Data-Rich Phenomena
Modelling, Analysing & Simulation Using Partial Differential Equations
14-16 December 2015, Cambridge

Statistics and Topological Data Analysis

Bertrand MICHEL
LSTA UPMC - GEOMETRICA INRIA Saclay

Introduction

Topological data analysis and topological inference

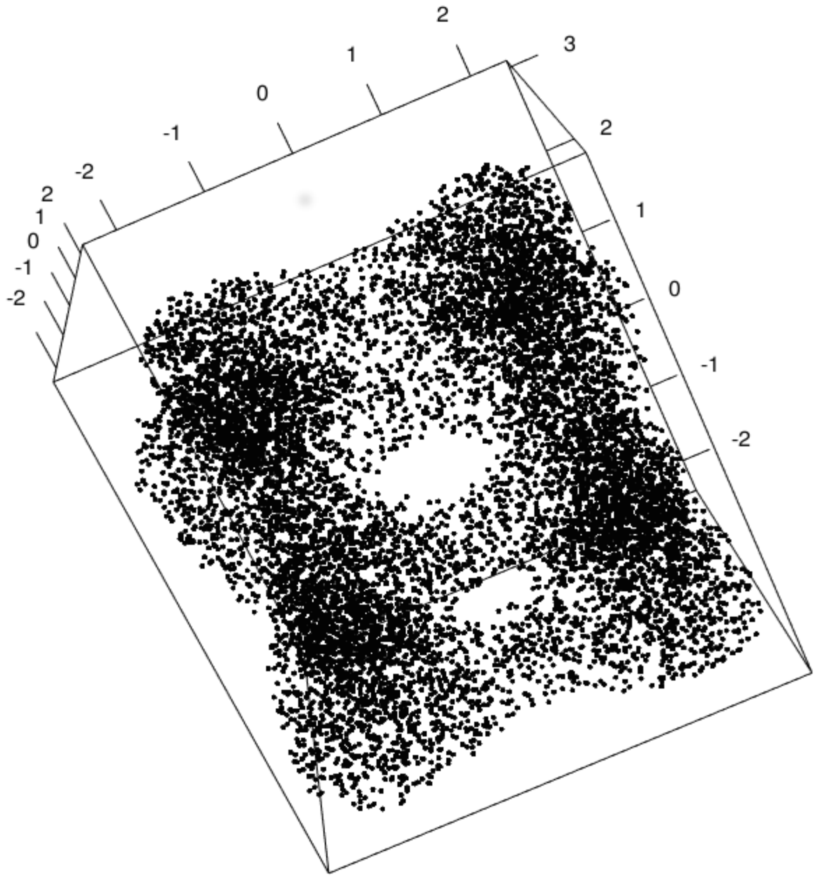
- **Geometric inference** and **algebraic topology tools**, **computational topology** has recently witnessed important developments with regards to data analysis, giving birth to the field of **topological data analysis** (TDA).

- The aim of TDA is to infer relevant, qualitative and quantitative **topological structures** (clusters, holes ...) directly from the data.

- The two popular methods in TDA : **Mapper algorithm** [Singh et al., 2007] and **persistent homology** [Edelsbrunner et al., 2002].

- TDA methods relies on **Topological Inference** methods / results.

- **Topological inference** methods aim to infer topological properties of an unknown topological space $\mathbb{X}$, typically from a point cloud $\mathbb{X}_n$ "close" to $\mathbb{X}$.



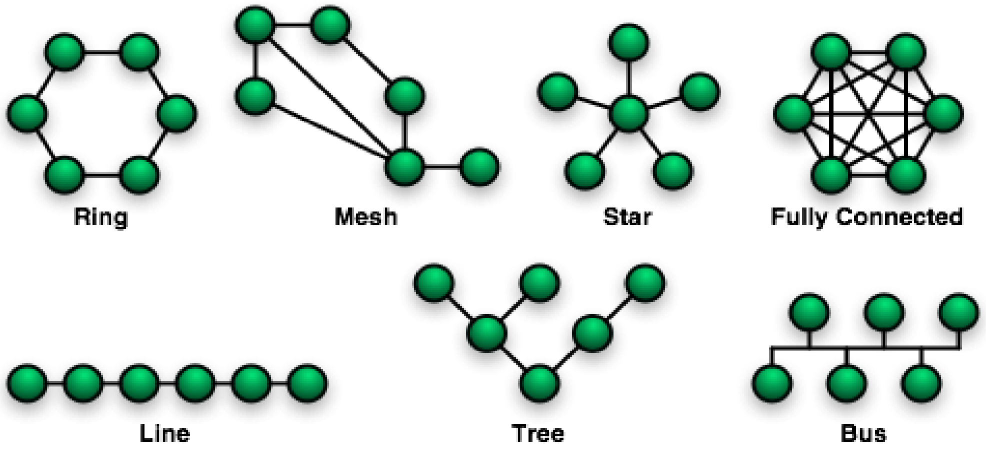
Properties of Topological Methods for Data Analysis

From [Carlson, 2013]: *The point of view on the study of shape which is particular to topology can be described in terms of three points.*

1. *The properties of shape studied by topology are independent of any particular coordinate representation of the shape in question, and instead depends only on the pairwise distances between the points making up the shape.*

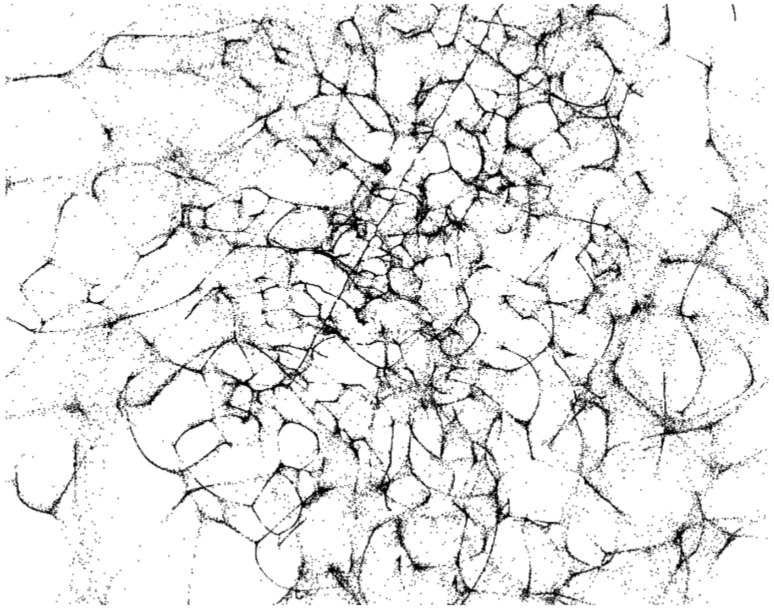
2. *Topological properties of shape are deformation invariant, i.e. they do not change if the shape is stretched or compressed.*

3. *Topology constructs compressed representations of shapes, which retain many interesting and useful qualitative features while ignoring some fine detail.*

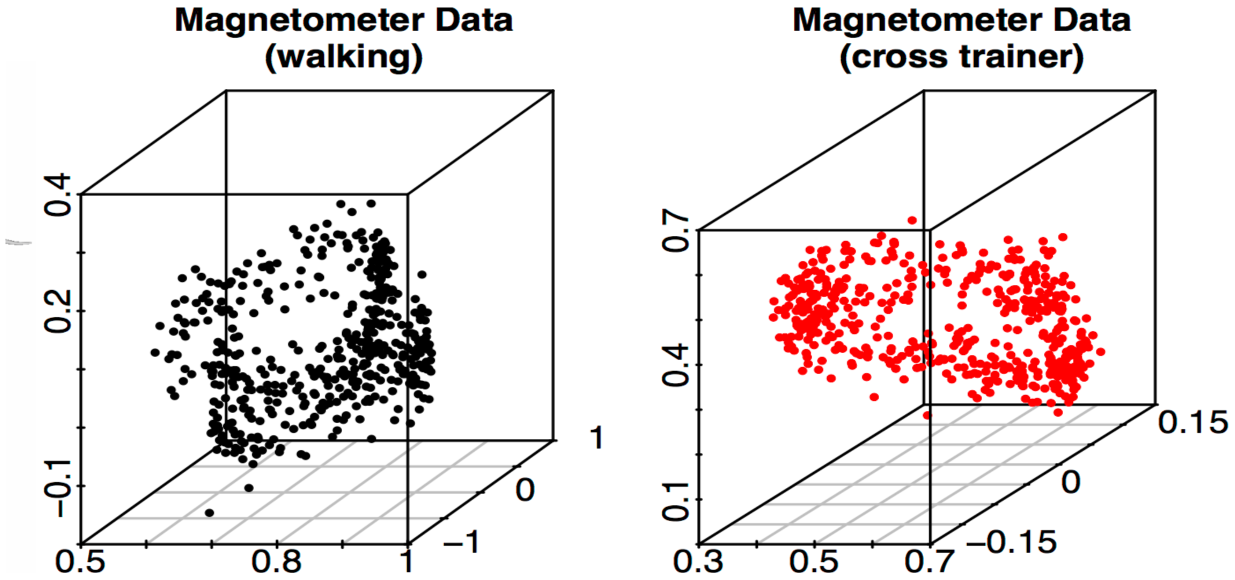


Application fields of TDA methods

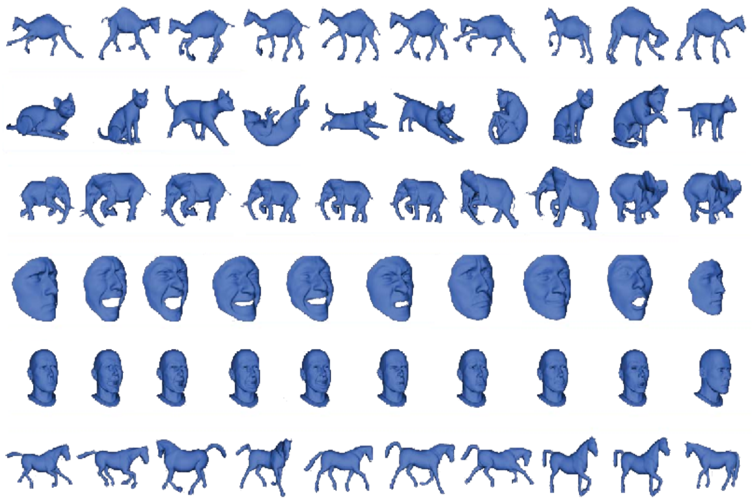
[distribution of galaxies]



[Magnetometer Data]

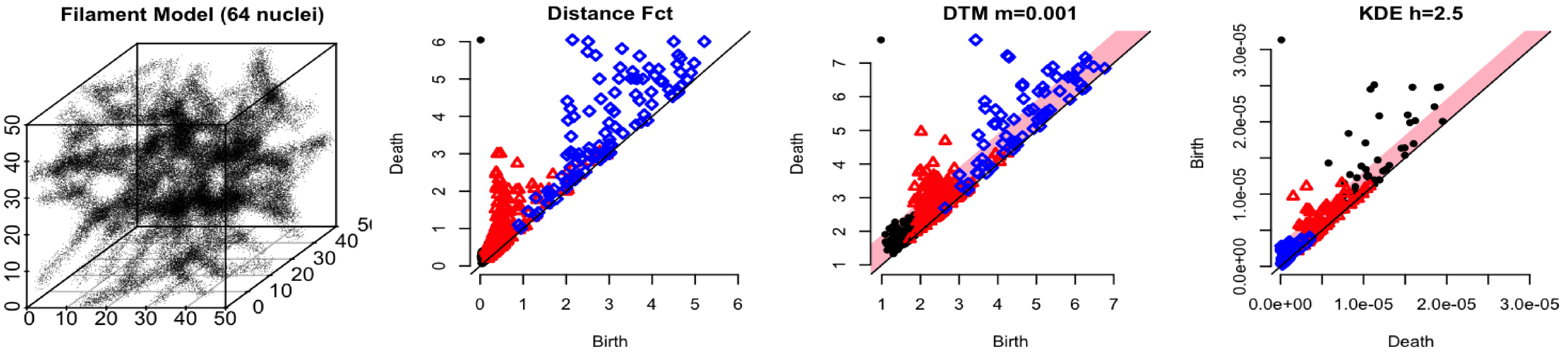


[3D shape database]



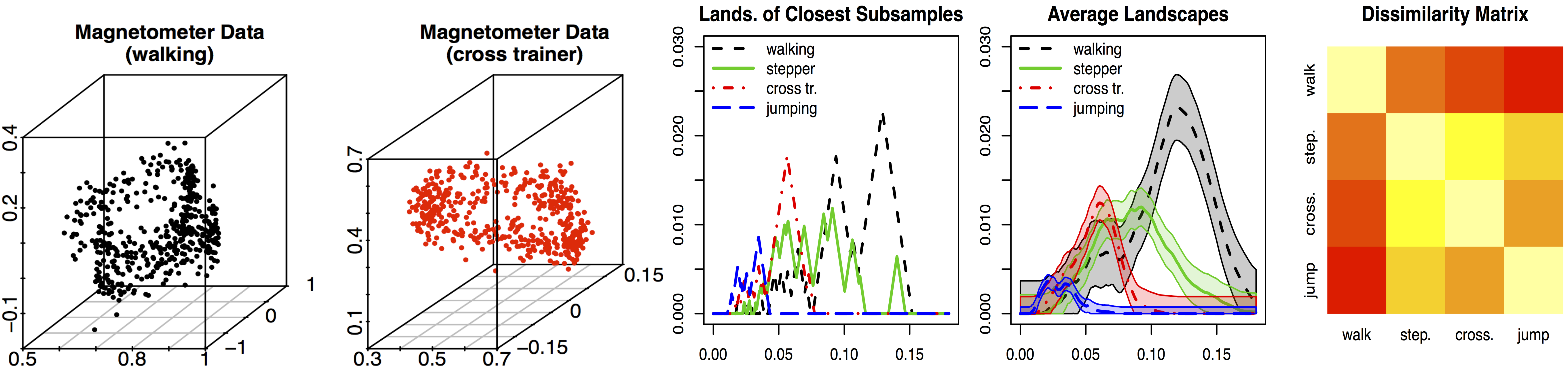
Topological data analysis methods can be used:

- For **exploratory analysis**, visualization:



[Chazal et al., 2014a]

- For **feature extraction** in supervised settings (prediction) :



[Chazal et al., 2015a]

Statistics and TDA

Until very recently, TDA and topological inference mostly relied on deterministic approaches. Alternatively, a *statistical approach to TDA* means that :

- we consider data as generated from an unknown distribution
- the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object.

Non exhaustive list of questions for a statistical approach to TDA :

- proving consistency of TDA
  methods.
- providing confidence regions for topological features and discussing the significance of the estimated topological quantities.
- selecting relevant scales at which the topological phenomenon should be considered.
- dealing with outliers and providing robust methods for TDA.

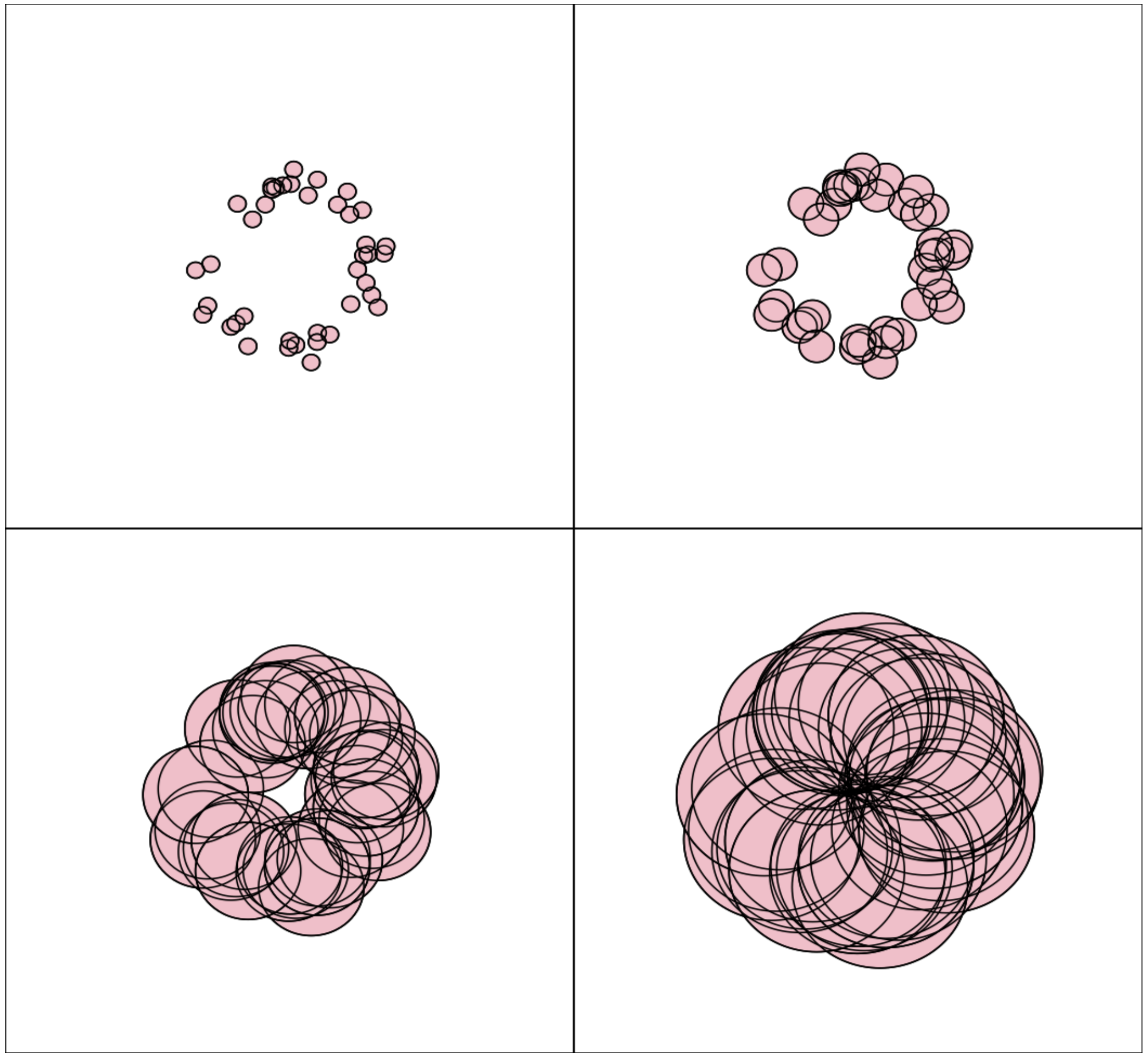
Homology
and
Persistent homology

Approximating models for TDA : Offsets and Simplicial Complexes

Point clouds in themselves do not carry any non trivial topological or geometric structure.

For a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ (or in a metric space),
the r-offset of $\mathbb{X}_n$ is defined by

$$\mathbb{X}_n^r = \bigcup_{x \in \mathbb{X}_n} B(x, r).$$



More generally, for any compact set $\mathbb{X}$ (in $\mathbb{R}^d$),

$$\mathbb{X}^r := \bigcup_{x \in \mathbb{X}} B(x, r) = d_{\mathbb{X}}^{-1}([0, r])$$

where the distance function $d_{\mathbb{X}}$ to $\mathbb{X}$ is

$$d_{\mathbb{X}}(y) = \inf_{x \in \mathbb{X}} \|x - y\|.$$

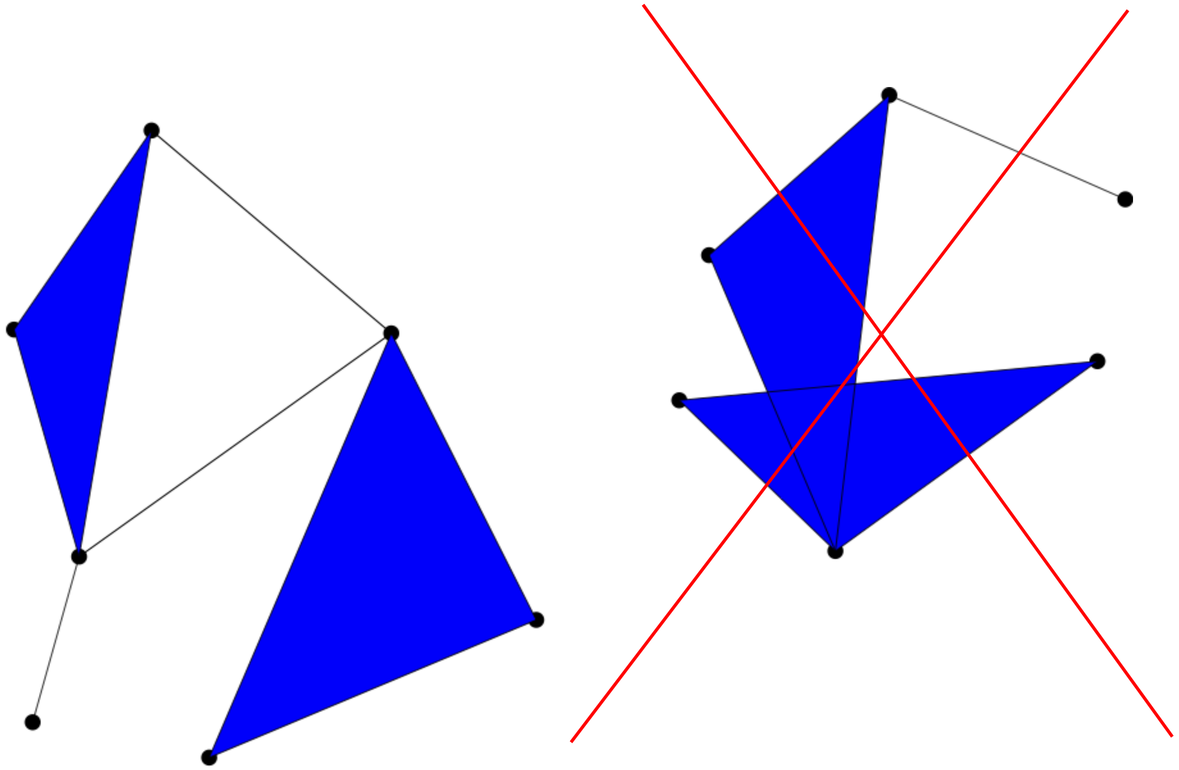
General idea: deduce from $(\mathbb{X}_n^r)_{r>0}$ some topological and geometric information of an underlying object.

Approximating models for TDA : Offsets and Simplicial Complexes

Non-discrete sets such as offsets, and also continuous mathematical shapes like curves, surfaces cannot easily be encoded as finite discrete structures.

A geometric simplicial complex $\mathcal{C}$ is a set of simplices such that:

- Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.
- The intersection of any two simplices s_1, $s_2 \in \mathcal{C}$ is either a face of both s_1 and s_2, or empty.

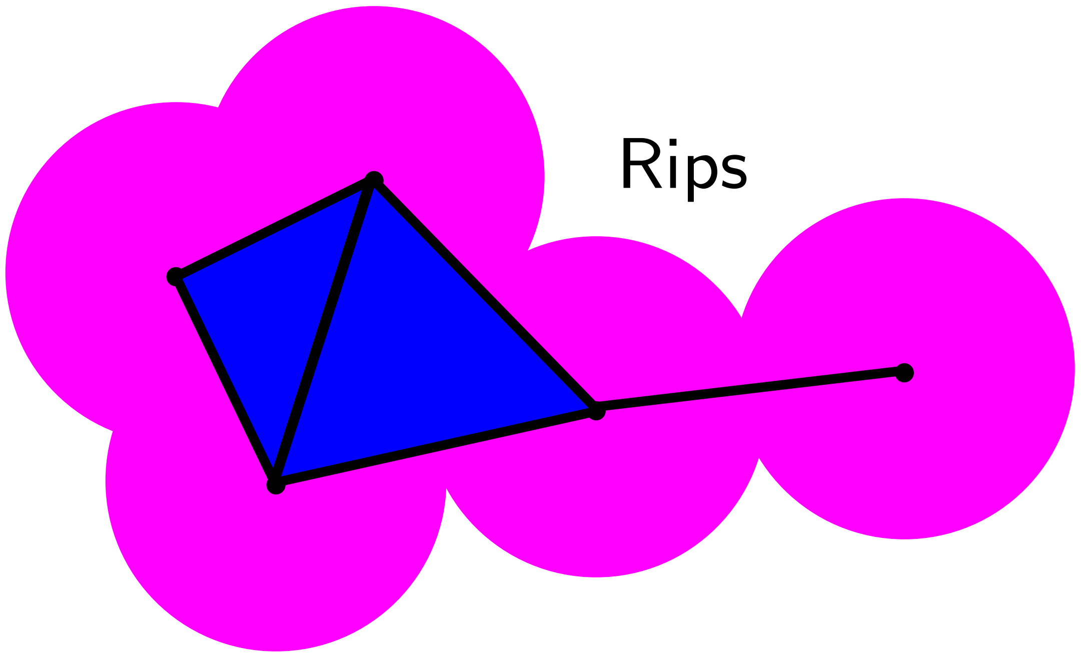


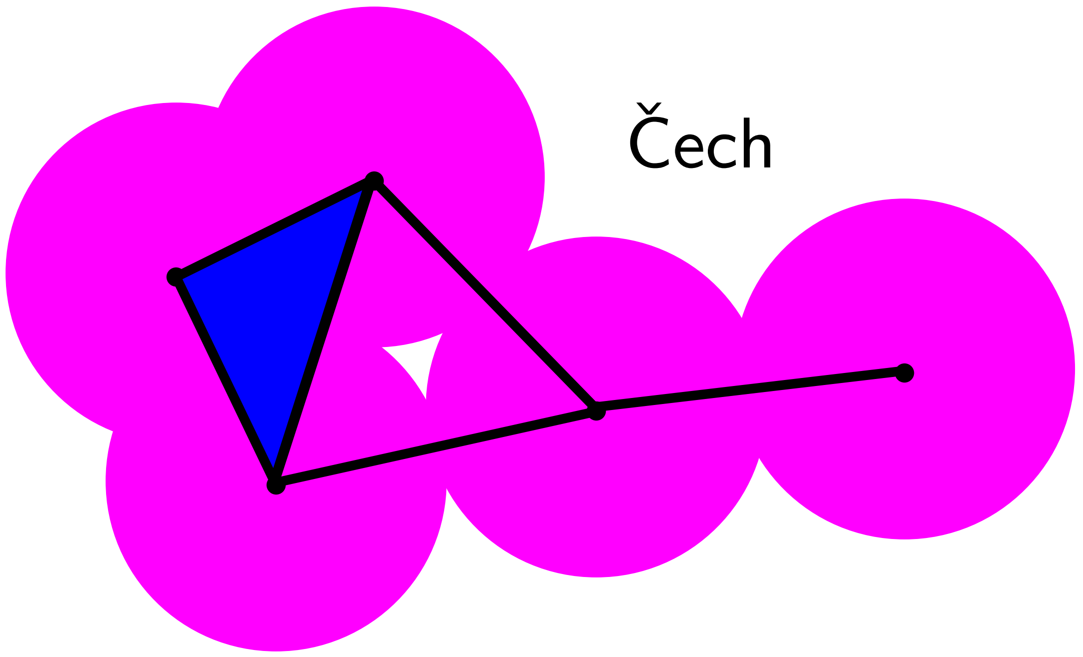
Approximating models for TDA : Offsets and Simplicial Complexes

Examples:

- A simplex $[x_0, x_1, \cdots, x_k]$ is in the Čech complex $\check{\mathbb{C}}\mathrm{ech}_\alpha(\mathbb{X}_n)$ if and only if $\bigcap_{j=0}^{k} B(x_j, \alpha) \neq \emptyset$.
- A simplex $[x_0, x_1, \cdots, x_k]$ is in the Rips complex $\mathbb{R}\mathrm{ips}_\alpha(\mathbb{X}_n)$ if and only if $\|x_j - x_{j'}\| \leq \alpha$ for all $j, j' \in \{1, \ldots, k\}$.

Can be also defined for a set of points in any metric space or for any compact metric space.

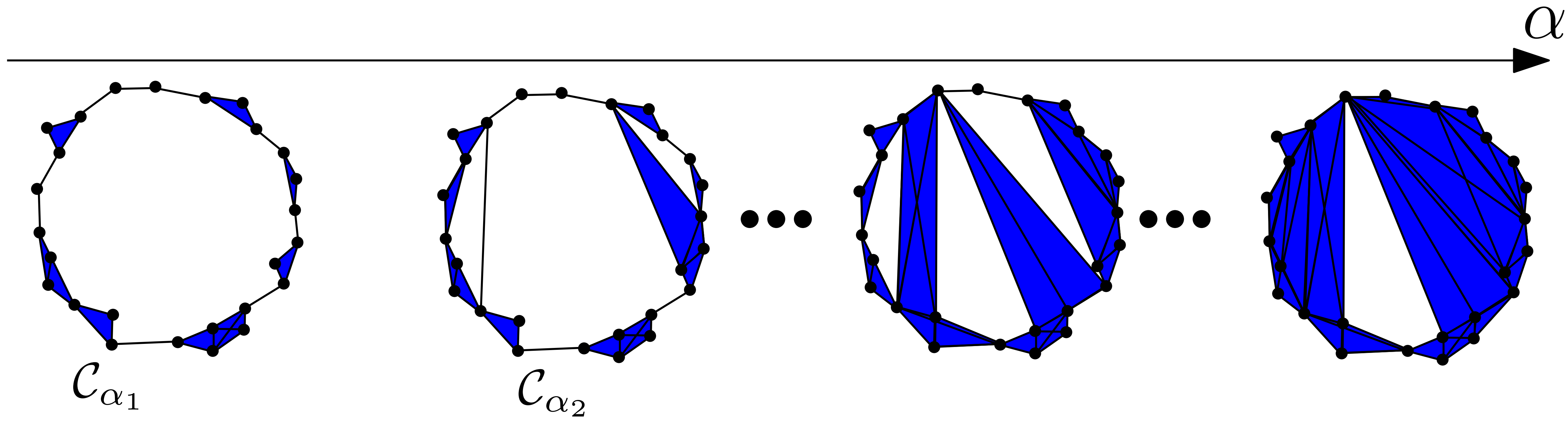




Nerve Theorem [Hatcher, 2001] : the offsets $\mathbb{X}_n^\alpha$ of a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ are homotopy equivalent to the Čech complex $\check{\mathbb{C}}\mathrm{ech}_\alpha(\mathbb{X}_n)$

Filtrations of simplicial complexes

Given a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$, we generally define a **filtration** of (nested simplicial) complexes by considering all the possibles scale parameters α : $(\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$



One difficult question : How can we choose a “convenient” scale parameter ?

Topological invariants

How topological spaces can be compared from a topological point of view ?

For comparing topological spaces, we condiser topological invariants (preserved by homeomorphism) : numbers, groups, polynomials.

Topological invariants

How topological spaces can be compared from a topological point of view ?

For comparing topological spaces, we condiser topological invariants (preserved by homeomorphism) : numbers, groups, polynomials.

Homotopy is weaker than homeomorphism but is preserves many topological invariants.

Two continous functions $f : X \to Y$ and $g : X \to Y$ are homotopic if there exists a continous application $H : X \times [0,1] \to Y$ such that $H(\cdot, 0) = f$ and $H(\cdot, 1) = g$.

Two topological spaces X and Y are homotopic if there exists two continous applications $f : X \to Y$ and $g : Y \to X$ such that

- $g \circ f$ is homotopic to id_X;
- $f \circ g$ is homotopic to id_Y;

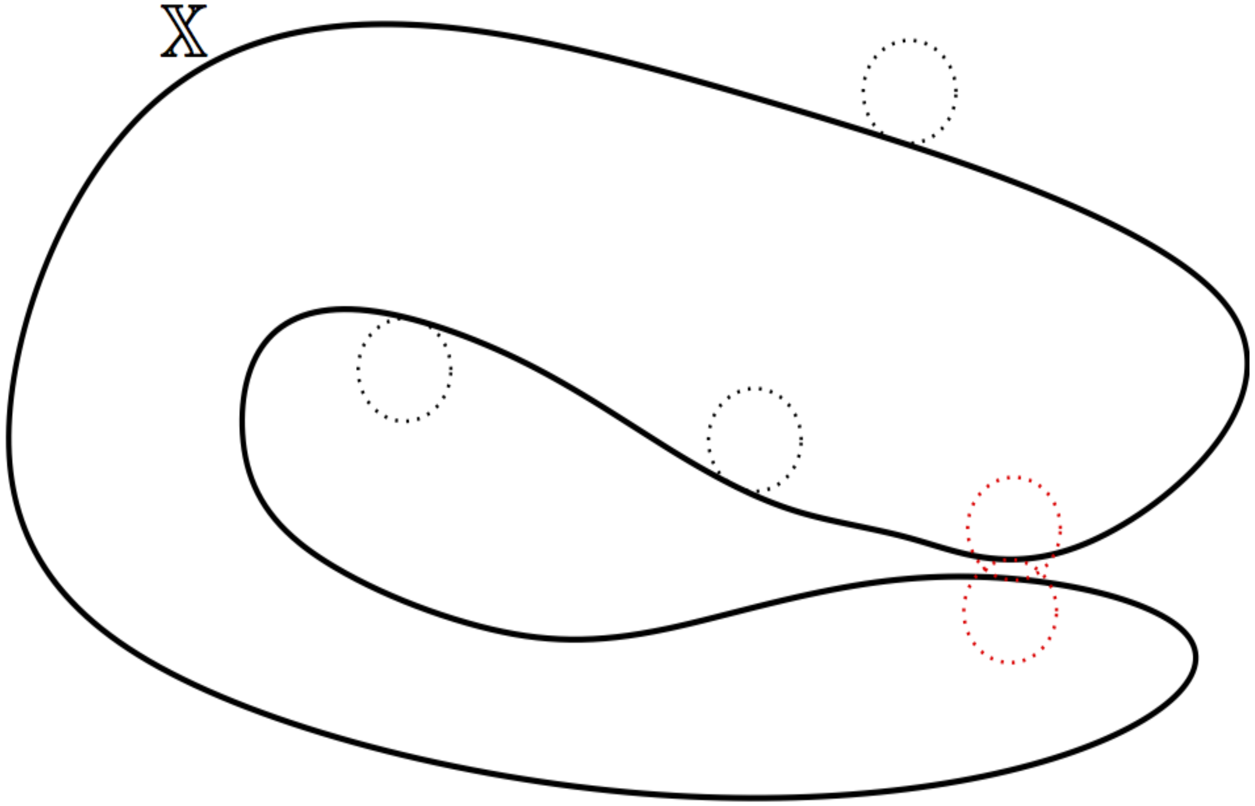
Topological Stability and Regularity

Topological inference : under “regularity assumptions”, topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

Topological Stability and Regularity

Topological inference : under "regularity assumptions", topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

- The *local feature size* is a local notion of regularity :
  For $x \in \mathbb{X}$, $\mathrm{lfs}_{\mathbb{X}}(x) := d\left(x, \mathcal{M}(\mathbb{X}^c)\right)$.



- The global version of the local feature size is the *reach* [Federer, 1959] :

$$\kappa(\mathbb{X}) = \inf_{x \in \mathbb{X}^c} \mathrm{lfs}_{\mathbb{X}}(x).$$

  The reach is small if either $\mathbb{X}$ is not smooth or if $\mathbb{X}$ is close to being self-intersecting.

- Weak feature size and its extensions [Chazal and Lieutier, 2007] (by considering the critical values of $d_{\mathbb{X}}$).

Topological Stability and Regularity

Topological inference : under “regularity assumptions”, topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

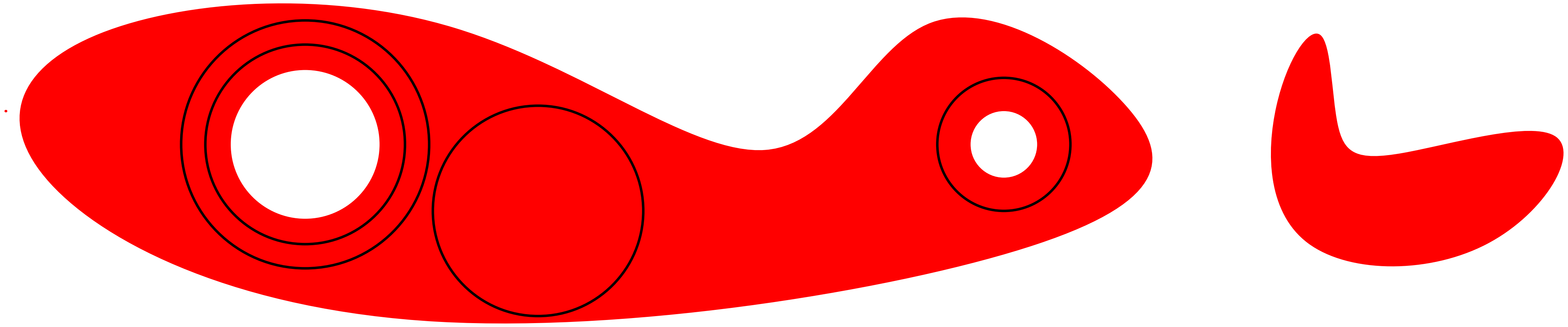
$$d_{\mathsf{H}}(\mathbb{X},\mathbb{Y}) = \inf\{\alpha \geq 0 \mid \mathbb{X} \subset \mathbb{Y}^{\alpha} \text{ and } \mathbb{Y} \subset \mathbb{X}^{\alpha}\}$$

Example :

Theorem [Chazal and Lieutier, 2007]: Let $\mathbb{X}$ and $\mathbb{Y}$ be two compact sets in $\mathbb{R}^d$ and let $\varepsilon > 0$ be such that $d_{\mathsf{H}}(\mathbb{X},\mathbb{Y}) < \varepsilon$, $\mathrm{wfs}(\mathbb{X}) > 2\varepsilon$ and $\mathrm{wfs}(\mathbb{Y}) > 2\varepsilon$. Then for any $0 < \alpha < 2\varepsilon$, $\mathbb{X}^{\alpha}$ and $\mathbb{Y}^{\beta}$ are homotopy equivalent.

Homology inference

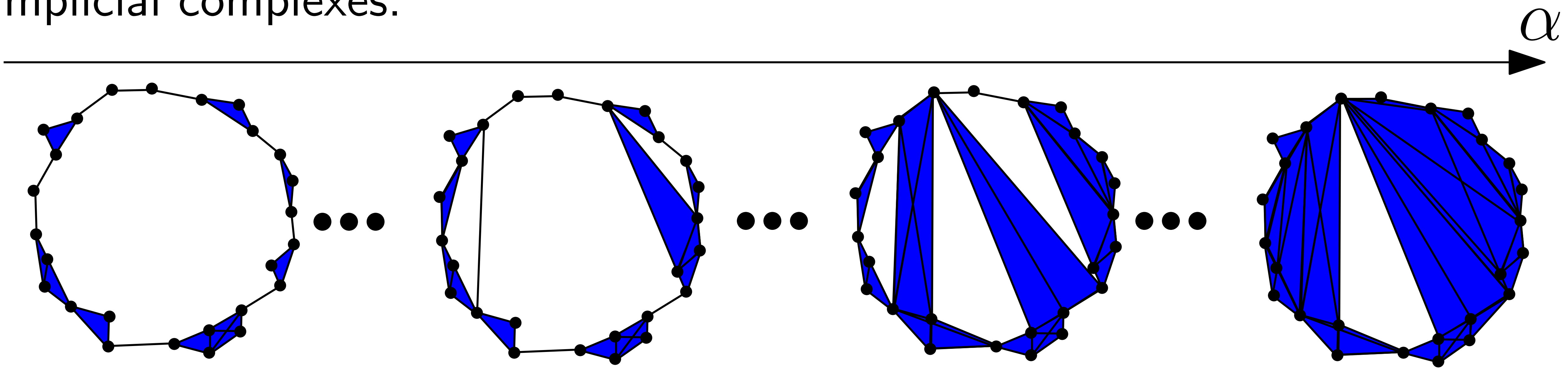
- **Homotopy** is not easy to compute in practice.
- **Singular homology** provides a algebraic description of “holes” in a geometric shape (connected components, loops, etc ...)
- **Betti number** β_k is the rank of the k-th homology group.
- **Computational Topology** : Betti numbers can be computed on simplicial complexes.



Homology inference [Niyogi et al., 2008 and 2011] [Balakrishnan et al., 2012] : The Betti number (actually the homotopy type) of Riemannian manifolds with positive reach can be recovered with high probability from offsets of a sample on (or close to) the manifold.

Persistent homology

Starting from a point cloud $\mathbb{X}_n$, let $\mathrm{Filt} = (\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$ be a fitration of nested simplicial complexes.

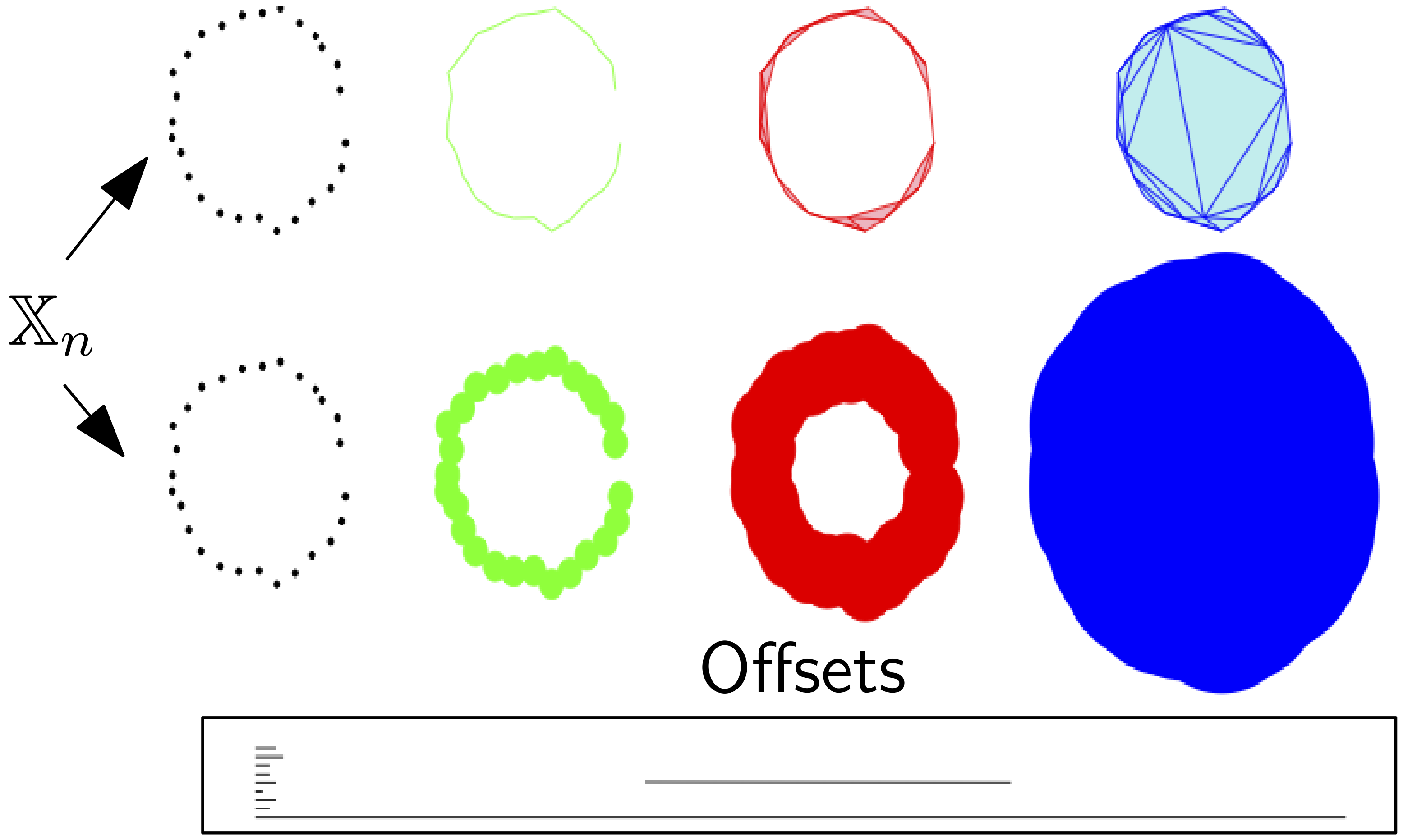


Persistent homology: identification of "persistent" topological features along the filtration.

- multiscale information ;
- more stable and more robust ;
- (but does not answer the scale selection problem...)

Barecodes and Persistence Diagrams

Filtration of simplicial complexes $\mathrm{Filt}(\mathbb{X}_n)$



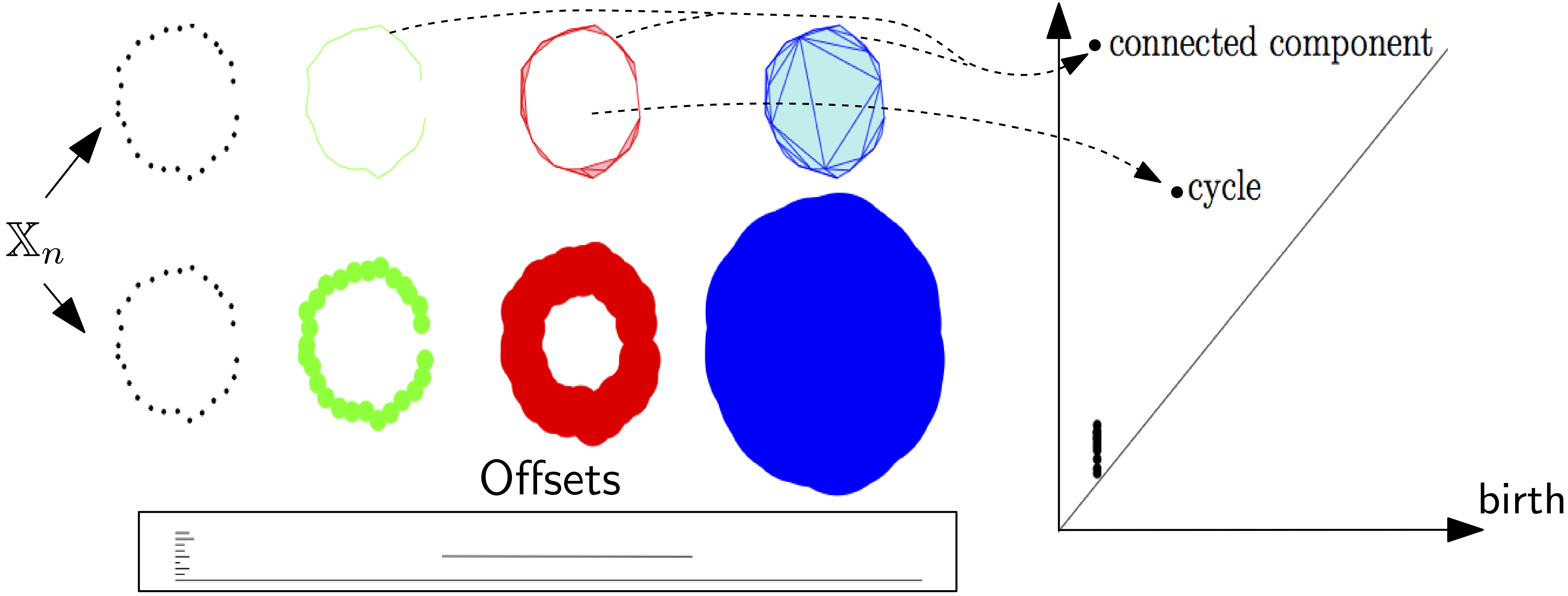
$\mathbb{X}_n$

Offsets

Barecode

Barecodes and Persistence Diagrams

Filtration of simplicial complexes $\mathrm{Filt}(\mathbb{X}_n)$

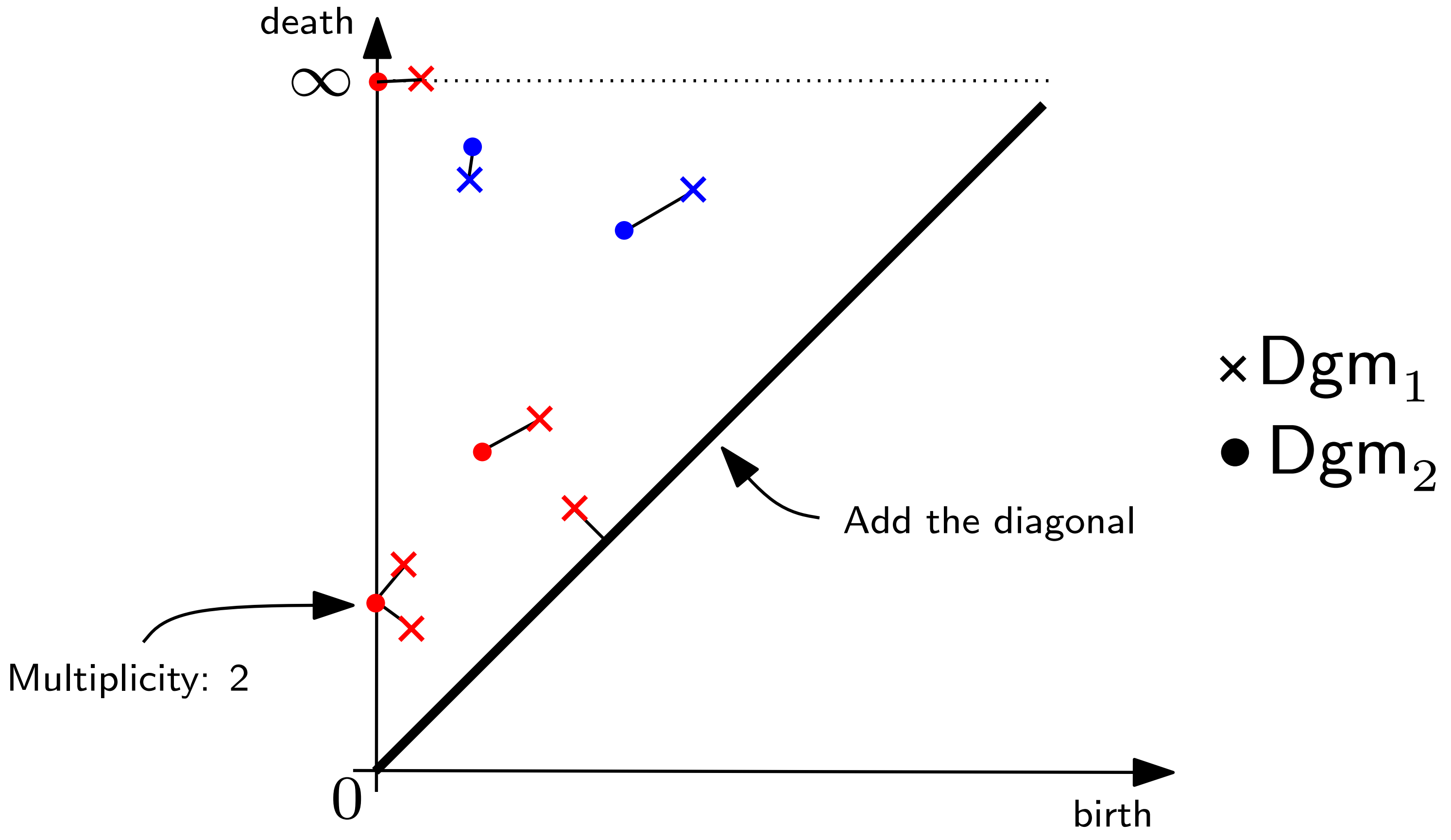


Offsets

Barecode

$\mathrm{Dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)$
Persistence diagram of the filtration $\mathrm{Filt}(\mathbb{X}_n)$ built on $\mathbb{X}_n$.

Distance between persistence diagrams and stability



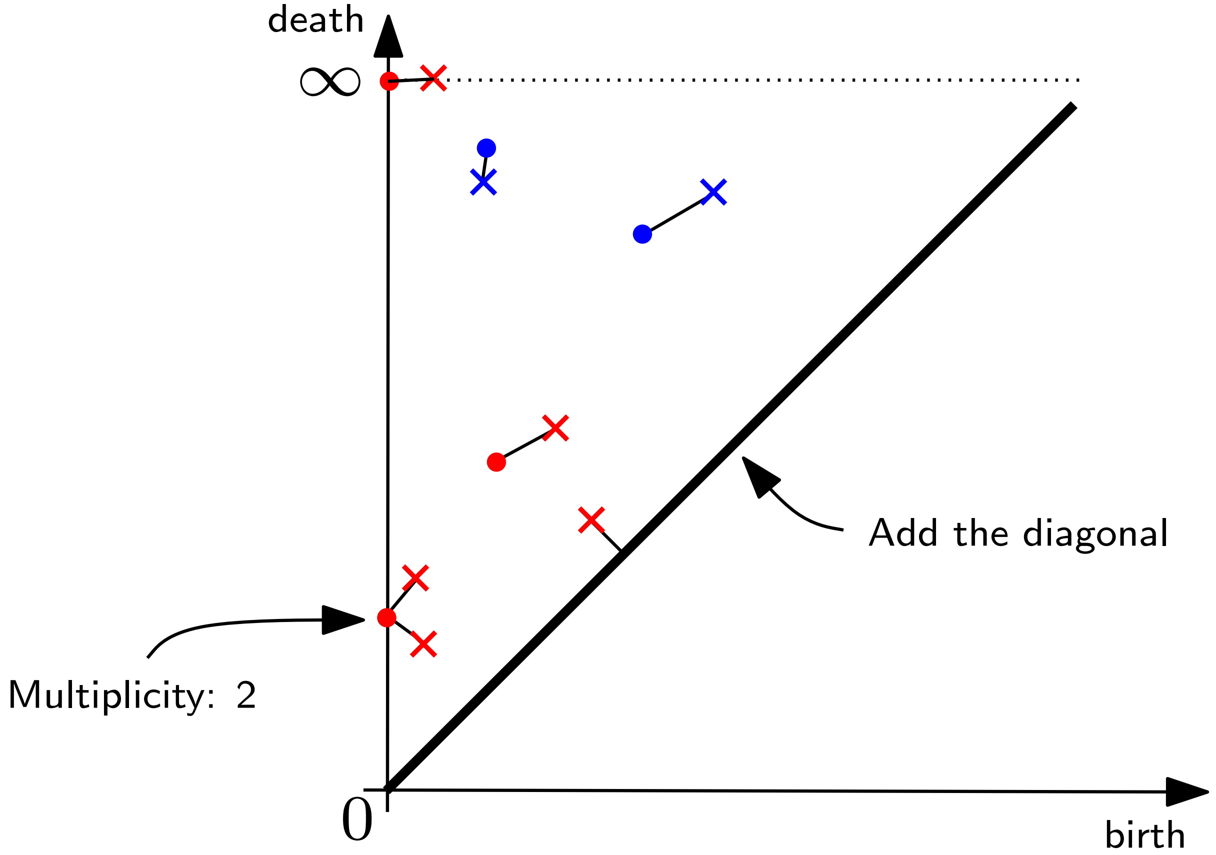
The bottleneck distance between two diagrams Dgm$_1$ and Dgm$_2$ is

$$\mathrm{d_b}(\mathsf{Dgm}_1, \mathsf{Dgm}_2) = \inf_{\gamma \in \Gamma} \sup_{p \in \mathsf{Dgm}_1} \|p - \gamma(p)\|_\infty$$

where Γ is the set of all the bijections between Dgm$_1$ and Dgm$_2$ and

$$\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|).$$

Distance between persistence diagrams and stability



× $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}))$

● $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{Y}))$

Theorem [Chazal et al., 2012]: For any compact metric spaces $(\mathbb{X}, \rho)$ and $(\mathbb{Y}, \rho')$,

$$\mathrm{d_b}\left(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{Dgm}(\mathrm{Filt}(\mathbb{Y}))\right) \leq 2\,\mathrm{d_{GH}}\left(\mathbb{X}, \mathbb{Y}\right).$$

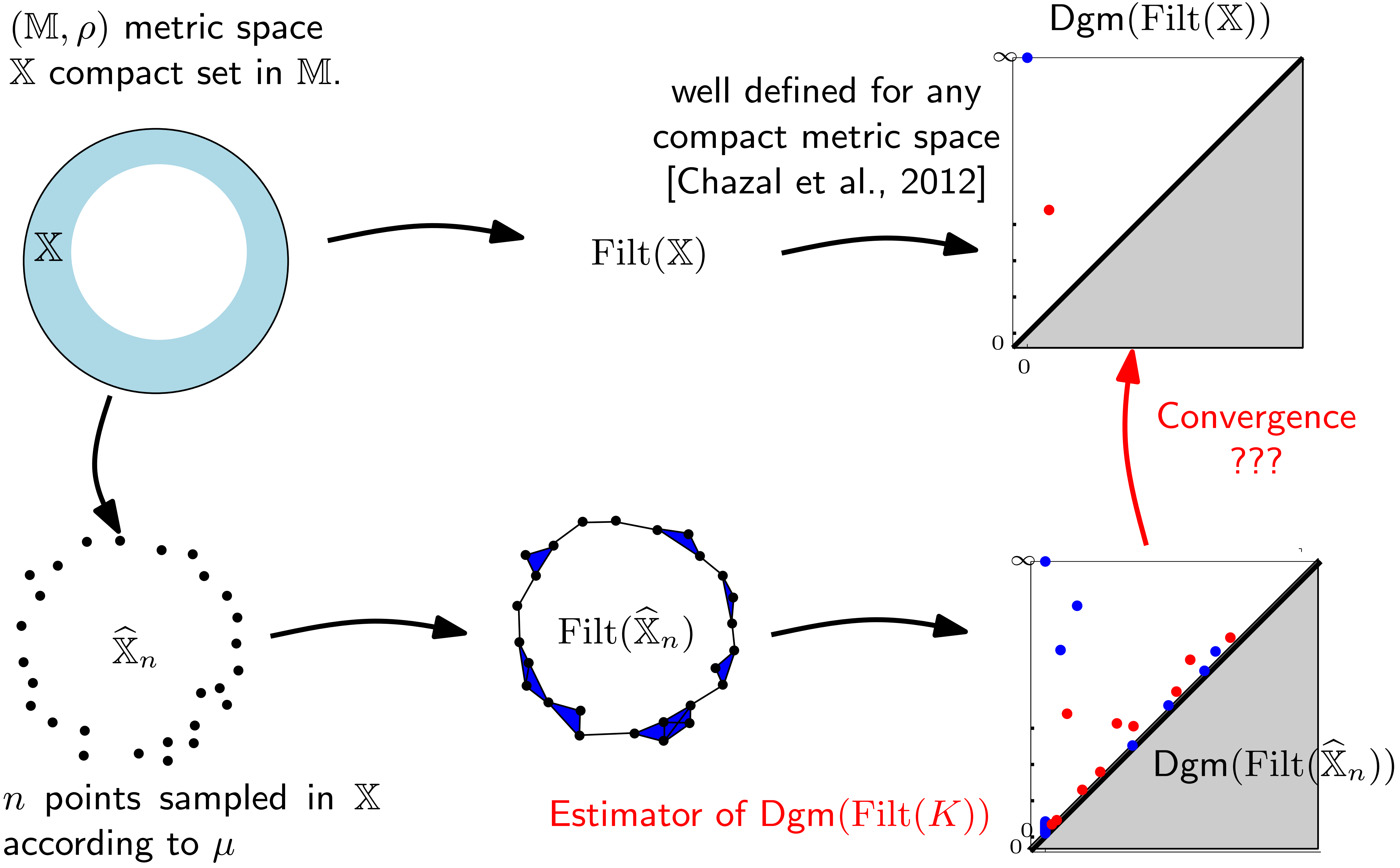
Consequently, if $\mathbb{X}$ and $\mathbb{Y}$ are embedded in the same metric space $(\mathbb{M}, \rho)$ then

$$\mathrm{d_b}\left(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{Dgm}(\mathrm{Filt}(\mathbb{Y}))\right) \leq 2\,\mathrm{d_H}\left(\mathbb{X}, \mathbb{Y}\right).$$

Statistics
and
Persistent homology

Persistence diagram inference [Chazal et al., 2014b]

Joint work with F. Chazal, M. Glisse and C. Labruère.



$(\mathbb{M}, \rho)$ metric space
$\mathbb{X}$ compact set in $\mathbb{M}$.

$\mathbb{X}$

$\mathrm{Filt}(\mathbb{X})$

well defined for any
compact metric space
[Chazal et al., 2012]

$\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}))$

Convergence
???

$\widehat{\mathbb{X}}_n$

$\mathrm{Filt}(\widehat{\mathbb{X}}_n)$

$\mathrm{Dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_n))$

n points sampled in $\mathbb{X}$
according to μ

Estimator of $\mathrm{Dgm}(\mathrm{Filt}(K))$

Persistence diagram inference [Chazal et al., 2014]

For $a, b > 0$, μ satisfies the (a, b)-standard assumption on its support $\mathbb{X}_\mu$ if for any $x \in X_\mu$ and any $r > 0$:

$$\mu(B(x, r)) \geq \min(ar^b, 1).$$

$\mathcal{P}(a, b, \mathbb{M})$: set of all the probability measures satisfying the (a, b)-standard assumption on the metric space $(\mathbb{M}, \rho)$.

Theorem: For $a, b > 0$:

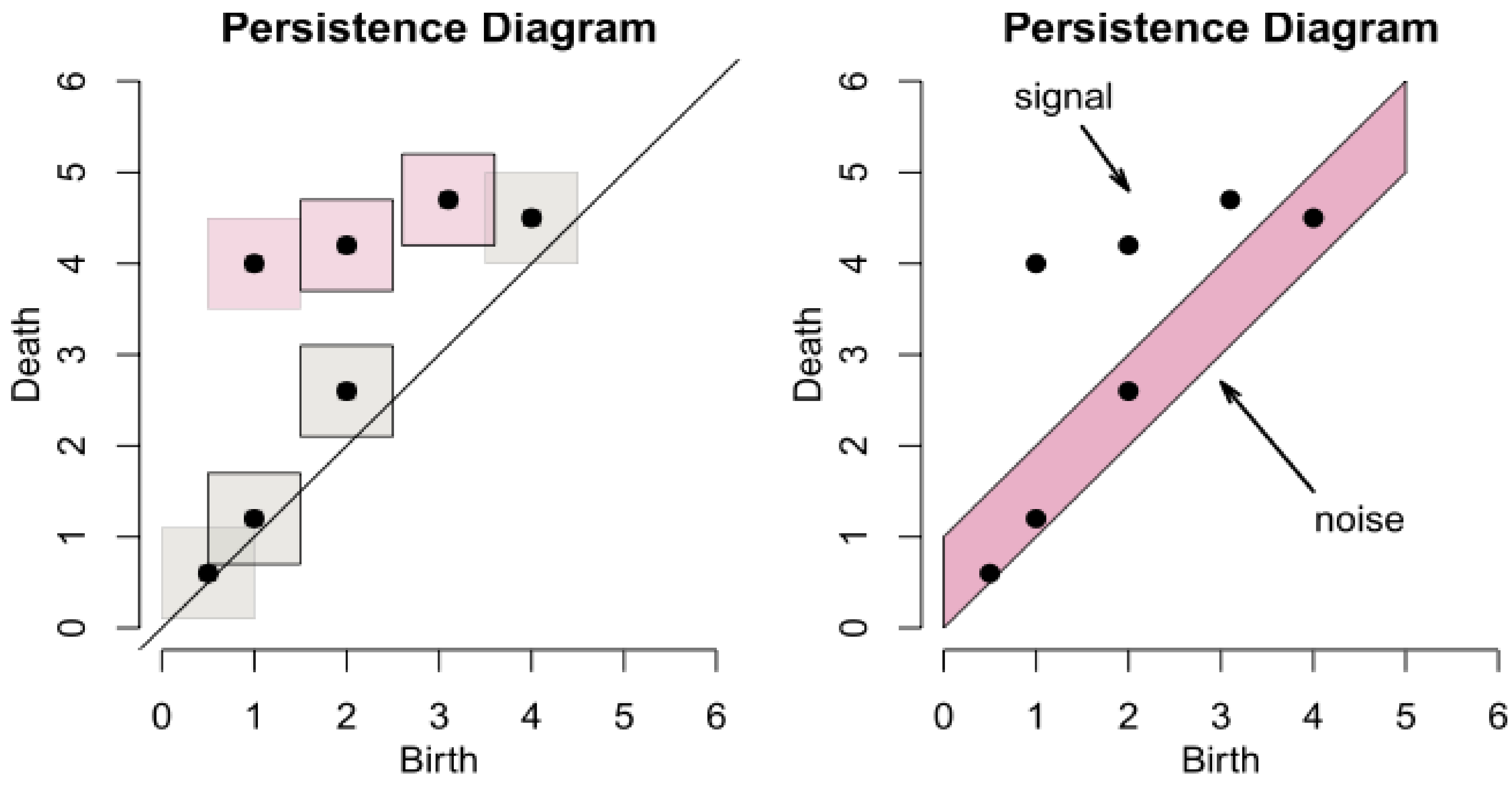
$$\sup_{\mu \in \mathcal{P}(a,b,\mathbb{M})} \mathbb{E}\left[\mathrm{d_b}(\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{Dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_n)))\right] \leq C \left(\frac{\ln n}{n}\right)^{1/b}$$

where C only depends on a and b.
Under additional technical hypotheses, for any estimator $\widehat{\mathsf{Dgm}}_n$ of $\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$:

$$\liminf_{n \to \infty} \sup_{\mu \in \mathcal{P}(a,b,\mathbb{M})} \mathbb{E}\left[\mathrm{d_b}(\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathsf{Dgm}}_n)\right] \geq C' n^{-1/b}$$

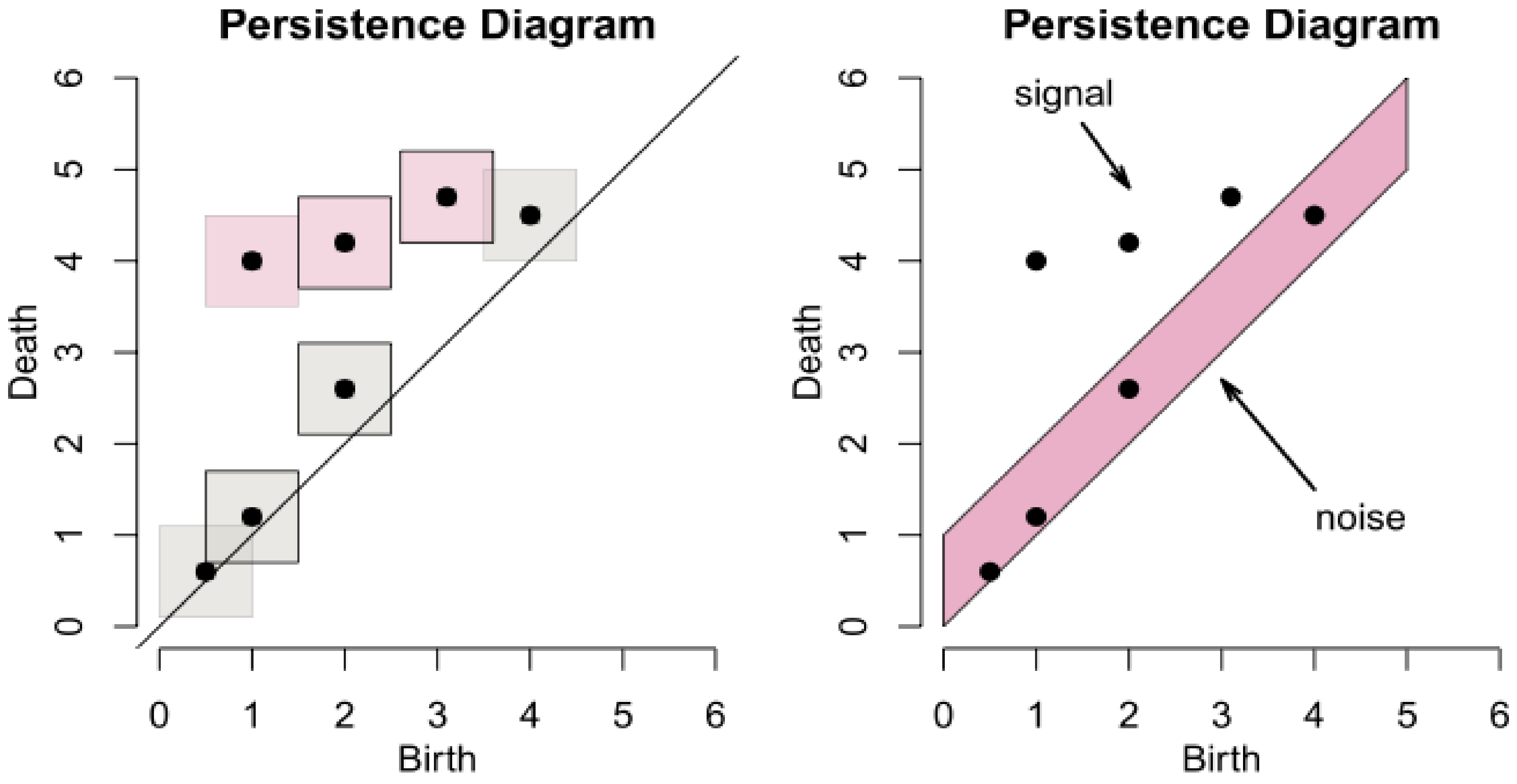
where C' is an absolute constant.

Confidence sets for persistence diagrams [Fasy et al., 2014]



$$P\left(\mathsf{Dgm}(\mathrm{Filt}(K)) \in \hat{\mathcal{R}}\right) \geq 1-\alpha \qquad ??$$

Confidence sets for persistence diagrams [Fasy et al., 2014]



Persistence Diagram

Persistence Diagram

$$P\left(\mathsf{Dgm}(\mathrm{Filt}(K)) \in \hat{\mathcal{R}}\right) \geq 1 - \alpha \qquad ??$$

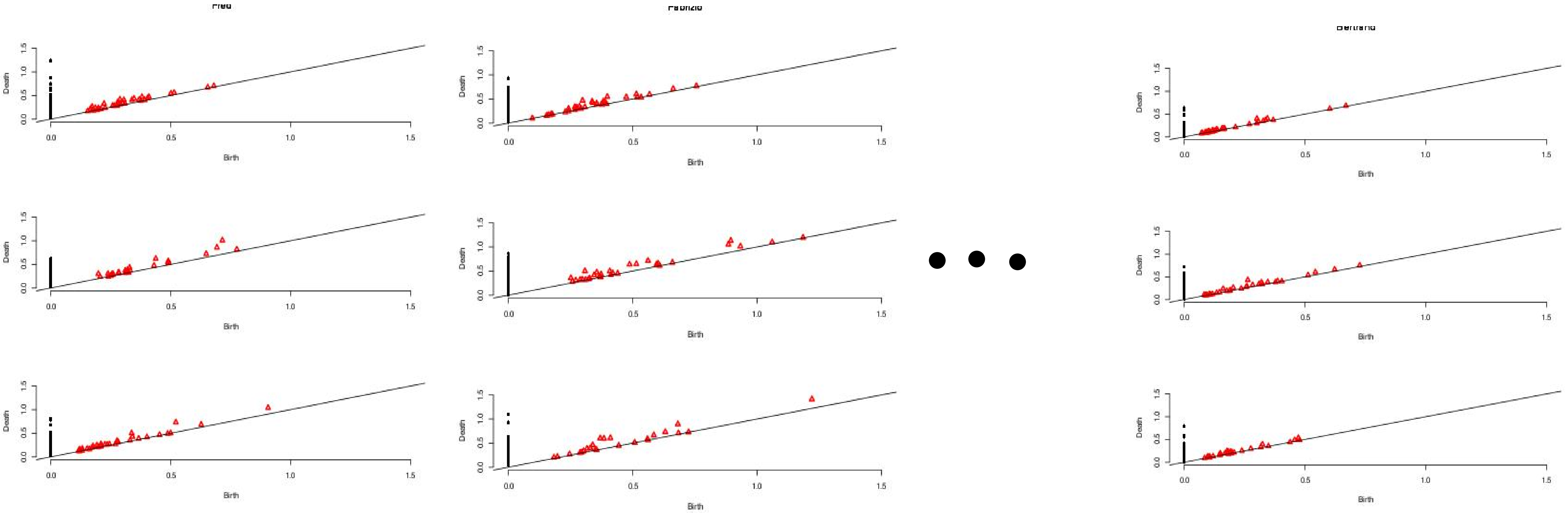
Using the Hausdorff stability, we can define confidence sets for persistence diagrams.

$$W_\infty\left(\mathsf{Dgm}\left(\mathrm{Filt}(K)\right), \mathsf{Dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)\right) \leq \mathrm{d_H}(K, \mathbb{X}_n)$$

it is sufficient to find c_n such that

$$\limsup_{n\to\infty}\left(\mathrm{d_H}(K, \mathbb{X}_n) > c_n\right) \leq \alpha.$$

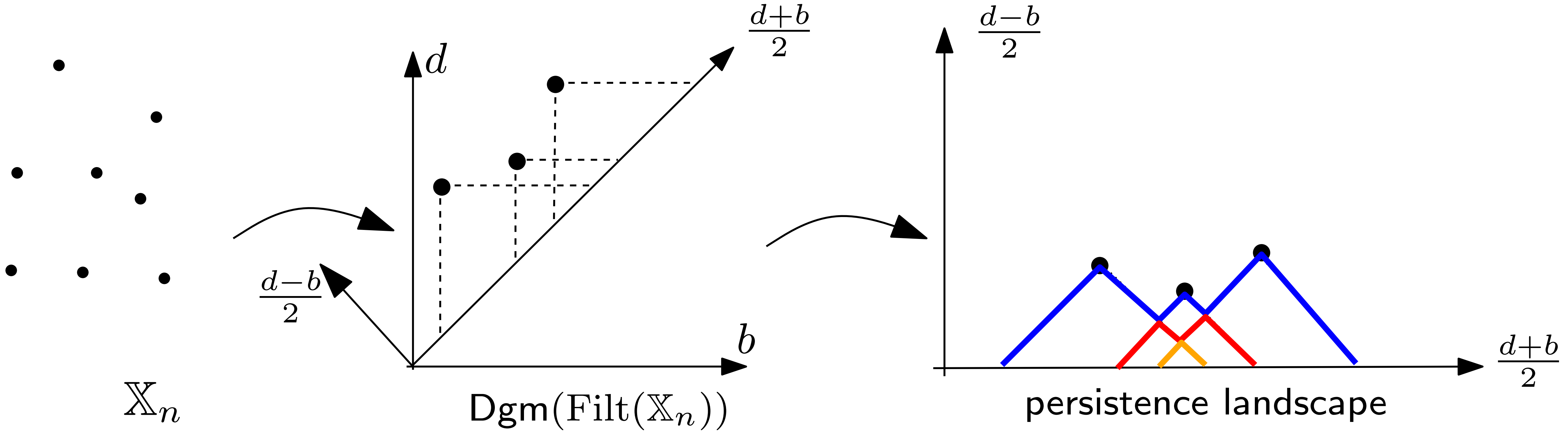
How can be defined the “mean” of a family of persistence diagram ?



- Frechet mean [Turner et al., 2014]
  Difficult to compute and no unicity.

- Use an alternative descriptor of persistence : Persistence landscapes [Bubenik, 2015]

Persistence landscapes [Bubnik, 2015]



$$\mathrm{Dgm} = \left\{ \left(\tfrac{d_i+b_i}{2}, \tfrac{d_i+b_i}{2}\right),\, i \in I \right\}$$

For $p = \left(\frac{b+d}{2}, \frac{d-b}{2}\right) \in \mathrm{Dgm}$,

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

Persistence landscape λ of Dgm:

$$\lambda(k,t) = \underset{p \in D}{\mathrm{kmax}}\ \Lambda_p(t), \quad t \in \mathbb{R},\ k \in \mathbb{N},$$

where kmax is k-th largest value in the set.

Stability: For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $|\lambda(k,t) - \lambda'(k,t)| \le \mathrm{d_b}(\mathrm{Dgm}, \mathrm{Dgm}')$.

Subsampling methods for pers. homology [Chazal et al., 2015]

joint work with F. Chazal, B. Fasy, F. Lecci, A. Rinaldo and L. Wasserman

- Let $X = \{X_1, \cdots, X_m\}$ sampled from μ.
- λ_X: corresponding persistence landscape.
- Ψ_μ^m: the measure induced by $\mu^{\otimes m}$ on the space of persistence landscapes.
- We consider the point-wise expectations of the (random) persistence landscape under this measure:

$$\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$$

- For $S_1^m, \ldots, S_\ell^m$ some independent samples of size m from $\mu^{\otimes m}$, the empirical counterpart of $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)]$ is

$$\overline{\lambda_\ell^m}(t) = \frac{1}{\ell} \sum_{i=1}^{\ell} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T],$$

Risk analysis of $\overline{\lambda_\ell^m}$ and as an estimator of $\lambda_{\mathbb{X}_\mu}$ in the context of (a, b)-standard measures.

Subsampling methods for pers. homology [Chazal et al., 2015]

Definition: The p-th Wasserstein distance between two measures μ, ν defined on $(\mathbb{M}, \rho)$ is

$$W_{\rho,p}(\mu,\nu) = \left(\inf_{\Pi} \int_{\mathbb{M}\times\mathbb{M}} [\rho(x,y)]^p d\Pi(x,y) \right)^{\frac{1}{p}},$$

where the infimum is taken over all measures on $\mathbb{M} \times \mathbb{M}$ with marginals μ and ν.

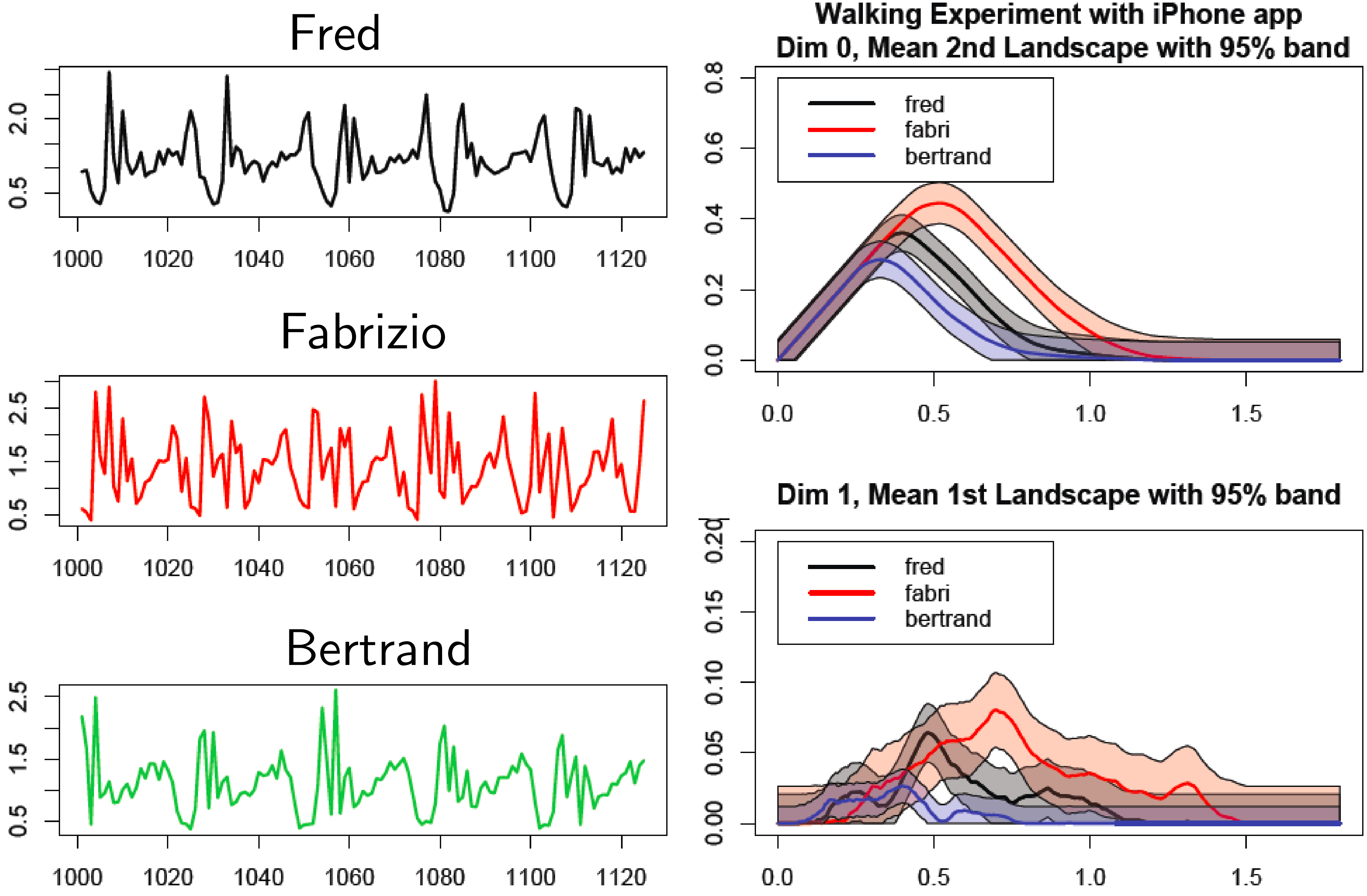
Stability of the average landscape:

Theorem: Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on $\mathbb{M}$. For any $p \geq 1$ we have

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2\, m^{\frac{1}{p}} W_{\rho,p}(\mu,\nu).$$

Subsampling methods for pers. homology [Chazal et al., 2015]

Application: Analysis of accelerometer data.

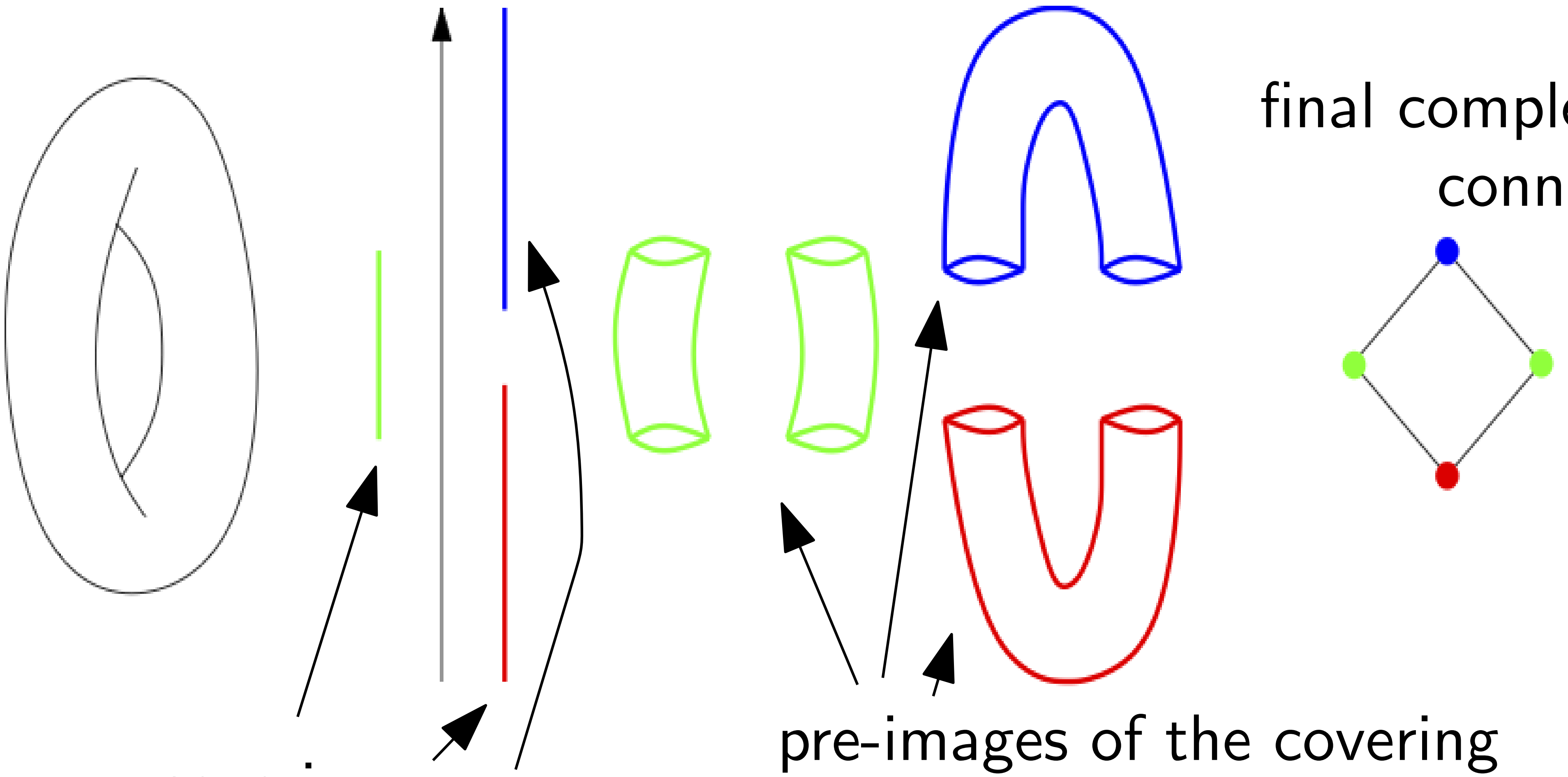


- topological features carry discriminative information
- no registration/calibration preprocessing step needed

Mapper in two slides

MAPPER (Singh etal., 2007)

filter function



final complex shows the connexions

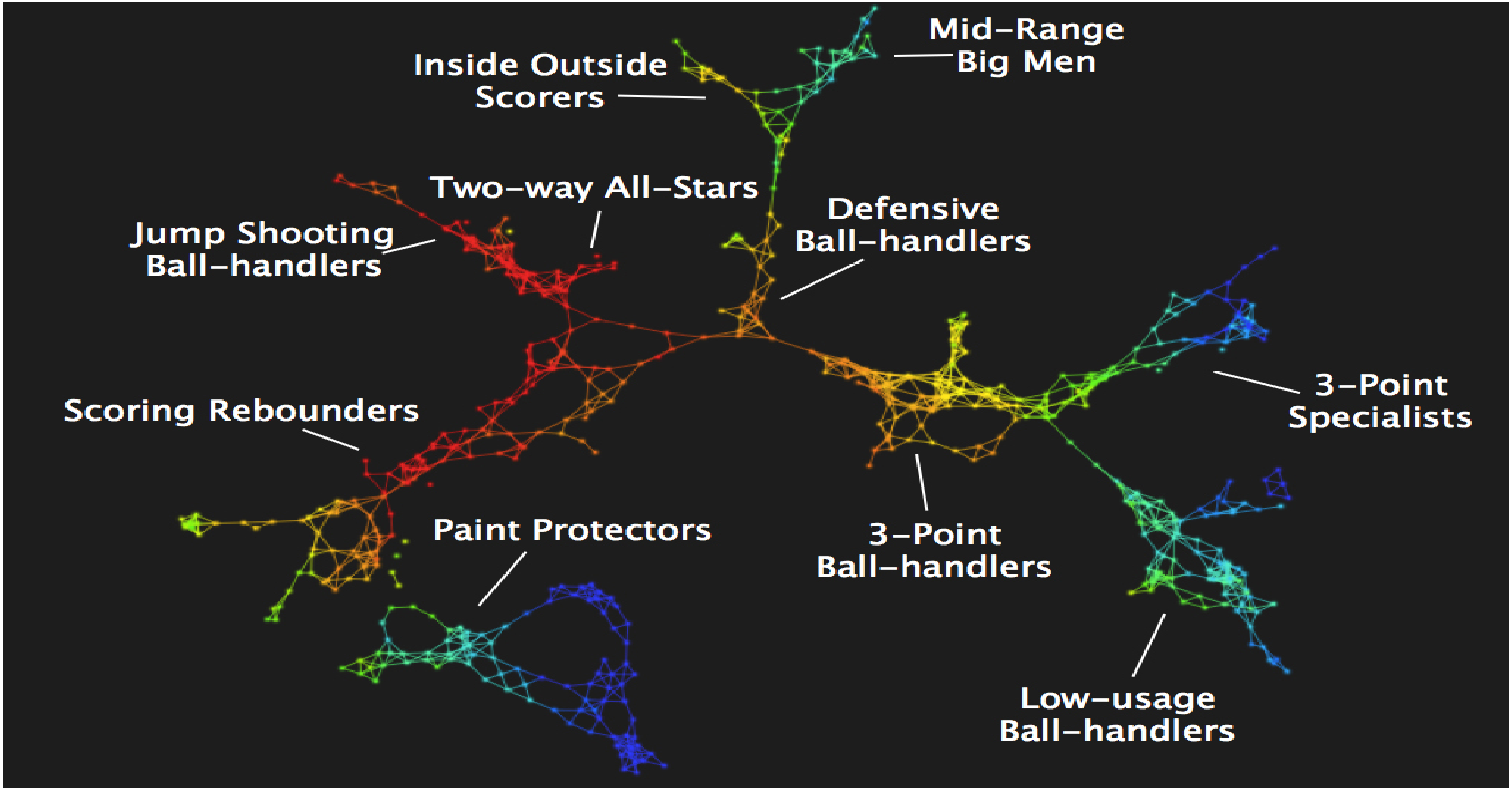
pre-images of the covering

covering

[credits : M. Carrière]

MAPPER (Singh etal., 2007)

Application to NBA players



[credits : AYASDI company]

Concluding remarks

- TDA methods focus on the topological properties (homology / persistent homology) of a shape.

- TDA methods can be used

  - as an “exploratory method”, in particuar when the point cloud is sampled on (close to) a real geometric object

  - as a “feature extraction” procedure, next these extracted features can be used for learning purposes.

- TDA is an emerging field, at the interface maths, computer sciences, stat

  Applications in many fields of sciences (medecine, biology, dynamic systems, astronomy, dynamical systems, physics ...)

- TDA methods need to bring together Geometric Inference, Computational Topology and Geometry, Statistics and Learning methods.

Thank you !

References

[Balakrishnan et al., 2012] Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A., and Wasserman, L. A.. Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72.

[Bubenik, 2015] Bubenik, P. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102.

[Carlsson, 2009] Carlsson, G. Topology and data. *AMS Bulletin*, 46(2):255–308.

[Chazal et al., 2014] Chazal, F., Glisse, M., Labruère, C., and Michel, B. Convergence rates for persistence diagram estimation in topological data analysis. To appear in *Journal of Machine Learning Research*.

[Chazal et al., 2014b] Chazal, F., Glisse, M., Labruère, C., and Michel, B. Convergence rates for persistence diagram estimation in topological data analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 163–171.

[Chazal and Lieutier, 2007] Chazal, F. and Lieutier, A. Stability and computation of topological invariants of solids in {\ Bbb R}^ n. *Discrete & Computational Geometry*, 37(4):601–617.

[Chazal et al., 2012] Chazal, F., de Silva, V., Glisse, M., and Oudot, S. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*.

[Chazal et al., 2015] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Subsampling methods for persistent homology. To appear in *Proceedings of the 32 st International Conference on Machine Learning (ICML-15)*.

References

[Fasy et al., 2014] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

[Niyogi et al., 2008] Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441.

[Singh et al., 2007] Singh, G., Mémoli, F., and Carlsson, G. E. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100.

[Turner et al., 2014] Turner, K., Mileyko, Y., Mukherjee, S. and Harer, J. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70 .