2nd Workshop on Topological Methods in Data Analysis
4th - 6th October 2021, Heidelberg University

Topological exploratory data analysis: nerves, Mappers and robust inference

Mathieu Carrière
INRIA Sophia-Antipolis
mathieu.carriere@inria.fr

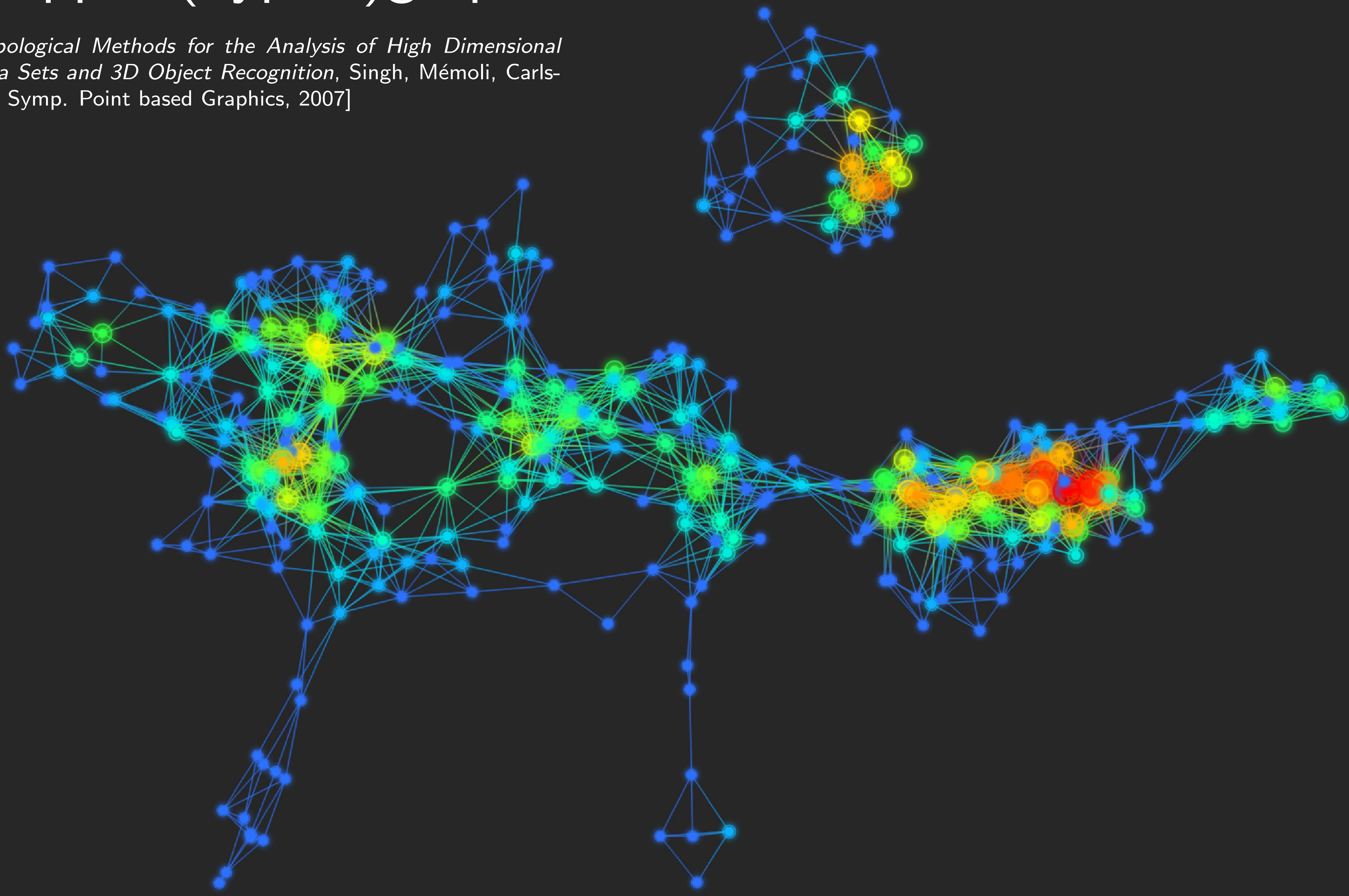
[*Structure and Stability of the One-Dimensional Mapper*, Carrière, Oudot, Found. Comput. Math., 2018]

[*Statistical Analysis and Parameter Selection for Mapper*, Carrière, Michel, Oudot, J. Machine Learning Research, 2018]

[*Statistical analysis of Mapper for stochastic and multivariate filters*, Carrière, Michel, J. Preprint, 2020]

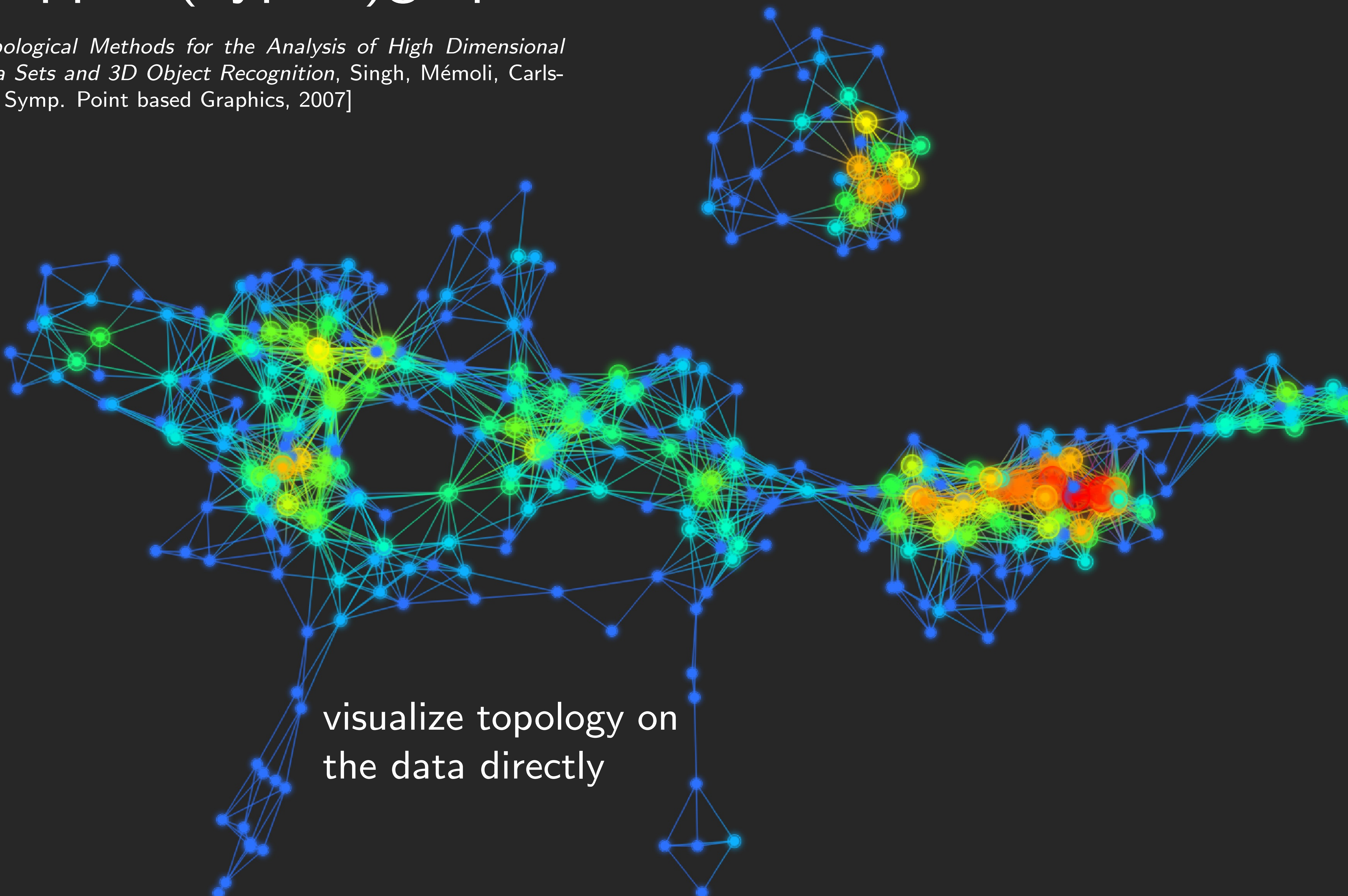
Mapper (hyper-)graphs

[*Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, Singh, Mémoli, Carlsson, Symp. Point based Graphics, 2007]



Mapper (hyper-)graphs

[*Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, Singh, Mémoli, Carlsson, Symp. Point based Graphics, 2007]



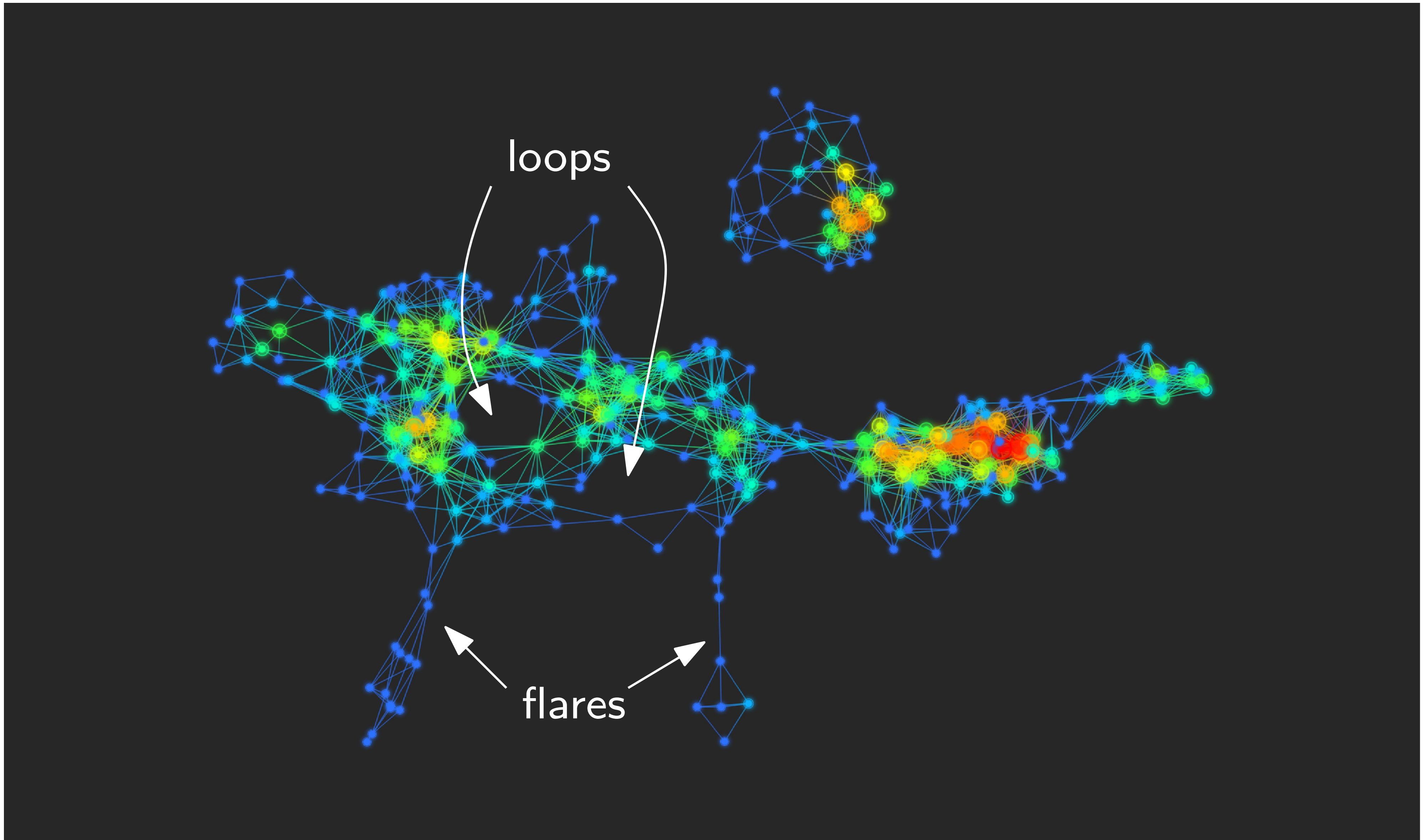
visualize topology on the data directly

Mapper in applications

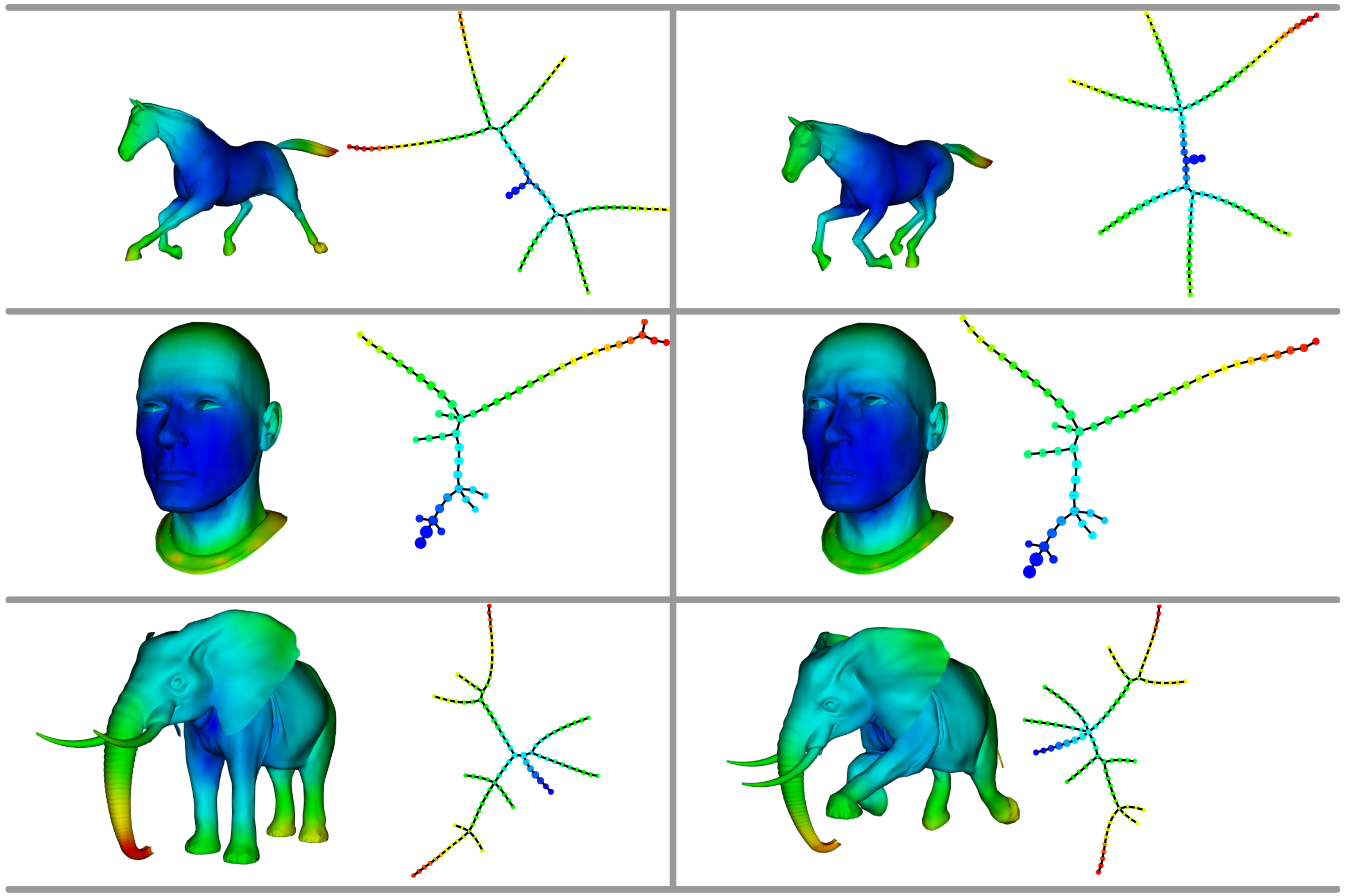
Two types of applications:

- clustering
- feature selection

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

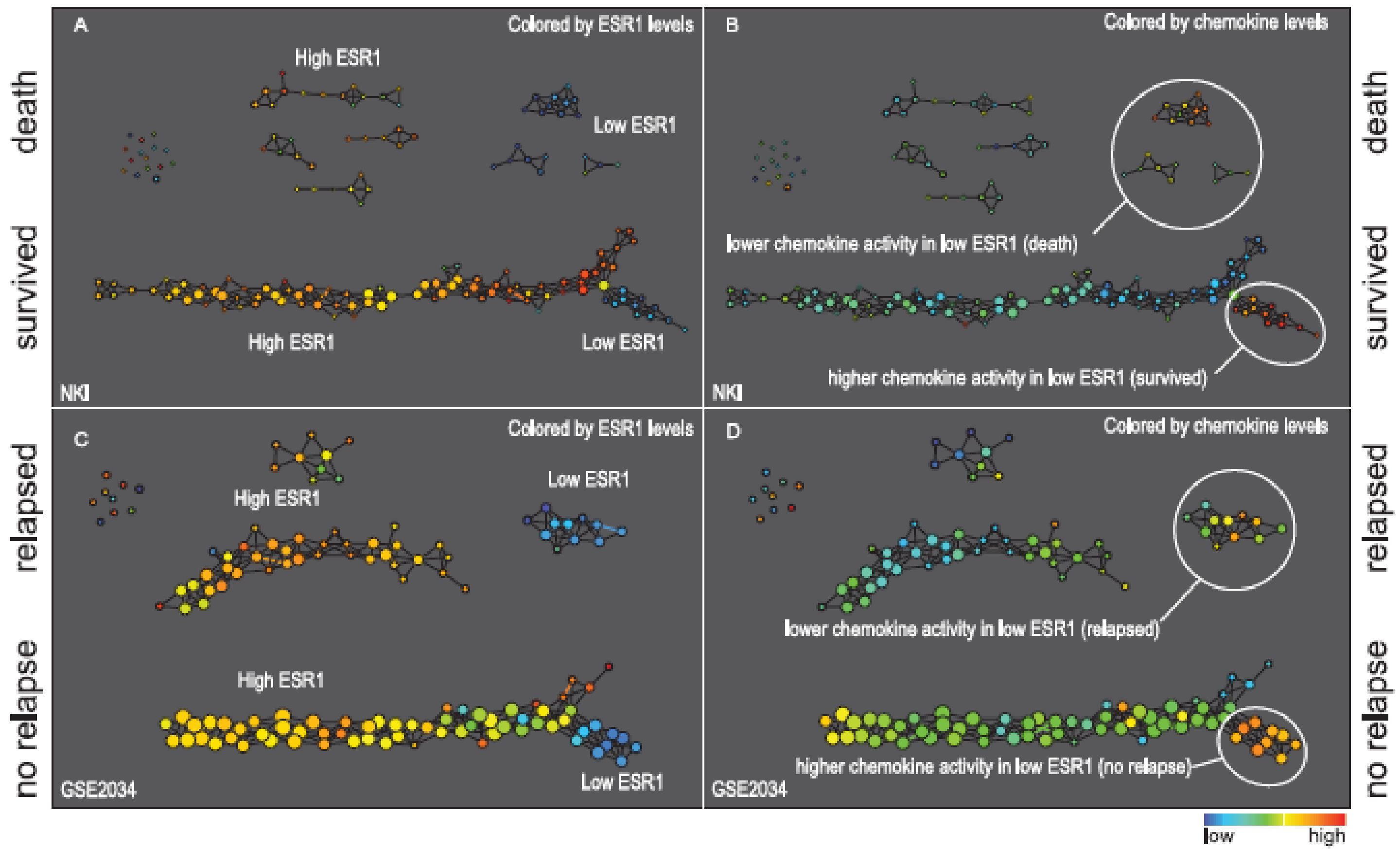


Mapper in applications



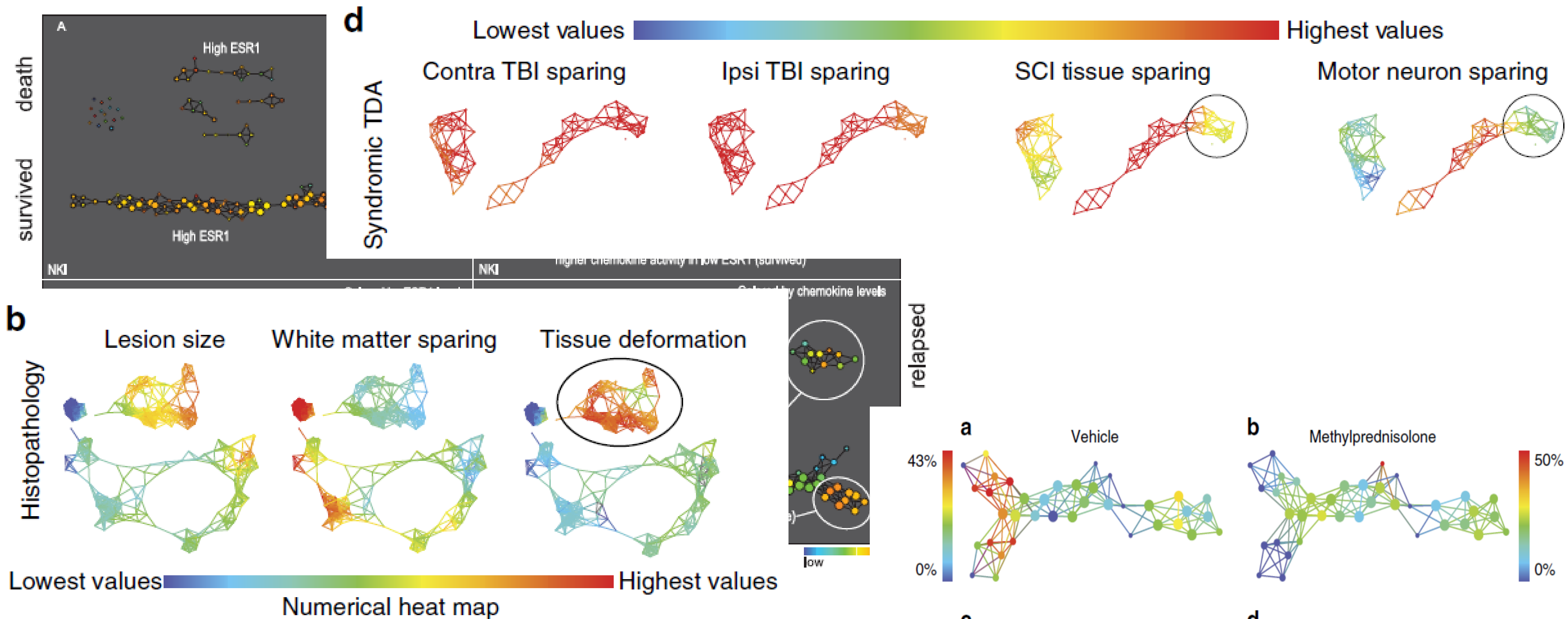
3d shapes classification

Mapper in applications

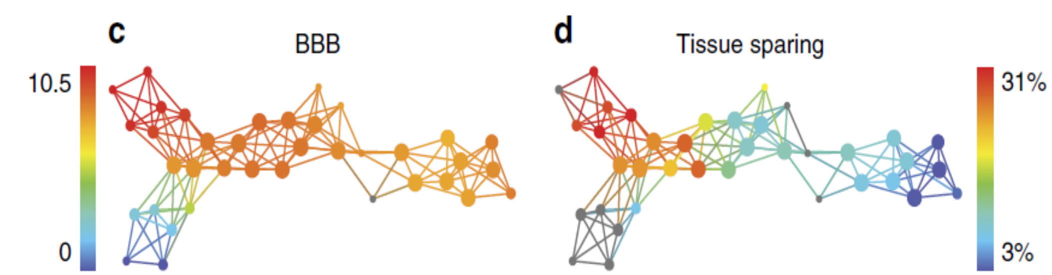


breast cancer subtype identification

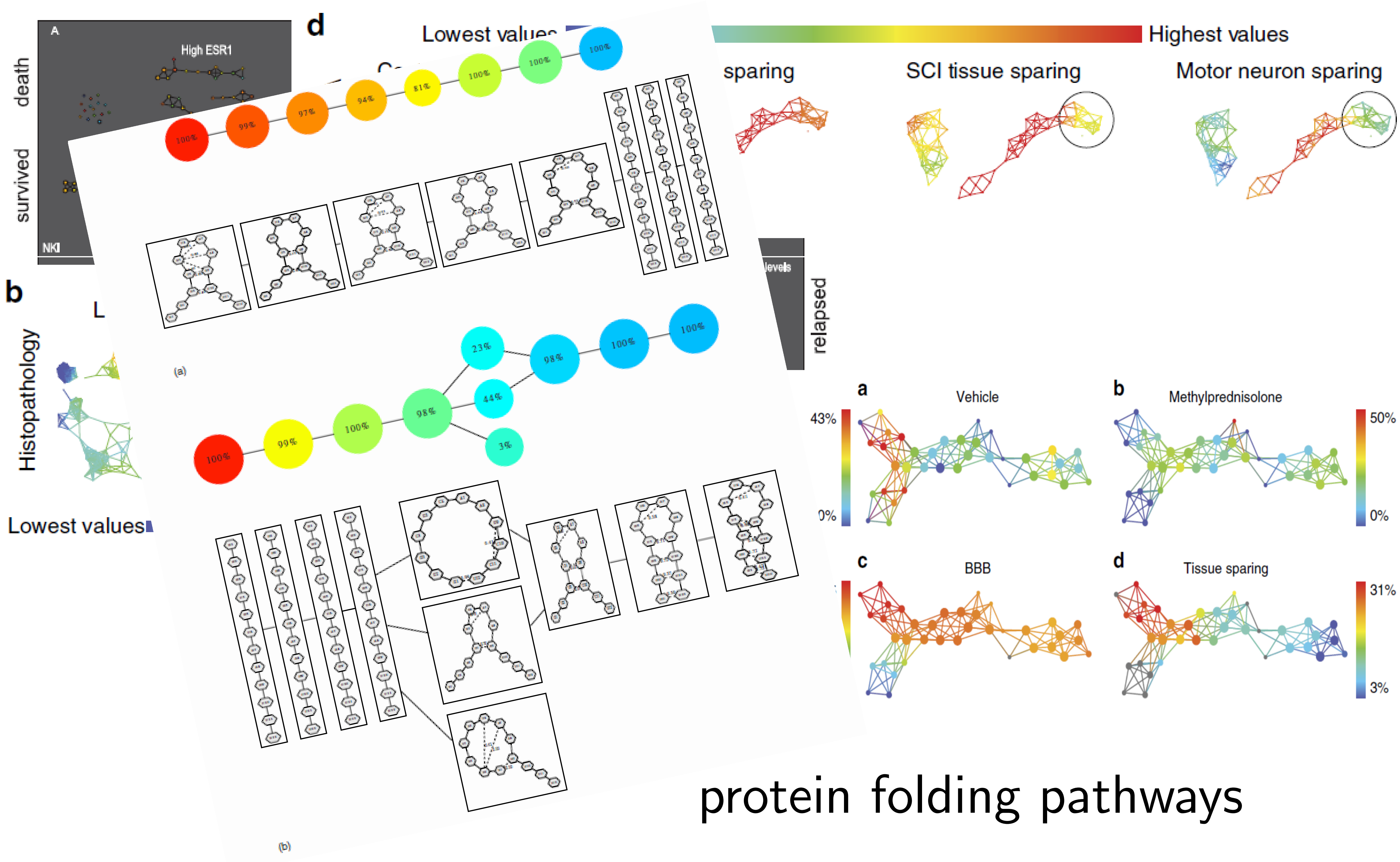
Mapper in applications



recovery from spinal cord injuries

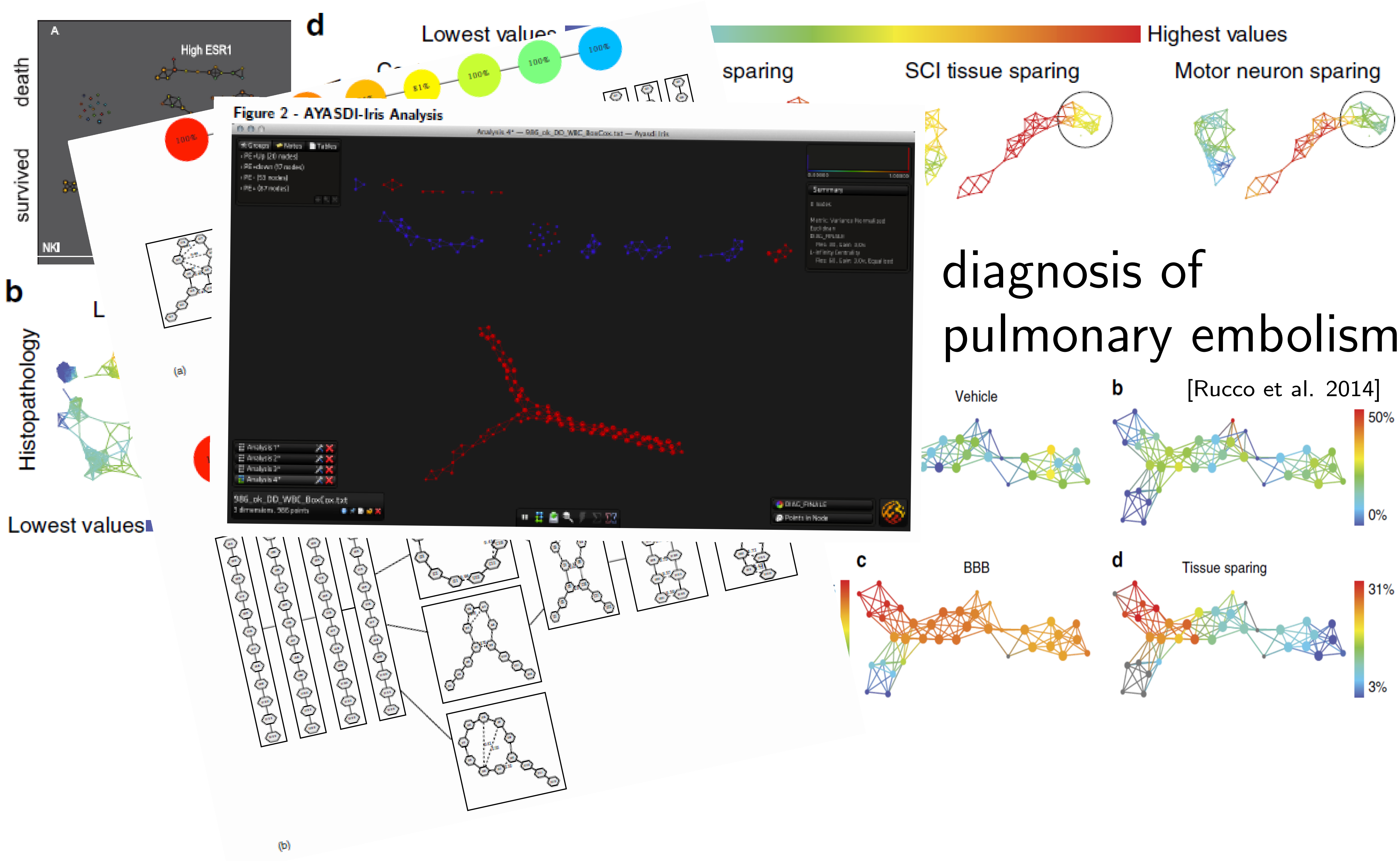


Mapper in applications



protein folding pathways

Mapper in applications

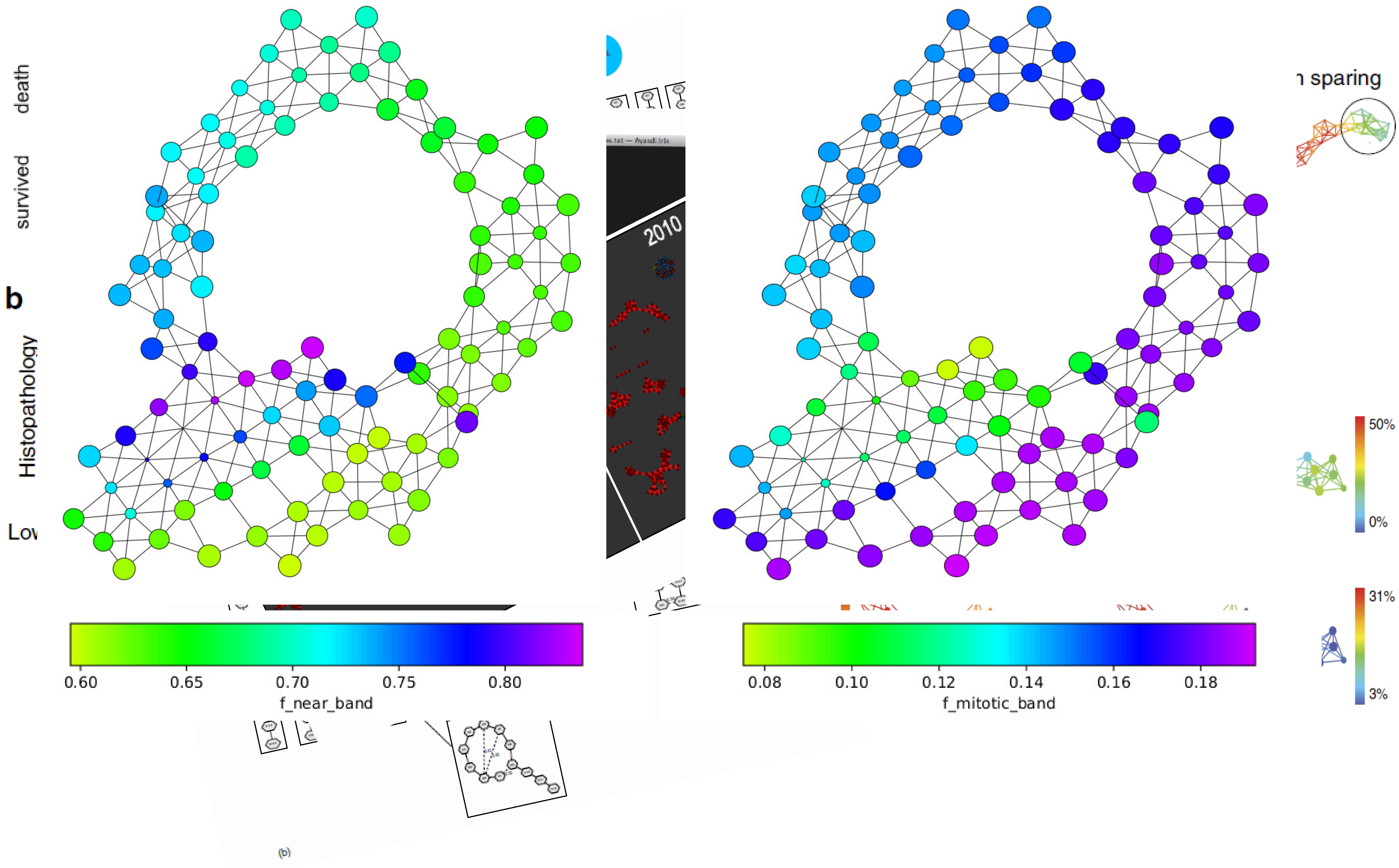


diagnosis of pulmonary embolism

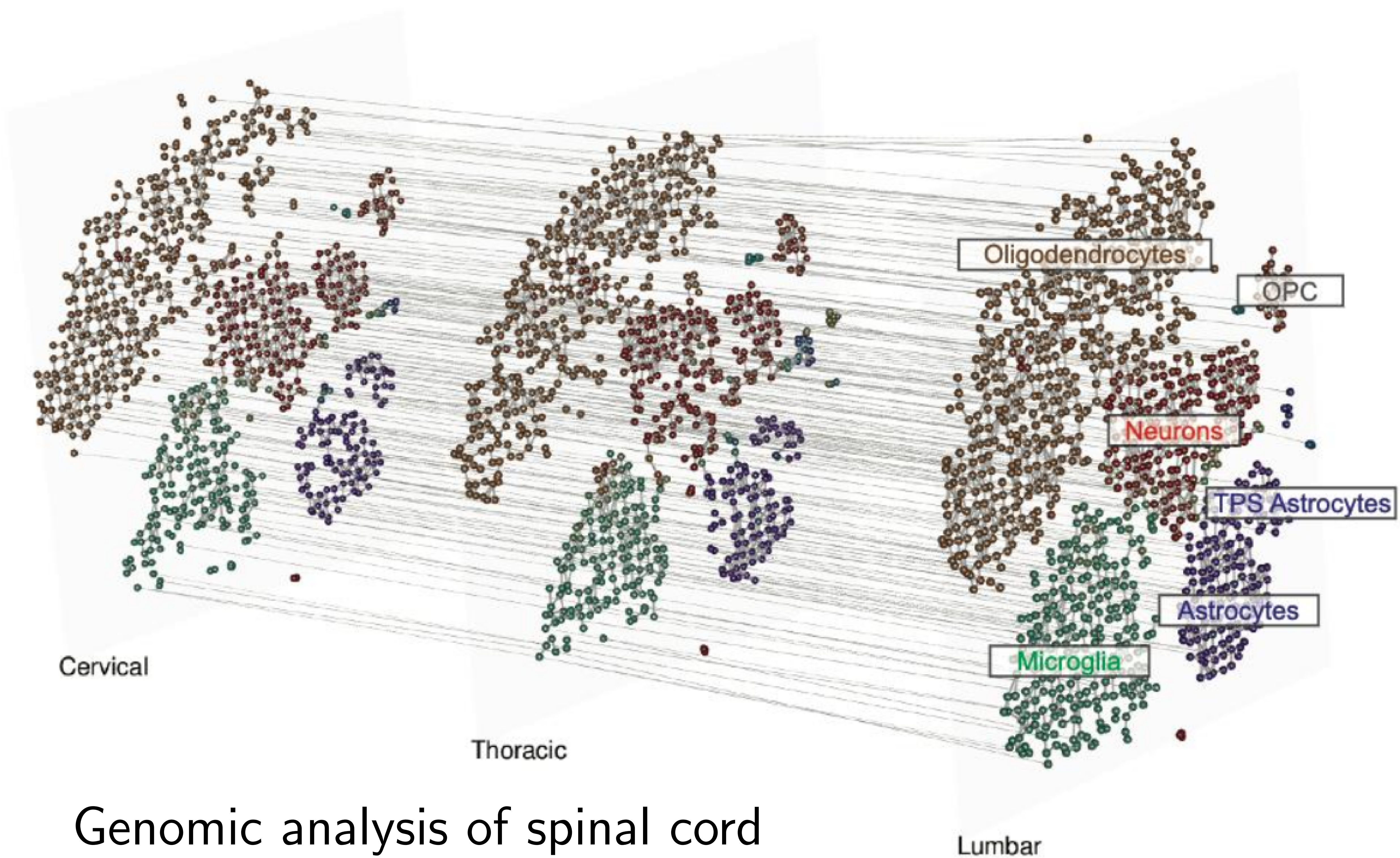
[Rucco et al. 2014]

Mapper in applications

Formal identification of cell cycle



Mapper in applications



Genomic analysis of spinal cord

Topological exploratory data analysis

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

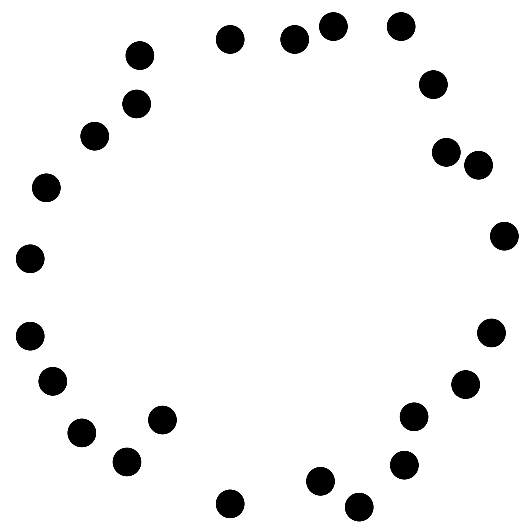
- Group data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of 'clusters'.

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

- Group data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of 'clusters'.

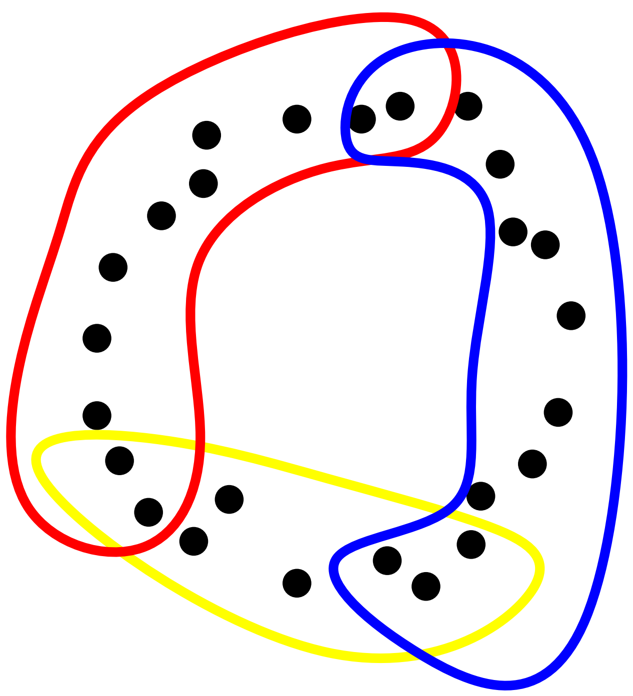


Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

- Group data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of 'clusters'.

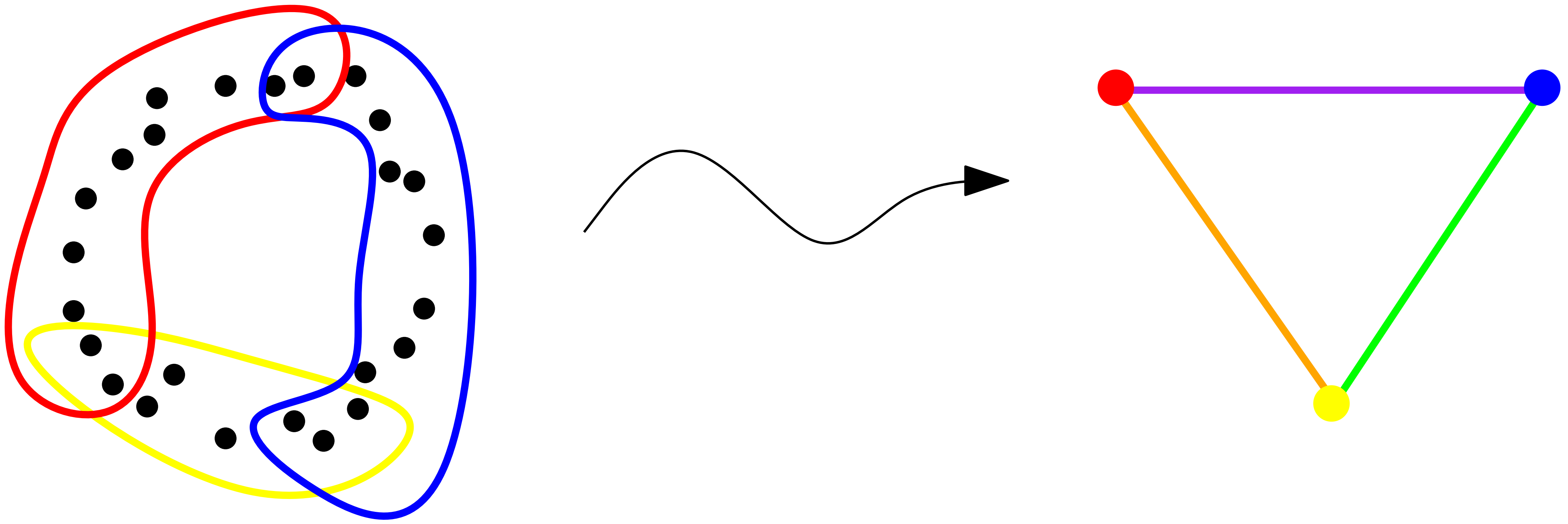


Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

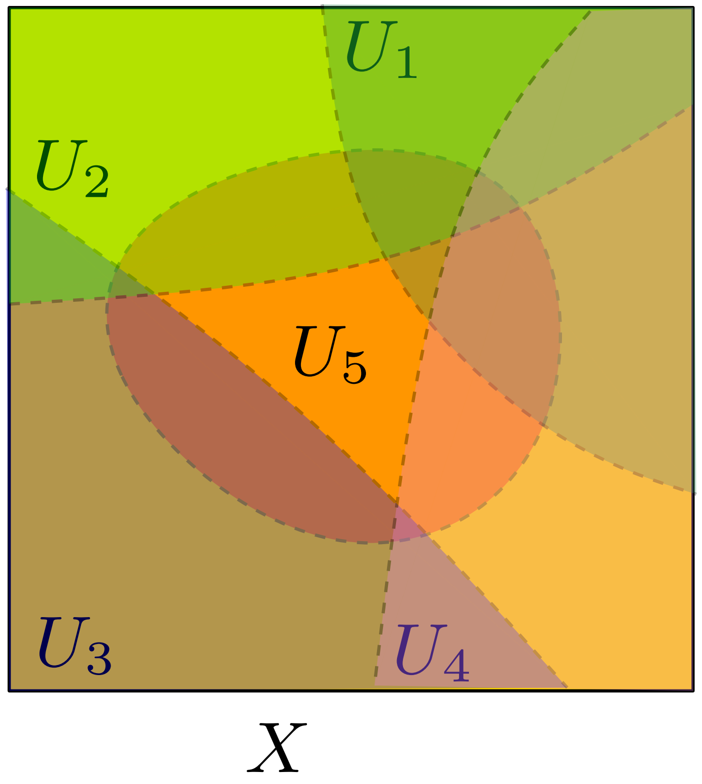
- Group data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of 'clusters'.



Topological exploratory data analysis

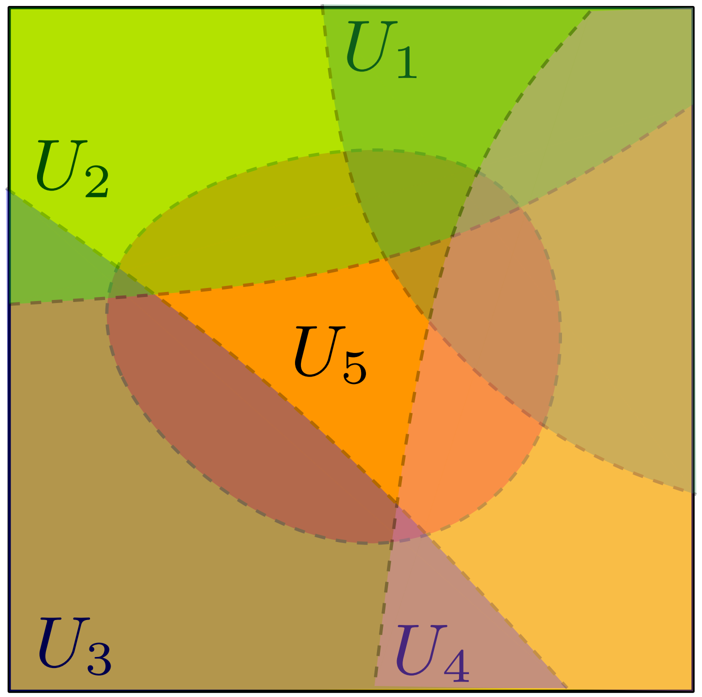
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

Topological exploratory data analysis



Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

Topological exploratory data analysis



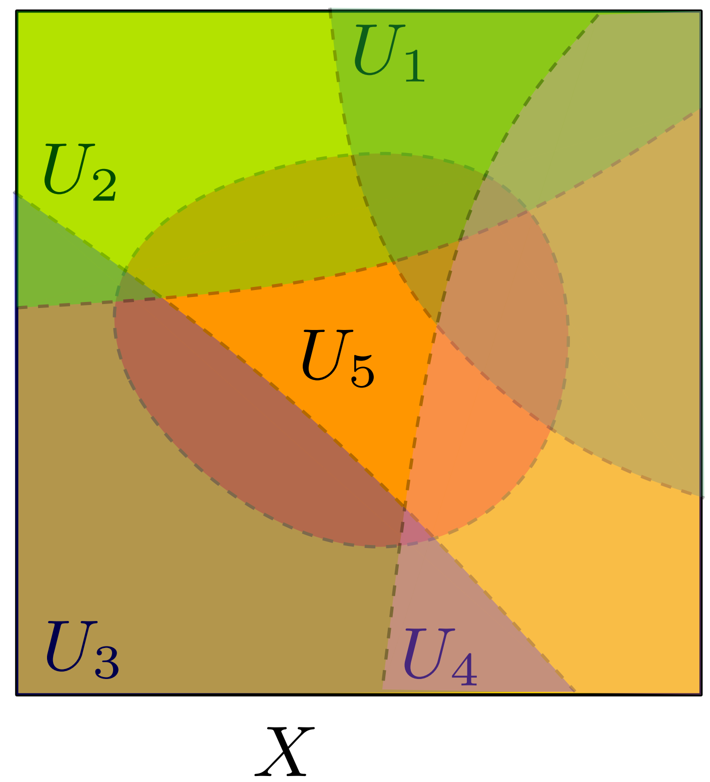
X

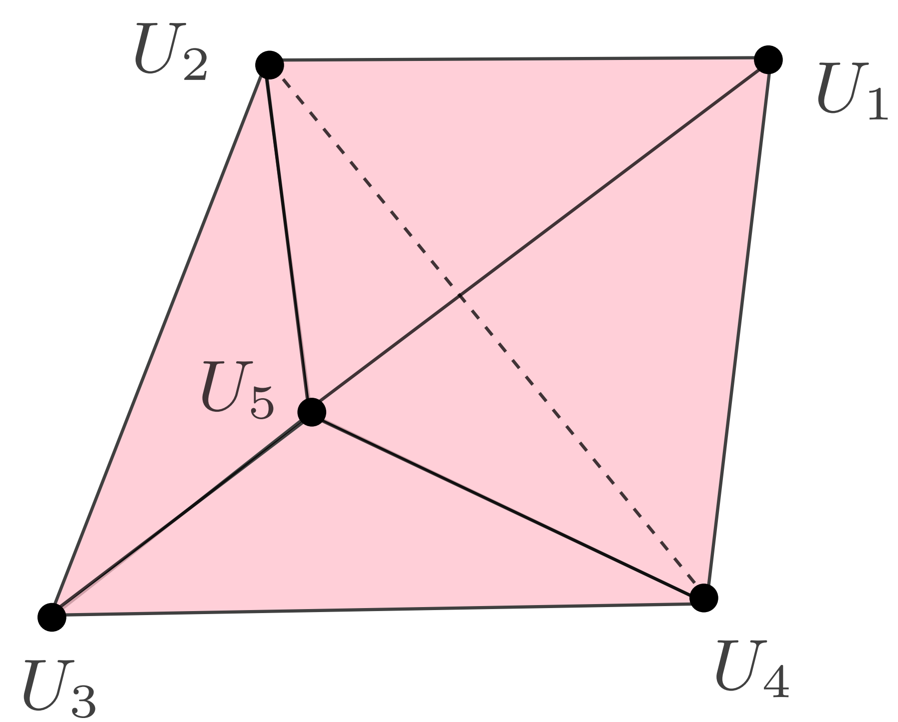
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

Def: Given a cover of a topological space X, $\mathcal{U} = (U_i)_{i \in I}$, its nerve is the abstract simplicial complex $C(\mathcal{U})$ whose vertex set is $\mathcal{U}$ and s.t.

$$\sigma = [U_{i_0}, U_{i_1}, \ldots, U_{i_k}] \in C(\mathcal{U}) \quad \text{if and only if} \quad \cap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

Topological exploratory data analysis





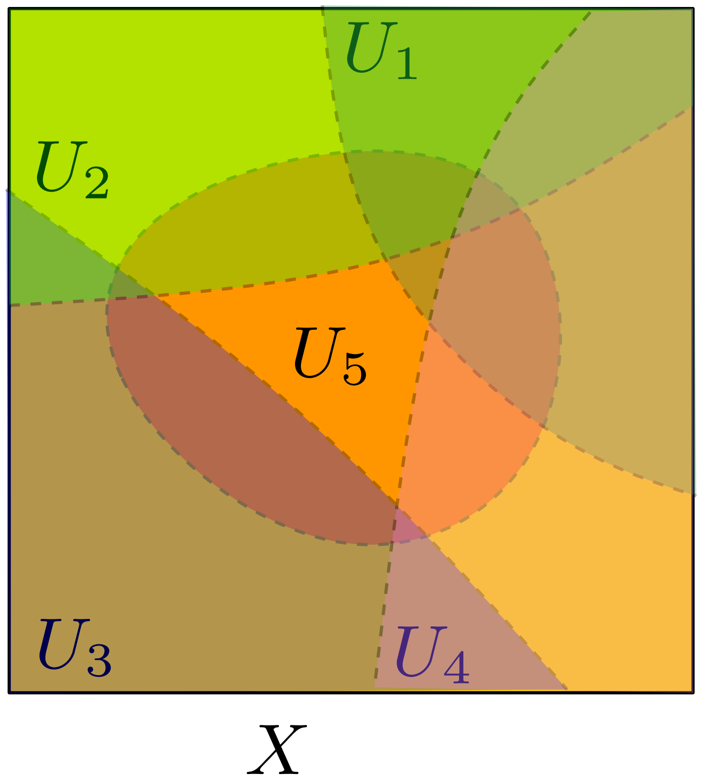
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

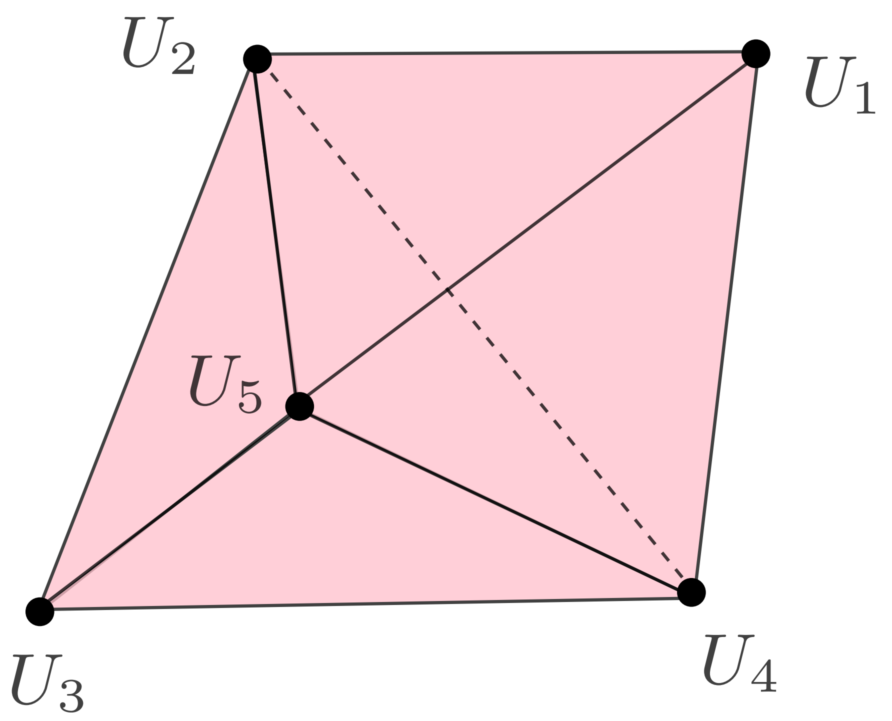
Def: Given a cover of a topological space X, $\mathcal{U} = (U_i)_{i \in I}$, its nerve is the abstract simplicial complex $C(\mathcal{U})$ whose vertex set is $\mathcal{U}$ and s.t.

$$\sigma = [U_{i_0}, U_{i_1}, \ldots, U_{i_k}] \in C(\mathcal{U}) \text{ if and only if } \cap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

Topological exploratory data analysis

[*On the imbedding of systems of compacta in simplicial complexes*, Borsuk, Fund. Math., 1948]





The Nerve Theorem: Let $\mathcal{U} = (U_i)_{i \in I}$ be a finite open cover of a subset X of $\mathbb{R}^d$ such that any intersection of the U_i's is either empty or contractible. Then X and $C(\mathcal{U})$ are homotopy equivalent. In particular, their homology groups are isomorphic.

For non-experts, you can replace:
- 'contractible' by 'convex',
- 'are homotopy equivalent' by 'same topological invariants'.

Topological exploratory data analysis

Q: How to build meaningful covers?

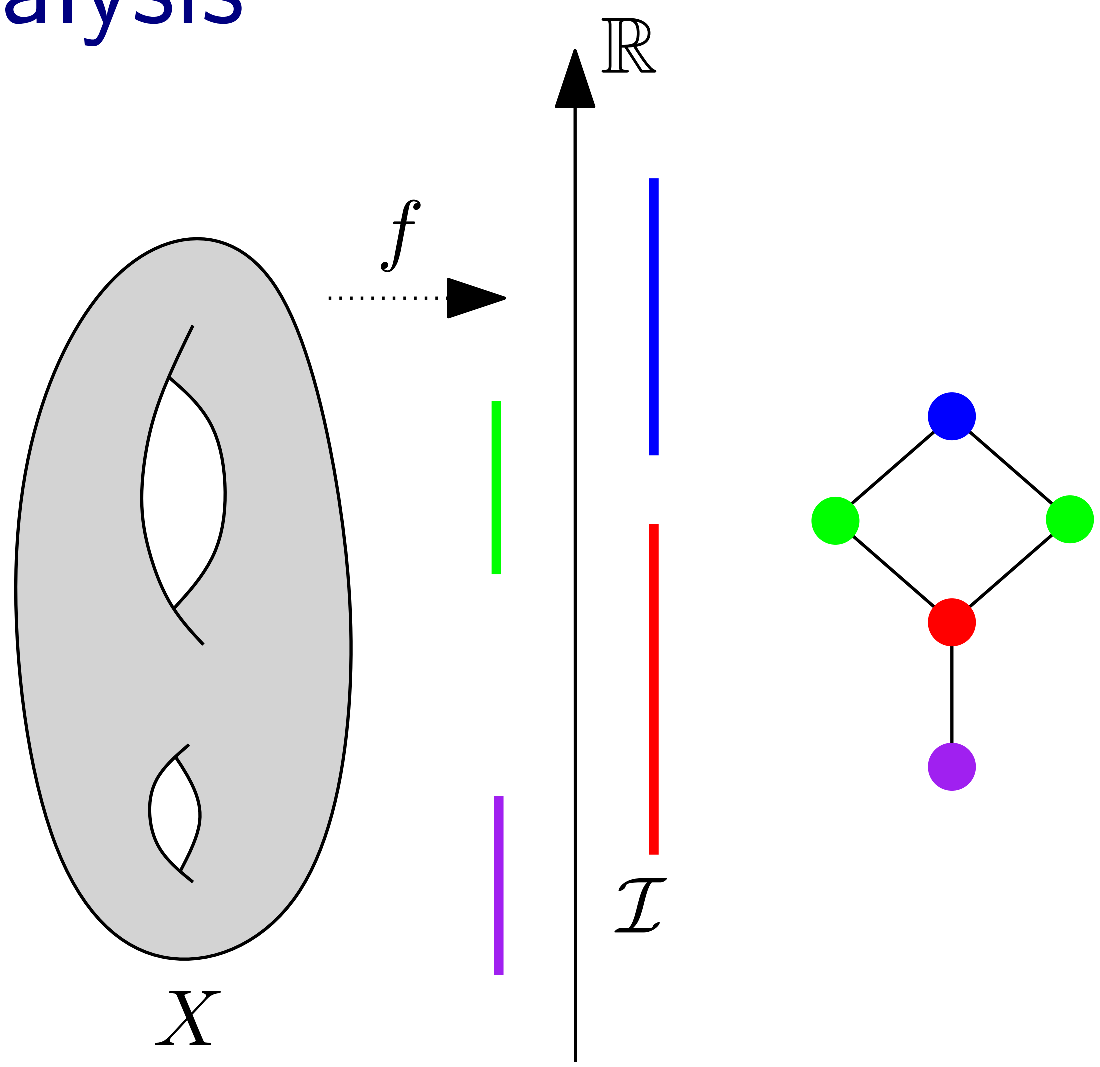
Two directions:

Topological exploratory data analysis

Q: How to build meaningful covers?

Two directions:

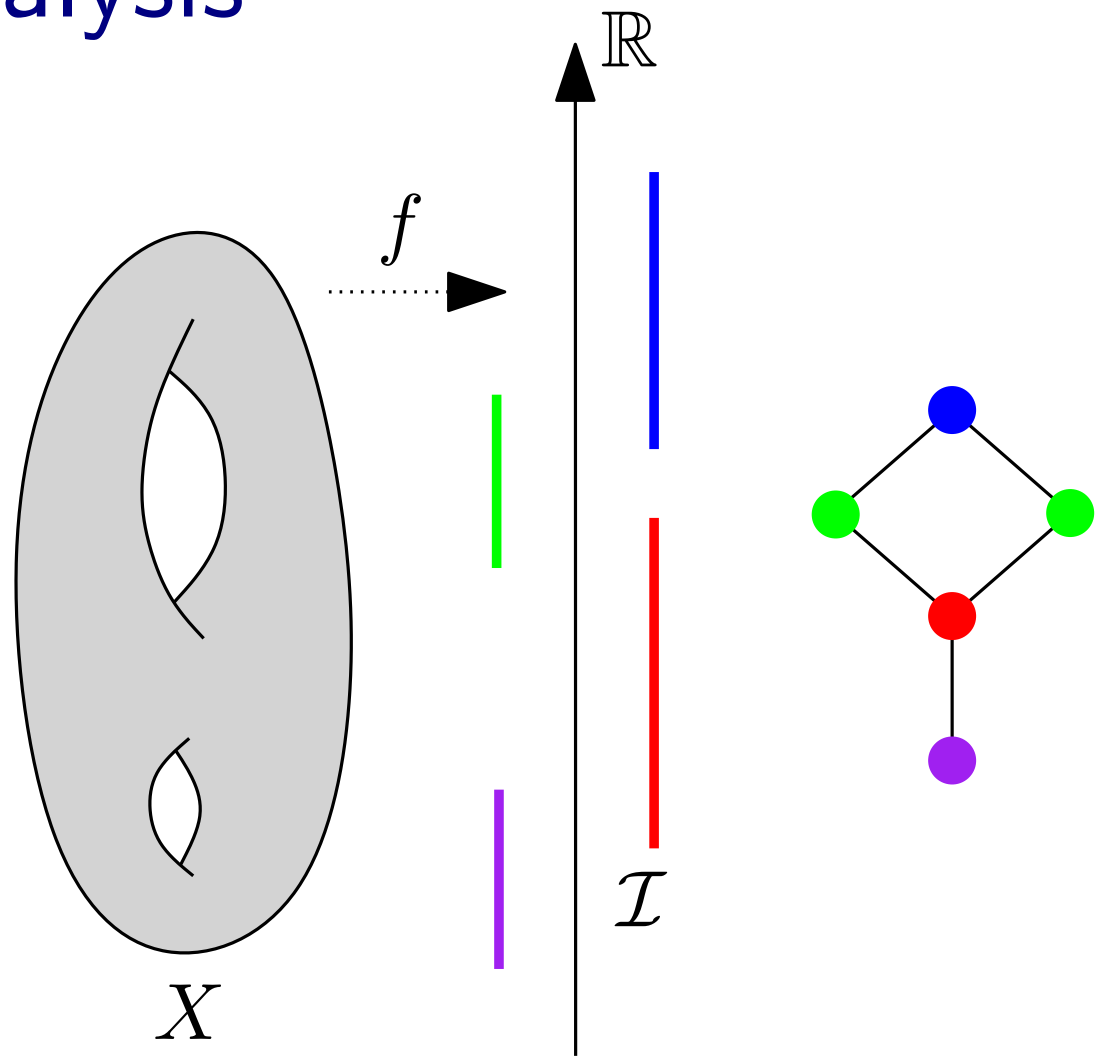
1. Using a function (lens) defined on the data:
$\rightarrow$ the Mapper algorithm
$\rightarrow$ exploratory data analysis



Topological exploratory data analysis

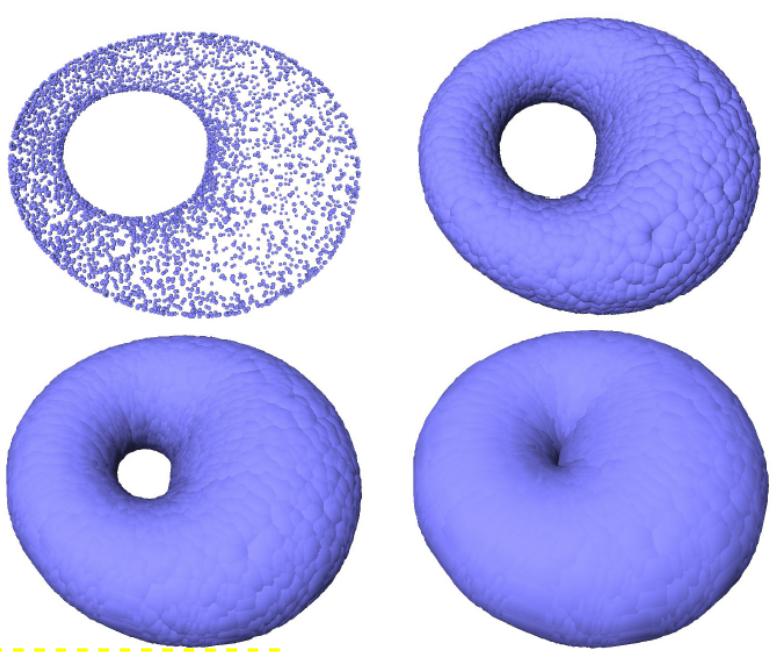
Q: How to build meaningful covers?

Two directions:



1. Using a function (lens) defined on the data:
$\rightarrow$ the Mapper algorithm
$\rightarrow$ exploratory data analysis

2. Covering data by balls:
$\rightarrow$ distance functions frameworks, persistence-based signatures,...
$\rightarrow$ geometric inference, provide a framework to establish various theoretical results in TDA.

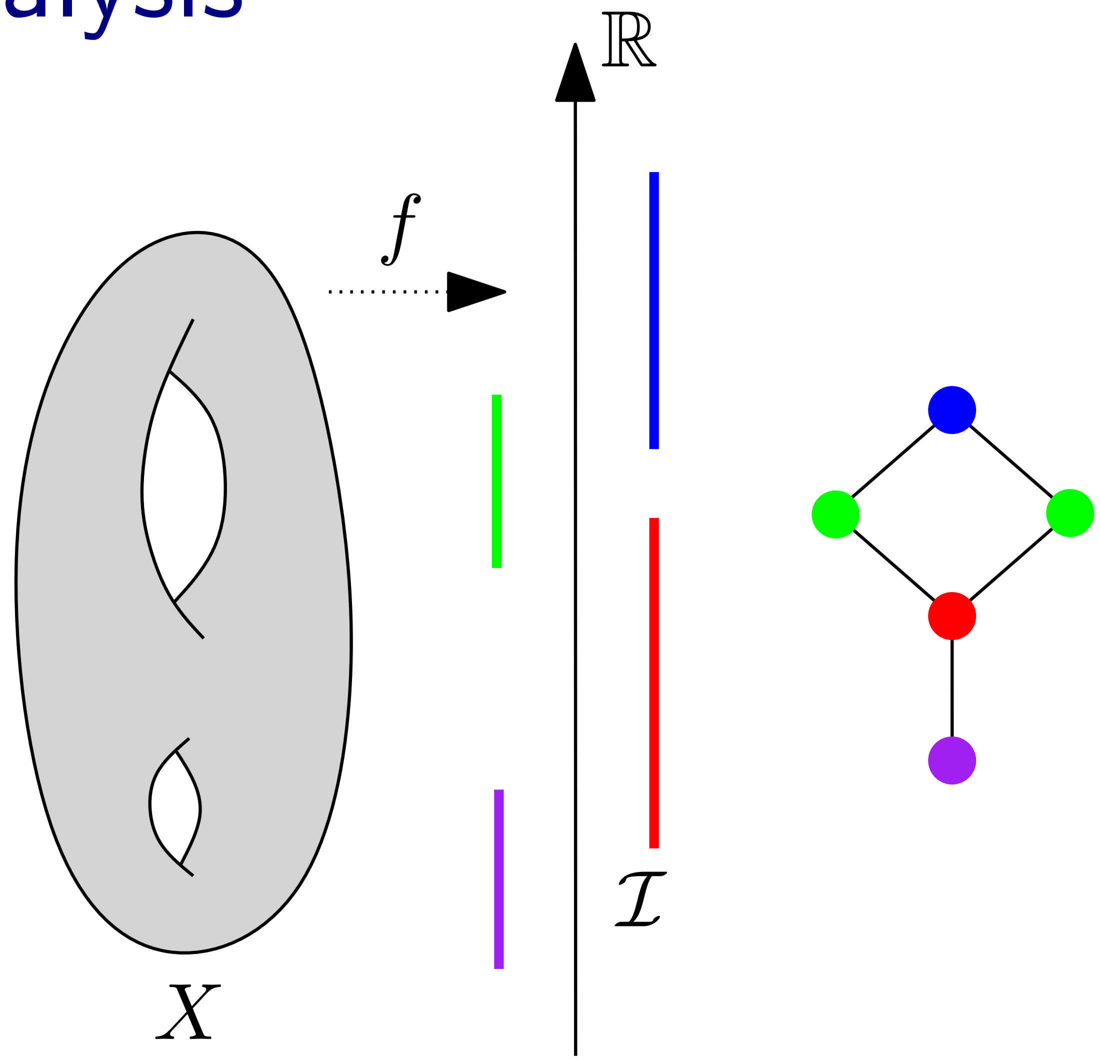


Topological exploratory data analysis

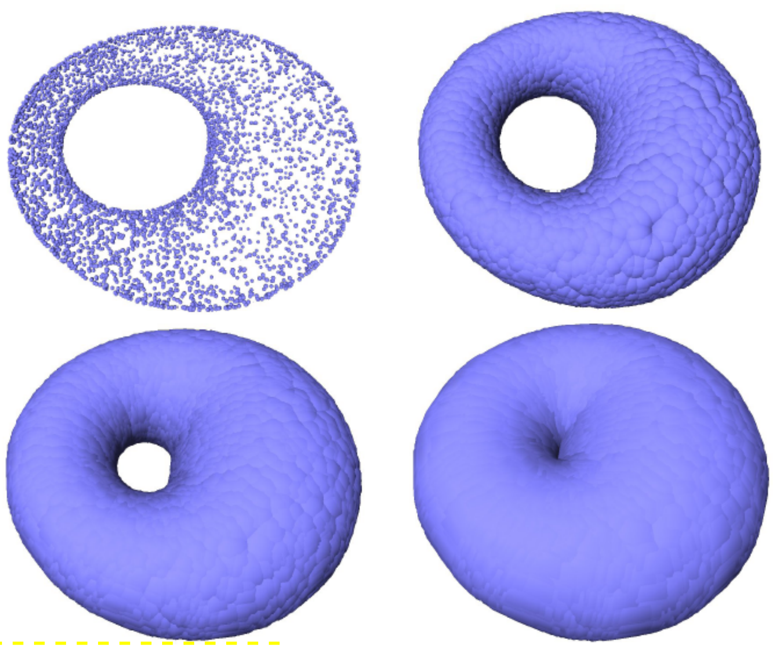
Q: How to build meaningful covers?

Two directions:

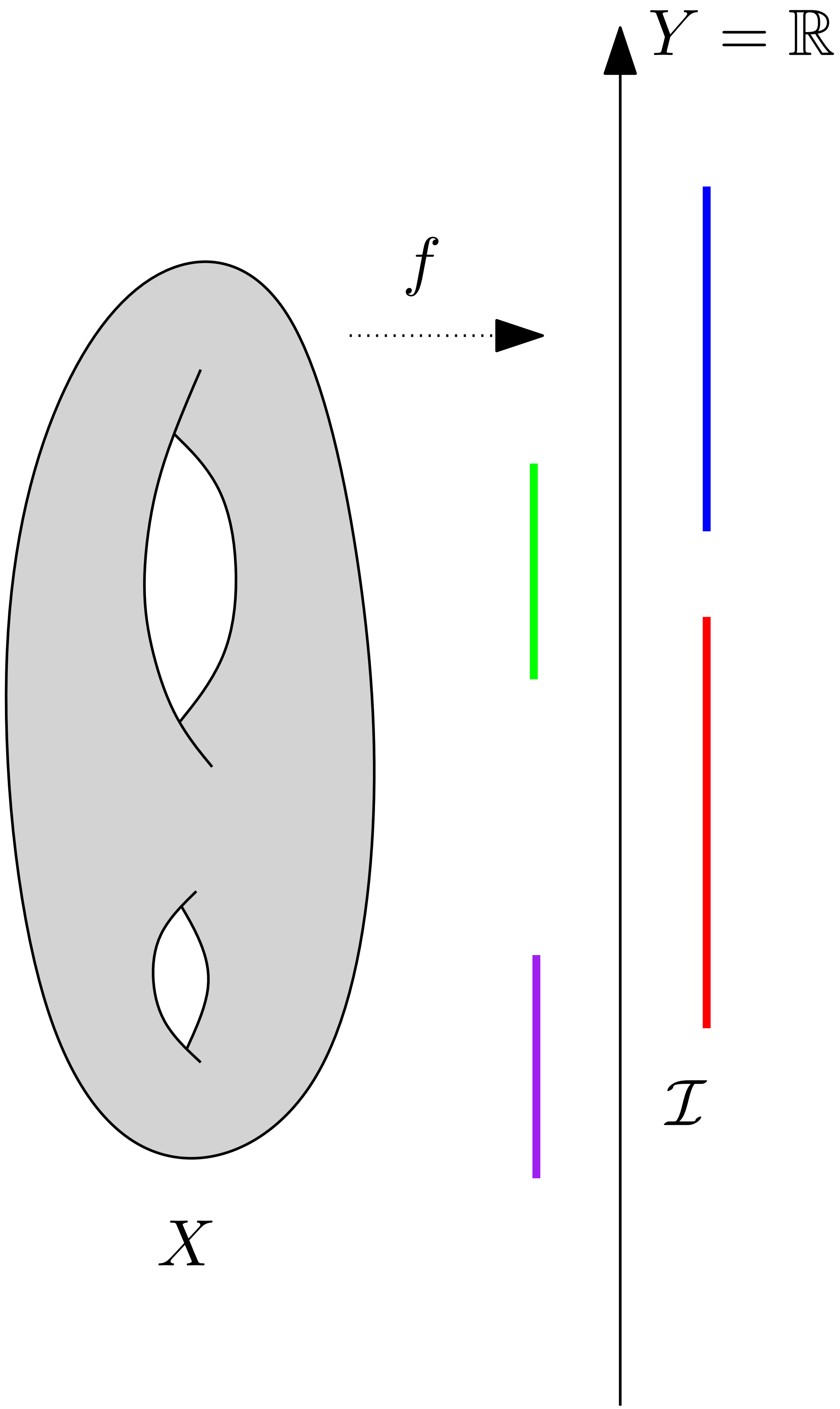
1. Using a function (lens) defined on the data:
$\rightarrow$ the Mapper algorithm
$\rightarrow$ exploratory data analysis



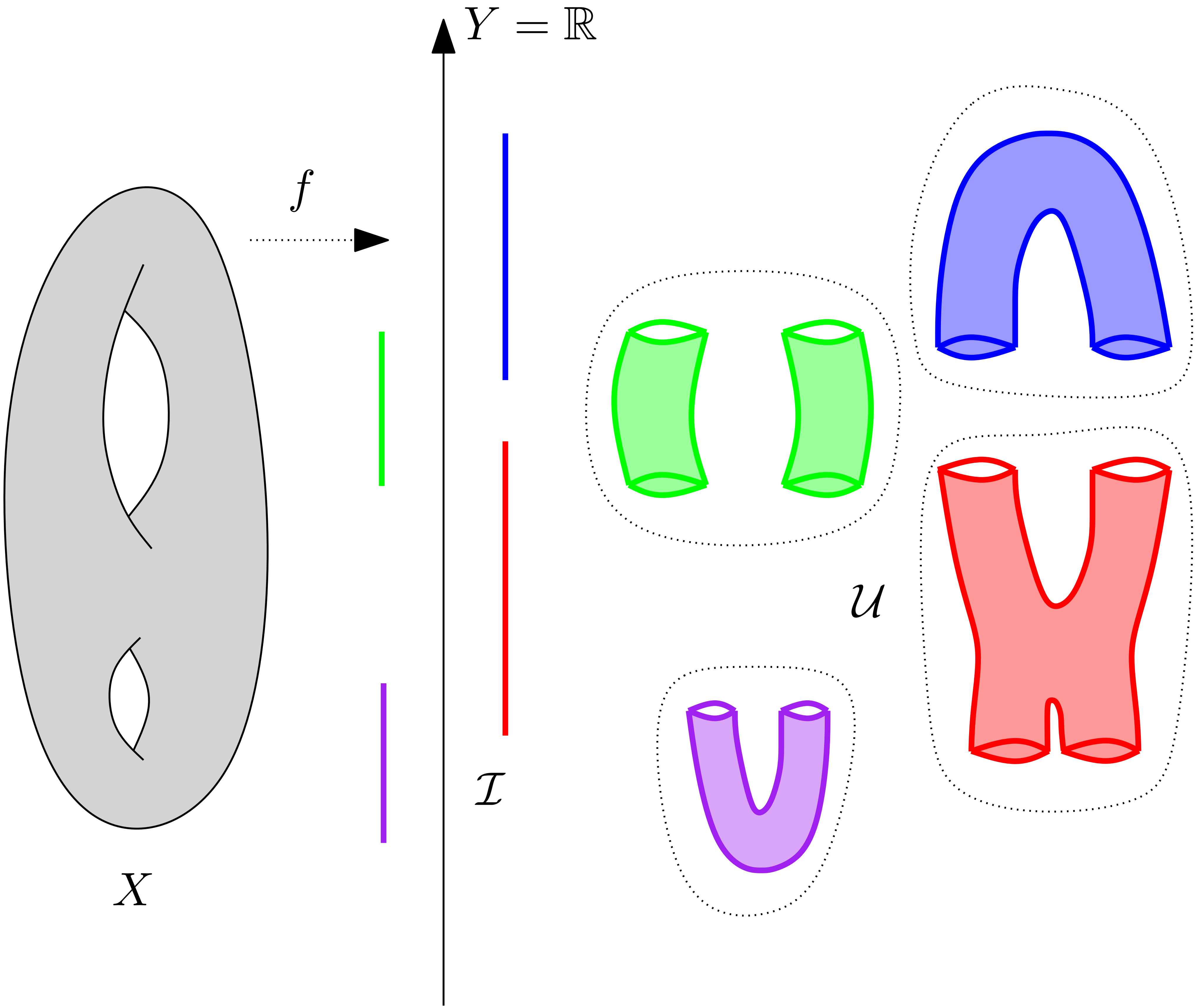
2. Covering data by balls:
$\rightarrow$ distance functions frameworks, persistence-based signatures,...
$\rightarrow$ geometric inference, provide a framework to establish various theoretical results in TDA.



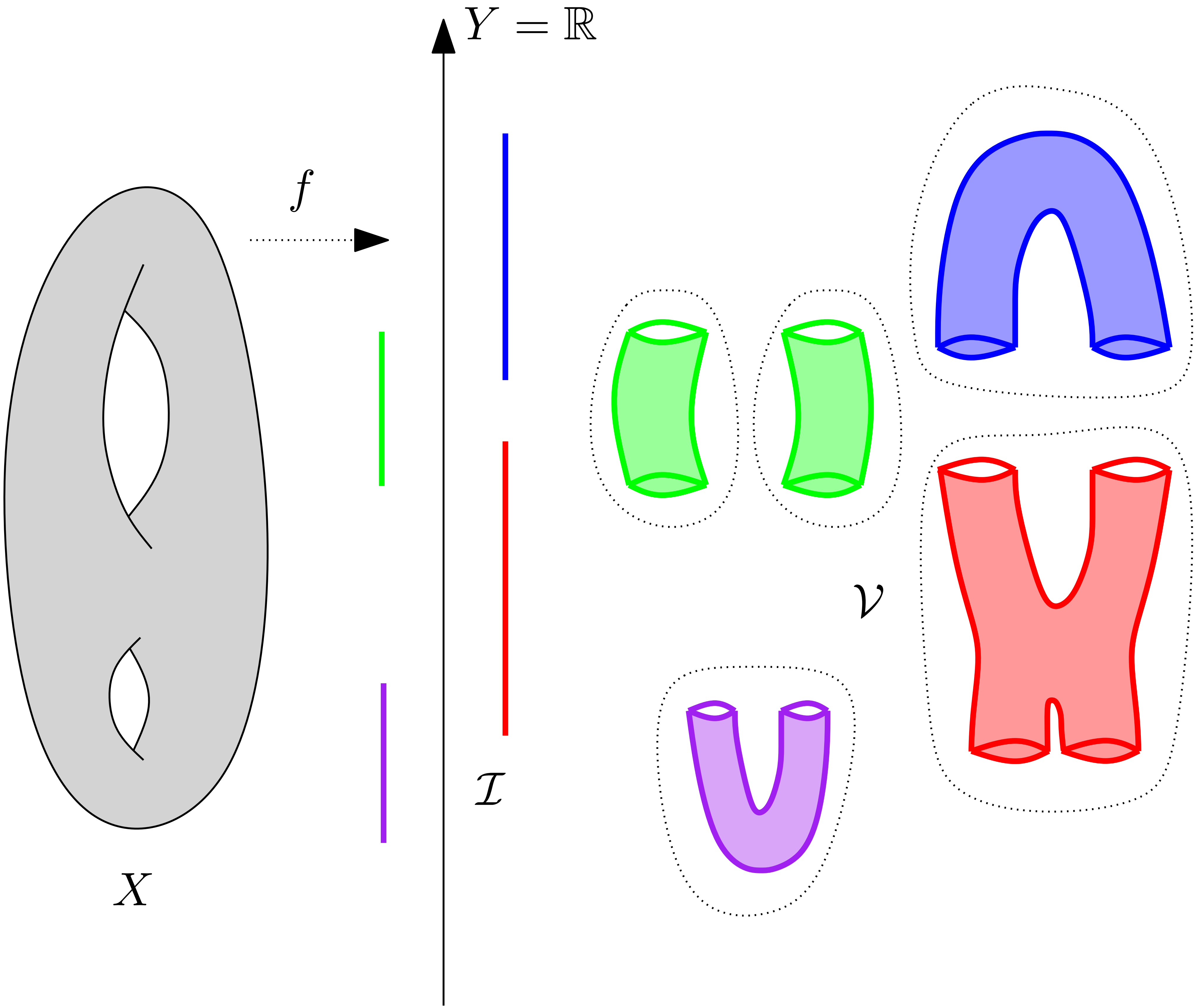
Mapper in the continuous setting



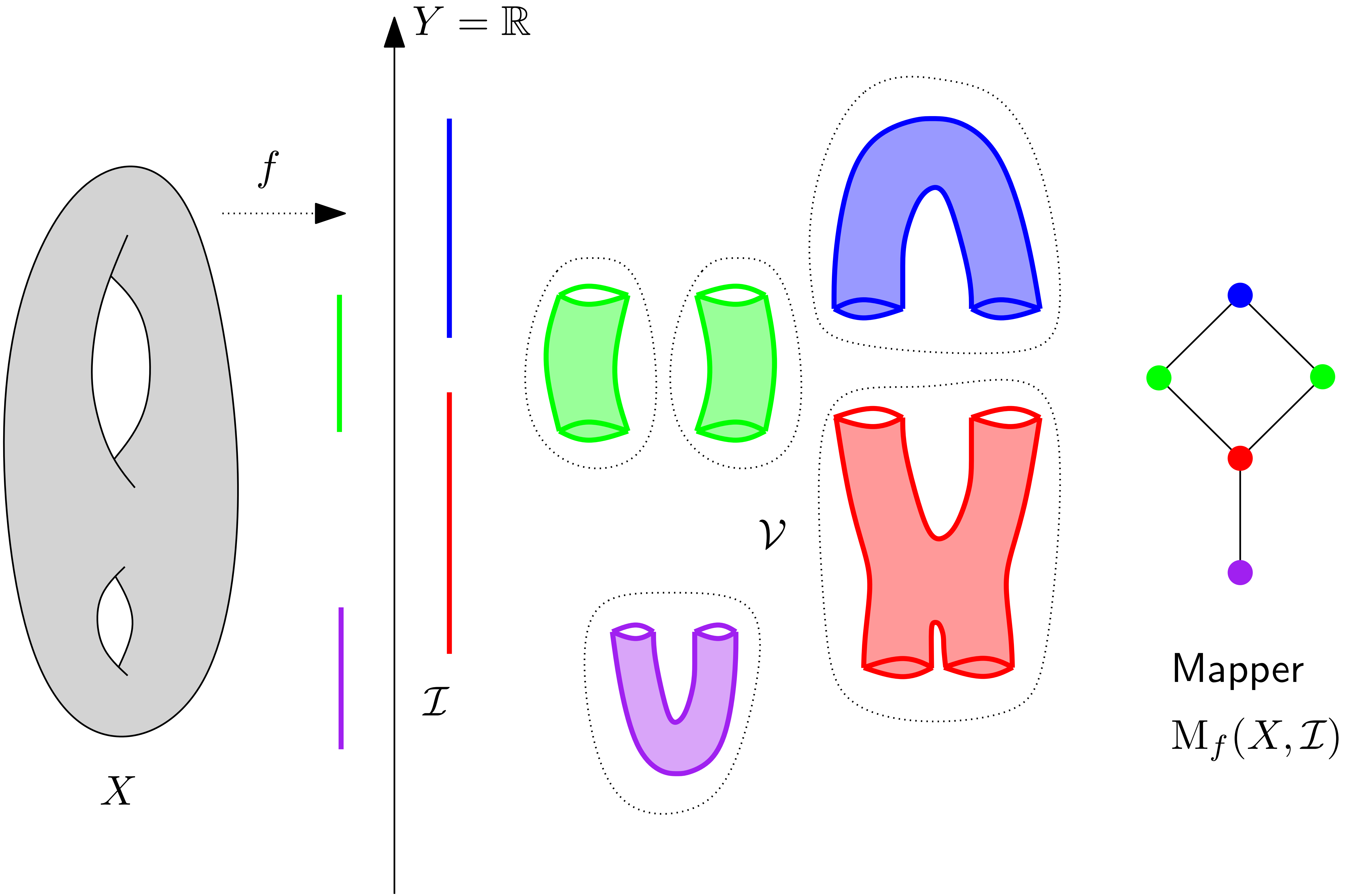
Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting

Input:

- topological space X
- continuous function $f : X \to Y$ $\quad$ ($Y = \mathbb{R}$ 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 *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 $f : P \to \mathbb{R}$
- 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 P: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$
- Refine $\mathcal{U}$ by separating each of its elements into its various clusters, as identified by a clustering algorithm $\to$ 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 are assessed by the presence of common data points

Mapper in practice

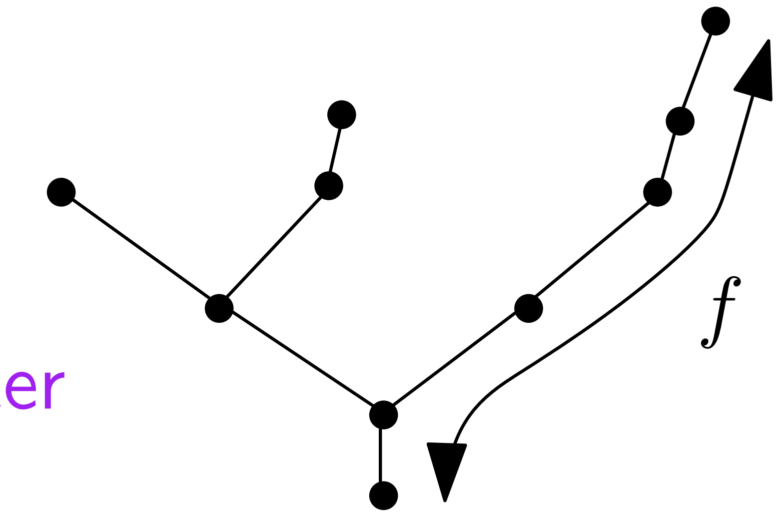
Parameters:

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

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

- clustering algorithm $\mathcal{C}$

Mapper in practice

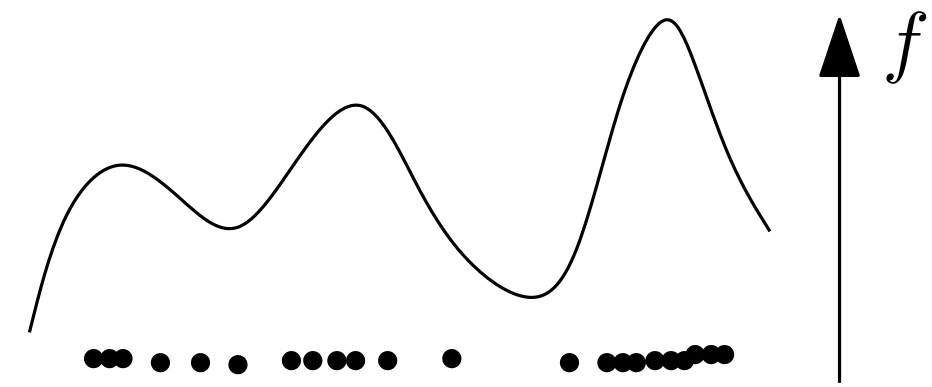


Parameters:

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

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

- clustering algorithm $\mathcal{C}$



Classical choices:

- density estimates
- centrality $f(x) = \sum_{y \in X} d(x, y)$
- eccentricity $f(x) = \max_{y \in X} d(x, y)$
- PCA coordinates

- Eigenfunctions of graph laplacians.
- Functions detecting outliers.
- Distance to a root point.
- Prior knowledge

Mapper in practice

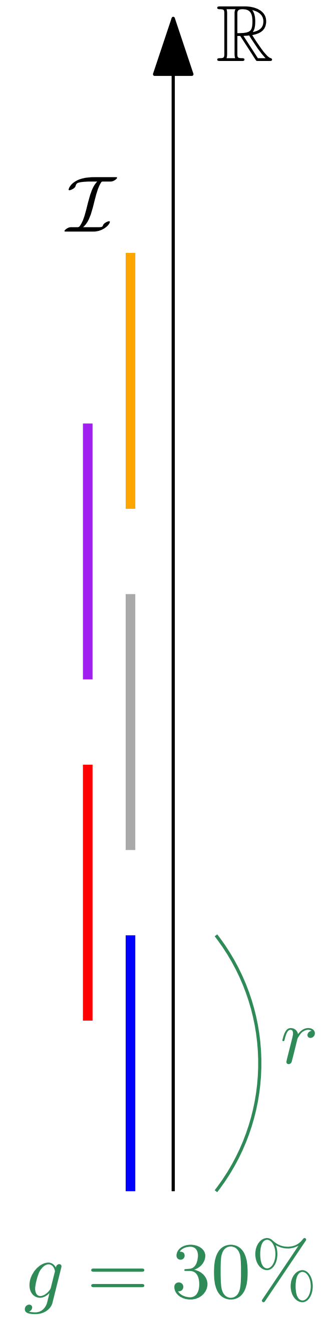
Parameters:

- function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- clustering algorithm $\mathcal{C}$

range scale

Uniform cover:

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

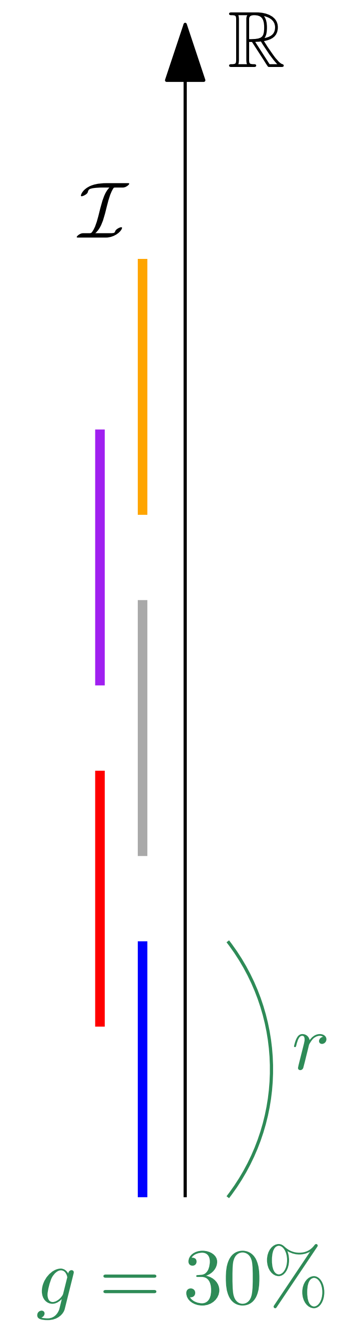


Mapper in practice

Parameters:

- function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- clustering algorithm $\mathcal{C}$

range scale



Uniform cover:

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

Intuition:

- small $r \to$ finer resolution, more nodes.
- large $r \to$ rougher resolution, less nodes.

- small $g \to$ less connectivity, nerve dimension small.
- large $g \to$ more connectivity, nerve dimension large.

Mapper in practice

Parameters:

- function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- clustering algorithm $\mathcal{C}$

Classical choices:

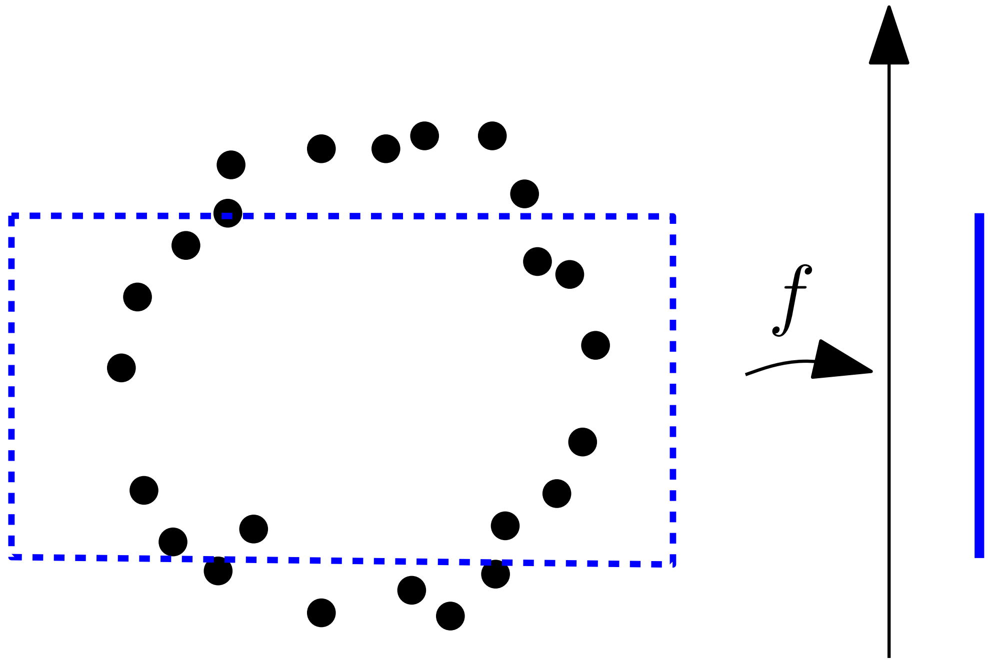
- any clustering algorithm works
- different clustering algorithms/parameters for each preimage
- for theoretical reasons, we prefer to work with hierarchical clustering with (predefined) neighborhood size δ

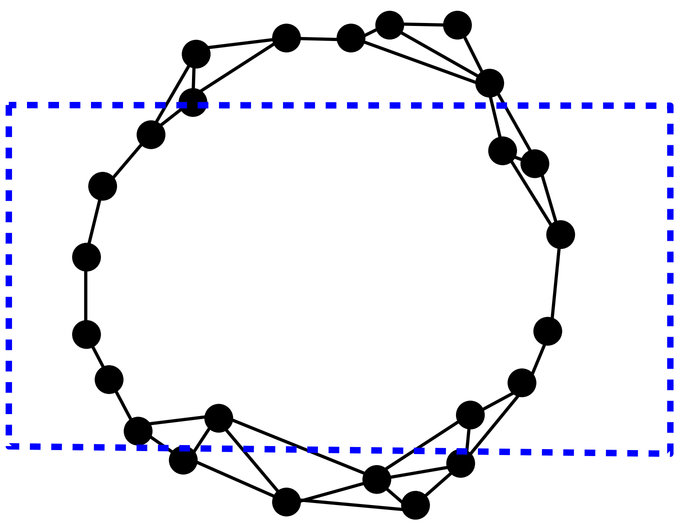
geometric scale

Mapper in practice

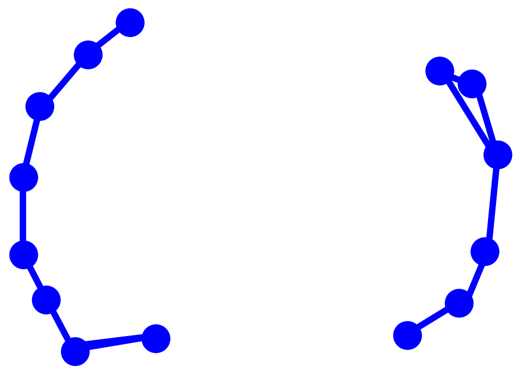
Parameters:

- function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- clustering algorithm $\mathcal{C}$



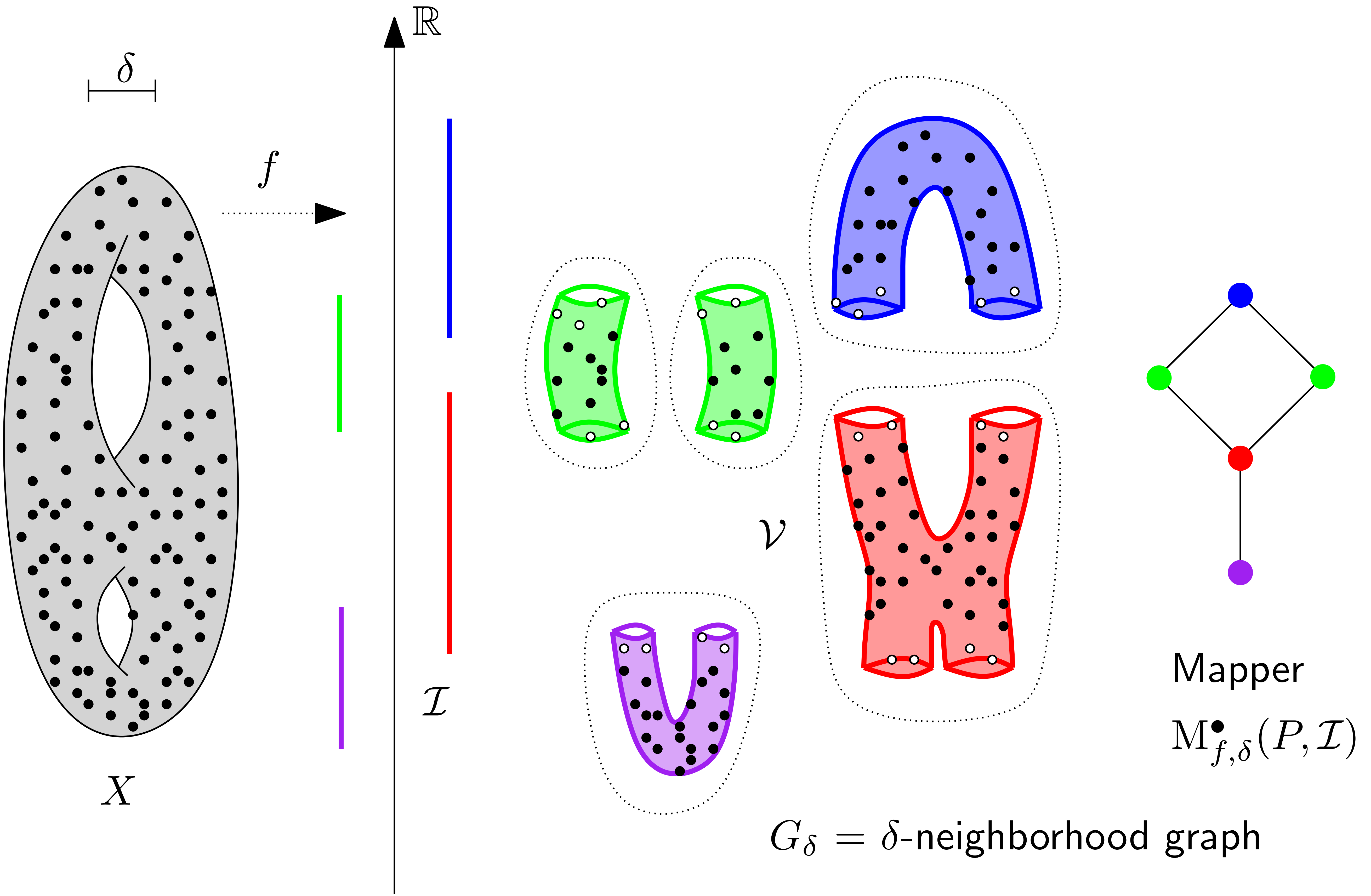


Build a neighboring graph (kNN,...)



Take the connected components of the subgraph spanned by the vertices in the preimage $f^{-1}(U)$.

Mapper in practice



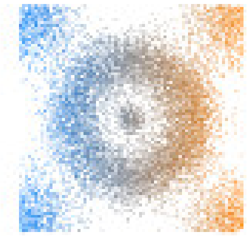
Choice of parameters

→ in practice: trial-and-error

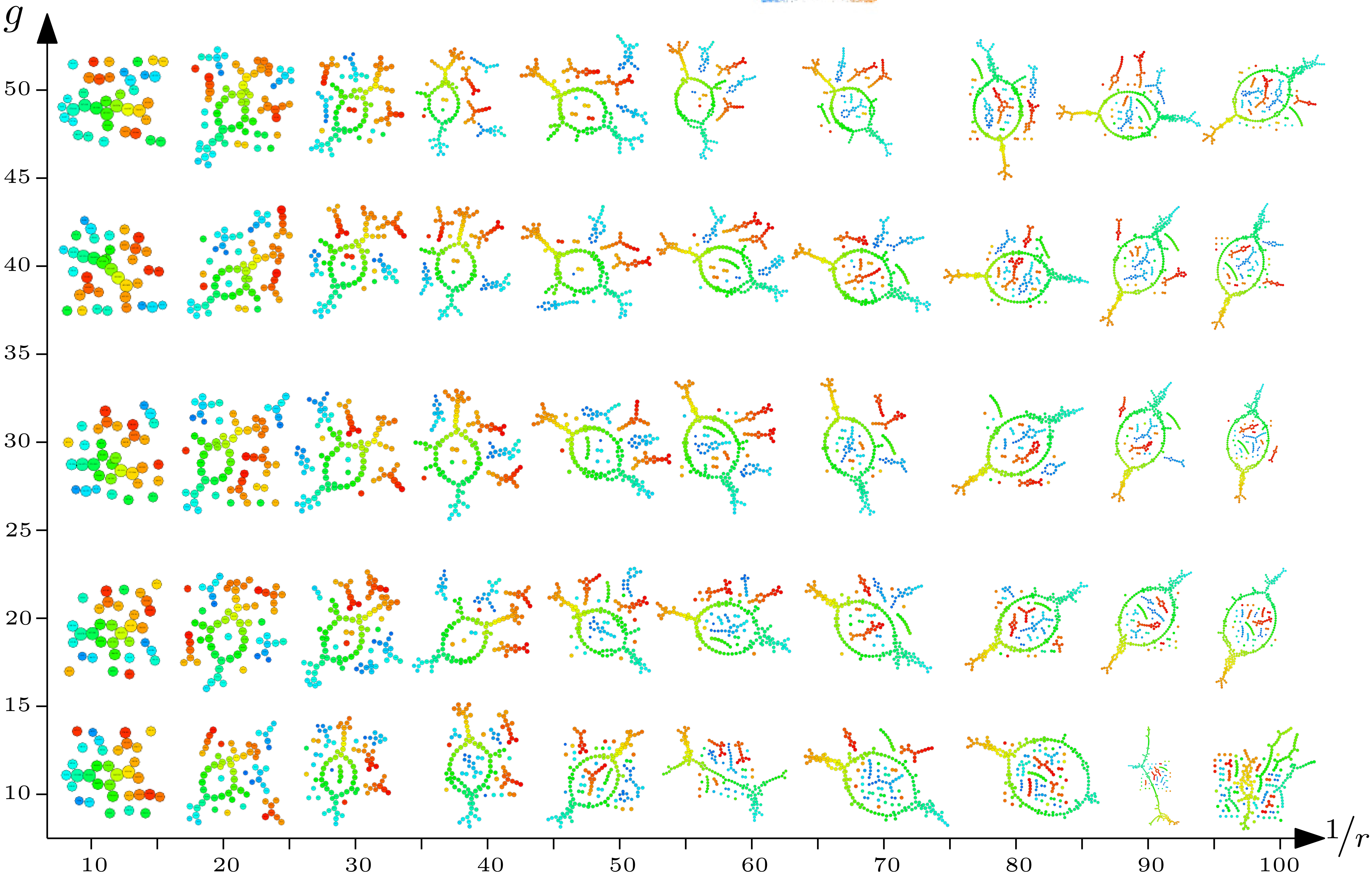
high-dimensional data sets[40,48]. This is performed automatically within the software, by deploying an ensemble machine learning algorithm that iterates through overlapping subject bins of different sizes that resample the metric space (with replacement), thereby using a combination of the metric location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

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

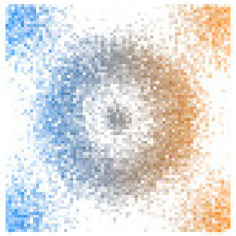
Choice of parameters



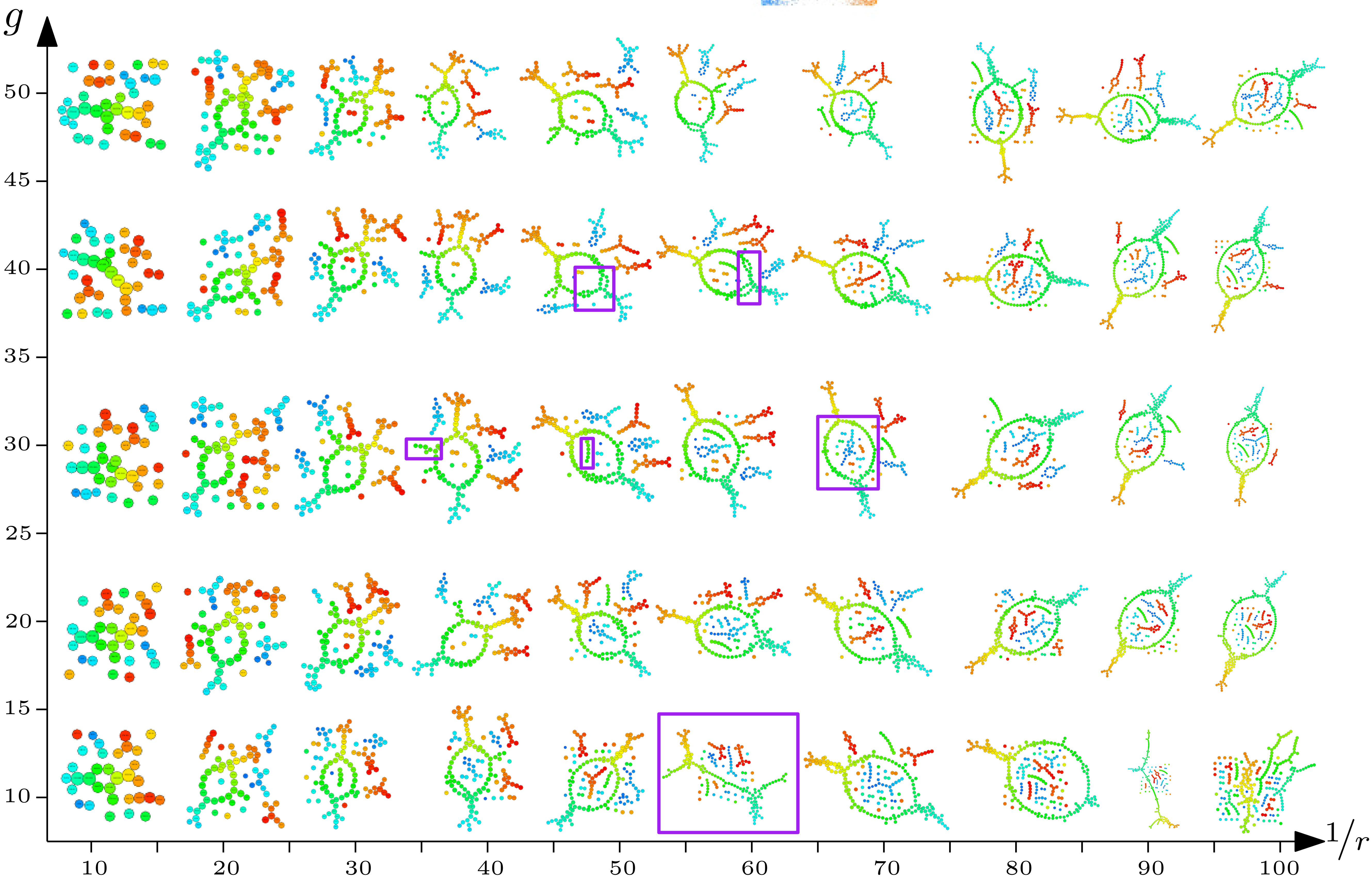
$f = f_x,\ \delta = 1\%$



Choice of parameters

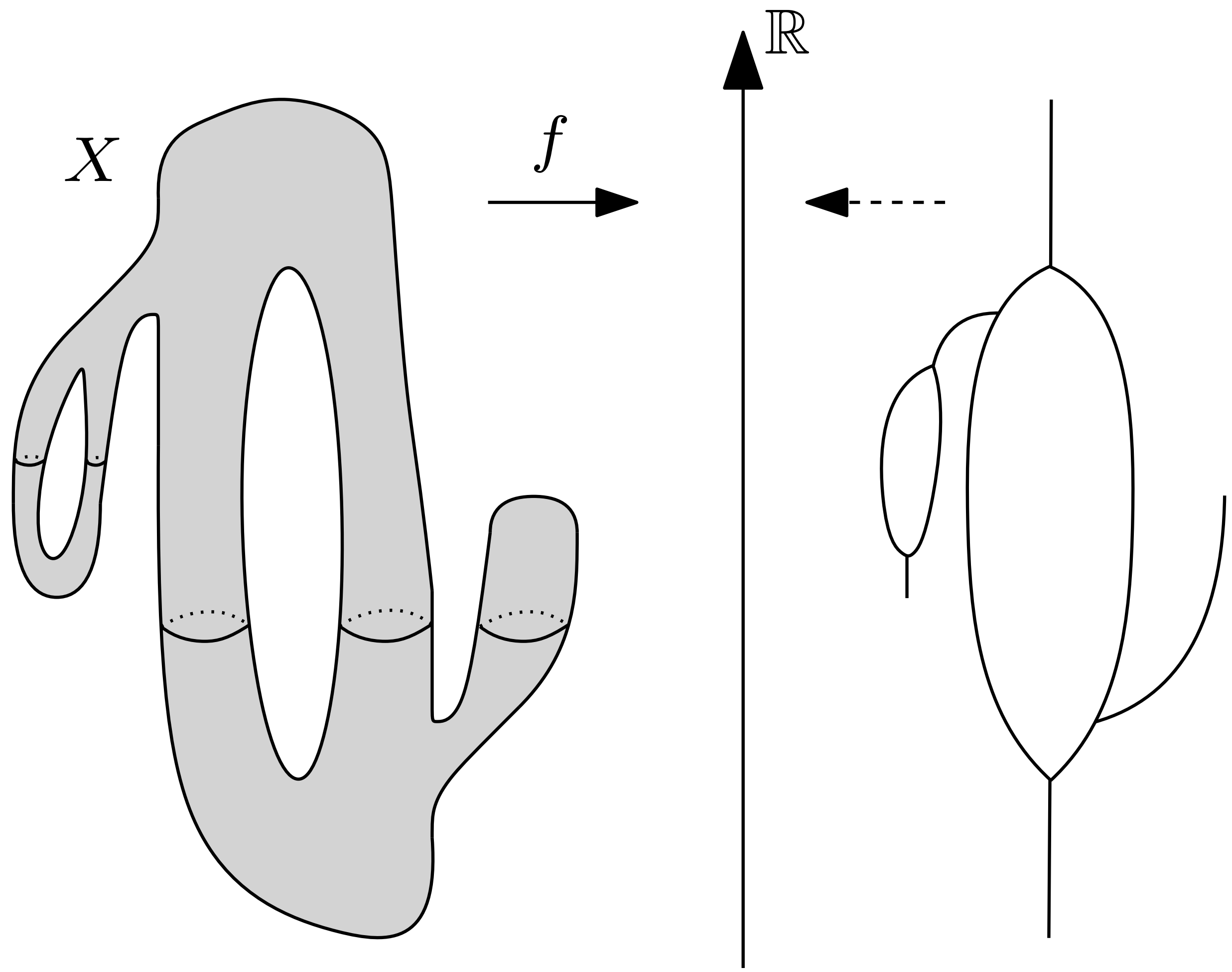


$f = f_x,\ \delta = 1\%$



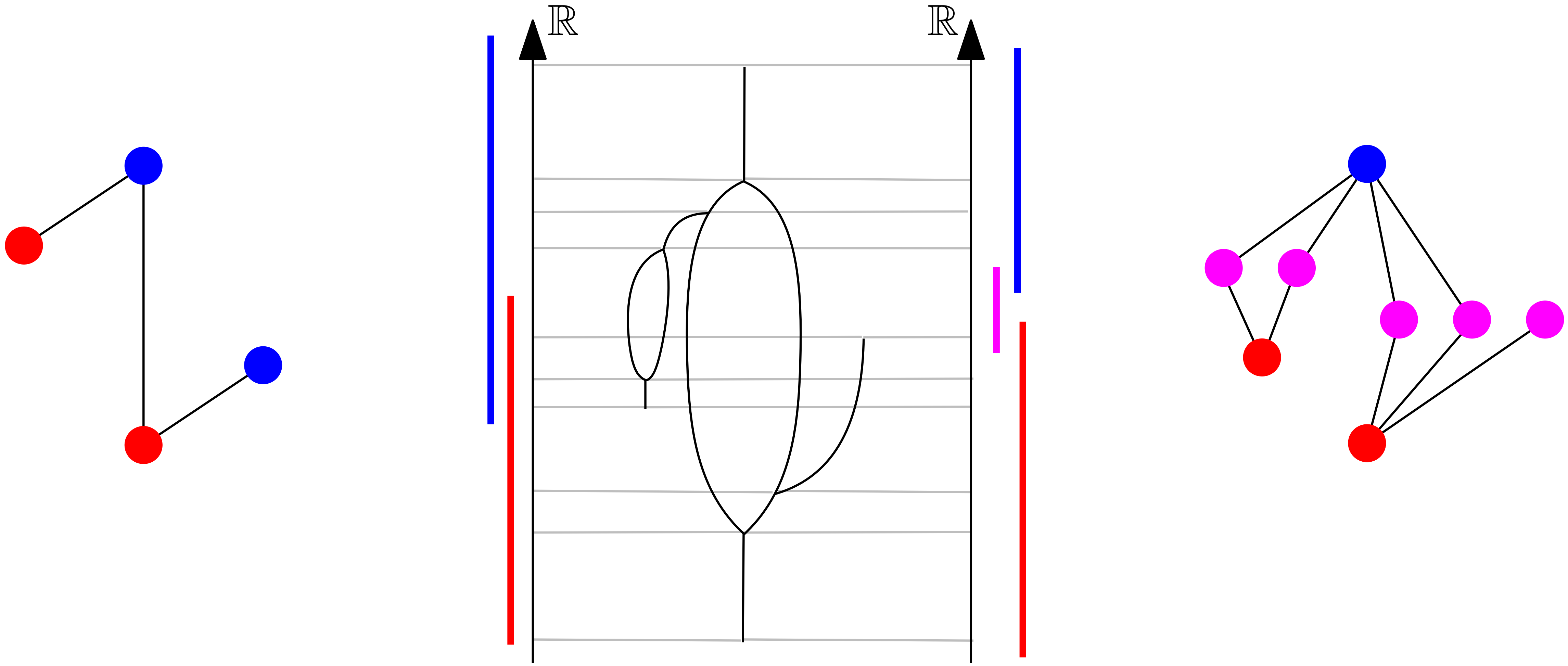
Reeb Graph

Reeb graph $\sim$ Mapper with extremely small resolution



Reeb Graph

Mapper $\sim$ *pixelized* Reeb graph

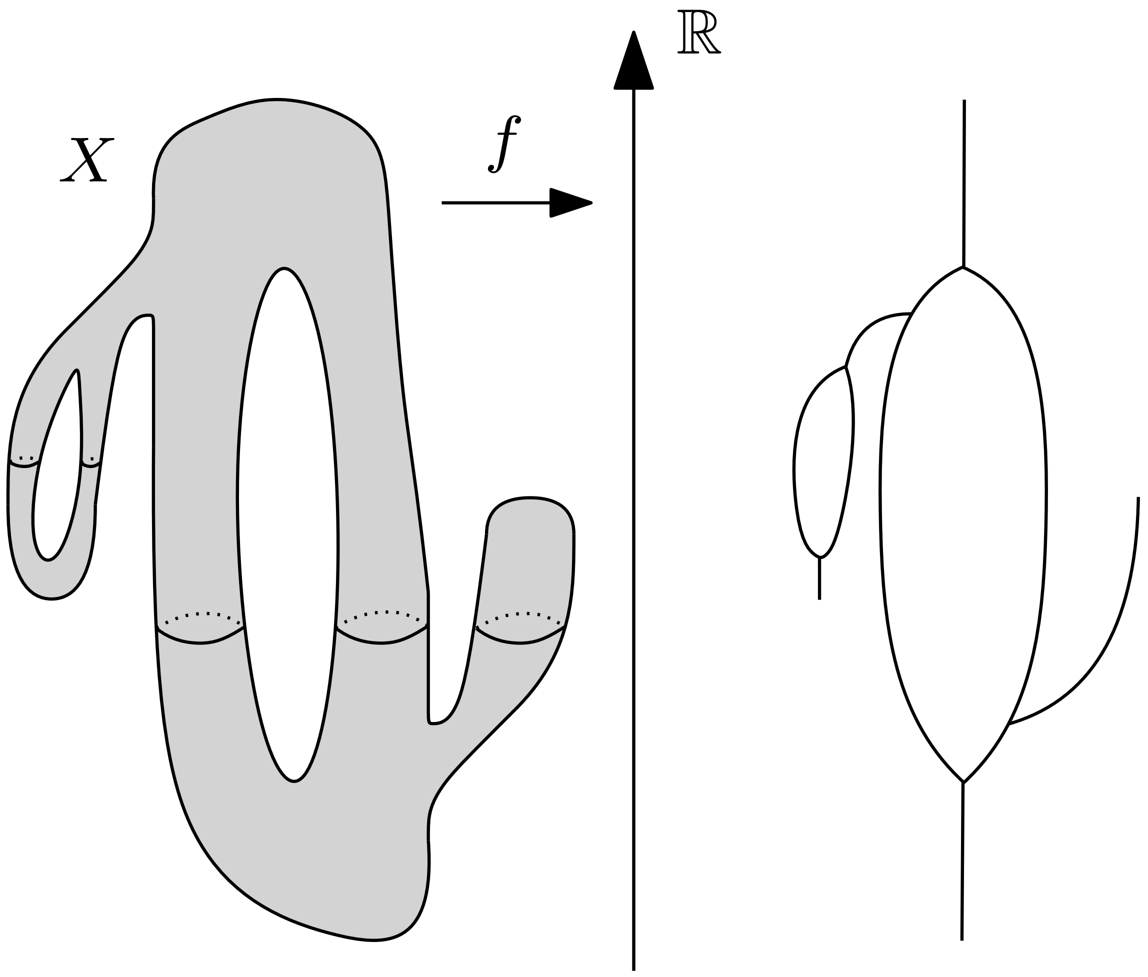


Reeb Graph

[*Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique*, Reeb, C. R. Acad. Sci. Paris, 1946]

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

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

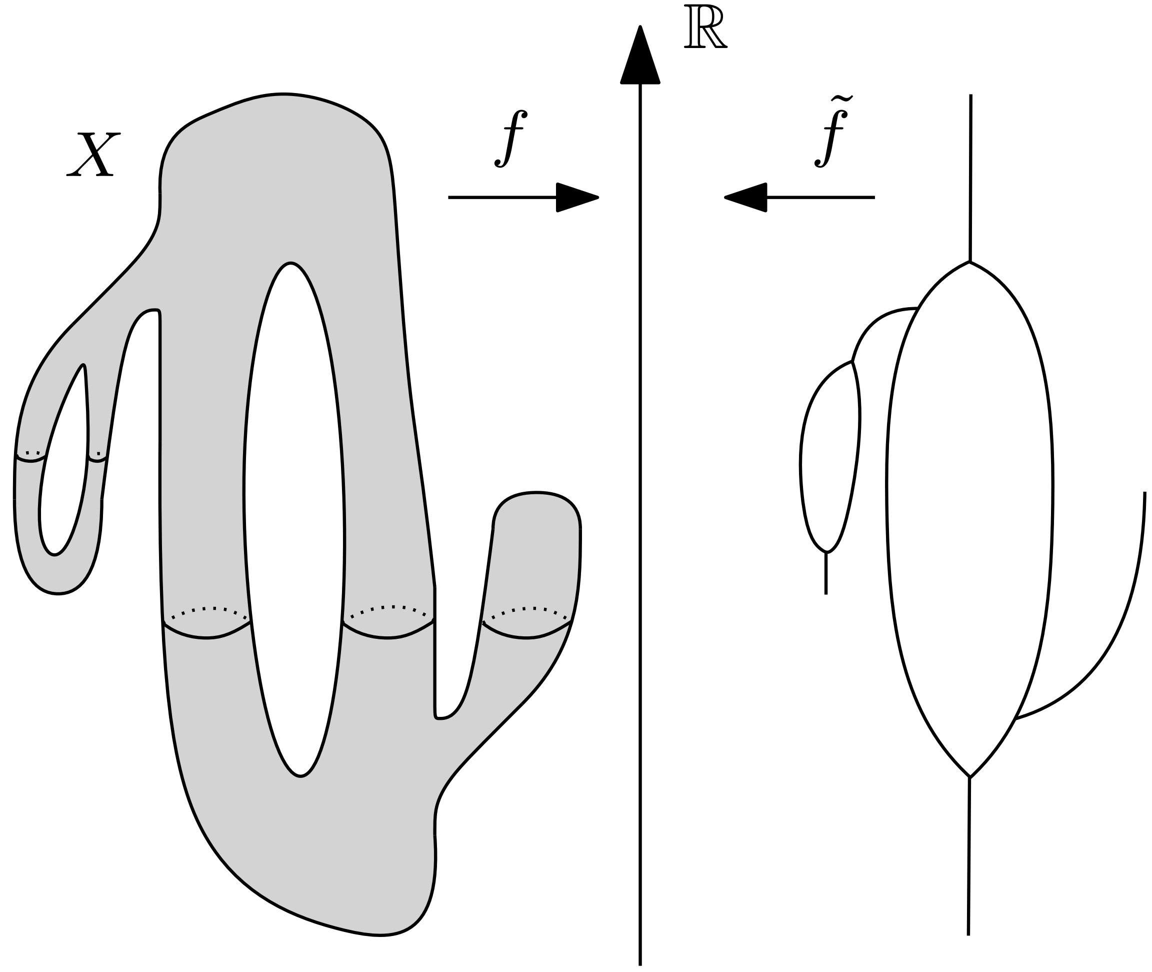


Reeb Graph

[*Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique*, Reeb, C. R. Acad. Sci. Paris, 1946]

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

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



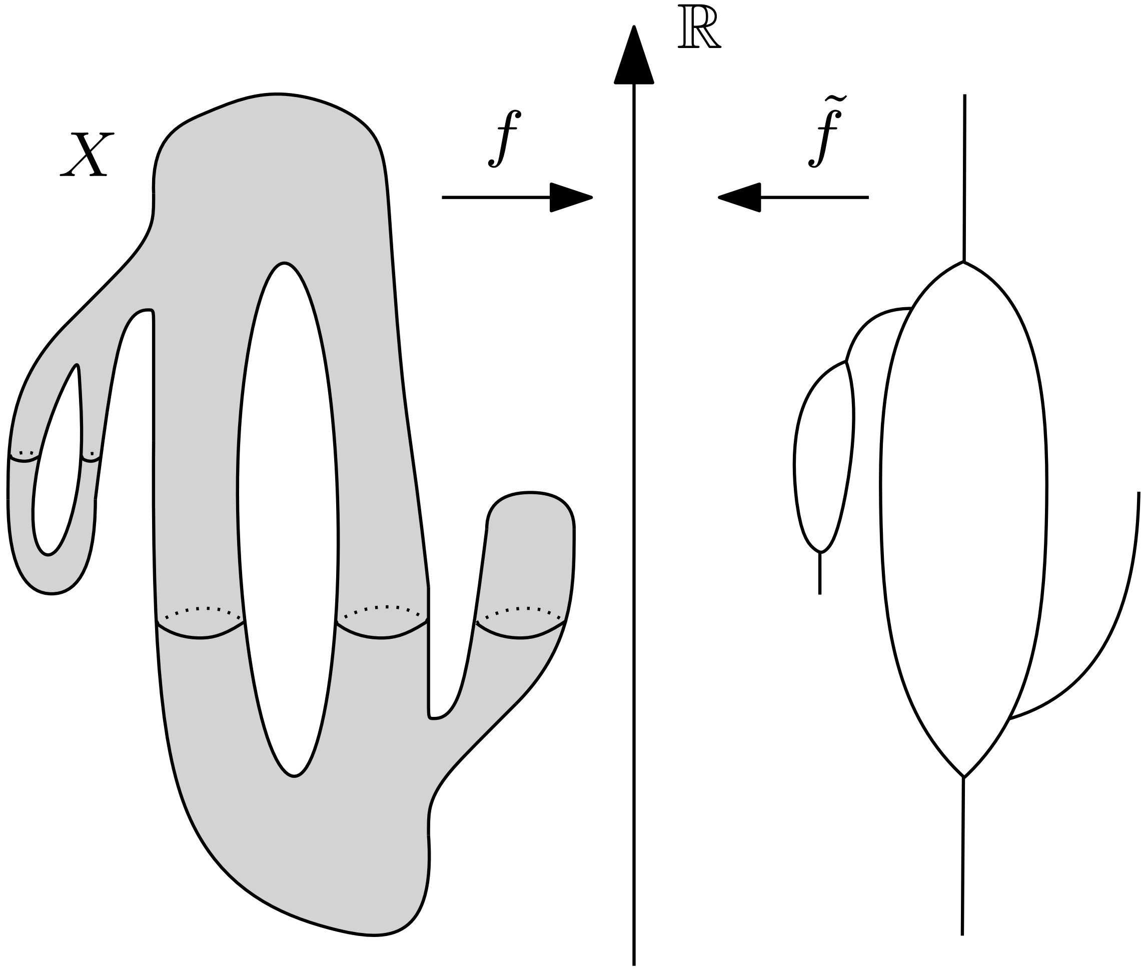
$$\begin{tikzcd} X \arrow[r, "f"] \arrow[d, "\pi"'] & \mathbb{R} \\ \mathrm{R}_f(X) \arrow[ur, "\tilde{f}"'] & \end{tikzcd}$$

Reeb Graph

[*Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique*, Reeb, C. R. Acad. Sci. Paris, 1946]

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

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



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

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

Prop: $H_*(\mathrm{R}_f(X)) = H_*(X)/\bar{H}_*(X)$.

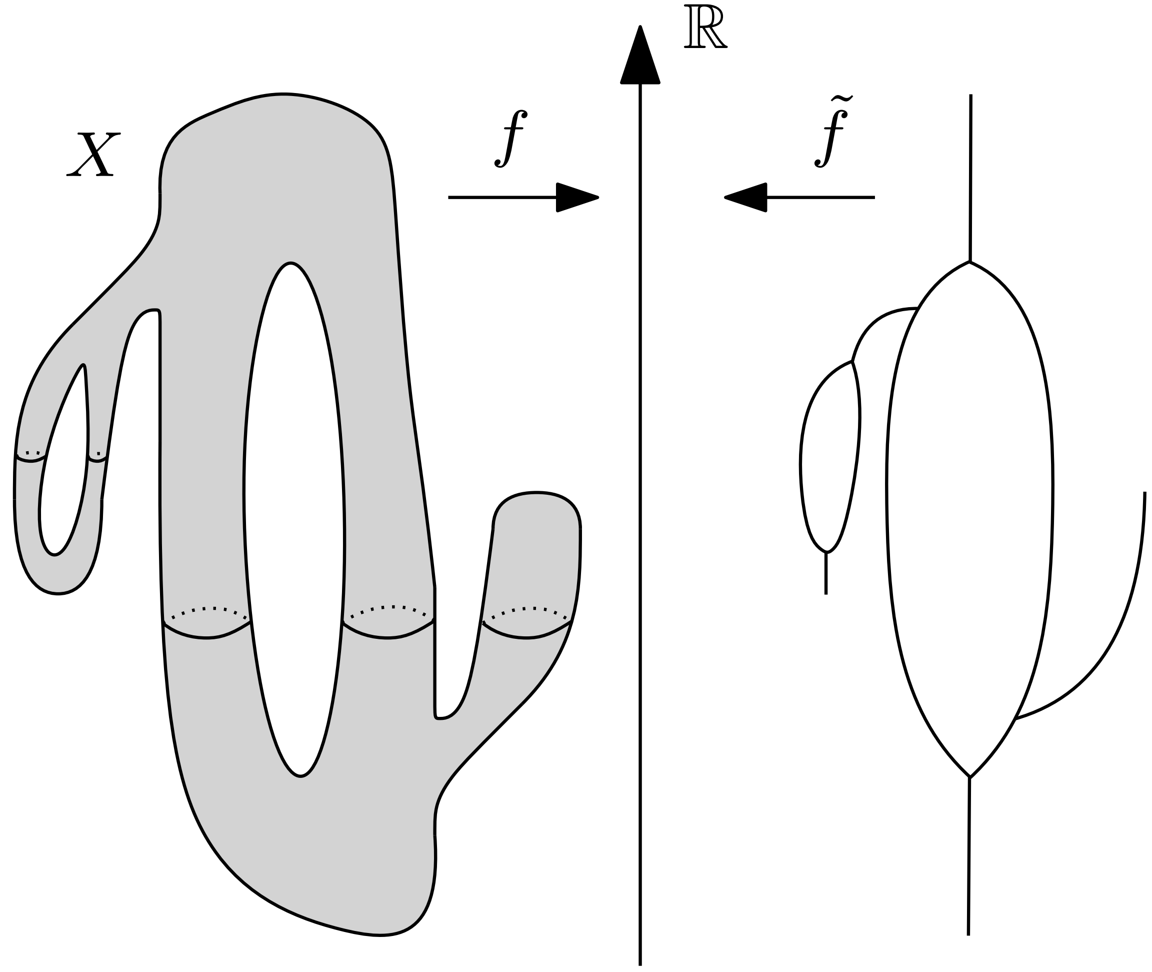
[*Reeb Graphs: Approximation and Persistence*, Dey, Wang, DCG, 2013]

Reeb Graph

[*Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique*, Reeb, C. R. Acad. Sci. Paris, 1946]

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

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



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

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

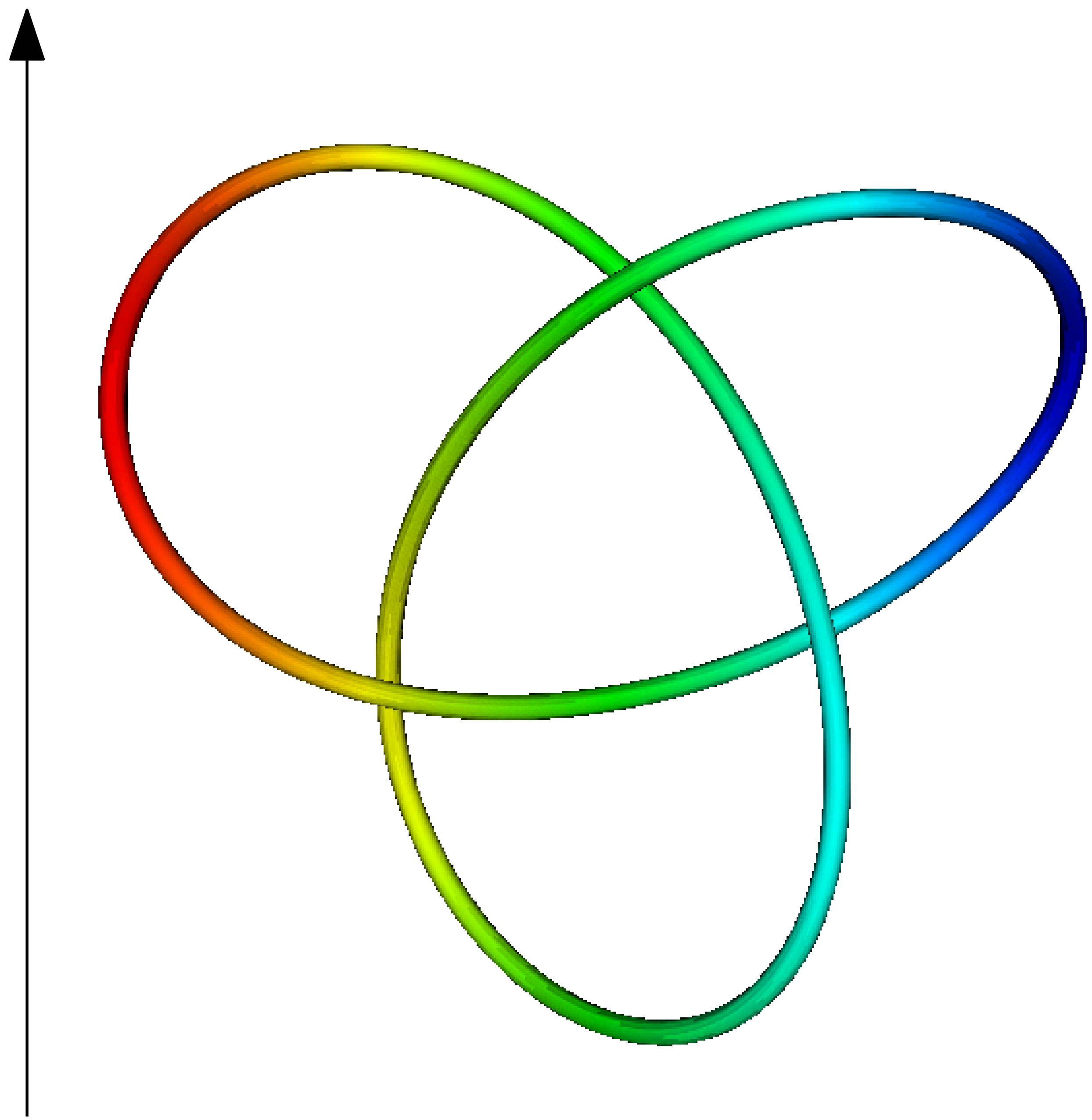
horizontal homology $\sim$ 'those homology classes that are included in a finite union of levelsets of f'

Prop: $H_*(\mathrm{R}_f(X)) = H_*(X)/\bar{H}_*(X)$.

[*Reeb Graphs: Approximation and Persistence*, Dey, Wang, DCG, 2013]

Reeb Graph

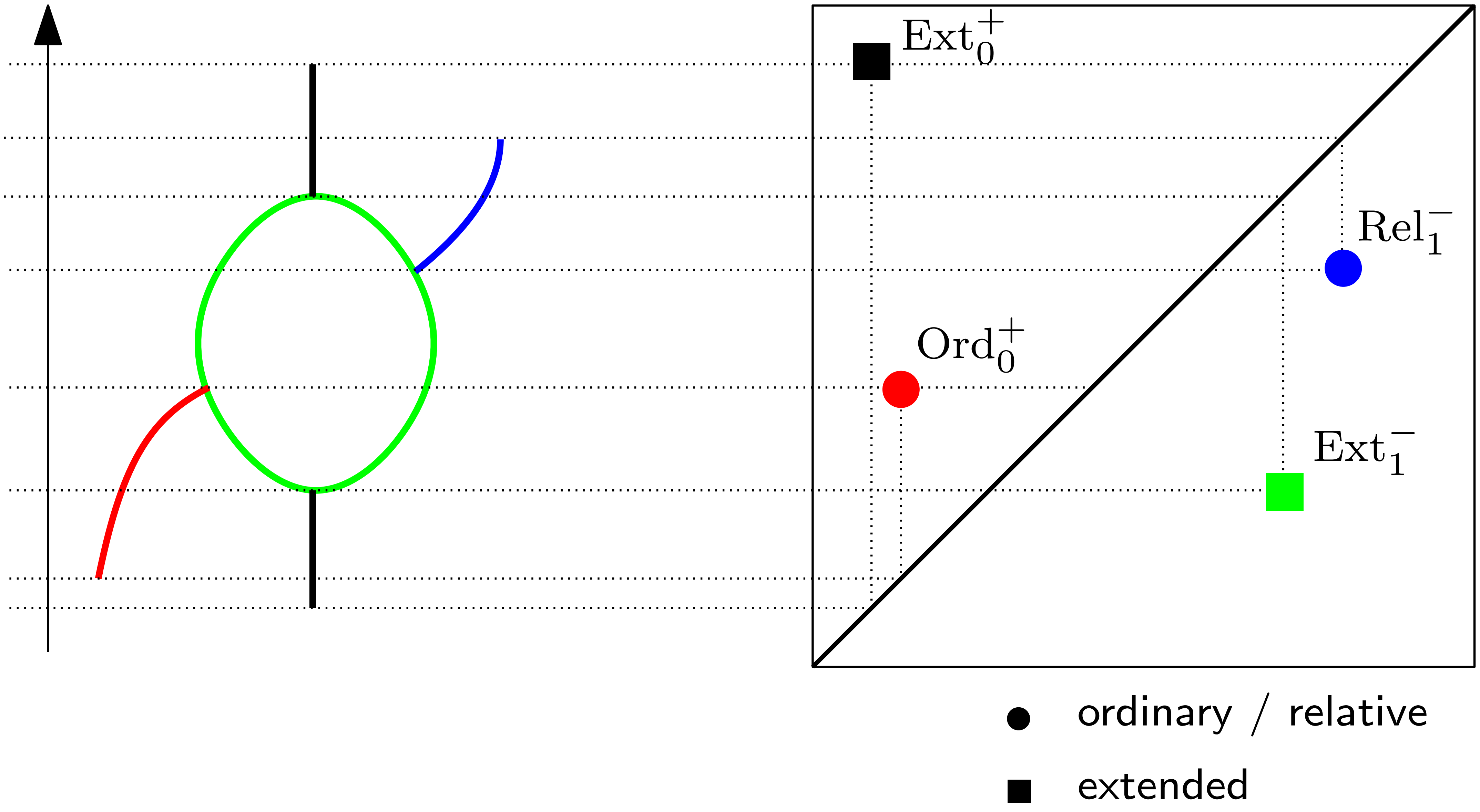
Q: What is the Reeb graph of the height function on the trefoil knot?



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

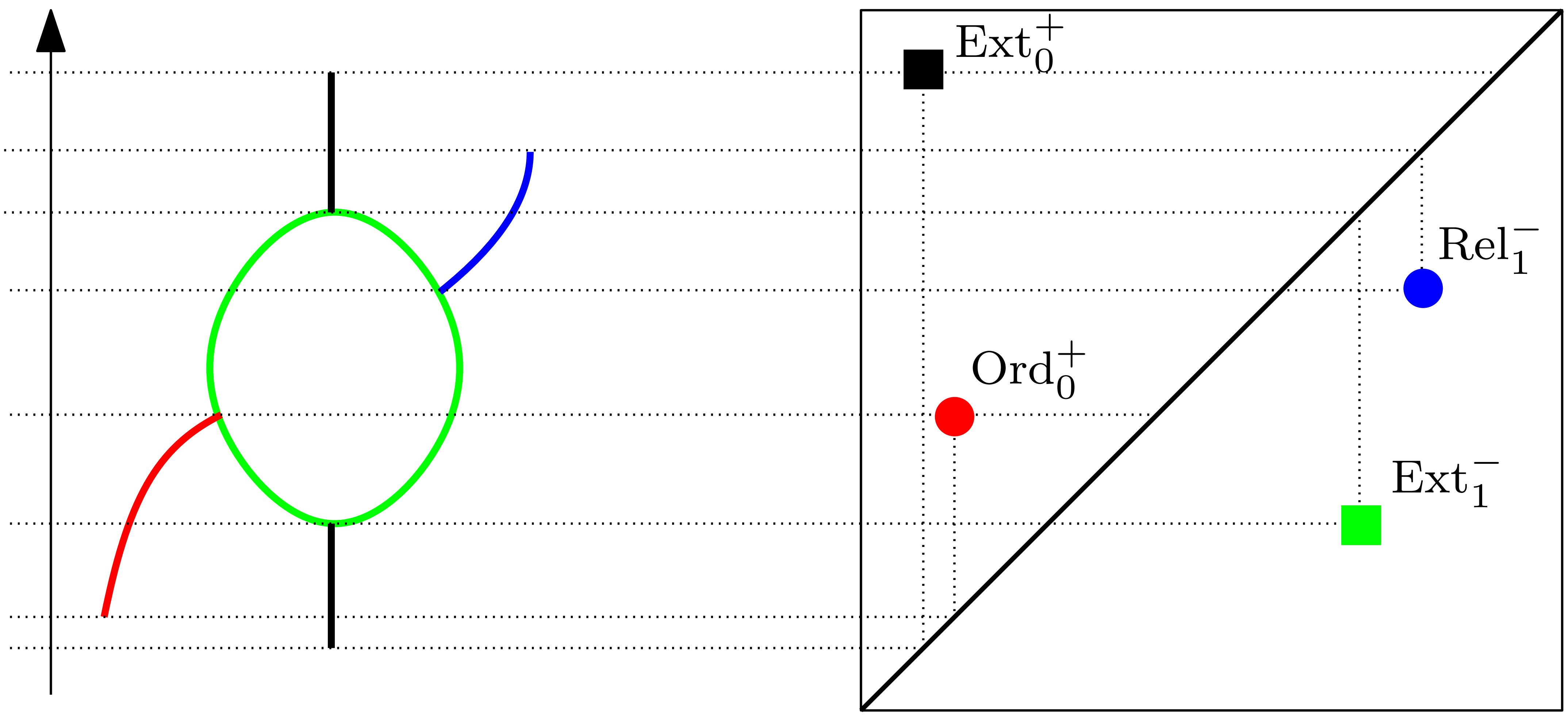
$\mathrm{Dg}\,\tilde{f}$ provides a **bag-of-features** descriptor for $\mathrm{R}_f(X)$:

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

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

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

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



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

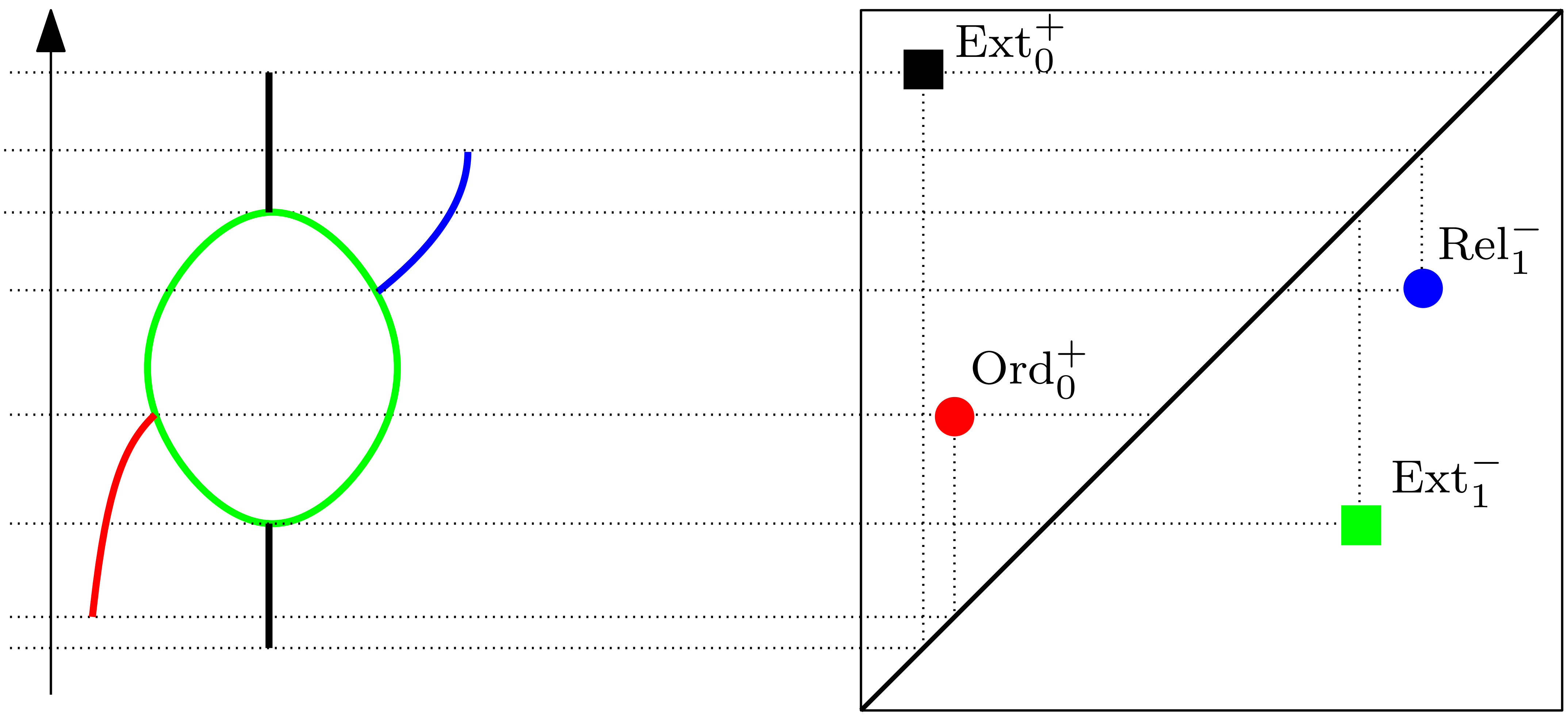
● ordinary / relative

■ extended

Graph Descriptor

Construction uses **extended persistence**,

using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



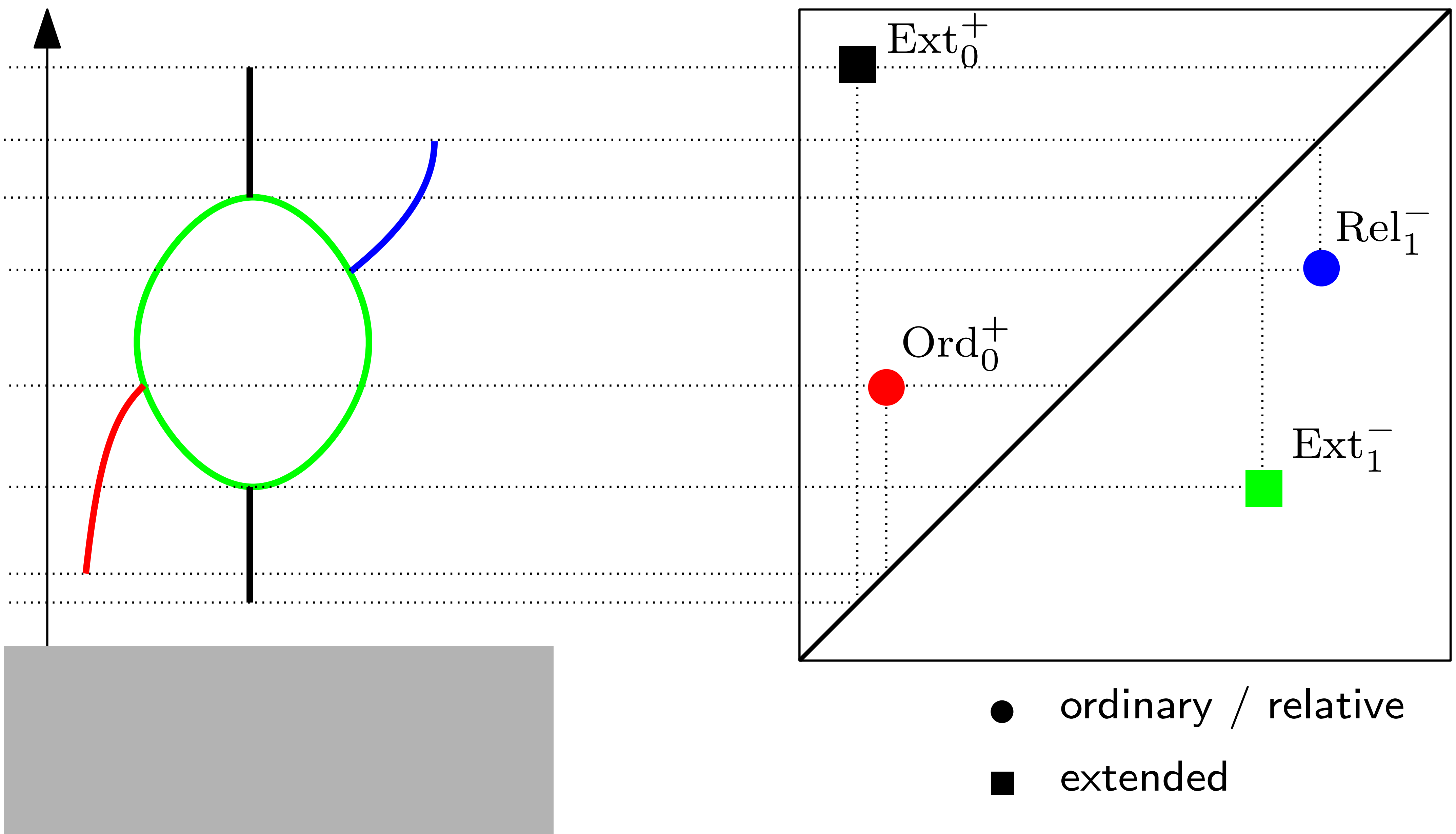
● ordinary / relative

■ extended

Graph Descriptor

Construction uses **extended persistence**,

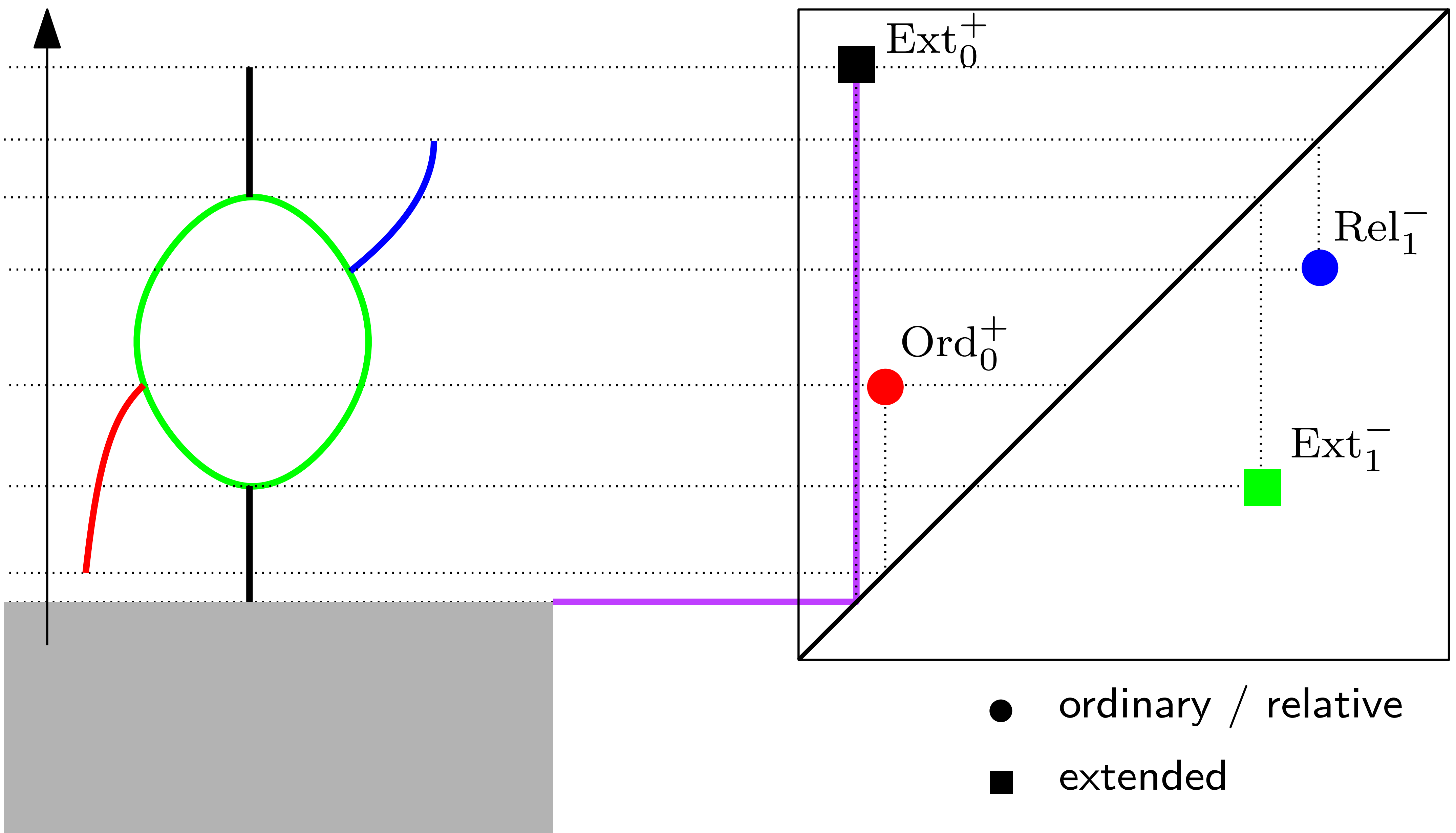
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

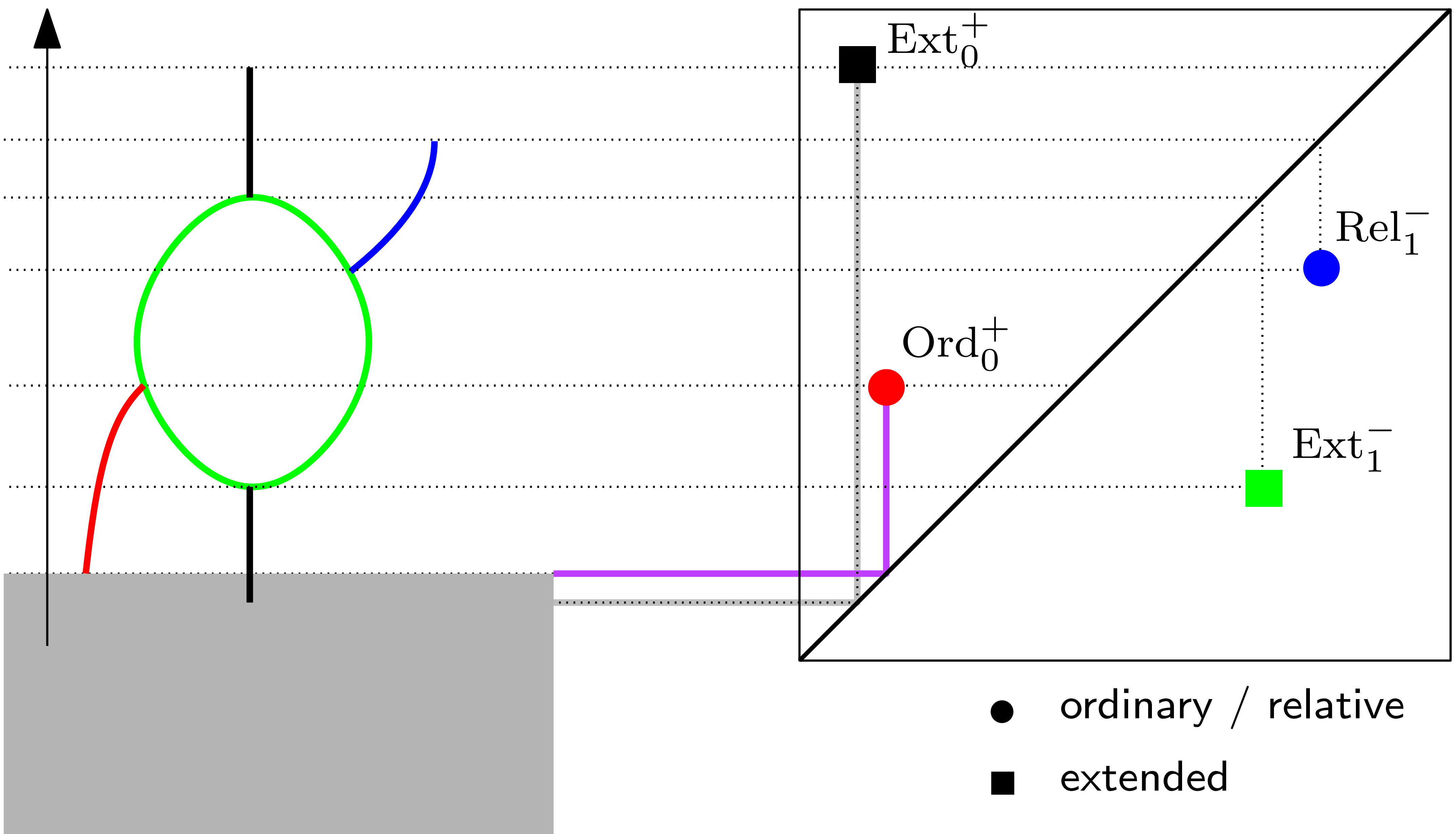
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

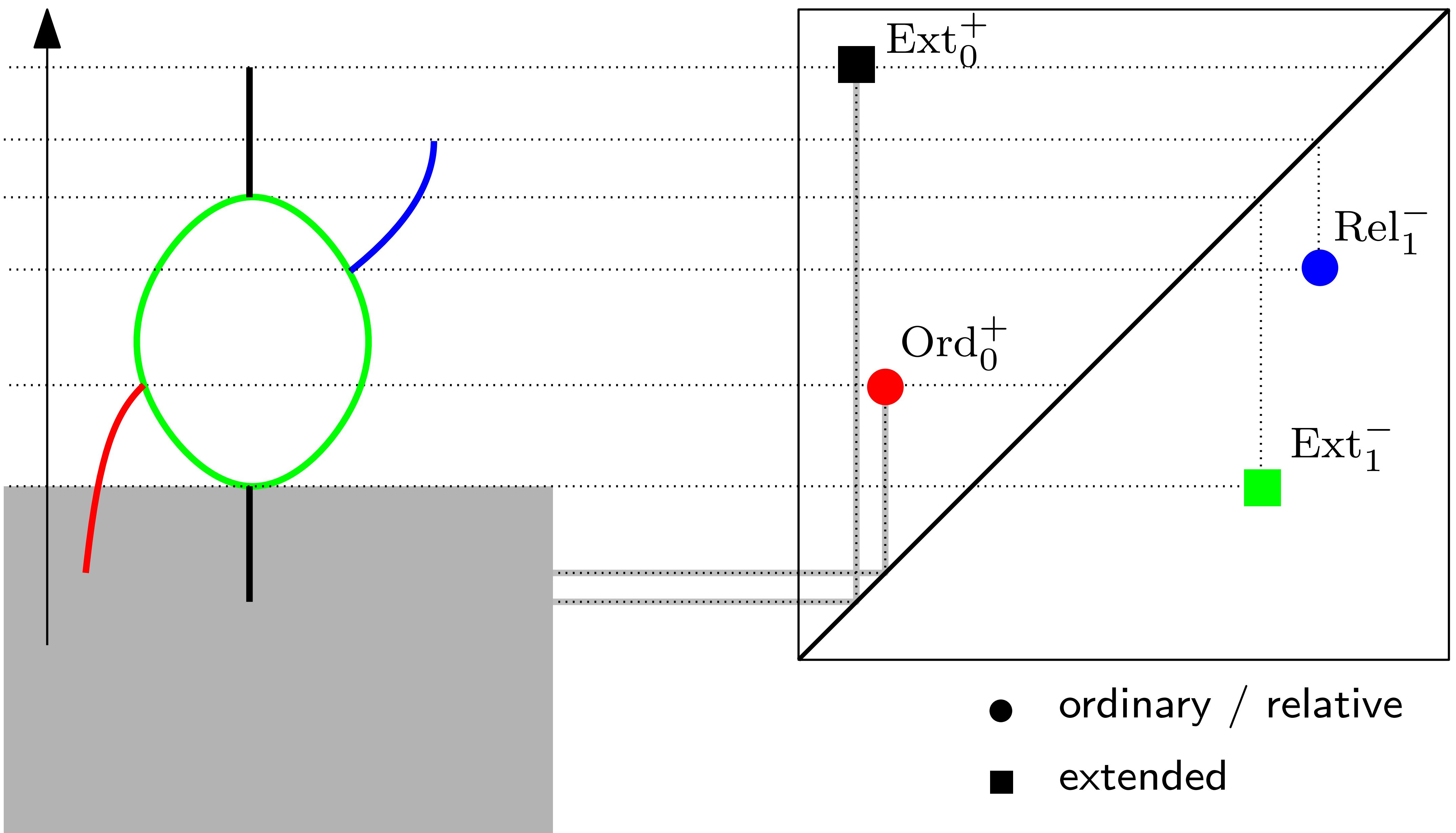
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

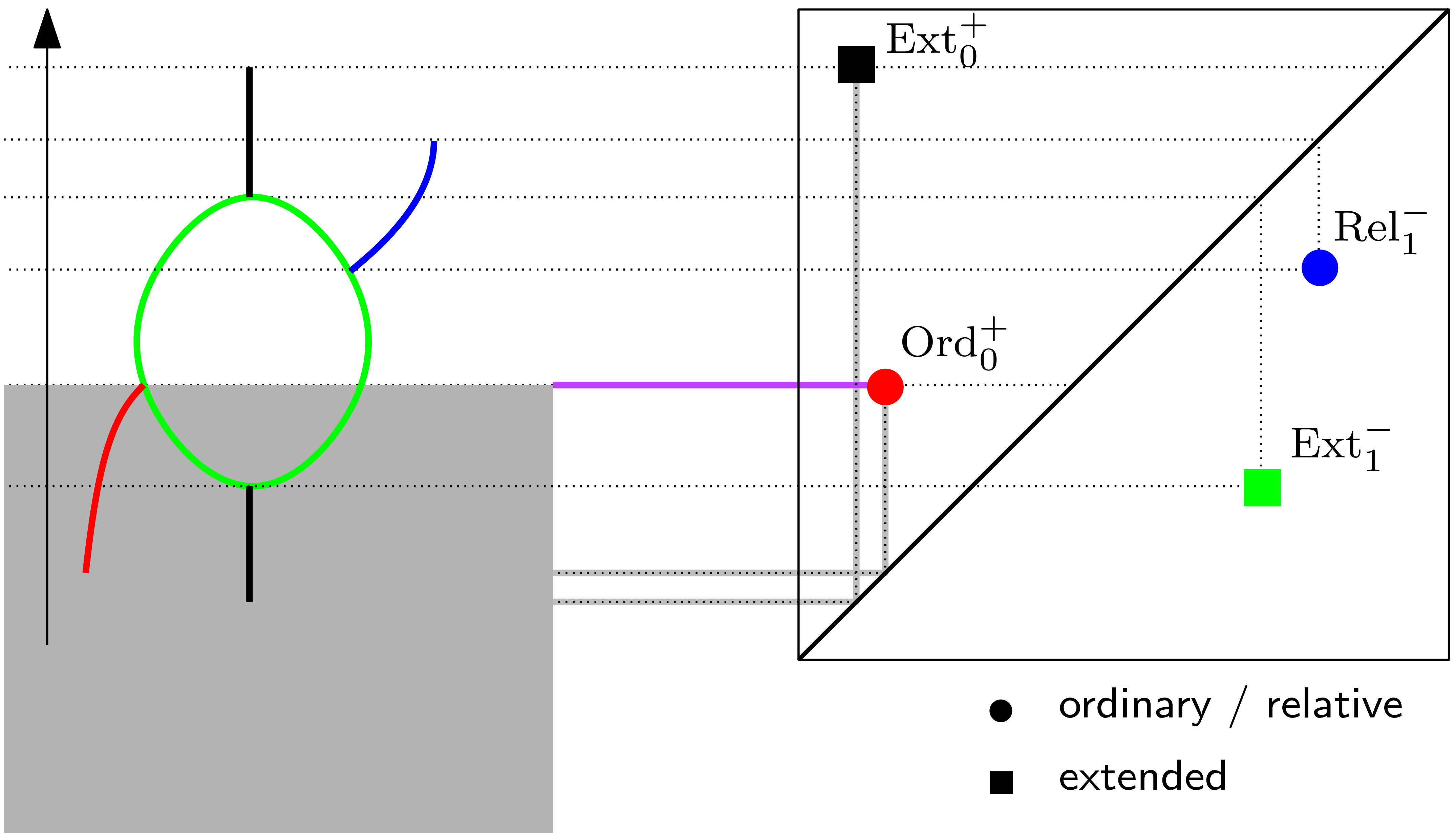
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

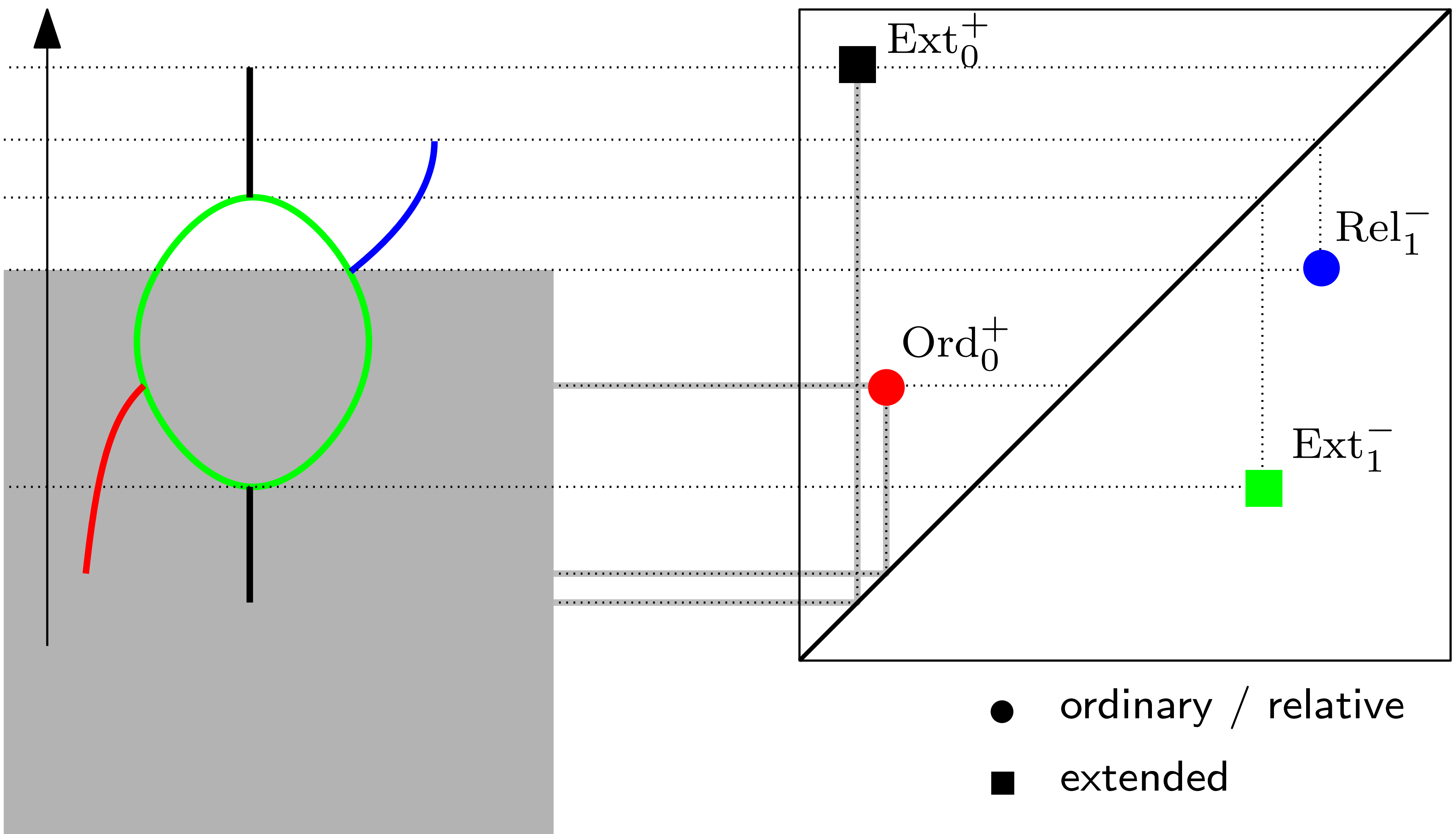
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

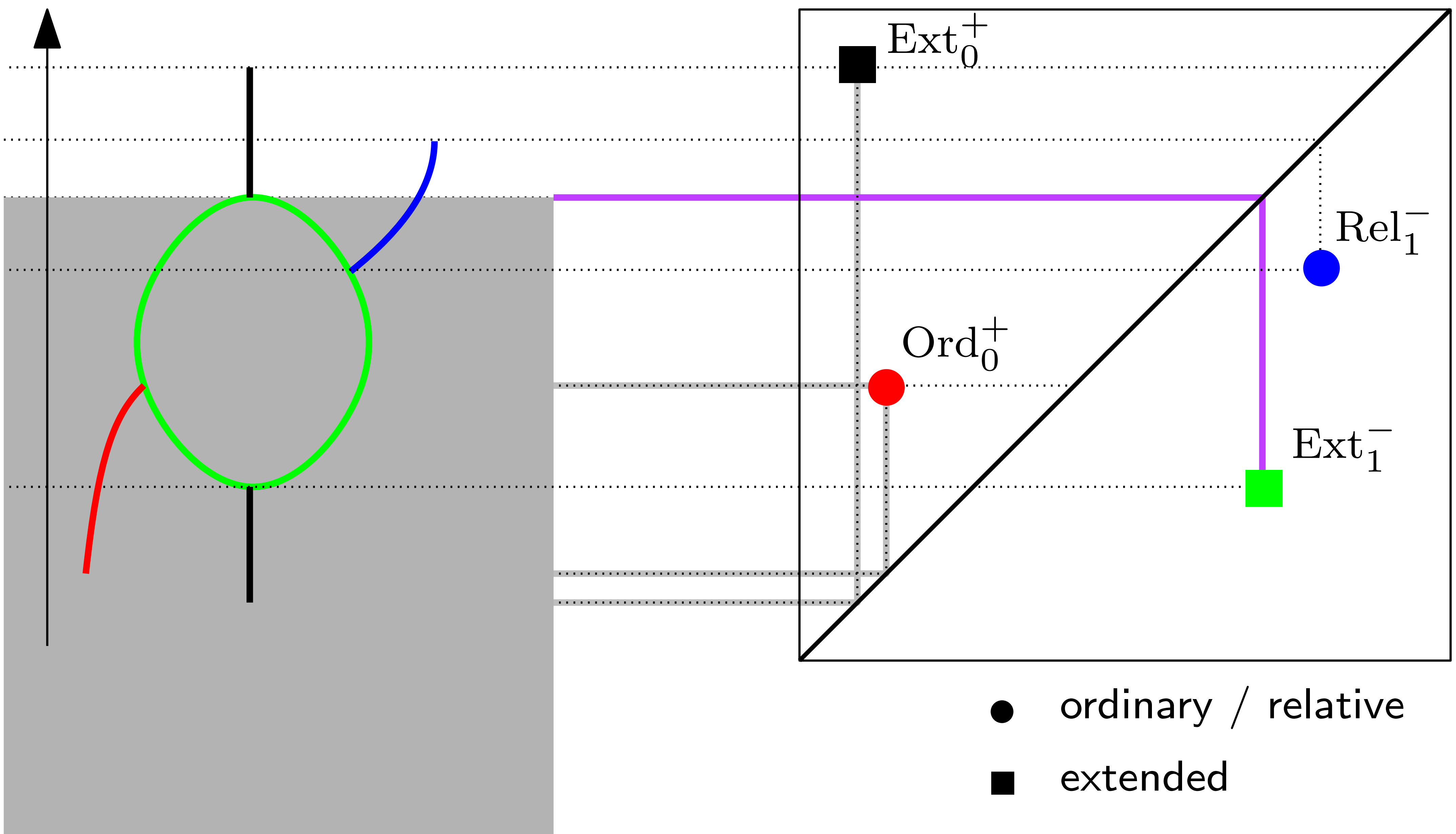
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

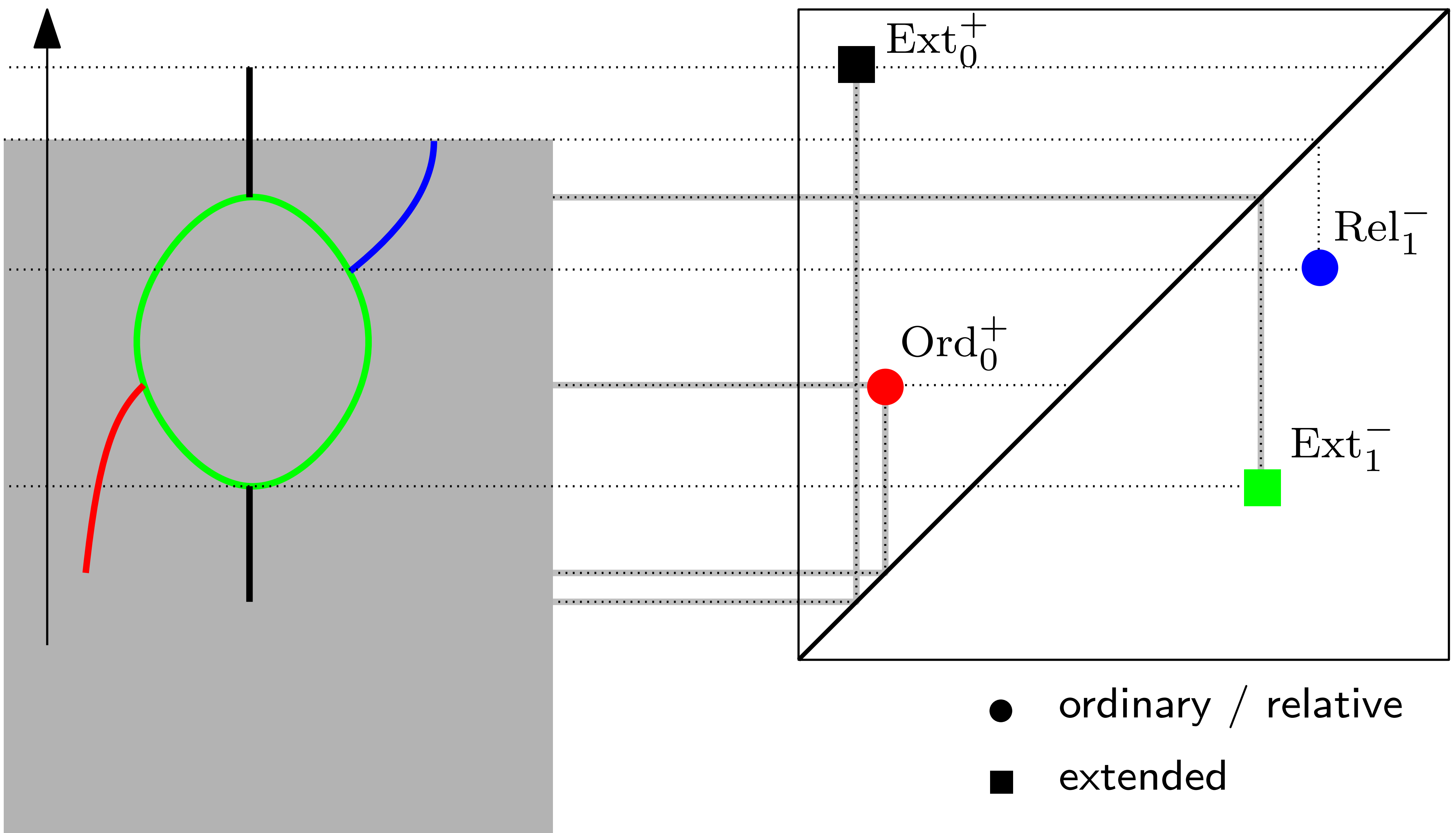
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

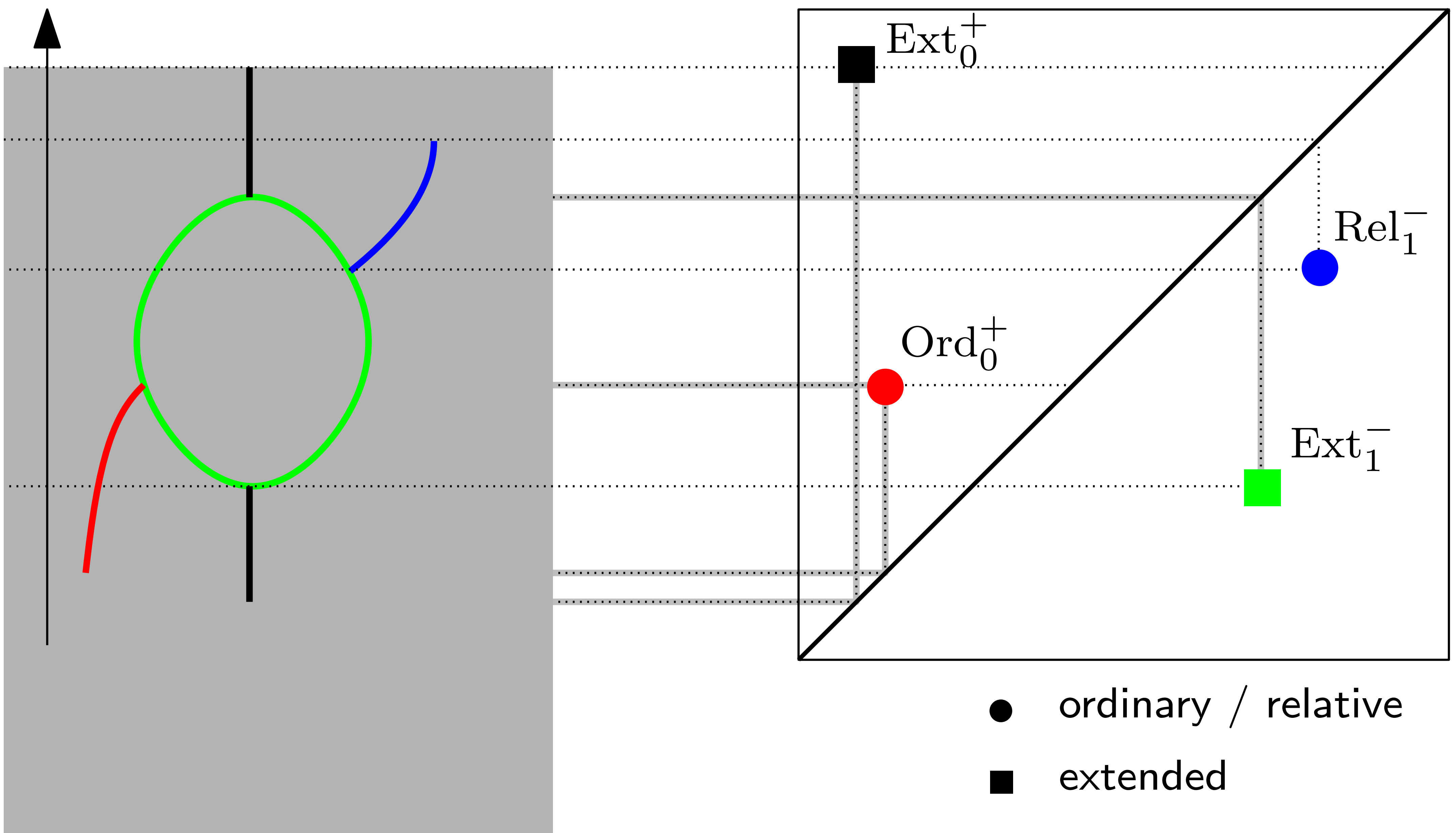
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

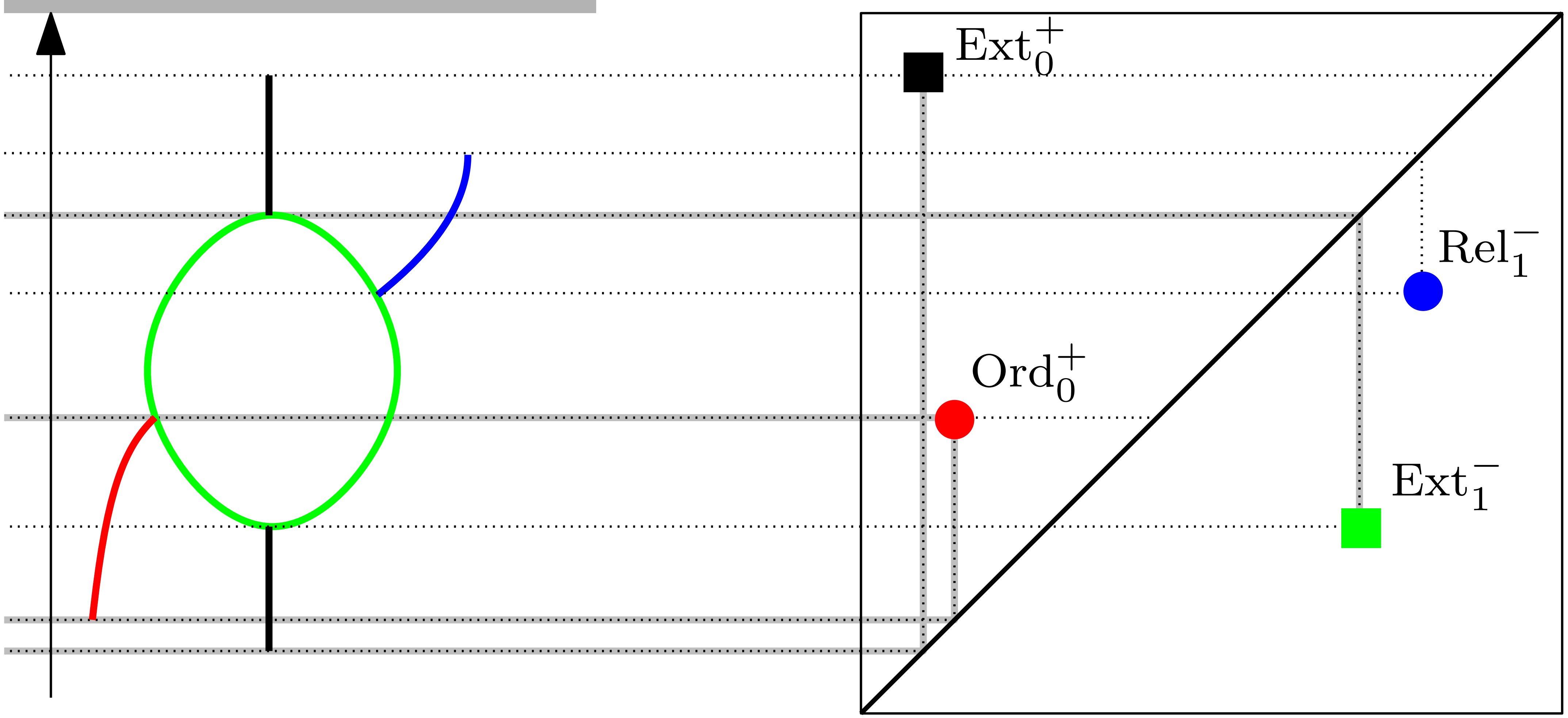
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph

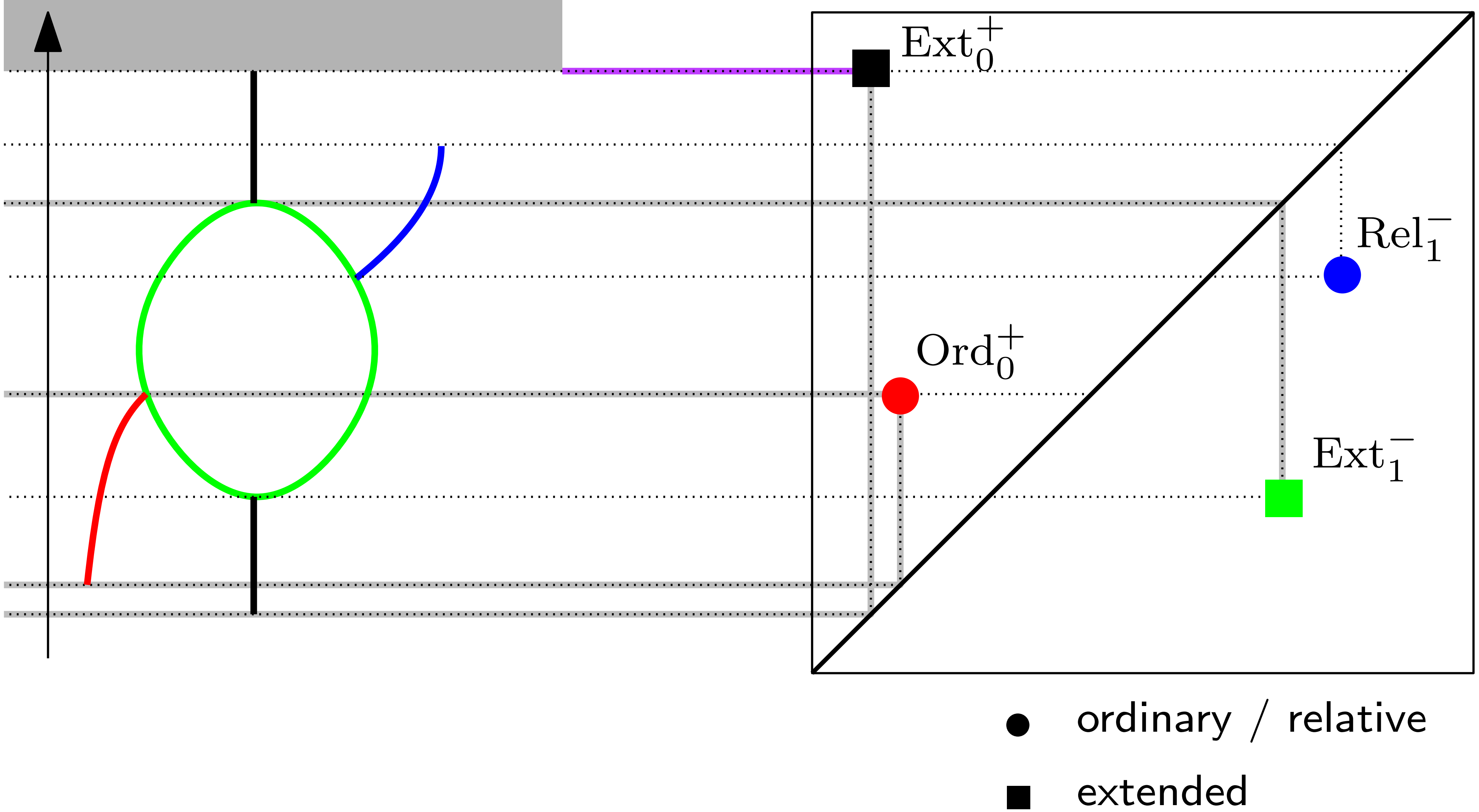


- ● ordinary / relative
- ■ extended

Graph Descriptor

Construction uses **extended persistence**,

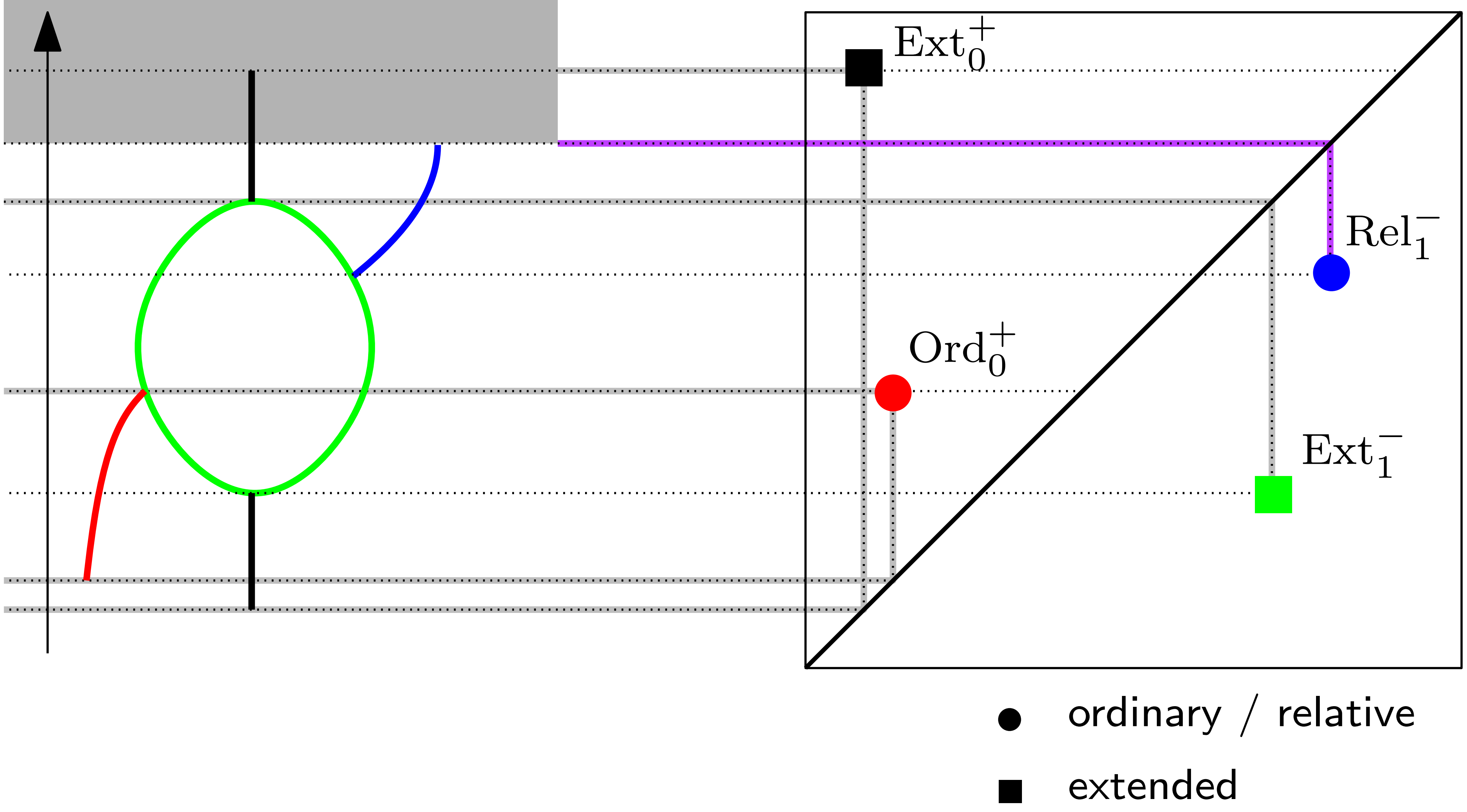
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

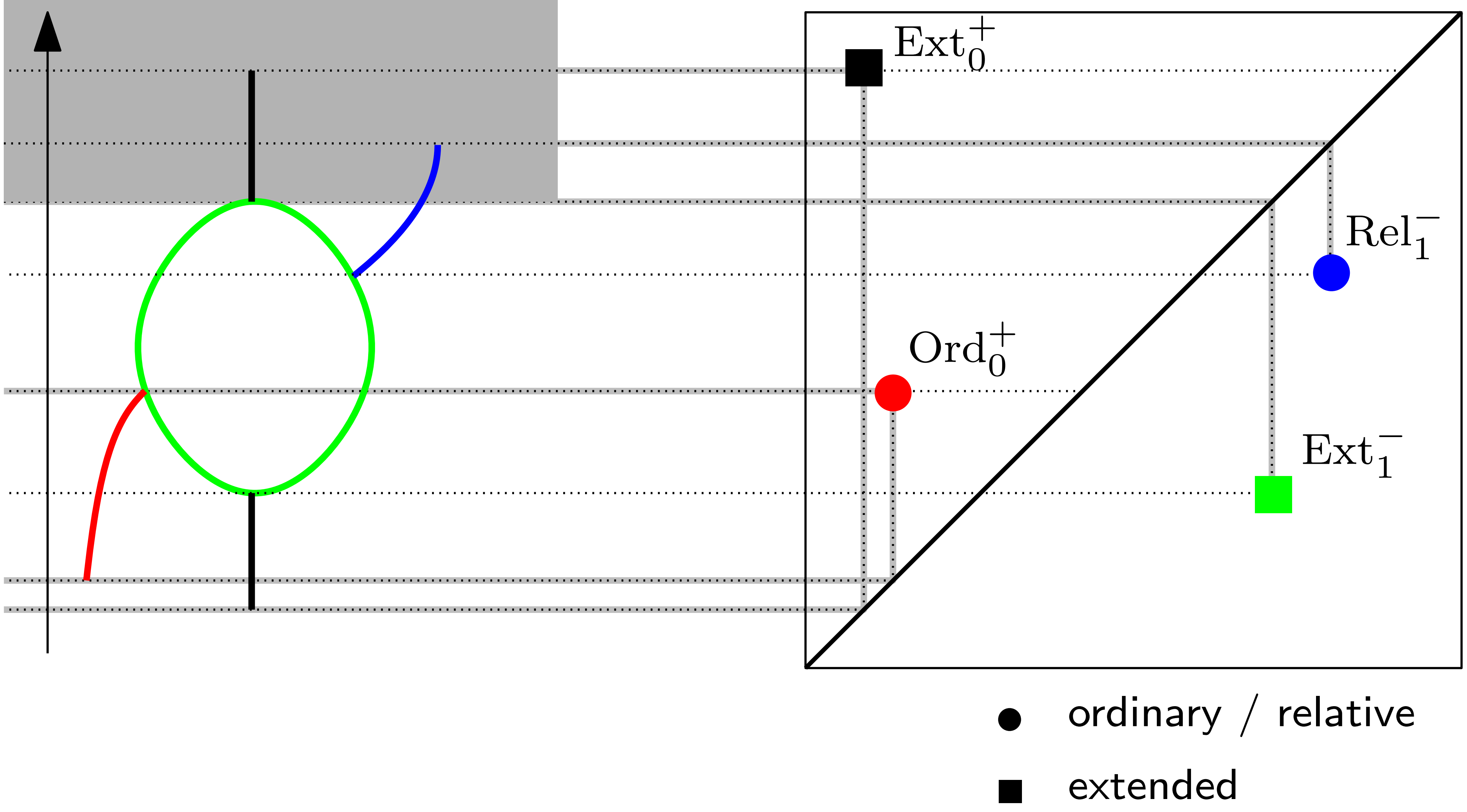
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

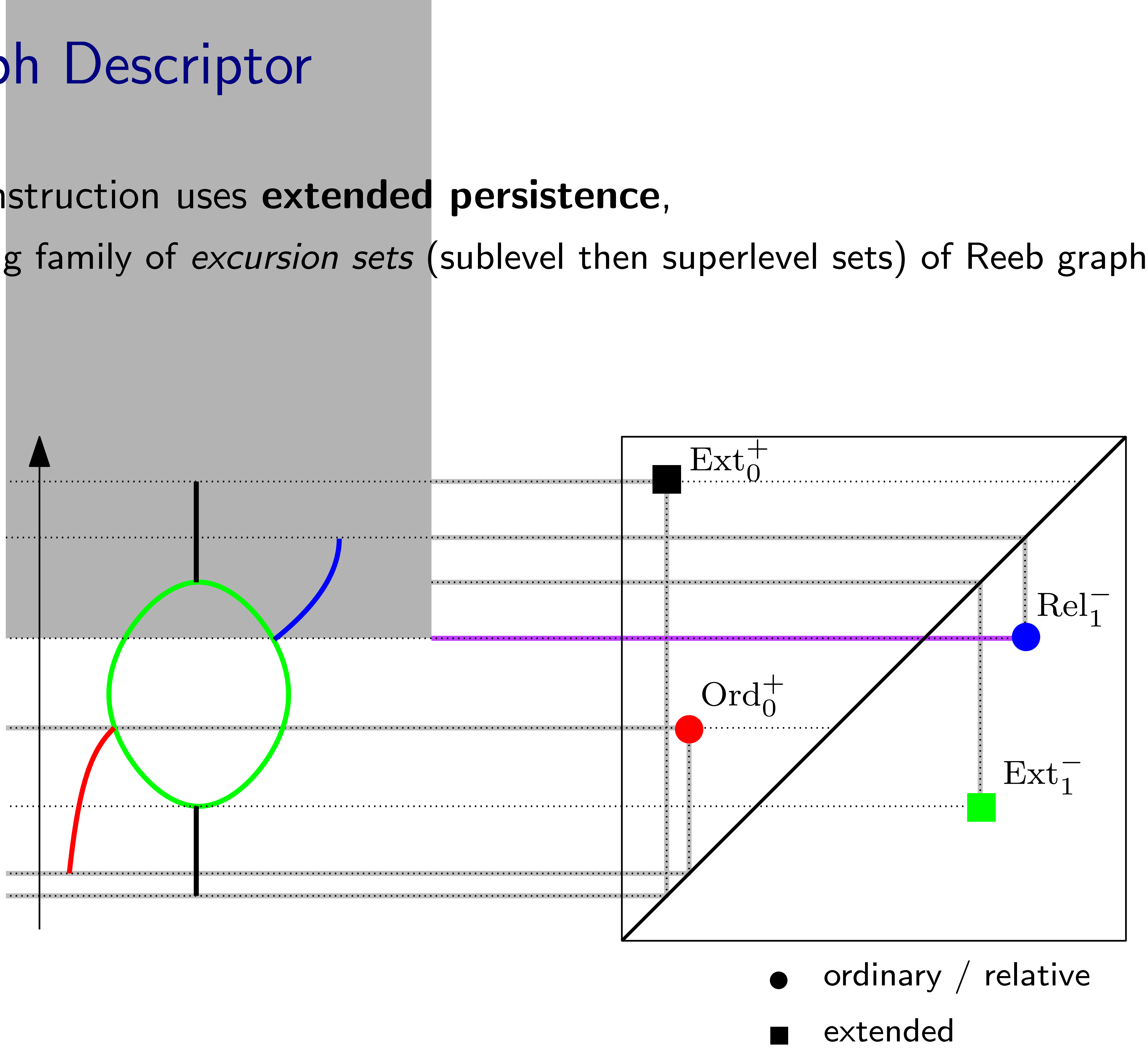
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

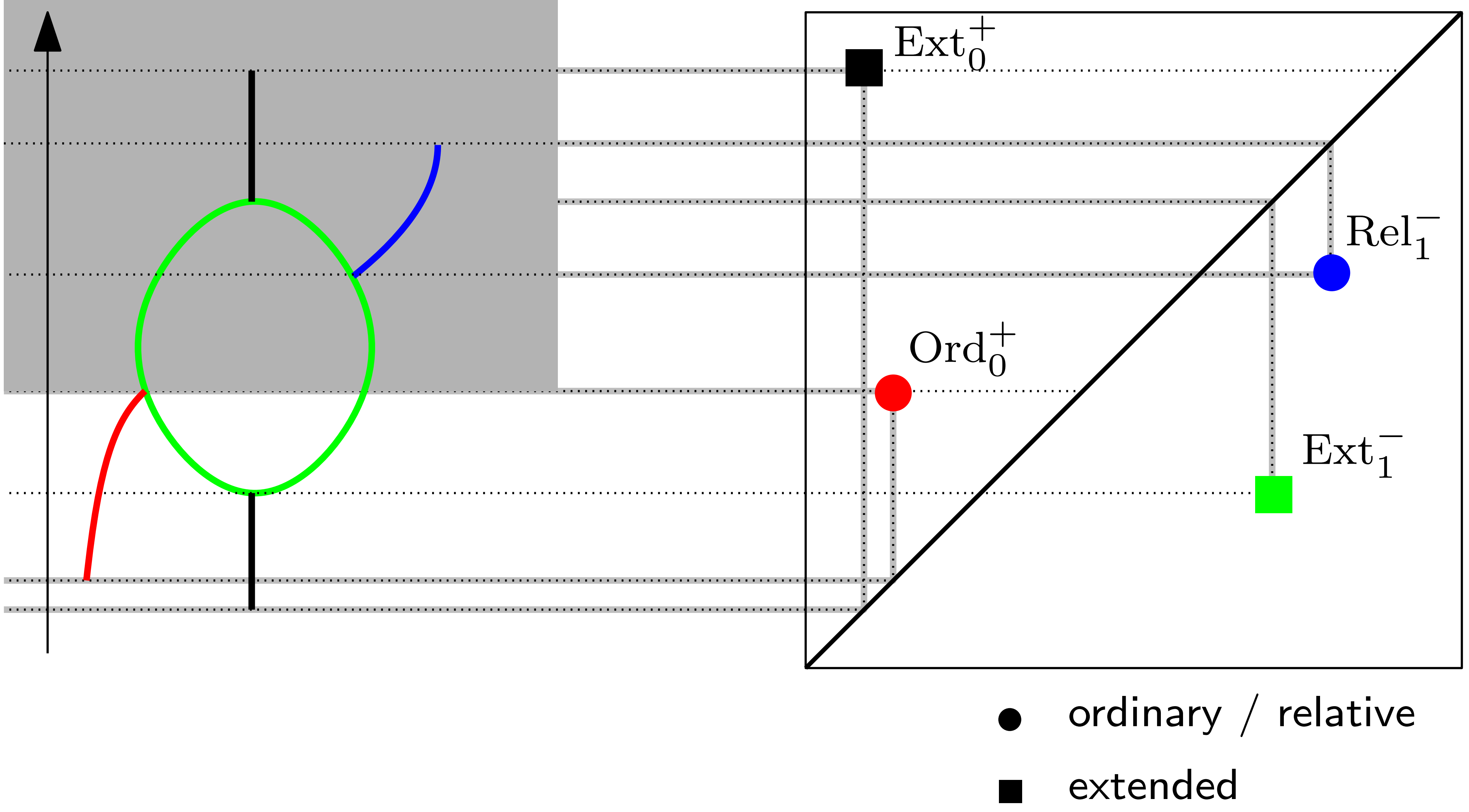
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

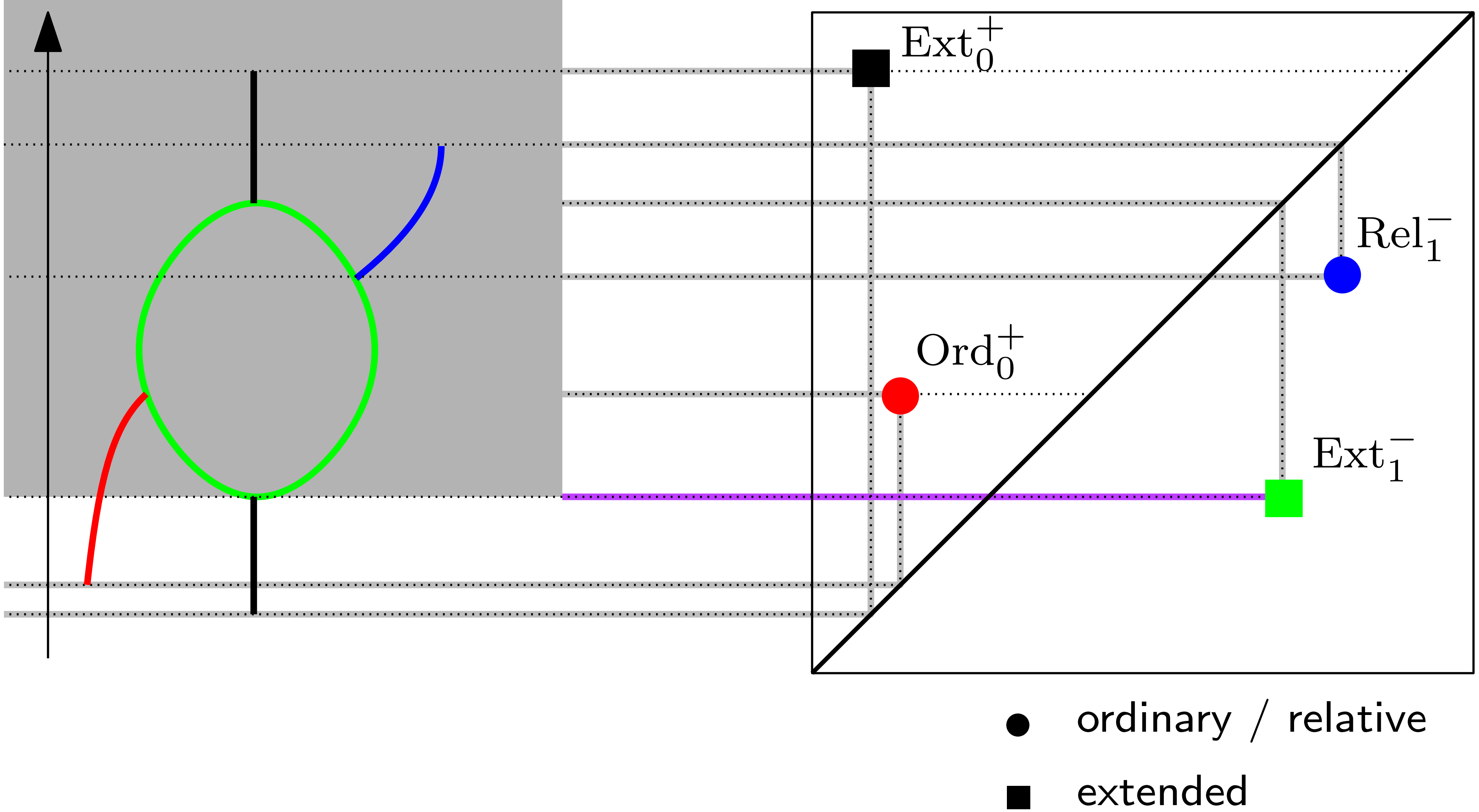
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

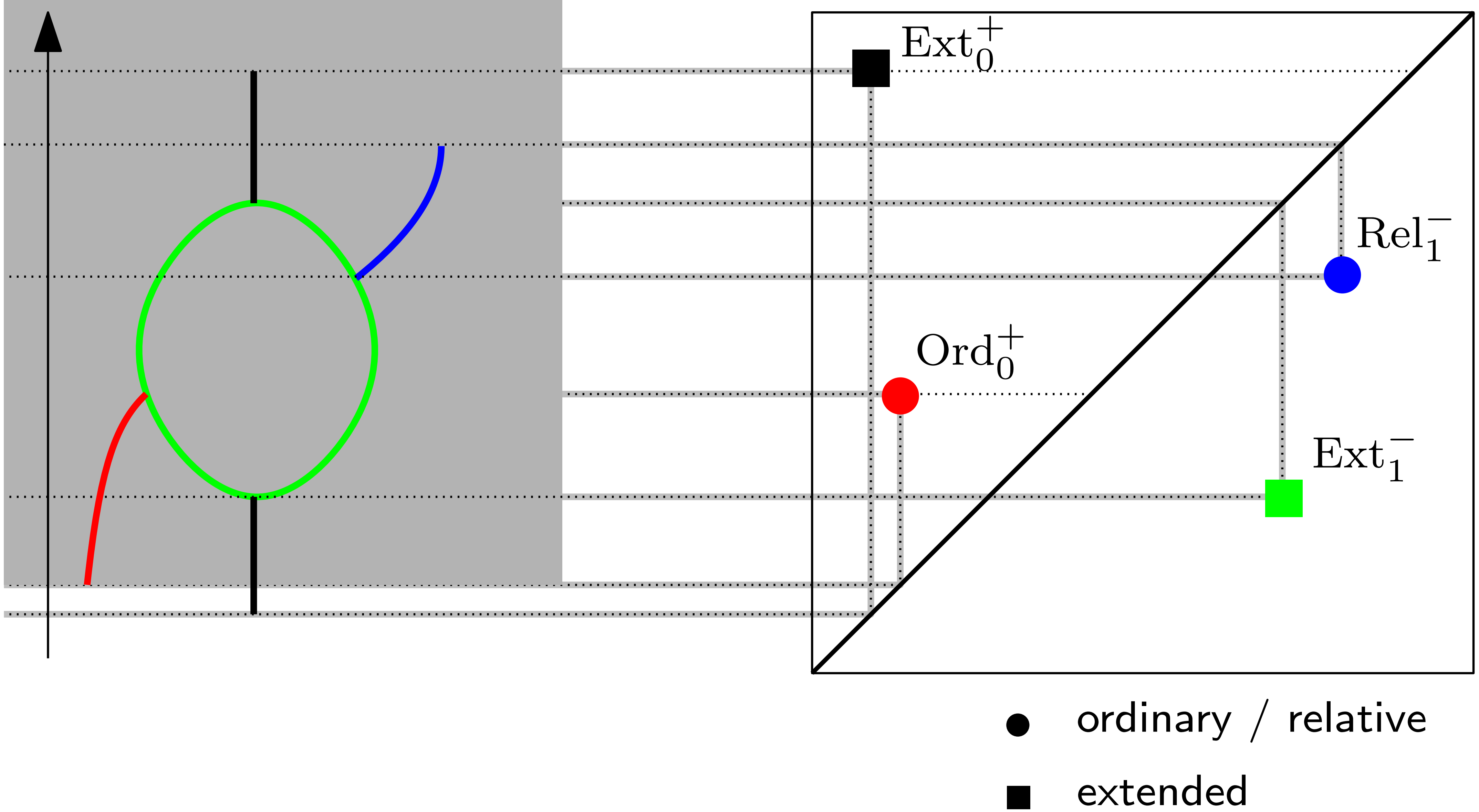
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

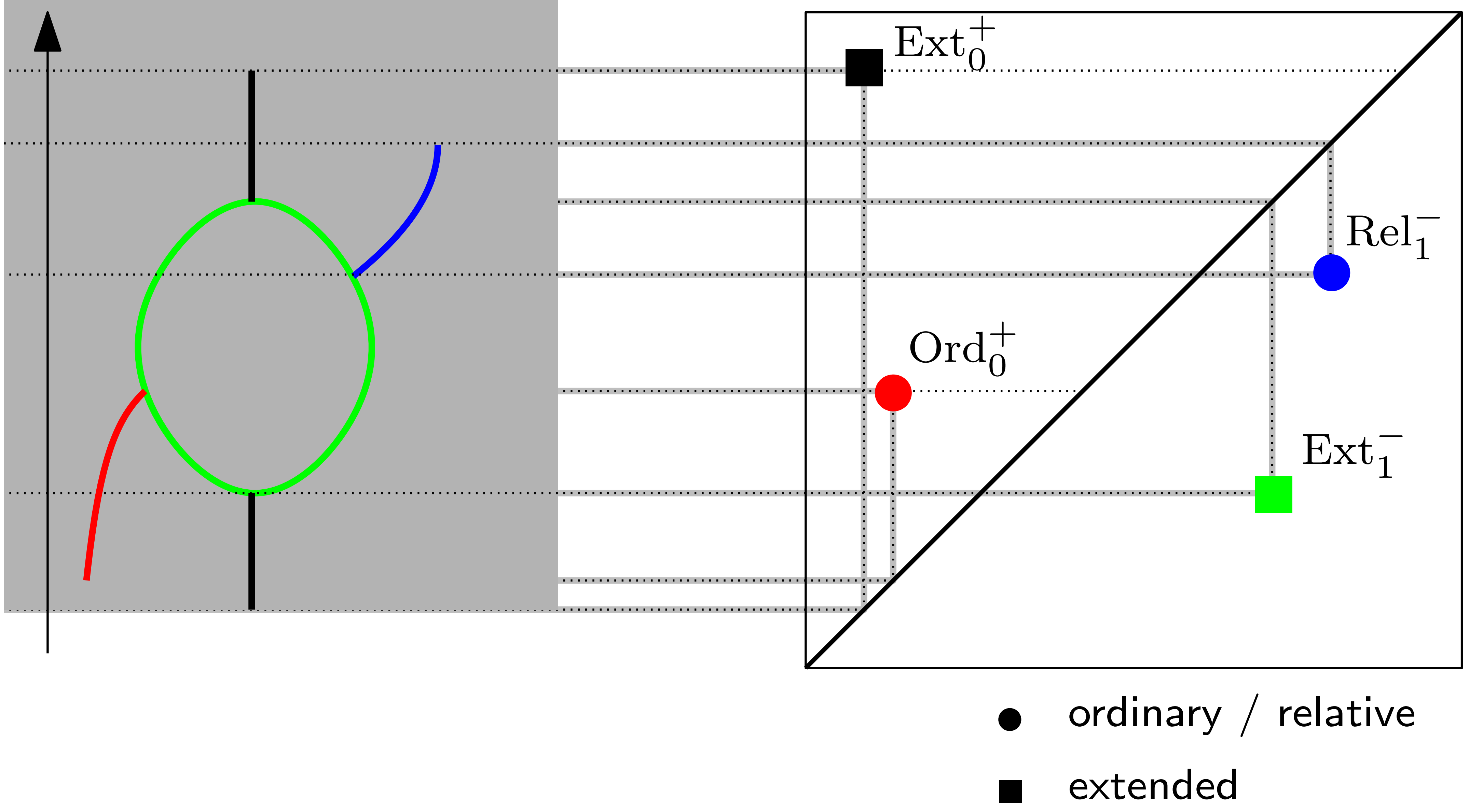
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

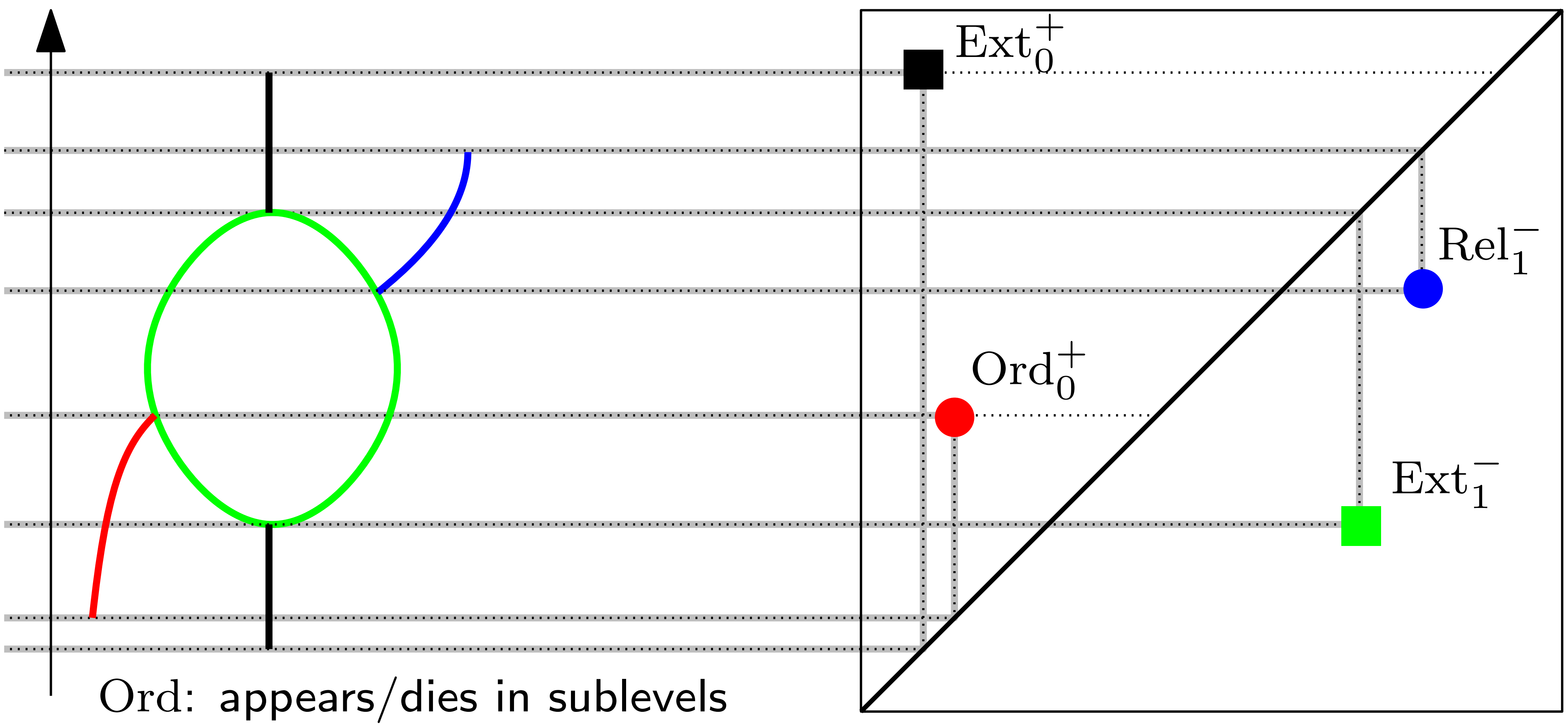
using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Graph Descriptor

Construction uses **extended persistence**,

using family of *excursion sets* (sublevel then superlevel sets) of Reeb graph



Ord: appears/dies in sublevels

Rel: appears/dies in superlevels

Ext: appears in sublevels, dies in superlevels

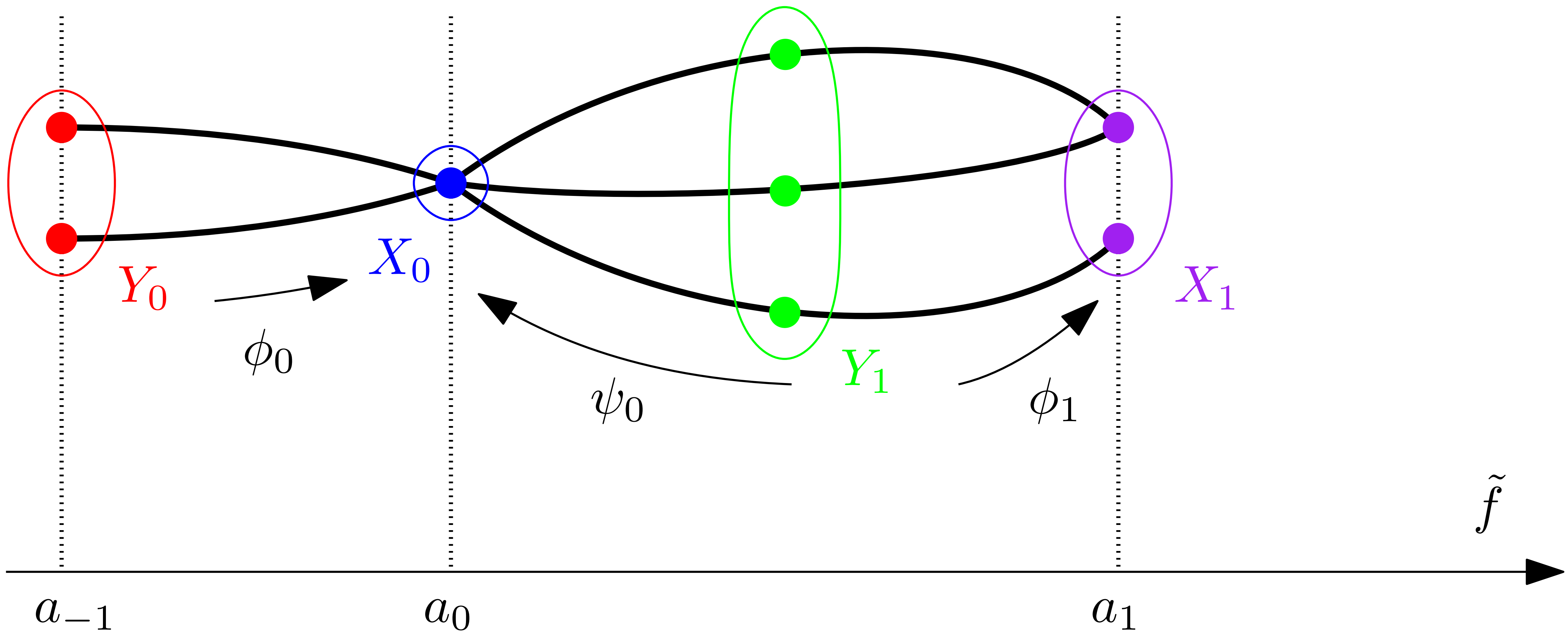
● ordinary / relative

■ extended

Graph Stratification

Reeb graph is a *telescope* (stratified space)

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



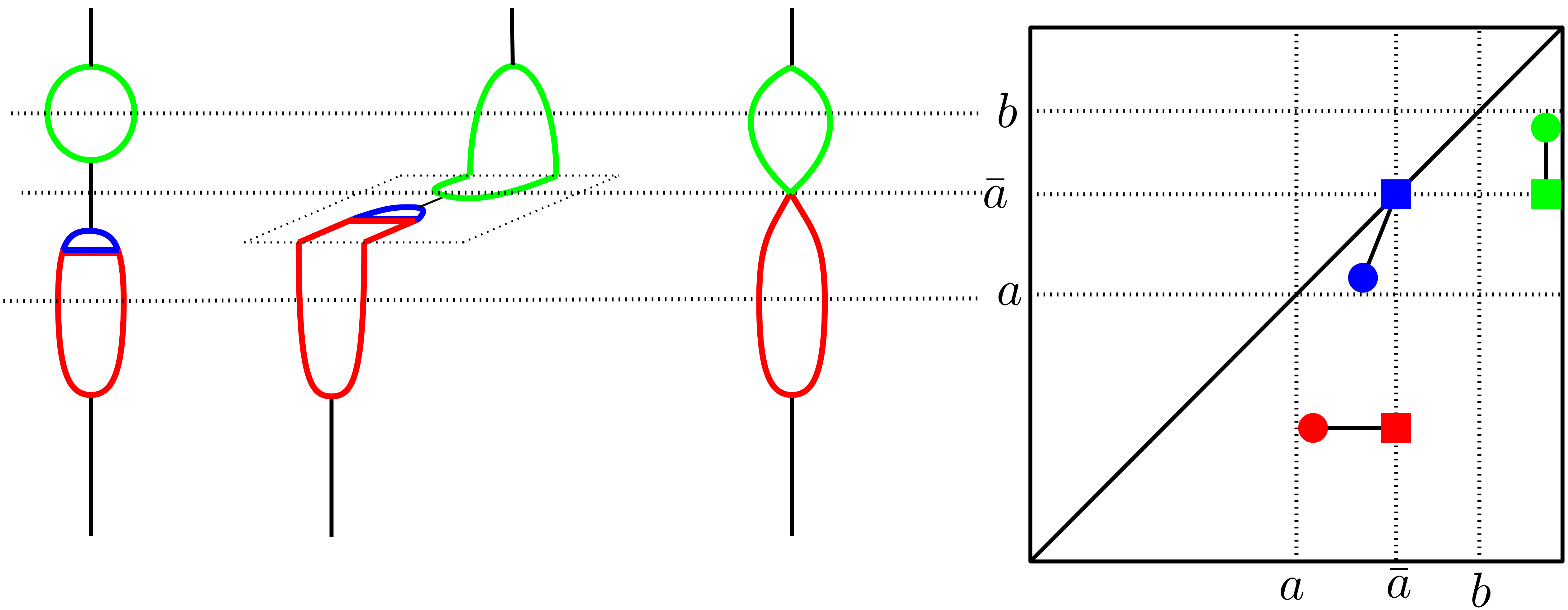
Idea: deform the Reeb graph so that it becomes the Mapper and track the changes in the persistence diagram

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

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

$$\downarrow$$

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

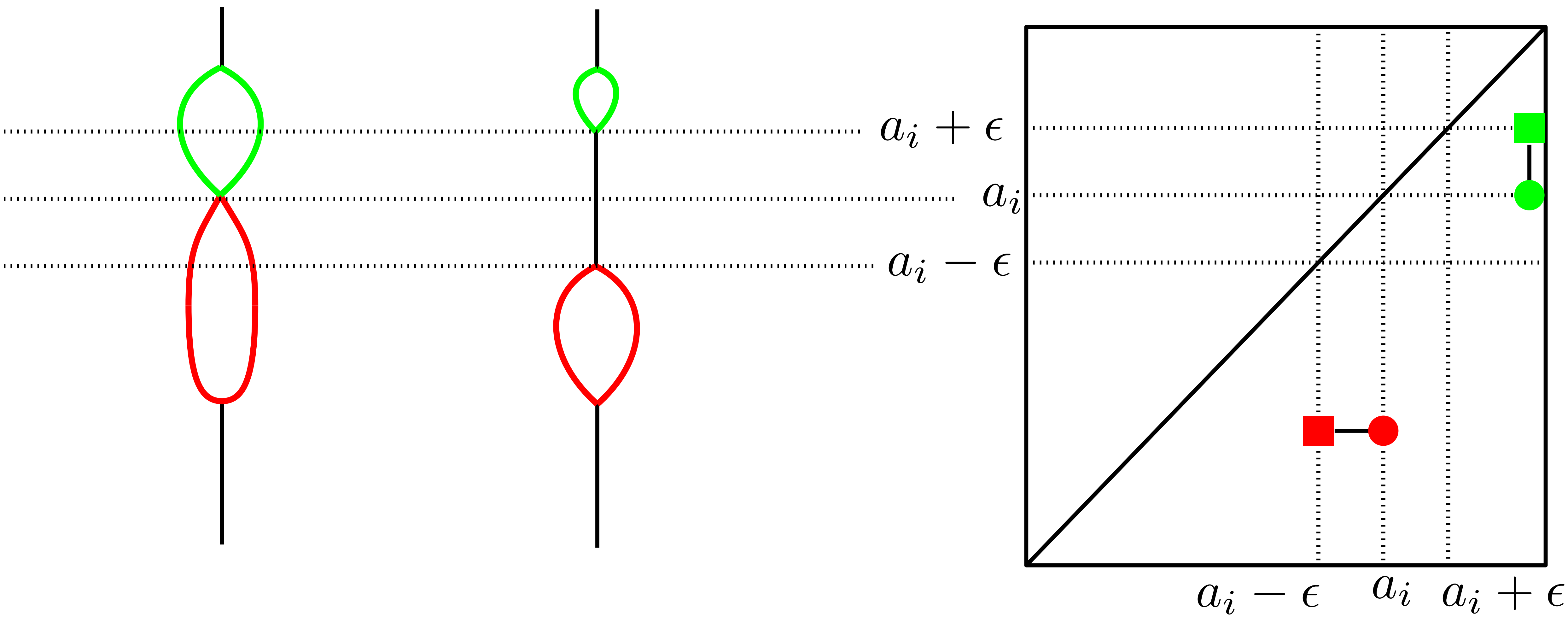


Operation 2: Split $Sp_{a_i,\epsilon}$

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

$$\downarrow$$

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

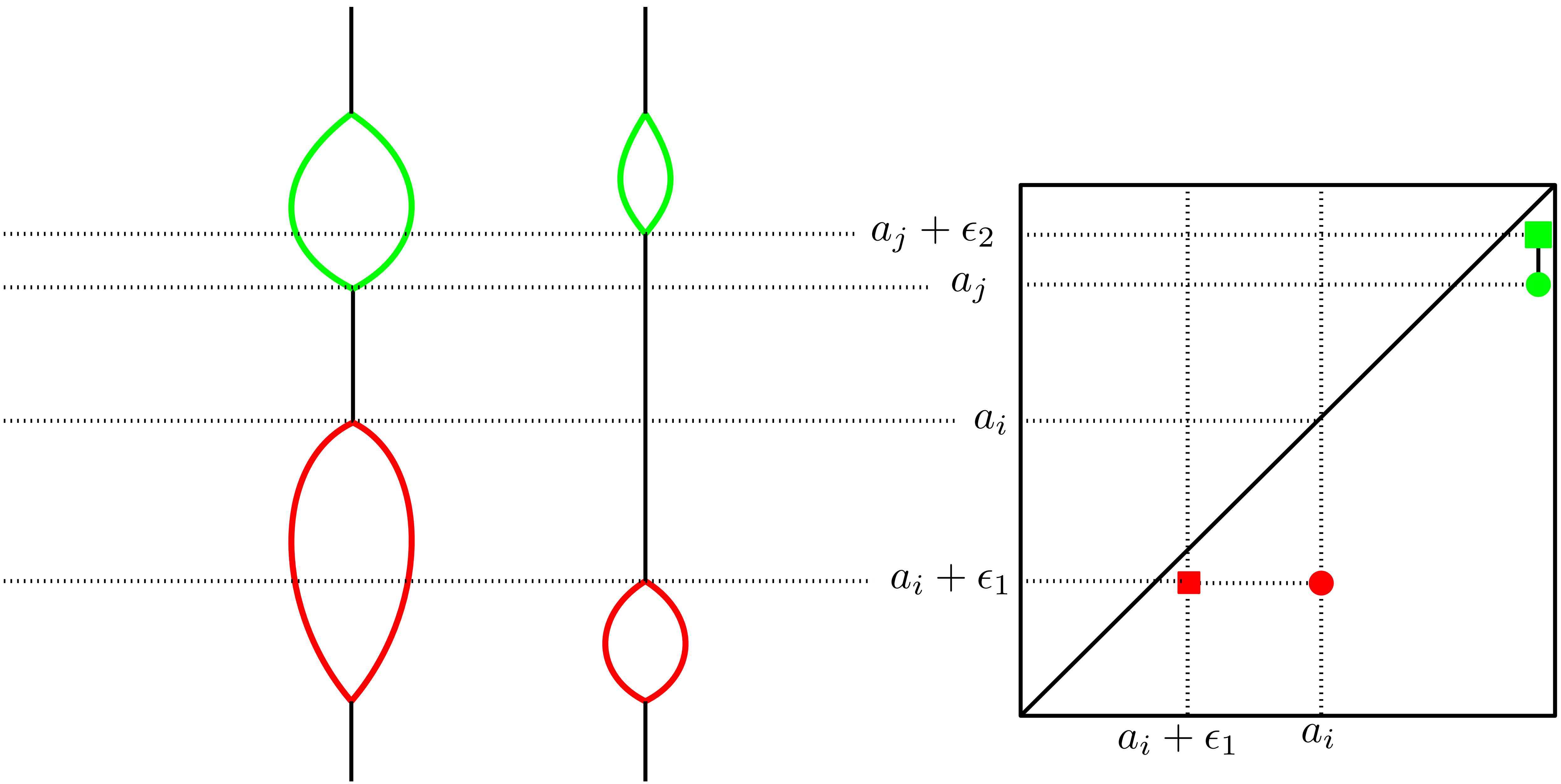


Operation 3: Shift $Sh_{a_i,\epsilon}$

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

$$\downarrow$$

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



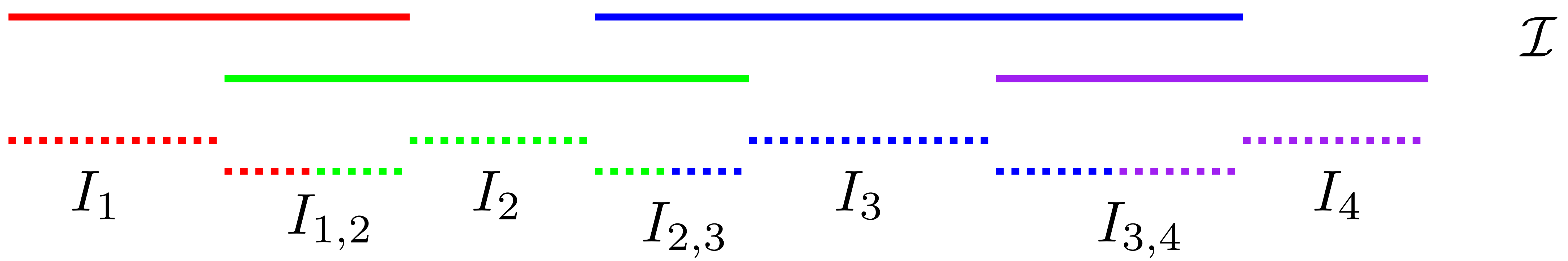
Formula Reeb graph $\rightarrow$ Mapper

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

Formula Reeb graph $\rightarrow$ Mapper

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

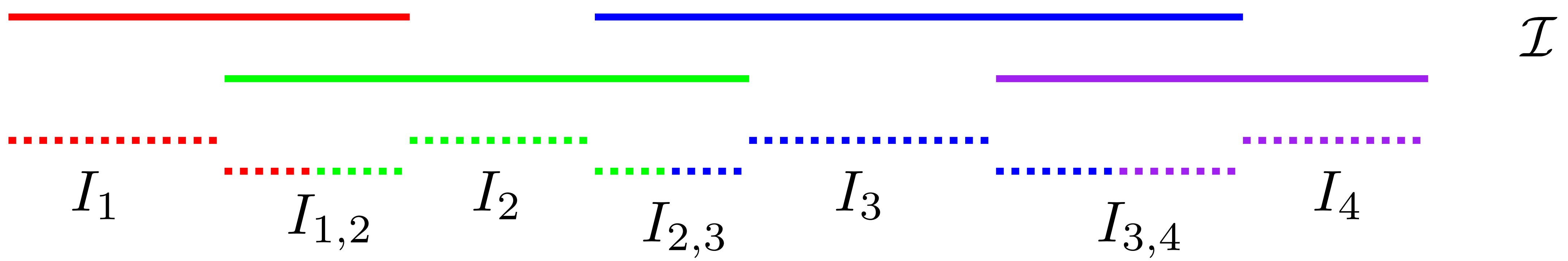
- $M_{\mathcal{I}}$ is the union of all M_{I_k} and $M_{I_{k,k+1}}$ for $I \in \mathcal{I}$



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 $I \in \mathcal{I}$

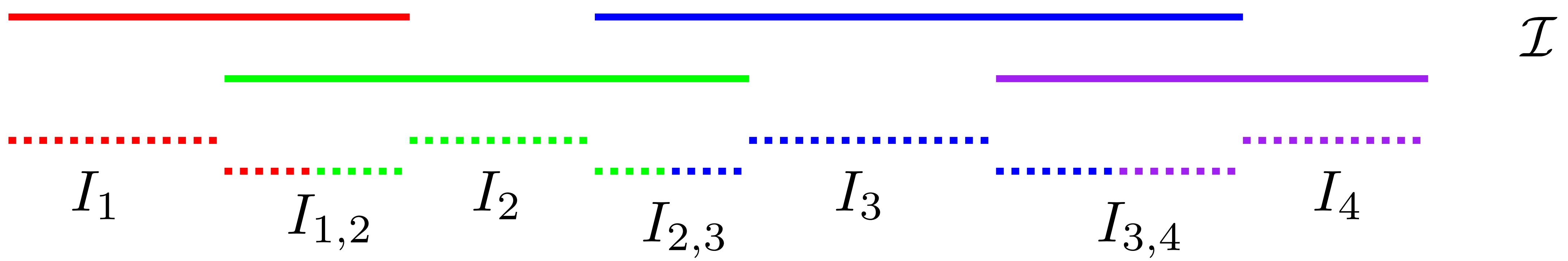


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small

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 $I \in \mathcal{I}$

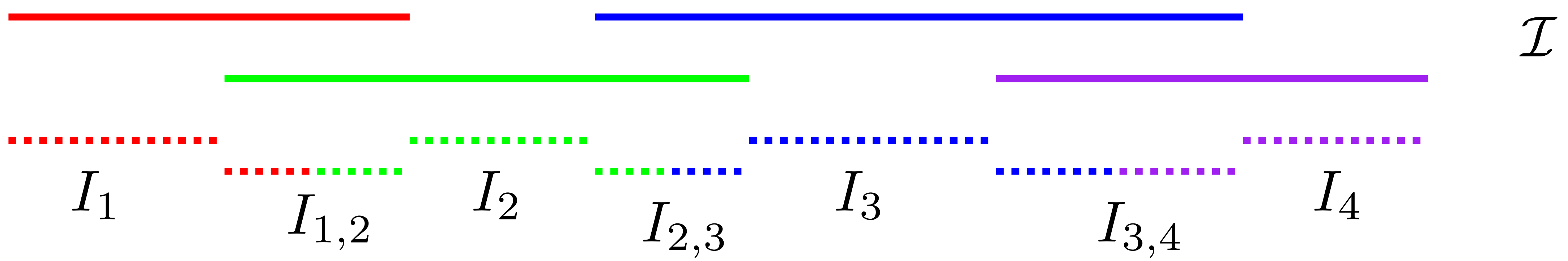


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small
- $Sh_{\mathcal{I}}$ is the union of all $Sh_{\epsilon_1,\bar{a}+\epsilon}$ and $Sh_{\epsilon_2,\bar{a}-\epsilon}$ with ϵ_1, ϵ_2 small

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 $I \in \mathcal{I}$

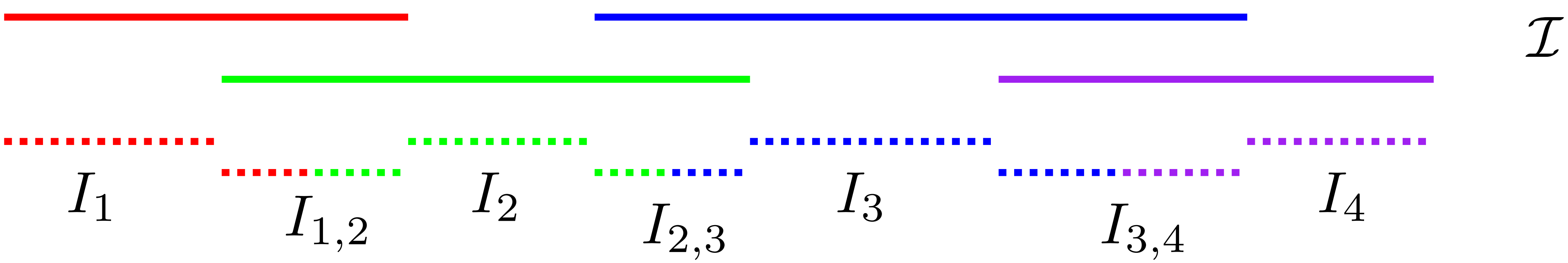


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small
- $Sh_{\mathcal{I}}$ is the union of all $Sh_{\epsilon_1,\bar{a}+\epsilon}$ and $Sh_{\epsilon_2,\bar{a}-\epsilon}$ with ϵ_1, ϵ_2 small
- $M'_{\mathcal{I}}$ is the union of all M_{I_k} for $I \in \mathcal{I}$

Formula Reeb graph $\rightarrow$ 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 $I \in \mathcal{I}$

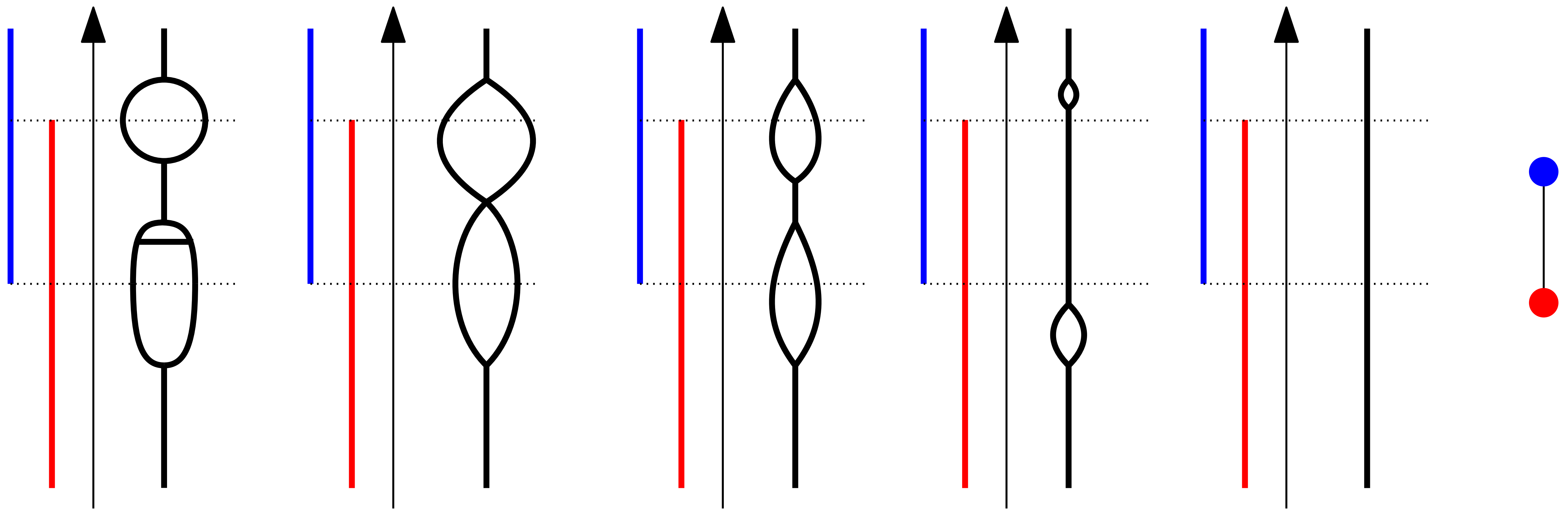


- $Sp_{\mathcal{I}}$ is the union of all $Sp_{\epsilon,\bar{a}}$ with ϵ small
- $Sh_{\mathcal{I}}$ is the union of all $Sh_{\epsilon_1,\bar{a}+\epsilon}$ and $Sh_{\epsilon_2,\bar{a}-\epsilon}$ with ϵ_1, ϵ_2 small
- $M'_{\mathcal{I}}$ is the union of all M_{I_k} for $I \in \mathcal{I}$

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

Formula Reeb graph $\to$ Mapper

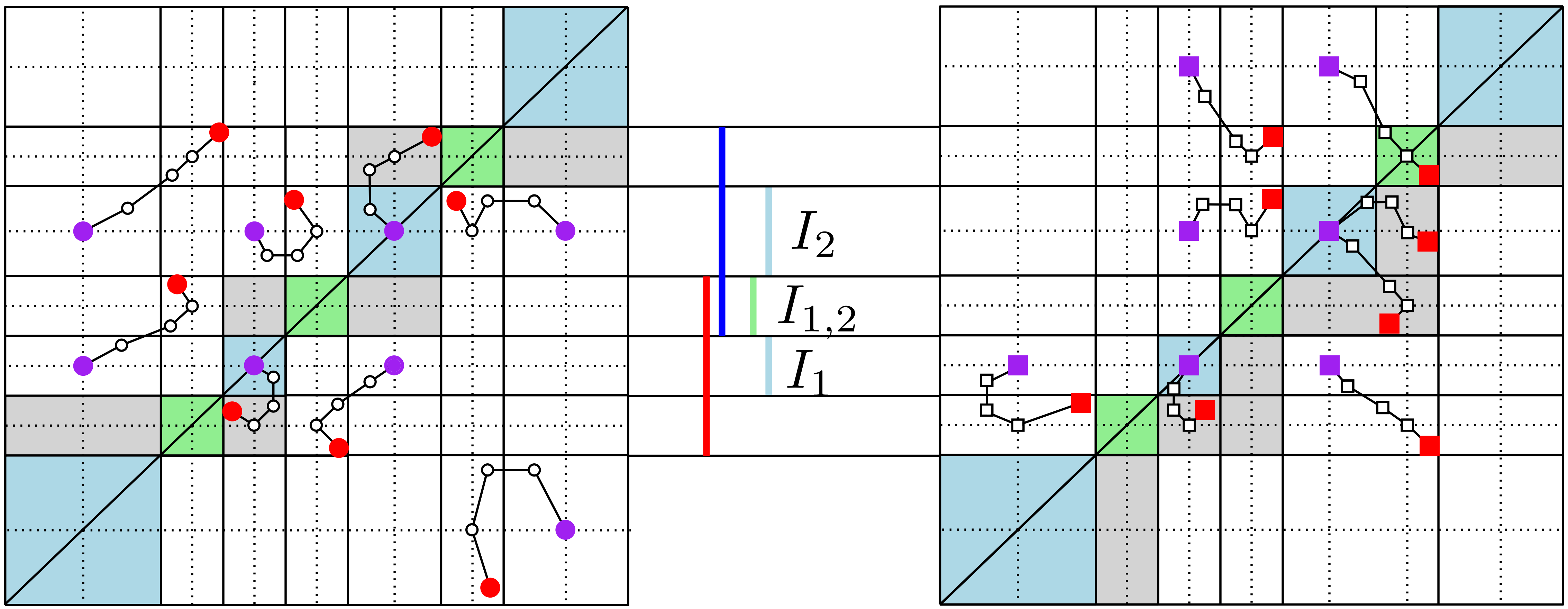
Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$



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

Formula Reeb graph → Mapper

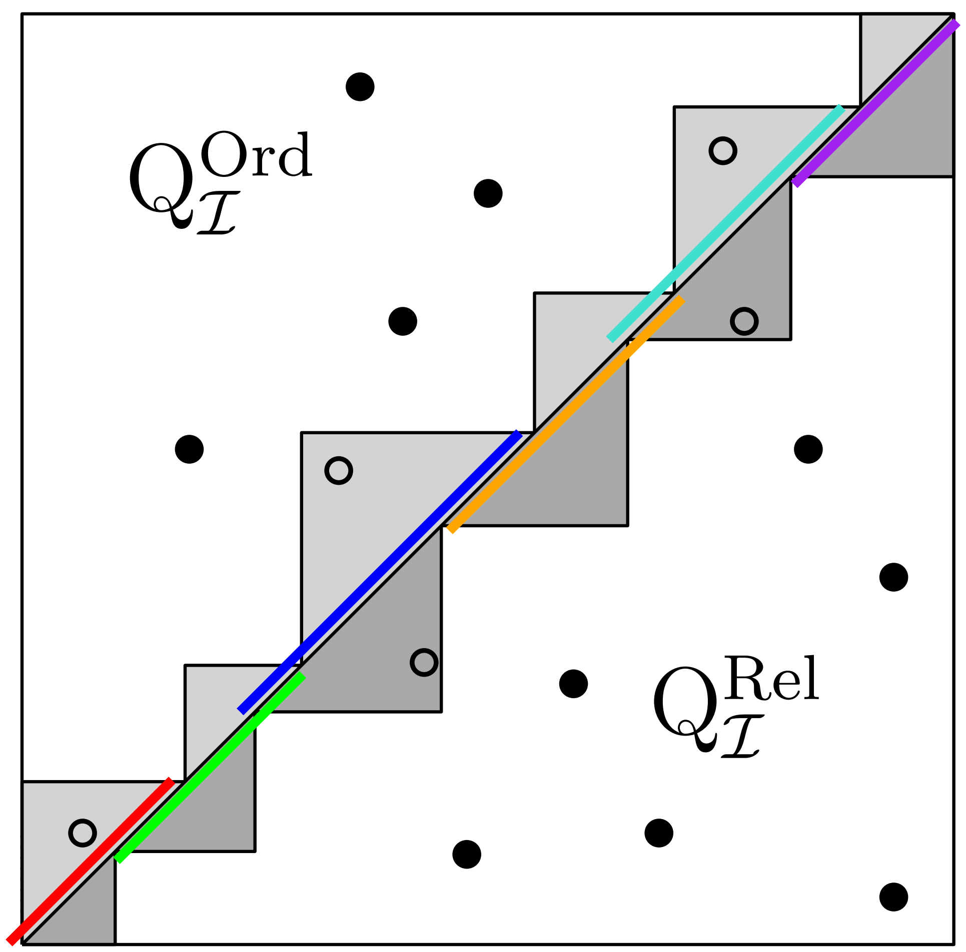
Let $\mathcal{I}$ be the cover of $\mathrm{im}(f)$

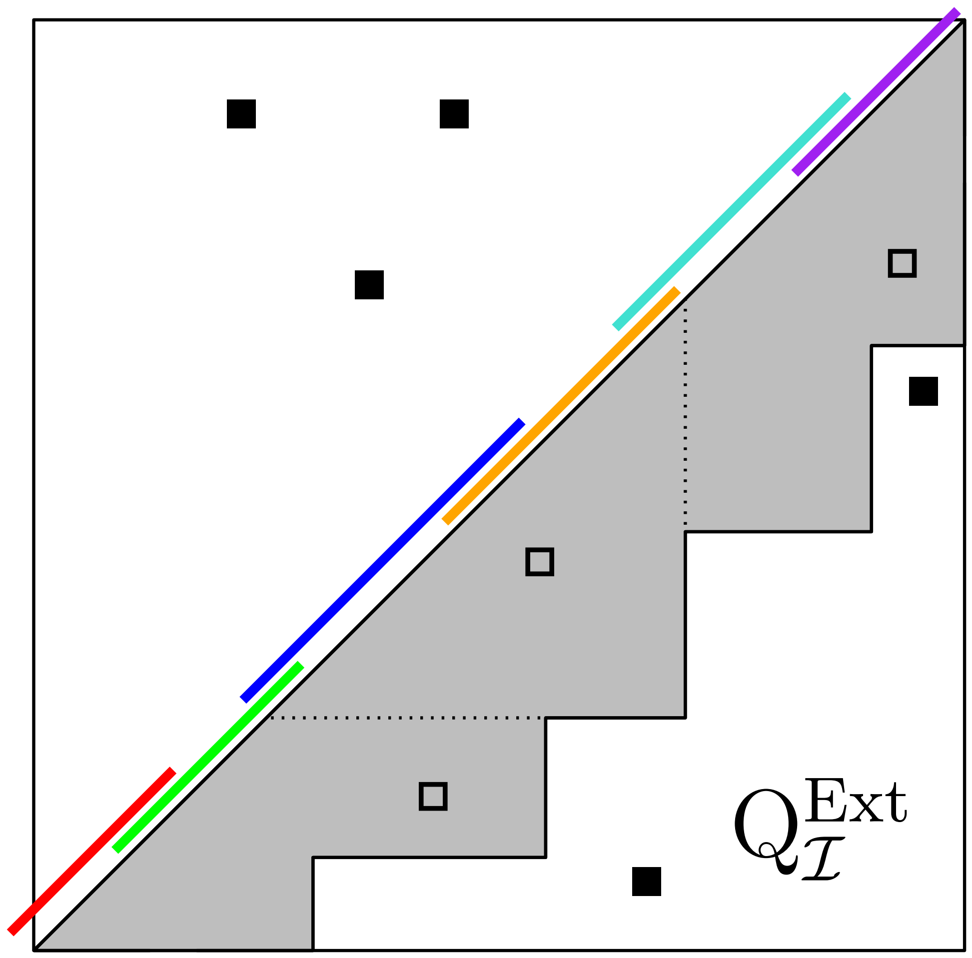


$$\mathrm{M}_f(X, \mathcal{I}) = M'_{\mathcal{I}} \circ Sh_{\mathcal{I}} \circ Sp_{\mathcal{I}} \circ M_{\mathcal{I}}(\mathrm{R}_f(X))$$

Descriptor for Mapper

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



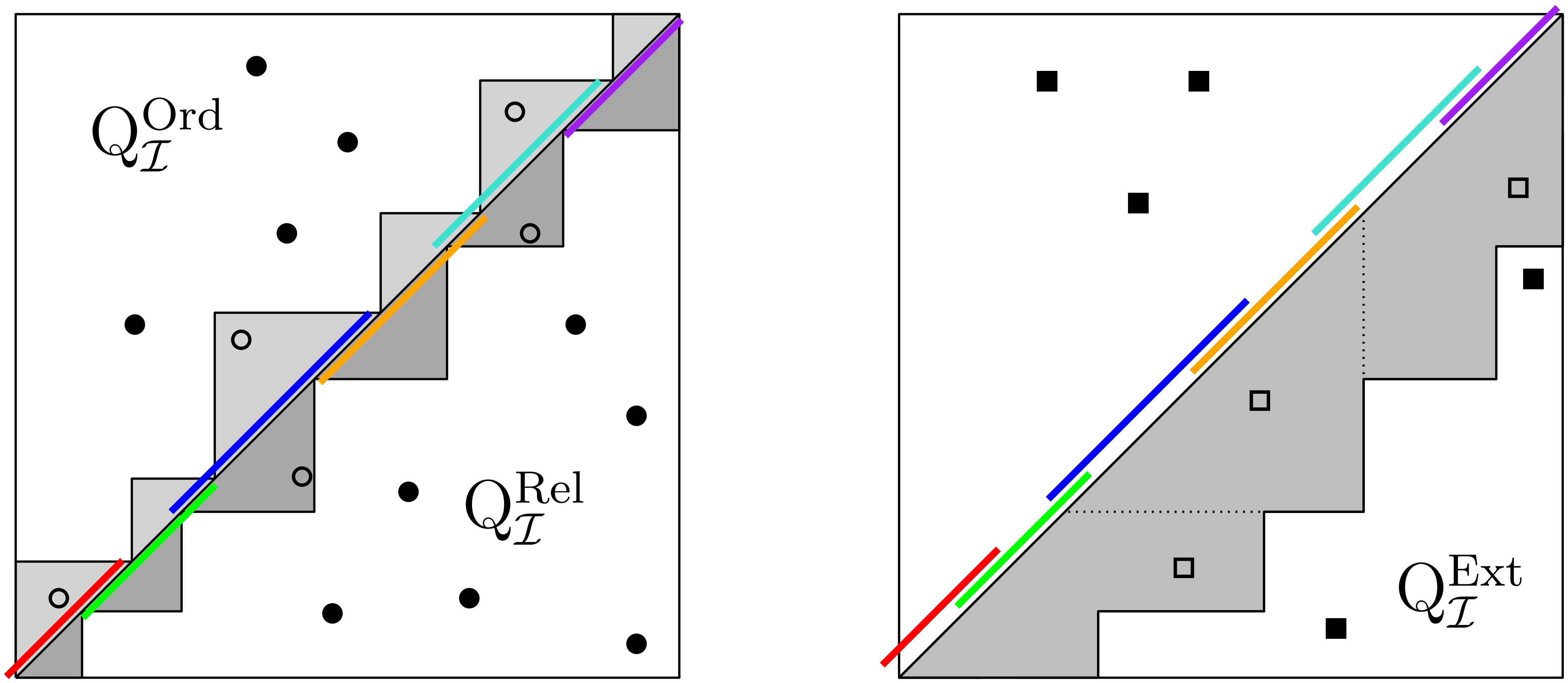


Descriptor for Mapper

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



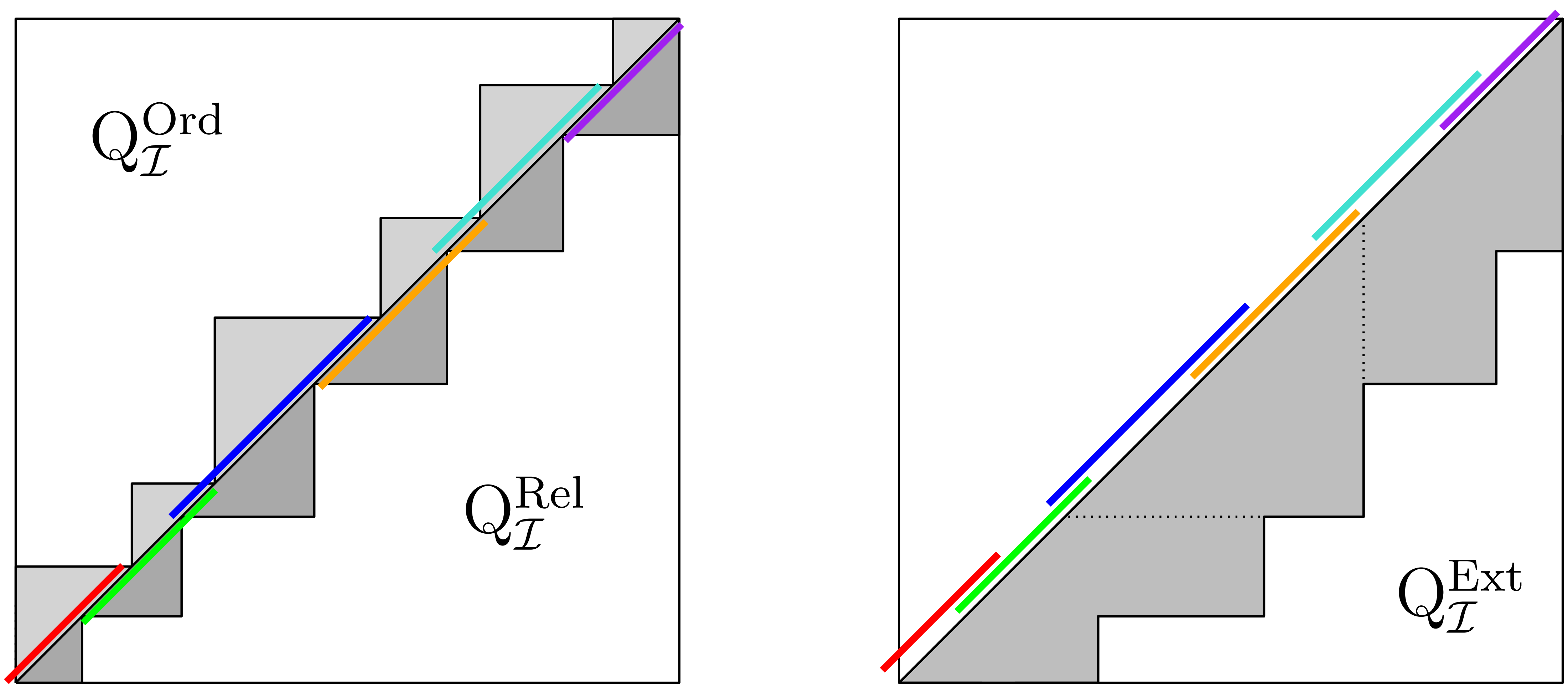
Descriptor for Mapper

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

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

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

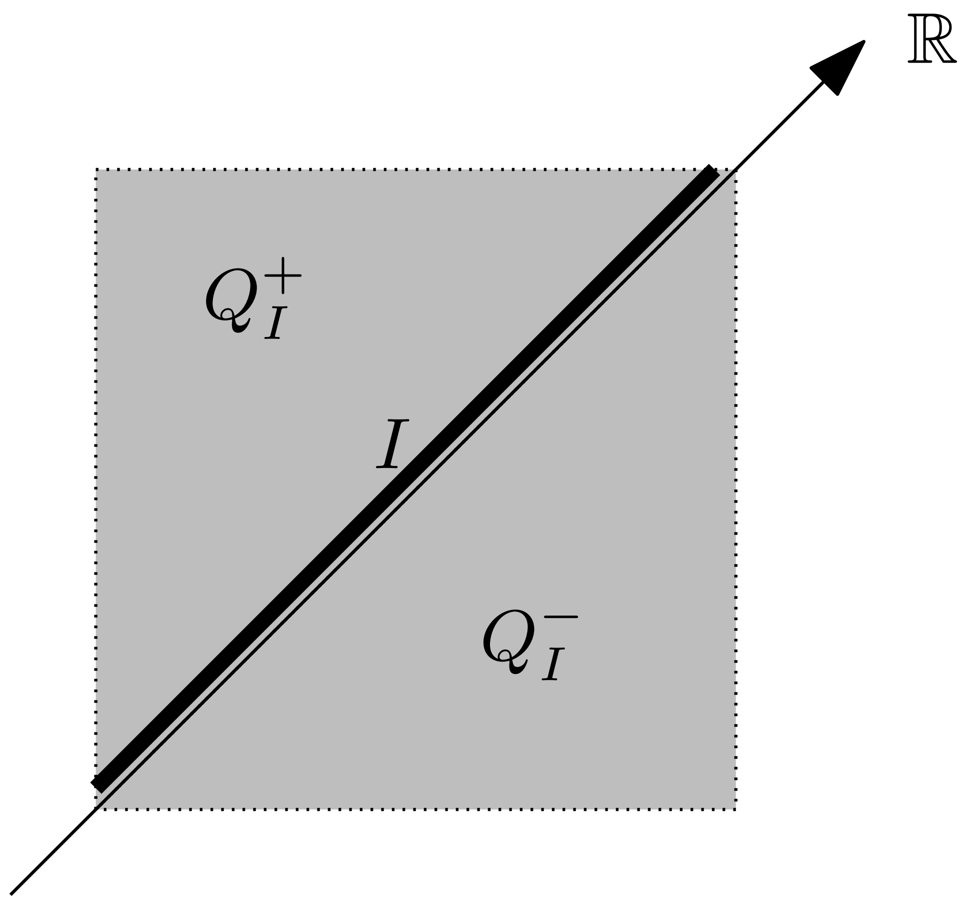
$$Q_{\mathcal{I}}^{\mathrm{Ext}} = \bigcup_{\substack{I, J \in \mathcal{I} \\ I \cap J \neq \emptyset}} Q^-_{I \cup J}$$

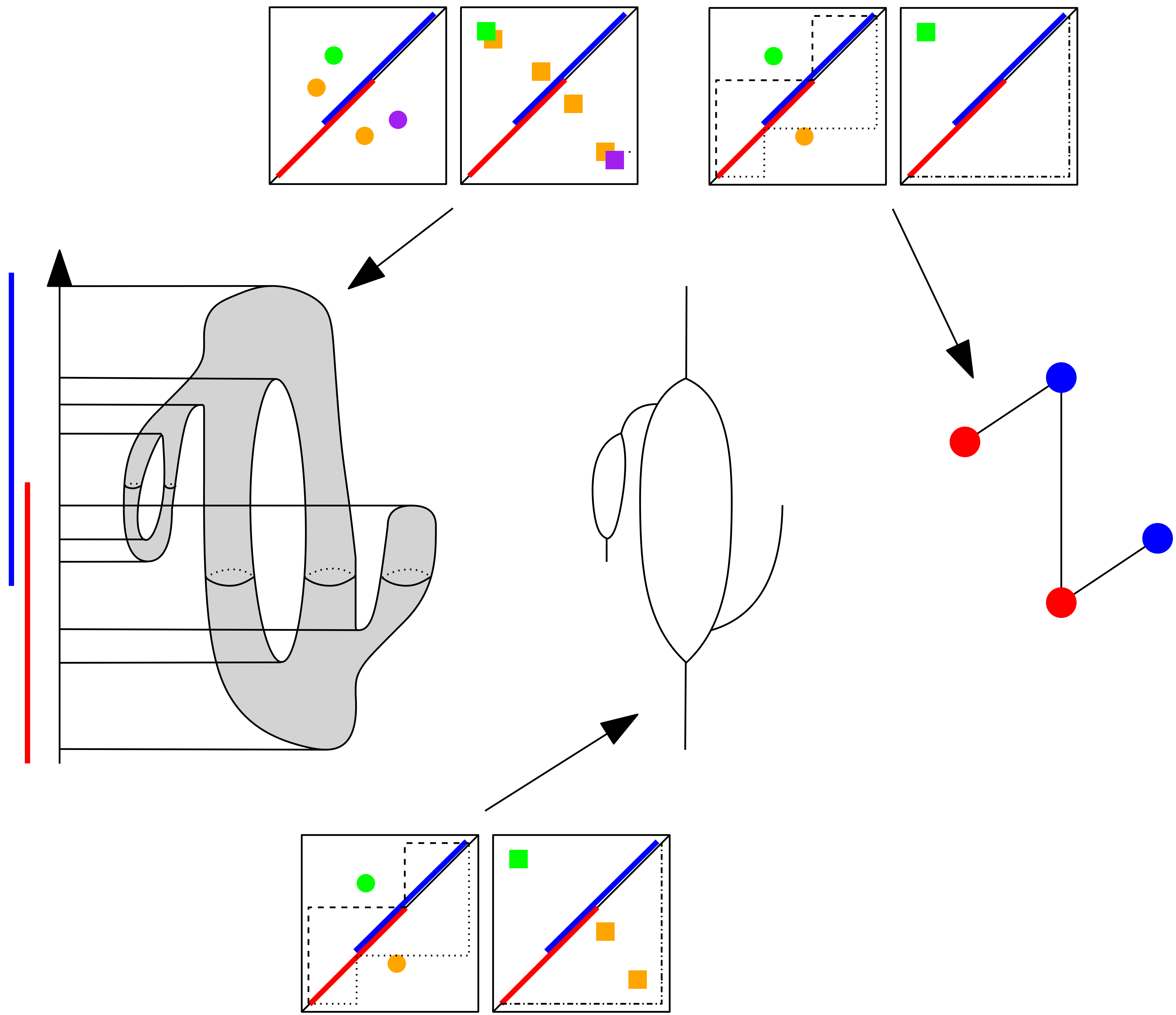


Descriptor for Mapper

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

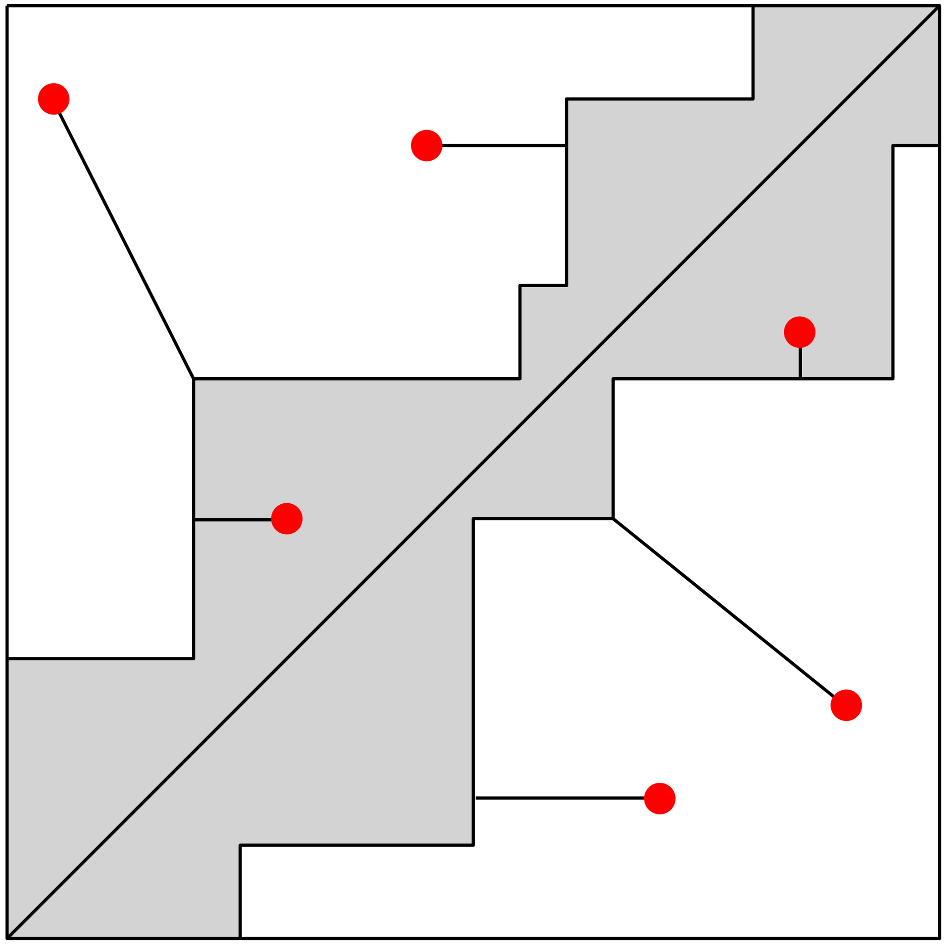
Stability of Mapper

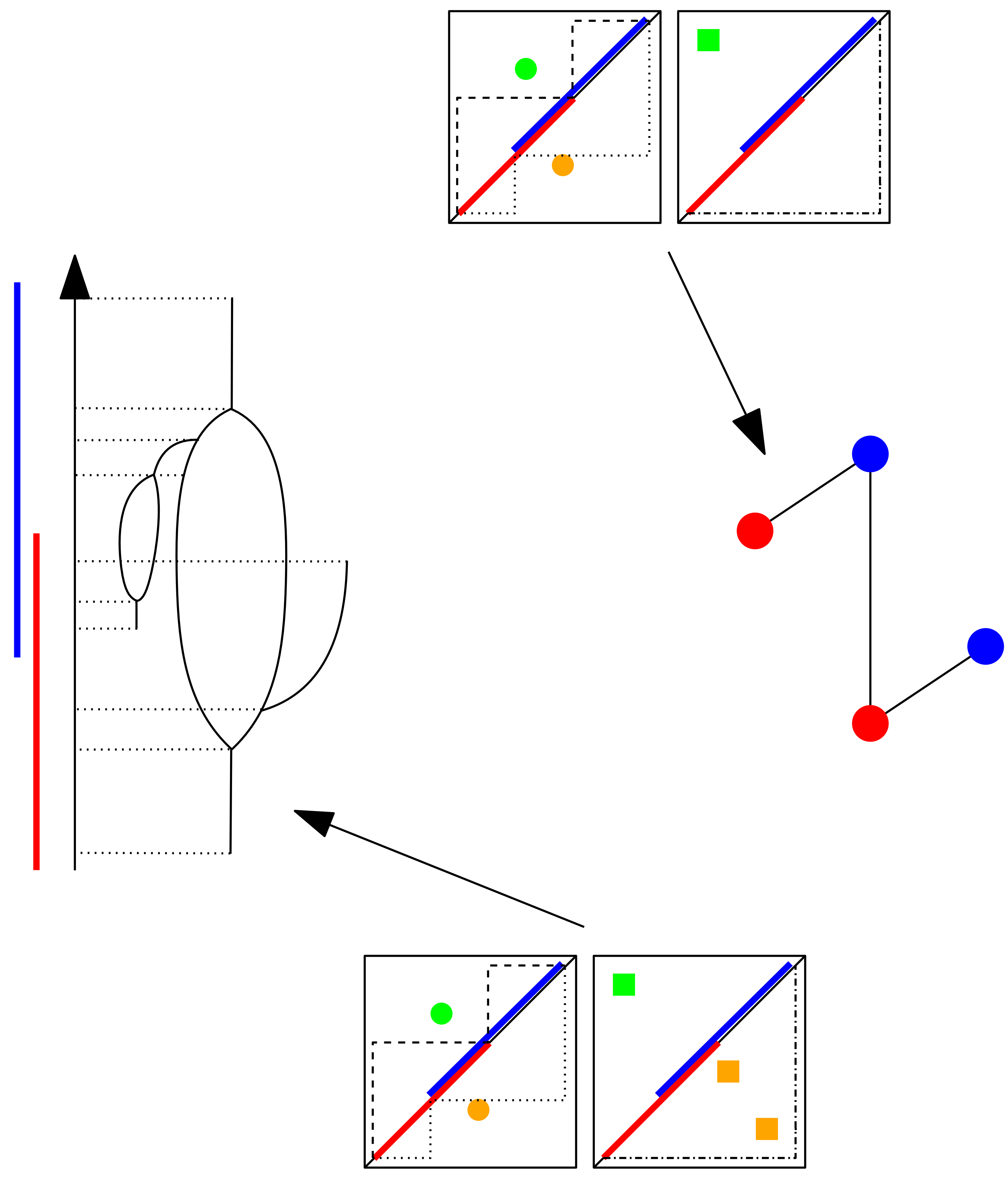
Def: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) := \mathrm{Ord}\,\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ord}} \cup \mathrm{Rel}\,\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Rel}} \cup \mathrm{Ext}\,\tilde{f} \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}$

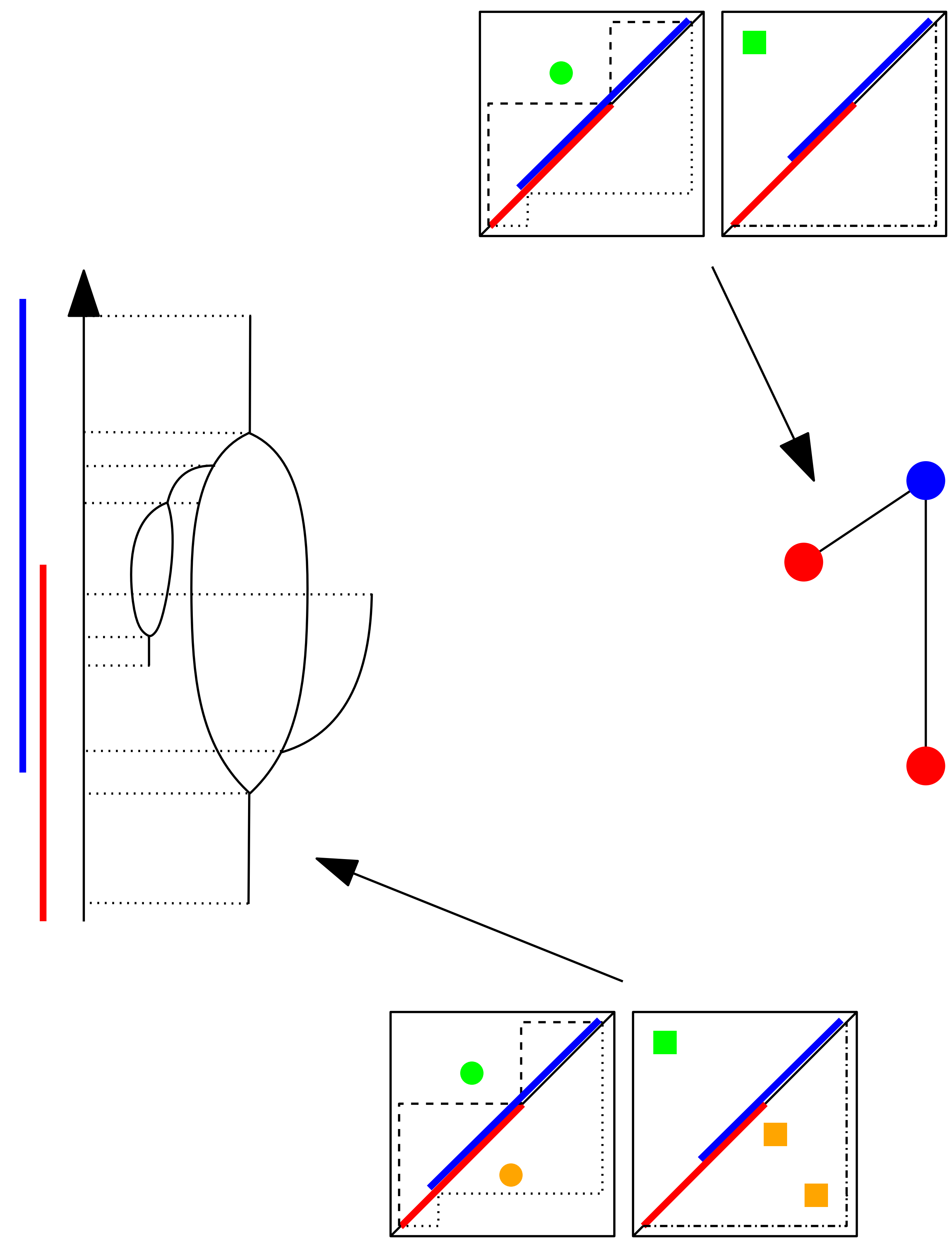
Thm: $\mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I})$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X,\mathcal{I})$:

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

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

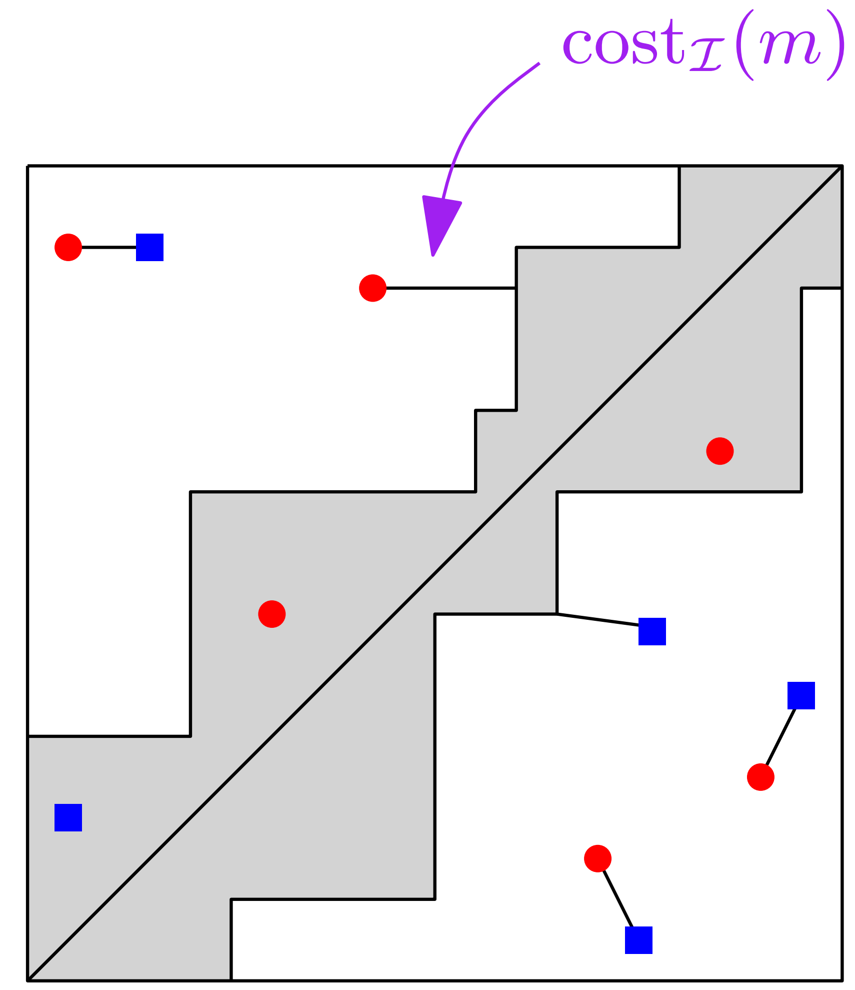






Stability of Mapper

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



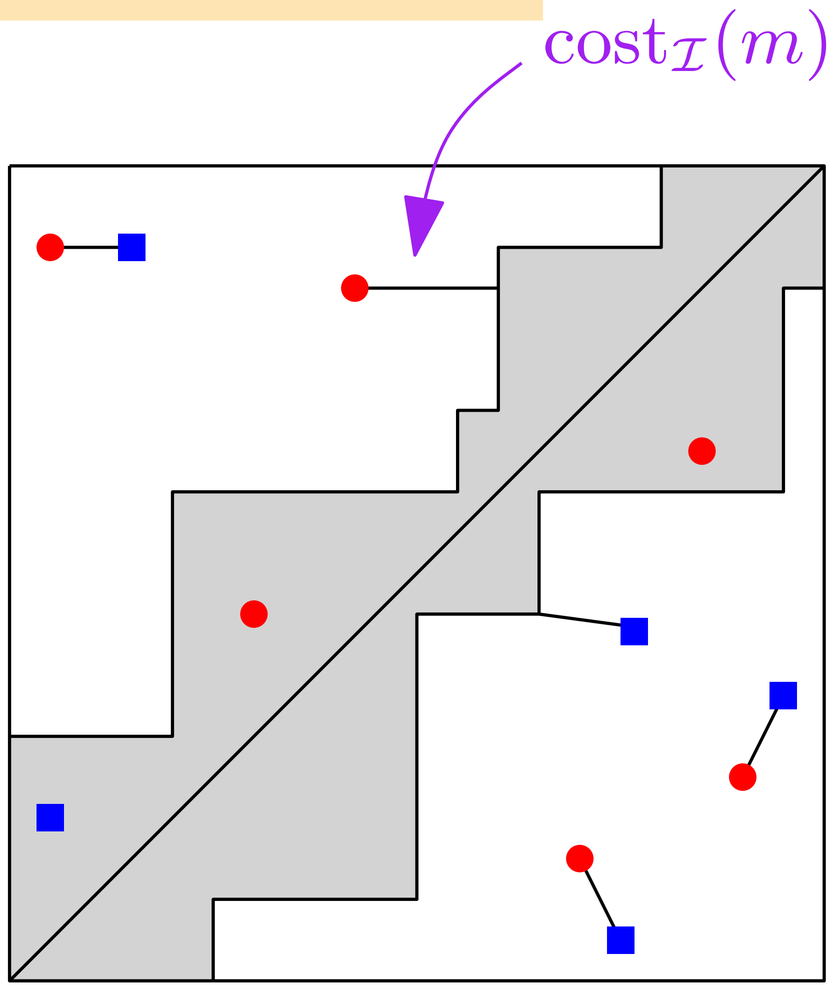
$m : \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})$

Stability of Mapper

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

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

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



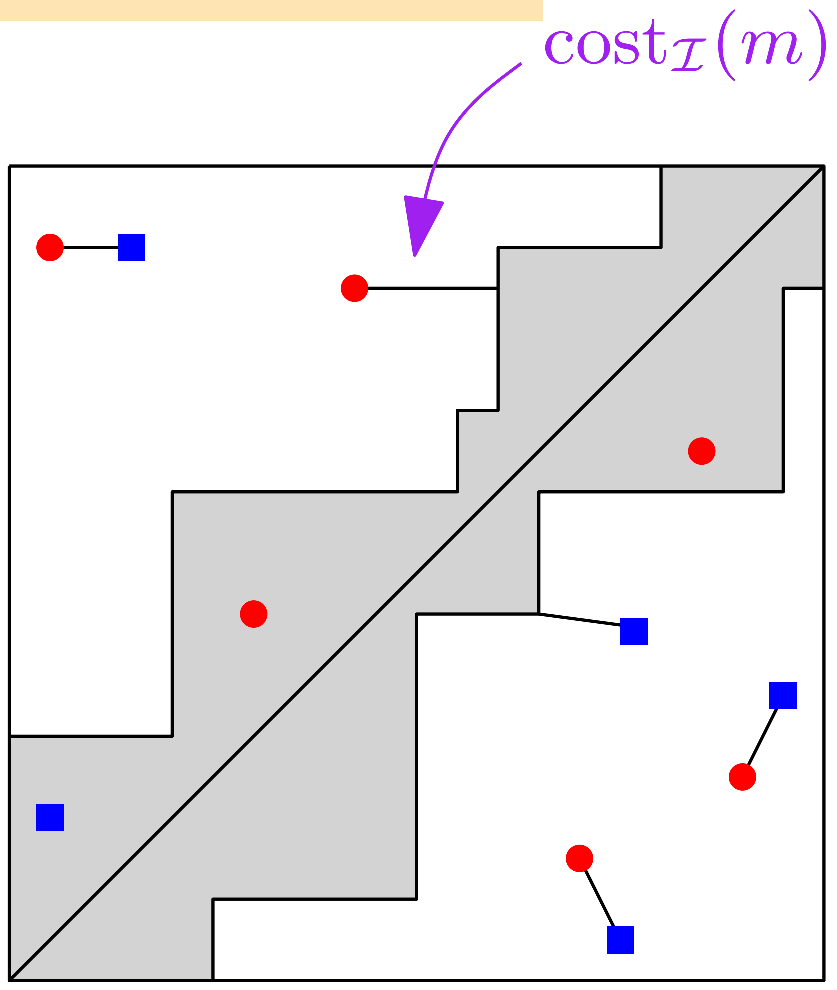
$m : \mathrm{Dg}\,\mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_{f'}(X,\mathcal{I})$

Stability of Mapper

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

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

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



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

Extensions to:

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

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 $\mathcal{V}$:

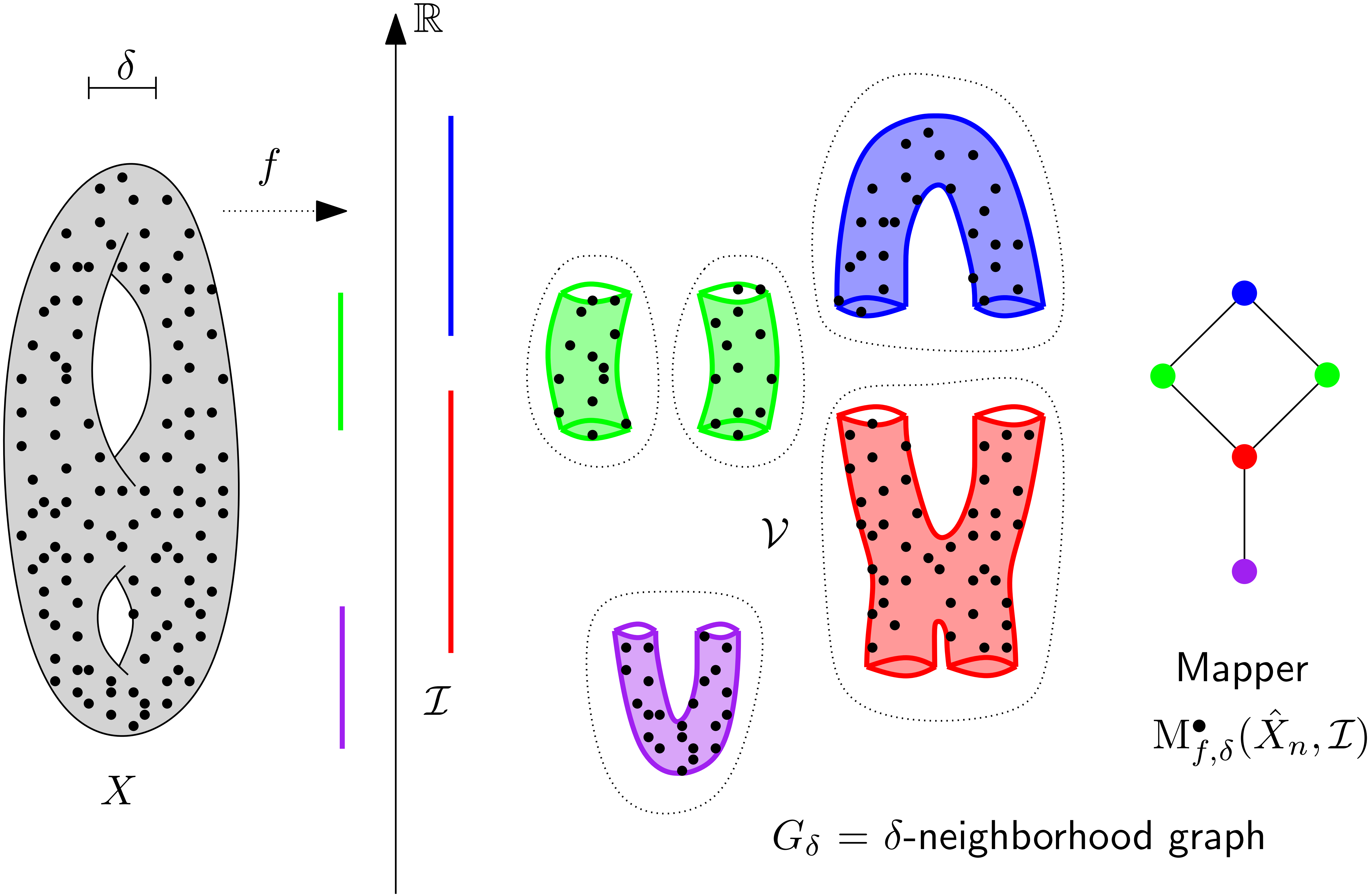
- 1 vertex per element $V \in \mathcal{V}$

- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$

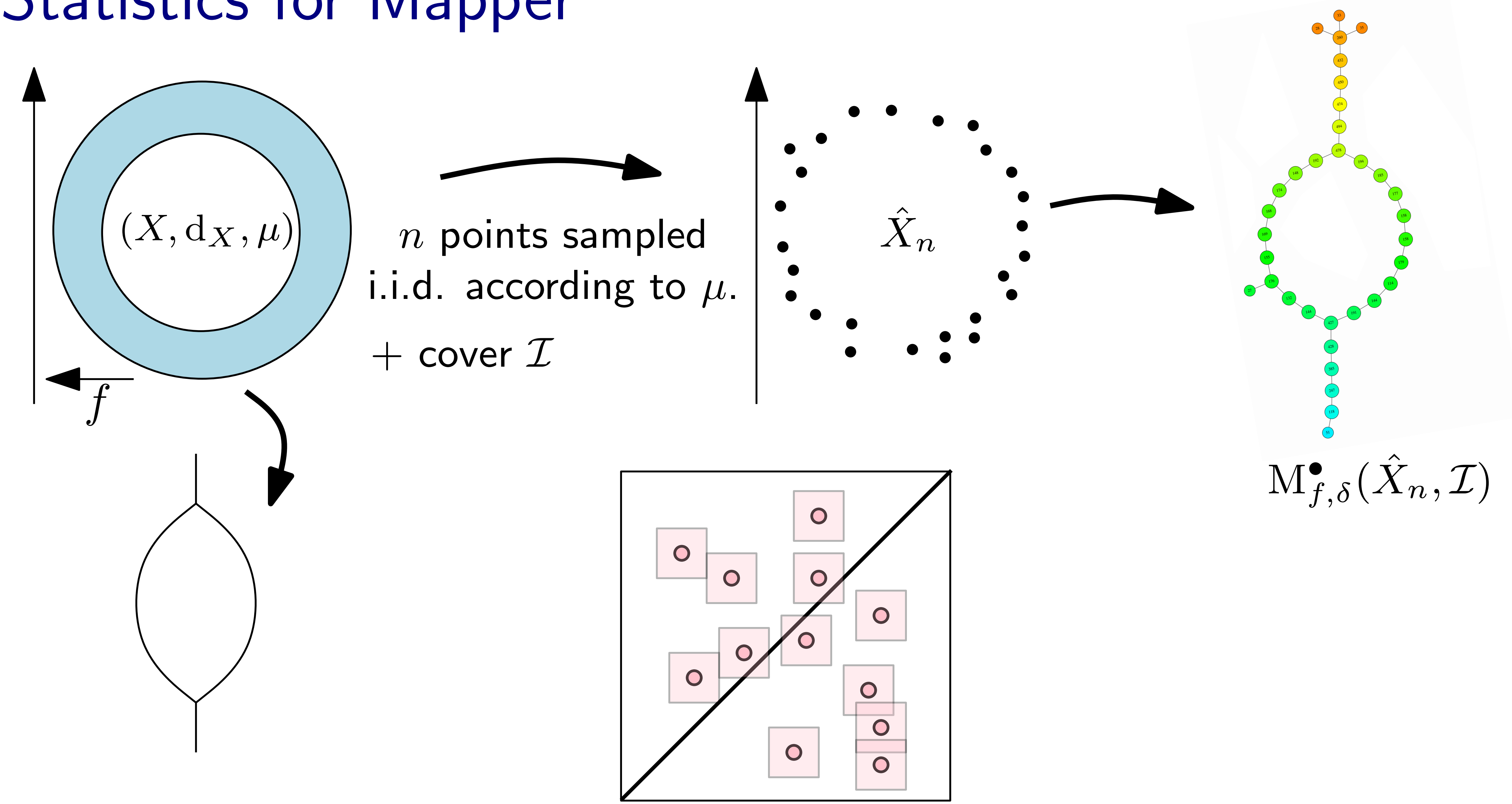
- 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

(intersections materialized by data points)

Mapper in practice



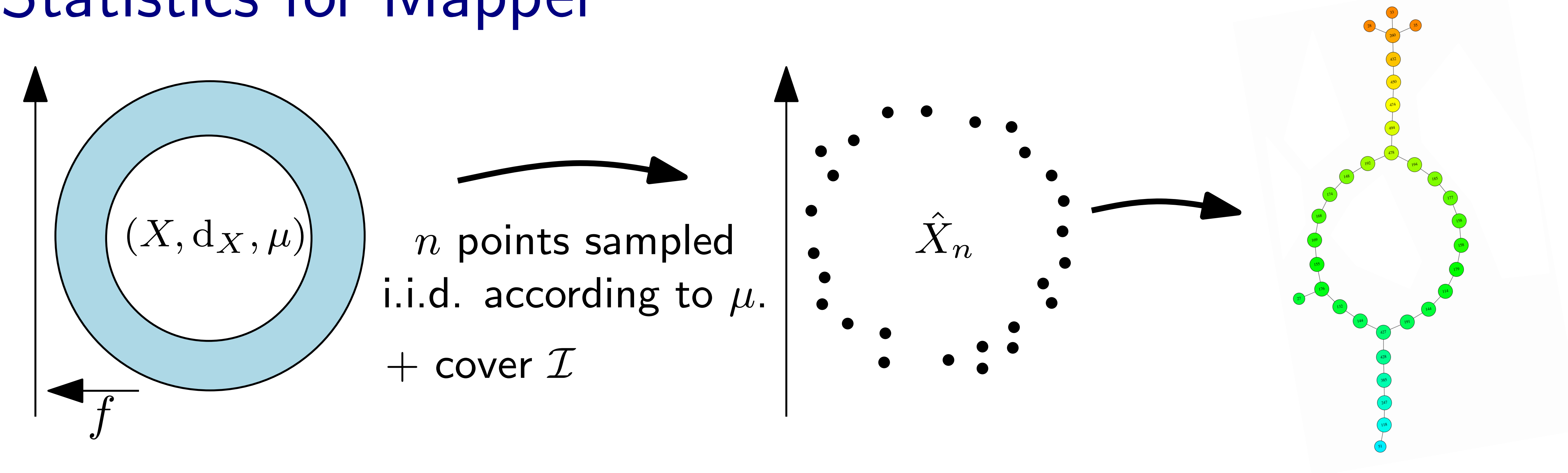
Statistics for Mapper



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 $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{V})$?

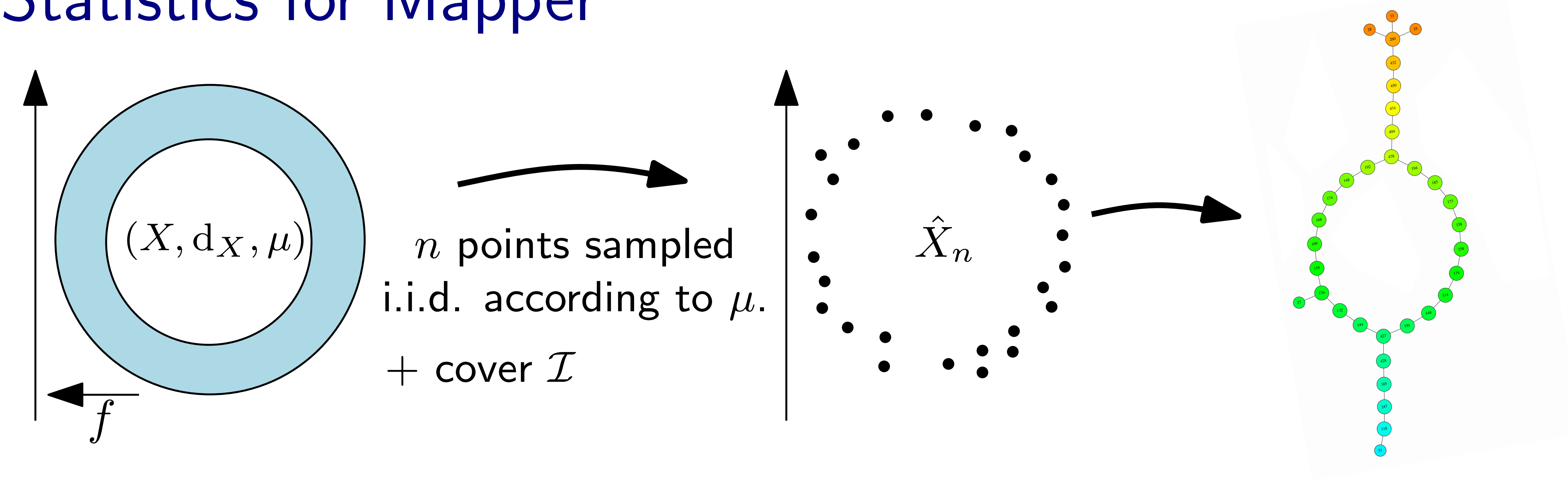
a. support $\rightarrow$ δ-neighborhood graph
$X \rightarrow G_\delta(\hat{X}_n)$

b. Reeb graph $\rightarrow$ Mapper

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

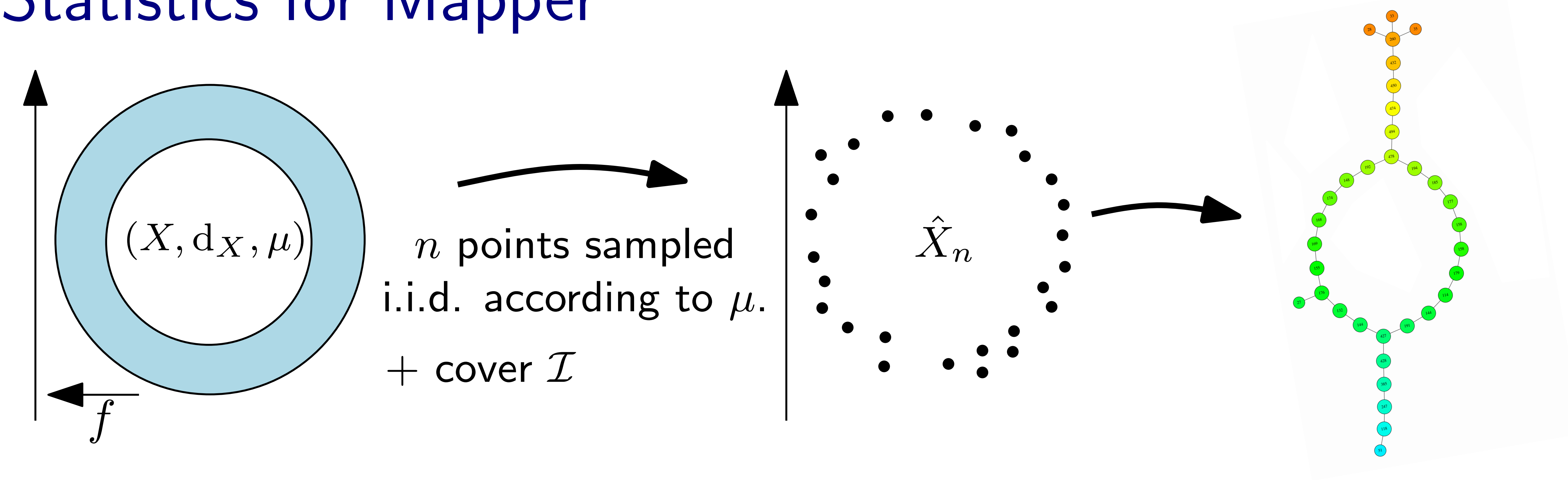
intersections given by metric graph $\rightarrow$ intersections given by points

Statistics for Mapper



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

Statistics for Mapper



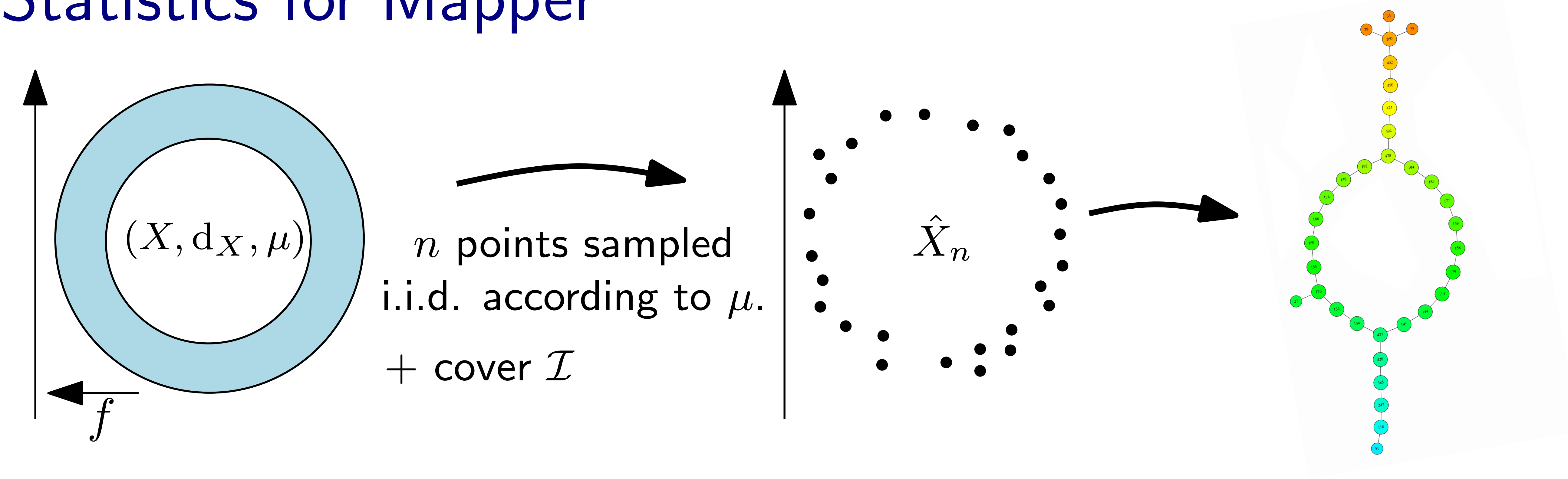
1. Link between $\mathrm{R}_f(X)$ and $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

support $\to$ δ-neighborhood graph

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

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

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

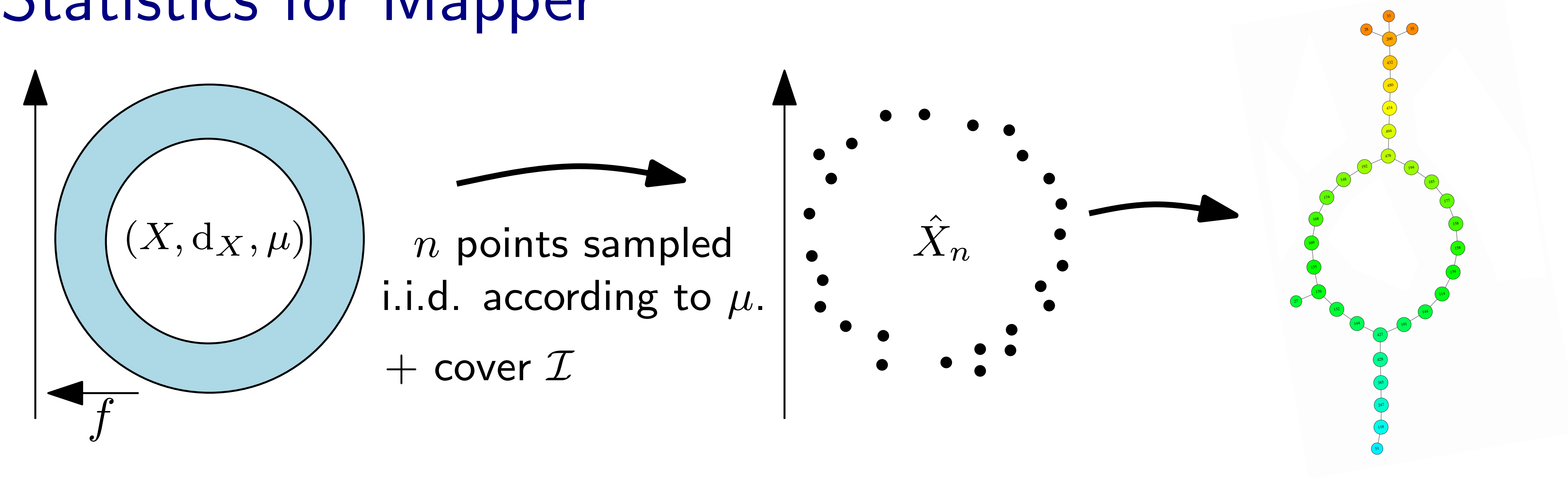
Thm: If $4d_H(X, \hat{X}_n) \leq \delta \leq \min\left\{\frac{1}{4}\mathrm{rch}(X), \frac{1}{4}\rho(X)\right\}$

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

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

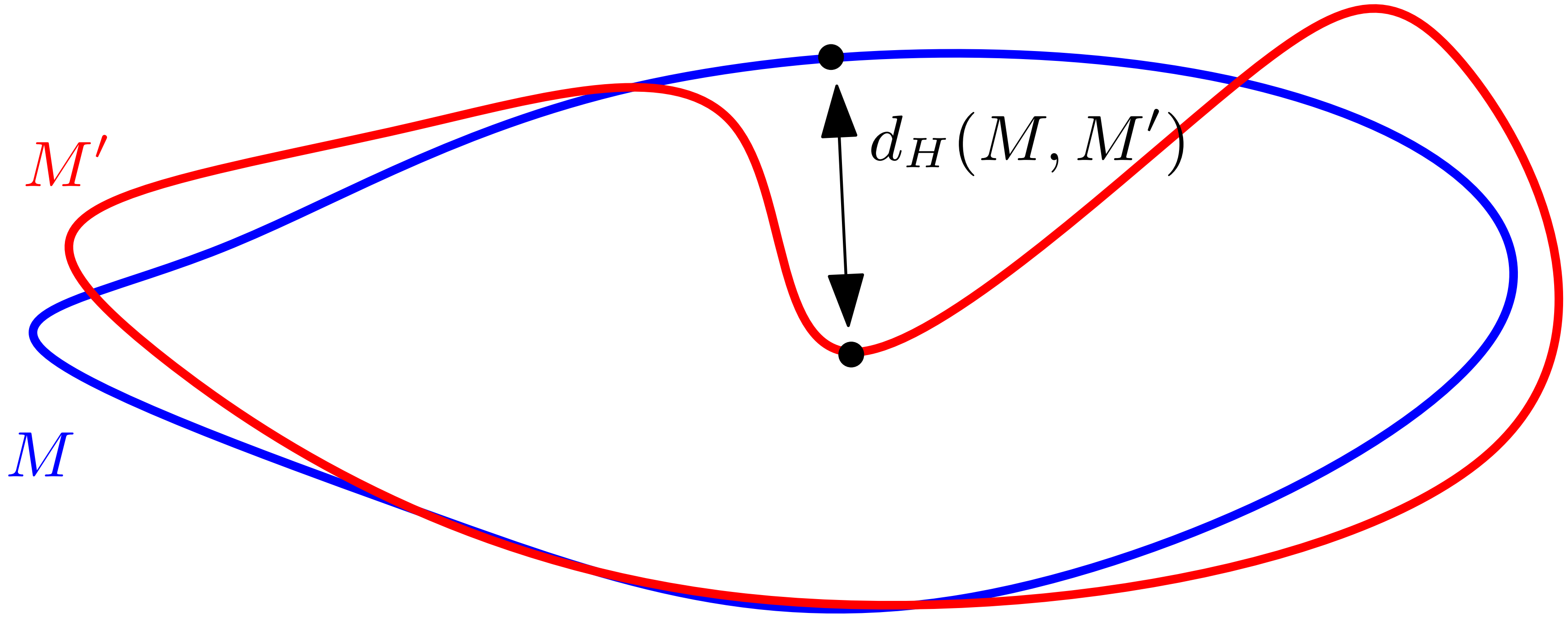
$$\omega : \delta \mapsto \sup\{|f(x) - f(y)| \, : \, d(x, y) \leq \delta\}$$

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



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

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

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

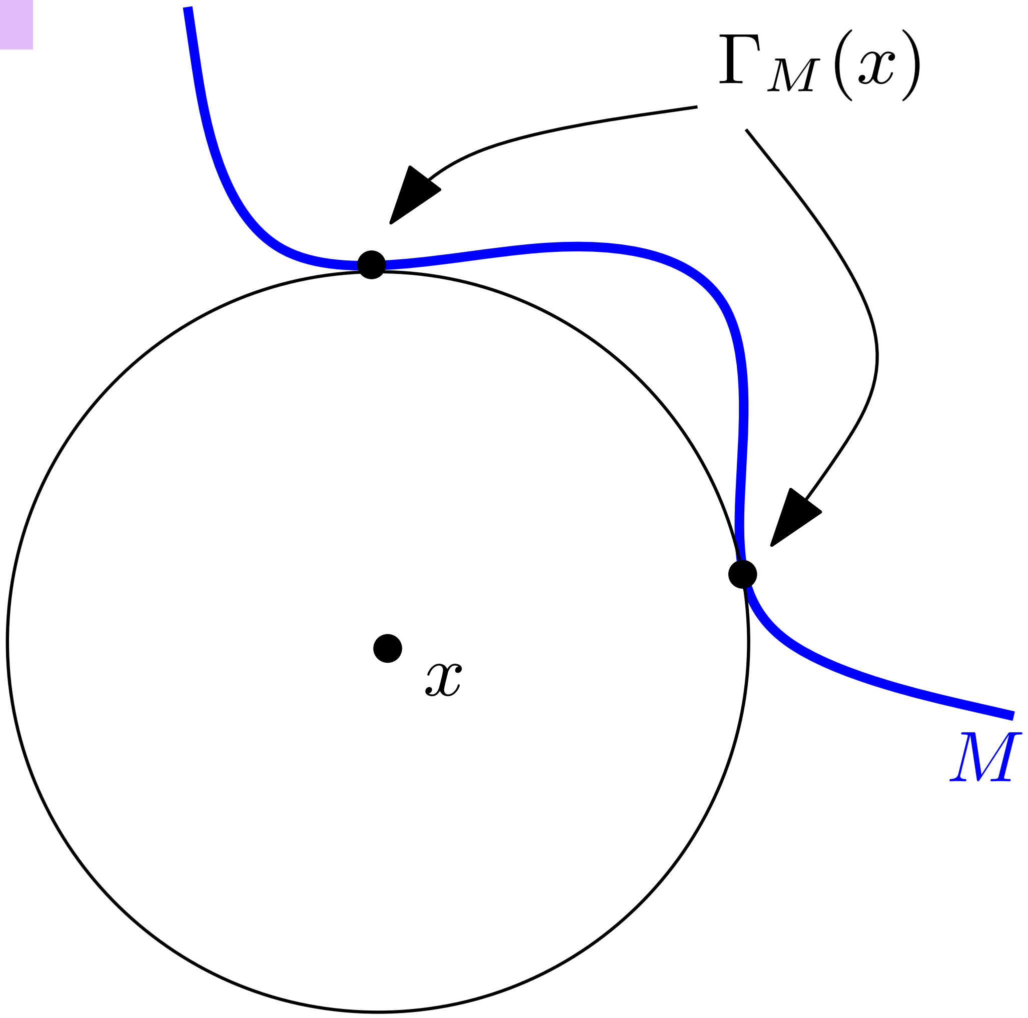
$$d_H(M, M') = \sup_{x \in \mathbb{R}^d} |d_M(x) - d_{M'}(x)|$$

Statistics for Mapper

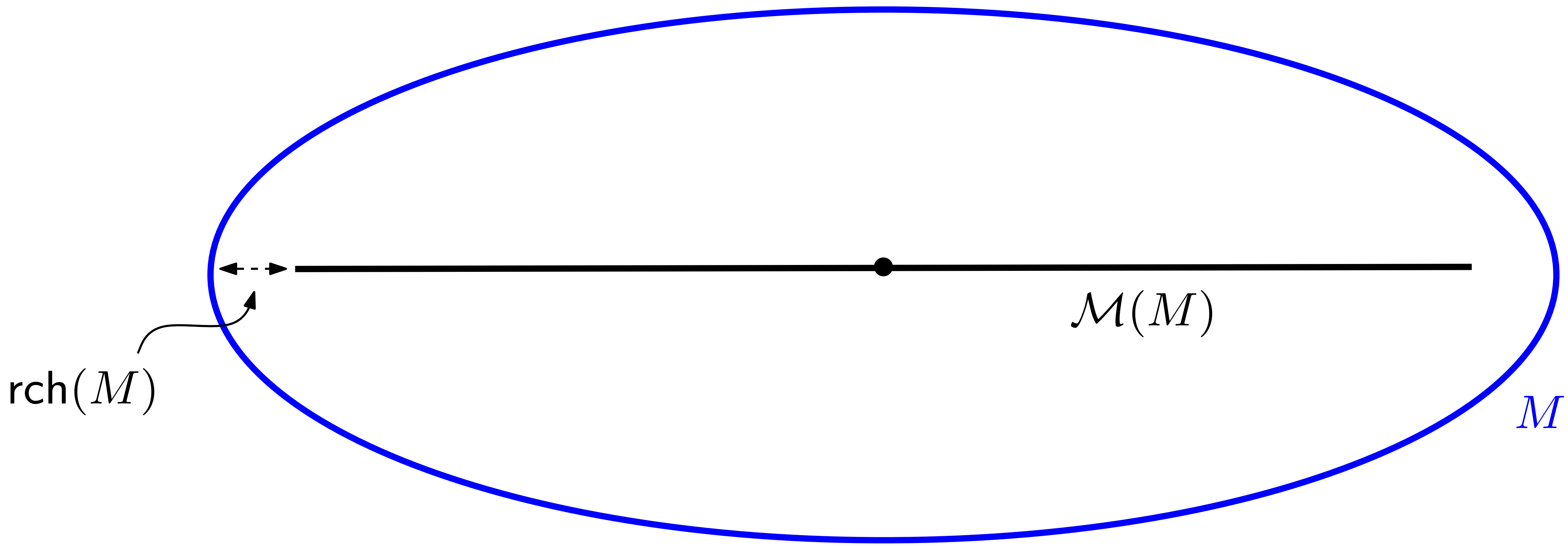
$$\Gamma_M(x) = \{y \in M : d_M(x) = \|x - y\|\}$$

Def: The medial axis of M:

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



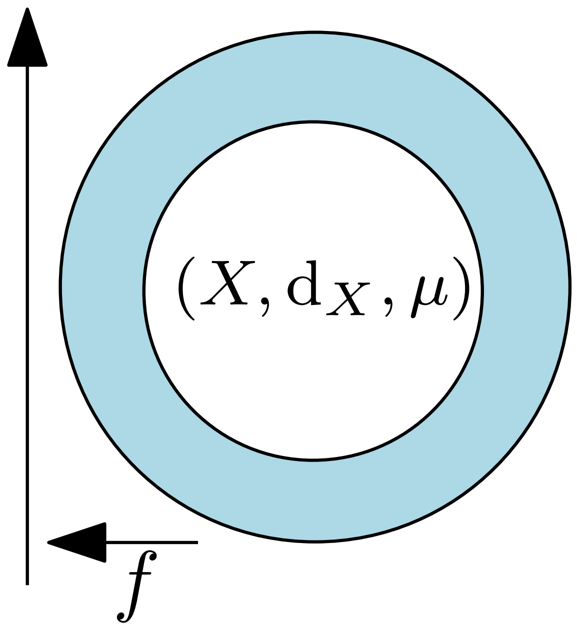
Statistics for Mapper



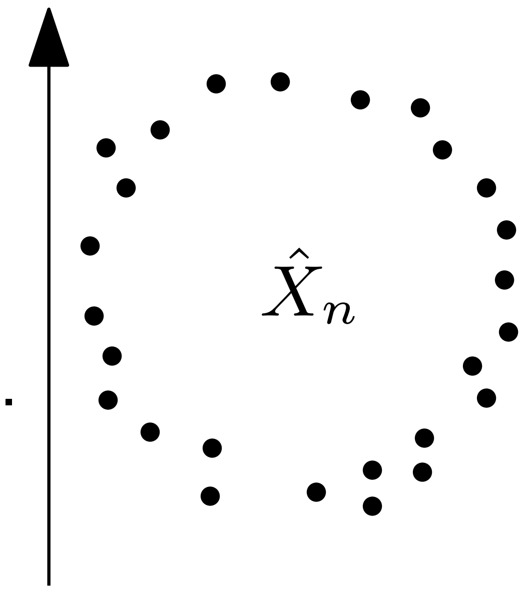
Def: The reach of M, rch(M) is the smallest distance from $\mathcal{M}(M)$ to M:

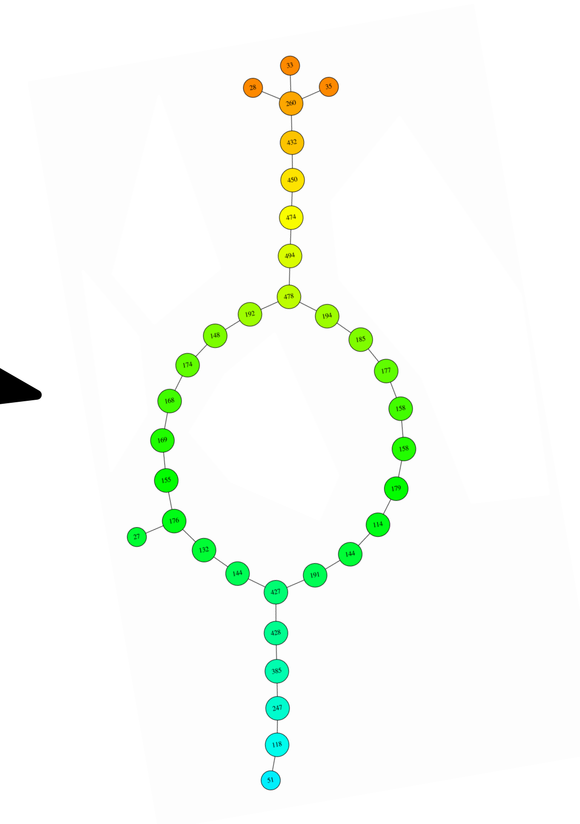
$$\mathrm{rch}(M) = \inf_{y \in \mathcal{M}(M)} d_M(y)$$

Statistics for Mapper



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





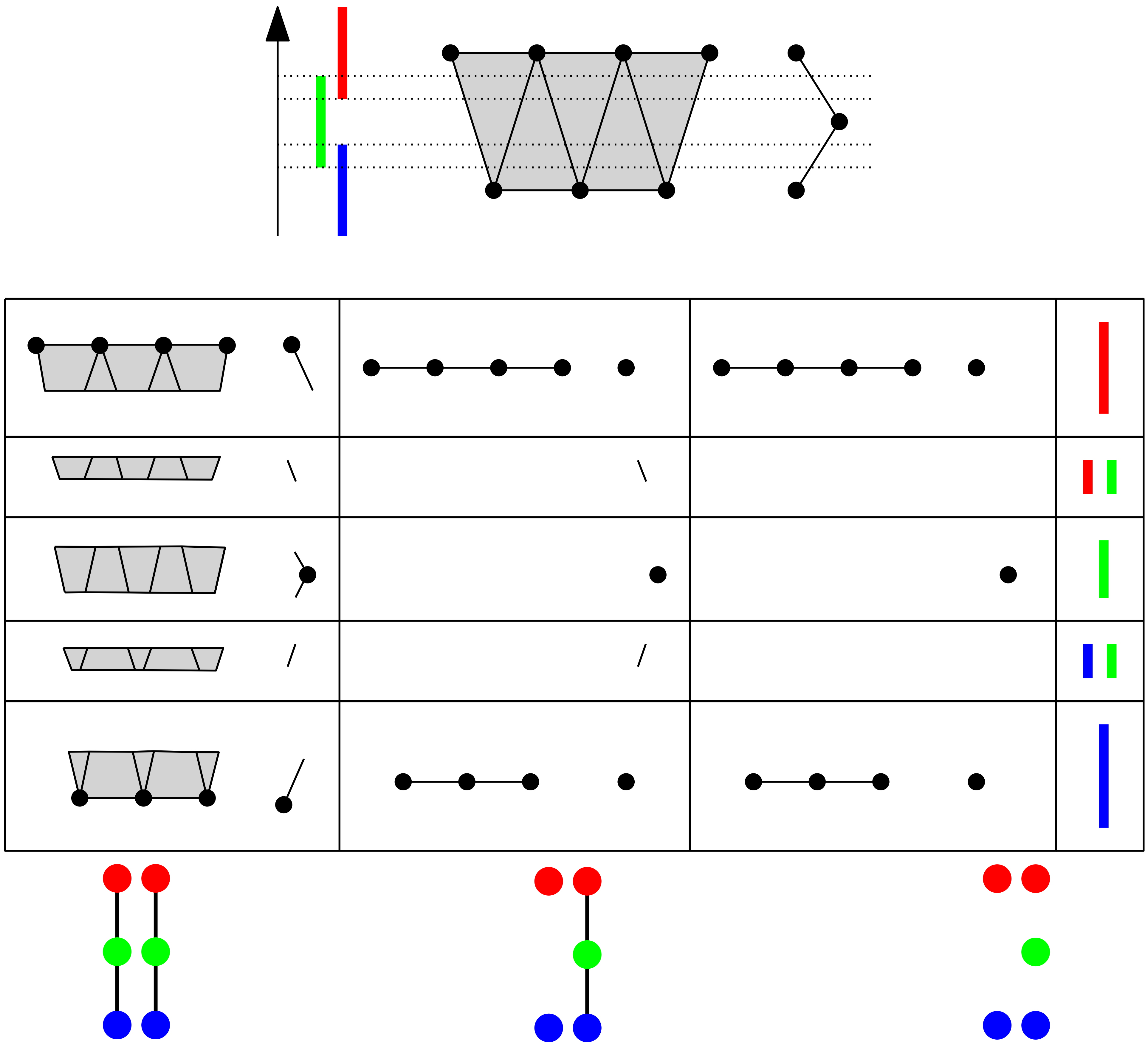
2. Link between $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{I})$ and $\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$?

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

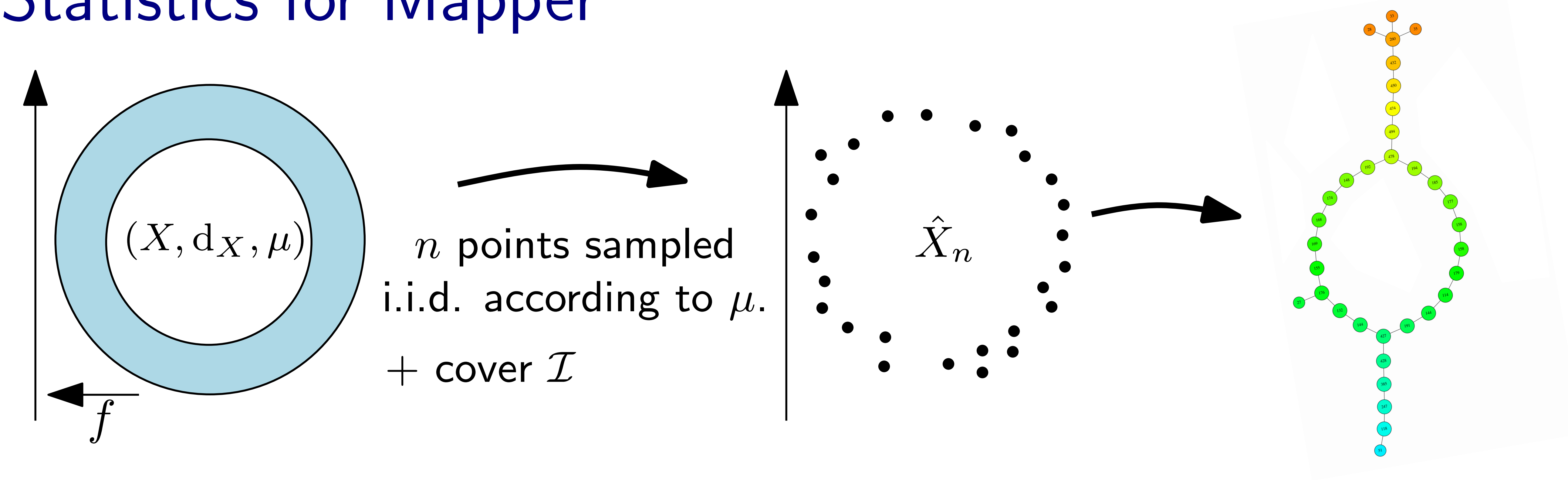
Thm: If there are no **intersection-crossing edges**, then

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

Statistics for Mapper



Statistics for Mapper



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

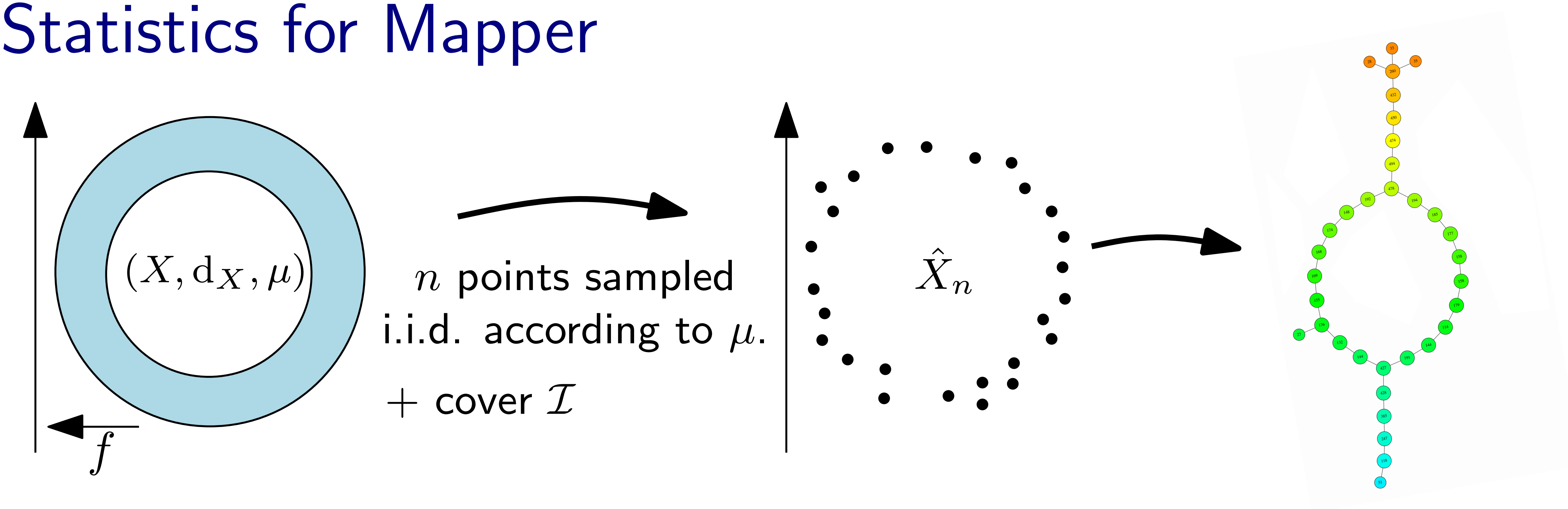
Hyp: μ is (a, b)-standard

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

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



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

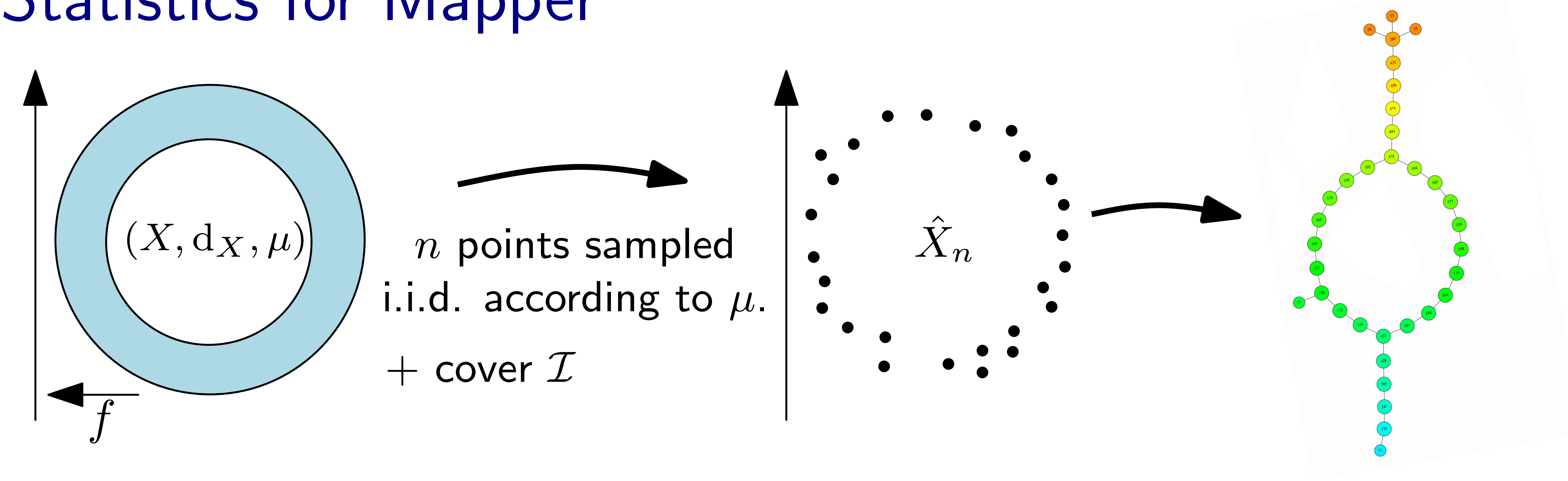
$$\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b},\ g_n \in \left(\tfrac{1}{3},\ \tfrac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n}, \qquad \text{one has } \forall \varepsilon > 0$$

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

where C depends only on a, b, c.

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

Statistics for Mapper



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.

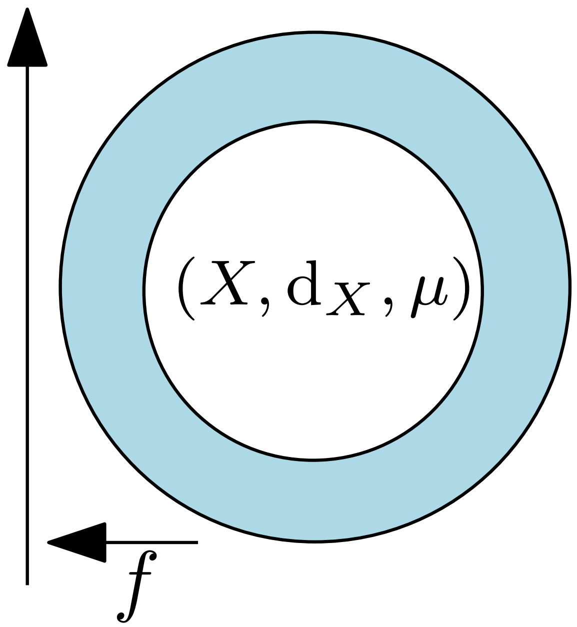
Thm: For any estimator $\widehat{\mathrm{R}}$, one has:

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

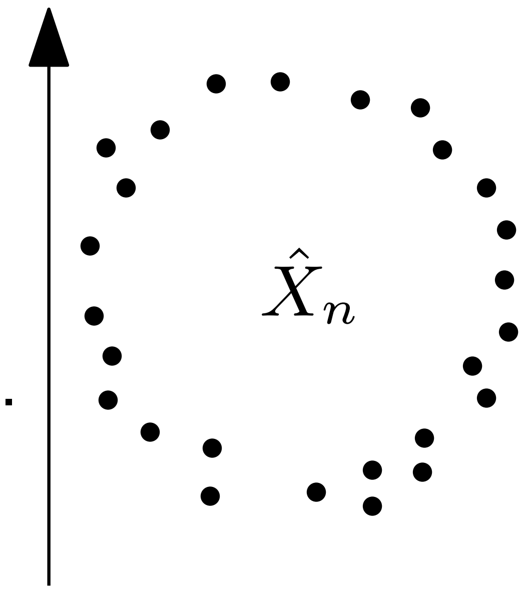
where C depends only on a, b.

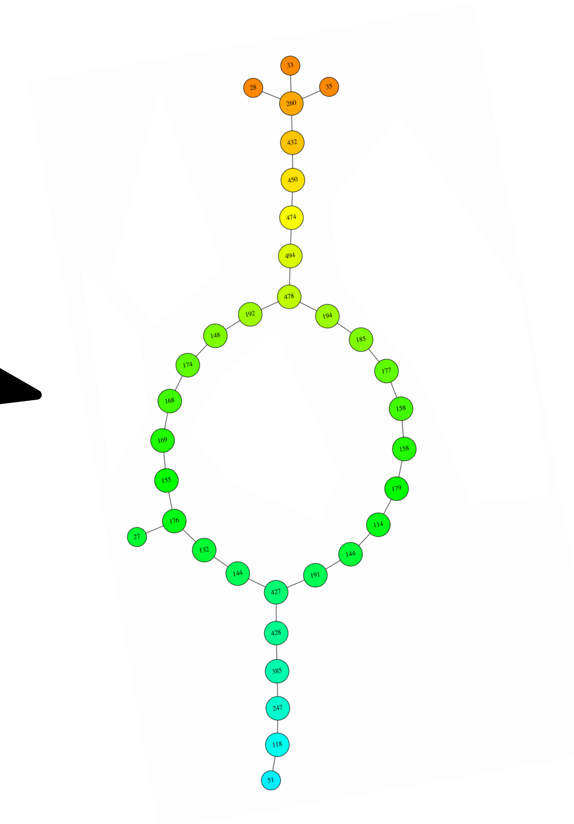
Consequence of Le Cam's lemma

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:

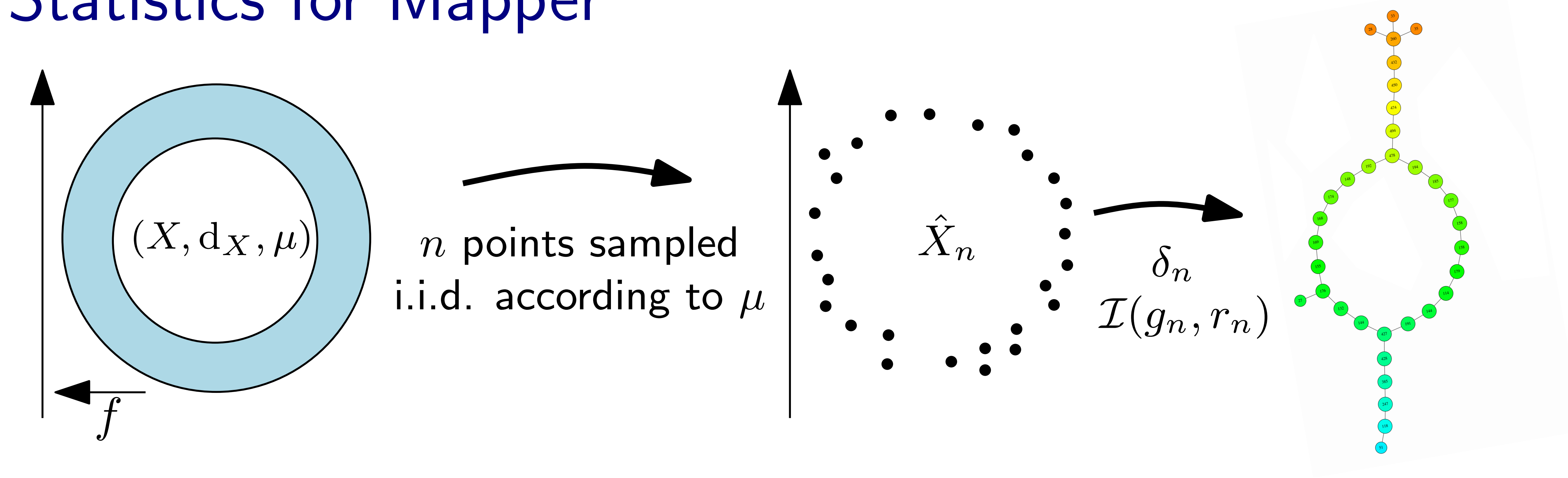
$$\delta_n = 4\left(\frac{2\log n}{\boxed{a}n}\right)^{1/\boxed{b}},\ g_n \in \left(\tfrac{1}{3},\ \tfrac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n}, \qquad \text{one has } \forall \varepsilon > 0$$

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

where C depends only on a, b, c.

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

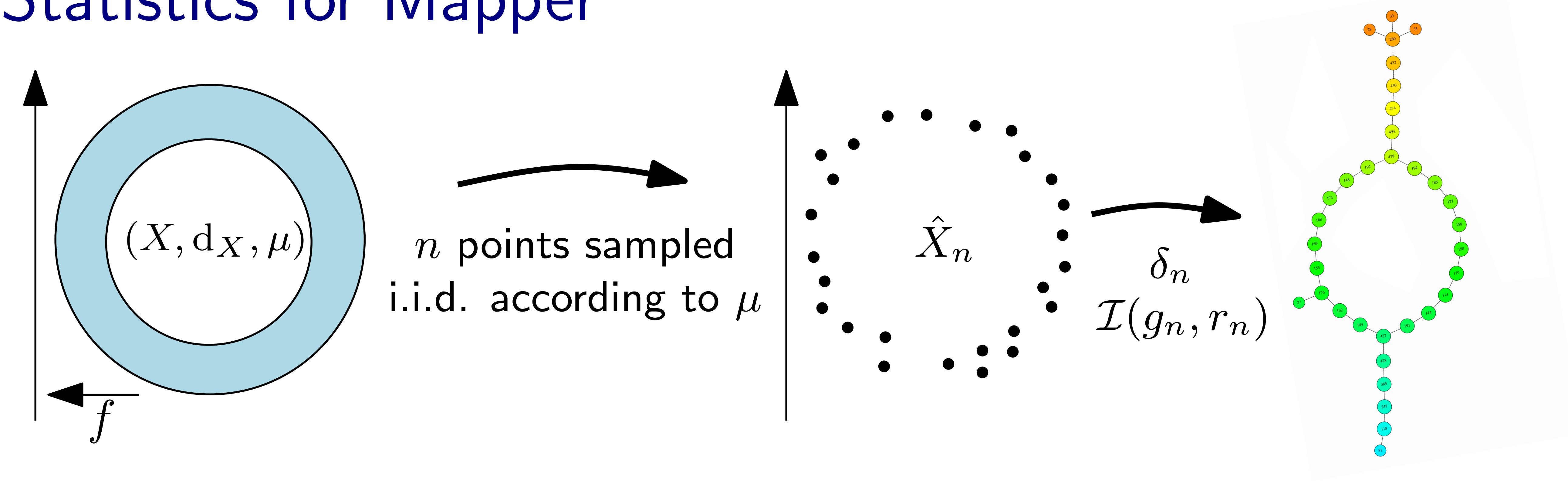
Statistics for Mapper



$\rightarrow$ 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 $s(n)$

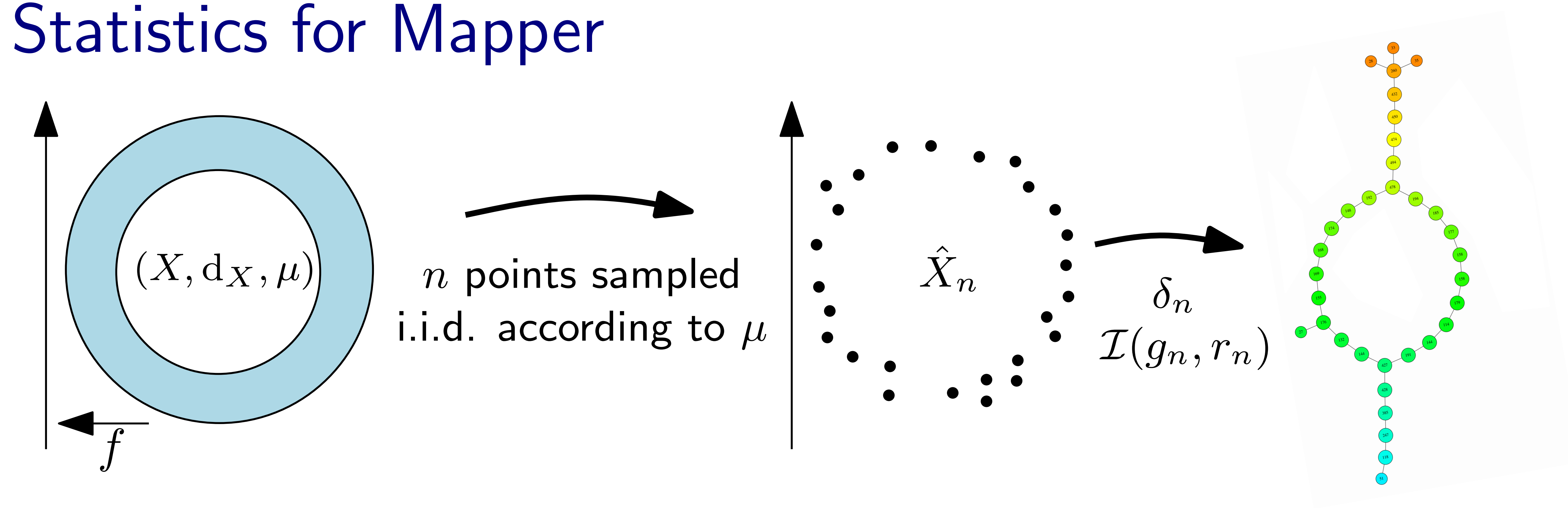
Thm: If μ is (a,b)-standard and f is c-Lipschitz, then for:

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

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

where C depends only on a, b, c.

Statistics for Mapper



Ex : PCA filter

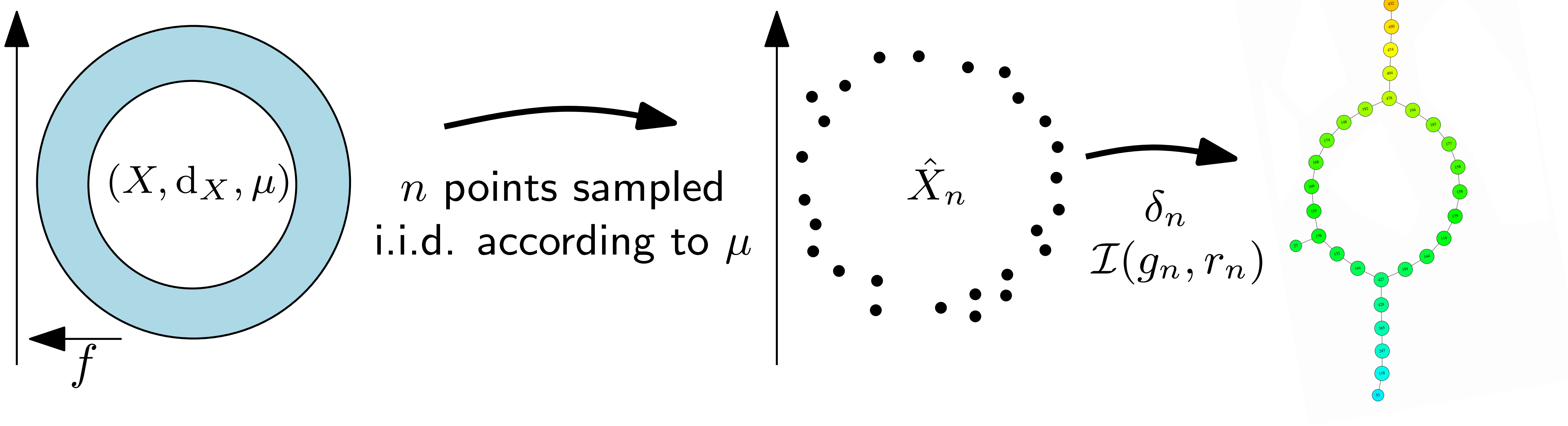
Π_1 : orthonormal projection onto first principal direction of covariance operator

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

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

[*PCA-Kernel Estimation*, Biau, Mas, *Statistics & Risk Modeling with Applications in Finance and Insurance*, 2012]

Statistics for Mapper



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

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

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

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

[*Topological Analysis of Nerves, Reeb Spaces, Mappers, and Multiscale Mappers*, Dey, Mémoli, Wang, SoCG, 2017]

Def: The *filter-based pseudometric* $d_f : M \times M \to \mathbb{R}$ is defined as

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

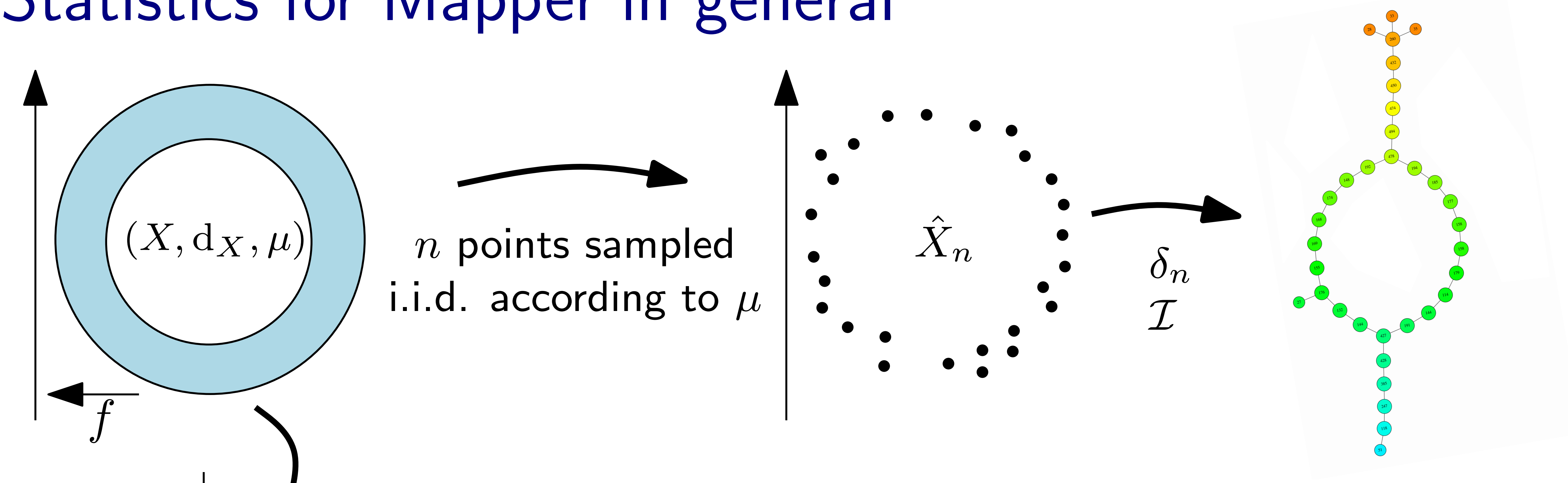
where $\Gamma(x, x')$ denotes the set of all continuous paths $\gamma : [0, 1] \to M$ such that $\gamma(0) = x$ and $\gamma(1) = x'$, and diam_Y denotes the *diameter* of a subset of Y.

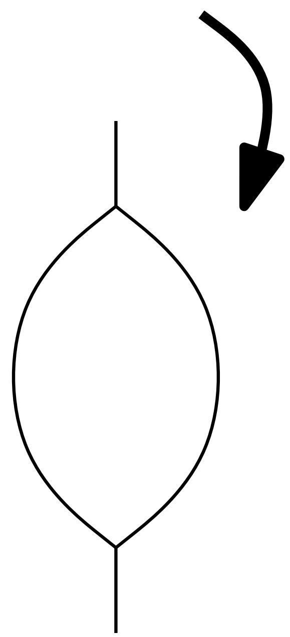
Def: The *Gromov-Hausdorff* metric d_{GH} between $(M, d_f), (M', d_{f'})$ is defined as

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

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



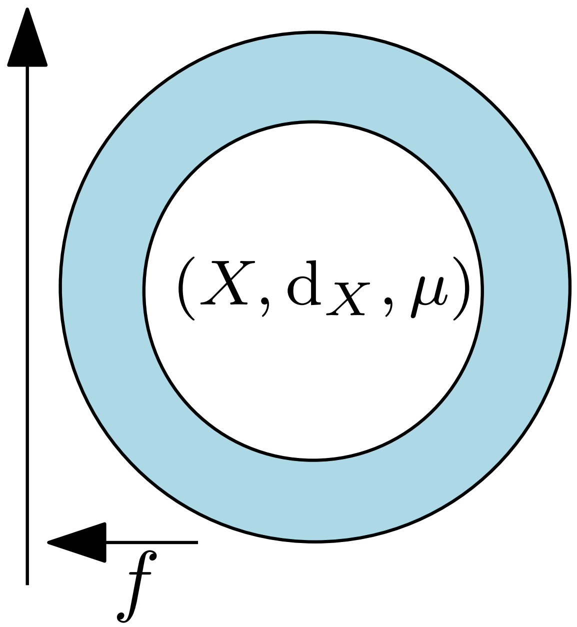


$$\mathbb{E}\left[d_{\mathrm{GH}}(\mathrm{M}_f, \mathrm{R}_f) \leq \,?\right] \geq 0.95$$

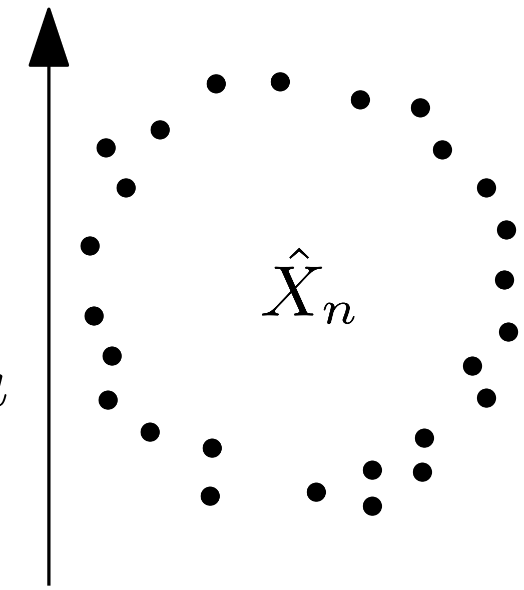
Question:

How to assess distance confidence?

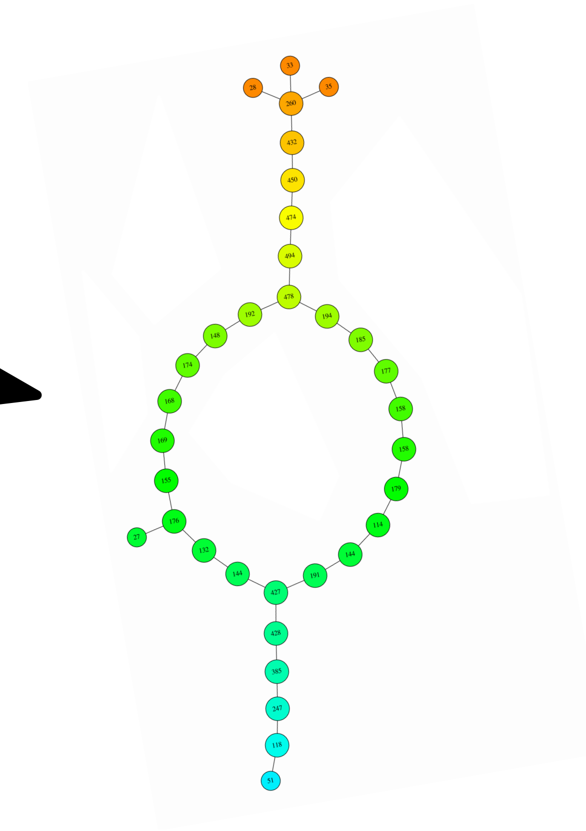
Statistics for Mapper in general



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



δ_n
$\mathcal{I}$



Thm: If μ and $f\#\mu$ are (a,b)-standard, then for δ_n as before, one has:

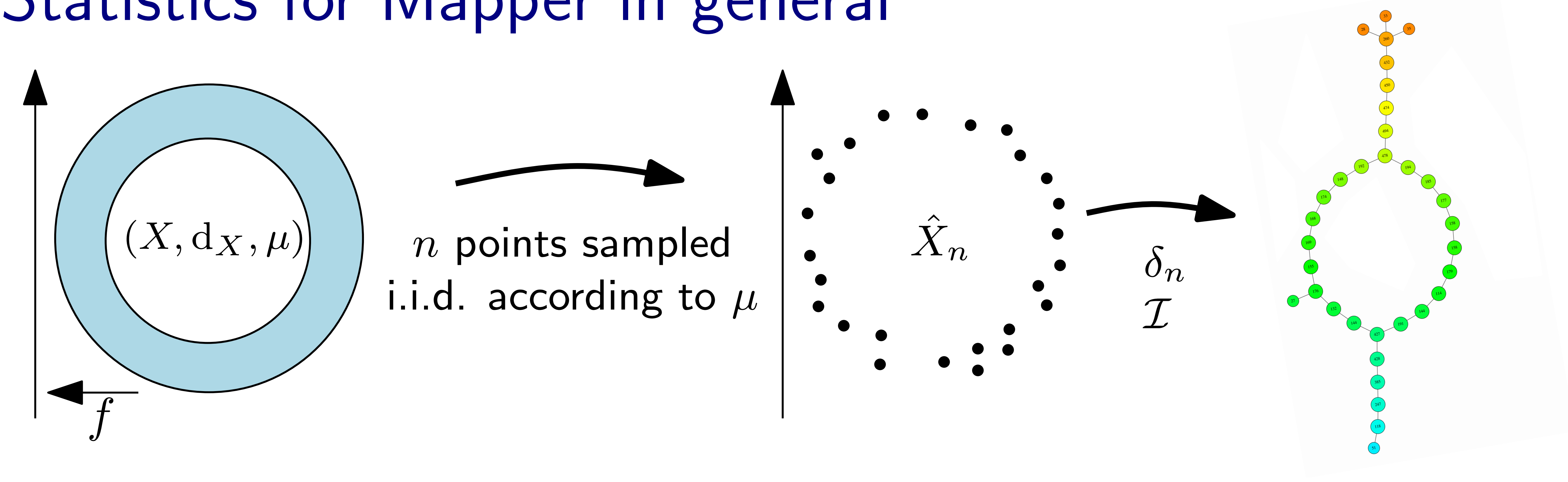
$$\mathbb{E}\left[d_{\mathrm{GH}}(\mathrm{M}^{\bullet}_{f,\delta_n}(\hat{X}_n,\mathcal{I}),\mathrm{R}_f(X))\right] \leq 5\cdot\mathbb{E}\left[\mathrm{res}(\mathcal{I})\right] + C\omega\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b},$$

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 allows to bound $\mathbb{E}\left[\mathrm{res}(\mathcal{I})\right]$.

[*A k-points-based distance for robust geometric inference*, Brecheteau, Levrard, *Bernouilli*, 2020]

Statistics for Mapper in general



Thm: 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 $\mathbb{R}^D$:

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

Other works

Another line of work is about the *interleaving distance* between Mappers and Reeb spaces *seen as cosheaves* **Open**$(\mathbb{R}^d) \to$ **Set**.

Prop: For $f : X \to \mathbb{R}^d$, $d_I(\mathcal{C}(\mathrm{R}_f(X)), \mathcal{C}(\mathrm{M}_f(X, \mathcal{I}))) \leq \mathrm{res}(\mathcal{I})$

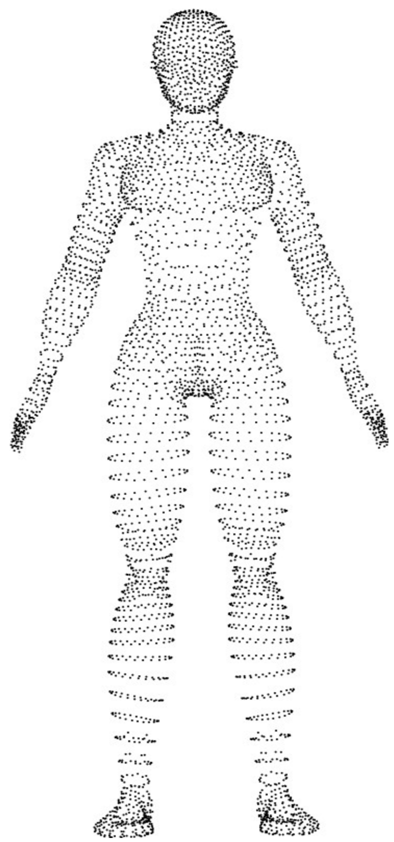
Prop: For $f : X \to \mathbb{R}$,

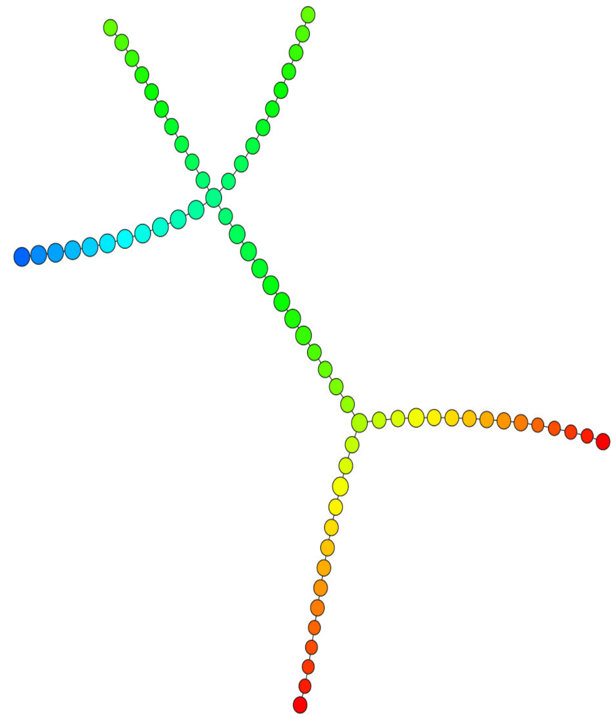
$$\lim_{n\to+\infty} \ \mathbb{P}\left(d_I(\mathcal{C}(\mathrm{R}_f(X)), \mathcal{C}(\mathrm{M}_f(\hat{X}_n, \mathcal{I}))) \leq \mathrm{res}(\mathcal{I})\right) = 1$$

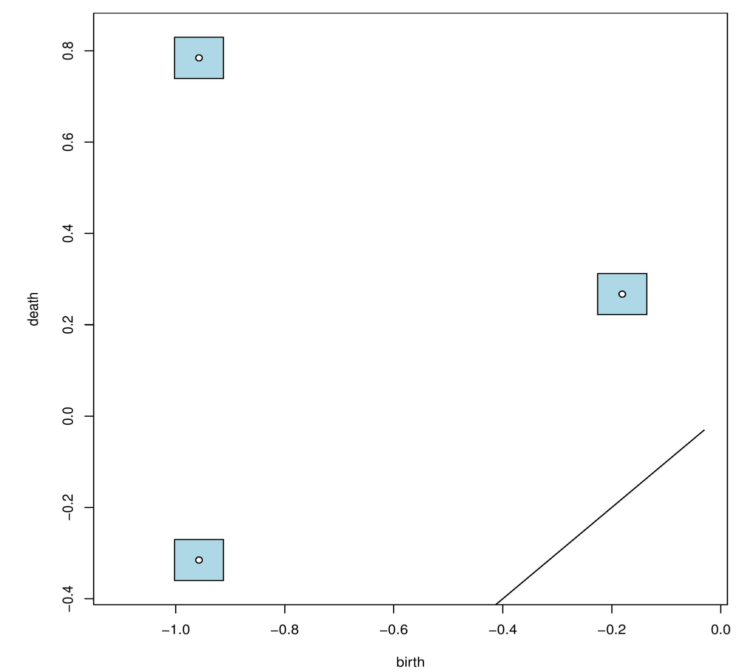
[*Convergence between categorical representations of Reeb space and Mapper*, Munch, Wang, SoCG, 2016]

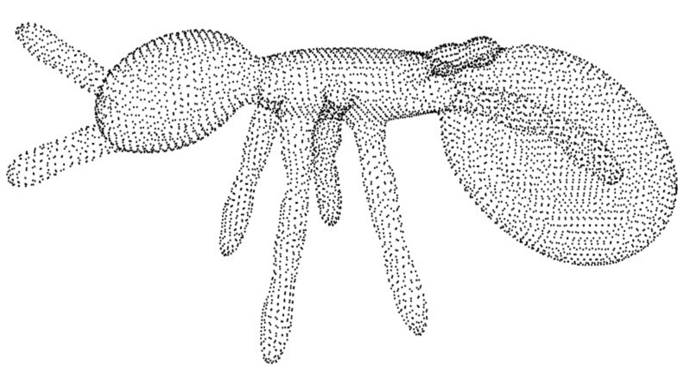
[*Probabilistic convergence and stability of random Mapper graphs*, Brown et al., JACT, 2020]

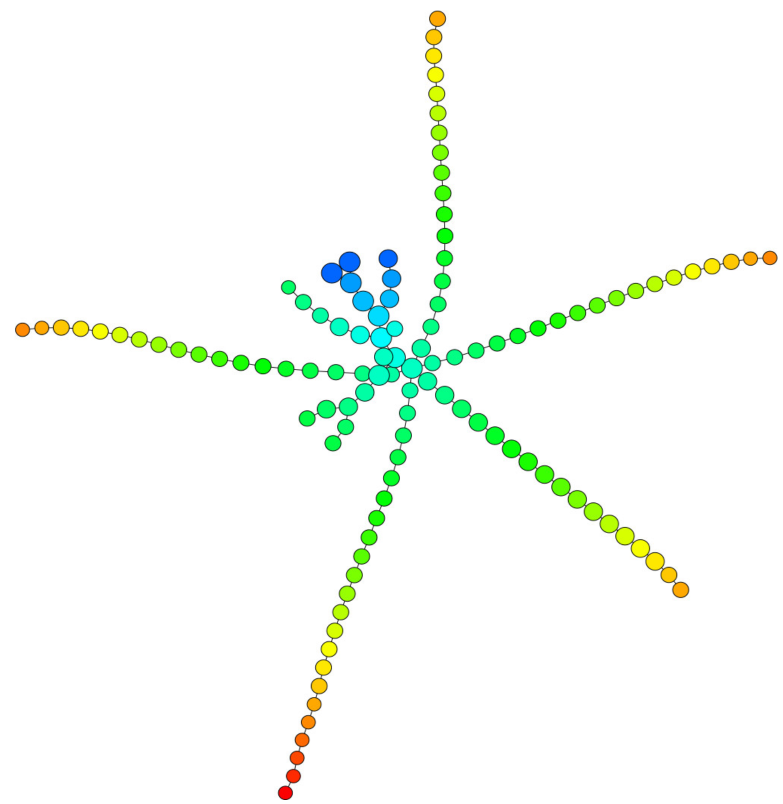
Experiments 85% confidence intervals

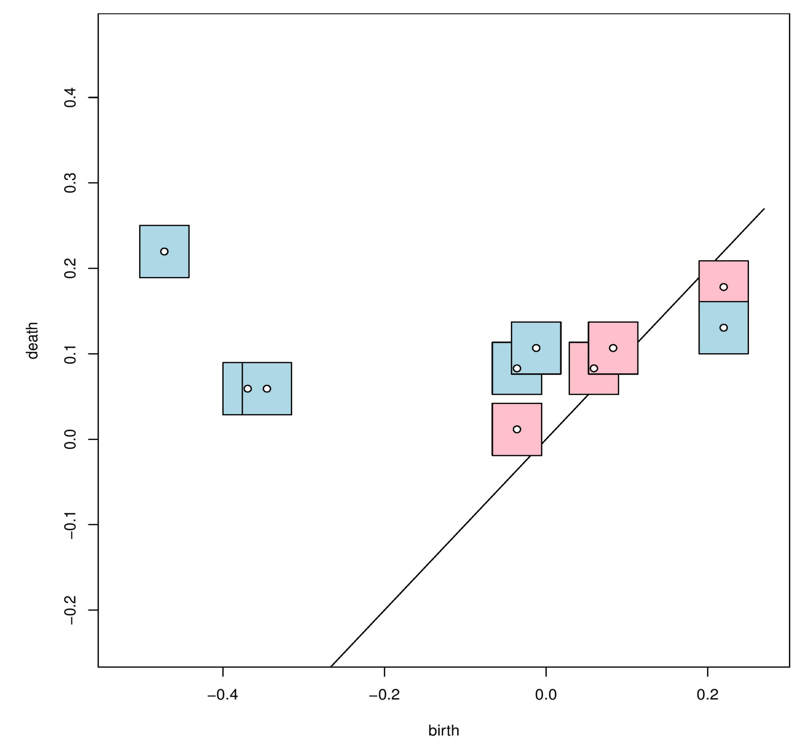




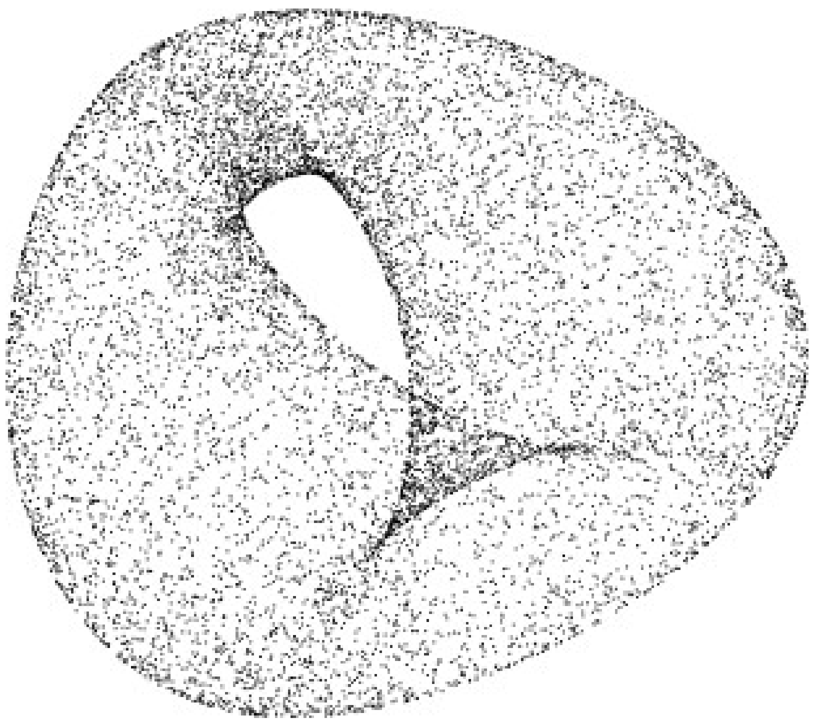


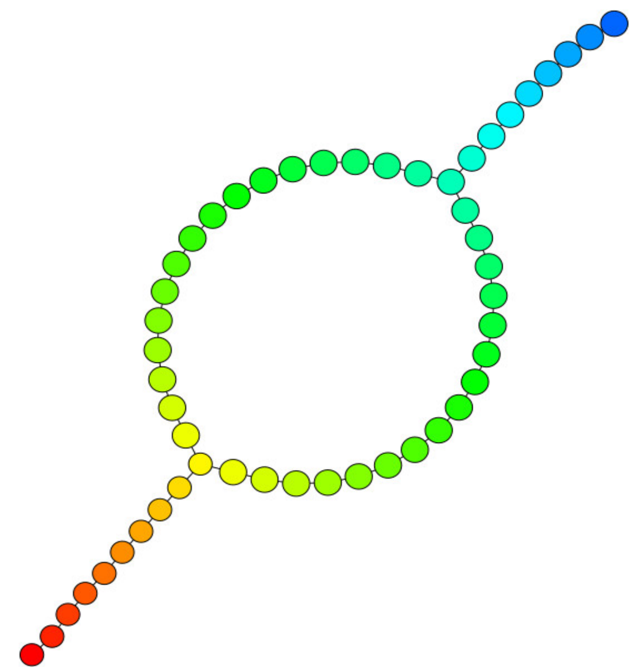


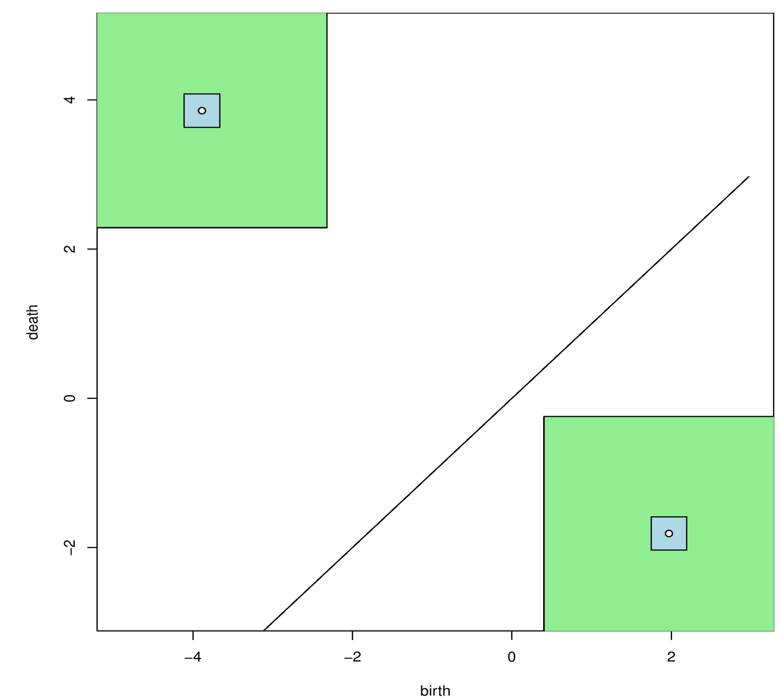


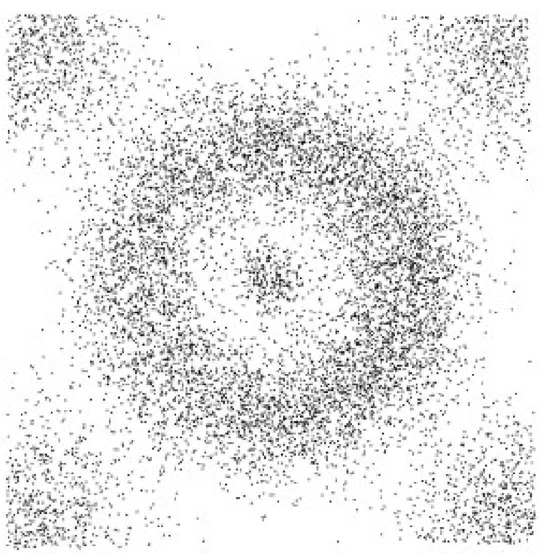


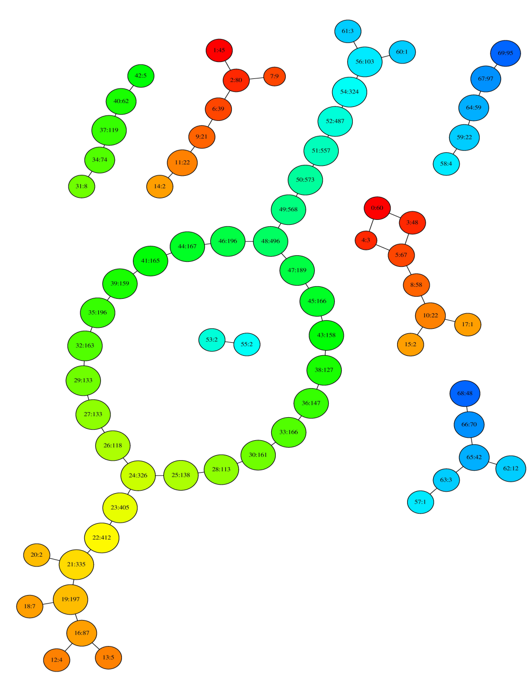
Experiments 85% confidence intervals

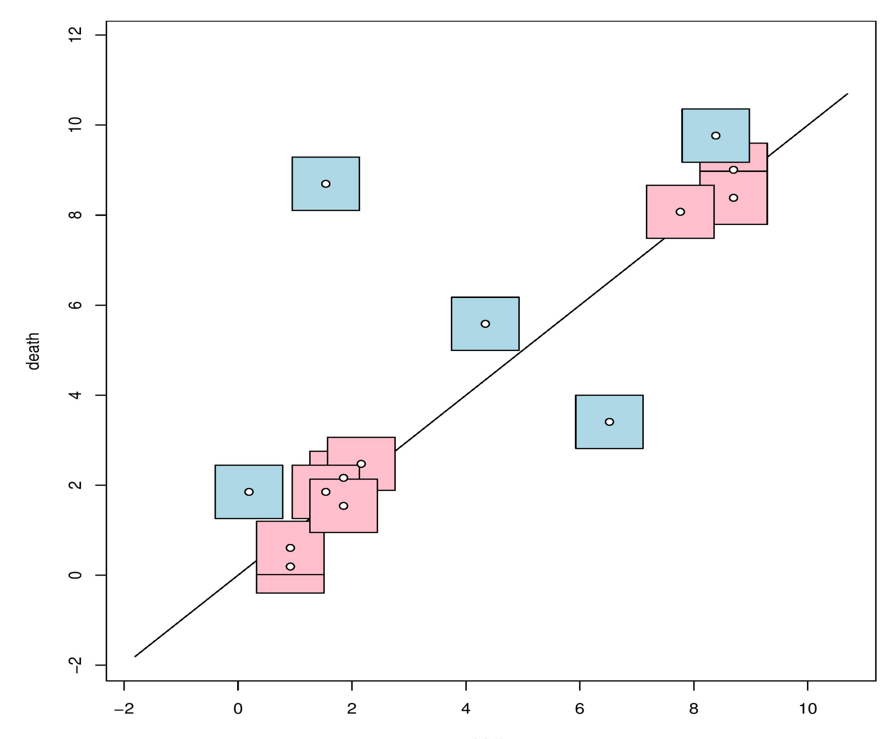




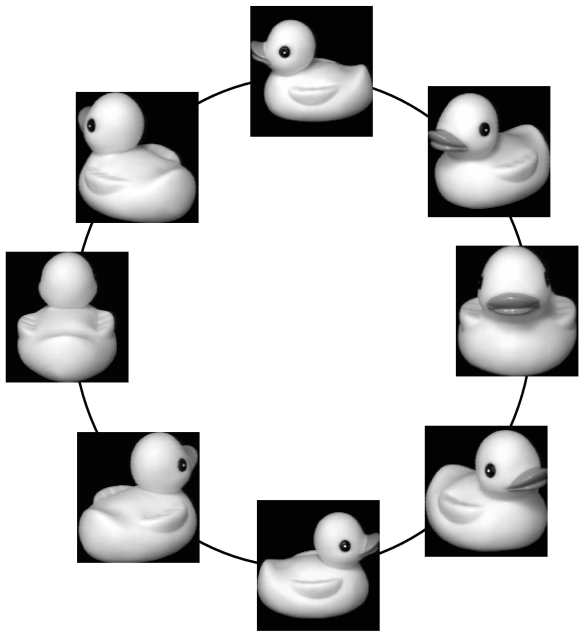


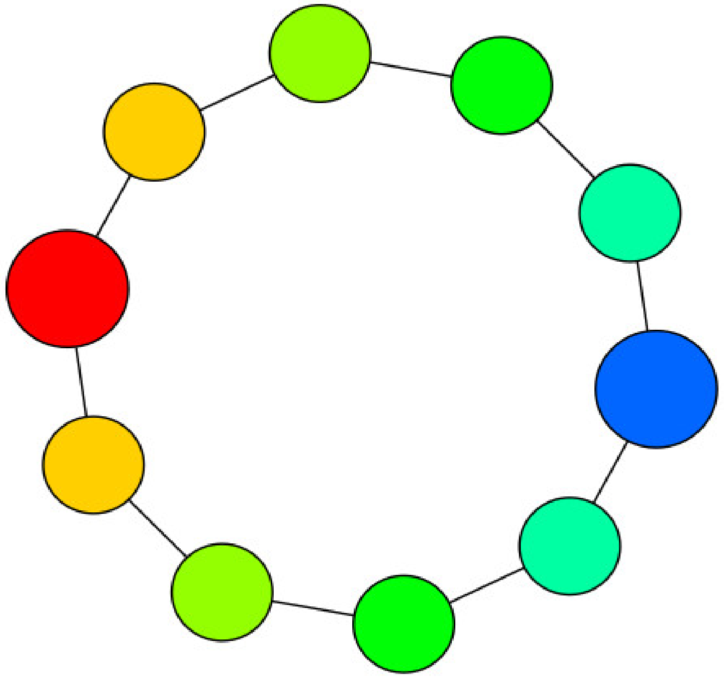


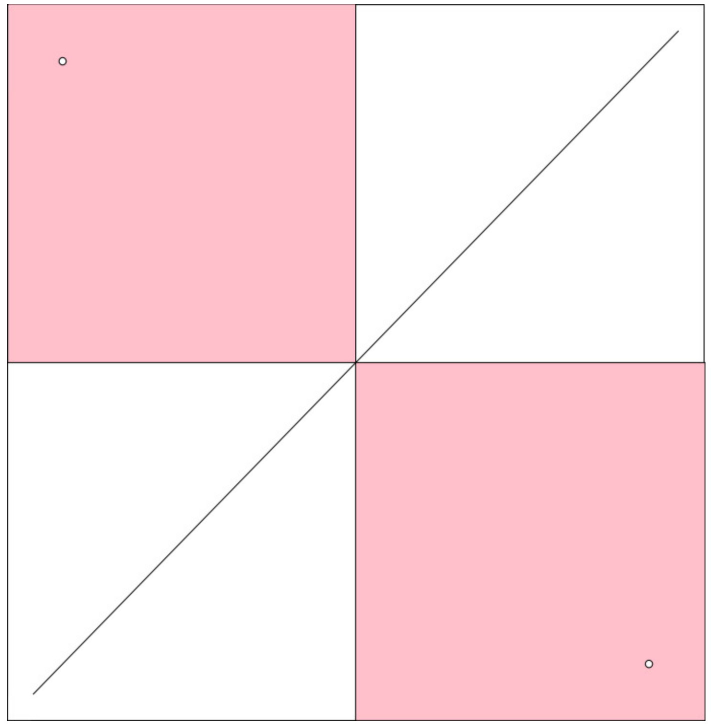


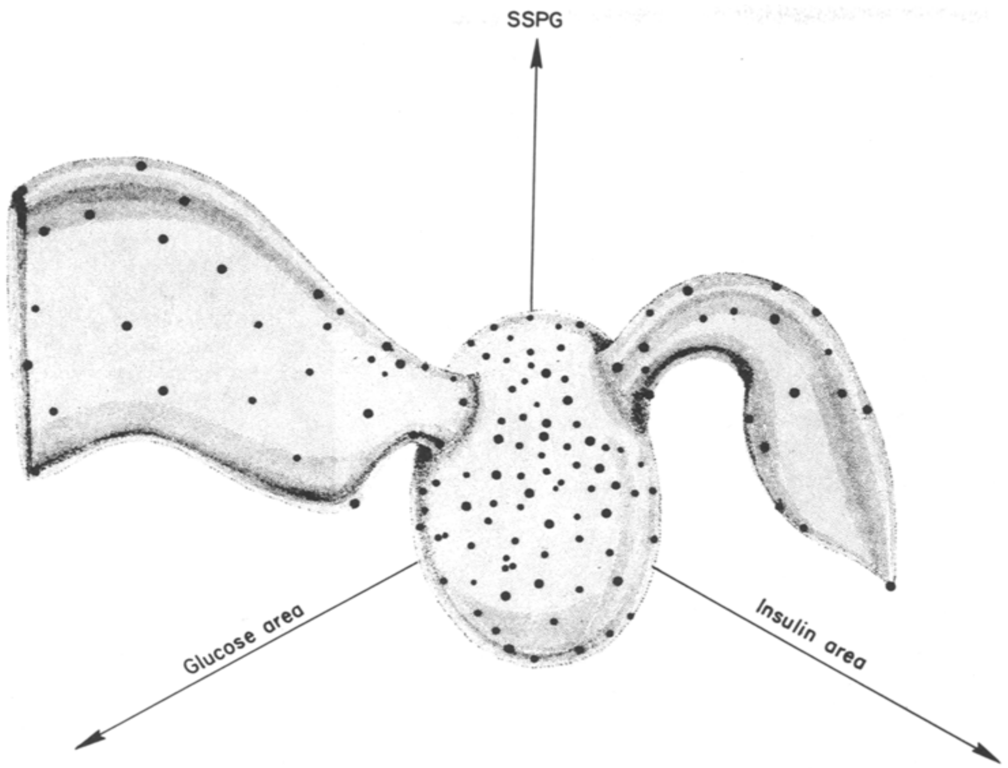


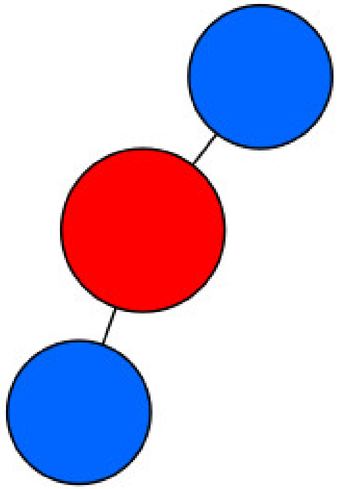
Experiments 85% confidence intervals

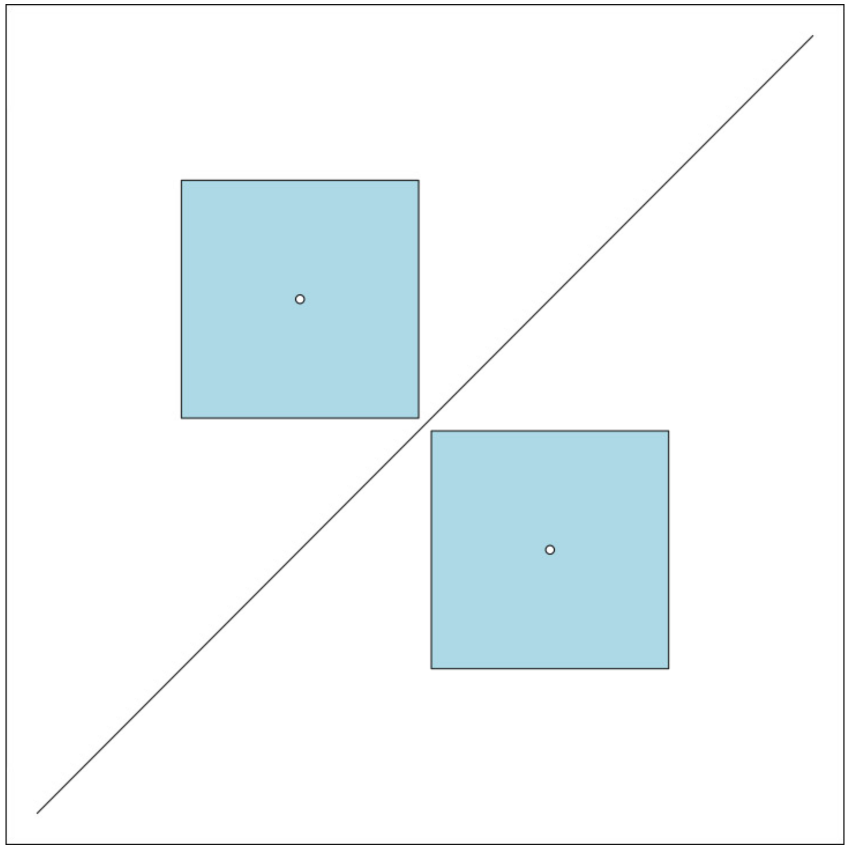




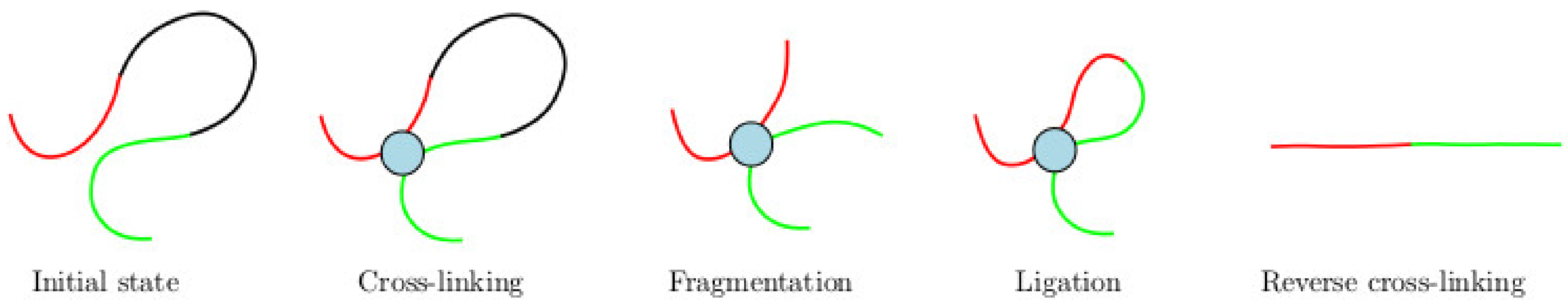


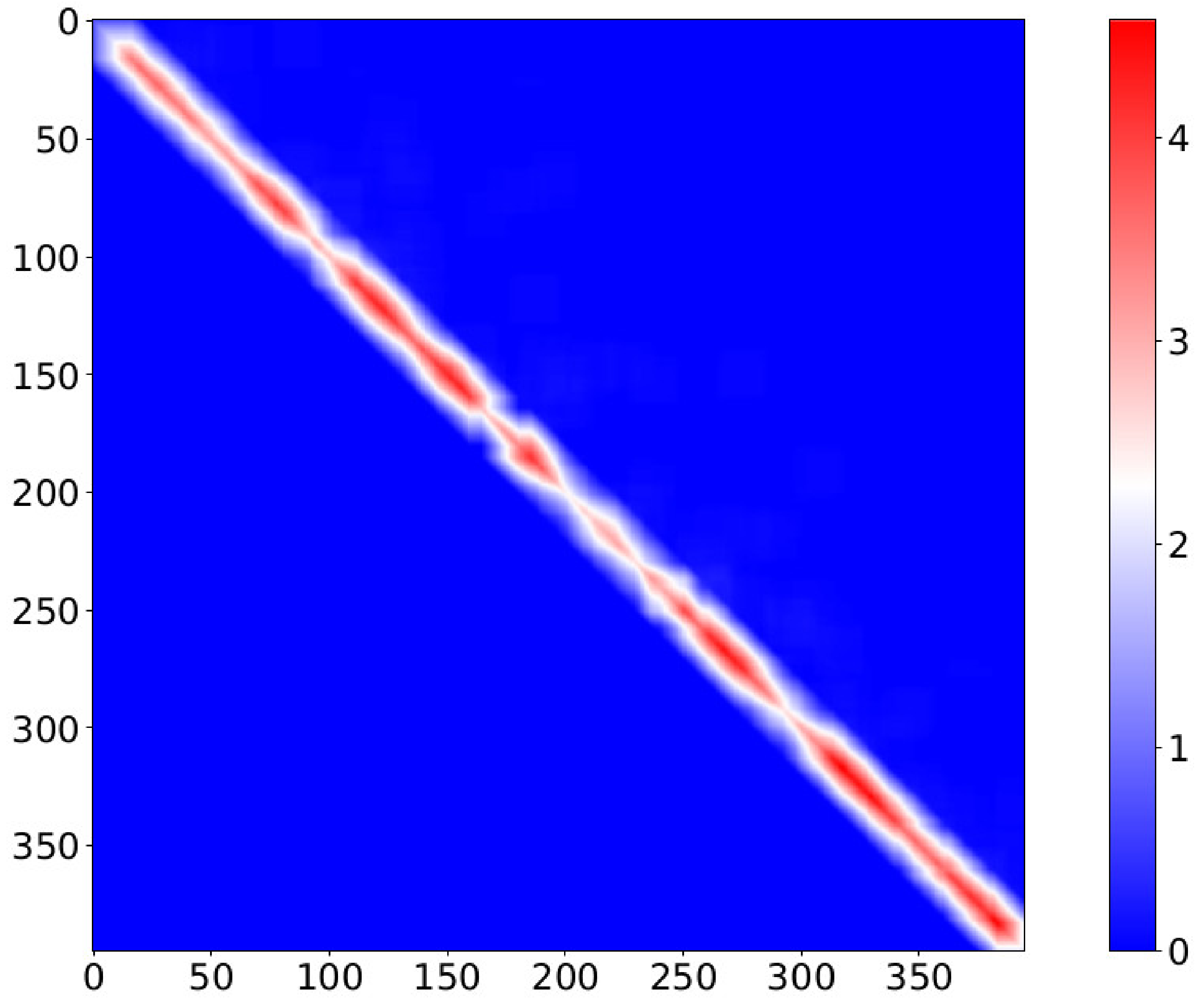




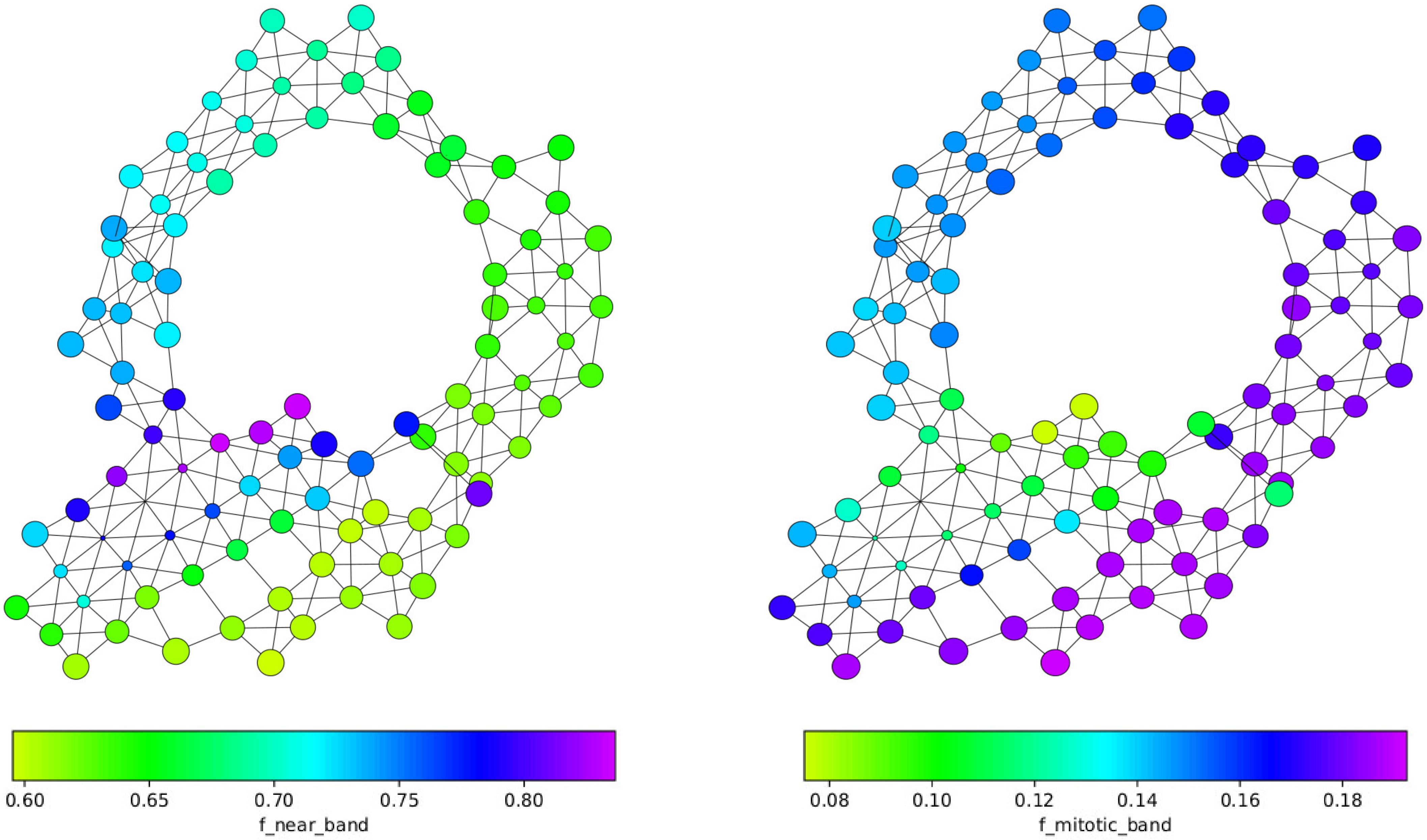


Experiments Chromosome conformation capture



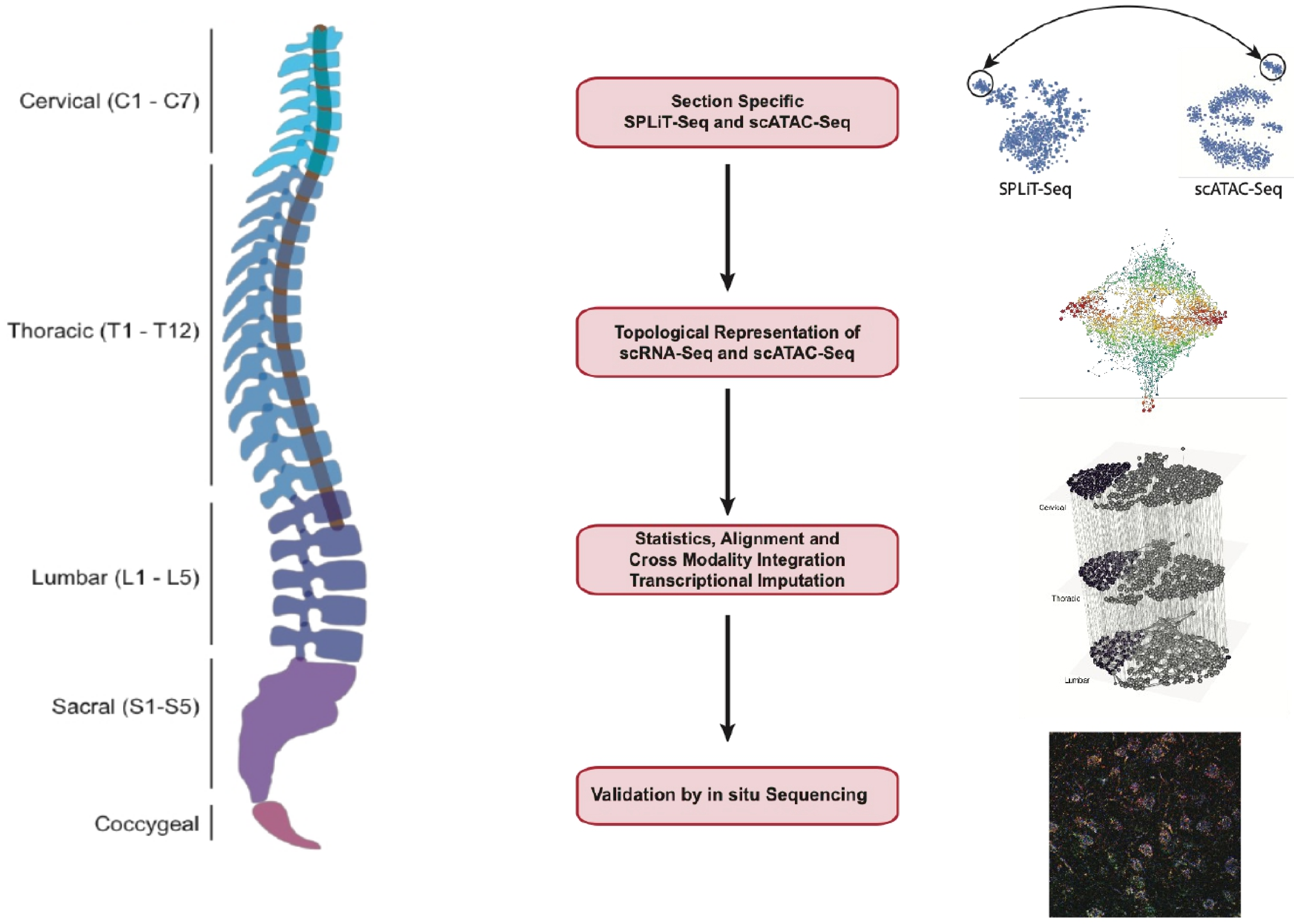


Experiments Chromosome conformation capture

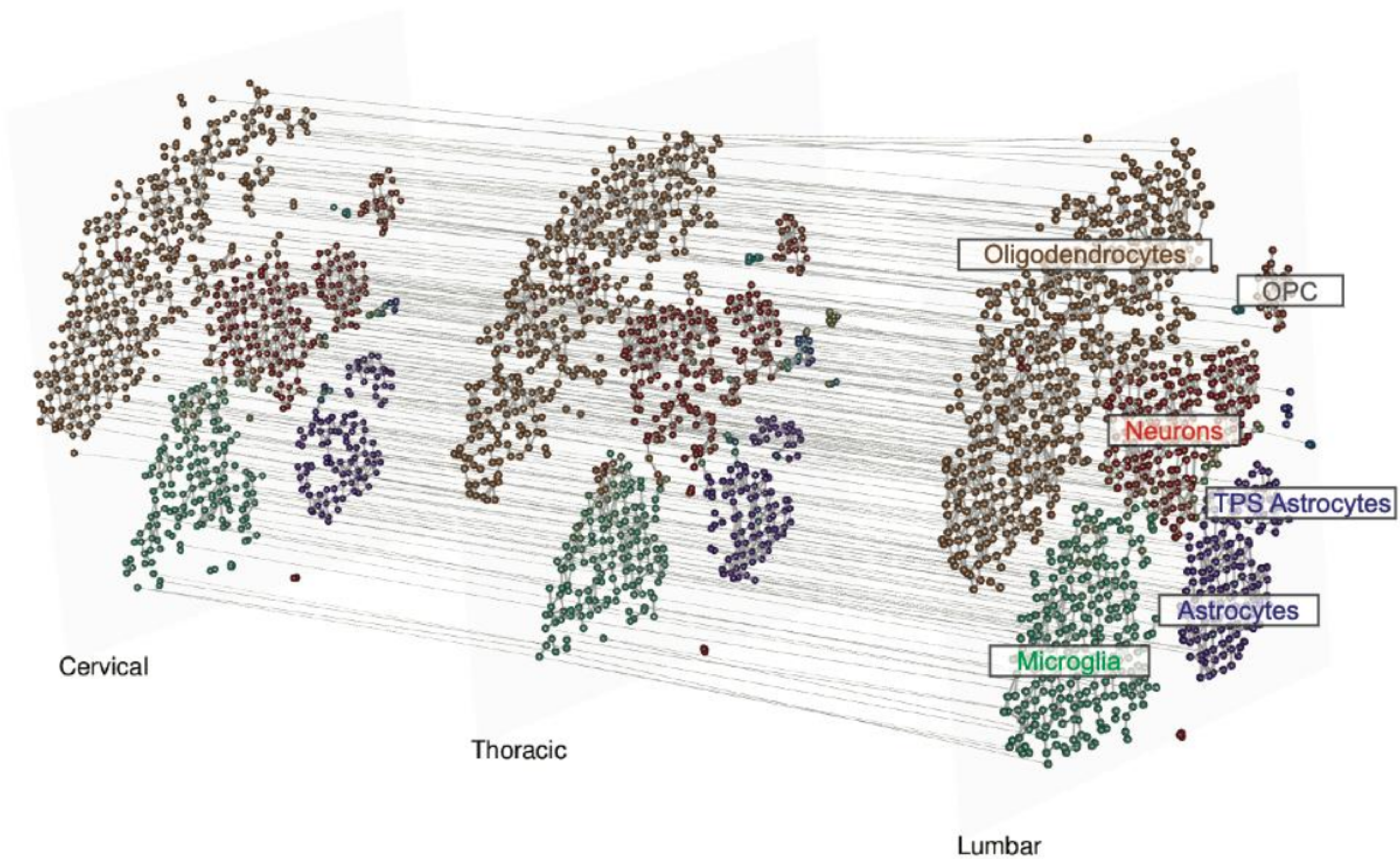


Formal identification of cell cycle with 95% confidence

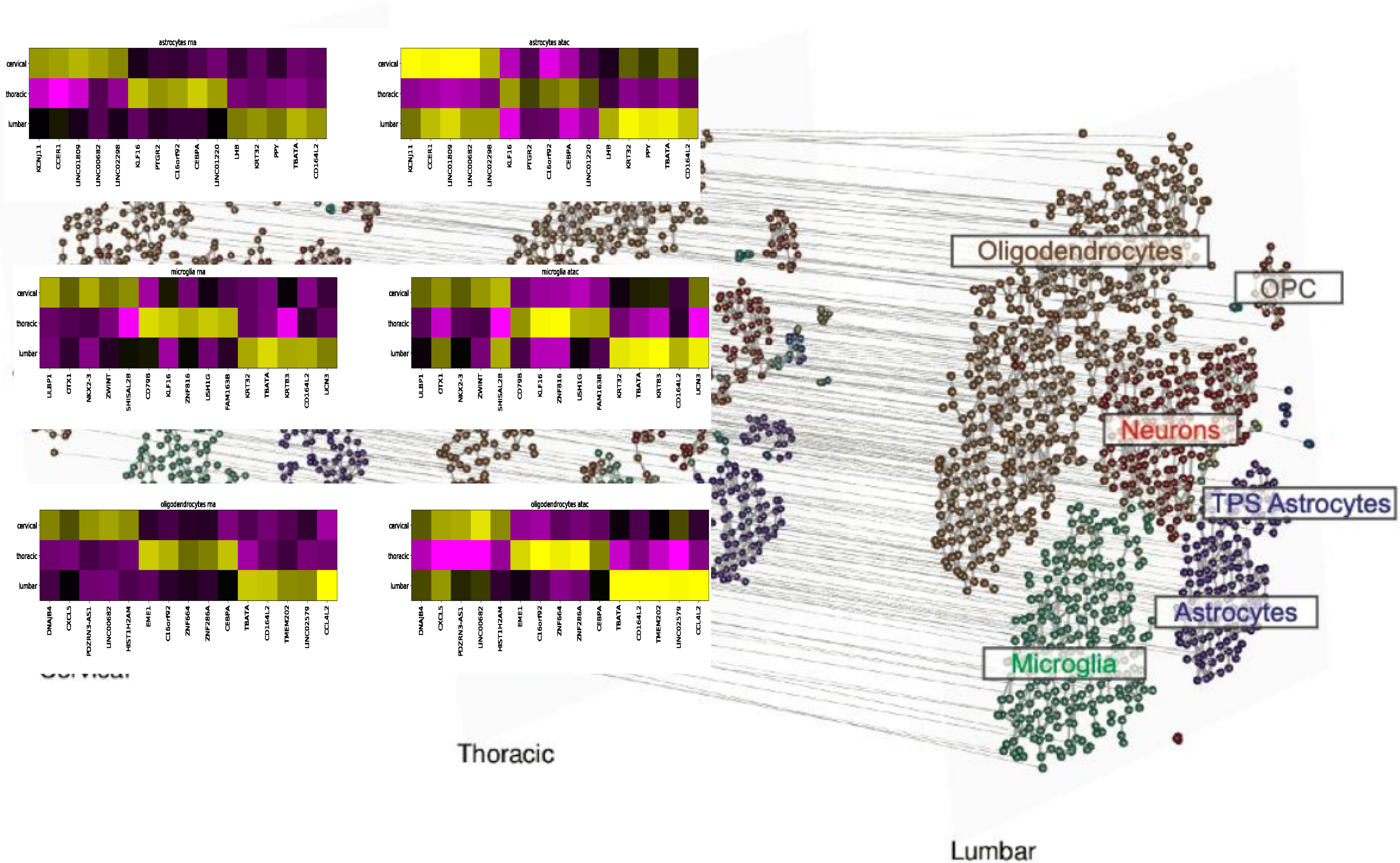
Experiments Spinal cord data



Experiments Spinal cord data

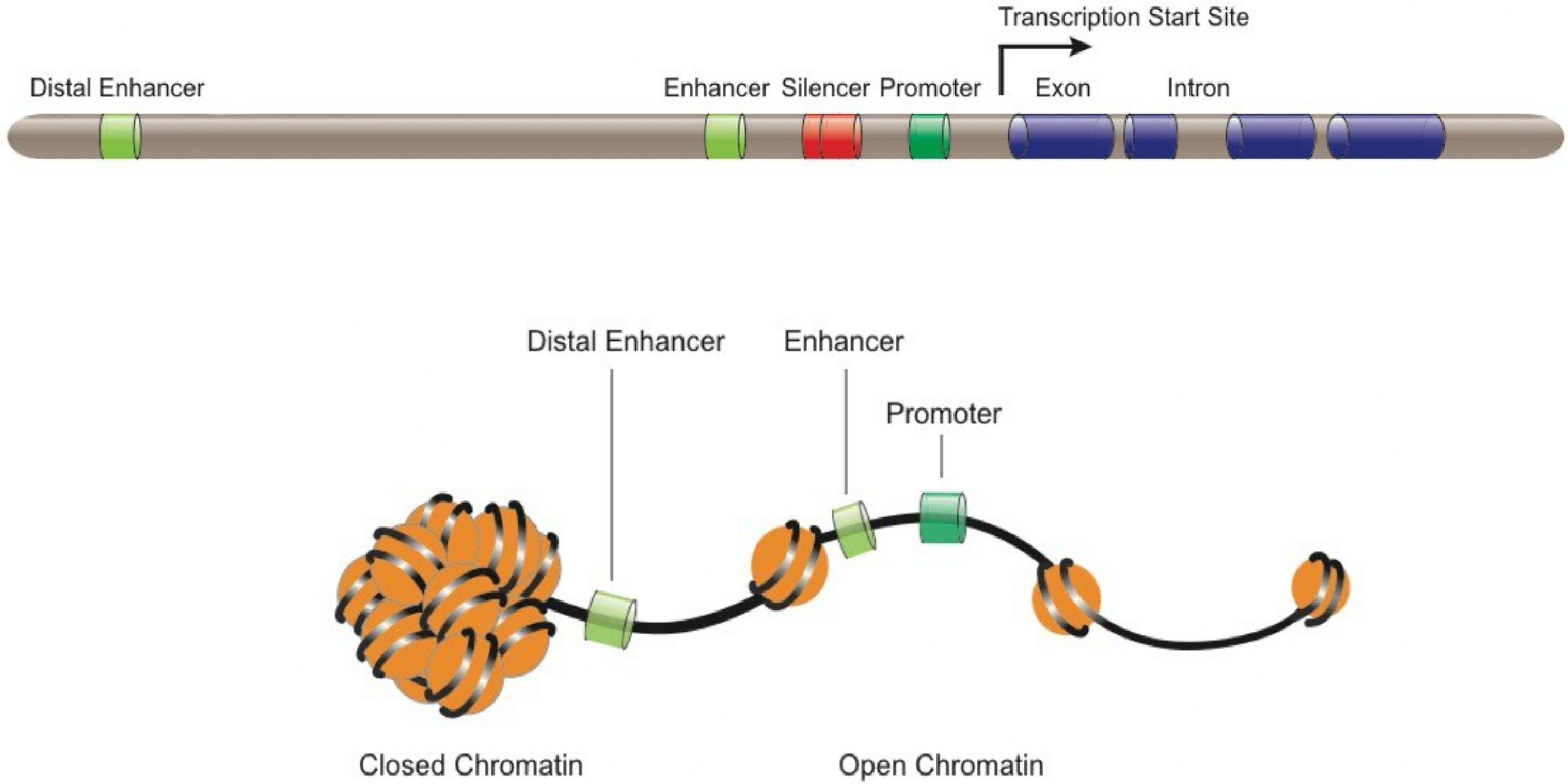


Experiments   Spinal cord data



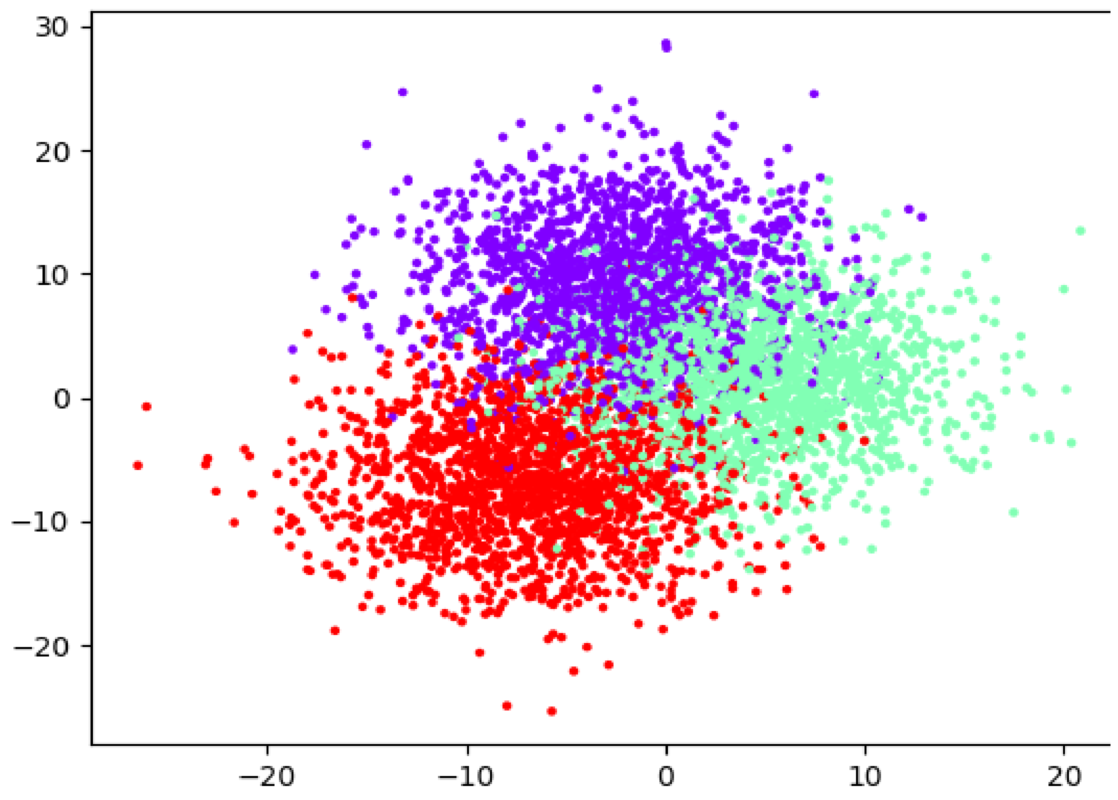
Experiments Spinal cord data

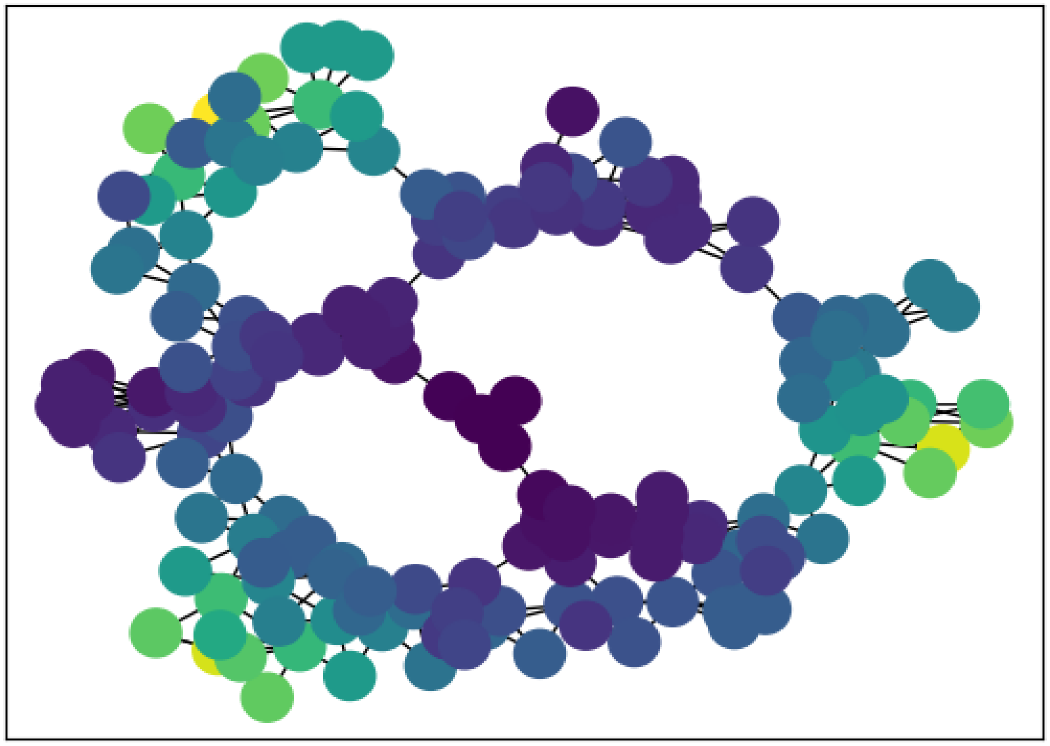
Gene expression (SPLiTseq) and gene accessibility (ATACseq) of single cells of one healthy individual for 3 sections of spinal cord



Experiments Machine learning classifier

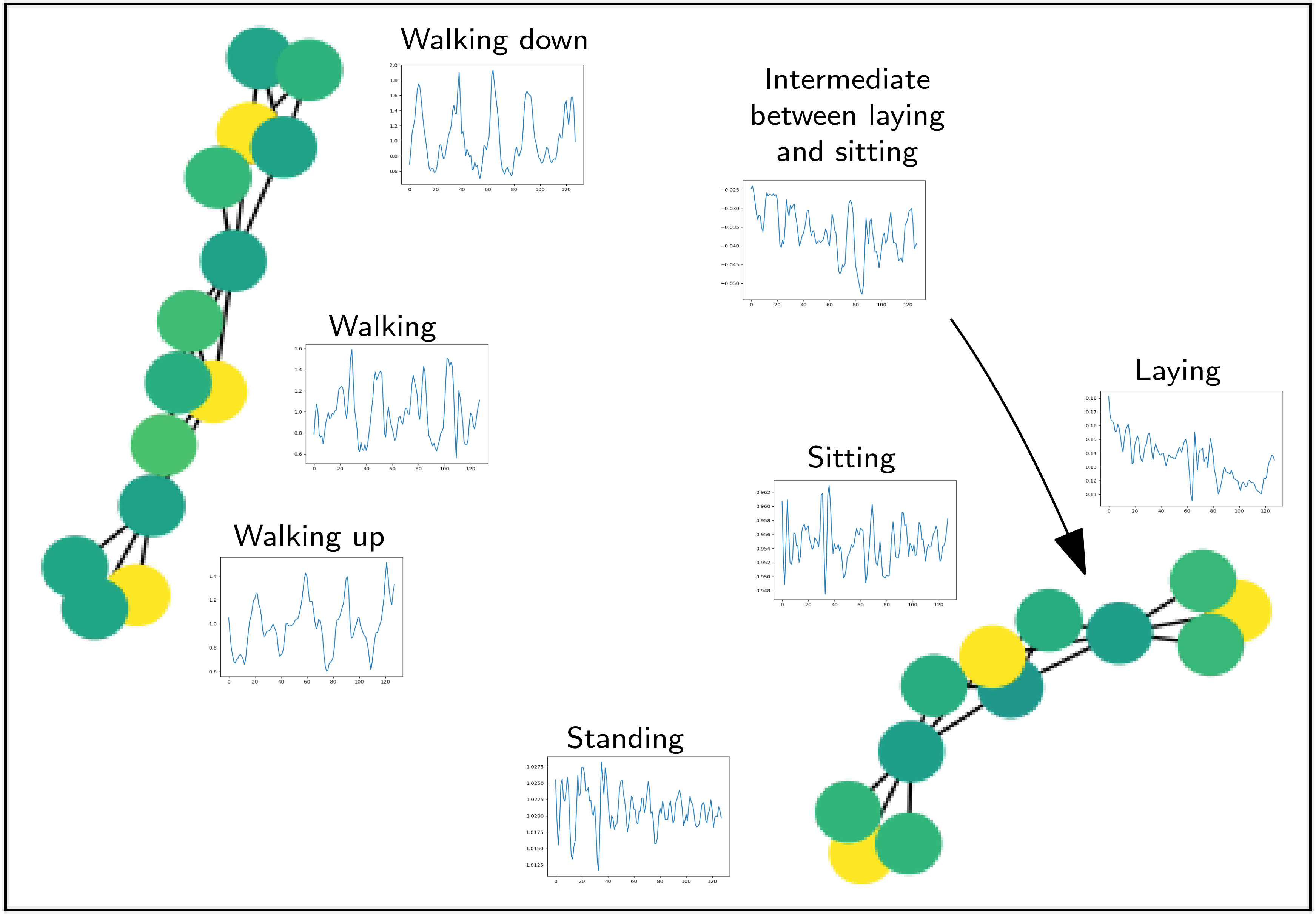
Filter = confidence of Random Forest classifier (in $\mathbb{R}^3$)





Experiments Machine learning classifier

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



Thanks!!