transformation requires accurate knowledge of the data flow and data dependencies, which can be uncovered by profiling a particular execution of the program.

In this work we develop Mickey, an end-to-end infrastructure for dynamic binary analysis, which produces feedback about the potential to apply structured transformations to uncover non-trivial parallelism and data locality via complex program rescheduling. Our tool can handle both inter- and intraprocedural aspects of the program in a unified way, thus providing structured interprocedural transformation feedback.

This work is the fruit of the collaboration 9.4.1.1 with CSU and the past collaboration Nano2017 with STMicroelectronics. It has been submitted for presentation at the ACM conference on Principles and Practice of Parallel Programming, PPoPP 2019.

## 7.2. Combined Scheduling and Register Allocation

**Participants:** Prashant Singh Rawah [OSU, USA], Aravind Sukumaran-Rajam [OSU, USA], Atanas Rountev [OSU, USA], Fabrice Rastello, Louis-Noël Pouchet [CSU, USA], Atanas Rountev [OSU, USA], P. Sadayappan [OSU, USA].

Register allocation is one of the most studied compiler optimization but its impact on performance is highly coupled with scheduling. Recent advances on computer simulation and artificial intelligence lead to application kernels with very high register pressure. Our contributions in this area consist in developing new scheduling schemes that both expose SIMD parallelism and register reuse.

### *7.2.1. Register Optimizations for Stencils on GPUs*

The recent advent of compute-intensive GPU architecture has allowed application developers to explore high-order 3D stencils for better computational accuracy. A common optimization strategy for such stencils is to expose sufficient data reuse by means such as loop unrolling, with the hope of register-level reuse. However, the resulting code is often highly constrained by register pressure. While the current state-of-the-art register allocators are satisfactory for most applications, they are unable to effectively manage register pressure for such complex high-order stencils, resulting in a sub-optimal code with a large number of register spills. In this work, we develop a statement reordering framework that models stencil computations as DAG of trees with shared leaves, and adapts an optimal scheduling algorithm for minimizing register usage for expression trees. The effectiveness of the approach is demonstrated through experimental results on a range of stencils extracted from application codes.

This work is the fruit of the collaboration 9.4.1.1 with OSU. It has been presented at the ACM/SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPoPP 2018.

### *7.2.2. Associative instruction reordering to alleviate register pressure*

Register allocation is generally considered a practically solved problem. For most applications, the register allocation strategies in production compilers are very effective in controlling the number of loads/stores and register spills. However, existing register allocation strategies are not effective and result in excessive register spilling for computation patterns with a high degree of many-to-many data reuse, e.g., high-order stencils and tensor contractions. We develop a source-to-source instruction reordering strategy that exploits the flexibility of reordering associative operations to alleviate register pressure. The developed transformation module implements an adaptable strategy that can appropriately control the degree of instruction-level parallelism, while relieving register pressure. The effectiveness of the approach is demonstrated through experimental results using multiple production compilers (GCC, Clang/LLVM) and target platforms (Intel Xeon Phi, and Intel x86 multi-core).

This work is the fruit of the collaboration 9.4.1.1 with OSU. It has been presented at ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis, SC 2018.

## 7.3. Runtime Verification and Monitoring

**Participants:** Raphael Jakse, Yliès Falcone, Jean Francois Mehaut, Srdan Krstic, Giles Reger, Dmitriy Traytel, Hosein Nazarpour, Mohamad Jaber, Marius Bozga, Saddek Bensalem, Salwa Kobeissi, Adnan Utayim.

We report on several contributions related with the runtime verification and monitoring of systems. We address several aspects such as the instrumentation, the understanding and classification of existing concepts and tools, the definition of frameworks for monitoring distributed systems and a case study on monitoring smart homes.

### *7.3.1. Interactive Runtime Verification: Formal Models, Algorithms, and Implementation*

Interactive runtime verification (i-RV) combines runtime verification and interactive debugging. Runtime verification consists in studying a system at runtime, looking for input and output events to discover, check or enforce behavioral properties. Interactive debugging consists in studying a system at runtime in order to discover and understand its bugs and fix them, inspecting interactively its internal state. We define an efficient and convenient way to check behavioral properties automatically on a program using a debugger. We aim at helping bug discovery and understanding by guiding classical interactive debugging techniques using runtime verification.

In this work, we provide a formal model for a program execution under a debugger, which we compose with a general model of a monitor and a scenario to model the interactively verified program. We provide guarantees on the verdicts issued by the monitor using the instrumentation provided by the debugger. We provide an algorithmic view of this model suitable for producing implementations, and we present Verde, an implementation based on GDB to interactively verify C programs. We built a set of experiments using Verde to assess usefulness of Interactive Runtime Verification and performance of our implementation. Our results show that though debugger-based instrumentation incurs non-trivial performance costs, i-RV is appliable performance-wise in a variety of cases and helps studying bugs.

This work has been submitted at the ACM Transactions on Software Engineering and Methodology (TOSEM).

### *7.3.2. A Taxonomy for Classifying Runtime Verification Tools*

Over the last 15 years Runtime Verification (RV) has grown into a diverse and active field, which has stimulated the development of numerous theoretical frameworks and tools. Many of the tools are at first sight very different and challenging to compare. Yet, there are similarities. In this work, we classify RV tools within a high-level taxonomy of concepts. We first present this taxonomy and discuss the different dimensions. Then, we survey RV tools and classify them according to the taxonomy. This work constitutes a snapshot of the current state of the art and enables a comparison of existing tools.

This work has been published in the proceedings of the 18th International Conference on Runtime Verification.

### *7.3.3. Bringing Runtime Verification Home*

We use runtime verification (RV) to check various specifications in a smart apartment. The specifications can be broken down into three types: be- havioral correctness of the apartment sensors, detection of specific user activities (known as activities of daily living), and composition of specifications of the previous types. The context of the smart apartment provides us with a complex system with a large number of components with two different hierarchies to group specifications and sensors: geographically within the same room, floor or globally in the apartment, and logically following the different types of specifications. We leverage a recent approach to decentralized RV of decentralized specifications, where monitors have their own specifications and communicate together to verify more general specifications. This allows us to re-use specifications, and combine them to: (1) scale beyond existing centralized RV techniques, and (2) greatly reduce computation and communication costs.

This work has been published in the proceedings of the 18th International Conference on Runtime Verification.

### *7.3.4. Tracing Distributed Component-Based Systems, a Brief Overview*

We overview a framework for tracing asynchronous distributed component-based systems with multiparty interactions managed by distributed schedulers. Neither the global state nor the total ordering of the system events is available at runtime. We instrument the system to retrieve local events from the local traces of the schedulers. Local events are sent to a global observer which reconstructs on-the-fly the global traces that are compatible with the local traces, in a concurrency-preserving and communication-delay insensitive fashion. The global traces are represented as an original lattice over partial states, such that any path of the lattice projected on a scheduler represents the corresponding lo- cal partial trace according to that scheduler (soundness), and all possible global traces of the system are recorded (completeness).

This work has been published in the proceedings of the 18th International Conference on Runtime Verification.

### *7.3.5. Can We Monitor All Multithreaded Programs?*

Runtime Verification(RV)is a lightweight formal method which consists in verifying that an execution of a program is correct wrt a specification. The specification formalizes with properties the expected correct behavior of the system. Programs are instrumented to extract necessary information from the execution and feed it to monitors tasked with checking the properties. From the perspective of a monitor, the system is a black box; the trace is the only system information provided. Parallel programs generally introduce an added level of complexity on the program execution due to concurrency. A concurrent execution of a parallel program is

best represented as a partial order. A large number of RV approaches generate monitors using formalisms that rely on total order, while more recent approaches utilize formalisms that consider multiple traces.

We made a tutorial where we review some of the main RV approaches and tools that handle multithreaded Java programs. We discuss their assumptions, limitations, expressiveness, and suitability when tackling parallel programs such as producer- consumer and readers-writers. By analyzing the interplay between specification formalisms and concurrent executions of programs, we identify four questions RV practitioners may ask themselves to classify and determine the situations in which it is sound to use the existing tools and approaches.

This work has been published in the proceedings of the 18th International Conference on Runtime Verification.

### *7.3.6. Facilitating the Implementation of Distributed Systems with Heterogeneous Interactions*

We introduce HDBIP an extension of the Behavior Interaction Priority (BIP) framework. BIP is a component-based framework with a rigorous operational semantics and high-level and expressive interaction model. HD-BIP extends BIP interaction model by allowing heterogeneous interactions targeting distributed systems. HD-BIP allows both multiparty and direct send/receive interactions that can be directly mapped to an underlying communication library. Then, we present a correct and efficient code generation from HDBIP to C++ implementation using Message Passing Interface (MPI). We present a non-trivial case study showing the effectiveness of HDBIP.

This work has been published in the proceedings of the 14th International Conference on integrated Formal Methods.

### *7.3.7. Modularizing Behavioral and Architectural Crosscutting Concerns in Formal Component-Based Systems*

We define a method to modularize crosscutting concerns in Component-Based Systems (CBSs) expressed using the Behavior Interaction Priority (BIP) framework. Our method is inspired from the Aspect Oriented Programming (AOP) paradigm which was initially conceived to support the separation of concerns during the development of monolithic systems. BIP has a formal operational semantics and makes a clear separation between architecture and behavior to allow for compositional and incremental design and analysis of systems. We distinguish local from global aspects. Local aspects model concerns at the component level and are used to refine the behavior of components. Global aspects model concerns at the architecture level, and hence refine communications (synchronization and data transfer) between components. We formalize local and global aspects as well as their composition and integration into a BIP system through rigorous transformation primitives. We present AOP-BIP, a tool for Aspect-Oriented Programming of BIP systems, demonstrate its use to modularize logging, security, and fault tolerance in a network protocol, and discuss its possible use in runtime verification of CBSs.

This work has been published in the Journal of Logical and Algebraic Methods in Programming.

## 7.4. Numa MeMory Analyzer

**Participants:** François Trahay [Télécom SudParis], Manuel Selva, Lionel Morel [CEA], Kevin Marquet [INSA Lyon].

Non Uniform Memory Access (NUMA) architectures are nowadays common for running High-Performance Computing (HPC) applications. In such architectures, several distinct physical memories are assembled to create a single shared memory. Nevertheless, because there are several physical memories, access times to these memories are not uniform depending on the location of the core performing the memory request and on the location of the target memory. Hence, threads and data placement are crucial to efficiently exploit such architectures. To help in taking decision about this placement, profiling tools are needed. NUMA MeMory Analyzer (NumaMMA) is a new profiling tool for understanding the memory access patterns of HPC applications. NumaMMA combines efficient collection of memory traces using hardware mechanisms with original visualization means allowing to see how memory access patterns evolve over time. The information reported by NumaMMA allows to understand the nature of these access patterns inside each object allocated

by the application. We show how NumaMMA can help understanding the memory patterns of several HPC applications in order to optimize them and get speedups up to 28% over the standard non optimized version.

This work has been published in the 47th International Conference on Parallel Processing - ICPP 2018.

## 7.5. Towards an Easier Way to Program FPGAs in an HPC Context

**Participants:** Georgios Christodoulis, Manuel Selva, Francois Broquedis, Frederic Desprez, Olivier Muller [TIMA].

Heterogeneity in HPC nodes appears as a promising solution to improve the execution of a wide range of scientific applications, regarding both performance and energy consumption. Unlike CPUs and GPUs, FPGAs can be configured to fit the application needs, making them an appealing target to extend traditional heterogeneous HPC architectures. However, exploiting them requires an in-depth knowledge of low-level hardware and high expertise on vendor-provided tools, which should not be the primary concern of HPC application programmers. In the context of the Persyval HEAVEN project, we proposed a framework enabling a more straightforward development of scientific applications over FPGA enhanced platforms. Our solution requires the minimum knowledge of the underlying architecture, as well as fewer changes to the existing code. To fulfill these requirements, we extended the StarPU task programming library that initially targets heterogeneous architectures to support FPGAs. We used Vivado HLS, a high-level synthesis tool to deliver efficient hardware implementations of the tasks from high-level languages like C/C++. Our solution, validated on a blocking version of the matrix multiplication algorithm, offers an easier way to exploit FPGAs from an HPC application. We also conducted some preliminary experiments to validate our proof-of-concept implementation regarding performance.

This work has been published in the 13th International Symposium on Reconfigurable Communication-centric Systems-on-Chip and obtained the best paper award.

## 7.6. Automatic IPC Profile Analysis to Detect Phases in HPC Application

**Participants:** Mathieu Stoffel, François Broquedis, Frederic Desprez, Abdelhafid Mazouz [Atos/Bull], Philippe Rols [Atos/Bull].

Mathieu Stoffel started his PhD in February 2018 on a CIFRE contract with Atos/Bull. The purpose of this work is to enhance the energy consumption of HPC applications on large-scale platforms. The first phase of the thesis project consists in an in-depth study of the evolution of the metrics characterizing the state of the supercomputer during the execution of a highly parallel application. Indeed, the utilization rates of the different components of the HPC system may demonstrate extreme variations during the execution of the aforementioned application. These variations are sometimes subject to repeat themselves on a regular basis during the application execution. We refer to this phenomena as application "phases". In this context, we already generated precise IPC profiles out of many benchmarks and real-life applications and we worked on a methodology to adapt the CPU frequency based on these profiles. This part of the thesis has been published in an IEEE Cluster workshop (HPCMASPA). Currently, we are working on a detection tool for the application phases. It will implement an automated reconfiguration of the parameters of the HPC system during the execution of the application, in relation with the type of phase being executed. By doing so, the tool will aim at optimizing the energy consumption associated with the execution of the application, by adapting the state of the HPC systems all along the aforesaid execution.

## 7.7. Teaching of Algorithms, Programming, and Debugging

**Participants:** Florent Bouchez-Tichadou, Theo Barollet, Aurelien Flori, Thomas Herve.

### 7.7.1. Teaching Algorithms using Problem and Challenge Based Learning

Teaching algorithms is always a challenge at any level of the CS curriculum, as it is often viewed as a theoretical field. While many exercises revolve around classical examples that illustrate interesting algorithmic points, they are often disconnected from reality, which is a major drawback for students trying to learn. During the last four years, we have been trying to reconnect the teaching of algorithms with their applicability in the real world to M1 and L2 students, by giving them actual problems that could arise in their life of future software engineers, challenging enough to force them to use particular algorithmic techniques or data structures—e.g., linked lists, binary trees, dynamic programming or approximation algorithms.

By assigning students in groups of 5 to 6 members, we wanted to create an environment where they function as a team trying to work together to solve a problem. This allowed them to help each other in their respective comprehension, and made them more autonomous in their learning. The effective materials was provided as online pdf files so they had to read and learn from them by themselves, while the class sessions with a tutor (teacher) where used for the problem-solving part, with guidance from the tutor (who is there to make sure the learning takes place).

After four years of experimentation with M1 students, we found that the student's grades were stable, in particular there was no decrease in exams performances compared to the classical course that was taught in the previous years. However, the students progressed in trans-disciplinary skills such a communication and the writing of essays. More importantly, students show a strong adhesion to the teaching method, 50% of them rating it as "excellent" (6) and 25% as "good" (resp. 6 and 5 on a scale from 1 (terrible) to 6 (excellent)). No student rated the course below average.

This work has been published in the 23rd International Conference on Innovation and Technology in Computer Science Education, ITiCSE 2018.

### 7.7.2. Data Structures Visualization at Runtime

Debuggers are powerful tools to observe a program behaviour and find bugs but they are not often used by developers and especially beginners because of the hard learning curve of such tools. They provide information on low level data but are not able to analyze higher level elements such as data structures. This work tries to provide a more intuitive representation of the program execution to ease debugging and algorithms understanding.

We have a basic prototype, Moly, which is a GDB extension (GNU Project Debugger) to explore a program runtime memory and analyze its data structures. It also provides an interface with an external visualizer, Lotos, through a formatted output. Running Moly along with a dedicated visualizer should allow a programmer to spot bugs easier by seeing the subsequent whole memory states of the program and some data structures information.

The current status of Moly allows a programmer to explore all attainable memory at any point during the debug process, and already provides minimum information about the possible properties of the data structures, such as recognizing graphs, trees, or linked lists. Future work includes recognizing access patterns to the structures to extract for instance visit patterns and higher-level properties (such as the breaking of data structure properties between break-points).

The external visualizer, Lotos, is still in its early stages of development and was enough to make a proof-of-concept that it is possible to display via a web browser the information gathered by Moly. Our plans is to redesign this part from scratch using the knowledge gaining during the writing of this prototype.

### 7.7.3. AppoLab: an Online Platform to Engage Students in Their Learning

Classical teaching of algorithms and low-level data structures at the L2 european level is often tedious and unappealing to students, with much of the time being spent on analysing and devising algorithms for textbook cases, such as sorting lists of integers, visiting linked lists or trees, etc.

Using Problem-Based Learning helps to alleviate this problem, by presenting more complex problems to handle, hence engaging more students in their learning. This work revolves around the design of a learning platform that includes gamification in PBL. AppoLab is in its core a server that has scripted "exercices". Students can communicate with the server either manually, using telnet; but ultimately, they will need to script the communication also from their side, since the server will gradually impose constraints on the problems such as timeouts or large input sizes.

This preliminary work was used this year in some parts of an Algorithm course at the L2 level, and has received positive feedback from the students. This encourages us to continue this development and study more precisely the impact it has on students' engagement in their learning.

## CTRL-A Project-Team

# 6. New Results

## 6.1. Programming support for Autonomic Computing

### *6.1.1. Reactive languages*

**Participants:** Gwenaël Delaval, Eric Rutten.

Our work in reactive programming for autonomic computing systems is focused on the specification and compilation of declarative control objectives, under the form of contracts, enforced upon classical mode automata as defined in synchronous languages. The compilation involves a phase of Discrete Controller Synthesis, integrating the tool ReaX, in order to obtain an imperative executable code. The programming language Heptagon / BZR (see Section Software and Platforms) integrates our research results [6].

Recent work concerns exploring new possibilities offered by logico-numeric control. We consider Symbolic Limited Lookahead Control for Best-effort Dynamic Computing Resource Management. We put forward a new modeling technique for Dynamic Resource Management (DRM) based on discrete events control for symbolic logico-numerical systems, especially Discrete Controller Synthesis (DCS). The resulting models involve state and input variables defined on an infinite domain (Integers), thereby no exact DCS algorithm exists for safety control. We thus formally define the notion of limited lookahead, and associated best-effort control objectives targeting safety and optimization on a sliding window for a number of steps ahead. We give symbolic algorithms, illustrate our approach on an example model for DRM, and report on performance results based on an implementation in the tool ReaX. This work is in cooperation with the Sumo team at Inria Rennes (Hervé Marchand) and University of Liverpool (Nicolas Berthier), and is published in the WODES 2018 conference [14].

We also have ongoing activities on abstraction methods for compilation using discrete controller synthesis (needed for example, in order to program the controllers for systems where the useful data for control can be of arbitrary types (integer, real, ...) , or also for systems which are naturally distributed, and require a decentralized controller) and on compilation and diagnosis for discrete controller synthesis (which is made special by the declarative nature of the compilation, where it is not easy to precisely diagnose cases where no solution can be found).

On the applicative side, we also consider such modular and logico-numeric approaches for the control of different targets in self-adaptive and reconfigurable systems (see below in Section 6.2.2.1 and 6.2.1.3 [20], [15].

### *6.1.2. Domain-specific languages*

**Participants:** Gwenaël Delaval, Soguy Mak Kare Gueye, Eric Rutten.

Our work in Domain-specific languages (DSLs) is founded on our work in component-based programming for autonomic computing systems as examplified by e.g., FRACTAL. We consider essentially the problem of specifying the control of components assembly reconfiguration, with an approach based on the integration within such a component-based framework of a reactive language as in Section 6.1.1 [5].

In recent work, we proposed an extension of a classical Software Architecture Description Languages (ADL) with Ctrl-F, DSL for the specification of dynamic reconfiguration behavior in a [1].

Based on this experience, we are working on a proposal for a DSL called Ctrl-DPR, allowing designers to easily generate Autonomic Managers for DPR FPGA systems (see Section 6.2.1.3 ). Users can describe their system and their management strategies, in terms of the entities composing the system : tasks, versions, applications, ressources, policies. The DSL relies on a behavioral modeling of these entities, targeted at the design of autonomic managers to control the reconfigurations in such a way as to enforce given policies and strategies. These model-based control techniques are embedded in a compiler, connected to the reactive language and discrete controller synthesis tool of Section 6.1.1 , which enables to generate a C implementation of the controller enforcing the management strategies. We apply our DSL for the management of a video application on a UAV. This work is in cooperation with LabSticc in Lorient (Jean-Philippe Diguet), and is published in the ICAC 2018 conference [16].

Ongoing work involves a generalization from our experiences in software components, DPR FPGA, as well as Rule-based autonomic manager as in Section 6.1.3 . As we observed a similarity in objects and structures (e.g., tasks, implmentation versions, resources, and upper-level application layer), we are considering a more general DSL, which could be specialized towards such different target domains, and where the compilation towards reactive models could be studied and improved, especially considering the features of Section 6.1.1 . This direction will also lead us to study the definition of architectural patterns for multiple loop Autonomic Managers, particularly hierarchical, with lower layers autonomy alleviating management burden from the upper layers.

### *6.1.3. Rule-based systems*

**Participants:** Adja Sylla, Gwenaël Delaval, Eric Rutten.

This work concerns a high-level language for safe rule-based programming in the LINC transactional rule-based platform developed at CEA [10]. Rule based middlewares such as LINC enable high level programming of distributed adaptive systems behaviours. LINC also provides the systems with transactional guarantees and hence ensures their reliability at runtime. However, the set of rules may contain design errors (e.g. conflicts, violations of constraints) that can bring the system in unsafe safe or undesirables states, despite the guarantees provided by LINC. On the other hand, automata based languages such as Heptagon/BZR enable formal verification and especially synthesis of discrete controllers to deal with design errors. Our work studies these two languages and combines their execution mechanisms, from a technical perspective. We target applications to the domain of Internet of Things and more particularly smart building, office or home (see Section 6.2.2.1 ).

This work is in cooperation with CEA LETI/DACLE (Maxime Louvel), it was the topic of the PhD of Adja Sylla at CEA, co-advised with M. Louvel, and aspects on applications of logico-numeric control are published in the CCTA 2018 conference [20].

## 6.2. Design methods for reconfiguration controller design in computing systems

We apply the results of the previous axes of the team's activity, as well as other control techniques, to a range of infrastructures of different natures, but sharing a transversal problem of reconfiguration control design. From this very diversity of validations and experiences, we draw a synthesis of the whole approach, towards a general view of Feedback Control as MAPE-K loop in Autonomic Computing [23] [9].

### *6.2.1. High-Performance Computing*

**Participants:** Agustin Yabo, Soguy Mak Kare Gueye, Gwenaël Delaval, Stéphane Mocanu, Bogdan Robu, Eric Rutten.

#### *6.2.1.1. Automated regulation and software transactional memory*

A parallel program needs to manage the trade-off between the time spent in synchronisation and computation. This trade-off is significantly affected by its parallelism degree. A high parallelism degree may decrease computing time while increasing synchronisation cost. We performed work on dynamic control of thread parallelism and mapping. We address concurrency issues via Software Transactional Memory (STM). We implement feedback control loops to automate management of threads and diminish program execution time.

This work was performed in the framework on the PhD of Naweiluo Zhou, and published in the journal on Concurrency and Computation: Practice and Experience [13].

*6.2.1.2. A Control-Theory based approach to minimize cluster underuse*

HPC systems are facing more and more variability in their behavior, related to e.g., performance and power consumption, and the fact that they are less predictable requires more runtime management. One such problem is found in the context of CiGri, a simple, lightweight, scalable and fault tolerant grid system which exploits the unused resources of a set of computing clusters. This work resulted in first results addressing the problem of automated resource management in an HPC infrastructure, using techniques from Control Theory to design a controller that maximizes cluster utilization while avoiding overload. We put in place a mechanism for feedback (Proportional Integral, PI) control system software, through a maximum number of jobs to be sent to the cluster, in response to system information about the current number of jobs processed. Additionally, we developed a Model-Predictive Controller to improve the performance of the system.

This work is done in cooperation with the Datamove team of Inria/LIG, and Gipsa-lab. It was the topic of the Master's thesis of Agustin Yabo [25]. Preliminary results were published in the AIScience workshop (Autonomous Infrastructure for Science) of the HPDC conference [19].

*6.2.1.3. Reconfiguration control in DPR FPGA*

6.2.1.3.1. DPR FPGA and discrete control for reconfiguration

Implementing self-adaptive embedded systems, such as UAV drones, involves an offline provisioning of the several implementations of the embedded functionalities with different characteristics in resource usage and performance in order for the system to dynamically adapt itself under uncertainties. We propose an autonomic control architecture for self-adaptive and self-reconfigurable FPGA-based embedded systems. The control architecture is structured in three layers: a mission manager, a reconfiguration manager and a scheduling manager. In this work we focus on the design of the reconfiguration manager. We propose a design approach using automata-based discrete control. It involves reactive programming that provides formal semantics, and discrete controller synthesis from declarative objectives.

This work is in the framework of the ANR project HPeC (see Section 8.2.1 ), and is published in the International Workshop on High Performance and Dynamic Reconfigurable Systems and Networks (DRSN 2018), part of the HPCS 2018 conference [17] ; for the evaluation of the apllication of logico-numeric control, in the CCTA 18 conference [15] ; for the proposal of a Domain Specific Language, in the ICAC 2018 conference [16].

6.2.1.3.2. Mission management and stochastic control

In the Mission Management workpackage of the ANR project HPeC, a concurrent control methodology is constructed for the optimal mission planning of a U.A.V. in stochastic environnement. The control approach is based on parallel ressource sharing Partially Observable Markov Decision Processes modeling of the mission. The parallel POMDP are reduced to discrete Markov Decision Models using Bayesian Networks evidence for state identification. The control synthesis is an iterative two step procedure : first MDP are solved for the optimisation of a finite horizon cost problem ; then the possible ressource conflicts between parallel actions are solved either by a priority policy or by a QoS degradation of actions, e.g., like using a lower resolution version of the image processing task if the ressource availability is critical.

This work was performed in the framework on the PhD of Chabha Hireche, and published in the journal on Sensors [24], [12].

### *6.2.2. IoT*

**Participants:** Neïl Ayeb, Adja Sylla, Gwenaël Delaval, Stéphane Mocanu, Eric Rutten.

*6.2.2.1. Control of smart buildings*

A smart environment is equipped with numerous devices (i.e., sensors, actuators) that are possibly distributed over different locations (e.g., rooms of a smart building). These devices are automatically controlled to achieve different objectives related, for instance, to comfort, security and energy savings. Our work proposes a design framework based on the combination of the rule based middleware LINC and the automata based language Heptagon/BZR (H/BZR). It consists of: an abstraction layer for the heterogeneity of devices, a transactional execution mechanism to avoid inconsistencies and a controller that, based on a generic model of the environment, makes appropriate decisions and avoids conflicts. A case study with concrete devices, in the field of building automation, is presented to illustrate the framework.

This work is in the framework of the cooperation with CEA (see Section 7.1 ), and is published in the CCTA 2018 conference [20].

*6.2.2.2. Device management*

The research topic is targeting an adaptative and decentralized management for the IoT. It will contribute design methods for processes in virtualized gateways in order to enhance IoT infrastructures. More precisely, it concerns Device Management in the case of large numbers of connected sensors and actuators, as can be found in Smart Home and Building, Smart Electricity grids, and industrial frameworks as in Industry 4.0.

This work is in the framework of the Inria/Orange labs joint laboratory (see Section 7.2.1 ), and supported by the CIFRE PhD thesis grant of Neïl Ayeb, starting dec. 2017.

*6.2.2.3. Security in SCADA industrial systems*

We focus mainly on vulnerability search, automatic attack vectors synthesis and intrusion detection. Model checking techniques are used for vulnerability search and automatic attack vectors construction. Intrusion detection is mainly based on process-oriented detection with a technical approach from run-time monitoring. The LTL formalism is used to express safety properties which are mined on an attack-free dataset. The resulting monitors are used for fast intrusion detections.

A demonstrator of attack/defense scenario in SCADA systems will be built on the existing G-ICS lab (hosted by ENSE3/Grenoble-INP).

This work is in the framework of the ANR project Sacade on cybersecurity of industrial systems (see Section 8.2.2 ) [18] [22] [21].

The work is also supported by Grenoble Alpes Cybersecurity Institute (see Section 8.1.1 ).

Ongoing work concerns the complementary topic of analysis and identification of reaction mechanisms for self-protection in cybersecurity, where, beyond classical defense mechanisms that detect intrusions and attacks or assess the kind of danger that is caused by them, we explore models and control techniques for the automated reaction to attacks, in order to use detection information to take the appropriate defense and repair actions.

**DANTE Project-Team**

# 7. New Results

## 7.1. Graph Signal Processing and Machine Learning

**Participants:** Paulo Gonçalves, Esteban Bautista Ruiz, Mikhail Tsitsvero, Sarah de Nigris.

### *7.1.1. Analytic signal in many dimensions*

In a series of two articles [30] and [54] (in collaboration with P. Borgnat), we extended analytic signal to the multidimensional case. First we showed how to obtain separate phase-shifted components and how to combine them into instantaneous amplitude and phase. Secondly we defined the proper hypercomplex analytic signal as a holomorphic hypercomplex function on the boundary of polydisk in the hypercomplex space. Next it was shown that the correct phase-shifted components can be obtained by positive frequency restriction of the Scheffers-Fourier transform based on the commutative and associative algebra generated by the set of elliptic hypercomplex numbers. Moreover we demonstrated that for d > 2 there is no corresponding Clifford-Fourier transform that allows to recover phase-shifted components correctly. Finally the euclidean-domain construction of instantaneous amplitude was extended to manifold and manifold-like graphs and point clouds.

### *7.1.2. BGP Zombies: an Analysis of Beacons Stuck Routes*

Joint work with Romain Fontugne (IIJ Research Lab, Japan) and Patrice Abry (CNRS, Physics Lab of ENS de Lyon) [25].

Network operators use the Border Gateway Protocol (BGP) to control the global visibility of their networks. When withdrawing an IP prefix from the Internet, an origin network sends BGP withdraw messages, which are expected to propagate to all BGP routers that hold an entry for that address space in their routing table. Yet network operators occasionally report issues where routers maintain routes to IP prefixes withdrawn by their origin network. We refer to this problem as BGP zombies and characterize their appearance using RIS BGP beacons, a set of prefixes withdrawn every four hours at predetermined times. Across the 27 monitored beacon prefixes, we observe usually more than one zombie outbreak per day. But their presence is highly volatile, on average a monitored peer misses 1.8% withdraws for an IPv4 beacon (2.7% for IPv6). We also discovered that BGP zombies can propagate to other ASes, for example, zombies in a transit network are inevitably affect- ing its customer networks. **We employ a graph-based semi-supervised machine learning technique to estimate the scope of zombies propagation**, and found that most of the observed zombie outbreaks are small (i.e. on average 10% of monitored ASes for IPv4 and 17% for IPv6). We also report some large zombie outbreaks with almost all monitored ASes affected.

### *7.1.3. Design of graph filters and filterbanks*

Book chapter [43], co-authored with Nicolas Tremblay (CNRS, UGA Gipsa-Lab) and Pierre Borgnat (CNRS, Physics Lab, ENS de Lyon).

Basic operations in graph signal processing consist in processing signals indexed on graphs either by filtering them or by changing their domain of representation, in order to better extract or analyze the important information they contain. The aim of this chapter is to review general concepts underlying such filters and representations of graph signals. We first recall the different Graph Fourier Transforms that have been developed in the literature, and show how to introduce a notion of frequency analysis for graph signals by looking at their variations. Then, we move to the introduction of graph filters, that are defined like the classical equivalent for 1D signals or 2D images, as linear systems which operate on each frequency band of a signal. Some examples of filters and of their implementations are given. Finally, as alternate representations of graph signals, we focus on multiscale transforms that are defined from filters. Continuous multiscale transforms such as spectral wavelets on graphs are reviewed, as well as the versatile approaches of filterbanks on graphs. Several variants of graph filterbanks are discussed, for structured as well as arbitrary graphs, with a focus on the central point of the choice of the decimation or aggregation operators.

## 7.2. Optimization

**Participant:** Marion Foare.

### *7.2.1. A new proximal method for joint image restoration and edge detection with the Mumford-Shah model*

Joint work with Nelly Pustelnik (CNRS, Physics Lab of ENS de Lyon) and Laurent Condat (CNRS, GIPSA Lab) [24].

In this paper, we propose an adaptation of the PAM algorithm to the minimization of a nonconvex functional designed for joint image denoising and contour detection. This new functional is based on the Ambrosio–Tortorelli approximation of the well-known Mumford–Shah functional. We motivate the proposed approximation, offering flexibility in the choice of the possibly non-smooth penalization, and we derive closed form expression for the proximal steps involved in the algorithm. We focus our attention on two types of penalization: 1-norm and a proposed quadratic-1 function. Numerical experiments show that the proposed method is able to detect sharp contours and to reconstruct piecewise smooth approximations with low computational cost and convergence guarantees. We also compare the results with state-of-the-art re-laxations of the Mumford–Shah functional and a recent discrete formulation of the Ambrosio–Tortorelli functional.

### *7.2.2. Semi-Linearized Proximal Alternating Minimization for a Discrete Mumford–Shah Model*

Joint work with Nelly Pustelnik (CNRS, Physics Lab of ENS de Lyon) and Laurent Condat (CNRS, GIPSA Lab) [51].

The Mumford–Shah model is a standard model in image segmentation and many approximations have been proposed in order to approximate it. The major interest of this functional is to be able to perform jointly image restoration and contour detection. In this work, we propose a general formulation of the discrete counterpart of the Mumford–Shah functional, adapted to nonsmooth penalizations, fitting the assumptions required by the Proximal Alternating Linearized Minimization (PALM), with convergence guarantees. A second contribution aims to relax some assumptions on the involved functionals and derive a novel Semi-Linearized Proximal Alternated Minimization (SL-PAM) algorithm, with proved convergence. We compare the performances of the algorithm with several nonsmooth penalizations, for Gaussian and Poisson denoising, image restoration and RGB-color denoising. We compare the results with state-of-the-art convex relaxations of the Mumford–Shah functional, and a discrete version of the Ambrosio–Tortorelli functional. We show that the SL-PAM algorithm is faster than the original PALM algorithm, and leads to competitive denoising, restoration and segmentation results.

### *7.2.3. Discrete Mumford-Shah on graph for mixing matrix estimation*

Joint work with Yacouba Kaloga (Physics Lab of ENS de Lyon), Nelly Pustelnik (CNRS, Physics Lab of ENS de Lyon) and Pablo Jensen (CNRS, Physics Lab of ENS de Lyon) [53].

The discrete Mumford-Shah formalism has been introduced for the image denoising problem, allowing to capture both smooth behavior inside an object and sharp transitions on the boundary. In the present work, we propose first to extend this formalism to graphs and to the problem of mixing matrix estimation. New algorithmic schemes with convergence guarantees relying on proximal alternating minimization strategies are derived and their efficiency (good estimation and robustness to initialization) are evaluated on simulated data, in the context of vote transfer matrix estimation.

## 7.3. Wireless & Wired Networks

**Participants:** Thomas Begin, Anthony Busson, Isabelle Guérin Lassous.

### 7.3.1. Conflict graph-based model for IEEE 802.11 networks: A Divide-and-Conquer approach

WLANs (Wireless Local Area Networks) based on the IEEE 802.11 standard have become ubiquitous in our daily lives. We typically augment the number of APs (Access Points) within a WLAN to extend its coverage and transmission capacity. This leads to network densification, which in turn demands some form of coordination between APs so as to avoid potential misconfigurations. In our article [20], we describe a performance modeling method that can provide guidance for configuring WLANs and be used as a decision-support tool by a network architect or as an algorithm embedded within a WLAN controller. The proposed approach estimates the attained throughput of each AP, as a function of the WLAN's conflict graph, the AP loads, the frame sizes, and the link transmission rates. Our modeling approach employs a Divide-and-Conquer strategy which breaks down the original problem into multiple sub-problems, whose solutions are then combined to provide the solution to the original problem. We conducted extensive simulation experiments using the ns-3 simulator that show the model's accuracy is generally good with relative errors typically less than 10%. We then explore two issues of WLAN configuration: choosing a channel allocation for the APs and enabling frame aggregation on APs.

### 7.3.2. Video on Demand in IEEE 802.11p-based Vehicular Networks: Analysis and Dimensioning

This is a joint work with A. Boukerche. In [31], we consider a VoD (Video on-Demand) platform designed for vehicles traveling on a highway or other major roadway. Typically, cars or buses would subscribe to this delivery service so that their passengers get access to a catalog of movies and series stored on a back-end server. Videos are delivered through IEEE 802.11p Road Side Units deployed along the highway. In this paper, we propose a simple analytical and yet accurate solution to estimate (at the speed of a click) two key performance parameters for a VoD platform: (i) the total amount of data down-loaded by a vehicle over its journey and (ii) the total "interruption time" , which corresponds to the time a vehicle spends with the playback of its video interrupted because of an empty buffer. After validating its accuracy against a set of simulations run with ns-3, we show an example of application of our analytical solution for the sizing of an IEEE 802.11p-based VoD platform.

### 7.3.3. An accurate and efficient modeling framework for the performance evaluation of DPDK-based virtual switches

This is a joint work with B. Baynat, G. Artero Gallardo and V. Jardin [4]. Data plane development kit (DPDK) works as a specialized library that enables virtual switches to accelerate the processing of incoming packets by, among other things, balancing the incoming flow of packets over all the CPU cores and processing packets by batches to make a better use of the CPU cache. Although DPDK has become a de facto standard, the performance modeling of a DPDK-based vSwitch remains a challenging problem. In this paper, we present an analytical queueing model to evaluate the performance of a DPDK-based vSwitch. Such a virtual equipment is represented by a complex polling system in which packets are processed by batches, i.e., a given CPU core processes several packets of one of its attached input queues before switching to the next one. To reduce the complexity of the associated model, we develop a general framework that consists in decoupling the polling system into several queueing subsystems, each one corresponding to a given CPU core. We resort to servers with vacation to capture the interactions between subsystems. Our proposed solution is conceptually simple, easy to implement and computationally efficient. Tens of comparisons against a discrete-event simulator show that our models typically deliver accurate estimates of the performance parameters of interest (e.g., attained throughput, packet latency or loss rate). We illustrate how our models can help in determining an adequate setting of the vSwitch parameters using several real-life case studies.

### 7.3.4. Association optimization in Wi-Fi networks

Densification of Wi-Fi networks has led to the possibility for a wireless station to choose between several access points (APs), improving coverage, wireless link quality and mobility. But densification of APs may generate interference, contention and decrease the global throughput as these APs have to share a limited number of channels. The recent trend in which Wi-Fi networks are managed in a centralized way offers

the opportunity to alleviate this problem through a global optimization of the resource usage. In particular, optimizing the association step between APs and stations can increase the overall throughput and fairness between stations. In this work, we propose an original solution to this optimization problem. First, we propose a mathematical model to evaluate and forecast the throughput achieved for each station for a given association. The best association is then defined as the one that maximizes a logarithmic utility function using the stations' throughputs predicted by the model. The use of a logarithmic utility function allows to achieve a good trade-off between overall throughput and fairness. A heuristic based on a local search algorithm is used to propose approximate solutions to this optimization problem. This approach has the benefit to be tuned according to the CPU and time constraints of the WLAN controller. A comparison between different heuristic versions and the optimum solution shows that the proposed heuristic offers solutions very close to the optimum with a significant gain of time.

In the first place, we consider a saturated network. Even if such traffic conditions are rare, the optimization of the association step under this assumption has the benefit to fairly share the bandwidth between stations. Nevertheless, traffic demands may be very different from one station to another and it may be more useful to optimize associations according to the stations' demands. In a second step, we propose an optimization of the association step based on the stations' throughputs and the channel busy time fraction (BTF). The latter is defined as the proportion of time the channel is sensed busy by an AP. We propose an analytical model that predicts BTF for any configuration. Associations are optimized in order to minimize the greatest BTF in the network. This original approach allows the Wi-Fi manager to unload the most congested AP, increase the throughput for most of the stations, and offer more bandwidth to stations that need it. We present a local search technique that finds local optima to this optimization problem. This heuristic relies on an analytical model that predicts BTF for any configuration. The model is based on a Markov network and a Wi-Fi conflict graph. NS-3 simulations including a large set of scenarios highlight the benefits of our approach and its ability to improve the performance in congested and non-congested Wi-Fi networks.

Lastly, we consider the latest amendments of the IEEE 802.11 standard. The main challenges are to propose models that take into account recent enhancements such as spatial multiplexing (MIMO) at the physical layer and frame aggregation mechanism at the MAC layer. To assess these new features, we derive an association optimization approach based on a new metric, named Hypothetical Busy Time Fraction (H-BTF), that combines the classical Busy Time Fraction (BTF) and the frame aggregation mechanism [3].

### *7.3.5. Transient analysis of idle time in VANETs using Markov-reward models*

The development of analytical models to analyze the behavior of vehicular ad hoc networks (VANETs) is a challenging aim. Adaptive methods are suitable for many algorithms (*e.g.* choice of forwarding paths, dynamic resource allocation, channel control congestion) and services (*e.g.* provision of multimedia services, message dissemination). These adaptive algorithms help the network to maintain a desired performance level. However, this is a difficult goal to achieve, especially in VANETs due to fast position changes of the VANET nodes. Adaptive decisions should be taken according to the current conditions of the VANET. Therefore, evaluation of transient measures is required for the characterization of VANETs. In the literature, different works address the characterization and measurement of the idle (or busy) time to be used in different proposals to attain a more efficient usage of wireless network. We focus on the idle time of the link between two VANET nodes. Specifically, we have developed an analytical model based on a straightforward Markov reward chain (MRC) to obtain transient measurements of this idle time. Numerical results from the analytical model fit well with simulation results [12].

## 7.4. Performance Evaluation of Communication Networks

**Participants:** Thomas Begin, Philippe Nain, Isabelle Guérin Lassous.

### *7.4.1. First-Come-First-Served Queues with Multiple Servers and Customer Classes*

This a joint work with A. Brandwajn [5]. We present a simple approach to the solution of a multi-server FCFS queueing system with several classes of customers and phase-type service time distributions. The proposed

solution relies on solving a single two-class model in which we distinguish one of the classes and we aggregate the remaining customer classes. We use a reduced state approximation to solve this two-class model. We propose two types of aggregation: exact, in which we merge the phase-type service time distributions exactly, and approximate, in which we simplify the phase-type distribution for the aggregated class by matching only its first two moments. The proposed approach uses simple mathematics and is highly scalable in terms of the number of servers, the number of classes, as well as the number of phases per class. Our approach applies both to queues with finite and infinite buffer space.

### 7.4.2. A study of systems with multiple operating levels, probabilistic thresholds and hysteresis

This a joint work with A. Brandwajn [6]. Current architecture of many computer systems relies on dynamic allocation of a pool of resources according to workload conditions to meet specific performance objectives while minimizing cost (e.g., energy or billing). In such systems, different levels of operation may be defined, and switching between operating levels occurs at certain thresholds of system congestion. To avoid rapid oscillations between levels of service, "hysteresis" is introduced by using different thresholds for increasing and decreasing workload levels, respectively. We propose a model of such systems with general arrivals, arbitrary number of servers and operating levels where each higher operating level may correspond to an arbitrary number of additional servers and soft (i.e. non-deterministic) thresholds to account for "inertia" in switching between operating levels. In our model, request service times are assumed to be memoryless and server processing rates may be a function of the current operating level and of the number of requests (users) in the system. Additionally, we allow for delays in the activation of additional operating levels. We use simple mathematics to obtain a semi-numerical solution of our model. We illustrate the versatility of our model using several case study examples inspired by features of real systems. In particular, we explore optimal thresholds as a tradeoff between performance and energy consumption.

### 7.4.3. Covert cycle stealing in an M/G/1 queue

Consider an M/G/1 queue where arriving jobs are under control of a party (Willie). There exists a second party, Alice who may or may not want to introduce a sequence of jobs to be serviced. Her goal is to prevent Willie from being able to distinguish between these two cases. The question that we address is: can Alice introduce her stream of jobs covertly, i.e., prevent Willie from distinguishing between the two possibilities, either her introducing the stream or not, and if so, at what rate can she introduce her jobs? We present a square-root law on the amount of service Alice can receive covertly. The covertness criterion is that the probabilities of false alarm and missed detection is arbitrarily close to one. One result we have established is the following: consider exponential service times for Alice's jobs and Willies' jobs with rate $\mu_1$ and $\mu_2$, respectively. During $n$ Willie's job busy periods, Alice can submit covertly $O(\sqrt{n})$ jobs if $\mu_1 < 2\mu_2$, $O(\sqrt{n/\log n})$ jobs if $\mu_1 = 2\mu_2$, and $O(n^{\mu_1/\mu_2})$ jobs if $\mu_1 > 2\mu_2$. This is the first time that such a phase transition has been observed in this context. This ongoing research, carried out by P. Nain in collaboration with D. Towsley (Univ. Massachusetts) and B. Jiang (Shanghai Jiao Tong Univ.), has various applications in the context of service level agreement.

### 7.4.4. LRU caches

The work on network caches operating under the standard Least-Recently-Used (LRU) management policy, initiated in 2017 (see 2017 Dante Activity Report), has been completed and published [13]. Under weak statistical assumptions on the content request process, this work establishes the validity of the so-called "Che's approximation" as the cache size and the number of content go to infinity.

### 7.4.5. Stochastic Multilayer Networks

A stochastic multilayer network is the aggregation of $M$ networks (one per layer) where each is a subgraph of a foundational network $G$. Each layer network is the result of probabilistically removing links and nodes from $G$. The resulting network includes any link that appears in at least $K$ layers. This model is an instance of a non-standard site-bond percolation model. Two sets of results are obtained in [28]: first, we derive the probability distribution that the $M$-layer network is in a given configuration for some particular graph structures (explicit results are provided for a line and an algorithm is provided for a tree), where a configuration is the collective state of all links (each either active or inactive). Next, we show that for appropriate scalings of the node and link

selection processes in a layer, links are asymptotically independent as the number of layers goes to infinity, and follow Poisson distributions. Numerical results are provided to highlight the impact of having several layers on some metrics of interest (including expected size of the cluster a node belongs to in the case of the line). This model finds applications in wireless communication networks with multichannel radios, multiple social networks with overlapping memberships, transportation networks, and, more generally, in any scenario where a common set of nodes can be linked via co-existing means of connectivity.

## 7.5. Computational Human Dynamics and Temporal Networks

**Participants:** Márton Karsai, Éric Fleury, Jean-Philippe Magué, Philippe Nain, Jean-Pierre Chevrot.

### *7.5.1. Correlations and dynamics of consumption patterns in social-economic networks*

In [16], we analyse a coupled dataset collecting the mobile phone communications and bank transactions history of a large number of individuals living in a Latin American country [16]. After mapping the social structure and introducing indicators of socioeconomic status, demographic features, and purchasing habits of individuals, we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of stratification in the social structure. In addition, we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Finally, by analysing individual consumption histories, we detect dynamical patterns in purchase behaviour and their correlations with the socioeconomic status, demographic characters and the egocentric social network of individuals. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in resource allocation, marketing, and recommendation system design.

### *7.5.2. Mapping temporal-network percolation to weighted, static event graphs*

The dynamics of diffusion-like processes on temporal networks are influenced by correlations in the times of contacts. This influence is particularly strong for processes where the spreading agent has a limited lifetime at nodes: disease spreading (recovery time), diffusion of rumors (lifetime of information), and passenger routing (maximum acceptable time between transfers). In [14], we introduce weighted event graphs as a powerful and fast framework for studying connectivity determined by time-respecting paths where the allowed waiting times between contacts have an upper limit. We study percolation on the weighted event graphs and in the underlying temporal networks, with simulated and real-world networks. We show that this type of temporal-network percolation is analogous to directed percolation, and that it can be characterized by multiple order parameters.

### *7.5.3. Randomized reference models for temporal networks*

Many real-world dynamical systems can successfully be analyzed using the temporal network formalism. Empirical temporal networks and dynamic processes that take place in these situations show heterogeneous, non-Markovian, and intrinsically correlated dynamics, making their analysis particularly challenging. Randomized reference models (RRMs) for temporal networks constitute a versatile toolbox for studying such systems. Defined as ensembles of random networks with given features constrained to match those of an input (empirical) network, they may be used to identify statistically significant motifs in empirical temporal networks (i.e. over-represented w.r.t. the null random networks) and to infer the effects of such motifs on dynamical processes unfolding in the network. However, the effects of most randomization procedures on temporal network characteristics remain poorly understood, rendering their use non-trivial and susceptible to misinterpretation. In the work presented in [52], we propose a unified framework for classifying and understanding microcanonical RRMs (MRRMs). We use this framework to propose a canonical naming convention for existing randomization procedures, classify them, and deduce their effects on a range of important temporal network features. We furthermore show that certain classes of compatible MRRMs may be applied in sequential composition to generate more than a hundred new MRRMs from existing ones surveyed in this article. We provide a tutorial for the use of MRRMs to analyze an empirical temporal network and we review applications of MRRMs found

in literature. The taxonomy of MRRMs we have developed provides a reference to ease the use of MRRMs, and the theoretical foundations laid here may further serve as a base for the development of a principled and systematic way to generate and apply randomized reference null models for the study of temporal networks.

### *7.5.4. Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis*

Our usage of language is not solely reliant on cognition but is arguably determined by myriad external factors leading to a global variability of linguistic patterns. This issue, which lies at the core of sociolinguistics and is backed by many small-scale studies on face-to-face communication, is addressed in [29], by constructing a dataset combining the largest French Twitter corpus to date with detailed socioeconomic maps obtained from national census in France. We show how key linguistic variables measured in individual Twitter streams depend on factors like socioeconomic status, location, time, and the social network of individuals. We found that (i) people of higher socioeconomic status, active to a greater degree during the daytime, use a more standard language; (ii) the southern part of the country is more prone to use more standard language than the northern one, while locally the used variety or dialect is determined by the spatial distribution of socioeconomic status; and (iii) individuals connected in the social network are closer linguistically than disconnected ones, even after the effects of status homophily have been removed. Our results inform sociolinguistic theory and may inspire novel learning methods for the inference of socioeconomic status of people from the way they tweet.

### *7.5.5. Threshold driven contagion on weighted networks*

Weighted networks capture the structure of complex systems where interaction strength is meaningful. This information is essential to a large number of processes, such as threshold dynamics, where link weights reflect the amount of influence that neighbours have in determining a node's behaviour. Despite describing numerous cascading phenomena, such as neural firing or social contagion, the modelling of threshold dynamics on weighted networks has been largely overlooked. We fill this gap in [21], by studying a dynamical threshold model over synthetic and real weighted networks with numerical and analytical tools. We show that the time of cascade emergence depends non-monotonously on weight heterogeneities, which accelerate or decelerate the dynamics, and lead to non-trivial parameter spaces for various networks and weight distributions. Our methodology applies to arbitrary binary state processes and link properties, and may prove instrumental in understanding the role of edge heterogeneities in various natural and social phenomena.

### *7.5.6. Link transmission centrality in large-scale social networks*

Understanding the importance of links in transmitting information in a network can provide ways to hinder or postpone ongoing dynamical phenomena like the spreading of epidemic or the diffusion of information. In our work [22], we propose a new measure based on stochastic diffusion processes, the*transmission centrality*, that captures the importance of links by estimating the average number of nodes to whom they transfer information during a global spreading diffusion process. We propose a simple algorithmic solution to compute transmission centrality and to approximate it in very large networks at low computational cost. Finally we apply transmission centrality in the identification of weak ties in three large empirical social networks, showing that this metric outperforms other centrality measures in identifying links that drive spreading processes in a social network.

### *7.5.7. Prepaid or Postpaid? That Is the Question: Novel Methods of Subscription Type Prediction in Mobile Phone Services*

In the paper [41], we investigate the behavioural differences between mobile phone customers with prepaid and postpaid subscriptions. Our study reveals that (a) postpaid customers are more active in terms of service usage and (b) there are strong structural correlations in the mobile phone call network as connections between customers of the same subscription type are much more frequent than those between customers of different subscription types. Based on these observations, we provide methods to detect the subscription type of customers by using information about their personal call statistics, and also their egocentric networks simultaneously. The key of our first approach is to cast this classification problem as a problem of graph labelling, which can be solved by max-flow min-cut algorithms. Our experiments show that, by using both

user attributes and relationships, the proposed graph labelling approach is able to achieve a classification accuracy of $\sim 87\%$, which outperforms by $\sim 7\%$ supervised learning methods using only user attributes. In our second problem, we aim to infer the subscription type of customers of external operators. We propose via approximate methods to solve this problem by using node attributes, and a two-way indirect inference method based on observed homophilic structural correlations. Our results have straightforward applications in behavioural prediction and personal marketing.

### *7.5.8. Service Adoption Spreading in Online Social Networks*

The collective behaviour of people adopting an innovation, product or online service is commonly interpreted as a spreading phenomenon throughout the fabric of society. This process is arguably driven by social influence, social learning and by external effects like media. Observations of such processes date back to the seminal studies by Rogers and Bass, and their mathematical modelling has taken two directions: One paradigm, called simple contagion, identifies adoption spreading with an epidemic process. The other one, named complex contagion, is concerned with behavioural thresholds and successfully explains the emergence of large cascades of adoption resulting in a rapid spreading often seen in empirical data. The observation of real-world adoption processes has become easier lately due to the availability of large digital social network and behavioural datasets. This has allowed simultaneous study of network structures and dynamics of online service adoption, shedding light on the mechanisms and external effects that influence the temporal evolution of behavioural or innovation adoption. These advancements have induced the development of more realistic models of social spreading phenomena, which in turn have provided remarkably good predictions of various empirical adoption processes. In our chapter [39], we review recent data-driven studies addressing real-world service adoption processes. Our studies provide the first detailed empirical evidence of a heterogeneous threshold distribution in adoption. We also describe the modelling of such phenomena with formal methods and data-driven simulations. Our objective is to understand the effects of identified social mechanisms on service adoption spreading, and to provide potential new directions and open questions for future research.

### *7.5.9. Attention on Weak Ties in Social and Communication Networks*

Granovetter's weak tie theory of social networks is built around two central hypotheses. The first states that strong social ties carry the large majority of interaction events; the second maintains that weak social ties, although less active, are often relevant for the exchange of especially important information (e.g., about potential new jobs in Granovetter's work). While several empirical studies have provided support for the first hypothesis, the second has been the object of far less scrutiny. A possible reason is that it involves notions relative to the nature and importance of the information that are hard to quantify and measure, especially in large scale studies. In our work [48], we search for empirical validation of both Granovetter's hypotheses. We find clear empirical support for the first. We also provide empirical evidence and a quantitative interpretation for the second. We show that attention, measured as the fraction of interactions devoted to a particular social connection, is high on weak ties—possibly reflecting the postulated informational purposes of such ties—but also on very strong ties. Data from online social media and mobile communication reveal network-dependent mixtures of these two effects on the basis of a platform's typical usage. Our results establish a clear relationships between attention, importance, and strength of social links, and could lead to improved algorithms to prioritize social media content.

**DATAMOVE Project-Team**

# 6. New Results

## 6.1. Integration of High Performance Computing and Data Analytics

### *6.1.1. I/O Survey*

First contribution is a comprehensive survey on parallel I/O in the HPC context [14]. As the available processing power and amount of data increase, I/O remains a central issue for the scientific community. This survey focuses on a traditional I/O stack, with a POSIX parallel file system. Through the comprehensive study of publications from the most important conferences and journals in a five-year time window, we discuss the state of the art of I/O optimization approaches, access pattern extraction techniques, and performance modeling, in addition to general aspects of parallel I/O research. This survey enables us to identify the general characteristics of the field and the main current and future research topics.

### *6.1.2. Task Based In Situ Processing*

One approach to bypass the I/O bottleneck is *in situ* processing, an important research topic at DataMove. The *in situ* paradigm proposes to reduce data movement and to analyze data while still resident in the memory of the compute node by co-locating simulation and analytics on the same compute node. The simplest approach consists in modifying the simulation timeloop to directly call analytics routines. However, several works have shown that an *asynchronous* approach where analytics and simulation run concurrently can lead to a significantly better performance. Today, the most efficient approach consists in running the analytics processes on a set of dedicated cores, called helper cores, to isolate them from the simulation processes. Simulation and analytics thus run concurrently on different cores but this static isolation can lead to underused resources if the simulation or the analytics do not fully use all the assigned cores.

In this work performed in collaboration with CEA, we developed TINS, a task-based in situ framework that implements a novel *dynamic helper core* strategy. TINS relies on a work stealing scheduler and on task-based programming. Simulation and analytics tasks are created concurrently and scheduled on a set of worker threads created by a single instance of the work stealing scheduler. Helper cores are assigned dynamically: some worker threads are dedicated to analytics when analytics tasks are available while they join the other threads for processing simulation tasks otherwise, leading to a better resource usage. We leverage the good compositionality properties of task-based programming to seamlessly keep the analytics and simulation codes well separated and a plugin system enables to develop parallel analytics codes outside of the simulation code.

TINS is implemented with the Intel Threading Building Blocks (TBB) library that provides a task-based programming model and a work stealing scheduler. The experiments are conducted with the hybrid MPI+TBB ExaStamp molecular dynamics code that we associate with a set of analytics representative of computational physics algorithms. We show up to $40\%$ performance improvement over various other approaches, including the standard helper core, on experiments on up to 14,336 Broadwell cores.

### *6.1.3. Stream Processing*

Stream processing is the Big Data equivalent of in situ processing. It consists in analyzing on-line incoming streams of data, often produced from sensors or social networks like Twitter. We investigated the convergence between both paradigms through different directions: how the programming environment developed specifically for stream processing can applied to the data produced by large parallel simulations [18]; Proposing a dynamics data structure to keep sorted data streams [12]; Evaluating the performance of the FlameMR framework on data produced from a parallel simulation[13]. We summarize here the 2 first contributions.

*6.1.3.1. Packed Memory QuadTree.*

Over the past years, several in-memory big-data management systems have appeared in academia and industry. In-memory databases systems avoid the overheads related to traditional I/O disk-based systems and have made possible to perform interactive data-analysis over large amounts of data. A vast literature of systems and research strategies deals with different aspects, such as the limited storage size and a multi-level memory-hierarchy of caches. Maintaining the right data layout that favors locality of accesses is a determinant factor for the performance of in-memory processing systems. Stream processing engines like Spark or Flink support the concept of *window*, which collects the latest events without a specific data organization. It is possible to trigger the analysis upon the occurrence of a given criterion (time, volume, specific event occurrence). After a window is updated, the system shifts the processing to the next batch of events. There is a need to go one step further to keep a live window continuously updated while having a fine grain data replacement policy to control the memory footprint. The challenge is the design of dynamic data structures to absorb high rate data streams, stash away the oldest data to stay in the allowed memory budget while enabling fast queries executions to update visual representations. A possible solution is the extension of database structures like R-trees used in SpatiaLite or PostGis, or to develop dedicated frameworks like Kit based on a pyramid structure.

We developed a novel self-organized cache-oblivious data structure, called PMQ, for in-memory storage and indexing of fixed length records tagged with a spatiotemporal index. We store the data in an array with a controlled density of gaps (*i.e.*, empty slots) that benefits from the properties of the *Packed Memory Arrays*. The empty slots guarantee that insertions can be performed with a low amortized number of data movements ($(O(\log^2(N)))$) while enabling efficient spatiotemporal queries. During insertions, we rebalance parts of the array when required to respect density constraints, and the oldest data is stashed away when reaching the memory budget. To spatially subdivide the data, we sort the records according to their Morton index, thus ensuring spatial locality in the array while defining an implicit, recursive quadtree, which leads to efficient spatiotemporal queries. We validate PMQ for consuming a stream of tweets to answer visual and range queries. PMQ significantly outperforms the widely adopted spatial indexing data structure R-tree, typically used by relational databases, as well as the conjunction of Geohash and $B^+$-tree, typically used by NoSQL databases.

*6.1.3.2. Flink based in situ Processing.*

We proposed to leverage Apache Flink, a scalable stream processing engine from the Big Data domain, in this HPC context. Flink enables to program analyses within a simple window based map/reduce model, while the runtime takes care of the deployment, load balancing and fault tolerance. We build a complete in transit analytics workflow, connecting an MD simulation to Apache Flink and to a distributed database, Apache HBase, to persist all the desired data. To demonstrate the expressivity of this programming model and its suitability for HPC scientific environments, two common analytics in the Molecular Dynamics field have been implemented. We assessed the performance of this framework, concluding that it can handle simulations of sizes used in the literature while providing an effective and versatile tool for scientists to easily incorporate on-line parallel analytics in their current workflows.

## 6.2. Data Aware Batch Scheduling

### *6.2.1. Batch Scheduling for Energy*

The project COSMIC [24], [22], [16], [17], in collaboration with Myriads team in Inria Rennes-Atlantique, targets the optimization of green energy usage in Clouds. The project considers a geographically distributed cloud, with each data center associated with a local photovoltaic (PV) farm. The objective is to maximize the photovoltaic energy by allocation the computing workload to the data centers according to its energy production. The production forecasting is modeled with a truncated normal law, permitting to consider the uncertainty of the forecast.

Chapter [24] considers a simple model with homogeneous Virtual Machines submitted at unpredictable rate. This study has resulted in a scheduling algorithm for task allocation. The chapter demonstrates the optimality of this algorithm at current time slot according to production forecast parameters.

Paper [22] extends these results to heterogeneous VM. Each VM is defined by its arrival date, its execution time, its memory requirement and its CPU usage. In this model, due to execution time durations, the possibility to migrate running VM was considered. An algorithm is detailed in the paper that is compared to standard algorithm through simulations.

A third study [16], [17] has carefully modeled the interactions between the Cloud and the energy supplier. Due to variability of PV production and workload submission, each data center will alternatively inject energy into the electricity grid or purchase energy. The energy model considers a virtual energy pool mitigating the surplus and deficit of the different data center, with reduced costs regarding the difference between electricity cost and electricity injection tariff. The algorithm detailed in this paper outperforms well-known round-robin approaches, as shown by simulations.

### 6.2.2. Learning Methods for Batch Scheduling

Most of Job Scheduling algorithms apply greedy tasks ordering, as First Come First Served (FCFS) or Shortest Processing time First (SPF). They give simple methods, highly practical with certain guarantees. They are however far from optimal. Mixed methods, combining many of this basic methods permit to improve their performance. DataMove has developed [27] a learning method permitting to adapt the Mixed method to benchmarks. An extensive experimental campaign has permitted to determine the possibilities of basic and mixed methods according to the benchmarks characteristics, enhancing the efficiency of mixed methods.

### 6.2.3. Reproducibility

Related to batch scheduling experimentation, DataMove has led investigations on reproducibility [23]. Existing approaches focus on repeatability, but this is only the first step to reproducibility: Continuing a scientific work from a previous experiment requires to be able to modify it. This ability is called reproducibility with Variation. We show that capturing the environment of execution is necessary but not sufficient ; we also need the environment of development. The variation also implies that those environments are subject to evolution, so the whole software development lifecycle needs to be considered. To take into account these evolutions, software environments need to be clearly defined, reconstructible with variation, and easy to share. In this context, we propose new way of seeing reproducibility through the scientific software development lifecycle. Each step in this lifecycle requires a software environment. We define a software environment by a set of applications and libraries, with all their dependencies, and their configurations, required to achieve a step in a scientific workflow.

### 6.2.4. Online Algorithms

Rob van Stee wrote a review of 2018 online algorithms including our recent contributions on resource augmentation [0] We quote him here:

*Progress was also made on scheduling to minimize weighted flow time on unrelated machines. In ESA 2016, Giorgio Lucarelli et al. [1] had considered a version where the online algorithm can reject some $\varepsilon_r > 0$ fraction (by weight) of the jobs and have machines that are $1 + \varepsilon_s$ as fast as the offline machines, for some $\varepsilon_s > 0$. They showed that this is already enough to achieve a competitive ratio of $O(1/(\varepsilon_s\varepsilon_r))$.*

*In SPAA 2018, Giorgio Lucarelli et al.[20] (a superset of the previous authors) showed that it is in fact sufficient to reject a $2\varepsilon$ fraction of the total number of jobs to achieve a competitive ratio of $2(\frac{1+\varepsilon}{\varepsilon})$ for minimizing the total flow time. This algorithm sometimes rejects a job other than the one that has just arrived. The authors show that this is necessary, as otherwise there is a lower bound of $\Omega(\Delta)$ even on a single machine. Here $\Delta$ is the size ratio (the ratio of largest to smallest job size). (Obviously this lower bound also holds if you cannot reject jobs at all.)*

*They also consider the speed scaling model, in which machines can be sped up if additional energy is invested, and the goal is to minimize the total weighted flow time plus energy usage. If the power function of machine $i$ is given by $P(s_i(t)) = s_i(t)^\alpha$, where $s_i(t)$ is the current speed of machine $i$, there is an algorithm which is $O((1 + 1/\varepsilon)^{\alpha/(\alpha-1)})$-competitive that rejects jobs of total weight at most a fraction $\varepsilon$ of the total weight of all*

[0] Rob van Stee. 2018. SIGACT News Online Algorithms Column 34: 2018 in review. SIGACT News 49, 4 (December 2018), 36-45.

*the jobs. They also give a positive result for jobs with hard deadlines, where the goal is to minimize the total energy usage and no job may be rejected.*

*In ESA 2018, the same set of authors [11] improved/generalized these results by showing that rejection alone is sufficient for an algorithm to be competitive even for weighted flow time. They presented an $O(1/\varepsilon^3)$-competitive algorithm that rejects at most $O(\varepsilon)$ of the total weight of the jobs. In this algorithm, jobs are assigned (approximately) greedily to machines, and each machine runs the jobs assigned to it using Highest Density First. A job may be rejected if it is running while much heavier jobs arrive or if it is in the queue while very many jobs arrive. The second rule simulates the resource augmentation on the speed.*

**DATASPHERE Team**

# 7. New Results

## 7.1. Political economy

We pursued our work on digital platforms and their impact on the structure of socio-economic systems, which results from the capacity to separate data or information from the actors of the physical world. In [9], we showed how the movement above ground of the intermediation activity transforms territories. A global analysis of the geopolitics of technology was presented in [3].

## 7.2. Anthropocene studies

We have investigated the possible similarities between biological systems and social systems facing shortage of resources, suggesting that the digital revolution might have something to do with the Anthropocene. More comprehensive approaches that rely on digital systems to control society and nudge citizens to adapt their behavior have been developed in Asia. We analyse in particular the social scoring system in China, and Society 5.0 in Japan [6]. An investigation of the world of images and photography in the time of algorithms was conducted in [2].

## 7.3. Laws and digital

The emergence of digital services affects the legal system. The law is always associated to a territory, while digital systems act remotely over large regions crossing borders to reach the population, imposing new norms. In [1], we suggest that a new framework is necessary to apprehend new phenomena, such as those resulting from the conflicts between global search engines and local rules with respect to the Right to be forgotten for instance.

## 7.4. Network data analytics

In collaboration with the Chinese Academy of Sciences, we worked on packet processing algorithmic for high speed network measurements. In [5] a packet capture archive system is developed and described. In [4] a theoretical analysis of the TCAM updates delay that is the main shortcoming of TCAM usage in high speed packet processors is presented. Quality of service for network functions were considered in [7].

**DRACULA Project-Team**

# 5. New Results

## 5.1. Oscillations and asymptotic convergence for a delay differential equation modeling platelet production

In [13], a model for platelet production is introduced for which the platelet count is described by a delay differential equation $P'(t) = -\gamma P(t) + f(P(t))g(P(t-r))$ where $f$ and $g$ are positive decreasing functions. First, the authors study the oscillation of the solutions around the unique equilibrium of the equation above, obtaining an inequality implying such an oscillation. They also obtain provide a condition such that this inequality is necessary and sufficient for oscillation. This result is compared to already existing results and the biological meaning of the inequality is studied. The authors also present a result on the asymptotic convergence of the solutions. This result depends on the behavior of the solution for $t \in [0, r]$, and the authors provide an analysis of the link between this behavior and the initial conditions in the case of a simpler model.

## 5.2. Meningioma growth dynamics assessed by radiocarbon retrospective birth dating

It is not known how long it takes from the initial neoplastic transformation of a cell to the detection of a tumor, which would be valuable for understanding tumor growth dynamics. We have assessed the age and growth dynamics in patients with WHO grade I meningiomas by combining retrospective birth-dating of cells by analyzing incorporation of nuclear-bomb-test-derived 14C, analysis of cell proliferation, cell density, MRI imaging and mathematical modeling. We provide an integrated model of the growth dynamics of benign meningiomas. The mean age of WHO grade I meningiomas was 22.1$\pm$ 6.5 years. We conclude that WHO grade I meningiomas are very slowly growing brain tumors, which are resected in average two decades after time of origination. [18]

## 5.3. Existence and stability of periodic solutions of an impulsive differential equation and application to CD8 T-cell differentiation

In this article [16], we study a scalar impulsive differential equation (IDE) with the aim of studying the effects of uneven molecular partitioning upon cell mitosis on CD8 T-cell differentiation. To do so, we introduce mathematical results that stand for a more general class of IDE, then apply them to our IDE and discuss those results with regard to the initial biological problem.

## 5.4. Investigating the role of the experimental protocol in phenylhydrazine-induced anemia on mice recovery

Erythropoiesis, the process of production of red blood cells, is performed through complex regulatory processes. We proposed an earlier model describing stress erythropoiesis in mice [33]. This model, based on the description of erythroid progenitor and erythrocyte dynamics using delay equations, led us to conclude on the quantitative importance of self-renewal. In [6], we refined this previous approaches by taking into account a more mechanistic description of the induction of anemia via phenylhydrazine injection. This led us to revisit some of our initial hypothesis regarding self-renewal regulation.

## 5.5. Generalizing a mathematical model of prion aggregation allows strain coexistence and co-stability by including a novel misfolded species

Prions are proteins capable of adopting misfolded conformations and transmitting these conformations to other normally folded proteins. A distinct feature of prion propagation is the existence of different phenotypical variants, called strains. In order to conform to biological observations of strain coexistence and co-stability, we develop in [19] an extension of the classical model by introducing a novel prion species consistent with biological studies.

## 5.6. Analysis and Numerical Simulation of a Polymerization Model with Possible Agglomeration Process

The purpose of [20] is to provide analytical and numerical results for a general polymerization model with lengthening process by agglomeration. 2D spatial diffusion of monomers is taken into account for the mass transfer between monomers and polymers. The analysis of the model is performed thanks to a double fixed point theorem. Adequate numerical scheme based on a generalization of the anti-dissipative method developed in Goudon (Math. Models Methods Appl. Sci. 23:1177–1215, 2013

## 5.7. The Origin of Species by Means of Mathematical Modelling

Darwin described biological species as groups of morphologically similar individuals. These groups of individuals can split into several subgroups due to natural selection, resulting in the emergence of new species. Some species can stay stable without the appearance of a new species, some others can disappear or evolve. In [10] we have developed a model which allows us to reproduce the principal patterns in Darwin's diagram. Some more complex evolutionary patterns are also observed. The relation between Darwin's definition of species, stated above, and Mayr's definition of species (group of individuals that can reproduce) is also discussed.

## 5.8. Improved duality estimates in the time discrete case for cross diffusion models

In [28], time discrete versions of the duality estimates derived by Canizo et al. for parabolic systems have been obtained. They allow the construction of solution with superquadratic reactions terms for cross diffusion models with bounded pressure.

**ELAN Team**

# 6. New Results

## 6.1. Inverse design of a suspended elastic rod

**Participants:** Florence Bertails-Descoubes, Victor Romero.

In collaboration with Alexandre Derouet-Jourdan (OLM Digital, Japan) and Arnaud Lazarus (UPMC, Laboratoire Jean le Rond d'Alembert), we have investigated the inverse design problem of a suspended elastic subject to gravity. We have proved that given an arbitrary space curve, there exists a unique solution for the natural configuration of the rod, which is independent of the initial framing of the input curve. Moreover, this natural configuration can be easily computed by solving three linear ODEs in sequence, starting from any input framing. This work has been published in Roy. Soc. Proc A [1] and physical aspects of this study have been communicated about in a mechanical congress [4].

## 6.2. Simulation of cloth contact with exact Coulomb friction

**Participants:** Florence Bertails-Descoubes, Laurence Boissieux.

In collaboration with Gilles Daviet (Weta Digital, New Zealand) and Rahul Narain's group (University of Minnesota and IIT Delhi), we have developed a new implicit solver for taking into account contact in cloth with Coulomb friction. Our key idea stems from the observation that for a nodal system like cloth, and in the case where each node is subject to at most one contacting constraint (either an external or self-contact), the frictional contact problem may be formulated based on velocities as primary variables, without having to compute the costly Delassus operator; then, by reversing the roles classically played by the velocities and the contact impulses, conical complementarity solvers of the literature may be leveraged to solve for compatible velocities at nodes. To handle the full complexity of cloth dynamics scenarios, we have extended this base algorithm in two ways: first, towards the accurate treatment of frictional contact at any location of the cloth, through an adaptive node refinement strategy; second, towards the handling of multiple constraints at each node, through the duplication of constrained nodes and the adding of pin constraints between duplicata. Our method proves to be both fast and robust, allowing us to simulate full-size garments with an unprecedent level of realism compared to former methods, while maintaining similar computational timings. Our work has been published at ACM Transactions on Graphics (ACM SIGGRAPH 2018) [2].

## 6.3. Inverse design of thin elastic shells

**Participants:** Mickaël Ly, Florence Bertails-Descoubes, Laurence Boissieux.

In collaboration with Romain Casati (former PhD student of F. Bertails-Descoubes) and Mélina Skouras (EPI IMAGINE), we have proposed an inverse strategy for modeling thin elastic shells physically, just from the observation of their geometry. Our algorithm takes as input an arbitrary target mesh, and interprets this configuration automatically as a stable equilibrium of a shell simulator under gravity and frictional contact constraints with a given external object. Unknowns are the natural shape of the shell (i.e., its shape without external forces) and the frictional contact forces at play, while the material properties (mass density, stiffness, friction coefficients) can be freely chosen by the user. Such an inverse problem formulates as an ill-posed nonlinear system subject to conical constraints. To select and compute a plausible solution, our inverse solver proceeds in two steps. In a first step, contacts are reduced to frictionless bilateral constraints and a natural shape is retrieved using the adjoint method. The second step uses this result as an initial guess and adjusts each bilateral force so that it projects onto the admissible Coulomb friction cone, while preserving global equilibrium. To better guide minimization towards the target, these two steps are applied iteratively using a degressive regularization of the shell energy. We validate our approach on simulated examples with reference material parameters, and show that our method still converges well for material parameters lying within a

reasonable range around the reference, and even in the case of arbitrary meshes that are not issued from a simulation. We finally demonstrate practical inversion results on complex shell geometries freely modeled by an artist or automatically captured from real objects, such as posed garments or soft accessories. Our work has been published at ACM Transactions on Graphics (ACM SIGGRAPH Asia 2018) [3] and has been selected for a Press Release of the ACM.

**ERABLE Project-Team**

# 6. New Results

## 6.1. General comments

We present in this section the main results obtained in 2018.

We tried to organise these along the four axes as presented above. Clearly, in some cases, a result obtained overlaps more than one axis. In such case, we chose the one that could be seen as the main one concerned by such results.

We did not indicate here the results on more theoretical aspects of computer science if it did not seem for now that they could be relevant in contexts related to computational biology. Actually, those on string [32], [33], [36], [11] and graph algorithms in general [2], [35], [38], [37], [39], [41], [54], [42], [44], [43], [40], [47], [5], [45], [48], [49], [53], [23], [24], or on more general algorithmic problems notably related to data structures are already relevant for life sciences (biology or ecology) or in the future could become more specifically so. We do in particular believe that dynamic graph approaches could be of great interest in the future for some of the enumeration problems we constantly meet in biology.

A few other results of 2018 are not mentioned in this report, not because the corresponding work is not important, but because it was likewise more specialised [52], or the work represented a survey *e.g.* [21]). Likewise, also for space reasons, we do not detail the results presented in some biological papers of the team when these did not require a mathematical or algorithmical input or are surveys [1], [4], [7], [8], [12], [19], [20], [22], [25], [31].

On the other hand, we do mention a couple of works that were submitted towards the end of 2018.

## 6.2. Axis 1: Genomics

**Genome hybrid assembly**
Long read sequencing technologies are considered to be the solution for handling genome repeats, allowing near reference-level reconstructions of large genomes. However, long read de novo assembly pipelines are computationally intense and require a considerable amount of coverage, thereby hindering their broad application to the assembly of large genomes. Alternatively, hybrid assembly methods that combine short and long read sequencing technologies can reduce the time and cost required to produce de novo assemblies of large genomes. In [10], we proposed a new method, called FAST-SG, that uses a new ultrafast alignment-free algorithm specifically designed for constructing a scaffolding graph using lightweight data structures. FAST-SG can construct the graph from either short or long reads. This allows the reuse of efficient algorithms designed for short read data and permits the definition of novel modular hybrid assembly pipelines. Using comprehensive standard datasets and benchmarks, we showed how FAST-SG outperforms the state-of-the-art short read aligners when building the scaffolding graph and can be used to extract linking information from either raw or error-corrected long reads. We also showed how a hybrid assembly approach using FAST-SG with shallow long-read coverage (5X) and moderate computational resources can produce long-range and accurate reconstructions of the genomes of *Arabidopsis thaliana* (Ler-0) and human (NA12878). We are currently working on the assembly process itself, using the scaffolding graphs obtained with FAST-SG. The results obtained so far are extremely promising and a paper is currently in preparation. This is part of the work done by Alex di Genova, postdoc in ERABLE.

**Variant annotation**

Genome-wide analyses estimate that more than 90% of multi exonic human genes produce at least two transcripts through a genomic variant called alternative splicing (AS). Various bioinformatics methods are available to analyse AS from RNAseq data. Most methods start by mapping the reads to an annotated reference genome, but some start by a *de novo* assembly of the reads. In [3], we presented a systematic comparison of a mapping-first approach (FARLINE) and an assembly-first approach (scKisSplice). We applied these methods to two independent RNAseq datasets and found that the predictions of the two pipelines overlapped (70% of exon skipping events were common), but with noticeable differences. The assembly-first approach allowed to find more novel variants, including novel unannotated exons and splice sites. It also predicted AS in recently duplicated genes. The mapping-first approach allowed to find more lowly expressed splicing variants, and splice variants overlapping repeats. This work demonstrated that annotating AS with a single approach leads to missing out a large number of candidates, many of which are differentially regulated across conditions and can be validated experimentally. We therefore advocate for the combined use of both mapping-first and assembly-first approaches for the annotation and differential analysis of AS from RNAseq datasets. This was part of the work of Clara Benoît-Pilven, postdoc at Inserm and in ERABLE, to which also participated other current or ex-members of ERABLE, namely Camille Marchet (during her stay as ADT engineer with ERABLE), Emilie Chautard (when she was postdoc Inserm and in ERABLE), Gustavo Sacomoto (when he was PhD and then for one year postdoc in ERABLE), and Leandro Lima (current PhD student of ERABLE).

Another type of variant, namely SNPs was also considered in [51]. In this paper, mutations are detected by eBWT (extended Burrows-Wheeler Transform). Indeed, we notices that eBWT of a collection of DNA fragments tend to cluster together the copies of nucleotides sequenced from a genome. We showed that it is thus possible to accurately predict how many copies of any nucleotide are expected inside each such cluster, and that a precise LCP array based procedure can locate these clusters in the eBWT. These theoretical insights were validated in practice with SNPs being clustered in the eBWT of a reads collection. We developed a tool for finding SNPs with a simple scan of the eBWT and LCP arrays. Preliminary results show that our method requires much less coverage than the state-of-the-art tools while drastically improving precision and sensitivity.

Both types of variants correspond to special types of $st$-paths in graphs, a topic that was also explored from a more purely theoretical point of view in two papers, one already accepted [46] and on that is about to be submitted and extends the results obtained in 2017 on bubble (as $st$-paths are also called in bioinformatics) generators in directed graphs.

**Full-length *de novo* viral quasispecies assembly through variation graph construction**
Viruses populate their hosts as a viral quasispecies: a collection of genetically related mutant strains. Viral quasispecies assembly refers to reconstructing the strain-specific haplotypes from read data, and predicting their relative abundances within the mix of strains, an important step for various treatment-related reasons. Reference-genome-independent ("de novo") approaches have yielded benefits over reference-guided approaches, because reference-induced biases can become overwhelming when dealing with divergent strains. While being very accurate, extant *de novo* methods only yield rather short contigs. It remains to reconstruct full-length haplotypes together with their abundances from such contigs. In [34], we first constructed a variation graph, a recently popular, suitable structure for arranging and integrating several related genomes, from the short input contigs, without making use of a reference genome. To obtain paths through the variation graph that reflect the original haplotypes, we solved a minimisation problem that yields a selection of maximal-length paths that is optimal in terms of being compatible with the read coverages computed for the nodes of the variation graph. We output the resulting selection of maximal length paths as the haplotypes, together with their abundances. Benchmarking experiments on challenging simulated data sets showed significant improvements in assembly contiguity compared to the input contigs, while preserving low error rates. As a consequence, our method outperforms all state-of-the-art viral quasispecies assemblers that aim at the construction of full-length haplotypes, in terms of various relevant assembly measures. The tool, called VIRUS-VG, is availabe at https://bitbucket.org/jbaaijens/virus-vg.

A member of ERABLE was also involved in the Second Annual Meeting of the European Virus Bioinformatics Center (EVBC), held in Utrecht, Netherlands, and whose focus was on computational approaches in virology, with topics including (but not limited to) virus discovery, diagnostics, (meta-)genomics, modeling, epidemiology, molecular structure, evolution, and viral ecology. Approximately 120 researchers from around the world attended the meeting this year. An overview of new developments and novel research findings that emerged during the meeting was published in the journal *Viruses* [16].

**Bacterial genome-wide association studies (GWAS)**
Genome-wide association study (GWAS) methods applied to bacterial genomes have shown promising results for genetic marker discovery or detailed assessment of marker effect. Recently, alignment-free methods based on $k$-mer composition have proven their ability to explore the accessory genome. However, they lead to redundant descriptions and results which are sometimes hard to interpret. In [17], we introduced DBGWAS, an extended $k$-mer-based GWAS method producing interpretable genetic variants associated with distinct phenotypes. Relying on compacted de Bruijn graphs (cDBG), our method gathers cDBG nodes, identified by the association model, into subgraphs defined from their neighbourhood in the initial cDBG. DBGWAS is alignment-free and only requires a set of contigs and phenotypes. In particular, it does not require prior annotation or reference genomes. It produces subgraphs representing phenotype-associated genetic variants such as local polymorphisms and mobile genetic elements (MGE). It offers a graphical framework which helps interpret GWAS results. Importantly, it is also computationally efficient (the experiments took one hour and a half on average). We validated our method using antibiotic resistance phenotypes for three bacterial species. DBGWAS recovered known resistance determinants such as mutations in core genes in *Mycobacterium tuberculosis*, and genes acquired by horizontal transfer in *Staphylococcus aureus* and *Pseudomonas aeruginosa* along with their MGE context. It also enabled us to formulate new hypotheses involving genetic variants not yet described in the antibiotic resistance literature. This is part of the work of Magali Jaillard, PhD student of Laurent Jacob who is an external collaborator of ERABLE, and of Leandro I. S. de Lima, PhD student co-supervised by three members of ERABLE.

## 6.3. Axis 2: Metabolism and post-transcriptional regulation

**Multi-objective metabolic mixed integer optimisation: with an application to yeast strain engineering**
In a paper submitted and already available in bioRxiv (https://www.biorxiv.org/content/early/2018/11/22/476689), we explored the concept of multi-objective optimisation in the field of metabolic engineering when both continuous and integer decision variables are involved in the model. In particular, we proposed a multi-objective model which may be used to suggest reaction deletions that maximise and/or minimise several functions simultaneously. The applications may include, among others, the concurrent maximisation of a bioproduct and of biomass, or maximisation of a bioproduct while minimising the formation of a given by-product, two common requirements in microbial metabolic engineering. Production of ethanol by the widely used cell factory *Saccharomyces cerevisiae* was adopted as a case study to demonstrate the usefulness of the proposed approach in identifying genetic manipulations that improve productivity and yield of this economically highly relevant bioproduct. We did an *in vivo* validation and we could show that some of the predicted deletions exhibit increased ethanol levels in comparison with the wild-type strain. The multi-objective programming framework we developed, called MOMO, is open-source and uses POLYSCIP as underlying multi-objective solver. This is part of the work of Ricardo de Andrade, postdoc at Unversity of São Paulo with Roberto Marcondes, and in ERABLE. It is joint work with Susana Vinga, external collaborator of ERABLE and partner of the Inria Associated Team Compasso.

**Metabolic shifts**

With the increasing availability of so-called 'omics data – transcriptomics, proteomics, and metabolics – there has been growing interest in various ways of integrating them with the metabolic network. When the network is represented by a graph, 'omics data can guide the extraction of subnetworks of interest to find metabolic pathways or sets of related genes. Within the framework of constraint-based modelling, 'omics data can be used to improve the prediction of metabolic behaviour and to build context-specific metabolic models. One interesting application of metabolic reconstructions in conjunction with 'omics data is to use the two to understand metabolic shifts. When an organism encounters a change in environmental conditions, often a re-organisation of metabolism follows. Comparative measurements of gene expression and metabolite concentrations can be used to gain insight into these changes but this data is "structureless", meaning it lacks the information about how the metabolic components relate to each other. A metabolic network on the other hand contains this information, and can thus greatly benefit such an analysis. We developed a new method, called MOOMIN, that combines the results of a differential expression analysis comparing the gene expression levels in two different conditions with a metabolic network to produce a hypothesis of a metabolic shift. The idea is to use the network structure to define feasible global changes in metabolism. These changes are then scored based on the gene expression data with the goal of finding the change that best agrees with the observations. Finding the best-scoring change is formulated into an optimisation problem that can be solved using Mixed-Integer Linear Programming. This is part of the work of Henri Taneli Pusa, co-supervised by 3 members of ERABLE, whose manuscript was submitted to the reviewers and who should be defending his PhD in early February 2019. The paper on MOOMIN will be submitted soon, and the software then made available. Participated also in this work Mariana G. Ferrarini, postdoc at Insa and in ERABLE, and Ricardo Andrade, postdoc at University of São Paulo with Roberto Marcondes and in ERABLE.

**Metabolic games**
The PhD of Taneli also investigated game theory in the context of metabolism. Game theory is a branch of applied mathematics that deals with interacting rational agents with conflicting goals. When rationality is replaced with natural selection, *evolutionary* game theory can be used to explain the "decisions" taken by even microscopic organisms. The PhD manuscript presents the idea of a *metabolic game*, a game theoretical model for the prediction of metabolic behaviour. In contrast to Flux Balance Analysis, where the metabolic state is predicted using simple optimisation, a metabolic game takes into account the fact that optimality is influenced by the surrounding members of a microbial community. By changing the availability of nutrients, or secreting beneficial or harmful molecules, microbes essentially create their own environment and make optimal behaviour context-specific. A paper is submitted that reviews the literature that has applied game theory to the study of microbes, with a focus on metabolism and especially games derived using metabolic networks and constraint-based modelling. In the PhD manuscript, Taneli further explains the idea behind a metabolic game and discusses different aspects of defining such games: the choice of players, actions, and payoffs.

## 6.4. Axis 3: (Co)Evolution

**Exploring the robustness of the parsimonious reconciliation method in host-symbiont cophylogeny**
Following our previous work on reconciliation methods for cophylogeny, in [29], we explored the robustness of the parsimonious host-symbiont tree reconciliation method under editing or small perturbations of the input. The editing involved making different choices of unique symbiont mapping to a host in the case where multiple associations exist. This is made necessary by the fact that the tree reconciliation model is currently unable to handle such associations. The analysis performed could however also address the problem of errors. The perturbations were re-rootings of the symbiont tree to deal with a possibly wrong placement of the root specially in the case of fast-evolving species. In order to do this robustness analysis, we introduced a simulation scheme specifically designed for the host-symbiont cophylogeny context, as well as a measure to compare sets of tree reconciliations, both of which are of interest by themselves. This work was also part of the PhD of a previous student of ERABLE, Laura Urbini.

**Geometric medians in reconciliation spaces**

Recently, there has been much interest in studying spaces of tree reconciliations (as used in cophylogenetic studies), which arise by defining some metric $d$ on the set $\mathcal{R}(P, H, \phi)$ of all possible reconciliations between two trees $P$ and $H$ where $\phi$ represents the map between the leaf-sets of $P$ and $H$ (corresponding to present-day associations). In [14], we studied the following question: how do we compute a *geometric median* for a given subset $\Psi$ of $\mathcal{R}(P, H, \phi)$ relative to $d$, *i.e.* an element $\psi_{med} \in \mathcal{R}(P, H, \phi)$ such that

$$\sum_{\psi' \in \Psi} d(\psi_{med}, \psi') \leq \sum_{\psi' \in \Psi} d(\psi, \psi')$$

holds for all $\psi \in \mathcal{R}(P, H, \phi)$? For a model where so-called host-switches or transfers are not allowed, and for a commonly used metric $d$ called the *edit-distance*, we showed that although the cardinality of $\mathcal{R}(P, H, \phi)$ can be super-exponential, it is still possible to compute a geometric median for a set $\Psi$ in $\mathcal{R}(P, H, \phi)$ in polynomial time. We expect that this result could be useful for computing a summary or consensus for a set of reconciliations (*e.g.* for a set of suboptimal reconciliations). The collaboration with Katharina Huber and Vincent Moulton from the School of Computing Sciences at the University of New Anglia was made possible by a Royal Society Grant obtained by the two partners (UNA and ERABLE).

**Exploring and Visualising Spaces of Tree Reconciliations**
A common approach to tree reconciliation involves specifying a model that assigns costs to certain events, such as cospeciation, and then tries to find a mapping between two specified phylogenetic trees which minimises the total cost of the implied events. For such models, it has been shown, including by the ERABLE members in previous papers, that there may be a huge number of optimal solutions, or at least solutions that are close to optimal. It is therefore of interest to be able to systematically compare and visualise whole collections of reconciliations between a specified pair of trees. In [13] , we considered various metrics on the set of all possible reconciliations between a pair of trees, some that have been defined before but also new metrics that we proposed. We showed that the diameter for the resulting spaces of reconciliations can in some cases be determined theoretically, information that we used to normalise and compare properties of the metrics. We also implemented the metrics and compared their behaviour on several host parasite datasets, including the shapes of their distributions. In addition, we showed that in combination with multidimensional scaling, the metrics can be useful for visualising large collections of reconciliations, much in the same way as phylogenetic tree metrics can be used to explore collections of phylogenetic trees. Implementations of the metrics can be downloaded from https://team.inria.fr/erable/en/team-members/blerina-sinaimeri/reconciliation-distances/. This work was also funded by a Royal Society Grant obtained by the two partners (at University of New Anglia and ERABLE).

**Variants of phylogenetic network problems**
Although not falling within the general topic of coevolution, phylogenetic networks are of great interest as another way of representing the evolution of a set of species. In the context of such representations, unrooted and root-uncertain variants of several well-known phylogenetic network problems were explored. The hybridisation number problem requires to embed a set of binary rooted phylogenetic trees into a binary rooted phylogenetic network such that the number of nodes with indegree two is minimised. However, from a biological point of view accurately inferring the root location in a phylogenetic tree is notoriously difficult and poor root placement can artificially inflate the hybridisation number. To this end, we studied in [30] a number of relaxed variants of this problem. We started by showing that the fundamental problem of determining whether an unrooted phylogenetic network displays (*i.e.* embeds) an unrooted phylogenetic tree, is NP-hard. On the positive side, we show that this problem is FPT in the reticulation number. In the rooted case, the corresponding FPT result is trivial, but here we required more subtle argumentation. Next we showed that the hybridisation number problem for unrooted networks (when given two unrooted trees) is equivalent to the problem of computing the tree bisection and reconnect distance of the two unrooted trees. In the third part of the paper, we considered the "root uncertain" variant of hybridisation number. Here we were free to choose the root location in each of a set of unrooted input trees such that the hybridisation number of the resulting rooted trees is minimised. On the negative side, we showed that this problem is APX-hard. On the positive side, we showed that the problem is FPT in the hybridisation number, via kernelisation, for any number of input trees.

## 6.5. Axis 4: Human, animal and plant health

**Hydrogen peroxide production and myo-inositol metabolism as important traits for virulence of Mycoplasma hyopneumoniae**

*Mycoplasma hyopneumoniae* is the causative agent of enzootic pneumonia. In a previous work, we had reconstructed the metabolic models of this species along with two other mycoplasmas from the respiratory tract of swine: *Mycoplasma hyorhinis*, considered less pathogenic but which nonetheless causes disease and *Mycoplasma flocculare*, a commensal bacterium. We had identified metabolic differences that partially explained their different levels of pathogenicity. One important trait was the production of hydrogen peroxide from the glycerol metabolism only in the pathogenic species. Another important feature was a pathway for the metabolism of myo-inositol in *M. hyopneumoniae*. In the paper accepted this year [9], we tested these traits to understand their relation to the different levels of pathogenicity, comparing not only the species but also pathogenic and attenuated strains of *M. hyopneumoniae*. Regarding the myo-inositol metabolism, we showed that only *M. hyopneumoniae* assimilated this carbohydrate and remained viable when myo-inositol was the primary energy source. Strikingly, only the two pathogenic strains of *M. hyopneumoniae* produced hydrogen peroxide in complex medium. We also showed that this production was dependent on the presence of glycerol. Although further functional tests are needed, this work enabled to identify two interesting metabolic traits of *M. hyopneumoniae* that might be directly related to its enhanced virulence. This is part of the work of Mariana G. Ferrarini, currently postdoc at Insa and in ERABLE, and of Scheila G. Mucha whose PhD (defended in Sept. 2018) was co-supervised by Arnaldo Zaha and by a member of ERABLE.

**Cancer**

A member of ERABLE continues deeply involved with the Centre Léon Bérard in Lyon, and in that context, a number of works are running, all related to cancer genomics. In the first [28], an integrated genomic study was performed of 25 tumour tissues from radical prostatectomy of aggressive (defined by the International Society of Urological Pathology) prostate cancer patients (10 African Caribbean and 15 French Caucasian) using single nucleotide polymorphism arrays, whole-genome sequencing, and RNA sequencing. The results showed that African Caribbean tumours are characterised by a more frequent deletion at 1q41-43 encompassing the DNA repair gene PARP1, and a higher proportion of intra-chromosomal rearrangements including duplications associated with CDK12 truncating mutations. Transcriptome analyses showed an over-expression of genes related to androgen receptor activity in African Caribbean tumours, and of PVT1, a long non-coding RNA located at 8q24 that confirms the strong involvement of this region in prostate tumours from men of African ancestry. In a second study [15], gene-expression profiling data was used to build and validate a predictive model of outcome for patients with follicular lymphoma. A robust 23-gene expression-based predictor of progression-free survival that is applicable to routinely available formalin-fixed, paraffin-embedded tumour biopsies from such patients was thus developed and validated. Applying this score could allow individualised therapy for patients according to their risk category. In a third study, an integrated analysis highlighted APC11 protein expression as a likely new independent predictive marker for colorectal cancer [6].

In a parallel work by another member of ERABLE [27], it was proposed that cancer Is not (only) a senescence problem. Age is indeed one of the strongest predictors of cancer and risk of death from cancer. Cancer is therefore generally viewed as a senescence-related malady. However, cancer also exists at subclinical levels in humans and other animals, but its earlier effects on the body are poorly known by comparison. What was argued in [27] is that cancer is a significant but ignored burden on the body and is likely to be a strong selective force from early during the lifetime of an organism. It was thus proposed that time has come to adopt this novel view of malignant pathologies to improve our understanding of the ways in which oncogenic phenomena influence the ecology and evolution of animals long before their negative impacts become evident and fatal.

**Xylella fastidiosa epidemiological model**

*Xylella fastidiosa* is a notorious plant pathogenic bacterium that represents a threat to crops worldwide. Its subspecies, *Xylella fastidiosa* subsp. *fastidiosa* is the causal agent of Pierce's disease of grapevines. Pierce's disease has presented a serious challenge for the grapevine industry in the United States and turned into an epidemic in Southern California due to the invasion of the insect vector *Homalodisca vitripennis*. In an attempt to minimize the effects of *Xylella fastidiosa* subsp. *fastidiosa* in vineyards, various studies have been developing and testing strategies to prevent the occurrence of Pierce's disease, *i.e.*, prophylactic strategies. Research has also been undertaken to investigate therapeutic strategies to cure vines infected by *Xylella fastidiosa* subsp. *fastidiosa*. In [18], we explicitly review all the strategies published to date and specifies their current status. Furthermore, an epidemiological model of *Xylella fastidiosa* subsp. *fastidiosa* is proposed and key parameters for the spread of Pierce's disease deciphered in a sensitivity analysis of all model parameters. Based on these results, it is concluded that future studies should prioritise therapeutic strategies, while investments should only be made in prophylactic strategies that have demonstrated promising results in vineyards. This is part of the PhD of Henri Taneli Pusa in the context of the H2020 ITN MicroWine, together with another PhD student of the ITN, Ifigeneia Kyrkou. Ifigeneia was the first author of the paper [18] but the mathematical model is the work of Taneli.

**IBIS Project-Team**

# 6. New Results

## 6.1. Analysis of fluorescent reporter gene data

The use of fluorescent and luminescent reporter genes allows real-time monitoring of gene expression, both at the level of individual cells and cell populations (Section 3.2 ). Over the years, many useful resources have appeared, such as libraries of reporter strains for model organisms and computer tools for designing reporter plasmids. Moreover, the widespread adoption of thermostated microplate readers in experimental laboratories has made it possible to automate and multiplex reporter gene assays on the population level. This has resulted in large time-series data sets, typically comprising $10^5 - 10^6$ measurements of absorbance, fluorescence, and luminescence for $10^3$ wells on the microplate. In order to fully exploit these data sets, we need sound mathematical methods to infer biologically relevant quantities from the primary data and computer tools to apply the methods in an efficient and user-friendly manner.

In the past few years we developed novel methods for the analysis of reporter gene data obtained in microplate experiments, based on the use of regularized linear inversion. This allows a range of estimation problems to be solved, notably the inference of growth rate, promoter activity, and protein concentration profiles. The linear inversion methods, published in *Bioinformatics* in 2015 [12], have been implemented in the Python package WELLFARE and integrated in the web application WELLINVERTER. Funded by a grant from the Institut Français de Bioinformatique (IFB), we improved WellInverter by developing a parallel computational architecture with a load balancer to distribute the analysis queries over several back-end servers, a new graphical user interface, and a plug-in system for defining high-level routines for parsing data files produced by microplate readers from different manufacturers. This has resulted in a scalable and user-friendly web service providing a guaranteed quality of service, in terms of availability and response time. This year the web service has been redeployed on the new IFB cloud and on an Inria server, accompanied by extensive user documentation, online help, and a tutorial. Moreover, we submitted a journal paper on WELLINVERTER illustrating the use of the tool by analyzing data of the expression of a fluorescent reporter gene controlled by a phage promoter in growing *Escherichia coli* populations. We notably show that the expression pattern in different growth media, supporting different growth rates, corresponds to the pattern expected for a constitutive gene.

Compared to most reporter gene assays based on fluorescence proteins, luciferase reporters have a superior signal-to-noise ratio, since they do not suffer from the high autofluorescence background of the bacterial cell. At the same time, however, luciferase reporters have the drawback of constant light emission, which leads to undesired cross-talk between neighbouring wells on a microplate. To overcome this limitation, Marco Mauri in collaboration with colleagues from the Philipps-Universität Marburg developed a computational method to correct for luminescence bleed-through and to estimate the "true" luminescence activity for each well of a microplate. As the sole input our algorithm uses the signals measured from a calibration plate, in which the light emitted from a single luminescent well serves as an estimate for the "light-spread function". We show that this light-spread function can be used to deconvolve any other measurement obtained under the same technical conditions. Our analysis demonstrates that the correction preserves low-level signals close to the background and shows that it is universally applicable to different kinds of microplate readers and plate types. A journal article on this work was submitted this year.

## 6.2. Microdomain formation of bacterial membrane proteins

Fluorescent reporters can be used not only to quantify gene expression, but also to localize proteins in different compartments of the cell, In particular, in bacteria proteins within the cytoplasmic membrane display distinct localization patterns and arrangements. While multiple models exist describing the dynamics of membrane proteins, to date there have been few systematic studies, particularly in bacteria, to evaluate how protein size, number of transmembrane domains, and temperature affect their diffusion, and if conserved localization patterns exist.

Marco Mauri in collaboration with colleagues from the Philipps-Universität in Marburg has used fluorescence microscopy, single-molecule tracking (SMT), and computer-aided visualization methods to obtain a better understanding of the three-dimensional organization of bacterial membrane proteins, using the model bacterium *Bacillus subtilis*. First, we carried out a systematic study of the localization of over 200 *B. subtilis* membrane proteins, tagged with monomeric mVenus-YFP at their original gene locus. Their subcellular localization could be discriminated in polar, septal, patchy, and punctate patterns. Almost 20% of membrane proteins specifically localized to the cell poles, and a vast majority of all proteins localized in distinct structures, which we term microdomains. Dynamics were analyzed for selected membrane proteins, using SMT. Diffusion coefficients of the analyzed transmembrane proteins did not correlate with protein molecular weight, but correlated inversely with the number of transmembrane helices, *i.e.*, transmembrane radius. We observed that temperature can strongly influence diffusion on the membrane, in that upon growth temperature upshift, diffusion coefficients of membrane proteins increased and still correlated inversely to the number of transmembrane domains, following the Saffman–Delbrück relation.

The vast majority of membrane proteins were observed to localize to distinct multimeric assemblies. Diffusion of membrane proteins can be suitably described by discriminating diffusion coefficients into two protein populations, one mobile and one immobile, the latter likely constituting microdomains. Moreover, in this study published in *BMC Biology* [18], we provided a method to correct the diffusion coefficient for the membrane curvature. Our results show there is high heterogeneity and yet structural order in the cell membrane, and provide a roadmap for our understanding of membrane organization in prokaryotes. Given the exceptionally richness of the data obtained, both further analysis on membrane lateral movements and a more detailed theory on the effect of membrane curvature is possible.

## 6.3. Stochastic modeling and identification of gene regulatory networks in bacteria

At the single-cell level, the processes that govern single-cell dynamics in general and gene expression in particular are better described by stochastic models rather than the deterministic models underlying the linear inversion methods discussed in Section 6.6 . Modern techniques for the real-time monitoring of gene expression in single cells enable one to apply stochastic modelling to study the origins and consequences of random noise in response to various environmental stresses, and the emergence of phenotypic variability. The potential impact of single-cell stochastic analysis and modelling ranges from a better comprehension of the biochemical regulatory mechanisms underlying cellular phenotypes to the development of new strategies for the (computer assisted or genetically engineered) control of cell populations and even of single cells.

Work in IBIS on gene expression and interaction dynamics at the level of individual cells is addressed in terms of identification of parametric intrinsic noise models, on the one hand, and the nonparametric inference of gene expression statistics, on the other hand, from population snapshot data. Along with modelling and inference, identifiability analysis is dedicated special attention. Other problems related with single-cell modelling, extracellular variability and inheritance of traits at cell division are considered also through external collaborations, as discussed below (Section 6.4 ).

Concerning identification of intrinsic noise dynamics in single cells, previous results on the contribution of stochasticity to parameter identifiability and on reconstruction of unknown gene regulatory networks have been taken further. For the case of population snapshot meaurements, where the dynamics of the population statistics are observed by simple time-lapse experiments, our earlier results showing that variance measurements may provide tremendous improvement in network reconstruction relative to sole mean measurements have been developed into a full-blown method for first-order gene network reconstruction. Additionally, parameter identifiability methods and results initially developed for gene expression models have been generalized to the whole class of first-order stochastic reaction networks. These developments have been presented and demonstrated by simulation in a paper published in the journal *Processes* [15].

Reconstruction of promoter activity statistics from reporter gene population snapshot data has been further investigated, leading to a full-blown spectral analysis and reconstruction method for reporter gene systems. In

the context of the ANR project MEMIP (Section 7.2 ), we have characterized reporter systems as noisy linear systems operating on a stochastic input (promoter activity), and developed an inversion method for estimation of promoter activation statistics from reporter population snapshots that is nonparametric, in the sense that it does not assume any parametric model for the unknown promoter dynamics. This analysis rests on a more general, original generalization of moment equations that we have developed for stochastic reaction networks with state-affine rates subject to random input processes. The theoretical results, together with a demonstration of the reporter gene inversion method on simulated data, have been accepted for publication in *Automatica* this year [16]. In addition to utilization of the method on real gene-expression data, the results lend themselves to several additional applications, among which the study of extrinsic noise and the optimal design of reporter systems.

## 6.4. Modelling and analysis of cellular trait dynamics over lineage trees

The investigation of cellular populations at a single-cell level has already led to the discovery of important phenomena, such as the occurrence of different phenotypes in an isogenic population. Nowadays, several experimental techniques, such as microscopy combined with the use of microfluidic devices, enable one to take investigation further by providing time-profiles of the dynamics of individual cells over entire lineage trees. The difficulty, and at the same time the opportunity, in exploiting these data and inferring mathematical models from them is the fact that the behavior of different cells is correlated because of inheritance.

From the modelling point of view, lineage trees are well described by structured branching population models where the life cycle of each cell depends on individual characteristics, such as size and internal protein dynamics, which play a key role in the mechanisms of cell division. One important aspect in the analysis of these population models consists in the investigation of biases arising from the sampling of a finite set of observed individuals. In order to characterize bias, we studied the dynamics of a structured branching population where the trait of each individual evolves in accordance with a Markov process. We assumed that the rate of division of each individual is a function of its trait and when a branching event occurs, the trait of the descendants at birth depends on their number and on the trait of the mother. We explicitly described the Markov process, named auxiliary process, corresponding to the dynamics of the trait of a "typical" individual by deriving its associated infinitesimal generator. In particular, we proved that this process characterizes exactly the process of the trait of a uniformly sampled individual in a large population approximation. This work, carried out by Aline Marguet, has been accepted for publication in the journal *Bernoulli* [19].

We also investigated the long-time behavior of the population and proved that a typical individual in the population asymptotically behaves like the auxiliary process previously introduced. These results have been submitted for publication [22]. Structured branching processes and their analysis also provide the basis for identification tools for lineage-tree data. In particular, in the context of a bifurcating Markov chain, where each individual is characterized by a trait evolving in accordance with a scalar diffusion, we proved that the maximum-likelihood estimator of the division rate is asymptotically efficient and demonstrate the method on simulated data. This work, in collaboration with M. Hoffmann at Univ Paris-Dauphine, has also been submitted for publication [21].

Along the same lines, modelling and identification of gene expression models with mother-daughter inheritance are being investigated in the context of the ANR project MEMIP. Starting from an earlier work of the group [7], with reference to an application on osmotic shock response by yeast, the key question is to what extent leveraging an inheritance model improves inference of individual cell dynamics as well as of inheritance dynamics themselves, relative to state-of-art approaches where inheritance is not accounted for at a modelling stage.

## 6.5. Models of carbon metabolism in bacteria

Adaptation of bacterial growth to changes in environmental conditions, such as the availability of specific carbon sources, is triggered at the molecular level by the reorganization of metabolism and gene expression: the concentration of metabolites is adjusted, as well as the concentration and activities of enzymes, the rate

of metabolic reactions, the transcription and translation rates, and the stability of proteins and RNAs. This reprogramming of the bacterial cell is carried out by i) specific interactions involving regulatory proteins or RNAs that specifically respond to the change of environmental conditions and ii) global regulation involving changes in the concentration of RNA polymerase, ribosomes, and metabolite pools that globally affect the rates of transcription, translation, and degradation of all RNAs and proteins.

A quantitative description and understanding of this complex network, cutting across metabolism, gene expression, and signalling, can be accessed through mathematical modelling only. In collaboration with Andreas Kremling, professor at TU München and former visiting scientist in the IBIS project-team, Hans Geiselmann, Delphine Ropers and Hidde de Jong developed an ensemble of variants of a simple core model of carbon catabolite repression. The model variants, with two substrate assimilation pathways and four intracellular metabolites only, differ from one another in only a single aspect, each breaking the symmetry between the two pathways in a different manner. Interestingly, all model variants are able to reproduce the data from a reference diauxic growth experiment. For each of the model variants, we predicted the behaviour in two new experimental conditions. When qualitatively comparing these predictions with experimental data, a number of models could be excluded while other model variants are still not discriminable. The best-performing model variants are based on inducer inclusion and activation of enzymatic genes by a global transcription factor, but the other proposed factors may complement these well-known regulatory mechanisms. The model ensemble, which was described in a study pubmished in *BMC Systems Biology* this year, offers a better understanding of the variety of mechanisms that have been proposed to play a role in carbon catabolite repression, but is also useful as an educational resource for systems biology.

The same focus on the dynamics of physiological processes has shaped a project on the post-transcriptional control of carbon central metabolism in *E. coli*. In the framework of the PhD thesis of Manon Morin, supported by a Contrat Jeune Scientifique INRA-Inria, the collaboration of Delphine Ropers with Muriel Cocaign-Bousquet and Brice Enjalbert at INRA/INSA Toulouse has demonstrated the key role played by the post-transcriptional regulatory system CSR in growth transitions in a series of publications in the past few years (*e.g.*, [9]). The collaboration with INRA/INSA de Toulouse is continued in the context of the PhD thesis of Thibault Etienne, funded by an INRA-Inria PhD grant, with the objective of developing models able to explain how cells coordinate their physiology and the functioning of the transcription, translation, and degradation machineries following changes in the availability of carbon sources in the environment. This work is further supported by the ANR project ECORIB accepted this year and an IXXI grant in collaboration with the ERABLE project-team (Section 7.2 ).

## 6.6. Modelling bacterial growth

Various mathematical approaches have been used in the literature to describe the networks of biochemical reactions involved in microbial growth. With various levels of detail, the resulting models provide an integrated view of these reaction networks, including the transport of nutrients from the environment and the metabolism and gene expression allowing the conversion of these nutrients into biomass. The models hence bridge the scale between individual reactions to the growth of cell populations. Analysing the dynamics of some of these models mentioned above becomes quickly intractable, when mathematical functions are for instance given by complex algebraic expressions resulting from the mass balance of biochemical reactions. In a paper published in the *Bulletin of Mathematical Biology* [13], Edith Grac, former post-doc in IBIS, Delphine Ropers, and Stefano Casagranda and Jean-Luc Gouzé from the BIOCORE project-team, have studied how monotone system theory and time-scale arguments can be used to reduce high-dimension models based on the mass-action law. Applying the approach to an important positive feedback loop regulating the expression of RNA polymerase in *E. coli*, made it possible to study the stability of the system steady states and relate the dynamical behaviour of the system to observations on the physiology of the bacterium *E. coli*.

In another paper published in *BMC Systems Biology* [14], Delphine Ropers and BIOCORE members Stefano Casagranda, Jean-Luc Gouzé, and Suzanne Touzeau, have developed a new approach to deal with model complexity. The approach, named Principle Process Analysis, allows to identify processes playing a key role in the model dynamics and to reduce the complex dynamics to these core processes, omitting processes that are

inactive. In particular, it has allowed the reduction of a well-known model of circadian rhythms in mammals into a succession of simpler submodels. Their analysis has resulted in the identification of the source of circadian oscillations, the main oscillator being the negative feedback loop involving proteins PER, CRY, CLOCK-BMAL1, in agreement with previous modelling and experimental studies.

Recent work has shown that coarse-grained models of resource allocation can account for a number of empirical regularities relating the the macromolecular composition of the cell to the growth rate. Some of these models hypothesize control strategies enabling microorganisms to optimize growth. While these studies focus on steady-state growth, such conditions are rarely found in natural habitats, where microorganisms are continually challenged by environmental fluctuations. In recent years, in the framework of the PhD thesis of Nils Giordano, we extended the study of microbial growth strategies to dynamical environments, using a self-replicator model. In collaboration with the BIOCORE project-team, we formulated dynamical growth maximization as an optimal control problem that can be solved using Pontryagin's Maximum Principle and we compared the theoretical results thus obtained with different possible implementations of growth control in bacterial cells [5]. The extension and experimental validation of some of these results are currently being carried out by Antrea Pavlou in the framework of her PhD project, funded by the ANR project Maximic (Section 7.2 ).

## 6.7. Growth control in bacteria and biotechnological applications

The ability to experimentally control the growth rate is crucial for studying bacterial physiology. It is also of central importance for applications in biotechnology, where often the goal is to limit or even arrest growth. Growth-arrested cells with a functional metabolism open the possibility to channel resources into the production of a desired metabolite, instead of wasting nutrients on biomass production. In recent years we obtained a foundation result for growth control in bacteria [6], in that we engineered an *E. coli* strain where the transcription of a key component of the gene expression machinery, RNA polymerase, is under the control of an inducible promoter. By changing the inducer concentration in the medium, we can adjust the RNA polymerase concentration and thereby switch bacterial growth between zero and the maximal growth rate supported by the medium. The publication also presented a biotechnological application of the synthetic growth switch in which both the wild-type *E. coli* strain and our modified strain were endowed with the capacity to produce glycerol when growing on glucose. Cells in which growth has been switched off continue to be metabolically active and harness the energy gain to produce glycerol at a twofold higher yield than in cells with natural control of RNA polymerase expression.

The experimental work underlying the growth switch has been continued in several directions in the context of the Maximic project by Célia Boyat. Moreover, in collaboration with colleagues from the BIOCORE project-team, we have formulated the maximization of metabolite production by means of the growth switch as a resource reallocation problem that can be analyzed by means of the self-replicator models of bacterial growth mentioned in Section 6.6 in combination with methods from optimal control theory. In a publication accepted for the *Journal of Mathematical Biology* this year [20], we study various optimal control problems by means of a combination of analytical and computational techniques. We show that the optimal solutions for biomass maximization and product maximization are very similar in the case of unlimited nutrient supply, but diverge when nutrients are limited. Moreover, external growth control overrides natural feedback growth control and leads to an optimal scheme consisting of a first phase of growth maximization followed by a second phase of product maximization. This two-phase scheme agrees with strategies that have been proposed in metabolic engineering. More generally, this work shows the potential of optimal control theory for better understanding and improving biotechnological production processes. Extensions concerning the effect on growth and bioproduction of the (biological or technological) costs associated with discontinuous control strategies, and of the time allotted to optimal substrate utilization, are described in a contribution to a control theory conference submitted this year.

**IMAGINE Project-Team**

# 7. New Results

## 7.1. Sculpting Mountains: Interactive Terrain Modeling Based on Subsurface Geology

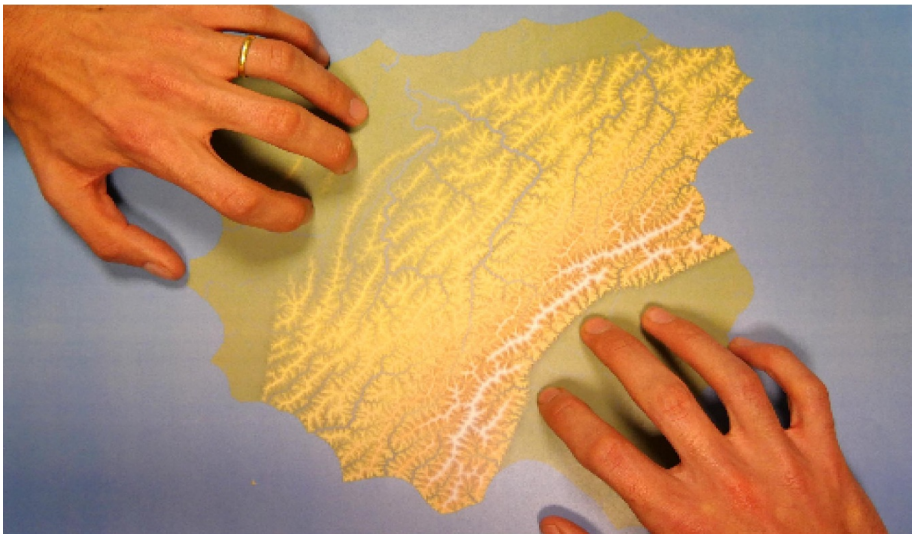


*Figure 2. Sculpting Mountains: Interactive Terrain Modeling Based on Subsurface Geology*

Most mountain ranges are formed by the compression and folding of colliding tectonic plates. Subduction of one plate causes large-scale asymmetry while their layered composition (or stratigraphy) explains the multi-scale folded strata observed on real terrains. As part of Guillaume Cordonnier's PhD thesis, we introduced a novel interactive modeling technique to generate visually plausible, large scale terrains that capture these phenomena (illustrated in Fig. 2 ). Our method draws on both geological knowledge for consistency and on sculpting systems for user interaction. The user is provided hands-on control on the shape and motion of tectonic plates, represented using a new geologically-inspired model for the Earth crust. The model captures their volume preserving and complex folding behaviors under collision, causing mountains to grow. It generates a volumetric uplift map representing the growth rate of subsurface layers. Erosion and uplift movement are jointly simulated to generate the terrain. The stratigraphy allows us to render folded strata on eroded cliffs. We validated the usability of our sculpting interface through a user study, and compare the visual consistency of the earth crust model with geological simulation results and real terrains.

## 7.2. Exploratory design of mechanical devices with motion constraints

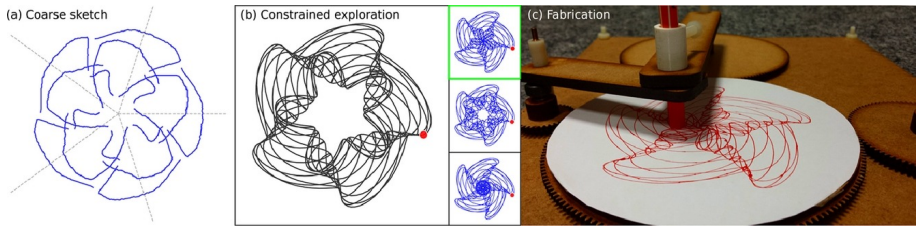



*Figure 3. Exploratory design of mechanical devices with motion constraints. (a) The user first selects a mechanically feasible drawing by providing a rough sketch. (b) The user is then able to interactively explore local alternatives (b) by defining visual constraints directly on the pattern (here, the cusp position). (c) The resulting machine is automatically exported to laser cutter profiles for fabrication.*

Mechanical devices are ubiquitous in our daily lives, and the motion they are able to transmit is often a critical part of their function. While digital fabrication devices facilitate their realization, motion-driven mechanism design remains a challenging task. We take drawing machines as a case study in exploratory design. Devices such as the Spirograph can generate intricate patterns from an assembly of simple mechanical elements. Trying to control and customize these patterns, however, is particularly hard, especially when the number of parts increases. We propose a novel constrained exploration method that enables a user to easily explore feasible drawings by directly indicating pattern preferences at different levels of control. This is (illustrated in Fig. 3 ). The user starts by selecting a target pattern with the help of construction lines and rough sketching, and then fine-tunes it by prescribing geometric features of interest directly on the drawing. The designed pattern can then be directly realized with an easy-to-fabricate drawing machine. The key technical challenge is to facilitate the exploration of the high dimensional configuration space of such fabricable machines. To this end, we propose a novel method that dynamically reparameterizes the local configuration space and allows the user to move continuously between pattern variations, while preserving user-specified feature constraints. We tested our framework on several examples, conducted a user study, and fabricated a sample of the designed examples.

## 7.3. Automatic Generation of Geological Stories from a Single Sketch

Describing the history of a terrain from a vertical geological cross-section is an important problem in geology, called geological restoration. Designing the sequential evolution of the geometry is usually done manually, involving many trials and errors. In this work, we recast this problem as a storyboarding problem, where the diifferent stages in the restoration are automatically generated as storyboard panels and displayed as geological stories. Our system allows geologists to interactively explore multiple scenarios by selecting plausible geological event sequences and backward simulating them at interactive rate, causing the terrain layers to be progressively un-deposited, un-eroded, un-compacted, unfolded and un-faulted. Storyboard sketches are generated along the way. When a restoration is complete, the storyboard panels can be used for automatically generating a forward animation of the terrain history, enabling quick visualization and validation of hypotheses. As a proof-of-concept, we describe how our system was used by geologists to restore and animate cross-sections in real examples at various spatial and temporal scales and with diffierent levels of complexity, including the Chartreuse region in the French Alps.

## 7.4. 3D Shape Decomposition and Sub-parts Classification

This paper (illustrated in Fig. 5 ) introduces a measure of significance on a curve skeleton of a 3D piecewise linear shape mesh, allowing the computation of both the shape's parts and their saliency. We begin by reformulating three existing pruning measures into a non-linear PCA along the skeleton. From this PCA,

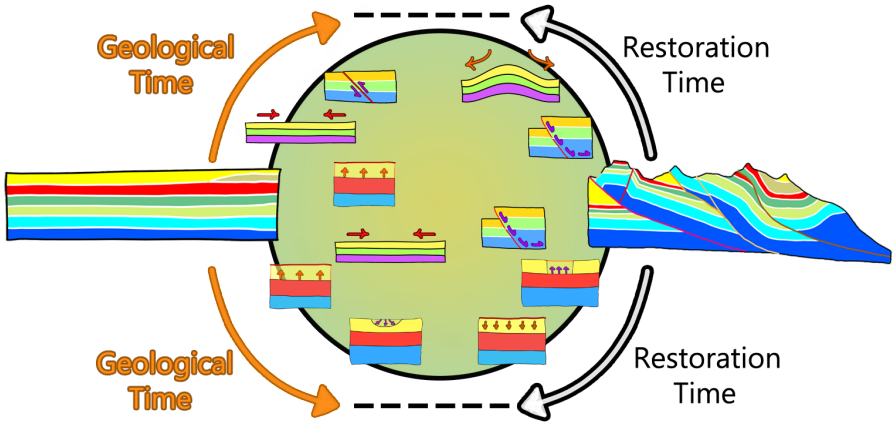



*Figure 4. Automatic Generation of Geological Stories from a Single Sketch. From left to right, the original terrain from several million years ago undergoes events that will transform it to its current state. From right to left, the current terrain is restored and undergoes undo events that will transform it back to its original state.*

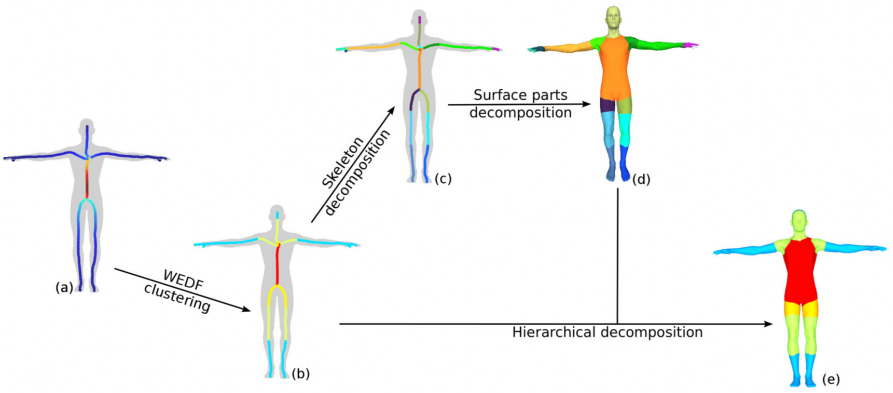



*Figure 5. 3D Shape Decomposition and Sub-parts Classification. Starting from a 3D shape and its curve skeleton, we compute a new measure called WEDF on the curve skeleton (a) and, by clustering WEDF values, we decompose the skeleton into hierarchical parts (b). To each connected part on the skeleton –shown with a different color (c)– a connected region of the surface mesh is assigned (d). Then, a salience value according to the hierarchy is assigned to each corresponding surface part (e) –parts of same importance get a similar color.*

we then derive a volume-based salience measure, the 3D WEDF, that determines the relative importance to the global shape of the shape part associated to a point of the skeleton. First, we provide robust algorithms for computing the 3D WEDF on a curve skeleton, independent on the number of skeleton branches. Then, we cluster the WEDF values to partition the curve skeleton, and coherently map the decomposition to the associated surface mesh. Thus, we develop an unsupervised hierarchical decomposition of the mesh faces into visually meaningful shape regions that are ordered according to their degree of perceptual salience. The shape analysis tools introduced in this paper are important for many applications including shape comparison, editing, and compression.

## 7.5. Interactive Generation of Time-evolving, Snow-Covered Landscapes with Avalanches

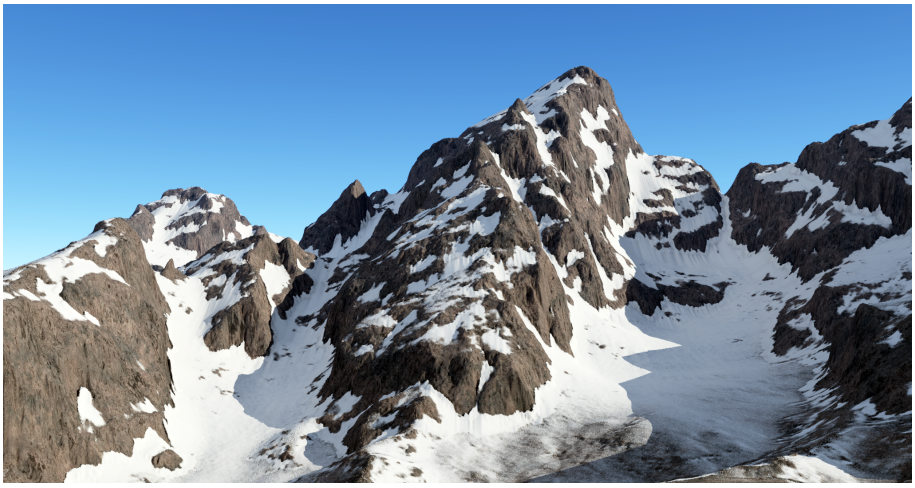


*Figure 6. Interactive Generation of Time-evolving, Snow-Covered Landscapes with Avalanches*

As part of Guillaume Cordonnier's PhD thesis, we also introduced a novel method for interactive generation of visually consistent, snow-covered landscapes, which provides control of their dynamic evolution over time. Our main contribution (illustrated in Fig. 6 ) was the real-time phenomenological simulation of avalanches and other user-guided events, such as tracks left by Nordic skiing, which can be applied to interactively sculpt the landscape. The terrain is modeled as a height field with additional layers for stable, compacted, unstable, and powdery snow, which behave in combination as a semi-viscous fluid. We incorporate the impact of several phenomena, including sunlight, temperature, prevailing wind direction, and skiing activities. The snow evolution includes snow-melt and snowdrift, which affect stability of the snow mass and the probability of avalanches. A user can shape landscapes and their evolution either with a variety of interactive brushes, or by prescribing events along a winter season time-line. Our optimized GPU-implementation allows interactive updates of snow type and depth across a large ($10 \times 10$ km) terrain, including real-time avalanches, making this suitable for visual assets in computer games. We evaluated our method through perceptual comparison against exiting methods and real snow-depth data.

**MAVERICK Project-Team**

# 7. New Results

## 7.1. Expressive Rendering

### *7.1.1. A workflow for designing stylized shading effects*

**Participants:** Alexandre Bléron, Romain Vergne, Thomas Hurtut, Joëlle Thollot.

In this report [18], we describe a workflow for designing stylized shading effects on a 3D object, targeted at technical artists. Shading design, the process of making the illumination of an object in a 3D scene match an artist vision, is usually a time-consuming task because of the complex interactions between materials, geometry, and lighting environment. Physically based methods tend to provide an intuitive and coherent workflow for artists, but they are of limited use in the context of non-photorealistic shading styles. On the other hand, existing stylized shading techniques are either too specialized or require considerable hand-tuning of unintuitive parameters to give a satisfactory result. Our contribution is to separate the design process of individual shading effects in three independent stages: control of its global behavior on the object, addition of procedural details, and colorization. Inspired by the formulation of existing shading models, we expose different shading behaviors to the artist through parametrizations, which have a meaningful visual interpretation. Multiple shading effects can then be composited to obtain complex dynamic appearances. The proposed workflow is fully interactive, with real-time feedback, and allows the intuitive exploration of stylized shading effects, while keeping coherence under varying viewpoints and light configurations (see Fig. 2 ). Furthermore, our method makes use of the deferred shading technique, making it easily integrable in existing rendering pipelines.

### *7.1.2. MNPR: A framework for real-time expressive non-photorealistic rendering of 3D computer graphics*

**Participants:** Santiago Montesdeoca, Hock Soon Seah, Amir Semmo, Pierre Bénard, Romain Vergne, Joëlle Thollot, Davide Benvenuti.

We propose a framework for expressive non-photorealistic rendering of 3D computer graphics: MNPR. Our work focuses on enabling stylization pipelines with a wide range of control, thereby covering the interaction spectrum with real-time feedback. In addition, we introduce control semantics that allow cross-stylistic art-direction, which is demonstrated through our implemented watercolor, oil and charcoal stylizations (see Fig. 3 ). Our generalized control semantics and their style-specific mappings are designed to be extrapolated to other styles, by adhering to the same control scheme. We then share our implementation details by breaking down our framework and elaborating on its inner workings. Finally, we evaluate the usefulness of each level of control through a user study involving 20 experienced artists and engineers in the industry, who have collectively spent over 245 hours using our system. MNPR is implemented in Autodesk Maya and open-sourced through this publication, to facilitate adoption by artists and further development by the expressive research and development community. This paper was presented at Expressive [13] and received the best paper award.

### *7.1.3. Motion-coherent stylization with screen-space image filters*

**Participants:** Alexandre Bléron, Romain Vergne, Thomas Hurtut, Joëlle Thollot.

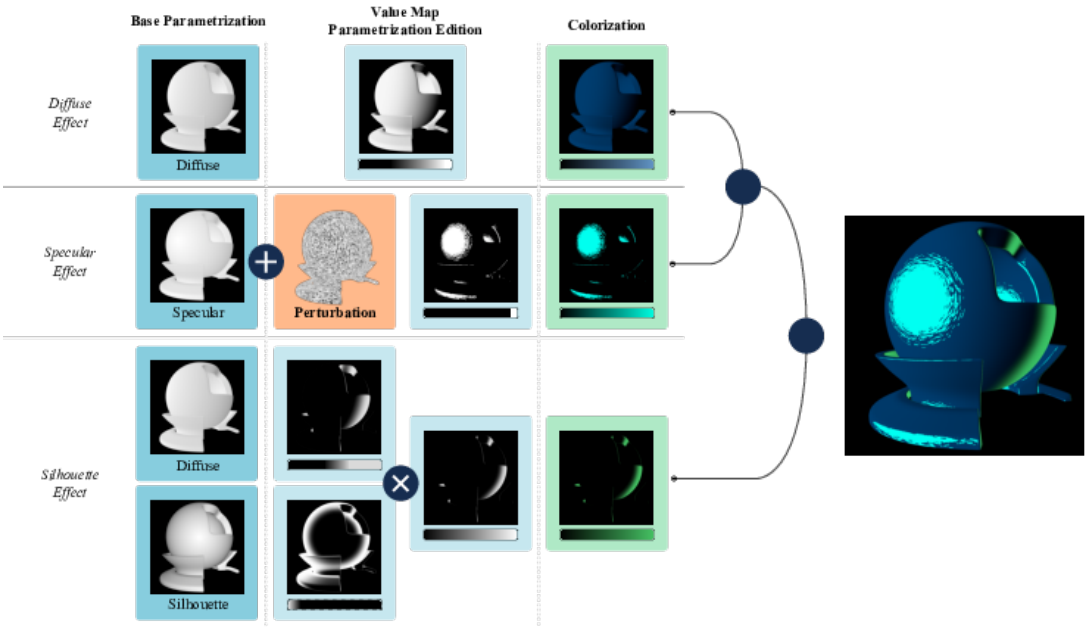



*Figure 2. Illustration of our workflow showing an example with three appearance effects. A user can modify and combine base parametrizations to design the shading behavior (blue nodes) of an appearance effect, using value maps and combination operations. A color map (green nodes) is then applied on the designed behavior to colorize the effect. Output effects are then composited to obtain the final appearance. Perturbations (orange nodes) can be attached to every operation in order to add procedural details to an effect. The orientation of the perturbation can be controlled by the gradient of a shading behavior (as shown here), or by an external vector field, such as a tangent map.*


*Figure 3. A scene rendered through MNPR in different styles. Baba Yaga's hut model, © Inuciian.*

One of the qualities sought in expressive rendering is the 2D impression of the resulting style, called flatness. In the context of 3D scenes, screen-space stylization techniques are good candidates for flatness as they operate in the 2D image plane, after the scene has been rendered into G-buffers. Various stylization filters can be applied in screen-space while making use of the geometrical information contained in G-buffers to ensure motion coherence. However, this means that filtering can only be done inside the rasterized surface of the object. This can be detrimental to some styles that require irregular silhouettes to be convincing. In this paper, we describe a post-processing pipeline that allows stylization filters to extend outside the rasterized footprint of the object by locally *inflating* the data contained in G-buffers (see Fig. 4 ). This pipeline is fully implemented on the GPU and can be evaluated at interactive rates. We show how common image filtering techniques, when integrated in our pipeline and in combination with G-buffer data, can be used to reproduce a wide range of *digitally-painted* appearances, such as directed brush strokes with irregular silhouettes, while keeping enough motion coherence. This paper was presented at Expressive [11].

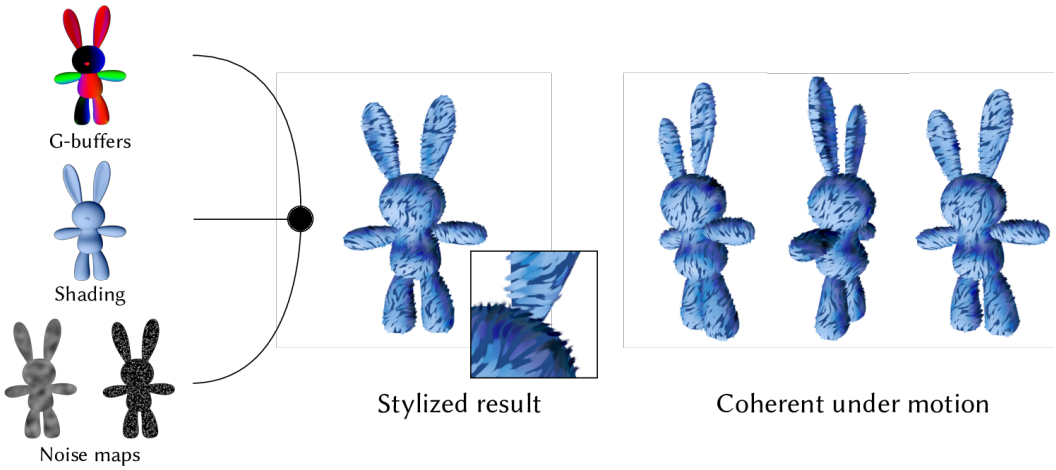



*Figure 4. Using standard G-buffers and auxiliary buffers (noise, shading) as input, our pipeline can reproduce stylization effects that extend outside the original rasterized footprint of the object. Visual features produced by the filters stay coherent under motion or viewpoint changes.*

## 7.2. Illumination simulation and materials

### *7.2.1. Rendering homogeneous participating media*

**Participants:** Beibei Wang, Nicolas Holzschuch, Liangsheng Ge, Lu Wang.

Illumination effects in translucent materials are a combination of several physical phenomena: refraction at the surface, absorption and scattering inside the material. Because refraction can focus light deep inside the material, where it will be scattered, practical illumination simulation inside translucent materials is difficult. We have worked on a Point-Based Global Illumination method for light transport on homogeneous translucent materials with refractive boundaries. We start by placing light samples inside the translucent material and organizing them into a spatial hierarchy. At rendering, we gather light from these samples for each camera ray. We compute separately the sample contributions for single, double and multiple scattering, and add them. Multiple scattering effects are precomputed and stores in a table, accessed at runtime. An illustration of our approach is given in Fig 5 . We present two implementations of our algorithm: an offline version for high-quality rendering and an interactive GPU implementation. The offline version provides significant speed-ups and reduced memory footprints compared to state-of-the-art algorithms, with no visible impact on quality.

The GPU version yields interactive frame rates: 30 fps when moving the viewpoint, 25 fps when editing the light position or the material parameters. This work was published in IEEE Transactions on Visualization and Computer Graphics [9].

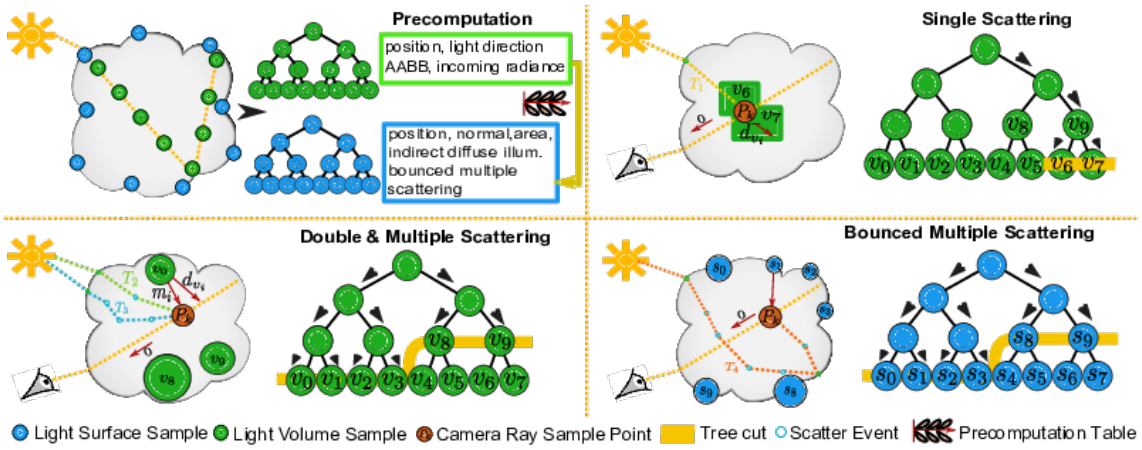



*Figure 5. Our algorithm: we begin by computing incoming light at volume and surface samples. We then compute Single-, Double- and Multiple scattering effects for each camera ray using these volume and surface samples.*

Storing the precomputed table for these multiple scattering effects is the largest memory cost for this algorithm. In a separate work, we used a neural network to encode these effects. We replaced the precomputed multiple scattering table with a trained neural network, with a cost of 6490 bytes (1623 floats). At runtime, the neural network is used to generate multiple scattering. The approach can be combined with many rendering algorithms, as illustrated in Fig. 6 . This work was published as a Siggraph Talk [12].

### *7.2.2. Fast global illumination with discrete stochastic microfacets using a filterable model*

**Participants:** Beibei Wang, Lu Wang, Nicolas Holzschuch.

Many real-life materials have a sparkling appearance, whether by design or by nature. Examples include metallic paints, sparkling varnish but also snow. These sparkles correspond to small, isolated, shiny particles reflecting light in a specific direction, on the surface or embedded inside the material. The particles responsible for these sparkles are usually small and discontinuous. These characteristics make it difficult to integrate them efficiently in a standard rendering pipeline, especially for indirect illumination. Existing approaches use a 4-dimensional hierarchy, searching for light-reflecting particles simultaneously in space and direction. The approach is accurate, but still expensive. We have shown that this 4-dimensional search can be approximated using separate 2-dimensional steps. This approximation allows fast integration of glint contributions for large footprints, reducing the extra cost associated with glints be an order of magnitude, as illustrated in Fig. 7 . This work was published in Computer Graphics Forum and presented at the Pacific Graphics conference [10].

### *7.2.3. Handling fluorescence in a uni-directional spectral path tracer*

**Participants:** Michal Mojzík, Alban Fichet, Alexander Wilkie

We present two separate improvements to the handling of fluorescence effects in modern uni-directional spectral rendering systems.

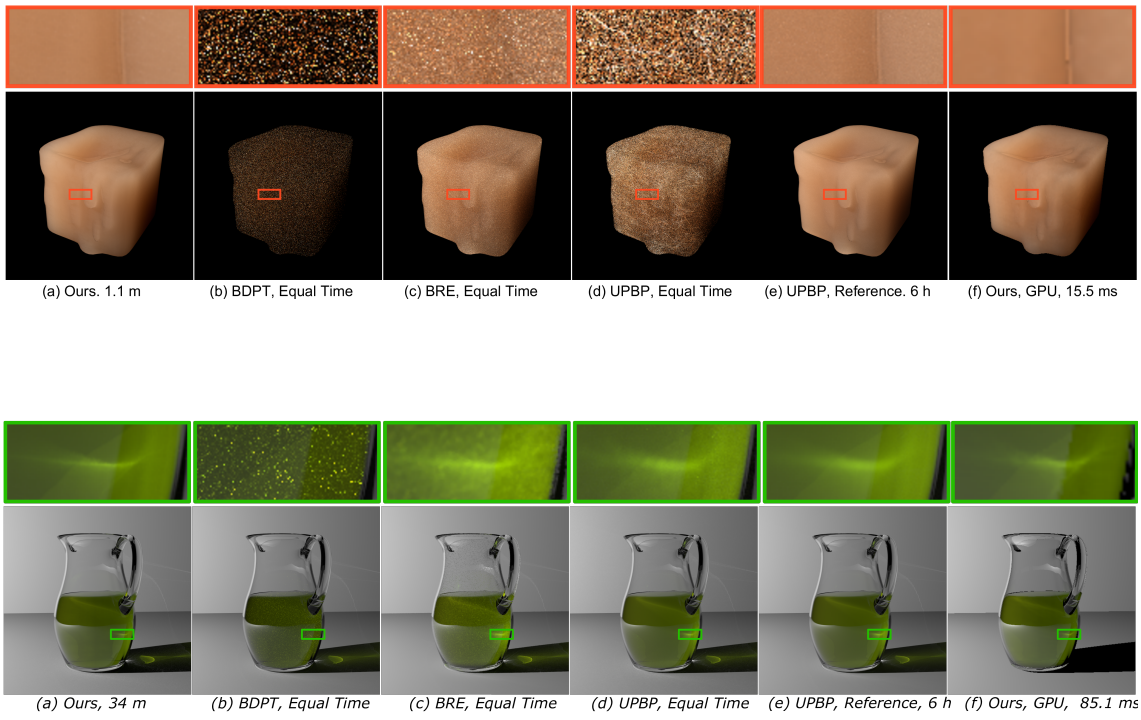



*Figure 6. comparison between our algorithm, other algorithms with equal time or equal quality and reference images. Top row: wax. For this material, with a large albedo and a small mean free path, multiple scattering effects dominate. Bottom row: olive oil. For this material with low albedo and large mean-free-path, low-order scattering effects dominate.*

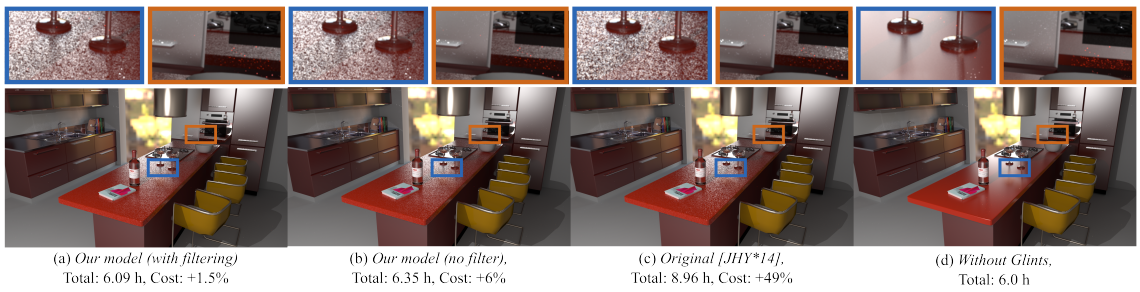
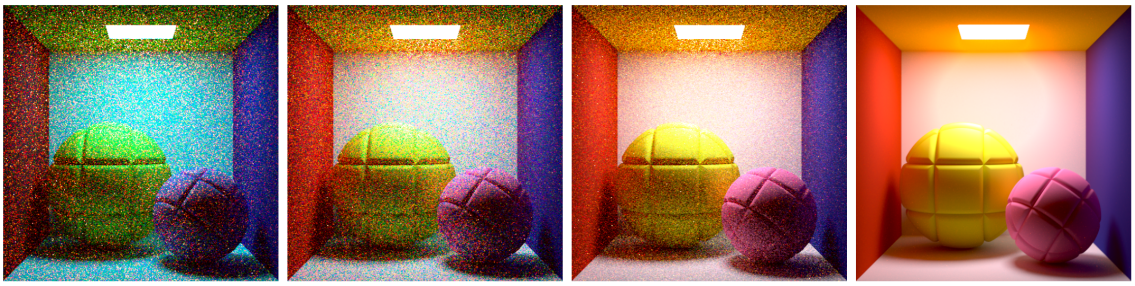



*Figure 7. Our algorithm, compared to the original Discrete Stochastic Microfacets model (c). Converting the 4D search to a product of 2D searches (b) produces almost identical results. This is the basis for our filterable model (a), which allows fast global illumination with negligible cost..*

The first is the formulation of a new distance tracking scheme for fluorescent volume materials which exhibit a pronounced wavelength asymmetry. Such volumetric materials are an important and not uncommon corner case of wavelength-shifting media behaviour, and have not been addressed so far in rendering literature. This new tracking scheme (figure 8 (b)) converges faster than a simple modification that can be added to the traditional exponential tracking (figure 8 (a)).

The second one is that we introduce an extension of Hero wavelength sampling which can handle fluorescence events, both on surfaces, and in volumes. Both improvements are useful by themselves, and can be used separately: when used together, they enable the robust inclusion of arbitrary fluorescence effects in modern uni-directional spectral MIS path tracers (figure 8 (c)). Our extension of Hero wavelength sampling is generally useful, while our proposed technique for distance tracking in strongly asymmetric media is admittedly not very efficient. However, it makes the most of a rather difficult situation, and at least allows the inclusion of such media in uni-directional path tracers, albeit at comparatively high cost. Which is still an improvement since up to now, their inclusion was not really possible at all, due to the inability of conventional tracking schemes to generate sampling points in such volume materials. This work was published in the journal Computer Graphics Forum [6].



(a) $\max(\sigma_t, \sigma_s)$, 16 spp, mono (b) Fluorescent-aware, 16 spp, mono (c) Fluo-aware, 16spp, HWSS (d) Fluo-aware, 65000 spp, HWSS

*Figure 8. Comparison of proposed techniques to improve rendering of fluorescence.*

### *7.2.4. A versatile parameterization for measured material manifolds*

**Participants:** Cyril Soler, Kartic Subr, Derek Nowrouzezahrai.

A popular approach for computing photorealistic images of virtual objects requires applying reflectance profiles measured from real surfaces, introducing several challenges: the memory needed to faithfully capture realistic material reflectance is large, the choice of materials is limited to the set of measurements, and image synthesis using the measured data is costly. Typically, this data is either compressed by projecting it onto a subset of its linear principal components or by applying non-linear methods. The former requires many components to faithfully represent the input reflectance, whereas the latter necessitates costly extrapolation algorithms. We learn an underlying, low-dimensional non-linear reflectance manifold amenable to rapid exploration and rendering of real-world materials. We can express interpolated materials as linear combinations of the measured data, despite them lying on an inherently non-linear manifold. This allows us to efficiently interpolate and extrapolate measured BRDFs, and to render directly from the manifold representation. We exploit properties of Gaussian process latent variable models and use our representation for high-performance and offline rendering with interpolated real-world materials. This work has been published in the journal Computer Graphics Forum [7], and presented at Eurographics 2018.

## 7.3. Complex scenes

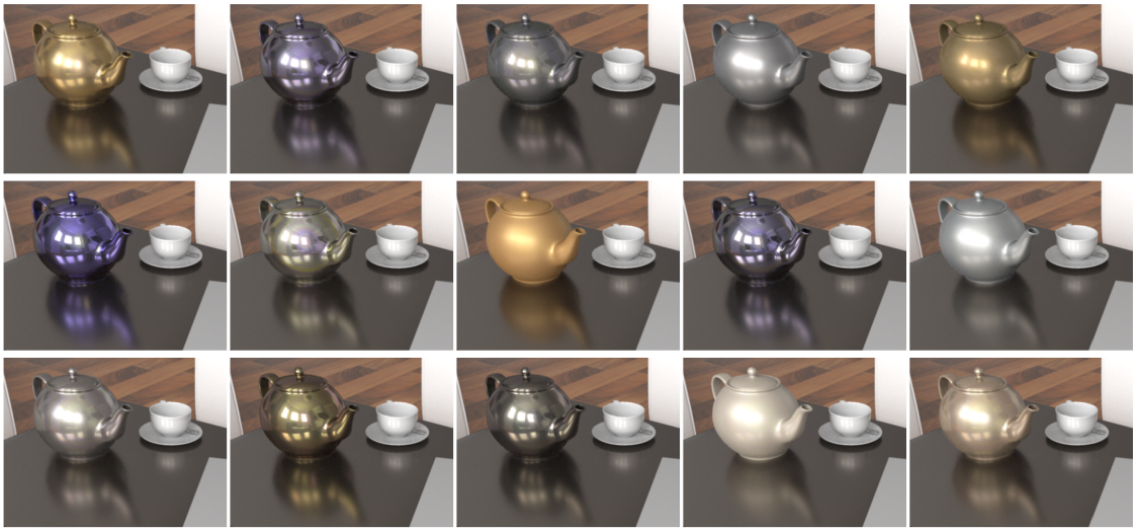


*Figure 9. Four of the images above (Number 2, 4, 6 and 12 in reading order) are rendered with measured BRDFs from the MERL dataset, the remaining 11 being rendered with BRDFs randomly picked from our parameterization of the non-linear manifold containing MERL materials. We explore this manifold interactively to produce high-quality BRDFs which retain the physical properties and perceptual aspect of real materials.*

### *7.3.1. A new microflake model with microscopic self-shadowing for accurate volume downsampling*

**Participants:** Guillaume Loubet, Fabrice Neyret.

In this work, we addressed the problem of representing the effect of internal self-shadowing in elements about to be filtered out at a given LOD, in the scope of volume of voxels containing density and phase-function (represented by a microflakes).

Naïve linear methods for downsampling high resolution microflake volumes often produce inaccurate results, especially when input voxels are very opaque. Preserving correct appearance at all resolutions requires taking into account inter- and intravoxel self-shadowing effects (see Figure 10 ). We introduce a new microflake model whose parameters characterize self-shadowing effects at the microscopic scale. We provide an anisotropic self-shadowing function and a microflake distribution for which scattering coefficients and phase functions of our model have closed-form expressions. We use this model in a new downsampling approach in which scattering parameters are computed from local estimations of self-shadowing in the input volume. Unlike previous work, our method handles datasets with spatially varying scattering parameters, semi-transparent volumes and datasets with intricate silhouettes. We show that our method generates LoDs with correct transparency and consistent appearance through scales for a wide range of challenging datasets, allowing for huge memory savings and efficient distant rendering without loss of quality. This work received the Best Paper Award at Eurographics 2018 and was published in the journal Computer Graphics Forum [5].

## 7.4. Texture synthesis

### *7.4.1. Gabor noise revisited*

**Participants:** Vincent Tavernier, Fabrice Neyret, Romain Vergne, Joëlle Thollot.

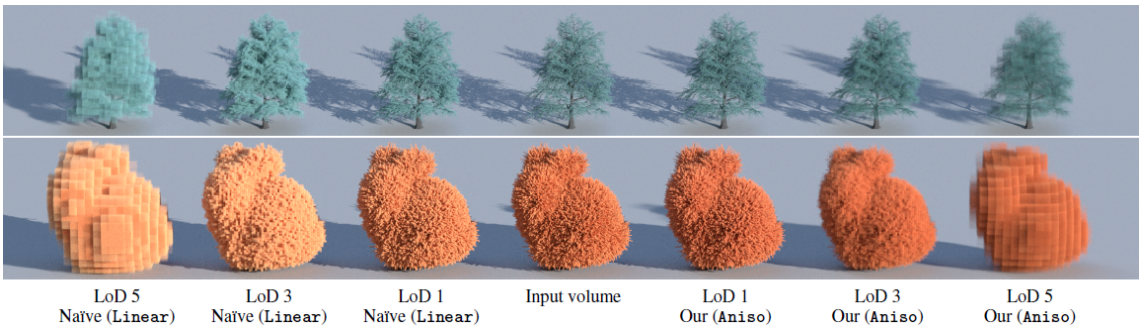
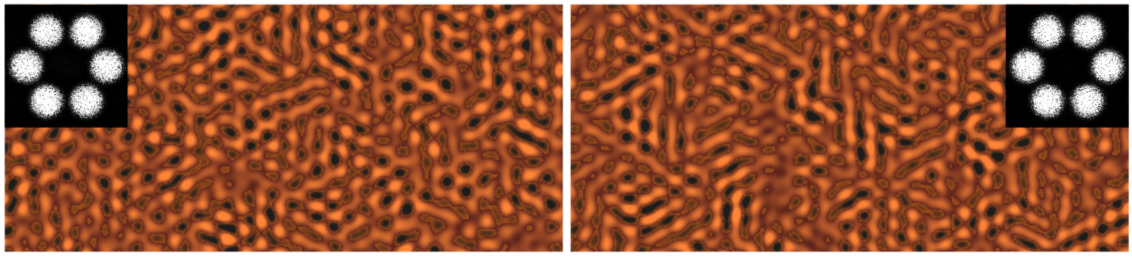



*Figure 10. Comparison between naïve downsampling of microflake volumes and our method ("Aniso"). Naïve dowsampling of volumes with dense voxels often lead to inaccurate results due to the loss of inter- and intra-voxel self-shadowing effects. Our method is based on a new participating medium model and on local estimations of self-shadowing. It generates LoDs with correct transparency and consistent appearance through scales. Rendered with volume path tracing in Mitsuba (http://www.mitsuba-renderer.org/): the trunk of the cedar is a mesh.*

Gabor Noise is a powerful procedural texture synthesis technique, but has two major drawbacks: It is costly due to the high required splat density and not always predictable because properties of instances can differ from those of the process. We bench performance and quality using alternatives for each Gabor Noise ingredient: point distribution, kernel weighting and kernel shape. For this, we introduce 3 objective criteria to measure process convergence, process stationarity, and instance stationarity. We show that minor implementation changes allow for $17-24\times$ speed-up with same or better quality (see Fig. 11 ).

This paper was presented at AFIG [17] and received the best paper award. An article has been submitted to Eurographics-short 2019.



(a) Seminal Gabor, $N = 45$ (b) Bernoulli+strat.+sin, $N = 3$

*Figure 11. Real case with complex power spectrum (3 kernels, cf. inset) and non-linear post-treatment. Our optimized set of ingredients achieves the same visual quality in $1/17^{th}$ of the time required by the seminal method.*

### *7.4.2. High-performance by-example noise using a histogram-preserving blending operator*

**Participants:** Eric Heitz, Fabrice Neyret.

We propose a new by-example noise algorithm that takes as input a small example of a stochastic texture and synthesizes an infinite output with the same appearance. It works on any kind of random-phase inputs as well as on many non-random-phase inputs that are stochastic and non-periodic, typically natural textures such as moss, granite, sand, bark, etc. Our algorithm achieves high-quality results comparable to state-of-the-art procedural-noise techniques but is more than 20 times faster. Our approach is conceptually simple: we partition the output texture space on a triangle grid and associate each vertex with a random patch from the input such that the evaluation inside a triangle is done by blending 3 patches. The key to this approach is the blending operation that usually produces visual artifacts such as ghosting, softened discontinuities and reduced contrast, or introduces new colors not present in the input. We analyze these problems by showing how linear blending impacts the histogram and show that a blending operator that preserves the histogram prevents these problems. The main requirement for a rendering application is to implement such an operator in a fragment shader without further post-processing, i.e. we need a histogram-preserving blending operator that operates only at the pixel level. Our insight for the design of this operator is that, with Gaussian inputs, histogram-preserving blending boils down to mean and variance preservation, which is simple to obtain analytically. We extend this idea to non-Gaussian inputs by "Gaussianizing" them with a histogram transformation and "de-Gaussianizing" them with the inverse transformation after the blending operation. We show how to precompute and store these histogram transformations such that our algorithm can be implemented in a fragment shader, as illustrated in Fig. 12 . This work received the Best Paper Award at High Performance Graphics 2018 [4].

## 7.5. Visualization

### *7.5.1. A "What if" approach for eco-feedback*

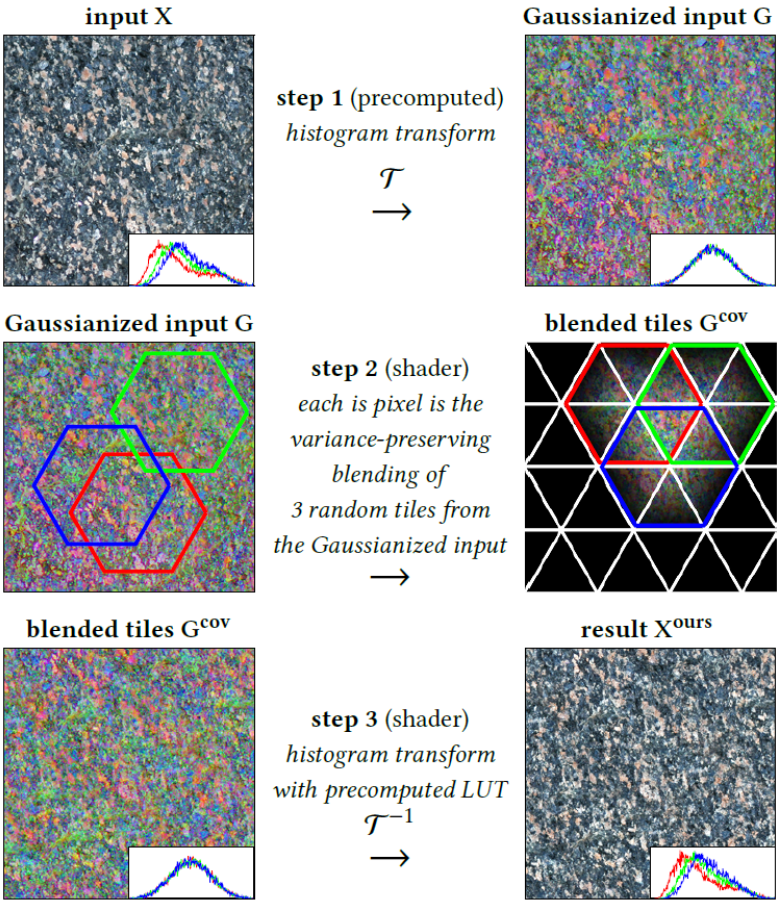
**Participants:** Jérémy Wambecke, Georges-Pierre Bonneau, Romain Vergne, Renaud Blanch.

Many households share the objective of reducing electricity consumption for either economic or ecological motivations. Eco-feedback technologies support this objective by providing users with a visualization of their consumption. However as pointed out by several studies, users encounter difficulties in finding concrete actions to reduce their consumption. To overcome this limitation, we introduce and evaluate Activelec, a system based on the visualization and interaction with user's behavior rather than raw consumption data. The user's behavior is modeled as the set of actions modifying the state of appliances over time. A key novelty of our solution is its focus on the What if approach applied to eco-feedback. Users can analyze and experiment scenarios by selecting and modifying their usage of electrical appliances over time and visualize the impact on the consumption, as illustrated in Fig. 13 . In [16] we conducted two laboratory user studies that evaluate the usability of Activelec and the relevance of the What if approach for electricity consumption. Our results show that users understand the interaction paradigm and can easily find relevant modifications in their usage of appliances. Moreover participants judge these changes of behavior would require little effort to be adopted. In [15] we conducted an in-situ evaluation of Activelec, confirming these results in a real setting.

### *7.5.2. Morphorider: a new way for Structural Monitoring via Shape Acquisition*

**Participants:** Tibor Stanko, Laurent Jouanet, Nathalie Saguin-Sprynski, Georges-Pierre Bonneau, Stefanie Hahmann.

In collaboration with CEA-Leti we introduce a new kind of monitoring device, illustrated in Fig. 14 , allowing the shape acquisition of a structure via a single mobile node of inertial sensors and an odometer. Previous approaches used devices placed along a network with fixed connectivity between the sensor nodes (lines, grid). When placed onto a shape, this sensor network provides local surface orientations along a curve network on the shape, but its absolute position in the world space is unknown. The new mobile device provides a novel way of structures monitoring: the shape can be scanned regularly, and following the shape or some specific parameters along time may afford the detection of early signs of failure. Here, we present a complete framework for 3D shape reconstruction. To compute the shape, our main insight is to formulate the reconstruction as a set of optimization problems. Using discrete representations, these optimization problems are resolved efficiently and at interactive time rates. We present two main contributions. First, we introduce a novel method for creating well-connected networks with cell-complex topology using only orientation and distance measurements and



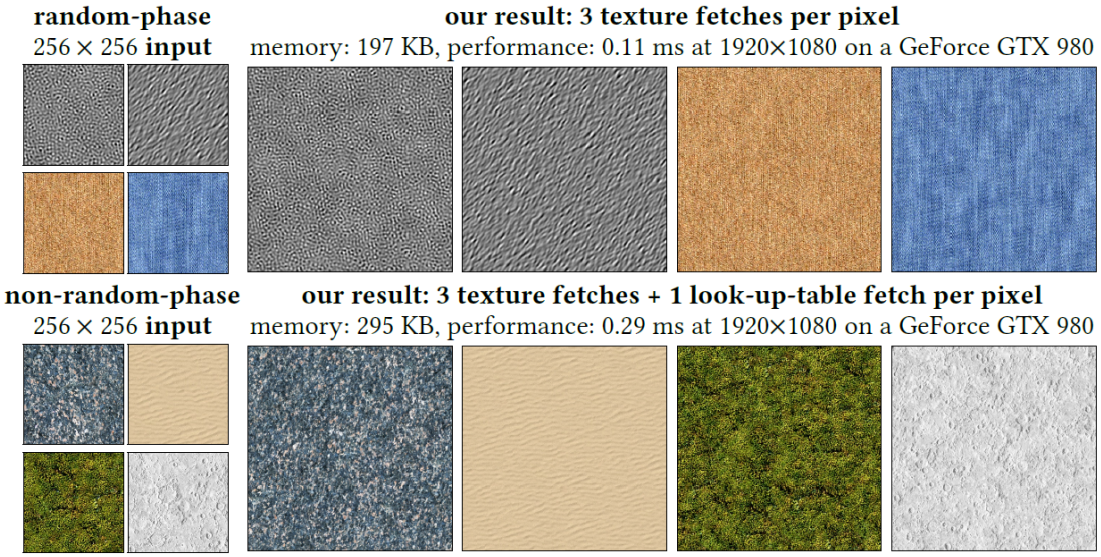




*Figure 12. Top: method overview. Bottom: results and performances.*

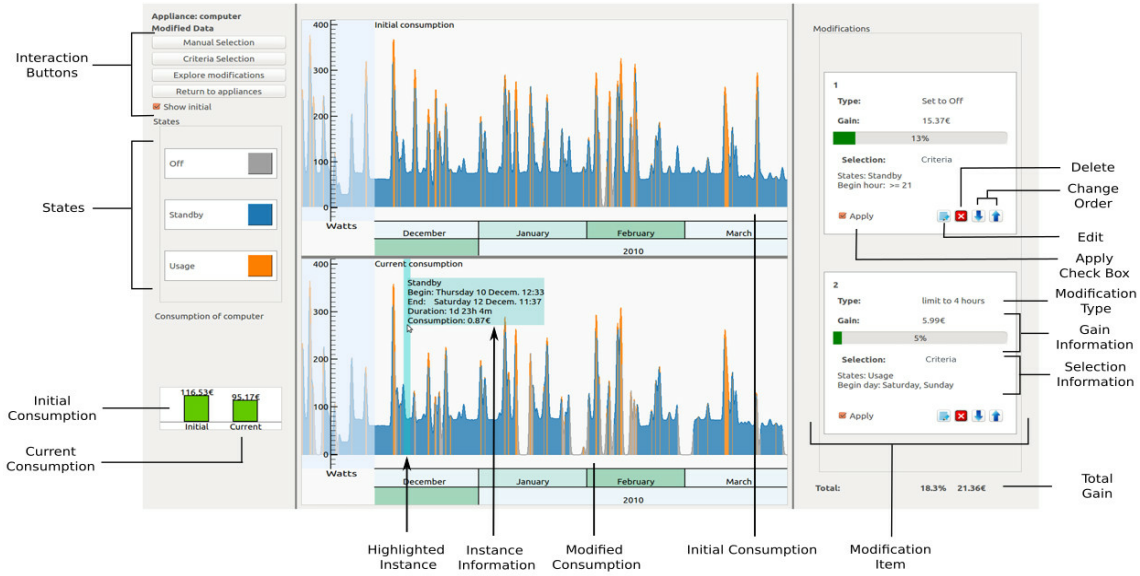



*Figure 13. Interface of our system. A computer has been chosen by the user, whose states are Usage (orange) and Standby (blue). At the right, we can see that the user has applied two modifications, the first one to remove instances of Standby after 9 P.M, and the second one to limit the instances of On to 4 hours during the weekend. When the user is selecting instances, this panel displays information about the selection.*

a set of user-defined constraints. Second, we address the problem of surfacing a closed 3D curve network with given surface normals. The normal input increases shape fidelity and allows to achieve globally smooth and visually pleasing shapes. The proposed framework was tested on experimental data sets acquired using our device. A quantitative evaluation was performed by computing the error of reconstruction for our own designed surfaces, thus with known ground truth. Even for complex shapes, the mean error remains around 1%. This work was published at the 9th European Workshop on Structural Health Monitoring [14].

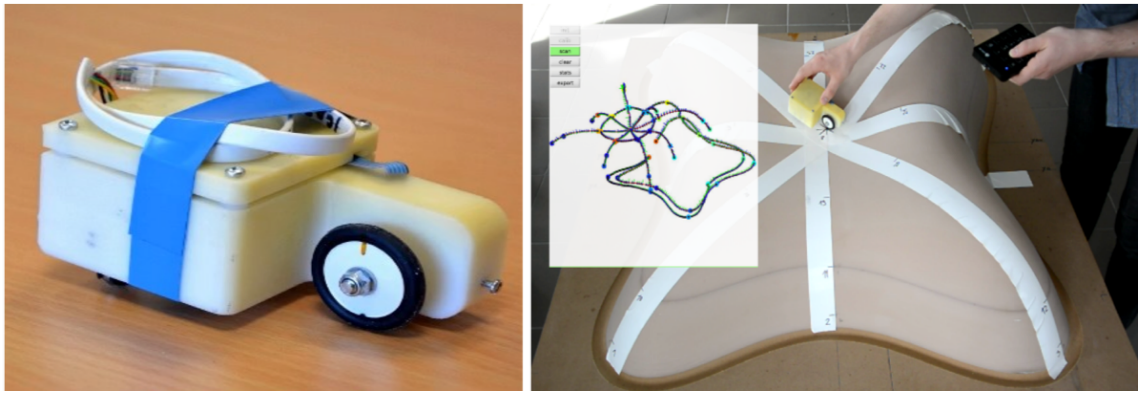


*Figure 14. Morphorider: Structural Monitoring via Shape Acquisition (right) with a mobile device (left) equipped with an inertial node of sensors and an odometer.*

**MISTIS Project-Team**

# 7. New Results

## 7.1. Mixture models

### *7.1.1. Hierarchical mixture of linear mappings in high dimension*

**Participant:** Florence Forbes.

**Joint work with**: Benjamin Lemasson from Grenoble Institute of Neuroscience, Naisyin Wang and Chun-Chen Tu from University of Michigan, Ann Arbor, USA.

Regression is a widely used statistical tool. A large number of applications consists of learning the association between responses and predictors. From such an association, different tasks, including prediction, can then be conducted. To go beyond simple linear models while maintaining tractability, non-linear mappings can be handled through exploration of local linearity. The non-linear relationship can be captured by a mixture of locally linear regression models as proposed in the so-called Gaussian Locally Linear Mapping (GLLiM) model [6] that assumes Gaussian noise models. In the past year, we have been working on several extensions and applications of GLLiM as described below and the next two subsections.

We proposed a structured mixture model called Hierarchical Locally Linear Mapping (HGLLiM), to predict low-dimensional responses based on high dimensional covariates when the associations between the responses and the covariates are non-linear. For tractability, HGLLiM adopts inverse regression to handle the high dimension and locally-linear mappings to capture potentially non-linear relations. Data with similar associations are grouped together to form a cluster. A mixture is composed of several clusters following a hierarchical structure. This structure enables shared covariance matrices and latent factors across smaller clusters to limit the number of parameters to estimate. Moreover, HGLLiM adopts a robust estimation procedure for model stability. We used three real-world datasets to demonstrate different features of HGLLiM. With the face dataset, HGLLiM shows the ability of modeling non-linear relationship through mixtures. With the orange juice dataset, we show the prediction performance of HGLLiM is robust to the presence of outliers. Moreover, we demonstrated that HGLLiM is capable of handling large-scale complex data using the data acquired from a magnetic resonance vascular fingerprinting (MRvF) study. These examples illustrate the wide applicability of HGLLiM on handling different aspects of a complex data structure in prediction. A preliminary version of this work under revision for JRSS-C can be found in [72].

### *7.1.2. Dictionary-free MR fingerprinting parameter estimation via inverse regression*

**Participants:** Florence Forbes, Fabien Boux, Julyan Arbel.

**Joint work with**: Emmanuel Barbier from Grenoble Institute of Neuroscience.

Magnetic resonance imaging (MRI) can map a wide range of tissue properties but is often limited to observe a single parameter at a time. In order to overcome this problem, Ma et al. introduced magnetic resonance fingerprinting (MRF), a procedure based on a dictionary of simulated couples of signals and parameters. Acquired signals called fingerprints are then matched to the closest signal in the dictionary in order to estimate parameters. This requires an exhaustive search in the dictionary, which even for moderately sized problems, becomes costly and possibly intractable . We propose an alternative approach to estimate more parameters at a time. Instead of an exhaustive search for every signal, we use the dictionary to learn the functional relationship between signals and parameters. This allows the direct estimation of parameters without the need of searching through the dictionary. We investigated the use of GLLiM [6] that bypasses the problems associated with high-to-low regression. The experimental validation of our method is performed in the context of vascular fingerprinting. The comparison between a standard grid search and the proposed approach suggest that MR Fingerprinting could benefit from a regression approach to limit dictionary size and fasten computation time. Preliminary tests and results have been presented at International Society for Magnetic Resonance in Medicine conference, ISMRM 2018 [35].

### 7.1.3. Massive analysis of multi-angular hyperspectral images of the planet Mars by inverse regression of physical models

**Participants:** Florence Forbes, Benoit Kugler.

**Joint work with**: Sylvain Douté from Institut de Planétologie et d'Astrophysique de Grenoble (IPAG).

In the starting PhD of Benoit Kugler, the objective is to develop a statistical learning technique capable of solving a complex inverse problem in planetary remote sensing. The challenges are 1) the large number of observations to to inverse, 2) their large dimension, 3) the need to provide predictions for correlated parameters and 4) the need to provide a quality index (eg. uncertainty). To achieve this goal, we have started to investigate a setting in which a physical model is available to provide simulations that can then be used for learning prior to inversion of real observed data. For the learning step to be as accurate as possible, an initial task is then to estimate the best fit of the theoretical model to the real data. We proposed an iterative procedure based on a combination of GLLiM [6] predictions and importance sampling steps.

### 7.1.4. Quantitative MRI Characterization of Brain Abnormalities in de novo Parkinsonian patients

**Participants:** Florence Forbes, Veronica Munoz Ramirez, Alexis Arnaud, Julyan Arbel.

**Joint work with**: Michel Dojat from Grenoble Institute of Neuroscience.

Currently there is an important delay between the onset of Parkinson's disease and its diagnosis. The detection of changes in physical properties of brain structures may help to detect the disease earlier. In this work, we proposed to take advantage of the informative features provided by quantitative MRI to construct statistical models representing healthy brain tissues. We used mixture models of non Gaussian distributions [8] to capture the non-standard shape of the data multivariate distribution. This allowed us to detect atypical values for these features in the brain of Parkinsonian patients following a procedure similar to that in [16]. Promising preliminary results demonstrate the potential of our approach in discriminating patients from controls and revealing the subcortical structures the most impacted by the disease. This work has been accepted at the IEEE International Symposium on Biological Imaging, ISBI 2019 [36].

### 7.1.5. No structural differences are revealed by voxel-based morphometry in de novo Parkinsonian patients

**Participants:** Florence Forbes, Veronica Munoz Ramirez.

**Joint work with**: Michel Dojat from Grenoble Institute of Neuroscience and Pierrick Coupé from Laboratoire Bordelais de Recherche en Informatique, UMR 5800, Univ. Bordeaux, Talence.

The identification of brain morphological alterations in newly diagnosed PD patients (i.e. de novo) could potentially serve as a biomarker and accelerate diagnosis. However, presently no consensus exists in the literature possibly due to several factors: small size cohorts, differences in segmentation techniques or bad control of false positive rates. In this study, we seek, using the Computational Anatomy Toolbox (CAT12) (University of Jena) pipeline, for morphological brain differences in gray and white matter of 66 controls and 144 de novo PD patients whose data were extracted from the PPMI (Parkinson Progressive Markers Initiative) database. Moreover, we searched for subcortical structure differences using the new online platform VolBrain (J. V. Manjón and P. Coupé, "volBrain: An Online MRI Brain Volumetry System," Front. Neuroinform., vol. 10, p. 30, Jul. 2016). We found no structural brain differences in this de novo Parkinsonian population, neither in tissues using a whole brain analysis nor in any of nine subcortical structures analyzed separately. We concluded that some results published in the literature appear as false positives and are not reproducible.

### 7.1.6. Characterization of daily glycemic variability in the patient with type 1 diabetes

**Participants:** Florence Forbes, Fei Zheng.

**Joint work with**: Stéphane Bonnet from CEA Leti and Pierre-Yves Benhamou, Manon Jalbert from CHU Grenoble Alpes.

Glycemic variability (GV) is an important component of glycemic control in patients with type 1 diabetes. Many metrics have been proposed to account for this variability but none is unanimous among physicians. One difficulty is that the variations in blood sugar levels are expressed very differently from one day to another in some subjects. Our goal was to develop and evaluate the performance of a daily GV index built by combining different known metrics (CV, MAGE, GVP etc). This in order to merge their descriptive power to obtain a more complete and more accurate index. This preliminary study will be presented at the Société Francophone du Diabète (SFD) in 2019 [46].

### *7.1.7. Glycemic variability improves after pancreatic islet transplantation in patients with type 1 diabetes*

**Participants:** Florence Forbes, Fei Zheng.

**Joint work with**: Stéphane Bonnet from CEA Leti and Pierre-Yves Benhamou, Manon Jalbert from CHU Grenoble Alpes.

Glycemic variability (GV) must be taken into account in the efficacy of treatment of type 1 diabetes because it determines the quality of glycemic control, the risk of complication of the patient's disease. Our goal in this study was to describe GV scores in patients with pancreatic islet transplantation (PIT) type 1 diabetes in the TRIMECO trial, and change of thresholds, for each index. predictive of success of PIT.

### *7.1.8. Dirichlet process mixtures under affine transformations of the data*

**Participant:** Julyan Arbel.

**Joint work with**: Riccardo Corradin from Milano Bicocca, Italy and Bernardo Nipoti from Trinity College Dublin, Ireland.

Location-scale Dirichlet process mixtures of Gaussians (DPM-G) have proved extremely useful in dealing with density estimation and clustering problems in a wide range of domains. Motivated by an astronomical application, in this work we address the robustness of DPM-G models to affine transformations of the data, a natural requirement for any sensible statistical method for density estimation. In [57], we first devise a coherent prior specification of the model which makes posterior inference invariant with respect to affine transformation of the data. Second, we formalize the notion of asymptotic robustness under data transformation and show that mild assumptions on the true data generating process are sufficient to ensure that DPM-G models feature such a property. As a by-product, we derive weaker assumptions than those provided in the literature for ensuring posterior consistency of Dirichlet process mixtures, which could reveal of independent interest. Our investigation is supported by an extensive simulation study and illustrated by the analysis of an astronomical dataset consisting of physical measurements of stars in the field of the globular cluster NGC 2419.

### *7.1.9. Applications of mixture models in Industry*

**Participant:** Julyan Arbel.

**Joint work with**: Kerrie Mengersen, Earl Duncan, Clair Alston-Knox and Nicole White.

A very wide range of commonly encountered problems in industry are amenable to statistical mixture modelling and analysis. These include process monitoring or quality control, efficient resource allocation, risk assessment, prediction, and so on. Commonly articulated reasons for adopting a mixture approach include the ability to describe non-standard outcomes and processes, the potential to characterize each of a set of multiple outcomes or processes via the mixture components, the concomitant improvement in interpretability of the results, and the opportunity to make probabilistic inferences such as component membership and overall prediction.

In [51], We illustrate the wide diversity of applications of mixture models to problems in industry, and the potential advantages of these approaches, through a series of case studies.

### 7.1.10. Approximation results regarding the multiple-output mixture of the Gaussian-gated linear experts model

**Participant:** Florence Forbes.

**Joint work with**: Hien Nguyen, La Trobe University Melbourne Australia and Faicel Chamroukhi, Caen University, France.

Mixture of experts (MoE) models are a class of artificial neural networks that can be used for functional approximation and probabilistic modeling. An important class of MoE models is the class of mixture of linear experts (MoLE) models, where the expert functions map to real topological output spaces. Recently, Gaussian-gated MoLE models have become popular in applied research. There are a number of powerful approximation results regarding Gaussian-gated MoLE models, when the output space is univariate. These results guarantee the ability of Gaussian-gated MoLE mean functions to approximate arbitrary continuous functions, and Gaussian-gated MoLE models themselves to approximate arbitrary conditional probability density functions. We utilized and extended upon the univariate approximation results in order to prove a pair of useful results for situations where the output spaces are multivariate. We do this by proving a pair of lemmas regarding the combination of univariate MoLE models, which are interesting in their own rights.

### 7.1.11. Models for ranking data

**Participant:** Marta Crispino.

within the BigInsight project, Oslo.

We developed a new method and algorithms for working with ranking data. This kind of data is particularly relevant in applications involving personalized recommendations. In particular, we have invented a new Bayesian approach based on extensions of the Mallows model, which allows making personalized recommendations equipped with a level of uncertainty.

The Mallows model (MM) is a popular parametric family of models for ranking data, based on the assumption that a modal ranking, which can be interpreted as the consensus ranking of the population, exists. The probability of observing a given ranking is then assumed to decay exponentially fast as its distance from the consensus grows. The MM is therefore a two-parameter distance-based family of models. The scale or precision parameter, controlling the concentration of the distribution determines the rate of decay of the probability of individual ranks. Individual models with different properties can be obtained depending on the choice of distance on the space of permutations. A major drawback of the MM is that its computational complexity has limited its use to a particular form based on Kendall distance. We develop new computationally tractable methods for Bayesian inference in Mallows models that work with any right-invariant distance. Our method performs inference on the consensus ranking of the items, also when based on partial rankings, such as top-k items or pairwise comparisons. When assessors are many or heterogeneous, we propose a mixture model for clustering them in homogeneous subgroups, with cluster specific consensus rankings. We develop approximate stochastic algorithms that allow a fully probabilistic analysis, leading to coherent quantifications of uncertainties, make probabilistic predictions on the class membership of assessors based on their ranking of just some items, and predict missing individual preferences, as needed in recommendation systems. The methodology has been published in the Journal of Machine Learning Research, JMLR, in early 2018.

A generalization of the model above involves dealing with non-transitive and heterogeneous pairwise comparison data, coming from an experiment within the musicology domain. We thus develop a mixture model extension of the Bayesian Mallows model able to handle non-transitive data, with a latent layer of uncertainty which captures the generation of preference misreporting. This paper was recently accepted for publication in the Annals of Applied Statistics, AoAS.

Within this project, we also write a survey paper, whose main goal is to compare the performance of our method with other existing methodologies, including the Plackett-Luce, the Bradley-Terry, the collaborative filtering methods, and some of their variations. We illustrate and discuss the use of these models by means of an experiment in which assessors rank potatoes, and with a simulation. The purpose of this paper is not to recommend the use of one best method, but to present a palette of different possibilities for different questions and different types of data. This was recently accepted on the Annual Review of Statistics and Its Applications, ARSIA.

## 7.2. Semi and non-parametric methods

### *7.2.1. Estimation of extreme risk measures*

**Participant:** Stéphane Girard.

**Joint work with:** A. Daouia (Univ. Toulouse), L. Gardes (Univ. Strasbourg), J. Elmethni (Univ. Paris 5) and G. Stupfler (Univ. Nottingham, UK).

One of the most popular risk measures is the Value-at-Risk (VaR) introduced in the 1990's. In statistical terms, the VaR at level $\alpha \in (0,1)$ corresponds to the upper $\alpha$-quantile of the loss distribution. The Value-at-Risk however suffers from several weaknesses. First, it provides us only with a pointwise information: VaR($\alpha$) does not take into consideration what the loss will be beyond this quantile. Second, random loss variables with light-tailed distributions or heavy-tailed distributions may have the same Value-at-Risk. Finally, Value-at-Risk is not a coherent risk measure since it is not subadditive in general. A first coherent alternative risk measure is the Conditional Tail Expectation (CTE), also known as Tail-Value-at-Risk, Tail Conditional Expectation or Expected Shortfall in case of a continuous loss distribution. The CTE is defined as the expected loss given that the loss lies above the upper $\alpha$-quantile of the loss distribution. This risk measure thus takes into account the whole information contained in the upper tail of the distribution. In [20], we investigate the extreme properties of a new risk measure (called the Conditional Tail Moment) which encompasses various risk measures, such as the CTE, as particular cases. We study the situation where some covariate information is available under some general conditions on the distribution tail. We thus has to deal with conditional extremes. However, the asymptotic normality of the empirical CTE estimator requires that the underlying distribution possess a finite variance; this can be a strong restriction in heavy-tailed models which constitute the favoured class of models in actuarial and financial applications. One possible solution in very heavy-tailed models where this assumption fails could be to use the more robust Median Shortfall, but this quantity is actually just a quantile, which therefore only gives information about the frequency of a tail event and not about its typical magnitude. In [65], we construct a synthetic class of tail $L_p$ -medians, which encompasses the Median Shortfall (for $p = 1$) and Conditional Tail Expectation (for $p = 2$). We show that, for $1 < p < 2$, a tail $L_p$ -median always takes into account both the frequency and magnitude of tail events, and its empirical estimator is, within the range of the data, asymptotically normal under a condition weaker than a finite variance. We extrapolate this estimator, along with another technique, to proper extreme levels using the heavy-tailed framework. The estimators are showcased on a simulation study and on a set of real fire insurance data showing evidence of a very heavy right tail.

A possible coherent alternative risk measure is based on expectiles [18], [63], [62]. Compared to quantiles, the family of expectiles is based on squared rather than absolute error loss minimization. The flexibility and virtues of these least squares analogues of quantiles are now well established in actuarial science, econometrics and statistical finance. Both quantiles and expectiles were embedded in the more general class of M-quantiles [19] as the minimizers of a generic asymmetric convex loss function. It has been proved very recently that the only M-quantiles that are coherent risk measures are the expectiles.

### *7.2.2. Extrapolation limits associated with extreme-value methods*

**Participants:** Clément Albert, Stéphane Girard.

**Joint work with:** L. Gardes (Univ. Strasbourg) and A. Dutfoy (EDF R&D).

The PhD thesis of Clément Albert (co-funded by EDF) is dedicated to the study of the sensitivity of extreme-value methods to small changes in the data and to their extrapolation ability. Two directions are explored:

(i) In [54], we investigate the asymptotic behavior of the (relative) extrapolation error associated with some estimators of extreme quantiles based on extreme-value theory. It is shown that the extrapolation error can be interpreted as the remainder of a first order Taylor expansion. Necessary and sufficient conditions are then provided such that this error tends to zero as the sample size increases. Interestingly, in case of the so-called Exponential Tail estimator, these conditions lead to a subdivision of Gumbel maximum domain of attraction into three subsets. In constrast, the extrapolation error associated with Weissman estimator has a common behavior over the whole Fréchet maximum domain of attraction. First order equivalents of the extrapolation error are then derived and their accuracy is illustrated numerically.

(ii) In [53], We propose a new estimator for extreme quantiles under the log-generalized Weibull-tail model, introduced by Cees de Valk. This model relies on a new regular variation condition which, in some situations, permits to extrapolate further into the tails than the classical assumption in extreme-value theory. The asymptotic normality of the estimator is established and its finite sample properties are illustrated both on simulated and real datasets.

### *7.2.3. Estimation of local intrinsic dimensionality with extreme-value methods*

**Participant:** Stéphane Girard.

**Joint work with:** L. Amsaleg (LinkMedia, Inria Rennes), O. Chelly (NII Japon), T. Furon (LinkMedia, Inria Rennes), M. Houle (NII Japon), K.-I. Kawarabayashi (NII Japon), M. Nett (Google).

This work is concerned with the estimation of a local measure of intrinsic dimensionality (ID). The local model can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the distribution of distances to a query point is modeled in terms of a continuous random variable. This form of intrinsic dimensionality can be particularly useful in search, classification, outlier detection, and other contexts in machine learning, databases, and data mining, as it has been shown to be equivalent to a measure of the discriminative power of similarity functions. In [14], several estimators of local ID are proposed and analyzed based on extreme value theory, using maximum likelihood estimation, the method of moments, probability weighted moments, and regularly varying functions. An experimental evaluation is also provided, using both real and artificial data.

### *7.2.4. Bayesian inference for copulas*

**Participants:** Julyan Arbel, Marta Crispino, Stéphane Girard.

We study in [58] a broad class of asymmetric copulas known as Liebscher copulas and defined as a combination of multiple—usually symmetric—copulas. The main thrust of this work is to provide new theoretical properties including exact tail dependence expressions and stability properties. A subclass of Liebscher copulas obtained by combining Fréchet copulas is studied in more details. We establish further dependence properties for copulas of this class and show that they are characterized by an arbitrary number of singular components. Furthermore, we introduce a novel iterative construction for general Liebscher copulas which *de facto* insures uniform margins, thus relaxing a constraint of Liebscher's original construction. Besides, we show that this iterative construction proves useful for inference by developing an Approximate Bayesian computation sampling scheme. This inferential procedure is demonstrated on simulated data.

In [22], we investigate the properties of a new transformation of copulas based on the co-copula and an univariate function. It is shown that several families in the copula literature can be interpreted as particular outputs of this transformation. Symmetry, association, ordering and dependence properties of the resulting copula are established.

### *7.2.5. Bayesian nonparametric clustering*

**Participant:** Julyan Arbel.

**Joint work with**: Riccardo Corradin from Milano Bicocca, Michal Lewandowski from Bocconi University, Milan, Italy, Caroline Lawless from Université Paris-Dauphine, France.

For a long time, the Dirichlet process has been the gold standard discrete random measure in Bayesian nonparametrics. The Pitman–Yor process provides a simple and mathematically tractable generalization, allowing for a very flexible control of the clustering behaviour. Two commonly used representations of the Pitman–Yor process are the stick-breaking process and the Chinese restaurant process. The former is a constructive representation of the process which turns out very handy for practical implementation, while the latter describes the partition distribution induced. Obtaining one from the other is usually done indirectly with use of measure theory. In contrast, we propose in [66] an elementary proof of Pitman–Yor's Chinese Restaurant process from its stick-breaking representation.

In the discussion paper [56], we propose a simulation study to emphasise the difference between Variation of Information and Binder's loss functions in terms of number of clusters estimated by means of the use of the Markov chain Monte Carlo output only and a "greedy" method.

The chapter [47] is part of a book edited by Stéphane Girard and Julyan Arbel. It presents a Bayesian nonparametric approach to clustering, which is particularly relevant when the number of components in the clustering is unknown. The approach is illustrated with the Milky Way's globulars, that are clouds of stars orbiting in our galaxy. Clustering globulars is key for better understanding the Milky Way's history. We define the Dirichlet process and illustrate some alternative definitions such as the Chinese restaurant process, the Pólya Urn, the Ewens sampling formula, the stick-breaking representation through some simple *R* code. The Dirichlet process mixture model is presented, as well as the *R* package *BNPmix* implementing Markov chain Monte Carlo sampling. Inference for the clustering is done with the variation of information loss function.

### *7.2.6. Multi sensor fusion for acoustic surveillance and monitoring*

**Participants:** Florence Forbes, Jean-Michel Bécu.

**Joint work with**: Pascal Vouagner and Christophe Thirard from ACOEM company.

In the context of the DGA-rapid WIFUZ project, we addressed the issue of determining the localization of shots from multiple measurements coming from multiple sensors. The WIFUZ project is a collaborative work between various partners: DGA, ACOEM and HIKOB companies and Inria. This project is at the intersection of data fusion, statistics, machine learning and acoustic signal processing. The general context is the surveillance and monitoring of a zone acoustic state from data acquired at a continuous rate by a set of sensors that are potentially mobile and of different nature. The overall objective is to develop a prototype for surveillance and monitoring that is able to combine multi sensor data coming from acoustic sensors (microphones and antennas) and optical sensors (infrared cameras) and to distribute the processing to multiple algorithmic blocs. As an illustration, the MISTIS contribution is to develop technical and scientific solutions as part of a collaborative protection approach, ideally used to guide the best coordinated response between the different vehicles of a military convoy. Indeed, in the case of an attack on a convoy, identifying the threatened vehicles and the origin of the threat is necessary to organize the best response from all members on the convoy. Thus it will be possible to react to the first contact (emergency detection) to provide the best answer for threatened vehicles (escape, lure) and for those not threatened (suppression fire, riposte fire). We developed statistical tools that make it possible to analyze this information (characterization of the threat) using fusion of acoustic and image data from a set of sensors located on various vehicles. We used Bayesian inversion and simulation techniques to recover multiple sources mimicking collaborative interaction between several vehicles.

### *7.2.7. Extraction and data analysis toward "industry of the future"*

**Participants:** Florence Forbes, Hongliang Lu, Fatima Fofana, Jaime Eduardo Arias Almeida.

**Joint work with**: J. F. Cuccaro and J. C Trochet from Vi-Technology company.

Industry as we know it today will soon disappear. In the future, the machines which constitute the manufacturing process will communicate automatically as to optimize its performance as whole. Transmitted information essentially will be of statistical nature. In the context of VISION 4.0 project with Vi-Technology, the role of MISTIS is to identify what statistical methods might be useful for the printed circuits boards assembly industry. The topic of F. Fofana's internship was to extract and analyze data from two inspection machines of a industrial process making electronic cards. After a first extraction step in the SQL database, the goal was to enlighten the statistical links between these machines. Preliminary experiments and results on the Solder Paste Inspection (SPI) step, at the beginning of the line, helped identifying potentially relevant variables and measurements (eg related to stencil offsets) to identify future defects and discriminate between them. More generally, we have access to two databases at both ends (SPI and Component Inspection) of the assembly process. The goal is to improve our understanding of interactions in the assembly process, find out correlations between defects and physical measures, generate proactive alarms so as to detect departures from normality.

### *7.2.8. Change point detection for the analysis of dynamic single molecules*

**Participants:** Florence Forbes, Theo Moins.

**Joint work with**: Virginie Stoppin-Mellet from Grenoble Institute of Neuroscience.

The objective of this study was to develop a statistical learning technique to analyze signals produced by molecules. The main difficulties are the noisy nature of the signals and the definition of a quality index to allow the elimination of poor-quality data and false positive signals. In collaboration with the GIN, we addressed the statistical analysis of intensity traces (2 month internship of Theo Moins, Ensimag 2A). Namely, the ImageJ Thunderstorm toolbox, which has been developed for the detection of single molecule in super resolution imaging, has been successfully used to detect immobile single molecules and generate time-dependent intensity traces. Then the R package Segmentor3IsBack, a fast segmentation algorithm based on 5 possible statistical models, proved efficient in the processing of the noisy intensity traces. This preliminary study led to a multidisciplinary project funded by the Grenoble data institute for 2 years in which we will also address additional challenges for the tracking of a large population of single molecules.

## 7.3. Graphical and Markov models

### *7.3.1. Fast Bayesian network structure learning using quasi-determinism screening*

**Participants:** Thibaud Rahier, Stéphane Girard, Florence Forbes.

**Joint work with:** Sylvain Marié, Schneider Electric.

Learning the structure of Bayesian networks from data is a NP-Hard problem that involves an optimization task on a super-exponential sized space. In this work, we show that in most real life datasets, a number of the arcs contained in the final structure can be prescreened at low computational cost with a limited impact on the global graph score. We formalize the identification of these arcs via the notion of quasi-determinism, and propose an associated algorithm that reduces the structure learning to a subset of the original variables. We show, on diverse benchmark datasets, that this algorithm exhibits a significant decrease in computational time and complexity for only a little decrease in performance score. A first version of this work can be found in [71] and has been presented at the JFRB 2018 workshop [41].

### *7.3.2. Robust structure learning using multivariate t-distributions*

**Participants:** Karina Ashurbekova, Florence Forbes.

**Joint work with:** Sophie Achard, senior researcher at CNRS, Gipsa-lab.

Structure learning is an active topic nowadays in different application areas, i.e. genetics, neuroscience. We addressed the issue of robust graph structure learning in continuous settings. We focused on sparse precision matrix estimation for its tractability and ability to reveal some measure of dependence between variables. For this purpose, we proposed to extract good features from existing methods, namely *tlasso* and CLIME procedures. The former is based on the observation that standard Gaussian modelling results in procedures that are too sensitive to outliers and proposes the use of $t$-distributions as an alternative. The latter is an alternative to the popular Lasso optimization principle which can handle some of its limitations. We then combined these ideas into a new procedure referred to as tCLIME that can be seen as a modified *tlasso* algorithm. Numerical performance was investigated using simulated data and reveals that tCLIME performs favorably compared to the other standard methods. This work was presented at the Journées de Statistiques de la Société Francaise de Statistique in Saclay, 2018, [39].

### *7.3.3. Structure learning via Hadamard product of correlation and partial correlation matrices*

**Participants:** Karina Ashurbekova, Florence Forbes.

**Joint work with:** Sophie Achard, senior researcher at CNRS, Gipsa-lab.

Classical conditional independences or marginal independences may not be sufficient to express complex relationships. In this work we introduced a new structure learning procedure where an edge in the graph corresponds to a non zero of both correlation and partial correlation. A theoretical study was derived which shows the good properties of the proposed graph estimator, illustrated also on a synthetic example.

### *7.3.4. Spatial mixtures of multiple scaled $t$-distributions*

**Participants:** Florence Forbes, Alexis Arnaud.

**Joint work with:** Steven Quinito Masnada, Inria Grenoble Rhone-Alpes

The goal is to implement an hidden Markov model version of our recently introduced mixtures of non standard multiple scaled $t$-distributions. The motivation for doing that is the application to multiparametric MRI data for lesion analysis. When dealing with MRI human data, spatial information is of primary importance. For our preliminary study on rat data [16], the results without spatial information were already quite smooth. The main anatomical structures can be identified. We suspect the reason is that the measured parameters already contain a lot of information about the underlying tissues. However, introducing spatial information is always useful and is our ongoing work. In the statistical framework we have developed (mixture models and EM algorithm), it is conceptually straightforward to introduce an additional Markov random field. In addition, when using a Markov random field it is easy to incorporate additional atlas information.

### *7.3.5. Spectral CT reconstruction with an explicit photon-counting detector model: a "one-step" approach*

**Participants:** Florence Forbes, Pierre-Antoine Rodesch.

**Joint work with:** Veronique Rebuffel and Clarisse Fournier from CEA-LETI Grenoble.

In the context of Pierre-Antoine Rodesh's PhD thesis, we investigate new statistical and optimization methods for tomographic reconstruction from non standard detectors providing multiple energy signals. Recent developments in energy-discriminating Photon-Counting Detector (PCD) enable new horizons for spectral CT. With PCDs, new reconstruction methods take advantage of the spectral information measured through energy measurement bins. However PCDs have serious spectral distortion issues due to charge-sharing, fluorescence escape, pileup effect Spectral CT with PCDs can be decomposed into two problems: a noisy geometric inversion problem (as in standard CT) and an additional PCD spectral degradation problem. The aim of this study is to introduce a reconstruction method which solves both problems simultaneously: a one-step approach. An explicit linear detector model is used and characterized by a Detector Response Matrix (DRM). The algorithm reconstructs two basis material maps from energy-window transmission data. The results prove that the simultaneous inversion of both problems is well performed for simulation data. For comparison, we also perform a standard two-step approach: an advanced polynomial decomposition of measured sinograms combined with a filtered-back projection reconstruction. The results demonstrate the potential uses of this method for medical imaging or for non-destructive control in industry. Preliminary results have been presented at the SPIE medical imaging 2018 conference in Houston, USA [37].

### *7.3.6. Non parametric Bayesian priors for hidden Markov random fields*

**Participants:** Florence Forbes, Julyan Arbel, Hongliang Lu.

Hidden Markov random field (HMRF) models are widely used for image segmentation or more generally for clustering data under spatial constraints. They can be seen as spatial extensions of independent mixture models. As for standard mixtures, one concern is the automatic selection of the proper number of components in the mixture, or equivalently the number of states in the hidden Markov field. A number of criteria exist to select this number automatically based on penalized likelihood (eg. AIC, BIC, ICL etc.) but they usually require to run several models for different number of classes to choose the best one. Other techniques (eg. reversible jump) use a fully Bayesian setting including a prior on the class number but at the cost of prohibitive computational times. In this work, we investigate alternatives based on the more recent field of Bayesian nonparametrics. In particular, Dirichlet process mixture models (DPMM) have emerged as promising candidates for clustering applications where the number of clusters is unknown. Most applications of DPMM involve observations which are supposed to be independent. For more complex tasks such as unsupervised image segmentation with spatial relationships or dependencies between the observations, DPMM are not satisfying. This work has been presented at the Joint Statistical Meeting in Vancouver Canada [29] and at the Journées de la Statistique in Saclay [40].

### *7.3.7. Hidden Markov models for the analysis of eye movements*

**Participants:** Jean-Baptiste Durand, Brice Olivier.

*This research theme is supported by a LabEx PERSYVAL-Lab project-team grant.*
**Joint work with**: Anne Guérin-Dugué (GIPSA-lab) and Benoit Lemaire (Laboratoire de Psychologie et Neurocognition)

In the last years, GIPSA-lab has developed computational models of information search in web-like materials, using data from both eye-tracking and electroencephalograms (EEGs). These data were obtained from experiments, in which subjects had to decide whether a text was related or not to a target topic presented to them beforehand. In such tasks, reading process and decision making are closely related. Statistical analysis of such data aims at deciphering underlying dependency structures in these processes. Hidden Markov models (HMMs) have been used on eye movement series to infer phases in the reading process that can be interpreted as steps in the cognitive processes leading to decision. In HMMs, each phase is associated with a state of the Markov chain. The states are observed indirectly though eye-movements. Our approach was inspired by Simola et al. (2008), but we used hidden semi-Markov models for better characterization of phase length distributions [80]. The estimated HMM highlighted contrasted reading strategies (ie, state transitions), with both individual and document-related variability. However, the characteristics of eye movements within each phase tended to be poorly discriminated. As a result, high uncertainty in the phase changes arose, and it could be difficult to relate phases to known patterns in EEGs.

This is why, as part of Brice Olivier's PhD thesis, we are developed integrated models coupling EEG and eye movements within one single HMM for better identification of the phases. Here, the coupling incorporates some delay between the transitions in both (EEG and eye-movement) chains, since EEG patterns associated to cognitive processes occur lately with respect to eye-movement phases. Moreover, EEGs and scanpaths were recorded with different time resolutions, so that some resampling scheme had to be added into the model, for the sake of synchronizing both processes. An associated EM algorithm for maximum likelihood parameter estimation was derived.

New results were obtained in the standalone analysis of the eye-movements. A comparison between the effects of three types of texts was performed, considering texts either closely related, moderately related or unrelated to the target topic.

Our goal for this coming year is to implement and validate our coupled model for jointly analyzing eye-movements and EEGs in order to improve the discrimination of the reading strategies.

### *7.3.8. Lossy compression of tree structures*

**Participant:** Jean-Baptiste Durand.

**Joint work with**: Christophe Godin and Romain Azaïs (Inria Mosaic)

The class of self-nested trees presents remarkable compression properties because of the systematic repetition of subtrees in their structure. The aim of our work is to achieve compression of any unordered tree by finding the nearest self-nested tree. Solving this optmization problem without more assumptions is conjectured to be an NP-complete or NP-hard problem. In [34], we firstly provided a better combinatorial characterization of this specific family of trees. In particular, we showed from both theoretical and practical viewpoints that complex queries can be quickly answered in self-nested trees compared to general trees. We also presented an approximation algorithm of a tree by a self-nested one that can be used in fast prediction of edit distance between two trees.

Our goal for this coming year is to apply this approach to quantify the degree of self-nestedness of several plant species and extend first results obtained on rice panicles stating that near self-nestedness is a fairly general pattern in plants.

### *7.3.9. Relations between structural characteristics in rose bush and visual sensory attributes for objective evaluation of the visual quality*

**Participant:** Jean-Baptiste Durand.

**Joint work with**: Gilles Galopin (QUASAV, Agrocampus Ouest)

Within ornamental horticulture context, visual quality of plants is a critical criterion for consumers looking for immediate decorative effect products. Studying links between architecture and its phenotypic plasticity in response to growing conditions and the resulting plant visual appearance represents an interesting lever to propose a new approach for managing product quality from specialized crops. Objectives of the present study were to determine whether architectural components may be identified across different growing conditions (1) to study the architectural development of a shrub over time; and (2) to predict sensory attributes data characterizing multiple visual traits of the plants. The approach addressed in this study stands on the sensory profile method using a recurrent blooming modern rose bush presented in rotation using video stimuli. Plants were cultivated under a shading gradient in three distinct environments (natural conditions, under 55% and 75% shading nets). Architecture and video of the plants were recorded during three stages, from 5 to 15 months after plant multiplication. Predictive models of visual quality were obtained with regression and variable transformation to encompass non-linear relationships [21]. The proposed approach is a way to gain a better insight into the architecture of shrub plants together with their visual appearance to target processes of interest in order to optimize growing conditions or select the most fitting genotypes across breeding programs, with respect to contrasted consumer preferences.

As a perspective, dynamic traits issued from hidden-Markov-based growth models should be used for a better characterization of visual quality, as well as identification of reiterated complexes, which are believed to play a major role in rose bush structure.

### *7.3.10. Bayesian neural networks*

**Participants:** Julyan Arbel, Mariia Vladimirova.

**Joint work with**: Pablo Mesejo from University of Granada, Spain.

We investigate in [45] and [44] deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust of the paper is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularisation scheme at the level of the units. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

**MOEX Project-Team**

# 4. New Results

## 4.1. Cultural knowledge evolution

Our cultural knowledge evolution work currently focusses on alignment evolution.

Agents may use ontology alignments to communicate when they represent knowledge with different ontologies: alignments help reclassifying objects from one ontology to the other. Such alignments may be provided by dedicated algorithms [7], but their accuracy is far from satisfying. Yet agents have to proceed. They can take advantage of their experience in order to evolve alignments: upon communication failure, they will adapt the alignments to avoid reproducing the same mistake.

We performed such repair experiments [2] and revealed that, by playing simple interaction games, agents can effectively repair random networks of ontologies or even create new alignments.

### *4.1.1. Strengthening modality for cultural alignment repair*

**Participants:** Jérôme Euzenat [Correspondent], Iris Lohja.

Our previous work on cultural alignment repair achieved 100% precision for all adaptation operators, i.e., all the correspondences in the alignments were correct, but were still missing some correspondences, and did not achieve 100% recall. We had conjectured that this was due to a phenomenon called reverse shadowing [2], avoiding to find specific correspondences.

This year we introduced a new adaptation modality, strengthening, to test this hypothesis. The strengthening modality replaces a successful correspondence by one of its subsumed correspondences covering the current instance. This modality is different from those developed so far, because it leads agents to adapt their alignment when the game played has been a success (previously, it was always when a failure occurred). We defined three alternative definitions of this modality depending on if the agent chooses the most general, most specific or a random such correspondence.

The strengthening modality has been implemented in our *Lazy lavender* software. We experimentally showed that it was not interferring with the other modalities as soon as the *add* operator was used. This means that all properties of the previous adaptation operators are preserved. Moreover, as expected, recall was greatly increased, to the point that some operators achieve 99% F-measure. However, the agents still do not reach 100% recall.

### *4.1.2. Experiment reproducibility through container technology*

**Participants:** Jérôme Euzenat [Correspondent], Bilal Lahmami.

Performing experiments and reporting them requires care in order for others to be able to repeat them.

We experimented with container technology in order to embed our experiments and offer to others to run them easily. To that extent, we developed scripts associated to the *Lazy lavender* software to specify, run, and analyse experiments. In particular, these scripts are able to generate a Docker container specification that can perform experiments in the same conditions or with updated software. The documentation of the experiments on our Wiki platform (https://gforge.inria.fr/plugins/mediawiki/wiki/lazylav/index.php/Lazy_Lavender) is also eased by this process.

## 4.2. Link keys

Link keys (§3.2 ) are explored following two directions:

- Extracting link keys;
- Reasoning with link keys.

### 4.2.1. Link key extraction with relational concept analysis

**Participants:** Manuel Atencia, Jérôme David [Correspondent], Jérôme Euzenat.

We have further investigated link key extraction using relational concept analysis and the associated prototype implementation [8]. In particular, we showed that that link keys extracted by formal concept analysis are equivalent to an extension of those which were extracted by our former algorithm [1]

### 4.2.2. Link key extraction under ontological constraints

**Participants:** Jérôme David [Correspondent], Jérôme Euzenat, Khadija Jradeh.

We investigated the use of link keys taking advantage of ontologies. This can be carried out in two different directions: exploiting the ontologies under which data sets are published, and extracting link keys using ontology constructors for combining attribute and class names.

Following the first approach, we extended our existing algorithms to extract link keys involving inverse ($^{-1}$), union ($\sqcup$), intersection ($\sqcap$) and paths ($\circ$) of properties. This helps providing link keys when it is not possible otherwise (without inverse, there is no possible correspondence if one data set is using parents and the other is using children). We showed how the paths could be normalised to reduce the search space. Extracting link keys under these conditions required to introduce better indexing techniques to avoid unnecessary link key generation and even looping.

We implemented this method and evaluated it by running experiments on two real data sets, this resulted in finding the correct link keys that were not found without them.

### 4.2.3. Tableau method for $\mathcal{ALC}$+Link key reasoning

**Participants:** Manuel Atencia [Correspondent], Jérôme Euzenat, Khadija Jradeh.

Link keys can also be thought of as axioms in a description logic. We further worked on the tableau method designed for the $\mathcal{ALC}$ description logic to support reasoning with link keys.

## 4.3. Semantic web queries

### 4.3.1. Evaluation of query transformations without data

**Participants:** Jérôme David, Jérôme Euzenat [Correspondent].

Query transformations are ubiquitous in semantic web query processing. For any situation in which transformations are not proved correct by construction, the quality of these transformations has to be evaluated. Usual evaluation measures are either overly syntactic and not very informative —the result being: correct or incorrect— or dependent from the evaluation sources. Moreover, both approaches do not necessarily yield the same result. We proposed to ground the evaluation on query containment [4]. This allows for a data-independent evaluation that is more informative than the usual syntactic evaluation. In addition, such evaluation modalities may take into account ontologies, alignments or different query languages as soon as they are relevant to query evaluation [6].

**MORPHEO Project-Team**

# 7. New Results

## 7.1. Surface Motion Capture Animation Synthesis

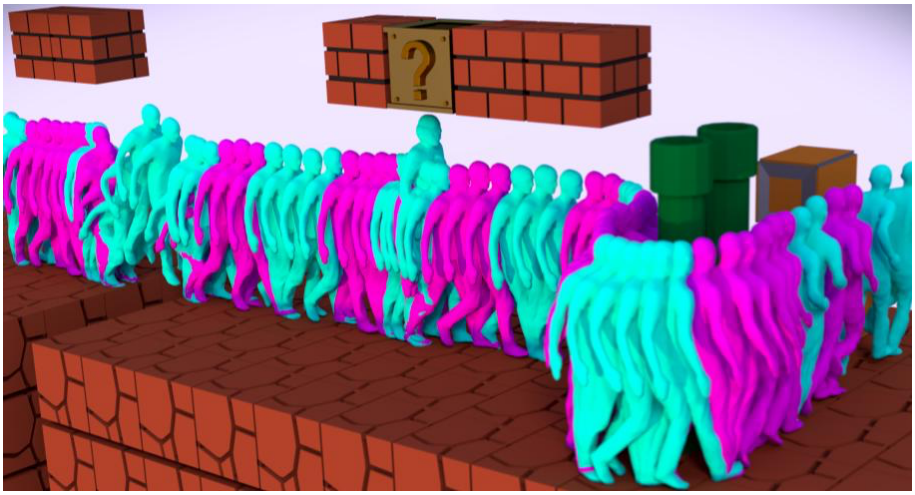


*Figure 3.*

We propose to generate novel animations from a set of elementary examples of video-based surface motion capture, under user-specified constraints. 4D surface capture animation is motivated by the increasing demand from media production for highly realistic 3D content. To this aim, data driven strategies that consider video-based information can produce animation with real shapes, kinematics and appearances. Our animations rely on the combination and the interpolation of textured 3D mesh data, which requires examining two aspects: (1) Shape geometry and (2) appearance. First, we propose an animation synthesis structure for the shape geometry, the Essential graph, that outperforms standard Motion graphs in optimality with respect to quantitative criteria, and we extend optimized interpolated transition algorithms to mesh data. Second, we propose a compact view-independent representation for the shape appearance. This representation encodes subject appearance changes due to viewpoint and illumination, and due to inaccuracies in geometric modelling independently. Besides providing compact representations, such decompositions allow for additional applications such as interpolation for animation (see figure 3 ).

This result was published in a prominent computer graphics journal, IEEE Transactions on Visualization and Computer Graphics [2].

## 7.2. A Multilinear Tongue Model Derived from Speech Related MRI Data of the Human Vocal Tract

We present a multilinear statistical model of the human tongue that captures anatomical and tongue pose related shape variations separately. The model is derived from 3D magnetic resonance imaging data of 11 speakers sustaining speech related vocal tract configurations. To extract model parameters, we use a minimally supervised method based on an image segmentation approach and a template fitting technique. Furthermore,

we use image denoising to deal with possibly corrupt data, palate surface information reconstruction to handle palatal tongue contacts, and a bootstrap strategy to refine the obtained shapes. Our evaluation shows that, by limiting the degrees of freedom for the anatomical and speech related variations, to 5 and 4, respectively, we obtain a model that can reliably register unknown data while avoiding overfitting effects. Furthermore, we show that it can be used to generate plausible tongue animation by tracking sparse motion capture data.

This result was published in Computer Speech and Language 51 [3].

## 7.3. CBCT of a Moving Sample from X-rays and Multiple Videos

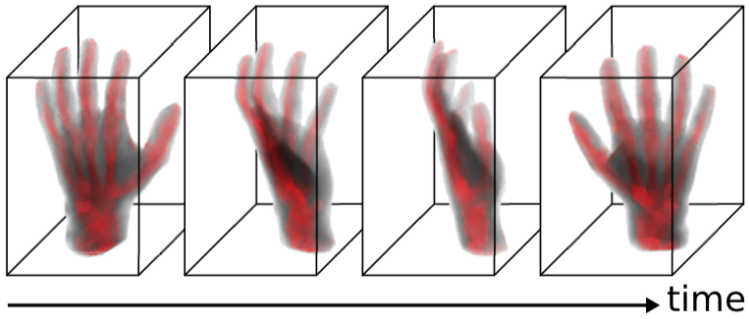



*Figure 4. Dense volumetric attenuation reconstruction from a rigidly moving sample captured by a single planar X-ray imaging device and a surface motion capture system. Higher attenuation (here bone structure) is highlighted in red.*

We consider dense volumetric modeling of moving samples such as body parts. Most dense modeling methods consider samples observed with a moving X-ray device and cannot easily handle moving samples. We propose instead a novel method to observe shape motion from a fixed X-ray device and to build dense in-depth attenuation information. This yields a low-cost, low-dose 3D imaging solution, taking benefit of equipment widely available in clinical environments. Our first innovation is to combine a video-based surface motion capture system with a single low-cost/low-dose fixed planar X-ray device, in order to retrieve the sample motion and attenuation information with minimal radiation exposure. Our second innovation is to rely on Bayesian inference to solve for a dense attenuation volume given planar radioscopic images of a moving sample. This approach enables multiple sources of noise to be considered and takes advantage of very limited prior information to solve an otherwise ill-posed problem. Results show that the proposed strategy is able to reconstruct dense volumetric attenuation models from a very limited number of radiographic views over time on synthetic and in-situ data, as illustrated in Figure 4 .

This result was published in a prominent medical journal, IEEE Transactions on Medical Imaging [4].

## 7.4. Automatic camera calibration using multiple sets of pairwise correspondences

We propose a new method to add an uncalibrated node into a network of calibrated cameras using only pairwise point correspondences (see figure 5 ). While previous methods perform this task using triple correspondences, these are often difficult to establish when there is limited overlap between different views. In such challenging cases we must rely on pairwise correspondences and our solution becomes more advantageous. Our method includes an 11-point minimal solution for the intrinsic and extrinsic calibration of a camera from pairwise correspondences with other two calibrated cameras, and a new inlier selection framework that extends the

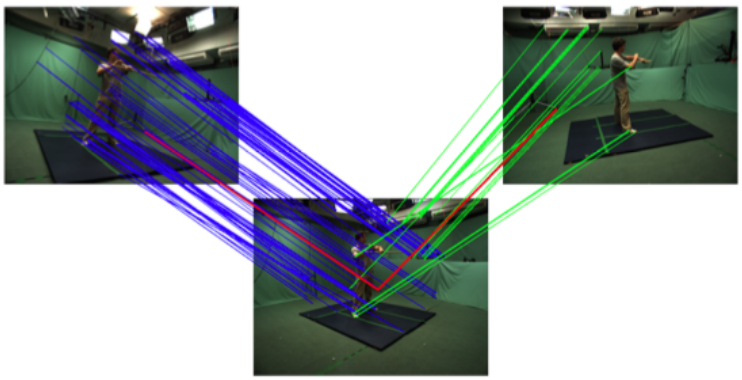


*Figure 5. Correspondences extracted from SIFT features. Given the wide baseline between the views there is a single reliable triple correspondence (red) while there are many reliable pairwise correspondences (blue and green).*

traditional RANSAC family of algorithms to sampling across multiple datasets. Our method is validated on different application scenarios where a lack of triple correspondences might occur: addition of a new node to a camera network; calibration and motion estimation of a moving camera inside a camera network; and addition of views with limited overlap to a Structure-from-Motion model.

This result was published in a prominent medical journal, IEEE Transactions on Pattern Analysis and Machine Intelligence [5].

## 7.5. Multilinear Autoencoder for 3D Face Model Learning

*Figure 6. Shape variations caused by different expressions of the same subject.*

Generative models have proved to be useful tools to represent 3D human faces and their statistical variations (see figure 6 ). With the increase of 3D scan databases available for training, a growing challenge lies in the ability to learn generative face models that effectively encode shape variations with respect to desired attributes, such as identity and expression, given datasets that can be diverse. This paper addresses this challenge by proposing a framework that learns a generative 3D face model using an autoencoder architecture, allowing hence for weakly supervised training. The main contribution is to combine a convolutional neural network-based en-coder with a multilinear model-based decoder, taking therefore advantage of both the convolutional network robustness to corrupted and incomplete data, and of the multilinear model capacity to effectively model and decouple shape variations. Given a set of 3D face scans with annotation labels for

the desired attributes, e.g. identities and expressions, our method learns an expressive multilinear model that decouples shape changes due to the different factors. Experimental results demonstrate that the proposed method outperforms recent approaches when learning multilinear face models from incomplete training data, particularly in terms of space decoupling, and that it is capable of learning from an order of magnitude more data than previous methods.

This result was published in IEEE Winter Conference on Applications of Computer Vision [6].

## 7.6. Spatiotemporal Modeling for Efficient Registration of Dynamic 3D Faces

We consider the registration of temporal sequences of 3D face scans. Face registration plays a central role in face analysis applications, for instance recognition or transfer tasks, among others. We propose an automatic approach that can register large sets of dynamic face scans without the need for landmarks or highly specialized acquisition setups. This allows for extended versatility among registered face shapes and deformations by enabling to leverage multiple datasets, a fundamental property when e.g. building statistical face models. Our approach is built upon a regression-based static registration method, which is improved by spatiotemporal modeling to exploit redundancies over both space and time. We experimentally demonstrate that accurate registrations can be obtained for varying data robustly and efficiently by applying our method to three standard dynamic face datasets.

This work has been published in 3D Vision 2018 [7].

## 7.7. Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency

*Figure 7. Challenging scene captured with Kinovis. (left) one input image, (center) reconstructions obtained with our previous work based on classical 2D features, (right) proposed solution. Our results validate the key improvement of a CNN-learned disparity to MVS for performance capture scenarios. Results particularly improve in noisy, very low contrast and low textured regions such as the arm, the leg or even the black skirt folds.*

The rise of virtual and augmented reality fuels an increased need for contents suitable to these new technologies including 3D contents obtained from real scenes (see figure 7 ). We consider in this paper the problem of 3D shape reconstruction from multi-view RGB images. We investigate the ability of learning-based strategies to effectively benefit the reconstruction of arbitrary shapes with improved precision and robustness. We especially target real life performance capture, containing complex surface details that are difficult to recover with existing approaches. A key step in the multi-view reconstruction pipeline lies in the search for matching features between viewpoints in order to infer depth information. We propose to cast the matching on a 3D receptive field along viewing lines and to learn a multi-view photoconsistency measure for that purpose. The intuition is that deep networks have the ability to learn local photometric configurations in a broad way, even with respect to different orientations along various viewing lines of the same surface point. Our results demonstrate this ability, showing that a CNN, trained on a standard static dataset, can help recover surface details on dynamic scenes that are not perceived by traditional 2D feature based methods. Our evaluation also shows that our solution compares on par to state of the art reconstruction pipelines on standard evaluation datasets, while yielding significantly better results and generalization with realistic performance capture data.

This work has been published in the European Conference on Computer Vision 2018 [9] and Reconnaissance des Formes, Image, Apprentissage et Perception 2018 [8].

## 7.8. FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis

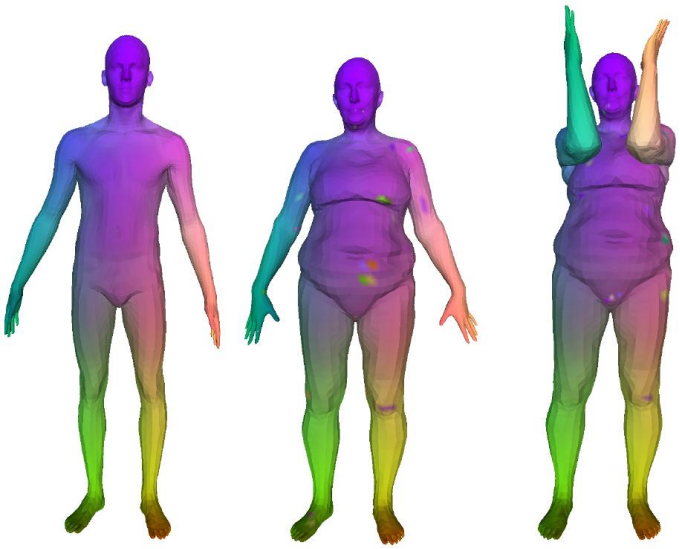


*Figure 8. Two examples of texture transfer from a reference shape in neutral pose (left) using shape correspondences predicted by FeaStNet (multi-scale architecture, without refinement).*

Convolutional neural networks (CNNs) have massively impacted visual recognition in 2D images, and are now ubiquitous in state-of-the-art approaches. CNNs do not easily extend, however, to data that are not represented by regular grids, such as 3D shape meshes or other graph-structured data, to which traditional local convolution operators do not directly apply. To address this problem, we propose a novel graph-convolution operator to establish correspondences between filter weights and graph neighborhoods with arbitrary connectivity. The key novelty of our approach is that these correspondences are dynamically computed from features learned by the network, rather than relying on predefined static coordinates over the graph as in previous work. We obtain excellent experimental results that significantly improve over previous state-of-the-art shape correspondence

results (see figure 8 ). This shows that our approach can learn effective shape representations from raw input coordinates, without relying on shape descriptors.

This work has been published in the IEEE Conference on Computer Vision and Pattern Recognition 2018 [11].

## 7.9. Analyzing Clothing Layer Deformation Statistics of 3D Human Motions

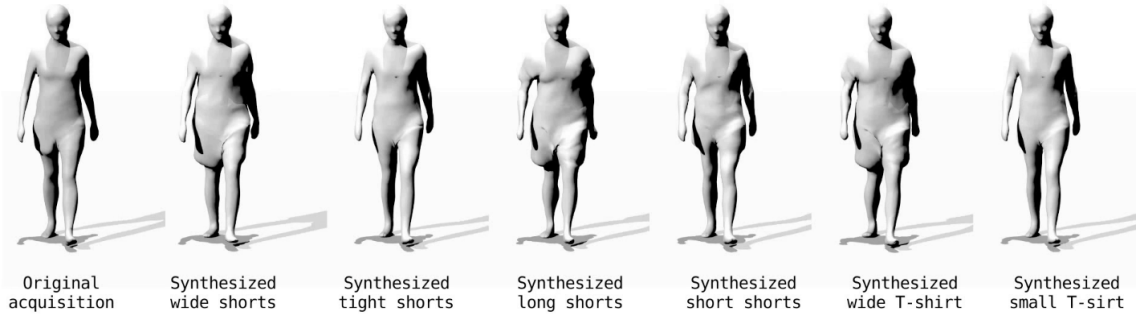



*Figure 9. Examples of clothing re-synthesis based on our clothing layer regression model.*

Recent capture technologies and methods allow not only to retrieve 3D model sequence of moving people in clothing, but also to separate and extract the underlying body geometry and motion component and separate the clothing as a geometric layer. So far this clothing layer has only been used as raw offsets for individual applications such as retargeting a different body capture sequence with the clothing layer of another sequence, with limited scope, e.g. using identical or similar motions. The structured, semantics and motion-correlated nature of the information contained in this layer has yet to be fully understood and exploited. To this purpose we propose a comprehensive analysis of the statistics of this layer with a simple two-component model, based on PCA subspace reduction of the layer information on one hand, and a generic parameter regression model using neural networks on the other hand, designed to regress from any semantic parameter whose variation is observed in a training set, to the layer parameterization space. We show that this model not only allows to reproduce previous motion retargeting works, but generalizes the data generation capabilities of the method to other semantic parameters such as clothing variation and size (see figure 9 ), or physical material parameters with synthetically generated training sequence, paving the way for many kinds of capture data-driven creation and augmentation applications.

This work has been published in the European Conference on Computer Vision 2018 [12].

**MOSAIC Team**

# 6. New Results

## 6.1. Dynamical characterization of morphogenesis at cellular scale

**Participants:** Guillaume Cerutti, Emmanuel Faure [External Collaborator], Christophe Godin, Bruno Leggio, Jonathan Legrand, Patrick Lemaire [External Collaborator], Grégoire Malandain [External Collaborator], Jan Traas [External Collaborator].

- Research Axes: **RA1** (*Representation of biological organisms and their forms in silico*) & **RA3** (*Plasticity & robustness of forms*)
- Key Modeling Challenges: **KMC3** (*Realistic integrated digital models*)

The modeling of morphogenesis requires to explore the interconnection of different spatial and temporal scales of developing organisms. Non-trivial questions such as whether the observed robustness of morphogenesis is rooted in some highly conserved properties at the cellular level or whether it emerges as a macroscopic phenomenon, necessitate precise, quantitative analyses of complex 3D dynamic structures. The study of dynamical properties at the cellular scale poses at the same time key technical challenges and fundamental theoretical questions. An example of the former category is how to characterize and follow the change of shape of cells within tissues and of tissues within organs, and how to couple this change with, for instance, gene expression dynamics; an illustration of the latter is how to define cell-scale variability of morphogenesis within and between species. Our team has produced this year several results in this context:

**Cells spatio-temporal properties and patterns characterization.** Over the past few years, we have achieved quantitative characterization of some of the cells physical properties, such as volumes or curvatures, in a developing tissue. Together with cell lineaging, it also enabled the quantification of temporal properties at cellular scale such as volumetric growth rate or strain patterns. To ease-up the analysis and to structure the previously described data, we have implemented a dedicated spatio-temporal graph structure, formalizing the cell network and its change in time.

To further characterize the tissue development, we developed clustering methods to identify cellular patterns based on a selection of quantified cell properties, including topology. Since such data are highly structured, both in time and space, we developed two complementary approaches:

1. spatial oriented: this approach use the cell neighborhood and a selection of cell descriptors to create pairwise distance maps latter clustered by a distance-based method, such as Ward's hierarchical clustering.
2. temporal oriented: this approach uses the lineage forest and a selection of cell descriptors to infer cell identities using Hidden Markov Tree (HMT) models.

Both approaches allow later characterization of the detected cluster or groups of cells based on their properties and should be published during the first half of 2019.

**Atlases.** One fundamental requirement to understand morphogenesis is the creation of atlases of different properties and different species. This year we have started creating two morphogenetic atlases: the atlas of gene expression patterns in the *Arabidopsis thaliana* flower development and the atlas of early embryonic development of the ascidian *Phallusia mammillata*.

*Phallusia mammillata* embryos develop with an invariant cellular lineage and with a relatively low number of cells ($\sim 700$) up until the end of neurulation. This allows the creation of atlases with cellular resolution. Developing embryos from in-vitro fertilised dechorionated eggs have been injected with mRNA to fluorescently label their cell membranes and imaged by light-sheet microscopy for several hours of development. Automated image reconstruction through the segmentation pipeline ASTEC allowed to collect a large number of wild-type and mutated development with single-cell resolution and with a temporal resolution of two minutes. Based on this amount of data and on the invariant early ascidian lineage, we started curating an atlas of wild-type cellular, tissue and embryonic properties. Each cell, classified by its unique name, is identified in each wild-type embryo and analyzed through the dedicated computational pipelines. The result of this work provides a comprehensive view on the variability (in time, within and between embryos) of properties such as cell volume, cell surface, cell and tissue shape, cell topology, length of cell cycle, cell position within its tissue and globally within the embryo, orientation of cell's cleavage plane. This cellular networks have been coupled via cell names with genetic data coming from the the ascidian genetic database (ANISEED) and a specific tool, Morphonet, has been developed to explore these morphodynamic atlases seamlessly within a web-browser (paper in revision).

On the other hand, developing digital atlases of organism or organs development is a complex challenge for organisms presenting a strong variability in the cellular layout. Indeed contrary to *C. Elegans* or *P. mammillata*, for instance, that posses a very strict cell lineage, the development of most organisms or organs is under the influence of robust genetic patterns but without a unique cellular layout. In that respect, proposing a cell-based atlas of flower development for instance is not straightforward and specific methods have been developed to choose a representative examples of the developing *A. thaliana* flower. Using this representative flower we have generated an atlas in which we have introduced manually the expression patterns of 27 genes. The knowledge generated by the creation of this atlas makes it possible to have a first quantitative (correlative) view on the relation between gene activity and growth.

Both these works should result in publications in 2019.

**Robustness of ascidian embryonic development.** The image segmentation pipeline ASTEC developed by the team allows the 3D dynamic reconstruction of early ascidian embryogenesis at cellular resolution. Based on the high-quality wild-type data of our ascidian morphogenetic atlas and on ANISEED, we investigated the robustness of ascidian embryonic development and established a model to explain its origin. Thanks to the image-analysis pipelines we developed, we could extract relevant information from data and to perform cell-to-cell comparisons between different embryos of the same species (Phallusia mammillata). Since embryos developing from dechorionated eggs are left-right symmetric, we assessed the degree of cell-level variability between two embryos with different genomes (genetic variability) by comparing it to the intrinsic left-right variability in cellular properties within each embryo (stochastic variability). We showed that the same degree of variability is observed within and between embryos, demonstrating how ascidian embryonic development is highly canalised, and that the high reproducibility of shapes observed during embryogenesis is rooted in the robustness of cellular geometry and topology. Based on these observations, we studied the dynamics of embryonic patterning by developing a quantitative mathematical model for cellular fate-restriction events based on kinetic equations describing biochemical signalling. This model suggests that the robustness of cell topology and geometry is necessary for cell-cell biochemical interactions to give rise to the correct fate restriction events, a phenomenon which might represent a strong evolutive constraint to cell-scale variability in ascidians.

These results gave rise to a work which is currently under review and published as a preprint [16].

**Digital reconstruction of developing *Arabidopsis* ovule.** The ovule is a relatively simple organ, with limited developmental variability, which makes it an excellent case study for the computational modeling of organ development. In order to test various hypotheses of cellular growth, we reconstructed a first 4D digital tissue structure of a developing ovule as a triangulated cellular complex. It can be used as an input for FEM-based simulations, and will allow to compare quantitatively the results of growth models with actual ovule development.

This work was part of the *Imago* project.

## 6.2. Reconstruction of macroscopic forms from images and characterization of their variability

**Participants:** Guillaume Cerutti, Christophe Godin, Jonathan Legrand, Katia Mirande.

- Research Axes: **RA1** (*Representations of forms in silico*) & **RA3** (*Plasticity & robustness of forms*)
- Key Modeling Challenges: **KMC3** (*Realistic integrated digital models*)

To study the variability of macroscopic forms resulting from development, it is necessary to both develop digital reconstruction methods, typically based on image acquisitions, and statistical tools to define notions of distance or average between these forms. The automatic inference of computational representations of forms or organ traits from images of different types is therefore an essential step, for which the use of prior knowledge can be very beneficial. Realistic synthetic models of forms can guide the reconstruction algorithms and/or assess their performances. Computational representations of forms can then be used to analyze how forms vary at the scale of a population, of a species or between species, with potential applications in species identification and genetic or environmental robustness estimation.

**Automatized characterization of 3D plant architecture.** The digital reconstruction of branching and organ forms and the quantification of phenotypic traits (lengths of internodes, angles between organs, leaf shapes) is of great interest for the analysis of plant morphology at population scale. We develop an automated processing pipeline that involves the 3D reconstruction of plant architecture from RGB image acquisitions performed by a robot, and the segmentation of the reconstructed plant into organs. To provide validation data for the pipeline, we designed a generative model of *Arabidopsis thaliana* simulating the development of the plant architecture at organ scale. This model was used to develop the method for the measurement of angles of organs and test its accuracy. In a second phase, the model will be used to generate training data for machine learning techniques introduced in the reconstruction methods.

This work is part of the *ROMI* project.

**Identification of plant species from morphological traits.** The description of morphological traits of the various organs of the plants (leaves, bark, flowers and fruits) is essential for the characterization of a phenotype, and is highly relevant in the context of species or variety identification. In the context of tree species identification from RGB images of their organs, we study methods to represent the morphological charateristics of the plant organs, and the way to combine those different sources of information to enhance the classification performance. We demonstrated that botany-inspired descriptors of bark improves tree species classification based on leaves [13]. We also explore the possibility of using deep learning techniques to train a system to extract botanically relevant information from images [3].

This work is part of the *ReVeRIES* ANR project, in which the team is not directly involved.

This work has led to a publication in *Ecological Informatics* and to a participation at the *International Worskhop on Image Analysis Methods for the Plant Sciences* in Nottingham in January 2018.

## 6.3. Analysis of tree data

**Participants:** Romain Azaïs, Christophe Godin, Florian Ingels, Clément Legrand.

- Related Research Axes: **RW1** (*Representations of forms in silico*)
- Related Key Modeling Challenges: **KMC1** (*A new paradigm for modeling tree structures in biology*)

Tree-structured data naturally appear at different scales and in various fields of biology where plants and blood vessels may be described by trees. In the team, we aim to investigate *a new paradigm for modeling tree structures in biology* in particular to solve complex problems related to the *representation of biological organisms and their forms in silico*.

In 2018, we investigated the following questions linked to the analysis of tree data. (i) How to control the complexity of the algorithms used to solve queries on tree structures? For example, computing the edit distance matrix of a dataset of large trees is numerically expensive. (ii) How to estimate the parameters within a stochastic model of trees? And finally, (iii) how to develop statistical learning algorithms adapted to tree data? In general, trees do not admit a Euclidean representation, while most of classification algorithms are only adapted to Euclidean data. Consequently, we need to study methods that are specific to tree data.

**Approximation of trees by self-nested trees.** Complex queries on tree structures (*e.g.*, computation of edit distance, finding common substructures, compression) are required to handle tree objects. A critical question is to control the complexity of the algorithms implemented to solve these queries. One way to address this issue is to approximate the original trees by simplified structures that achieve good algorithmic properties. One can expect good algorithmic properties from structures that present a high level of redundancy in their substructures. Indeed, one can take account these repetitions to avoid redundant computations on the whole structure. In the team, we think that the class of self-nested trees, that are the most compressed trees by DAG compression scheme, is a good candidate to be such an approximation class.

In [7], we have proved the algorithmic efficiency of self-nested trees through different questions (compression, evaluation of recursive functions, evaluation of edit distance) and studied their combinatorics. In particular, we have established that self-nested trees are roughly exponentially less frequent than general trees. This combinatorics can be an asset in exhaustive search problems. Nevertheless, this result also says that one can not always take advantage of the remarkable algorithmic properties of self-nested trees when working with general trees. Consequently, our aim is to investigate how general trees can be approximated by simplified trees in the class of self-nested trees from both theoretical and numerical perspectives.

We conjecture that the problem of optimal approximation by a self-nested tree is NP-hard. Despite a substantial work in 2018 (internship of Clément Legrand), this remains an open question. Consequently, we have developed a suboptimal approximation algorithm based on the *height profile* of a tree that can be used to very rapidly predict the edit distance between two trees, which is a usual but costly operation for comparing tree data in computational biology [7]. Another algorithm based on the simulation of Gibbs measures on the space of trees is currently under development. This work should result in a publication next year.

**Statistical inference.** The main objective of statistical inference is to retrieve the unknown parameters of a stochastic model from observations. A Galton-Watson tree is the genealogical tree of a population starting from one initial ancestor in which each individual gives birth to a random number of children according to the same probability distribution, independently of each other. In a recent work [12], we have focused on Galton-Watson trees conditional on their number of nodes. Several main classes of random trees can be seen as conditioned Galton-Watson trees. For instance, an ordered tree picked uniformly at random in the set of all ordered trees of a given size is a conditioned Galton-Watson tree with offspring distribution the geometric law with parameter $1/2$. Statistical methods were developed for conditioned Galton-Watson trees in [19]. We have introduced new estimators and stated their consistency. Our techniques improve the existing results both theoretically and numerically. A simulation study shows the good behavior of our procedure on finite-sample sizes and from missing or noisy data.

In a very different context, a substantial work has been made on statistical inference for piecewise-deterministic processes [2], [9], [8].

**Kernel methods for tree data.** In statistical learning, one aims to build a decision rule of a qualitative variable $Y$ as a function of a feature $X$ (typically a vector of $\mathbb{R}^d$) from a training dataset $(X_i, Y_i)_{1\leq i\leq n}$. We assume that $X$ is a tree, ordered or not, with or without labels. This framework is quite original since the state space of $X$ is not endowed with a canonical inner product. Kernel methods are particularly adapted to this setting since they enable to transform the raw data into a Hilbert space. In this context, the main issue is related to the construction of a *good kernel*. A kernel function adapted to trees is the subtree kernel introduced [24]. While the literature has never been focused on the weight function involved in the subtree kernel, we have shown that this function is crucial in prediction problems. We have proposed a new algorithm for computing the subtree kernel. It has been designed to allow learning the weight function directly from the data. On some difficult datasets, the prediction error is dramatically decreased from $> 50\%$ to $3\%$.

This work is part of the *ROMI* project, that aims to develop an open and lightweight robotics platform for microfarms. This project requires to investigate advanced analysis and modeling techniques for plant structures. A main issue that arises in this context is to predict a feature of the plant (species, health status, etc) from its topology.

**Invited talk on tree structures and algorithms** Christophe Godin gave a invited talk entitled *Can we manipulate tree forms like numbers?* that was prepared with Romain Azaïs at the workshop on *Mathematics for Developmental Biology* organized at the Banff International Research Station for Mathematical Innovation and Discovery, organized by P. Prusinkiewicz and E. Mjolsness (Banff, Canada, December 2017).

*Abstract*: Tree-forms are ubiquitous in nature and recent observation technologies make it increasingly easy to capture their details, as well as the dynamics of their development, in 3 dimensions with unprecedented accuracy. These massive and complex structural data raise new conceptual and computational issues related to their analysis and to the quantification of their variability. Mathematical and computational techniques that usually successfully apply to traditional scalar or vectorial datasets fail to apply to such structural objects: How to define the average form of a set of tree-forms? How to compare and classify tree-forms? Can we solve efficiently optimization problems in tree-form spaces? How to approximate tree-forms? Can their intrinsic exponential computational curse be circumvented? In this talk, we presented a recent work to approach these questions from a new perspective, in which tree-forms show properties similar to that of numbers or real functions: they can be decomposed, approximated, averaged, and transformed in dual spaces where specific computations can be carried out more efficiently. We will discuss how these first results can be applied to the analysis and simulation of tree-forms in developmental biology (https://www.birs.ca/events/2017/5-day-workshops/17w5164).

## 6.4. Mechanics of tissue morphogenesis

**Participants:** Olivier Ali, Arezki Boudaoud [External Collaborator], Guillaume Cerutti, Ibrahim Cheddadi [External Collaborator], Christophe Godin, Bruno Leggio, Jonathan Legrand, Hadrien Oliveri, Jan Traas [External Collaborator].

- Research Axes: **RA2** (*Data-driven models*) & **RA3** (*Plasticity & robustness of forms*)
- Key Modeling Challenges: **KMC2** (*Efficient computational mechanical models of growing tissues*) & **KMC3** (*Realistic integrated digital models*)

As deformations supporting morphogenesis require the production of mechanical work within tissues, the ability to simulate accurately the mechanical behavior of growing living tissues is a critical issue of the MOSAIC project. From a macroscopic perspective, tissues mechanics can be formalized through the framework of continuum mechanics. However, the fact that they are composed, at the microscopic level, by active building blocks out of equilibrium (namely cells) offers genuine modeling challenges and opportunities. This section describes the team's efforts on integrating cellular behaviors such as mechano-sensitivity, intercellular fluxes of materials and cell division into a macroscopic mechanical picture of morphogenesis.

**Mechanical influence of inner tissues.** Mechanical stress patterns within plant tissues emerge from the balance between inner-pressure-induced forces and the elastic response of the cell wall [0] over the entire tissue. Being able to derive, from a specific cellular architecture, the corresponding pattern of stresses within a tissue is crucial for the study of morphogenesis. It requires a precise description of the tissue as a network of connected cells and the ability to run numerical simulations of force balance on such heterogeneous structures.

To that end, we developed numerical methods to generate finite element meshes from: i) 3D microscopic images with sub-cellular resolution (referred to as *bio-inspired* structures) and ii) 3D cellularized geometrical volumes (referred to as *artificial* structures). Combined with a FEM-based simulation framework previously developed within the team [20], we generated quantitative maps of stress distributions in multilayered reconstructed tissues. The combined analysis of stress patterns on *bio-inspired* and *artificial* structures showed how mechanical stresses experienced by cells convey geometrical information to cells about the global shape of the tissue as well as the local shape of cells.

[0] A thick protective exoskeleton surrounding plant cells

This work was part of the *Morphogenetics* IPL and Jan Traas ERC grant *Morphodynamics*.

This work is currently under review in the *Bulletin of Mathematical Biology* and has been presented at the *19th International Conference of System Biology* held in Lyon at fall.

**Shape regulation.** Reproducible and robust morphogenesis requires growth coordination of thousands of cells. How such coordination can be "implemented" in living organism is a core question for *RA3*. One identified mechanism in plant to coordinate growth rely on cells mechano-sensitivity [0]. Combined with the geometrical dependency of mechanical stress (*c.f.* previous subsection), this suggests the existence of a feedback mechanism that regulates tissue shape changes. We have been investigating closely the consequences of such a mechanism.

To that end, we first modeled the bio-molecular pathway relating mechanical stress experienced by cells to actual modification of their mechanical properties (*e.g.* cell wall stiffness). This work enabled us to describe plant tissues as an active material featuring large-scale properties, such as stress stiffening [0], emerging from sub-cellular dynamics. This work has been published in *Journal of Mathematical Biology* [5].

In parallel, we modeled the influence of cell wall elasticity (value, orientation) on the growth dynamics of tissues. This was done in the context of plant organogenesis, in close collaboration with biologists investigating the effect of cell-wall-related mutations on plant organ initiation. Our modeling approach was based on our previously developed *strain-based growth* model [20]. This joint study has been published in *Development* earlier this year [1].

We then studied how initial spherical symmetry (e.g. dome-shaped primordia) can be potentially broken during development in such active tissues and lead to elongated or flat shapes. For this, we integrated the *stress feedback* model with the *strain-based growth* model to investigate how their interplay could influence the morphogenesis of 3D cellularized structures. In particular, we showed that a stress-based feedback mechanism can maintain the typical plant growth modes (*i.e.* axial elongation or 2D flat expansion) and amplify asymmetries. This computational approach to symmetry breaking in growing tissues has been developed in parallel to experimental investigations addressing the shape evolution of sepals [0].

This work was part of the *Mophogenetics* IPL and Jan Traas ERC grant *Morphodynamics*.

The whole story has been presented at the *9th International Plant Biomechanics Conference* in Montreal this summer. A journal article combining both our modeling approach and experimental work in the context of symmetry breaking during plant organogenesis is currently being written.

**Influence of water fluxes on plant morphogenesis.** Since pressure appears as the "engine" behind growth-related deformation in plants, its regulation by cells is a major control mechanism of morphogenesis. We developed 2D computational models to investigate the morphological consequences of the interplay between cell expansion, water fluxes between cells and tissue mechanics. This interdisciplinary work, combining experiments and modeling, addresses the influence of turgor pressure heterogeneities on relative growth rate between cells. We showed that the coupling between fluxes and mechanics allows us to predict observed morphological heterogeneities without any *ad hoc* assumption. It also reveals the existence of a putative inhibitory action of organ growth on growth in immediatemy neighboring regions, due to the hydraulic coupling between cells during growth.

This work was part of the Agropolis fundation project *MecaFruit3D* and Arezki Boudaoud's ERC *PhyMorph.*

Two papers report the results of this work (one currently under review in *Nature Physics* [21] and a second one that is about to be submitted. These results have also been presented last summer at the *9th International Plant Biomechanics Conference* in Montreal.

**Influence of dividing cells on tissue mechanics during morphogenesis in ascidians.** The control of cell division orientation is of prime importance for patterning and shape emergence, especially in animal embryos where the first developmental stages happen at constant volume. In recent years, the Hetwig's rule appeared

[0] the ability to probe mechanical stress around them and to modify accordingly their growth behavior
[0] the ability of the tissue to re-enforce itself in the directions of high mechanical solicitations
[0] leaf-like organs surrounding and protecting flowers

as a physical model accounting for orientation of cell division. Within animal tissues it has been shown that the coupling of externally induced strain and Hertwig's rule leads to the orientation of cell divisions with the main stress direction.

We investigated through modeling the consequences, in a multicellular context, of such stress-based regulation of cell division orientation. To that end, we developed a theoretical standpoint on the many-body energetic thermodynamics of cell divisions in the presence of external anisotropic stress. We showed that Hertwig's rule emerges as a limiting-case behavior and how anisotropic mechanical stresses can provide important cues to guide cell divisions. Our model accounts for the division pattern observed in the epidermis of the embryo of ascidian *Phallusia mammillata*, including those reproducible observed deviations from Hertwig's rule which have so far eluded explanation.

This work was part of the *Digem* project.

This work has been presented in two national conferences: the *IBC Scientific Days* and the *Cell Cycle Days* both held in Montpellier. A paper is currently being written.

**Automatic quantification of adhesion defects in microscopy images.** Direct measurements of mechanical stresses experienced by living tissues are not yet feasible. To circumvent this limitation, we developed an indirect method based on measurements of cracks in tissues: Our biologist colleagues developed cell-adhesion mutants in which strong connections between epithelial cells are impaired. As a consequence, mechanical stresses within the tissue produce cracks. Distribution and orientation of these cracks can be related to the main directions of the mechanical forces at play. We developed a 2D image analysis pipeline to detect and quantify these cracks in microscopy projections of epithelia, and deduce the magnitude and orientation of tensions in organs and tissues. This tool has been used to evidence new mechanical signaling mechanisms in *Arabidopsis*.

This analysis pipeline has been published in [6] and used by collaborators in the analysis performed in [26].

## 6.5. Signaling and transport for tissue patterning

**Participants:** Romain Azaïs, Guillaume Cerutti, Christophe Godin, Bruno Leggio, Jonathan Legrand, Teva Vernoux [External Collaborator].

- Research Axes: **RA1** (*Representations of forms in silico*) & **RA2** (*Data-driven models*)
- Key Modelling Challenges: **KMC3** (*Realistic integrated digital models*)

One central mechanism in the shaping of biological forms is the definition of regions with different genetic identities or physiological properties through bio-chemical processes operating at cellular level. Such patterning of the tissue is often controlled by the action of molecular signals for which active or passive transport mechanisms determine patterning spatial precision. The shoot apical meristem (SAM) of flowering plants is a remarkable example of such finely controlled system where the dynamic interplay between the hormone auxin and the polarization of efflux carriers PIN1 during growth governs the rhythmic patterning of organs, and the consequent emergence of phyllotaxis. Using *Arabidopsis thaliana* as a model system, we developed an integrated view of the meristem as a self-organizing dynamical form by reconstructing the dynamics of physiological processes from living tissues, and by proposing computational models integrating transport and signaling to study tissue patterning *in silico*.

**Automatic quantification of auxin transport polarities.** Time-lapse imaging of living SAM tissues marked with various fluorescent proteins allows monitoring the dynamics of cell-level molecular processes. Using a co-visualization of functional fluorescent auxin transporter (PIN1-GFP) with a dye staining of cell walls with propidium iodide (PI), we developed a method to quantify in 3D the polarization of auxin transport for every anticlinal wall of the first layer of cells. The digitally reconstructed network evidenced an overall stable convergence of PIN1 polarities towards the center of the meristem, with local front lines matching dynamic accumulations of auxin [15]. It also showed that the apparent crescent shape often thought to indicate polarities in cells might sometimes be misleading, and opens the way for a new view of how auxin transport is regulated.

**Temporal auxin signaling in meristem organ patterning.** Morphogenetic signals such as auxin define spatial distributions that are thought to control tissue patterning, but it has been proposed in animals that they also carry temporal information in their dynamics. A recent model developed by our group has postulated the existence of a stochastic mechanism to explain disturbed phyllotaxis patterns. This model assumes that organ initiation results from a temporal integration of a morphogenetic signal that buffers molecular noise [22]. Using a quantitative analysis of the dynamics of auxin distribution and response, we provide evidence that organ initiation in the SAM is indeed dependent on the temporal integration of the auxin signal [15]. The duration of cell exposition to auxin is used to differentiate temporally sites of organ initiation, and provide robustness to the rhythmic organ patterning.

**Computational models of integrated transport and signaling.** To interpret these new observations of auxin signaling and transport in the meristem, we investigate theoretical and computational models to study dynamic auxin distributions and the consequent organ patterning at the level of the meristem. Building on existing models of auxin transport [23], [25], we investigate different competing hypotheses on the auxin-PIN interplay, through numerical simulations based on rate equations for molecular transport and efflux carrier polarization. Quantitative comparisons with in vivo observations will provide cues on how the system responses are linked to memory effects and information exchanges between auxin and PINs.

These works were part of the *BioSensors* HFSP project and are carried out in the *Phyllo* ENS-Lyon project and gave rise to a journal article submitted for publication. These results have been presented at the *International Worskhop on Image Analysis Methods for the Plant Sciences* in Nottingham in January 2018 and in several invited talks given by Teva Vernoux and Christophe Godin.

## 6.6. Regulation of branching mechanisms in plants

**Participants:** Romain Azaïs, Frédéric Boudon [External Collaborator], Christophe Godin.

- Research Axes: **RA2** (*Data-driven models*) & **RA3** (*Plasticity & robustness of forms*)
- Key Modelling Challenges: **KMC3** (*Realistic integrated digital models*)

Branching in plants results from the development of apical meristems that recursively produce lateral meristems. These meristems may be more or less differentiated with respect to the apical meristem from which they originate, potentially leading to different types of lateral branches or organs. They also can undergo a more or less long period of inactivation, due to systemic regulation. The understanding of branching systems morphogenesis in plants thus relies on the analysis of the regulatory mechanisms that control both meristem differentiation and inactivation.

**Analysis of the diversity of inflorescence architecture in different rice species.** Rice is a major cereal for world food security and understanding the genetic and environmental determinants of its branching habits is a timely scientific challenge. The domestication, i.e., the empirical selection by humans, of rice began 10 000 years ago in Asia and 3 000 years ago in Africa. It thus provides a short-term model of the processes of evolution of plants.

Hélène Adam and Stéphane Jouannic from the group Evo-Devo de l'Inflorescence of UMR DIADE at IRD (Montpellier) have collected for years on the different continents an outstanding database of panicle-type inflorescence phenotypes in Asian and African, cultivated and wild, rice species. Classical statistical analysis based on the extraction of characteristic traits for each individual branching system were able to separate wild species from cultivated ones, but could not discriminate between wild species, suggesting that the entire branching structure should be used for classification methods to operate. For this, we are currently developing statistical methods on tree structures (see section 6.3) that should allow us to achieve better discrimination between panicles, based on their branching topology in addition to geometric traits. By coupling the quantitative study of the panicles to genomic analyses carried out by the IRD group, we should be able to highlight which regulation pathways have been selected or altered during the domestication process.

**The role of sugars in apical dominance.** The outgrowth of axillary buds is a key process in plant branching and which is often shown to be suppressed by the presence of auxin in nodal stems. However, local auxin levels are not always sufficient to explain bud outgrowth inhibition. Recent studies have also identified a contribution of sugar deprivation to this phenomenon. Whether sugars act independently of auxin or other hormones auxin regulates is unknown. Auxin has been shown to induce a decrease of cytokinin levels and to upregulate strigolactone biosynthesis in nodes. Based on rose and pea experiments, both in vitro and in planta, with our collaborators Jessica Bertheloot, Soulaiman Sakr from Institut de Recherche en Horticulture et Semences (IRHS) in Angers, we have shown that sucrose and auxin act antagonistically, dose-dependently, and non-linearly to modulate bud outgrowth. The Angers group provided experimental evidence that sucrose represses bud response to strigolactones but does not markedly affect the action of auxin on cytokinin levels. Using a modeling approach, we tested the ability of this complex regulatory network to explain the observed phenotypes. The computational model can account for various combinations of sucrose and hormones on bud outgrowth in a quantitative manner and makes it possible to express bud outgrowth delay as a simple function of auxin and sucrose levels in the stem. These results provide a simple auxin-sucrose-cytokinin-strigolactone network that accounts for plant adaptation to growing conditions. A paper relating this work is currently under review.

**The fractal nature of plants.** Inflorescence branching systems are complex and diverse. They result from the interaction between meristem growth and gene regulatory networks that control the flowering transition during morphogenesis. To study these systems, we focused on cauliflower mutants, in which the meristem repeatedly fails in making a complete transition to the flower and for which a complete mechanistic explanation is still lacking.

In collaboration with Eugenio Azpeitia and François Parcy's group in Grenoble, we have developed a first model of the control of floral initiation by genes, refining previous networks from the literature so that they can integrate our hypotheses about the emergence of cauliflower phenotypes. The complete network was validated by multiple analyses, including sensitivity analyses, stable state analysis, mutant analysis, among others. It was then coupled with an architectural model of plant development using L-systems. The coupled model was used to study how changes in gene dynamics and expression could impact in different ways the architectural properties of plants. The model was then used to study how changes in certain parameters could generate different curd morphologies, including the normal and the fractal-like Romanesco. A paper reporting this work is currently being written.

**NANO-D Project-Team**

# 6. New Results

## 6.1. Generating conformational transition paths with low potential-energy barriers for proteins

**Participants:** Minh Khoa Nguyen, Léonard Jaillet and Stéphane Redon.
**Publication:** Journal of Computer-Aided Molecular Design, 2018 [66].

The knowledge of conformational transition paths in proteins can be useful for understanding protein mechanisms. Recently, we have introduced the As-Rigid-As-Possible (ARAP) interpolation method, for generating interpolation paths between two protein conformations. The method was shown to preserve well the rigidity of the initial conformation along the path. However, because the method is totally geometry-based, the generated paths may be inconsistent because the atom interactions are ignored. Therefore, we introduce a new method to generate conformational transition paths with low potential-energy barriers for proteins. The method is composed of three processing stages. First, ARAP interpolation is used for generating an initial path. Then, the path conformations are enhanced by a clash remover. Finally, Nudged Elastic Band, a path-optimization method, is used to produce a low-energy path. Large energy reductions are found in the paths obtained from the method than in those obtained from the ARAP interpolation method alone. The results also show that ARAP interpolation is a good candidate for generating an initial path because it leads to lower potential-energy paths than two other common methods for path interpolation (see Figure 1 for an example of optimized transition path).

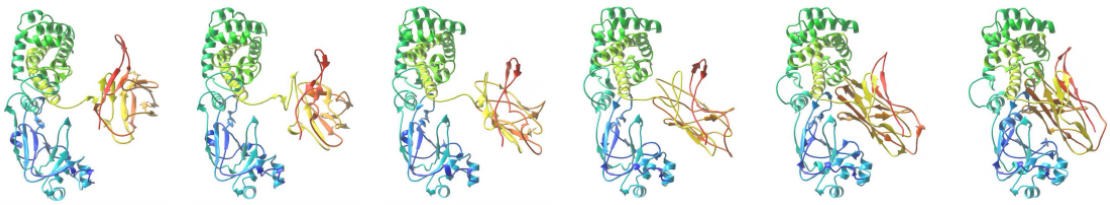


*Figure 1. The path for diphtheria toxin after ARAP interpolation and NEB optimization.*

## 6.2. ART-RRT: As-Rigid-As-Possible Exploration of Ligand Unbinding Pathways

**Participants:** Minh Khoa Nguyen, Leonard Jaillet, Stephane Redon.
**Publication:** Journal of Computational Chemistry, 2018 [65].

We have proposes a method to efficiently generate approximate ligand unbinding pathways. It combines an efficient tree-based exploration method with a morphing technique from Computer Graphics for dimensionality reduction. This method is computationally cheap and, unlike many existing approaches, does not require a reaction coordinate to guide the search. It can be used for finding pathways with known or unknown directions beforehand. The approach is evaluated on several benchmarks and the obtained solutions are compared with the results from other state-of-the-art approaches. We show that the method is time-efficient and produces pathways in good agreement with other state-of-the-art solutions. These paths can serve as first approximations that can be used, analyzed, or improved with more specialized methods (see Figure 2 ).

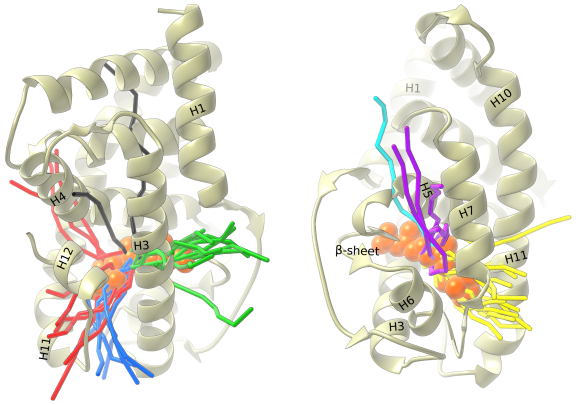



*Figure 2. Families of paths (in colored sticks) obtained with ART-RRT for the unbinding of retinoic acid hormone from its receptor. The protein is represented by ribbons and the ligand by orange balls.*

## 6.3. Atomistic modelling and simulation of transmission electron microscopy images: application to intrinsic defects of graphene

**Participants:** Cyril Guedj, Léonard Jaillet, François Rousse and Stéphane Redon.
**Publication:** Proceedings of 8th International Conference on Simulation and Modeling Methodologies, Technologies and Applications - Volume 1: SIMULTECH [53].
The characterization of advanced materials and devices in the nanometer range requires complex tools, and the data analysis at the atomic level is required to understand the precise links between structure and properties. We have demonstrated that the atomic-scale modelling of graphene-based defects may be performed efficiently for various structural arrangements using the Brenner module of the SAMSON software platform (cf Figure 3 ). The signatures of all kinds of defects are computed in terms of energy and scanning transmission electron microscopy simulated images. The results are in good agreement with all theoretical and experimental data available. This original methodology is an excellent compromise between the speed and the precision required by the semiconductor industry and opens the possibility of realistic in-silico research conjugated to experimental nanocharacterisation of these promising materials.

## 6.4. Impact of hydrogen on graphene-based materials: atomistic modeling and simulation of HRSTEM images.

**Participants:** Cyril Guedj, Léonard Jaillet, François Rousse and Stéphane Redon.
**Oral presentation:** AVS 65th International Symposium and Exhibition.
**Summary:** The hydrogen energy transition is highly probable, because hydrogen is the most abundant element in the universe and represents an ideal “green” source of energy. Meanwhile, the safe hydrogen production and storage remains a major challenge still in progress. To understand and optimize the device efficiency and the interface engineering, it is advantageous to perform advanced nanocharacterizations, linked to numerical modelling and simulations. This task is particularly difficult, because hydrogen is labile and prone to rapid reorganization. This structural evolution may be monitored with transmission electron microscopy (TEM) techniques, but in spite of significant progresses, the direct detection of hydrogen with High Resolution Scanning Transmission Electron Microscopy (HRSTEM) or energy-loss spectroscopy still remains a serious challenge.

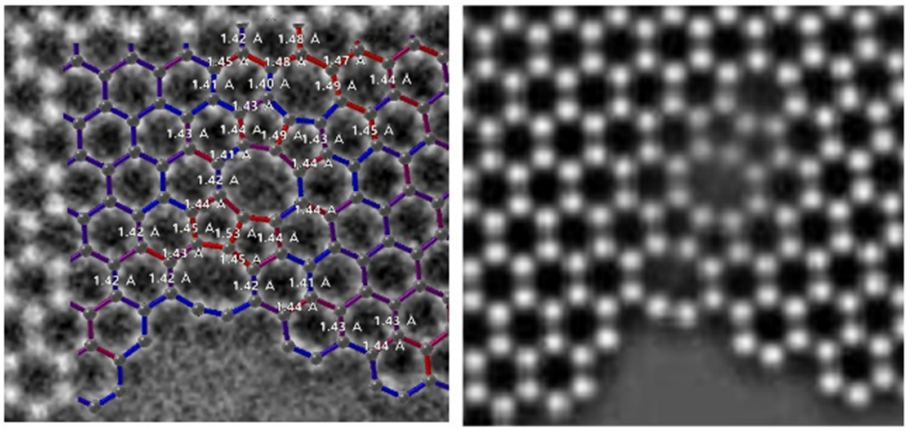


*Figure 3. Left: atomistic model of the extended defect 88-7-5555 defect superimposed to the experimental HRTEM image entitled "SALVE-III-project-HRTEM-graphene-vacancy-characteristic-defects.png" (Salve, 2018). Right: corresponding simulated HRTEM image.*

We investigate here the interaction of hydrogen with graphene using the Brenner module of the SAMSON software platform and we propose an original methodology to characterize its structural arrangement at the atomic scale by simulating HRSTEM images to interpret experimental results. In particular, we compare the effect of hydrogen on dark field (DF), bright field (BF), high-angle annular dark filed (HAADF) and annular bright field (ABF) images, to estimate the best technique suited to hydrogen detection. In addition, we present the effect of carbon vacancies and adatoms on the stability of hydrogen coverage, associated to the HRSTEM signatures of the most stable configurations. These results provide the necessary building blocks to analyze the structure and energetics of hydrogenated graphene-based materials at the atomic scale.

## 6.5. Atomistic modelling of diamond-type $Si_xGe_yC_zSn_{1-x-y-z}$ crystals for realistic transmission electron microscopy image simulations

**Participants:** Leonard Jaillet and Cyril Guedj.
The realistic simulations of transmission electron microscopy (TEM) images requires an accurate definition of the positions of all atoms, which are linked to the mechanical properties of the material. We are working on an approach to build optimized models to represent the lattice parameters and elastic properties of Si, Ge, diamond, alpha-tin and related diamond alloys.

In order to compute precisely the complex $Si_xGe_yC_zSn_{1-x-y-z}$ diamond crystals, a dedicated parametrization of the Keating force field has been proposed. An original periodic boundary strategy has also been provided. Our tool can be used to interpret experimental TEM with a speed several orders of magnitude higher than for ab-initio methods. The method predicts the correct lattice parameters and elastic constants for published experimental results with low deviation. Finally, we have shown that subsequent Monte Carlo simulations predict original self-ordering effects in C in good agreement with the theory. A publication is in preparation on this topic.

## 6.6. Analytical symmetry detection method AnAnaS

**Participants:** Guillaume Pagès, Sergei Grudinin, Elvira Kinzina.
**Publications:** Journal of Structural Biology, 2018 [21], Journal of Structural Biology, 2018 [20].

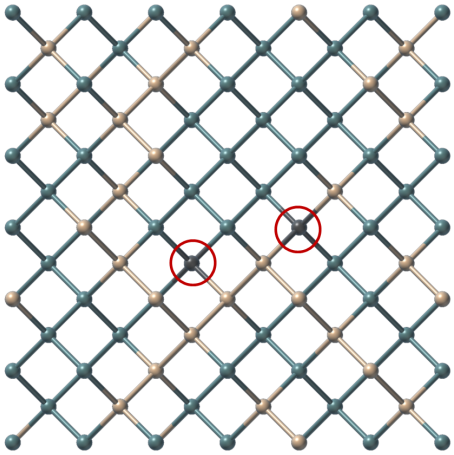


*Figure 4. Crystal of $Si_{40}Ge_{60}$ where two carbon atoms (circled in red) have been inserted. The properties of the crystal such as its lattice parameter can be characterized in function of the position of the carbon atoms.*

Macromolecules are generally not rigid bodies at physiological temperature and they adopt different conformational states. Thus, if one considers a macromolecular assembly made of $N$ subunits, do we expect that all the units will be structurally identical to each other? Most probably not, since at any given moment of time, each unit may be sampling a different conformational state. For example, there are plenty of X-ray structures of homo-dimers, where the individual monomers are not structurally identical.

In order to quantitatively assess these differences, we developed a method for Analytical Analysis of Symmetries (AnAnaS) in protein complexes. The method is extremely fast, robust and accurate. Two papers describing the method were published [21], [20]. This method is available on the website of the team (https://team.inria.fr/nano-d/software/ananas/).

## 6.7. Deep Learning for Symmetry detection

**Participants:** Guillaume Pagès, Sergei Grudinin.
**Publication:** arXiv preprint, 2018 [29].

We worked on a fully-structural method for detecting symmetries in molecular structures. This allowed us to detect tandem repeats, or even symmetry in density maps. We created a method based on neural network and deep learning, inspired by the advances in computer vision in the past decade. According to our tests on simulated examples, our method is able to detect the order of a cyclic symmetry (which can be 1 for asymmetric structure) with a 92% accuracy, and guesses the direction of the axis of symmetry with an average error of 3°. A manuscript describing this method has been submitted for publication and is available on arXiv [29].

## 6.8. New method for protein model quality assessment Ornate

**Participants:** Benoit Charmettant, Guillaume Pagès, Sergei Grudinin.
**Publication:** bioRxiv preprint, 2018 [28].

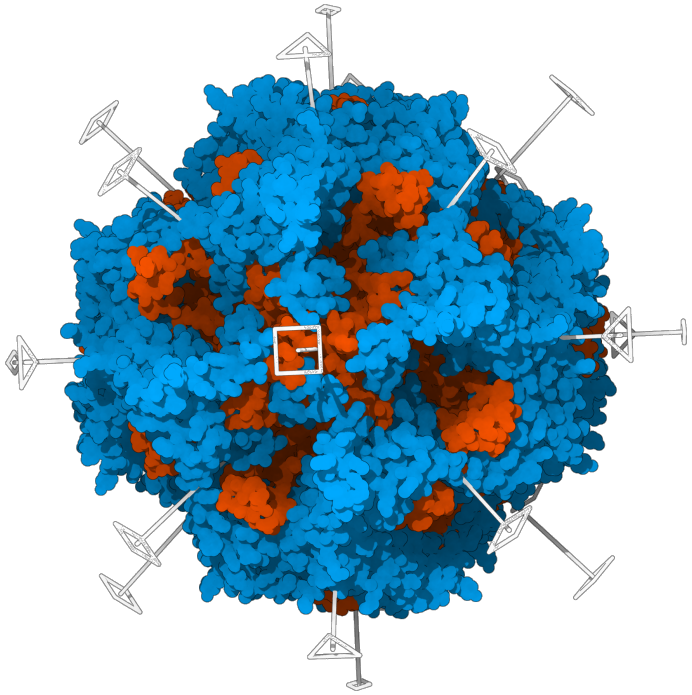


*Figure 5. System with an octahedral symmetry, with the symmetry axes displayed in SAMSON.*

Protein model quality assessment (QA) is a crucial and yet open problem in structural bioinformatics. It consists of estimating a score to assess whether a given three-dimensional structure is correctly folded or not. The current best methods for single-model QA typically combine results from different approaches, each based on different input features constructed by experts in the field. Then, the prediction model is trained using a machine-learning algorithm. Recently, with the development of convolutional neural networks (CNN), the training paradigm has changed. In computer vision, the expert-developed features have been significantly overpassed by automatically trained convolutional filters. This motivated us to apply a three-dimensional (3D) CNN to the problem of protein model QA.

We developed a novel method for single-model QA called Ornate. Ornate (Oriented Routed Neural network with Automatic Typing) is a residue-wise scoring function that takes as input 3D density maps. It predicts the local (residue-wise) and the global model quality through a deep 3D CNN. Specifically, Ornate aligns the input density map, corresponding to each residue and its neighborhood, with the backbone topology of this residue. This circumvents the problem of ambiguous orientations of the initial models. Also, Ornate includes automatic identification of atom types and dynamic routing of the data in the network. Established benchmarks (CASP 11 and CASP 12) demonstrate the state-of-the-art performance of our approach among single-model QA methods. A manuscript describing this method has been submitted for publication and is available on bioRxiv [28].

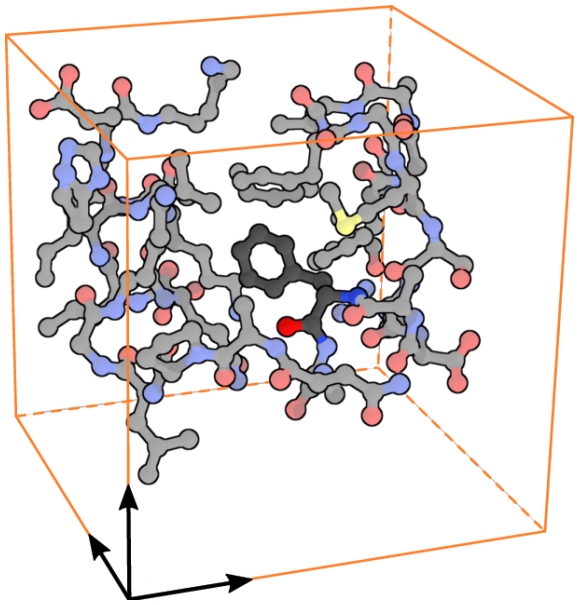


*Figure 6. Example of input given to the 3D CNN Ornate.*

**NECS Project-Team**

# 7. New Results

## 7.1. Network systems: modeling, analysis, and estimation

### *7.1.1. Network reduction towards a scale-free structure preserving physical properties*

**Participants:** N. Martin, P. Frasca, C. Canudas de Wit [Contact person].

In the context of the ERC project, we are addressing a problem of graph reduction, where a given arbitrary weighted graph is reduced to a (smaller) scale-free graph while preserving a consistency with the initial graph and some physical properties. This problem can be formulated as a minimization problem. We give specifications to this general problem to treat a particular case: to this end we define a metric to measure the scale-freeness of a graph and another metric to measure the similarity between two graphs with different dimensions, based on a notion of spectral centrality. Moreover, through the reduction we also preserve a property of mass conservation (essentially, Kirchoff's first law). We study the optimization problem and, based on the gained insights, we derive an algorithm allowing to find an approximate solution. Finally, we have simulated the algorithm both on synthetic networks and on real-world examples of traffic networks that represent the city of Grenoble. These results are presented in [57] and in [31]. We also developed an application to the control of epidemics [58].

### *7.1.2. Cyber-Physical Systems: a control-theoretic approach to privacy and security*

**Participants:** F. Garin [Contact person], A. Kibangou, S. Gracy.

Cyber-physical systems are composed of many simple components (agents) with interconnections giving rise to a global complex behaviour. Interesting recent research has been exploring how the graph describing interactions affects control-theoretic properties such as controllability or observability, namely answering the question whether a small group of agents would be able to drive the whole system to a desired state, or to retrieve the state of all agents from the observed local states only.

A related problem is observability in the presence of an unknown input, where the input can represent a failure or a malicious attack, aiming at disrupting the normal system functioning while staying undetected. We study linear network systems, and we aim at characterizing input and state observability (ISO), namely the conditions under which both the whole network state and the unknown input can be reconstructed from some measured local states. We complement the classical algebraic characterizations with novel structural results, which depend only on the graph of interactions (equivalently, on the zero pattern of the system matrices). More precisely, we obtain two kinds of results (see [24], [25] and the PhD thesis of S. Gracy): structural results, true for almost all interaction weights, and strongly structural results, true for all non-zero interaction weights. We consider both the case where the system graph is time-invariant, and the case where it varies in time.

When the conditions for ISO are satisfied, one can run algorithms in the same vein as a Kalman filter, in order to reconstruct the state and the unknown input from noisy measurements. These algorithms are known for the case where the input can be reconstructed with only one time-step of delay with respect to the measurements; in [54] we propose a (suboptimal) filter for the case when this is not possible, i.e., more measurements are needed for the input reconstruction.

### *7.1.3. Heterogeneity and uncertainty in distributed estimation from relative measurements*

**Participants:** C. Ravazzi, N. K. Chan, P. Frasca [Contact person].

This work, presented in [34], has studied the problem of estimation from relative measurements in a graph, in which a vector indexed over the nodes has to be reconstructed from pairwise measurements of differences between its components associated to nodes connected by an edge. In order to model heterogeneity and uncertainty of the measurements, we assume them to be affected by additive noise distributed according to a Gaussian mixture. In this original setup, we formulate the problem of computing the Maximum-Likelihood (ML) estimates and we design two novel algorithms, based on Least Squares regression and Expectation-Maximization (EM). The first algorithm (LSEM) is centralized and performs the estimation from relative measurements, the soft classification of the measurements, and the estimation of the noise parameters. The second algorithm (Distributed LS-EM) is distributed and performs estimation and soft classification of the measurements, but requires the knowledge of the noise parameters. We provide rigorous proofs of convergence for both algorithms and we present numerical experiments to evaluate their performance and compare it with solutions from the literature. The experiments show the robustness of the proposed methods against different kinds of noise and, for the Distributed LS-EM, against errors in the knowledge of noise parameters.

### *7.1.4. Average state estimation in large-scale multi-cluster networks*

**Participants:** U. Niazi, A. Kibangou, C. Canudas de Wit [Contact person].

In the context of the ERC project, we are addressing the problem of estimation of a functional of non-observed states. Indeed, large-scale network systems can be unobservable from the dedicated state measurements at few nodes. By resorting to an aggregation of multiple clusters of unmeasured nodes, we are investigating the observability and detectability of average states of the clusters. The approach is to obtain a reduced network system whose state vector contains the average states of the clusters. The notion of average observability is defined with respect to the observability of this reduced network system. For average observability, we have stated a necessary condition and a sufficient condition depending solely on the structure of the network. Average detectability, which is a milder notion than average observability, is also studied and a sufficient condition, under which an open-loop average state observer converges, is provided. This condition requires clusters of unmeasured nodes to have negatively balanced local outflow centrality.

## 7.2. Control of multi-agent systems and opinion dynamics

### *7.2.1. Open multi-agent systems: Dynamic consensus*

**Participants:** W. S. Rossi, P. Frasca [Contact person].

In [53] we investigate a dynamic consensus problem for an open multi-agent system. Open multi-agent systems are characterized by a time-varying set of agents connected by a network: agents may leave and new agents may join the network at any time, thus the term “open”. The dynamic consensus problem consists in achieving agreement about the time-varying average of a set of reference signals that are assumed to be the agents’ inputs. Dynamic consensus has recently found application in the context of distributed estimation for electric demand-side management, where a large population of connected domestic appliances needs to estimate its future average power consumption. Since the considered network of devices changes as new appliances log in and out, there is a need to develop and characterize dynamic consensus algorithms for these open scenarios. In this paper we give several initial contributions both to a general theory of open multi-agent systems and to the specific problem of dynamic consensus within this context. On the theoretical side, we propose a formal definition of open multi-agent system, a suitable notion of stability, and some sufficient conditions to establish it. On the applied side, we design a novel dynamic consensus algorithm, the Open Proportional Dynamic Consensus algorithm. We characterize some of its convergence properties in the proposed open-multi-agent systems framework and we illustrate its evolution by numerical simulations.

### *7.2.2. Robust average consensus over unreliable networks*

**Participants:** F. Acciani, P. Frasca [Contact person], G. Heijenk, A. Stoorvogel.

Packet loss is a serious issue in wireless consensus networks, as even few failures might prevent a network to converge to the desired consensus value. In the last four years, we have devised some possible ways to compensate for the errors caused by packet collisions, by modifying the updating weights. Since these modifications may result in a reduced convergence speed, a gain parameter is used to increase the convergence speed, and an analysis of the stability of the network is performed, leading to a criterion to choose such gain to guarantee network stability. For the implementation of the compensation method, we propose a new communication algorithm, which uses both synchronous and asynchronous mechanisms to achieve average consensus and to deal with uncertainty in packet delivery. The paper [14] provides a complete account of our results.

### 7.2.3. *Asynchronous opinion dynamics on the $k$-nearest-neighbors graph*

**Participants:** W. S. Rossi, P. Frasca [Contact person].

This work is about a new model of opinion dynamics with opinion-dependent connectivity. We assume that agents update their opinions asynchronously and that each agent's new opinion depends on the opinions of the k agents that are closest to it. In the paper [63], we show that the resulting dynamics is substantially different from comparable models in the literature, such as bounded-confidence models. We study the equilibria of the dynamics, observing that they are robust to perturbations caused by the introduction of new agents. We also prove that if the number of agents n is smaller than 2k, the dynamics converge to consensus. This condition is only sufficient.

### 7.2.4. *Quantization effects in opinion dynamics*

**Participants:** F. Ceragioli, P. Frasca [Contact person].

This work deals with continuous-time opinion dynamics that feature the interplay of continuous opinions and discrete behaviors. In our model, the opinion of one individual is only influenced by the behaviors of fellow individuals. The key technical difficulty in the study of these dynamics is that the right-hand sides of the equations are discontinuous and thus their solutions must be intended in some generalized sense: in our analysis, we consider both Carathéodory and Krasovskii solutions. We first prove the existence and completeness of Carathéodory solutions from every initial condition and we highlight a pathological behavior of Carathéodory solutions, which can converge to points that are not (Carathéodory) equilibria. Notably, such points can be arbitrarily far from consensus and indeed simulations show that convergence to nonconsensus configurations is common. In order to cope with these pathological attractors, we study Krasovskii solutions. We give an estimate of the asymptotic distance of all Krasovskii solutions from consensus and we prove its tightness by an example of equilibrium such that this distance is quadratic in the number of agents. This fact implies that quantization can drastically destroy consensus. However, consensus is guaranteed in some special cases, for instance, when the communication among the individuals is described by either a complete or a complete bipartite graph. These results are reported in details in [19], whereas the book chapter [66] puts them in the broader context of consensus-seeking dynamics with discontinuous right-hand side.

### 7.2.5. *Message-passing computation of harmonic influence in social networks*

**Participants:** W. S. Rossi, P. Frasca [Contact person].

In the study of networks, identifying the most important nodes is of capital importance. The concept of Harmonic Influence has been recently proposed as a metric for the importance of nodes in a social network. This metric evaluates the ability for one node to sway the opinions of the other nodes in the network, under the assumption of a linear diffusion of opinions in the network. A distributed message passing algorithm for its computation has been proposed by Vassio et al., 2014, but its convergence guarantees were limited to trees and regular graphs. In [36], we prove that the algorithm converges on general graphs. In [64], we offer two additional contributions to its study. We evaluate how the presence of communities in the network impacts the algorithm performance, and how the algorithm performs on networks which change topology during the execution of the algorithm.

### 7.2.6. *Distributed control and game theory: self-optimizing systems*

**Participants:** F. Garin [Contact person], B. Gaujal [POLARIS], S. Durand.

The design of distributed algorithms for a networked control system composed of multiple interacting agents, in order to drive the global system towards a desired optimal functioning, can benefit from tools and algorithms from game theory. This is the motivation of the Ph.D. thesis of Stéphane Durand, a collaboration between POLARIS and NECS teams.

The focus of this thesis is on the complexity of the best response algorithm to find a Nash equilibrium for potential games. Best response is a simple greedy algorithm, known to converge to a Nash equilibrium if players play one after the other in a round-robin way, but with a worst-case complexity which is exponential in the number of players. We consider instead its average complexity over the ensemble of random potential games, showing that such average complexity is surprisingly low, only linear in the number of players. Then we focus on removing the need of a centralised scheduler enforcing the round robin order of play. In [52], [21] we consider agents activated according to independent local Poisson clocks, and we show that (despite the possible overlaps of the computations of some players), we can still obtain convergence, with an average complexity of order $n \log n / \log \log n$, where $n$ is the number of players. In [51] we show how to take advantage of the structure of the interactions between players in a network game: noninteracting players can play simultaneously. This improves best response algorithm, both in the centralized and in the distributed case.

### *7.2.7. Control of switched interconnected large-scale systems*

**Participants:** H. Fourati [Contact person], D. Belkhiat, D. Jabri.

We proposed in [27] a new design of a decentralized output-feedback tracking control for a class of switched large-scale systems with external bounded disturbances. The controller proposed herein is synthesized to satisfy the robust $H_\infty$ tracking performance with local disturbance attenuation levels. Based on multiple switched Lyapunov functions, sufficient conditions proving the existence of the proposed controller are formulated in terms of Linear Matrix Inequalities (LMI).

## 7.3. Transportation networks and vehicular systems

### *7.3.1. Density and flow reconstruction in urban traffic networks*

**Participants:** C. Canudas de Wit [Contact person], H. Fourati, A. Kibangou, A. Ladino, M. Rodriguez-Vega.

In [56], we consider the problem of joint reconstruction of flow and density in a urban traffic network using heterogeneous sources of information. The traffic network is modeled within the framework of macroscopic traffic models, where we adopt Lighthill-Whitham-Richards model (LWR) conservation equation characterized by a piecewise linear fundamental diagram. The estimation problem considers two key principles. First, the error minimization between the measured and reconstructed flows and densities, and second the equilibrium state of the network which establishes flow propagation within the network. Both principles are integrated together with the traffic model constraints established by the supply/demand paradigm. Finally the problem is cast as a constrained quadratic optimization with equality constraints in order to shrink the feasible region of estimated variables. Some simulation scenarios based on synthetic data for a manhattan grid network are provided in order to validate the performance of the proposed algorithm.

In [62], we addressed the conditions imposed on the number and location of fixed sensors such that all flows in the network can be uniquely reconstructed. We determine the minimum number of sensors needed to solve the problem given partial information of turning ratios, and then we propose a linear time algorithm for their allocation in a network. Using these results in addition to floating car data, we propose a method to reconstruct all traffic density and flow.

### *7.3.2. Discrete-time system optimal dynamic traffic assignment (SO-DTA) with partial control for horizontal queuing networks*

**Participants:** S. Samaranayake, J. Reilly, W. Krichene, M. L. Delle Monache [Contact person], P. Goatin [Acumes, Inria], A. Bayen.

Dynamic traffic assignment (DTA) is the process of allocating time-varying origin-destination (OD) based traffic demand to a set of paths on a road network. There are two types of traffic assignment that are generally considered, the user equilibrium or Wardrop equilibrium allocation (UE-DTA), in which users minimize individual travel-time in a selfish manner, and the system optimal allocation (SODTA) where a central authority picks the route for each user and seeks to minimize the aggregate total travel-time over all users. It can be shown that the price of anarchy (PoA), the worst-case ratio of the system delay caused by the selfish behavior over the system optimal solution, may be arbitrarily large even in simple networks. System optimal (SO) traffic assignment on the other hand leads to optimal utilization of the network resources, but is hard to achieve in practice since the overriding objective for individual drivers in a road network is to minimize their own travel-time. It is well known that setting a toll on each road segment corresponding to the marginal delay of the demand moves the user equilibrium towards a SO allocation. In [37], we formulate the system optimal dynamic traffic assignment problem with partial control (SO-DTAPC), using a Godunov discretization of the Lighthill-Williams-Richards (LWR) partial differential equation (PDE) with a triangular flux function. We propose solving the SO-DTA-PC problem with the non-convex traffic dynamics and limited OD data with complete split ratios as a non-linear optimal control problem. This formulation generalizes to multiple sources and multiple destinations. We show that the structure of our dynamical system allows for very efficient computation of the gradient via the discrete adjoint method.

### *7.3.3. Priority-based Riemann solver for traffic flow on networks*

**Participants:** M. L. Delle Monache [Contact person], P. Goatin [Acumes, Inria], B. Piccoli.

In [20] we introduce a novel solver for traffic intersection which considers priorities among the incoming roads as the first criterion and maximization of flux as the second. The main idea is that the road with the highest priority will use the maximal flow taking into account also outgoing roads constraints. If some room is left for additional flow then the road with the second highest priority will use the left space and so on. A precise definition of the new Riemann solver, called Priority Riemann Solver, is based on a traffic distribution matrix , a priority vector and requires a recursion method. The general existence theorem for Riemann solvers on junctions can not be applied in the present case.Therefore, we achieve existence via a new set of general properties.

### *7.3.4. Dissipation of stop-and-go waves via control of autonomous vehicles*

**Participants:** R. Stern, S. Cui, M. L. Delle Monache [Contact person], R. Bhadani, M. Bunting, M. Churchill, N. Hamilton, R. Haulcy, H. Pohlmann, F. Wu, B. Piccoli, B. Seibold, J. Sprinkle, D. B. Work.

Traffic waves are phenomena that emerge when the vehicular density exceeds a critical threshold. Considering the presence of increasingly automated vehicles in the traffic stream, a number of research activities have focused on the influence of automated vehicles on the bulk traffic flow. In [38], we demonstrate experimentally that intelligent control of an autonomous vehicle is able to dampen stop-and-go waves that can arise even in the absence of geometric or lane changing triggers. Precisely, our experiments on a circular track with more than 20 vehicles show that traffic waves emerge consistently, and that they can be dampened by controlling the velocity of a single vehicle in the flow. We compare metrics for velocity, braking events, and fuel economy across experiments. These experimental findings suggest a paradigm shift in traffic management: flow control will be possible via a few mobile actuators (less than $5\%$) long before a majority of vehicles have autonomous capabilities.

### *7.3.5. Cooperative adaptive cruise control over unreliable networks*

**Participants:** F. Acciani, P. Frasca [Contact person], G. Heijenk, E. Semsar-Kazerooni, A. Stoorvogel.

Cooperative Adaptive Cruise Control (CACC) is a promising technique to increase highway throughput, safety and comfort for vehicles. Enabled by wireless communication, CACC allows a platoon of vehicles to achieve better performance than Adaptive Cruise Control; however, since wireless is employed, problems related to communication unreliability arise. In [45], we design a digital controller to achieve platoon stability, enhanced by an observer to increase robustness against packet losses. Our results confirms the interest of using an observer in combination with a local and cooperative digital controller.

### 7.3.6. Heterogeneity in synchronization: an adaptive control approach, with applications to vehicle platooning

**Participants:** S. Baldi, P. Frasca [Contact person].

Heterogeneity is a substantial obstacle to achieve synchronisation of interconnected systems (that is, in control ) In order to overcome heterogeneity, advanced control techniques are needed, such as the use of "internal models" or of adaptive techniques. In a series of papers motivated by multi-vehicle platooning and coordinated autonomous driving, we have explored the application of adaptive control techniques. Our results cover both the cases of state-feedback [15] and of output-feedback [16], under the assumption that the topology of the interconnections has no circuits. Further investigation has shown that restrictive assumption can be relaxed (at least for state-feedback on some specific topologies) [47]. This understanding paves the road to use these techniques not only to stabilise heterogeneous platoons, but also to manage their merging or splitting operations [48].

### 7.3.7. Modeling traffic on roundabout

**Participants:** M. L. Delle Monache [Contact person], A. Rat, S. Hammond, B. Piccoli.

In [50] we introduce a Riemann solver for traffic flow on a roundabout with two lanes. The roundabout is modeled as a sequence of junctions. The Riemann solver provides a solution at junctions by taking into consideration traffic distribution, priorities, and the maximization of through flux. We prove existence and uniqueness of the solution of the Riemann problem and show some results numerically. This work stems from the fact that there is a general notion among transportation professionals that having a longer additional lane length at a double-lane roundabout entry yields better performances. In [55], we investigate this notion using Lighthill-Whitham-Richards Model. Using Lighthill- Whitham-Richards model, a double-lane roundabout with additional lane design at the entry is analyzed. The additional lane lengths are varied at the entry in order to study the effect of different additional lane lengths on roundabout performance. The results obtained with the PDE model were then compared with similar lane length variations in VISSIM.

### 7.3.8. Two dimensional models for traffic

**Participants:** S. Mollier, M. L. Delle Monache, C. Canudas de Wit [Contact person], B. Seibold.

The work deals with the problem of modeling traffic flow in urban area, e. g. a town. More precisely, the goal is to design a two-dimensional macroscopic traffic flow model suitable to model large network as the one of a city. Macroscopic traffic models are inspired from fluid dynamic. They represent vehicles on the road by a density and describe their evolution with partial differential equations. Usually, these models are one dimensional models and, for instance, give a good representation of the evolution of traffic states in highway. The extension of these 1D models to a network is possible thanks to models of junction but can be tedious according to the number of parameters to fit. In the last few years, the idea of models based on a two dimensional conservation laws arose in order to represent traffic flow in large and dense networks. This study starts with a simple model [33] for homogeneous network and where a preferred direction of traffic exists. Our aim is to extend gradually this model by adding complexity. As this approach is uncommon, we investigate a way to compare the results of this model with microsimulation in [73] using Aimsun. Then, in the literature, the network is mainly assumed to be homogeneous. However, in a large-scale scenario, it is unlikely that the traffic network characteristics–such as speed limit, number of lanes, or the network geometry–remain constant throughout the network. Therefore, we introduce a first extension [59] where the fundamental diagram is space-dependent and varies with respect to the area considered. Finally, we have studied more recently a possible way to relax the assumption of a preferred direction of flow by considering several layers of density such that each layer describe a different direction of flow. In this case, the model becomes a system of conservation and is hyperbolic-elliptic which imply special caution in the choice of the numerical method.

## 7.4. Multisensor data fusion for navigation

### 7.4.1. Sensors fusion for attitude estimation

**Participants:** H. Fourati [Contact person], Z. Zhou, J. Wu.

Attitude estimation consists in the determination of rigid body orientation in 3D space (principally in terms of Euler angles, rotation matrix, or quaternion). As a key problem for multisensor attitude determination, Wahba's problem has been studied for almost 50 years. In [42], we present a novel linear approach to solve this problem. We name the proposed method the fast linear attitude estimator (FLAE) because it is faster than known representative algorithms. The original Wahba's problem is extracted to several 1-D equations based on quaternions. They are then investigated with pseudoinverse matrices establishing a linear solution to n-D equations, which are equivalent to the conventional Wahba's problem. To obtain the attitude quaternion in a robust manner, an eigenvalue-based solution is proposed. Symbolic solutions to the corresponding characteristic polynomial are derived, showing higher computation speed. Also, to verify the feasibility in embedded application, an experiment on the accelerometer–magnetometer combination is carried out where the algorithms are compared via C++ programming language. From other side, the integration of the Accelerometer and Magnetometer (AM) provides continuous, stable and accurate attitude information for land-vehicle navigation without magnetic distortion and external acceleration. However, magnetic disturbance and linear acceleration strongly degrade the overall system performance. As an important complement, the Global Navigation Satellite System (GNSS) produces the heading estimates, thus it can potentially benefit the AM system. Such a GNSS/AM system for attitude estimation is mathematically converted to a multi-observation vector pairs matching problem in [44]. The optimal and sub-optimal attitude determination and their time-varying recursive variants are all comprehensively investigated and discussed. The developed methods are named as the Optimal Linear Estimator of Quaternion (OLEQ), Suboptimal-OLEQ (SOLEQ) and Recursive-OLEQ (ROLEQ) for different application scenarios. The theory is established based on our previous contributions, and the multi-vector matrix multiplications are decomposed with the eigenvalue factorization. Some analytical results are proven and given, which provides the reader with a brand new viewpoint of the attitude determination and its evolution. With the derivations of the two-vector case, the n-vector case is then naturally formed. The algorithms are then implemented using the C++ programming language on the designed hardware with a GNSS module, three-axis accelerometer and three-axis magnetometer, giving an effective validation of them in real-world applications. In [39], a super fast attitude solution is obtained for consumer electronics accelerometer-magnetometer combination. The quaternion parameterizing the orientation is analytically derived from a least-square optimization that maintains very simple form. Like previously developed approaches, this algorithm does not require predetermined magnetometer reference vector. In [41], we present a novel sequential multiplicative quaternion attitude estimation method from various vector sensor outputs. The unique linear constitution of the algorithm leads to its specific name of Recursive Linear Quaternion Estimator (RLQE). The algorithm's architecture is designed to use each single pair of vector observation linearly so that the vector observations can be arbitrarily chosen and fused. The closed-form covariance of the RLQE is derived that builds up the existence of a highly reliable RLQE Kalman filter (RLQE-KF). In [65], to generate the virtual-gyro output in the case of gyroscope failures, virtual-gyro Kalman filter is established for angular rate estimation base on attitude estimation results.

### *7.4.2. Attitude estimation applied in augmented reality*

**Participants:** H. Fourati [Contact person], T. Michel, P. Genevès, N. Layaïda.

We investigate the precision of attitude estimation algorithms in the particular context of pedestrian navigation with commodity smartphones and their inertial/magnetic sensors. A particular attention was paid to the study of attitude estimation in the context of augmented reality motions when using smartphones [32]. We report on an extensive comparison and experimental analysis of existing algorithms. We focus on typical motions of smartphones when carried by pedestrians. We use a precise ground truth obtained from a motion capture system. We test state-of-the-art and built-in attitude estimation techniques with several smartphones, in the presence of magnetic perturbations typically found in buildings. We discuss the obtained results, analyze advantages and limits of current technologies for attitude estimation in this context. Furthermore, we propose a new technique for limiting the impact of magnetic perturbations with any attitude estimation algorithm used in this context.

### *7.4.3. Attitude determination for satellite*

**Participants:** H. Fourati [Contact person], S. Pourtakdoust, Csug Team, E. Kerstel.

Recently, we started to work on attitude estimation for satellites. In [29], we are focused on the development and verification of a heat attitude model (HAM) for satellite attitude determination. Within this context, the Sun and the Earth are considered as the main external sources of radiation that could affect the satellite surface temperature changes. Assuming that the satellite orbital position (navigational data) is known, the proposed HAM provides the satellite surface temperature with acceptable accuracy and also relates the net heat flux (NHF) of three orthogonal satellite surfaces to its attitude via the inertial to satellite transformation matrix. The proposed HAM simulation results are verified through comparison with commercial thermal analysis tools. The proposed HAM has been successfully utilized in some researches for attitude estimation, and further studies for practical implementations are still ongoing. Actually, we are establishing a project around quantum communication experiments under Nanobob CubeSat mission [28]. Some attitude estimation algorithms will be deployed to orient the satellite to the ground station.

### *7.4.4. Sensors fusion for distance measurement in pedestrian navigation*

**Participants:** H. Fourati [Contact person], Z. Zhou, J. Wu.

We developed in [43] a foot-mounted pedestrian navigation system prototype with the emphasis on distance measuring with an inertial measurement unit (IMU) which implies the characteristics of pedestrian gait cycle and thus can be used as a crucial step indicator for distance calculation. An adaptive time- and frequency-domains joint distance measuring method is proposed by utilizing the means of behaviors classification. Two key issues are studied: step detection and step length determination. For the step detection part, first behavior classification along with state transition strategy is designed to identify typical pedestrian behaviors including standing still, walking, running and irregular swing. Then a four-stage step detection method is proposed to adaptively determine both step frequency and threshold in a flexible window. Based on the behavior classification results, a two-segment functional based step length model is established to adapt the walking and running behaviors.

# NUMED Project-Team (section vide)

**PERCEPTION Project-Team**

# 6. New Results

## 6.1. Multichannel Speech Separation and Enhancement Using the Convolutive Transfer Function

We addressed the problem of speech separation and enhancement from multichannel convolutive and noisy mixtures, *assuming known mixing filters*. We proposed to perform the speech separation and enhancement tasks in the short-time Fourier transform domain, using the convolutive transfer function (CTF) approximation [39]. Compared to time-domain filters, CTF has much less taps, consequently it has less near-common zeros among channels and less computational complexity. The work proposes three speech-source recovery methods, namely: (i) the multichannel inverse filtering method, i.e. the multiple input/output inverse theorem (MINT), is exploited in the CTF domain, and for the multi-source case, (ii) a beamforming-like multichannel inverse filtering method applying single source MINT and using power minimization, which is suitable whenever the source CTFs are not all known, and (iii) a constrained Lasso method, where the sources are recovered by minimizing the $\ell_1$-norm to impose their spectral sparsity, with the constraint that the $\ell_2$-norm fitting cost, between the microphone signals and the mixing model involving the unknown source signals, is less than a tolerance. The noise can be reduced by setting a tolerance onto the noise power. Experiments under various acoustic conditions are carried out to evaluate the three proposed methods. The comparison between them as well as with the baseline methods is presented.

## 6.2. Speech Dereverberation and Noise Reduction Using the Convolutive Transfer Function

We address the problems of blind multichannel identification and equalization for *joint speech dereverberation and noise reduction*. The standard time-domain cross-relation methods are hardly applicable for blind room impulse response identification due to the near-common zeros of the long impulse responses. We extend the cross-relation formulation to the short-time Fourier transform (STFT) domain, in which the time-domain impulse response is approximately represented by the convolutive transfer function (CTF) with much less coefficients. For the oversampled STFT, CTFs suffer from the common zeros caused by the non-flat-top STFT window. To overcome this, we propose to identify CTFs using the STFT framework with oversampled signals and critically sampled CTFs, which is a good trade-off between the frequency aliasing of the signals and the common zeros problem of CTFs. The phases of the identified CTFs are inaccurate due to the frequency aliasing of the CTFs, and thus only their magnitudes are used. This leads to a non-negative multichannel equalization method based on a non-negative convolution model between the STFT magnitude of the source signal and the CTF magnitude. To recover the STFT magnitude of the source signal and to reduce the additive noise, the $\ell_2$-norm fitting error between the STFT magnitude of the microphone signals and the non-negative convolution is constrained to be less than a noise power related tolerance. Meanwhile, the $\ell_1$-norm of the STFT magnitude of the source signal is minimized to impose the sparsity [38].
Website: https://team.inria.fr/perception/research/ctf-dereverberation/.

## 6.3. Speech Enhancement with a Variational Auto-Encoder

We addressed the problem of enhancing speech signals in noisy mixtures using a source separation approach. We explored the use of neural networks as an alternative to a popular speech variance model based on supervised non-negative matrix factorization (NMF). More precisely, we use a variational auto-encoder as a speaker-independent supervised generative speech model, highlighting the conceptual similarities that this approach shares with its NMF-based counterpart. In order to be free of generalization issues regarding the noisy recording environments, we follow the approach of having a supervised model only for the target

speech signal, the noise model being based on unsupervised NMF. We developed a Monte Carlo expectation-maximization algorithm for inferring the latent variables in the variational auto-encoder and estimating the unsupervised model parameters. Experiments show that the proposed method outperforms a semi-supervised NMF baseline and a state-of-the-art fully supervised deep learning approach.
Website: https://team.inria.fr/perception/research/ieee-mlsp-2018/.

## 6.4. Audio-Visual Speaker Tracking and Diarization

We are particularly interested in modeling the interaction between an intelligent device and a group of people. For that purpose we develop audio-visual person tracking methods [36]. As the observed persons are supposed to carry out a conversation, we also include speaker diarization into our tracking methodology. We cast the diarization problem into a tracking formulation whereby the active speaker is detected and tracked over time. A probabilistic tracker exploits the spatial coincidence of visual and auditory observations and infers a single latent variable which represents the identity of the active speaker. Visual and auditory observations are fused using our recently developed weighted-data mixture model [12], while several options for the speaking turns dynamics are fulfilled by a multi-case transition model. The modules that translate raw audio and visual data into image observations are also described in detail. The performance of the proposed method are tested on challenging datasets that are available from recent contributions which are used as baselines for comparison [36].
Websites:
https://team.inria.fr/perception/research/wdgmm/,
https://team.inria.fr/perception/research/speakerloc/,
https://team.inria.fr/perception/research/speechturndet/, and
https://team.inria.fr/perception/research/avdiarization/.

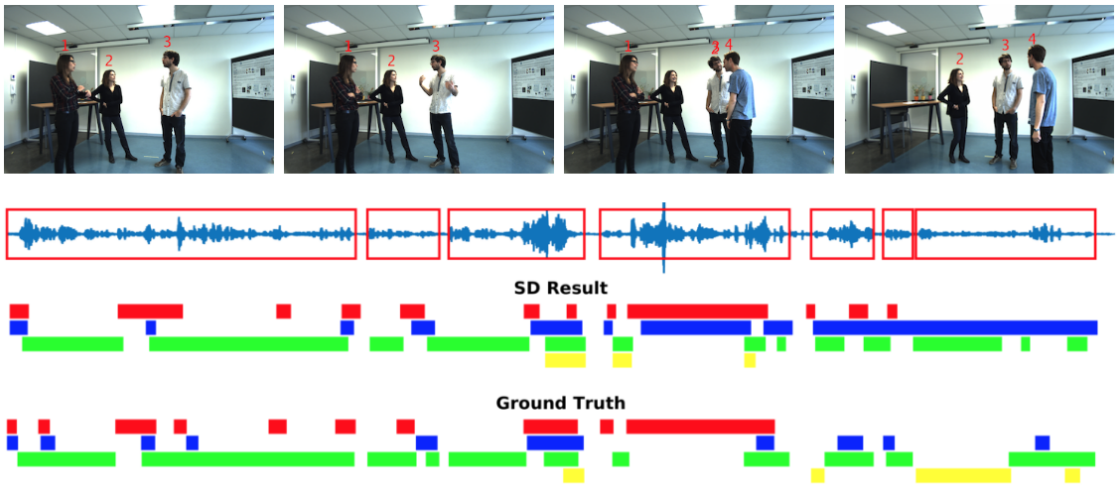



*Figure 3. This figure illustrates the audiovisual tracking and diarization method that we have recently developed. First row: A number is associated with each tracked person. Second row: diarization result. Third row: the ground truth diarization. Fourth row: acoustic signal recorded by one of the two microphones.*

## 6.5. Tracking Eye Gaze and of Visual Focus of Attention

The visual focus of attention (VFOA) has been recognized as a prominent conversational cue. We are interested in estimating and tracking the VFOAs associated with multi-party social interactions. We note that in this type of situations the participants either look at each other or at an object of interest; therefore their eyes are not always visible. Consequently both gaze and VFOA estimation cannot be based on eye detection and tracking. We propose a method that exploits the correlation between eye gaze and head movements. Both VFOA and gaze are modeled as latent variables in a Bayesian switching state-space model (also named switching Kalman filter). The proposed formulation leads to a tractable learning method and to an efficient online inference procedure that simultaneously tracks gaze and visual focus. The method is tested and benchmarked using two publicly available datasets, Vernissage and LAEO, that contain typical multi-party human-robot and human-human interactions [42].
Website: https://team.inria.fr/perception/research/eye-gaze/.

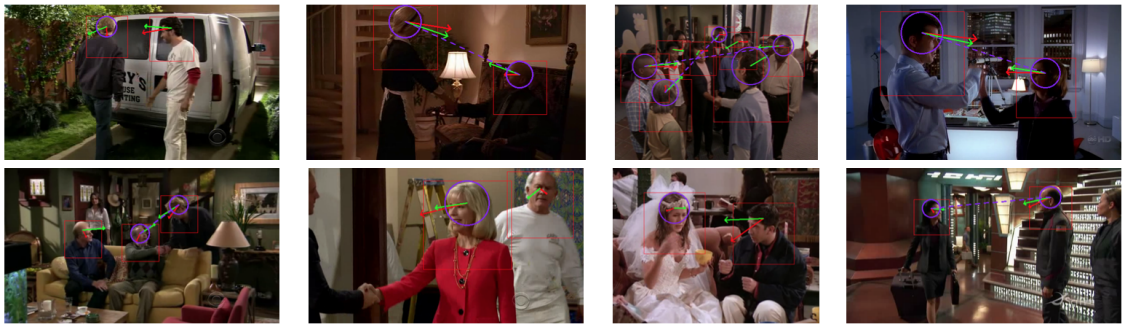


*Figure 4. This figure shows some results obtained with the LAEO dataset. The top row shows results obtained with coarse head orientation and the bottom row shows results obtained with fine head orientation. Head orientations are shown with red arrows. The algorithm infers gaze directions (green arrows) and VFOAs (blue circles). People looking at each others are shown with a dashed blue line.*

## 6.6. Variational Bayesian Inference of Multiple-Person Tracking

We addressed the problem of tracking multiple speakers using audio information or via the fusion of visual and auditory information. We proposed to exploit the complementary nature of these two modalities in order to accurately estimate smooth trajectories of the tracked persons, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status – either speaking or silent – of each tracked person along time, e.g. Figure 1 . We proposed to cast the problem at hand into a generative audio-visual fusion (or association) model formulated as a latent-variable temporal graphical model. This may well be viewed as the problem of maximizing the posterior joint distribution of a set of continuous and discrete latent variables given the past and current observations, which is intractable. We propose a variational inference model which amounts to approximate the joint distribution with a factorized distribution. The solutions take the form of closed-form expectation maximization procedures using Gaussian distributions [44], [58], [56] or the von Mises distribution for circular variables [55]. We described in detail the inference algorithms, we evaluate their performance and we compared them with several baseline methods. These experiments show that the proposed audio and audio-visual trackers perform well in informal meetings involving a time-varying number of people.
Websites:
https://team.inria.fr/perception/research/var-av-track/,

https://team.inria.fr/perception/research/audiotrack-vonm/.

## 6.7. High-Dimensional and Deep Regression

One of the most important achievements for the last years has been the development of high-dimensional to low-dimensional regression methods. The motivation for investigating this problem raised from several problems that appeared both in audio signal processing and in computer vision. Indeed, often the task in data-driven methods is to recover low-dimensional properties and associated parameterizations from high-dimensional observations. Traditionally, this can be formulated as either an unsupervised method (dimensionality reduction of manifold learning) or a supervised method (regression). We developed a learning methodology at the crossroads of these two alternatives: the output variable can be either fully observed or partially observed. This was cast into the framework of linear-Gaussian mixture models in conjunction with the concept of inverse regression. It gave rise to several closed-form and approximate inference algorithms [8]. The method is referred to as *Gaussian locally linear mapping*, or GLLiM. As already mentioned, high-dimensional regression is useful in a number of data processing tasks because the sensory data often lies in high-dimensional spaces. Each one of these tasks required a special-purpose version of our general framework. Sound-source localization was the first to benefit from our formulation. Nevertheless, the sparse nature of speech spectrograms required the development of a GLLiM version that is able to with full-spectrum sounds and to test with sparse-spectrum ones [9]. This could be immediately applied to audio-visual alignment and to sound-source separation and localization [7].

In conjunction with our computer vision work, high-dimensional regression is a very useful methodology since visual features, obtained either by hand-crafted feature extraction methods or using convolutional neural networks, lie in high-dimensional spaces. Such properties as object pose lie in low-dimensional spaces and must be extracted from features. We took such an approach and proposed a head pose estimator [10]. Visual tracking can also benefit from GLLiM. Indeed, it is not practical to track objects based on high-dimensional features. We therefore combined GLLiM with switching linear dynamic systems. In 2018 we proposed a robust deep regression method [46]. In parallel we thoroughly benchmarked and analyzed deep regression tasks using several CNN architectures [57].

## 6.8. Human-Robot Interaction

Audio-visual fusion raises interesting problems whenever it is implemented onto a robot. Robotic platforms have their own hardware and software constraints. In addition, commercialized robots have economical constraints which leads to the use of cheap components. A robot must be reactive to changes in its environment and hence it must take fast decisions. This often implies that most of the computing resources must be onboard of the robot.

Over the last decade we have tried to do our best to take these constraints into account. Starting from our scientific developments, we put a lot of efforts into robotics implementations. For example, the audio-visual fusion method described in [2] used a specific robotic middleware that allowed fast communication between the robot and an external computing unit. Subsequently we developed a powerful software package that enables distributed computing. We also put a lot of emphasis on the implementation of low-level audio and visual processing algorithms. In particular, our single- and multiple audio source methods were implemented in real time onto the humanoid robot NAO [25], [50]. The multiple person tracker [4] was also implemented onto our robotic platforms [5], e.g. Figure 5 .

More recently, we investigated the use of reinforcement learning (RL) as an alternative to sensor-based robot control [45], [37]. The robotic task consists of turning the robot head (gaze control) towards speaking people. The method is more general in spirit than visual (or audio) servoing because it can handle an arbitrary number of speaking or non speaking persons and it can improve its behavior online, as the robot experiences new situations. An overview of the proposed method is shown in Fig. 6 . The reinforcement learning formulation enables a robot to learn where to look for people and to favor speaking people via a trial-and-error strategy.

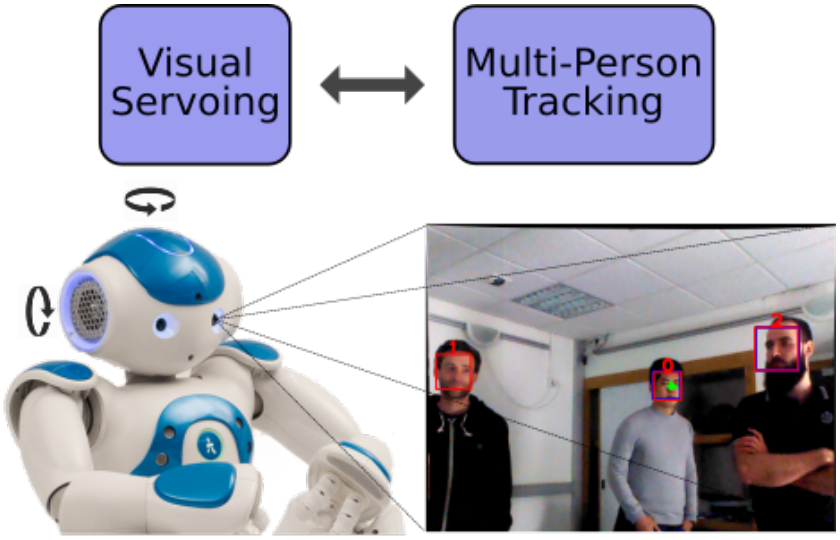



*Figure 5. The multi-person tracking method is combined with a visual servoing module. The latter estimates the optimal robot commands and the expected impact of the tracked person locations. The multi-person tracking module refines the locations of the persons with the new observations and the information provided by the visual servoing.*

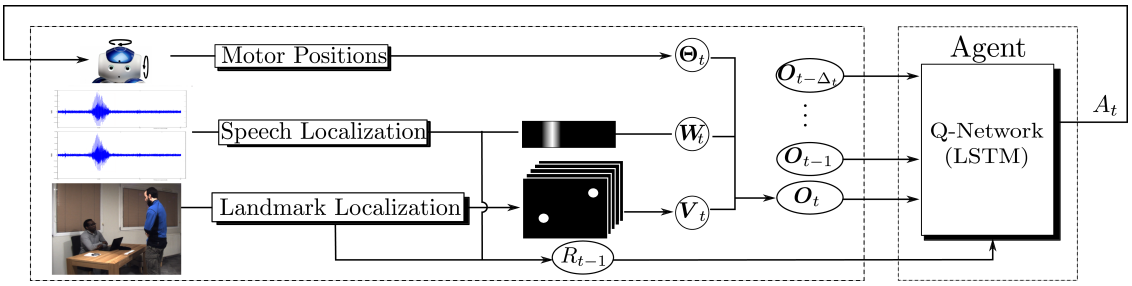



*Figure 6. Overview of the proposed deep RL method for controlling the gaze of a robot. At each time index $t$, audio and visual data are represented as binary maps which, together with motor positions, form the set of observations $\mathbf{O}_t$. A motor action $A_t$ (rotate the head left, right, up, down, or stay still) is selected based on past and present observations via maximization of current and future rewards. The rewards $R$ are based on the number of visible persons as well as on the presence of speech sources in the camera field of view. We use a deep Q-network (DQN) model that can be learned both off-line and on-line. Please consult [45], [37] for further details.*

Past, present and future HRI developments require datasets for training, validation, test as well as for benchmarking. HRI datasets are challenging because it is not easy to record realistic interactions between a robot and users. RL avoids systematic recourse to annotated datasets for training. In [45], [37] we proposed the use of a simulated environment for pre-training the RL parameters, thus avoiding spending hours of tedious interaction.
Websites:
https://team.inria.fr/perception/research/deep-rl-for-gaze-control/,
https://team.inria.fr/perception/research/mot-servoing/.

## 6.9. Generation of Diverse Behavioral Data

We target the automatic generation of visual data depicting human behavior, and in particular how to design a method able to learn the generation of *data diversity*. In particular, we focus on smiles, because each smile is unique: one person surely smiles in different ways (e.g. closing/opening the eyes or mouth). We wonder if given one input image of a neutral face, we can generate multiple smile videos with distinctive characteristics. To tackle this one-to-many video generation problem, we propose a novel deep learning architecture named Conditional MultiMode Network (CMM-Net). To better encode the dynamics of facial expressions, CMM-Net explicitly exploits facial landmarks for generating smile sequences. Specifically, a variational auto-encoder is used to learn a facial landmark embedding. This single embedding is then exploited by a conditional recurrent network which generates a landmark embedding sequence conditioned on a specific expression (e.g. spontaneous smile), implemented as a Conditional LSTM. Next, the generated landmark embeddings are fed into a multi-mode recurrent landmark generator, producing a set of landmark sequences still associated to the given smile class but clearly distinct from each other, we call that a Multi-Mode LSTM. Finally, these landmark sequences are translated into face videos. Our experimental results, see Figure 7 , demonstrate the effectiveness of our CMM-Net in generating realistic videos of multiple smile expressions [52].

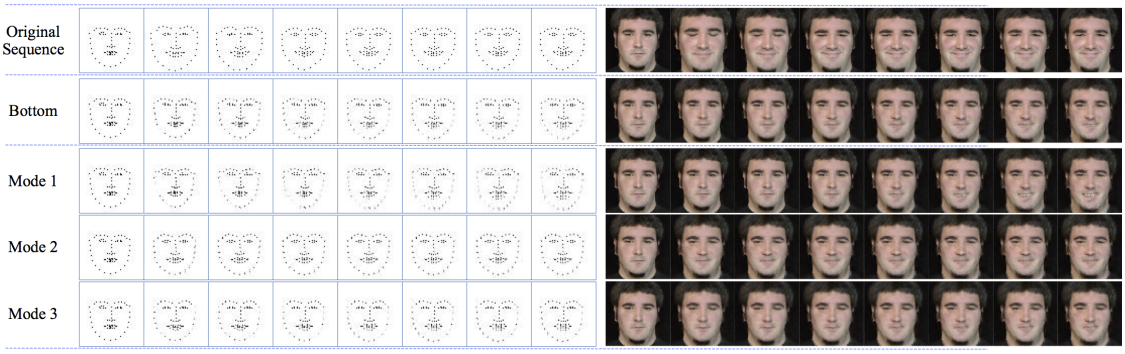



*Figure 7. Multi-mode generation example with a sequence: landmarks (left) and associated face images (right) after the landmark-to-image decoding step based on Variational Auto-Encoders. The rows correspond to the original sequence (first), output of the Conditional LSTM (second), and output of the Multi-Mode LSTM (last three rows).*

## 6.10. Registration of Multiple Point Sets

We have also addressed the rigid registration problem of multiple 3D point sets. While the vast majority of state-of-the-art techniques build on pairwise registration, we proposed a generative model that explains jointly registered multiple sets: back-transformed points are considered realizations of a single Gaussian mixture model (GMM) whose means play the role of the (unknown) scene points. Under this assumption, the joint registration problem is cast into a probabilistic clustering framework. We formally derive an expectation-maximization procedure that robustly estimates both the GMM parameters and the rigid transformations that map each individual cloud onto an under-construction reference set, that is, the GMM means. GMM variances carry rich information as well, thus leading to a noise- and outlier-free scene model as a by-product. A second version of the algorithm is also proposed whereby newly captured sets can be registered online. A thorough discussion and validation on challenging data-sets against several state-of-the-art methods confirm the potential of the proposed model for jointly registering real depth data [35].
Website: https://team.inria.fr/perception/research/jrmpc/

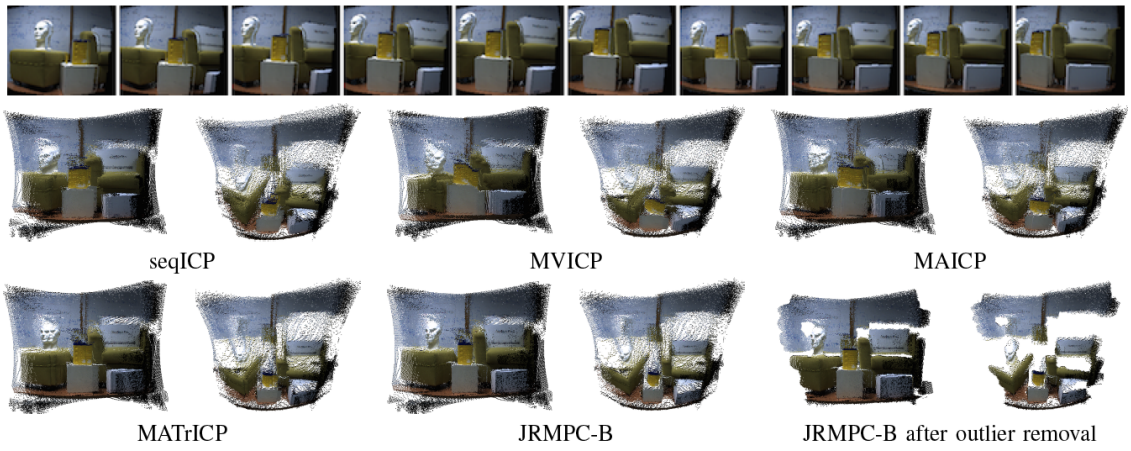



*Figure 8. Integrated point clouds from the joint registration of 10 TOF images that record a static scene (EXBI data-set). Top: color images that roughly show the scene content of each range image (occlusions due to cameras baseline may cause texture artefacts). Bottom: front-view and top-view of integrated sets after joint registration. The results obtained with the proposed method (JRMPC-B) are compared with several other methods.*

**PERVASIVE Project-Team**

# 7. New Results

## 7.1. Using Attention to Address Human-Robot Motion

**Participants:** Thierry Fraichard, Rémi Paulin, Patrick Reignier.

To capture the specificity of robot motion among people, we choose the term **Human-Robot Motion** (HRM) [0], to denote the study of how robots should move among people. HRM is about designing robots whose motions are deemed socially **acceptable** from a human point of view while remaining **safe**.

After 15 years of research on HRM, the main concept that has emerged is that of *social spaces*, *i.e.* regions of the environment that people consider as psychologically theirs [33], any intrusion in their social space will be a source of discomfort. Such social spaces are characterized by the position of the person, *i.e.* "Personal Space", or the activity they are currently engaged in, *i.e.* "Interaction Space" and "Activity Space". The most common approach in HRM is to define costmaps on such social spaces: the higher the cost, the less desirable it is to be there. The costmaps are then used for navigation purposes, *e.g.* [37] and [36].

Social spaces are of course relevant to HRM but they have limitations. First, it is not straightforward to define them; what is their shape or size, especially in cluttered environments? Second, it seems obvious that there is more to acceptability than geometry only: the appearance of a robot and its velocity will also influence the way it is perceived by people. Finally, social spaces can be conflicting because when a robot needs to interact with a person, it is very likely that it will have to penetrate a social space.

To complement social spaces, we have started to explore whether human attention could be useful to address HRM vis-à-vis the acceptability aspect. Why attention? The answer is straightforward: the acceptability of a robot motion is directly related to the way it is perceived by a person hence our interest in human attention. For a person, attention is a cognitive mechanism for filtering the person's sensory information (to avoid an overwhelming amount of information) [35]. It controls where and to what the person's attentional resources are allocated.

In 2014, we introduced the concept of **attention field**, *i.e.* a predictor of the amount of attention that a person allocates to the robot when the robot is in a given state. In [32], the attention field was computed thanks to a computational model of attention proposed in [34] in the context of ambient applications and pervasive systems. In this model, attentional resources are focused on a single specific area of the person's visual space (as per the zoom lens model [31]). Later studies have demonstrated that the situation is more complex and that attentional resources can be distributed over multiple objects in the visual space [35].

In 2018, we have developed a novel **computational model of attention** that takes this property into account. This model is used to compute the attention field for a robot. The attention field is then used to define different **attentional properties** for the robot's motions such as distraction or surprise. The relevance of the attentional properties for HRM have been demonstrated on a proof-of-concept **acceptable motion planner** on various case studies where a robot is assigned different tasks. The multi-criteria nature of motion planning in the context of HRM led to the design of an acceptable motion planner based upon a state-of-the-art many-objective optimization algorithm. It shows how to compute acceptable motions that are non-distracting and non-surprising, but also motions that convey the robot's intention to interact with a person. All these contributions have been presented in the PhD of Rémi Paulin [6] and the conference article [26].

## 7.2. Simulating Haptic Sensations

**Participants**: Jingtao Chen, Sabine Coquillart

[0] In reference to Human-Robot Interaction (HRI), *i.e.* the study of the interactions, in the broad sense of the word, between people and robots.

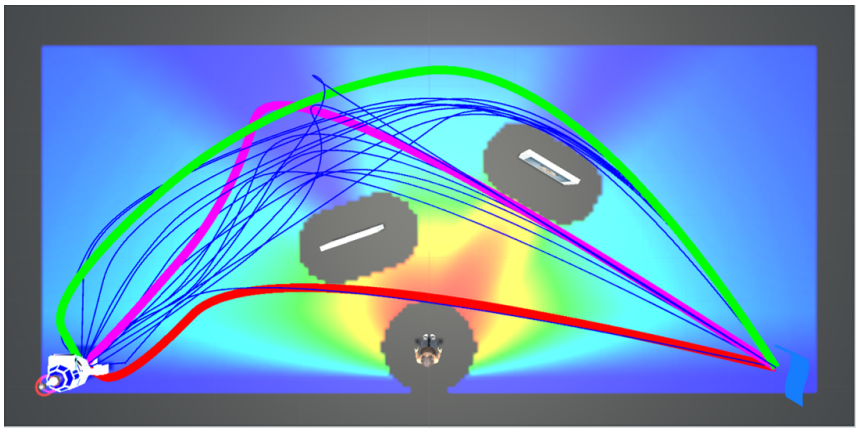


*Figure 1. Motions with different attentional properties in a scenario featuring a person watching at paintings in a museum and a robot which is to travel from left to right: less distracting (green) vs. shortest (red) motions are depicted. The purple motion is a trade-off solution.*

**Partners**: Inria GRA, LIG, GIPSA, G-SCOP

Pseudo-haptic feedback is a technique aiming to simulate haptic sensations without active haptic feedback devices. Peudo-haptic techniques have been used to simulate various haptic feedbacks such as stiffness, torques, and mass. In the framework of the Persyval project, a novel pseudo-haptic experiment has been set up. The aim of this experiment is to study the force and EMG signals during a pseudo-haptic task. A stiffness discrimination task similar to the one published in Lecuyer's PhD thesis has been chosen. The experimental set-up has been developed, as well as the software controlling the experiment. Pre-tests have been conducted. They have been followed by formal tests with subjects.

## 7.3. Observing and Modeling Awareness and Expertise During Problem Solving

**Participants**: Thomas Guntz, Dominique Vaufreydaz, James Crowley, Philippe Dessus, Raffaella Balzarini

### *7.3.1. Observing and Modelling Competence and Awareness from Eye-gaze and Emotion*

We have constructed an instrument for capturing and interpreting multimodal signals of humans engaged in solving challenging problems. Our instrument captures eye gaze, fixations, body postures, and facial expressions signals from humans engaged in interactive tasks on a touch screen. We use a 23 inch Touch-Screen computer, a Kinect 2.0 mounted 35 cm above the screen to observe the subject, a 1080p Webcam for a frontal view, a Tobii Eye-Tracking bar (Pro X2-60 screen-based) and two adjustable USB-LED for lighting condition control. A wooden structure is used to rigidly mount the measuring equipment in order to assure identical sensor placement and orientation for all recordings.

As a pilot study, we observed expert chess players engaged in solving problems of increasing difficulty [Guntz et al 18a]. Our initial hypothesis was that we could directly detect awareness of significant configurations of chess pieces (chunks) from eye-scan and physiological measurements of emotion in reaction to game situation. The pilot experiment demonstrated that this initial hypothesis was overly simplistic.

In order to better understand the phenomena observed in our pilot experiment, we have constructed a model of the cognitive processes involved, using theories from cognitive science and classic (symbolic) artificial intelligence. This model is a very partial description that allows us to ask questions and make predictions

to guide future experiments. Our model posits that experts reason with a situation model that is strongly constrained by limits to the number of entities and relations that may be considered at a time. This limitation forces subjects to construct abstract concepts (chunks) to describe game play, in order to explore alternative moves. Expert players retain associations of situations with emotions in long-term memory. The rapid changes in emotion correspond to recognition of previously encountered situations during exploration of the game tree. Recalled emotions guide selection of situation models for reasoning. This hypothesis is in accordance with Damasio's Somatic Marker hypothesis, which posits that emotions guide behavior, particularly when cognitive processes are overloaded [Damasio 91].

Our hypothesis is that the subject uses the evoked emotions to select from the many possible situations for reasoning about moves during orientation and exploration. With this interpretation, the player rapidly considers partial descriptions as situations composed of a limited number of perceived chunks. Recognition of situations from experience evokes emotions that are displayed as face expressions and body posture.

With this hypothesis, valence, arousal and dominance are learned from experience and associated with chess situations in long-term memory to guide reasoning in chess. Dominance corresponds to the degree of experience with the recognized situation. As players gain experience with alternate outcomes for a situation, they become more assured in their ability to spot opportunities and avoid dangers. Valence corresponds to whether the situation is recognized as favorable (providing opportunities) or unfavorable (creating threats). Arousal corresponds to the imminence of a threat or opportunity. A defensive player will give priority to reasoning about unfavorable situations and associated dangers. An aggressive player will seek out high valence situations. All players will give priority to situations that evoke strong arousal. The amount of effort that player will expend exploring a situation can determined by dominance.

In 2019 we will conduct an additional experiment designed confirm and explore this hypothesis. Results will be reported in a journal paper (under preparation) as well as in the doctoral thesis of Thomas Guntz, to be defended in late 2019.

### *7.3.2. Bibliography*

[Damasio 91] Damasio, A., Somatic Markers and the Guidance of Behavior. New York: Oxford University Press. pp. 217–299, 1991.

[Guntz et al. 18a] T. Guntz, R. Balzarini, D. Vaufreydaz, and J.L. Crowley, "Multimodal Observation and Classification of People Engaged in Problem Solving: Application to Chess Players". Multimodal Technologies and Interaction, Vol 2 No. 2, p11, 2018.

[Guntz et al. 18b] T. Guntz, J.L. Crowley, D. Vaufreydaz, R. Balzarini, P. Dessus, The Role of Emotion in Problem Solving: first results from observing chess, Workshop on Modeling Cognitive Processes from Multimodal Data, at the 2018 ACM International Conference on Multimodal Interaction, ICMI 2018, Oct 2018.

## 7.4. Learning Routine Patterns of Activity in the Home

**Participants**: Julien Cumin, James Crowley

**Other Partners**: Fano Ramparany, Greg Lefevre (Orange Labs)

During the month of February 2017, we have collected 4 weeks of data on daily activities within the Amiqual4Home Smart Home Living lab apartment. This dataset was presented at the international Conference on Ubiquitous Computing and Ambient Intelligence, UCAmI 2017, at Bethlehem PA, in Nov 2017 and is currently available for download from the Amiqual4Home web server (http://amiqual4home.inria.fr/en/orange4home/)

The objective of this research action is to develop a scalable approach to learning routine patterns of activity in a home using situation models. Information about user actions is used to construct situation models in which key elements are semantic time, place, social role, and actions. Activities are encoded as sequences of situations. Recurrent activities are detected as sequences of activities that occur at a specific time and place

each day. Recurrent activities provide routines that can be used to predict future actions and anticipate needs and services. An early demonstration has been to construct an intelligent assistant that can respond to and filter inter-personal communications.

## 7.5. Bayesian Reasoning

**Participants:** Emmanuel Mazer, Raphael Frisch, Marvin Faix, Augustin Lux, Didier Piau, Jeremy Belot.

To overcome the ever growing needs in computing power, alternative computing paradigms have been developed such as stochastic architectures. These latter have found substantial interests for energy efficient implementations in artificial intelligence. In particular, mixing stochastic computing with Bayesian models makes a promising paradigm for non-conventional computational architectures dedicated to Bayesian inference. The ability to deal with uncertainty and adapt its computational accuracy is some of the advantages of these computing approaches.

During 2018 we have designed a first hardware prototype to localize a sound source with a stochastic machine. The goal of this project was to provide a proof of concept of stochastic machines by implementing an autonomous platform of sound source localization. It includes an sound acquisition module, a pre-processing circuit, and the stochastic machine. The platform has been implemented on an Altera Cyclone V FPGA and validated functionally with digital simulations. Several optimization to improve size and power consumption have been proposed. Results in terms of computation time, power and used FPGA resources allowed to assess their impact on future design. The same architecture of stochastic machine was also analyzed in simulation to provide design guidelines for our next design [25].

Further, we have proposed a way to reduce the memory needs of our architecture by sharing a memory between the processing units (in collaboration with TIMA and C2M -Université Paris Sud ). This optimization reduces the area and the cost of our architecture. However, its impact on power consumption is not obvious. Therefore, we designed an integrated circuit (ASIC) with our original and optimized proposals. We synthesized the VHDL description of the circuit in the FDSOI 28nm technology from STMicroelectronics. Notice that the memory has been implemented thanks to a SRAM memory compiler. The results highlight that the optimized machine significantly reduces both the circuit area (by 30% ) and the power consumption (by 35% ). Nevertheless, the simulations showed that, in the optimized version, the memory represents nearly 60 % of our circuit area and more than 55% of the power consumption. According to the latest literature, the Magnetic Random Access Memory (MRAM) technology provides some promising features and would approximately reduce by a factor of 20 the memory area. Moreover, this feature should drastically impact the power consumption. Thus, our future works will focus on the implementation of Bayesian machines using MRAM instead of SRAM. A poster describing this work was presented at the International Conference on rebooting Computing.

We have proposed (in collaboration with ISIR - Université Paris Sorbonne)a new way to localize several sound sources using a Bayesian model. This multi-source localization algorithm is fast and can readily be implemented on our stochastic machine (Paper submitted at ICASSP 2019). The Figure 2 shows the location of the source and of the microphones in the simulated environment. The Figure 3 shows the posterior distribution of the location of one source using a short frame and the Figure 4 shows the result using fifty frames. As the frame are very short the localization of the two sources is readily obtained and it is used as a bootstrap for the source separation algorithm .

We devised and successfully tested a Bayesian model for the source separation problem. The model assumes the localization of the sources are known. The inference - retrieving the sound emitted by each source from the mixed signals obtained with several microphones - takes place in a very high dimensional space. Nevertheless, the Gibbs algorithm is well suited to solve the problem when the location of the sources are known. A very efficient implementation of this algorithm was tested with a realistic sound simulator using human voices. The algorithm can be implemented on a sampling machine and the corresponding stochastic architecture has been devised. It is currently implemented on an FPGA.

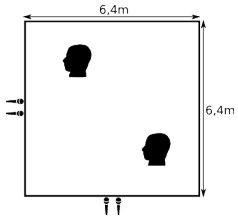
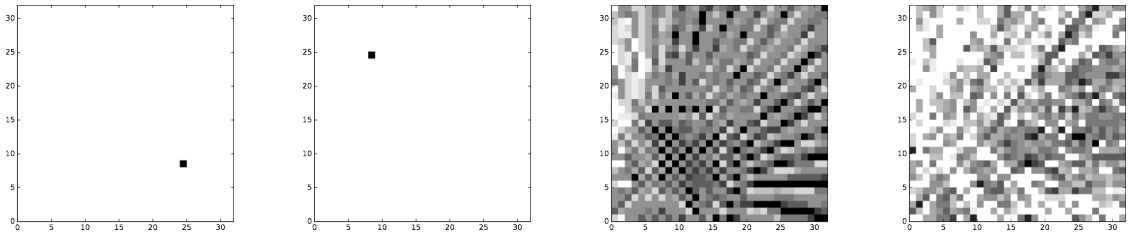



*Figure 2. Simulated room setup.*



(a) Frame 1 (b) Frame 3 (c) Frame 29 (d) Frame 46

*Figure 3. Posterior distribution maps for a single source obtained for 4 very short time-frames of a given 50-frame bloc.*

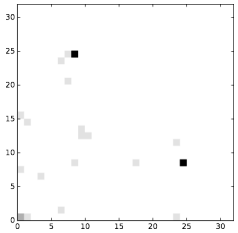


*Figure 4. Final distribution map after fusion over 50 frames. The two black squares correspond to the actual positions of the two sources.*

**POLARIS Project-Team**

# 7. New Results

## 7.1. Design of Experiments

A large amount of resources is spent writing, porting, and optimizing scientific and industrial High Performance Computing applications, which makes autotuning techniques fundamental to lower the cost of leveraging the improvements on execution time and power consumption provided by the latest software and hardware platforms. Despite the need for economy, most autotuning techniques still require a large budget of costly experimental measurements to provide good results, while rarely providing exploitable knowledge after optimization. In [40], we present a user-transparent (white-box) autotuning technique based on Design of Experiments that operates under tight budget constraints by significantly reducing the measurements needed to find good optimizations. Our approach enables users to make informed decisions on which optimizations to pursue and when to stop. We present an experimental evaluation of our approach and show it is capable of leveraging user decisions to find the best global configuration of a GPU Laplacian kernel using half of the measurement budget used by other common autotuning techniques. We show that our approach is also capable of finding speedups of up to $50\times$, compared to gcc's-O3, for some kernels from the SPAPT benchmark suite, using up to $10\times$ less measurements than random sampling.

## 7.2. Experimenting with Fog Infrastructures

To this day, the Internet of Things (IoT) continues its explosive growth. Nevertheless, with the exceptional evolution of traffic demand, existing infrastructures are struggling to resist. In this context, Fog computing is shaping the future of IoT applications. Fog computing provides computing, storage and communication resources at the edge of the network, near the physical world. This section describes two independent contributions on how to study and develop FOG infrastructures. These contributions take place in the context of the Inria/Orange Labs joint laboratory.

- Despite its several advantages, Fog computing raises new challenges which slow its adoption down. In particular, there are currently few practical solutions allowing to exploit such infrastructure and to evaluate potential strategies. In [42], we propose a prototype orchestration architecture building on both Grid5000 and Fit-IoT lab (SILECS). This experimental testbed allows to realistically and rigorously evaluate orchestration strategies. In [20], we propose FITOR, an orchestration system for IoT applications in the Fog environment, which extends the actor-model based Calvin framework to cope with Fog environments while offering efficient orchestration mechanisms. In order to optimize the provisioning of Fog-Enabled IoT applications, FITOR relies on O-FSP, an optimized fog service provisioning strategy which aims to minimize the provisioning cost of IoT applications, while meeting their requirements. Based on extensive experiments, the results obtained show that O-FSP optimizes the placement of IoT applications and outperforms the related strategies in terms of i) provisioning cost ii) resource usage and iii) acceptance rate.
- End devices nearing the physical world can have interesting properties such as short delays, responsiveness, optimized communications and privacy. However, these end devices have low stability and are prone to failures. There is consequently a need for failure management protocols for IoT applications in the Fog. The design of such solutions is complex due to the specificities of the environment, i.e., (i) dynamic infrastructure where entities join and leave without synchronization, (ii) high heterogeneity in terms of functions, communication models, network, processing and storage capabilities, and, (iii) cyber-physical interactions which introduce non-deterministic and physical world's space and time dependent events. In [29], [37], we present a fault tolerance approach taking into account these three characteristics of the Fog-IoT environment. Fault tolerance is achieved by saving the state of the application in an uncoordinated way. When a failure is

detected, notifications are propagated to limit the impact of failures and dynamically reconfigure the application. Data stored during the state saving process are used for recovery, taking into account consistency with respect to the physical world. The approach was validated through practical experiments on a smart home platform.

## 7.3. HPC Application Analysis and Visualization

- Programming paradigms in High-Performance Computing have been shifting towards task-based models which are capable of adapting readily to heterogeneous and scalable supercomputers. The performance of task-based application heavily depends on the runtime scheduling heuristics and on its ability to exploit computing and communication resources. Unfortunately, the traditional performance analysis strategies are unfit to fully understand task-based runtime systems and applications: they expect a regular behavior with communication and computation phases, while task-based applications demonstrate no clear phases. Moreover, the finer granularity of task-based applications typically induces a stochastic behavior that leads to irregular structures that are difficult to analyze. Furthermore, the combination of application structure, scheduler, and hardware information is generally essential to understand performance issues. The papers [36], [6] presents a flexible framework that enables one to combine several sources of information and to create custom visualization panels allowing to understand and pinpoint performance problems incurred by bad scheduling decisions in task-based applications. Three case-studies using StarPU-MPI, a task-based multi-node runtime system, are detailed to show how our framework can be used to study the performance of the well-known Cholesky factorization. Performance improvements include a better task partitioning among the multi-(GPU,core) to get closer to theoretical lower bounds, improved MPI pipelining in multi-(node,core,GPU) to reduce the slow start, and changes in the runtime system to increase MPI bandwidth, with gains of up to 13% in the total makespan.
- In the context of multi-physics simulations on unstructured and heterogeneous meshes, generating well-balanced partitions is not trivial. The computing cost per mesh element in different phases of the simulation depends on various factors such as its type, its connectivity with neighboring elements or its layout in memory with respect to them, which determines the data locality. Moreover, if different types of discretization methods or computing devices are combined, the performance variability across the domain increases. Due to all these factors, evaluate a representative computing cost per mesh element, to generate well-balanced partitions, is a difficult task. Nonetheless, load balancing is a critical aspect of the efficient use of extreme scale systems since idle-times can represent a huge waste of resources, particularly when a single process delays the overall simulation. In this context, we present in [16] some improvements carried out on an in-house geometric mesh partitioner based on the Hilbert Space-Filling Curve. We have previously tested its effectiveness by partitioning meshes with up to 30 million elements in a few tenths of milliseconds using up to 4096 CPU cores, and we have leveraged its performance to develop an autotuning approach to adjust the load balancing according to runtime measurements. In this paper, we address the problem of having different load distributions in different phases of the simulation, particularly in the matrix assembly and in the solution of the linear system. We consider a multi-partition approach to ensure a proper load balance in all the phases. The initial results presented show the potential of this strategy.

## 7.4. Energy Optimization and Smart Grids Simulation

Large-scale decentralized photovoltaic (PV) generators are currently being installed in many low-voltage distribution networks. Without grid reinforcements or production curtailment, they might create current and/or voltage issues. In [13], [45], we consider the use the advanced metering infrastructure (AMI) as the basis for PV generation control. We show that the advanced metering infrastructure may be used to infer some knowledge about the underlying network, and we show how this knowledge can be used by a simple feed-forward controller to curtail the solar production efficiently.

We developed a environment for co-simulating electrical networks, telecommunication networks and online learning algorithms [3]. One the outputs of this work was to allow us to perform realistic numerical simulations of active distribution networks. We used this simulator to compare our proposed controller with two other controller structures: open-loop, and feedback P (U) and Q(U). We demonstrate that our feed-forward controller –that requires no prior knowledge of the underlying electrical network– brings significant performance improvements as it can effectively suppress over-voltage and over-current while requiring low energy curtailment. This method can be implemented at low cost and require no specific information about the network on which it is deployed.

Finally, we study demand-Response (DR) programs, whereby users of an electricity network are encouraged by economic incentives to rearrange their consumption in order to reduce production costs. Such mechanisms are envisioned to be a key feature of the smart grid paradigm. Several recent works proposed DR mechanisms and used analytical models to derive optimal incentives. Most of these works, however, rely on a macroscopic description of the population that does not model individual choices of users. In in[4], we conduct a detailed analysis of those models and we argue that the macroscopic descriptions hide important assumptions that can jeopardize the mechanisms' implementation (such as the ability to make personalized offers and to perfectly estimate the demand that is moved from a timeslot to another). Then, we start from a microscopic description that explicitly models each user's decision. We introduce four DR mechanisms with various assumptions on the provider's capabilities. Contrarily to previous studies, we find that the optimization problems that result from our mechanisms are complex and can be solved numerically only through a heuristic. We present numerical simulations that compare the different mechanisms and their sensitivity to forecast errors. At a high level, our results show that the performance of DR mechanisms under reasonable assumptions on the provider's capabilities are significantly lower than those suggested by previous studies, but that the gap reduces when the population's flexibility increases.

## 7.5. Simulation of HPC Applications

Beside continuous development and contribution to the SimGrid project, the two following contributions have been published this year. Both build on the SMPI interface which allows to efficiently predict the performance of MPI applications.

- Finite-difference methods are commonplace in High Performance Computing applications. Despite their apparent regularity, they often exhibit load imbalance that damages their efficiency. In [9], we characterize the spatial and temporal load imbalance of Ondes3D, a typical finite-differences application dedicated to earthquake modeling. Our analysis reveals imbalance originating from the structure of the input data, and from low-level CPU optimizations. Ondes3D was successfully ported to AMPI/CHARM++ using over-decomposition and MPI process migration techniques to dynamically rebalance the load. However, this approach requires careful selection of the over-decomposition level, the load balancing algorithm, and its activation frequency. These choices are usually tied to application structure and platform characteristics. We have thus proposed a workflow that leverages the capabilities of SimGrid to conduct such study at low experimental cost. We rely on a combination of emulation, simulation, and application modeling that requires minimal code modification and manages to capture both spatial and temporal load imbalance to faithfully predict the performance of dynamic load balancing. We evaluate the quality of our simulation by comparing simulation results with the outcome of real executions and demonstrate how this approach can be used to quickly find the optimal load balancing configuration for a given application/hardware configuration.
- It is typical in High Performance Computing (HPC) courses to give students access to HPC platforms so that they can benefit from hands-on learning opportunities. Using such platforms, however, comes with logistical and pedagogical challenges. For instance, a logistical challenge is that access to representative platforms must be granted to students, which can be difficult for some institutions or course modalities; and a pedagogical challenge is that hands-on learning opportunities are constrained by the configurations of these platforms. A way to address these challenges is to

instead simulate program executions on arbitrary HPC platform configurations. In [19] we focus on simulation in the specific context of distributed-memory computing and MPI programming education. While using simulation in this context has been explored in previous works, our approach offers two crucial advantages. First, students write standard MPI programs and can both debug and analyze the performance of their programs in simulation mode. Second, large-scale executions can be simulated in short amounts of time on a single standard laptop computer. This is possible thanks to SMPI, an MPI simulator provided as part of SimGrid. After detailing the challenges involved when using HPC platforms for HPC education and providing background information about SMPI, we present SMPI Courseware. SMPI Courseware is a set of in-simulation assignments that can be incorporated into HPC courses to provide students with hands-on experience for distributed-memory computing and MPI programming learning objectives. We describe some these assignments, highlighting how simulation with SMPI enhances the student learning experience.

## 7.6. Mean Field and Refined Mean Field Methods

Mean field approximation is a popular means to approximate stochastic models that can be represented as a system of $N$ interacting objects. It is know to be exact as $N$ goes to infinity. In a recent series of paper, [24], [25], [7], we establish theoretical results and numerical methods that allows us to define an approximation that is much more accurate than the classical mean field approximation. This new approximation, that we call the *refined mean field approximation*, is based on the computation of an expansion term of the order $1/N$. By considering a variety of applications, that include coupon collector, load balancing and bin packing problems, we illustrate that the proposed refined mean field approximation is significantly more accurate that the classic mean field approximation for small and moderate values of $N$: relative errors are often below $1\%$ for systems with $N = 10$.

In [23], [8], we improve this result in two directions. First, we show how to obtain the same result for the transient regime. Second, we provide a further refinement by expanding the term in $1/N^2$ (both for transient and steady-state regime). Our derivations are inspired by moment-closure approximation, a popular technique in theoretical biochemistry. We provide a number of examples that show: (1) that this new approximation is usable in practice for systems with up to a few tens of dimensions, and (2) that it accurately captures the transient and steady state behavior of such systems.

## 7.7. Optimization of Networks and Communication

This section describes two independent contributions on the analysis and optimization of networks and communication.

- Telecommunication networks are converging to a massively distributed cloud infrastructure interconnected with software defined networks. In the envisioned architecture, services will be deployed flexibly and quickly as network slices. Our paper [26] addresses a major bottleneck in this context, namely the challenge of computing the best resource provisioning for network slices in a robust and efficient manner. With tractability in mind, we propose a novel optimization framework which allows fine-grained resource allocation for slices both in terms of network bandwidth and cloud processing. The slices can be further provisioned and auto-scaled optimally based on a large class of utility functions in real-time. Furthermore, by tuning a slice-specific parameter, system designers can trade off traffic-fairness with computing-fairness to provide a mixed fairness strategy. We also propose an iterative algorithm based on the alternating direction method of multipliers (ADMM) that provably converges to the optimal resource allocation and we demonstrate the method's fast convergence in a wide range of quasi-stationary and dynamic settings.
- Distributed power control schemes in wireless networks have been well-examined, but standard methods rarely consider the effect of potentially random delays, which occur in almost every real-world network. We present in paper [33] Robust Feedback Averaging, a novel power control algorithm that is capable of operating in delay-ridden and noisy environments. We prove optimal convergence of this algorithm in the presence of random, time-varying delays, and present numerical

simulations that indicate that Robust Feedback Averaging outperforms the ubiquitous Foschini-Miljanic algorithm in several regimes.

## 7.8. Privacy, Fairness, and Transparency in Online Social Medias

Bringing transparency to algorithmic decision making systems and guaranteeing that the system satisfies properties of fairness and privacy is crucial in today's world. To start tackling this broad challenge, we focused on the case of online advertising and we had the following contributions.

- *Transparency properties for social media advertising and audit of Facebook's explanations.* In [15], we took a first step towards exploring the transparency mechanisms provided by social media sites, focusing on the two processes for which Facebook provides transparency mechanisms: the process of how Facebook infers data about users, and the process of how advertisers use this data to target users. We call explanations about those two processes *data explanations* and *ad explanations*, respectively.

  We identify a number of *properties* that are key for different types of explanations aimed at bringing transparency to social media advertising. We then evaluate empirically how well Facebook's explanations satisfy these properties and discuss the implications of our findings in view of the possible purposes of explanations. In particular, for *ad explanations*, we define five key properties: *personalization*, *completeness*, *correctness* (and the companion property of *misleadingness*), *consistency*, and *determinism*, and we show that Facebook's ad explanations are often *incomplete* and sometimes *misleading*. In particular, we observe that Facebook reveals only the most prevalent attribute used by the advertisers, which may allow malicious advertisers to easily obfuscate ad explanations from ad campaigns that are discriminatory or that target privacy-sensitive attributes. For *data explanations*, we define four key properties of the explanations: *specificity*, *snapshot completeness*, *temporal completeness*, and *correctness*; and we show that Facebook's explanations are *incomplete* and often *vague*; hence potentially limiting user control.

  Overall, our study provides a first step towards better understanding and improving transparency in social media advertising. During this work, we developed the tool AdAnalyst (https://adanalyst.mpi-sws.org/), which was instrumental for the study but also provides a transparency tool on its own for the large public, and is anticipated to be the basis of a number of further research studies in transparency.
- *Potential for discrimination in social media advertising.* Recently, online targeted advertising platforms like Facebook have been criticized for allowing advertisers to discriminate against users belonging to sensitive groups, i.e., to exclude users belonging to a certain race or gender from receiving their ads. Such criticisms have led, for instance, Facebook to disallow the use of attributes such as ethnic affinity from being used by advertisers when targeting ads related to housing or employment or financial services. In our paper [30], we systematically investigate the different targeting methods offered by Facebook (traditional attribute- or interest-based targeting, custom audience and lookalike audience) for their ability to enable discriminatory advertising and showed that a malicious advertiser can create highly discriminatory ads without using sensitive attributes (hence banning those features is inefficient to solve the problem). We argue that discrimination measures should be based on the targeted population and not on the attributes used for targeting and propose a discrimination metric in this direction.
- *Identification and resolution of privacy leakages in the Facebook's advertising platform.* In paper [31] we discovered that the information provided to advertisers through the custom audience feature (where an advertisers can upload PIIs (Personally Identifiable Information) of their customers and Facebook matches those with their users) was very severely leaking personal information. Specifically, it was making it possible for a malicious advertiser knowing the email address of a user to discover its phone number. Perhaps even worse, it was allowing a malicious advertiser to de-anonymize visitors of a website he controls. We discovered that the problem was due to the way Facebook computes estimates of the number of users matching a list of PIIs and proposed a solution based on not de-duplicating records with different PIIs belonging to the same users; and

we proved the robustness of our solution theoretically. Our work led to Facebook implementing a solution inspired by the one we proposed.

## 7.9. Optimization Methods

This section describes four independent contributions on optimization.

- In view of solving convex optimization problems with noisy gradient input, we analyze in the paper [11] the asymptotic behavior of gradient-like flows under stochastic disturbances. Specifically, we focus on the widely studied class of mirror descent schemes for convex programs with compact feasible regions, and we examine the dynamics' convergence and concentration properties in the presence of noise. In the vanishing noise limit, we show that the dynamics converge to the solution set of the underlying problem (a.s.). Otherwise, when the noise is persistent, we show that the dynamics are concentrated around interior solutions in the long run, and they converge to boundary solutions that are sufficiently "sharp". Finally, we show that a suitably rectified variant of the method converges irrespective of the magnitude of the noise (or the structure of the underlying convex program), and we derive an explicit estimate for its rate of convergence.
- We examine in paper [12] a class of stochastic mirror descent dynamics in the context of monotone variational inequalities (including Nash equilibrium and saddle-point problems). The dynamics under study are formulated as a stochastic differential equation driven by a (single-valued) monotone operator and perturbed by a Brownian motion. The system's controllable parameters are two variable weight sequences that respectively pre- and post-multiply the driver of the process. By carefully tuning these parameters, we obtain global convergence in the ergodic sense, and we estimate the average rate of convergence of the process. We also establish a large deviations principle showing that individual trajectories exhibit exponential concentration around this average.
- We develop in [17] a new stochastic algorithm with variance reduction for solving pseudo-monotone stochastic variational inequalities. Our method builds on Tseng's forward-backward-forward algorithm, which is known in the deterministic literature to be a valuable alternative to Korpelevich's extragradient method when solving variational inequalities over a convex and closed set governed with pseudo-monotone and Lipschitz continuous operators. The main computational advantage of Tseng's algorithm is that it relies only on a single projection step, and two independent queries of a stochastic oracle. Our algorithm incorporates a variance reduction mechanism, and leads to a.s. convergence to solutions of a merely pseudo-monotone stochastic variational inequality problem. To the best of our knowledge, this is the first stochastic algorithm achieving this by using only a single projection at each iteration.
- One of the most widely used training methods for large-scale machine learning problems is distributed asynchronous stochastic gradient descent (DASGD). However, a key issue in its implementation is that of delays: when a "worker" node asynchronously contributes a gradient update to the "master", the global model parameter may have changed, rendering this information stale. In massively parallel computing grids, these delays can quickly add up if a node is saturated, so the convergence of DASGD is uncertain under these conditions. Nevertheless, by using a judiciously chosen quasilinear step-size sequence, we show in [35] that it is possible to amortize these delays and achieve global convergence with probability 1, even under polynomially growing delays, reaffirming in this way the successful application of DASGD to large-scale optimization problems.

## 7.10. Multi-agent Learning and Distributed Best Response

This section describes several independent contributions on multi-agent learning.

- In [5], [22], [21], we study how fast can simple algorithms compute Nash equilibria. We study the case of random potential games for which we have designed and analyzed distributed algorithms to compute a Nash equilibrium. Our algorithms are based on best-response dynamics, with suitable revision sequences (orders of play). We compute the average complexity over all potential games

of best response dynamics under a random i.i.d. revision sequence, since it can be implemented in a distributed way using Poisson clocks. We obtain a distributed algorithm whose execution time is within a constant factor of the optimal centralized one. We also showed how to take advantage of the structure of the interactions between players in a network game: non- interacting players can play simultaneously. This improves best response algorithm, both in the centralized and in the distributed case.

- In [10], we study a class of evolutionary game dynamics defined by balancing a gain determined by the game's payoffs against a cost of motion that captures the difficulty with which the population moves between states. Costs of motion are represented by a Riemannian metric, i.e., a state-dependent inner product on the set of population states. The replicator dynamics and the (Euclidean) projection dynamics are the archetypal examples of the class we study. Like these representative dynamics, all Riemannian game dynamics satisfy certain basic desiderata, including positive correlation and global convergence in potential games. Moreover, when the underlying Riemannian metric satisfies a Hessian integrability condition, the resulting dynamics preserve many further properties of the replicator and projection dynamics. We examine the close connections between Hessian game dynamics and reinforcement learning in normal form games, extending and elucidating a well-known link between the replicator dynamics and exponential reinforcement learning.
- The paper [18] examines the long-run behavior of learning with bandit feedback in non-cooperative concave games. The bandit framework accounts for extremely low-information environments where the agents may not even know they are playing a game; as such, the agents' most sensible choice in this setting would be to employ a no-regret learning algorithm. In general, this does not mean that the players' behavior stabilizes in the long run: no-regret learning may lead to cycles, even with perfect gradient information. However, if a standard monotonicity condition is satisfied, our analysis shows that no-regret learning based on mirror descent with bandit feedback converges to Nash equilibrium with probability 1. We also derive an upper bound for the convergence rate of the process that nearly matches the best attainable rate for single-agent bandit stochastic optimization.
- In [34], we consider a game-theoretical multi-agent learning problem where the feedback information can be lost during the learning process and rewards are given by a broad class of games known as variationally stable games. We propose a simple variant of the classical online gradient descent algorithm, called reweighted online gradient descent (ROGD) and show that in variationally stable games, if each agent adopts ROGD, then almost sure convergence to the set of Nash equilibria is guaranteed, even when the feedback loss is asynchronous and arbitrarily corrrelated among agents. We then extend the framework to deal with unknown feedback loss probabilities by using an estimator (constructed from past data) in its replacement. Finally, we further extend the framework to accomodate both asynchronous loss and stochastic rewards and establish that multi-agent ROGD learning still converges to the set of Nash equilibria in such settings. Together, these results contribute to the broad lanscape of multi-agent online learning by significantly relaxing the feedback information that is required to achieve desirable outcomes.
- Regularized learning is a fundamental technique in online optimization, machine learning and many other fields of computer science. A natural question that arises in these settings is how regularized learning algorithms behave when faced against each other. In the paper [27], we study a natural formulation of this problem by coupling regularized learning dynamics in zero-sum games. We show that the system's behavior is Poincaré recurrent, implying that almost every trajectory revisits any (arbitrarily small) neighborhood of its starting point infinitely often. This cycling behavior is robust to the agents' choice of regularization mechanism (each agent could be using a different regularizer), to positive-affine transformations of the agents' utilities, and it also persists in the case of networked competition, i.e., for zero-sum polymatrix games.

## 7.11. Blotto games

The Colonel Blotto game is a famous game commonly used to model resource allocation problems in many domains ranging from security to advertising. Two players distribute a fixed budget of resources on multiple

battlefields to maximize the aggregate value of battlefields they win, each battlefield being won by the player who allocates more resources to it. The continuous version of the game –where players can choose any fractional allocation– has been extensively studied, albeit only with partial results to date. Recently, the discrete version –where allocations can only be integers– started to gain traction and algorithms were proposed to compute the equilibrium in polynomial time; but these remain computationally impractical for large (or even moderate) numbers of battlefields. In [32], [46], we propose an algorithm to compute very efficiently an approximate equilibrium for the discrete Colonel Blotto game with many battlefields. We provide a theoretical bound on the approximation error as a function of the game's parameters, in particular number of battlefields and resource budgets. We also propose an efficient dynamic programming algorithm to compute the best-response to any strategy that allows computing for each game instance the actual value of the error. We perform numerical experiments that show that the proposed strategy provides a fast and good approximation to the equilibrium even for moderate numbers of battlefields.

**PRIVATICS Project-Team**

# 6. New Results

## 6.1. Fine-Grained Control over Tracking to Support the Ad-Based Web Economy

**Participant:** Claude Castelluccia.

The intrusiveness of Web tracking and the increasing invasiveness of digital advertising have raised serious concerns regarding user privacy and Web usability, leading a substantial chunk of the populace to adopt ad-blocking technologies in recent years. The problem with these technologies, however, is that they are extremely limited and radical in their approach, and they completely disregard the underlying economic model of the Web, in which users get content free in return for allowing advertisers to show them ads. Nowadays, with around 200 million people regularly using such tools, said economic model is in danger. In this article, we investigate an Internet technology that targets users who are not, in general, against advertising, accept the trade-off that comes with the "free" content, but—for privacy concerns—they wish to exert fine-grained control over tracking. Our working assumption is that some categories of web pages (e.g., related to health or religion) are more privacy-sensitive to users than others (e.g., about education or science). Capitalizing on this, we propose a technology that allows users to specify the categories of web pages that are privacy-sensitive to them and block the trackers present on such web pages only. As tracking is prevented by blocking network connections of third-party domains, we avoid not only tracking but also third-party ads. Since users continue receiving ads on those web pages that belong to non-sensitive categories, our approach may provide a better point of operation within the trade-off between user privacy and the Web economy. To test the appropriateness and feasibility of our solution, we implemented it as a Web-browser plug-in, which is currently available for Google Chrome and Mozilla Firefox. Experimental results from the collected data of 746 users during one year show that only 16.25% of ads are blocked by our tool, which seems to indicate that the economic impact of the ad-blocking exerted by privacy-sensitive users could be significantly reduced.

## 6.2. Differentially Private Mixture of Generative Neural Networks

**Participant:** Claude Castelluccia.

Generative models are used in a wide range of applications building on large amounts of contextually rich information. Due to possible privacy violations of the individuals whose data is used to train these models, however, publishing or sharing generative models is not always viable. In this paper, we present a novel technique for privately releasing generative models and entire high-dimensional datasets produced by these models. We model the generator distribution of the training data with a mixture of k generative neural networks. These are trained together and collectively learn the generator distribution of a dataset. Data is divided into k clusters, using a novel differentially private kernel k-means, then each cluster is given to separate generative neural networks, such as Restricted Boltzmann Machines or Variational Autoencoders, which are trained only on their own cluster using differentially private gradient descent. We evaluate our approach using the MNIST dataset, as well as call detail records and transit datasets, showing that it produces realistic synthetic samples, which can also be used to accurately compute arbitrary number of counting queries.

## 6.3. On the Cost-Effectiveness of Mass Surveillance

**Participant:** Claude Castelluccia.

In recent times, we have witnessed an increasing concern by governments and intelligence agencies to deploy mass-surveillance systems that help them fight terrorism. Although a government may be perfectly legitimate to do so, it is questionable whether a preventive-surveillance state is rational and cost-effective. In this paper, we conduct a theoretical analysis of the cost of such surveillance systems. Our analysis starts with a fairly well-known result in statistics, namely, the false-positive paradox. We propose a quantitative measure of the total cost of a monitoring program, and study a detection system that is designed to minimize it, subject to a constraint in the percentage of terrorists the agency wishes to capture. Our formulation is first illustrated by means of several simple albeit insightful examples of terrorist and innocent profiles. Then, we conduct an extensive experimental study from real-world socio-demographic data of jihadist terrorism in the U.K. and Spain, and provide insight into the rationality and cost-effectiveness of two countries with two of the biggest defense budgets in the world.

## 6.4. To Extend or not to Extend: on the Uniqueness of Browser Extensions and Web Logins

**Participants:** Claude Castelluccia, Gabor Gulyas.

Recent works showed that websites can detect browser extensions that users install and websites they are logged into. This poses significant privacy risks, since extensions and Web logins that re ect user's behavior, can be used to uniquely identify users on the Web. This paper reports on the rst large-scale behavioral uniqueness study based on 16,393 users who visited our website. We test and detect the presence of 16,743 Chrome extensions, covering 28% of all free Chrome extensions. We also detect whether the user is connected to 60 different websites. We analyze how unique users are based on their behavior, and nd out that 54.86% of users that have installed at least one detectable extension are unique; 19.53% of users are unique among those who have logged into one or more detectable websites; and 89.23% are unique among users with at least one extension and one login. We use an advanced ngerprinting algorithm and show that it is possible to identify a user in less than 625 milliseconds by selecting the most unique combinations of extensions. Because privacy extensions contribute to the uniqueness of users, we study the trade-o between the amount of trackers blocked by such extensions and how unique the users of these extensions are. We have found that privacy extensions should be considered more useful than harmful. The paper concludes with possible counter- measures.

## 6.5. Privacy-Preserving Release of Spatio-Temporal Density

**Participants:** Claude Castelluccia, Gergely Acs.

In today's digital society, increasing amounts of contextually rich spatio-temporal information are collected and used, e.g., for knowledge-based decision making, research purposes, optimizing operational phases of city management, planning infrastructure networks, or developing timetables for public transportation with an increasingly autonomous vehicle fleet. At the same time, however, publishing or sharing spatio-temporal data, even in aggregated form, is not always viable owing to the danger of violating individuals' privacy, along with the related legal and ethical repercussions. In this chapter, we review some fundamental approaches for anonymizing and releasing spatio-temporal density, i.e., the number of individuals visiting a given set of locations as a function of time. These approaches follow different privacy models providing different privacy guarantees as well as accuracy of the released anonymized data. We demonstrate some sanitization (anonymization) techniques with provable privacy guarantees by releasing the spatio-temporal density of Paris, in France. We conclude that, in order to achieve meaningful accuracy, the sanitization process has to be carefully customized to the application and public characteristics of the spatio-temporal data.

## 6.6. Algorithmic Decision Systems in the Health and Justice Sectors: Certification and Explanations for Algorithms in European and French Law

**Participant:** Daniel Le Metayer.

Algorithmic decision systems are already used in many everyday tools and services on the Internet, and they also play an increasing role in many situations in which people's lives and rights are strongly affected, such as job and loans applications, but also medical diagnosis and therapeutic choices, or legal advice and court decisions. This evolution gives rise to a whole range of questions. In this paper, we argue that certification and explanation are two complementary means of strengthening the European legal framework and enhancing trust in algorithmic decision systems. The former can be seen as the delegation of the task of checking certain criteria to an authority, while the latter allows the stakeholders themselves (for example, developers, users and decision-subjects) to understand the results or the logic of the system. We explore potential legal requirements of accountability in this sense and their effective implementation. These two aspects are tackled from the perspective of the European and French legal frameworks. We focus on two particularly sensitive application domains, namely the medical and legal sectors.

## 6.7. Capacity: an Abstract Model of Control over Personal Data

**Participant:** Daniel Le Metayer.

While the control of individuals over their personal data is increasingly seen as an essential component of their privacy, the word "control" is usually used in a very vague way, both by lawyers and by computer scientists. This lack of precision may lead to misunderstandings and makes it difficult to check compliance. To address this issue, we propose a formal framework based on capacities to specify the notion of control over personal data and to reason about control properties. We illustrate our framework with social network systems and show that it makes it possible to characterize the types of control over personal data that they provide to their users and to compare them in a rigorous way.

## 6.8. Biometric Systems Private by Design: Reasoning about privacy properties of biometric system architectures

**Participant:** Daniel Le Metayer.

In  is to show the applicability of the privacy by design approach to biometric systems and the benefit of using formal methods to this end. We build on a general framework for the definition and verification of privacy architectures introduced at STM 2014 and show how it can be adapted to biometrics. The choice of particular techniques and the role of the components (central server, secure module, biometric terminal, smart card, etc.) in the architecture have a strong impact on the privacy guarantees provided by a biometric system. Some architectures have already been analysed but on a case by case basis, which makes it difficult to draw comparisons and to provide a rationale for the choice of specific options. In this paper, we describe the application of a general privacy architecture framework to specify different design options for biometric systems and to reason about them in a formal way.

## 6.9. Privacy Risk Analysis to Enable Informed Privacy Settings

**Participant:** Daniel Le Metayer.

is a contribution to enhancing individual control over personal data which is promoted, inter alia, by the new EU General Data Protection Regulation. We propose a method to enable better informed choices of privacy preferences or privacy settings. The method relies on a privacy risk analysis framework parameterized with privacy settings. The user can express his choices, visualize their impact on the privacy risks through a user-friendly interface, and decide to revise them as necessary to reduce risks to an acceptable level.

## 6.10. Enhancing Transparency and Consent in the IoT

**Participants:** Daniel Le Metayer, Claude Castelluccia, Mathieu Cunche, Victor Morel.

The development of the IoT raises specific questions in terms of privacy, especially with respect to information to users and consent. We argue that (1) all necessary information about collected data and the collecting devices should be communicated electronically to all data subjects in their range and (2) data subjects should be able to reply also electronically and express their own privacy choices. In this position paper, we take some examples of technologies and initiatives to illustrate our position (including direct and registry-based communications) and discuss them in the light of the GDPR and the WP29 recommendations.

## 6.11. Toward privacy in IoT mobile devices for activity recognition

**Participant:** Antoine Boutet.

Recent advances in wireless sensors for personal healthcare allow to recognise human real-time activities with mobile devices. While the analysis of those datastream can have many benefits from a health point of view, it can also lead to privacy threats by exposing highly sensitive information. In this work, we propose a privacy-preserving framework for activity recognition. This framework relies on a machine learning technique to efficiently recognise the user activity pattern, useful for personal healthcare monitoring, while limiting the risk of re-identification of users from biometric patterns that characterizes each individual. To achieve that, we first deeply analysed different features extraction schemes in both temporal and frequency domain. We show that features in temporal domain are useful to discriminate user activity while features in frequency domain lead to distinguish the user identity. On the basis of this observation, we second design a novel protection mechanism that processes the raw signal on the user's smartphone and transfers to the application server only the relevant features unlinked to the identity of users. In addition, a generalisation-based approach is also applied on features in frequency domain before to be transmitted to the server in order to limit the risk of re-identification. We extensively evaluate our framework with a reference dataset: results show an accurate activity recognition (87%) while limiting the re-identification rate (33%). This represents a slightly decrease of utility (9%) against a large privacy improvement (53%) compared to state-of-the-art baselines, while reducing the computational cost on the application server.

## 6.12. The Long Road to Computational Location Privacy: A Survey

**Participant:** Antoine Boutet.

The widespread adoption of continuously connected smartphones and tablets developed the usage of mobile applications, among which many use location to provide geolocated services. These services provide new prospects for users: getting directions to work in the morning, leaving a check-in at a restaurant at noon and checking next day's weather in the evening are possible right from any mobile device embedding a GPS chip. In these location-based applications, the user's location is sent to a server, which uses them to provide contextual and personalised answers. However, nothing prevents the latter from gathering, analysing and possibly sharing the collected information, which opens the door to many privacy threats. Indeed, mobility data can reveal sensitive information about users, among which one's home, work place or even religious and political preferences. For this reason, many privacy-preserving mechanisms have been proposed these last years to enhance location privacy while using geolocated services. This work surveys and organises contributions in this area from classical building blocks to the most recent developments of privacy threats and location privacy-preserving mechanisms. We divide the protection mechanisms between online and offline use cases, and organise them into six categories depending on the nature of their algorithm. Moreover, this work surveys the evaluation metrics used to assess protection mechanisms in terms of privacy, utility and performance. Finally, open challenges and new directions to address the problem of computational location privacy are pointed out and discussed.

## 6.13. CYCLOSA: Decentralizing Private Web Search Through SGX-Based Browser Extensions

**Participant:** Antoine Boutet.

By regularly querying Web search engines, users (unconsciously) disclose large amounts of their personal data as part of their search queries, among which some might reveal sensitive information (e.g. health issues, sexual, political or religious preferences). Several solutions exist to allow users querying search engines while improving privacy protection. However, these solutions suffer from a number of limitations: some are subject to user re-identification attacks, while others lack scalability or are unable to provide accurate results. This contribution presents CYCLOSA, a secure, scalable and accurate private Web search solution. CYCLOSA improves security by relying on trusted execution environments (TEEs) as provided by Intel SGX. Further, CYCLOSA proposes a novel adaptive privacy protection solution that reduces the risk of user re-identification. CYCLOSA sends fake queries to the search engine and dynamically adapts their count according to the sensitivity of the user query. In addition, CYCLOSA meets scalability as it is fully decentralized, spreading the load for distributing fake queries among other nodes. Finally, CYCLOSA achieves accuracy of Web search as it handles the real query and the fake queries separately, in contrast to other existing solutions that mix fake and real query results.

## 6.14. ACCIO: How to Make Location Privacy Experimentation Open and Easy

**Participant:** Antoine Boutet.

The advent of mobile applications collecting and exploiting the location of users opens a number of privacy threats. To mitigate these privacy issues, several protection mechanisms have been proposed this last decade to protect users' location privacy. However, these protection mechanisms are usually implemented and evaluated in monolithic way, with heterogeneous tools and languages. Moreover, they are evaluated using different methodologies, metrics and datasets. This lack of standard makes the task of evaluating and comparing protection mechanisms particularly hard. In this work, we present ACCIO, a unified framework to ease the design and evaluation of protection mechanisms. Thanks to its Domain Specific Language, ACCIO allows researchers and practitioners to define and deploy experiments in an intuitive way, as well as to easily collect and analyse the results. ACCIO already comes with several state-of-the-art protection mechanisms and a toolbox to manipulate mobility data. Finally, ACCIO is open and easily extensible with new evaluation metrics and protection mechanisms. This openness, combined with a description of experiments through a user-friendly DSL, makes ACCIO an appealing tool to reproduce and disseminate research results easier. In this work, we present ACCIO's motivation and architecture, and demonstrate its capabilities through several use cases involving multiples metrics, state-of-the-art protection mechanisms, and two real-life mobility datasets collected in Beijing and in the San Francisco area.

## 6.15. Collaborative Filtering Under a Sybil Attack: Similarity Metrics do Matter!

**Participant:** Antoine Boutet.

Recommendation systems help users identify interesting content, but they also open new privacy threats. In this contribution, we deeply analyze the effect of a Sybil attack that tries to infer information on users from a user-based collaborative-filtering recommendation systems. We discuss the impact of different similarity metrics used to identity users with similar tastes in the trade-off between recommendation quality and privacy. Finally, we propose and evaluate a novel similarity metric that combines the best of both worlds: a high recommendation quality with a low prediction accuracy for the attacker. Our results, on a state-of-the-art recommendation framework and on real datasets show that existing similarity metrics exhibit a wide range of behaviors in the presence of Sybil attacks, while our new similarity metric consistently achieves the best trade-off while outperforming state-of-the-art solutions.

## 6.16. Automatic Privacy and Utility Preservation of Mobility Data: A Nonlinear Model-Based Approach

**Participant:** Antoine Boutet.

The widespread use of mobile devices and location-based services has generated massive amounts of mobility databases. While processing these data is highly valuable, privacy issues can occur if personal information is revealed. The prior art has investigated ways to protect mobility data by providing a large range of Location Privacy Protection Mechanisms (LPPMs). However, the privacy level of the protected data significantly varies depending on the protection mechanism used, its configuration and on the characteristics of the mobility data. Meanwhile, the protected data still needs to enable some useful processing. To tackle these issues, in this work we present PULP, a framework that finds the suitable protection mechanism and automatically configures it for each user in order to achieve user-defined objectives in terms of both privacy and utility. PULP uses nonlinear models to capture the impact of each LPPM on data privacy and utility levels. Evaluation of our framework is carried out with two protection mechanisms of the literature and four real-world mobility datasets. Results show the efficiency of PULP, its robustness and adaptability. Comparisons between LPPMs' configurator and the state of the art further illustrate that PULP better realizes users' objectives and its computations time is in orders of magnitude faster.

## 6.17. Privacy Preserving Analytics

**Participant:** Mathieu Cunche.

As communications-enabled devices are becoming more ubiquitous, it becomes easier to track the movements of individuals through the radio signals broadcasted by their devices. Thus, while there is a strong interest for physical analytics platforms to leverage this information for many purposes, this tracking also threatens the privacy of individuals. To solve this issue, we propose a privacy-preserving solution for collecting aggregate mobility patterns while satisfying the strong guarantee of $\varepsilon$-differential privacy. More precisely, we introduce a sanitization mechanism for efficient, privacy-preserving and non-interactive approximate distinct counting for physical analytics based on perturbed Bloom filters called Pan-Private BLIP. We also extend and generalize previous approaches for estimating distinct count of events and joint events (i.e., intersection and more generally t-out-of-n cardinalities). Finally, we evaluate expirementally our approach and compare it to previous ones on real datasets.

## 6.18. Detecting smartphone state changes through a Bluetooth based timing attack

**Participants:** Mathieu Cunche, Guillaume Celosia.

Bluetooth is a popular wireless communication technology that is available on most mobile devices. Although Bluetooth includes security and privacy preserving mechanisms, we show that a Bluetooth harmless inherent request-response mechanism can taint users privacy. More specifically, we introduce a timing attack that can be triggered by a remote attacker in order to infer information about a Bluetooth device state. By observing the L2CAP layer ping mechanism timing variations, it is possible to detect device state changes, for instance when the device goes in or out of the locked state. Our experimental results show that change point detection analysis of the timing allows to detect device state changes with a high accuracy. Finally, we discuss applications and countermeasures.

## 6.19. Analyzing Ultrasound-based Physical Tracking Systems

**Participant:** Mathieu Cunche.

A trending application of ultrasound communication is the implementation of ultrasound beacons to track owners of mobile phones in stores and shopping centers. We present the analysis of an Ultrasound-based tracking application. By analyzing several mobile applications along with the network communication and sample of the original audio signal, we were able to reverse engineer the ultrasonic communications and some other elements of the system. Based on those finding we show how arbitrary ultrasonic signal can be generated and how to perform jamming. Finally we analyze a real world deployment and discuss privacy implications.

**ROMA Project-Team**

# 7. New Results

## 7.1. Birkhoff–von Neumann decomposition

The well-known Birkhoff-von Neumann (BvN) decomposition expresses a doubly stochastic matrix as a convex combination of a number of permutation matrices. For a given doubly stochastic matrix, there are many BvN decompositions, and finding the one with the minimum number of permutation matrices is NP-hard. There are heuristics to obtain BvN decompositions for a given doubly stochastic matrix. A family of heuristics are based on the original proof of Birkhoff and proceed step by step by subtracting a scalar multiple of a permutation matrix at each step from the current matrix, starting from the given matrix. At every step, the subtracted matrix contains nonzeros at the positions of some nonzero entries of the current matrix and annihilates at least one entry, while keeping the current matrix nonnegative. Our first result, which supports a claim of Brualdi [68], shows that this family of heuristics can miss optimal decompositions. We also investigate the performance of two heuristics from this family theoretically. The findings are published in a journal [10].

## 7.2. Parallel sparse matrix-vector multiply

There are three common parallel sparse matrix-vector multiply algorithms: 1D 3 row-parallel, 1D column-parallel and 2D row-column-parallel. The 1D parallel algorithms offer the advantage of having only one communication phase. On the other hand, the 2D parallel algorithm is more scalable but it suffers from two communication phases. In this work, we introduce a novel concept of heterogeneous messages where a heterogeneous message may contain both input-vector entries and partially computed output-vector entries. This concept not only leads to a decreased number of messages, but also enables fusing the input-and output-communication phases into a single phase. These findings are exploited to propose a 1.5D parallel sparse matrix-vector multiply algorithm which is called local row-column-parallel. This proposed algorithm requires a constrained fine-grain partitioning in which each fine-grain task is assigned to the processor that contains either its input-vector entry, or its output-vector entry, or both. We propose two methods to carry out the constrained fine-grain partitioning. We conduct our experiments on a large set of test matrices to evaluate the partitioning qualities and partitioning times of these proposed 1.5D methods. The findings are published in a journal [14].

## 7.3. Scheduling series-parallel task graphs to minimize peak memory

We consider a variant of the well-known, NP-complete problem of minimum cut linear arrangement for directed acyclic graphs. In this variant, we are given a directed acyclic graph and we are asked to find a topological ordering such that the maximum number of cut edges at any point in this ordering is minimum. In our variant, the vertices and edges have weights, and the aim is to minimize the maximum weight of cut edges in addition to the weight of the last vertex before the cut. There is a known, polynomial time algorithm [78] for the cases where the input graph is a rooted tree. We focus on the instances where the input graph is a directed series-parallel graph, and propose a polynomial time algorithm, thus expanding the class of graphs for which a polynomial time algorithm is known. Directed acyclic graphs are used to model scientific applications where the vertices correspond to the tasks of a given application and the edges represent the dependencies between the tasks. In such models, the problem we address reads as minimizing the peak memory requirement in an execution of the application. Our work, combined with Liu's work on rooted trees addresses this practical problem in two important classes of applications. The findings are published in a journal [15].

## 7.4. Parallel Candecomp/Parafac decomposition of sparse tensors using dimension trees

Tensor factorization has been increasingly used to address various problems in many fields such as signal processing, data compression, computer vision, and computational data analysis. CANDECOMP/PARAFAC (CP) decomposition of sparse tensors has successfully been applied to many well-known problems in web search, graph analytics, recommender systems, health care data analytics, and many other domains. In these applications, computing the CP decomposition of sparse tensors efficiently is essential in order to be able to process and analyze data of massive scale. For this purpose, we investigate an efficient computation and parallelization of the CP decomposition for sparse tensors. We provide a novel computational scheme for reducing the cost of a core operation in computing the CP decomposition with the traditional alternating least squares (CP-ALS) based algorithm. We then effectively parallelize this computational scheme in the context of CP-ALS in shared and distributed memory environments, and propose data and task distribution models for better scalability. We implement parallel CP-ALS algorithms and compare our implementations with an efficient tensor factorization library, using tensors formed from real-world and synthetic datasets. With our algorithmic contributions and implementations, we report up to 3.95x, 3.47x, and 3.9x speedups in sequential, shared memory parallel, and distributed memory parallel executions over the state of the art, and up to 1466x overall speedup over the sequential execution using 4096 cores on an IBM BlueGene/Q supercomputer. The findings are published in a journal [13].

## 7.5. Approximation algorithms for maximum matchings in undirected graphs

We propose heuristics for approximating the maximum cardinality matching on undirected graphs. Our heuristics are based on the theoretical body of a certain type of random graphs, and are made practical for real-life ones. The idea is based on judiciously selecting a subgraph of a given graph and obtaining a maximum cardinality matching on this subgraph. We show that the heuristics have an approximation guarantee of around $0.866 - \log(n)/n$ for a graph with $n$ vertices. Experiments for verifying the theoretical results in practice are provided. The findings are published in a conference proceedings [25].

## 7.6. SINA: A Scalable iterative network aligner

Given two graphs, network alignment asks for a potentially partial mapping between the vertices of the two graphs. This arises in many applications where data from different sources need to be integrated. Recent graph aligners use the global structure of input graphs and additional information given for the edges and vertices. We present SINA, an efficient, shared memory parallel implementation of such an aligner. Our experimental evaluations on a 32-core shared memory machine showed that SINA scales well for aligning large real-world graphs: SINA can achieve up to $28.5\times$ speedup, and can reduce the total execution time of a graph alignment problem with 2M vertices and 100M edges from 4.5 hours to under 10 minutes. To the best of our knowledge, SINA is the first parallel aligner that uses global structure and vertex and edge attributes to handle large graphs. The findings are published in a conference proceedings [34].

## 7.7. Acyclic partitioning of large directed acyclic graphs

We investigate the problem of partitioning the vertices of a directed acyclic graph into a given number of parts. The objective function is to minimize the number or the total weight of the edges having end points in different parts, which is also known as edge cut. The standard load balancing constraint of having an equitable partition of the vertices among the parts should be met. Furthermore, the partition is required to be acyclic, i.e., the inter-part edges between the vertices from different parts should preserve an acyclic dependency structure among the parts. In this work, we adopt the multilevel approach with coarsening, initial partitioning, and refinement phases for acyclic partitioning of directed acyclic graphs. We focus on two-way partitioning (sometimes called bisection), as this scheme can be used in a recursive way for multi-way partitioning. To ensure the acyclicity of the partition at all times, we propose novel and efficient coarsening and refinement heuristics. The quality of the computed acyclic partitions is assessed by computing the edge cut. We also

propose effective ways to use the standard undirected graph partitioning methods in our multilevel scheme. We perform a large set of experiments on a dataset consisting of (i) graphs coming from an application and (ii) some others corresponding to matrices from a public collection. We report improvements, on average, around 59% compared to the current state of the art. The findings are published in a research report [50].

## 7.8. Effective heuristics for matchings in hypergraphs

The problem of finding a maximum cardinality matching in a $d$-partite $d$-uniform hypergraph is an important problem in combinatorial optimization and has been theoretically analyzed by several researchers. In this work, we first devise heuristics for this problem by generalizing the existing cheap graph matching heuristics. Then, we propose a novel heuristic based on tensor scaling to extend the matching via judicious hyperedge selections. Experiments on random, synthetic and real-life hypergraphs show that this new heuristic is highly practical and superior to the others on finding a matching with large cardinality. The findings are published in a research report [46].

## 7.9. Scaling matrices and counting the perfect matchings in graphs

We investigate efficient randomized methods for approximating the number of perfect matchings in bipartite graphs and general graphs. Our approach is based on assigning probabilities to edges. The findings are published in a research report [47].

## 7.10. A scalable clustering-based task scheduler for homogeneous processors using DAG partitioning

When scheduling a directed acyclic graph (DAG) of tasks on computational platforms, a good trade-off between load balance and data locality is necessary. List-based scheduling techniques are commonly used greedy approaches for this problem. The downside of list-scheduling heuristics is that they are incapable of making short-term sacrifices for the global efficiency of the schedule. In this work, we describe new list-based scheduling heuristics based on clustering for homogeneous platforms. Our approach uses an acyclic partitioner for DAGs for clustering. The clustering enhances the data locality of the scheduler with a global view of the graph. Furthermore, since the partition is acyclic, we can schedule each part completely once its input tasks are ready to be executed. We present an extensive experimental evaluation showing the trade-offs between the granularity of clustering and the parallelism, and how this affects the scheduling. Furthermore, we compare our heuristics to the best state-of-the-art list-scheduling and clustering heuristics, and obtain better performance in cases with many communications. The findings are published in a research report [53].

## 7.11. Data-Locality Aware Dynamic Schedulers for Independent Tasks with Replicated Inputs

In this work we concentrate on a crucial parameter for efficiency in Big Data and HPC applications: data locality. We focus on the scheduling of a set of independant tasks, each depending on an input file. We assume that each of these input files has been replicated several times and placed in local storage of different nodes of a cluster, similarly of what we can find on HDFS system for example. We consider two optimization problems, related to the two natural metrics: makespan optimization (under the constraint that only local tasks are allowed) and communication optimization (under the constraint of never letting a processor idle in order to optimize makespan). For both problems we investigate the performance of dynamic schedulers, in particular the basic greedy algorithm we can for example find in the default MapReduce scheduler. First we theoretically study its performance, with probabilistic models, and provide a lower bound for communication metric and asymptotic behaviour for both metrics. Second we propose simulations based on traces from a Hadoop cluster to compare the different dynamic schedulers and assess the expected behaviour obtained with the theoretical study.

These findings have been presented at the CEBDA workshop [19].

## 7.12. Parallel scheduling of DAGs under memory constraints.

Scientific workflows are frequently modeled as Directed Acyclic Graphs (DAG) of tasks, which represent computational modules and their dependencies, in the form of data produced by a task and used by another one. This formulation allows the use of runtime systems which dynamically allocate tasks onto the resources of increasingly complex and heterogeneous computing platforms. However, for some workflows, such a dynamic schedule may run out of memory by exposing too much parallelism. This work focuses on the problem of transforming such a DAG to prevent memory shortage, and concentrates on shared memory platforms. We first propose a simple model of DAG which is expressive enough to emulate complex memory behaviors. We then exhibit a polynomial-time algorithm that computes the maximum peak memory of a DAG, that is, the maximum memory needed by any parallel schedule. We consider the problem of reducing this maximum peak memory to make it smaller than a given bound by adding new fictitious edges, while trying to minimize the critical path of the graph. After proving this problem NP-complete, we provide an ILP solution as well as several heuristic strategies that are thoroughly compared by simulation on synthetic DAGs modeling actual computational workflows. We show that on most instances, we are able to decrease the maximum peak memory at the cost of a small increase in the critical path, thus with little impact on quality of the final parallel schedule.

This work has been presented at the IPDPS 2018 conference [31] and an extended version has been submitted to the Elsevier JPDC journal [52].

## 7.13. Online Scheduling of Task Graphs on Hybrid Platforms.

Modern computing platforms commonly include accelerators. We target the problem of scheduling applications modeled as task graphs on hybrid platforms made of two types of resources, such as CPUs and GPUs. We consider that task graphs are uncovered dynamically, and that the scheduler has information only on the available tasks, i.e., tasks whose predecessors have all been completed. Each task can be processed by either a CPU or a GPU, and the corresponding processing times are known. Our study extends a previous $4\sqrt{m/k}$-competitive online algorithm [61], where $m$ is the number of CPUs and $k$ the number of GPUs ($m \geq k$). We prove that no online algorithm can have a competitive ratio smaller than $\sqrt{m/k}$. We also study how adding flexibility on task processing, such as task migration or spoliation, or increasing the knowledge of the scheduler by providing it with information on the task graph, influences the lower bound. We provide a $(2\sqrt{m/k}+1)$-competitive algorithm as well as a tunable combination of a system-oriented heuristic and a competitive algorithm; this combination performs well in practice and has a competitive ratio in $\Theta(\sqrt{m/k})$. Finally, simulations on different sets of task graphs illustrate how the instance properties impact the performance of the studied algorithms and show that our proposed tunable algorithm performs the best among the online algorithms in almost all cases and has even performance close to an offline algorithm.

This work has been presented at the EuroPar 2018 conference [24].

## 7.14. Memory-aware tree partitioning on homogeneous platforms

Scientific applications are commonly modeled as the processing of directed acyclic graphs of tasks, and for some of them, the graph takes the special form of a rooted tree. This tree expresses both the computational dependencies between tasks and their storage requirements. The problem of scheduling/traversing such a tree on a single processor to minimize its memory footprint has already been widely studied. Hence, we move to parallel processing and study how to partition the tree for a homogeneous multiprocessor platform, where each processor is equipped with its own memory. We formally state the problem of partitioning the tree into subtrees such that each subtree can be processed on a single processor and the total resulting processing time is minimized. We prove that the problem is NP-complete, and we design polynomial-time heuristics to address it. An extensive set of simulations demonstrates the usefulness of these heuristics.

This work has been presented as a short paper in the PDP 2018 conference [27].

## 7.15. Reliability-aware energy optimization for throughput-constrained applications on MPSoC.

Multi-Processor System-on-Chip (MPSoC) has emerged as a promising platform to meet the increasing performance demand of embedded applications. However, due to limited energy budget, it is hard to guarantee that applications on MPSoC can be accomplished on time with a required throughput. The situation becomes even worse for applications with high reliability requirements, since extra energy will be inevitably consumed by task re-executions or duplicated tasks. Based on Dynamic Voltage and Frequency Scaling (DVFS) and task duplication techniques, this paper presents a novel energy-efficient scheduling model, which aims at minimizing the overall energy consumption of MPSoC applications under both throughput and reliability constraints. The problem is shown to be NP-complete, and several polynomial-time heuristics are proposed to tackle this problem. Comprehensive simulations on both synthetic and real application graphs show that our proposed heuristics can meet all the given constraints, while reducing the energy consumption.

This findings have been presented at the ICPADS 2018 conference [26].

## 7.16. Malleable task-graph scheduling with a practical speed-up model

Scientific workloads are often described by Directed Acyclic task Graphs. Indeed, DAGs represent both a theoretical model and the structure employed by dynamic runtime schedulers to handle HPC applications. A natural problem is then to compute a makespan-minimizing schedule of a given graph. In this paper, we are motivated by task graphs arising from multifrontal factorizations of sparse matrices and therefore work under the following practical model. Tasks are malleable (i.e., a single task can be allotted a time-varying number of processors) and their speedup behaves perfectly up to a first threshold, then speedup increases linearly, but not perfectly, up to a second threshold where the speedup levels off and remains constant.

After proving the NP-hardness of minimizing the makespan of DAGs under this model, we study several heuristics. We propose model-optimized variants for PROPSCHEDULING, widely used in linear algebra application scheduling, and FLOWFLEX. GREEDYFILLING is proposed, a novel heuristic designed for our speedup model, and we demonstrate that PROPSCHEDULING and GREEDYFILLING are 2-approximation algorithms. In the evaluation, employing synthetic data sets and task graphs arising from multifrontal factorization, the proposed optimized variants and GREEDYFILLING significantly outperform the traditional algorithms, whereby GREEDYFILLING demonstrates a particular strength for balanced graphs.

These findings have been published in the IEEE TPDS journal [16].

## 7.17. Performance and scalability of the block low-rank multifrontal factorization on multicore architectures

Matrices coming from elliptic Partial Differential Equations have been shown to have a low-rank property which can be efficiently exploited in multifrontal solvers to provide a substantial reduction of their complexity. Among the possible low-rank formats, the Block Low-Rank format (BLR) is reasonably easy to use in a general purpose multifrontal solver and its potential compared to standard (full-rank) solvers has been demonstrated. Recently, new variants have been introduced and it was proved that they can further reduce the complexity but their performance remained to be analyzed. We develop a multithreaded BLR factorization, and analyze its efficiency and scalability in shared-memory multicore environments. We identify the challenges posed by the use of BLR approximations in multifrontal solvers and put forward several algorithmic variants of the BLR factorization that overcome these challenges by improving its efficiency and scalability. We illustrate the performance analysis of the BLR multifrontal factorization with numerical experiments on a large set of problems coming from a variety of real-life applications.

This work has been accepted for publication in the ACM Transactions on Mathematical Software [5].

## 7.18. On exploiting sparsity of multiple right-hand sides in sparse direct solvers

The cost of the solution phase in sparse direct methods is sometimes critical. It can be larger than that of the factorization in applications where systems of linear equations with thousands of right-hand sides (RHS) must be solved. In this work, we focus on the case of multiple sparse RHS with different nonzero structures in each column. In this setting, vertical sparsity reduces the number of operations by avoiding computations on rows that are entirely zero, and horizontal sparsity goes further by performing each elementary solve operation only on a subset of the RHS columns. To maximize the exploitation of horizontal sparsity, we propose a new algorithm to build a permutation of the RHS columns. We then propose an original approach to split the RHS columns into a minimal number of blocks, while reducing the number of operations down to a given threshold. Both algorithms are motivated by geometric intuitions and designed using an algebraic approach, so that they can be applied to general systems. We demonstrate the effectiveness of our algorithms on systems coming from real applications and compare them to other standard approaches. We also give some perspectives and possible applications.

This work has been accepted for publication in the SIAM Journal on Scientific Computing [6].

## 7.19. Efficient use of sparsity by direct solvers applied to 3D controlled-source EM problems

Controlled-source electromagnetic (CSEM) surveying becomes a widespread method for oil and gas exploration, which requires fast and efficient software for inverting large-scale EM datasets. In this context, one often needs to solve sparse systems of linear equations with a *large* number of *sparse* right-hand sides, each corresponding to a given transmitter position. Sparse direct solvers are very attractive for these problems, especially when combined with low-rank approximations which significantly reduce the complexity and the cost of the factorization. In the case of thousands of right-hand sides, the time spent in the sparse triangular solve tends to dominate the total simulation time and here we propose several approaches to reduce it. A significant reduction is demonstrated for marine CSEM application by utilizing the sparsity of the right-hand sides (RHS) and of the solutions that results from the geometry of the problem. Large gains are achieved by restricting computations at the forward substitution stage to exploit the fact that the RHS matrix might have empty rows (*vertical sparsity*) and/or empty blocks of columns within a non-empty row (*horizontal sparsity*). We also adapt the parallel algorithms that were designed for the factorization to solve-oriented algorithms and describe performance optimizations particularly relevant for the very large numbers of right-hand sides of the CSEM application. We show that both the operation count and the elapsed time for the solution phase can be significantly reduced. The total time of CSEM simulation can be divided by approximately a factor of 3 on all the matrices from our set (from 3 to 30 million unknowns, and from 4 to 12 thousands RHSs).

These findings are described in a technical report [37] and will be submitted for publication.

## 7.20. A Generic Approach to Scheduling and Checkpointing Workflows

We dealt with scheduling and checkpointing strategies to execute scientific workflows on failure-prone large-scale platforms. To the best of our knowledge, this work was the first to target fail-stop errors for arbitrary workflows. Most previous work addresses soft errors, which corrupt the task being executed by a processor but do not cause the entire memory of that processor to be lost, contrarily to fail-stop errors. We revisited classical mapping heuristics such as HEFT and MINMIN and complement them with several checkpointing strategies. The objective was to derive an efficient trade-off between checkpointing every task (CKPTALL), which is an overkill when failures are rare events, and checkpointing no task (CKPTNONE), which induces dramatic re-execution overhead even when only a few failures strike during execution. Contrarily to previous work, our approach applies to arbitrary workflows, not just special classes of dependence graphs such as M-SPGs (Minimal Series-Parallel Graphs). Extensive experiments report significant gain over both CKPTALL and CKPTNONE, for a wide variety of workflows.

This findings have been presented at the ICPP 2018 conference [28].

## 7.21. Scheduling independent stochastic tasks under deadline and budget constraints

We studied scheduling strategies for the problem of maximizing the expected number of tasks that can be executed on a cloud platform within a given budget and under a deadline constraint. The execution times of tasks follow IID probability laws. The main questions are how many processors to enroll and whether and when to interrupt tasks that have been executing for some time. We provide complexity results and an asymptotically optimal strategy for the problem instance with discrete probability distributions and without deadline. We extend the latter strategy for the general case with continuous distributions and a deadline and we design an efficient heuristic which is shown to outperform standard approaches when running simulations for a variety of useful distribution laws.

This findings have been presented at the SBAC-PAD 2018 conference [23].

**SOCRATE Project-Team**

# 6. New Results

## 6.1. Multi-User Communications

Activities in axis 2 primarily focus on communicating multi-user systems. They represent the core of the research activity that will be pursued in Maracas team.

The first pillar of our research concerns the evaluation of fundamental limits of wireless systems (e.g. capacity) often express as a fundamental tradeoff : energy efficiency - spectral efficiency tradeoff, rate versus reliability, information versus energy transfert,... Our work relies mostly on information theory, signal processing, estimation theory and game theory.

The second pillar concerns the evaluation of real systems and their performance is confronted to the above mentioned fundamental limits. These activities rely on strong collaborations with industry (Nokia, Orange, SigFox, Sequans, SPIE-ICS,...) We also manage the FIT/CorteXlab testbed offering a remote access to a worldwide unique platform.

Beyond these two pillars, we also explore new research areas where our background is relevant. These prospective activities are performed with external collaborations and prepare the future activity of Maracas team. This year we explored molecular communications (supported by an Inria exploratory project), smart grids in collaboration with Sheffield, VLC in association with Agora team or Privacy preservation in collaboration with Privatics team.

### *6.1.1. Fundamental limits in communications*

#### *6.1.1.1. Variations on point to point capacity and related tools*

In [31] discrete approximations of the capacity are introduced where the input distribution is constrained to be discrete in addition to any other constraints on the input. For point-to-point memoryless additive noise channels, rates of convergence to the capacity of the original channel are established for a wide range of channels for which the capacity is finite. These results are obtained by viewing discrete approximations as a capacity sensitivity problem, where capacity losses are studied when there are perturbations in any of the parameters describing the channel. In particular, it is shown that the discrete approximation converges arbitrarily close to the channel capacity at rate $O(\Delta)$, where $\Delta$ is the discretization level of the approximation. Examples of channels where this rate of convergence holds are also given, including additive Cauchy and inverse Gaussian noise channels.

In [30] the properties of finite frames are explored. Finite frames are sequences of vectors in finite dimensional Hilbert spaces that play a key role in signal processing and coding theory. We studied the class of tight unit-norm frames for $\mathbb{C}^d$ that also form regular schemes, called tight regular schemes (TRS). Many common frames that arise in applications such as equiangular tight frames and mutually unbiased bases fall in this class. We investigate characteristic properties of TRSs and prove that for many constructions, they are intimately connected to weighted 1-designs—arising from quadrature rules for integrals over spheres in $\mathbb{C}^d$ with weights dependent on the Voronoi regions of each frame element.

#### *6.1.1.2. Interference channel with feedback*

The interference channel is a well-known model used to represent simultaneous transmissions in a wireless environment. In the framework of Victor Quintero's PhD, we explored the performance of this model with noisy feedbacks.

In [35], an achievable $\eta$-Nash equilibrium ($\eta$-NE) region for the two-user Gaussian interference channel with noisy channel-output feedback is presented for all $\eta \geq 1$. This result is obtained in the scenario in which each transmitter-receiver pair chooses its own transmit-receive configuration in order to maximize its own individual information transmission rate. At an $\eta$-NE, any unilateral deviation by either of the pairs does not increase the corresponding individual rate by more than $\eta$ bits per channel use.

In [6], the capacity region of the linear deterministic interference channel with noisy channel-output feedback (LD-IC-NF) is fully characterized. The proof of achievability is based on random coding arguments and rate splitting; block-Markov superposition coding; and backward decoding. The proof of converse reuses some of the existing outer bounds and includes new ones obtained using genie-aided models. Following the insight gained from the analysis of the LD-IC-NF, an achievability region and a converse region for the two-user Gaussian interference channel with noisy channel-output feedback (GIC-NF) are presented. Finally, the achievability region and the converse region are proven to approximate the capacity region of the G-IC-NF to within $4.4$ bits.

*6.1.1.3. Wiretap channel*

The Wiretap channel allows to address the secrecy constraint in an information theory framework. In [13], an analysis of an input distribution that achieves the secrecy capacity of a general degraded additive noise wiretap channel is presented. In particular, using convex optimization methods, an input distribution that achieves the secrecy capacity is characterized by conditions expressed in terms of integral equations. The new conditions are used to study the structure of the optimal input distribution for three different additive noise cases: vector Gaussian; scalar Cauchy; and scalar exponential.

*6.1.1.4. Simultaneous Information and Energy Transmission*

Simultaneous information and energy transmission (SIET) is an active research problem and aims at providing energy and information simultaneously from transmitters to receivers. We explore the optimal trade-offs in different settings.

In [34], a non-asymptotic analysis of the fundamental limits of simultaneous energy and information transmission (SEIT) is presented. The notion of information-capacity region, i.e., the largest set of simultaneously achievable information and energy rates, is revisited in a context in which transmissions occur within a finite number of channel uses and strictly positive error decoding probability and energy shortage probability are tolerated. The focus is on the case of one transmitter, one information receiver and one energy harvester communicating through binary symmetric memoryless channels. In this case, the information-capacity region is approximated and the trade-off between information rate and energy rate is thoroughly studied.

In [5], the fundamental limits of simultaneous information and energy transmission (SIET) in the two-user Gaussian interference channel (G-IC) with and without perfect channel-output feedback are approximated by two regions in each case, i.e., an achievable region and a converse region. When the energy transmission rate is normalized by the maximum energy rate the approximation is within a constat gap. In the proof of achievability, the key idea is the use of power-splitting between two signal components: an information-carrying component and a no-information component. The construction of the former is based on random coding arguments, whereas the latter consists in a deterministic sequence known by all transmitters and receivers. The proof of the converse is obtained via cut-set bounds, genie-aided channel models, Fano's inequality and some concentration inequalities considering that channel inputs might have a positive mean. Finally, the energy transmission enhancement due to feedback is quantified and it is shown that feedback can at most double the energy transmission rate at high signal to noise ratios.

*6.1.1.5. Modeling Interference in Large-Scale Uplink SCMA*

Massive connectivity is a fundamental challenge for IoT, as discussed in the next section from a practical perspective. From a theoretical perspective, we propose to relax the assumption of Gaussian interference.

Fast varying active transmitter sets with very short length transmissions arise in communications for the Internet of Things. As a consequence, the interference is dynamic, leading to non-Gaussian statistics. At the same time, the very high density of devices is motivating non-orthogonal multiple access (NOMA) techniques, such as sparse code multiple access (SCMA). In [2], we study the statistics of the dynamic interference from devices using SCMA. In particular, we show that the interference is $\alpha-$stable with non-trivial dependence structure for large scale networks modeled via Poisson point processes. Moreover, the interference on each frequency band is shown to be sub-Gaussian $\alpha-$stable in the special case of disjoint SCMA codebooks. We investigate the impact of the $\alpha-$stable interference on achievable rates and on the optimal density of devices. Our analysis suggests that ultra dense networks are desirable even with $\alpha-$stable interference.

This contribution is a good introduction of the next section where the performance of IoT access techniques are evaluated.

*6.1.1.6. General Massive Machine Type Communications Uplink*

Non Orthogonal Multiple Access (NOMA) is expected to play an important role for IoT networks, allowing to reduce signaling overheads and to maximize the capacity of dense networks with multiple packets simultaneous transmission. In the uplink, NOMA can improve significantly the performance of an ALOHA random access if the receiver implements a multiuser detection algorithm. In [11], we compared the performance of a code domain NOMA with a classical ALOHA protocol, through simulations. The code domain NOMA uses random Gaussian codes at the transmitters and exploits compressive sensing at the receiver to maximize users detection and to minimize symbol error rates.

As the number of machine type communications increases at an exponential rate, new solutions have to be found in order to deal with the uplink traffic. At the same time, new types of Base Stations (BS) that use a high number of antennas are being designed, and their beamforming capabilities can help to separate signals that have different angles of arrivals. In [15], we consider a network where a BS serves a high number of nodes that lacks a receive chain, and we analyze the evolution of the outage probability as a function of the number of antennas at the BS. We then study the effect of an angle offset between the main beam and the desired node's direction in order to provide realistic results in a beam-switching scenario.

*6.1.1.7. Multiple Base Stations Diversity for UNB Systems*

In the framework of the long-term collaboration with Sigfox, the PhD of Yuqi Mo defended mast December, explored the performance of Ultra Narrow Band (UNB) with a focus on sophisticated signal processing techniques such as multi-BS processing or successive interference cancellation (SIC). UNB (Ultra Narrow Band) is one of the technologies dedicated to low-power wide-area communication for IoT, currently exploited by SigFox

In [33], [18], the specificity of UNB is the Aloha-type channel access scheme, asynchronous in both time and frequency domain. This randomness can cause partial spectral interference. In this paper, we take advantage of the spatial diversity of multiple base stations to improve the UNB performance, by using selection combining. In the presence of pathloss and spectral randomness of UNB, the channels are considered correlated. A theoretical analysis of outage probability is demonstrated by considering this correlation, for the case of 2 base stations. This methodology of probability computing can be extended to $K$ BSs. The diversity of multiple receivers is proved to be beneficial in enhancing the performance of UNB networks. This gain is shown to be related to the density of the base stations, as well as the distance between each of them. In [8], we propose to apply signal combining and interference cancellation technologies across multiple base stations in UNB networks, in order to take advantage of their spatial diversity. We evaluate and compare the performance enhancement of each technology, compared to single BS case. These technologies exploiting multi-BS diversity are proved to be significantly beneficial in improving UNB networks' scalability. We can gain until 28 times better performance with one iteration global SIC. We highlight that these results provide us a choice among the technologies according to the improvement needs and the implementation complexity.

### *6.1.2. Contributions in other application fields*

*6.1.2.1. Molecular communications*

Molecular communications is emerging as a technique to support coordination in nanonetworking, particularly in biochemical systems. In complex biochemical systems such as in the human body, it is not always possible to view the molecular communication link in isolation as chemicals in the system may react with chemicals used for the purpose of communication. There are two consequences: either the performance of the molecular communication link is reduced; or the molecular link disrupts the function of the biochemical system. As such, it is important to establish conditions when the molecular communication link can coexist with a biochemical system. In [4], we develop a framework to establish coexistence conditions based on the theory of chemical reaction networks. We then specialize our framework in two settings: an enzyme-aided molecular communication system; and a low-rate molecular communication system near a general biochemical

system. In each case, we prove sufficient conditions to ensure coexistence. In [29], we develop a general framework for the coexistence problem by drawing an analogy to the cognitive radio problem in wireless communication systems. For the particularly promising underlay strategy, we propose a formalization and outline key consequences.

Another key challenge in nanonetworking is to develop a means of coordinating a large number of nanoscale devices. Devices in molecular communication systems—once information molecules are released— are typically viewed as passive, not reacting chemically with the information molecules. While this is an accurate model in diffusion-limited links, it is not the only scenario. In particular, the dynamics of molecular communication systems are more generally governed by reaction-diffusion, where the reaction dynamics can also dominate. This leads to the notion of reaction-limited molecular communication systems, where the concentration profiles of information molecules and other chemical species depends largely on reaction kinetics. In this regime, the system can be approximated by a chemical reaction network. In [14], we exploit this observation to design new protocols for both point-to-point links with feedback and networks for event detection. In particular, using connections between consensus and advection theory and reaction networks lead to simple characterizations of equilibrium concentrations, which yield simple—but accurate— design rules even for networks with a large number of devices.

*6.1.2.2. Smart Grids*

Smart grids is another application field where information theory and signal processing can be useful. During 2018, we addressed security issues. In [41], random attacks that jointly minimize the amount of information acquired by the operator about the state of the grid and the probability of attack detection are presented. The attacks minimize the information acquired by the operator by minimizing the mutual information between the observations and the state variables describing the grid. Simultaneously, the attacker aims to minimize the probability of attack detection by minimizing the Kullback-Leibler (KL) divergence between the distribution when the attack is present and the distribution under normal operation. The resulting cost function is the weighted sum of the mutual information and the KL divergence mentioned above. The trade-off between the probability of attack detection and the reduction of mutual information is governed by the weighting parameter on the KL divergence term in the cost function. The probability of attack detection is evaluated as a function of the weighting parameter. A sufficient condition on the weighting parameter is given for achieving an arbitrarily small probability of attack detection. The attack performance is numerically assessed on the IEEE 30-Bus and 118-Bus test systems.

*6.1.2.3. Privacy and tracking*

In a joint work with Privatics team, we presented in [40] the analysis of an Ultrasound-based tracking application. By analyzing several mobile applications along with the network communication and sample of the original audio signal, we were able to reverse engineer the ultrasonic communications and some other elements of the system. Based on those finding we show how arbitrary ultrasonic signal can be generated and how to perform jamming. Finally we analyze a real world deployment and discuss privacy implications.

*6.1.2.4. VLC*

In a joint work with Agora, we present in [12] our efforts to design a communication system between an ordinary RGB light emitting diode and a smart-phone. This work in progress presents our preliminary findings obtained investigating this poorly known and unusual channel. We give engineering insights on driving an RGB light emitting diode for camera communication and discuss remaining challenges. Finally, we propose possible solutions to cope with these issues that are blockers for a user ready implementation.

*6.1.2.5. Intelligent Transport*

On-demand transport has been disrupted by Uber and other providers, which are challenging the traditional approach adopted by taxi services. Instead of using fixed passenger pricing and driver payments, there is now the possibility of adaptation to changes in demand and supply. Properly designed, this new approach can lead to desirable tradeoffs between passenger prices, individual driver profits and provider revenue. However, pricing and allocations - known as mechanisms - are challenging problems falling in the intersection of economics and computer science. In [3], we develop a general framework to classify mechanisms in on-demand transport.

Moreover, we show that data is key to optimizing each mechanism and analyze a dataset provided by a real-world on-demand transport provider. This analysis provides valuable new insights into efficient pricing and allocation in on-demand transport.

## 6.2. Flexible Radio Front-End

Activities in this axis could globally be divided in two main topics: low-power wireless sensors (with applications in wearable devices, guided propagation for ventilation systems, and tag-to-tag RFID), and optimization of waveforms (for wake-up radio receivers and wireless power transfer).

### *6.2.1. Low-Power WSN*

Wearable sensors for health monitoring can enable the early detection of various symptoms, and hence rapid remedial actions may be undertaken. In particular, the monitoring of cardiac events by using such wearable sensors can provide real-time and more relevant diagnosis of cardiac arrhythmia than classical solutions. However, such devices usually use batteries, which require regular recharging to ensure long-term measurements. In the framework of a local collaborative project, we therefore designed and evaluated a connected sensor for the ambulatory monitoring of cardiac events, which can be used as an autonomous device without the need of a battery. Even when using off-the-shelf, low-cost integrated circuits, by optimizing both the hardware and software embedded in the device, we were able to reduce the energy consumption of the entire system to below 0.4 mW while measuring and storing the ECG on a non-volatile memory. Moreover, in this project, a power-management circuit able to store energy collected from the radio communication interface is proposed, able to make the connected sensor fully autonomous. Initial results show that this sensor could be suitable for a truly continuous and long-term monitoring of cardiac events [32].

In collaboration with Atlantic, we have done here a preliminary study [37], [23] of wireless transmissions using the ventilation metallic ducts as waveguides. Starting from the waveguide theory, we deeply studied in simulation the actual attenuation encountered by radiowaves in such a specific medium. This kind of wireless link appears to be really efficient, and therefore highly promising to implement Internet of Things (IoT) in old buildings to make them smarter. This study also expresses a very simple empirical model in order to ease dimensioning a wireless network in such conditions and a specific antenna design enabling both good performance and high robustness to the influence of the environment.

The Spie ICS- INSA Lyon chair on IoT has granted us for a PhD thesis on Scatter Radio and RFID tag-to-tag communications. Some seminal results have shown that it is actually possible to create a communication between two RFID tags, just using ambient radiowaves or a dedicated distant radio source, without the need of generating a signal from the tag itself.

### *6.2.2. Optimization of waveforms for wake-up radio and energy harvesting*

First Filter Bank Multi Carrier (FBMC) signals are employed in order to improve the performance of a quasi-passive wake-up radio receiver (WuRx) for which the addressing is performed by the means of a frequency fingerprint. The feasibility of such kind of WuRx was already demonstrated by using orthogonal frequency-division multiplexing (OFDM) signals to form the identifiers. Together with the main advantage of this approach (i.e. no base band processing needed and consequently a reduced energy consumption), one of the drawbacks is their low sensitivity. Through a set of circuit-system co-simulations, it is shown that by their characteristics, especially high Peak to Average Power Ratio (PAPR) and high out of band attenuation, FBMC signals manage to boost the sensitivity and moreover to enhance the robustness of this kind of WuRx. Moreover, we introduced robust wake-up IDs for quasi-passive wake-up receivers in an Internet of Things context[16]. These IDs can address single devices and are based on the Hadamard codes. Further a novel wake-up threshold is implemented to make the device more sensitive and robust against false wake-ups (FWUs). The wake-up procedure is simulated with a tap delay line (TDL) model for a line of sight (LOS) channel and a non line of sight (NLOS) channel. In both scenarios sufficient wake-up distances are reached with low false wake-up probabilitys (FWUPs). Additionally, the system is tested against the influence of an external bandwidth use. Finally, a recommendation for the global system is given.

In [21], we are proposing a way to maximize the DC power collected in the case of a wireless power transfer (WPT) scenario. Three main aspects are taken into account: the RF (radio frequency) source, the propagation channel and the rectifier as the main part of the energy collecting circuit. This problem is formulated as a convex optimization one. Then, as a first step towards solving this problem, a rectifier circuit was simulated by using Keysight's ADS software and, by using a classical model identification strategy i.e. Vector Fitting algorithm, the state-space model of the passive parts of this rectifier were extracted. In order to verify the extracted model, S11 input reflection coefficients and DC output voltages of the original circuit and the state-space model are compared.

### *6.2.3. UWB for localization*

Ultra Wide Band (UWB) is a wireless communication technology that is characterized, in its *impulse radio* scheme [55], by very short duration waveforms called *pulses* (in the order of few nanoseconds), using a wide band and low power spectral density. Among the many advantages offered by this technology is the fact that the arrival time of a pulse can be determined quite precisely, giving the opportunity to measure the distance between two communicating devices by estimating the flight time of the signal.

Although this technology has been known for a long time, it is only recently that cheap UWB chips have been commercialized for civilian applications. As the UWB technology is sensitive to many parameters, the effective performance of localization systems based on UWB may vary a lot compared to what is announced in datasheets. Some accuracy studies have been performed [47], [48] but few of them focus on rapid movement of the transceivers.

Indeed, indoor ranging is in itself dependent on many parameter and very difficult to evaluate objectively, but when the transceivers are moving fast (say as if they were attached to dancer's wrists), more parameters are to be taken into account: transceiver calibration, random errors, presence of obstacle, antenna orientation etc.

In [20], we study experimentally the precision of UWB ranging for rapid movements in an indoor environment, based on the technology proposed by Decawave (DW1000 [45]) whose chips have already been integrated in many commercial devices. We show in particular how to improve the precision of the distance measured by averaging the ranging over successive samples.

## 6.3. Software Radio Programming Model

### *6.3.1. Non Uniform Memory Access Analyzer*

Non Uniform Memory Access (NUMA) architectures are nowadays common for running High-Performance Computing (HPC) applications. In such architectures, several distinct physical memories are assembled to create a single shared memory. Nevertheless, because there are several physical memories, access times to these memories are not uniform depending on the location of the core performing the memory request and on the location of the target memory. Hence, threads and data placement are crucial to efficiently exploit such architectures. To help in taking decision about this placement, profiling tools are needed. In [36], we propose NUMA MeMory Ana-lyzer (NumaMMA), a new profiling tool for understanding the memory access patterns of HPC applications. NumaMMA combines efficient collection of memory traces using hardware mechanisms with original visualization means allowing to see how memory access patterns evolve over time. The information reported by NumaMMA allows to understand the nature of these access patterns inside each object allocated by the application. We show how NumaMMA can help understanding the memory patterns of several HPC applications in order to optimize them and get speedups up to 28% over the standard non optimized version.

### *6.3.2. Environments for transiently powered devices*

An important research initiative is being followed in Socrate today: the study of the new NVRAM technology and its use in ultra-low power context. NVRAM stands for Non-Volatile Radom Access Memory. Non-Volatile memory has been extising for a while (Nand Flash for instance) but was not sufficiently fast to be used as main memory. Many emerging technologies are forseen for Non-Volatile RAM to replace current RAM [50].

Socrate has started a work on the applicability of NVRAM for *transiantly powered systems*, i.e. systems which may undergo power outage at any time. This study resulted in the Sytare software presented at the NVMW conference [25] and also to the starting of an Inria Project Lab [39]: ZEP.

The Sytare software introduces a checkpointing system that takes into account peripherals (ADC, leds, timer, radio communication, etc.) present on all embedded system. Checkpointing is the natural solution to power outage: regularly save the state of the system in NVRAM so as to restore it when power is on again. However, no work on checkpointing took into account the restoration of the states of peripherals, Sytare provides this possibility. A complete description of Sytare has been accepted to IEEE Transaction on Computers [1], special issue on NVRAM.

### 6.3.3. Dynamic memory allocation for heterogeneous memory systems

In a low power system-on-chip the memory hierarchy is traditionally composed of Static RAM (SRAM) and NOR flash. The main feature of SRAM is a fast access time, while Flash memory is dense, and also non-volatile i.e. it does not require power to retain data. Because of its low writing speed, Flash memory is mostly used in a read-only fashion (e.g. for code) and the amount of SRAM is kept to a minimum in order to lower leakage power.

Emerging memory technologies exhibit different trade-offs and more heterogeneity. Non-Volatile RAM technologies like MRAM (Magnetic RAM) or RRAM (Resistive RAM) open new perspectives on power-management since they can be switched on or off at very little cost. Their characteristics are very dependent on the technology used, but it is now widely known that they will provide a high integration density and fast read access time to persistent data. NVRAM is usually not as fast as SRAM and some technologies have a limited endurance hence are not suited to store frequently modified data. In addition, most NVRAM technologies have asymmetric access times, writes being slower than reads.

In the context of embedded systems, the hardware architecture is evolving towards a model where different memory banks, with different hardware characteristics, are directly exposed to software, as it has been the case for scratchpad memories (SPM). This raises questions including:

- What is the expected performance impact of adding fast memory to a system based on NVRAM? In particular: will the addition of a small amount of fast memory result in significant performance improvement?
- How should one adapt and optimize their software memory management to leverage these new technologies?

In [10], [28], we study these questions in the perpective of dynamic memory allocation. In this first study we show, with extensive profiling how much can be gained with a clever dynamic memory allocation in the context of heterogeneous memory. We limit the study to two different memories, RAM and NVRAM for instance. This gain can go up to 15% of performance, depending of course of the performances of the different memories used. These results will be helpfull to design a clever dynamic allocator for these new architectures and also will help in the design process of new architecture for low power systems that will include NVRAM for normaly-off systems for instance.

### 6.3.4. Arithmetic for signal processing

Linear Time Invariant (LTI) filters are often specified and simulated using high-precision software, before being implemented in low-precision fixed-point hardware. A problem is that the hardware does not behave exactly as the simulation due to quantization and rounding issues. The article [7] advocates the construction of LTI architectures that behave as if the computation was performed with infinite accuracy, then converted to the low-precision output format with an error smaller than its least significant bit. This simple specification guarantees the numerical quality of the hardware, even for critical LTI systems. Besides, it is possible to derive the optimal values of all the internal data formats that ensure that the specification is met. This requires a detailed error analysis that captures not only the quantization and rounding errors, but also their infinite accumulation in recursive filters. This generic methodology is detailed for the case of low-precision LTI filters in the Direct Form I implemented in FPGA logic. It is demonstrated by a fully automated and open-source architecture generator tool integrated in FloPoCo, and validated on a range of Infinite Impulse Response filters.

### 6.3.5. Karatsuba multipliers on modern FPGAs

The Karatsuba method is a well-known technique to reduce the complexity of large multiplications. However it is poorly suited to the rectangular 17x25-bit multipliers embedded in recent Xilinx FPGAs: The traditional Karatsuba approach must under-use them as square 18x18 ones. In [17], the Karatsuba method is extended to efficiently use such rectangular multipliers to build larger multipliers. Rectangular multipliers can be efficiently exploited if their input word sizes have a large greatest common divider. In the Xilinx FPGA case, this can be obtained by using the 17x25 embedded multipliers as 16x24. The obtained architectures are implemented with due detail to architectural features such as the pre-adders and post-adders available in Xilinx DSP blocks. They are synthesized and compared with traditional Karatsuba, but also with (non-Karatsuba) state-of-the-art tiling techniques that make use of the full rectangular multipliers. The proposed technique improves resource consumption and performance for multipliers of numbers larger than 64 bits.

### 6.3.6. PyGA: a Python to FPGA compiler prototype

In a collaboration with Intel, Yohann Uguen has worked on a compiler of Python to FPGA [22]. Based on the Numba Just-In-Time (JIT) compiler for Python and the Intel FPGA SDK for OpenCL, it allows any Python user to use a FPGA card as an accelerator for Python seamlessly, albeit with limited performance so far.

### 6.3.7. General computer arithmetic

A second edition of the Handbook for Floating-Point Arithmetic has been published [38].

With colleagues from Aric, we have worked on a critical review [42] of the Posit system, a proposed alternative to the prevalent floating-point format.

# SPADES Project-Team

# 6. New Results

## 6.1. Design and Programming Models

**Participants:** Pascal Fradet, Alain Girault, Gregor Goessler, Xavier Nicollin, Christophe Prévot, Sophie Quinton, Arash Shafiei, Jean-Bernard Stefani, Martin Vassor, Souha Ben Rayana.

### *6.1.1. A multiview contract theory for cyber-physical system design and verification*

The design and verification of critical cyber-physical systems is based on a number of models (and corresponding analysis techniques and tools) representing different viewpoints such as function, timing, security and many more. Overall correctness is guaranteed by mostly informal, and therefore basic, arguments about the relationship between these viewpoint-specific models. More precisely, the assumptions that a viewpoint-specific analysis makes on the other viewpoints remain mostly implicit, and whenever explicit they are handled mostly manually. In [11], we argue that the current design process over-constrains the set of possible system designs and that there is a need for methods and tools to formally relate viewpoint-specific models and corresponding analysis results. We believe that a more flexible contract-based approach could lead to easier integration, to relaxed assumptions, and consequently to more cost efficient systems while preserving the current modelling approach and its tools.

The framework we have in mind would provide viewpoint specific contract patterns guaranteeing inter-viewpoint consistency in a flexible manner. At this point, most of the work remains to be done. On the application side, we need a more complete picture of existing inter-viewpoint models. We also need the theory required for the correctness proofs, but it should be based on the needs on the application side.

### *6.1.2. End-to-end worst-case latencies of task chains for flexibility analysis*

In collaboration with Thales, we address the issue of change during design and after deployment in safety-critical embedded system applications. More precisely, we focus on timing aspects with the objective to anticipate, at design time, future software evolutions and identify potential schedulability bottlenecks. The work presented in this section is the PhD topic of Christophe Prévot, in the context of a collaboration with Thales TRT, and our algorithms are being implemented in the Thales tool chain, in order to be used in industry.

This year, we have completed our work on the analysis of end-to-end worst-case latencies of task chains [10] that was needed to extend our approach for quantifying the flexibility, with respect to timing, of real-time systems made of chains of tasks. In a nutshell, flexibility is the property of a given system to accommodate changes in the future, for instance the modification of some of the parameters of the system, or the addition of a new task in the case of a real-time system.

One major issue that hinders the use of performance analysis in industrial design processes is the pessimism inherent to any analysis technique that applies to realistic system models (*e.g.*, , systems with task chains). Indeed, such analyses may conservatively declare unschedulable systems that will in fact never miss any deadlines. The two main avenues for improving this are (i) computing tighter upper bounds on the worst-case latencies, and (ii) measuring the pessimism, which requires to compute also guaranteed lower bounds. A lower bound is guaranteed by providing an actual system execution exhibiting a behavior as close to the worst case as possible. As a first step, we focus in [10] on uniprocessor systems executing a set of sporadic or periodic hard real-time task chains. Each task has its own priority, and the chains are scheduled according to the fixed-priority preemptive scheduling policy. Computing the worst-case end-to-end latency of each chain is complex because of the intricate relationship between the task priorities. Compared to state of the art analyses, we propose here tighter upper bounds, as well as lower bounds on these worst-case latencies. Our experiments show the relevance of lower bounds on the worst-case behavior for the industrial design of real-time embedded systems.

Based on our end-to-end latency analysis for task chains, we have also proposed an extension of the concept of slack to task chains and shown how it can be used to perform flexibility analysis and sensitivity analysis. This solution is particularly relevant for industry as it provides means by which the system designer can anticipate the impact on timing of software evolutions, at design time as well as after deployment.

### 6.1.3. Location graphs

We have introduced the location graph model [58] as an expressive framework for the definition of component-based models able to deal with dynamic software configurations with sharing and encapsulation constraints. We have completed a first study of the location graph behavioral theory (under submission), initiated its formalization in Coq, and an implementation of the location framework with an emphasis of the expression of different isolation and encapsulation constraints.

We are now studying conservative extensions to the location graph framework to support the compositional design of heterogeneous hybrid dynamical systems and their attendant notions of approximate simulations [60].

In collaboration with the Spirals team at Inria Lille – Nord Europe, we have applied the location framework for the definition of a pivot model for the description of software configurations in a cloud computing environment. We have shown how to interpret in our pivot model several configuration management models and languages including TOSCA, OCCI, Docker Compose, Aeolus, OpenStack HOT.

### 6.1.4. Dynamicity in dataflow models

Recent dataflow programming environments support applications whose behavior is characterized by dynamic variations in resource requirements. The high expressive power of the underlying models (*e.g.*, Kahn Process Networks or the CAL actor language) makes it challenging to ensure predictable behavior. In particular, checking *liveness* (*i.e.*, no part of the system will deadlock) and *boundedness* (*i.e.*, the system can be executed in finite memory) is known to be hard or even undecidable for such models. This situation is troublesome for the design of high-quality embedded systems. In the past few years, we have proposed several parametric dataflow models of computation (MoCs) [40], [31], we have written a survey providing a comprehensive description of the existing parametric dataflow MoCs [34], and we have studied *symbolic* analyses of dataflow graphs [35]. More recently, we have proposed an original method to deal with lossy communication channels in dataflow graphs [39].

We are now studying models allowing dynamic reconfigurations of the *topology* of the dataflow graphs. In particular, many modern streaming applications have a strong need for reconfigurability, for instance to accommodate changes in the input data, the control objectives, or the environment.

We have proposed a new MoC called Reconfigurable Dataflow (RDF) [15]. RDF extends SDF with transformation rules that specify how the topology and actors of the graph may be reconfigured. Starting from an initial RDF graph and a set of transformation rules, an arbitrary number of new RDF graphs can be generated at runtime. The major quality of RDF is that it can be statically analyzed to guarantee that all possible graphs generated at runtime will be connected, consistent, and live. This is the research topic of Arash Shafiei's PhD, in collaboration with Orange Labs.

### 6.1.5. Monotonic prefix consistency in distributed systems

We have studied the issue of data consistency in distributed systems. Specifically, we have considered a distributed system that replicates its data at multiple sites, which is prone to partitions, and which is assumed to be available (in the sense that queries are always eventually answered). In such a setting, strong consistency, where all replicas of the system apply synchronously every operation, is not possible to implement. However, many weaker consistency criteria that allow a greater number of behaviors than strong consistency, are implementable in available distributed systems. We have focused on determining the strongest consistency criterion that can be implemented in a convergent and available distributed system that tolerates partitions, and we have shown that no criterion stronger than Monotonic Prefix Consistency (MPC [61], [44]) can be implemented [18].

## 6.2. Certified Real-Time Programming

**Participants:** Pascal Fradet, Alain Girault, Gregor Goessler, Xavier Nicollin, Sophie Quinton, Xiaojie Guo, Maxime Lesourd.

### *6.2.1. Time predictable programming languages and architectures*

Time predictability (PRET) is a topic that emerged in 2007 as a solution to the ever increasing unpredictability of today's embedded processors, which results from features such as multi-level caches or deep pipelines [37]. For many real-time systems, it is mandatory to compute a strict bound on the program's execution time. Yet, in general, computing a tight bound is extremely difficult [64]. The rationale of PRET is to simplify both the programming language and the execution platform to allow more precise execution times to be easily computed [27].

We have extended the PRET-C compiler [25] in order to make it energy aware. To achieve this, we use dynamic voltage and frequency scaling (DFVS) and we insert DVFS control points in the control flow graph of the PRET-C program. Several difficulties arise: (i) the control flow graph is concurrent, (ii) the resulting optimization problem is a time and energy multi-criteria problem, and (iii) since we consider PRET-C programs, we actually address the Worst-Case Execution Time (WCET) and the Worst-Case Energy Consumption (WCEC). Thanks to a novel ILP formulation and to a bicriteria heuristic, we are able to address the two objectives jointly and to compute, for each PRET-C program, the Pareto front of the non-dominated solutions in the 2D space (WCET,WCEC) [63]. We have recently improved this result to reduce the complexity of the algorithm and to produce the *optimal* Pareto front. This is the topic of Jia Jie Wang's postdoc.

Moreover, within the CAPHCA project, we have proposed a new approach for predictable inter-core communication between tasks allocated on different cores. Our approach is based on the execution of synchronous programs written in the FOREC programming language on deterministic architectures called PREcision Timed. The originality resides in the time-triggered model of computation and communication that allows for a very precise control over the thread execution. Synchronisation is done via configurable Time Division Multiple Access (TDMA) arbitrations (either physical or conceptual) where the optimal size and offset of the time slots are computed to reduce the inter-core synchronization costs. Results show that our model guarantees time-predictable inter-core communication, the absence of concurrent accesses (without relying on hardware mechanisms), and allows for optimized execution throughput. This is the topic of Nicolas Hili's postdoc.

### *6.2.2. Schedulability of weakly-hard real-time systems*

We focus on the problem of computing tight deadline miss models for real-time systems, which bound the number of potential deadline misses in a given sequence of activations of a task. In practical applications, such guarantees are often sufficient because many systems are in fact not hard real-time [4]. A weakly-hard real-time guarantee specifies an upper bound on the maximum number m of deadline misses of a task in a sequence of k consecutive executions. Based on our previous work on Typical Worst-Case Analysis [4], [8], we have introduced in [13] the first verification method which is able to provide weakly-hard real-time guarantees for tasks and task chains in systems with multiple resources under partitioned scheduling with fixed priorities. All existing weakly-hard real-time verification techniques are restricted today to systems with a single resource. Our verification method is applied in the context of switched networks with traffic streams between nodes, and we demonstrate its practical applicability on an automotive case study.

### *6.2.3. Synthesis of switching controllers using approximately bisimilar multiscale abstractions*

The use of discrete abstractions for continuous dynamics has become standard in hybrid systems design (see *e.g.*, [60] and the references therein). The main advantage of this approach is that it offers the possibility to leverage controller synthesis techniques developed in the areas of supervisory control of discrete-event systems [56]. The first attempts to compute discrete abstractions for hybrid systems were based on traditional systems behavioral relationships such as simulation or bisimulation, initially proposed for discrete systems most notably in the area of formal methods. These notions require inclusion or equivalence of observed

behaviors which is often too restrictive when dealing with systems observed over metric spaces. For such systems, a more natural abstraction requirement is to ask for closeness of observed behaviors. This leads to the notions of approximate simulation and bisimulation introduced in [42]. These approaches are based on sampling of time and space where the sampling parameters must satisfy some relation in order to obtain abstractions of a prescribed precision. In particular, the smaller the time sampling parameter, the finer the lattice used for approximating the state-space; this may result in abstractions with a very large number of states when the sampling period is small. However, there are a number of applications where sampling has to be fast; though this is generally necessary only on a small part of the state-space.

We are currently investigating an approach using mode sequences as symbolic states for our abstractions. By using mode sequences of variable length we are able to adapt the granularity of our abstraction to the dynamics of the system, so as to automatically trade off precision against controllability of the abstract states.

### 6.2.4. A Markov Decision Process approach for energy minimization policies

In the context of independent real-time sporadic jobs running on a single-core processor equipped with Dynamic Voltage and Frequency Scaling (DVFS), we have proposed a Markov Decision Process approach (MDP) to compute the scheduling policy that dynamically chooses the voltage and frequency level of the processor such that each job meets its deadline and the total energy consumption is minimized. We distinguish two cases: the finite case (there is a fixed time horizon) and the infinite case. In the finite case, several *offline* solutions exist, which all use the complete knowledge of all the jobs that will arrive within the time horizon [65], *i.e.*, their size and deadlines. But clearly this is unrealistic in the embedded context where the characteristics of the jobs are not known in advance. Then, an optimal offline policy called Optimal Available (OA) has been proposed in [65]. Our goal was to improve this result by taking into account the *statistical characteristics* of the upcoming jobs. When such information is available (for instance by profiling the jobs based on execution traces), we have proposed several speed policies that optimize the *expected* energy consumption. We have shown that this general constrained optimization problem can be modeled as an unconstrained MDP by choosing a proper state space that also encodes the constraints of the problem. In particular, this implies that the optimal speed at each time can be computed using a *dynamic programming* algorithm (under a finite horizon), and that the optimal speed at any time $t$ will be a deterministic function of the current state at time $t$ [41]. Under an infinite horizon, we use a *Value Iteration* algorithm.

This work led us to compare several existing speed policies with respect to their feasibility. Indeed, the policies (OA) [65], (AVR) [65], and (BKP) [29] all assume that the maximal speed $S_{max}$ available on the processor is infinite, which is an unrealistic assumption. For these three policies and for our (MDP) policy, we have established necessary and sufficient conditions on $S_{max}$ guaranteeing that no job will ever miss its deadline.

This is the topic of Stephan Plassart's PhD, funded by the CASERM Persyval project.

### 6.2.5. Formal proofs for schedulability analysis of real-time systems

We have started to lay the foundations for computer-assisted formal verification of real-time systems analyses. Specifically, we contribute to Prosa [23], a Coq library of reusable concepts and proofs for real-time systems analysis. A key scientific challenge is to achieve a modular structure of proofs, *e.g.*, for response time analysis. Our goal is to use this library for:

1. a better understanding of the role played by some assumptions in existing proofs;
2. a formal verification and comparison of different analysis techniques; and
3. the certification of results of existing (*e.g.*, industrial) analysis tools.

Our first major result [16] is a task model that generalizes the digraph model [59] and its corresponding analysis for fixed-priority scheduling with limited preemption. The motivation for this work, which is not yet fully proven in Coq, is to obtain a formally verified schedulability analysis for a very expressive task model. In the context of computer assisted verification, it permits to factorize the correctness proofs of a large number of analyses. The digraph task model seems a good candidate due to its powerful expressivity. Alas, its ability to capture dependencies between arrival and execution times of jobs of different tasks is very limited. Our extended model can capture dependencies between jobs of the same task as well as jobs of different tasks. We

provide a correctness proof of the analysis that is written in a way amenable to its formalization in the Coq proof assistant. Despite being much more general, the Response Time Analysis (RTA) for our model is not significantly more complex than the original one. Also, it underlines similarities between existing analyses, in particular the analysis for the digraph model and Tindell's offset model [62].

A second major result is CertiCAN, a tool produced using Coq for the formal certification of CAN analysis results. Result certification is a process that is light-weight and flexible compared to tool certification, which makes it a practical choice for industrial purposes. The analysis underlying CertiCAN is based on a combined use of two well-known CAN analysis techniques [62] that makes it computationally efficient. Experiments demonstrate that CertiCAN is able to certify the results of RTaW-Pegase, an industrial CAN analysis tool, even for large systems. This result paves the way for a broader acceptance of formal tools for the certification of real-time systems analysis results. Beyond CertiCAN, we believe that this work is significant in that it demonstrates the advantage of result certification over tool certification for the RTA of CAN buses. In addition, the underlying technique can be reused for any other system model for which there exist RTAs with different levels of precision. This work will be presented at RTAS 2019.

In parallel, we have completed and published in [17] a Coq formalization of Typical Worst-Case Analysis (TWCA) [4], [8], an analysis technique for weakly-hard real-time systems. Our generic analysis is based on an abstract model that characterizes the exact properties needed to make TWCA applicable to any system model. Our results are formalized and checked using the Coq proof assistant along with the Prosa schedulability analysis library. This work opens up new research directions for TWCA by providing a formal framework for the trade-off that must be found between time efficiency and precision of the analysis. Hopefully, our generic proof will make it easier to extend TWCA to more complex models in the future. In addition, our experience with formalizing real-time systems analyses shows that it is not only a way to increase confidence in the results of the analyses; it also helps understanding their key intermediate steps, the exact assumptions required, and how they can be generalized.

### 6.2.6. Logical execution time

In collaboration with TU Braunschweig and Daimler, we have worked on the application of the Logical Execution Time (LET) paradigm [50], according to which data are read and written at predefined time instants, to the automotive industry. The LET paradigm was considered until recently by the automotive industry as not efficient enough in terms of buffer space and timing performance. The shift to embedded multicore processors has represented a game changer: The design and verification of multicore systems is a challenging area of research that is still very much in progress. Predictability clearly is a crucial issue which cannot be tackled without changes in the design process. Several OEMs and suppliers have come to the conclusion that LET might be a key enabler and a standardization effort is already under way in the automotive community to integrate LET into AUTOSAR. We have organized a Dagstuhl seminar [9] to discuss and sketch solutions to the problems raised by the use of LET in multicore systems. A white paper on the topic is under preparation.

So far, LET has been applied only at the ECU (Electronic Control Unit) level by the automotive industry. Recent developments in electric powertrains and autonomous vehicle functions raise parallel programming from the multicore level to the vehicle level where the standard LET approach cannot apply directly. We have proposed System Level LET [21], an extension of LET with relaxed synchronization requirements which allows separating network design from ECU design and makes LET applicable to automotive distributed systems.

### 6.2.7. Scheduling under multiple constraints and Pareto optimization

We have continued our work on multi-criteria scheduling, in two directions. First, in the context of dynamic applications that are launched and terminated on an embedded homogeneous multi-core chip, under execution time and energy consumption constraints, we have proposed a two layer adaptive scheduling method [26]. In the first layer, each application (represented as a DAG of tasks) is scheduled statically on subsets of cores: 2 cores, 3 cores, 4 cores, and so on. For each size of these sets (2, 3, 4, ...), there may be only one topology or several topologies. For instance, for 2 or 3 cores there is only one topology (a "line"), while for 4 cores there are three distinct topologies ("line", "square", and "T shape"). Moreover, for each topology,

we generate statically several schedules, each one subject to a different total energy consumption constraint, and consequently with a different Worst-Case Reaction Time (WCRT). Coping with the energy consumption constraints is achieved thanks to Dynamic Frequency and Voltage Scaling (DVFS). In the second layer, we use these pre-generated static schedules to reconfigure dynamically the applications running on the multi-core each time a new application is launched or an existing one is stopped. The goal of the second layer is to perform a dynamic global optimization of the configuration, such that each running application meets a pre-defined quality-of-service constraint (translated into an upper bound on its WCRT) and such that the total energy consumption be minimized. For this, we *(i)* allocate a sufficient number of cores to each active application, *(ii)* allocate the unassigned cores to the applications yielding the largest gain in energy, and *(iii)* choose for each application the best topology for its subset of cores (*i.e.*, better than the by default "line" topology). This is a joint work with Ismail Assayad (U. Casablanca, Morocco) who visited the team in 2018.

Second, we have proposed the first of its kind multi-criteria scheduling heuristics for a DAG of tasks onto an homogeneous multi-core chip. Given an application modeled as a Directed Acyclic Graph (DAG) of tasks and a multicore architecture, we produce a set of non-dominated (in the Pareto sense) static schedules of this DAG onto this multicore. The criteria we address are the execution time, reliability, power consumption, and peak temperature. These criteria exhibit complex antagonistic relations, which make the problem challenging. For instance, improving the reliability requires adding some redundancy in the schedule, which penalizes the execution time. To produce Pareto fronts in this 4-dimension space, we transform three of the four criteria into constraints (the reliability, the power consumption, and the peak temperature), and we minimize the fourth one (the execution time of the schedule) under these three constraints. By varying the thresholds used for the three constraints, we are able to produce a Pareto front of non-dominated solutions. Each Pareto optimum is a static schedule of the DAG onto the multicore. We propose two algorithms to compute static schedules. The first is a ready list scheduling heuristic called ERPOT (Execution time, Reliability, POwer consumption and Temperature). ERPOT actively replicates the tasks to increase the reliability, uses Dynamic Voltage and Frequency Scaling to decrease the power consumption, and inserts cooling times to control the peak temperature. The second algorithm uses an Integer Linear Programming (ILP) program to compute an optimal schedule. However, because our multi-criteria scheduling problem is NP-complete, the ILP algorithm is limited to very small problem instances. Comparisons showed that the schedules produced by ERPOT are on average only 10% worse than the optimal schedules computed by the ILP program, and that ERPOT outperforms the PowerPerf-PET heuristic from the literature on average by 33%. This is a joint work with Athena Abdi and Hamid Zarandi from Amirkabir University in Tehran, Iran.

## 6.3. Fault Management and Causal Analysis

**Participants:** Pascal Fradet, Alain Girault, Gregor Goessler, Jean-Bernard Stefani, Martin Vassor.

### *6.3.1. Fault Ascription in Concurrent Systems*

The failure of one component may entail a cascade of failures in other components; several components may also fail independently. In such cases, elucidating the exact scenario that led to the failure is a complex and tedious task that requires significant expertise.

The notion of causality *(did an event $e$ cause an event $e'$?)* has been studied in many disciplines, including philosophy, logic, statistics, and law. The definitions of causality studied in these disciplines usually amount to variants of the counterfactual test "$e$ is a cause of $e'$ if both $e$ and $e'$ have occurred, and in a world that is as close as possible to the actual world but where $e$ does not occur, $e'$ does not occur either". In computer science, almost all definitions of logical causality — including the landmark definition of [48] and its derivatives — rely on a causal model that. However, this model may not be known, for instance in presence of black-box components. For such systems, we have been developing a framework for blaming that helps us establish the causal relationship between component failures and system failures, given an observed system execution trace. The analysis is based on a formalization of counterfactual reasoning [6].

We are currently working on a revised version of our general semantic framework for fault ascription in [46] that satisfies a set of formally stated requirements — such as its behavior under several notions of abstraction and refinement —, and on its instantiation to acyclic models of computation, in order to compare our approach with the standard definition of *actual causality* proposed by Halpern and Pearl.

### *6.3.2. Fault Management in Virtualized Networks*

From a more applied point of view we are investigating, in the context of Sihem Cherrared's PhD thesis, approaches for fault explanation and localization in virtualized networks. In essence, Network Function Virtualization (NFV), widely adopted by the industry and the standardization bodies, is about running network functions as software workloads on commodity hardware to optimize deployment costs and simplify the life-cycle management of network functions. However, it introduces new fault management challenges including dynamic topology and multi-tenant fault isolation that we discuss in [14]. As a first step to tackle those challenges, we have extended the classical fault management process to the virtualized functions by introducing LUMEN: a Global Fault Management Framework. Our approach aims at providing the availability and reliability of the virtualized 5G end-to-end service chain. LUMEN includes the canonical steps of the fault management process and proposes a monitoring solution for all types of Network virtualization Environments. Our framework is based on open source solutions and could easily be integrated with other existing autonomic management models.

## STEEP Project-Team

# 6. New Results

## 6.1. Calibration of the Tranus Land Use Module: Optimisation-Based Algorithms, their Validation, and Parameter Selection by Statistical Model Selection

Instantiating land use and transport integrated models (LUTI modelling) is a complicated task, requiring substantial data collection, parameter estimation and expert analysis. In this work, we present a partial effort towards the automation of the calibration of Tranus, one of the most popular LUTI models. First, we give a detailed mathematical description of the activity module and the usual calibration approach. Secondly, we reformulate the estimation of the endogenous parameters called shadow prices as an optimisation problem. We also propose an optimisation algorithm for the calibration of the substitution submodel, setting a base for future fully integrated calibration. We analyse the case of transportable and non-transportable economic sectors and propose a detailed mathematical scheme for each case. We also discuss how to validate calibration results and propose to use synthetic data generated from real world problems in order to assess convergence properties and accuracy of calibration methods. Results of this methodology are presented for realistic scenarios. Finally, we propose a model selection scheme to reduce the number of shadow prices that need to be calibrated, with the aim of reducing the risk of overfitting to data. This work is published in [2].

## 6.2. Convolutional neural networks for disaggregated population mapping using open data

High resolution population count data are vital for numerous applications such as urban planning, transportation model calibration, and population growth impact measurements, among others. In this work, we present and evaluate an end-to-end framework for computing disaggregated population mapping employing convolutional neural networks (CNNs). Using urban data extracted from the OpenStreetMap database, a set of urban features are generated which are used to guide population density estimates at a higher resolution. A population density grid at a 200 by 200 meter spatial resolution is estimated, using as input gridded population data of 1 by 1 kilometer. Our approach relies solely on open data with a wide geographical coverage, ensuring replicability and potential applicability to a great number of cities in the world. Fine-grained gridded population data is used for 15 French cities in order to train and validate our model. A stand-alone city is kept out for the validation procedure. The results demonstrate that the neural network approach using massive OpenStreetMap data outperforms other approaches proposed in related works. This work is published in [5].

## 6.3. Uncertainties of Domestic Road Freight Statistics: Insights for Regional Material Flow Studies

Freight statistics are at the core of many studies in the field of industrial ecology because they depict the physical inter-dependencies of territories and allow to link worldwide productions and consumptions. Recent studies have been increasingly focusing on subnational scales, often relying on domestic freight data. In this perspective, this article analyses the uncertainties of the French domestic road freight survey, road being by far the most common mode of transport in the country. Based on a statistical analysis of the survey, we propose a model to estimate the uncertainty of any given domestic road transport flow. We also assess uncertainty reduction when averaging the flows over several years, and obtain for instance a 30% reduction for a 3-year average. We then study the impact of the uncertainties on regional material flow studies such as the Economy-Wide Material Flow Analysis of the Bourgogne region. Overall the case studies advocate for a systematic assessment of freight uncertainties, as neither the disaggregation level nor the quantities traded are good enough predictors. This justifies the need for an easy-to-implement estimation model. Finally, basic comparison with the German and Swedish surveys tend to indicate that the main conclusions presented in this article are likely to be valid in other European countries. This work is published in [3].

## 6.4. A method for downscaling open population data

To extend our ongoing work on urban sprawl indicators (see above), we have developed a method to perform dissagregated population estimations at building level using open data. Our goal is to estimate the number of people living at the fine level of individual households by using open urban data and coarse-scaled population data. First, a fine scale description of residential land use per building is built using OpenStreetMap. Then, using coarse-scale gridded population data, we perform the down-scaling for each household given their containing area for residential usage. We rely solely on open data in order to ensure replicability, and to be able to apply our method to any city in the world, as long as sufficient data exists. The evaluation is carried out using fine-grained census block data for cities in France as ground-truth.

This work is published in [6] and the associated software implementation is made available as open source code at https://github.com/lgervasoni/urbansprawl.

## 6.5. Modelling the relationships between urban morphology, pollutant generation and concentration in the air using PLS path modelling

We have simultaneously modelled the factors that contribute to shaping the urban environment in terms of population density and activities and the level of land use mix on the one hand, and the mechanisms through which this urban morphology is linked to the emission of pollutants and their concentration in the air in the municipalities of the Auvergne-Rhône-Alpes region. To do this, we used the PLS path modelling approach, which is a method of estimating structural equations to model the relationships between latent variables obtained by extracting the information contained in the multidimensional data used to measure them. This work was carried out as part of Diop Samba's internship [8].

## 6.6. Implementation of the World3 model in Python for parametric exploration

The 'World3' model is a digital tool for simulating long-term interactions between population, industrial growth, food production and the boundaries of terrestrial ecosystems. This model was developed in the 1970s. We have ported this model to a modern infrastructure (Python3 + related libraries), in order to be able to apply parameter learning, data analysis and sensitivity study techniques to it. This work was carried out as part of the internship of Aina Rasoldier.

# THOTH Project-Team

# 7. New Results

## 7.1. Visual Recognition in Images and Videos

### *7.1.1. Actor and Observer: Joint Modeling of First and Third-Person Videos*

**Participants:** Gunnar Sigurdsson [CMU], Abhinav Gupta [CMU], Cordelia Schmid, Ali Farhadi [AI2, Univ. Washington], Karteek Alahari.

Several theories in cognitive neuroscience suggest that when people interact with the world, or simulate interactions, they do so from a first-person egocentric perspective, and seamlessly transfer knowledge between third-person (observer) and first-person (actor). Despite this, learning such models for human action recognition has not been achievable due to the lack of data. Our work in [33] takes a step in this direction, with the introduction of Charades-Ego, a large-scale dataset of paired first-person and third-person videos, involving 112 people, with 4000 paired videos. This enables learning the link between the two, actor and observer perspectives. Thereby, we address one of the biggest bottlenecks facing egocentric vision research, providing a link from first-person to the abundant third-person data on the web. We use this data to learn a joint representation of first and third-person videos, with only weak supervision, and show its effectiveness for transferring knowledge from the third-person to the first-person domain.

### *7.1.2. Learning to Segment Moving Objects*

**Participants:** Pavel Tokmakov, Cordelia Schmid, Karteek Alahari.

We study the problem of segmenting moving objects in unconstrained videos [14]. Given a video, the task is to segment all the objects that exhibit independent motion in at least one frame. We formulate this as a learning problem and design our framework with three cues: (i) independent object motion between a pair of frames, which complements object recognition, (ii) object appearance, which helps to correct errors in motion estimation, and (iii) temporal consistency, which imposes additional constraints on the segmentation. The framework is a two-stream neural network with an explicit memory module. The two streams encode appearance and motion cues in a video sequence respectively , while the memory module captures the evolution of objects over time, exploiting the temporal consistency. The motion stream is a convolutional neural network trained on synthetic videos to segment independently moving objects in the optical flow field. The module to build a visual memory in video, i.e., a joint representation of all the video frames, is realized with a convolutional recurrent unit learned from a small number of training video sequences. For every pixel in a frame of a test video, our approach assigns an object or background label based on the learned spatio-temporal features as well as the 'visual memory' specific to the video. We evaluate our method extensively on three benchmarks, DAVIS, Freiburg-Berkeley motion seg-mentation dataset and SegTrack. In addition, we provide an extensive ablation study to investigate both the choice of the training data and the influence of each component in the proposed framework. An overview of our model is shown in Figure 1 .

### *7.1.3. Unsupervised Learning of Artistic Styles with Archetypal Style Analysis*

**Participants:** Daan Wynen, Cordelia Schmid, Julien Mairal.

In [36], we introduce an unsupervised learning approach to automatically discover, summarize, and manipulate artistic styles from large collections of paintings. Our method (summarized in Figure 2 ) is based on archetypal analysis, which is an unsupervised learning technique akin to sparse coding with a geometric interpretation. When applied to neural style representations from a collection of artworks, it learns a dictionary of archetypal styles, which can be easily visualized. After training the model, the style of a new image, which is characterized by local statistics of deep visual features, is approximated by a sparse convex combination of archetypes. This enables us to interpret which archetypal styles are present in the input image, and in which proportion. Finally, our approach allows us to manipulate the coefficients of the latent archetypal decomposition, and achieve various special effects such as style enhancement, transfer, and interpolation between multiple archetypes.

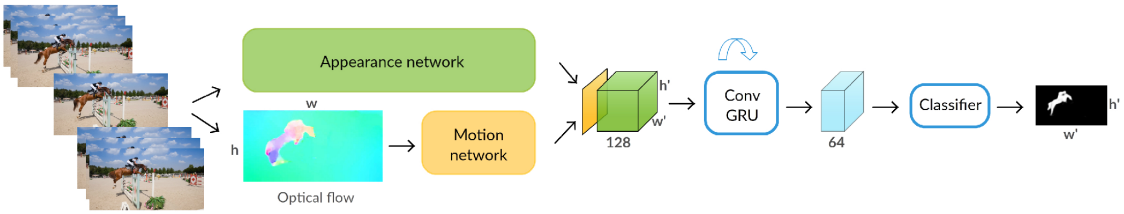



*Figure 1. Overview of our segmentation approach [14]. Each video frame is processed by the appearance (green) and the motion (yellow) networks to produce an intermediate two-stream representation. The ConvGRU module combines this with the learned visual memory to compute the final segmentation result. The width (w') and height (h') of the feature map and the output are $w/8$ and $h/8$ respectively.*

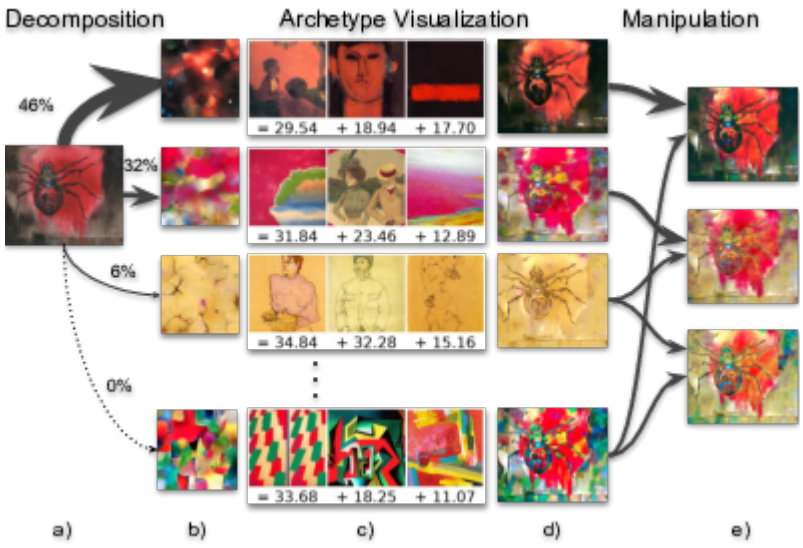



*Figure 2. Using deep archetypal style analysis, we can represent the style of an artwork (a) as a convex combination of archetypes. The archetypes can be visualized as synthesized textures (b), as a convex combination of artworks (c) or, when analyzing a specific artwork, as stylized versions of that artwork itself (d). Free recombination of the archetypal styles then allows for novel stylizations of the input.*

### 7.1.4. Learning from Web Videos for Event Classification

**Participants:** Nicolas Chesneau, Karteek Alahari, Cordelia Schmid.

Traditional approaches for classifying event videos rely on a manually curated training dataset. While this paradigm has achieved excellent results on benchmarks such as TrecVid multimedia event detection (MED) challenge datasets, it is restricted by the effort involved in careful annotation. Recent approaches have attempted to address the need for annotation by automatically extracting images from the web, or generating queries to retrieve videos. In the former case, they fail to exploit additional cues provided by video data, while in the latter, they still require some manual annotation to generate relevant queries. We take an alternate approach in [4], leveraging the synergy between visual video data and the associated textual metadata, to learn event classifiers without manually annotating any videos. Specifically, we first collect a video dataset with queries constructed automatically from textual description of events, prune irrelevant videos with text and video data, and then learn the corresponding event classifiers. We evaluate this approach in the challenging setting where no manually annotated training set is available, i.e., EK0 in the TrecVid challenge, and show state-of-the-art results on MED 2011 and 2013 datasets.

### 7.1.5. How good is my GAN?

**Participants:** Konstantin Shmelkov, Cordelia Schmid, Karteek Alahari.

Generative adversarial networks (GANs) are one of the most popular methods for generating images today. While impressive results have been validated by visual inspection, a number of quantitative criteria have emerged only recently. We argue here that the existing ones are insufficient and need to be in adequation with the task at hand. In [32] introduce two measures based on image classification—GAN-train and GAN-test (illustrated in Figure 3 ), which approximate the recall (diversity) and precision (quality of the image) of GANs respectively. We evaluate a number of recent GAN approaches based on these two measures and demonstrate a clear difference in performance. Furthermore, we observe that the increasing difficulty of the dataset, from CIFAR10 over CIFAR100 to ImageNet, shows an inverse correlation with the quality of the GANs, as clearly evident from our measures.

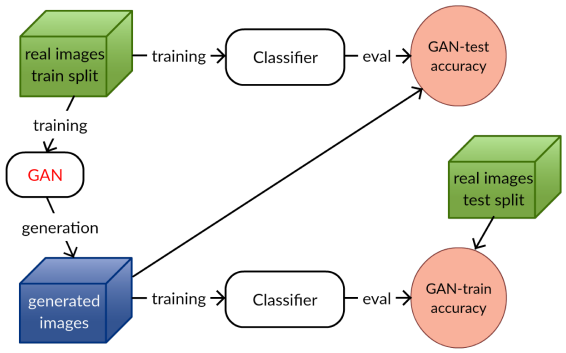



*Figure 3. Illustration of GAN-train and GAN-test. GAN-train learns a classifier on GAN generated images and measures the performance on real test images. This evaluates the diversity and realism of GAN images. GAN-test learns a classifier on real images and evaluates it on GAN images. This measures how realistic GAN images are.*

### 7.1.6. Modeling Visual Context is Key to Augmenting Object Detection Datasets

**Participants:** Nikita Dvornik, Julien Mairal, Cordelia Schmid.

Performing data augmentation for learning deep neural networks is well known to be important for training visual recognition systems. By artificially increasing the number of training examples, it helps reducing overfitting and improves generalization. For object detection, classical approaches for data augmentation consist of generating images obtained by basic geometrical transformations and color changes of original training images. In [23], we go one step further and leverage segmentation annotations to increase the number of object instances present on training data. For this approach to be successful, we show that modeling appropriately the visual context surrounding objects is crucial to place them in the right environment. Otherwise, we show that the previous strategy actually hurts. Clear difference between the two approaches can is presented in Figure 4 . With our context model, we achieve significant mean average precision improvements when few labeled examples are available on the VOC'12 benchmark.

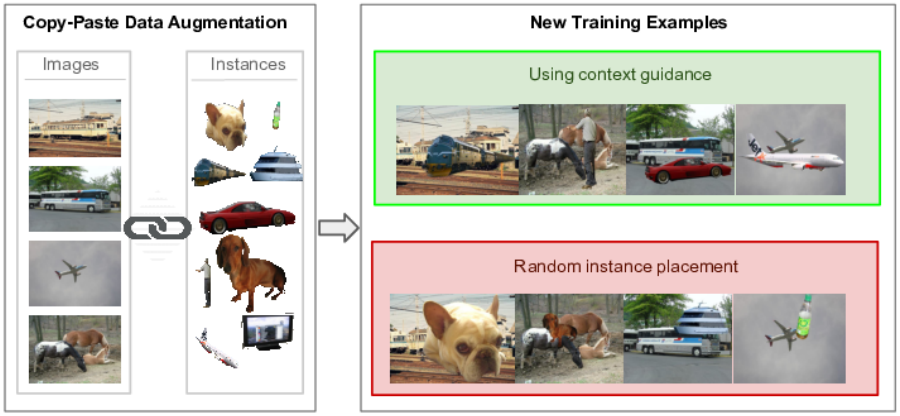



*Figure 4.* ***Examples of data-augmented training examples produced by our approach.*** *Images and objects are taken from the VOC'12 dataset that contains segmentation annotations. We compare the output obtained by pasting the objects with our context model vs. those obtained with random placements. Even though the results are not perfectly photorealistic and display blending artefacts, the visual context surrounding objects is more often correct with the explicit context model.*

### *7.1.7. On the Importance of Visual Context for Data Augmentation in Scene Understanding*

**Participants:** Nikita Dvornik, Julien Mairal, Cordelia Schmid.

Performing data augmentation for learning deep neural networks is known to be important for training visual recognition systems. By artificially increasing the number of training examples, it helps reducing overfitting and improves generalization. While simple image transformations such as changing color intensity or adding random noise can already improve predictive performance in most vision tasks, larger gains can be obtained by leveraging task-specific prior knowledge. In [42], we consider object detection and semantic segmentation and augment the training images by blending objects in existing scenes, using instance segmentation annotations. We observe that randomly pasting objects on images hurts the performance, unless the object is placed in the right context. To resolve this issue, we propose an explicit context model by using a convolutional neural network, which predicts whether an image region is suitable for placing a given object or not. In our experiments, we show that by using copy-paste data augmentation with context guidance we are able to improve detection and segmentation on the PASCAL VOC12 and COCO datasets, with significant gains when few labeled examples are available. The way to augment for different tasks and annotations is presented

in Figure 5 . We also show that the method is not limited to datasets that come with expensive pixel-wise instance annotations and can be used when only bounding box annotations are available, by employing weakly-supervised learning for instance masks approximation.

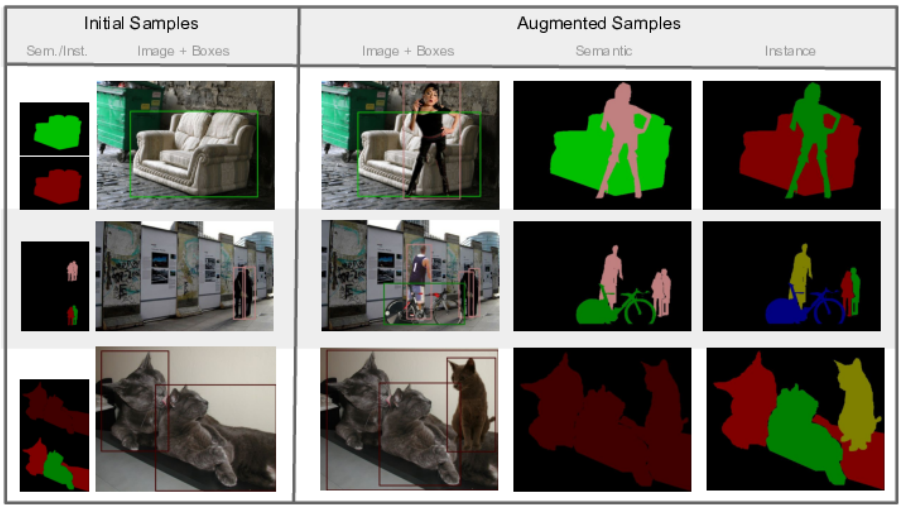



*Figure 5.* ***Data augmentation for different types of annotations.*** *The first column contains samples from the training dataset with corresponding semantic/instance segmentation and bounding box annotations. Columns 2-4 present the result of applying context-driven augmentation to the initial sample with corresponding annotations.*

### *7.1.8. Predicting future instance segmentation by forecasting convolutional features*

**Participants:** Pauline Luc, Camille Couprie [Facebook AI Research], Yann Lecun [Facebook AI Research], Jakob Verbeek.

Anticipating future events is an important prerequisite towards intelligent behavior. Video forecasting has been studied as a proxy task towards this goal. Recent work has shown that to predict semantic segmentation of future frames, forecasting at the semantic level is more effective than forecasting RGB frames and then segmenting these. In [28], we consider the more challenging problem of future instance segmentation, which additionally segments out individual objects. To deal with a varying number of output labels per image, we develop a predictive model in the space of fixed-sized convolutional features of the Mask R-CNN instance segmentation model. We apply the "detection head" of Mask R-CNN on the predicted features to produce the instance segmentation of future frames. Experiments show that this approach significantly improves over strong baselines based on optical flow and repurposed instance segmentation architectures. We show an overview of the proposed method in Figure 6 .

### *7.1.9. Joint Future Semantic and Instance Segmentation Prediction*

**Participants:** Camille Couprie [Facebook AI Research], Pauline Luc, Jakob Verbeek.

The ability to predict what will happen next from observing the past is a key component of intelligence. Methods that forecast future frames were recently introduced towards better machine intelligence. However, predicting directly in the image color space seems an overly complex task, and predicting higher level representations using semantic or instance segmentation approaches were shown to be more accurate. In [20], we introduce a novel prediction approach that encodes instance and semantic segmentation information in a single representation based on distance maps. Our graph-based modeling of the instance segmentation

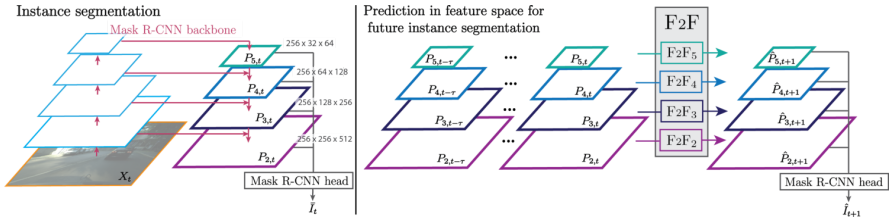



*Figure 6. For future instance segmentation, we extract a pyramid of features from frames $t - \tau$ to $t$, and use them to predict the pyramid features for frame $t + 1$. We learn separate feature-to-feature prediction models for each level of the pyramid. The predicted features are then given as input to a downstream network to produce future instance segmentation.*

prediction problem allows us to obtain temporal tracks of the objects as an optimal solution to a watershed algorithm. Our experimental results on the Cityscapes dataset present state-of-the-art semantic segmentation predictions, and instance segmentation results outperforming a strong baseline based on optical flow. We show an overview of the proposed method in Figure 7 .

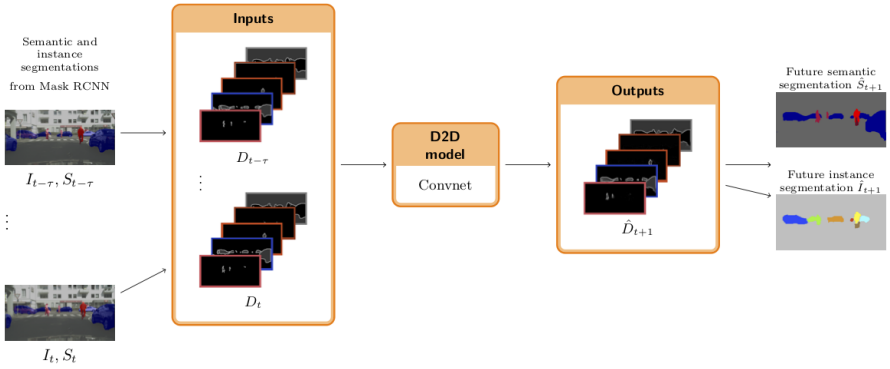



*Figure 7. Our representation enables both future semantic and instance segmentation prediction. It relies on distance maps from the different objects contours: For each channel of an input segmentation, corresponding to a specific class, the segmentation is decomposed into zeros for background, ones for objects and high values for contours. Then a convnet is trained to predict the future representation. Taking its argmax lets us recover the future semantic segmentation, and computing a watershed from it leads to the future instance segmentation.*

### *7.1.10. Depth-based Hand Pose Estimation: Methods, Data, and Challenges*

**Participants:** James S. Supancic [UC Irvine], Grégory Rogez, Yi Yang [Baidu Research], Jamie Shotton [Microsoft Research], Deva Ramanan [Carnegie Mellon University].

Hand pose estimation has matured rapidly in recent years. The introduction of commodity depth sensors and a multitude of practical applications have spurred new advances. In [13], we provide an extensive analysis of the state-of-the-art, focusing on hand pose estimation from a single depth frame. We summarize important conclusions here: (1) Pose estimation appears roughly solved for scenes with isolated hands. However,

methods still struggle to analyze cluttered scenes where hands may be interacting with nearby objects and surfaces. To spur further progress we introduce a challenging new dataset with diverse, cluttered scenes. (2) Many methods evaluate themselves with disparate criteria , making comparisons difficult. We define a consistent evaluation criteria, rigorously motivated by human experiments. (3) We introduce a simple nearest-neighbor baseline that outperforms most existing systems (see results in Fig. 8 ). This implies that most systems do not generalize beyond their training sets. This also reinforces the under-appreciated point that training data is as important as the model itself. We conclude with directions for future progress.

*Figure 8. We evaluate a broad collection of hand pose estimation algorithms on different training and testsets under consistent criteria. Test sets which contained limited variety, in pose and range, or which lacked complex backgrounds were notably easier. To aid our analysis, we introduce a simple 3D exemplar (nearest-neighbor) baseline that both detects and estimates pose suprisingly well, outperforming most existing systems. We show the best-matching detection window in (middle) and the best-matching exemplar in (bottom). We use our baseline to rank dataset difficulty, compare algorithms, and show the importance of training set design.*

### *7.1.11. Image-based Synthesis for Deep 3D Human Pose Estimation*

**Participants:** Grégory Rogez, Cordelia Schmid.

In [11], we address the problem of 3D human pose estimation in the wild. A significant challenge is the lack of training data, i.e., 2D images of humans annotated with 3D poses. Such data is necessary to train state-of-the-art CNN architectures. Here, we propose a solution to generate a large set of photorealistic synthetic images of humans with 3D pose annotations. We introduce an image-based synthesis engine that artificially augments a dataset of real images with 2D human pose annotations using 3D Motion Capture (MoCap) data. Given a candidate 3D pose our algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. See examples in Figure 9 . The resulting images are used to train an end-to-end CNN for full-body 3D pose estimation. We cluster the training data into a large number of pose classes and tackle pose estimation as a K-way classification problem. Such an approach is viable only with large training sets such as ours. Our method outperforms the state of the art in terms of 3D pose estimation in controlled environments (Human3.6M) and shows promising results for in-the-wild images (LSP). This demonstrates that CNNs trained on artificial images generalize well to real images. Compared to data generated from more classical rendering engines, our synthetic images do not require any domain adaptation or fine-tuning stage.

### *7.1.12. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images*

**Participants:** Grégory Rogez, Philippe Weinzaepfel [Naver Labs Europe], Cordelia Schmid.

*Figure 9. Given a candidate 3D pose, our algorithm selects for each joint an image whose annotated 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. We show 6 examples corresponding to the same 3D pose observed from 6 different camera viewpoints.*

In [12], we propose an end-to-end architecture for joint 2D and 3D human pose estimation in natural images. Key to our approach is the generation and scoring of a number of pose proposals per image, which allows us to predict 2D and 3D pose of multiple people simultaneously. See example in Figure 10 . Hence, our approach does not require an approximate localization of the humans for initialization. Our architecture, named LCR-Net, contains 3 main components: 1) the pose proposal generator that suggests potential poses at different locations in the image; 2) a classifier that scores the different pose proposals ; and 3) a regressor that refines pose proposals both in 2D and 3D. All three stages share the convolutional feature layers and are trained jointly. The final pose estimation is obtained by integrating over neighboring pose hypotheses , which is shown to improve over a standard non maximum suppression algorithm. Our approach significantly outperforms the state of the art in 3D pose estimation on Human3.6M, a controlled environment. Moreover, it shows promising results on real images for both single and multi-person subsets of the MPII 2D pose benchmark and demonstrates satisfying 3D pose results even for multi-person images.

### *7.1.13. Link and code: Fast indexing with graphs and compact regression codes*

**Participants:** Matthijs Douze [Facebook AI Research], Alexandre Sablayrolles, Hervé Jégou [Facebook AI Research].

Similarity search approaches based on graph walks have recently attained outstanding speed-accuracy trade-offs, taking aside the memory requirements. In [21], we revisit these approaches by considering, additionally, the memory constraint required to index billions of images on a single server. This leads us to propose a method based both on graph traversal and compact representations. We encode the indexed vectors using quantization and exploit the graph structure to refine the similarity estimation, see Figure 11 . In essence, our method takes the best of these two worlds: the search strategy is based on nested graphs, thereby providing high precision with a relatively small set of comparisons. At the same time it offers a significant memory compression. As a result, our approach outperforms the state of the art on operating points considering 64–128 bytes per vector, as demonstrated by our results on two billion-scale public benchmarks.

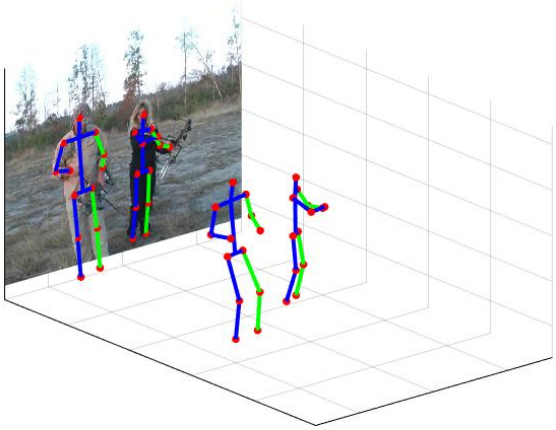


*Figure 10. Examples of joint 2D-3D pose detections in a natural image. Even in case of occlusion or truncation, we estimate the joint locations by reasoning in term of full-body 2D-3D poses.*

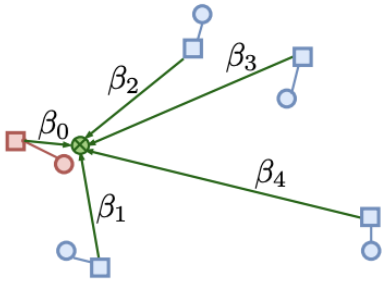



*Figure 11. Illustration of our approach: we adopt a graph traversal strategy that maintains a connectivity between all database points. We further improve the estimate by regressing each database vector from its encoded neighbors.*

### 7.1.14. Sparse weakly supervised models for object localization in road environment

**Participants:** Valentina Zadrija [Univ. Zagreb], Josip Krapac [Univ. Zagreb], Sinisa Segvic [Univ. Zagreb], Jakob Verbeek.

In [16] we propose a novel weakly supervised object localization method based on Fisher-embedding of low-level features (CNN, SIFT), and model sparsity at the component level. Fisher-embedding provides an interesting alternative to raw low-level features, since it allows fast and accurate scoring of image subwindows with a model trained on entire images. Model sparsity reduces overfitting and enables fast evaluation. We also propose two new techniques for improving performance when our method is combined with nonlinear normalizations of the aggregated Fisher representation of the image. These techniques are i) intra-component metric normalization and ii) first-order approximation to the score of a normalized image representation. We evaluate our weakly supervised localization method on real traffic scenes acquired from driver's perspective. The method dramatically improves the localization AP over the dense non-normalized Fisher vector baseline (16 percentage points for zebra crossings, 21 percentage points for traffic signs) and leads to a huge gain in execution speed (91× for zebra crossings, 74× for traffic signs). See Figure 12 for several example outputs.

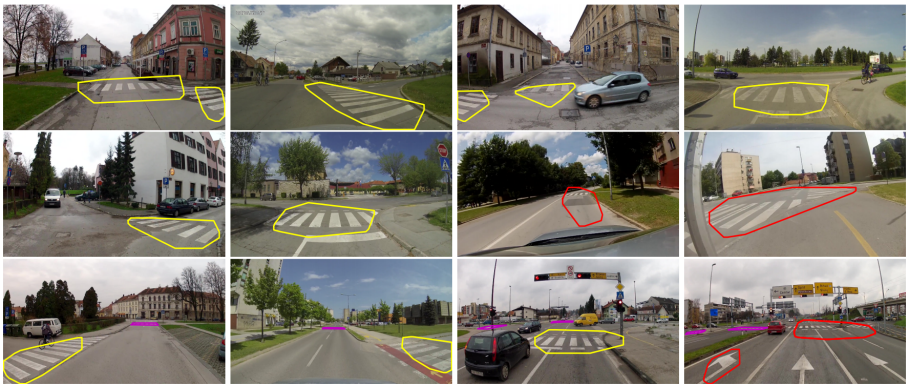


*Figure 12. Localization results on test images: correct localization polygons (yellow), false postive responses (red), and ground-truth polygons for false negatives (magenta).*

### 7.1.15. Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization

**Participants:** Xiaotian Li [Aalto Univ.], Juha Ylioinas [Aalto Univ.], Jakob Verbeek, Juho Kannala [Univ. Oulu].

Image-based camera relocalization is an important problem in computer vision and robotics. Recent works utilize convolutional neural networks (CNNs) to regress for pixels in a query image their corresponding 3D world coordinates in the scene. The final pose is then solved via a RANSAC-based optimization scheme using the predicted coordinates, see Figure 13 . Usually, the CNN is trained with ground truth scene coordinates, but it has also been shown that the network can discover 3D scene geometry automatically by minimizing single-view reprojection loss. However, due to the deficiencies of reprojection loss, the network needs to be carefully initialized. In [27], we present a new angle-based reprojection loss which resolves the issues of the original reprojection loss. With this new loss function, the network can be trained without careful initialization, and the system achieves more accurate results. The new loss also enables us to utilize available multi-view constraints, which further improve performance.

### 7.1.16. FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis

**Participants:** Nitika Verma, Edmond Boyer [Inria, MORPHEO], Jakob Verbeek.

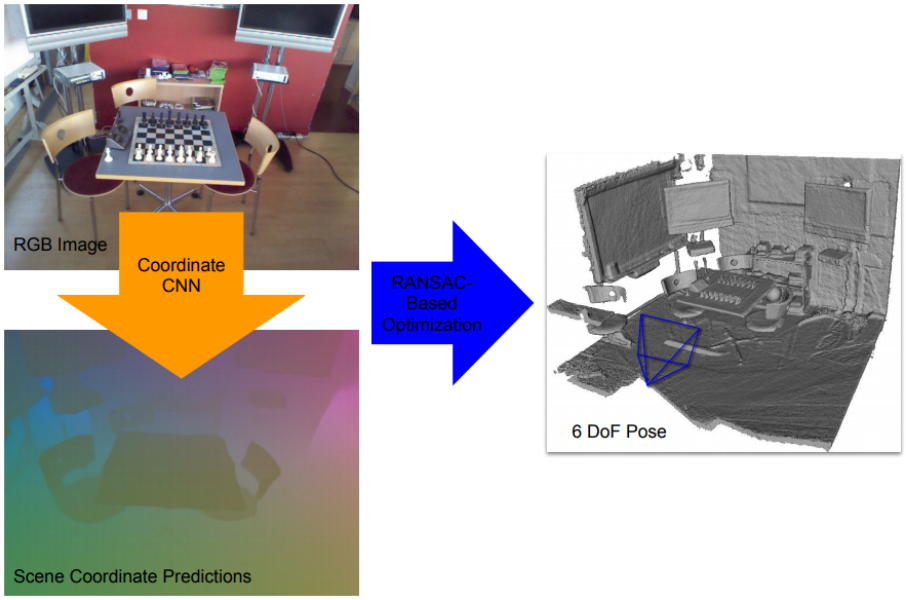



*Figure 13. Localization pipeline. In this two-stage pipeline, a coordinate CNN first produces scene coordinate predictions from an RGB image, and then the predicted correspondences are fed into a RANSAC-based solver to determine the camera pose.*

Convolutional neural networks (CNNs) have massively impacted visual recognition in 2D images, and are now ubiquitous in state-of-the-art approaches. While CNNs naturally extend to other domains, such as audio and video, where data is also organized in rectangular grids, they do not easily generalize to other types of data such as 3D shape meshes, social network graphs or molecular graphs. In our recent paper [35], we propose a novel graph-convolutional network architecture to handle such data. The architecture builds on a generic formulation that relaxes the 1-to-1 correspondence between filter weights and data elements around the center of the convolution, see Figure 14 for an illustration. The main novelty of our architecture is that the shape of the filter is a function of the features in the previous network layer, which is learned as an integral part of the neural network. Experimental evaluations on digit recognition and 3D shape correspondence yield state-of-the-art results, significantly improving over previous work for shape correspondence.

## 7.2. Statistical Machine Learning

### *7.2.1. Modulated Policy Hierarchies*

**Participants:** Alexander Pashevich, Danijar Hafner [Google Brain], James Davidson [Vernalis (R&D) Ltd.], Rahul Sukthankar [Google], Cordelia Schmid.

Solving tasks with sparse rewards is a main challenge in reinforcement learning. While hierarchical controllers are an intuitive approach to this problem, current methods often require manual reward shaping, alternating training phases, or manually defined sub tasks. In [45], we introduce modulated policy hierarchies (MPH), that can learn end-to-end to solve tasks from sparse rewards. To achieve this, we study different modulation signals and exploration for hierarchical controllers. Specifically, we find that communicating via bit-vectors is more efficient than selecting one out of multiple skills, as it enables mixing between them (see Figure 15 ). To facilitate exploration, MPH uses its different time scales for temporally extended intrinsic motivation at each level of the hierarchy. We evaluate MPH on the robotics tasks of pushing and sparse block stacking, where it outperforms recent baselines.

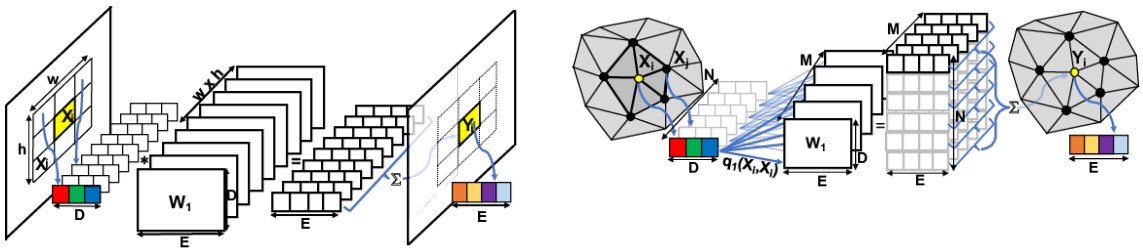



*Figure 14. Left: Illustration of a standard CNN, representing the parameters as a set of $M = w \times h$ weight matrices, each of size $D \times E$. Each weight matrix is associated with a single relative position in the input patch. Right: Our graph convolutional network, where each node in the input patch is associated in a soft manner to each of the $M$ weight matrices based on its features using the weight $q_m(\mathbf{x}_i, \mathbf{x}_j)$.*

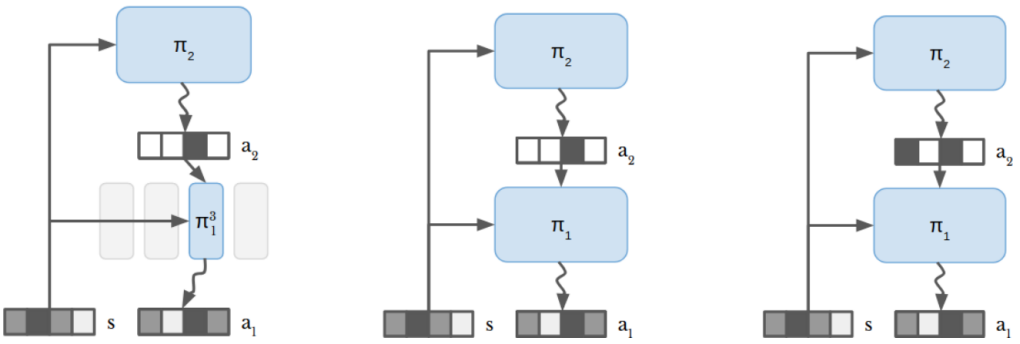



*Figure 15. Overview of hierarchical policies. (a) The options agent selects between separate skill networks using a categorical master policy. (b) The one-hot agent combines the skills into a single network and is modulated by a 1-hot signal. (c) Our modulated policy hierarchy sends a binary vector, allowing for richer communication and mixing of skills.*

### 7.2.2. Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations

**Participants:** Alberto Bietti, Julien Mairal.

The success of deep convolutional architectures is often attributed in part to their ability to learn multiscale and invariant representations of natural signals. However, a precise study of these properties and how they affect learning guarantees is still missing. In [38], we consider deep convolutional representations of signals; we study their invariance to translations and to more general groups of transformations, their stability to the action of diffeomorphisms, and their ability to preserve signal information. This analysis is carried by introducing a multilayer kernel based on convolutional kernel networks and by studying the geometry induced by the kernel mapping. We then characterize the corresponding reproducing kernel Hilbert space (RKHS), showing that it contains a large class of convolutional neural networks with homogeneous activation functions. This analysis allows us to separate data representation from learning, and to provide a canonical measure of model complexity, the RKHS norm, which controls both stability and generalization of any learned model. In addition to models in the constructed RKHS, our stability analysis also applies to convolutional networks with generic activations such as rectified linear units, and we discuss its relationship with recent generalization bounds based on spectral norms.

### 7.2.3. A Contextual Bandit Bake-off

**Participants:** Alberto Bietti, Alekh Agarwal [Microsoft Research], John Langford [Microsoft Research].

Contextual bandit algorithms are essential for solving many real-world interactive machine learning problems. Despite multiple recent successes on statistically and computationally efficient methods, the practical behavior of these algorithms is still poorly understood. In [37], we leverage the availability of large numbers of supervised learning datasets to compare and empirically optimize contextual bandit algorithms, focusing on practical methods that learn by relying on optimization oracles from supervised learning. We find that a recent method using optimism under uncertainty works the best overall. A surprisingly close second is a simple greedy baseline that only explores implicitly through the diversity of contexts, followed by a variant of Online Cover which tends to be more conservative but robust to problem specification by design. Along the way, we also evaluate and improve several internal components of contextual bandit algorithm design. Overall, this is a thorough study and review of contextual bandit methodology.

### 7.2.4. Learning Disentangled Representations with Reference-Based Variational Autoencoders

**Participants:** Adria Ruiz, Oriol Martinez [Universitat Pompeu Fabra, Barcelona], Xavier Binefa [Universitat Pompeu Fabra, Barcelona], Jakob Verbeek.

Learning disentangled representations from visual data, where different high-level generative factors are independently encoded, is of importance for many computer vision tasks. Supervised approaches, however, require a significant annotation effort in order to label the factors of interest in a training set. To alleviate the annotation cost, in [47] we introduce a learning setting which we refer to as "reference-based disentangling". Given a pool of unlabelled images, the goal is to learn a representation where a set of target factors are disentangled from others. The only supervision comes from an auxiliary "reference set" that contains images where the factors of interest are constant. See Fig. 16 for illustrative examples. In order to address this problem, we propose reference-based variational autoencoders, a novel deep generative model designed to exploit the weak supervisory signal provided by the reference set. During training, we use the variational inference framework where adversarial learning is used to minimize the objective function. By addressing tasks such as feature learning, conditional image generation or attribute transfer, we validate the ability of the proposed model to learn disentangled representations from minimal supervision.

### 7.2.5. On Regularization and Robustness of Deep Neural Networks

**Participants:** Alberto Bietti, Grégoire Mialon, Julien Mairal.

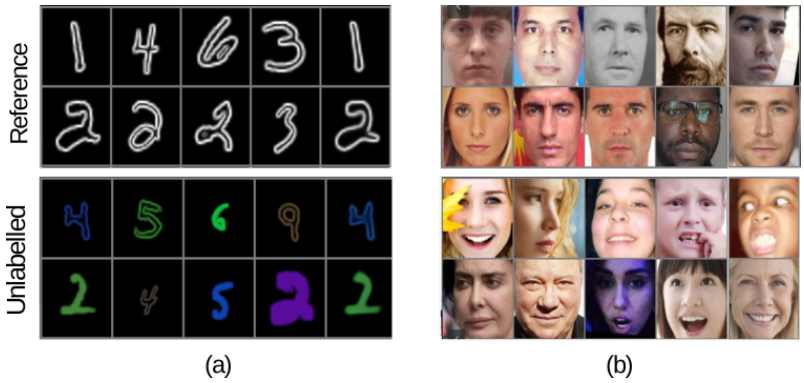



*Figure 16. Illustration of different reference-based disentangling problems. (a) Disentangling style from digits. The reference distribution is composed by numbers with a fixed style (b) Disentangling factors of variations related with facial expressions. Reference images correspond to neutral faces. Note that pairing information between unlabelled and reference images is not available during training.*

For many supervised learning tasks, deep neural networks are known to work well when large amounts of annotated data are available. Yet, Despite their success, deep neural networks suffer from several drawbacks: they lack robustness to small changes of input data known as "adversarial examples" and training them with small amounts of annotated data is challenging. In [39], we study the connection between regularization and robustness of deep neural networks by viewing them as elements of a reproducing kernel Hilbert space (RKHS) of functions and by regularizing them using the RKHS norm. Even though this norm cannot be computed, we consider various approximations based on upper and lower bounds. These approximations lead to new strategies for regularization, but also to existing ones such as spectral norm penalties or constraints, gradient penalties, or adversarial training. Besides, the kernel framework allows us to obtain margin-based bounds on adversarial generalization. We show that our new algorithms lead to empirical benefits for learning on small datasets and learning adversarially robust models. We also discuss implications of our regularization framework for learning implicit generative models.

### *7.2.6. Mixed batches and symmetric discriminators for GAN training*

**Participants:** Thomas Lucas, Corentin Tallec [Inria, TAU], Jakob Verbeek, Yann Ollivier [Facebook AI Research].

Generative adversarial networks (GANs) are powerful generative models based on providing feedback to a generative network via a discriminator network. However, the discriminator usually assesses individual samples. This prevents the discriminator from accessing global distributional statistics of generated samples, and often leads to *mode dropping*: the generator models only part of the target distribution. In [29] we propose to feed the discriminator with *mixed batches* of true and fake samples, and train it to predict the ratio of true samples in the batch. The latter score does not depend on the order of samples in a batch. Rather than learning this invariance, we introduce a generic permutation-invariant discriminator architecture, which is illustrated in Figure 17 . This architecture is provably a universal approximator of all symmetric functions. Experimentally, our approach reduces mode collapse in GANs on two synthetic datasets, and obtains good results on the CIFAR10 and CelebA datasets, both qualitatively and quantitatively.

### *7.2.7. Auxiliary Guided Autoregressive Variational Autoencoders*

**Participants:** Thomas Lucas, Jakob Verbeek.

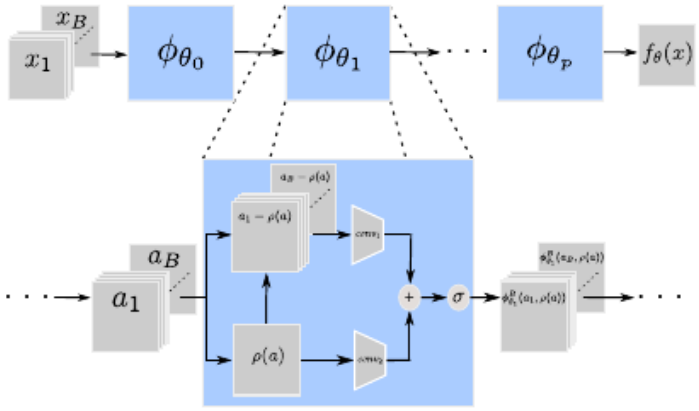



*Figure 17. Graphical representation of our discriminator architecture. Each convolutional layer of an otherwise classical CNN architecture is modified to include permutation invariant batch statistics, denoted $\rho(x)$. This is repeated at every layer so that the network gradually builds up more complex statistics.*

Generative modeling of high-dimensional data is a key problem in machine learning. Successful approaches include latent variable models and autoregressive models. The complementary strengths of these approaches, to model global and local image statistics respectively, suggest hybrid models combining the strengths of both. Our contribution in [30] is to train such hybrid models using an auxiliary loss function that controls which information is captured by the latent variables and what is left to the autoregressive decoder, as illustrated in Figure 18 . In contrast, prior work on such hybrid models needed to limit the capacity of the autoregressive decoder to prevent degenerate models that ignore the latent variables and only rely on autoregressive modeling. Our approach results in models with meaningful latent variable representations, and which rely on powerful autoregressive decoders to model image details. Our model generates qualitatively convincing samples, and yields state-of-the-art quantitative results.

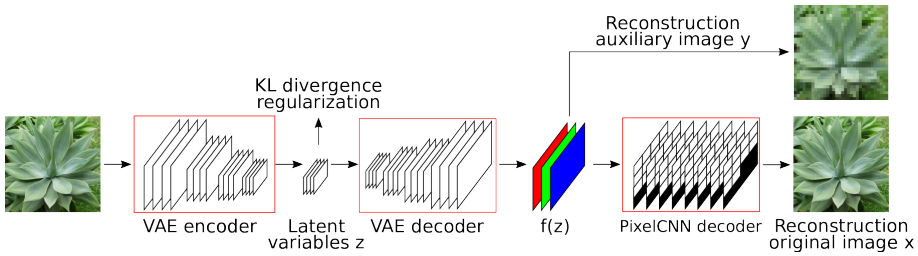



*Figure 18. Schematic illustration of our auxiliary guided autoregressive variational autoencoder (AGAVE). An input image is encoded into a latent representation and decoded back into an image. This first reconstruction is guided by an auxiliary maximum likelihood loss and regularized with a Kullback-Liebler divergence. An autoregressive model is then conditionned on the auxiliary reconstruction and also trained with maximum likelihood.*

### 7.2.8. End-to-End Incremental Learning

**Participants:** Francisco Castro [Univ. Malaga], Manuel Marin-Jimenez [Univ. Cordoba], Nicolas Guil [Univ. Malaga], Cordelia Schmid, Karteek Alahari.

Although deep learning approaches have stood out in recent years due to their state-of-the-art results, they continue to suffer from catastrophic forgetting, a dramatic decrease in overall performance when training with new classes added incrementally. This is due to current neural network architectures requiring the entire dataset, consisting of all the samples from the old as well as the new classes, to update the model—a requirement that becomes easily unsustainable as the number of classes grows. We address this issue with our approach [17] to learn deep neural networks incrementally, using new data and only a small exemplar set corresponding to samples from the old classes. This is based on a loss composed of a distillation measure to retain the knowledge acquired from the old classes, and a cross-entropy loss to learn the new classes. Our incremental training is achieved while keeping the entire framework end-to-end, i.e., learning the data representation and the classifier jointly, unlike recent methods with no such guarantees. We evaluate our method extensively on the CIFAR-100 and ImageNet (ILSVRC 2012) image classification datasets, and show state-of-the-art performance.

## 7.3. Large-scale Optimization for Machine Learning

### 7.3.1. Stochastic Subsampling for Factorizing Huge Matrices

**Participants:** Julien Mairal, Arthur Mensch [Inria, Parietal], Gael Varoquaux [Inria, Parietal], Bertrand Thirion [Inria, Parietal].

In [10], we present a matrix-factorization algorithm that scales to input matrices with both huge number of rows and columns. Learned factors may be sparse or dense and/or non-negative, which makes our algorithm suitable for dictionary learning, sparse component analysis, and non-negative matrix factorization. Our algorithm streams matrix columns while subsampling them to iteratively learn the matrix factors. At each iteration, the row dimension of a new sample is reduced by subsampling, resulting in lower time complexity compared to a simple streaming algorithm. Our method comes with convergence guarantees to reach a stationary point of the matrix-factorization problem. We demonstrate its efficiency on massive functional Magnetic Resonance Imaging data (2 TB), and on patches extracted from hyperspectral images (103 GB). For both problems, which involve different penalties on rows and columns, we obtain significant speed-ups compared to state-of-the-art algorithms. The main principle of the method is illustrated in Figure 19 .

### 7.3.2. An Inexact Variable Metric Proximal Point Algorithm for Generic Quasi-Newton Acceleration

**Participants:** Hongzhou Lin, Julien Mairal, Zaid Harchaoui [Univ. Washington].

In [43], we propose a generic approach to accelerate gradient-based optimization algorithms with quasi-Newton principles. The proposed scheme, called QuickeNing, can be applied to incremental first-order methods such as stochastic variance-reduced gradient (SVRG) or incremental surrogate optimization (MISO). It is also compatible with composite objectives, meaning that it has the ability to provide exactly sparse solutions when the objective involves a sparsity-inducing regularization. QuickeNing relies on limited-memory BFGS rules, making it appropriate for solving high-dimensional optimization problems. Besides, it enjoys a worst-case linear convergence rate for strongly convex problems. We present experimental results where QuickeNing gives significant improvements over competing methods for solving large-scale high-dimensional machine learning problems, see Figure 20 for example.

### 7.3.3. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice

**Participants:** Hongzhou Lin, Julien Mairal, Zaid Harchaoui [Univ. Washington].

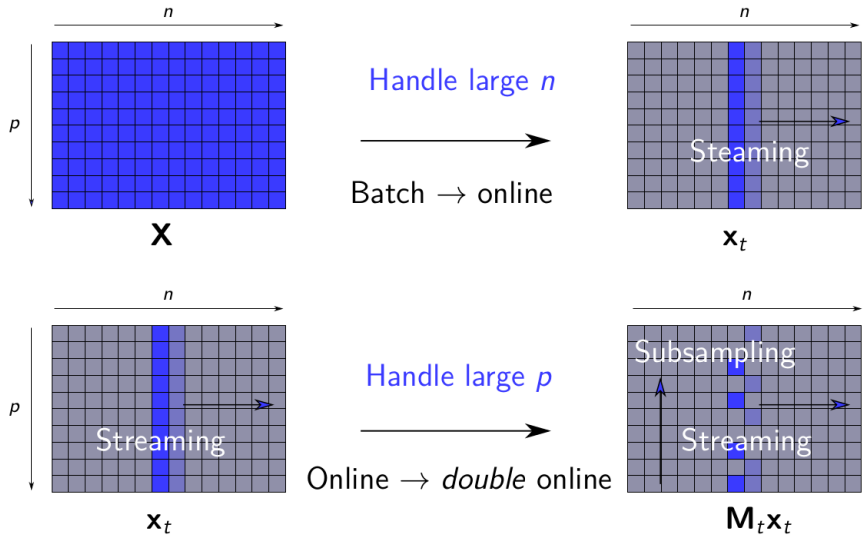



*Figure 19. Illustration of the matrix factorization algorithm, which streams columns in one dimension while subsampling them.*

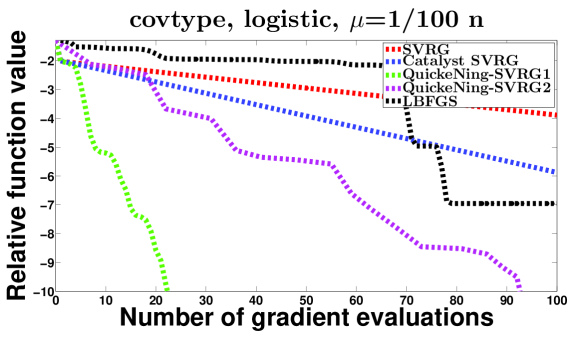



*Figure 20. An illustration of the minimization of logistic regression. Significant improvement is observed by applying QuickeNing.*

In [9], we introduce a generic scheme for accelerating gradient-based optimization methods in the sense of Nesterov. The approach, called Catalyst, builds upon the inexact accelerated proximal point algorithm for minimizing a convex objective function, and consists of approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. One of the key to achieve acceleration in theory and in practice is to solve these sub-problems with appropriate accuracy by using the right stopping criterion and the right warm-start strategy. In this work, we give practical guidelines to use Catalyst and present a comprehensive theoretical analysis of its global complexity. We show that Catalyst applies to a large class of algorithms, including gradient descent, block coordinate descent, incremental algorithms such as SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. We conclude with extensive experiments showing that acceleration is useful in practice, especially for ill-conditioned problems.

### *7.3.4. Catalyst Acceleration for Gradient-Based Non-Convex Optimization*

**Participants:** Courtney Paquette [Univ. Washington], Hongzhou Lin, Dmitriy Drusvyatskiy [Univ. Washington], Julien Mairal, Zaid Harchaoui [Univ. Washington].

In [31], we introduce a generic scheme to solve nonconvex optimization problems using gradient-based algorithms originally designed for minimizing convex functions. When the objective is convex, the proposed approach enjoys the same properties as the Catalyst approach of Lin et al, 2015. When the objective is nonconvex, it achieves the best known convergence rate to stationary points for first-order methods. Specifically, the proposed algorithm does not require knowledge about the convexity of the objective; yet, it obtains an overall worst-case efficiency of $O(\epsilon^{-2})$ and, if the function is convex, the complexity reduces to the near-optimal rate $O(\epsilon^{-2/3})$. We conclude the paper by showing promising experimental results obtained by applying the proposed approach to SVRG and SAGA for sparse matrix factorization and for learning neural networks (see Figure 21 ).

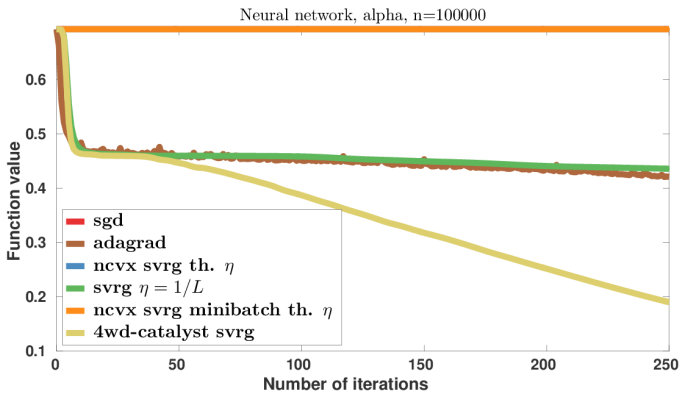



*Figure 21. Comparison of different algorithms for the minimization of a two-layer neural network. Applying our method provides a clear acceleration in terms of function value.*

## 7.4. Pluri-disciplinary Research

### *7.4.1. Biological Sequence Modeling with Convolutional Kernel Networks*

**Participants:** Dexiong Chen, Laurent Jacob [CNRS, LBBE Laboratory], Julien Mairal.

The growing number of annotated biological sequences available makes it possible to learn genotype-phenotype relationships from data with increasingly high accuracy. When large quan- tities of labeled samples are available for training a model, convolutional neural networks can be used to predict the phenotype of unannotated sequences with good accuracy. Unfortunately, their performance with medium- or small-scale datasets is mitigated, which requires inventing new data-efficient approaches. In [40], we introduce a hybrid approach between convolutional neural networks and kernel methods to model biological sequences. Our method 22 enjoys the ability of convolutional neural networks to learn data representations that are adapted to a specific task, while the kernel point of view yields algorithms that perform significantly better when the amount of training data is small. We illustrate these advantages for transcription factor binding prediction and protein homology detection, and we demonstrate that our model is also simple to interpret, which is crucial for discovering predictive motifs in sequences. The source code is freely available at https://gitlab.inria.fr/dchen/CKN-seq.

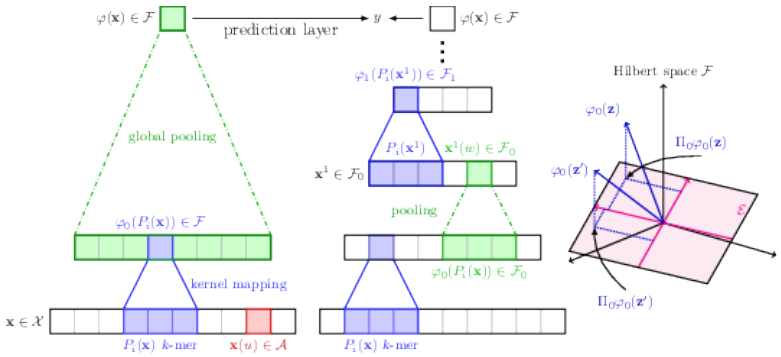



*Figure 22. Construction of single-layer (left) and multilayer (middle) CKN-seq and the approximation of one layer (right). For a single-layer model, each $k$-mer $P_i(\mathbf{x})$ is mapped to $\varphi_0(P_i(\mathbf{x}))$ in $\mathcal{F}$ and projected to $\Pi\varphi_0(P_i(\mathbf{x}))$ parametrized by $\psi_0(P_i(\mathbf{x}))$. Then, the final finite-dimensional sequence is obtained by the global pooling, $\psi(\mathbf{x}) = \frac{1}{m}\sum_{i=0}^{m}\psi_0(P_i(\mathbf{x}))$. The multilayer construction is similar, but relies on intermediate maps, obtained by local pooling.*

### *7.4.2. Token-level and sequence-level loss smoothing for RNN language models*

**Participants:** Maha Elbayad, Laurent Besacier [LIG], Jakob Verbeek.

In [25] we investigate the limitations of the maximum likelihood estimation (MLE) used when training recurrent neural network language models. First, the MLE treats all sentences that do not match the ground truth as equally poor, ignoring the structure of the output space. Second, it suffers from "exposure bias": during training tokens are predicted given ground-truth sequences, while at test time prediction is conditioned on generated output sequences. To overcome these limitations we build upon the recent reward augmented maximum likelihood approach i.e., sequence-level smoothing that encourages the model to predict sentences close to the ground truth according to a given performance metric. We extend this approach to token-level loss smoothing, and propose improvements to the sequence-level smoothing approach. Our experiments on two different tasks, image captioning (see Fig. 23 ) and machine translation, show that token-level and sequence-level loss smoothing are complementary, and significantly improve results.

### *7.4.3. Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction*

**Participants:** Maha Elbayad, Laurent Besacier [LIG], Jakob Verbeek.

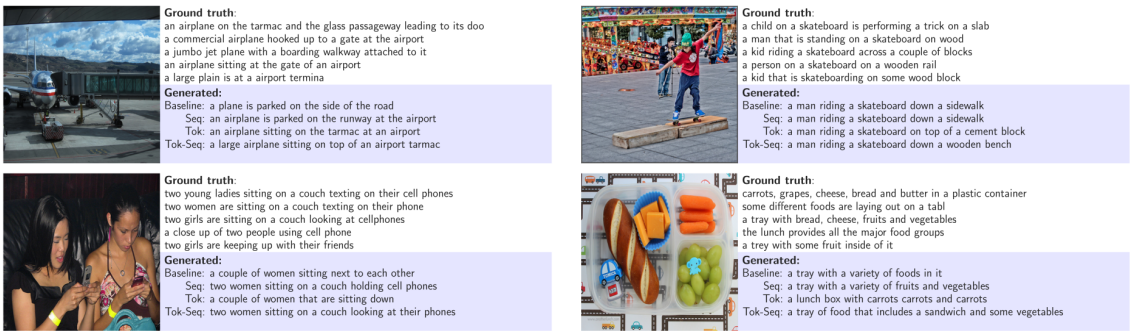



*Figure 23. Examples of generated captions with the baseline MLE and our models with attention.*

Current state-of-the-art machine translation systems are based on encoder-decoder architectures, that first encode the input sequence, and then generate an output sequence based on the input encoding. Both are interfaced with an attention mechanism that recombines a fixed encoding of the source tokens based on the decoder state. In [24], we propose an alternative approach which instead relies on a single 2D convolutional neural network across both sequences as illustrated in Figure 24 . Each layer of our network re-codes source tokens on the basis of the output sequence produced so far. Attention-like properties are therefore pervasive throughout the network. Our model yields excellent results, outperforming state-of-the-art encoder-decoder systems, while being conceptually simpler and having fewer parameters.

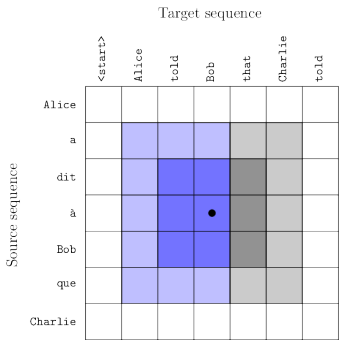



*Figure 24. Convolutional layers in our model use masked 3×3 filters so that features are only computed from previous output symbols. Illustration of the receptive fields after one (dark blue) and two layers (light blue), together with the masked part of the field of view of a normal 3×3 filter (gray)*

### *7.4.4. Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis*

**Participant:** Ghislain Durif.

The development of high-throughput biology technologies now allows the investigation of the genome-wide diversity of transcription in single cells. This diversity has shown two faces: the expression dynamics (gene to gene variability) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. Second, the cell-to-cell variability is high, with a low proportion of cells expressing the same gene at the same

time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent and to provide a summarized view of single-cell expression data. PCA is one of the most powerful framework to provide a suitable representation of high dimensional datasets, by searching for latent directions catching the most variability in the data. Unfortunately, classical PCA is based on Euclidean distances and projections that work poorly in presence of over-dispersed counts that show drop-out events (zero-inflation) like single-cell expression data. In [22], we propose a probabilistic Count Matrix Factorization (pCMF) approach for single-cell expression data analysis, that relies on a sparse Gamma-Poisson factor model. This hierarchical model is inferred using a variational EM algorithm. We show how this probabilistic framework induces a geometry that is suitable for single-cell data visualization, and produces a compression of the data that is very powerful for clustering purposes. Our method is competed to other standard representation methods like t-SNE, and we illustrate its performance for the representation of zero-inflated over-dispersed count data. We also illustrate our work with results on a publicly available data set, being single-cell expression profile of neural stem cells. Our work is implemented in the pCMF R-package.

### *7.4.5. Extracting Universal Representations of Cognition across Brain-Imaging Studies*

**Participants:** Arthur Mensch [Inria, Parietal], Julien Mairal, Bertrand Thirion [Inria, Parietal], Gael Varoquaux [Inria, Parietal].

We show in [44] how to extract shared brain representations that predict mental processes across many cognitive neuroimaging studies. Focused cognitive-neuroimaging experiments study precise mental processes with carefully-designed cognitive paradigms; however the cost of imaging limits their statistical power. On the other hand, large-scale databasing efforts increase considerably the sample sizes, but cannot ask precise cognitive questions. To address this tension, we develop new methods that turn the heterogeneous cognitive information held in different task-fMRI studies into common-universal-cognitive models. Our approach does not assume any prior knowledge of the commonalities shared by the studies in the corpus; those are inferred during model training. The method uses deep-learning techniques to extract representations - task-optimized networks - that form a set of basis cognitive dimensions relevant to the psychological manipulations, as illustrated in Figure 25 . In this sense, it forms a novel kind of functional atlas, optimized to capture mental state across many functional-imaging experiments. As it bridges information on the neural support of mental processes, this representation improves decoding performance for 80% of the 35 widely-different functional imaging studies that we consider. Our approach opens new ways of extracting information from brain maps, increasing statistical power even for focused cognitive neuroimaging studies, in particular for those with few subjects.

### *7.4.6. Loter: Inferring local ancestry for a wide range of species*

**Participants:** Thomas Dias-Alves, Julien Mairal, Michael Blum [CNRS, TIMC Laboratory].

Admixture between populations provides opportunity to study biological adaptation and phenotypic variation. Admixture studies can rely on local ancestry inference for admixed individuals, which consists of computing at each locus the number of copies that originate from ancestral source populations, as illustrated in Figure 26 . Existing software packages for local ancestry inference are tuned to provide accurate results on human data and recent admixture events. In [5], we introduce Loter, an open-source software package that does not require any biological parameter besides haplotype data in order to make local ancestry inference available for a wide range of species. Using simulations, we compare the performance of Loter to HAPMIX, LAMP-LD, and RFMix. HAPMIX is the only software severely impacted by imperfect haplotype reconstruction. Loter is the less impacted software by increasing admixture time when considering simulated and admixed human genotypes. LAMP-LD and RFMIX are the most accurate method when admixture took place 20 generations ago or less; Loter accuracy is comparable or better than RFMix accuracy when admixture took place of 50 or more generations; and its accuracy is the largest when admixture is more ancient than 150 generations. For simulations of admixed Populus genotypes, Loter and LAMP-LD are robust to increasing admixture times by contrast to RFMix. When comparing length of reconstructed and true ancestry tracts, Loter and LAMP-LD provide results whose accuracy is again more robust than RFMix to increasing admixture times. We apply Loter to admixed Populus individuals and lengths of ancestry tracts indicate that admixture took place around

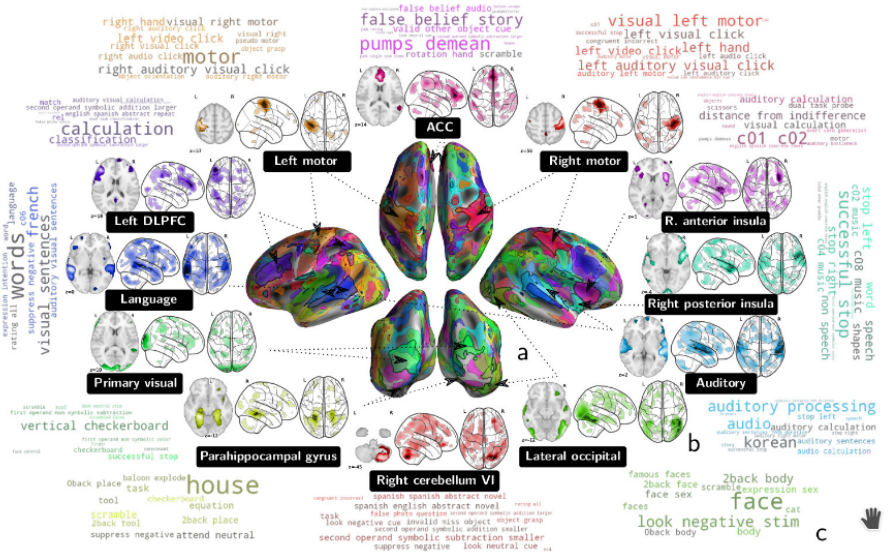



*Figure 25. Visualization of some of task-optimized networks. Our approach allows to learn networks that are important for inter-subject decoding across studies. These networks, individually focal and collectively well spread across the cortex, are readily associated with the cognitive tasks that they contribute to predict. We display a selection of these networks, named with the salient anatomical brain region they recruit, along with a word-cloud representation of the stimuli whose likelihood increases with the network activation.*

100 generations ago. The Loter software package and its source code are available at https://github.com/bcm-uga/Loter.

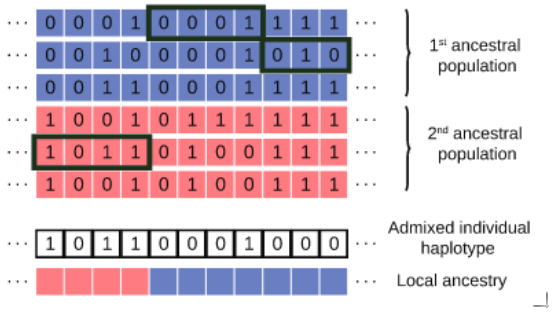



*Figure 26. Graphical description of Local Ancestry Inference as implemented in the software Loter. Given a collection of parental haplotypes from the source populations depicted in blue and red, Loter assumes that an haplotype of an admixed individuals is modeled as a mosaic of existing parental haplotypes.*

**TRIPOP Team**

# 6. New Results

## 6.1. Nonlinear waves in granular chains

**Participants:** Guillaume James, Bernard Brogliato, Kirill Vorotnikov.

Granular chains made of aligned beads interacting by contact (e.g. Newton's cradle) are widely studied in the context of impact dynamics and acoustic metamaterials. In order to describe the response of such systems to impacts or vibrations, it is important to analyze different wave effects such as the propagation of localized compression pulses (solitary waves) or oscillations (traveling breathers), or the scattering of vibrations through the chain. Such phenomena are strongly influenced by contact nonlinearities (Hertz force), spatial inhomogeneities and dissipation.

In the work [22], we analyze the Kuwabara-Kono (KK) model for contact damping, and we develop new approximations of this model which are efficient for the simulation of multiple impacts. The KK model is a simplified viscoelastic contact model derived from continuum mechanics, which allows for simpler calibration (using material parameters instead of phenomenological ones), but its numerical simulation requires a careful treatment due to its non-Lipschitzian character. Using different dissipative time-discretizations of the conservative Hertz model, we show that numerical dissipation can be tuned properly in order to reproduce the physical dissipation of the KK model and associated wave effects. This result is obtained analytically in the limit of small time steps (using methods from backward analysis) and is numerically validated for larger time steps. The resulting schemes turn out to provide good approximations of impact propagation even for relatively large time steps.

In reference [8], we analyze the discrete $p$-Schrödinger equation, an envelope equation that describes small oscillations in a Newton's cradle. In the limit when the exponent of the contact force lies slightly above unity, we derive three different continuum limits of the model which allow us to approximate the profiles of traveling breather solutions. One model consists of a logarithmic nonlinear Schrödinger equation which leads to a Gaussian approximation, and the two other are fully nonlinear degenerate Schrödinger equations which provide compacton approximations. These approximations are numerically validated by Newton-type computations. In the opposite (vibroimpact) limit when the exponent of the contact force is large, we obtain an analytical approximation of solitary waves in the form of a compacton.

## 6.2. Periodic motions of coupled impact oscillators

**Participants:** Guillaume James, Vincent Acary, Franck Pérignon.

In the work [17], we study the existence and stability of time-periodic oscillations in an infinite chain of linearly coupled impact oscillators, for rigid impacts without energy dissipation. We reformulate the search of periodic solutions as a boundary value problem incorporating unilateral constraints. This formulation, together with an appropriate notion of nondegenerate modes, allows us to construct nonsmooth modes of oscillations (spatially localized or extended) when the oscillators are weakly coupled (this approach is an adaptation of the idea of 'anticontinuum' limit to the nonsmooth setting). In this framework, we show the existence of exact solutions (in particular, we check the condition of non-penetration of the obstacle) for an arbitrary number of impacting particles. Different solution branches corresponding to stable or unstable breathers, multibreathers and nonsmooth normal modes are found. We provide a formula for the monodromy matrix that determines spectral stability of nonsmooth modes in the presence of simple impacts. These results are completed by a numerical computation of the time-periodic solutions at larger coupling, and the Siconos software is used to simulate the system and explore dynamical instabilities. The above approach is much more effective than numerical continuation of periodic solutions based on stiff compliant models, which leads to stiff ODEs and costly numerical continuation.

## 6.3. Solitary waves in the excitable Burridge-Knopoff model

**Participants:** Guillaume James, Jose Eduardo Morales Morales, Arnaud Tonnelier.

The Burridge-Knopoff model is a lattice differential equation describing a chain of blocks connected by springs and pulled over a surface. This model was originally introduced to investigate nonlinear effects arising in the dynamics of earthquake faults. One of the main ingredients of the model is a nonlinear velocity-dependent friction force between the blocks and the fixed surface. We introduce a simplified piecewise linear friction law (reminiscent of the McKean nonlinearity for excitable cells) which allows us to obtain analytical expression of solitary waves and study some of their qualitative properties, such as wavespeed and propagation failure. These results have been published in [11].
We have obtained an existence theorem for solitary waves in the Burridge-Knopoff model. Our approach uses a piecewise-linear friction force combined with a weak coupling strength. Using asymptotic arguments, we show that trial solutions, obtained semi-analytically, satisfy, for some parameter set, the inequality constraints associated with the threshold conditions. An approximation of the wave profile is obtained and a minimal wave speed is derived.

## 6.4. Signal propagation along excitable chains

**Participant:** Arnaud Tonnelier.

Nonlinear self-sustained waves, or *autowaves*, have been identified in a large class of discrete excitable media. We have proposed a simple continuous-time threshold model for wave propagation in excitable media. The ability of the resulting transmission line to convey a one-bit signal is investigated. Existence and multistability of signals where two successive units share the same waveform is established. We show that, depending on the connectivity of the transmission line, an arbitrary number of distinct signals can be transmitted. More precisely, we prove that, for a one-dimensional information channel with $n$th-neighbor interactions, a $n$-fold degeneracy of the speed curve induces the coexistence of $2n$ propagating signals, $n$ of which are stable and allow $n$ distinct symbols transmission. The influence of model parameters (time constants, coupling strength and connectivity) on the traveling signal properties is analyzed. This work is almost finished and is going to be submitted.

## 6.5. Numerical analysis of multibody mechanical systems with constraints

This scientific theme concerns the numerical analysis of mechanical systems with bilateral and unilateral constraints, with or without friction [1]. They form a particular class of dynamical systems whose simulation requires the development of specific methods for analysis and dedicated simulators [6].

### *6.5.1. Multibody systems with clearances (dynamic backlash)*

**Participants:** Vincent Acary, Bernard Brogliato.

The PhD thesis of N. Akadkhar under contract with Schneider Electric has concerned the numerical simulation of mechanical systems with unilateral constraints and friction, where the presence of clearances in imperfect joints plays a crucial role. A first work deals with four-bar planar mechanisms with clearances at the joints, which induce unilateral constraints and impacts, rendering the dynamics nonsmooth. The objective is to determine sets of parameters (clearance value, restitution coefficients, friction coefficients) such that the system's trajectories stay in a neighborhood of the ideal mechanism (*i.e.* without clearance) trajectories. The analysis is based on numerical simulations obtained with the projected Moreau-Jean time-stepping scheme. Circuits breakers with 3D joint clearances have been studied in [3] [41] where it is demonstrated that the nonsmooth dynamics approach as coded in our software SICONOS, allows a very good prediction of the system's dynamics, with experimental validation. An overview of various approaches for the feedback control of multibody systems with joint clearances is proposed in [4].

### *6.5.2. Generalized-$\alpha$ scheme for nonsmooth multibody systems.*

**Participant:** Vincent Acary.

This work [16] concerns a formalism for the transient simulation of nonsmooth dynamic mechanical systems composed of rigid and flexible bodies, kinematic joints and frictionless contact conditions. The proposed algorithm guarantees the exact satisfaction of the bilateral and unilateral constraints both at position and velocity levels. Thus, it significantly differs from penalty techniques since no penetration is allowed. The numerical scheme is obtained in two main steps. Firstly, a splitting method is used to isolate the contributions of impacts, which shall be integrated with only first-order accuracy, from smooth contributions which can be integrated using a higher order scheme. Secondly, following the idea of Gear, Gupta and Leimkuhler, the equation of motion is reformulated so that the bilateral and unilateral constraints appear both at position and velocity levels. After time discretization, the equation of motion involves two complementarity conditions and it can be solved at each time step using a monolithic semi-smooth Newton method. The numerical behaviour of the proposed method is studied and compared to other approaches for a number of numerical examples. Itis shown that the formulation offers a unified and valid approach for the description of contactconditions between rigid bodies as well as between flexible bodies.

### *6.5.3. Mechanics of musical instruments with contact and impacts.*

**Participants:** Vincent Acary, Franck Pérignon.

Collisions in musical string instruments play a fundamental role in explaining the sound production in various instruments such as sitars, tanpuras and electric basses. Contacts occuring during the vibration provide a nonlinear effect which shapes a specific tone due to energy transfers and enriches the hearing experience. As such, they must be carefully simulated for the purpose of physically-based sound synthesis. Most of the numerical methods presented in the literature rely on a compliant modeling of the contact force between the string and the obstacle. In this contribution, numerical methods from nonsmooth contact dynamics are used to integrate the problem in time. A Moreau-Jean time-stepping scheme is combined with an exact scheme for phases with no contact, thus controlling the numerical dispersion. Results for a two-point bridge mimicking a tanpura and an electric bass are presented, showing the ability of the method to deal efficiently with such problems while invoking, as compared to a compliant approach, less modelling parameters and a reduced computational burden [7].

### *6.5.4. Numerical solvers for frictional contact problems.*

**Participants:** Vincent Acary, Maurice Brémond.

In [15] report, we review several formulations of the discrete frictional contact problem that arises in space and time discretized mechanical systems with unilateral contact and three-dimensional Coulomb's friction. Most of these formulations are well–known concepts in the optimization community, or more generally, in the mathematical programming community. To cite a few, the discrete frictional contact problem can be formulated as variational inequalities, generalized or semi–smooth equations, second–order cone complementarity problems, or as optimization problems such as quadratic programming problems over second-order cones. Thanks to these multiple formulations, various numerical methods emerge naturally for solving the problem. We review the main numerical techniques that are well-known in the literature and we also propose new applications of methods such as the fixed point and extra-gradient methods with self-adaptive step rules for variational inequalities or the proximal point algorithm for generalized equations. All these numerical techniques are compared over a large set of test examples using performance profiles. One of the main conclusion is that there is no universal solver. Nevertheless, we are able to give some hints to choose a solver with respect to the main characteristics of the set of tests

#### *6.5.4.1. Impact laws in chains of aligned balls*

In [18] several "classical" multiple-impact laws are compared on chains of aligned balls: Moreau's law, the binary collision law, and the LZB approach [2]. Short analyses of these laws are made, and thorough comparisons are led numerically. It is concluded that both Moreau and the binary collision laws, furnish good results (in terms of predictability) only in very particular cases of elasticity coefficien, contact stiffenesses ratios, and mass ratios.

## 6.6. Analysis and Control of Set-Valued Systems

**Participants:** Bernard Brogliato, Christophe Prieur, Alexandre Vieira.

### *6.6.1. Higher-order sweeping process*

This work [5] continues our previous results in [33], to the case when exogeneous terms are present in both the unilateral constraint, and in the dynamics. A suitable change of state variables allows one to recast the dynamics in a format that is close to the autonomous case, so that the well-posedness issues (existence and uniqueness of solutions) is shown (see the preprint [20] for a complete analysis, which in fact differs only slightly from the original one in [33]). The link with switching DAEs is made.

### *6.6.2. Robust sliding-mode control: continuous and discrete-time*

This work [10] concerns the robust control of linear time-invariant systems, subjected to nonlinear varying state dependent disturbances as well as parameter uncertainties. A specific set-valued class of sliding-mode controllers is designed, and its discretization (with the implicit method introduced in [32], [34]) is analysed. One difficulty is that the parameter uncertainties, as well as the discretization, create unmatched disturbances. Stability and convergence results are proved. Let us mention also [9] that corrects a slight mistake in [80]. In the same way it is worth citing [14], [14] which continues the analysis of the implicit discretization of set-valued systems, this time oriented towards the consistency of time-discretizations for homogeneous systems, with one discontinuity at zero (sometimes called quasi-continuous, strangely enough).

### *6.6.3. Evolution variational inequalities*

In [13] we continue our previous works on well-posedness and stabilization/control of a class of set-valued systems, that take the form of evolution variational inequalities. Dissipativity is then a key property. Regulation with state and output feedback, viability issues, are solved, with absolutely continuous and bounded variations solutions. Applications are in power converters.

### *6.6.4. Optimal control of LCS*

The quadratic and minimum time optimal control of LCS as in (6 ) is tackled in [24], [25]. This work relies on the seminal results by Guo and ye (SIAM 2016), and aims at particularizing their results for LCS, so that they become numerically tractable and one can compute optimal controllers and optimal trajectories. The basic idea is to take advantage of the complementarity, to construct linear complementarity problems in the Pontryagin's necessary conditions which can then be integrated numerically, without having to guess a priori the switching instants (the optimal controller can eb discontinuous and the optimal trajectories can visit several modes of the complementarity conditions).

**TYREX Project-Team**

# 6. New Results

## 6.1. On the Optimization of Recursive Relational Queries

Graph databases have received a lot of attention recently as they are particularly useful in many applications such as social networks or for the semantic web. Various languages have emerged to query such graph databases. At the heart of many of those query languages, there is a construction to navigate through the graph which allows some form of recursion.The relational model has benefited from a huge body of research in the last half century and that is why many graph databases either rely on, or have adopted the techniques of, relational-based query engines. Since its introduction, the relational model has seen various attempts to extend it with recursion and it is now possible to use recursion in several SQL- or Datalog-based database systems. The optimization of recursive queries remains, however, a challenge. In this work, we introduce $\mu$-RA, a variation of the Relational Algebra that allows for the expression of relational queries with recursion. $\mu$-RA can express unions of conjunctive regular path queries as well as certain non-regular properties. We present its syntax, semantics and the rewriting rules we specifically devised to tackle the optimization of recursive queries. A prototype evaluator implementing these rewriting rules is shown to be more efficient than previous approaches.

These results were presented at the BDA 2018 conference [14].

## 6.2. A Multi-Criteria Experimental Ranking of Distributed SPARQL Evaluators

SPARQL is the standard language for querying RDF data. There exists a variety of SPARQL query evaluation systems implementing different architectures for the distribution of data and computations. Differences in architectures coupled with specific optimizations, for e.g. preprocessing and indexing, make these systems incomparable from a purely theoretical perspective. This results in many implementations solving the SPARQL query evaluation problem while exhibiting very different behaviours, not all of them being adapted to any context. We provide a new perspective on distributed SPARQL evaluators, based on multi-criteria experimental rankings. Our suggested set of 5 features (namely velocity, immediacy, dynamicity, parsimony, and resiliency) provides a more comprehensive description of the behaviours of distributed evaluators when compared to traditional runtime performance metrics. We show how these features help in more accurately evaluating to which extent a given system is appropriate for a given use case. For this purpose, we systematically benchmarked a panel of 10 state-of-the-art implementations. We ranked them using a reading grid that helps in pinpointing the advantages and limitations of current technologies for the distributed evaluation of SPARQL queries.

These results were presented at the IEEE Big Data 2018 conference [13].

## 6.3. SPARQL Query Containment under Schema

Query containment is defined as the problem of determining if the result of a query is included in the result of another query for any dataset. It has major applications in query optimization and knowledge base verification. The main objective of this work is to provide sound and complete procedures to determine containment of SPARQL queries under expressive description logic schema axioms. Beyond that, these procedures are experimentally evaluated. To date, testing query containment has been performed using different techniques: containment mapping, canonical databases, automata theory techniques and through a reduction to the validity problem in logic. In this work, we use the latter technique to test containment of SPARQL queries using an expressive modal logic called $\mu$-calculus. For that purpose, we define an RDF graph encoding as a transition system which preserves its characteristics. In addition, queries and schema axioms are encoded as $\mu$-calculus formulae. Thereby, query containment can be reduced to testing validity in the logic. We identify

various fragments of SPARQL and description logic schema languages for which containment is decidable. Additionally, we provide theoretically and experimentally proven procedures to check containment of these decidable fragments. Finally, we propose a benchmark for containment solvers which is used to test and compare the current state-of-the-art containment solvers.

These results were published in the Journal on Data Semantics [4].

## 6.4. Selectivity Estimation for SPARQL Triple Patterns with Shape Expressions

ShEx (Shape Expressions) is a language for expressing constraints on RDF graphs. In this work we optimize the evaluation of conjunctive SPARQL queries, on RDF graphs, by taking advantage of ShEx constraints. Our optimization is based on computing and assigning ranks to query triple patterns, dictating their order of execution. We first define a set of well-formed ShEx schemas that possess interesting characteristics for SPARQL query optimization. We then define our optimization method by exploiting information extracted from a ShEx schema. We finally report on evaluation results performed showing the advantages of applying our optimization on the top of an existing state-of-the-art query evaluation system.

These results were presented at the 2018 International Conference on Web Engineering [9].

## 6.5. Evaluation of Query Transformations without Data

Query transformations are ubiquitous in semantic web query processing. For any situation in which transformations are not proved correct by construction, the quality of these transformations has to be evaluated. Usual evaluation measures are either overly syntactic and not very informative — the result being: correct or incorrect — or dependent from the evaluation sources. Moreover, both approaches do not necessarily yield the same result. We suggest that grounding the evaluation on query containment allows for a data-independent evaluation that is more informative than the usual syntactic evaluation. In addition, such evaluation modalities may take into account ontologies, alignments or different query languages as soon as they are relevant to query evaluation.

These results were presented at a workshop of the 2018 International Conference on World Wide Web [10].

## 6.6. Graph Queries: From Theory to Practice

In this work, we review various graph query language fragments that are both theoretically tractable and practically relevant. We focus on the most expressive one that retains these properties and use it as a stepping stone to examine the underpinnings of graph query evaluation along graph view maintenance. Further broadening the scope of the discussion, we then consider alternative processing techniques for graph queries, based on graph summarization and path query learning. We conclude by pinpointing the open research directions in this emerging area. These results were published in Sigmod Record Journal [3].

## 6.7. Query-based Linked Data Anonymization

In this work, we introduce and develop a declarative framework for privacy-preserving Linked Data publishing in which privacy and utility policies are specified as SPARQL queries. Our approach is data independent and leads to inspect only the privacy and utility policies in order to determine the sequence of anonymization operations applicable to any graph instance for satisfying the policies. We prove the soundness of our algorithms and gauge their performance through experiments.

These results were presented in the International Semantic Web Conference (ISWC 2018) [11].

## 6.8. Querying Graphs

Graph data modeling and querying arises in many practical application domains such as social and biological networks where the primary focus is on concepts and their relationships and the rich patterns in these complex webs of interconnectivity. In this book, we present a concise unified view on the basic challenges which arise over the complete life cycle of formulating and processing queries on graph databases. To that purpose, we present all major concepts relevant to this life cycle, formulated in terms of a common and unifying ground: the property graph data model — the predominant data model adopted by modern graph database systems.

In this book [17], we aim especially to give a coherent and in-depth perspective on current graph querying and an outlook for future developments. Our presentation is self-contained, covering the relevant topics from: graph data models, graph query languages and graph query specification, graph constraints, and graph query processing. We conclude by indicating major open research challenges towards the next generation of graph data management systems.

## 6.9. Backward Type Inference for XML Queries

Although XQuery is a statically typed, functional query language for XML data, some of its features such as upward and horizontal XPath axes are typed imprecisely. The main reason is that while the XQuery data model allows to navigate upwards and between siblings from a given XML node, the type model, e.g., regular tree types, can describe only the subtree structure of the given node. Recently, Giuseppe Castagna and our team independently proposed in 2015 a precise forward type inference system for XQuery using an extended type language that can describe not only a given XML node but also its context. In this work, as a complementary method to such forward type inference systems, we propose an enhanced backward type inference system for XQuery, based on an extended type language. Results include an exact type system for XPath axes and a sound type system for XQuery expressions [19].

## 6.10. Scalable and Interpretable Predictive Models for Electronic Health Records

Early identification of patients at risk of developing complications during their hospital stay is currently one of the most challenging issues in healthcare. Complications include hospital-acquired infections, admissions to intensive care units, and in-hospital mortality. Being able to accurately predict the patients' outcomes is a crucial prerequisite for tailoring the care that certain patients receive, if it is believed that they will do poorly without additional intervention. We consider the problem of complication risk prediction, such as patient mortality, from the electronic health records of the patients. We study the question of making predictions on the first day at the hospital, and of making updated mortality predictions day after day during the patient's stay. We develop distributed models that are scalable and interpretable. Key insights include analysing diagnoses known at admission and drugs served, which evolve during the hospital stay. We leverage a distributed architecture to learn interpretable models from training datasets of gigantic size. We test our analyses with more than one million of patients from hundreds of hospitals, and report on the lessons learned from these experiments.

These results were presented at the 2018 International Conference on Data Science and Applications [12].

## 6.11. Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission

We show how the analysis of very large amounts of drug prescription data make it possible to detect, on the day of hospital admission, patients at risk of developing complications during their hospital stay. We explore, for the first time, to which extent volume and variety of big prescription data help in constructing predictive models for the automatic detection of at-risk profiles. Our methodology is designed to validate our claims that: (1) drug prescription data on the day of admission contain rich information about the patient's situation and perspectives of evolution, and (2) the various perspectives of big medical data (such as veracity, volume, variety) help in extracting this information.We build binary classification models to identify at-risk patient

profiles. We use a distributed architecture to ensure scalability of model construction with large volumes of medical records and clinical data. We report on practical experiments with real data of millions of patients and hundreds of hospitals. We demonstrate how the fine-grained analysis of such big data can improve the detection of at-risk patients, making it possible to construct more accurate predictive models that significantly benefit from volume and variety, while satisfying important criteria to be deployed in hospitals.

These results were published in the Big Data Research journal [6].

## 6.12. ProvSQL: Provenance and Probability Management in PostgreSQL

This demonstration showcases ProvSQL, an open-source module for the PostgreSQL database management system that adds support for computation of provenance and probabilities of query results. A large range of provenance formalisms are supported, including all those captured by provenance semirings, provenance semirings with monus, as well as where-provenance. Probabilistic query evaluation is made possible through the use of knowledge compilation tools, in addition to standard approaches such as enumeration of possible worlds and Monte-Carlo sampling. ProvSQL supports a large subset of non-aggregate SQL queries.

These results were published in the PVLDB journal [8].

## 6.13. A Method to Quantitatively Evaluate Geo Augmented Reality Applications

We propose a method for quantitatively assessing the quality of Geo AR browsers. Our method aims at measuring the impact of attitude and position estimations on the rendering precision of virtual features. We report on lessons learned by applying our method on various AR use cases with real data. Our measurement technique allows shedding light on the limits of what can be achieved in Geo AR with current technologies. This also helps in identifying interesting perspectives for the further development of high-quality Geo AR applications.

These results were presented at the ISMAR 2018 conference [15].

## 6.14. Attitude Estimation for Indoor Navigation and Augmented Reality with Smartphones

We investigate the precision of attitude estimation algorithms in the particular context of pedestrian navigation with commodity smartphones and their inertial/magnetic sensors. We report on an extensive comparison and experimental analysis of existing algorithms. We focus on typical motions of smartphones when carried by pedestrians. We use a precise ground truth obtained from a motion capture system. We test state-of-the-art and built-in attitude estimation techniques with several smartphones, in the presence of magnetic perturbations typically found in buildings. We discuss the obtained results, analyze advantages and limits of current technologies for attitude estimation in this context. Furthermore, we propose a new technique for limiting the impact of magnetic perturbations with any attitude estimation algorithm used in this context. We show how our technique compares and improves over previous works. A particular attention was paid to the study of attitude estimation in the context of augmented reality motions when using smartphones.

These results were published in the Pervasive and Mobile Computing journal [7].

## 6.15. A Hybrid Approach for Spatio-Temporal Validation of Declarative Multimedia

Declarative multimedia documents represent the description of multimedia applications in terms of media items and relationships among them. Relationships specify how media items are dynamically arranged in time and space during runtime. Although a declarative approach usually facilitates the authoring task, authors can still make mistakes due to incorrect use of language constructs or inconsistent or missing relationships in a document. In order to properly support multimedia application authoring, it is important to provide tools with

validation capabilities. Document validation can indicate possible inconsistencies in a given document to an author so that it can be revised before deployment. Although very useful, multimedia validation tools are not often provided by authoring tools. This work proposes a multimedia validation approach that relies on a formal model called Simple Hyper-media Model (SHM). SHM is used for representing a document for the purpose of validation. An SHM document is validated using a hybrid approach based on two complementary techniques. The first one captures the document's spatio-temporal layout in terms of its state throughout its execution by means of a rewrite theory, and validation is performed through model checking. The second one captures the document's layout in terms of intervals and event occurrences by means of Satisfiability Modulo Theories (SMT) formulas, and validation is performed through SMT solving. Due to different characteristics of both approaches, each validation technique complements the other in terms of expressiveness of SHM and tests to be checked. We briefly present validation tools that use our approach. They were evaluated with real NCL documents and by usability tests.

These results were published in the ACM Transactions on Multimedia Computing, Communications and Applications journal [5].