Topological Machine Learning Seminar

14 January 2022

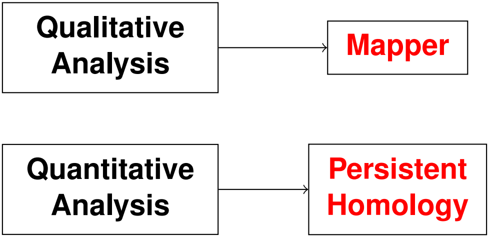
A theoretical and practical overview of homological persistence

Carles Casacuberta

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



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

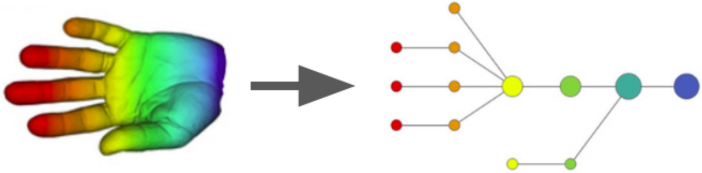
Mapper

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, e.g., single linkage.

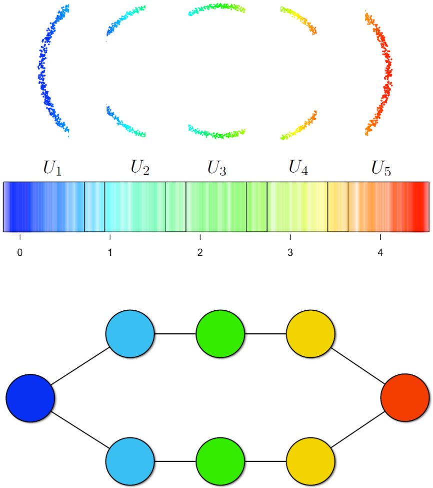
Output:

- A coloured graph.



Mapper

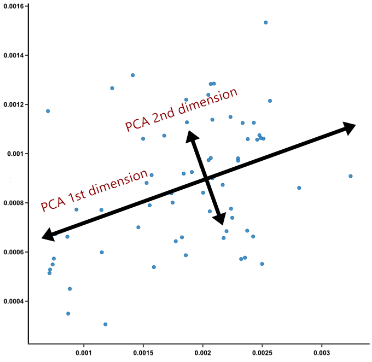

Filter Range : [0-4.2]
Interval Length : 1
Overlap : 20%


U_1
U_2
U_3
U_4
U_5
0
1
2
3
4

Mapper

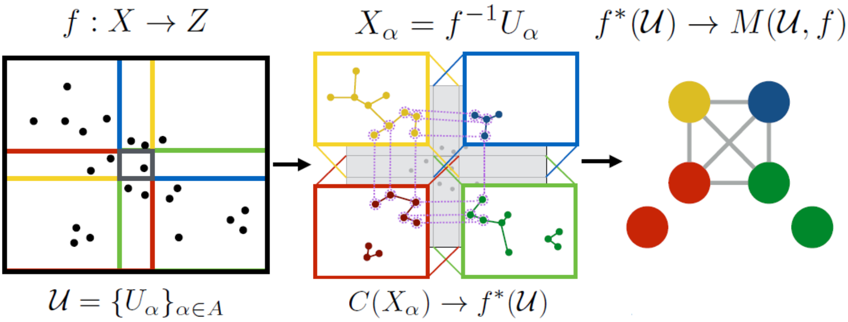
Choice of a filter function

Use dimensionality reduction methods such as **principal component analysis** (PCA).



Mapper

An example where the parameter space *Z* is 2-dimensional:

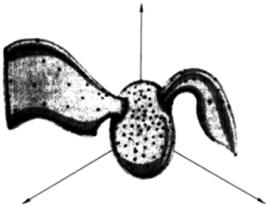


Source: M. Piekenbrock, 2018

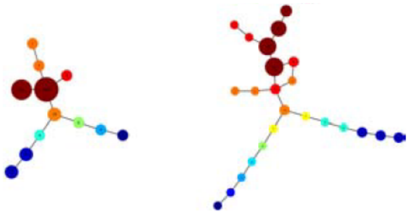
Mapper

Example: The Miller–Reaven diabetes study (1985)

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

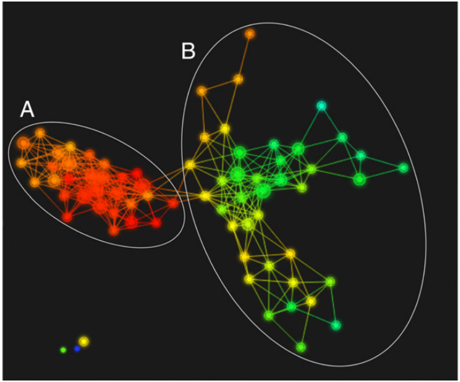


3-D image in the original study using projection and pursuit. Flares are type I and type II diabetes.



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

Mapper

Kepler Mapper

https://kepler-mapper.scikit-tda.org

Python Mapper

http://danifold.net/mapper/

TDAview (Mapper online)

https://voineagulab.github.io/TDAview/

H. J. van Veen et al. (2019)
Kepler Mapper: A flexible Python implementation of the Mapper algorithm, Journal of Open Source Software 4(42), 1315

K. Walsh, M. A. Voineagu, F. Vafaee, I. Voineagu (2020)
TDAview: an online visualization tool for topological data analysis, Bioinformatics 36, 4805–4809

Mapper

Work in progress

Comparison between endocardial and epicardial resynchronization in a model of non-ischaemic cardiomyopathy

Prevalence of peripheral arterial disease and associated cardiovascular risk factors in elderly people

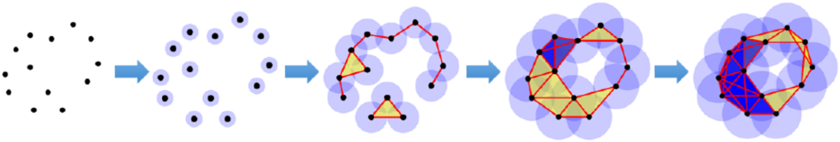
A. Ferrà, J. Guich, M. Vilasís, J. Vives, C. Casacuberta (UB)

in collaboration with

G. Amorós, J. M. Guerra, T. Puig (Hospital de Sant Pau)

Persistent Homology

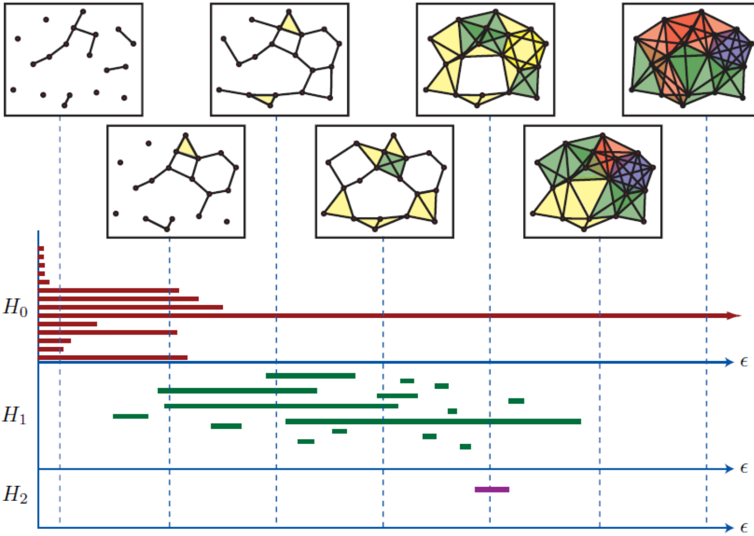
Data Set → Filtered Simplicial Complex → **Persistent Homology** → **Barcode**



Homology groups of a simplicial complex X:

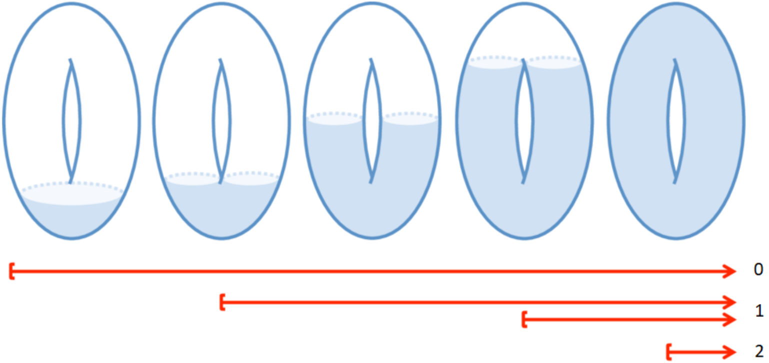
- $H_0(X)$ counts connected components of X;
- $H_1(X)$ counts 1-dimensional cycles in X;
- $H_2(X)$ counts 2-dimensional cavities in X; etc.

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,

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

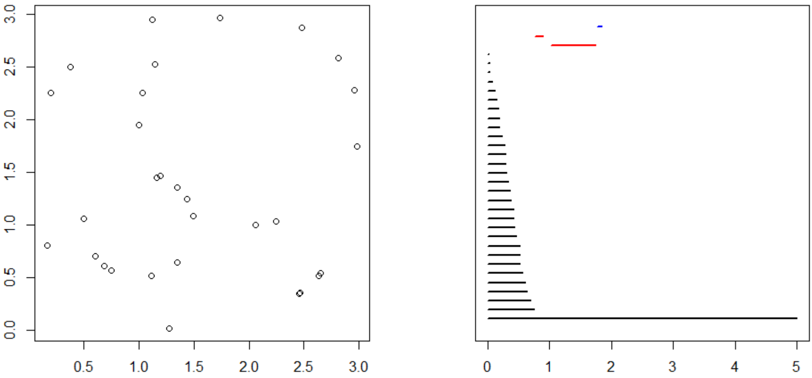
where

- $B(X)$ 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:

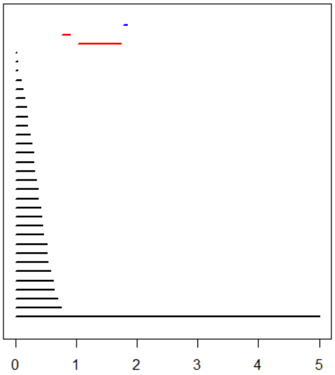
$$\boldsymbol{W_\infty(B(f), B(g)) \leq \|f - g\|_\infty}.$$

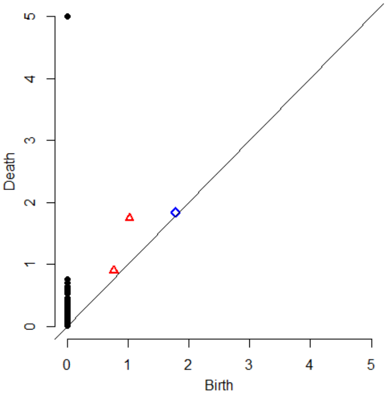
Barcodes



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

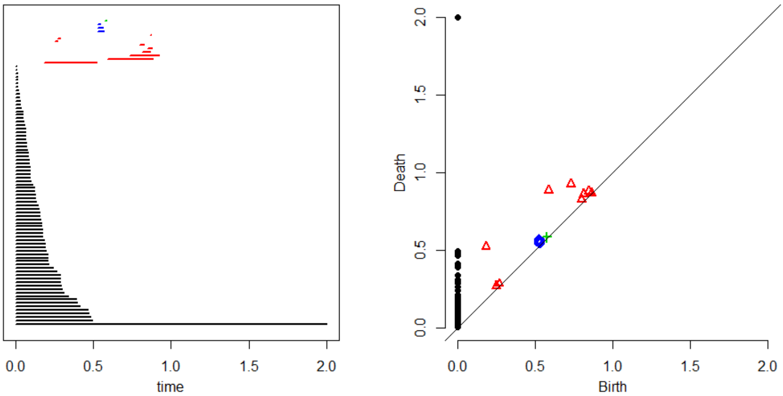
Persistence Diagrams





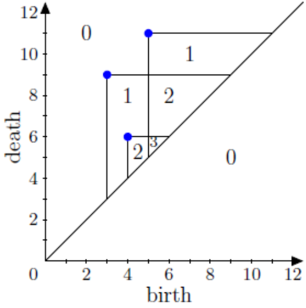
The coordinates (b, d) of each point in a **persistence diagram** correspond to *birth* and *death* of a homology generator.

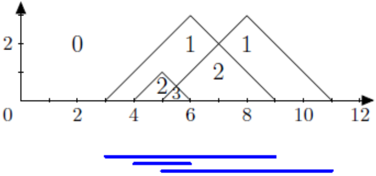
Persistence Diagrams

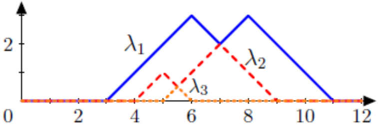


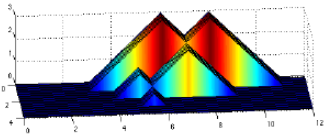
Points near the diagonal are generally viewed as *noise*.

Landscapes

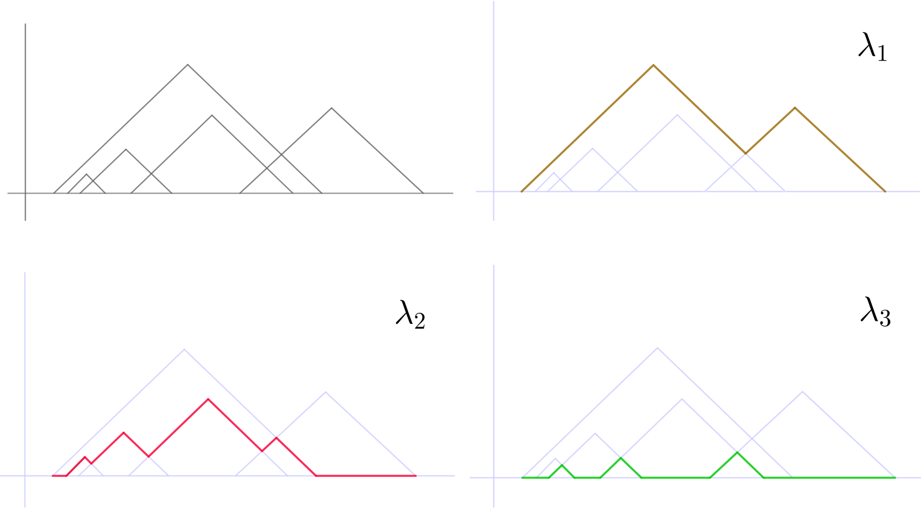

death
birth
0 2 4 6 8 10 12
0 1 2 3


0 1 2 3
0 2 4 6 8 10 12

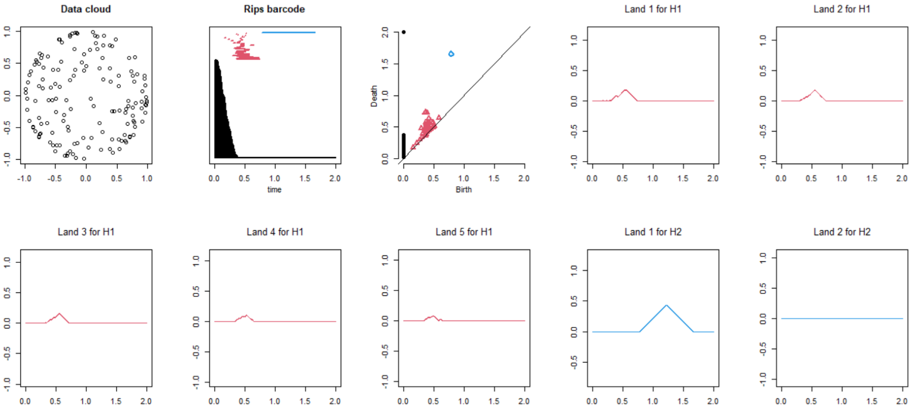

λ_1
λ_2
λ_3
0 2 4 6 8 10 12



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

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

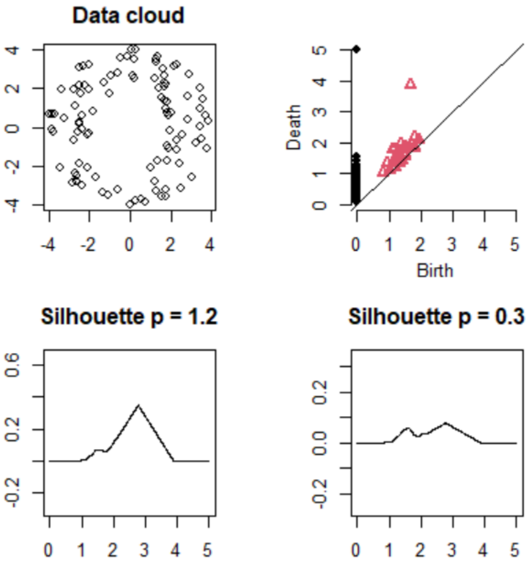
where $\{w_i\}$ are weights to be chosen, and

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

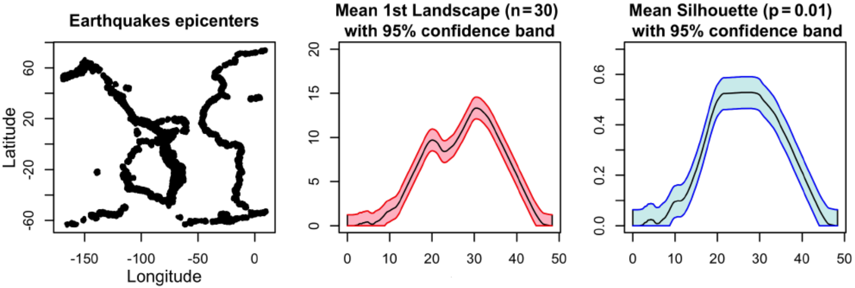
A frequent choice is $\boldsymbol{w_i = (d_i - b_i)^p}$ where p is optional:

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

Silhouettes



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

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/

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

Persistence Descriptors

R. Ballester et al. (2021)

Towards explaining the generalization gap in neural networks using topological data analysis, preprint

F. P. Nobbe (2021)

The brain network of motivation: A topological approach, Master's Thesis, UB

B. T. Fasy, Y. Qin, B. Summa, C. Wenk (2020)

Comparing distance metrics on vectorized persistence summaries, Topological Data Analysis and Beyond, 34th Conference on Neural Information Processing Systems (NeurIPS 2020)

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).

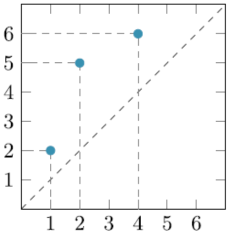
Vectorized Summaries

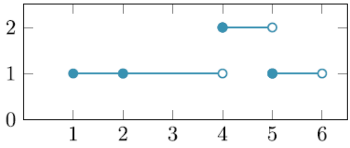
Betti curves

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

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

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





Vectorized Summaries

Persistence images

For a given persistence diagram, consider a function

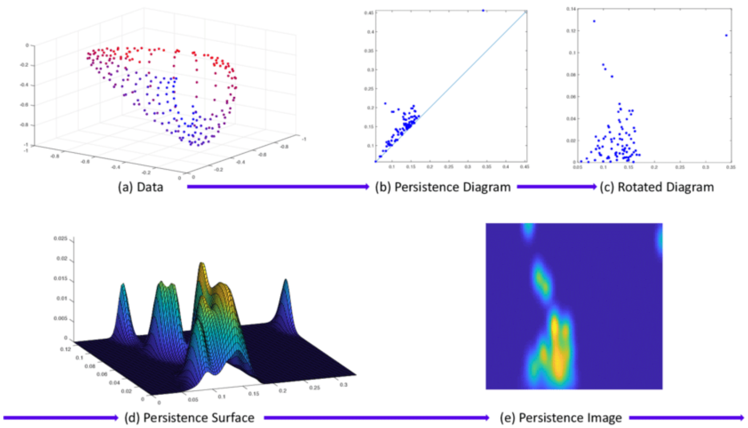
$$\Phi(s,t) = \sum_{i=1}^{n} w_i \, G_i(s,t)$$

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.

Vectorized Summaries

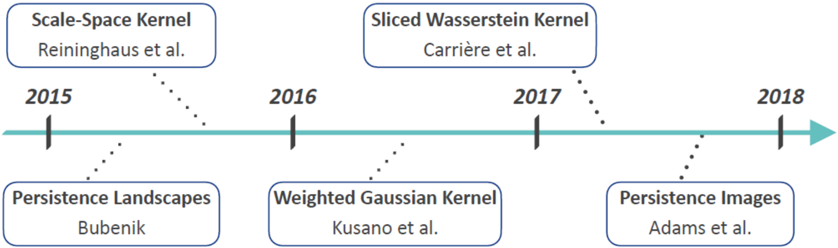


Generate a surface by centering 2D Gaussian distributions at each point, and generate a **persistence image** by summing the volume under the Gaussian distributions over the area of each pixel.

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 (CVPR), 4741–4748



Source: U. Fugacci, CNR-IMATI, Genova

Kernels

A function $K\colon X \times X \to \mathbb{R}$ on a set X is a **kernel** if there exist a Hilbert space H and a map $\Phi\colon X \to H$ such that

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

for all x, y. The Hilbert space H is called **feature space** and the map Φ is called **feature map.**

Every kernel $K\colon X \times X \to \mathbb{R}$ induces a pseudometric on X corresponding to the norm distance on the feature space:

$$d_K(x, y) = \|\Phi(x) - \Phi(y)\|.$$

Example:

- Gaussian kernel: $K(x, y) = e^{-\|x-y\|^2/2\sigma^2}$.

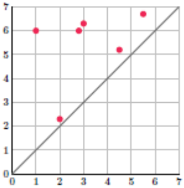
Kernels

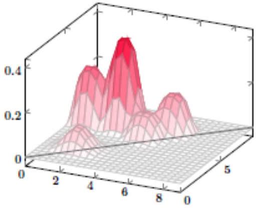
Scale-space kernel

$K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ where $\mathcal{D}$ is the set of all persistence diagrams,

$$K_\sigma(D_1, D_2) = \frac{1}{8\pi\sigma} \sum_{p \in D_1,\, q \in D_2} e^{-\|p-q\|^2/8\sigma} - e^{-\|p-\bar{q}\|^2/8\sigma}.$$

To each persistence diagram $D \in \mathcal{D}$ one assigns a sum of Dirac deltas on the points (b_i, d_i) as initial condition for a heat diffusion problem with a boundary condition on the diagonal:





Kernels

Classification performance

The following percentages were obtained over a range of 10 time parameters t_i using the landscape kernel K^L and the scale-space kernel K_σ with an SVM classifier on SHREC 2014:

HKS t_i	k^L	k_σ	Δ
t_1	68.0 ± 3.2	94.7 ± 5.1	+26.7
t_2	$\mathbf{88.3} \pm 3.3$	$\mathbf{99.3} \pm 0.9$	+11.0
t_3	61.7 ± 3.1	96.3 ± 2.2	+34.7
t_4	81.0 ± 6.5	97.3 ± 1.9	+16.3
t_5	84.7 ± 1.8	96.3 ± 2.5	+11.7
t_6	70.0 ± 7.0	93.7 ± 3.2	+23.7
t_7	73.0 ± 9.5	88.0 ± 4.5	+15.0
t_8	81.0 ± 3.8	88.3 ± 6.0	+7.3
t_9	67.3 ± 7.4	88.0 ± 5.8	+20.7
t_{10}	55.3 ± 3.6	91.0 ± 4.0	+35.7

Source: Reininghaus et al. (2015)

TDA Classifiers

Is it feasible to develop a classifier based solely on persistence descriptors?

Wait a few minutes