Probabilistic and Statistical Analysis of the Mapper algorithm in Topological Data Analysis

Journal Club on Topological Data Analysis

Universität Heidelberg

Mathieu Carrière—joint work with S. Oudot, B. Michel, R. Rabadan

M.C., S. Oudot, *Structure and stability of the 1-dimensional Mapper*. Foundations of Computational Mathematics, 2017.

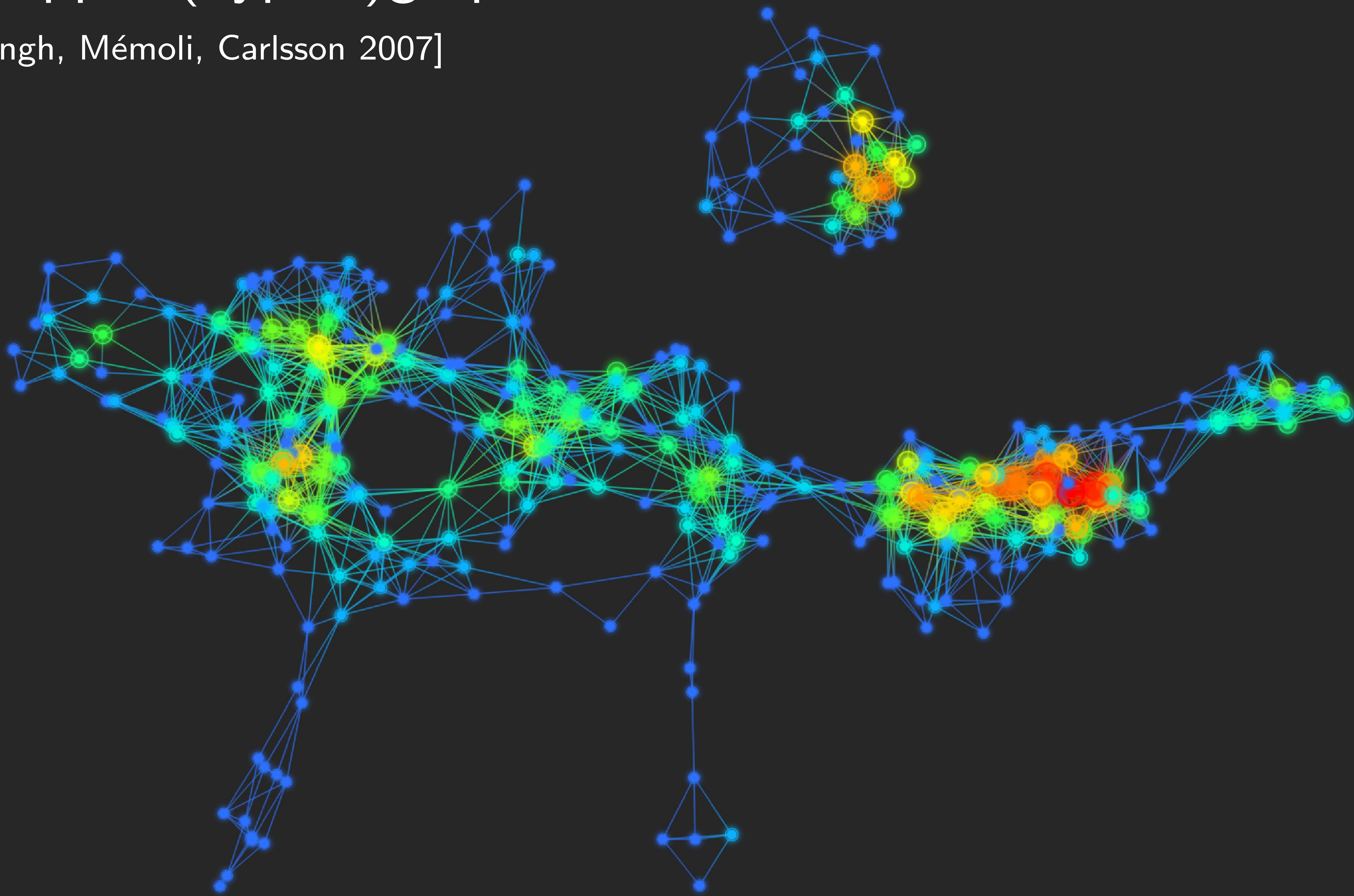
M.C., B. Michel, S. Oudot, *Statistical analysis and parameter selection for the Mapper*. Journal of Machine Learning Research, 2018.

INVENTEURS DU MONDE NUMÉRIQUE

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

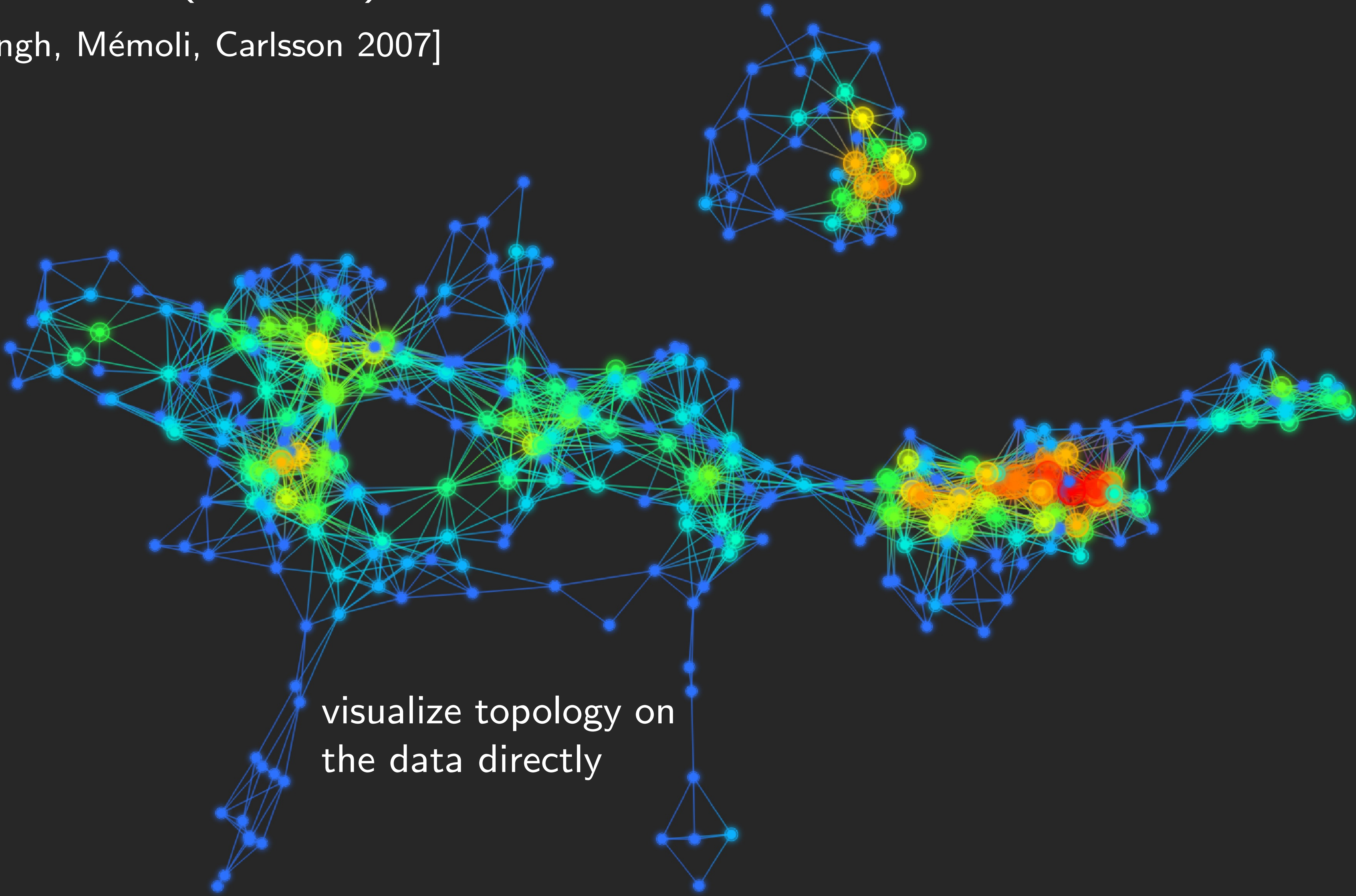
Mapper (hyper-)graphs

[Singh, Mémoli, Carlsson 2007]



Mapper (hyper-)graphs

[Singh, Mémoli, Carlsson 2007]



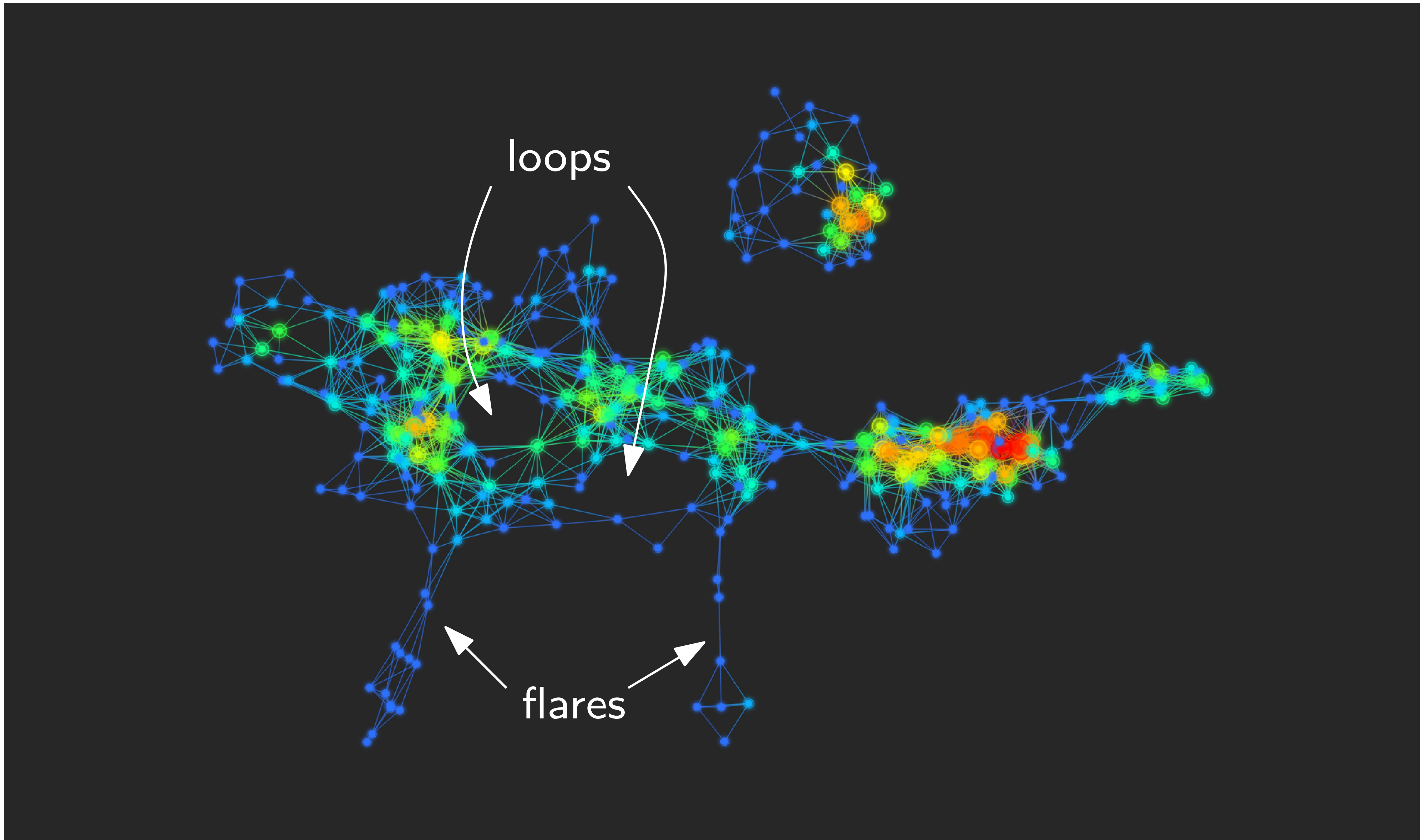
visualize topology on
the data directly

Mapper in applications

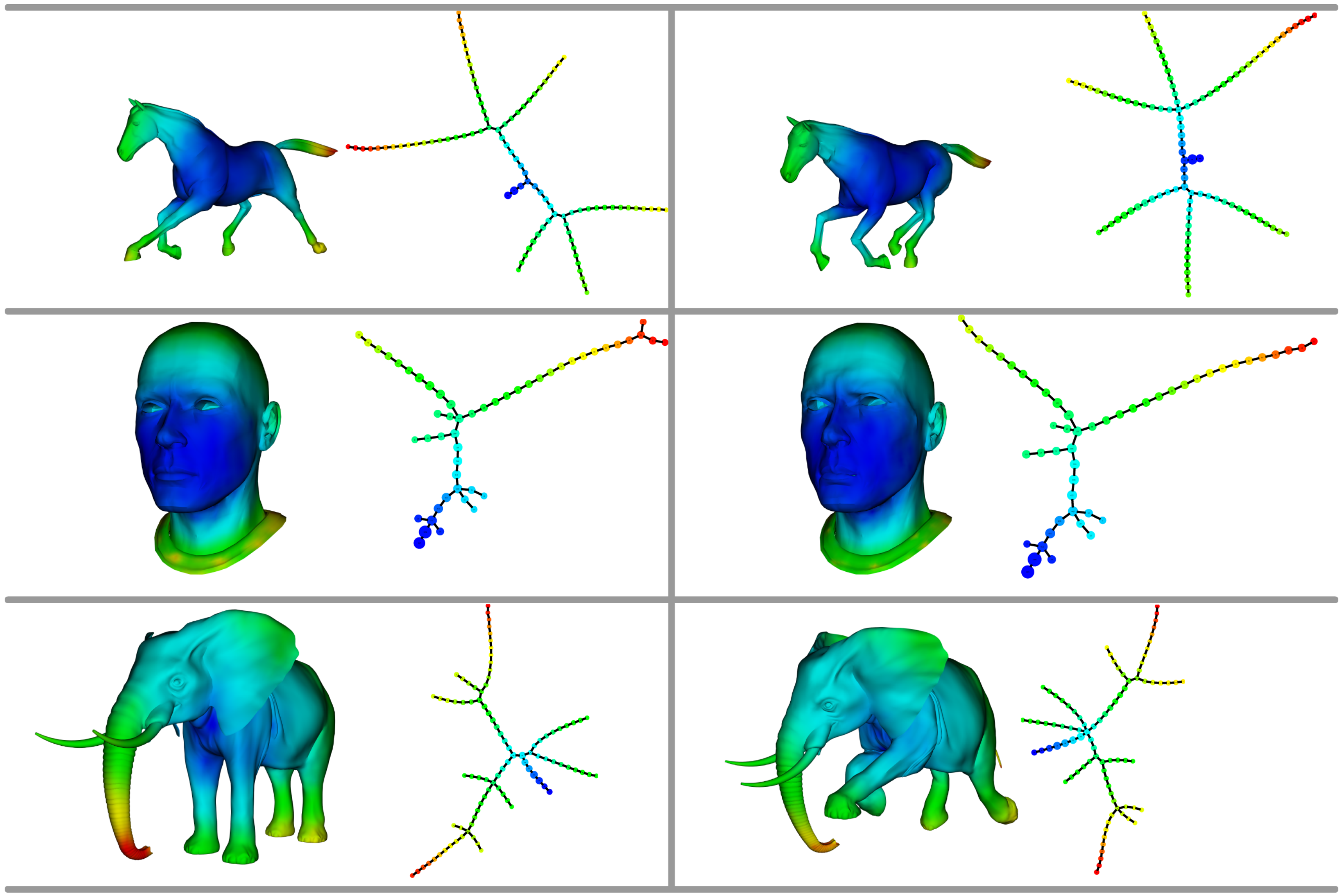
Two types of applications:

- clustering
- feature selection

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



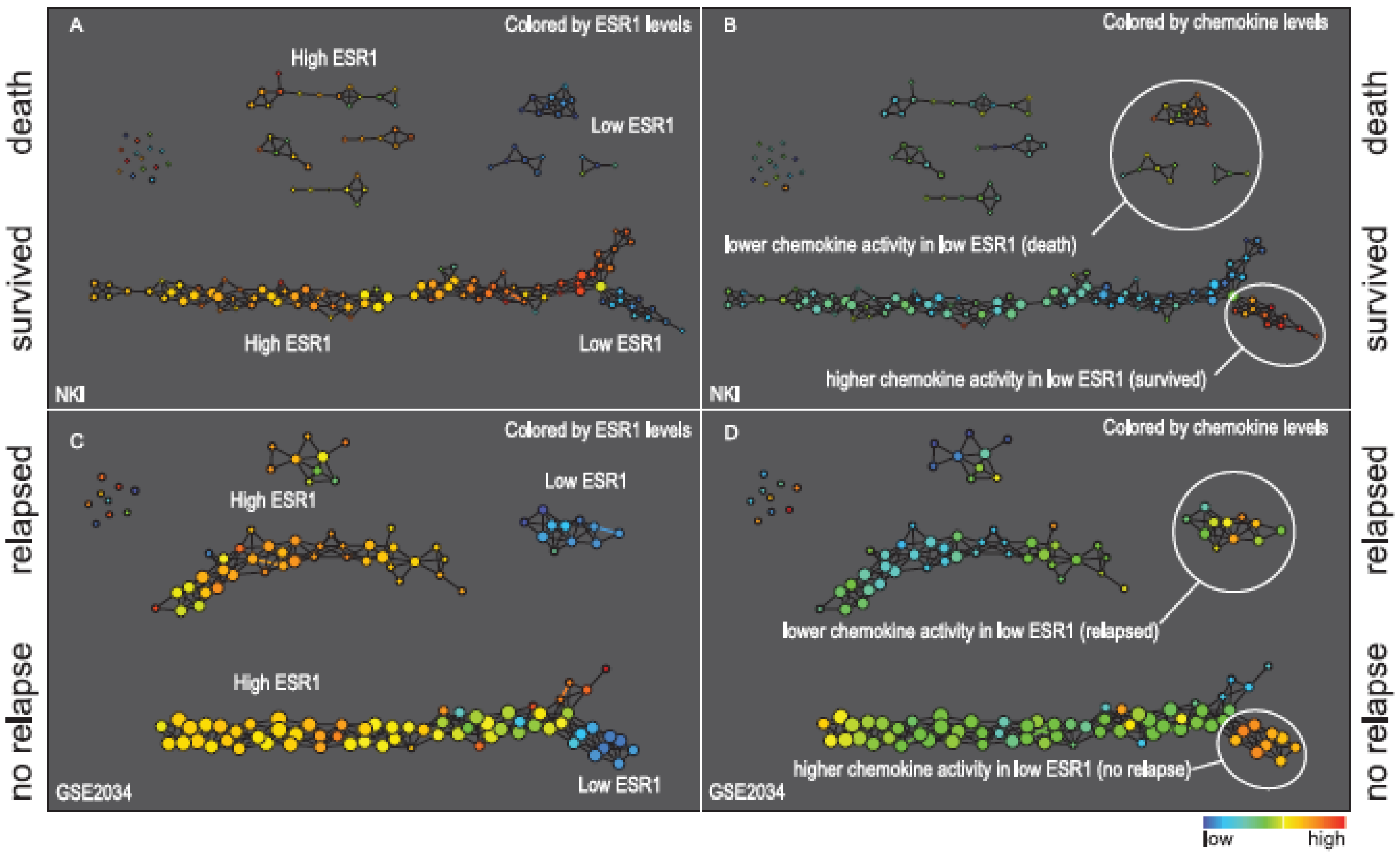
Mapper in applications



3d shapes classification

[Singh, Mémoli, Carlsson 2007]

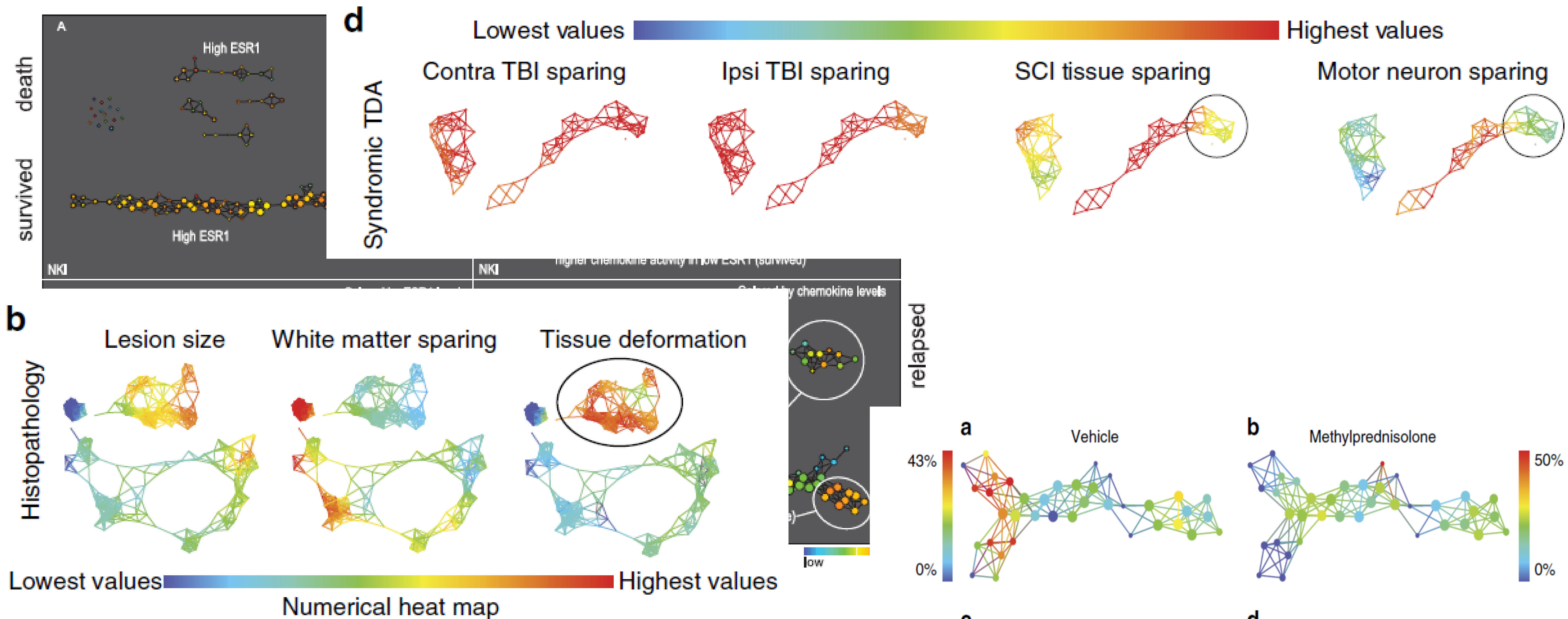
Mapper in applications



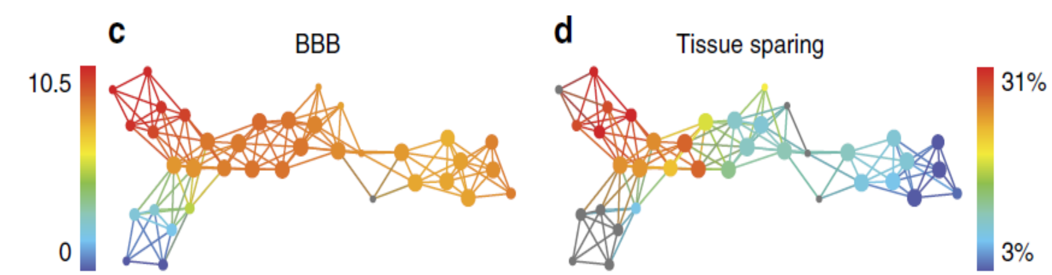
breast cancer subtype identification

[Nicolau et al. 2011]

Mapper in applications

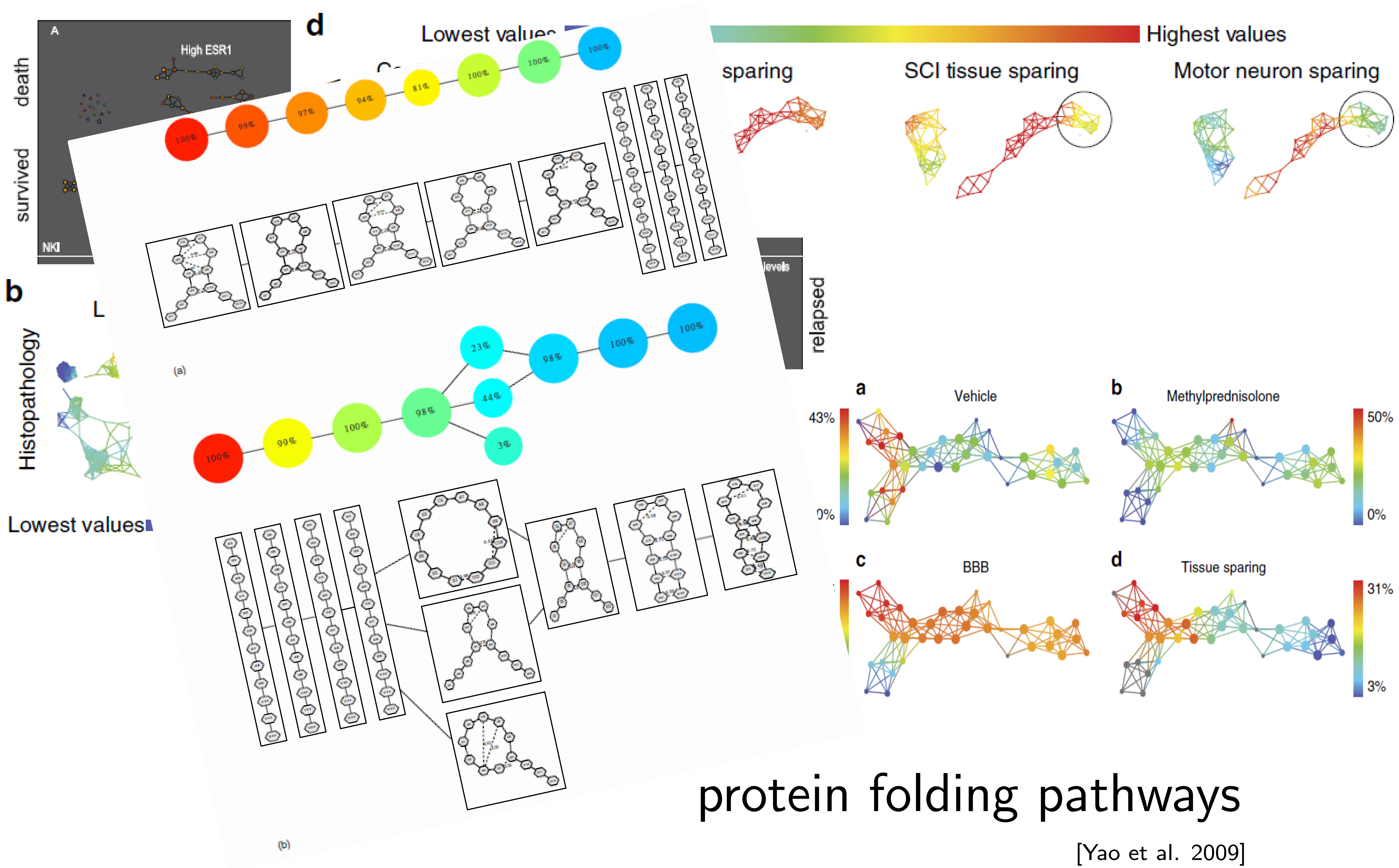


recovery from spinal cord injuries



[Nielson et al. 2015]

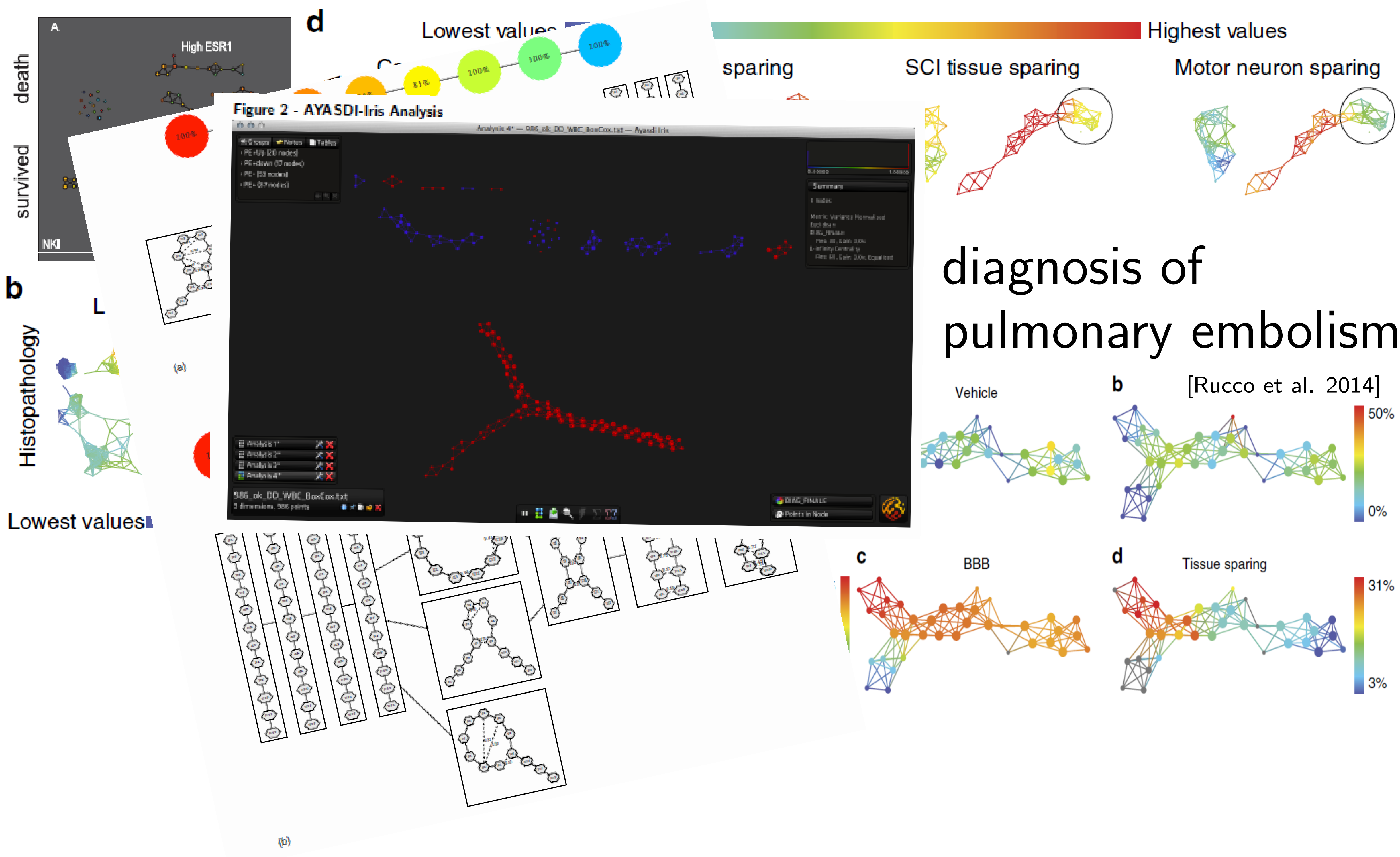
Mapper in applications



protein folding pathways

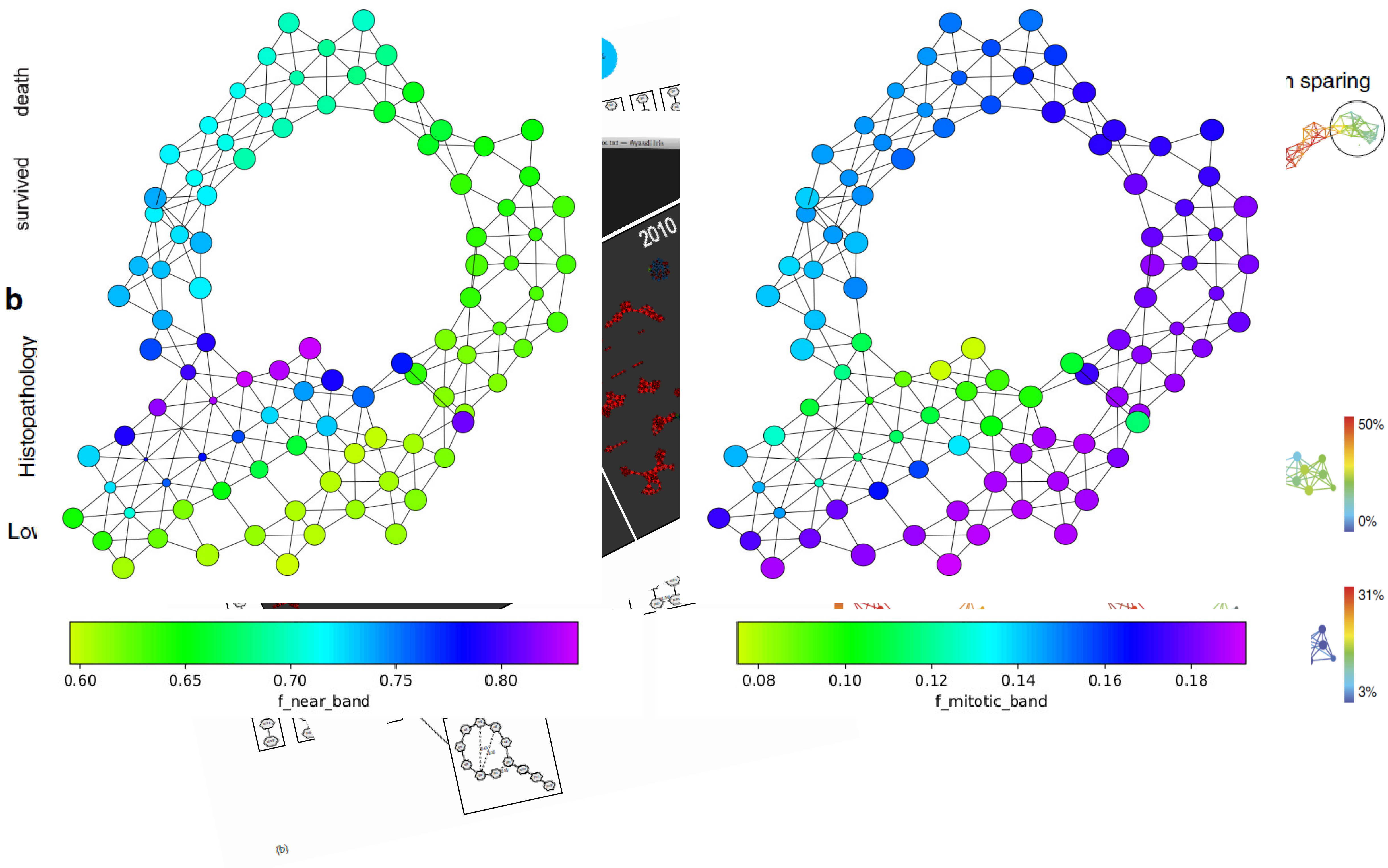
[Yao et al. 2009]

Mapper in applications

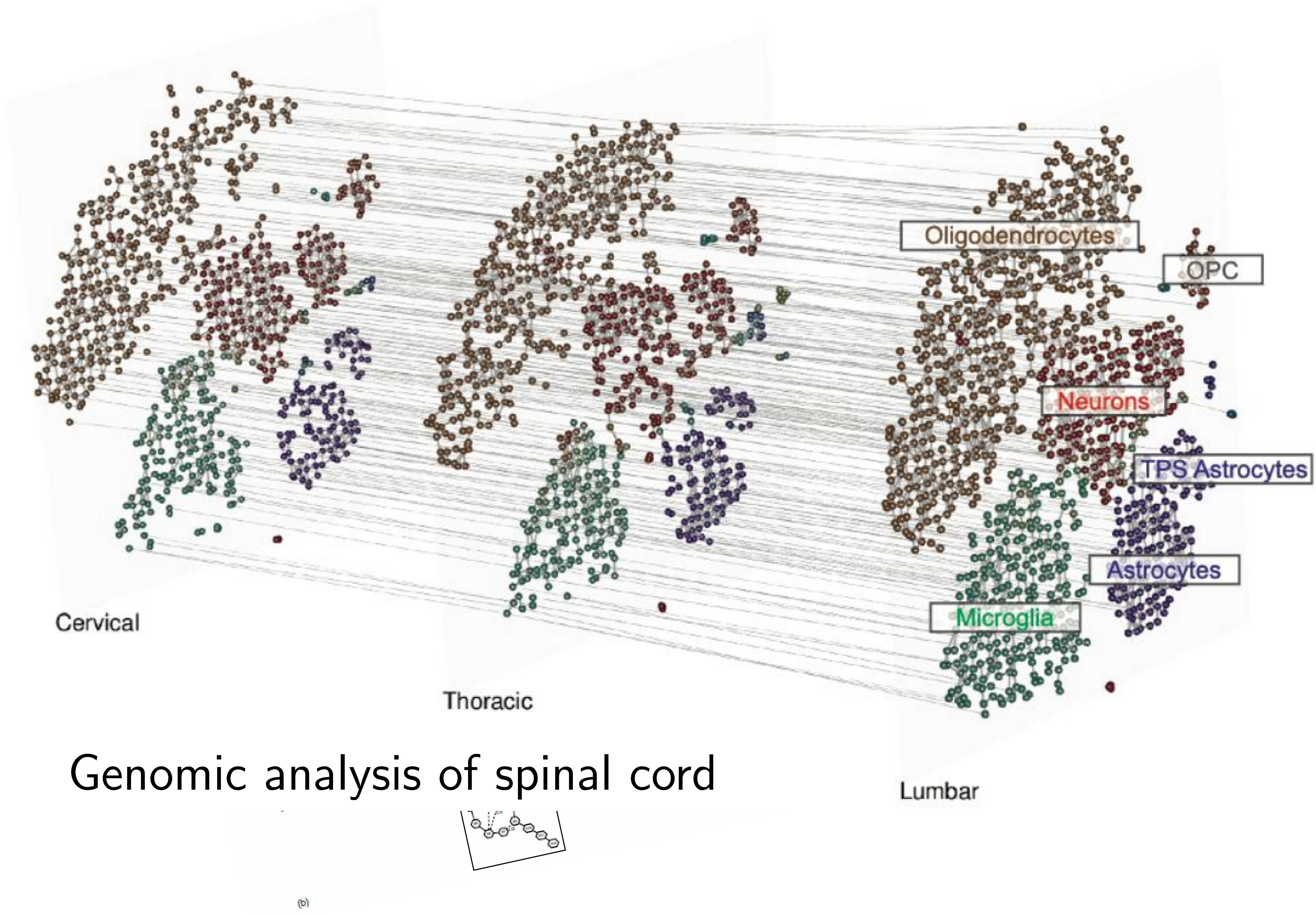


Mapper in applications

Formal identification of cell cycle

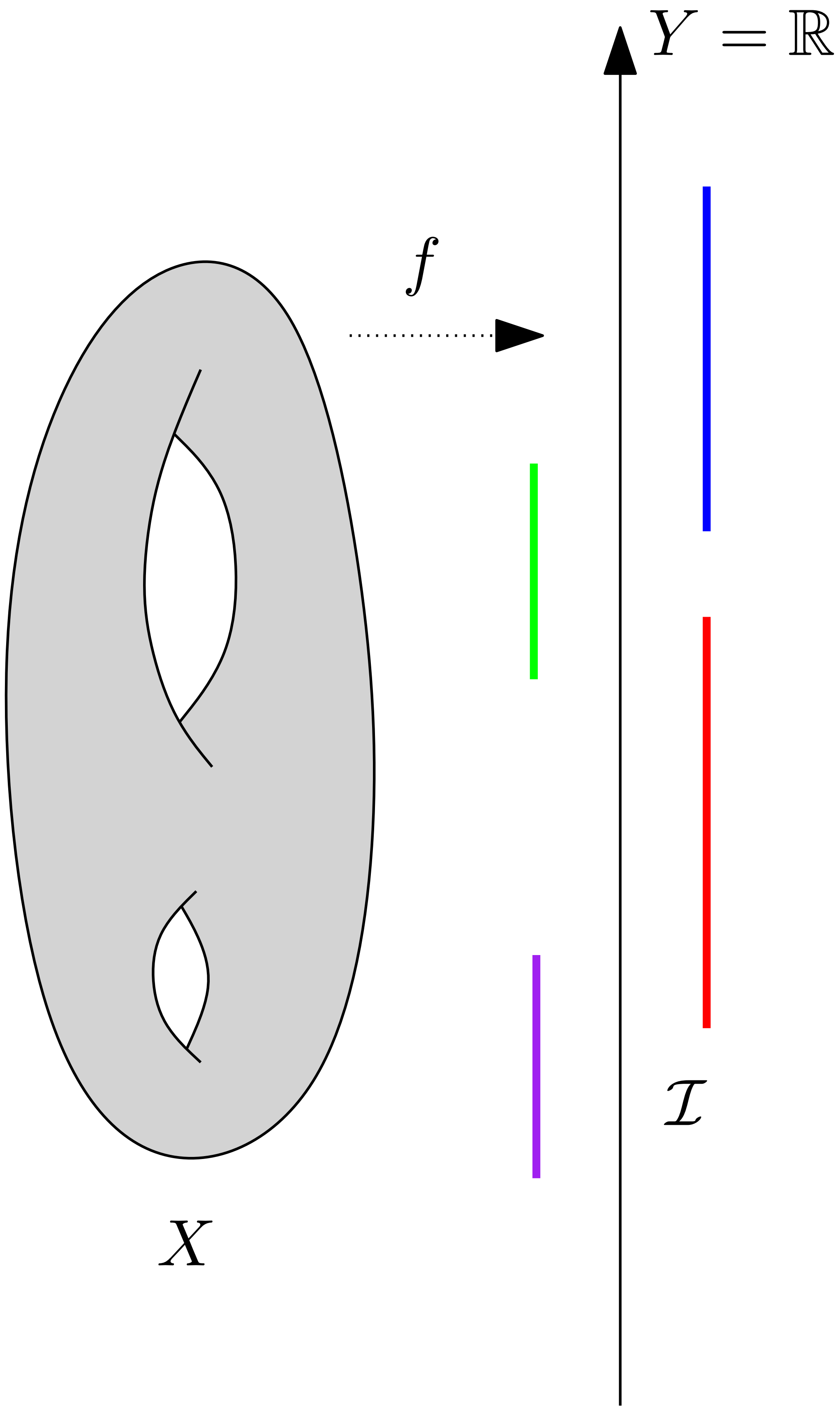


Mapper in applications

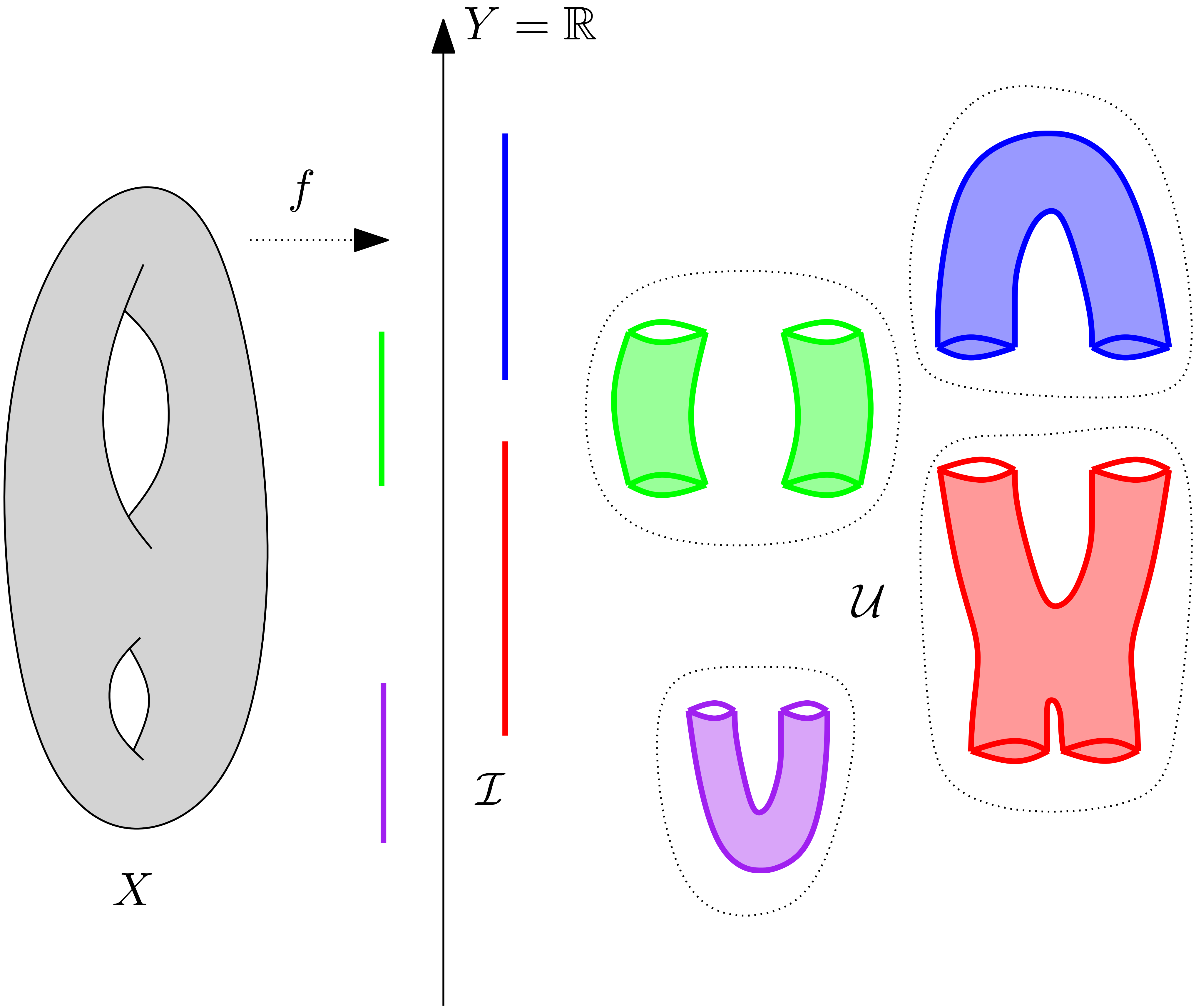


Genomic analysis of spinal cord

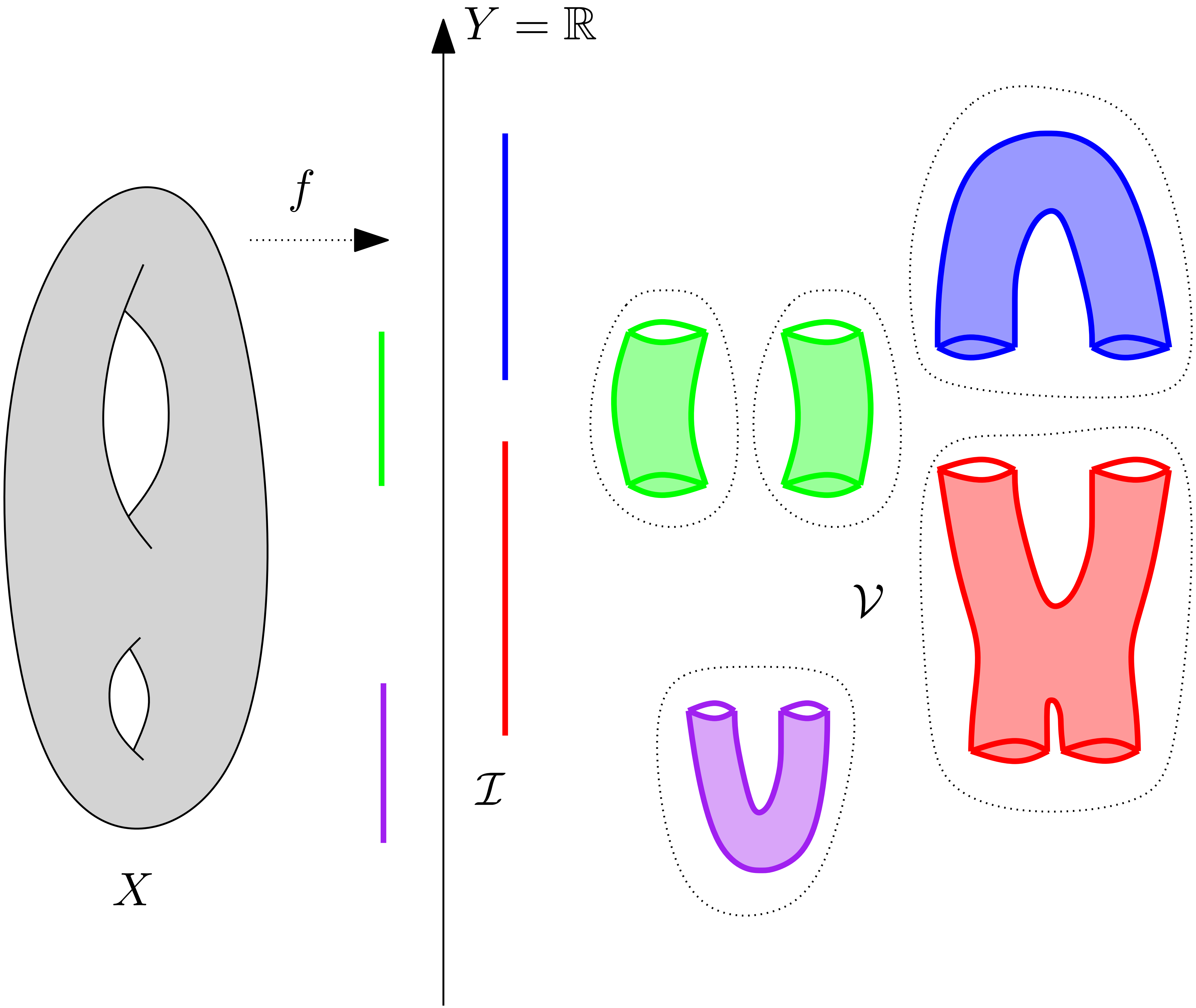
Mapper in the continuous setting



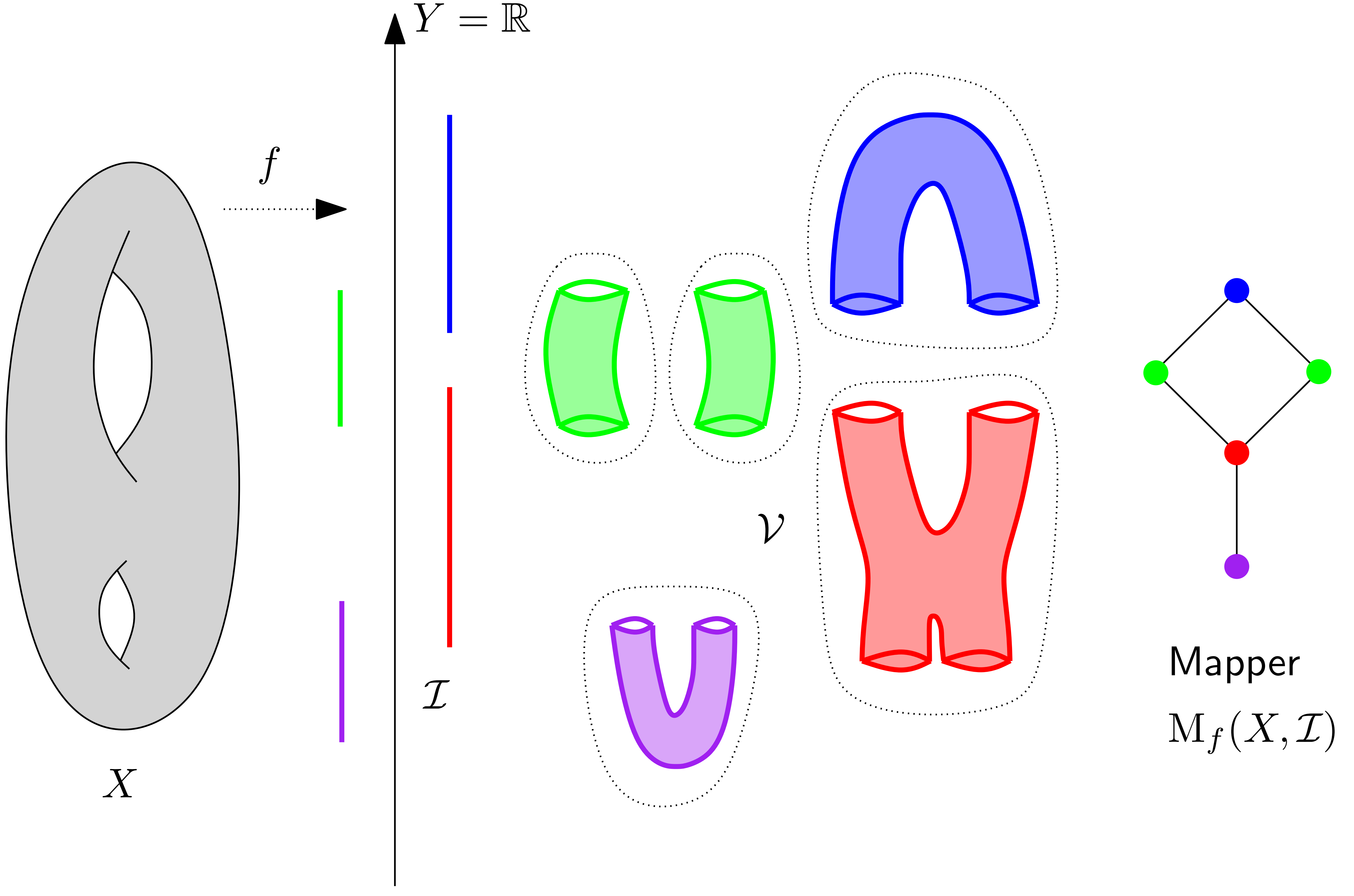
Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting

Input:

- topological space X
- continuous function $f : X \to Y$ ($Y = \mathbb{R}$ in this talk)
- 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 $\to$ 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 $f : P \to Y$ $\quad$ ($Y = \mathbb{R}$ in this talk)
- 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: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$

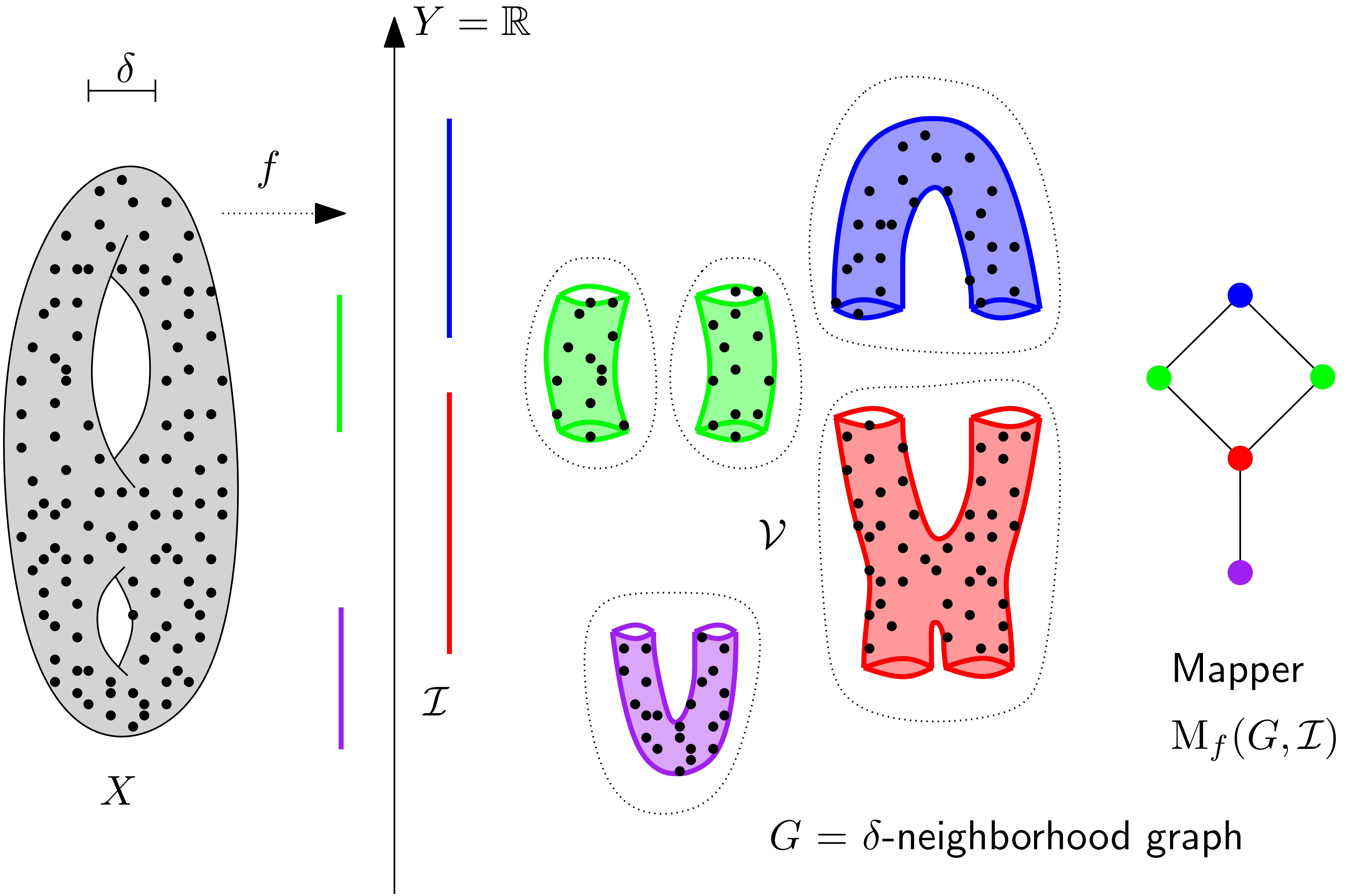
• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in G $\to$ connected cover $\mathcal{V}$

• 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 $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

(intersections materialized by data points)

Mapper in practice

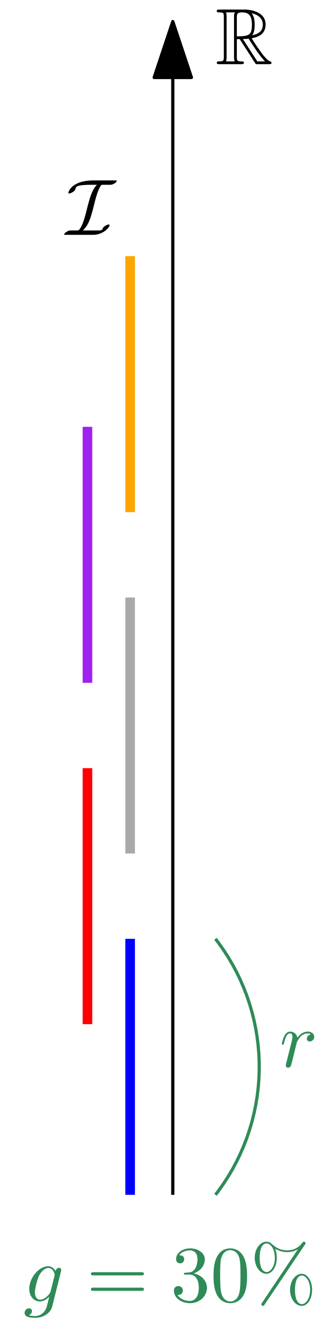


Choice of parameters

Parameters:

- function $f : P \to \mathbb{R}$ ← lens | filter
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals ← range scale
- neighborhood size δ ← geometric scale

Choice of parameters



Parameters:

- function $f : P \to \mathbb{R}$ ← lens | filter
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals ← range scale
- neighborhood size δ ← geometric scale

$\rightarrow$ uniform cover $\mathcal{I}$: ← range scale

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

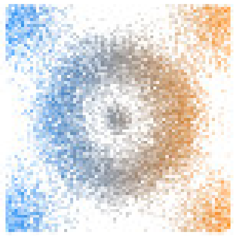
Choice of parameters

→ in practice: trial-and-error

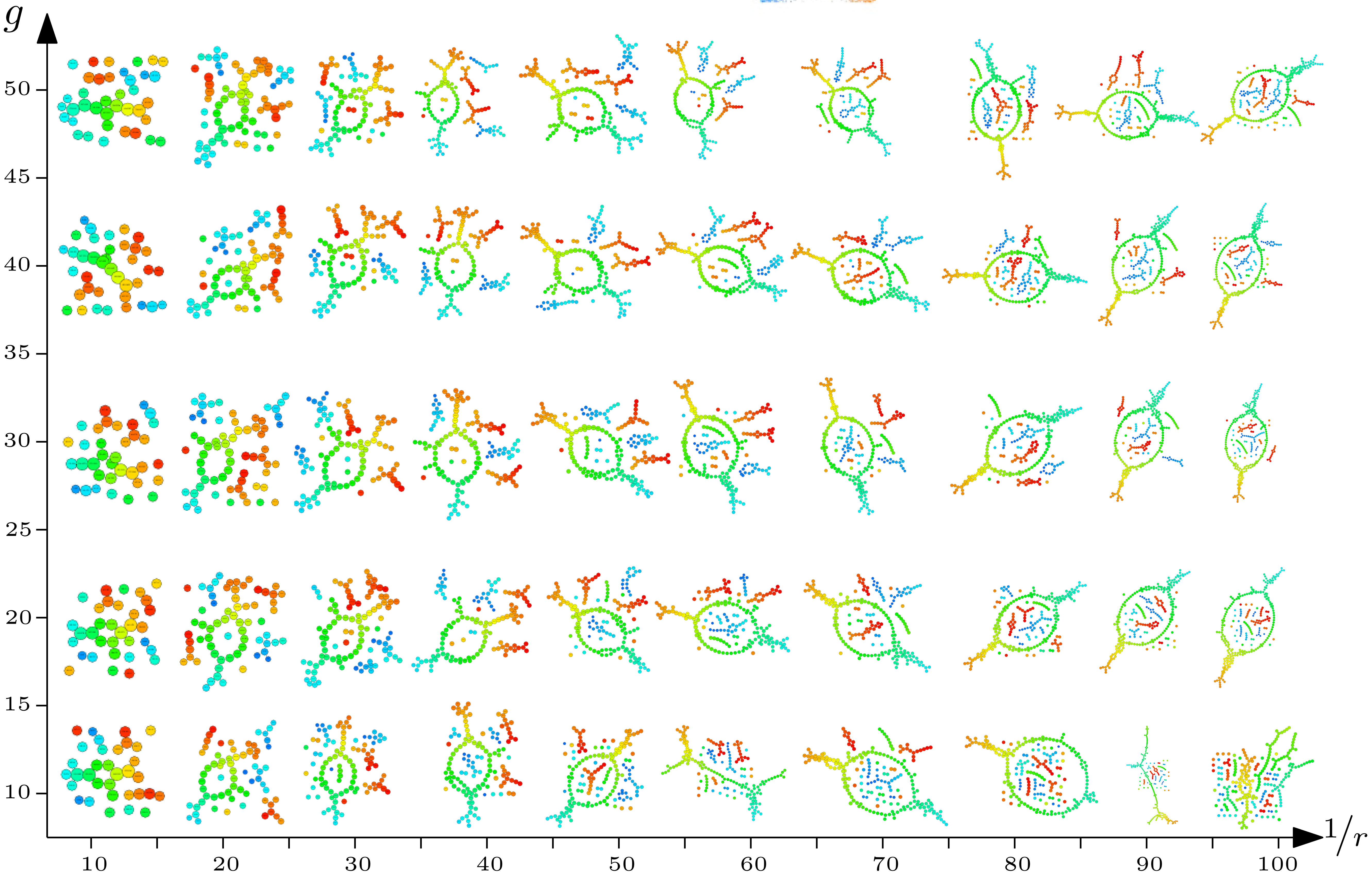
high-dimensional data sets[40,48]. This is performed automatically within the software, by deploying an ensemble machine learning algorithm that iterates through overlapping subject bins of different sizes that resample the metric space (with replacement), thereby using a combination of the metric location 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

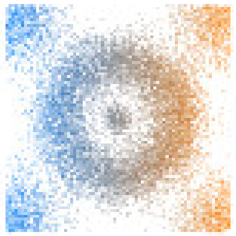
Choice of parameters



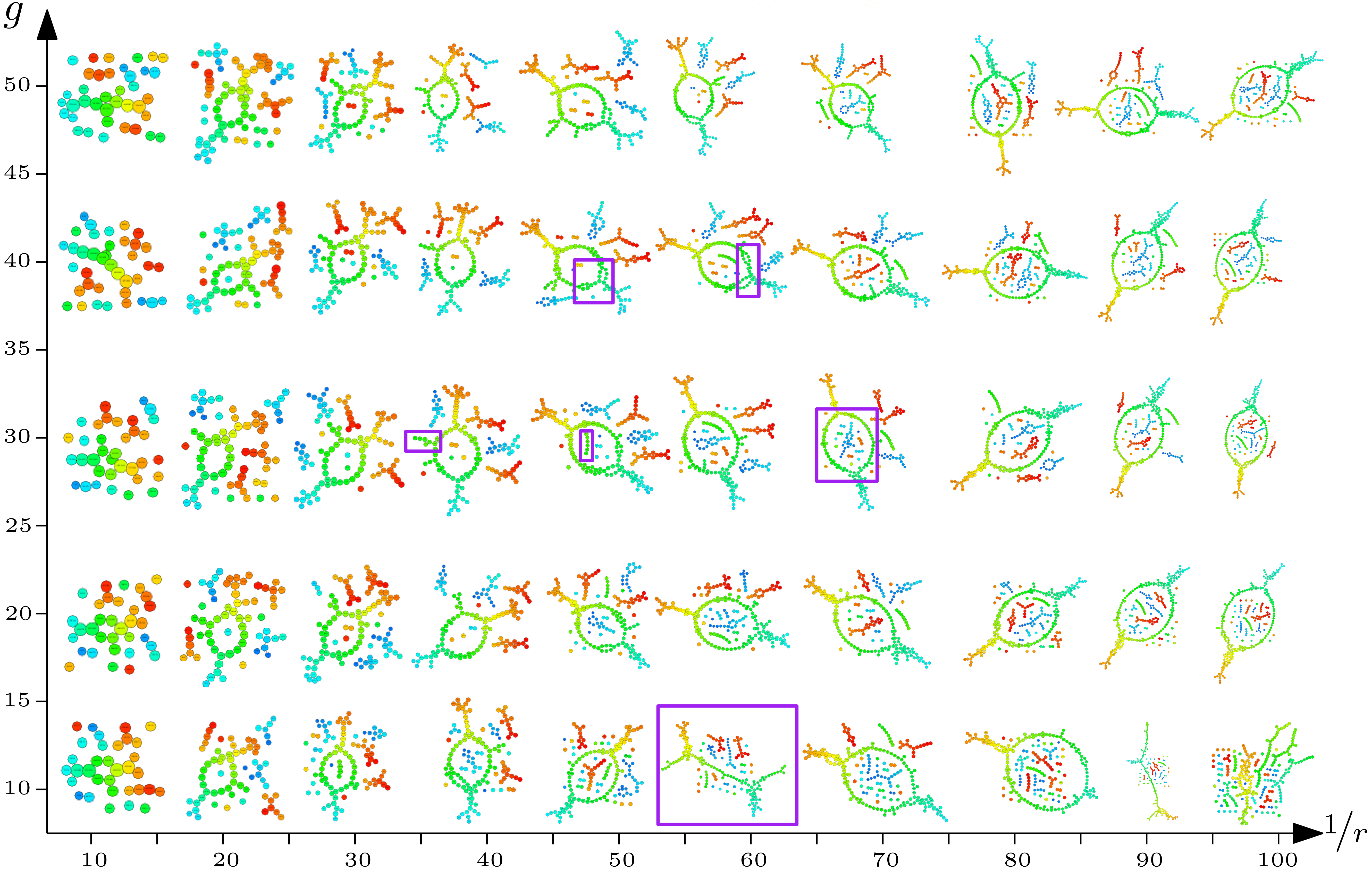
$f = f_x,\ \delta = 1\%$



Choice of parameters

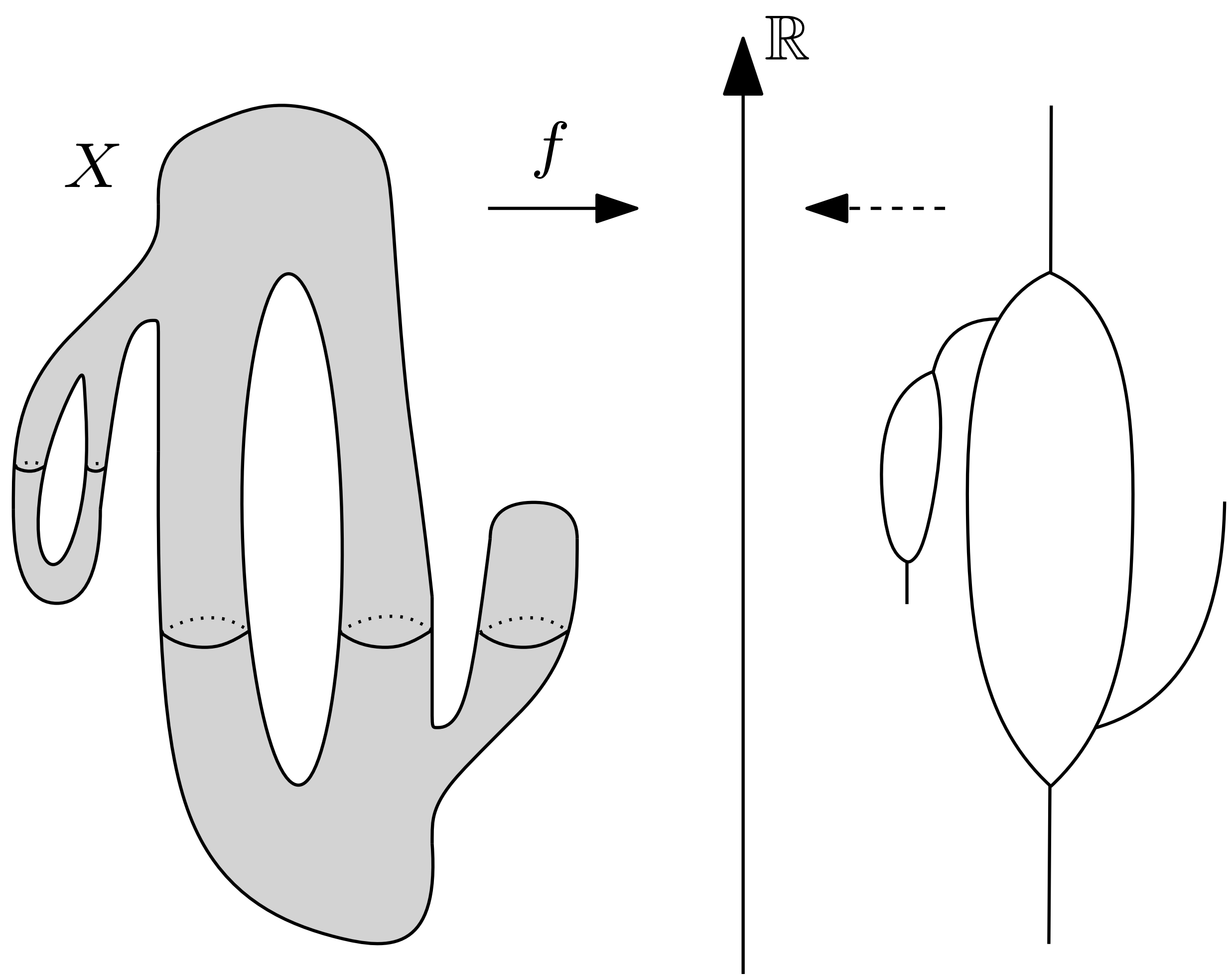


$f = f_x,\ \delta = 1\%$



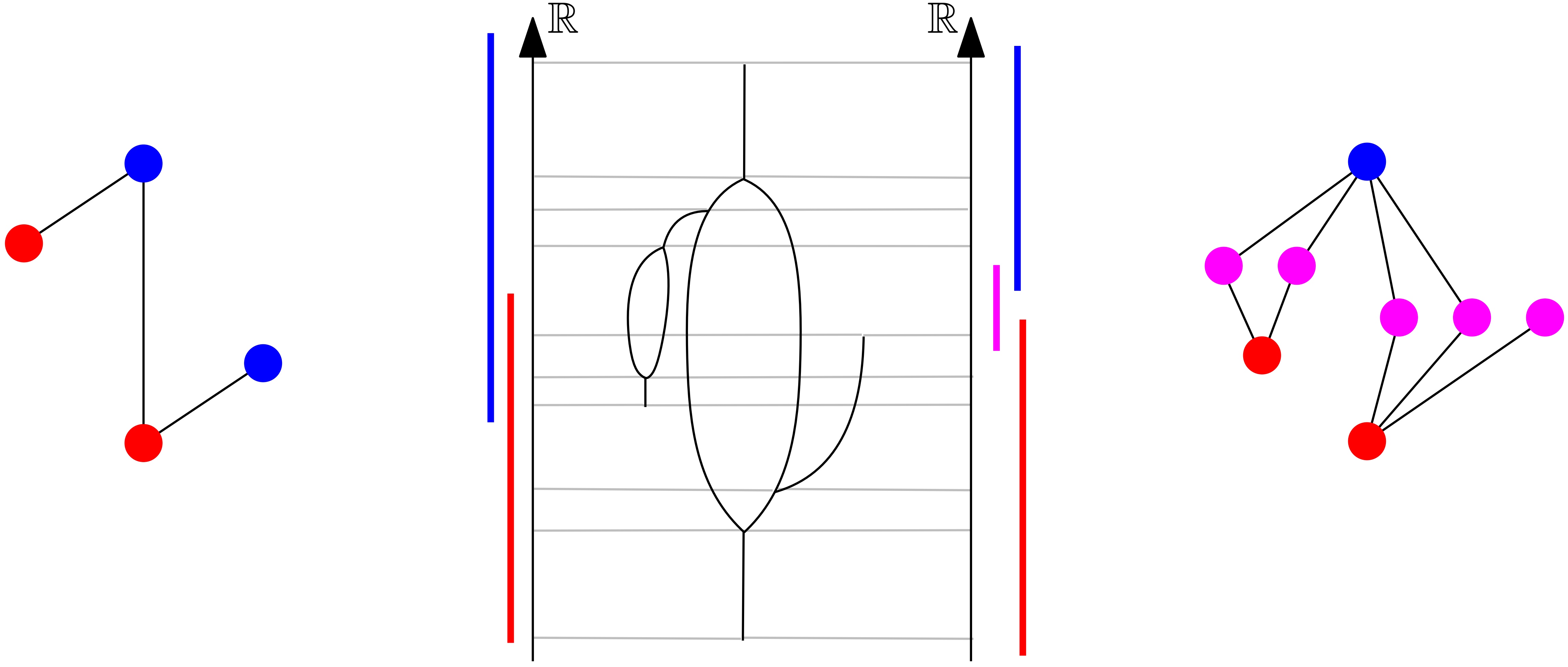
Reeb Graph

Reeb graph $\sim$ Mapper with extremely small resolution



Reeb Graph

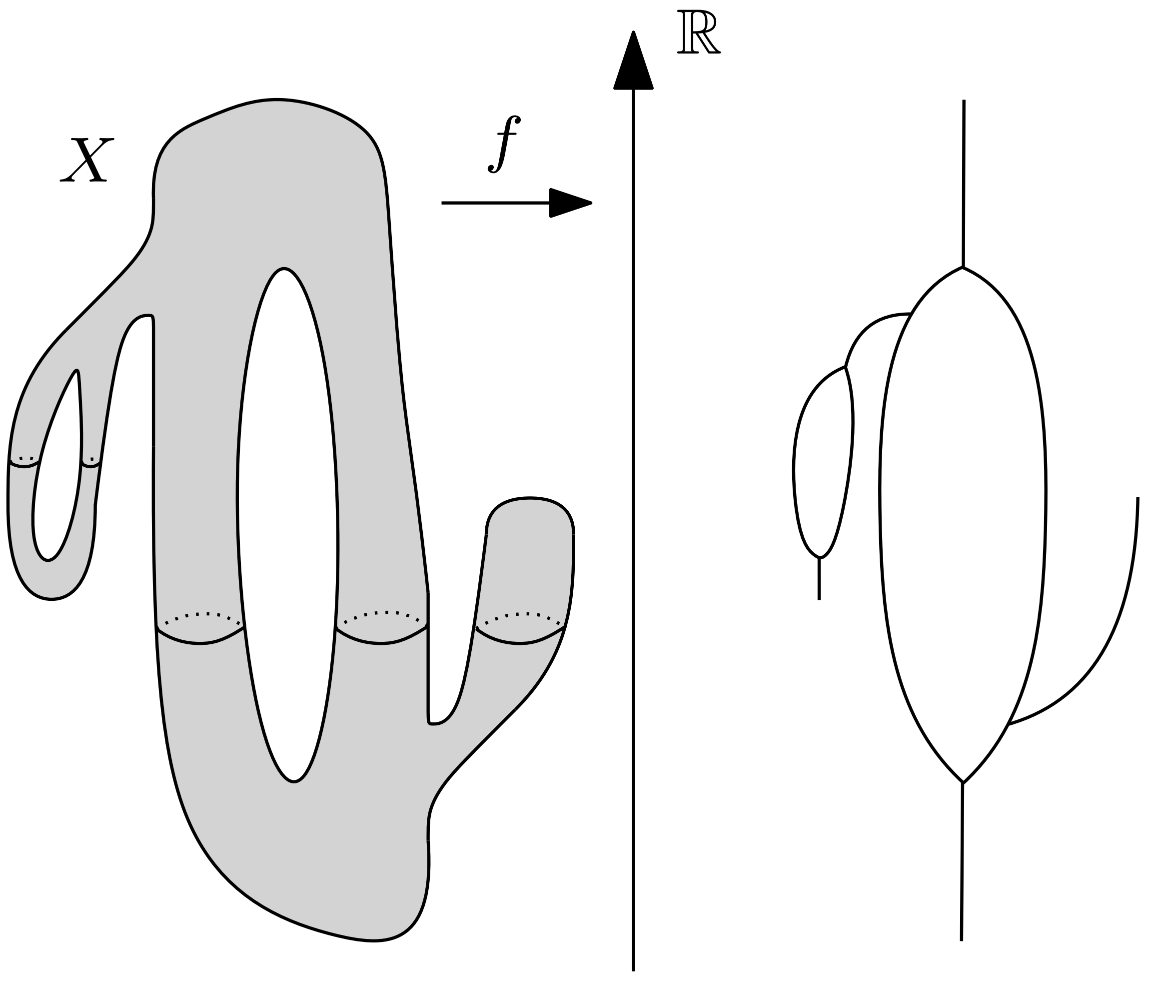
Mapper ∼ *pixelized* Reeb graph



Reeb Graph

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

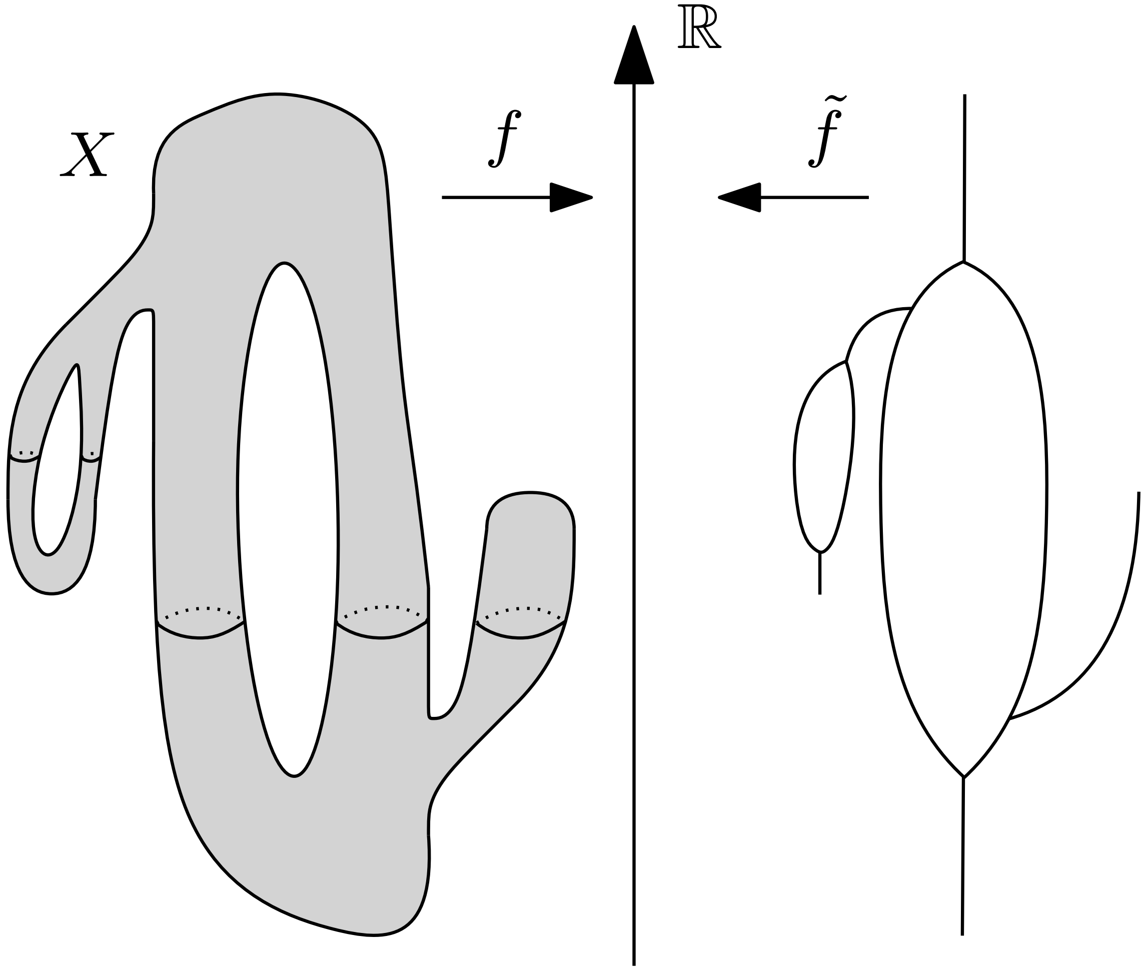
$$\mathrm{R}_f(X) := X/\sim$$

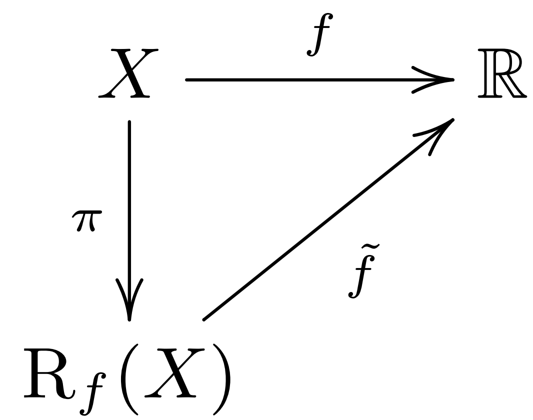


Reeb Graph

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

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

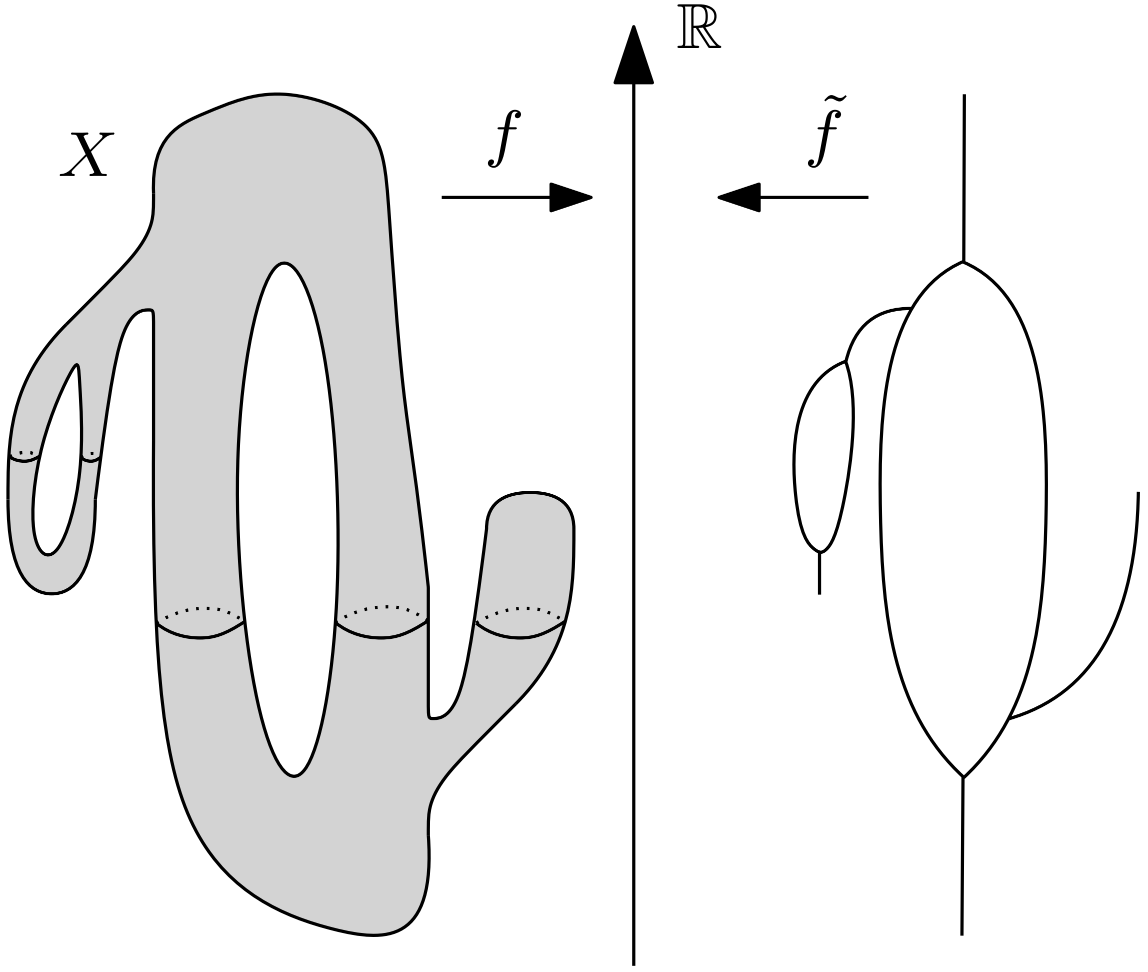


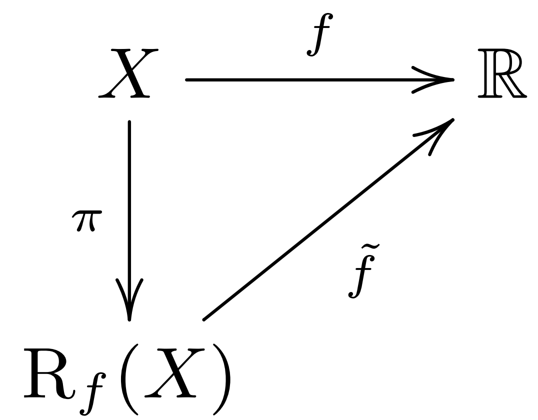


Reeb Graph

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

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



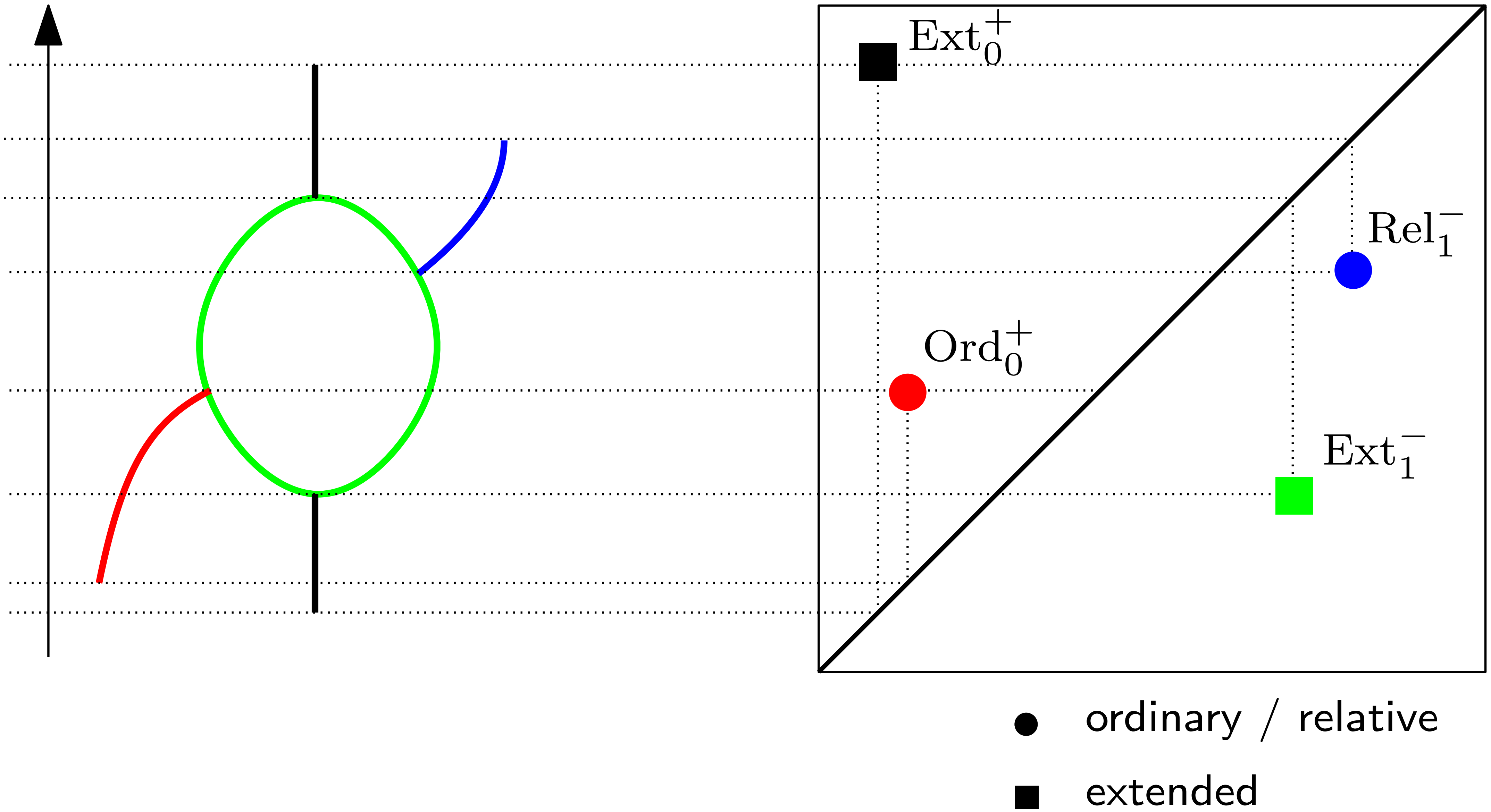


Prop: $\mathrm{R}_f(X)$ is a graph when (X, f) is Morse or of **Morse type**

Graph Descriptor

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

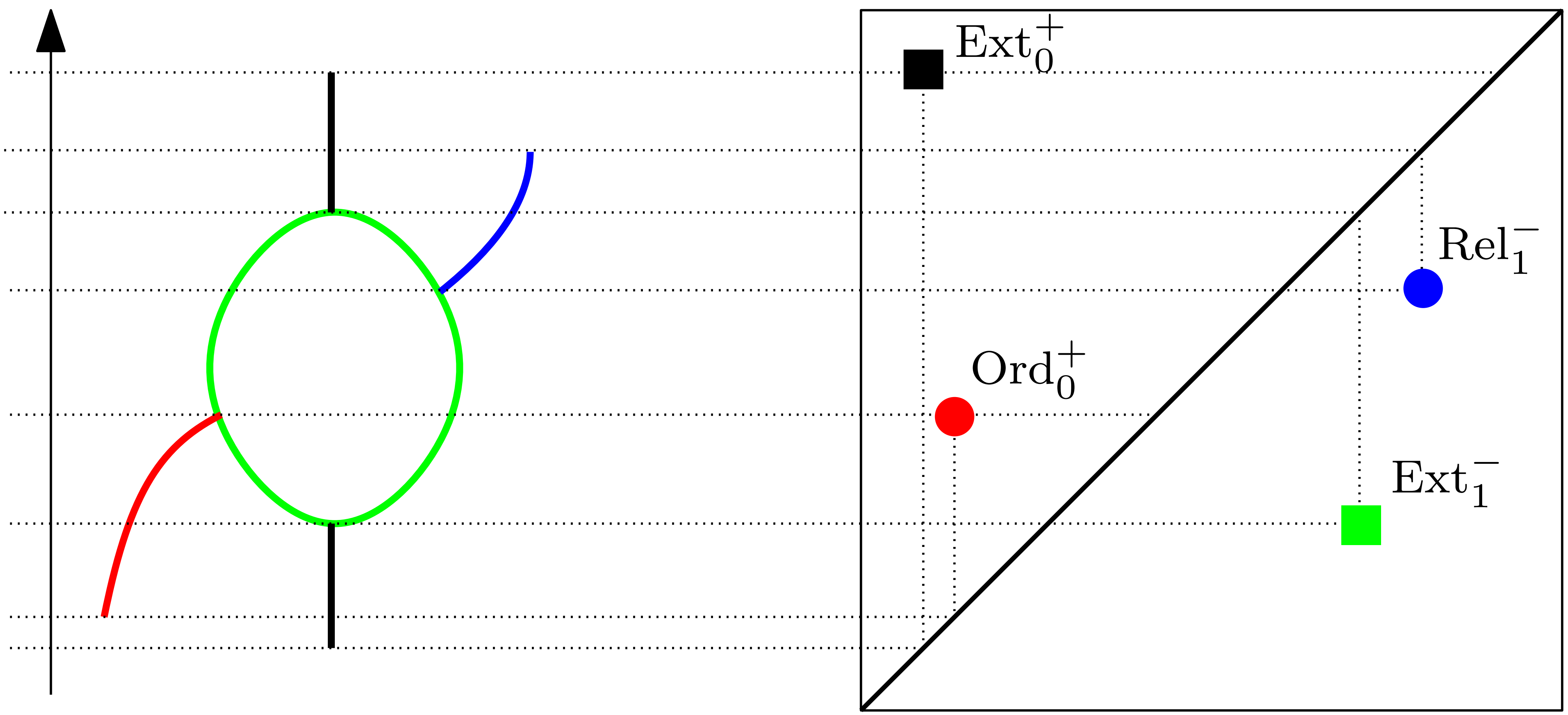
$\mathrm{Ord}_0\,\tilde{f} \longleftrightarrow$ downward branches	$\mathrm{Ext}_0\,\tilde{f} \longleftrightarrow$ trunks (cc)
$\mathrm{Rel}_1\,\tilde{f} \longleftrightarrow$ upward branches	$\mathrm{Ext}_1\,\tilde{f} \longleftrightarrow$ loops



Graph Descriptor

$\mathrm{Dg}\,\tilde{f}$ 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)

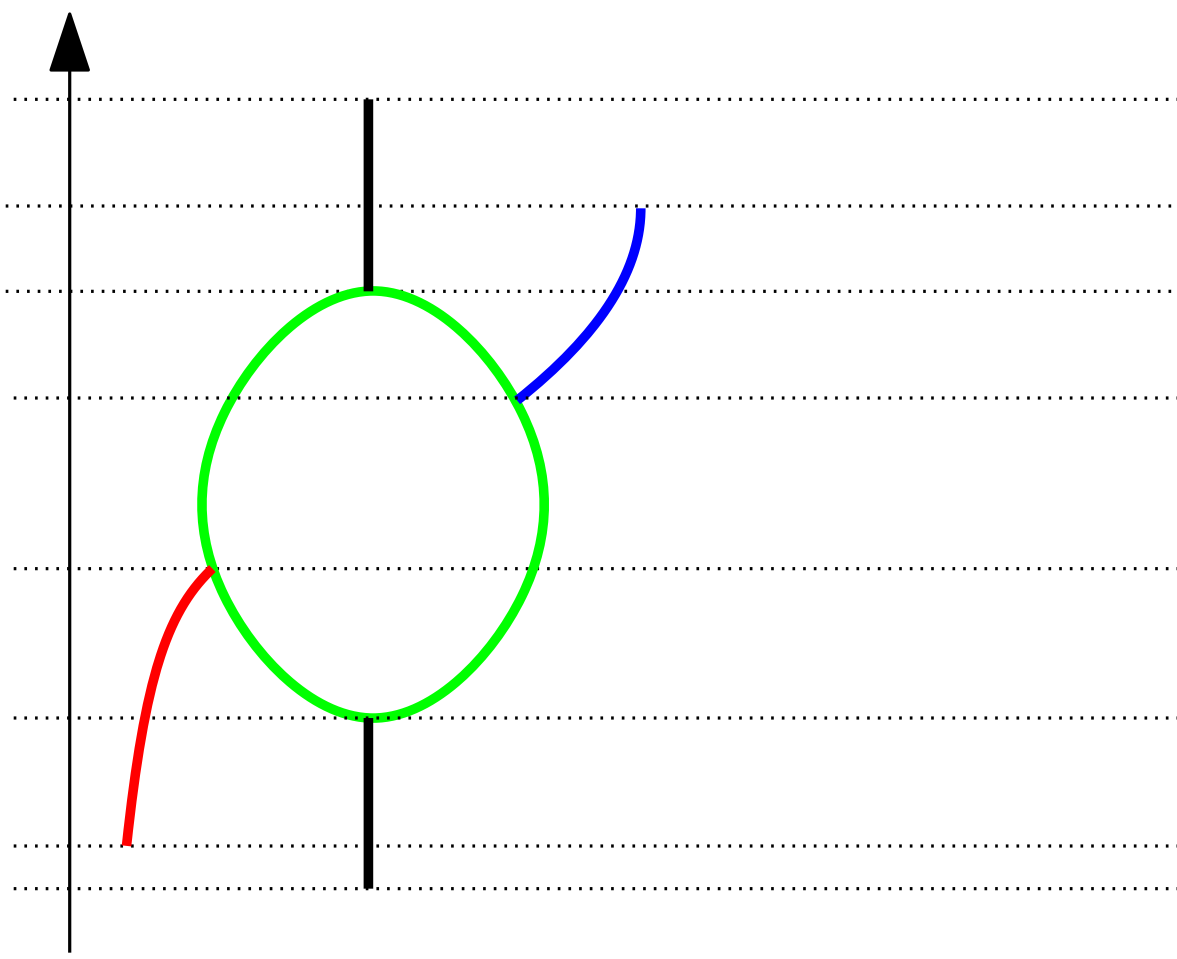
● ordinary / relative

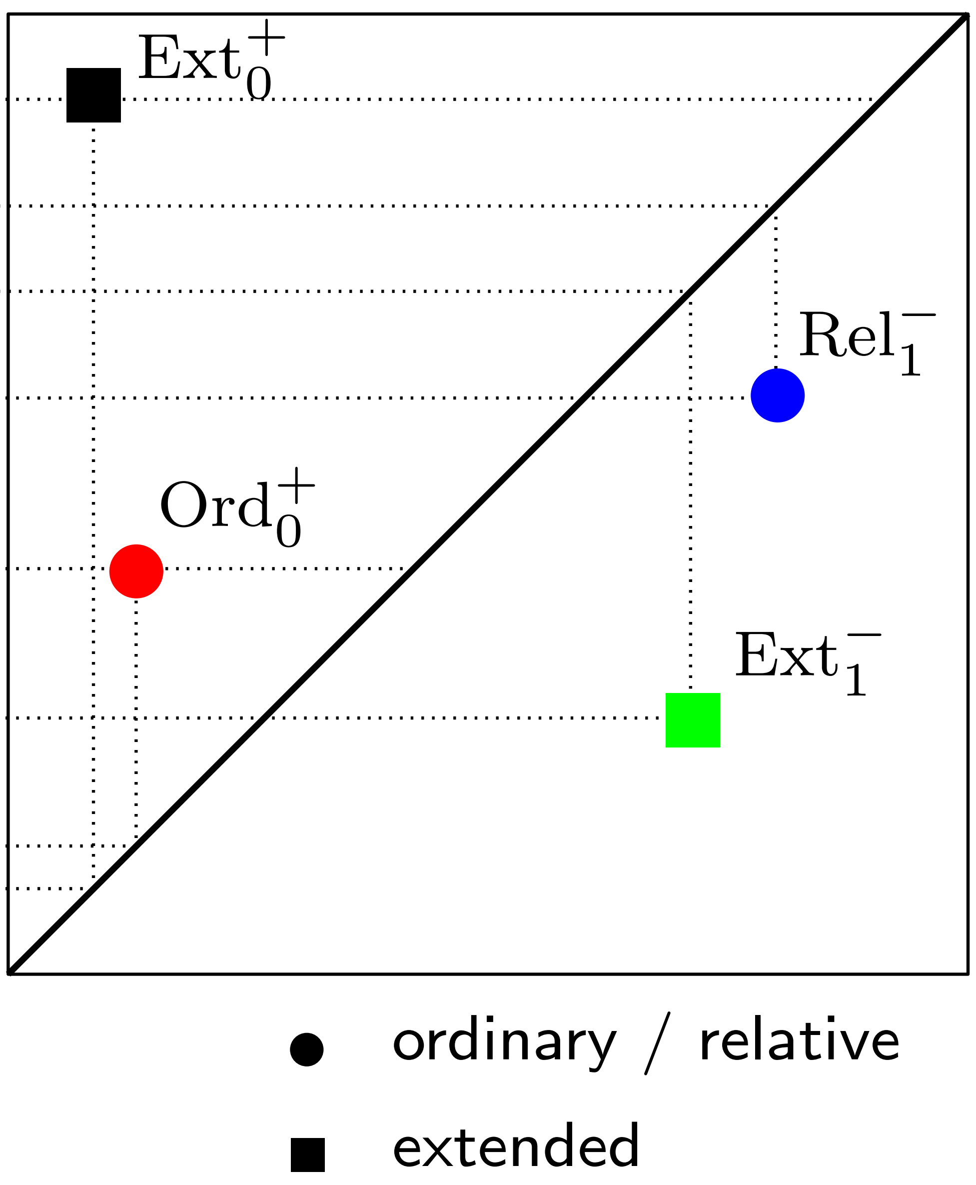
■ extended

Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family

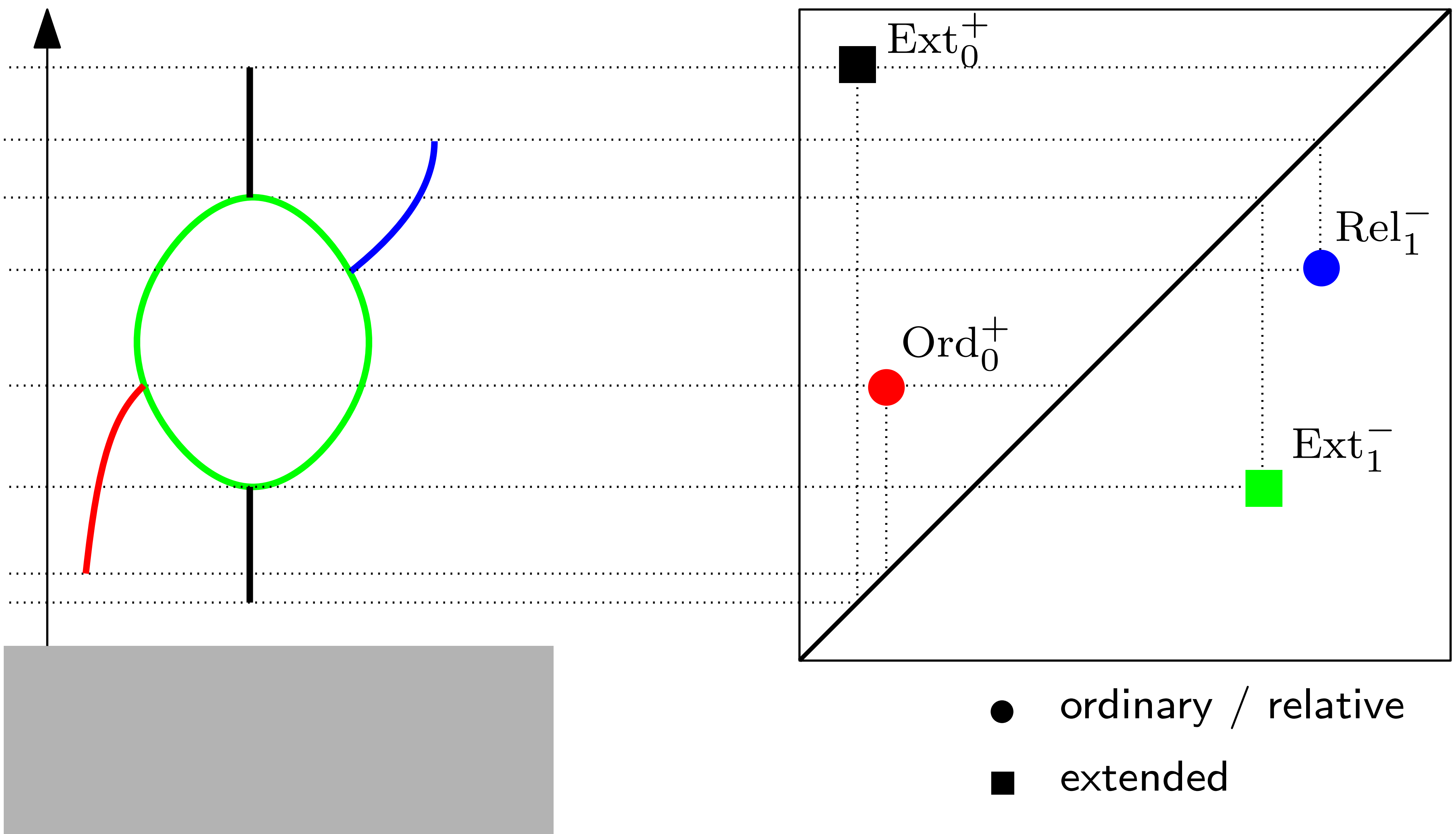




Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

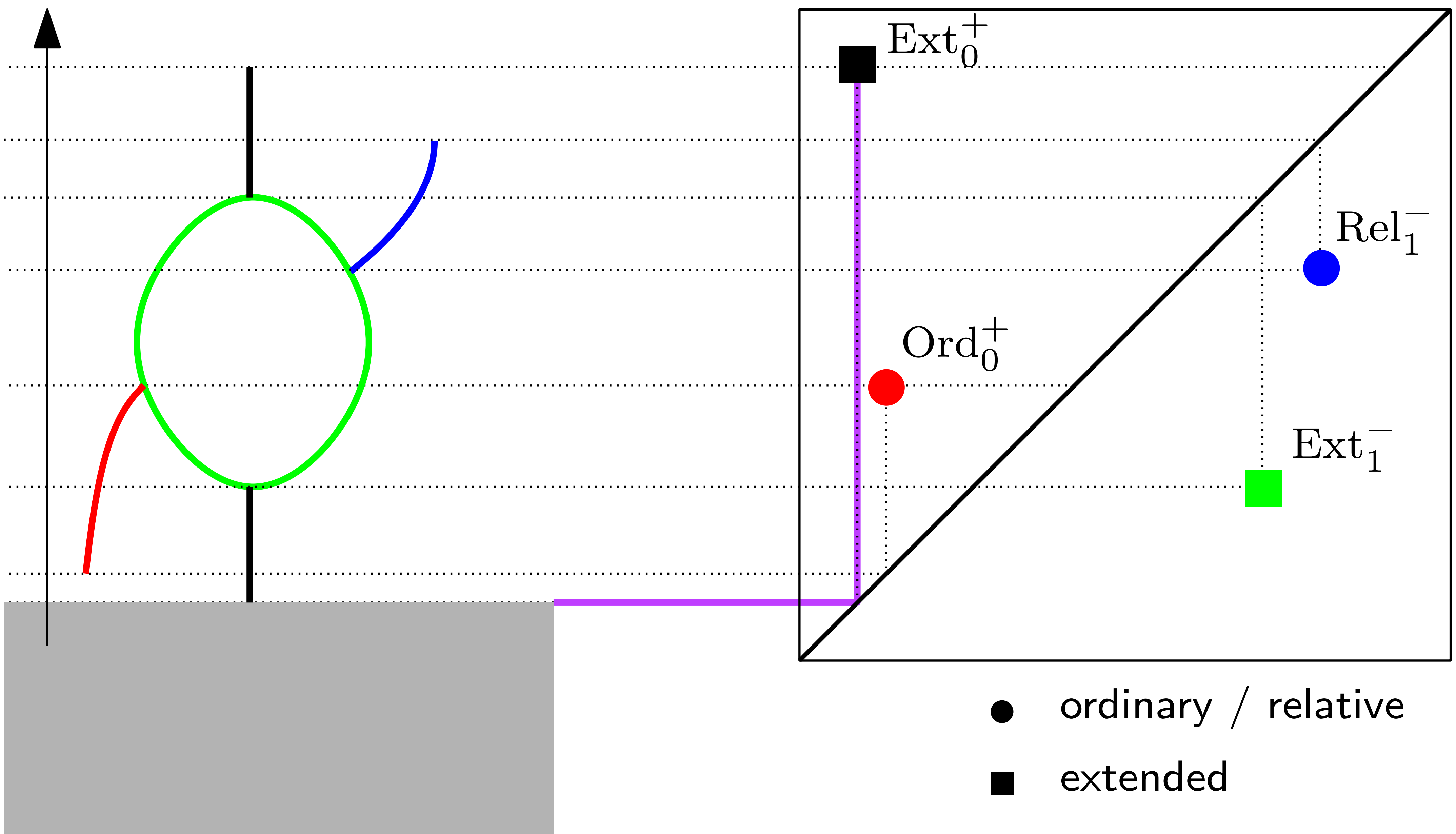
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

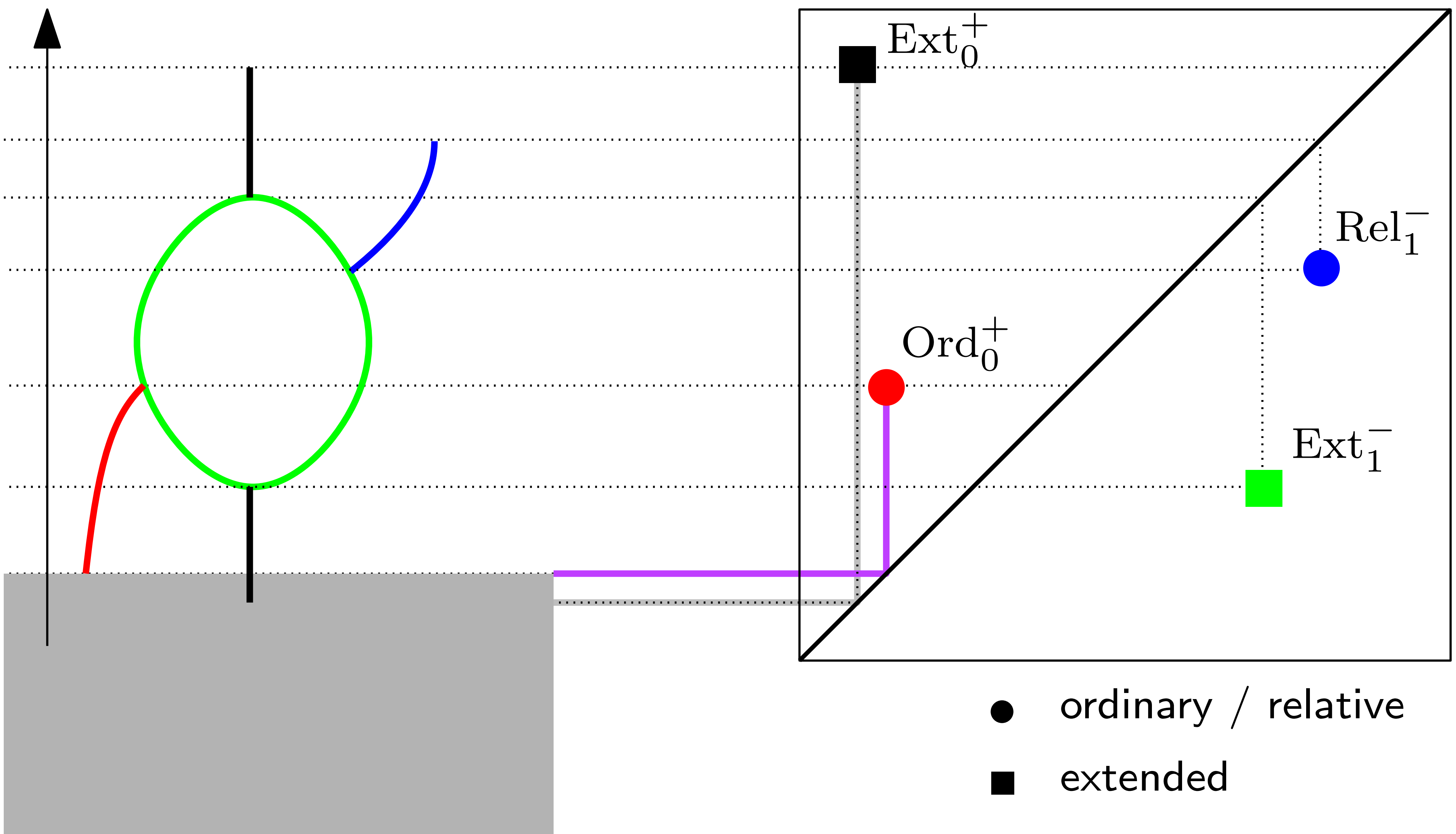
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

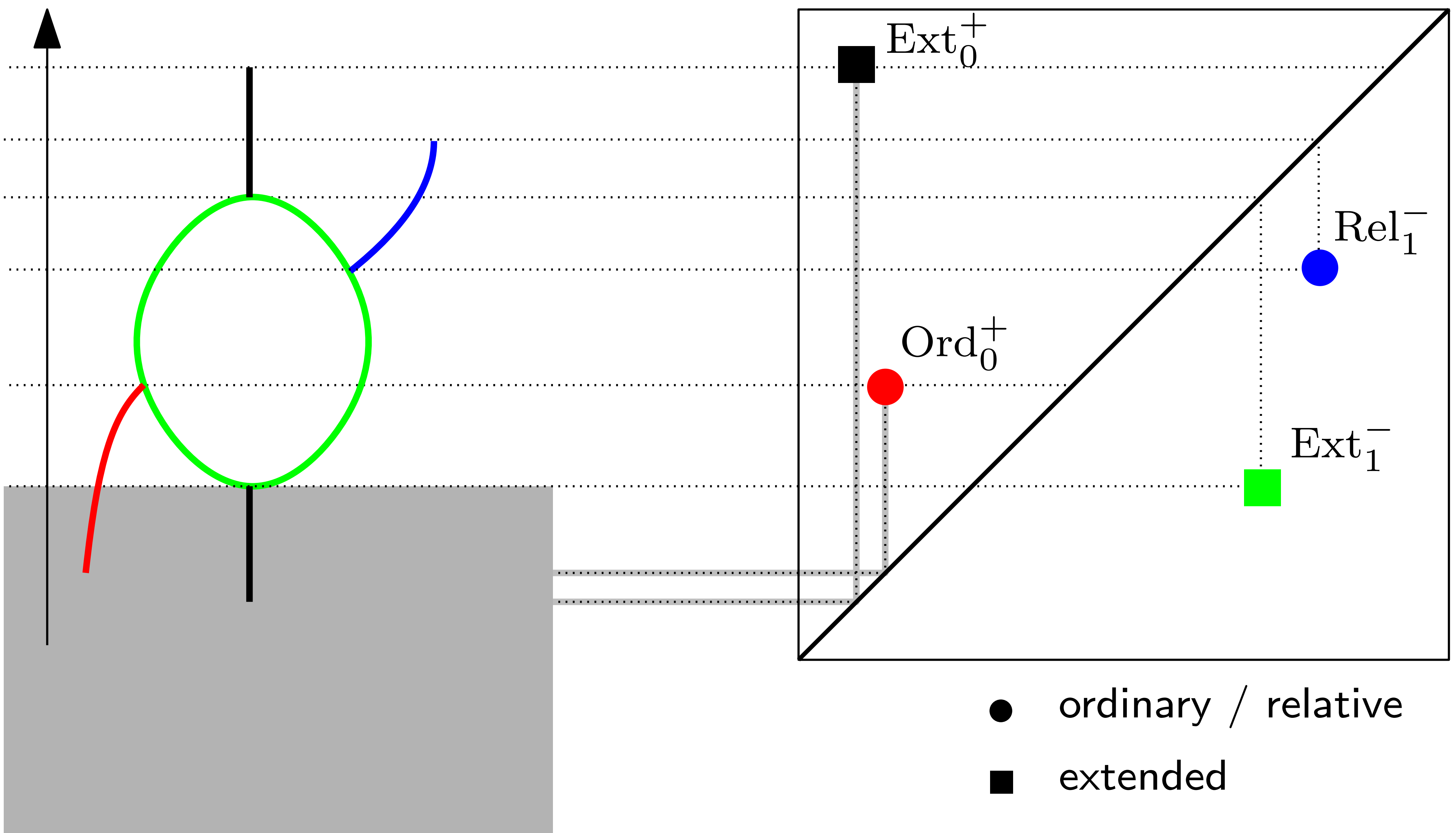
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

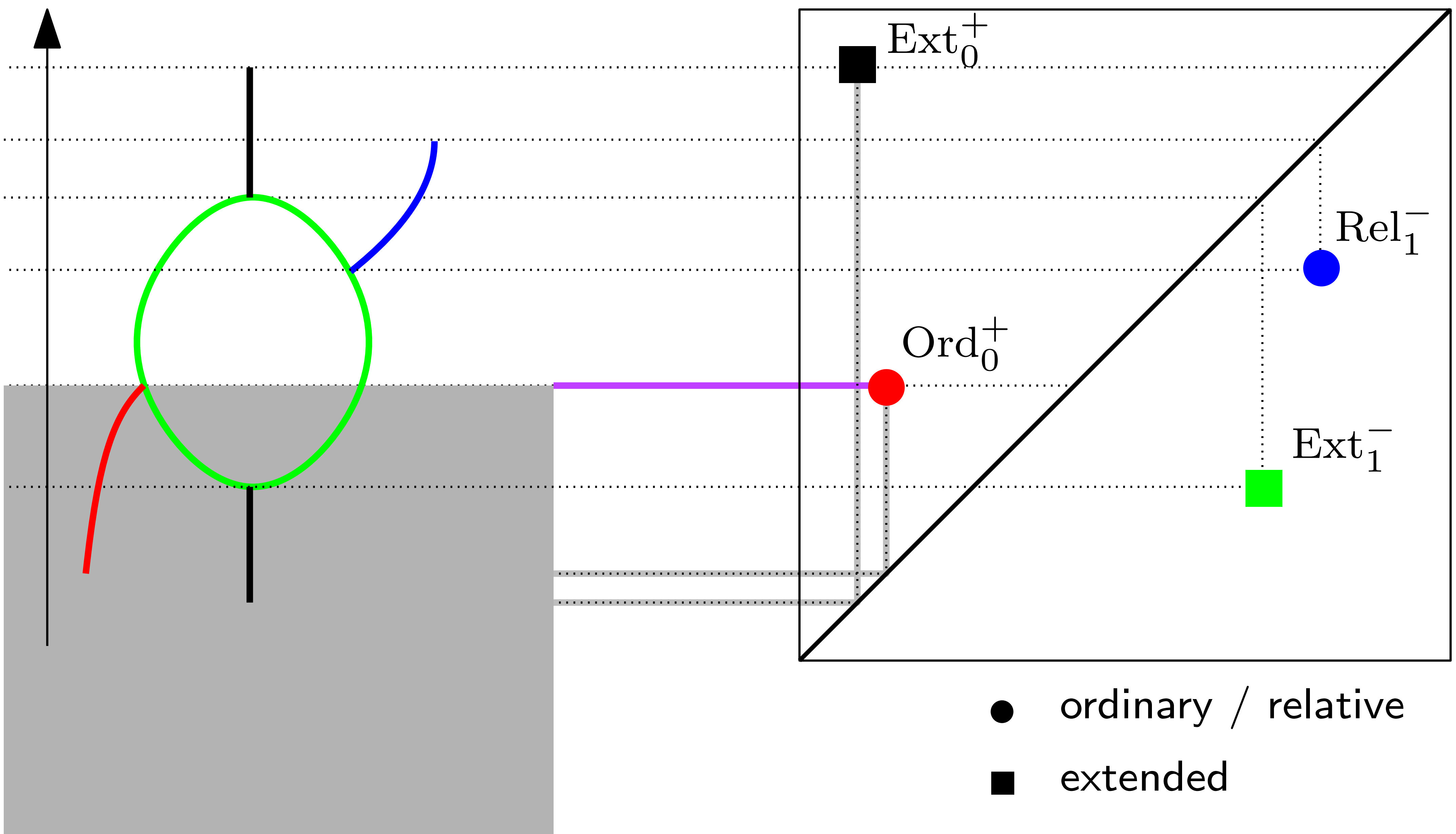
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

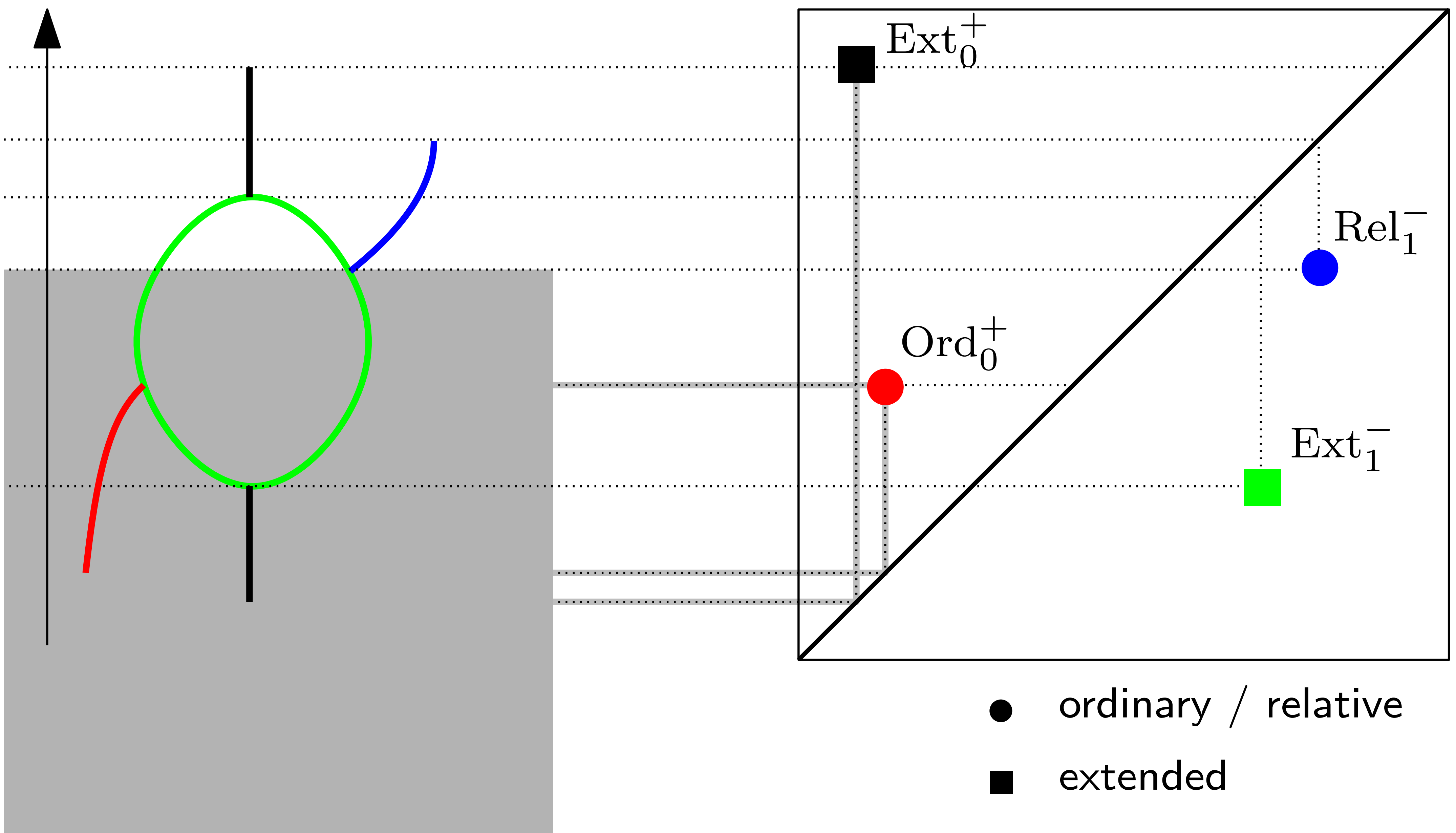
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

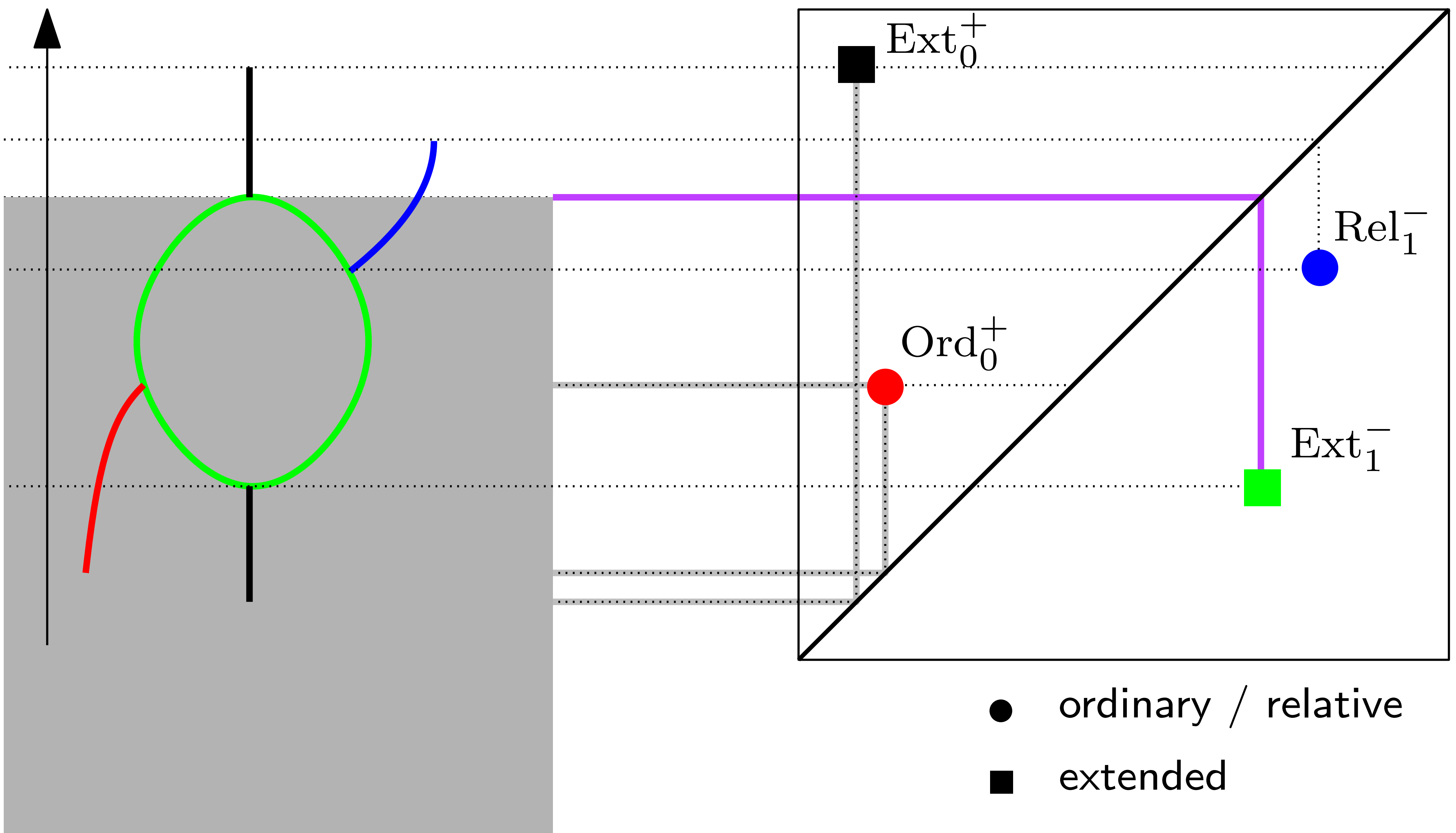
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

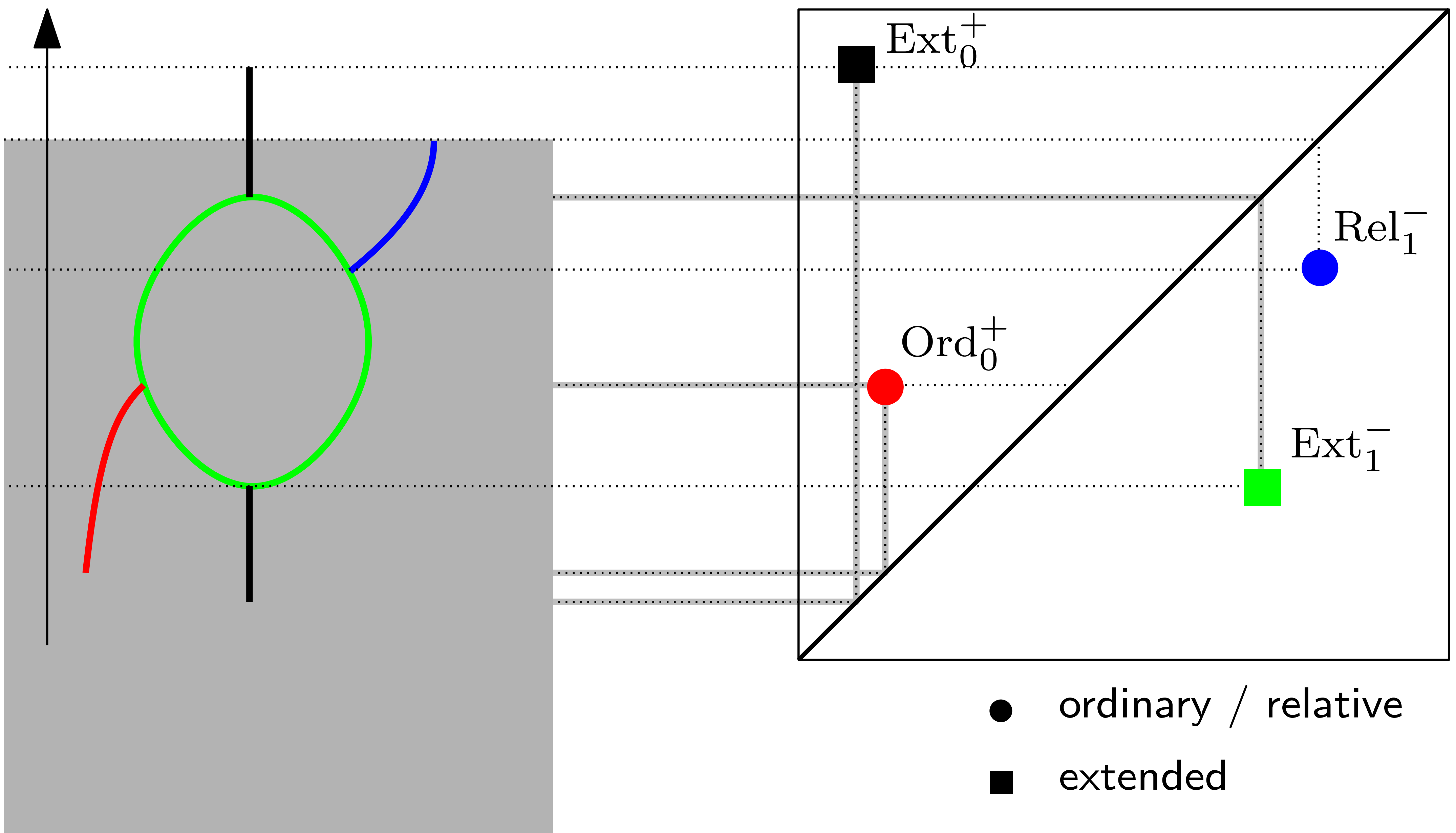
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

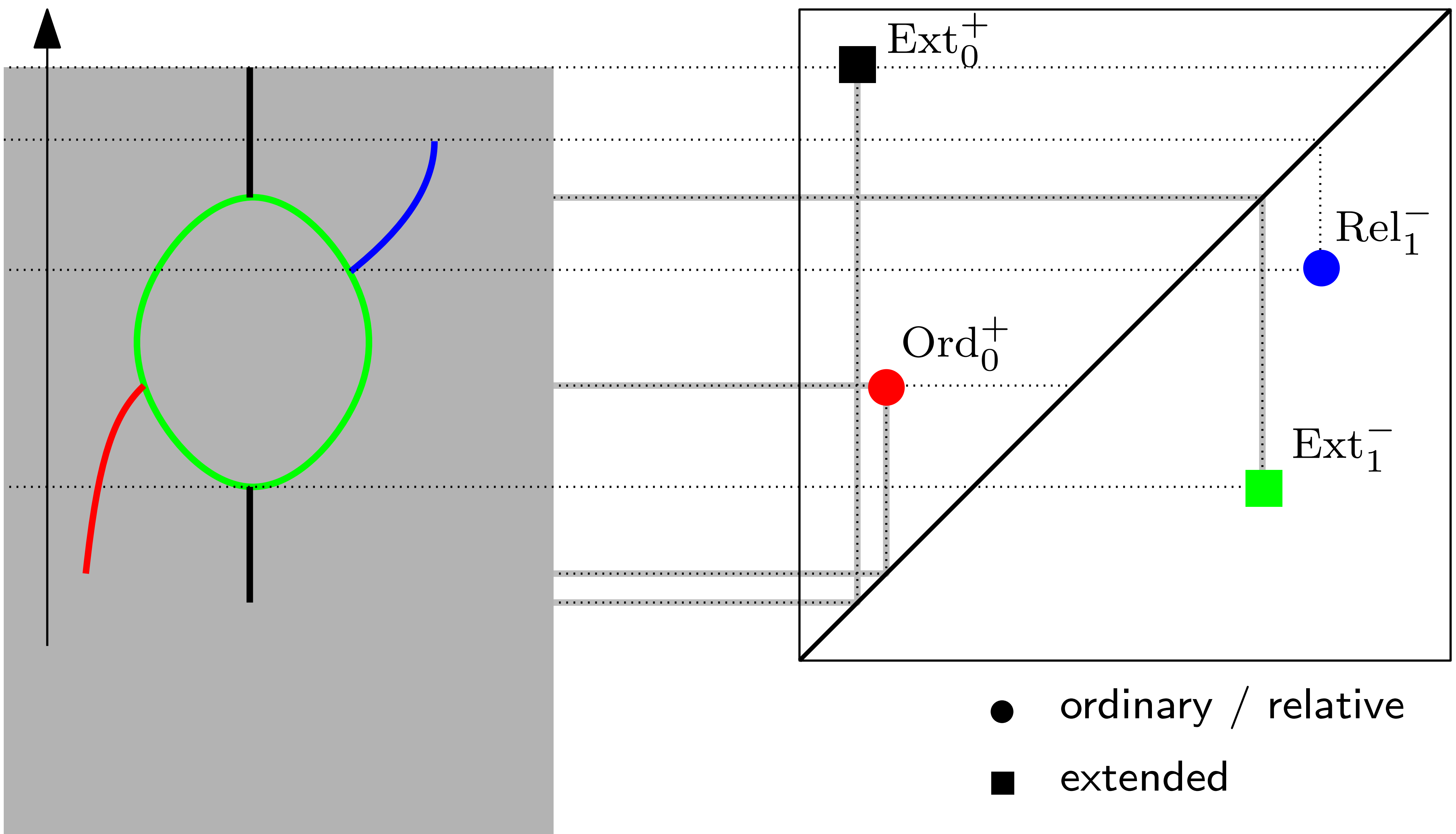
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

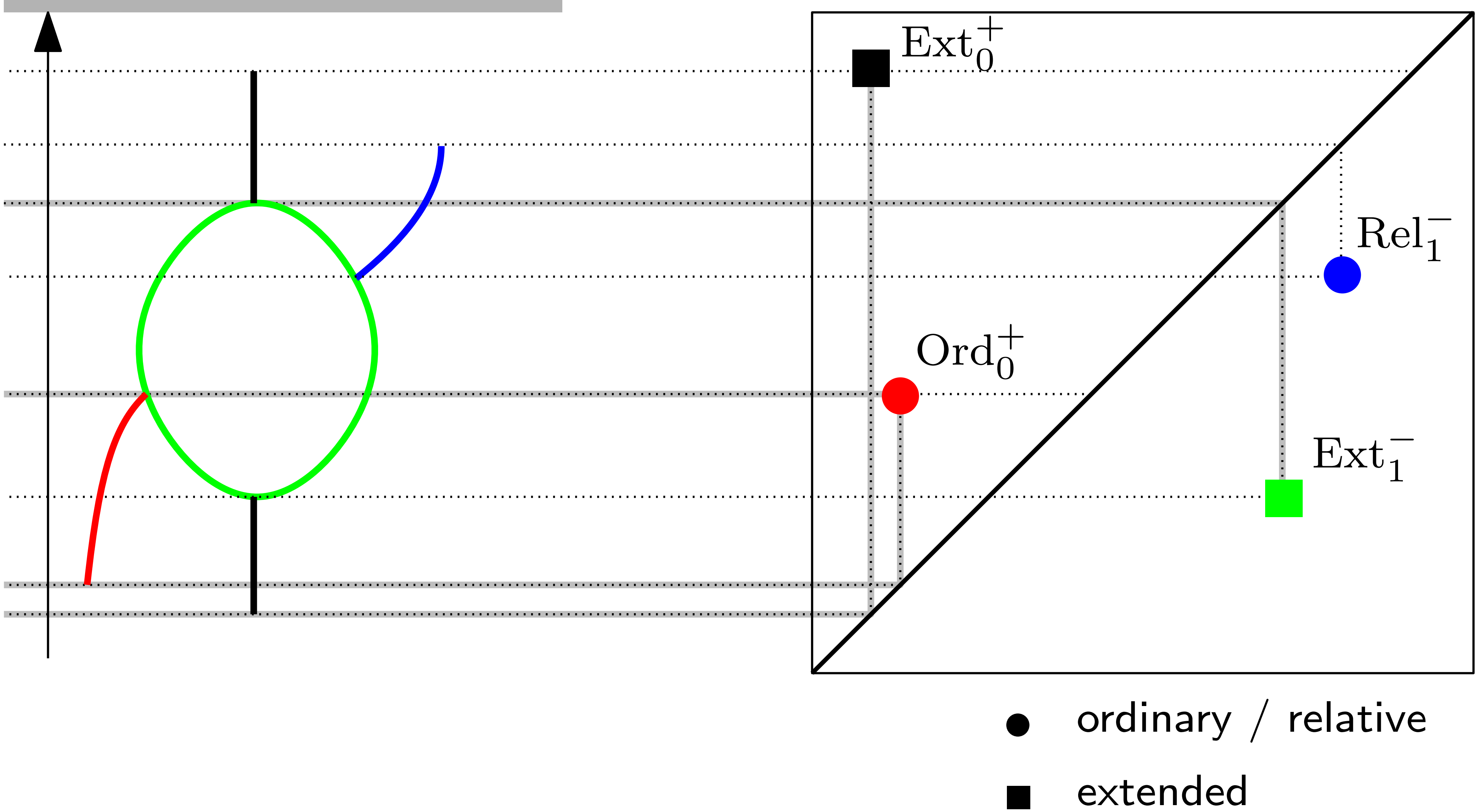
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

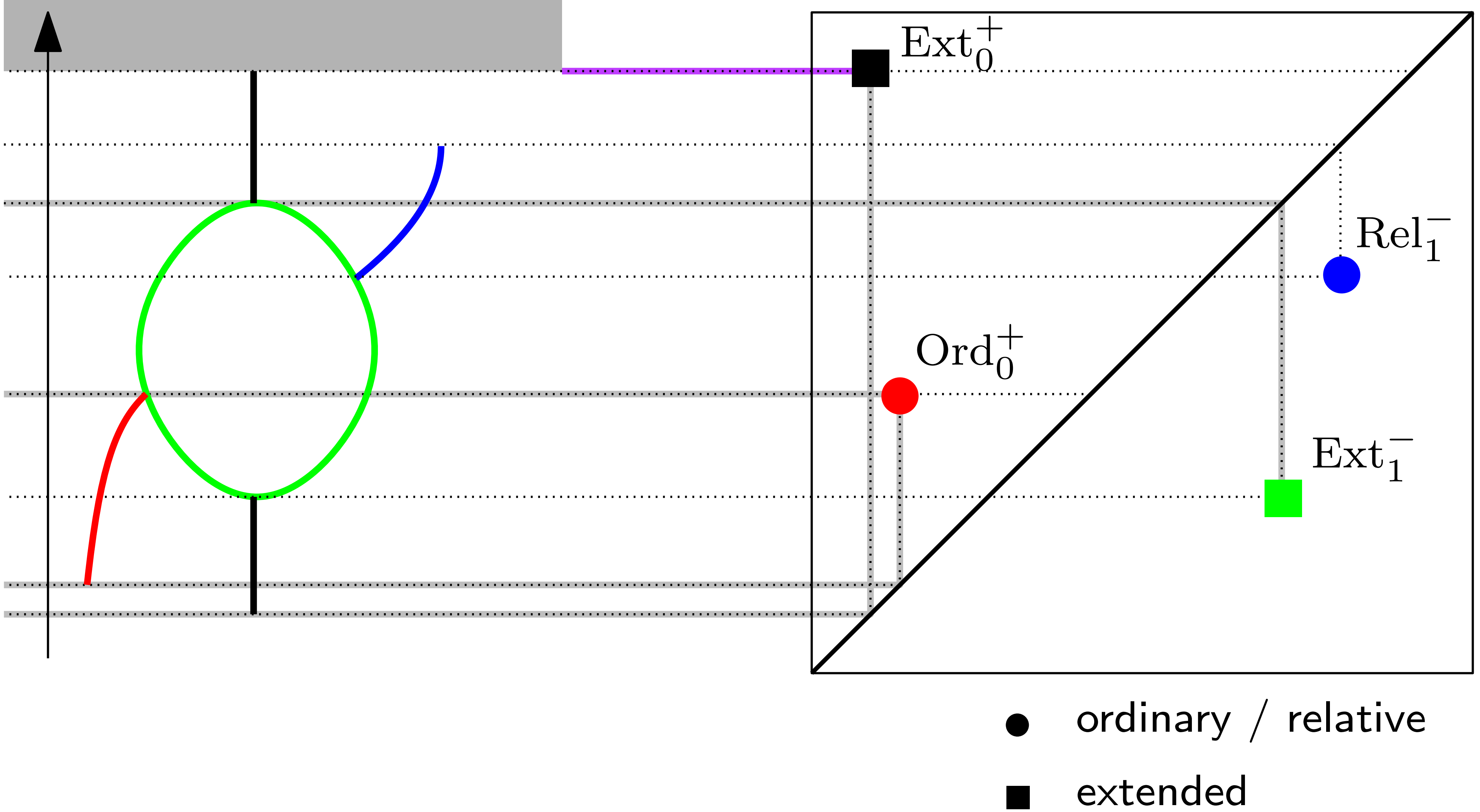
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

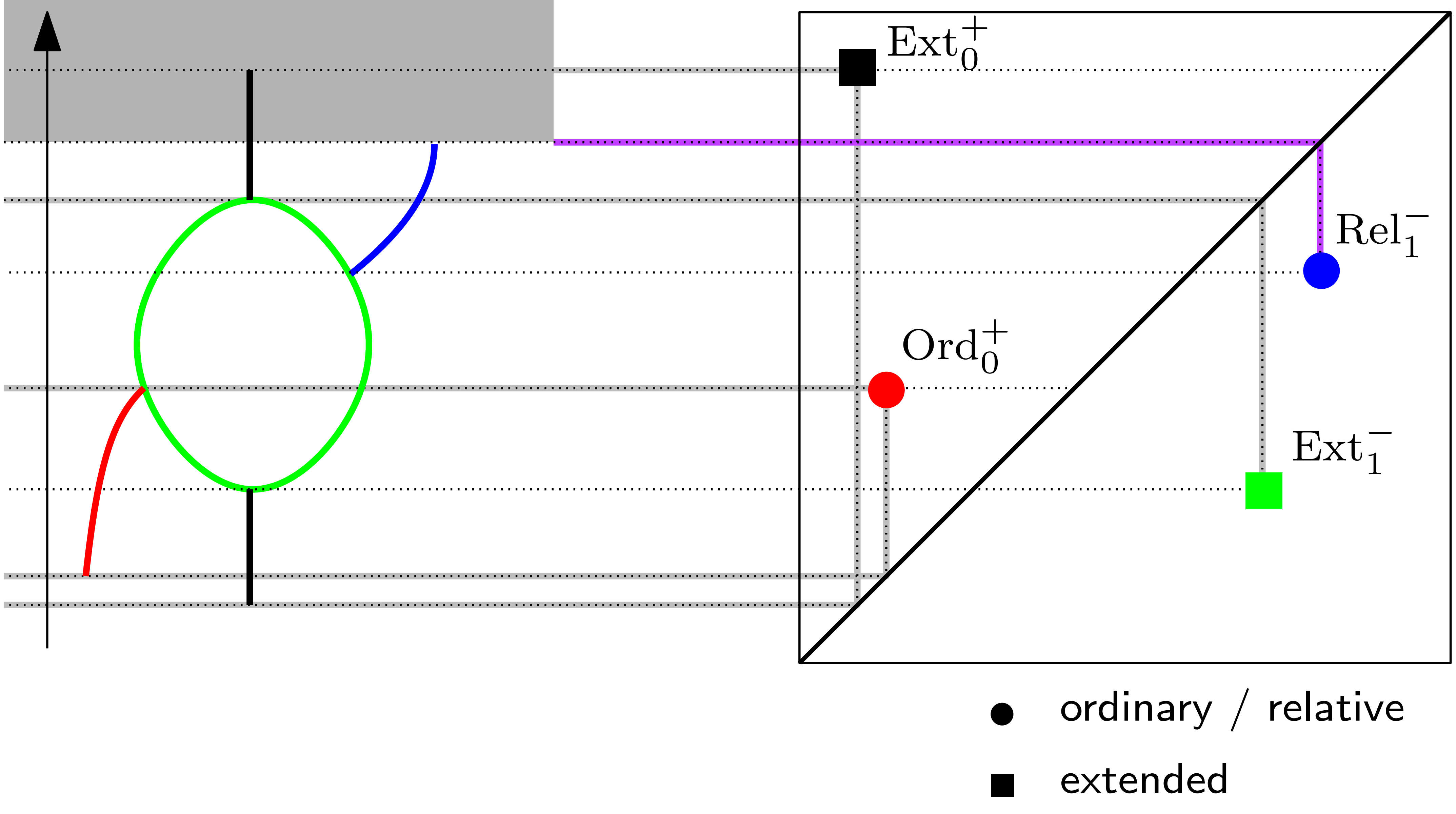
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

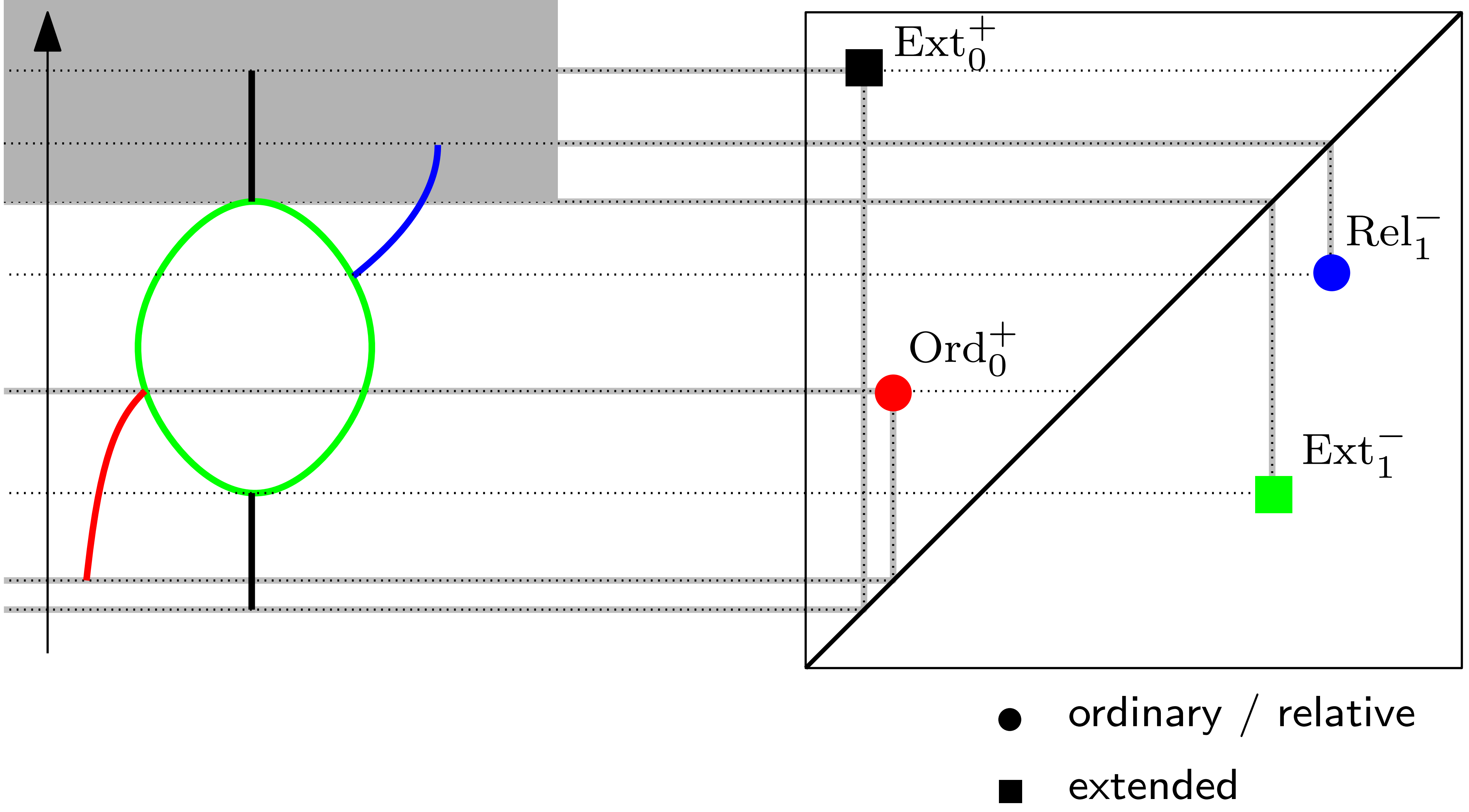
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

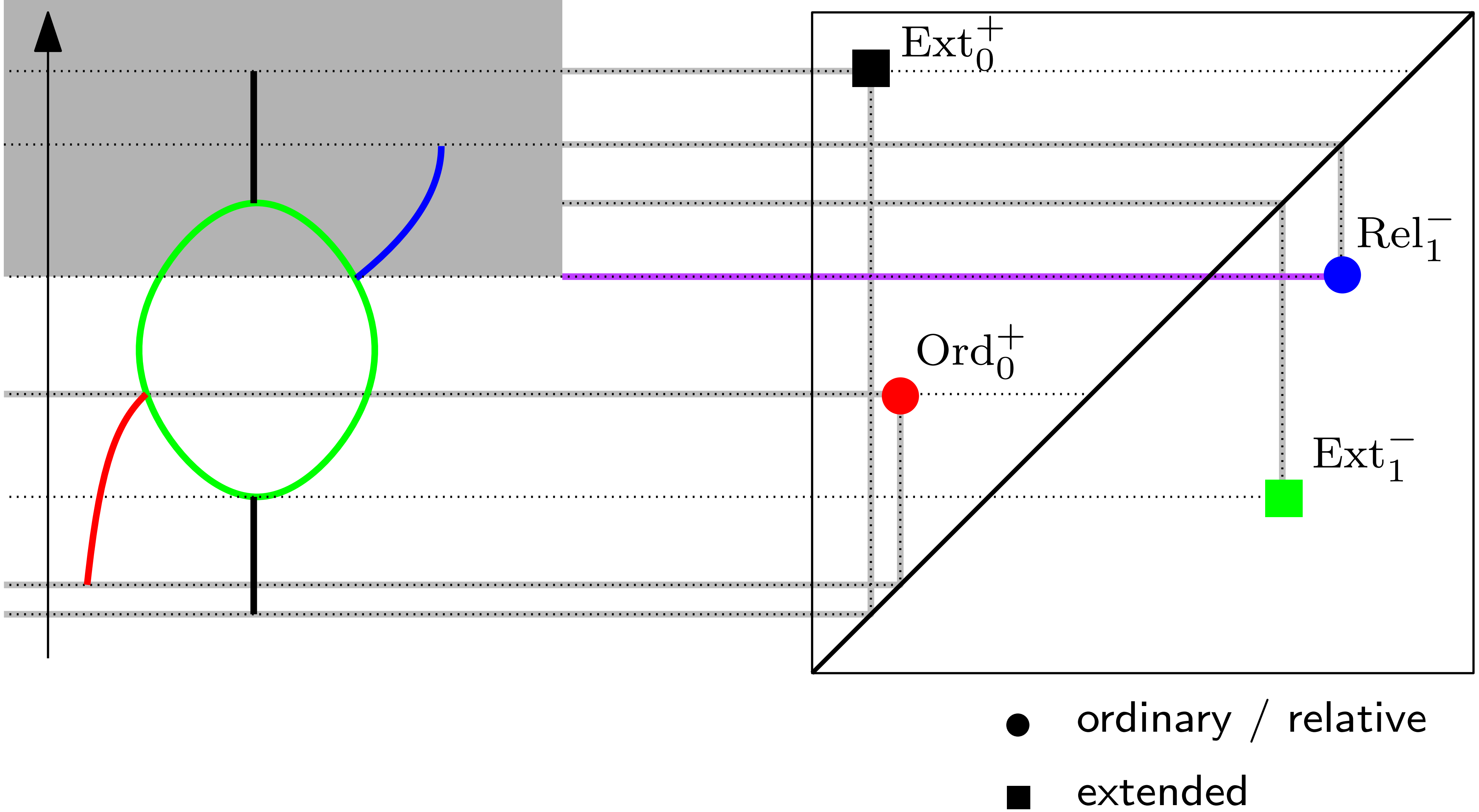
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

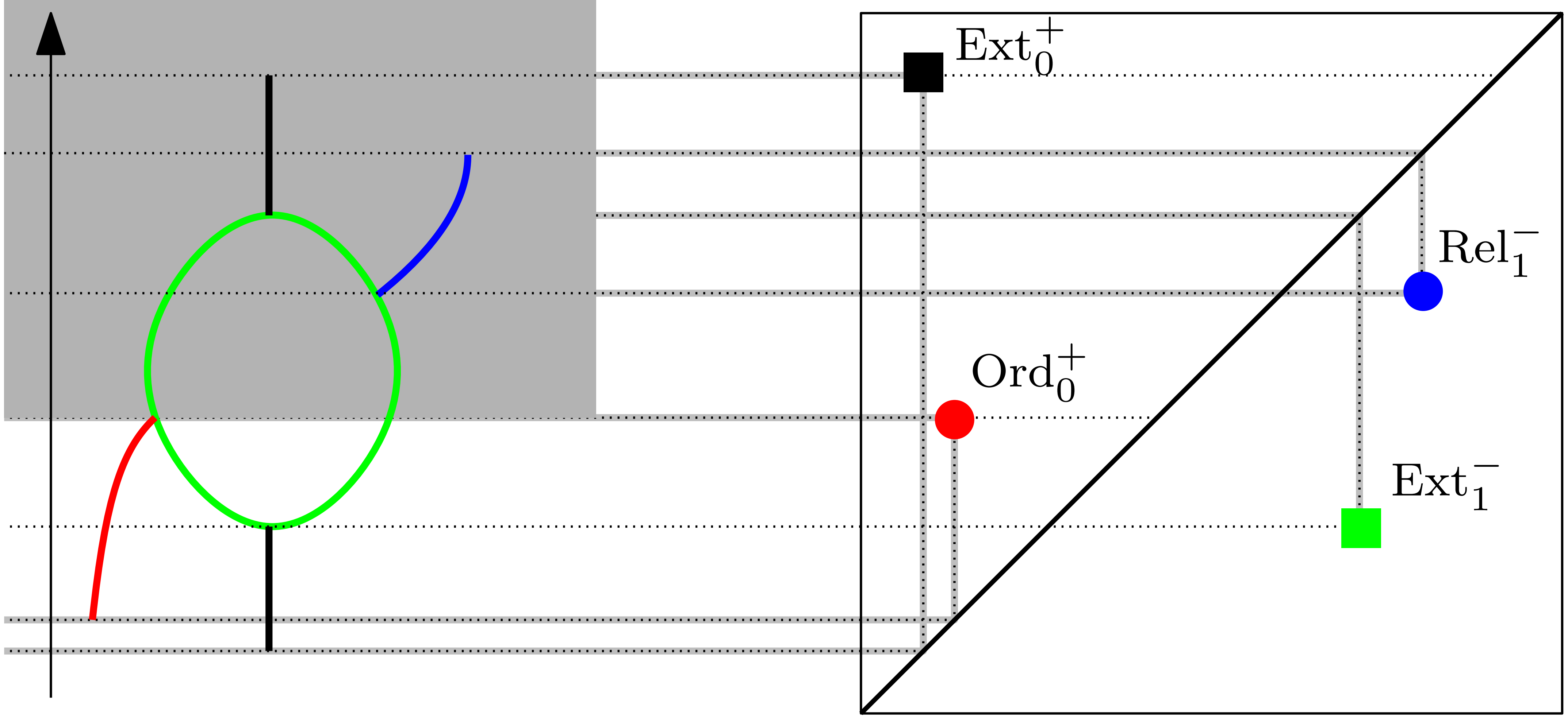
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



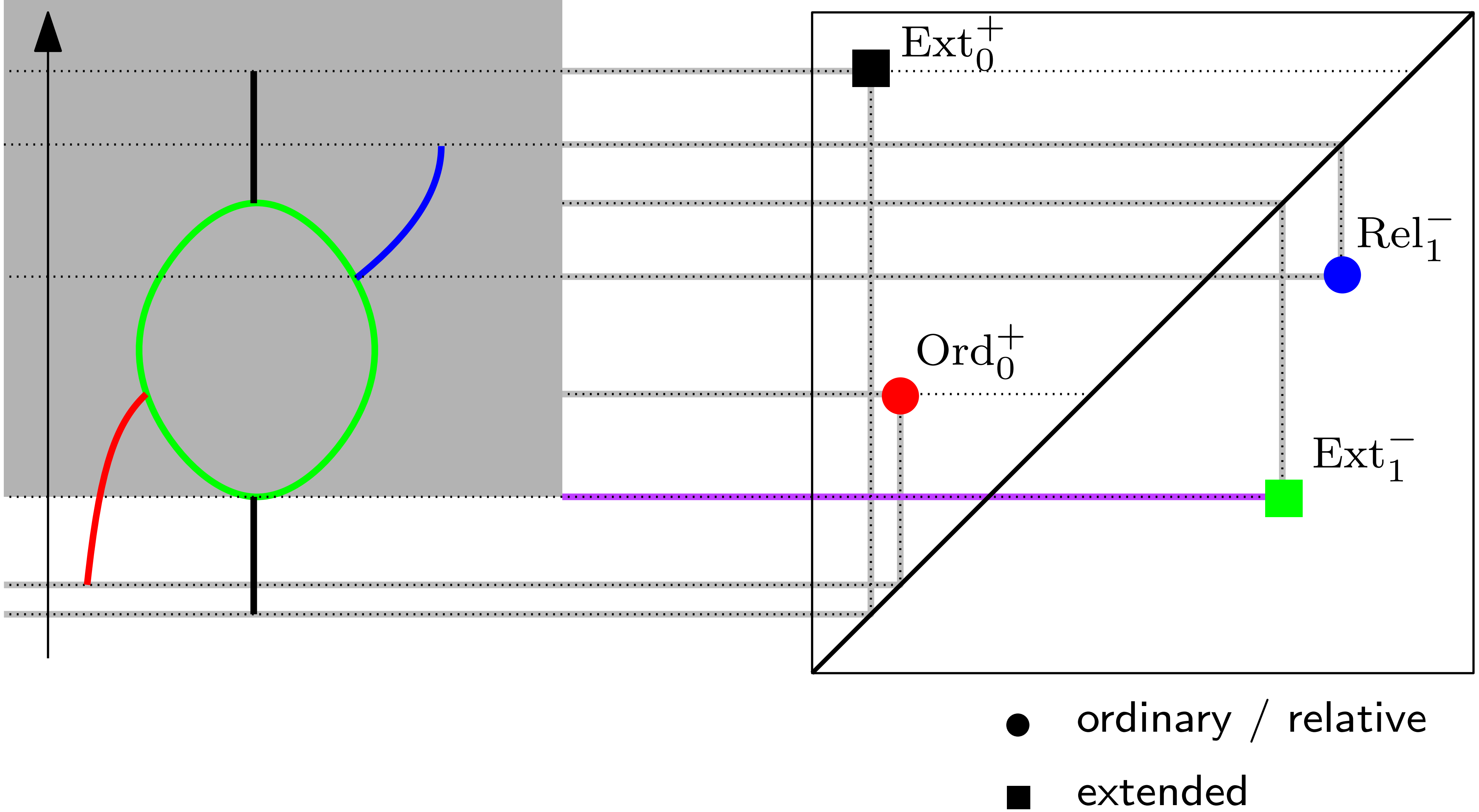
● ordinary / relative

■ extended

Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

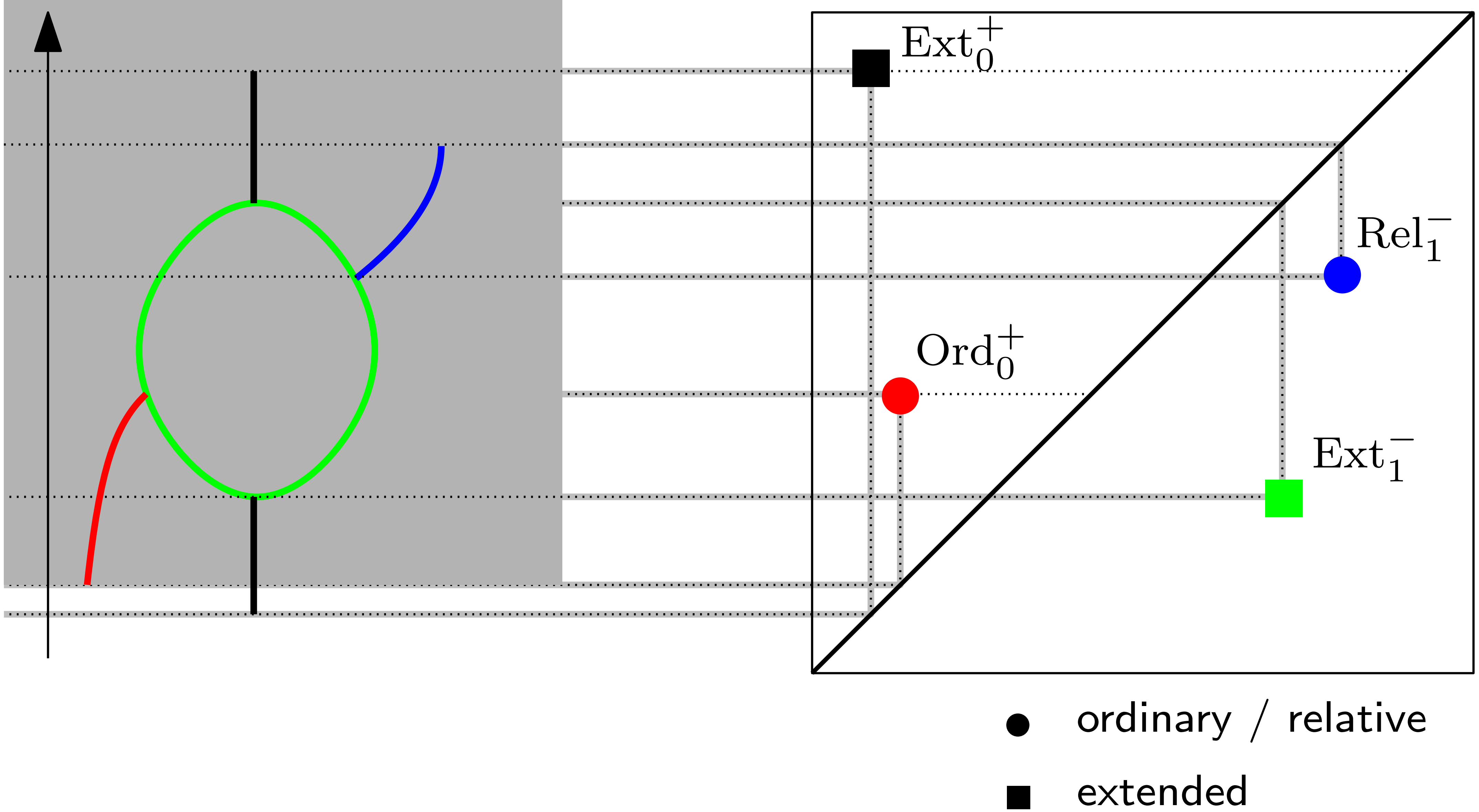
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

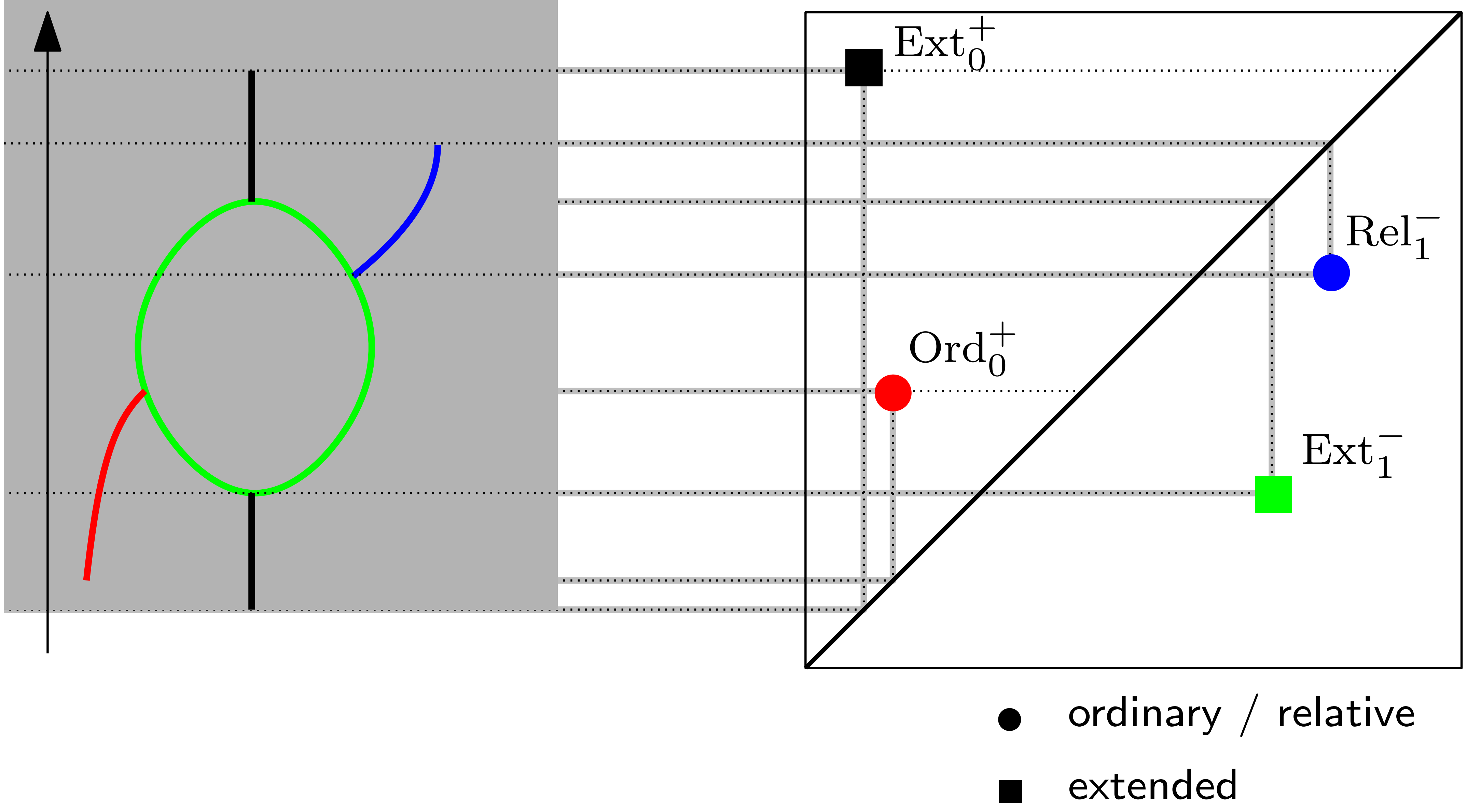
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

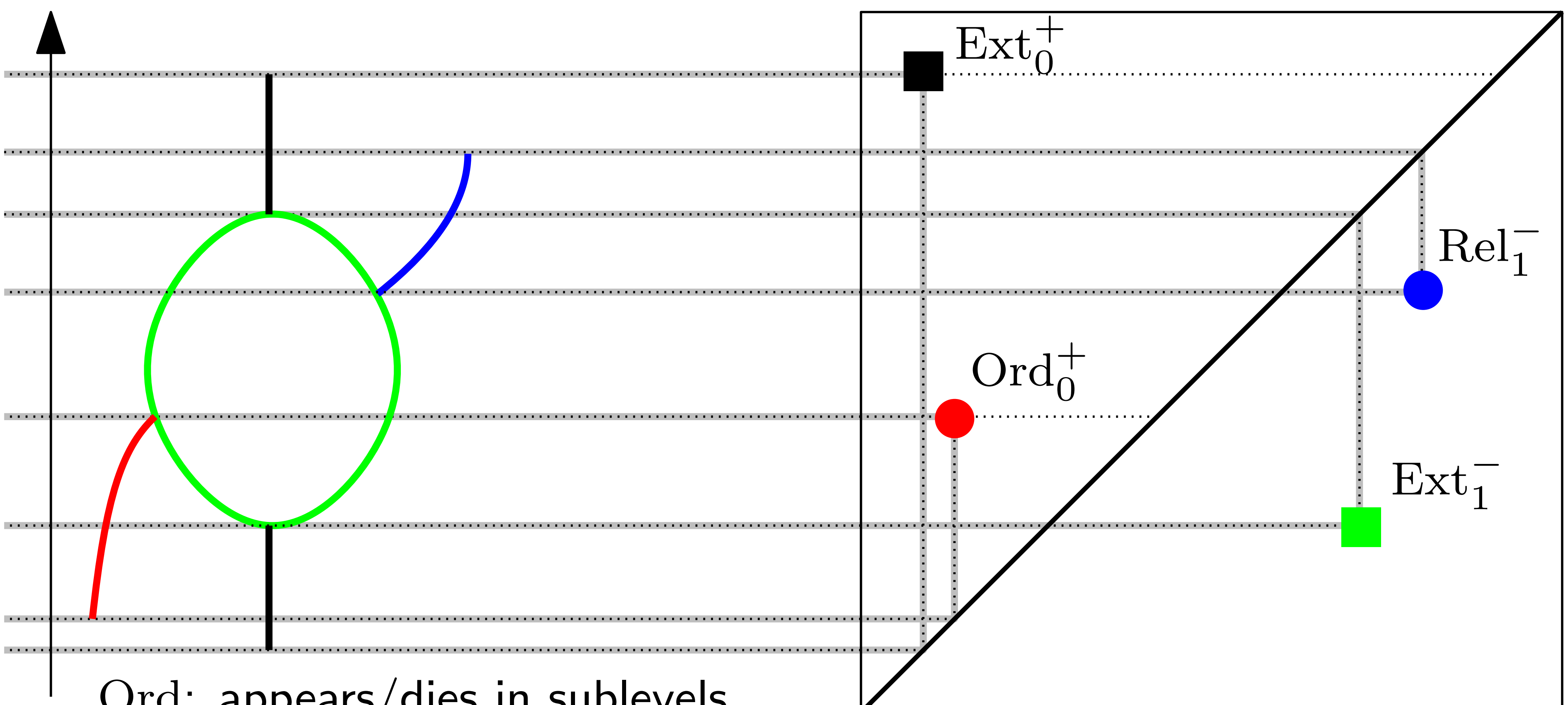
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Graph Descriptor

Construction uses **extended persistence**: [Cohen-Steiner, Edelsbrunner, Harer 2008]

- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Ord: appears/dies in sublevels

Rel: appears/dies in superlevels

Ext: appears in sublevels, dies in superlevels

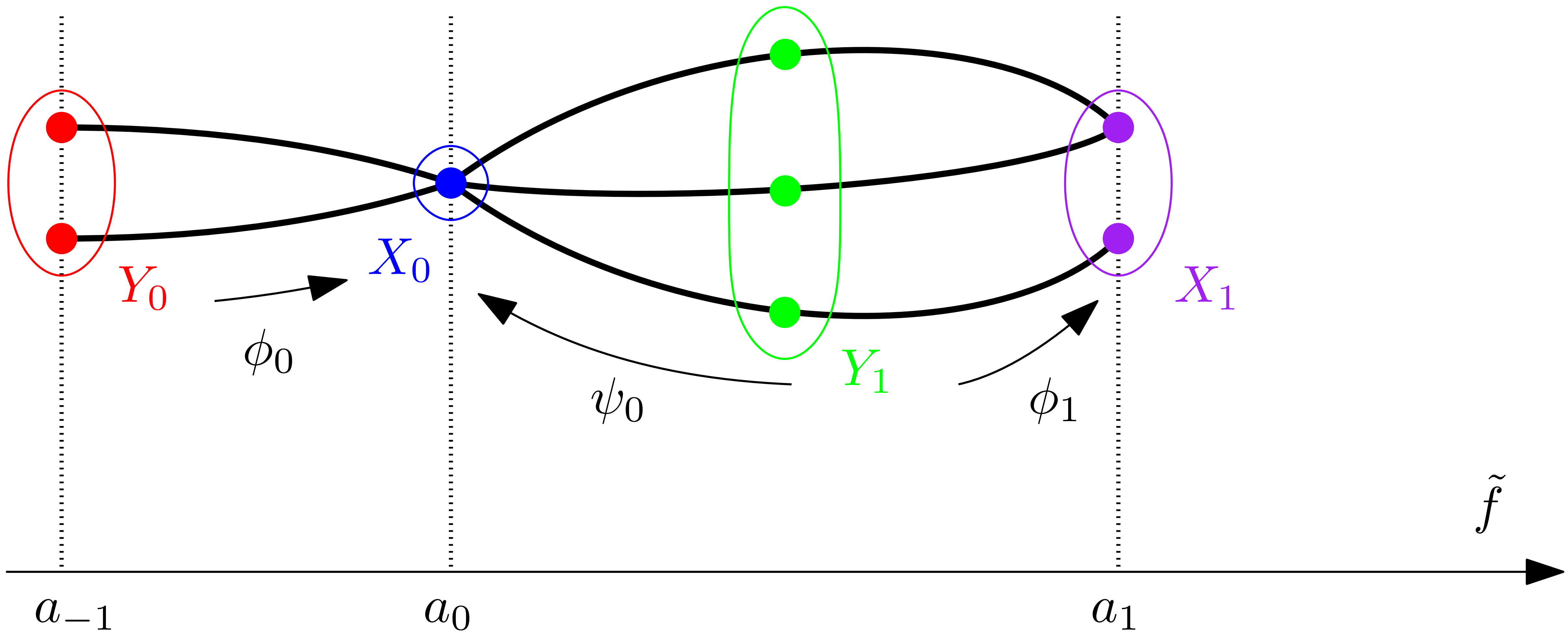
● ordinary / relative

■ extended

Graph Stratification

Reeb graph is a *telescope* (stratified space)

$$Y_0 \times [a_{-1}, a_0] \cup_{\psi_{-1}} X_0 \times \{a_0\} \cup_{\phi_0} Y_1 \times [a_0, a_1] \cup_{\psi_0} X_1 \times \{a_1\} \cup_{\phi_1} \ldots$$



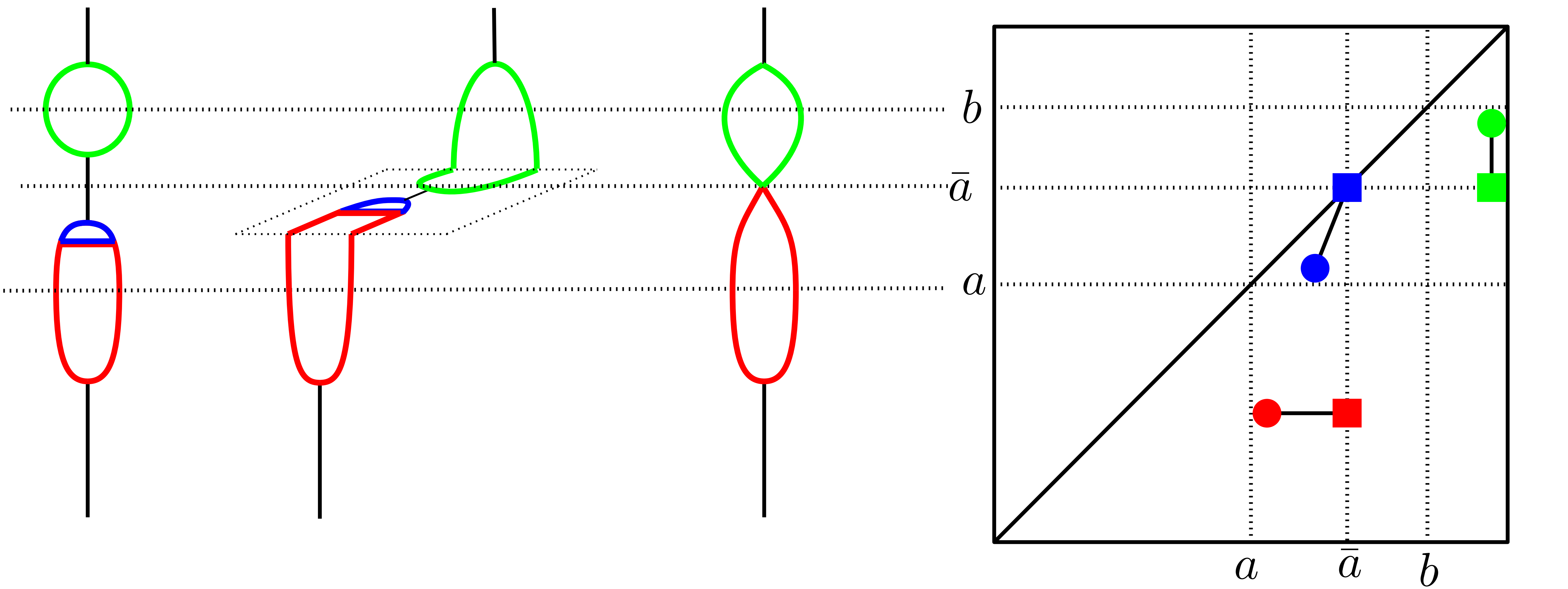
Idea: deform the Reeb graph so that it becomes the Mapper and track the changes in the persistence diagram

Operation 1: Merge $M_{a,b}$

$$(Y_{i-1} \times [a_{i-1}, a_i]) \cup_{\psi_{i-1}} (X_i \times \{a_i\}) \cup_{\phi_i} \ldots \cup_{\psi_{j-1}} (X_j \times \{a_j\}) \cup_{\phi_j} (Y_j \times [a_j, a_{j+1}])$$

$$\downarrow$$

$$(Y_{i-1} \times [a_{i-1}, \bar{a}]) \cup_{f_{i-1}} (\tilde{f}^{-1}([a,b]) \times \{\bar{a}\}) \cup_{g_j} (Y_j \times [\bar{a}, a_{j+1}])$$

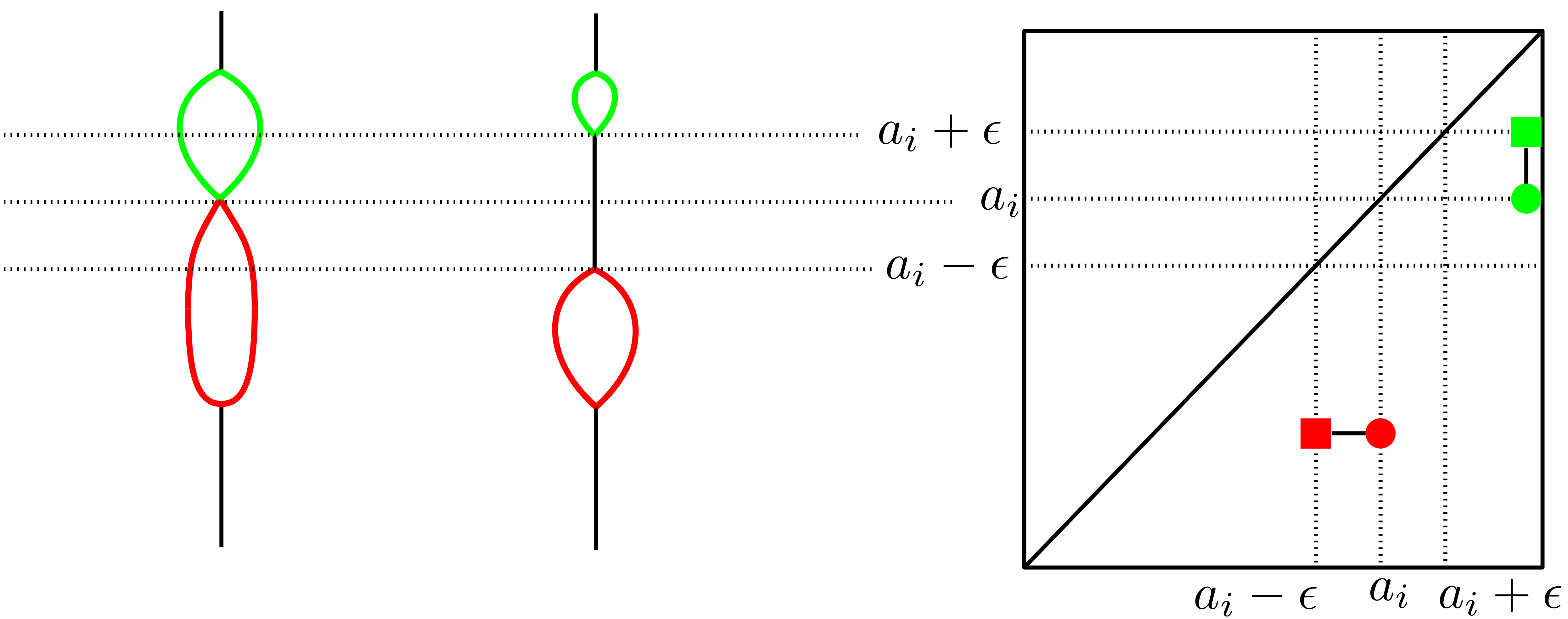


Operation 2: Split $Sp_{a_i,\epsilon}$

$$(Y_{i-1} \times [a_{i-1}, a_i]) \cup_{\psi_{i-1}} (X_i \times \{a_i\}) \cup_{\phi_i} (Y_i \times [a_i, a_{i+1}])$$

$$\downarrow$$

$$(Y_{i-1} \times [a_{i-1}, a_i - \epsilon]) \cup_{\psi_{i-1}^{a_i-\epsilon}} (X_i \times \{a_i - \epsilon\}) \cup_{\mathrm{id}} (X_i \times [a_i - \epsilon, a_i + \epsilon]) \cup_{\mathrm{id}} (X_i \times \{a_i + \epsilon\}) \cup_{\phi_i^{a_i+\epsilon}} (Y_i \times [a_i + \epsilon, a_{i+1}])$$

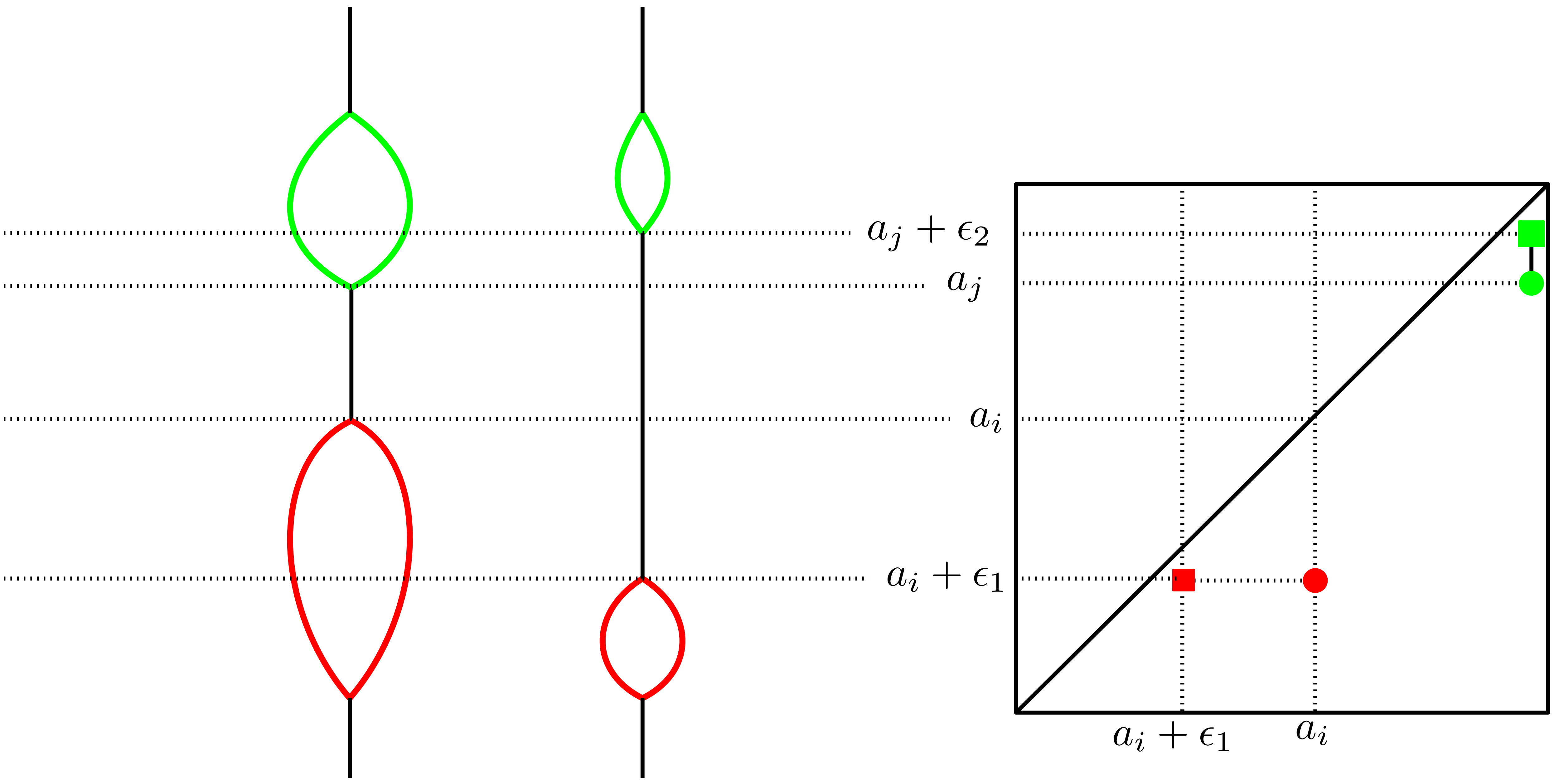


Operation 3: Shift $Sh_{a_i,\epsilon}$

$$(Y_{i-1} \times [a_{i-1}, a_i]) \cup_{\psi_{i-1}} (X_i \times \{a_i\}) \cup_{\phi_i} (Y_i \times [a_i, a_{i+1}])$$

$$\downarrow$$

$$(Y_{i-1} \times [a_{i-1}, a_i + \epsilon]) \cup_{\psi_{i-1}} (X_i \times \{a_i + \epsilon\}) \cup_{\phi_i} (Y_i \times [a_i + \epsilon, a_{i+1}])$$



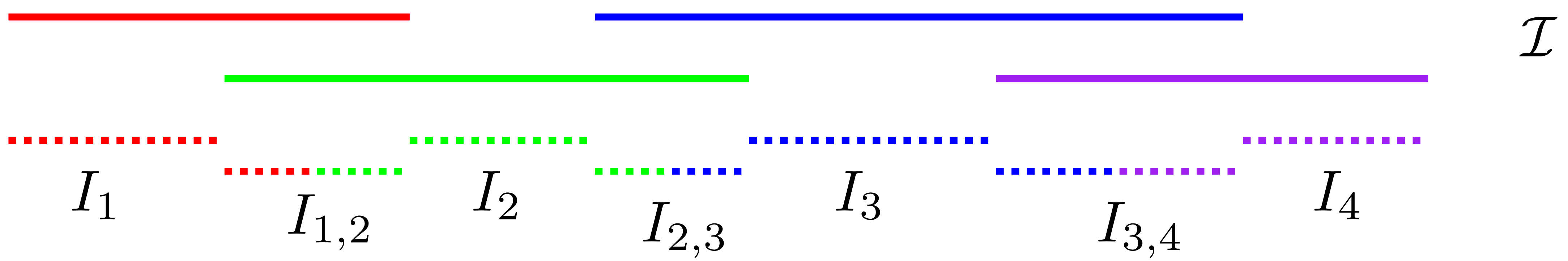
Formula Reeb graph $\to$ Mapper

Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

Formula Reeb graph $\to$ Mapper

Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

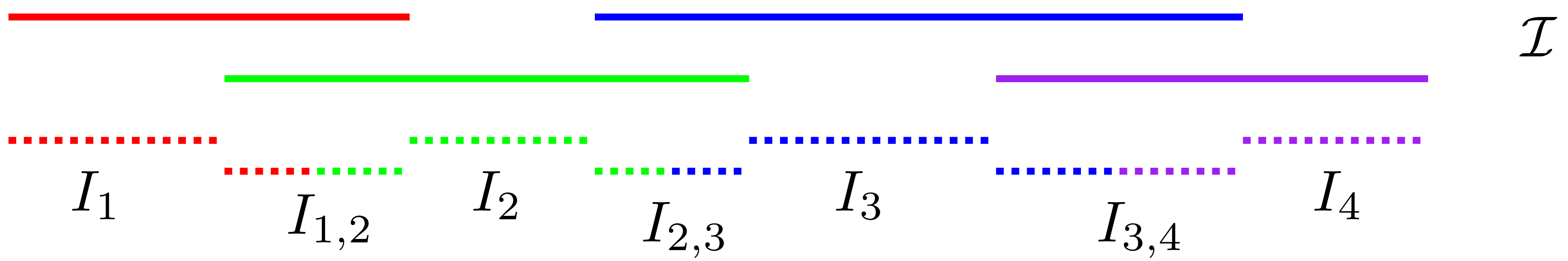
- $M_{\mathcal{I}}$ is the union of all M_{I_k} and $M_{I_{k,k+1}}$ for $I \in \mathcal{I}$



Formula Reeb graph $\to$ Mapper

Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

- $M_{\mathcal{I}}$ is the union of all M_{I_k} and $M_{I_{k,k+1}}$ for $I \in \mathcal{I}$

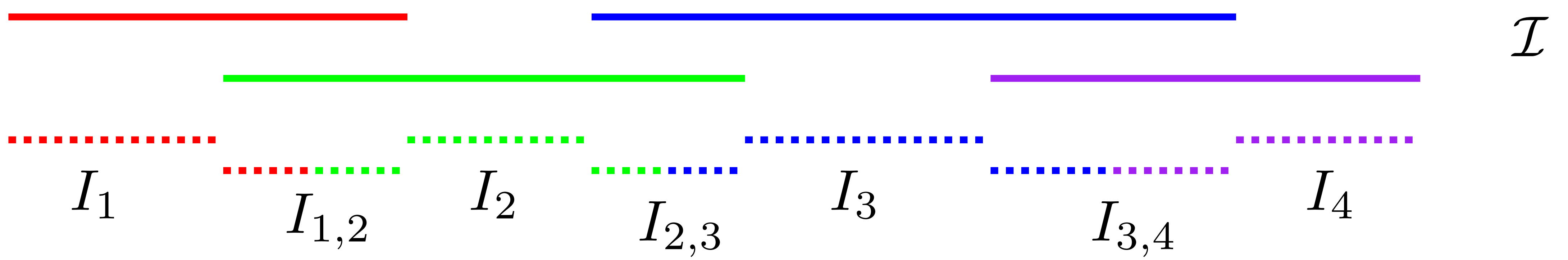


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small

Formula Reeb graph $\rightarrow$ Mapper

Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

- $M_{\mathcal{I}}$ is the union of all M_{I_k} and $M_{I_{k,k+1}}$ for $I \in \mathcal{I}$

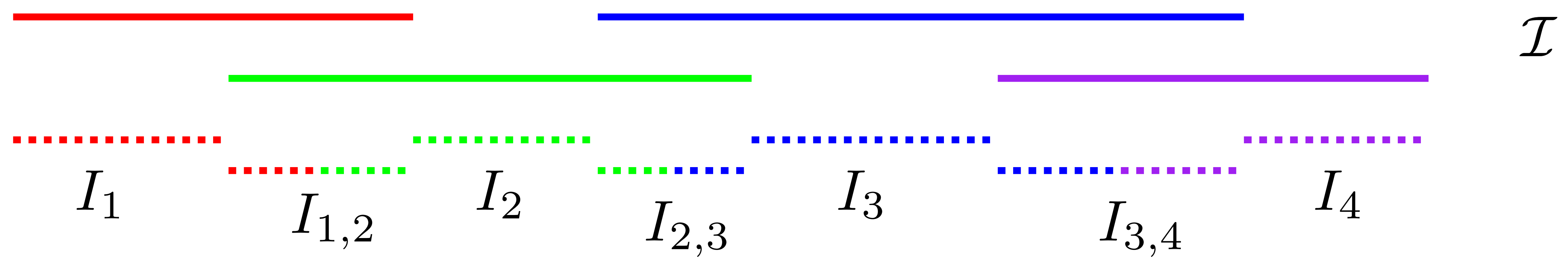


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small
- $Sh_{\mathcal{I}}$ is the union of all $Sh_{\epsilon_1,\bar{a}+\epsilon}$ and $Sh_{\epsilon_2,\bar{a}-\epsilon}$ with ϵ_1, ϵ_2 small

Formula Reeb graph $\to$ Mapper

Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

- $M_{\mathcal{I}}$ is the union of all M_{I_k} and $M_{I_{k,k+1}}$ for $I \in \mathcal{I}$

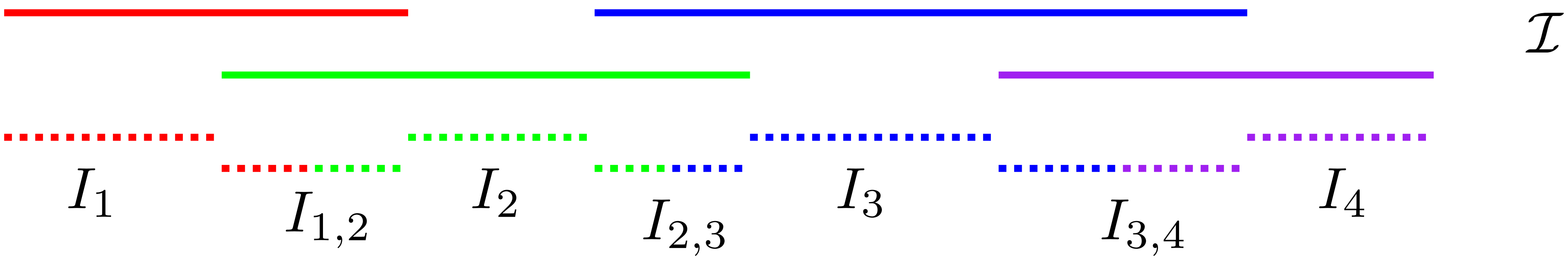


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small
- $Sh_{\mathcal{I}}$ is the union of all $Sh_{\epsilon_1,\bar{a}+\epsilon}$ and $Sh_{\epsilon_2,\bar{a}-\epsilon}$ with ϵ_1, ϵ_2 small
- $M'_{\mathcal{I}}$ is the union of all M_{I_k} for $I \in \mathcal{I}$

Formula Reeb graph → Mapper

Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

- $M_{\mathcal{I}}$ is the union of all M_{I_k} and $M_{I_{k,k+1}}$ for $I \in \mathcal{I}$

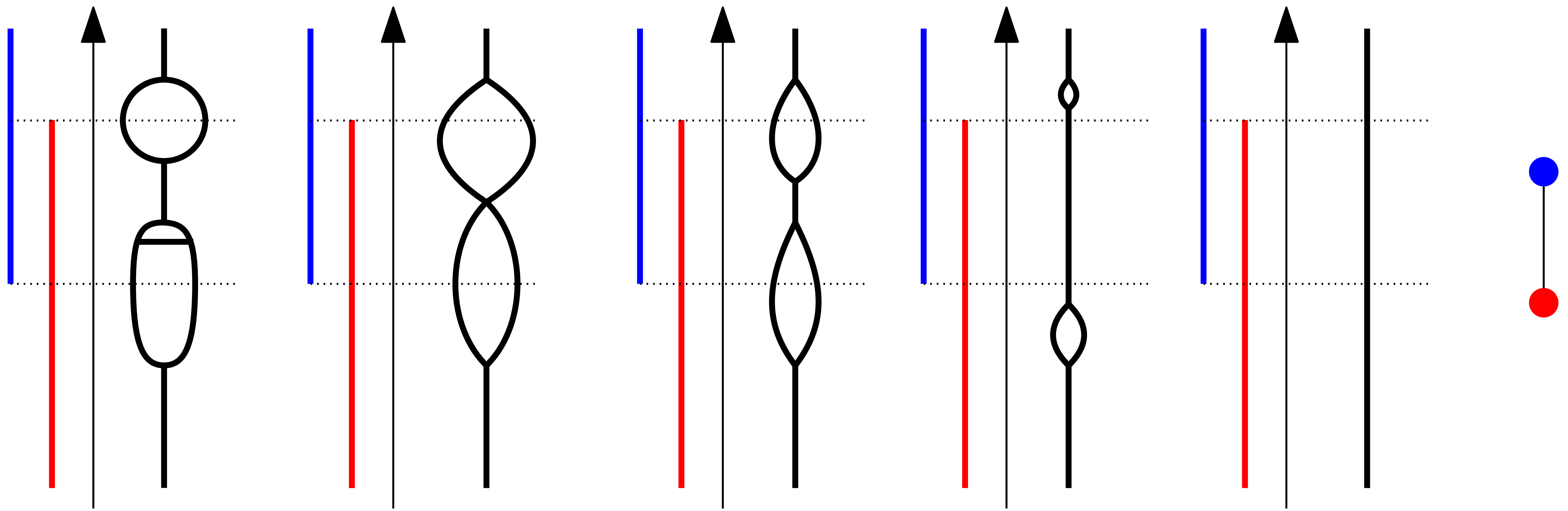


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small
- $Sh_{\mathcal{I}}$ is the union of all $Sh_{\epsilon_1,\bar{a}+\epsilon}$ and $Sh_{\epsilon_2,\bar{a}-\epsilon}$ with ϵ_1, ϵ_2 small
- $M'_{\mathcal{I}}$ is the union of all M_{I_k} for $I \in \mathcal{I}$

$$\mathrm{M}_f(X, \mathcal{I}) = M'_{\mathcal{I}} \circ Sh_{\mathcal{I}} \circ Sp_{\mathcal{I}} \circ M_{\mathcal{I}}(\mathrm{R}_f(X))$$

Formula Reeb graph → Mapper

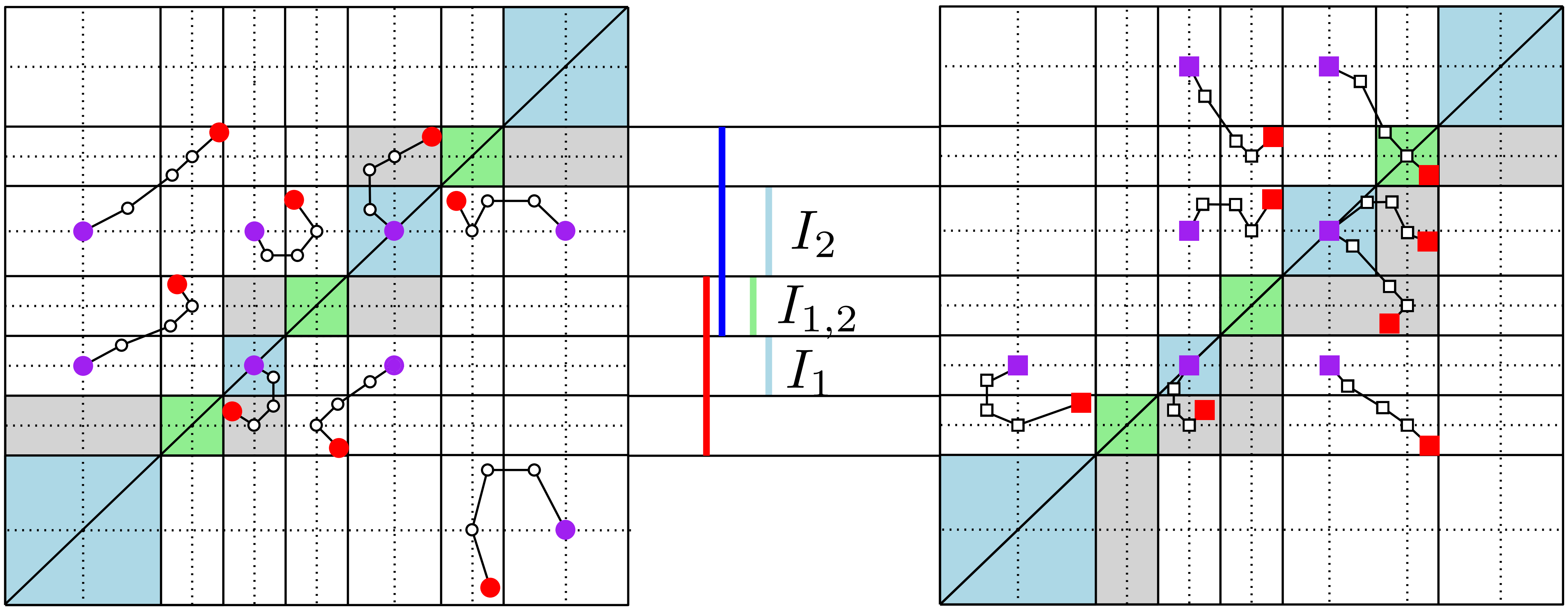
Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$



$$\mathrm{M}_f(X, \mathcal{I}) = M'_{\mathcal{I}} \circ Sh_{\mathcal{I}} \circ Sp_{\mathcal{I}} \circ M_{\mathcal{I}}(\mathrm{R}_f(X))$$

Formula Reeb graph → Mapper

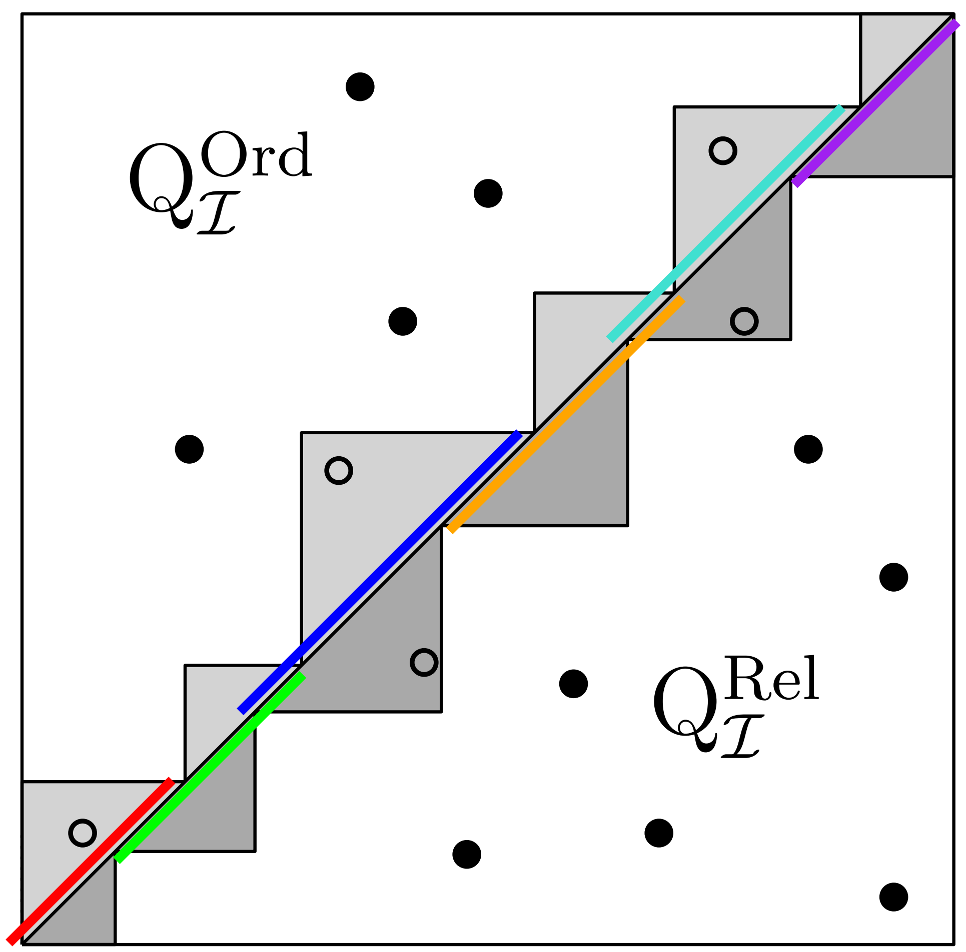
Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

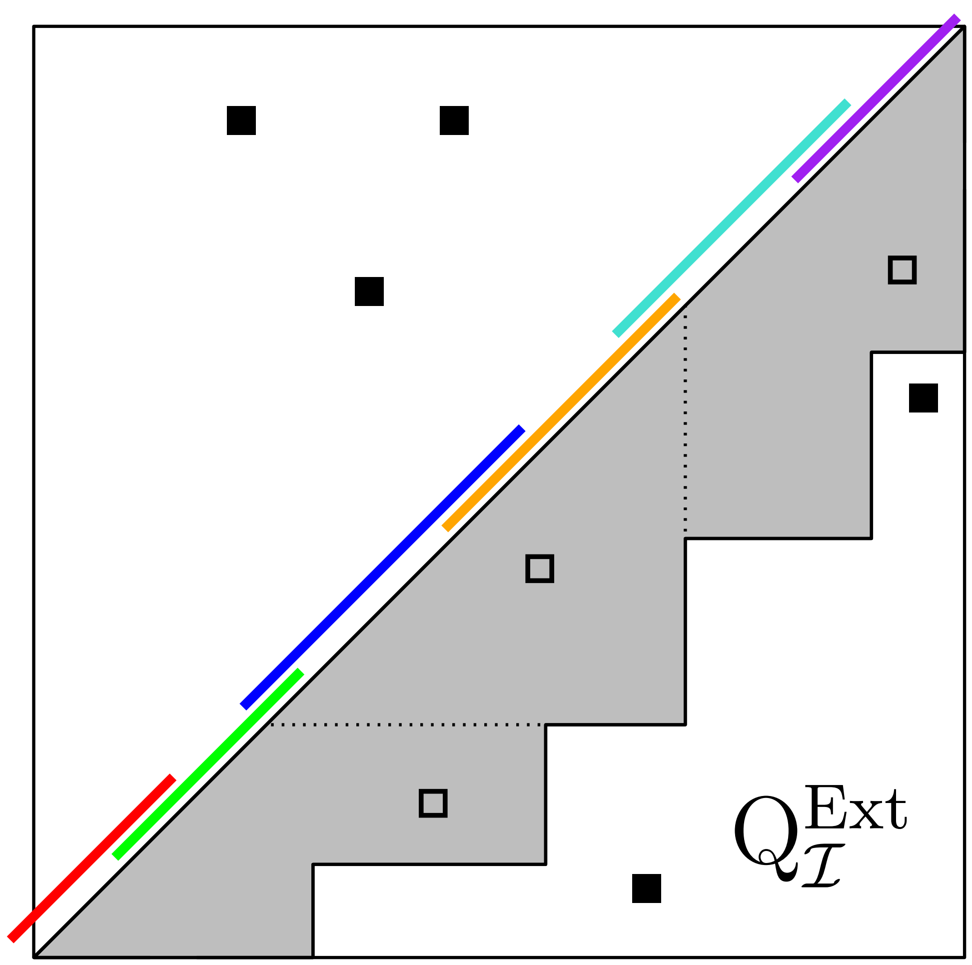


$$\mathrm{M}_f(X, \mathcal{I}) = M'_{\mathcal{I}} \circ Sh_{\mathcal{I}} \circ Sp_{\mathcal{I}} \circ M_{\mathcal{I}}(\mathrm{R}_f(X))$$

Descriptor for Mapper

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





Descriptor for Mapper

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

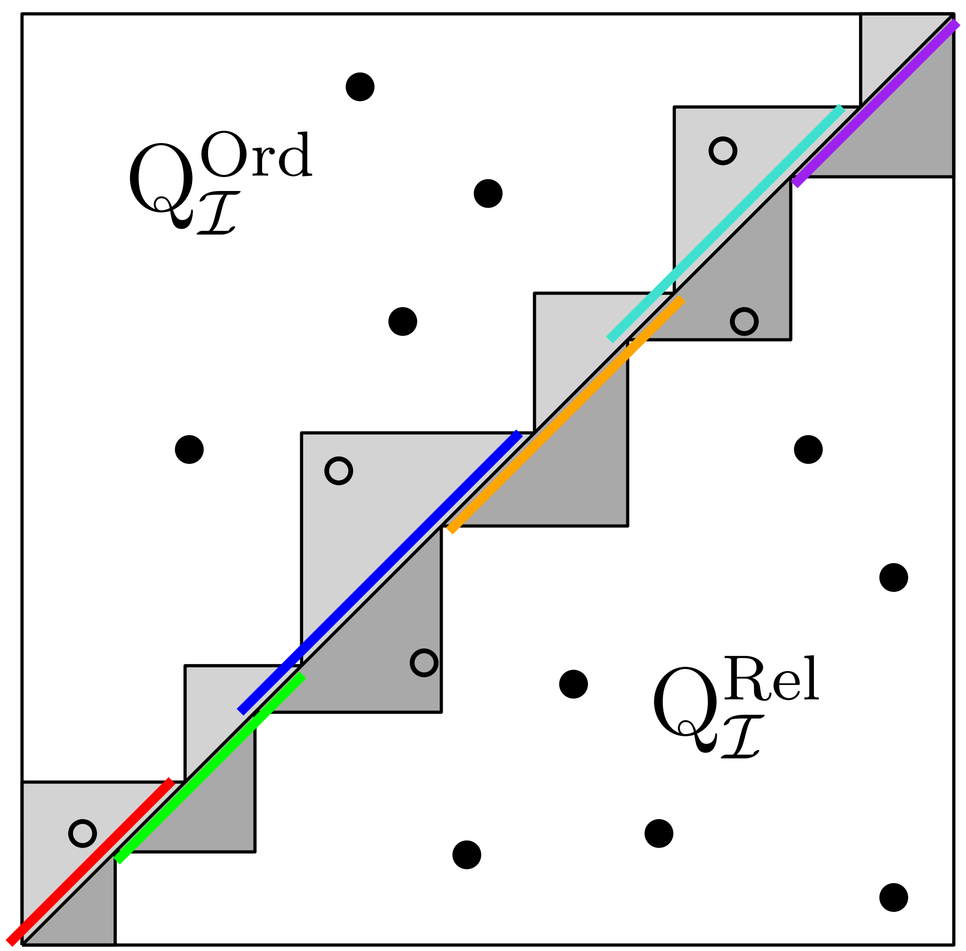
Thm: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I})$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X,\mathcal{I})$:

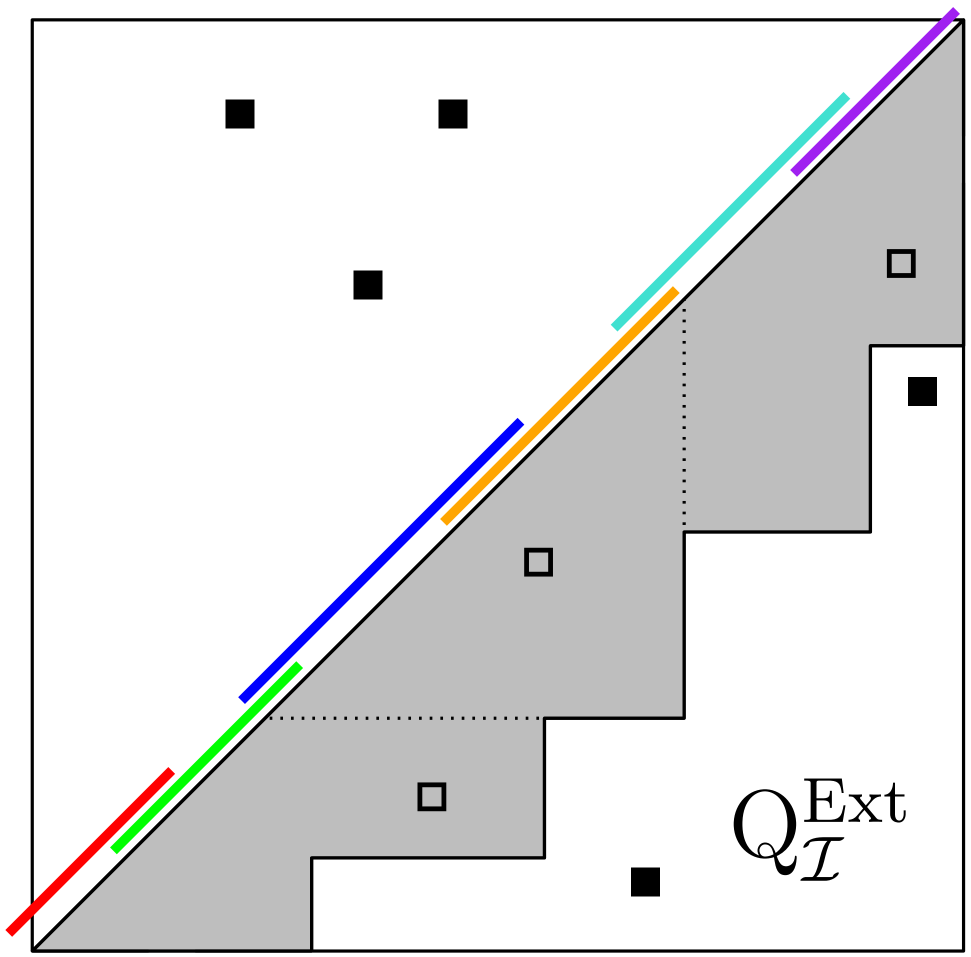
$\mathrm{Ord}_0 \longleftrightarrow$ downward branches

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

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

$\mathrm{Ext}_1 \longleftrightarrow$ loops





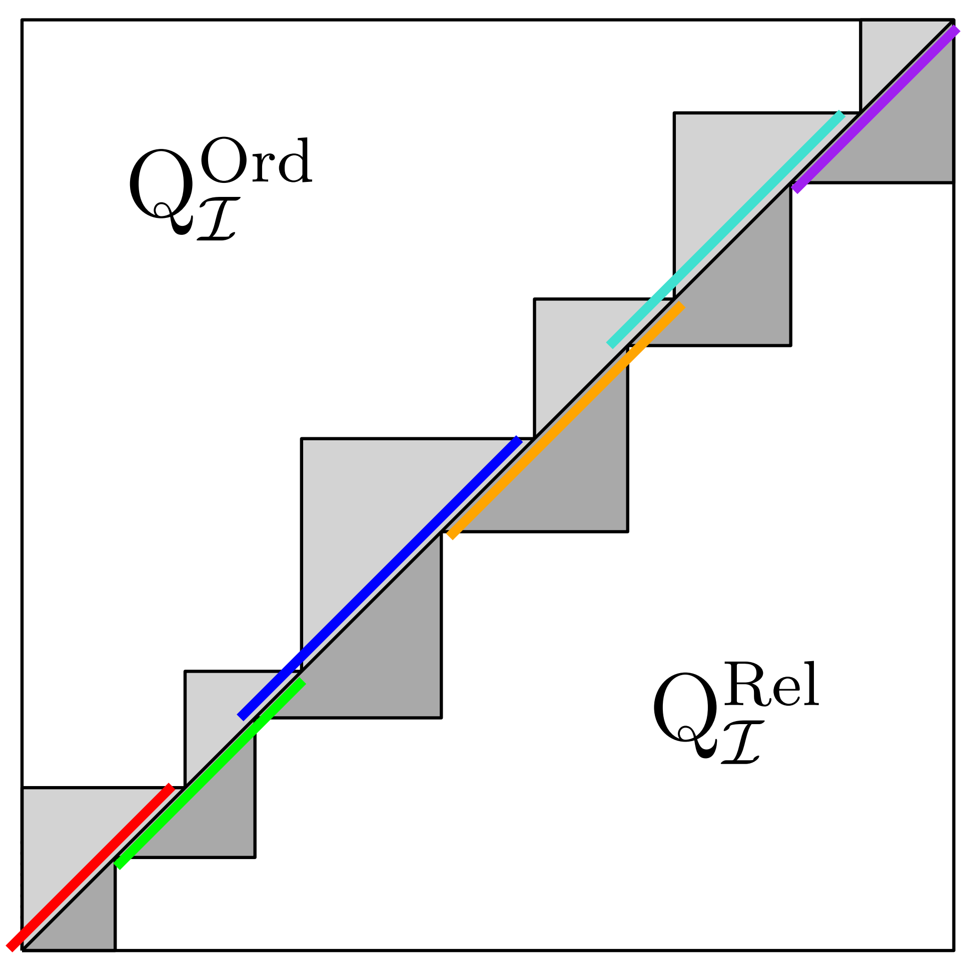
Descriptor for Mapper

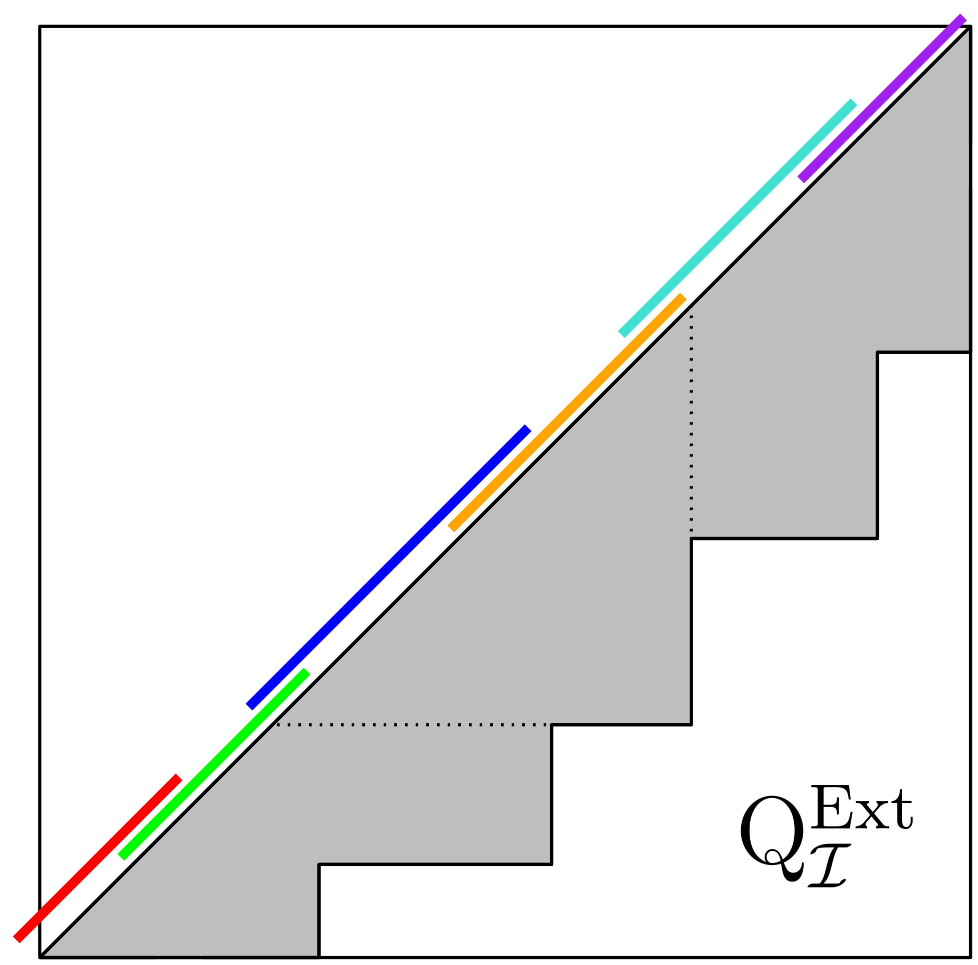
Let $\mathcal{I}$ minimal cover of $\mathrm{Im} f \subseteq \mathbb{R}$. For $I \in \mathcal{I}$, let $I = I^- \sqcup \tilde{I} \sqcup I^+$

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

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

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



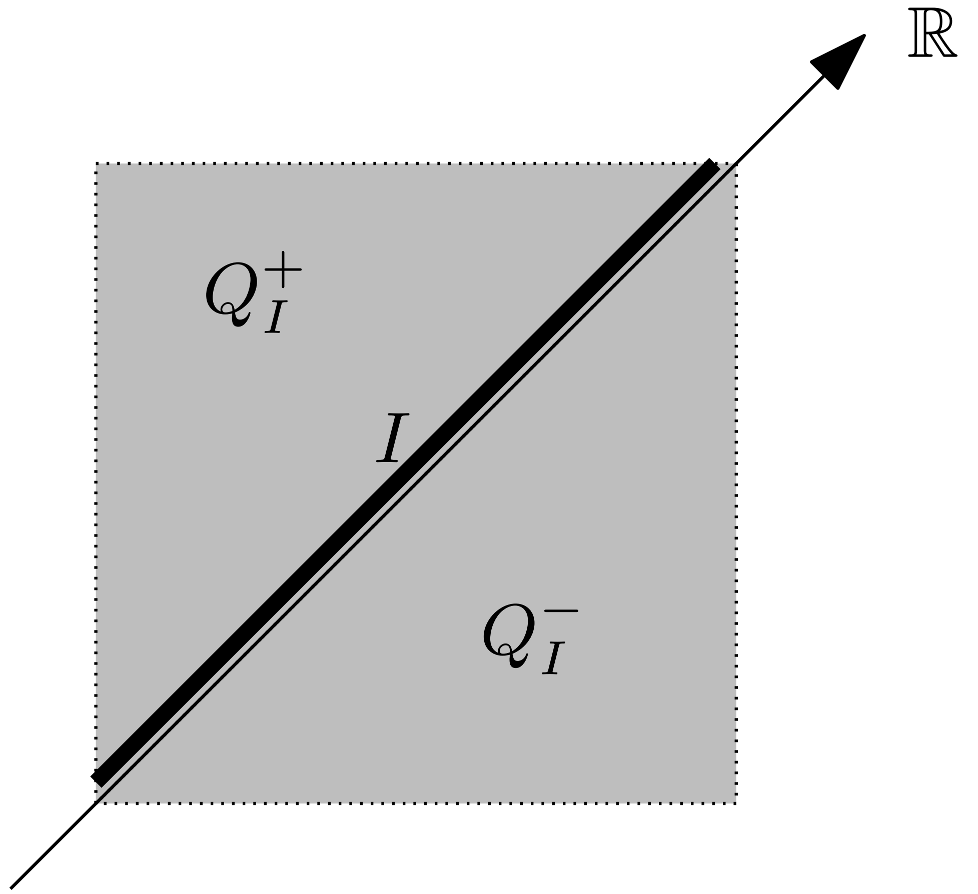


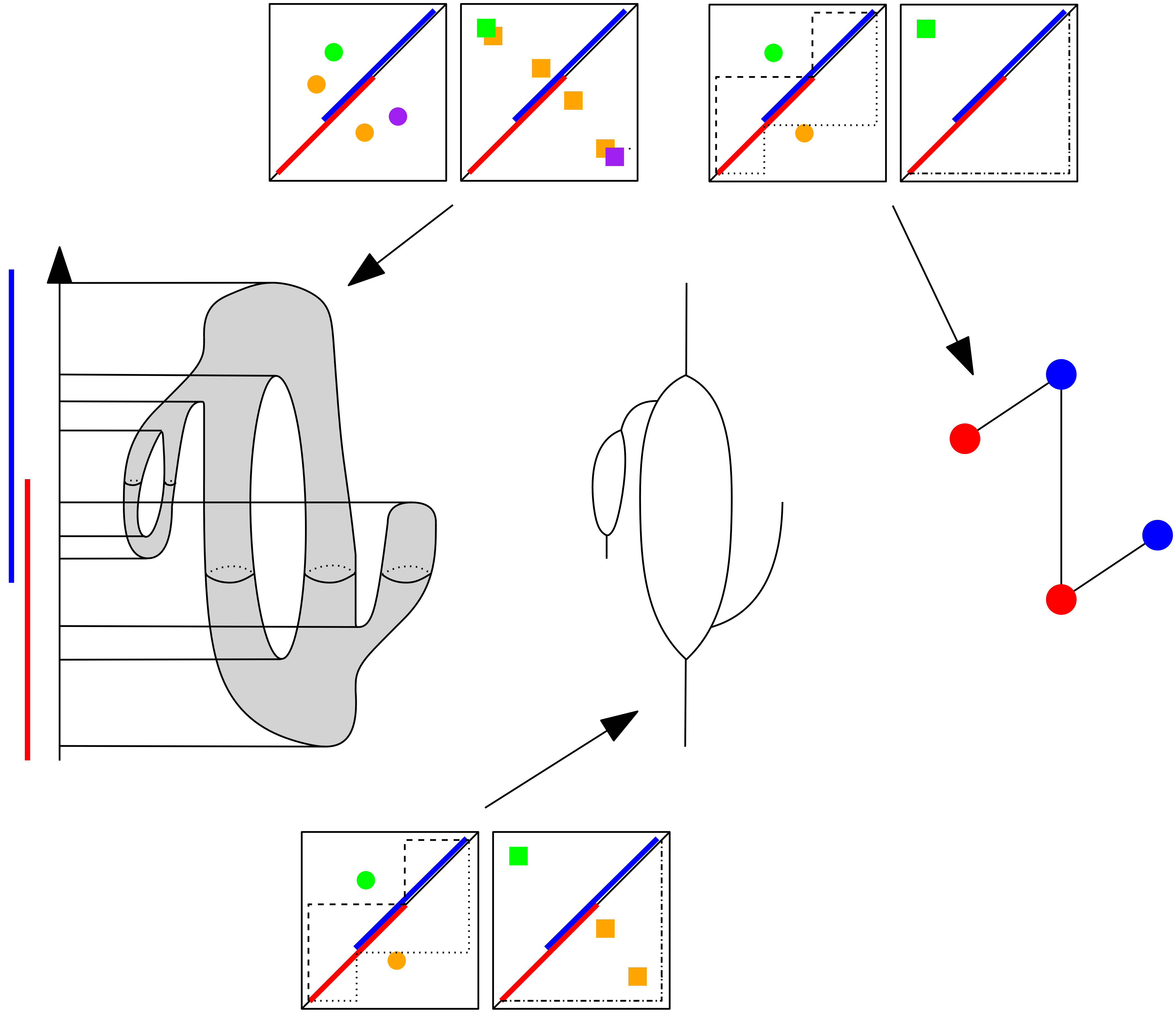
Descriptor for Mapper

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

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

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

Cor: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) = \mathrm{Dg}\,\tilde{f}$ 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

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

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

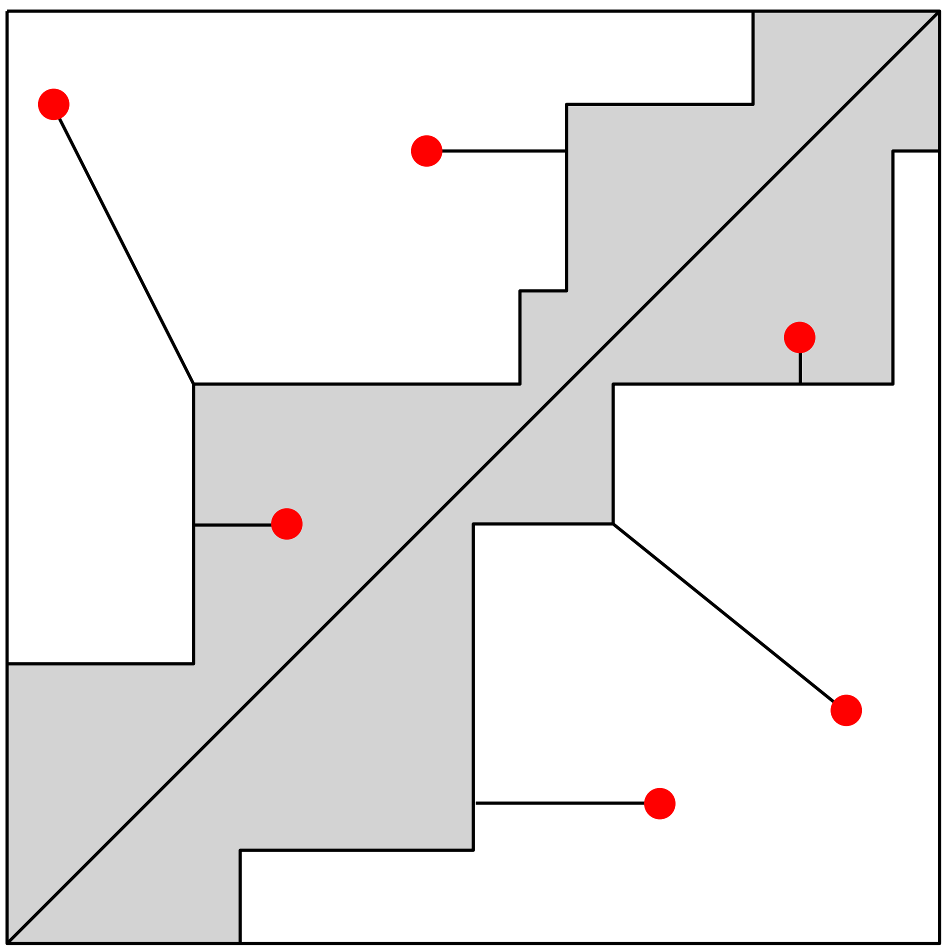
$\mathrm{Ord}_0 \longleftrightarrow$ downward branches

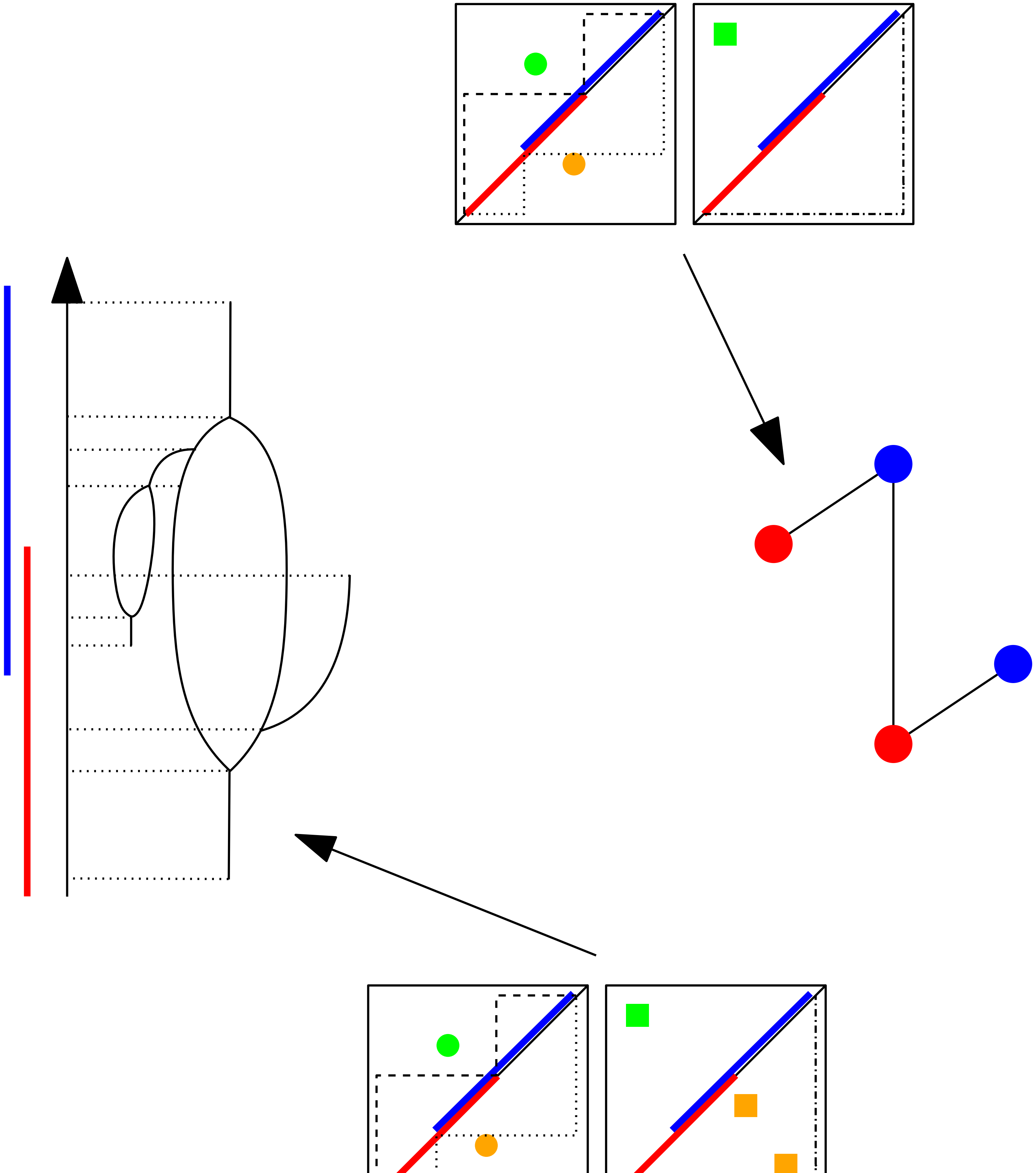
$\mathrm{Rel}_1 \longleftrightarrow$ upward branches

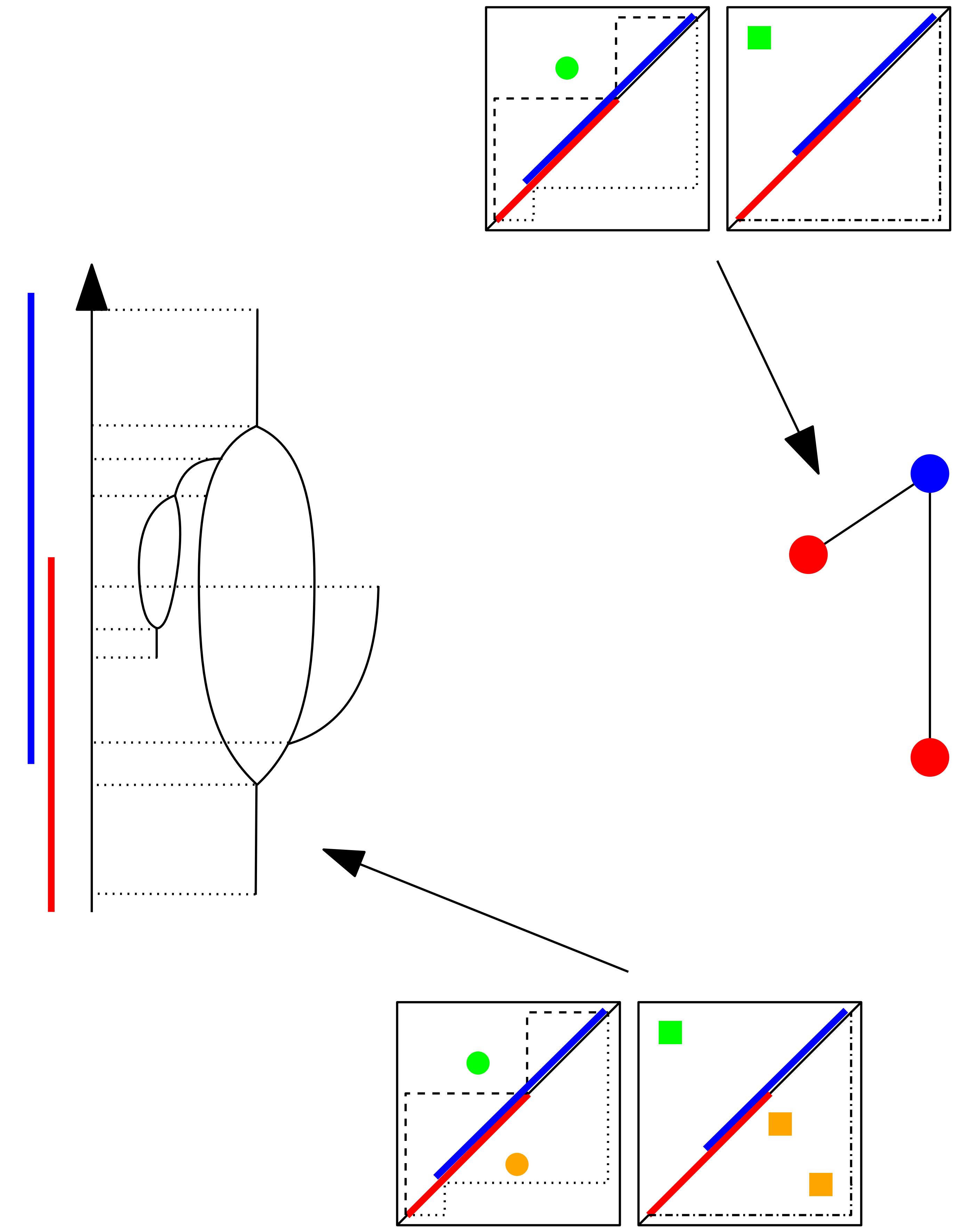
$\mathrm{Ext}_0 \longleftrightarrow$ trunks (cc)

$\mathrm{Ext}_1 \longleftrightarrow$ loops

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

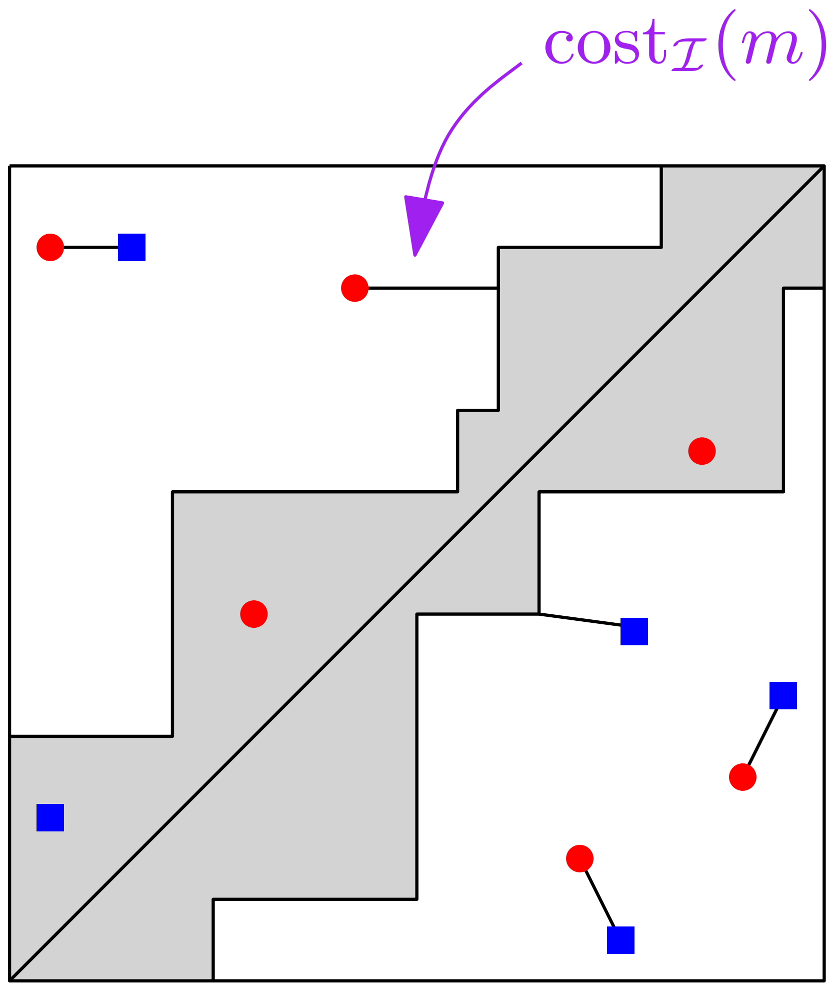






Stability of Mapper

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



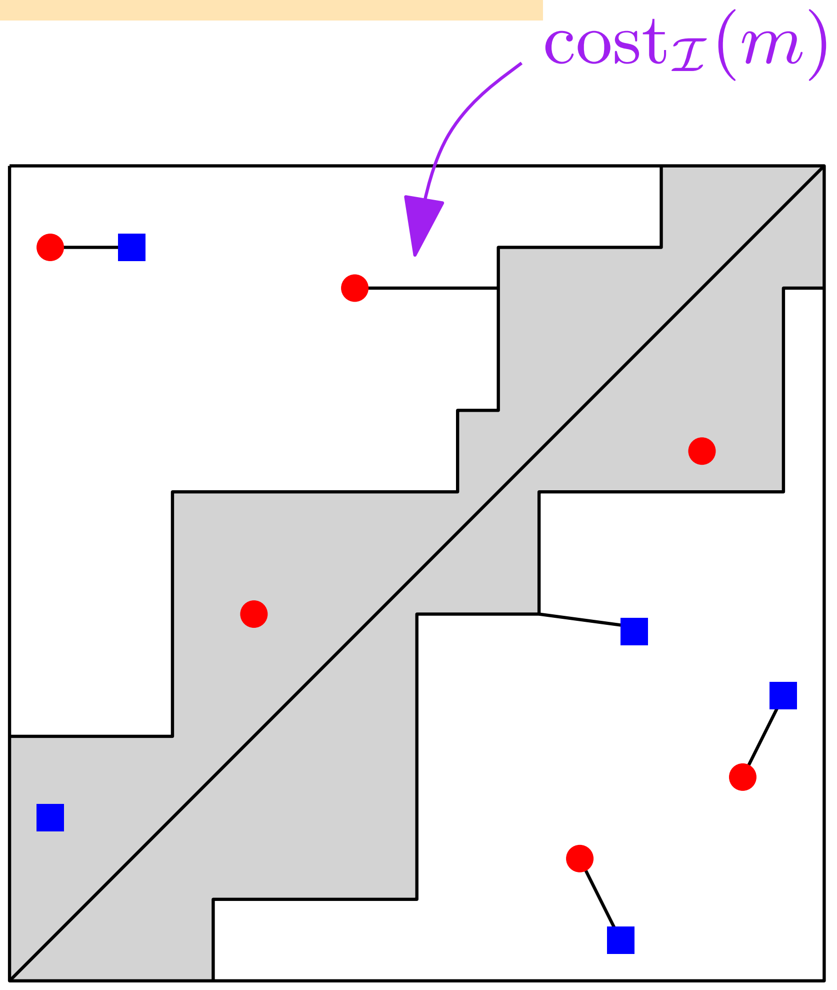
$m:\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})$

Stability of Mapper

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

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

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



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

Stability of Mapper

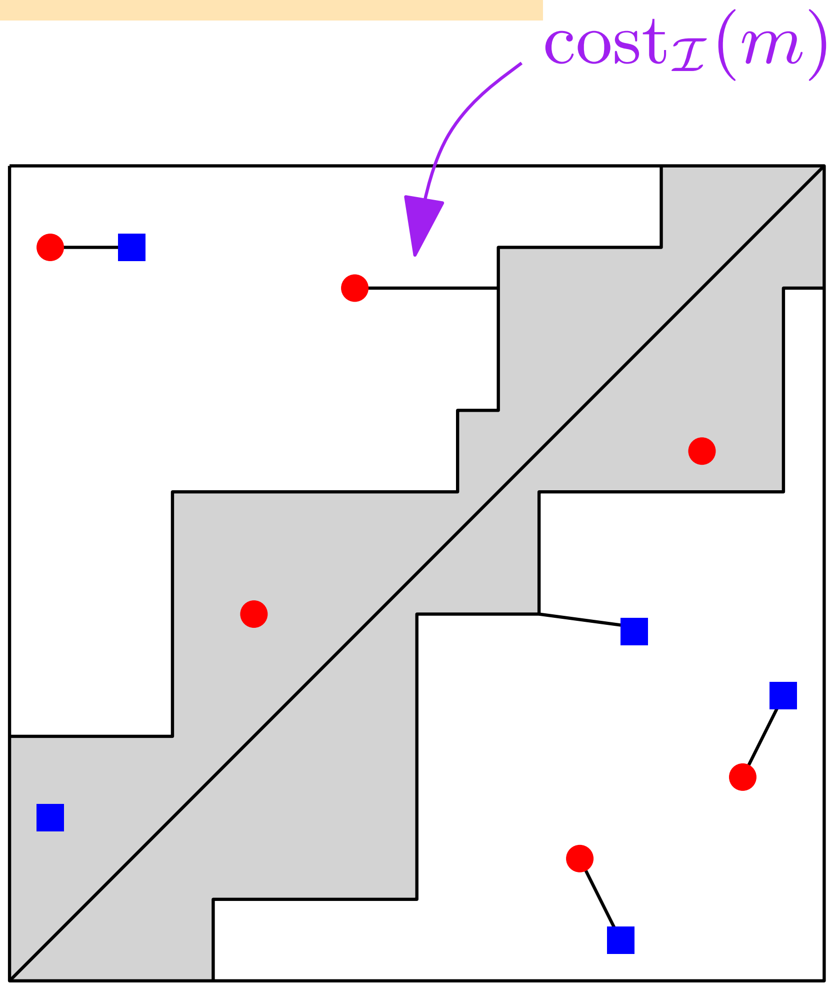
Def: $\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I})) := \inf_m \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$

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 $\mathcal{I}$



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

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: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$

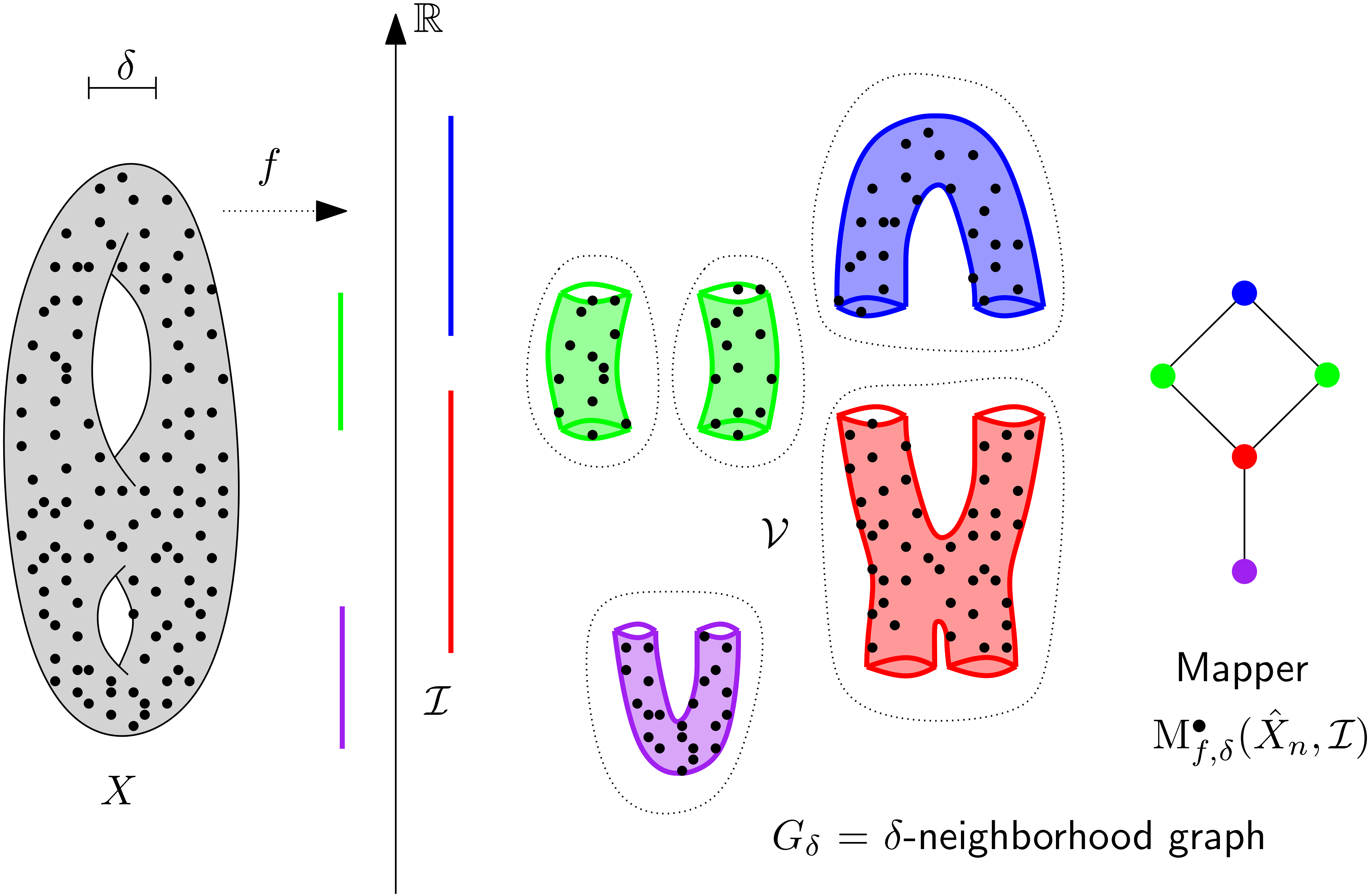
• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in G $\to$ connected cover $\mathcal{V}$

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

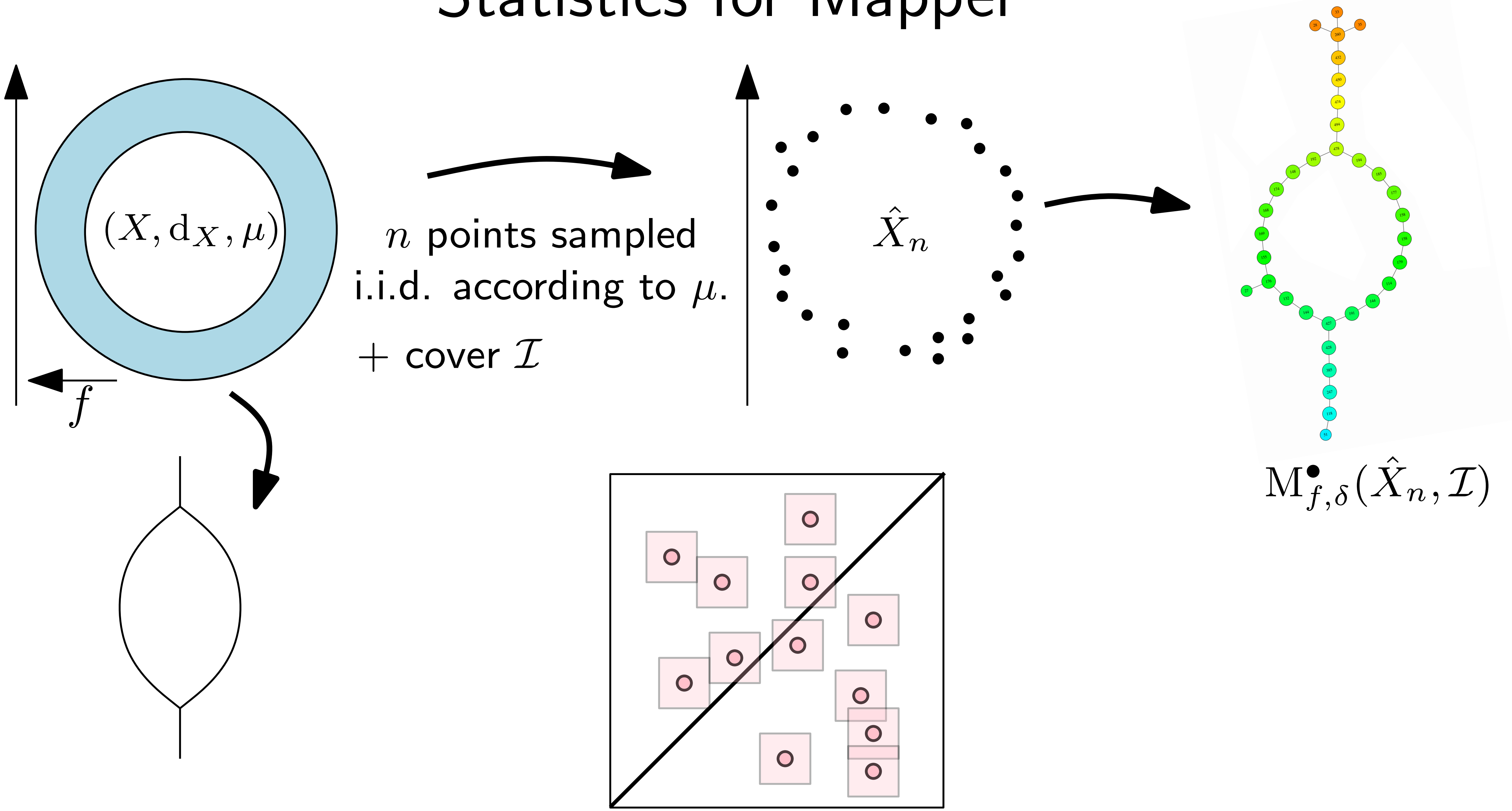
- 1 vertex per element $V \in \mathcal{V}$
- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$
- 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

(intersections materialized by data points)

Mapper in practice



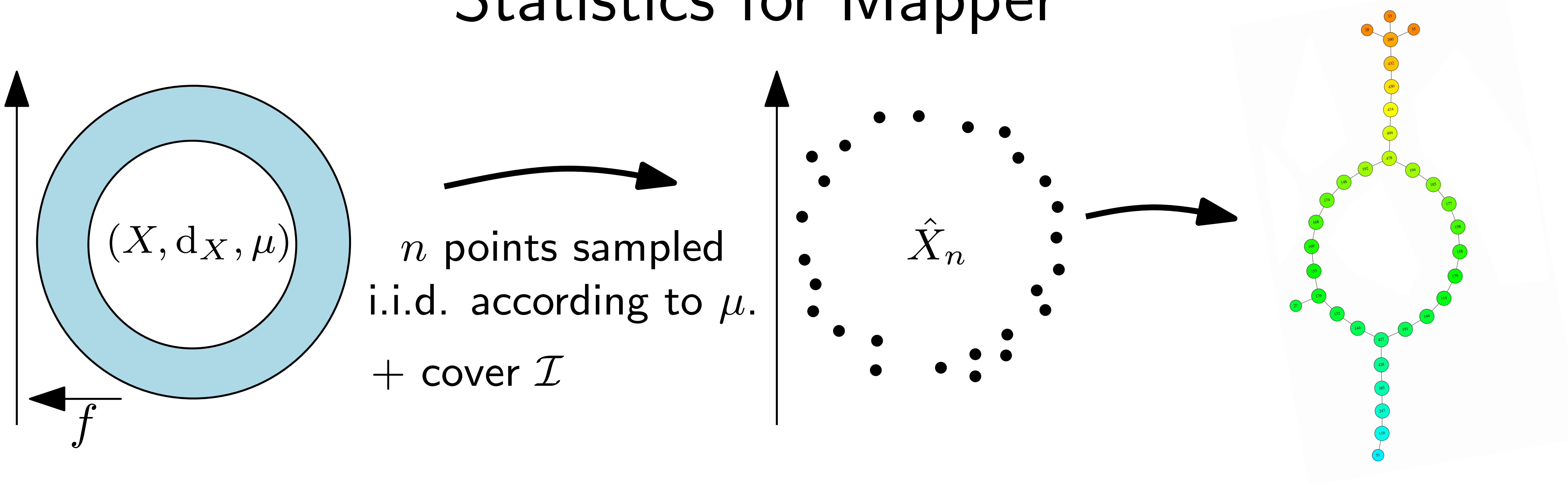
Statistics for Mapper



Questions:

- Statistical properties of the estimator $\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$?
- Convergence to the ground truth $\mathrm{R}_f(X)$ in d_B? Deviation bounds?

Statistics for Mapper



Let $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$ denote $\mathrm{M}_f(G_\delta, \mathcal{I})$

1. Link between $\mathrm{R}_f(X)$ and $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{V})$?

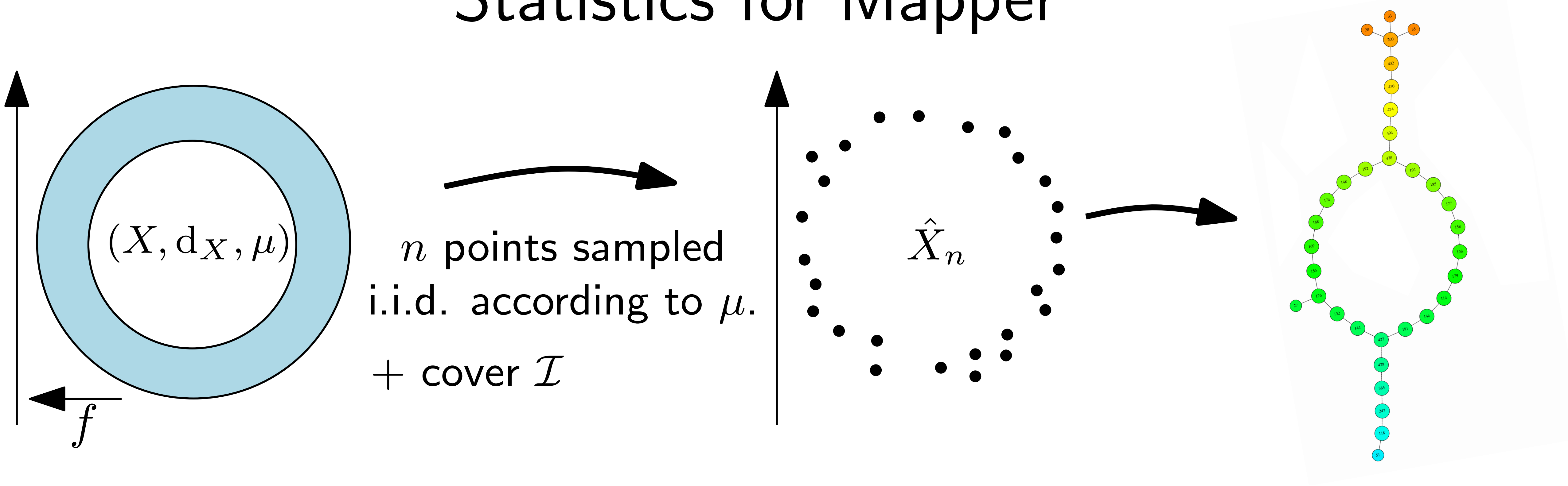
a. support $\rightarrow$ δ-neighborhood graph
$X \rightarrow G_\delta(\hat{X}_n)$

b. Reeb graph $\rightarrow$ Mapper

2. Link between $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$ and $\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

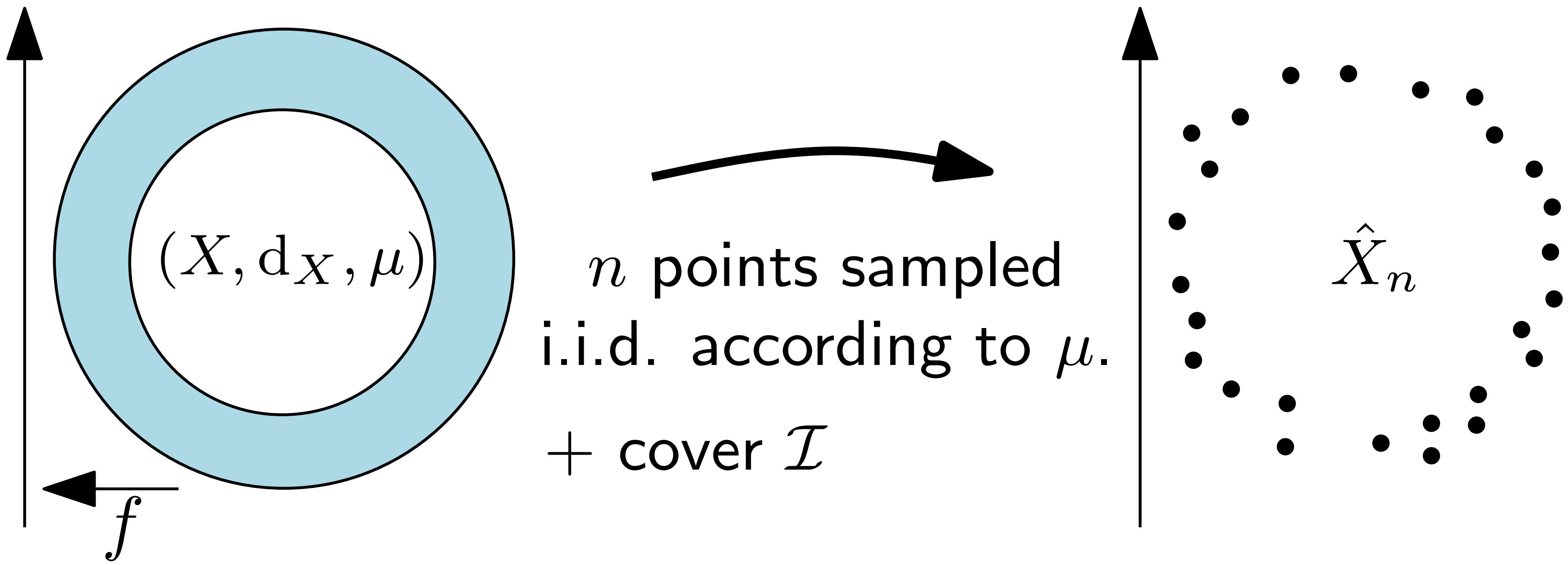
intersections given by metric graph $\rightarrow$ intersections given by points

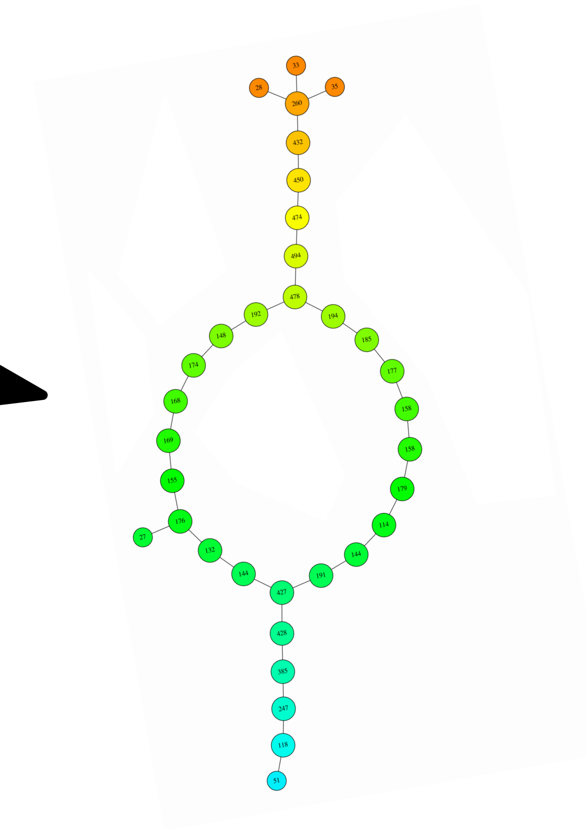
Statistics for Mapper



1. Link between $\mathrm{R}_f(X)$ and $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

Statistics for Mapper





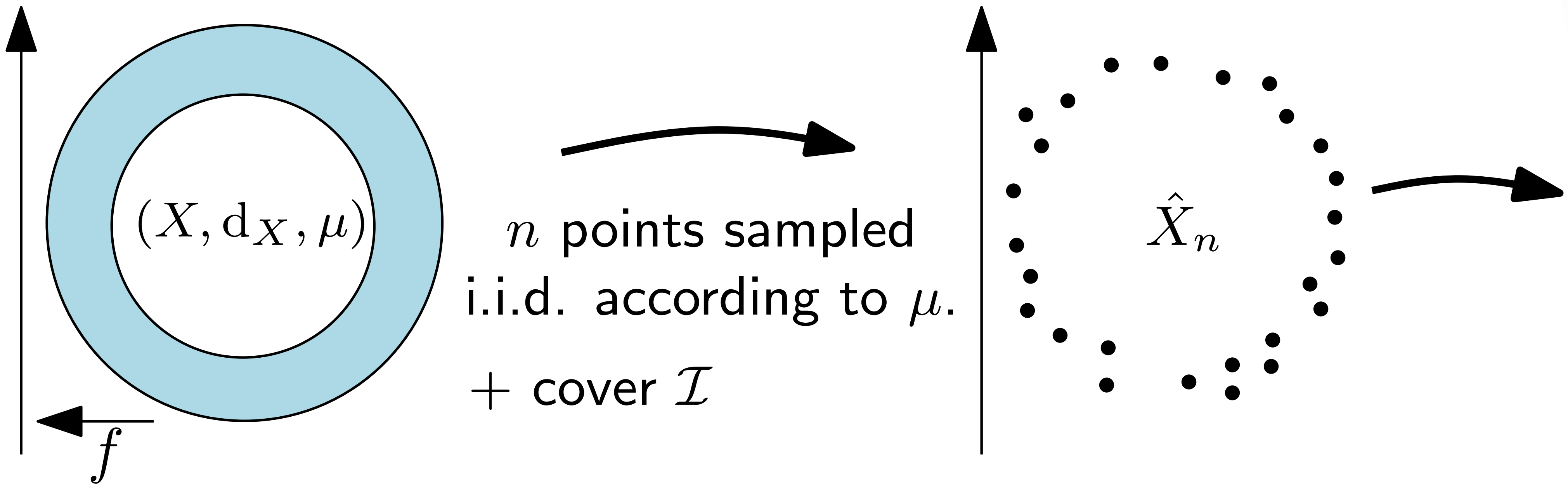
1. Link between $\mathrm{R}_f(X)$ and $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

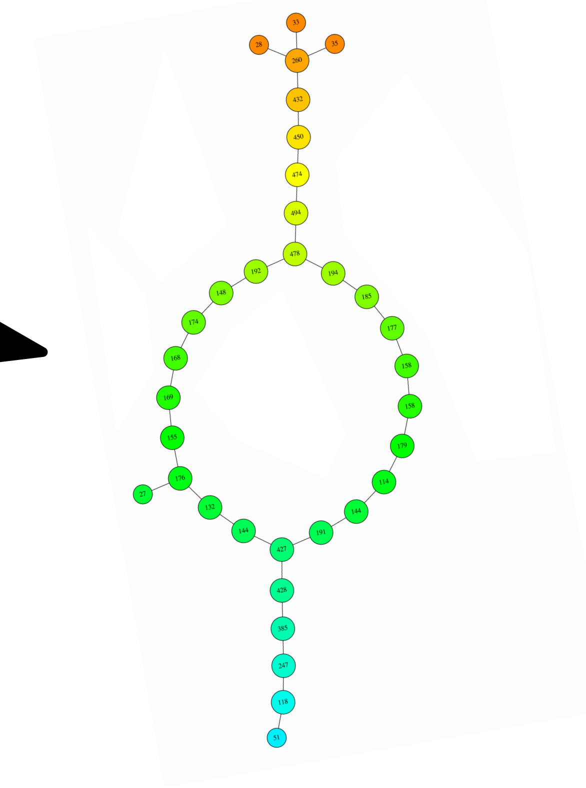
support $\rightarrow$ δ-neighborhood graph

Thm: If $4d_H(X, \widehat{X}_n) \leq \delta \leq \min\left\{\frac{1}{4}\mathrm{rch}(X), \frac{1}{4}\rho(X)\right\}$

$$d_B(\mathrm{Dg}\,\mathrm{R}_f(X), \mathrm{Dg}\,\mathrm{R}_f(G_\delta(\widehat{X}_n))) \leq 2\omega(\delta)$$

Statistics for Mapper





1. Link between $\mathrm{R}_f(X)$ and $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

support $\rightarrow$ δ-neighborhood graph

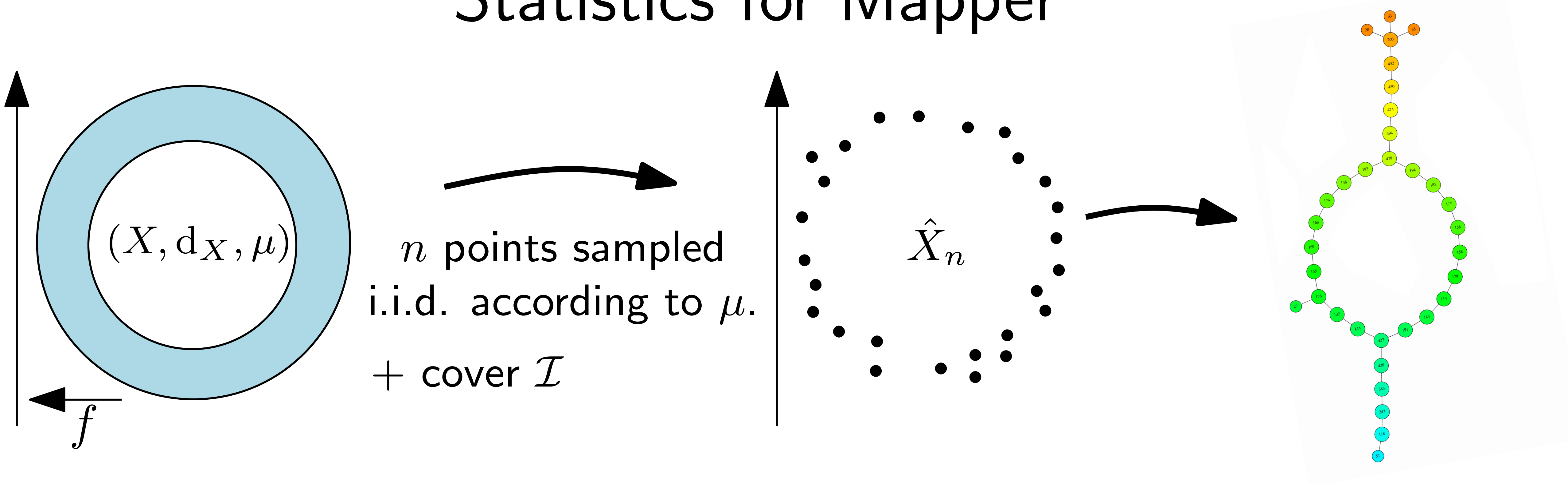
Thm: If $4d_H(X, \widehat{X}_n) \leq \delta \leq \min\left\{\frac{1}{4}\mathrm{rch}(X), \frac{1}{4}\rho(X)\right\}$

$$d_B(\mathrm{Dg}\,\mathrm{R}_f(X), \mathrm{Dg}\,\mathrm{R}_f(G_\delta(\widehat{X}_n))) \leq 2\omega(\delta)$$

Reeb graph $\rightarrow$ Mapper

Thm: $d_B(\mathrm{Dg}\,\mathrm{R}_f(G_\delta(\hat{X}_n)), \mathrm{Dg}\,\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})) \leq r$

Statistics for Mapper



1. Link between $\mathrm{R}_f(X)$ and $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

ω: modulus of continuity of f

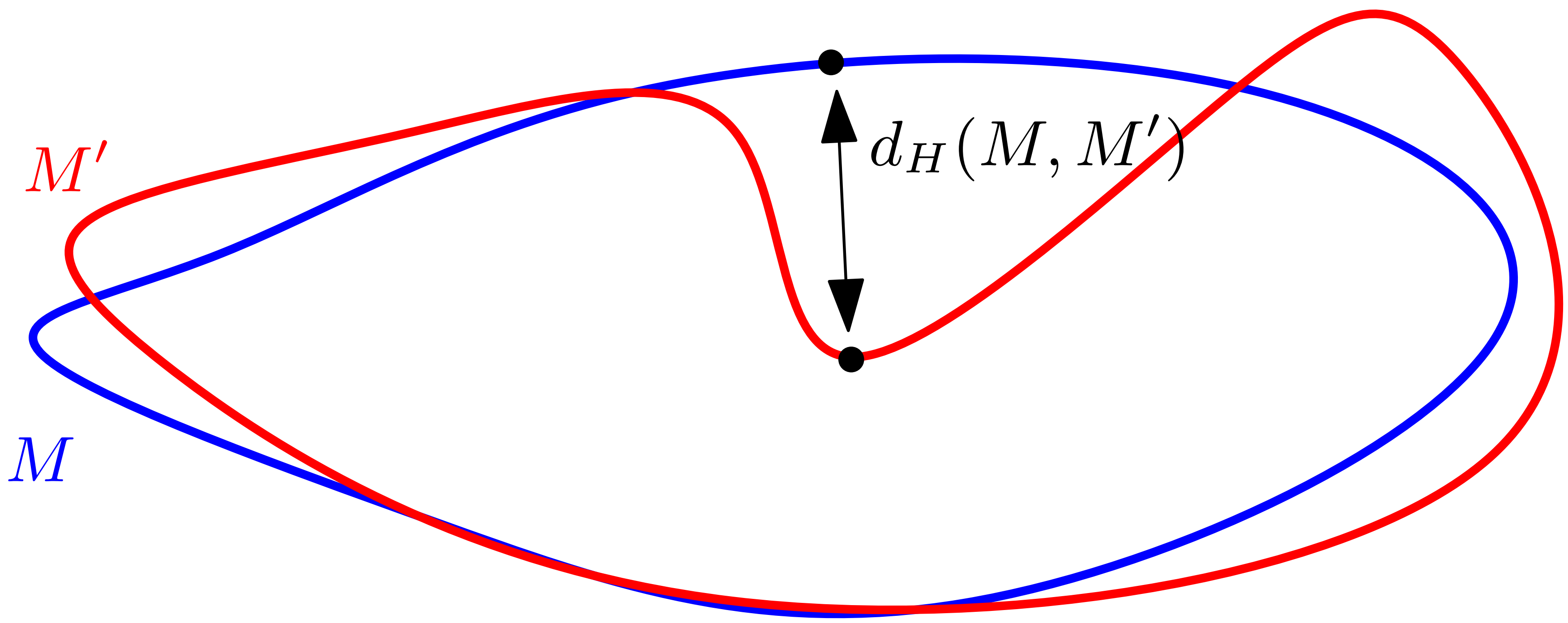
$$\omega : \delta \mapsto \sup\{|f(x) - f(y)| \, : \, d(x, y) \leq \delta\}$$

rch: reach of X.

ρ: radius of convexity of X: largest r s.t. geodesic balls of radius r are convex.

d_H: Hausdorff distance.

Statistics for Mapper



The distance function to a compact $M \subset \mathbb{R}^d$, $d_M : \mathbb{R}^d \to \mathbb{R}_+$ is defined by:

$$d_M(x) = \inf_{p \in M} \|x - p\|$$

The Hausdorff distance between two compact sets $M, M' \subset \mathbb{R}^d$ is:

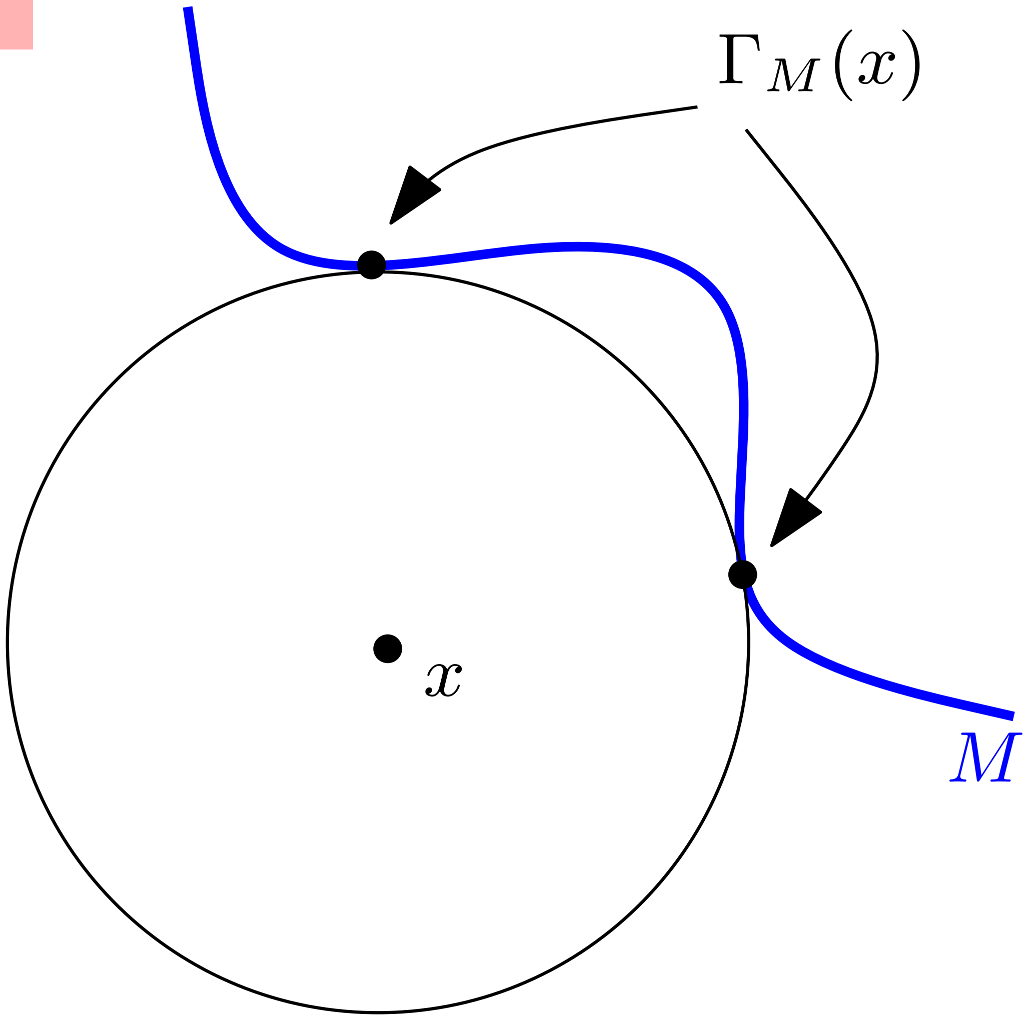
$$d_H(M, M') = \sup_{x \in \mathbb{R}^d} |d_M(x) - d_{M'}(x)|$$

Statistics for Mapper

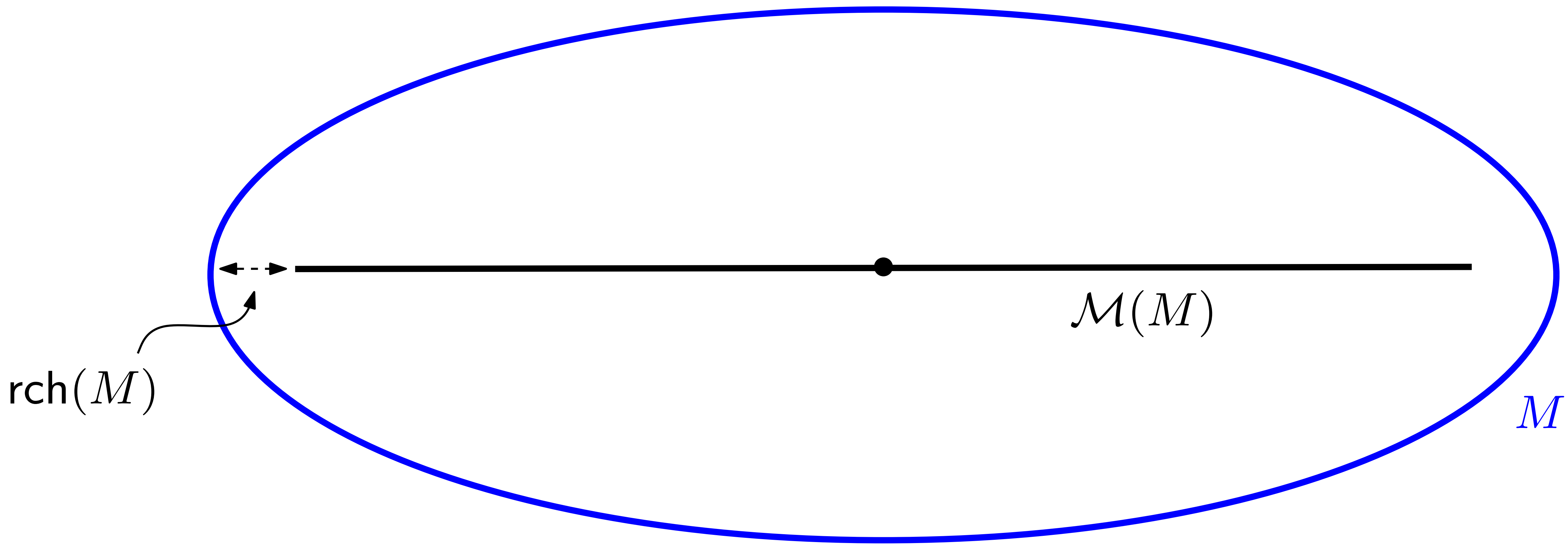
$$\Gamma_M(x) = \{y \in M : d_M(x) = \|x - y\|\}$$

Def: The medial axis of M:

$$\mathcal{M}(M) = \{x \in \mathbb{R}^d : |\Gamma_M(x)| \geq 2\}$$



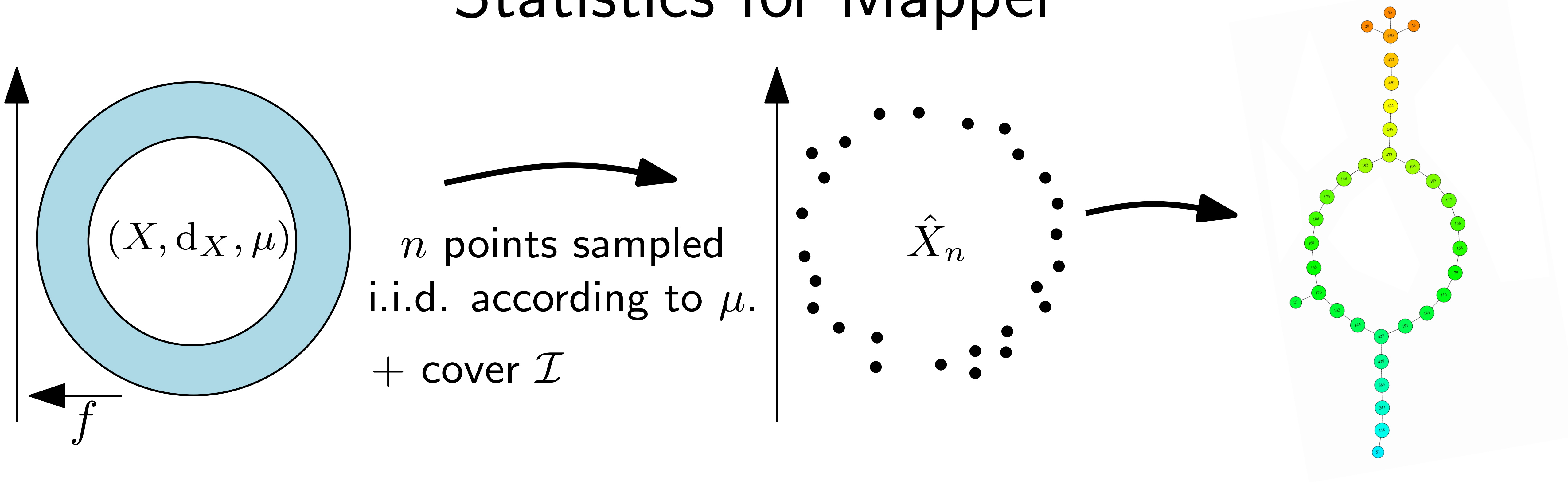
Statistics for Mapper



Def: The reach of M, rch(M) is the smallest distance from $\mathcal{M}(M)$ to M:

$$\mathrm{rch}(M) = \inf_{y \in \mathcal{M}(M)} d_M(y)$$

Statistics for Mapper



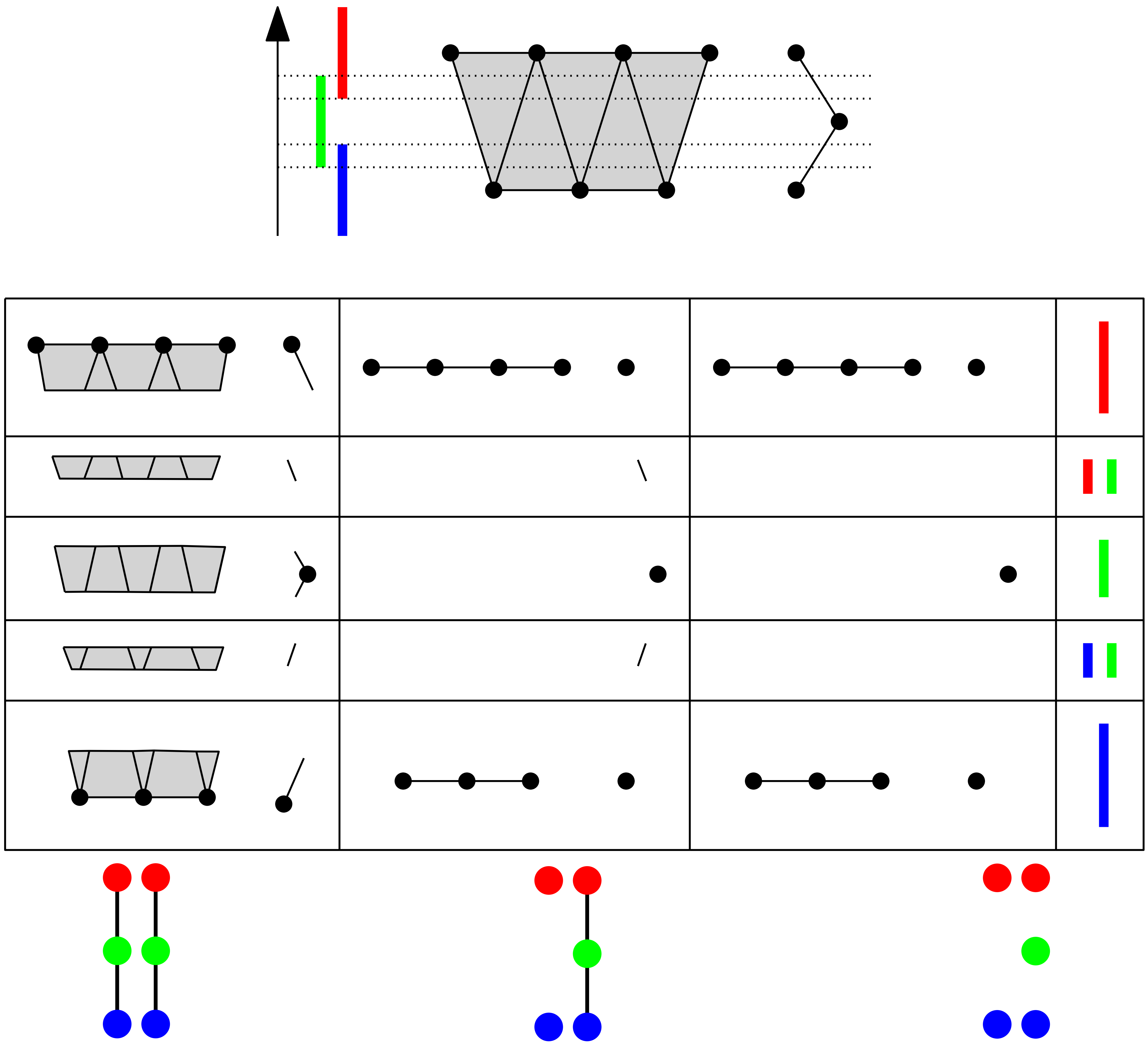
2. Link between $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$ and $\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

intersections given by metric graph $\rightarrow$ intersections given by points

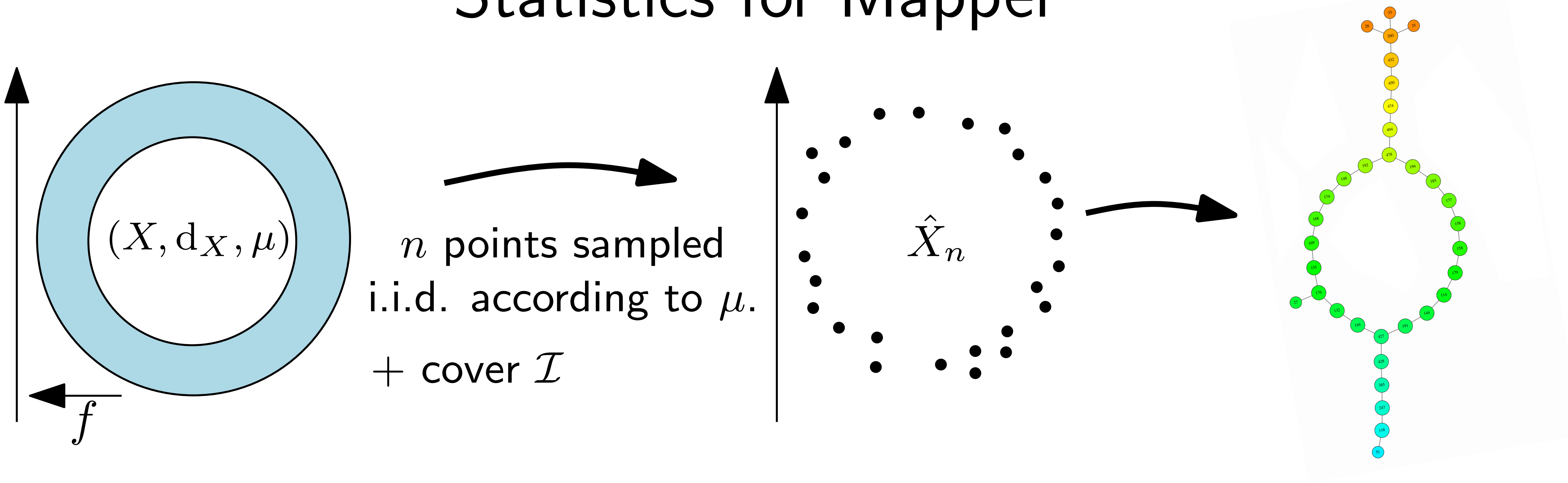
Thm: If there are no **intersection-crossing edges**, then

$$\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I}) = \mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$$

Statistics for Mapper



Statistics for Mapper



$\hat{X}_n$ is random $\Rightarrow d_H(X, \hat{X}_n)$ is random

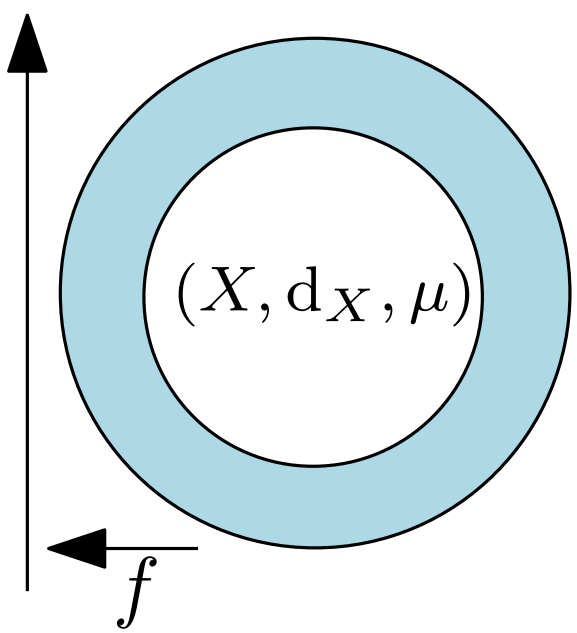
Hyp: μ is (a, b)-standard

$$\mu(B(x, r)) \geq \min\{1, ar^b\} \text{ for all } x \in X \text{ and } r > 0$$

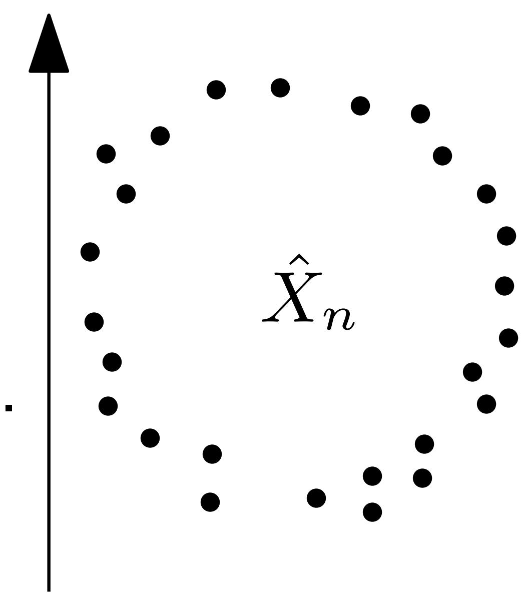
Then it is known that, for n sufficiently large, one has with high probability:

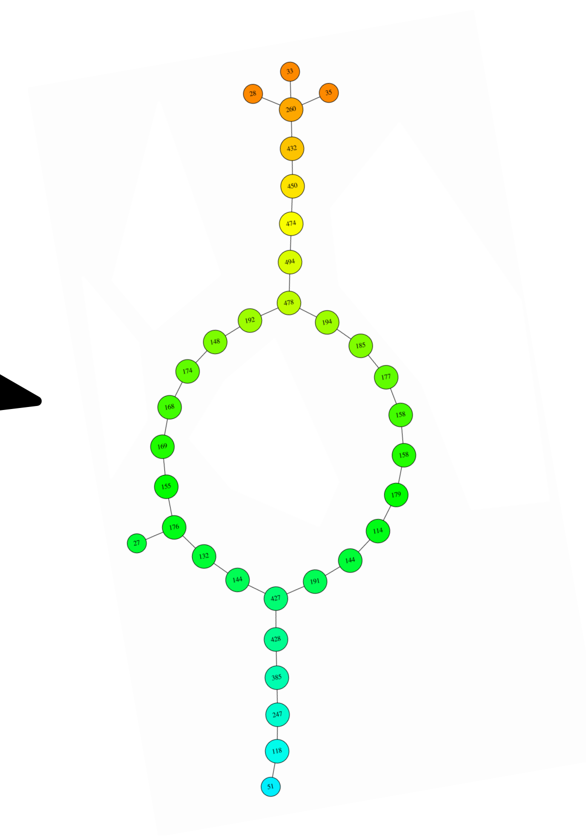
$$d_H(X, \hat{X}_n) \leq \left(\frac{2 \log n}{an}\right)^{1/b}$$

Statistics for Mapper



n points sampled
i.i.d. according to μ.
$+$ cover $\mathcal{I}$





Thm: If μ is (a, b)-standard and f is c-Lipschitz then for:

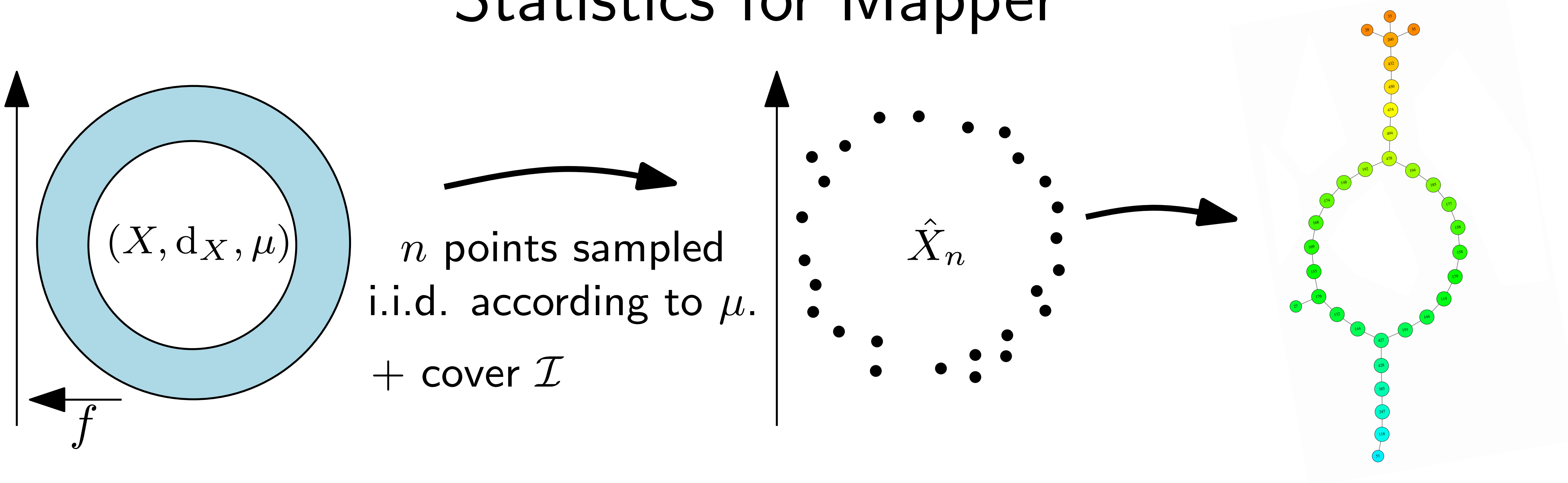
$$\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b},\ g_n \in \left(\frac{1}{3},\ \frac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n}, \qquad \text{one has } \forall \varepsilon > 0$$

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

where C depends only on a, b, c.

More generally: $r_n = \omega(\delta_n)/g_n$

Statistics for Mapper



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.

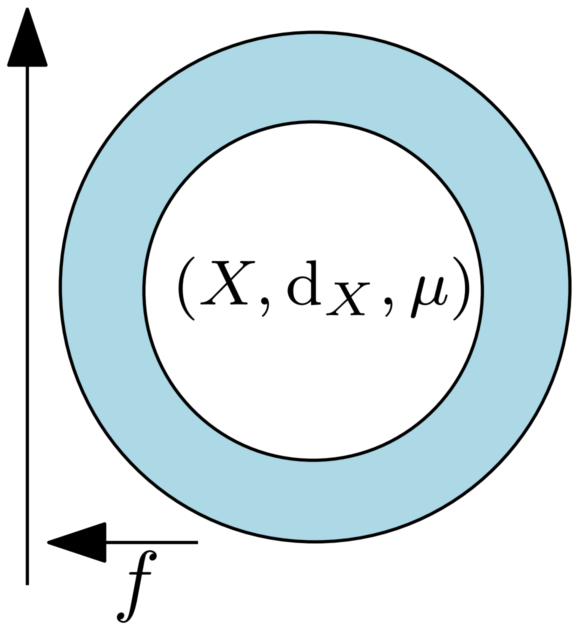
Thm: For any estimator $\widehat{\mathrm{R}}$, one has:

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

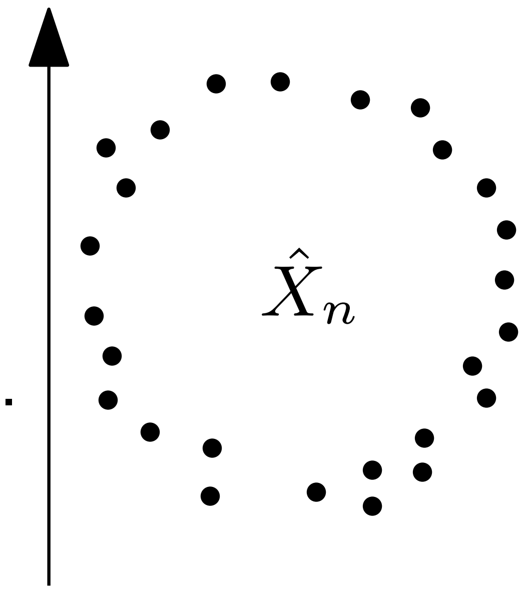
where C depends only on a, b.

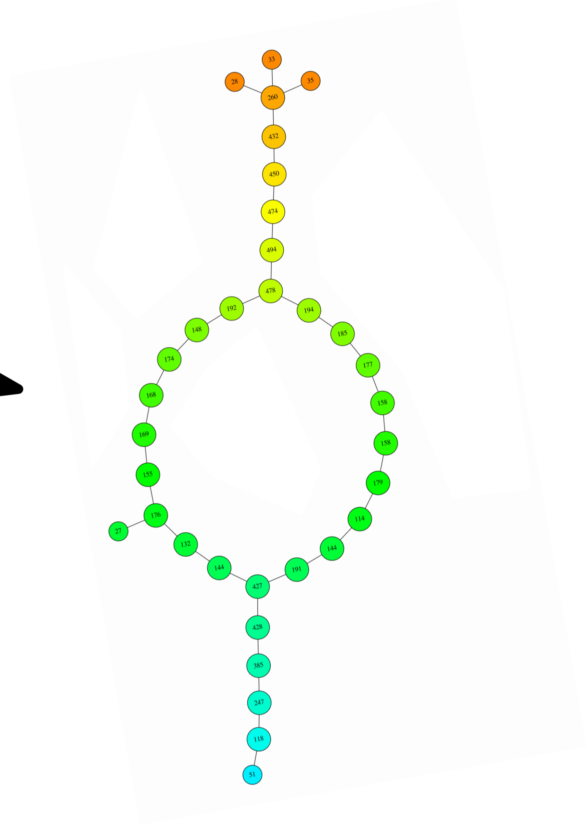
Consequence of Le Cam's lemma

Statistics for Mapper



n points sampled
i.i.d. according to μ.
$+$ cover $\mathcal{I}$





Thm: If μ is (a, b)-standard and f is c-Lipschitz then for:

$$\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b},\ g_n \in \left(\frac{1}{3},\ \frac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n},$$

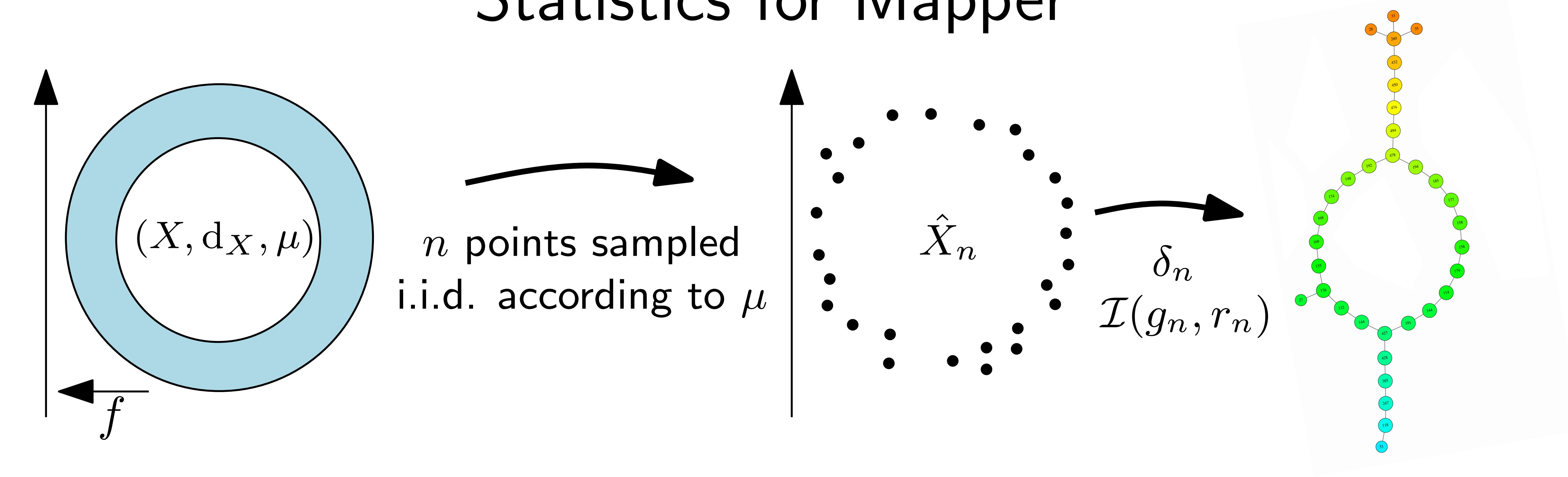
one has $\forall \varepsilon > 0$

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

where C depends only on a, b, c.

More generally: $r_n = \omega(\delta_n)/g_n$

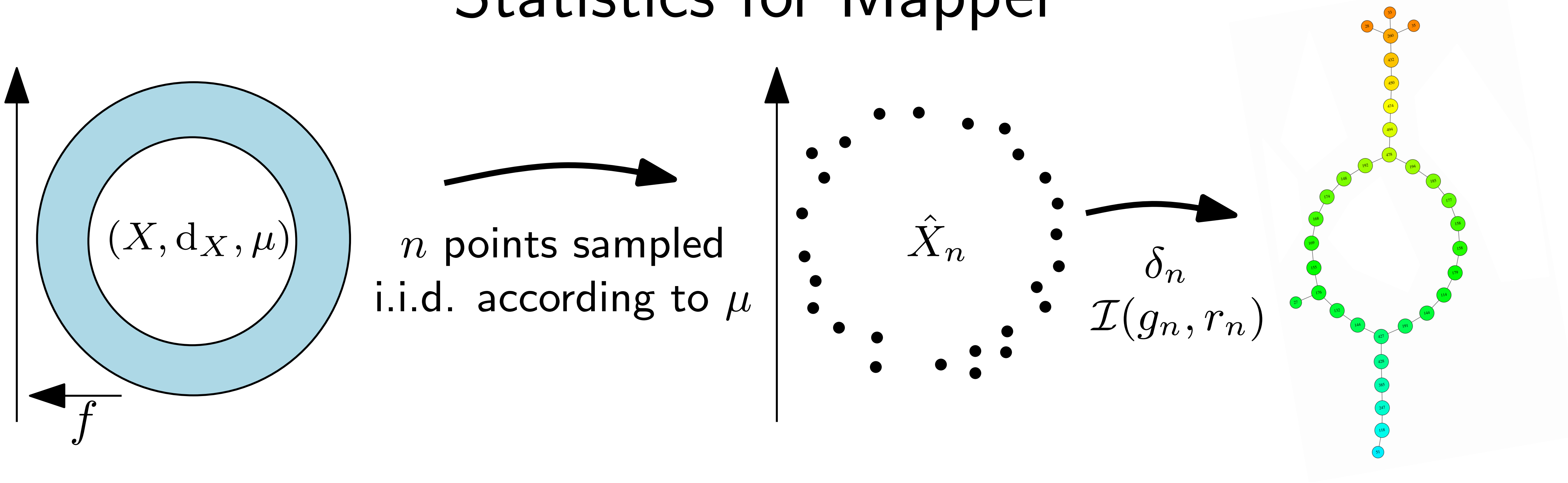
Statistics for Mapper



$\rightarrow$ subsampling to tune δ_n: let $\beta > 0$ and take $s(n) = \frac{n}{\log(n)^{1+\beta}}$

$\delta_n := d_H(\hat{X}_n^{s(n)}, \hat{X}_n)$ where $\hat{X}_n^{s(n)}$ is a subset of $\hat{X}_n$ of size $s(n)$

Statistics for Mapper



$\rightarrow$ subsampling to tune δ_n: let $\beta > 0$ and take $s(n) = \frac{n}{\log(n)^{1+\beta}}$

$\delta_n := d_H(\hat{X}_n^{s(n)}, \hat{X}_n)$ where $\hat{X}_n^{s(n)}$ is a subset of $\hat{X}_n$ of size $s(n)$

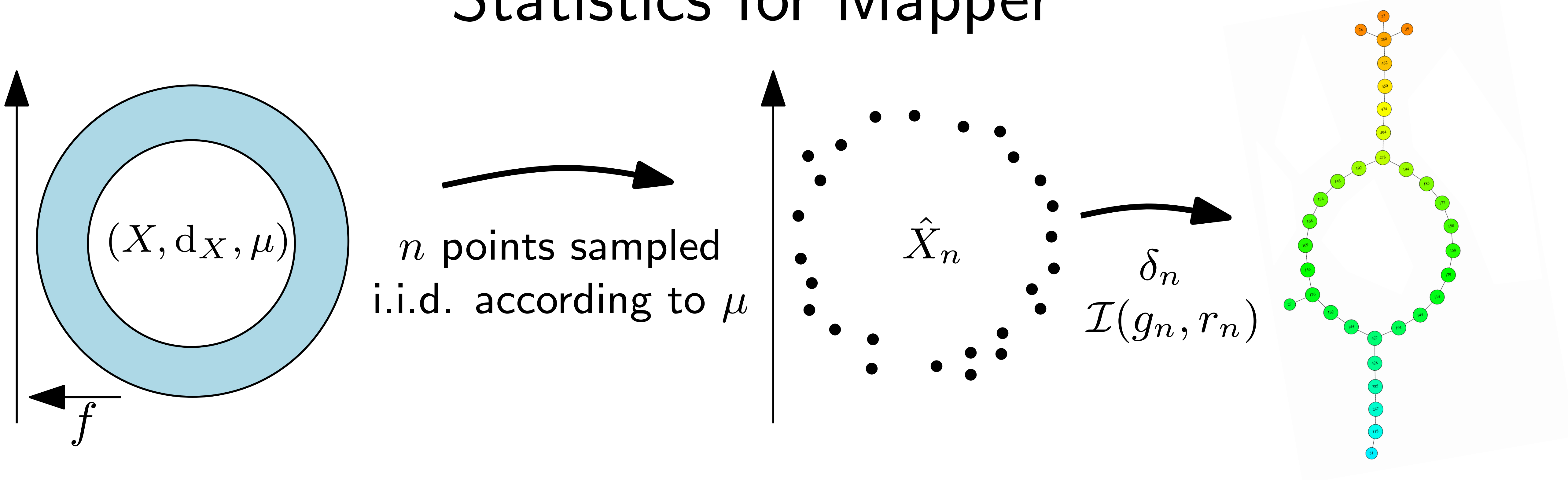
Thm: If μ is (a, b)-standard and f is c-Lipschitz, then for:

$$\delta_n = d_H(\hat{X}_n^{s(n)}, \hat{X}_n),\ g_n \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n}, \qquad \text{one has } \forall \varepsilon > 0$$

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathrm{M}^{\bullet}_{f,\delta_n}(\hat{X}_n, \mathcal{I}(g_n, r_n)),\ \mathrm{Dg}\,\mathrm{R}_f(X)\right)\right] \leq C\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b},$$

where C depends only on a, b, c.

Statistics for Mapper



Ex : PCA filter

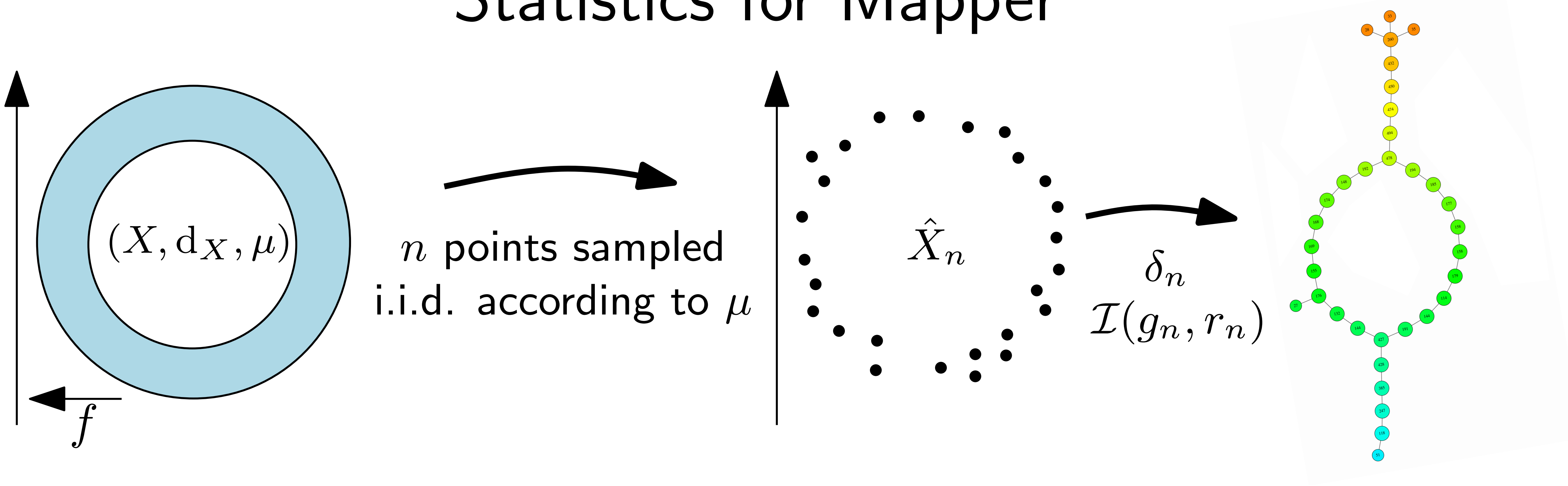
Π_1 : orthonormal projection onto first principal direction of the covariance operator

$\widehat{\Pi}_1$: orthonormal projection onto first principal direction of the empirical covariance operator

Using [Biau et. al. 2012]:

$$\mathbb{E}\left[d_B\left(\mathrm{R}_{\Pi_1}(\mathcal{X}), \mathrm{M}^{\bullet}_{\widehat{\Pi}_1(\widehat{X}_n),\delta_n}(\widehat{X}_n, \mathcal{I}(g_n, r_n))\right)\right] \lesssim \left(\frac{(\log(n))^{2+\beta}}{n}\right)^{1/b} \vee \frac{1}{\sqrt{n}}$$

Statistics for Mapper



Thm: If μ is (a, b)-standard and f is c-Lipschitz, then for:

$$\delta_n = d_H(\hat{X}_n^{s(n)}, \hat{X}_n),\ g_n \in \left(\tfrac{1}{3},\ \tfrac{1}{2}\right),\ r_n = \tfrac{c\delta_n}{g_n}, \qquad \text{one has } \forall \varepsilon > 0$$

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathrm{M}^{\bullet}_{f,\delta_n}(\hat{X}_n, \mathcal{I}(g_n, r_n)),\ \mathrm{Dg}\,\mathrm{R}_f(X)\right)\right] \leq C\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b},$$

where C depends only on a, b, c.

Get confidence region with $\mathbb{E}\left[d(\cdot,\cdot)\right] = \int_\alpha \mathbb{P}(d(\cdot,\cdot) \geq \alpha)\mathrm{d}\alpha$

Multivariate case: filter-based pseudometric

Def: [Dey Mémoli Wang *SoCG* 2017]:
The *filter-based pseudometric* $d_f : M \times M \to \mathbb{R}$ is defined as

$$d_f(x, x') = \inf_{\gamma \in \Gamma(x,x')} \mathrm{diam}_Y(f \circ \gamma),$$

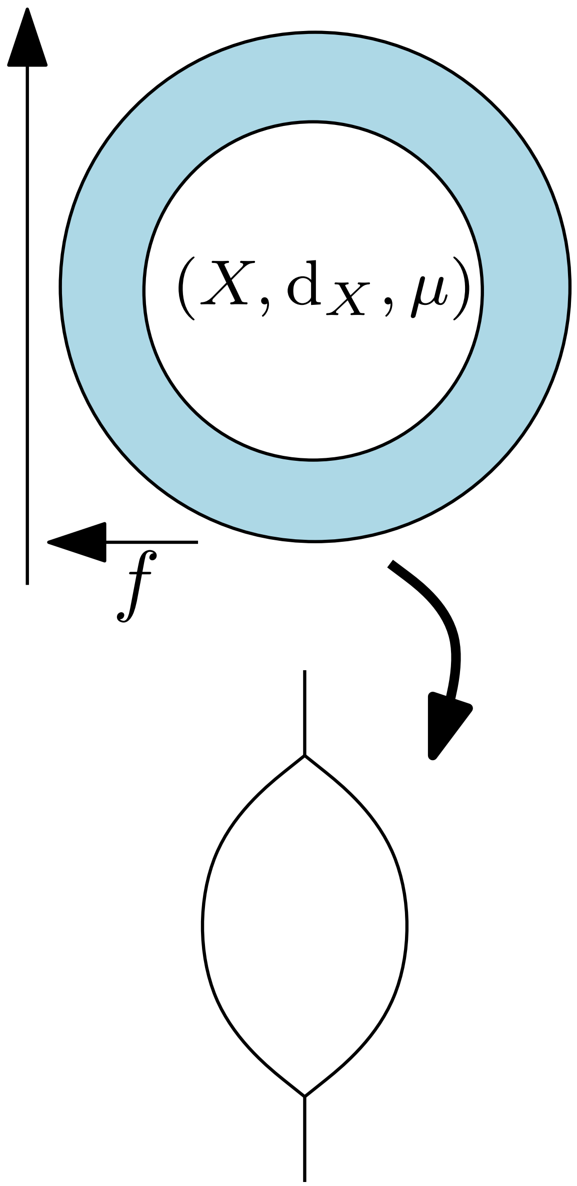
where $\Gamma(x, x')$ denotes the set of all continuous paths $\gamma : [0,1] \to M$ such that $\gamma(0) = x$ and $\gamma(1) = x'$, and diam_Y denotes the *diameter* of a subset of Y

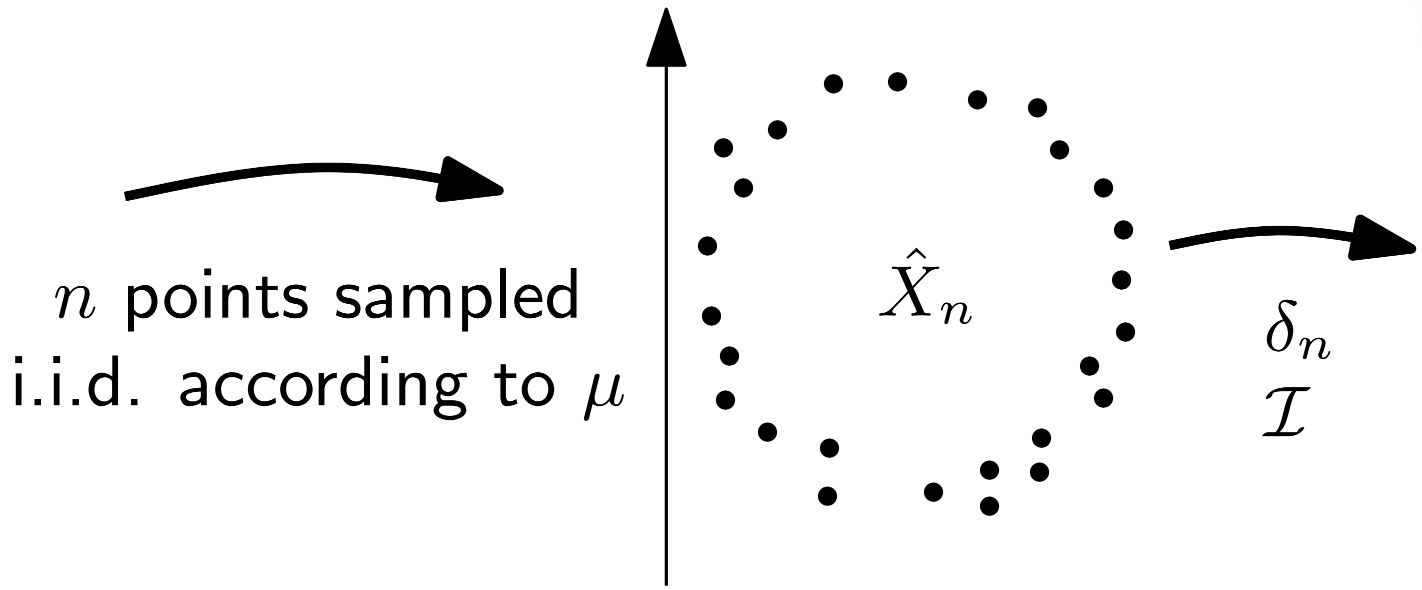
Def:
The *Gromov-Hausdorff* metric d_{GH} between $(M, d_f), (M', d_{f'})$ is defined as

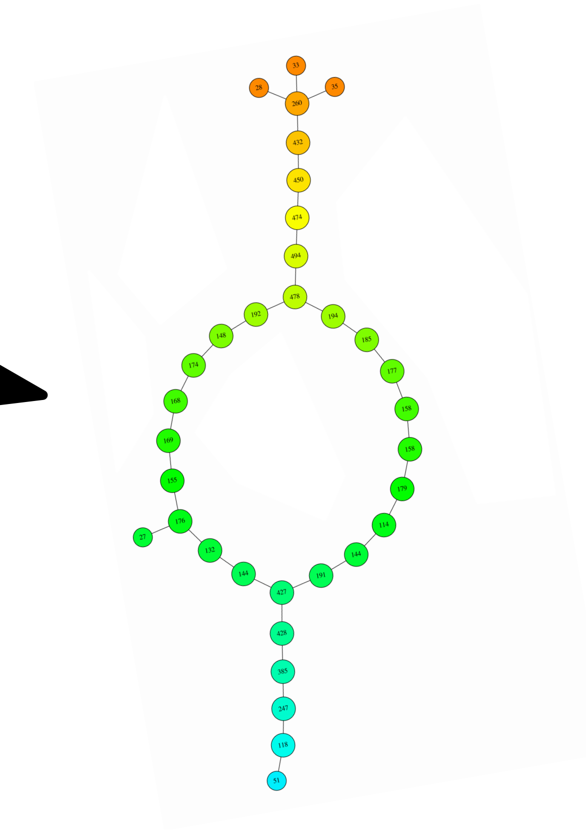
$$d_{\mathrm{GH}}(M, M') = \frac{1}{2} \inf_C \sup_{(x,x'),(y,y') \in C} |d_f(x,y) - d_{f'}(x', y')|,$$

where C denotes the set of all correspondences between M and M' (subsets of $M \times M'$ s.t. projections onto M and M' are surjective)

Statistics for Mapper in general





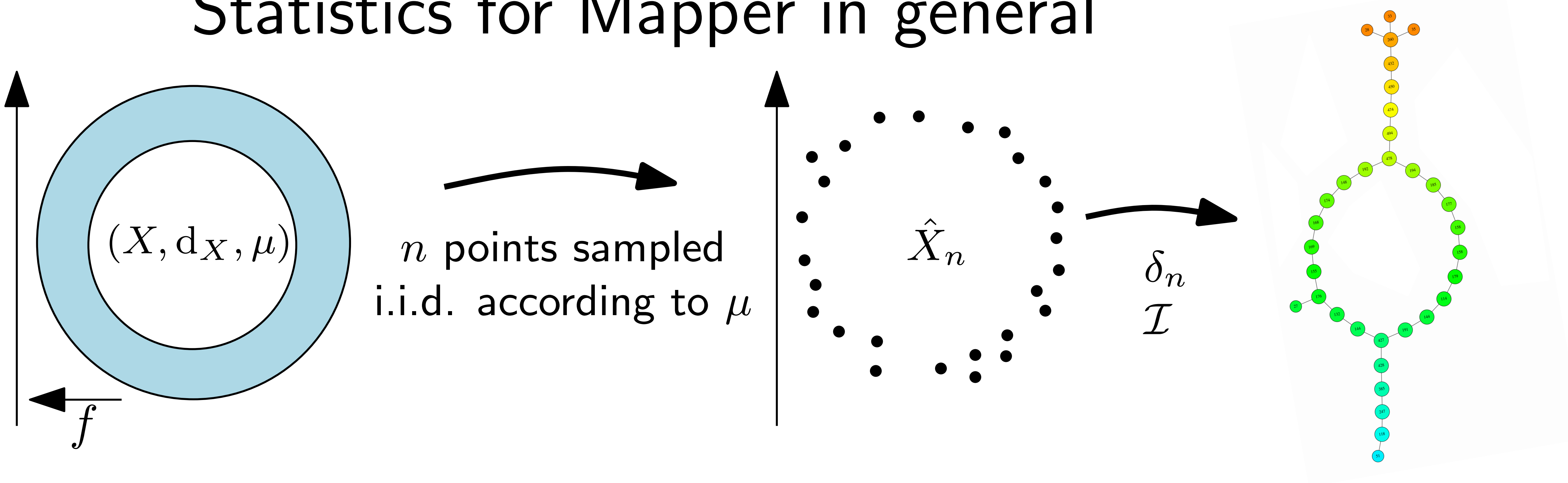


$$\mathbb{E}\left[d_{\mathrm{GH}}(\mathrm{M}_f, \mathrm{R}_f) \leq \,?\right] \geq 0.95$$

Question:

How to assess distance confidence?

Statistics for Mapper in general



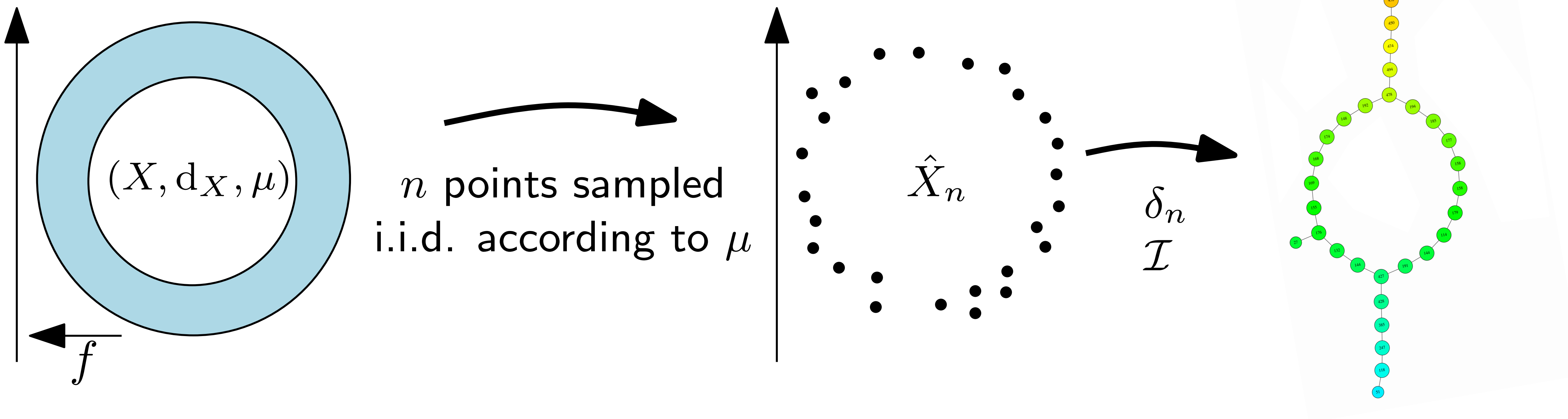
Thm: [C. Michel *Preprint* 2020]
If μ and $f\#\mu$ are (a,b)-standard, then for δ_n as before, one has:

$$\mathbb{E}\left[d_{\mathrm{GH}}(\mathrm{M}^{\bullet}_{f,\delta_n}(\hat{X}_n, \mathcal{I}), \mathrm{R}_f(X))\right] \leq 5 \cdot \mathbb{E}\left[\mathrm{res}(\mathcal{I})\right] + C\omega\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b},$$

where C depends only on a, b, and res denotes the *resolution* of the cover $\mathcal{I}$, i.e., the diameter of its elements

Moreover, using covers with hypercubes or K-means, or quantized Distance-to-Measure [Brecheteau Levrard *Bernouilli* 2020] allows to bound $\mathbb{E}\left[\mathrm{res}(\mathcal{I})\right]$.

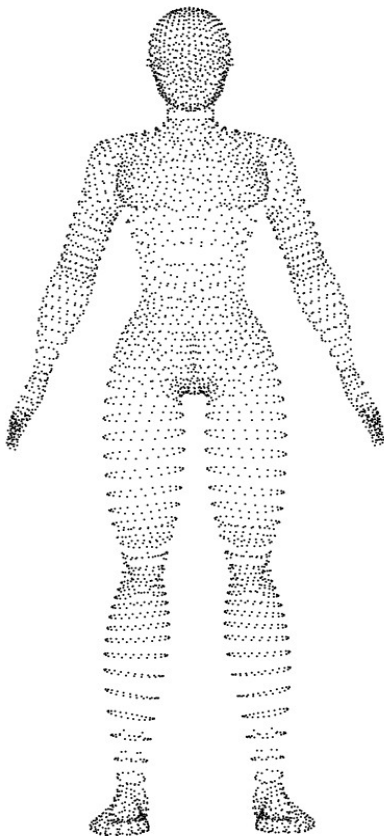
Statistics for Mapper in general

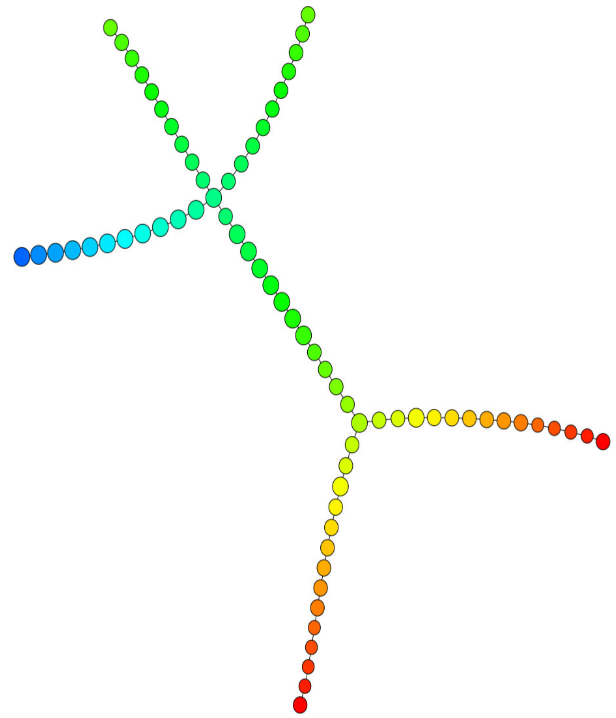


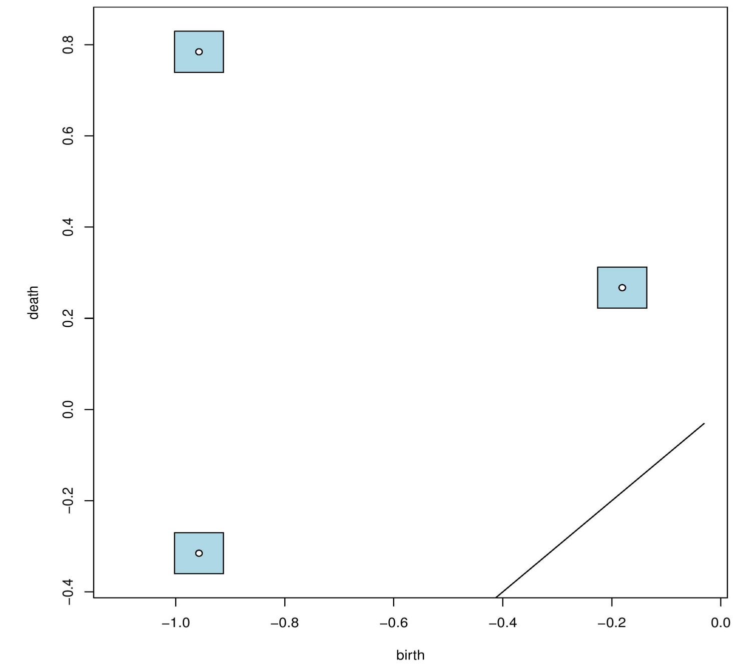
Thm: [C. Michel *Preprint* 2020]
If $w(u) \leq cu^\gamma$ for some $c > 0, \gamma \in (0,1)$, and for a cover $\mathcal{I}$ given by thickening a K-means partition in $\mathbb{R}^D$:

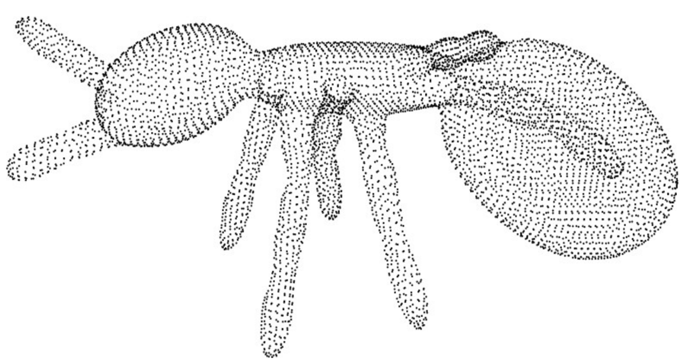
$$\mathbb{E}\left[\mathrm{res}(\mathcal{I})\right] \leq K^{-(2\gamma^2)/(2\gamma b + b^2)} + \left(\frac{KD}{n}\right)^{\gamma/(2b+4\gamma)}$$

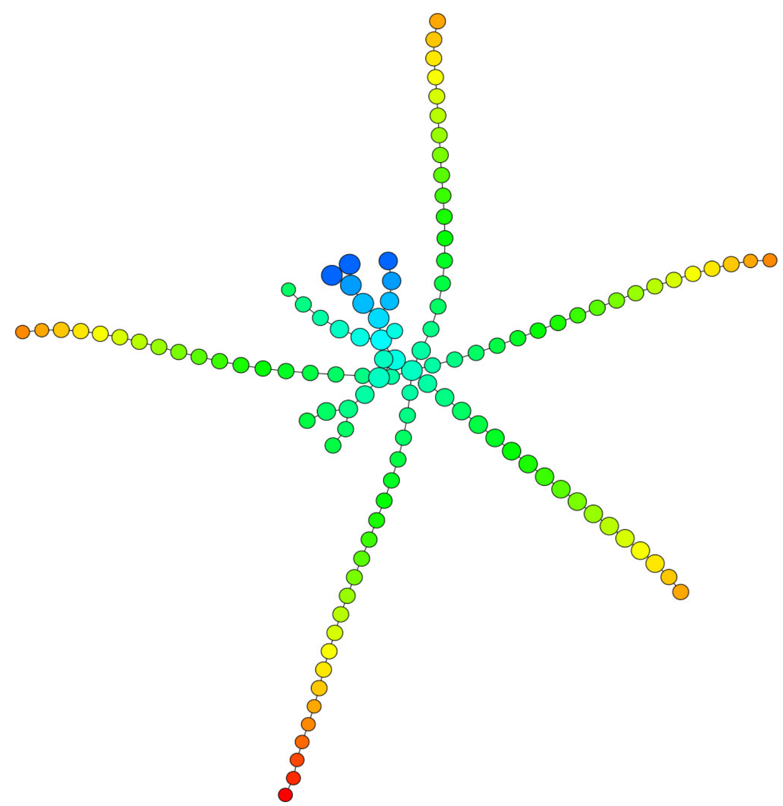
Experiments 85% confidence intervals

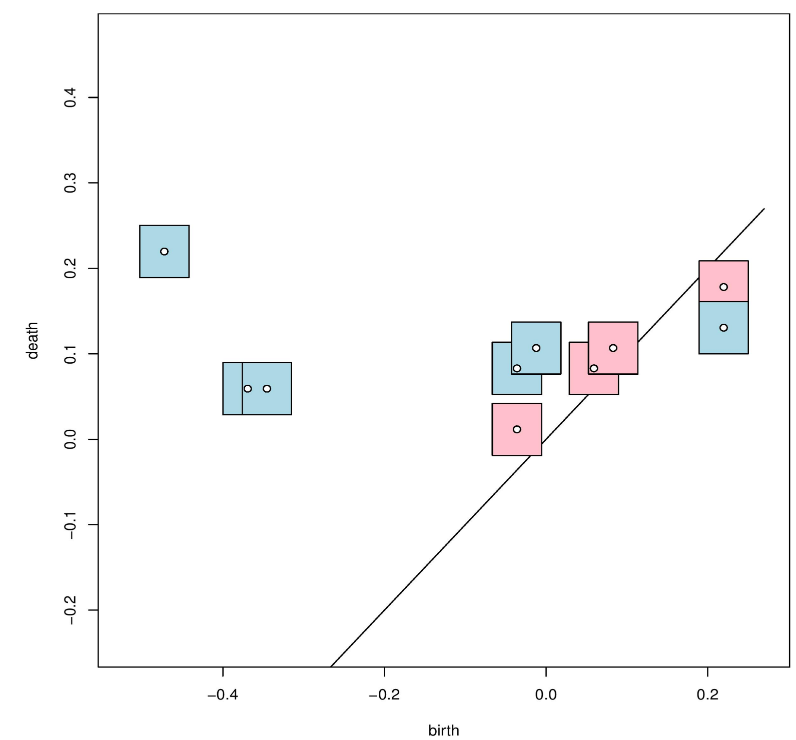




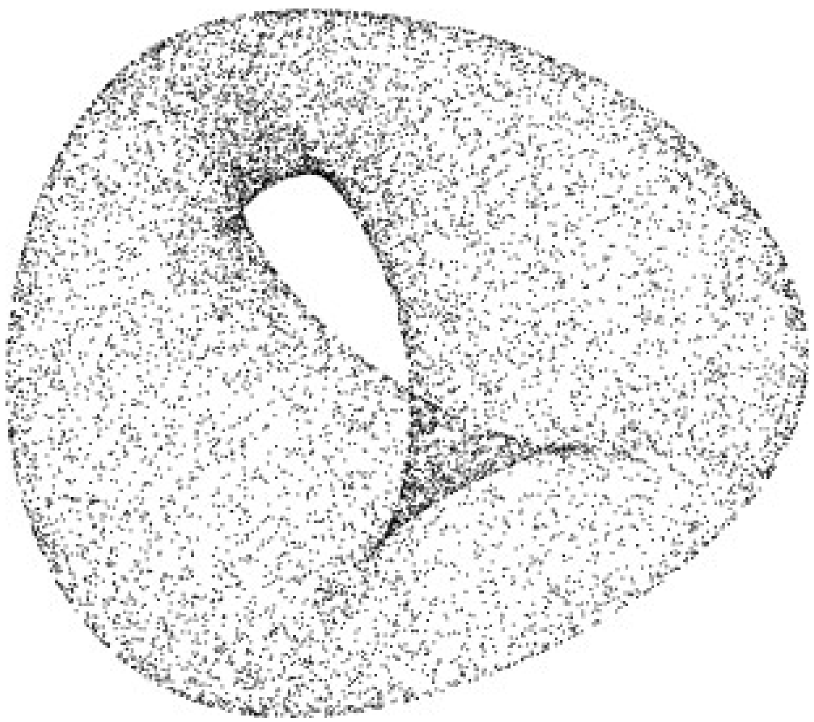


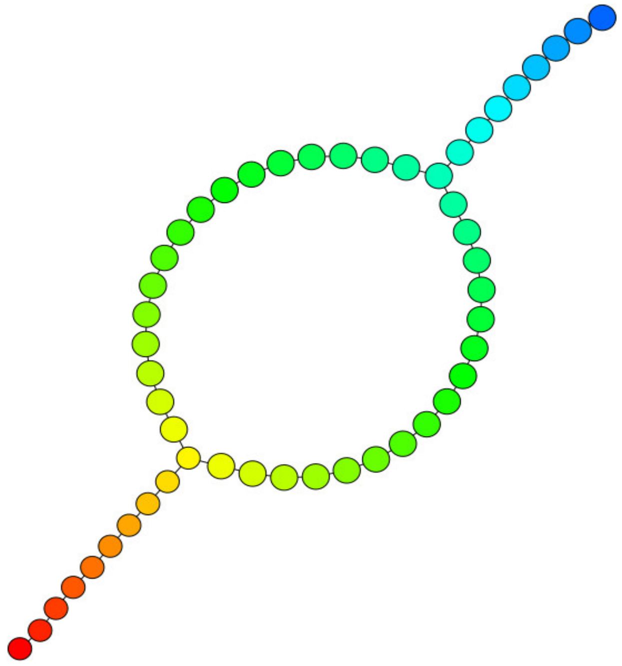


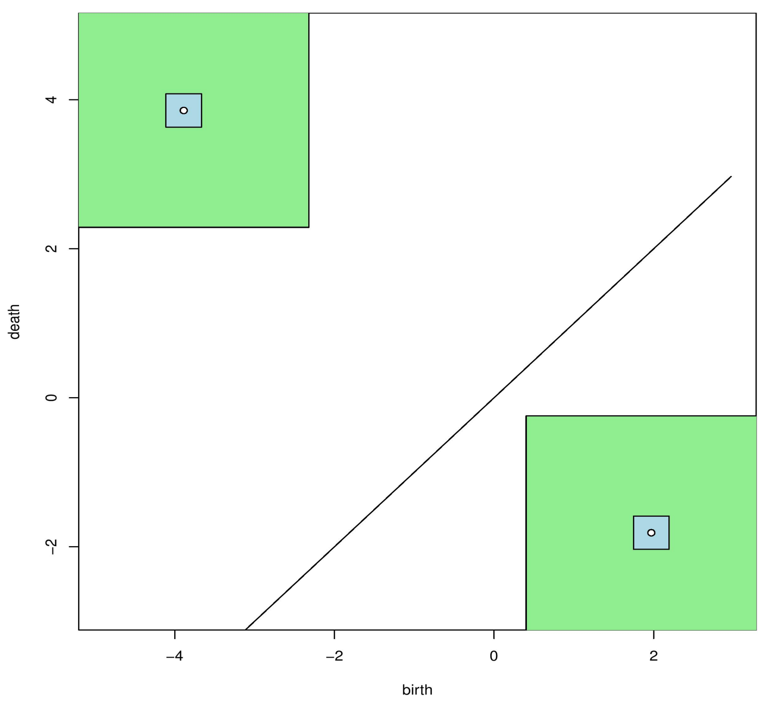


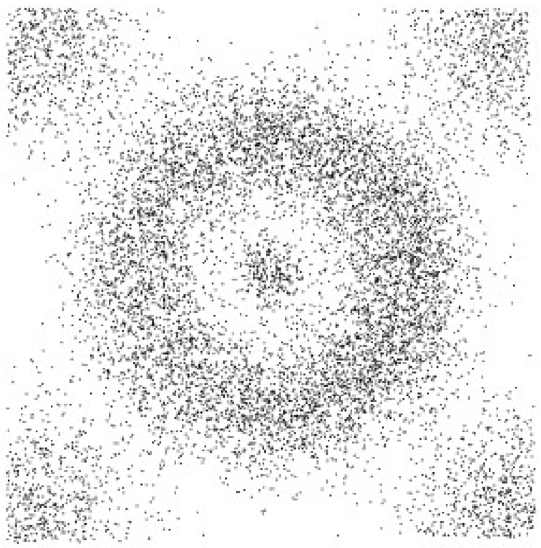


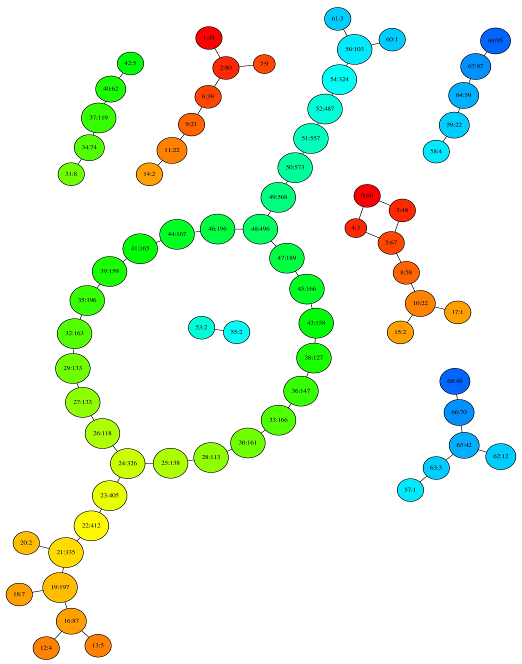
Experiments 85% confidence intervals

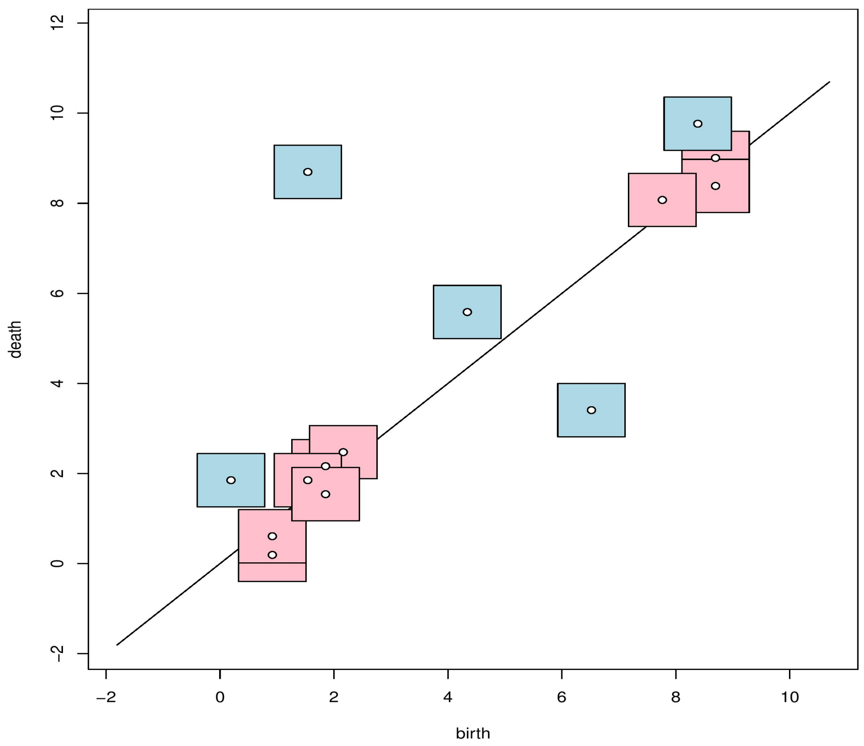




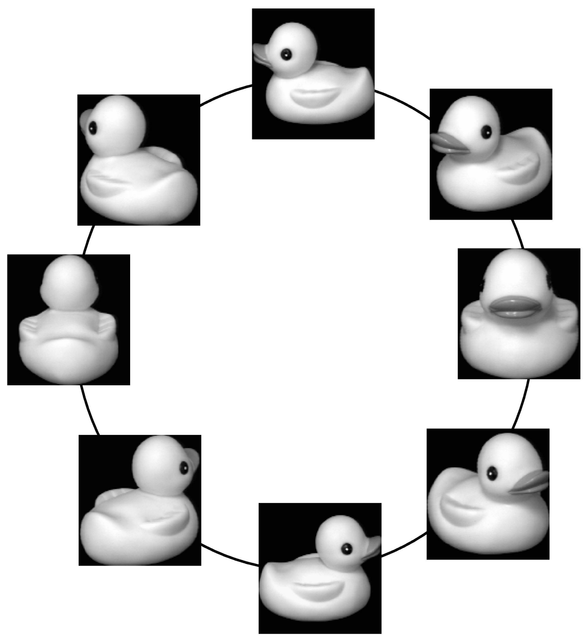


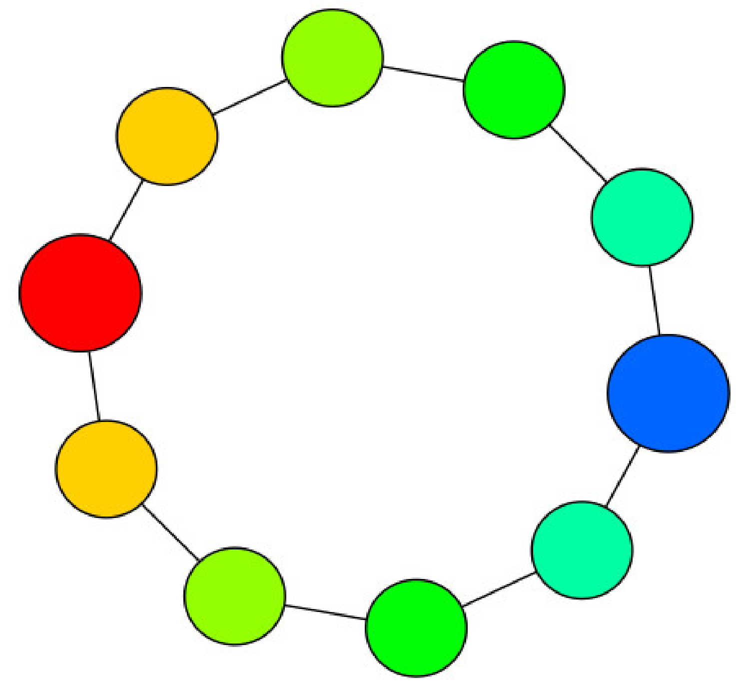


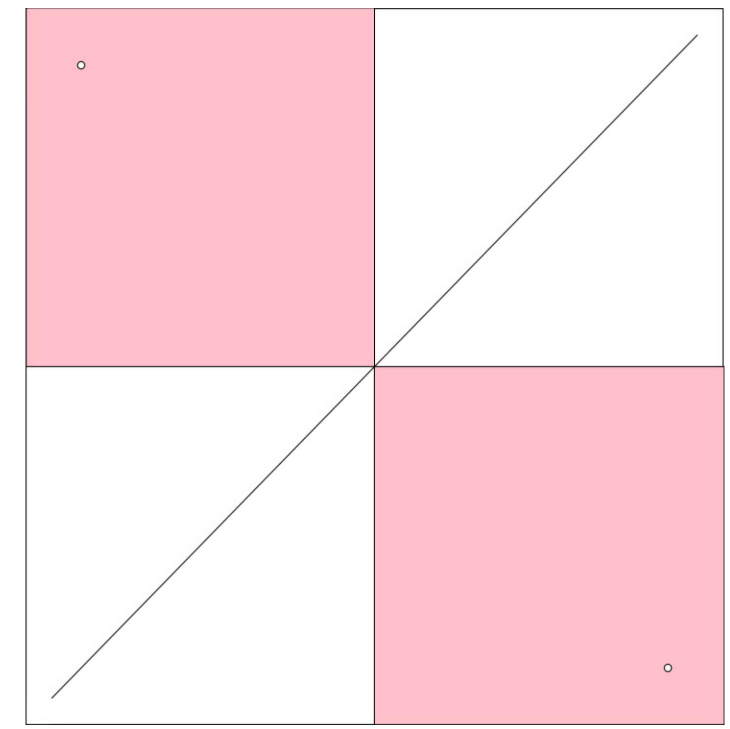


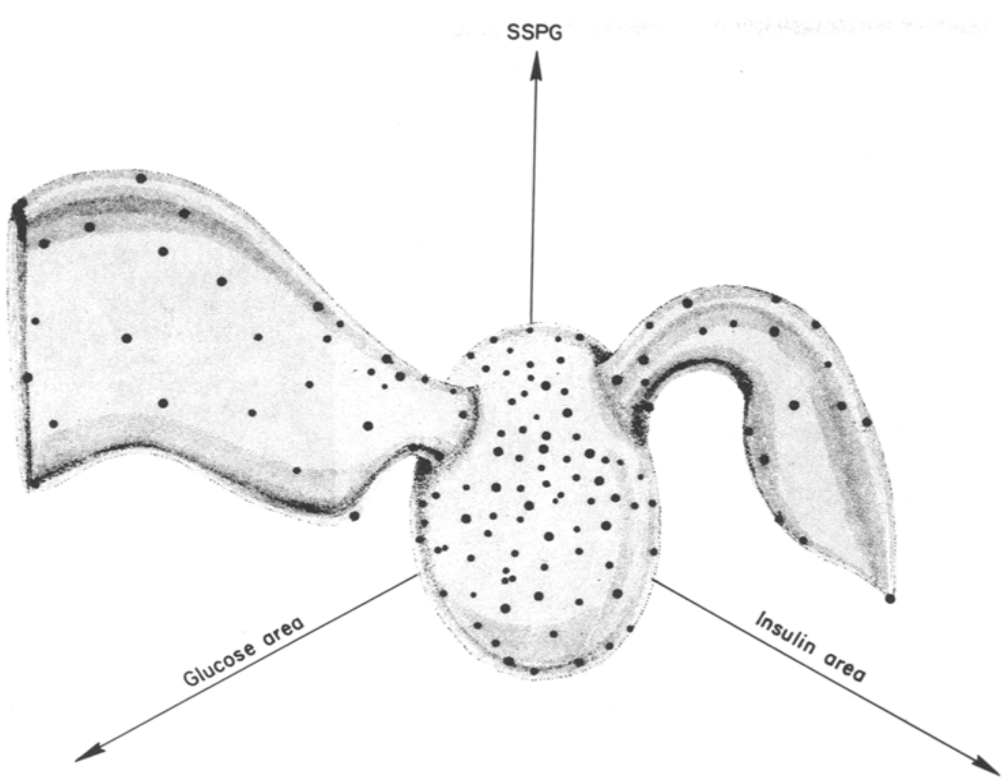


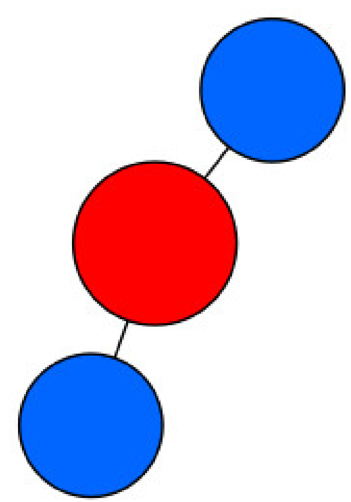
Experiments 85% confidence intervals

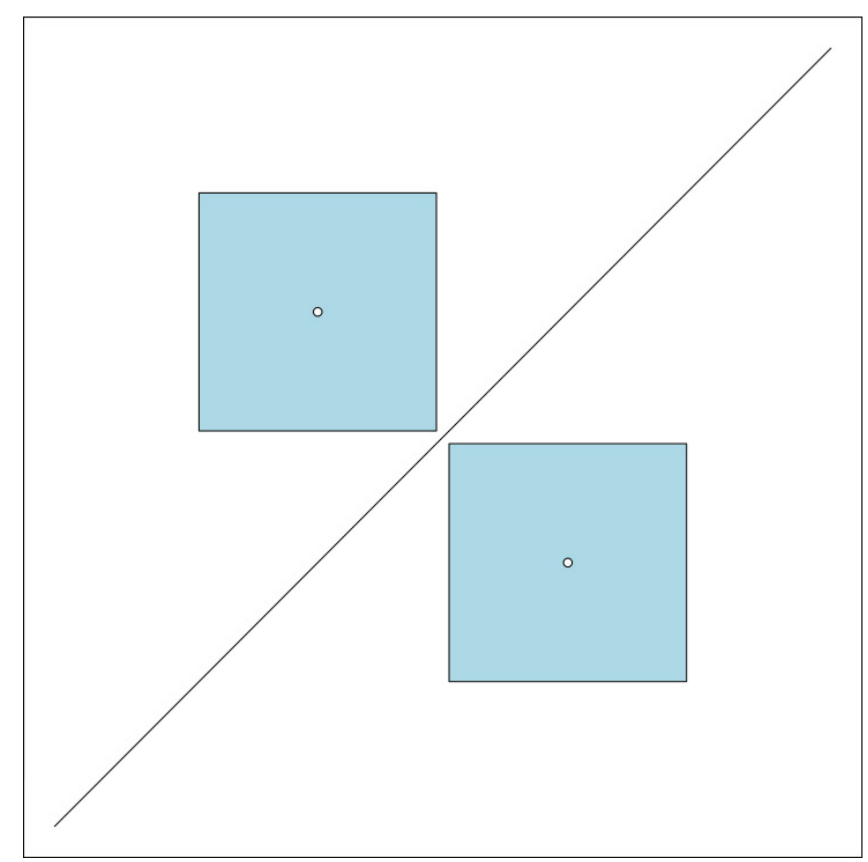






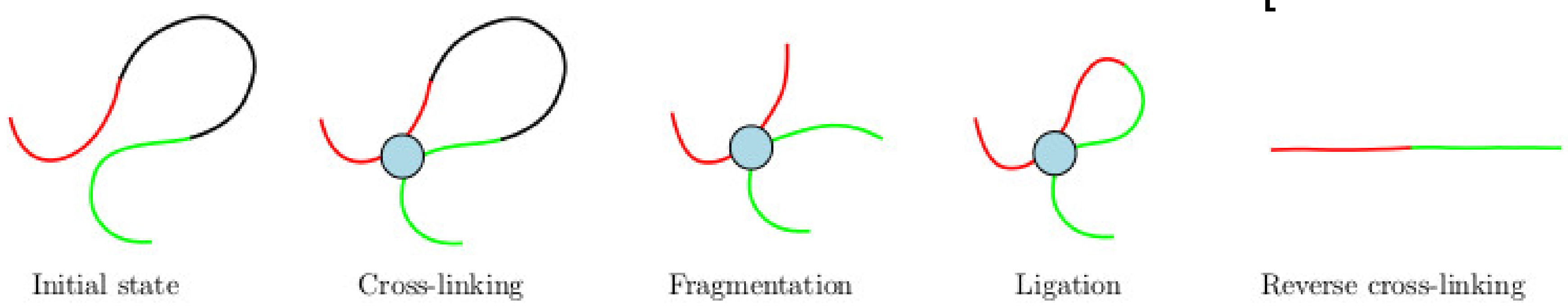


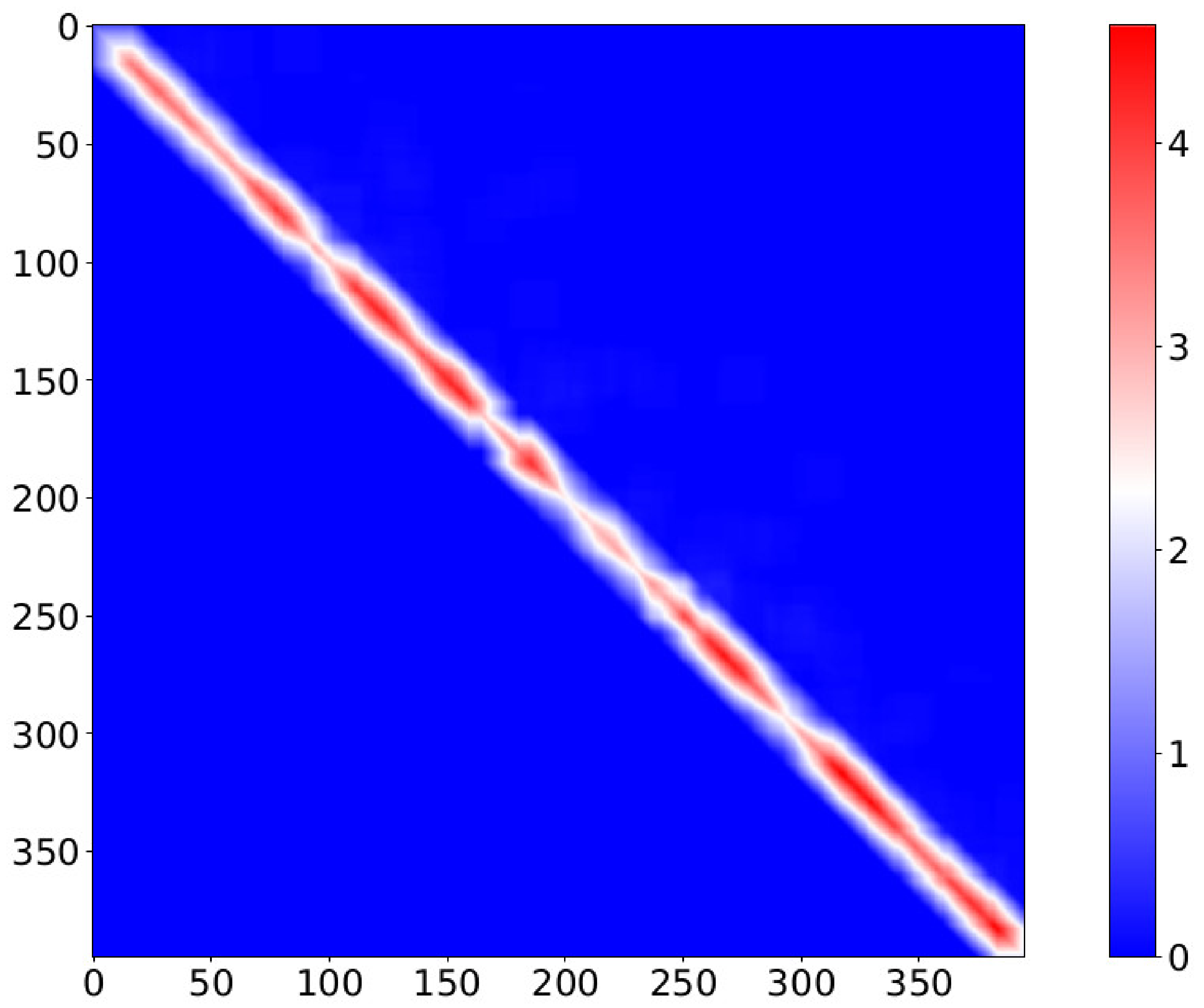




Experiments   Chromosome conformation capture

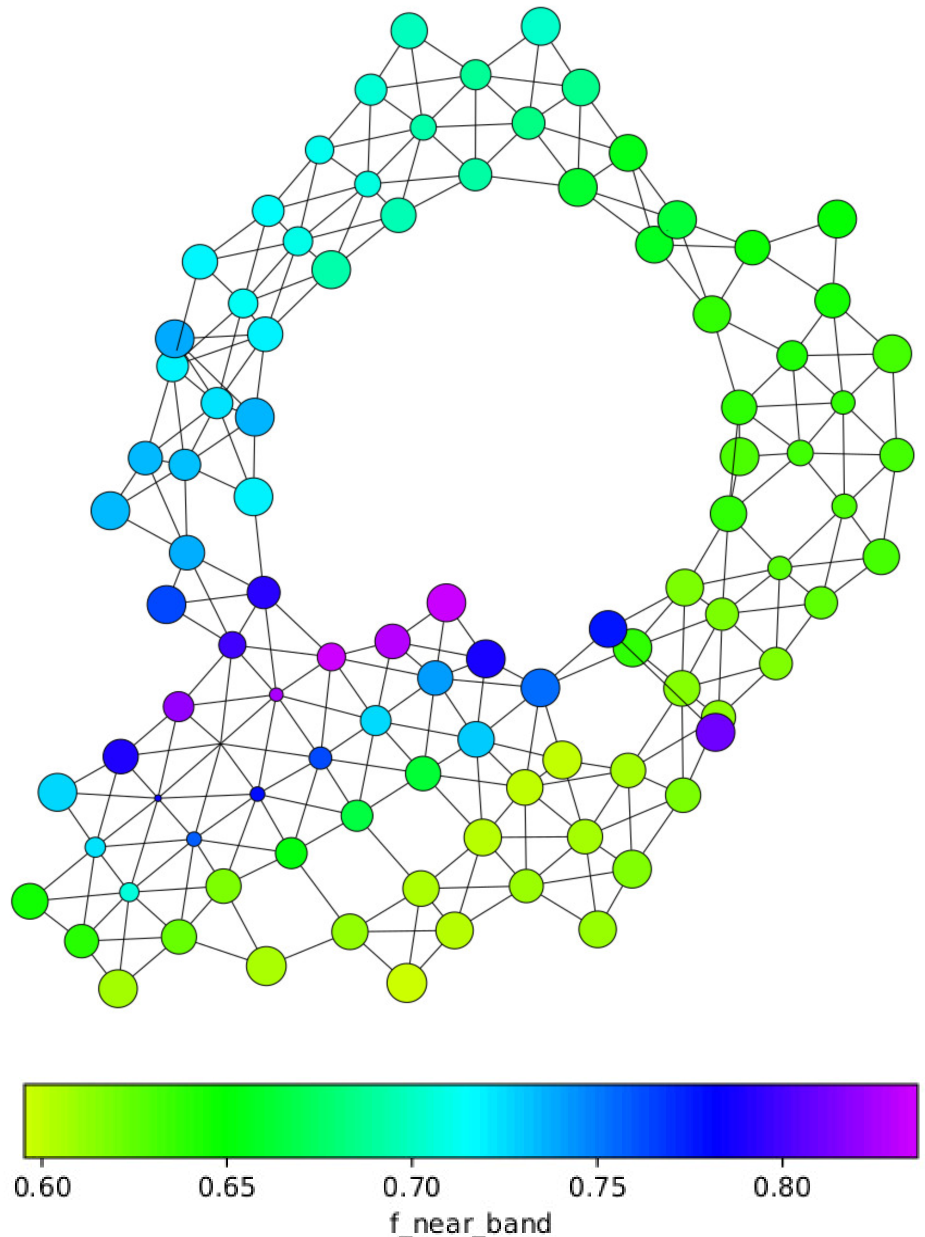
[C. Rabadan *Abel* 2018]

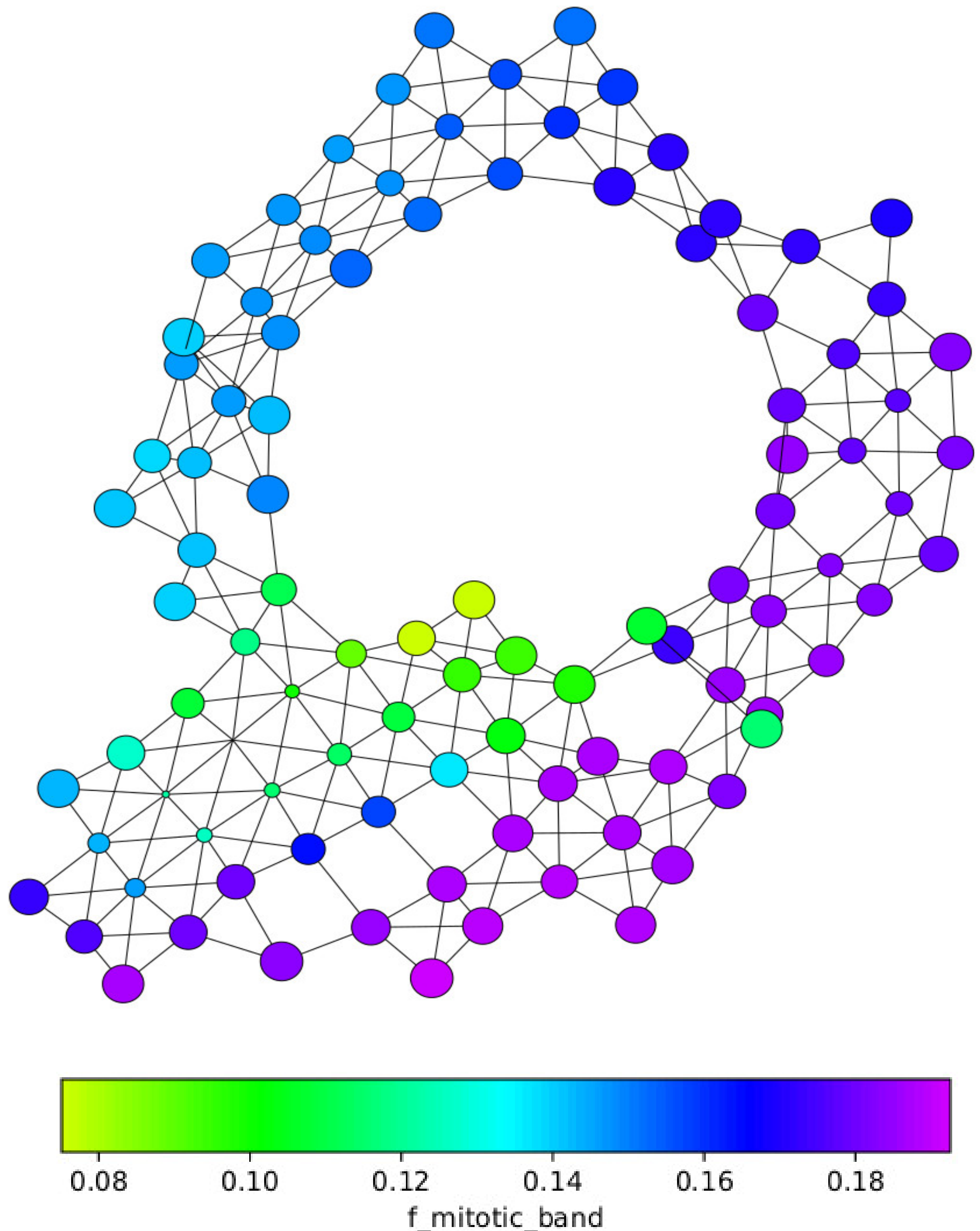




Experiments Chromosome conformation capture

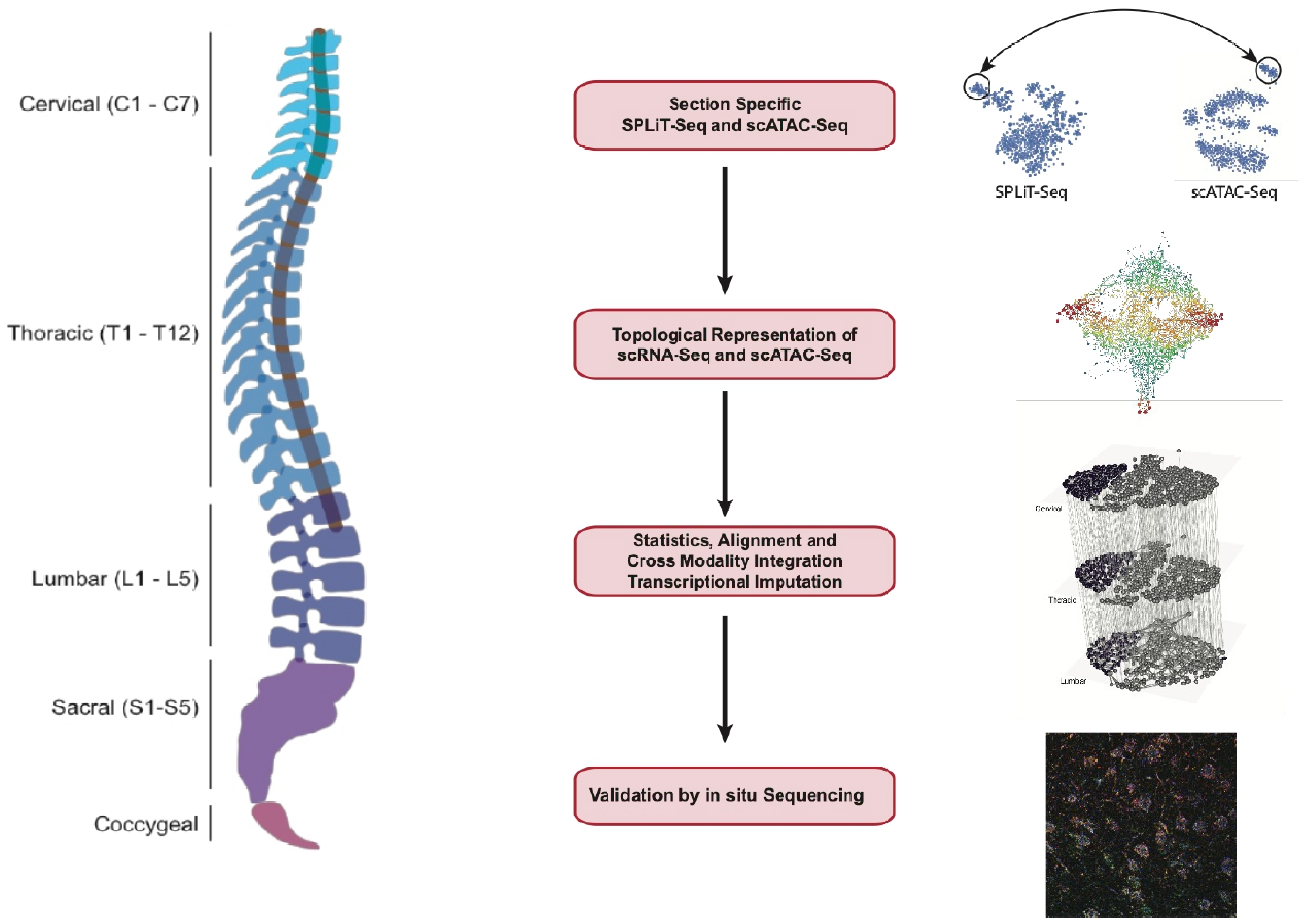
[C. Rabadan *Abel* 2018]



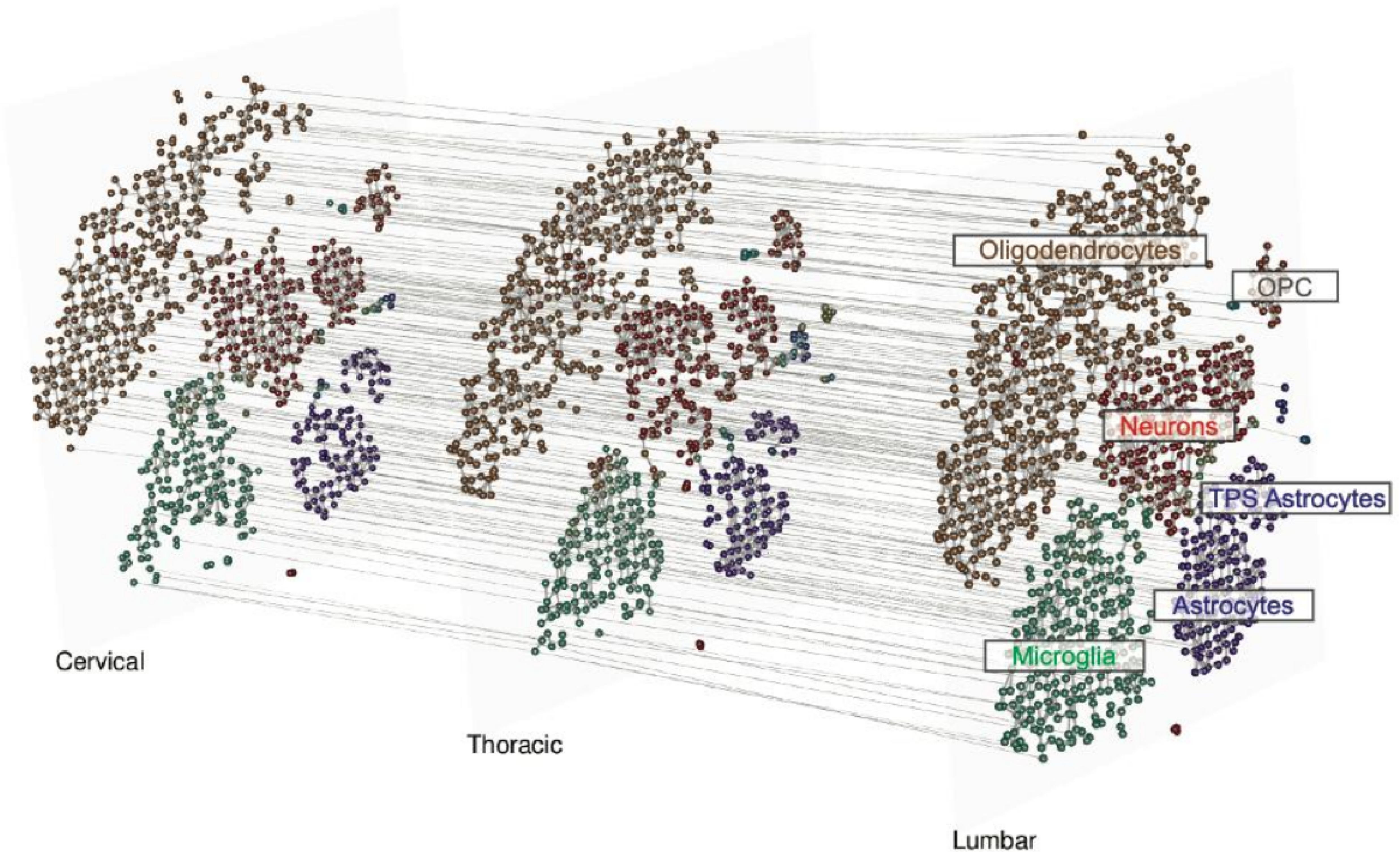


Formal identification of cell cycle with 95% confidence

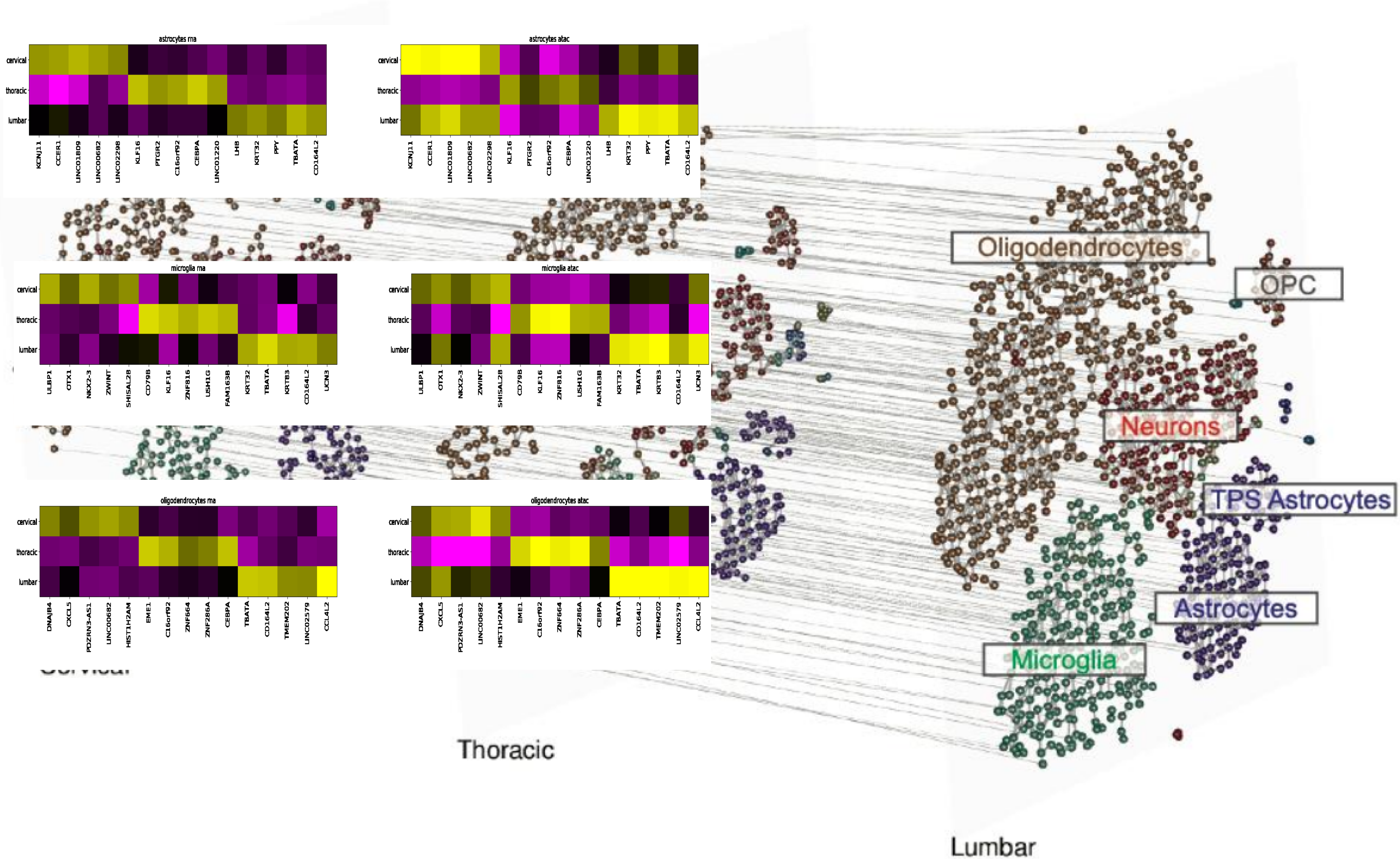
Experiments   Spinal cord data   Joint work with Rizvi Rabadan 2020



Experiments Spinal cord data Joint work with Rizvi Rabadan 2020

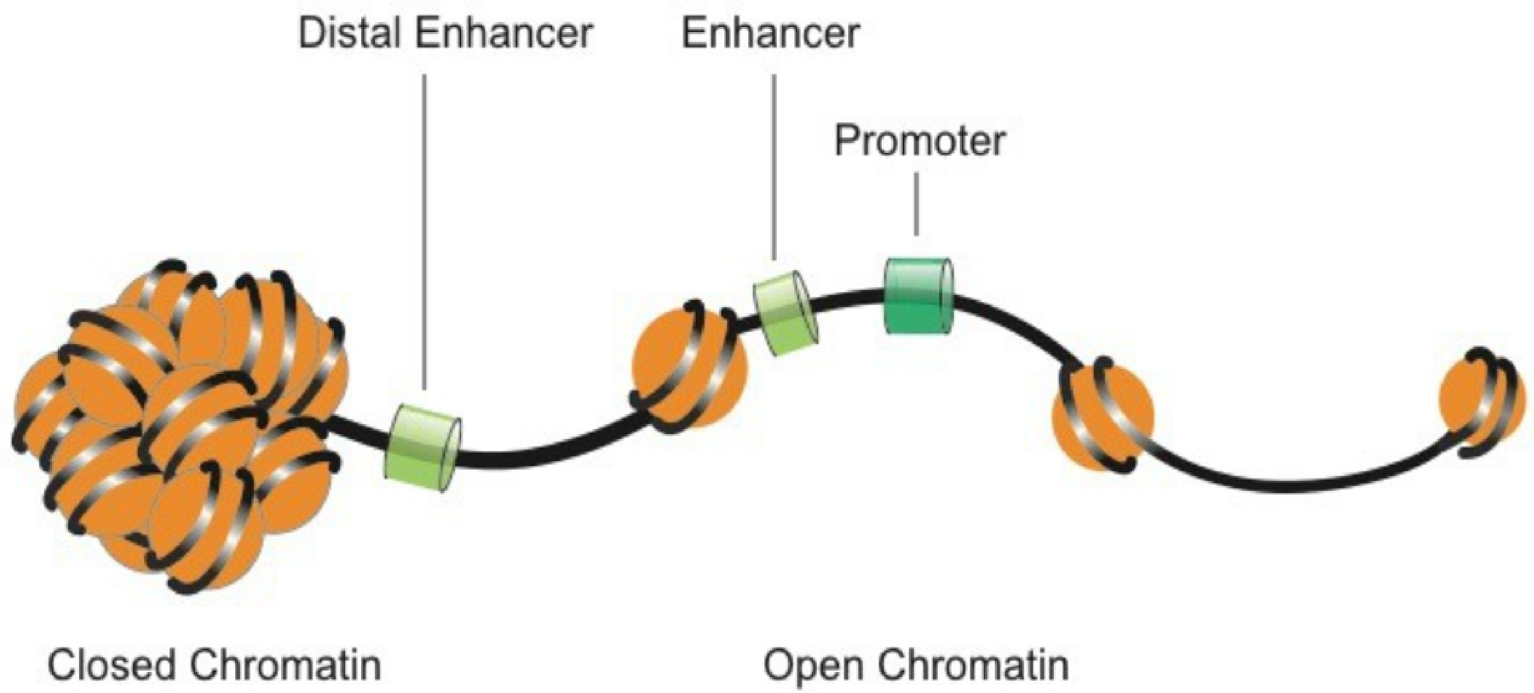


Experiments Spinal cord data Joint work with Rizvi Rabadan 2020



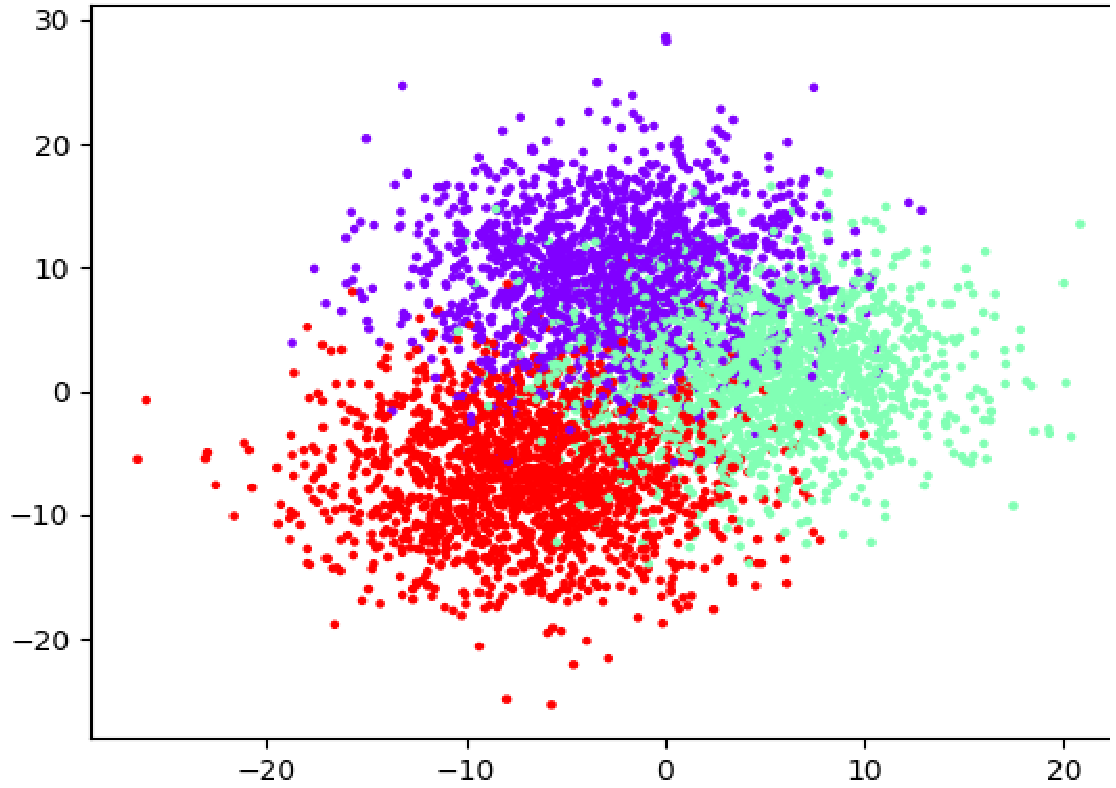
Experiments Spinal cord data Joint work with Rizvi Rabadan 2020

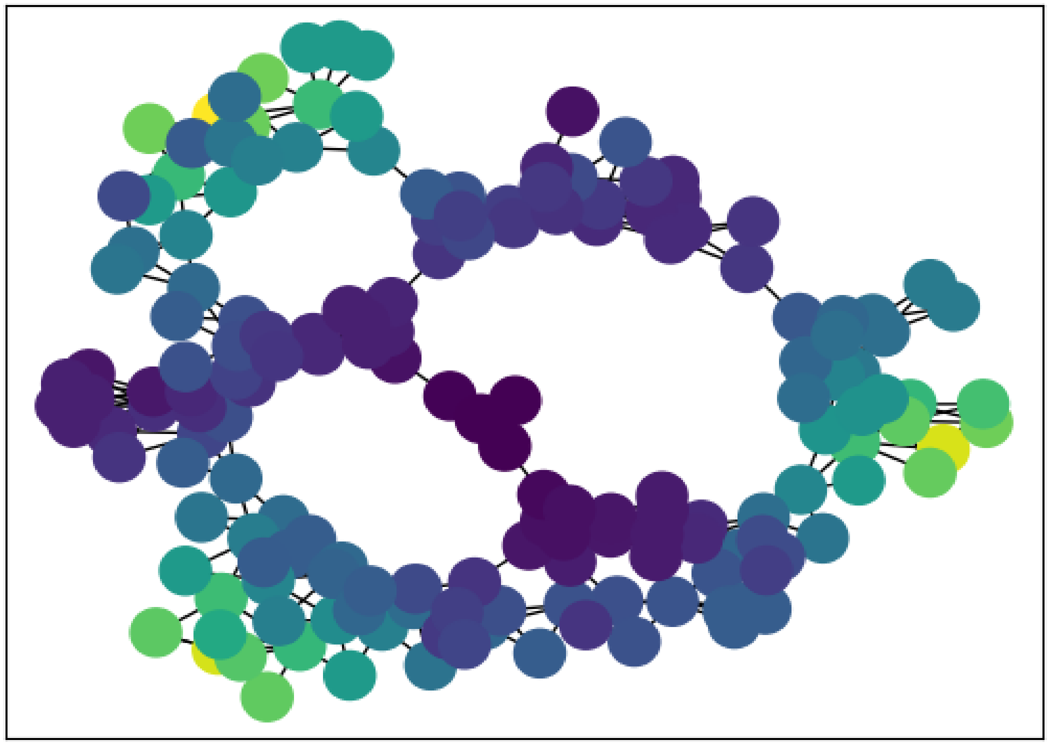
Gene expression (SPLiTseq) and gene accessibility (ATACseq) of single cells of one healthy individual for 3 sections of spinal cord



Experiments Machine learning classifier

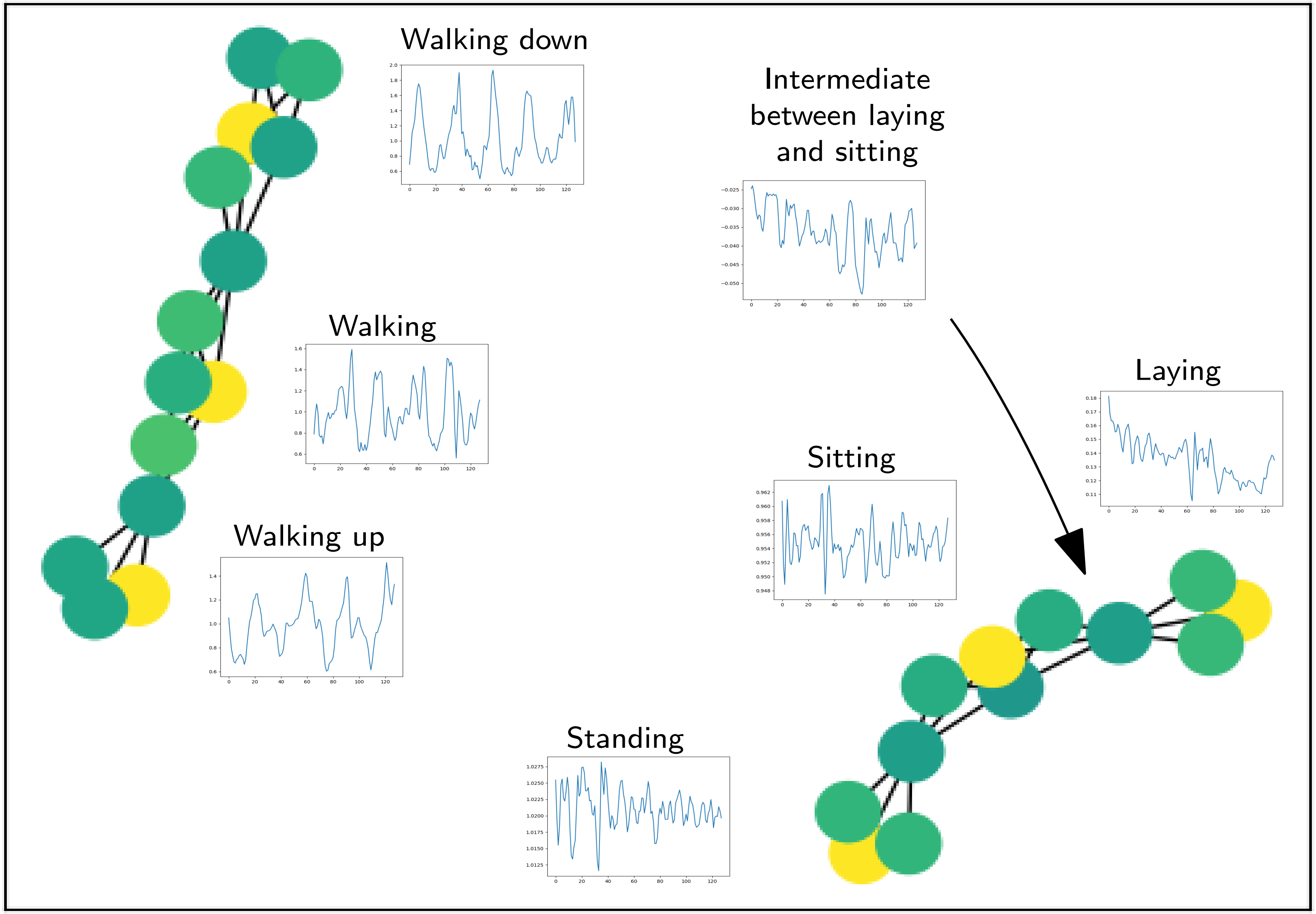
Filter = confidence of Random Forest classifier (in $\mathbb{R}^3$)





Experiments   Machine learning classifier

Filter = confidence of Random Forest classifier (in $\mathbb{R}^6$)



Thanks!!