Towards multivariate persistence for ML

David Loiseaux

3IA PhD student (M. Carrière, F. Cazals), DataShape research group

Chair of Jean-Daniel Boissonnat

Inria Sophia Antipolis - Méditerranée & Université Côte d'Azur, France

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 (see [2]). This motivates their generalization : *multiparameter persistence modules*.

References, Packages

[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

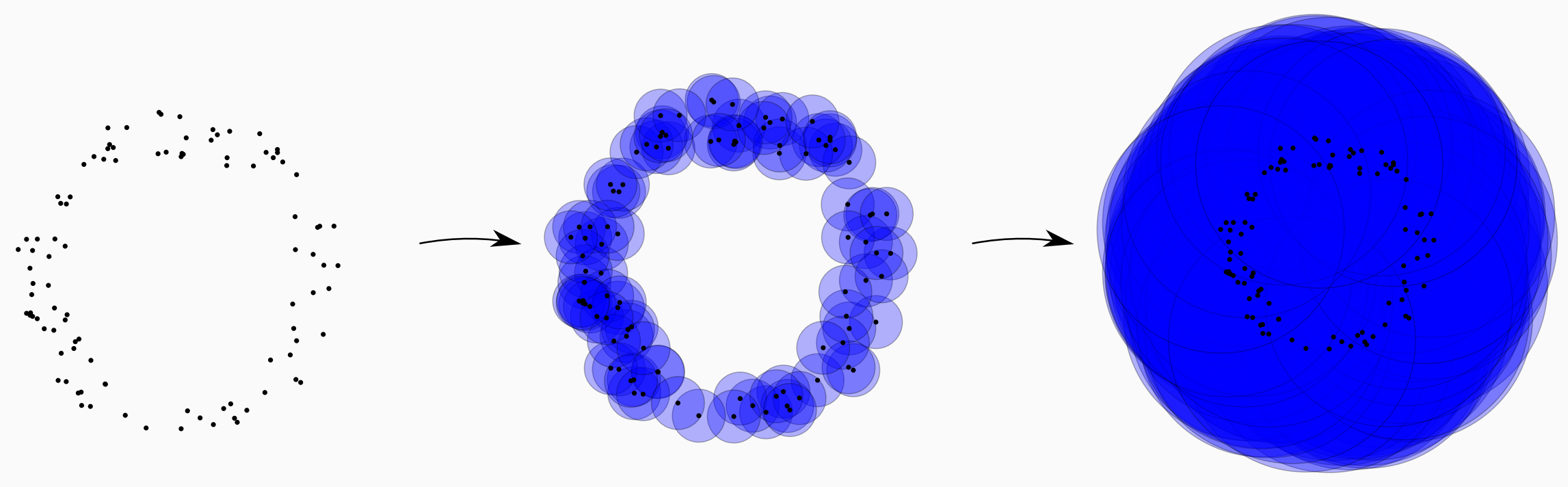
[7] Hatcher A. Algebraic topology

- गुढी GUDHI Geometry Understanding in Higher Dimensions is a C++ , python and R library driven by Inria.
- Ripser is a minimalistic and efficient library (C++, python, julia, R) meant to compute persistent homology.
- Dionysus, PHAT are also good alternatives.

Idea

Take a point cloud. We want to recover topological features of the underlying space.

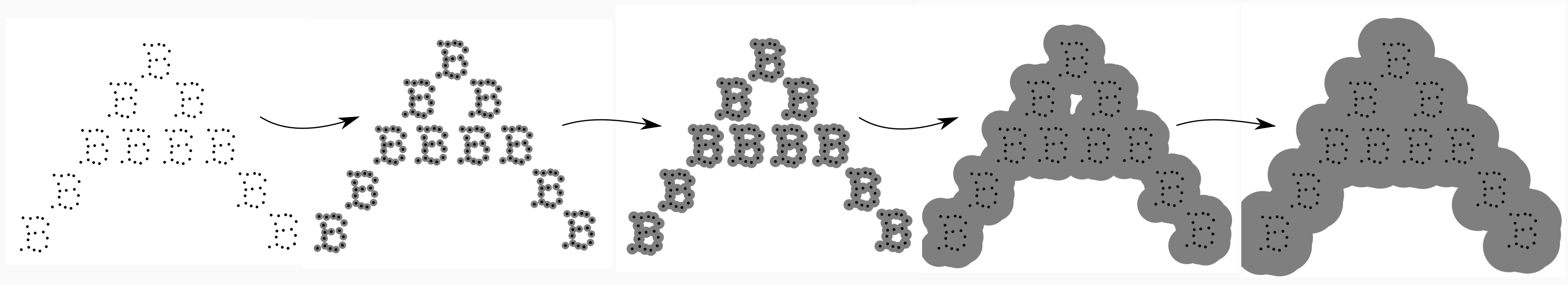
Increase the radii of the balls around the points, and keep in memory the topological features of this space over time.



As one can see in the above figure, the blue circle was born at step 2 and died at step 3; and the number of connected components went from 100 at step 1, to 1 at step 2.

Scale property

This allows to capture topological features at every scale, at once !

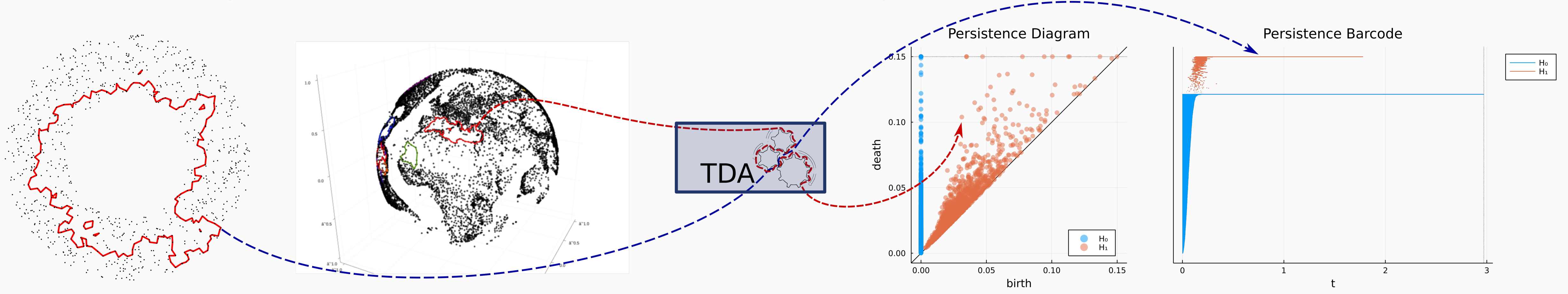


Filter functions

It is more efficient and convenient in practice to work with *filters functions*; and look at the topology of the sublevelsets of this function. In that context, "growing balls" ≅ sublevelset of distance filter function. Those are more abstract and general constructions than the previous ones.

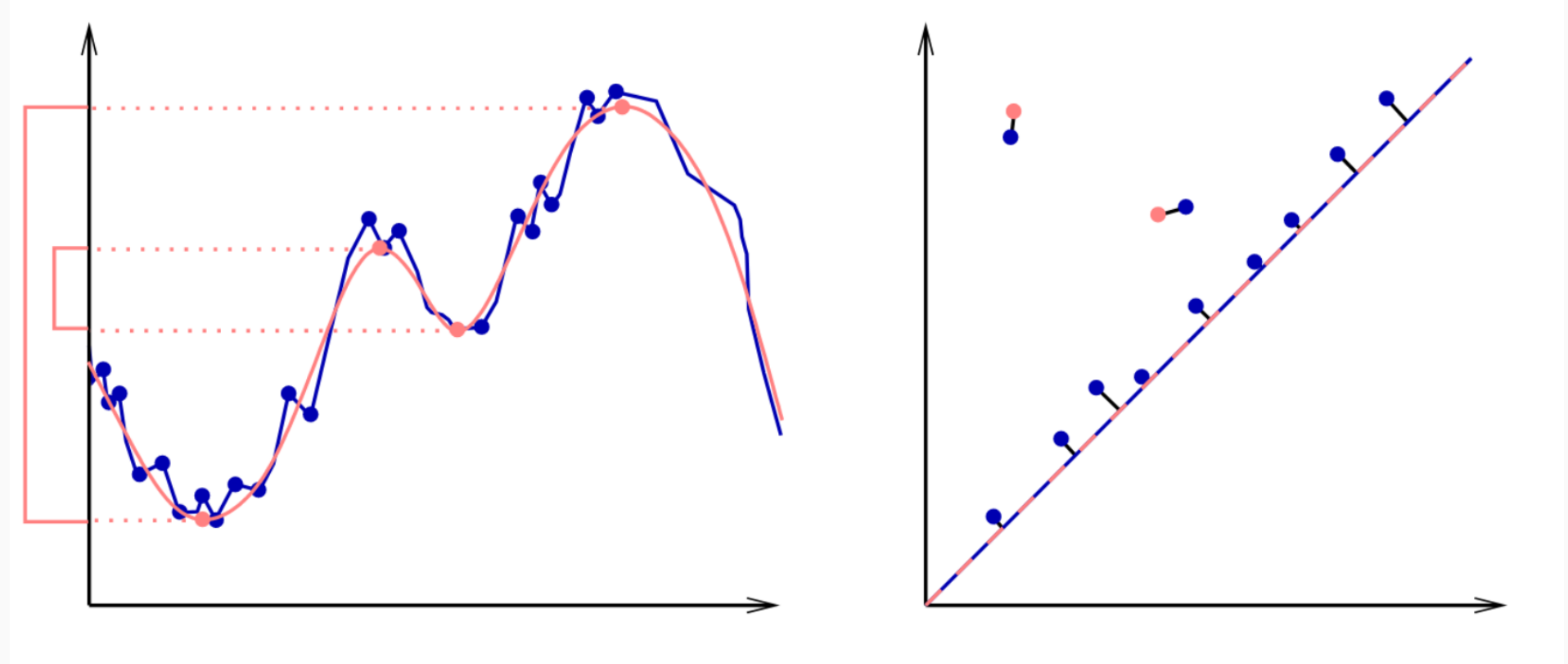
Byproducts of persistence modules

We encode topological features in *persistence barcodes* or *persistence diagrams* by representing their *birth* and *death times*.



Some nice results

Stability, convergence, compact descriptors (taken from [4]). $d_b(\mathrm{Dgm}\, f, \mathrm{Dgm}\, g) \leq \|f-g\|_\infty$.



A successful method

[1] Nicolau M, Levine AJ, Carlsson G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. 2011

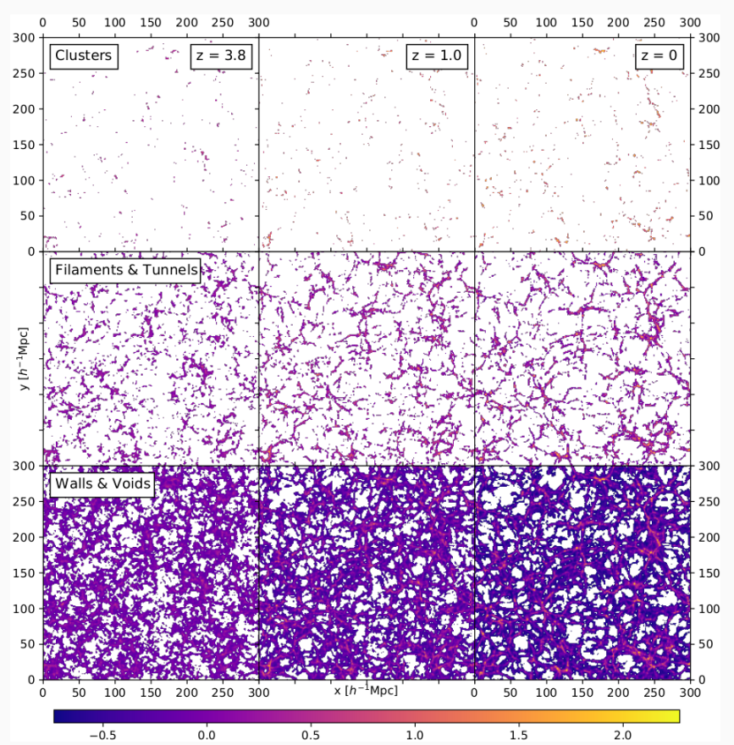
[2] Bleher M. Hahn L. Patino-Galindo J. Carriere M. Bauer U. Rabadan R. Ott A. Topology identifies emerging adaptive mutations in SARS-CoV-2. 2021

[3] Stolz B. Kaeppler J. Markelc B. Mech F. Lipsmeier F. Muschel R. Byrne H. Harrington H. Multiscale Topology Characterises Dynamic Tumour Vascular Networks 2020

[4] Horn M. de Brouwer E. MoorM. Moreau Y. Rieck B. Borgwardt K. Topological Graph Neural Networks. 2021

[5] Wilding G. Nevenzeel K. van de Weygaert R. Vegter G.Pranav P. Jones B. Efstathiou K. Feldbrugge J. Persistent homology of the cosmic web. I: Hierarchical topology in ΛCDM cosmologies. 2021

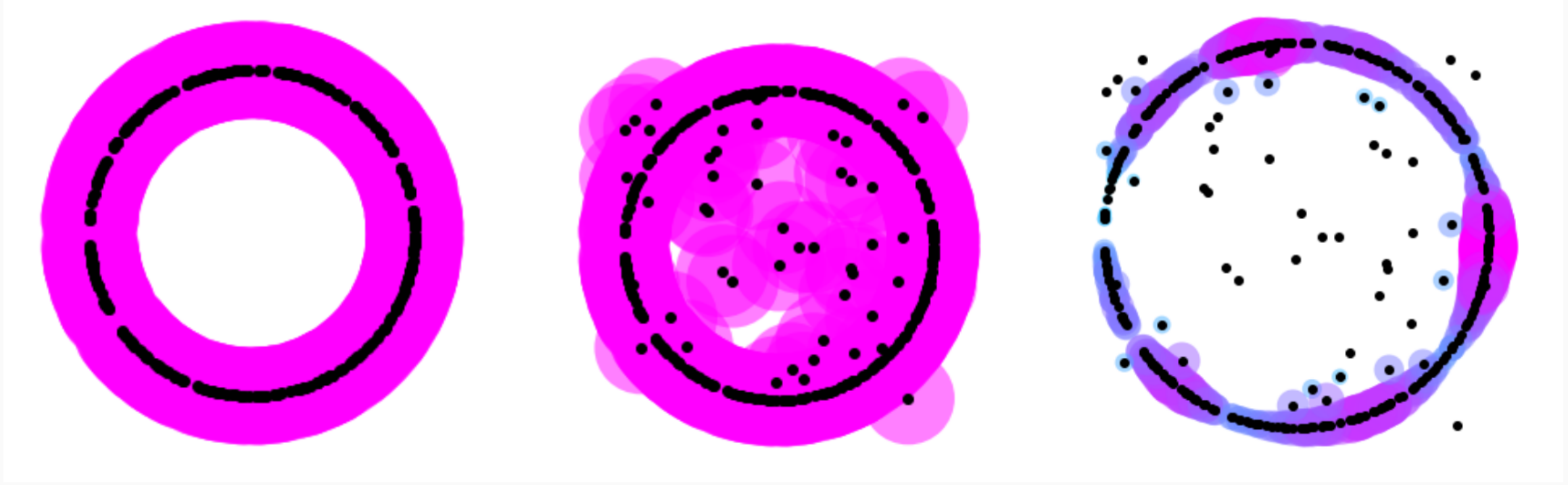
[6] Moor M. Horn M. Rieck B. Borgwardt K. Topological Autoencoders 2020



…and *much* more ! Take a look at the Zotero group TDA-Applications. Picture taken from [5].

Limitations of 1 dimensional persistence

- Not stable with respect to outliers (from [5]).

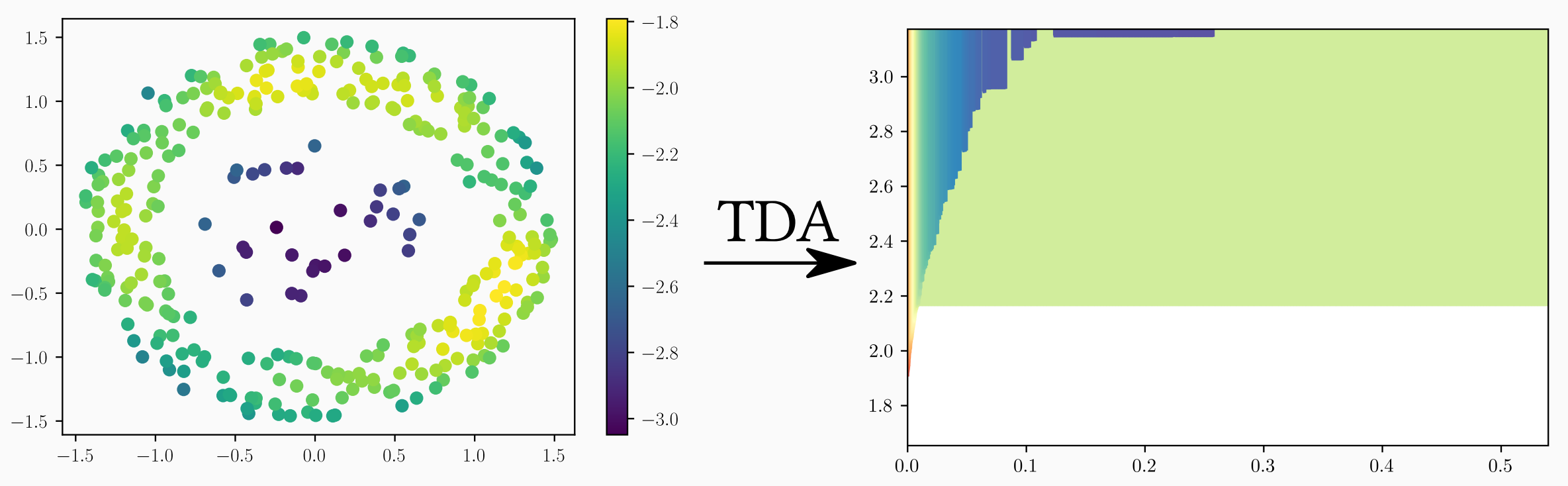


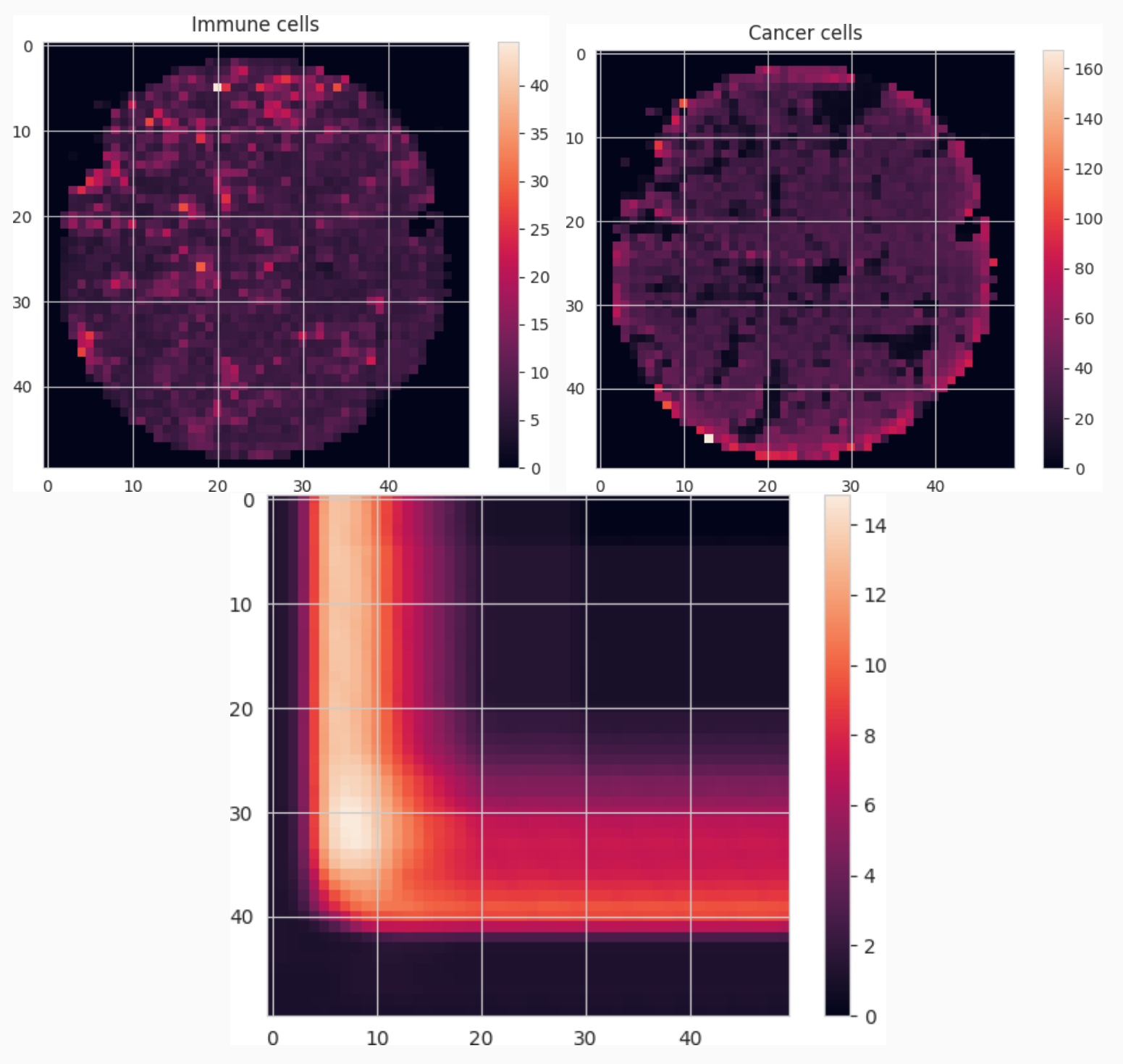
- Depends on *a priori* choices.
- Can be limited in some cases ⟹ need to compute separately different *points of view* of the data set.

⟹ More general pipeline : *multiparameter persistence*; which can handle many filters at once !

Multipersistence : a work in progress

As the theory of multiparameter persistence is new, there is still no proper way to compute it in the general case. We are currently working on the theory and implementation of such a pipeline.





Taking density into account recovers the circle feature, represented as the big green rectangle.

It has already been successfully applied to characterize various cell types in immunofluorescence images. See figure on the right from [2].

LaTeX TikZposter