Towards Multivariate Persistence For Machine Learning

David Loiseaux

3IA PhD Student (M. Carrière, F. Cazals), DataShape Research group

Chair of Jean-Daniel Boissonnat

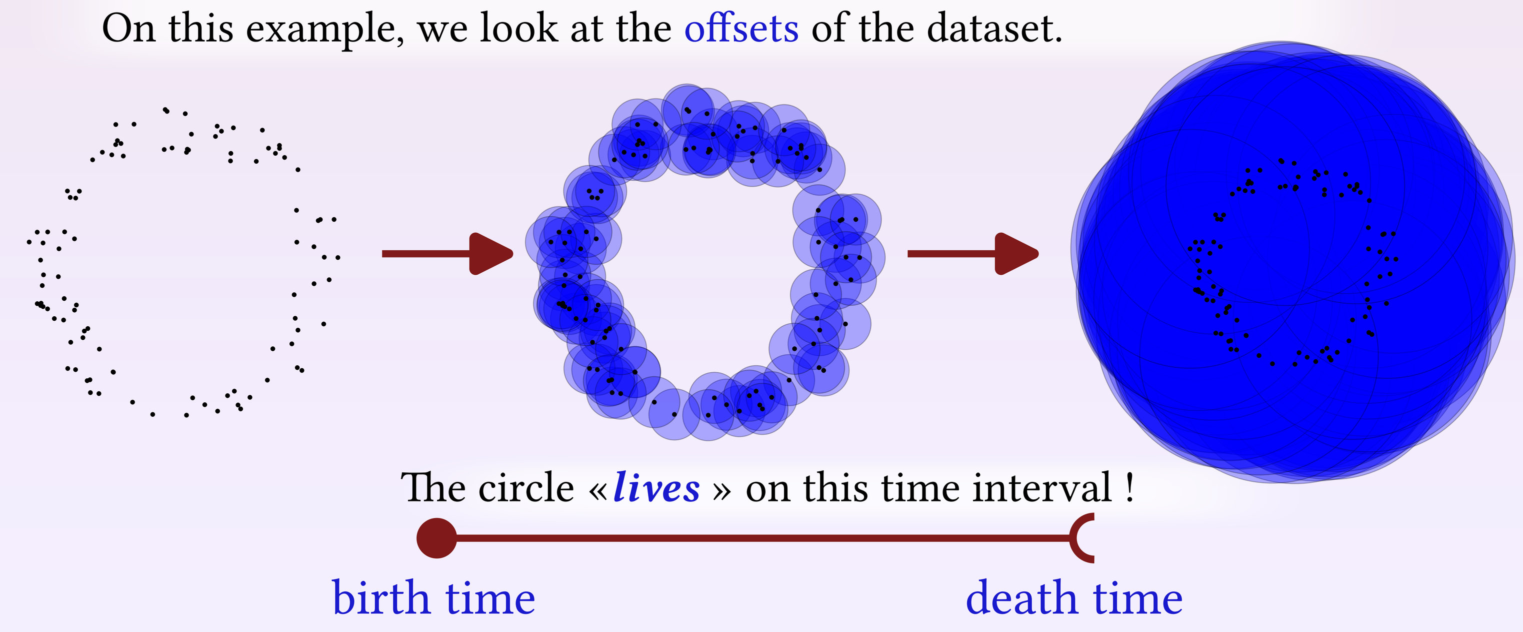
Centre Inria d'Université Côte d'Azur

Motivation

The huge variety of dataset in the wild has brought many difficulties from a statistical point of view, such as the so-called « *curse of the dimensionality* ».
Fortunately, data sets usually lie close to some hidden structure; which, if taken into account in the learning pipeline, can help mitigate this effect.
Topological Data Analysis (**TDA**) is a strategy that aims for a solution to this challenge, by providing compact descriptors inferring the topological features of this hidden structure, such as connectivity, loops, cavities; with nice guarantees.
However, these main descriptors, the *persistent modules*, still suffers from some technical limitations; particularly, in computational biology, there are, in some cases no « bayesian way » to compute them, as the input is too large. This motivates their generalization : multiparameter persistence modules.

Idea

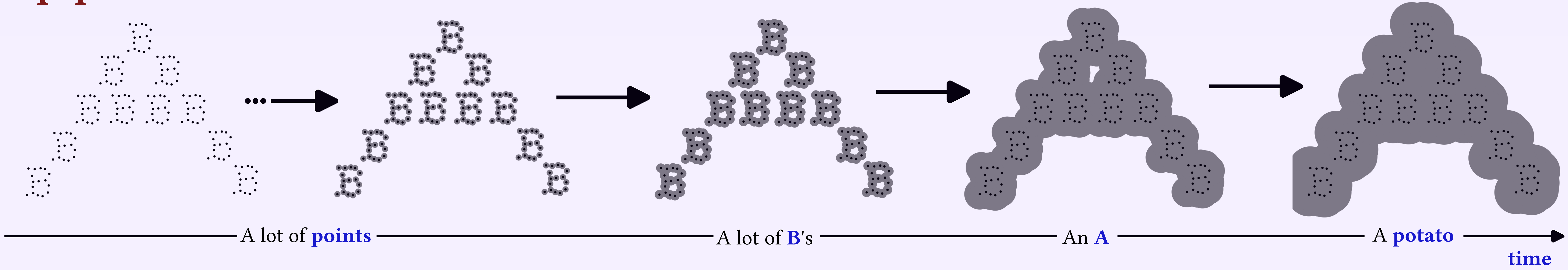
A standard goal is to recover « *topological features* » from a dataset.
E.g., from points sampled on a circle, we want to retrieve this circle.
On this example, we look at the offsets of the dataset.



This construction can be generalized to « *filter functions* »; we look at the topology of the sublevelsets $(\{x \in \mathcal{X} \mid f(x) \leq t\})_{t \in \mathbb{R}}$ for a function $f : \mathcal{X} \to \mathbb{R}$

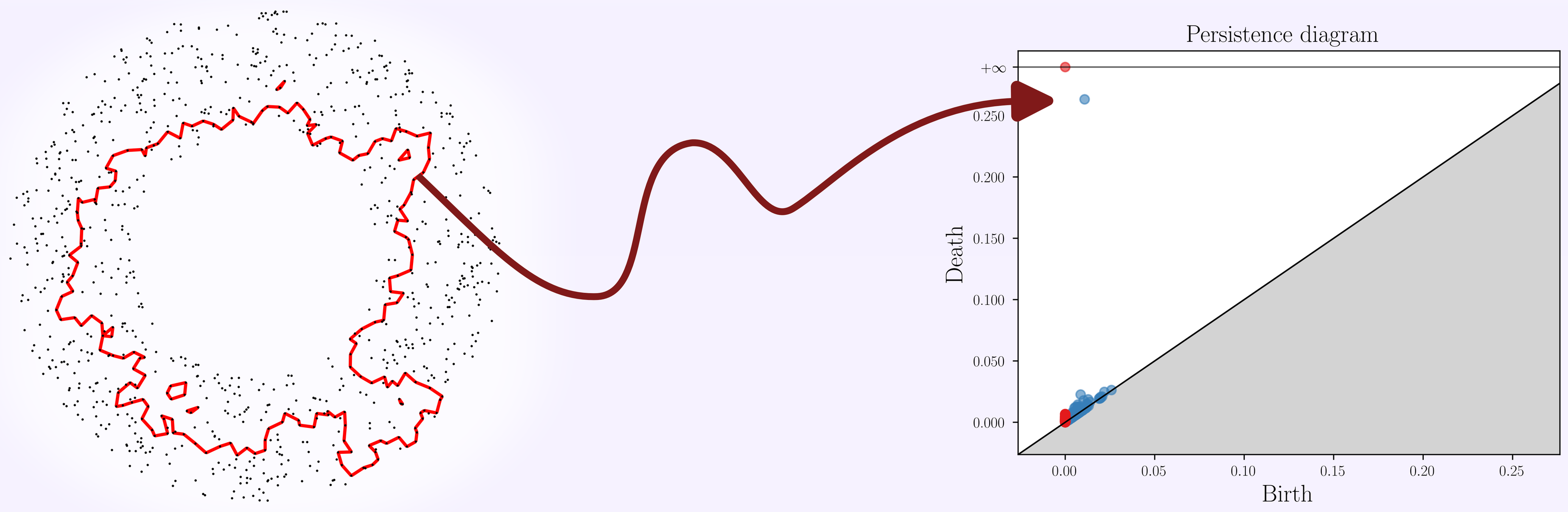
In that context, « *growing balls* » ≈ take the sublevelsets of the distance to the dataset function.

This pipeline catches all scales at once !



Byproducts of TDA

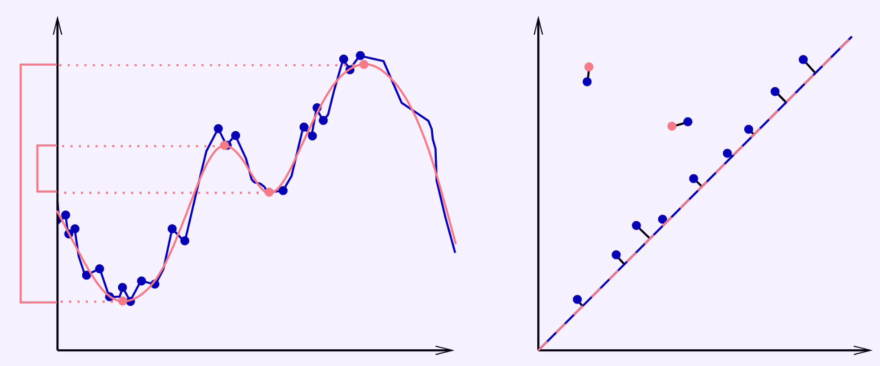
The mathematical structure behind is the *persistent homology*, which encodes birth and death time of each topological feature. It can be represented as a *persistence diagram* (**Dgm**).



Nice properties

Stability

For two functions $f, g : \mathcal{X} \to \mathbb{R}$,
$d_b\left(\mathrm{Dgm}(f), \mathrm{Dgm}(g)\right) \leq \|f - g\|_\infty$.



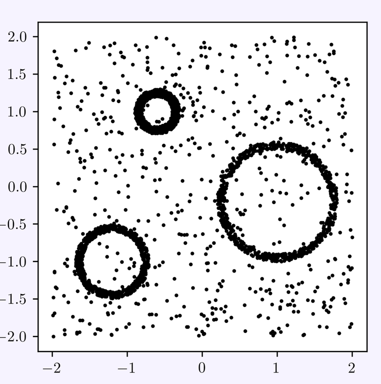
Convergence

If $X^{(n)} = (X_1, \ldots, X_n) \sim \mu^{\otimes n}$ is a nice sampling of a space X_μ,
Then $\mathrm{Dgm}\left(X^{(n)}\right) \xrightarrow[n \to \infty]{d_b} \mathrm{Dgm}\left(X_\mu\right)$

Compact descriptors

(Almost) any dataset can be used in this pipeline, and the output is very dense in informations.

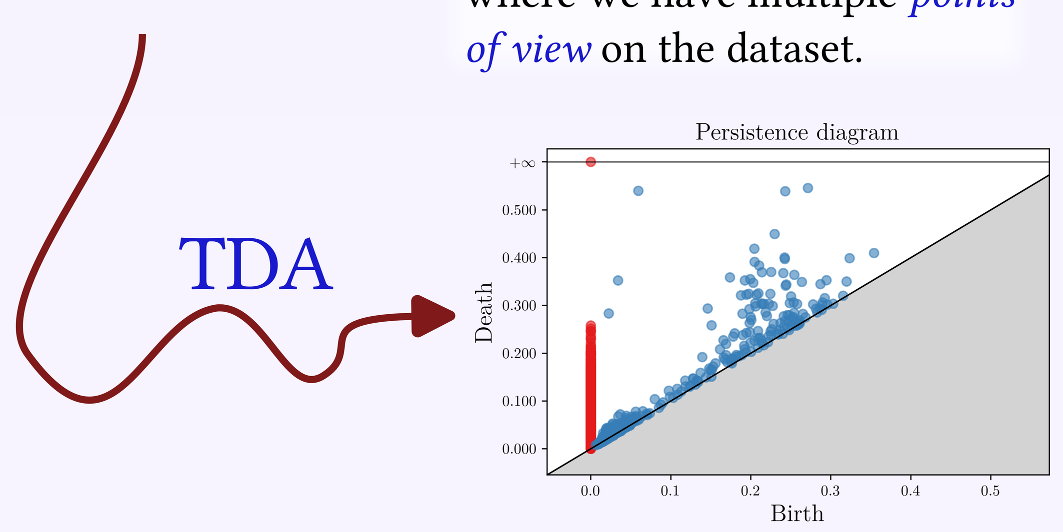
Limitations of 1D persistence



Can be unstable w.r.t. outliers.

Can depend on *a priori* choices.

Can be limited in some cases, where we have multiple *points of view* on the dataset.

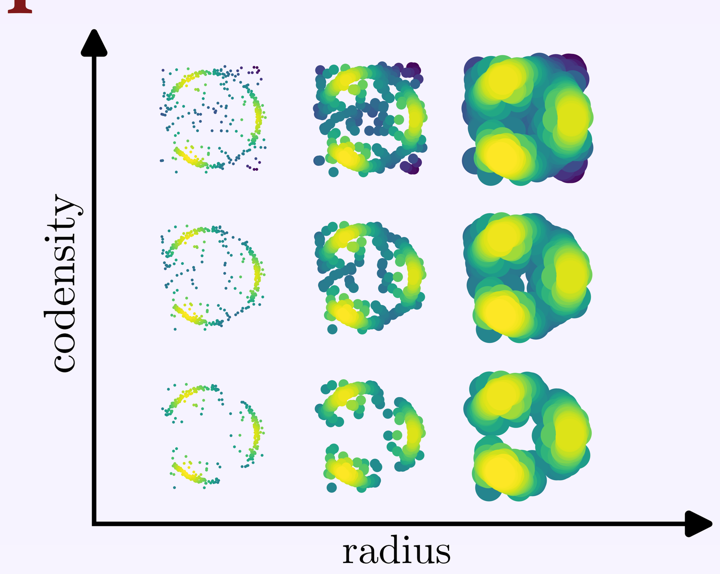


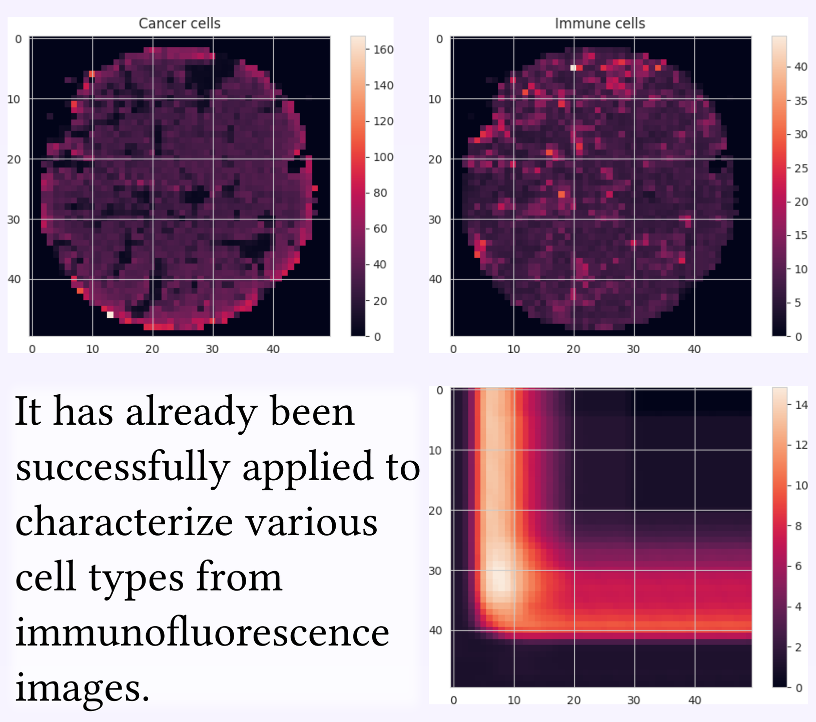
Multiparameter persistence

This is a generalization of the previous pipeline; it allows filters of multiple parameters

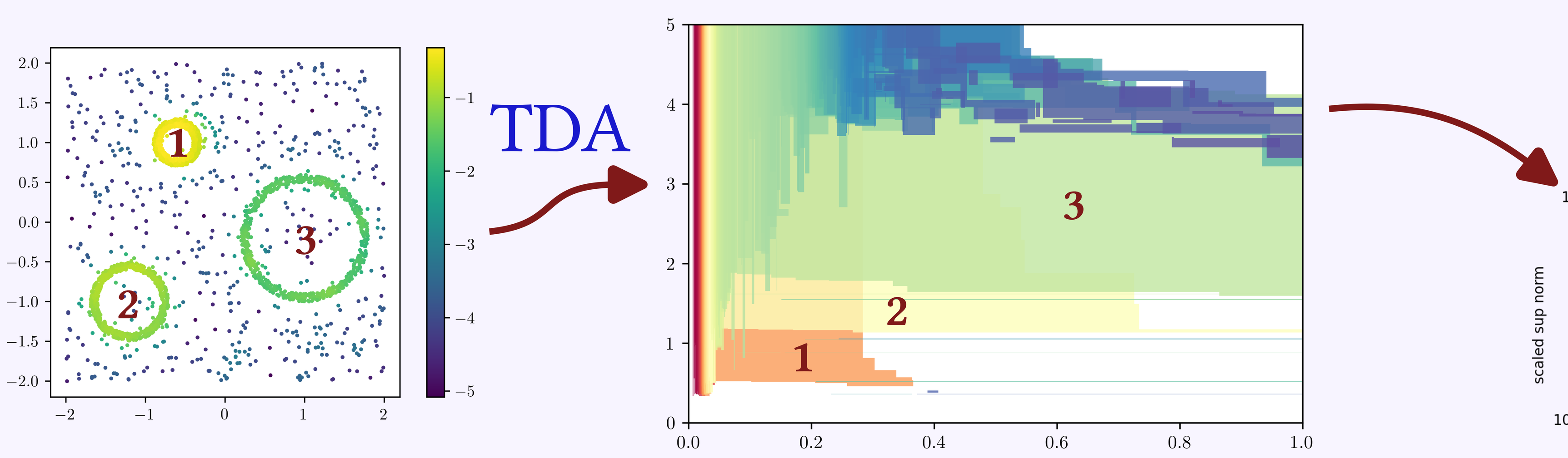
$$f : \mathcal{X} \to \mathbb{R}^d$$

This allows to study the topological variations of, e.g., scale and density, or multiple marker genes jointly.

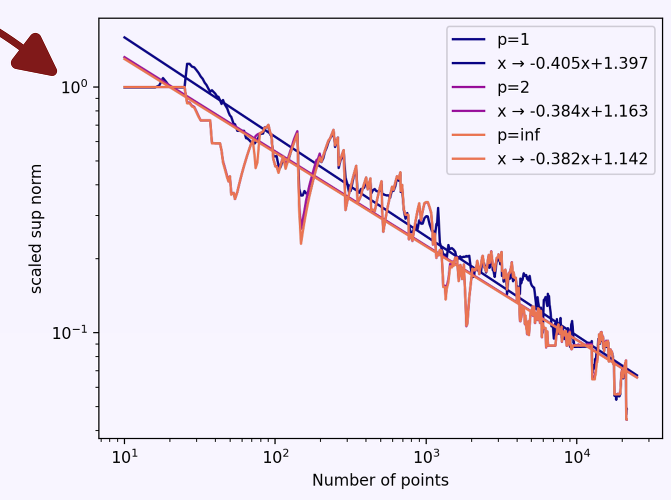




It has already been successfully applied to characterize various cell types from immunofluorescence images.



Convergence



References

[1] Chazal F. de Silva V. Glisse M. Oudot S. The Structure and Stability of Persistence Modules
[2] Carrière M. Blumberg A. Multiparameter Persistence Images for Topological Machine Learning
[3] Chazal F. Glisse M. Labruère C. Michel B. Convergence Rates for Persistence Diagram Estimation in Topological Data Analysis
[4] Cohen-Steiner D. Edelsbrunner H. Harer J. Stability of persistence diagrams
[5] Hirokazu A. Chazal F. Glisse M. Yuichi I. Hiroya I. Raphaël T. Yuhei U. DTM-based Filtrations
[6] Blumberg A. Lesnick M. Stability of 2-Parameter Persistent Homology
[7] Botnan, M. B., Oppermann, S. & Oudot, S. Signed barcodes for multi-parameter persistence via rank decompositions and rank-exact resolutions

Packages

MMA : https://gitlab.inria.fr/dloiseau/multipers
Rivet : https://github.com/rivetTDA/rivet/
GUDHI Geometry Understanding in Higher Dimensions