
RESEARCH CENTER
**Lille - Nord Europe**

FIELD

# Activity Report 2014

# Section Scientific Foundations

Edition: 2015-03-24

**ATEAMS Project-Team**

# 3. Research Program

## 3.1. Research method

We are inspired by formal methods and logic to construct new tools for software analysis, transformation and generation. We try and proof the correctness of new algorithms using any means necessary.

Nevertheless we mainly focus on the study of existing (large) software artifacts to validate the effectiveness of new tools. We apply the scientific method. To (in)validate our hypothesis we often use detailed manual source code analysis, or we use software metrics, and we have started to use more human subjects (programmers).

Note that we maintain ties with the CWI spinoff "Software Improvement Group" which services most of the Dutch software industry and government and many European companies as well. This provides access to software systems and information about software systems that is valuable in our research.

## 3.2. Software analysis

This research focuses on source code; to analyze it, transform it and generate it. Each analysis or transformation begins with fact extraction. After that we may analyze specific software systems or large bodies of software systems. Our goal is to improve software systems by understanding and resolving the causes of software complexity. The approach is captured in the EASY acronym: Extract Analyze SYnthesize. The first step is to extract facts from source code. These facts are then enriched and refined in an analysis phase. Finally the result is synthesized in the form of transformed or generated source code, a metrics report, a visualization or some other output artifact.

The mother and father of fact extraction techniques are probably Lex, a scanner generator, and AWK, a language intended for fact extraction from textual records and report generation. Lex is intended to read a file character-by-character and produce output when certain regular expressions (for identifiers, floating point constants, keywords) are recognized. AWK reads its input line-by-line and regular expression matches are applied to each line to extract facts. User-defined actions (in particular print statements) can be associated with each successful match. This approach based on regular expressions is in wide use for solving many problems such as data collection, data mining, fact extraction, consistency checking, and system administration. This same approach is used in languages like Perl, Python, and Ruby. Murphy and Notkin have specialized the AWK-approach for the domain of fact extraction from source code. The key idea is to extend the expressivity of regular expressions by adding context information, in such a way that, for instance, the begin and end of a procedure declaration can be recognized. This approach has, for instance, been used for call graph extraction but becomes cumbersome when more complex context information has to be taken into account such as scope information, variable qualification, or nested language constructs. This suggests using grammar-based approaches as will be pursued in the proposed project. Another line of research is the explicit instrumentation of existing compilers with fact extraction capabilities. Examples are: the GNU C compiler GCC, the CPPX C++ compiler, and the Columbus C/C++ analysis framework. The Rigi system provides several fixed fact extractors for a number of languages. The extracted facts are represented as tuples (see below). The CodeSurfer source code analysis tool extracts a standard collection of facts that can be further analyzed with built-in tools or user-defined programs written in Scheme. In all these cases the programming language as well as the set of extracted facts are fixed thus limiting the range of problems that can be solved.

The approach we are exploring is the use of syntax-related program patterns for fact extraction. An early proposal for such a pattern-based approach consisted of extending a fixed base language (either C or PL/1 variant) with pattern matching primitives. In our own previous work on RScript we have already proposed a query algebra to express direct queries on the syntax tree. It also allows the querying of information that is attached to the syntax tree via annotations. A unifying view is to consider the syntax tree itself as "facts" and to represent it as a relation. This idea is already quite old. For instance, Linton proposes to represent all syntactic as well as semantic aspects of a program as relations and to use SQL to query them. Due to the lack of expressiveness of SQL (notably the lack of transitive closure) and the performance problems encountered, this approach has not seen wider use.

Parsing is a fundamental tool for fact extraction for source code. Our group has longstanding contributions in the field of Generalized LR parsing and Scannerless parsing. Such generalized parsing techniques enable generation of parsers for a wide range of existing (legacy) programming languages, which is highly relevant for experimental research and validation.

Extracted facts are often refined, enriched and queried in the analysis phase. We propose to use a relational formalization of the facts. That is, facts are represented as sets of tuples, which can then be queried using relational algebra operators (e.g., domain, transitive closure, projection, composition etc.). This relational representation facilitates dealing with graphs, which are commonly needed during program analysis, for instance when processing control-flow or data-flow graphs. The Rascal language integrates a relational sub-language by providing comprehensions over different kinds of data types, in combination with powerful pattern matching and built-in primitives for computing (transitive/reflexive) closures and fixpoint computations (equation solving).

### *3.2.1. Goals*

The main goal is to replace labour-intensive manual programming of fact extractors by automatic generation based on concise and formal specification. There is a wide open scientific challenge here: to create a uniform and generic framework for fact extraction that is superior to current more ad-hoc approaches, yet flexible enough to be customized to the analysis case at hand. We expect to develop new ideas and techniques for generic (language-parametric) fact extraction from source code and other software artifacts.

Given the advances made in fact extraction we are starting to apply our techniques to observe source code and analyze it in detail.

## 3.3. Refactoring and Transformation

The second goal, to be able to safely refactor or transform source code can be realized in strong collaboration with extraction and analysis.

Software refactoring is usually understood as changing software with the purpose of increasing its readability and maintainability rather than changing its external behavior. Refactoring is an essential tool in all agile software engineering methodologies. Refactoring is usually supported by an interactive refactoring tool and consists of the following steps:

- Select a code fragment to refactor.
- Select a refactoring to apply to it.
- Optionally, provide extra parameter needed by the refactoring (e.g., a new name in a renaming).

The refactoring tool will now test whether the preconditions for the refactoring are satisfied. Note that this requires fact extraction from the source code. If this fails the user is informed. The refactoring tool shows the effects of the refactoring before effectuating them. This gives the user the opportunity to disable the refactoring in specific cases.The refactoring tool applies the refactoring for all enabled cases. Note that this implies a transformation of the source code. Some refactorings can be applied to any programming language (e.g., rename) and others are language specific (e.g., Pull Up Method). At http://www.refactoring.com an extensive list of refactorings can be found.

There is hardly any general and pragmatic theory for refactoring, since each refactoring requires different static analysis techniques to be able to check the preconditions. Full blown semantic specification of programming languages have turned out to be infeasible, let alone easily adaptable to small changes in language semantics. On the other hand, each refactoring is an instance of the extract, analyze and transform paradigm. Software transformation regards more general changes such as adding functionality and improving non-functional properties like performance and reliability. It also includes transformation from/to the same language (source-to-source translation) and transformation between different languages (conversion, translation). The underlying techniques for refactoring and transformation are mostly the same. We base our source code transformation techniques on the classical concept of term rewriting, or aspects thereof. It offers simple but powerful pattern matching and pattern construction features (list matching, AC Matching), and type-safe heterogenous data-structure traversal methods that are certainly applicable for source code transformation.

### *3.3.1. Goals*

Our goal is to integrate the techniques from program transformation completely with relational queries. Refactoring and transformation form the Achilles Heel of any effort to change and improve software. Our innovation is in the strict language-parametric approach that may yield a library of generic analyses and transformations that can be reused across a wide range of programming and application languages. The challenge is to make this approach scale to large bodies of source code and rapid response times for precondition checking.

## 3.4. The Rascal Meta-programming language

The Rascal Domain-Specific Language for Source code analysis and Transformation is developed by ATeams. It is a language specifically designed for any kind of meta programming.

Meta programming is a large and diverse area both conceptually and technologically. There are plentiful libraries, tools and languages available but integrated facilities that combine both source code analysis and source code transformation are scarce. Both domains depend on a wide range of concepts such as grammars and parsing, abstract syntax trees, pattern matching, generalized tree traversal, constraint solving, type inference, high fidelity transformations, slicing, abstract interpretation, model checking, and abstract state machines. Examples of tools that implement some of these concepts are ANTLR, ASF+SDF, CodeSurfer, Crocopat, DMS, Grok, Stratego, TOM and TXL. These tools either specialize in analysis or in transformation, but not in both. As a result, combinations of analysis and transformation tools are used to get the job done. For instance, ASF+SDF relies on RScript for querying and TXL interfaces with databases or query tools. In other approaches, analysis and transformation are implemented from scratch, as done in the Eclipse JDT. The TOM tool adds transformation primitives to Java, such that libraries for analysis can be used directly. In either approach, the job of integrating analysis with transformation has to be done over and over again for each application and this requires a significant investment.

We propose a more radical solution by completely merging the set of concepts for analysis and transformation of source code into a single language called Rascal. This language covers the range of applications from pure analyses to pure transformations and everything in between. Our contribution does not consist of new concepts or language features *per se*, but rather the careful collaboration, integration and cross-fertilization of existing concepts and language features.

### *3.4.1. Goals*

The goals of Rascal are: (a) to remove the cognitive and computational overhead of integrating analysis and transformation tools, (b) to provide a safe and interactive environment for constructing and experimenting with large and complicated source code analyses and transformations such as, for instance, needed for refactorings, and (c) to be easily understandable by a large group of computer programming experts. Rascal is not limited to one particular object programming language, but is generically applicable. Reusable, language specific, functionality is realized as libraries. As an end-result we envision Rascal to be a one-stop shop for source code analysis, transformation, generation and visualization.

## 3.5. Domain-specific Languages

Our final goal is centered around Domain-specific languages (DSLs), which are software languages tailored to a specific problem domain. DSLs can provide orders of magnitude improvement in terms of software quality and productivity. However, the implementation of DSLs is challenging and requires not only thorough knowledge of the problem domain (e.g., finance, digital forensics, insurance, auditing etc.), but also knowledge of language implementation (e.g., parsing, compilation, type checking etc.). Tools for language implementation have been around since the archetypical parser generator YACC. However, many of such tools are characterized by high learning curves, lack of integration of language implementation facets, and lead to implementations that are hard to maintain. This line of research focuses on two topics: improve the practice and experience of DSL implementation, and evaluate the success of DSLs in industrial practice.

Language workbenches [4] are integrated environments to facilitate the development of all aspects of DSLs. This includes IDE support (e.g., syntax coloring, outlining, reference resolving etc.) for the defined languages. Rascal can be seen as a language workbench that focuses on flexibility, programmability and modularity. DSL implementation is, in essence, an instance of source code analysis and transformation. As a result, Rascal's features for fact extraction, analysis, tree traversal and synthesis are an excellent fit for this area. An important aspect in this line of research is bringing the IDE closer to the source code. This will involve investigation of heterogeneous representations of source code, by integrating graphical, tabular or forms-based user interface elements. As a result, we propose Rascal as a feature-rich workbench for model-driven software development.

The second component of this research is concerned with evaluating DSLs in industrial contexts. This means that DSLs constructed using Rascal will be applied in real-life environments so that expected improvements in quality, performance, or productivity can be observed. We already have experience with this in the domain of digital forensics, computational auditing and games.

### *3.5.1. Goals*

The goal of this research topic is to improve the practice of DSL-based software development through language design and tool support. A primary focus is to extend the IDE support provided by Rascal, and to facilitate incremental, and iterative design of DSLs. The latter is supported by new (meta-)language constructs for extending existing language implementations. This will require research into extensible programming and composition of compilers, interpreters and type checkers. Finally, a DSL is never an island: it will have to integrate with (third-party) source code, such as host language, libraries, runtime systems etc. This leads to the vision of multi-lingual programming environments [15].

**BONSAI Project-Team**

# 3. Research Program

## 3.1. Combinatorial discrete models and algorithms

Our research is driven by biological questions. At the same time, we have in mind to develop well-founded models and efficient algorithms. Biological macromolecules are naturally modelled by various types of discrete structures: String, trees, and graphs. String algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years. Members of the team also have a strong expertise in text indexing and compressed index data structures, such as BWT. Such methods are widely-used for the analysis of biological sequences because they allow a data set to be stored and queried efficiently. Ordered trees and graphs naturally arise when dealing with structures of molecules, such as RNAs or non-ribosomal peptides. The underlying questions are: How to compare molecules at structural level, how to search for structural patterns ? String, trees and graphs are also useful to study genomic rearrangements: Neighborhoods of genes can be modelled by oriented graphs, genomes as permutations, strings or trees. High-performance computing is another tool that we use to achieve our goals.

## 3.2. Discrete statistics and probability

At a lower level, our work relies on a basic background on discrete statistics and probability. When dealing with large input data sets, it is essential to be able to discriminate between noisy features observed by chance from those that are biologically relevant. The aim here is to introduce a probabilistic model and to use sound statistical methods to assess the significance of some observations about these data. Examples of such observations are the length of a repeated region, the number of occurrences of a motif (DNA or RNA), the free energy of a conserved RNA secondary structure, etc. Probabilistic models are also used to describe genome evolution. In this context, Bayesian models and MCMC sampling allow to approximate probability distributions over free parameters and to describe biologically relevant models.

**DOLPHIN Project-Team**

# 3. Research Program

## 3.1. Hybrid multi-objective optimization methods

The success of metaheuristics is based on their ability to find efficient solutions in a reasonable time [58]. But with very large problems and/or multi-objective problems, efficiency of metaheuristics may be compromised. Hence, in this context it is necessary to integrate metaheuristics in more general schemes in order to develop even more efficient methods. For instance, this can be done by different strategies such as cooperation and parallelization.

The DOLPHIN project deals with *"a posteriori"* multi-objective optimization where the set of Pareto solutions (solutions of best compromise) have to be generated in order to give the decision maker the opportunity to choose the solution that interests him/her.

Population-based methods, such as evolutionary algorithms, are well fitted for multi-objective problems, as they work with a set of solutions [53], [57]. To be convinced one may refer to the list of references on Evolutionary Multi-objective Optimization maintained by Carlos A. Coello [0], which contains more than 5500 references. One of the objectives of the project is to propose advanced search mechanisms for intensification and diversification. These mechanisms have been designed in an adaptive manner, since their effectiveness is related to the landscape of the MOP and to the instance solved.

In order to assess the performances of the proposed mechanisms, we always proceed in two steps: first, we carry out experiments on academic problems, for which some best known results exist; second, we use real industrial problems to cope with large and complex MOPs. The lack of references in terms of optimal or best known Pareto set is a major problem. Therefore, the obtained results in this project and the test data sets will be available at the URL http://dolphin.lille.inria.fr/ at 'benchmark'.

### *3.1.1. Cooperation of metaheuristics*

In order to benefit from the various advantages of the different metaheuristics, an interesting idea is to combine them. Indeed, the hybridization of metaheuristics allows the cooperation of methods having complementary behaviors. The efficiency and the robustness of such methods depend on the balance between the exploration of the whole search space and the exploitation of interesting areas.

Hybrid metaheuristics have received considerable interest these last years in the field of combinatorial optimization. A wide variety of hybrid approaches have been proposed in the literature and give very good results on numerous single objective optimization problems, which are either academic (traveling salesman problem, quadratic assignment problem, scheduling problem, etc) or real-world problems. This efficiency is generally due to the combinations of single-solution based methods (iterative local search, simulated annealing, tabu search, etc) with population-based methods (genetic algorithms, ants search, scatter search, etc). A taxonomy of hybridization mechanisms may be found in [62]. It proposes to decompose these mechanisms into four classes:

- *LRH class - Low-level Relay Hybrid*: This class contains algorithms in which a given metaheuristic is embedded into a single-solution metaheuristic. Few examples from the literature belong to this class.
- *LTH class - Low-level Teamwork Hybrid*: In this class, a metaheuristic is embedded into a population-based metaheuristic in order to exploit strengths of single-solution and population-based metaheuristics.

[0] http://www.lania.mx/~ccoello/EMOO/EMOObib.html

- *HRH class - High-level Relay Hybrid*: Here, self contained metaheuristics are executed in a sequence. For instance, a population-based metaheuristic is executed to locate interesting regions and then a local search is performed to exploit these regions.
- *HTH class - High-level Teamwork Hybrid*: This scheme involves several self-contained algorithms performing a search in parallel and cooperating. An example will be the island model, based on GAs, where the population is partitioned into small subpopulations and a GA is executed per subpopulation. Some individuals can migrate between subpopulations.

Let us notice that, hybrid methods have been studied in the mono-criterion case, their application in the multi-objective context is not yet widely spread. The objective of the DOLPHIN project is to integrate specificities of multi-objective optimization into the definition of hybrid models.

### *3.1.2. Cooperation between metaheuristics and exact methods*

Until now only few exact methods have been proposed to solve multi-objective problems. They are based either on a Branch-and-bound approach, on the algorithm $A^{\star}$, or on dynamic programming. However, these methods are limited to two objectives and, most of the time, cannot be used on a complete large scale problem. Therefore, sub search spaces have to be defined in order to use exact methods. Hence, in the same manner as hybridization of metaheuristics, the cooperation of metaheuristics and exact methods is also a main issue in this project. Indeed, it allows us to use the exploration capacity of metaheuristics, as well as the intensification ability of exact methods, which are able to find optimal solutions in a restricted search space. Sub search spaces have to be defined along the search. Such strategies can be found in the literature, but they are only applied to mono-objective academic problems.

We have extended the previous taxonomy for hybrid metaheuristics to the cooperation between exact methods and metaheuristics. Using this taxonomy, we are investigating cooperative multi-objective methods. In this context, several types of cooperations may be considered, according to the way the metaheuristic and the exact method cooperate. For instance, a metaheuristic can use an exact method for intensification or an exact method can use a metaheuristic to reduce the search space.

Moreover, a part of the DOLPHIN project deals with studying exact methods in the multi-objective context in order: i) to be able to solve small size problems and to validate proposed heuristic approaches; ii) to have more efficient/dedicated exact methods that can be hybridized with metaheuristics. In this context, the use of parallelism will push back limits of exact methods, which will be able to explore larger size search spaces [55].

### *3.1.3. Goals*

Based on the previous works on multi-objective optimization, it appears that to improve metaheuristics, it becomes essential to integrate knowledge about the problem structure. This knowledge can be gained during the search. This would allow us to adapt operators which may be specific for multi-objective optimization or not. The goal here is to design auto-adaptive methods that are able to react to the problem structure. Moreover, regarding the hybridization and the cooperation aspects, the objectives of the DOLPHIN project are to deepen these studies as follows:

- *Design of metaheuristics for the multi-objective optimization:* To improve metaheuristics, it becomes essential to integrate knowledge about the problem structure, which we may get during the execution. This would allow us to adapt operators that may be specific for multi-objective optimization or not. The goal here is to design auto-adaptive methods that are able to react to the problem structure.
- *Design of cooperative metaheuristics:* Previous studies show the interest of hybridization for a global optimization and the importance of problem structure study for the design of efficient methods. It is now necessary to generalize hybridization of metaheuristics and to propose adaptive hybrid models that may evolve during the search while selecting the appropriate metaheuristic. Multi-objective aspects have to be introduced in order to cope with the specificities of multi-objective optimization.

- *Design of cooperative schemes between exact methods and metaheuristics:* Once the study on possible cooperation schemes is achieved, we will have to test and compare them in the multi-objective context.
- *Design and conception of parallel metaheuristics:* Our previous works on parallel metaheuristics allow us to speed up the resolution of large scale problems. It could be also interesting to study the robustness of the different parallel models (in particular in the multi-objective case) and to propose rules that determine, given a specific problem, which kind of parallelism to use. Of course these goals are not disjoined and it will be interesting to simultaneously use hybrid metaheuristics and exact methods. Moreover, those advanced mechanisms may require the use of parallel and distributed computing in order to easily make cooperating methods evolve simultaneously and to speed up the resolution of large scale problems.
- *Validation:* In order to validate the obtained results we always proceed in two phases: validation on academic problems, for which some best known results exist and use on real problems (industrial) to cope with problem size constraints.

  Moreover, those advanced mechanisms are to be used in order to integrate the distributed multi-objective aspects in the ParadisEO platform (see the paragraph on software platform).

## 3.2. Parallel multi-objective optimization: models and software frameworks

Parallel and distributed computing may be considered as a tool to speedup the search to solve large MOPs and to improve the robustness of a given method. Moreover, the joint use of parallelism and cooperation allows improvements on the quality of the obtained Pareto sets. Following this objective, we will design and implement parallel models for metaheuristics (evolutionary algorithms, tabu search approach) and exact methods (branch-and-bound algorithm, branch-and-cut algorithm) to solve different large MOPs.

One of the goals of the DOLPHIN project is to integrate the developed parallel models into software frameworks. Several frameworks for parallel distributed metaheuristics have been proposed in the literature. Most of them focus only either on evolutionary algorithms or on local search methods. Only few frameworks are dedicated to the design of both families of methods. On the other hand, existing optimization frameworks either do not provide parallelism at all or just supply at most one parallel model. In this project, a new framework for parallel hybrid metaheuristics is proposed, named *Parallel and Distributed Evolving Objects (ParadisEO)* based on EO. The framework provides in a transparent way the hybridization mechanisms presented in the previous section, and the parallel models described in the next section. Concerning the developed parallel exact methods for MOPs, we will integrate them into well-known frameworks such as COIN.

### *3.2.1. Parallel models*

According to the family of addressed metaheuristics, we may distinguish two categories of parallel models: parallel models that manage a single solution, and parallel models that handle a population of solutions. The major single solution-based parallel models are the following: the *parallel neighborhood exploration model* and the *multi-start model*.

- *The parallel neighborhood exploration model* is basically a "low level" model that splits the neighborhood into partitions that are explored and evaluated in parallel. This model is particularly interesting when the evaluation of each solution is costly and/or when the size of the neighborhood is large. It has been successfully applied to the mobile network design problem (see Application section).
- *The multi-start model* consists in executing in parallel several local searches (that may be heterogeneous), without any information exchange. This model raises particularly the following question: is it equivalent to execute $k$ local searches during a time $t$ than executing a single local search during $k \times t$? To answer this question we tested a multi-start Tabu search on the quadratic assignment problem. The experiments have shown that the answer is often landscape-dependent. For example, the multi-start model may be well-suited for landscapes with multiple basins.

Parallel models that handle a population of solutions are mainly: the *island model*, the *central model* and *the distributed evaluation of a single solution*. Let us notice that the last model may also be used with single-solution metaheuristics.

- In *the island model*, the population is split into several sub-populations distributed among different processors. Each processor is responsible of the evolution of one sub-population. It executes all the steps of the metaheuristic from the selection to the replacement. After a given number of generations (synchronous communication), or when a convergence threshold is reached (asynchronous communication), the migration process is activated. Then, exchanges of solutions between sub-populations are realized, and received solutions are integrated into the local sub-population.
- *The central (Master/Worker) model* allows us to keep the sequentiality of the original algorithm. The master centralizes the population and manages the selection and the replacement steps. It sends sub-populations to the workers that execute the recombination and evaluation steps. The latter returns back newly evaluated solutions to the master. This approach is efficient when the generation and evaluation of new solutions is costly.
- *The distributed evaluation model* consists in a parallel evaluation of each solution. This model has to be used when, for example, the evaluation of a solution requires access to very large databases (data mining applications) that may be distributed over several processors. It may also be useful in a multi-objective context, where several objectives have to be computed simultaneously for a single solution.

As these models have now been identified, our objective is to study them in the multi-objective context in order to use them advisedly. Moreover, these models may be merged to combine different levels of parallelism and to obtain more efficient methods [56], [61].

### *3.2.2. Goals*

Our objectives focus on these issues are the following:

- *Design of parallel models for metaheuristics and exact methods for MOPs*: We will develop parallel cooperative metaheuristics (evolutionary algorithms and local search algorithms such as the Tabu search) for solving different large MOPs. Moreover, we are designing a new exact method, named PPM (Parallel Partition Method), based on branch and bound and branch and cut algorithms. Finally, some parallel cooperation schemes between metaheuristics and exact algorithms have to be used to solve MOPs in an efficient manner.
- *Integration of the parallel models into software frameworks*: The parallel models for metaheuristics will be integrated in the ParadisEO software framework. The proposed multi-objective exact methods must be first integrated into standard frameworks for exact methods such as COIN and BOB++. A *coupling* with ParadisEO is then needed to provide hybridization between metaheuristics and exact methods.
- *Efficient deployment of the parallel models on different parallel and distributed architectures including GRIDs*: The designed algorithms and frameworks will be efficiently deployed on non-dedicated networks of workstations, dedicated cluster of workstations and SMP (Symmetric Multi-processors) machines. For GRID computing platforms, peer to peer (P2P) middlewares (XtremWeb-Condor) will be used to implement our frameworks. For this purpose, the different optimization algorithms may be re-visited for their efficient deployment.

**DREAMPAL Team**

# 3. Research Program

## 3.1. New Models for New Technologies

Over the past 25 years there have been several hardware-architecture generations dedicated to massively parallel computing. We have contributed to them in the past, and shall continue doing so in the Dreampal project. The three generations, chronologically ordered, are:

- Supercomputers from the 80s and 90s, based on massively parallel architectures that are more or less distributed (from the Cray T3D or Connection Machine CM2 to GRID 5000). Computer scientists have proposed methods and tools for mapping sequential algorithms to those parallel architectures in order to extract maximum power from them. We have contributed in this area in the past: http://www.lifl.fr/west/team.html.
- Parallelism pervades the chips! A new challenge appears: hardware/software co-design, in order to obtain performance gains by designing algorithms together with the parallel architectures of chips adapted to the algorithms. During the previous decade many studies, including ours in the Inria DaRT team, were dedicated to this type of co-design. DaRT has contributed to the development of the OMG MARTE standard (http://www.omgmarte.org) and to its implementation on several parallel platforms. Gaspard2, our implementation of this concept, was identified as one of the key software tools developed at Inria: http://www.inria.fr/en/centre/lille/research/platforms-and-flagship-software/flagship-software.
- The new challenge of the 2010s is, in our opinion, the integration of dynamic reconfiguration and massive parallelism. New circuits with high-density integration and supporting dynamic hardware reconfiguration have been proposed. In such architectures one can dynamically change the architecture while an algorithm is running on it. The Dynamic Partial Reconfiguration (DPR) feature offered by recent FPGA boards even allows, in theory, to generate optimized hardware at runtime, by adding, removing, and replacing components on a by-need basis. This integration of dynamic reconfiguration and massive parallelism induces a new degree of complexity, which we, as computer scientists, need to understand and deal with order to make possible the design of applications running on such architectures. This is the main challenge that we address in the Dreampal project. We note that we adress these problems as computer scientists; we do, however, collaborate with electronics specialists in order to benefit from their expertise in 3-D FPGAs.

Excerpt from the HiPEAC vision 2011/12

> *"The advent of 3D stacking enables higher levels of integration and reduced costs for off-chip communications. The overall complexity is managed due to the separation in different dies, independently designed."*

FPGAs (Field Programmable Gate Arrays) are configurable circuits that have emerged as a privileged target platform for intensive signal processing applications. FPGAs take advantage of the latest technological developments in circuits. For example, the Virtex7 from Xilinx offers a 28-nanometer integration, which is only one or two generations behind the latest general-purpose processors. 3D-Stacked Integrated Circuits (3D SICs) consist of two or more conventional 2D circuits stacked on the top of each other and built into the same IC. Recently, 3D SICs have been released by Xilinx for the Virtex 7 FPGA family. 3D integration will vastly increase the integration capabilities of FPGA circuits. The convergence of massive parallelism and dynamic reconfiguration in inevitable: we believe it is one of the main challenges in computing for the current decade.

By incorporating the configuration and/or data/program memory on the top of the FPGA fabric, with fast and numerous connections between memory and elementary logic blocks (∼10000 connections between dies), it will be possible to obtain dynamically reconfigurable computing platforms with a very high reconfiguration rate. Such a rate was not possible before, due to the serial nature of the interface between the configuration memory and the FPGA fabric itself. The FPGA technology also enables massively parallel architectures due to the large number of programmable logic fabrics available on the chip. For instance, Xilinx demonstrated 3600 8-bit picoBlaze softcore processors running simultaneously on the Virtex-7 2000T FPGA. For specific applications, picoBlaze can be replaced by specialized hardware accelerators or other IPs (Intellectual Property) components. This opens the possibility of creating massively parallel IP-based machines.

## 3.2. Multi-softcore on 3D FPGA

From the 2010 Xilinx white paper on FPGAs:

> *"Unlike a processor, in which architecture of the ALU is fixed and designed in a general-purpose manner to execute various operations, the CLBs (configurable logic blocks ) can be programmed with just the operations needed by the application... The FPGA architecture provides the flexibility to create a massive array of application-specific ALUs..The new solution enables high-bandwidth connectivity between multiple die by providing a much greater number of connections... enabling the integration of massive quantities of interconnect logic resources within a single package"*

Softcore processors are processors implemented using hardware synthesis. Proprietary solutions include PicoBlaze, MicroBlaze, Nios, and Nios II; open-source solutions include Leon, OpenRisk, and FC16. The choice is wide and many new solutions emerge, including multi-softcore implementations on FPGAs. An alternative to softcores are hardware accelerators on FPGAs, which are dedicated circuits that are an order of magnitude faster than softcores. Between these two approaches, there are other various approaches that connect IPs to softcores, in which, the processor's machine-code language is extended, and IP invocations become new instructions. We envisage a new class of softcores (we call them reflective softcores [0]), where almost everything is implemented in IPs; only the control flow is assigned to the softcore itself. The partial dynamic reconfiguration of next-generation FPGAs makes such dynamic IP management possible in practice. We believe that efficient reflective softcores on the new 3D-FPGAs should be as small as possible: low-performance generic hardware components (ALU, registers, memory, I/O...) should be replaced by dedicated high-performance IPs.

We are developing a sofcore processor called HoMade (http://www.lifl.fr/~dekeyser/Homade) following these ideas.

In the multi-reflective softcores that we develop, some softcores will be slaves and others will be masters. Massively parallel dynamically reconfigurable architectures of softcores can thus be envisaged. This requires, additionally, a parallel management of the partial dynamic reconfiguration system. This can be done, for example, on a given subset of softcores: a massively parallel reconfiguration will replace the current replication of a given IP with the replication of a new IP. Thanks to the new 3D-FPGAs this task can be performed efficiently and in parallel using the large number of 3D communication links (Through-Silicon-Vias). Our roadmap for HoMade is to evolve towards this multi-reflective softcore model.

## 3.3. When Hardware Meets Software

HIPEAC vision 2011/12: *"The number of cores and instruction set extensions increases with every new generation, requiring changes in the software to effectively exploit the new features."*

[0] Hereafter, by reflective system, we mean a system that is able to modify its own structure and behaviour while it is running. A reflective softcore thus dynamically adds, removes, and replaces IPs in the application running on it, and is able to dynamically modify its own program memory, thereby dynamically altering the program it is executing.

When the new massively parallel dynamically reconfigurable architectures become reality users will need languages for programming software applications on them. The languages will be themselves dynamic and parallel, in order to reflect and to fully exploit the dynamicity and parallelism of the architectures. Thus, developers will be able to invoke reconfiguration and call parallel instructions in their programs. This expressiveness comes with a cost, however, because new classes of bugs can be induced by the interaction between dynamic reconfiguration and parallelism; for example, deadlocks due to waiting for output from an IP that does not exist any more due to a reconfiguration. The detection and elimination of such bugs before deployment is paramount for cost-effectiveness and safety reasons.

Thus, we shall build an environment for developing software on parallel, dynamically reconfigurable architectures that will include languages and adequate formal analyses and verification tools for them, in addition to more traditional tools (emulators, compilers, etc). To this end we shall be using formal-semantics frameworks associated with easy-to-use formal verification tools in order to formally define our languages of interest and allow users to formally verify their programs. The K semantic framework (http://k-framework.org), developed jointly by Univs. Urbana Champaign, USA, and Iasi, Romania) is one such framework, which is mature enough (it has allowed defining a formal semantics of the largest subset of the C language to date, as well as many other languages from essentially all programming paradigms) and is familiar to us from previous work. In K, one can rapidly prototype a language definition and try several versions of the syntax and semantics of instructions. This is important in our project, where the proposed programming languages (in particular, the HoMade assembly language) will go through several versions before being stabilized. Moreover, once a language is defined in K one gets an interpreter of the language and one gains access to formal verification tools for free. We are also developing new analysis verification tools for K (in collaboration with the K team), which will be adapted and used in the Dreampal project.

**FUN Project-Team**

# 3. Research Program

## 3.1. Introduction

The research area of FUN research group is represented in Figure 1 . FUN research group will address every item of Figure 1 starting from the highest level of the figure, *i.e.* in area of homogeneous FUNs to the lowest one. Going down brings more applications and more issues to solve. Results achieved in the upper levels can be re-used in the lower ones. Current networks encountered nowadays are the ones at the higher level, without any interaction between them. In addition, solutions provided for such networks are rarely directly applicable in realistic networks because of the impact of the wireless medium.

FUN research group intends to fill the scientific gap and extend research performed in the area of wireless sensor and actor networks and RFID systems in two directions that are complementary and should be performed in parallel:

- **From theory to experimentation and reciprocally** On one hand, FUN research group intends to investigate new self-organization techniques for these future networks that take into account realistic parameters, emphasizing experimentation and considering mobility.
- **Towards heterogeneous FUNs** On the other hand, FUN research group intends to investigate techniques to allow heterogeneous FUNs to work together in a transparent way for the user. Indeed, new applications integrating several of these components are very much in demand (i.e. smart building) and thus these different technologies need to cooperate.

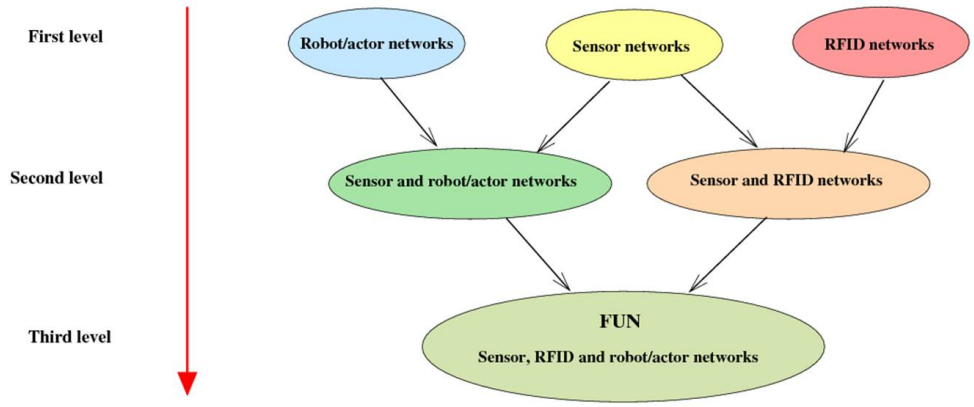


*Figure 1. Panorama of FUN.*

## 3.2. From theory to experimentation and reciprocally

Nowadays, even if some powerful and efficient propositions arise in the literature for each of these networks, very few are validated by experimentations. And even when this is the case, no lesson is learnt from it to improve the algorithms. FUN research group needs to study the limits of current assumptions in realistic and mobile environments.

Solutions provided by the FUN research group will mainly be algorithmic. These solutions will first be studied theoretically, principally by using stochastic geometry (like in [47]) or self-stabilization [49] tools in order to derive algorithm behavior in ideal environment. Theory is not an end in itself but only a tool to help in the characterization of the solution in the ideal world. For instance, stochastic geometry will allow quantifying changes in neighborhood or number of hops in a routing path. Self-stabilization will allow measuring stabilization times.

Those same solutions will then be confronted to realistic environments and their 'real' behavior will be analyzed and compared to the expected ones. Comparing theory, simulation and experimentation will allow will allow the influence of a realistic environment be better measured. From this and from the analysis of the information really available for nodes, FUN research group will investigate some means either to counterbalance these effects or to take advantage of them. New solutions provided by the FUN research group will take into consideration the vagaries of a realistic wireless environment and the node mobility. New protocols will take as inputs environmental data (as signal strength or node velocity/position, etc) and node characteristics (the node may have the ability to move in a controlled way) when available. FUN research group will thus adopt a **cross-layered** approach between hardware, physical environment, application requirements, self-organizing and routing techniques. For instance, FUN research group will study how the controlled node mobility can be exploited to enhance the network performance at lowest cost.

Solutions will follow the building process presented by Figure 2 . Propositions will be analyzed not only theoretically and by simulation but also by experimentation to observe the impact of the realistic medium on the behavior of the algorithms. These observations should lead to the derivation of cross-layered models. Experimentation feedbacks will be re-injected in solution design in order to propose algorithms that best fit the environment, and so on till getting satisfactory behavior in both small and large scale environments. All this should be done in such a way that the resulting propositions fit the hardware characteristics (low memory, CPU and energy capacity) and easy to deploy to allow their use by non experts. Since solutions should take into account application requirements as well as hardware characteristics and environment, solutions should be generic enough and then able to self-configure to adapt their environment settings.

In order to achieve this experimental environments, the FUN research group will maintain its strong activity on platform deployment such as SensLAB [52], FIT [25] and Aspire [44]. Next steps will be to experiment not only on testbeds but also on real use cases. These latter will be given through different collaborations.

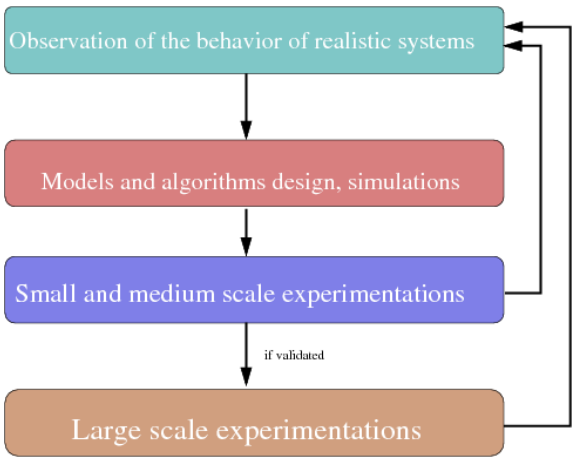


*Figure 2. Methodology applied in the FUN research group.*

**FUN research group will investigate self-organizing techniques for FUNs by providing cross-layered solutions that integrate in their design the adaptability to the realistic environment features. Every solution will be validated with regards to specific application requirements and in realistic environments.**

**Facing the medium instability.** The behavior of wireless propagation is very depending of the surrounding environment (in-door vs outdoor, night vs day, etc) and is very instable. Many experiments in different environment settings should be conducted. Experiment platforms such as SensLAB, FIT, our wifiBot as robots and actuators and our RFID devices will be used offering ways to experiment easily and quickly in different environments but might not be sufficient to experiment every environment.
**Adaptability and flexibility.** Since from one application to another one, requirements and environments are different, solutions provided by FUN research group should be **generic** enough and **self-adapt** to their environment. Algorithm design and validation should also take into account the targeted applications brought for instance by our industrial partners like Etineo. All solution designs should keep in mind the devices constrained capacities. Solutions should consume low resources in terms of memory, processor and energy to provide better performances and scale. All should be self-adaptive.

FUN research group will try to take advantage of some observed features that could first be seen as drawbacks. For instance, the broadcast nature of wireless networks is first an inconvenient since the use of a link between two nodes inhibits every other communication in the same transmission area. But algorithms should exploit that feature to derive new behaviors and a node blocked by another transmission should overhear it to get more information and maybe to limit the overall information to store in the network or overhead communication.

## 3.3. Towards unified heterogeneous FUNs

The second main direction to be followed by the FUN research group is to merge networks from the upper layer in Fig. 1 into networks from the lowest level. Indeed, nowadays, these networks are still considered as separated issues. But considering mixed networks bring new opportunities. Indeed, robots can deploy, replace, compensate sensor nodes. They also can collect periodically their data, which avoids some long and multi-hop communications between sensor nodes and thus preserving their resources. Robots can also perform many additional tasks to enhance network performance like positioning themselves on strategic points to ensure area coverage or reduce routing path lengths. Similarly, coupling sensors and RFID tags also bring new opportunities that are more and more in-demand from the industrial side. Indeed, an RFID reader may be a sensor in a wireless sensor network and data hold by RFID tags and collected by readers might need to be reported to a sink. This will allow new applications and possibilities such as the localization of a tagged object in an environment be covered by sensors.

When at last all components are gathered, this leads us to a new era in which every object is autonomous. Let's consider for instance a smart home equipped with sensors and RFID reader. An event triggered by a sensor (*i.e.* an increase of the temperature) or a RFID reader (*i.e.* detection of a tag hold by a person) will trigger actions from actuators (*i.e.* lowering of stores, door opening). Possibilities are huge. But with all these new opportunities come new technological issues with other constraints. Every entity is considered as an object possibly mobile which should be dynamically identified and controlled. To support this dynamics, protocols should be localized and distributed. Model derived from experiment observations should be unified to fit all these classes of devices.

**FUN research group will investigate new protocols and communication paradigms that allow the technologies to be transparently merged. Objects and events might interconnect while respecting on-going standards and building an autonomic and smart network while being compliant with hardware resources and environment.**

Technologies such as wireless sensors, wireless robots/actuators and RFID tags/ readers, although presenting many common points are still part of different disciplines that have evolved in parallel ways. Every branch is at different maturity levels and has developed its own standards. Nevertheless, making all these devices part of a single unified network leverages technological issues (partly addressed in the former objective) but also regarding to on-going standards and data formatting. FUN research group will have to study current standards

of every area in order to propose compliant solutions. Such works have been initiated in the POPS research group in the framework of the FP7 ASPIRE project. Members of FUN research group intend to continue and enlarge these works.

Todays' EPCGlobal compliant RFID readers must comply to some rules and be configurable through an ALE (Application Level Event) [42]. While a fixed and connected RFID reader is easily configurable, configuring remotely a mobile RFID reader might be very difficult since it implies to first locate it and then send configuration data through a wireless dynamic network. FUN research group will investigate some tools that make the configuration easy and transparent for the user. This remote configuration of mobile readers through the network should consider application requirements and network and reader characteristics to choose the best trade-off relative to the software part embedded in the reader. The biggest part embedded, the lowest bandwidth overhead (data can be filtered and aggregated in the reader) and the greater mobility (readers are still fully operational even when disconnected) but the more difficult to set up and the more powerful readers. All these aspects will be studied within the FUN research group.

# LINKS Team (section vide)

**MAGNET Team**

# 3. Research Program

## 3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data. We consider information networks in which the data are vectorial data and texts. We model such information networks as (multiple) (hyper)graphs wherein nodes correspond to entities (documents, spans of text, users, ...) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship, ...). Our main research goal is to propose new learning algorithms to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Hence, we will investigate new learning algorithms for node clustering and node classification, link classification and link prediction. Also, we will search for the best hidden graph structure to be generated for solving a given learning task. We will base our research on generative models for graphs, on machine learning for graphs and on machine learning for texts. The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. An additional challenge will be to design scalable methods for large information networks. Hence, we will explore how sampling and randomization can be used in new machine learning algorithms. Also, active machine learning algorithms for graphs will be investigated.

On the first hand we want to design machine learning algorithms on graphs to solve problems in networks of texts and documents in natural language. The main originality of this research is to consider and take advantage of the setting of networked data exploiting the relationships between different data entities and, overall, the graph topology. On the second hand, in a concomitant way, we want to develop prediction models for graph-like data. This includes prediction, ranking and classification of links and nodes in an on-line or batch setting. The two objectives are intertwined, enrich each other and raise important scientific questions we want to focus on. Our research proposal is organized according to the following questions:

1. How to go beyond vectorial classification models in natural language oriented tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create network from observations of information diffusion processes?
3. How to design methods able to achieve very good predictive accuracy without giving up on scalability?
4. How to go beyond strict node homophilic/similarity assumptions in graph-based learning methods?

## 3.2. Beyond vectorial models for NLP

One of our overall research objectives is to derive graph-based machine learning algorithms for natural language and text information extraction tasks. This section discusses the motivations behind the use of graph-based ML approaches for these tasks, the main challenges associated with it, as well as some concrete projects. Some of the challenges go beyond NLP problems and will be further developed in the next sections. An interesting aspect of the project is that we anticipate some important cross-fertilizations between NLP and ML graph-based techniques, with NLP not only benefiting from but also pushing ML graph-based approaches into new directions.

Motivations for resorting to graph-based algorithms for texts are at least threefold. First, online texts are organized in networks. With the advent of the web, and the development of forums, blogs, and micro-blogging, and other forms of social media, text productions have become strongly connected. Thus, documents on the web are linked through hyperlinks, forum posts and emails are organized in threads, tweets can be retweeted, etc. Additional connections can be made through users connections (co-authorship, friendship, follower, etc.). Interestingly, NLP research has been rather slow in coming to terms with this situation, and most work still focus on document-based or sentence-based predictions (wherein inter-document or inter-sentence structure is not exploited). Furthermore, several multi-document tasks exist in NLP (such as multi-document summarization and cross-document coreference resolution), but most existing work typically ignore document boundaries and simply apply a document-based approach, therefore failing to take advantage of the multi-document dimension [26], [28].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NL tasks often involve discovering a relational structure over a set of text spans (words, phrases, clauses, sentences, etc.). Furthermore, the *input* of numerous NLP tasks is also a graph; indeed, most end-to-end NLP systems are conceived as pipelines wherein the output of one processor is in the input of the next. For instance, several tasks take POS tagged sequences or dependency trees as input. But this structured input is often converted to a vectorial form, which inevitably involves a loss of information.

Finally, graph-based representations and learning methods in principle appear to address some core problems faced by NLP, such as the fact that textual data are typically not independent and identically distributed, they often live on a manifold, they involve very high dimensionality, and their annotations is costly and scarce. As such, graph-based methods represent an interesting alternative, or at least complement, to structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. While structured output approaches are able to model local dependencies (e.g., between neighboring words or sentences), they cannot efficiently capture long distance dependencies, like forcing a particular $n$-gram to receive the same labeling in different sentences or documents for instance. On the other hand, graph-based models provide a natural way to capture global properties of the data through the exploitation of walks and neighborhood in graphs. Graph-based methods, like label propagation, have also been shown to be very effective in semi-supervised settings, and have already given some positive results on a few NLP tasks [9], [30].

Given the above motivations, our first line of research will be to investigate how one can leverage an underlying network structure (e.g., hyperlinks, user links) between documents, or text spans in general, to enhance prediction performances for several NL tasks. We think that a "network effect", similar to the one that took place in Information Retrieval (with the Page Rank algorithm), could also positively impact NLP research. A few recent papers have already opened the way, for instance in attempting to exploit Twitter follower graph to improve sentiment classification [29].

Part of the challenge in this work will be to investigate how adequately and efficiently one can model these problems as instances of more general graph-based problems, such as node clustering/classification or link prediction discussed in the next sections. In a few cases, like text classification or sentiment analysis, graph modeling appears to be straightforward: nodes correspond to texts (and potentially users), and edges are given by relationships like hyperlinks, co-authorship, frienship, or thread membership. Unfortunately, modeling NL problems as networks is not always that obvious. From the one hand, the right level of representation will probably vary depending on the task at hand: the nodes will be sentences, phrases, words, etc. From the other hand, the underlying graph will typically not be given a priori, which in turn raises the question of how we construct it. Of course, there are various well-known ways to obtain similarity measures between text contents (and its associated vectorial data), and graphs can be easily constructed from those combined with some sparsification method. But we would like our similarity to be tailored to the task objective. An additional problem with many NLP problems is that features typically live in different types of spaces (e.g., binary, discrete, continuous). A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [9], [33]. We identify the issue of adaptative graph construction as an important scientific challenge for machine learning on graphs in general, and we will discuss it further in Section 3.3 .

As noted above, many NLP tasks have been recast as structure prediction problems, allowing to capture (some of the) output dependencies. Structure prediction can be viewed as (set of) link prediction with global loss or dependencies, which means that graph-based learning methods can handle (at least, approximately) output prediction dependencies, and they can in principle capture additional more global dependencies given the right graph structure. How to best combine structured output and graph-based ML approaches is another challenge that we intend to address. We will initially investigate this question within a semi-supervised context, concentrating on graph based regularization and graph propagation methods. Within such approaches, labels are typically binary or they correspond to small finite set. Our objective is to explore how one propagates an exponential number of *structured labels* (like a sequence of tags or a dependency tree) through graphs. Recent attempts at blending structured output models with graph-based models are investigated in [30], [17]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specificied tag sequence or tree) and use the graph structure to complete the output structure. This last question is very relevant to NL problems where human annotations are costly; being able to learn from partial annotations could therefore allow for more targeted annotations and in turn reduced costs [18].

The NL tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisonned in both document and cross-document settings, although we expect to exploit inter-document links either way. Choices for these particular tasks is guided by the fact that are still open problems for the NLP community, they potentially have a high impact for industrial applications (like information retrieval, question answering, etc.), and we already have some expertise on these tasks in the team. As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [32].

We have already initiated some work on the coreference resolution problem in the context of ML graph-based approaches. We cast this problem as a spectral clustering problem. Given than features can be numerical or nominal, the definition of a good similarity measure between entities is not straightforward. As a first solution, we consider only numerical attributes to build a $k$-nn graph of mentions so that graph clustering methods can be applied. Nominal attributes and relations are introduced by means of soft constraints on this clustering. Constraints can have various forms and have the ability of going beyond homophily assumptions, taking into account for instance dissimilarity relationships. From this setting we derive new graph-based learning methods. We propose to study the modification of graph clustering and spectral embeddings to satisfy certain constraints induced by several types of supervision: (i) nodes belong to the same group or to different groups, and (ii) some groups are fully known while others have to be discovered. This semi-supervised graph clustering problem is studied in a batch and transductive setting. But interesting extensions can be investigated in an online and active setting.

## 3.3. Adaptive Graph Construction

In most applications, edge weights are computed through a complex data-modeling process and convey crucially important information for classifying nodes, which makes it possible to infer information related to each data sample even exploiting the graph topology solely. In fact, a widespread approach to the solution of several classification problems is representing the data through an undirected weighted graph in which edge weights quantify the similarity between data points. This technique for coding input data has been applied to several domains, including classification of genomic data ( [27]), face recognition ( [16]), and text categorization ( [21]).

In some cases, the full adjacency matrix is generated by employing suitable similarity functions chosen through a deep understanding of the problem structure. For example TF-IDF representation of documents, the affinity between pairs of samples is often estimated through the cosine measure or the $\chi^2$ distance. After the generation of the full adjacency matrix, the second phase for obtaining the final graph consists in an edge sparsification/reweighting operation. Some of the edges of the clique obtained in the first step are pruned and the remaining ones can be reweighted to meet the specific requirements of the given classification problem. Constructing a graph with these methods obviously entails various kinds of loss of information. However, in

problems like node classification, the use of graphs generated from several datasets can lead to an improvement in accuracy performance ( [34], [10], [11]). Hence, the transformation of a dataset into a graph may, at least in some cases, partially remove various kinds of irregularities present in the original datasets, while keeping some of the most useful information for classifying the data samples. Moreover, it is often possible to accomplish classification tasks on the obtained graph using a running time remarkably lower than is needed by algorithms exploiting the initial datasets, and a suitable sparse graph representation can be seen as a compressed version of the original data. This holds even when input data are provided in a online/stream fashion, so that the resulting graph evolves over time.

In this project we will address the problem of adaptive graph construction towards several directions. One is the question of choosing the best similarity measure given the objective learning task. This question is related to the question of similarity learning ( [12]) which has not been considered in the context of graph based learning. In the context of structured prediction, we will develop approaches where output structures are organized in graphs whose similarity is given by top-$k$ outcomes of greedy algorithms.

A different way we envision adaptative graph construction is in the context of semi-supervised learning. Partial supervision can take various forms and an interesting and original setting is governed by two currently studied applications: detection of brain anomaly from connectome data and polls recommendation in marketing. Indeed, for these two applications, a partial knowledge of the information diffusion process can be observed while the network is unknown or only partially known. An objective is to construct (or complete) the network structure from some local diffusion information. The problem can be formalized as a graph construction problem from partially observed diffusion processes. It has been studied very recently in [23]. In our case, the originality comes either from the existence of different sources of observations or from the large impact of node contents in the network.

We will study how to combine graphs defined by networked data and graphs built from flat data to solve a given task. This is of major importance for information networks because, as said above, we will have to deal with multiple relations between entities (texts, spans of texts, ...) and also use textual data and vectorial data. We have started to work on combining graphs in a semi supervised setting for node classification problems along the PhD thesis of T. Ricatte. Future work include combination geared by semi-supervision on link prediction tasks. This can be studied in an active learning setting. But one important issue is to design scalable approaches, thus to exploit locality given by the network. Doing this we address another objective to build non uniformly parameterized combinations.

## 3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provides a rich representation of data. Often enough the data is only partially available and the graph representation is very helpful in predicting the unobserved elements. We are interested in problems where the complete structure of the graph needs to be recover and only a fraction of the links is observed. The link prediction problem falls into this category. We are also interested in the recommendation and link classification problems which can be seen as graphs where the structure is complete but some labels on the links (weights or signs) are missing. Finally we are also interested in labelling the nodes of the graph, with class or cluster memberships or with a real value, provided that we have (some information about) the labels for some of the nodes.

The semi-supervised framework will be also considered. A midterm research plan is to study how graph-based regularization models help for structured prediction problems. This question will be studied in the context of NLP tasks, as noted in Section 3.2 , but we also plan to develop original machine learning algorithms that have a more general applicability. Inputs are networks whose nodes (texts) have to be labeled by structures. We assume that structures lie in some manifold and we want to study how labels can propagate in the network. One approach is to find smooth labeling function corresponding to an harmonic function on both manifolds in input and output. We also plan to extend our results on spectral clustering with must-link and cannot-link constraints in two directions. We have proposed a batch method with an optimization problem based on an adaptive spectral embedding with respects to constraints. We want to extend this approach to an on-line and

active setting where a flow of graphs (each one is a document) is given as input. In the case of large graphs, we also consider the case where partial supervision consists in the knowledge of few clusters.

Scalability is one of the main issue in the design of new prediction algorithms working on networked data. It has gained more and more importance in recent years, because of the growing size of the most popular networked data that are now used by millions of people. In such contexts, learning algorithms whose computation time scales quadratically, or slower, in the number of considered data objects (usually nodes or vertices, depending on the given task) should be considered impractical.

These observations lead to the idea of using graph sparsification techniques in order to work on a part of the original network for getting results that can be easily extended and used for the whole original input. A sparsified version of the original graph can often be seen as a subset of the initial input, i.e. a suitably selected input subgraph which forms the training set (or, more in general, it is included in the training set). This holds even for the active setting.

A simple example could be to find a spanning tree of the input graph, possibly using randomization techniques, with properties such that we are allowed to obtain interesting results for the initial graph dataset. We have started to explore this research direction for instance in [31]. This approach leaves us with the problem of choosing a good spanning tree, taking into account that the setting could be adversarial (e.g, in the online case the presentation and the assignment of the labels are both arbitrary). A suitable use of the randomization power becomes therefore remarkably significant. Moreover, it is interesting to observe that running a prediction algorithm on a sparsified version of the input dataset allows the parallelization of prediction tasks. In fact, given a prediction task for a networked dataset, in a preliminary phase one could run a randomized graph sparsification method in parallel on different machines. For example, in the case of the spanning tree use, one could then draw several spanning trees at the same time, each on a different computer. This way it is possible to simultaneously run different prediction experiments on the same task and aggregating the obtained results at the end, with several methods (e.g. simply by majority vote) in order to increase the robustness and accuracy predictions.

At the level of the mathematical foundations, the key issue to be addressed in the study of (large-scale) random networks also concerns the segmentation of network data into sets of independent and identically distributed observations. If we identify the data sample with the whole network, as it has been done in previous approaches [22], we typically end up with a set of observations (such as nodes or edges) which are highly interdependent and hence overly violate the classic i.i.d. assumption. In this case, the data scale can be so large and the range of correlations can be so wide, that the cost of taking into account the whole data and their dependencies is typically prohibitive. On the contrary, if we focus instead on a set of subgraphs independently drawn from a (virtually infinite) target network, we come up with a set of independent and identically distributed observations—namely the subgraphs themselves, where subgraph sampling is the underlying ergodic process [13]. Such an approach is one principled direction for giving novel statistical foundations to random network modeling. At the same time, because one shifts the focus from the whole network to a set of subgraphs, complexity issues can be restricted to the number of subgraphs and their size. The latter quantities can be controlled much more easily than the overall network size and dependence relationships, thus allowing to tackle scalability challenges through a radically redesigned approach.

We intend to develop new learning models for link prediction problems. We have already proposed a conditional model in [20] with statistics based on Fiedler values computed on small subgraphs. We will investigate the use of such a conditional model for link prediction. We will also extend the conditional probabilistic models to the case of graphs with textual and vectorial data by defining joint conditional models. Indeed, an important challenge for information networks is to introduce node contents in link ranking and link prediction methods that usually rely solely on the graph structure. A first step in this direction was already proposed in [19] where we learn a mapping of node content to a new representation constrained by the existing link structure and applied it for link recommendation. This approach opens a different view on recommendation by means of link ranking problems for which we think that non parametric approaches should be fruitful.

Regarding link classification problems, we plan to devise a whole family of active learning strategies, which could be based on spanning trees or sparse input subgraphs, that exploit randomization and the structure of the graph in order to offset the adversarial label assignment. We expect these active strategies to exhibit good accuracies with a remarkably small number of queried edges, where passive learning methods typically break down. The theoretical findings can be supported by experiments run on both synthetic and real-world (Slashdot, Epinions, Wikipedia, and others) datasets.

We are interested in studying generative models for graph labeling, exploiting the results obtained in p-stochastic model for link classification (investigated in [15]) and statistical model for node label assignment which can be related to tree-structured Markov random fields [24].

In developing our algorithms, we focus on providing theoretical guarantees on prediction accuracy and, at the same time, on computational efficiency. The development of methods that simultaneously guarantee optimal accuracy and computational efficiency is a very challenging goal. In fact, the accuracy of most methods in the literature is not rigorously analyzed from a theoretical point of view. Likewise, tight time and space complexity bounds are not generally provided. This contrasts with the need to manage extremely large relational datasets like, e.g., snapshots of the World Wide Web.

## 3.5. Beyond Homophilic Relationships

In many cases, the algorithms devised for solving node classification problems are driven by the following assumption: linked entities tend to be assigned to the same class. This assumption, in the context of social networks, is known as homophily ( [14], [25]) and involves ties of every type, including friendship, work, marriage, age, gender, and so on. In social networks, homophily naturally implies that a set of individuals can be parted into subpopulations that are more cohesive. In fact, the presence of homogeneous groups sharing interests is one of the most significant reasons for affinity among interconnected individuals, which suggests that, in spite of its simplicity, this principle turns out to be very powerful for node classification problems in general networks.

Recently, however, researchers have started to consider networked data where connections may also carry a negative meaning. For instance, disapproval or distrust in social networks, negative endorsements on the Web. Concrete examples are provided by certain types of online social networks. Users of Slashdot can tag other users as friends or foes. Similarly, users of Epinions can give positive or negative ratings not only to products but also to other users. Even in the social network of Wikipedia administrators, votes cast by an admin in favor or against the promotion of another admin can be viewed as positive or negative links. More examples of signed links are found in other domains, such as the excitatory or inhibitory interactions between genes or gene products in biological networks.

Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical object, called signed graph, has an unexpectedly rich additional complexity. For example, the spectral properties of signed graphs, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of their unsigned counterparts. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting the sign of links in a given graph. In online social networks, this may be viewed as a form of sentiment analysis, since we would like to semantically categorize the relationship between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [4]. A bipartite hyperedge connects two ends which are disjoint subsets of nodes. Bipartite hyperedges is a way to relate two collections of (possibly heterogeneous) entities represented by nodes. In the NLP setting, while hyperedges can be used to model bags of words, bipartite hyperedges are associated with relationships between bags of words. But each end of bipartite hyperedges is also a way to represent complex entities, gathering several attribute values (nodes) into hyperedges viewed as records. Our hypergraph notion naturally extends directed and undirected weighted graph. We have defined a spectral theory for this new class of hypergraphs and opened a way to smooth labeling on sets of nodes. The weighting

scheme permits to weight the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This is exactly that equilibrium condition that provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes. (Theoretical analysis of our hypergraphs exhibits tight relationships with signed graphs). Following this competition idea, bipartite hyperedges are like matches between two teams and examples of applications are team creation. The basic tasks in which we are interested in are hyperedge classification, hyperedge prediction, node weight prediction. Finally, hypergraphs also represent a way to summarize or compress large graphs in which there exists highly connected couples of (large) subsets of nodes.

To conclude, we plan to go beyond the homophilic bias from the algorithmic as well as from the modeling point of view. We will consider new kind of modeling and learning biases provided by graphs with negative weights (signed graphs) and hypergraphs. We will study their spectral properties, smoothness measures of (node or edge) labeling. Sampling and walking also need to be reconsidered. From the machine learning perspective, we will study edge and node labeling in batch and online settings. In connection with our main targeted applications, we will mainly consider unsupervised and semi-supervised situations. We think that allowing negative weights and advanced relationships on nodes will also lead to space efficient representations of graphs.

**MEPHYSTO Team**

# 3. Research Program

## 3.1. From statistical physics to continuum mechanics

Whereas numerical methods in nonlinear elasticity are well-developed and reliable, constitutive laws used for rubber in practice are phenomenological and generally not very precise. On the contrary, at the scale of the polymer-chain network, the physics of rubber is very precisely described by statistical physics. The main challenge in this field is to understand how to derive macroscopic constitutive laws for rubber-like materials from statistical physics.

At the continuum level, rubber is modelled by an energy $E$ defined as the integral over a domain $D$ of $\mathbb{R}^d$ of some energy density $W$ depending only locally on the gradient of the deformation $u$: $E(u) = \int_D W(\nabla u(x))dx$. At the microscopic level (say 100nm), rubber is a network of cross-linked and entangled polymer chains (each chain is made of a sequence of monomers). At this scale the physics of polymer chains is well-understood in terms of statistical mechanics: monomers thermally fluctuate according to the Boltzmann distribution [46]. The associated Hamiltonian of a network is typically given by a contribution of the polymer chains (using self-avoiding random bridges) and a contribution due to steric effects (rubber is packed and monomers are surrounded by an excluded volume). The main challenge is to understand how this statistical physics picture yields rubber elasticity. Treloar assumed in [56] that for a piece of rubber undergoing some macroscopic deformation, the cross-links do not fluctuate and follow the macroscopic deformation, whereas between two cross-links, the chains fluctuate. This is the so-called affine assumption. Treloar's model is in rather good agreement with mechanical experiments in small deformation. In large deformation however, it overestimates the stress. A natural possibility to relax Treloar's model consists in relaxing the affine assumption while keeping the network description, which allows one to distinguish between different rubbers. This can be done by assuming that the deformation of the cross-links minimizes the free energy of the polymer chains, the deformation being fixed at the boundary of the macroscopic domain $D$. This gives rise to a "variational model". The analysis of the asymptotic behavior of this model as the typical length of a polymer chain vanishes has the same flavor as the homogenization theory of integral functionals in nonlinear elasticity (see [41], [52] in the periodic setting, and [42] in the random setting).

Our aim is to relate qualitatively and quantitatively the (precise but unpractical) statistical physics picture to explicit macroscopic constitutive laws that can be used for practical purposes.

In collaboration with R. Alicandro (Univ. Cassino, Italy) and M. Cicalese (Univ. Munich, Germany), A. Gloria analyzed in [1] the (asymptotic) $\Gamma$-convergence of the variational model for rubber, in the case when the polymer chain network is represented by some ergodic random graph. The easiest such graph is the Delaunay tessellation of a point set generated as follows: random hard spheres of some given radius $\rho$ are picked randomly until the domain is jammed (the so-called random parking measure of intensity $\rho$). With M. Penrose (Univ. Bath, UK), A. Gloria studied this random graph in this framework [6]. With P. Le Tallec (Mechanics department, Ecole polytechnique, France), M. Vidrascu (project-team REO, Inria Paris-Rocquencourt), and A. Gloria introduced and tested in [15] a numerical algorithm to approximate the homogenized energy density, and observed that this model compares well to rubber elasticity qualitatively.

These preliminary results show that the variational model has the potential to explain qualitatively and quantitatively how rubber elasticity emerges from polymer physics. In order to go further and obtain more quantitative results and rigorously justify the model, we have to address several questions of analysis, modelling, scientific computing, inverse problems, and physics.

## 3.2. Quantitative stochastic homogenization

Whereas the approximation of homogenized coefficients is an easy task in periodic homogenization, this is a highly nontrivial task for stochastic coefficients. This is in order to analyze numerical approximation methods of the homogenized coefficients that F. Otto (MPI for mathematics in the sciences, Leipzig, Germany) and A. Gloria obtained the first quantitative results in stochastic homogenization [4]. The development of a complete stochastic homogenization theory seems to be ripe for the analysis and constitutes the second major objective of this section.

In order to develop a quantitative theory of stochastic homogenization, one needs to quantitatively understand the corrector equation (3 ). Provided $A$ is stationary and ergodic, it is known that there exists a unique random field $\phi_\xi$ which is a distributional solution of (3 ) almost surely, such that $\nabla\phi_\xi$ is a stationary random field with bounded second moment $\langle|\nabla\phi_\xi|^2\rangle < \infty$, and with $\phi(0) = 0$. Soft arguments do not allow to prove that $\phi_\xi$ may be chosen stationary (this is wrong in dimension $d = 1$). In [4], [5] F. Otto and A. Gloria proved that, in the case of discrete elliptic equations with iid conductances, there exists a unique stationary corrector $\phi_\xi$ with vanishing expectation in dimension $d > 2$. Although it cannot be bounded, it has bounded finite moments of any order:

$$\langle|\phi_\xi|^q\rangle < \infty \text{ for all } q \geq 1. \tag{1}$$

They also proved that the variance of spatial averages of the energy density $(\xi + \nabla\phi_\xi) \cdot A(\xi + \nabla\phi_\xi)$ on balls of radius $R$ decays at the rate $R^{-d}$ of the central limit theorem. These are the *first optimal quantitative results* in stochastic homogenization.

The proof of these results, which is inspired by [53], is based on the insight that coefficients such as the Poisson random inclusions are special in the sense that the associated probability measure satisfies a spectral gap estimate. Combined with elliptic regularity theory, this spectral gap estimate quantifies ergodicity in stochastic homogenization. This systematic use of tools from statistical physics has opened the way to the quantitative study of stochastic homogenization problems, which we plan to fully develop.

## 3.3. Nonlinear Schrödinger equations

As well known, the (non)linear Schrödinger equation

$$\partial_t\varphi(t,x) = -\Delta\varphi(t,x) + \lambda V(x)\varphi(t,x) + g|\varphi|^2\varphi(t,x), \quad \varphi(0,x) = \varphi_0(x) \tag{2}$$

with coupling constants $g \in \mathbb{R}, \lambda \in \mathbb{R}_+$ and real potential V (possibly depending also on time) models many phenomena of physics.

When in the equation (5 ) above one sets $\lambda = 0, g \neq 0$, one obtains the nonlinear (focusing of defocusing) Schrödinger equation. It is used to model light propagation in optical fibers. In fact, it then takes the following form:

$$i\partial_z\varphi(t,z) = -\beta(z)\partial_t^2\varphi(t,z) + \gamma(z)|\varphi(t,z)|^2\varphi(z,t), \tag{3}$$

where $\beta$ and $\gamma$ are functions that characterize the physical properties of the fiber, $t$ is time and $z$ the position along the fiber. Several issues are of importance here. Two that will be investigated within the MEPHYSTO project are: the influence of a periodic modulation of the fiber parameters $\beta$ and $\gamma$ and the generation of so-called “rogue waves” (which are solutions of unusually high amplitude) in such systems.

If $g = 0, \lambda \neq 0$, $V$ is a random potential, and $\varphi_0$ is deterministic, this is the standard random Schrödinger equation describing for example the motion of an electron in a random medium. The main issue in this setting is the determination of the regime of Anderson localization, a property characterized by the boundedness in time of the second moment $\int x^2 |\varphi(t,x)|^2 dx$ of the solution. If this second moment remains bounded in time, the solution is said to be localized. Whereas it is known that the solution is localized in one dimension for all (suitable) initial data, both localized and delocalized solutions exist in dimension 3 and it remains a major open problem today to prove this, cf. [44].

If now $g \neq 0, \lambda \neq 0$ and $V$ is still random, but $|g| \ll \lambda$, a natural question is whether, and in which regime, one-dimensional Anderson localization perdures. Indeed, Anderson localization can be affected by the presence of the nonlinearity, which corresponds to an interaction between the electrons or atoms. Much numerical and some analytical work has been done on this issue (see for example [47] for a recent work at PhLAM, Laser physics department, Univ. Lille 1), but many questions remain, notably on the dependence of the result on the initial conditions, which, in a nonlinear system, may be very complex. The cold atoms team of PhLAM (Garreau-Szriftgiser) is currently setting up an experiment to analyze the effect of the interactions in a Bose-Einstein condensate on a closely related localization phenomenon called "dynamical localization", in the kicked rotor, see below.

## 3.4. Dynamical localization and kicked rotors

The kicked rotor is a unitary discrete time dynamics proposed in the seventies in the context of studies on quantum chaos, and used recently as a "quantum simulator" for the Anderson model. It is a quantum equivalent of the standard map and is obtained by integrating a time-dependent linear Schrödinger equation with a time-periodic, very singular (delta comb) potential. It continues to pose considerable mathematical challenges, in particular the so-called "quantum suppression of classical chaos" in the presence of a strong potential, which remains an open problem from the mathematical point of view. It can be rephrased as follows: show that the $H^1$ norm of the solution is uniformly bounded in time (see [36] for more background). In more recent years, the question has arisen how the behavior of this system would change in the presence of a nonlinear term in the Schrödinger equation.

This problem displays both numerical and analytical challenges, in particular because of the difficulty to obtain long time simulations of the system and because of the presence of instabilities due to the nonlinearity. Preliminary theoretical results motivate some conjectures on the behavior of these systems, that we plan to validate empirically in a first step. Indeed, reliable long-time simulations of the system should allow us to get more insight into the behavior of the exact solutions in the unstable cases. One of the main difficulties for the numerical simulation is the intrinsic instability of the system, which magnifies quite rapidly the numerical error due to machine precision. This requires the use of multiprecision techniques in order to handle reasonably long times, even for moderate nonlinearities, and of the transparent boundary conditions recently introduced by members of the former SIMPAF project-team.

**MINT Project-Team**

# 3. Research Program

## 3.1. Human-Computer Interaction

The scientific approach that we follow considers user interfaces as means, not an end: our focus is not on interfaces, but on interaction considered as a phenomenon between a person and a computing system [46]. We *observe* this phenomenon in order to understand it, i.e. *describe* it and possibly *explain* it, and we look for ways to significantly *improve* it. HCI borrows its methods from various disciplines, including Computer Science, Psychology, Ethnography and Design. Participatory design methods can help determine users' problems and needs and generate new ideas, for example [52]. Rapid and iterative prototyping techniques allow to decide between alternative solutions [47]. Controlled studies based on experimental or quasi-experimental designs can then be used to evaluate the chosen solutions [54]. One of the main difficulties of HCI research is the doubly changing nature of the studied phenomenon: people can both adapt to the system and at the same time adapt it for their own specific purposes [51]. As these purposes are usually difficult to anticipate, we regularly *create* new versions of the systems we develop to take into account new theoretical and empirical knowledge. We also seek to *integrate* this knowledge in theoretical frameworks and software tools to disseminate it.

## 3.2. Numerical and algorithmic real-time gesture analysis

Whatever is the interface, user provides some curves, defined over time, to the application. The curves constitute a gesture (positionnal information, yet may also include pressure). Depending on the hardware input, such a gesture may be either continuous (e.g. data-glove), or not (e.g. multi-touch screens). User gesture can be multi-variate (several fingers captured at the same time, combined into a single gesture, possibly involving two hands, maybe more in the context of co-located collaboration), that we would like, at higher-level, to be structured in time from simple elements in order to create specific command combinations. One of the scientific foundations of the research project is an algorithmic and numerical study of gesture, which we classify into three points:

- *clustering*, that takes into account intrinsic structure of gesture (multi-finger/multi-hand/multi-user aspects), as a lower-level treatment for further use of gesture by application;
- *recognition*, that identifies some semantic from gesture, that can be further used for application control (as command input). We consider in this topic multi-finger gestures, two-handed gestures, gesture for collaboration, on which very few has been done so far to our knowledge. On the contrary, in the case of single gesture case (i.e. one single point moving over time in a continuous manner), numerous studies have been proposed in the current literature, and interestingly, are of interest in several communities: HMM [55], Dynamic Time Warping [57] are well-known methods for computer-vision community, and hand-writing recognition. In the computer graphics community, statistical classification using geometric descriptors has previously been used [53]; in the Human-Computer interaction community, some simple (and easy to implement) methods have been proposed, that provide a very good compromise between technical complexity and practical efficiency [56].
- *mapping to application*, that studies how to link gesture inputs to application. This ranges from transfer function that is classically involved in pointing tasks [48], to the question to know how to link gesture analysis and recognition to the algorithmic of application content, with specific reference examples.

We ground our activity on the topic of numerical algorithm, expertise that has been previously achieved by team members in the physical simulation community (within which we think that aspects such as elastic deformation energies evaluation, simulation of rigid bodies composed of unstructured particles, constraint-based animation... will bring up interesting and novel insights within HCI community).

## 3.3. Design and control of haptic devices

Our scientific approach in the design and control of haptic devices is focused on the interaction forces between the user and the device. We search of controlling them, as precisely as possible. This leads to different designs compared to other systems which control the deformation instead. The research is carried out in three steps:

- *identification:* we measure the forces which occur during the exploration of a real object, for example a surface for tactile purposes. We then analyze the record to deduce the key components – *on user's point of view* – of the interaction forces.
- *design:* we propose new designs of haptic devices, based on our knowledge of the key components of the interaction forces. For example, coupling tactile and kinesthetic feedback is a promising design to achieve a good simulation of actual surfaces. Our goal is to find designs which leads to compact systems, and which can stand close to a computer in a desktop environment.
- *control:* we have to supply the device with the good electrical conditions to accurately output the good forces.

**MODAL Project-Team**

# 3. Research Program

## 3.1. Generative model design

The first objective of MODAL consists in designing, analyzing, estimating and evaluating new generative parametric models for multivariate and/or heterogeneous data. It corresponds typically to continuous and categorical data but it includes also other widespread ones like ordinal, functional, ranks,...Designed models have to take into account potential correlations between variables while being (1) justifiable and realistic, (2) meaningful and parsimoniously parameterized, (3) of low computational complexity. The main purpose is to identify a few theoretical and general principles for model generation, loosely dependent on the variable nature. In this context, we propose two concurrent approaches which could be general enough for dealing with correlation between many types of homogeneous or heterogeneous variables:

- Designs general models by combining two extreme models (full dependent and full independent) which are well-defined for most of variables;
- Uses kernels as a general way for dealing with multivariate and heterogeneous variables.

## 3.2. Data visualization

The second objective of MODAL is to propose meaningful and quite accurate low dimensional visualizations of data typically in two-dimensional (2D) spaces, less frequently in one-dimensional (1D) or three-dimensional (3D) spaces, by using the generative models designed in the first objective. We propose also to visualize simultaneously the data and the model. All visualizations will depend on the aim at hand (typically clustering, classification or density estimation). The main originality of this objective lies in the use of models for visualization, a strategy from which we expect to have a better control on the subjectivity necessarily induced by any graphical display. In addition, the proposed approach has to be general enough to be independent on the variable nature. Note that the visualization objective is consistent with the dissemination of our methodologies through specific softwares. Indeed, displaying data is an important step in the data analysis process.

**NON-A Project-Team**

# 3. Research Program

## 3.1. General annihilators

Estimation is quite easy in the absence of perturbations. It becomes challenging in more realistic situations, faced to measurement noises or other unknown inputs. In our works, as well as in the founding text of *Non-A*, we have shown how our estimation techniques can successfully get rid of perturbations of the so-called *structured* type, which means the ones that can be annihilated by some linear differential operator (called the annihilator). *ALIEN* already defined such operators by integral operators, but using more general convolution operators is an alternative to be analyzed, as well as defining the "best way to kill" perturbations. Open questions are:

**OQ1)** Does a normal form exist for such annihilators?

**OQ2)** Or, at least, does there exist an adequate basis representation of the annihilator in some adequate algebra?

**OQ3)** And lastly, can the annihilator parameters be derived from efficient tuning rules?

*The two first questions will directly impact Indicators 1 (time) and 2 (complexity), whereas the last one will impact indicator 3 (robustness).*

## 3.2. Numerical differentiation

Estimating the derivative of a (noisy) signal with a sufficient accuracy can be seen as a key problem in domains of control and diagnosis, as well as signal and image processing. At the present stage of our research, the estimation of the $n$-th order time derivatives of noisy signals (including noise filtering for $n = 0$) appears as a common area for the whole project, either as a research field, or as a tool that is used both for model-based and model-free techniques. *One of the open questions is about the robustness issues (Indicator 3) with respect to the annihilator, the parameters and the numerical implementation choices.*

Two classes of techniques are considered here (**Model-based** and **Model-free**), both of them aiming at non-asymptotic estimation.

In what we call *model-based techniques*, the derivative estimation is regarded as an observation problem, which means the software-based reconstruction of unmeasured variables and, more generally, a left inversion problem [0]. This involves linear/homogeneous/nonlinear state models, including ordinary equations, systems with delays, hybrid systems with impulses or switches [0], which still has to be exploited in the finite-time and fixed-time context. Power electronics is already one of the possible applications.

*Model-free techniques* concern the works initiated by *ALIEN*, which rely on the only information contained in the output signal and its derivatives. The corresponding algorithms rely on our algebraic annihilation viewpoint. *One open question is: How to provide an objective comparison analysis between Model-based and Model-free estimation techniques? For this, we will only concentrate on Non-Asymptotic ones. This comparison will have to be based on the three Indicators 1 (time), 2 (complexity) and 3 (robustness).*

[0]Left invertibility deals with the question of recovering the full state of a system ("observation") together with some of its inputs ("unknown input observers"), and also refers to algebraic structural conditions.

[0]Note that hybrid dynamical systems (HDS) constitute an important field of investigation since, in this case, the discrete state can be considered as an unknown input.

## 3.3. Model-free control

Industry is keen on simple and powerful controllers: the tuning simplicity of the classical PID controller explains its omnipresence in industrial control systems, although its performances drop when working conditions change. The last challenge we consider is to define control techniques which, instead of using sophisticated models (the development of which may be expensive), use the information contained in the output signal and its estimated derivatives, which can be regarded as "signal-based" controllers. *Such design should take into account the Indicators 1 (time), 2 (complexity) and 3 (robustness).*

## 3.4. Applications

Keeping in mind that we will remain focused at developing and applying fundamental methods for non-asymptotic estimation, we intend to deal with 4 main domains of application (see the lower part of Figure 1 ). The Lille context offers interesting opportunities in WSAN (wireless sensor and actuator networks and, more particularly, networked robots) at Inria, as well as nano/macro machining at ENSAM. A power electronics platform will be developed in ENSEA Cergy. Last, in contact with companies, several grants, patents and collaborations are expected from the applications of $i-$PID. Each of these four application domains was presented in the *Non-A* proposal:

- Networked robots, WSAN [Lille]
- Nano/macro machining [Lille]
- Multicell chopper [Lille and Cergy]
- *i*-PID for industry

In the present period, we choose to give a particular focus to the first item (Networked robots), which already received some development. It can be considered as the objective 4.

These applications are described with more details below.

**RMOD Project-Team**

# 3. Research Program

## 3.1. Software Reengineering

Strong coupling among the parts of an application severely hampers its evolution. Therefore, it is crucial to answer the following questions: How to support the substitution of certain parts while limiting the impact on others? How to identify reusable parts? How to modularize an object-oriented application?

Having good classes does not imply a good application layering, absence of cycles between packages and reuse of well-identified parts. Which notion of cohesion makes sense in presence of late-binding and programming frameworks? Indeed, frameworks define a context that can be extended by subclassing or composition: in this case, packages can have a low cohesion without being a problem for evolution. How to obtain algorithms that can be used on real cases? Which criteria should be selected for a given remodularization?

To help us answer these questions, we work on enriching Moose, our reengineering environment, with a new set of analyses [56], [55]. We decompose our approach in three main and potentially overlapping steps:

1. Tools for understanding applications,
2. Remodularization analyses,
3. Software Quality.

### *3.1.1. Tools for understanding applications*

**Context and Problems.** We are studying the problems raised by the understanding of applications at a larger level of granularity such as packages or modules. We want to develop a set of conceptual tools to support this understanding.

Some approaches based on Formal Concept Analysis (FCA) [84] show that such an analysis can be used to identify modules. However the presented examples are too small and not representative of real code.

**Research Agenda.**

FCA provides an important approach in software reengineering for software understanding, design anomalies detection and correction, but it suffers from two problems: (i) it produces lattices that must be interpreted by the user according to his/her understanding of the technique and different elements of the graph; and, (ii) the lattice can rapidly become so big that one is overwhelmed by the mass of information and possibilities [45]. We look for solutions to help people putting FCA to real use.

### *3.1.2. Remodularization analyses*

**Context and Problems.** It is a well-known practice to layer applications with bottom layers being more stable than top layers [72]. Until now, few works have attempted to identify layers in practice: Mudpie [86] is a first cut at identifying cycles between packages as well as package groups potentially representing layers. DSM (dependency structure matrix) [85], [80] seems to be adapted for such a task but there is no serious empirical experience that validates this claim. From the side of remodularization algorithms, many were defined for procedural languages [68]. However, object-oriented programming languages bring some specific problems linked with late-binding and the fact that a package does not have to be systematically cohesive since it can be an extension of another one [87], [59].

As we are designing and evaluating algorithms and analyses to remodularize applications, we also need a way to understand and assess the results we are obtaining.

**Research Agenda.** We work on the following items:

Layer identification. We propose an approach to identify layers based on a semi-automatic classification of package and class interrelationships that they contain. However, taking into account the wish or knowledge of the designer or maintainer should be supported.

Cohesion Metric Assessment. We are building a validation framework for cohesion/coupling metrics to determine whether they actually measure what they promise to. We are also compiling a number of traditional metrics for cohesion and coupling quality metrics to evaluate their relevance in a software quality setting.

### *3.1.3. Software Quality*

**Research Agenda.** Since software quality is fuzzy by definition and a lot of parameters should be taken into account we consider that defining precisely a unique notion of software quality is definitively a Grail in the realm of software engineering. The question is still relevant and important. We work on the two following items:

Quality models. We studied existing quality models and the different options to combine indicators — often, software quality models happily combine metrics, but at the price of losing the explicit relationships between the indicator contributions. There is a need to combine the results of one metric over all the software components of a system, and there is also the need to combine different metric results for any software component. Different combination methods are possible that can give very different results. It is therefore important to understand the characteristics of each method.

Bug prevention. Another aspect of software quality is validating or monitoring the source code to avoid the emergence of well known sources of errors and bugs. We work on how to best identify such common errors, by trying to identify earlier markers of possible errors, or by helping identifying common errors that programmers did in the past.

## 3.2. Language Constructs for Modular Design

While the previous axis focuses on how to help remodularizing existing software, this second research axis aims at providing new language constructs to build more flexible and recomposable software. We will build on our work on traits [82], [57] and classboxes [46] but also start to work on new areas such as isolation in dynamic languages. We will work on the following points: (1) Traits and (2) Modularization as a support for isolation.

### *3.2.1. Traits-based program reuse*

**Context and Problems.** Inheritance is well-known and accepted as a mechanism for reuse in object-oriented languages. Unfortunately, due to the coarse granularity of inheritance, it may be difficult to decompose an application into an optimal class hierarchy that maximizes software reuse. Existing schemes based on single inheritance, multiple inheritance, or mixins, all pose numerous problems for reuse.

To overcome these problems, we designed a new composition mechanism called Traits [82], [57]. Traits are pure units of behavior that can be composed to form classes or other traits. The trait composition mechanism is an alternative to multiple or mixin inheritance in which the composer has full control over the trait composition. The result enables more reuse than single inheritance without introducing the drawbacks of multiple or mixin inheritance. Several extensions of the model have been proposed [54], [76], [47], [58] and several type systems were defined [60], [83], [77], [70].

Traits are reusable building blocks that can be explicitly composed to share methods across unrelated class hierarchies. In their original form, traits do not contain state and cannot express visibility control for methods. Two extensions, stateful traits and freezable traits, have been proposed to overcome these limitations. However, these extensions are complex both to use for software developers and to implement for language designers.

**Research Agenda: Towards a pure trait language.** We plan distinct actions: (1) a large application of traits, (2) assessment of the existing trait models and (3) bootstrapping a pure trait language.

- To evaluate the expressiveness of traits, some hierarchies were refactored, showing code reuse [49]. However, such large refactorings, while valuable, may not exhibit all possible composition problems, since the hierarchies were previously expressed using single inheritance and following certain patterns. We want to redesign from scratch the collection library of Smalltalk (or part of it). Such a redesign should on the one hand demonstrate the added value of traits on a real large and redesigned library and on the other hand foster new ideas for the bootstrapping of a pure trait-based language.

  In particular we want to reconsider the different models proposed (stateless [57], stateful [48], and freezable [58]) and their operators. We will compare these models by (1) implementing a trait-based collection hierarchy, (2) analyzing several existing applications that exhibit the need for traits. Traits may be flattened [75]. This is a fundamental property that confers to traits their simplicity and expressiveness over Eiffel's multiple inheritance. Keeping these aspects is one of our priority in forthcoming enhancements of traits.
- Alternative trait models. This work revisits the problem of adding state and visibility control to traits. Rather than extending the original trait model with additional operations, we use a fundamentally different approach by allowing traits to be lexically nested within other modules. This enables traits to express (shared) state and visibility control by hiding variables or methods in their lexical scope. Although the traits' "flattening property" no longer holds when they can be lexically nested, the combination of traits with lexical nesting results in a simple and more expressive trait model. We formally specify the operational semantics of this combination. Lexically nested traits are fully implemented in AmbientTalk, where they are used among others in the development of a Morphic-like UI framework.
- We want to evaluate how inheritance can be replaced by traits to form a new object model. For this purpose we will design a minimal reflective kernel, inspired first from ObjVlisp [53] then from Smalltalk [63].

### *3.2.2. Reconciling Dynamic Languages and Isolation*

**Context and Problems.** More and more applications require dynamic behavior such as modification of their own execution (often implemented using reflective features [67]). For example, F-script allows one to script Cocoa Mac-OS X applications and Lua is used in Adobe Photoshop. Now in addition more and more applications are updated on the fly, potentially loading untrusted or broken code, which may be problematic for the system if the application is not properly isolated. Bytecode checking and static code analysis are used to enable isolation, but such approaches do not really work in presence of dynamic languages and reflective features. Therefore there is a tension between the need for flexibility and isolation.

**Research Agenda: Isolation in dynamic and reflective languages.** To solve this tension, we will work on *Sure*, a language where isolation is provided by construction: as an example, if the language does not offer field access and its reflective facilities are controlled, then the possibility to access and modify private data is controlled. In this context, layering and modularizing the meta-level [50], as well as controlling the access to reflective features [51], [52] are important challenges. We plan to:

- Study the isolation abstractions available in erights (http://www.erights.org) [74], [73], and Java's class loader strategies [69], [64].
- Categorize the different reflective features of languages such as CLOS [66], Python and Smalltalk [78] and identify suitable isolation mechanisms and infrastructure [61].
- Assess different isolation models (access rights, capabilities [79]...) and identify the ones adapted to our context as well as different access and right propagation.
- Define a language based on
  - the decomposition and restructuring of the reflective features [50],

- the use of encapsulation policies as a basis to restrict the interfaces of the controlled objects [81],
- the definition of method modifiers to support controlling encapsulation in the context of dynamic languages.

An open question is whether, instead of providing restricted interfaces, we could use traits to grant additional behavior to specific instances: without trait application, the instances would only exhibit default public behavior, but with additional traits applied, the instances would get extra behavior. We will develop *Sure*, a modular extension of the reflective kernel of Smalltalk (since it is one of the languages offering the largest set of reflective features such as pointer swapping, class changing, class definition...) [78].

**SEQUEL Project-Team**

# 3. Research Program

## 3.1. In Short

SEQUEL is primarily grounded on two domains:

- the problem of decision under uncertainty,
- statistical analysis and statistical learning, which provide the general concepts and tools to solve this problem.

To help the reader who is unfamiliar with these questions, we briefly present key ideas below.

## 3.2. Decision-making Under Uncertainty

The phrase "Decision under uncertainty" refers to the problem of taking decisions when we do not have a full knowledge neither of the situation, nor of the consequences of the decisions, as well as when the consequences of decision are non deterministic.

We introduce two specific sub-domains, namely the Markov decision processes which models sequential decision problems, and bandit problems.

### *3.2.1. Reinforcement Learning*

Sequential decision processes occupy the heart of the SEQUEL project; a detailed presentation of this problem may be found in Puterman's book [48].

A Markov Decision Process (MDP) is defined as the tuple $(\mathcal{X}, \mathcal{A}, P, r)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $P$ is the probabilistic transition kernel, and $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function. For the sake of simplicity, we assume in this introduction that the state and action spaces are finite. If the current state (at time $t$) is $x \in \mathcal{X}$ and the chosen action is $a \in \mathcal{A}$, then the Markov assumption means that the transition probability to a new state $x' \in \mathcal{X}$ (at time $t+1$) only depends on $(x, a)$. We write $p(x'|x, a)$ the corresponding transition probability. During a transition $(x, a) \to x'$, a reward $r(x, a, x')$ is incurred.

In the MDP $(\mathcal{X}, \mathcal{A}, P, r)$, each initial state $x_0$ and action sequence $a_0, a_1, ...$ gives rise to a sequence of states $x_1, x_2, ...$, satisfying $\mathbb{P}(x_{t+1} = x'|x_t = x, a_t = a) = p(x'|x, a)$, and rewards [0]$r_1, r_2, ...$ defined by $r_t = r(x_t, a_t, x_{t+1})$.

The history of the process up to time $t$ is defined to be $H_t = (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t)$. A policy $\pi$ is a sequence of functions $\pi_0, \pi_1, ...$, where $\pi_t$ maps the space of possible histories at time $t$ to the space of probability distributions over the space of actions $\mathcal{A}$. To follow a policy means that, in each time step, we assume that the process history up to time $t$ is $x_0, a_0, ..., x_t$ and the probability of selecting an action $a$ is equal to $\pi_t(x_0, a_0, ..., x_t)(a)$. A policy is called stationary (or Markovian) if $\pi_t$ depends only on the last visited state. In other words, a policy $\pi = (\pi_0, \pi_1, ...)$ is called stationary if $\pi_t(x_0, a_0, ..., x_t) = \pi_0(x_t)$ holds for all $t \geq 0$. A policy is called deterministic if the probability distribution prescribed by the policy for any history is concentrated on a single action. Otherwise it is called a stochastic policy.

[0]Note that for simplicity, we considered the case of a deterministic reward function, but in many applications, the reward $r_t$ itself is a random variable.

We move from an MD process to an MD problem by formulating the goal of the agent, that is what the sought policy $\pi$ has to optimize? It is very often formulated as maximizing (or minimizing), in expectation, some functional of the sequence of future rewards. For example, an usual functional is the infinite-time horizon sum of discounted rewards. For a given (stationary) policy $\pi$, we define the value function $V^\pi(x)$ of that policy $\pi$ at a state $x \in \mathcal{X}$ as the expected sum of discounted future rewards given that we state from the initial state $x$ and follow the policy $\pi$:

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x, \pi\right], \tag{4}$$

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function $V^\pi$ gives an evaluation of the performance of a given policy $\pi$. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [43]) and average reward settings. Note also that, here, we considered the problem of maximizing a reward functional, but a formulation in terms of minimizing some cost or risk functional would be equivalent.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [41], which introduces the optimal value function $V^*(x)$, defined as the optimal expected sum of rewards when the agent starts from a state $x$. We have $V^*(x) = \sup_\pi V^\pi(x)$. Now, let us give two definitions about policies:

- We say that a policy $\pi$ is optimal, if it attains the optimal values $V^*(x)$ for any state $x \in \mathcal{X}$, *i.e.*, if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. Under mild conditions, deterministic stationary optimal policies exist [42]. Such an optimal policy is written $\pi^*$.
- We say that a (deterministic stationary) policy $\pi$ is greedy with respect to (w.r.t.) some function $V$ (defined on $\mathcal{X}$) if, for all $x \in \mathcal{X}$,

$$\pi(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V(x')\right].$$

  where $\arg\max_{a \in \mathcal{A}} f(a)$ is the set of $a \in \mathcal{A}$ that maximizes $f(a)$. For any function $V$, such a greedy policy always exists because $\mathcal{A}$ is finite.

The goal of Reinforcement Learning (RL), as well as that of dynamic programming, is to design an optimal policy (or a good approximation of it).

The well-known Dynamic Programming equation (also called the Bellman equation) provides a relation between the optimal value function at a state $x$ and the optimal value function at the successors states $x'$ when choosing an optimal action: for all $x \in \mathcal{X}$,

$$V^*(x) = \max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{5}$$

The benefit of introducing this concept of optimal value function relies on the property that, from the optimal value function $V^*$, it is easy to derive an optimal behavior by choosing the actions according to a policy greedy w.r.t. $V^*$. Indeed, we have the property that a policy greedy w.r.t. the optimal value function is an optimal policy:

$$\pi^*(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{6}$$

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ( [54]):

- Bellman's dynamic programming approach, based on the introduction of the value function. It consists in learning a "good" approximation of the optimal value function, and then using it to derive a greedy policy w.r.t. this approximation. The hope (well justified in several cases) is that the performance $V^\pi$ of the policy $\pi$ greedy w.r.t. an approximation $V$ of $V^*$ will be close to optimality. This approximation issue of the optimal value function is one of the major challenges inherent to the reinforcement learning problem. **Approximate dynamic programming** addresses the problem of estimating performance bounds (*e.g.* the loss in performance $||V^* - V^\pi||$ resulting from using a policy $\pi$-greedy w.r.t. some approximation $V$- instead of an optimal policy) in terms of the approximation error $||V^* - V||$ of the optimal value function $V^*$ by $V$. Approximation theory and Statistical Learning theory provide us with bounds in terms of the number of sample data used to represent the functions, and the capacity and approximation power of the considered function spaces.
- Pontryagin's maximum principle approach, based on sensitivity analysis of the performance measure w.r.t. some control parameters. This approach, also called **direct policy search** in the Reinforcement Learning community aims at directly finding a good feedback control law in a parameterized policy space without trying to approximate the value function. The method consists in estimating the so-called **policy gradient**, *i.e.* the sensitivity of the performance measure (the value function) w.r.t. some parameters of the current policy. The idea being that an optimal control problem is replaced by a parametric optimization problem in the space of parameterized policies. As such, deriving a policy gradient estimate would lead to performing a stochastic gradient method in order to search for a local optimal parametric policy.

Finally, many extensions of the Markov decision processes exist, among which the Partially Observable MDPs (POMDPs) is the case where the current state does not contain all the necessary information required to decide for sure of the best action.

### *3.2.2. Multi-arm Bandit Theory*

Bandit problems illustrate the fundamental difficulty of decision making in the face of uncertainty: A decision maker must choose between what seems to be the best choice ("exploit"), or to test ("explore") some alternative, hoping to discover a choice that beats the current best choice.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when the effectiveness of the treatments are initially unknown and the patients arrive sequentially. These bandit problems became popular with the seminal paper [49], after which they have found applications in diverse fields, such as control, economics, statistics, or learning theory.

Formally, a K-armed bandit problem ($K \geq 2$) is specified by K real-valued distributions. In each time step a decision maker can select one of the distributions to obtain a sample from it. The samples obtained are considered as rewards. The distributions are initially unknown to the decision maker, whose goal is to maximize the sum of the rewards received, or equivalently, to minimize the regret which is defined as the loss compared to the total payoff that can be achieved given full knowledge of the problem, *i.e.*, when the arm giving the highest expected reward is pulled all the time.

The name "bandit" comes from imagining a gambler playing with K slot machines. The gambler can pull the arm of any of the machines, which produces a random payoff as a result: When arm k is pulled, the random payoff is drawn from the distribution associated to k. Since the payoff distributions are initially unknown, the gambler must use exploratory actions to learn the utility of the individual arms. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, to play well, the gambler must carefully balance exploration and exploitation. Auer *et al.* [40] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most

at a logarithmic rate with the number of trials, and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions).

## 3.3. Statistical analysis of time series

Many of the problems of machine learning can be seen as extensions of classical problems of mathematical statistics to their (extremely) non-parametric and model-free cases. Other machine learning problems are founded on such statistical problems. Statistical problems of sequential learning are mainly those that are concerned with the analysis of time series. These problems are as follows.

### *3.3.1. Prediction of Sequences of Structured and Unstructured Data*

Given a series of observations $x_1, \cdots, x_n$ it is required to give forecasts concerning the distribution of the future observations $x_{n+1}, x_{n+2}, \cdots$; in the simplest case, that of the next outcome $x_{n+1}$. Then $x_{n+1}$ is revealed and the process continues. Different goals can be formulated in this setting. One can either make some assumptions on the probability measure that generates the sequence $x_1, \cdots, x_n, \cdots$, such as that the outcomes are independent and identically distributed (i.i.d.), or that the sequence is a Markov chain, that it is a stationary process, etc. More generally, one can assume that the data is generated by a probability measure that belongs to a certain set $\mathcal{C}$. In these cases the goal is to have the discrepancy between the predicted and the "true" probabilities to go to zero, if possible, with guarantees on the speed of convergence.

Alternatively, rather than making some assumptions on the data, one can change the goal: the predicted probabilities should be asymptotically as good as those given by the best reference predictor from a certain pre-defined set.

Another dimension of complexity in this problem concerns the nature of observations $x_i$. In the simplest case, they come from a finite space, but already basic applications often require real-valued observations. Moreover, function or even graph-valued observations often arise in practice, in particular in applications concerning Web data. In these settings estimating even simple characteristics of probability distributions of the future outcomes becomes non-trivial, and new learning algorithms for solving these problems are in order.

### *3.3.2. Hypothesis testing*

Given a series of observations of $x_1, \cdots, x_n, \cdots$ generated by some unknown probability measure $\mu$, the problem is to test a certain given hypothesis $H_0$ about $\mu$, versus a given alternative hypothesis $H_1$. There are many different examples of this problem. Perhaps the simplest one is testing a simple hypothesis "$\mu$ is Bernoulli i.i.d. measure with probability of 0 equals 1/2" versus "$\mu$ is Bernoulli i.i.d. with the parameter different from 1/2". More interesting cases include the problems of model verification: for example, testing that $\mu$ is a Markov chain, versus that it is a stationary ergodic process but not a Markov chain. In the case when we have not one but several series of observations, we may wish to test the hypothesis that they are independent, or that they are generated by the same distribution. Applications of these problems to a more general class of machine learning tasks include the problem of feature selection, the problem of testing that a certain behaviour (such as pulling a certain arm of a bandit, or using a certain policy) is better (in terms of achieving some goal, or collecting some rewards) than another behaviour, or than a class of other behaviours.

The problem of hypothesis testing can also be studied in its general formulations: given two (abstract) hypothesis $H_0$ and $H_1$ about the unknown measure that generates the data, find out whether it is possible to test $H_0$ against $H_1$ (with confidence), and if yes then how can one do it.

### *3.3.3. Change Point Analysis*

A stochastic process is generating the data. At some point, the process distribution changes. In the "offline" situation, the statistician observes the resulting sequence of outcomes and has to estimate the point or the points at which the change(s) occurred. In online setting, the goal is to detect the change as quickly as possible.

These are the classical problems in mathematical statistics, and probably among the last remaining statistical problems not adequately addressed by machine learning methods. The reason for the latter is perhaps in that the problem is rather challenging. Thus, most methods available so far are parametric methods concerning piece-wise constant distributions, and the change in distribution is associated with the change in the mean. However, many applications, including DNA analysis, the analysis of (user) behaviour data, etc., fail to comply with this kind of assumptions. Thus, our goal here is to provide completely non-parametric methods allowing for any kind of changes in the time-series distribution.

### *3.3.4. Clustering Time Series, Online and Offline*

The problem of clustering, while being a classical problem of mathematical statistics, belongs to the realm of unsupervised learning. For time series, this problem can be formulated as follows: given several samples $x^1 = (x_1^1, \cdots, x_{n_1}^1), \cdots, x^N = (x_N^1, \cdots, x_{n_N}^N)$, we wish to group similar objects together. While this is of course not a precise formulation, it can be made precise if we assume that the samples were generated by $k$ different distributions.

The online version of the problem allows for the number of observed time series to grow with time, in general, in an arbitrary manner.

### *3.3.5. Online Semi-Supervised Learning*

Semi-supervised learning (SSL) is a field of machine learning that studies learning from both labeled and unlabeled examples. This learning paradigm is extremely useful for solving real-world problems, where data is often abundant but the resources to label them are limited.

Furthermore, *online* SSL is suitable for adaptive machine learning systems. In the classification case, learning is viewed as a repeated game against a potentially adversarial nature. At each step $t$ of this game, we observe an example $\mathbf{x_t}$, and then predict its label $\widehat{y}_t$.

The challenge of the game is that we only exceptionally observe the true label $y_t$. In the extreme case, which we also study, only a handful of labeled examples are provided in advance and set the initial bias of the system while unlabeled examples are gathered online and update the bias continuously. Thus, if we want to adapt to changes in the environment, we have to rely on indirect forms of feedback, such as the structure of data.

## 3.4. Statistical Learning and Bayesian Analysis

Before detailing some issues in these fields, let us remind the definition of a few terms.

**Machine learning** refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness.

**Statistical learning** is an approach to machine intelligence that is based on statistical modeling of data. With a statistical model in hand, one applies probability theory and decision theory to get an algorithm. This is opposed to using training data merely to select among different algorithms or using heuristics/"common sense" to design an algorithm.

**Bayesian Analysis** applies to data that could be seen as observations in the more general meaning of the term. These data may not only come from classical sensors but also from any *device* recording information. From an operational point of view, like for statistical learning, uncertainty about the data is modeled by a probability measure thus defining the so-called likelihood functions. This last one depends upon parameters defining the state of the world we focus on for decision purposes. Within the Bayesian framework the uncertainty about these parameters is also modeled by probability measures, the priors that are subjective probabilities. Using probability theory and decision theory, one then defines new algorithms to estimate the parameters of interest and/or associated decisions. According to the International Society for Bayesian Analysis (source: http://bayesian.org), and from a more general point of view, this overall process could be

summarize as follows: one assesses the current state of knowledge regarding the issue of interest, gather new data to address remaining questions, and then update and refine their understanding to incorporate both new and old data. Bayesian inference provides a logical, quantitative framework for this process based on probability theory.

**Kernel method.** Generally speaking, a kernel function is a function that maps a couple of points to a real value. Typically, this value is a measure of dissimilarity between the two points. Assuming a few properties on it, the kernel function implicitly defines a dot product in some function space. This very nice formal property as well as a bunch of others have ensured a strong appeal for these methods in the last 10 years in the field of function approximation. Many classical algorithms have been "kernelized", that is, restated in a much more general way than their original formulation. Kernels also implicitly induce the representation of data in a certain "suitable" space where the problem to solve (classification, regression, ...) is expected to be simpler (non-linearity turns to linearity).

The fundamental tools used in SEQUEL come from the field of statistical learning [45]. We briefly present the most important for us to date, namely, kernel-based non parametric function approximation, and non parametric Bayesian models.

### *3.4.1. Non-parametric methods for Function Approximation*

In statistics in general, and applied mathematics, the approximation of a multi-dimensional real function given some samples is a well-known problem (known as either regression, or interpolation, or function approximation, ...). Regressing a function from data is a key ingredient of our research, or to the least, a basic component of most of our algorithms. In the context of sequential learning, we have to regress a function while data samples are being obtained one at a time, while keeping the constraint to be able to predict points at any step along the acquisition process. In sequential decision problems, we typically have to learn a value function, or a policy.

Many methods have been proposed for this purpose. We are looking for suitable ones to cope with the problems we wish to solve. In reinforcement learning, the value function may have areas where the gradient is large; these are areas where the approximation is difficult, while these are also the areas where the accuracy of the approximation should be maximal to obtain a good policy (and where, otherwise, a bad choice of action may imply catastrophic consequences).

We particularly favor non parametric methods since they make quite a few assumptions about the function to learn. In particular, we have strong interests in $l_1$-regularization, and the (kernelized-)LARS algorithm. $l_1$-regularization yields sparse solutions, and the LARS approach produces the whole regularization path very efficiently, which helps solving the regularization parameter tuning problem.

### *3.4.2. Nonparametric Bayesian Estimation*

Numerous problems may be solved efficiently by a Bayesian approach. The use of Monte-Carlo methods allows us to handle non–linear, as well as non–Gaussian, problems. In their standard form, they require the formulation of probability densities in a parametric form. For instance, it is a common usage to use Gaussian likelihood, because it is handy. However, in some applications such as Bayesian filtering, or blind deconvolution, the choice of a parametric form of the density of the noise is often arbitrary. If this choice is wrong, it may also have dramatic consequences on the estimation quality. To overcome this shortcoming, one possible approach is to consider that this density must also be estimated from data. A general Bayesian approach then consists in defining a probabilistic space associated with the possible outcomes of the *object* to be estimated. Applied to density estimation, it means that we need to define a probability measure on the probability density of the noise: such a measure is called a *random measure*. The classical Bayesian inference procedures can then been used. This approach being by nature non parametric, the associated frame is called *Non Parametric Bayesian*.

In particular, mixtures of Dirichlet processes [44] provide a very powerful formalism. Dirichlet Processes are a possible random measure and Mixtures of Dirichlet Processes are an extension of well-known finite mixture models. Given a mixture density $f(x|\theta)$, and $G(d\theta) = \sum_{k=1}^{\infty} \omega_k \delta_{U_k}(d\theta)$, a Dirichlet process, we define a mixture of Dirichlet processes as:

$$F(x) = \int_{\Theta} f(x|\theta) G(d\theta) = \sum_{k=1}^{\infty} \omega_k f(x|U_k) \tag{7}$$

where $F(x)$ is the density to be estimated. The class of densities that may be written as a mixture of Dirichlet processes is very wide, so that they really fit a very large number of applications.

Given a set of observations, the estimation of the parameters of a mixture of Dirichlet processes is performed by way of a Monte Carlo Markov Chain (MCMC) algorithm. Dirichlet Process Mixture are also widely used in clustering problems. Once the parameters of a mixture are estimated, they can be interpreted as the parameters of a specific cluster defining a class as well. Dirichlet processes are well known within the machine learning community and their potential in statistical signal processing still need to be developed.

### *3.4.3. Random Finite Sets for multisensor multitarget tracking*

In the general multi-sensor multi-target Bayesian framework, an unknown (and possibly varying) number of targets whose states $x_1, ...x_n$ are observed by several sensors which produce a collection of measurements $z_1, ..., z_m$ at every time step $k$. Well-known models to this problem are track-based models, such as the joint probability data association (JPDA), or joint multi-target probabilities, such as the joint multi-target probability density. Common difficulties in multi-target tracking arise from the fact that the system state and the collection of measures from sensors are unordered and their size evolve randomly through time. Vector-based algorithms must therefore account for state coordinates exchanges and missing data within an unknown time interval. Although this approach is very popular and has resulted in many algorithms in the past, it may not be the optimal way to tackle the problem, since the sate and the data are in fact *sets* and not vectors.

The random finite set theory provides a powerful framework to deal with these issues. Mahler's work on finite sets statistics (FISST) provides a mathematical framework to build multi-object densities and derive the Bayesian rules for state prediction and state estimation. Randomness on object number and their states are encapsulated into random finite sets (RFS), namely multi-target(state) sets $X = \{x_1, ..., x_n\}$ and multi-sensor (measurement) set $Zk = \{z_1, ..., z_m\}$. The objective is then to propagate the multitarget probability density $f_{k|k}(X|Z(k))$ by using the Bayesian set equations at every time step $k$:

$$\begin{aligned} f_{k+1|k}(X|Z^{(k)}) &= \int f_{k+1|k}(X|W) f_{k|k}(W|Z^{(k)}) \delta W \\ f_{k+1|k+1}(X|Z^{(k+1)}) &= \frac{f_{k+1}(Z_{k+1}|X) f_{k+1|k}(X|Z^{(k)})}{\int f_{k+1}(Z_{k+1}|W) f_{k+1|k}(W|Z^{(k)}) \delta W} \end{aligned} \tag{8}$$

where:

- $X = \{x_1, ..., x_n\}$ is a multi-target state, *i.e.* a finite set of elements $x_i$ defined on the single-target space $\mathcal{X}$; [0]
- $Z_{k+1} = \{z_1, ..., z_m\}$ is the current multi-sensor observation, *i.e.* a collection of measures $z_i$ produced at time $k+1$ by all the sensors;
- $Z^{(k)} = \bigcup_{t \leqslant k} Z_t$ is the collection of observations up to time $k$;
- $f_{k|k}(W|Z^{(k)})$ is the current multi-target posterior density in state $W$;
- $f_{k+1|k}(X|W)$ is the current multi-target Markov transition density, from state $W$ to state $X$;
- $f_{k+1}(Z|X)$ is the current multi-sensor/multi-target likelihood function.

[0] The state $x_i$ of a target is usually composed of its position, its velocity, etc.

Although equations (5 ) may seem similar to the classical single-sensor/single-target Bayesian equations, they are generally intractable because of the presence of the *set integrals*. For, a RFS $\Xi$ is characterized by the family of its Janossy densities $j_{\Xi,1}(x_1)$, $j_{\Xi,2}(x_1,x_2)$... and not just by one density as it is the case with vectors. Mahler then introduced the PHD, defined on single-target state space. The PHD is the quantity whose integral on any region $S$ is the expected number of targets inside $S$. Mahler proved that the PHD is the first-moment density of the multi-target probability density. Although defined on single-state space X, the PHD encapsulates information on both target number and states.

**SPIRALS Team**

# 3. Research Program

## 3.1. Introduction

Our research program on self-adaptive software targets two key properties that are detailed in the remainder of this section: self-healing and self-optimization.

## 3.2. Objective #1: Self-healing - Mining software artifacts to automatically evolve systems

Software systems are under the pressure of changes all along their lifecycle. Agile development blurs the frontier between design and execution and requires constant adaptation. The size of systems (millions of lines of code) multiplies the number of bugs by the same order of magnitude. More and more systems, such as sensor network devices, live in "surviving" mode, in the sense that they are neither rebootable nor upgradable.

Software bugs are hidden in source code and show up at development-time, testing-time or worse, once deployed in production. Except for very specific application domains where formal proofs are achievable, bugs can not be eradicated. As an order of magnitude, on 16 Dec 2011, the Eclipse bug repository contains 366,922 bug reports. Software engineers and developers work on bug fixing on a daily basis. Not all developers spend the same time on bug fixing. In large companies, this is sometimes a full-time role to manage bugs, often referred to as Quality Assurance (QA) software engineers. Also, not all bugs are equal, some bugs are analyzed and fixed within minutes, others may take months to be solved [123].

In terms of research, this means that: (i) one needs means to automatically adapt the design of the software system through automated refactoring and API extraction, (ii) one needs approaches to automate the process of adapting source code in order to fix certain bugs, (iii) one needs to revisit the notion of error-handling so that instead of crashing in presence of errors, software adapts itself to continue with its execution, e.g., in degraded mode.

There is no one-size-fits-all solution for each of these points. However, we think that novel solutions can be found by using **data mining and machine learning techniques tailored for software engineering** [124]. This body of research consists of mining some knowledge about a software system by analyzing the source code, the version control systems, the execution traces, documentation and all kinds of software development and execution artifacts in general. This knowledge is then used within recommendation systems for software development, auditing tools, runtime monitors, frameworks for resilient computing, etc.

The novelty of our approach consists of using and tailoring data mining techniques for analyzing software artifacts (source code, execution traces) in order to achieve the **next level of automated adaptation** (*e.g.*, automated bug fixing). Technically, we plan to mix unsupervised statistical learning techniques (*e.g.* frequent item set mining) and supervised ones (*e.g.* training classifiers such as decision trees). This research is currently not being performed by data mining research teams since it requires a high level of domain expertise in software engineering, while software engineering researchers can use off-the-shelf data mining libraries, such as Weka [98].

We now detail the two directions that we propose to follow to achieve this objective.

### *3.2.1. Learning from software history how to design software and fix bugs*

The first direction is about mining techniques in software repositories (*e.g.*, CVS, SVN, Git). Best practices can be extracted by data mining source code and the version control history of existing software systems. The design and code of expert developers significantly vary from the artifacts of novice developers. We will learn to differentiate those design characteristics by comparing different code bases, and by observing the semantic refactoring actions from version control history. Those design rules can then feed the test-develop-refactor constant adaptation cycle of agile development.

**Fault localization of bugs reported in bug repositories.** We will build a solid foundation on empirical knowledge about bugs reported in bug repository. We will perform an empirical study on a set of representative bug repositories to identify classes of bugs and patterns of bug data. For this, we will build a tool to browse and annotate bug reports. Browsing will be helped with two kinds of indexing: first, the tool will index all textual artifacts for each bug report; second it will index the semantic information that is not present by default in bug management software (*i.e.*, "contains a stacktrace"). Both indexes will be used to find particular subsets of bug reports, for instance "all bugs mentioning invariants and containing a stacktrace". Note that queries with this kind of complexity and higher are mostly not possible with the state-of-the-art of bug management software. Then, analysts will use annotation features to annotate bug reports. The main outcome of the empirical study will be the identification of classes of bugs that are appropriate for automated localization. Then, we will run machine learning algorithms to identify the latent links between the bug report content and source code features. Those algorithms would use as training data the existing traceability links between bug reports and source code modifications from version control systems. We will start by using decision trees since they produce a model that is explicit and understandable by expert developers. Depending on the results, other machine learning algorithms will be used. The resulting system will be able to locate elements in source code related to a certain bug report with a certain confidence.

**Automated bug fix generation with search-based techniques.** Once a location in code is identified as being the cause of the bug, we can try to automatically find a potential fix. We envision different techniques: (1) infer fixes from existing contracts and specifications that are violated; (2) infer fixes from the software behavior specified as a test suite; (3) try different fix types one-by-one from a list of identified bug fix patterns; (4) search fixes in a fix space that consists of combinations of atomic bug fixes. Techniques 1 and 2 are explored in [91] and [122]. We will focus on the latter techniques. To identify bug fix patterns and atomic bug fixes, we will perform a large-scale empirical study on software changes (also known as changesets when referring to changes across multiple files). We will develop tools to navigate, query and annotate changesets in a version control system. Then, a grounded theory will be built to master the nature of fixes. Eventually, we will decompose change sets in atomic actions using clustering on changeset actions. We will then use this body of empirical knowledge to feed search-based algorithms (*e.g.* genetic algorithms) that will look for meaningful fixes in a large fix space. To sum up, our research on automated bug fixing will try not only to point to source code locations responsible of a bug, but to search for code patterns and snippets that may constitute the skeleton of a valid patch. Ultimately, a blend of expert heuristics and learned rules will be able to produce valid source code that can be validated by developers and committed to the code base.

### *3.2.2. Run-time self-healing*

The second proposed research direction is about inventing a self-healing capability at run-time. This is complementary to the previous objective that mainly deals with development time issues. We will achieve this in two steps. First, we want to define frameworks for resilient software systems. Those frameworks will help to maintain the execution even in the presence of bugs, *i.e.* to let the system survive. As exposed below, this may mean for example to switch to some degraded modes. Next, we want to go a step further and to define solutions for automated runtime repair, that is, not simply compensating the erroneous behavior, but also determining the correct repair actions and applying them at run-time.

**Mining best effort values.** A well-known principle of software engineering is the "fail-fast" principle. In a nutshell, it states that as soon as something goes wrong, software should stop the execution before entering incorrect states. This is fine when a human user is in the loop, capable of understanding the error or at least rebooting the system. However, the notion of "failure-oblivious computing" [112] shows that in certain domains, software should run in a resilient mode (*i.e.* capable of recovering from errors) and/or best-effort mode (*i.e.* a slightly imprecise computation is better than stopping). Hence, we plan to investigate data mining techniques in order to learn best-effort values from past executions (*i.e.* somehow learning what is a correct state, or the opposite what is not a completely incorrect state). This knowledge will then be used to adapt the software state and flow in order to mitigate the error consequences, the exact opposite of fail-fast for systems with long-running cycles.

**Embedding search based algorithms at runtime.** Harman recently described the field of search-based software engineering [99]. We think that certain search based approaches can be embedded at runtime with the goal of automatically finding solutions that avoid crashing. We will create software infrastructures that allow automatically detecting and repairing faults at run-time. The methodology for achieving this task is based on three points: (1) empirical study of runtime faults; (2) learning approaches to characterize runtime faults; (3) learning algorithms to produce valid changes to the software runtime state. An empirical study will be performed to analyze those bug reports that are associated with runtime information (*e.g.* core dumps or stacktraces). After this empirical study, we will create a system that learns on previous repairs how to produce small changes that solve standard runtime bugs (*e.g.* adding an array bound check to throw a handled domain exception rather than a spurious language exception). To achieve this task, component models will be used to (1) encapsulate the monitoring and reparation meta-programs in appropriate components and (2) support runtime code modification using scripting, reflective or bytecode generation techniques.

## 3.3. Objective #2: Self-optimization - Sharing runtime behaviors to continuously adapt software

Complex distributed systems have to seamlessly adapt to a wide variety of deployment targets. This is due to the fact that developers cannot anticipate all the runtime conditions under which these systems are immersed. A major challenge for these software systems is to develop their capability to continuously reason about themselves and to take appropriate decisions and actions on the optimizations they can apply to improve themselves. This challenge encompasses research contributions in different areas, from environmental monitoring to realtime symptoms diagnosis, to automated decision making. The variety of distributed systems, the number of optimization parameters, and the complexity of decisions often resign the practitioners to design monolithic and static middleware solutions. However, it is now globally acknowledged that the development of dedicated building blocks does not contribute to the adoption of sustainable solutions. This is confirmed by the scale of actual distributed systems, which can—for example—connect several thousands of devices to a set of services hosted in the Cloud. In such a context, the lack of support for smart behaviours at different levels of the systems can inevitably lead to its instability or its unavailability. In June 2012, an outage of Amazon's Elastic Compute Cloud in North Virginia has taken down Netflix, Pinterest, and Instagram services. During hours, all these services failed to satisfy their millions of customers due to the lack of integration of a self-optimization mechanism going beyond the boundaries of Amazon.

The research contributions we envision within this area will therefore be organized as a reference model for engineering **self-optimized distributed systems** autonomously driven by *adaptive feedback control loops*, which will automatically enlarge their scope to cope with the complexity of the decisions to be taken. This solution introduces a multi-scale approach, which first privileges local and fast decisions to ensure the homeostasis [0] property of a single node, and then progressively propagates symptoms in the network in order to reason on a longer term and a larger number of nodes. Ultimately, domain experts and software developers can be automatically involved in the decision process if the system fails to find a satisfying solution. The research program for this objective will therefore focus on the study of mechanisms for **monitoring, taking decisions, and automatically reconfiguring software at runtime and at various scales**. As stated in the self-healing objective, we believe that there is no one-size-fits-all mechanism that can span all the scales of the system. We will therefore study and identify an optimal composition of various adaptation mechanisms in order to produce long-living software systems.

The novelty of this objective is to exploit the wisdom of crowds to define new middleware solutions that are able to continuously adapt software deployed in the wild. We intend to demonstrate the applicability of this approach to distributed systems that are deployed from mobile phones to cloud infrastructures. The key scientific challenges to address can be summarized as follows: *How does software behave once deployed in the wild? Is it possible to automatically infer the quality of experience, as it is perceived by users? Can the*

[0] Homeostasis is the property of a system that regulates its internal environment and tends to maintain a stable, relatively constant condition of properties [Wikipedia].

*runtime optimizations be shared across a wide variety of software? How optimizations can be safely operated on large populations of software instances?*

The remainder of this section further elaborates on the opportunities that can be considered within the frame of this objective.

### 3.3.1. Monitoring software in the wild

Once deployed, developers are generally no longer aware of how their software behave. Even if they heavily use testbeds and benchmarks during the development phase, they mostly rely on the bugs explicitly reported by users to monitor the efficiency of their applications. However, it has been shown that contextual artifacts collected at runtime can help to understand performance leaks and optimize the resilience of software systems [125]. Monitoring and understanding the context of software at runtime therefore represent the first building block of this research challenge. Practically, we intend to investigate crowdsensing approaches, to smartly collect and process runtime metrics (*e.g.*, request throughput, energy consumption, user context). Crowdsensing can be seen as a specific kind of crowdsourcing activity, which refers to the capability of lifting a (large) diffuse group of participants to delegate the task of retrieving trustable data from the field. In particular, crowdsensing covers not only *participatory sensing* to involve the user in the sensing task (*e.g.*, surveys), but also *opportunistic sensing* to exploit mobile sensors carried by the user (*e.g.*, smartphones).

While reported metrics generally enclose raw data, the monitoring layer intends to produce meaningful indicators like the *Quality of Experience* (QoE) perceived by users. This QoE reflects representative symptoms of software requiring to trigger appropriate decisions in order to improve its efficiency. To diagnose these symptoms, the system has to process a huge variety of data including runtime metrics, but also history of logs to explore the sources of the reported problems and identify opportunities for optimizations. The technics we envision at this level encompass machine learning, principal component analysis, and fuzzy logic [111] to provide enriched information to the decision level.

### 3.3.2. Collaborative decision-making approaches

Beyond the symptoms analysis, decisions should be taken in order to improve the *Quality of Service* (QoS). In our opinion, collaborative approaches represent a promising solution to effectively converge towards the most appropriate optimization to apply for a given symptom. In particular, we believe that exploiting the wisdom of the crowd can help the software to optimize itself by sharing its experience with other software instances exhibiting similar symptoms. The intuition here is that the body of knowledge that supports the optimization process cannot be specific to a single software instance as this would restrain the opportunities for improving the quality and the performance of applications. Rather, we think that any software instance can learn from the experience of others.

With regard to the state-of-the-art, we believe that a multi-levels decision infrastructure, inspired from distributed systems like Spotify [95], can be used to build a decentralized decision-making algorithm involving the surrounding peers before requesting a decision to be taken by more central control entity. In the context of collaborative decision-making, peer-based approaches therefore consist in quickly reaching a consensus on the decision to be adopted by a majority of software instances. Software instances can share their knowledge through a micro-economic model [89], that would weight the recommandations of experienced instances, assuming their age reflects an optimal configuration.

Beyond the peer level, the adoption of algorithms inspired from evolutionary computations, such as genetic programming, at an upper level of decision can offer an opportunity to test and compare several alternative decisions for a given symptom and to observe how does the crowd of applications evolves. By introducing some diversity within this population of applications, some instances will not only provide a satisfying QoS, but will also become naturally resilient to unforeseen situations.

### 3.3.3. Smart reconfigurations in the large

Any decision taken by the crowd requires to propagate back to and then operated by the software instances. While simplest decisions tend to impact software instances located on a single host (*e.g.*, laptop, smartphone),

this process can also exhibit more complex reconfiguration scenarios that require the orchestration of various actions that have to be safely coordinated across a large number of hosts. While it is generally acknowledged that centralized approaches raise scalability issues, we think that self-optimization should investigate different reconfiguration strategies to propagate and apply the appropriate actions. The investigation of such strategies can be addressed in two steps: the consideration of *scalable data propagation protocols* and the identification of *smart reconfiguration mechanisms*.

With regard to the challenge of scalable data propagation protocols, we think that research opportunities encompass not only the exploitation of gossip-based protocols [94], but also the adoption of publish/subscribe abstractions [101] in order to decouple the decision process from the reconfiguration. The fundamental issue here is the definition of a communication substrate that can accomodate the propagation of decisions with relaxed properties, inspired by *Delay Tolerant Networks* (DTN), in order to reach weakly connected software instances. We believe that the adoption of asynchronous communication protocols can provide the sustainable foundations for addressing various execution environments including harsh environments, such as developing countries, which suffer from a partial connectivity to the network. Additionally, we are interested in developing the principle of *social networks of applications* in order to seamlessly group and organize software instances according to their similarities and acquaintances. The underlying idea is that grouping application instances can contribute to the identification of optimization profiles not only contributing to the monitoring layer, but also interested in similar reconfigurations. Social networks of applications can contribute to the anticipation of reconfigurations by exploiting the symptoms of similar applications to improve the performance of others before that problems actually happen.

With regard to the challenge of smart reconfiguration mechanisms, we are interested in building on our established experience of adaptive middleware [8] in order to investigate novel approaches to efficient application reconfigurations. In particular, we are interested in adopting seamless micro-updates and micro-reboot technics to provide in-situ reconfiguration of pieces of software. Additionally, the provision of safe and secured reconfiguration mechanisms is clearly a key issue that requires to be carefully addressed in order to avoid malicious exploitation of dynamic reconfiguration mechanisms against the software itself. In this area, although some reconfiguration mechanisms integrate transaction models [102], most of them are restricted to local reconfigurations, without providing any support for executing distributed reconfiguration transactions. Additionally, none of the approached published in the literature include security mechanisms to preserve from unauthorized or malicious reconfigurations.