RESEARCH CENTRES

Inria Saclay Centre at Université Paris-Saclay

Inria Centre at Université Côte d'Azur

IN PARTNERSHIP WITH:

Université Paris-Saclay, CNRS

2024 ACTIVITY REPORT

Project-Team DATASHAPE

Understanding the shape of data

IN COLLABORATION WITH: Laboratoire de mathématiques d'Orsay de l'Université de Paris-Sud (LMO)

DOMAIN

Algorithmics, Programming, Software and Architecture

THEME

Algorithmics, Computer Algebra and Cryptology

Inria

Contents

Project-Team DATASHAPE

Creation of the Project-Team: 2020 October 01

Keywords

Computer sciences and digital sciences

A3. – Data and knowledge

A3.4. – Machine learning and statistics

A7.1. – Algorithms

A8. – Mathematics of computing

A8.1. – Discrete mathematics, combinatorics

A8.3. – Geometry, Topology

A9. – Artificial intelligence

Other research topics and application domains

B1. – Life sciences

B2. – Health

B5. – Industry of the future

B9. – Society and Knowledge

B9.5. – Sciences

1 Team members, visitors, external collaborators

Research Scientists

- Frederic Chazal [Team leader, INRIA, Senior Researcher]
- Charles Arnal [INRIA, Starting Research Position, until Oct 2024]
- Jean-Daniel Boissonnat [INRIA, Emeritus]
- Mathieu Carrière [INRIA, Researcher]
- David Cohen-Steiner [INRIA, Researcher]
- Marc Glisse [INRIA, Researcher]
- Clément Maria [INRIA, Researcher]
- Nina Lisann Otter [INRIA, ISFP]
- Mathijs Wintraecken [INRIA, ISFP]

Faculty Members

- Gilles Blanchard [UNIV PARIS SACLAY, Professor]
- Blanche Buet [UNIV PARIS SACLAY, Associate Professor]
- Rémi Leclercq [UNIV PARIS SACLAY, Associate Professor, from Oct 2024]
- Pierre Pansu [UNIV PARIS SACLAY, Emeritus, from Oct 2024]
- Pierre Pansu [UNIV PARIS SACLAY, Professor, until Sep 2024]

Post-Doctoral Fellows

- Daniele Cannarsa [INRIA, Post-Doctoral Fellow, from Sep 2024]
- Francesco Conti [INRIA, Post-Doctoral Fellow, from Sep 2024]
- Corentin Lunel [INRIA, Post-Doctoral Fellow, from Oct 2024]
- Renata Turkes [INRIA, Post-Doctoral Fellow, from Jun 2024]

PhD Students

- Charly Boricaud [UNIV. PARIS SACLAY, until Sep 2024]
- Charly Boricaud [UNIV PARIS SACLAY, ATER, from Sep 2024]
- Jeremie Capitao-Miniconi [UNIV PARIS SACLAY, until May 2024]
- Antoine Commaret [UNIV COTE D'AZUR, from Dec 2024, ingénieur de recherche Nematic team at the LJAD]
- Antoine Commaret [INRIA, from Sep 2024 until Nov 2024]
- Antoine Commaret [UNIV COTE D'AZUR, until Aug 2024]
- Bastien Dussap [UNIV PARIS SACLAY, until Sep 2024]
- Myriam Frikha [ERICSSON, CIFRE, from Oct 2024]
- Alexandre Guerin [Sysnav, until Aug 2024]

- Hugo Henneuse [UNIV PARIS SACLAY]
- Antonio Lage De Sousa Leitao [Scuola Normale Superiore di Pisa, from Nov 2024]
- David Loiseaux [INRIA]
- Henrique Lovisi Ennes [UNIV COTE D'AZUR]

Technical Staff

- Vincent Rouvreau [INRIA, Engineer]
- Hannah Schreiber [INRIA, Engineer]

Interns and Apprentices

- Sean Bontemps [INRIA, Intern, from Aug 2024 until Aug 2024]
- Ezechiel Jimenez [UNIV PARIS SACLAY, Intern, from Mar 2024 until Aug 2024]
- Jérôme Taupin [ENS Paris, Intern, from Sep 2024]
- Jérôme Taupin [ENS PARIS, Intern, from Apr 2024 until Jul 2024]

Administrative Assistants

- Aissatou-Sadio Diallo [INRIA]
- Sophie Honnorat [INRIA]

Visiting Scientists

- Marzieh Eidi [Max Planck Institute for Mathematics in the Sciences, Leipzig, from Oct 2024]
- Musashi Koyama [Australian National University, from Oct 2024 until Oct 2024, as part of TopTime EA]
- Marina Meila-Predoviciu [UNIV WASHINGTON ST-LOUIS, from Oct 2024]
- Adam Onus [Queen Mary University of London, from Oct 2024 until Oct 2024]
- Matteo Pegoraro [UNIV AALBORG, from Mar 2024 until Apr 2024]

External Collaborator

- Bertrand Michel [CENTRALE NANTES]

2 Overall objectives

During the last two decades, building on solid theoretical and algorithmic bases, geometric inference and computational topology have experienced important developments towards data analysis. New mathematically well-founded theories gave birth to the field of Topological Data Analysis (TDA), which is now arousing interest from both academia and industry. Although one can trace back geometric approaches for data analysis quite far in the past, TDA really started as a field with the pioneering works of H. Edelsbrunner et al. and G. Carlsson et al. in persistent homology at the beginning of the century. TDAis mainly motivated by the idea that topology and geometry provide a powerful approach to infer robust qualitative, and sometimes quantitative, information about the structure of data. It aims at providing mathematical results and methods to infer, analyze and exploit complex data (point clouds, graphs,

images, 3D shapes, time series...). It also intends to give access to robust and efficient data structures and algorithms to represent these data and that are amenable to precise analysis.

The overall objective of DataShape is three-fold:

1. to settle the **mathematical, statistical and algorithmic foundations of** TDA, and, more generally to contribute to the development of topological and geometric approaches in Machine Learning and AI;

2. to develop a new family of well-founded and efficient data structures, algorithms and methods to uncover and exploit the geometry of data through the **development of a state-of-the-art and easy-to-use open source software**;

3. to disseminate and promote TDA research and outcomes among the data science community through **collaborations with other domains of science and industrials**.

The approach of DataShape relies on the conviction that, to reach these objectives, combining statistical, topological/geometric and computational approaches in a common framework is mandatory. For that purpose, DataShape became a joint team with the Laboratoire de Mathématiques d'Orsay in 2020 and now gathers a wide variety of expertise, going from fundamental mathematics to software development and industrial applications. The team also considers that TDA needs to be combined with other data sciences approaches and tools, in particular statistical learning, to lead to successful real applications. Significant efforts have been made during the evaluation period to develop several long term industrial research collaborations in data science and AI.

The research program of DataShape is organized around four strongly correlated axes reflecting our will to address TDA challenges in a global and unified framework.
The first axis focuses on *the algorithmic aspects of* TDA *and geometric inference* as well as the *mathematical foundations* of the fields. Fundamental problems are the construction, processing and analysis of discrete representations of complex and possibly high dimensional shapes.
The second axis is dedicated to *the statistical aspects of* TDA . It is dedicated to the study of the properties of topological information inferred from data from a statistical perspective and intends to propose new models and approaches for the development of TDA in well-founded probabilistic and statistical settings. This axis also includes the analysis and development of general-purpose statistical learning approaches and tools that are currently active in the community and of relevance for Datashape's scientific goals.
The third axis is driven by the problems raised by the use of *topological and geometric approaches in machine learning*. It aims at better understanding the role of topological and geometric structures in machine learning problems and at applying TDA tools to develop specialized topological approaches to be used in combination with other machine learning methods.
The fourth axis is dedicated to *software development and experimental research*, mainly through the GUDHI platform. GUDHI is intended to provide a high quality state-of-the-art implementation of data structures and algorithms dedicated to TDAthrough an easy-to-use open source software.
Each DATASHAPE member is involved in several research axes ensuring strong connections and interactions between them. Last, although the above 4 axes concentrate the main research activities of the team, DATASHAPE always remains open and encourages its members to explore new directions and approaches related to geometric and topological methods in data analysis and machine learning. The past experience of the team has shown that such a strategy is often very fruitful and may lead to innovative and new research directions.

3 Research program

3.1 Algorithmic aspects and new mathematical directions for topological and geometric data analysis

TDA requires to construct and manipulate appropriate representations of complex and high dimensional shapes. A major difficulty comes from the fact that the complexity of data structures and algorithms used

to approximate shapes rapidly grows as the dimensionality increases, which makes them intractable in high dimensions. We focus our research on simplicial complexes which offer a convenient representation of general shapes and generalize graphs and triangulations. Our work includes the study of simplicial complexes with good approximation properties and the design of compact data structures to represent them.

In low dimensions, effective shape reconstruction techniques exist that can provide precise geometric approximations very efficiently and under reasonable sampling conditions. Extending those techniques to higher dimensions as is required in the context of TDA is problematic since almost all methods in low dimensions rely on the computation of a subdivision of the ambient space. A direct extension of those methods would immediately lead to algorithms whose complexities depend exponentially on the ambient dimension, which is prohibitive in most applications. A first direction to by-pass the curse of dimensionality is to develop algorithms whose complexities depend on the intrinsic dimension of the data (which most of the time is small although unknown) rather than on the dimension of the ambient space. Another direction is to resort to cruder approximations that only captures the homotopy type or the homology of the sampled shape. The recent theory of persistent homology provides a powerful and robust tool to study the homology of sampled spaces in a stable way.

3.2 Statistical aspects of topological and geometric data analysis

The wide variety of larger and larger available data - often corrupted by noise and outliers - requires to consider the statistical properties of their topological and geometric features and to propose new relevant statistical models for their study.

There exist various statistical and machine learning methods intending to uncover the geometric structure of data. Beyond manifold learning and dimensionality reduction approaches that generally do not allow to assert the relevance of the inferred topological and geometric features and are not well-suited for the analysis of complex topological structures, set estimation methods intend to estimate, from random samples, a set around which the data is concentrated. In these methods, that include support and manifold estimation, principal curves/manifolds and their various generalizations to name a few, the estimation problems are usually considered under losses, such as Hausdorff distance or symmetric difference, that are not sensitive to the topology of the estimated sets, preventing these tools to directly infer topological or geometric information.

Regarding purely topological features, the statistical estimation of homology or homotopy type of compact subsets of Euclidean spaces, has only been considered recently, most of the time under the quite restrictive assumption that the data are randomly sampled from smooth manifolds.

In a more general setting, with the emergence of new geometric inference tools based on the study of distance functions and algebraic topology tools such as persistent homology, computational topology has recently seen an important development offering a new set of methods to infer relevant topological and geometric features of data sampled in general metric spaces. The use of these tools remains widely heuristic and until recently there were only a few preliminary results establishing connections between geometric inference, persistent homology and statistics. However, this direction has attracted a lot of attention over the last three years. In particular, stability properties and new representations of persistent homology information have led to very promising results to which the DATASHAPE members have significantly contributed. These preliminary results open many perspectives and research directions that need to be explored.

Our goal is to build on our first statistical results in TDA to develop the mathematical foundations of Statistical Topological and Geometric Data Analysis. Combined with the other objectives, our ultimate goal is to provide a well-founded and effective statistical toolbox for the understanding of topology and geometry of data.

3.3 Topological and geometric approaches for machine learning

This objective is driven by the problems raised by the use of topological and geometric approaches in machine learning. The goal is both to use our techniques to better understand the role of topological and geometric structures in machine learning problems and to apply our TDA tools to develop specialized topological approaches to be used in combination with other machine learning methods.

3.4 Experimental research and software development

We develop a high quality open source software platform called GUDHI which is becoming a reference in geometric and topological data analysis in high dimensions. The goal is not to provide code tailored to the numerous potential applications but rather to provide the central data structures and algorithms that underlie applications in geometric and topological data analysis.

The development of the GUDHI platform also serves to benchmark and optimize new algorithmic solutions resulting from our theoretical work. Such development necessitates a whole line of research on software architecture and interface design, heuristics and fine-tuning optimization, robustness and arithmetic issues, and visualization. We aim at providing a full programming environment following the same recipes that made up the success story of the CGAL library, the reference library in computational geometry.

Some of the algorithms implemented on the platform will also be interfaced to other software platforms, such as the R software for statistical computing, and languages such as Python in order to make them usable in combination with other data analysis and machine learning tools. A first attempt in this direction has been done with the creation of an R package called TDA in collaboration with the group of Larry Wasserman at Carnegie Mellon University (Inria Associated team CATS) that already includes some functionalities of the GUDHI library and implements some joint results between our team and the CMU team. A similar interface with the Python language is also considered a priority. To go even further towards helping users, we will provide utilities that perform the most common tasks without requiring any programming at all.

4 Application domains

Our work is mostly of a fundamental mathematical and algorithmic nature but finds a variety of applications in data analysis, e.g., in material science, biology, sensor networks, 3D shape analysis and processing, to name a few.

More specifically, DATASHAPE has developed and is still developing a strong expertise on new TDA methods for Machine Learning and Artificial Intelligence for complex data and (complex) time-dependent data. This includes, for example:

- the analysis of high dimensional point cloud data (PhD of Bastien Dussap with Metafora),
- the analysis of trajectories obtained from inertial sensors (PhD theses of Wojtek Riese and Alexandre Guérin with Sysnav),
- domain adaptation problems for time series (PhD of Myriam Frikha with Ericsson),
- anomaly detection (with IRT Systemx and Confiance.AI program),
- the statistical significance of biological phenomena (cell cycle, stem cell differentiation, immune system responses) that occur in large scale single-cell RNAseq and spatial transcriptomics data sets (collaboration with Rabadan Lab, Columbia University),
- the analysis of gene regulatory networks for plant-pathogen interactions (collaboration with INRAE),
- the analysis of satellite imaging and cartography data sets (collaboration with Thalès Alenia Space).

5 Social and environmental responsibility

5.1 Footprint of research activities

The weekly research seminar of DATASHAPE is now taking place in hybrid mode. The travels for the team members have decreased a lot these years to take care of the environmental footprint of the team.

6 Highlights of the year

6.1 Events

- We organized a one week team workshop in the week of the 29th of April 2024, giving the opportunity to all the PhD students, post-doc and researchers of the team to present their work and discuss scientific questions all together.

6.2 PhD defenses

- Antoine Commaret, supervised by David Cohen-Steiner and Indira Chatterji. December 20th, 2024.
- David Loiseaux, supervised by Mathieu Carrière and Frédéric Cazals. December 6th, 2024.
- Bastien Dussap, supervised by Gilles Blanchard and Marc Glisse. October 1st, 2024.

7 New software, platforms, open data

7.1 New software

7.1.1 GUDHI

Name: Geometric Understanding in Higher Dimensions

Keywords: Computational geometry, Topology, Clustering

Scientific Description: The Gudhi library is an open source library for Computational Topology and Topological Data Analysis (TDA). It offers state-of-the-art algorithms to construct various types of simplicial complexes, data structures to represent them, and algorithms to compute geometric approximations of shapes and persistent homology.

The GUDHI library offers the following interoperable modules:

. Complexes: + Cubical + Simplicial: Rips, Witness, Alpha and Čech complexes + Cover: Nerve and Graph induced complexes . Data structures and basic operations: + Simplex tree, Skeleton blockers and Toplex map + Construction, update, filtration and simplification . Topological descriptors computation . Manifold reconstruction . Topological descriptors tools: + Bottleneck and Wasserstein distance + Statistical tools + Persistence diagram and barcode

Functional Description: The GUDHI open source library will provide the central data structures and algorithms that underly applications in geometry understanding in higher dimensions. It is intended to both help the development of new algorithmic solutions inside and outside the project, and to facilitate the transfer of results in applied fields.

News of the Year: Below is a list of changes made since GUDHI 3.9.0:

- Persistence matrix > Matrix API is in a beta version and may change in incompatible ways in the near future. . Matrix structure for filtered complexes with multiple functionnalities related to persistence homology, such as representative cycles computation or vineyards.

- Rips complex . Rips complex persistence scikit-learn like interface

- Čech complex . A new utility to compute the Delaunay-Čech filtration on a Delaunay triangulation.

URL: https://gudhi.inria.fr/

Publication: hal-01108461

Contact: Marc Glisse

Participants: Clément Maria, François Godi, David Salinas, Jean-Daniel Boissonnat, Marc Glisse, Marietta Yvinec, Pawel Dlotko, Siargey Kachanovich, Vincent Rouvreau, Mathieu Carrière, Clément Jamin, Siddharth Pritam, Frederic Chazal, Steve Oudot, Wojciech Reise, Hind Montassif, Hannah Schreiber, Martin Royer, David Loiseaux

Partners: Université Côte d'Azur (UCA), Fujitsu

7.1.2 Multipers

Name: Multiparameter Persistence for Machine Learning

Keywords: Topology, Machine learning

Functional Description: multipers is a Python library for Topological Data Analysis, focused on Multiparameter Persistence computation and visualizations for Machine Learning. It features several efficient computational and visualization tools, with integrated, easy to use, auto-differentiable Machine Learning pipelines, that can be seamlessly interfaced with scikit-learn and PyTorch. This library is meant to be usable for non-experts in Topological or Geometrical Machine Learning. Performance-critical functions are implemented in C++ or in Cython, are parallelizable with TBB, and have Python bindings and interface. It can handle a very diverse range of datasets that can be framed into a (finite) multi-filtered simplicial or cell complex, including, e.g., point clouds, graphs, time series, images, etc.

URL: https://davidlapous.github.io/multipers/

Publication: hal-04801544

Contact: David Loiseaux

Participants: David Loiseaux, Hannah Schreiber

8 New results

8.1 Algorithmic aspects and new mathematical directions for topological and geometric data analysis

8.1.1 Generalized Morse theory for tubular neighborhoods

Participant: Antoine Commaret.

We define a notion of Morse function and establish Morse theory-like theorems over offsets of any compact set in a Euclidean space at regular values of their distance function. Using non-smooth analysis and tools from geometric measure theory, we prove that the homotopy type of the sublevels sets of these Morse functions changes at a critical value by gluing exactly one cell around each critical point.

8.1.2 Persistent Intrinsic Volumes

Participant: David Cohen-Steiner, Antoine Commaret.

We develop a new method to estimate the area, and more generally the intrinsic volumes, of a compact subset X of $\mathbb{R}^d$ from a set Y that is close in the Hausdorff distance. This estimator enjoys a linear rate of convergence as a function of the Hausdorff distance under mild regularity conditions on X. Our approach combines tools from both geometric measure theory and persistent homology, extending the noise filtering properties of persistent homology from the realm of topology to geometry. Along the way, we obtain a stability result for intrinsic volumes.

8.1.3 Critical points of the distance function to a generic submanifold

Participant: Charles Arnal, David Cohen-Steiner.

In collaboration with Vincent Divol (CEREMADE)

In general, the critical points of the distance function d_M to a compact submanifold $M \subset \mathbb{R}^D$ can be poorly behaved. In this article, we show that this is generically not the case by listing regularity conditions on the critical and µ-critical points of a submanifold and by proving that they are generically satisfied and stable with respect to small C^2 perturbations. More specifically, for any compact abstract manifold M , the set of embeddings $i : M \to \mathbb{R}^D$ such that the submanifold i(M) satisfies those conditions is open and dense in the Whitney C^2-topology. When those regularity conditions are fulfilled, we prove that the critical points of the distance function to an ε-dense subset of the submanifold (e.g. obtained via some sampling process) are well-behaved. We also provide many examples that showcase how the absence of these conditions can result in pathological cases.

8.1.4 On Edge Collapse of Random Simplicial Complexes

Participant: Jean-Daniel Boissonnat.

In collaboration with Kunal Dutta (Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland), Soumik Dutta (Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland), and Siddharth Pritam (Chennai Mathematical Institute, India).

We consider the edge collapse (introduced in [66]) process on the Erdős-Rényi random clique complex $X(n, c/\sqrt{n})$ on n vertices with edge probability $c/\sqrt{n}$ such that $c > \sqrt{\eta_2}$, where

$$\eta_2 = \inf\{\eta | x = e^{-\eta(1-x^2)} \; has\; a\; solution\; in\; (0,1)\}.$$

For a given $c > \sqrt{\eta_2}$, we show that after t iterations of maximal edge collapsing phases, the remaining subcomplex, or t-core, has at most

$$(1+o(1))\binom{n}{2}(1-c^2/3)(1-(1-\gamma_t)^3))$$

and at least

$$(1+o(1))\binom{n}{2}p(1-c^2/3)(1-\gamma_{t+1}-c^2(1-\gamma_t)^2)$$

edges asymptotically almost surely (a.a.s.), where $\{\gamma_t\}_{|t\geq 0}$ is recursively determined by $\gamma_{t+1} = e^{-c^2(1-\gamma_t)^2}$ and $\gamma_0 = 0$. We also determine the upper and lower bound on the final core with explicit formulas. If $c < \sqrt{\eta_2}$, then we show that the final core contains $o(n\sqrt{n})$ edges. On the other hand, if, instead of c being a constant with respect to n, $c > \sqrt{2\log n}$, then the edge collapse process is no more effective in reducing the size of the complex. Our proof is based on the notion of local weak convergence [65] together with two new components. Firstly, we identify the critical combinatorial structures that control the outcome of the edge collapse process. By controlling the expected number of these structures during the edge collapse process, we establish a.a.s. bounds on the size of the core. We also give a new concentration inequality for typically Lipschitz functions on random graphs which improves on the bound of [71] and is, therefore, of independent interest. The proof of our lower bound is via the recursive technique of [68] to simulate cycles in infinite trees. These are the first theoretical results proved for edge collapses on random (or non-random) simplicial complexes.

8.1.5 On Edge Collapse of Random Simplicial Complexes

Participant: Jean-Daniel Boissonnat.

In collaboration with Kunal Dutta (Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland.

Computing persistent homology of large datasets using Gaussian kernels is useful in the domains of topological data analysis and machine learning as shown by Phillips, Wang and Zheng [69]. However, unlike in the case of persistent homology computation using the Euclidean distance or the k-distance, using Gaussian kernels involves significantly higher overhead, as all distance computations are in terms of the Gaussian kernel distance which is computationally more expensive. Further, most algorithmic implementations (e.g. Gudhi, Ripser, etc.) are based on Euclidean distances, so the question of finding a Euclidean embedding -preferably low-dimensional -that preserves the persistent homology computed with Gaussian kernels, is quite important. We consider the Gaussian kernel power distance (GKPD) given by Phillips, Wang and Zheng. Given an n-point dataset and a relative error parameter $\epsilon \in (0,1]$, we show that the persistent homology of the Čech filtration of the dataset computed using the GKPD can be approximately preserved using $\mathcal{O}(\epsilon^{-2}\log n)$ dimensions, under a high stable rank condition. Our results also extend to the Delaunay filtration and the (simpler) case of the weighted Rips filtrations constructed using the GKPD. Compared to the Euclidean embedding for the Gaussian kernel function in $\sim n$ dimensions, which uses the Cholesky decomposition of the matrix of the kernel function applied to all pairs of data points, our embedding may also be viewed as dimensionality reduction -reducing the dimensionality from n to $\sim \log n$ dimensions.

Our proof utilizes the embedding of Chen and Phillips [67], based on the Random Fourier Functions of Rahimi and Recht [70], together with two novel ingredients. The first one is a new decomposition of the squared radii of Čech simplices computed using the GKPD, in terms of the pairwise GKPDs between the vertices, which we state and prove. The second is a new concentration inequality for sums of cosine functions of Gaussian random vectors, which we call Gaussian cosine chaoses. We believe these are of independent interest and will find other applications in future.

8.1.6 Sparsification of the generalized persistence diagrams for scalability through gradient descent

Participant: Mathieu Carrière.

In collaboration with Seunghyun Kim, Woojin Kim (KAIST, South Korea)

The generalized persistence diagram (GPD) is a natural extension of the classical persistence barcode to the setting of multi-parameter persistence and beyond. The GPD is defined as an integer-valued function whose domain is the set of intervals in the indexing poset of a persistence module, and is known to be able to capture richer topological information than its single-parameter counterpart. However, computing the GPD is computationally prohibitive due to the sheer size of the interval set. Restricting the GPD to a subset of intervals provides a way to manage this complexity, compromising discriminating power to some extent. However, identifying and computing an effective restriction of the domain that minimizes the loss of discriminating power remains an open challenge.

In this work, we introduce a novel method for optimizing the domain of the GPD through gradient descent optimization. To achieve this, we introduce a loss function tailored to optimize the selection of intervals, balancing computational efficiency and discriminative accuracy. The design of the loss function is based on the known erosion stability property of the GPD. We showcase the efficiency of our sparsification method for dataset classification in supervised machine learning. Experimental results demonstrate that our sparsification method significantly reduces the time required for computing the GPDs associated to several datasets, while maintaining classification accuracies comparable to

those achieved using full GPDs. Our method thus opens the way for the use of GPD-based methods to applications at an unprecedented scale.

8.1.7 multipers: Multiparameter Persistence for Machine Learning

Participant: David Loiseaux, Hannah Schreiber.

multipers is a Python library for Topological Data Analysis, focused on Multiparameter Persistence computation and visualizations for Machine Learning. It features several efficient computational and visualization tools, with integrated, easy to use, auto-differentiable Machine Learning pipelines, that can be seamlessly interfaced with scikit-learn and PyTorch. This library is meant to be usable for non-experts in Topological or Geometrical Machine Learning. Performance-critical functions are implemented in C++ or in Cython, are parallelizable with TBB, and have Python bindings and interface. It can handle a very diverse range of datasets that can be framed into a (finite) multi-filtered simplicial or cell complex, including, e.g., point clouds, graphs, time series, images, etc.

8.1.8 Brillouin Zones of Integer Lattices and Their Perturbations

Participant: Mathijs Wintraecken.

In collaboration with Herbert Edelsbrunner (Institute of Science and Technology Austria) , Alexey Garber (Department of Mathematics, The University of Texas at Brownsville) , Mohadese Ghafari (Northeastern University [Boston]) , Teresa Heiss (Institute of Science and Technology Austria) , Morteza Saghafian (Institute of Science and Technology Austria)

For a locally finite set, $A \subseteq \mathbb{R}^d$, the kth Brillouin zone of $a \in A$ is the region of points $x \in \mathbb{R}^d$ for which $\|x-a\|$ is the kth smallest among the Euclidean distances between x and the points in A. If A is a lattice, the kth Brillouin zones of the points in A are translates of each other, and together they tile space. Depending on the value of k, they express medium- or long-range order in the set. In [12], we study fundamental geometric and combinatorial properties of Brillouin zones, focusing on the integer lattice and its perturbations. Our results include the stability of a Brillouin zone under perturbations, a linear upper bound on the number of chambers in a zone for lattices in $\mathbb{R}^2$, and the convergence of the maximum volume of a chamber to zero for the integer lattice.

8.1.9 Tight Bounds for the Learning of Homotopy à la Niyogi, Smale, and Weinberger for Subsets of Euclidean Spaces and of Riemannian Manifolds.

Participant: Mathijs Wintraecken.

In collaboration with Dominique Attali (Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab), Hana Dal Poz Kouřimská (Institute of Science and Technology Austria), Christopher Fillmore (Institute of Science and Technology Austria), Ishika Ghosh (Institute of Science and Technology Austria, Michigan State University [East Lansing]), André Lieutier (No Affiliation), Elizabeth Stephenson (Institute of Science and Technology Austria)

In [22] we extend and strengthen the seminal work by Niyogi, Smale, and Weinberger on the learning of the homotopy type from a sample of an underlying space. In their work, Niyogi, Smale, and Weinberger studied samples of C^2 manifolds with positive reach embedded in $\mathbb{R}^d$. We extend their results in the following ways:

- As the ambient space we consider both $\mathbb{R}^d$ and Riemannian manifolds with lower bounded sectional curvature.
- In both types of ambient spaces, we study sets of positive reach — a significantly more general setting than C^2 manifolds — as well as general manifolds of positive reach.
- The sample P of a set (or a manifold) $\mathscr{S}$ of positive reach may be noisy. We work with two one-sided Hausdorff distances — ε and δ — between P and $\mathscr{S}$. We provide tight bounds in terms of ε and δ, that guarantee that there exists a parameter r such that the union of balls of radius r centred at the sample P deformation-retracts to $\mathscr{S}$. We exhibit their tightness by an explicit construction.

We carefully distinguish the roles of δ and ε. This is not only essential to achieve tight bounds, but also sensible in practical situations, since it allows one to adapt the bound according to sample density and the amount of noise present in the sample separately.

8.1.10 The Ultimate Frontier: An Optimality Construction for Homotopy Inference (Media Exposition)

Participant: Mathijs Wintraecken.

In collaboration with Dominique Attali (Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab), Hana Dal Poz Kouřimská (Institute of Science and Technology Austria), Christopher Fillmore (Institute of Science and Technology Austria), Ishika Ghosh (Institute of Science and Technology Austria, Michigan State University [East Lansing]), André Lieutier (No Affiliation), Elizabeth Stephenson (Institute of Science and Technology Austria)

In our companion paper 'Tight bounds for the learning of homotopy à la Niyogi, Smale, and Weinberger for subsets of Euclidean spaces and of Riemannian manifolds' [22] we gave optimal bounds (in terms of the two one-sided Hausdorff distances) on a sample P of an input shape $\mathscr{S}$ (either manifold or general set with positive reach) such that one can infer the homotopy of $\mathscr{S}$ from the union of balls with some radius centred at P, both in Euclidean space and in a Riemannian manifold of bounded curvature. The construction showing the optimality of the bounds is not straightforward. The purpose of this video [63] is to visualize and thus elucidate said construction in the Euclidean setting.

Note that a media contribution consists of a small description [21] (which is part of the proceedings of the conference) and the video itself [63].

8.1.11 The Medial Axis of Any Closed Bounded Set Is Lipschitz Stable with Respect to the Hausdorff Distance Under Ambient Diffeomorphisms

Participant: Mathijs Wintraecken.

In collaboration with Hana Dal Poz Kouřimská (Institute of Science and Technology Austria), André Lieutier (No Affiliation)

In [28] we prove that the medial axis of closed sets is Hausdorff stable in the following sense: Let $\mathscr{S} \subseteq \mathbb{R}^d$ be a fixed closed set that contains a bounding sphere. That is, the bounding sphere is part of the set $\mathscr{S}$. Consider the space of $C^{1,1}$ diffeomorphisms of $\mathbb{R}^d$ to itself, which keep the bounding sphere invariant. The map from this space of diffeomorphisms (endowed with a Banach norm) to the space of closed subsets of $\mathbb{R}^d$ (endowed with the Hausdorff distance), mapping a diffeomorphism F to the closure of the medial axis of $F(\mathscr{S})$, is Lipschitz. This extends a previous stability result of Chazal and Soufflet on the stability of the medial axis of C^2 manifolds under C^2 ambient diffeomorphisms.

8.1.12 The distance function to a finite set is a topological Morse function

Participant: Charles Arnal.

In [34], we show that the distance function to any finite set $X \subset \mathbb{R}^n$ is a topological Morse function, regardless of whether X is in general position. We also precisely characterize its topological critical points and their indices, and relate them to the differential critical points of the function.

8.1.13 Distance-from-flat persistent homology transform

Participant: Nina Otter, Renata Turkeš.

In collaboration with Adam Onus (Queen Mary University of London).

In [60] we introduce a generalisation of the persistent homology transform (PHT) in which we consider arbitrary parameter spaces and sublevel sets with respect to any function. In particular, we study transforms, defined on the Grassmannian $\mathbb{AG}(m,n)$ of affine subspaces of $\mathbb{R}^n$, which allow to scan a shape by probing it with all possible affine m-dimensional subspaces $P \subset \mathbb{R}^n$, for fixed dimension m, and by then computing persistent homology of sublevel set filtrations of the function $\mathrm{dist}(\cdot, P)$ encoding the distance from the flat P. We call such transforms "distance-from-flat" PHTs. We show that these transforms are injective and continuous and that they provide computational advantages over the classical PHT. In particular, we show that it is enough to compute homology only in degrees up to $m-1$ to obtain injectivity; for $m = 1$ this provides a very powerful and computationally advantageous tool for examining shapes, which in a previous work by Nina Otter and Renata Turkeš has proven to significantly outperform state-of-the-art neural networks for shape classification tasks.

8.1.14 Time-optimal persistent homology representatives for univariate time series

Participant: António Leitão, Nina Otter.

In [58] we introduce time-optimal PH representatives for time-varying data, that allow one to extract representatives that are close in time in an appropriate sense. We illustrate our methods on quasi-periodic synthetic time series, as well as time series arising from climate models, and we show that our methods provide optimal PH representatives that are better suited for these types of problems than existing optimality notions, such as length-optimal PH representatives.

8.2 Statistical aspects of topological and geometric data analysis

8.2.1 Wasserstein convergence of Čech persistence diagrams for samplings of submanifolds

Participant: Charles Arnal, David Cohen-Steiner.

In collaboration with Vincent Divol (CEREMADE, Université Paris-Dauphine PSL)

Čech Persistence diagrams (PDs) are topological descriptors routinely used to capture the geometry of complex datasets. They are commonly compared using the p-Wasserstein distances; however, the extent to which PDs are stable with respect to these metrics remains poorly understood. We partially close this gap by focusing on the case where datasets are sampled on an m-dimensional submanifold of $\mathbb{R}^d$. Under

this manifold hypothesis, we show that convergence with respect to the p-Wasserstein metric happens exactly when $p > m$. We also provide improvements upon the bottleneck stability theorem in this case and prove new laws of large numbers for the total α-persistence of PDs. Finally, we show how these theoretical findings shed new light on the behavior of the feature maps on the space of PDs that are used in ML-oriented applications of Topological Data Analysis.

8.2.2 Statistical estimation of sparsity and efficiency for molecular codes

Participant: Mathieu Carrière.

In collaboration with Jun Hou Fung, Andrew Blumberg (Columbia University, USA)

A fundamental biological question is to understand how cell types and functions are determined by genomic and proteomic coding. A basic form of this question is to ask if small families of genes or proteins code for cell types. For example, it has been shown that the collection of homeodomain proteins can uniquely delineate all 118 neuron classes in the nematode C. elegans. However, unique characterization is neither robust nor rare. Our goal in this paper is to develop a rigorous methodology to characterize molecular codes.

We show that in fact for information-theoretic reasons almost any sufficiently large collection of genes is able to disambiguate cell types, and that this property is not robust to noise. To quantify the discriminative properties of a molecular codebook in a more refined way, we develop new statistics – partition cardinality and partition entropy – borrowing ideas from coding theory. We prove these are robust to data perturbations, and then apply these in the C. elegans example and in cancer. In the worm, we show that the homeodomain transcription factor family is distinguished by coding for cell types sparsely and efficiently compared to a control of randomly selected family of genes. Furthermore, the resolution of cell type identities defined using molecular features increases as the worm embryo develops. In cancer, we perform a pan-cancer study where we use our statistics to quantify interpatient tumor heterogeneity and we identify the chromosome containing the HLA family as sparsely and efficiently coding for melanoma.

8.2.3 Subsampling, aligning, and averaging to find circular coordinates in recurrent time series

Participant: Mathieu Carrière.

In collaboration with Jun Hou Fung, Andrew Blumberg (Columbia University, USA) and Michael Mandell (Indiana University, USA)

We introduce a new algorithm for finding robust circular coordinates on data that is expected to exhibit recurrence, such as that which appears in neuronal recordings of C. elegans. Techniques exist to create circular coordinates on a simplicial complex from a dimension 1 cohomology class, and these can be applied to the Rips complex of a dataset when it has a prominent class in its dimension 1 cohomology. However, it is known this approach is extremely sensitive to uneven sampling density.

Our algorithm comes with a new method to correct for uneven sampling density, adapting our prior work on averaging coordinates in manifold learning. We use rejection sampling to correct for inhomogeneous sampling and then apply Procrustes matching to align and average the subsamples. In addition to providing a more robust coordinate than other approaches, this subsampling and averaging approach has better efficiency.

We validate our technique on both synthetic data sets and neuronal activity recordings. Our results reveal a topological model of neuronal trajectories for C. elegans that is constructed from loops in which different regions of the brain state space can be mapped to specific and interpretable macroscopic behaviors in the worm.

8.2.4 Resampling and averaging coordinates on data

Participant: Mathieu Carrière.

In collaboration with Jun Hou Fung, Andrew Blumberg (Columbia University, USA) and Michael Mandell (Indiana University, USA)

We introduce algorithms for robustly computing intrinsic coordinates on point clouds. Our approach relies on generating many candidate coordinates by subsampling the data and varying hyperparameters of the embedding algorithm (e.g., manifold learning). We then identify a subset of representative embeddings by clustering the collection of candidate coordinates and using shape descriptors from topological data analysis. The final output is the embedding obtained as an average of the representative embeddings using generalized Procrustes analysis. We validate our algorithm on both synthetic data and experimental measurements from genomics, demonstrating robustness to noise and outliers.

8.2.5 Topological Analysis for Detecting Anomalies (TADA) in Time Series

Participant: Frédéric Chazal.

In collaboration with Clément Levrard (Univ.Rennes) and Martin Royer (IRT System X).

In [11], we introduce a new methodology based on the field of Topological Data Analysis for detecting anomalies in multivariate time series, that aims to detect global changes in the dependency structure between channels. The proposed approach is lean enough to handle large scale datasets, and extensive numerical experiments back the intuition that it is more suitable for detecting global changes of correlation structures than existing methods. Some theoretical guarantees for quantization algorithms based on dependent time sequences are also provided.

8.2.6 Topological signatures of periodic-like signals

Participant: Wojciech Riese, Frédéric Chazal.

In collaboration with Bertrand Michel (Ecole Centrale Nantes)

In [19], we present a method to construct signatures of periodic-like data. Based on topological considerations, our construction encodes information about the order and values of local extrema. Its main strength is robustness to reparametrisation of the observed signal, so that it depends only on the form of the periodic function. The signature converges as the observation contains increasingly many periods. We show that it can be estimated from the observation of a single time series using bootstrap techniques.

8.2.7 Support and distribution inference from noisy data

Participant: Jérémie Capitao-Miniconi.

In collaboration E. Gassiat (LMO, Univ. Paris-Saclay) and L Lehéricy (Univ. Côte d'Azur)

In [46], we consider noisy observations of a distribution with unknown support. In the deconvolution model, it has been proved recently that, under very mild assumptions, it is possible to solve the deconvolution problem without knowing the noise distribution and with no sample of the noise. We first give

general settings where the theory applies and provide classes of supports that can be recovered in this context. We then exhibit classes of distributions over which we prove adaptive minimax rates (up to a log log factor) for the estimation of the support in Hausdorff distance. Moreover, for the class of distributions with compact support, we provide estimators of the unknown (in general singular) distribution and prove maximum rates in Wasserstein distance. We also prove an almost matching lower bound on the associated minimax risk.

8.2.8 Deconvolution of repeated measurements corrupted by unknown noise

Participant: Jérémie Capitao-Miniconi.

In collaboration E. Gassiat (LMO, Univ. Paris-Saclay) and L Lehéricy (Univ. Côte d'Azur)

Recent advances have demonstrated the possibility of solving the deconvolution problem without prior knowledge of the noise distribution. In [45], we study the repeated measurements model, where information is derived from multiple measurements of X perturbed independently by additive errors. Our contributions include establishing identifiability without any assumption on the noise except for coordinate independence. We propose an estimator of the density of the signal for which we provide rates of convergence, and prove that it reaches the minimax rate in the case where the support of the signal is compact. Additionally, we propose a model selection procedure for adaptive estimation. Numerical simulations demonstrate the effectiveness of our approach even with limited sample sizes.

8.2.9 A Geometric Approach for Multivariate Jumps Detection

Participant: Hugo Henneuse.

[53] addresses the inference of jumps (i.e. sets of discontinuities) within multivariate signals from noisy observations in the non-parametric regression setting. Departing from standard analytical approaches, we propose a new framework, based on geometric control over the set of discontinuities. This allows to consider larger classes of signals, of any dimension, with potentially wild discontinuities (exhibiting, for example, self-intersections and corners). We study a simple estimation procedure relying on histogram differences and show its consistency and near-optimality for the Hausdorff distance over these new classes. Furthermore, exploiting the assumptions on the geometry of jumps, we design procedures to infer consistently the homology groups of the jumps locations and the persistence diagrams from the induced offset filtration.

8.2.10 Persistence Diagram Estimation : Beyond Plug-in Approaches

Participant: Hugo Henneuse.

Persistent homology is a tool from Topological Data Analysis (TDA) used to summarize the topology underlying data. It can be conveniently represented through persistence diagrams. Observing a noisy signal, common strategies to infer its persistence diagram involve plug-in estimators, and convergence properties are then derived from sup-norm stability. This dependence on the sup-norm convergence of the preliminary estimator is restrictive, as it essentially imposes to consider regular classes of signals. Departing from these approaches, in [54] we design an estimator based on image persistence. In the context of the Gaussian white noise model, and for large classes of piecewise-constant signals, we prove that the proposed estimator is consistent and achieves parametric rates.

8.2.11 Persistence Diagram Estimation of Multivariate Piecewise Hölder-continuous Signals

Participant: Hugo Henneuse.

To our knowledge, the analysis of convergence rates for persistence diagram estimation from noisy signals had predominantly relied on lifting signal estimation results through sup norm (or other functional norm) stability theorems. We believe that moving forward from this approach can lead to considerable gains. In [55], we illustrate it in the setting of Gaussian white noise model. We examine from a minimax perspective, the inference of persistence diagram (for sublevel sets filtration). We show that for piecewise Hölder-continuous functions, with control over the reach of the discontinuities set, taking the persistence diagram coming from a simple histogram estimator of the signal, permit to achieve the minimax rates known for Hölder-continuous functions.

8.2.12 Persistence-based Modes Inference

Participant: Hugo Henneuse.

In [56], we address the problem of estimating multiple modes of a multivariate density, using persistent homology, a key tool from Topological Data Analysis. We propose a procedure, based on a preliminary estimation of the H_0-persistence diagram, to estimate the number of modes, their locations, and the associated local maxima. For large classes of piecewise-continuous functions, we show that these estimators are consistent and achieve nearly minimax rates. These classes involve geometric control over the set of discontinuities of the density, and differ from commonly considered function classes in mode(s) inference. Interestingly, we do not suppose regularity or even continuity in any neighborhood of the modes.

8.2.13 Transductive conformal inference with adaptive scores

Participant: Gilles Blanchard.

In collaboration with Ulysse Gazin (U. Paris-Sorbonne), Etienne Roquain (U. Paris-Sorbonne)

Conformal inference is a fundamental and versatile tool that provides distribution-free guarantees for many machine learning tasks. In [30], we consider the transductive setting, where decisions are made on a test sample of m new points, giving rise to m conformal p-values. While classical results only concern their marginal distribution, we show that their joint distribution follows a Pólya urn model, and establish a concentration inequality for their empirical distribution function. The results hold for arbitrary exchangeable scores, including adaptive ones that can use the covariates of the test+calibration samples at training stage for increased accuracy. We demonstrate the usefulness of these theoretical results through uniform, in-probability guarantees for two machine learning tasks of current interest: interval prediction for transductive transfer learning and novelty detection based on two-class classification.

8.2.14 Moment inequalities for sums of weakly dependent random fields

Participant: Gilles Blanchard.

In collaboration A. Carpentier (U. Potsdam), O. Zadorozhnyi (TU München)

In [8], we derive both Azuma-Hoeffding and Burkholder-type inequalities for partial sums over a rectangular grid of dimension d of a random field satisfying a weak dependency assumption of projective type: the difference between the expectation of an element of the random field and its conditional expectation given the rest of the field at a distance more than δ is bounded, in L_p distance, by a known decreasing function of δ. The analysis is based on the combination of a multi-scale approximation of random sums by martingale difference sequences, and of a careful decomposition of the domain. The obtained results extend previously known bounds under comparable hypotheses, and do not use the assumption of commuting filtrations.

8.3 Topological and geometric approaches for machine learning

8.3.1 Diffeomorphic interpolation for efficient persistence-based topological optimization

Participant: Mathieu Carrière.

In collaboration with Marc Theveneau (McGill University, Canada) and Théo Lacombe (Université Gustave Eiffel)

Topological Data Analysis (TDA) provides a pipeline to extract quantitative topological descriptors from structured objects. This enables the definition of topological loss functions, which assert to what extent a given object exhibits some topological properties. These losses can then be used to perform topological optimizationvia gradient descent routines. While theoretically sounded, topological optimization faces an important challenge: gradients tend to be extremely sparse, in the sense that the loss function typically depends on only very few coordinates of the input object, yielding dramatically slow optimization schemes in this http URL on the central case of topological optimization for point clouds, we propose in this work to overcome this limitation using diffeomorphic interpolation, turning sparse gradients into smooth vector fields defined on the whole space, with quantifiable Lipschitz constants. In particular, we show that our approach combines efficiently with subsampling techniques routinely used in TDA, as the diffeomorphism derived from the gradient computed on a subsample can be used to update the coordinates of the full input object, allowing us to perform topological optimization on point clouds at an unprecedented scale. Finally, we also showcase the relevance of our approach for black-box autoencoder (AE) regularization, where we aim at enforcing topological priors on the latent spaces associated to fixed, pre-trained, black-box AE models, and where we show thatlearning a diffeomorphic flow can be done once and then re-applied to new data in linear time (while vanilla topological optimization has to be re-run from scratch). Moreover, reverting the flow allows us to generate data by sampling the topologically-optimized latent space directly, yielding better interpretability of the model.

8.3.2 Differentiability and optimization of multiparameter persistent homology

Participant: Mathieu Carrière, David Loiseaux.

In collaboration with Siddharth Setlur (University of Edinburgh), Luis Scoccola (U. Sherbrooke) and Steve Oudot (Geomerix, Inria Saclay)

Real-valued functions on geometric data – such as node attributes on a graph – can be optimized using descriptors from persistent homology, allowing the user to incorporate topological terms in the loss function. When optimizing a single real-valued function (the one-parameter setting), there is a canonical choice of descriptor for persistent homology: the barcode. The operation mapping a real-valued function to its barcode is differentiable almost everywhere, and the convergence of gradient descent for losses using

barcodes is relatively well understood. When optimizing a vector-valued function (the multiparameter setting), there is no unique choice of descriptor for multiparameter persistent homology, and many distinct descriptors have been proposed. This calls for the development of a general framework for differentiability and optimization that applies to a wide range of multiparameter homological descriptors. In this article, we develop such a framework and show that it encompasses well-known descriptors of different flavors, such as signed barcodes and the multiparameter persistence landscape. We complement the theory with numerical experiments supporting the idea that optimizing multiparameter homological descriptors can lead to improved performances compared to optimizing one-parameter descriptors, even when using the simplest and most efficiently computable multiparameter descriptors.

8.3.3 Differentiable Mapper for topological optimization of data representation

Participant: Mathieu Carrière.

In collaboration with Ziyad Oulhaj, Bertrand Michel (Ecole Centrale de Nantes)

Unsupervised data representation and visualization using tools from topology is an active and growing field of Topological Data Analysis (TDA) and data science. Its most prominent line of work is based on the so-called Mapper graph, which is a combinatorial graph whose topological structures (connected components, branches, loops) are in correspondence with those of the data itself. While highly generic and applicable, its use has been hampered so far by the manual tuning of its many parameters-among these, a crucial one is the so-called filter: it is a continuous function whose variations on the data set are the main ingredient for both building the Mapper representation and assessing the presence and sizes of its topological structures. However, while a few parameter tuning methods have already been investigated for the other Mapper parameters (i.e., resolution, gain, clustering), there is currently no method for tuning the filter itself. In this work, we build on a recently proposed optimization framework incorporating topology to provide the first filter optimization scheme for Mapper graphs. In order to achieve this, we propose a relaxed and more general version of the Mapper graph, whose convergence properties are investigated. Finally, we demonstrate the usefulness of our approach by optimizing Mapper graph representations on several datasets, and showcasing the superiority of the optimized representation over arbitrary ones.

8.3.4 Choosing the parameter of the Fermat distance: navigating geometry and noise.

Participant: Frédéric Chazal, Laure Ferraris.

In collaboration with P. Groisman, M. Jonckheere, F. Pascal and F. Sapienza

The Fermat distance has been recently established as a useful tool for machine learning tasks when a natural distance is not directly available to the practitioner or to improve the results given by Euclidean distances by exploding the geometrical and statistical properties of the dataset. This distance depends on a parameter α that greatly impacts the performance of subsequent tasks. Ideally, the value of α should be large enough to navigate the geometric intricacies inherent to the problem. At the same, it should remain restrained enough to sidestep any deleterious ramifications stemming from noise during the process of distance estimation. In [10], we study both theoretically and through simulations how to select this parameter.

8.3.5 MAGDiff: Covariate Data Set Shift Detection via Activation Graphs of Deep Neural Networks

Participant: Felix Hensel, Charles Arnal, Mathieu Carrière, Frédéric Chazal.

In collaboration with T. Lacombe (Univ. G. Eiffel), H. Kurihara (Fujitsu), Y. Ike (Kyushu Univ.)

Despite their successful application to a variety of tasks, neural networks remain limited, like other machine learning methods, by their sensitivity to shifts in the data: their performance can be severely impacted by differences in distribution between the data on which they were trained and that on which they are deployed. In [14], we propose a new family of representations, called MAGDiff, that we extract from any given neural network classifier and that allows for efficient covariate data shift detection without the need to train a new model dedicated to this task. These representations are computed by comparing the activation graphs of the neural network for samples belonging to the training distribution and to the target distribution, and yield powerful data- and task-adapted statistics for the two-sample tests commonly used for data set shift detection. We demonstrate this empirically by measuring the statistical powers of two-sample Kolmogorov-Smirnov (KS) tests on several different data sets and shift types, and showing that our novel representations induce significant improvements over a state-of-the-art baseline relying on the network output.

8.3.6 Topological phase estimation method for reparameterized periodic functions

Participant: Frédéric Chazal, Wojciech reise.

In collaboration with Thomas Bonis (Univ. G. Eiffel) and Bertrand Michel (Ecole Centrale Nantes).

In [9], we consider a signal composed of several periods of a periodic function, of which we observe a noisy reparametrisation. The phase estimation problem consists of finding that reparametrisation, and, in particular, the number of observed periods. Existing methods are well-suited to the setting where the periodic function is known, or at least, simple. We consider the case when it is unknown and we propose an estimation method based on the shape of the signal. We use the persistent homology of sublevel sets of the signal to capture the temporal structure of its local extrema. We infer the number of periods in the signal by counting points in the persistence diagram and their multiplicities. Using the estimated number of periods, we construct an estimator of the reparametrisation. It is based on counting the number of sufficiently prominent local minima in the signal. This work is motivated by a vehicle positioning problem, on which we evaluated the proposed method.

8.3.7 ML Model Coverage Assessment by Topological Data Analysis Exploration

Participant: Martin Royer.

In collaboration with Ayman Fakhouri (IRT SystemX), Faouzi Adjed (IRT SystemX), Martin Gonzalez (IRT SystemX).

The increasing complexity of deep learning models necessitates advanced methods for model coverage assessment, a critical factor for their reliable deployment. In [29], we introduce a novel approach leveraging topological data analysis to evaluate the coverage of a couple dataset & classification model. By using tools from topological data analysis, our method identifies underrepresented regions within the data, thereby enhancing the understanding of both model performances and data completeness. This approach simultaneously evaluates the dataset and the model, highlighting areas of potential risk. We

report experimental evidence demonstrating the effectiveness of this topological framework in providing a comprehensive and interpretable coverage assessment. As such, we aim to open new avenues for improving the reliability and trustworthiness of classification models, laying the groundwork for future research in this domain.

8.3.8 Euler Characteristic Tools for Topological Data Analysis

[Added by Gilles]

Participant: Olympio Hacquard, Vadim Lebovici.

Note: the authors are no longer members of DATASHAPE as of 2024, but this work is the outcome of work done during their Ph.D. in DATASHAPE.

In [13], we study Euler characteristic techniques in topological data analysis. Pointwise computing the Euler characteristic of a family of simplicial complexes built from data gives rise to the so-called Euler characteristic profile. We show that this simple descriptor achieves state-of-the-art performance in supervised tasks at a meagre computational cost. Inspired by signal analysis, we compute hybrid transforms of Euler characteristic profiles. These integral transforms mix Euler characteristic techniques with Lebesgue integration to provide highly efficient compressors of topological signals. As a consequence, they show remarkable performances in unsupervised settings. On the qualitative side, we provide numerous heuristics on the topological and geometric information captured by Euler profiles and their hybrid transforms. Finally, we prove stability results for these descriptors as well as asymptotic guarantees in random settings.

8.4 Miscellaneous

8.4.1 Supervised Contamination Detection, with Flow Cytometry Application

Participant: Gilles Blanchard, Frédéric Chazal, Solenne Gaucher.

In [52], The contamination detection problem aims to determine whether a set of observations has been contaminated, i.e. whether it contains points drawn from a distribution different from the reference distribution. Here, we consider a supervised problem, where labeled samples drawn from both the reference distribution and the contamination distribution are available at training time. This problem is motivated by the detection of rare cells in flow cytometry. Compared to novelty detection problems or two-sample testing, where only samples from the reference distribution are available, the challenge lies in efficiently leveraging the observations from the contamination detection to design more powerful tests. In this article, we introduce a test for the supervised contamination detection problem. We provide non-asymptotic guarantees on its Type I error, and characterize its detection rate. The test relies on estimating reference and contamination densities using histograms, and its power depends strongly on the choice of the corresponding partition. We present an algorithm for judiciously choosing the partition that results in a powerful test. Simulations illustrate the good empirical performances of our partition selection algorithm and the efficiency of our test. Finally, we showcase our method and apply it to a real flow cytometry dataset.

8.4.2 Iteration Head: A Mechanistic Study of Chain-of-Thought.

Participant: Charles Arnal.

In collaboration with Vivien Cabannes (FAIR, Meta AI), Wassim Bouaziz (FAIR, Meta AI), Alice Yang (FAIR, Meta AI), Francois Charton (FAIR, Meta AI), Julia Kempe (Courant University and Center for Data Science, NYU & FAIR, Meta AI).

Chain-of-Thought (CoT) reasoning is known to improve Large Language Models both empirically and in terms of theoretical approximation power. However, our understanding of the inner workings and conditions of apparition of CoT capabilities remains limited. [26] helps fill this gap by demonstrating how CoT reasoning emerges in transformers in a controlled and interpretable setting. In particular, we observe the appearance of a specialized attention mechanism dedicated to iterative reasoning, which we coined "iteration heads". We track both the emergence and the precise working of these iteration heads down to the attention level, and measure the transferability of the CoT skills to which they give rise between tasks.

8.4.3 Touring sampling with pushforward maps

Participant: Charles Arnal.

In collaboration with Vivien Cabannes (FAIR, Meta AI).

The number of sampling methods could be daunting for a practitioner looking to cast powerful machine learning methods to their specific problem. [25] takes a theoretical stance to review and organize many sampling approaches in the “generative modeling” setting, where one wants to generate new data that are similar to some training examples. By revealing links between existing methods, it might prove useful to overcome some of the current challenges in sampling with diffusion models, such as long inference time due to diffusion simulation, or the lack of diversity in generated samples.

8.4.4 Prompt Selection Matters: Enhancing Text Annotations for Social Sciences with Large Language Models

Participant: Charles Arnal.

In collaboration with Louis Abraham (Université Paris 1 Panthéon-Sorbonne) and Antoine Marie (Institut Jean Nicod, Ecole Normale Supérieure, PSL-EHESS-CNRS).

Large Language Models have recently been applied to text annotation tasks from social sciences, equalling or surpassing the performance of human workers at a fraction of the cost. However, no inquiry has yet been made on the impact of prompt selection on labelling accuracy. In [33], we show that performance greatly varies between prompts, and we apply the method of automatic prompt optimization to systematically craft high quality prompts. We also provide the community with a simple, browser-based implementation of the method at github.

8.4.5 Scaling Laws with Hidden Structure

Participant: Charles Arnal.

In collaboration with Clement Berenfeld (Postdam Univ.), Simon Rosenberg (C3IA) and Vivien Cabannes (FAIR, Meta AI).

Statistical learning in high-dimensional spaces is challenging without a strong underlying data structure. Recent advances with foundational models suggest that text and image data contain such hidden structures, which help mitigate the curse of dimensionality. Inspired by results from nonparametric statistics, we hypothesize that this phenomenon can be partially explained in terms of decomposition of complex tasks into simpler subtasks. In [35], we present a controlled experimental framework to test whether neural networks can indeed exploit such "hidden factorial structures." We find that they do leverage these latent patterns to learn discrete distributions more efficiently, and derive scaling laws linking model sizes, hidden factorizations, and accuracy. We also study the interplay between our structural assumptions and the models' capacity for generalization.

8.4.6 Mode Estimation with Partial Feedback

Participant: Charles Arnal.

In collaboration with Vivien Cabannes (FAIR, Meta AI) and Vianney Perchet (Center for Research in Economics and Statistics (CREST), ENSAE, Palaiseau, France).

The combination of lightly supervised pre-training and online finetuning has played a key role in recent AI developments. These new learning pipelines call for new theoretical frameworks. In [36], we formalize core aspects of weakly supervised and active learning with a simple problem: the estimation of the mode of a distribution using partial feedback. We show how entropy coding allows for optimal information acquisition from partial feedback, develop coarse sufficient statistics for mode identification, and adapt bandit algorithms to our new setting. Finally, we combine those contributions into a statistically and computationally efficient solution to our problem.

8.4.7 False discovery proportion envelopes with m-consistency

Participant: Gilles Blanchard.

In collaboration with Iqraa Meah (U. Paris-Cité and INSERM), Etienne Roquain (U. Paris-Sorbonne)

In [17] We provide new nonasymptotic false discovery proportion (FDP) confidence envelopes in several multiple testing settings relevant for modern high dimensional-data methods. We revisit the multiple testing scenarios considered in the recent work of Katsevich and Ramdas (2020): top-k, preordered (including knockoffs), online. Our emphasis is on obtaining FDP confidence bounds that both have nonasymptotical coverage and are asymptotically accurate in a specific sense, as the number m of tested hypotheses grows. Namely, we introduce and study the property (which we call m-consistency) that the confidence bound converges to or below the desired level α when applied to a specific reference α-level false discovery rate (FDR) controlling procedure. In this perspective, we derive new bounds that provide improvements over existing ones, both theoretically and practically, and are suitable for situations where at least a moderate number of rejections is expected. These improvements are illustrated with numerical experiments and real data examples. In particular, the improvement is significant in the knockoffs setting, which shows the impact of the method for a practical use. As side results, we introduce a new confidence envelope for the empirical cumulative distribution function of i.i.d. uniform variables and we provide new power results in sparse cases, both being of independent interest.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

- **Participants:** Alexandre Guerin, Frédéric Chazal.

Collaboration with **Sysnav**, a French SME with world leading expertise in navigation and geopositioning in extreme environments, on TDA, geometric approaches and machine learning for the analysis of movements of pedestrians and patients equipped with inertial sensors (CIFRE PhD of Alexandre Guérin).

- **Participants:** Bastien Dussap, Marc Glisse, Gilles Blanchard.

Research collaboration with **MetaFora** on the development of new TDA-based and statistical methods for the analysis of cytometric data (started in Nov. 2019).

- **Participants:** David Cohen-Steiner.

Collaboration with **Dassault Systèmes** and Inria team **Geomerix** (Saclay) on the applications of methods from geometric measure theory to the modelling and processing of complex 3D shapes (PhD of Lucas Brifault, started in May 2022).

- **Participants:** Frédéric Chazal, Myriam Frikha.

Research collaboration with **Ericsson** on transfer learning for temporal data with applications in telecommunications (PhD of Myriam Frikha, started in November 2024).

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL or an international program

TopTime

Title: Topological and statistical methods for time series data

Duration: 2024 - 2026

Coordinator: Nina Otter

Partners:

- Australian National University Canberra (Australie) — coordinator: Katharine Turner (katharine.turner@anu.edu.au)

Summary: Methods from the mathematical area of topology have been applied successfully to data arising from a variety of domains. The DataShape team and the ANU team have both established mathematical foundations for this area, generally known as Topological Data Analysis. While there exist empirical studies of the usefulness of TDA methods for dynamic data (i.e., data changing over time), there is to date no principled understanding of topological properties of dynamic data. In this project we will develop principled mathematical foundations to study dynamical data from a topological point of view, and address specific challenges in real-world applications, as well as develop state-of-the-art software of the methods, accessible to a broad audience of data scientists.

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Marzieh Eidi

Status post-Doc

Institution of origin: Max Planck Institute for Mathematics in the Sciences

Country: Germany

Dates: 01.10.2024-10.12.2024

Context of the visit: ALTO exchange programme between MPG and CNRS on “Role and positional analysis for simplicial complexes: a random-walk approach” (participants: Marzieh Eidi and Nina Otter)

Mobility programme: research visit

Musashi Koyama

Status PhD student

Institution of origin: Australian National University

Country: Australia

Dates: 01.10.2024-31.10.2024

Context of the visit: Equipe Associée TopTime

Mobility programme: research visit

10.2.2 Visits to international teams

Research stays abroad

Mathijs Wintraecken

Visited institution: University of Notre Dame

Country: United States of America, State of Indiana

Dates: 03.11.2024-15.11.2024

Context of the visit: Collaboration with Erin Chambers as well as Chris Fillmore, and Elizabeth Stephenson

Mobility program/type of mobility: research stay

10.3 National initiatives

Extended visit

Participants: Corentin Lunel, Clément Maria.

- Duration : 1 years from Septembre 2024
- Coordinator : Clément Maria
- Location : Institut Montpelliérain Alexandre Grothendieck (IMAG) - Université de Montpellier

The visit consists of federating mathematicians from IMAG working on low dimensional and quantum topology together with computer scientists from Datashape, to work at the interface of the two fields.

10.3.1 ANR

ANR Chair in AI

Participants: Frédéric Chazal, Marc Glisse, Louis Pujol, Wojciech Riese.

- Acronym : TopAI
- Type : ANR Chair in AI.
- Title : Topological Data Analysis for Machine Learning and AI
- Coordinator : Frédéric Chazal
- Duration : 4 years from September 2020 to August 2024.
- Others Partners: Two industrial partners, the French SME Sysnav and the French start-up MetaFora.
- Abstract:

The TopAI project aims at developing a world-leading research activity on topological and geometric approaches in Machine Learning (ML) and AI with a double academic and industrial/societal objective. First, building on the strong expertise of the candidate and his team in TDA, TopAI aims at designing new mathematically well-founded topological and geometric methods and tools for Data Analysis and ML and to make them available to the data science and AI community through state-of-the-art software tools. Second, thanks to already established close collaborations and the strong involvement of French industrial partners, TopAI aims at exploiting its expertise and tools to address a set of challenging problems with high societal and economic impact in personalized medicine and AI-assisted medical diagnosis.

ANR ALGOKNOT

Participants: Clément Maria.

- Acronym : ALGOKNOT.
- Type : ANR Jeune Chercheuse Jeune Chercheur.
- Title : Algorithmic and Combinatorial Aspects of Knot Theory.
- Coordinator : Clément Maria.
- Duration : 2020 – 2025 (5 years).
- Abstract: The project AlgoKnot aims at strengthening our understanding of the computational and combinatorial complexity of the diverse facets of knot theory, as well as designing efficient algorithms and software to study their interconnections.
- See also: Clément Maria and ANR AlgoKnot.

ANR GeMfaceT

Participants: Blanche Buet.

- Acronym: GeMfaceT.
- Type: ANR JCJC -CES 40 – Mathématiques
- Title: A bridge between Geometric Measure and Discrete Surface Theories
- Coordinator: Blanche Buet.
- Duration: 48 months, starting October 2021.
- Abstract: This project positions at the interface between geometric measure and discrete surface theories. There has recently been a growing interest in non-smooth structures, both from theoretical point of view, where singularities occur in famous optimization problems such as Plateau problem or geometric flows such as mean curvature flow, and applied point of view where complex high dimensional data are no longer assumed to lie on a smooth manifold but are more singular and allow crossings, tree-structures and dimension variations. We propose in this project to strengthen and expand the use of geometric measure concepts in discrete surface study and complex data modelling and also, to use those possible singular disrcete surfaces to compute numerical solutions to the aforementioned problems.

10.3.2 Collaboration with other national research institutes

Confiance.ai / IRT SystemX

Participants: Frédéric Chazal.

Research collaboration on anomaly detection for multivariate time series using TDA and ML approaches.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

- Blanche Buet was a member of the organisation committee of the conference Geometric Sciences in Action: from Geometric Statistics to Shape Analysis (CIRM, may 2024).
- Nina Otter was a co-organiser of an American Mathematical Society Mathematics Research Community (AMS MRC) on “Climate science at the interface between Topological Data Analysis and Dynamical Systems Theory”, held in Buffalo, NY, in June 2024.
- Clément Maria was co-organizer of the *Geometry & Computing conference* held at CIRM, Marseille, October 2024.
- Clément Maria was co-organizer of the *QuantAzur Days*, Nice, October 2024.
- DataShape seminar: the seminar is held in person either at the LMO or at Inria Sophia Antipolis; all talks are recorded and can be accessed at: bbb2.imo.universite-paris-saclay.fr.

11.1.2 Scientific events: selection

Member of the conference program committees

- David Cohen-Steiner was a member of the program committee of SGP 2024 (Symposium on Geometry Processing, Cambridge USA, 2024)

- Gilles Blanchard was Senior Area Chair for the NeurIPS'24 conference.
- Clément Maria was member of the program committee of the International Symposium on Computational Geometry (SOCG) 2024.

11.1.3 Journal

Member of the editorial boards

- Gilles Blanchard was member of the following journal editorial boards: Annals of Statistics (IMS), Electronic Journal of Statistics (IMS), Bernoulli (ISI/Bernoulli Society)
- Gilles Blanchard is co-editor-in-chief for the "Mathematics and Applications" collection (Springer, SMAI)
- Frédéric Chazal is a member of the following journal editorial boards: Discrete and Computational Geometry (Springer), Graphical Models (Elsevier).
- Frédéric Chazal is the Editor-in-Chief of the Journal of Applied and Computational Topology (Springer).
- Clément Maria is a co-editor of the CGTA Special Issue on Algorithmic Aspects of Computational and Applied Topology.

11.1.4 Invited talks

- Gilles Blanchard gave an invited talk at the International Conference on Statistics and Data Science (ICSDS) organized by the IMS, Nice, Dec. 2024.
- Gilles Blanchard gave an invited talk at the International Symposium on Nonparametric Statistics (ISNP) organized by the IMS in Braga, Portugal, June 2024.
- Vincent Rouvreau gave a presentation and a practical session of the GUDHI library at the Jeunes Chercheur.e.s en Géométrie (jcgeo24) organized by the GDR IG-RV and the GDR IFM in ESIEE Paris, Université Gustave Eiffel, June 2024.
- Clément Maria gave an invited keynote talk at the ComPerWorkshop 2024, Graz, Austria.
- Nina Otter was a plenary speaker at the Spires 2024 Conference, University of Oxford, UK.

11.1.5 Leadership within the scientific community

- Frédéric Chazal has been the Scientific Director of the DATAIA Institute at Université Paris-Saclay until Sept.1, 2024.
- Frédéric Chazal is a member of the board of directors of the DIM project AI4IDF of the Région Ile-de-France until Sept.1, 2024.
- Frédéric Chazal is a member of the Scientific Advisory Board of the Centre for Topological Data Analysis of the Mathematical Institute at Oxford.

11.1.6 Scientific expertise

- Frédéric Chazal is a member of the "commission prospective de l'I2M" (Institut de Mathématiques de Marseille).
- Clément Maria was a jury member for the UCA-DS4H PhD grant allocation scheme for 2024.

11.1.7 Research administration

- Frédéric Chazal is co-responsible of the “programme Mathématiques et IA” of the Fondation Mathématique Jacques Hadamard, Paris-Saclay Univ.
- Frédéric Chazal is a member of the council of the Graduate School in Mathematics, Paris-Saclay Univ.
- Blanche Buet is member of the laboratoire council and CCUPS of LMO, Paris-Saclay Univ.
- Gilles Blanchard is part of the Sustainable Development Commission at the LMO, U. Paris-Saclay.
- Clément Maria is co-responsible of the CNRS-Groupe de Travail GeoAlgo.
- Clément Maria is a member of the steering committee of the QuantAzur federative institute.
- Clément Maria is a member of the Steering Committee of the Année de la Géométrie du GdR IFM (Informatique Fondamentale et ses Mathématiques).
- Marc Glisse is president of the CDT at Inria Saclay.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Master: Frédéric Chazal, Analyse Topologique des Données, 30h eq-TD, Université Paris-Sud, France.
- Master: Marc Glisse, Computational Geometry Learning, 18h eq-TD, M2, MPRI, France.
- Master: Frédéric Cazals and Mathieu Carrière, Foundations of Geometric Methods in Data Analysis, 24h eq-TD, M2, École CentraleSupélec, France.
- Master: Frédéric Cazals and Jean-Daniel Boissonnat and Mathieu Carrière, Geometric and Topological Methods in Data Analysis, with Applications in Biology and Medecine, 24h eq-TD, M2, Université Côte d’Azur, France.
- Master: Mathieu Carrière, Basic Algebra for Data Analysis, 15h eq-TD, M1, Université Côte d’Azur, France.
- Master: Frédéric Chazal and Julien Tierny, Topological Data Analysis, 38h eq-TD, M2, Mathématiques, Vision, Apprentissage (MVA), ENS Paris-Saclay, France.
- Master: Gilles Blanchard, Mathematics for Artificial Intelligence 1, 70h eq-TD, IMO, Université Paris-Saclay, France.
- Master: Gilles Blanchard, Kernel and operator-theoretic methods in machine learning, 32h eq-TD, Université Paris-Saclay, France.
- Master: Blanche Buet, TD-Distributions et analyse de Fourier, 60h eq-TD, M1, Université Paris-Saclay, France.
- Master: Marc Glisse, Conception et analyse d’algorithmes, 44h eq-TD, M1, École Polytechnique, France.
- Master: Nina Otter, Méthodes de modellisation statistique, ENSTA Paris, 25.5h eq-TD. (2024)
- Master: Mathijs Wintraecken, Geometrie et topologie (colles), 16h eq-TD, M1 Mathématiques, Laboratoire Jean Alexandre Dieudonné, Nice, France.

11.2.2 Supervision

- PhD: Bastien Dussap, comparaison de données cytométriques. Defended in October 2024. Gilles Blanchard and Marc Glisse
- PhD: Jérémie Capitao-Miniconi, deconvolution for singular measures with geometric support. Defended in October 2024. Frédéric Chazal and Elisabeth Gassiat.
- PhD: Jean-Baptiste Fermanian, Estimation de Kernel Mean Embedding et tests multiples en grande dimension. Defended in November 2024. Gilles Blanchard and Magalie Fromont-Renoir.
- PhD: David Loiseaux, Multivariate topological data analysis for statistical machine learning. Defended in December 2024. Mathieu Carrière and Frédéric Cazals.
- PhD: Antoine Commaret, Persistent Geometry. Defended in December 2024. David Cohen-Steiner and Indira Chatterji.
- PhD in progress: Myriam Frikha, Domain adaptation for temporal data. Started in Nov. 2024. Frédéric Chazal.
- PhD in progress: Hannah Marienwald (TU Berlin), Transfer learning in high dimension. Started September 2019. Gilles Blanchard and Klaus-Robert Müller (TU Berlin).
- PhD in progress: Lucas Brifault, Théorie de la mesure géométrique appliquée pour la modélisation de formes complexes. Started May 2022. David Cohen-Steiner and Mathieu Desbrun.
- PhD in progress: Charly Boricaud, Geometric inference for Data analysis: a Geometric Measure Theory perspective. Started on October 2021. Blanche Buet, Gian Paolo Leonardi et Simon Masnou.
- PhD in progress: Hugo Henneuse. Statistical Foundations of Topological Data Analysis for multidimensional random fields. Started on October 2022. Frédéric Chazal and Pascal Massart.
- PhD in progress: António Leitão, started on November 2024. Nina Otter and Fosca Gianotti (Scuola Normale Superiore di Pisa).
- PhD in progress: Henrique Lovisi Ennes, started October 2023. Topological approach to Quantum Complexity. Clément Maria and Nicolas Nisse (INRIA UniCA, EPI COATI).

11.2.3 Juries

- Nina Otter was an external jury member for Renata Turkeš's' PhD defense at the University of Antwerp.

11.3 Popularization

11.3.1 Others science outreach relevant activities

- Clément Maria gave a talk on quantum computing at the INRIA UniCA C@fé-In, March 2024
- Clément Maria gave a talk "Portrait de chercheur" at the C@fé ADSTIC Sophia Tech.

12 Scientific production

12.1 Major publications

[1] G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee and C. Scott. 'Domain Generalization by Marginal Transfer Learning'. In: *Journal of Machine Learning Research* 22.2 (2021), pp. 1–55. DOI: 10.48550/arXiv.1711.07910. URL: https://hal.science/hal-02974216.

[2] J.-D. Boissonnat and M. Wintraecken. 'The Topological Correctness of PL Approximations of Isomanifolds'. In: *Foundations of Computational Mathematics* 22 (13th July 2021), pp. 967–1012. DOI: 10.1007/s10208-021-09520-0. URL: https://inria.hal.science/hal-03760378.

[3] M. Carriere and A. J. Blumberg. 'Multiparameter Persistence Images for Topological Machine Learning'. In: NeurIPS 2020 - 33rd Conference on Neural Information Processing Systems. Vancouver / Virtuel, Canada, 6th Dec. 2020. URL: https://inria.hal.science/hal-03112442.

[4] M. Carriere, F. Chazal, M. Glisse, Y. Ike and H. Kannan. 'Optimizing persistent homology based functions'. In: ICML 2021 - 38th International Conference on Machine Learning. Vol. PMLR 139. Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Virtual conference, United States, 18th July 2021, pp. 1294–1303. URL: https://inria.hal.science/hal-02969305.

[5] D. Cohen-Steiner, A. Lieutier and J. Vuillamy. 'Lexicographic Optimal Homologous Chains and Applications to Point Cloud Triangulations'. In: *Discrete and Computational Geometry* 68 (13th Sept. 2022). DOI: 10.1007/s00454-022-00432-6. URL: https://hal.science/hal-03870128.

[6] R. Gribonval, G. Blanchard, N. Keriven and Y. Traonmilin. 'Compressive Statistical Learning with Random Feature Moments'. In: *Mathematical Statistics and Learning* 3.2 (21st Aug. 2021), pp. 113–164. DOI: 10.4171/msl/20. URL: https://inria.hal.science/hal-01544609.

[7] C. Maria and J. Spreer. 'A Polynomial-Time Algorithm to Compute Turaev–Viro Invariants $TV_{4,q}$ of 3-Manifolds with Bounded First Betti Number'. In: *Foundations of Computational Mathematics* 20.5 (11th Nov. 2019), pp. 1013–1034. DOI: 10.1007/s10208-019-09438-8. URL: https://hal.science/hal-04296224.

12.2 Publications of the year

International journals

[8] G. Blanchard, A. Carpentier and O. Zadorozhnyi. 'Moment inequalities for sums of weakly dependent random fields'. In: *Bernoulli* 30.3 (1st Aug. 2024). DOI: 10.3150/23-BEJ1682. URL: https://hal.science/hal-04150509 (cit. on p. 18).

[9] T. Bonis, F. Chazal, B. Michel and W. Reise. 'Topological phase estimation method for reparameterized periodic functions'. In: *Advances in Computational Mathematics* 50.4 (8th July 2024), p. 66. DOI: 10.1007/s10444-024-10157-0. URL: https://hal.science/hal-03687686 (cit. on p. 20).

[10] F. Chazal, L. Ferraris, P. Groisman, M. Jonckheere, F. Pascal and F. Sapienza. 'Choosing the parameter of the Fermat distance: navigating geometry and noise'. In: *Transactions on Machine Learning Research Journal* (2024), pp. 2835–8856. DOI: 10.48550/arXiv.2311.18663. URL: https://hal.science/hal-04317396 (cit. on p. 19).

[11] F. Chazal, C. Levrard and M. Royer. 'Topological Analysis for Detecting Anomalies (TADA) in dependent sequences: application to Time Series'. In: *Journal of Machine Learning Research* 25 (1st Dec. 2024), pp. 1–49. URL: https://hal.science/hal-04604083 (cit. on p. 15).

[12] H. Edelsbrunner, A. Garber, M. Ghafari, T. Heiss, M. Saghafian and M. Wintraecken. 'Brillouin Zones of Integer Lattices and Their Perturbations'. In: *SIAM Journal on Discrete Mathematics* 38.2 (7th June 2024), pp. 1784–1807. DOI: 10.1137/22M1489071. URL: https://hal.science/hal-04628684 (cit. on p. 11).

[13] O. Hacquard and V. Lebovici. 'Euler Characteristic Tools For Topological Data Analysis'. In: *Journal of Machine Learning Research* 25 (July 2024). URL: https://hal.science/hal-04143938 (cit. on p. 21).

[14] F. Hensel, C. Arnal, M. Carrière, T. Lacombe, H. Kurihara, Y. Ike and F. Chazal. 'MAGDiff: Covariate Data Set Shift Detection via Activation Graphs of Deep Neural Networks'. In: *Transactions on Machine Learning Research Journal* (May 2024). URL: https://hal.science/hal-04103272 (cit. on p. 20).

[15] S. Liang, R. Turkes, J. Li, N. Otter and G. Montufar. 'Pull-back Geometry of Persistent Homology Encodings'. In: *Transactions on Machine Learning Research Journal* (2024). URL: https://hal.science/hal-04708462.

[16] D. Loiseaux and H. Schreiber. 'multipers: Multiparameter Persistence for Machine Learning'. In: *Journal of Open Source Software* 9.103 (14th Nov. 2024), p. 6773. DOI: 10.21105/joss.06773. URL: https://inria.hal.science/hal-04801544.

[17] I. Meah, G. Blanchard and E. Roquain. 'False discovery proportion envelopes with m-consistency'. In: *Journal of Machine Learning Research* 25.270 (15th Oct. 2024), pp. 1–52. URL: https://hal.science/hal-04727618 (cit. on p. 23).

[18] M. Michaud, A. Guérin, M. Dejean de la Bâtie, L. Bancel, L. Oudre and A. Tricot. 'The Analytical Validity of Stride Detection and Gait Parameters Reconstruction Using the Ankle-Mounted Inertial Measurement Unit Syde®'. In: *Sensors* 24.8 (10th Apr. 2024), p. 2413. DOI: 10.3390/s24082413. URL: https://hal.science/hal-04733572.

[19] W. Reise, B. Michel and F. Chazal. 'Topological signatures of periodic-like signals'. In: *Bernoulli* (2024). URL: https://hal.science/hal-04140929. In press (cit. on p. 15).

International peer-reviewed conferences

[20] C. Arnal, D. Cohen-Steiner and V. Divol. 'Wasserstein convergence of Čech persistence diagrams for samplings of submanifolds'. In: NeurIPS 2024. Vancouver (Canada), Canada, 9th Dec. 2024. URL: https://hal.science/hal-04617508.

[21] D. Attali, H. Dal Poz Kouřimská, C. Fillmore, I. Ghosh, A. Lieutier, E. Stephenson and M. Wintraecken. 'The Ultimate Frontier: An Optimality Construction for Homotopy Inference (Media Exposition)'. In: *Leibniz International Proceedings in Informatics*. SoCG 2024 - 40th International Symposium on Computational Geometry. Athènes, Greece: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 87:1–87:6. DOI: 10.4230/LIPIcs.SoCG.2024.87. URL: https://hal.science/hal-04628855 (cit. on p. 12).

[22] D. Attali, H. Dal Poz Kouřimská, C. Fillmore, I. Ghosh, A. Lieutier, E. Stephenson and M. Wintraecken. 'Tight Bounds for the Learning of Homotopy à la Niyogi, Smale, and Weinberger for Subsets of Euclidean Spaces and of Riemannian Manifolds'. In: *https://drops.dagstuhl.de/entities/volume/LIPIcs-volume-293*. SoCG 2024 - 40th International Symposium on Computational Geometry. Vol. 293. Athènes, Greece: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. DOI: 10.4230/LIPIcs.SoCG.2024.11. URL: https://hal.science/hal-04628805 (cit. on pp. 11, 12).

[23] J.-D. Boissonnat and K. Dutta. 'A Euclidean Embedding for Computing Persistent Homology with Gaussian Kernels'. In: *Leibniz International Proceedings in Informatics*. ESA 2024 - European Symposium on Algorithms. Vol. LIPIcs-308. 32nd Annual European Symposium on Algorithms (ESA 2024). Egham, United Kingdom: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 29:1–29:18. DOI: 10.4230/LIPICS.ESA.2024.29. URL: https://inria.hal.science/hal-04744438.

[24] J.-D. Boissonnat, K. Dutta, S. Dutta and S. Pritam. 'On Edge Collapse of Random Simplicial Complexes'. In: *Leibniz International Proceedings in Informatics*. SoCG 2024 - 40th International Symposium on Computational Geometry. Vol. LIPIcs-293. 40th International Symposium on Computational Geometry (SoCG 2024). Athens, Greece: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 21:1–21:16. DOI: 10.4230/LIPIcs.SoCG.2024.21. URL: https://inria.hal.science/hal-04744459.

[25] V. Cabannes and C. Arnal. 'Touring sampling with pushforward maps'. In: ICASSP 2024 - International Conference on Acoustics, Speech, and Signal Processing. Seoul, South Korea, 14th Apr. 2024. DOI: 10.48550/ARXIV.2311.13845. URL: https://inria.hal.science/hal-04471440 (cit. on p. 22).

[26] V. Cabannes, C. Arnal, W. Bouaziz, A. Yang, F. Charton and J. Kempe. 'Iteration Head: A Mechanistic Study of Chain-of-Thought'. In: NeurIPS 2024. Vancouver (Canada), Canada, 4th June 2024. URL: https://hal.science/hal-04620726 (cit. on p. 22).

[27] M. Carriere, M. Theveneau and T. Lacombe. 'Diffeomorphic interpolation for efficient persistence-based topological optimization'. In: Advances in Neural Information Processing Systems 37 (NeurIPS). Vancouver, Canada, Dec. 2024. URL: https://hal.science/hal-04587467.

[28] H. Dal Poz Kouřimská, A. Lieutier and M. Wintraecken. 'The Medial Axis of Any Closed Bounded Set Is Lipschitz Stable with Respect to the Hausdorff Distance Under Ambient Diffeomorphisms'. In: *Leibniz International Proceedings in Informatics*. SoCG 2024 - 40th International Symposium on Computational Geometry. Vol. LIPIcs-293. 40th International Symposium on Computational Geometry. Athens, Greece: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. DOI: 10.4230/LIPIcs.SoCG.2024.69. URL: https://hal.science/hal-04628829 (cit. on p. 12).

[29] A. Fakhouri, F. Adjed, M. Gonzalez and M. Royer. 'ML Model Coverage Assessment by Topological Data Analysis Exploration'. In: ATRACC workshop 2024 - AI Trustworthiness and Risk Assessment for Challenged Contexts / AAAI 2024 Fall Symposium. Arlington (VA), United States, 10th Nov. 2024. URL: https://hal.science/hal-04717675 (cit. on p. 20).

[30] U. Gazin, G. Blanchard and E. Roquain. 'Transductive conformal inference with adaptive scores'. In: *Proceedings of Machine Learning Research*. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. Vol. 238. Proceedings of Machine Learning Research (PMLR). Valencia, Spain: PMLR, 2024, pp. 1504–1512. URL: https://hal.science/hal-04266605 (cit. on p. 17).

[31] Z. Oulhaj, M. Carrière and B. Michel. 'Differentiable mapper for topological optimization of data representation'. In: The Forty-first International Conference on Machine Learning (ICML 2024). Wien, Austria, 14th Feb. 2024. URL: https://hal.science/hal-04457645.

[32] L. Scoccola, S. Setlur, D. Loiseaux, M. Carrière and S. Oudot. 'Differentiability and Optimization of Multiparameter Persistent Homology'. In: ICML 2024 - The Forty-first International Conference on Machine Learning. Wien, Austria, 11th June 2024. URL: https://hal.science/hal-04609272.

Reports & preprints

[33] L. Abraham, C. Arnal and A. Marie. *Prompt Selection Matters: Enhancing Text Annotations for Social Sciences with Large Language Models*. 15th July 2024. URL: https://hal.science/hal-04657856 (cit. on p. 22).

[34] C. Arnal. *The distance function to a finite set is a topological Morse function*. 22nd July 2024. URL: https://hal.science/hal-04657849 (cit. on p. 13).

[35] C. Arnal, C. Berenfeld, S. Rosenberg and V. Cabannes. *Scaling Laws with Hidden Structure*. 2nd Nov. 2024. URL: https://hal.science/hal-04768311 (cit. on p. 23).

[36] C. Arnal, V. Cabannes and V. Perchet. *Mode Estimation with Partial Feedback*. 2024. DOI: 10.48550/arXiv.2402.13079. URL: https://inria.hal.science/hal-04471461 (cit. on p. 23).

[37] C. Arnal, D. Cohen-Steiner and V. Divol. *Critical points of the distance function to a generic submanifold*. 2024. DOI: 10.48550/arXiv.2312.13147. URL: https://inria.hal.science/hal-04471485.

[38] D. Attali, H. Dal Poz Kouřimská, C. Fillmore, I. Ghosh, A. Lieutier, E. Stephenson and M. Wintraecken. *Supplementary material: The ultimate frontier: An optimality construction for homotopy inference*. 2024. URL: https://hal.science/hal-04501285.

[39] D. Attali, H. D. P. Kouřimská, C. Fillmore, I. Ghosh, A. Lieutier, E. Stephenson and M. Wintraecken. *Tight Bounds for the Learning of Homotopy à la Niyogi, Smale, and Weinberger for Subsets of Euclidean Spaces and of Riemannian Manifolds*. 7th Mar. 2024. URL: https://hal.science/hal-03721463.

[40] A. Baldi, B. Franchi and P. Pansu. *Continuous primitives for higher degree differential forms in Euclidean spaces, Heisenberg groups and applications*. 15th Mar. 2024. URL: https://hal.science/hal-04518692.

[41] G. Blanchard, G. Durand, A. Marandon-Carlhian and R. Périer. *FDR control and FDP bounds for conformal link prediction*. 2nd Apr. 2024. URL: https://hal.science/hal-04529648.

[42] G. Blanchard, J.-B. Fermanian and H. Marienwald. *Estimation of multiple mean vectors in high dimension*. 21st Mar. 2024. URL: https://hal.science/hal-04515801.

[43] A. J. Blumberg, M. Carriere, J. H. Fung and M. A. Mandell. *Resampling and averaging coordinates on data*. 2nd Aug. 2024. URL: https://inria.hal.science/hal-04706757.

[44] A. J. Blumberg, M. Carrière, J. H. Fung and M. A. Mandell. *Subsampling, aligning, and averaging to find circular coordinates in recurrent time series*. 24th Dec. 2024. URL: https://inria.hal.science/hal-04855856.

[45] J. Capitao-Miniconi, É. Gassiat and L. Lehéricy. *Deconvolution of repeated measurements corrupted by unknown noise*. 3rd Sept. 2024. URL: https://hal.science/hal-04685944 (cit. on p. 16).

[46] J. Capitao-Miniconi, É. Gassiat and L. Lehéricy. *Support and distribution inference from noisy data*. 24th May 2024. URL: https://hal.science/hal-04073724 (cit. on p. 15).

[47] M. Carriere, S. Kim and W. Kim. *Sparsification of the Generalized Persistence Diagrams for Scalability through Gradient Descent*. 9th Dec. 2024. URL: https://hal.science/hal-04826755.

[48] D. Cohen-Steiner and A. Commaret. *Persistent intrinsic volumes*. 19th July 2024. URL: https://hal.science/hal-04660971.

[49] A. Commaret. *Generalized Morse theory for tubular neighborhoods*. 2024. URL: https://hal.science/hal-04747884.

[50] H. Dal Poz Kouřimská, A. Lieutier and M. Wintraecken. *A free lunch: manifolds of positive reach can be smoothed without decreasing the reach*. 4th Dec. 2024. URL: https://inria.hal.science/hal-04816535.

[51] J. H. Fung, M. Carrière and A. Blumberg. *Statistical estimation of sparsity and efficiency for molecular codes*. 15th Aug. 2024. DOI: 10.1101/2024.08.13.607773. URL: https://inria.hal.science/hal-04706760.

[52] S. Gaucher, G. Blanchard and F. Chazal. *Supervised Contamination Detection, with Flow Cytometry Application*. 5th Apr. 2024. URL: https://hal.science/hal-04535142 (cit. on p. 21).

[53] H. Henneuse. *A Geometric Approach for Multivariate Jumps Detection*. 4th Oct. 2024. URL: https://hal.science/hal-04721009 (cit. on p. 16).

[54] H. Henneuse. *Persistence Diagram Estimation : Beyond Plug-in Approaches*. 30th May 2024. URL: https://hal.science/hal-04593678 (cit. on p. 16).

[55] H. Henneuse. *Persistence Diagram Estimation of Multivariate Piecewise Hölder-continuous Signals*. 28th Mar. 2024. URL: https://hal.science/hal-04524998 (cit. on p. 17).

[56] H. Henneuse. *Persistence-based Modes Inference*. 22nd July 2024. URL: https://hal.science/hal-04655461 (cit. on p. 17).

[57] E. K. Kandror, A. Wang, M. Carriere, A. Peterson, W. Liao, A. Tjarnberg, J. H. Fung, K. T. Mahbubani, J. Loper, W. Pangburn, Y. Xu, K. Saeb-Parsy, R. Rabadan, T. Maniatis and A. H. Rizvi. *Enhancer Dynamics and Spatial Organization Drive Anatomically Restricted Cellular States in the Human Spinal Cord*. 11th Jan. 2025. DOI: 10.1101/2025.01.10.632483. URL: https://inria.hal.science/hal-04881805.

[58] A. Leitao and N. Otter. *Time-optimal persistent homology representatives for univariate time series*. 2024. URL: https://hal.science/hal-04844440 (cit. on p. 13).

[59] A. Lieutier and M. Wintraecken. *Manifolds of positive reach, differentiability, tangent variation, and attaining the reach*. 4th Dec. 2024. URL: https://inria.hal.science/hal-04816588.

[60] A. Onus, N. Otter and R. Turkeš. *Shoving tubes through shapes gives a sufficient and efficient shape statistic*. 2024. URL: https://hal.science/hal-04856362 (cit. on p. 13).

[61] E. M. Saad, G. Blanchard and S. Arlot. *Online Orthogonal Matching Pursuit*. 7th Oct. 2024. URL: https://hal.science/hal-03141061.

[62] D. Williams and M. Wintraecken. *Quasi-optimal interpolation of gradients and vector-fields on protected Delaunay meshes in $\mathbb{R}^d$*. 2024. URL: https://hal.science/hal-04822711.

Other scientific publications

[63] D. Attali, H. D. P. Kouřimská, C. Fillmore, I. Ghosh, A. Lieutier, E. Stephenson and M. Wintraecken. *The Ultimate Frontier: An Optimality Construction for Homotopy Inference*. 2024. URL: https://hal.science/hal-04579406 (cit. on p. 12).

Scientific popularization

[64] D. Faranda, T. Lacombe, N. Otter and K. Strommen. 'Climate Science at the Interface Between Topological Data Analysis and Dynamical Systems Theory'. In: *Notices of the American Mathematical Society* 71 (1st Feb. 2024), pp. 267–271. DOI: 10.1090/noti2864. URL: https://inria.hal.science/hal-04396161.

12.3 Cited publications

[65] D. Aldous and J. M. Steele. 'The objective method: probabilistic combinatorial optimization and local weak convergence'. In: *Probability on discrete structures*. Springer, 2004, pp. 1–72 (cit. on p. 9).

[66] J.-D. Boissonnat and S. Pritam. 'Edge Collapse and Persistence of Flag Complexes'. In: *36th International Symposium on Computational Geometry (SoCG 2020)*. Ed. by S. Cabello and D. Z. Chen. Vol. 164. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020, 19:1–19:15. DOI: 10.4230/LIPIcs.SoCG.2020.19. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SoCG.2020.19 (cit. on p. 9).

[67] D. Chen and J. M. Phillips. 'Relative error embeddings of the Gaussian kernel distance'. In: *International Conference on Algorithmic Learning Theory*. PMLR. 2017, pp. 560–576 (cit. on p. 10).

[68] N. Linial and Y. Peled. 'Random simplicial complexes: around the phase transition'. In: *A Journey Through Discrete Mathematics: A Tribute to Jiří Matoušek* (2017), pp. 543–570 (cit. on p. 9).

[69] J. M. Phillips, B. Wang and Y. Zheng. 'Geometric Inference on Kernel Density Estimates'. In: *31st International Symposium on Computational Geometry (SoCG 2015)*. Ed. by L. Arge and J. Pach. Vol. 34. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015, pp. 857–871. DOI: 10.4230/LIPIcs.SOCG.2015.857. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SOCG.2015.857 (cit. on p. 10).

[70] A. Rahimi and B. Recht. 'Random Features for Large-Scale Kernel Machines'. In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer and S. Roweis. Vol. 20. Curran Associates, Inc., 2007. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf (cit. on p. 10).

[71] L. Warnke. 'On the Method of Typical Bounded Differences'. In: *Combinatorics, Probability and Computing* 25.2 (2016), pp. 269–299. DOI: 10.1017/S0963548315000103 (cit. on p. 9).