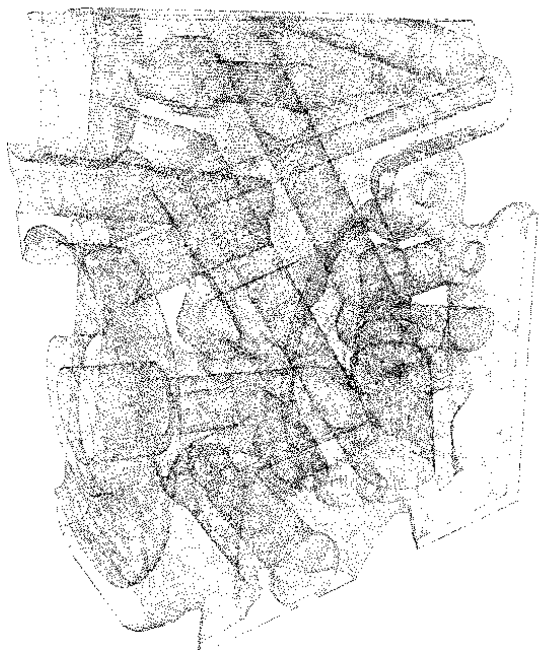
Seminar@SystemX
June 20, 2019

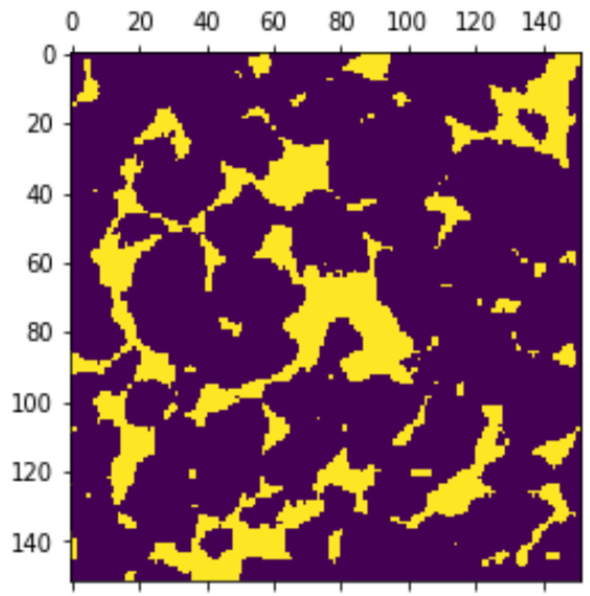
Understanding the shape of data: a brief introduction to Topological Data Analysis

Frédéric Chazal
DataShape team
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr

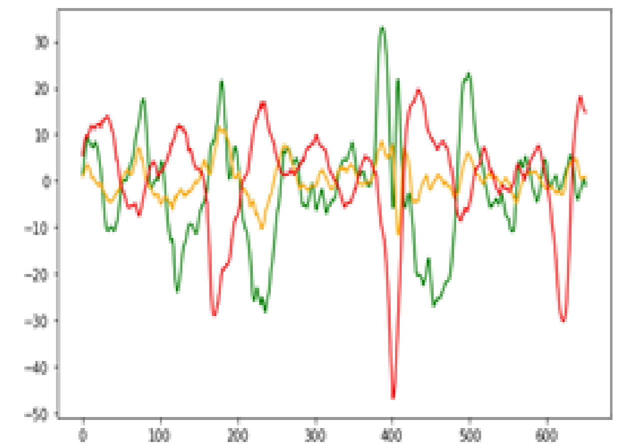
What is topological structure of data?

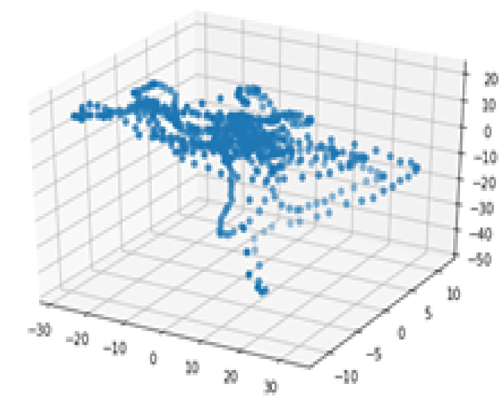


[Scanned 3D object]



[3D images (porous rocks)]

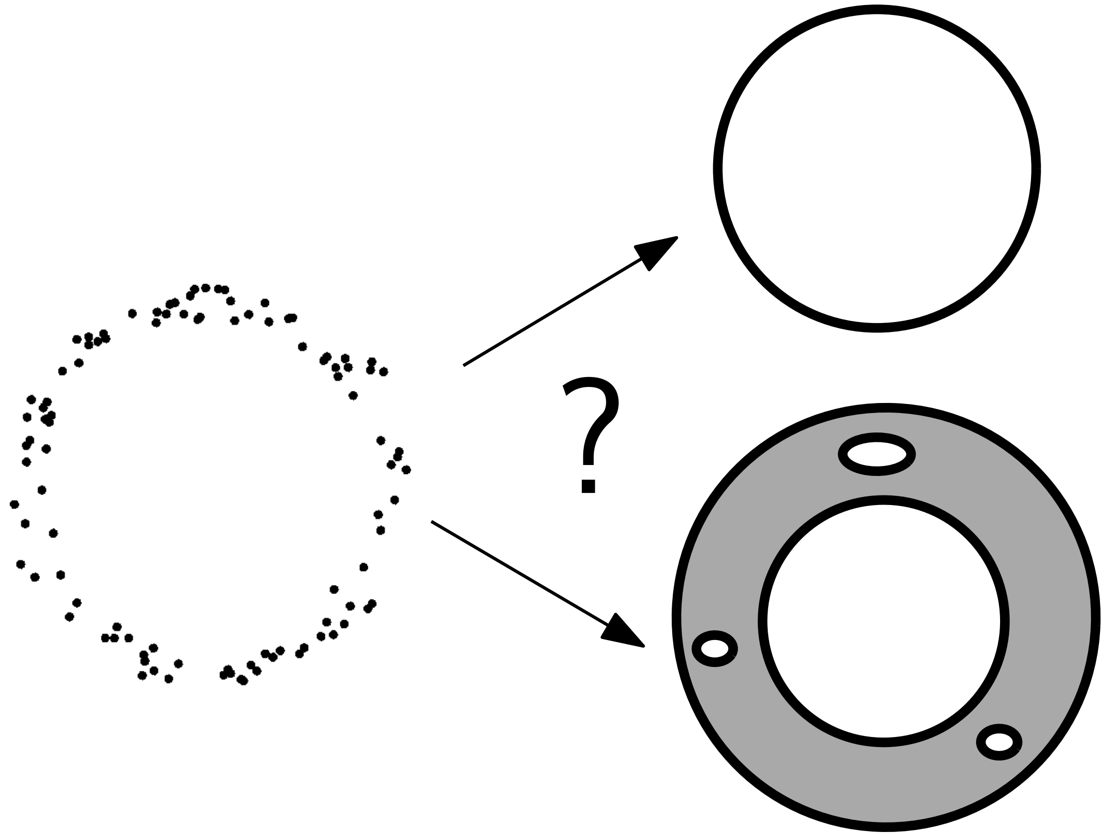


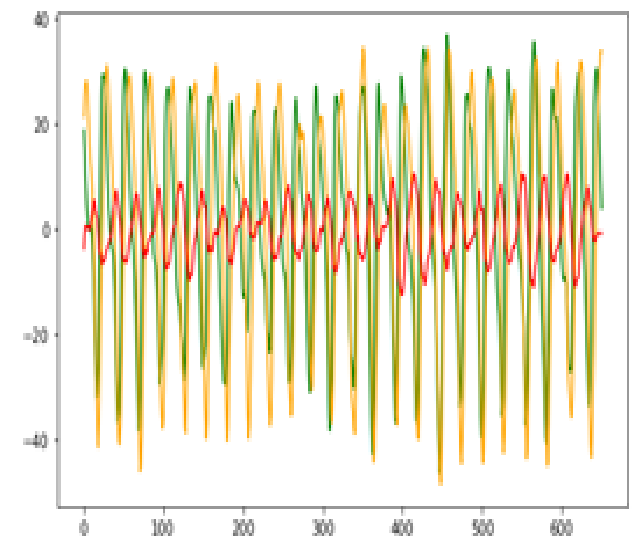


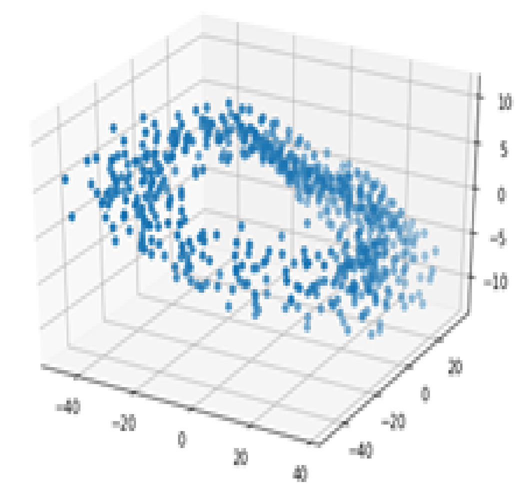
[Sensors (Sysnav courtesy)]

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

What is topological structure of data?

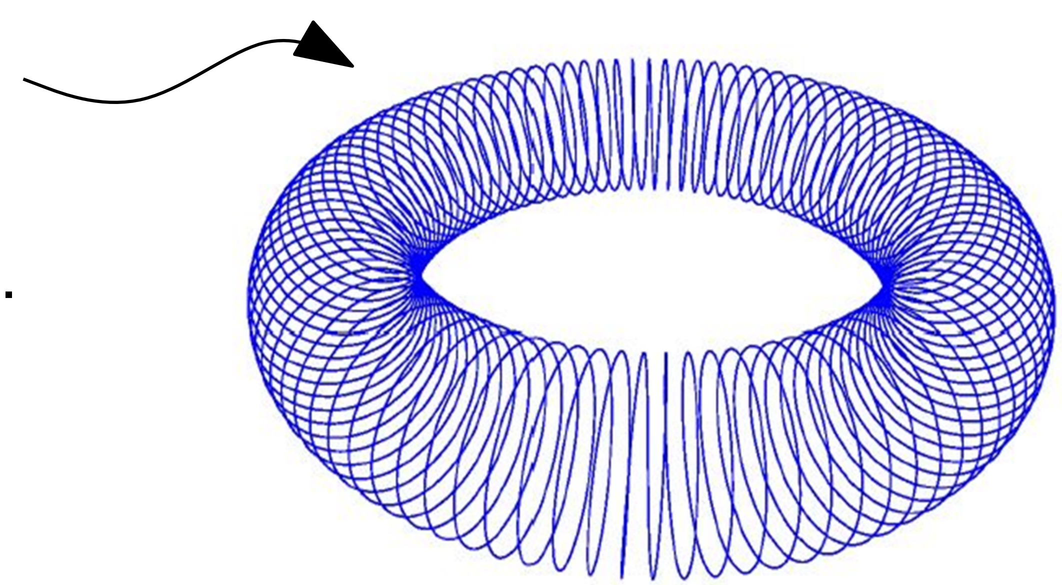






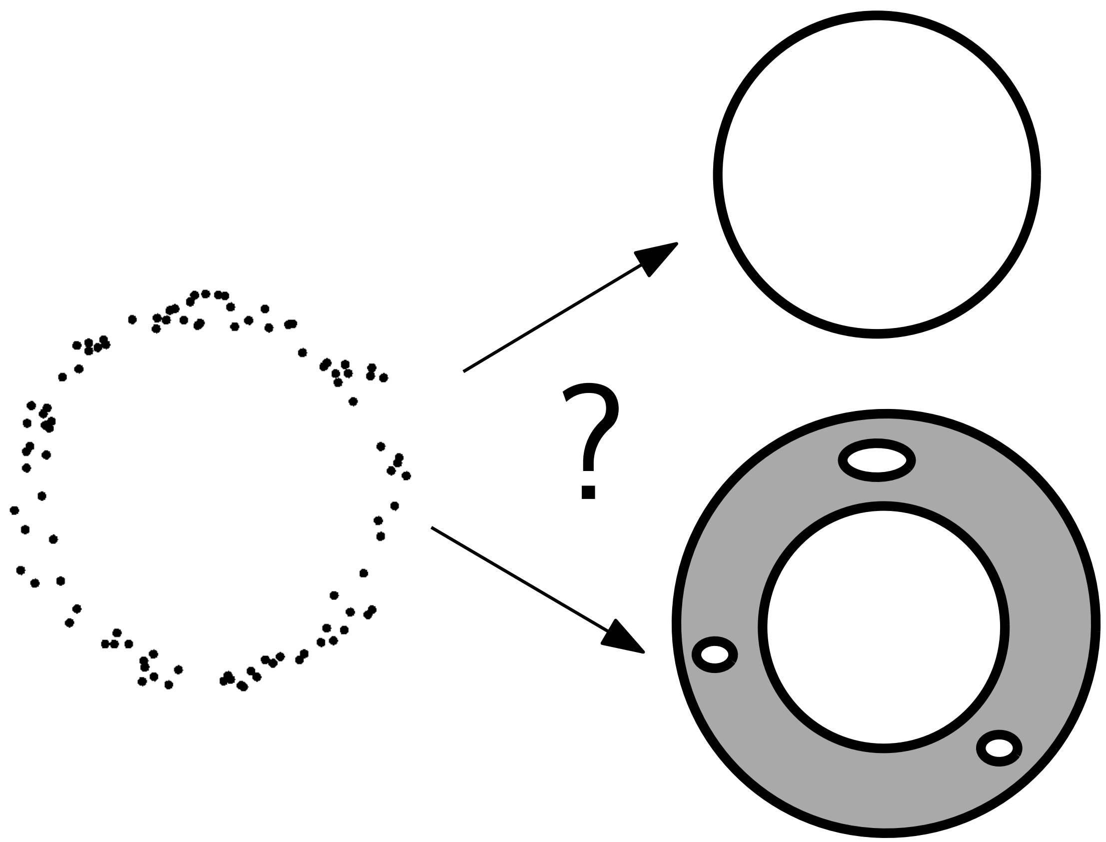
A non obvious problem:

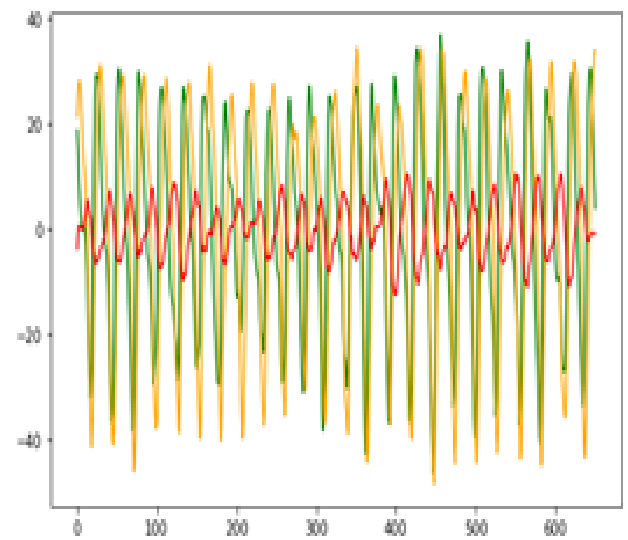
- → no direct access to topological/geometric information: need of intermediate constructions (simplicial complexes);
- → distinguish topological "signal" from noise;
- → topological information may be multiscale;
- → statistical analysis of topological information.

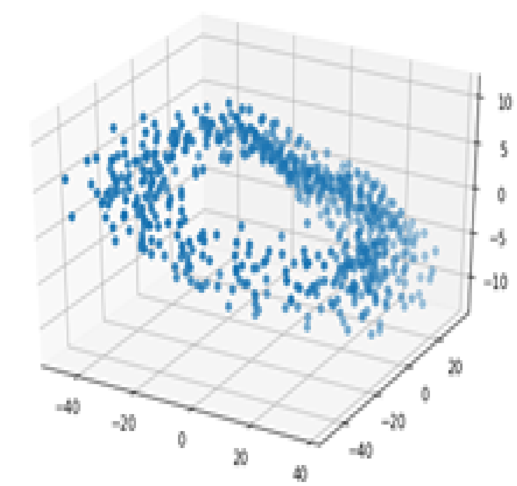


Topological Data Analysis (TDA)
Persistent homology!

What is topological structure of data?

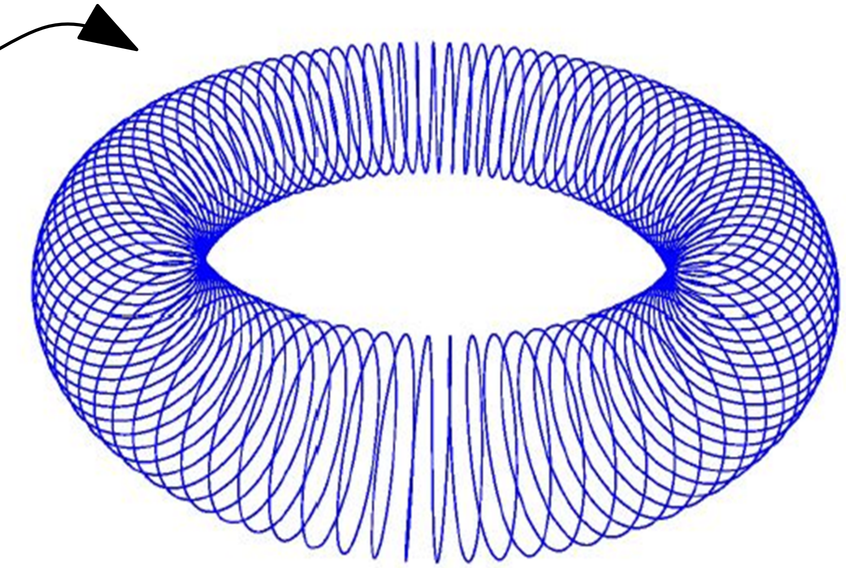






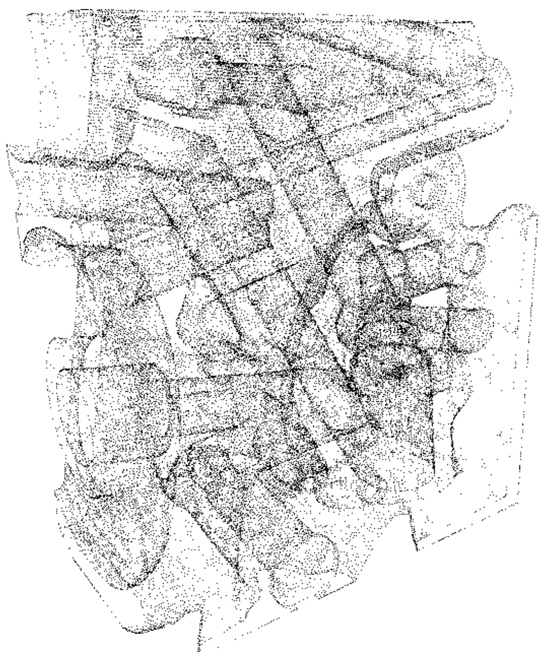
A non obvious problem:

- → no direct access to topological/geometric information: need of intermediate constructions (simplicial complexes);
- → distinguish topological "signal" from noise;
- → topological information may be multiscale;
- → statistical analysis of topological information.

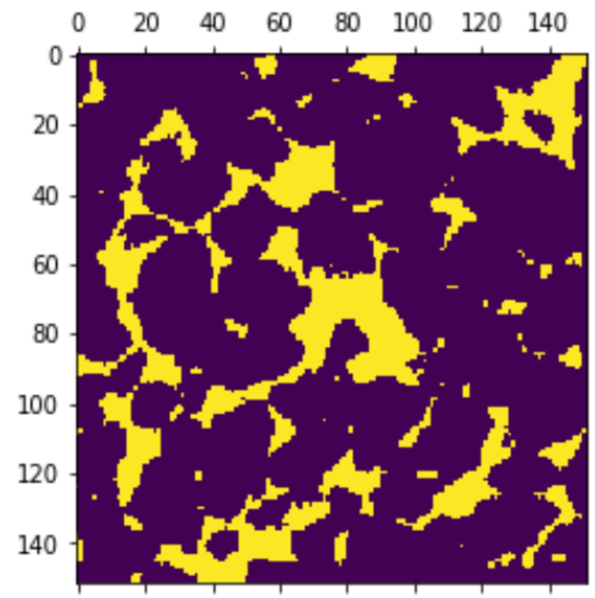


Topological Data Analysis (TDA)
Persistent homology!

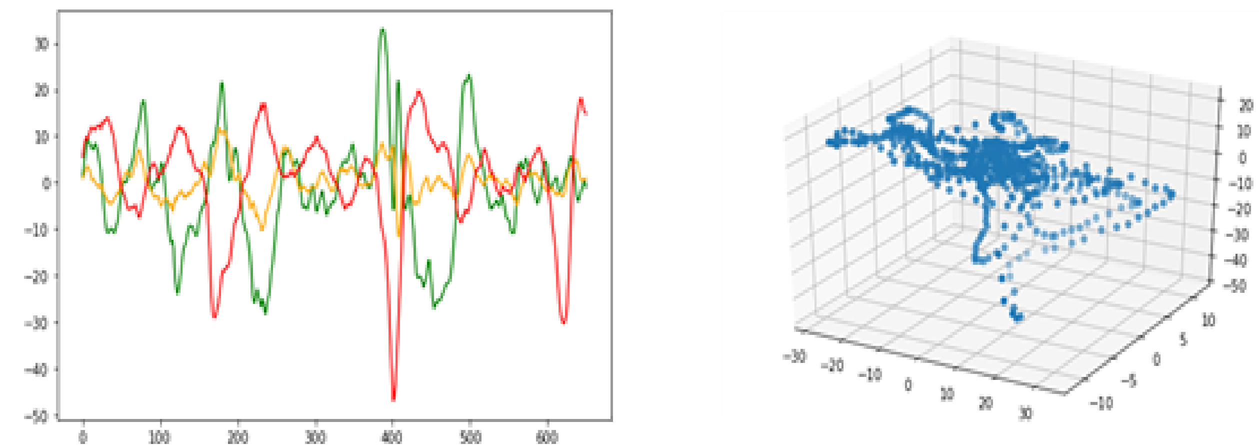
What is Topological Data Analysis (TDA)?



[Scanned 3D object]



[3D images (porous rocks)]



[Sensors (Sysnav courtesy)]

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

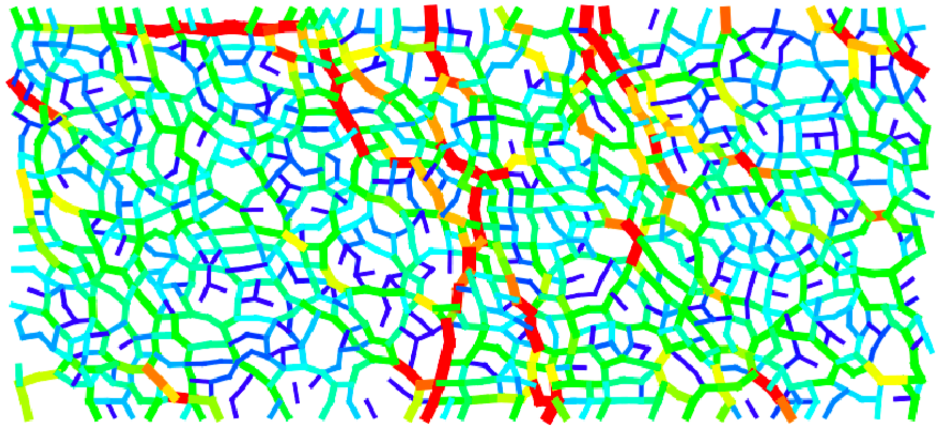
- infer relevant topological and geometric features from complex data,
- take advantage of topological/geometric information for further Data Analysis, Machine Learning and AI tasks.

For what kind of data is TDA useful?

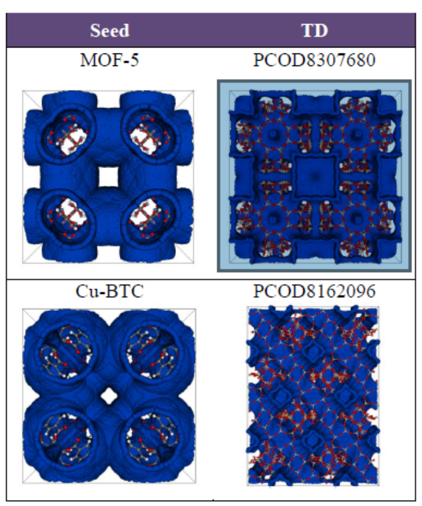
- Complex data!

For what kind of data is TDA useful?

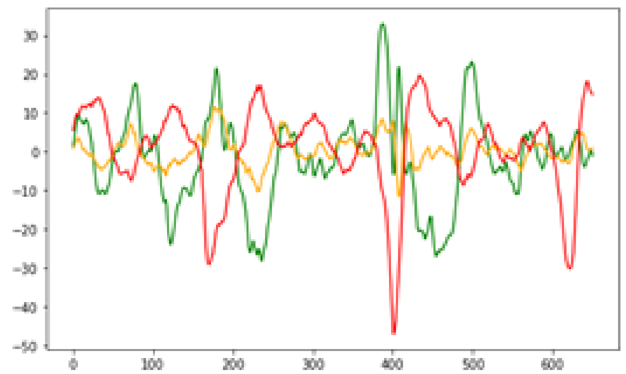
- Complex data!
- Examples (where TDA brings real added value):

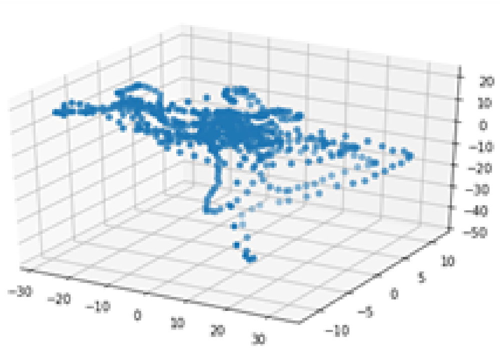


Force fields in granular media



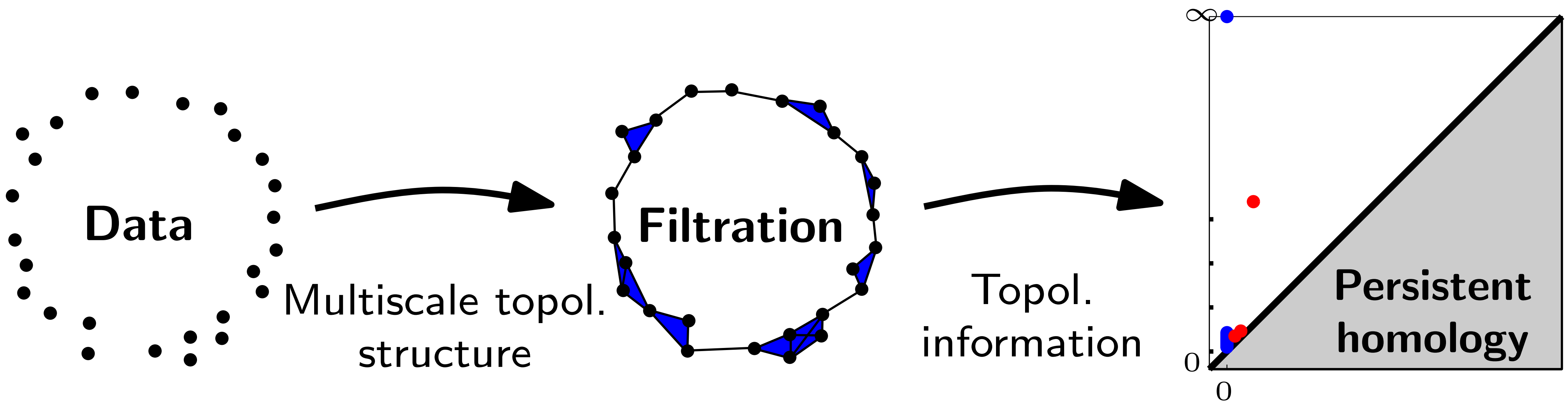
Nanomaterial design

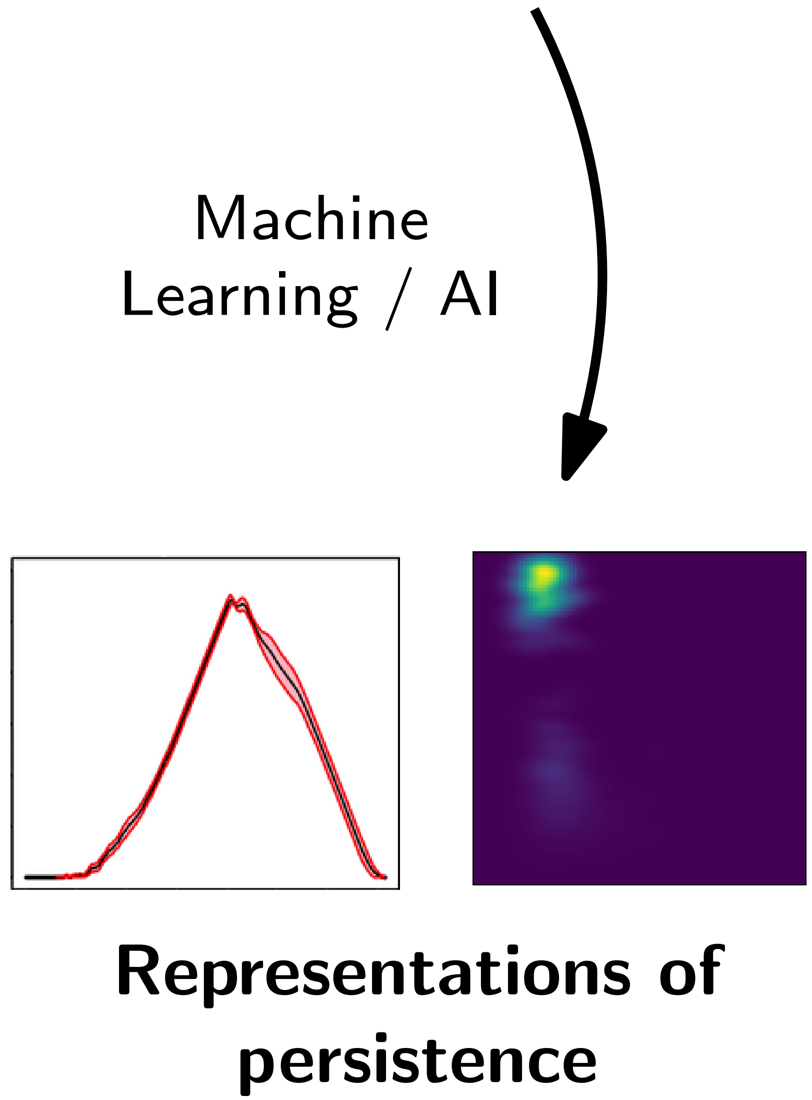




(Chaotic) time-dependent data - see later in the talk

The classical TDA pipeline





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

Persistent homology

The theory of persistence

A recent theory that is subject to intense research activities:

- from the mathematical perspective:

- general algebraic framework (persistence modules) and general stability results.
- extensions and generalizations of persistence (zig-zag persistence, multi-persistence, etc...)
- Statistical analysis of persistence.

- from the algorithmic and computational perspective:

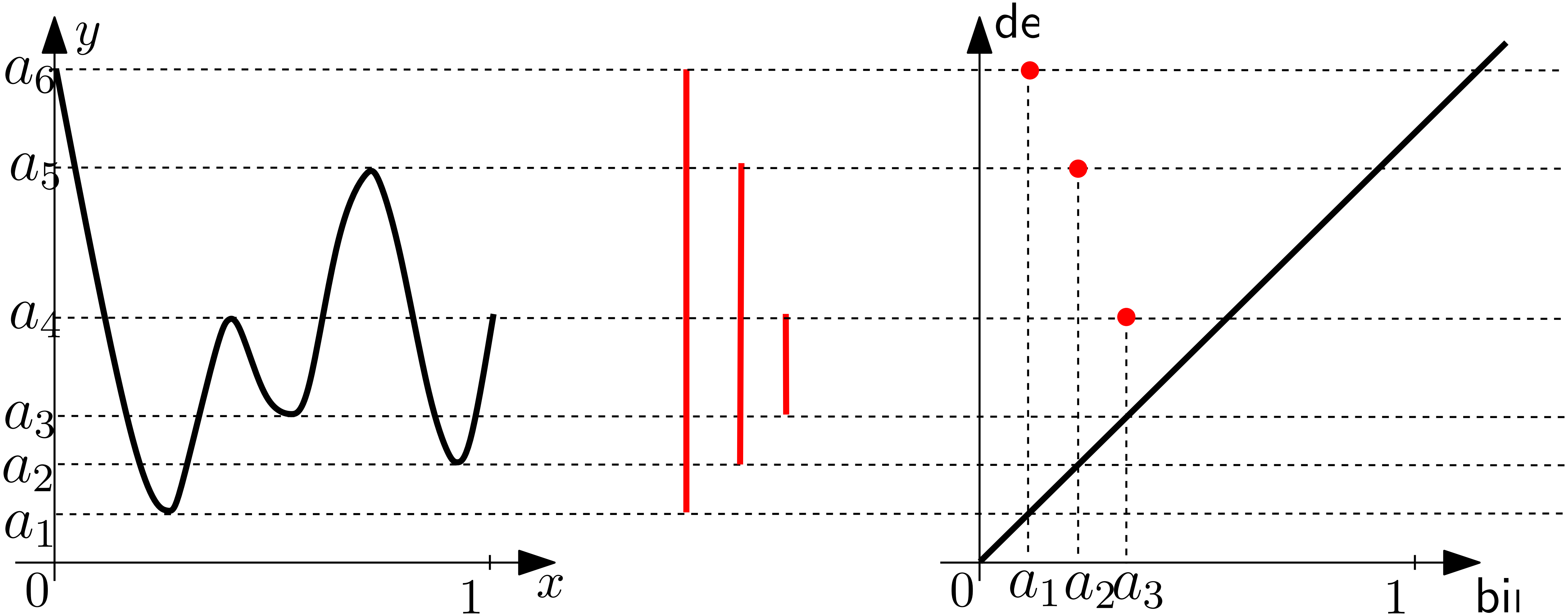
- efficient algorithms to compute persistence and some of its variants.
- efficient software libraries (in particular, Gudhi: https://project.inria.fr/gudhi/).

- from the data science perspective:

- representations of persistence that are suitable for Machine Learning
- Topological/geometric information in combination with other features

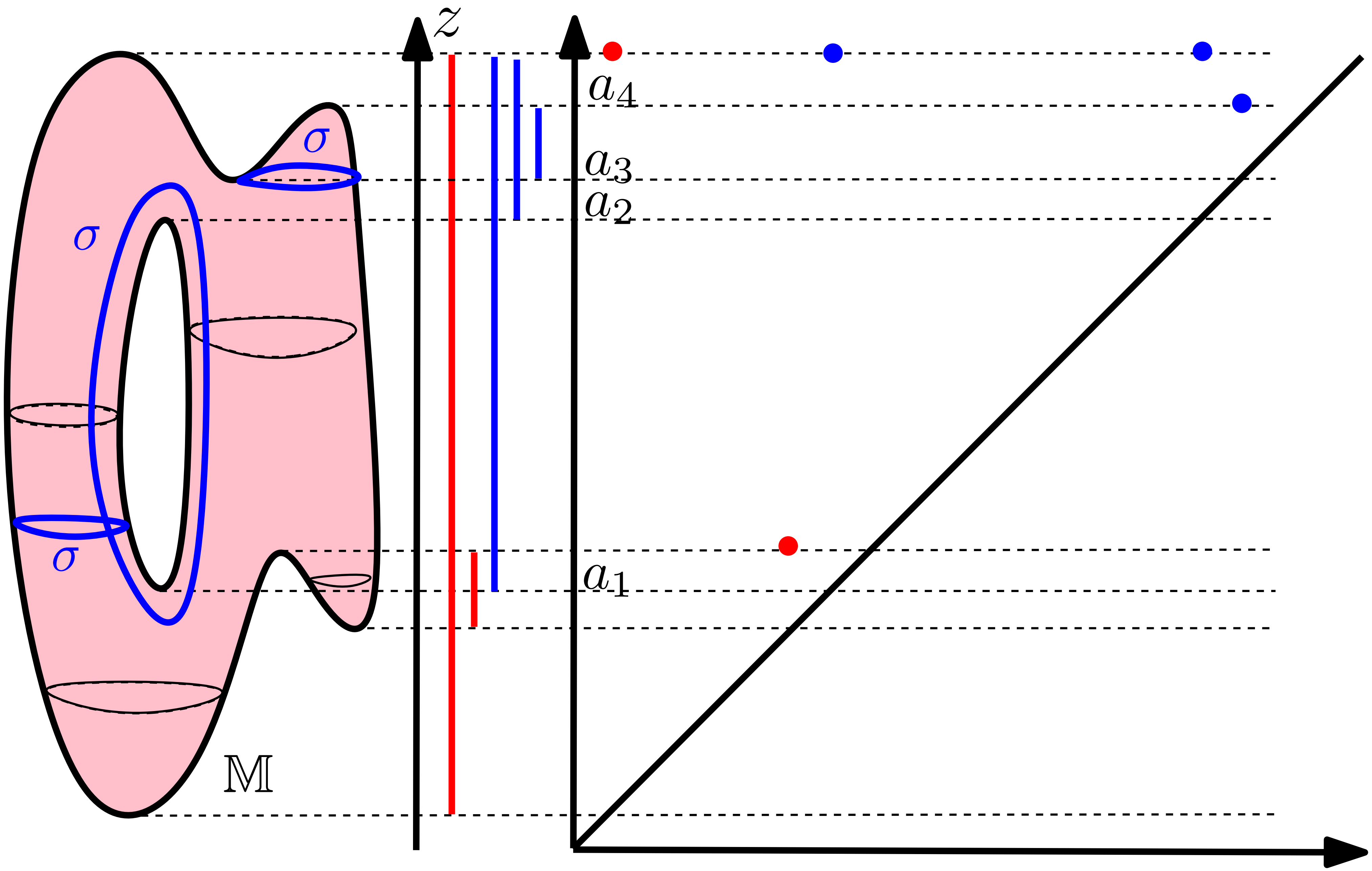
A whole machinery at the crossing of mathematics and computer science!

Persistent homology for functions



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

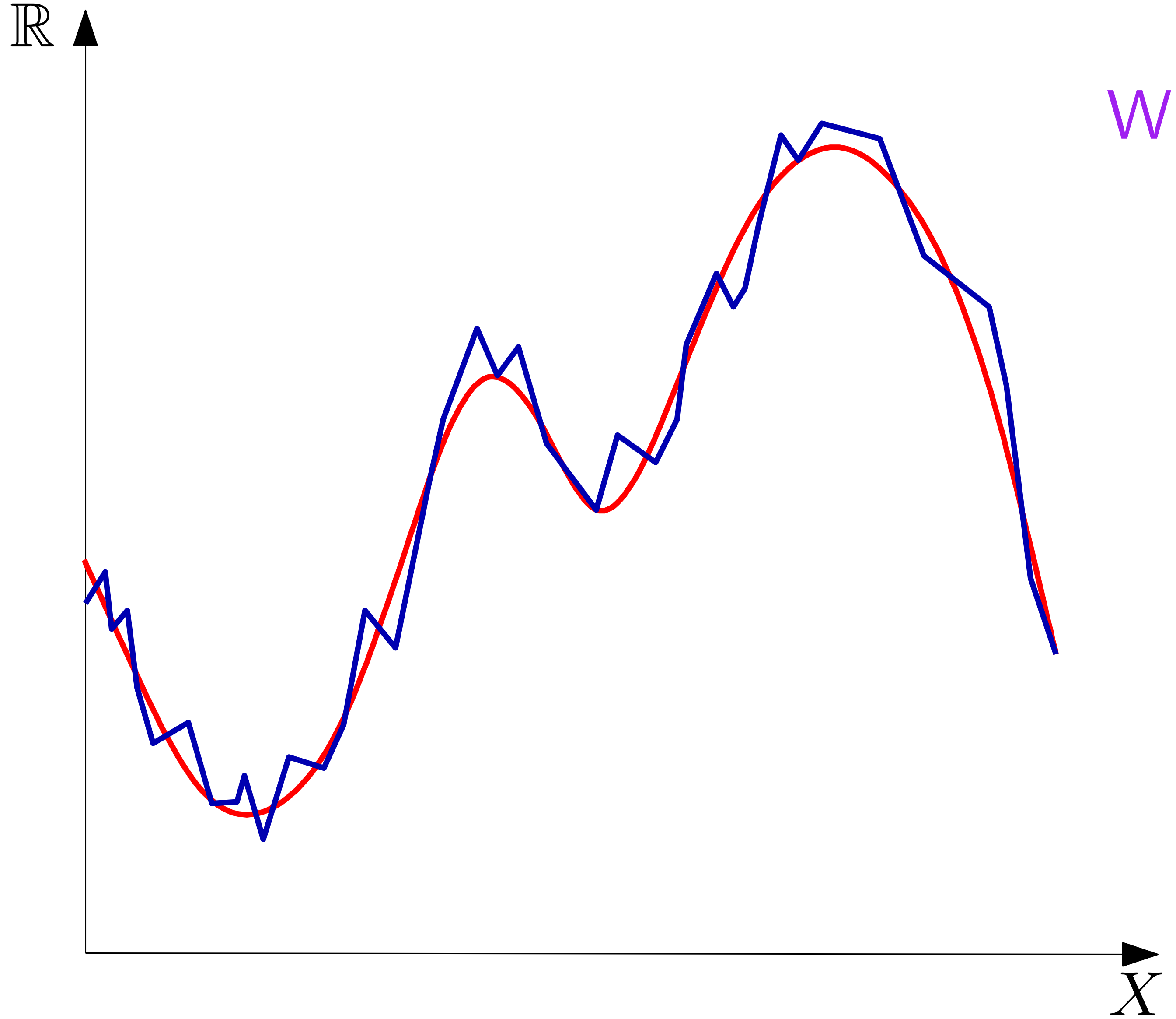
Persistent homology for functions



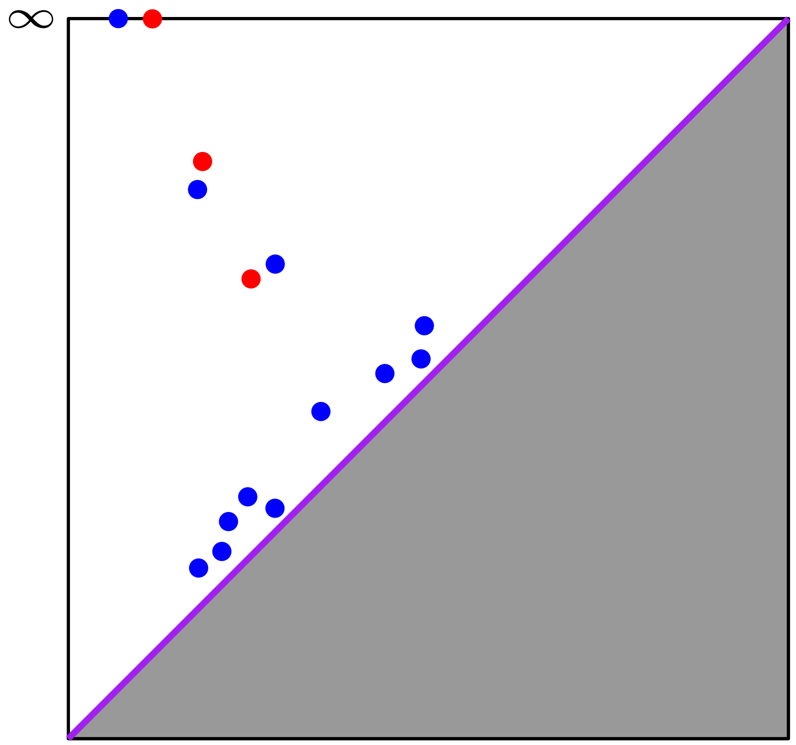
Tracking and encoding the evolution of the connected components (0-dimensional homology) and cycles (1-dimensional homology) of the sublevel sets.

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).

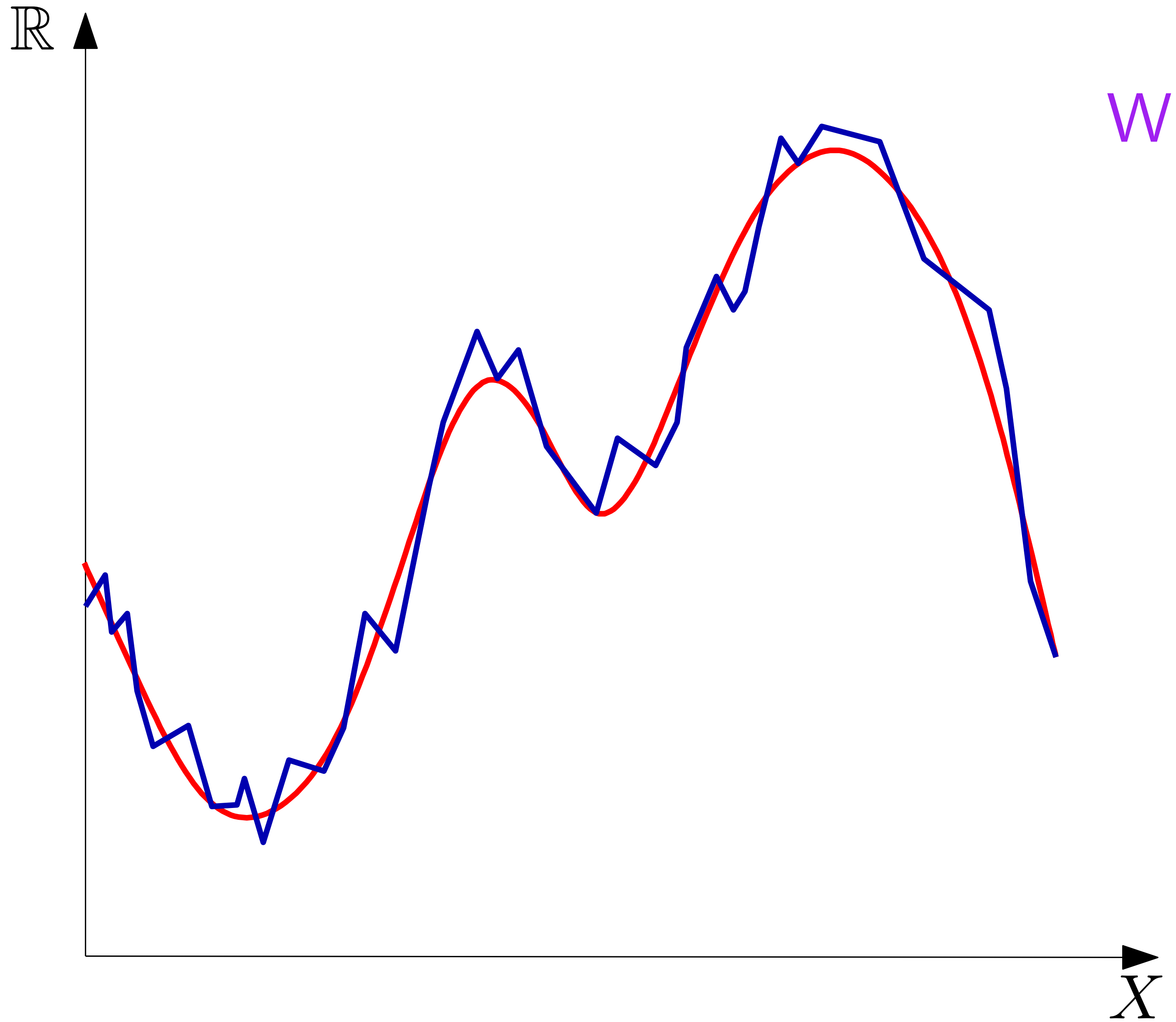
Stability properties



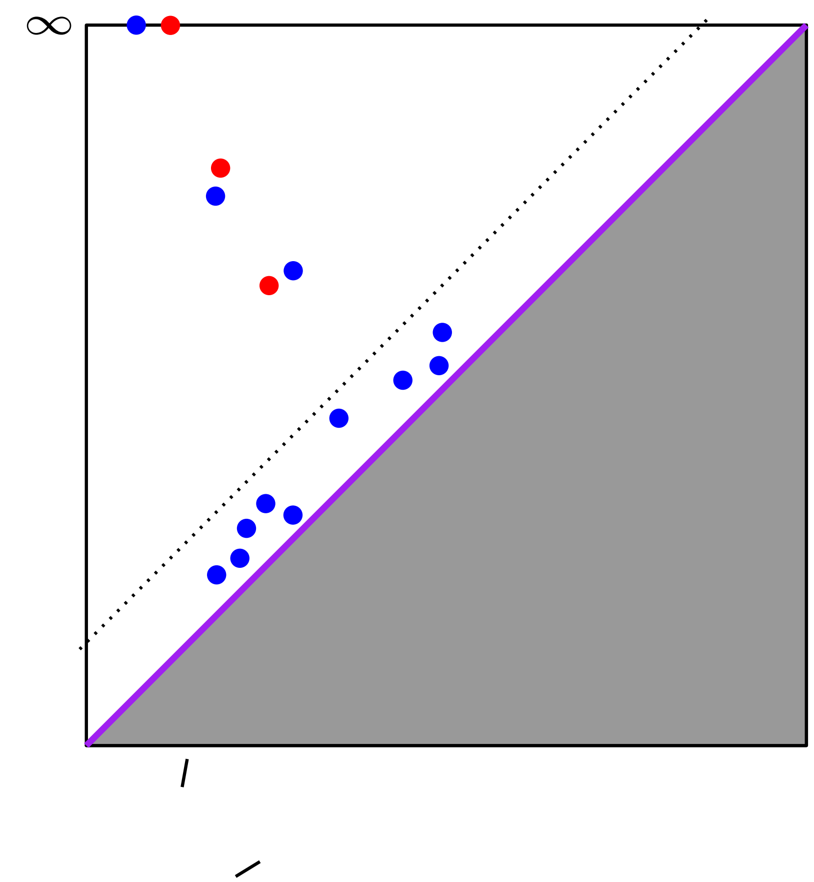
What if f is slightly perturbed?



Stability properties



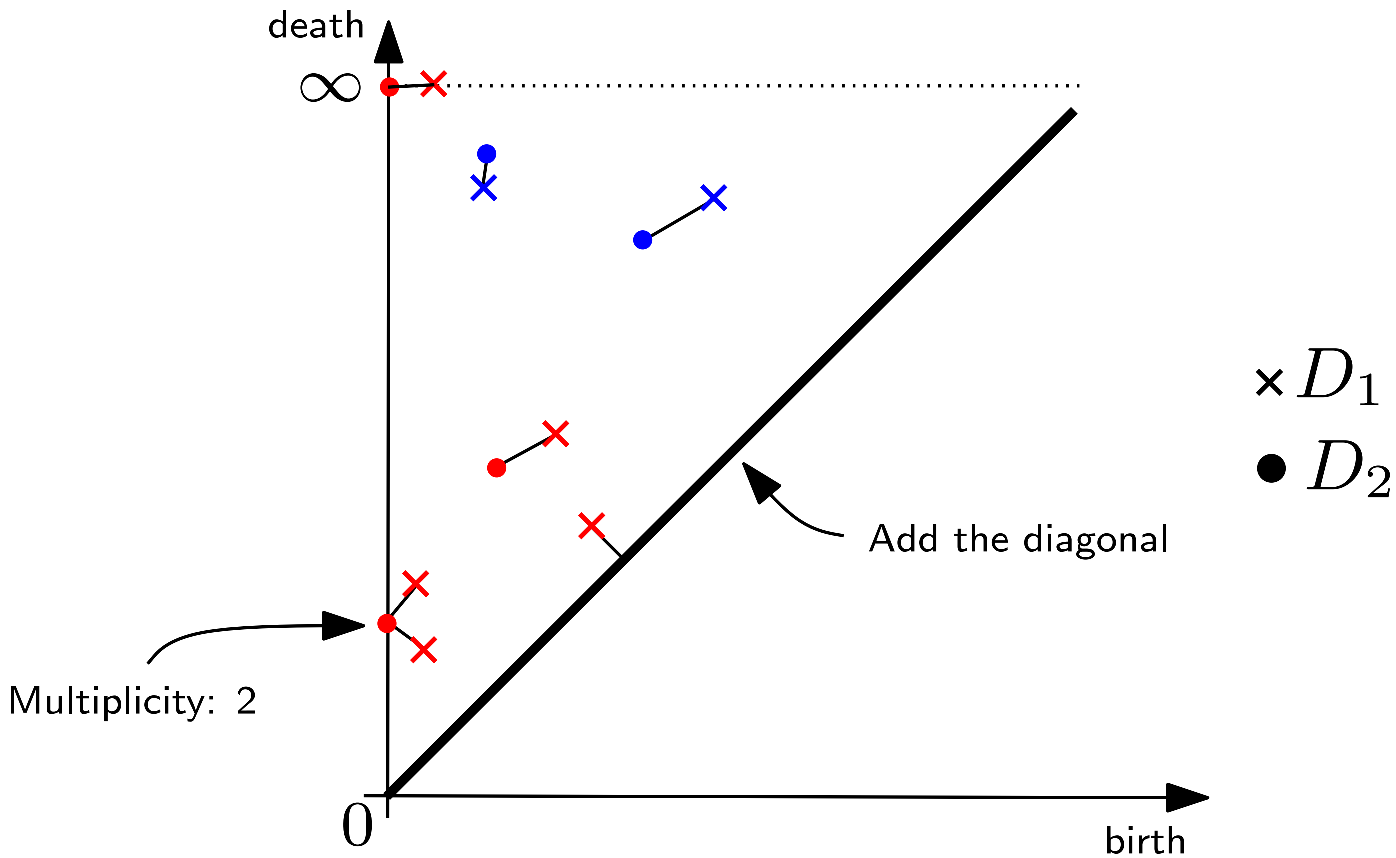
What if f is slightly perturbed?



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



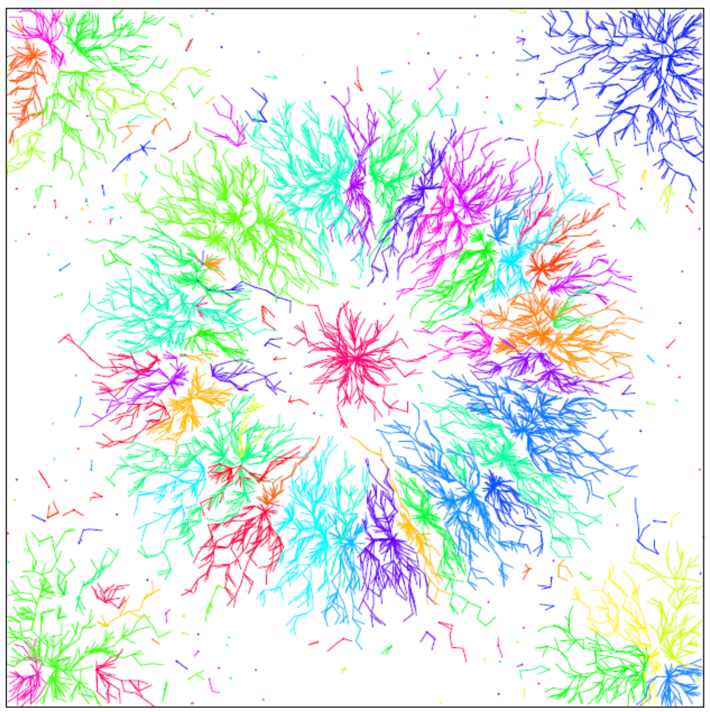
The bottleneck distance between two diagrams D_1 and D_2 is

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

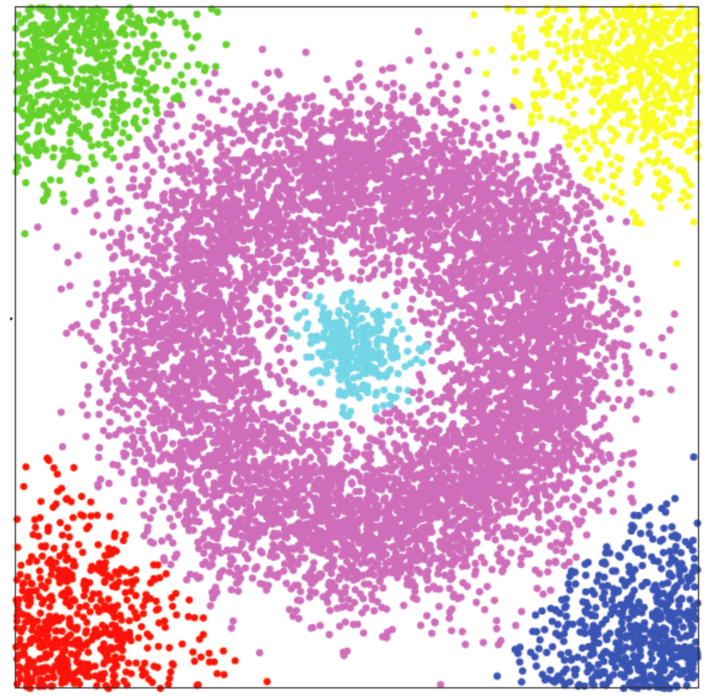
where Γ is the set of all the bijections between D_1 and D_2 and $\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|)$.

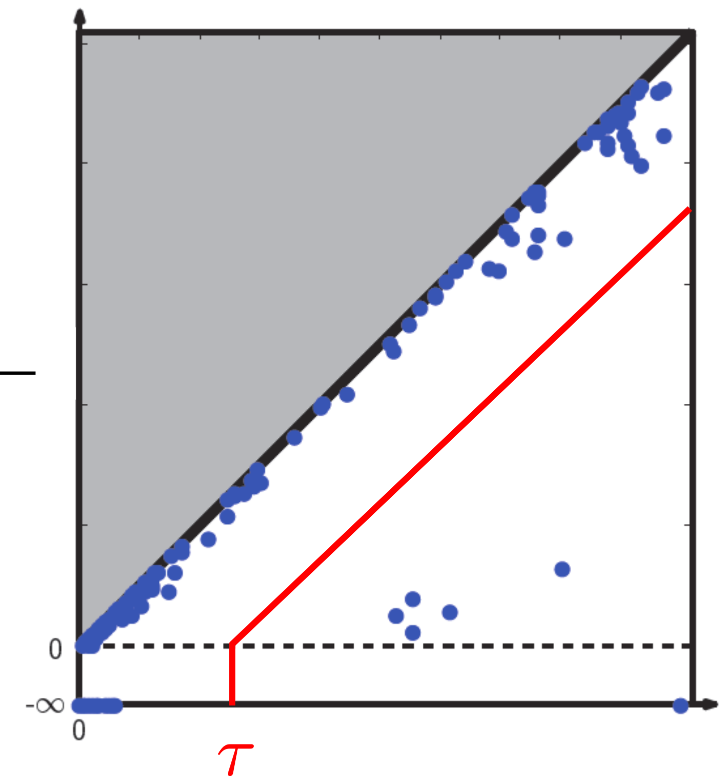
Some applications (illustrations)

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

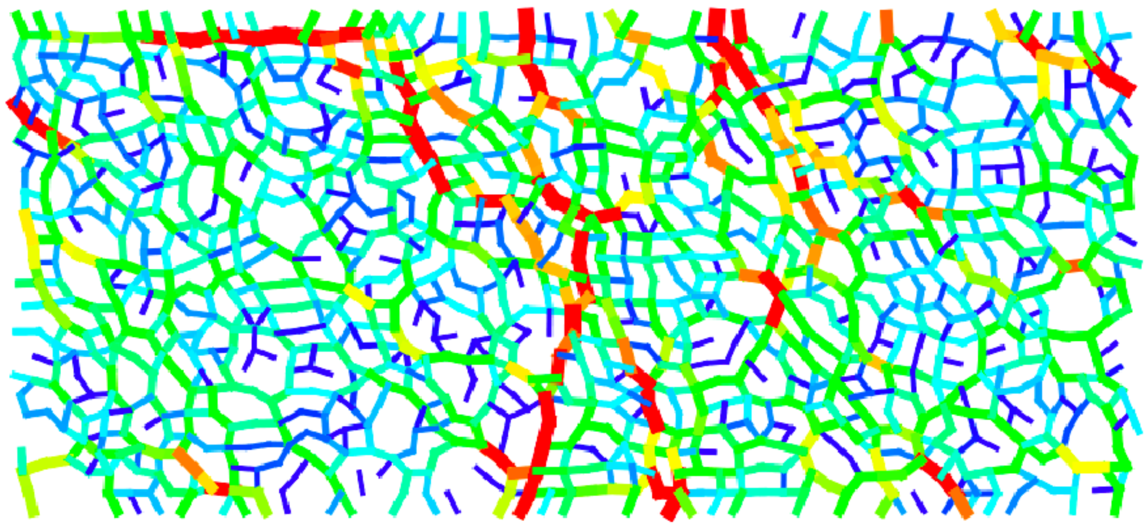


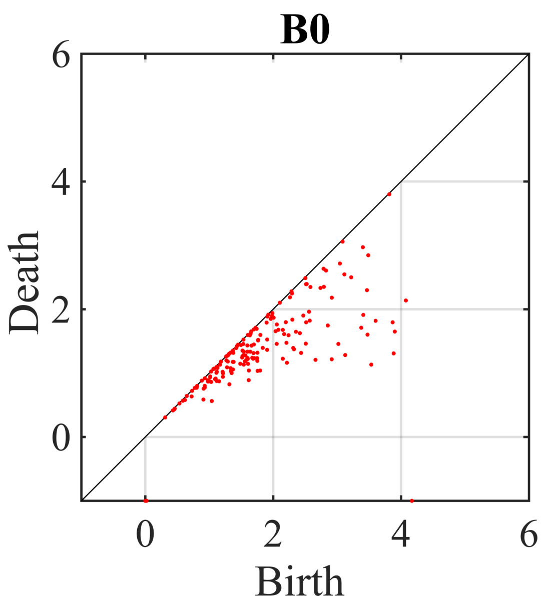
$\tau = 0$





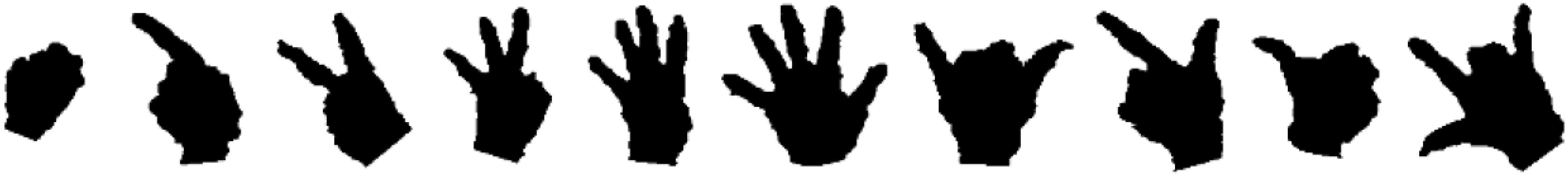
- Analysis of force fields in granular media [Kramar, Mischaikow et al]

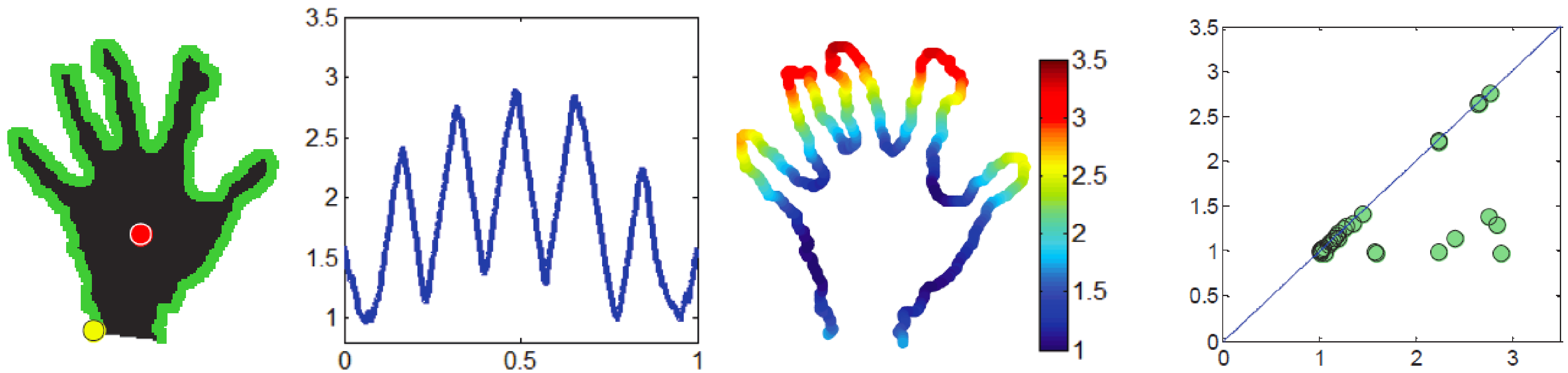




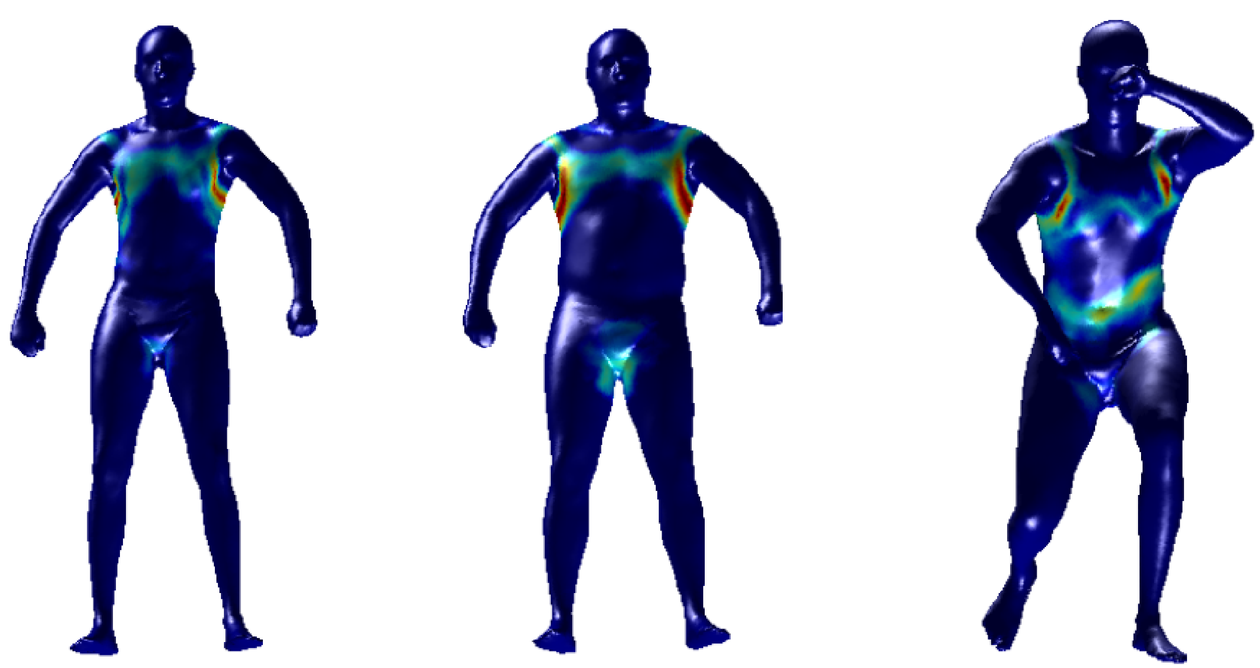
Some applications (illustrations)

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

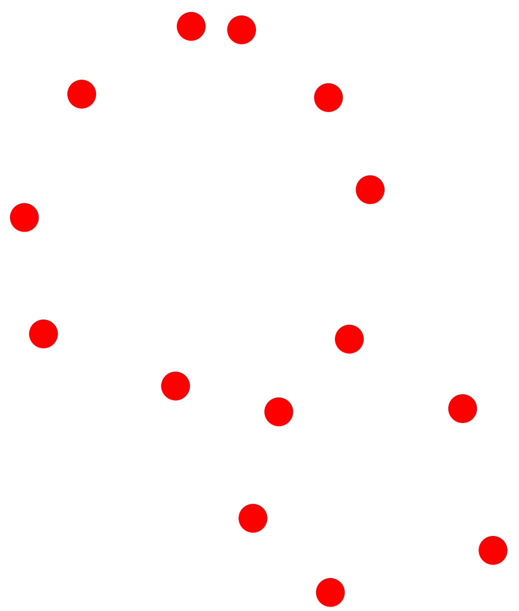




- Persistence-based pooling for shape recognition [Bonis, Ovsjanikov, Oudot, C. 2016]

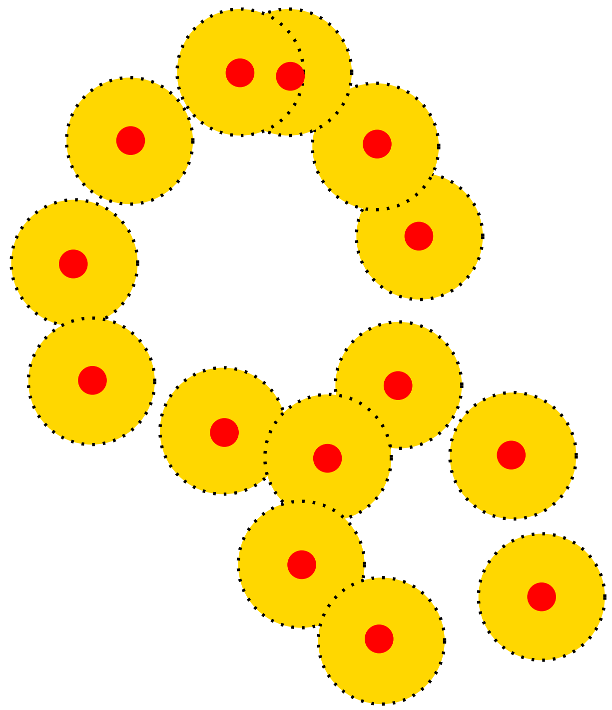


Persistent homology for point cloud data



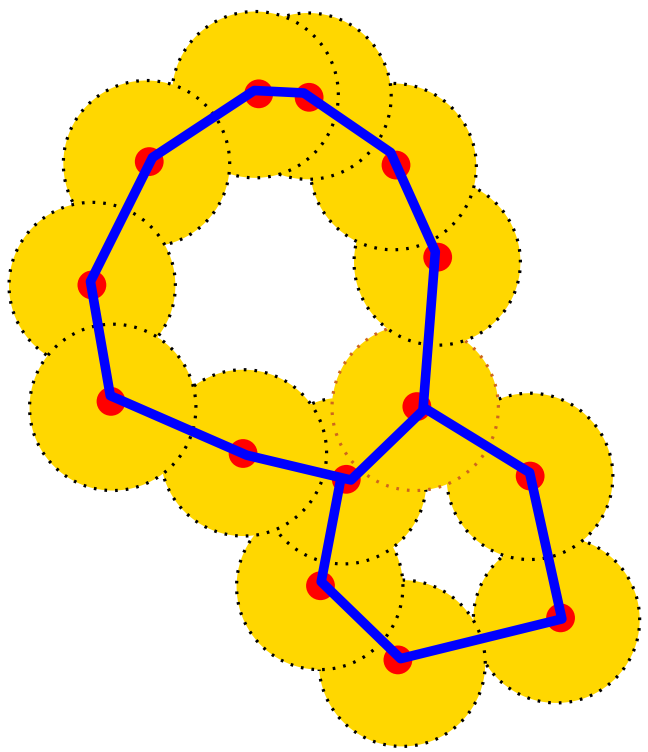
- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

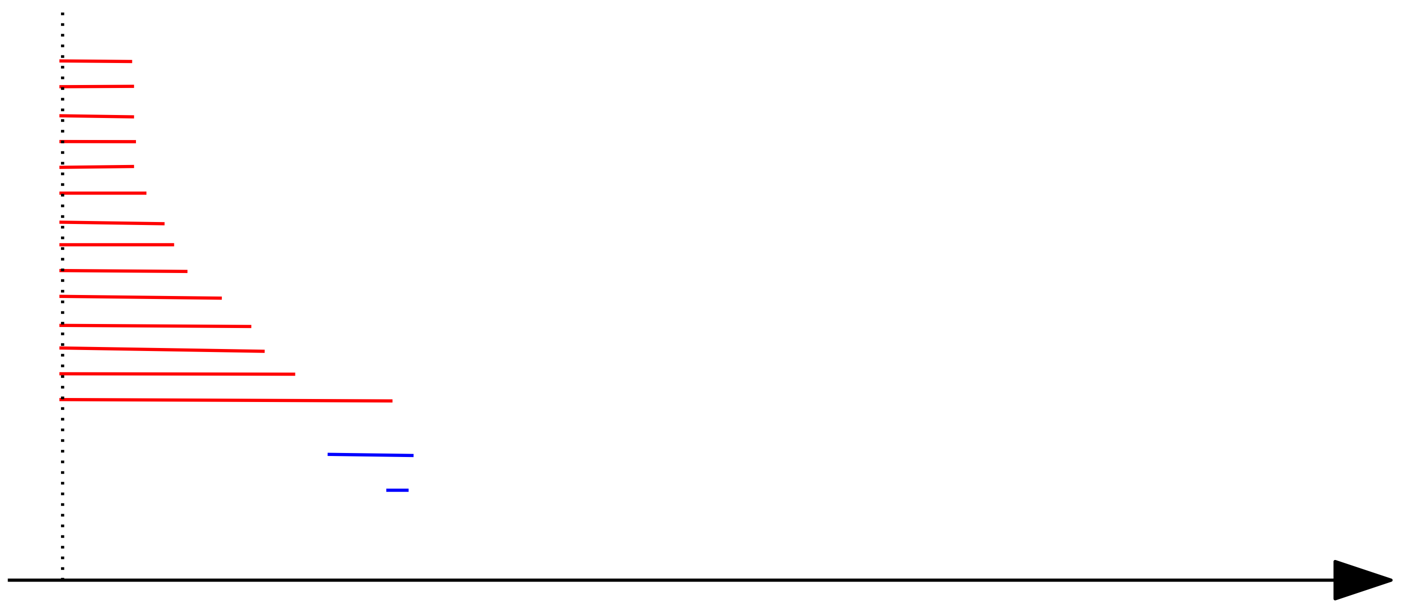
Persistent homology for point cloud data



- 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.

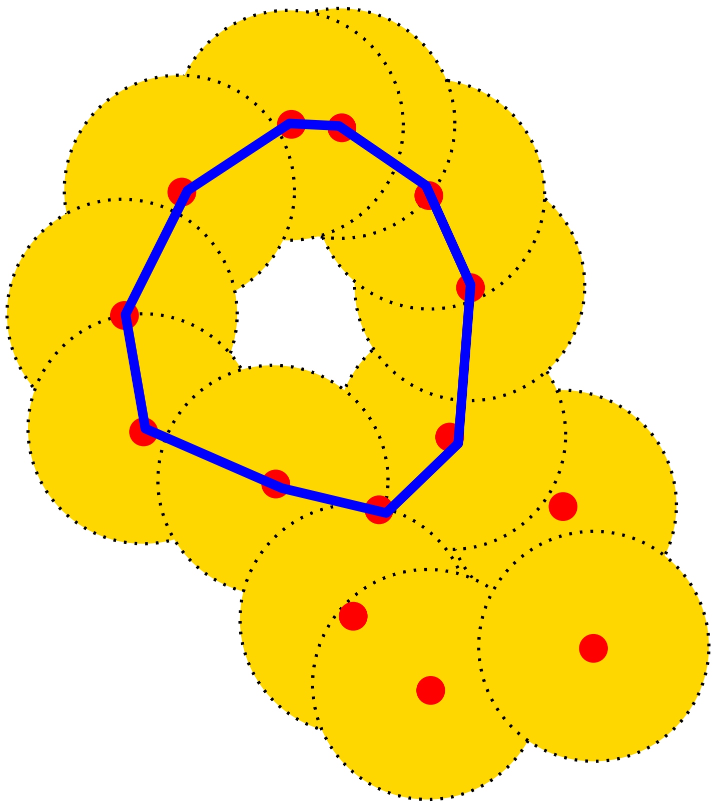
Persistent homology for point cloud data

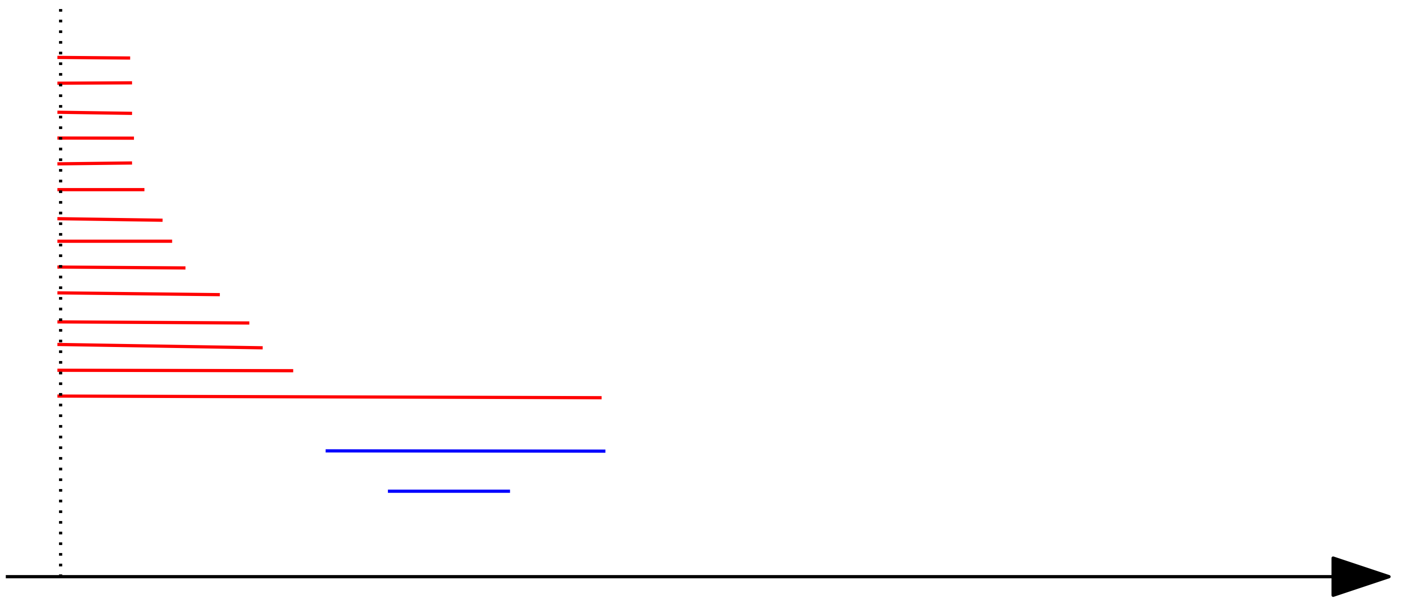




- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

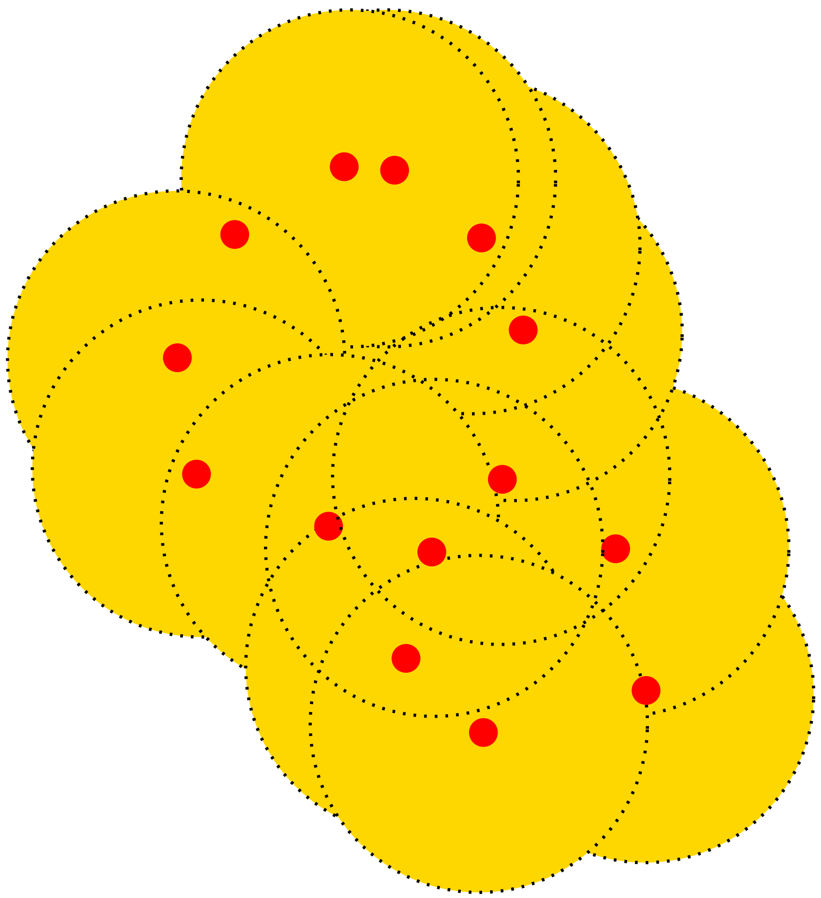
Persistent homology for point cloud data

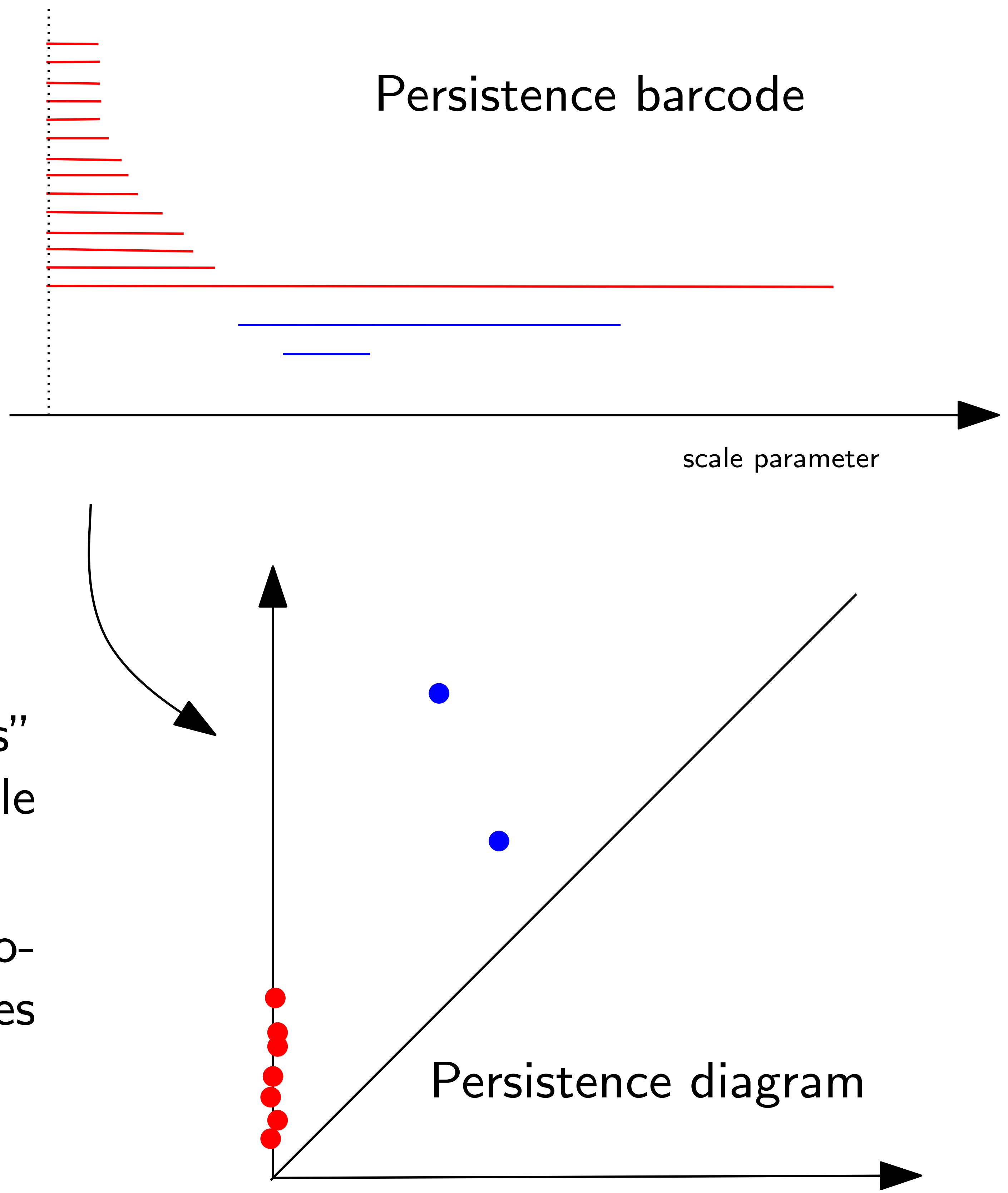




- Filtrations allow to construct "shapes" representing the data in a multiscale way.
- Persistent homology: encode the evolution of the topology across the scales $\rightarrow$ multi-scale topological signatures.

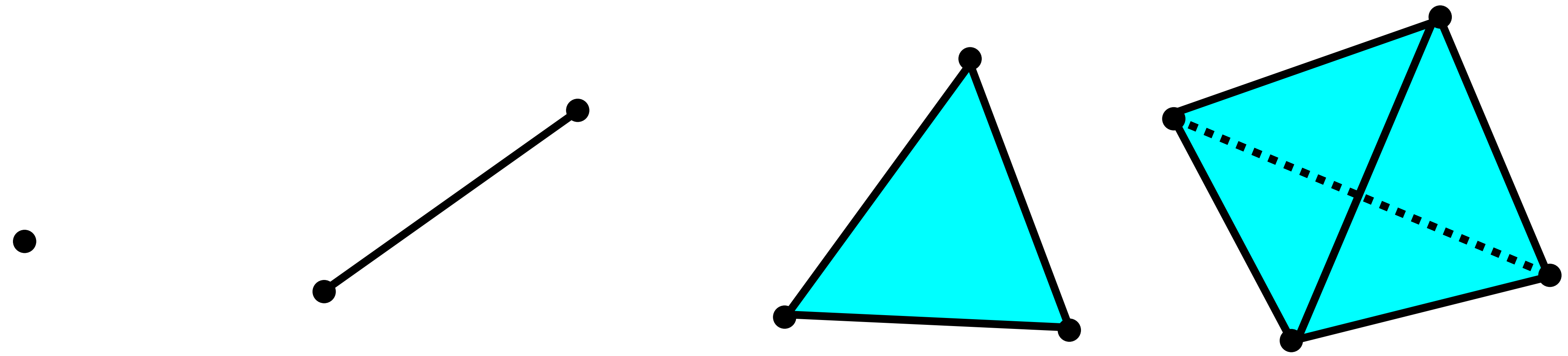
Persistent homology for point cloud data





- 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



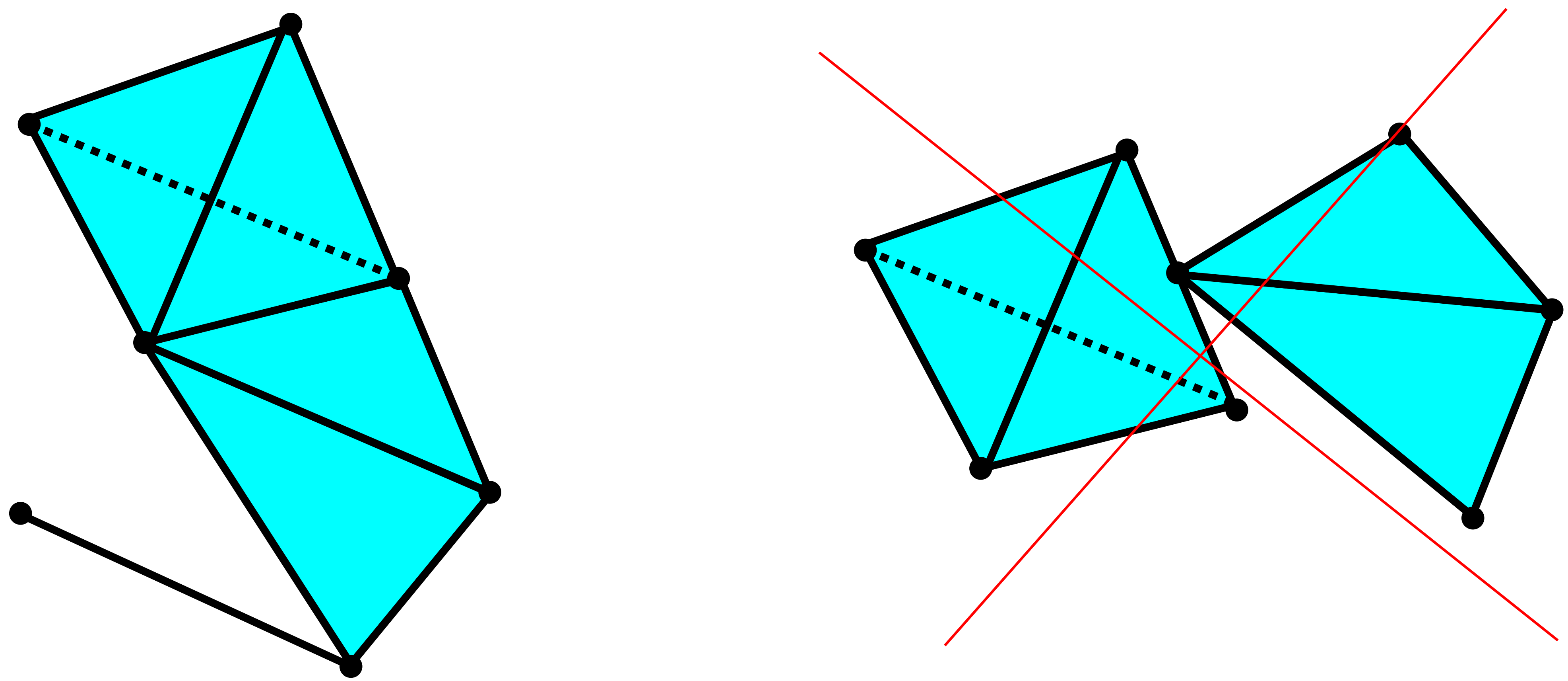
0-simplex: vertex | 1-simplex: edge | 2-simplex: triangle | 3-simplex: tetrahedron | etc...

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:

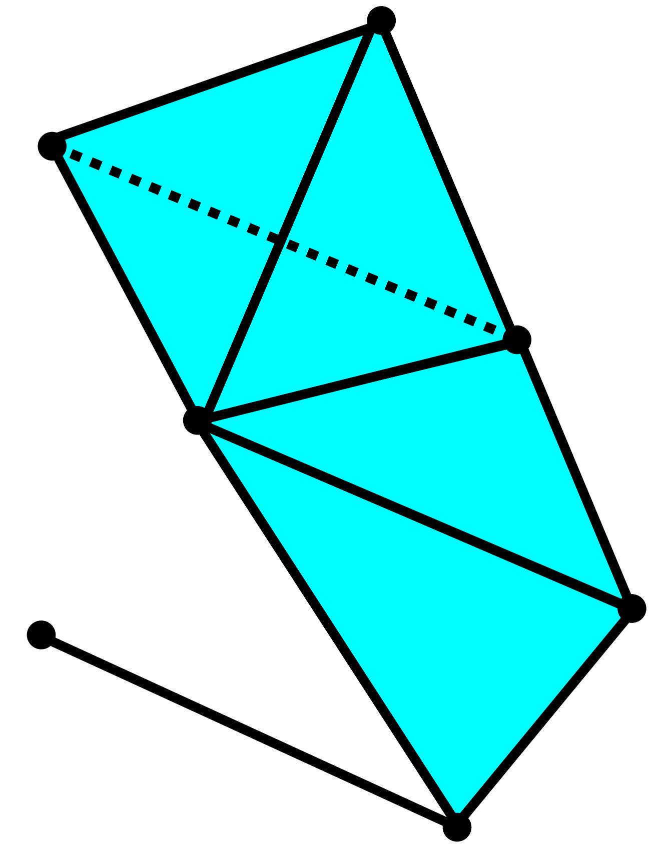
1. any face of a simplex of K is a simplex of K,
2. the intersection of any two simplices of K is either empty or a common face of both.

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:

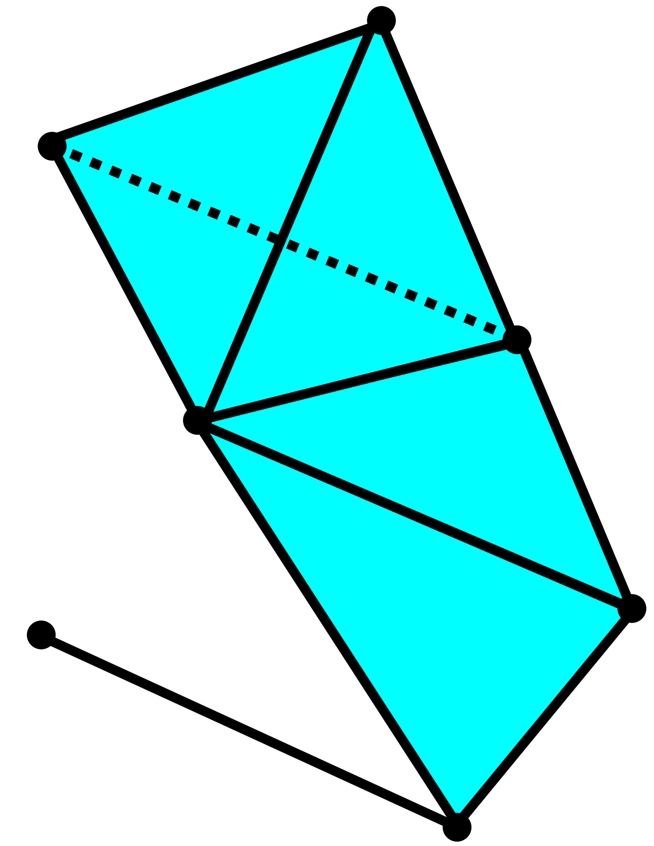
$$[e_{i_0} \cdots e_{i_k}] \in |K| \text{ iff } \{p_{i_0}, \cdots , p_{i_k}\} \in K$$

$|K|$ 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 $\sigma \in K$.

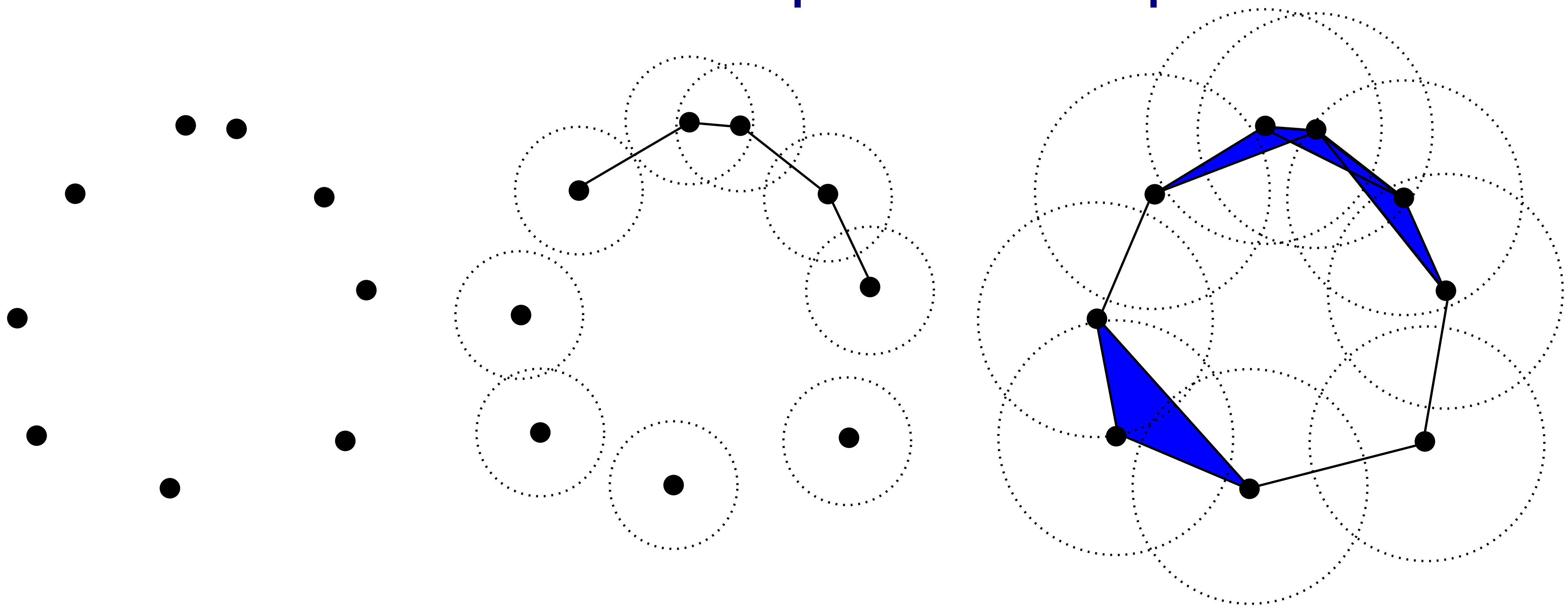


The elements of K are the simplices.

IMPORTANT

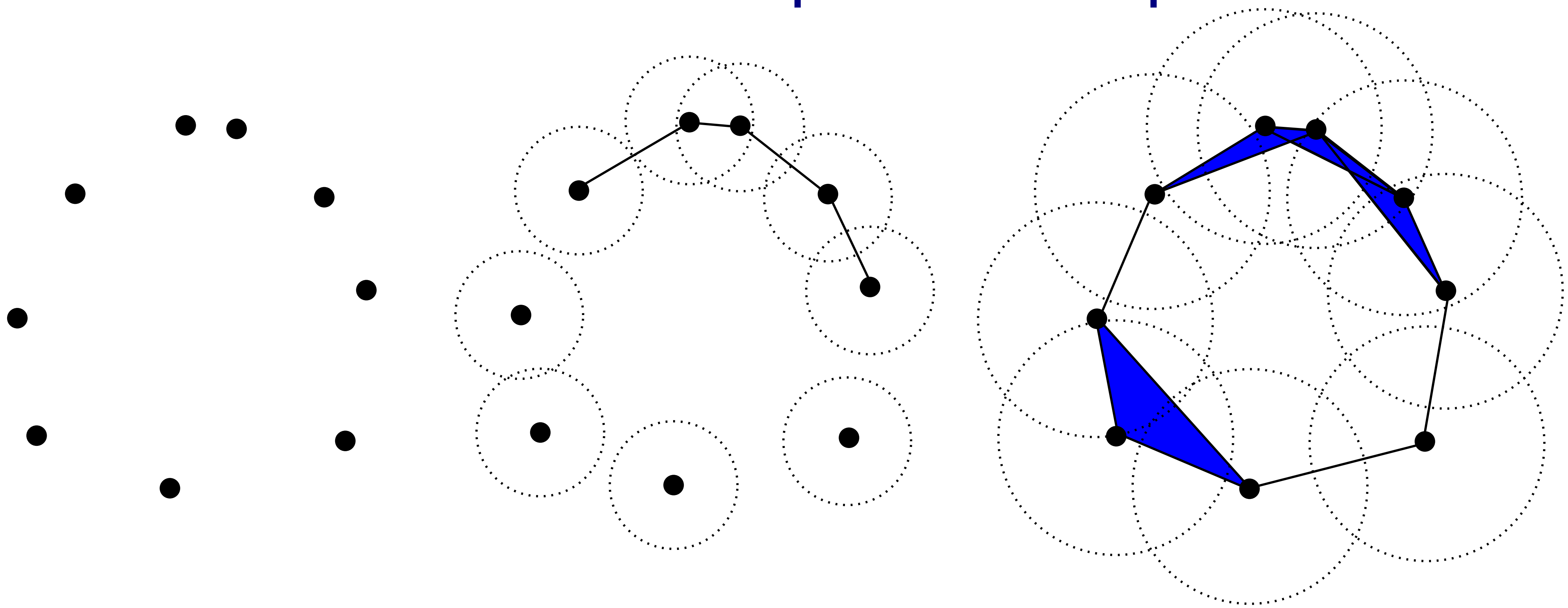
Simplicial complexes can be seen at the same time as geometric/topological spaces (good for top./geom. inference) and as combinatorial objects (abstract simplicial complexes, good for computations).

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



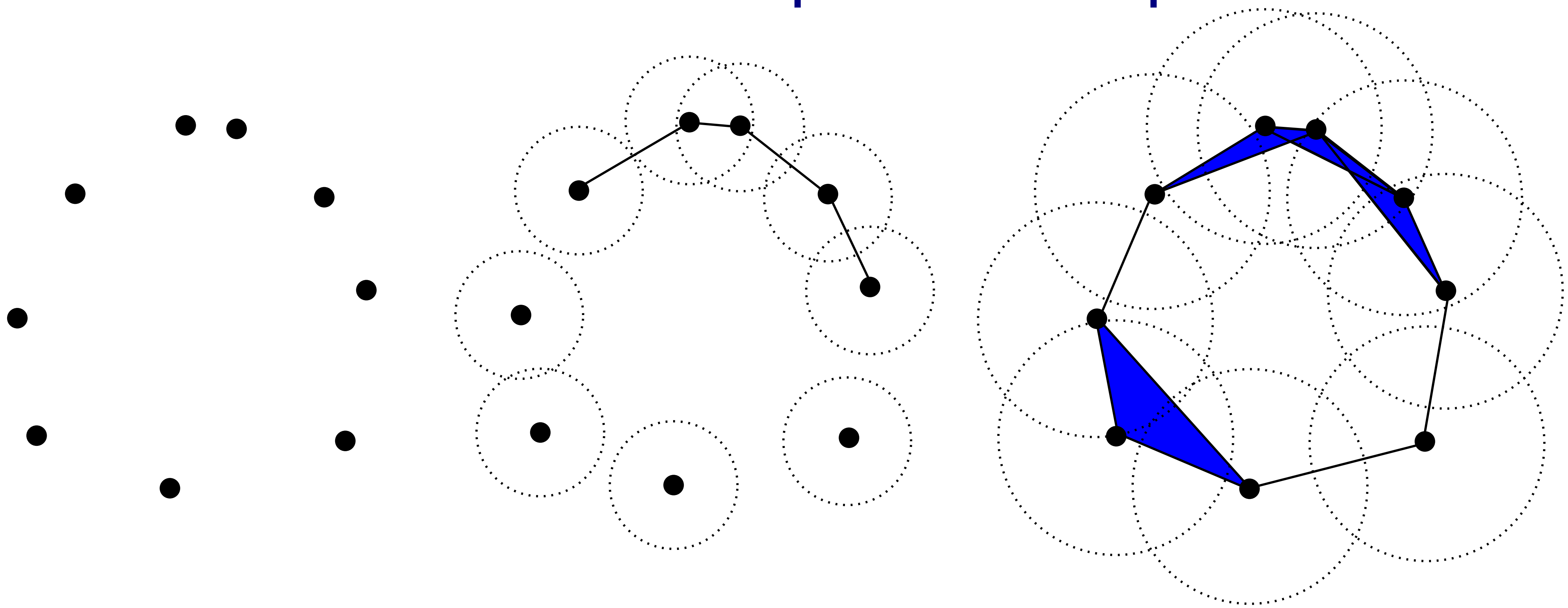
- A filtered simplicial complex (or a filtration) $\mathbb{S}$ built on top of a set $\mathbb{X}$ is a family $(\mathbb{S}_a \mid a \in \mathbf{R})$ of subcomplexes of some fixed simplicial complex $\overline{\mathbb{S}}$ with vertex set X s. t. $\mathbb{S}_a \subseteq \mathbb{S}_b$ for any $a \leq b$.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

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

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

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

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

Bottleneck distance

Gromov-Hausdorff distance

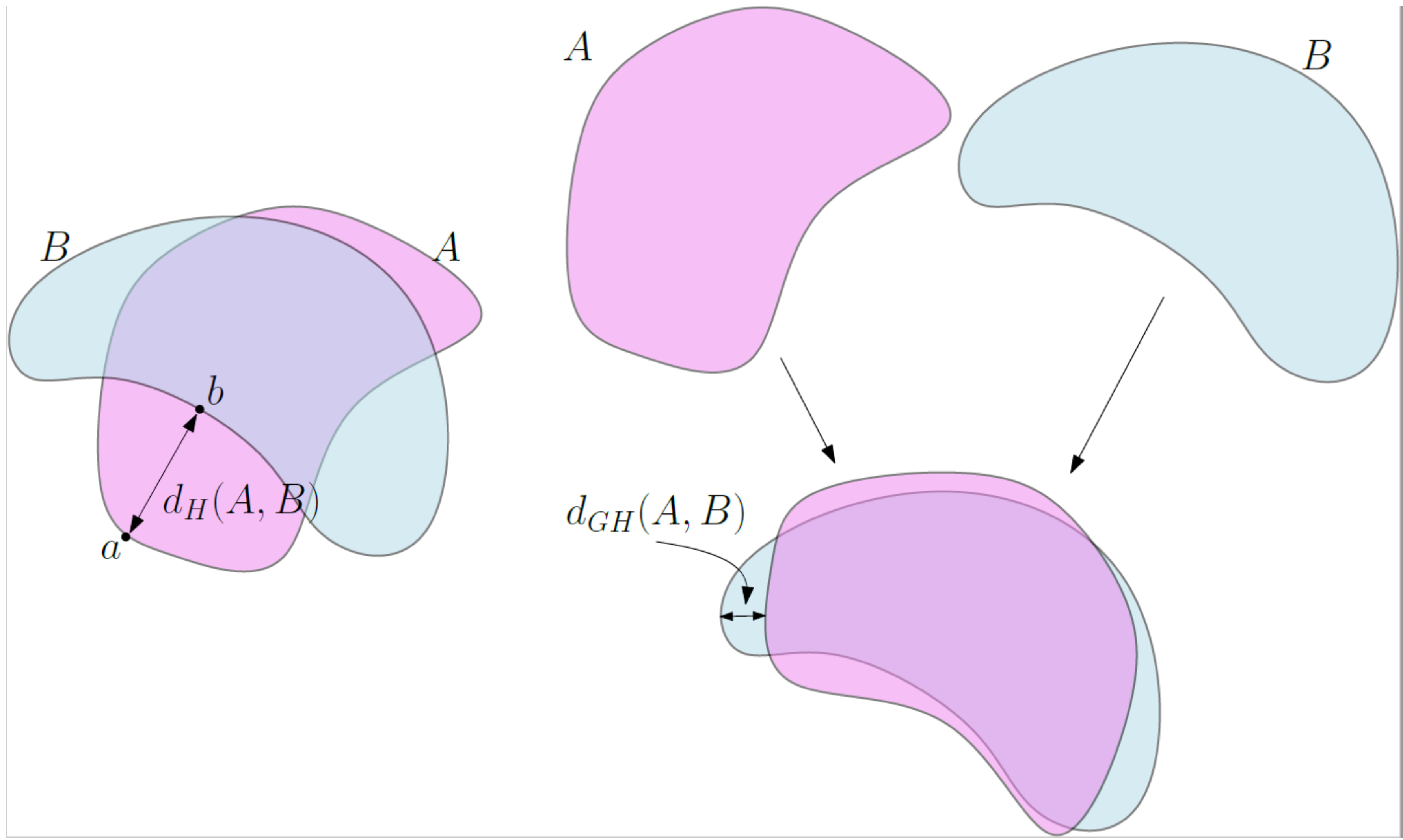
$$d_{GH}(\mathbb{X}, \mathbb{Y}) := \inf_{\mathbb{Z}, \gamma_1, \gamma_2} d_H(\gamma_1(\mathbb{X}), \gamma_2(\mathbb{X}))$$

$\mathbb{Z}$ 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



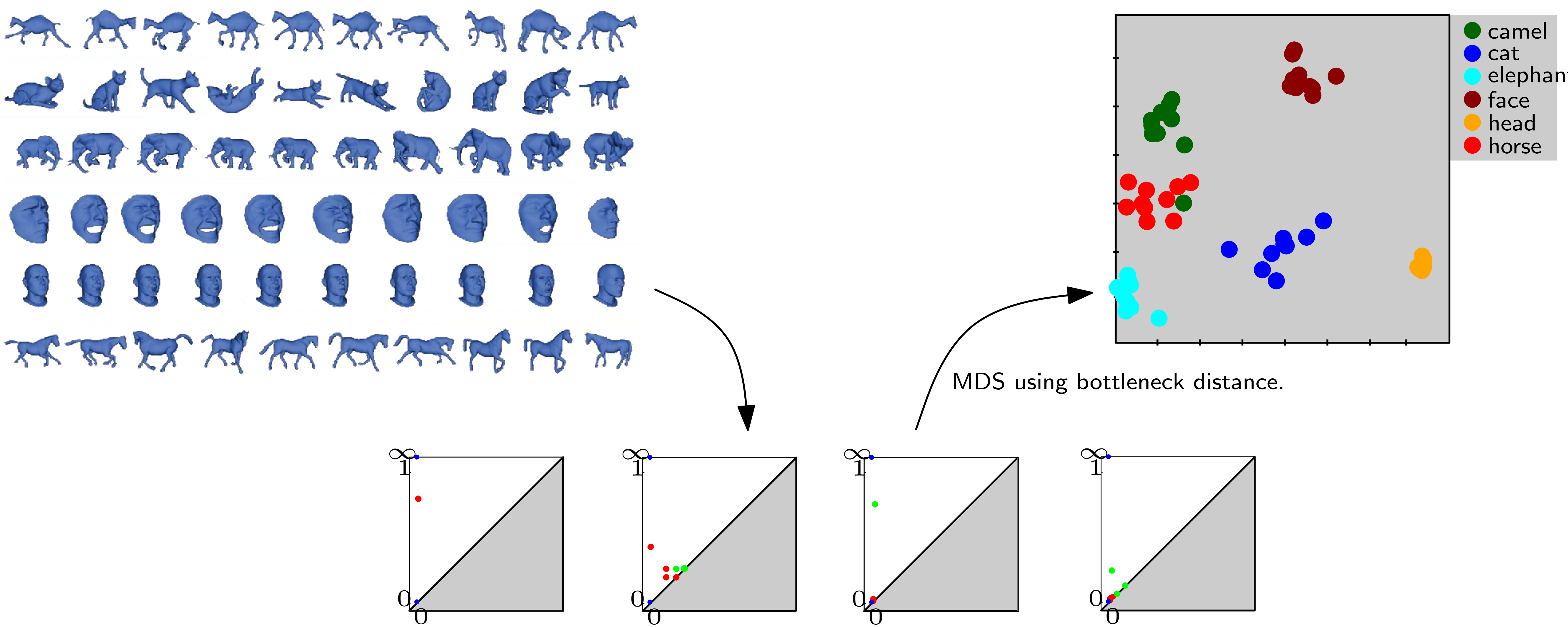
Let $A, B \subset M$ be two compact subsets of a metric space (M, d)

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

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

Application: non rigid shape classification

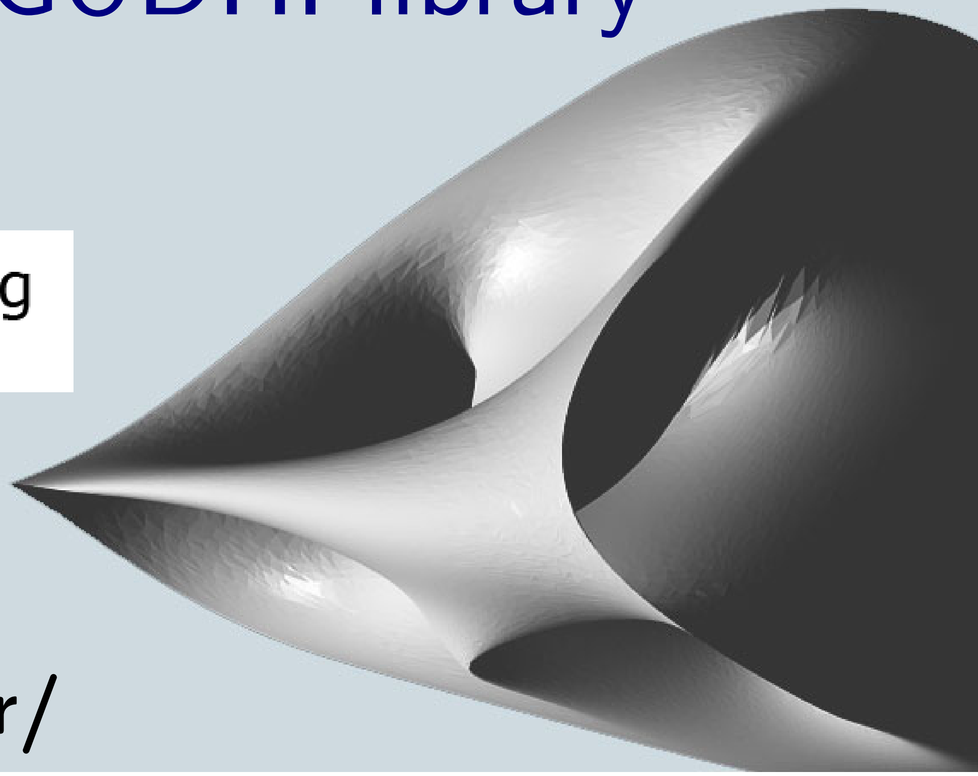
[C., Cohen-Steiner, Guibas, Mémoli, Oudot - SGP '09]



- 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

गुढी GUDHI Geometry Understanding in Higher Dimensions



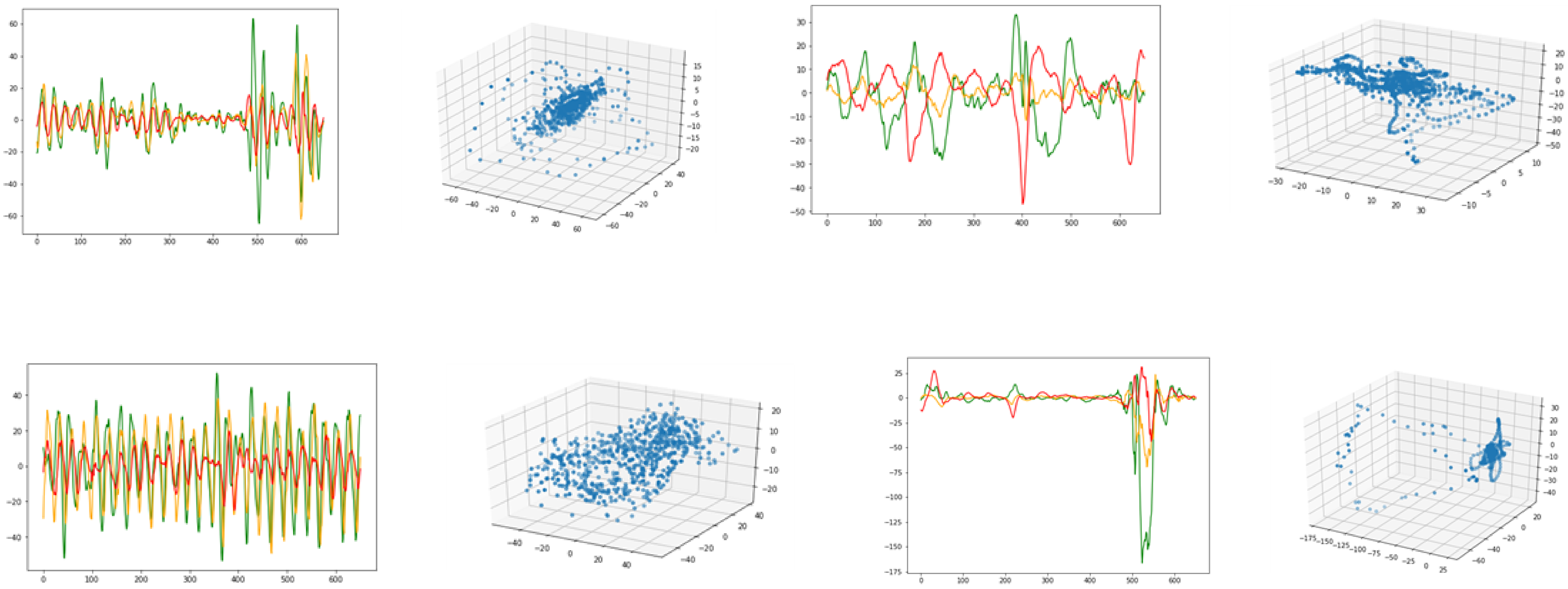
http://gudhi.gforge.inria.fr/

GUDHI :

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

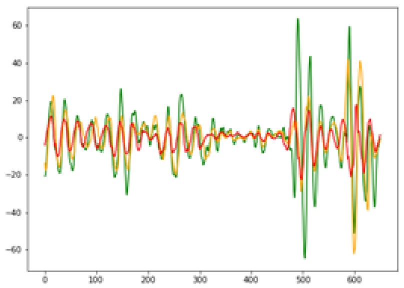
TDA and Machine Learning: some illustrative examples on real applications

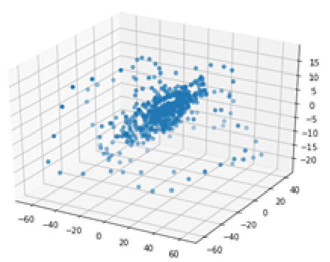
TDA and Machine Learning for sensor data



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

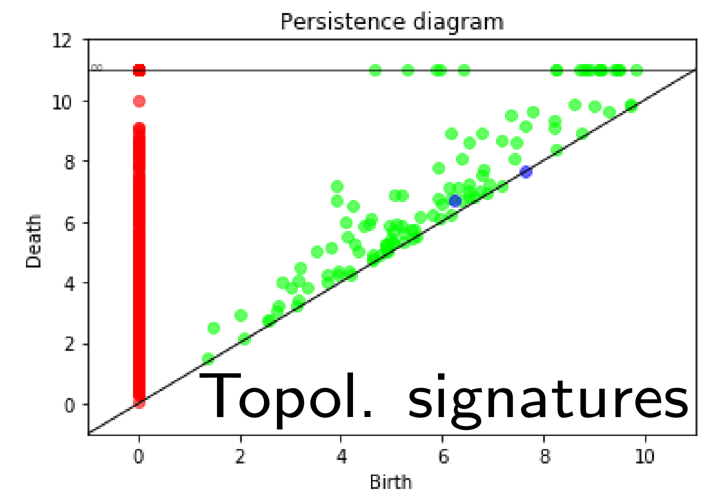
TDA and Machine Learning for sensor data





TDA pipeline

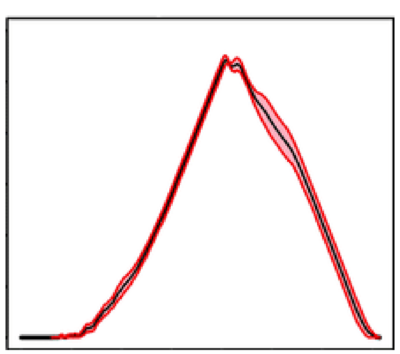
GUDHI software



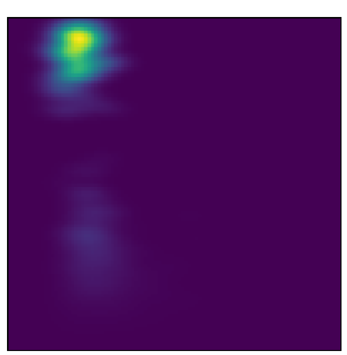
Topol. signatures

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

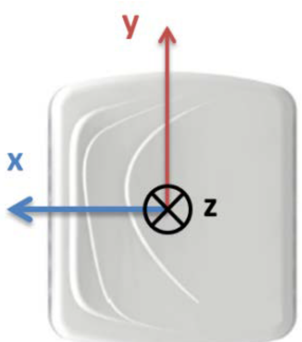
A joint industrial research project between

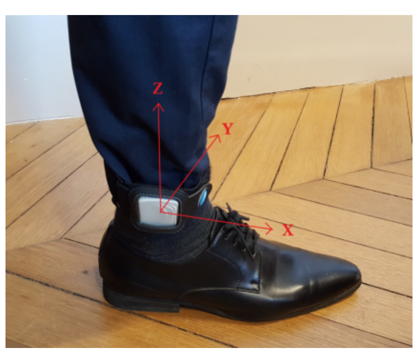
sysnav
NAVIGATION TECHNOLOGIES

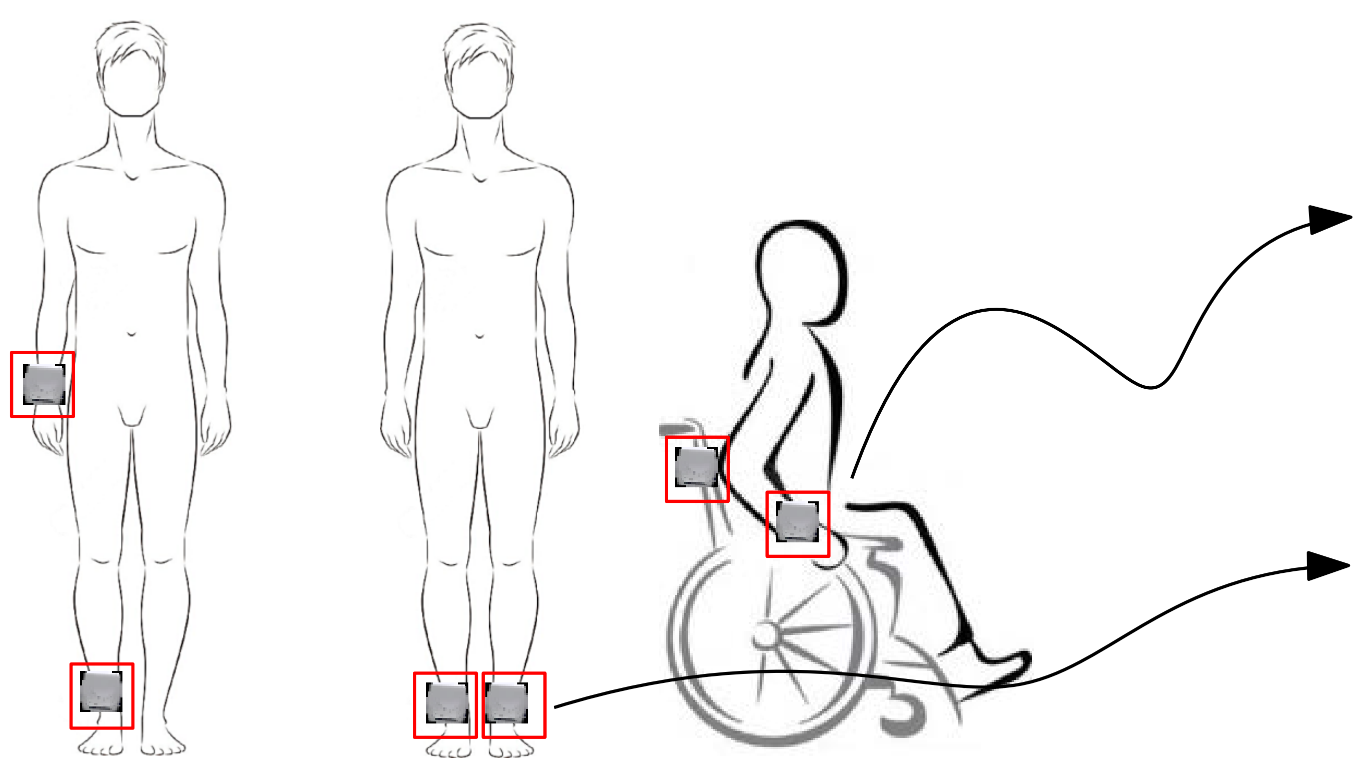
and

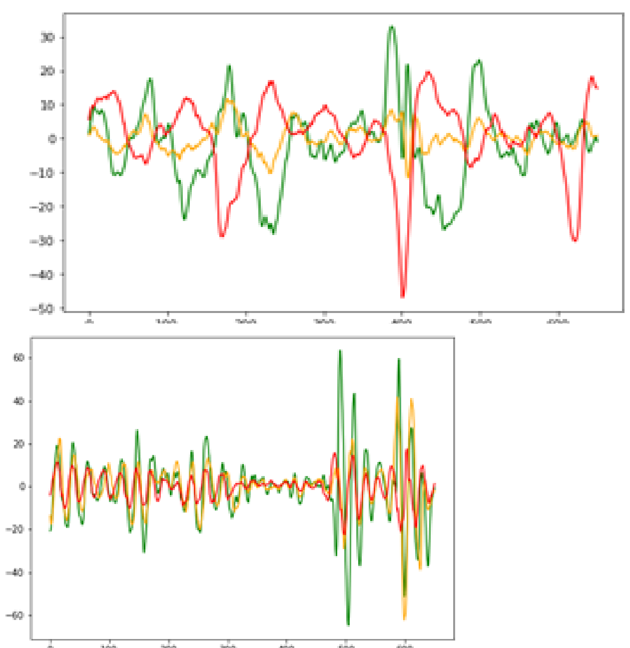
Inria informatics mathematics

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









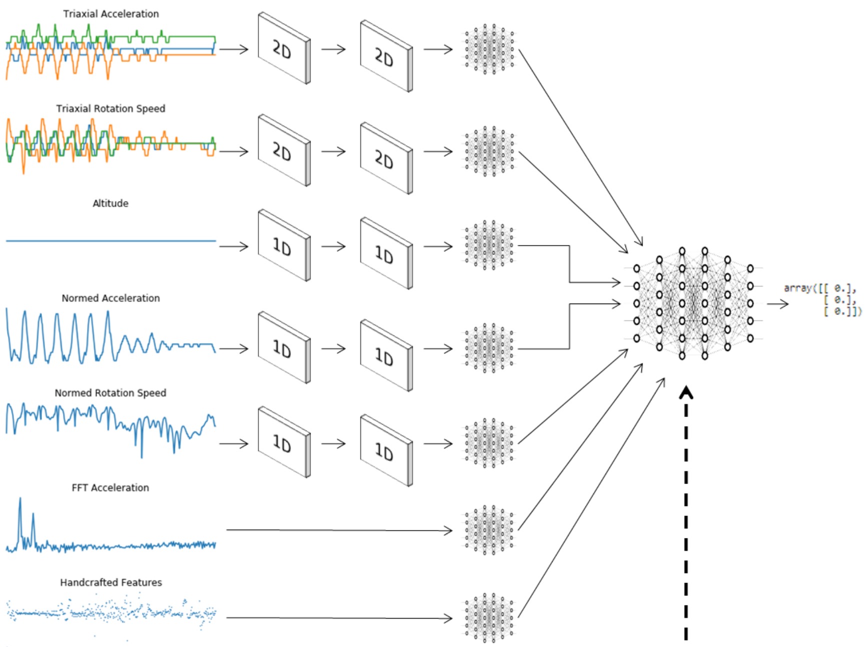
“Chaotic” time-dependent data

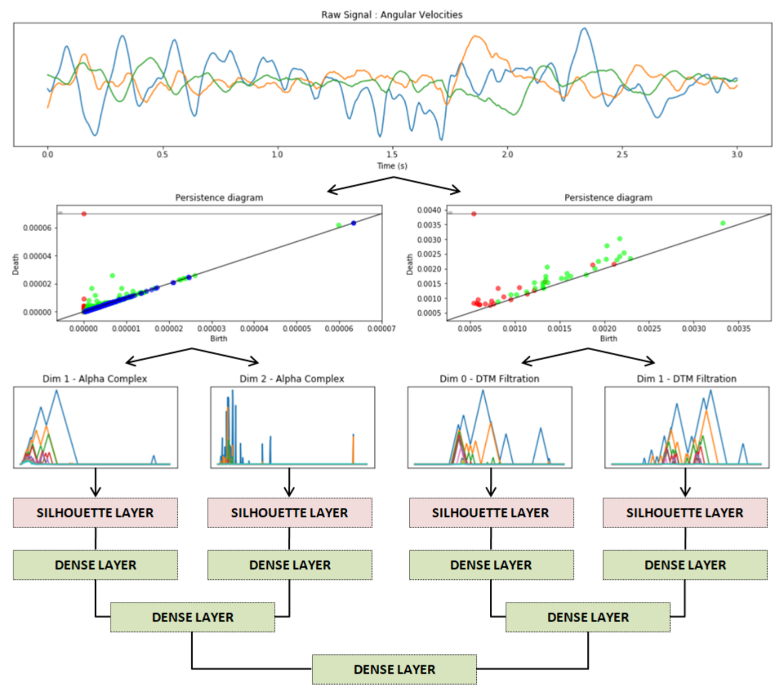
Objective: precise analysis of movements and activities of pedestrians.

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.

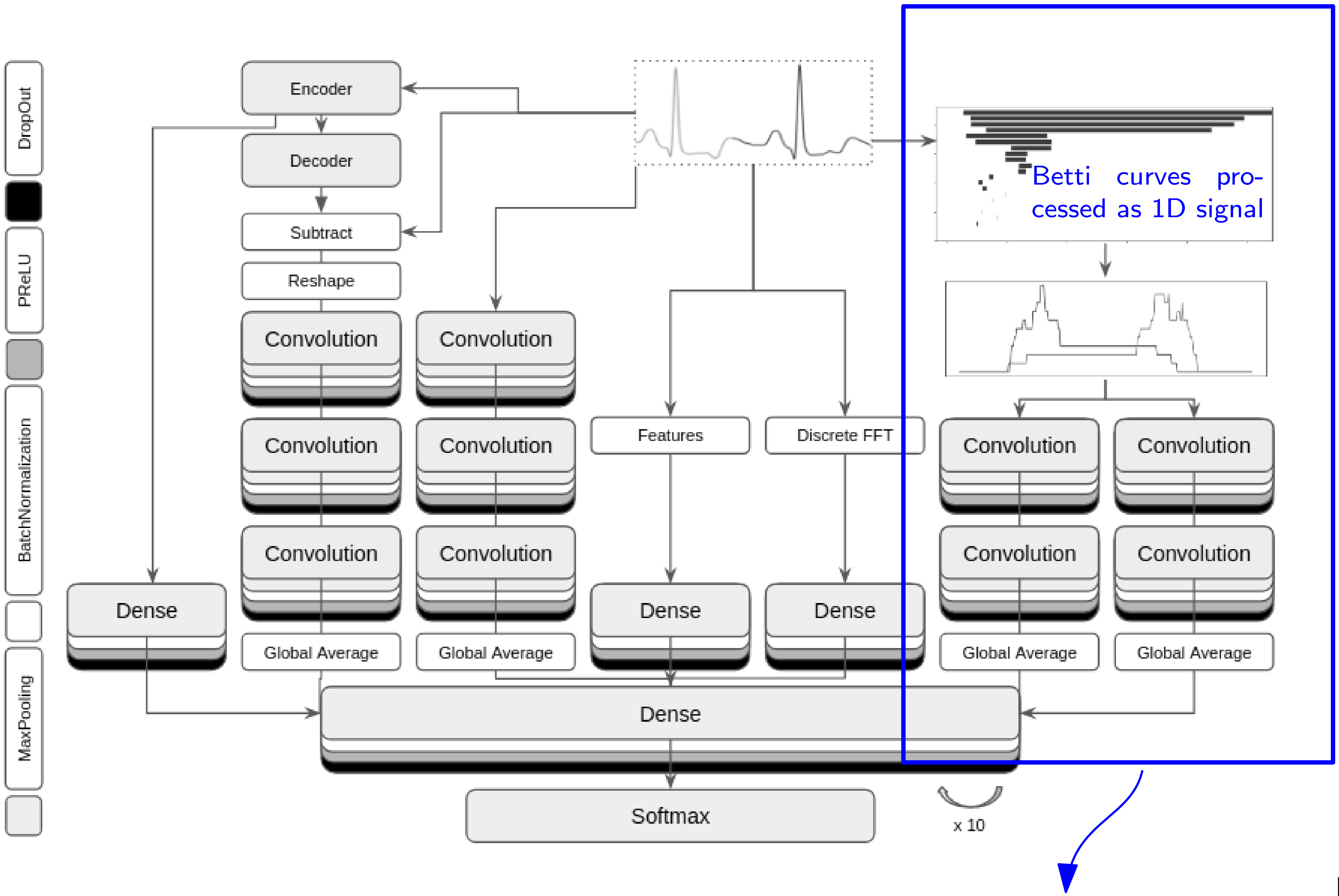
- Data collected in non controlled environments (home) are very chaotic.
- Data registration (uncertainty in sensors orientation/position).
- Reliable and robust information is mandatory.
- Events of interest are often rare and difficult to characterize.

sysnav NAVIGATION TECHNOLOGIES

Inria informatics mathematics

TDA-DL pipeline for arrhythmia detection

Objective: Arrythmia detection from ECG data.



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

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

FUJITSU Inria informatics mathematics

Thank you for your attention!