Persistent Homology in Data Science

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

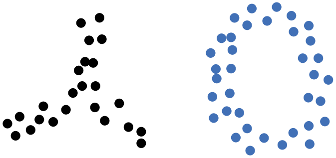
Salzburg University of Applied Sciences, Austria

Data Science Lunch Talk
Nov 9, 2021

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

Data has shape

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

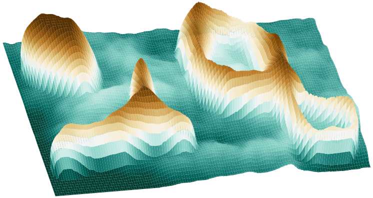


- Persistent homology gives us the notion of components, holes, tunnels, cavities, and so on and quantifies their “significance”.

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

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?

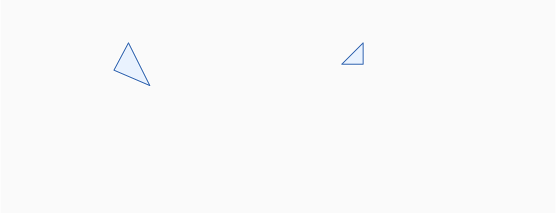
Topological evolution

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.

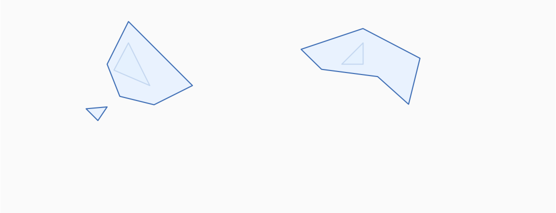
Topological evolution

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:

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

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



Persistent homology keeps track of the topological evolution of U_c.

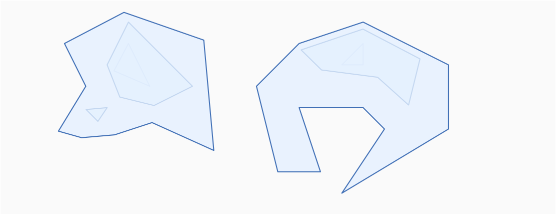
Topological evolution

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 $c = 1$.



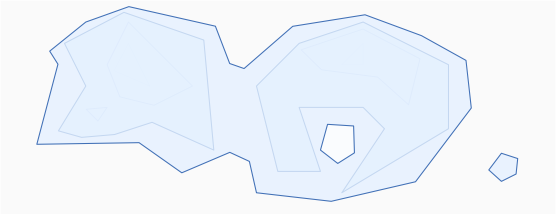
Persistent homology keeps track of the topological evolution of U_c.

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 $c = 1$.



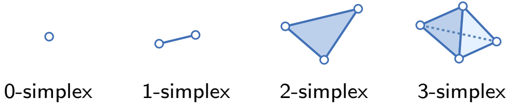
Persistent homology keeps track of the topological evolution of U_c.

Forming spaces

Notion of a space

We need notion of a "space" that is tangible for an algorithmic treatment.
Answer: Simplicial complex

An n-simplex σ in $\mathbb{R}^d$ is the convex hull of $n+1$ points.[1] And n is called dimension.



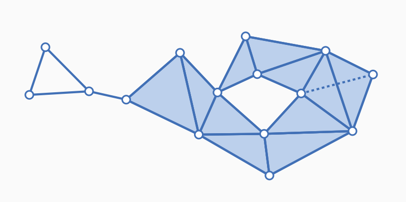
Note

A face of a simplex is a simplex, too.

[1] We also assume the points are affinely independent, i.e., the $n+1$ points do not lie in an affine-linear $(n-1)$-dimensional subspace.

As simplicial complex $\mathcal{S}$ is a “nice” set of simplices:

- If $\sigma \in \mathcal{S}$ then all faces of σ belong to $\mathcal{S}$, too.
- If $\sigma_1, \sigma_2 \in \mathcal{S}$ then $\sigma_1 \cap \sigma_2$ belongs to $\mathcal{S}$, too.[2]



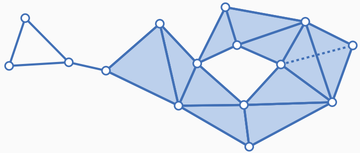
The dimension of $\mathcal{S}$ is the largest dimension of its simplices.

2 Unless the intersection would be empty.

Evolution of spaces

Notion of evolution

Answer: Filtration



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 simplices.

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

Evolution of spaces

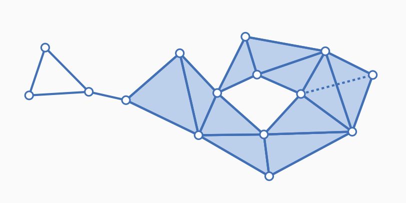
Notion of evolution

Answer: Filtration

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

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

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

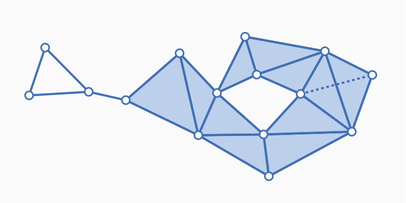


Think of $(\mathcal{S}_i)$ as iteratively adding simplices.

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

Notion of evolution

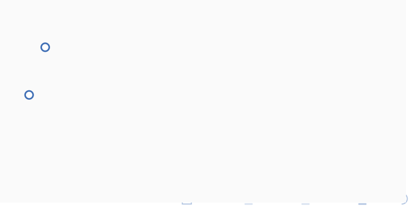
Answer: Filtration



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

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

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

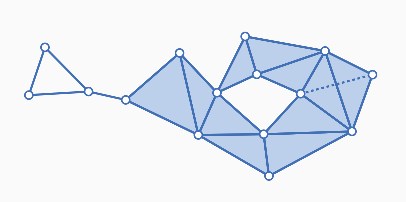


Think of $(\mathcal{S}_i)$ as iteratively adding simplices.

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

Notion of evolution

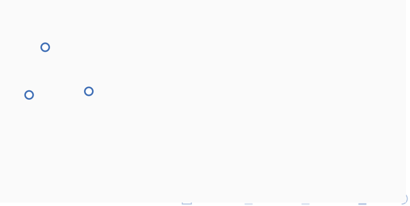
Answer: Filtration



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 simplices.

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

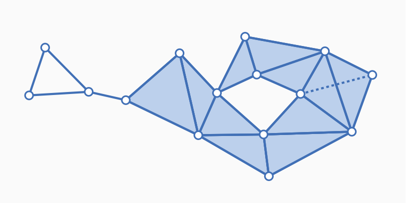
Notion of evolution

Answer: Filtration

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

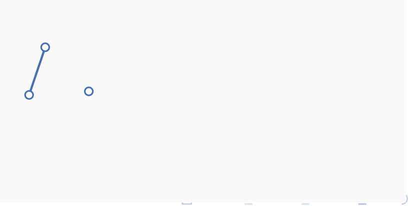
- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

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



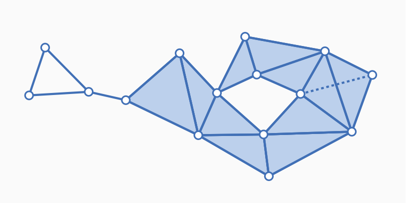
Think of $(\mathcal{S}_i)$ as iteratively adding simplices.

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



Notion of evolution

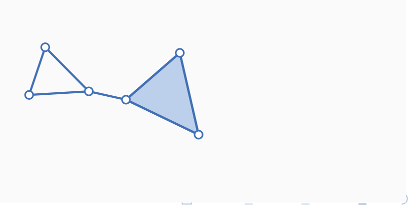
Answer: Filtration



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

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

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

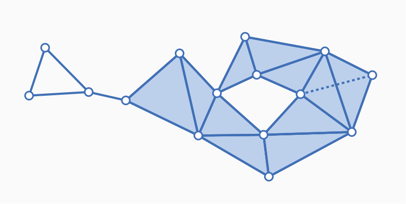


Think of $(\mathcal{S}_i)$ as iteratively adding simplices.

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

Notion of evolution

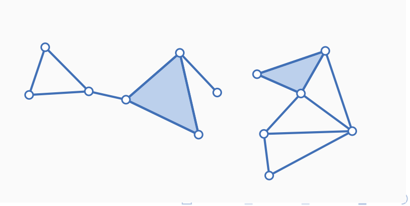
Answer: Filtration



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

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

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



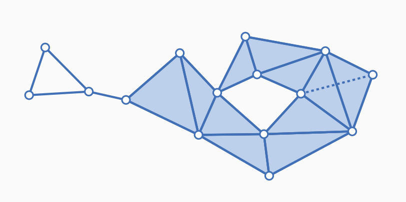
Think of $(\mathcal{S}_i)$ as iteratively adding simplices.

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

Evolution of spaces

Notion of evolution

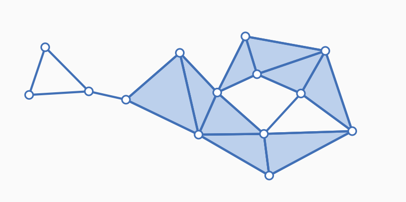
Answer: Filtration



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 simplices.

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

Evolution of spaces

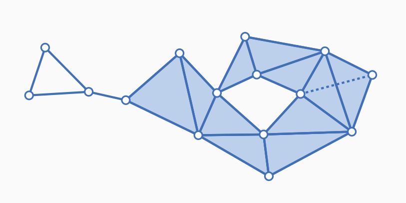
Notion of evolution

Answer: Filtration

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

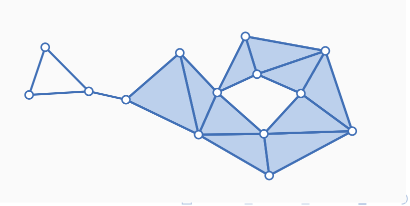
- A filtration $(\mathcal{S}_i)$ is a sequence of simplicial complexes

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



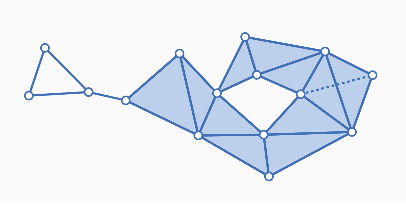
Think of $(\mathcal{S}_i)$ as iteratively adding simplices.

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



Notion of evolution

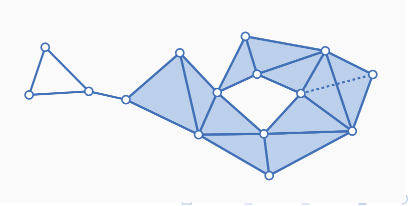
Answer: Filtration



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

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

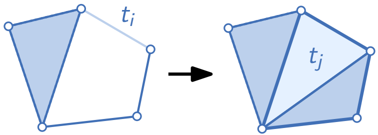
$$\emptyset = \mathcal{S}_0 \subset \cdots \subset \mathcal{S}_m = \mathcal{S}$$



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

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

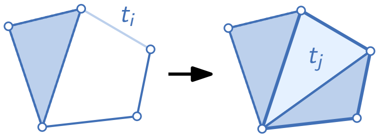
Persistence diagram

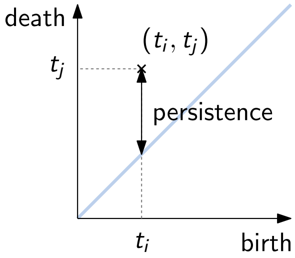


We associate a timestamp $t_i \in \mathbb{R}$ to the i-th step in the filtration $(\mathcal{S}_i)$ with

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

- 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.





We associate a 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.

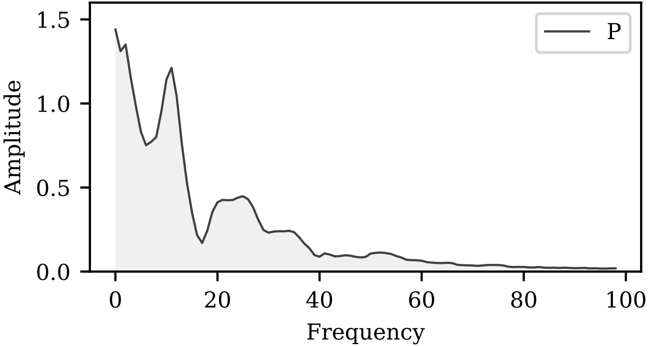
The p-th persistence diagram is a summary description:

- 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

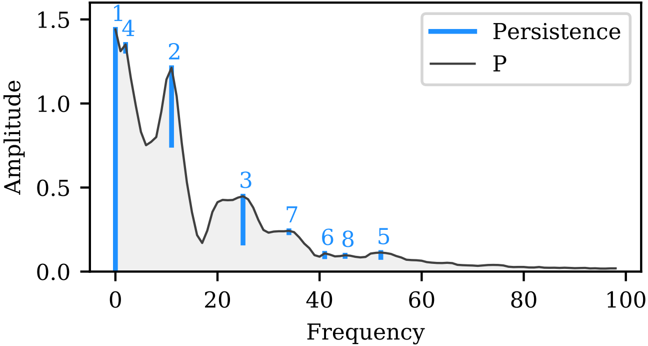
The function P stems from a system identification for a closed-loop controller in motion control.

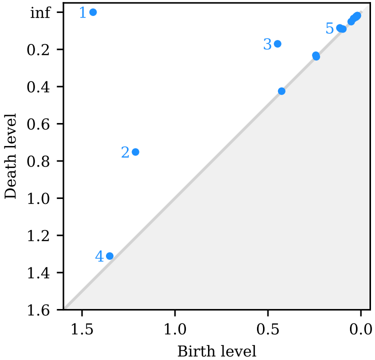
- Task: Detect peak at non-zero frequency, which is the natural frequency of the system.



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

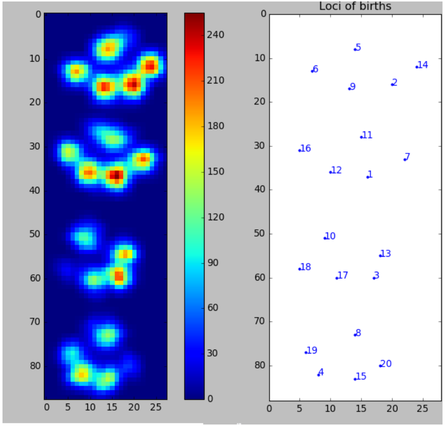
- Task: Detect peak at non-zero frequency, which is the natural frequency of the system.
- 0-th persistence diagram of super-levelset filtration of P.
- Can be computed in a few dozen lines of code in C, as fast as sorting numbers.

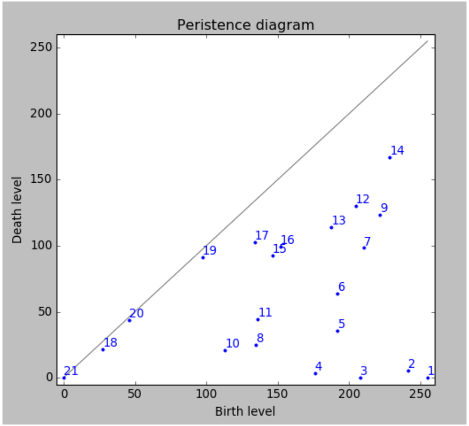




Application: Image analysis

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

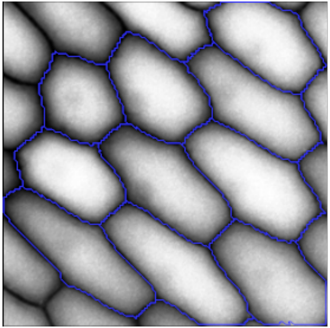


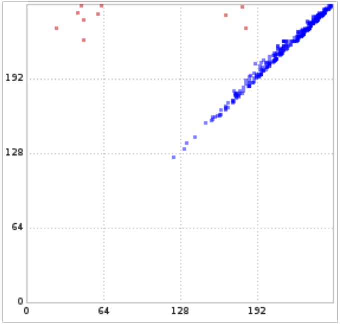


Application: Image analysis

Segmentation of cell boundaries.

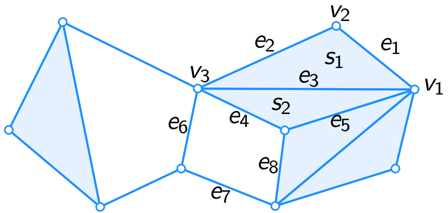
- Chosen 1-dimensional features (cycles) by thresholding in 1st persistence diagram.
- Like finding volcanoes in geographic height maps.



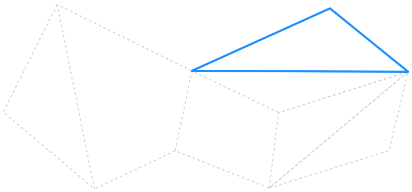


p-chain

We define a p-chain of $\mathcal{S}$ as a formal sum $\sum_i \lambda_i \sigma_i$ of p-simplices $\sigma_i \in \mathcal{S}$.

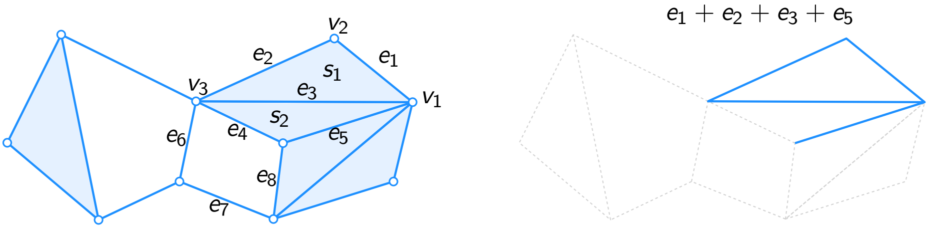


$e_1 + e_2 + e_3$



p-chain

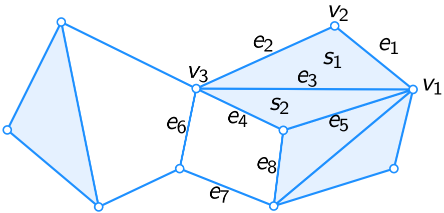
We define a p-chain of $\mathcal{S}$ as a formal sum $\sum_i \lambda_i \sigma_i$ of p-simplices $\sigma_i \in \mathcal{S}$.



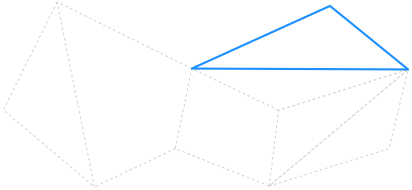
p-chain

We define a p-chain of $\mathcal{S}$ as a formal sum $\sum_i \lambda_i \sigma_i$ of p-simplices $\sigma_i \in \mathcal{S}$.

- Here, the coefficients λ_i are in $\mathbb{Z}_2$, i.e., we count modulo two.

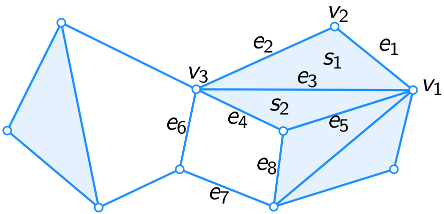


$$e_1 + e_2 + e_3 + e_5 + e_5 = e_1 + e_2 + e_3$$

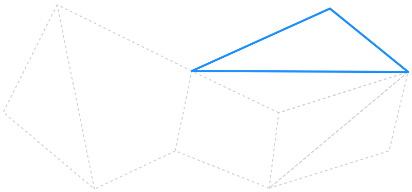


Boundaries

- The boundary $\partial\sigma$ of a p-simplex σ is formed by the $(p-1)$-dimensional faces of σ.
- The boundary ∂c of a p-chain c is the sum of boundaries of its simplices: $\partial\left(\sum_i \sigma_i\right) = \sum_i \partial\sigma_i$



$$\partial s_1 = e_1 + e_2 + e_3$$

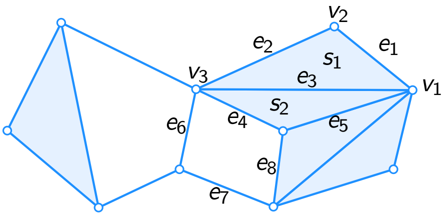


- Note that for a p-chain c the ∂c forms a $(p-1)$-chain.

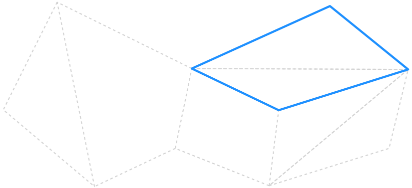
Under the mathematical hood

Boundaries

- The boundary $\partial\sigma$ of a p-simplex σ is formed by the $(p-1)$-dimensional faces of σ.
- The boundary ∂c of a p-chain c is the sum of boundaries of its simplices: $\partial\left(\sum_i \sigma_i\right) = \sum_i \partial\sigma_i$



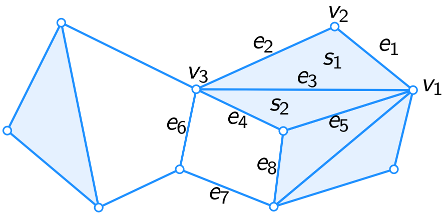
$$\partial(s_1 + s_2) = e_1 + e_2 + e_4 + e_5$$



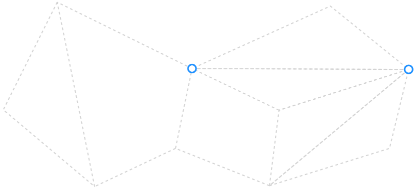
- Note that for a p-chain c the ∂c forms a $(p-1)$-chain.

Boundaries

- The boundary $\partial\sigma$ of a p-simplex σ is formed by the $(p-1)$-dimensional faces of σ.
- The boundary ∂c of a p-chain c is the sum of boundaries of its simplices: $\partial\left(\sum_i \sigma_i\right) = \sum_i \partial\sigma_i$



$$\partial(e_1 + e_2) = v_1 + v_2 + v_2 + v_3 = v_1 + v_3$$

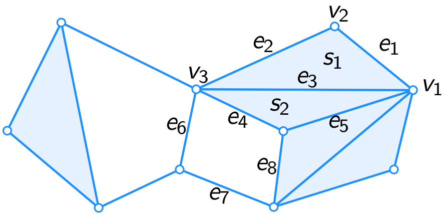


- Note that for a p-chain c the ∂c forms a $(p-1)$-chain.

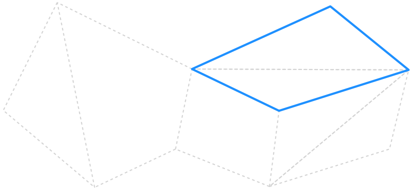
Under the mathematical hood

Cycle

A cycle c is a boundary-less p-chain, i.e., where $\partial c = 0$.



$$\partial(s_1 + s_2) = e_1 + e_2 + e_4 + e_5$$

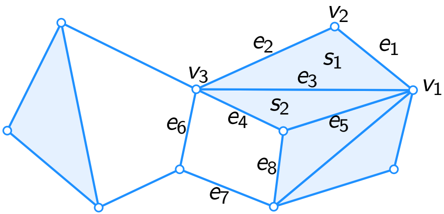


Note that $\partial\partial c = 0$ for any p-chain c, i.e., every boundary is a cycle, but not vice versa.

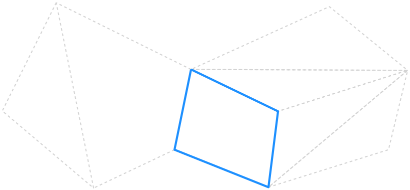
Under the mathematical hood

Say we want to capture the 1-dimensional features of $\mathcal{S}$, i.e., holes in $\mathcal{S}$.

- We search for certain 1-cycles, like $c = e_4 + e_6 + e_7 + e_8$.
- However, we regard c and $c + \partial s_2$ as the same in the sense that we can "continuously" transform the one cycle into the other.



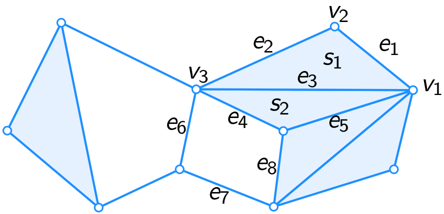
$e_4 + e_6 + e_7 + e_8$



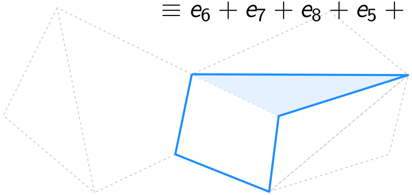
Under the mathematical hood

Say we want to capture the 1-dimensional features of $\mathcal{S}$, i.e., holes in $\mathcal{S}$.

- We search for certain 1-cycles, like $c = e_4 + e_6 + e_7 + e_8$.
- However, we regard c and $c + \partial s_2$ as the same in the sense that we can "continuously" transform the one cycle into the other.
- Ergo, we defining an equivalence relation $\equiv$ on p-chains by defining $c_1 \equiv c_2$ if there is a $(p+1)$-chain c^* such that $c_1 = c_2 + \partial c^*$.
- By features we mean the equivalence classes of $\equiv$, which are called homology classes.



$$\begin{aligned} e_4 + e_6 + e_7 + e_8 &\equiv e_4 + e_6 + e_7 + e_8 + \partial s_2 \\ &\equiv e_6 + e_7 + e_8 + e_5 + e_3 \end{aligned}$$



Under the mathematical hood

Ingredients to cook homology groups on $\mathcal{S}$:

- By C_p we denote its p-chains.
- By $\partial_p \colon C_p \to C_{p-1}$ we denote the p-th boundary operator.
- $Z_p = \ker \partial_p$ we denote are p-cycles and $B_p = \operatorname{im} \partial_{p+1}$ are p-boundaries.

Observations:

- C_p, B_p, Z_p form groups with $+$ as group action.
- We have $\partial_p \partial_{p+1} = 0$ and $B_p \subset Z_p \subset C_p$.

Homology group

We define the quotient group $H_p = Z_p / B_p = \ker \partial_p / \operatorname{im} \partial_{p+1}$ as the p-th homology group and its rank as the p-th Betti number β_p.

Under the mathematical hood

The persistent homology groups capture homology classes that survived a "time span" $[i, j]$ in a filtration $(\mathcal{S}_k)_{k=0}^m$.

- Let us denote by C_p^i the p-cycles in $\mathcal{S}_i$. Then we have this commutative diagram:

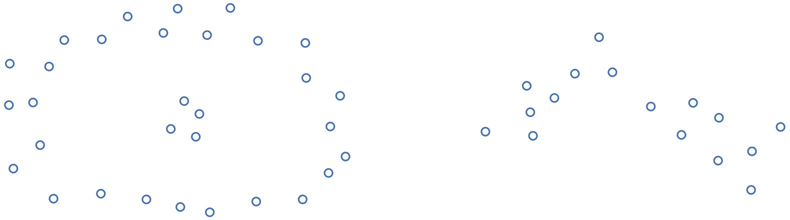
$$
\begin{tikzcd}
\cdots \arrow[r, "\partial_3^0"] & C_2^0 \arrow[r, "\partial_2^0"] \arrow[d, hook] & C_1^0 \arrow[r, "\partial_1^0"] \arrow[d, hook] & C_0^0 \arrow[r, "\partial_0^0"] \arrow[d, hook] & 0 \\
\cdots \arrow[r, "\partial_3^1"] & C_2^1 \arrow[r, "\partial_2^1"] \arrow[d, hook, dotted] & C_1^1 \arrow[r, "\partial_1^1"] \arrow[d, hook, dotted] & C_0^1 \arrow[r, "\partial_0^1"] \arrow[d, hook, dotted] & 0 \\
\cdots \arrow[r, "\partial_3^m"] & C_2^m \arrow[r, "\partial_2^m"] & C_1^m \arrow[r, "\partial_1^m"] & C_0^m \arrow[r, "\partial_0^m"] & 0
\end{tikzcd}
$$

Persistent homology group

The p-th persistent homology group $H_p^{i,j}$ is defined as $\ker \partial_p^i / (\operatorname{im} \partial_{p+1}^j \cap \ker \partial_p^i)$.

Application: Shape analysis of points

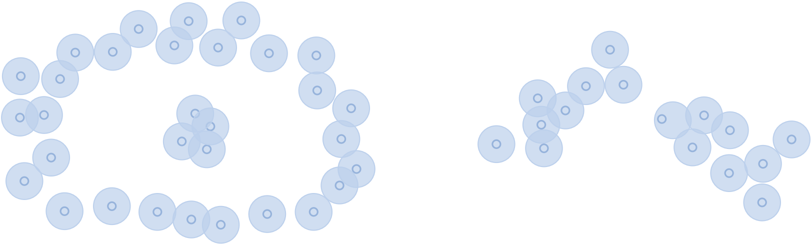
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- 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

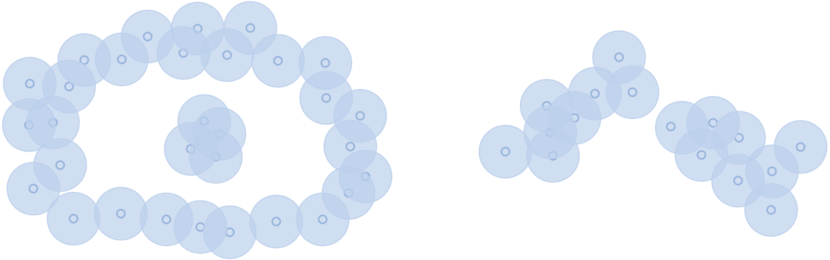
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- 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

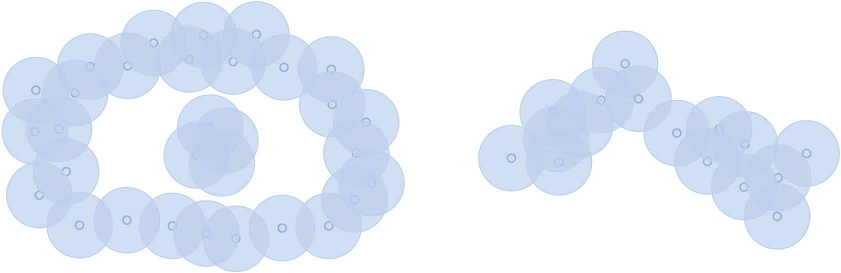
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- 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

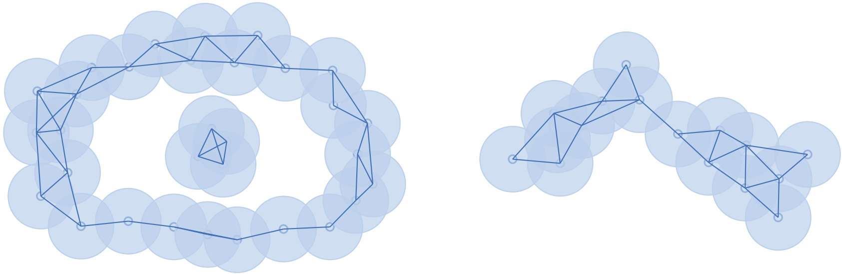
- Place a ball B_t of radius t around each point and consider the union P_t.
  - The connected components of P_t build clusters.
- The sequence (P_t) forms a filtration.



- 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 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.
- Actually, a (homotopy) equivalent simplicial complex is formed.

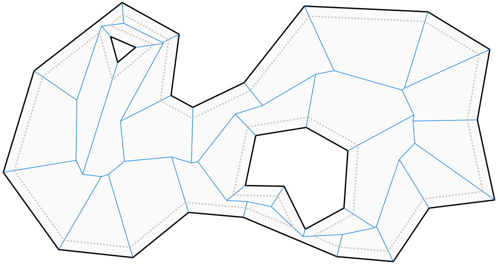


- The 0-th persistence diagram encodes the evolution and significance of clusters.
- Higher dimensional persistence diagram gives us additional information about holes.

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

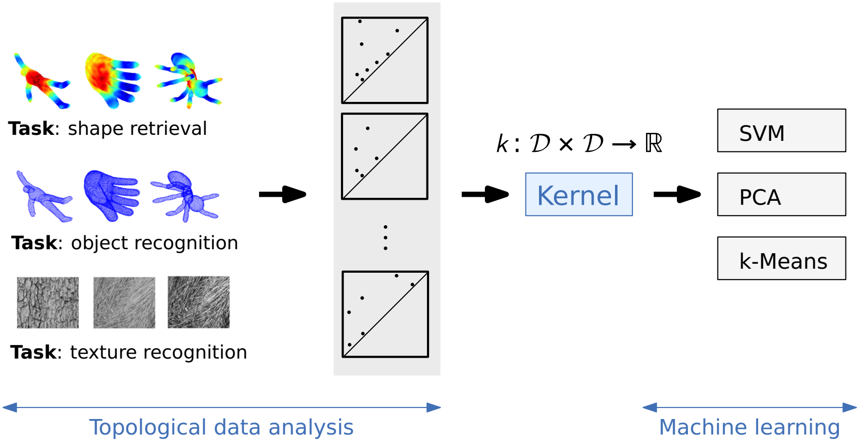
Persistence diagrams are a summary description of topological features.

- How to use this topological information for machine learning?

Application: Topological machine learning

Persistence diagrams are a summary description of topological features.

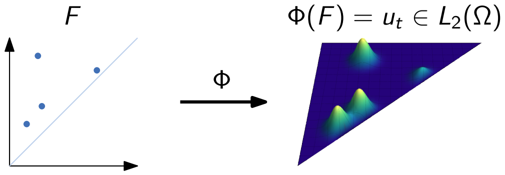
- How to use this topological information for machine learning?



Application: Topological machine learning

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

- Solution at time t denoted by $u_t \colon \Omega \to \mathbb{R}$.
- 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].

Where to go next?

- The textbook on Computational Topology:
  H. Edelsbrunner and J. Harer. *Computational Topology – An Introduction*. ISBN 978-0-8218-4925-5. American Mathematical Society, 2010
- A brief introduction into Peristent Homology in Data Science:
  Stefan Huber. "Persistent Homology in Data Science." In: *Proc. 3rd Int. Data Sci. Conf. (iDSC '20)*. Data Science – Analytics and Applications. Dornbirn, Austria (virtual), May 2020. DOI: 10.1007/978-3-658-32182-6_13
- A textbook on Topological Data Analysis and machine learning:
  `https://www.cs.purdue.edu/homes/tamaldey/book/CTDAbook/CTDAbook.html`
- Various software packages for R, Python, C and more.
  For instance one that shall be simple to understand:
  `https://www.sthu.org/code/libstick/`
- Wikipedia:
  `https://en.wikipedia.org/wiki/Persistent_homology`
  `https://en.wikipedia.org/wiki/Topological_data_analysis`

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

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.

[EH10] H. Edelsbrunner and J. Harer. *Computational Topology – An Introduction*. ISBN 978-0-8218-4925-5. American Mathematical Society, 2010.

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

[Hub20] Stefan Huber. "Persistent Homology in Data Science." In: *Proc. 3rd Int. Data Sci. Conf. (iDSC '20)*. Data Science – Analytics and Applications. Dornbirn, Austria (virtual), May 2020. DOI: 10.1007/978-3-658-32182-6_13.

[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.