Persistent Homology in Data Science

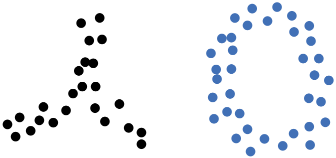
Stefan Huber <stefan.huber@fh-salzburg.ac.at>

Salzburg University of Applied Sciences, Austria

iDSC 2020[1] — 127.0.0.1
May 13, 2020

[1] Not at Dornbirn, Austria due to COVID-19. Partially supported by Digitiales Transferzentrum, Salzburg.

Topological Data Analysis: Often data displays some shape that carries valuable information.

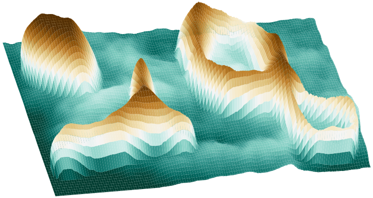


- Persistent homology gives us the notion of components, holes, tunnels, cavities, and so on and quantifies their "significance".

$$\text{Fourier analysis} : \text{signal} \;\widehat{=}\; \text{persistent homology} : \text{shape}$$

An intuitive approach: Mountains and volcanoes

Let $f\colon [0,1]^2 \to [0,1]$ be in $\mathcal{C}^0$, say, a height profile of a geographic map.



What mathematical notion is natural to capture "mountains" or "volcanoes"?

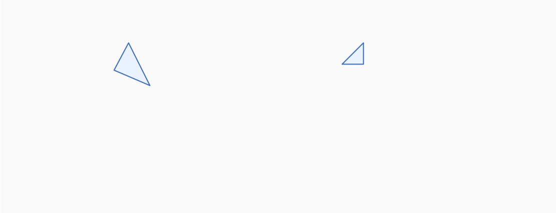
- Mountains are local maxima in f. Data has noise. How to filter to get "real mountains"?
- What about significance, which is not height? What about volcanoes?

In our simple setting, the method of persistent homology is known as watershed transformation:

- The super-level set U_c is the landmass above sea level c:

$$U_c = f^{-1}([c, 1]) = \{x \in [0, 1]^2 \colon f(x) \geq c\}$$

- U_c grows as c declines, starting at $c = 1$.



Persistent homology keeps track of the topological evolution of U_c.

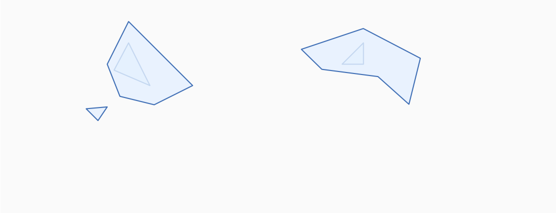
Topological evolution

In our simple setting, the method of persistent homology is known as watershed transformation:

- The super-level set U_c is the landmass above sea level c:

$$U_c = f^{-1}([c, 1]) = \{x \in [0, 1]^2 \colon f(x) \geq c\}$$

- U_c grows as c declines, starting at $c = 1$.



Persistent homology keeps track of the topological evolution of U_c.

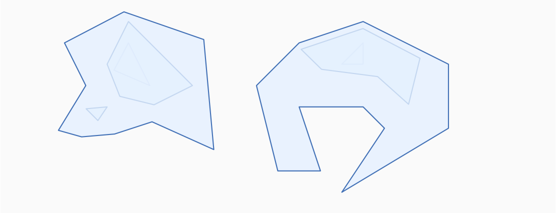
Topological evolution

In our simple setting, the method of persistent homology is known as watershed transformation:

- The super-level set U_c is the landmass above sea level c:

$$U_c = f^{-1}([c, 1]) = \{x \in [0, 1]^2 : f(x) \geq c\}$$

- U_c grows as c declines, starting at $c = 1$.



Persistent homology keeps track of the topological evolution of U_c.

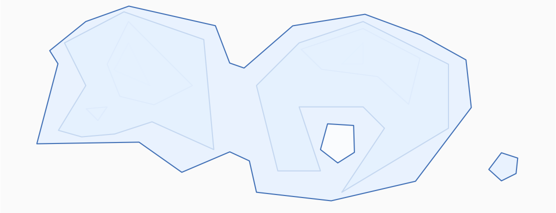
Topological evolution

In our simple setting, the method of persistent homology is known as watershed transformation:

- The super-level set U_c is the landmass above sea level c:

$$U_c = f^{-1}([c,1]) = \{x \in [0,1]^2 \colon f(x) \geq c\}$$

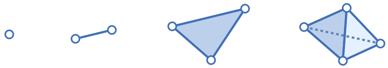
- U_c grows as c declines, starting at $c = 1$.

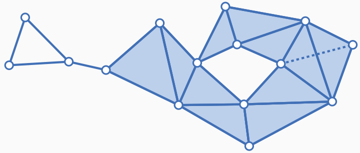


Persistent homology keeps track of the topological evolution of U_c.

General setting

An n-simplex is the convex hull of n points:





We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

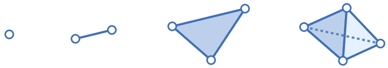
Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

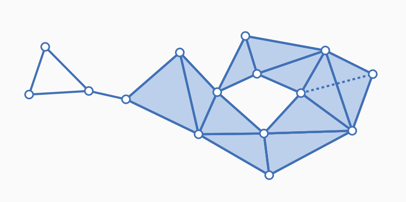
- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, ...) is its significance.

[1] Independent classes in the persistent homology group.

General setting

An n-simplex is the convex hull of n points:





We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

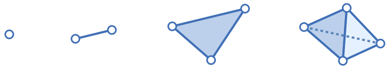
Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

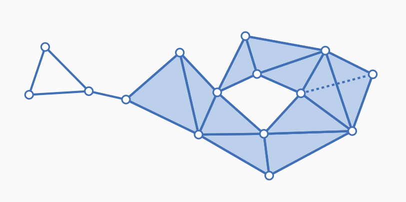
- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, . . .) is its significance.

[1] Independent classes in the persistent homology group.

General setting

An n-simplex is the convex hull of n points:





We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

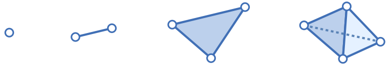
Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, . . .) is its significance.

[1] Independent classes in the persistent homology group.

General setting

An n-simplex is the convex hull of n points:



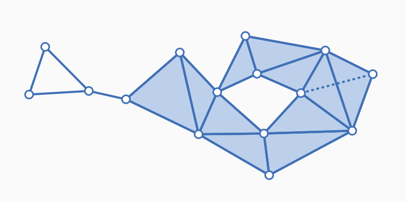
We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

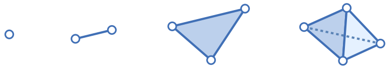
- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, . . .) is its significance.



[1] Independent classes in the persistent homology group.

General setting

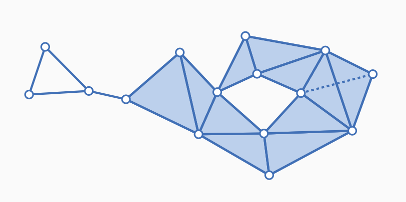
An n-simplex is the convex hull of n points:



We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$



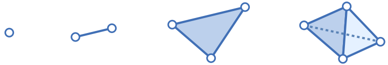
Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, . . .) is its significance.

[1] Independent classes in the persistent homology group.

General setting

An n-simplex is the convex hull of n points:



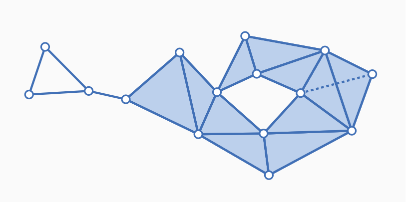
We have a simplicial complex $\mathcal{S}$ as underlying space.

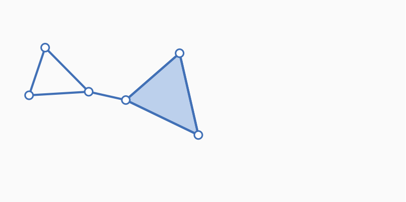
- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, ...) is its significance.

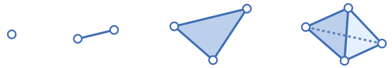




[1] Independent classes in the persistent homology group.

General setting

An n-simplex is the convex hull of n points:



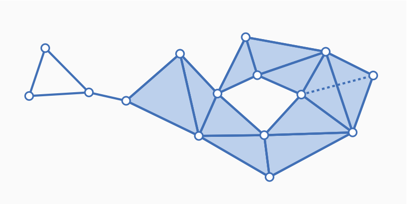
We have a simplicial complex $\mathcal{S}$ as underlying space.

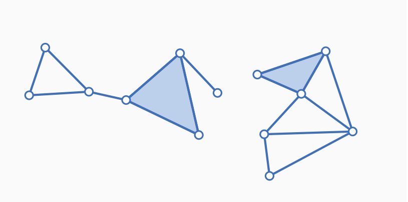
- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, ...) is its significance.

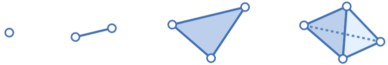




[1] Independent classes in the persistent homology group.

General setting

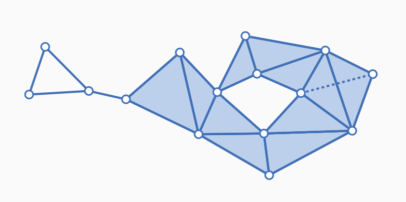
An n-simplex is the convex hull of n points:

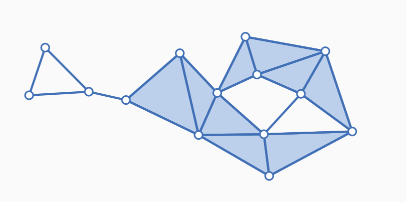


We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$





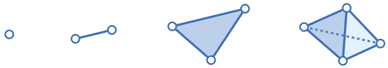
Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

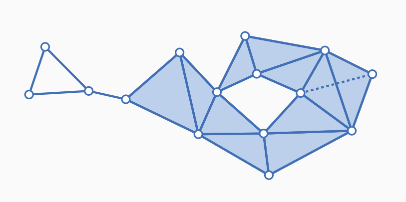
- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, ...) is its significance.

[1] Independent classes in the persistent homology group.

General setting

An n-simplex is the convex hull of n points:

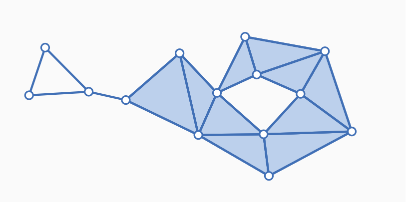




We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$

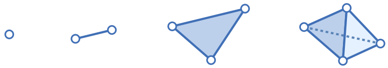


Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, ...) is its significance.

1 Independent classes in the persistent homology group.

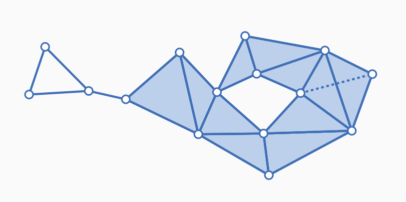
An n-simplex is the convex hull of n points:

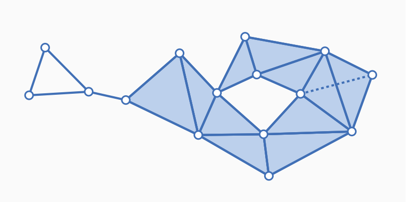


We have a simplicial complex $\mathcal{S}$ as underlying space.

- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$



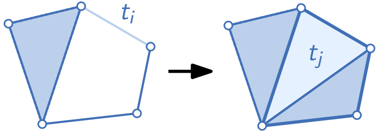


Think of $(\mathcal{S}_i)$ as iteratively adding adding simplices.

- At each step a feature is born or dies.
- The lifespan of a feature (component, hole, . . .) is its significance.

[1] Independent classes in the persistent homology group.

Persistence diagram

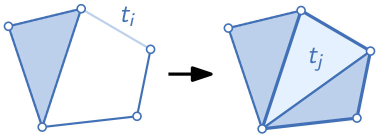


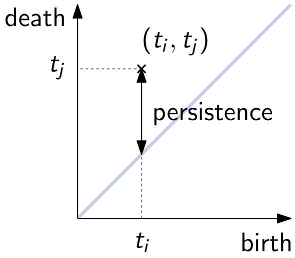
We associate at timestamp $t_i \in \mathbb{R}$ to the i-th step in the filtration $(\mathcal{S}_i)$ with

$$t_0 \leq t_1 \leq \cdots \leq t_m$$

- The persistent Betti number $\mu_p^{i,j}$ counts how many p-dimensional features were born at time t_i and died at time t_j.

Persistence diagram





We associate at timestamp $t_i \in \mathbb{R}$ to the i-th step in the filtration $(\mathcal{S}_i)$ with

$$t_0 \leq t_1 \leq \cdots \leq t_m$$

- The persistent Betti number $\mu_p^{i,j}$ counts how many p-dimensional features were born at time t_i and died at time t_j.

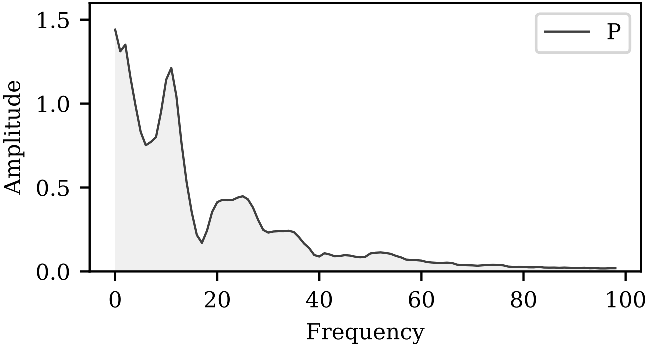
The p-th persistence diagram is a summary description:

- We place a point (t_i, t_j) with multiplicity $\mu_p^{i,j}$.
- Persistence is $t_j - t_i$.

Application: Peak detection for signal analysis

The function P stems from a system identification for a closed-loop controller in motion control.

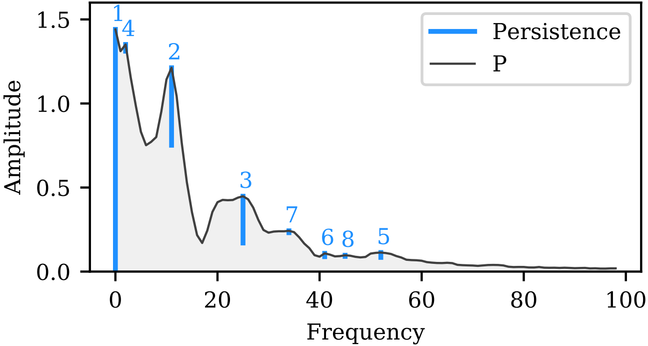
- Task: Detect peak at non-zero frequency, which is the natural frequency of the system.

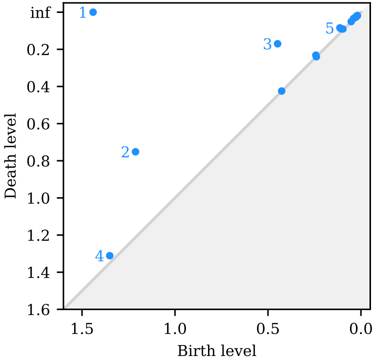


Application: Peak detection for signal analysis

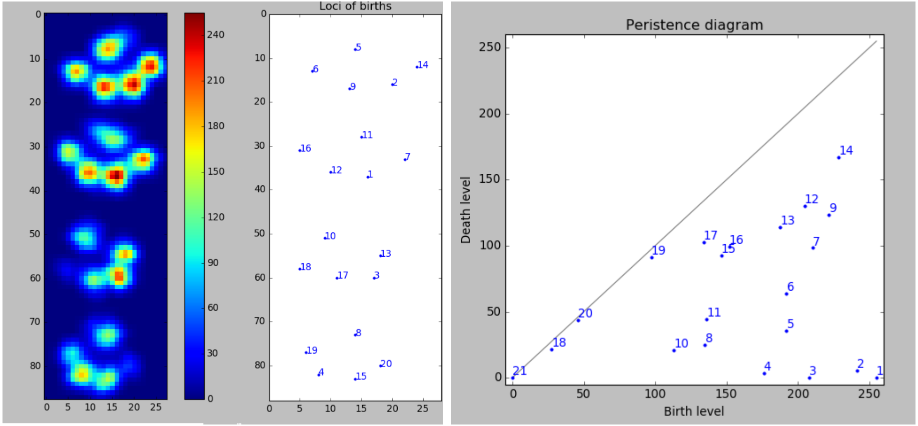
The function P stems from a system identification for a closed-loop controller in motion control.

- Task: Detect peak at non-zero frequency, which is the natural frequency of the system.
- 0-th persistence diagram of super-levelset filtration of P.
- Can be computed in a few dozen lines of code in C, as fast as sorting numbers.



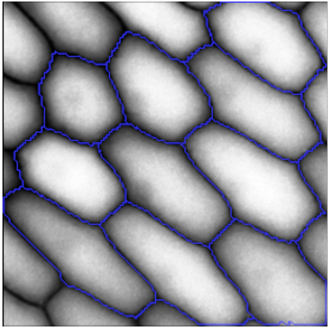


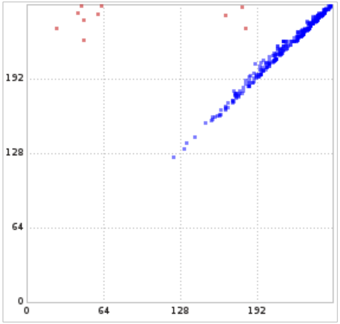
The 20 most persistent 0-dimensional features to detect animal paws.



Segmentation of cell boundaries.

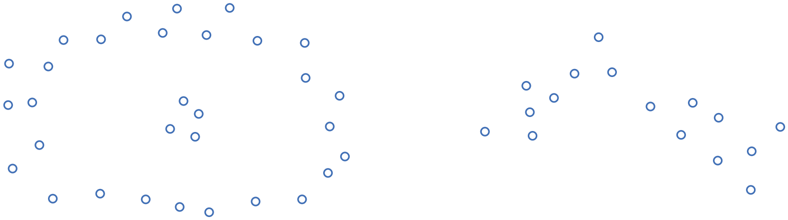
- Chosen 1-dimensional features (cycles) by thresholding in 1st persistence diagram.
- Like finding volcanoes in geographic height maps.





Application: Shape analysis of points

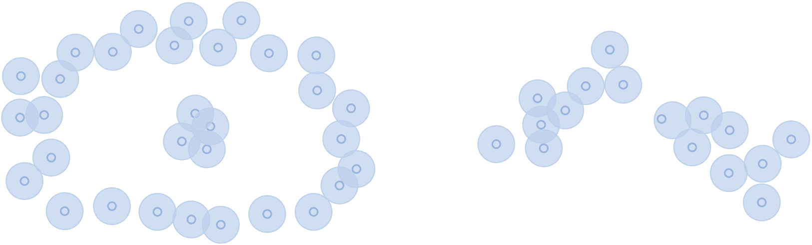
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- The 0-th persistence diagram encodes the evolution and significance of clusters.
- Higher dimensional persistence diagram gives us additional information about holes.

Application: Shape analysis of points

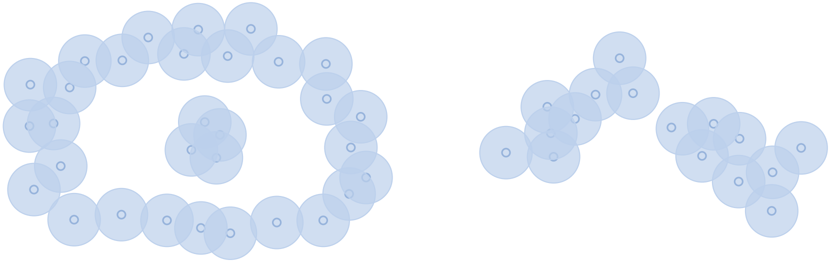
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- The 0-th persistence diagram encodes the evolution and significance of clusters.
- Higher dimensional persistence diagram gives us additional information about holes.

Application: Shape analysis of points

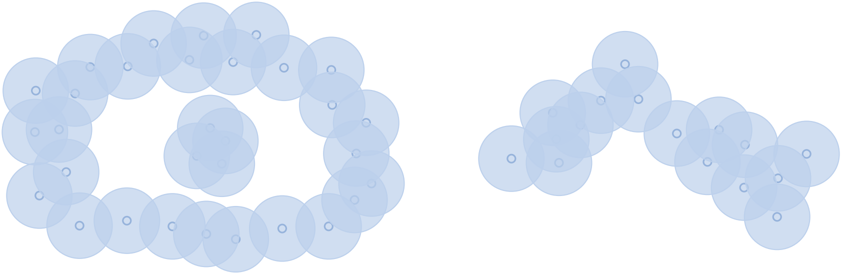
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- The 0-th persistence diagram encodes the evolution and significance of clusters.
- Higher dimensional persistence diagram gives us additional information about holes.

Application: Shape analysis of points

- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.

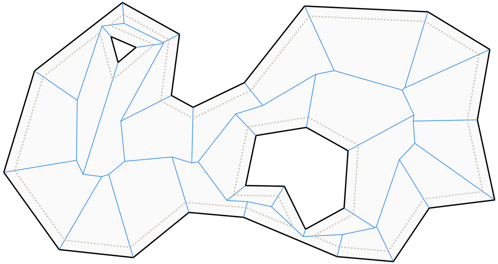


- The 0-th persistence diagram encodes the evolution and significance of clusters.
- Higher dimensional persistence diagram gives us additional information about holes.

Application: Shape analysis of polygons

Geometric shapes are often modeled as polygons, possibly with holes.

- A filtration is obtained by a (reversed) offset process, e.g., Minkowski offsets or mitered offsets.



- [Hub18] gave efficient algorithms to compute persistent homology based on Voronoi diagrams and straight skeletons by proving homotopy equivalence.
- Applications: Polygon decomposition, e.g., for high-speed NC-machining.

Application: Topological machine learning

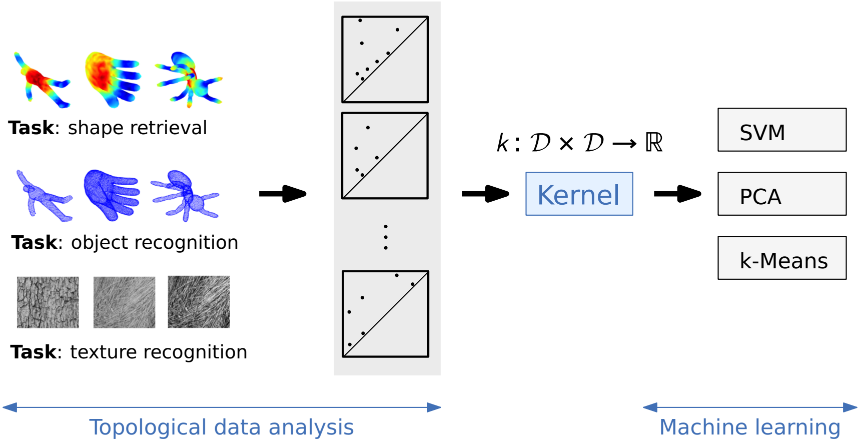
Persistence diagrams are a summary description of topological features.

- How to use this topological information for machine learning?

Application: Topological machine learning

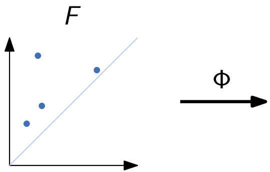
Persistence diagrams are a summary description of topological features.

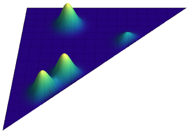
- How to use this topological information for machine learning?



Idea of [Rei+15]: Given F, solve a heat-diffusion PDE on $\Omega = \{\{x, y\} \in \mathbb{R}^2 : y \geq x\}$

- Solution at time t denoted by $u_t : \Omega \to \mathbb{R}$.
- Initial condition $u_0 = \sum_{p \in F} \delta_p$ with Dirac delta δ_p.
- Boundary condition $u_t = 0$ on $\partial\Omega$, as points on diagonal shall have no influence.





We directly constructed a feature map $\Phi : \mathcal{D} \to L_2(\Omega)$ on the set $\mathcal{D}$ of persistence diagrams.

- The kernel is given by $k(F, G) = \langle \Phi(F), \Phi(G) \rangle$.
- Important: The resulting kernel is stable, i.e., Lipschitz-continuous.

Conclusion

Persistent homology turns out to be useful:

- Clustering, image analysis, shape recognition, image segmentation, time series analysis, analysis of biological structures (drug molecules, roots, . . .), material analysis, . . .

It contributes to data science in two ways:

1. Persistent diagrams make various methods of data science applicable.
2. It is a tool within data science to help understanding methods.
   - E.g., explainable AI based on persistence of the inter-layer mapping in feed forward nets. [CG18].

Interreg Österreich-Bayern project

KI-Net – Bausteine für KI-basierte Optimierungen in der industriellen Fertigung:

- Lead: SCCH Hagenberg (OÖ)
- FH Salzburg
- TH Rosenheim
- Universität Innsbruck
- Hochschule Kempten

Interreg
Österreich–Bayern 2014–2020
Europäische Union – Europäischer Fonds für Regionale Entwicklung

Bibliography I

[CG18] Gunnar E. Carlsson and Rickard Brüel Gabrielsson. "Topological Approaches to Deep Learning." In: *CoRR* abs/1811.01122 (2018). arXiv: 1811.01122. URL: http://arxiv.org/abs/1811.01122.

[Hub18] Stefan Huber. "The Topology of Skeletons and Offsets." In: *Proc. 34th Europ. Workshop on Comp. Geom. (EuroCG '18)*. Mar. 2018.

[Rei+15] J. Reininghaus et al. "A Stable Multi-Scale Kernel for Topological Machine Learning." In: *Proc. 2015 IEEE Conf. Comp. Vision & Pat. Rec. (CVPR '15)*. Boston, MA, USA, June 2015, pp. 4741–4748.