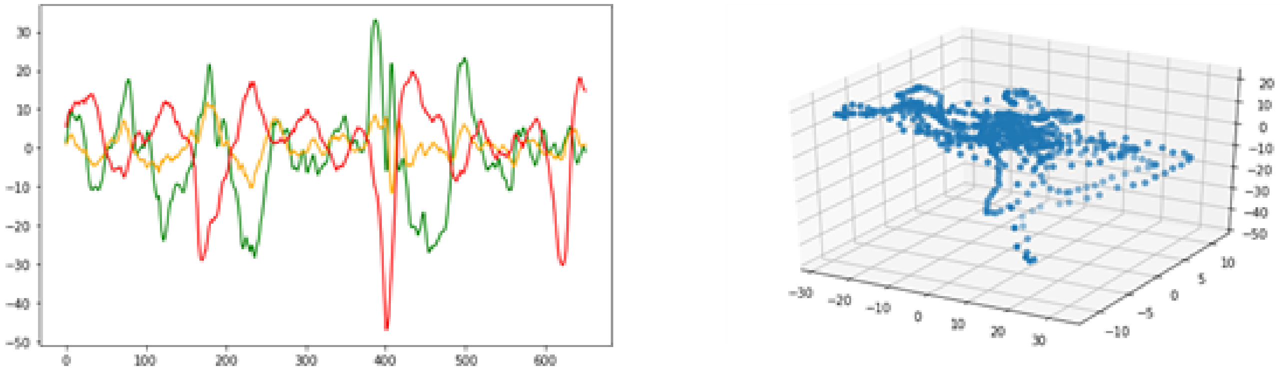
Ecole GEOMDATA
Fréjus - Sept 2018

A short introduction to Topologial Data Analysis

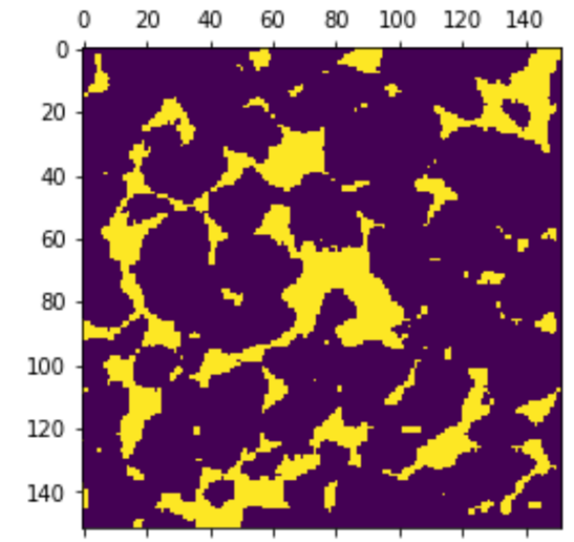
Frédéric Chazal and Marc Glisse
DataShape team
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr marc.glisse@inria.fr

informatics mathematics

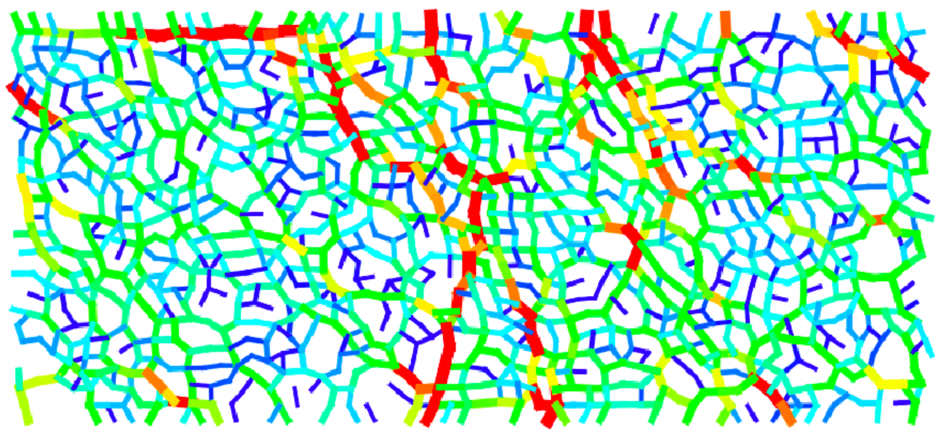
Introduction



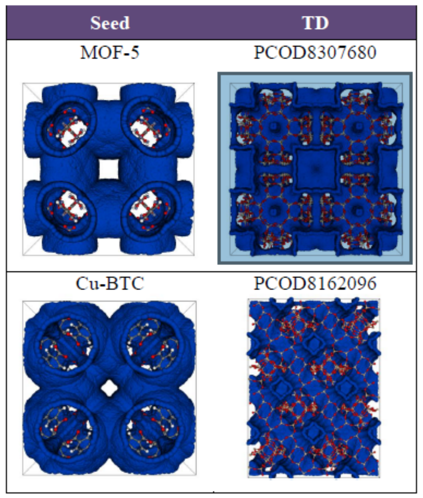
[Sensors]



[3D images (porous rocks)]



[Force fields in granular media]

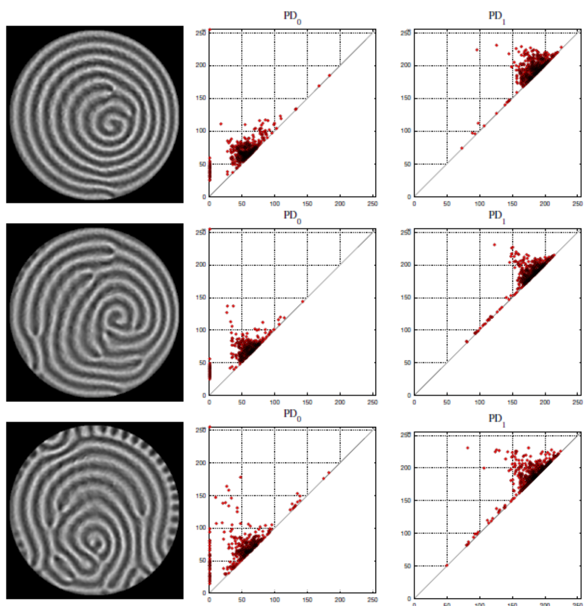
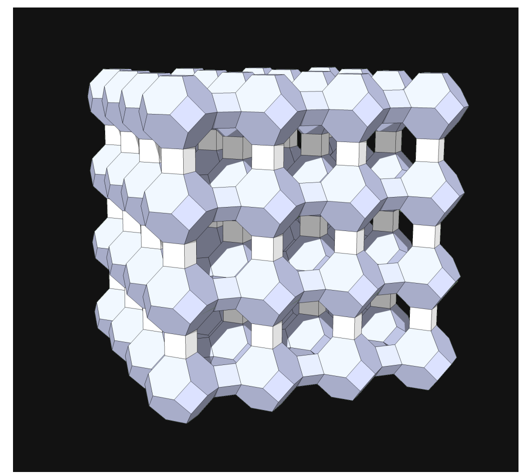
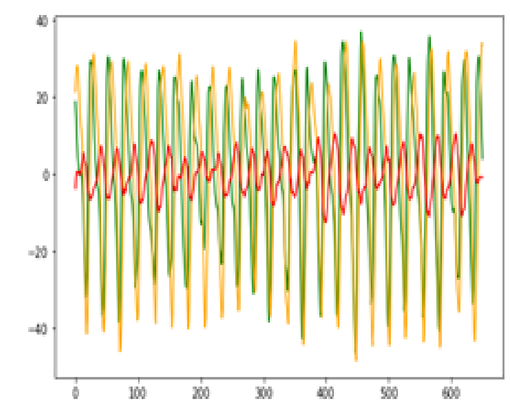
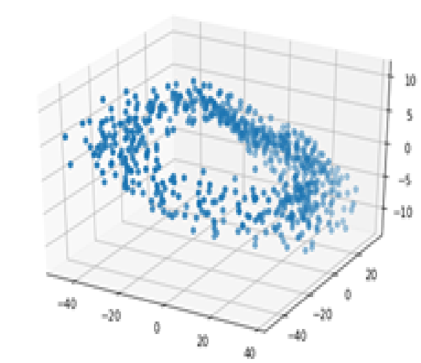


[Nano-materials - Li et al 2017]

Data often come as (sampling of) metric spaces or sets/spaces endowed with a similarity measure with, possibly complex, topological/geometric structure.

What is Topological Data Analysis (TDA)?

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

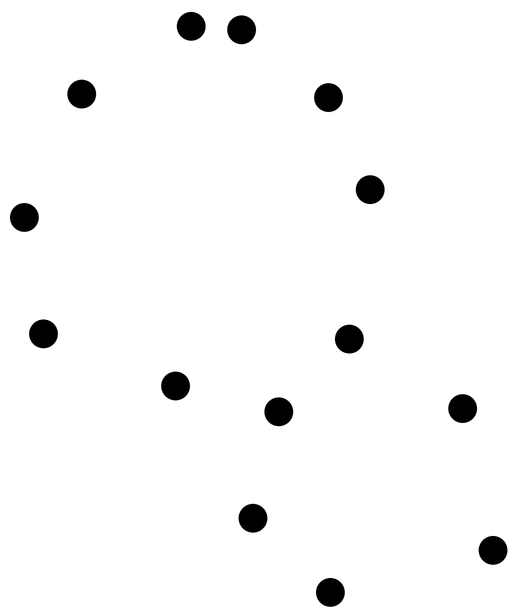
- Well-founded mathematical methods to infer and exploit relevant topological and geometric features (feature engineering) from data for exploratory data analysis, Machine Learning,...

- New and innovative tools for complex data to be used with or in complement of other ML and AI tools.

- High quality, efficient and easy-to-use software for TDA tools.

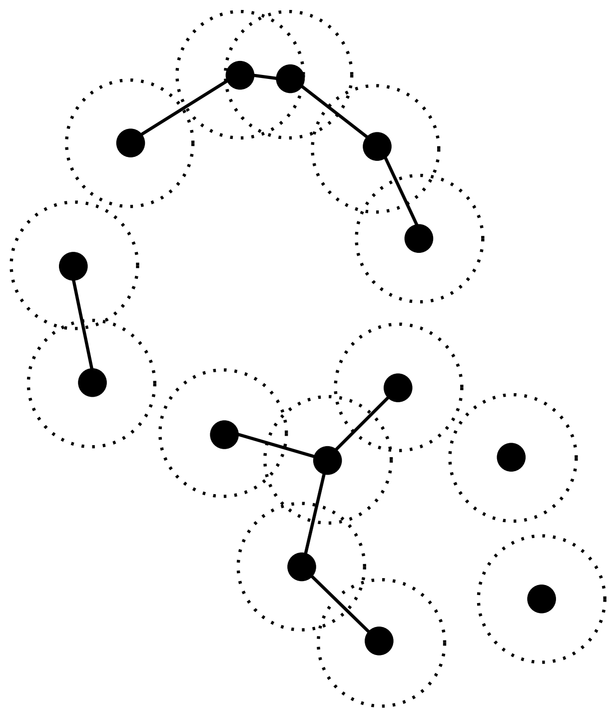
A recent and very active research field with already many successful applications

The standard TDA pipeline



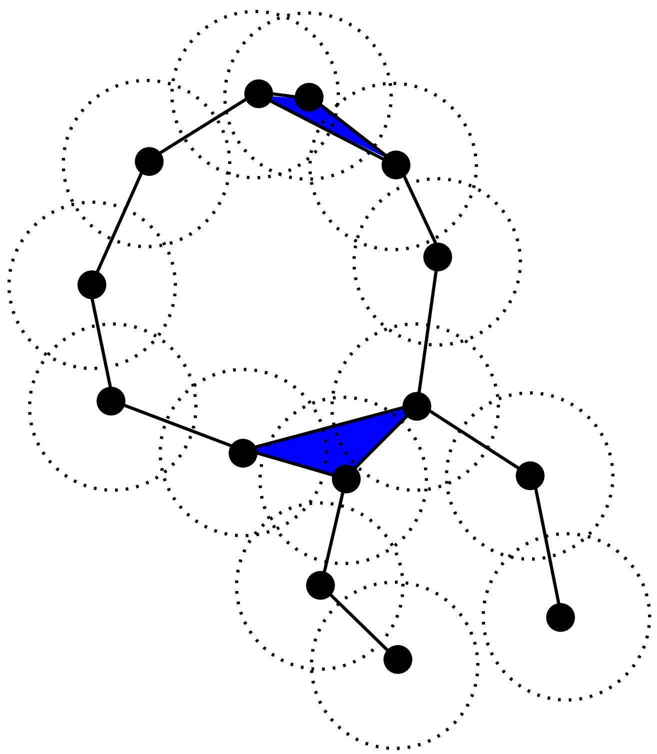
- Build a nested family of spaces (filtered simplicial complex) on top of data $\to$ multiscale topol. structure.
- Compute the persistent homology of the complex $\to$ multiscale topol. signature/features.
- Compare the signatures of “close” data sets $\to$ robustness and stability results.
- Statistical properties of signatures/features.

The standard TDA pipeline



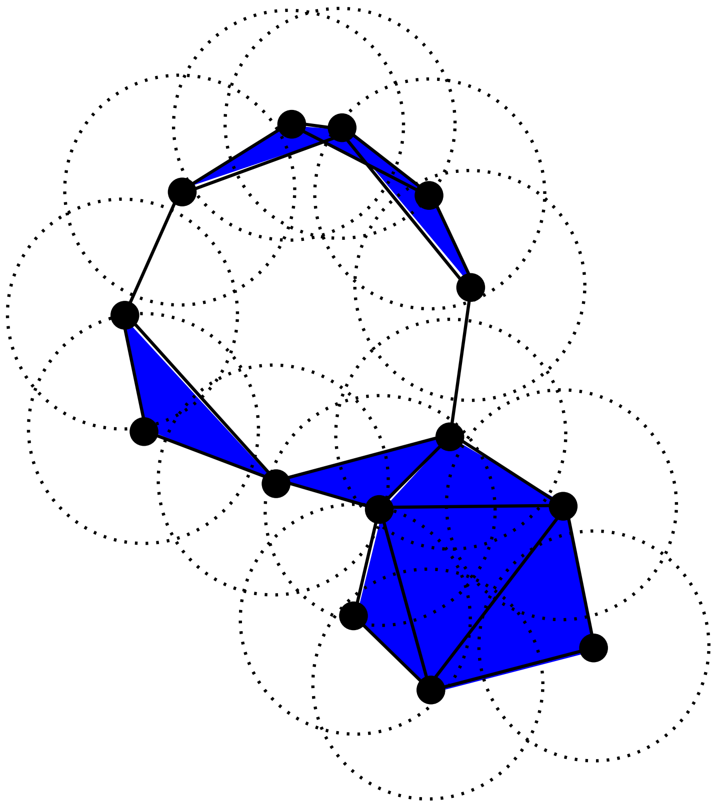
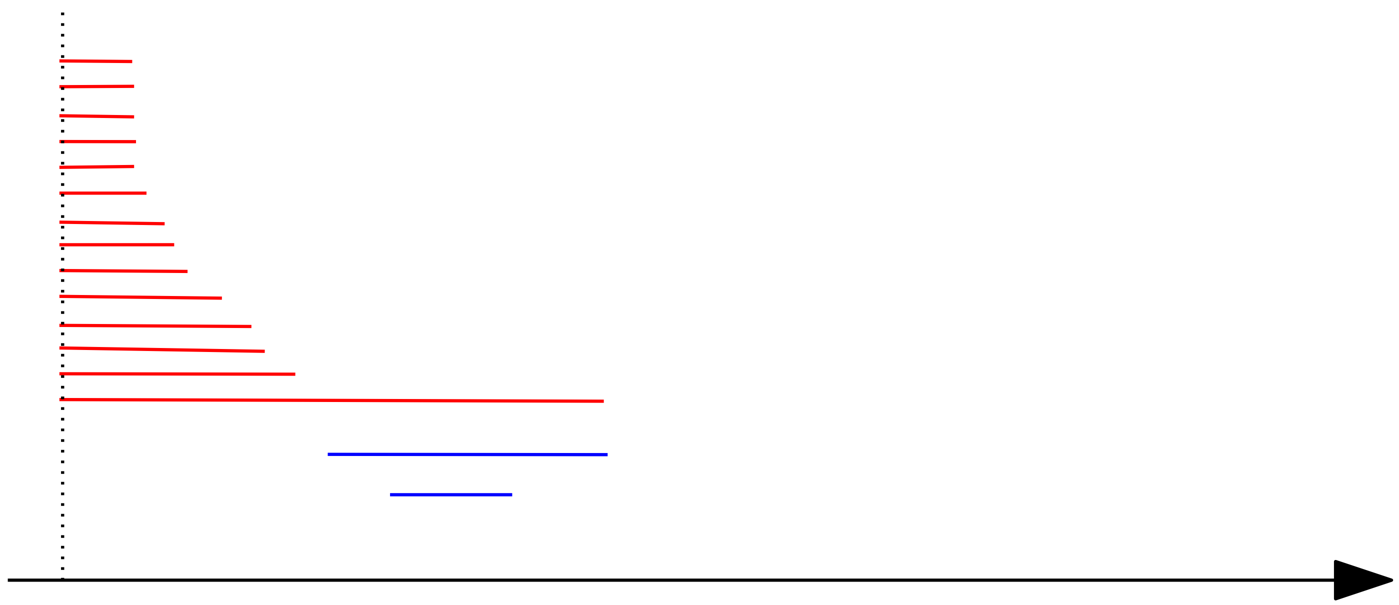
- Build a nested family of spaces (filtered simplicial complex) on top of data $\rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature/features.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures/features.

The standard TDA pipeline



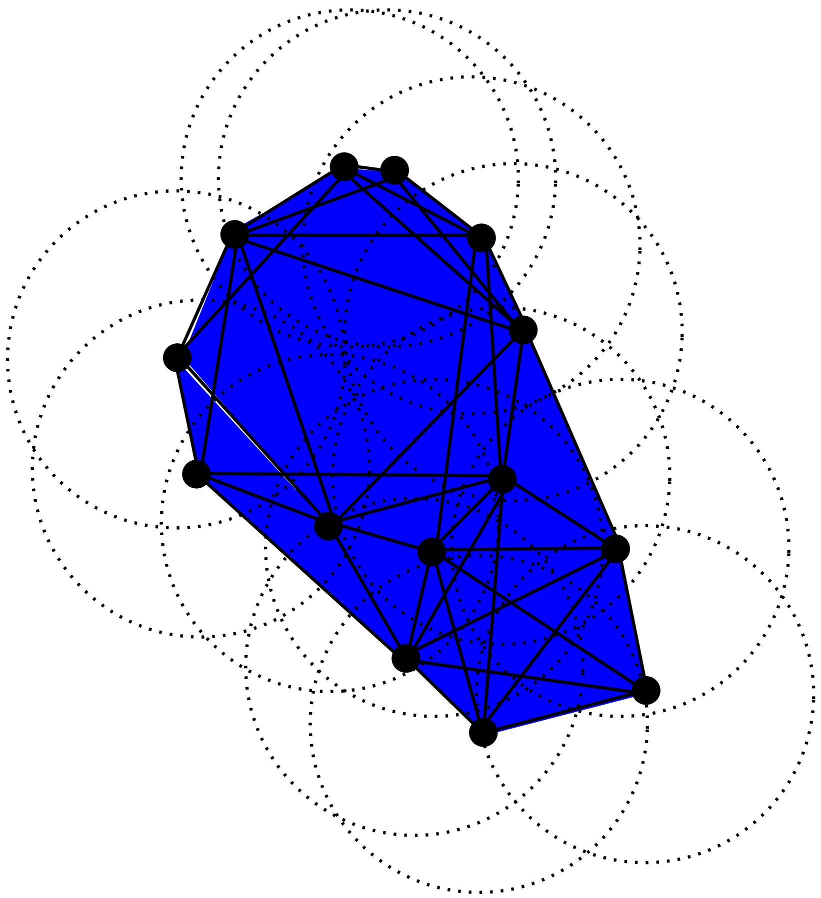
- Build a nested family of spaces (filtered simplicial complex) on top of data $\rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature/features.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures/features.

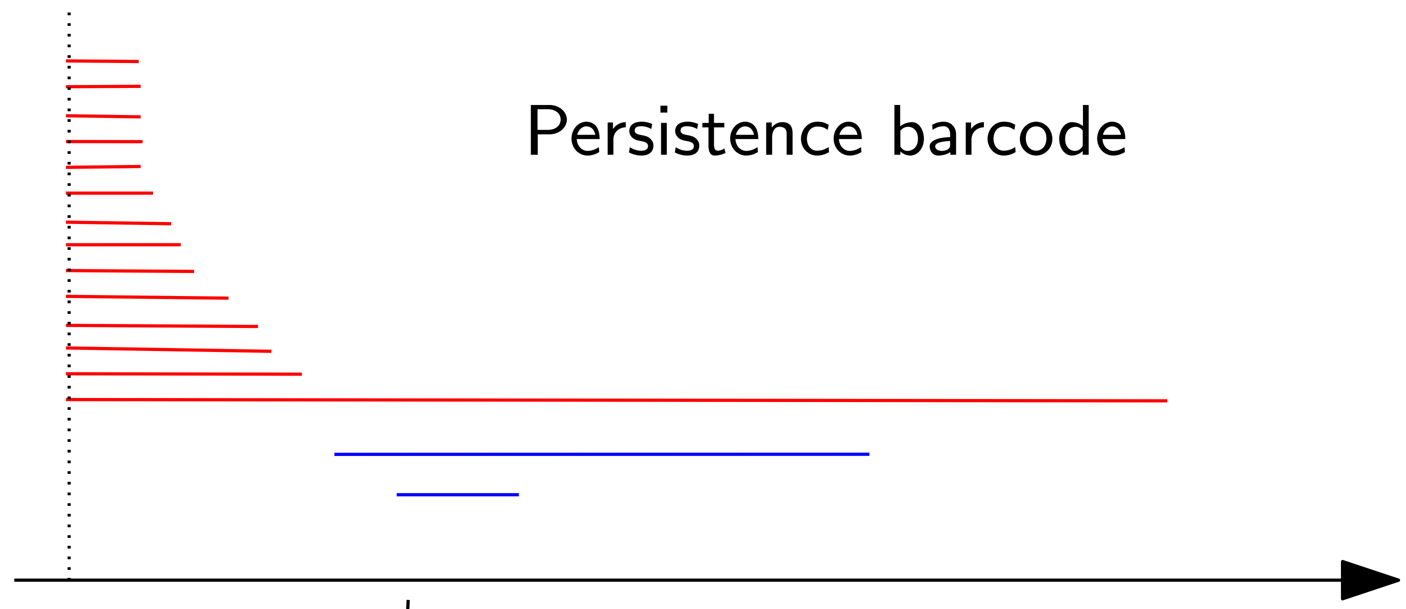
The standard TDA pipeline

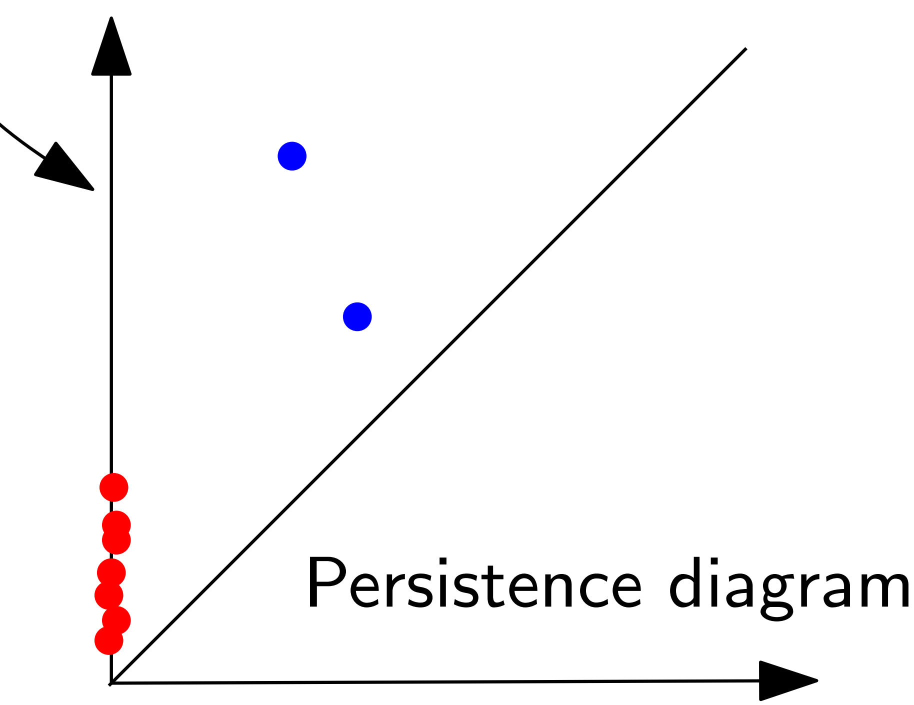
 

- Build a nested family of spaces (filtered simplicial complex) on top of data $\rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature/features.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures/features.

The standard TDA pipeline

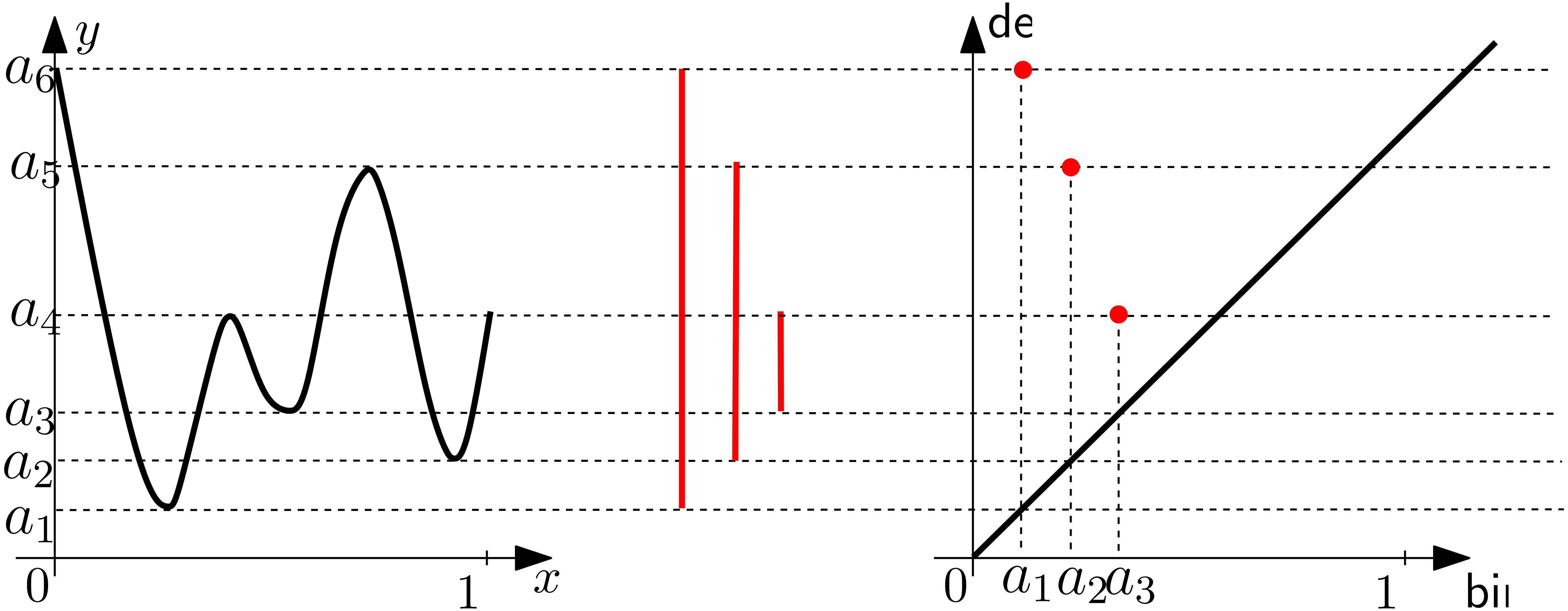






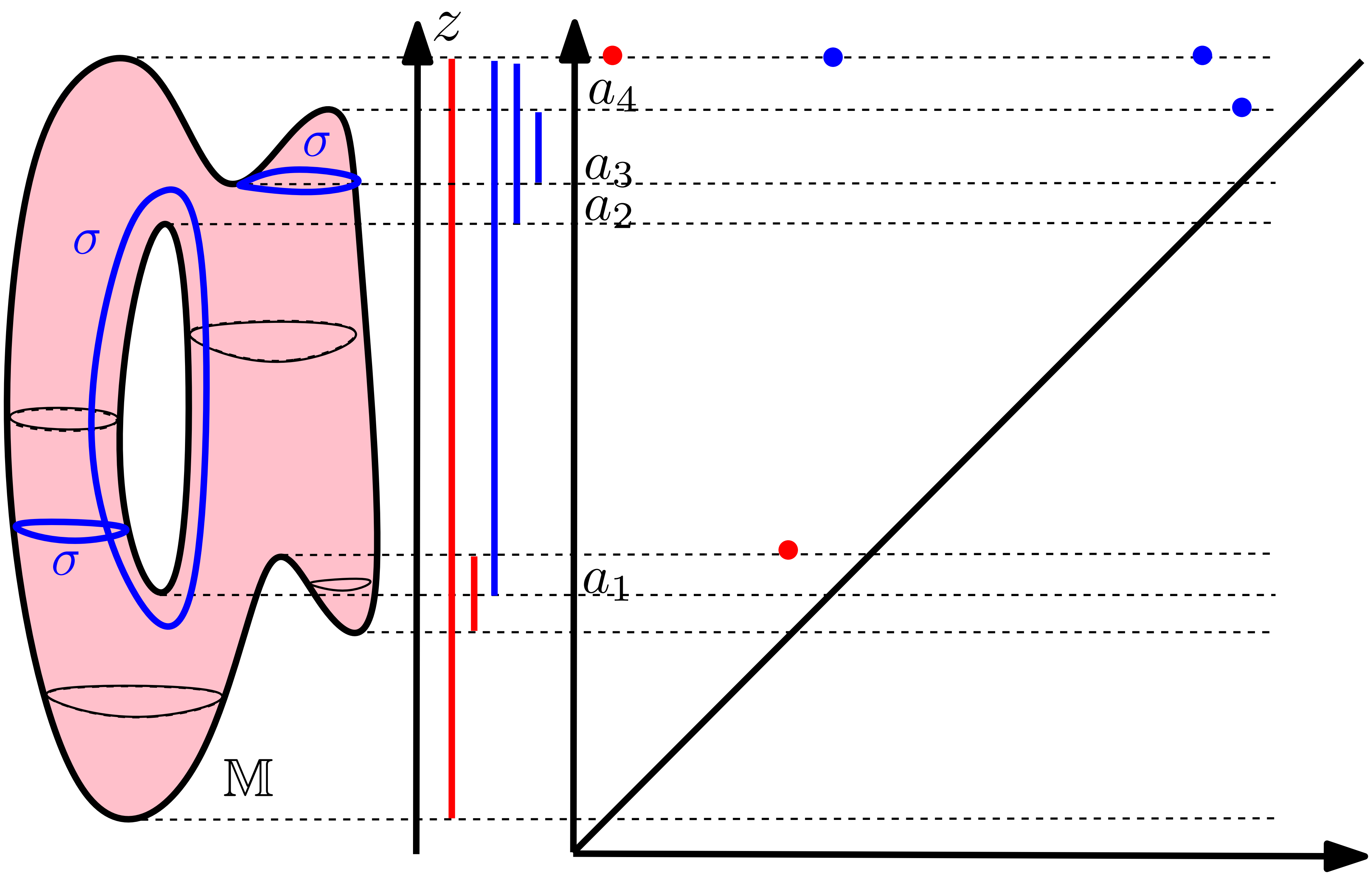
- Build a nested family of spaces (filtered simplicial complex) on top of data $\to$ multiscale topol. structure.
- Compute the persistent homology of the complex $\to$ multiscale topol. signature/features.
- Compare the signatures of "close" data sets $\to$ robustness and stability results.
- Statistical properties of signatures/features.

Persistent homology for functions



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

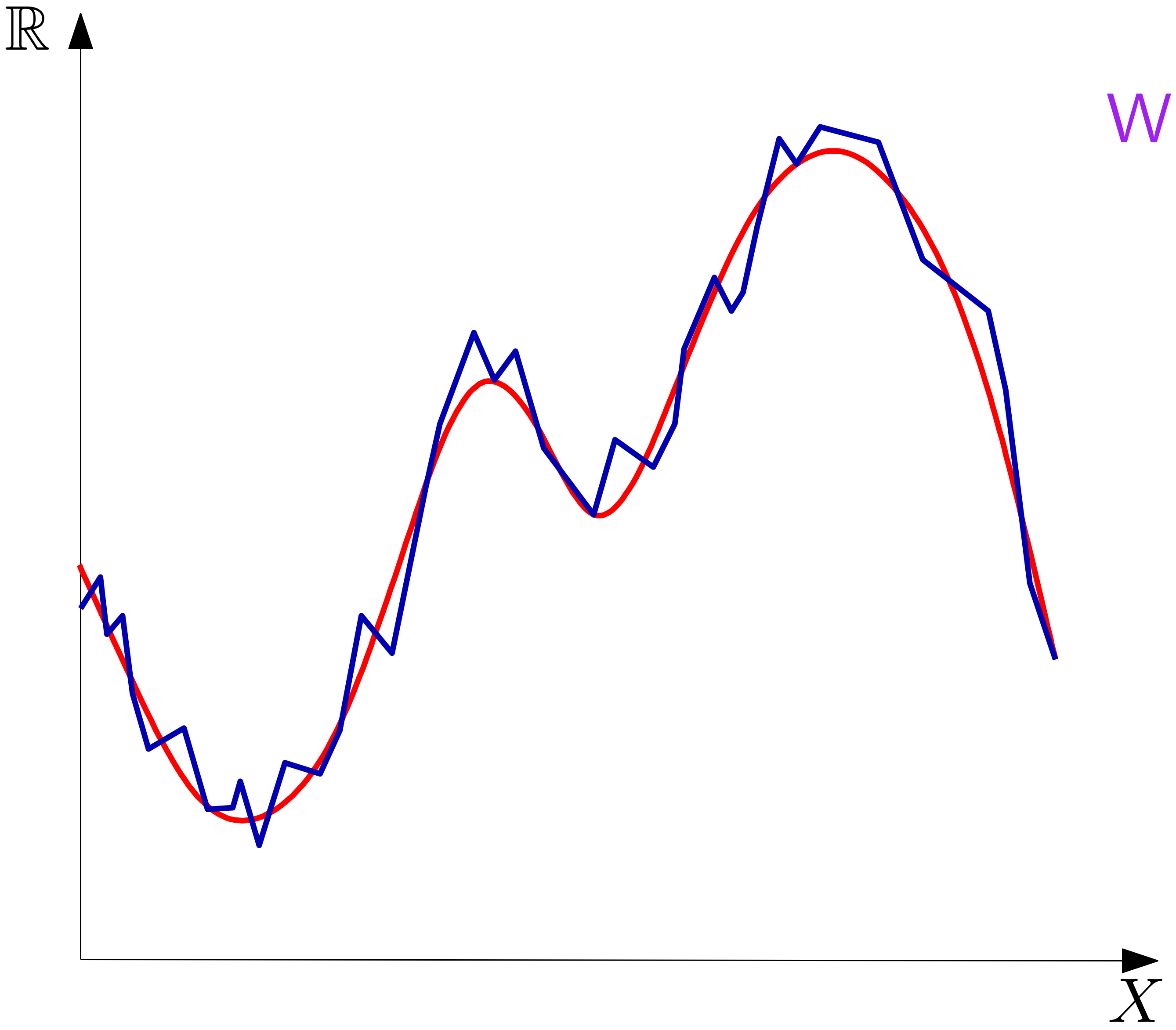
Persistent homology for functions



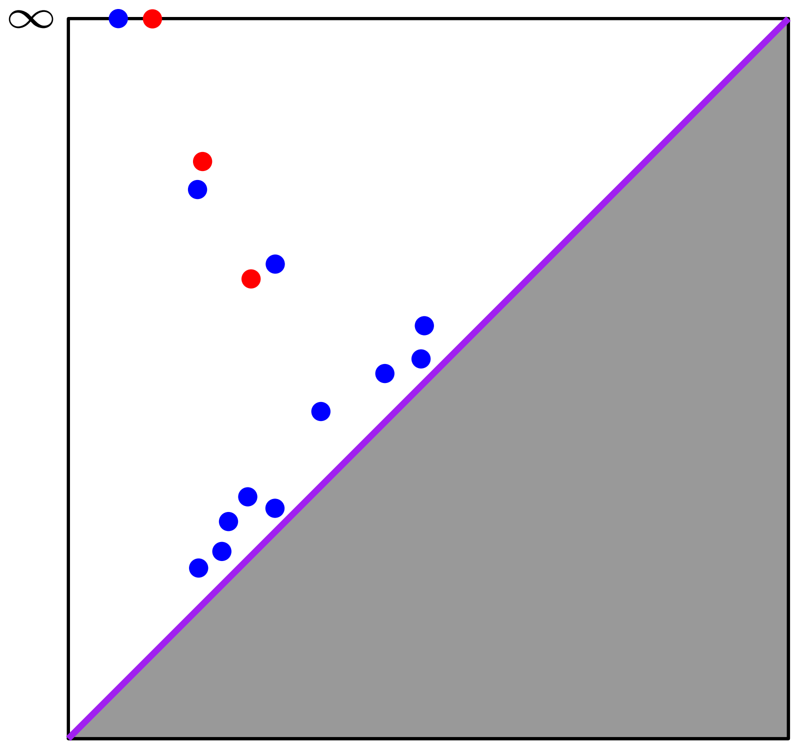
Tracking and encoding the evolution of the connected components (0-dimensional homology) and cycles (1-dimensional homology) of the sublevel sets.

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).

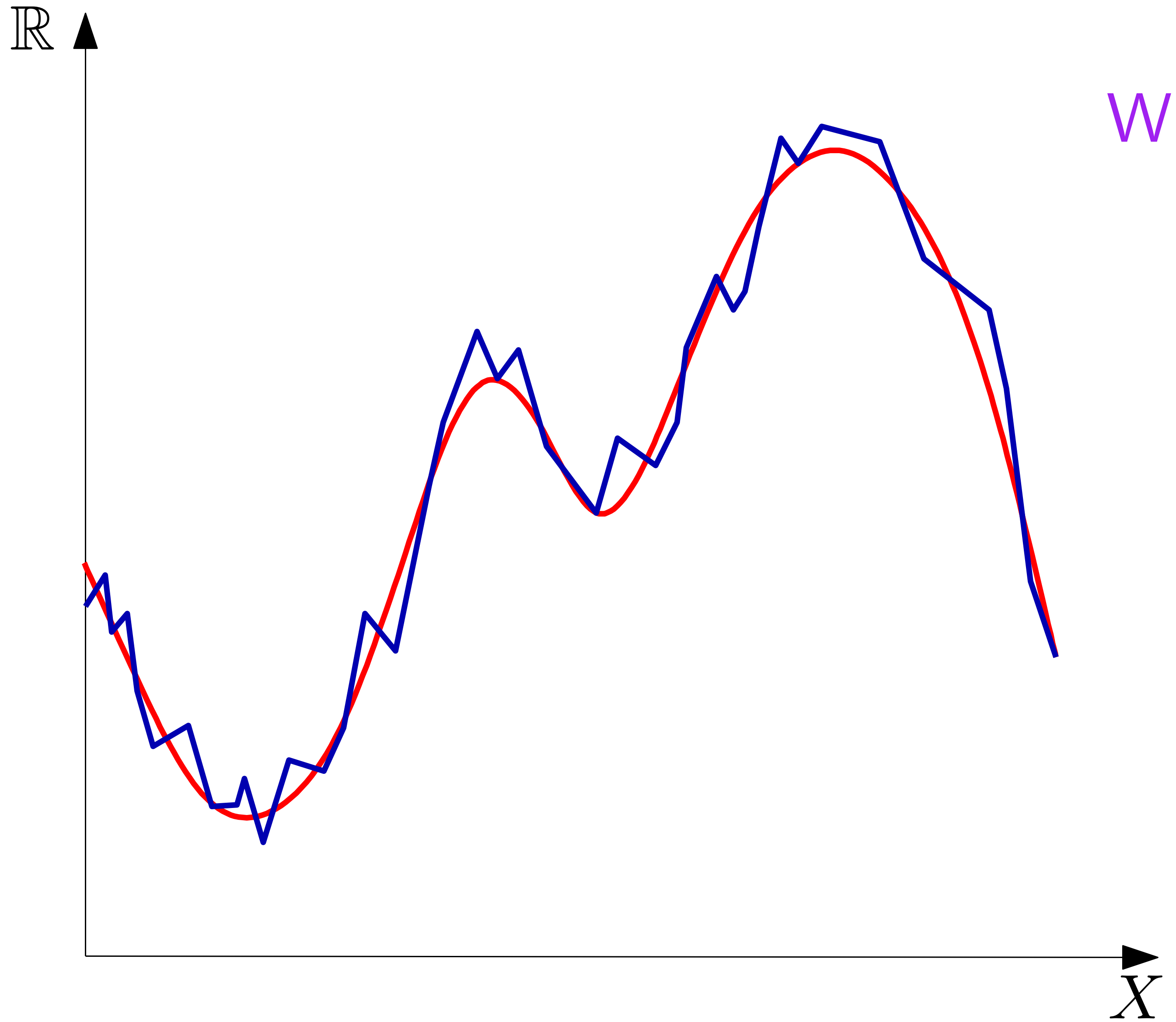
Stability properties



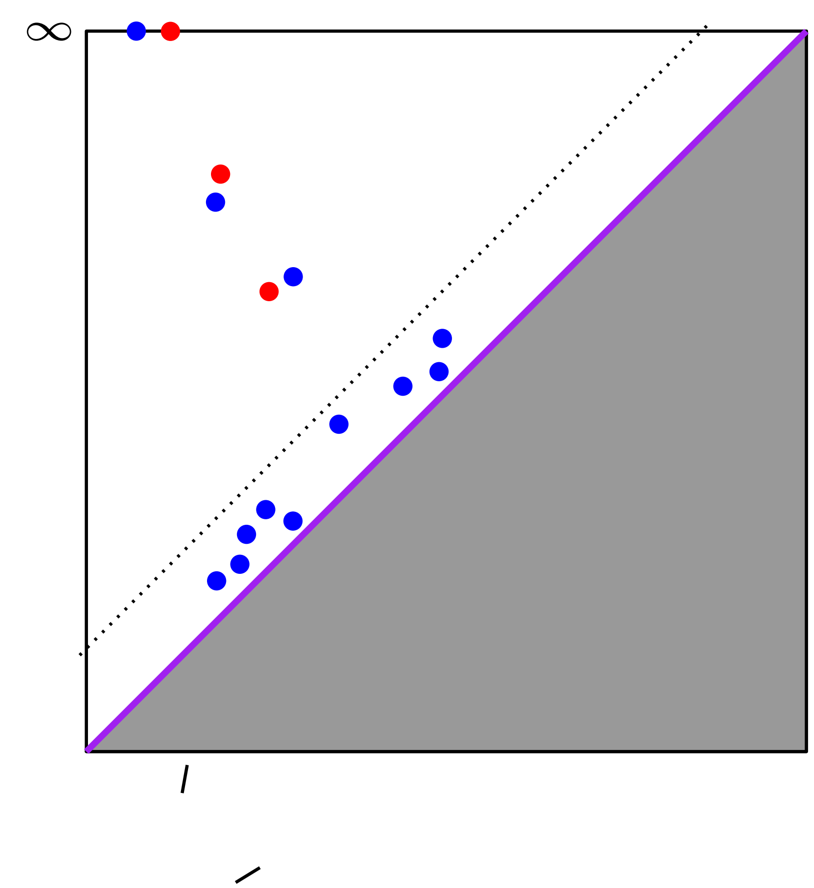
What if f is slightly perturbed?



Stability properties



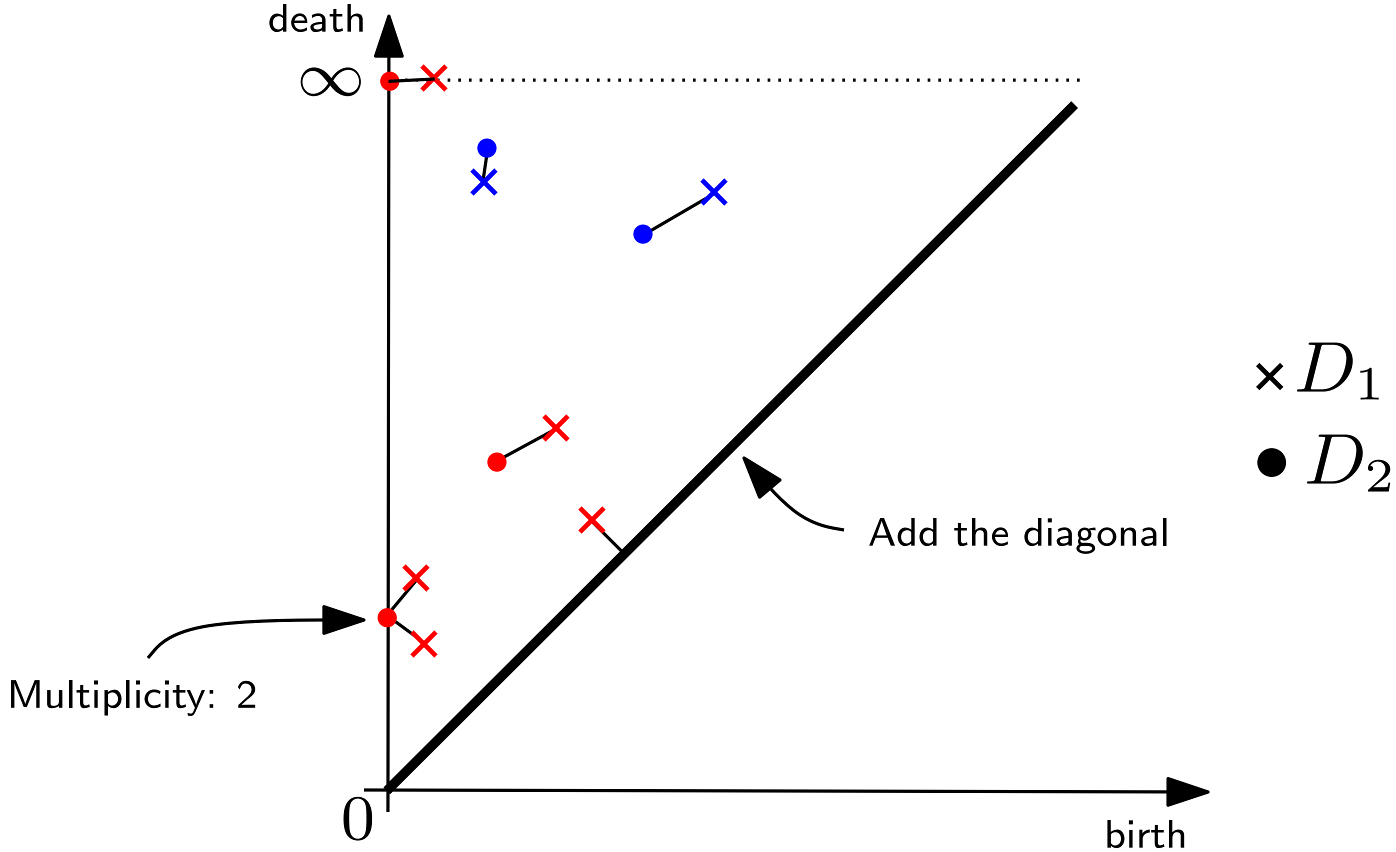
What if f is slightly perturbed?



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]

Comparing persistence diagrams



The bottleneck distance between two diagrams D_1 and D_2 is

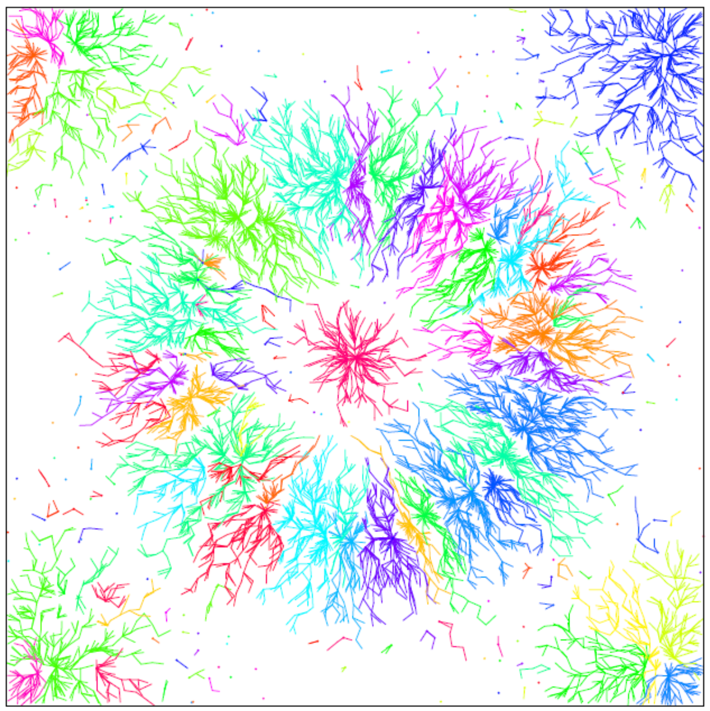
$$d_B(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty$$

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

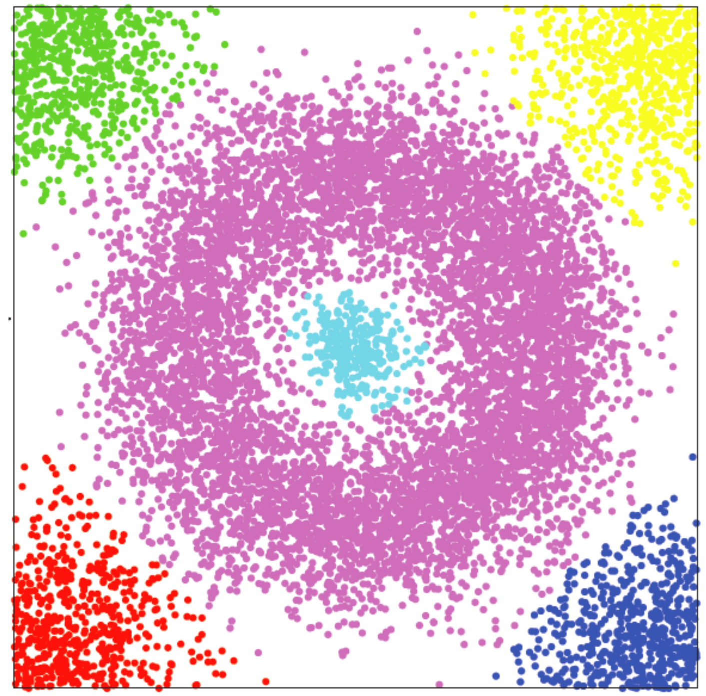
$\rightarrow$ Persistence diagrams provide easy to compare topological signatures.

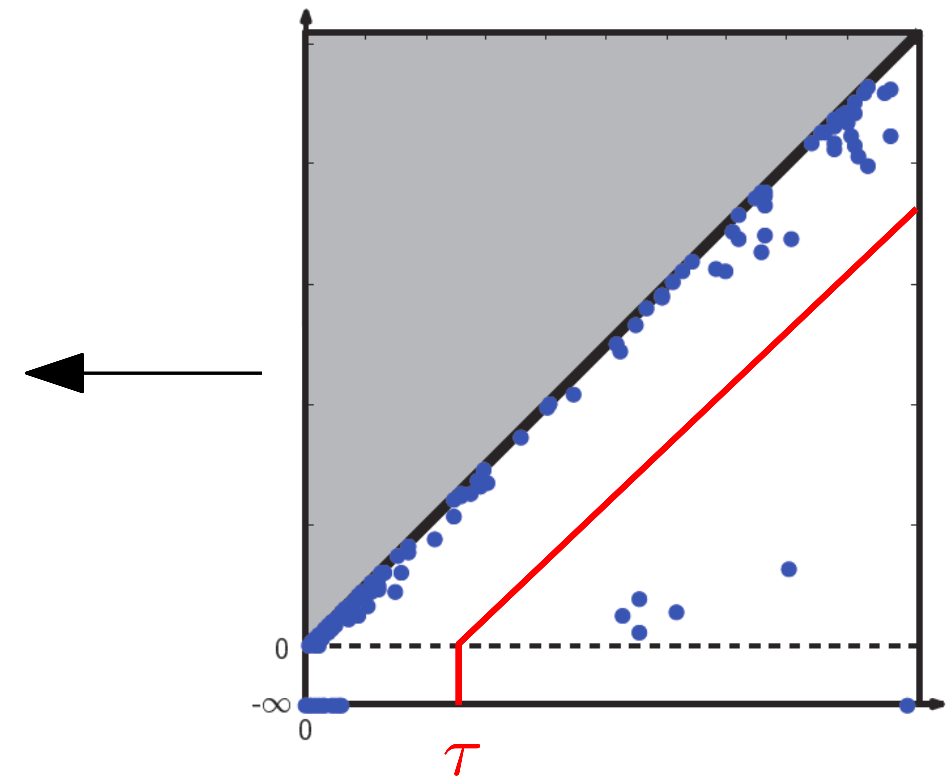
Some examples of applications

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

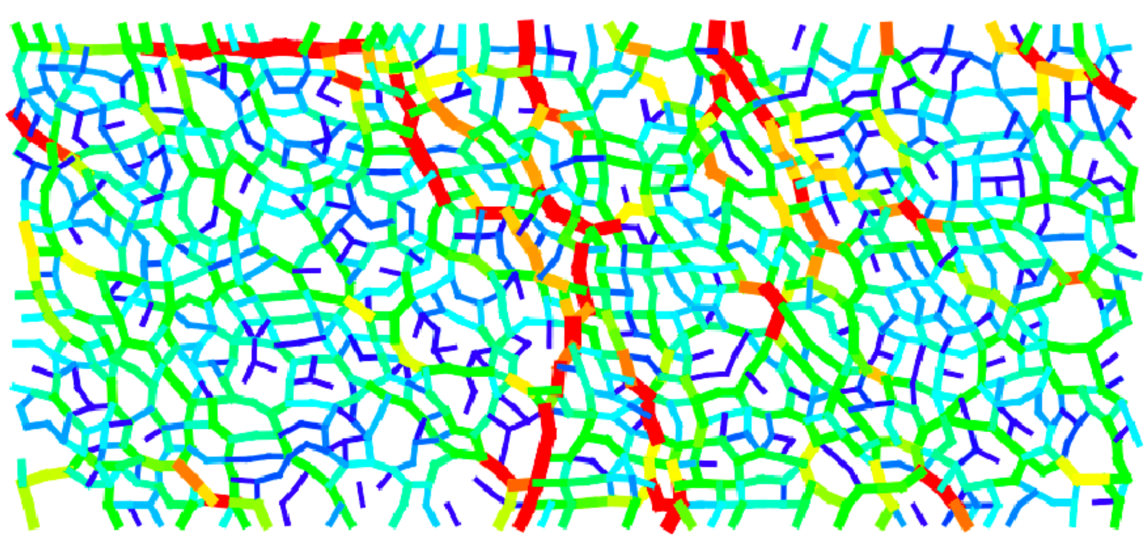


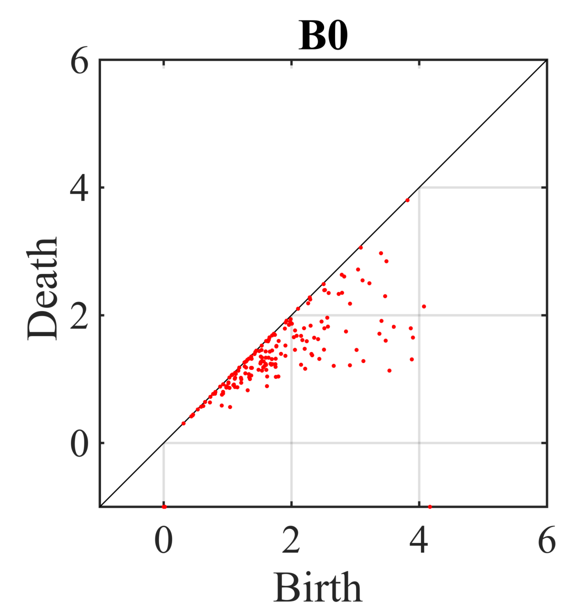
$\tau = 0$





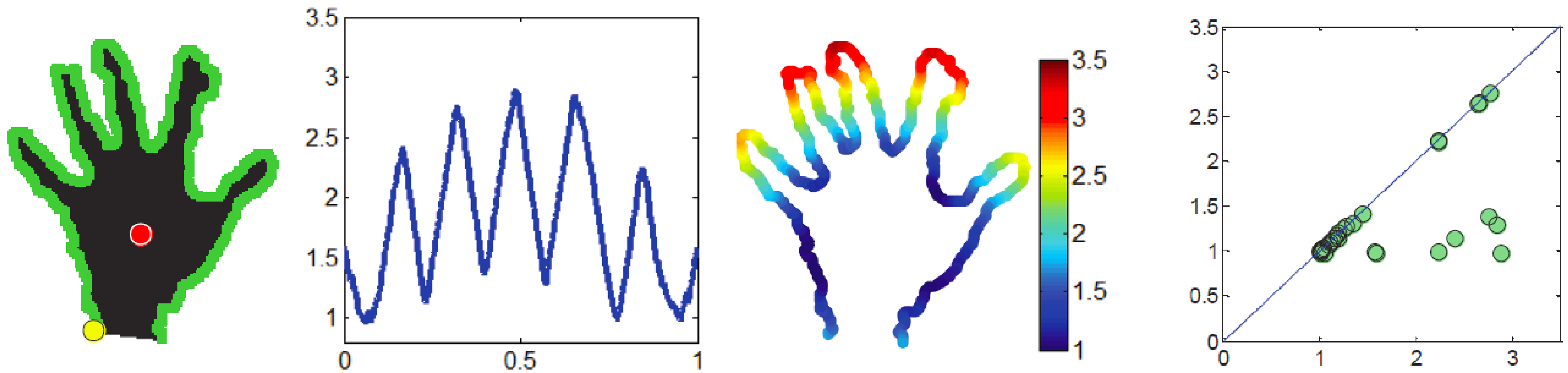
- Analysis of force fields in granular media [Kramar, Mischaikow et al]



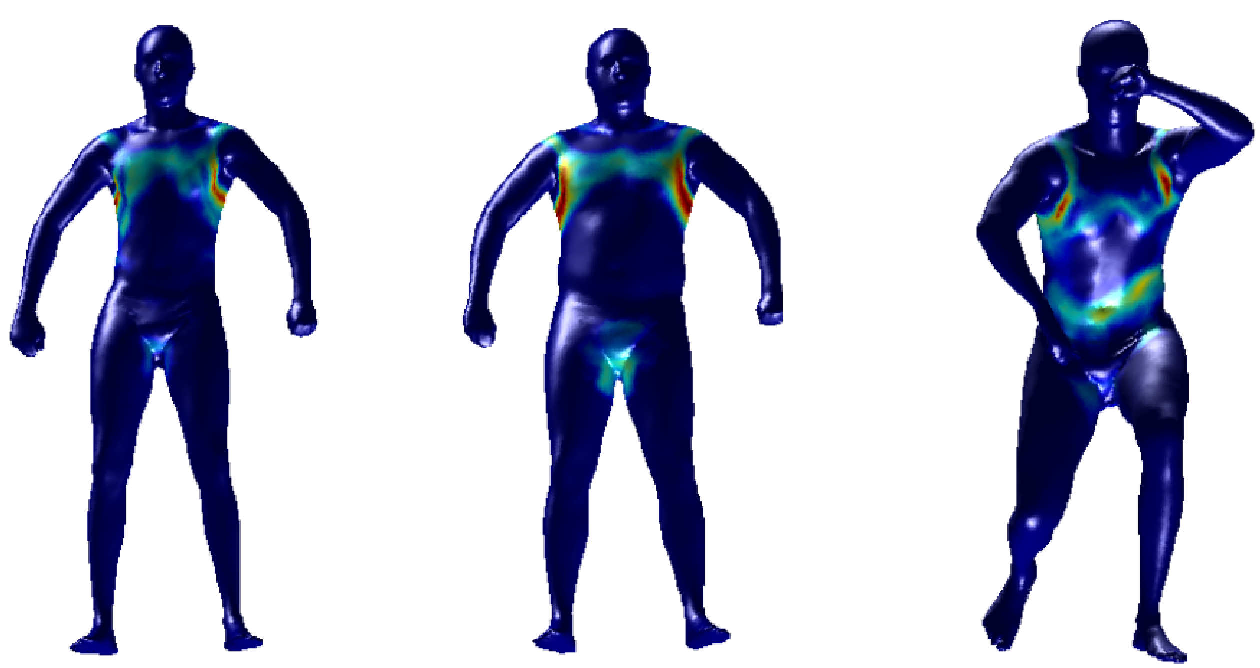


Some examples of applications

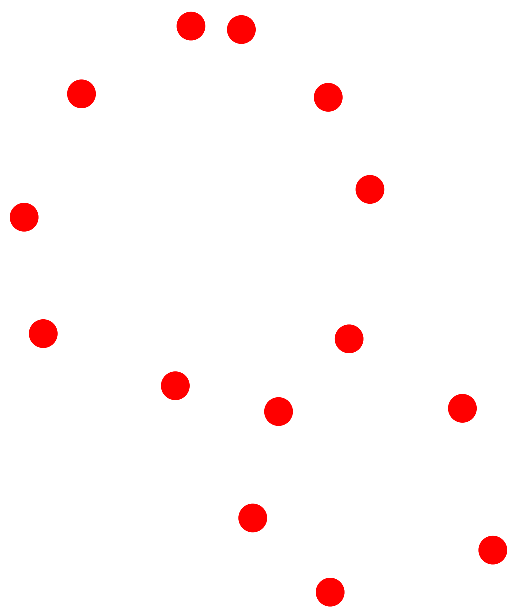
- Hand gesture recognition [Li, Ovsjanikov, C. - CVPR'14]



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

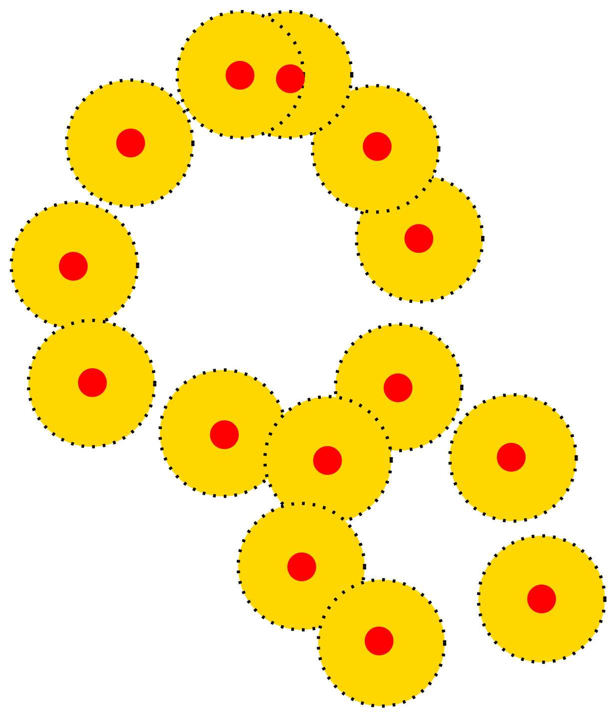


Persistent homology for point cloud data



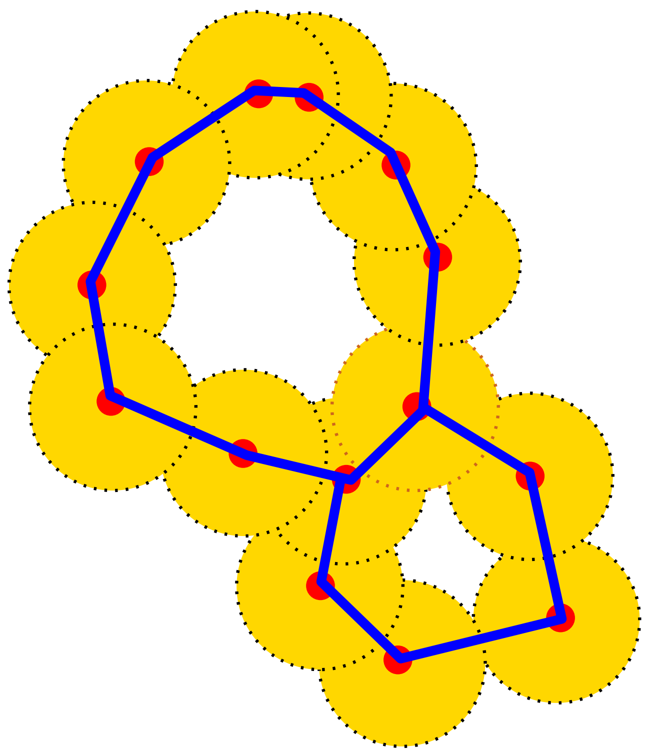
- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

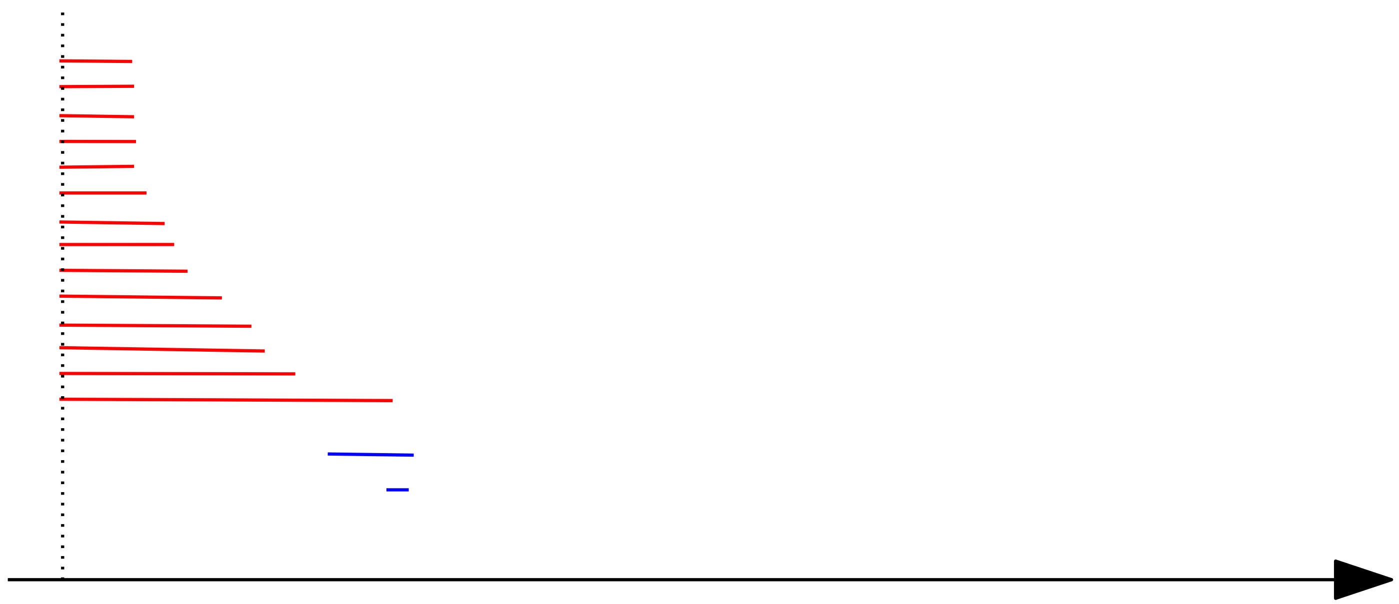
Persistent homology for point cloud data



- 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.

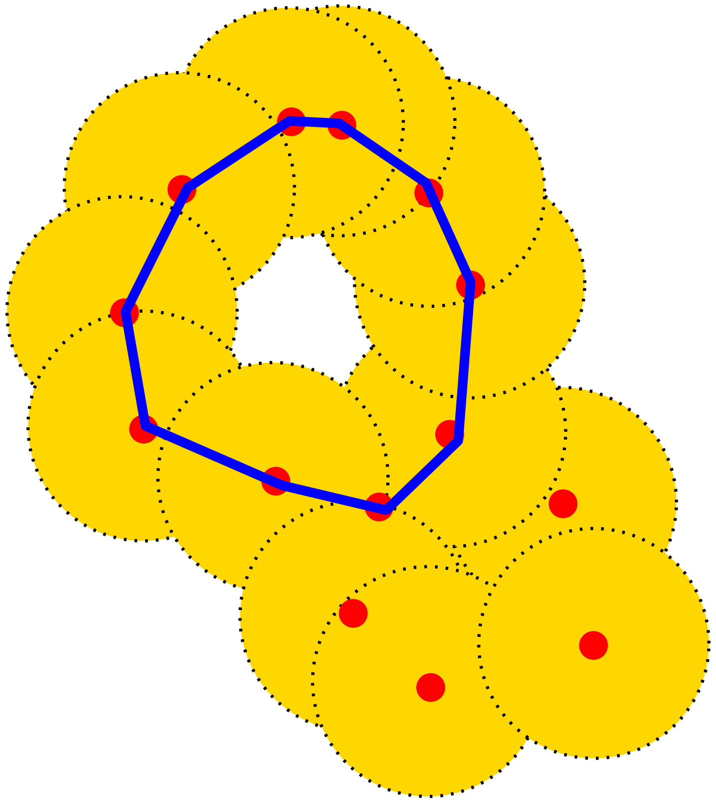
Persistent homology for point cloud data

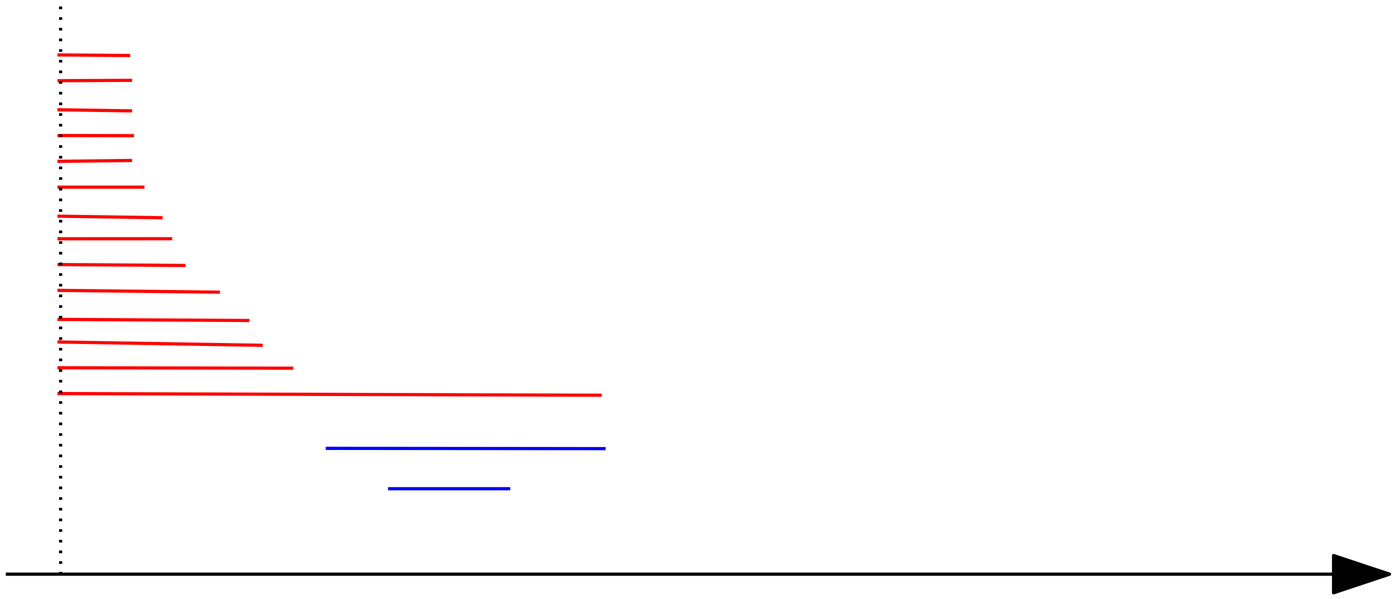




- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

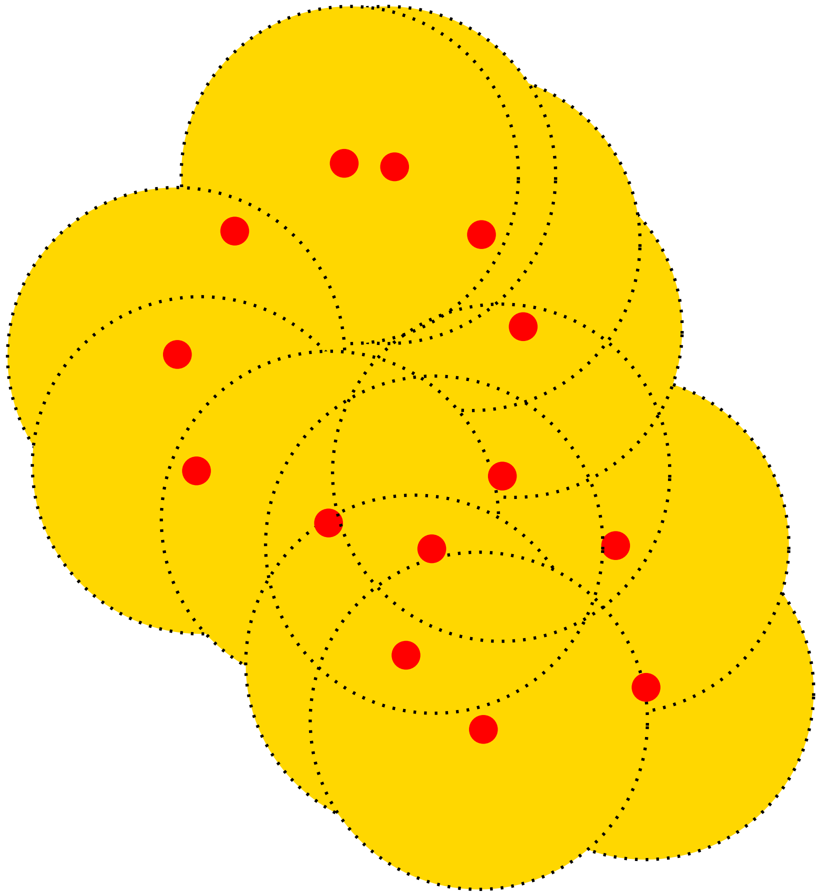
Persistent homology for point cloud data

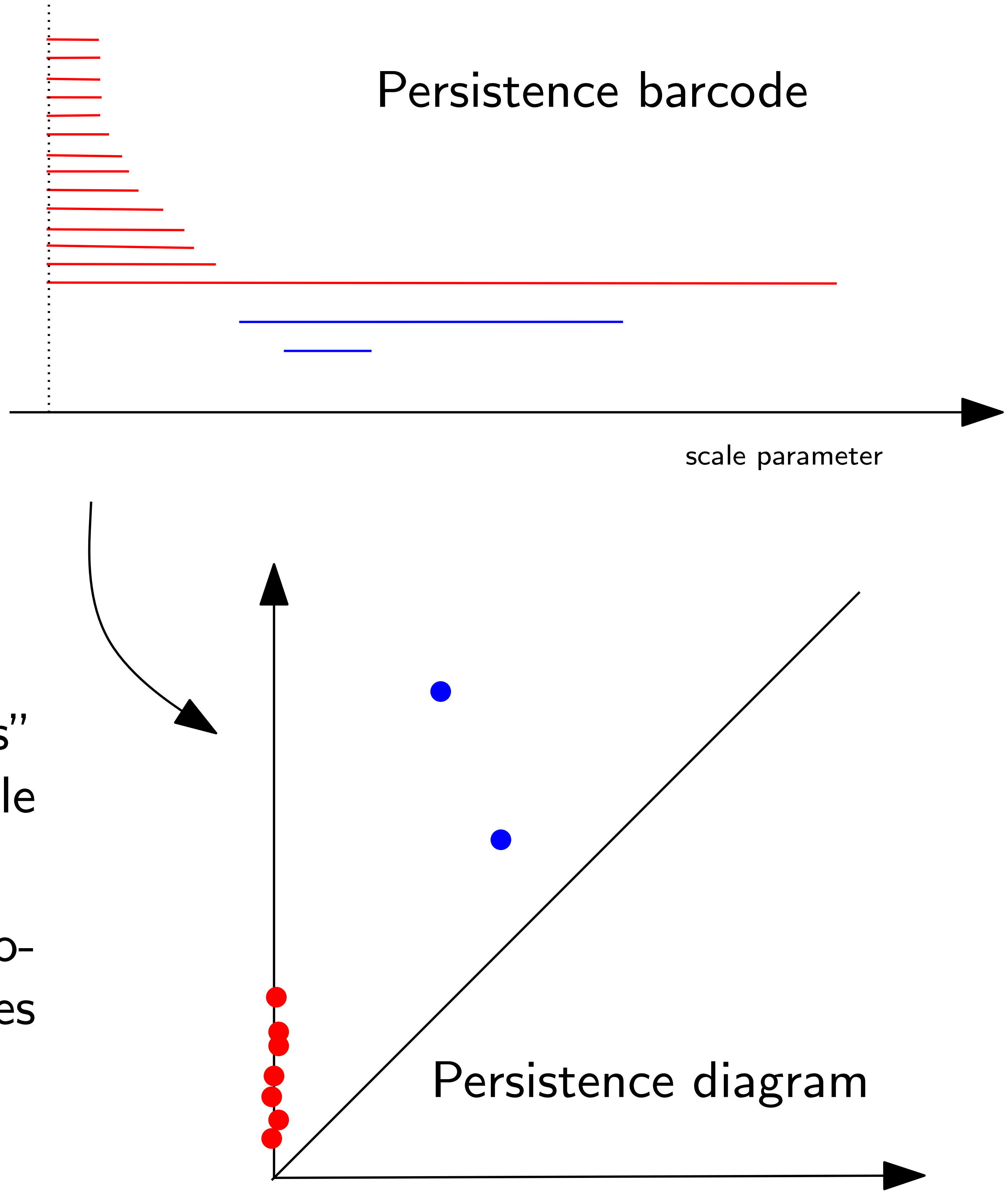




- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

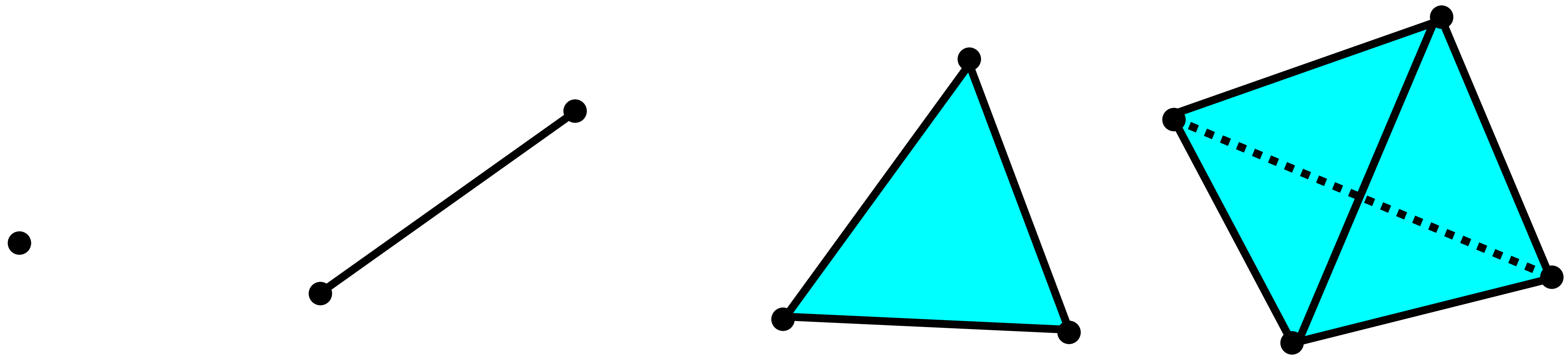
Persistent homology for point cloud data





- 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.

Simplicial complexes



0-simplex: vertex

1-simplex: edge

2-simplex: triangle

3-simplex: tetrahedron

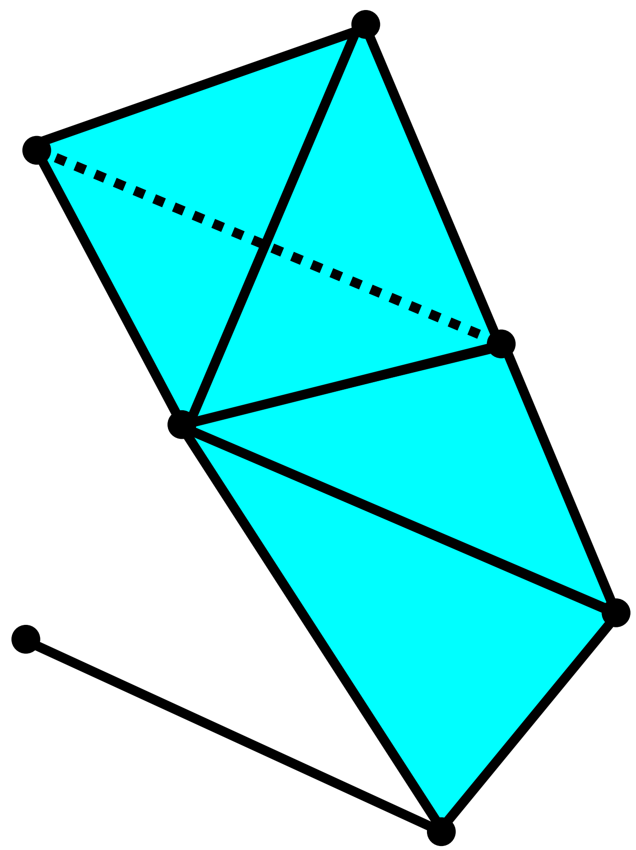
etc...

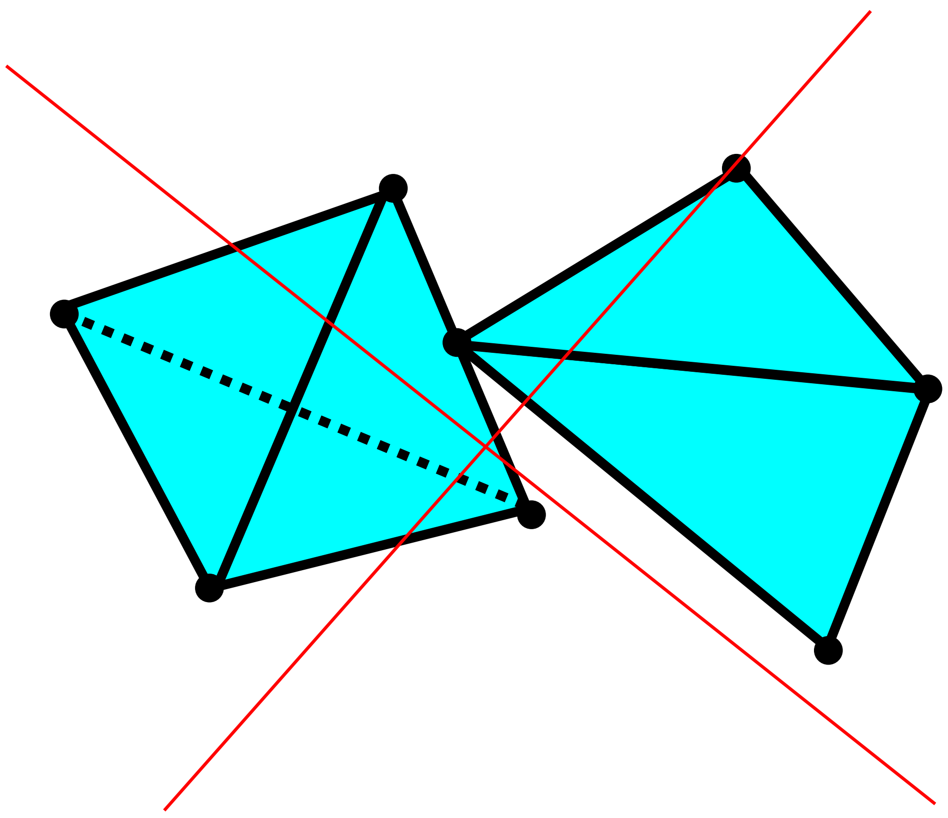
Given a set $P = \{p_0, \ldots, p_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the k-dimensional simplex σ, or k-simplex for short, spanned by P is the set of convex combinations

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

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:

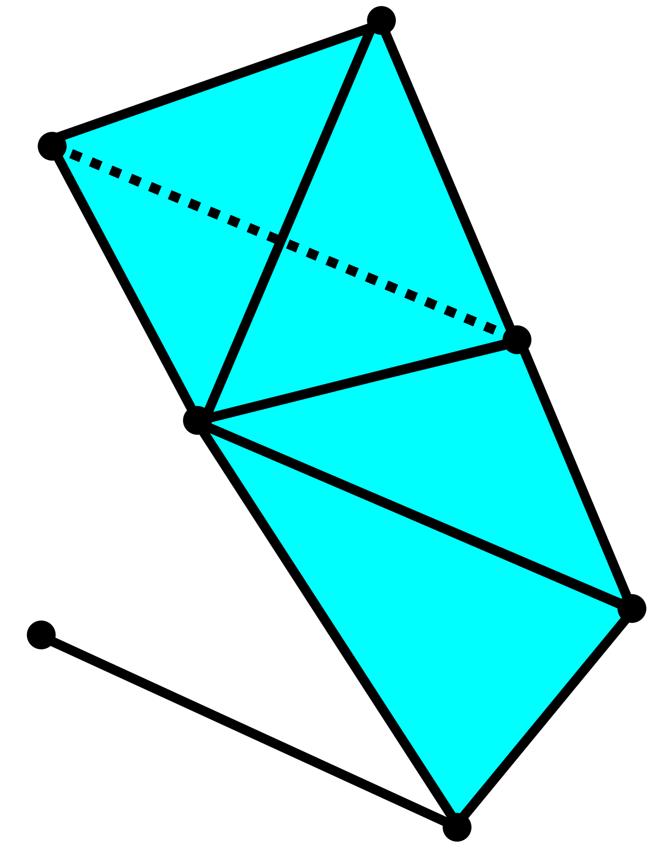
1. any face of a simplex of K is a simplex of K,
2. the intersection of any two simplices of K is either empty or a common face of both.

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

Abstract simplicial complexes

Let $P = \{p_1, \cdots p_n\}$ be a (finite) set. An abstract simplicial complex K with vertex set P is a set of 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.

Let $\{e_1, \cdots e_n\}$ a basis of $\mathbb{R}^n$. "The" geometric realization of K is the (geometric) subcomplex $|K|$ of the simplex spanned by $e_1, \cdots e_n$ such that:

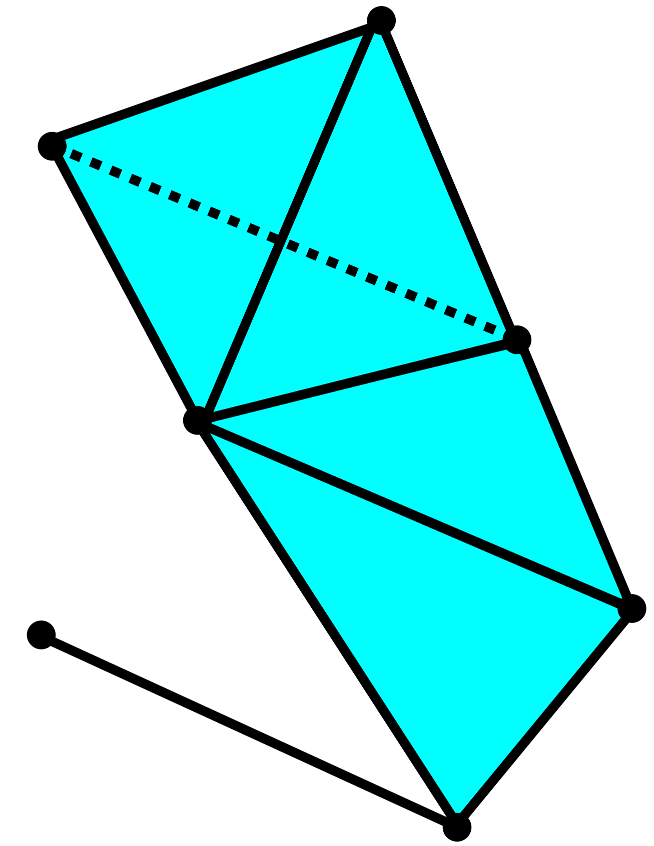
$$[e_{i_0} \cdots e_{i_k}] \in |K| \text{ iff } \{p_{i_0}, \cdots, p_{i_k}\} \in K$$

$|K|$ is a topological space (subspace of an Euclidean space)!

Abstract simplicial complexes

Let $P = \{p_1, \cdots p_n\}$ be a (finite) set. An abstract simplicial complex K with vertex set P is a set of 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 $\sigma \in K$.

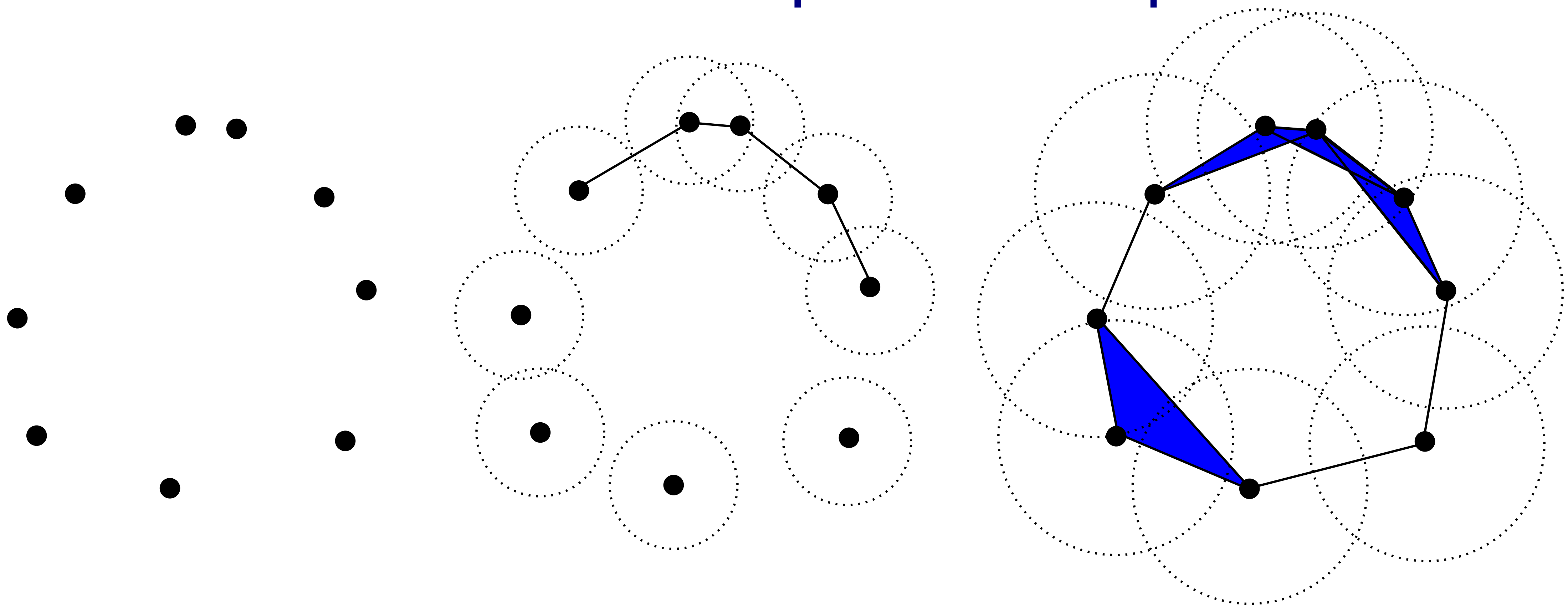


The elements of K are the simplices.

IMPORTANT

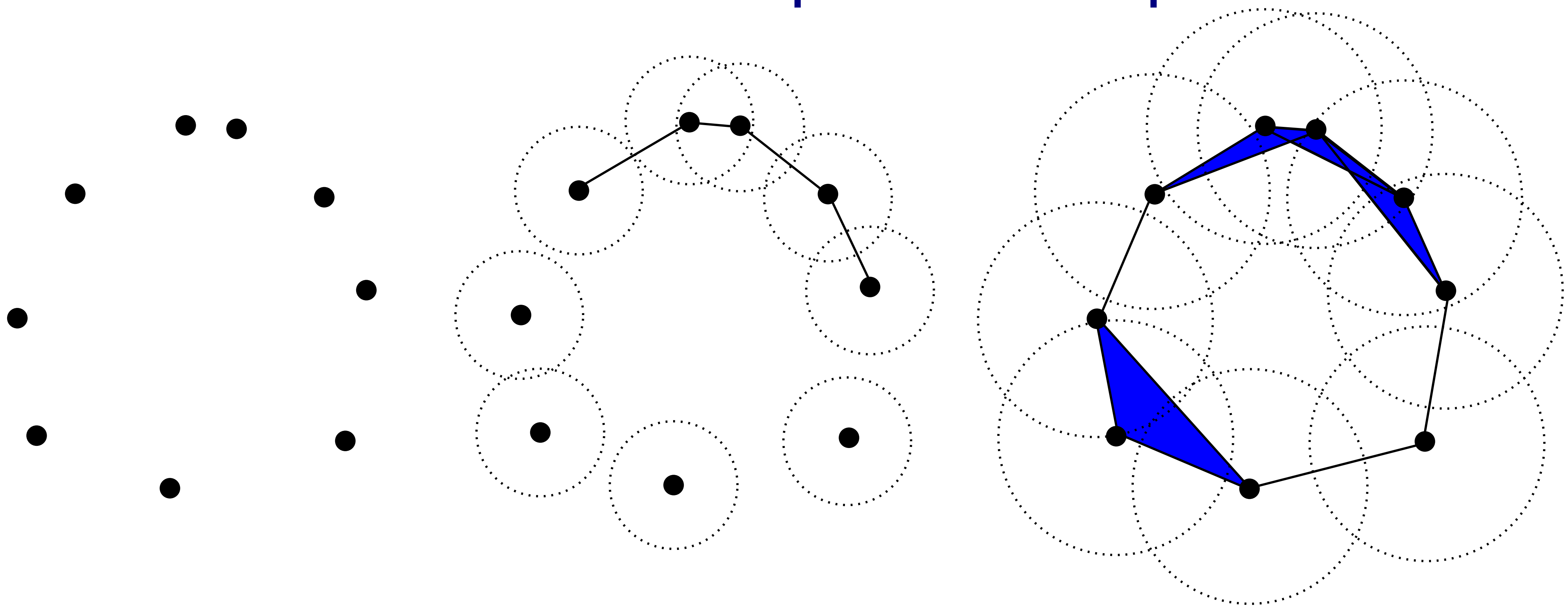
Simplicial complexes can be seen at the same time as geometric/topological spaces (good for top./geom. inference) and as combinatorial objects (abstract simplicial complexes, good for computations).

Filtrations of simplicial complexes



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

Filtrations of simplicial complexes



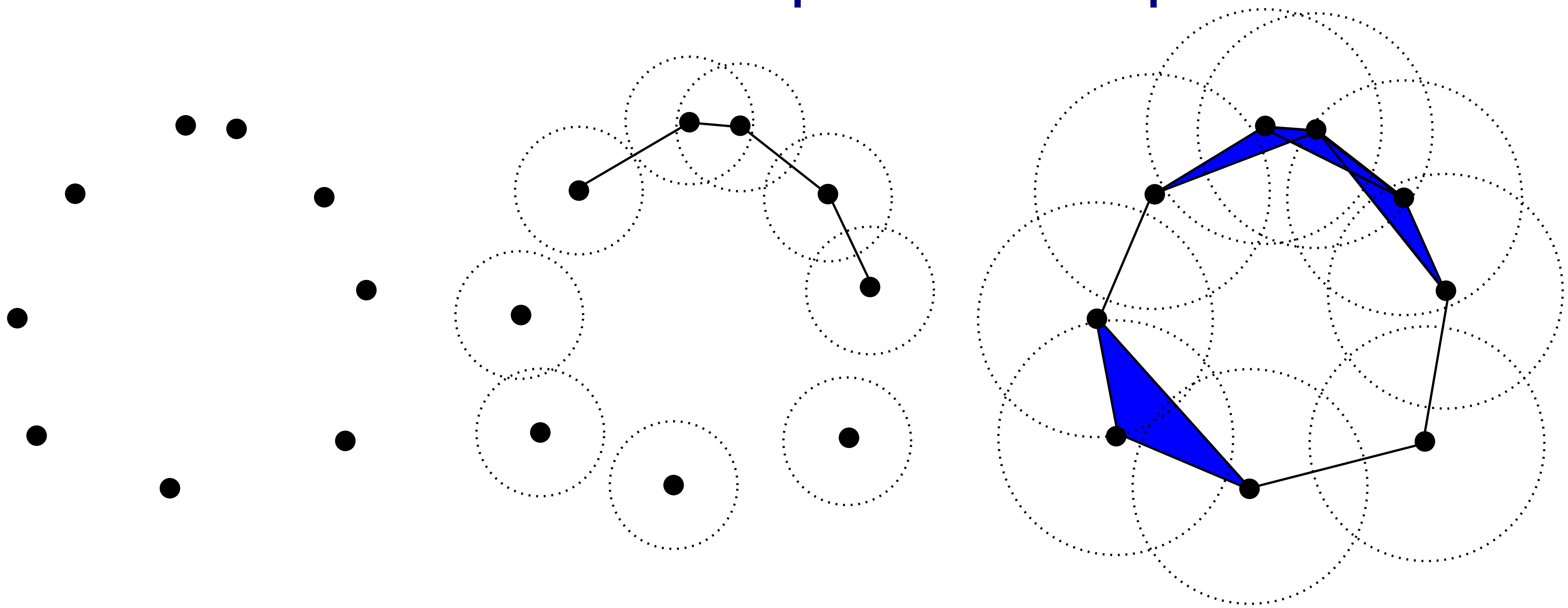
- A filtered simplicial complex (or a filtration) $\mathbb{S}$ built on top of a set $\mathbb{X}$ is a family $(\mathbb{S}_a \mid a \in \mathbf{R})$ of subcomplexes of some fixed simplicial complex $\overline{\mathbb{S}}$ with vertex set X s. t. $\mathbb{S}_a \subseteq \mathbb{S}_b$ for any $a \leq b$.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

Example: Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space.

- The Vietoris-Rips filtration is the filtered simplicial complexe defined by: for $a \in \mathbf{R}$,

$$[x_0, x_1, \cdots, x_k] \in \mathrm{Rips}(\mathbb{X}, a) \Leftrightarrow d_{\mathbb{X}}(x_i, x_j) \leq a, \quad \text{for all } i, j.$$

Filtrations of simplicial complexes



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

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

$\rightarrow$ See practical session with GUDHI

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

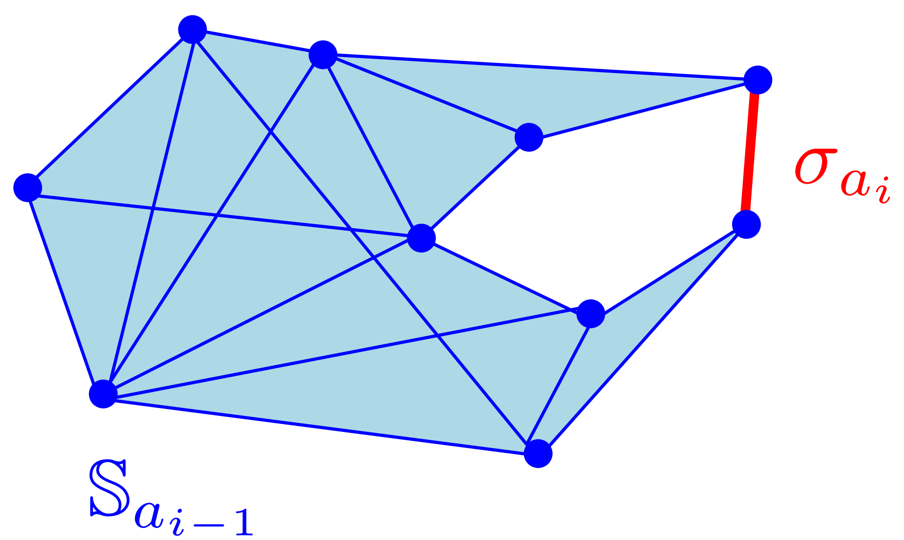
Persistent homology of filtered simplicial complexes

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.

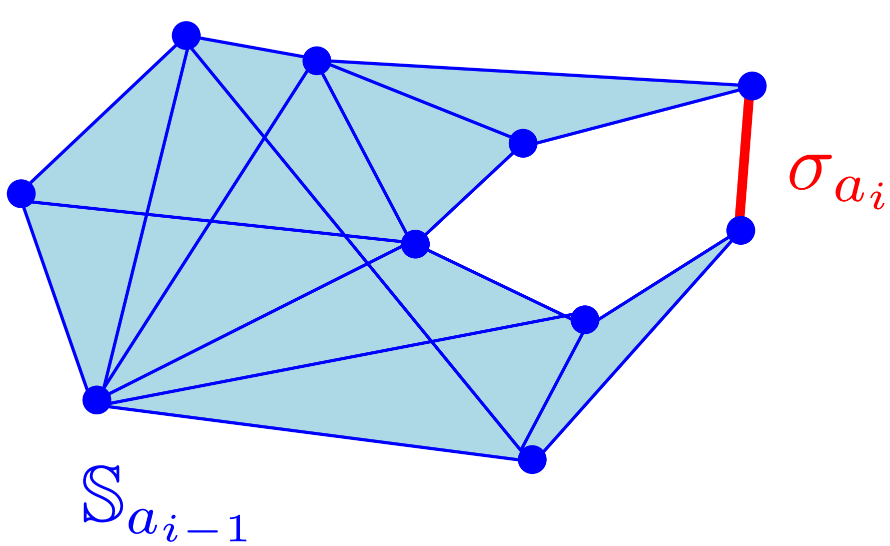
Persistent homology of filtered simplicial complexes

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:

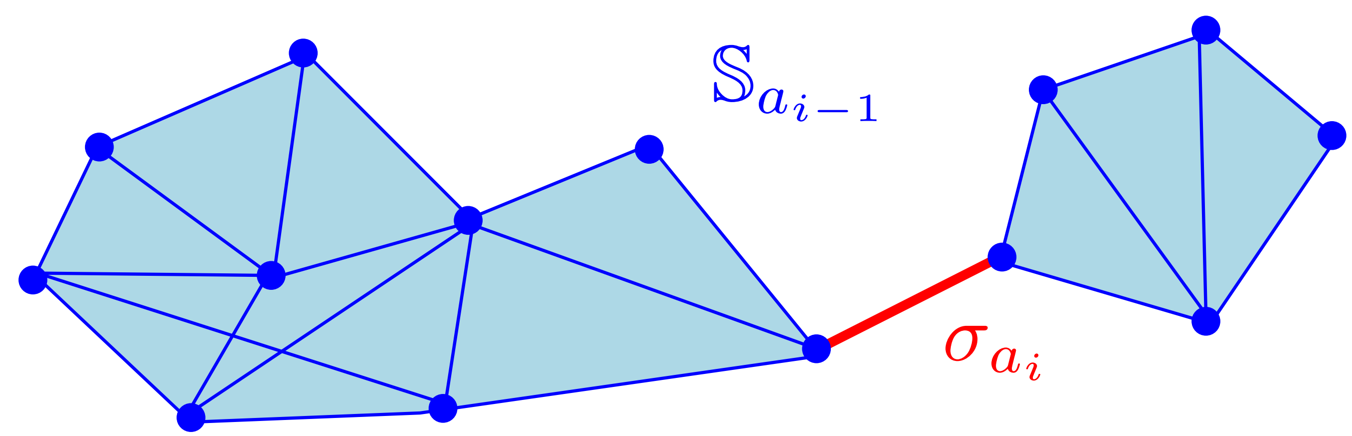
Let $k = \dim \sigma_{a_i}$ (ie. $\sigma_{a_i} = [v_0, \cdots, v_k]$)

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.

Stability properties

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

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

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

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

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).

Stability properties

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

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

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

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

Bottleneck distance

Gromov-Hausdorff distance

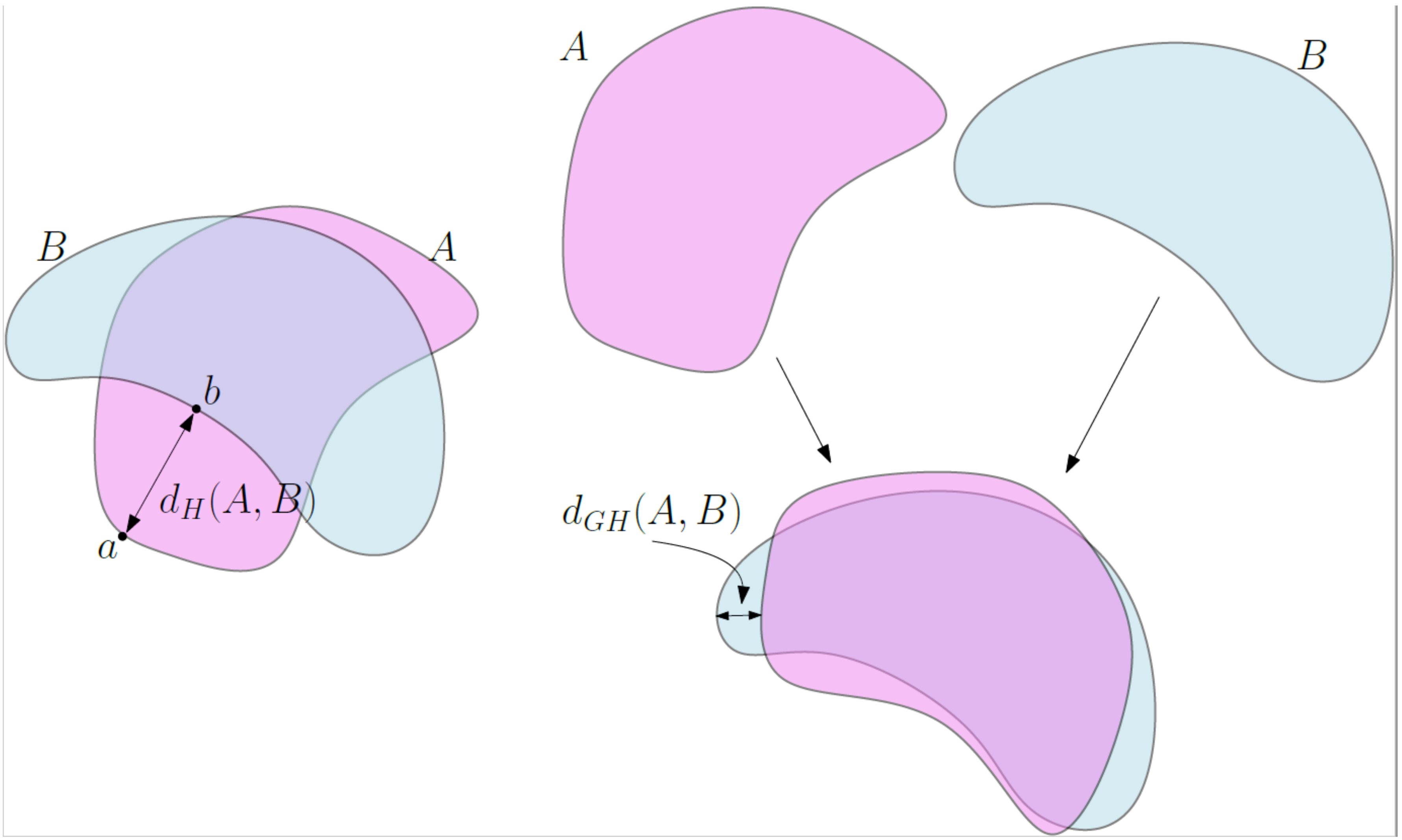
$$d_{GH}(\mathbb{X}, \mathbb{Y}) := \inf_{\mathbb{Z}, \gamma_1, \gamma_2} d_H(\gamma_1(\mathbb{X}), \gamma_2(\mathbb{X}))$$

$\mathbb{Z}$ 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).

From a statistical perspective, when $\mathbb{X}$ is a random point cloud, such result links the study of statistical properties of persistence diagrams to support estimation problems.

Hausdorff distance



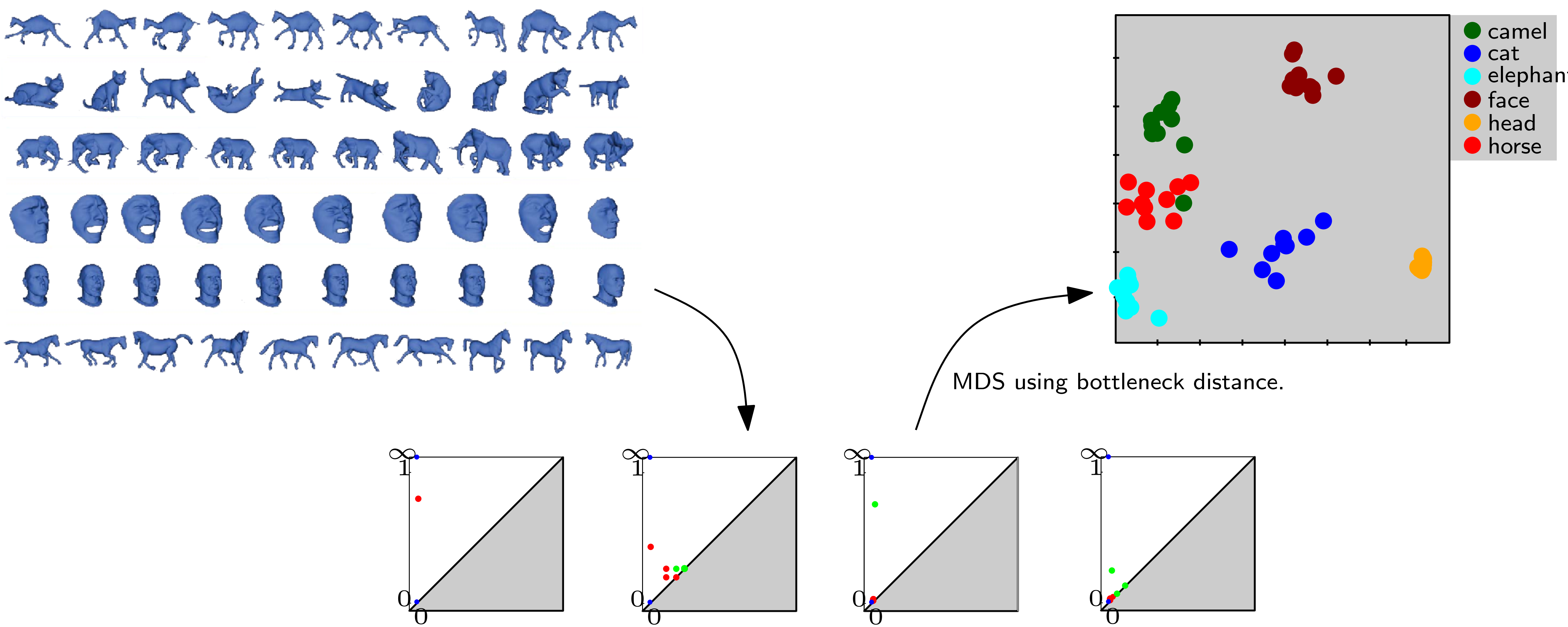
Let $A, B \subset M$ be two compact subsets of a metric space (M, d)

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

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

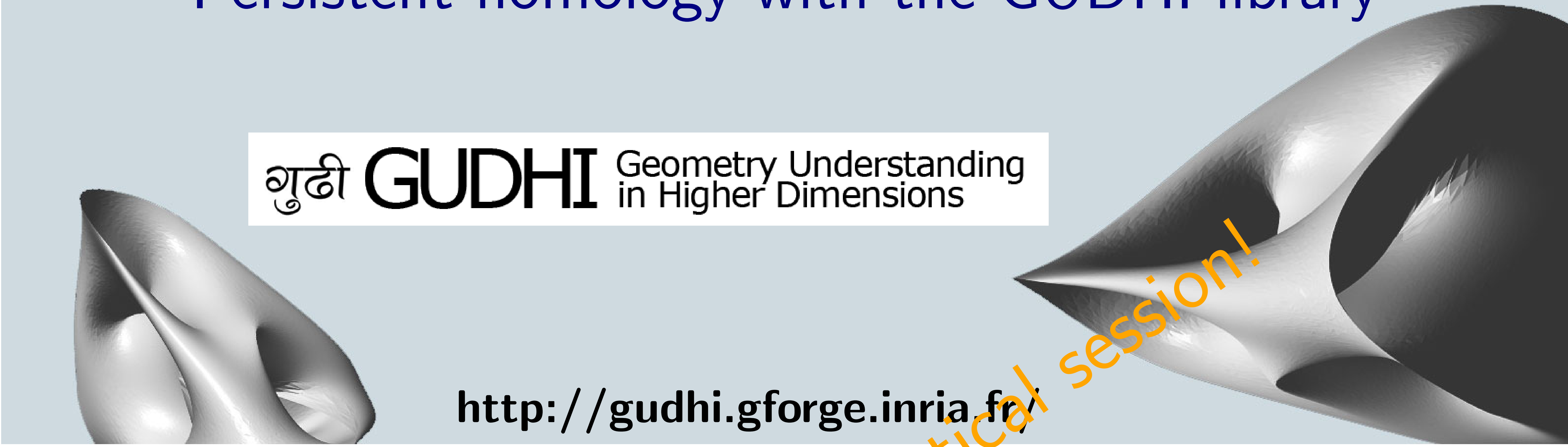
Application: non rigid shape classification

[C., Cohen-Steiner, Guibas, Mémoli, Oudot - SGP '09]



- Non rigid shapes in a same class are almost isometric, but computing Gromov-Hausdorff distance between shapes is extremely expensive.
- Compare diagrams of sampled shapes instead of shapes themselves.

Persistent homology with the GUDHI library



Advertisement for the practical session!

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.

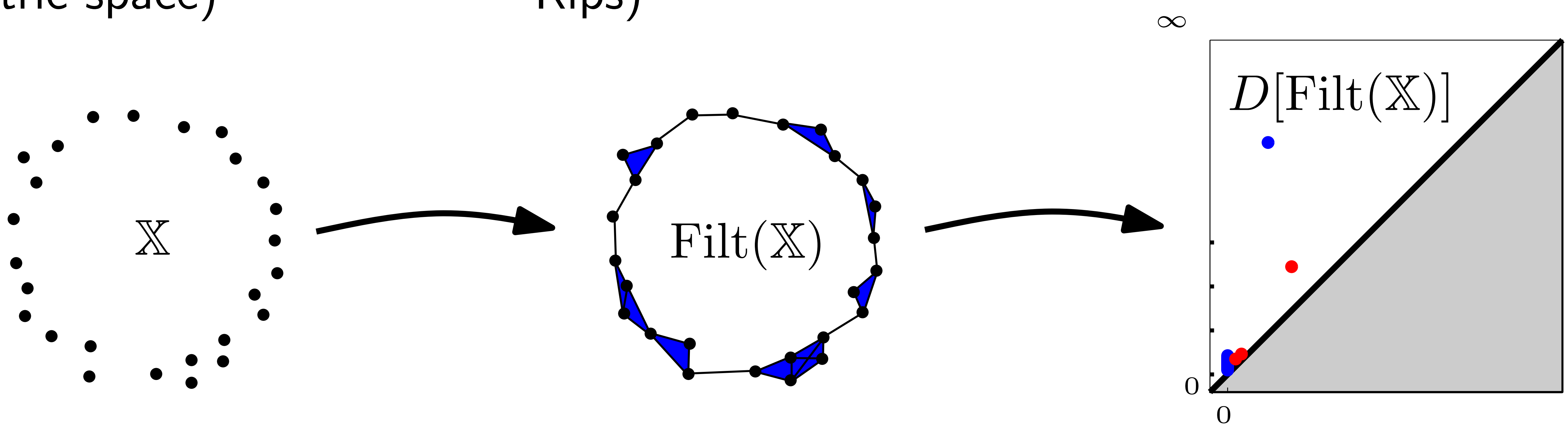
Statistical properties and features extraction from persistence diagrams

Statistical setting and "linear representations"

$\mathbb{X}$ is now a random point coud (in some metric space)

Filt is a deterministic filtration (e.g. Rips)

$D[\mathrm{Filt}(\mathbb{X})]$ becomes random

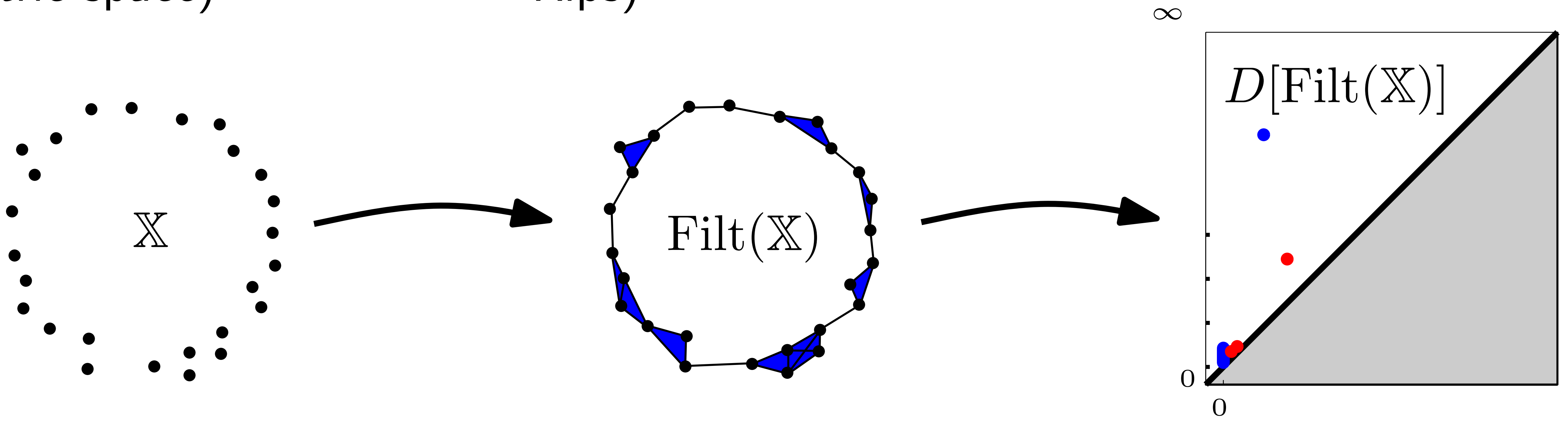


Statistical setting and "linear representations"

$\mathbb{X}$ is now a random point coud (in some metric space)

Filt is a deterministic filtration (e.g. Rips)

$D[\mathrm{Filt}(\mathbb{X})]$ becomes random



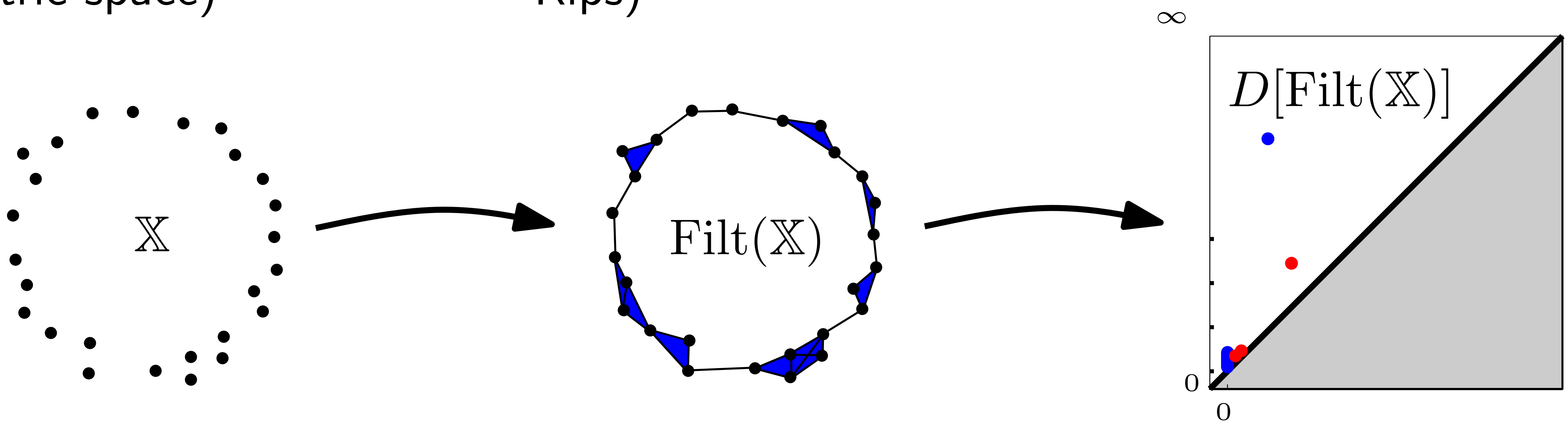
What can be said about the distribution of diagrams $D[\mathrm{Filt}(\mathbb{X})]$?

Statistical setting and "linear representations"

$\mathbb{X}$ is now a random point coud (in some metric space)

Filt is a deterministic filtration (e.g. Rips)

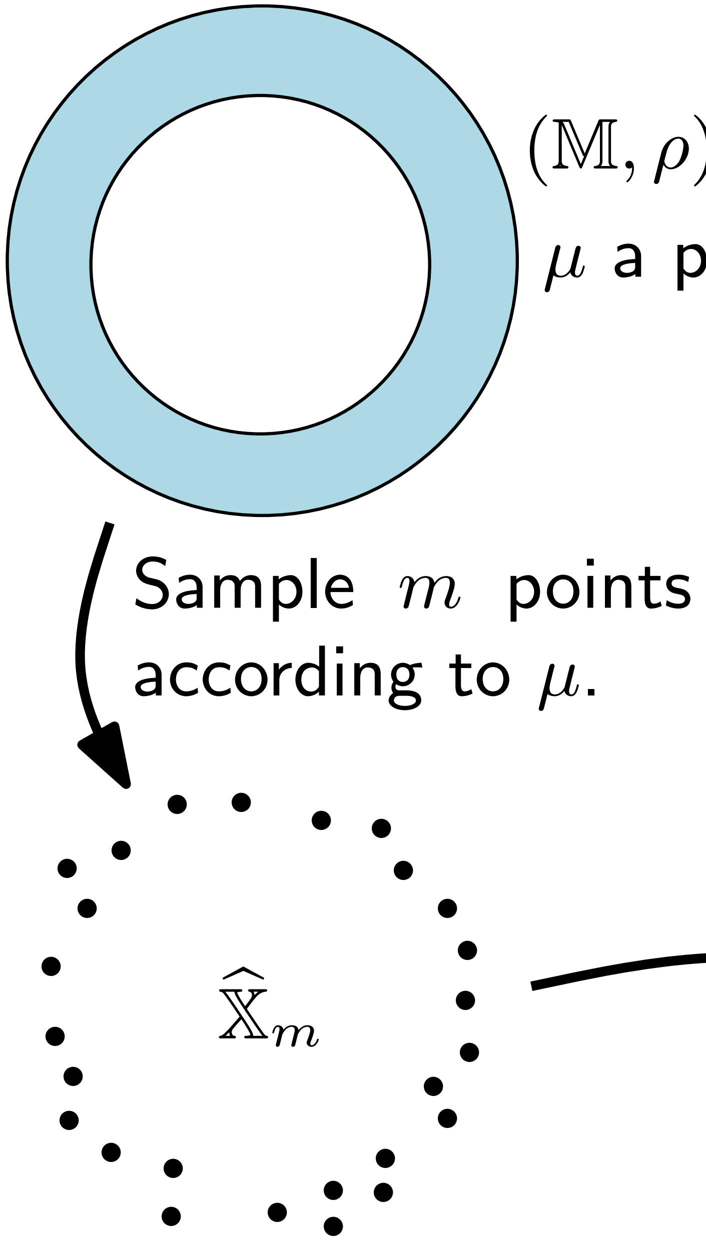
$D[\mathrm{Filt}(\mathbb{X})]$ becomes random



What can be said about the distribution of diagrams $D[\mathrm{Filt}(\mathbb{X})]$?

- Stability properties $\Rightarrow$ asymptotic properties, confidence bands, Wasserstein stability,...
- Other representation of persistence (landscapes, Betti curves, pers. images, kernels,...)

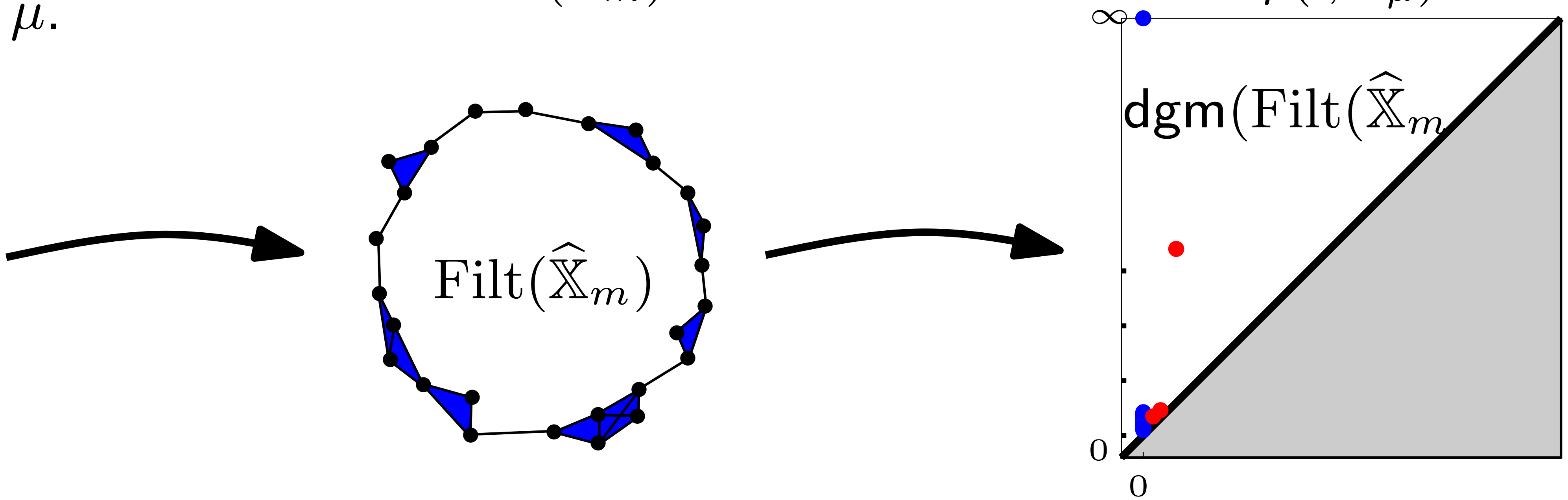
Statistical setting



$(\mathbb{M}, \rho)$ metric space

μ a probability measure with compact support $\mathbb{X}_\mu$.

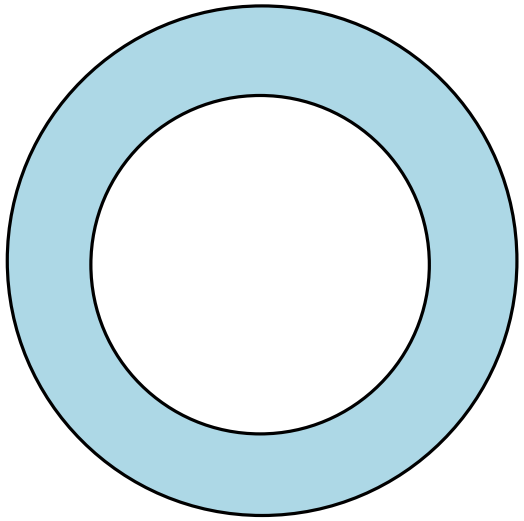
Examples:
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \mathrm{Rips}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \check{\mathrm{C}}\mathrm{ech}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) =$ sublevelset filtration of $\rho(., \mathbb{X}_\mu)$.



Questions:

- Statistical properties of dgm$(\mathrm{Filt}(\widehat{\mathbb{X}}_m))$? dgm$(\mathrm{Filt}(\widehat{\mathbb{X}}_m)) \to ?$ as $m \to +\infty$?

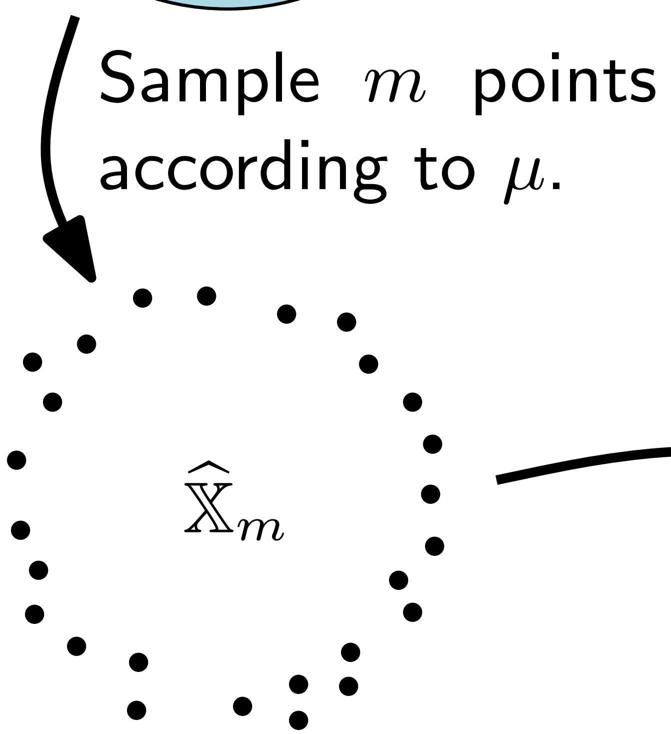
Statistical setting

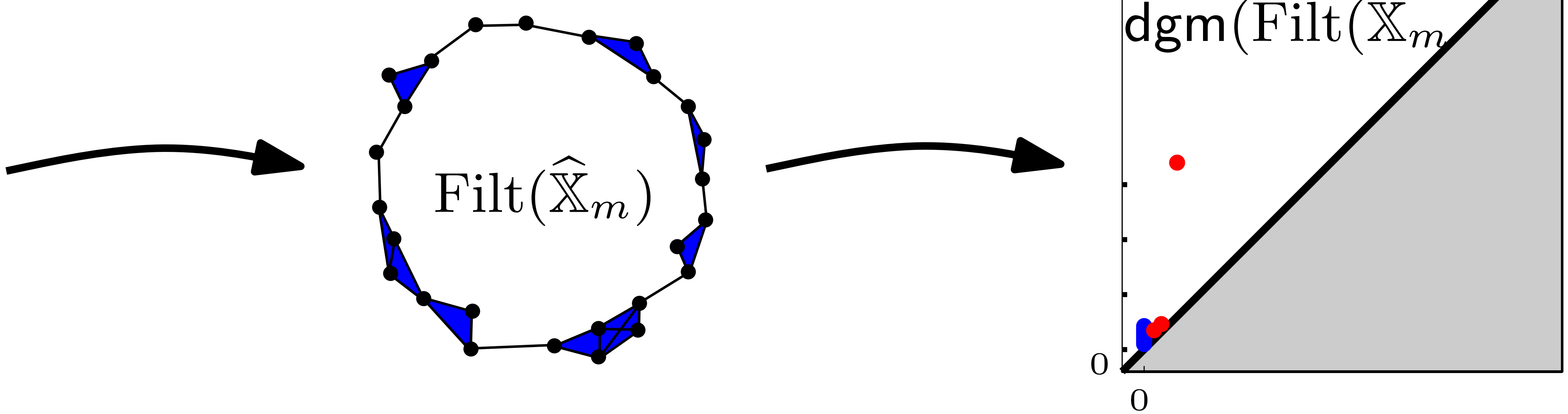


$(\mathbb{M}, \rho)$ metric space

μ a probability measure with compact support $\mathbb{X}_\mu$.

Examples:
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \mathrm{Rips}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \check{\mathrm{C}}\mathrm{ech}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) =$ sublevelset filtration of $\rho(., \mathbb{X}_\mu)$.

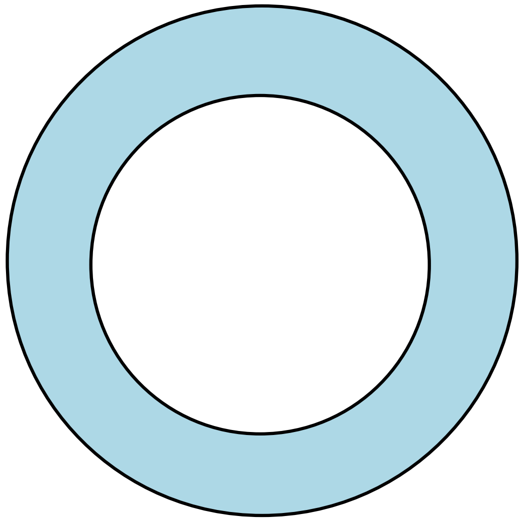




Questions:

- Statistical properties of dgm$(\mathrm{Filt}(\widehat{\mathbb{X}}_m))$? dgm$(\mathrm{Filt}(\widehat{\mathbb{X}}_m)) \to ?$ as $m \to +\infty$?
- Can we do more statistics with persistence diagrams? What can be said about distributions of diagrams?

Statistical setting

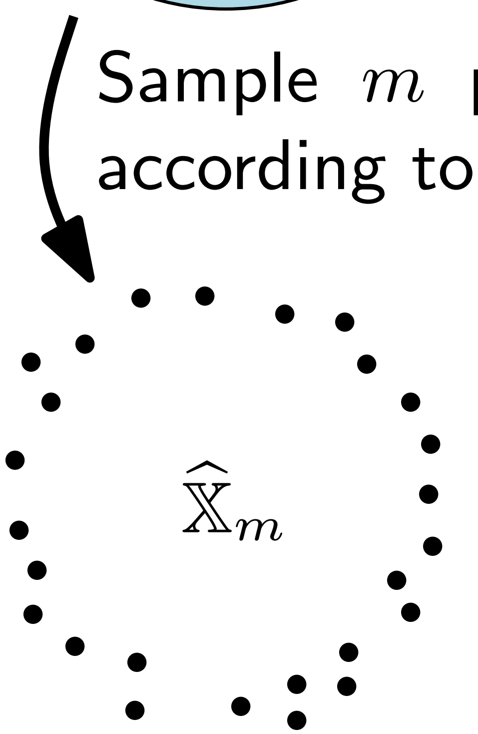


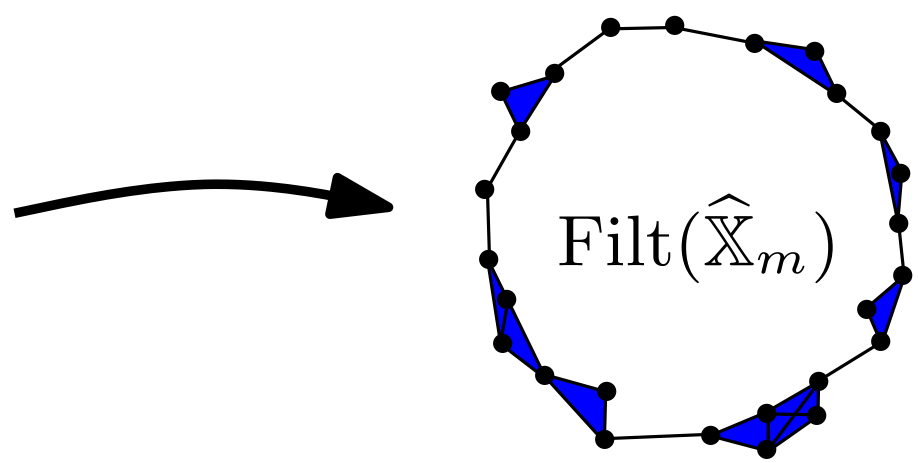
$(\mathbb{M}, \rho)$ metric space

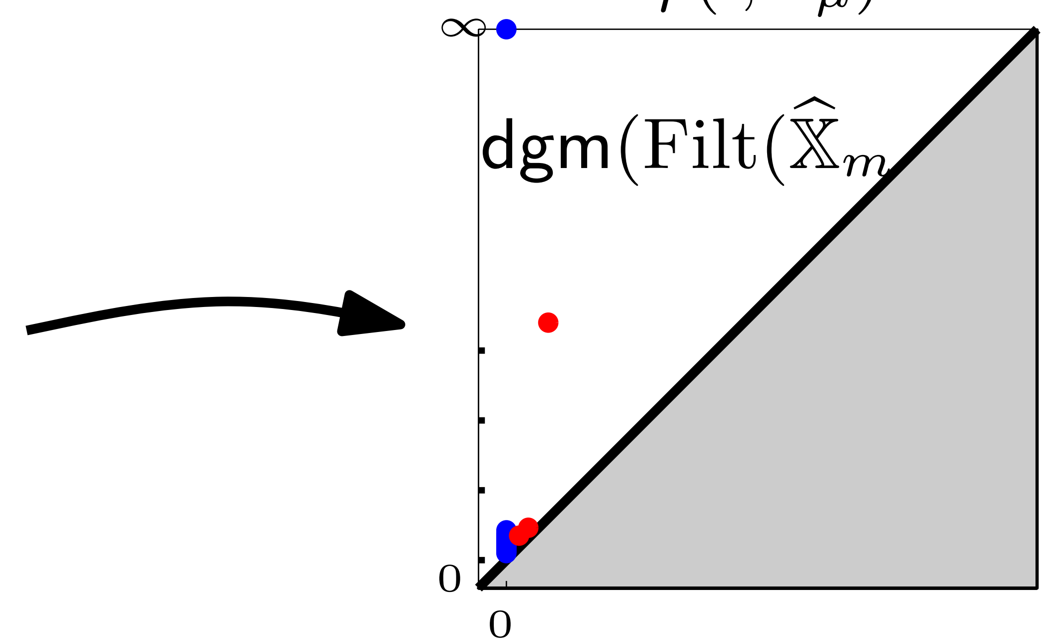
μ a probability measure with compact support $\mathbb{X}_\mu$.

Examples:
- $\text{Filt}(\widehat{\mathbb{X}}_m) = \text{Rips}_\alpha(\widehat{\mathbb{X}}_m)$
- $\text{Filt}(\widehat{\mathbb{X}}_m) = \check{\text{C}}\text{ech}_\alpha(\widehat{\mathbb{X}}_m)$
- $\text{Filt}(\widehat{\mathbb{X}}_m)$ = sublevelset filtration of $\rho(., \mathbb{X}_\mu)$.

Sample m points according to μ.







Stability thm: $\mathrm{d_b}(\text{dgm}(\text{Filt}(\mathbb{X}_\mu)), \text{dgm}(\text{Filt}(\widehat{\mathbb{X}}_m))) \leq 2 d_{GH}(\mathbb{X}_\mu, \widehat{\mathbb{X}}_m)$

So, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\mathrm{d_b}\left(\text{dgm}(\text{Filt}(\mathbb{X}_\mu)), \text{dgm}(\text{Filt}(\widehat{\mathbb{X}}_m))\right) > \varepsilon\right) \leq \mathbb{P}\left(d_{GH}(\mathbb{X}_\mu, \widehat{\mathbb{X}}_m) > \frac{\varepsilon}{2}\right)$$

Deviation inequality and rate of convergence

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]

For $a, b > 0$, μ satisfies the (a,b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x,r)) \geq \min(ar^b, 1)$.

Deviation inequality and rate of convergence

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]

For $a, b > 0$, μ satisfies the (a,b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x,r)) \geq \min(ar^b, 1)$.

Theorem: If μ satisfies the (a,b)-standard assumption, then for any $\varepsilon > 0$:

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))\right) > \varepsilon\right) \leq \min(\frac{8^b}{a\varepsilon^b}\exp(-ma\varepsilon^b), 1).$$

Deviation inequality and rate of convergence

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]

For $a, b > 0$, μ satisfies the (a,b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x,r)) \geq \min(ar^b, 1)$.

Theorem: If μ satisfies the (a,b)-standard assumption, then for any $\varepsilon > 0$:

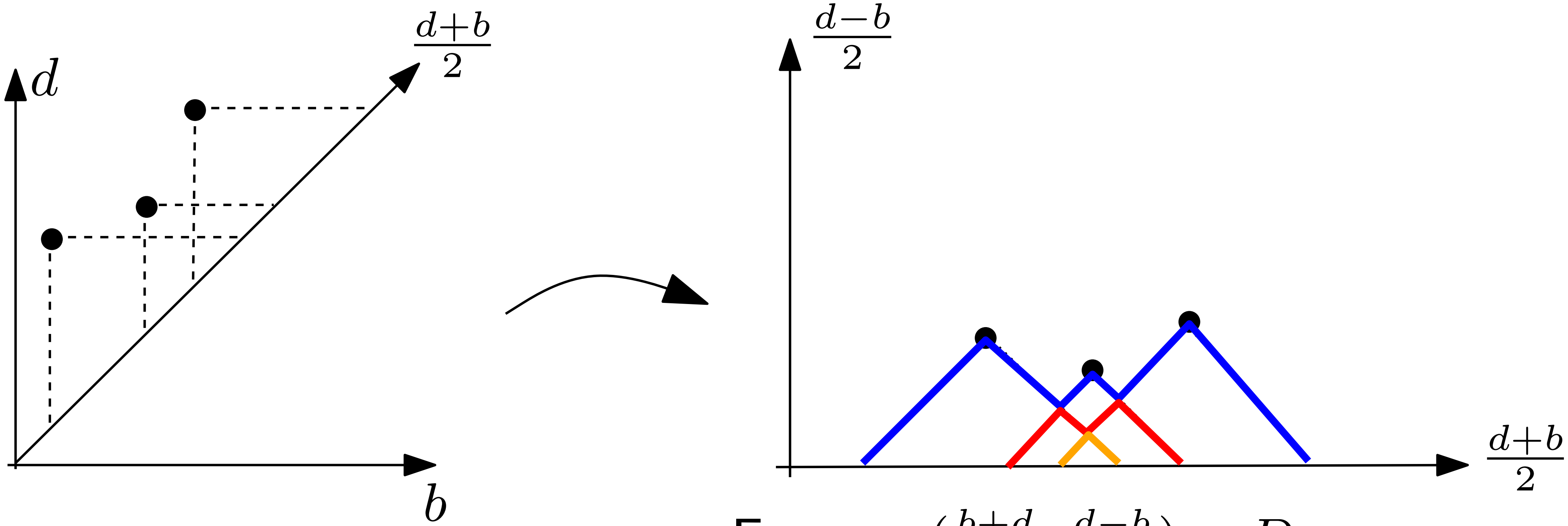
$$\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))\right) > \varepsilon\right) \leq \min(\frac{8^b}{a\varepsilon^b}\exp(-ma\varepsilon^b), 1).$$

Corollary: Let $\mathcal{P}(a, b, \mathbb{M})$ be the set of (a,b)-standard proba measures on $\mathbb{M}$. Then:

$$\sup_{\mu \in \mathcal{P}(a,b,\mathbb{M})} \mathbb{E}\left[\mathrm{d_b}(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m)))\right] \leq C\left(\frac{\ln m}{m}\right)^{1/b}$$

where the constant C only depends on a and b (not on $\mathbb{M}$!). Moreover, the upper bound is tight (in a minimax sense)!

Persistence landscapes



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

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

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

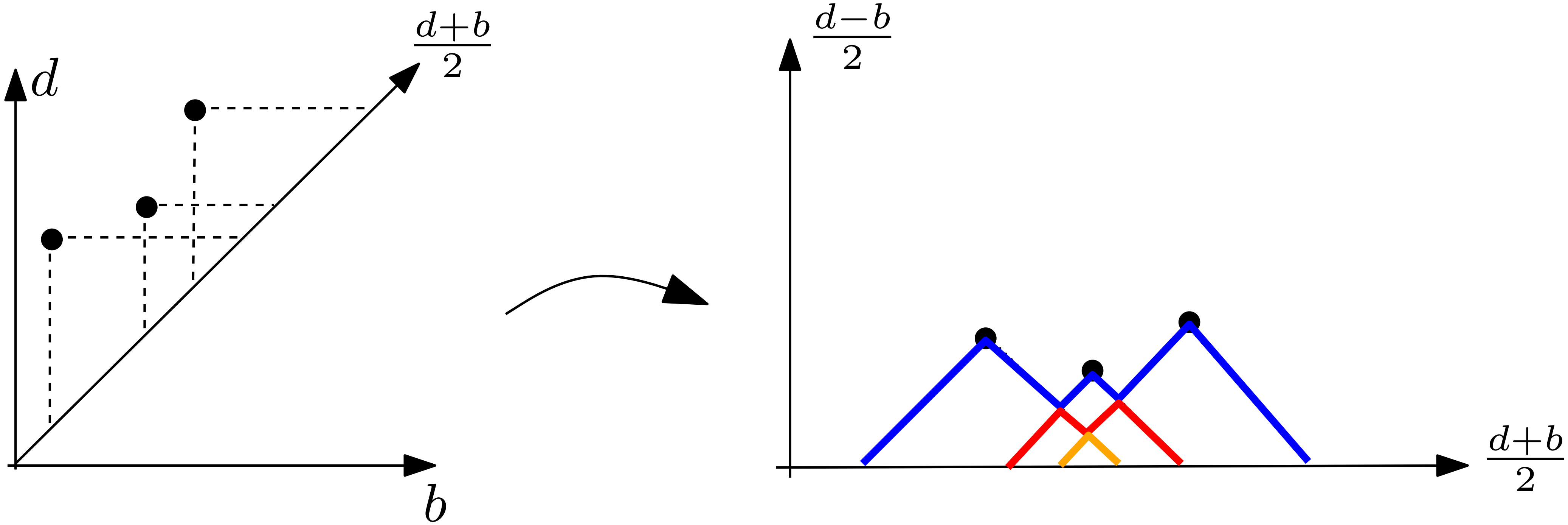
Persistence landscape [Bubenik 2012]:

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

where kmax is the kth largest value in the set.

Many other ways to “linearize” persistence diagrams: intensity functions, image persistence, Betti curves, kernels,...

Persistence landscapes



Persistence landscape [Bubenik 2012]:

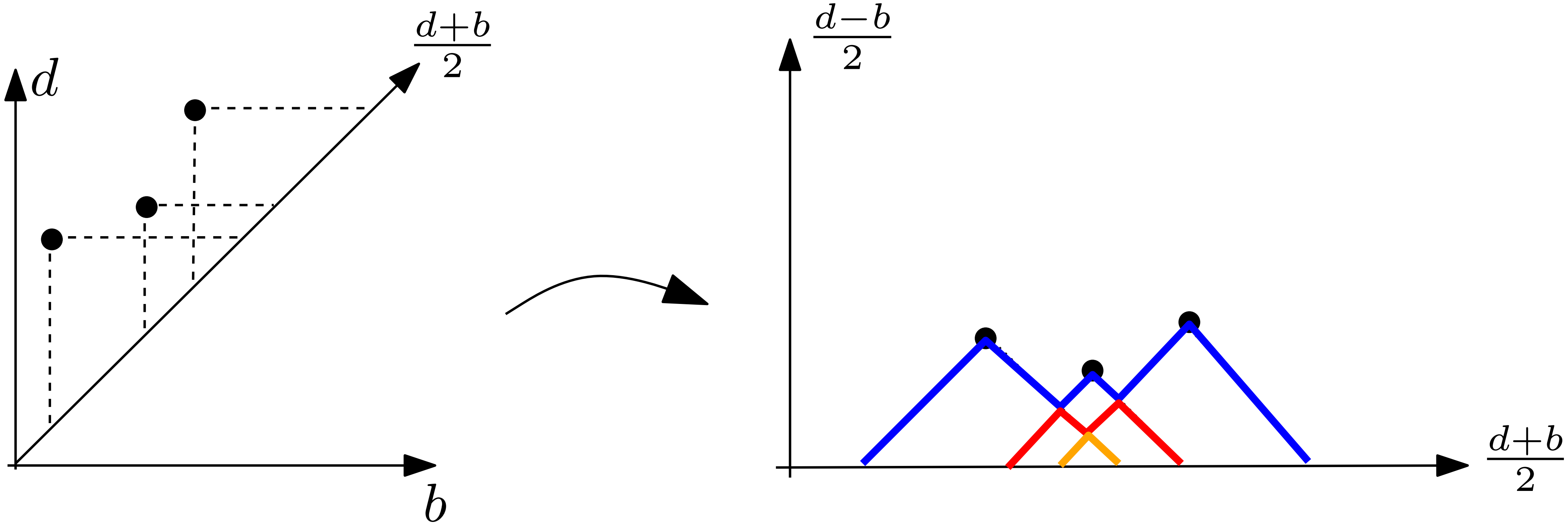
$$\lambda_D(k,t) = \underset{p \in \text{dgm}}{\text{kmax}}\ \Lambda_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N},$$

Properties

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

stability properties of persistence landscapes

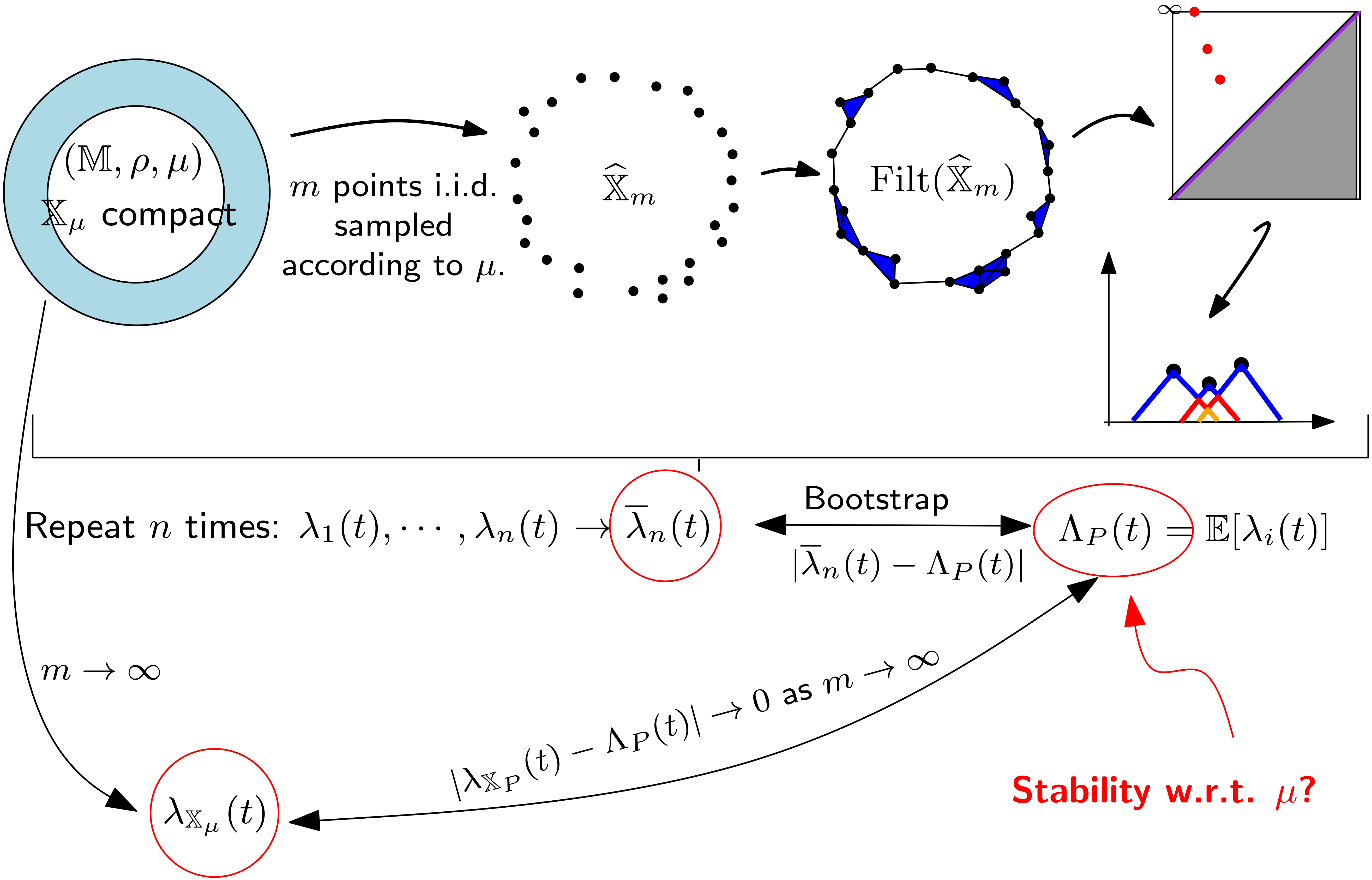
Persistence landscapes



- 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]

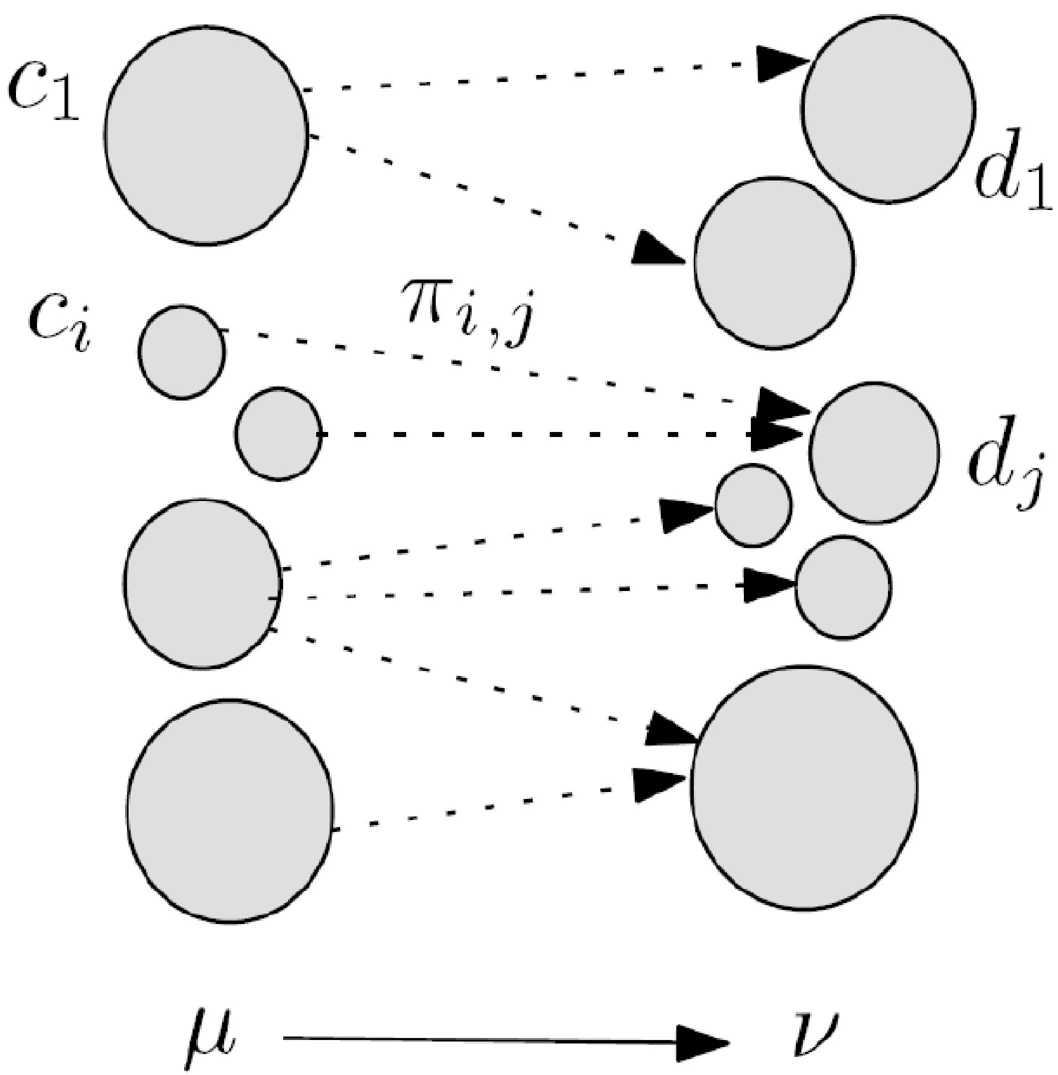
To summarize



Wasserstein distance

Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be probability measures on $\mathbb{M}$ with finite p-moments ($p \geq 1$).

"The" Wasserstein distance $W_p(\mu, \nu)$ quantifies the optimal cost of pushing μ onto ν, the cost of moving a small mass dx from x to y being $\rho(x, y)^p dx$.



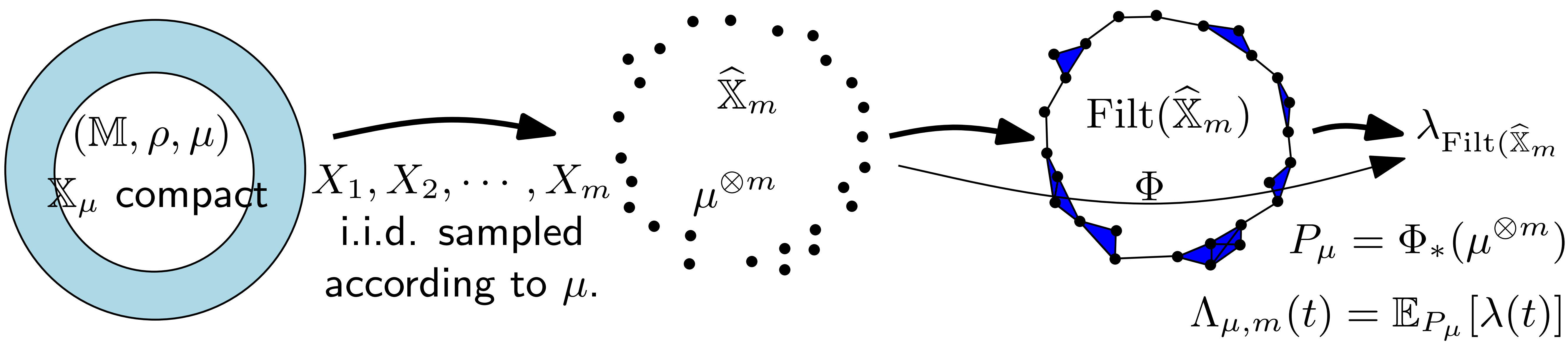
- Transport plan: Π a proba measure on $M \times M$ such that $\Pi(A \times \mathbb{R}^d) = \mu(A)$ and $\Pi(\mathbb{R}^d \times B) = \nu(B)$ for any borelian sets $A, B \subset M$.
- Cost of a transport plan:

$$C(\Pi) = \left(\int_{M \times M} \rho(x, y)^p d\Pi(x, y) \right)^{\frac{1}{p}}$$

- $W_p(\mu, \nu) = \inf_{\Pi} C(\Pi)$

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]



Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be proba measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

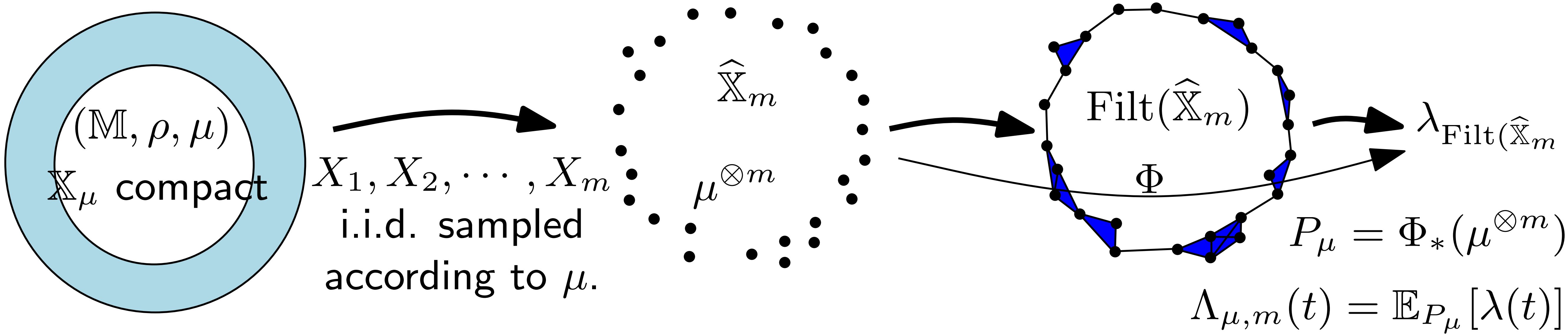
where W_p denotes the Wasserstein distance with cost function $\rho(x,y)^p$.

Remarks:

- similar results by Blumberg et al (2014) in the (Gromov-)Prokhorov metric (for distributions, not for expectations) ;
- Extended to point process setting y L. Decreusefond et al;
- $m^{\frac{1}{p}}$ cannot be replaced by a constant.

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]



Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be proba measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

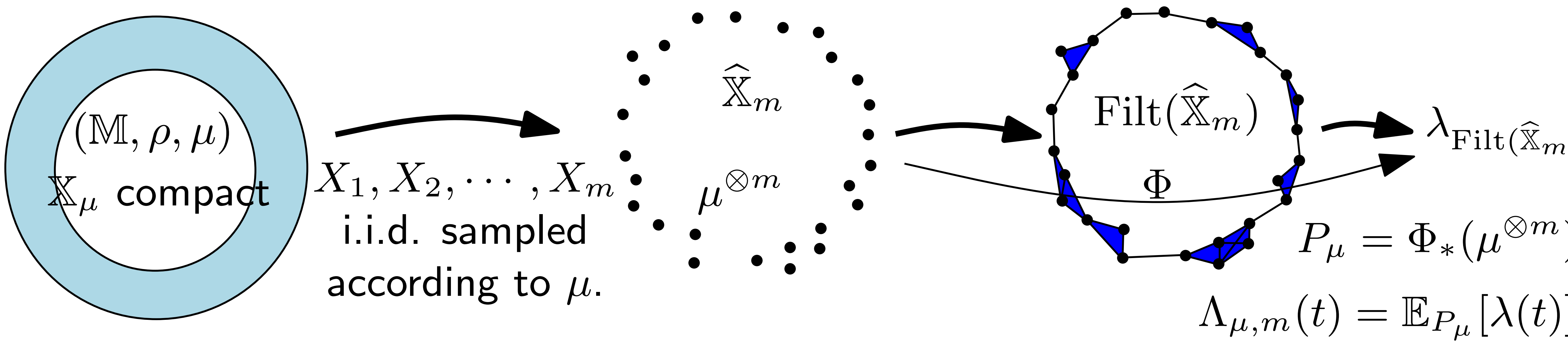
where W_p denotes the Wasserstein distance with cost function $\rho(x,y)^p$.

Consequences:

- Subsampling: efficient and easy to parallelize algorithm to infer topol. information from huge data sets.
- Robustness to outliers.
- R package TDA +Gudhi library: https://project.inria.fr/gudhi/software/

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]



Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be proba measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

where W_p denotes the Wasserstein distance with cost function $\rho(x,y)^p$.

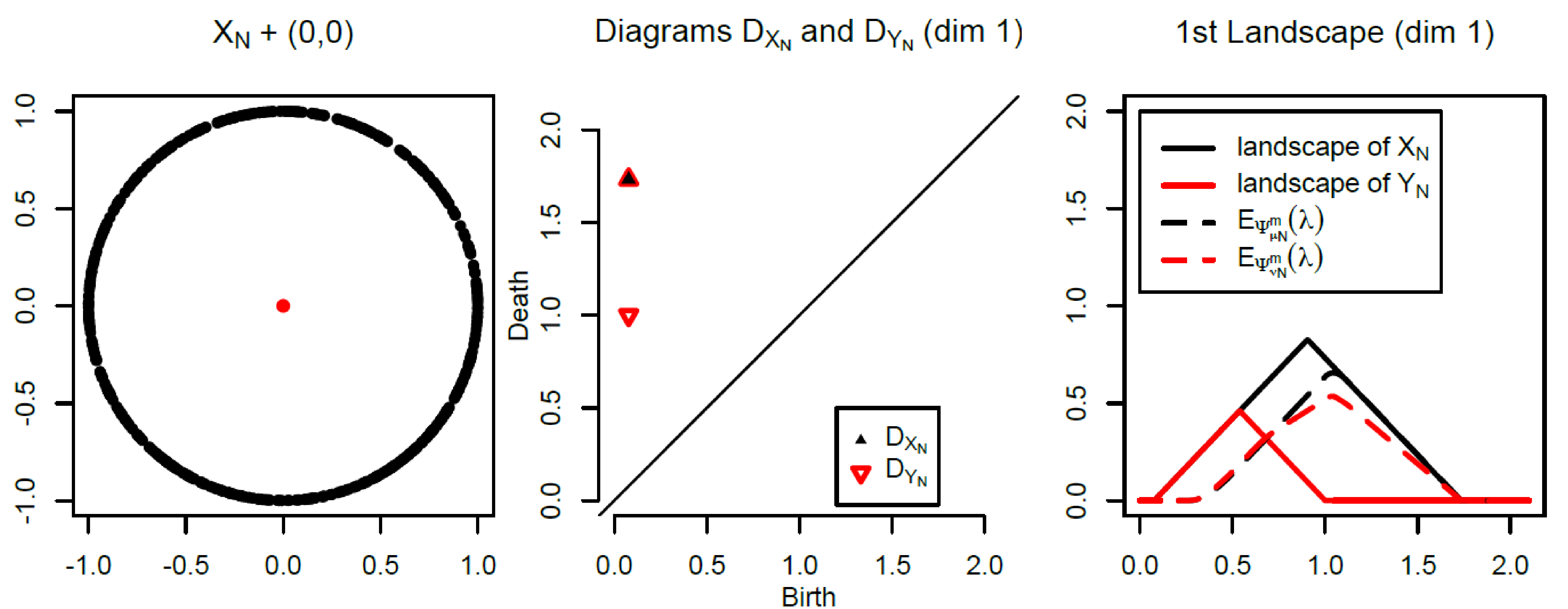
Proof:

1. $W_p(\mu^{\otimes m}, \nu^{\otimes m}) \leq m^{\frac{1}{p}} W_p(\mu, \nu)$
2. $W_p(P_\mu, P_\nu) \leq W_p(\mu^{\otimes m}, \nu^{\otimes m})$ (stability of persistence!)
3. $\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq W_p(P_\mu, P_\nu)$ (Jensen's inequality)

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]

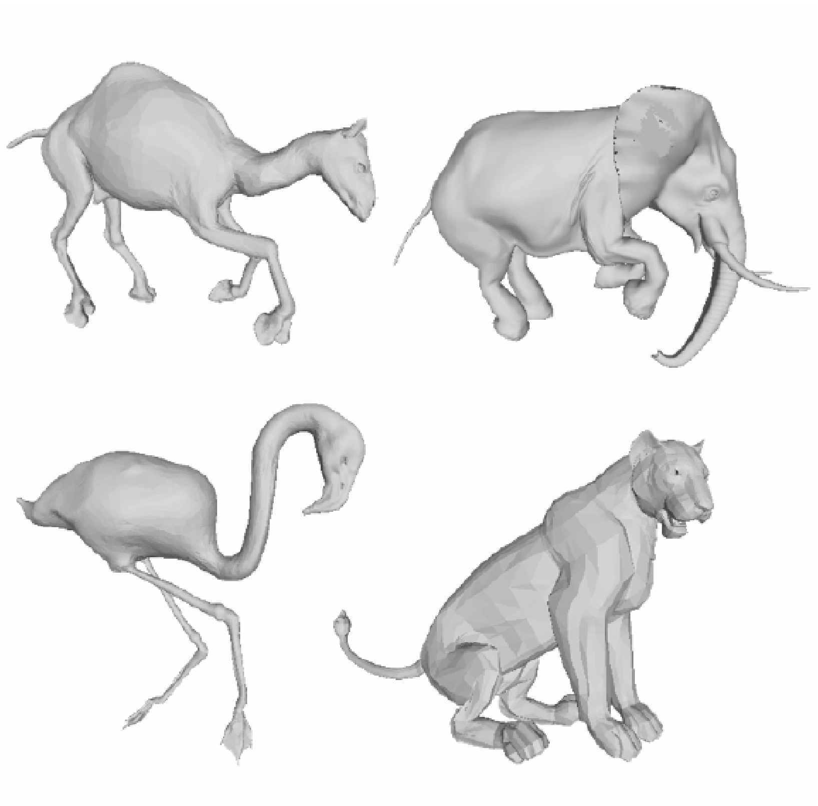
Example: Circle with one outlier.

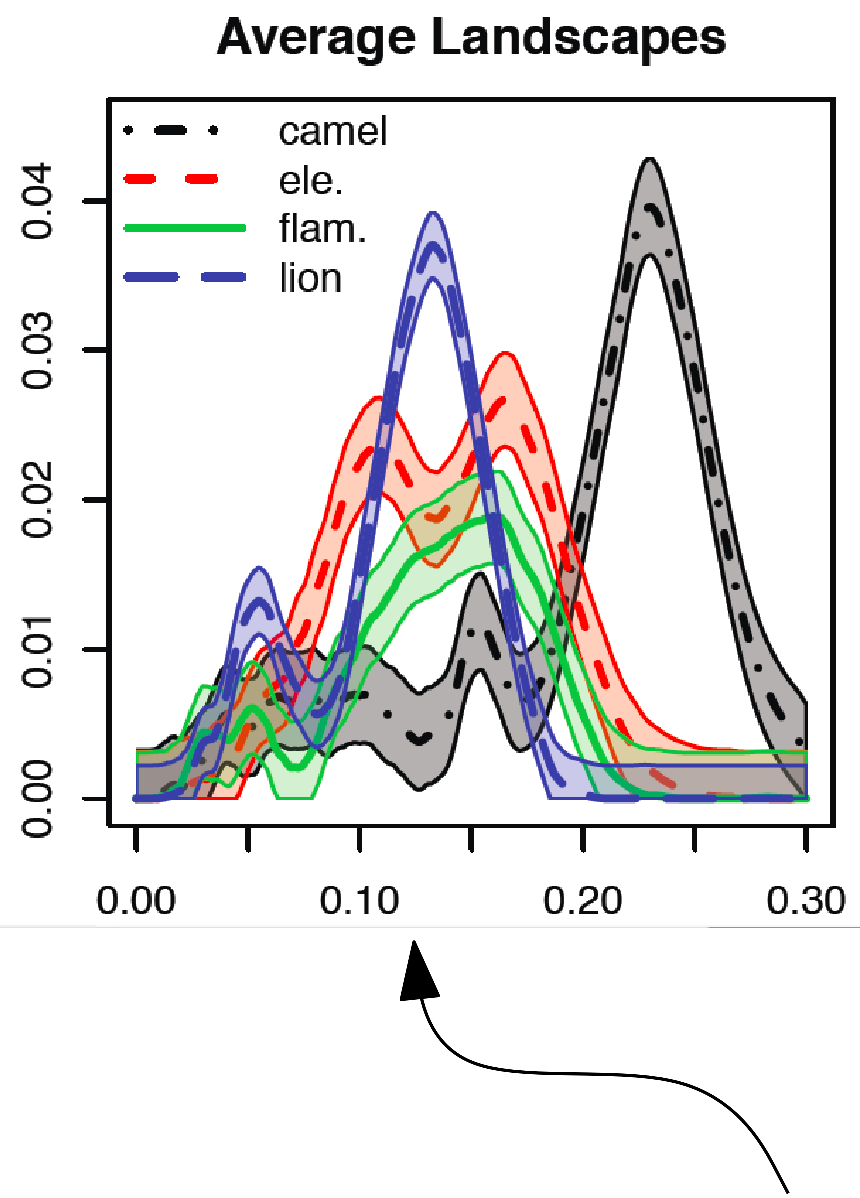


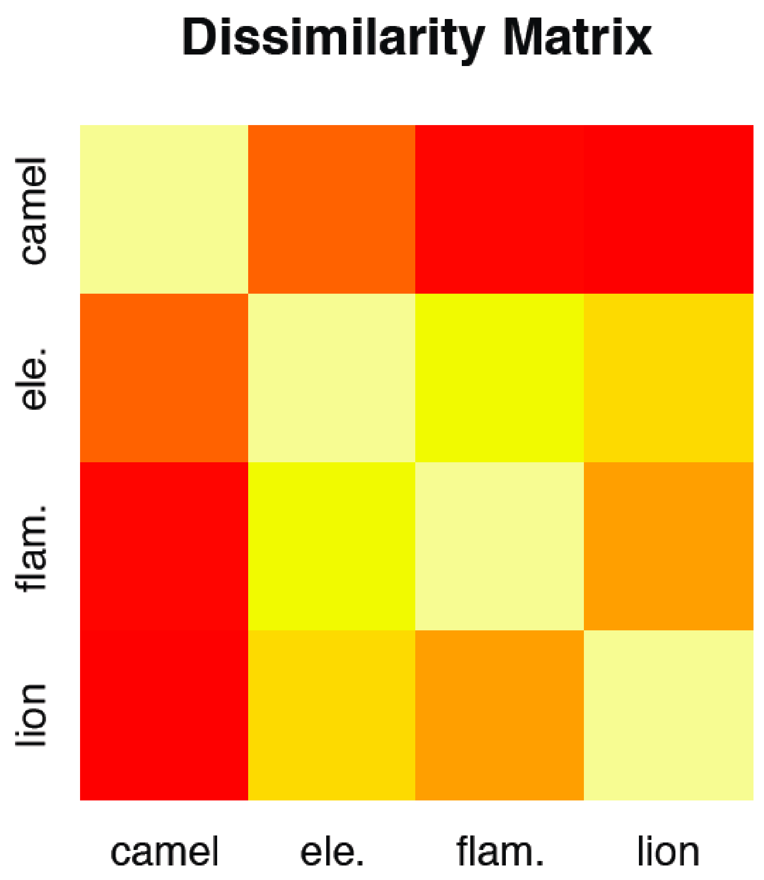
(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]

Example: 3D shapes





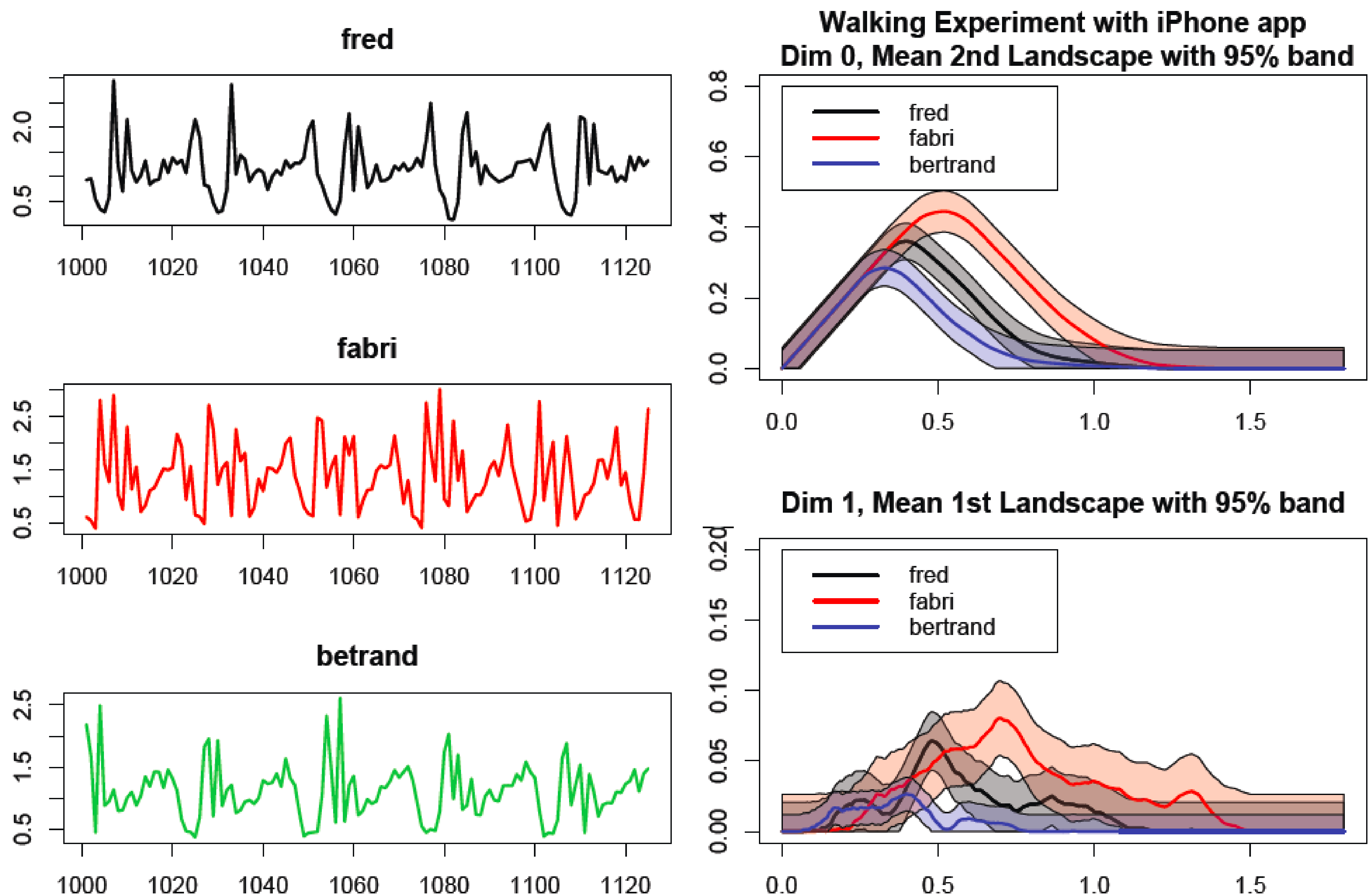


From $n = 100$ subsamples of size $m = 300$

(Sub)sampling and stability of expected landscapes

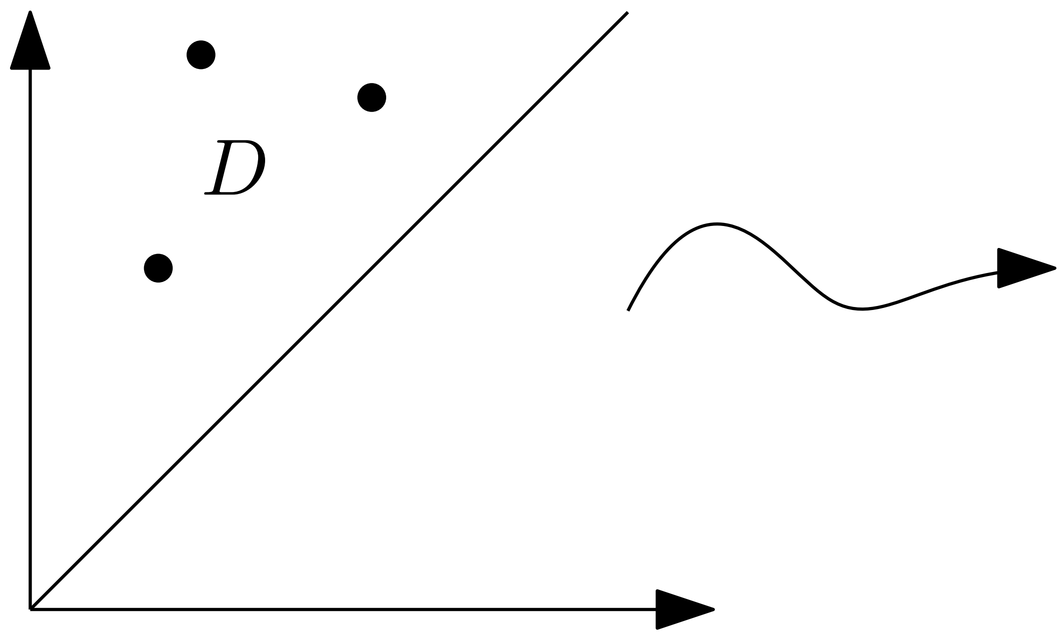
[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]

(Toy) Example: Accelerometer data from smartphone.



- spatial time series (accelerometer data from the smarphone of users).
- no registration/calibration preprocessing step needed to compare!

Persistence diagrams as discrete measures



$$D := \sum_{\mathbf{r} \in D} \delta_{\mathbf{r}}$$

Motivations:

- The space of measures is much nicer that the space of P. D. !
- In the "standard" algebraic persistence theory, persistence diagrams naturally appear as discrete measures in the plane (over rectangles).

  [Chazal, de Silva, Glisse, Oudot 16]
- Many persistence representations can be expressed as

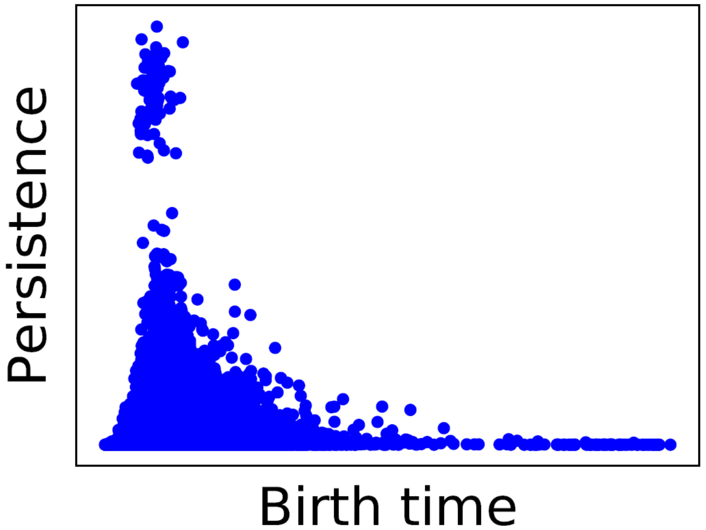
$$D(\phi) = \sum_{\mathbf{r} \in D} \phi(\mathbf{r}) = \int \phi(\mathbf{r}) dD(\mathbf{r})$$

Representation of Persistence diagrams

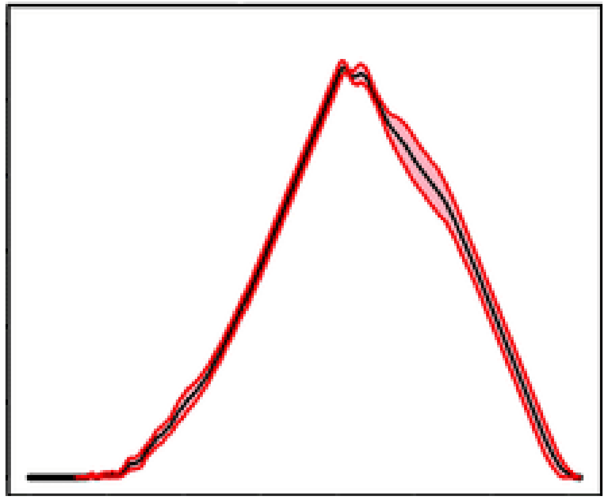
A representation is called linear if there exists $\phi : \mathbb{R}^2_> \to \mathcal{H}$ such that

$$\Phi(D) = \sum_{\mathbf{r} \in D} \phi(r) := D(\phi) = \int \phi(\mathbf{r})\, dD(\mathbf{r})$$

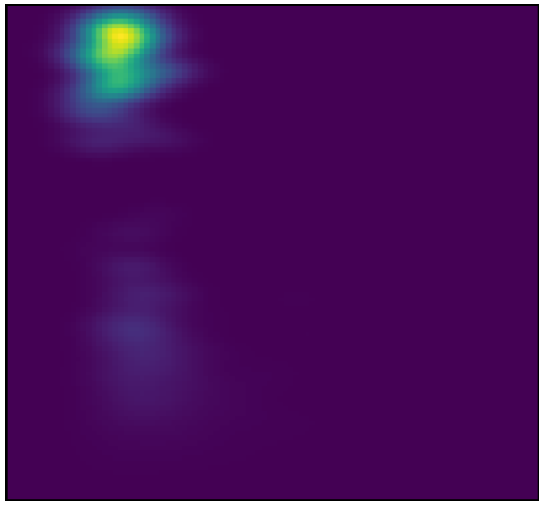
- Many existing representations among the literature:



Distrib. of life span, total persistence,...



Persistent silhouette
[Chazal & al, 2013]



Persistent surface
[Adams & al, 2016]

...

- Linear representations of persistence diagrams are well-suited to be learned from data. [e.g., Hofer et al, NIPS 2017]

Representation of Persistence diagrams

- D is a random persistence diagram (coming from some phenomenon).
- $E[D]$ is a deterministic measure on $\mathbb{R}^2_>$ defined by

$$\forall A \subset \mathbb{R}^2_>,\ E[D](A) = E[D(A)].$$

- $D_1, \ldots, D_N$ i.i.d.

$$\begin{aligned} \overline{\Phi} &= \frac{\Phi(D_1) + \cdots + \Phi(D_N)}{N} \\ &= \overline{\mu}(\phi) \\ &\approx E[D](\phi) \\ E[D](\phi) &= \int_{\mathbb{R}^2_>} \phi(\mathbf{r}) p(\mathbf{r}) d\,\mathbf{r} \end{aligned}$$

Under mild assumptions, $E[D]$ has a density w.r.t. Lebesgue measure in $\mathbb{R}^2$

The density of expected persistence diagrams

[C. - Divol, 2018]

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 some (not very strong) assumptions.

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 $\mathbb{R}^2_> = \{(b, d) \in \mathbb{R}^2 : b \leq d\}$.

The density of expected persistence diagrams

[C. - Divol, 2018]

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

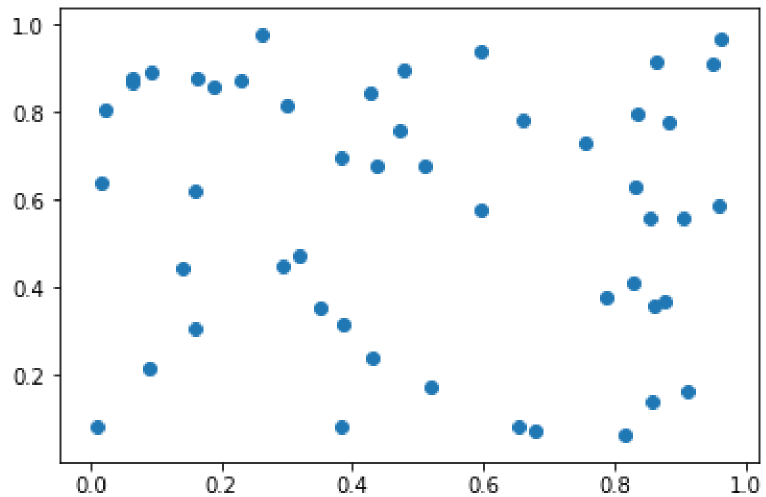
- M is a real analytic (compact) d-dimensional connected submanifold possibly with boundary,
- $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$,
- $\mathcal{K}$ satisfies some (not very strong) assumptions.

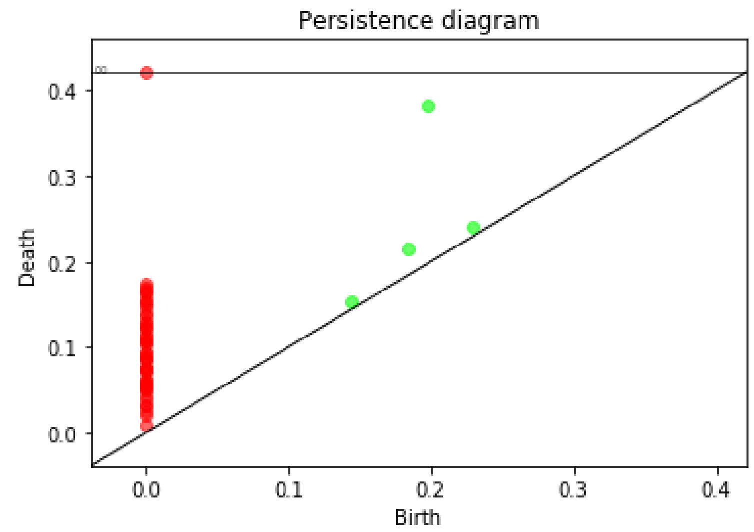
Then, for $s \geq 0$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on the half plane $\mathbb{R}^2_> = \{(b,d) \in \mathbb{R}^2 : b \leq d\}$.

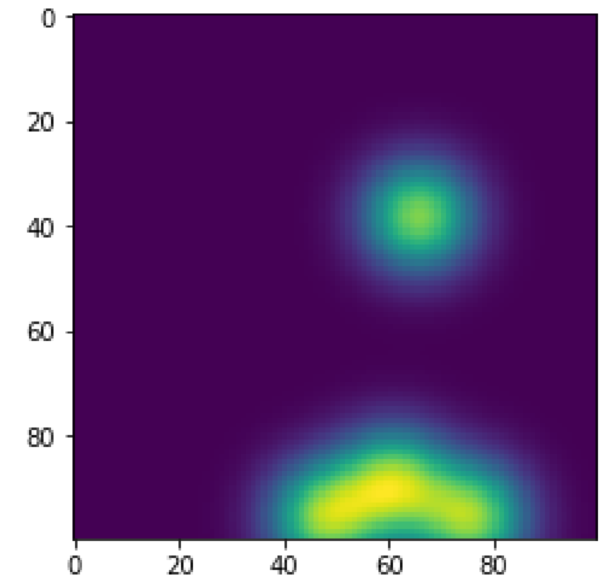
Theorem [smoothness]: Under the assumption of previous theorem, if moreover $\mathbb{X} \in M^n$ has a density of class C^k with respect to $\mathcal{H}_{nd}$. Then, for $s \geq 0$, the density of $E[D_s[\mathcal{K}(\mathbb{X})]]$ is of class C^k.

Persistence images

[Adams et al, JMLR 2017]







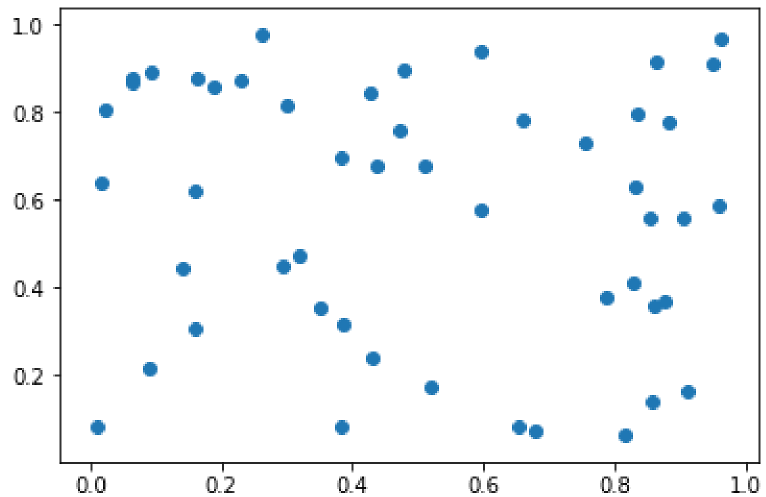
For $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel and H a bandwidth matrix (e.g. a symmetric positive definite matrix), pose for $u \in \mathbb{R}^2$, $K_H(z) = |H|^{-1/2} K(H^{-1/2} \cdot u)$

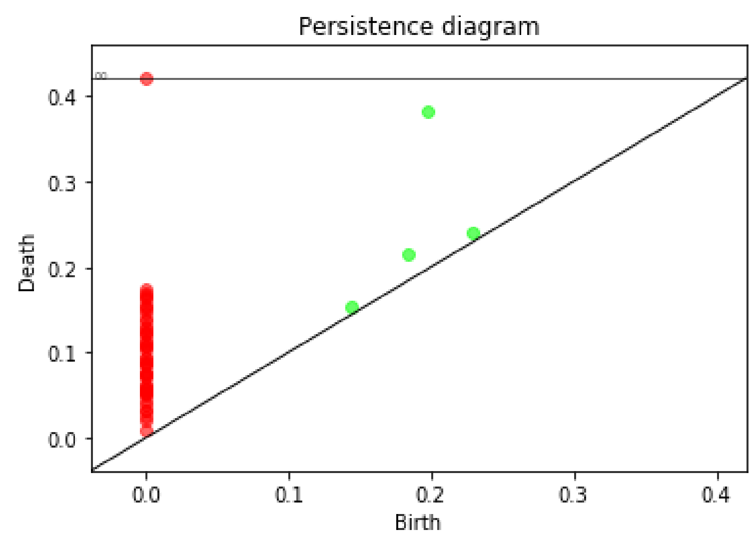
For $D = \sum_i \delta_{\mathbf{r}_i}$ a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by:

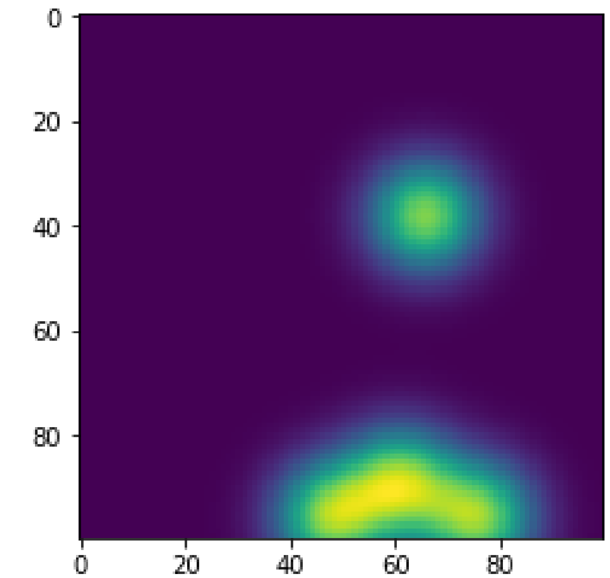
$$\forall z \in \mathbb{R}^2,\ \rho(D)(u) = \sum_i w(\mathbf{r}_i) K_H(u - \mathbf{r}_i) = D(w K_H(u - \cdot))$$

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

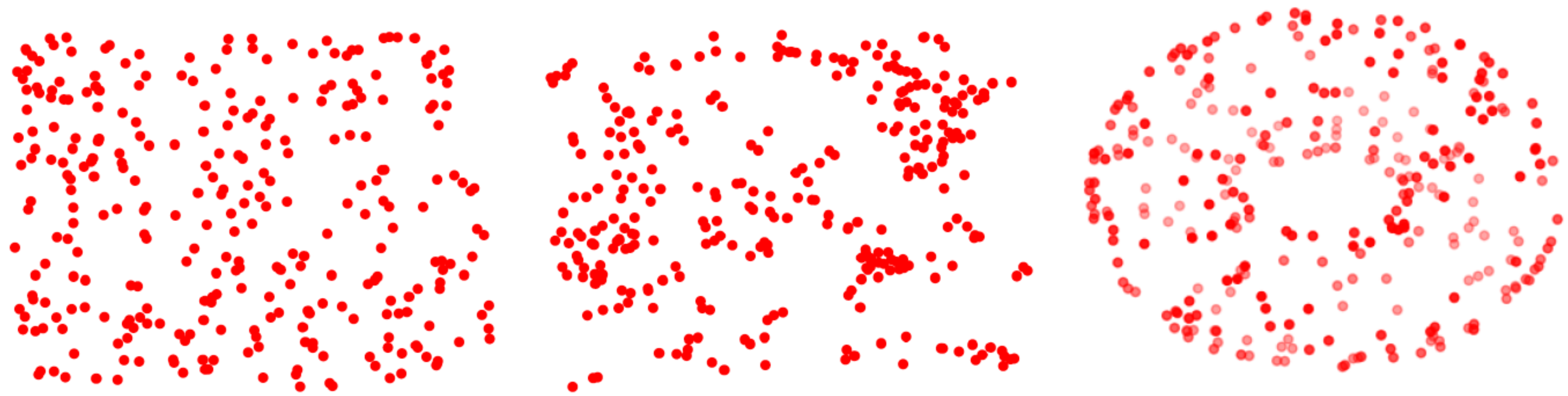
For $D = \sum_i \delta_{\mathbf{r}_i}$ a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by:

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

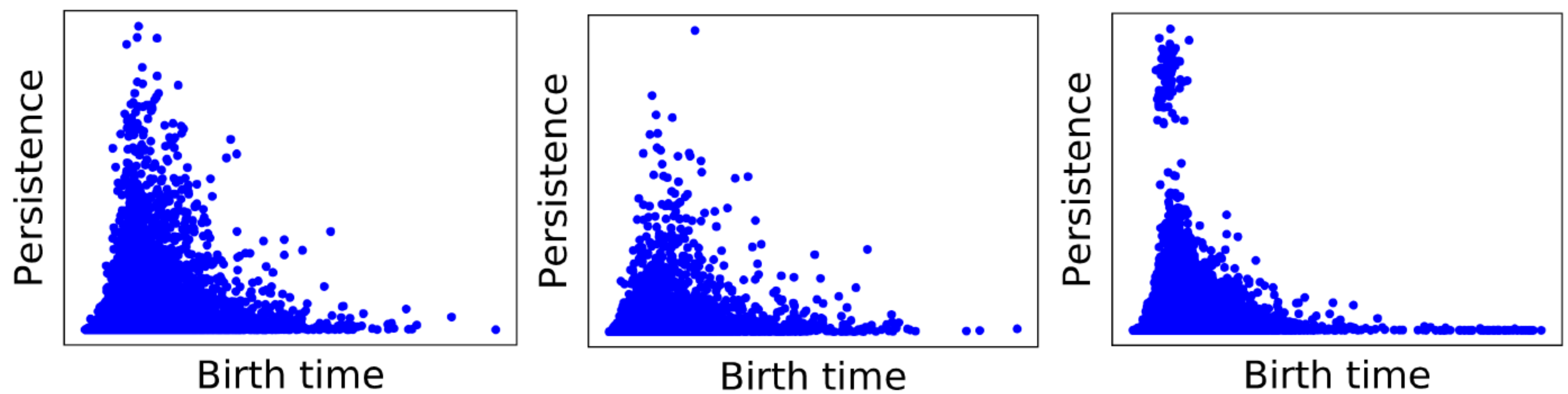
$\Rightarrow$ persistence surfaces can be seen as kernel based estimators of $E[D_s[\mathcal{K}(\mathbb{X})]]$.

Persistence images

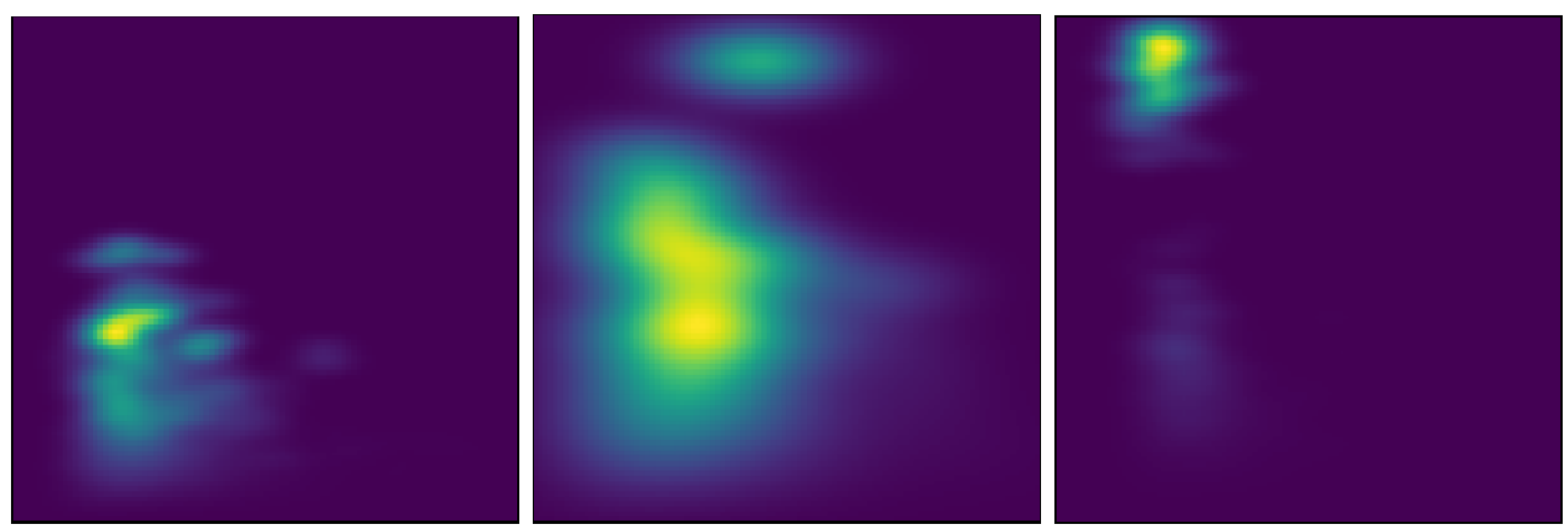
The realization of 3 different processes



The overlay of 40 different persistence diagrams

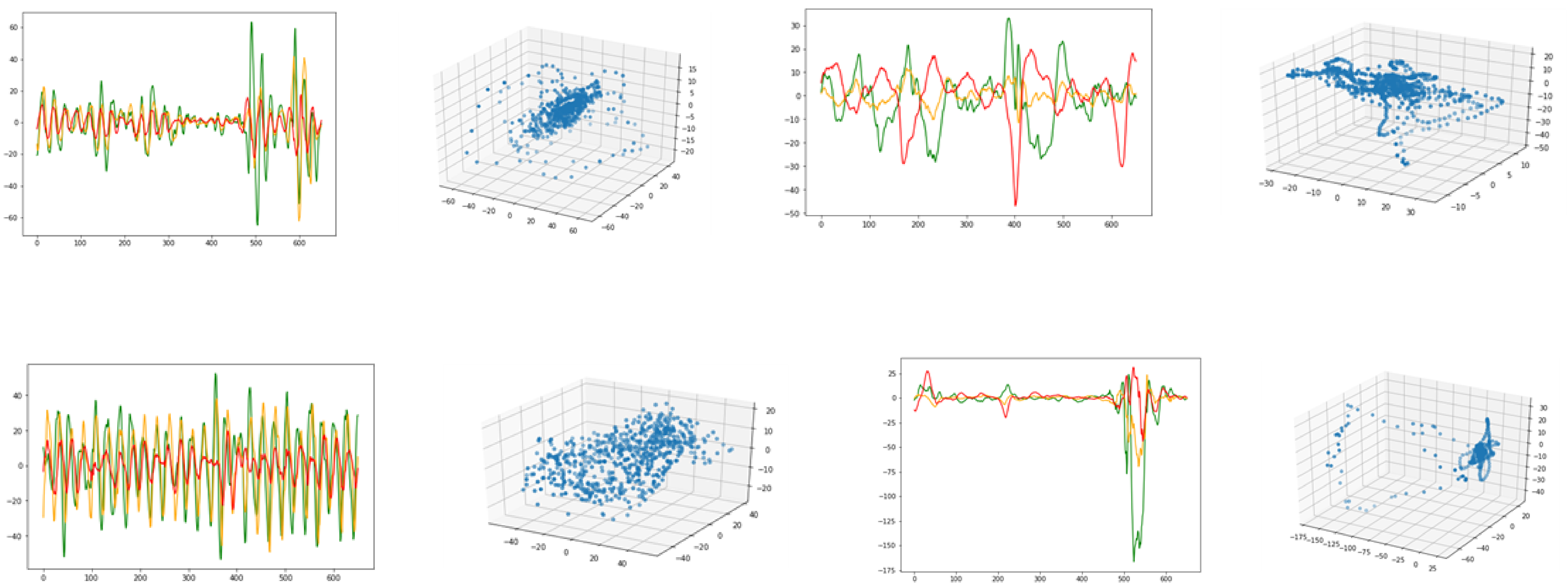


The persistence images with weight function $w(\mathbf{r}) = (r_2 - r_1)^3$ and bandwith selected using cross-validation.



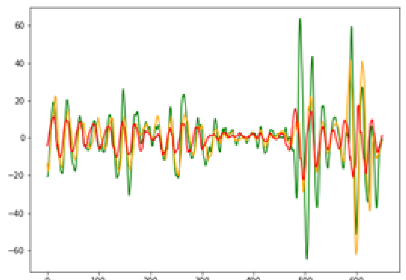
TDA and Machine Learning: some examples and illustrations.

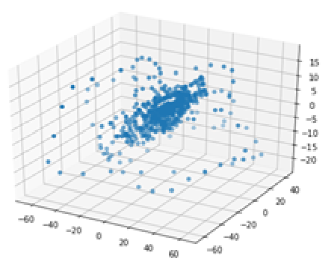
TDA and Machine Learning for time-dependent data



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

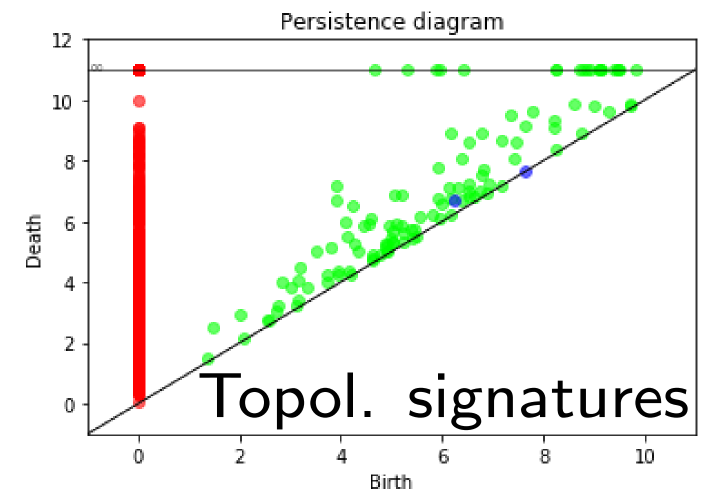
TDA and Machine Learning for time-dependent data





TDA pipeline

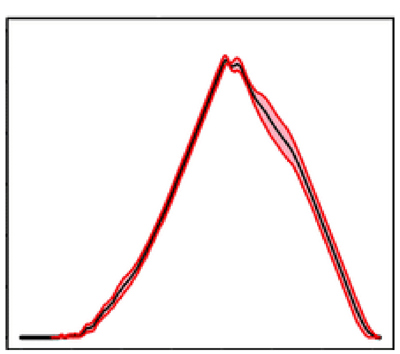
GUDHI software



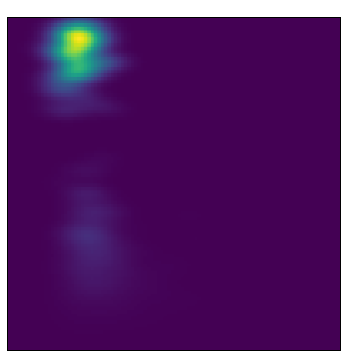
Topol. signatures

Feature engineering

Representations of persistence (linearization):



Persistent silhouette
[Chazal & al, 2013]



Persistent surface
[Adams & al, 2016]

. . .

ML/AI

Features extraction

Random forests

Deep learning

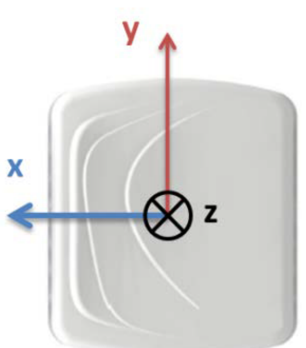
Etc...
(combined with other features)

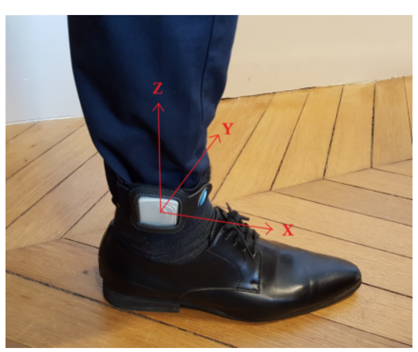
With landscapes: patient monitoring

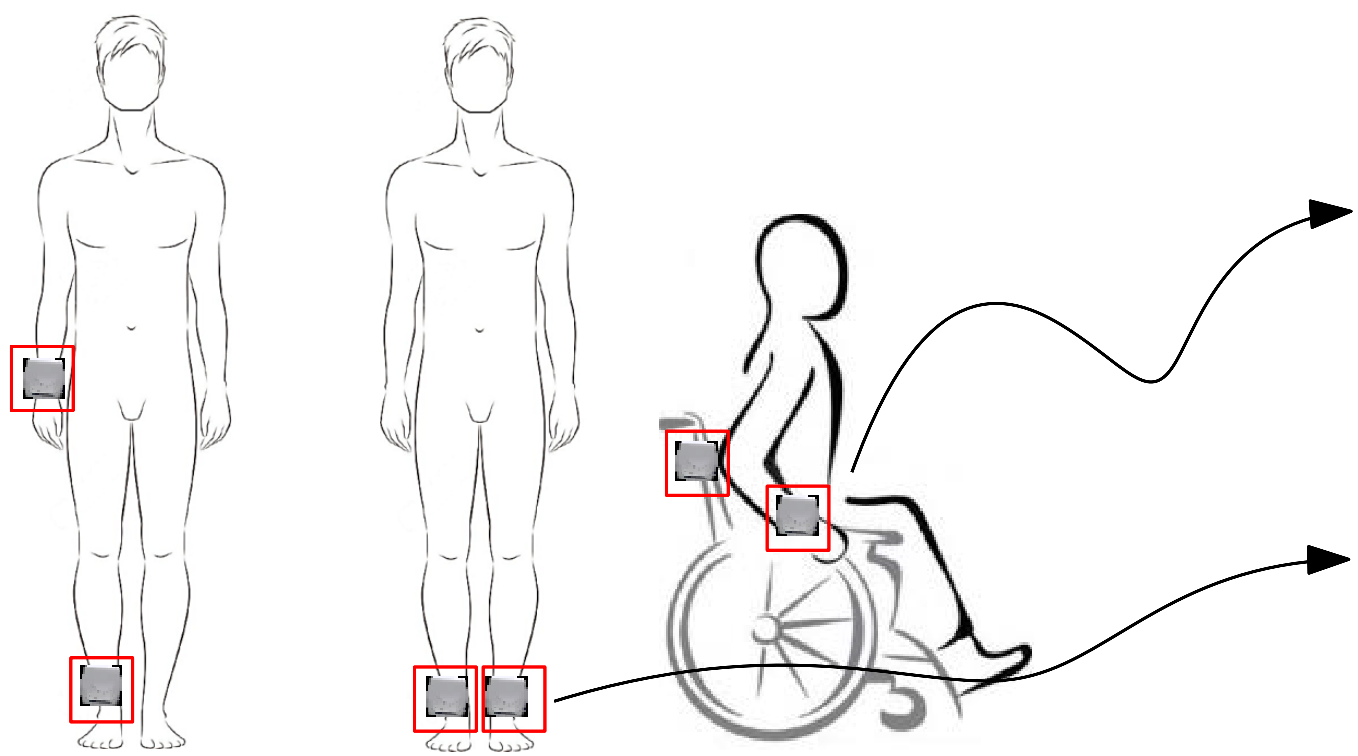
A joint industrial research project between

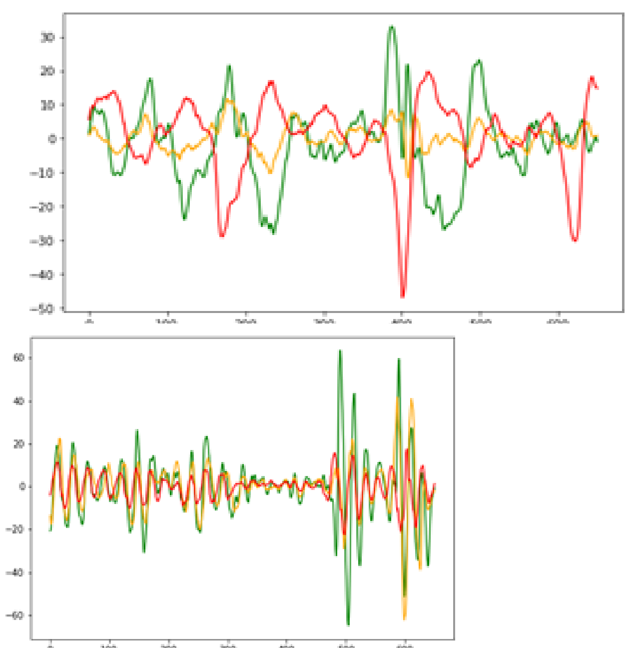
sysnav NAVIGATION TECHNOLOGIES and Inria informatics mathematics

A French SME with innovating technology to reconstruct pedestrian trajectories from inertial sensors (ActiMyo)









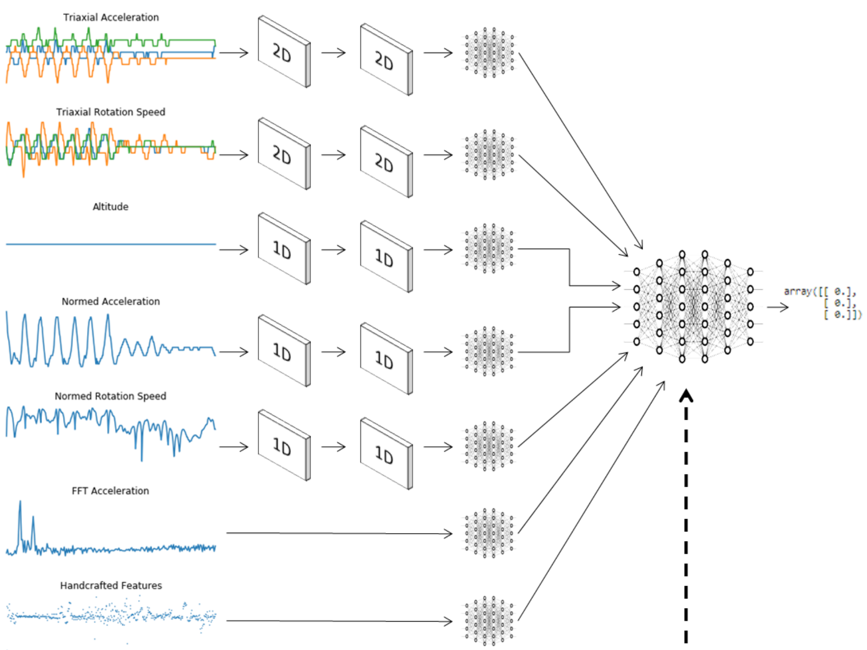
"Chaotic" time-dependent data

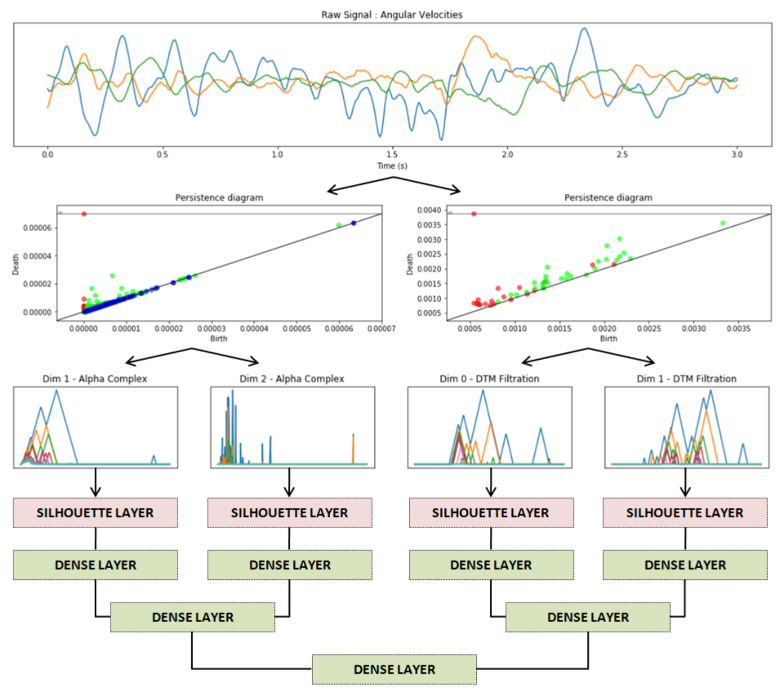
Objective: precise analysis of movements and activities of pedestrians.

Targeted applications: personal healthcare; medical studies; defense.

With landscapes: patient monitoring

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.

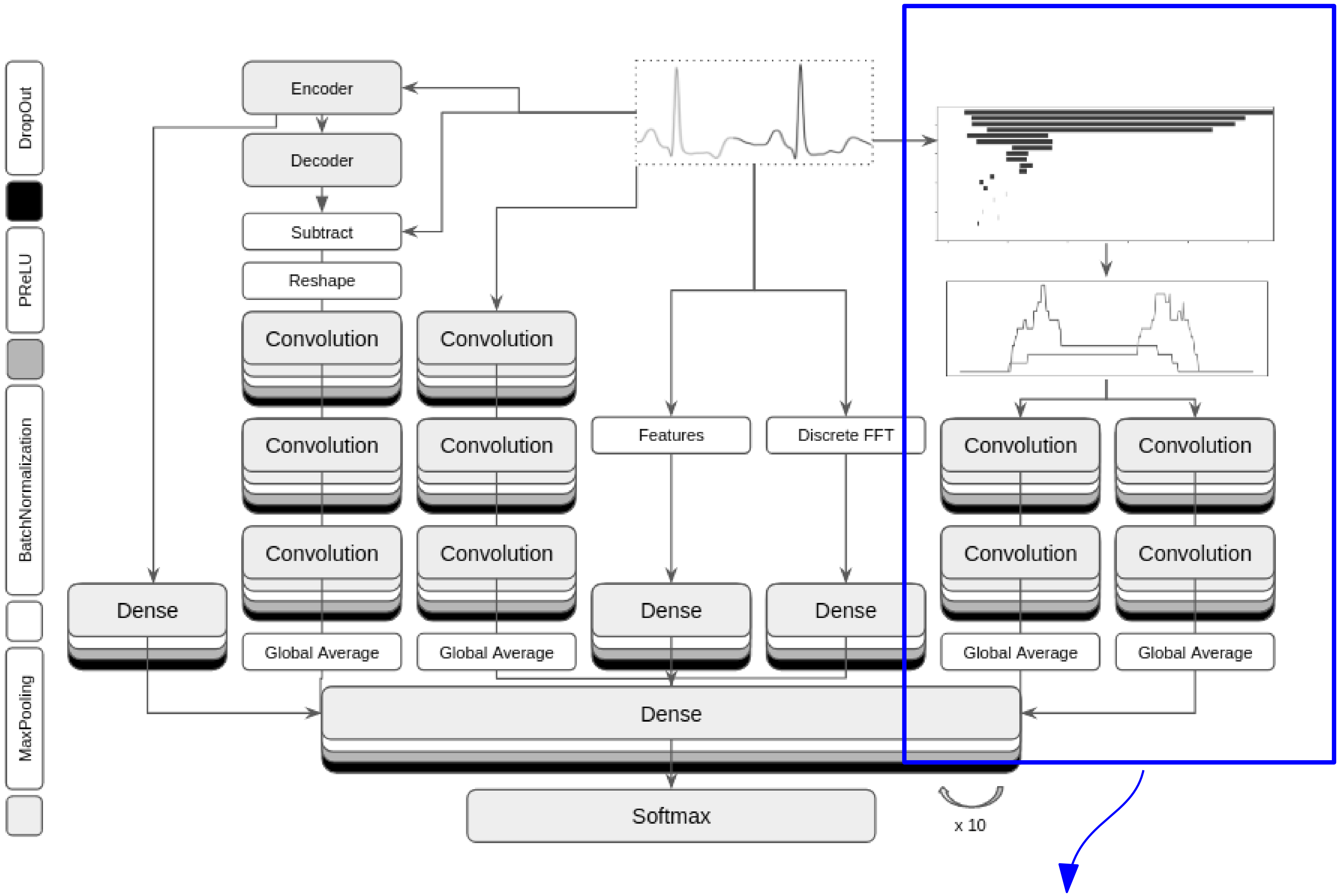
sysnav NAVIGATION TECHNOLOGIES

Inria informatics mathematics

With Betti curves: arrhythmia detection

Joint research project between Inria DataShape and Fujitsu

Objective: Arrythmia detection from ECG data.



TDA channel: Betti curves processed as 1D signal

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

FUJITSU Inria informatics mathematics

Thank you for your attention!

To get more details and more references:

- F. Chazal, B. Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. https://arxiv.org/abs/1710.04019
- J.-D. Boissonnat, F. Chazal, M. Yvinec. Geometric and Topological Inference. Cambridge University Press, 2018.

Software:

- The Gudhi library (C++/Python): https://project.inria.fr/gudhi/software/
- R package TDA