Applications of Topological Data Analysis
ARAMIS Lab Seminar

Antoine Commaret

07/10/2022

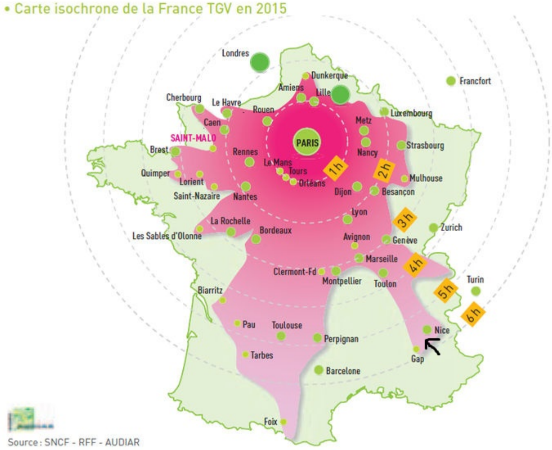
Inria

ICM Institut du Cerveau

ARAMIS LAB BRAIN DATA SCIENCE

About me

- PhD Student at INRIA Sophia-Antipolis in Persistent Homology and Geometric Measure Theory under the supervision of David Cohen-Steiner in the DataShape team.



About DataShape

We want to understand the *shape of data*.

- Geometrical and topological Inference
- Persistent Homology Theory
- Applications in Machine Learning and Biology

Today's Presentation

Goals

Today's goals :

- Briefly explain/remind you the key principles of Persistent Homology.
- Give you examples of its use in Data Analysis, to see if it can be of use in your work.

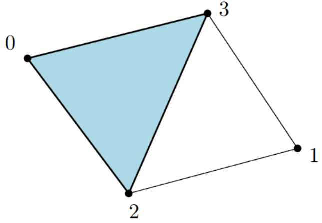
About Homology

Question : What is Homology ?

Intuitively, no need for exhausting definitions.

About Homology

Answer : a way to count holes or connected components !



General case

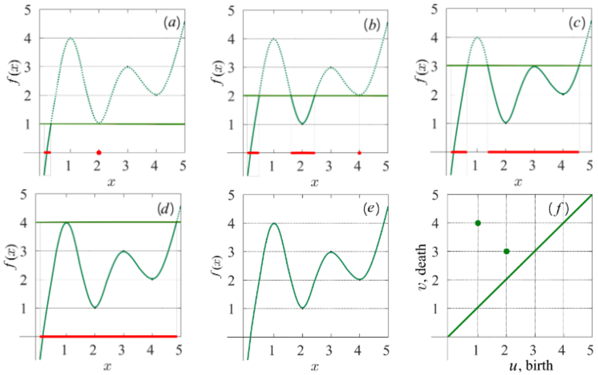
Filtration

Say you have sets $(X_t)_{t\in\mathbb{R}}$ increasing with t, that is

$$s < t \implies X_s \subset X_t$$

Keep track of the evolution of the **homology** (i.e the number of holes/connected components) of X_t.
Generally, given $f : X \to \mathbb{R}$ we take $X_t = f^{-1}(]-\infty, t])$.

General Case : sublevel filtration



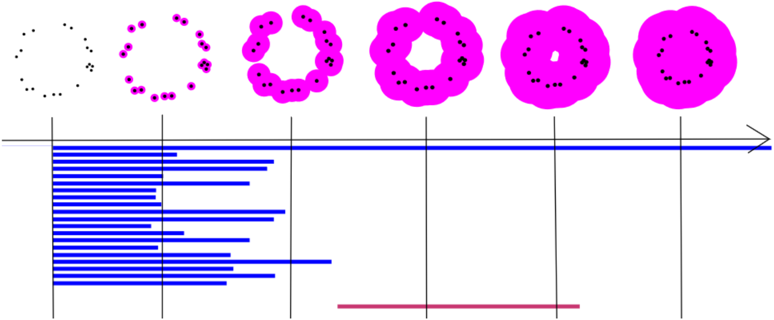
Working with distance functions

Let X be a point cloud.

Offset filtration

We can study the topology of the set X^r of points at distance to X smaller than r.

Working with distance functions



Working with distance functions

Persistent Homology is not always pertinent

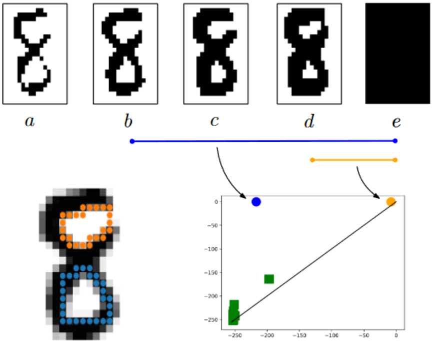
The data needs to be **clean**. Otherwise, we will not extract as much information.

Intensity of Pixels

Let X be a picture with one dimensional intensity (i.e grayscale).

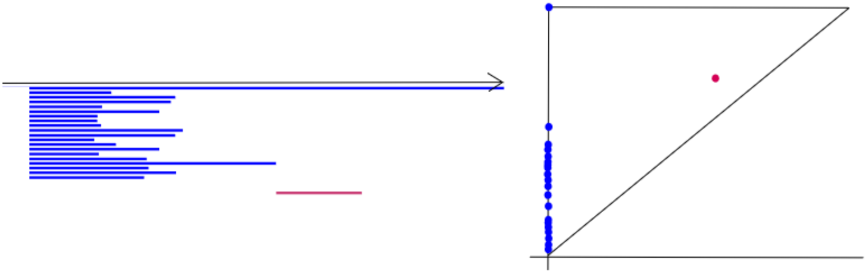
Intensity Filtration

We can study the topology of sets only containing the pixels with intensity smaller than r and watch how it evolves.



Persistent Diagrams

Each bar I is caracterized by its birth-death couple (b_I, d_I).
We can see it in the plane $\mathbb{R}^2$:



Comparing Diagrams

As set of points in $\mathbb{R}^2$, we can quantitatively compare diagrams through the so-called **bottleneck distance**.

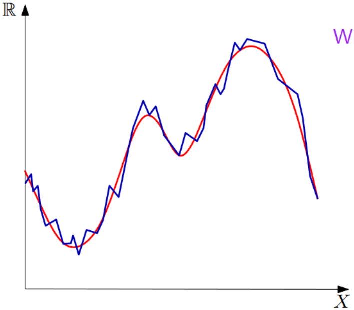
-> Allows to do **statistics** on topology!!

Comparing Diagrams

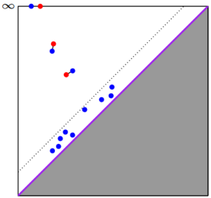
As set of points in $\mathbb{R}^2$, we can quantitatively compare diagrams through the so-called **bottleneck distance**.

Theorem (Stability):
For any *tame* functions $f, g : \mathbb{X} \to \mathbb{R}$, $\mathrm{d}_{\mathrm{B}}^{\infty}(\mathrm{D}_f, \mathrm{D}_g) \leq \|f - g\|_{\infty}$.

[Cohen-Steiner, Edelsbrunner, Harer 05], [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG 09], [C., de Silva, Glisse, Oudot 12]



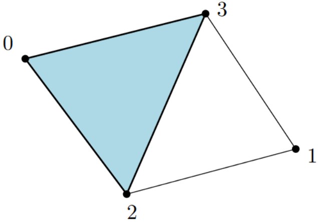
What if f is slightly perturbed?



2

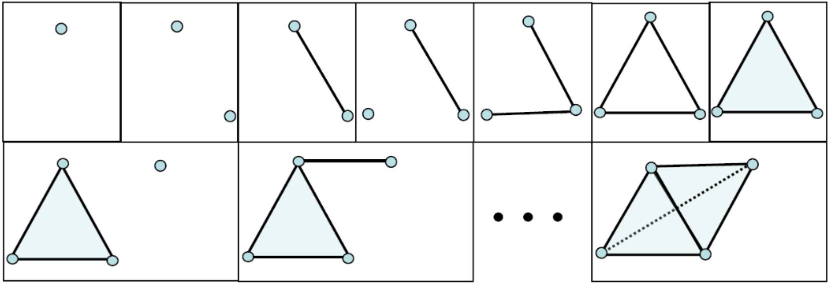
What does the computer do ?

The computer works with **simplicial complexes** which is a data structure.
It stores both **the vertices** and **how the vertices are linked**.

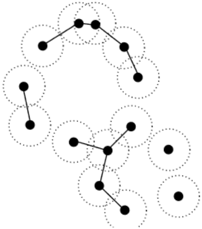


What does the computer do ?

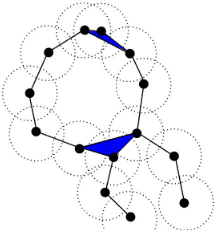
The computer works with **simplicial complexes** which is a data structure.
It stores **the vertices** and **how the vertices are linked**.



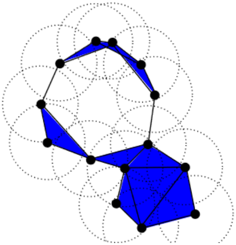
What does the computer do ?

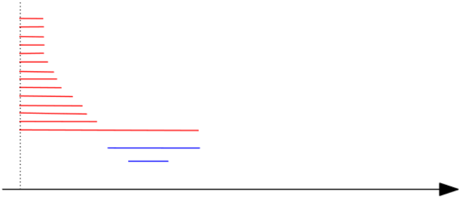


What does the computer do ?

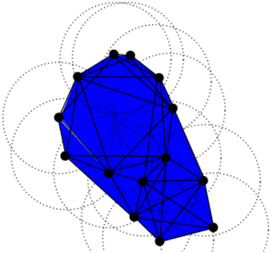


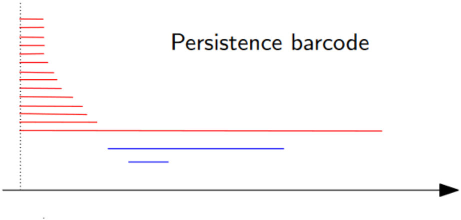
What does the computer do ?





What does the computer do ?





Worst case Complexity

Computing a persistence diagram requires **at most** $O(n^3)$ operations, where n is the number of simplices of the filtration.

Everyday Complexity

In practice, most persistence diagrams computations require **only** roughly $O(n)$ operations, where n is the number of simplices.

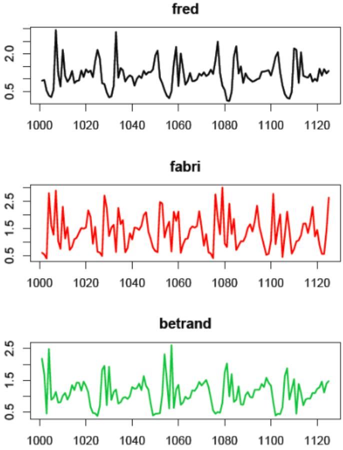
How to use it ?

There are different libraries for TDA.
Totally unbiased, I suggest the GUDHI package from the DataShape team !

Written in C++ for efficiency, with a high-level Python interface.

Toy example : Phones in a pocket

Fred, Fabrice and Bertrand have their own way to walk.



Toy example : Phones in a pocket

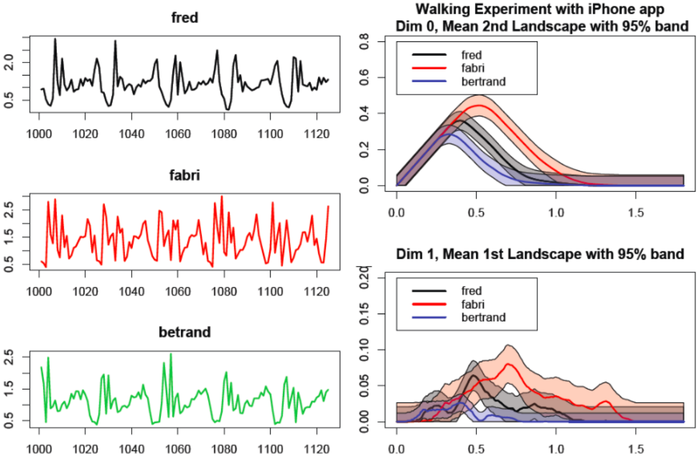
Fred, Fabrice and Bertrand have their own way to walk.

Idea : Using persistence over the functions

The functions are given by the accelerometer over walks.
-> *We cut the one-hour walk in small parts to do statistics*

Toy example : Phones in a pocket

Fred, Fabrice and Bertrand have their own way to walk.



An application to classification of Breast Cancer

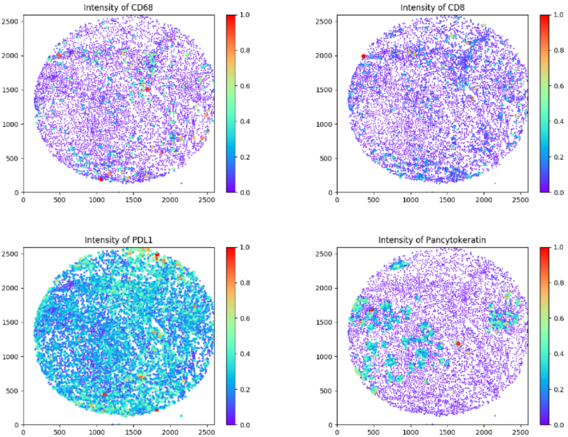
From a study using TDA to do statistical tests about breast cancer.

The data

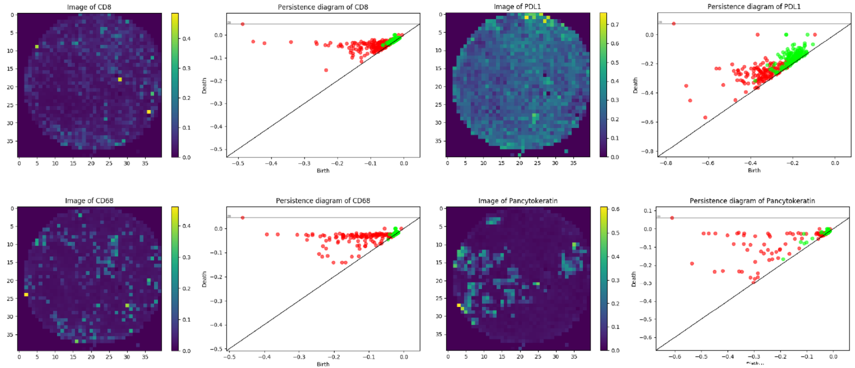
They had access to breast tissue samples and the knowledge of the disease's evolution.
Goal : to predict cancer subtype from the tissue.

An application to classification of Breast Cancer

Examples of a subsample :



An application to classification of Breast Cancer



An application to classification of Breast Cancer

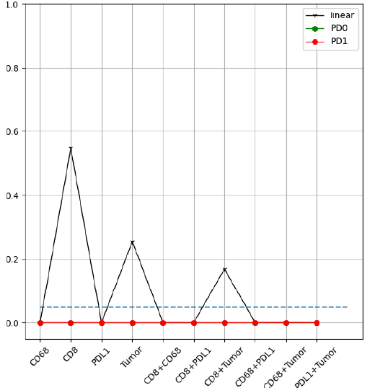
Idea

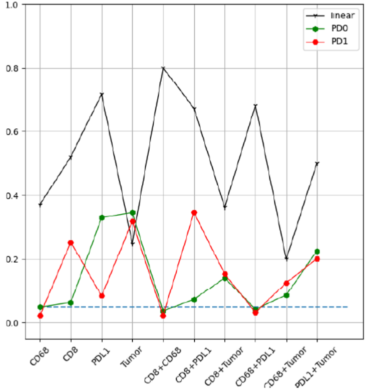
The authors' idea : combining different persistence diagrams to cleverly extract data.

An application to classification of Breast Cancer

Result

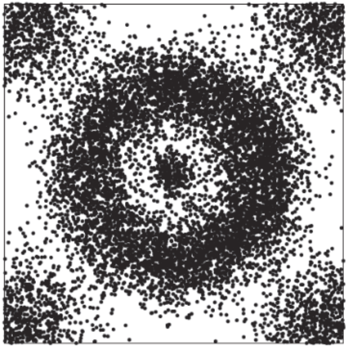
It ranked better than the state-of-the-art at the time!





Clustering using persistence

Clustering is hard when working with big points clouds.

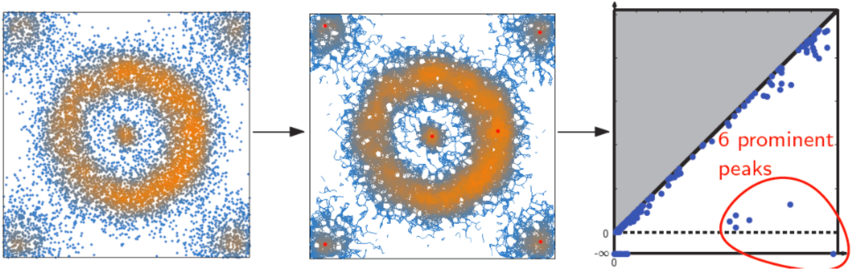


Luckily, there is ToMATo! *(Topological Mode Analysis Tool)*

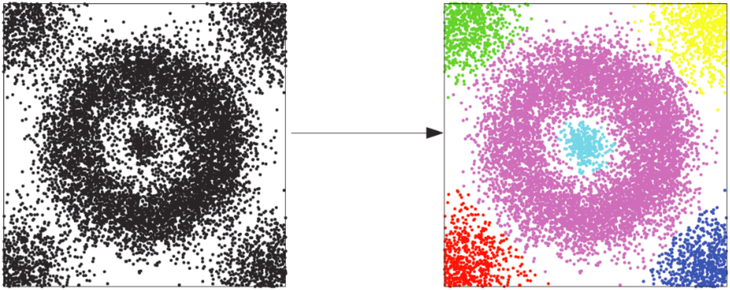
Clustering using persistence

Idea

We can filter **by the density of points** !



Clustering using persistence



Complexity

The complexity is $O(n\log(n))$ where n is the number of points.

Differentiating persistence diagrams - a quick overview

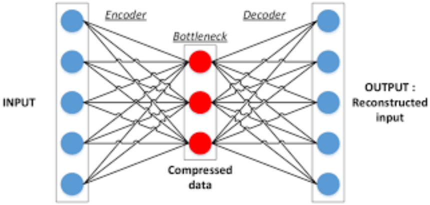
- You want to compute the persistence diagram of a picture.
- Its persistence diagram is a function of the intensity of the pixels

-> You can differentiate any persistence diagram !

Applications : Loss in a Neural Network

Given a Neural Network framework, topology can become important if you find a way to write a part of your loss as a function over the persistence diagrams.

Example : Topological Autoencoders



Loss : the distance between the persistence diagrams of the input/output.

Remerciements

Thank you for listening !

Feel free to ask any questions !