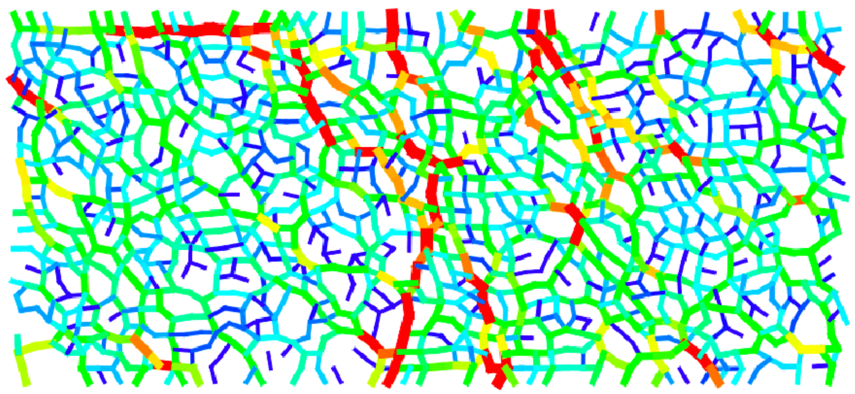
Journées Réduction de modèles et traitement géométrique des données"
November 2021

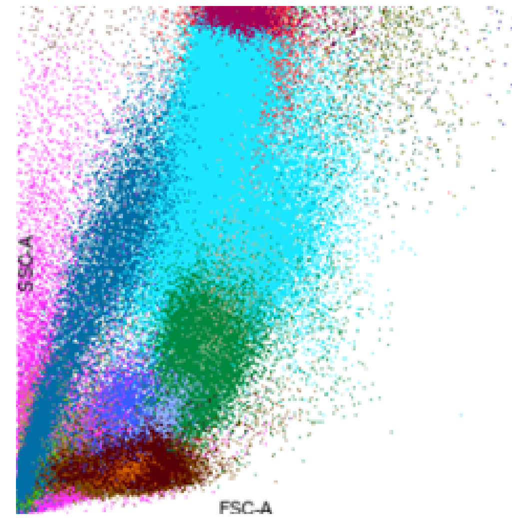
A brief introduction to Persistent Homology

Frédéric Chazal
DataShape team
INRIA & Laboratoire de Mathématiques d'Orsay
frederic.chazal@inria.fr

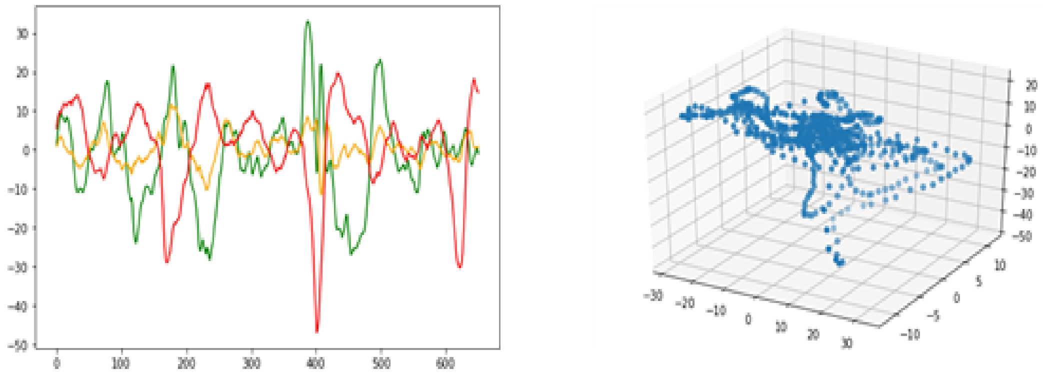
What is Topological Data Analysis (TDA)?



[Force fields in granular media (Krama et al)]



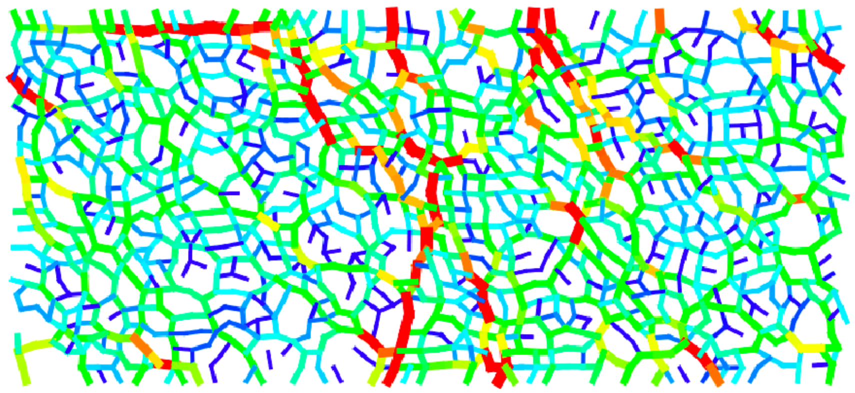
[Cell population - cytometry - MetaFora courtesy]



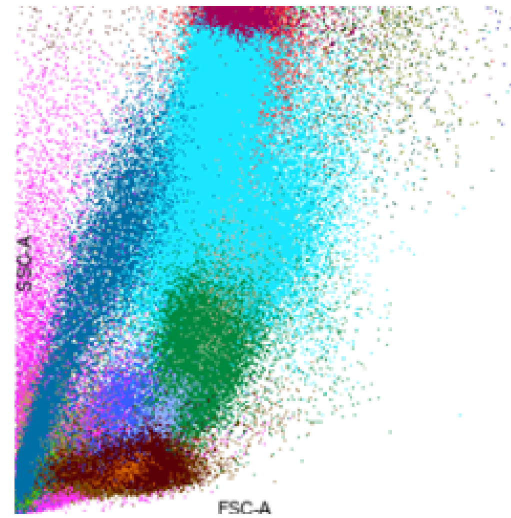
[Sensors (Sysnav courtesy)]

Modern data carry complex, but important, geometric/topological structure!

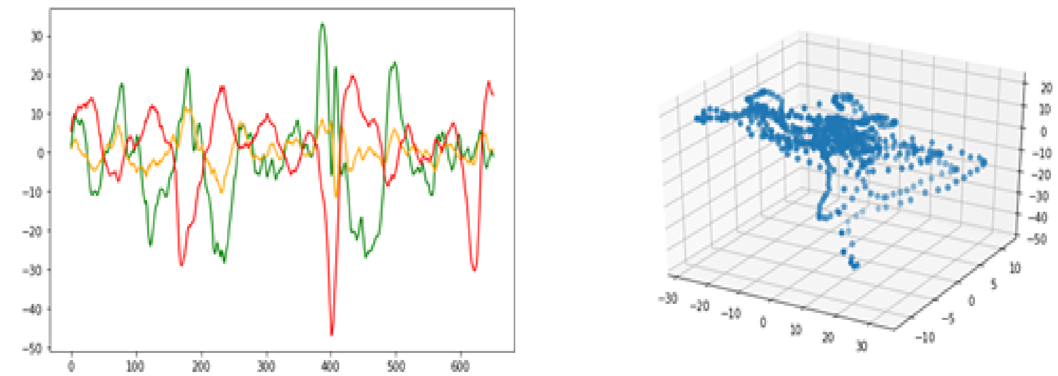
What is Topological Data Analysis (TDA)?



[Force fields in granular media (Krama et al)]



[Cell population - cytometry - MetaFora courtesy]

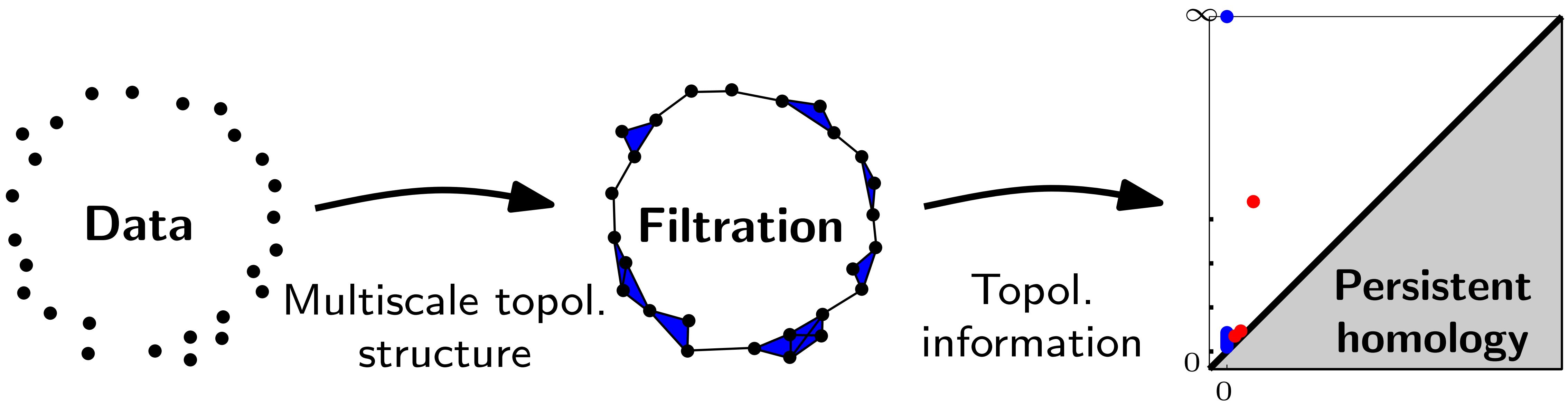


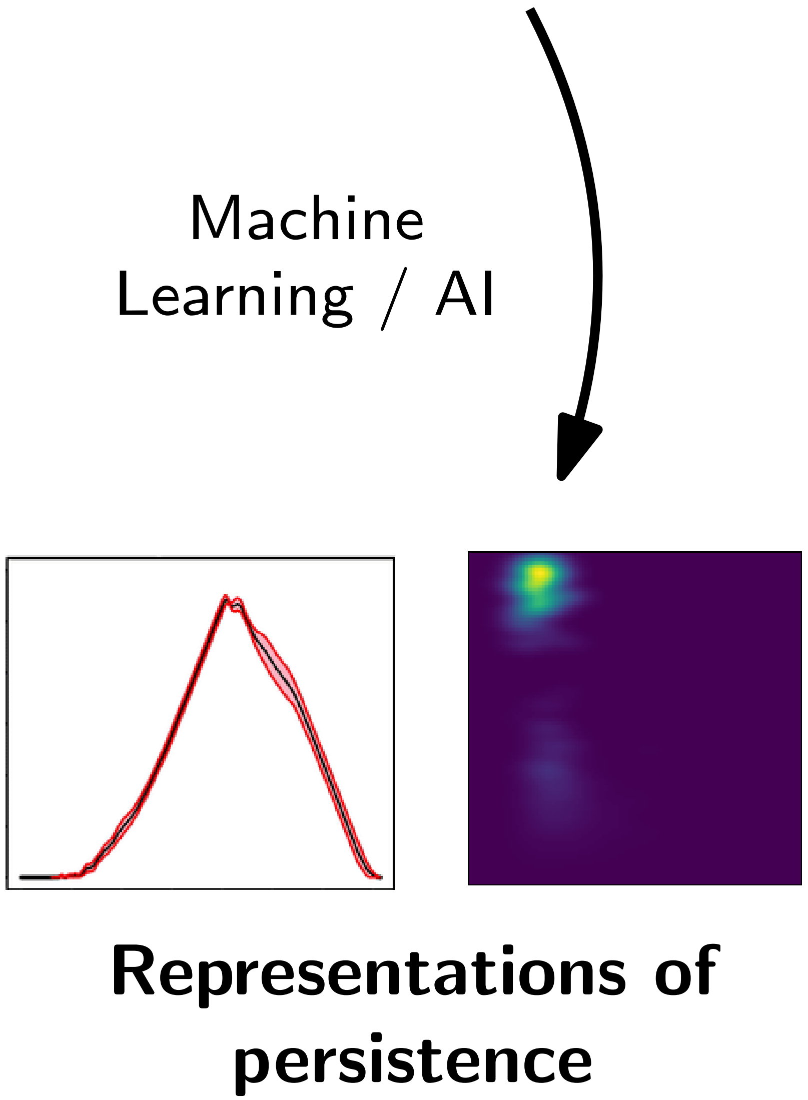
[Sensors (Sysnav courtesy)]

Topological Data Analysis (TDA) is a recent field whose aim is to:

- infer relevant topological and geometric features from complex data,
- take advantage of topological/geometric information for further Data Analysis, Machine Learning and AI tasks:
  - using topological features in ML pipelines,
  - taking advantage of topological information to improve ML pipelines.

The classical TDA pipeline





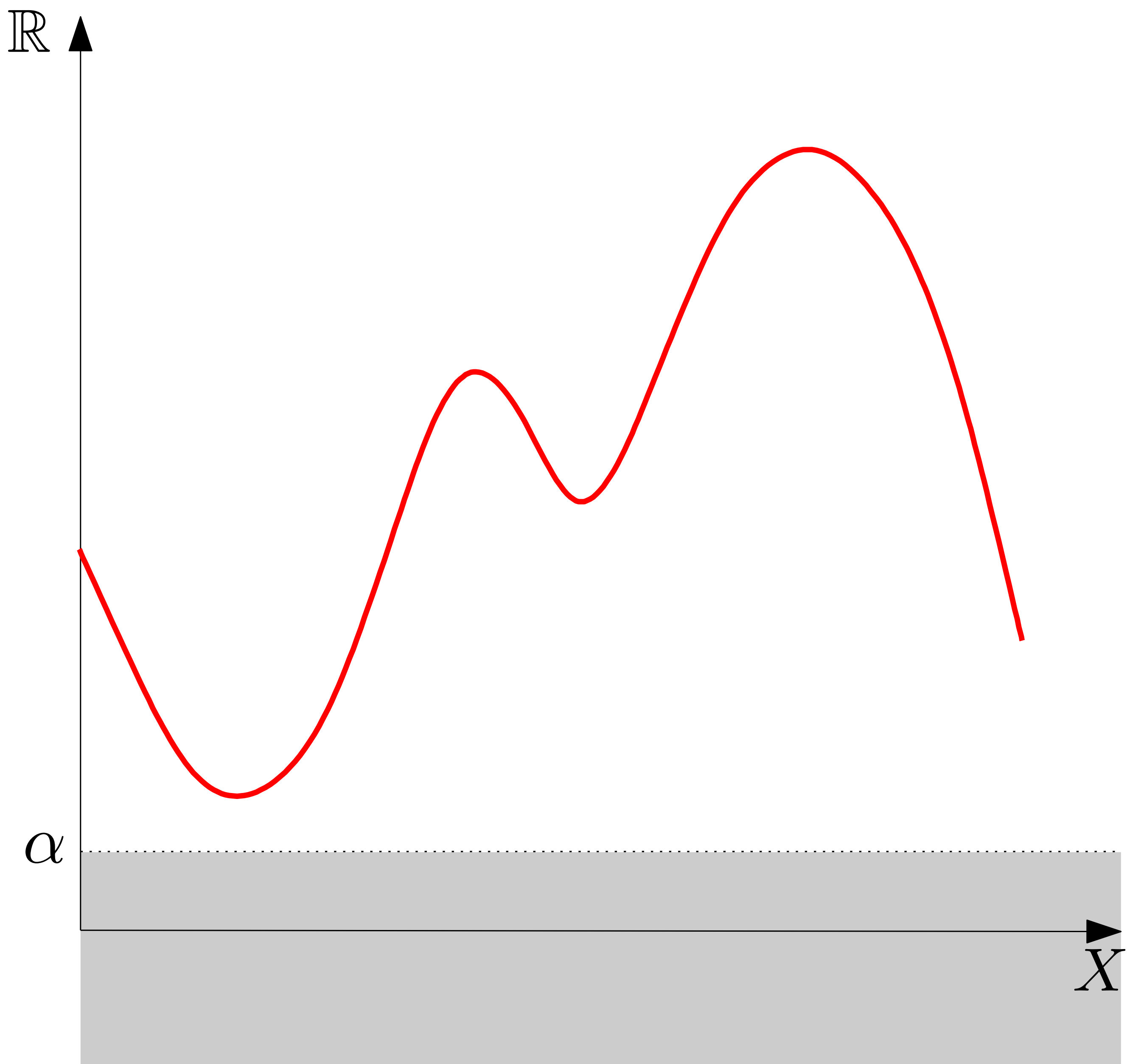
1. Build a multiscale topol. structure on top of data: filtrations.
2. Compute multiscale topol. signatures: persistent homology
3. Take advantage of the signature for further Machine Learning and AI tasks: Statistical aspects and representations of persistence

Persistent homology
Starting with a few examples

- $90's$: size theory (P. Frosini et al), framed Morse complex and stability (S.A. Barannikov).
- $2002 - 2005$: persistent homology (H. Edelsbrunner et al, Carlsson et al).
- important mathematical and practical developments since the 2000's.

Persistent homology for functions

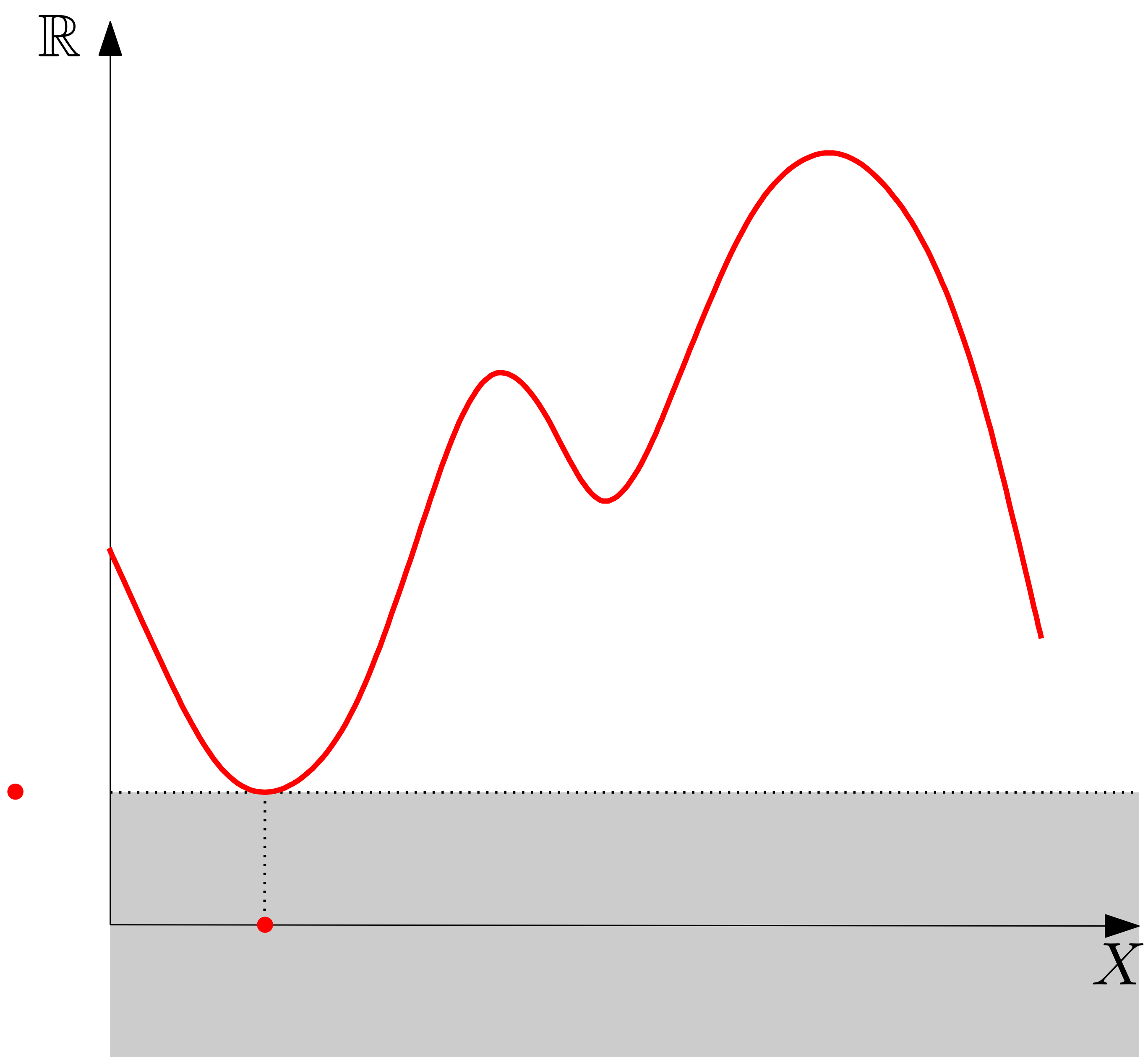
- Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function
- The family of sublevel sets of a function is an example of filtration.



Persistent homology for functions

• Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function

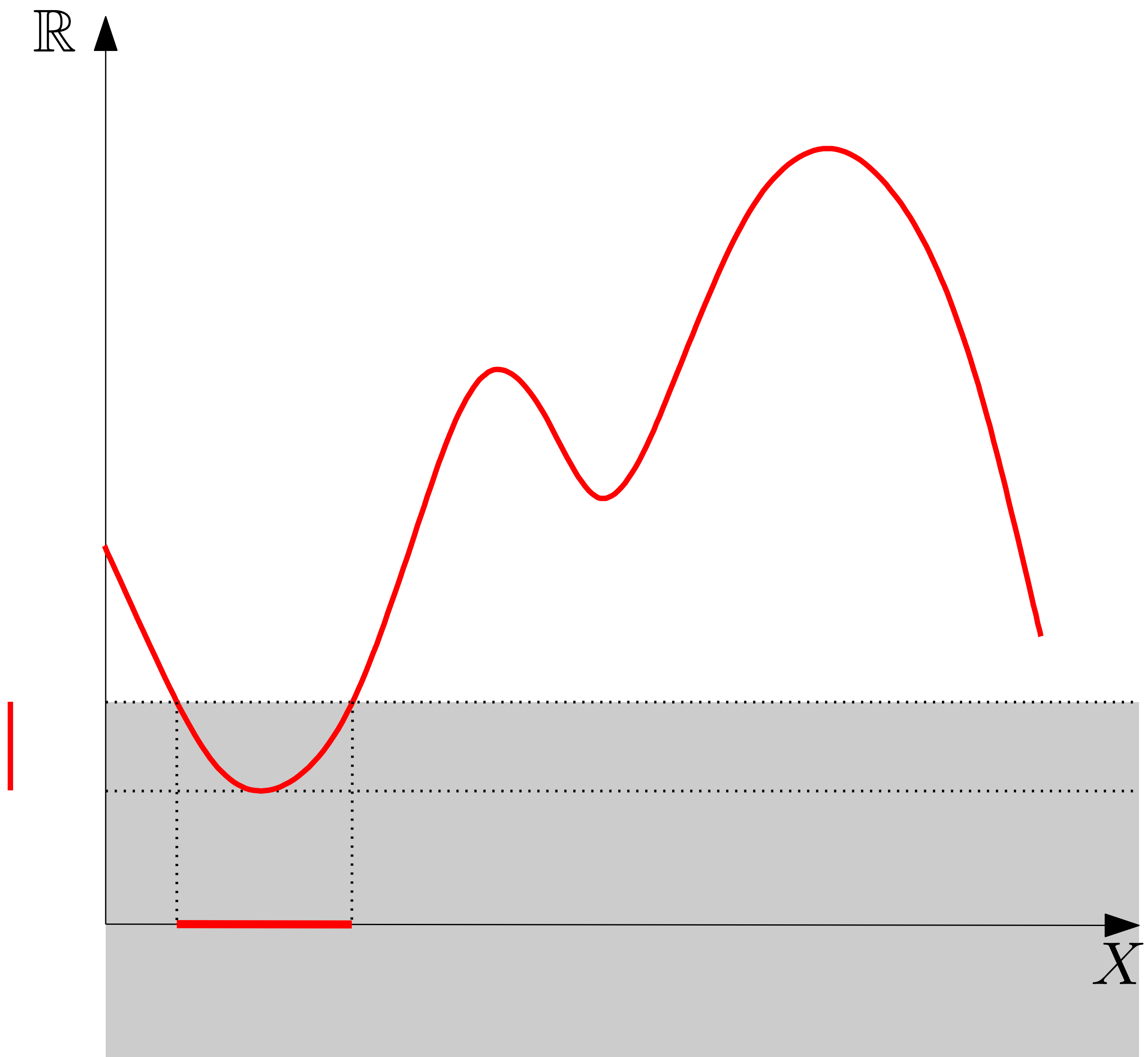
• The family of sublevel sets of a function is an example of filtration.



Persistent homology for functions

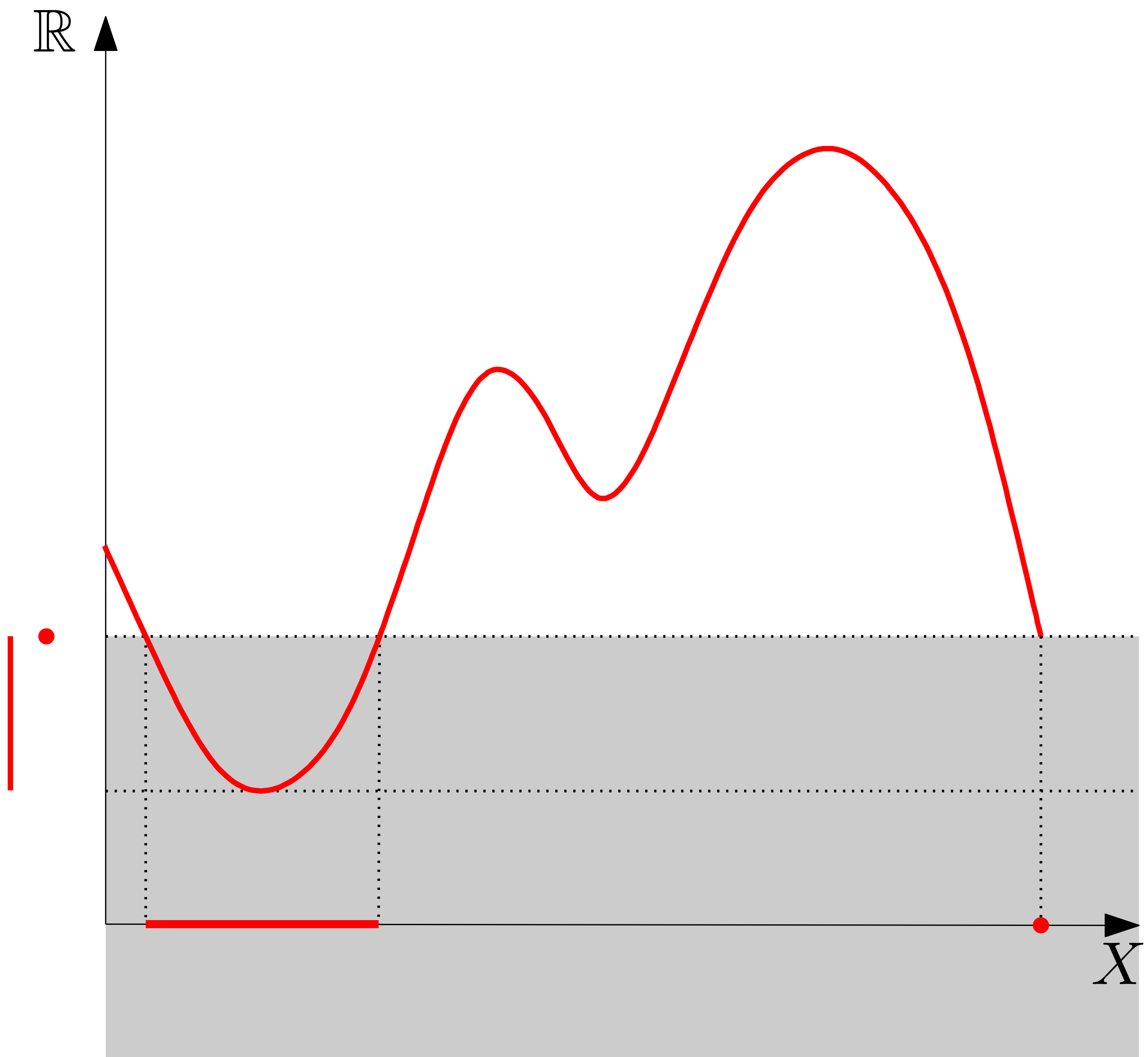
• Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function

• The family of sublevel sets of a function is an example of filtration.



Persistent homology for functions

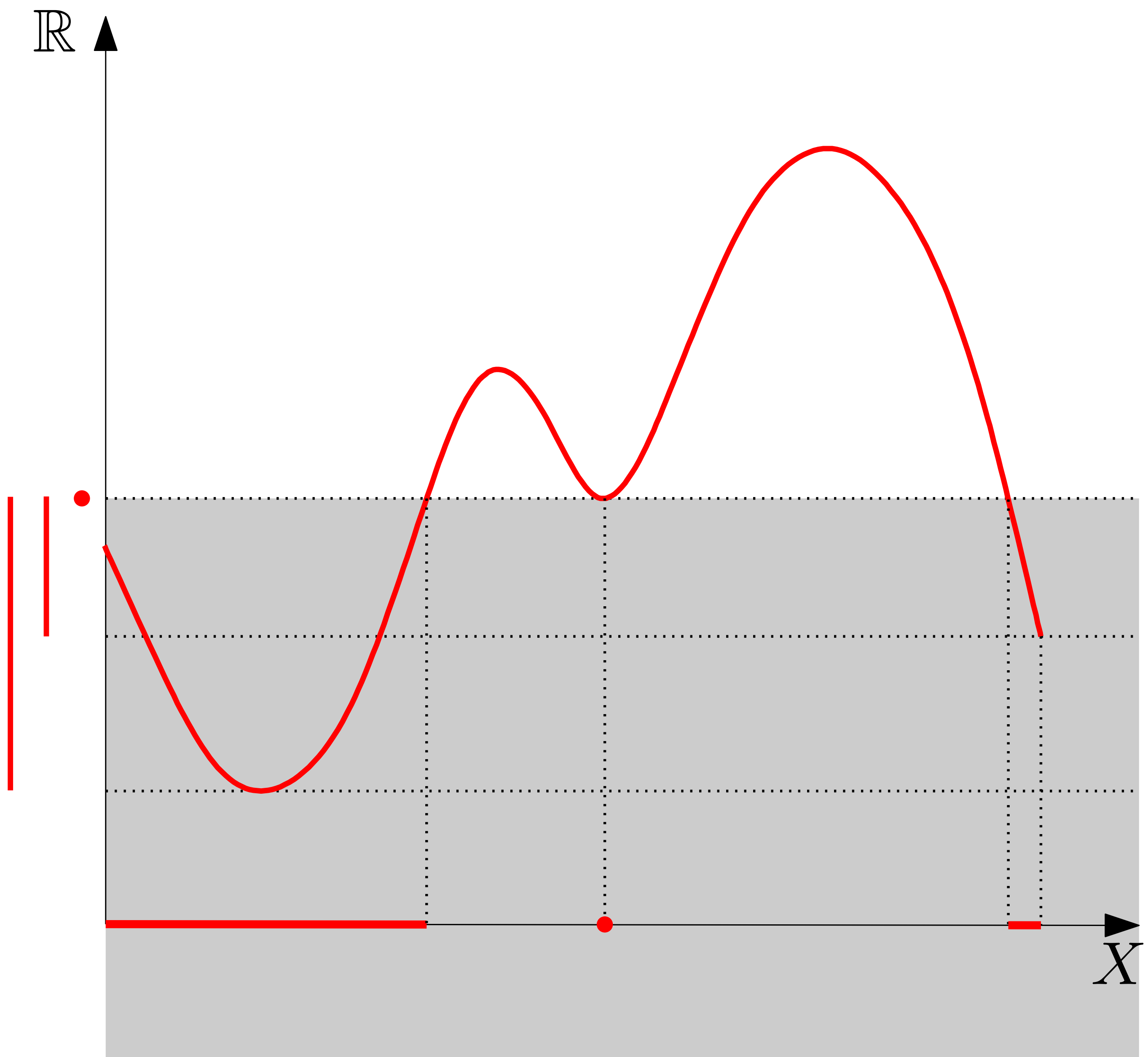
- Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function
- The family of sublevel sets of a function is an example of filtration.



Persistent homology for functions

• Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function

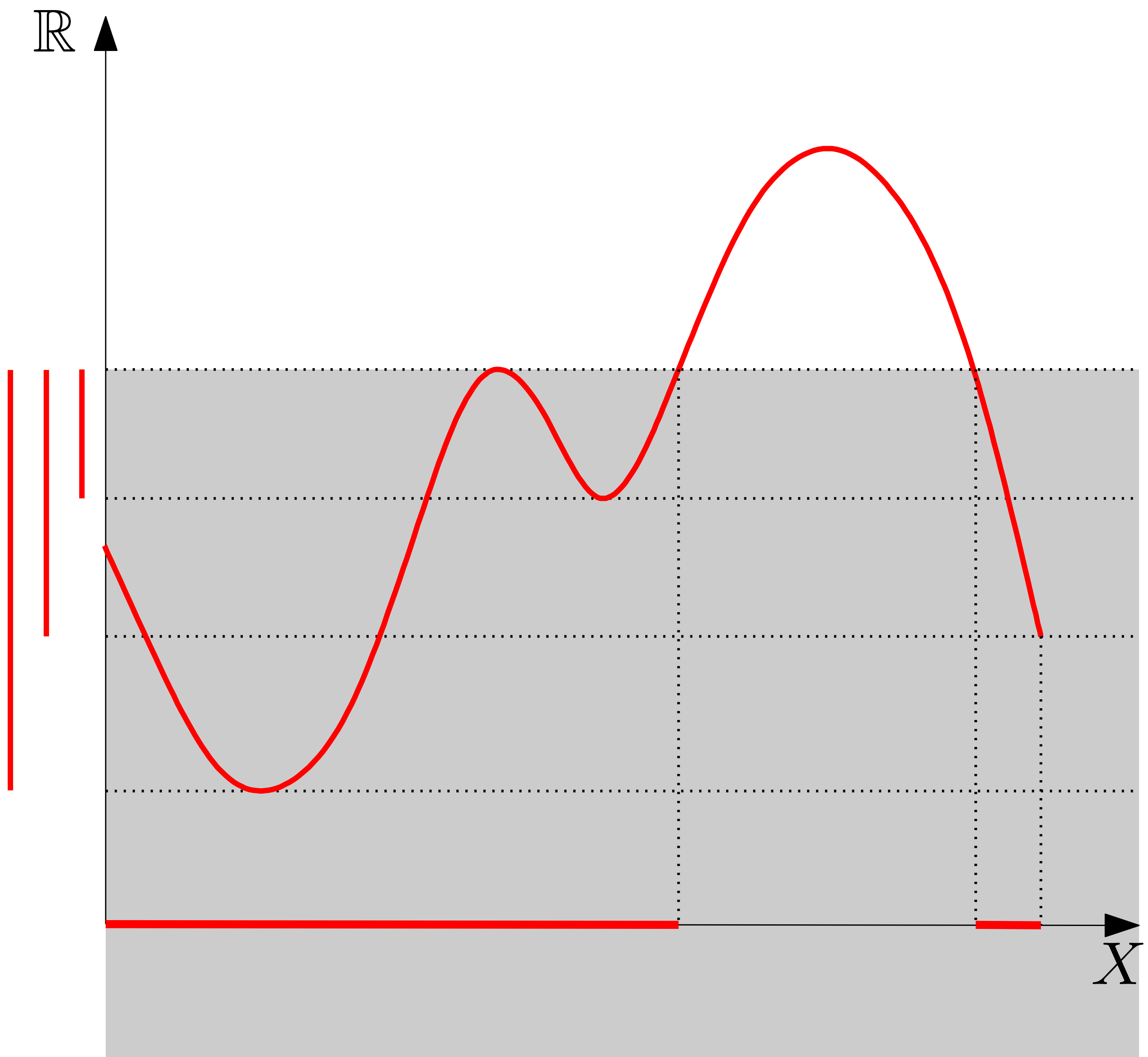
• The family of sublevel sets of a function is an example of filtration.



Persistent homology for functions

• Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function

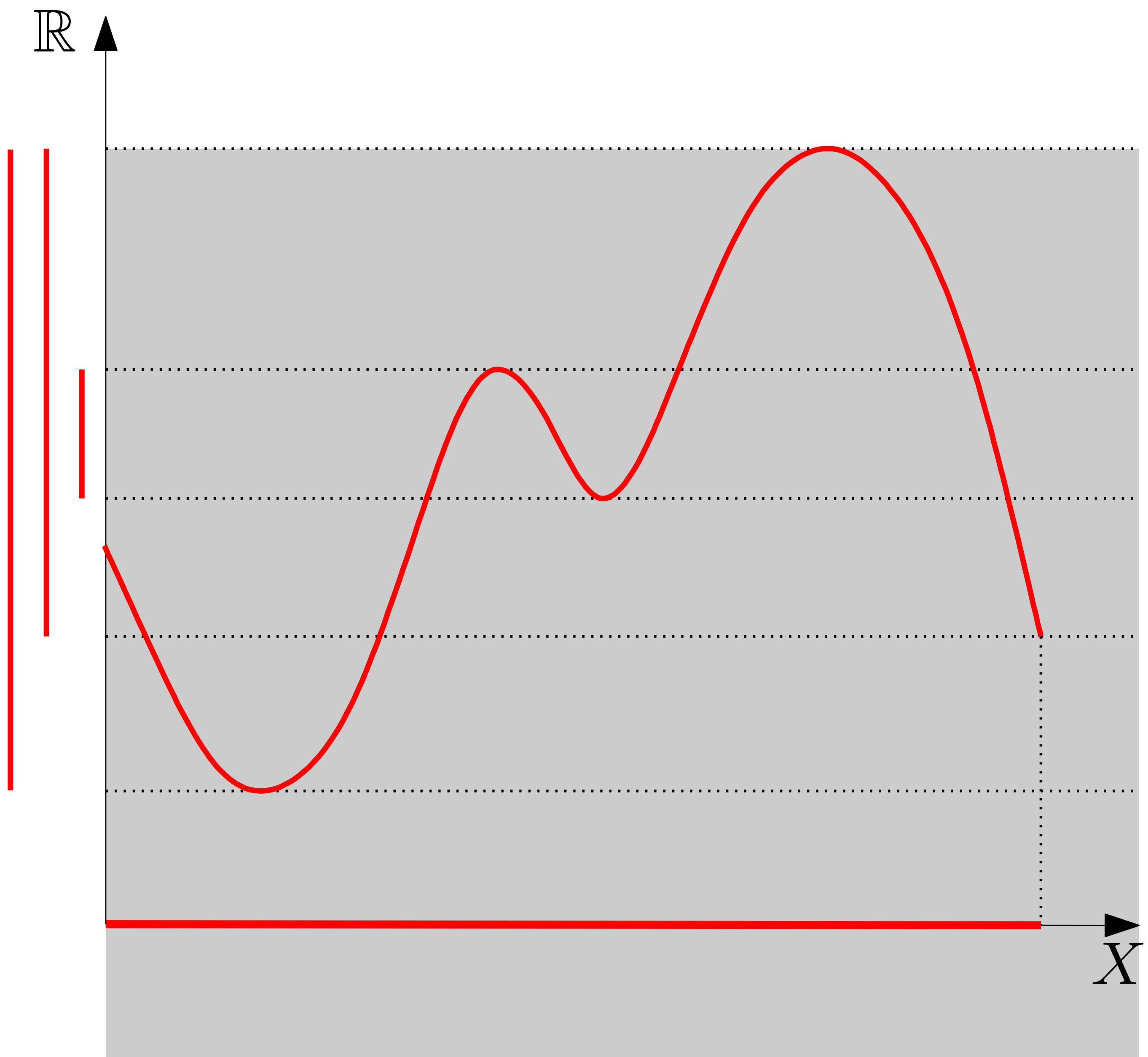
• The family of sublevel sets of a function is an example of filtration.



Persistent homology for functions

• Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function

• The family of sublevel sets of a function is an example of filtration.

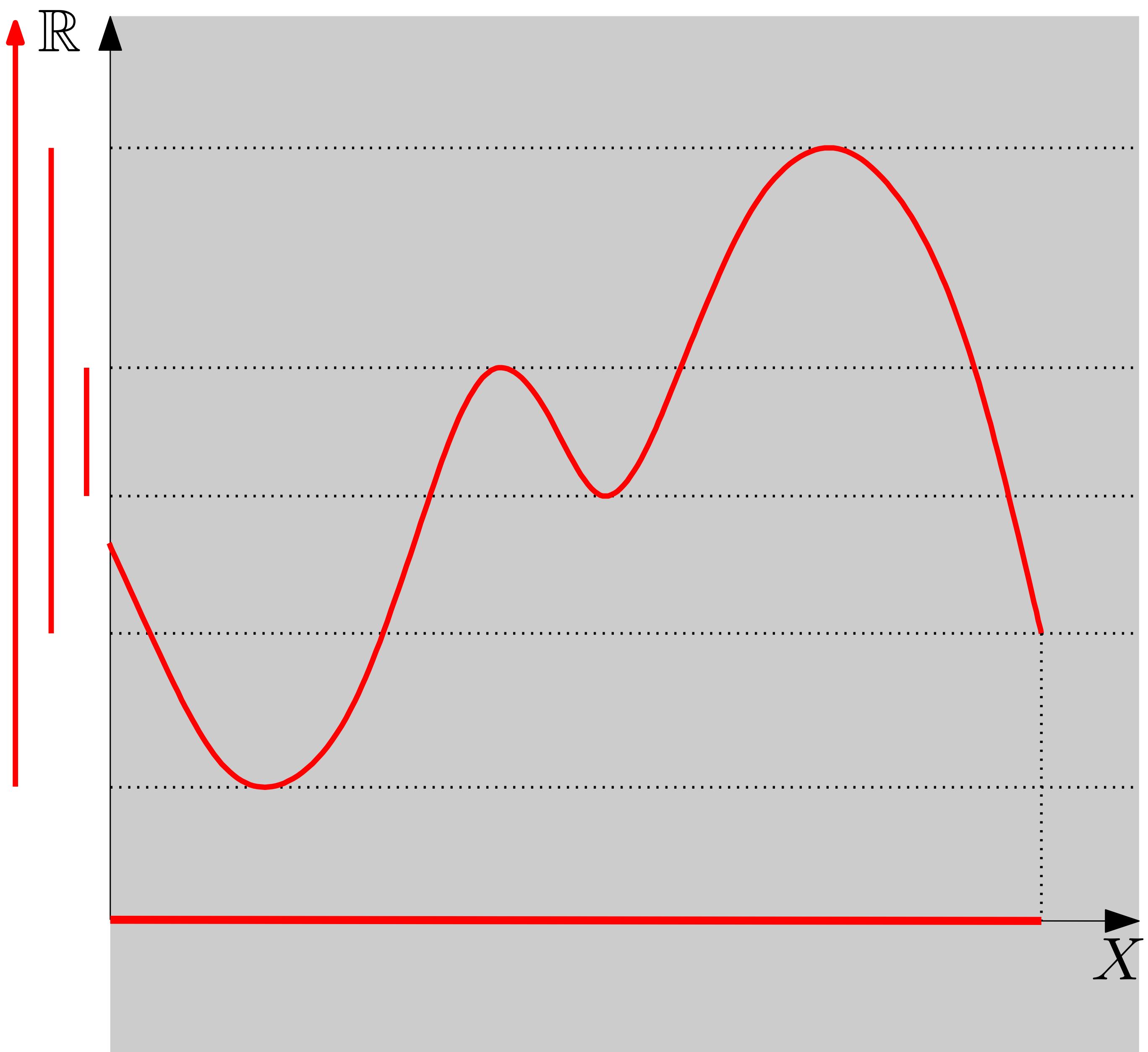


Persistent homology for functions

• Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function

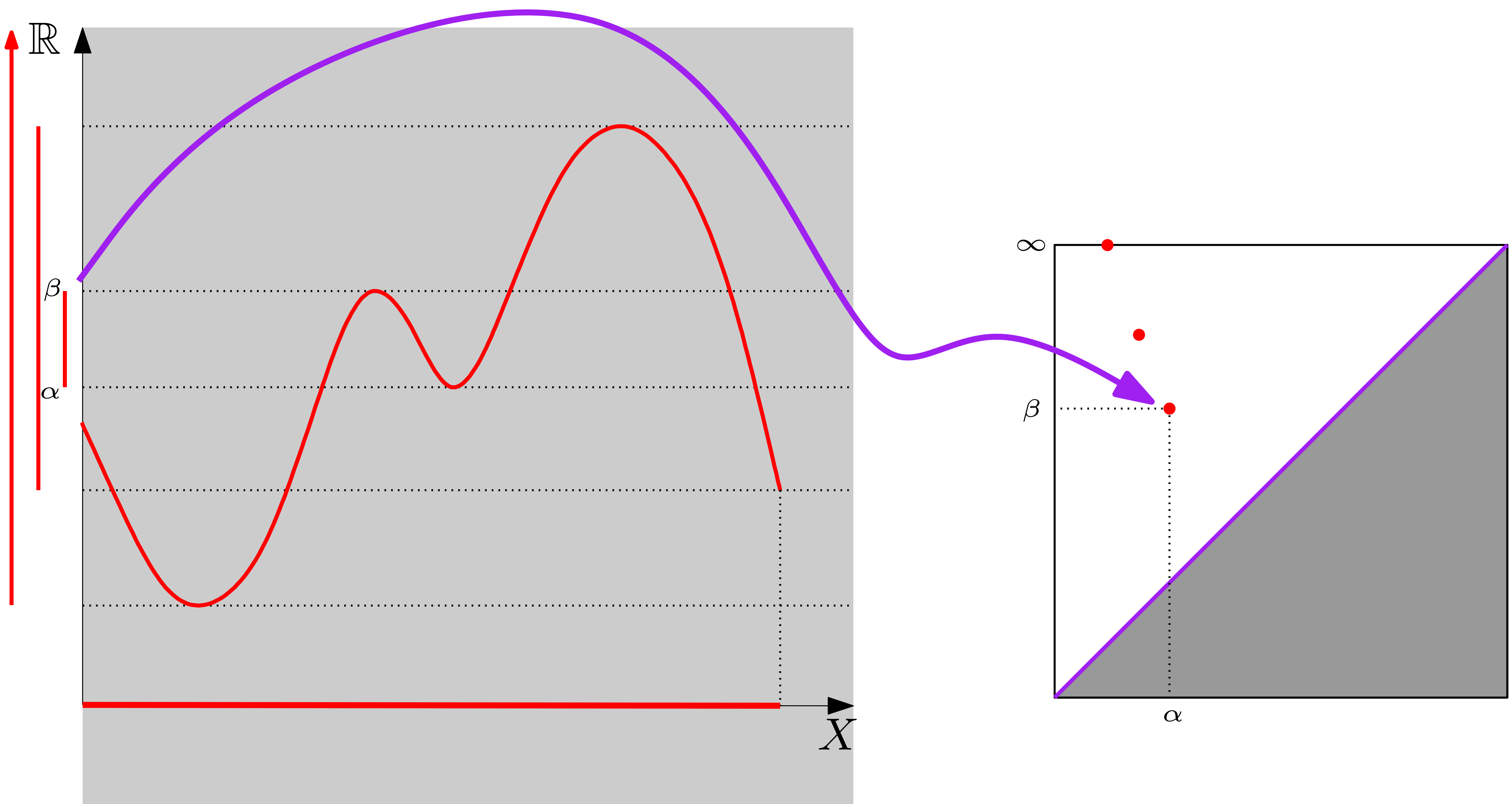
• The family of sublevel sets of a function is an example of filtration.

• Finite set of intervals (barcode) encodes births/deaths of topological features.



Persistent homology for functions

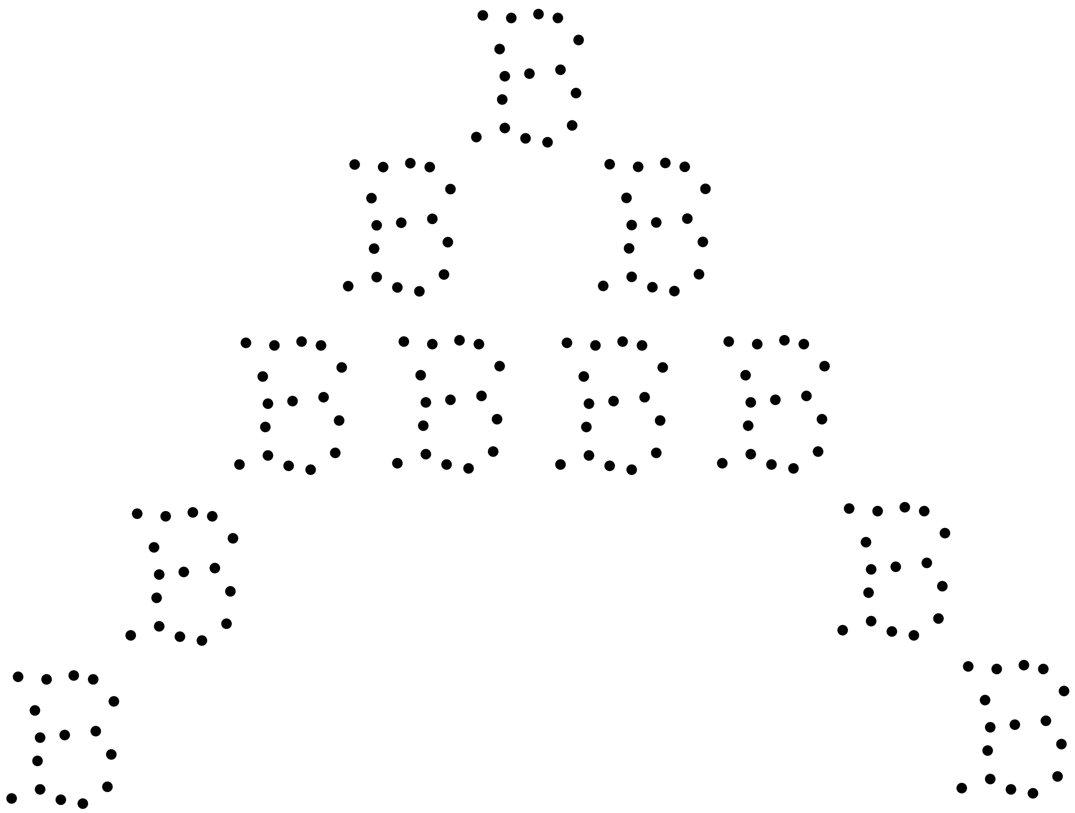
- Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function
- The family of sublevel sets of a function is an example of filtration.
- Finite set of intervals (barcode) encodes births/deaths of topological features.



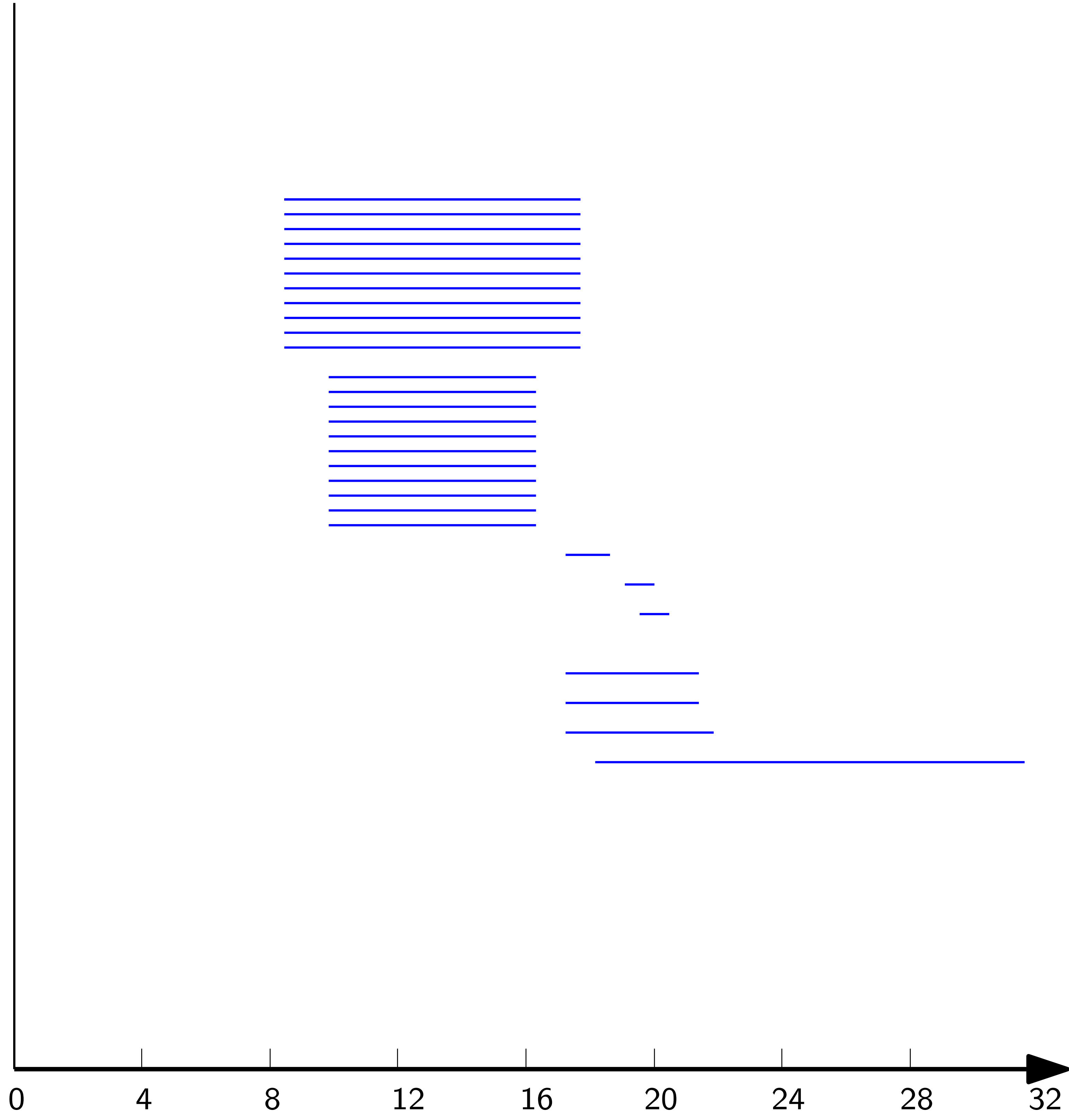
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



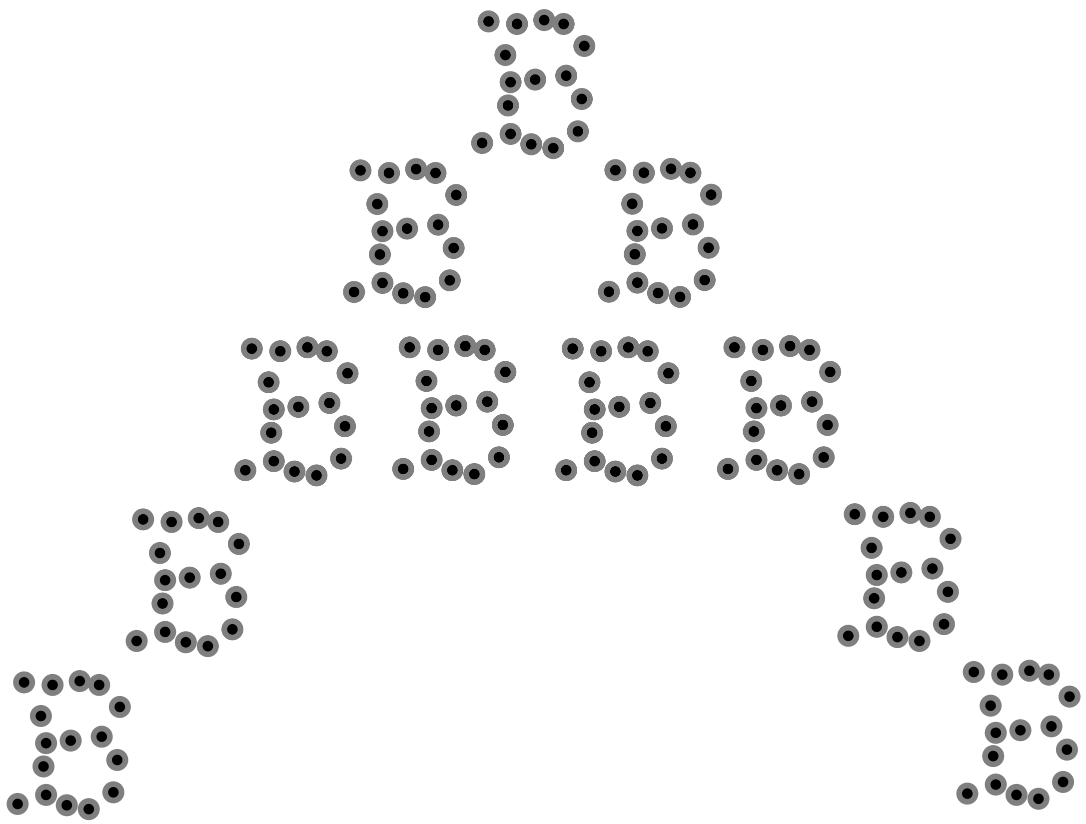
barcode for holes (1-d homology)



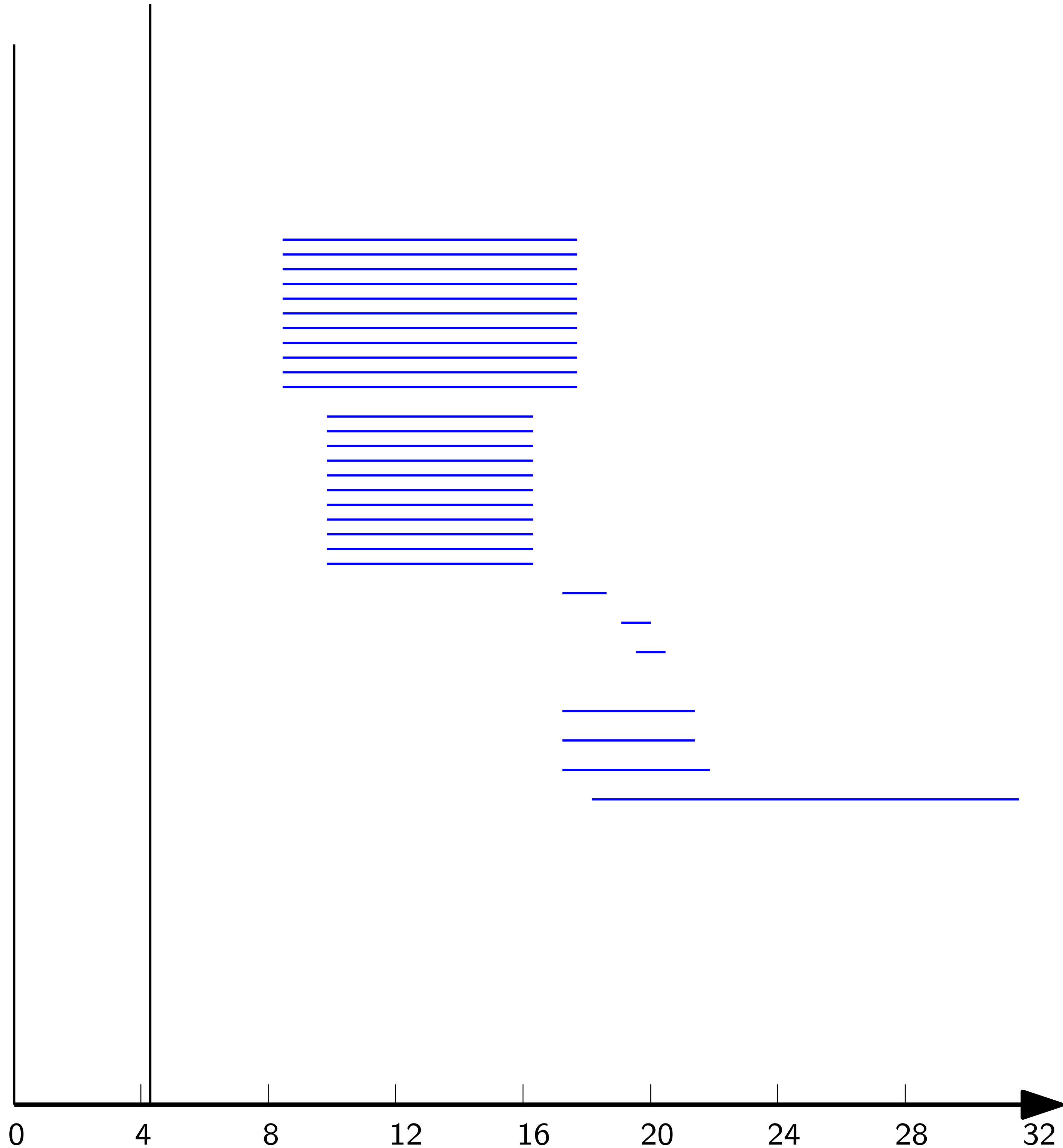
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



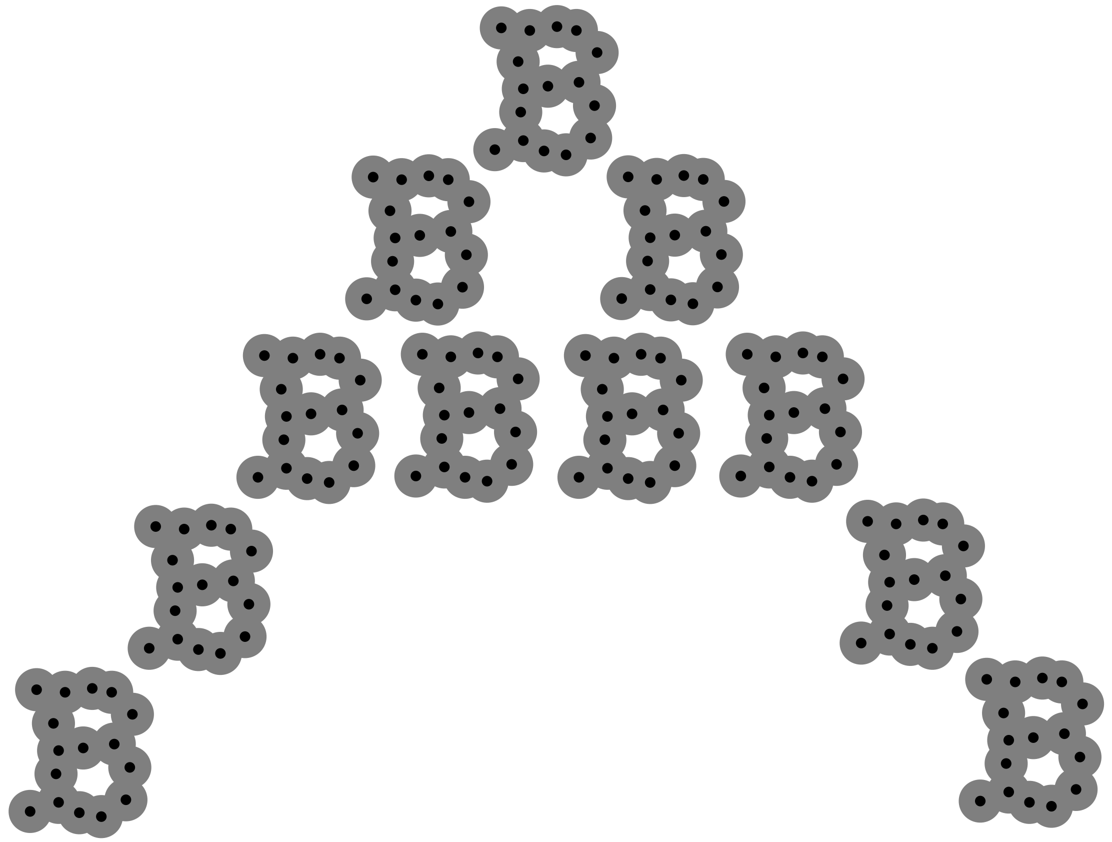
barcode for holes (1-d homology)



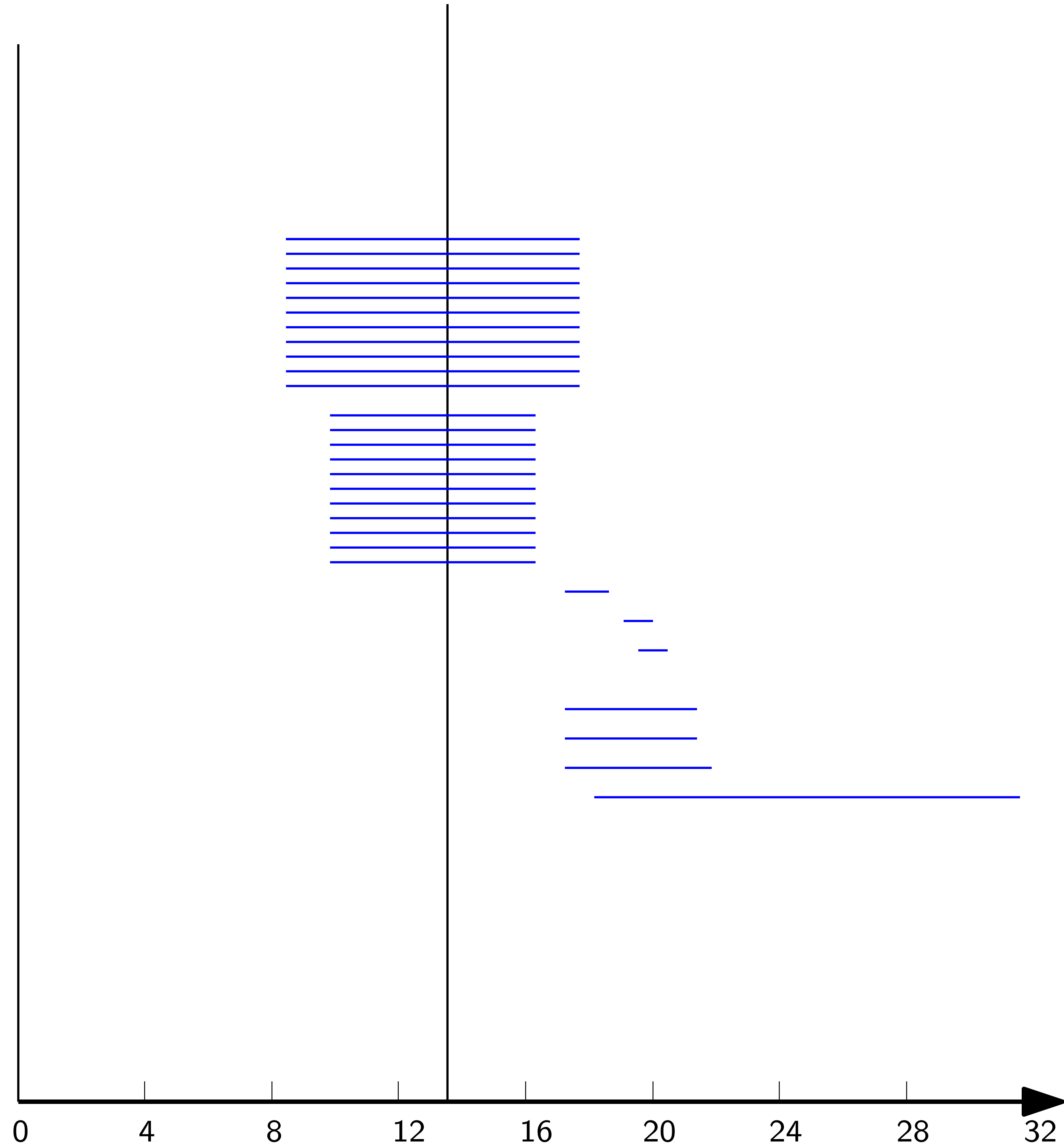
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



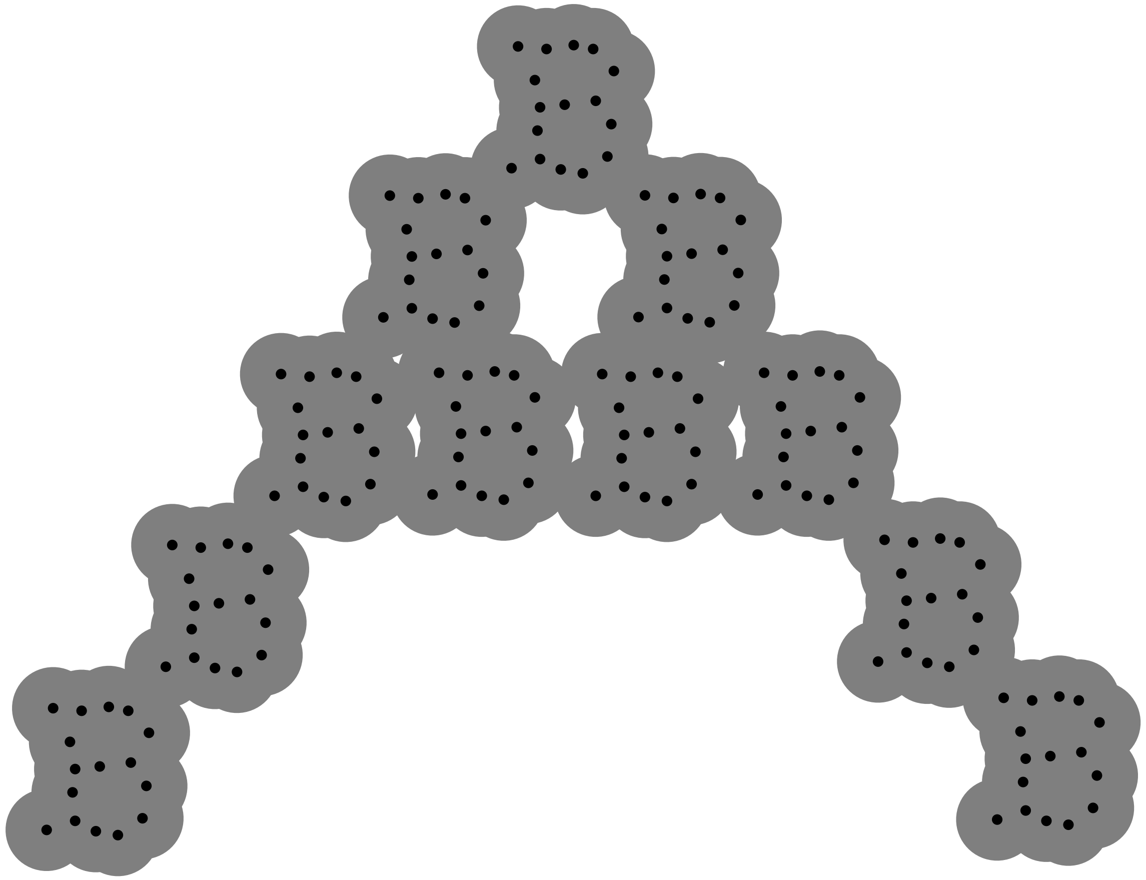
barcode for holes (1-d homology)



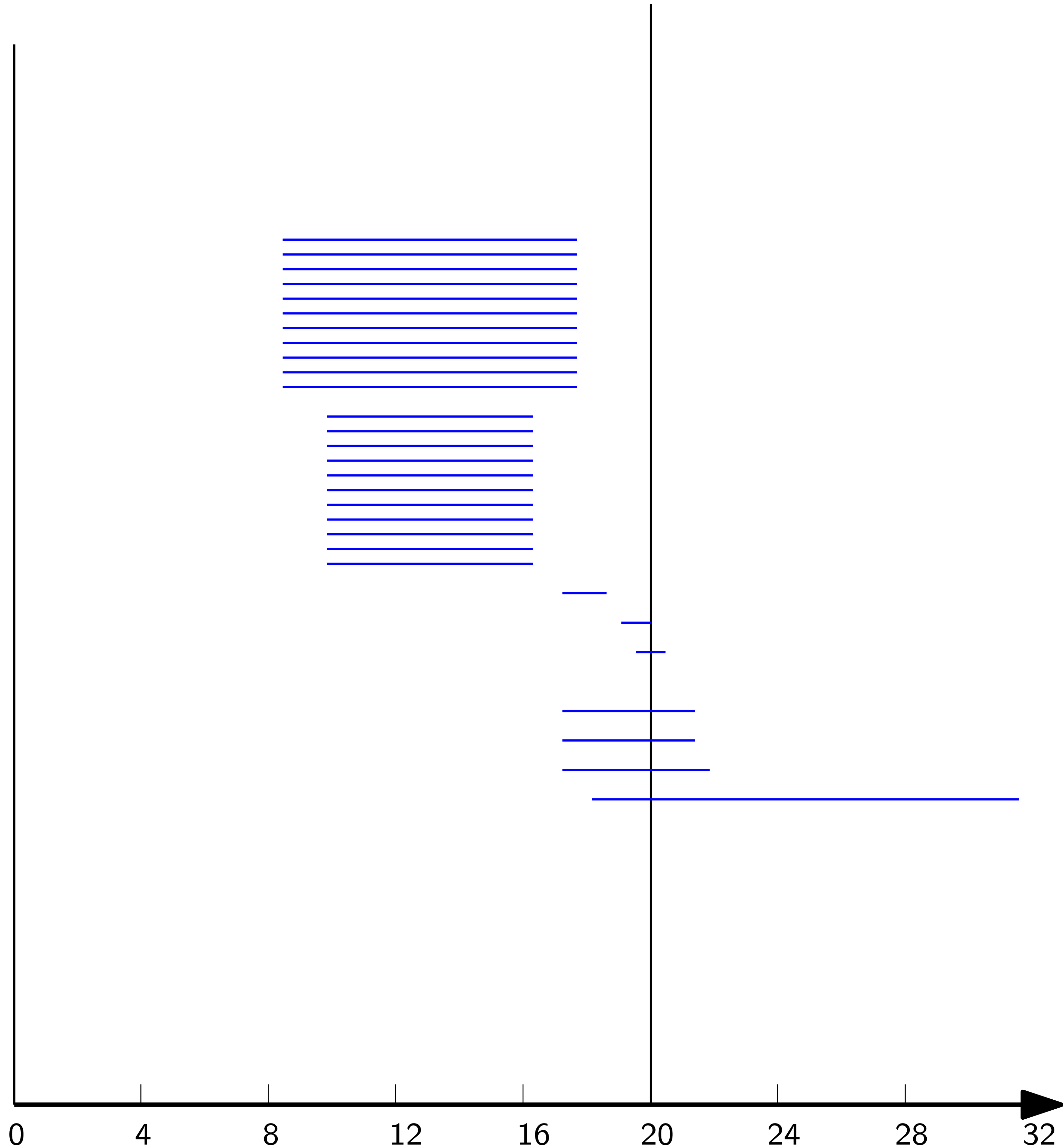
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



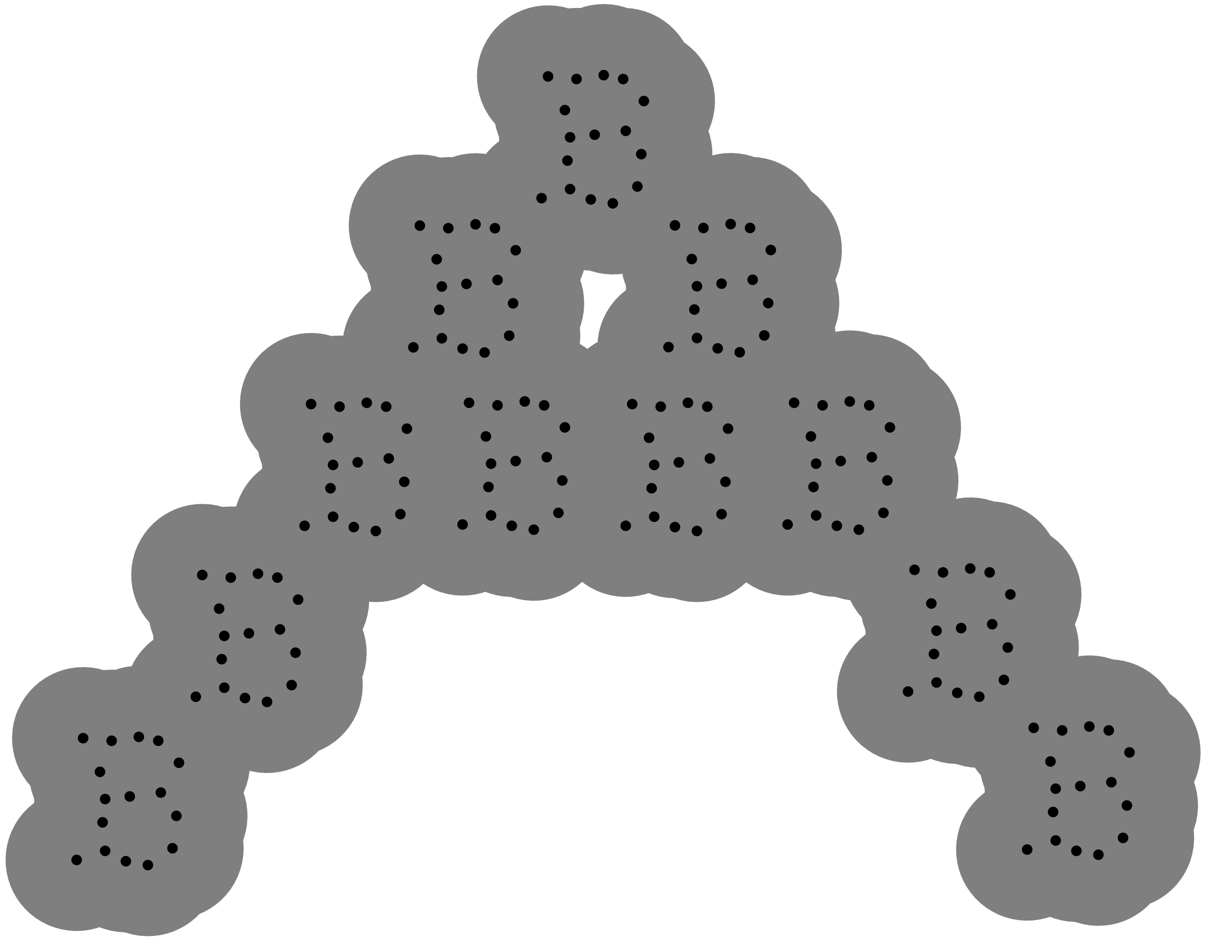
barcode for holes (1-d homology)



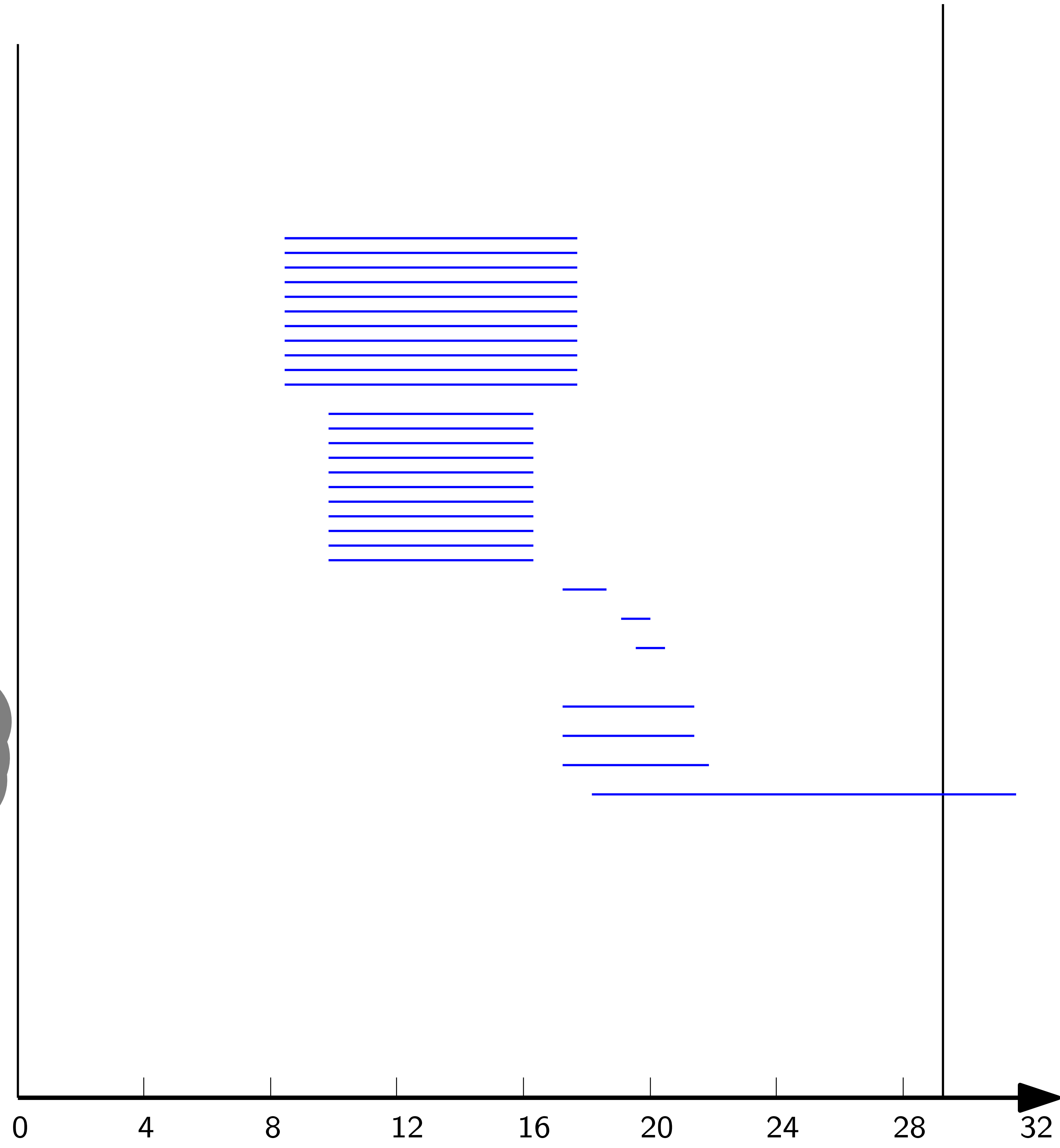
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



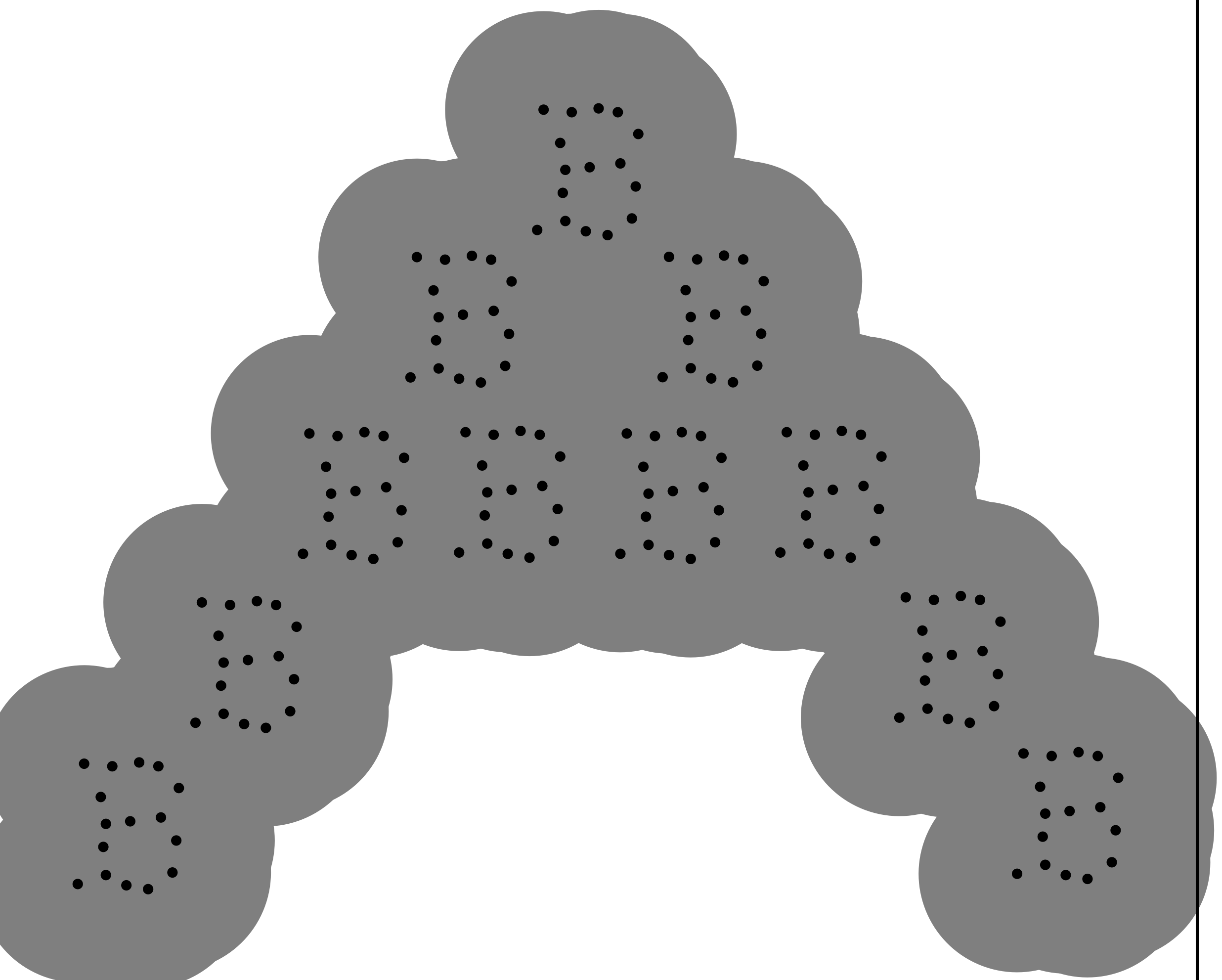
barcode for holes (1-d homology)



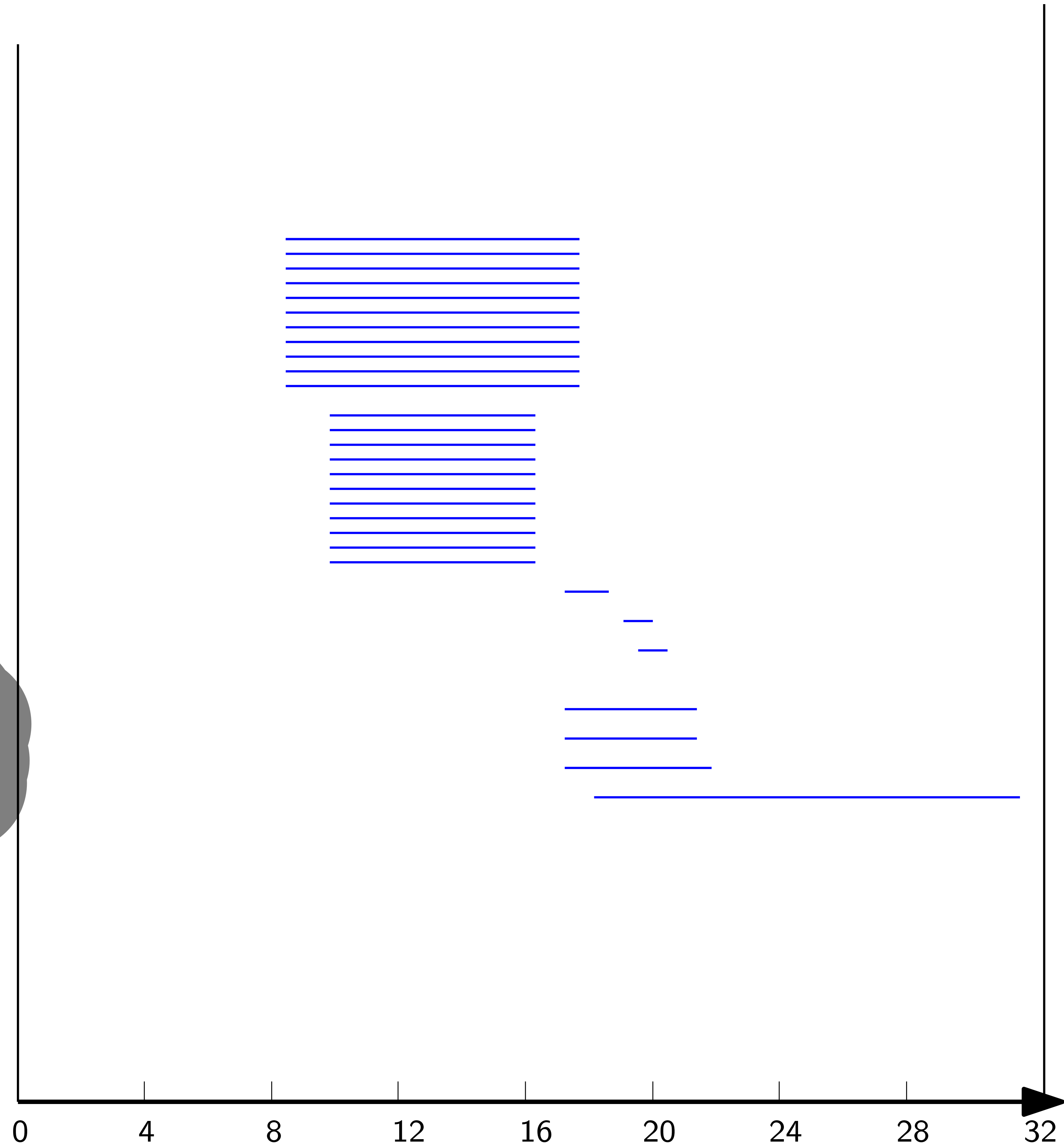
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

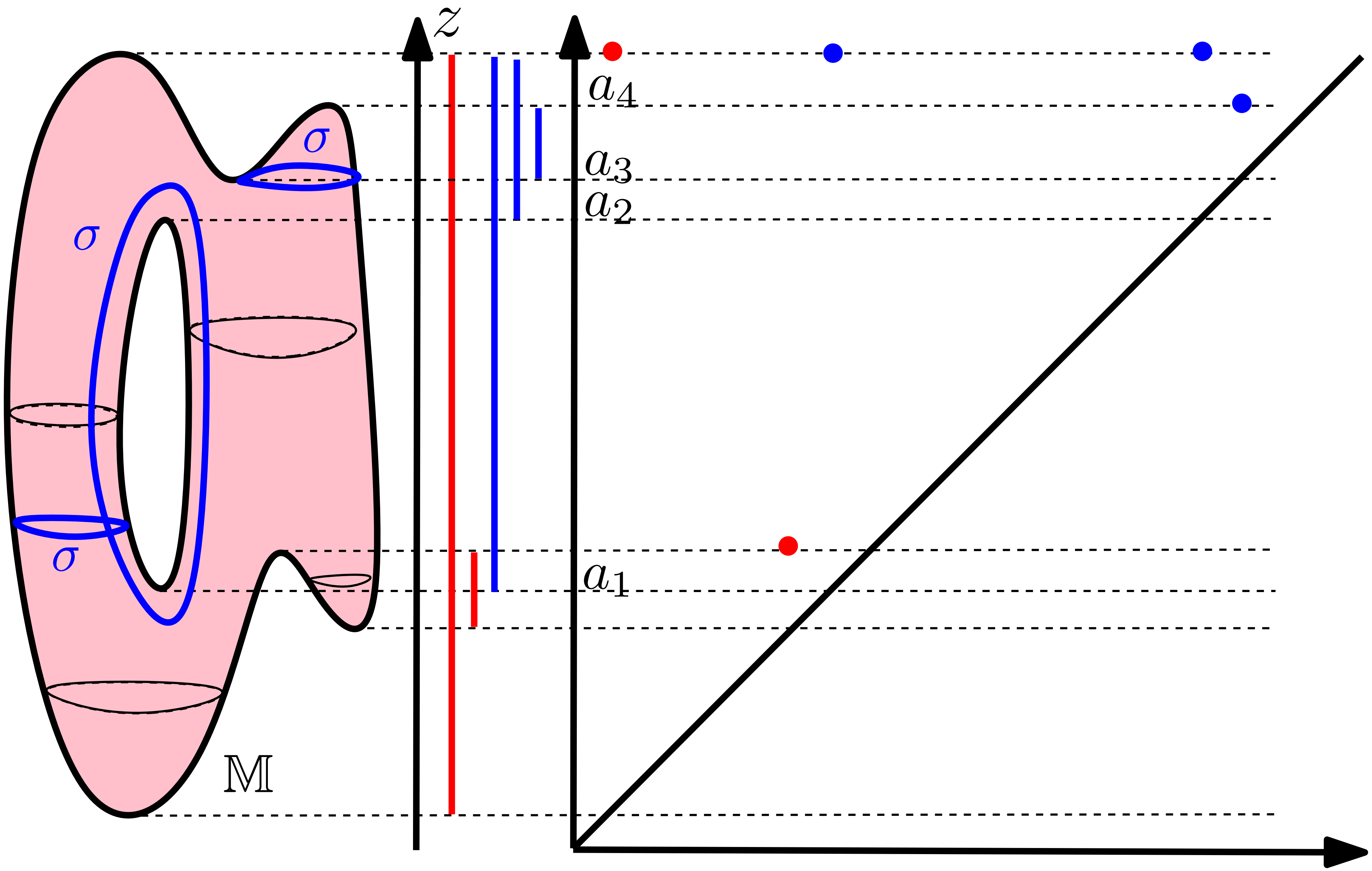
$$x \to \min_{p \in P} \|x - p\|_2$$



barcode for holes (1-d homology)



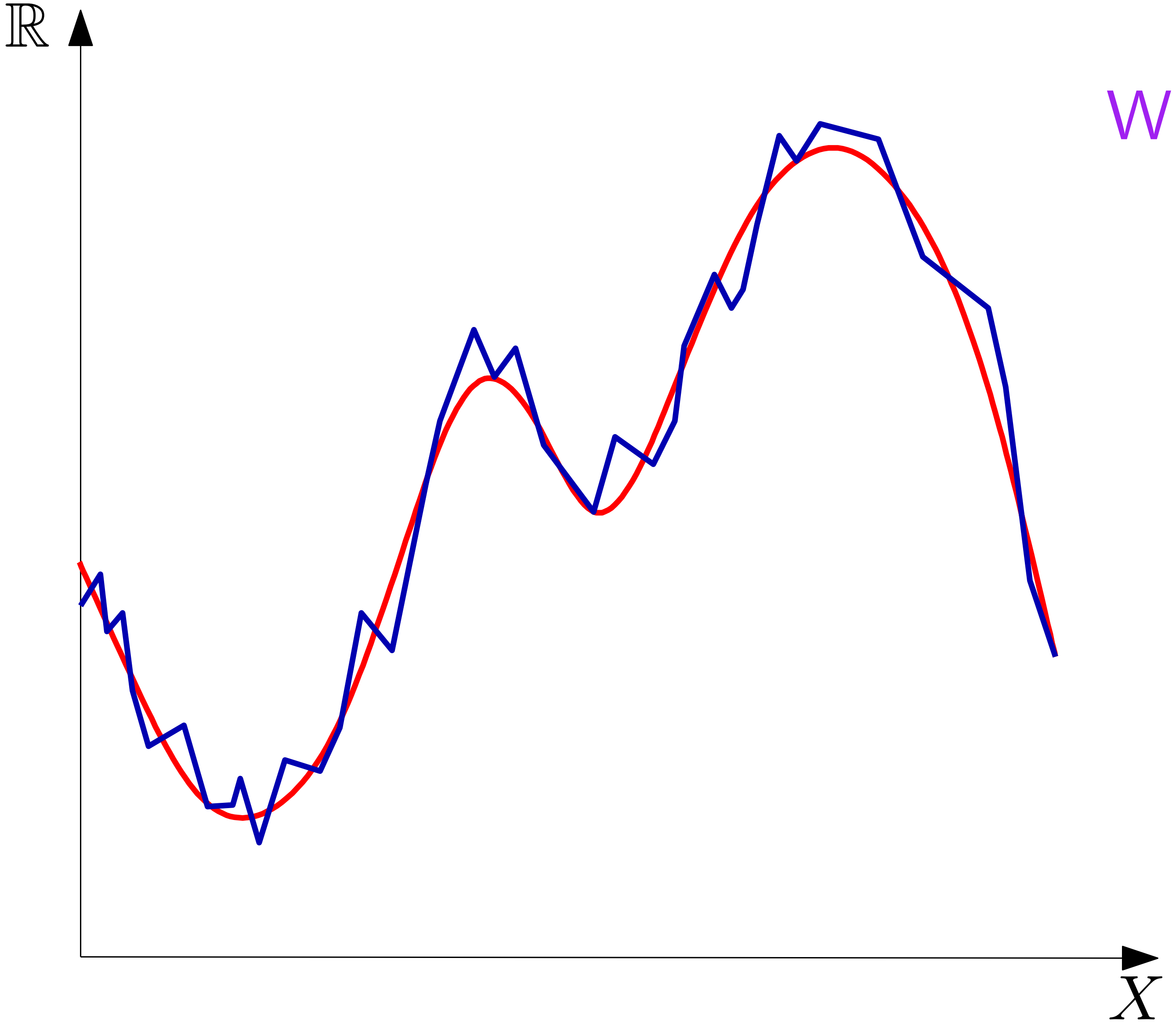
Persistent homology for functions



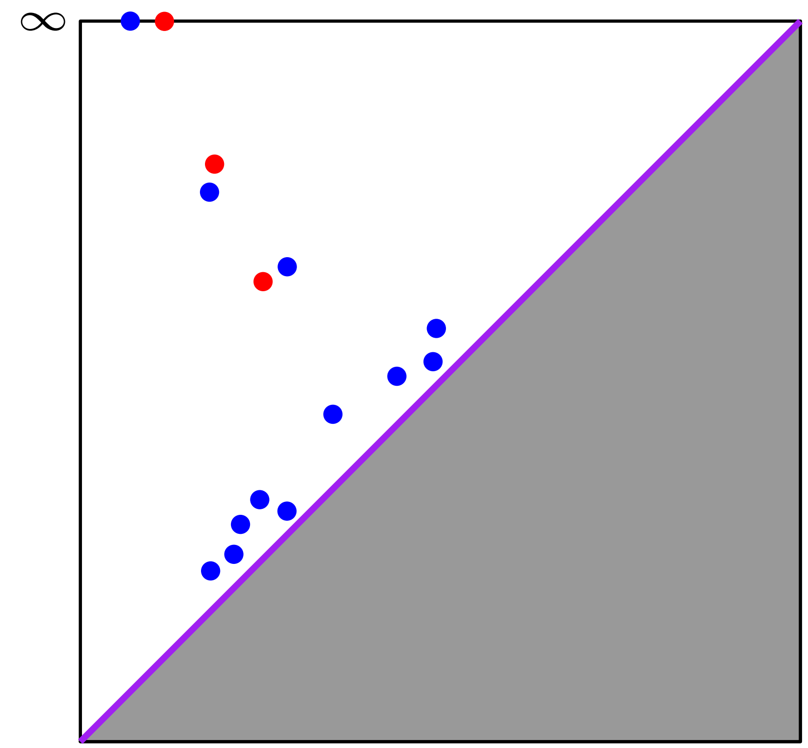
Tracking and encoding the evolution of the connected components (0-dimensional homology) and cycles (1-dimensional homology) of the sublevel sets.

Homology: an algebraic way to rigorously formalize the notion of k-dimensional cycles through a vector space (or a group), the homology group whose dimension is the number of "independent" cycles (the Betti number).

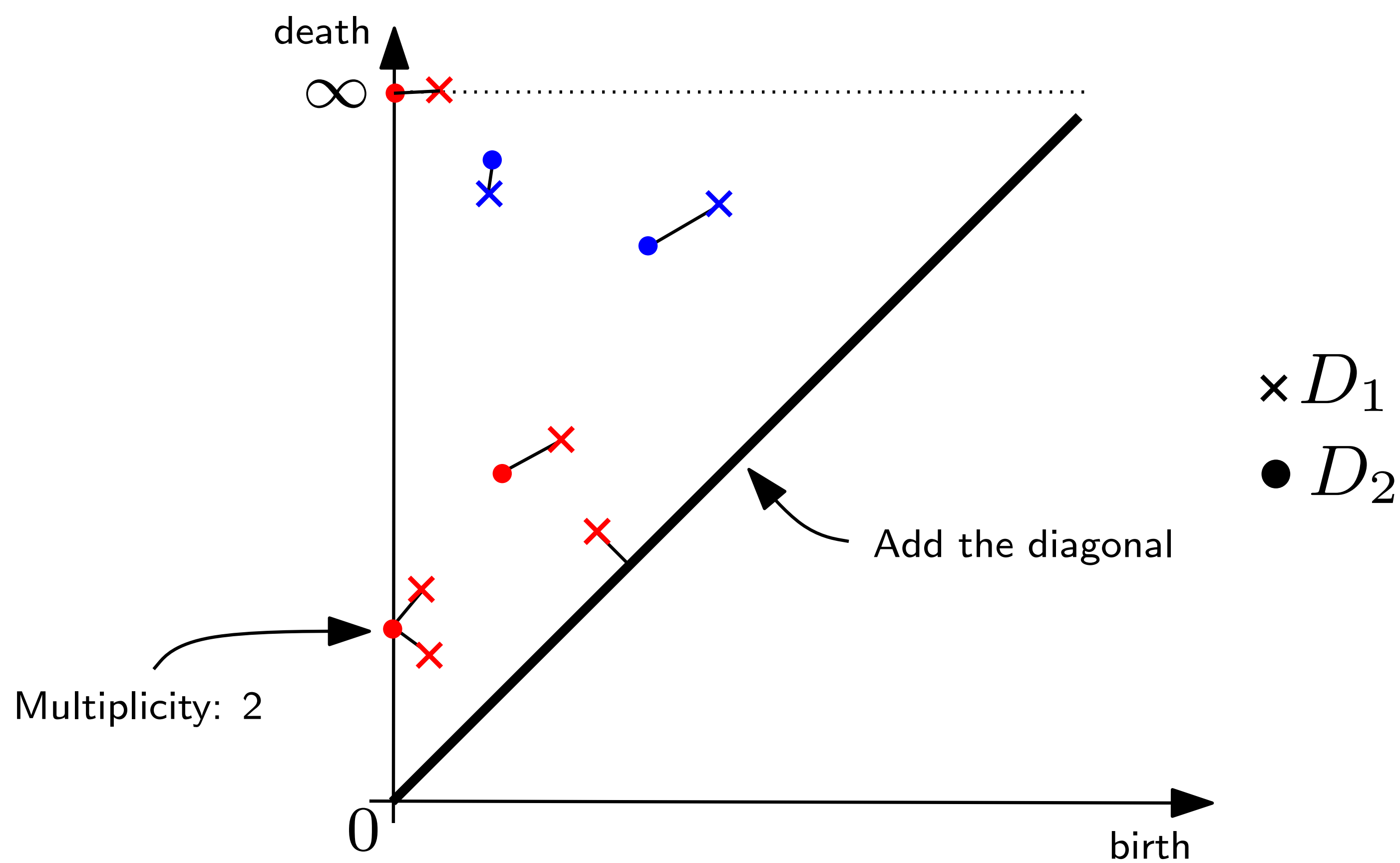
Stability properties



What if f is slightly perturbed?



Comparing persistence diagrams



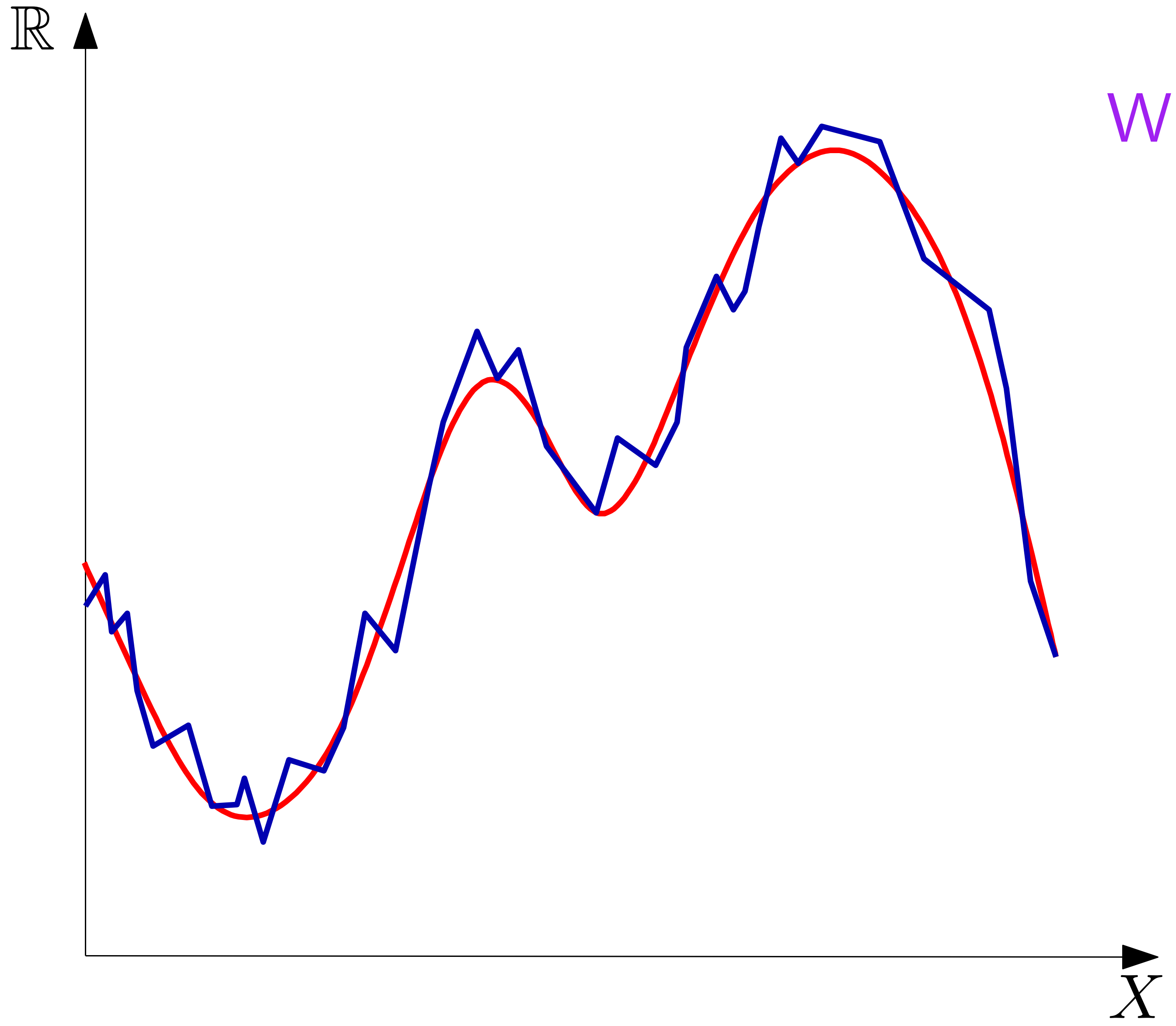
The bottleneck distance between two diagrams D_1 and D_2 is

$$d_B(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{m \in D_1} \|m - \gamma(m)\|_\infty$$

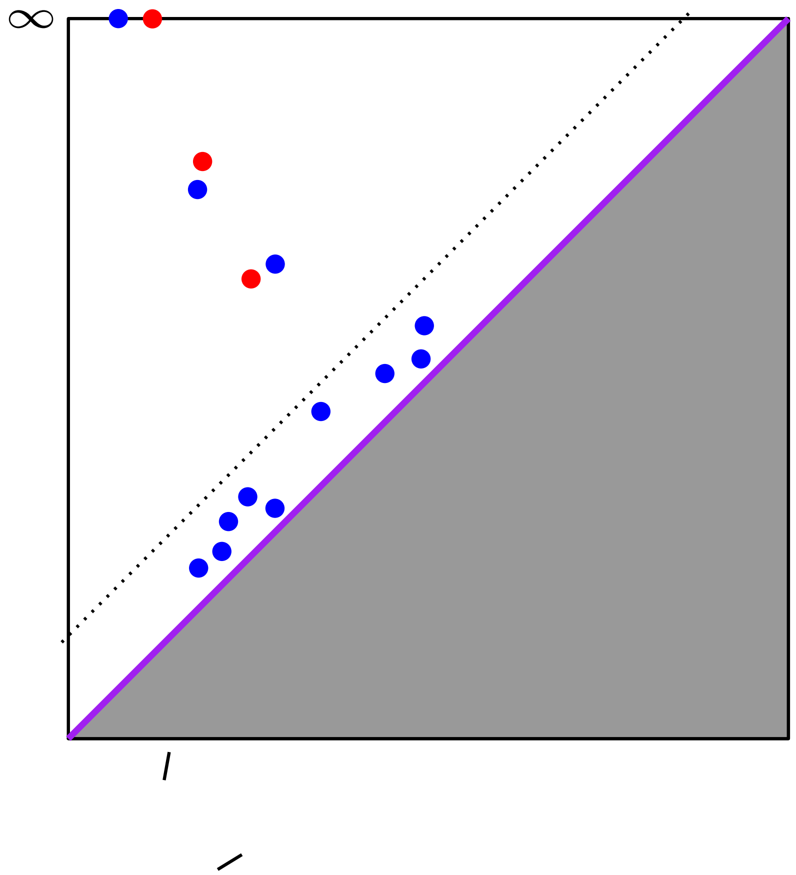
where Γ is the set of all the bijections between D_1 and D_2 and $\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|)$.

Rmk: in many applications, the so-called Wasserstein distance between diagrams is used: $W_p(D_1, D_2) = \inf_{\gamma \in \Gamma} \sum_{m \in D_1} \|m - \gamma(m)\|$, $p \geq 1$.

Stability properties



What if f is slightly perturbed?

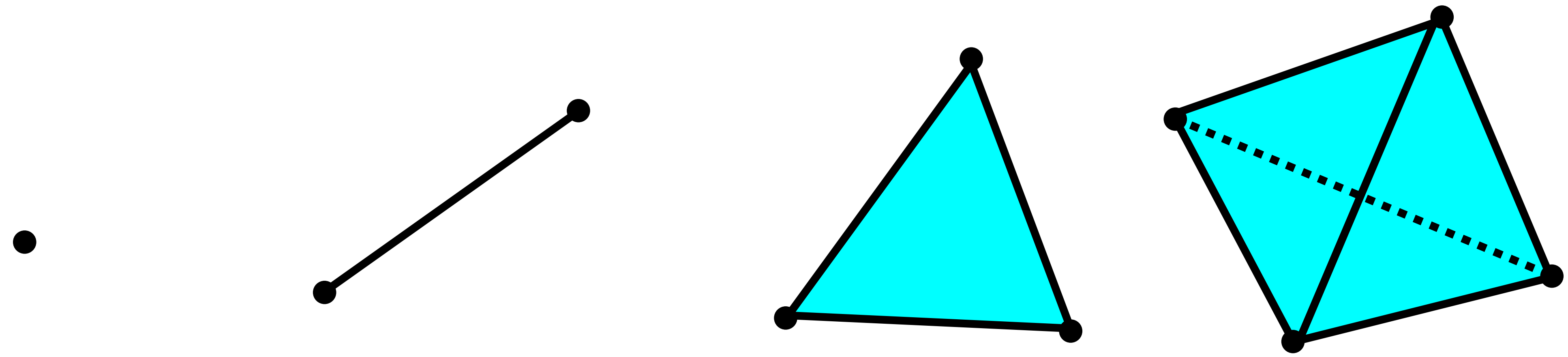


Theorem (Stability):
For any *tame* functions $f, g : \mathbb{X} \to \mathbb{R}$, $d_B(\mathrm{D}_f, \mathrm{D}_g) \leq \|f - g\|_\infty$.

[Cohen-Steiner, Edelsbrunner, Harer 05], [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG 09], [C., de Silva, Glisse, Oudot 12]

Simplicial complexes, filtrations, homology and persistent homology

Simplicial complexes



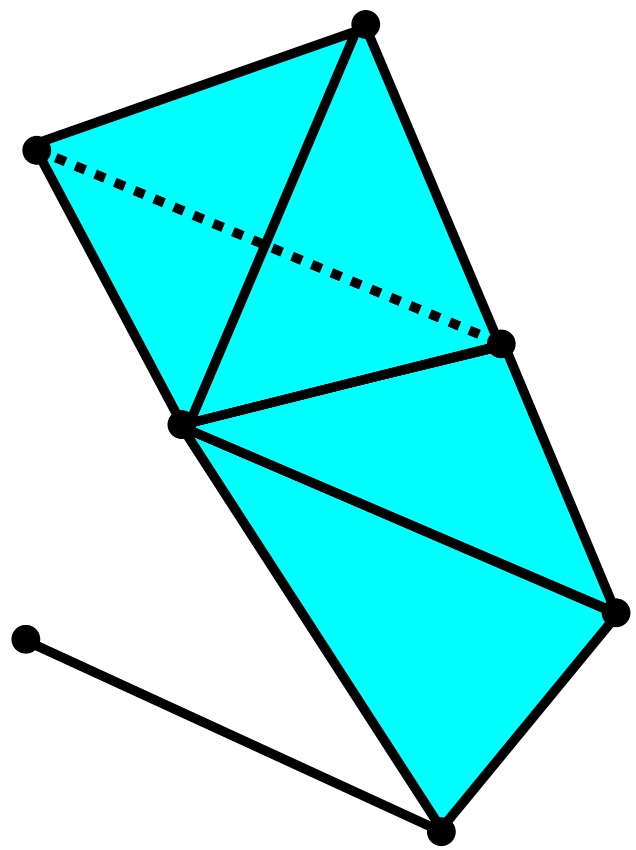
0-simplex: vertex | 1-simplex: edge | 2-simplex: triangle | 3-simplex: tetrahedron | etc...

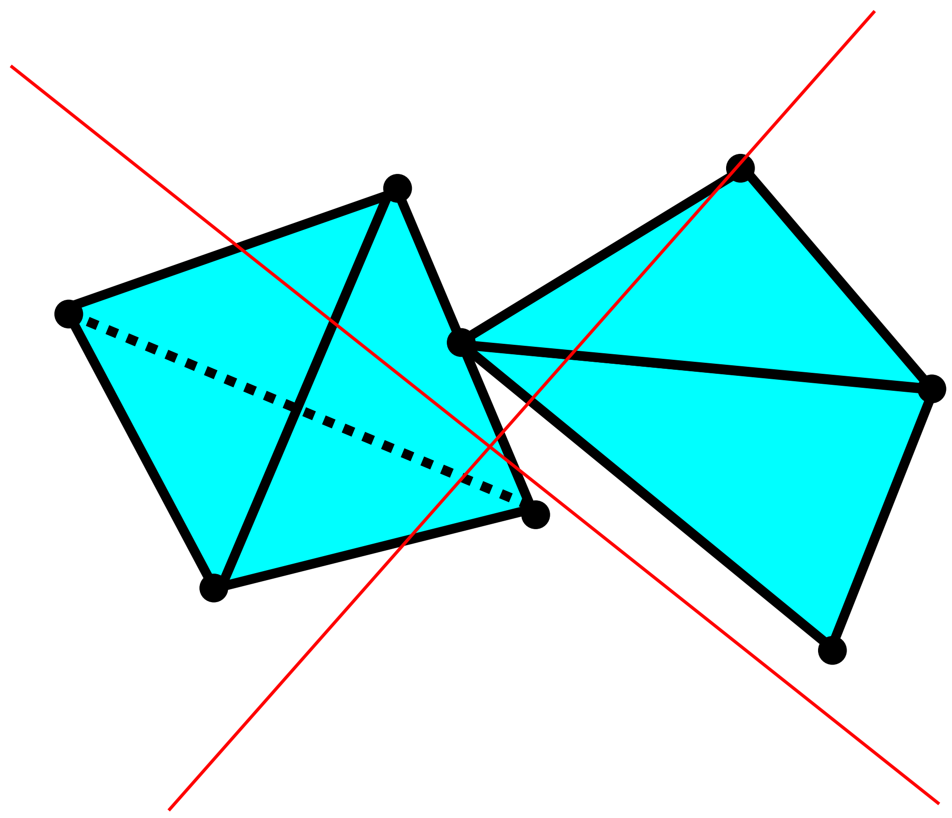
Given a set $P = \{p_0, \ldots, p_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the k-dimensional simplex σ, or k-simplex for short, spanned by P is the set of convex combinations

$$\sum_{i=0}^{k} \lambda_i \, p_i, \quad \text{with} \quad \sum_{i=0}^{k} \lambda_i = 1 \quad \text{and} \quad \lambda_i \geq 0.$$

The points $p_0, \ldots, p_k$ are called the vertices of σ.

Simplicial complexes





A (finite) simplicial complex K in $\mathbb{R}^d$ is a (finite) collection of simplices such that:

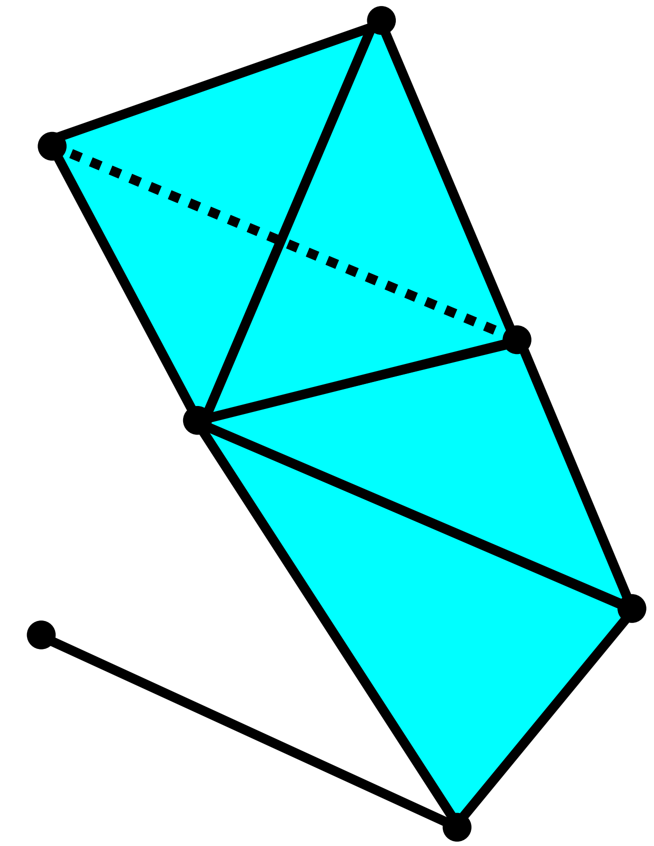
1. any face of a simplex of K is a simplex of K,
2. the intersection of any two simplices of K is either empty or a common face of both.

The underlying space of K, denoted by $|K| \subset \mathbb{R}^d$ is the union of the simplices of K.

Abstract simplicial complexes

Let P be a set. An abstract simplicial complex K with vertex set P is a set of finite subsets of P satisfying the two conditions :

1. The elements of P belong to K.
2. If $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$.



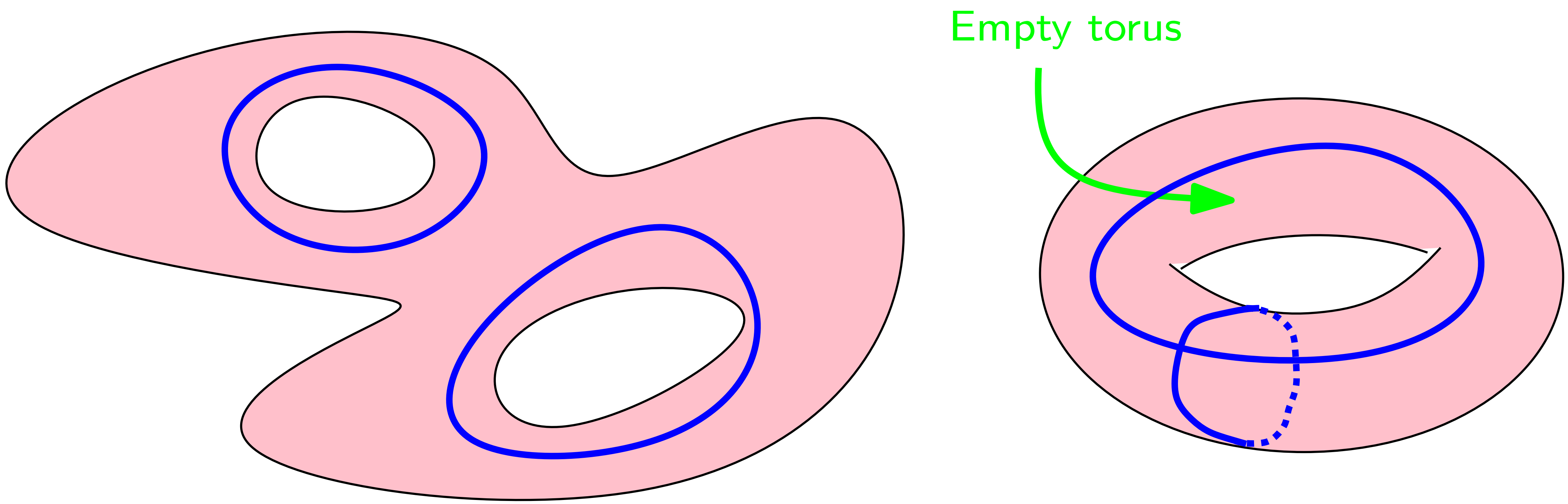
The elements of K are the simplices.

IMPORTANT

Simplicial complexes can be seen at the same time as geometric/topological spaces (good for top./geom. inference) and as combinatorial objects (abstract simplicial complexes, good for computations).

Homology in a nutshell (with coeff. in $\mathbb{Z}/2\mathbb{Z}$)

Formalize the notion of connected components, cycles/holes, voids... in a topological space.



- 2 connected components (0-dim homology)
- 4 cycles (1-dim homology)
- 1 void (2-dim homology)

Homology in a nutshell (with coeff. in $\mathbb{Z}/2\mathbb{Z}$)

The space of k-chains:

Let K be a d-dimensional simplicial complex. Let $k \in \{0, 1, \cdots, d\}$ and $\{\sigma_1, \cdots, \sigma_p\}$ be the set of k-simplices of K.

k-chain:

$$c = \sum_{i=1}^{p} \varepsilon_i \sigma_i \quad \text{with} \quad \varepsilon_i \in \mathbb{Z}/2\mathbb{Z} = \{0, 1\}$$

Sum of k-chains:

$$c + c' = \sum_{i=1}^{p} (\varepsilon_i + \varepsilon_i') \sigma_i \quad \text{and} \quad \lambda . c = \sum_{i=1}^{p} (\lambda \varepsilon_i') \sigma_i$$

where the sums $\varepsilon_i + \varepsilon_i'$ and the products $\lambda \varepsilon_i$ are modulo 2.

Homology in a nutshell (with coeff. in $\mathbb{Z}/2\mathbb{Z}$)

The boundary operator:

The boundary $\partial\sigma$ of a k-simplex σ is the sum of its $(k-1)$-faces. This is a $(k-1)$-chain.

$$If \sigma = [v_0, \cdots, v_k] \ \text{ then } \ \partial_k \sigma = \sum_{i=0}^{k} (-1)^i [v_0 \cdots \hat{v}_i \cdots v_k]$$

The boundary operator is the linear map defined by

$$\begin{array}{rccc} \partial_k : & \mathcal{C}_k(K) & \to & \mathcal{C}_{k-1}(K) \\ & c & \to & \partial_k c = \sum_{\sigma \in c} \partial_k \sigma \end{array}$$

$$\partial_k \partial_{k+1} := \partial_k \circ \partial_{k+1} = 0$$

Homology in a nutshell (with coeff. in $\mathbb{Z}/2\mathbb{Z}$)

Cycles and boundaries:

The chain complex associated to a complex K of dimension d

$$\emptyset \to \mathcal{C}_d(K) \xrightarrow{\partial} \mathcal{C}_{d-1}(K) \xrightarrow{\partial} \cdots \mathcal{C}_{k+1}(K) \xrightarrow{\partial} \mathcal{C}_k(K) \xrightarrow{\partial} \cdots \mathcal{C}_1(K) \xrightarrow{\partial} \mathcal{C}_0(K) \xrightarrow{\partial}$$

k-cycles:

$$Z_k(K) := \ker(\partial : \mathcal{C}_k \to \mathcal{C}_{k-1}) = \{c \in \mathcal{C}_k : \partial c = \emptyset\}$$

k-boundaries:

$$B_k(K) := im(\partial : \mathcal{C}_{k+1} \to \mathcal{C}_k) = \{c \in \mathcal{C}_k : \exists c' \in \mathcal{C}_{k+1}, c = \partial c'\}$$

$$B_k(K) \subset Z_k(K) \subset \mathcal{C}_k(K)$$

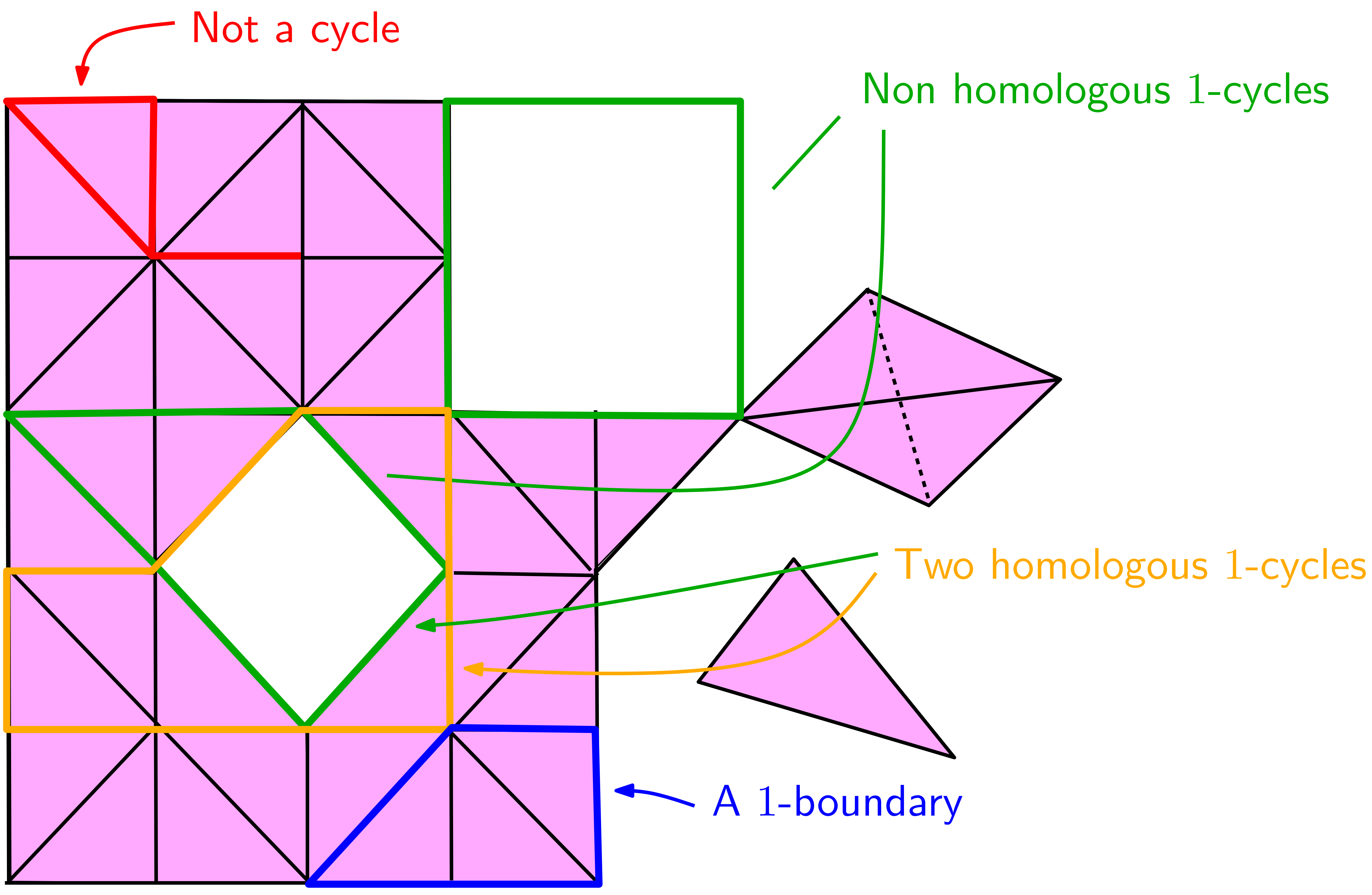
Homology in a nutshell (with coeff. in $\mathbb{Z}/2\mathbb{Z}$)

Homology groups and Betti numbers:

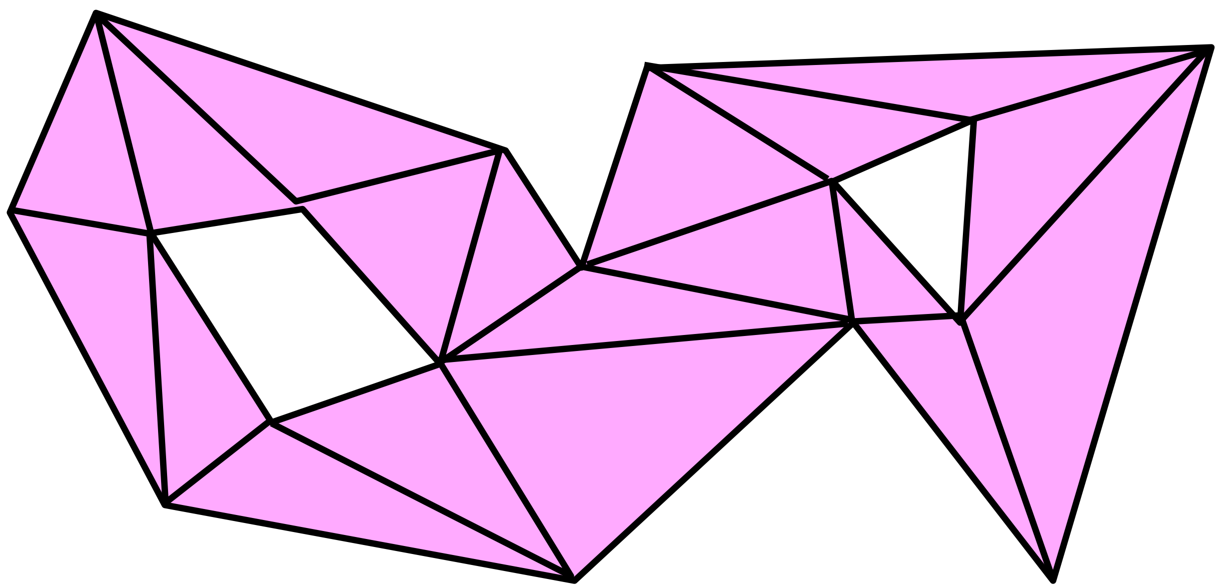
$$B_k(K) \subset Z_k(K) \subset \mathcal{C}_k(K)$$

- The k^{th} homology group of K: $H_k(K) = Z_k/B_k$
- Tout each cycle $c \in Z_k(K)$ corresponds its homology class $c+B_k(K) = \{c + b : b \in B_k(K)\}$.
- Two cycles c, c' are homologous if they are in the same homology class: $\exists b \in B_k(K)$ s. t. $b = c' - c (= c' + c)$.
- The k^{th} Betti number of K: $\beta_k(K) = \dim(H_k(K))$.

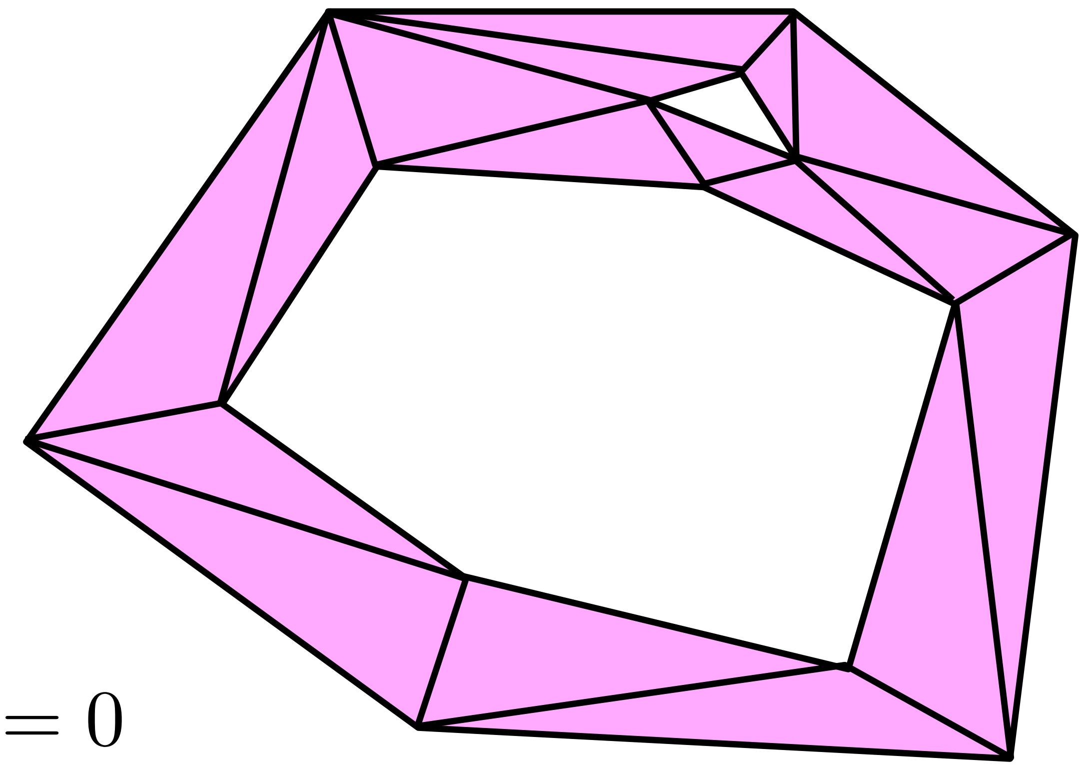
Cycles and boundaries



Topological invariance and singular homology



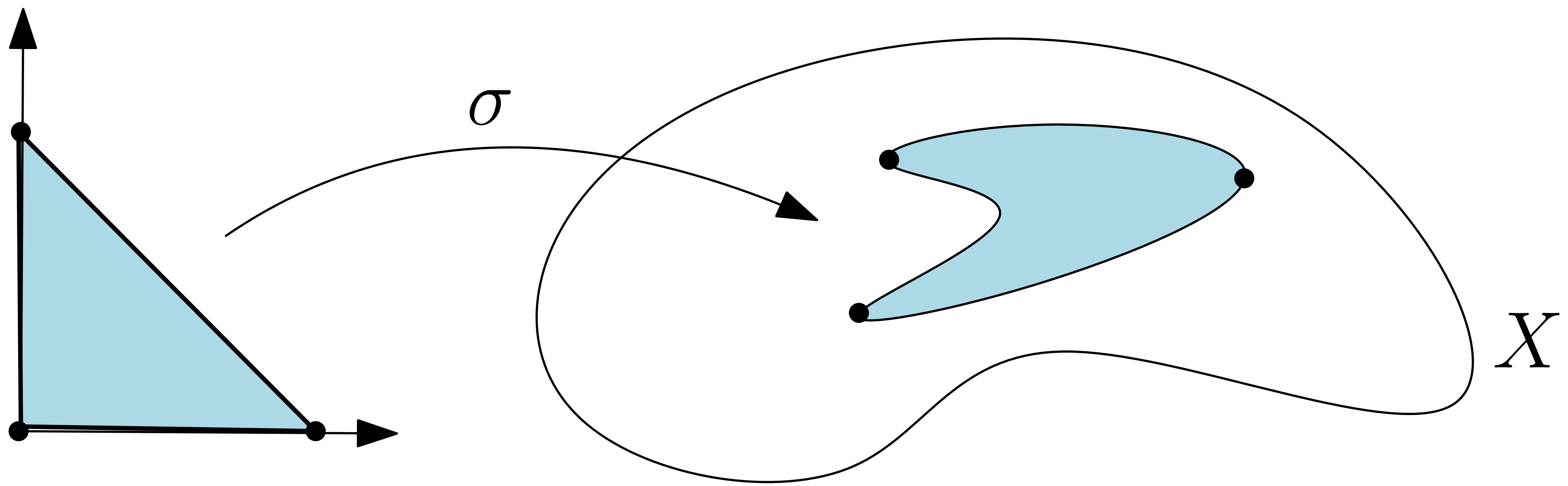
$\beta_0 = 1,\ \beta_1 = 2, \beta_2 = 0$



Theorem: If K and K' are two simplicial complexes with homeomorphic supports then their homology groups are isomorphic and their Betti numbers are equal.

- This is a classical result in algebraic topology but the proof is not obvious.
- Rely on the notion of singular homology $\rightarrow$ defined for any topological space.

Topological invariance and singular homology



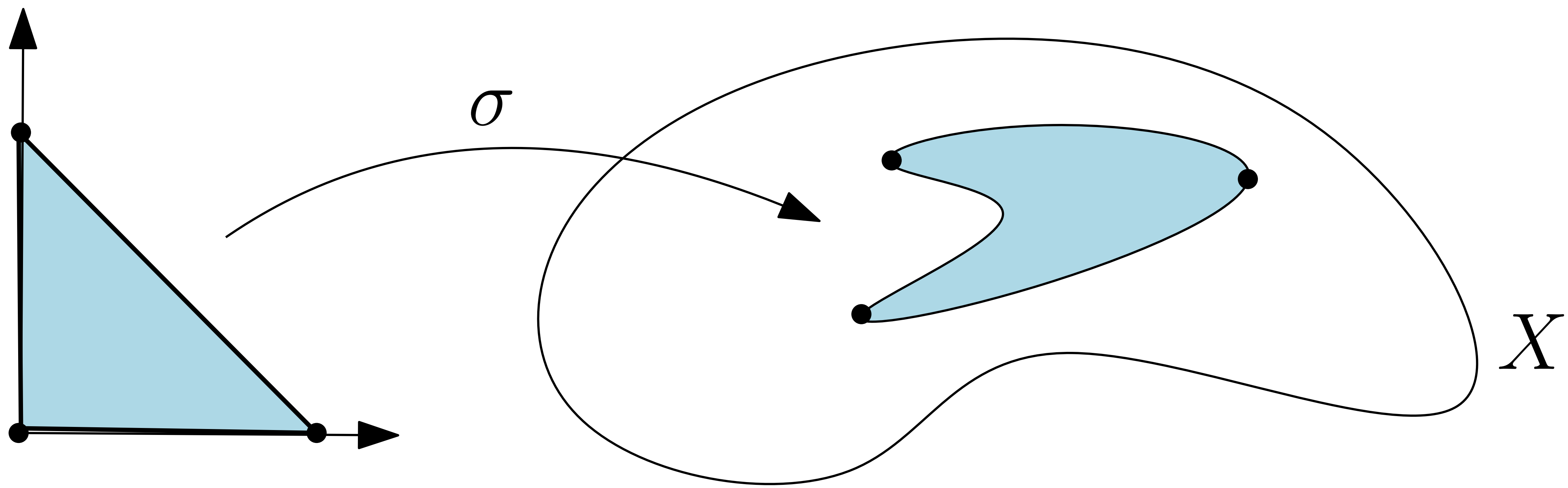
Let Δ_k be the standard simplex in $\mathbb{R}^k$. A singular k-simplex in a topological space X is a continuous map $\sigma : \Delta_k \to X$.

The same construction as for simplicial homology can be done with singular complexes $\to$ Singular homology

Important properties:

- Singular homology is defined for any topological space X.
- If X is homotopy equivalent to the support of a simplicial complex, then the singular and simplicial homology coincide!

Topological invariance and singular homology



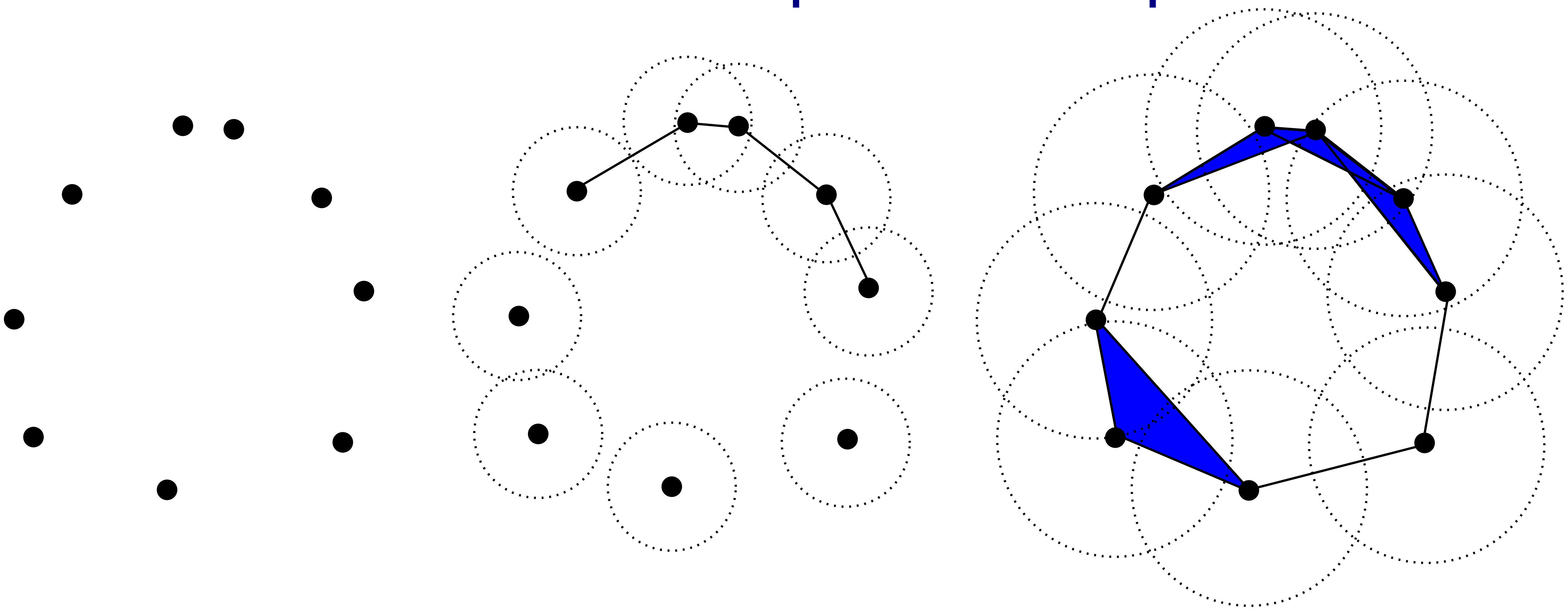
Homology and continuous maps:

- if $f : X \to Y$ is a continuous map and $\sigma : \Delta_k \to X$ a simplex in X, then $f \circ \sigma : \Delta_k \to Y$ is a simplex in $Y \Rightarrow f$ induces a linear maps between homology groups:

$$f_\sharp : H_k(X) \to H_k(Y)$$

- if $f : X \to Y$ is an homeomorphism or an homotopy equivalence then $f_\sharp$ is an isomorphism.

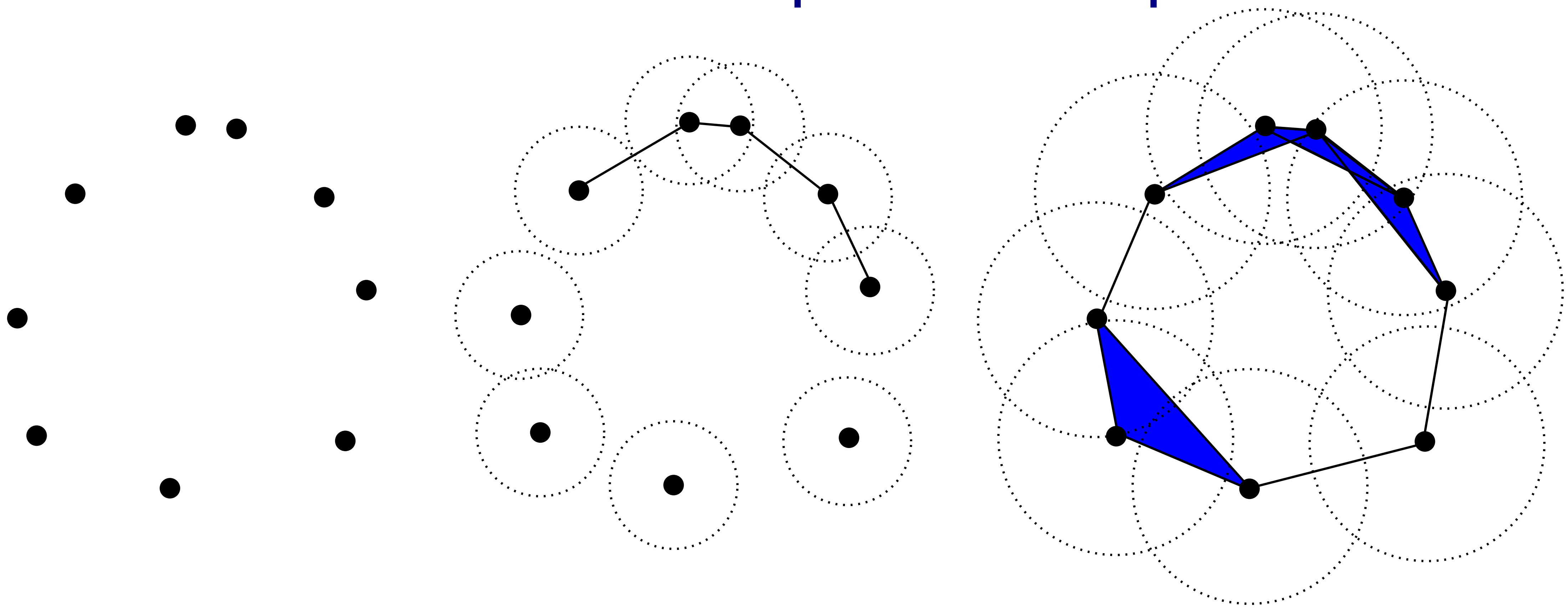
Filtrations of simplicial complexes



- A filtered simplicial complex (or a filtration) $\mathbb{S}$ built on top of a set $\mathbb{X}$ is a family $(\mathbb{S}_a \mid a \in \mathbf{T})$, $\mathbf{T} \subseteq \mathbb{R}$, of subcomplexes of some fixed simplicial complex $\mathbb{S}$ with vertex set X s. t. $\mathbb{S}_a \subseteq \mathbb{S}_b$ for any $a \leq b$.
- More generaly, filtration $=$ nested family of topological spaces indexed by $\mathbf{T}$.

Persistent homology of a filtered simplicial complexe encodes the evolution of the homology of the subcomplexes.

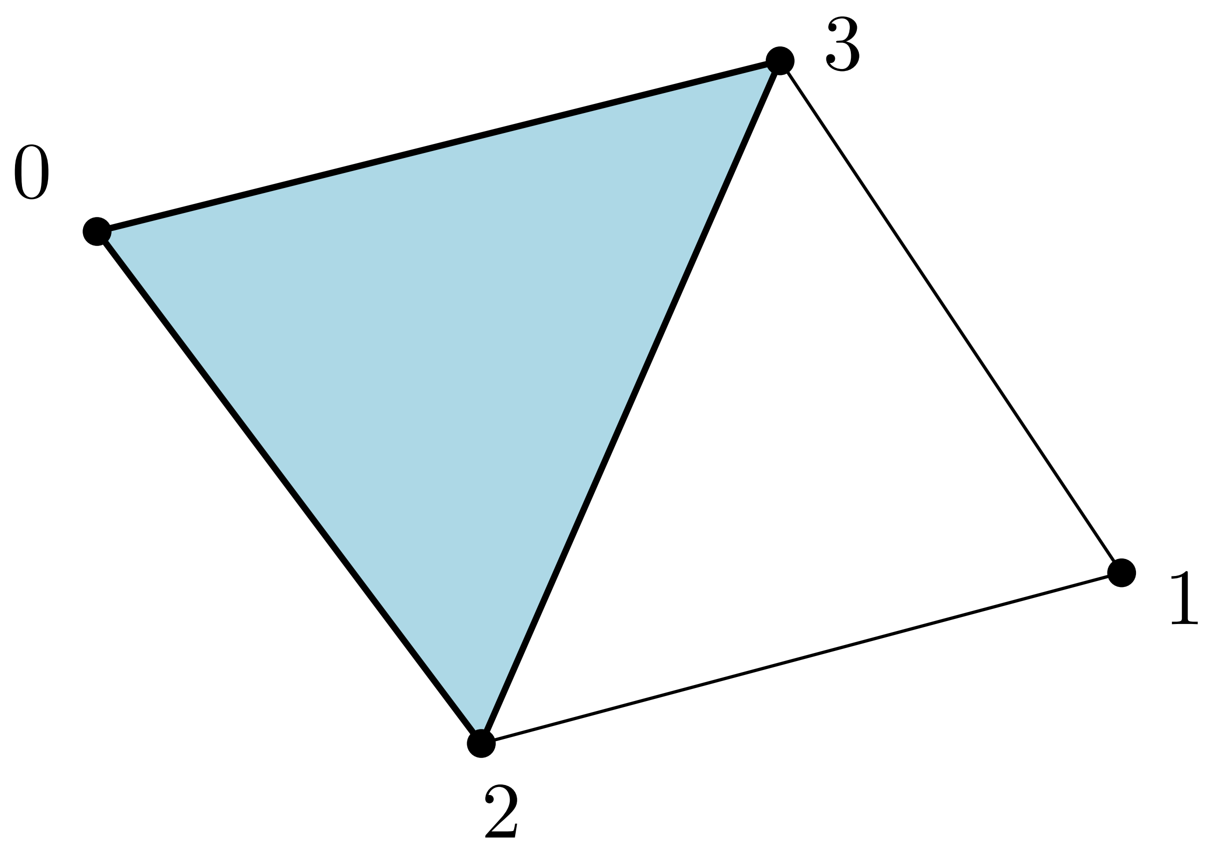
Filtrations of simplicial complexes



- A filtered simplicial complex (or a filtration) $\mathbb{S}$ built on top of a set $\mathbb{X}$ is a family $(\mathbb{S}_a \mid a \in \mathbf{T})$, $\mathbf{T} \subseteq \mathbb{R}$, of subcomplexes of some fixed simplicial complex $\mathbb{S}$ with vertex set X s. t. $\mathbb{S}_a \subseteq \mathbb{S}_b$ for any $a \leq b$.
- More generaly, filtration $=$ nested family of topological spaces indexed by $\mathbf{T}$.

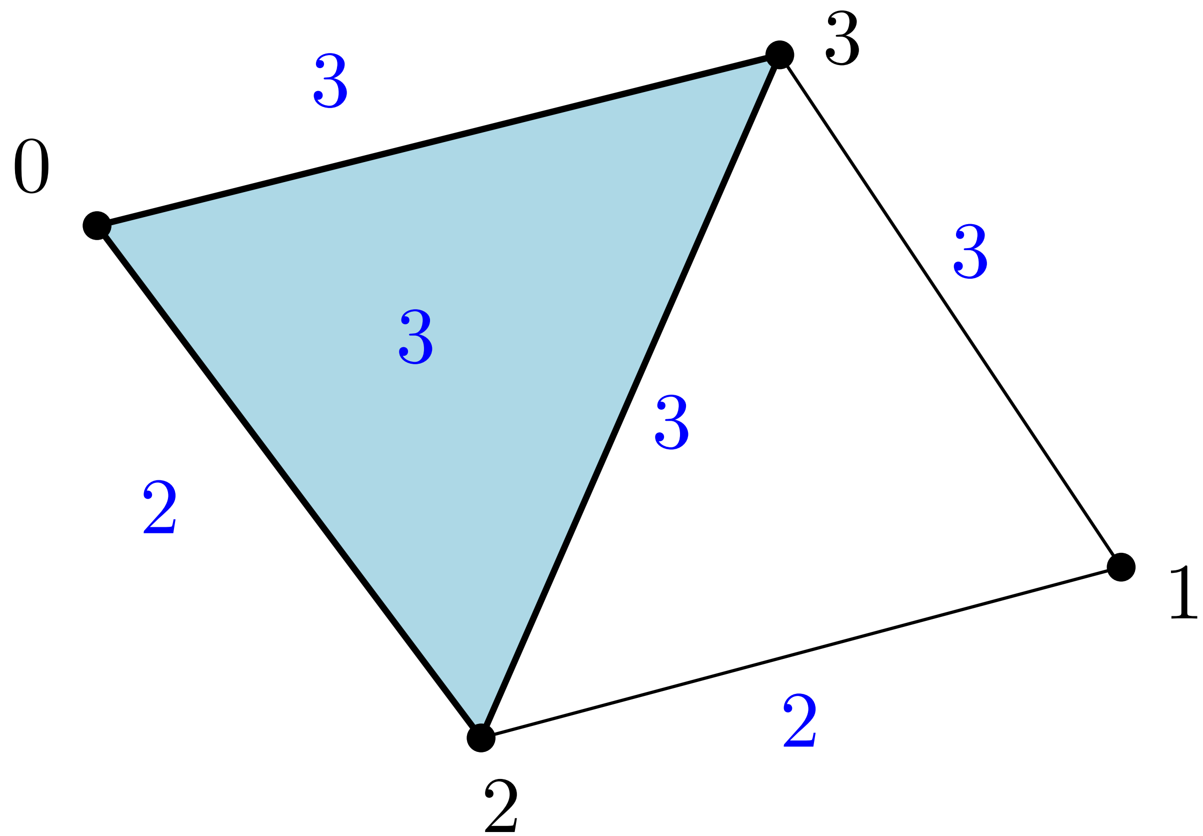
Many examples and ways to design filtrations depending on the application and targeted objectives : sublevel and upperlevel sets, Čech complex,...

Sublevel set filtration associated to a function



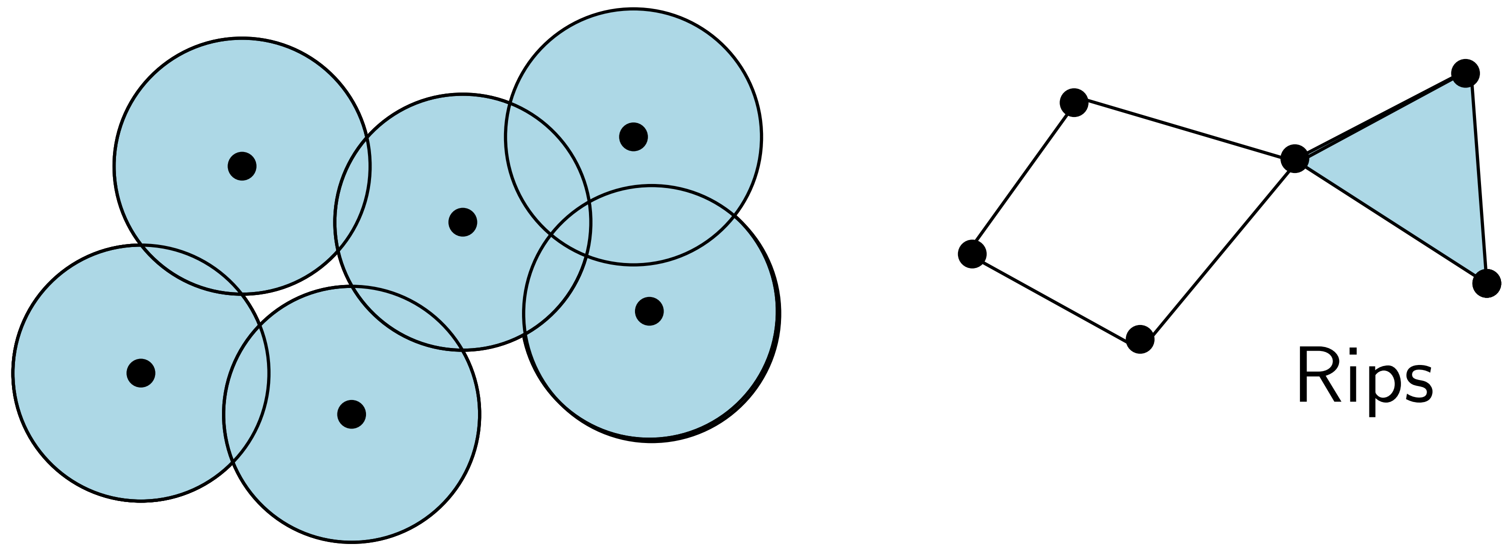
- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

Sublevel set filtration associated to a function



- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

The Vietoris-Rips filtration



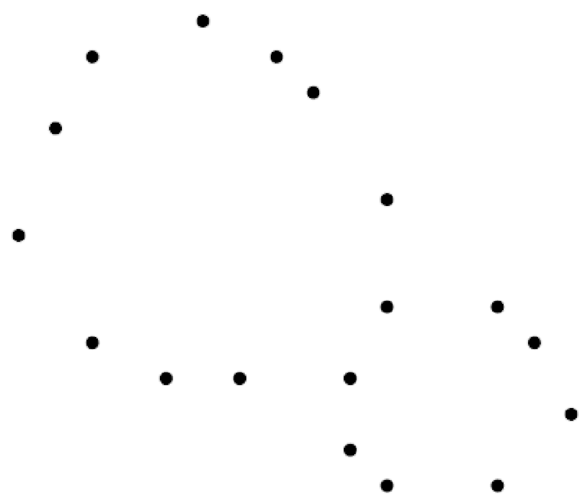
Let V be a point cloud (in a metric space (X, d)).

The Vietoris-Rips complex $\mathrm{Rips}(V)$ is the filtered simplicial complex indexed by $\mathbb{R}$ whose vertex set is V and defined by:

$$\sigma = [p_0 p_1 \cdots p_k] \in \mathrm{Rips}(V, \alpha) \text{ iff } \forall i, j \in \{0, \cdots, k\},\ d(p_i, p_j) \leq \alpha$$

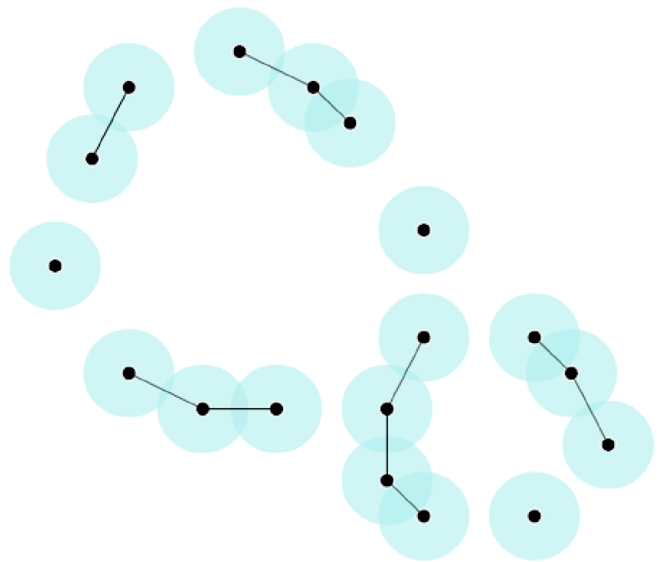
Easy to compute and fully determined by its 1-skeleton

Persistent homology for point cloud data



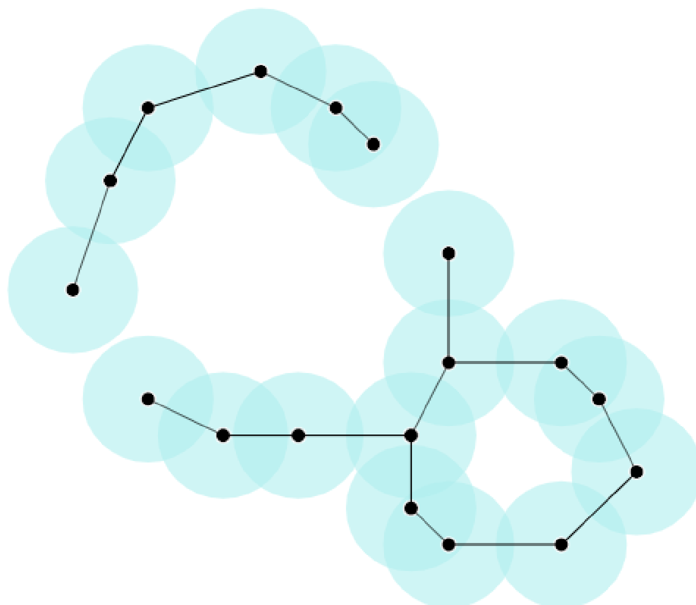
- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

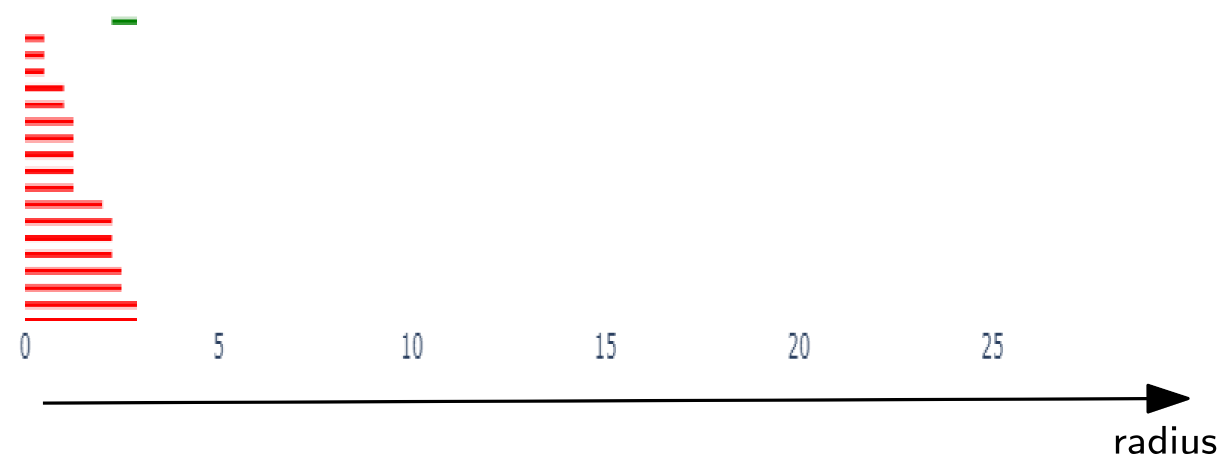
Persistent homology for point cloud data



- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

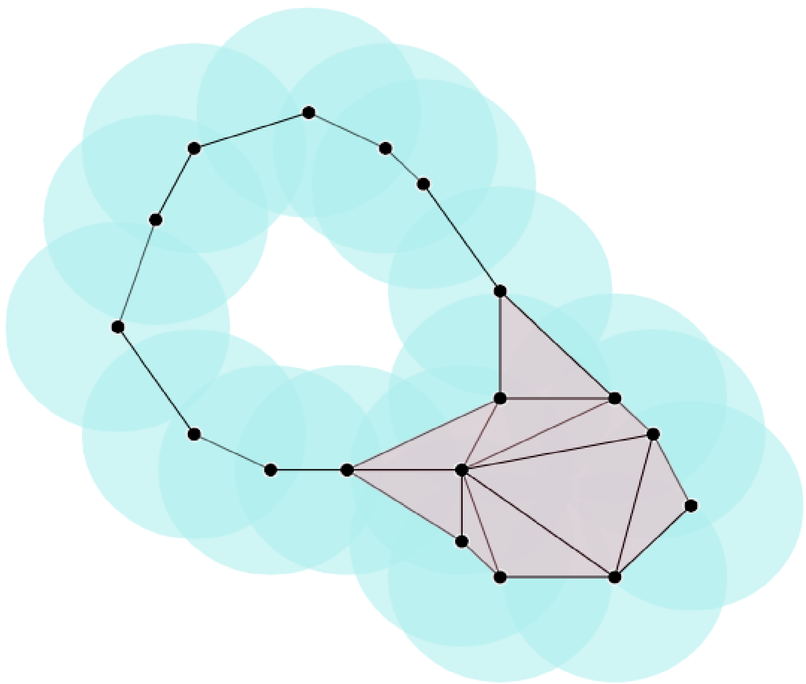
Persistent homology for point cloud data

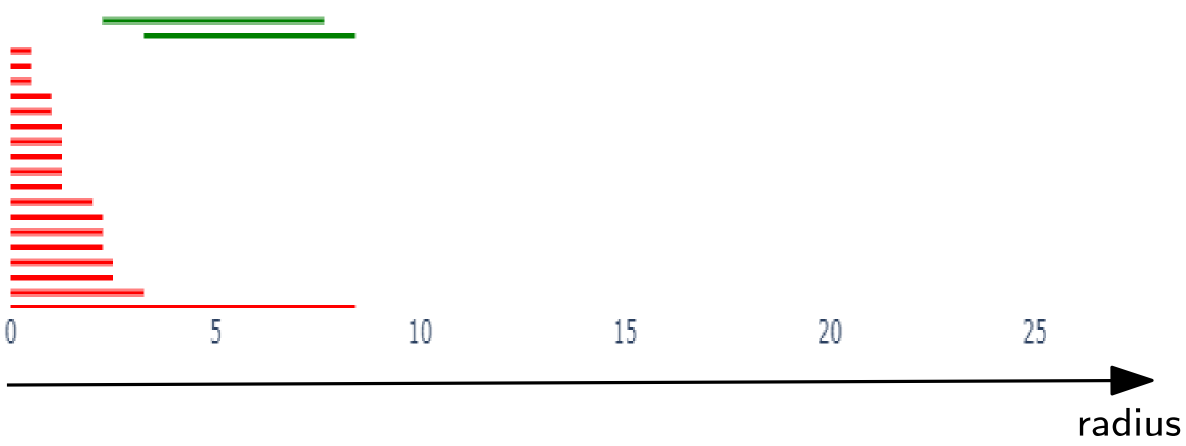




- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

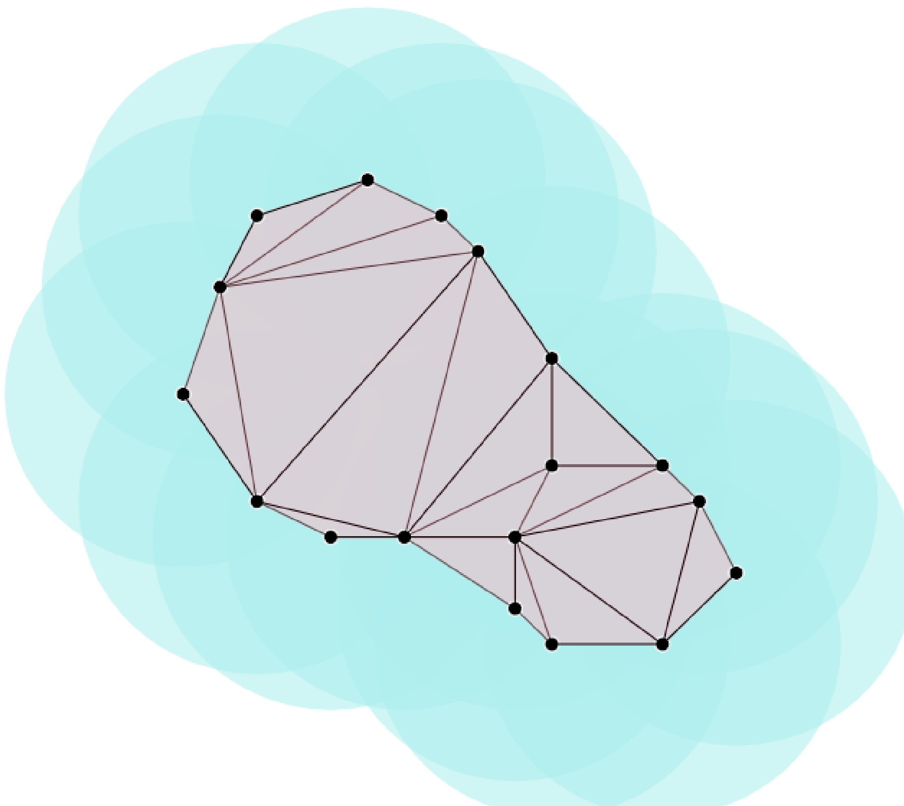
Persistent homology for point cloud data

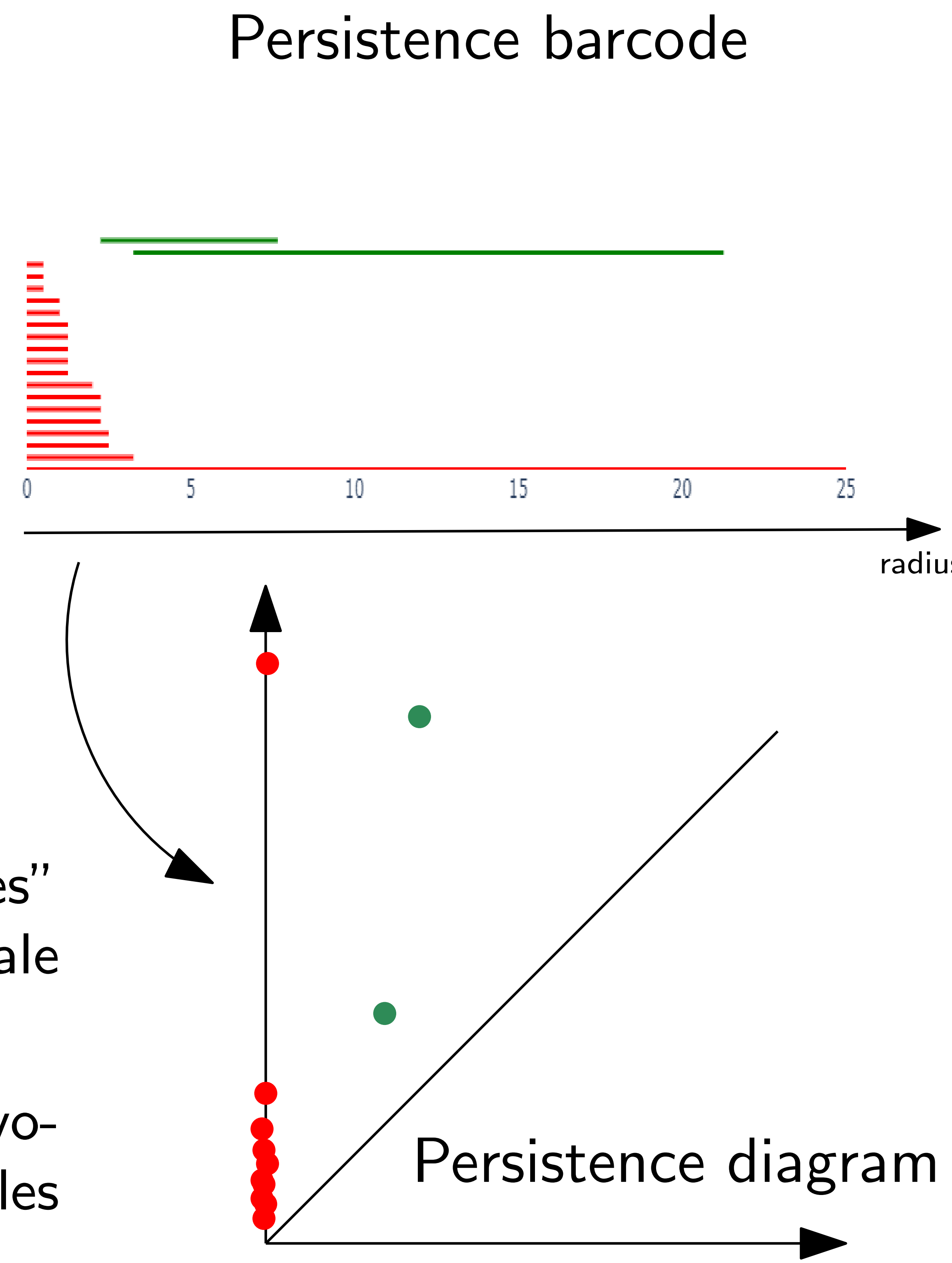




- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

Persistent homology for point cloud data





- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

Stability properties

"Stability theorem": Close spaces/data sets have close persistence diagrams!

[C., de Silva, Oudot - Geom. Dedicata 2013].

If $\mathbb{X}, \mathbb{Y}$ are compact metric spaces, then

$$\mathrm{d_b}(\mathsf{dgm}(\mathrm{Rips}(\mathbb{X})), \mathsf{dgm}(\mathrm{Rips}(\mathbb{Y}))) \leq 2 d_{GH}(\mathbb{X}, \mathbb{Y}).$$

Bottleneck distance

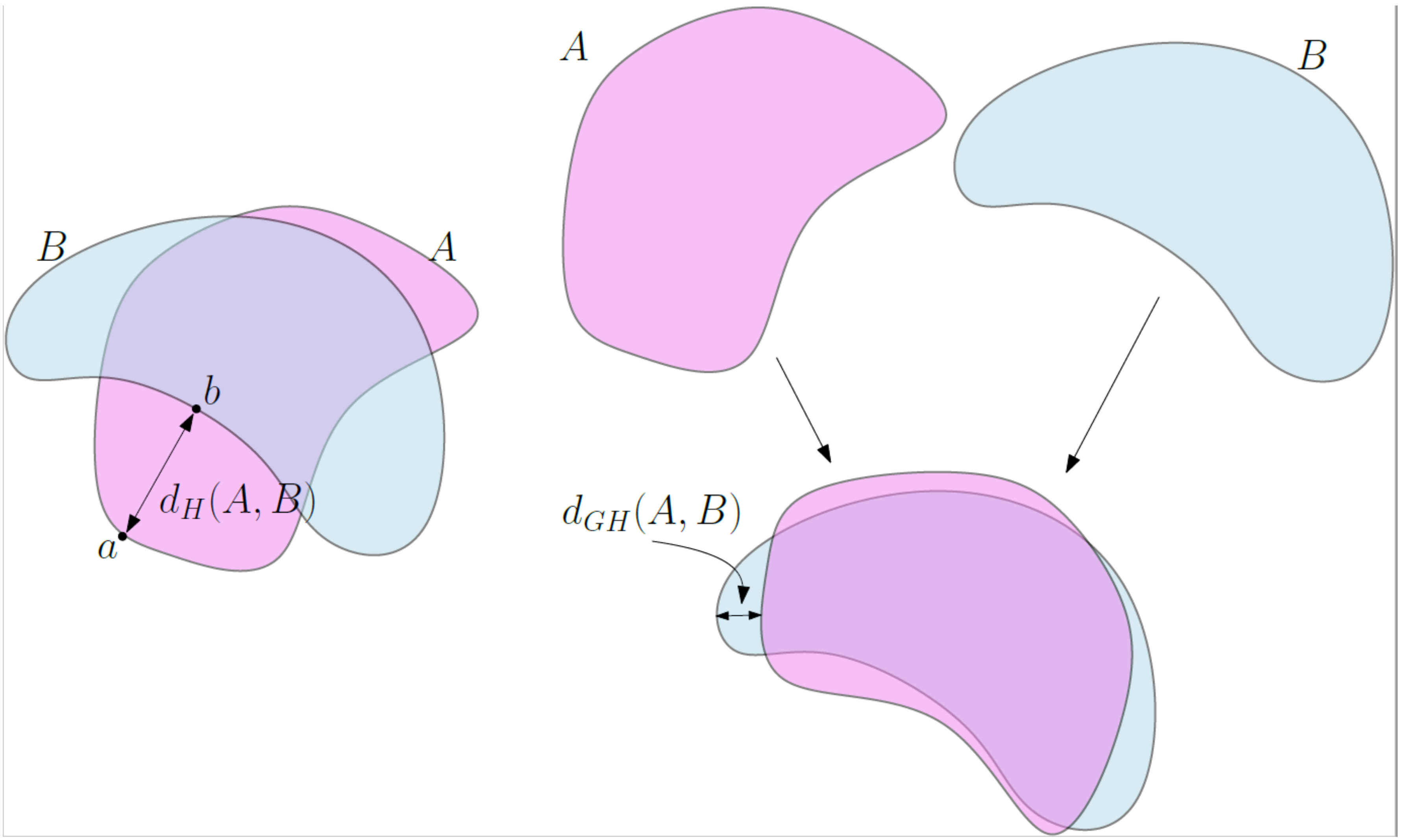
Gromov-Hausdorff distance

$$d_{GH}(\mathbb{X}, \mathbb{Y}) := \inf_{\mathbb{Z}, \gamma_1, \gamma_2} d_H(\gamma_1(\mathbb{X}), \gamma_2(\mathbb{X}))$$

$\mathbb{Z}$ metric space, $\gamma_1 : \mathbb{X} \to \mathbb{Z}$ and $\gamma_2 : \mathbb{Y} \to \mathbb{Z}$ isometric embeddings.

Rem: This result also holds for other families of filtrations (particular case of a more general thm).

Hausdorff distance



Let $A, B \subset M$ be two compact subsets of a metric space (M, d)

$$d_H(A,B) \quad = \quad \max\{\sup_{b \in B} d(b,A), \sup_{a \in A} d(a,B)\}$$

where $d(b, A) = \sup_{a \in A} d(b, a)$.

Persistent homology of filtered simplicial complexes

Let $\mathbb{S} = (\mathbb{S}_a \mid a \in \mathbf{R})$ be a finite filtered simplicial complex with N simplicices and let $\mathbb{S}_{a_1} \subset \mathbb{S}_{a_2} \subset \cdots \subset \mathbb{S}_{a_N}$ be the discrete filtration induced by the entering times of the simplices: $\mathbb{S}_{a_i} \setminus \mathbb{S}_{a_{i-1}} = \sigma_{a_i}$.

Persistent homology of filtered simplicial complexes

Let $\mathbb{S} = (\mathbb{S}_a \mid a \in \mathbf{R})$ be a finite filtered simplicial complex with N simplicices and let $\mathbb{S}_{a_1} \subset \mathbb{S}_{a_2} \subset \cdots \subset \mathbb{S}_{a_N}$ be the discrete filtration induced by the entering times of the simplices: $\mathbb{S}_{a_i} \setminus \mathbb{S}_{a_{i-1}} = \sigma_{a_i}$.

Process the simplices according to their order of entrance in the filtration:

Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

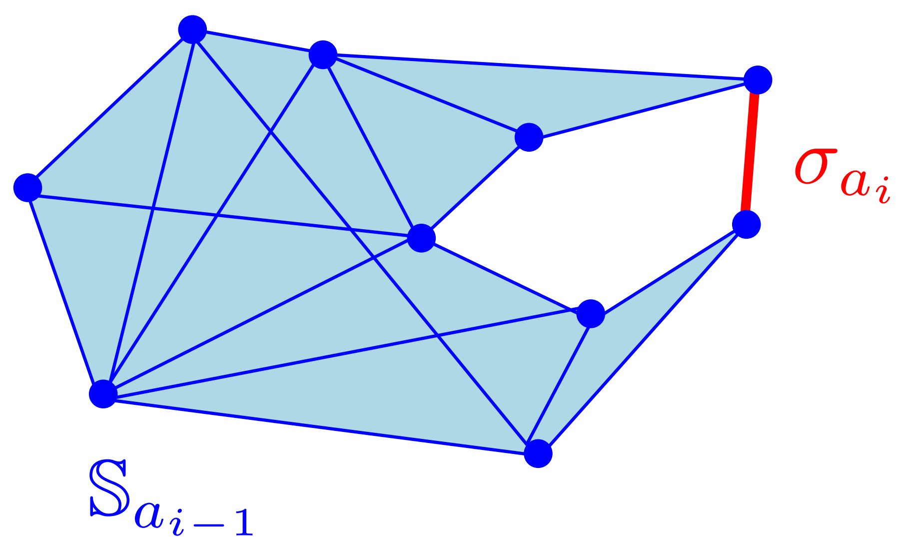
Persistent homology of filtered simplicial complexes

Let $\mathbb{S} = (\mathbb{S}_a \mid a \in \mathbf{R})$ be a finite filtered simplicial complex with N simplicices and let $\mathbb{S}_{a_1} \subset \mathbb{S}_{a_2} \subset \cdots \subset \mathbb{S}_{a_N}$ be the discrete filtration induced by the entering times of the simplices: $\mathbb{S}_{a_i} \setminus \mathbb{S}_{a_{i-1}} = \sigma_{a_i}$.

Process the simplices according to their order of entrance in the filtration:

Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

Case 1: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ creates a new k-dimensional topological feature in $\mathbb{S}_{a_i}$ (new homology class in H_k).



$\Rightarrow$ the birth of a k-dim feature is registered.

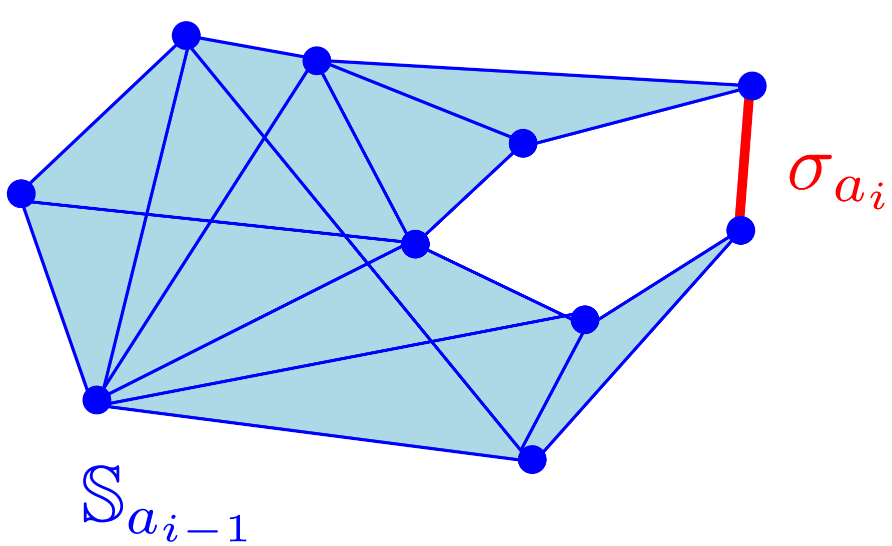
Persistent homology of filtered simplicial complexes

Let $\mathbb{S} = (\mathbb{S}_a \mid a \in \mathbf{R})$ be a finite filtered simplicial complex with N simplicices and let $\mathbb{S}_{a_1} \subset \mathbb{S}_{a_2} \subset \cdots \subset \mathbb{S}_{a_N}$ be the discrete filtration induced by the entering times of the simplices: $\mathbb{S}_{a_i} \setminus \mathbb{S}_{a_{i-1}} = \sigma_{a_i}$.

Process the simplices according to their order of entrance in the filtration:

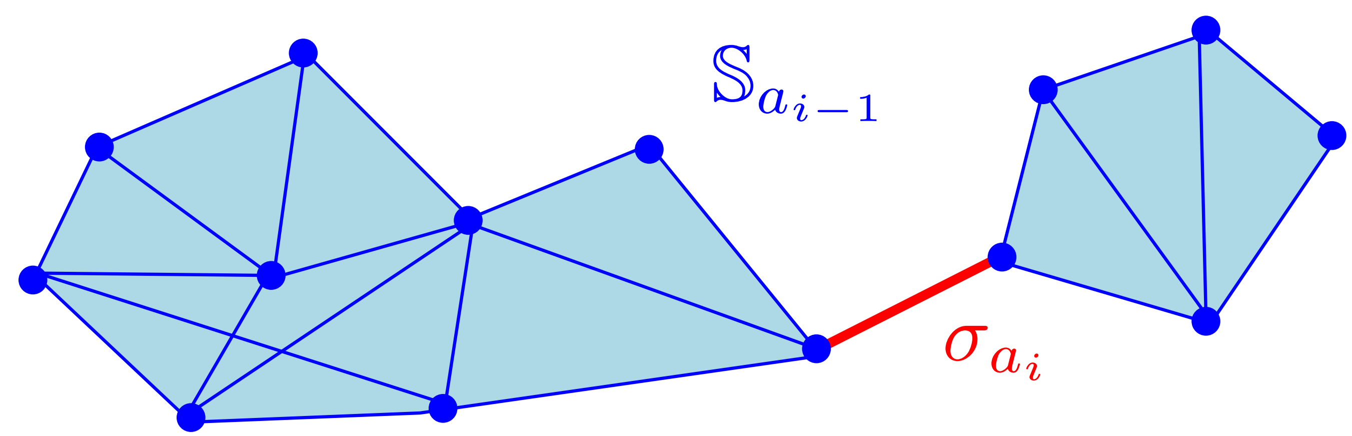
Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

Case 1: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ creates a new k-dimensional topological feature in $\mathbb{S}_{a_i}$ (new homology class in H_k).



$\Rightarrow$ the birth of a k-dim feature is registered.

Case 2: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ kills a $(k-1)$-dimensional topological feature in $\mathbb{S}_{a_i}$ (homology class in H_{k-1}).



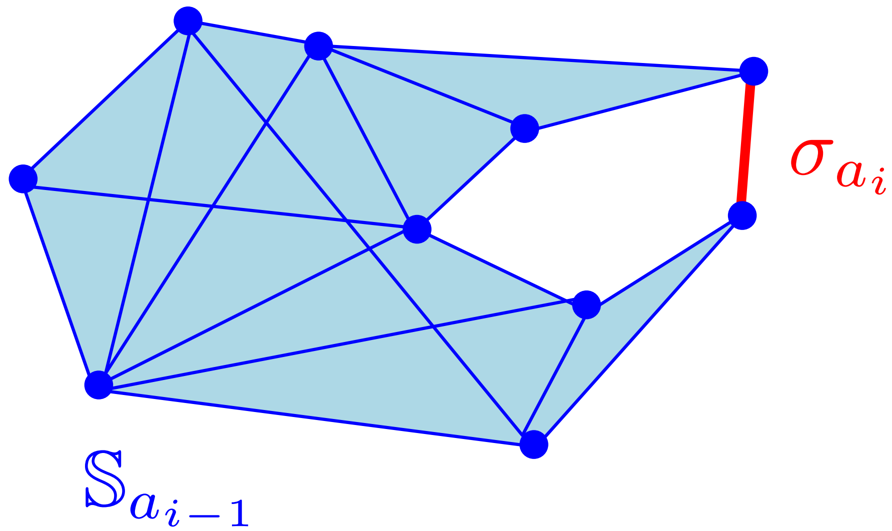
$\Rightarrow$ persistence algo. pairs the simplex σ_{a_i} to the simplex σ_{a_j} that gave birth to the killed feature.

Persistent homology of filtered simplicial complexes

Process the simplices according to their order of entrance in the filtration:

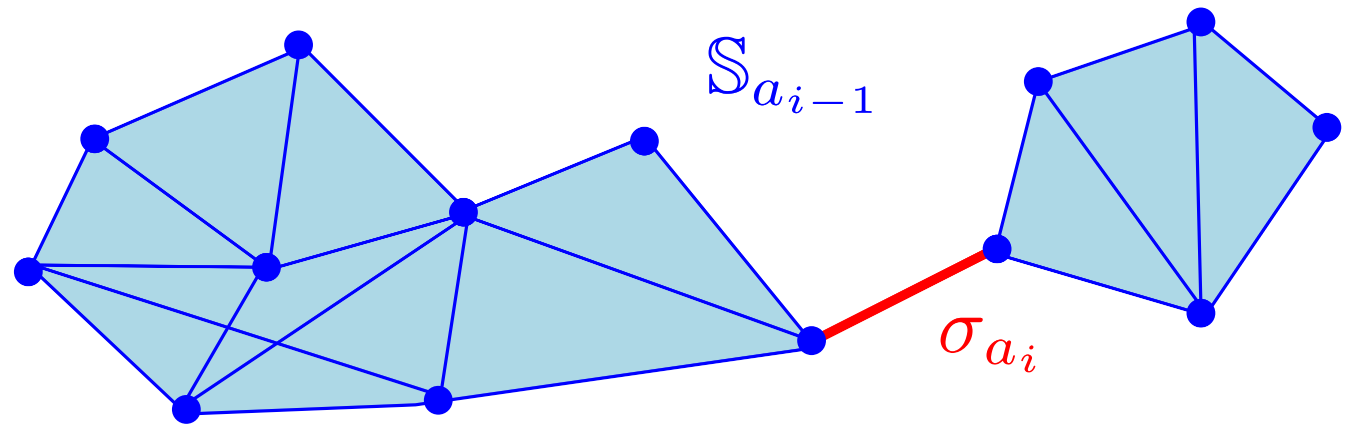
Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

Case 1: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ creates a new k-dimensional topological feature in $\mathbb{S}_{a_i}$ (new homology class in H_k).



$\Rightarrow$ the birth of a k-dim feature is registered.

Case 2: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ kills a $(k-1)$-dimensional topological feature in $\mathbb{S}_{a_i}$ (homology class in H_{k-1}).



$\Rightarrow$ persistence algo. pairs the simplex σ_{a_i} to the simplex σ_{a_j} that gave birth to the killed feature.

$\rightarrow$ $(\sigma_{a_j}, \sigma_{a_i})$: persistence pair

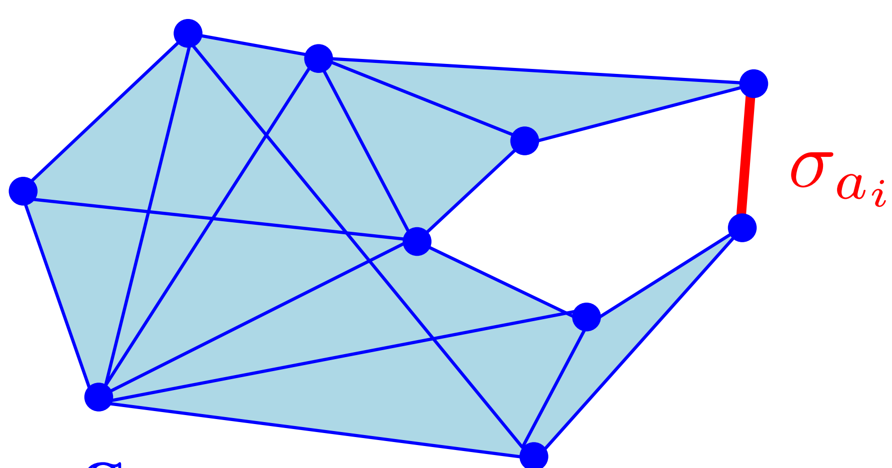
$\rightarrow$ $(a_j, a_i) \in \mathbb{R}^2$: point in the persistence diagram

Persistent homology of filtered simplicial complexes

Process the simplices according to their order of entrance in the filtration:

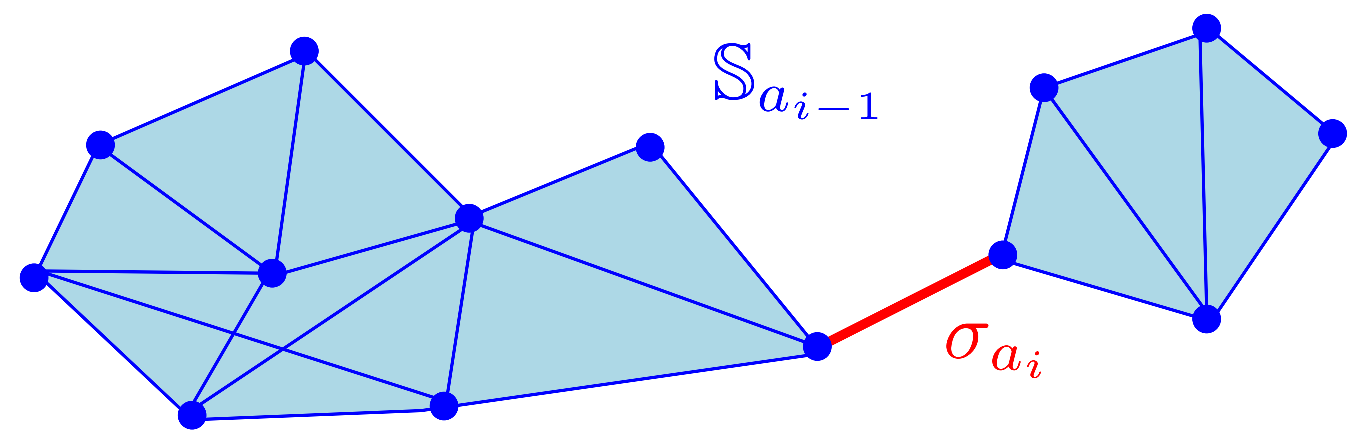
Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

Case 1: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ creates a new k-dimensional topological feature in $\mathbb{S}_{a_i}$ (new homology class in H_k).



$\Rightarrow$ the birth of a k-dim feature is registered.

Case 2: adding σ_{a_i} to $\mathbb{S}_{a_{i-1}}$ kills a $(k-1)$-dimensional topological feature in $\mathbb{S}_{a_i}$ (homology class in H_{k-1}).



$\Rightarrow$ persistence algo. pairs the simplex σ_{a_i} to the simplex σ_{a_j} that gave birth to the killed feature.

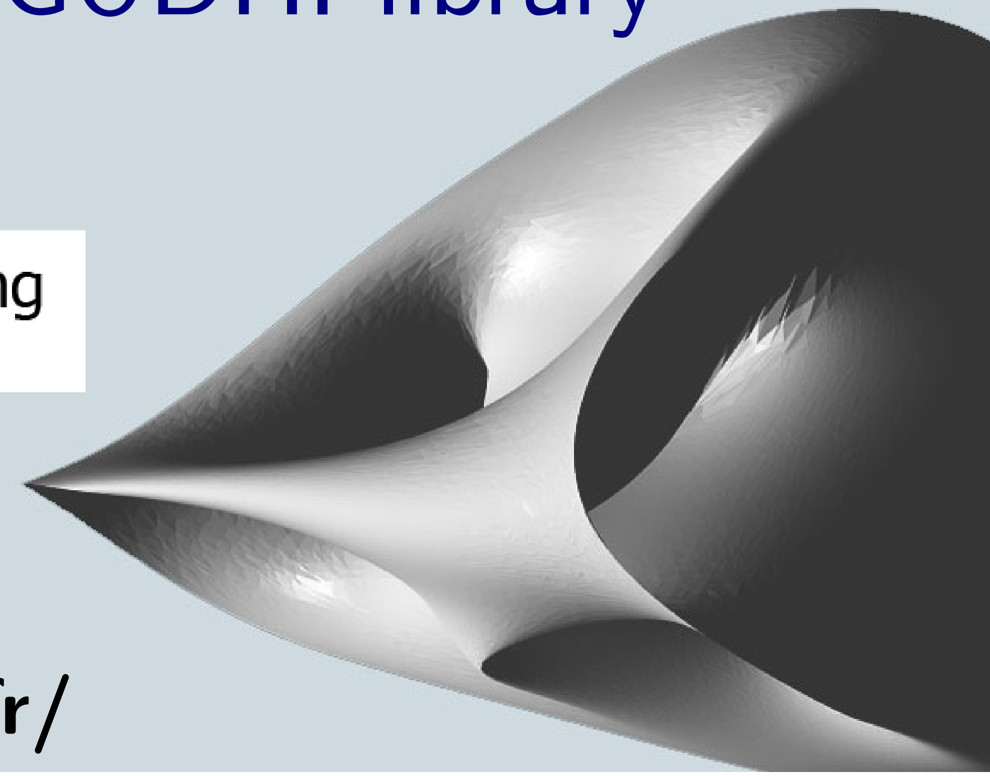
$\rightarrow$ $(\sigma_{a_j}, \sigma_{a_i})$: persistence pair

$\rightarrow$ $(a_j, a_i) \in \mathbb{R}^2$: point in the persistence diagram

Important to remember: the persistence pairs are determined by the order on the simplices; the corresponding points in the diagrams are determined by the indices.

Persistent homology with the GUDHI library

GUDHI Geometry Understanding in Higher Dimensions



http://gudhi.gforge.inria.fr/

GUDHI :

- a C++/Python open source software library for TDA,
- a developers team, an editorial board, open to external contributions,
- provides state-of-the-art TDA data structures and algorithms : design of filtrations, computation of pre-defined filtrations, persistence diagrams,...
- part of GUDHI is interfaced to R through the TDA package.

Algebraic perspective

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Examples: homology of filtrations gives rise to persistence modules.

- Let $\mathbb{S}$ be a filtered simplicial complex. If $V_a = \mathrm{H}(\mathbb{S}_a)$ and $v_a^b : \mathrm{H}(\mathbb{S}_a) \to \mathrm{H}(\mathbb{S}_b)$ is the linear map induced by the inclusion $\mathbb{S}_a \hookrightarrow \mathbb{S}_b$ then $(\mathrm{H}(\mathbb{S}_a) \mid a \in \mathbf{R})$ is a persistence module.
- Given a metric space $(\mathbb{X}, d_{\mathbb{X}})$, $\mathrm{H}(\mathrm{Rips}(\mathbb{X}))$ is a persistence module.
- If $f : X \to \mathbf{R}$ is a function, then the filtration defined by the sublevel sets of f, $\mathbb{F}_a = f^{-1}((-\infty, a])$, induces a persistence module at homology level.

Algebraic perspective

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Definition: A persistence module $\mathbb{V}$ is q-tame if for any $a < b$, v_a^b has a finite rank.

Theorem: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG'09], [C., de Silva, Glisse, Oudot 12]

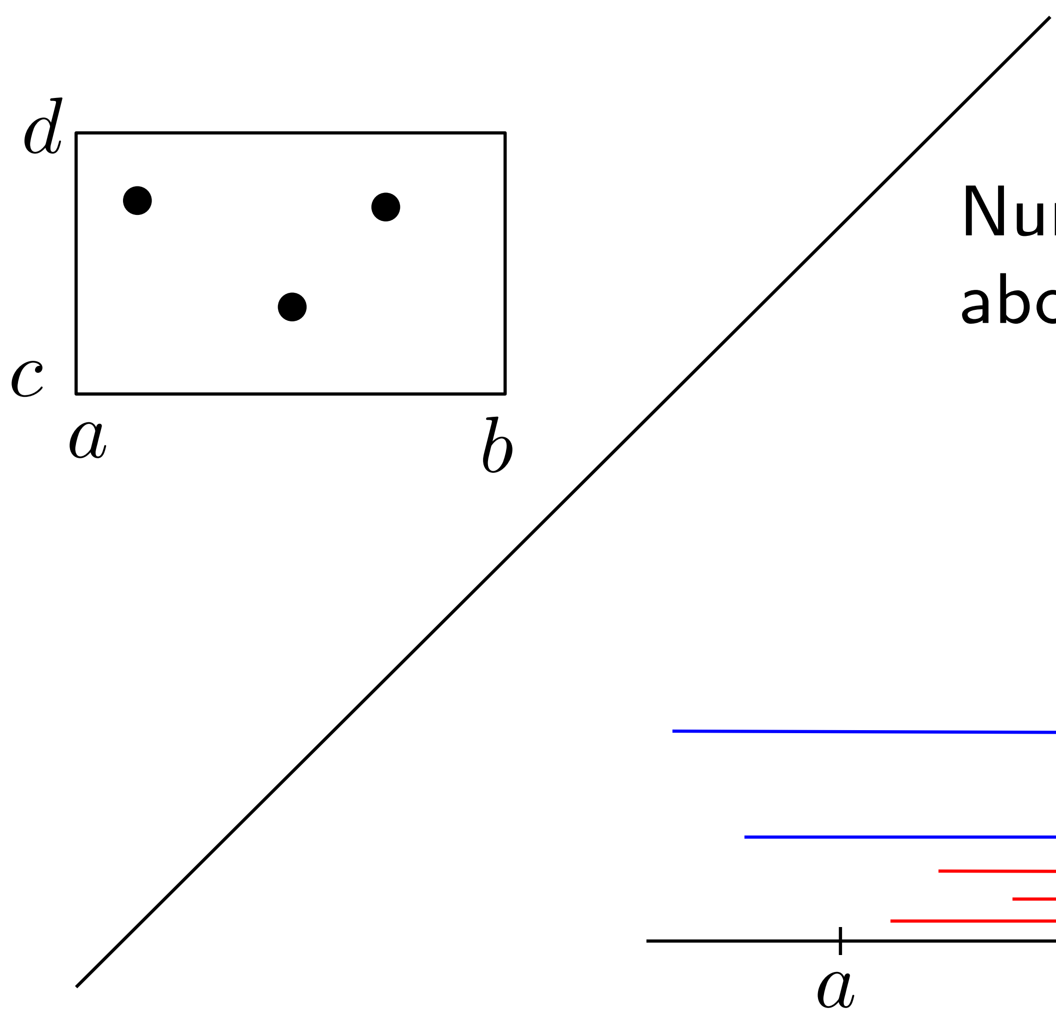
q-tame persistence modules have well-defined persistence diagrams.

Example: Let $\mathbb{X}$ be a compact metric space. Then $\mathrm{H}(\mathrm{Rips}(\mathbb{X}))$ is q-tame.

Algebraic perspective

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

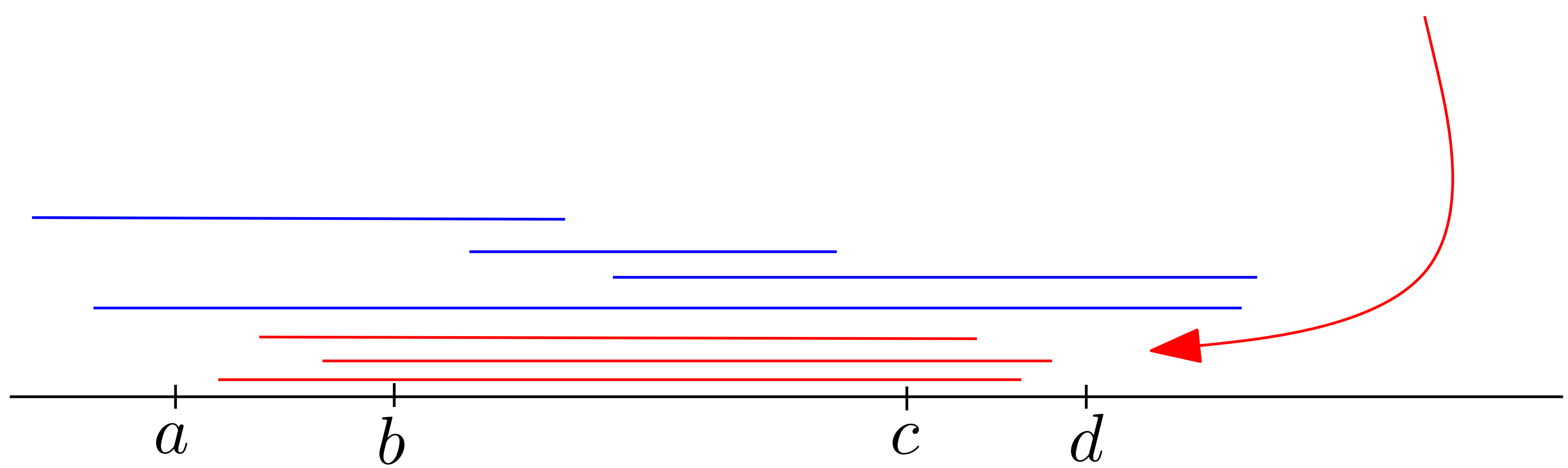
An idea about the definition of persistence diagrams:



Measures on rectangles:

Number of points in any rectangle $[a, b] \times [c, d]$ above the diagonal:

$$rk(v_b^c) - rk(v_b^d) + rk(v_a^d) - rk(v_a^c)$$



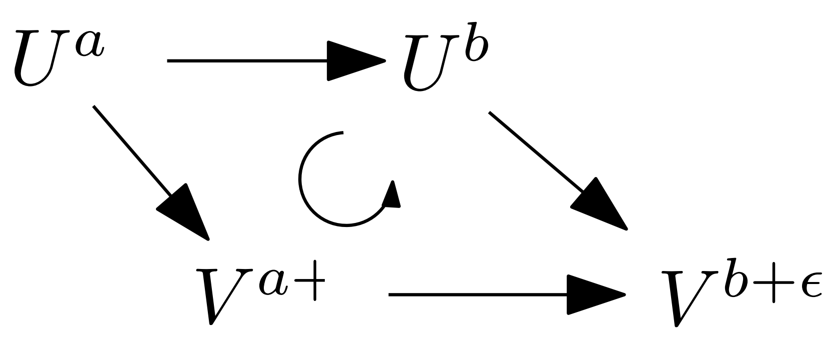
Algebraic perspective

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \le b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \le b \le c$, and where v_a^a is the identity map on V_a.

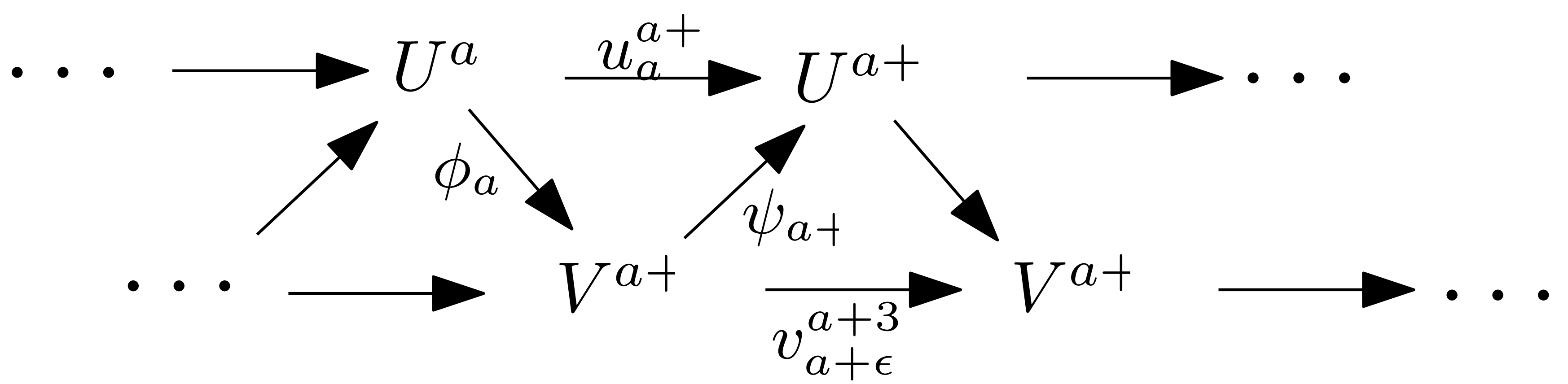
A homomorphism of degree ϵ between two persistence modules $\mathbb{U}$ and $\mathbb{V}$ is a collection Φ of linear maps

$$(\phi_a : U_a \to V_{a+\epsilon} \mid a \in \mathbf{R})$$

such that $v_{a+\epsilon}^{b+\epsilon} \circ \phi_a = \phi_b \circ u_a^b$ for all $a \le b$.



An ε-interleaving between $\mathbb{U}$ and $\mathbb{V}$ is specified by two homomorphisms of degree ϵ $\Phi : \mathbb{U} \to \mathbb{V}$ and $\Psi : \mathbb{V} \to \mathbb{U}$ s.t. $\Phi \circ \Psi$ and $\Psi \circ \Phi$ are the "shifts" of degree 2ϵ between $\mathbb{U}$ and $\mathbb{V}$.



Algebraic perspective

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Stability Thm: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG '09], [C., de Silva, Glisse, Oudot 12]

If $\mathbb{U}$ and $\mathbb{V}$ are q-tame and ϵ-interleaved for some $\epsilon \geq 0$ then

$$d_B(\mathrm{dgm}(\mathbb{U}), \mathrm{dgm}(\mathbb{V})) \leq \epsilon$$

Algebraic perspective

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Stability Thm: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG '09], [C., de Silva, Glisse, Oudot 12]

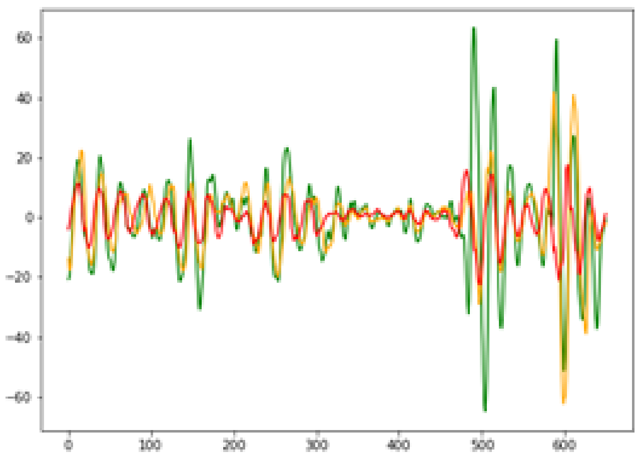
If $\mathbb{U}$ and $\mathbb{V}$ are q-tame and ϵ-interleaved for some $\epsilon \geq 0$ then

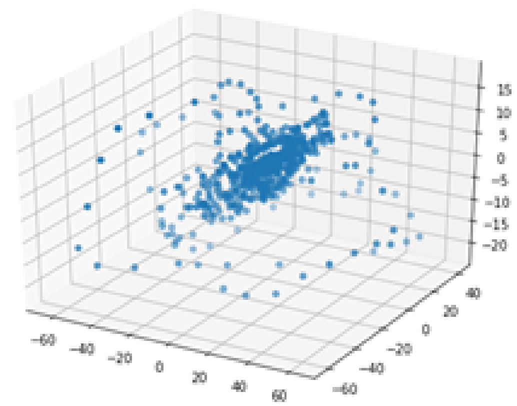
$$d_B(\text{dgm}(\mathbb{U}), \text{dgm}(\mathbb{V})) \leq \epsilon$$

Strategy: build filtrations that induce **q-tame** homology persistence modules and that turn out to be ϵ**-interleaved** when the considered spaces/functions are $O(\epsilon)$-*close*.

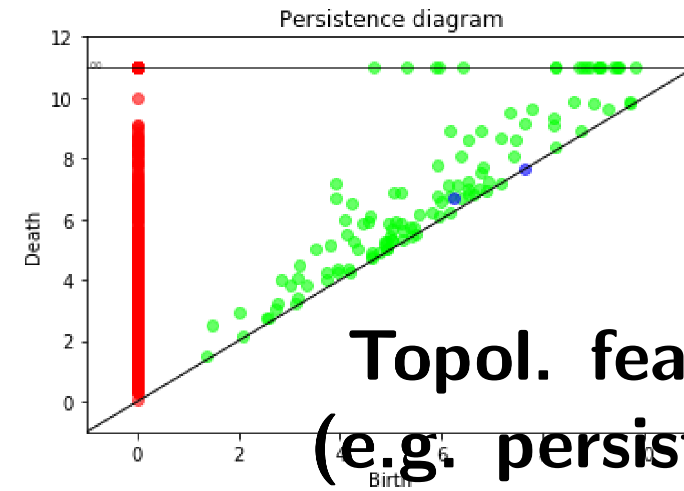
Exploiting persistent homology information

Motivation





Data



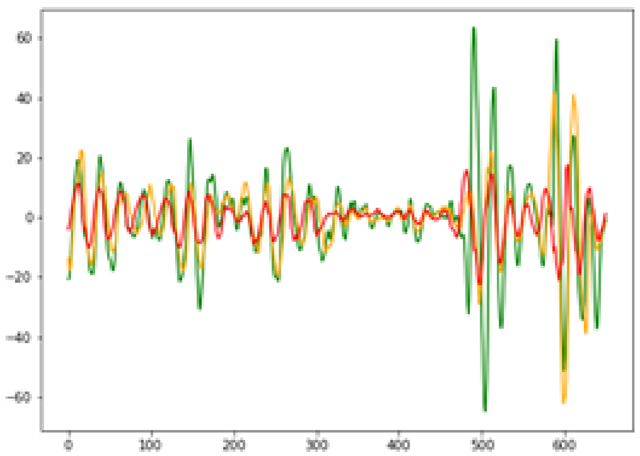
Topol. features (e.g. persistence)

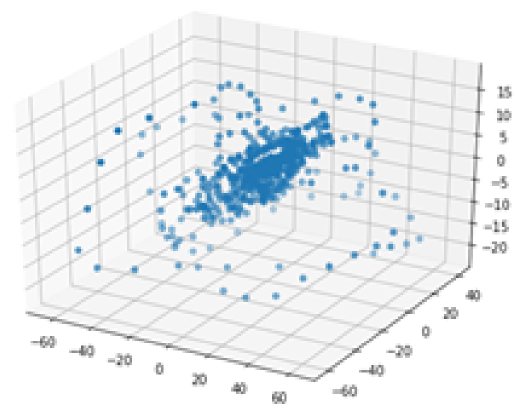
Strengths of topological information:

- robust, multiscale information,
- interpretable information (in some cases),
- benefit from a solid mathematical framework (e.g. persistent homology theory).

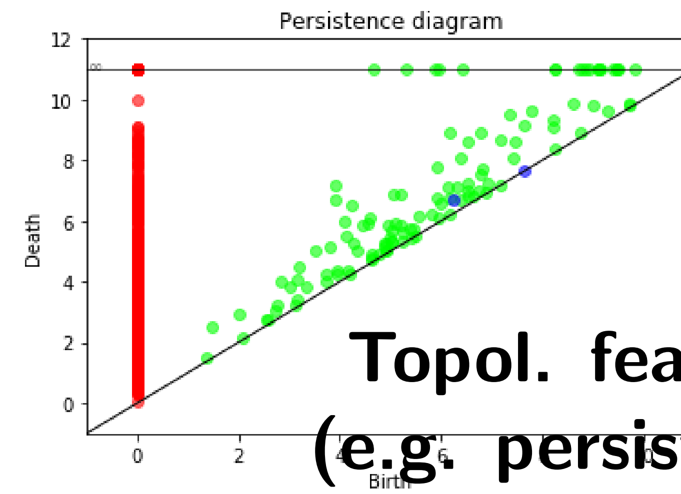
But...

Motivation



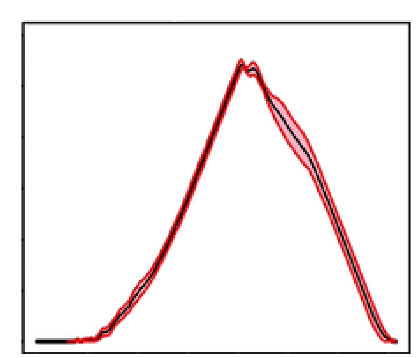


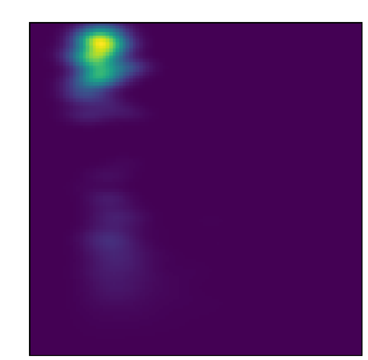
Data



Topol. features (e.g. persistence)

Machine Learning / AI

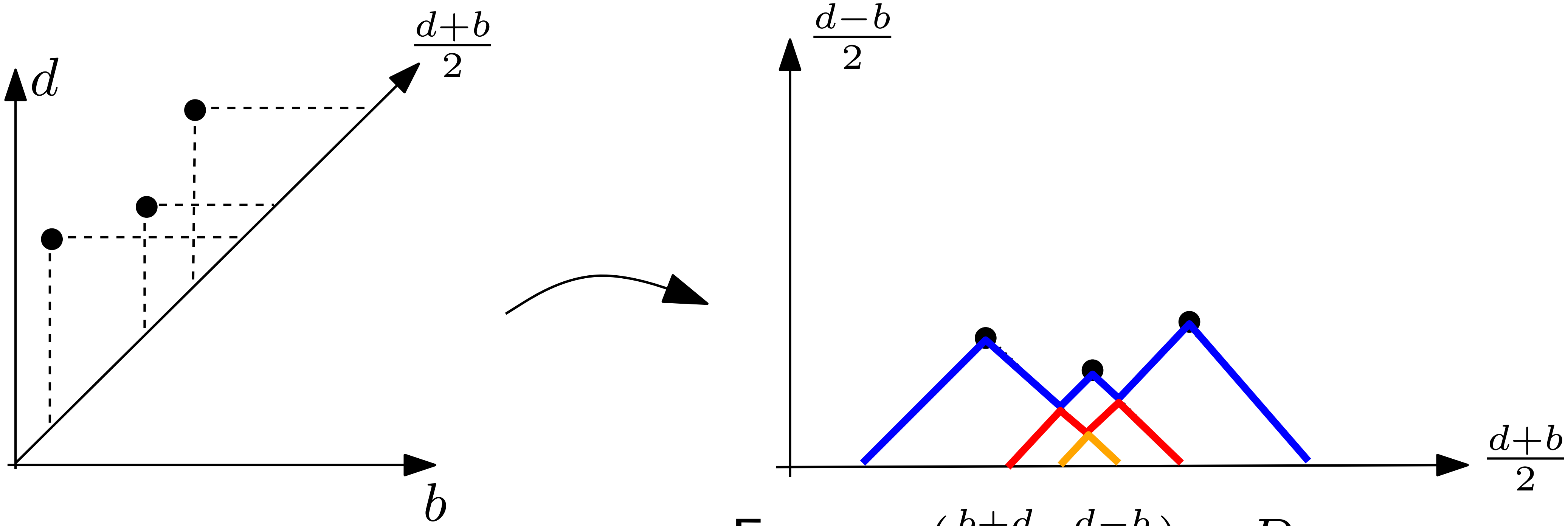




Representations of persistence

- Topological information is usually not well-suited for direct statistical analysis and ML algorithms (e.g. the space of PD is highly non linear)
- Not always clear which part of the topological features carries the relevant information.
- An active research area, but still at a very early stage (in particular regarding math. aspects).

Persistence landscapes



$D = \{(\frac{d_i+b_i}{2}, \frac{d_i+b_i}{2})\}i \in I$

For $p = (\frac{b+d}{2}, \frac{d-b}{2}) \in D$,

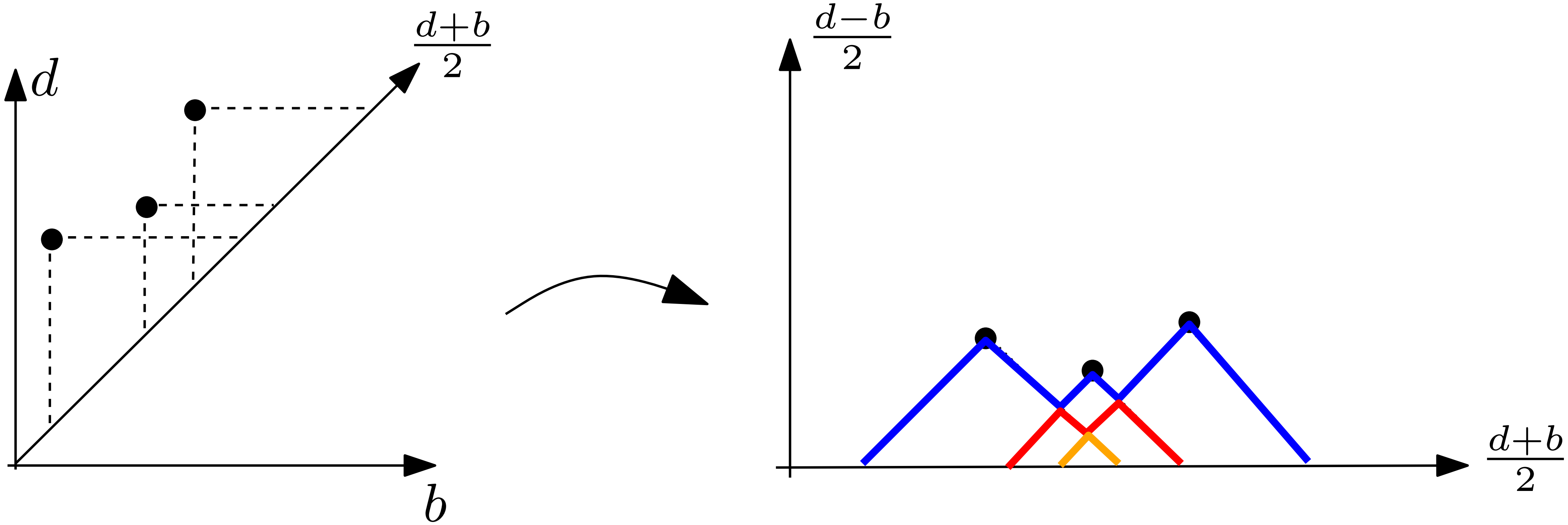
$$\Lambda_p(t) = \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases}$$

Persistence landscape [Bubenik 2012]:

$$\lambda_D(k,t) = \underset{p\in \mathsf{dgm}}{\mathsf{kmax}}\, \Lambda_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N},$$

where kmax is the kth largest value in the set.

Persistence landscapes



Persistence landscape [Bubenik 2012]:

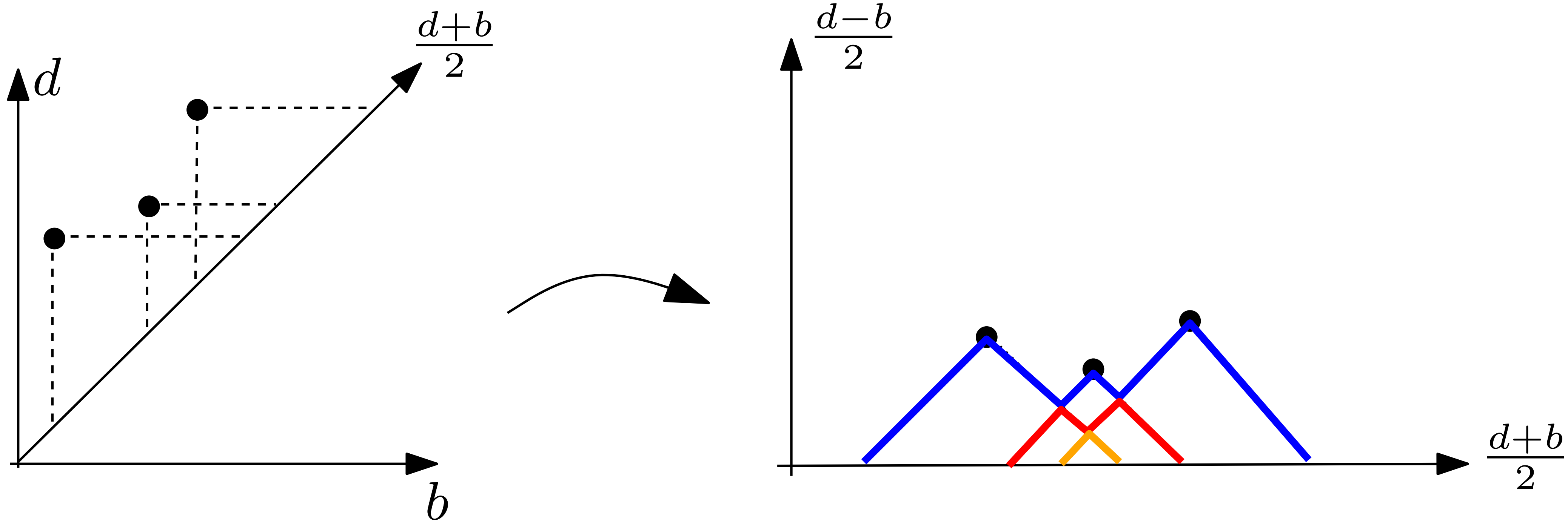
$$\lambda_D(k,t) = \underset{p \in \mathsf{dgm}}{\mathsf{kmax}}\ \Lambda_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N},$$

Properties

- For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $0 \le \lambda_D(k,t) \le \lambda_D(k+1,t)$.
- For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $|\lambda_D(k,t) - \lambda_{D'}(k,t)| \le d_B(D,D')$ where $d_B(D,D')$ denotes the bottleneck distance between D and D'.

stability properties of persistence landscapes

Persistence landscapes

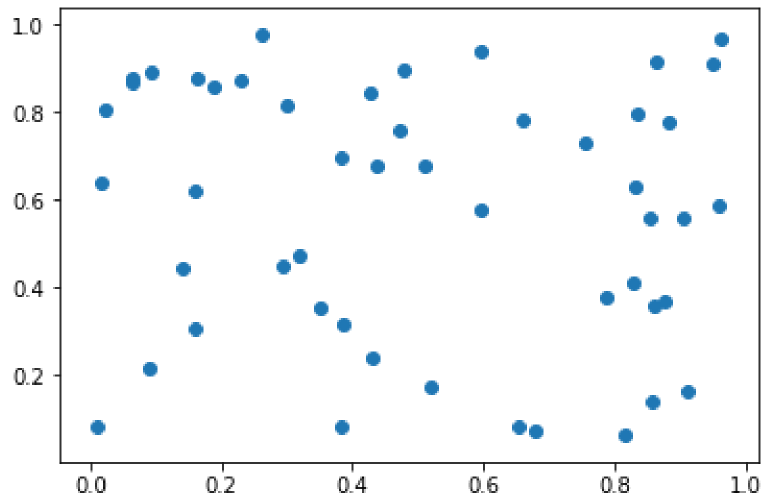


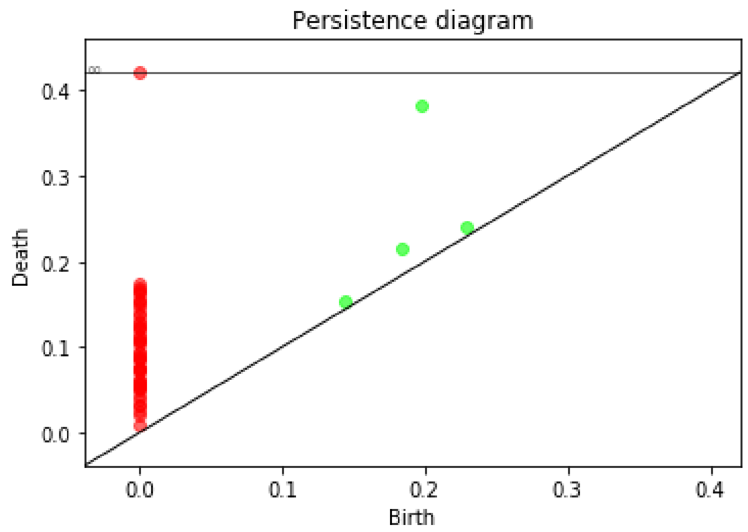
- Persistence encoded as an element of a functional space (vector space!).
- Expectation of distribution of landscapes is well-defined and can be approximated from average of sampled landscapes.
- Process point of view: convergence results and convergence rates $\rightarrow$ confidence intervals can be computed using bootstrap.

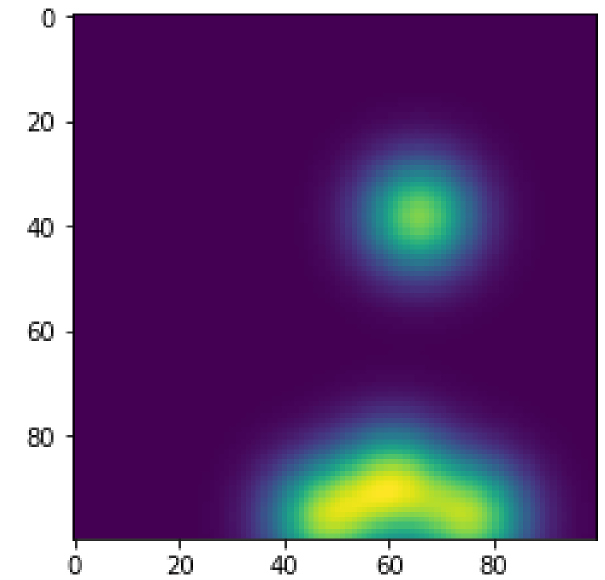
  [C., Fasy, Lecci, Rinaldo, Wasserman SoCG 2014]
- Provide a convenient way to process persistence information in deep neural networks. [Kim, Kim, Zaheer, Kim, C., Wasserman NeurIPS 2020, Carrière, C., Ike, Lacombe, Royer, Umeda AISTAT 2020]

Persistence images

[Adams et al, JMLR 2017]







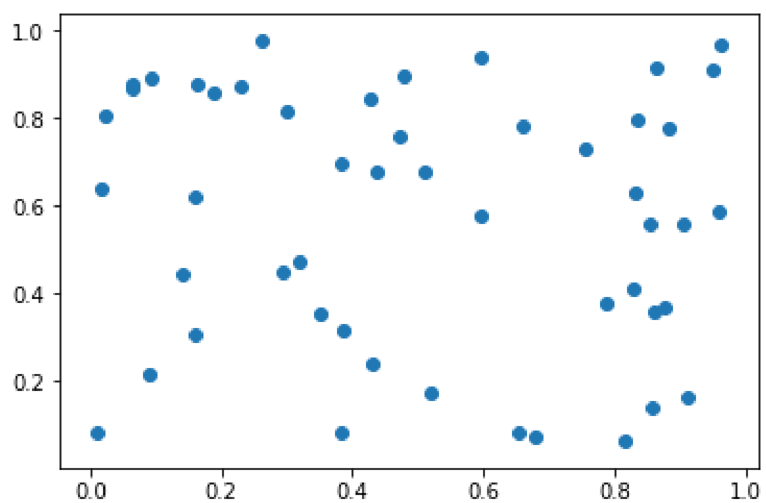
For $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel and H a bandwidth matrix (e.g. a symmetric positive definite matrix), pose for $u \in \mathbb{R}^2$, $K_H(z) = |H|^{-1/2} K(H^{-1/2} \cdot u)$

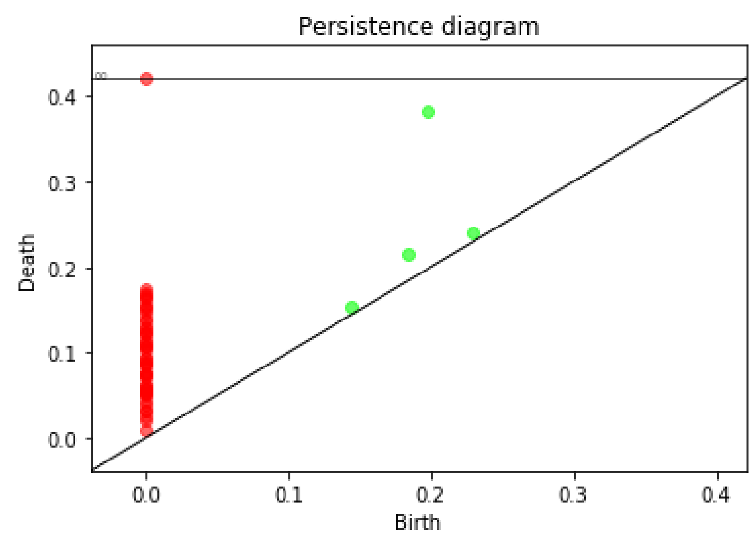
For $D = \sum_i \delta_{\mathbf{r}_i}$ a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by:

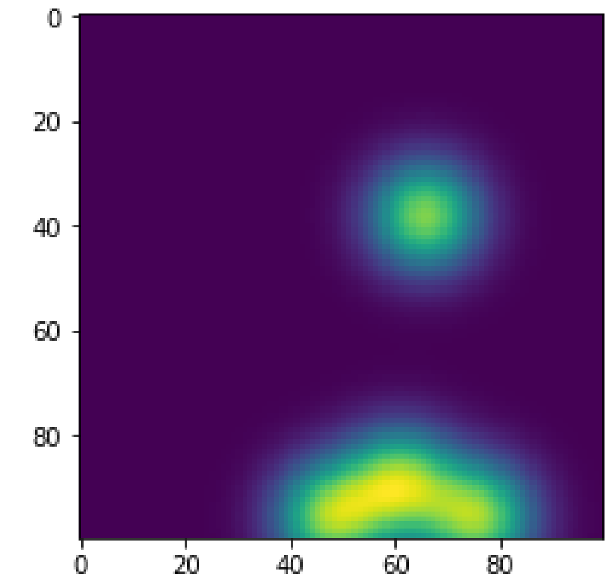
$$\forall z \in \mathbb{R}^2,\ \rho(D)(u) = \sum_i w(\mathbf{r}_i) K_H(u - \mathbf{r}_i) = D(wK_H(u - \cdot))$$

Persistence images

[Adams et al, JMLR 2017]







For $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel and H a bandwidth matrix (e.g. a symmetric positive definite matrix), pose for $u \in \mathbb{R}^2$, $K_H(z) = |H|^{-1/2} K(H^{-1/2} \cdot u)$

For $D = \sum_i \delta_{\mathbf{r}_i}$ a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by:

$$\forall z \in \mathbb{R}^2,\ \rho(D)(u) = \sum_i w(\mathbf{r}_i) K_H(u - \mathbf{r}_i) = D(w K_H(u - \cdot))$$

$\Rightarrow$ persistence surfaces can be seen as kernel estimates of $E[D_s[\mathcal{K}(\mathbb{X})]]$.

A zoo of representations of persistence

(non exhaustive list - see also Gudhi representations)

- Collections of 1D functions

  $\rightarrow$ landscapes [Bubenik 2012]

  $\rightarrow$ Betti curves [Umeda 2017]

- discrete measures: (interesting statistical properties [Chazal, Divol 2018])

  $\rightarrow$ persistence images [Adams et al 2017]

  $\rightarrow$ convolution with Gaussian kernel [Reininghaus et al. 2015] [Chepushtanova et al. 2015] [Kusano Fukumisu Hiraoka 2016-17] [Le Yamada 2018]

  $\rightarrow$ sliced on lines [Carrière Oudot Cuturi 2017]

- finite metric spaces [Carrière Oudot Ovsjanikov 2015]

- polynomial roots or evaluations [Di Fabio Ferri 2015] [Kališnik 2016]

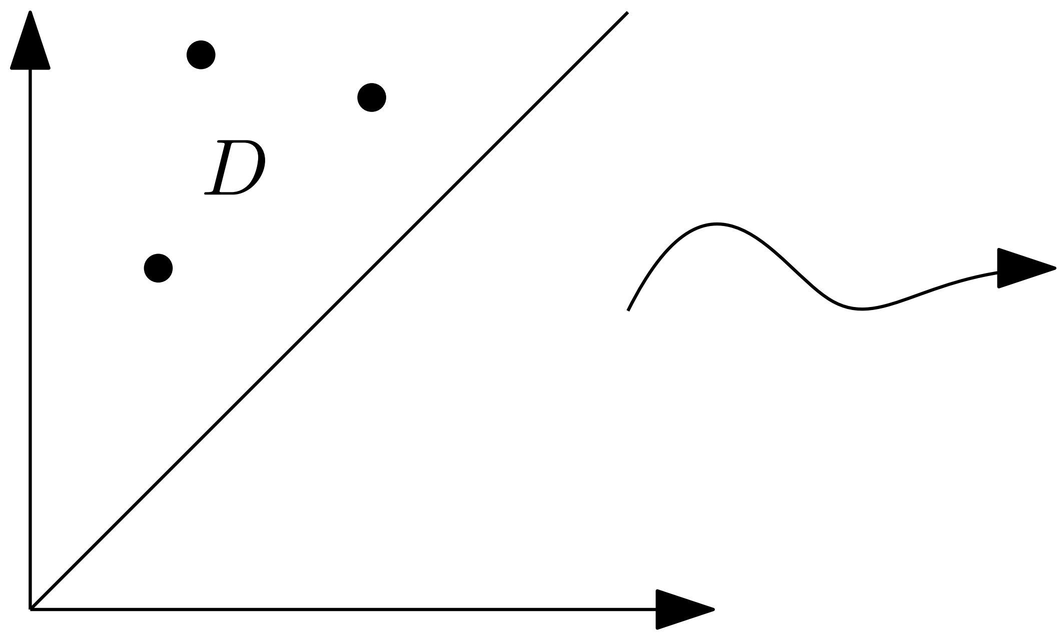
A zoo of representations of persistence

(non exhaustive list - see also Gudhi representations)

- Collections of 1D functions
  - $\rightarrow$ landscapes [Bubenik 2012]
  - $\rightarrow$ Betti curves [Umeda 2017]
- discrete measures: (interesting statistical properties [Chazal, Divol 2018])
  - $\rightarrow$ persistence images [Adams et al 2017]
  - $\rightarrow$ convolution with Gaussian kernel [Reininghaus et al. 2015] [Chepushtanova et al. 2015] [Kusano Fukumisu Hiraoka 2016-17] [Le Yamada 2018]
  - $\rightarrow$ sliced on lines [Carrière Oudot Cuturi 2017]
- finite metric spaces [Carrière Oudot Ovsjanikov 2015]
- polynomial roots or evaluations [Di Fabio Ferri 2015] [Kališnik 2016]

Problem: How to chose the right representation?

Persistence diagrams as discrete measures



$$D := \sum_{p \in D} \delta_p$$

Motivations:

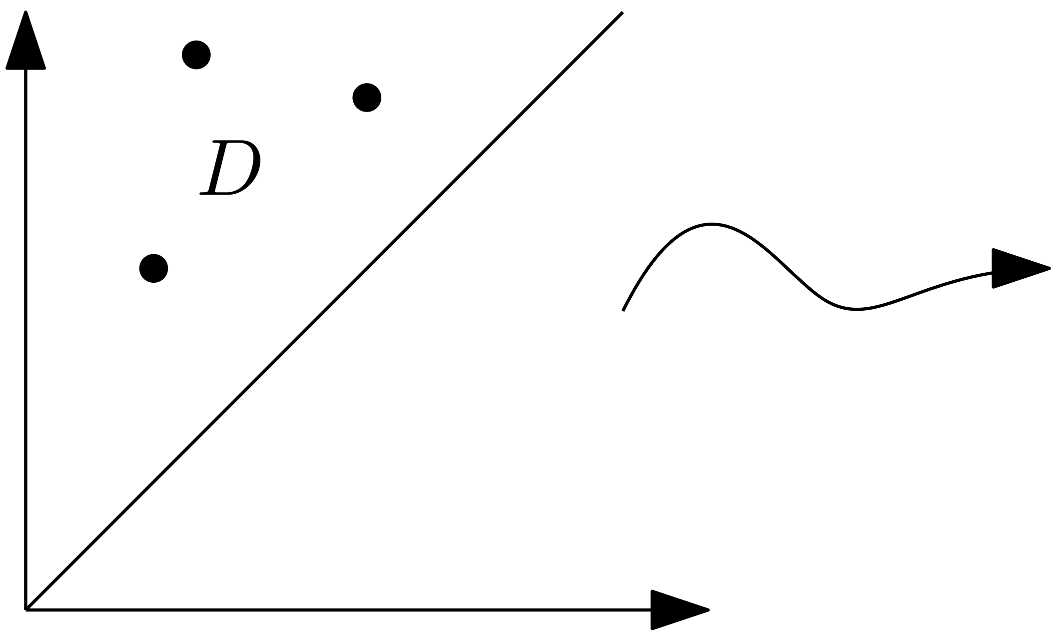
- The space of measures is much nicer that the space of P. D. !
- In the general algebraic persistence theory, persistence diagrams naturally appears as discrete measures in the plane.

[C., de Silva, Glisse, Oudot 16]

- Many persistence representations can be expressed as

$$D(f) = \sum_{p \in D} f(p) = \int f dD$$

Persistence diagrams as discrete measures



$$D := \sum_{p \in D} \delta_p$$

Benefits:

- Interesting statistical properties
- Data-driven selection of well-adapted representations (supervised and unsupervised, coming with guarantees)
- Optimisation of persistence-based functions

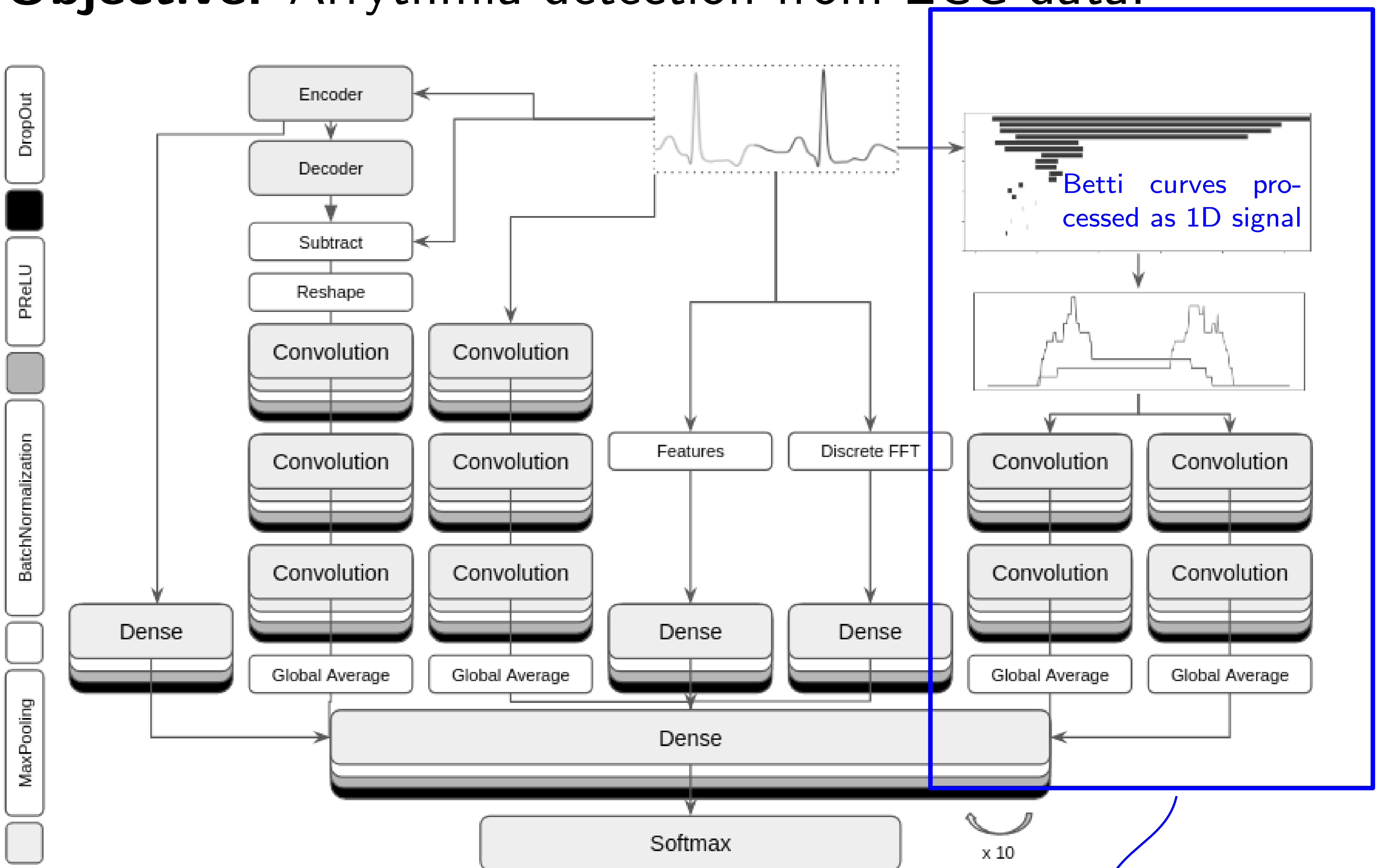
Many tools available and implemented in the GUDHI library

A few illustrative applications

Example of application: arrhythmia detection

[Dindin, Umeda, C. Can. Conf. AI 2020]

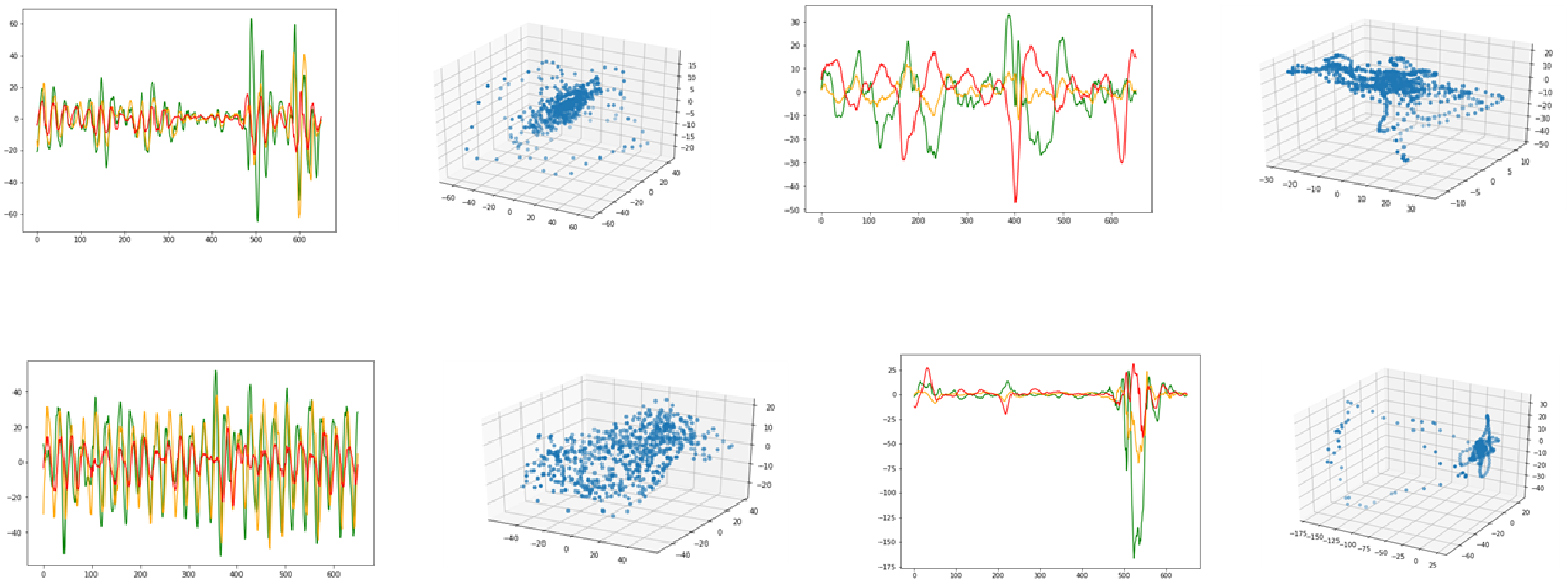
Objective: Arrythmia detection from ECG data.



- Improvement over state-of-the-art.
- Better generalization.

	Accuracy[%]
UCLA (2018)	93.4
Li et al. (2016)	94.6
Inria-Fujitsu (2018)*	**98.6**

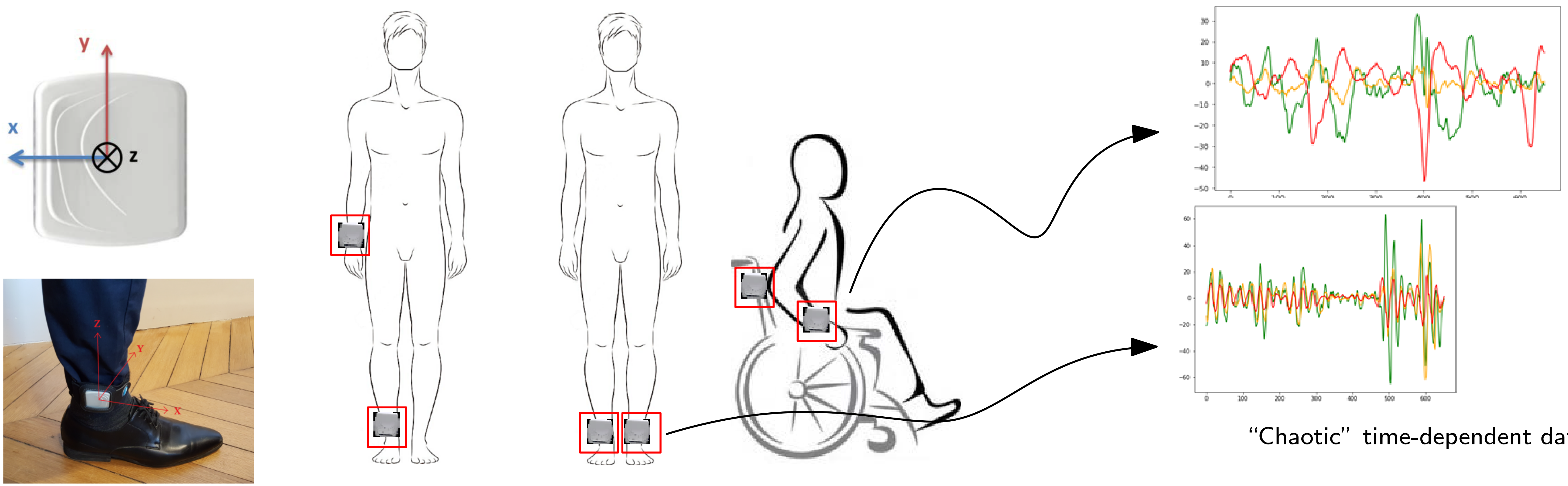
TDA and Machine Learning for sensor data



(Multivariate) time-dependent data can be converted into point clouds: sliding window, time-delay embedding,...

With landscapes: patient monitoring

[Beaufils, C., Dindin, Grelet, Michel 2018]



"Chaotic" time-dependent data

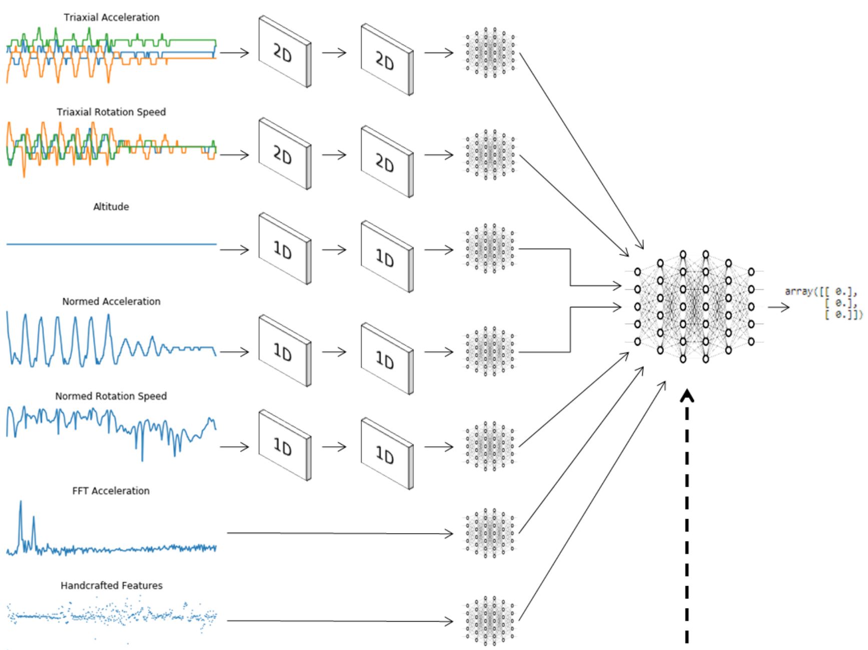
Objective: precise analysis of movements and activities of pedestrians.

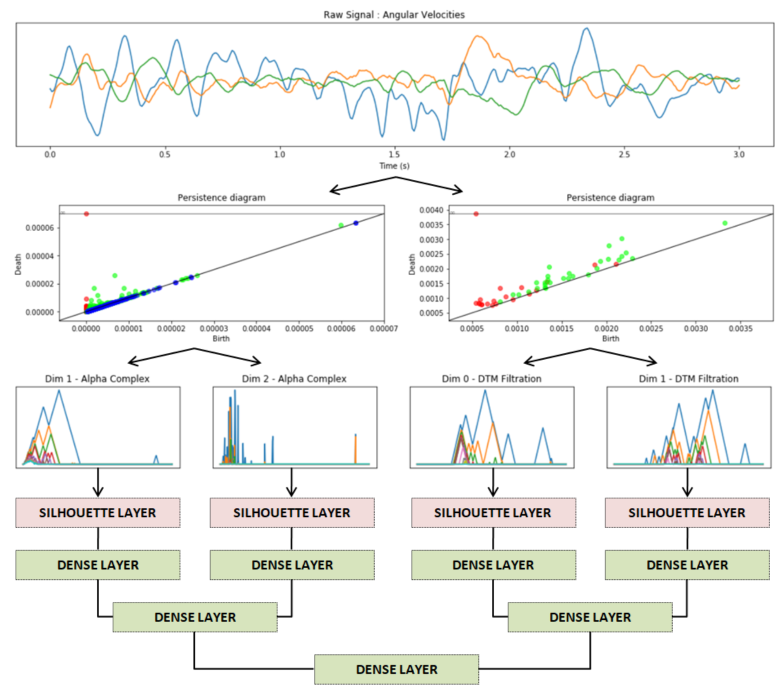
Applications: personal healthcare; medical studies; defense.

With landscapes: patient monitoring

[Beaufils, C., Dindin, Grelet, Michel 2018]

Example: Dyskinesia crisis detection and activity recognition:





Class	Naive	Multi	FEA	QUA	TDA
Walking	97.6	98.4	99.3	99.0	**99.5**
Upstairs	97.2	**99.8**	97.8	98.0	97.7
Downstairs	99.6	**99.7**	99.0	98.4	98.3
Sitting	87.1	93.1	89.7	91.8	**96.5**
Standing	87.0	97.7	97.2	97.2	**98.1**
Laying	92.4	**100.**	99.8	99.9	**100.**
Stand-Sit	90.8	**95.6**	89.1	91.3	93.4
Sit-Stand	**100.**	99.9	**100.**	**100.**	**100.**
Sit-Lie	87.1	81.1	84.2	90.0	**95.1**
Lie-Sit	81.4	81.8	85.9	**91.8**	87.9
Stand-Lie	74.2	**87.6**	86.5	87.4	81.5
Lie-Stand	80.4	72.1	**83.2**	77.7	**83.2**

Multi-channels CNN + TDA neural network

Results on publicly available data set (HAPT) - improve the state-of-the-art.

- Data collected in non controlled environments (home) are very chaotic.
- Data registration (uncertainty in sensors orientation/position).
- Reliable and robust information is mandatory.
- Events of interest are often rare and difficult to characterize.

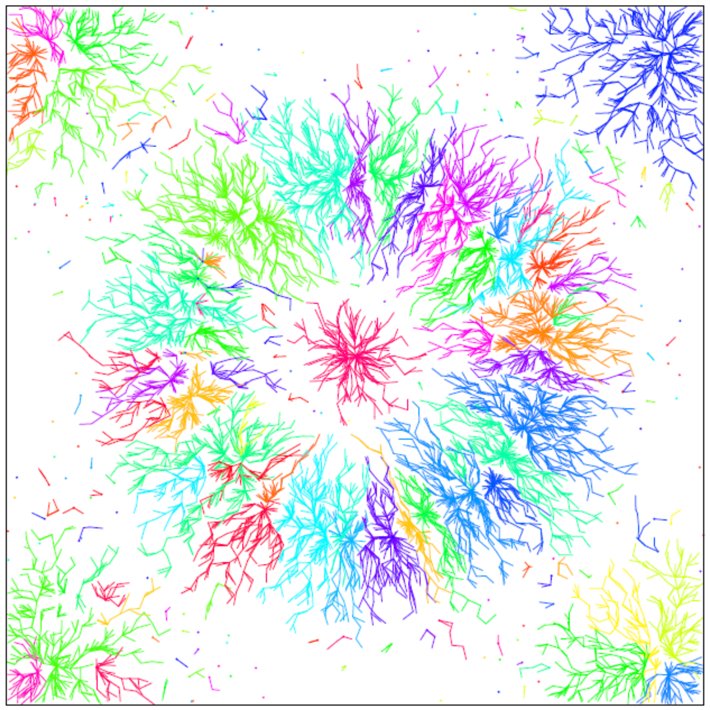
sysnav NAVIGATION TECHNOLOGIES

Inria informatics mathematics

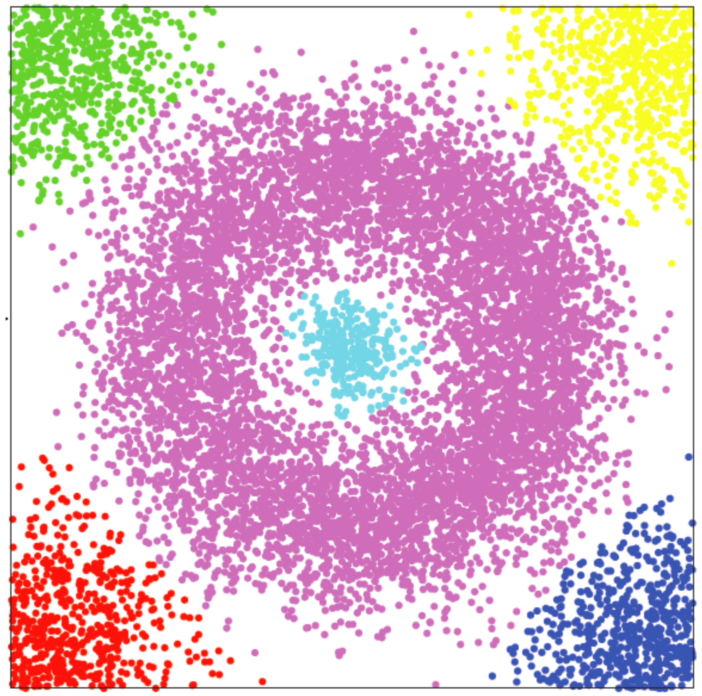
Thank you for your attention!

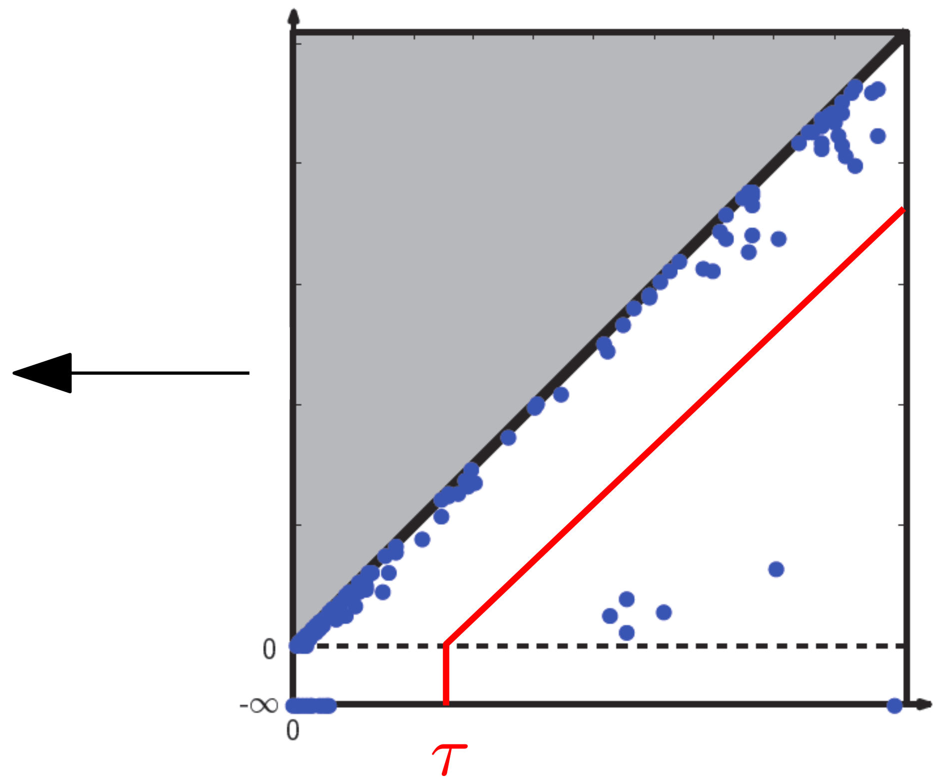
Some applications (illustrations)

- Persistence-based clustering [C.,Guibas,Oudot,Skraba - J. ACM 2013]

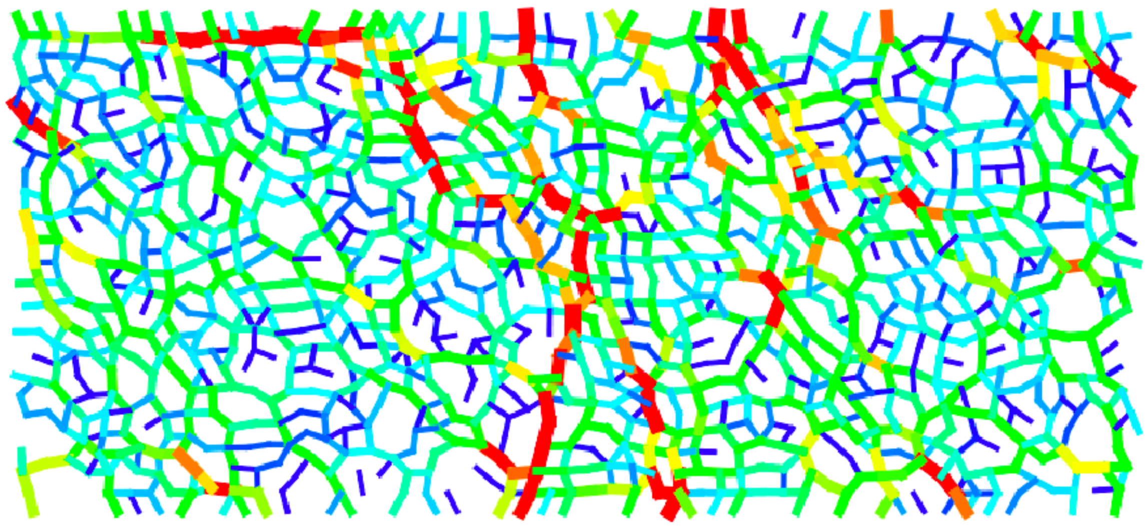


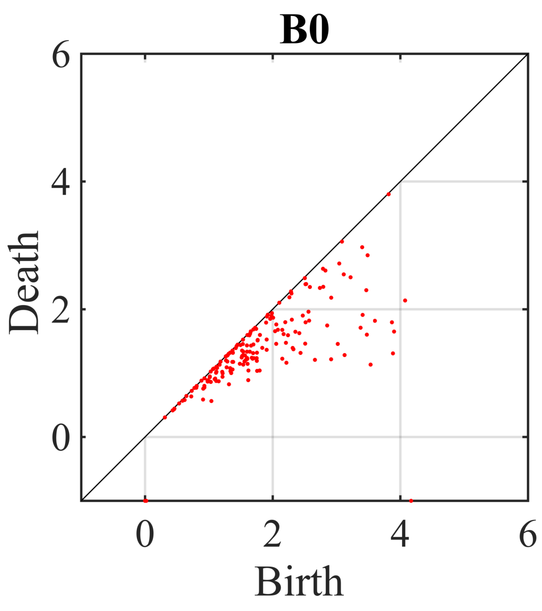
$\tau = 0$





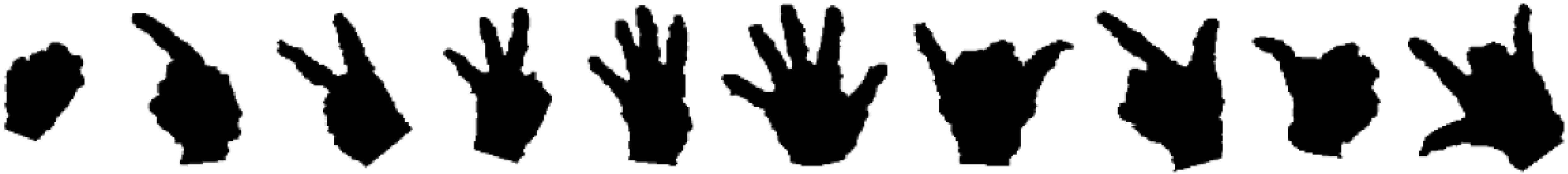
- Analysis of force fields in granular media [Kramar, Mischaikow et al]

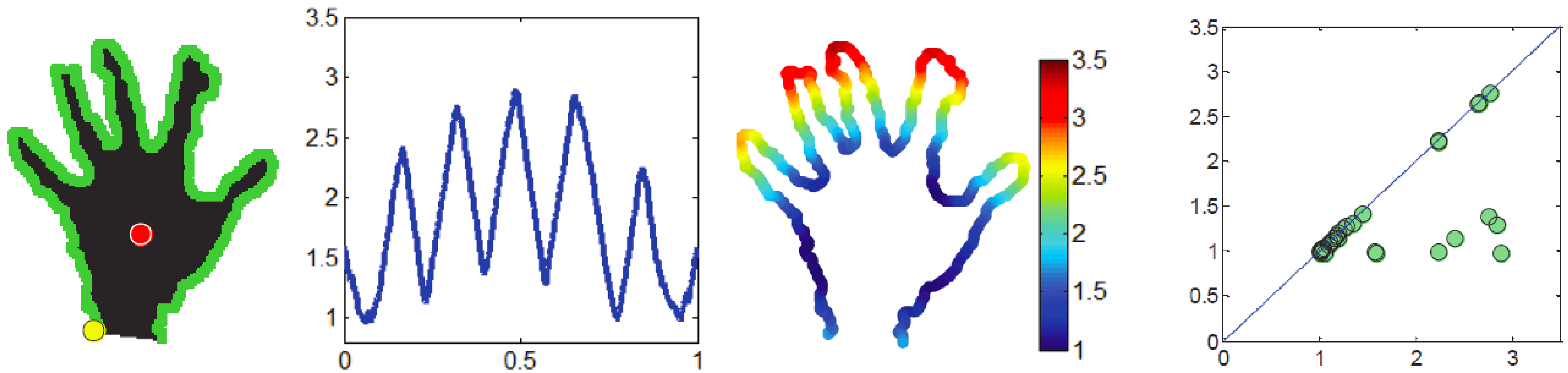




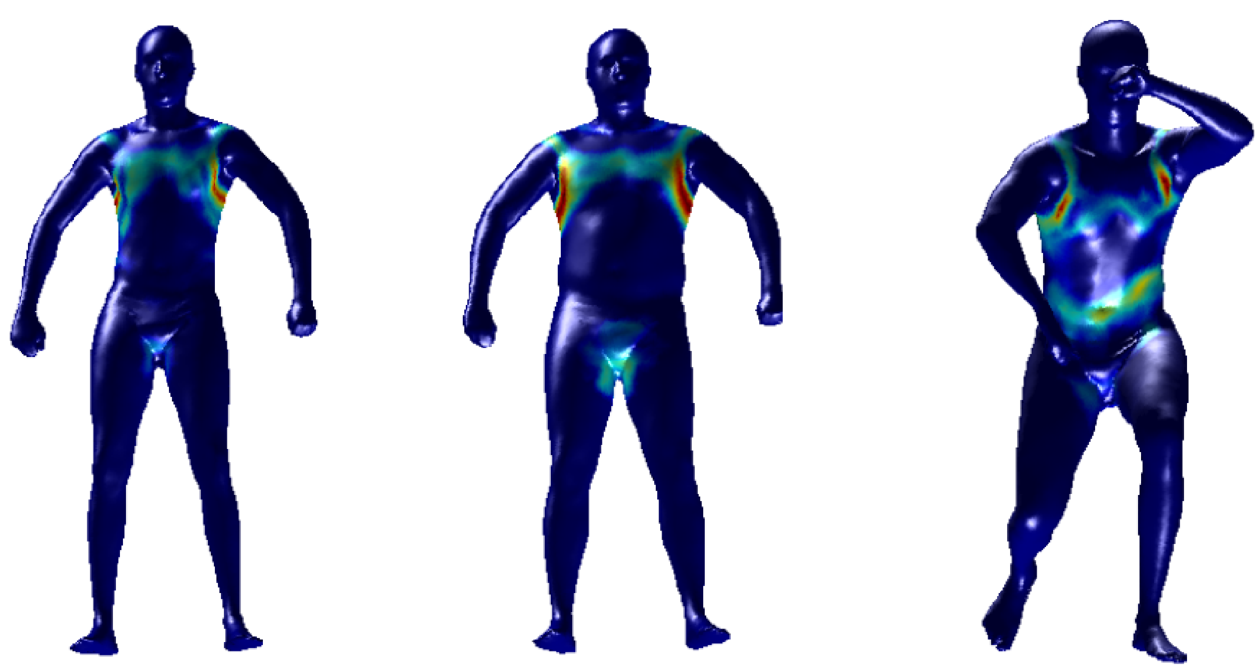
Some applications (illustrations)

- Hand gesture recognition [Li, Ovsjanikov, C. - CVPR'14]





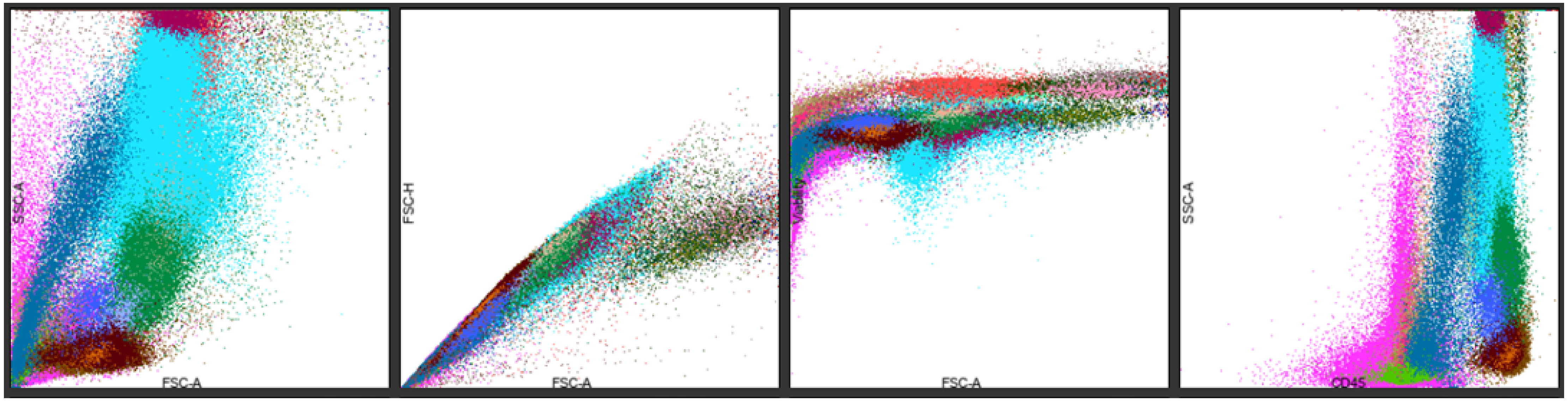
- Persistence-based pooling for shape recognition [Bonis, Ovsjanikov, Oudot, C. 2016]



Topology-based unsupervised classification and anomaly detection on cytometry data for medical diagnosis

METAFORA biosystems

An innovative start-up specialized in biological diagnosis from cytometry data.



Objective: unsupervised learning in large point clouds (several millions) in medium/high dimensions ($\approx 4 \to 80$)

Applications: medical diagnosis from blood samples (1 point $=$ 1 blood cell)

Methodology: TDA based approaches, combined with dim. reduction methods to identify relevant patterns and subsamples.