Saclay, July 8, 2014
DataSense Research day

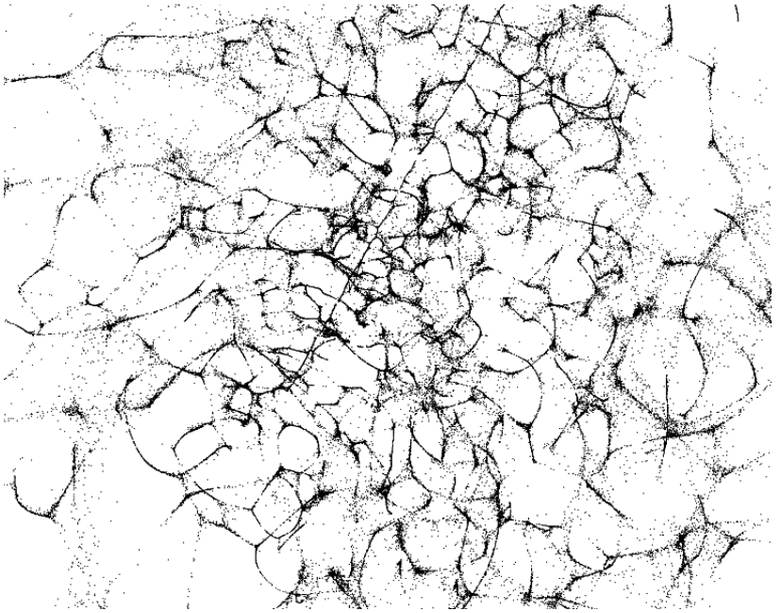
# Statistical properties of topological information inferred from data

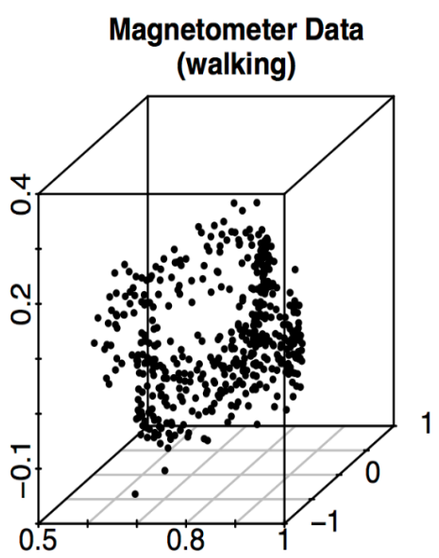
Frédéric Chazal
INRIA Saclay

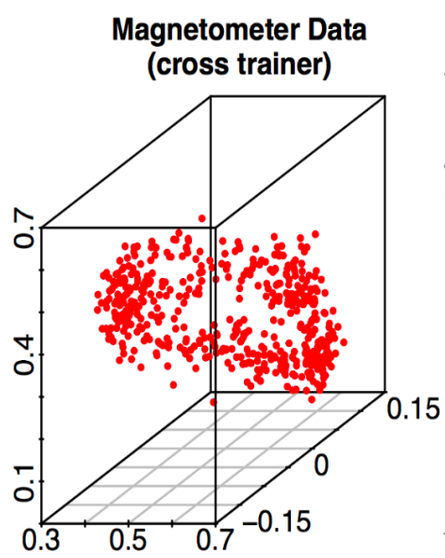
Joint works with B.T. Fasy (CMU), M. Glisse (INRIA), C. Labruère (Univ. Bourgogne), F. Lecci (CMU), B. Michel (LSTA Paris 6), A. Rinaldo (CMU), L. Wasserman (CMU).


# Introduction

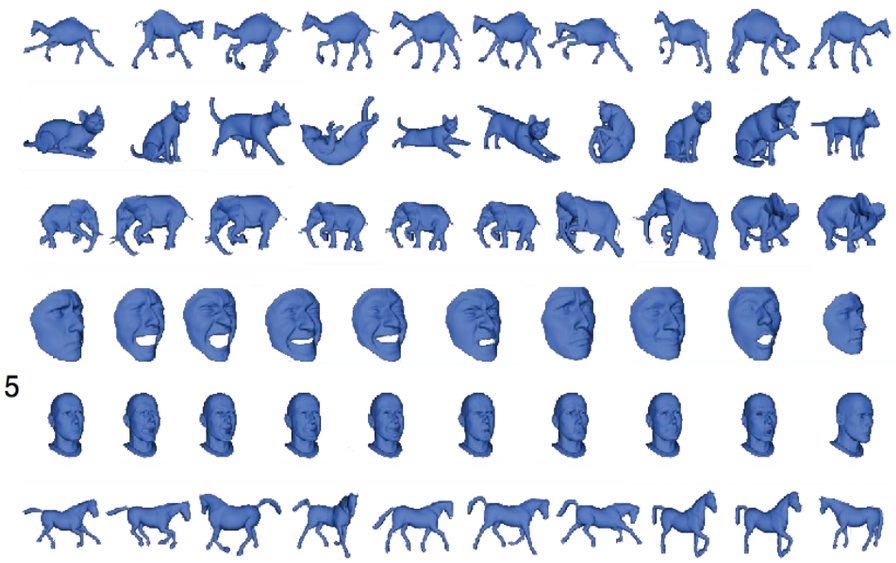
[distribution of galaxies]





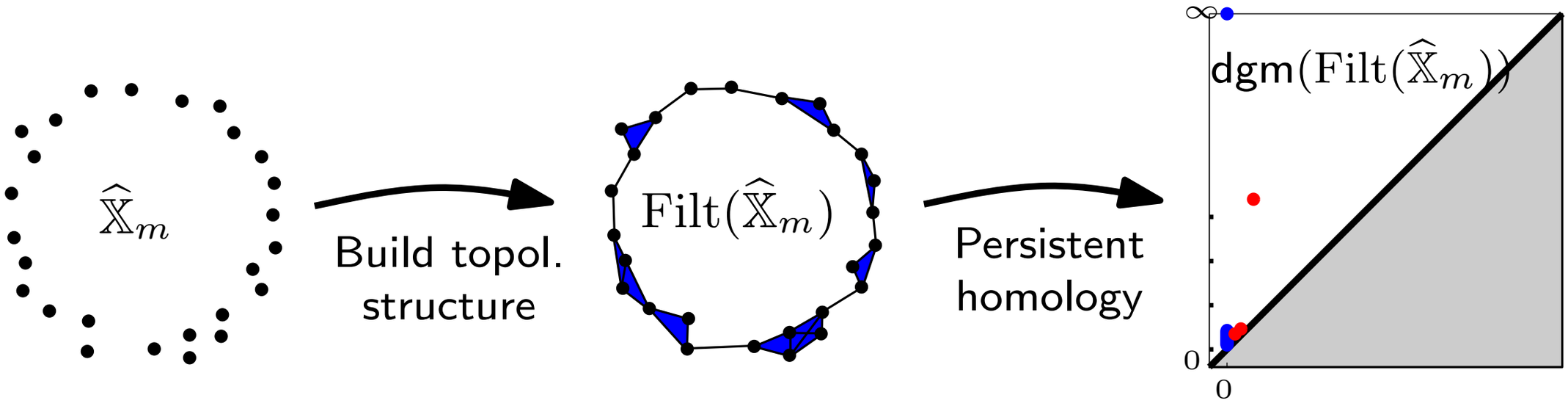




[Non-rigid shape database]



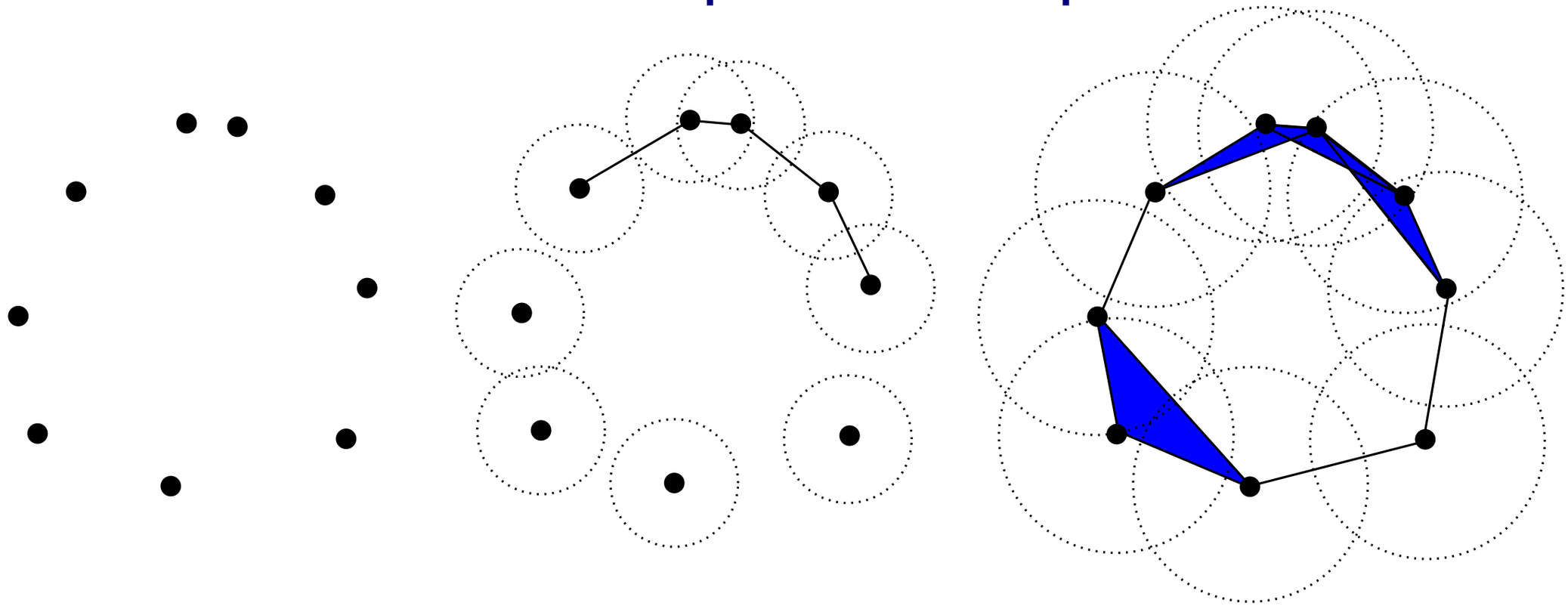
- Data often come as (sampling of) metric spaces or sets/spaces endowed with a similarity measure with, possibly complex, topological/geometric structure.
- Topological Data Analysis (TDA):
  - infer relevant topological and geometric features of these spaces.
  - take advantage of topol./geom. information for further processing of data (classification, recognition, learning, clestering,...).

# Introduction




- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

# Filtered simplicial complexes



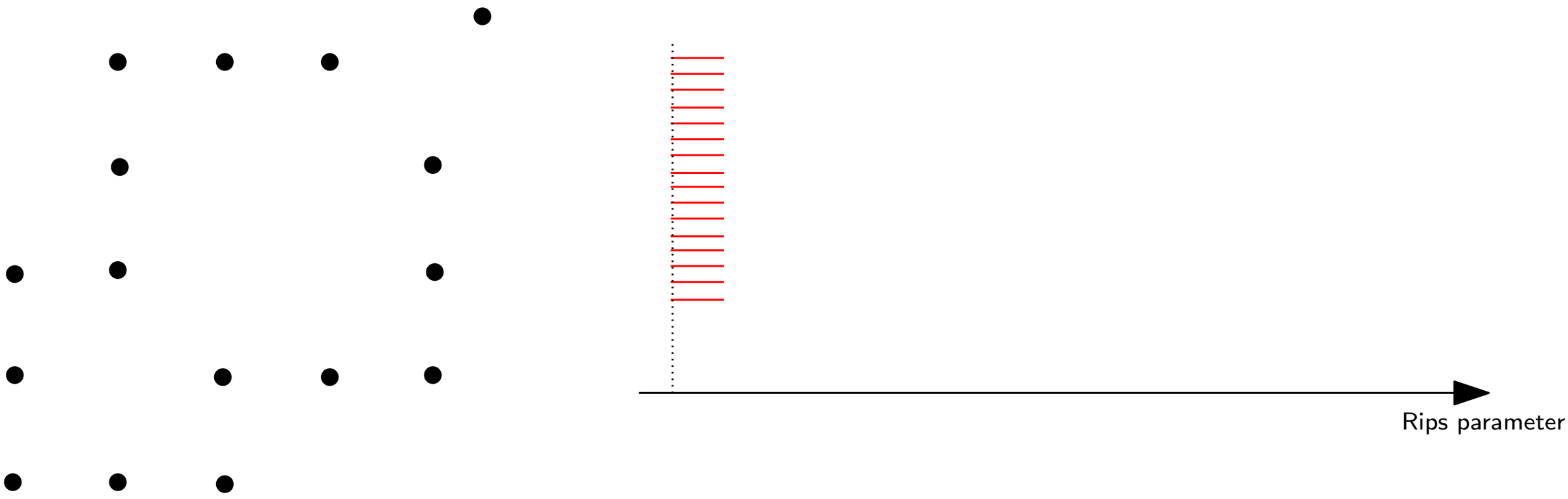
- simplicial complex: generalizes the notion of neighboring graph to higher dimensions by adding triangles,tetrahedra,...

- filtered simplicial complex $=$ nested family of simplicial complexes indexed by $\mathbf{R}$

**Example (used all along the talk):** Let $(X, d_X)$ be a metric space.
The Vietoris-Rips complex $\mathbb{R}\mathrm{ips}(X)$ is defined by: for $a \in \mathbf{R}$,

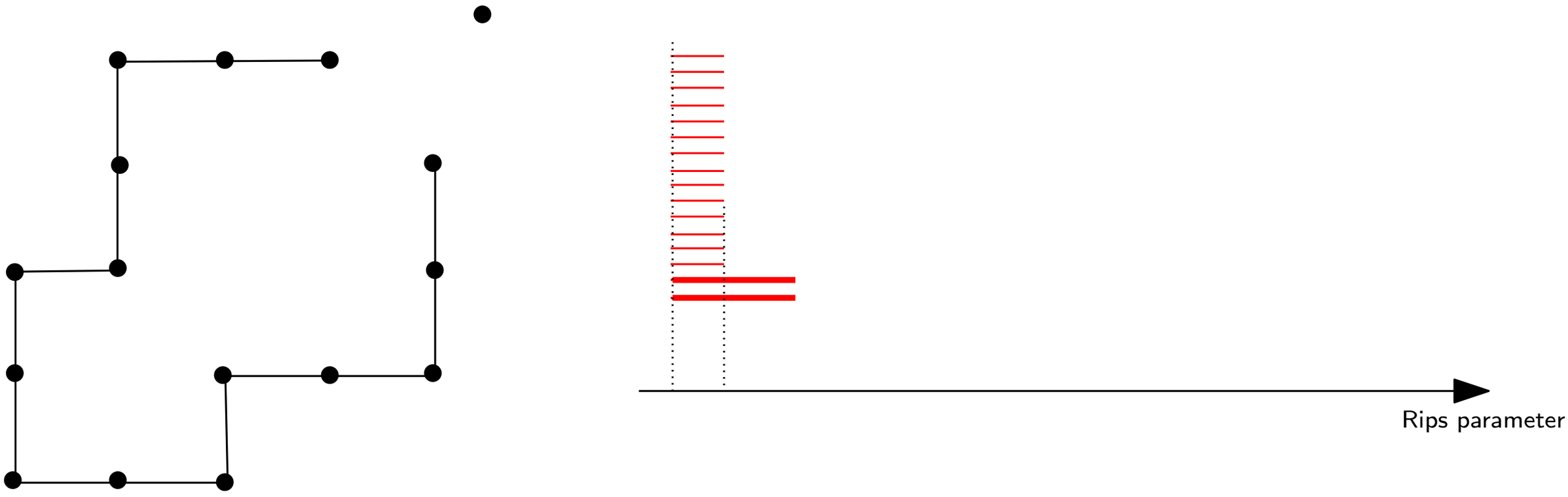
$$[x_0, x_1, \cdots, x_k] \in \mathrm{Rips}(X, a) \Leftrightarrow d_X(x_i, x_j) \leq a, \quad \text{for all } i, j$$

# Persistent homology of filtered complexes



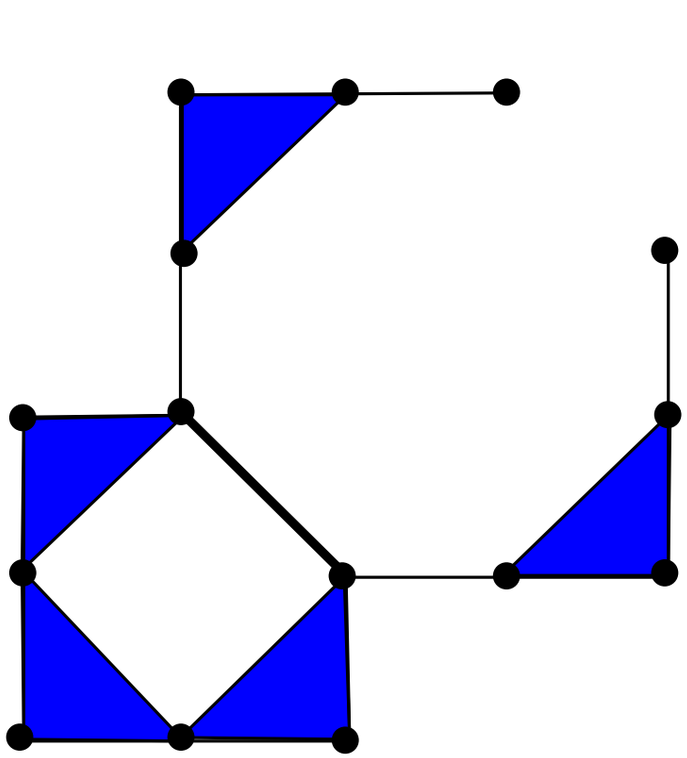

- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

# Persistent homology of filtered complexes



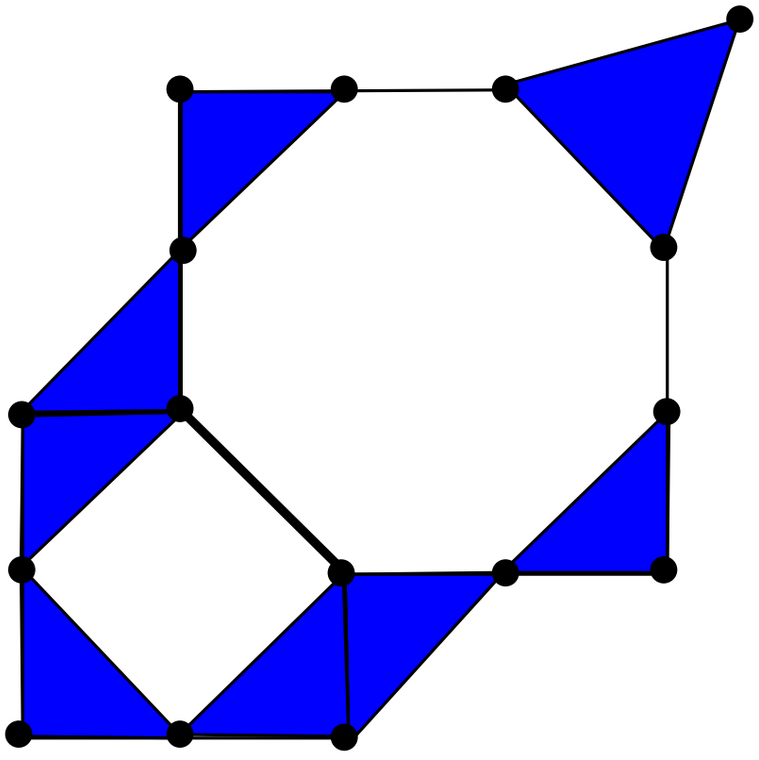

- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

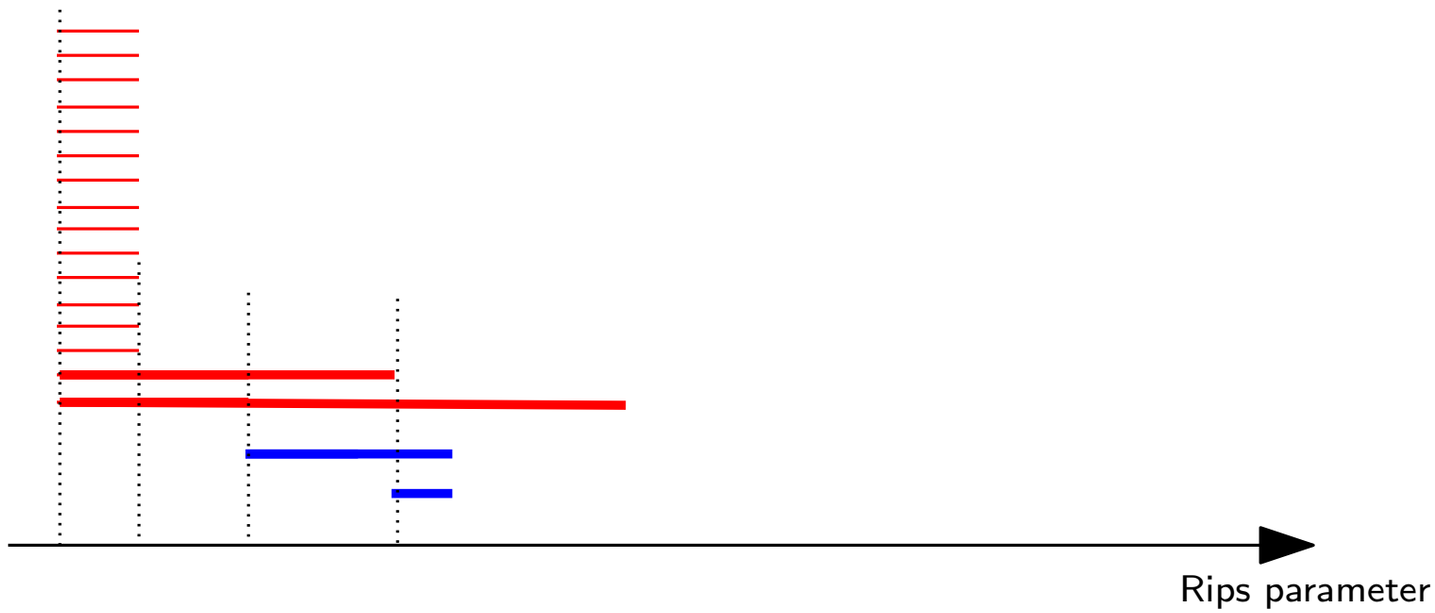
# Persistent homology of filtered complexes




- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

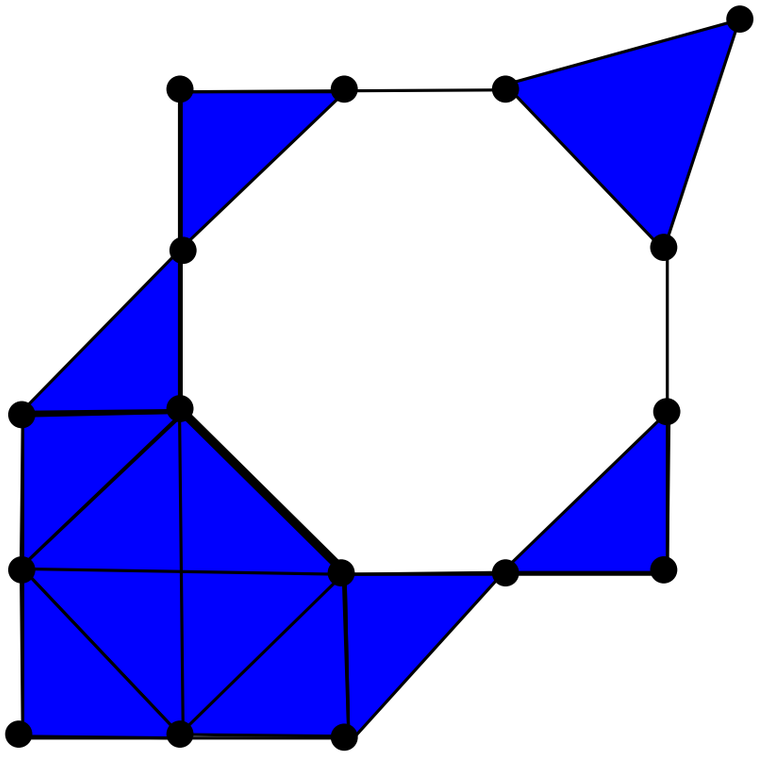
# Persistent homology of filtered complexes

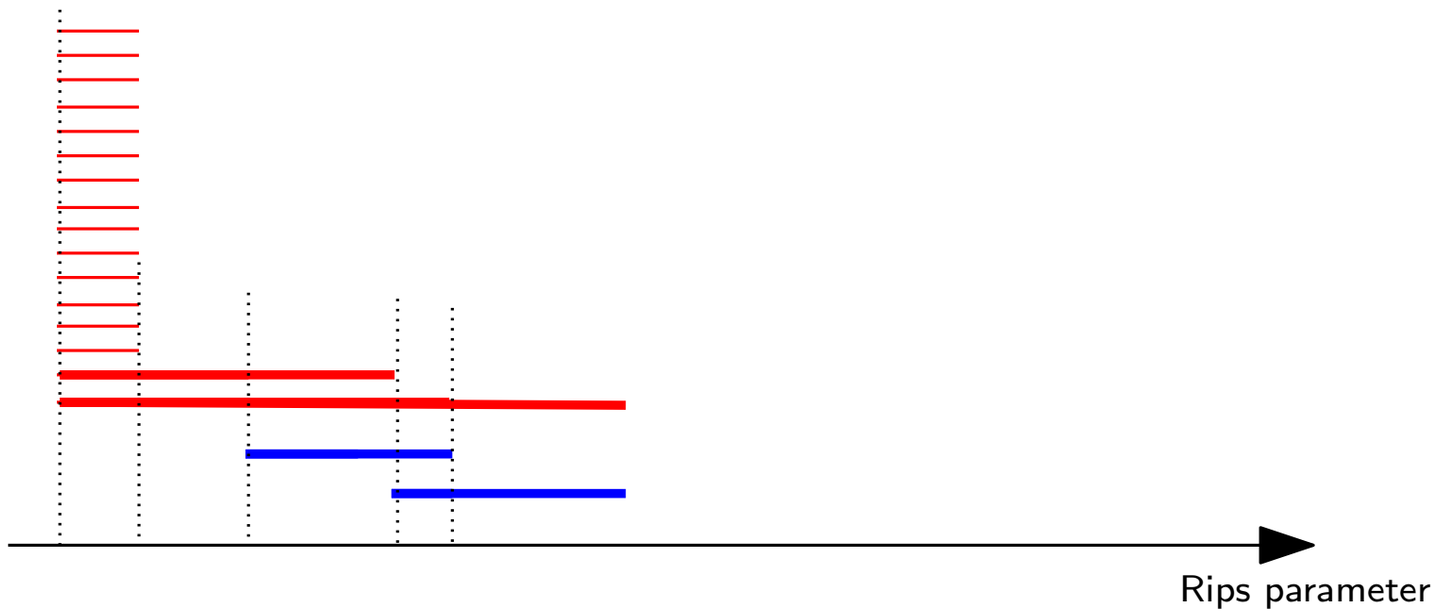





- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

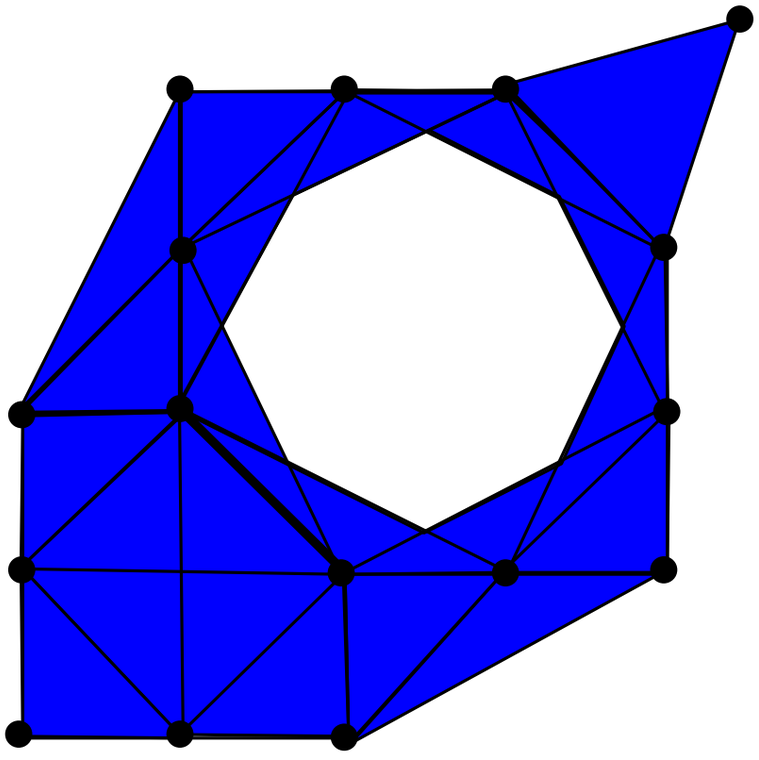
# Persistent homology of filtered complexes

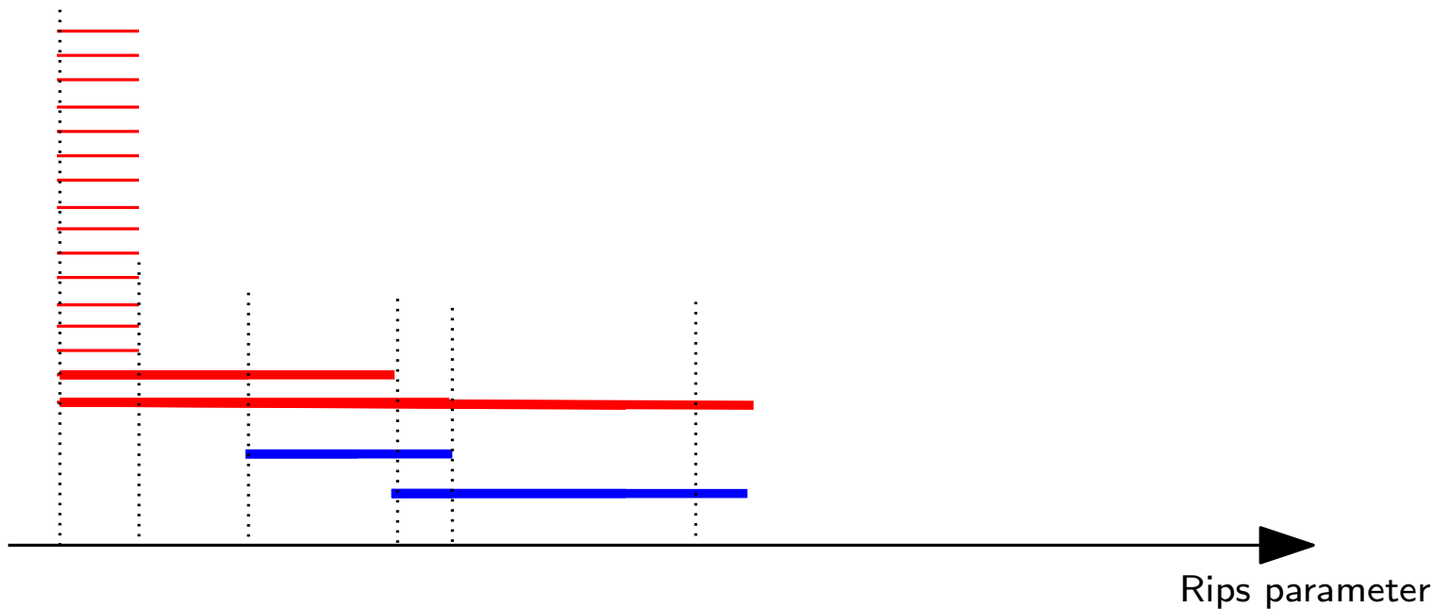





- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

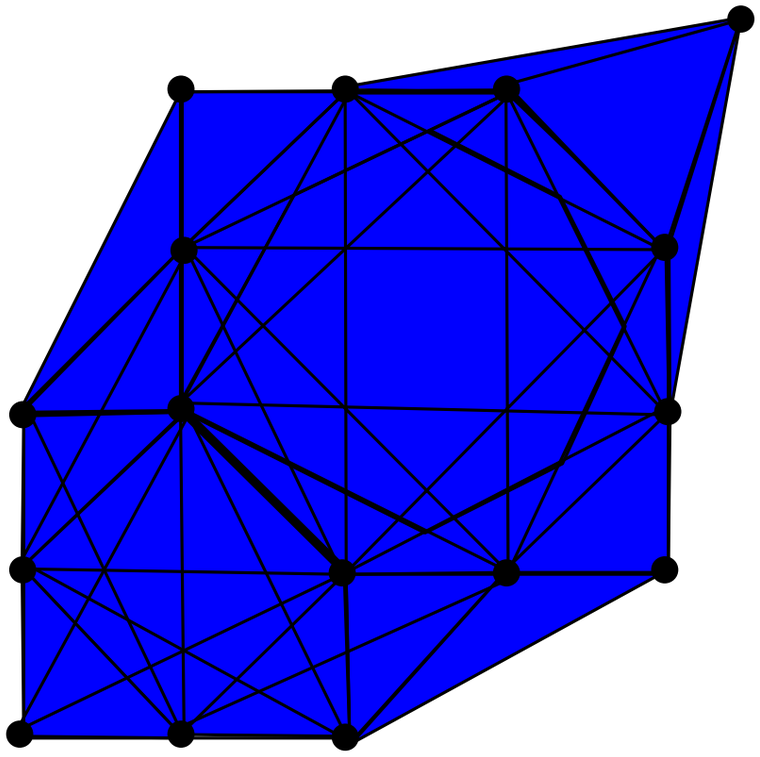
# Persistent homology of filtered complexes

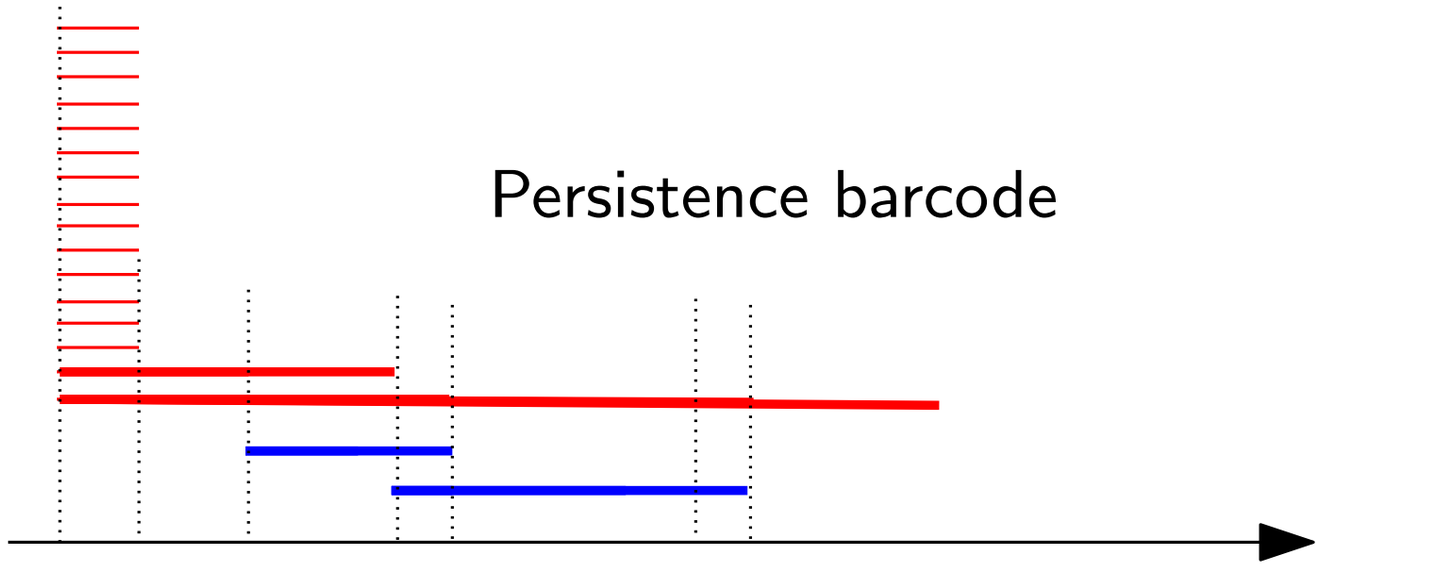


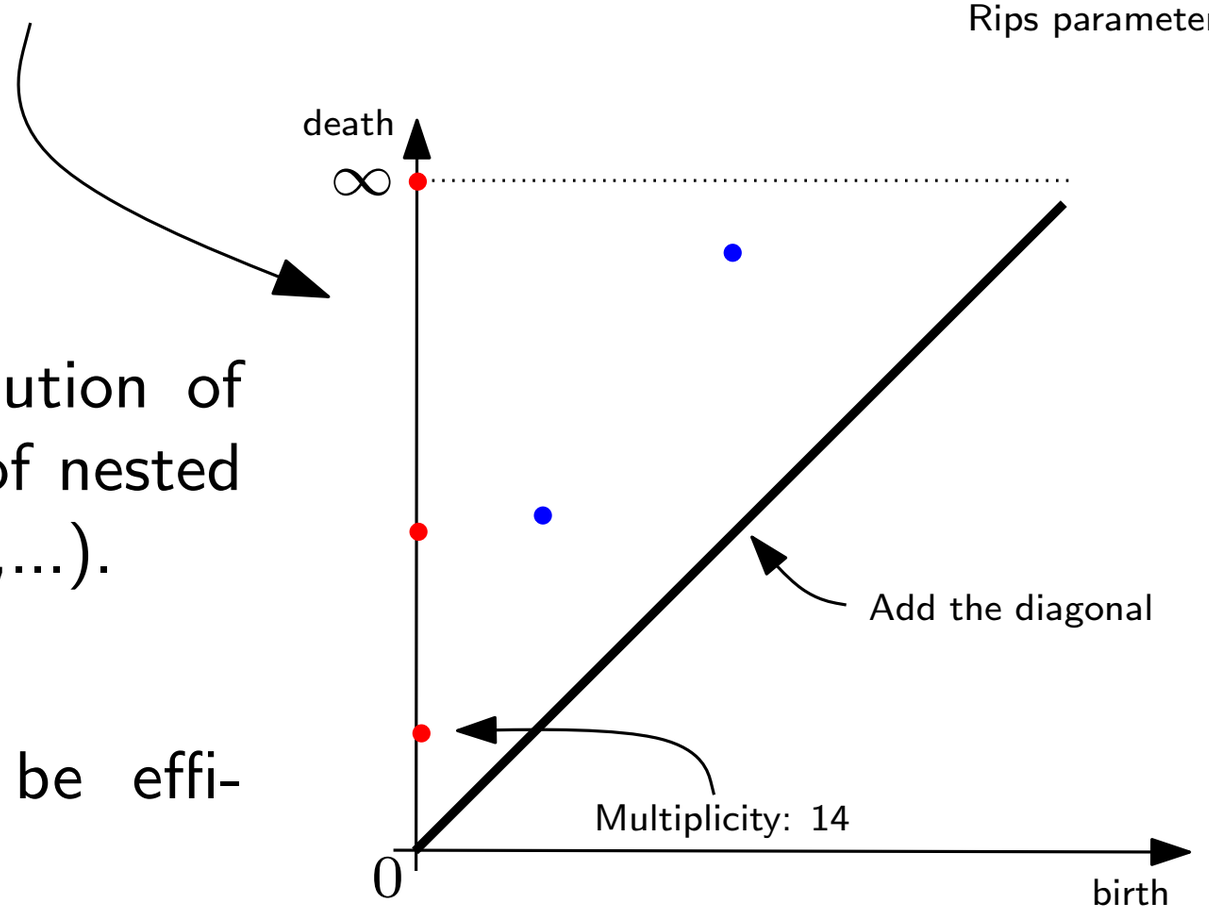



- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

# Persistent homology of filtered complexes






Persistence diagram

- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

# Stability properties

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

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

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

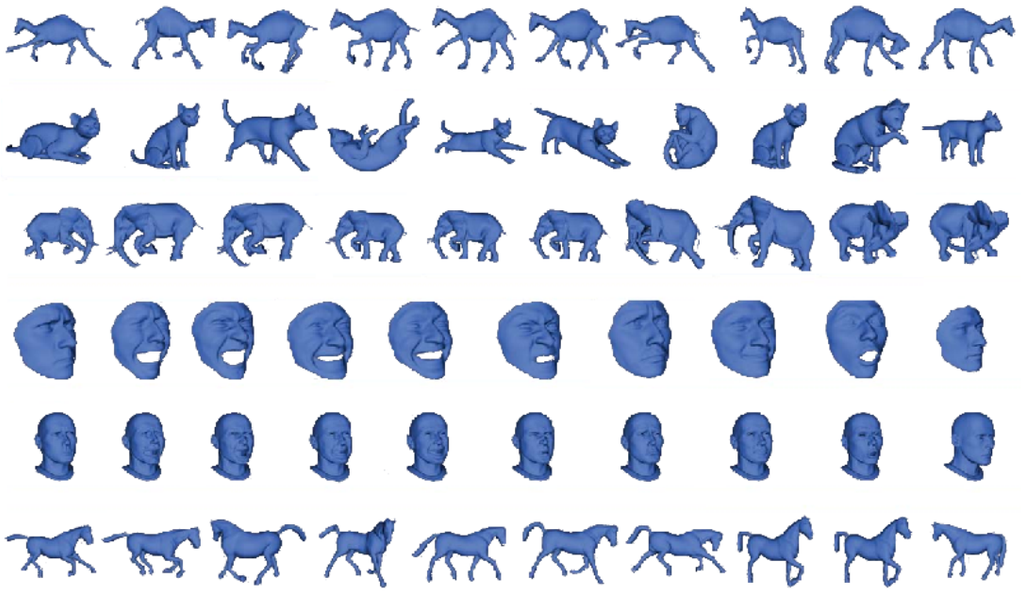
Bottleneck distance: a matching distance between diagrams

Gromov-Hausdorff distance

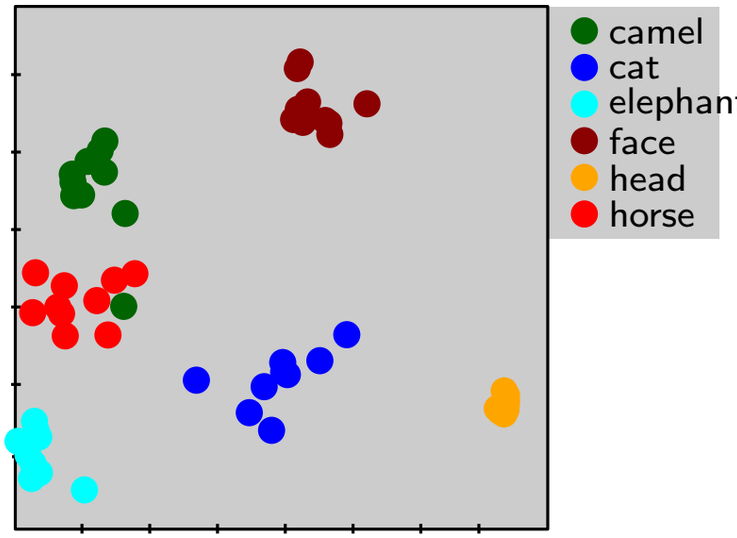
**Rem:** this a particular case of a more general theorem [C.-de Silva-Oudot 2013].

# Stability properties

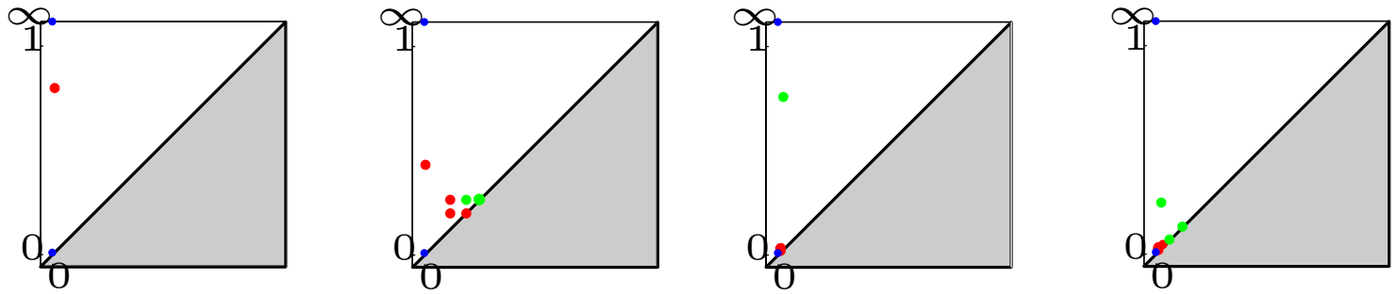
**Example:** Application to non rigid shape classification.



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



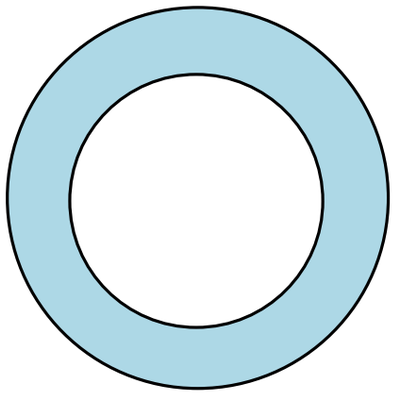

Use the metric on the space of persistence diagrams.




- 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.

$\rightarrow$ Other applications in image classifications, object recognition, clustering,...

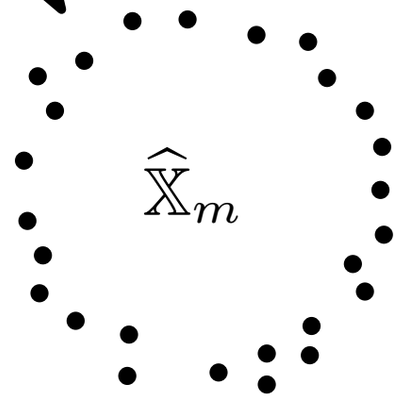
# Convergence rates of persistence diagrams



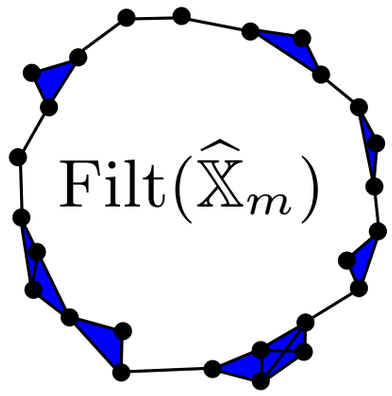
$(\mathbb{M}, \rho)$ metric space

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

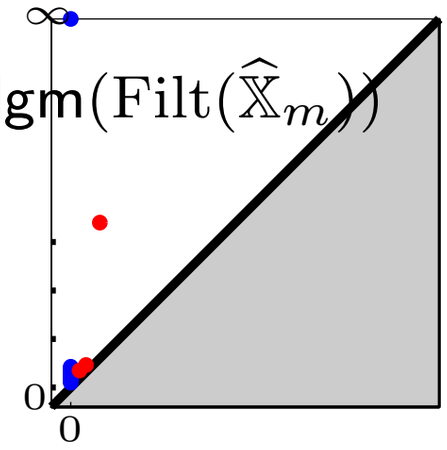
Sample $m$ points according to $\mu$.




Build topol. structure




Persistent homology

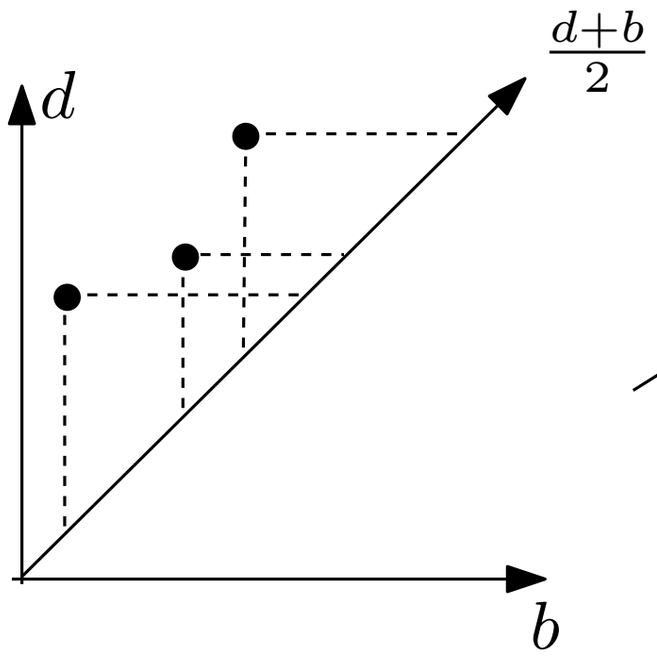


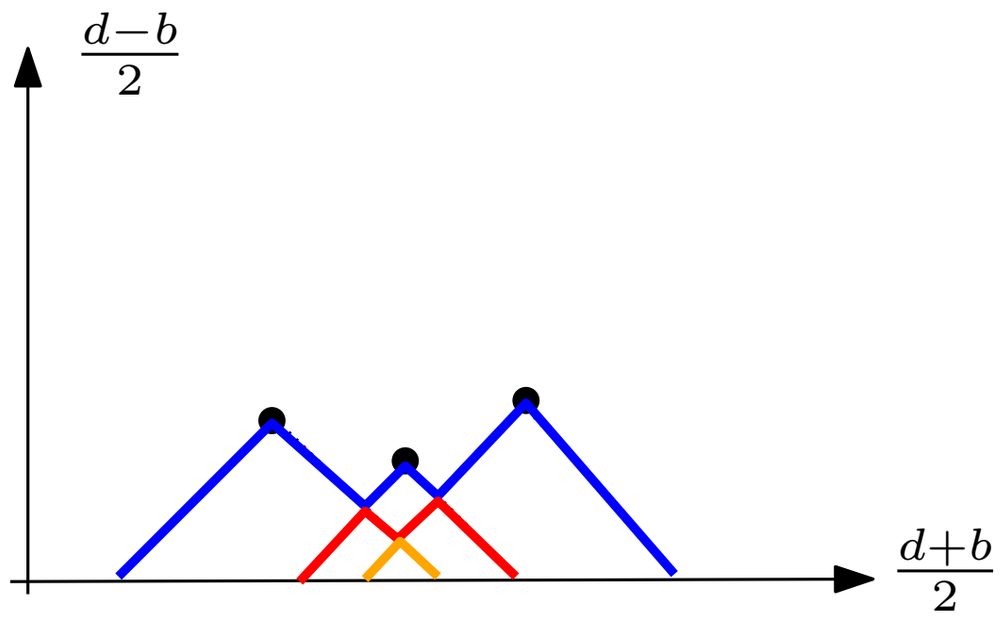

Assume that $\mu$ is $(a, b)$-standard: $\forall x \in \mathbb{X}_\mu,\ \forall r > 0,\ \mu(B(x, r)) \geq \min(ar^b, 1)$.

$$\mathbb{E}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Rips}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Rips}(\widehat{\mathbb{X}}_m)))\right] \leq C\left(\frac{\ln m}{m}\right)^{1/b}$$

The convergence rate is optimal among $(a, b)$ standard measures, whatever the choice of the estimator of $\mathrm{dgm}(\mathrm{Rips}(\mathbb{X}_\mu))$ (minimax convergence rate).

# To do more statistics: persistence landscapes




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

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

Persistence landscape [Bubenik 2012]:

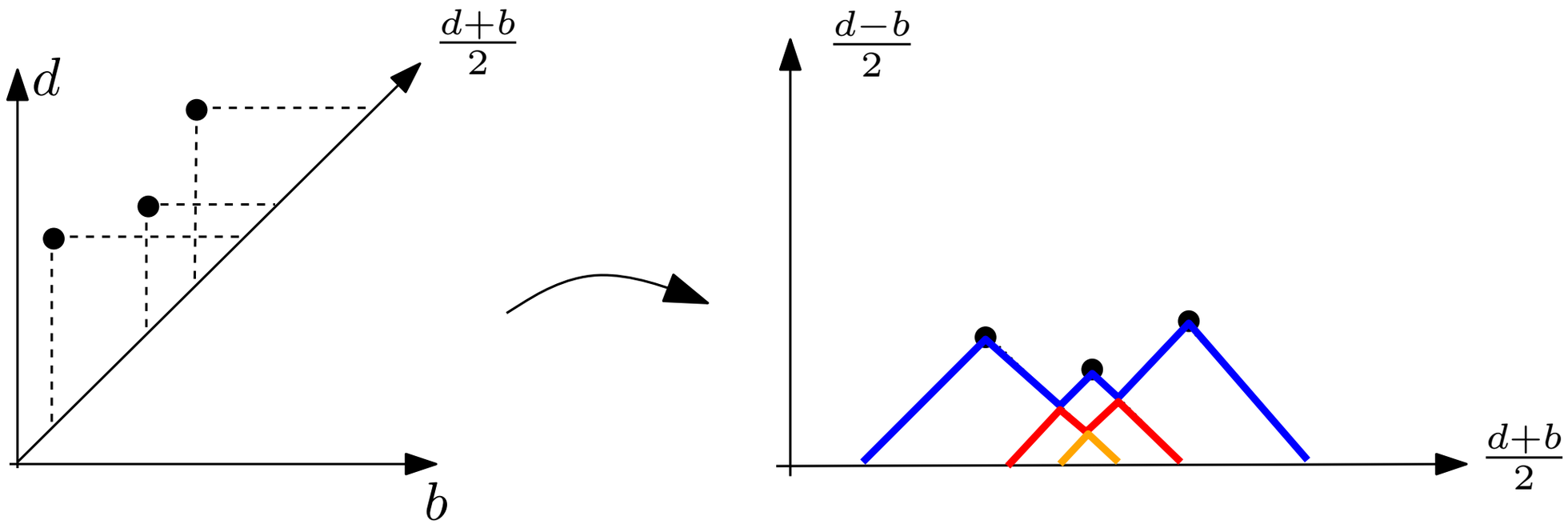
$$\Lambda_p(t) = \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases}$$

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

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

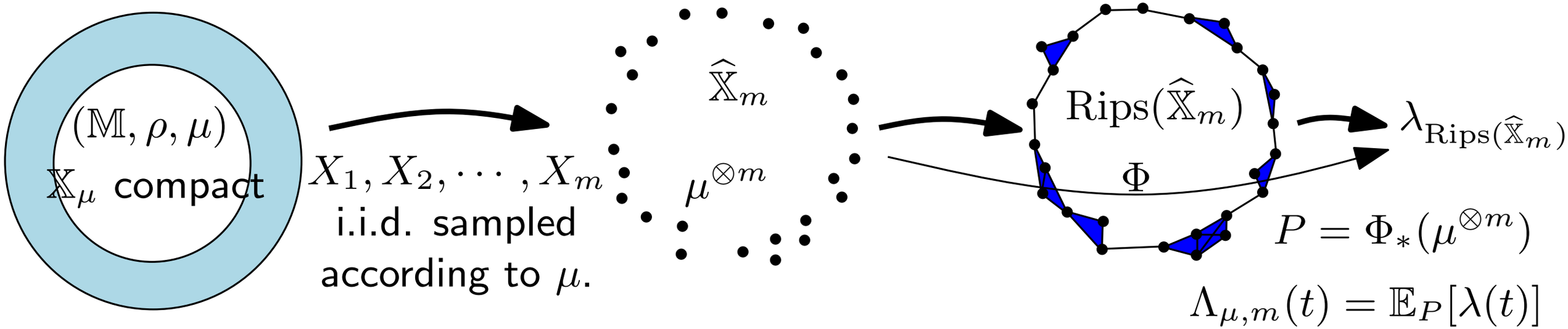
**Stability:** For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $|\lambda_D(k,t) - \lambda_{D'}(k,t)| \leq d_B(D, D')$.

# To do more statistics: 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.

# (Sub)sampling and stability of expected landscapes




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

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

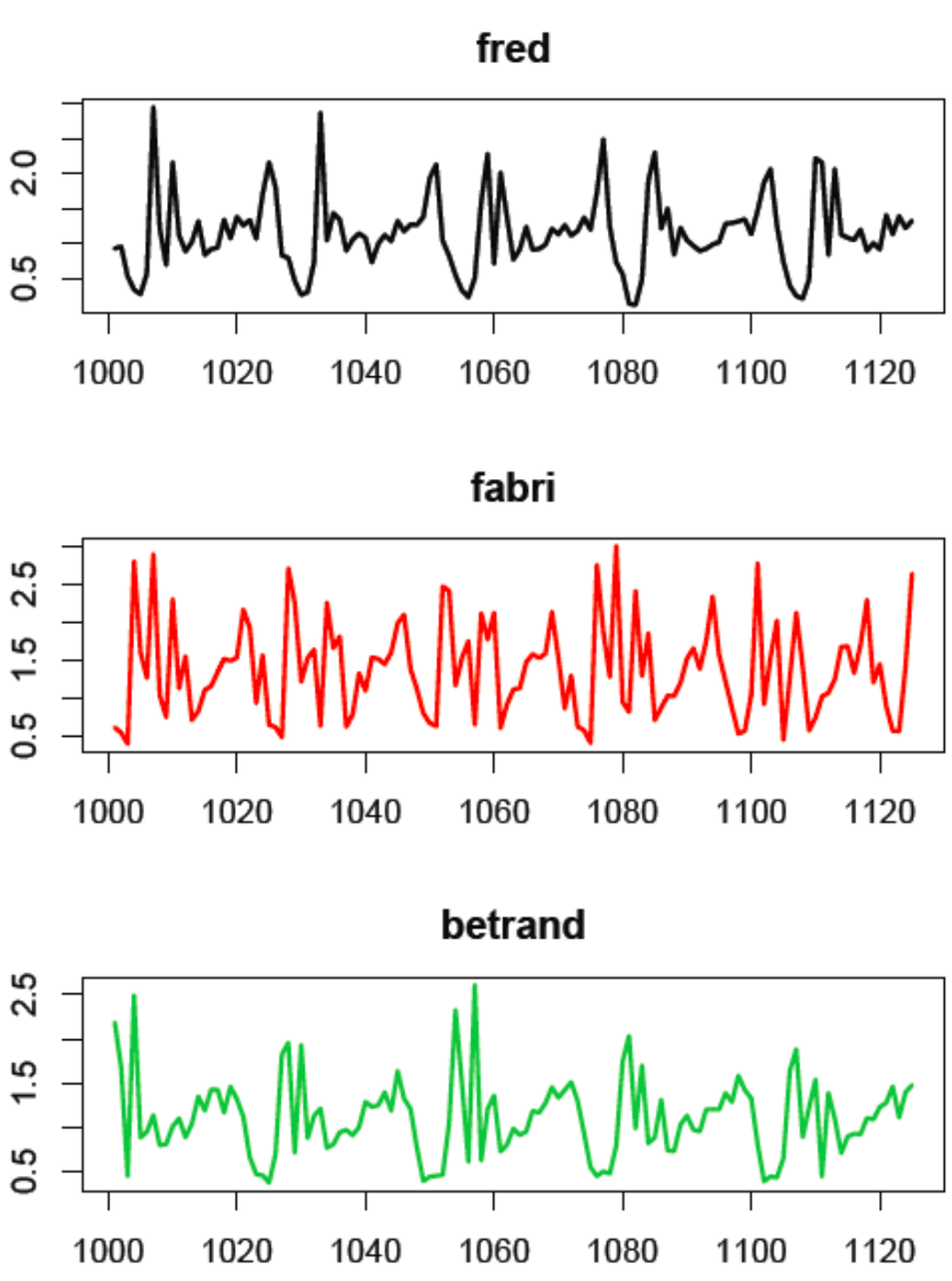
where $W_p$ denotes the Wasserstein distance with cost function $\rho(x, y)^p$.

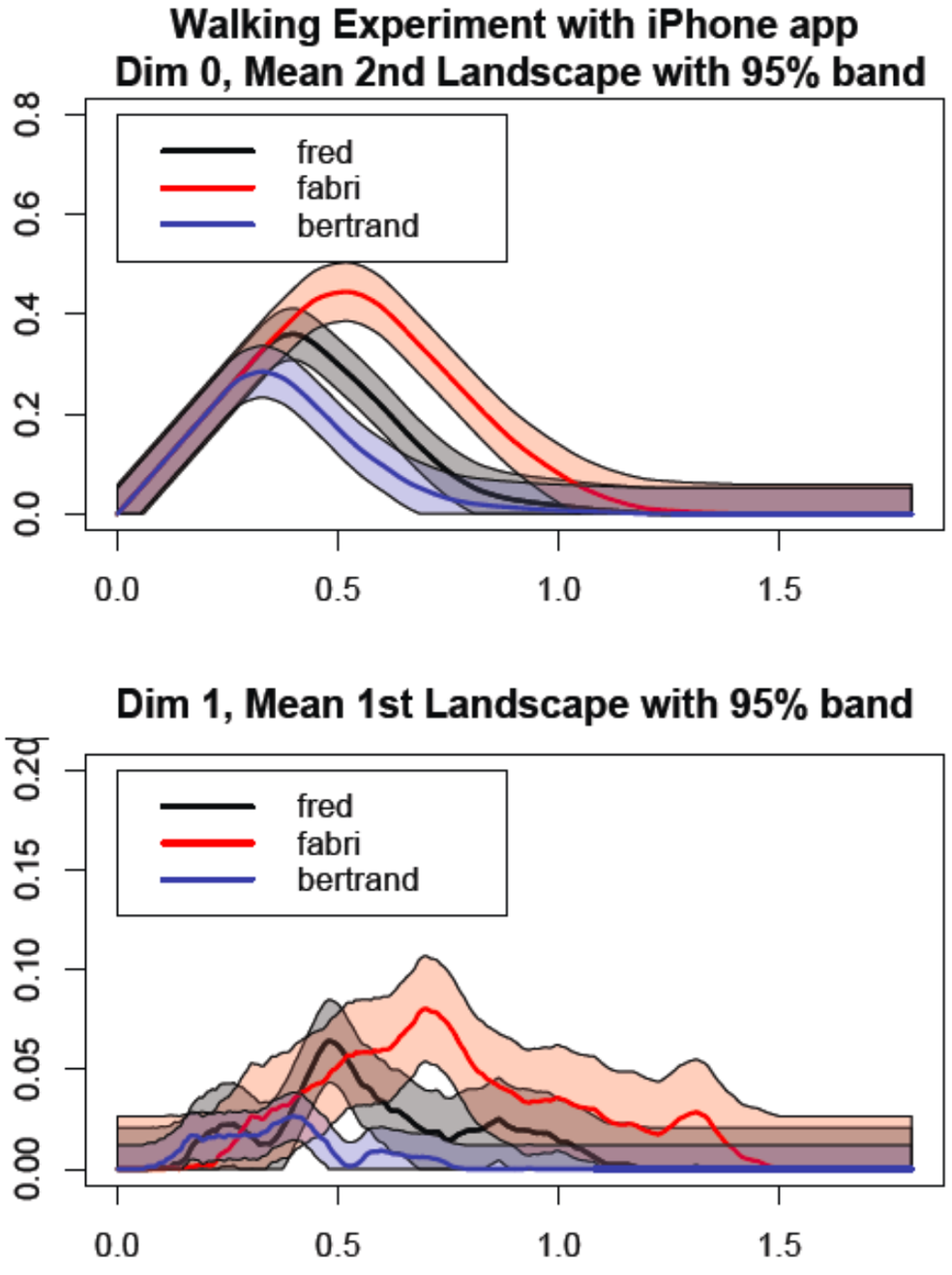
**Consequences:**

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

# (Sub)sampling and stability of expected landscapes

**(Toy) Example:** Accelerometer data from smartphone.




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

# Conclusion

- Persistent diagrams of geometric complexes built on top of data provide a very general, flexible way to infer relevant multiscale topological information.
- Although they live in a "non-friendly" metric space, persistence diagrams have good statistical properties.
- The convergence results are indeed more general:
  - extend to other families of filtered simplicial complexes (Čech-complexes, witness complexes,...)
  - extend to non-metric spaces endowed with a similarity measure.
- Subsampling, averaging:
  $\rightarrow$ robust topological inference under perturbations of the measure $\mu$.
  $\rightarrow$ very fast and easy to parallelize computation of topological feature for huge data.

# Conclusion

- Persistent diagrams of geometric complexes built on top of data provide a very general, flexible way to infer relevant multiscale topological information.
- Although they live in a "non-friendly" metric space, persistence diagrams have good statistical properties.
- The convergence results are indeed more general:
  - extend to other families of filtered simplicial complexes (Čech-complexes, witness complexes,...)
  - extend to non-metric spaces endowed with a similarity measure.
- Subsampling, averaging:
  $\rightarrow$ robust topological inference under perturbations of the measure $\mu$.
  $\rightarrow$ very fast and easy to parallelize computation of topological feature for huge data.

Thank you for your attention!

**References:**


- F. Chazal, M. Glisse, C. Labruère, B. Michel, Convergence rates for persistence diagram estimation in Topological Data Analysis, in Int. Conf. on Machine Learning 2014 (ICML 2014).
- C. Li, M. Ovsjanikov, F. Chazal, Persistence-based Structural Recognition, in proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014).
- F. Chazal, B. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, Stochastic Convergence of Persistence Landscapes and Silhouettes, in ACM Symposium on Computational Geometry 2014.
- F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, L. Wasserman, Subsampling Methods for Persistent Homology, arXiv:1406.1901, June 2014.
- F. Chazal, V. de Silva, S. Oudot, Persistence Stability for Geometric complexes, Geometria Dedicata 2014 (online first Dec. 2013).
- F. Chazal, V. de Silva, M. Glisse, S. Oudot, The Structure and Stability of Persistence Modules, arXiv:1207.3674, July 2012.