Advanced Mathematics for Scientific Challenges

2022–2023

Topological Data Analysis

Opening Presentation

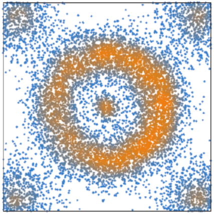
Topological Data Analysis

Goal: To analyze datasets possibly high-dimensional and noisy

Method: Detect and represent shape features such as connectivity, loops, cavities, flares, or clusters

Qualitative Analysis → **Mapper**

Quantitative Analysis → **Persistent Homology**



Mapper

Mapper is a data visualization algorithm combining

- dimensionality reduction,
- clustering,
- graph analytics.

G. Singh, F. Mémoli, G. Carlsson (2007), *Topological methods for the analysis of high dimensional data sets and 3D object recognition*, Eurographics Symposium on Point-Based Graphics

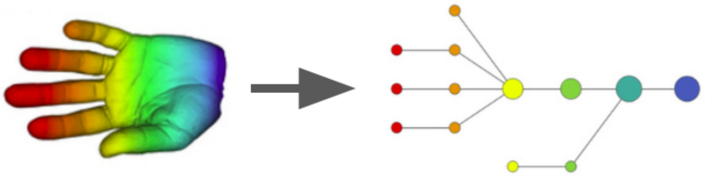
Mapper

Input:

- A data set X,
- a parameter space Z (a subset of $\mathbb{R}$ or $\mathbb{R}^2$),
- a function $f\colon X \to Z$, called a **filter function,**
- and a clustering algorithm.

Output:

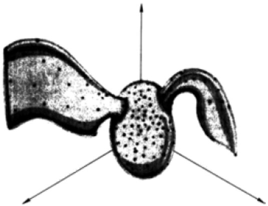
- A coloured graph.

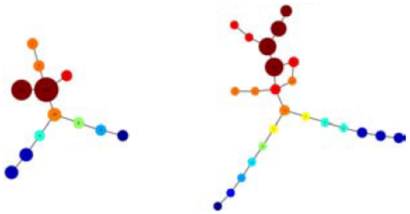


Mapper

Example: The Miller–Reaven diabetes study (1985)

Six variables were measured in a sample of 145 patients, yielding a 6-dimensional data set.

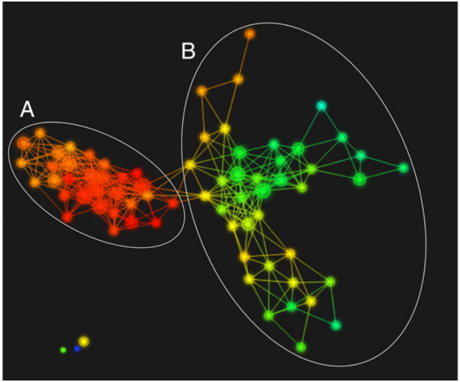




3-D image in the original study using projection and pursuit. Flares are type I and type II diabetes.

Mapper graphs with 3 and 4 filter intervals. Size of nodes indicate size of clusters. Colours indicate density. The blue ends represent the flares.

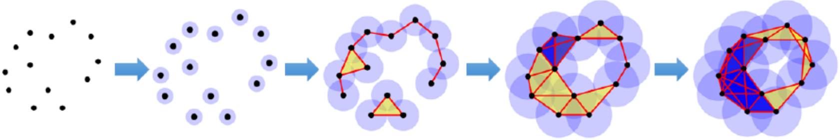
Mapper



J. L. Bruno et al. (2017), *Longitudinal identification of clinically distinct neurophenotypes in young children with fragile X syndrome*, PNAS 114(40), 10767–10772

Persistent Homology

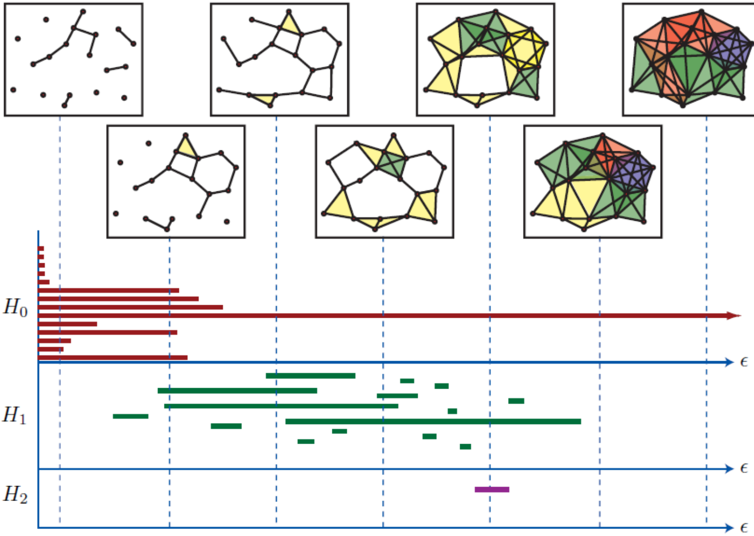
Data Set → Filtered Simplicial Complex → **Persistent Homology** → **Barcode**



Homology groups of a simplicial complex X:

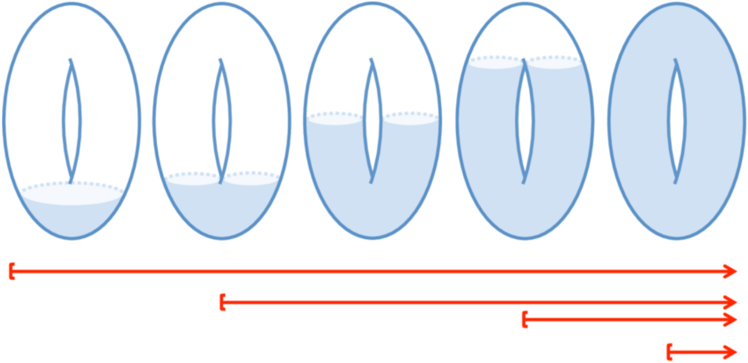
- $H_0(X)$ counts connected components of X;
- $H_1(X)$ counts 1-dimensional cycles in X;
- $H_2(X)$ counts 2-dimensional cavities in X; etc.

Barcodes



Barcodes

Morse functions on compact manifolds also yield barcodes:



Each homology generator is *born* at a certain height.

Barcodes

Stability Theorem

For two point clouds X and Y in the same ambient space,

$$\boldsymbol{W_\infty(B(X), B(Y)) \leq 2\, d_{GH}(X, Y)},$$

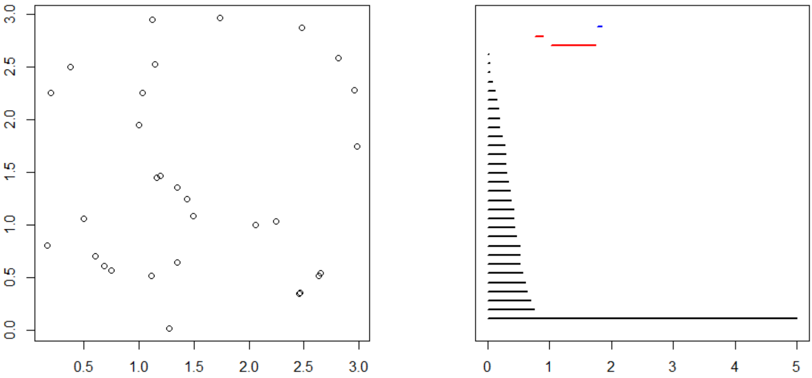
where

- $B(X)$ and $B(Y)$ denote the barcodes of X and Y;
- W_∞ is the **bottleneck distance** between barcodes;
- d_{GH} is the **Gromov–Hausdorff distance.**

A similar formula holds for barcodes of Morse functions f and g:

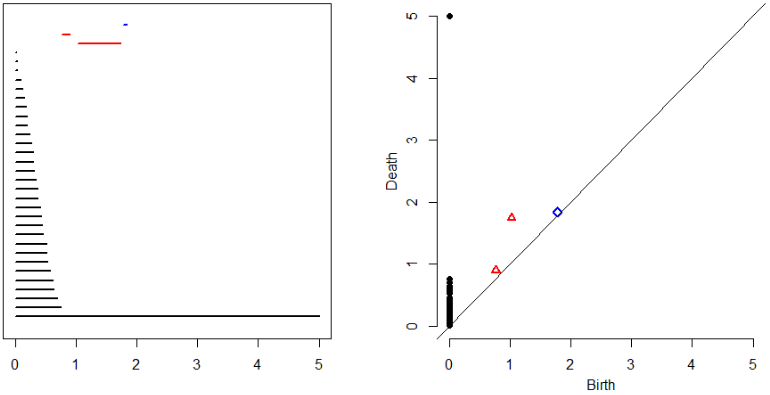
$$\boldsymbol{W_\infty(B(f), B(g)) \leq \|f - g\|_\infty}.$$

Barcodes



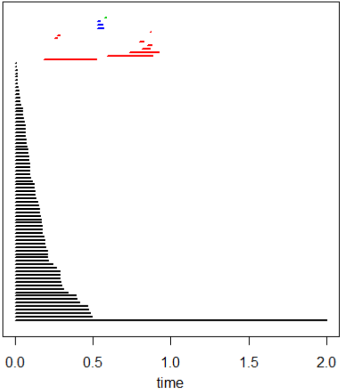
Persistence barcode for a point cloud with $N = 30$. There are homology generators in dimensions 0 (black), 1 (red) and 2 (blue).

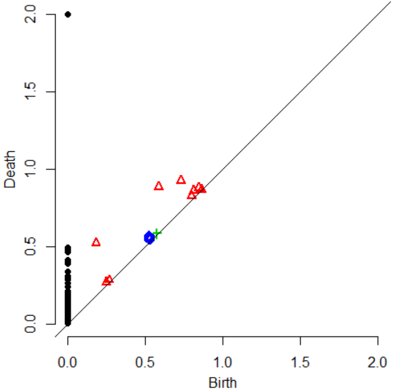
Persistence Diagrams



The coordinates (b, d) of each point in a **persistence diagram** correspond to *birth* and *death* of a homology generator.

Persistence Diagrams





Points near the diagonal are generally viewed as *noise*.

Persistence Descriptors

A **persistence descriptor** is a numerical summary or a vectorized summary from persistence diagrams.

Numerical summaries

- Average life
- Average midlife
- Entropy

Vectorized summaries

- Betti curves
- Landscapes and silhouettes
- Persistence images
- Kernels

Numerical Summaries

Average life:

$$\frac{1}{n}\sum_{i=1}^{n}(d_i - b_i)$$

Average midlife:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{b_i + d_i}{2}$$

Entropy:

$$-\sum_{i=1}^{n}\frac{d_i - b_i}{L}\log_2\left(\frac{d_i - b_i}{L}\right), \quad \text{where} \quad L = \sum_{i=1}^{n}(d_i - b_i).$$

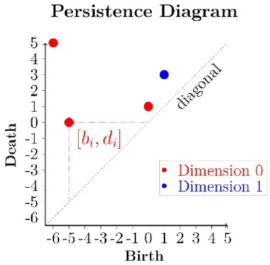
The **entropy** of a random variable is the average level of uncertainty inherent in its outcomes (Shannon, 1948).

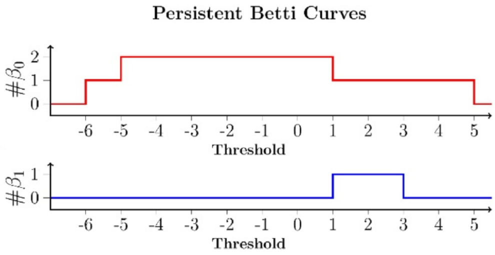
Betti Curves

For each $k \geq 0$, let $\beta_k : \mathbb{R} \to \mathbb{R}$ be defined as

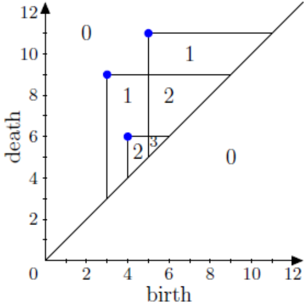
$$\boldsymbol{\beta_k(t) = \#\{(b, d) \mid b \leq t \leq d\},}$$

where (b, d) ranges over the points in a given persistence diagram for homological dimension k.

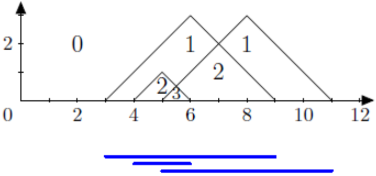




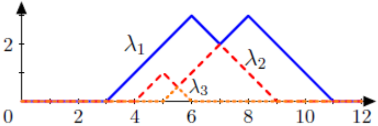
Landscapes



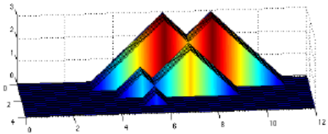
death
birth
0
2
4
6
8
10
12
0
1
1
2
2
3
0



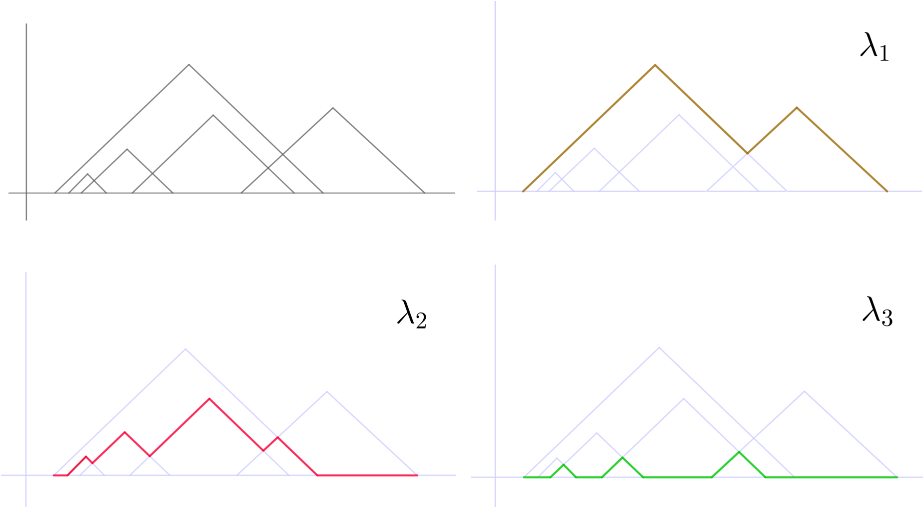
0
2
4
6
8
10
12
2
0
1
1
2
2
3



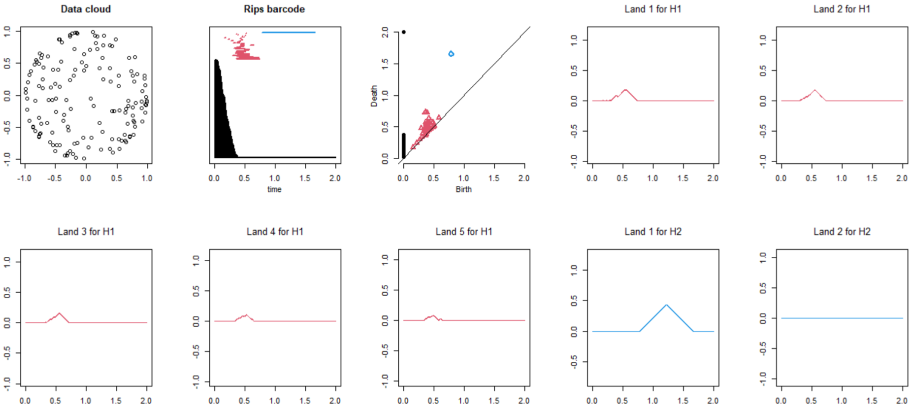
λ_1
λ_2
λ_3
0
2
4
6
8
10
12
2



Landscapes



Landscapes



Data cloud
Rips barcode
time
Death
Birth
Land 1 for H1
Land 2 for H1
Land 3 for H1
Land 4 for H1
Land 5 for H1
Land 1 for H2
Land 2 for H2

Silhouettes

A **silhouette** of a persistence diagram with m points (b_i, d_i) is a weighted average of landscape tent functions

$$\phi(t) = \frac{\sum_{i=1}^{m} w_i \, \Lambda_{(b_i,d_i)}(t)}{\sum_{i=1}^{m} w_i}$$

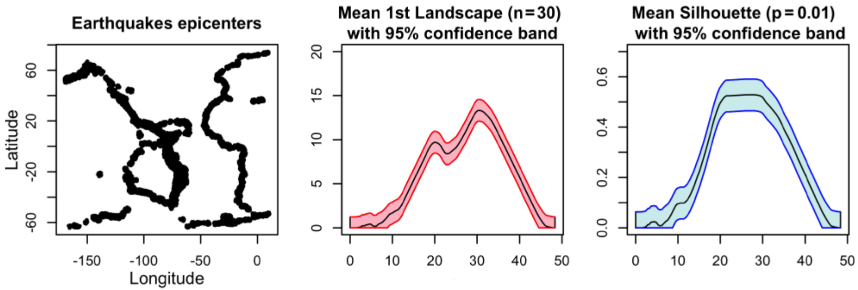
where $\{w_i\}$ are weights to be chosen, and

$$\Lambda_{(b,d)}(t) = \max\{0, \min\{t - b, d - t\}\}.$$

A frequent choice is $\boldsymbol{w_i = (d_i - b_i)^p}$ where p is optional:

- Choosing p small enhances low-persistence features.
- Choosing p large enhances highly persistent features.

Silhouettes



F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman (2014), *Stochastic convergence of persistence landscapes and silhouettes*, SOCG'14: Proceedings of the Thirtieth Annual Symposium on Computational Geometry, 474–483

P. Bubenik (2015), *Statistical topological data analysis using persistence landscapes*, J. Mach. Learn. Res. 16, 77–102

Persistence Images

For a given persistence diagram, consider a function

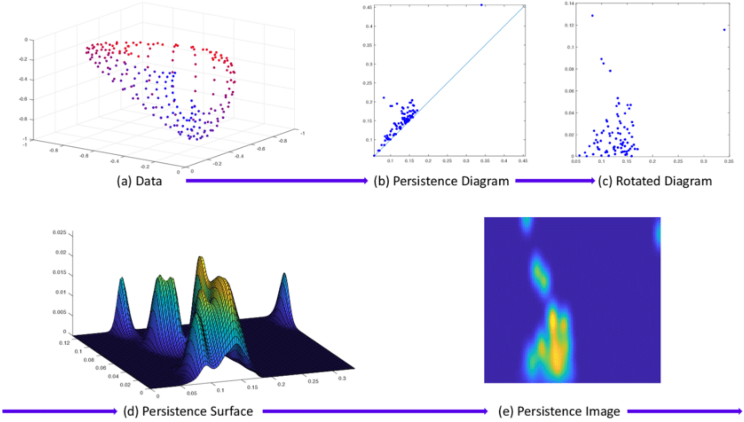
$$\Phi(s, t) = \sum_{i=1}^{n} w_i \, G_i(s, t)$$

for (s, t) in a square, where each w_i is a weight and G_i is a 2-dimensional Gaussian function centered at (b_i, d_i).

This yields a smoothing of the persistence diagram called a **persistence surface.**

A **persistence image** is a discretization of Φ on a grid overlay.

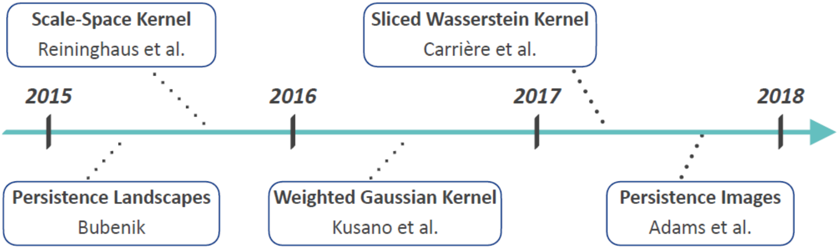
Persistence Images



Generate a surface by centering 2D Gaussian distributions at each point, and generate a **persistence image** by summing the volume under the Gaussian distributions over the area of each pixel.

Kernels

J. Reininghaus, S. Huber, U. Bauer, R. Kwitt (2015), *A stable multi-scale kernel for topological machine learning*, 2015 IEEE Conference on Computer Vision and Pattern Recognition, 4741–4748



Kernels provide a **dissimilarity measure** between persistence diagrams.

Biomedical Sciences

F. Belchí, M. Pirashvili, J. Conway et al. (2018),
Lung topology characteristics in patients with chronic obstructive pulmonary disease, Scientific Reports 8, 5341

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease characterized by chronic inflammation of the bronchi and the lung parenchyma.

Objectives: To develop, by means of Topological Data Analysis, a set of new radiomic features that can distinguish between healthy non-smokers, healthy smokers, and patients with COPD.

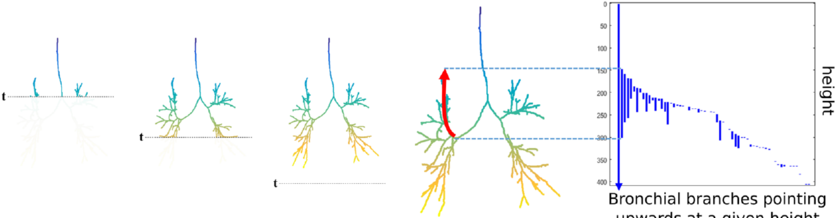
Population: For 30 participants (8 healthy non-smokers, 9 healthy smokers, 8 mild COPD and 5 moderate COPD), both inspiratory and expiratory tomography scans were obtained.

Biomedical Sciences

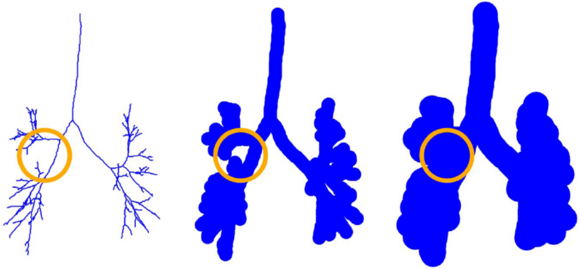
Methodology: Persistent homology in degrees 0, 1 and 2 was used in different ways to obtain different kinds of clinical insight.

- In degree 0, it was used to define **upwards complexity.**
- Persistent homology in degree 1 was used to measure **branch-to-branch proximity.**
- The degree 2 was used to overcome the limitation of the low spatial resolution of tomography scans by including information about the space between the airways and the outer boundary of the lobes.

Biomedical Sciences



Bronchial branches pointing upwards at a given height



Machine Learning

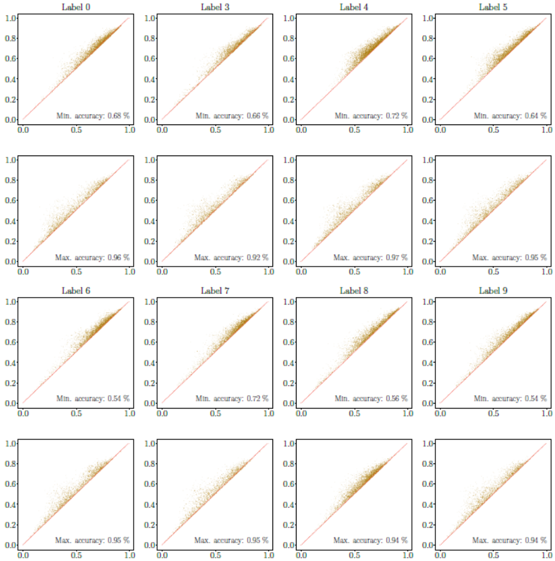
R. Ballester, X. Arnal, C. Casacuberta et al. (2021),
Predicting the generalization gap in neural networks using topological data analysis, arXiv:2203.12330 [cs.LG]

Method: Compute persistence diagrams of weighted graphs constructed from neuron activation correlations in a deep neural network after a training phase with a given dataset.

Goal: To capture patterns that are linked to the generalization capacity of the neural network.

Results: The generalization gap can be consistently predicted using persistence descriptors extracted from functional graphs.

Machine Learning



TDA Software

- **GUDHI** (*Geometry Understanding in Higher Dimensions*)
  http://gudhi.gforge.inria.fr
- **Dionysus**
  https://mrzv.org/software/dionysus2/
- **Ripser**
  https://live.ripser.org/
- The **R** package **TDAstats**
  https://cran.r-project.org/web/packages/TDAstats/index.html
- The **Matlab** library **JavaPlex**
  http://appliedtopology.github.io/javaplex/