Journées Modélisation Aléatoire et Statistique
Grenoble, August 30th, 2016

Topological Data Analysis
Focus: Mapper

Steve Oudot — Inria Saclay – Île-de-France

Resources:

- H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Society, 2010.
- S.O. *Persistence Theory: from Quiver Representations to Data Analysis*. AMS Mathematical Surveys and Monographs (209), 2015.

Journées Modélisation Aléatoire et Statistique
Grenoble, August 30th, 2016

Topological Data Analysis ①

Focus: Mapper ②

Steve Oudot — Inria Saclay – Île-de-France

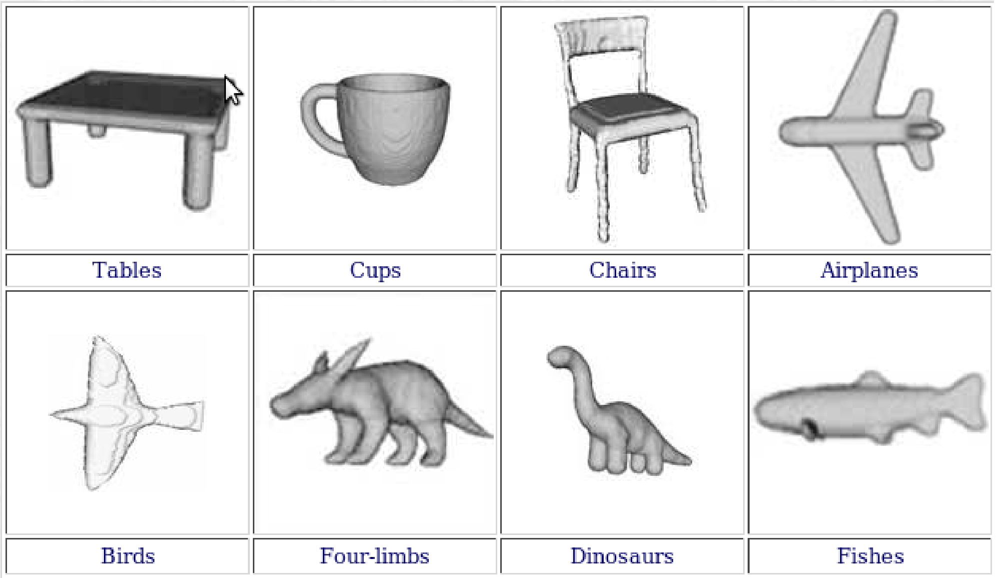
Resources:

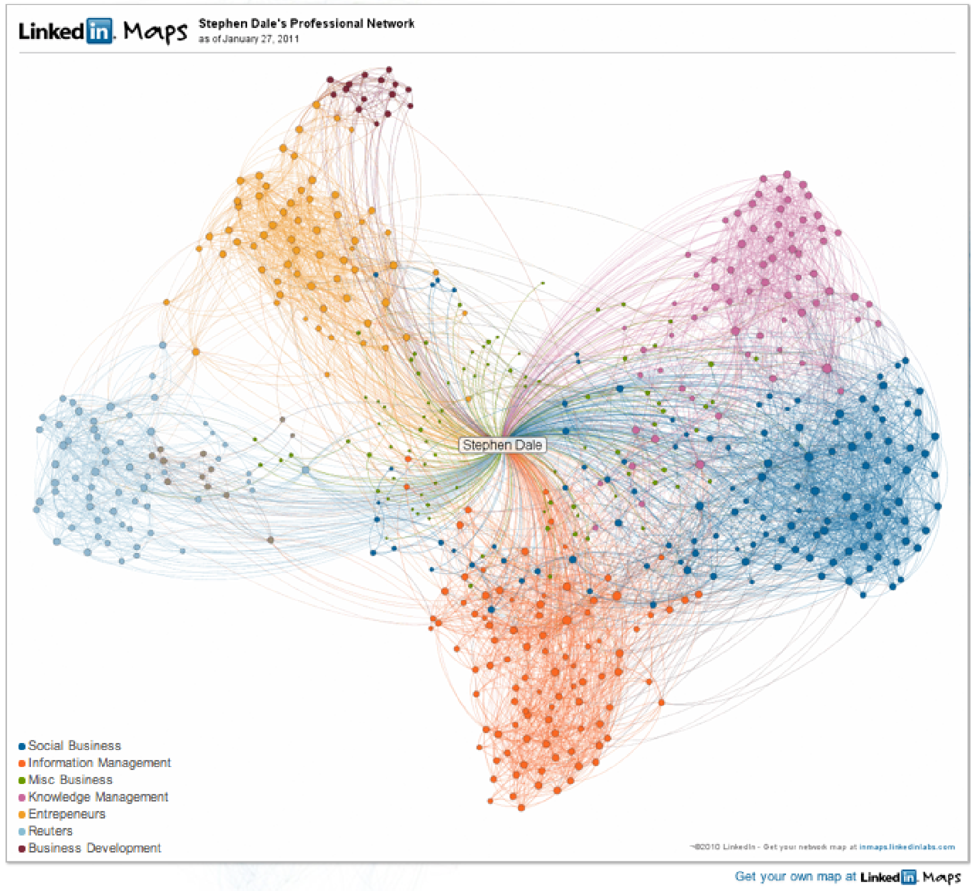
- H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Society, 2010.
- S.O. *Persistence Theory: from Quiver Representations to Data Analysis*. AMS Mathematical Surveys and Monographs (209), 2015.

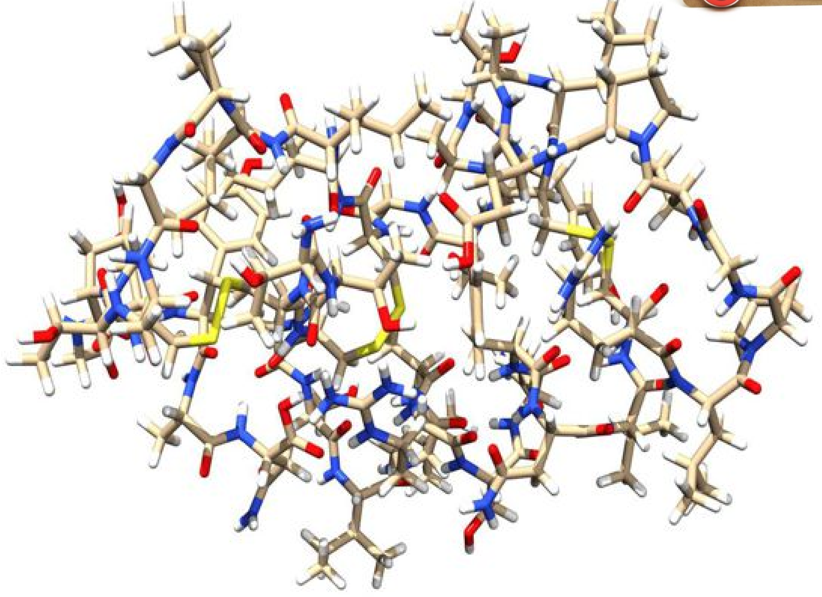
Geometric Data

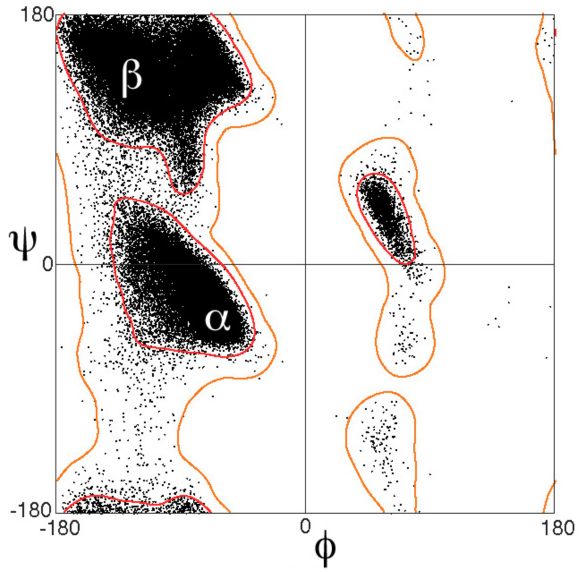
Input: point cloud equipped with a metric or (dis-)similarity measure

data point $\equiv$ image/patch, geometric shape, protein conformation, patient, LinkedIn user...





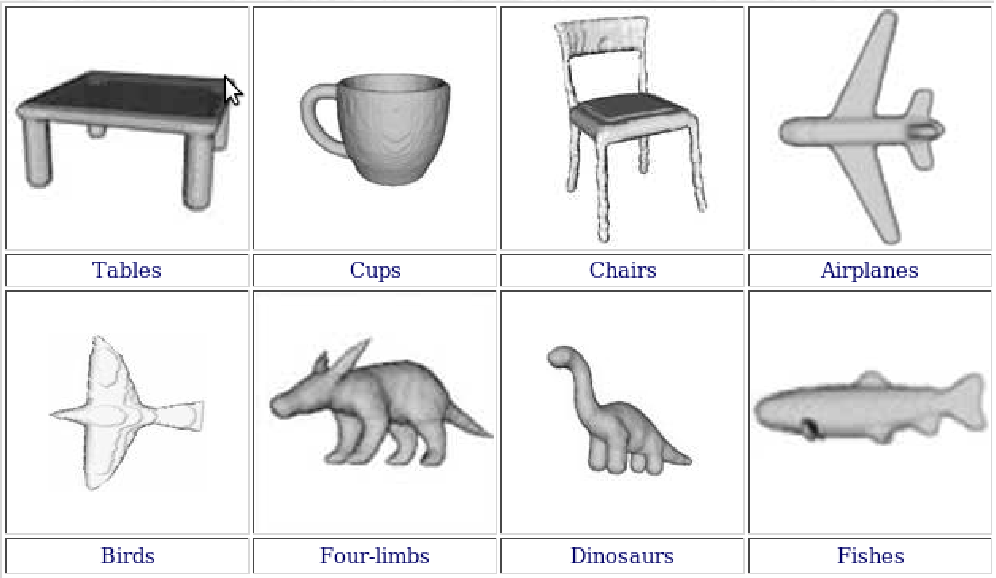


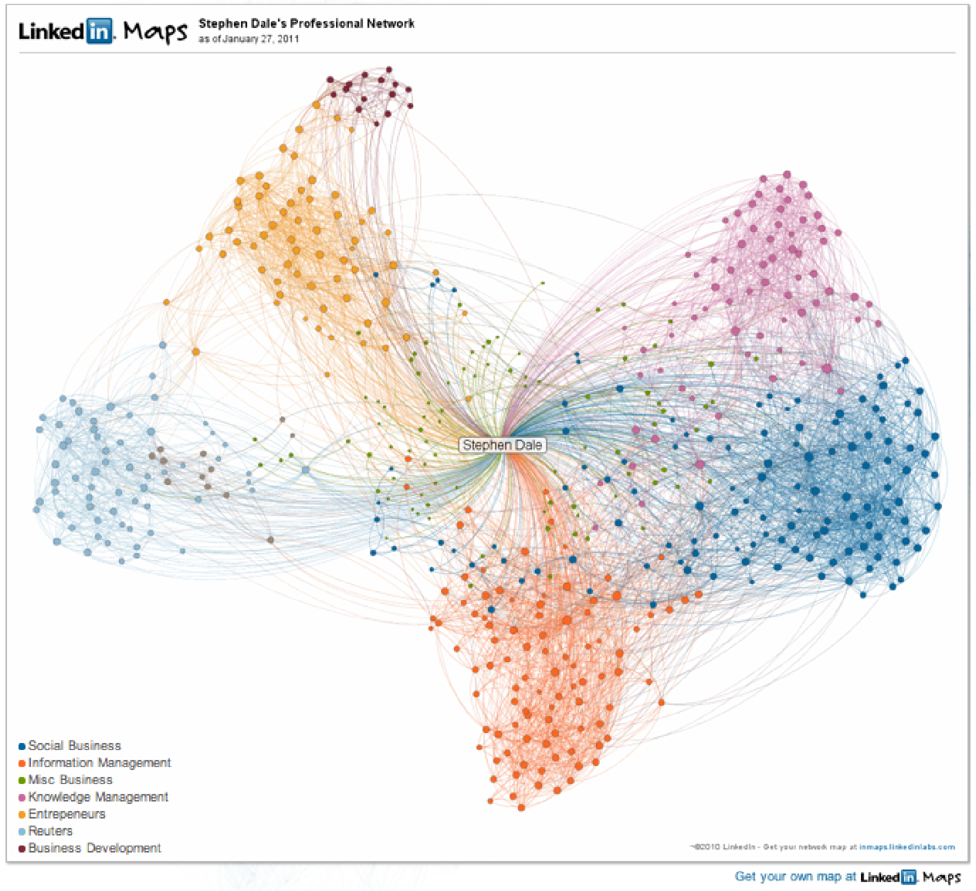


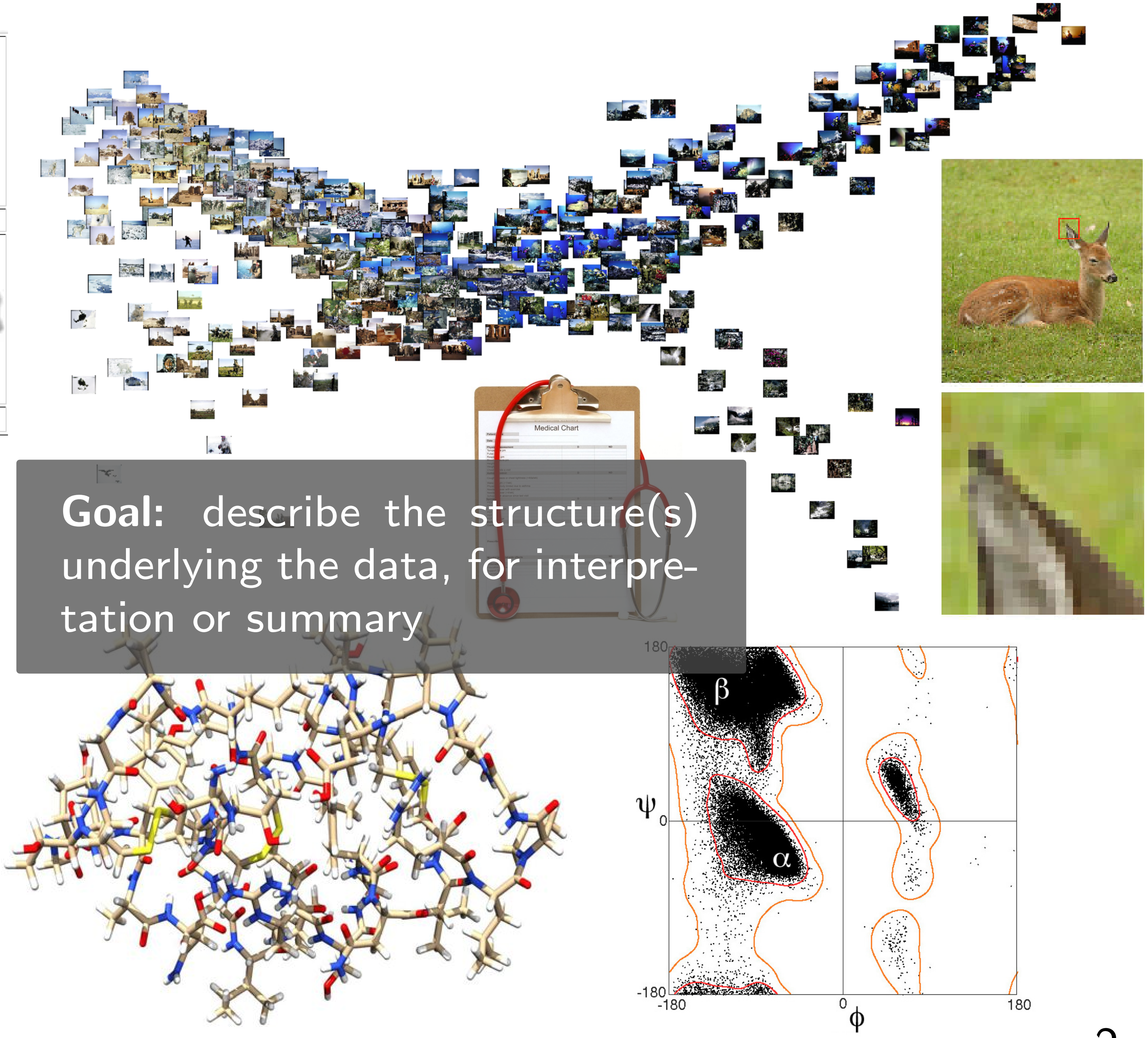
Geometric Data

Input: point cloud equipped with a metric or (dis-)similarity measure

data point ≡ image/patch, geometric shape, protein conformation, patient, LinkedIn user...

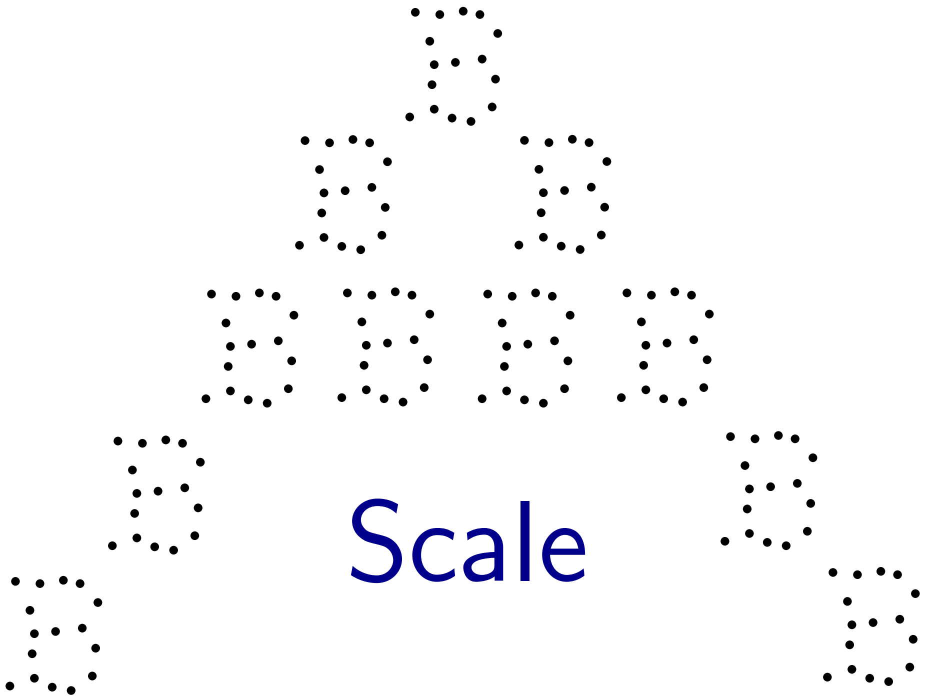




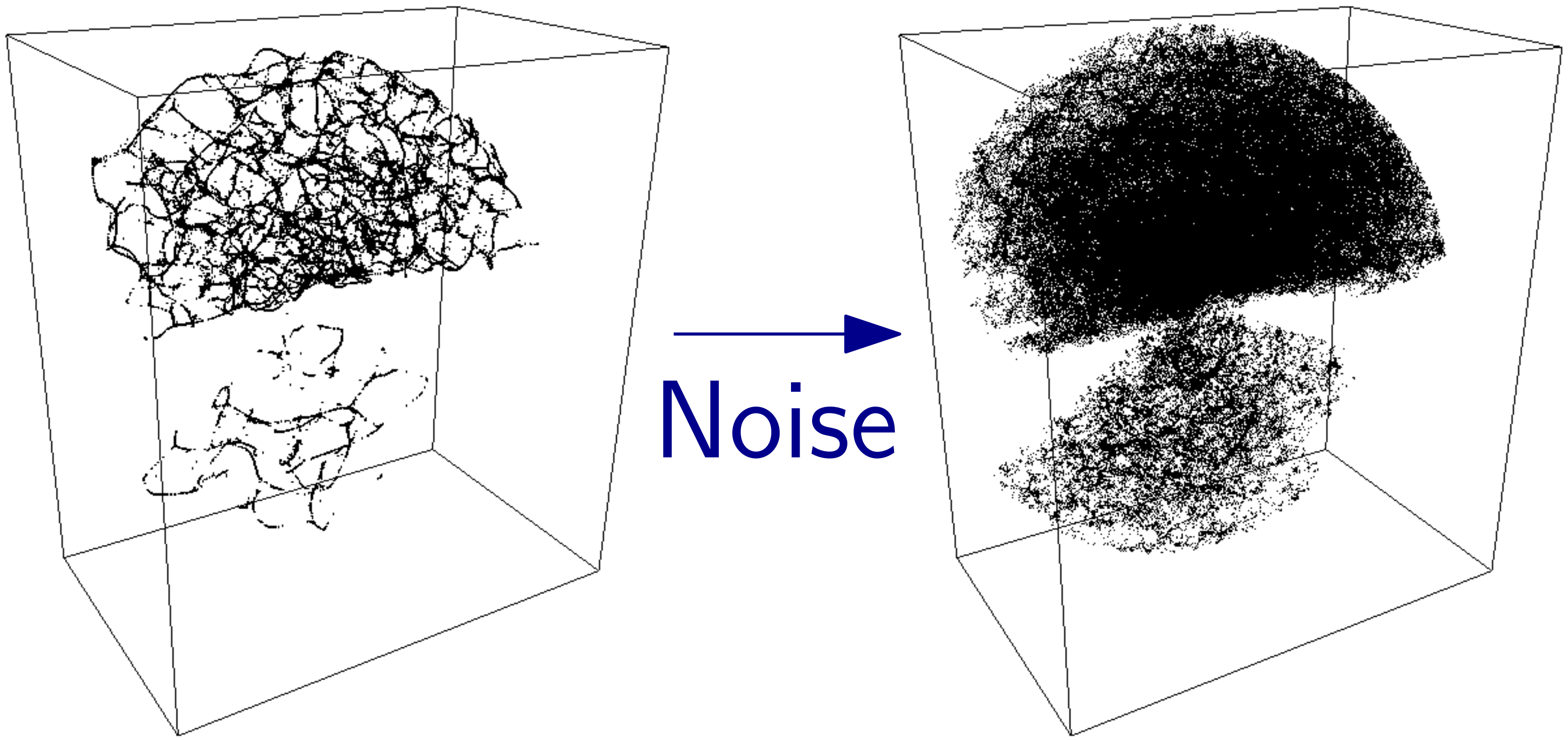


Goal: describe the structure(s) underlying the data, for interpretation or summary

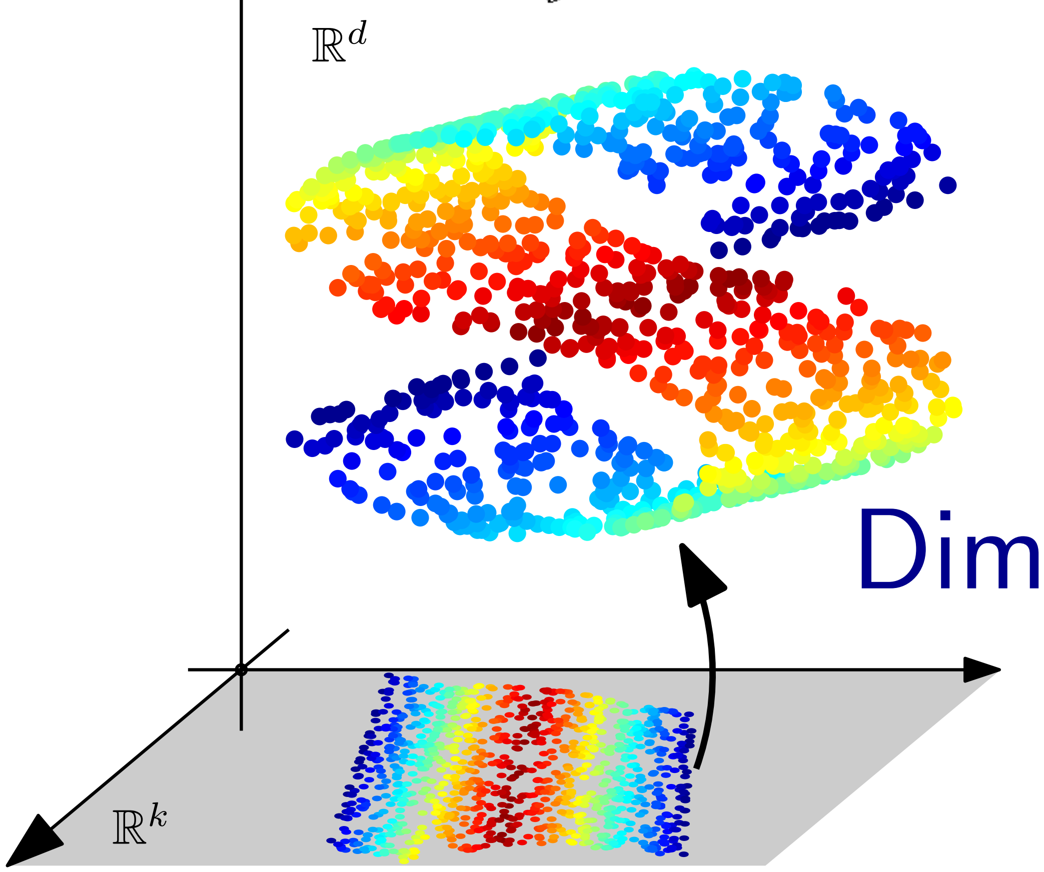
Challenges



Scale

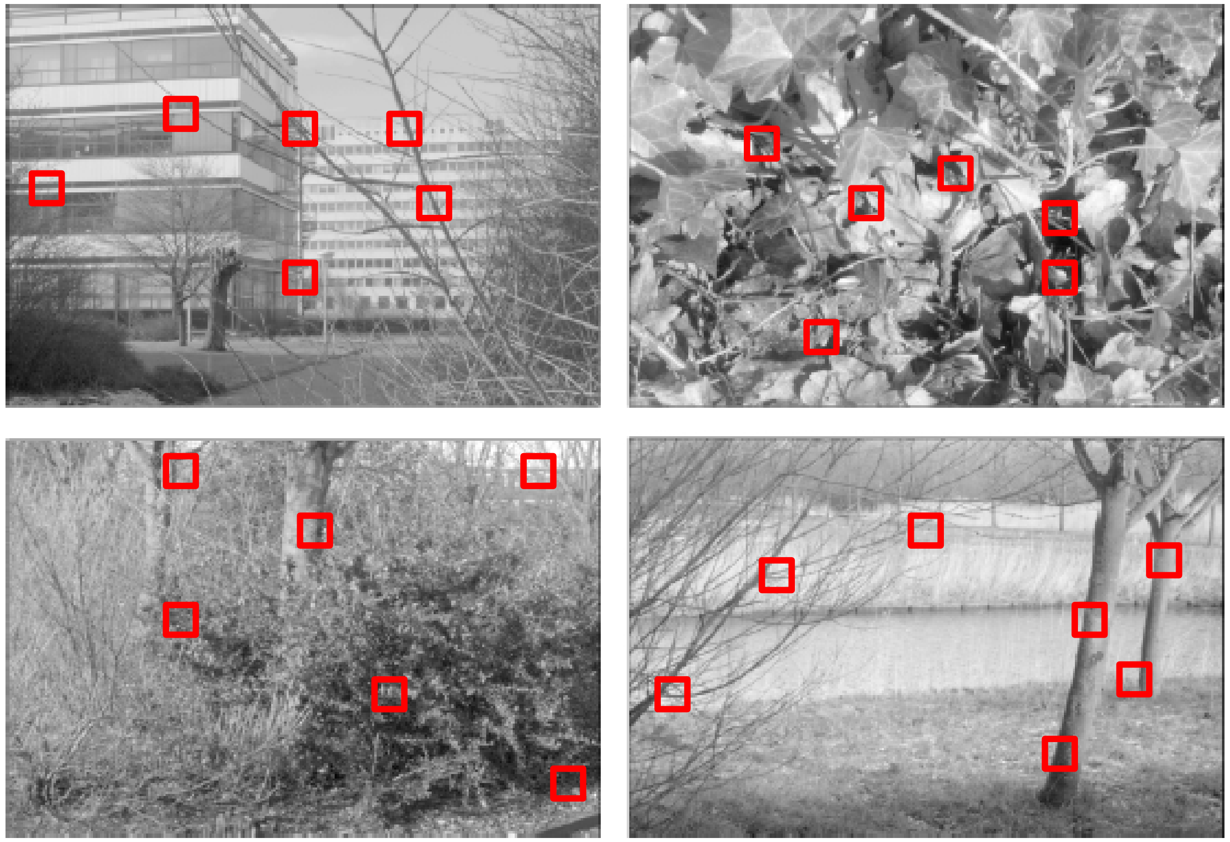


Noise



Dimensionality

Challenges

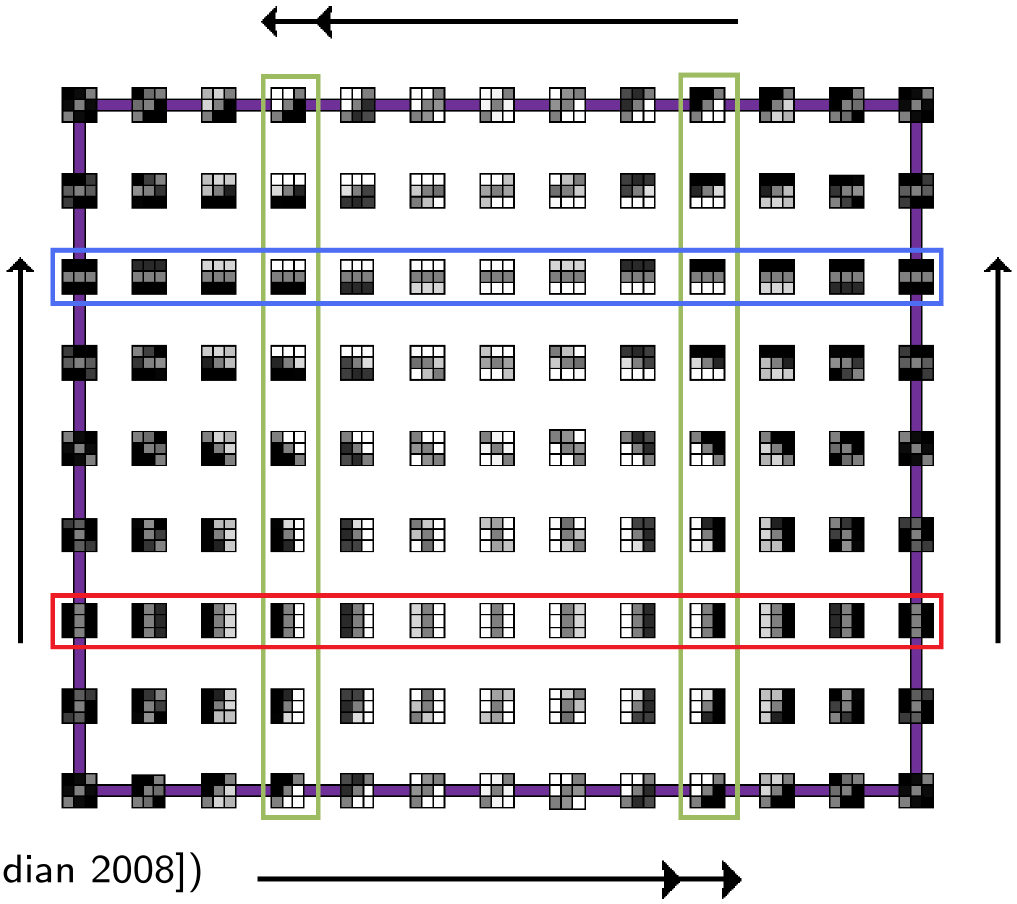


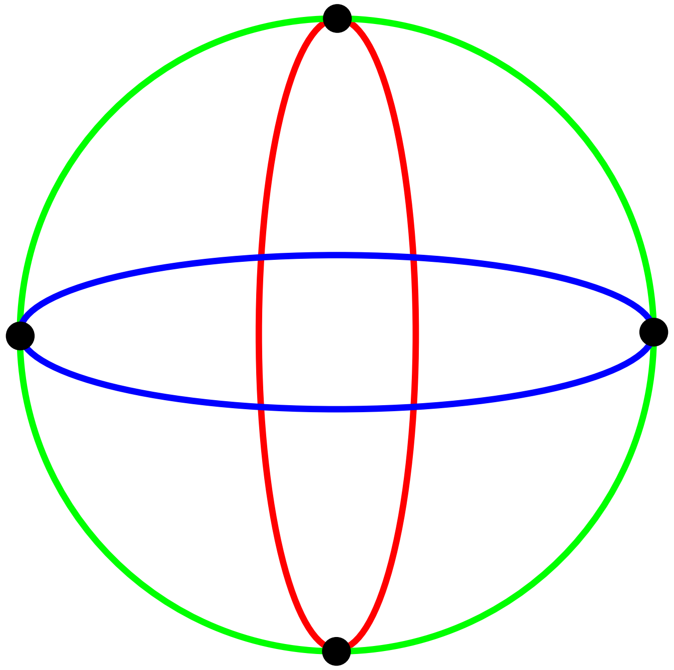
4 million data points in $\mathbb{R}^9$

(source: [Lee, Pederson, Mumford 2003])

Motivation: study cognitive representation of space of images

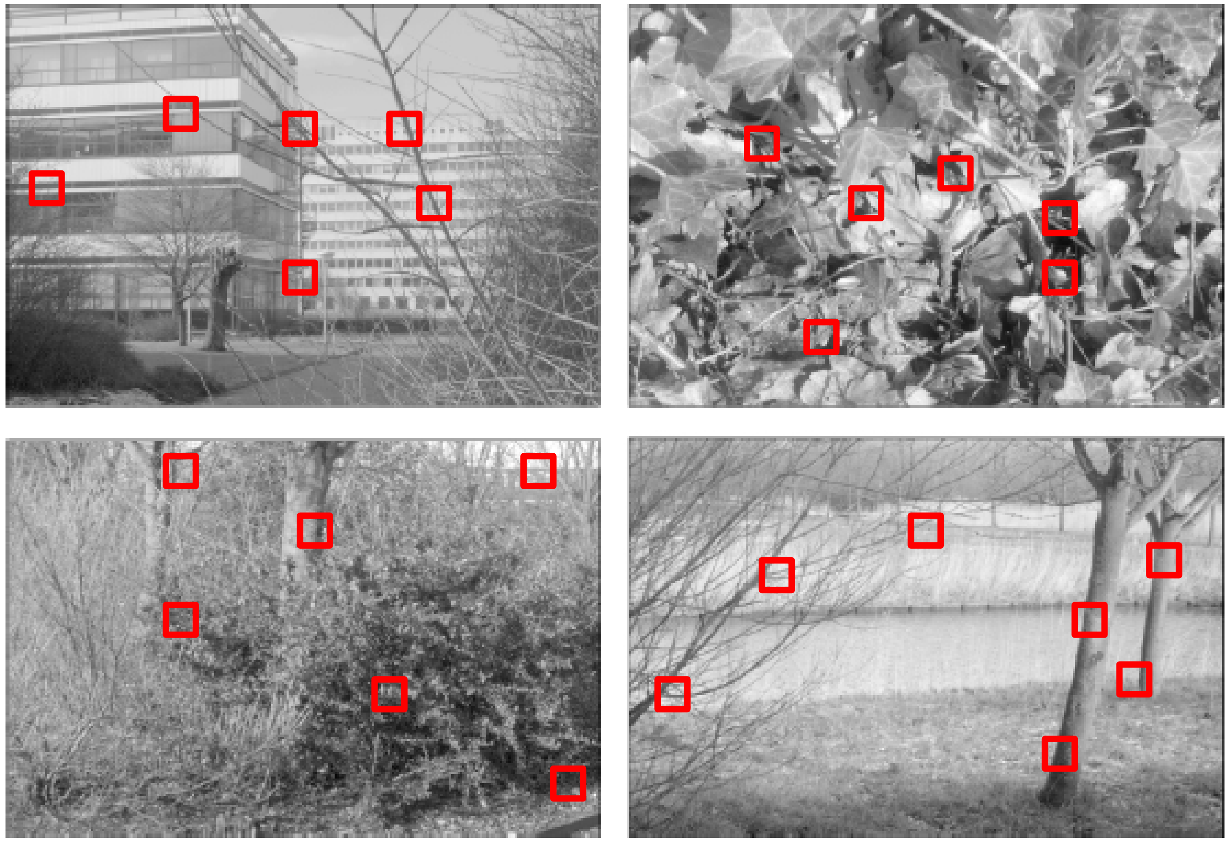
Topology





(source: [Carlsson, Ishkhanov, de Silva, Zomorodian 2008])

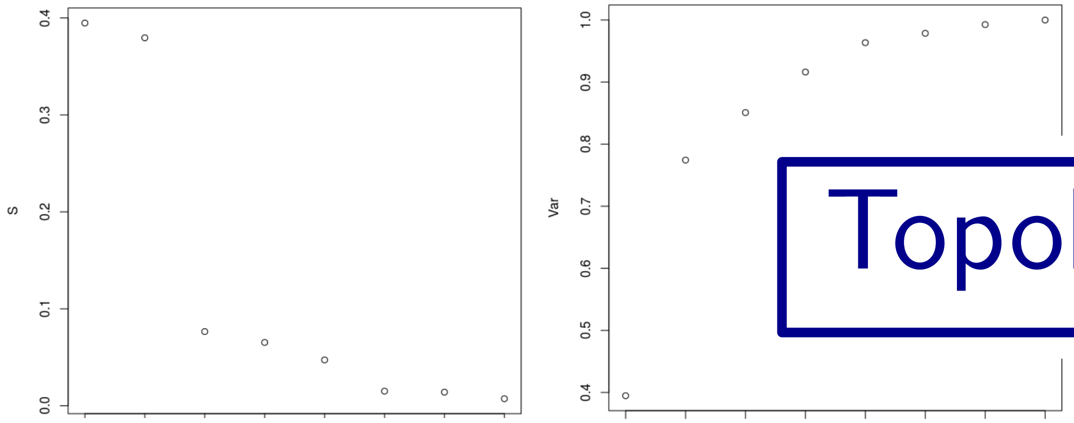
Challenges



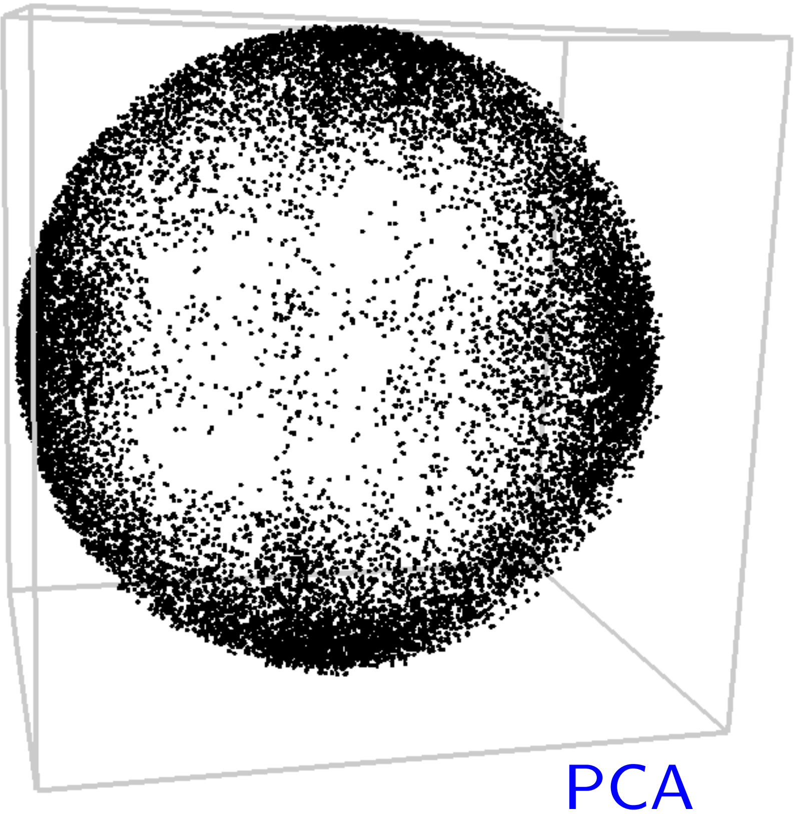
4 million data points in $\mathbb{R}^9$

(source: [Lee, Pederson, Mumford 2003])

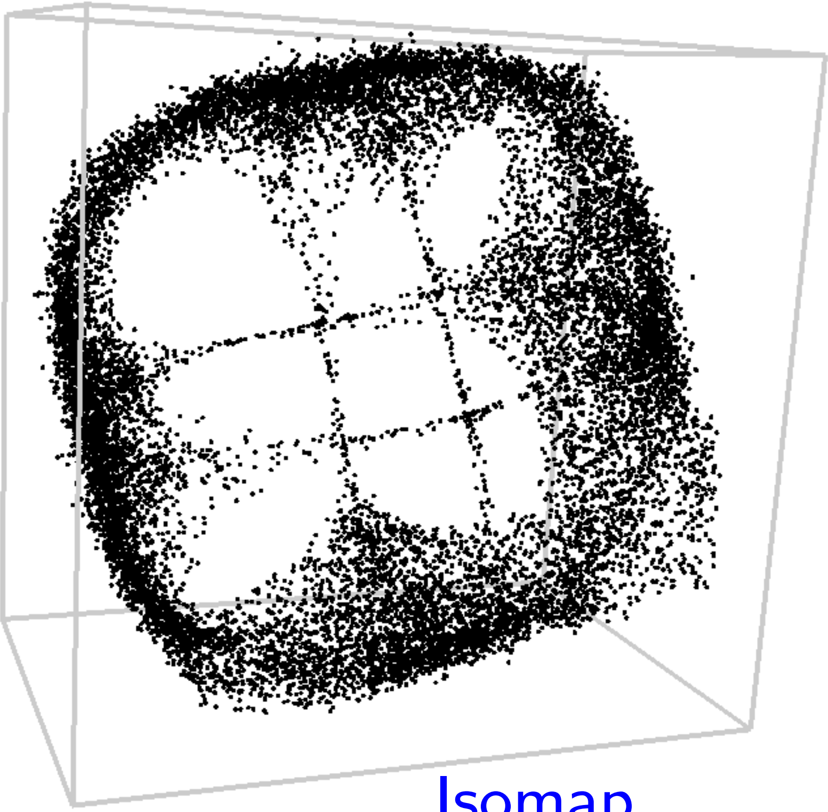
Motivation: study cognitive representation of space of images



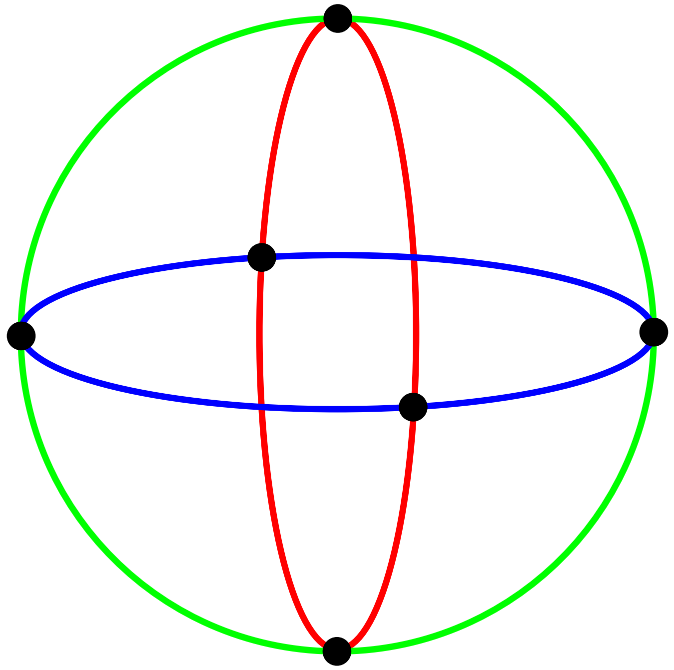
Topology



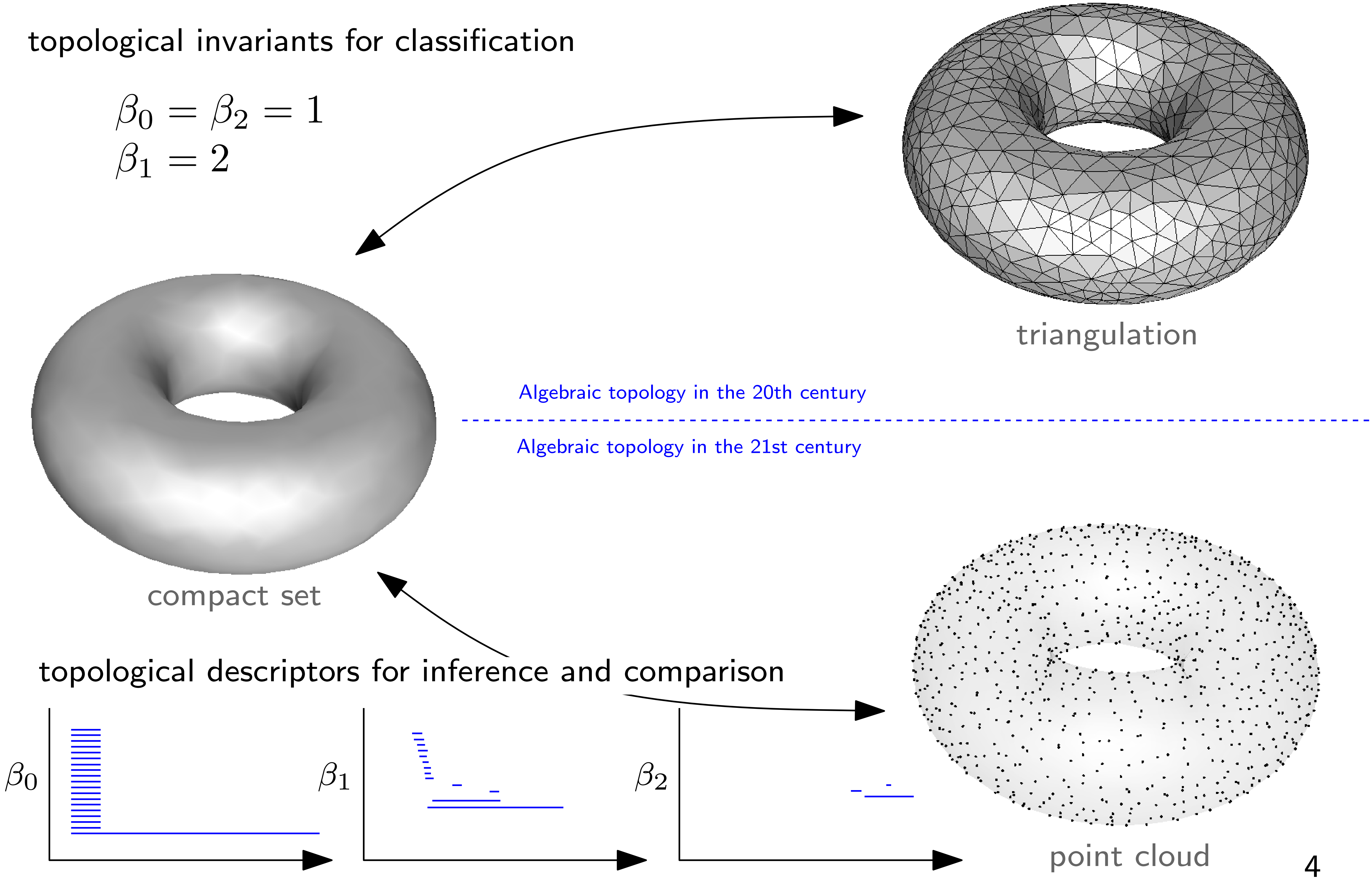
PCA



Isomap



Topological Data Analysis (TDA)



topological invariants for classification

$$\beta_0 = \beta_2 = 1$$
$$\beta_1 = 2$$

triangulation

compact set

Algebraic topology in the 20th century

Algebraic topology in the 21st century

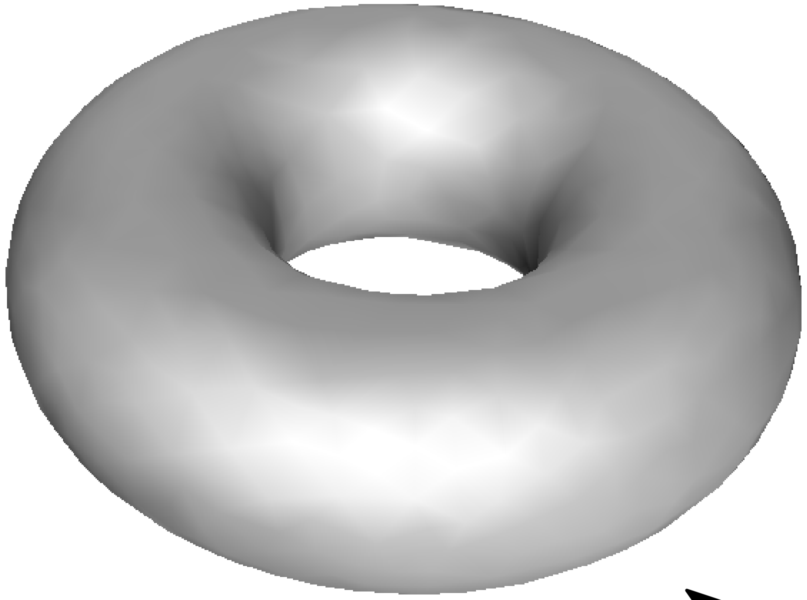
point cloud

topological descriptors for inference and comparison

Topological Data Analysis (TDA)

Properties of topological descriptors:

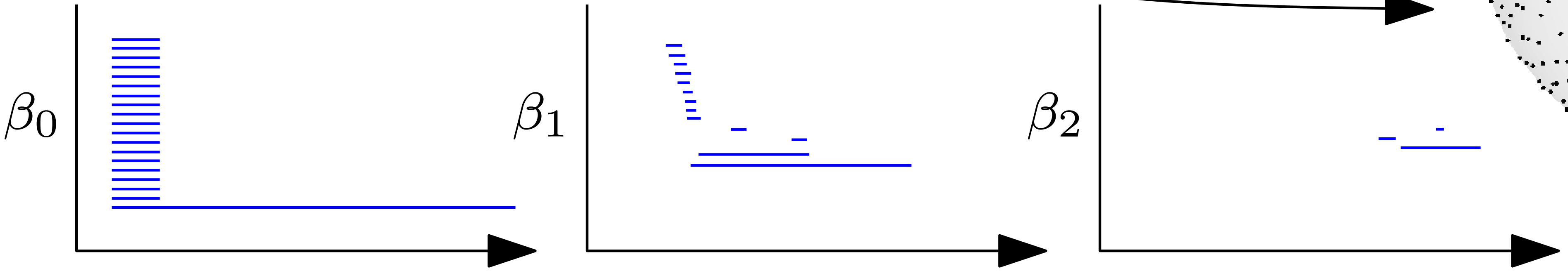
- invariant under coordinate changes
- stable with respect to perturbations
- informative

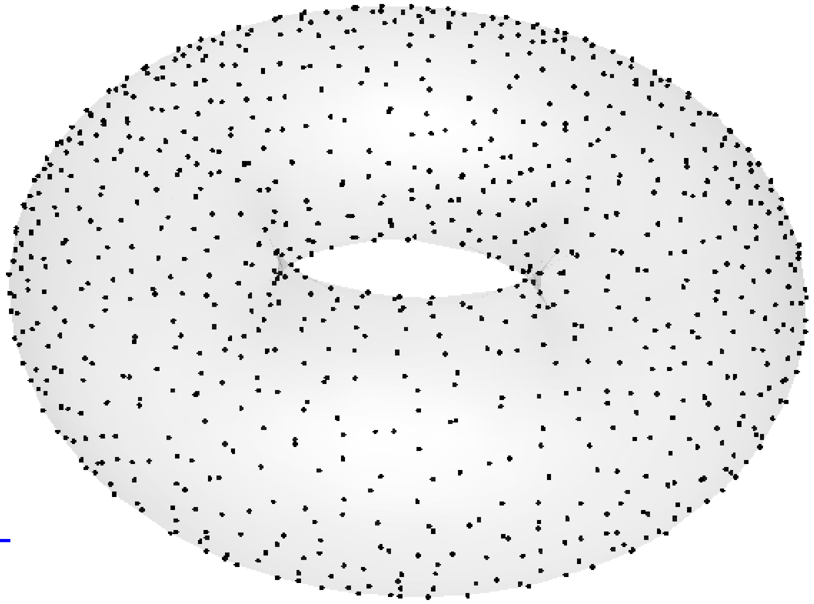


compact set

Algebraic topology in the 21st century

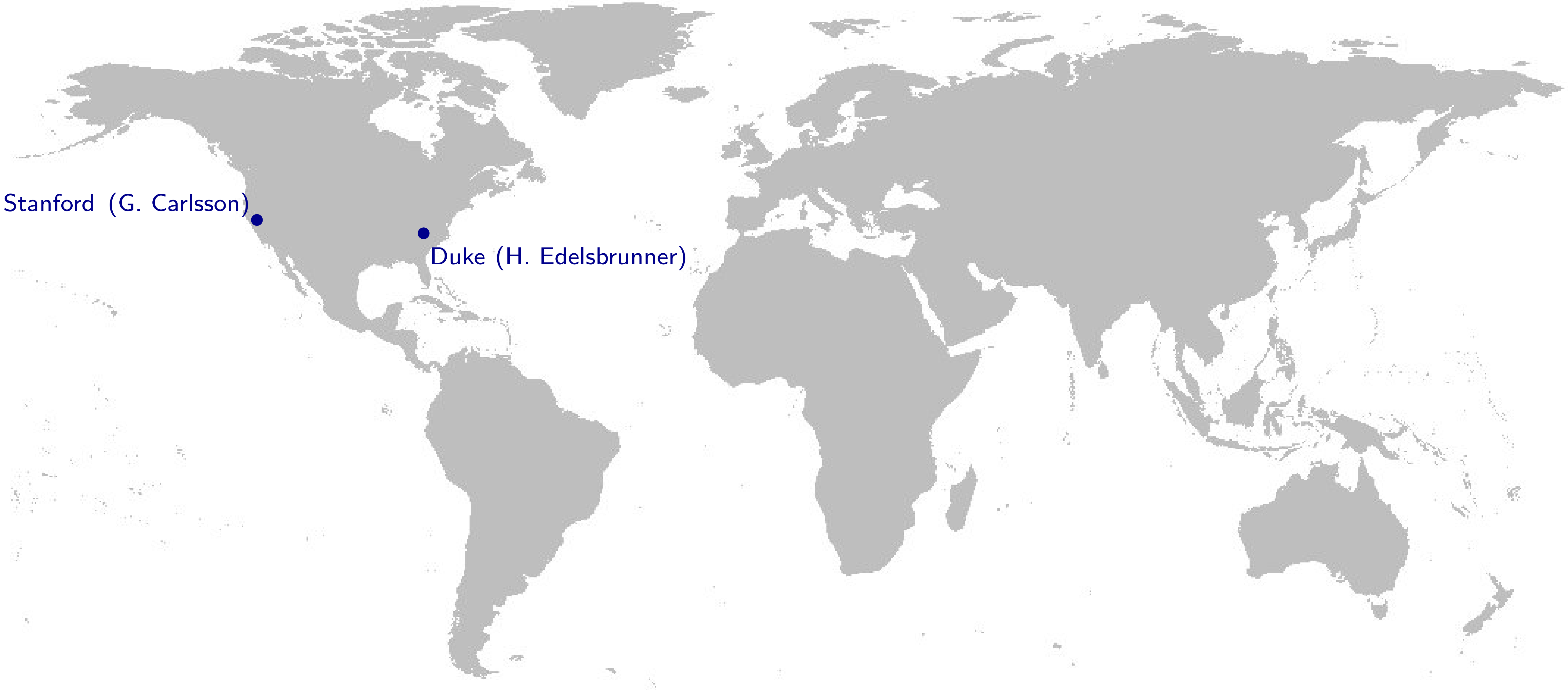
topological descriptors for inference and comparison





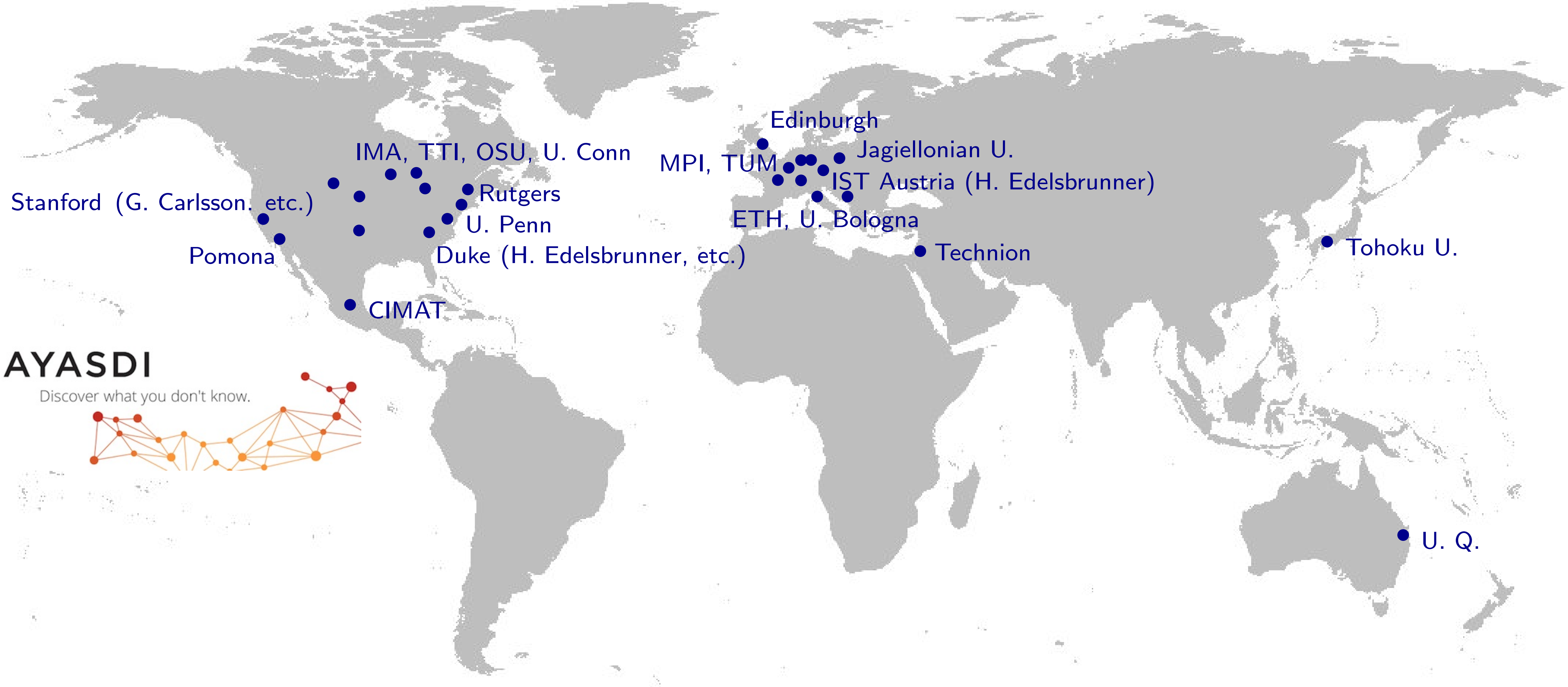
point cloud

The TDA community (as of 2002)



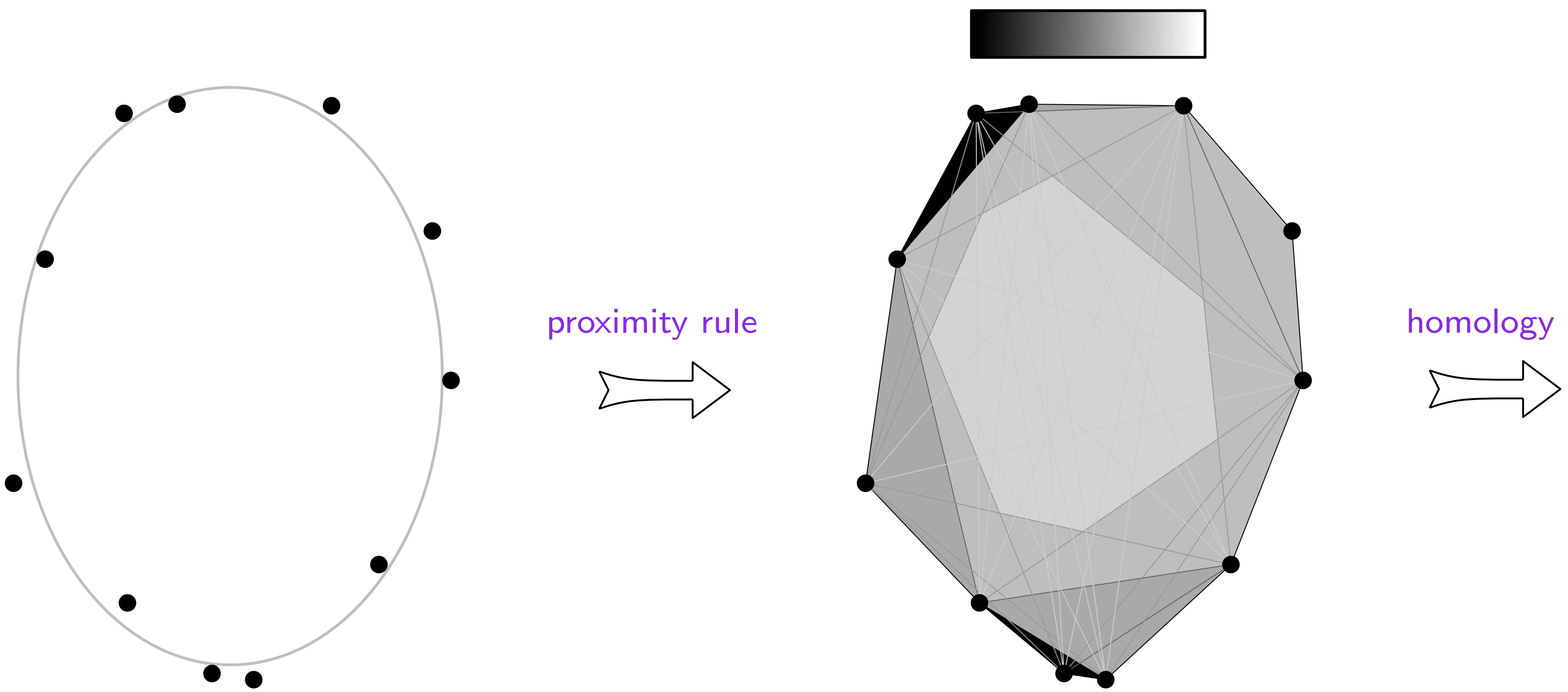
- 2 research groups (5-10 researchers)

The TDA community (as of 2016)



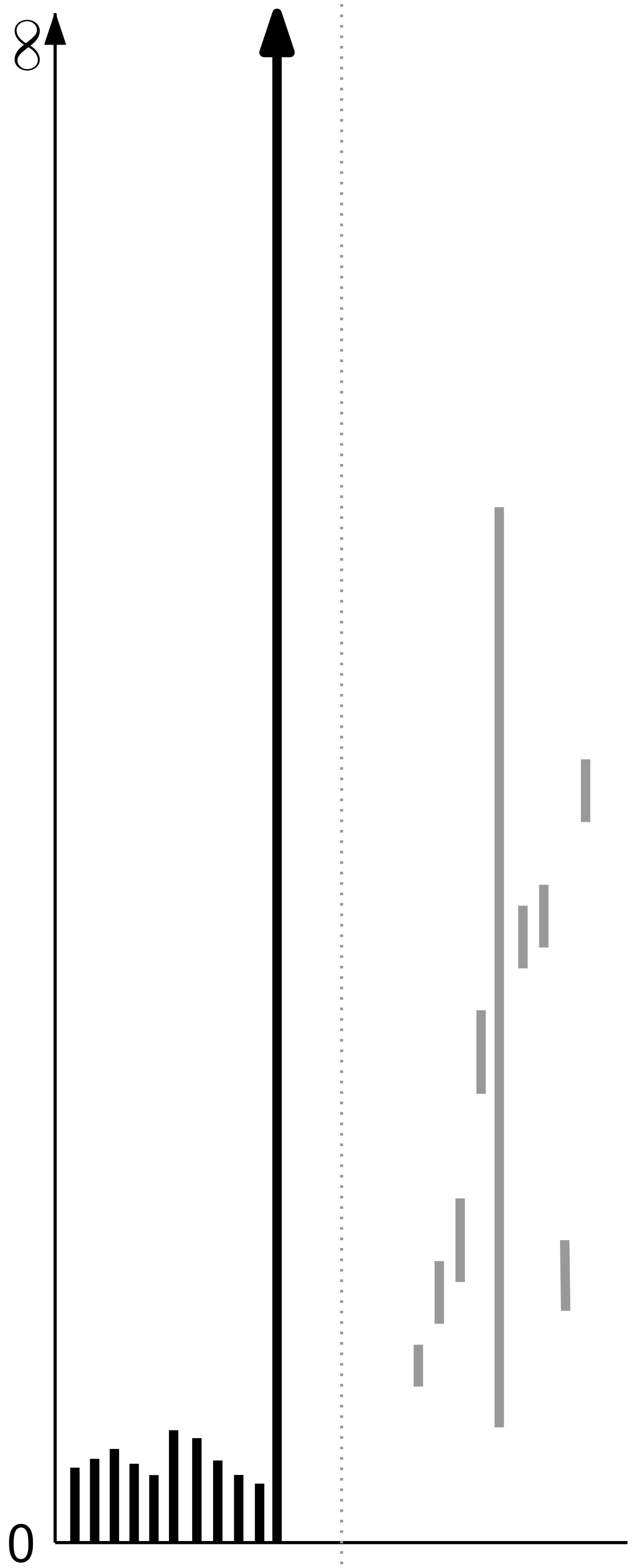
- 50-100 researchers working on theoretical foundations
- 200-300 researchers at the interface with applications
- very successful applications and company (Ayasdi)

The TDA pipeline at a glance



point cloud

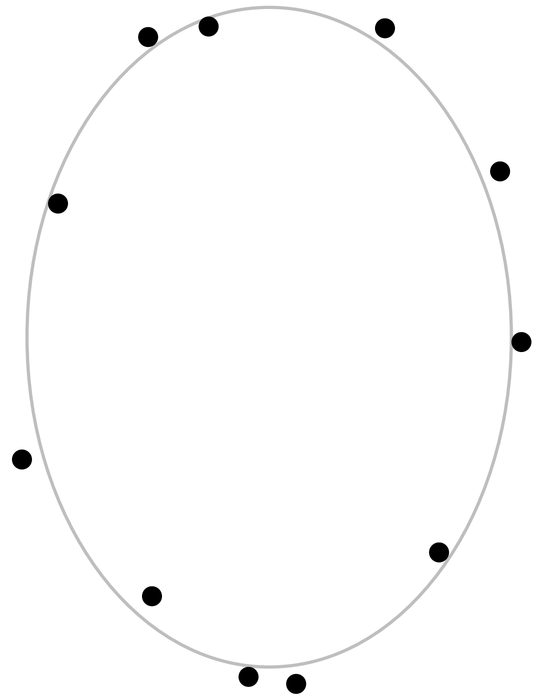
filtration (nested family of spaces)



descriptor (barcode)

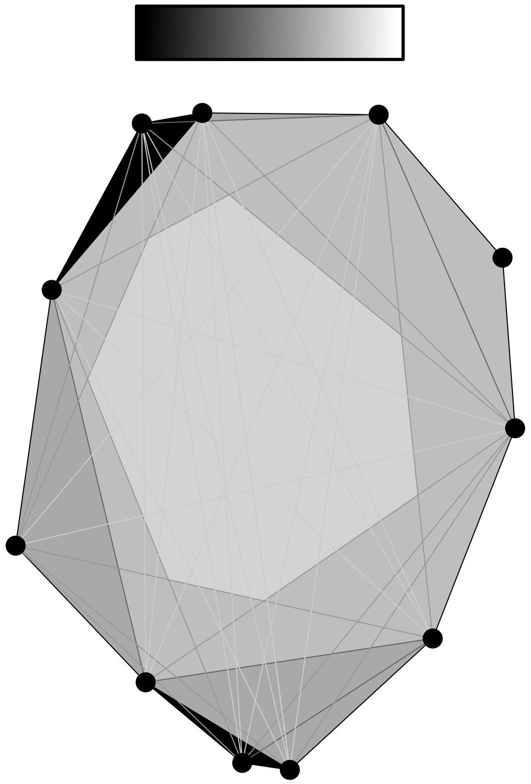
- visualization (unsupervised)
- summary (supervised)

The TDA pipeline at a glance



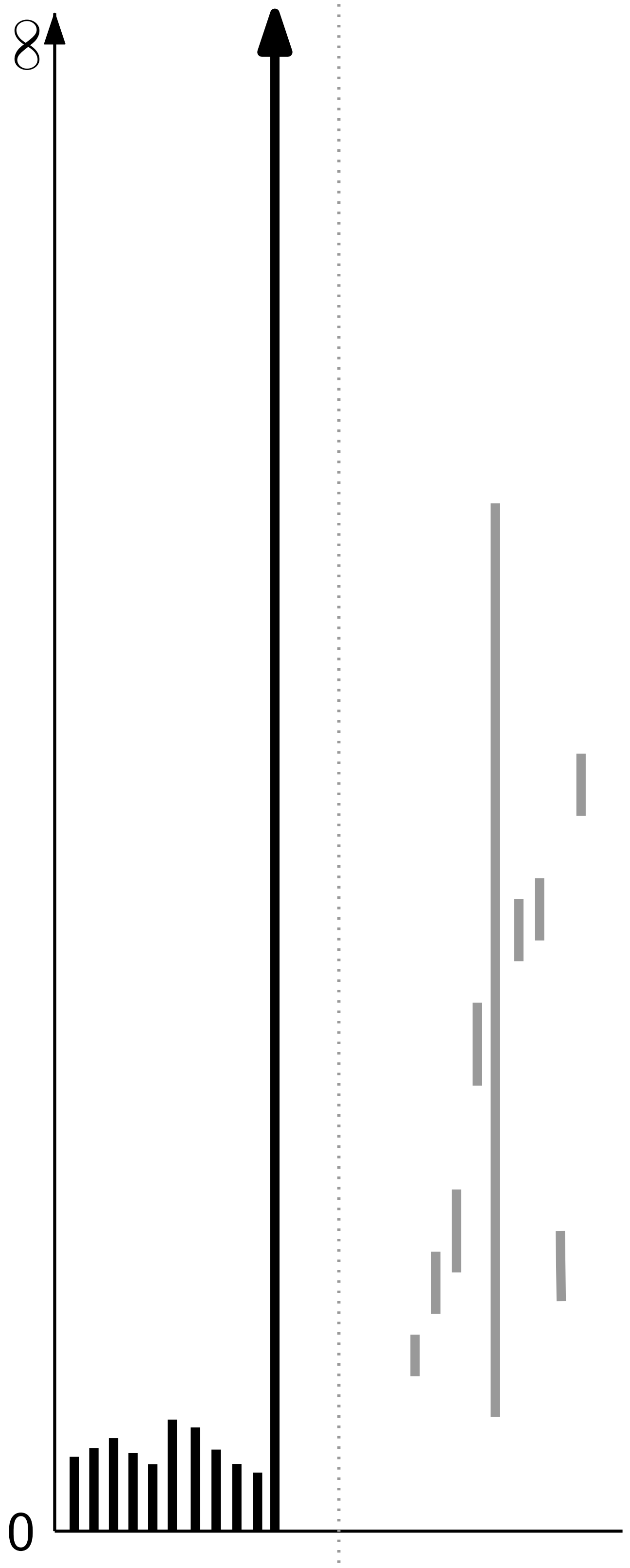
point cloud

proximity rule ⇒



filtration (nested family of spaces)

homology ⇒



descriptor (barcode)

3 pillars to the theory (topological persistence):

- decomposition theorems (∃ barcodes)
- algorithms (computation of barcodes)
- stability theorems (barcodes as stable descriptors)

Topological persistence in a nutshell

X topological space

$f : X \to \mathbb{R}$

↓

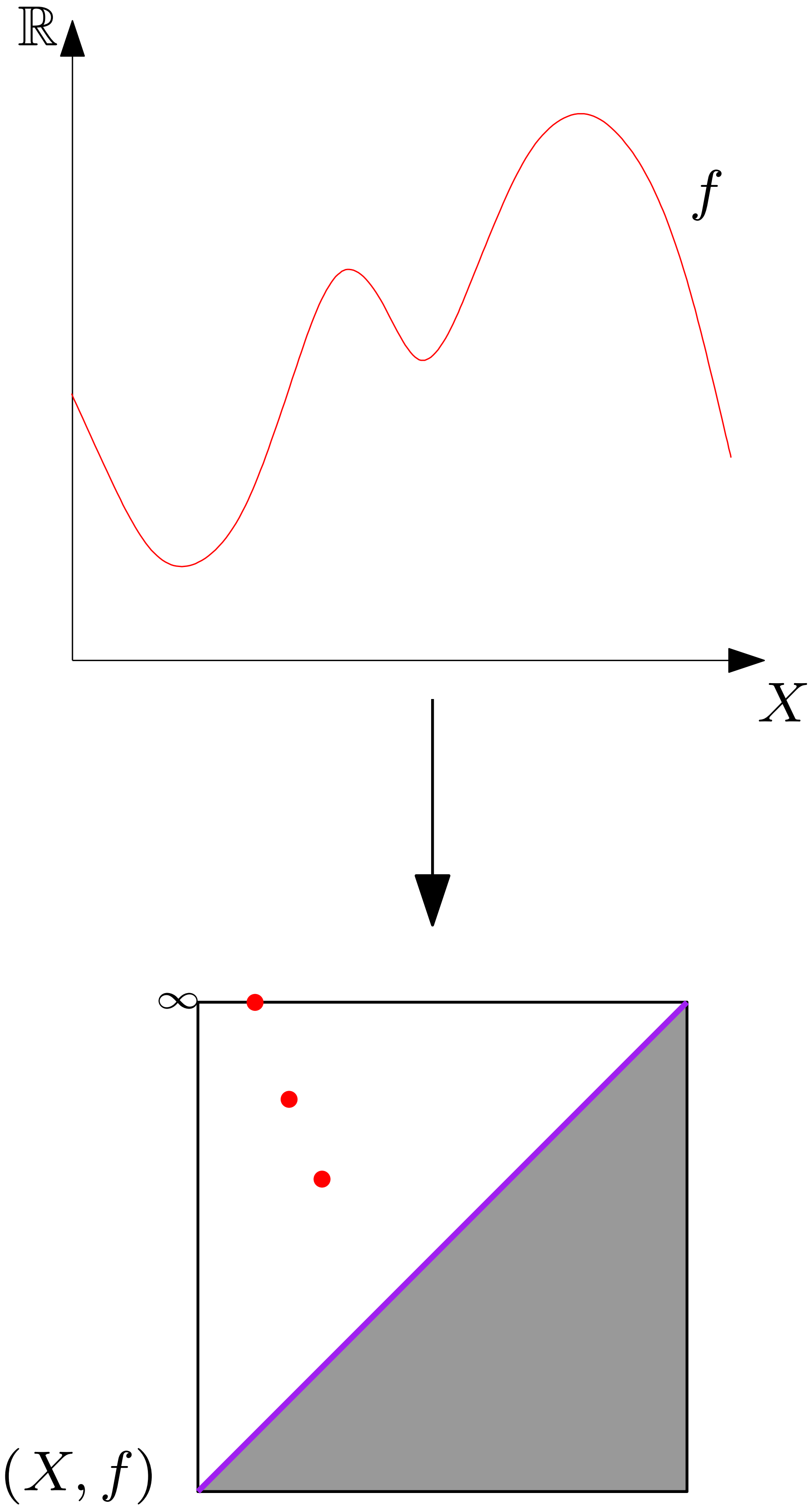
persistence

↓

$\mathrm{Dg}\, f$

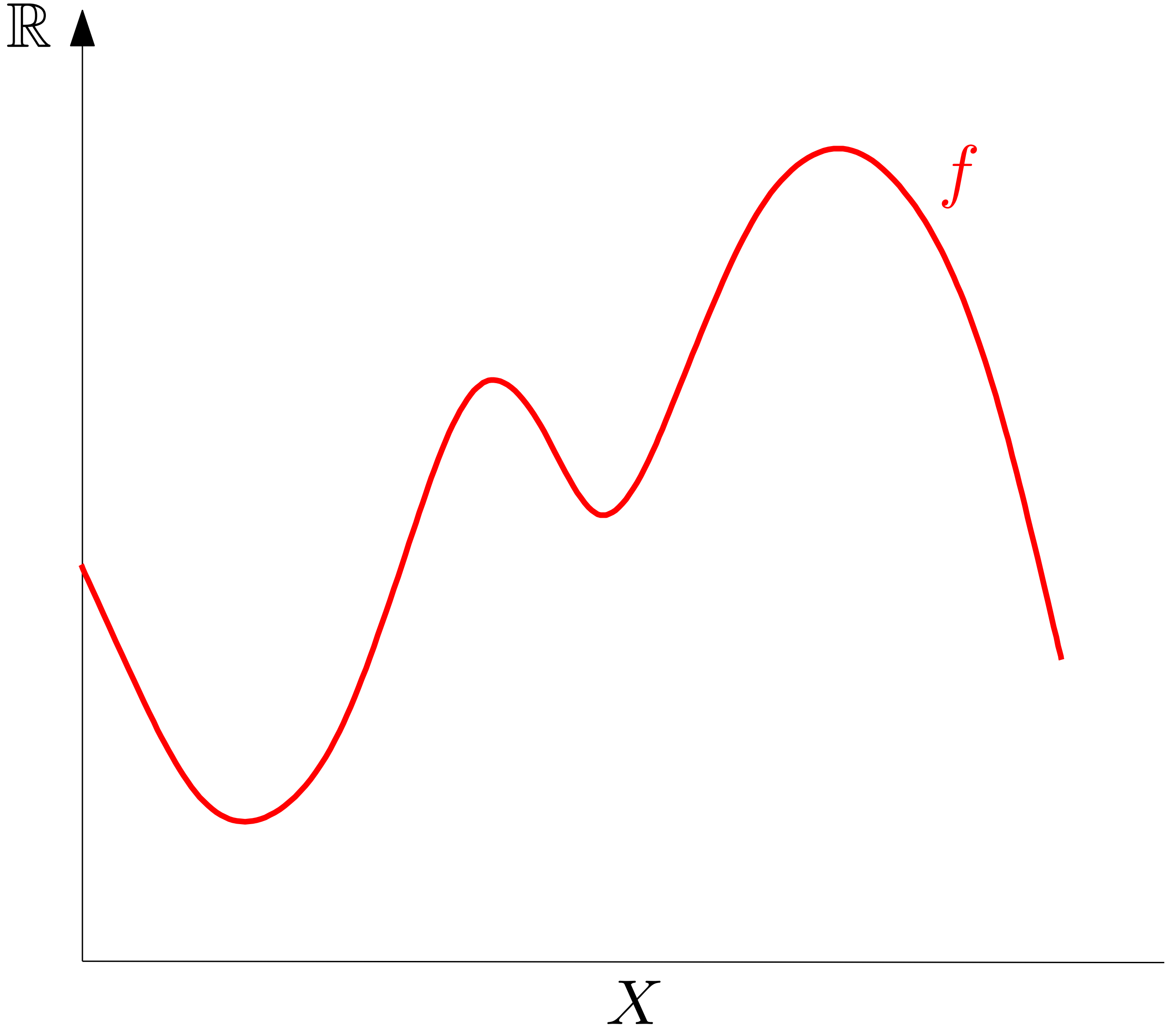
signature: *persistence diagram*

encodes the topological structure of the pair (X, f)



Topological persistence in a nutshell

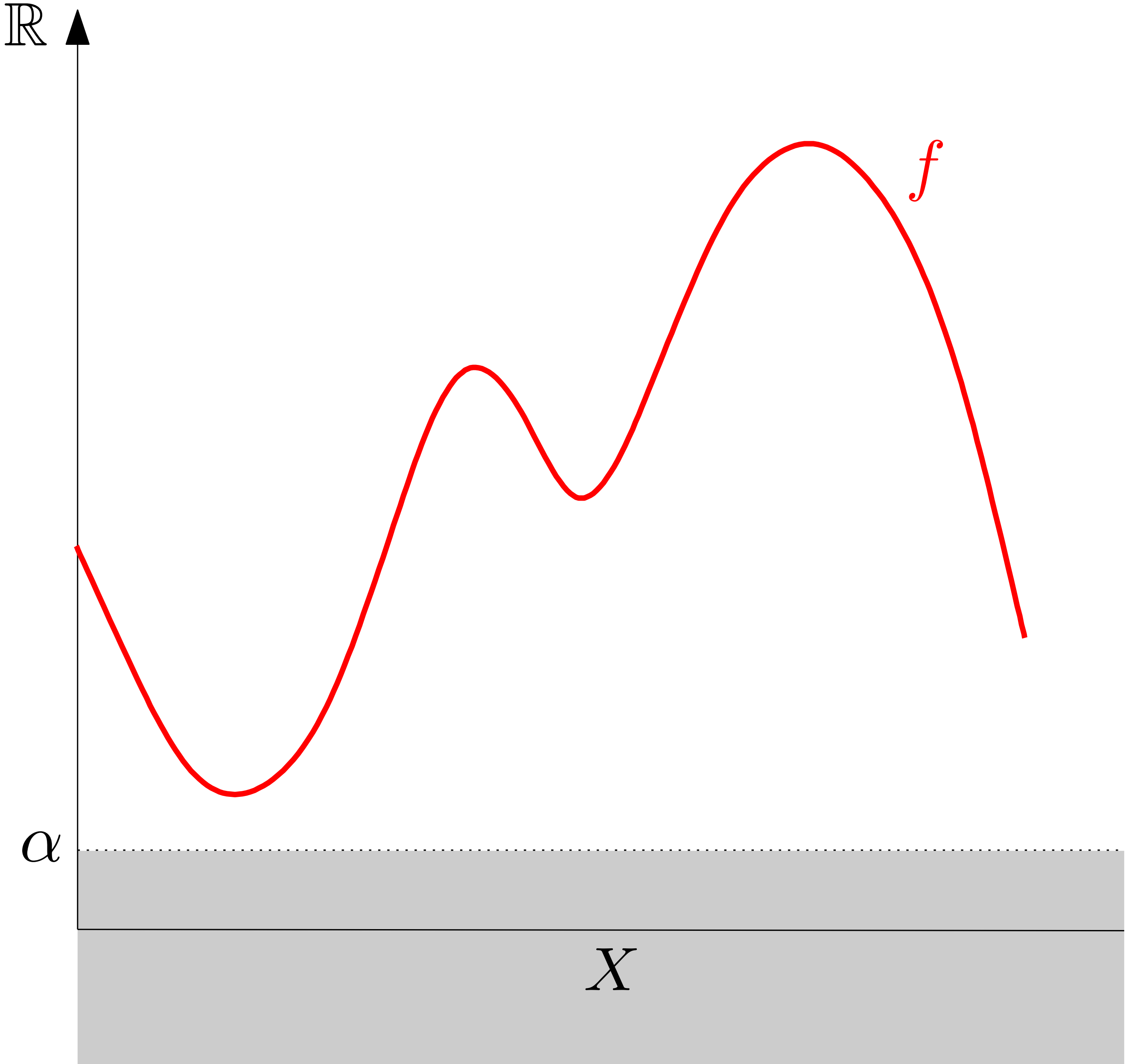
Inside the black box (X topological space, $f : X \to \mathbb{R}$):



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

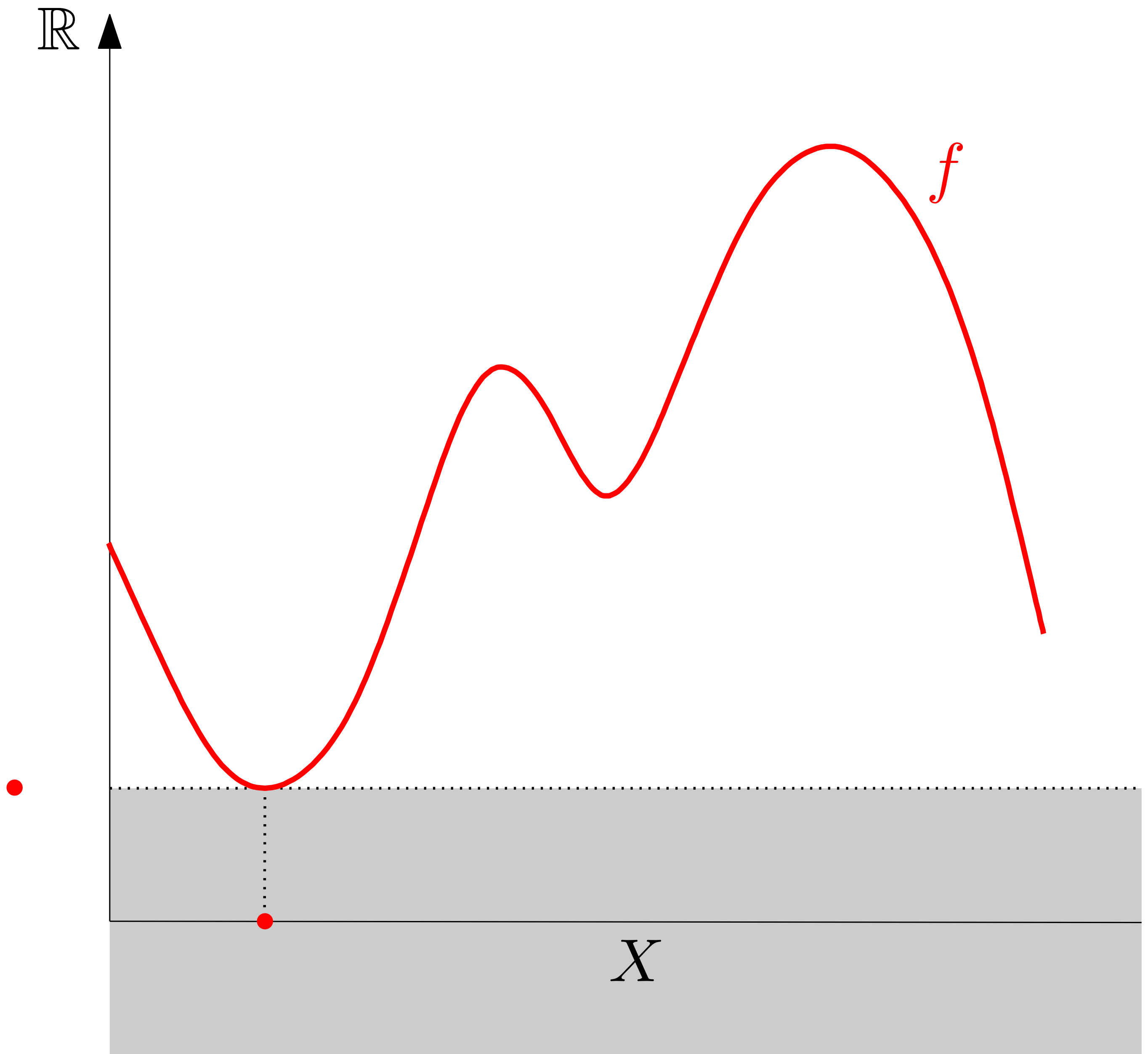
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

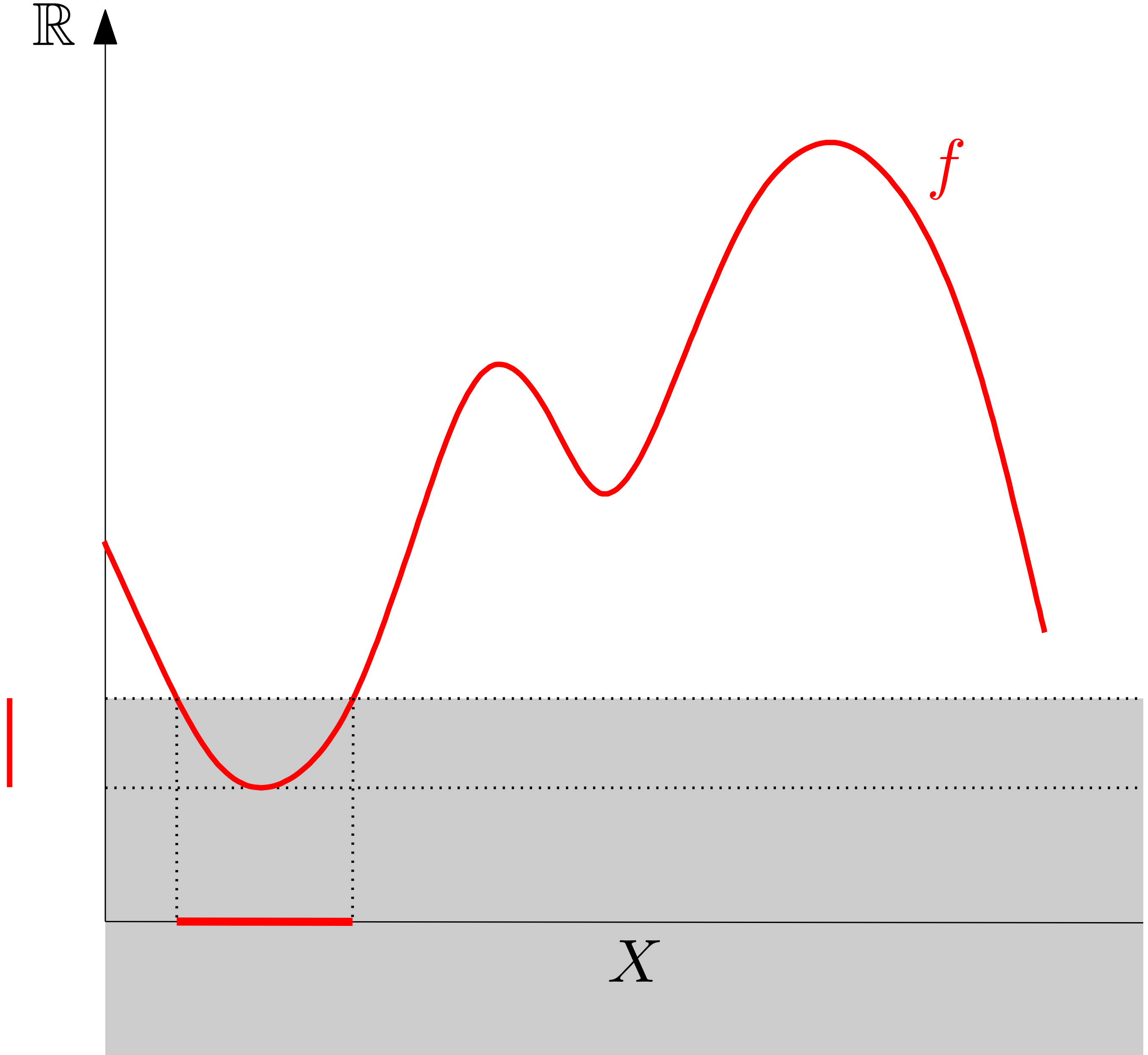
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

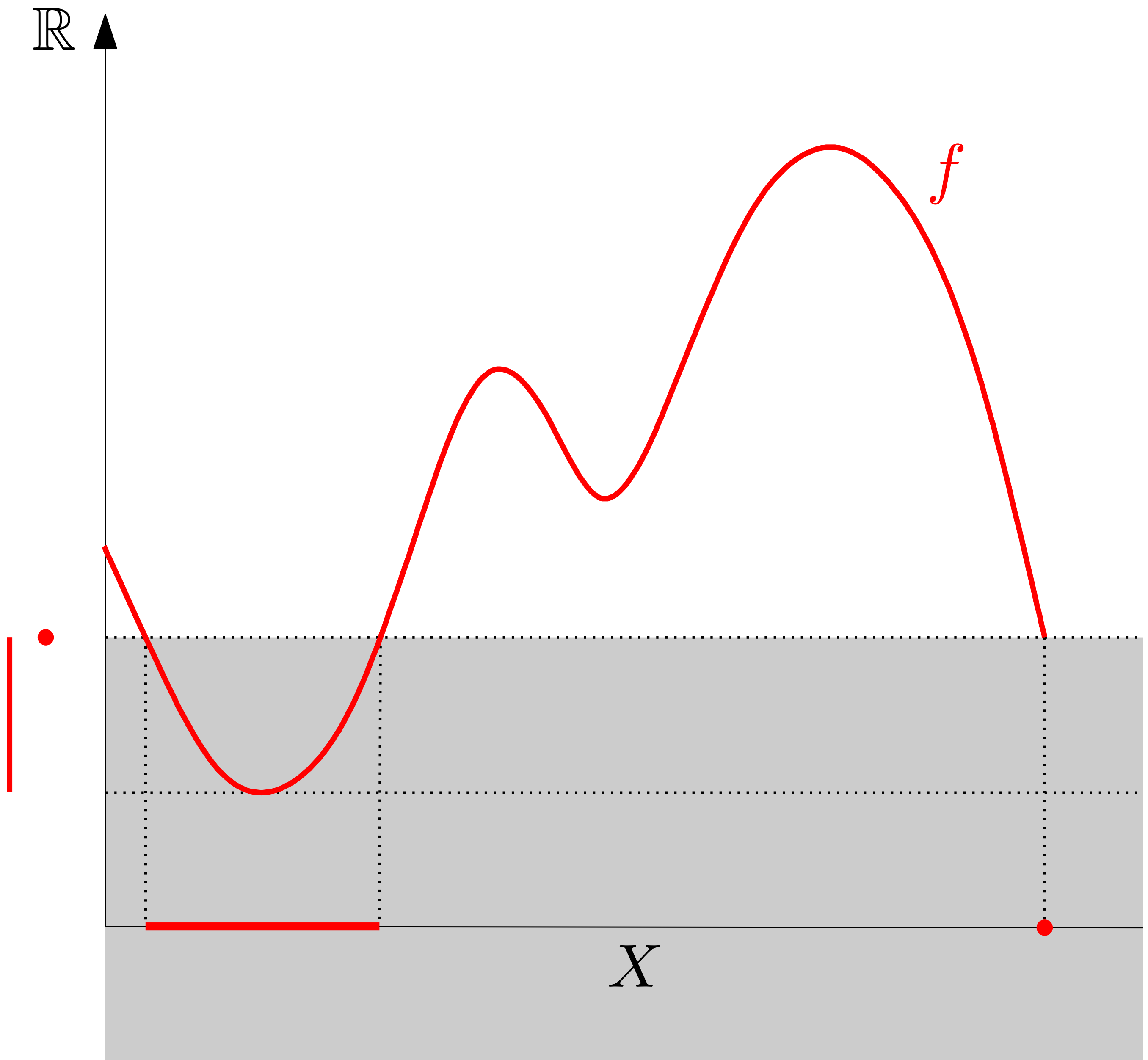
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

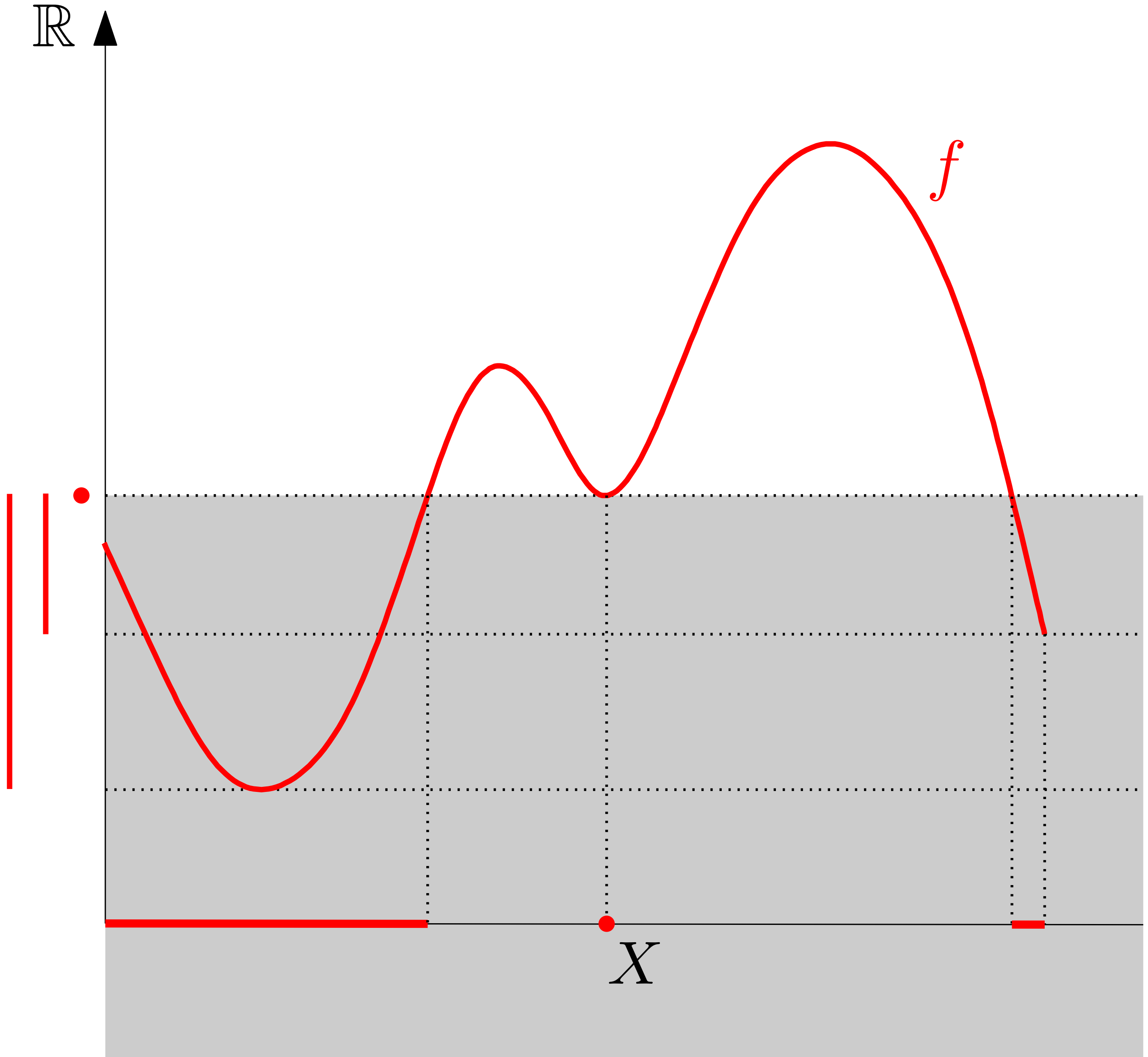
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

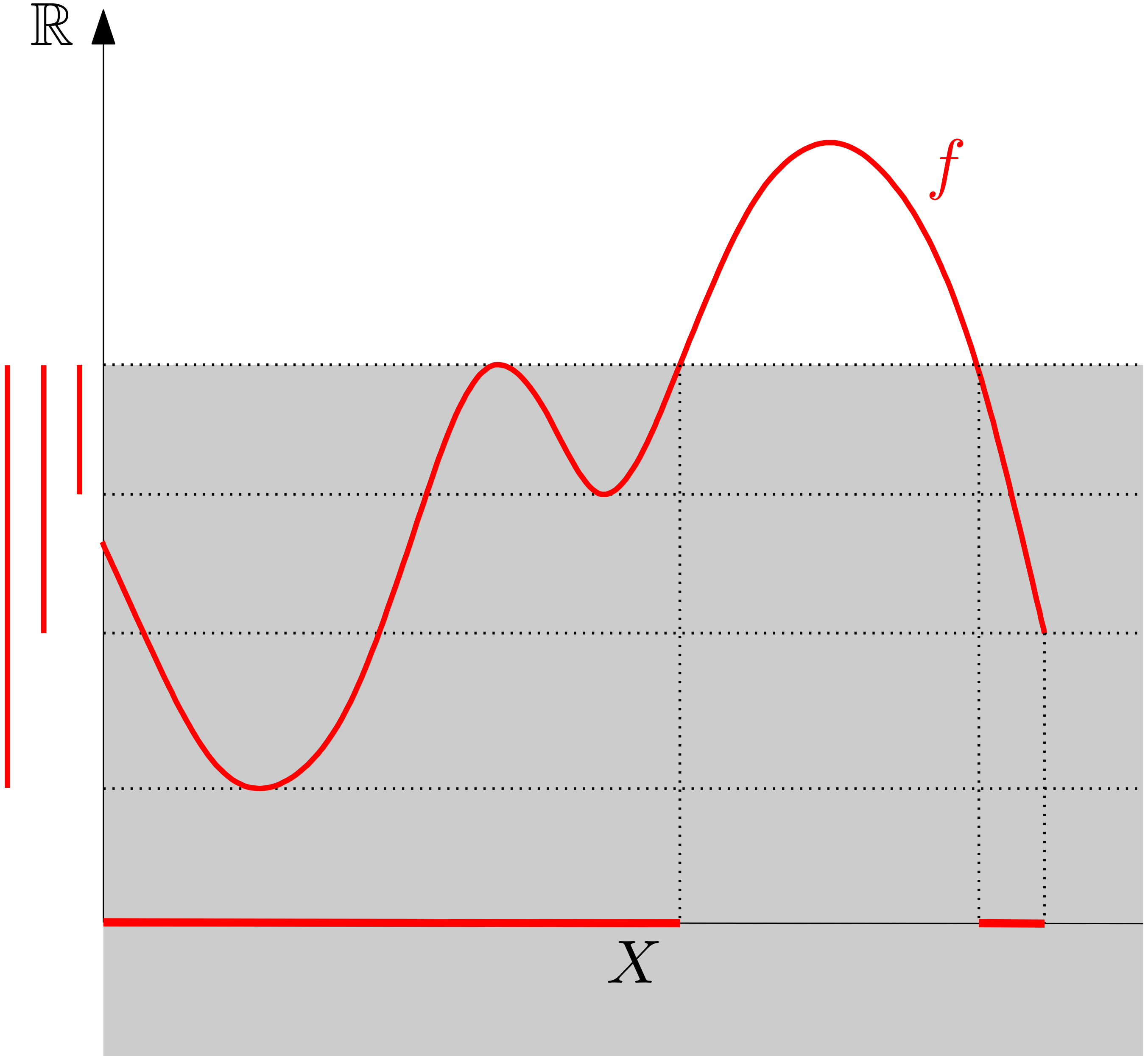
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

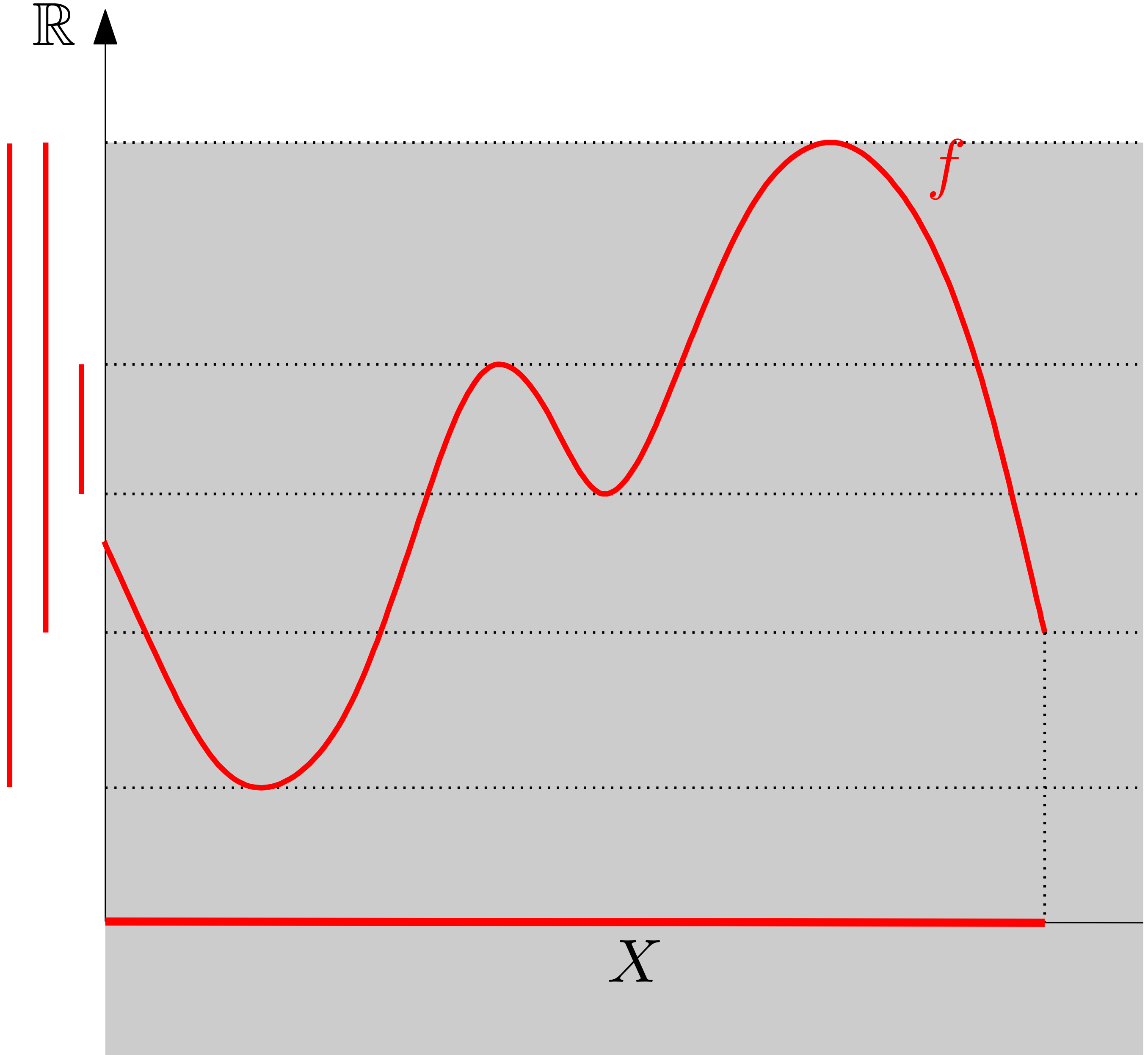
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

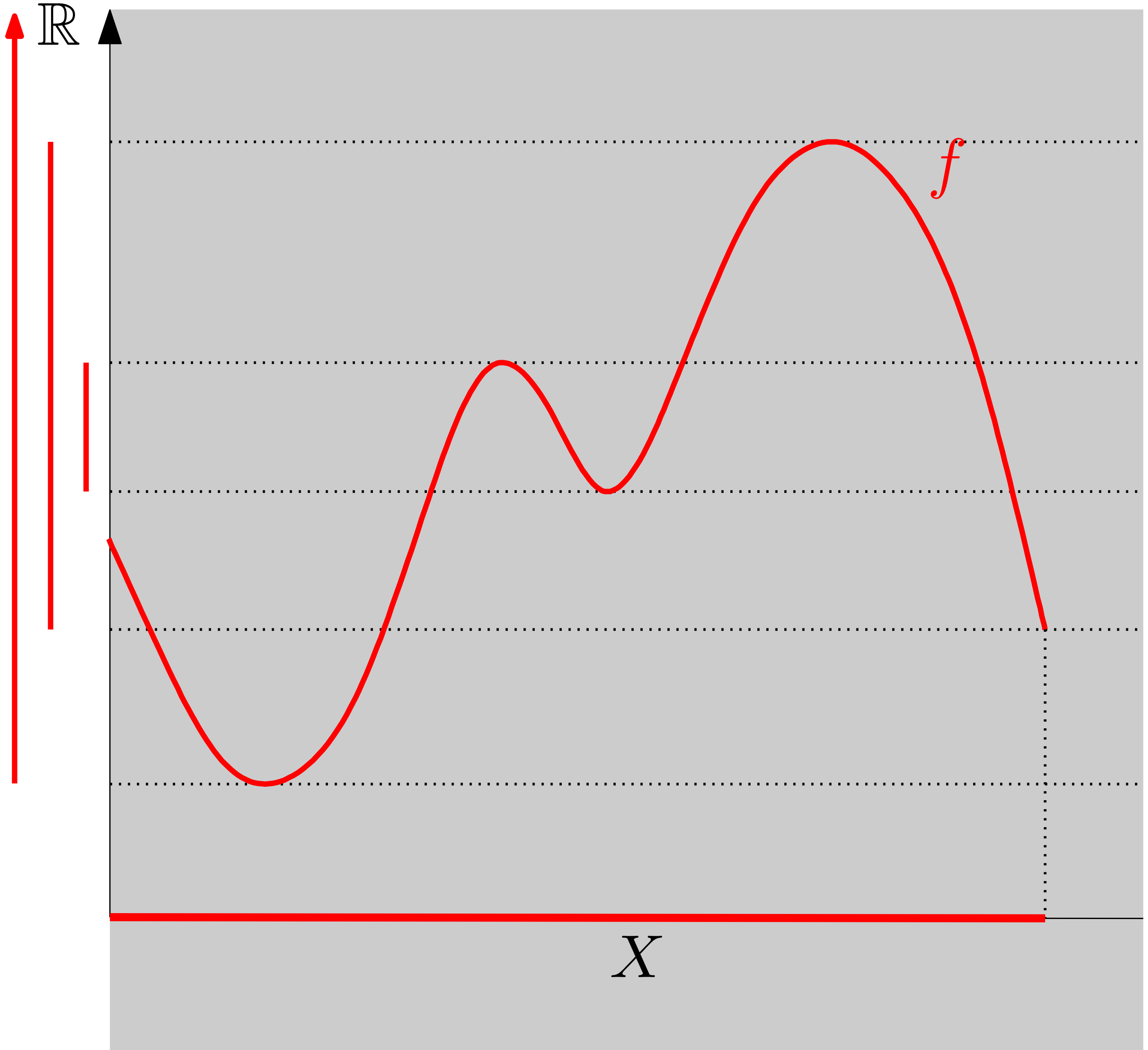
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

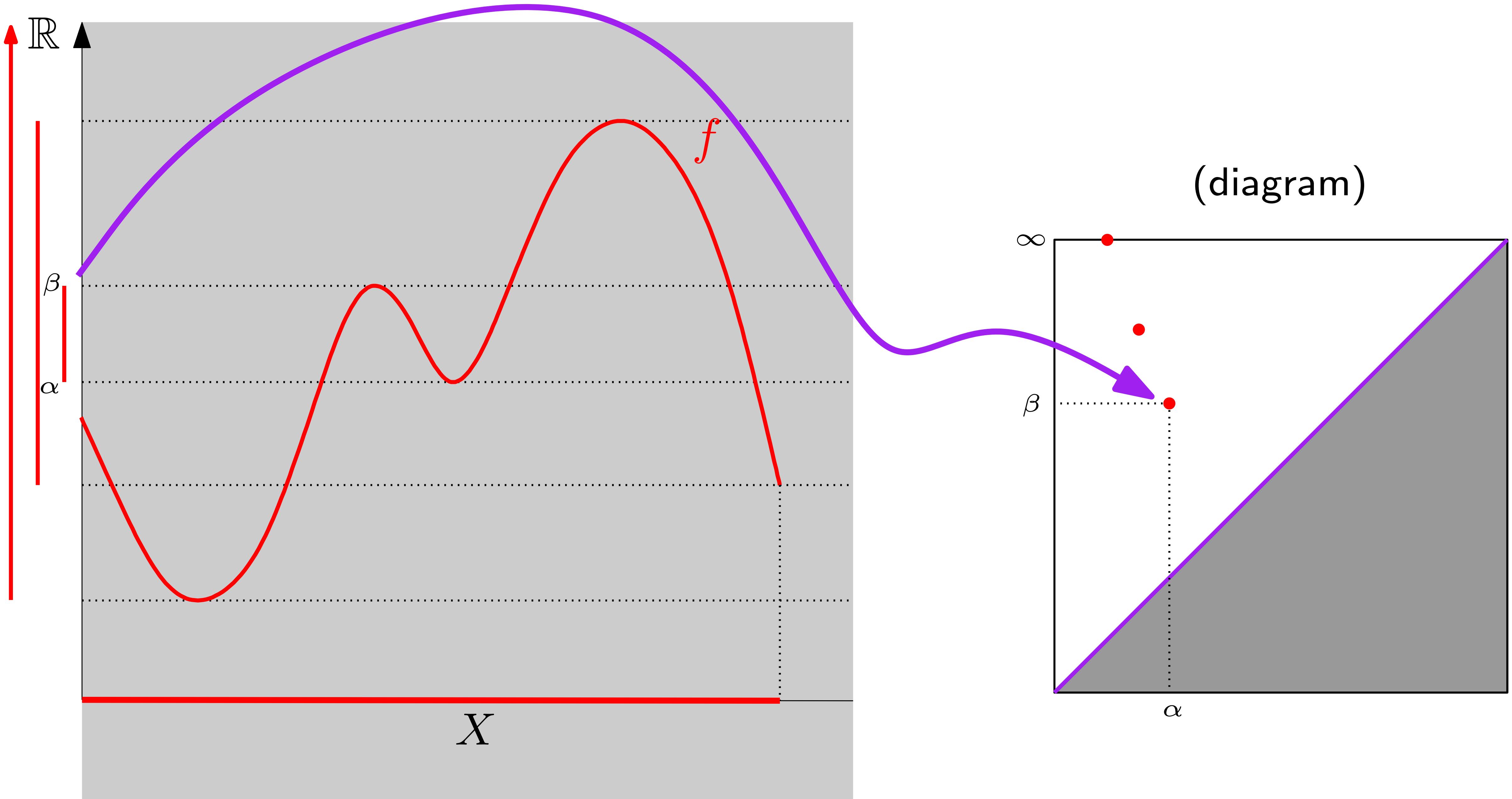
- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.
- Finite set of intervals (barcode) encodes births/deaths of topological features.



Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.
- Finite set of intervals (barcode) encodes births/deaths of topological features.



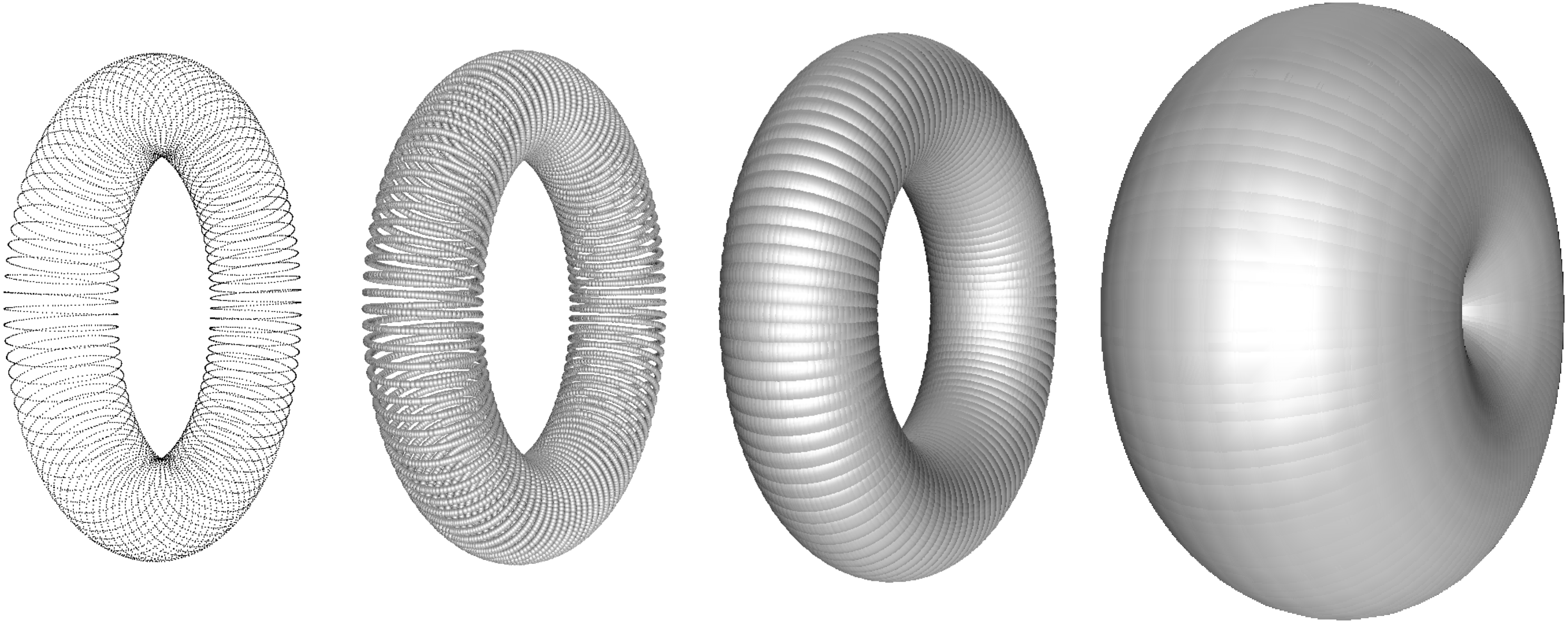
Topological persistence in a nutshell

Example:

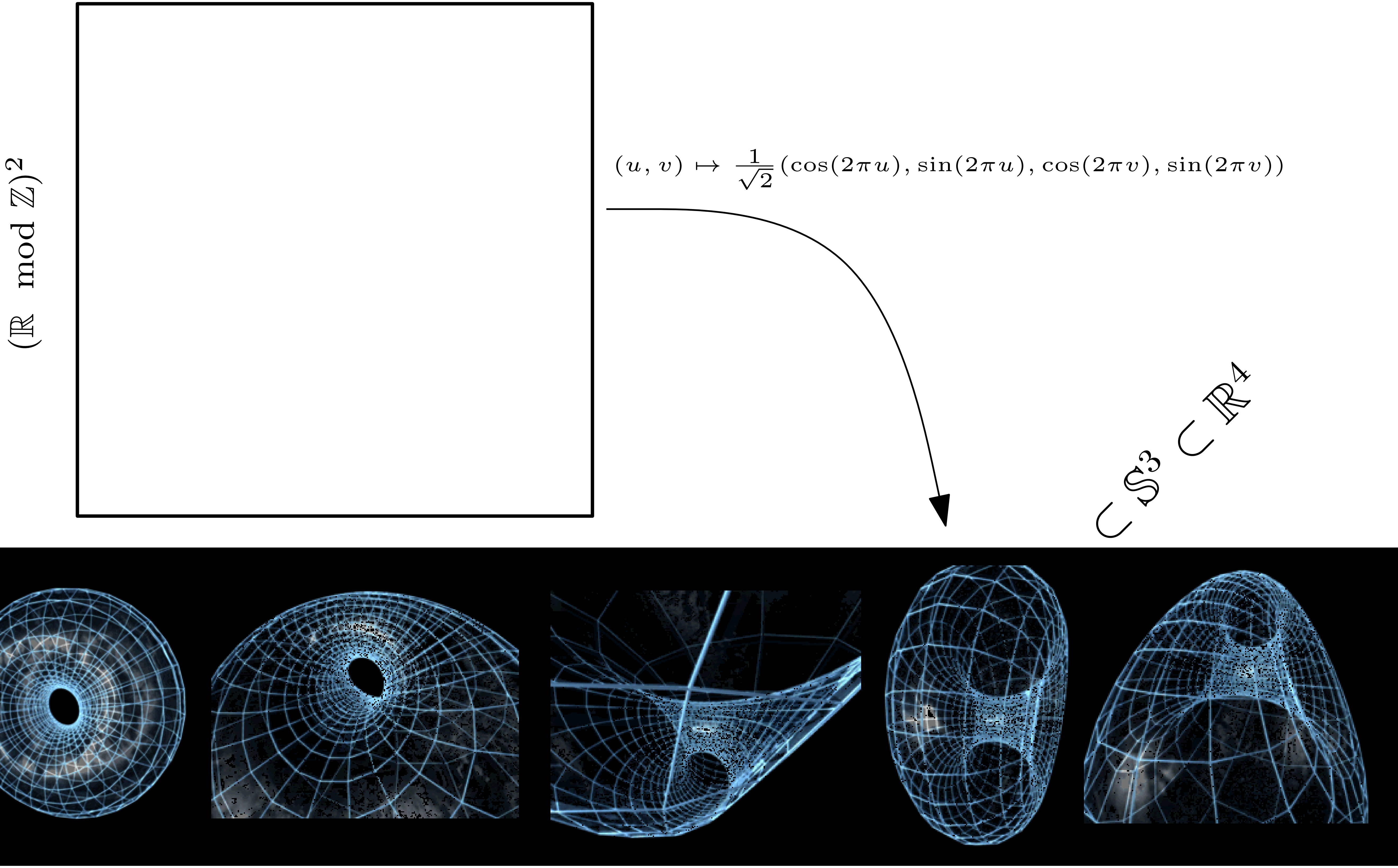
$X = \mathbb{R}^n$, $K \subset \mathbb{R}^n$ compact

$f : \mathbb{R}^n \to \mathbb{R}$

$x \mapsto \min_{y \in K} \|x - y\|_2$



Topological persistence in a nutshell



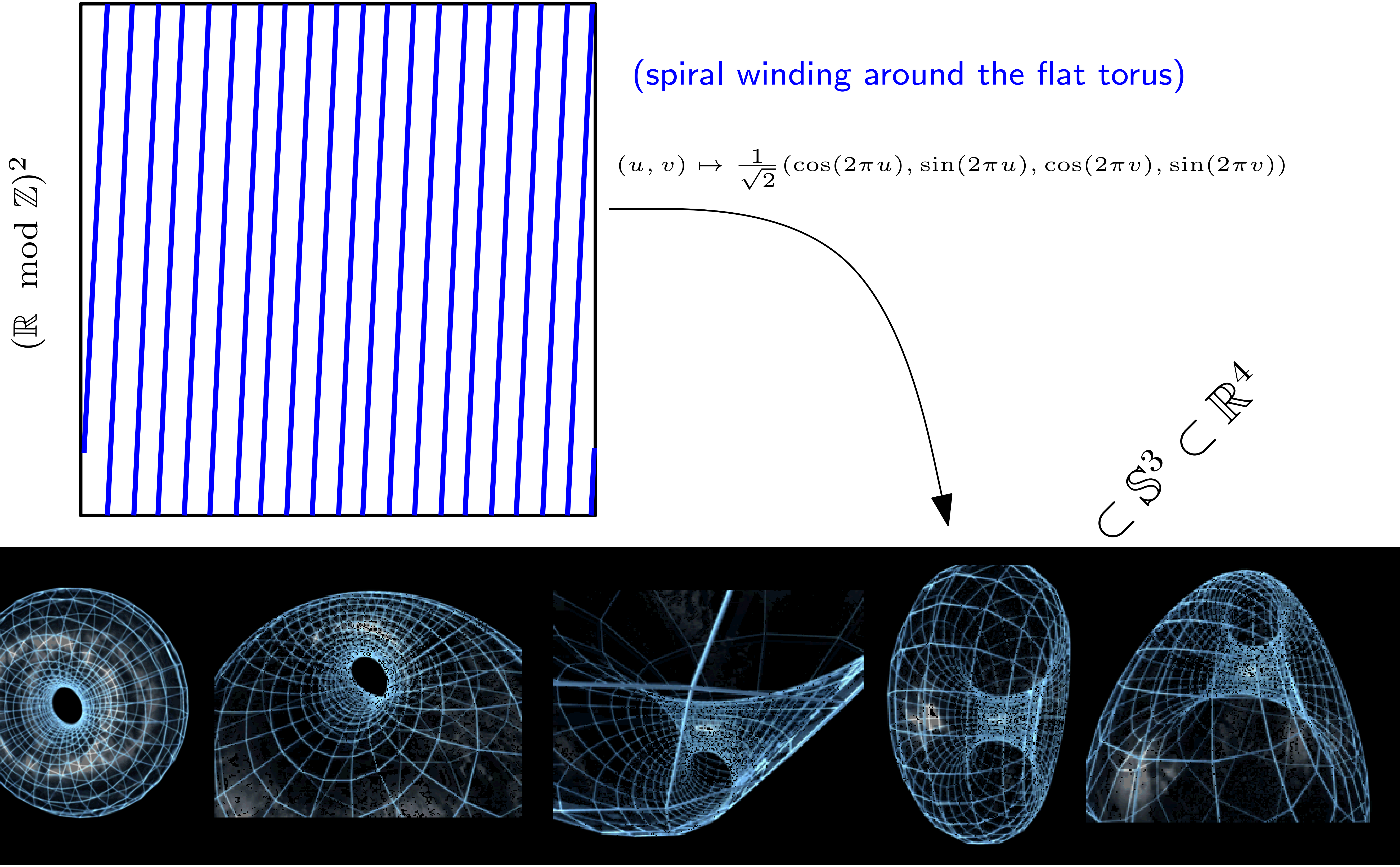
$(\mathbb{R} \mod \mathbb{Z})^2$

$(u, v) \mapsto \frac{1}{\sqrt{2}}(\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))$

$\subset \mathbb{S}^3 \subset \mathbb{R}^4$

source: `http://http://en.wikipedia.org/wiki/Clifford_torus`

Topological persistence in a nutshell

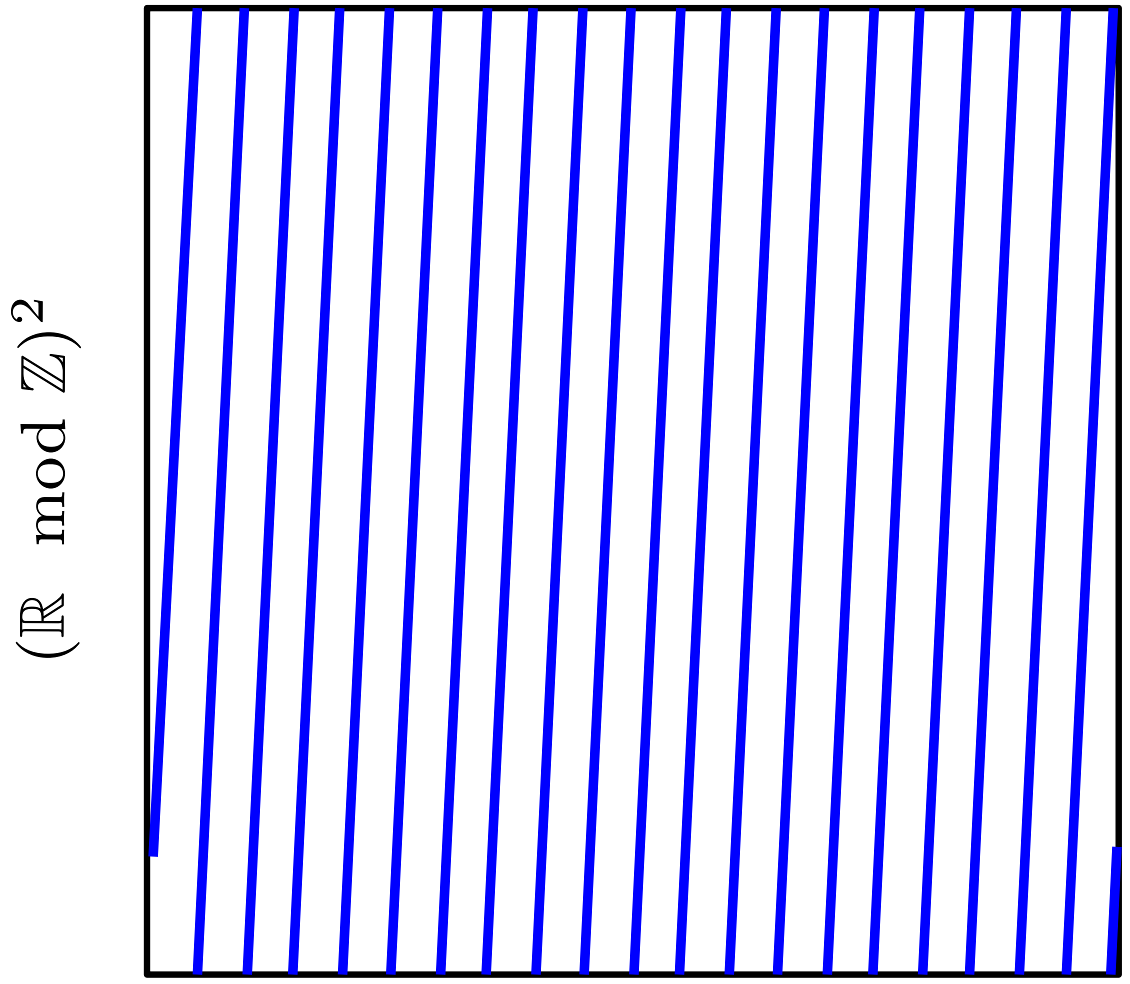


(spiral winding around the flat torus)

$$(u, v) \mapsto \frac{1}{\sqrt{2}}(\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))$$

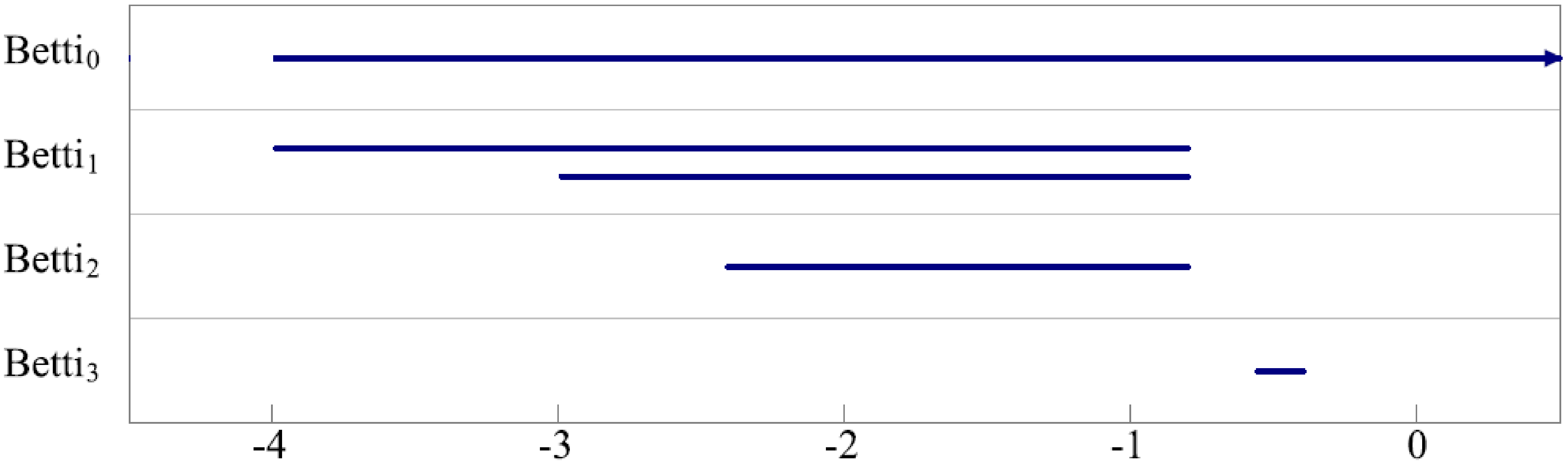
source: `http://http://en.wikipedia.org/wiki/Clifford_torus`

Topological persistence in a nutshell



(spiral winding around the flat torus)

$$(u, v) \mapsto \frac{1}{\sqrt{2}}(\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))$$

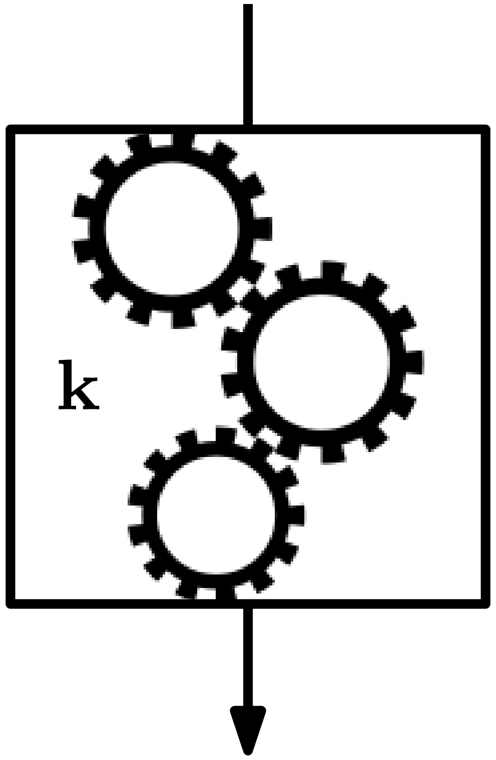


Topological persistence in a nutshell

Inside the black box (X topological space, $f : X \to \mathbb{R}$):

- Nested family (filtration) of sublevel-sets $F_\alpha = f^{-1}((-\infty, \alpha])$ for $\alpha \in \mathbb{R}$.
- Track evolution of topology throughout the family.
- Finite set of intervals (barcode) encodes births/deaths of topological features.

filtration: $F_1 \subseteq F_2 \subseteq F_3 \subseteq F_4 \subseteq F_5 \cdots$



(homology functor)

topological level

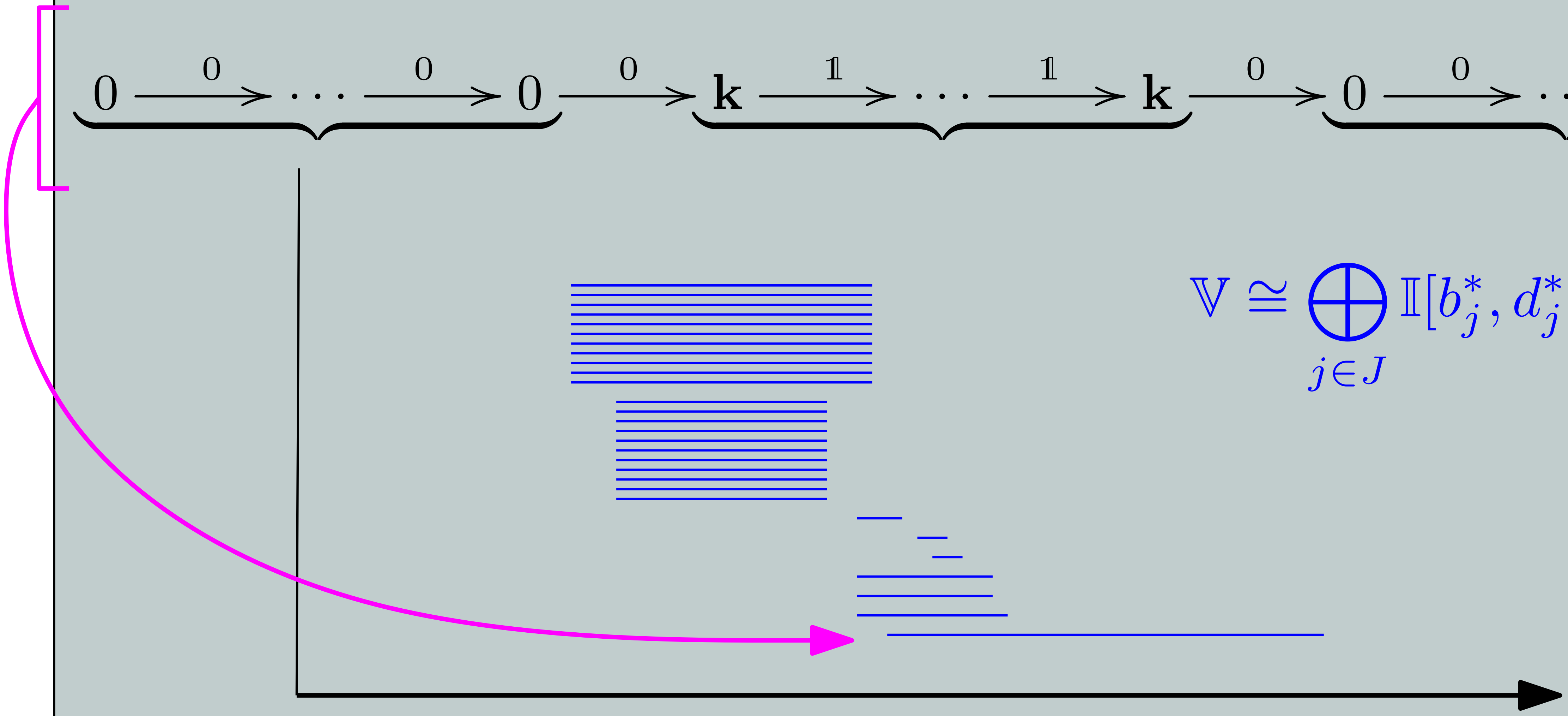
algebraic level

(persistence) module: $H_*(F_1) \to H_*(F_2) \to H_*(F_3) \to H_*(F_4) \to H_*(F_5) \cdots$

Topological persistence in a nutshell

Theorem. [Gabriel '72, Auslander '74, $\cdots$, Webb '85, $\cdots$, Crawley-Boevey '12]
Under some technical conditions, a persistence module $\mathbb{V}$ decomposes as a direct sum of **interval modules** $\mathbb{I}[b^*, d^*]$:

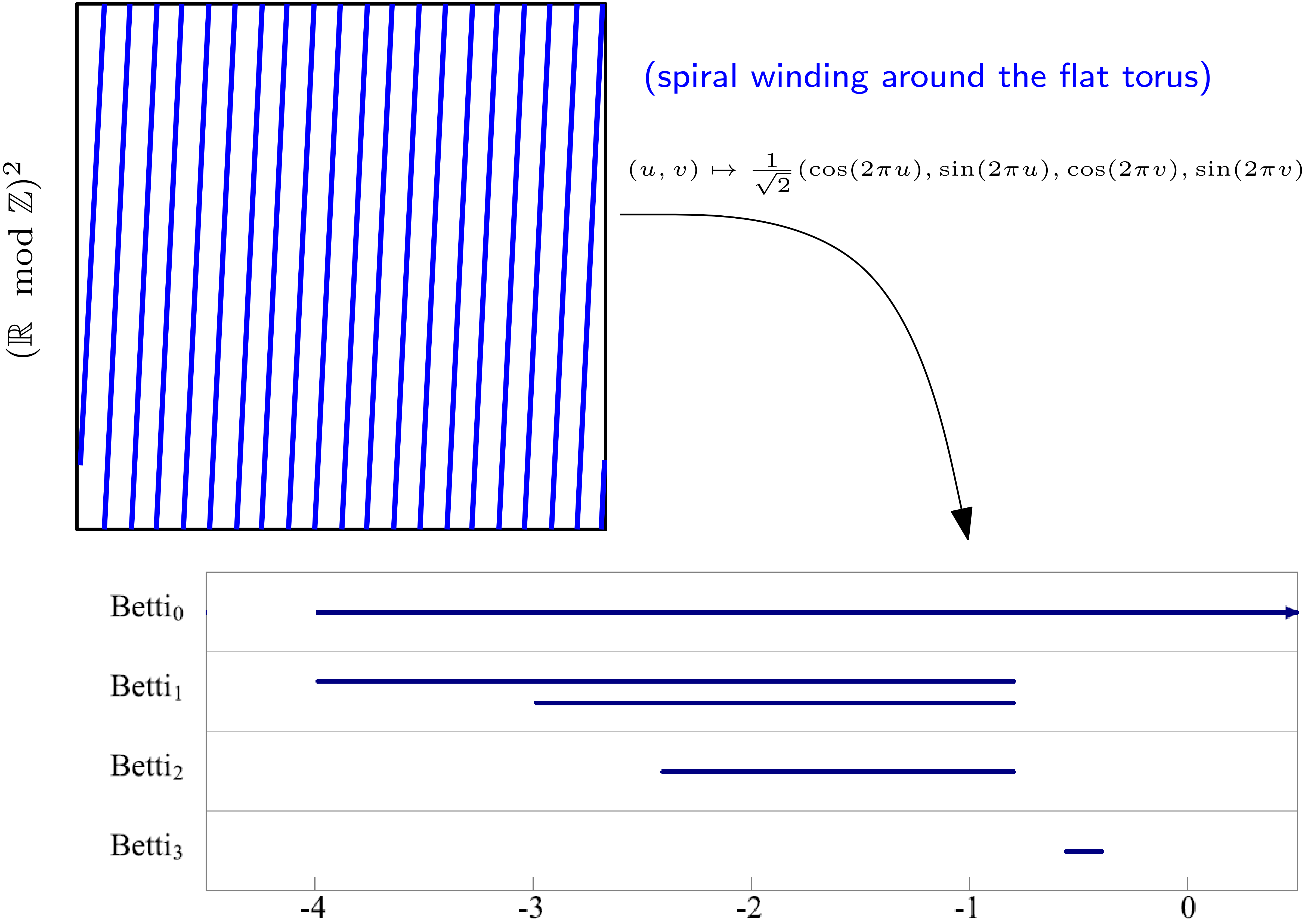
$$\underbrace{0 \xrightarrow{0} \cdots \xrightarrow{0} 0}_{} \xrightarrow{0} \underbrace{\mathbf{k} \xrightarrow{\mathbb{1}} \cdots \xrightarrow{\mathbb{1}} \mathbf{k}}_{} \xrightarrow{0} \underbrace{0 \xrightarrow{0} \cdots \xrightarrow{0} 0}_{}$$



$$\mathbb{V} \cong \bigoplus_{j \in J} \mathbb{I}[b_j^*, d_j^*]$$

(persistence) module: $H_*(F_1) \to H_*(F_2) \to H_*(F_3) \to H_*(F_4) \to H_*(F_5) \cdots$

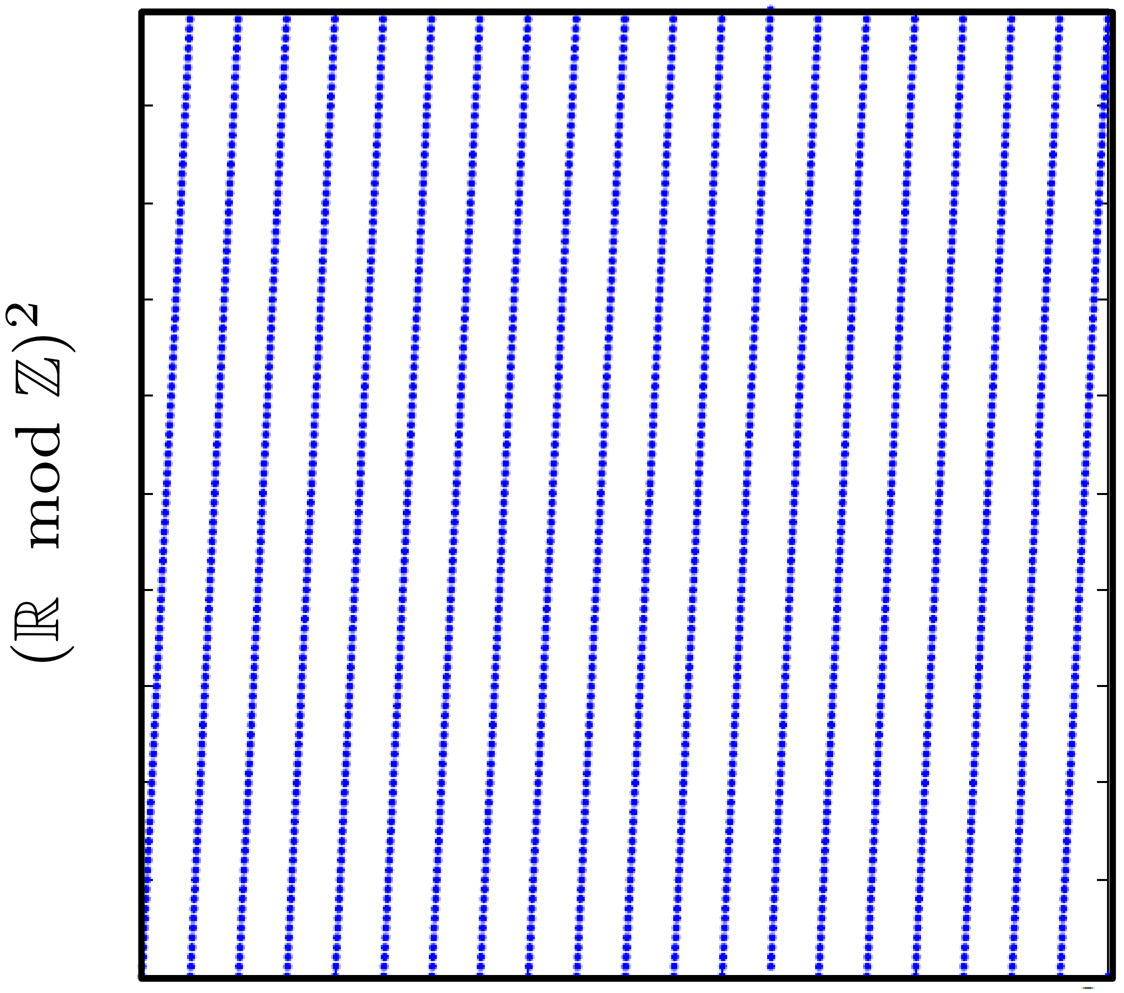
Topological persistence in a nutshell



(spiral winding around the flat torus)

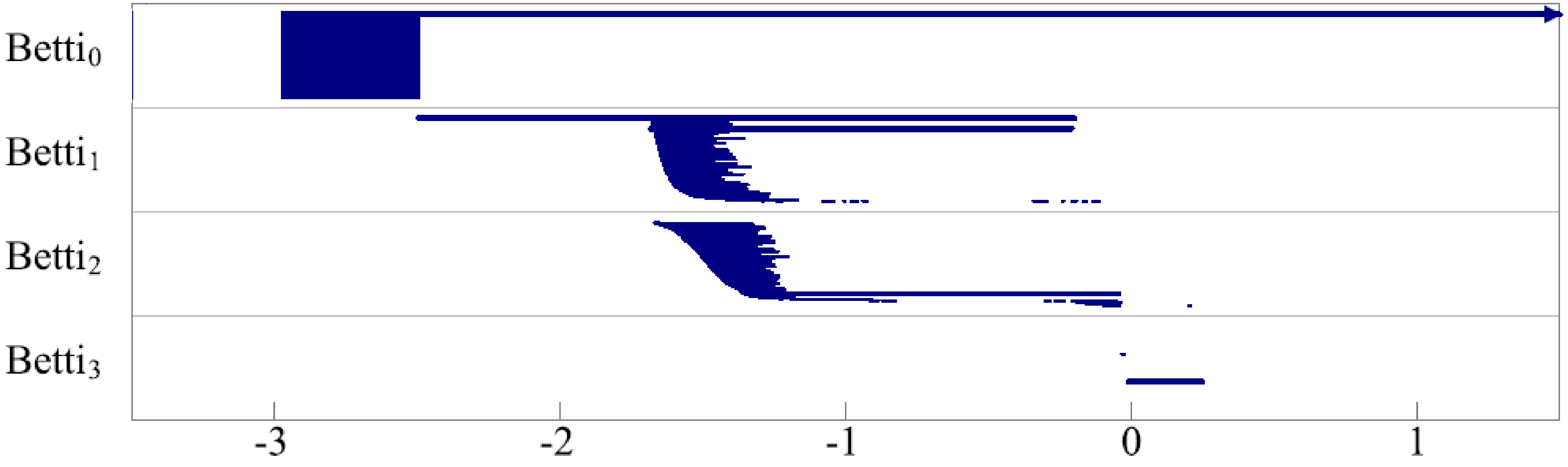
$$(u, v) \mapsto \frac{1}{\sqrt{2}}(\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))$$

Topological persistence in a nutshell



(2000 data points)

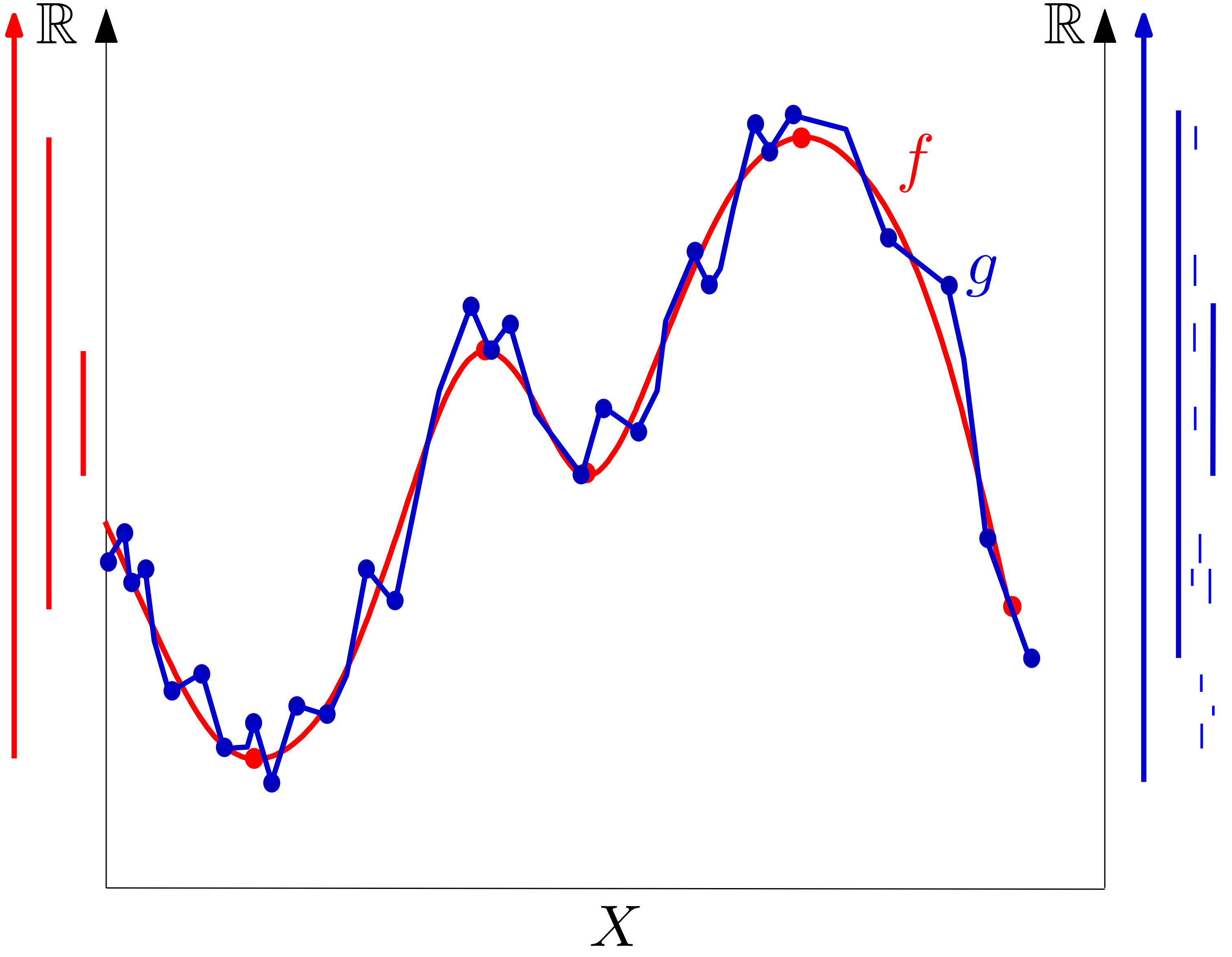
$(u, v) \mapsto \frac{1}{\sqrt{2}}(\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))$



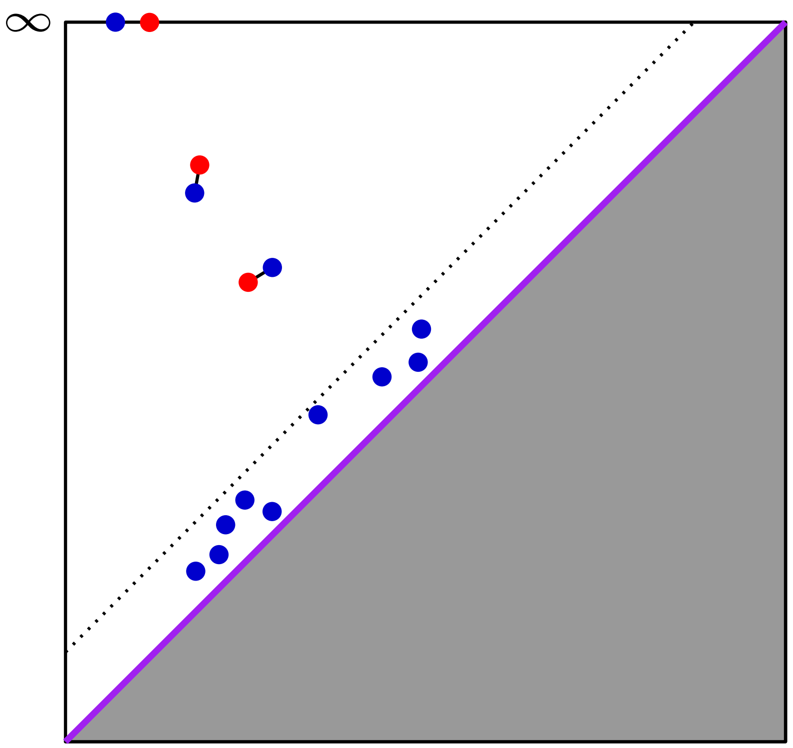
Topological persistence in a nutshell

Theorem: [Cohen-Steiner et al. 2005, Chazal et al. 2009]
For any *tame* functions $f, g : X \to \mathbb{R}$,

$$\mathrm{d}_B^\infty(\mathrm{Dg}\, f, \mathrm{Dg}\, g) \leq \|f - g\|_\infty$$



(bottleneck distance)



(W_∞ *with a twist*)

Statistics on the space of persistence diagrams

Note: the space of persistence diagrams is not a linear space

View persistence diagrams as discrete measures

Consider the p-th Wasserstein distance *with a twist*

Defining means (Fréchet means):

Given $D_1, \cdots, D_n$: persistence diagrams, is the following set empty?

$$\arg\min_{D} \sum_{i=1}^{n} W_p(D, D_i)^2$$

Statistics on the space of persistence diagrams

Note: the space of persistence diagrams is not a linear space

View persistence diagrams as discrete measures

Consider the p-th Wasserstein distance *with a twist*

Defining means (Fréchet means):

Given $D_1, \cdots, D_n$: persistence diagrams, is the following set empty?

$$\arg\min_{D} \sum_{i=1}^{n} W_p(D, D_i)^2$$

Theorem [Turner et al. 2012]: For any finite $p \geq 1$, the space of persistence diagrams equipped with W_p is *Polish* (complete and separable).

Corollary: Fréchet means exist

Statistics on the space of persistence diagrams

Note: the space of persistence diagrams is not a linear space

View persistence diagrams as discrete measures

Consider the p-th Wasserstein distance *with a twist*

Defining means (Fréchet means):

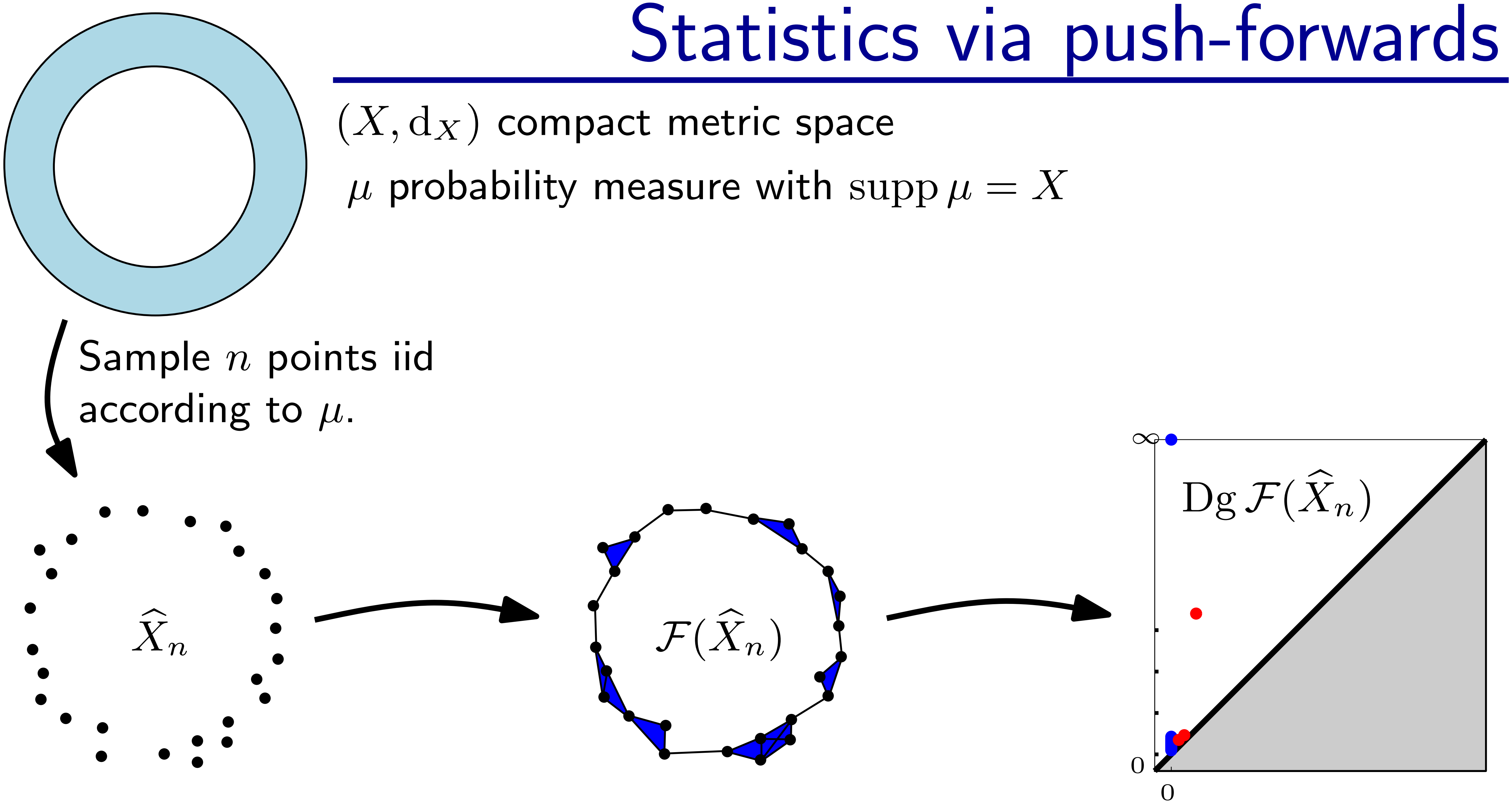
Given $D_1, \cdots, D_n$: persistence diagrams, is the following set empty?

$$\arg\min_{D} \sum_{i=1}^{n} W_p(D, D_i)^2$$

Theorem [Turner et al. 2012]: For any finite $p \geq 1$, the space of persistence diagrams equipped with W_p is *Polish* (complete and separable).

Corollary: Fréchet means exist... but are not unique (+hard to compute)

Statistics via push-forwards



(X, d_X) compact metric space

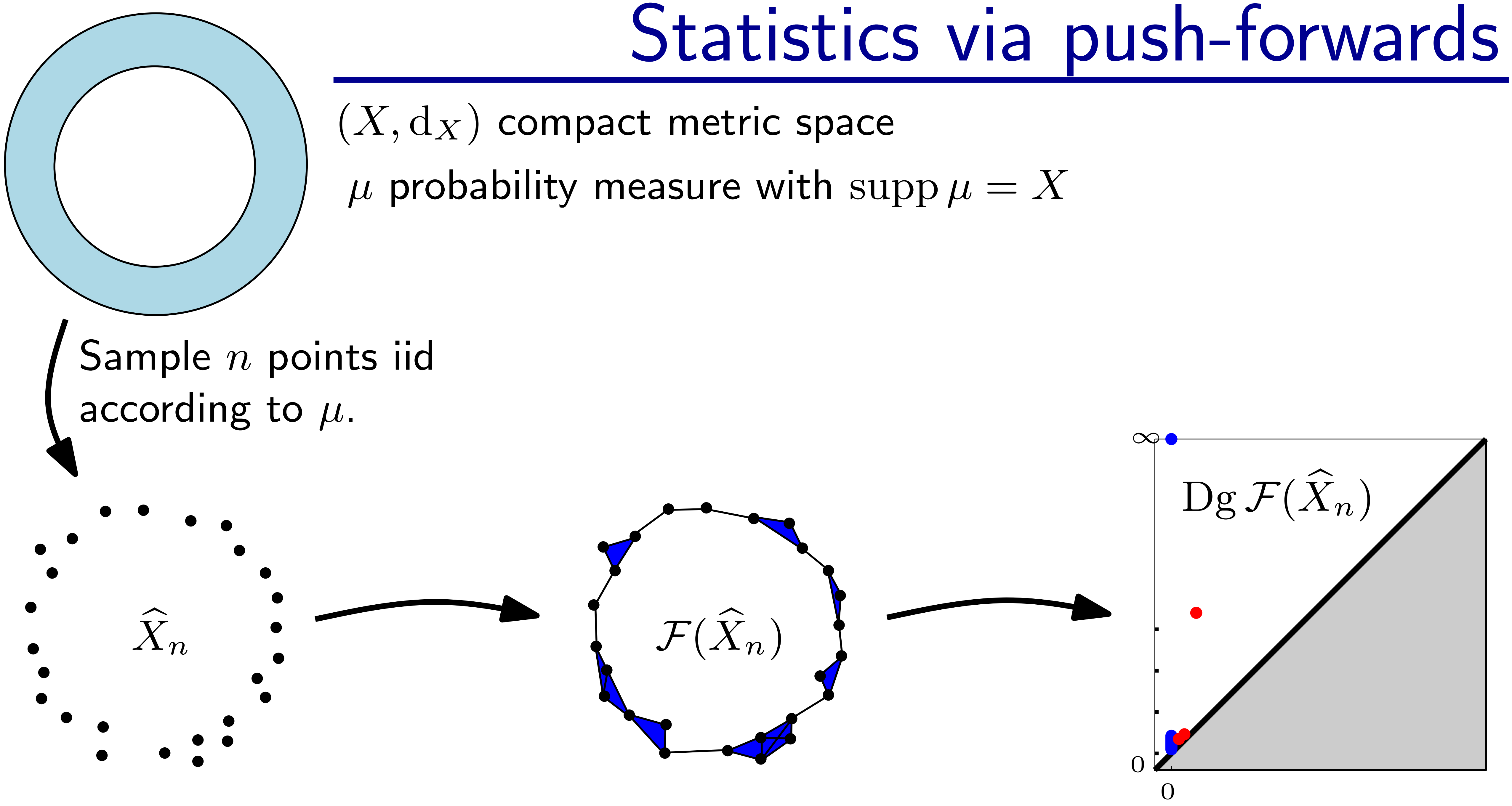
μ probability measure with $\operatorname{supp}\mu = X$

Sample n points iid according to μ.

Questions:

- Statistical properties of the estimator $\operatorname{Dg}\mathcal{F}(\widehat{X}_n)$?
- Convergence to the ground truth $\operatorname{Dg}\mathcal{F}(X)$? Deviation bounds?

Statistics via push-forwards



(X, d_X) compact metric space

μ probability measure with $\operatorname{supp}\mu = X$

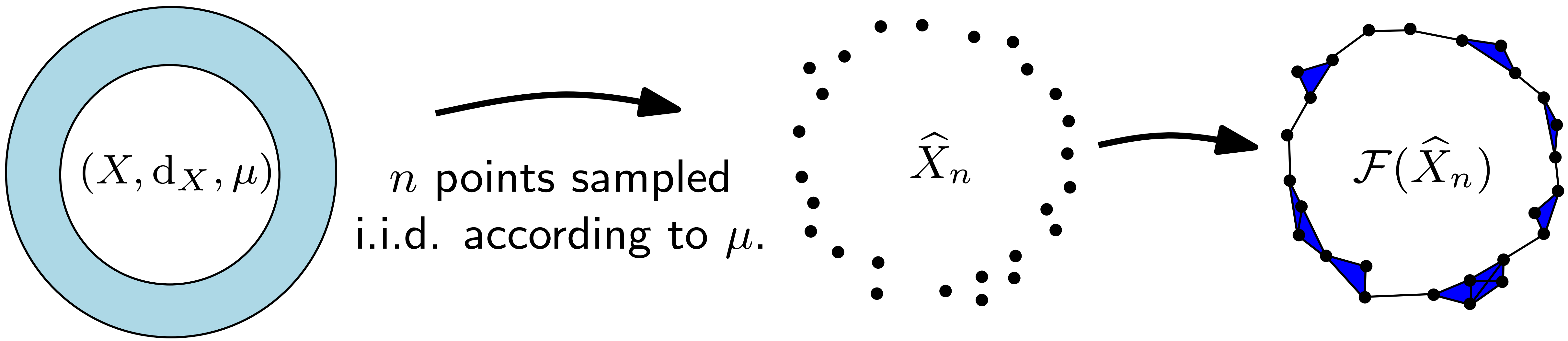
Sample n points iid according to μ.

Stability thm: $\mathrm{d}_B(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)) \leq 2\mathrm{d}_{\mathrm{H}}(\widehat{X}_n, X)$

$\Rightarrow$ for any $\varepsilon > 0$,

$$\mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X),\right) > \varepsilon\right) \leq \mathbb{P}\left(\mathrm{d}_{\mathrm{H}}(\widehat{X}_n, X) > \frac{\varepsilon}{2}\right)$$

Deviation inequality / rate of convergence



n points sampled i.i.d. according to μ.

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

Theorem [Chazal, Glisse, Labruère, Michel 2014-15]:

If μ is (a, b)-standard then for any $\varepsilon > 0$:

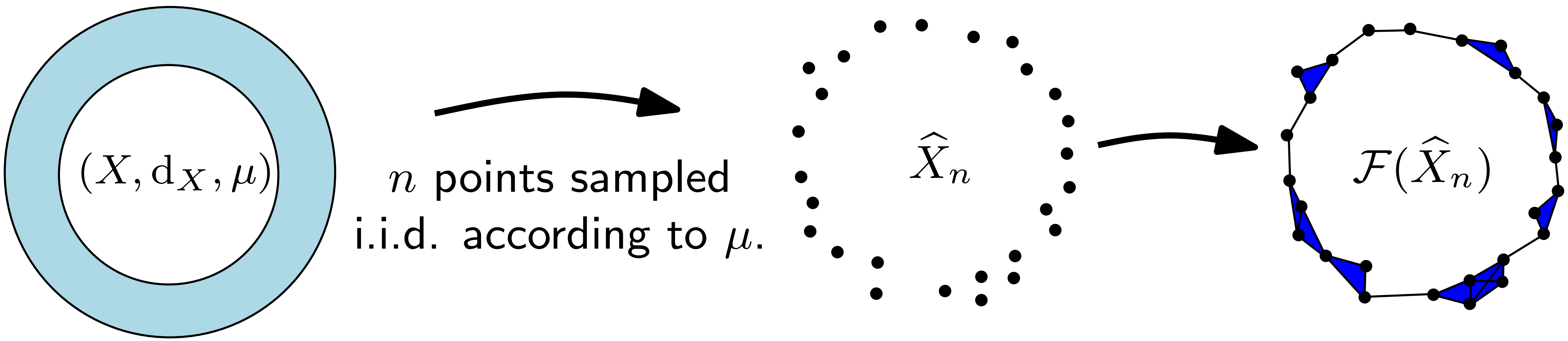
$$\mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > \varepsilon\right) \leq \frac{8^b}{a\varepsilon^b}\exp(-na\varepsilon^b)$$

Corollary [Chazal, Glisse, Labruère, Michel 2014-15]:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n),\ \mathrm{Dg}\,\mathcal{F}(X)\right)\right] \leq C\,\left(\frac{\log n}{n}\right)^{1/b},$$

where C depends only on a, b. Moreover, the estimator $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ is **minimax optimal** (up to a $\log n$ factor) on the space $\mathcal{P}$ of (a, b)-standard probability measures on X.

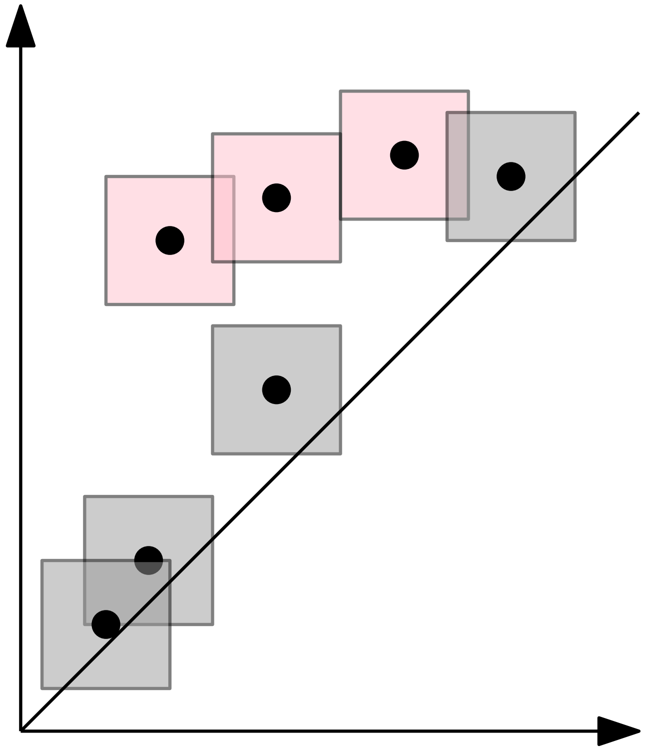
Deviation inequality / rate of convergence

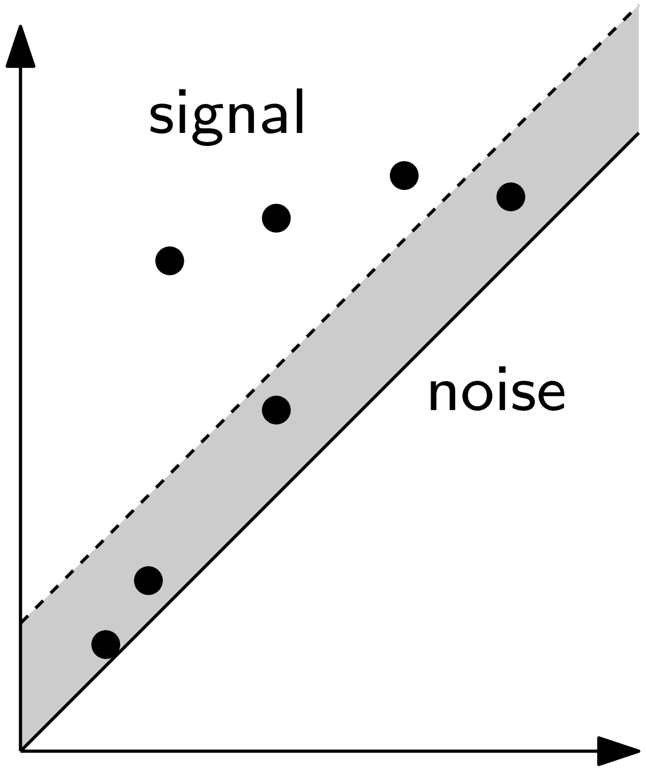


n points sampled
i.i.d. according to μ.

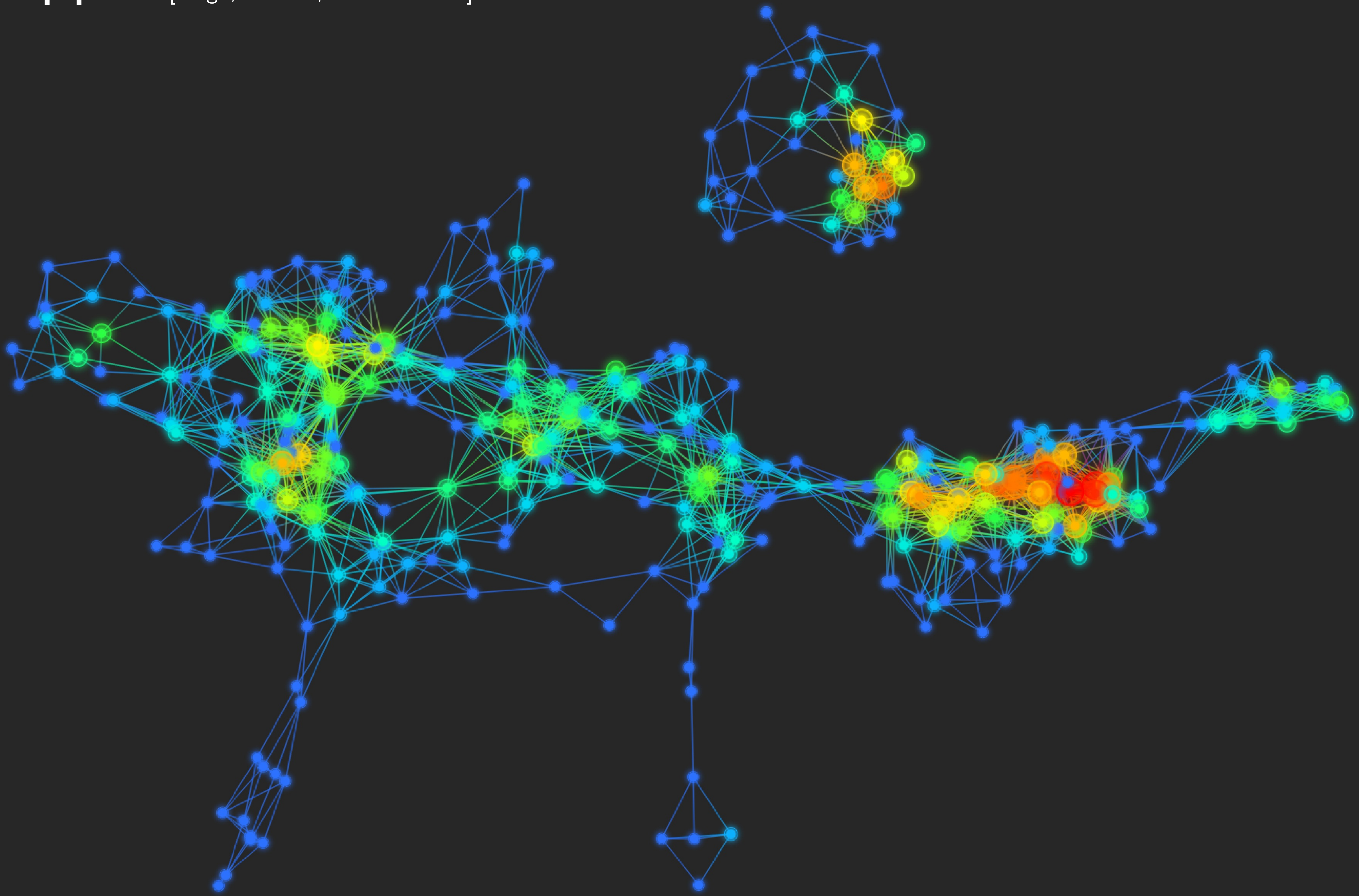
For $a, b > 0$, μ satisfies the (a, b)-**standard** assumption if for any $x \in X$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

→ subsampling / bootstrap → confidence regions, etc.



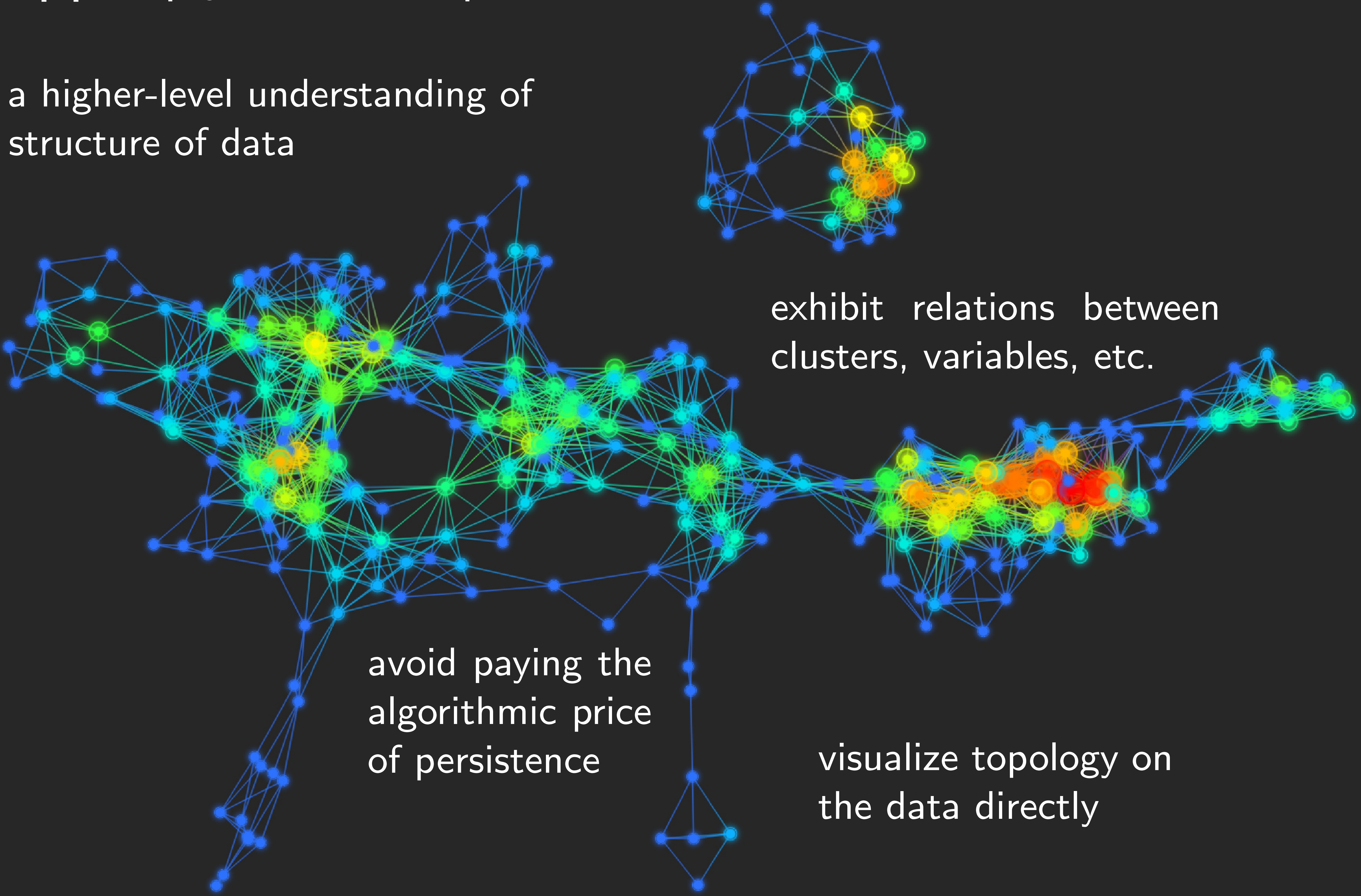


Mapper [Singh, Mémoli, Carlsson 2007]



same principle as persistence: summarize the topological structure of a map $f : X \to \mathbb{R}$

Mapper [Singh, Mémoli, Carlsson 2007]



get a higher-level understanding of
the structure of data

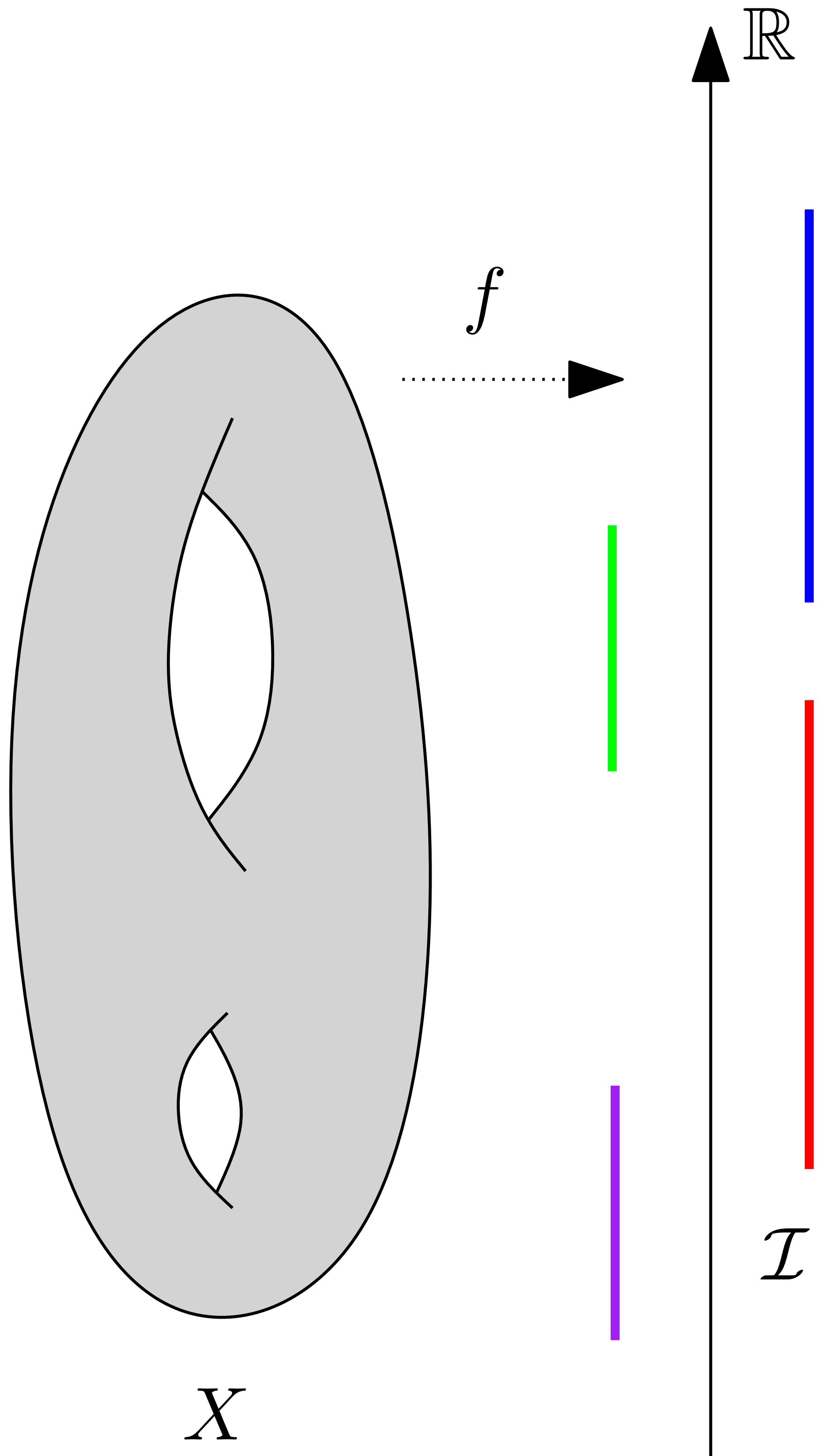
exhibit relations between
clusters, variables, etc.

avoid paying the
algorithmic price
of persistence

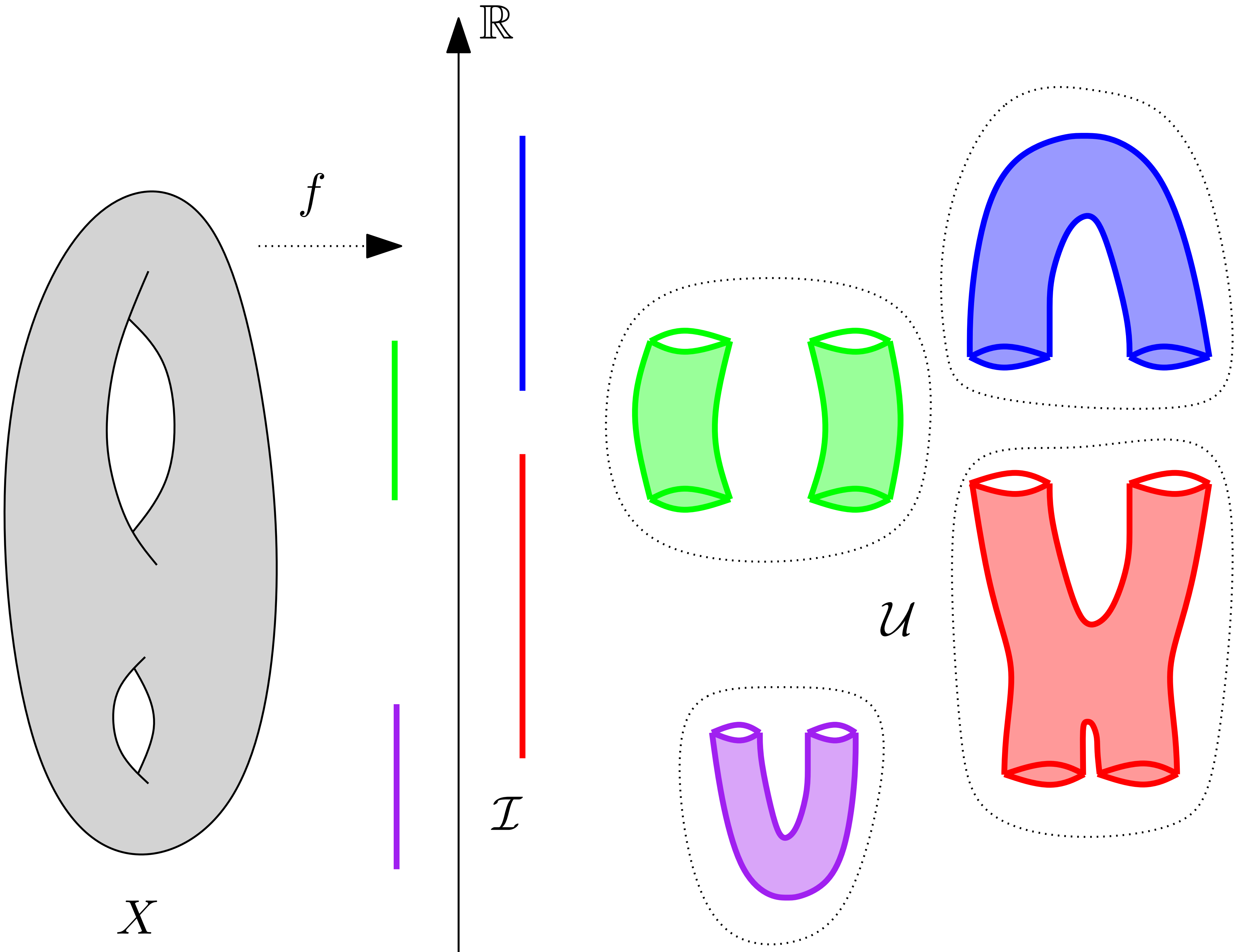
visualize topology on
the data directly

same principle as persistence: summarize the topological structure of a map $f : X \to \mathbb{R}$

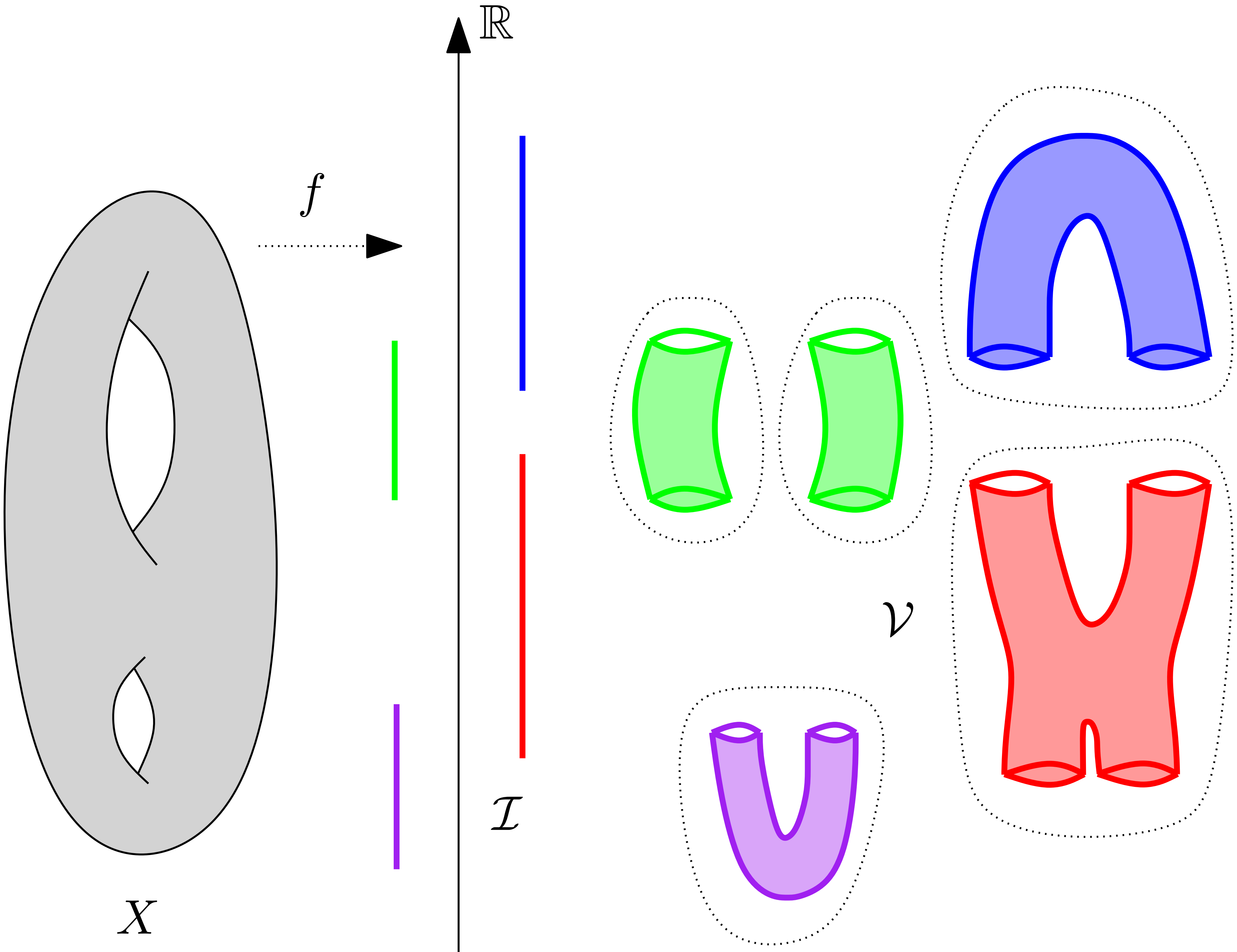
Mapper in the continuous setting



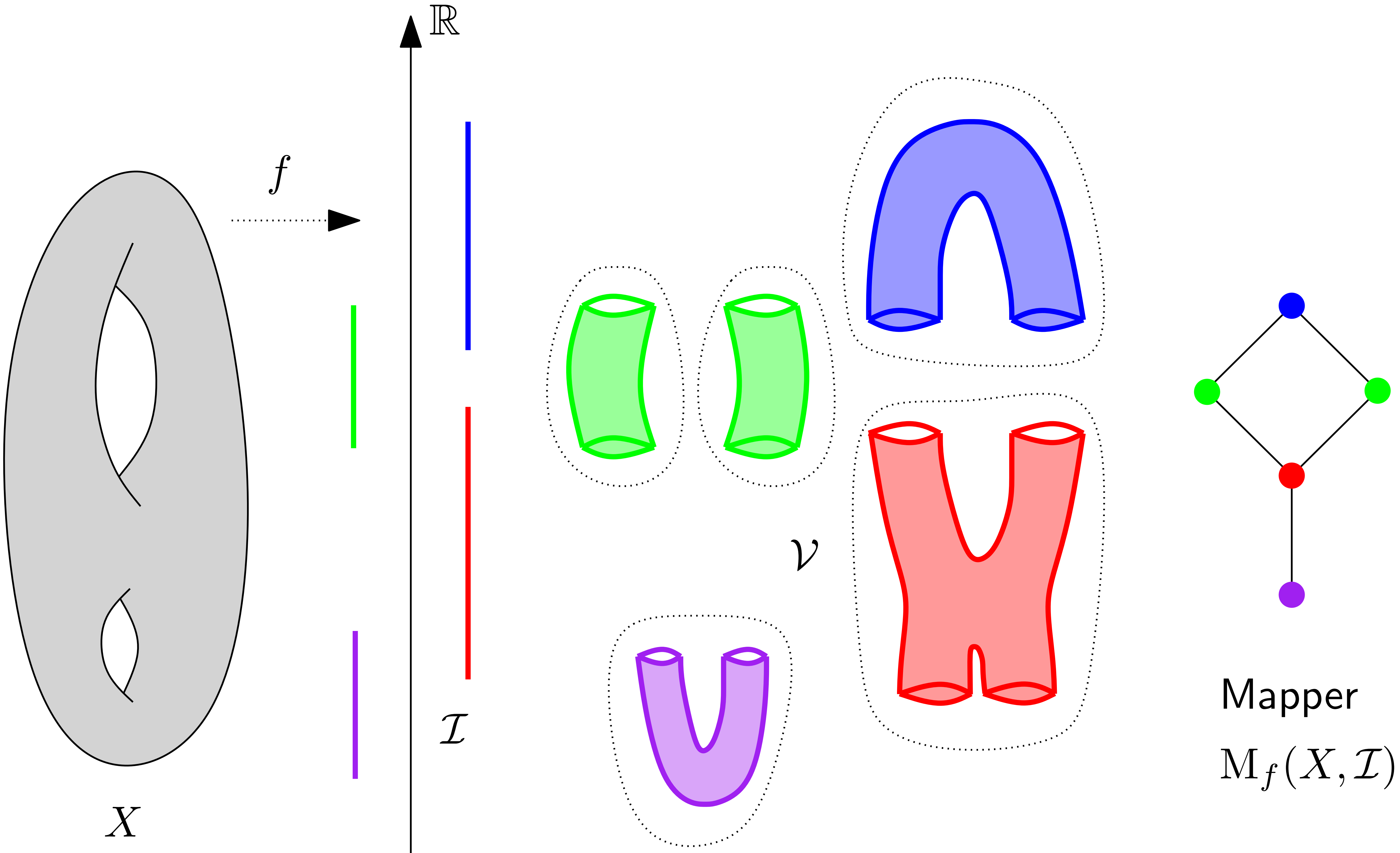
Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting

Input:

- topological space X
- continuous function $f : X \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals: $\mathrm{im} f \subseteq \bigcup_{I \in \mathcal{I}} I$

Method:

- Compute *pullback cover* $\mathcal{U}$ of X: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$
- Refine $\mathcal{U}$ by separating each of its elements into its various connected components in X → connected cover $\mathcal{V}$
- The Mapper is the *nerve* of $\mathcal{V}$:
  - 1 vertex per element $V \in \mathcal{V}$
  - 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$
  - 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

Mapper in practice

Input:

- point cloud $P \subseteq X$ with metric d_P
- continuous function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals: $\mathrm{im} f \subseteq \bigcup_{I \in \mathcal{I}} I$

Method: • Compute neighborhood graph $G = (P, E)$

• Compute *pullback cover* $\mathcal{U}$ of P: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$

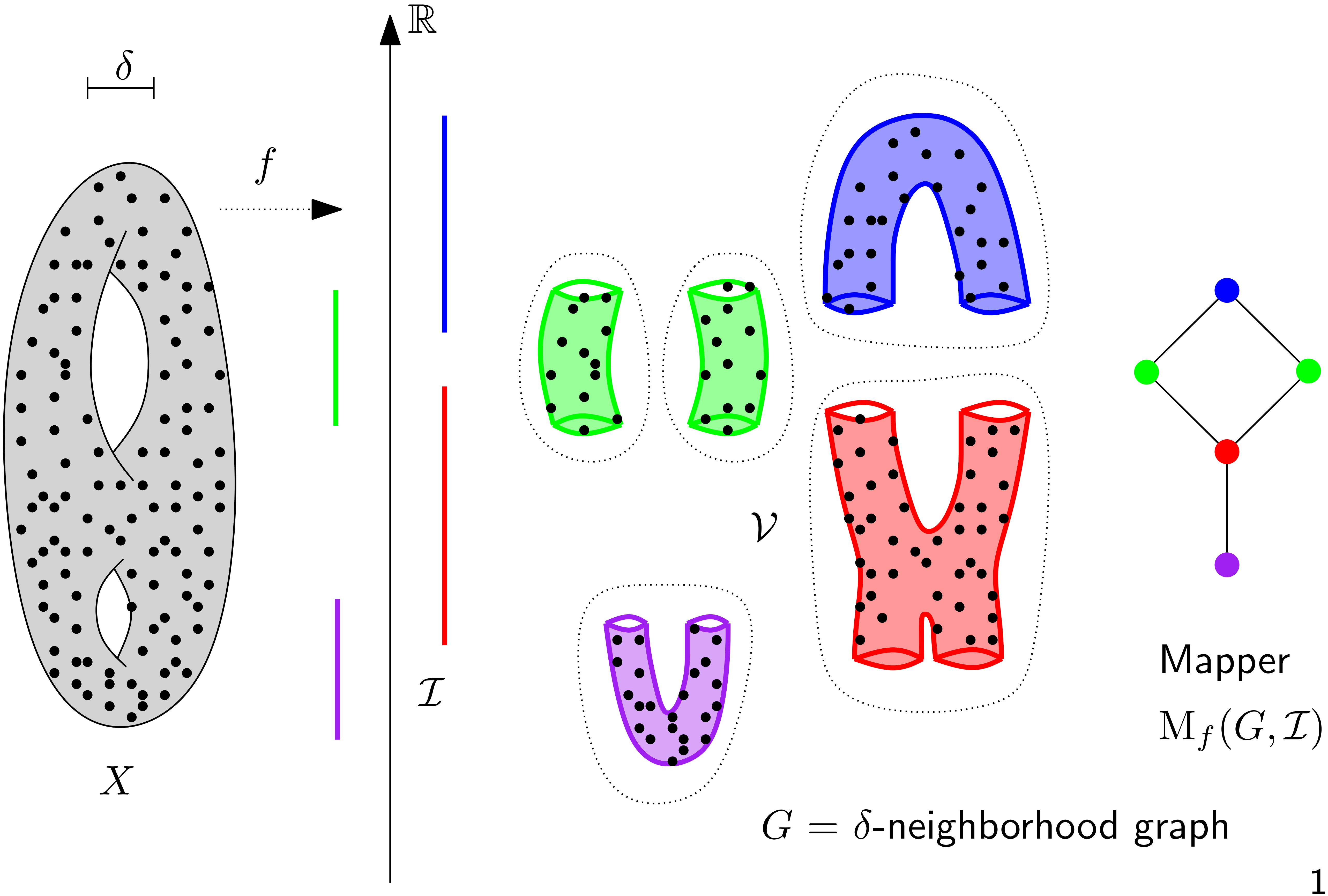
• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in G → connected cover $\mathcal{V}$

• The Mapper is the *nerve* of $\mathcal{V}$:

(intersections materialized by data points)

- 1 vertex per element $V \in \mathcal{V}$
- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$
- 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

Mapper in practice



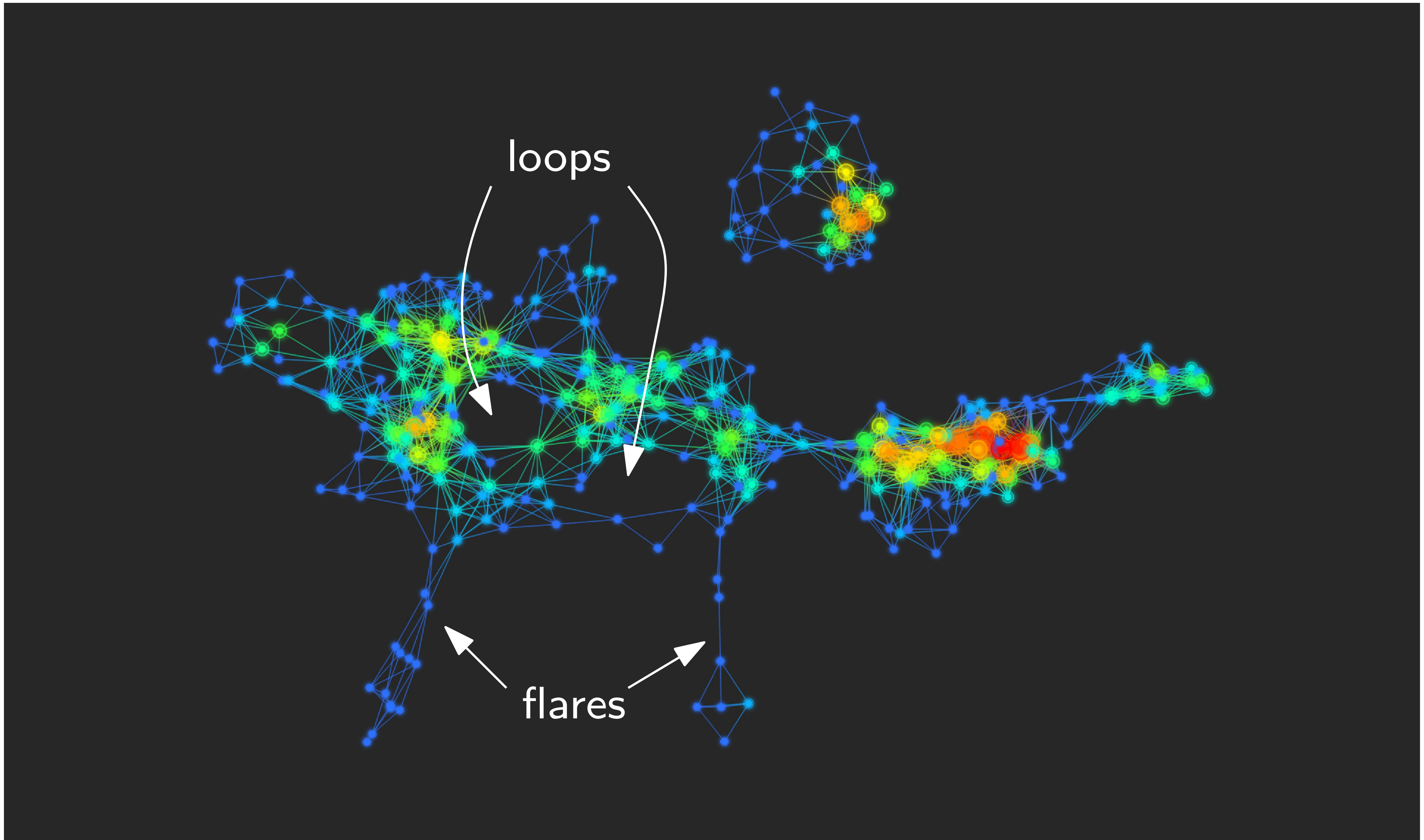
$G = \delta$-neighborhood graph

Mapper in applications

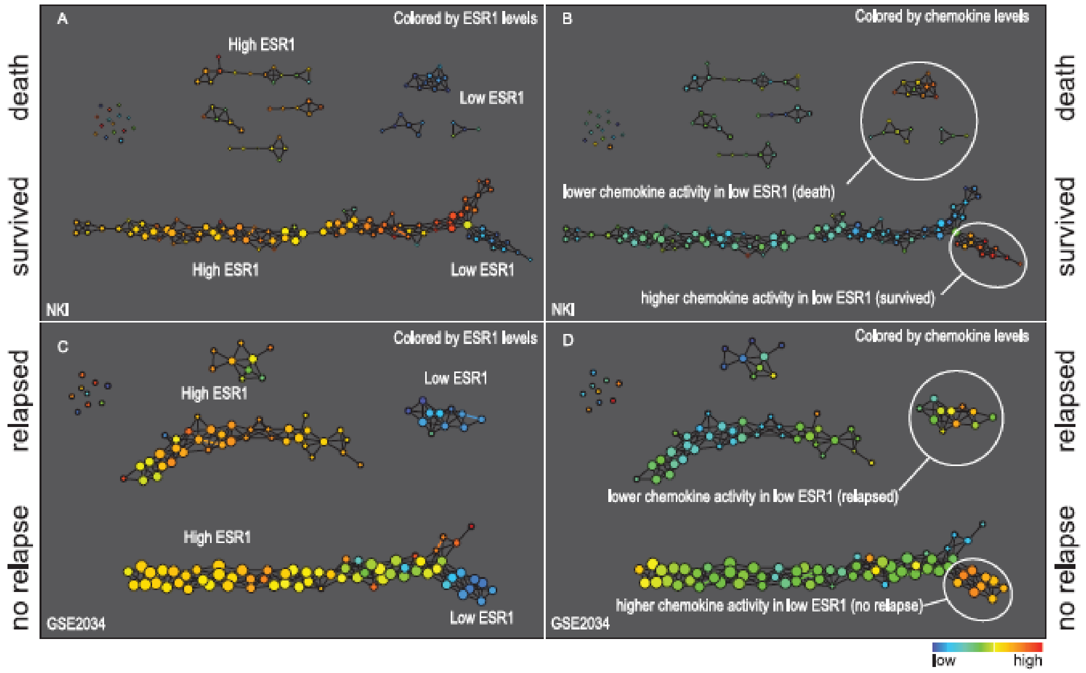
Two types of applications:

- clustering
- feature selection

principle: identify statistically relevant sub-populations through patterns (flares, loops)



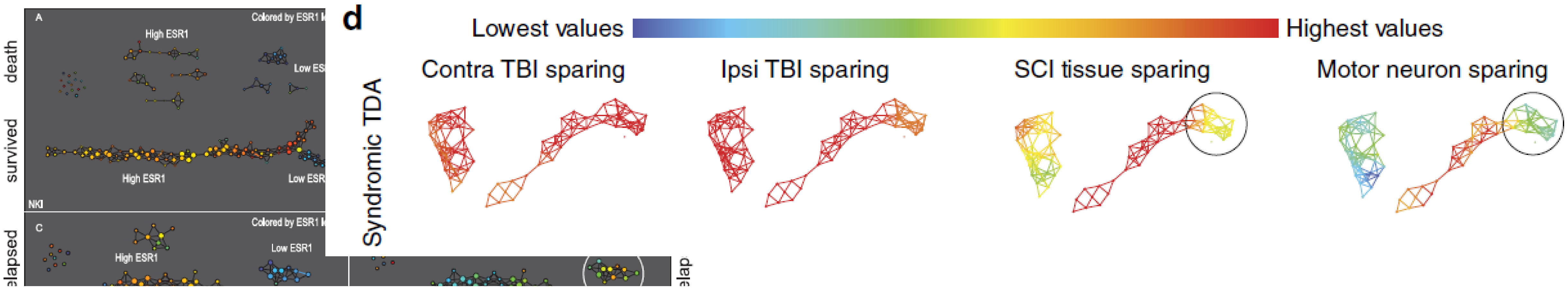
Mapper in applications

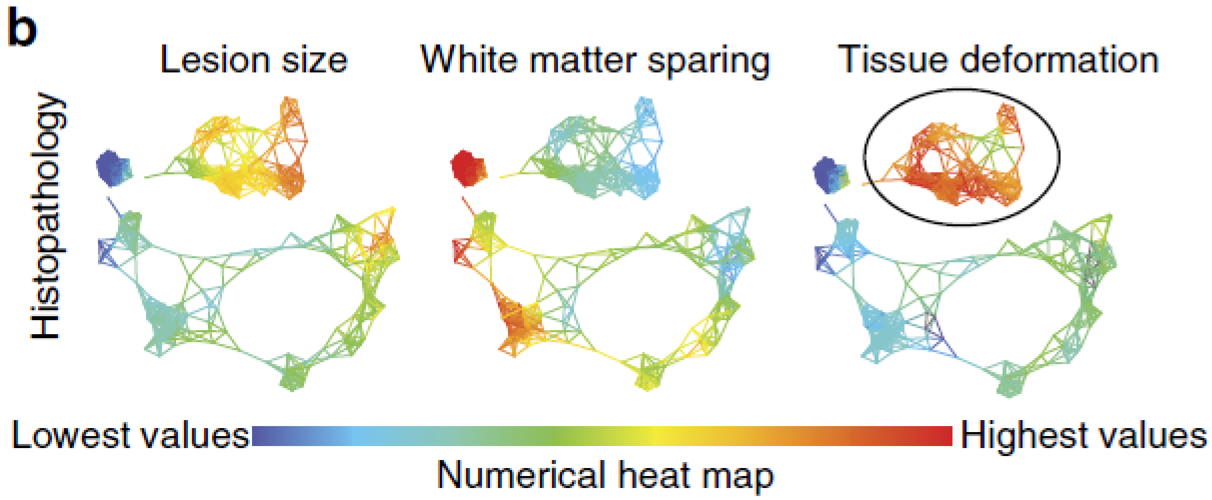


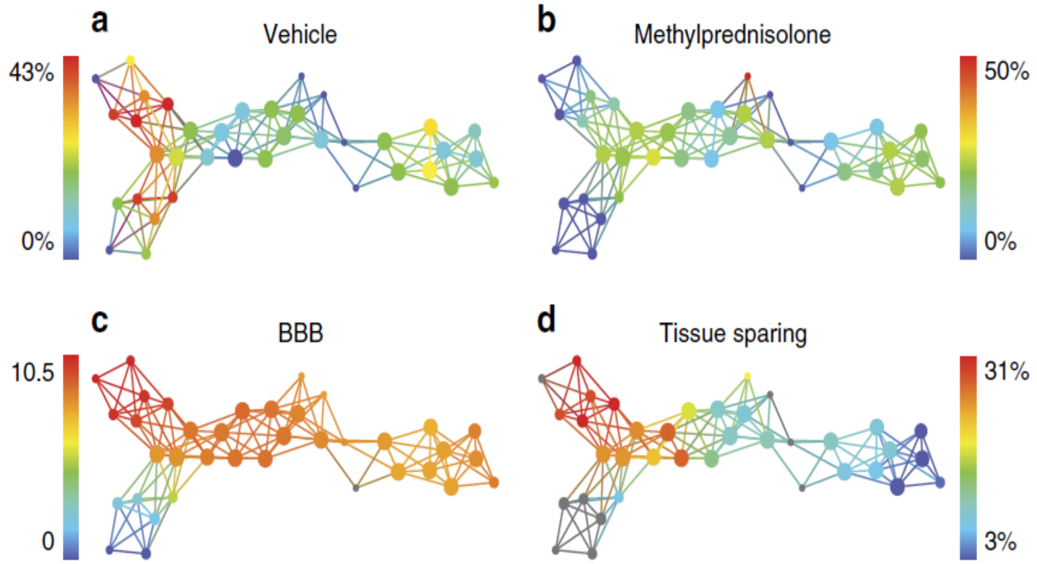
breast cancer subtype

[Nicolau et al. 2011]

Mapper in applications



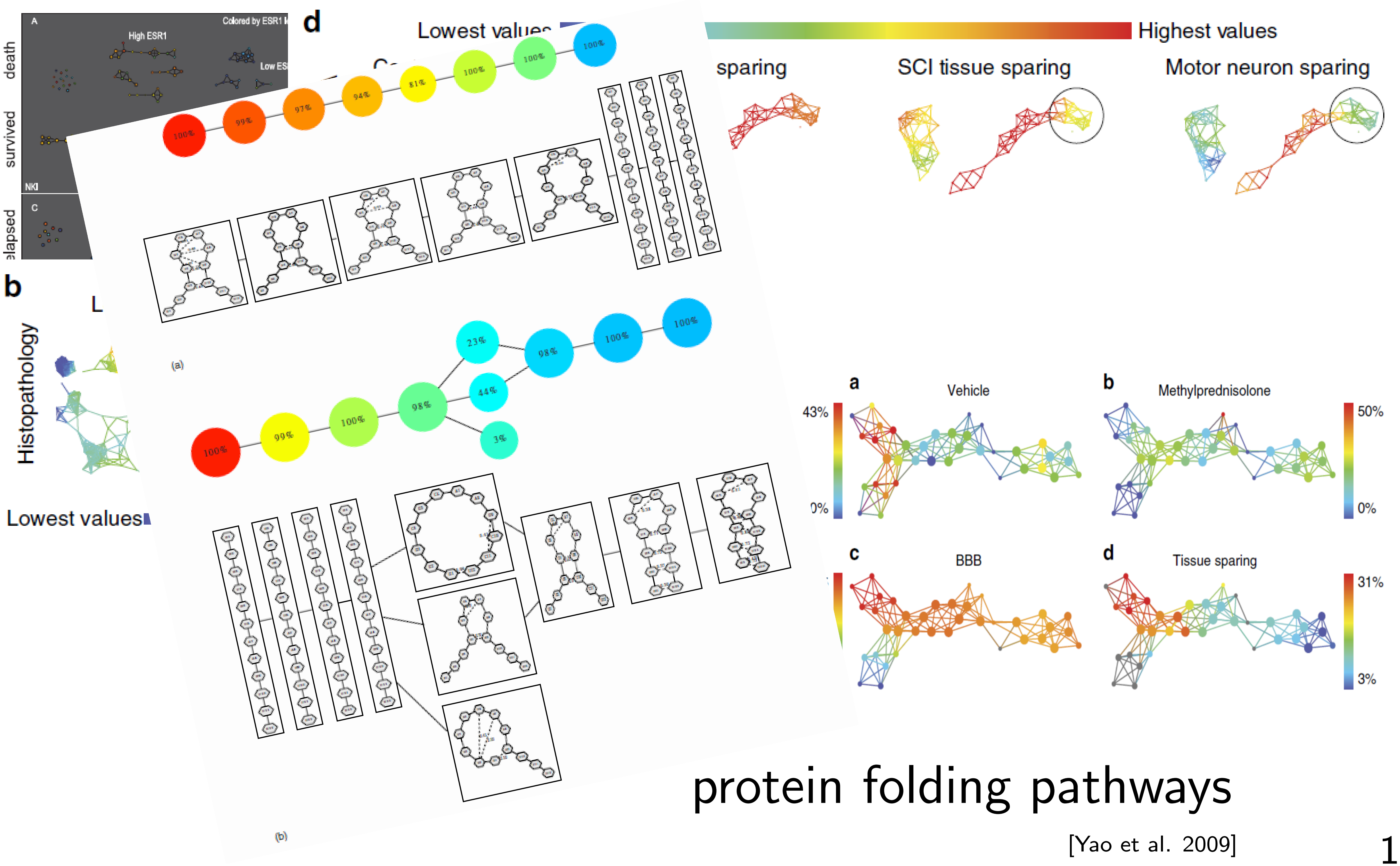




recovery from spinal cord injuries

[Nielson et al. 2015]

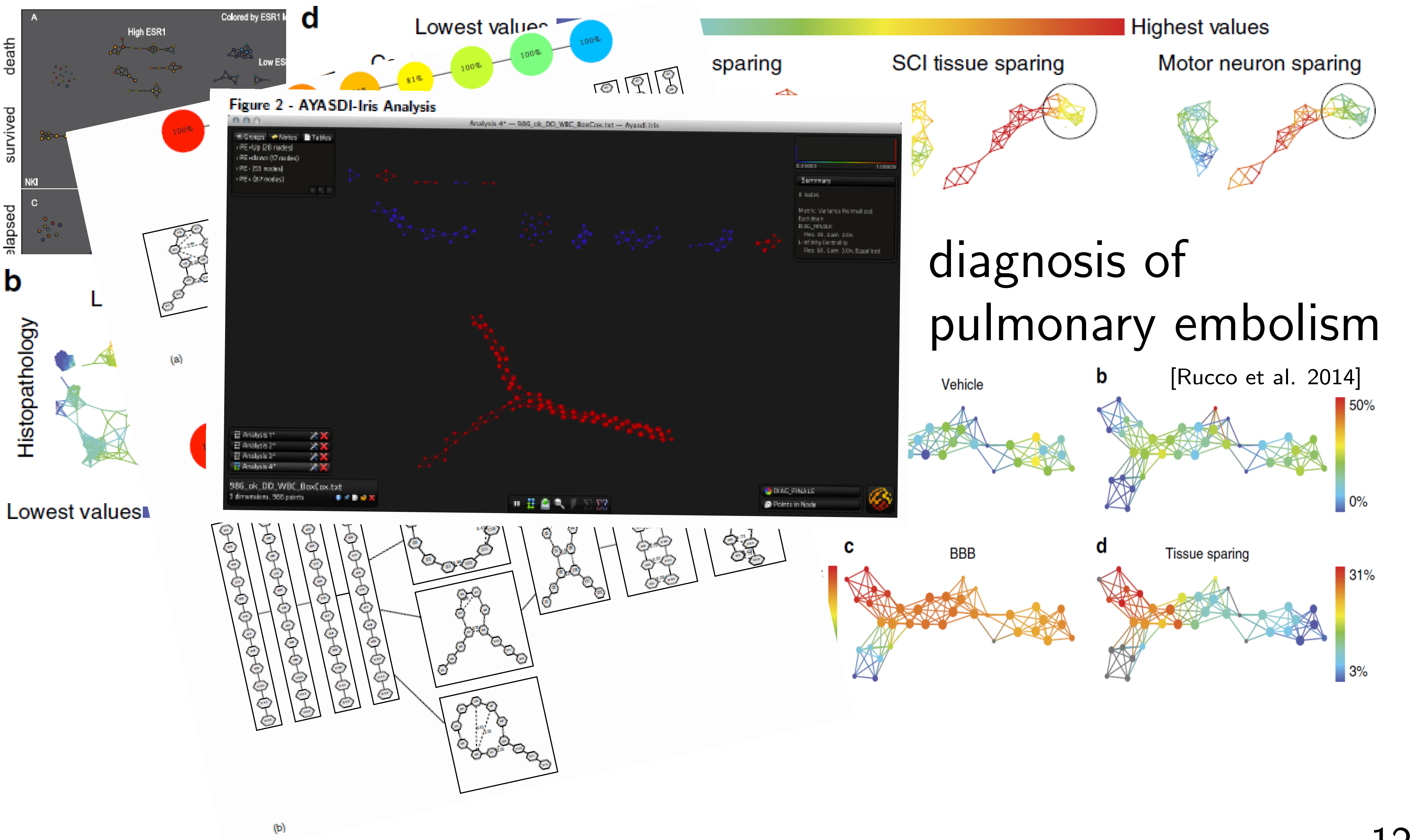
Mapper in applications



protein folding pathways

[Yao et al. 2009]

Mapper in applications



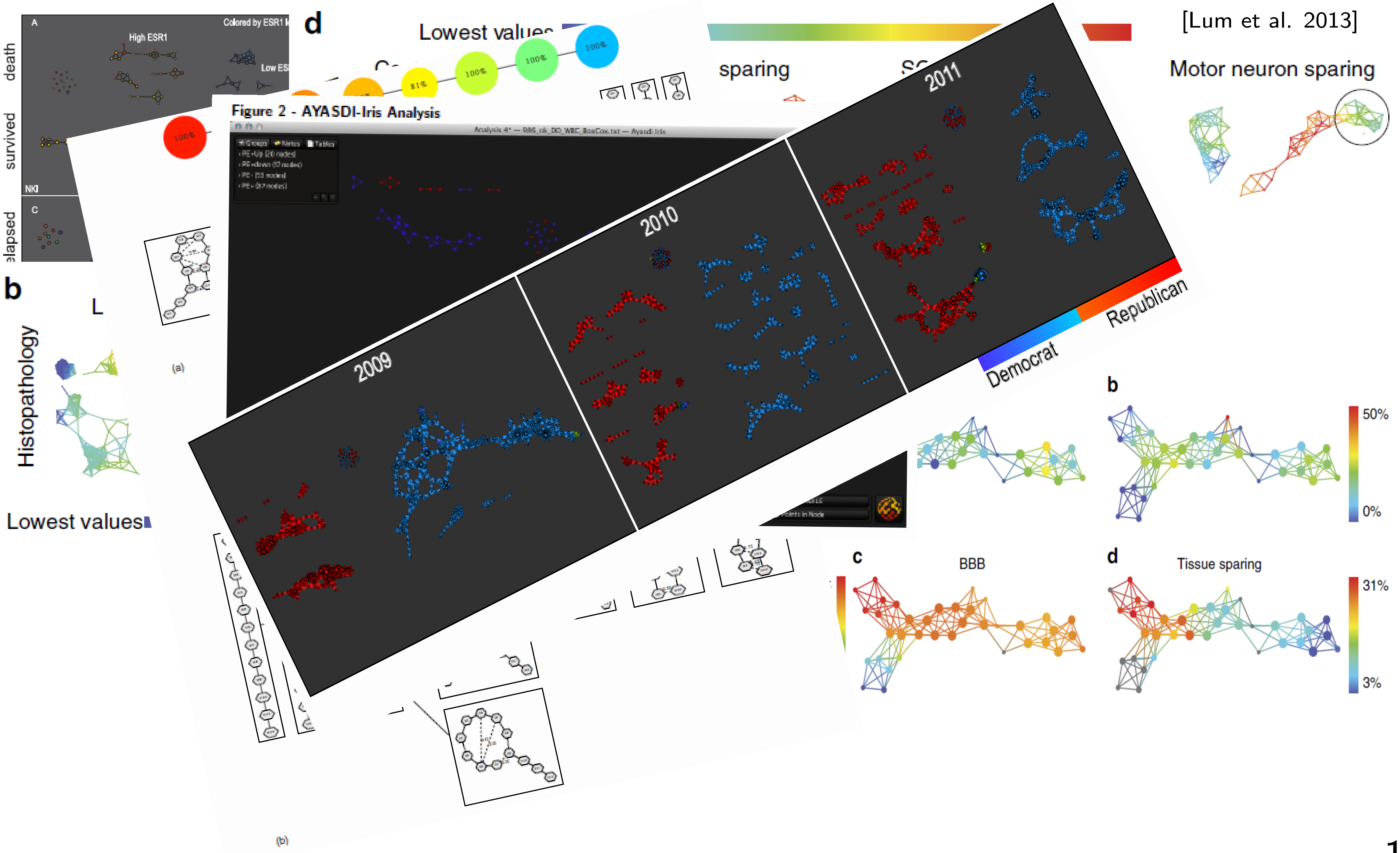
diagnosis of
pulmonary embolism

[Rucco et al. 2014]

Mapper in applications

implicit networks in the US house of representatives

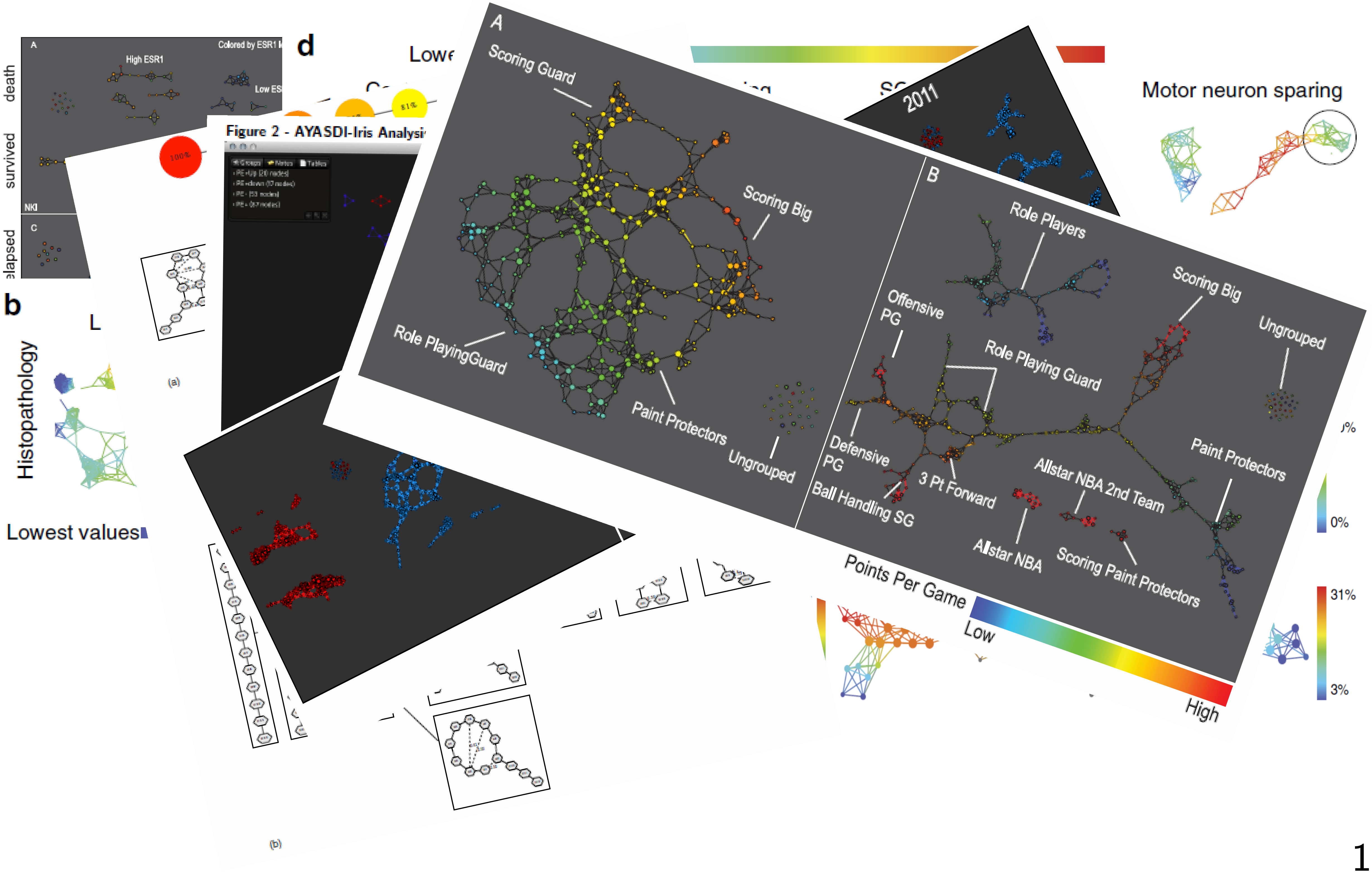
[Lum et al. 2013]



Mapper in applications

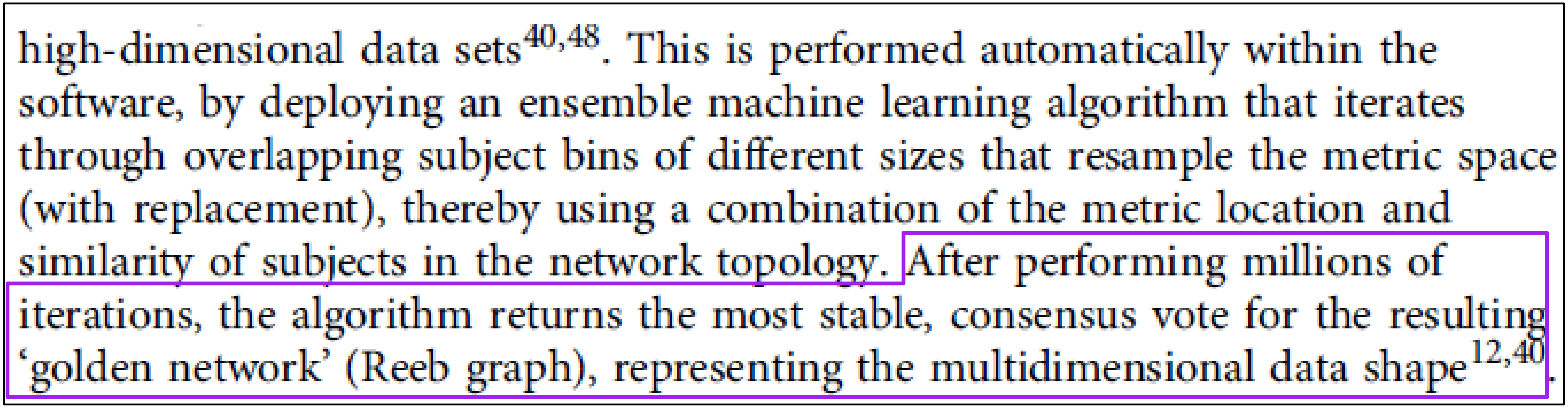
classification of NBA players

[Alagappan 2012]



Choice of parameters

→ in practice: trial-and-error



high-dimensional data sets[40,48]. This is performed automatically within the software, by deploying an ensemble machine learning algorithm that iterates through overlapping subject bins of different sizes that resample the metric space (with replacement), thereby using a combination of the metric location and similarity of subjects in the network topology. After performing millions of iterations, the algorithm returns the most stable, consensus vote for the resulting 'golden network' (Reeb graph), representing the multidimensional data shape[12,40].

Nielson et al.: *Topological Data Analysis for Discovery in Preclinical Spinal Cord Injury and Traumatic Brain Injury*, Nature, 2015

Choice of parameters

Parameters:

lens | filter

- function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- neighborhood size δ

range scale

geometric scale

Choice of parameters

Parameters:

- function $f : P \to \mathbb{R}$

lens | filter

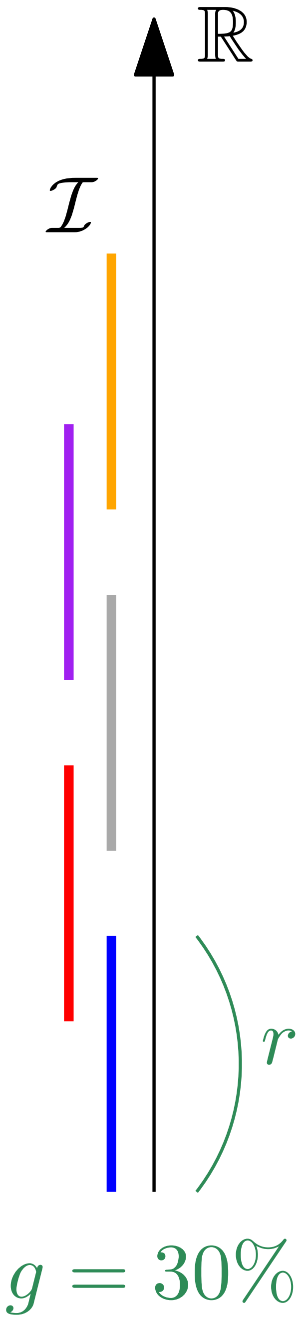
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- neighborhood size δ

range scale

geometric scale

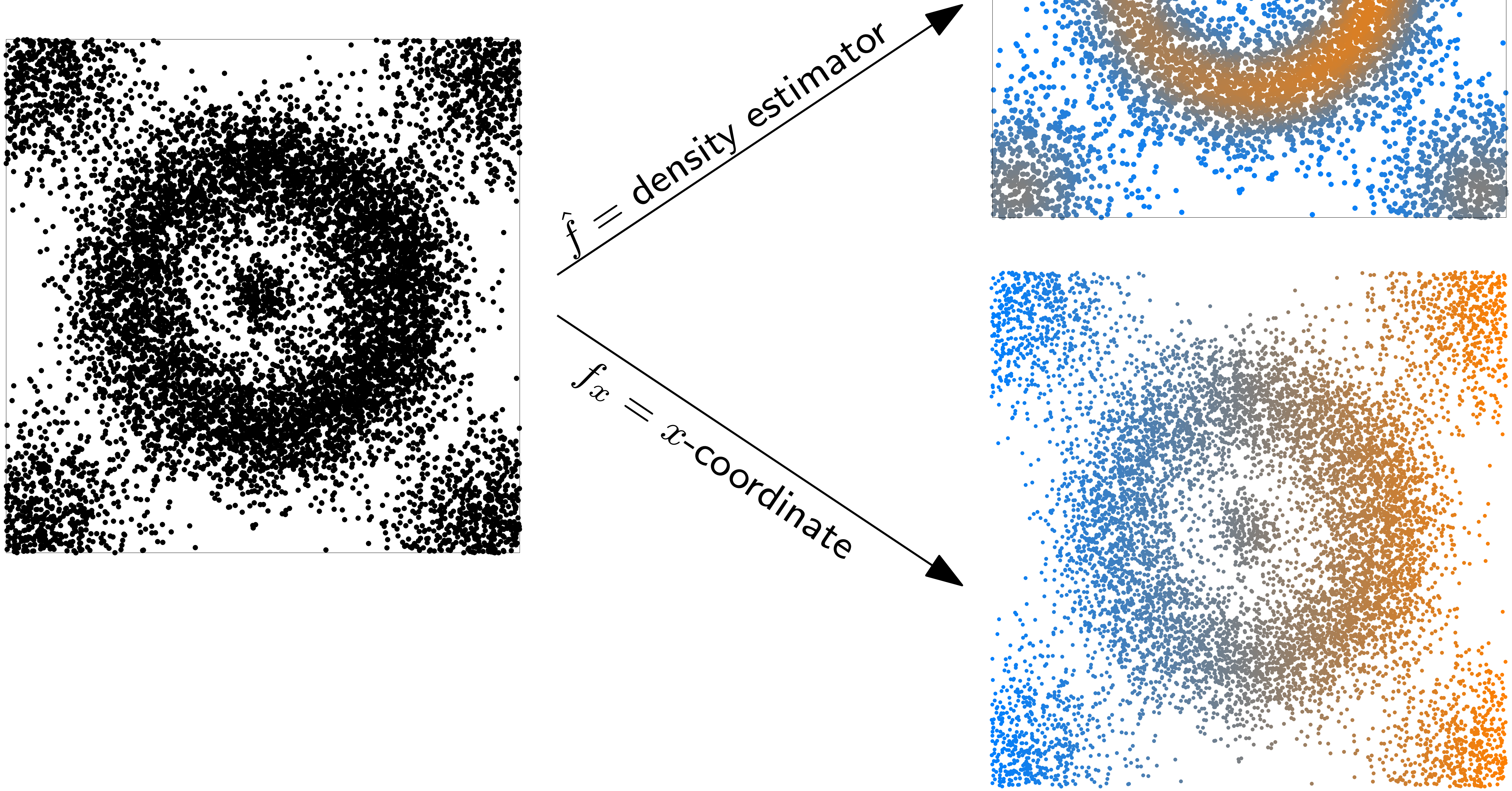
$\to$ uniform cover $\mathcal{I}$:

  - resolution / granularity: r (diameter of intervals)
  - gain: g (percentage of overlap)



Choice of parameters

Example: $P \subset \mathbb{R}^2$ sampled from a known probability distribution



$\hat{f}$ = density estimator

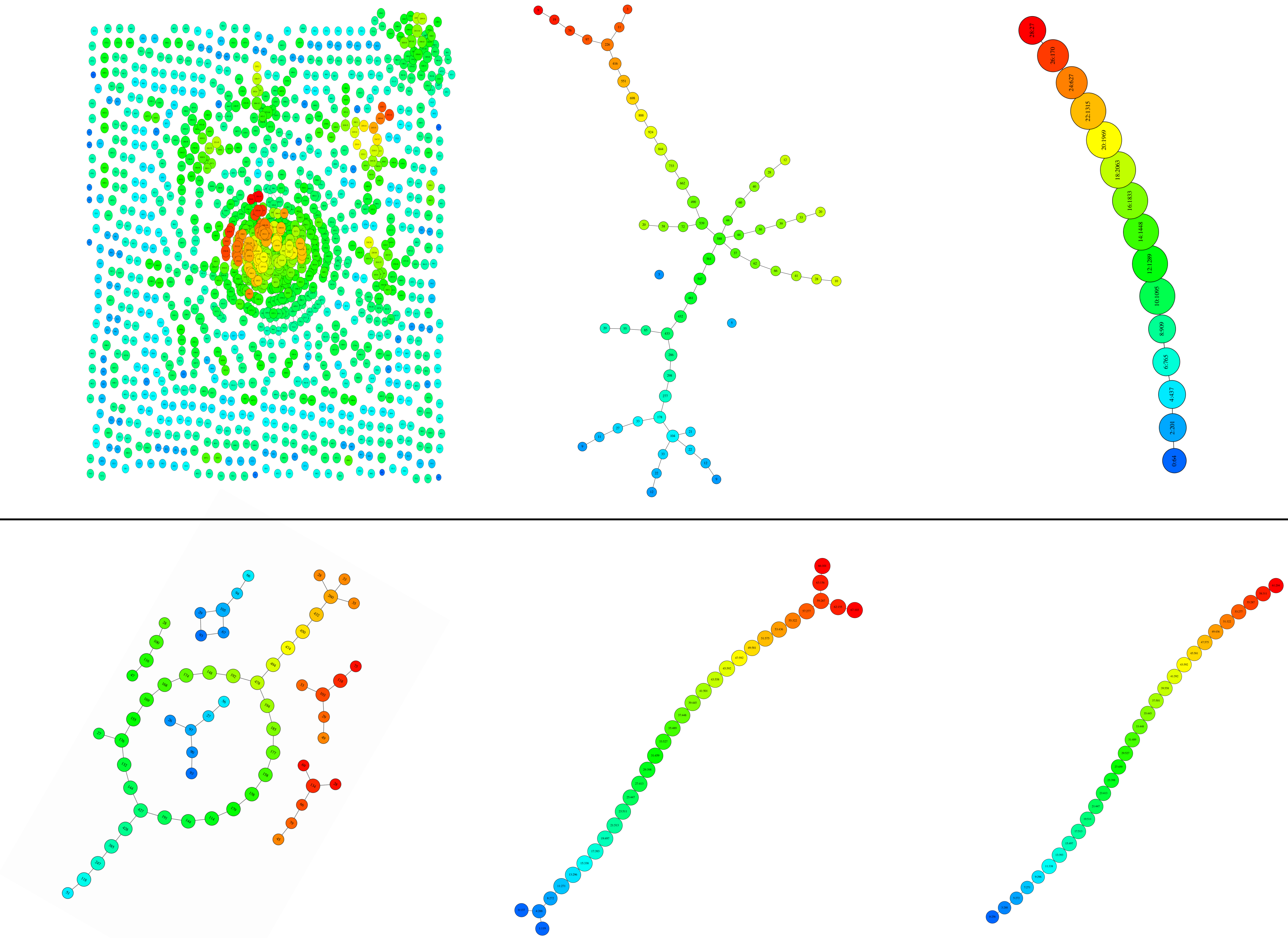
$f_x = x$-coordinate

Choice of parameters

$r = 30,\ g = 20\%$

$f = \hat{f}$

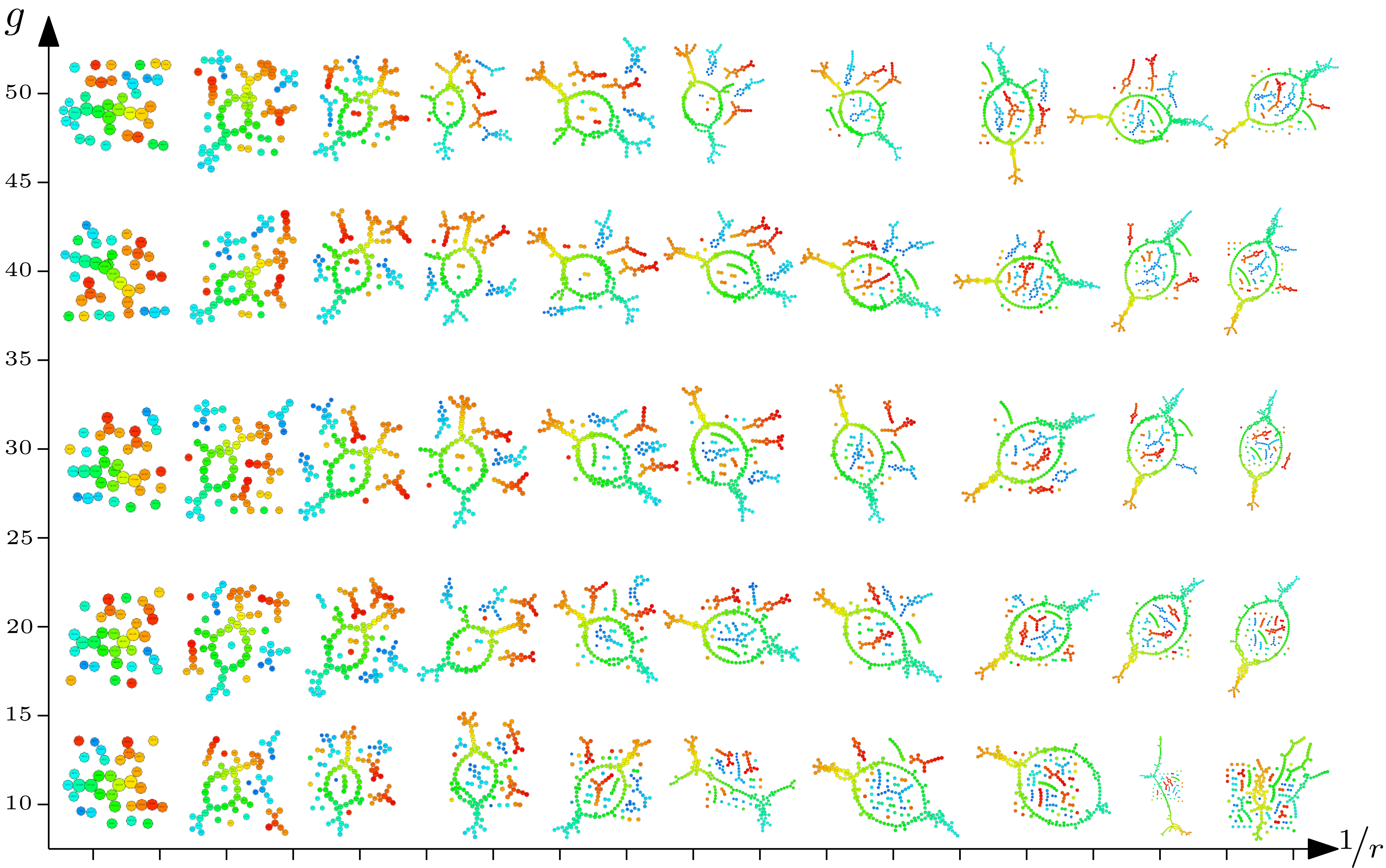
$f = f_x$



$\delta = 1\%$ $\qquad \delta = 10\%$ $\qquad \delta = 25\%$

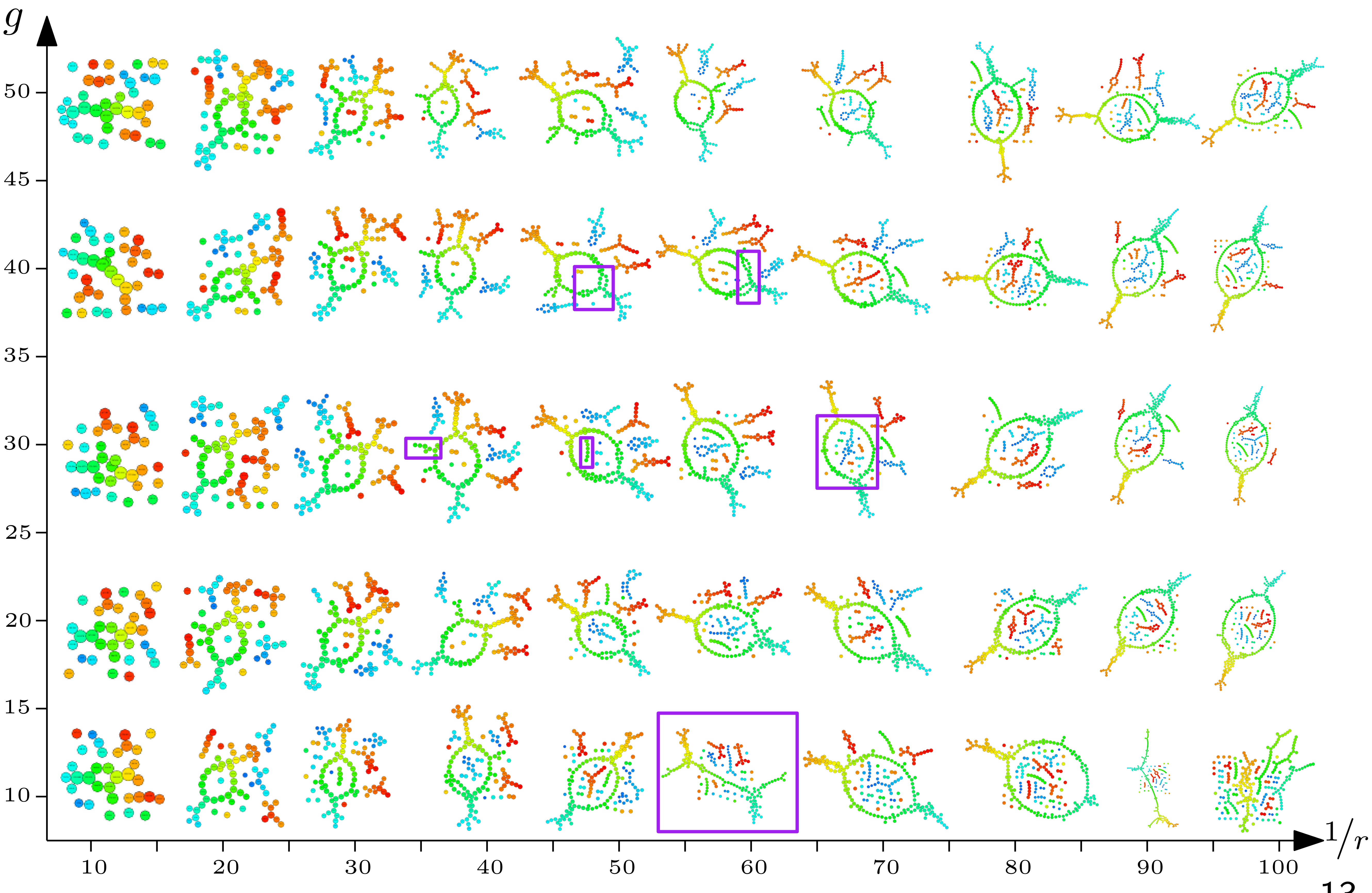
Choice of parameters

$f = f_x$, $\delta = 1\%$

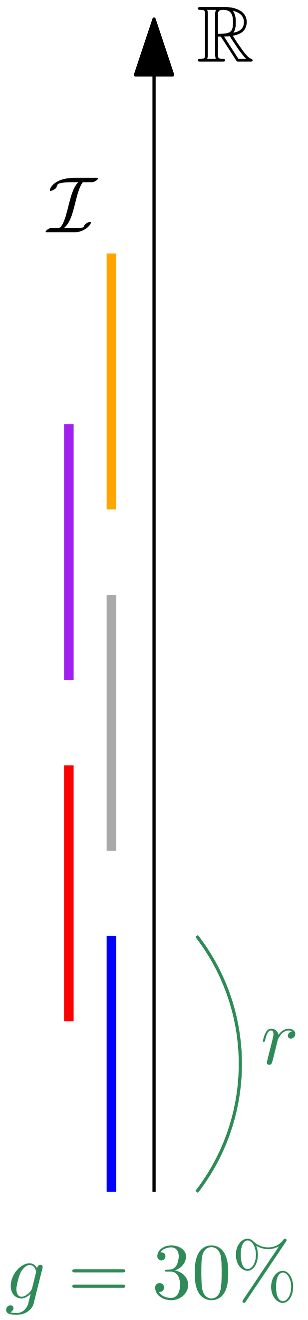


Choice of parameters

$f = f_x$, $\delta = 1\%$



Choice of parameters



Parameters:

- function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- neighborhood size δ

lens | filter

range scale

geometric scale

$\to$ uniform cover $\mathcal{I}$:

- resolution / granularity: r (diameter of intervals)
- gain: g (percentage of overlap)

Choice of parameters

Recent contributions:

$\rightarrow$ clarify the roles of r and g in the continuous setting

$\rightarrow$ introduce metrics between mappers

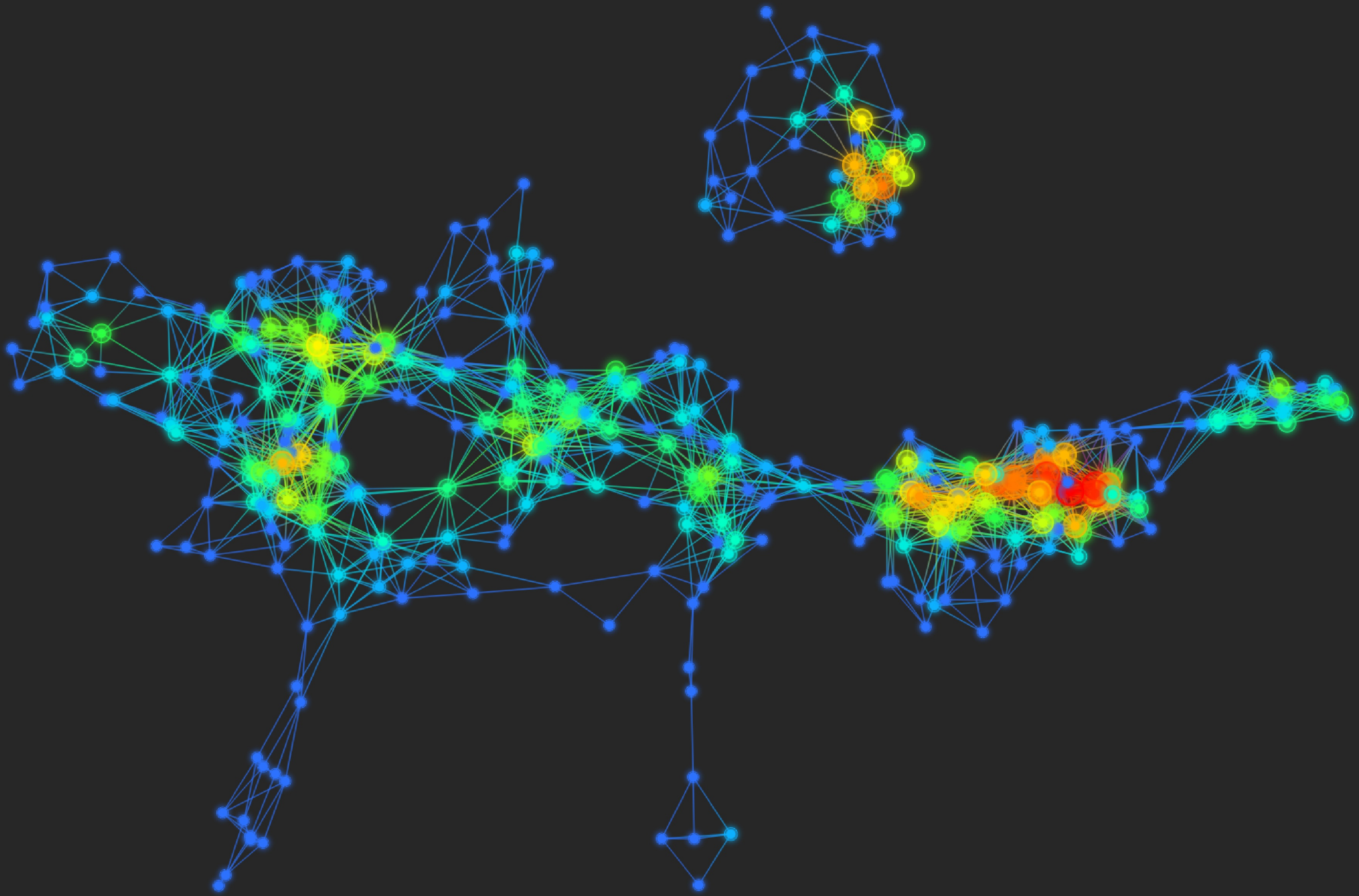
$\rightarrow$ establish stability and convergence results for Mappers

$\rightarrow$ relate discrete and continuous Mappers under conditions on δ

M. Carrière and S. O. *Structure and Stability of the 1-Dimensional Mapper*. 2016

E. Munch and B Wang. *Convergence between Categorical Representations of Reeb Space and Mapper*. 2016

V. de Silva, E. Munch and A. Patel. *Categorified Reeb Graphs*. 2015

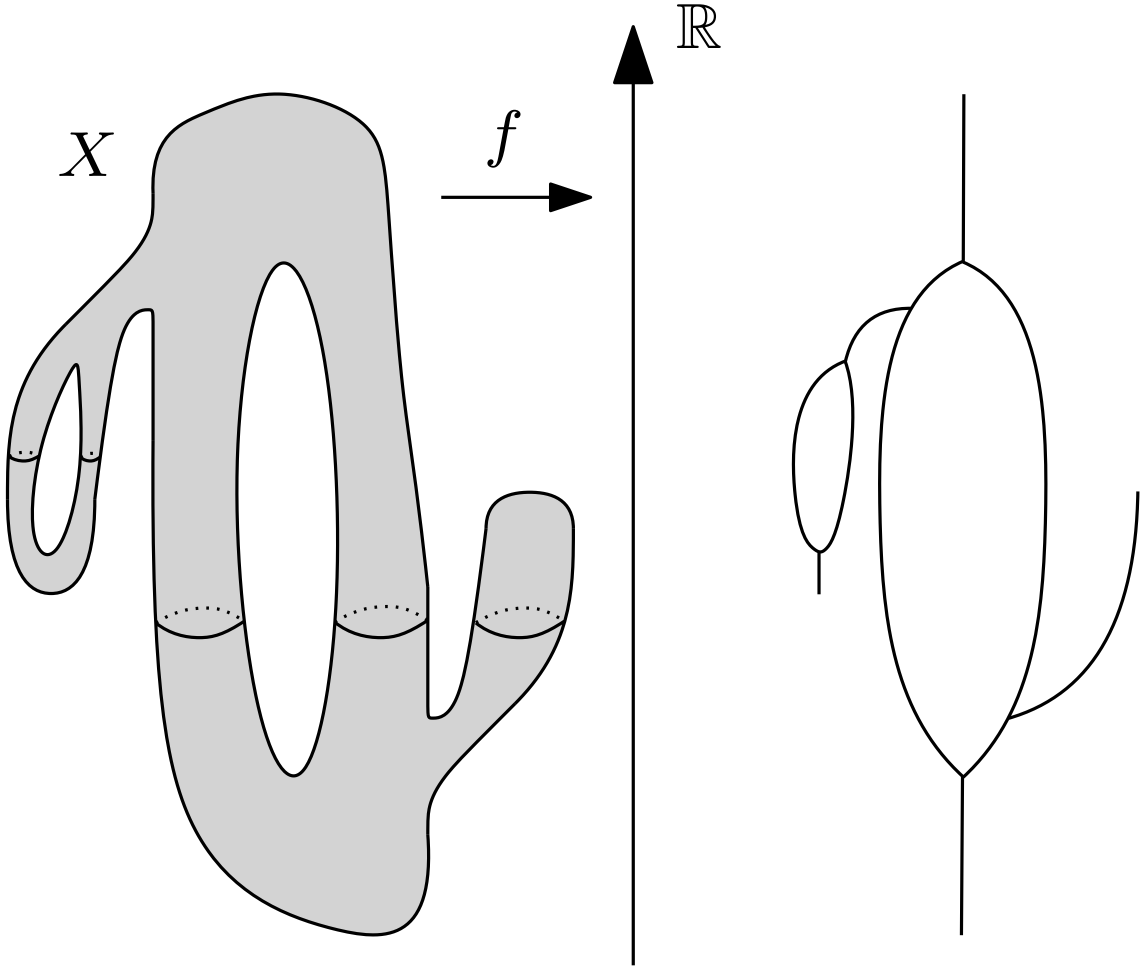


principle: summarize the topological structure of a map $f : X \to \mathbb{R}$ through a graph

Reeb Graph

$x \sim y \iff [\ f(x) = f(y)$ and x, y belong to same cc of $f^{-1}(\{f(x)\})\ ]$

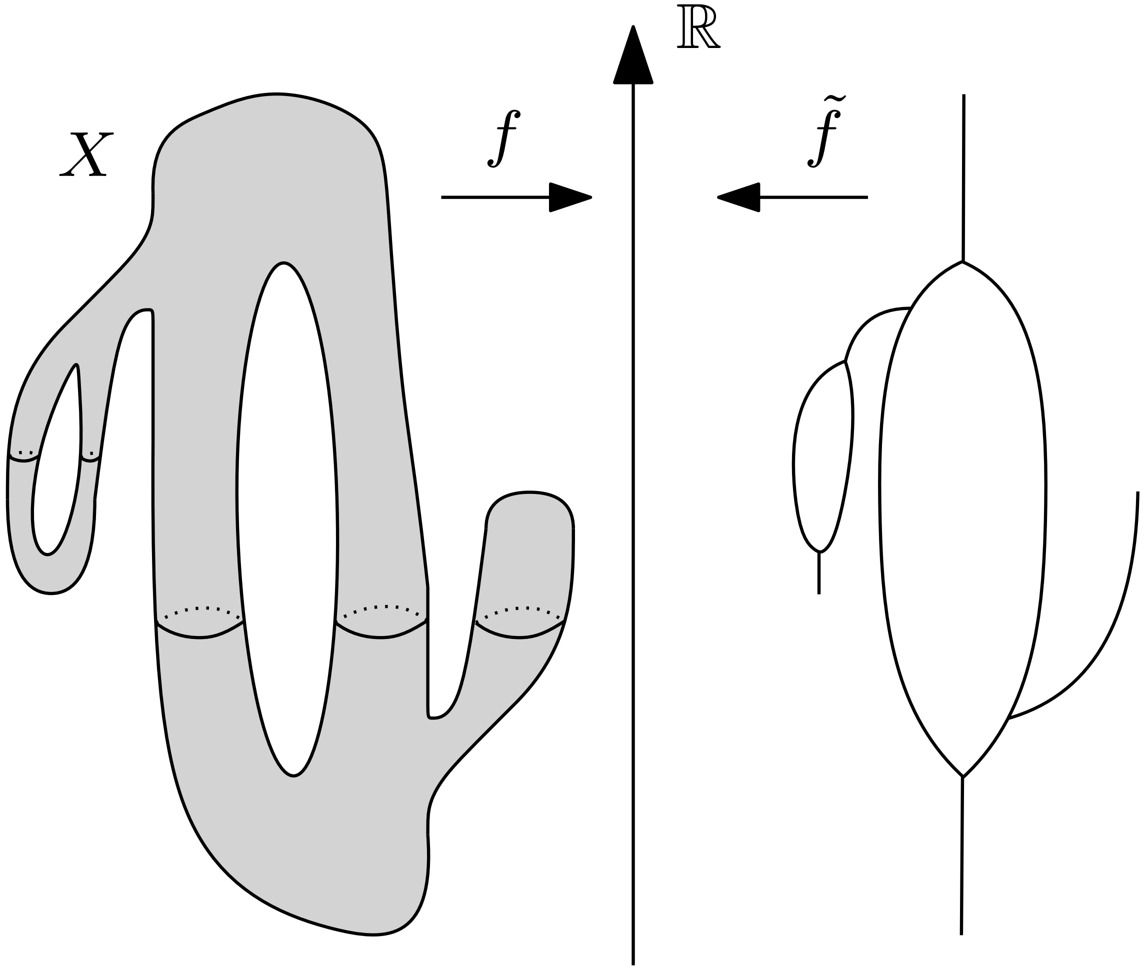
$\mathrm{R}_f(X) := X/\sim$



Reeb Graph

$x \sim y \iff [\ f(x) = f(y)$ and x, y belong to same cc of $f^{-1}(\{f(x)\})\ ]$

$\mathrm{R}_f(X) := X/\sim$



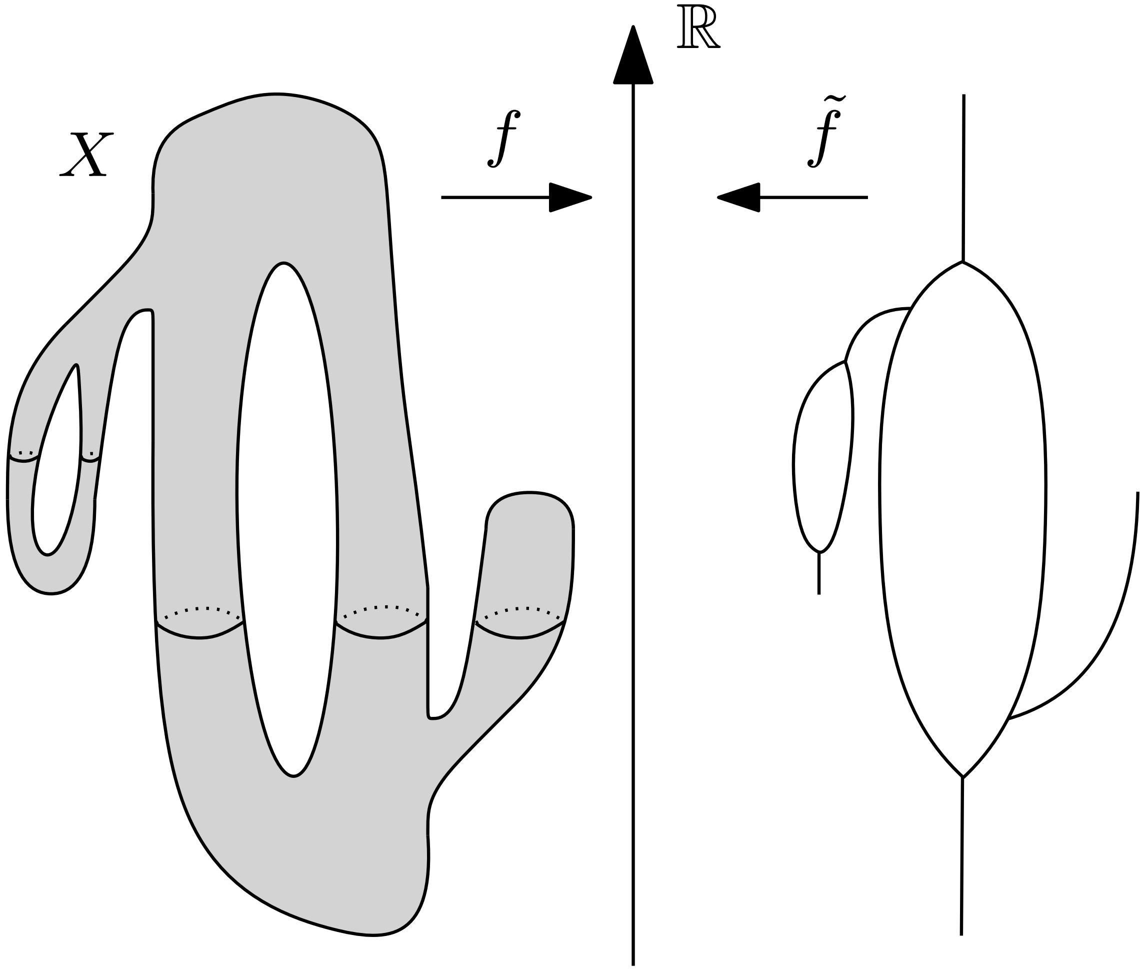
$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "\pi"'] & \mathbb{R} \\
\mathrm{R}_f(X) \arrow[ur, "\tilde{f}"'] &
\end{tikzcd}$$

Reeb Graph

$x \sim y \iff [\ f(x) = f(y)$ and x, y belong to same cc of $f^{-1}(\{f(x)\})\ ]$

$\mathrm{R}_f(X) := X/\sim$

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "\pi"'] & \mathbb{R} \\
\mathrm{R}_f(X) \arrow[ur, "\tilde{f}"'] &
\end{tikzcd}$$



Prop: $\mathrm{R}_f(X)$ is a (multi-)graph when (X, f) is Morse or more generally of **Morse type**

→ build a **descriptor** for this graph

Descriptor for Reeb graph

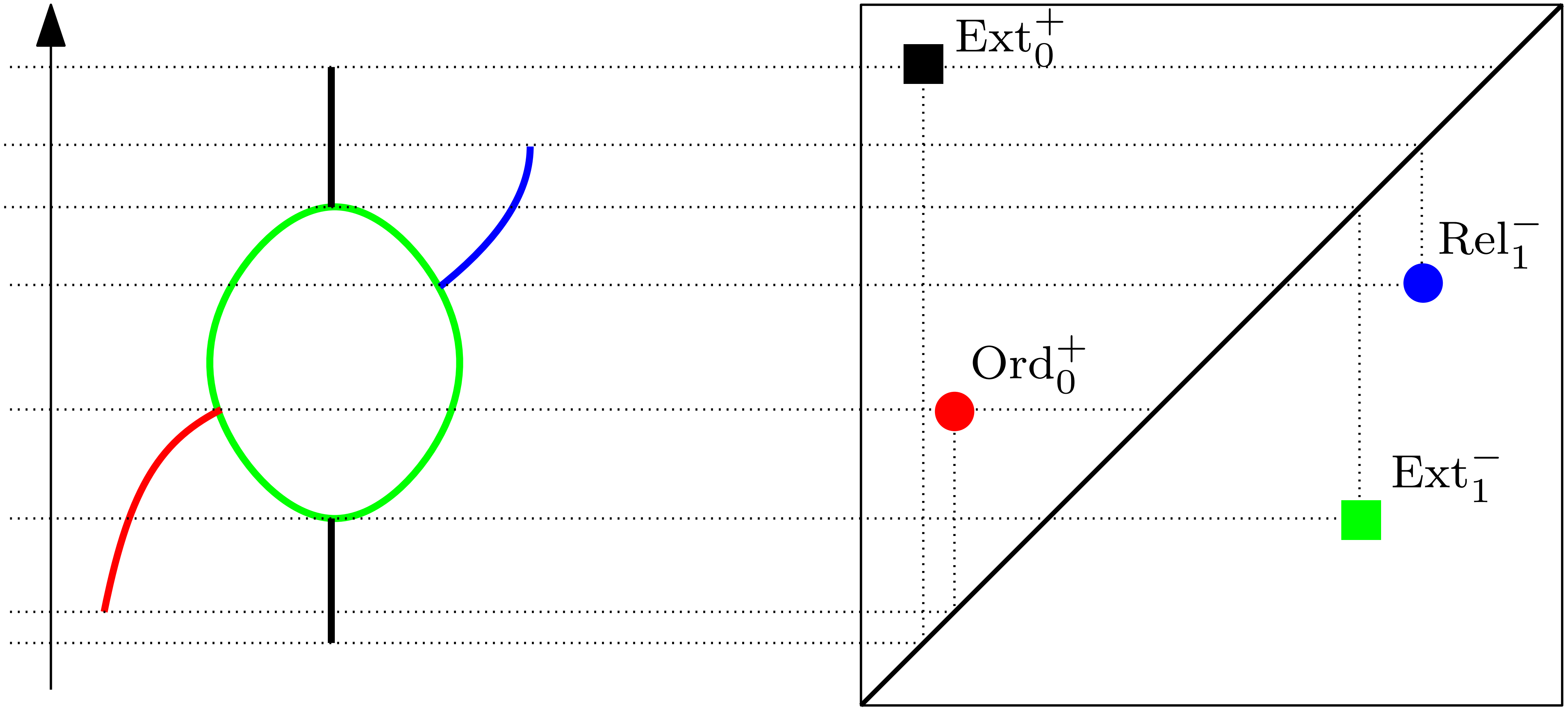
$\mathrm{Dg}\,\tilde{f}$ provides a **bag-of-features** descriptor for $\mathrm{R}_f(X)$:

$\mathrm{Ord}_0\,\tilde{f} \longleftrightarrow$ downward branches

$\mathrm{Rel}_1\,\tilde{f} \longleftrightarrow$ upward branches

$\mathrm{Ext}_0\,\tilde{f} \longleftrightarrow$ trunks (cc)

$\mathrm{Ext}_1\,\tilde{f} \longleftrightarrow$ loops



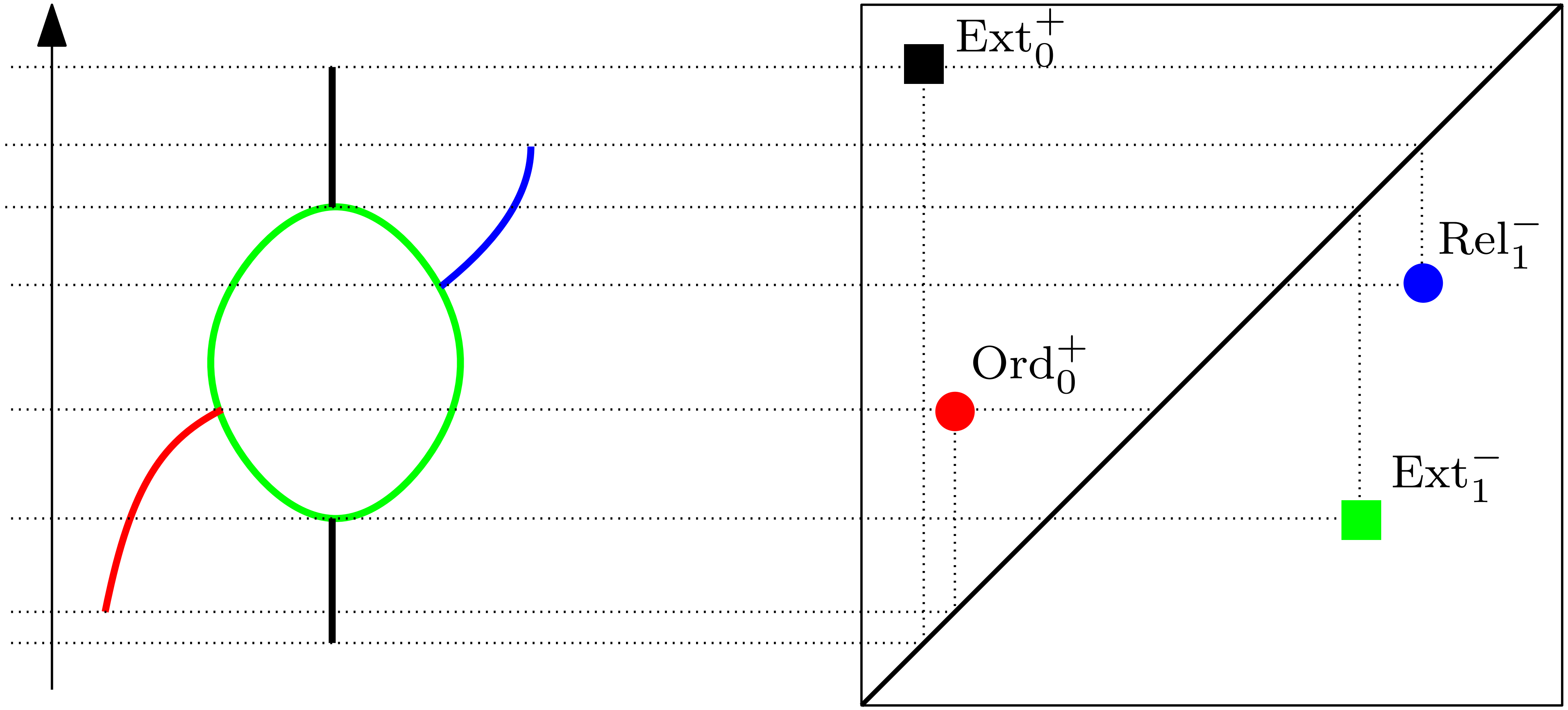
● ordinary / relative

■ extended

Descriptor for Reeb graph

$\mathrm{Dg}\,\tilde f$ provides a **bag-of-features** descriptor for $\mathrm{R}_f(X)$:

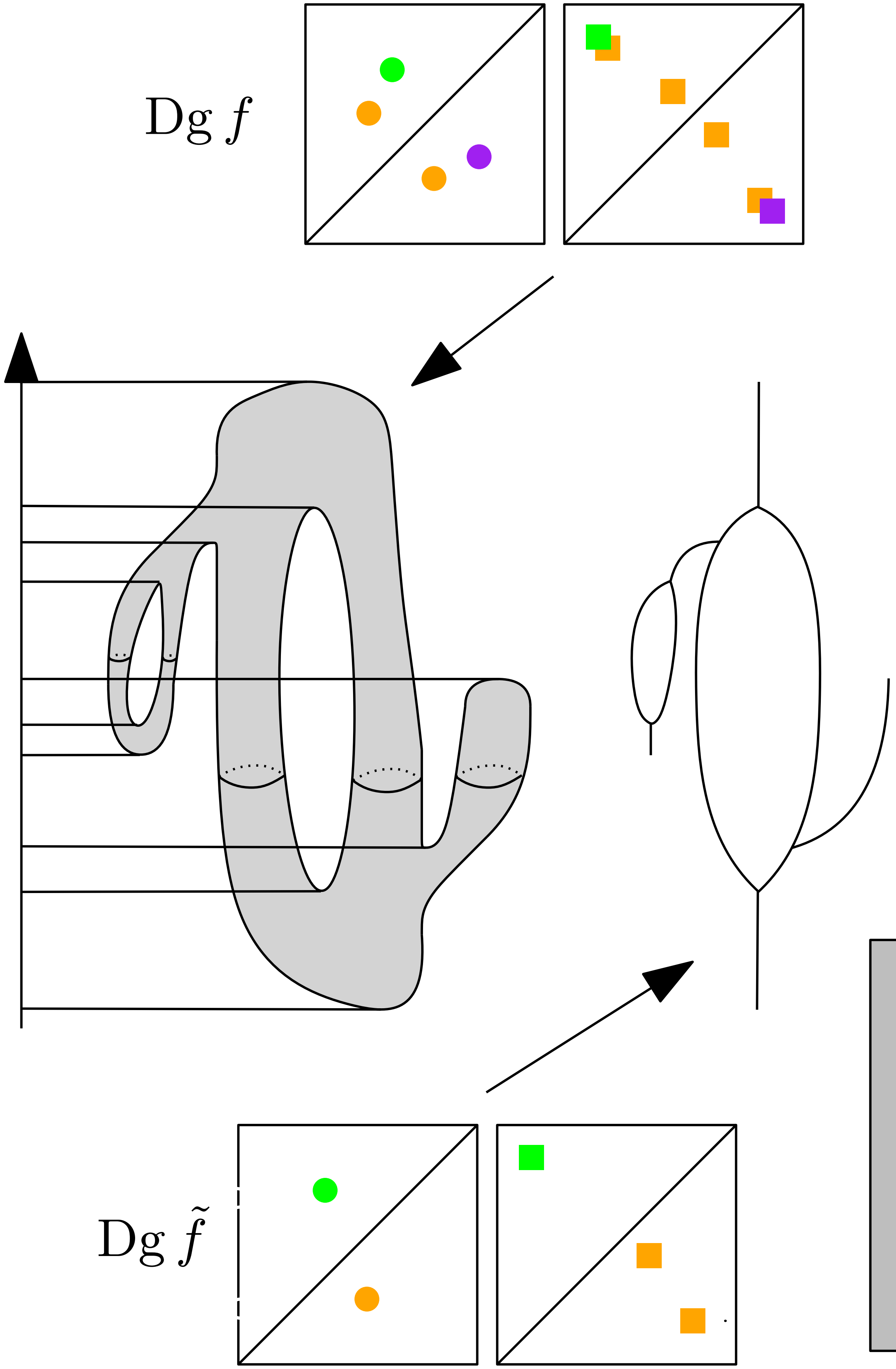
$\mathrm{Ord}_0\,\tilde f \longleftrightarrow$ downward branches	$\mathrm{Ext}_0\,\tilde f \longleftrightarrow$ trunks (cc)
$\mathrm{Rel}_1\,\tilde f \longleftrightarrow$ upward branches	$\mathrm{Ext}_1\,\tilde f \longleftrightarrow$ loops



... and distance to diagonal measures the (in-)stability of each feature w.r.t. perturbations of (X, f)

● ordinary / relative

■ extended



● ordinary / relative

■ extended

— H_0

— H_1

— H_2

Theorem: $\mathrm{Dg}\, \tilde{f} \subseteq \mathrm{Dg}\, f$. Indeed:

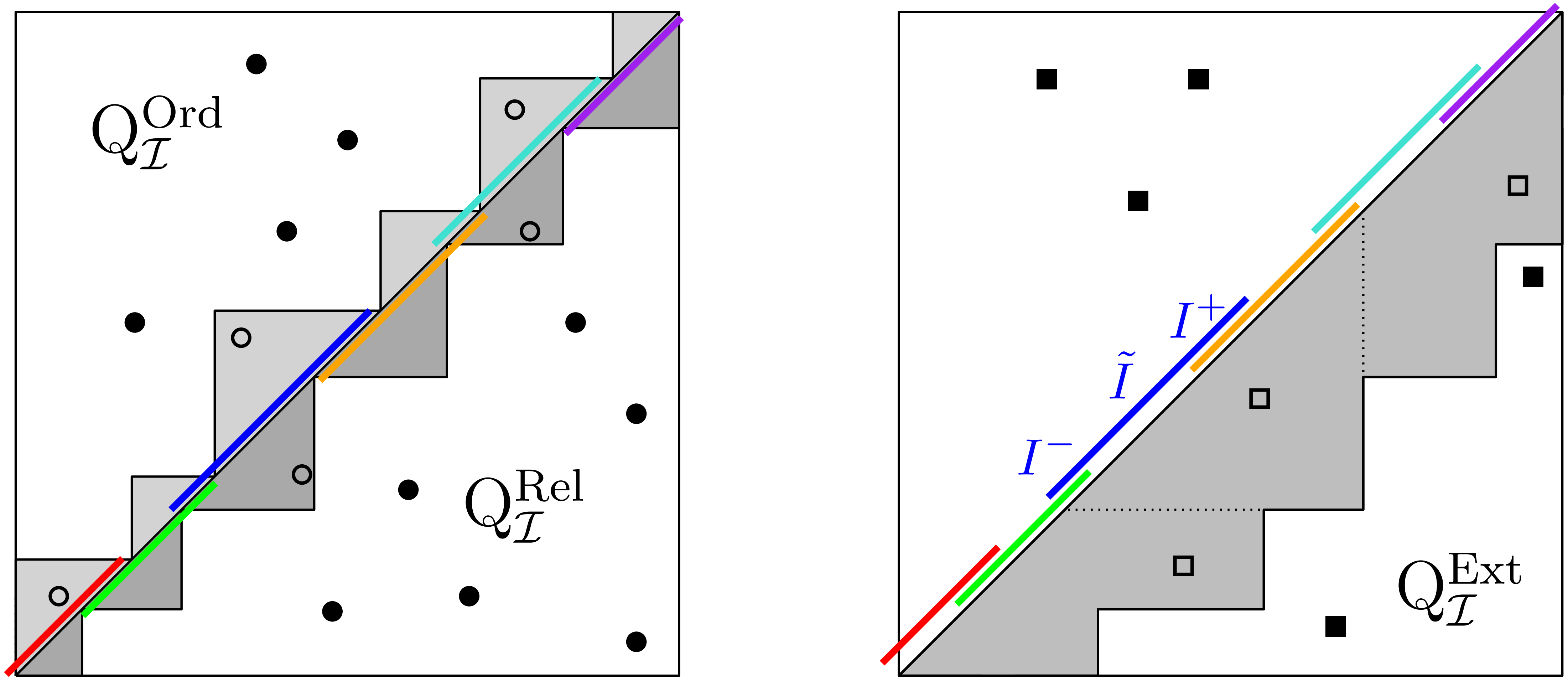
$$\mathrm{Dg}_0\, \tilde{f} = \mathrm{Dg}_0\, f$$

$$\mathrm{Dg}_1\, \tilde{f} = \mathrm{Dg}_1\, f \setminus (\mathrm{Ext}_1^+\, f \cup \mathrm{Ord}_1^+\, f)$$

$\mathrm{Dg}_i\, \tilde{f} = \emptyset$ for $i \geq 2$

Descriptor for Mapper

Definition: $\operatorname{Dg} \mathrm{M}_f(X, \mathcal{I}) := \operatorname{Ord} \tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ord}} \cup \operatorname{Rel} \tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Rel}} \cup \operatorname{Ext} \tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}$

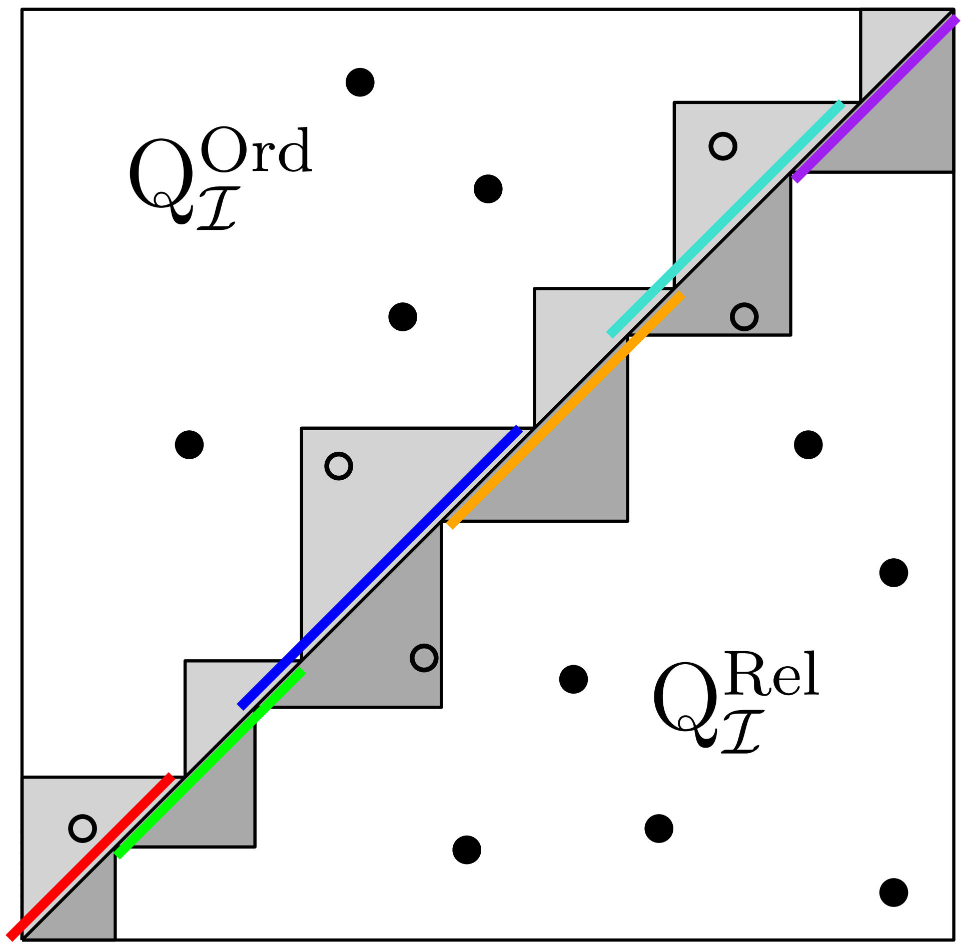


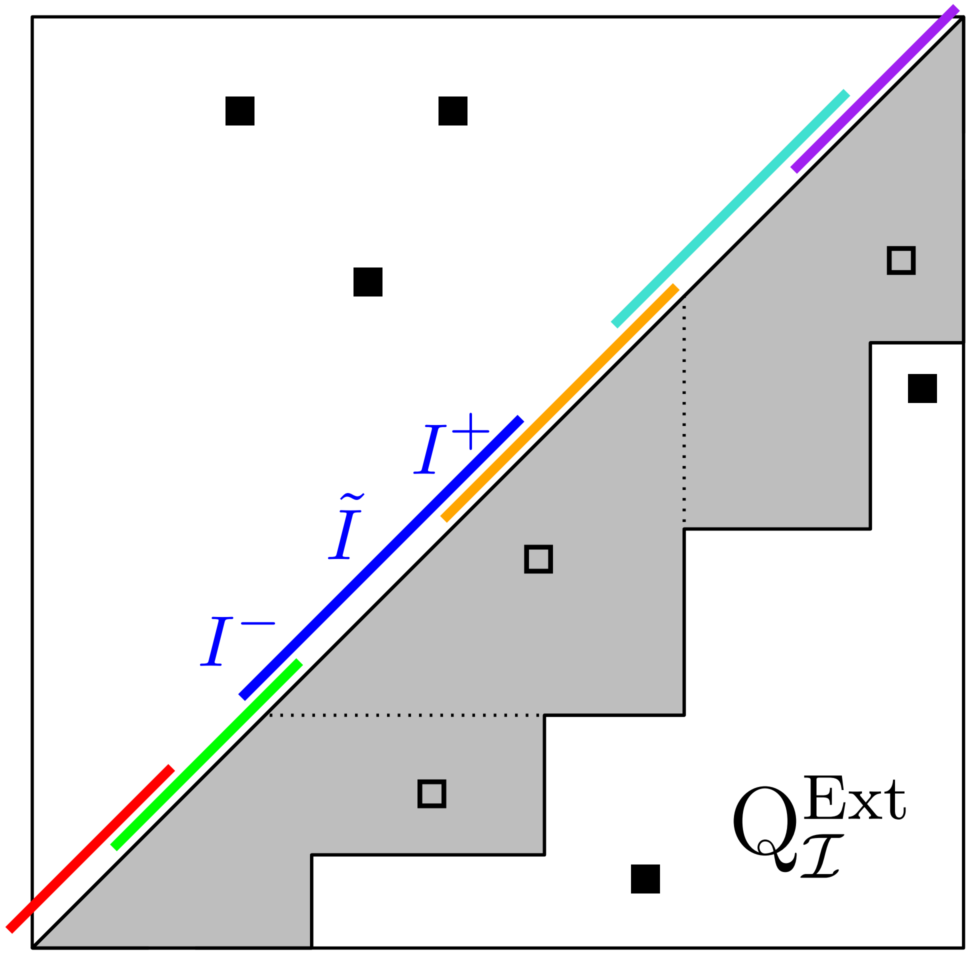
Descriptor for Mapper

Definition: $\operatorname{Dg} \mathrm{M}_f(X, \mathcal{I}) := \operatorname{Ord} \tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ord}} \cup \operatorname{Rel} \tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Rel}} \cup \operatorname{Ext} \tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}$

Thm: $\operatorname{Dg}(\mathrm{M}_f(X, \mathcal{I}))$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X, \mathcal{I})$:

$\mathrm{Ord}_0 \longleftrightarrow$ downward branches	$\mathrm{Ext}_0 \longleftrightarrow$ trunks (cc)
$\mathrm{Rel}_1 \longleftrightarrow$ upward branches	$\mathrm{Ext}_1 \longleftrightarrow$ loops



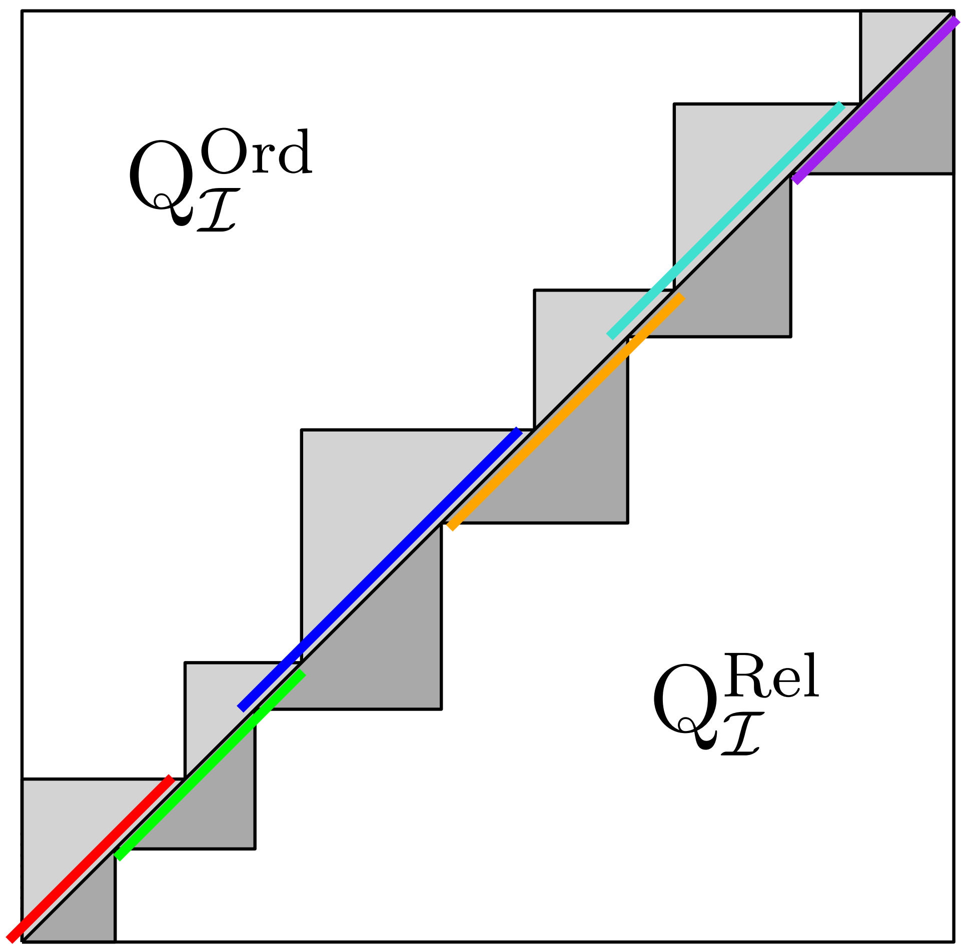


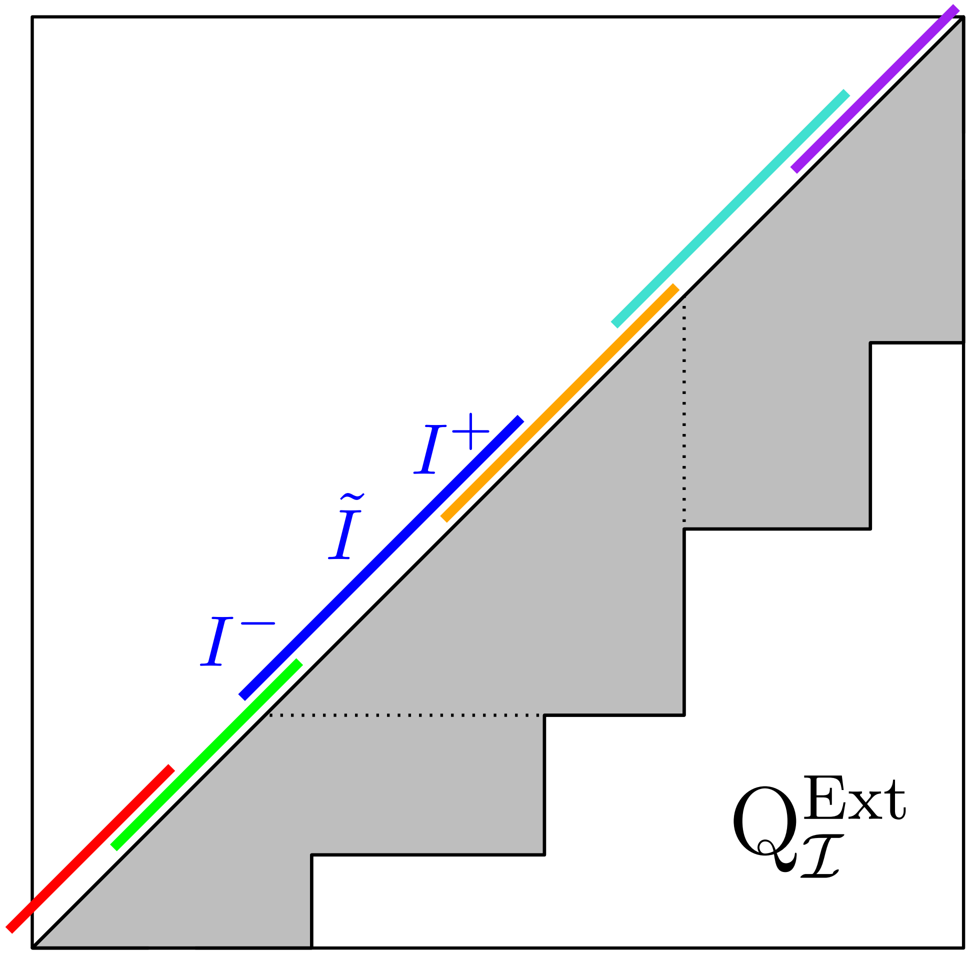
Descriptor for Mapper

Let $\mathcal{I}$ minimal cover of $\mathrm{Im} f \subseteq \mathbb{R}$. For $I \in \mathcal{I}$, let $I = I^- \sqcup \tilde{I} \sqcup I^+$

$$Q_{\mathcal{I}}^{\mathrm{Ord}} = \bigcup_{I \in \mathcal{I}} Q^+_{\tilde{I} \cup I^+} \qquad Q_{\mathcal{I}}^{\mathrm{Ext}} = \bigcup_{\substack{I,J \in \mathcal{I} \\ I \cap J \neq \emptyset}} Q^-_{I \cup J}$$

$$Q_{\mathcal{I}}^{\mathrm{Rel}} = \bigcup_{I \in \mathcal{I}} Q^-_{I^- \cup \tilde{I}}$$

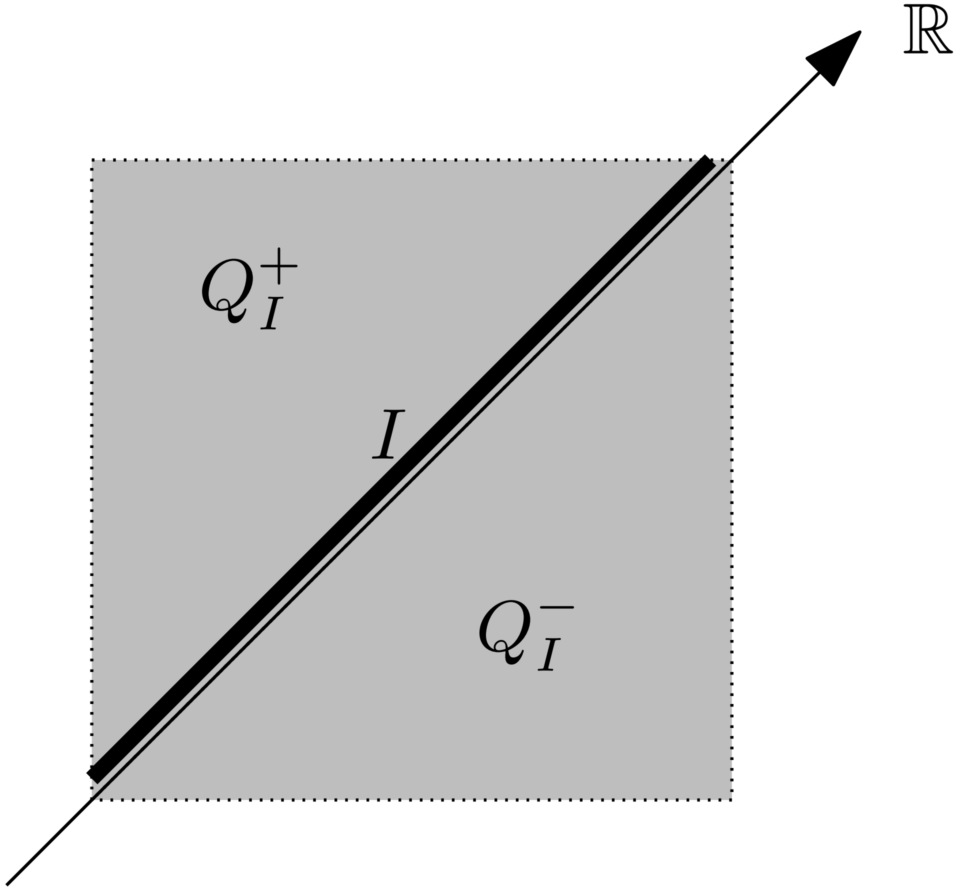


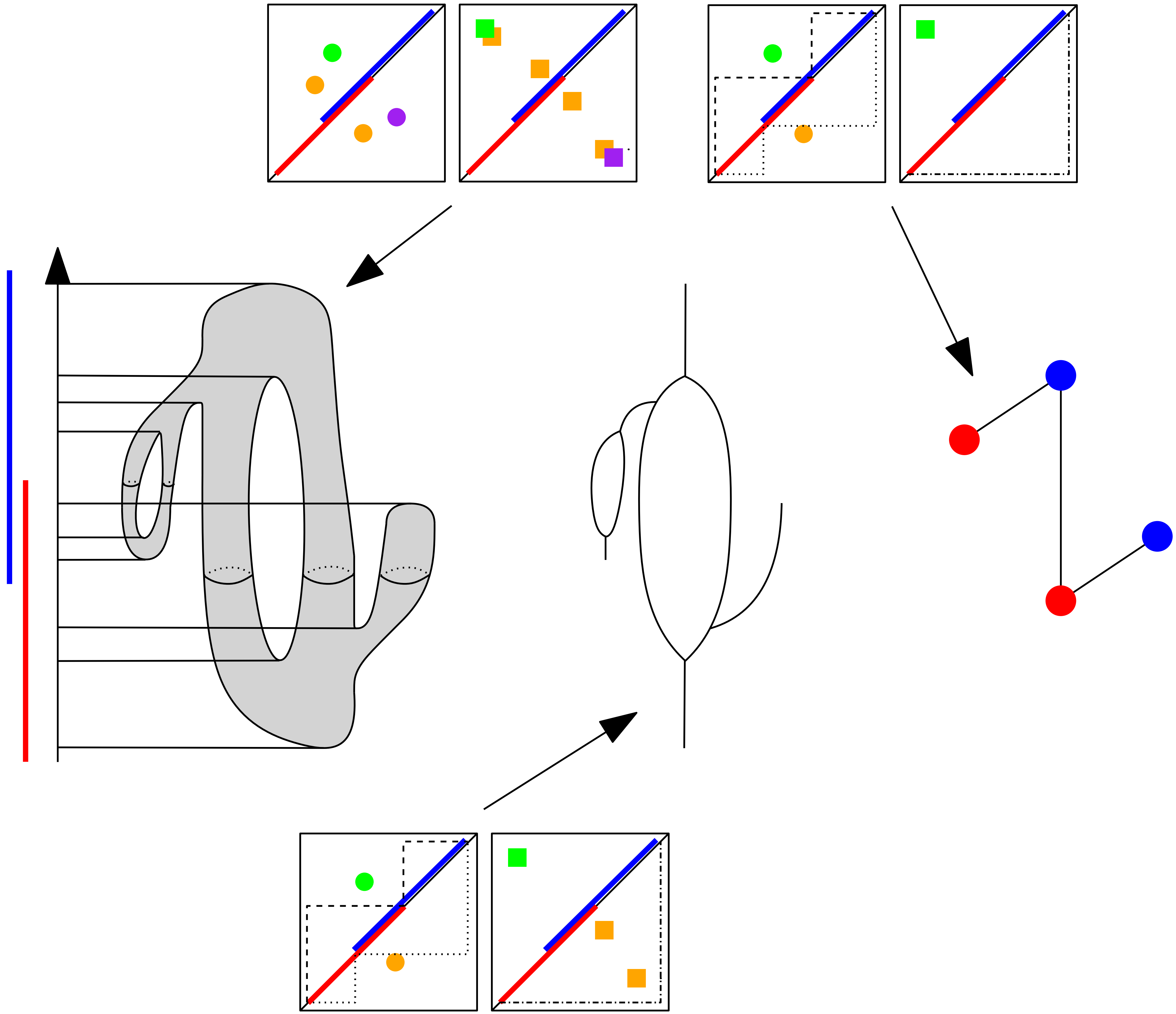


Descriptor for Mapper

Let $I \subseteq \mathbb{R}$ interval

$$Q_I^+ = \{(x, y) \in \mathbb{R}^2 \mid x \leq y \in I\}$$
$$Q_I^- = \{(x, y) \in \mathbb{R}^2 \mid y < x \in I\}$$





Structure of Mapper

Definition: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) := \mathrm{Ord}\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ord}} \cup \mathrm{Rel}\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Rel}} \cup \mathrm{Ext}\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}$

Thm: $\mathrm{Dg}\,(\mathrm{M}_f(X,\mathcal{I}))$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X,\mathcal{I})$:

$\mathrm{Ord}_0 \longleftrightarrow$ downward branches

$\mathrm{Rel}_1 \longleftrightarrow$ upward branches

$\mathrm{Ext}_0 \longleftrightarrow$ trunks (cc)

$\mathrm{Ext}_1 \longleftrightarrow$ loops

Corollary: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) = \mathrm{Dg}\,\tilde{f}$ whenever the resolution r of $\mathcal{I}$ is smaller than the smallest distance from $\mathrm{Dg}\,\tilde{f} \setminus \Delta$ to the diagonal Δ.

Stability of Mapper

Definition: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) := \mathrm{Ord}\,\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ord}} \cup \mathrm{Rel}\,\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Rel}} \cup \mathrm{Ext}\,\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}$

Thm: $\mathrm{Dg}\,(\mathrm{M}_f(X,\mathcal{I}))$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X,\mathcal{I})$:

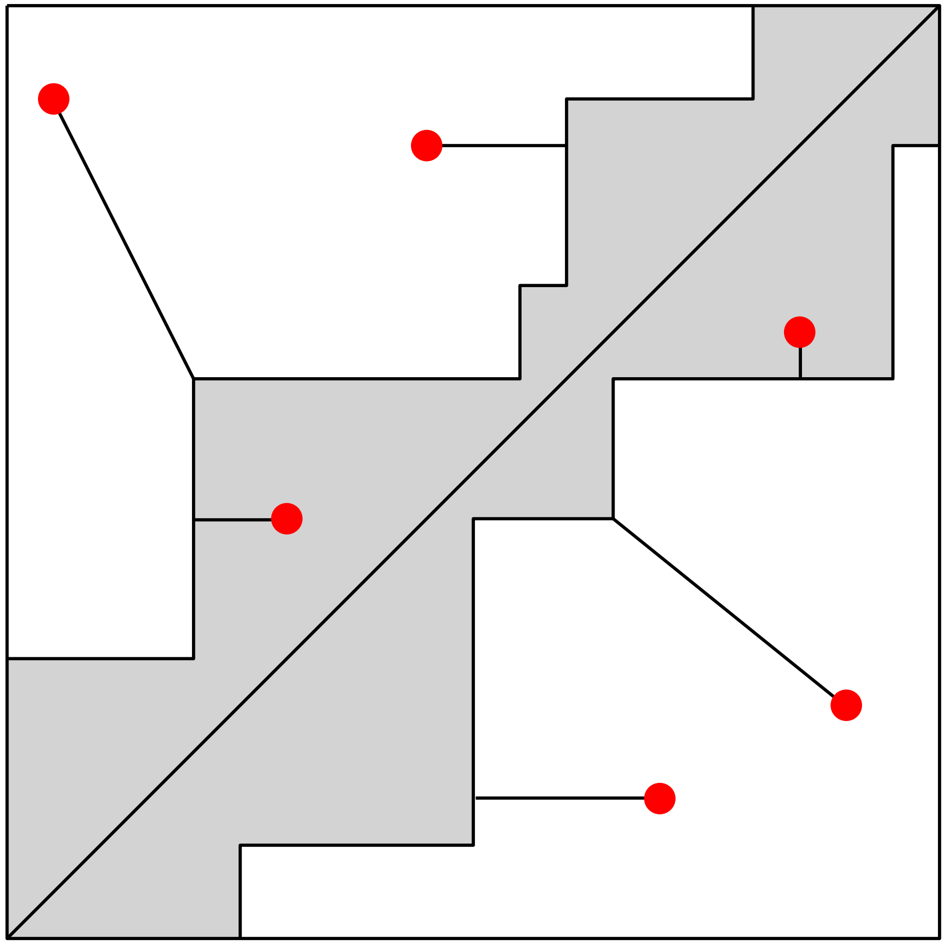
$\mathrm{Ord}_0 \longleftrightarrow$ downward branches

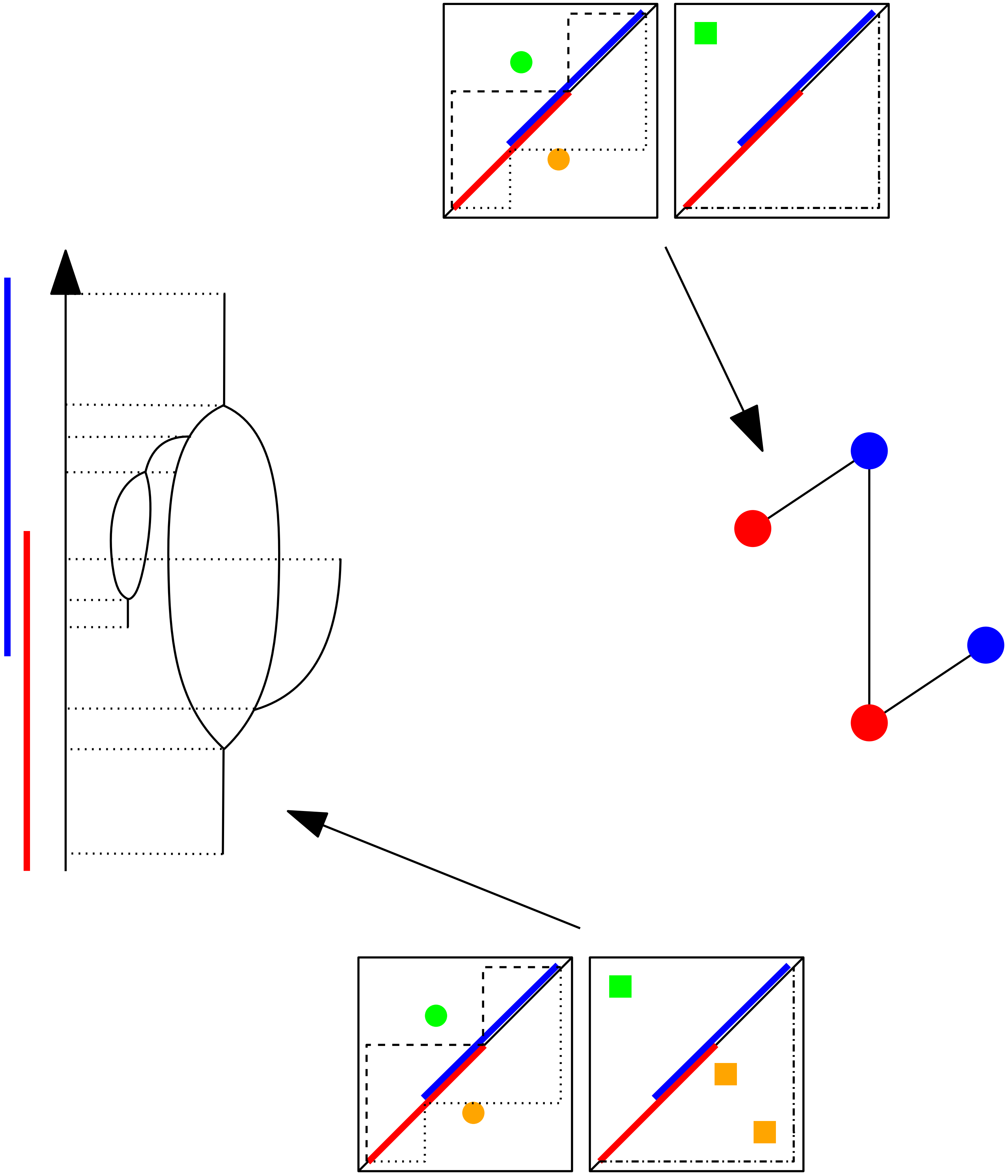
$\mathrm{Rel}_1 \longleftrightarrow$ upward branches

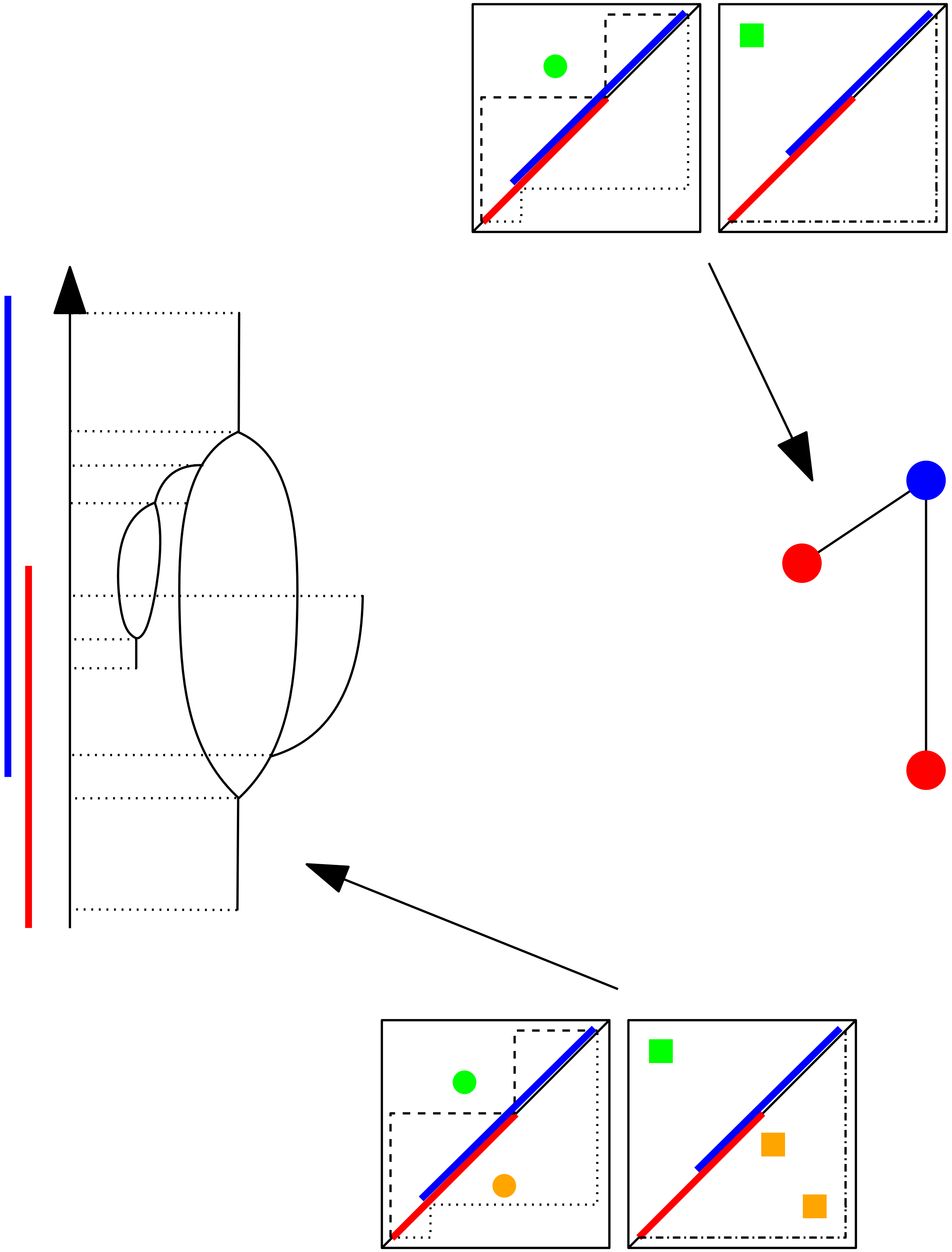
$\mathrm{Ext}_0 \longleftrightarrow$ trunks (cc)

$\mathrm{Ext}_1 \longleftrightarrow$ loops

... and distance to staircase boundary measures (in-)stability of each feature w.r.t. perturbations of $(X, f, \mathcal{I})$

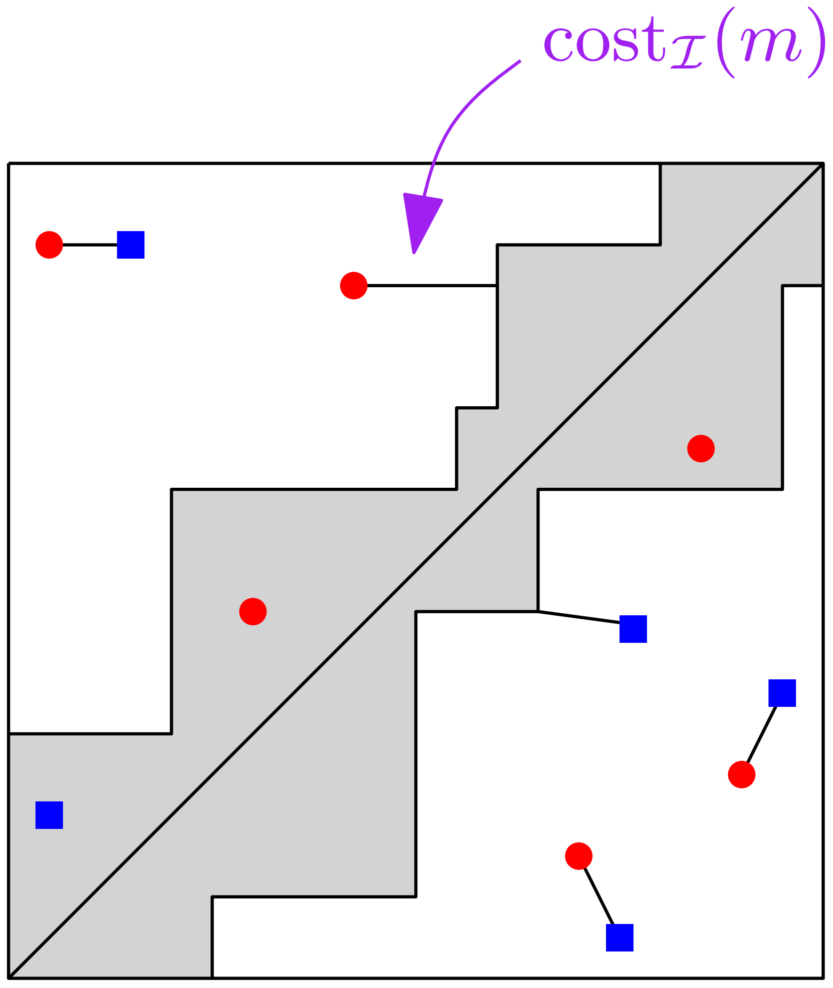






Stability of Mapper

Definition: $\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I})) := \inf_m \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$



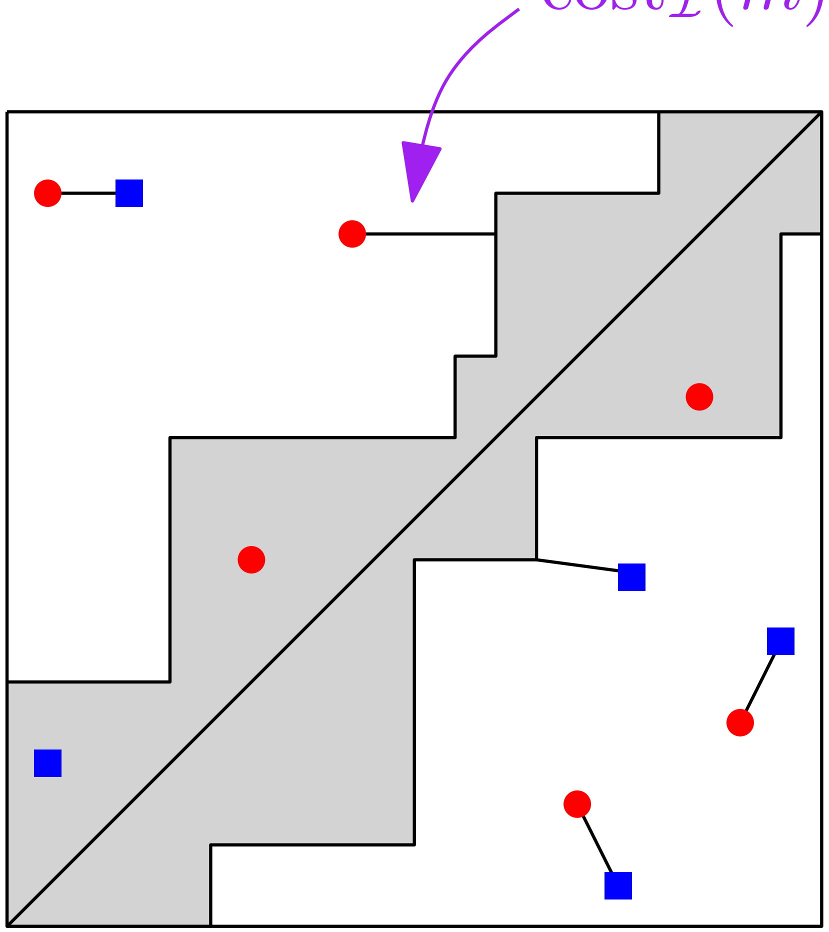
$m :\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})$

Stability of Mapper

Definition: $\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I})) := \inf_m \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$

Thm: For any functions $f, f' : X \to \mathbb{R}$ of Morse type,

$$\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})) \leq \|f - f'\|_\infty$$



$\mathrm{cost}_{\mathcal{I}}(m)$

$m :\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})$

Stability of Mapper

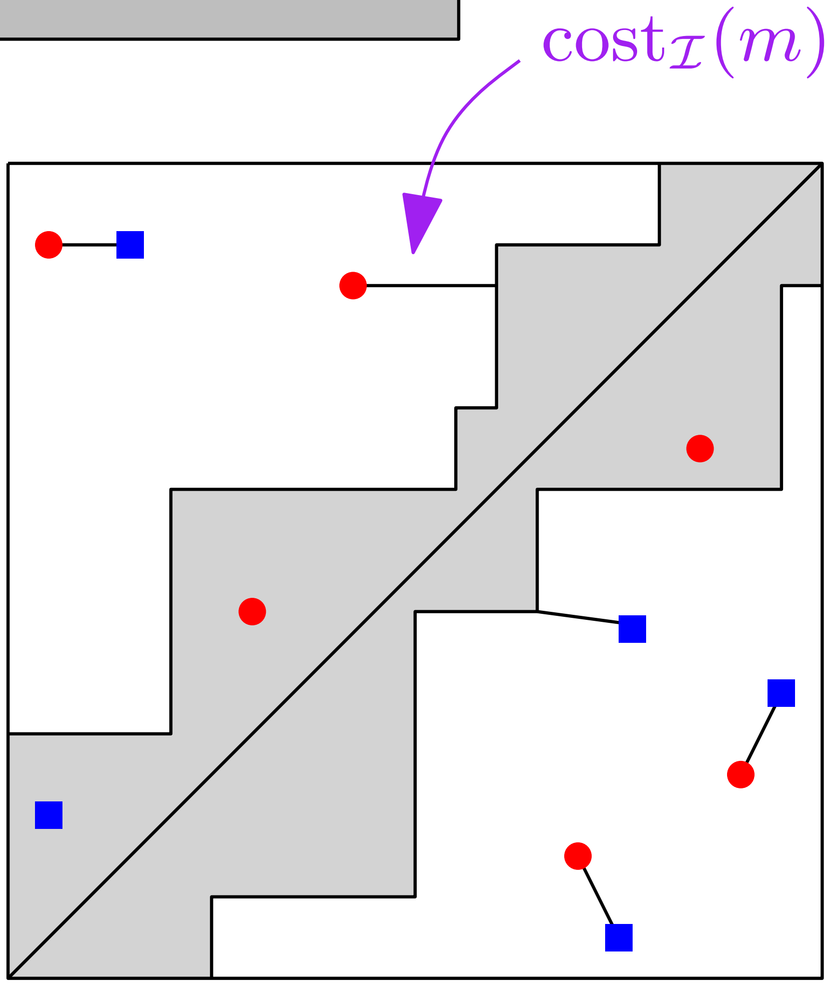
Definition: $\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I})) := \inf_m \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$

Thm: For any functions $f, f' : X \to \mathbb{R}$ of Morse type,

$$\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})) \leq \|f - f'\|_\infty$$

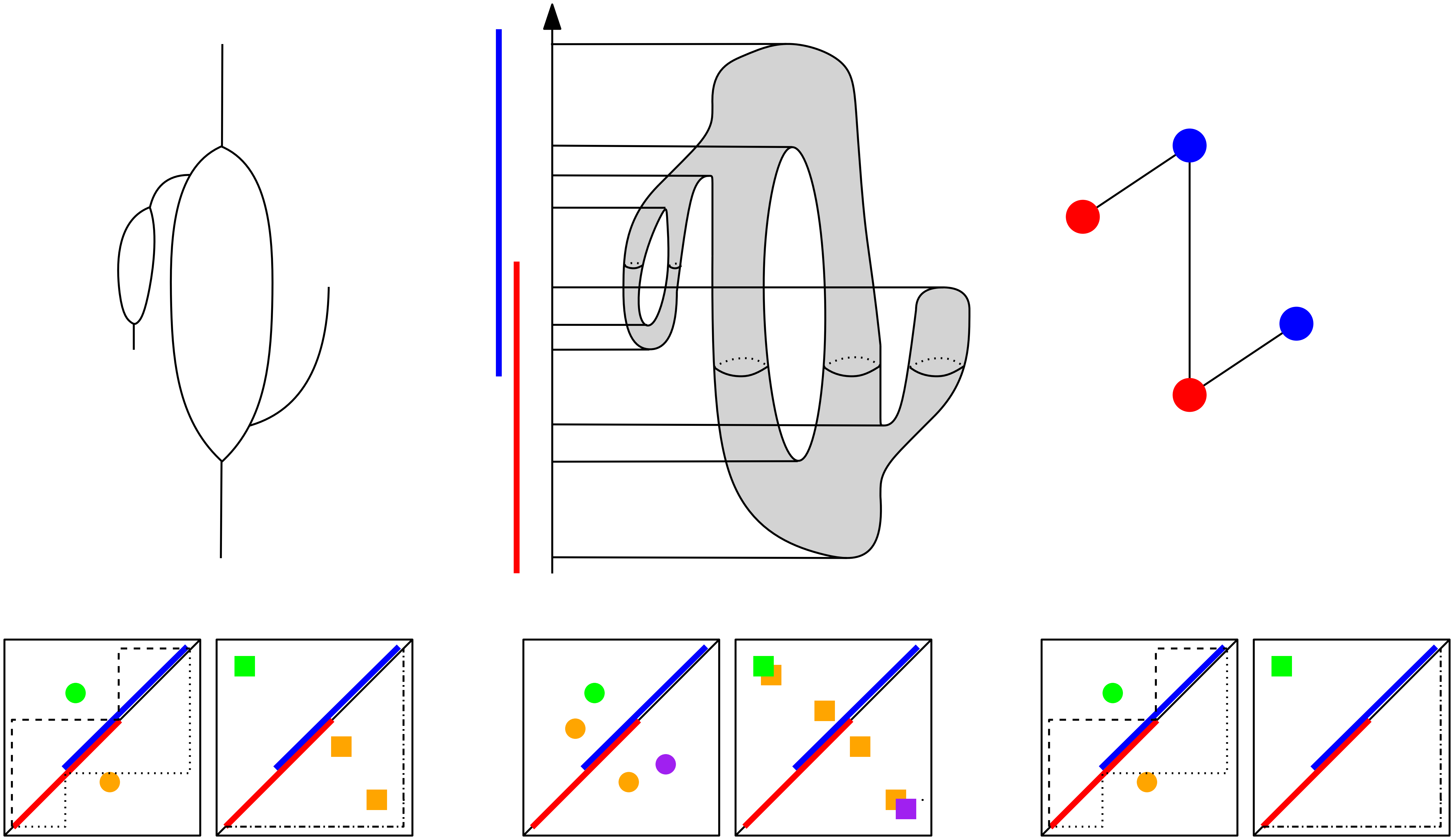
Extensions to:

- perturbations of X
- perturbations of $\mathcal{I}$

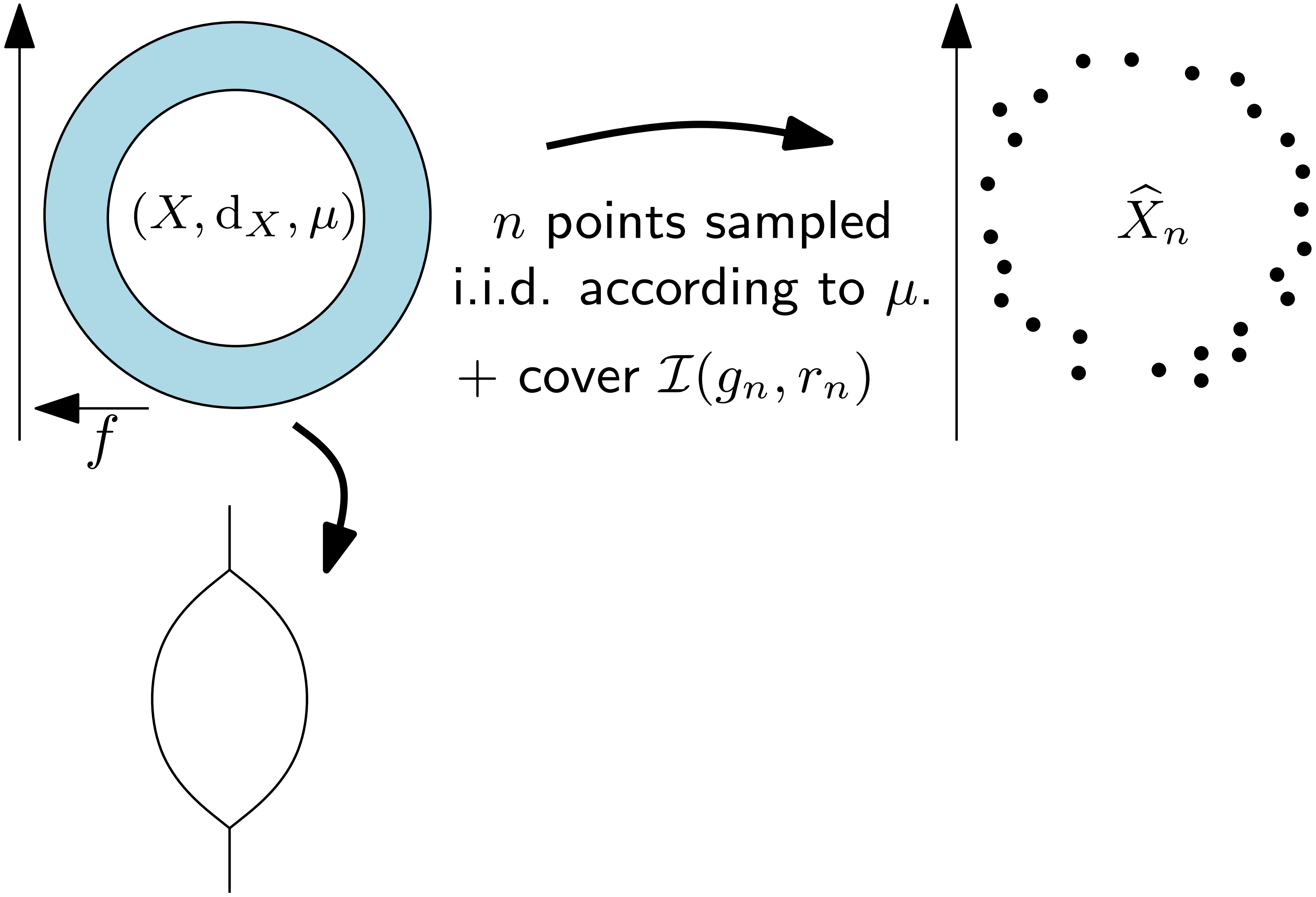


$m :$ $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})$

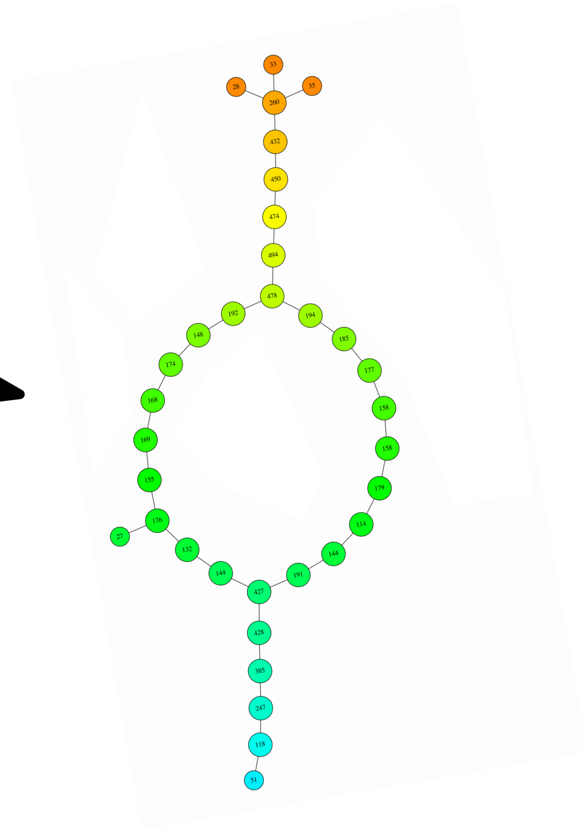
Summary



Back to statistics



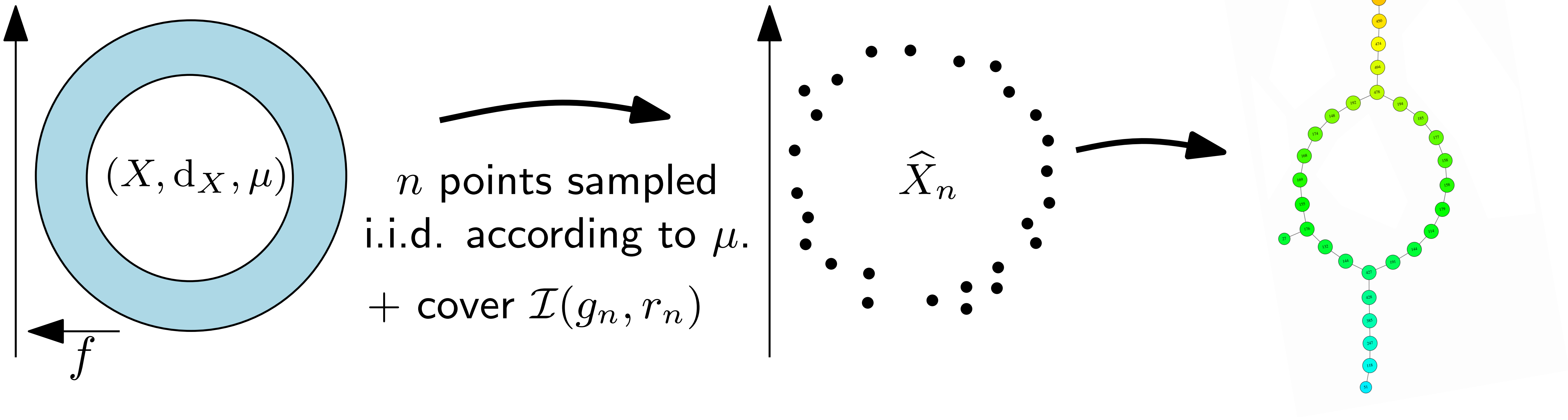
n points sampled
i.i.d. according to μ.
$+$ cover $\mathcal{I}(g_n, r_n)$



Questions:

- Statistical properties of the estimator $\mathrm{M}_f(\widehat{X}_n, \mathcal{I}(g_n, r_n))$?
- Convergence to the ground truth $\mathrm{R}_f(X)$ in d_B? Deviation bounds?

Back to statistics



n points sampled
i.i.d. according to μ.
+ cover $\mathcal{I}(g_n, r_n)$

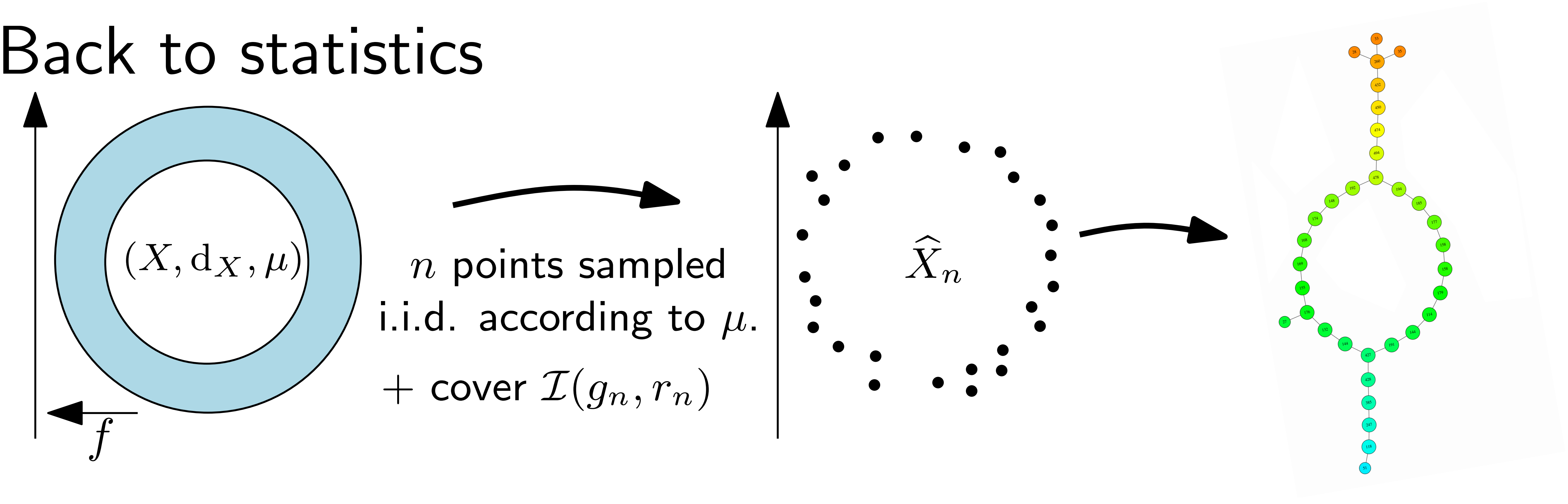
Theorem [Carrière, Michel, O. 2016]:

If μ is (a,b)-standard and $\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b}$, $g_n \in \left(\frac{1}{3},\, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, then $\forall \varepsilon > 0$:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathrm{M}_f(\widehat{X}_n, \mathcal{I}(g_n, r_n)),\ \mathrm{Dg}\,\mathrm{R}_f(X)\right)\right] \leq C\ \left(\frac{\log n}{n}\right)^{1/b},$$

where C depends only on a, b. Moreover, the estimator $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ is **minimax optimal** (up to a $\log n$ factor) on the space $\mathcal{P}$ of (a,b)-standard probability measures on X.

Back to statistics



n points sampled
i.i.d. according to μ.
+ cover $\mathcal{I}(g_n, r_n)$

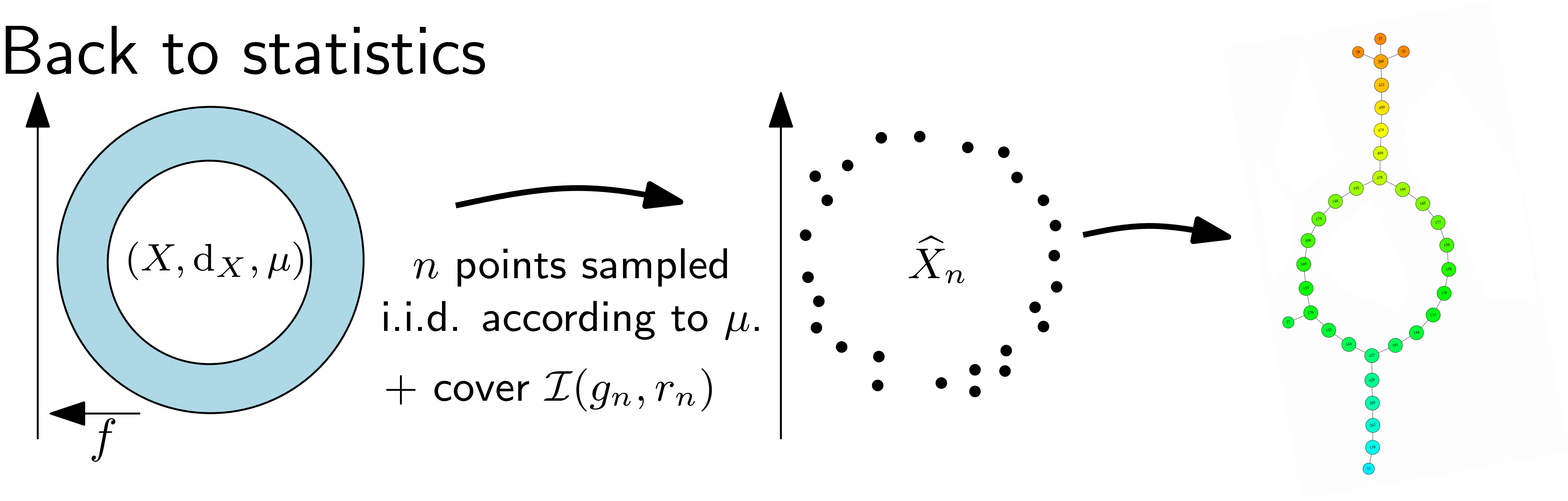
Theorem [Carrière, Michel, O. 2016]:

If μ is (a,b)-standard and $\delta_n = 4\left(\frac{2\log n}{\boxed{a}n}\right)^{1/\boxed{b}}$, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, then $\forall \varepsilon > 0$:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathrm{M}_f(\widehat{X}_n, \mathcal{I}(g_n, r_n)),\ \mathrm{Dg}\,\mathrm{R}_f(X)\right)\right] \leq C\left(\frac{\log n}{n}\right)^{1/b},$$

where C depends only on a, b. Moreover, the estimator $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ is **minimax optimal** (up to a $\log n$ factor) on the space $\mathcal{P}$ of (a,b)-standard probability measures on X.

Back to statistics



n points sampled
i.i.d. according to μ.
+ cover $\mathcal{I}(g_n, r_n)$

Theorem [Carrière, Michel, O. 2016]:

If μ is (a, b)-standard and $\delta_n = 4\left(\frac{2\log n}{\boxed{a}n}\right)^{1/\boxed{b}}$, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, then $\forall \varepsilon > 0$:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathrm{M}_f(\widehat{X}_n, \mathcal{I}(g_n, r_n)),\ \mathrm{Dg}\,\mathrm{R}_f(X)\right)\right] \leq C\left(\frac{\log n}{n}\right)^{1/b},$$

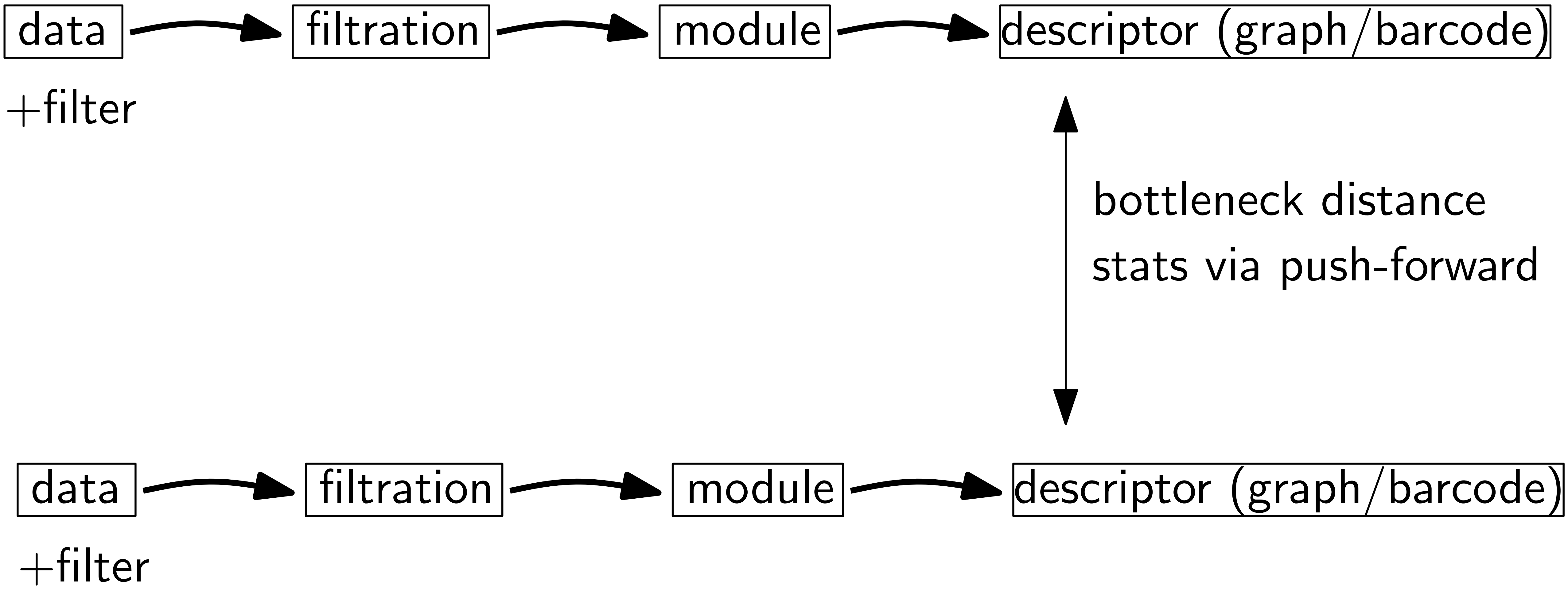
where C depends only on a, b. Moreover, the estimator $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ is **minimax optimal** (up to a $\log n$ factor) on the space $\mathcal{P}$ of (a, b)-standard probability measures on X.

→ subsampling to tune δ_n: take $(s_n)_{n\in\mathbb{N}} \to +\infty$ s.t. $s_n = o\left(\frac{n}{\log n}\right)$

$$\delta_n = \mathrm{average}_{\{\text{possible subsamples } Y \text{ of } \hat{X}_n \text{ of size } s_n\}}\, \mathrm{d_H}(Y, \hat{X}_n)$$

Wrap-up

TDA pipeline:



data → filtration → module → descriptor (graph/barcode)

+filter

bottleneck distance

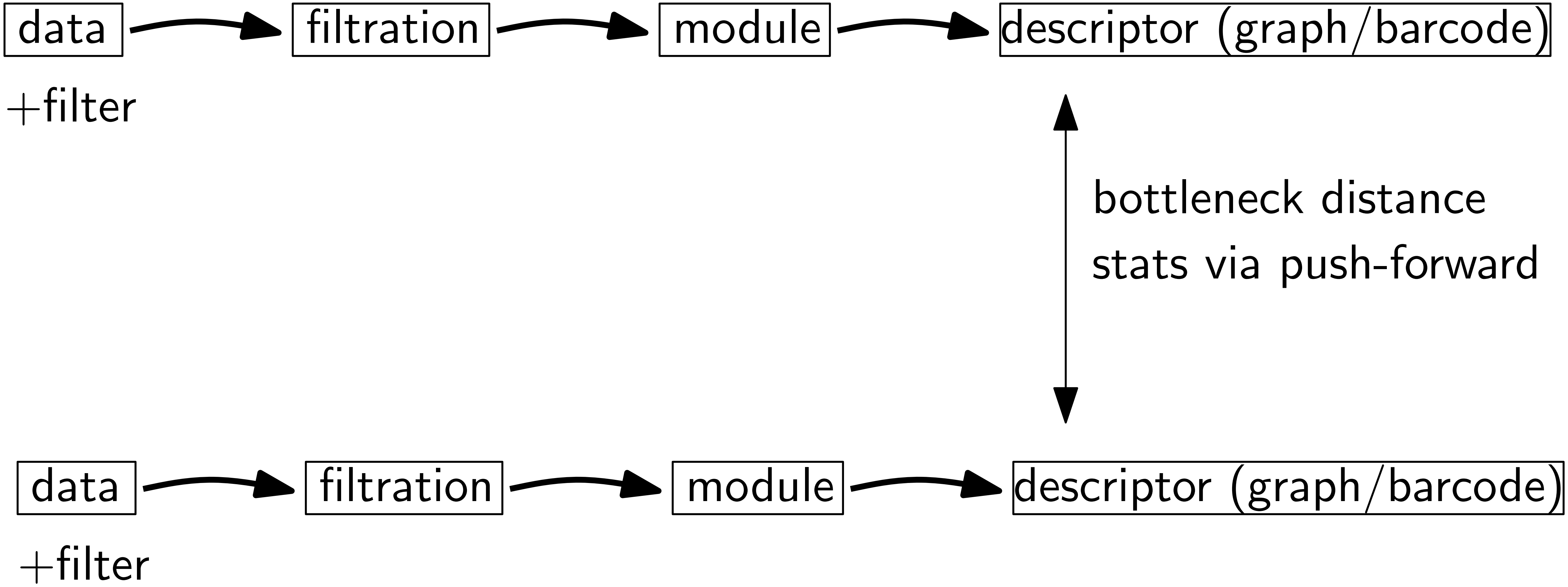
stats via push-forward

data → filtration → module → descriptor (graph/barcode)

+filter

Wrap-up

TDA pipeline:



data → filtration → module → descriptor (graph/barcode)

+filter

bottleneck distance

stats via push-forward

data → filtration → module → descriptor (graph/barcode)

+filter

→ filter selection?

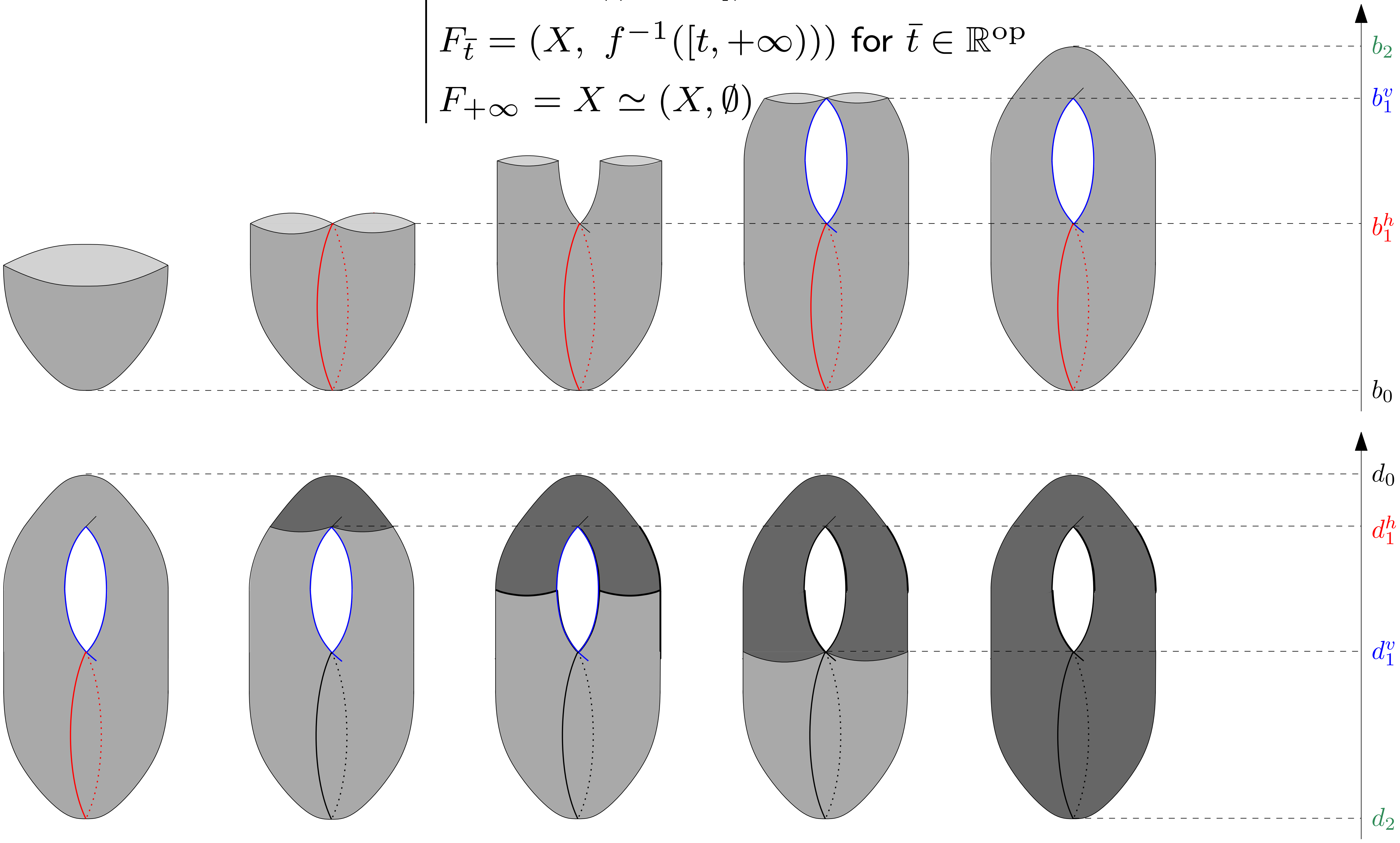
→ barcode/graph interpretation?

→ vector-valued functions?

Extended persistence

$$T = \mathbb{R} \cup \{+\infty\} \cup \mathbb{R}^{\mathrm{op}}$$

$$\left|\begin{array}{l} F_t = f^{-1}((-\infty, t]) \text{ for } t \in \mathbb{R} \\ F_{\bar{t}} = (X,\ f^{-1}([t, +\infty))) \text{ for } \bar{t} \in \mathbb{R}^{\mathrm{op}} \\ F_{+\infty} = X \simeq (X, \emptyset) \end{array}\right.$$



Extended persistence

$$T = \mathbb{R} \cup \{+\infty\} \cup \mathbb{R}^{\mathrm{op}}$$

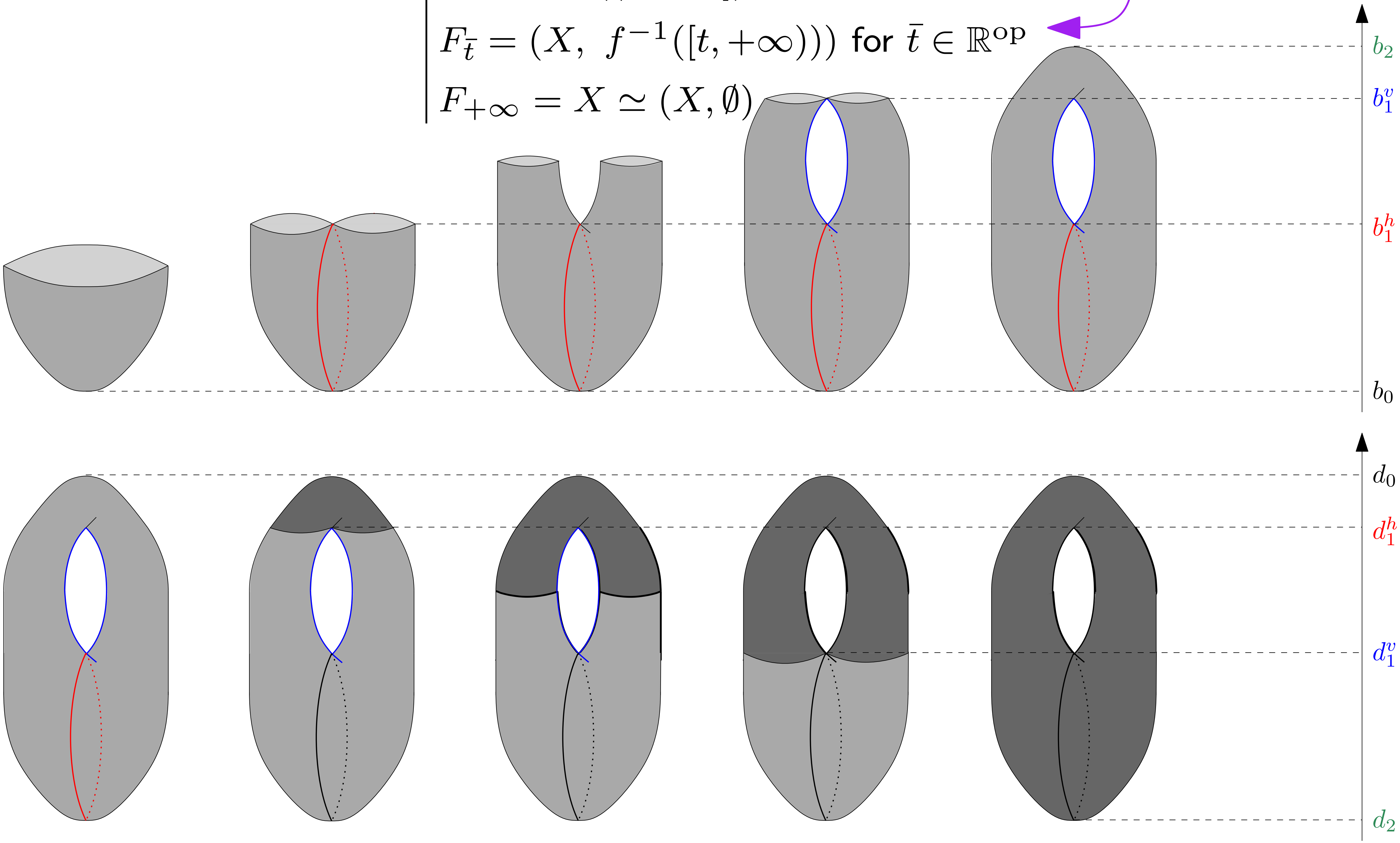
$F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$

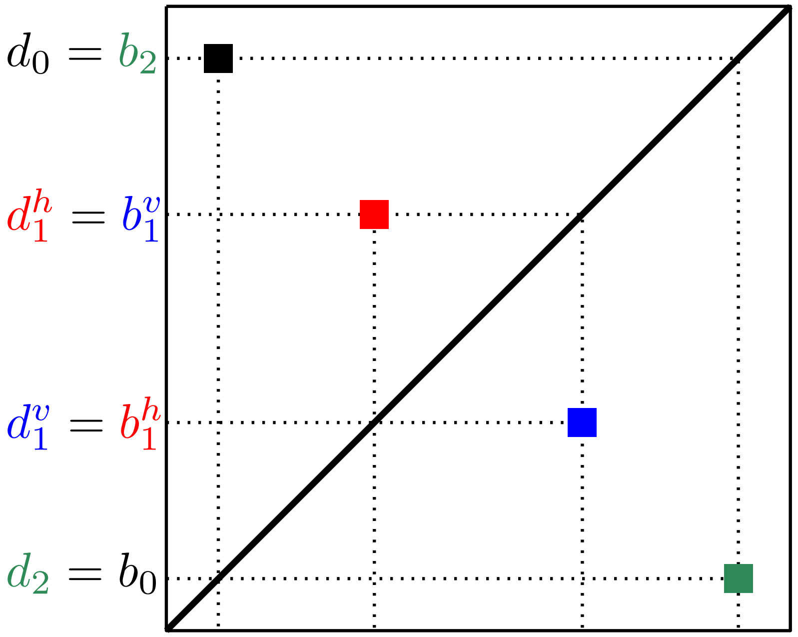
$F_{\bar{t}} = (X,\ f^{-1}([t, +\infty)))$ for $\bar{t} \in \mathbb{R}^{\mathrm{op}}$

$F_{+\infty} = X \simeq (X, \emptyset)$

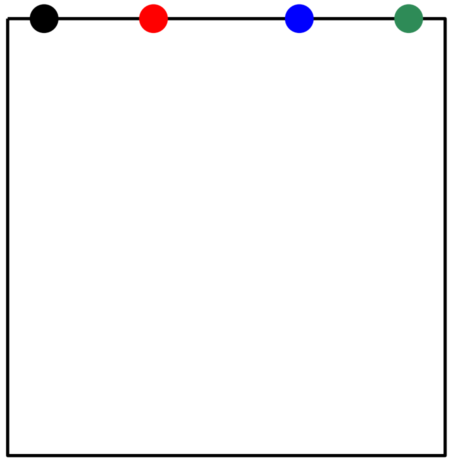
ordinary part

extended part



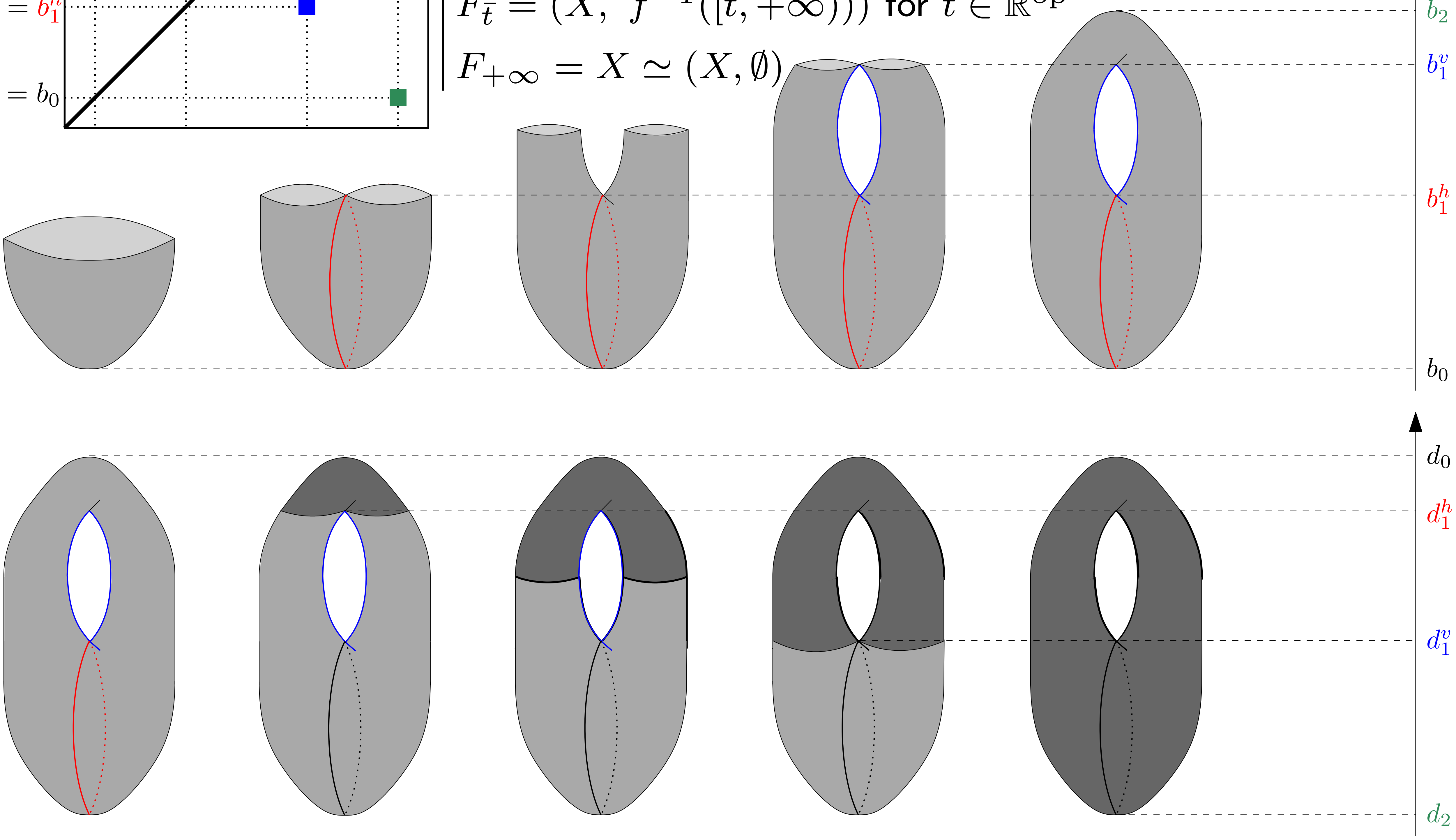


Extended persistence

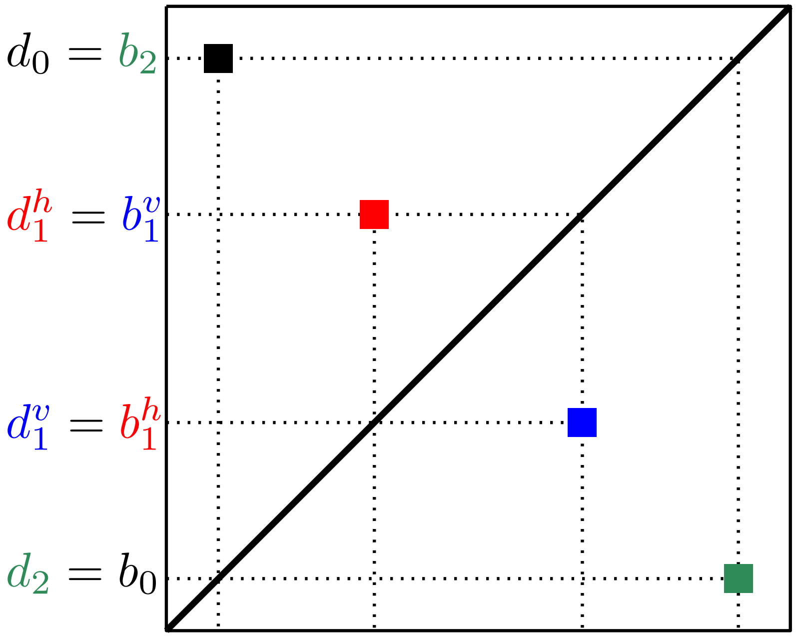


$$T = \mathbb{R} \cup \{+\infty\} \cup \mathbb{R}^{\mathrm{op}}$$

$$\left|\begin{array}{l} F_t = f^{-1}((-\infty, t]) \text{ for } t \in \mathbb{R} \\ F_{\bar{t}} = (X,\ f^{-1}([t, +\infty))) \text{ for } \bar{t} \in \mathbb{R}^{\mathrm{op}} \\ F_{+\infty} = X \simeq (X, \emptyset) \end{array}\right.$$



Extended persistence

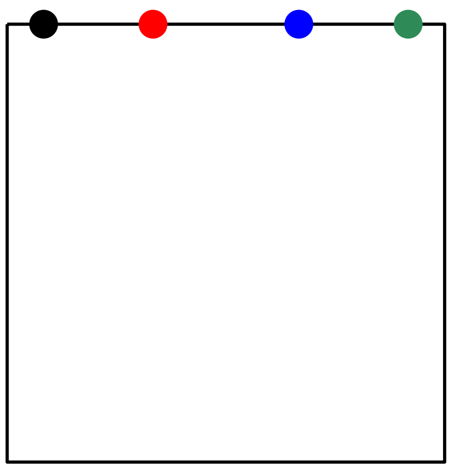


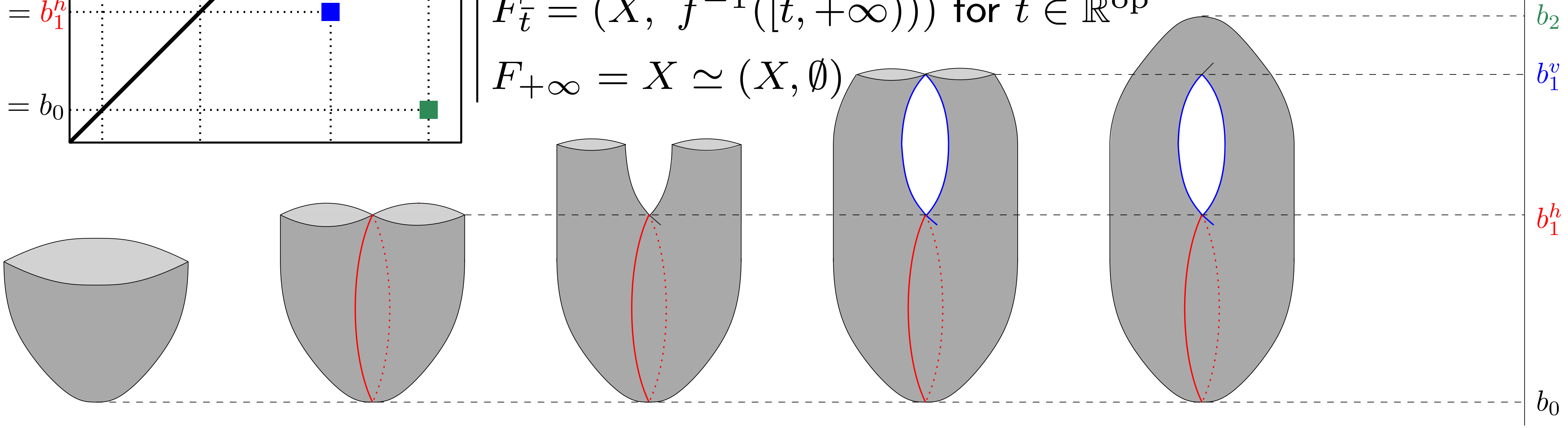
$$T = \mathbb{R} \cup \{+\infty\} \cup \mathbb{R}^{\mathrm{op}}$$

$F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$

$F_{\bar{t}} = (X,\ f^{-1}([t, +\infty)))$ for $\bar{t} \in \mathbb{R}^{\mathrm{op}}$

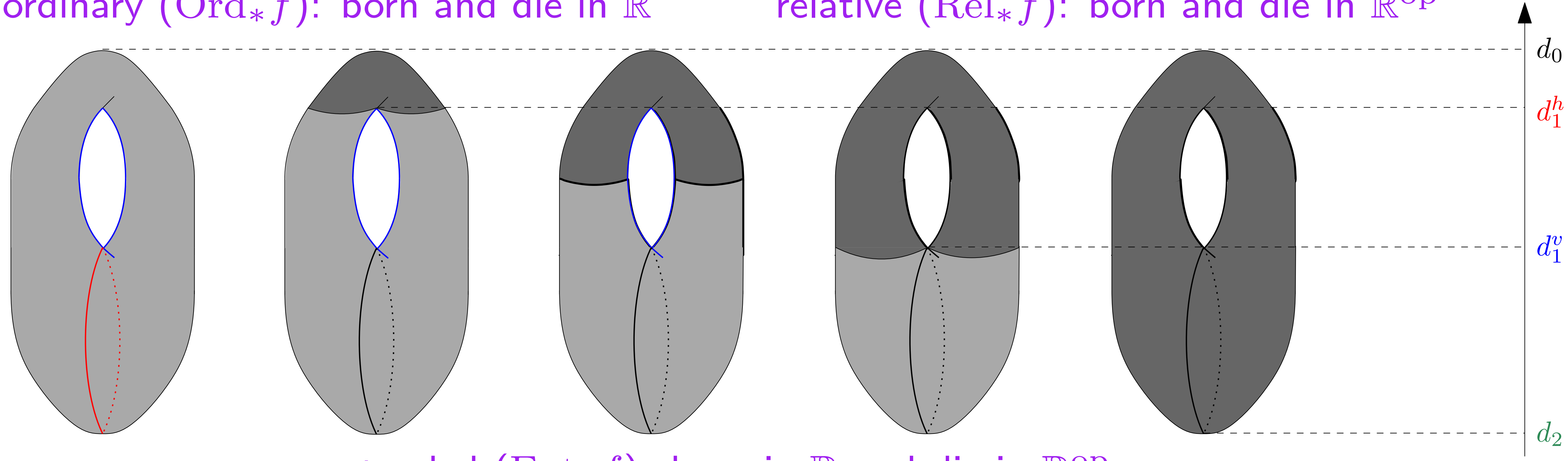
$F_{+\infty} = X \simeq (X, \emptyset)$





ordinary $(\mathrm{Ord}_* f)$: born and die in $\mathbb{R}$

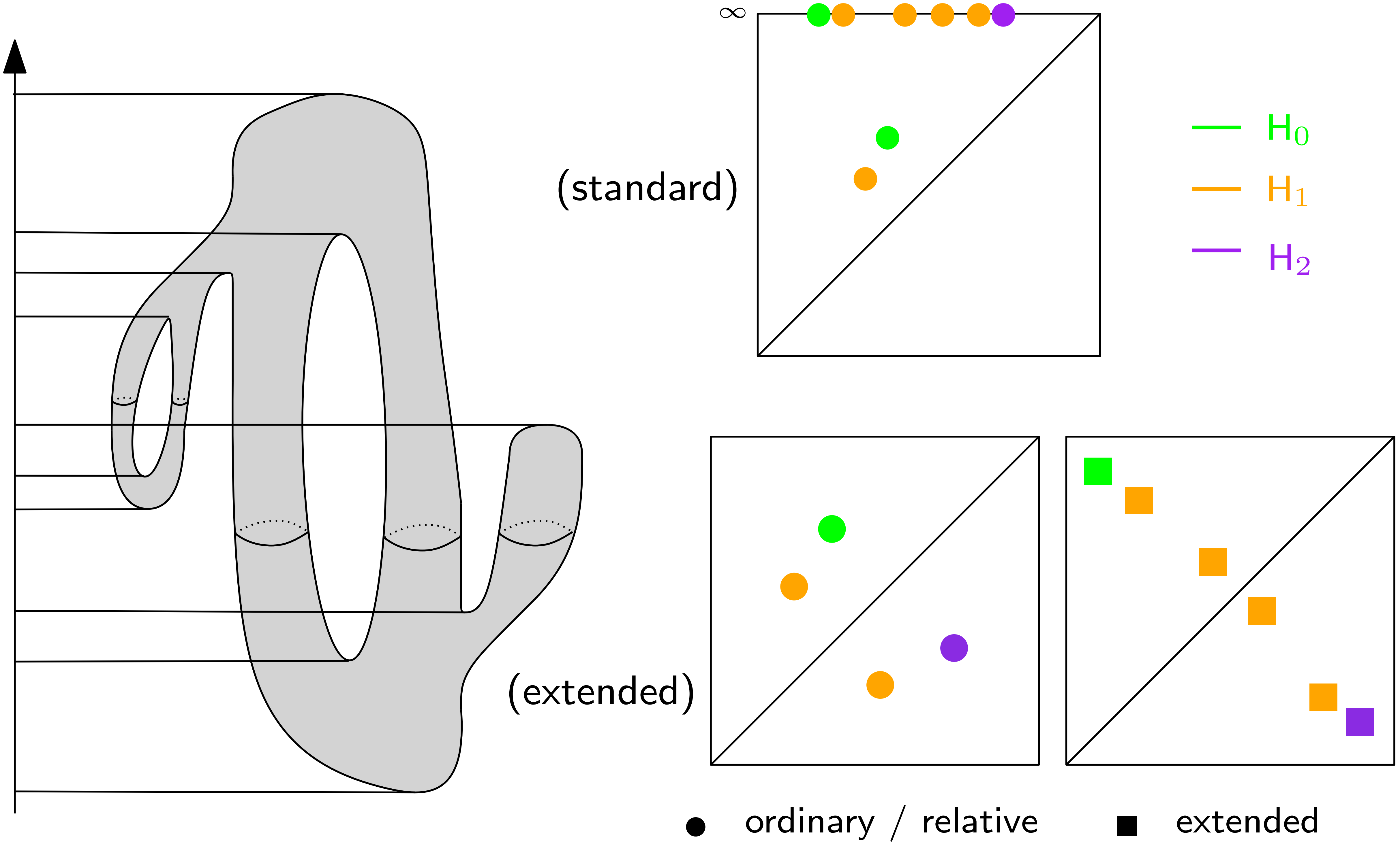
relative $(\mathrm{Rel}_* f)$: born and die in $\mathbb{R}^{\mathrm{op}}$



extended $(\mathrm{Ext}_* f)$: born in $\mathbb{R}$ and die in $\mathbb{R}^{\mathrm{op}}$

Extended persistence

Example (X surface in $\mathbb{R}^3$, f height function):



Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,
- analysis of 2d cortical thickness data,
- time series analysis,

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,
- analysis of 2d cortical thickness data,
- time series analysis,
- refinement of the classification of NBA players,
- discrimination of electroencephalogram signals recorded before and during epileptic seizures,
- statistical analysis of orthodontic data,
- measurement of structural changes during lipid vesicle fusion,
- characterization of the frequency and scale of lateral gene transfer in pathogenic bacteria,
- pattern detection in gene expression data,
- study of the cosmic web and its filamentary structure,

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,
- analysis of 2d cortical thickness data,
- time series analysis,
- refinement of the classification of NBA players,
- discrimination of electroencephalogram signals recorded before and during epileptic seizures,
- statistical analysis of orthodontic data,
- measurement of structural changes during lipid vesicle fusion,
- characterization of the frequency and scale of lateral gene transfer in pathogenic bacteria,
- pattern detection in gene expression data,
- study of the cosmic web and its filamentary structure,

\- large variety of applications

\- 2 reasons for using TDA:

- invariance + stability (good)
- fashionable (bad)