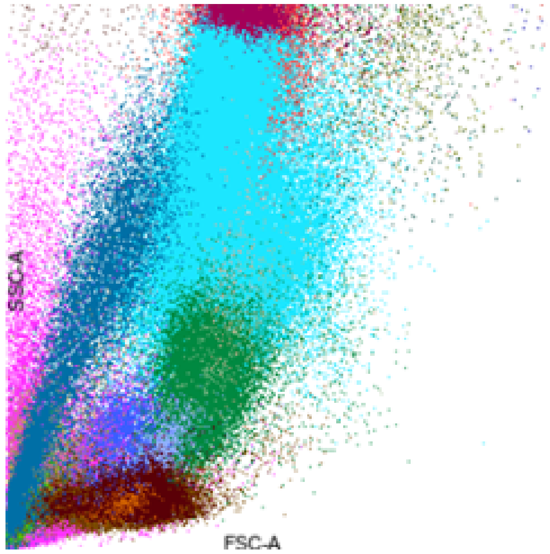
Rouen - September 22, 2022

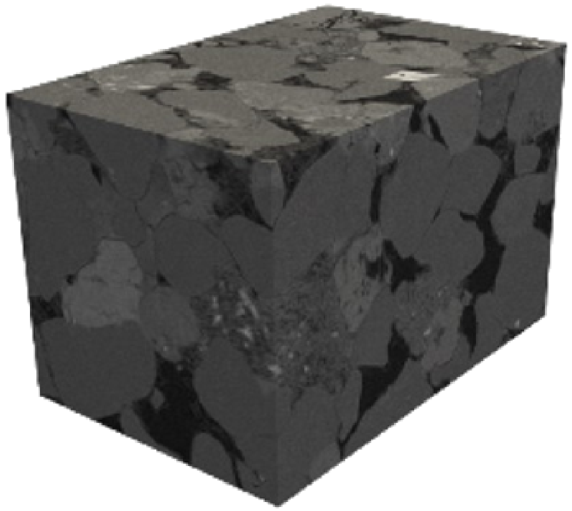
Comprendre la structure topologique des données : une introduction à l'homologie persistante.

Frédéric Chazal
DataShape team
Inria & Laboratoire de Mathématiques d'Orsay
Institut DATAIA Université Paris-Saclay

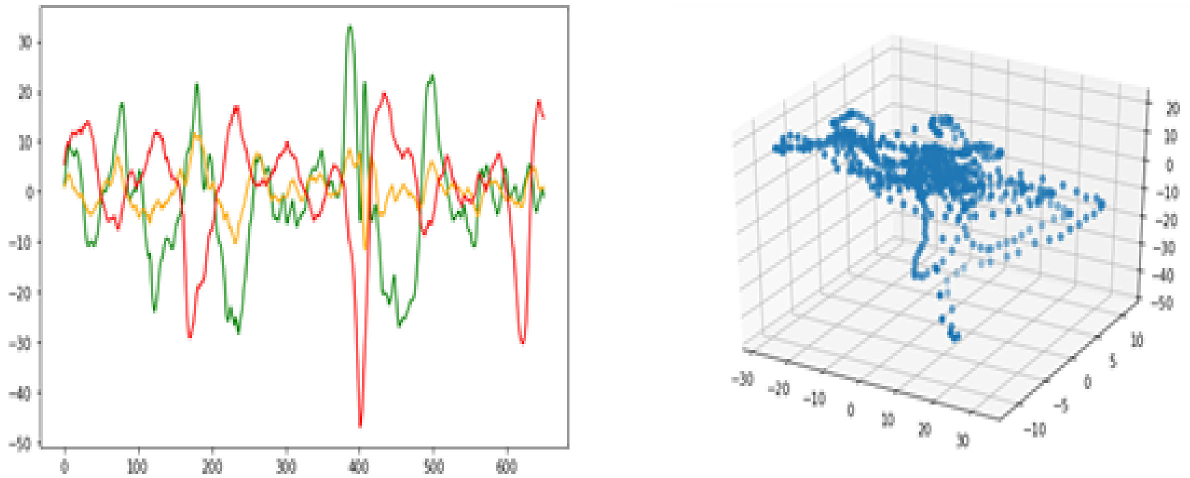
What is Topological Data Analysis (TDA) ?



[Cell population - cytometry - MetaFora courtesy]



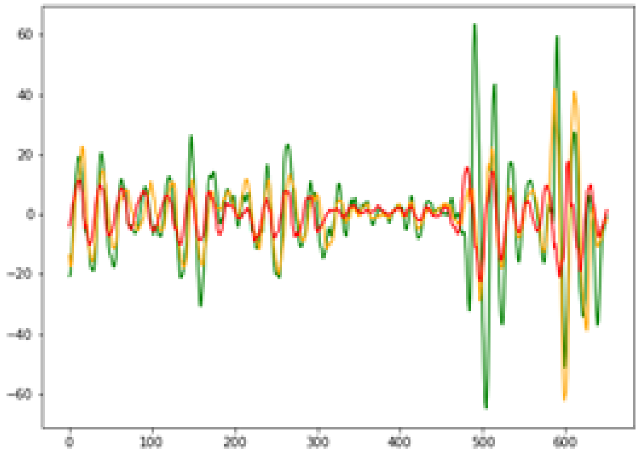
[Porous material (IFPEN courtesy)]

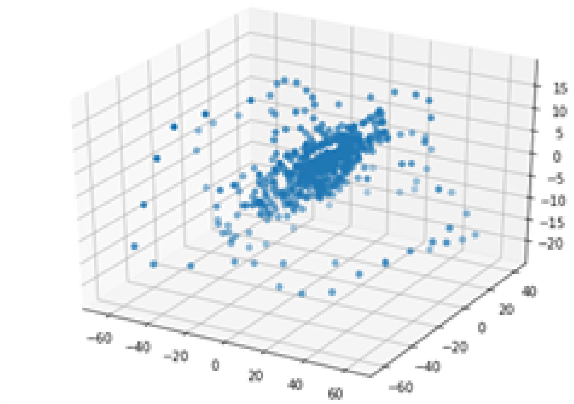


[Sensors (Sysnav courtesy)]

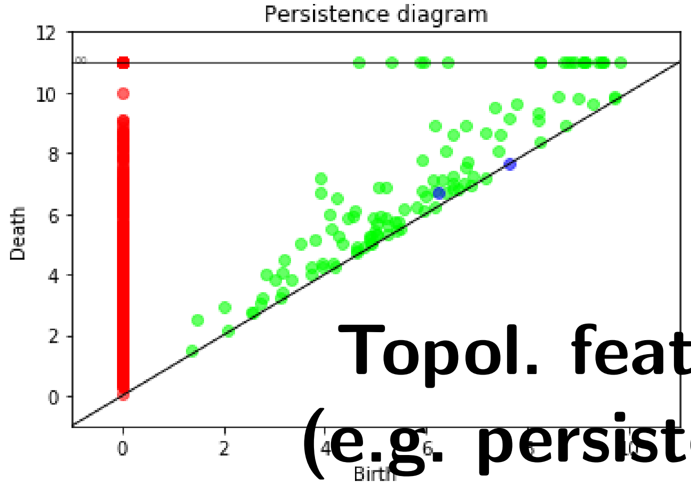
Modern data carry complex, but important, geometric/topological structure !

What is Topological Data Analysis (TDA) ?





Data

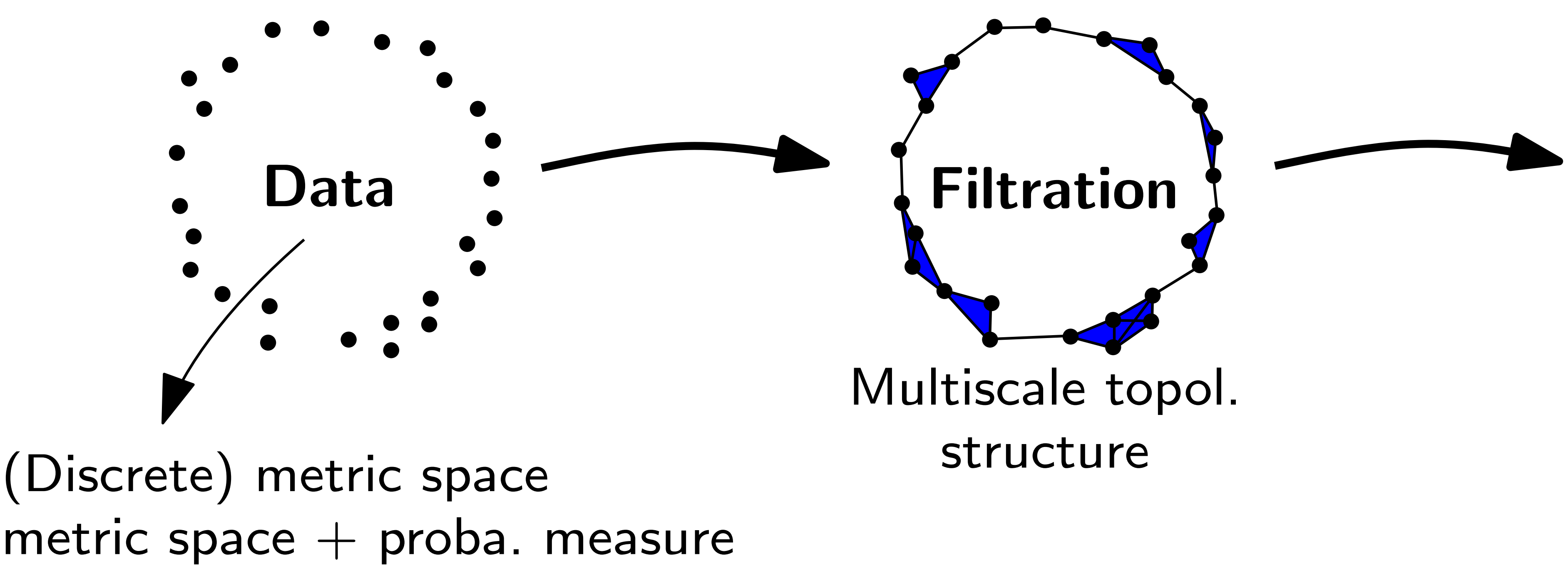


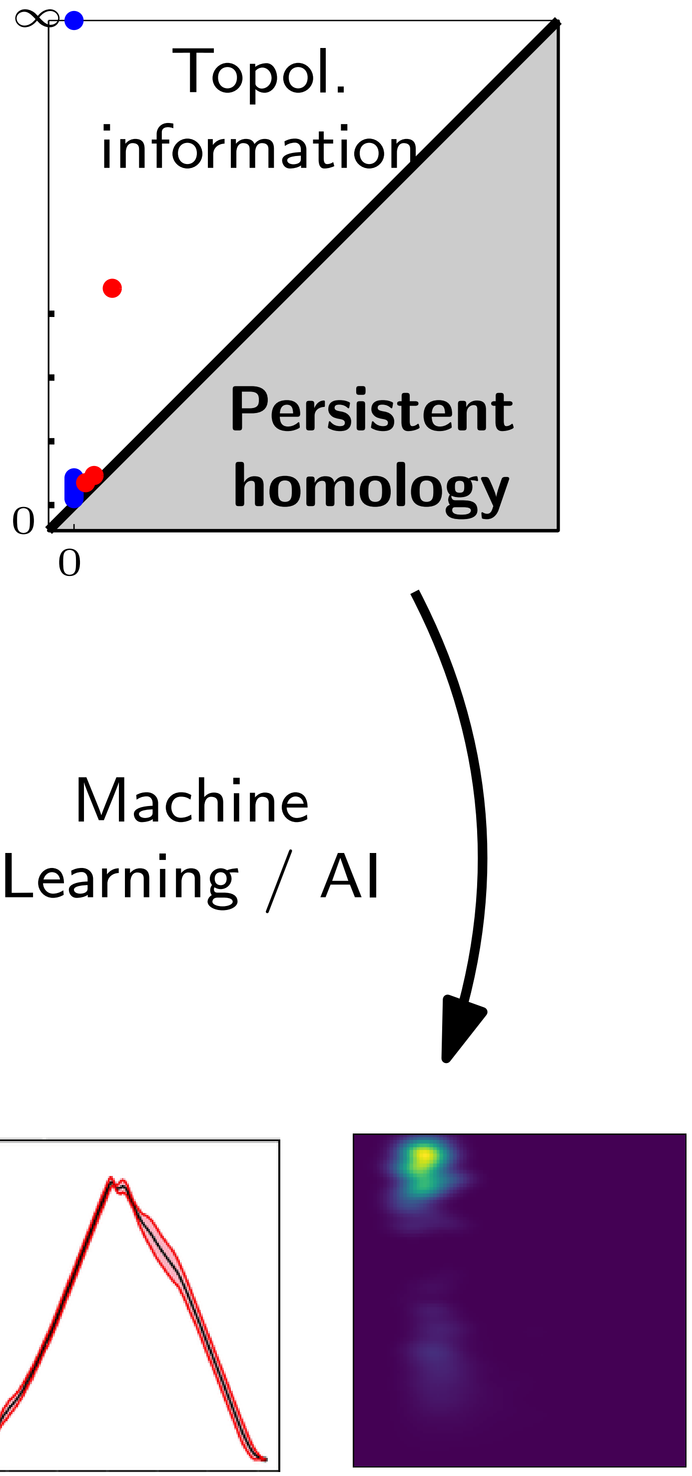
Topol. features (e.g. persistence)

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.

A classical TDA pipeline





Representations of persistence

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

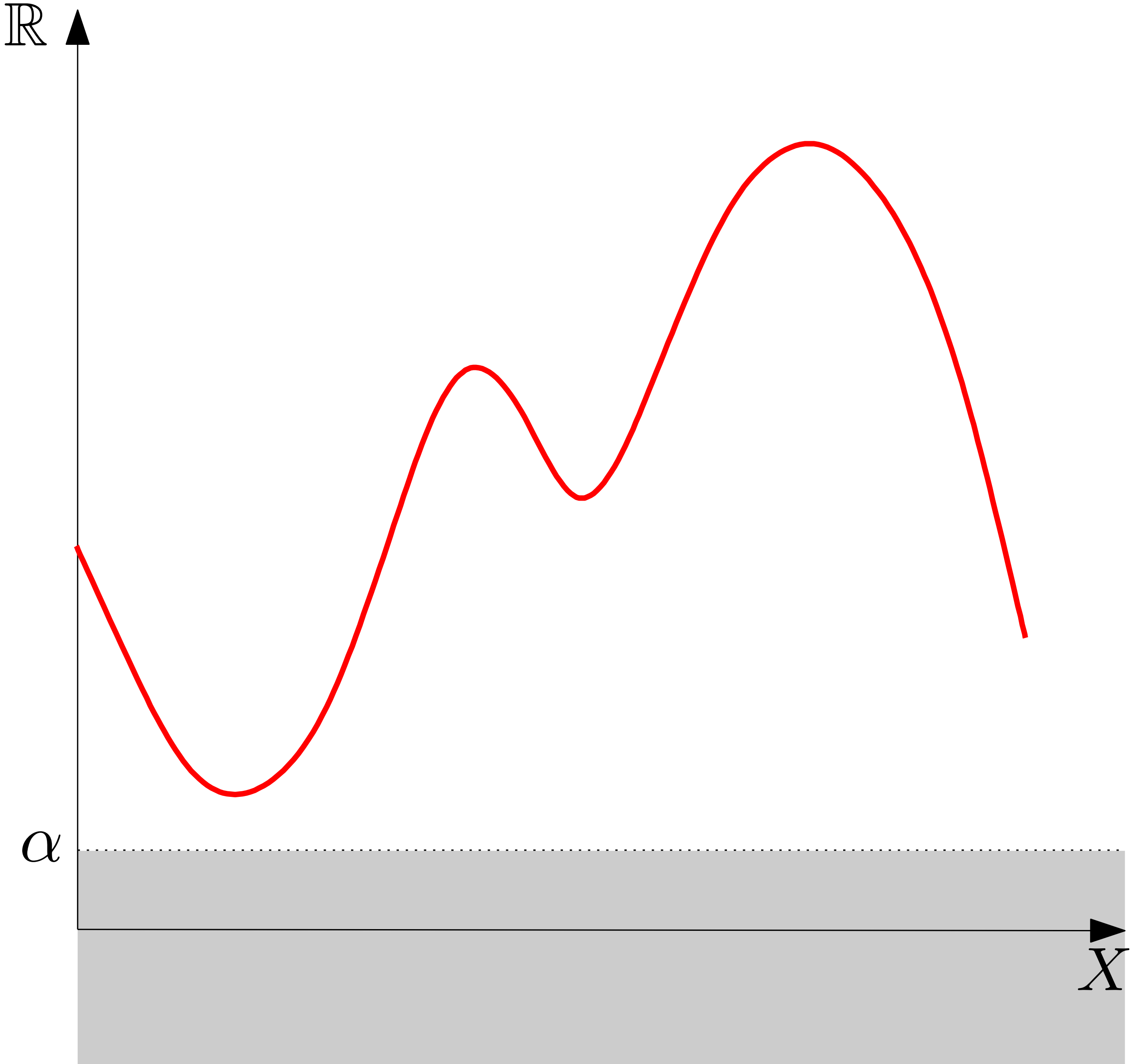
Persistent homology
Starting with a few examples

A general mathematical framework to encode the evolution of the topology (homology) of families of nested spaces (filtrations).

- $90's$: 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.

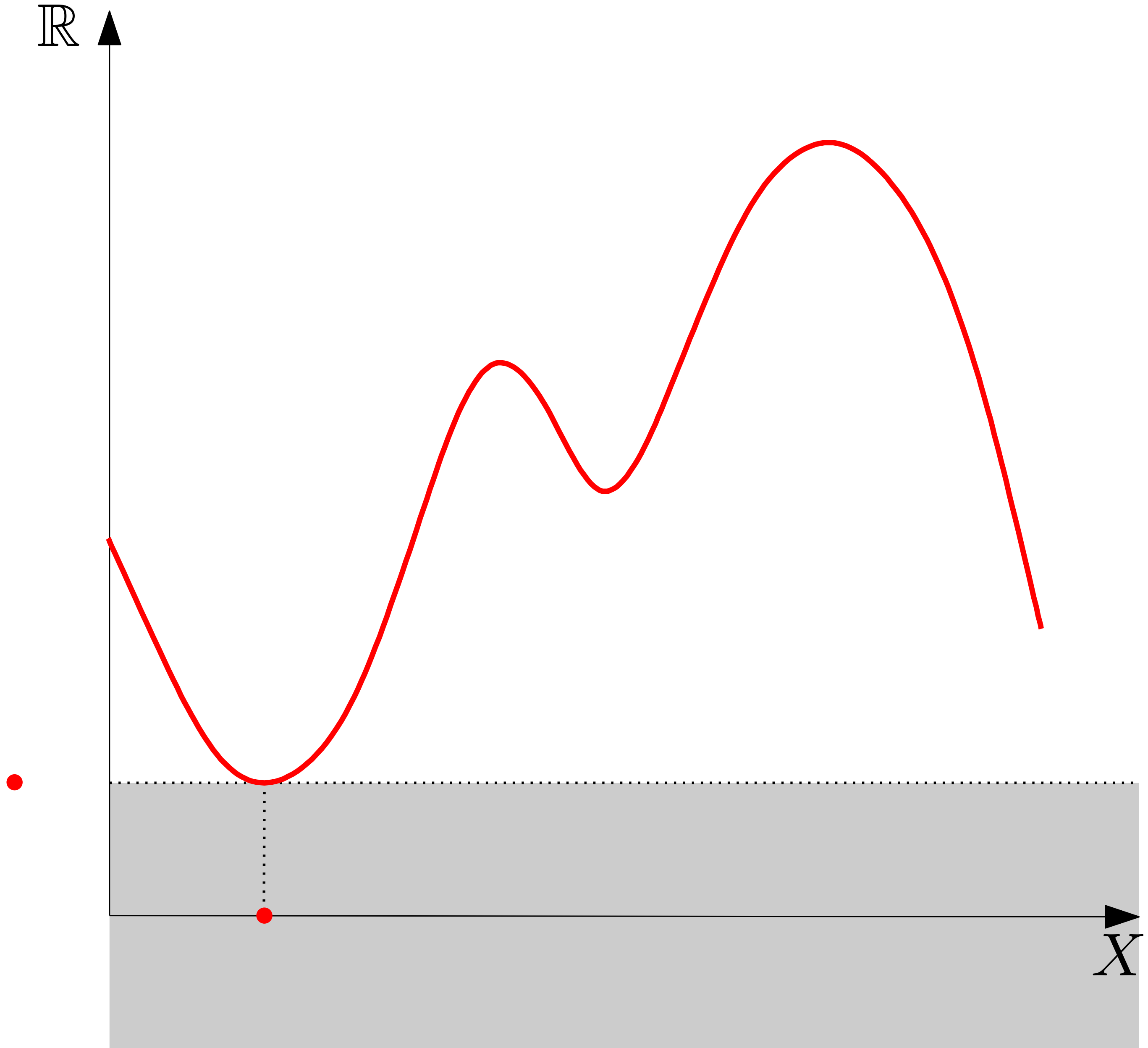
Persistent homology for functions

- 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.



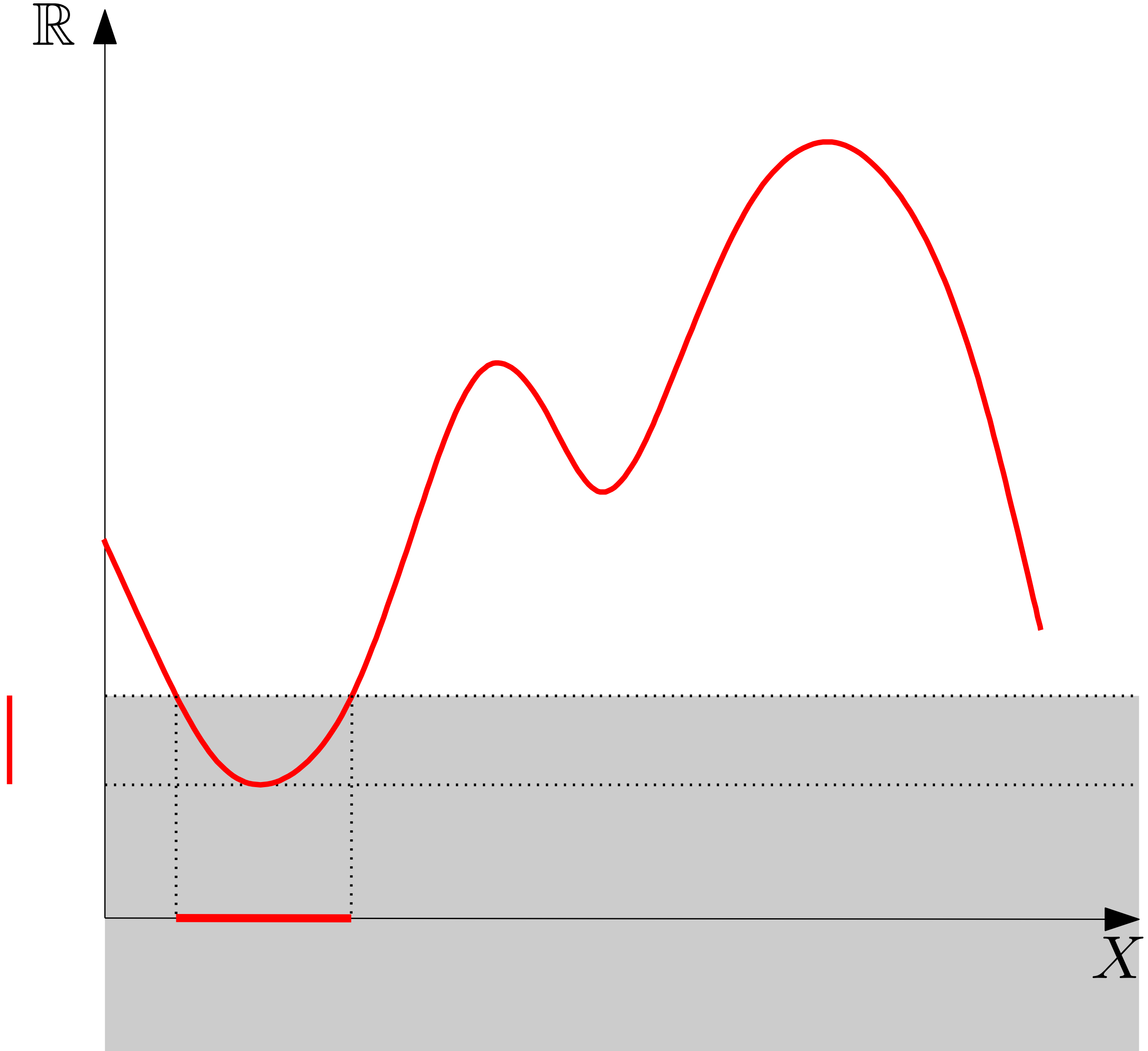
Persistent homology for functions

- 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.



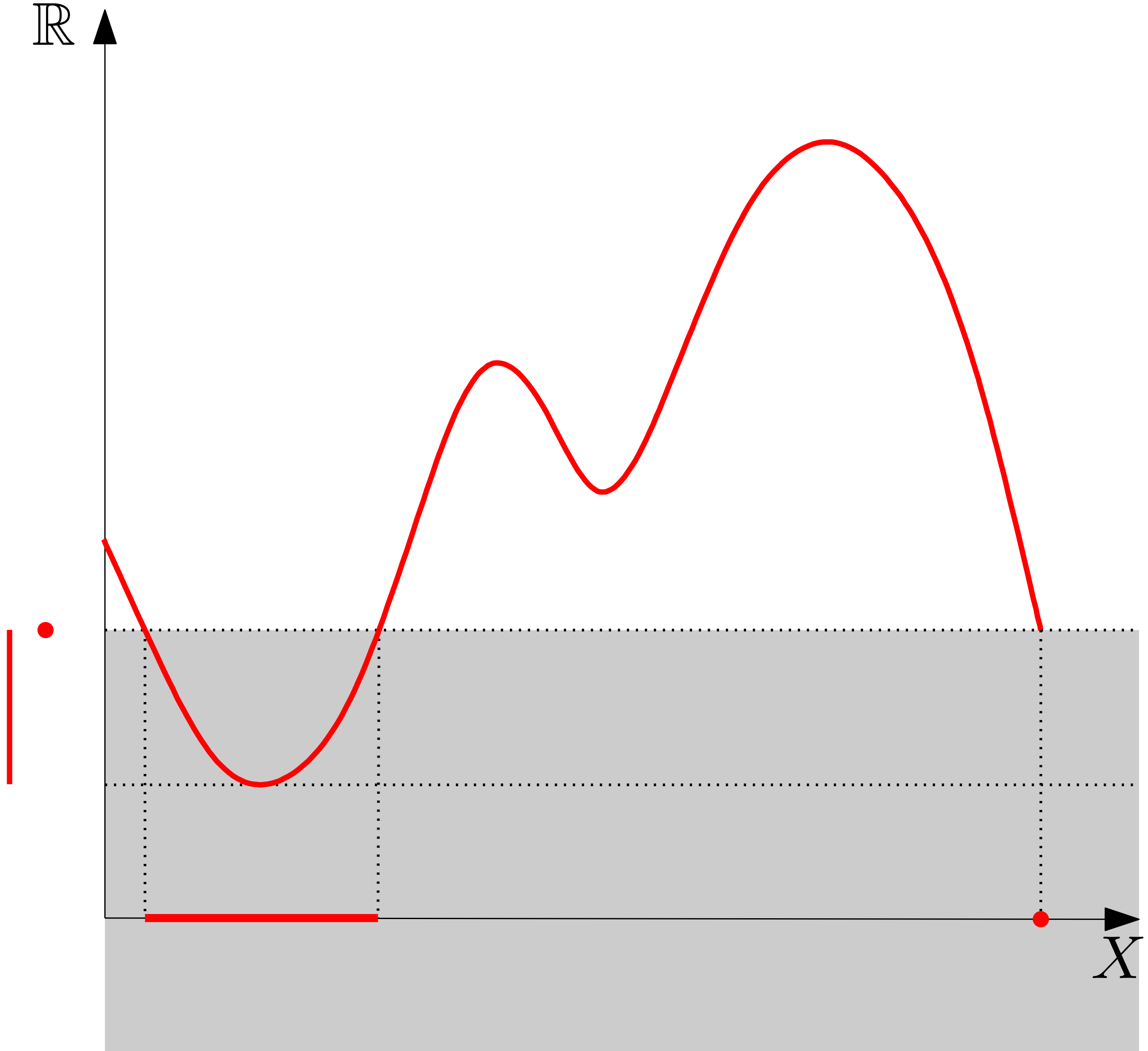
Persistent homology for functions

- 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.



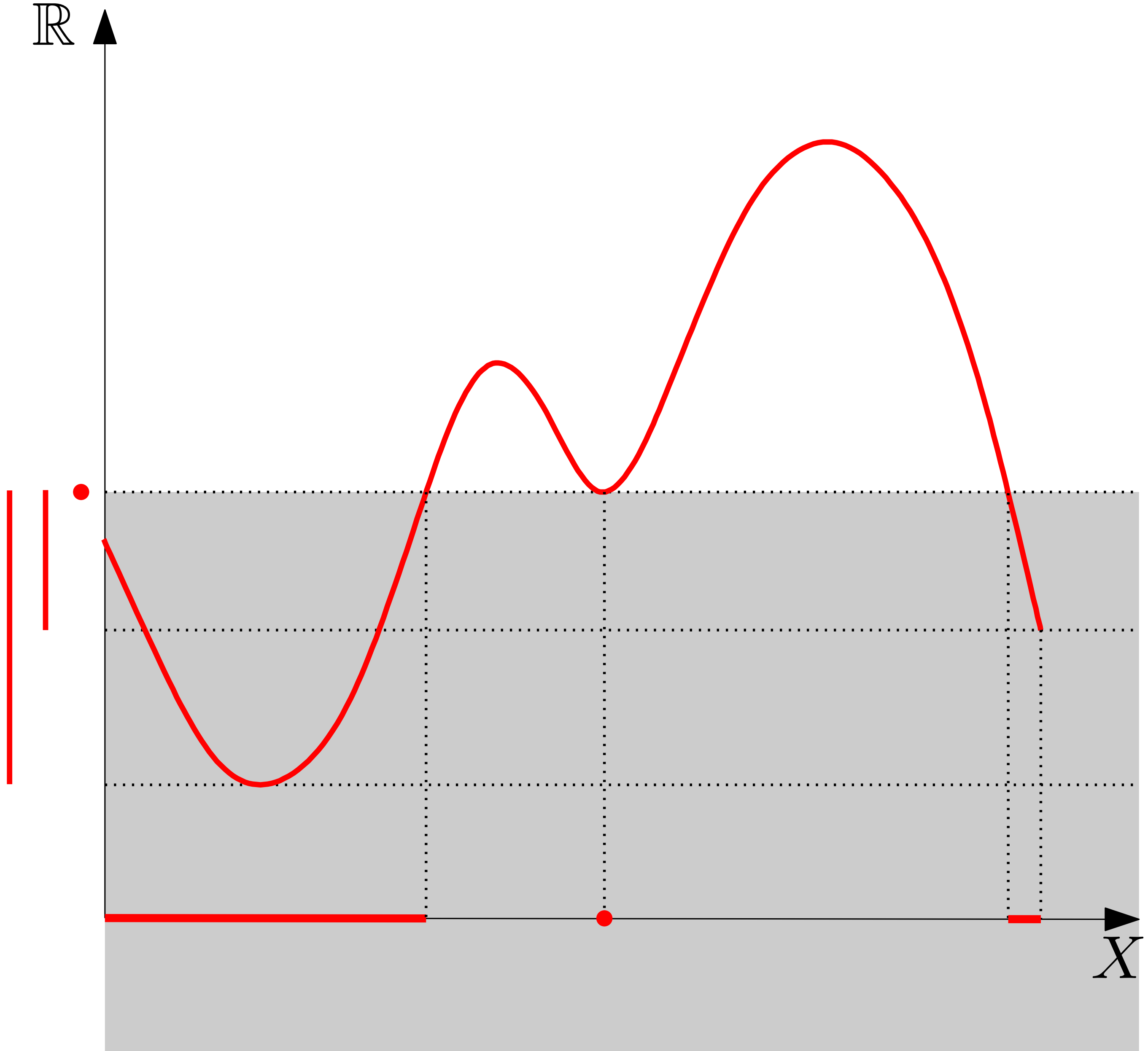
Persistent homology for functions

- 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.



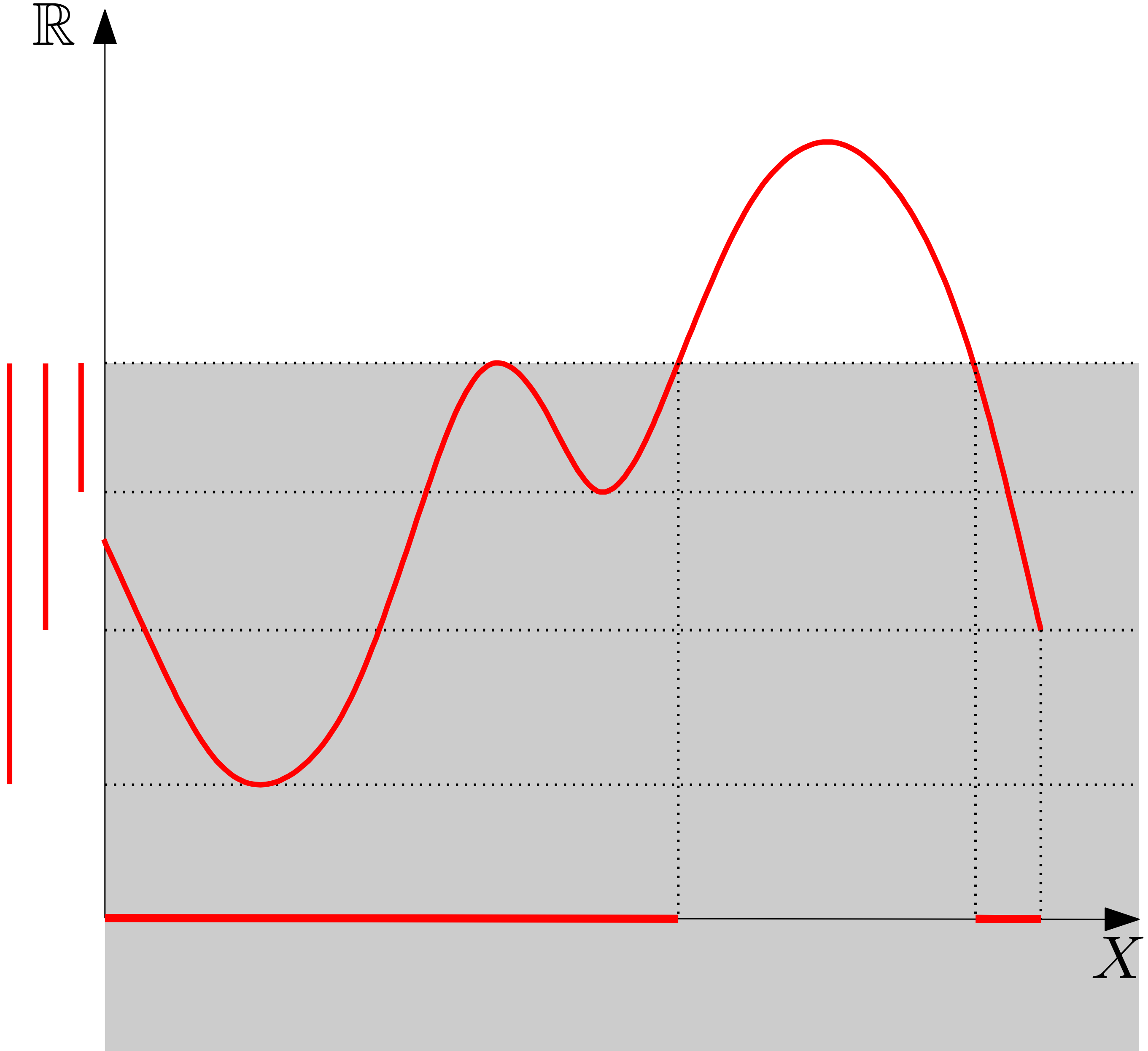
Persistent homology for functions

- 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.



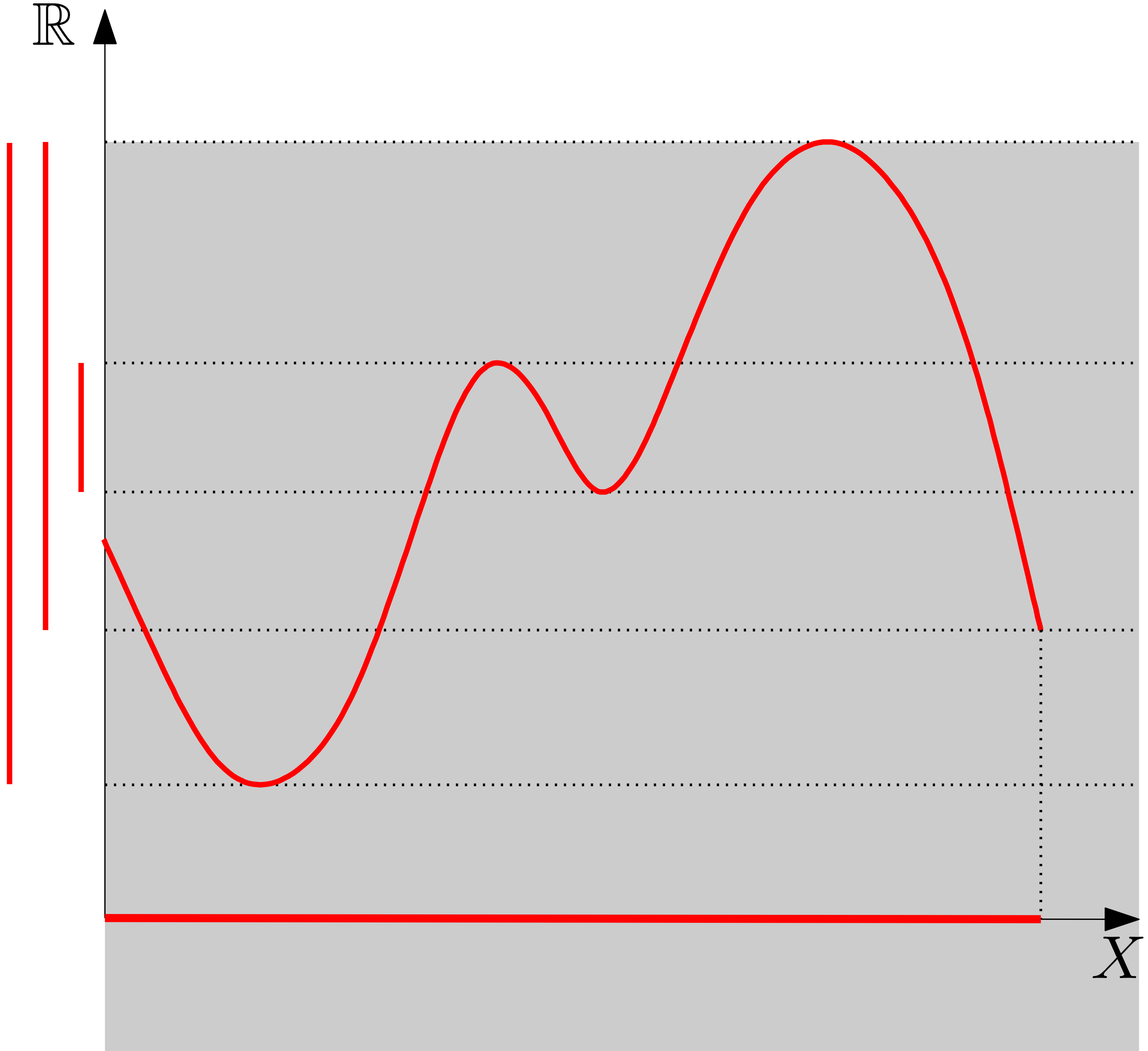
Persistent homology for functions

- 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.



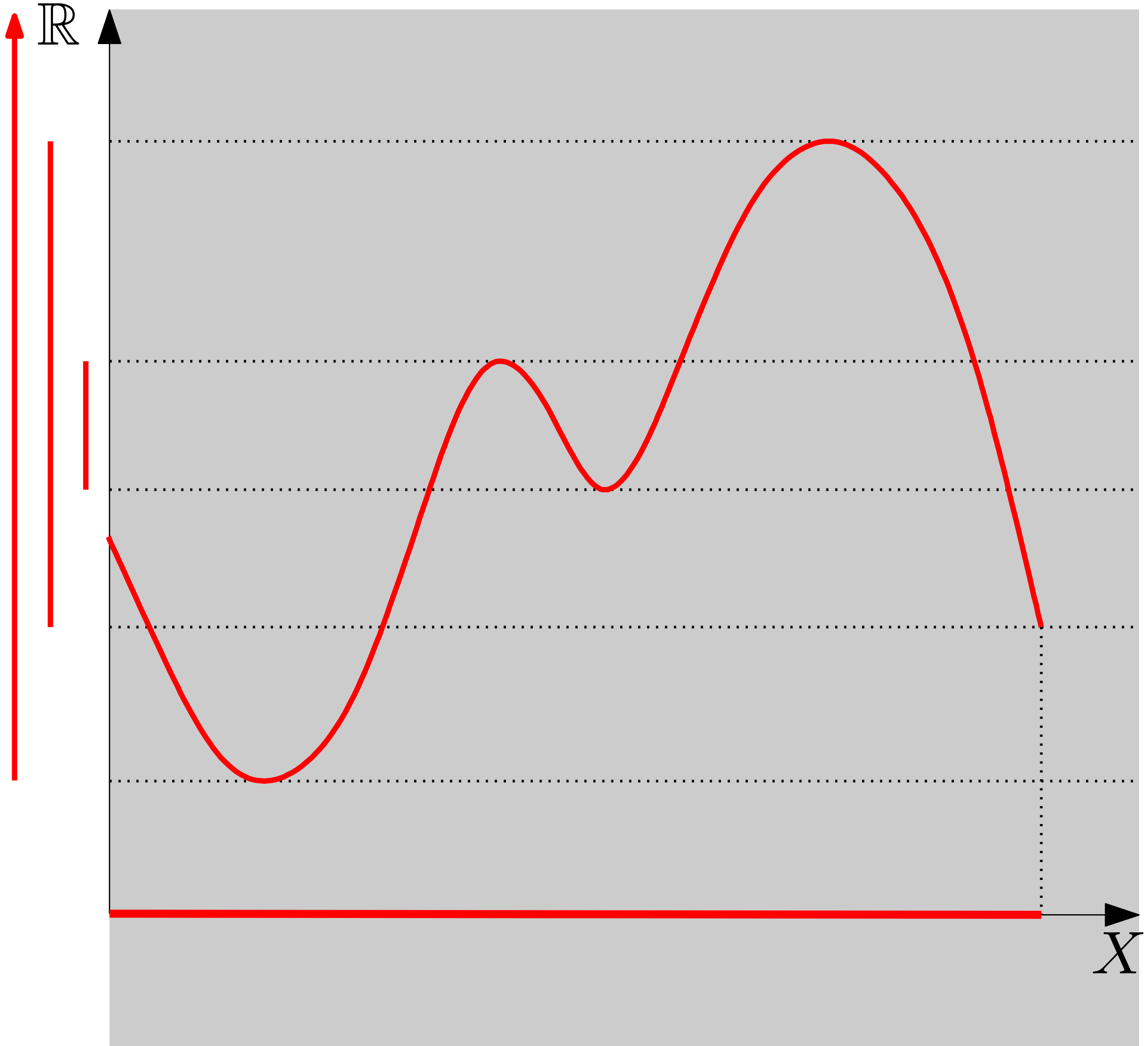
Persistent homology for functions

- 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.



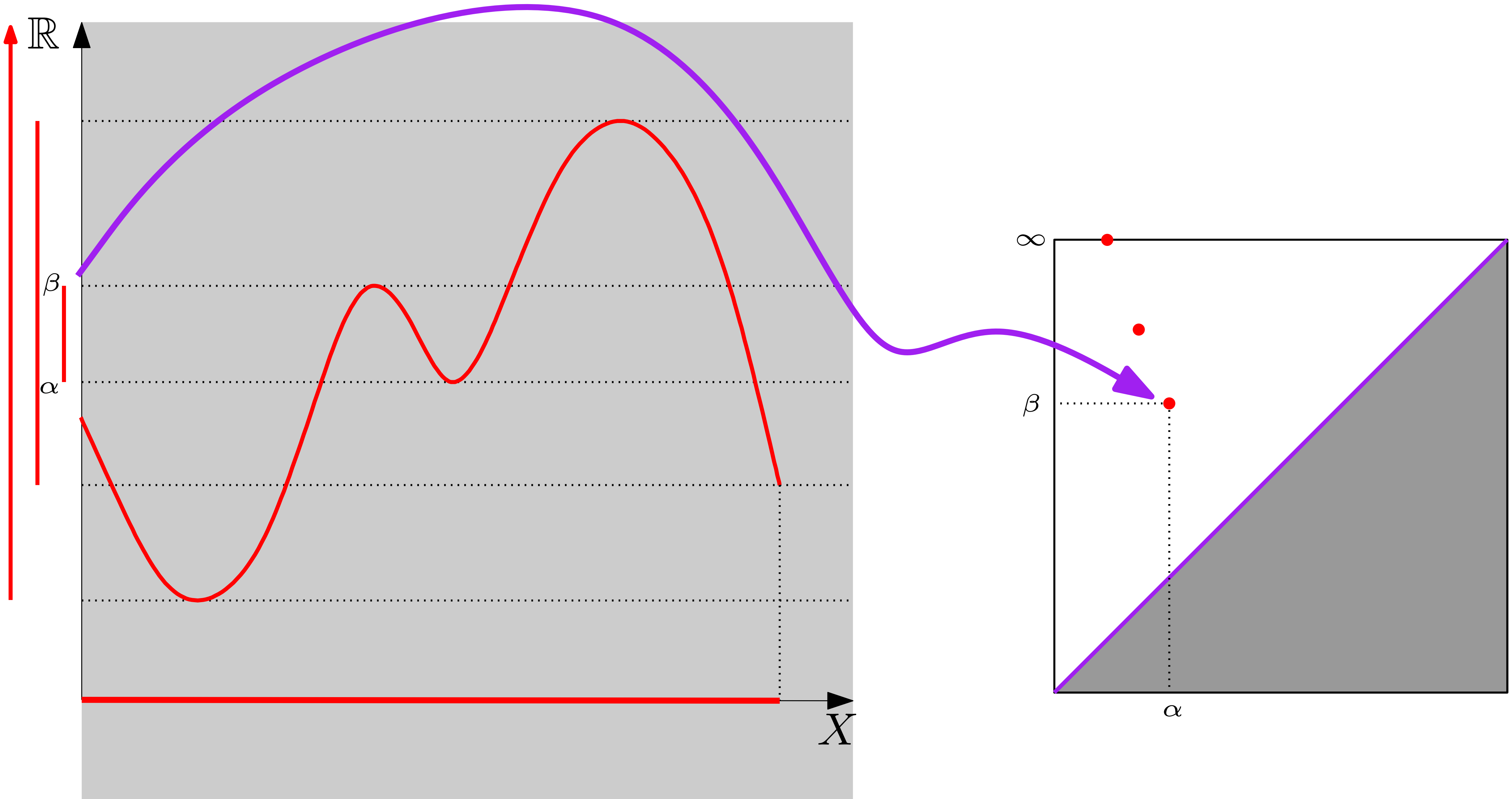
Persistent homology for functions

- 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

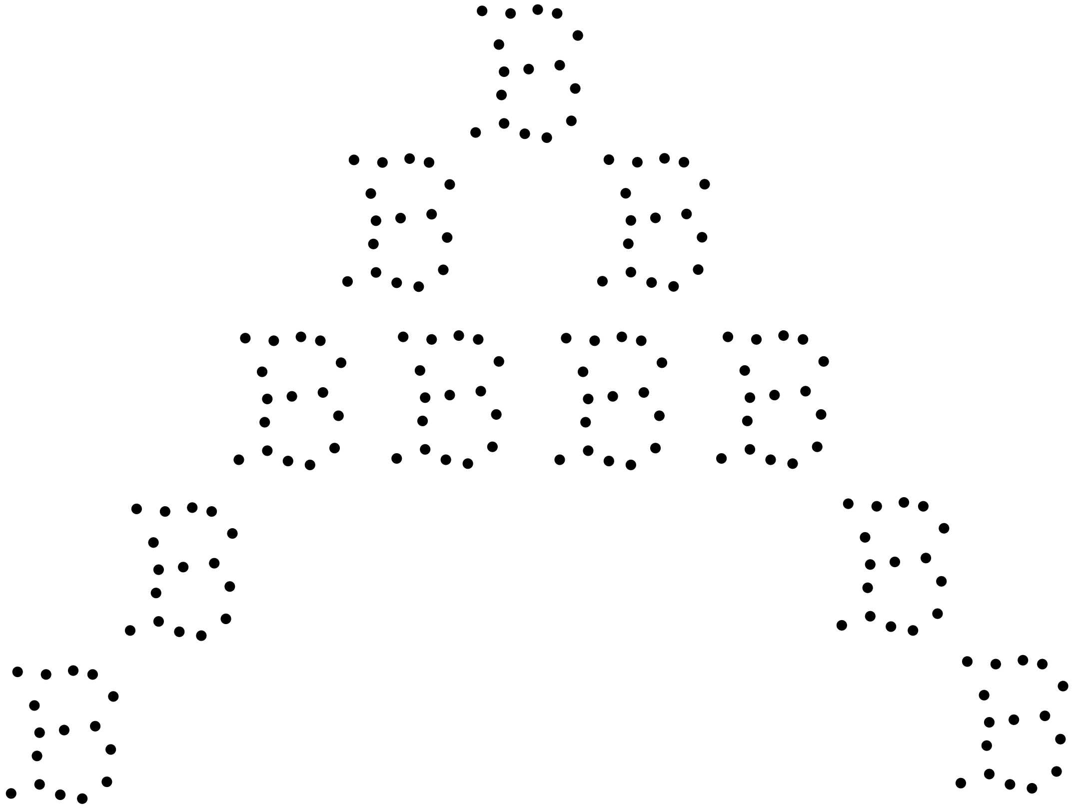
- Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function
- The family of sublevel sets of a function is an example of filtration.
- Finite set of intervals (barcode) encodes births/deaths of topological features.



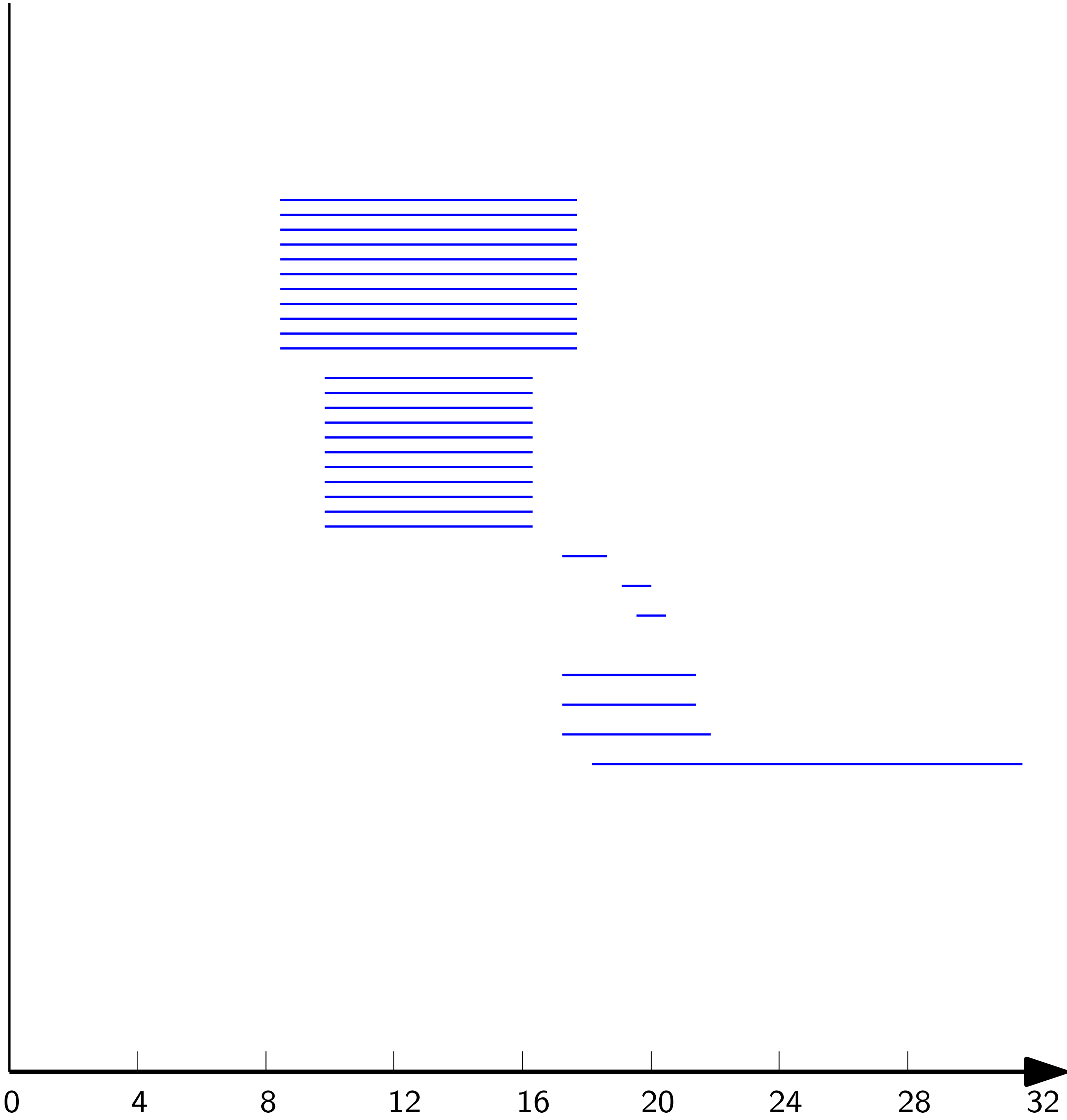
Persistent homology for functions

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

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



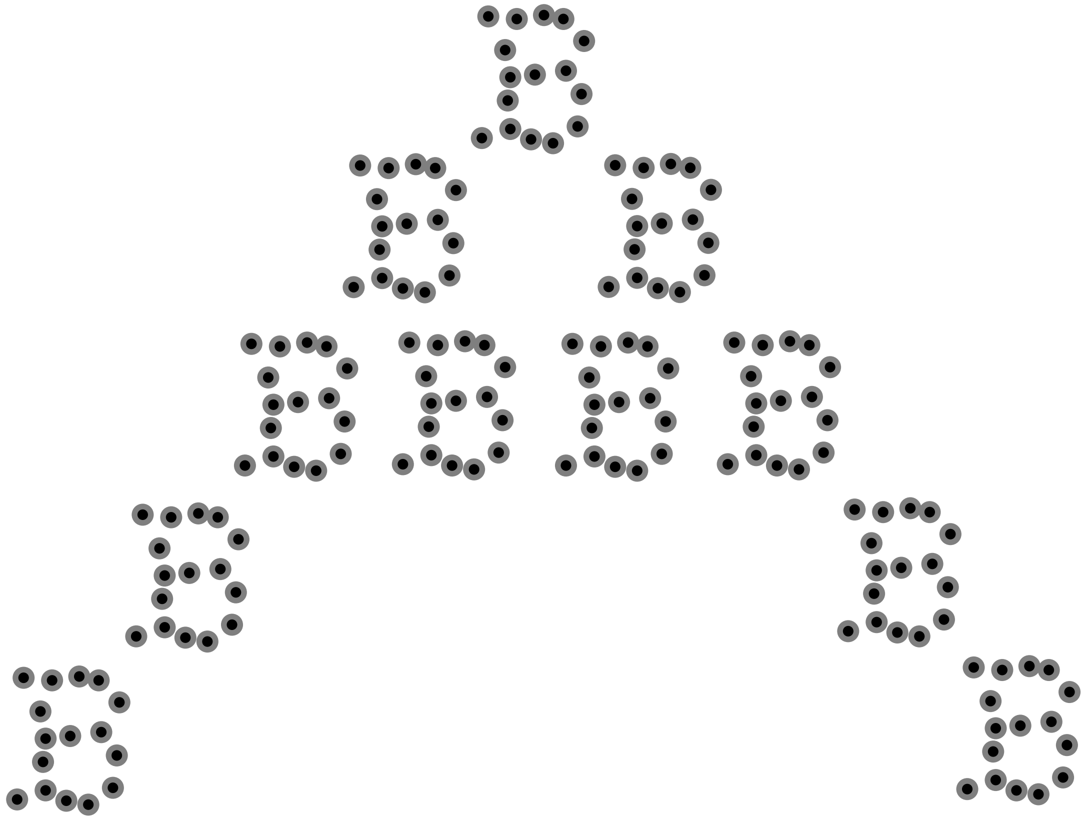
barcode for holes (1-d homology)



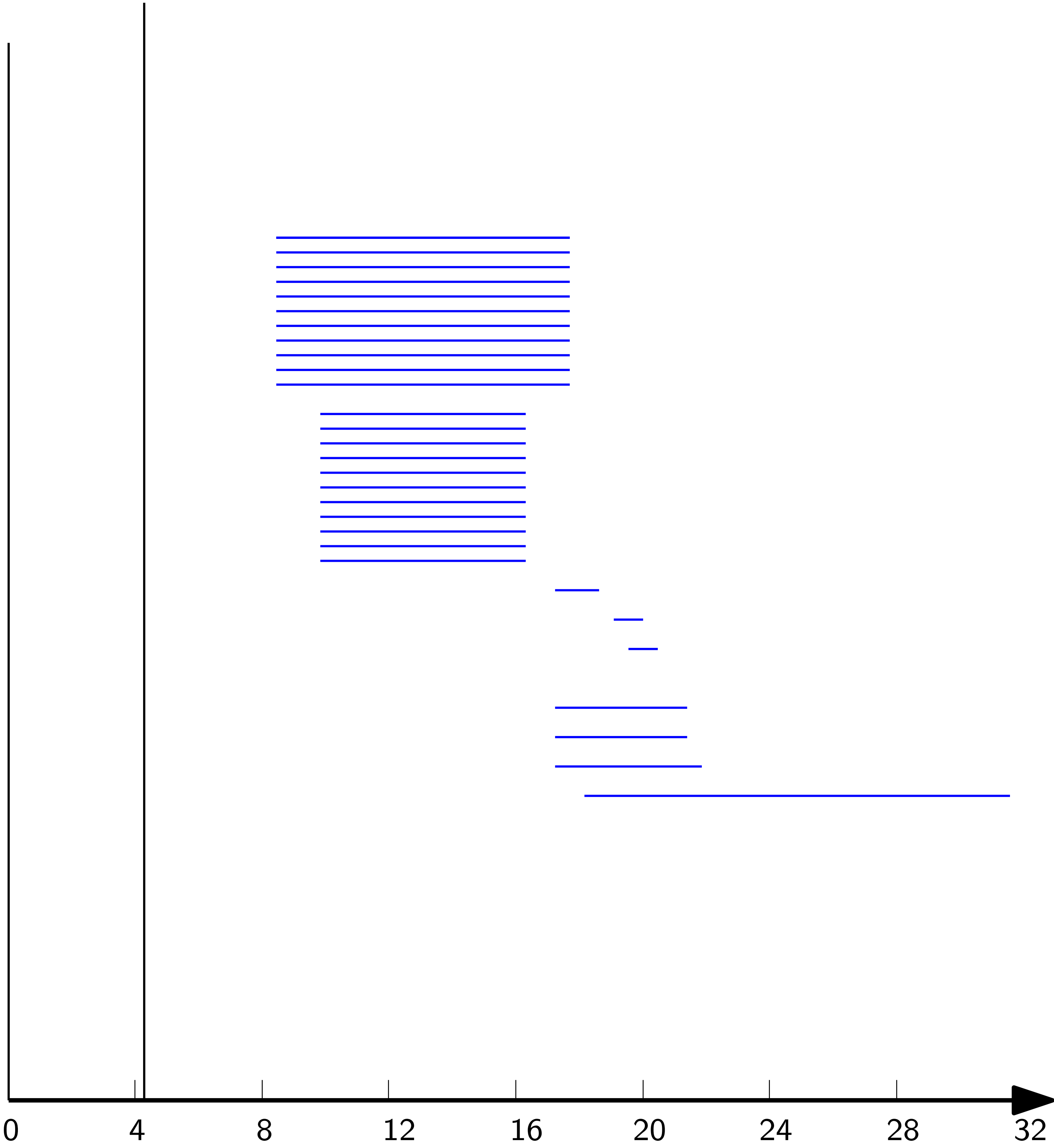
Persistent homology for functions

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

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



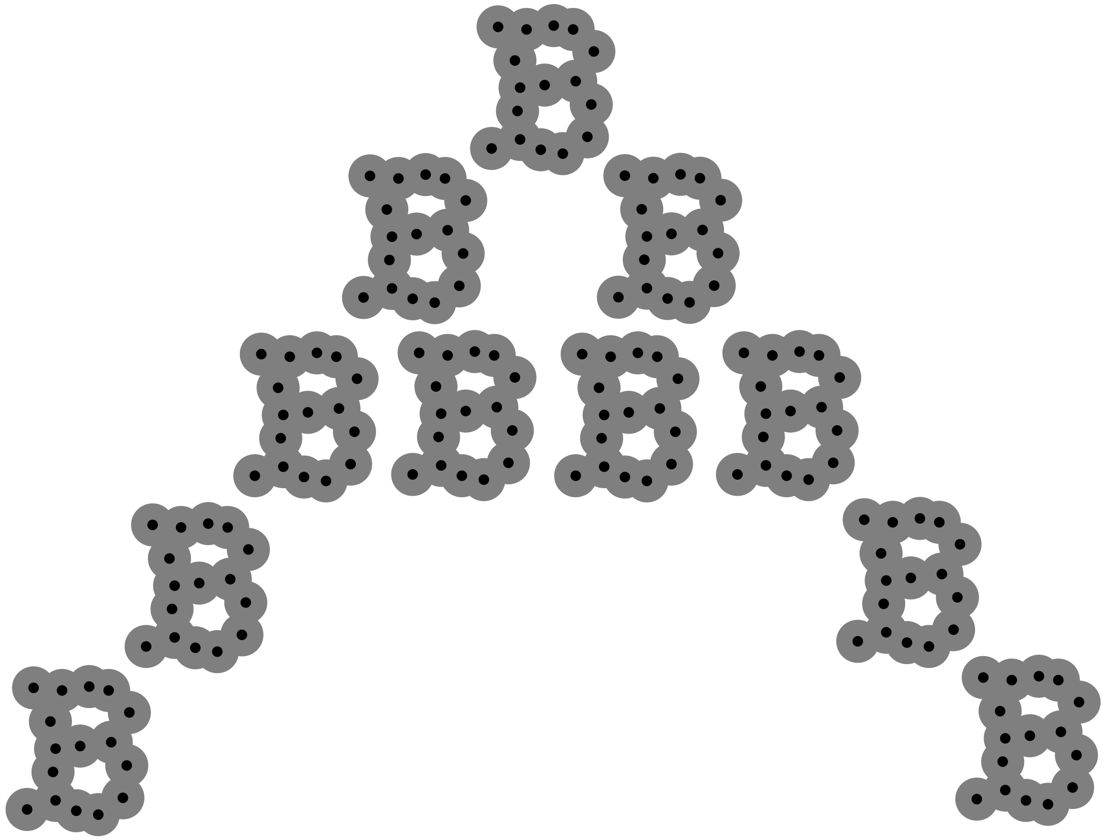
barcode for holes (1-d homology)



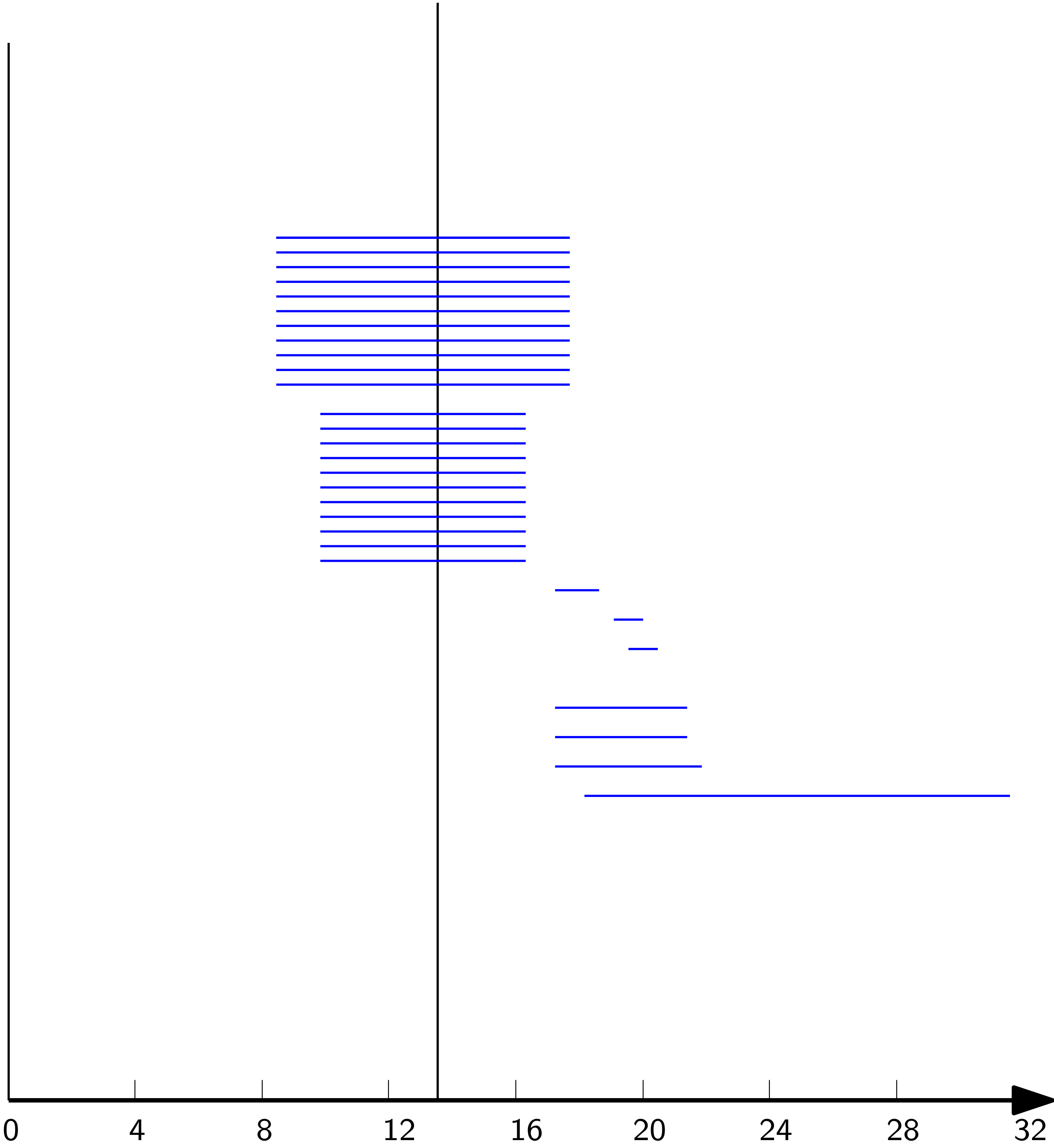
Persistent homology for functions

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

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



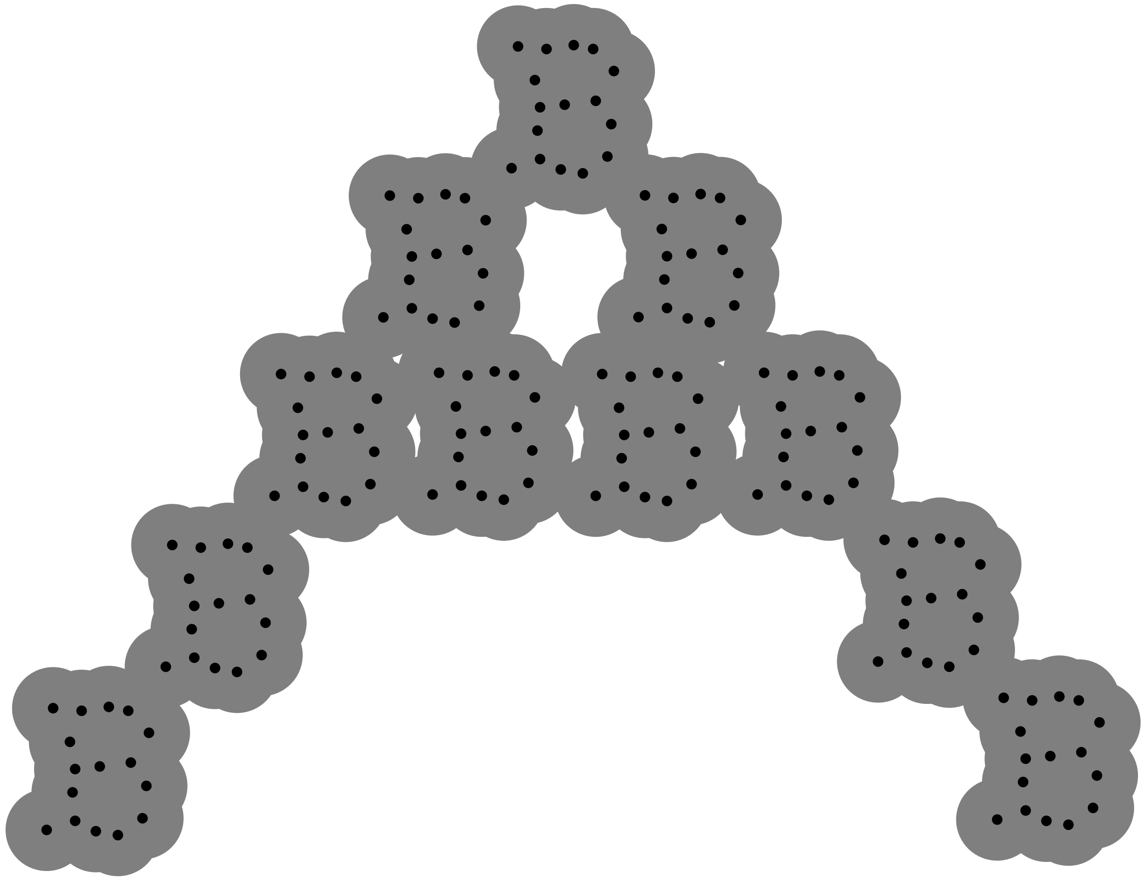
barcode for holes (1-d homology)

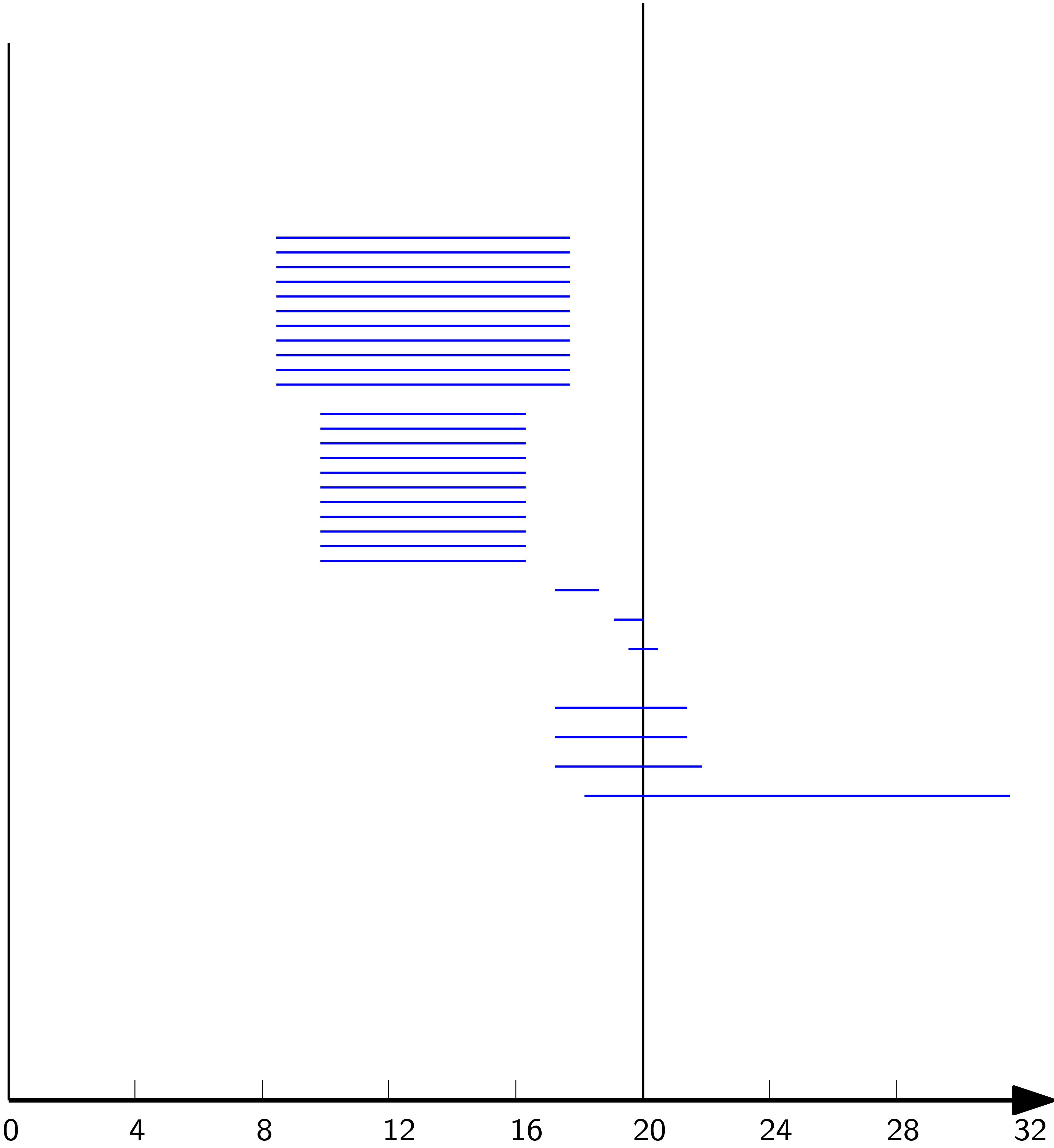


Persistent homology for functions

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

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



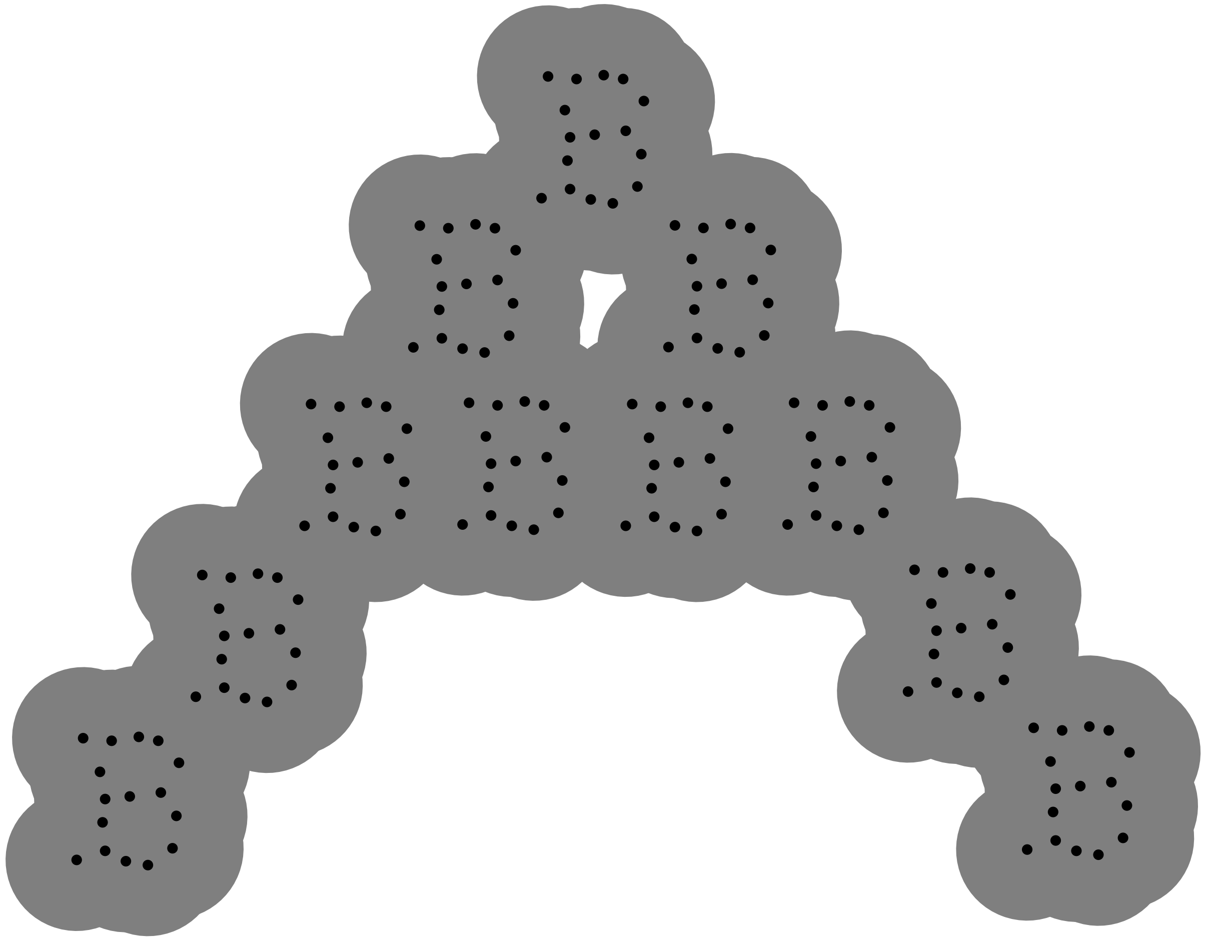


barcode for holes (1-d homology)

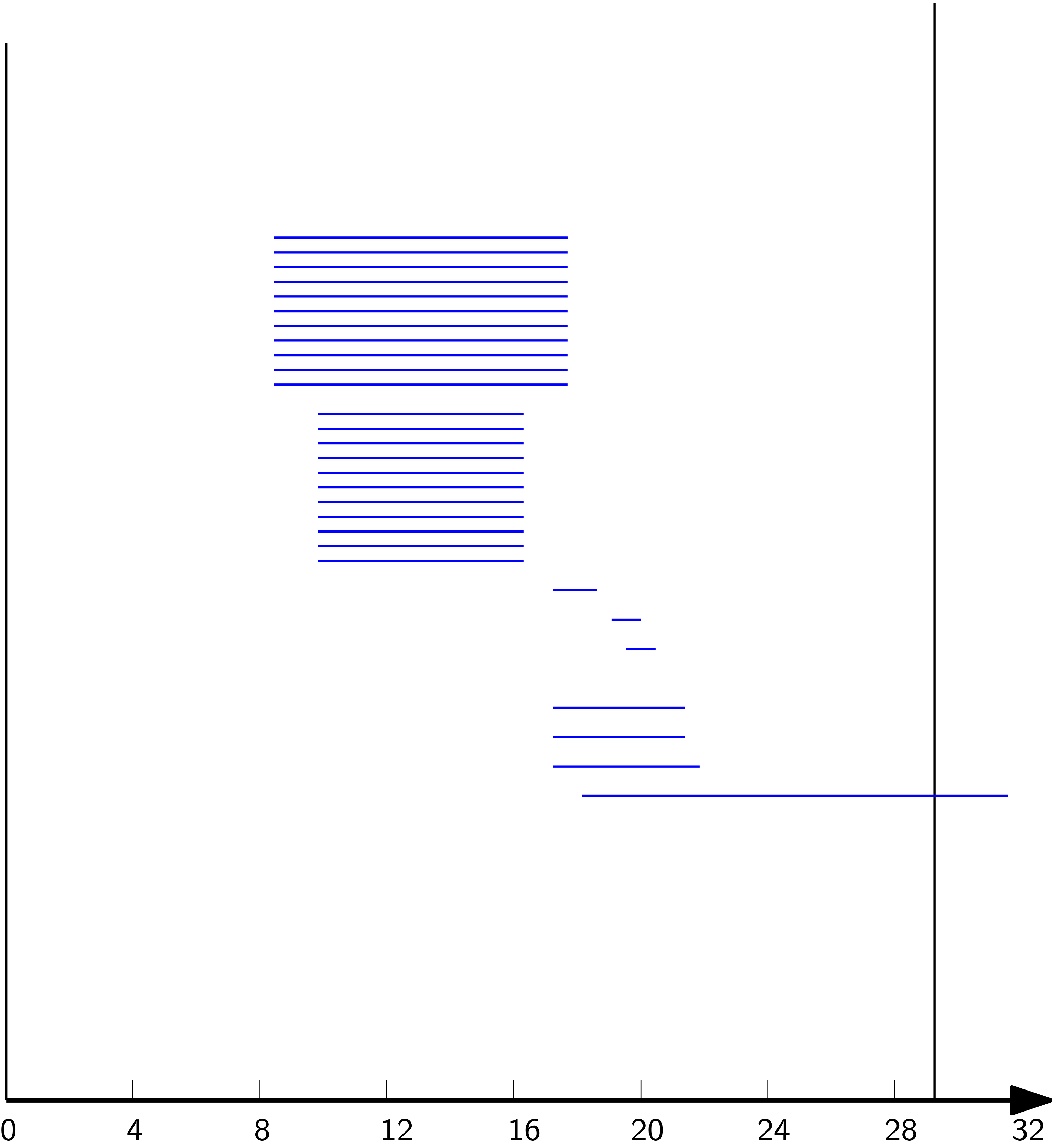
Persistent homology for functions

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

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



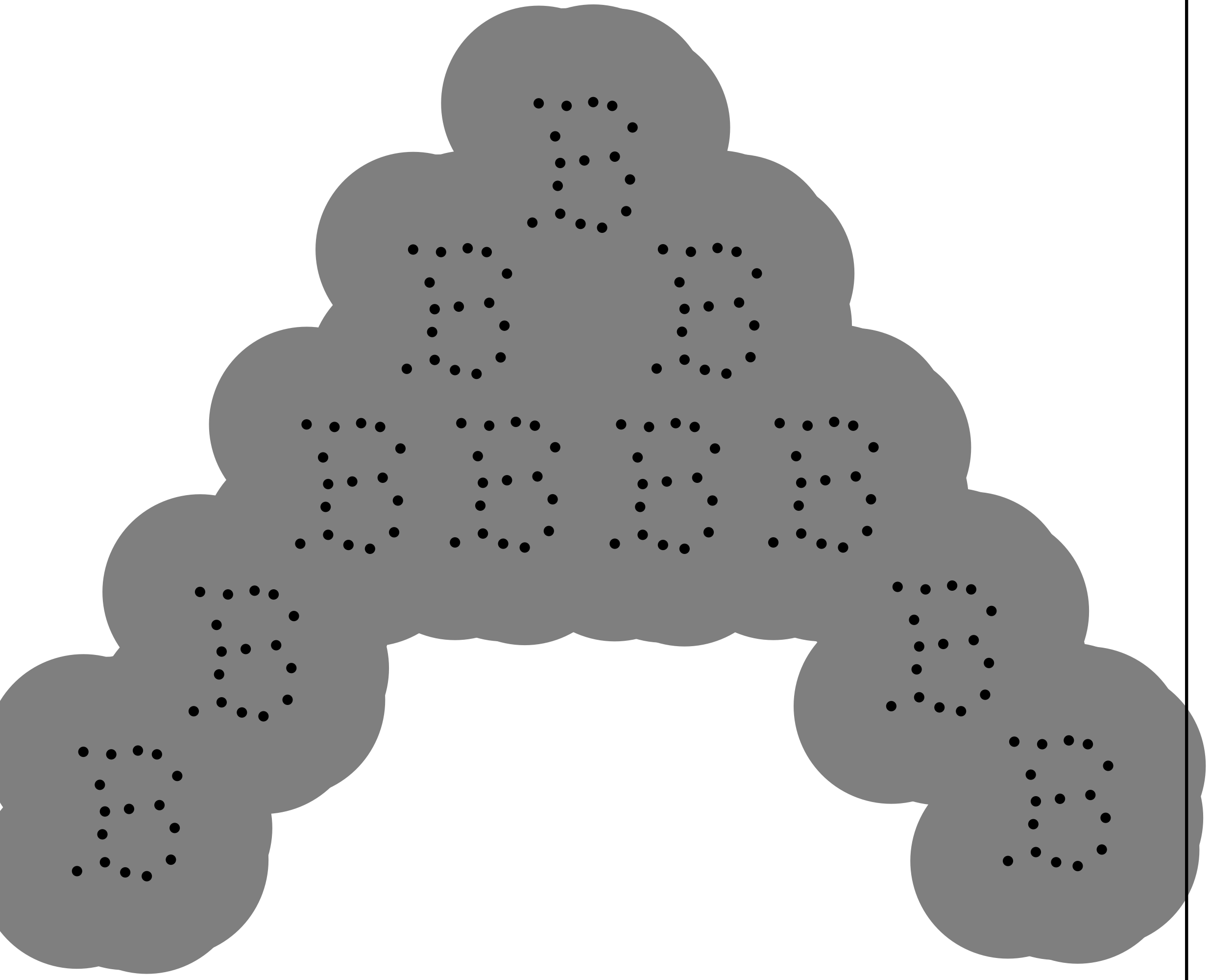
barcode for holes (1-d homology)



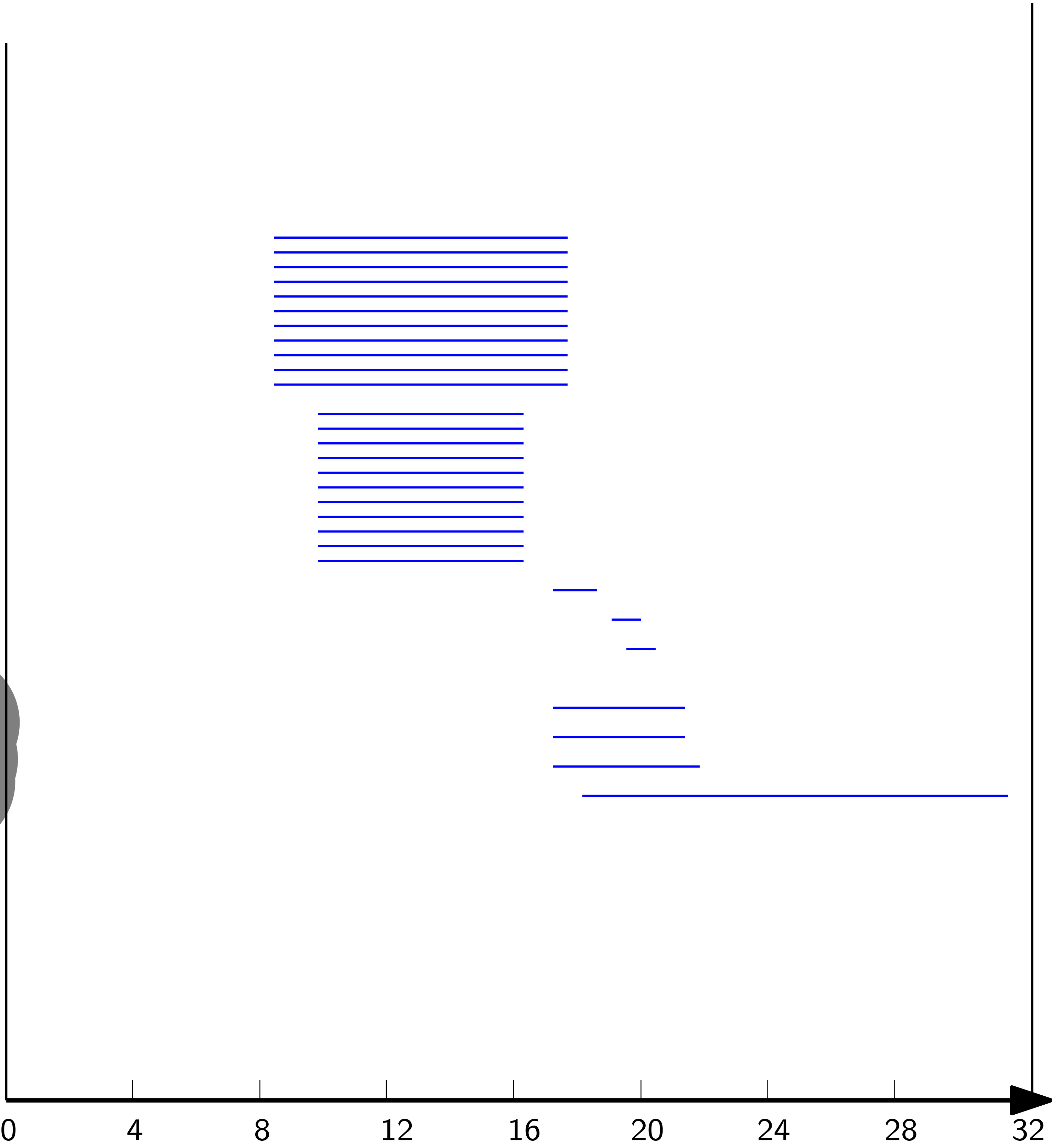
Persistent homology for functions

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

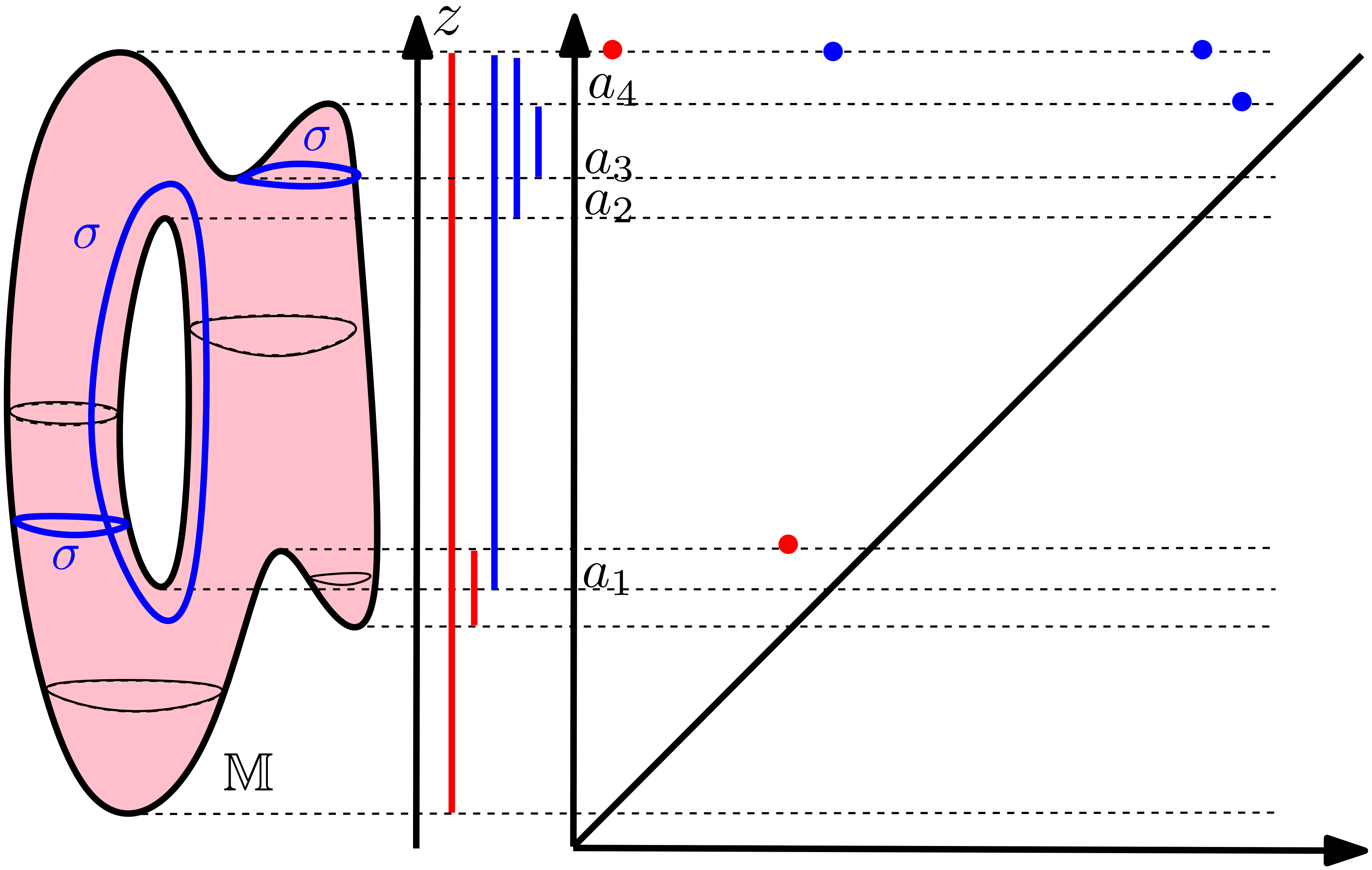
$$x \to \min_{p \in P} \|x - p\|_2$$



barcode for holes (1-d homology)



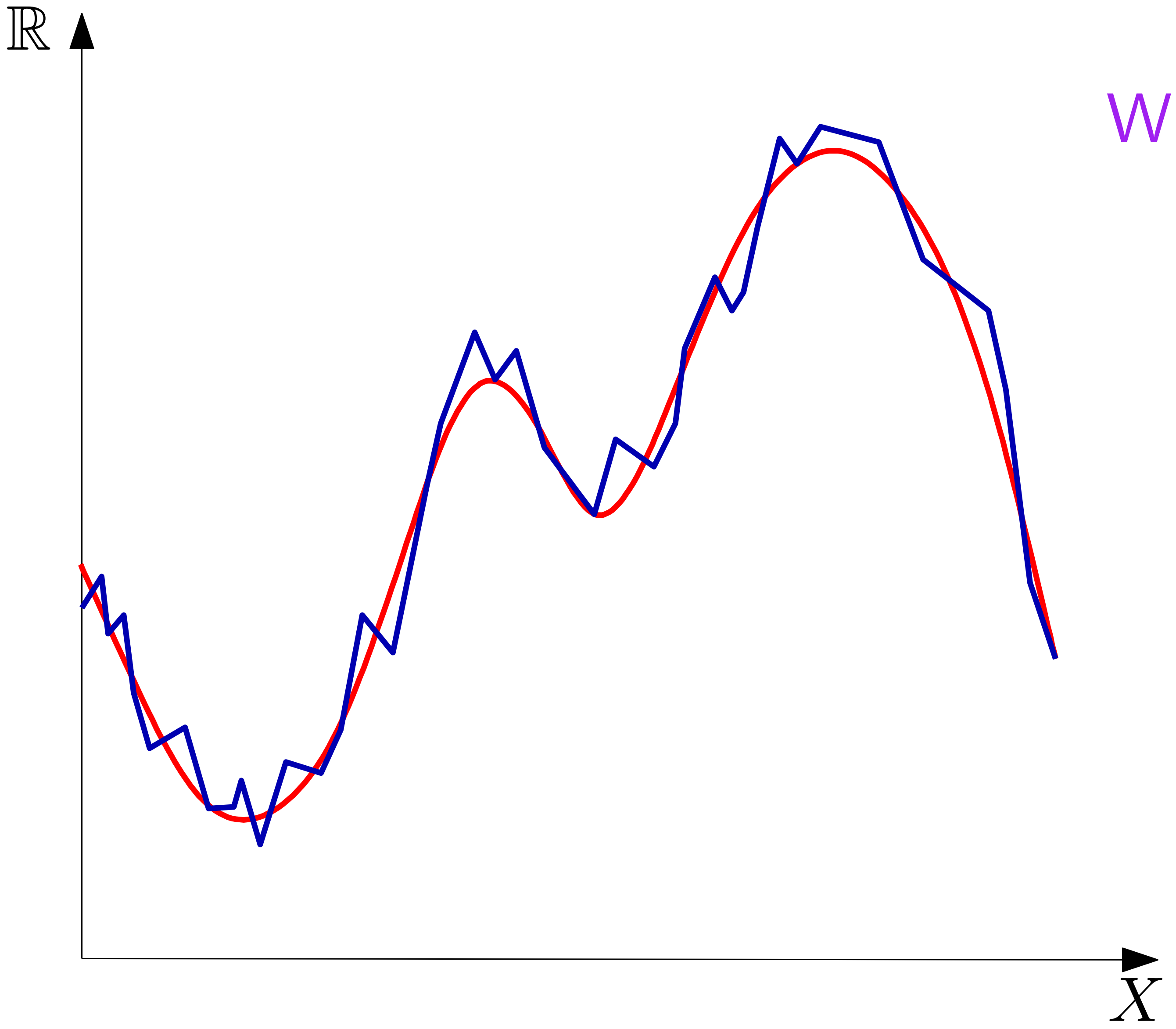
Persistent homology for functions



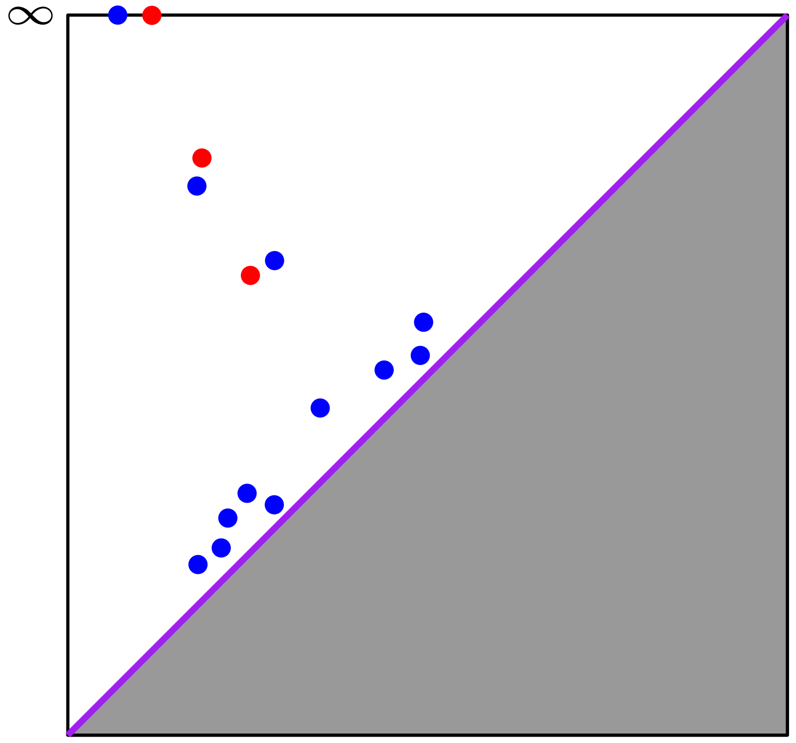
Tracking and encoding the evolution of the connected components (0-dimensional homology) and cycles (1-dimensional homology) of the sublevel sets.

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).

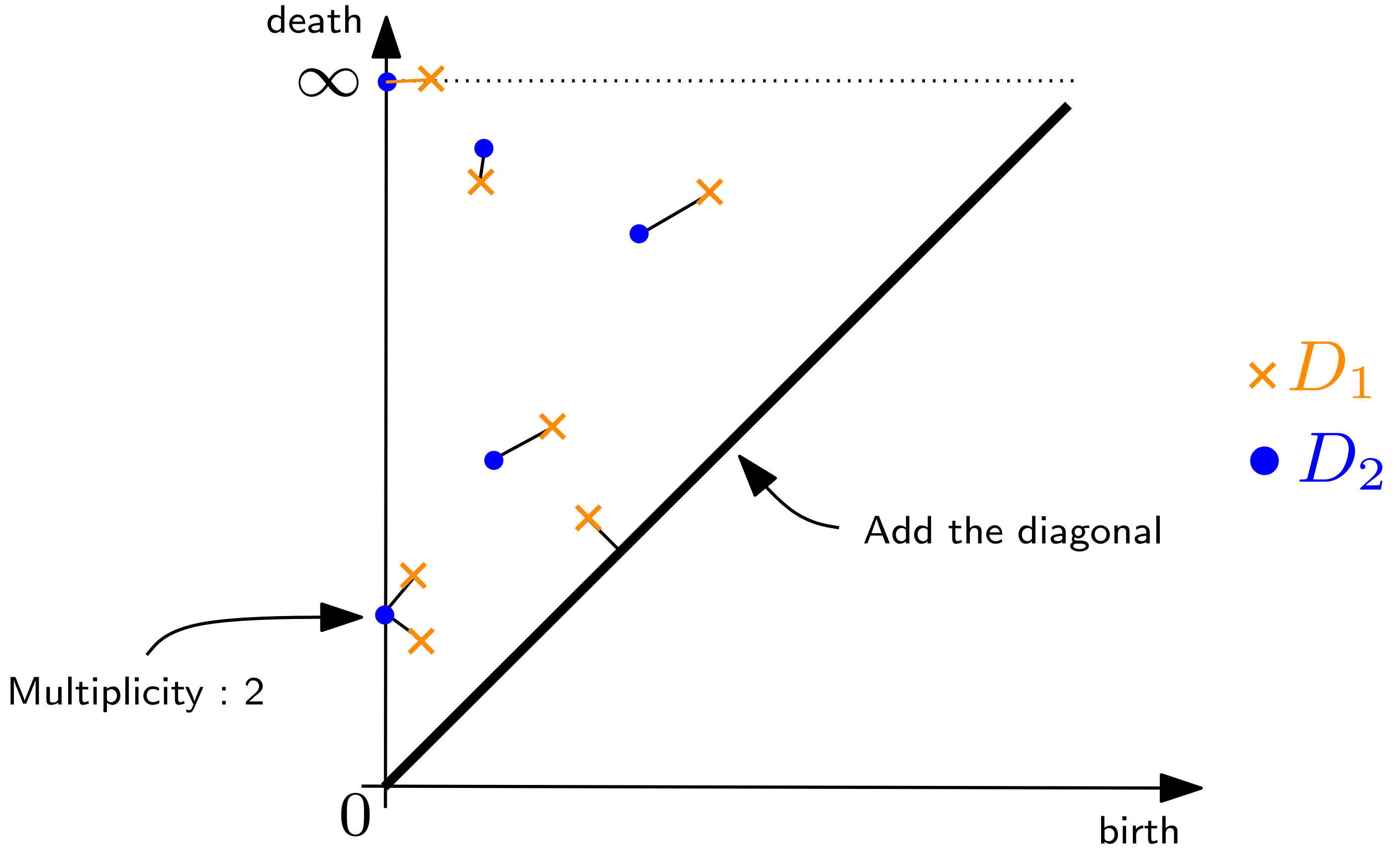
Stability properties



What if f is slightly perturbed ?



Distance between persistence diagrams

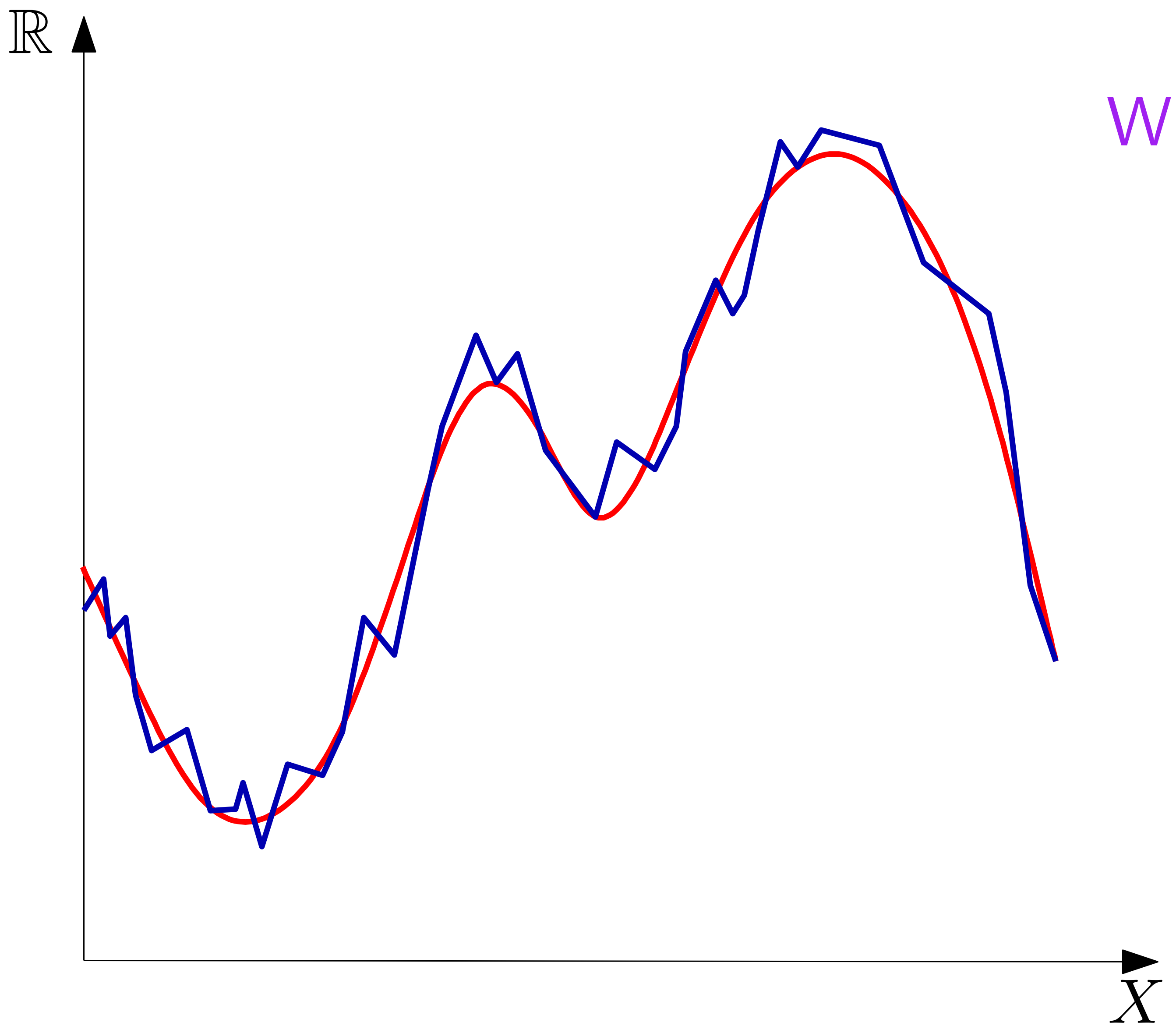


The bottleneck distance between two diagrams D_1 and D_2 is

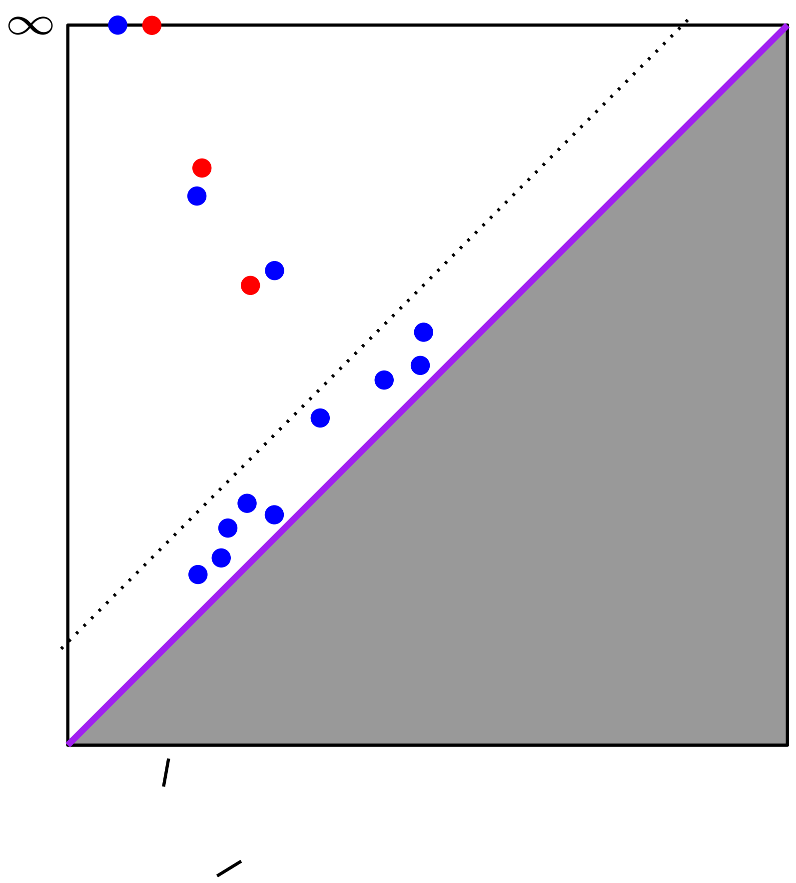
$$d_B(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty$$

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

Stability properties



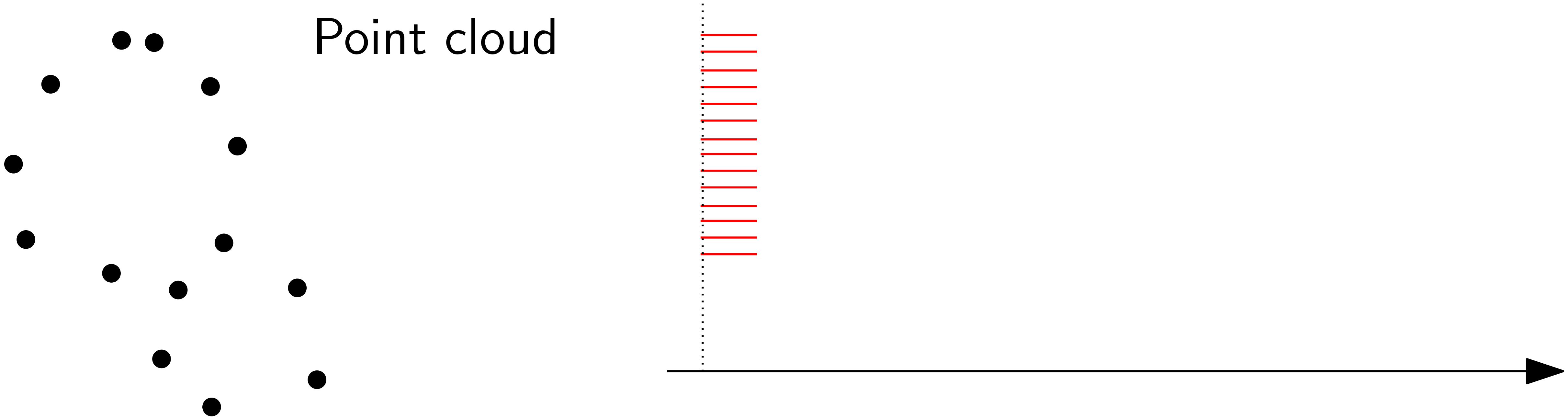
What if f is slightly perturbed ?



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

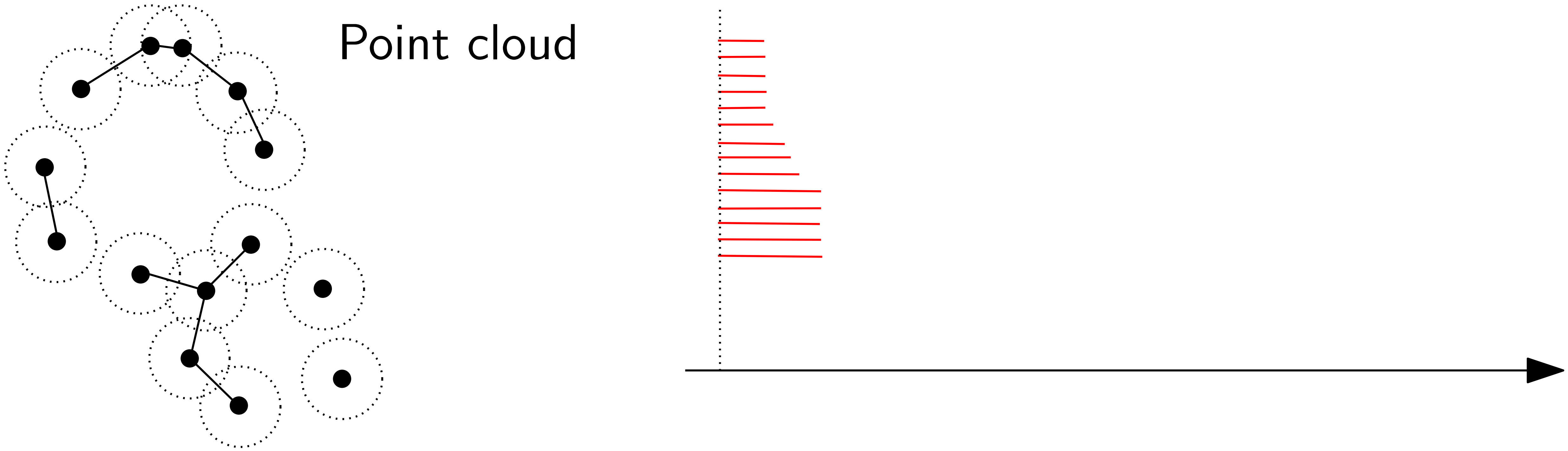
[Baranikov 94], [Cohen-Steiner, Edelsbrunner, Harer 05], [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG 09], [C., de Silva, Glisse, Oudot 12]

Persistent homology for point clouds



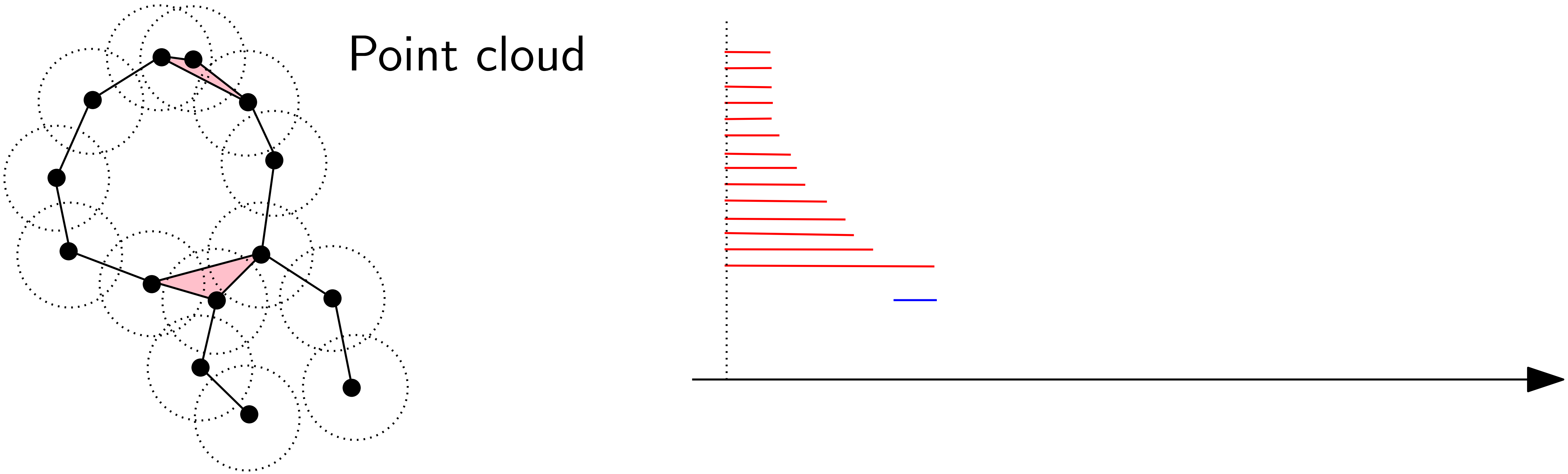
- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology : encode the evolution of the topology across the scales → multi-scale topological signatures.
- A general and well-studied mathematical framework.

Persistent homology for point clouds



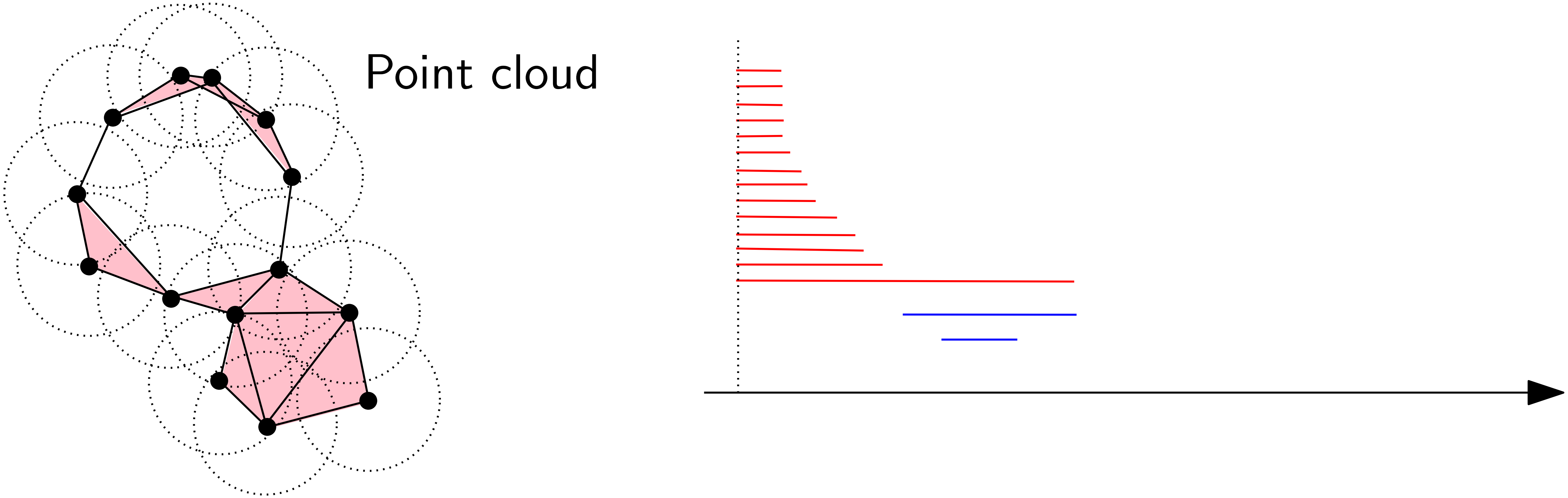
- Filtrations allow to construct “shapes” representing the data in a multiscale way.
- Persistent homology : encode the evolution of the topology across the scales → multi-scale topological signatures.
- A general and well-studied mathematical framework.

Persistent homology for point clouds



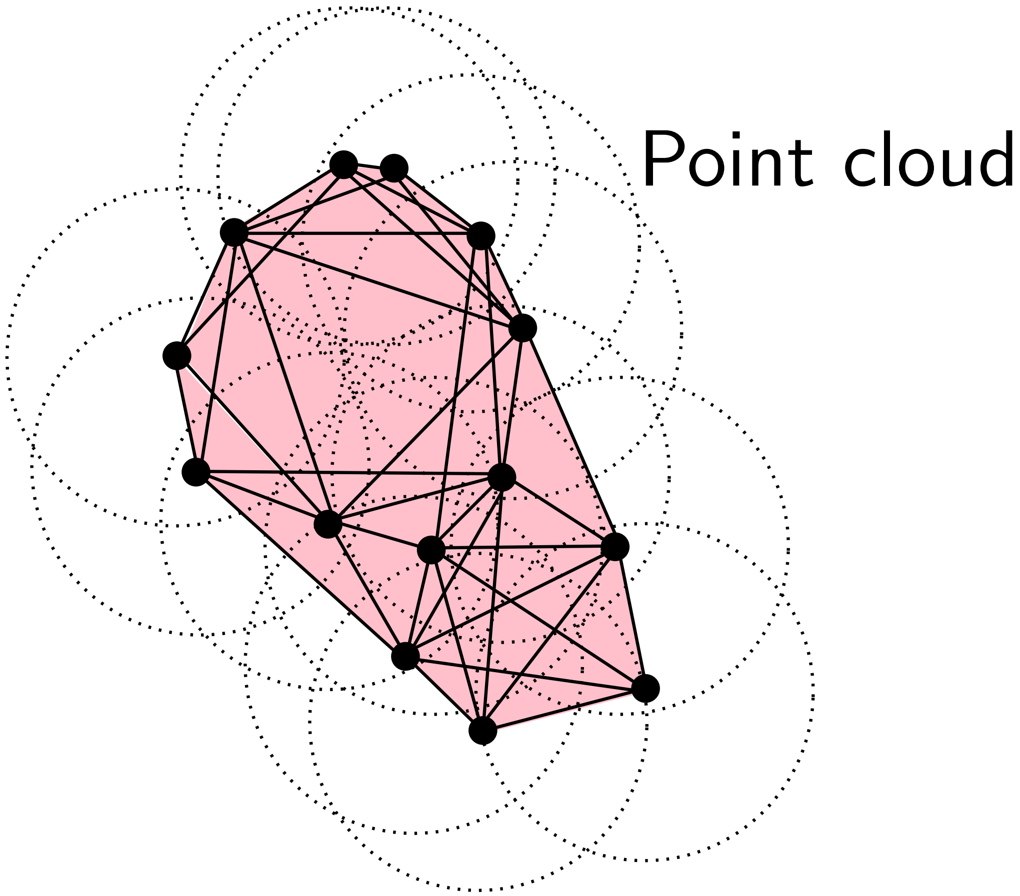
- Filtrations allow to construct “shapes” representing the data in a multiscale way.
- Persistent homology : encode the evolution of the topology across the scales → multi-scale topological signatures.
- A general and well-studied mathematical framework.

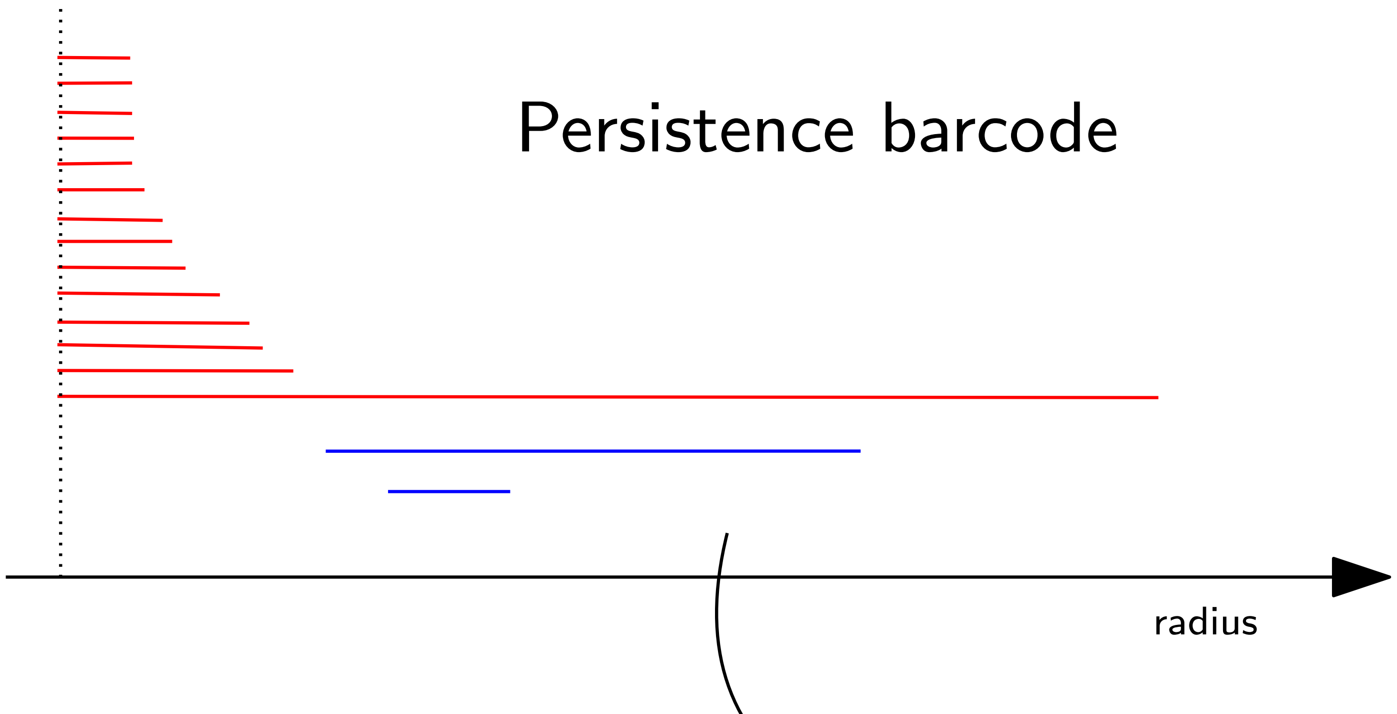
Persistent homology for point clouds

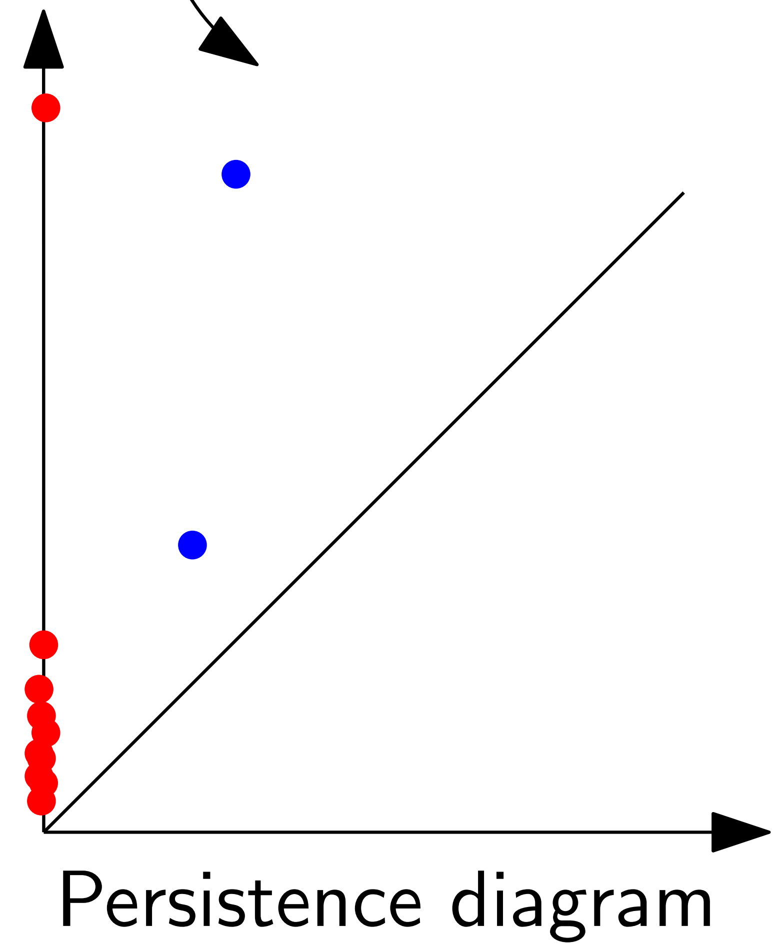


- Filtrations allow to construct “shapes” representing the data in a multiscale way.
- Persistent homology : encode the evolution of the topology across the scales → multi-scale topological signatures.
- A general and well-studied mathematical framework.

Persistent homology for point clouds



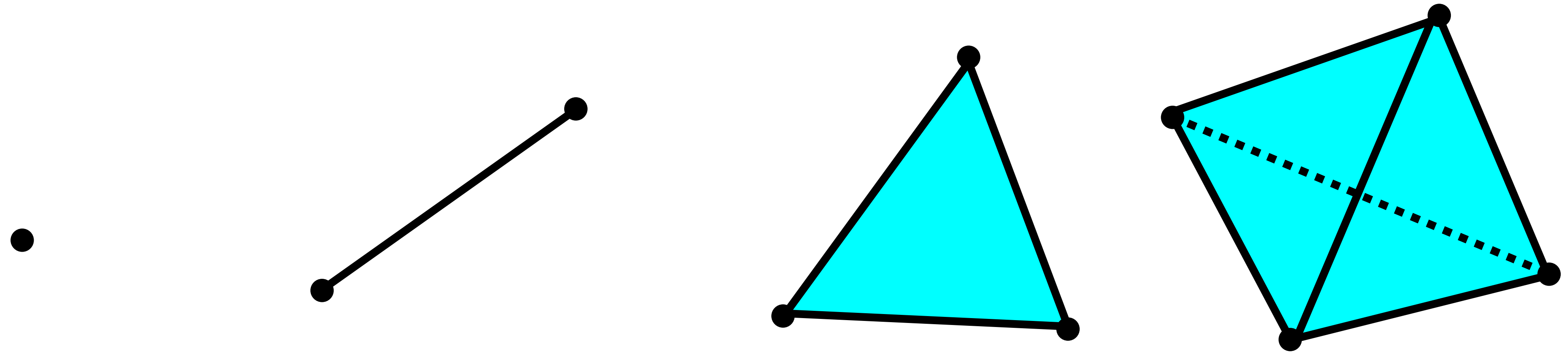




- Filtrations allow to construct “shapes” representing the data in a multiscale way.
- Persistent homology : encode the evolution of the topology across the scales → multi-scale topological signatures.
- A general and well-studied mathematical framework.

Simplicial complexes, filtrations, homology and persistent homology

Simplicial complexes



0-simplex : vertex

1-simplex : edge

2-simplex : triangle

3-simplex : tetrahedron

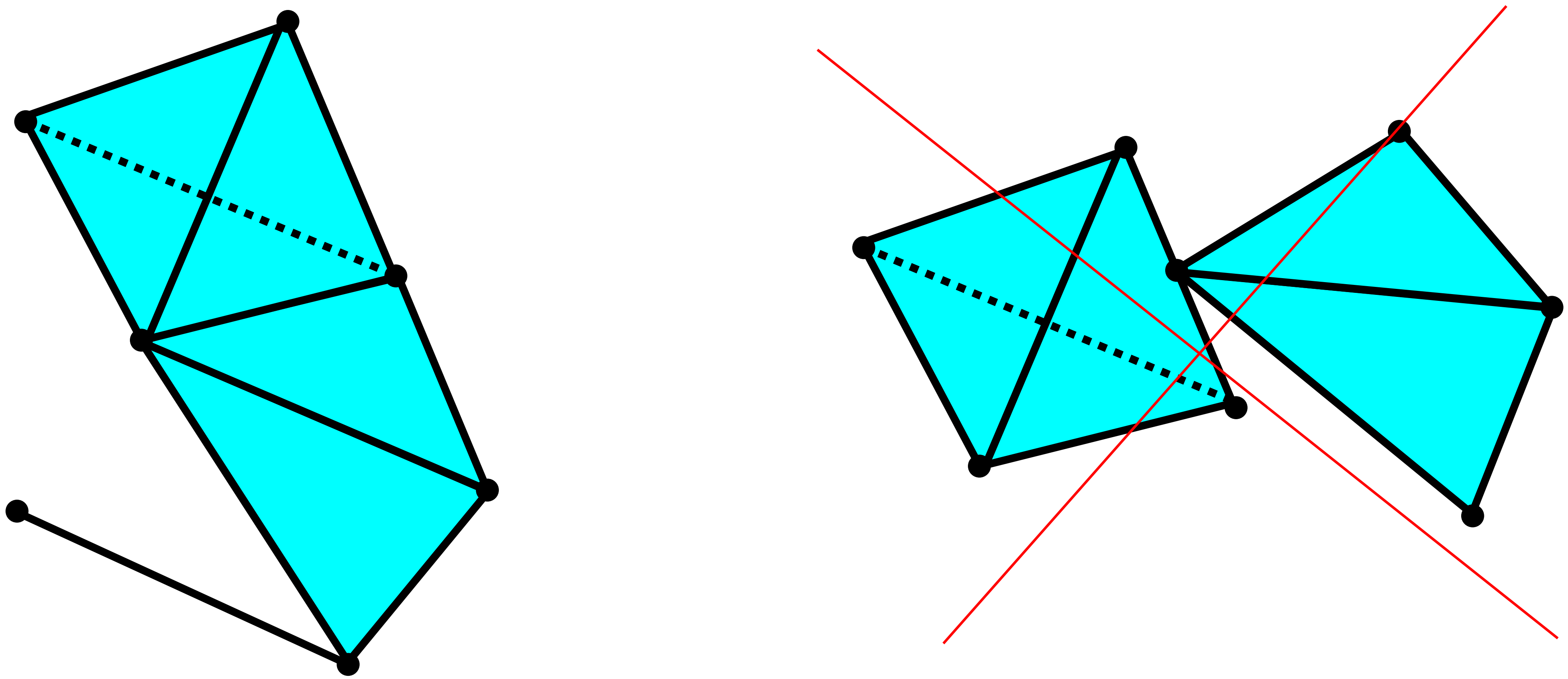
etc...

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

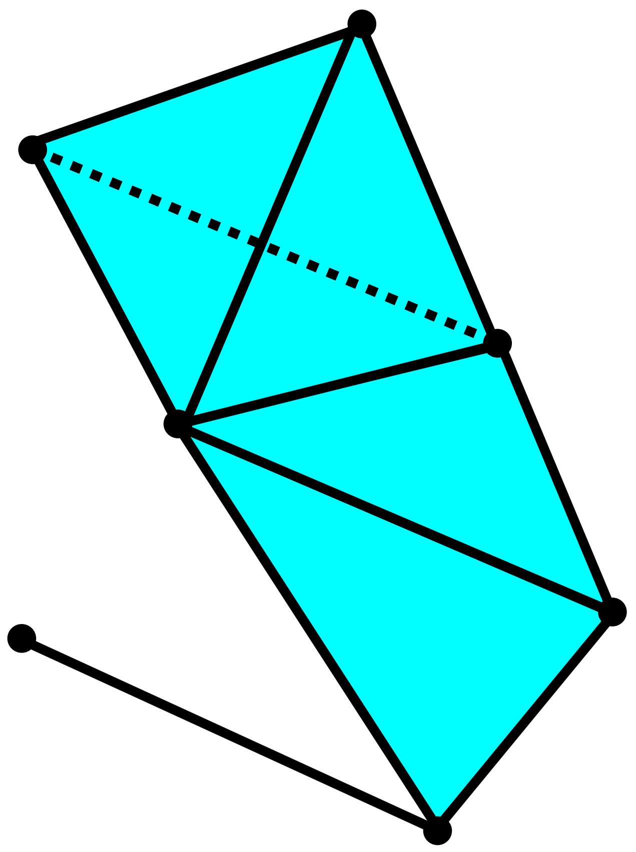


A (finite) simplicial complex K in $\mathbb{R}^d$ is a (finite) collection of simplices such that :

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 $\sigma \in K$.

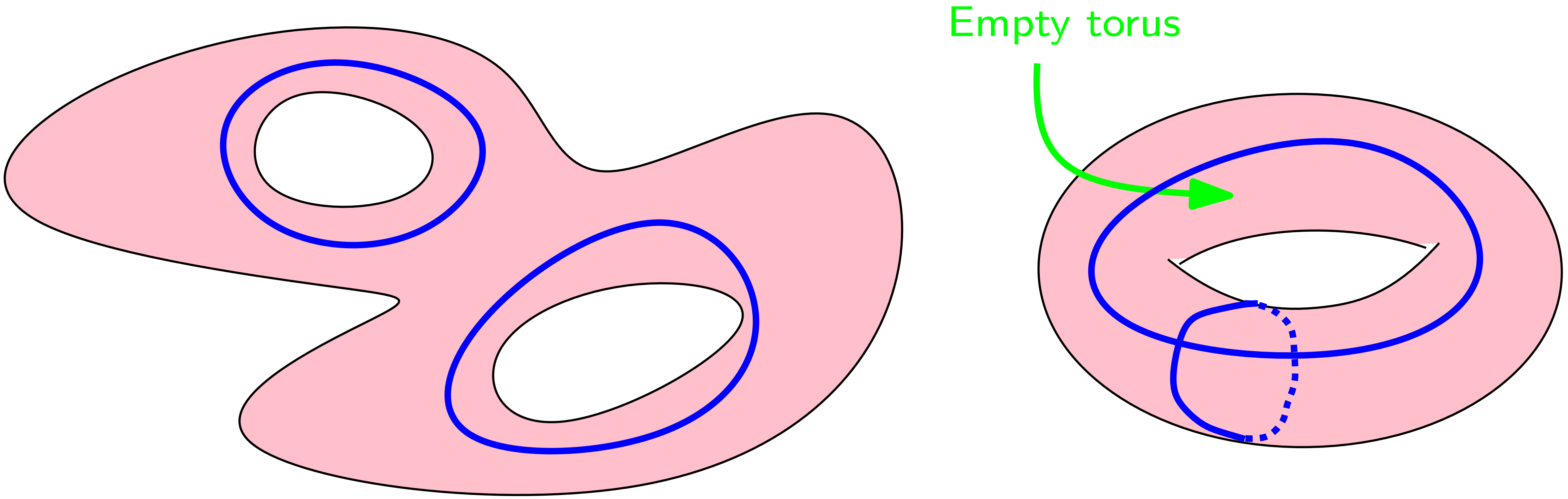
The elements of K are the simplices.

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 (here we will restrict to simplicial complexes).



- 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

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

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

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

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

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 $\mathbb{Z}/2\mathbb{Z}$)

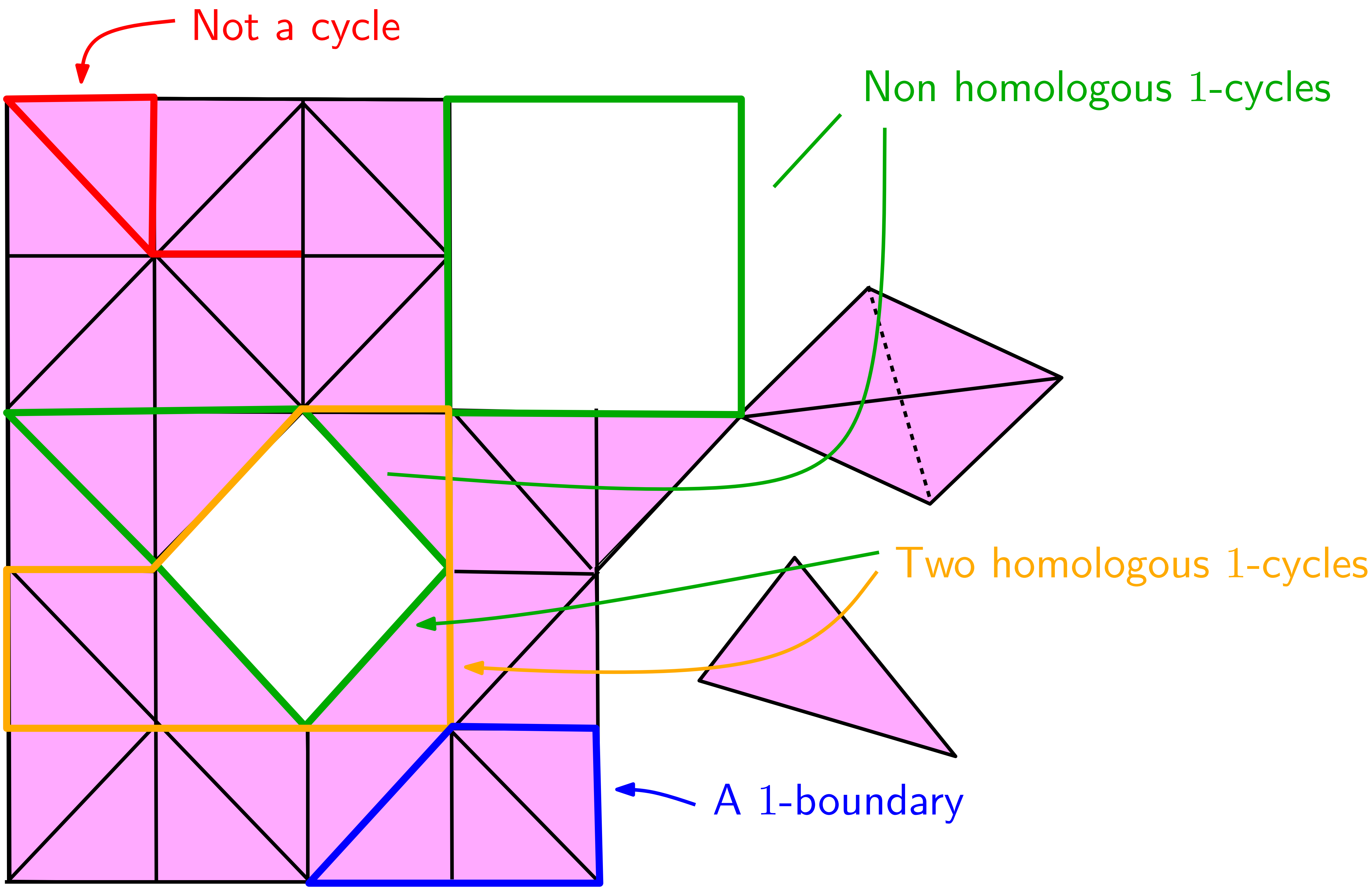
Homology groups and Betti numbers :

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

- 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 : $\beta_k(K) = \dim(H_k(K))$.

Remark : $\beta_0(K) =$ number of connected components of K.

Cycles and boundaries



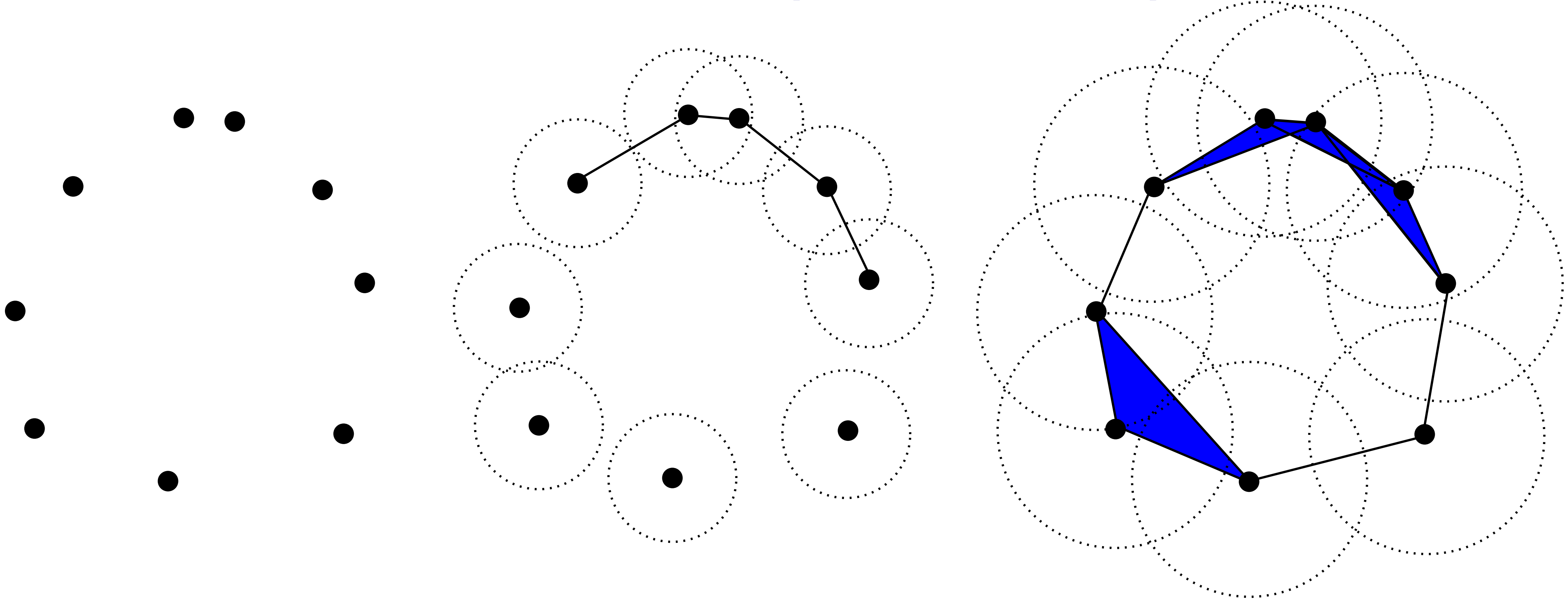
Not a cycle

Non homologous 1-cycles

Two homologous 1-cycles

A 1-boundary

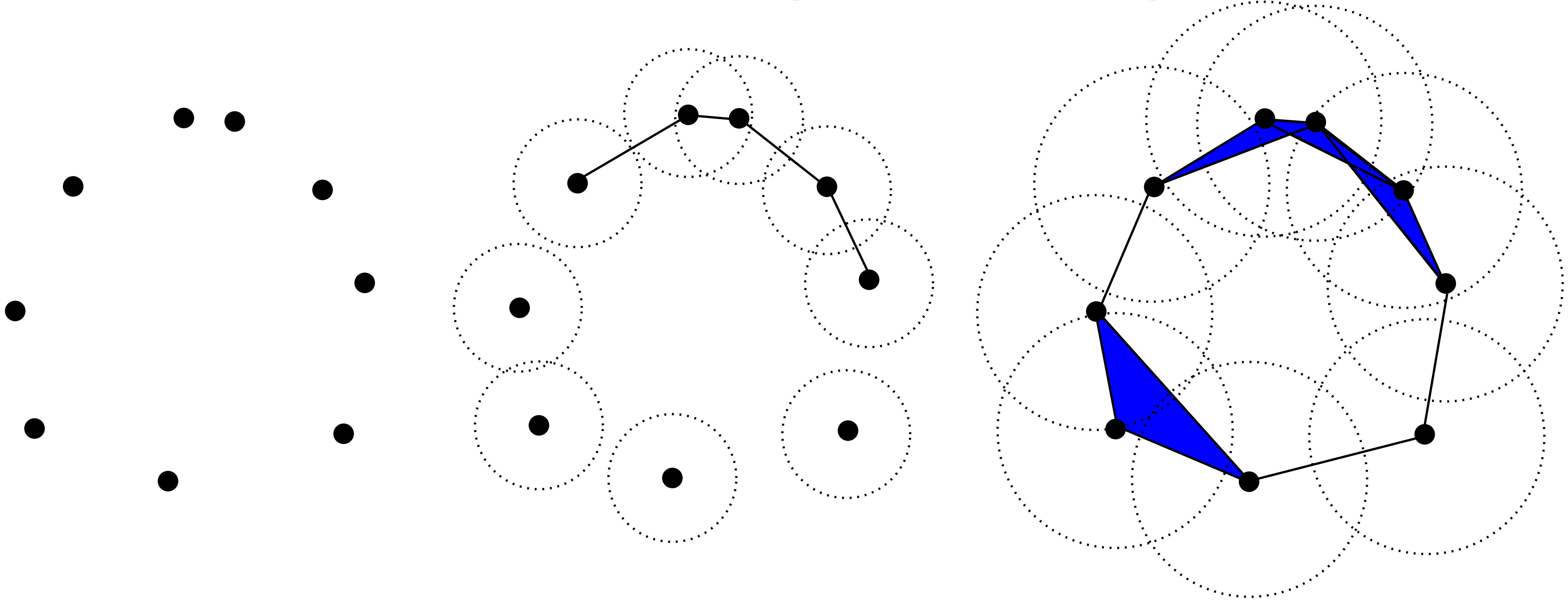
Filtrations of simplicial complexes



- A filtered simplicial complex (or a filtration) $\mathbb{K}$ built on top of a set X is a family $(K_a \mid a \in \mathbf{T})$, $\mathbf{T} \subseteq \mathbb{R}$, of subcomplexes of some fixed simplicial complex K with vertex set X s. t. $K_a \subseteq K_b$ 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.

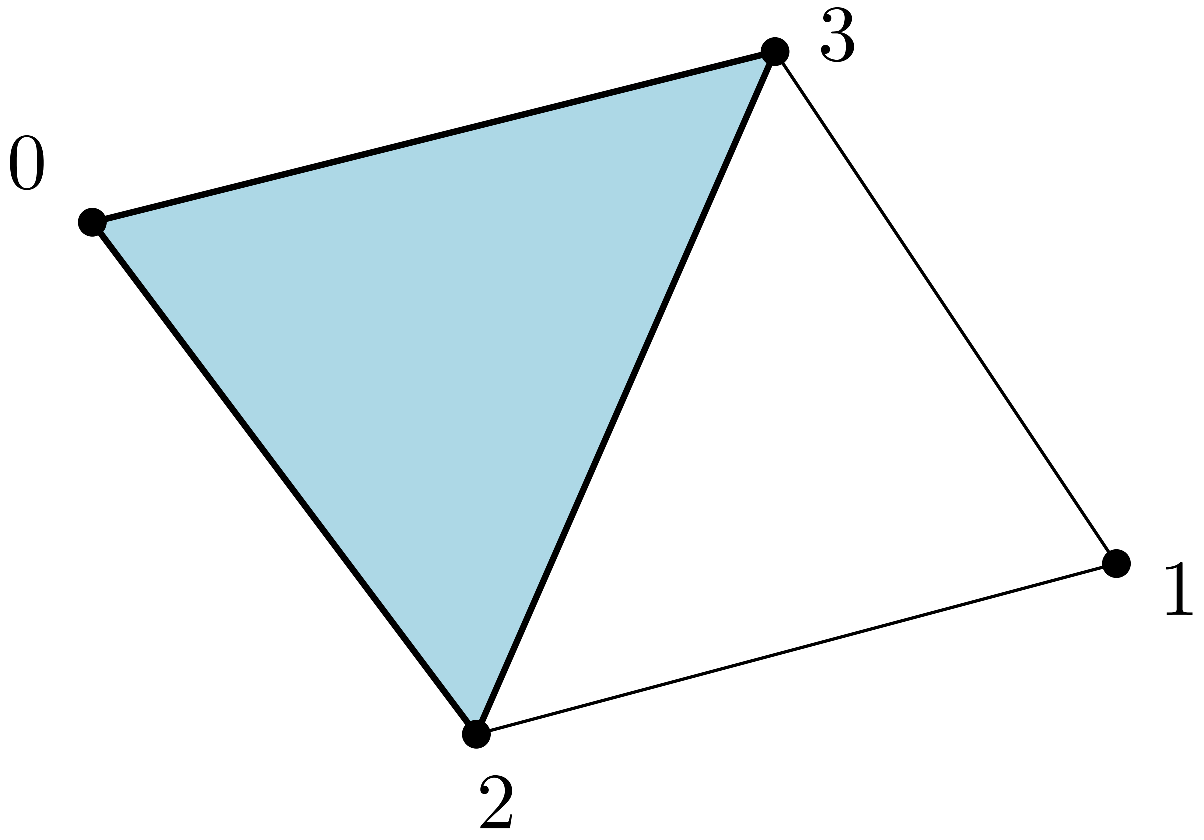
Filtrations of simplicial complexes



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

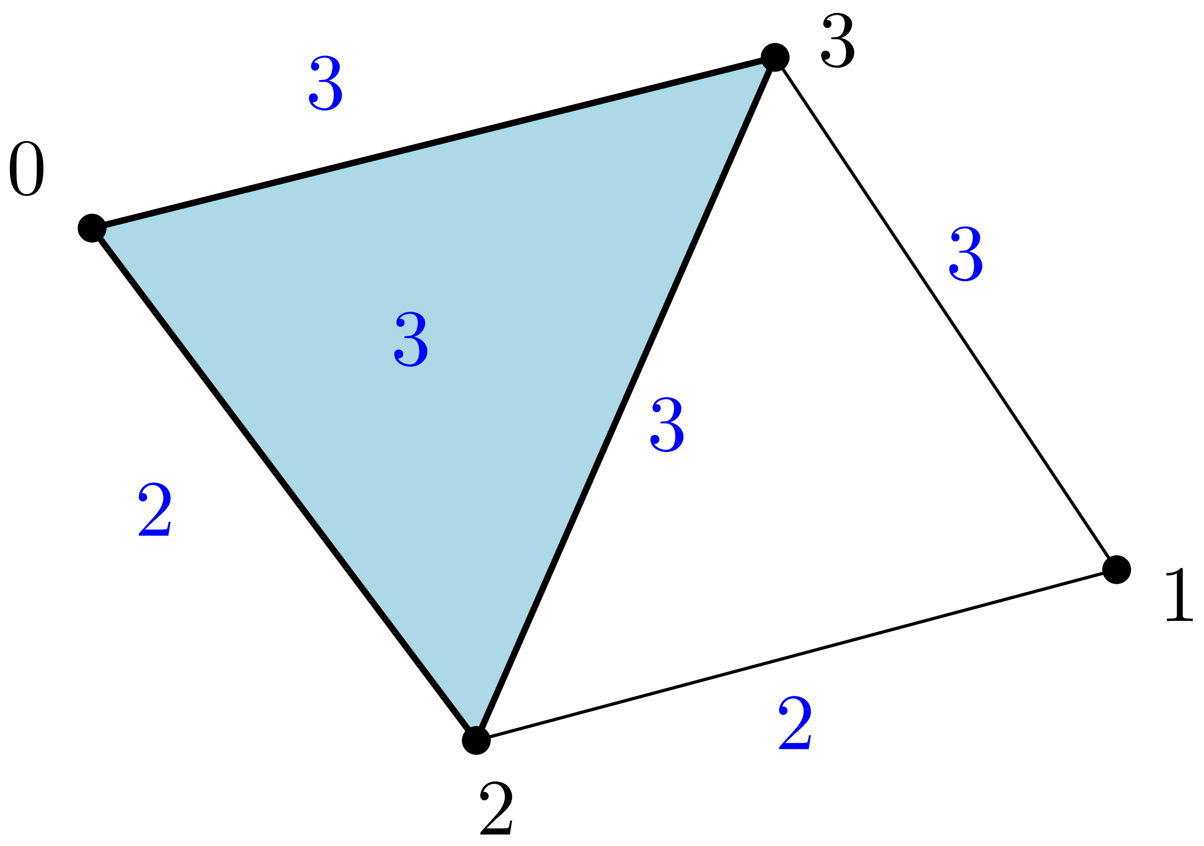
Many examples and ways to design filtrations depending on the application and targeted objectives : sublevel and upperlevel sets, Čech complex,...

Sublevel set filtration associated to a function



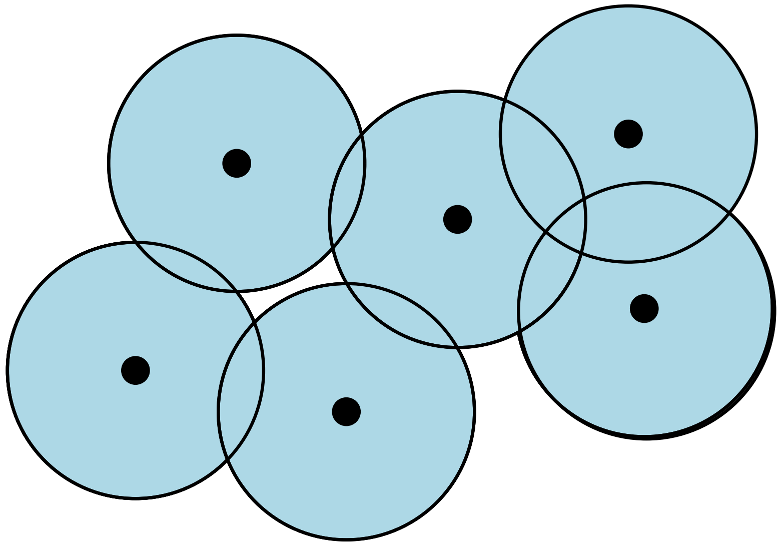
- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

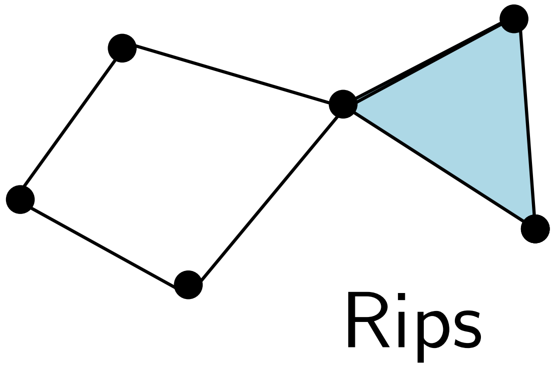
Sublevel set filtration associated to a function



- 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).

Example : 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

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

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

Bottleneck distance

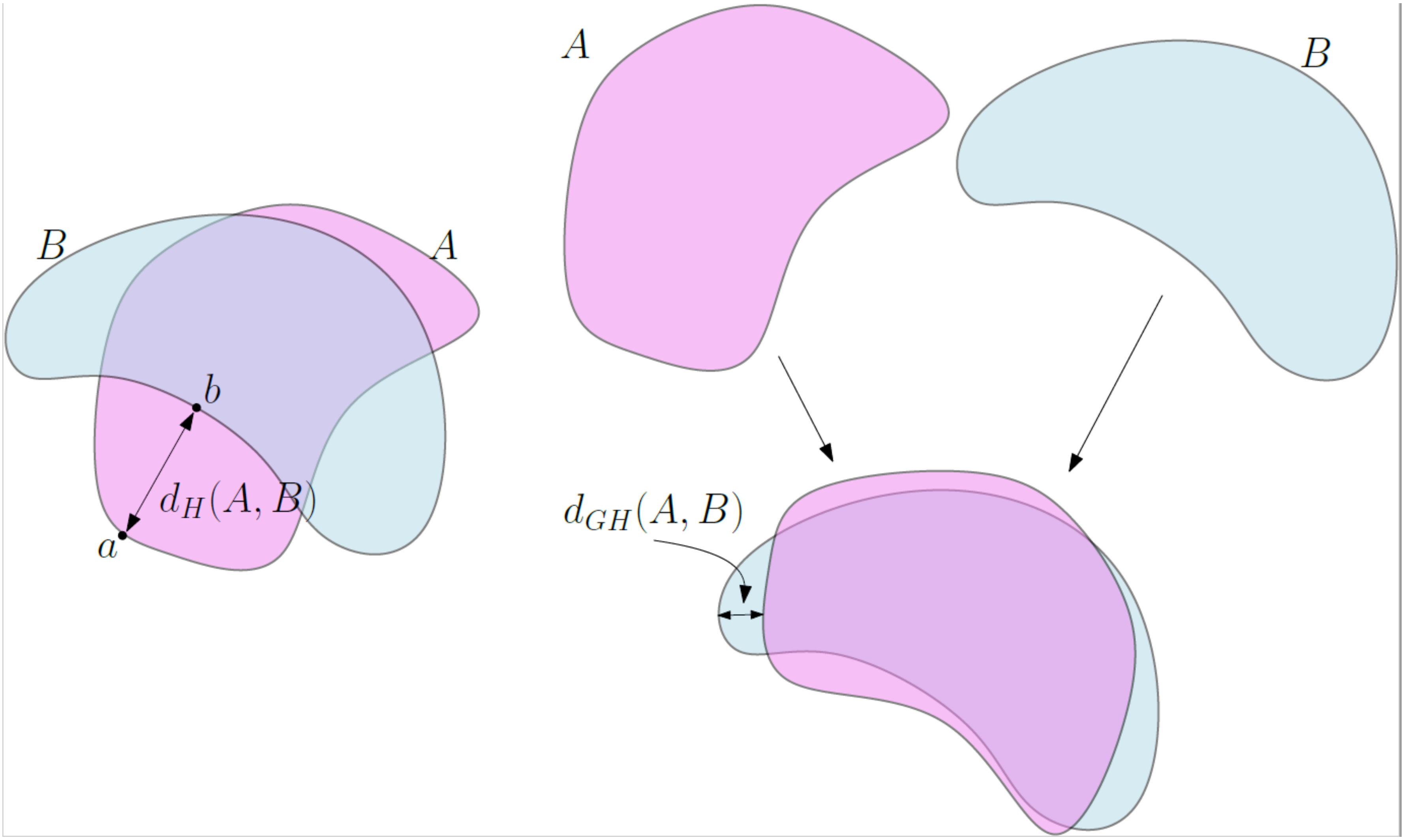
Gromov-Hausdorff distance

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

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

An algorithm to compute Betti numbers

Input : A filtration of a simplicial complex $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$, s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

Output : The Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K.

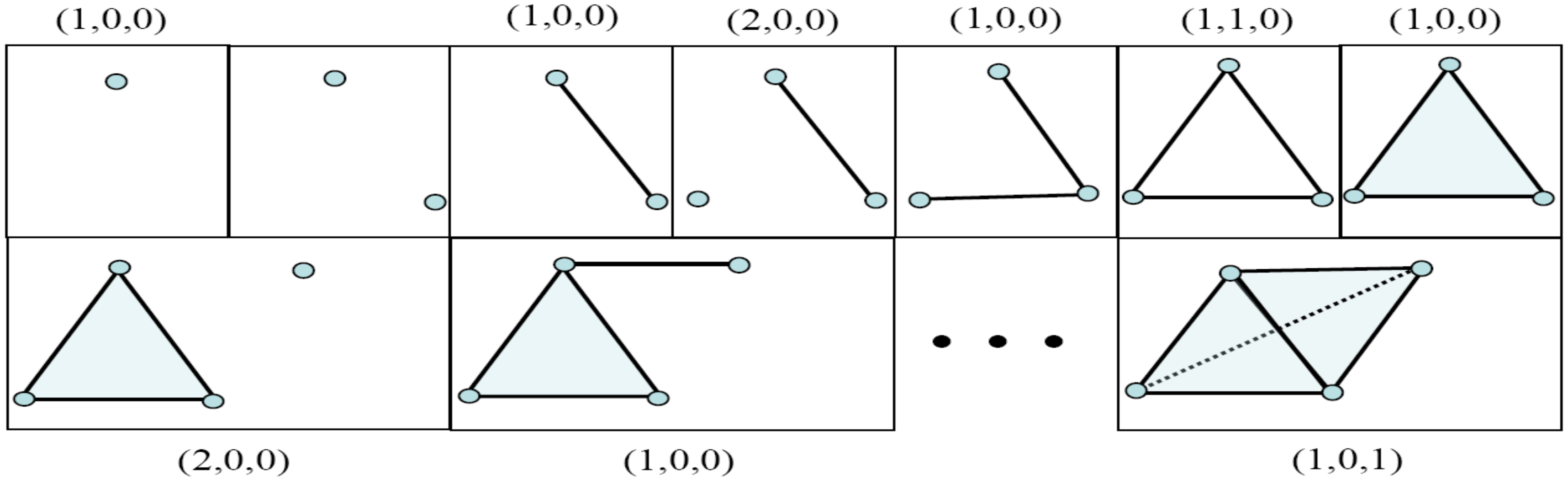
```
$\beta_0 = \beta_1 = \cdots = \beta_d = 0$ ;
for $i = 1$ to $m$
  $k = \dim \sigma^i - 1$ ;
  if $\sigma^i$ is contained in a $(k+1)$-cycle in $K^i$
    then $\beta_{k+1} = \beta_{k+1} + 1$ ;
  else $\beta_k = \beta_k - 1$ ;
  end if ;
end for ;
output $(\beta_0, \beta_1, \cdots, \beta_d)$ ;
```

An algorithm to compute Betti numbers

Input : A filtration of a simplicial complex $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$, s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

Output : The Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K.

```
$\beta_0 = \beta_1 = \cdots = \beta_d = 0$;
for $i = 1$ to $m$
  $k = \dim \sigma^i - 1$;
  if $\sigma^i$ is contained in a $(k+1)$-cycle in $K^i$
    then $\beta_{k+1} = \beta_{k+1} + 1$;
  else $\beta_k = \beta_k - 1$;
  end if;
end for;
output $(\beta_0, \beta_1, \cdots, \beta_d)$;
```



An algorithm to compute Betti numbers

Input : A filtration of a simplicial complex $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$, s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

Output : The Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K.

```
$\beta_0 = \beta_1 = \cdots = \beta_d = 0$ ;
for $i = 1$ to $m$
  $k = \dim \sigma^i - 1$ ;
  if $\sigma^i$ is contained in a $(k+1)$-cycle in $K^i$
    then $\beta_{k+1} = \beta_{k+1} + 1$ ;
  else $\beta_k = \beta_k - 1$ ;
  end if ;
end for ;
output $(\beta_0, \beta_1, \cdots, \beta_d)$ ;
```

Remark : At the i^{th} step of the algorithm, the vector $(\beta_0, \cdots, \beta_d)$ stores the Betti numbers of K^i.

An algorithm to compute Betti numbers

Input : A filtration of a simplicial complex $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$, s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

Output : The Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K.

```
$\beta_0 = \beta_1 = \cdots = \beta_d = 0$ ;
for $i = 1$ to $m$
  $k = \dim \sigma^i - 1$ ;
  if $\sigma^i$ is contained in a $(k+1)$-cycle in $K^i$
    then $\beta_{k+1} = \beta_{k+1} + 1$ ;
  else $\beta_k = \beta_k - 1$ ;
  end if ;
end for ;
output $(\beta_0, \beta_1, \cdots, \beta_d)$ ;
```

Definition : A $(k+1)$-simplex σ^i is positive if it is contained in a $(k+1)$-cycle in K^i. It is negative otherwise.

Create a new $(k+1)$-cycle in K^i

Destroy a k-cycle in K^i

$$\beta_k(K) = \sharp(\text{positive simplices}) \; - \; \sharp(\text{negative simplices})$$

From homology to persistent homology

Input : A filtration of a simplicial complex $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$, s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

Output : The Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K.

```
$\beta_0 = \beta_1 = \cdots = \beta_d = 0$ ;
for $i = 1$ to $m$
  $k = \dim \sigma^i - 1$ ;
  if $\sigma^i$ is contained in a $(k+1)$-cycle in $K^i$
    then $\beta_{k+1} = \beta_{k+1} + 1$ ;
  else $\beta_k = \beta_k - 1$ ;
  end if ;
end for ;
output $(\beta_0, \beta_1, \cdots, \beta_d)$ ;
```

The algorithm can be easily adapted to keep track of an homology basis and pairs positive simplices (birth of a new homological class) to negative simplices (death of an existing homology class).

Persistent homology of filtered simplicial complexes

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

Persistent homology of filtered simplicial complexes

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

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

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

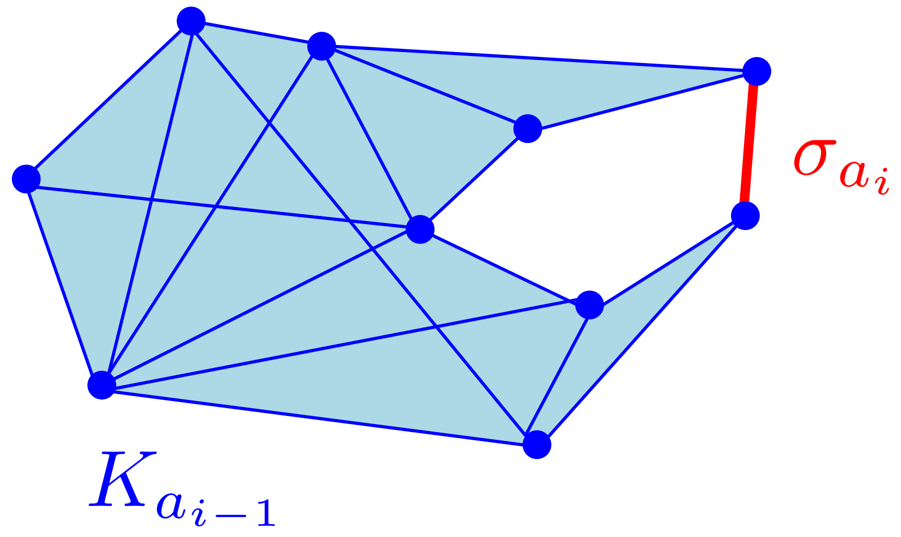
Persistent homology of filtered simplicial complexes

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

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 $K_{a_{i-1}}$ creates a new k-dimensional topological feature in K_{a_i} (new homology class in H_k).



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

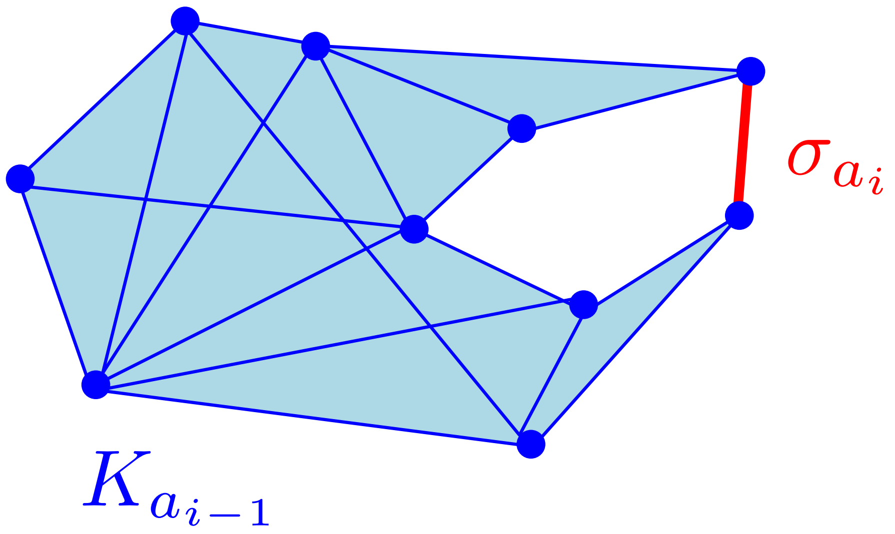
Persistent homology of filtered simplicial complexes

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

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

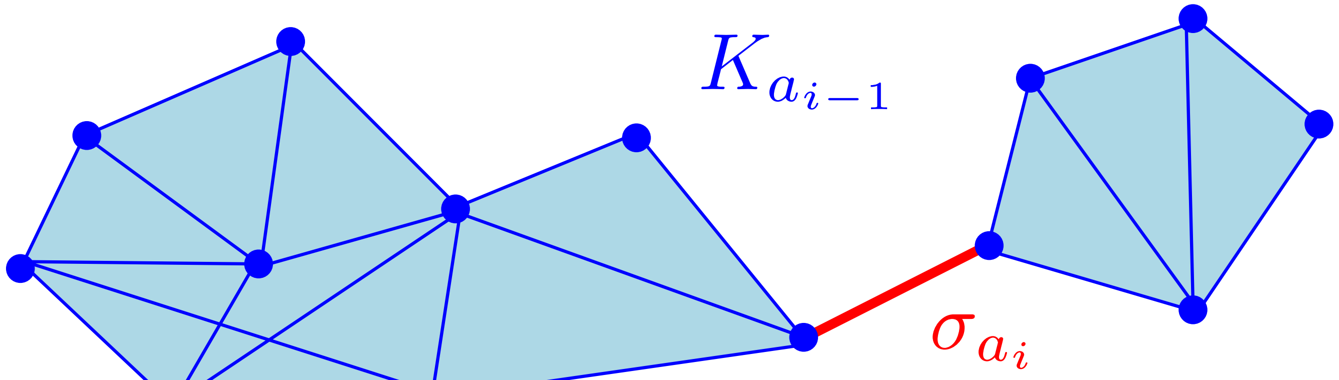
Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

Case 1 : adding σ_{a_i} to $K_{a_{i-1}}$ creates a new k-dimensional topological feature in K_{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 $K_{a_{i-1}}$ kills a $(k-1)$-dimensional topological feature in K_{a_i} (homology class in H_{k-1}).



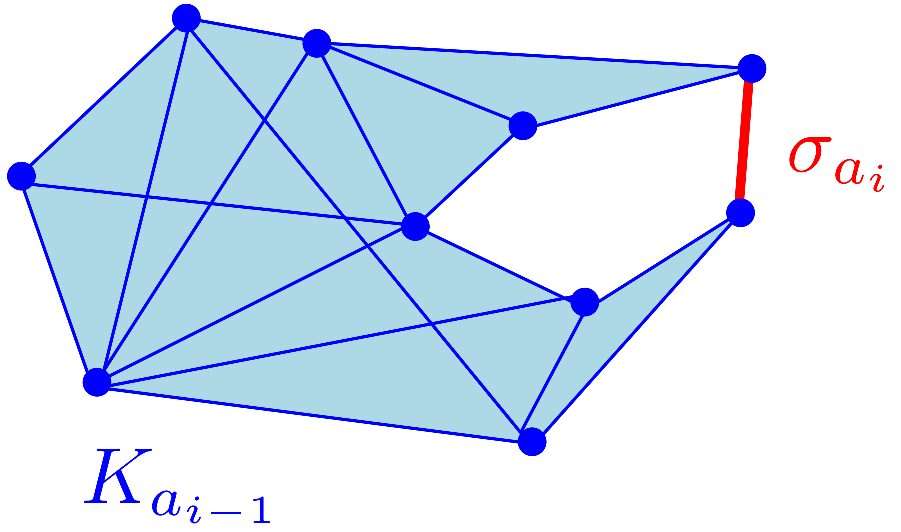
$\Rightarrow$ persistence algo. pairs the simplex σ_{a_i} to the simplex σ_{a_j} that gave birth to the killed feature.

Persistent homology of filtered simplicial complexes

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

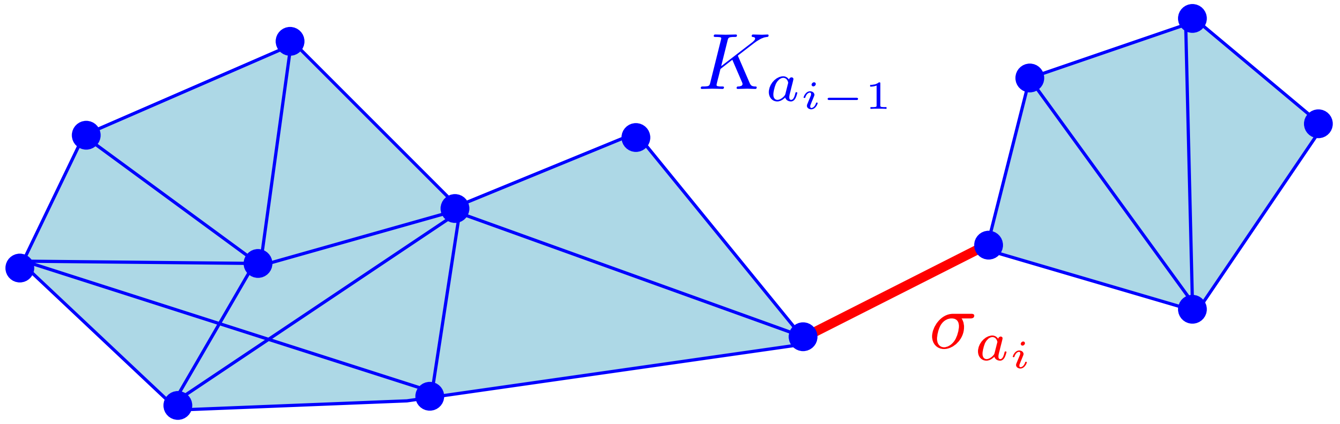
Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

Case 1 : adding σ_{a_i} to $K_{a_{i-1}}$ creates a new k-dimensional topological feature in K_{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 $K_{a_{i-1}}$ kills a $(k-1)$-dimensional topological feature in K_{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

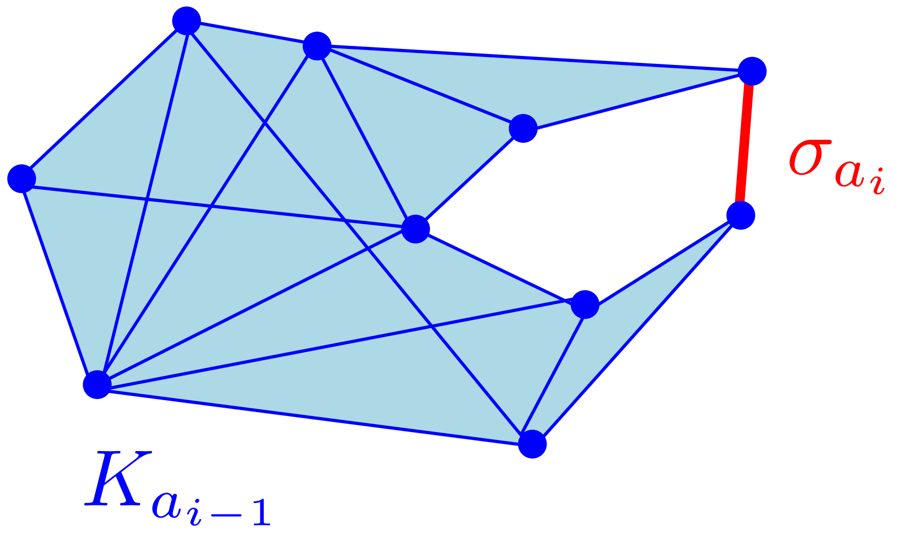
$\rightarrow$ $(a_j, a_i) \in \mathbb{R}^2$: point in the persistence diagram

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. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

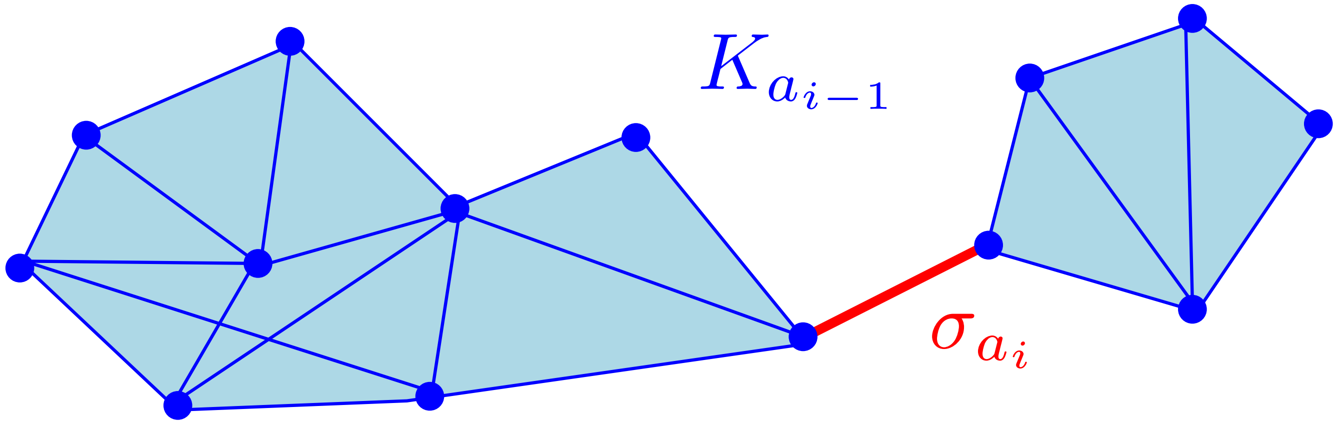
Case 1 : adding σ_{a_i} to $K_{a_{i-1}}$ creates a new k-dimensional topological feature in K_{a_i} (new homology class in H_k).



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

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.

Case 2 : adding σ_{a_i} to $K_{a_{i-1}}$ kills a $(k-1)$-dimensional topological feature in K_{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.

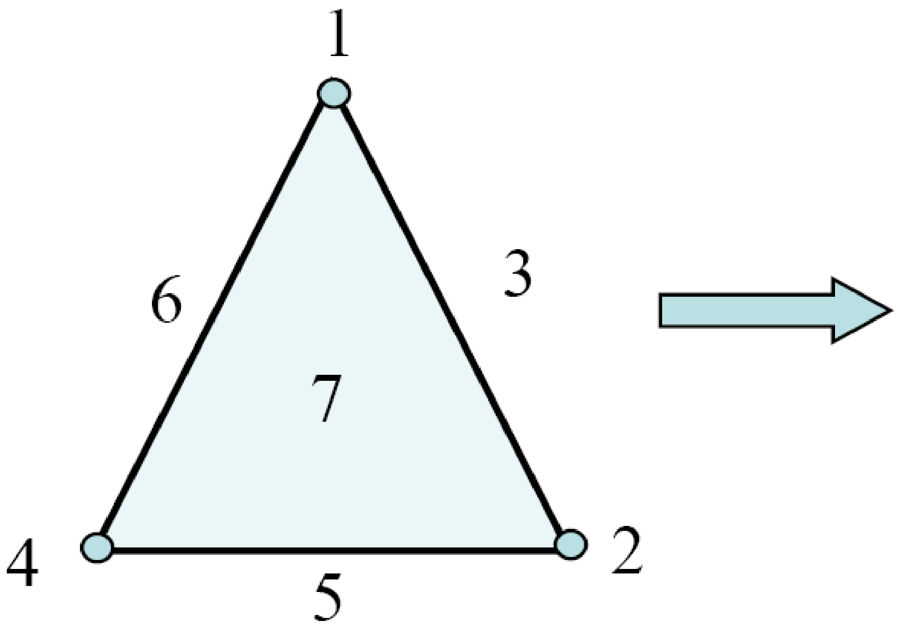
$\rightarrow$ $(\sigma_{a_j}, \sigma_{a_i})$: persistence pair

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

The persistence algorithm : matrix version

Input : $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$ a d-dimensional filtration of a simplicial complex K s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

The matrix of the boundary operator :



$$\begin{array}{cccccccc} & & & & j & & & \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & \boxed{1} & 1 & 0 & l(j) \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \end{array}$$

— $M = (m_{ij})_{i,j=1,\cdots,m}$ with coefficient in $\mathbb{Z}/2$ defined by

$$m_{ij} = 1 \text{ if } \sigma^i \text{ is a face of } \sigma^j \text{ and } m_{ij} = 0 \text{ otherwise}$$

— For any column C_j, $l(j)$ is defined by

$$(i = l(j)) \Leftrightarrow (m_{ij} = 1 \text{ and } m_{i'j} = 0 \ \ \forall i' > i)$$

The persistence algorithm : matrix version

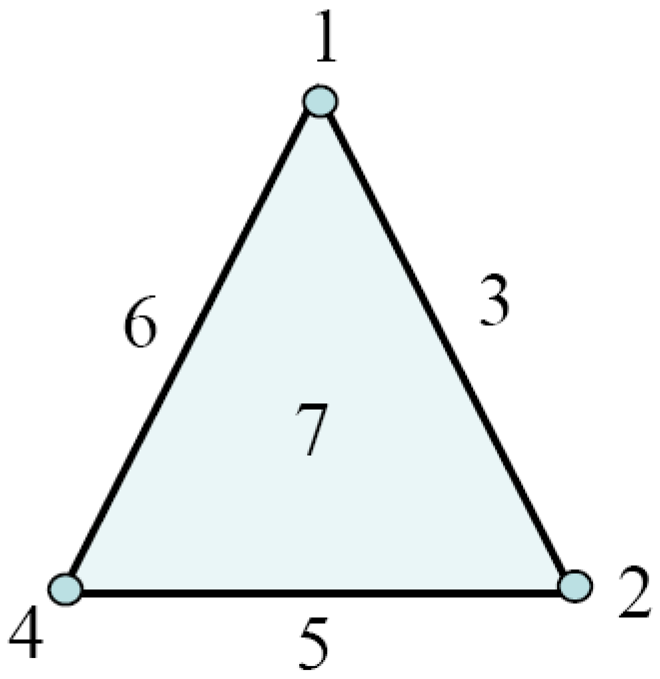
Input : $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$ a d-dimensional filtration of a simplicial complex K s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

```
Compute the matrix of the boundary operator $M$
For $j = 0$ to $m$
  While (there exists $j' < j$ such that $l(j') == l(j)$)
    $C_j = C_j + C_{j'}$ mod$(2)$ ;
  End while
End for
Output the pairs $(l(j), j)$ ;
```

Remark : The worst case complexity of the algorithm is $O(m^3)$ but much lower in most practical cases.

The persistence algorithm : matrix version

A simple example :

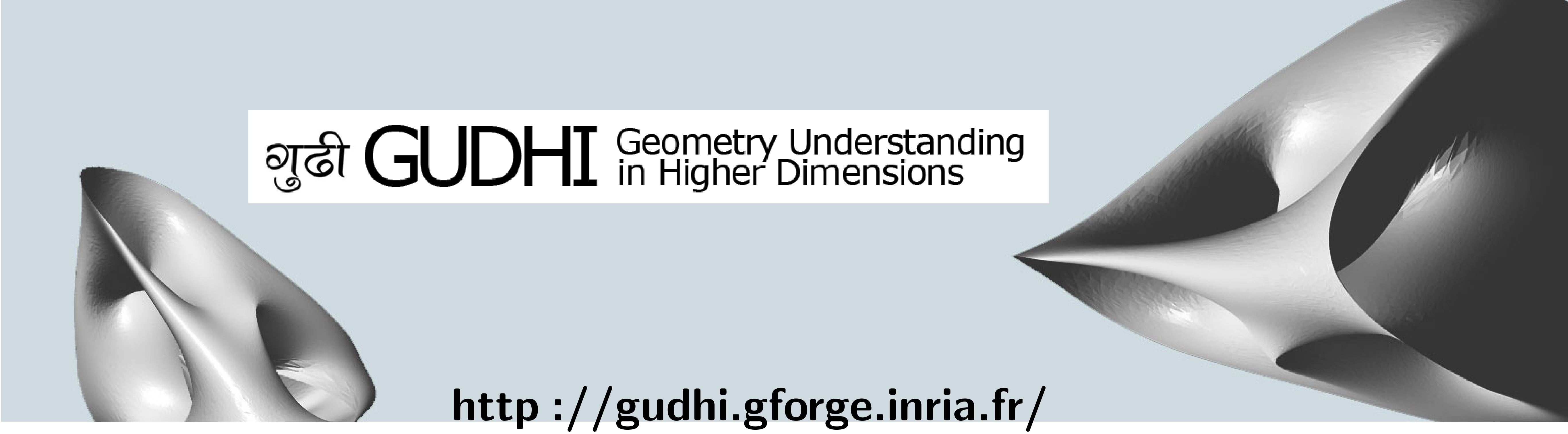


$$\Rightarrow \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & \boxed{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \begin{array}{c} \begin{array}{ccccccc} & & & & & C_5 + C_6 & \\ & & & & & \downarrow & \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & \boxed{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\Rightarrow \begin{array}{c} \begin{array}{ccccccc} & & & & & C_3 + C_6 & \\ & & & & & \downarrow & \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Paires : (2,3) (4,5) (6,7)

Persistent homology with the GUDHI library



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,...
- algorithms and tools for TDA and ML.

If there is some time left...

Persistence from an 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 :

- 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.

Persistence from an 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.

Persistence from an 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 precompact metric space. Then $\mathrm{H}(\mathrm{Rips}(\mathbb{X}))$ is q-tame.

Recall that a metric space $(\mathbb{X}, \rho)$ is precompact if for any $\epsilon > 0$ there exists a finite subset $F_\epsilon \subset \mathbb{X}$ such that $d_H(\mathbb{X}, F_\epsilon) < \epsilon$ (i.e. $\forall x \in X, \exists p \in F_\epsilon$ s.t. $\rho(x,p) < \epsilon$).

Persistence from an 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.

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 $a \leq b$.

$$\begin{tikzcd}
U^a \arrow[rr] \arrow[dr] & & U^b \arrow[dr] & \\
& V^{a+} \arrow[rr] \arrow[ur, phantom, "\circlearrowright"] & & V^{b+\epsilon}
\end{tikzcd}$$

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 $\mathbb{V}$.

$$\begin{tikzcd}
\cdots \arrow[rr] & & U^a \arrow[rr, "u_a^{a+}"] \arrow[dr, "\phi_a"] & & U^{a+} \arrow[rr] \arrow[dr] & & \cdots & \\
& \cdots \arrow[ur] \arrow[rr] & & V^{a+} \arrow[rr, "v_{a+\epsilon}^{a+3}"'] \arrow[ur, "\psi_{a+}"] & & V^{a+} \arrow[rr] & & \cdots
\end{tikzcd}$$

Persistence from an 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

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

Persistence from an 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

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

Exercise : Show the stability theorem for (tame) functions :
let $\mathbb{X}$ be a topological space and let $f, g : \mathbb{X} \to \mathbb{R}$ be two *tame* functions. Then

$$\mathrm{d_B}(\mathrm{D}_f, \mathrm{D}_g) \leq \|f - g\|_\infty.$$

Persistence from an 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

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

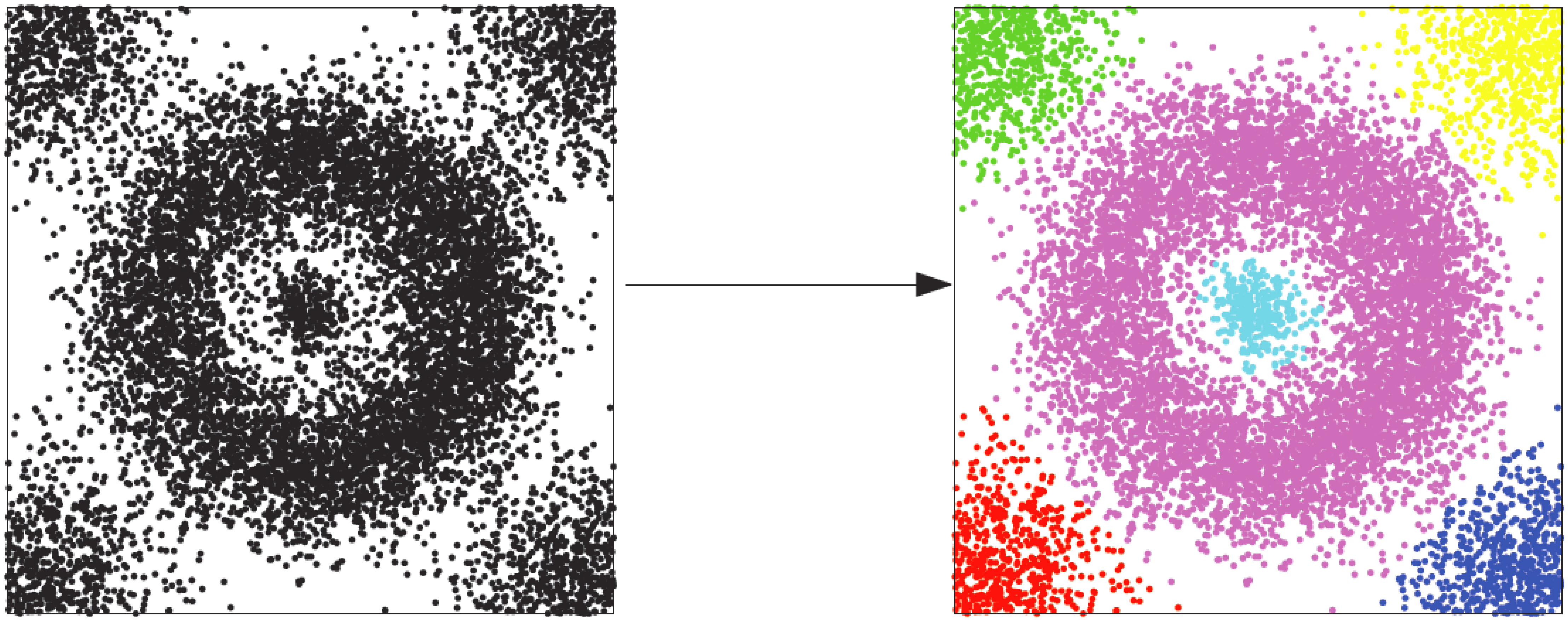
Strategy : build filtrations that induce **q-tame** homology persistence modules and that turn out to be **ϵ-interleaved** when the considered spaces/functions are *$O(\epsilon)$-close*.

A few applications of persistence

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

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



Input :

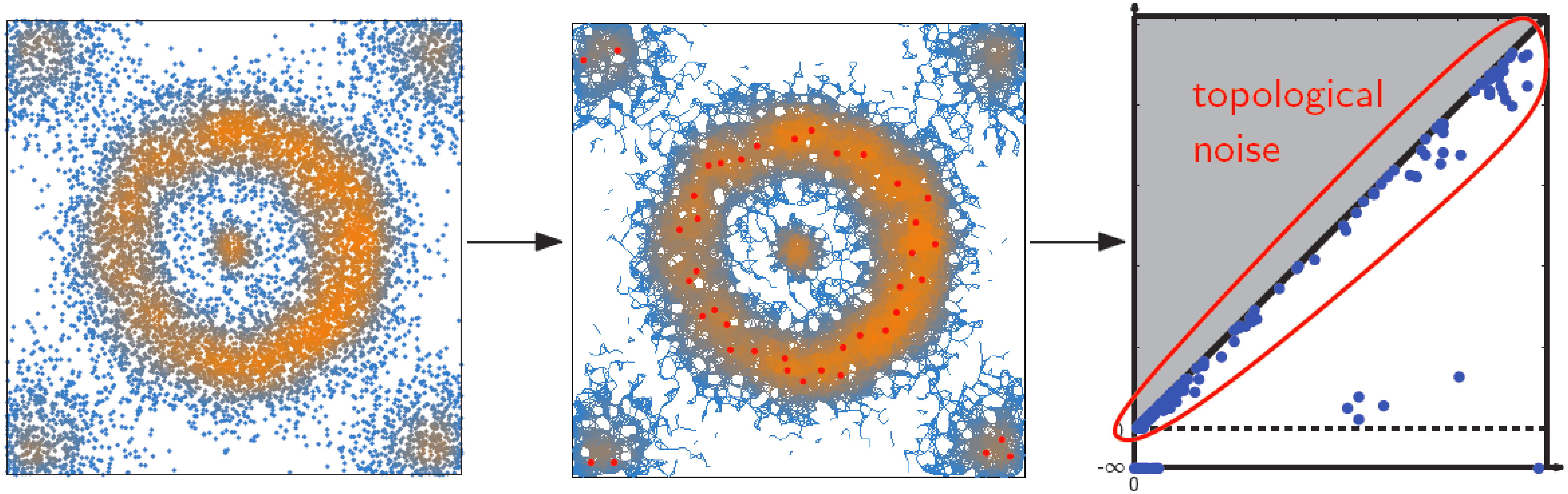
1. A finite set X of observations (point cloud with coordinates or pairwise distance matrix),
2. A real valued function f defined on the observations (e.g. density estimate).

Goal : Partition the data according to the basins of attraction of the peaks of f

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

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



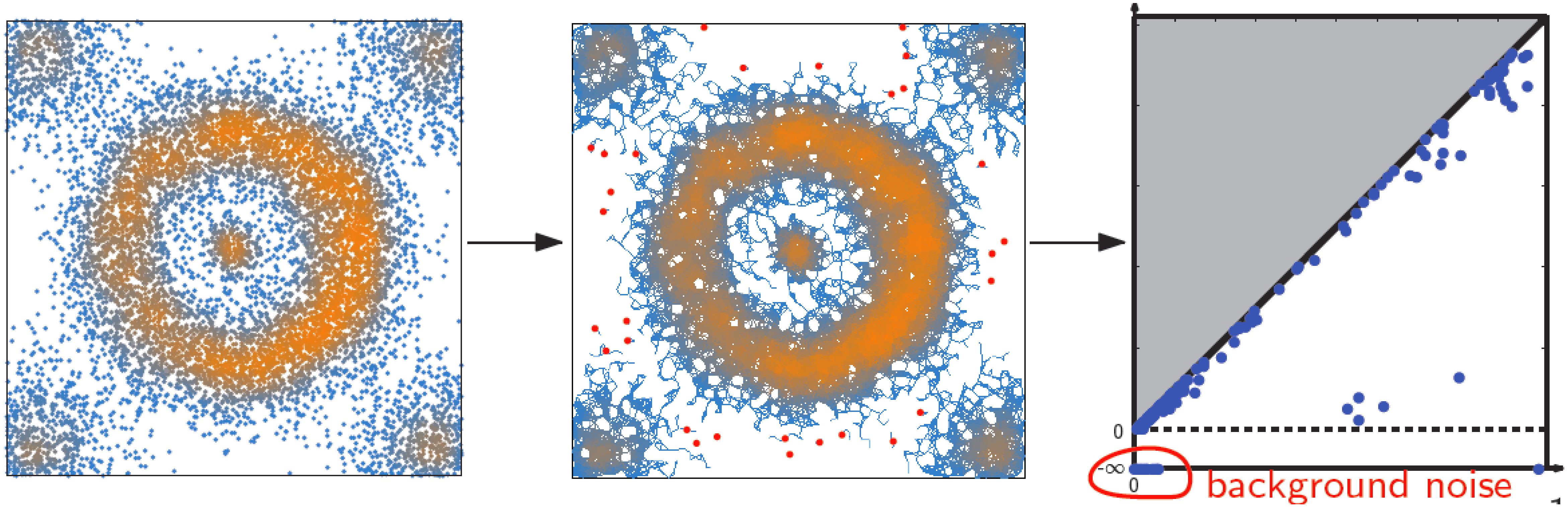
1. Build a neighborhing graph G on top of X.

2. Compute the (0-dim) persistence of f to identify prominent peaks → number of clusters (union-find algorithm).

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

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



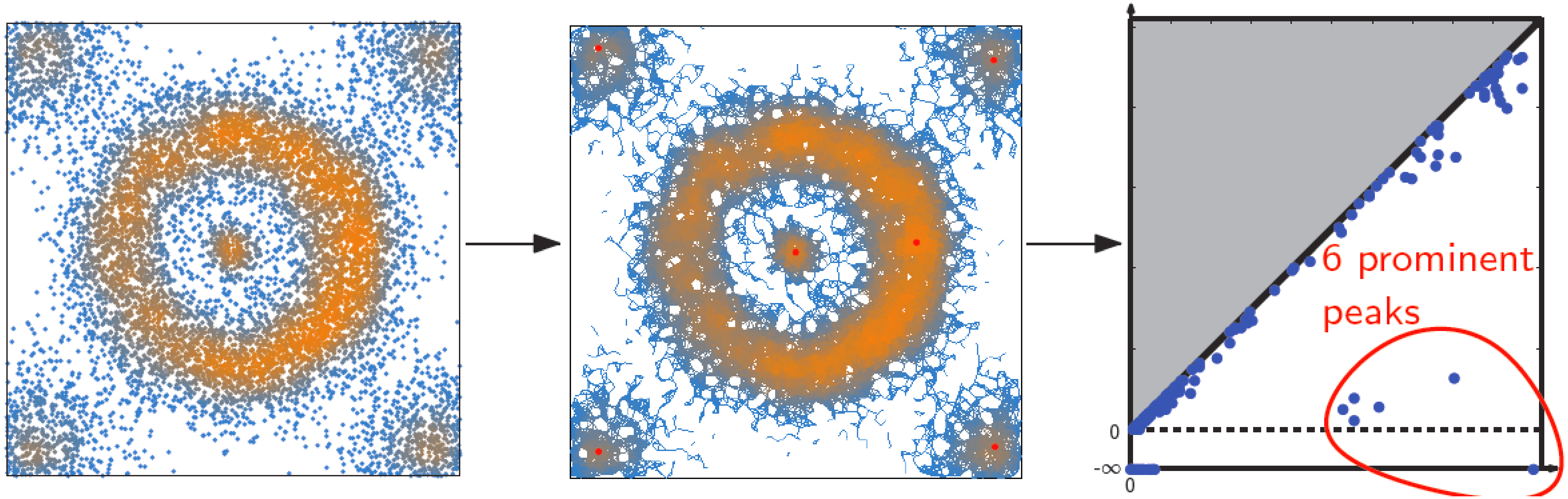
1. Build a neighborhing graph G on top of X.

2. Compute the (0-dim) persistence of f to identify prominent peaks → number of clusters (union-find algorithm).

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

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



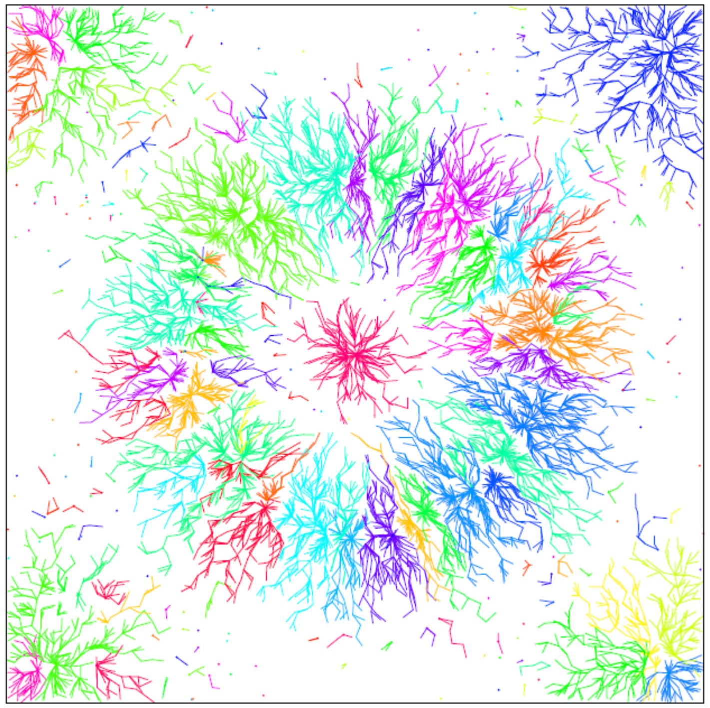
1. Build a neighborhing graph G on top of X.

2. Compute the (0-dim) persistence of f to identify prominent peaks → number of clusters (union-find algorithm).

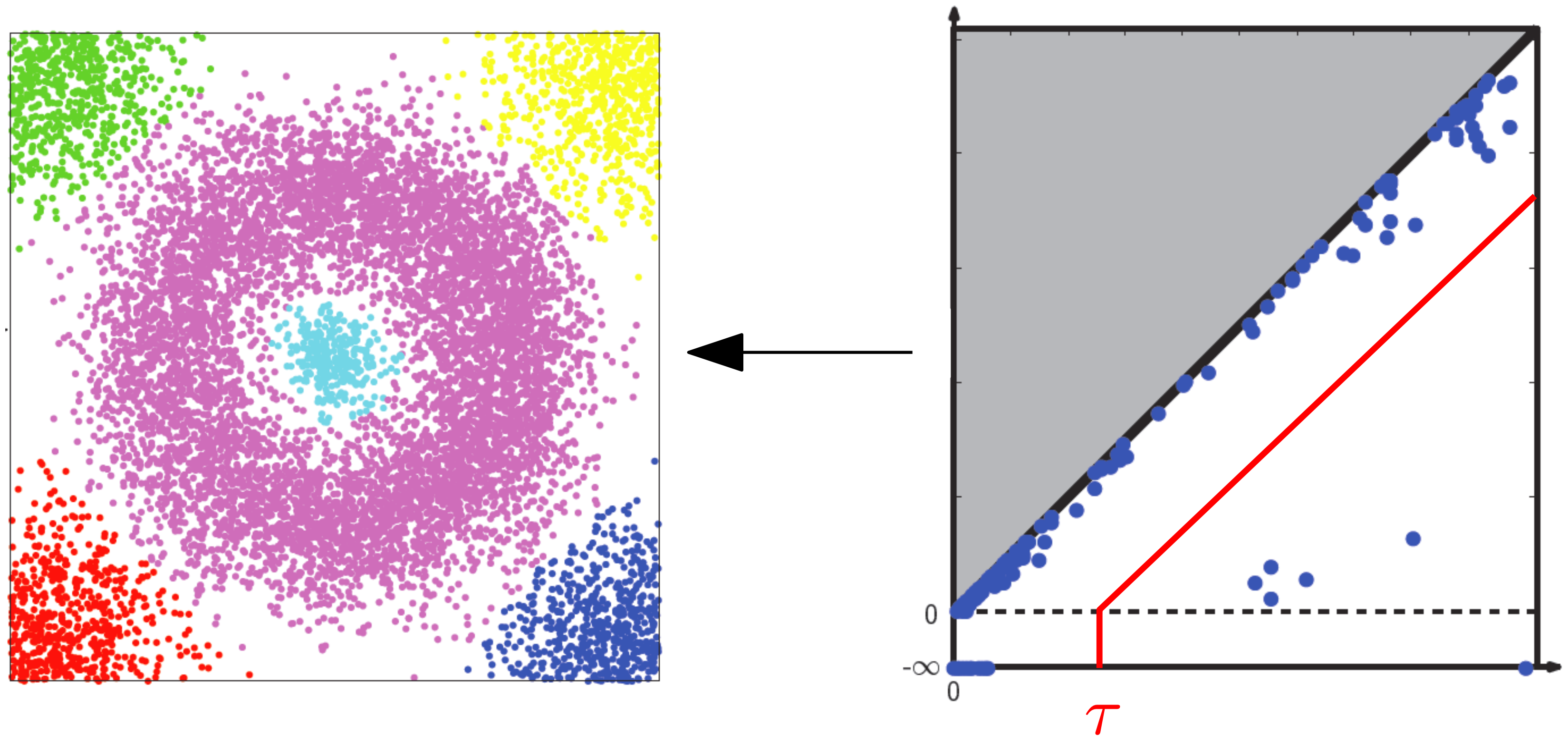
Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

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



$\tau = 0$



1. Build a neighborhing graph G on top of X.

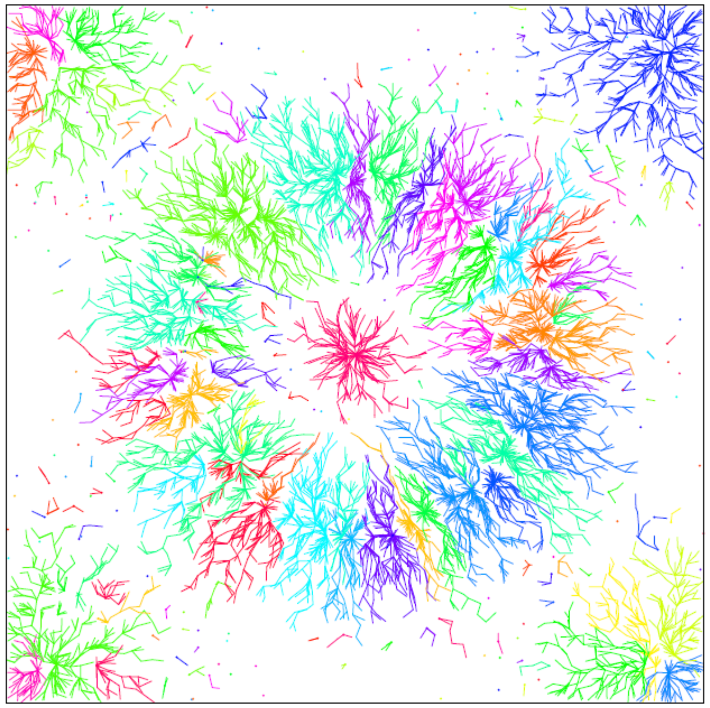
2. Compute the (0-dim) persistence of f to identify prominent peaks → number of clusters (union-find algorithm).

3. Chose a threshold $\tau > 0$ and use the persistence algorithm to merge components with prominence less than τ.

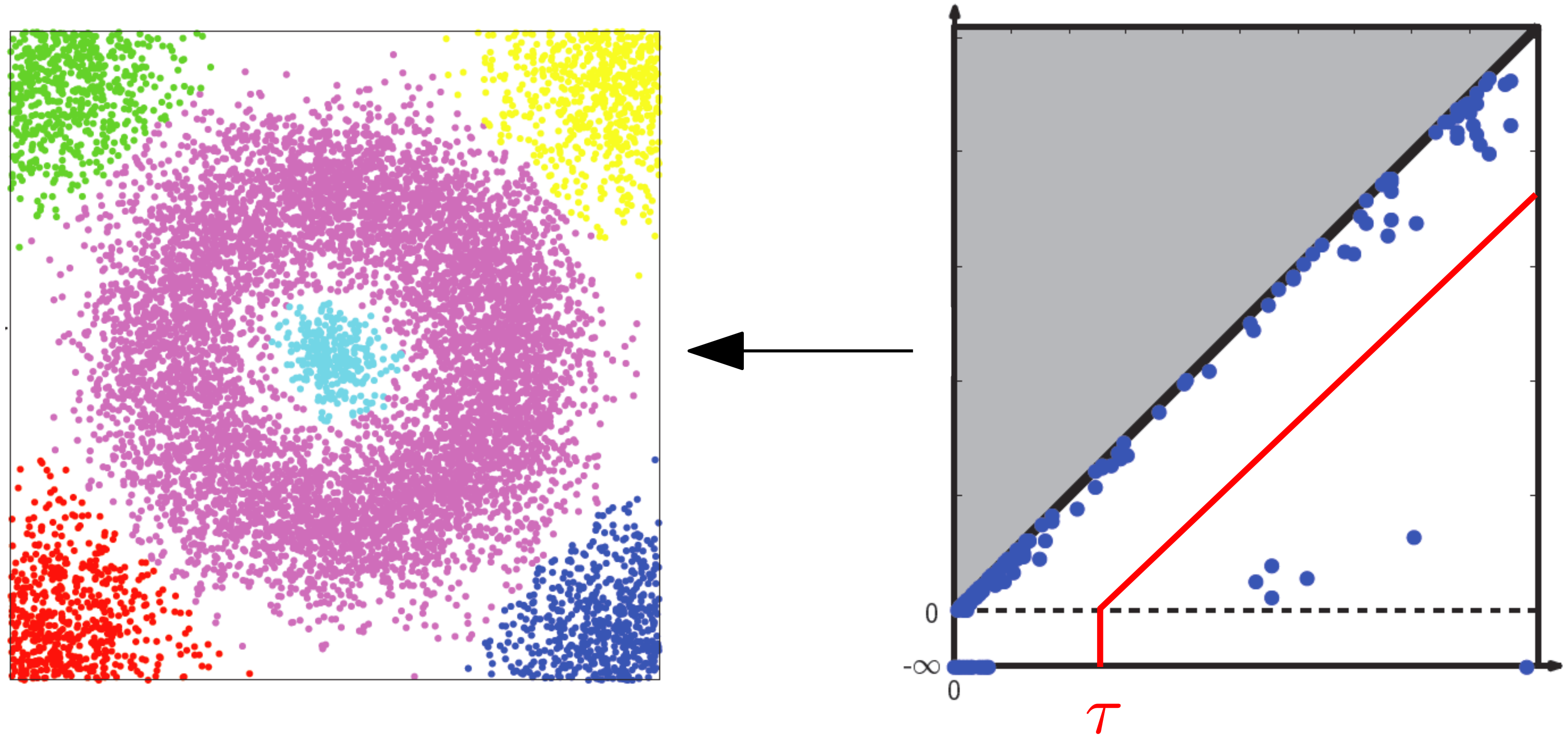
Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

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



$\tau = 0$



Complexity of the algorithm : $O(n \log n)$

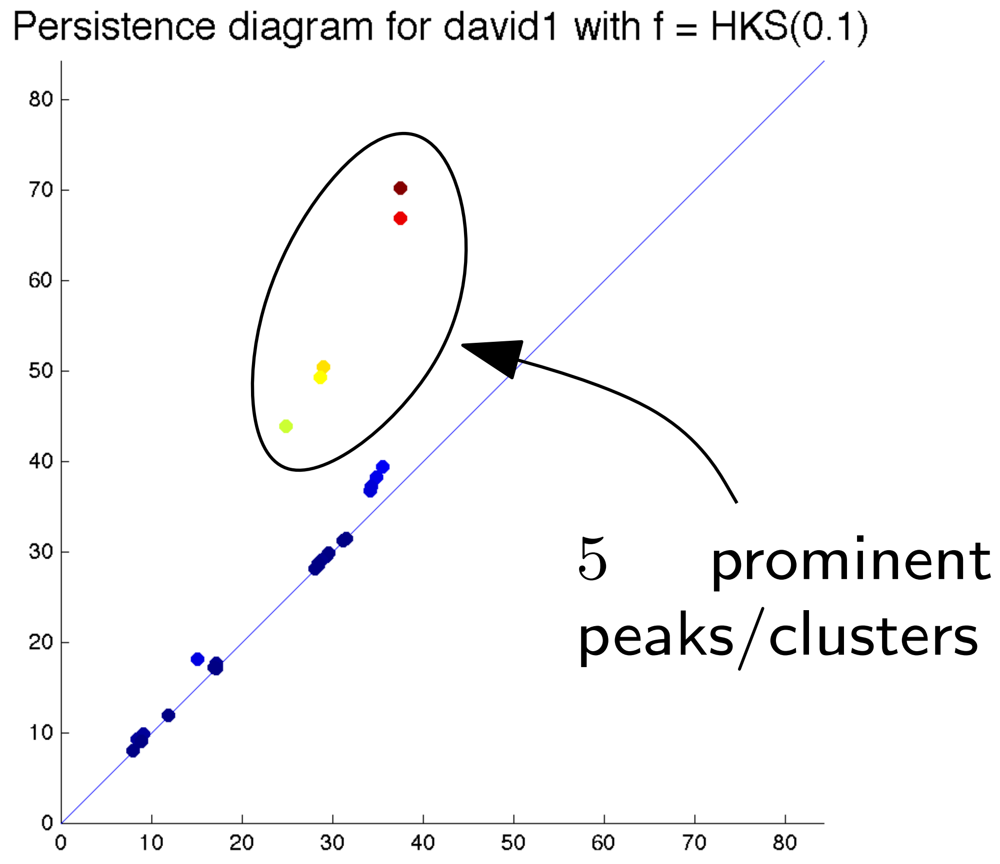
Theoretical guarantees :

- Stability of the number of clusters (w.r.t. perturbations of X and f).
- Partial stability of clusters : well identified stable parts in each cluster.

“soft ” clustering

Application to non-rigid shape segmentation

[Skraba, Ovsjanikov, C.,Guibas, NORDIA 10]



Persistence diagram for david1 with f = HKS(0.1)

5 prominent peaks/clusters

X : a 3D shape
f = HKS function on X

Problem : some part of clusters are unstable → dirty segments

Application to non-rigid shape segmentation

[Skraba, Ovsjanikov, C.,Guibas, NORDIA 10]

Problem : some part of clusters are unstable → dirty segments

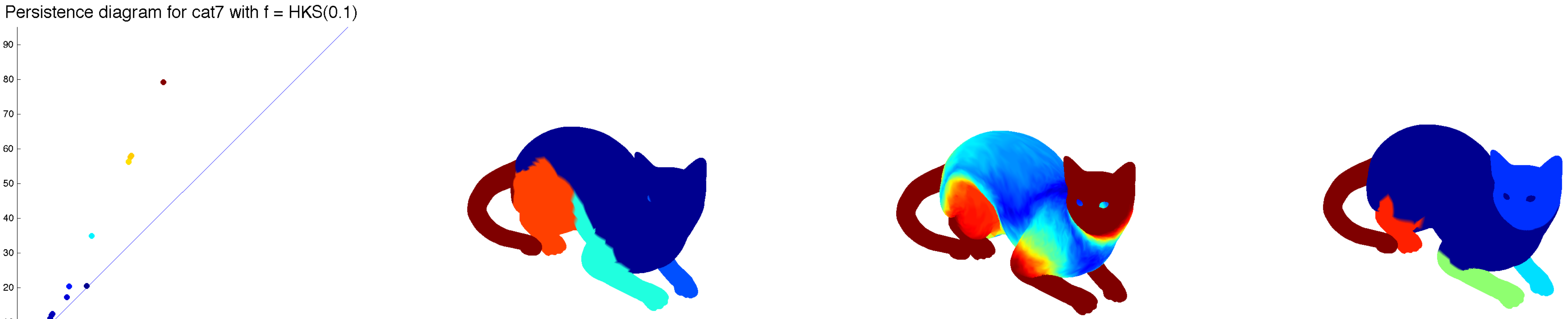
Idea :

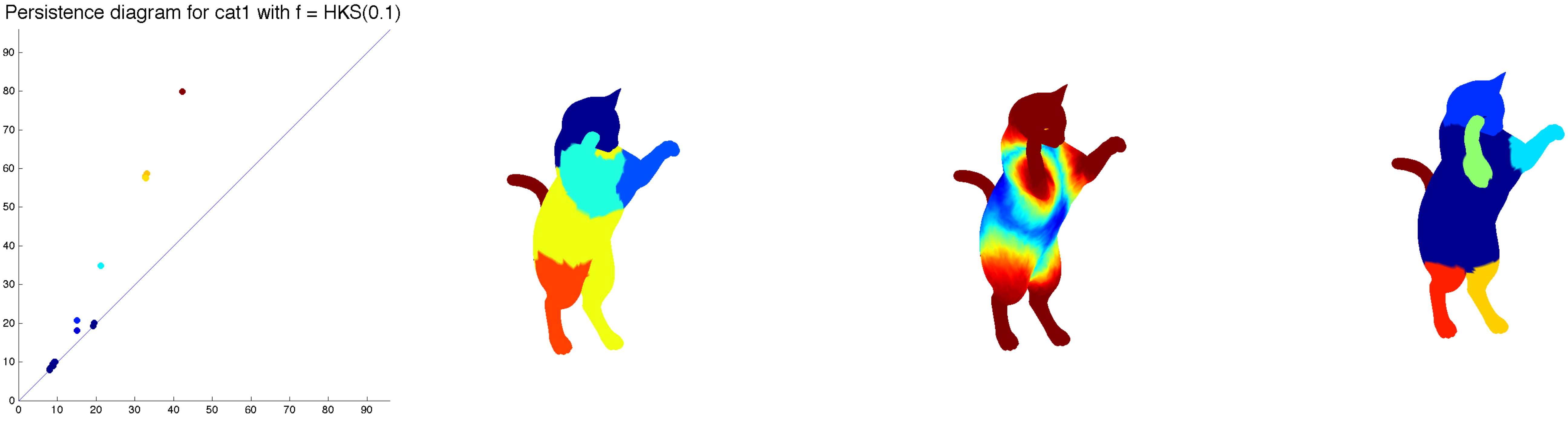
- Run the persistence based algorithm several times on random perturbations of f (size bounded by the “persistence” gap).

- Partial stability of clusters allows to establish correspondences between clusters across the different runs → for any $x \in X$, a vector giving the probability for x to belong to each cluster.

Application to non-rigid shape segmentation

[Skraba, Ovsjanikov, C.,Guibas, NORDIA 10]



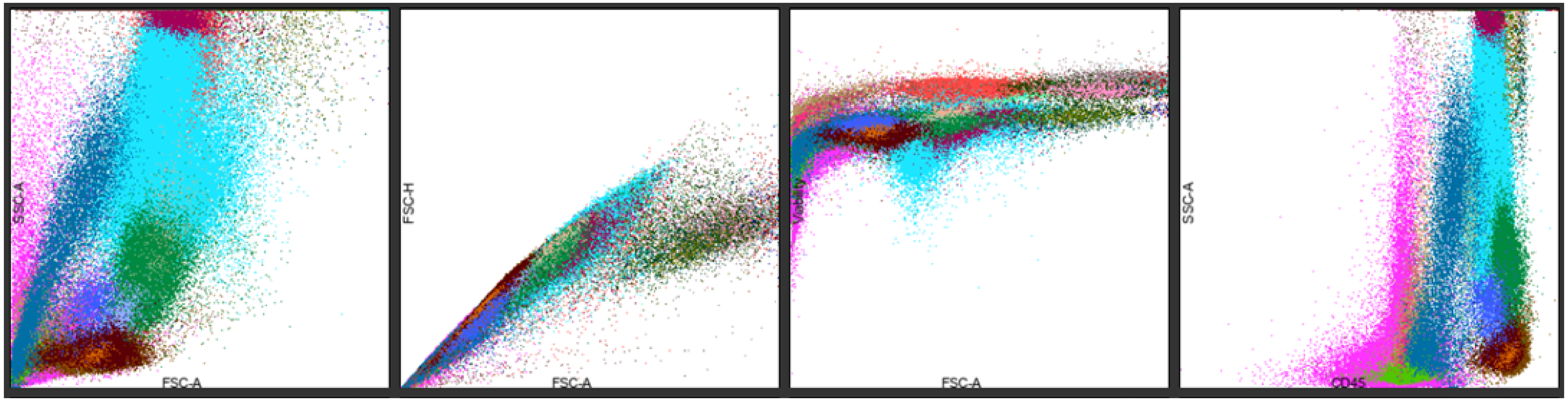


Topology-based unsupervised classification and anomaly detection on cytometry data for medical diagnosis

[M. Glisse, L. Pujol et al 2022]

METAFORA biosystems

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

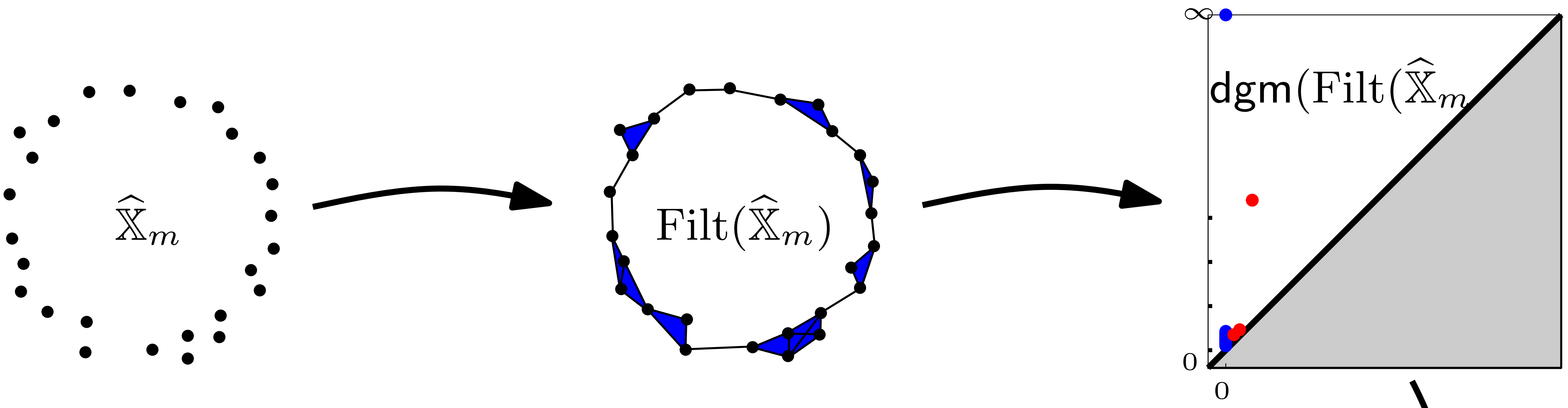


Objective : unsupervised learning in large point clouds (several millions) in medium/high dimensions ($\approx 4 \rightarrow 80$)

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

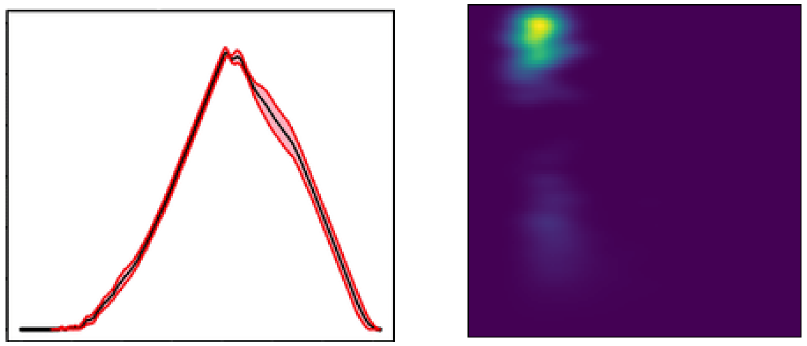
Methodology : TDA based approaches, combined with dim. reduction methods to identify relevant patterns and subsamples.

The problem of representation of persistence



Persistence diagrams are not well-suited for classical ML algorithms (the space of PD is highly non linear)
Not always clear which part of the diagrams carries the relevant information.

Machine Learning / AI



Representations of persistence

A zoo of representations of persistence

(non exhaustive list - see also Gudhi representations)

- Collections of 1D functions

  → landscapes [Bubenik 2012]

  → Betti curves [Umeda 2017]

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

  → persistence images [Adams et al 2017]

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

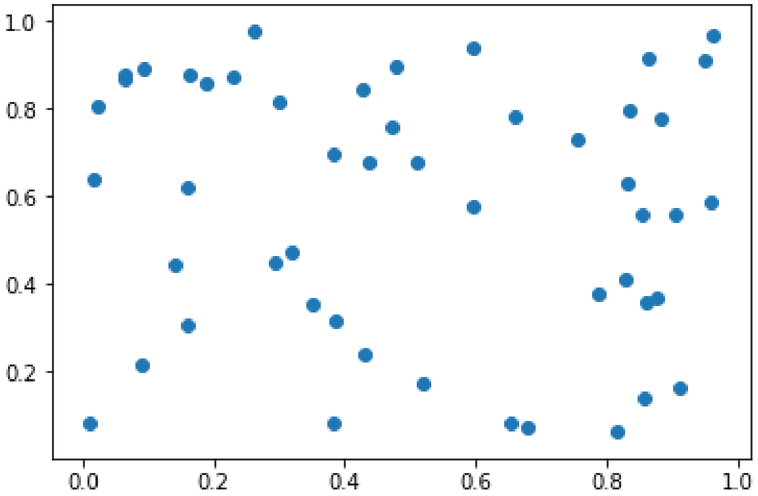
  → sliced on lines [Carrière Oudot Cuturi 2017]

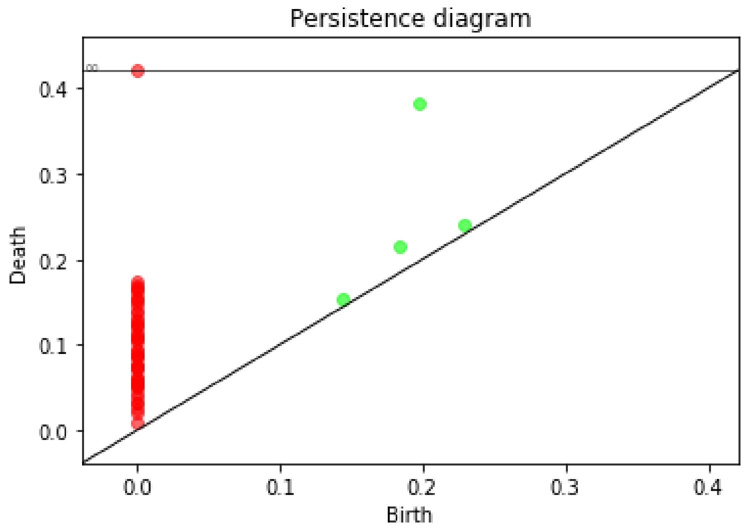
- finite metric spaces [Carrière Oudot Ovsjanikov 2015]

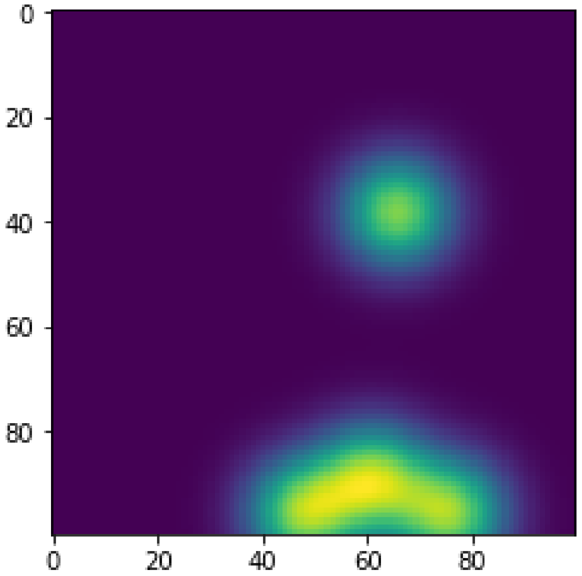
- polynomial roots or evaluations [Di Fabio Ferri 2015] [Kališnik 2016]

Persistence images

[Adams et al, JMLR 2017]





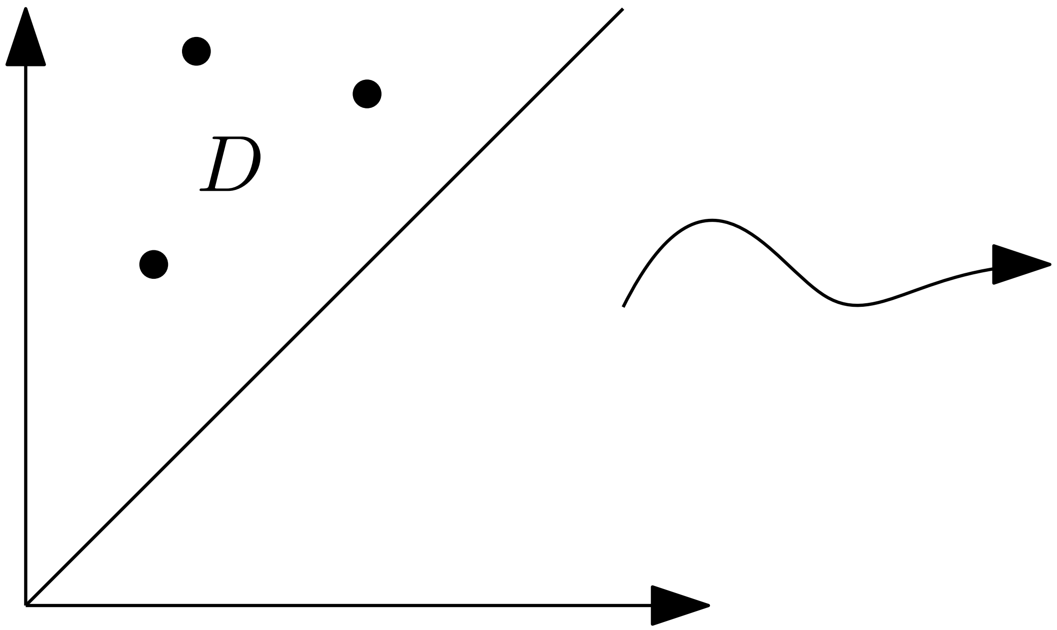


For $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel and H a bandwidth matrix (e.g. a symmetric positive definite matrix), pose for $u \in \mathbb{R}^2$, $K_H(u) = |H|^{-1/2} K(H^{-1/2} \cdot u)$

For $D = \sum_i \delta_{p_i}$ 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 :

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

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.

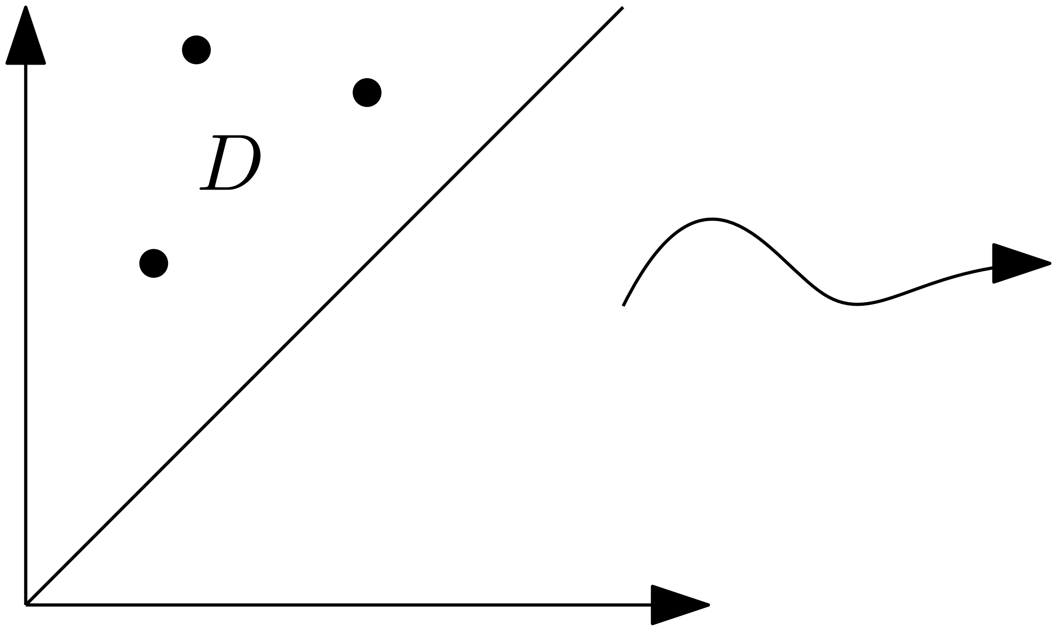
[C., de Silva, Glisse, Oudot 16]

- Many persistence representations can be expressed as

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

for well-chosen functions $f : \mathbb{R}^2 \to \mathcal{H}$.

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

Filtrations revisited

Let $n > 0$ be an integer,
$\mathcal{F}_n$: the collection of non-empty subsets of $\{1, \dots, n\}$,
M : a real analytic compact d-dim. connected manifold (poss. with boundary).

Filtering function :

$$\varphi = (\varphi[J])_{J \in \mathcal{F}_n} : M^n \to \mathbb{R}^{|\mathcal{F}_n|}$$

satisfiying the following conditions :

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$.

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.

Given $x = (x_1, \cdots, x_n)$, $\varphi(x)$ induces an order on the faces of the simplex with n vertices that is a filtration $\mathcal{K}(x)$:

$$\forall J \in \mathcal{F}_n,\ J \in \mathcal{K}(x, r) \Longleftrightarrow \varphi[J](x) \leq r.$$

Filtrations revisited

Not : for $x = (x_1, \dots, x_n) \in M^n$ and for J a simplex, $x(J) := (x_j)_{j \in J}$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$.

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$.

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \le \varphi[J']$.

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \dots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.

(K5) *Smoothness :* The function φ is subanalytic and the gradient of each of its entries (which is defined a.s.e.) is non vanishing a.s.e..

Filtrations revisited

Let $n > 0$ be an integer,
$\mathcal{F}_n$: the collection of non-empty subsets of $\{1, \dots, n\}$,
M : a real analytic compact d-dim. connected manifold (poss. with boundary).

Filtering function :

$$\varphi = (\varphi[J])_{J \in \mathcal{F}_n} : M^n \to \mathbb{R}^{|\mathcal{F}_n|}$$

satisfiying the following conditions :

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$.

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.

Given $x = (x_1, \cdots, x_n)$, $\varphi(x)$ induces an order on the faces of the simplex with n vertices that is a filtration $\mathcal{K}(x)$:

$$\forall J \in \mathcal{F}_n,\ J \in \mathcal{K}(x, r) \Longleftrightarrow \varphi[J](x) \leq r.$$

Filtrations revisited

Not : for $x = (x_1, \dots, x_n) \in M^n$ and for J a simplex, $x(J) := (x_j)_{j \in J}$

- (K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$.
- (K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$.
- (K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.
- (K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \dots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.
- (K5) *Smoothness :* The function φ is subanalytic and the gradient of each of its entries (which is defined a.s.e.) is non vanishing a.s.e..

The example of the Vietoris-Rips filtration

$$\varphi[J](x) = \max_{i,j \in J} d(x_i, x_j)$$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$.

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \ldots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \ldots, x_{\tau(n)}) = \varphi[J](x_1, \ldots, x_n)$.

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \ldots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.

(K5') *Smoothness :* The function φ is subanalytic and the gradient of each of its entries J of size larger than 1 is non vanishing a.e. and for $J = \{j\}$, $\varphi[\{j\}] \equiv 0$.

The example of the Vietoris-Rips filtration

$$\varphi[J](x) = \max_{i,j \in J} d(x_i, x_j)$$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$. ✓

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \ldots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \ldots, x_{\tau(n)}) = \varphi[J](x_1, \ldots, x_n)$.

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \ldots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.

(K5') *Smoothness :* The function φ is subanalytic and the gradient of each of its entries J of size larger than 1 is non vanishing a.e. and for $J = \{j\}$, $\varphi[\{j\}] \equiv 0$.

The example of the Vietoris-Rips filtration

$$\varphi[J](x) = \max_{i,j \in J} d(x_i, x_j)$$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$. ✓

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \ldots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \ldots, x_{\tau(n)}) = \varphi[J](x_1, \ldots, x_n)$. ✓

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \ldots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.

(K5') *Smoothness :* The function φ is subanalytic and the gradient of each of its entries J of size larger than 1 is non vanishing a.e. and for $J = \{j\}$, $\varphi[\{j\}] \equiv 0$.

The example of the Vietoris-Rips filtration

$$\varphi[J](x) = \max_{i,j \in J} d(x_i, x_j)$$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$. ✓

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$. ✓

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$. ✓

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \dots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.

(K5') *Smoothness :* The function φ is subanalytic and the gradient of each of its entries J of size larger than 1 is non vanishing a.e. and for $J = \{j\}$, $\varphi[\{j\}] \equiv 0$.

The example of the Vietoris-Rips filtration

$$\varphi[J](x) = \max_{i,j \in J} d(x_i, x_j)$$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$. ✓

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \ldots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \ldots, x_{\tau(n)}) = \varphi[J](x_1, \ldots, x_n)$. ✓

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$. ✓

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \ldots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U. ✓

(K5') *Smoothness :* The function φ is subanalytic and the gradient of each of its entries J of size larger than 1 is non vanishing a.e. and for $J = \{j\}$, $\varphi[\{j\}] \equiv 0$.

The example of the Vietoris-Rips filtration

$$\varphi[J](x) = \max_{i,j \in J} d(x_i, x_j)$$

(K1) *Absence of interaction :* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$. ✓

(K2) *Invariance by permutation :* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of the entries having support included in J, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$. ✓

(K3) *Monotony :* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$. ✓

(K4) *Compatibility :* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \dots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U. ✓

(K5') *Smoothness :* The function φ is subanalytic and the gradient of each of its entries J of size larger than 1 is non vanishing a.e. and for $J = \{j\}$, $\varphi[\{j\}] \equiv 0$. ✓

The density of expected persistence diagrams

Theorem : Fix $n \geq 1$. Assume that :

- M is a real analytic compact d-dimensional connected submanifold possibly with boundary,
- $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$,
- $\mathcal{K}$ satisfies the assumptions (K1)-(K5).

Then, for $s \geq 0$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on the half plane $\Delta = \{(b, d) \in \mathbb{R}^2 : b \leq d\}$.

The density of expected persistence diagrams

Theorem : Fix $n \geq 1$. Assume that :

- M is a real analytic compact d-dimensional connected riemannian manifold possibly with boundary,
- $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$,
- $\mathcal{K}$ satisfies the assumptions (K1)-(K4) and (K5').

Then, for $s \geq 1$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ. Moreover, $E[D_0[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on the vertical line $\{0\} \times [0, \infty)$.

The density of expected persistence diagrams

Theorem : Fix $n \geq 1$. Assume that :

- M is a real analytic compact d-dimensional connected riemannian manifold possibly with boundary,
- $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$,
- $\mathcal{K}$ satisfies the assumptions (K1)-(K4) and (K5').

Then, for $s \geq 1$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ. Moreover, $E[D_0[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on the vertical line $\{0\} \times [0, \infty)$.

Theorem [smoothness] : Under the assumption of previous theorem, if moreover $\mathbb{X} \in M^n$ has a density of class C^k with respect to $\mathcal{H}_{nd}$. Then, for $s \geq 0$, the density of $E[D_s[\mathcal{K}(\mathbb{X})]]$ is of class C^k.

Sketch of proof

1. There exists a partition of the complement of a (subanalytic) set of measure 0 in M^n by open sets $V_1, \cdots, V_R$ such that :

- the order of the simplices of $\mathcal{K}(x)$ is constant on each V_r,
- for any $r = 1, \cdots, R$, and any $x \in V_r$,

  $$D_s[\mathcal{K}(x)] = \sum_{i=1}^{N_r} \delta_{\mathbf{r}_i}$$

  with $\mathbf{r}_i = (\varphi[J_{i_1}](x), \varphi[J_{i_2}](x))$ where N_r, J_{i_1}, J_{i_2} only depends on V_r.
- J_{i_1}, J_{i_2} can be chosen so that the differential of

  $$\Phi_{ir} : x \in V_r \to \mathbf{r}_i = (\varphi[J_{i_1}](x), \varphi[J_{i_2}](x))$$

  has maximal rank (2).

Sketch of proof

2. The expected diagram can be written as

$$\begin{aligned} E[D_s[\mathcal{K}(\mathbb{X})]] &= \sum_{r=1}^{R} E\left[\mathbb{1}\{\mathbb{X} \in V_r\} D_s[\mathcal{K}(\mathbb{X})]\right] = \sum_{r=1}^{R} E\left[\mathbb{1}\{\mathbb{X} \in V_r\} \sum_{i=1}^{N_r} \delta_{\mathbf{r}_i}\right] \\ &= \sum_{r=1}^{R} \sum_{i=1}^{N_r} E\left[\mathbb{1}\{\mathbb{X} \in V_r\} \delta_{\mathbf{r}_i}\right] \end{aligned}$$

Sketch of proof

2. The expected diagram can be written as

$$\begin{aligned}
E[D_s[\mathcal{K}(\mathbb{X})]] &= \sum_{r=1}^{R} E\left[\mathbb{1}\{\mathbb{X}\in V_r\}D_s[\mathcal{K}(\mathbb{X})]\right] = \sum_{r=1}^{R} E\left[\mathbb{1}\{\mathbb{X}\in V_r\}\sum_{i=1}^{N_r}\delta_{\mathbf{r}_i}\right] \\
&= \sum_{r=1}^{R}\sum_{i=1}^{N_r} \underbrace{E\left[\mathbb{1}\{\mathbb{X}\in V_r\}\delta_{\mathbf{r}_i}\right]}_{\mu_{ir}}
\end{aligned}$$

3. Use the co-area formula :

Density of $\mathbb{X}$

$$\begin{aligned}
\mu_{ir}(B) &= P(\Phi_{ir}(\mathbb{X})\in B, \mathbb{X}\in V_r) \\
&= \int_{V_r} \mathbb{1}\{\Phi_{ir}(x)\in B\}\kappa(x)d\mathcal{H}_{nd}(x) \\
&= \int_{u\in B} \underbrace{\int_{x\in\Phi_{ir}^{-1}(u)} (J\Phi_{ir}(x))^{-1}\kappa(x)d\mathcal{H}_{nd-2}(x)}_{\text{Density of } \mu_{ir}} du.
\end{aligned}$$

The Hausdorff measure and the co-area formula

Definition : Let k be a non-negative number. For $A \subset \mathbb{R}^D$, and $\delta > 0$, consider

$$\mathcal{H}_k^\delta(A) := \inf \left\{ \sum_i \text{diam}(U_i)^k, A \subset \bigcup_i U_i \text{ and } \text{diam}(U_i) < \delta \right\}.$$

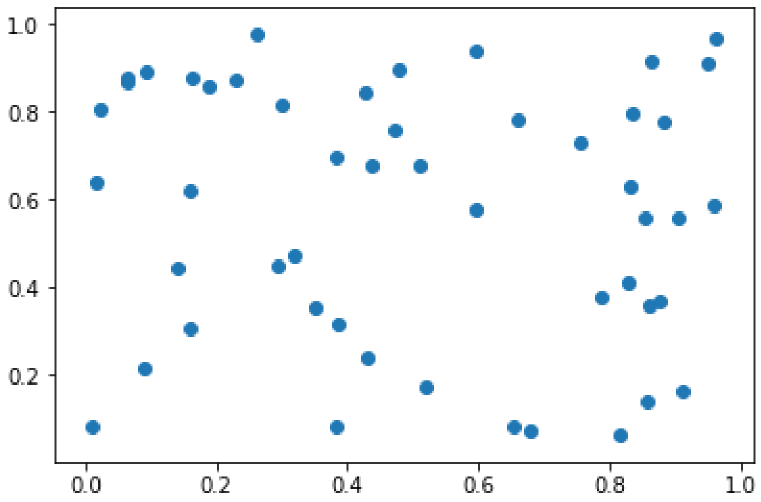
The *k-dimensional Haussdorf measure* on $\mathbb{R}^D$ of A is defined by $\mathcal{H}_k(A) := \lim_{\delta \to 0} \mathcal{H}_k^\delta(A)$.

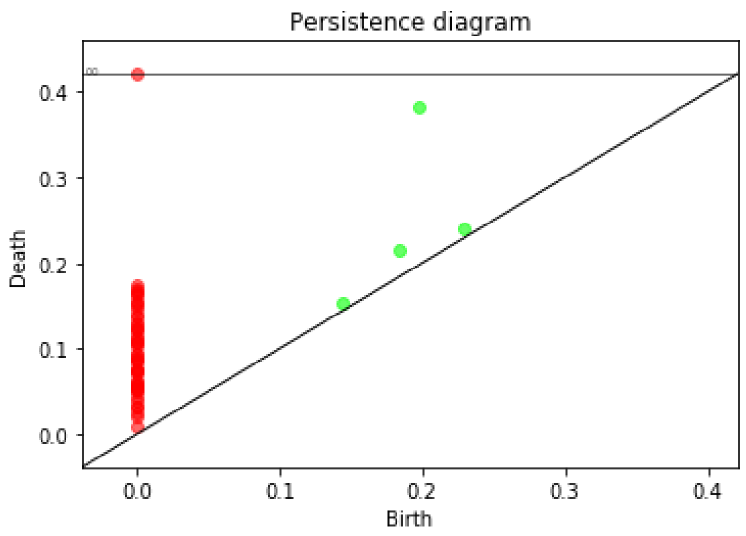
Theorem [Co-area formula] : Let M (resp. N) be a smooth Riemannian manifold of dimension m (resp n). Assume that $m \geq n$ and let $\Phi : M \to N$ be a differentiable map. Denote by $D\Phi$ the differential of Φ. The Jacobian of Φ is defined by $J\Phi = \sqrt{\det((D\Phi) \times (D\Phi)^t)}$. For $f : M \to N$ a positive measurable function, the following equality holds :

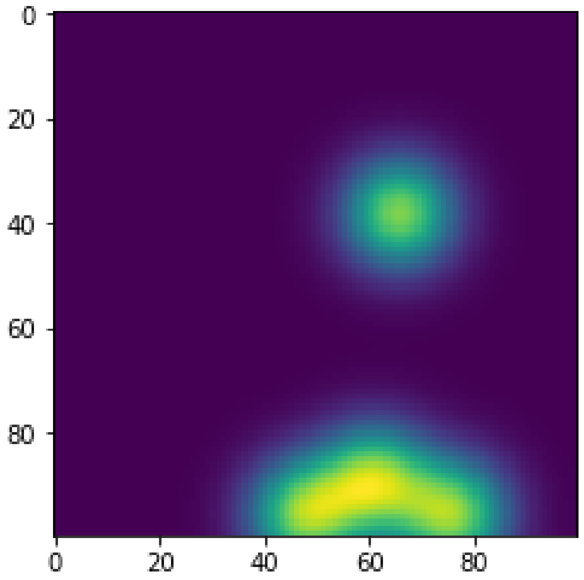
$$\int_M f(x) J\Phi(x) d\mathcal{H}_m(x) = \int_N \left(\int_{x \in \Phi^{-1}(\{y\})} f(x) d\mathcal{H}_{m-n}(x) \right) d\mathcal{H}_n(y).$$

Persistence images

[Adams et al, JMLR 2017]







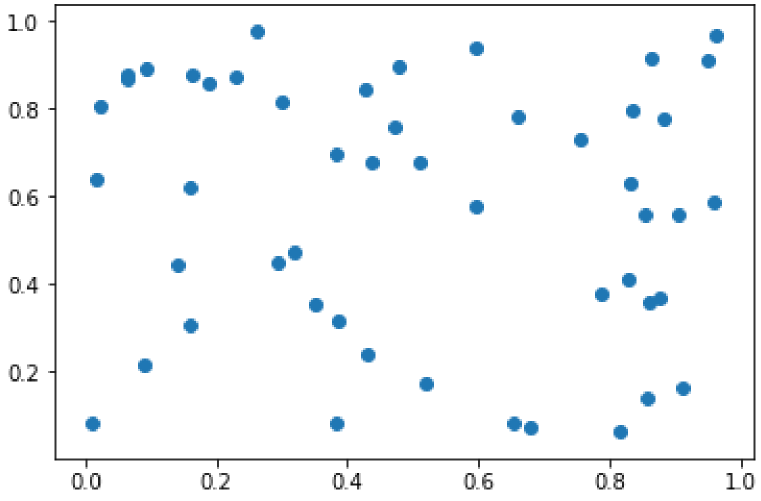
For $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel and H a bandwidth matrix (e.g. a symmetric positive definite matrix), pose for $u \in \mathbb{R}^2$, $K_H(z) = |H|^{-1/2} K(H^{-1/2} \cdot u)$

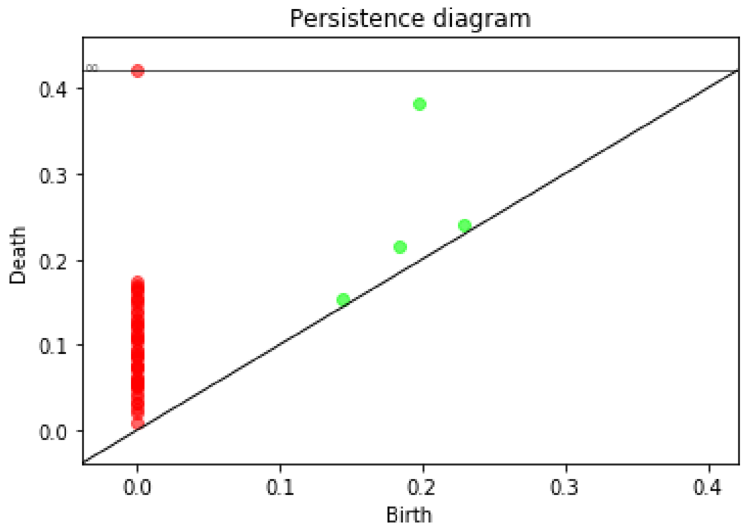
For $D = \sum_i \delta_{\mathbf{r}_i}$ a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by :

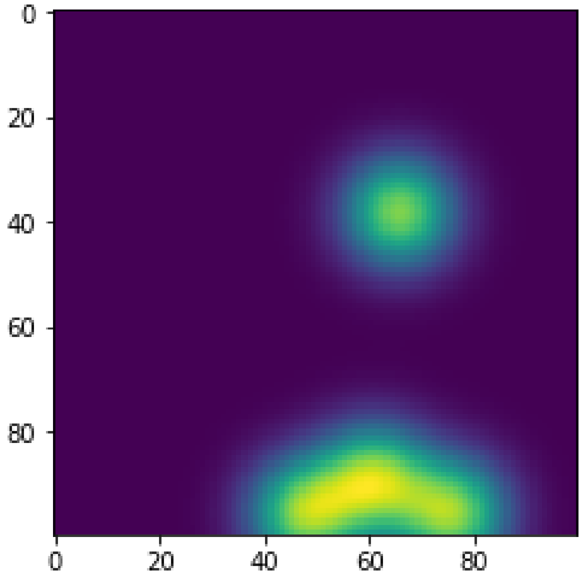
$$\forall z \in \mathbb{R}^2,\ \rho(D)(u) = \sum_i w(\mathbf{r}_i) K_H(u - \mathbf{r}_i) = D(wK_H(u - \cdot))$$

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 $u \in \mathbb{R}^2$, $K_H(z) = |H|^{-1/2} K(H^{-1/2} \cdot u)$

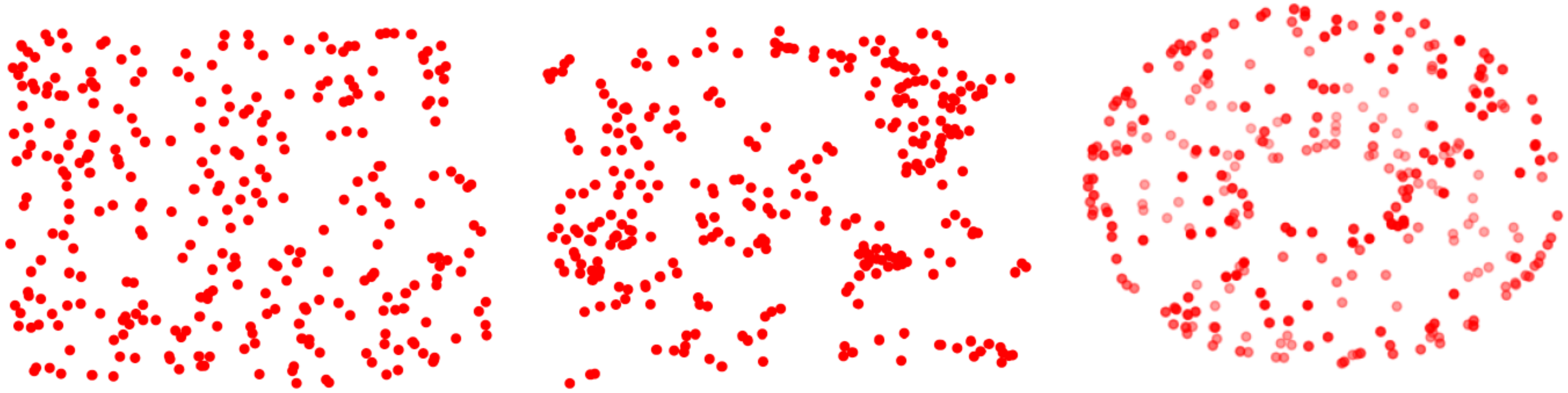
For $D = \sum_i \delta_{\mathbf{r}_i}$ a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by :

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

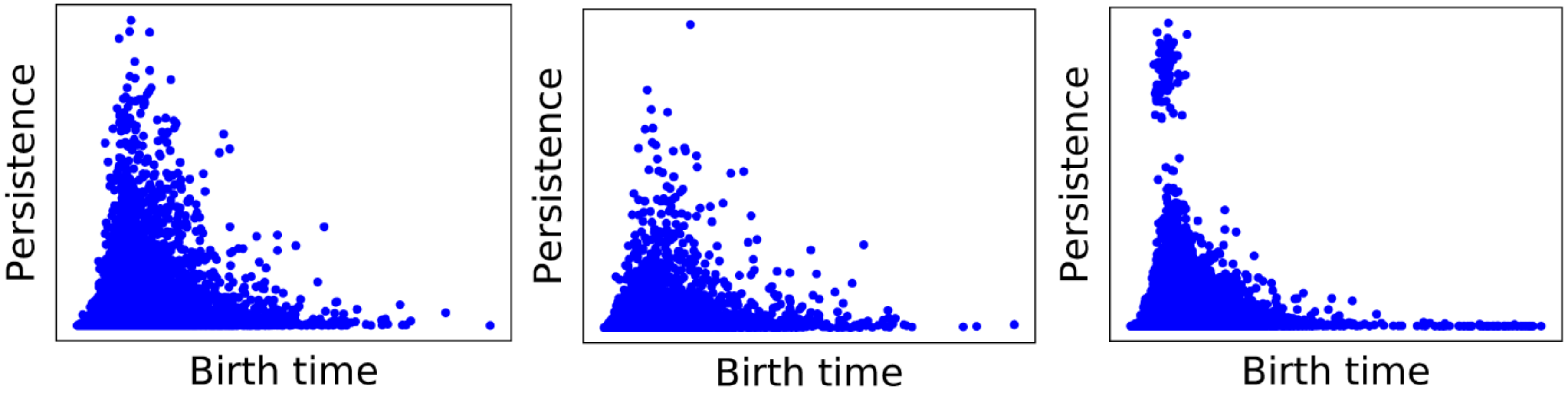
$\Rightarrow$ persistence surfaces can be seen as kernel estimates of $E[D_s[\mathcal{K}(\mathbb{X})]]$.

Persistence images

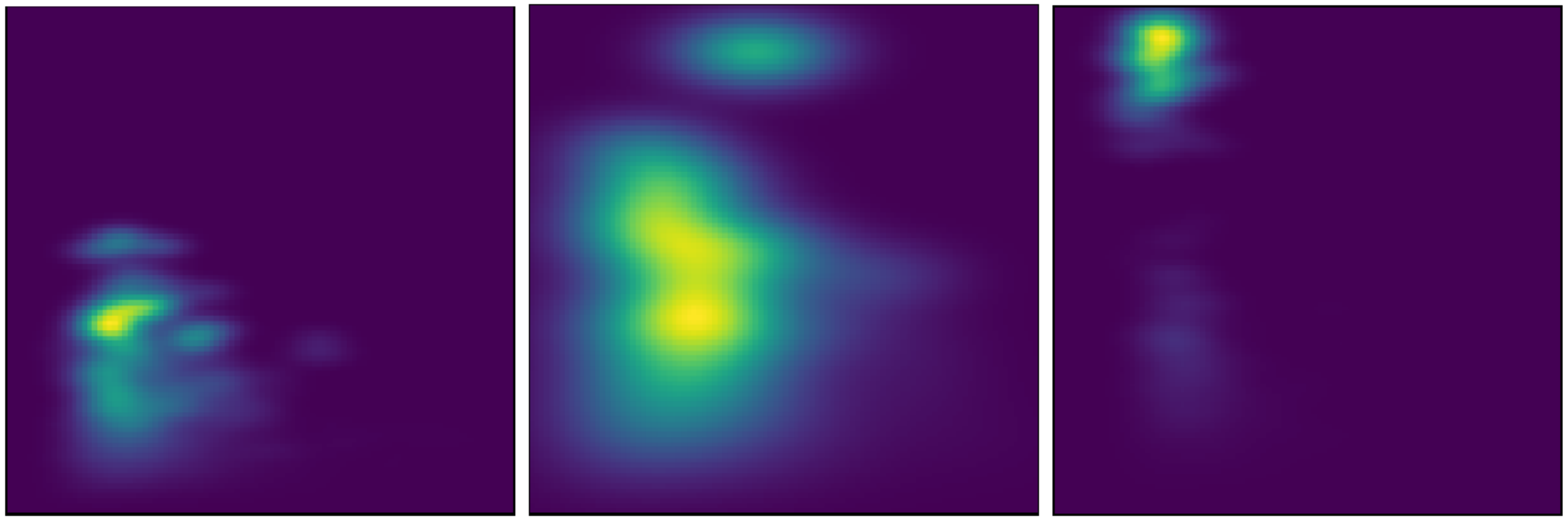
The realization of 3 different processes



The overlay of 40 different persistence diagrams



The persistence images with weight function $w(\mathbf{r}) = (r_2 - r_1)^3$ and bandwith selected using cross-validation.



Thank you for your attention !