COMPUTATIONAL TOPOLOGY FOR DATA ANALYSIS

Javier Perera-Lago

University of Seville, Spain

jperera@us.es

About me

Javier Perera-Lago

- Double University Degree in Mathematics and Statistics (University of Seville)
- Master's University in Mathematics (University of Seville)
- Enrolled in the PhD program of Mathematics since October 2023
- Mail: jperera@us.es

Topology

Topology is a branch of mathematics that studies the properties of shapes and surfaces that remain invariant under continuous transformations (strecthing, bending, twisting...).

It is more flexible and more general than geometry.

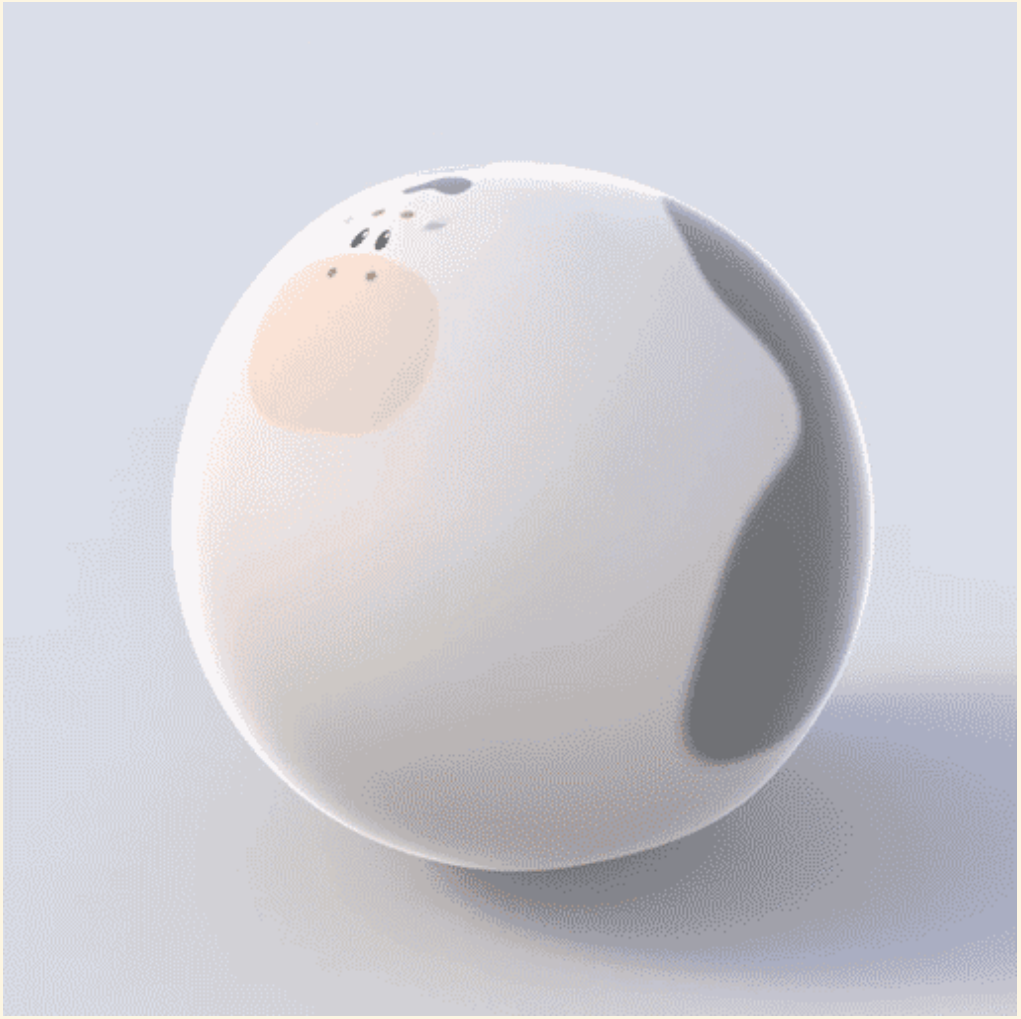
Geometrical properties	**Topological properties**
Area	Compactness
Volume	N° of connected pieces
Angles	Euler characteristic
Curvature	Fundamental group of loops
⋮	⋮

Topology

In geometry, these two objects are different.

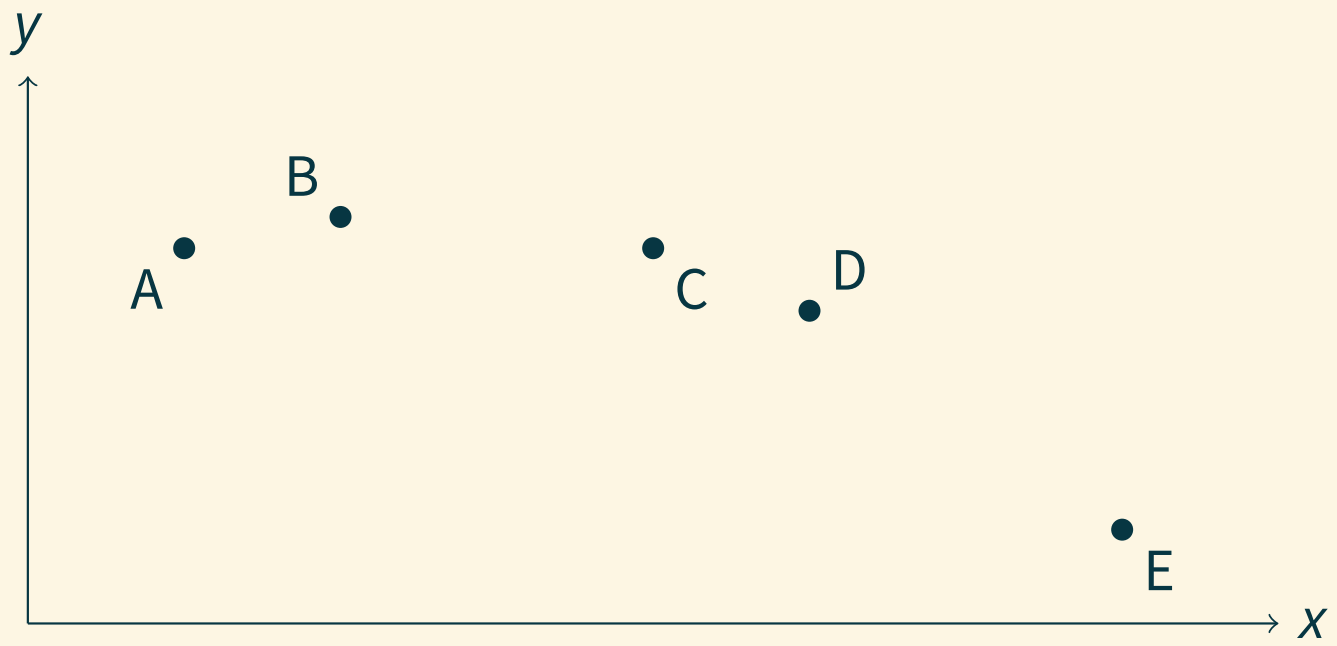
Topology

But they are ***exactly the same*** in topology!



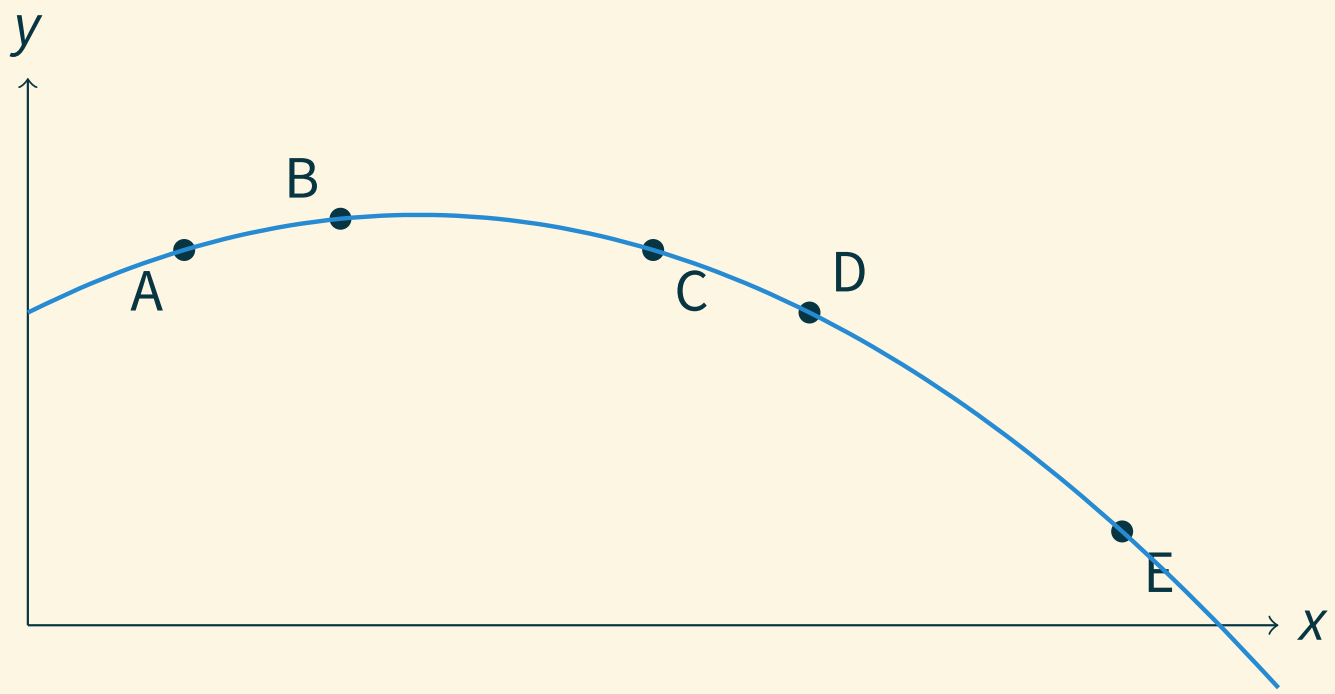
TDA: Topological Data Analysis

TDA is the branch within Data Analysis that uses concepts and tools from topology to extract information about the "shape of data".



TDA: Topological Data Analysis

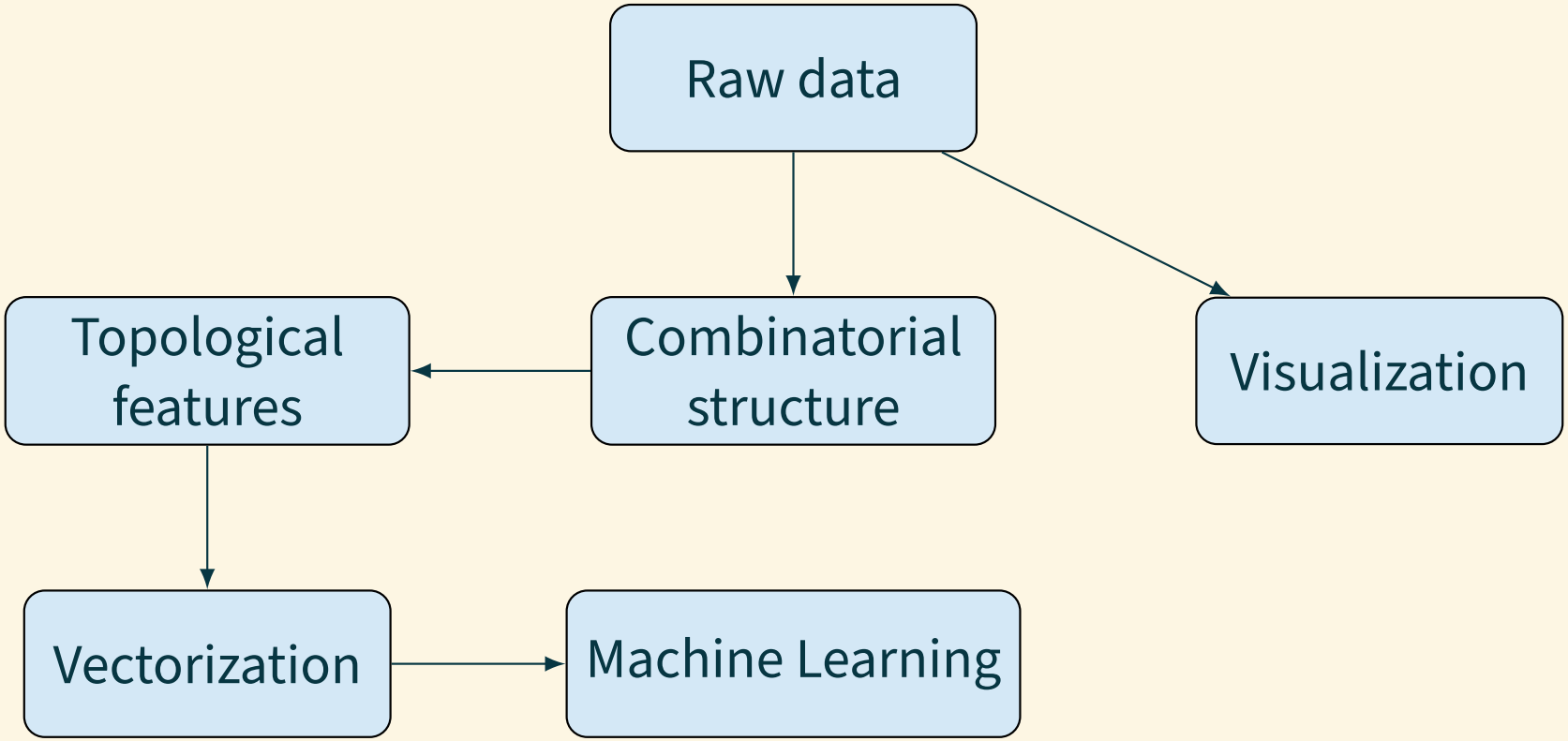
TDA is the branch within Data Analysis that uses concepts and tools from topology to extract information about the “shape of data”.



Manifold hypothesis: there exists an unknown manifold to which the dataset points belong.

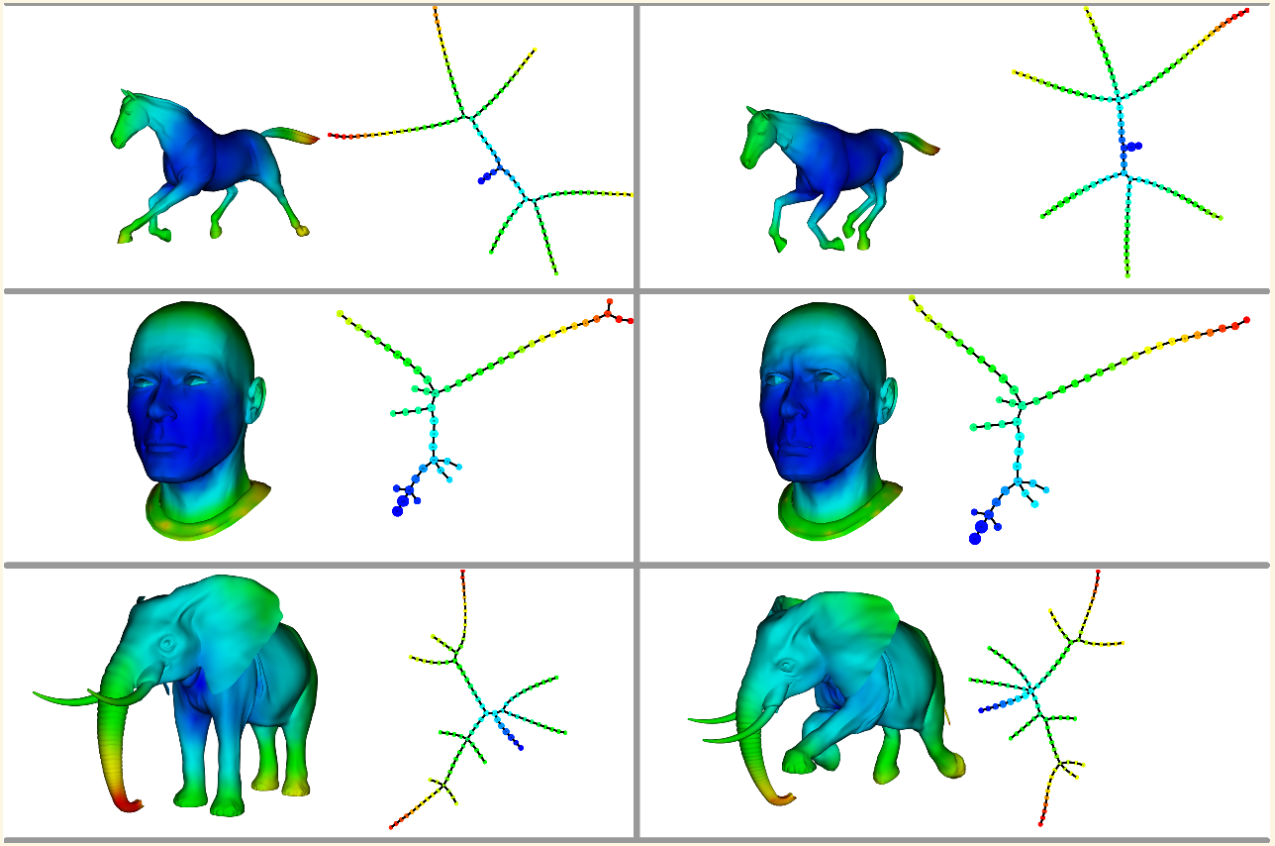
TDA: Topological Data Analysis

The general pipeline in TDA is:



Visualization: Mapper graph

Mapper is an algorithm that helps to visualize high-dimensional datasets in the form of a graph by using a filter function.

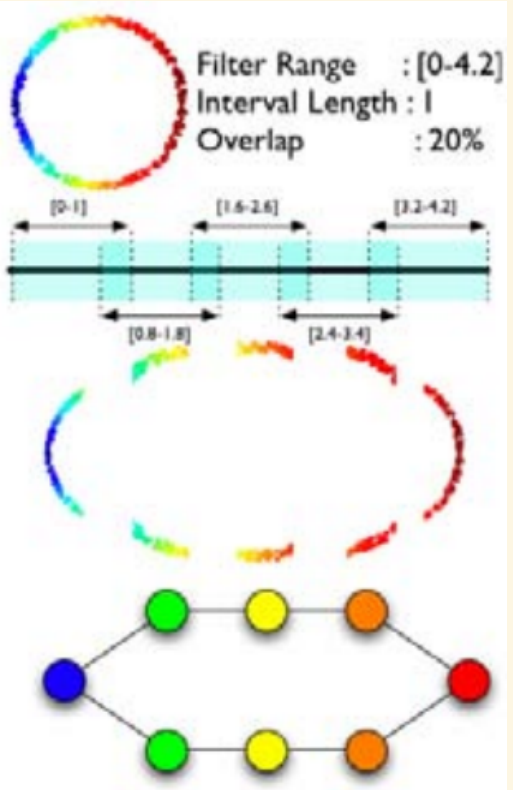


[Singh et al., 2007]

Visualization: Mapper graph

Given a dataset X, the pipeline to build a Mapper graph on it is:

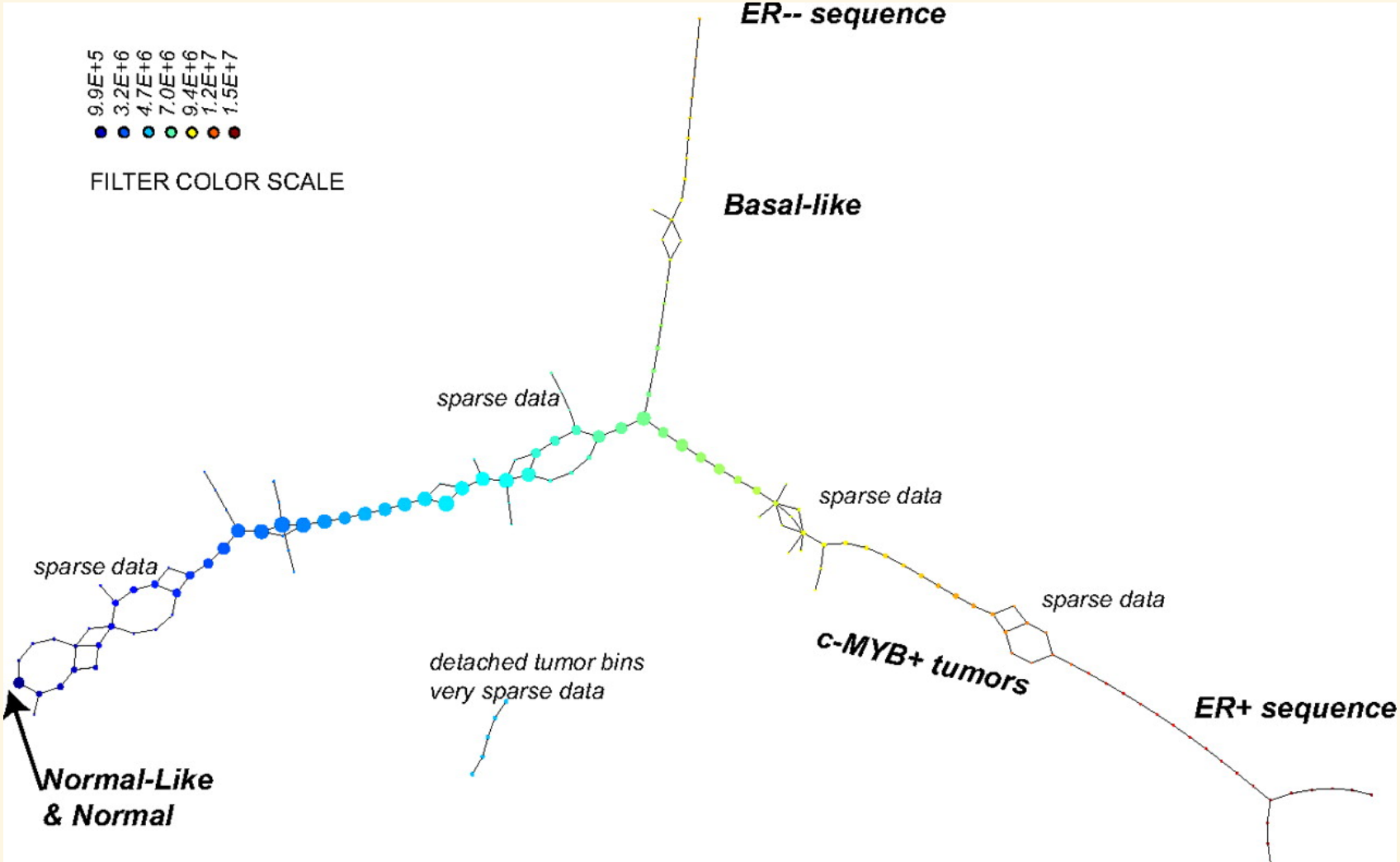
1. Consider a function $f : X \to \mathbb{R}$.
2. Cover the real line $\mathbb{R}$ with a set of overlapping intervals $\{I_a\}_{a \in A}$.
3. For each $a \in A$, build the sub-dataset $X_a = \{x \in X | \, f(x) \in I_a\}$.
4. For each $a \in A$, divide X_a into clusters.
5. Draw a node for each cluster.
6. Draw an edge between two clusters if there is any point in common.



[Singh et al., 2007]

Visualization: Mapper Graph

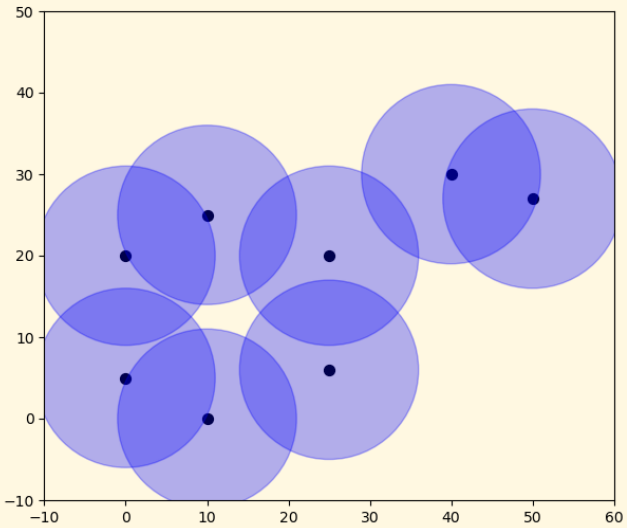
In 2011, the Mapper graph helped to discover a new subgroup of breast cancers with a unique mutational profile and excellent survival [Nicolau et al., 2011]

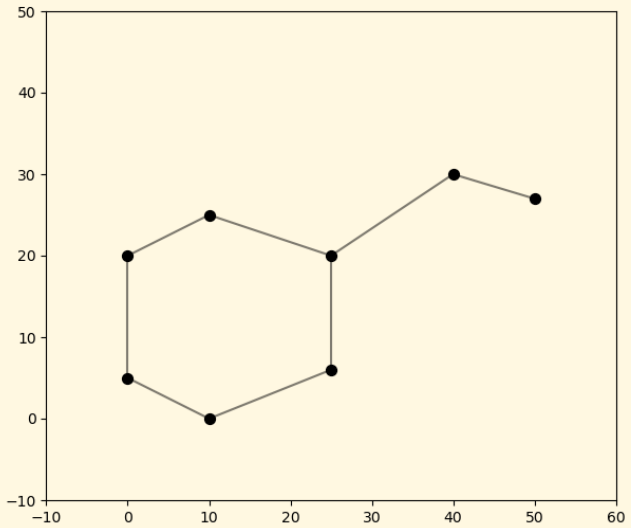


Combinatorial structures

The manifold hypothesis states that the dataset points lie on a manifold.

We approximate it by building a ***simplicial complex***, which is a combinatorial structure composed by points, edges, triangles, tetrahedra... (it's a generalization of graphs)





This is called a Cech complex.

Simplicial homology

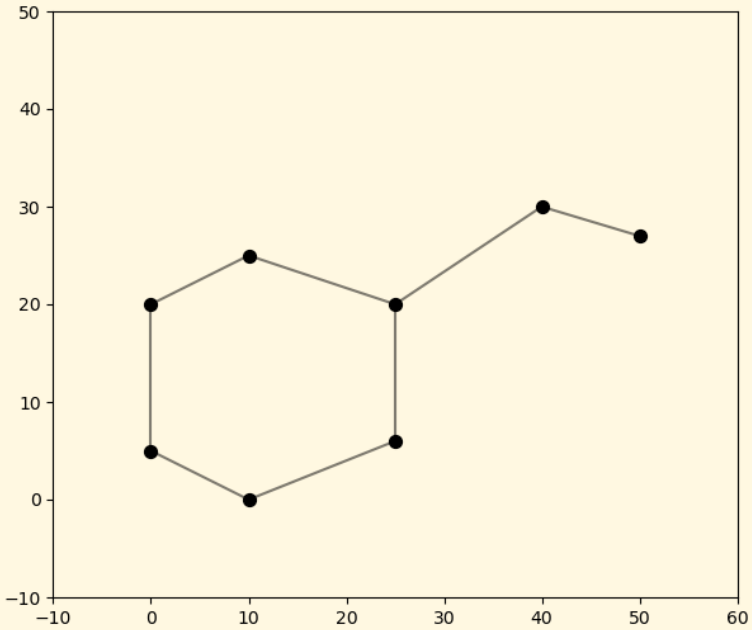
Given a simplicial complex, we compute its ***homology groups*** to extract topological information. There is one homology group H_n for each possible dimension n.

The rank of the group H_n is called the ***Betti number*** β_n. These numbers have an intuitive interpretation:

- β_0 is equal to the number of connected components.
- β_1 is equal to the number of distinct loops.
- β_2 is equal to the number of distinct voids.
- And so on...

Simplicial homology

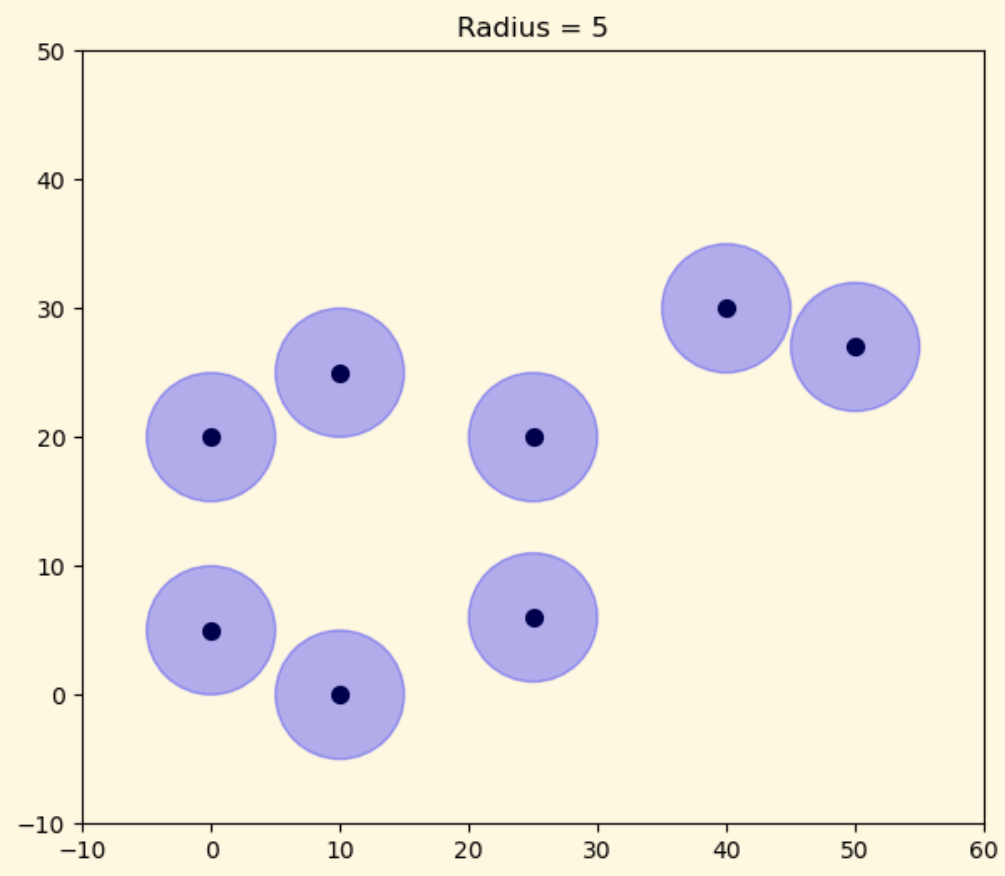
For example, this complex has the following Betti numbers:



$$\begin{cases} \beta_0 = 1 & \text{(there is one connected component)} \\ \beta_1 = 1 & \text{(there is one loop)} \\ \beta_n = 0 \; \forall n \geqslant 2 & \text{(there are no more topological features)} \end{cases}$$

Combinatorial structures

It is impossible to know a priori which is the best radius to build a Cech complex, so we can build a flitration of nested complexes by varying the radius.

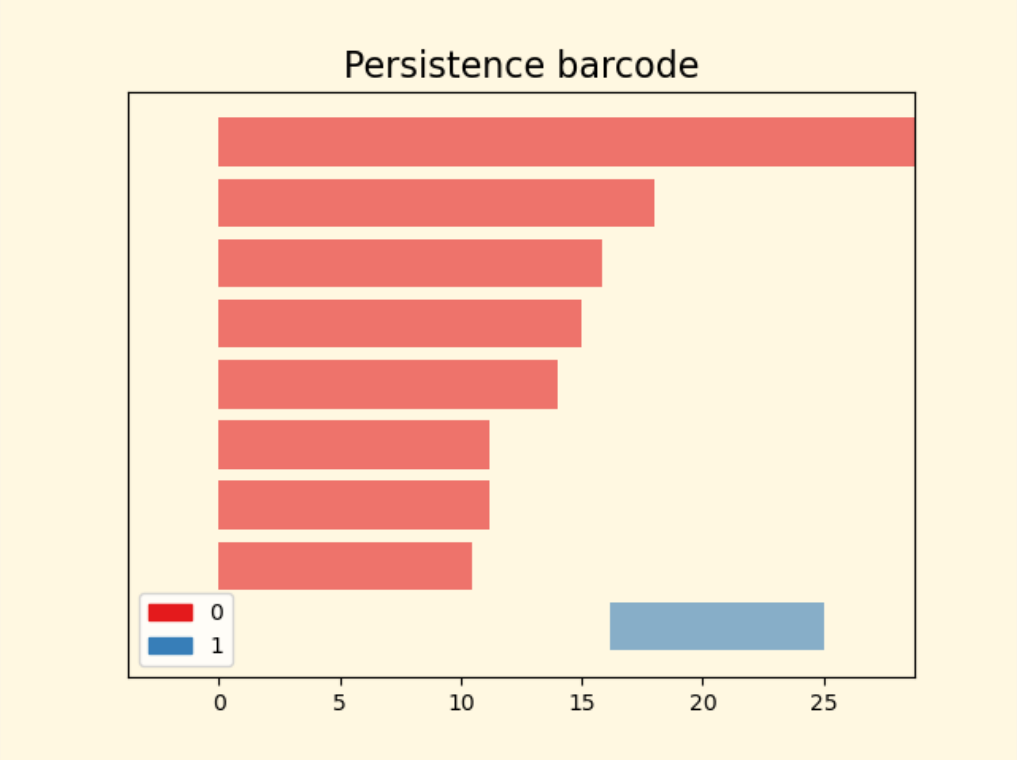


$$\beta_0 = 8.0$$

$$\beta_1 = 0.0$$

Persistent homology

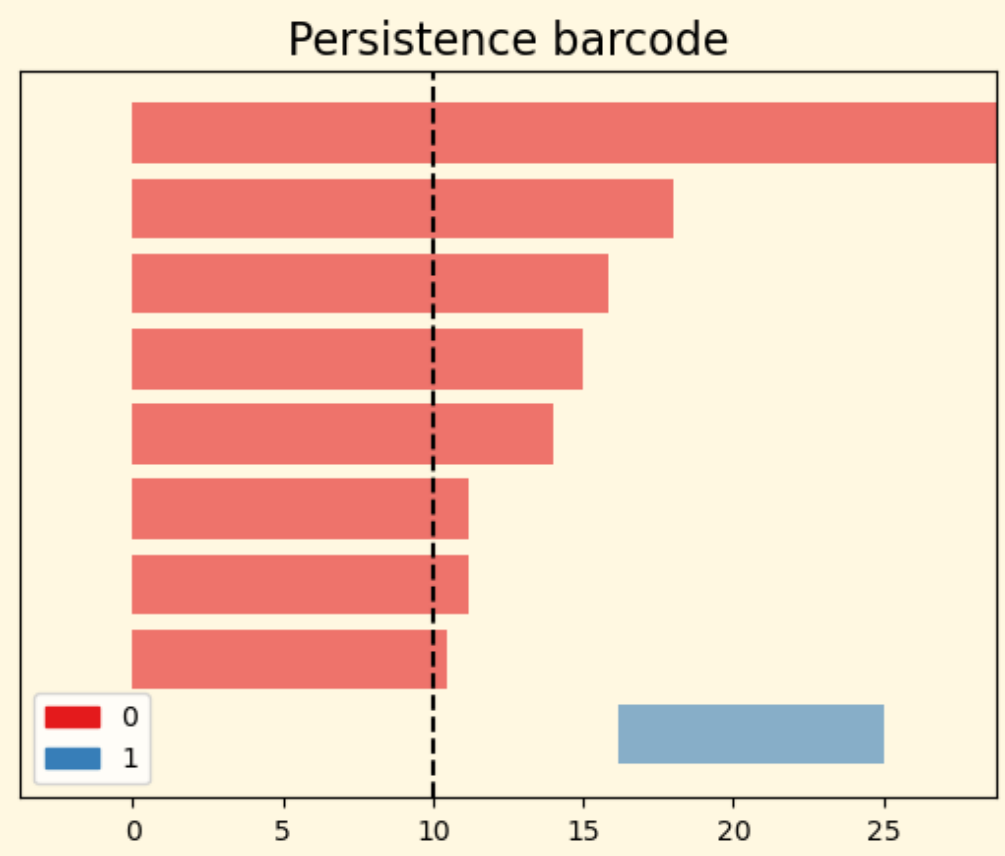
Persistent homology computes the homology groups for each complex in the filtration and merges them in a single algebraic object.



This is the ***barcode*** of the persistent homology.

Persistent homology

Persistent homology gives us the homology of each complex, but it contains more information.

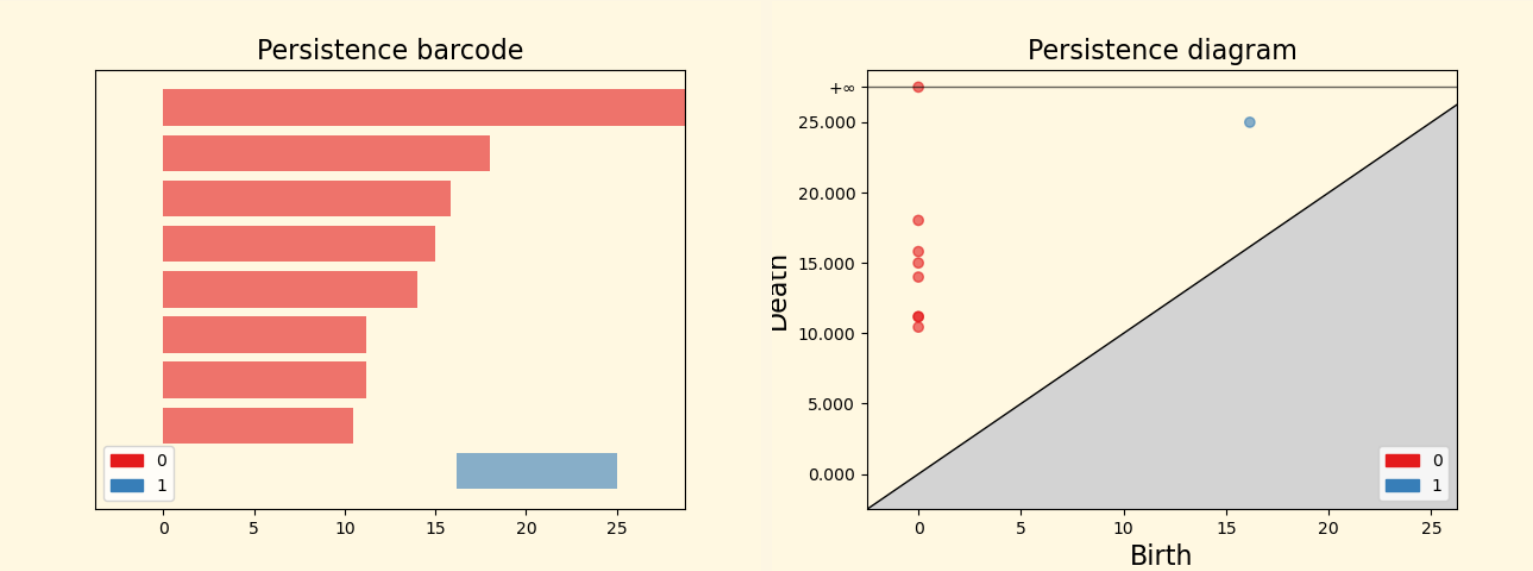


$$\beta_0 = 8.0$$

$$\beta_1 = 0.0$$

Persistent homology

If for any interval $[b, d]$ we draw the point (b, d) in the Cartesian axis, we get the ***persistent diagram***, which is a more geometrical representation of persistence.



Persistent homology

Stability theorem for persistent diagrams
[Cohen-Steiner et al., 2005].

The bottleneck distance between two diagrams is:

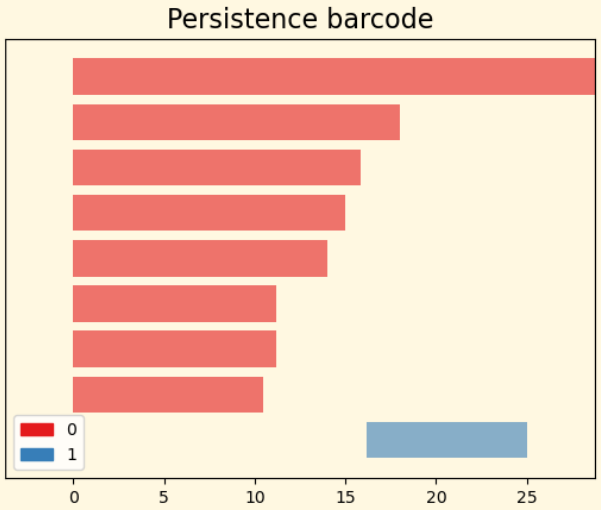
$$d_B(D, D') = \inf_{\varphi : D \to D'} \sup_{p \in D} \|p - \varphi(p)\|$$

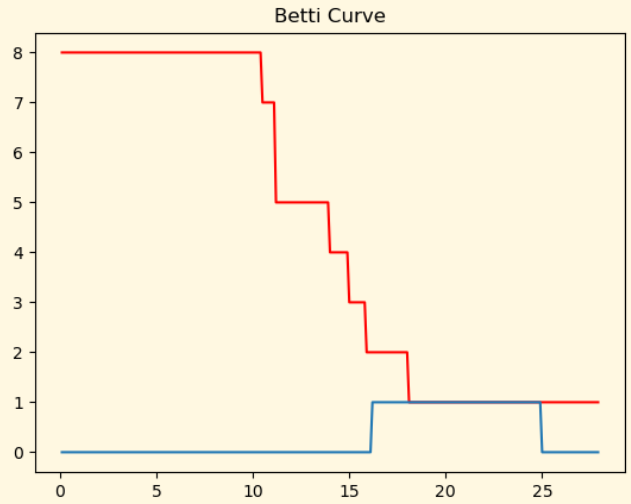
Let X be a triangulable space with continuous tame functions $f, g : X \to \mathbb{R}$. Then the persistence diagrams satisfy

$$d_B(D(f), D(g)) \leqslant \|f - g\|_\infty.$$

Vectorization

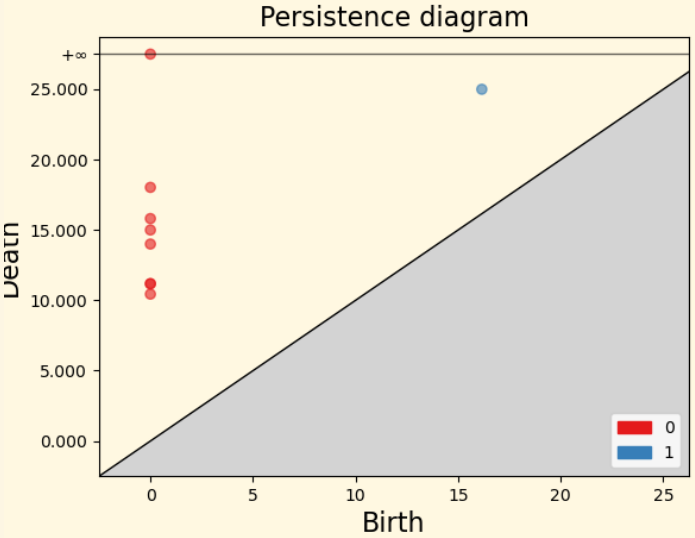
Ways to vectorize a barcode: ***Betti curves***

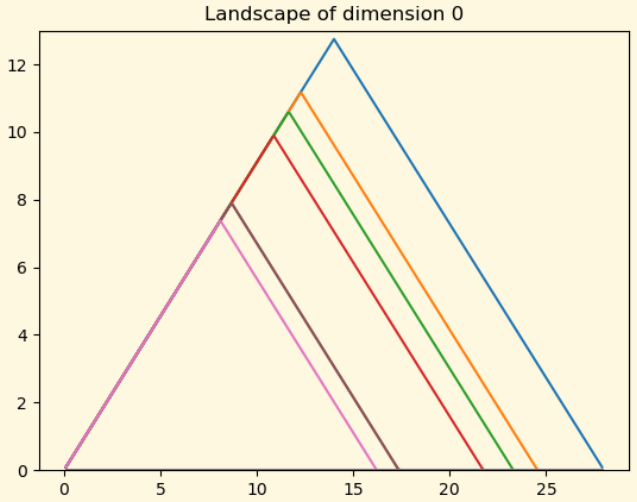


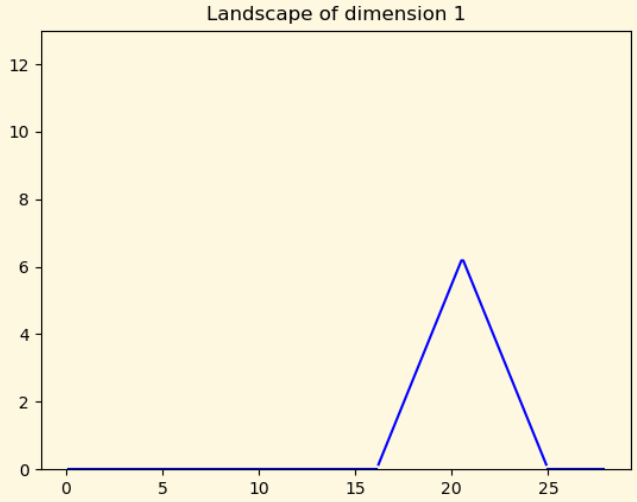


Vectorization

Ways to vectorize a diagram: ***Landscapes***

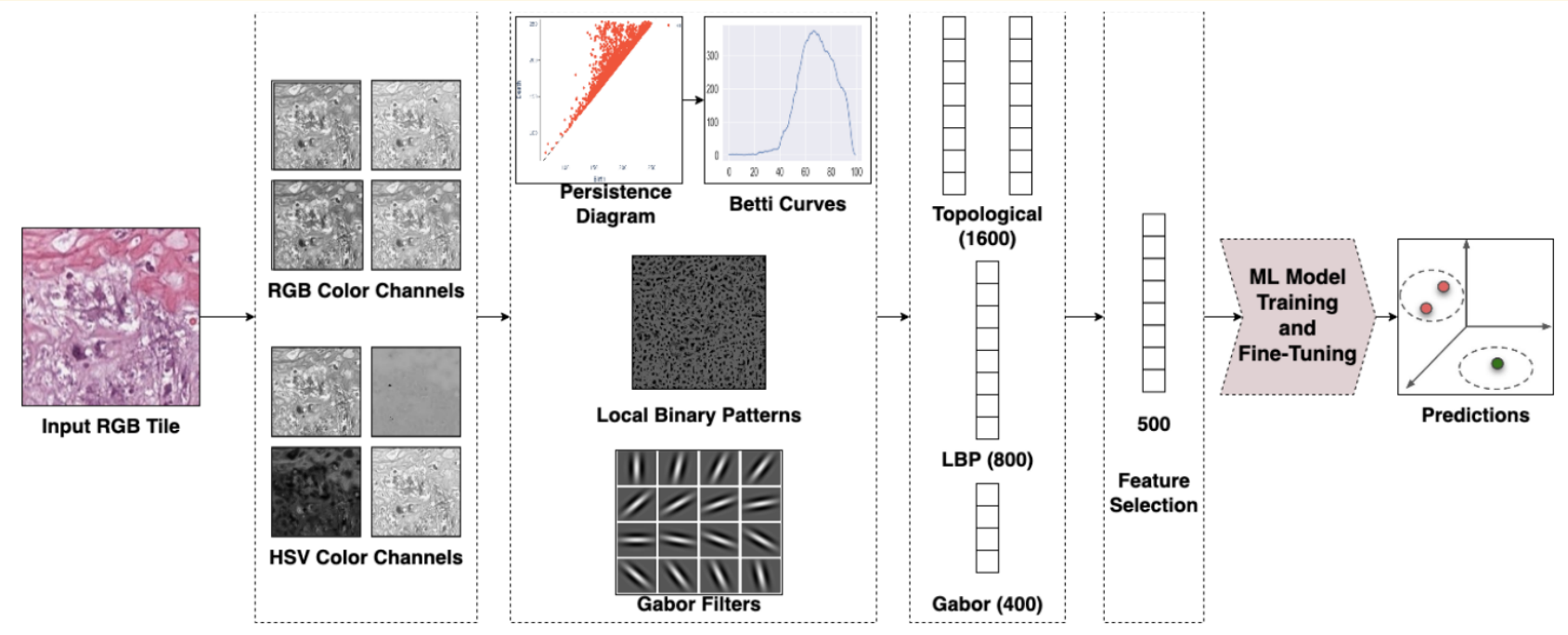






Vectorization

Vectorization has been used recently to detect cancer by analyzing histopathological images [Yadav et al., 2023].



Vectorization

More information on vectorization can be found in the article:

A survey of vectorization methods in topological data analysis.
Ali, D., Asaad, A., Jimenez, M. J., Nanda, V., Paluzo-Hidalgo, E., & Soriano-Trigueros, M. (2023). IEEE Transactions on Pattern Analysis and Machine Intelligence.

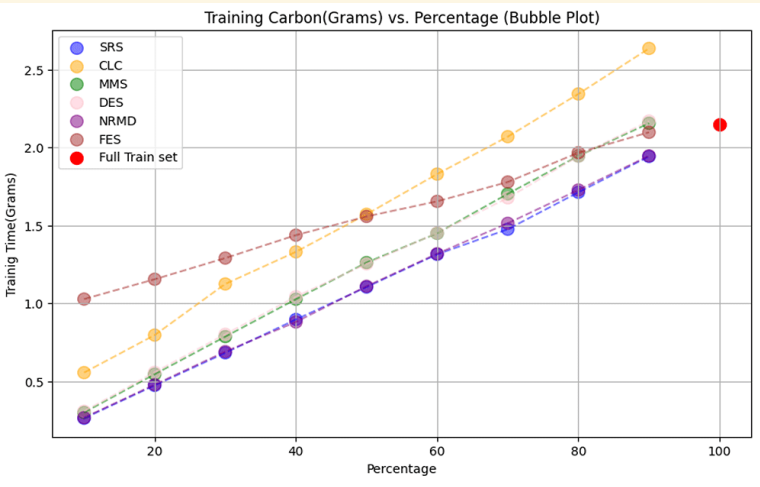
REXASIPRO

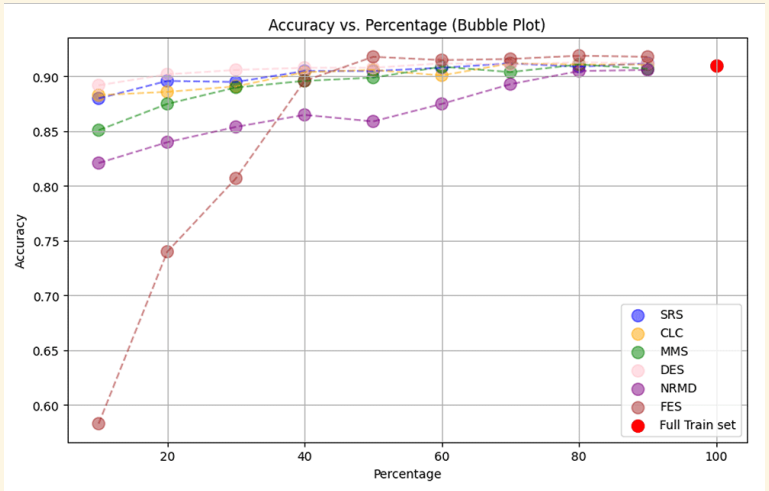
TDA techniques are helping mainly in two tasks of the ***REXASIPRO*** project:

- Task T6.2: Topology-based energy consumption optimization of Pedestrian Detection algorithm
- Task T6.3: : Topology-based optimization of robot fleet behavior

REXASIPRO: T6.2

Topology to assess data reduction.



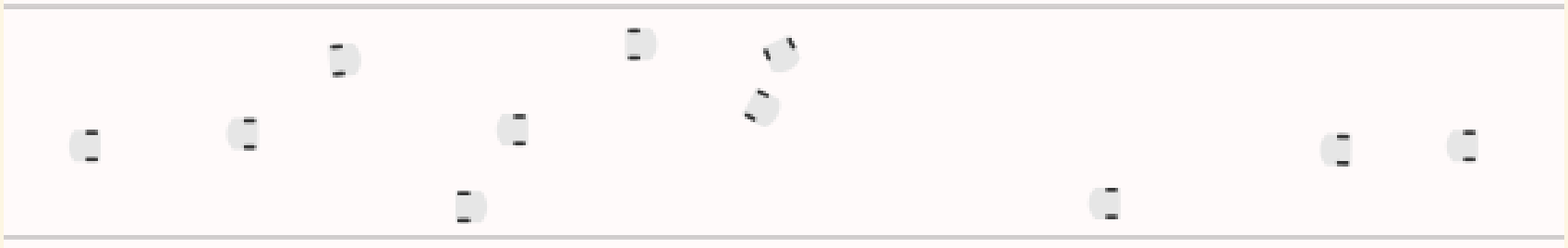


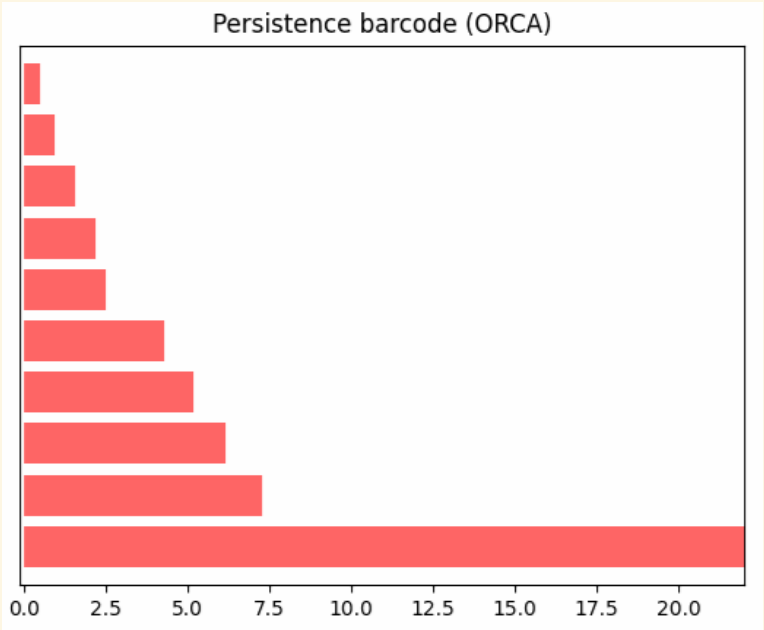
An in-depth analysis of data reduction methods for sustainable deep learning. Perera-Lago, J., Toscano-Duran, V., Paluzo-Hidalgo, E., Gonzalez-Diaz, R., Gutiérrez-Naranjo, M. A., & Rucco, M. (2024). Open Research Europe, 4(101), 101.

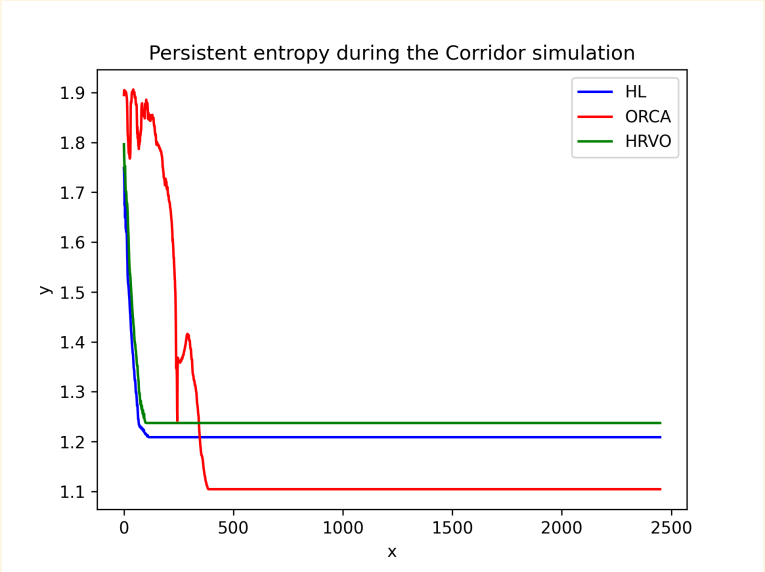
YouTube video

REXASIPRO: T6.3

Using TDA to compare different agent behaviors (work in progress).







References

- Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2005). Stability of persistence diagrams. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 263–271.
- Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.
- Singh, G., Mémoli, F., Carlsson, G. E., et al. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2:091–100.
- Yadav, A., Ahmed, F., Daescu, O., Gedik, R., and Coskunuzer, B. (2023). Histopathological cancer detection with topological signatures.