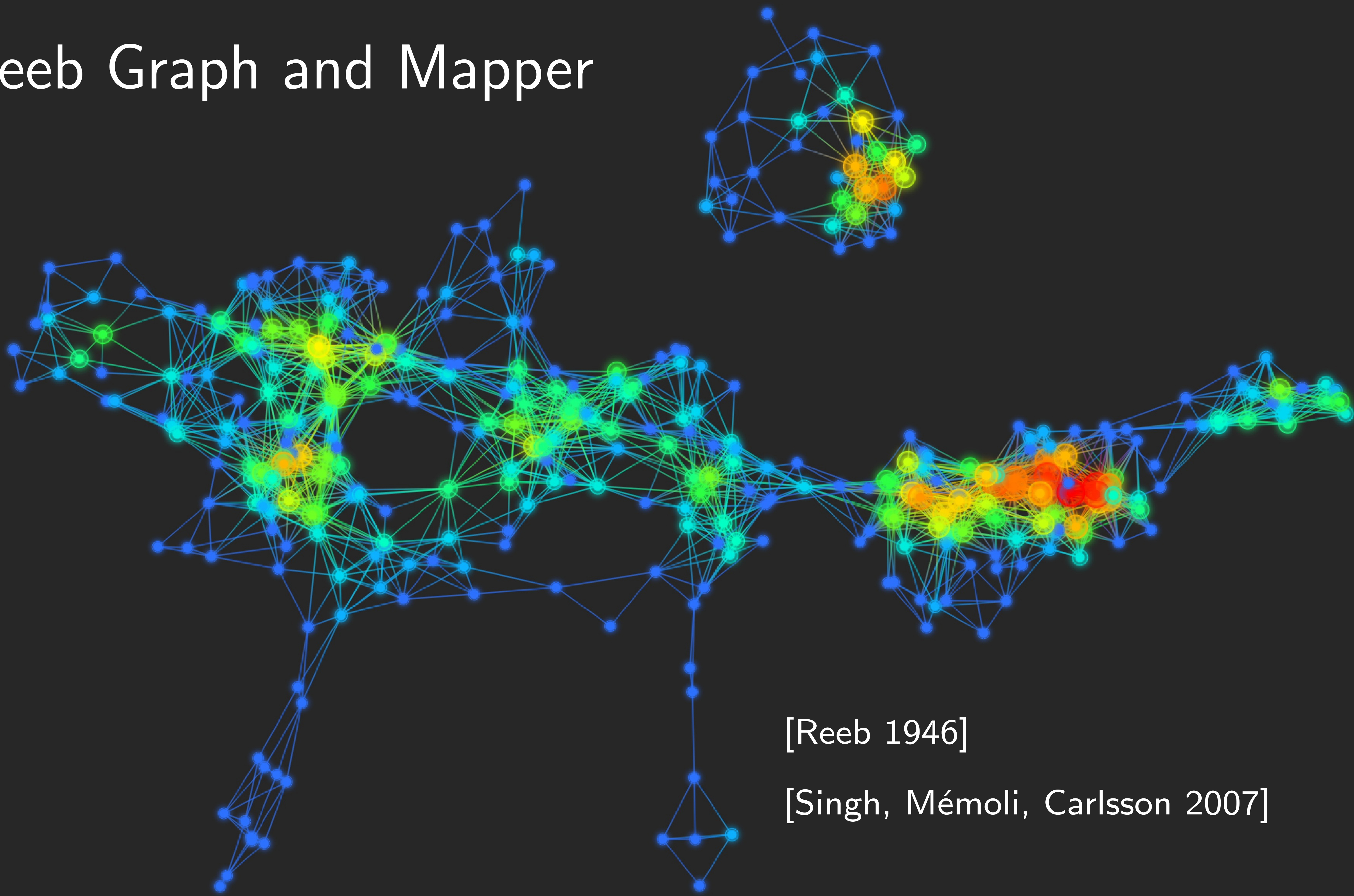
Reeb Graph and Mapper



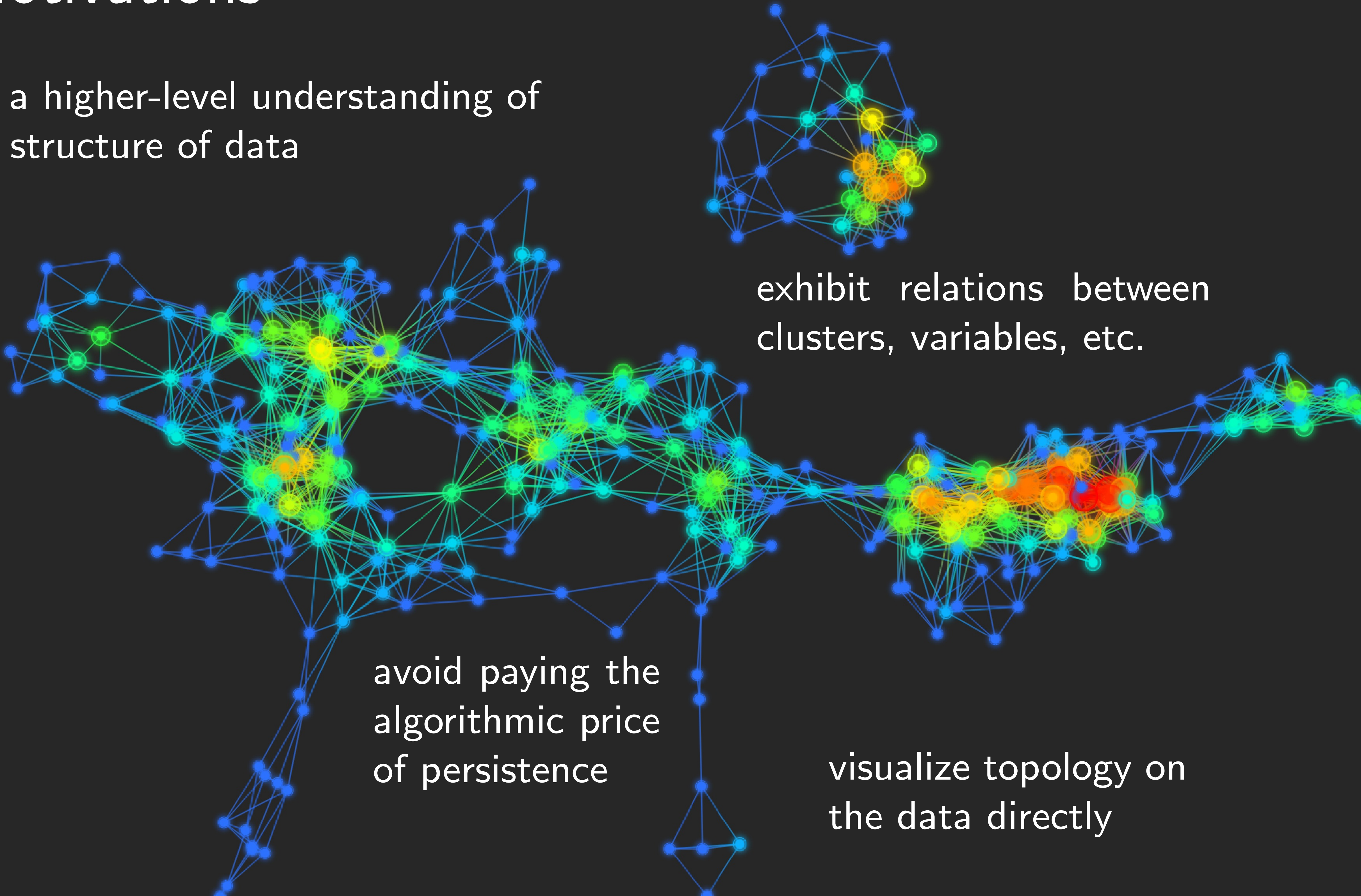
[Reeb 1946]

[Singh, Mémoli, Carlsson 2007]

Motivations

get a higher-level understanding of the structure of data

exhibit relations between clusters, variables, etc.



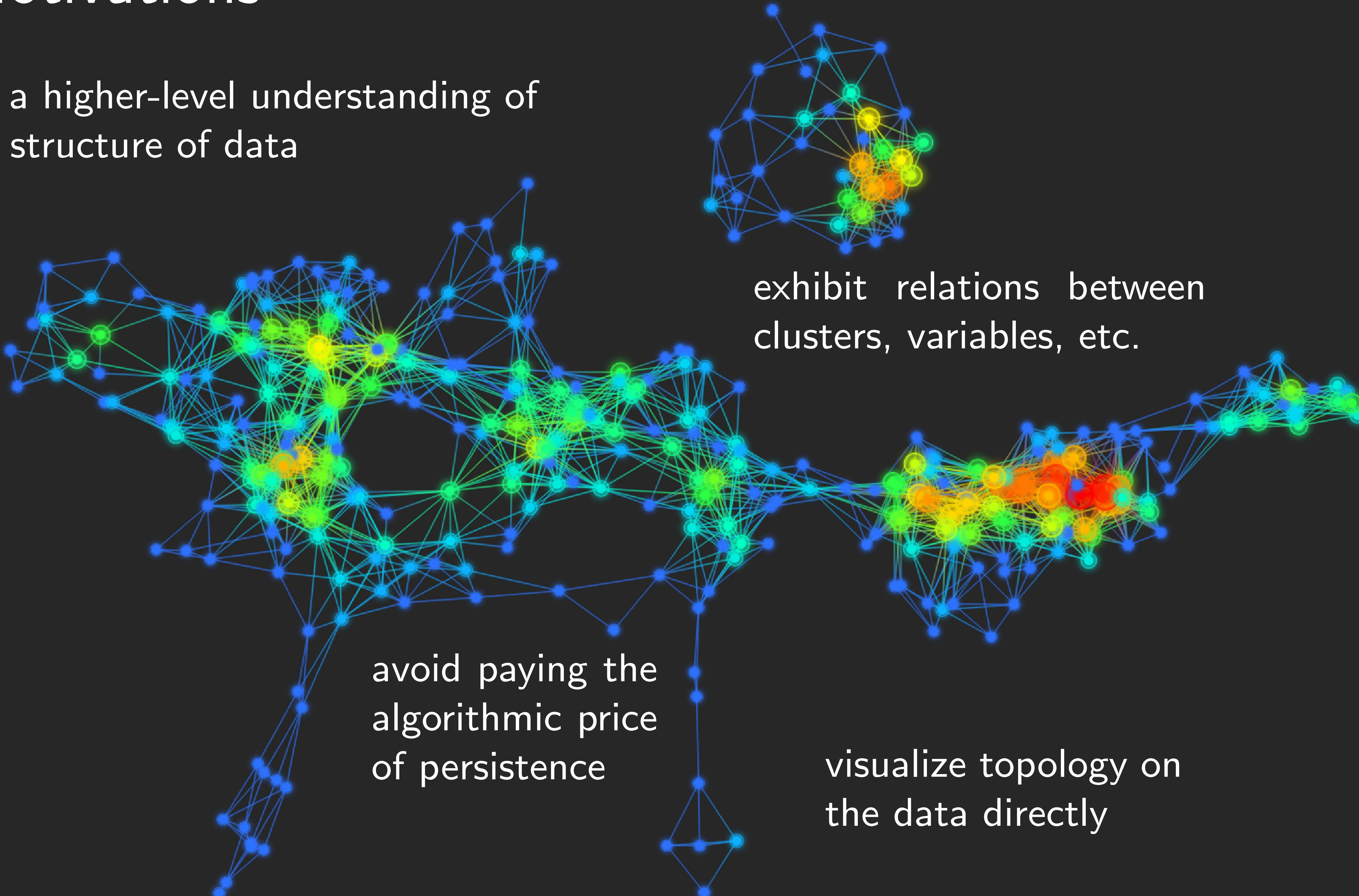
avoid paying the algorithmic price of persistence

visualize topology on the data directly

Motivations

get a higher-level understanding of the structure of data

exhibit relations between clusters, variables, etc.

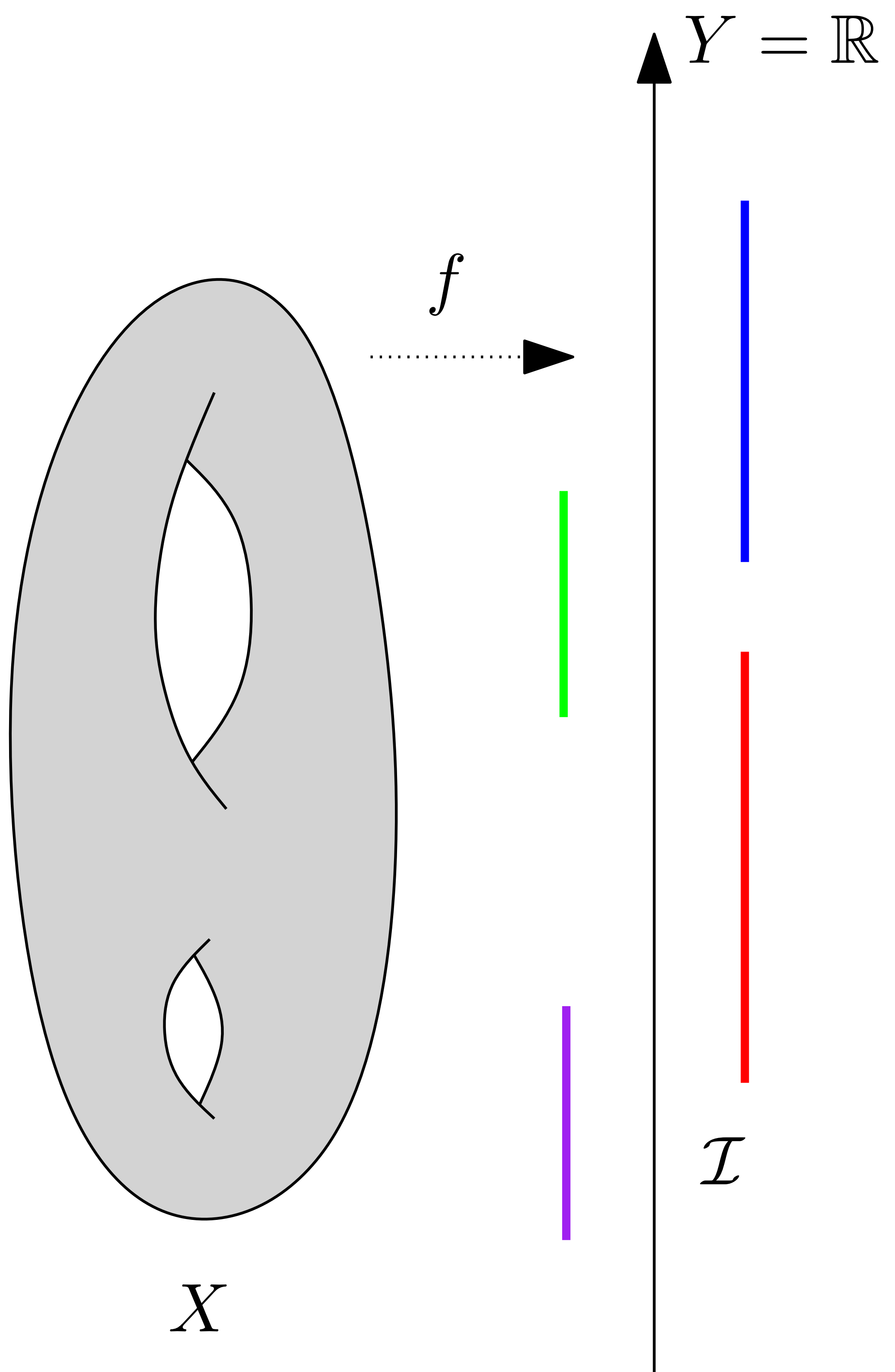


avoid paying the algorithmic price of persistence

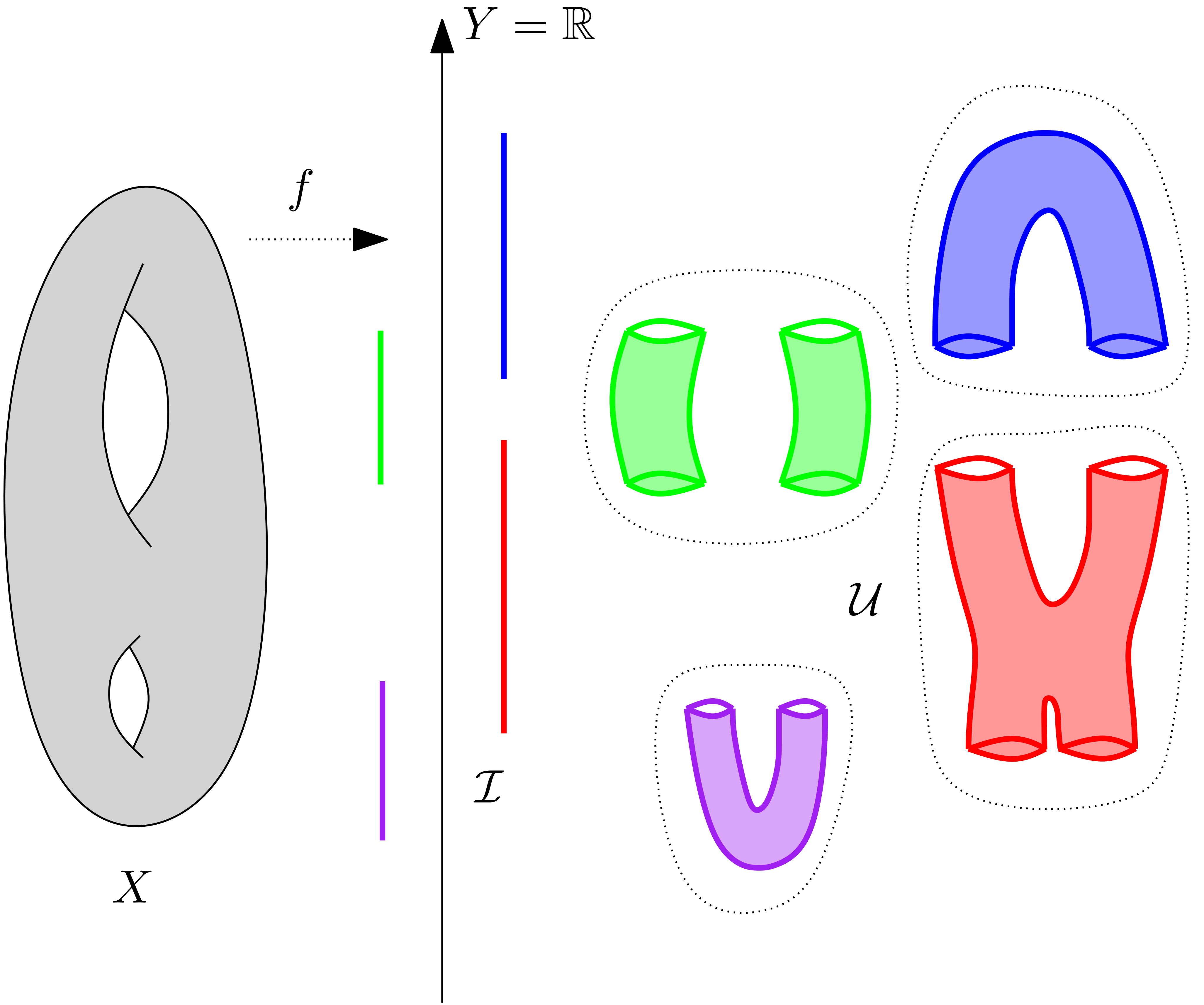
visualize topology on the data directly

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

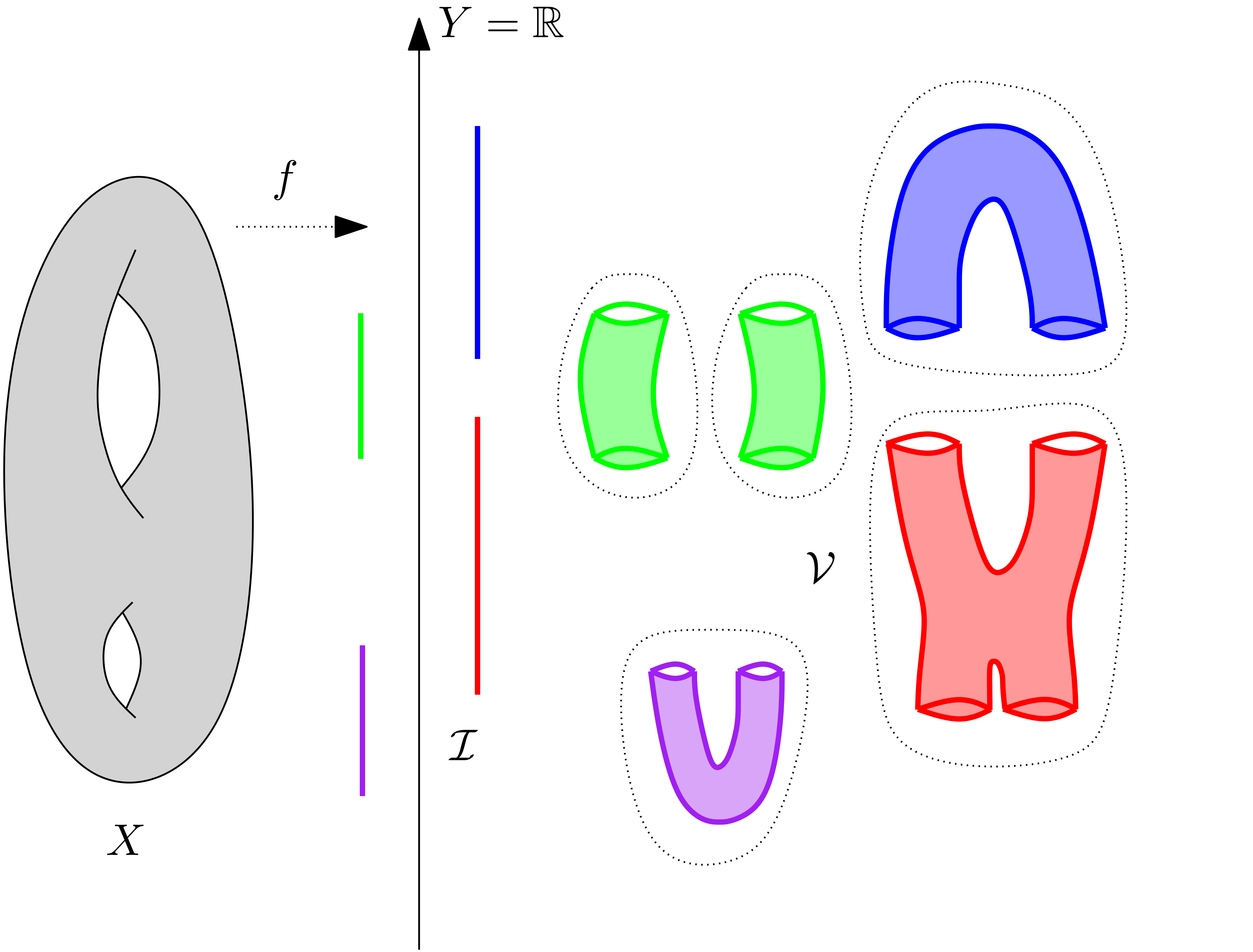
Mapper in the continuous setting



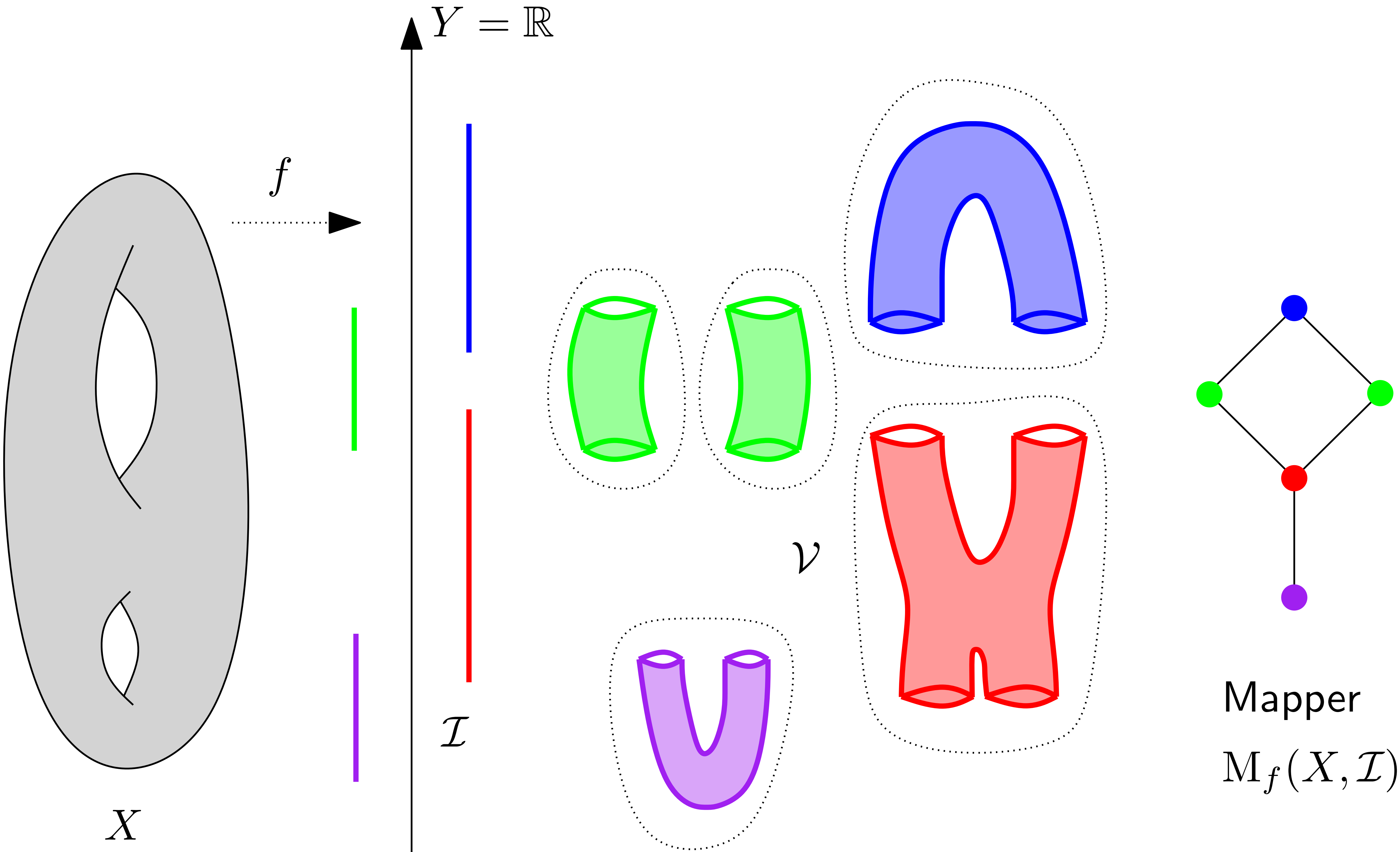
Mapper in the continuous setting



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 $f : P \to Y$ ($Y = \mathbb{R}$ 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 neighborhood graph $G = (P, E)$

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

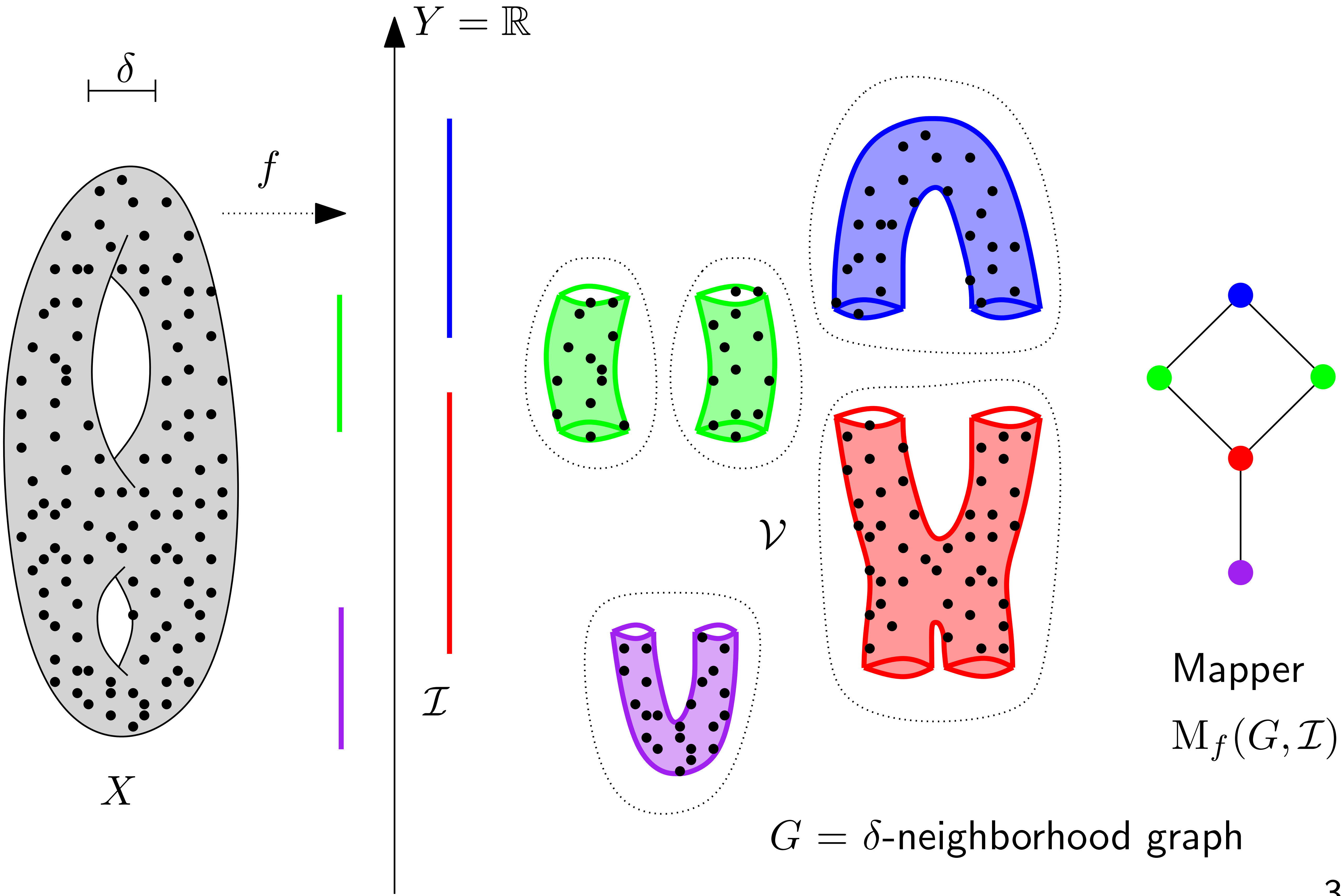
• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in G $\to$ connected cover $\mathcal{V}$

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

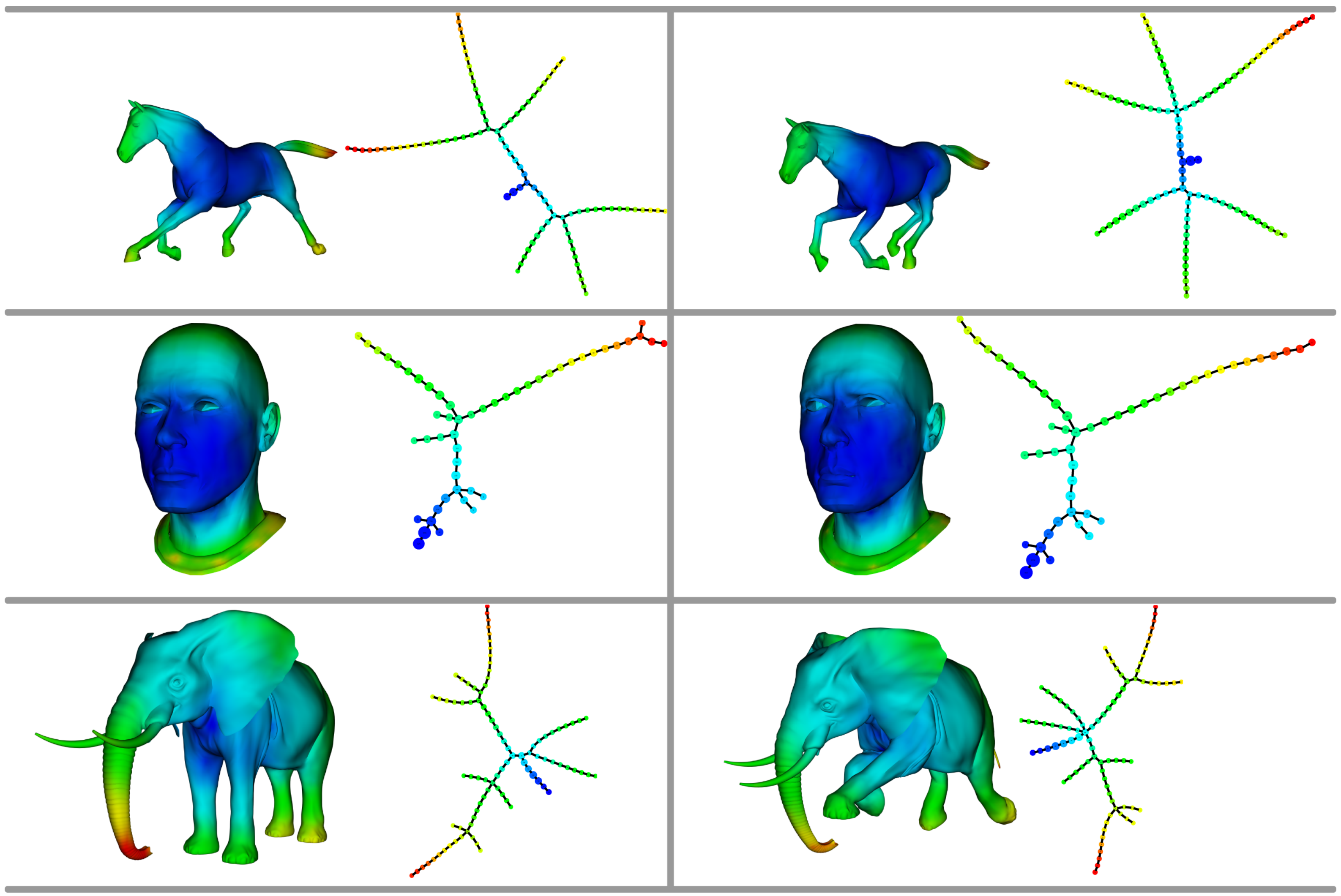
- 1 vertex per element $V \in \mathcal{V}$
- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$
- 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

(intersections materialized by data points)

Mapper in practice



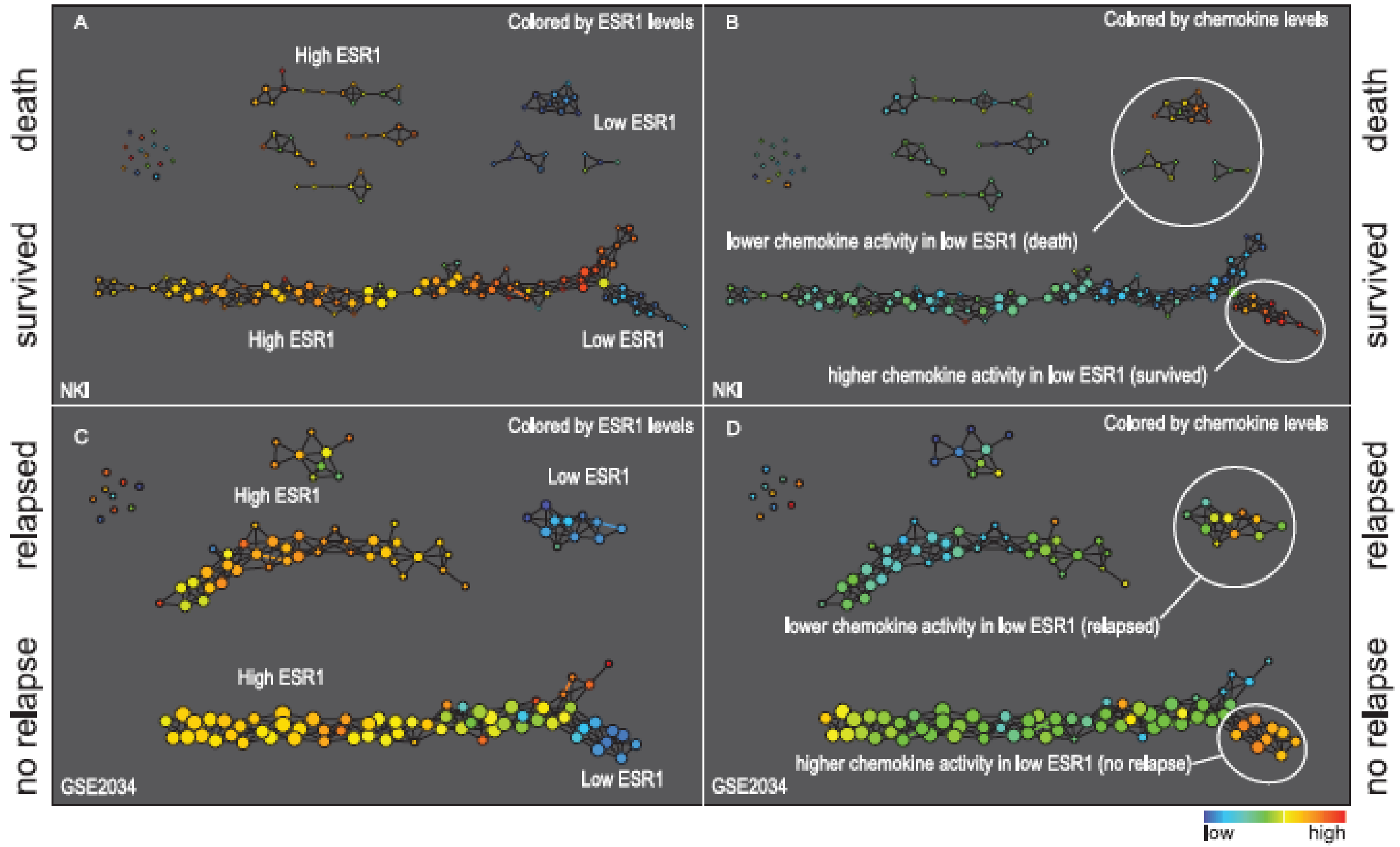
Mapper in applications

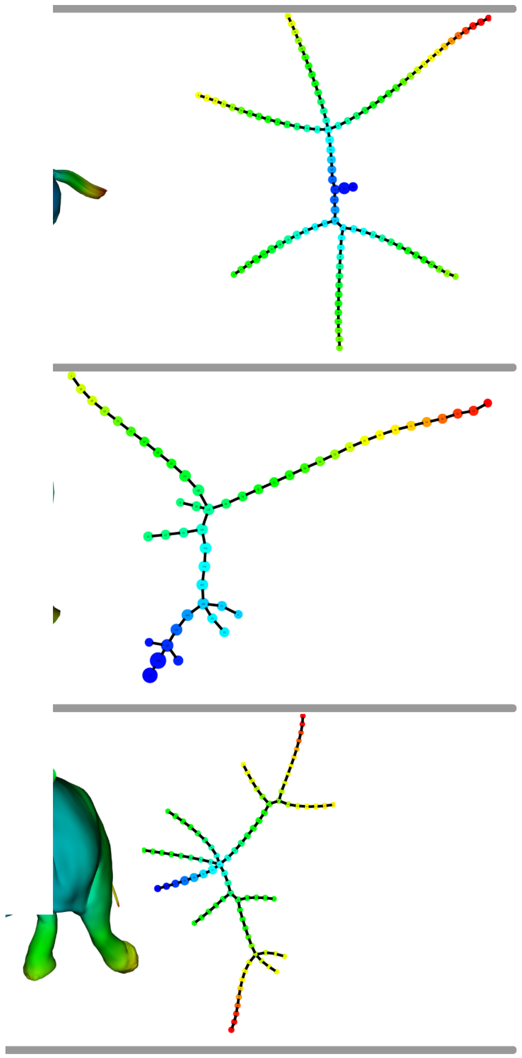


3d shapes classification

[Singh, Mémoli, Carlsson 2007]

Mapper in applications

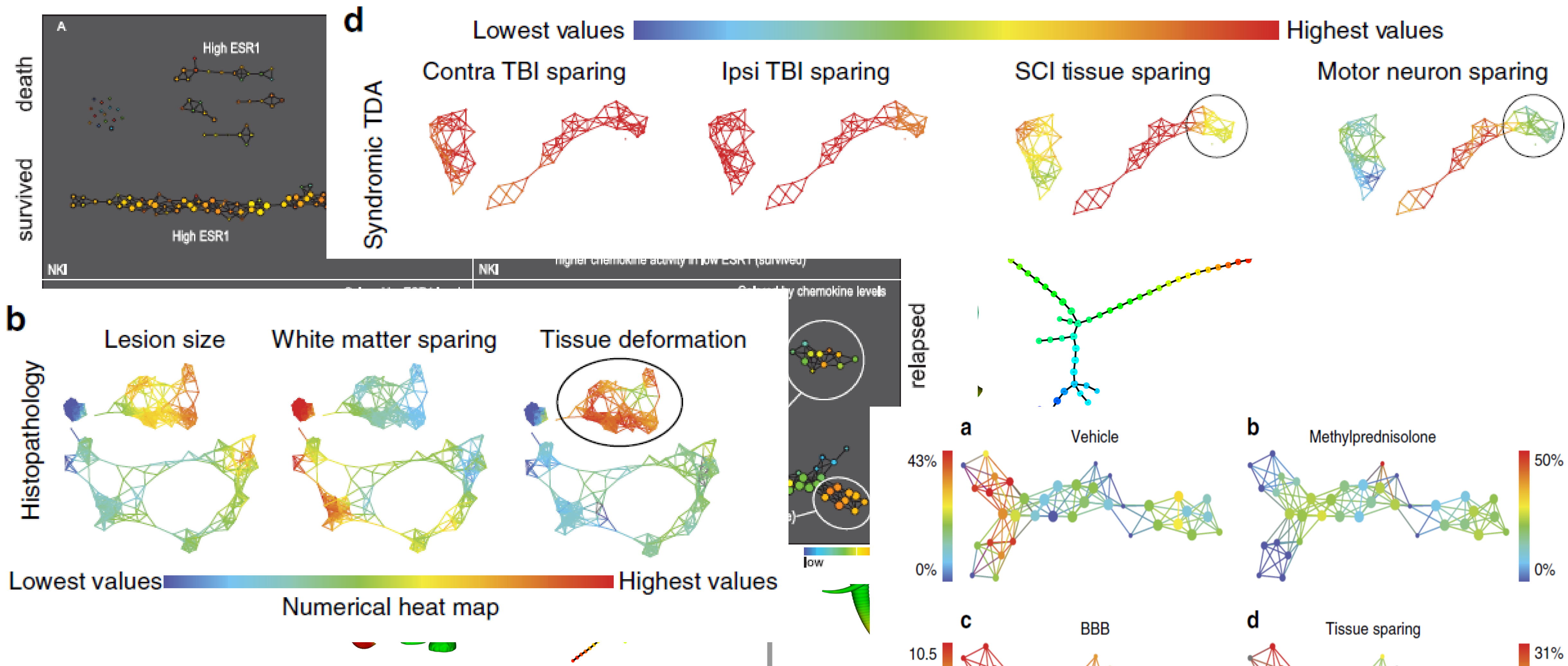




breast cancer subtype identification

[Nicolau et al. 2011]

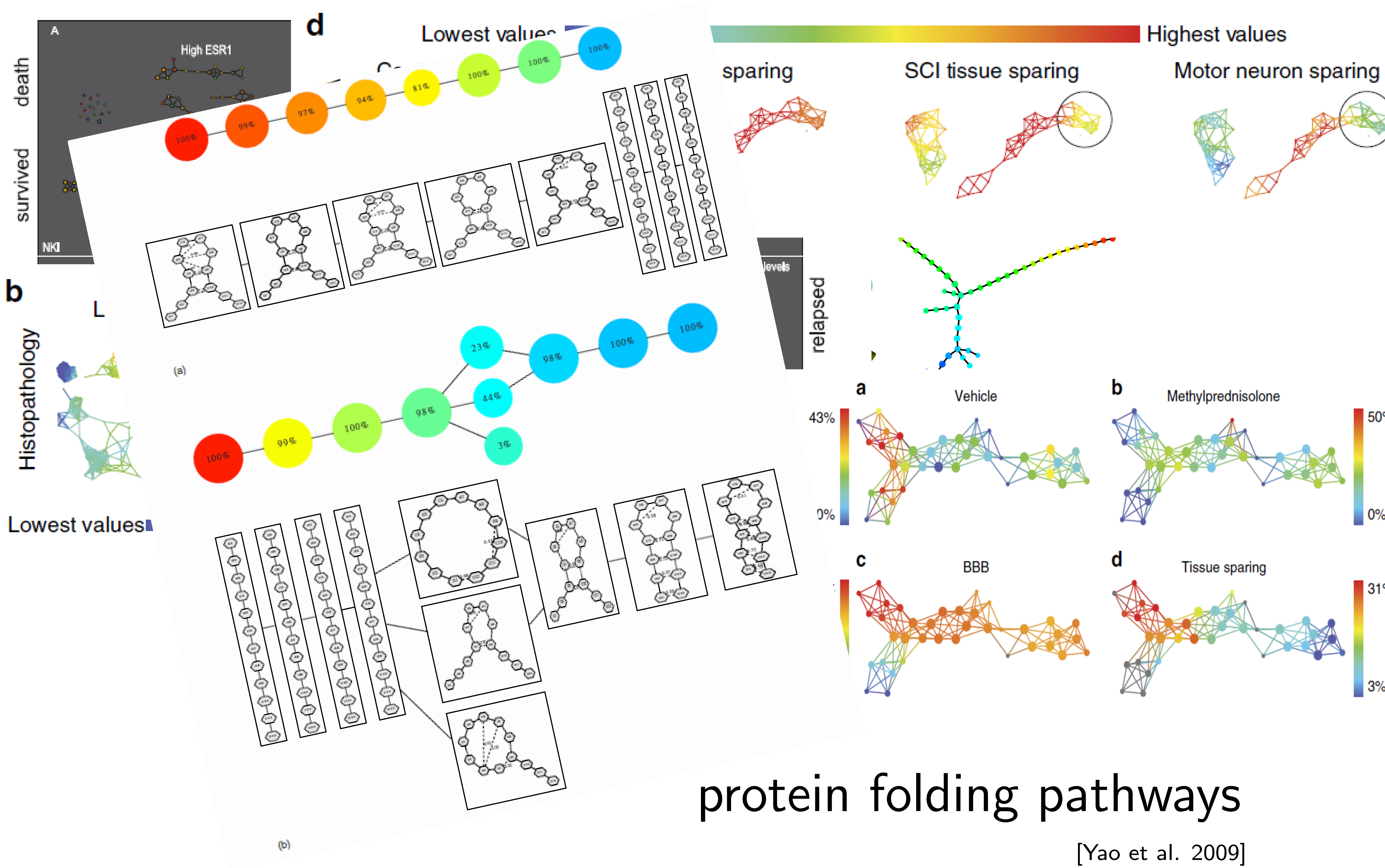
Mapper in applications



recovery from spinal cord injuries

[Nielson et al. 2015]

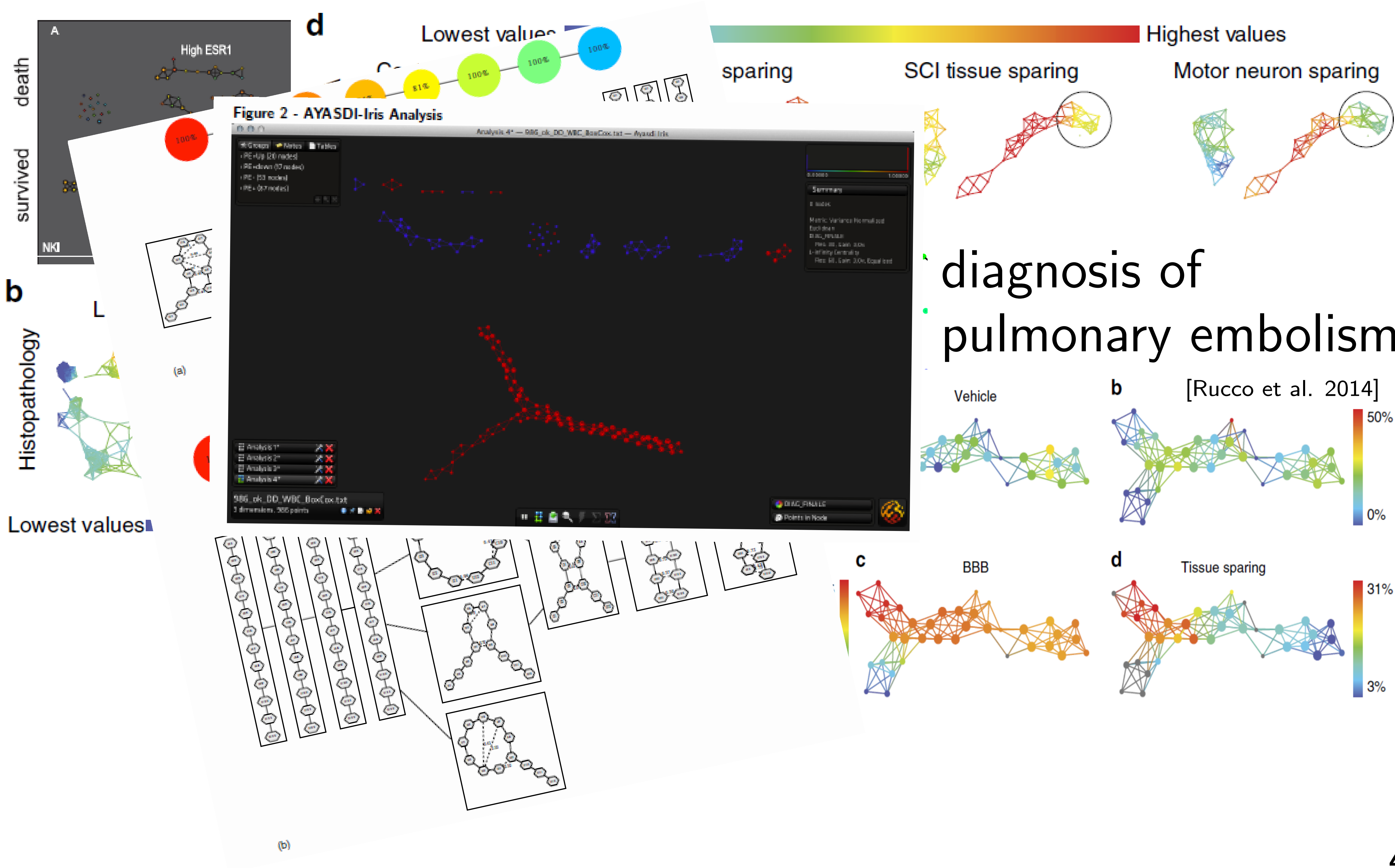
Mapper in applications



protein folding pathways

[Yao et al. 2009]

Mapper in applications



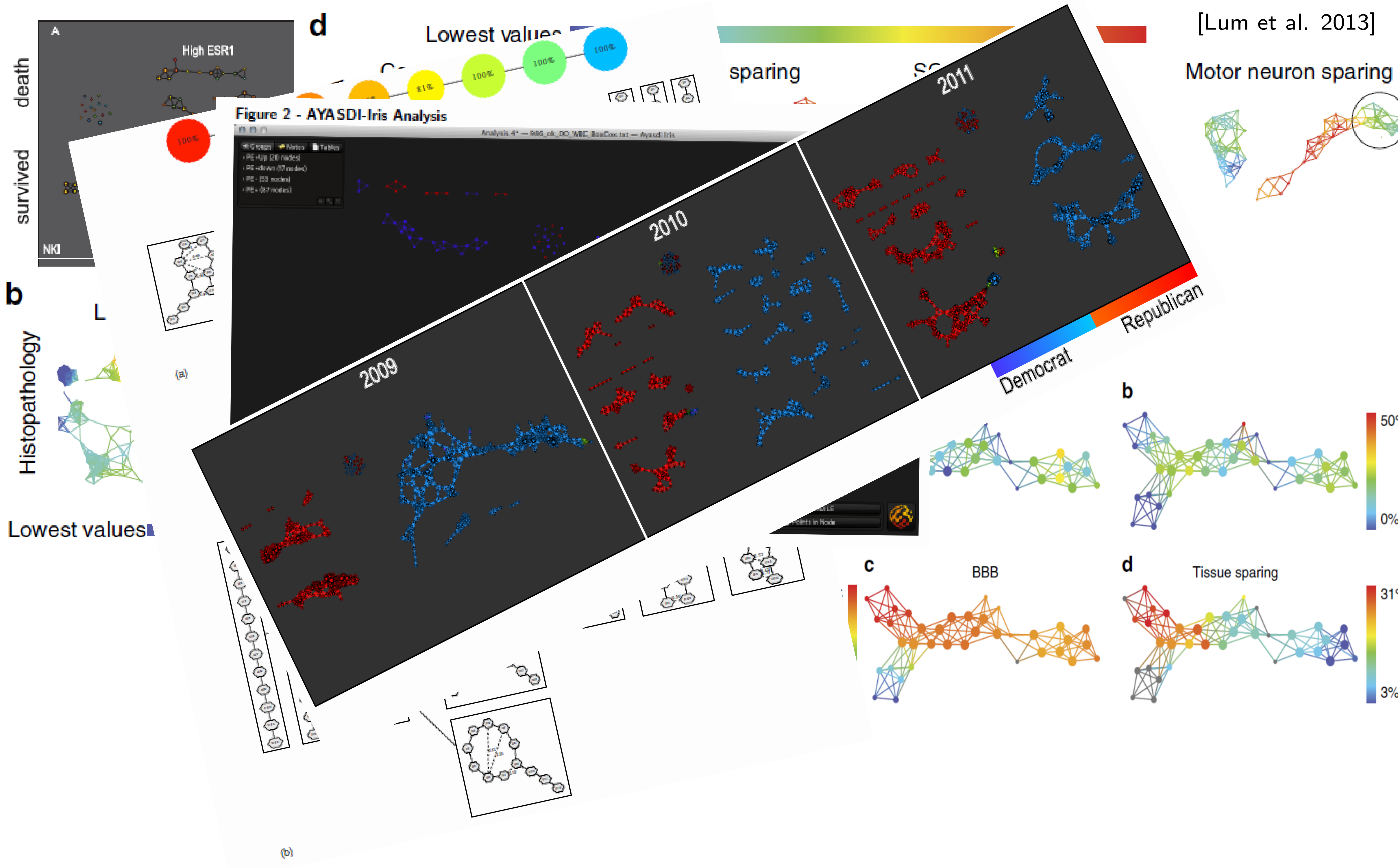
diagnosis of pulmonary embolism

[Rucco et al. 2014]

Mapper in applications

implicit networks in the US house of representatives

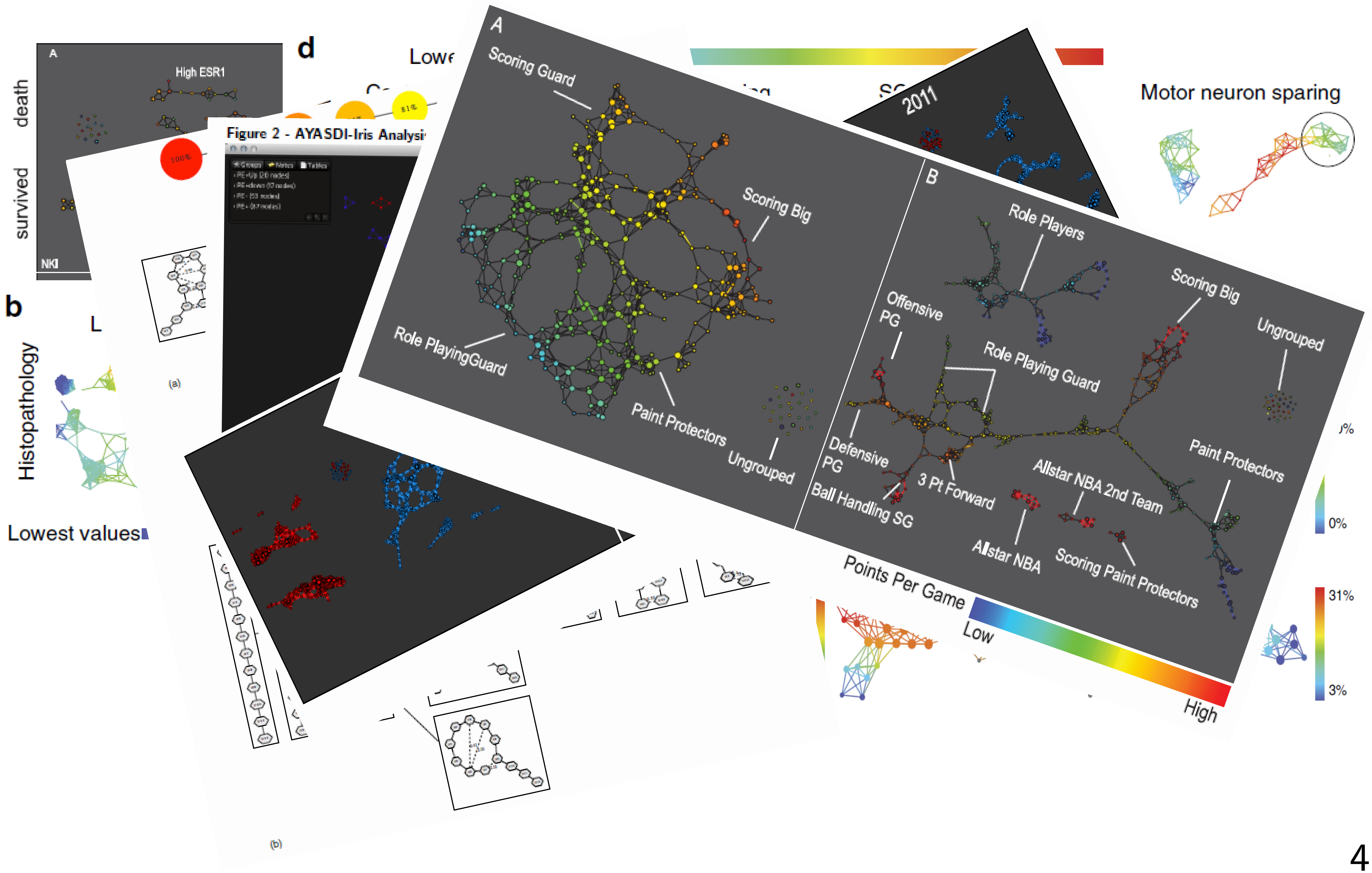
[Lum et al. 2013]



Mapper in applications

classification of NBA players

[Alagappan 2012]



Mapper in applications

Extracting insights from the shape of complex data using topology, Lum et al., Nature, 2013

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

Using Topological Data Analysis for Diagnosis Pulmonary Embolism, Rucco et al., arXiv preprint, 2014

Topological Methods for Exploring Low-density States in Biomolecular Folding Pathways, Yao et al., J. Chemical Physics, 2009

CD8 T-cell reactivity to islet antigens is unique to type 1 while CD4 T-cell reactivity exists in both type 1 and type 2 diabetes, Sarikonda et al., J. Autoimmunity, 2013

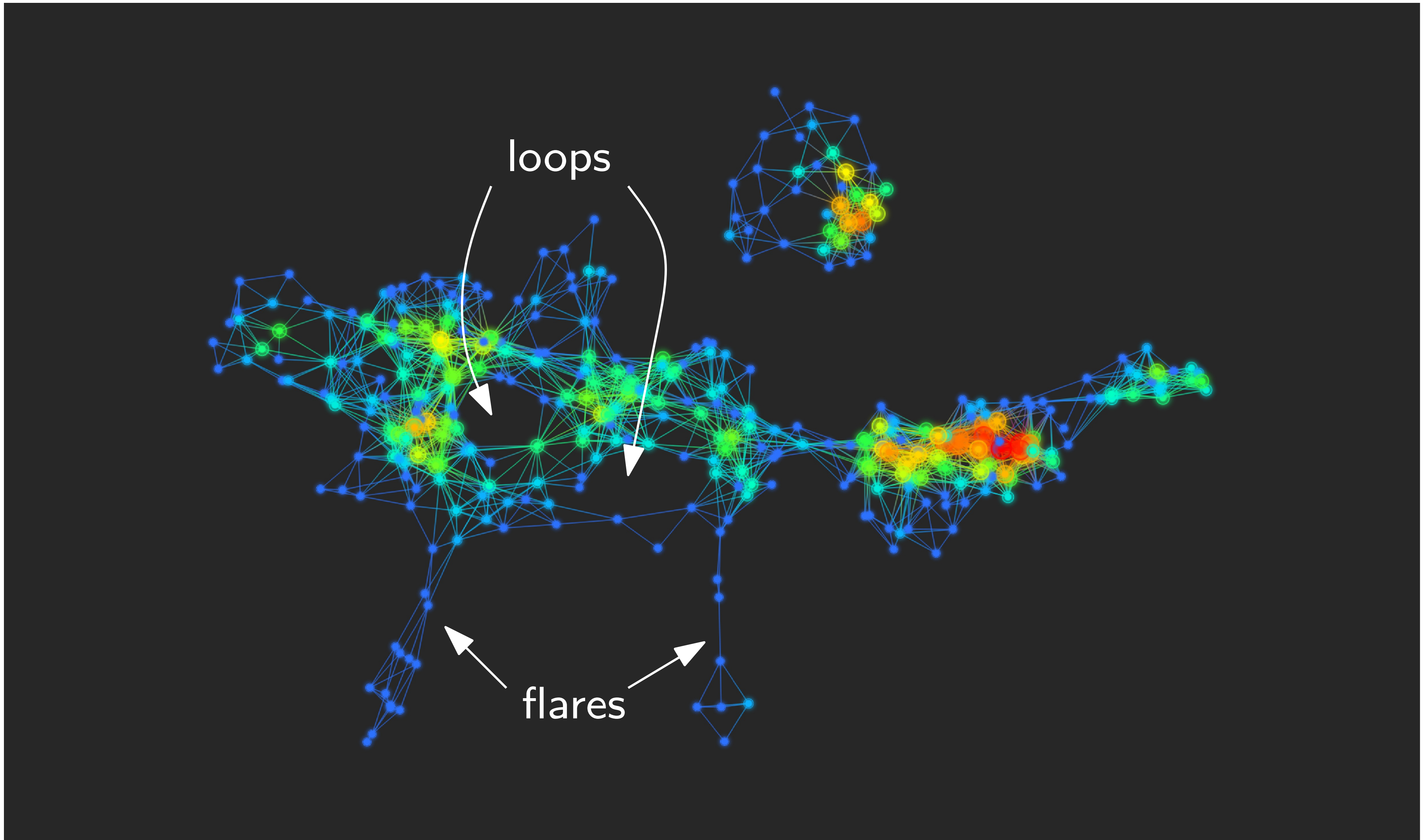
Innate and adaptive T cells in asthmatic patients: Relationship to severity and disease mechanisms, Hinks et al., J. Allergy Clinical Immunology, 2015

Mapper in applications

Two types of applications:

- clustering
- feature selection

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



Mapper in applications

1. clustering

Scheme:

compute the Mapper of your data

detect topological patterns ("loops", "flares") / subpopulations

use subpopulations to cluster data

Mapper in applications

1. clustering

Scheme:

compute the Mapper of your data

$\rightarrow$ selection of parameters

detect topological patterns ("loops", "flares") / subpopulations

$\rightarrow$ done by hand in general

$\rightarrow$ [Lum et al. 13] use persistence of eccentricity on Mapper graph

use subpopulations to cluster data

$\rightarrow$ visualize various features on the Mapper, check subpopulations for having the same feature level

$\rightarrow$ [Lum et al. 13] also use Monte-Carlo simulations with multivariate Gaussian distributions to validate the presence of flares

Extracting insights from the shape of complex data using topology, Lum et al., Nature, 2013

Goal: detect clusters in the US House of Representatives

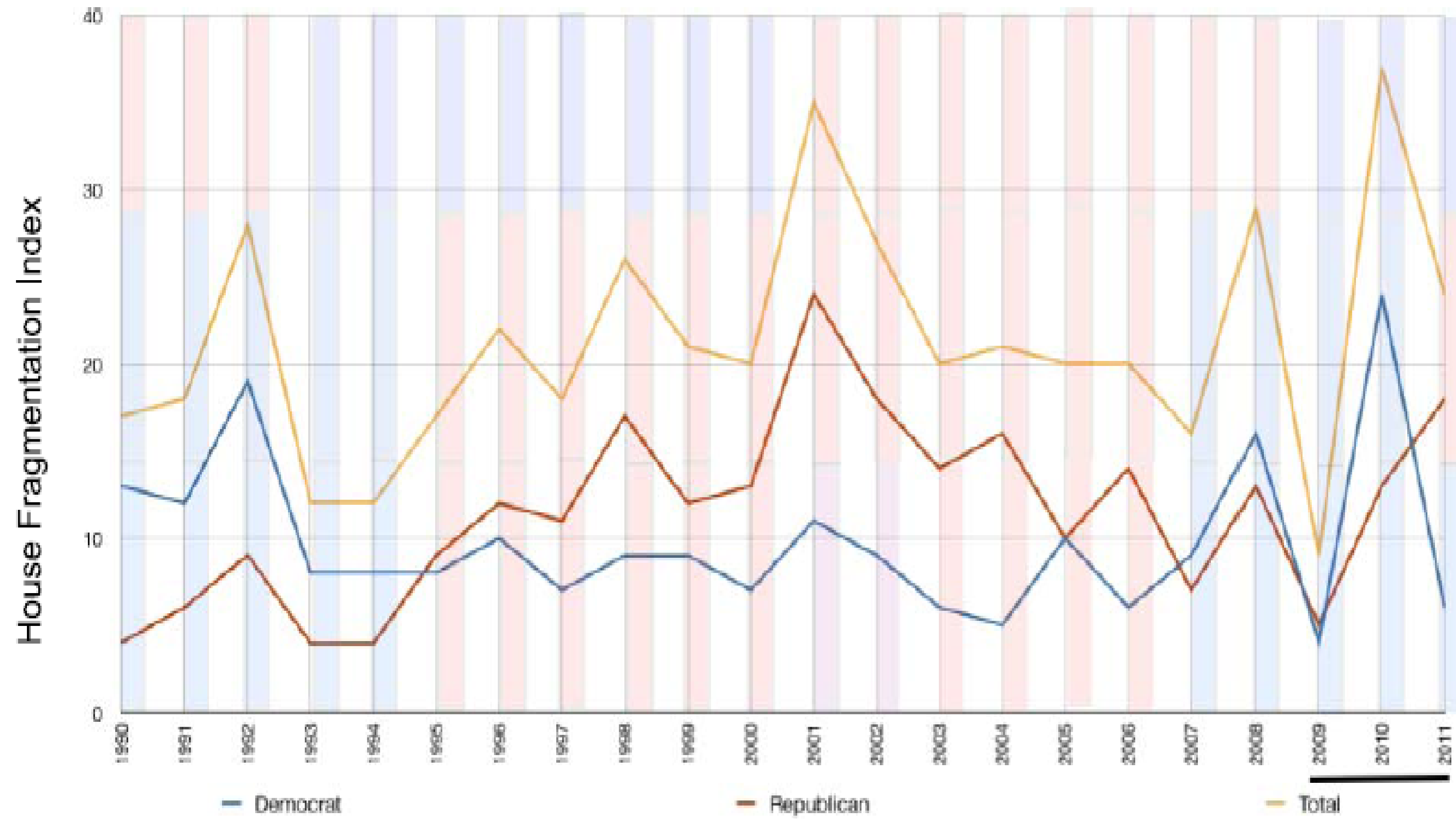
Points: member of the House

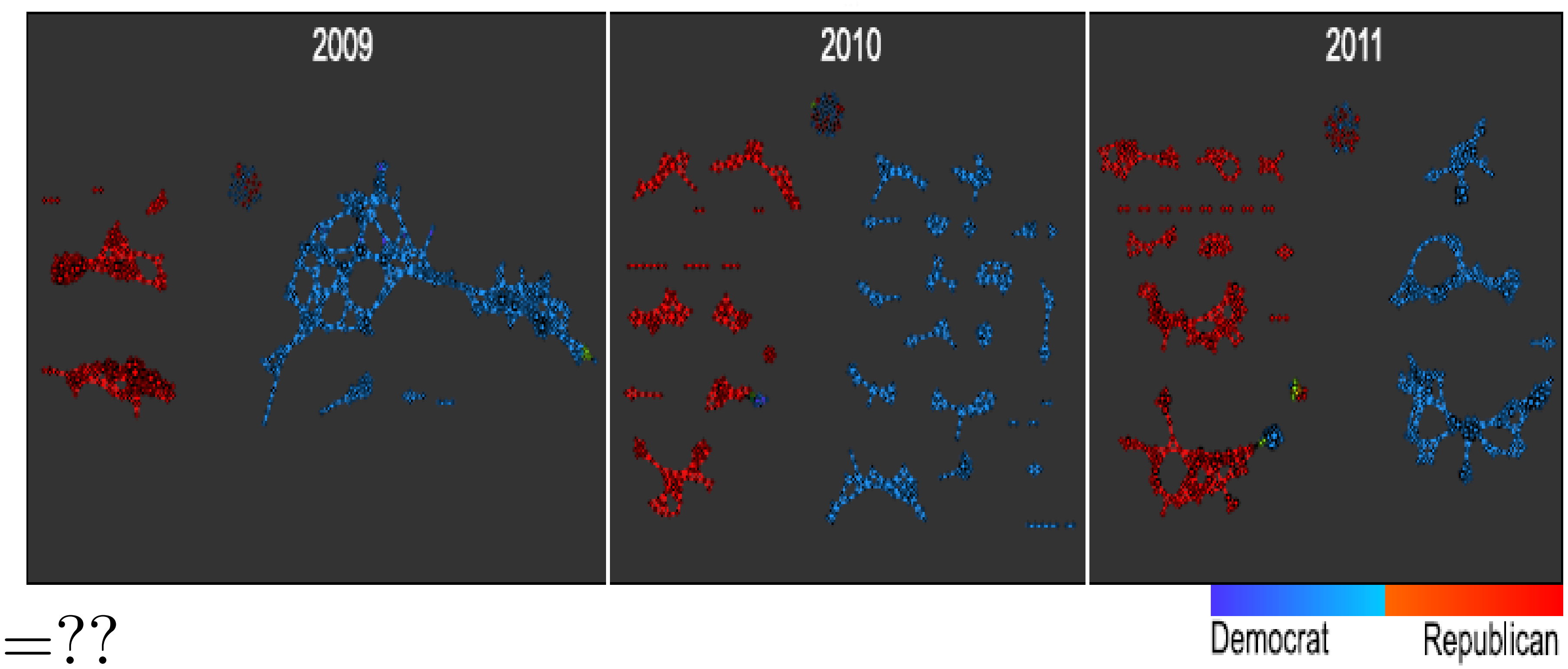
Filters: 1st and 2nd eigenvectors of the SVD of the coordinate matrix

Mapper colored by Republican/Democrat

Number of clusters for each political party through the years

PCA was only able to show the Republican/Democrat divide

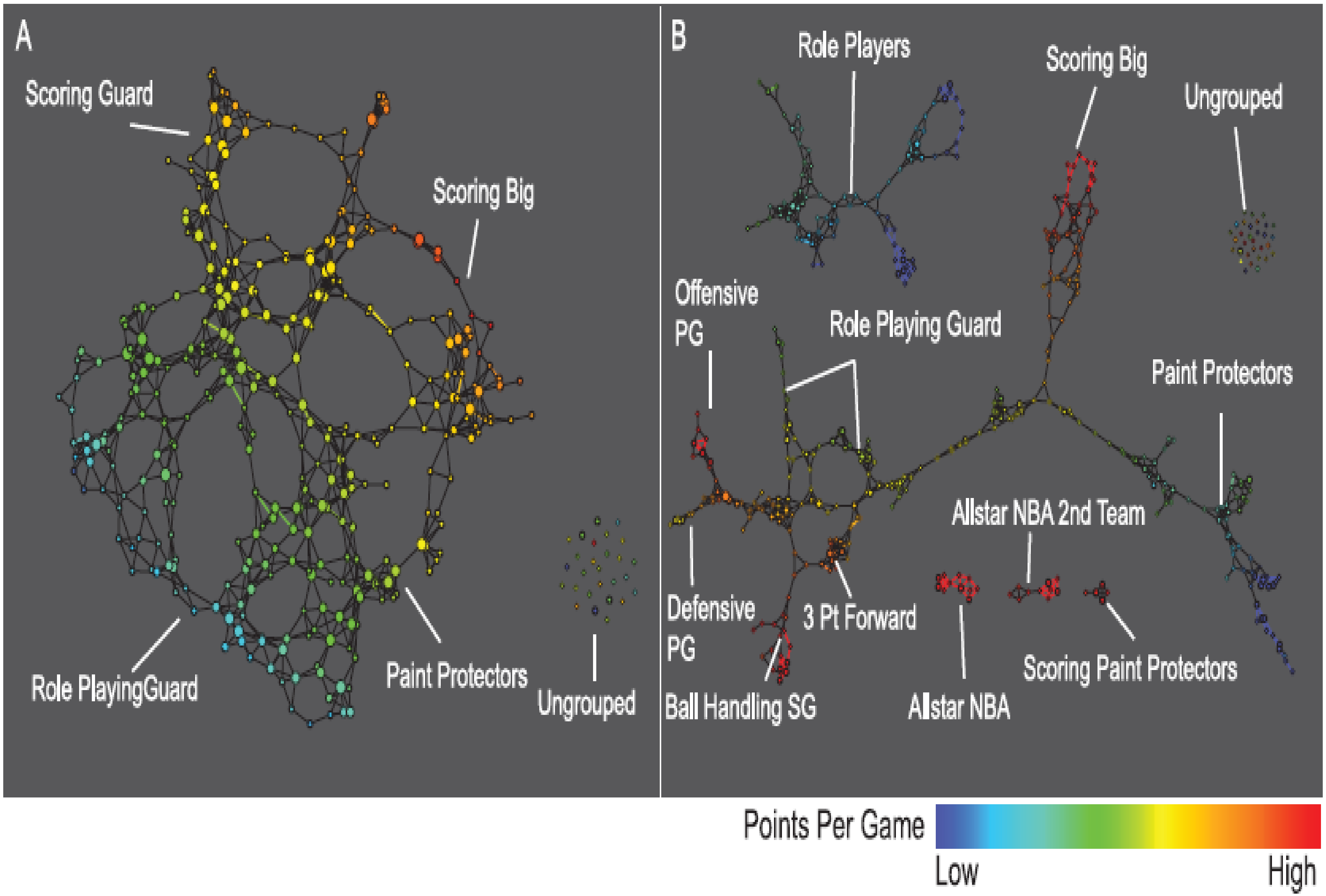




f: 1st and 2nd ev

$r = 1/120, g = 22\%, k = ??$

Same scheme: detect new clusters for NBA players (same paper)



Mapper in applications

2. feature selection

Scheme:

compute the Mapper of your data

detect topological patterns ("loops", "flares")

select features that best discriminate the corresponding subpopulations

Mapper in applications

2. feature selection

Scheme:

compute the Mapper of your data

$\rightarrow$ selection of parameters

detect topological patterns ("loops", "flares")

$\rightarrow$ done mostly by hand

$\rightarrow$ [Lum et al. 13] use persistence of eccentricity on Mapper graph

select features that best discriminate the corresponding subpopulations

$\rightarrow$ use 2-sample tests (typically Kolmogorov-Smirnov) on feature(substructure) vs feature(whole data set), then select features with low p-value (best discriminate subpopulation)

Extracting insights from the shape of complex data using topology, Lum et al., Nature, 2013

Goal: detect factors that influence survival after therapy in breast cancer patients

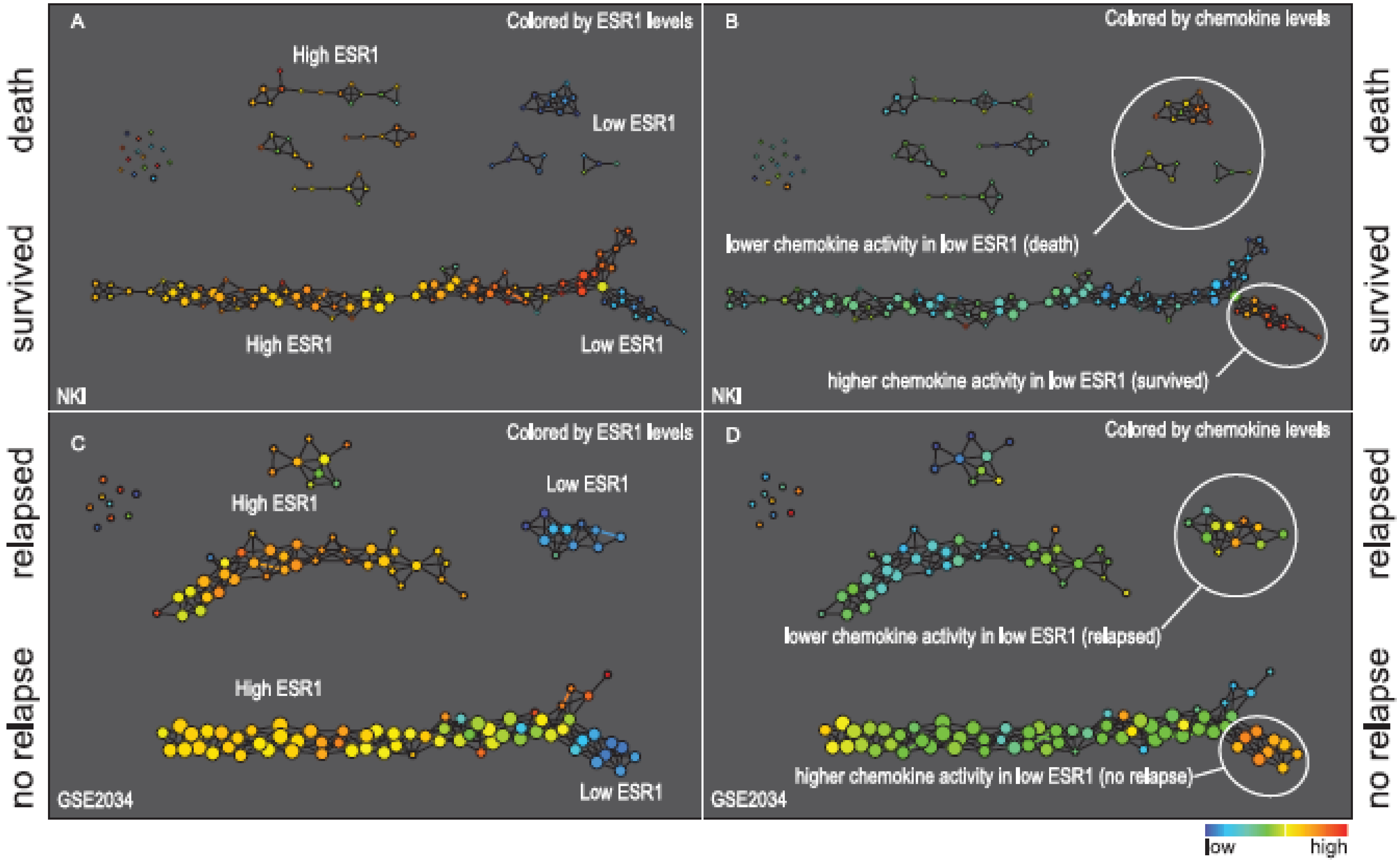
Points: breast cancer patients that went through specific therapy

Filters: eccentricity

Mapper colored by ESR1 level since it is understood that low-ESR1 groups are correlated to poor prognosis

f: eccentricity

$$r = 1/30, g = 33\%, k = ??$$



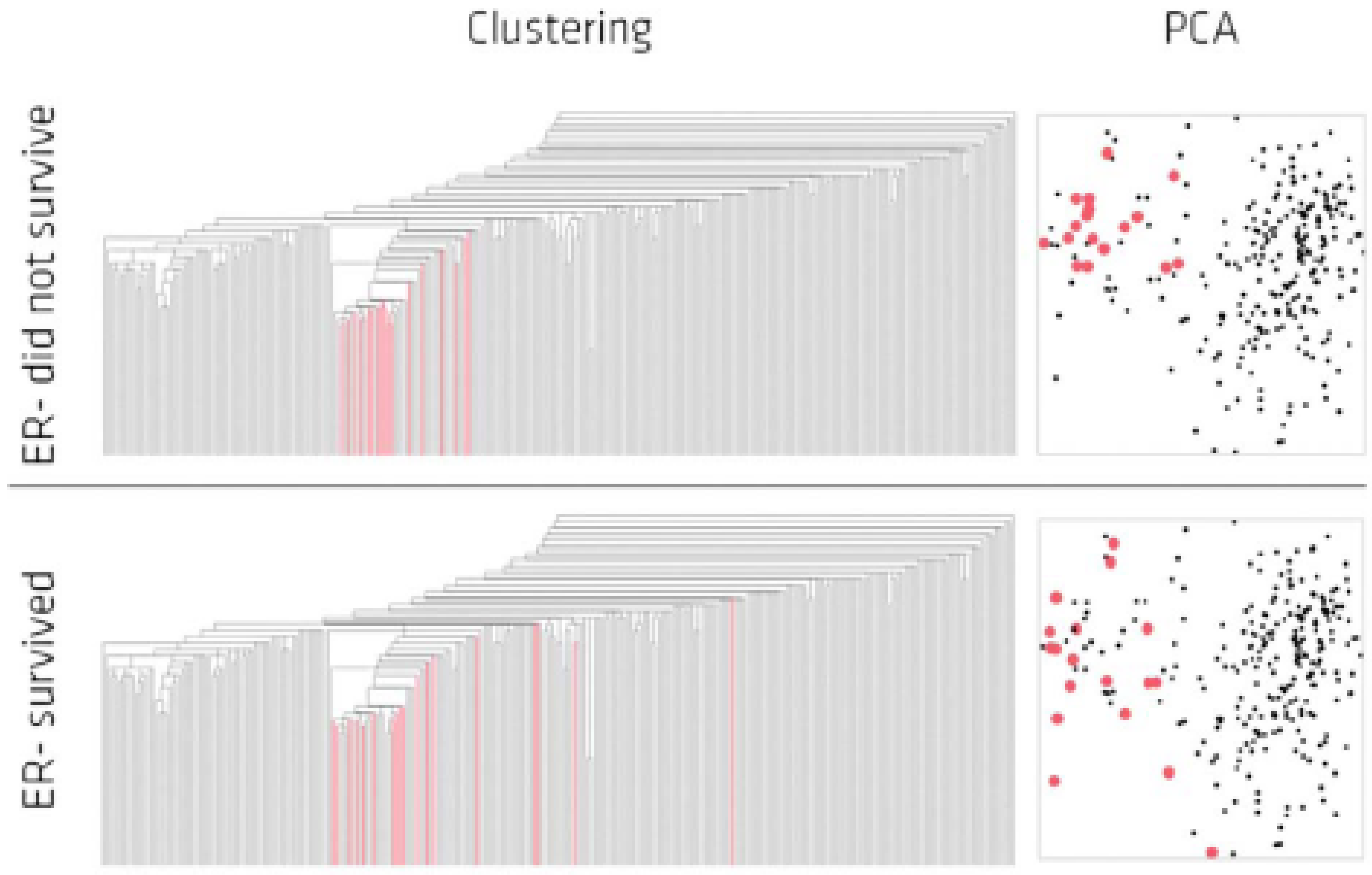
"Y" letter for survivors and ccs for non-survivors indicate structure

coloring with ESR1 level exhibits subcluster of survivors with low-ESR1 level (lower arm of the "Y")

genes with lowest p-value after KS test are the ones responsible for chemokine

coloring with chemokine level confirms this

PCA/Single-linkage clustering cannot see this



Choice of parameters

Parameters:

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

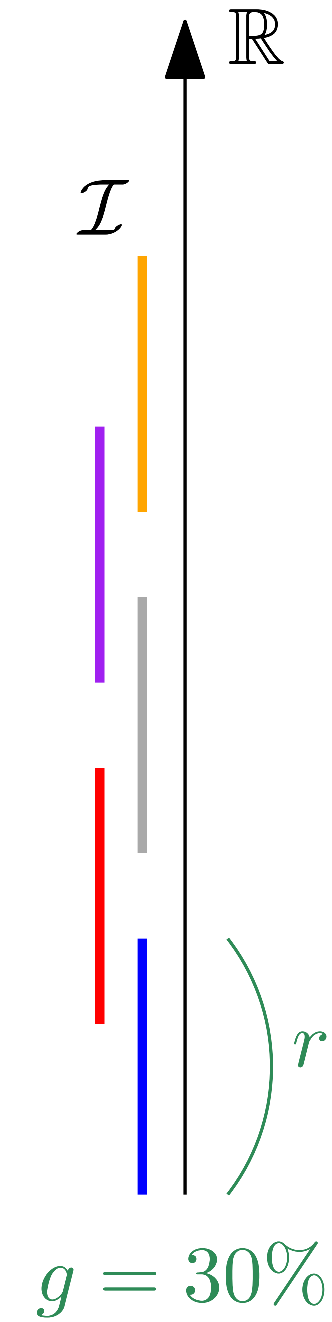
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals

- neighborhood size δ

range scale

geometric scale

Choice of parameters



Parameters:

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

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

- neighborhood size δ

geometric scale

range scale

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

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

Choice of parameters

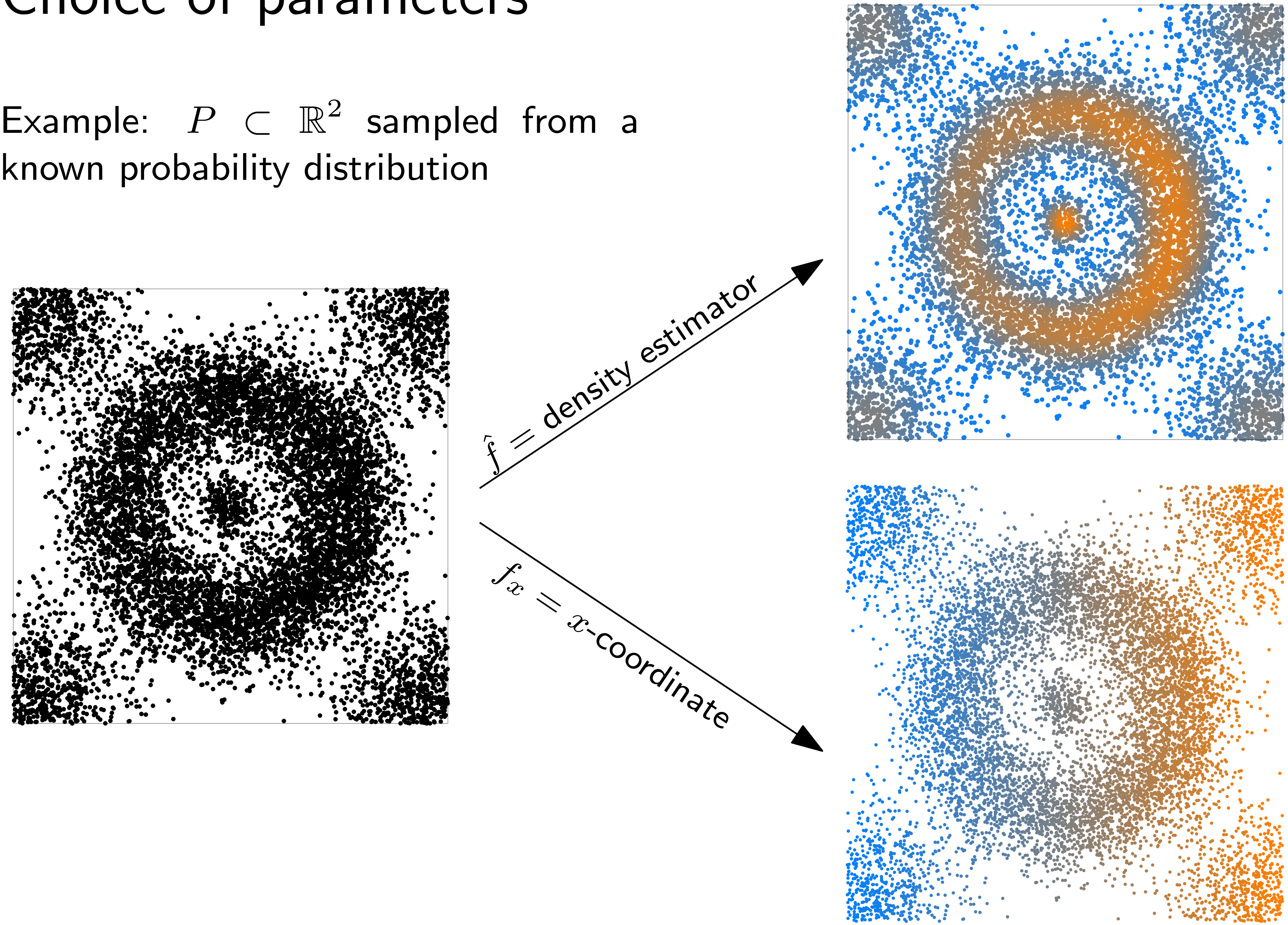
→ in practice: trial-and-error

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

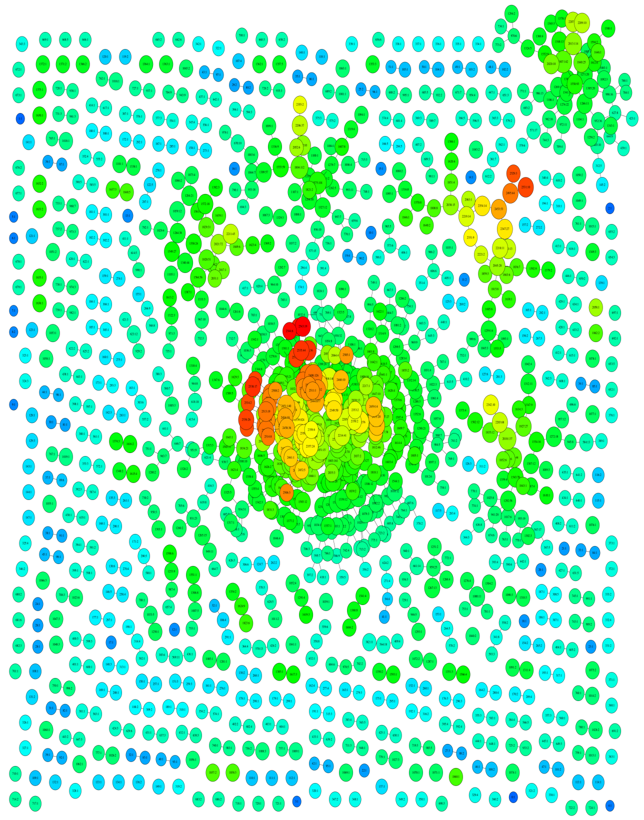
Example: $P \subset \mathbb{R}^2$ sampled from a known probability distribution

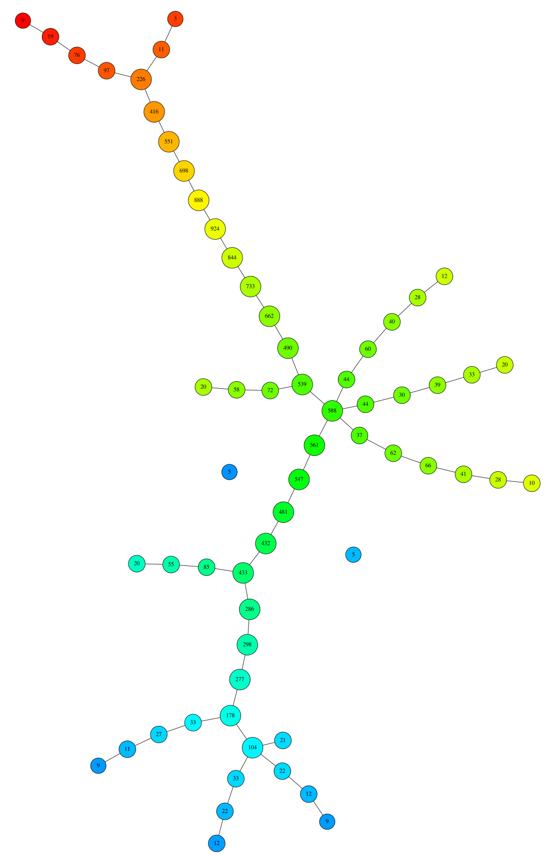


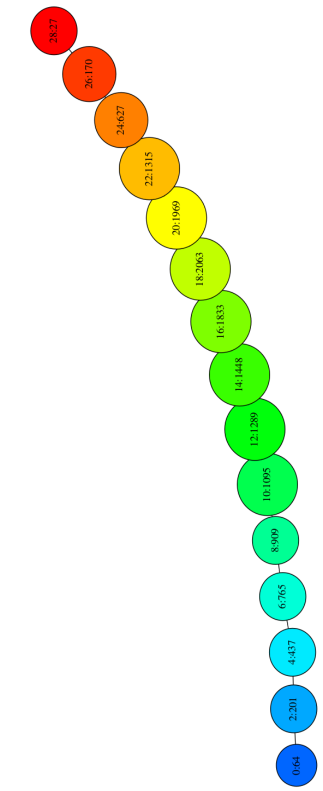
Choice of parameters

$r = 0.3,\ g = 20\%$

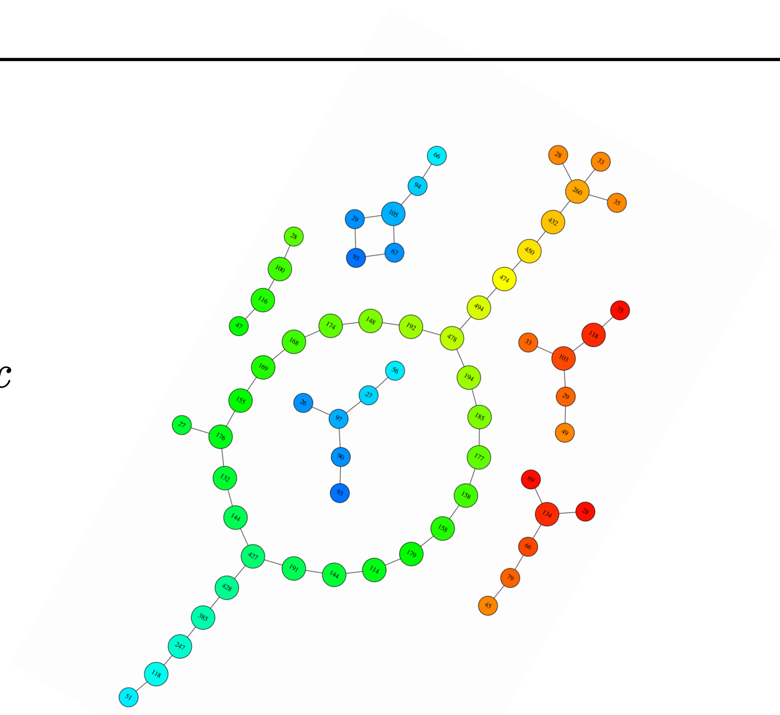
$f = \hat{f}$

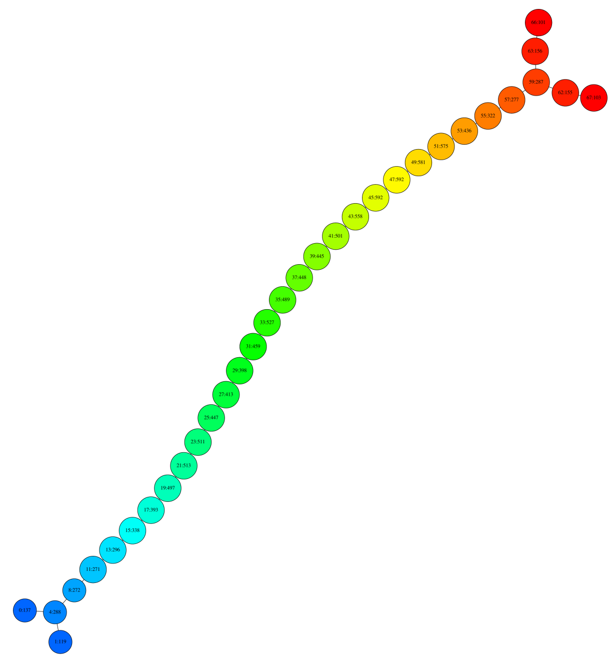


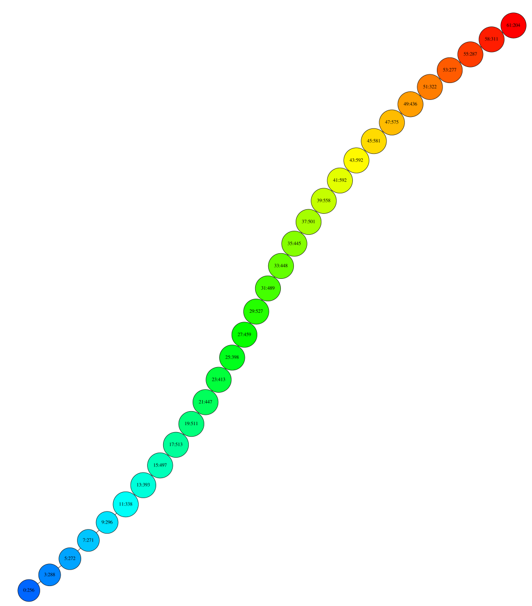




$f = f_x$



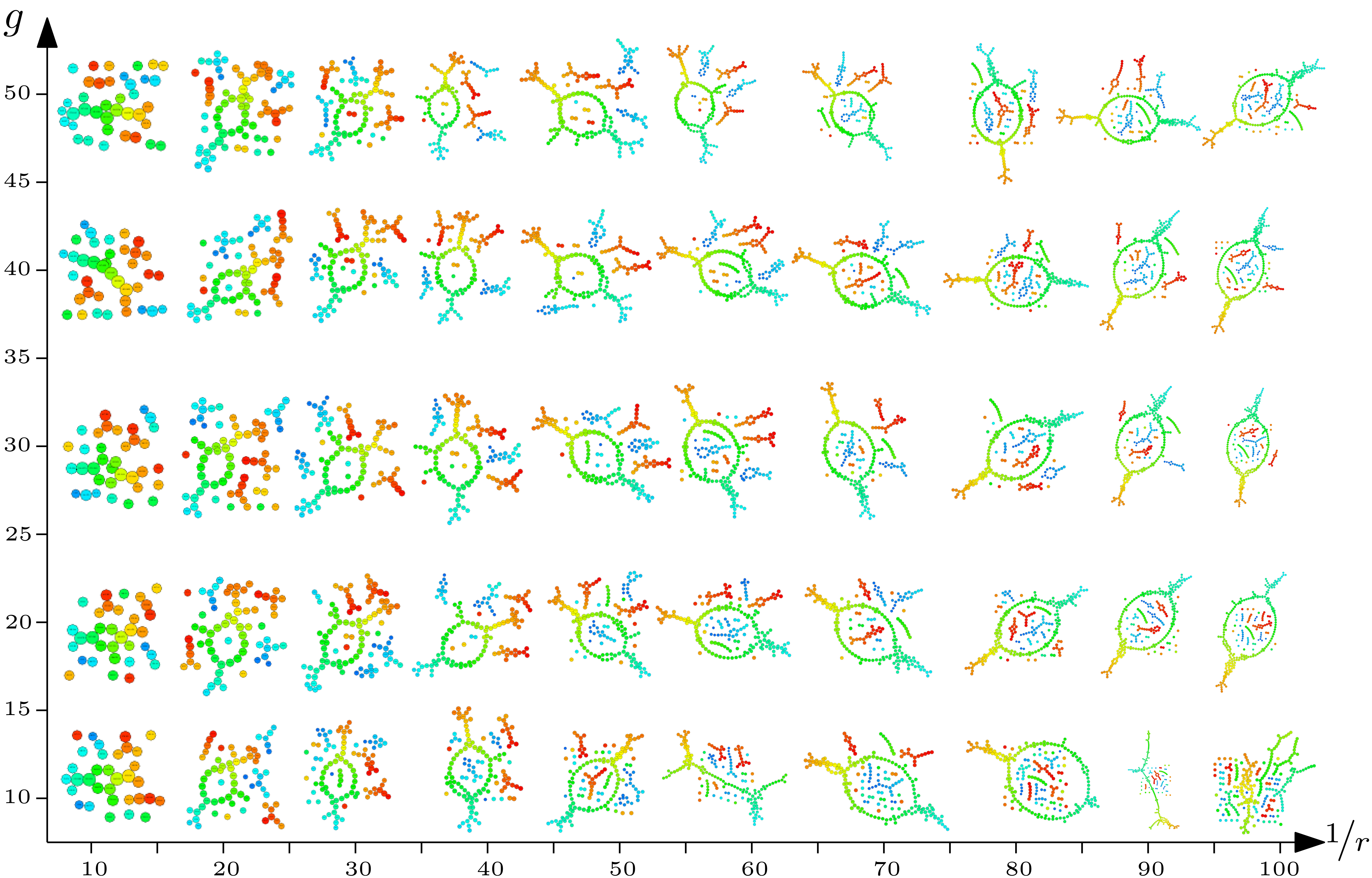




$\delta = 1\%$ $\delta = 10\%$ $\delta = 25\%$

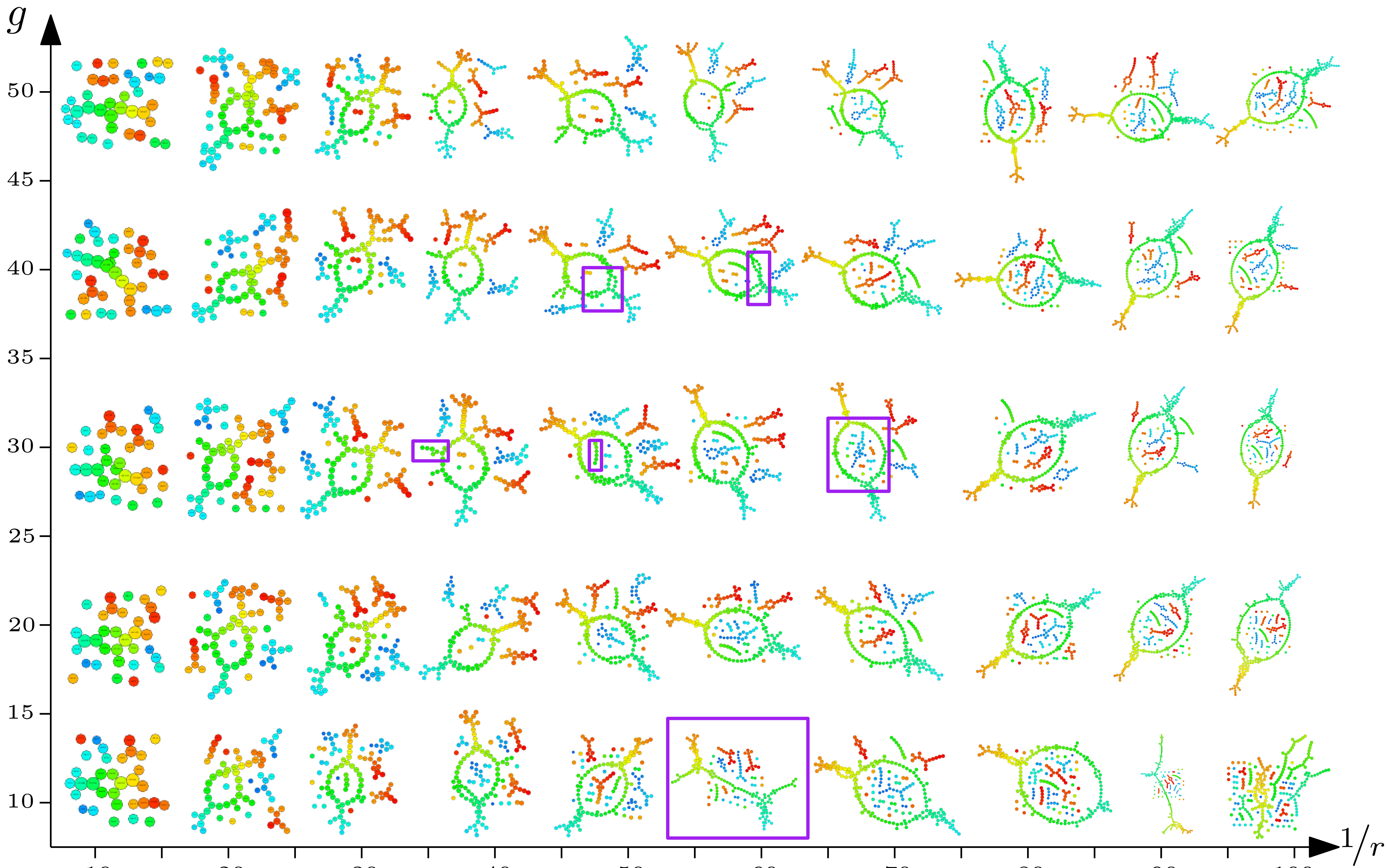
Choice of parameters

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



Choice of parameters

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



Choice of parameters

Recent contributions:

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

$\rightarrow$ introduce metrics between mappers

$\rightarrow$ establish stability and convergence results for Mappers

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

---

2 approaches:

• connection to topological persistence and representation theory
[Carrière, O. 2016] $<$ [Bauer, Ge, Wang 2013] [Cohen-Steiner, Edelsbrunner, Harer 2008]

• connection to constructible cosheaves in Sets and stratification theory
[Munch, Wang 2016] $<$ [de Silva, Munch, Patel 2015]

Choice of parameters

Recent contributions:

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

$\rightarrow$ introduce metrics between mappers

$\rightarrow$ establish stability and convergence results for Mappers

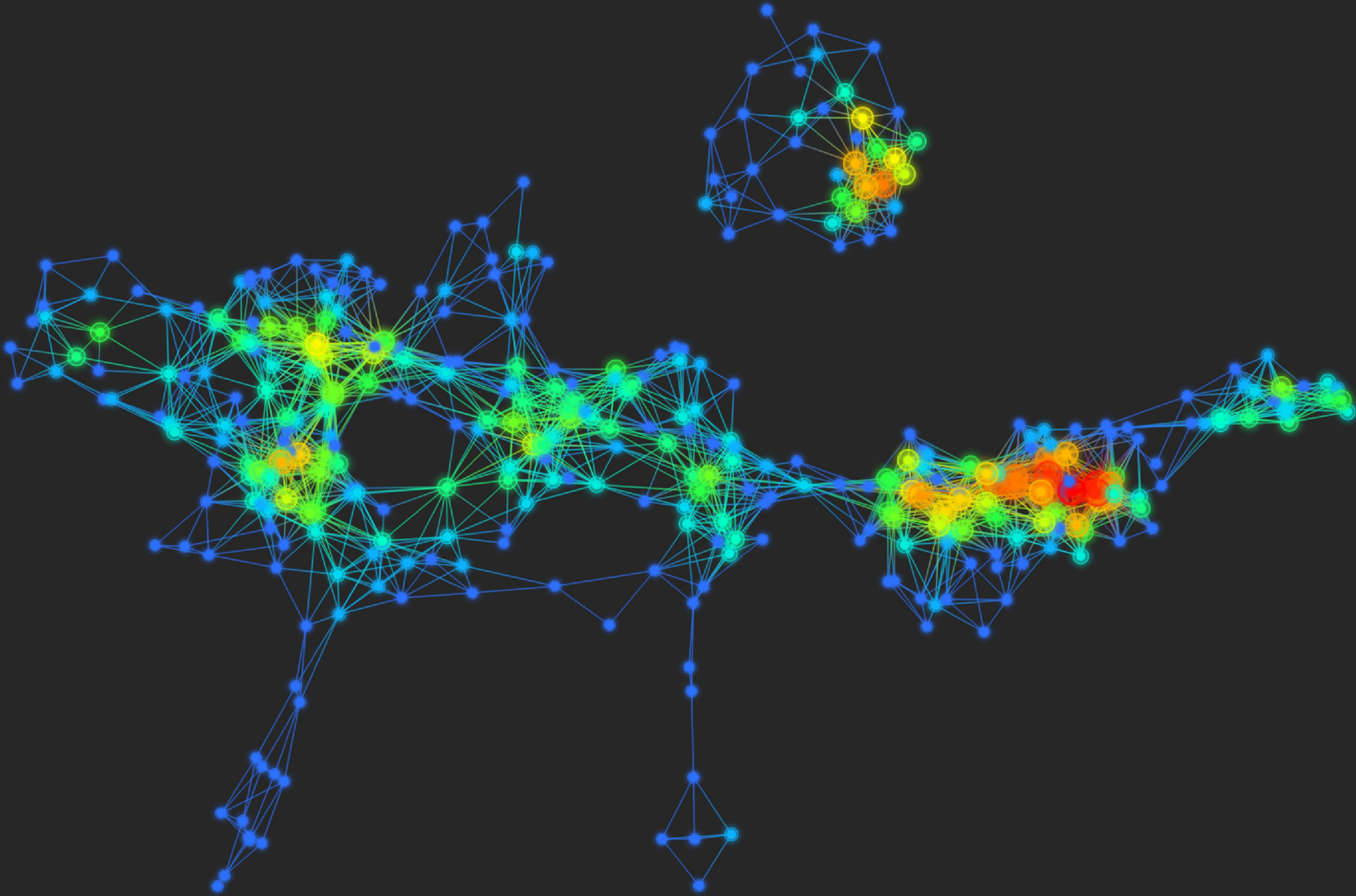
$\rightarrow$ relate discrete and continuous Mappers under conditions on δ

---

2 approaches:

• connection to topological persistence and representation theory
[Carrière, O. 2016] $<$ [Bauer, Ge, Wang 2013] [Cohen-Steiner, Edelsbrunner, Harer 2008]

• connection to constructible cosheaves in Sets and stratification theory
[Munch, Wang 2016] $<$ [de Silva, Munch, Patel 2015]

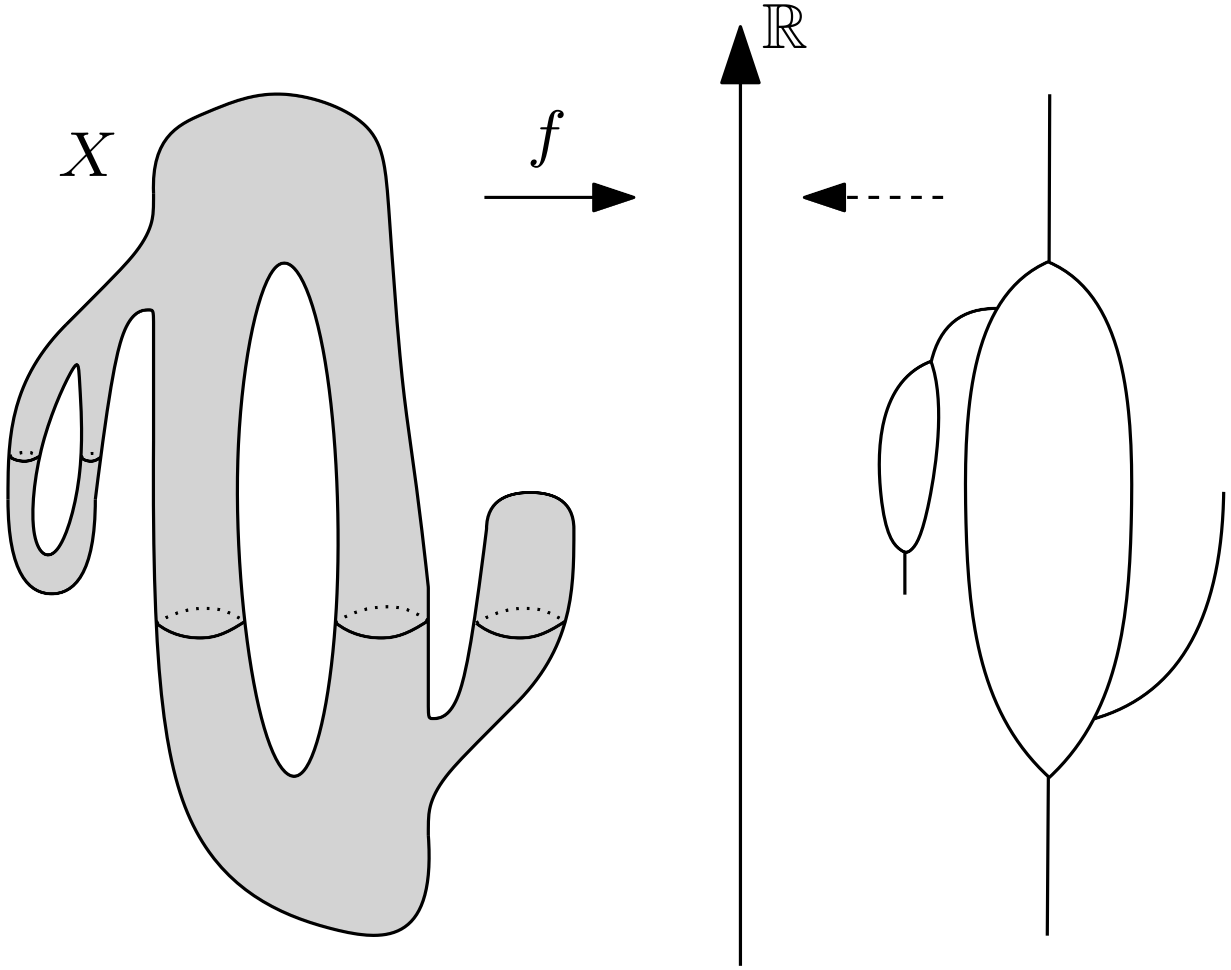


principle: summarize the topological structure of a map $f : X \to \mathbb{R}$ through a graph

Reeb Graph

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

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



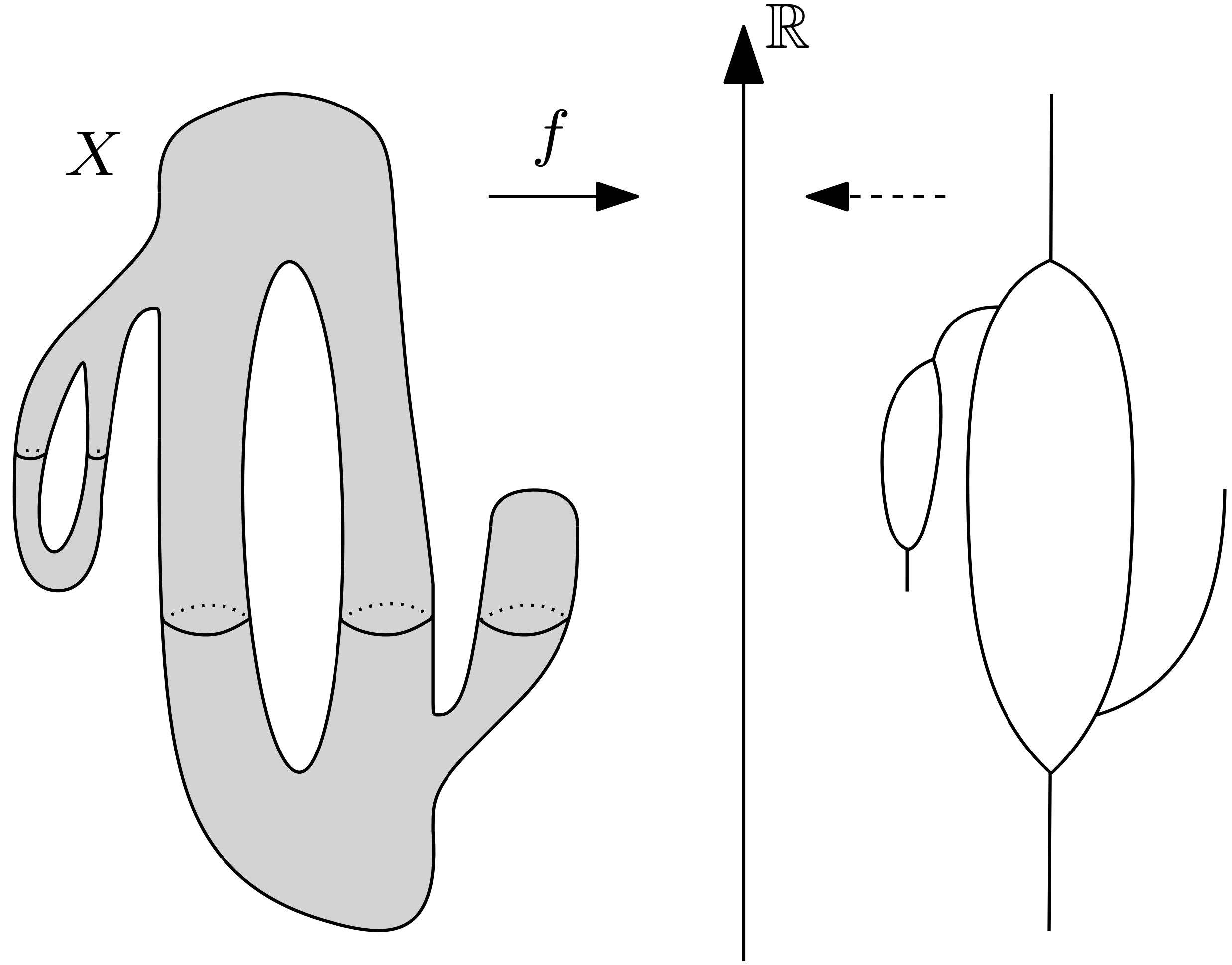
$$\begin{tikzcd} X \arrow[r, "f"] \arrow[d, "\sim"'] & \mathbb{R} \\ \mathrm{R}_f(X) \arrow[ur, dashed] & \end{tikzcd}$$

Reeb Graph

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

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

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

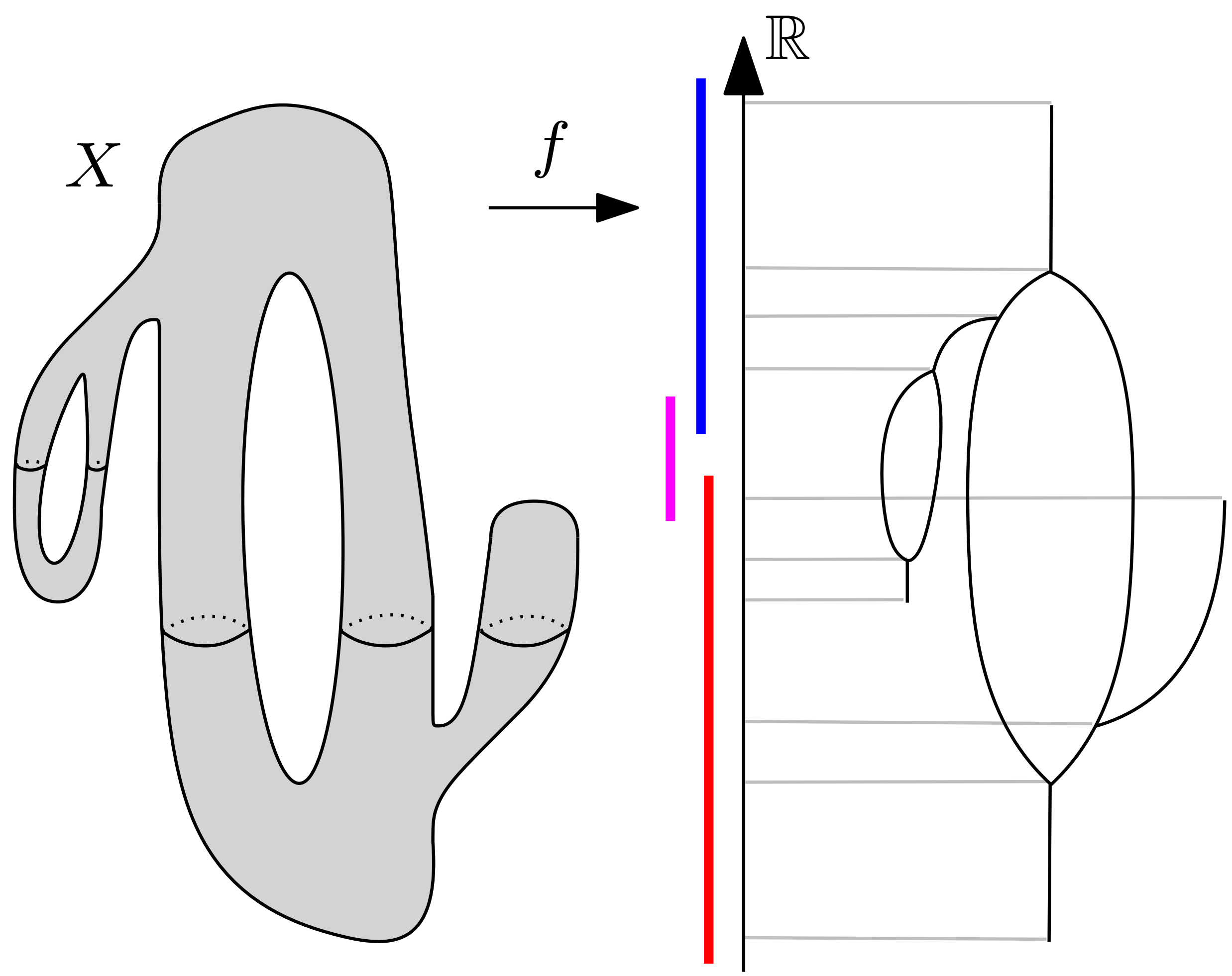


Prop: $\mathrm{R}_f(X)$ is a 1-d stratified space (*graph*) e.g. when (X, f) is Morse, or more generally of **Morse type**

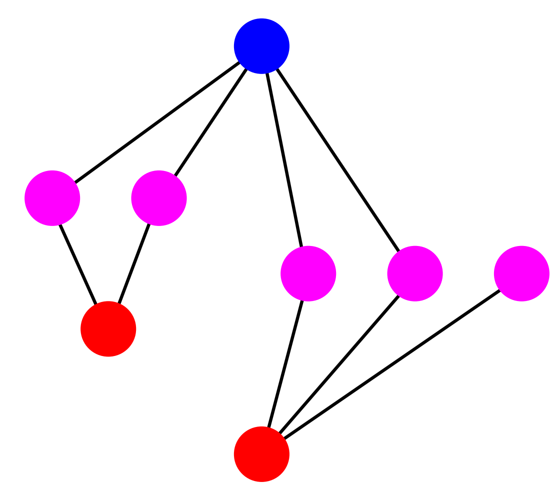
Reeb Graph

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

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



mapper $\equiv$ *pixelized* Reeb graph

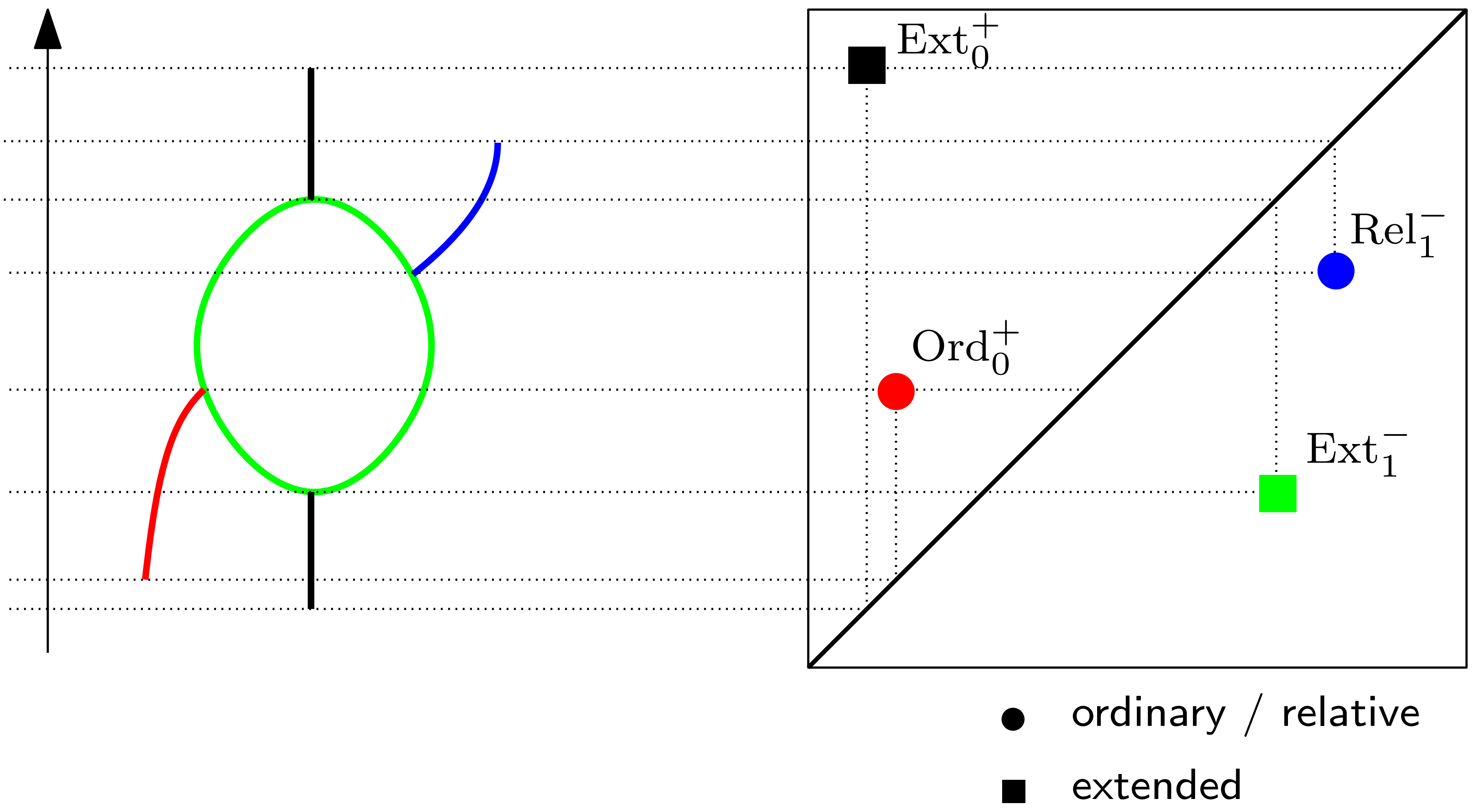


$\rightarrow$ build a **descriptor** for Reeb graphs

Descriptor for Reeb graph

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

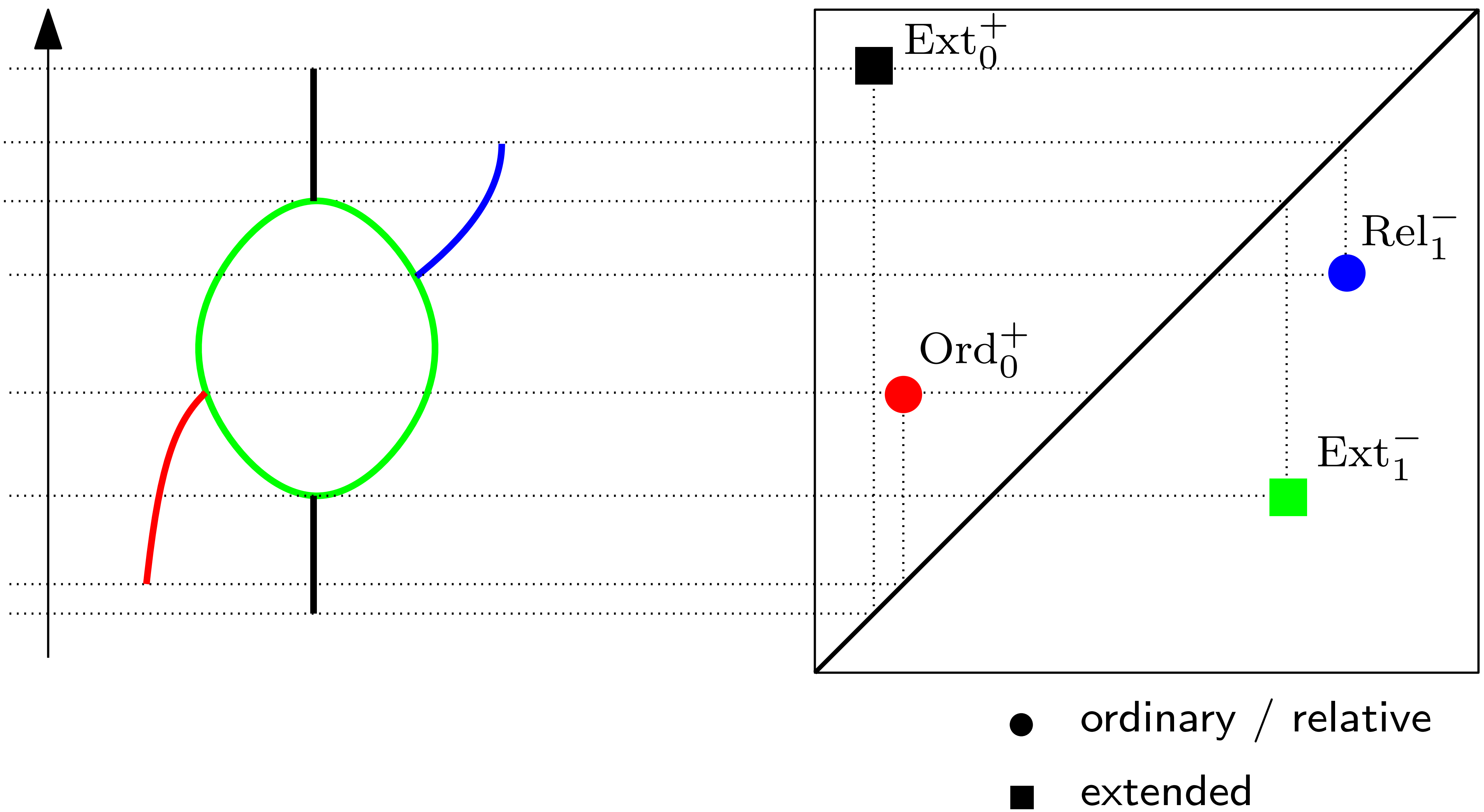
$\mathrm{Ord}_0\mathrm{R}_f \longleftrightarrow$ downward branches	$\mathrm{Ext}_0\mathrm{R}_f \longleftrightarrow$ trunks (cc)
$\mathrm{Rel}_1\mathrm{R}_f \longleftrightarrow$ upward branches	$\mathrm{Ext}_1\mathrm{R}_f \longleftrightarrow$ loops



Descriptor for Reeb graph

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

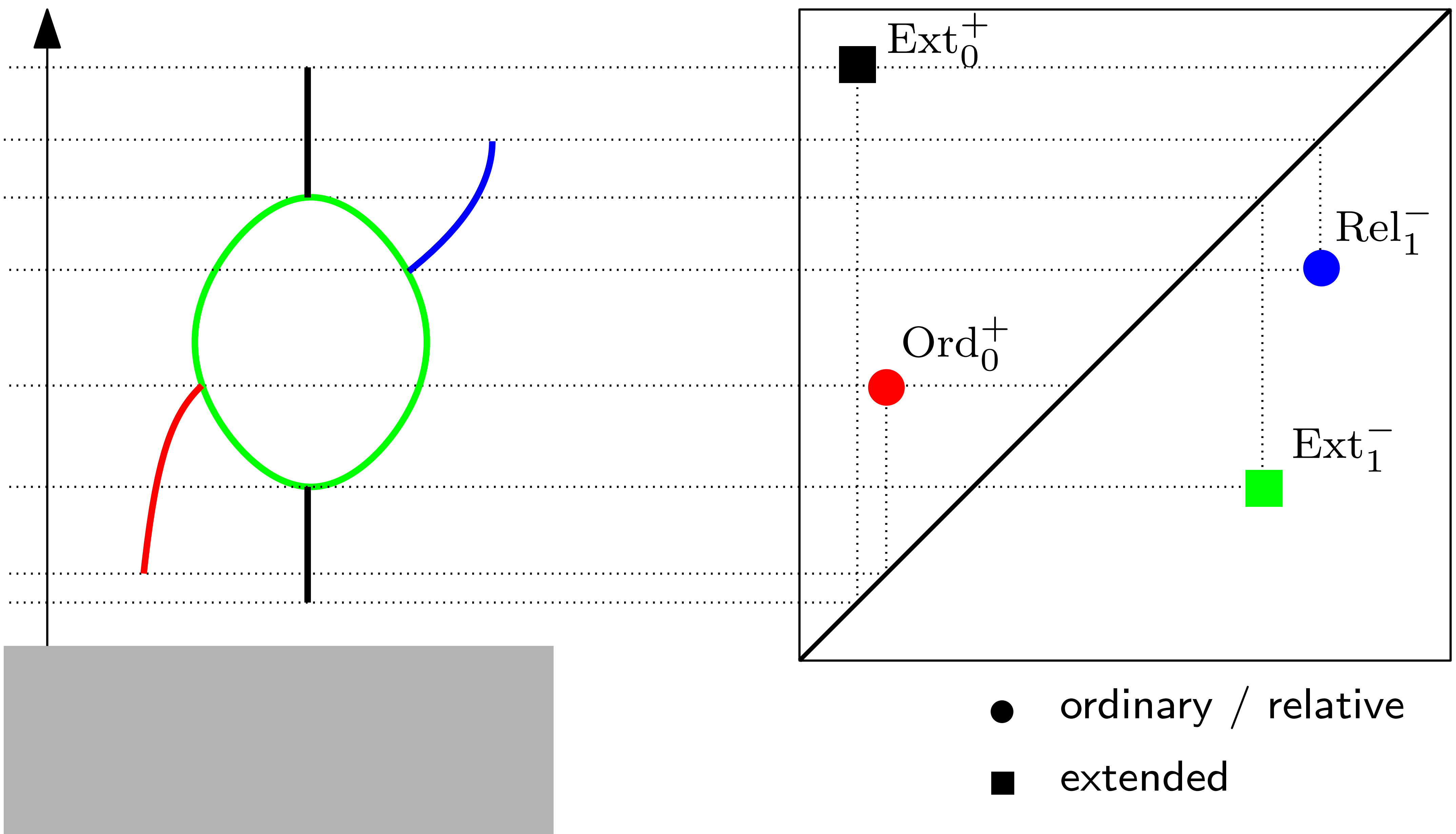
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

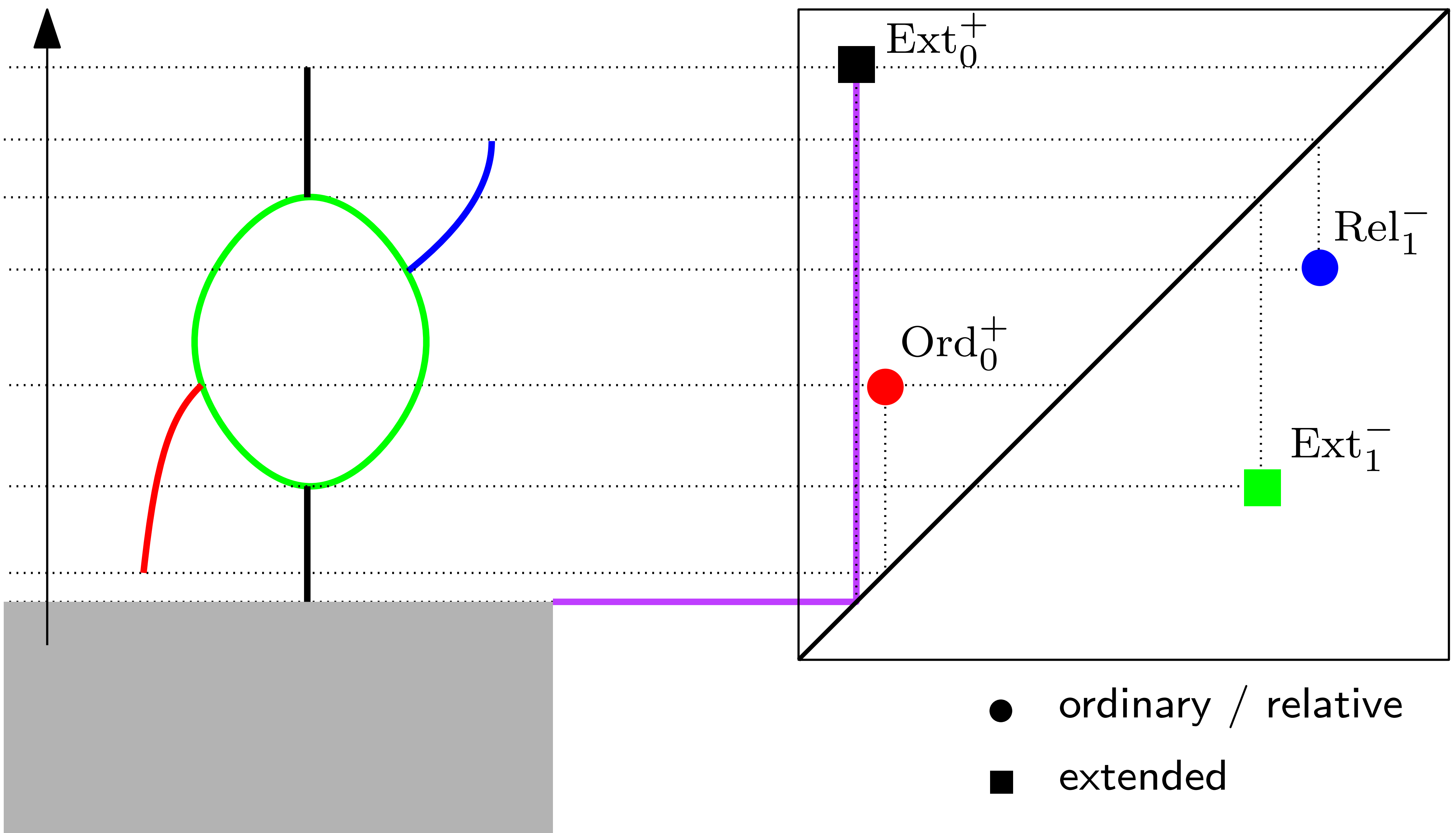
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

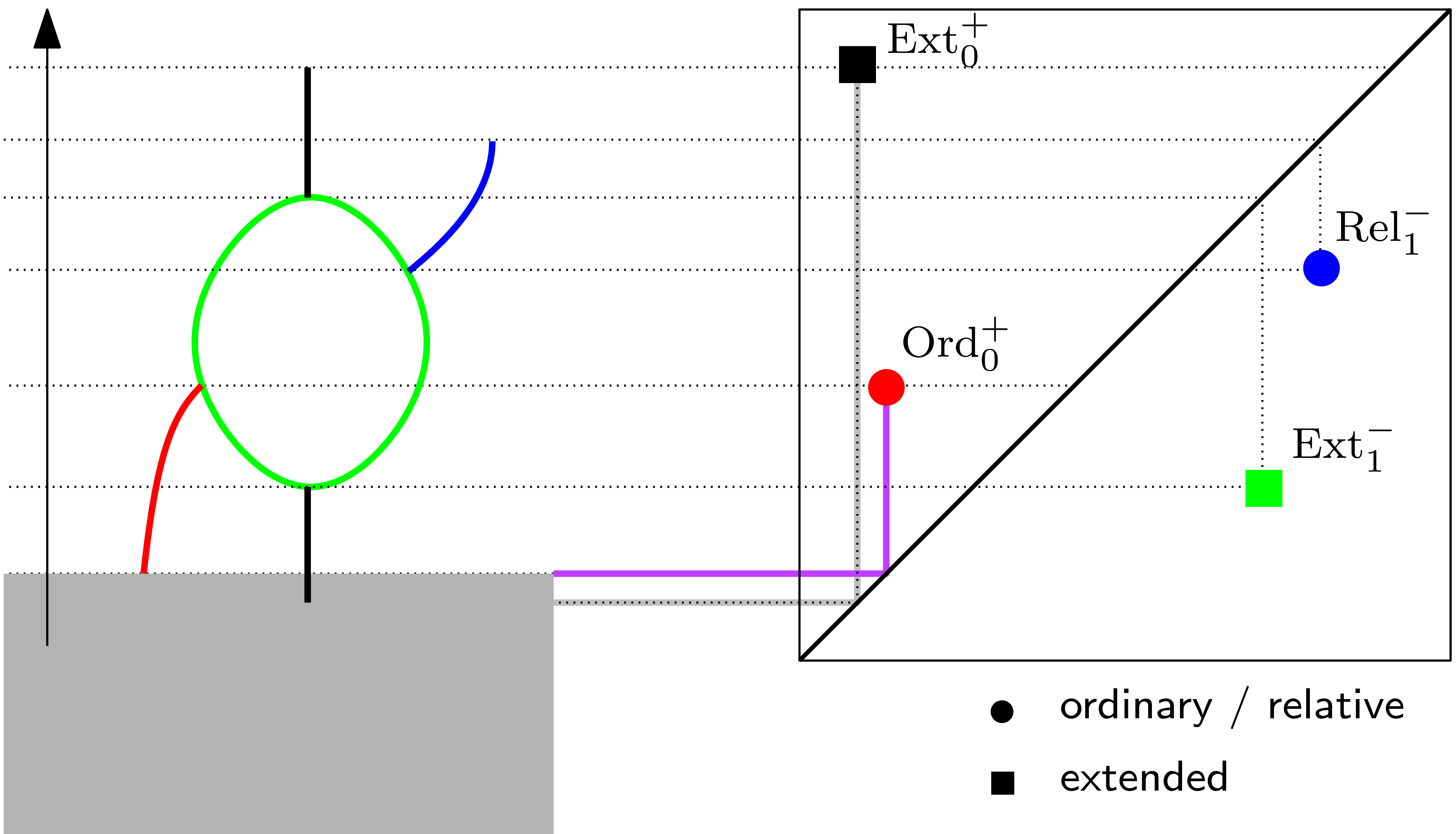
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

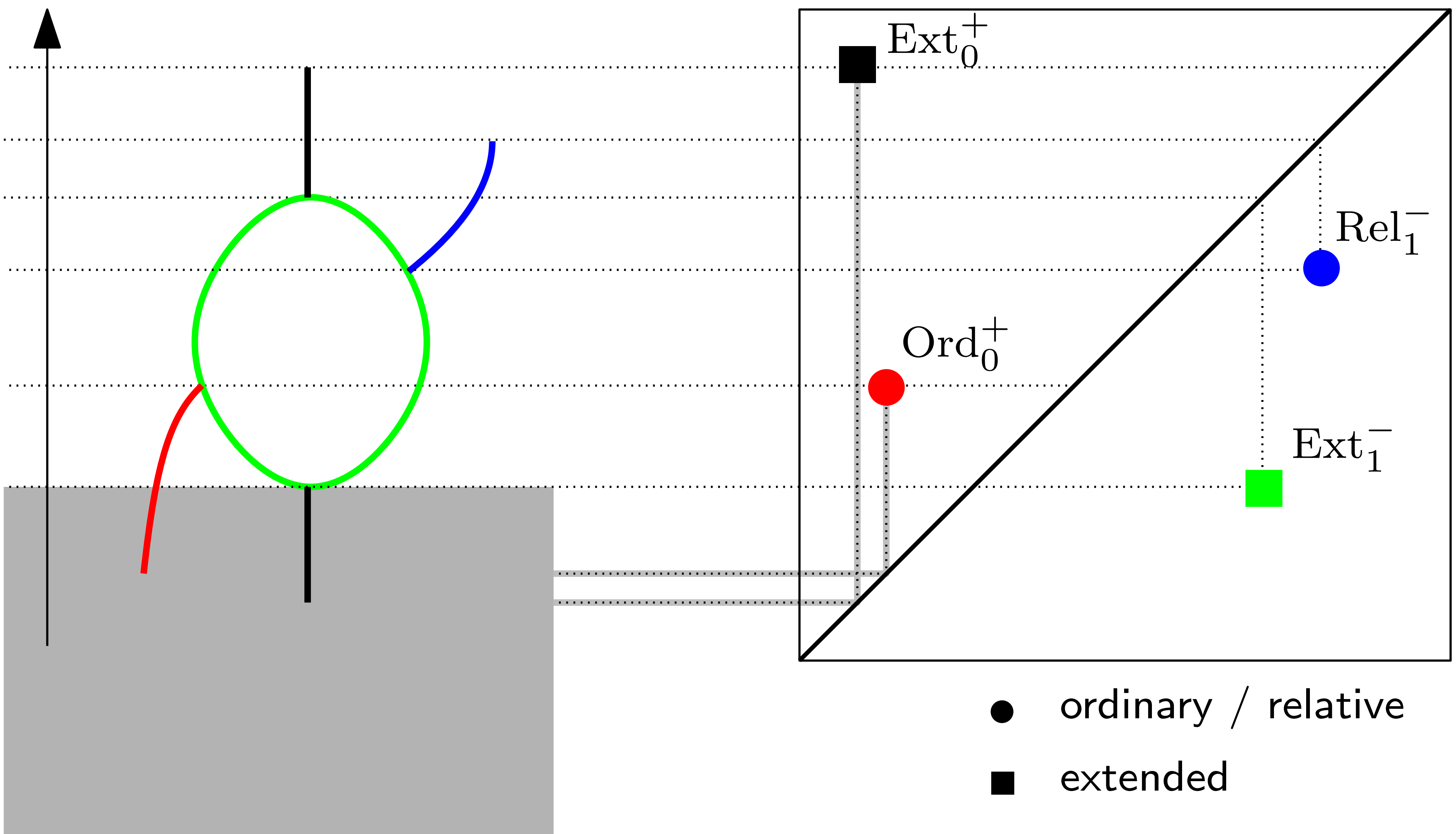
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

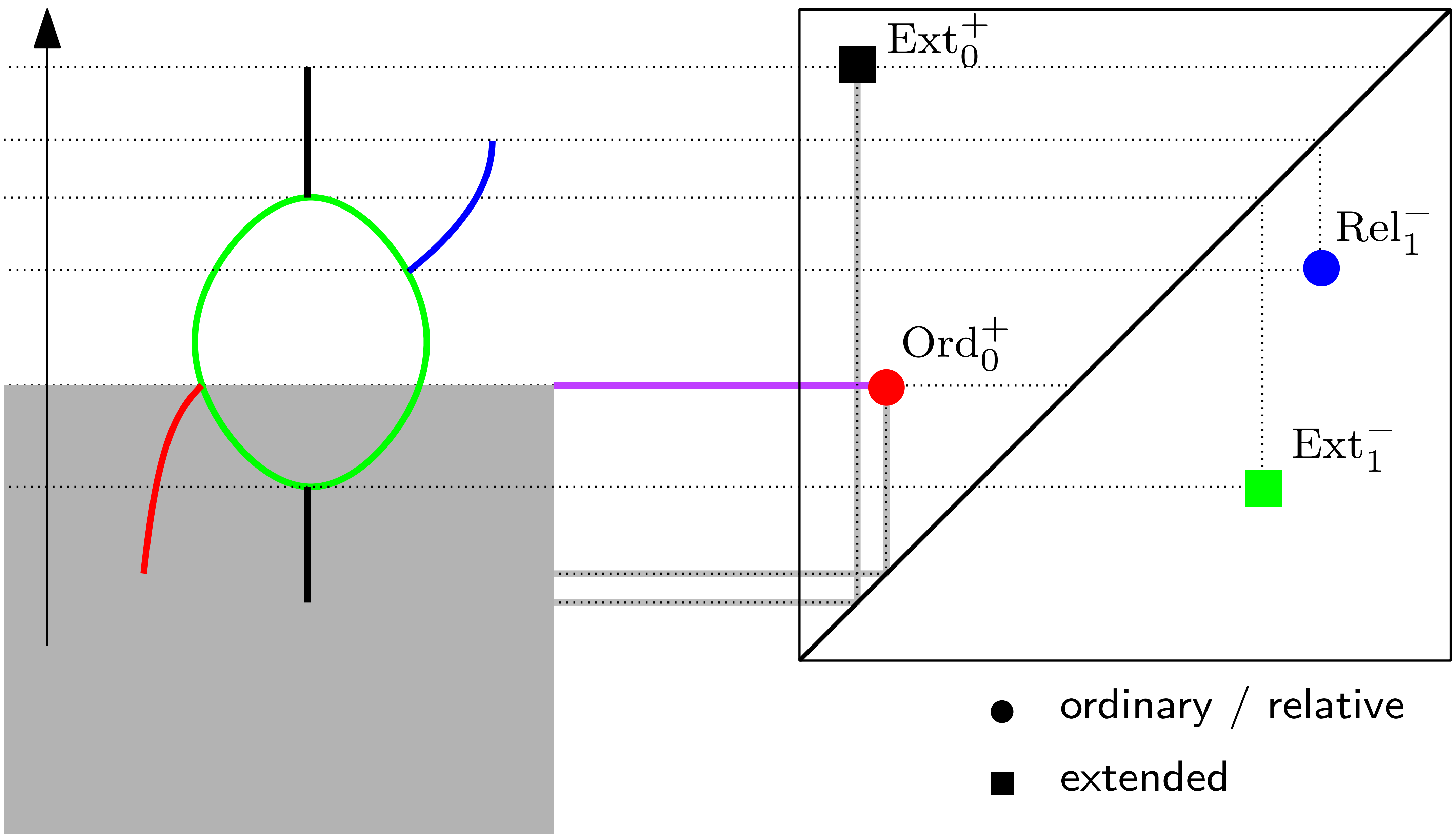
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

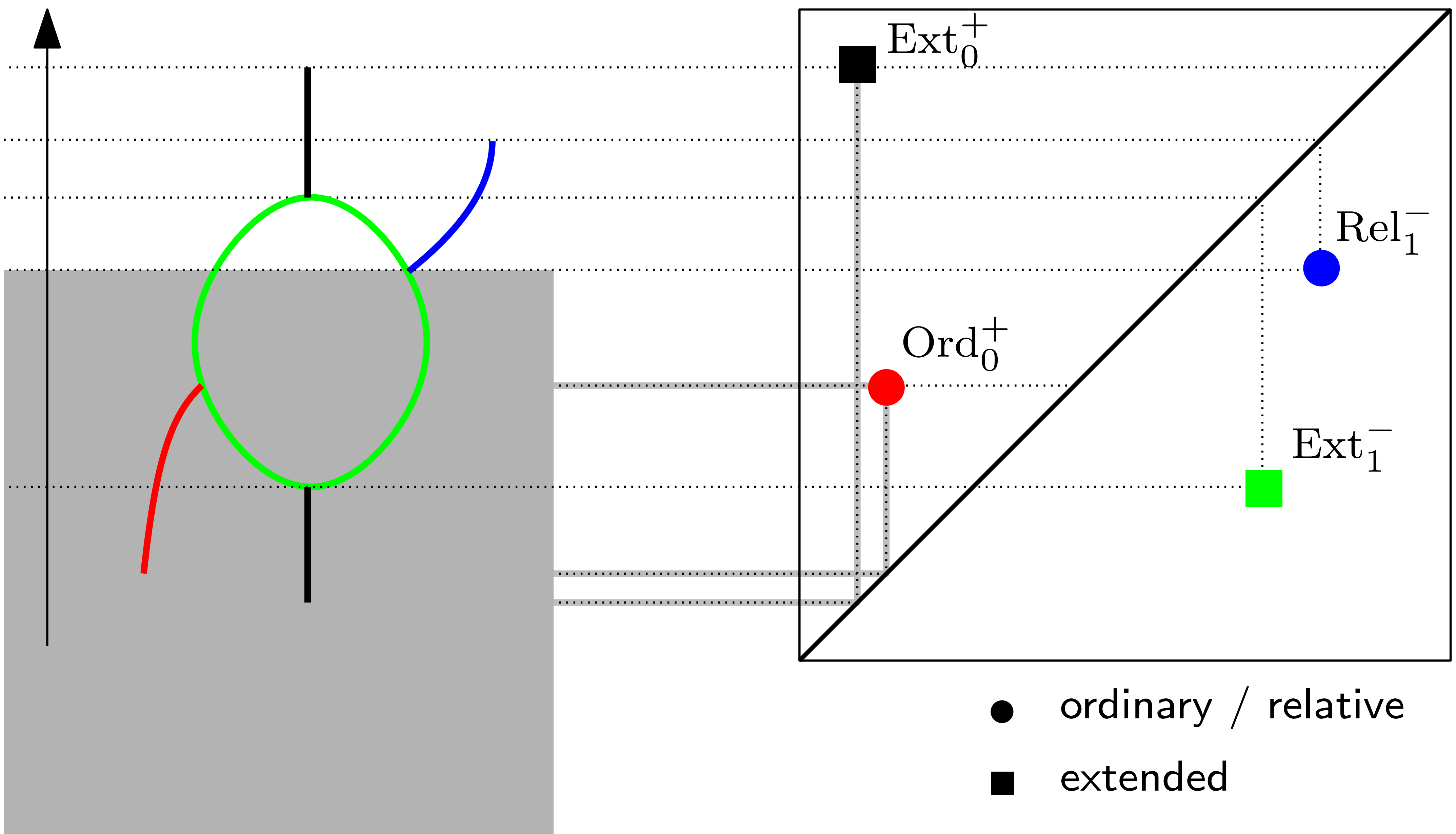
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

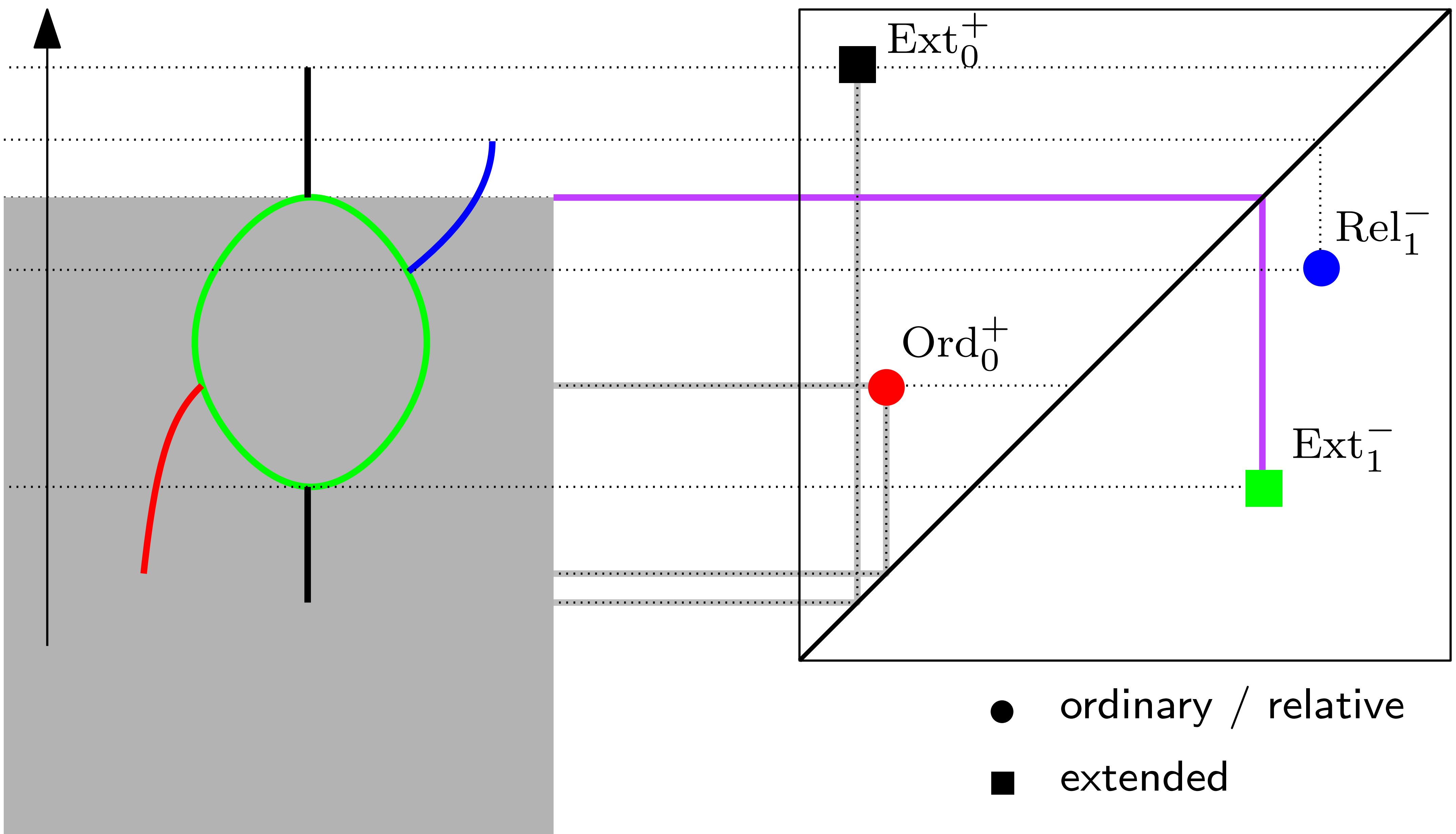
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

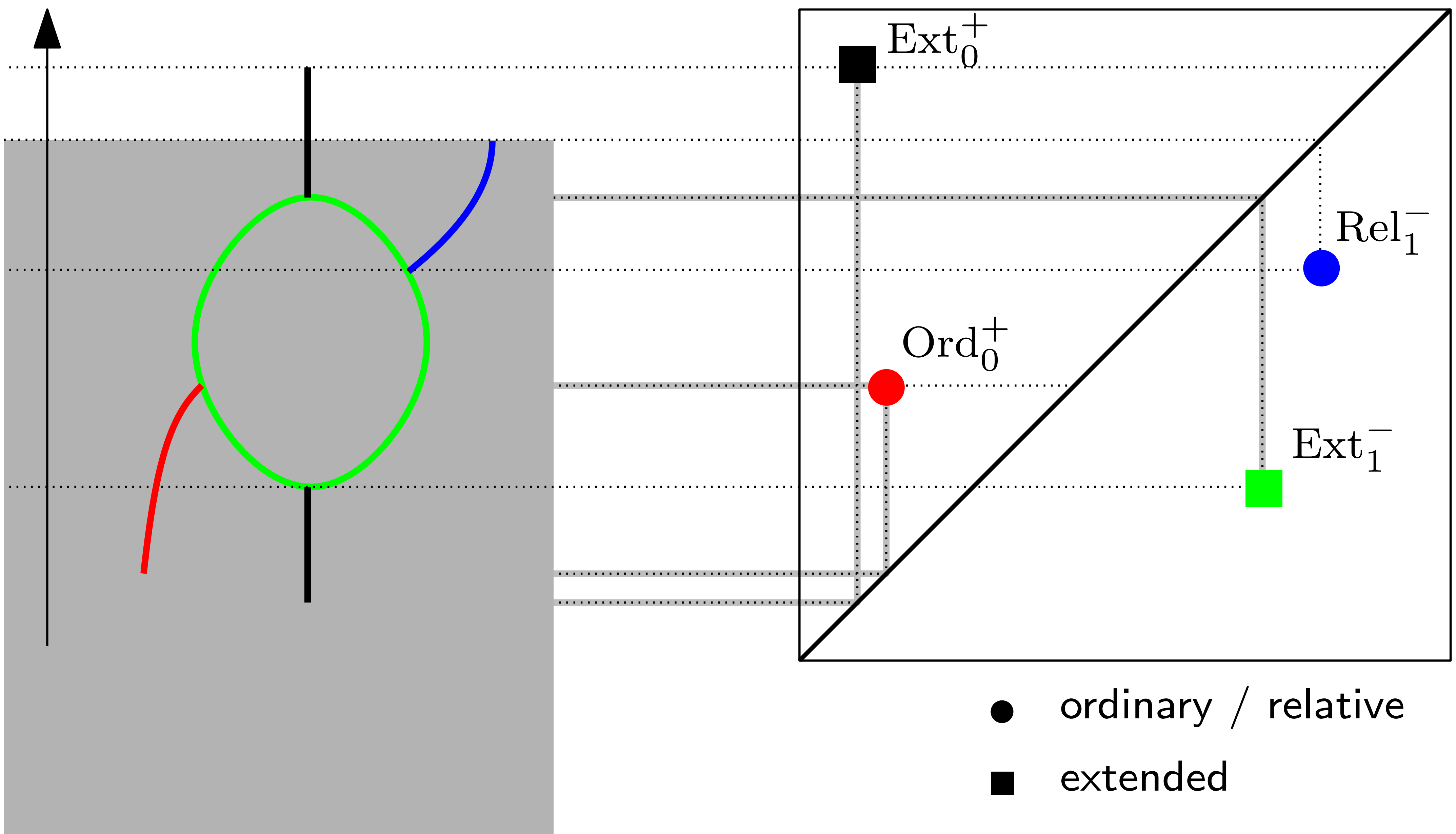
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

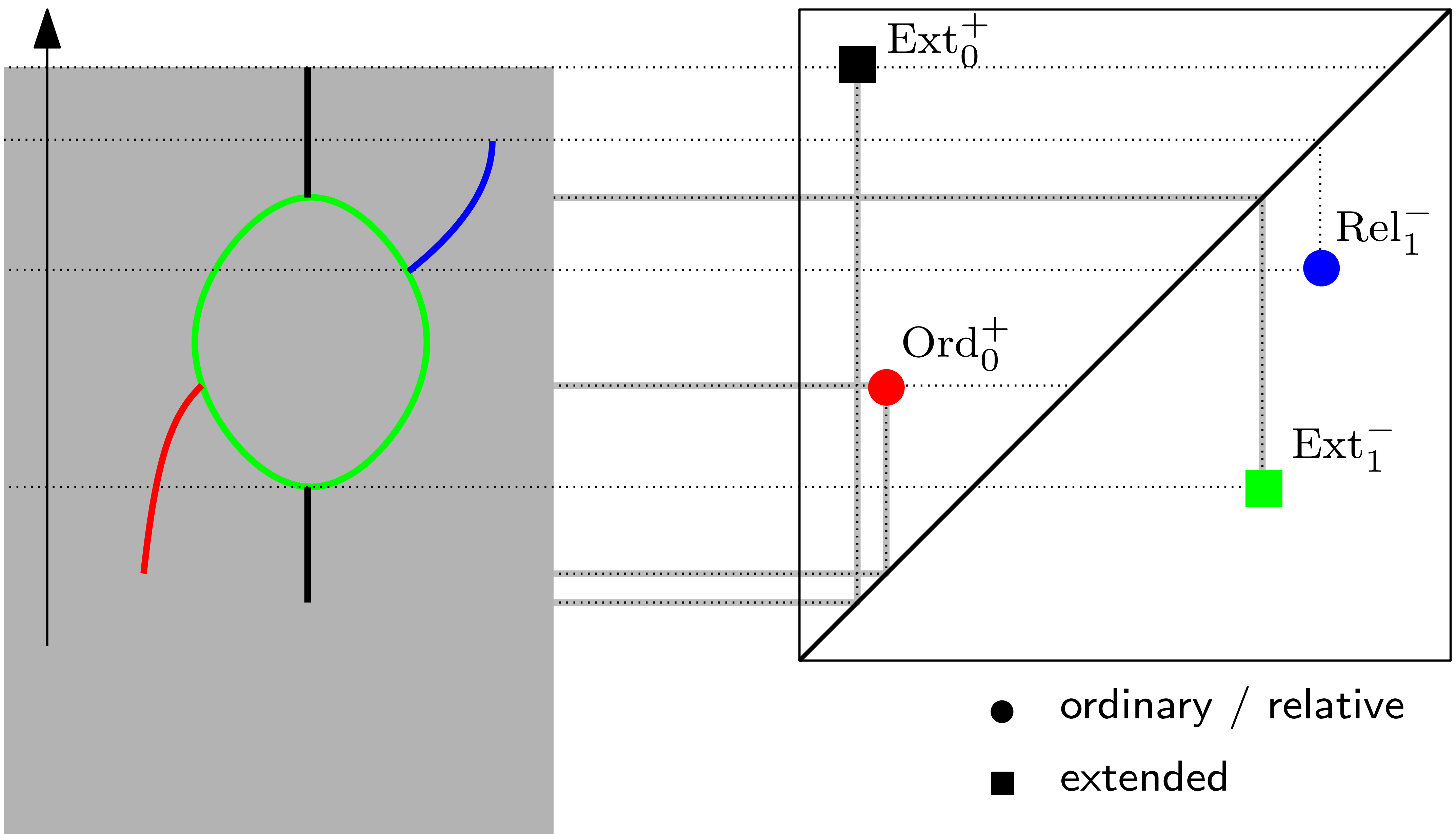
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

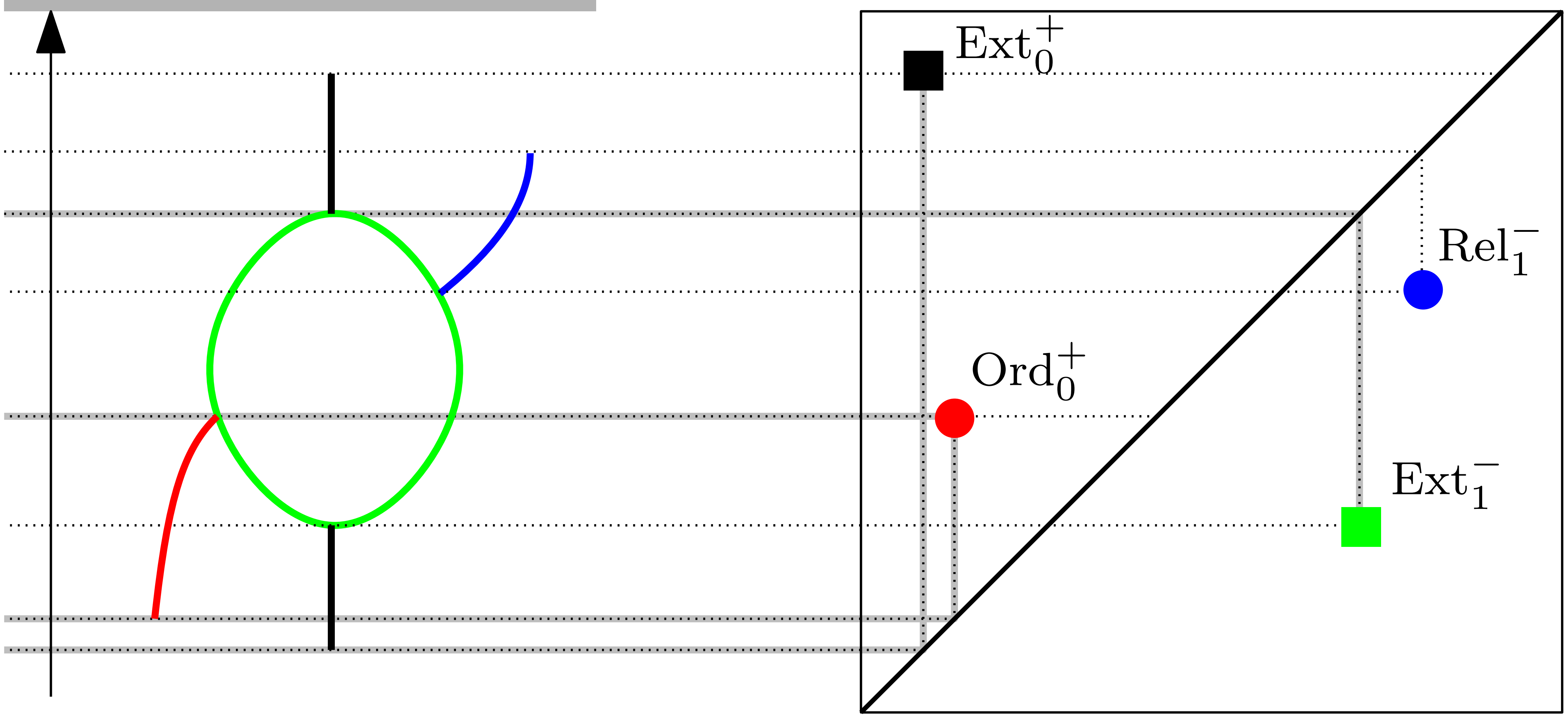
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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



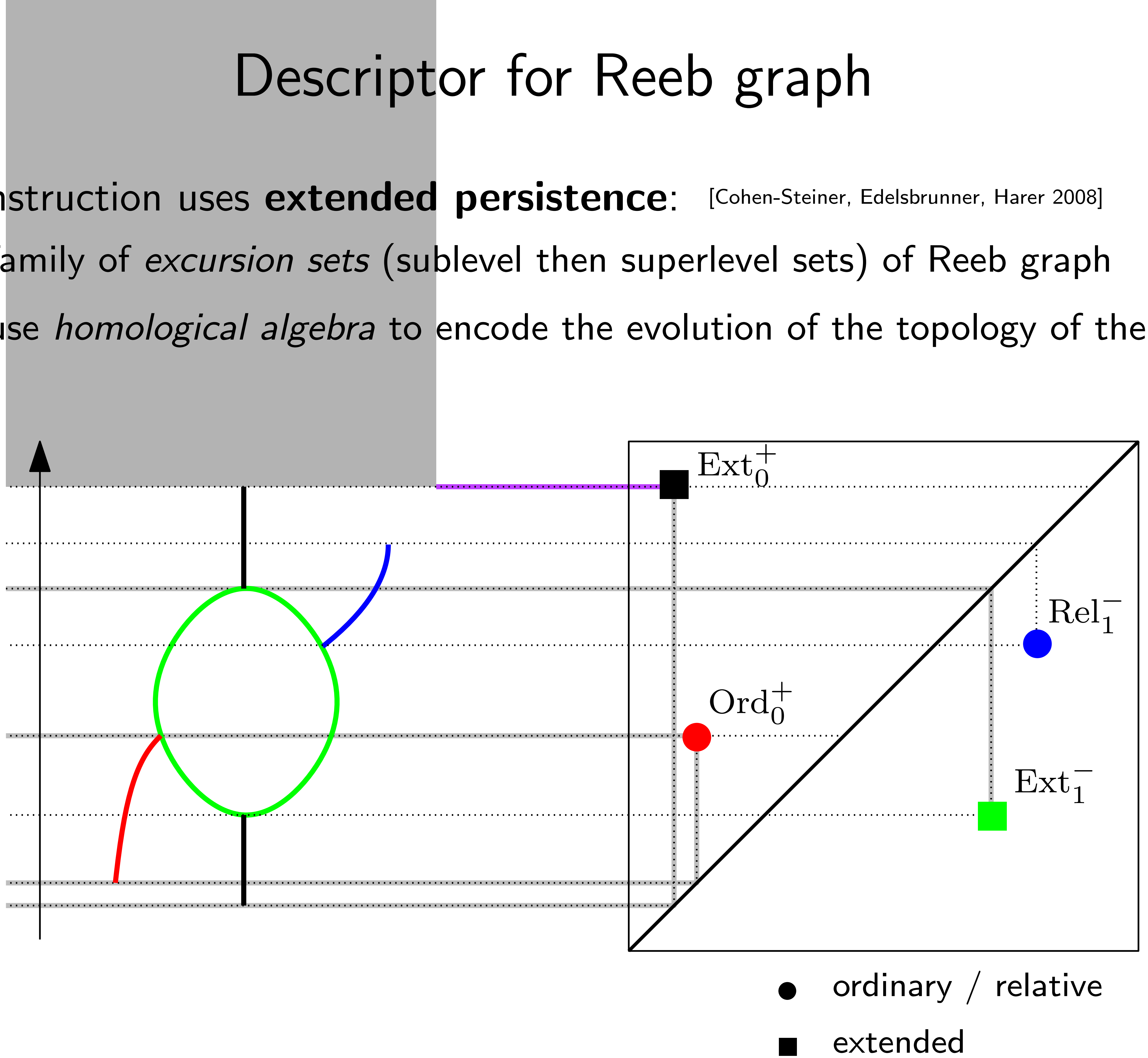
● ordinary / relative

■ extended

Descriptor for Reeb graph

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

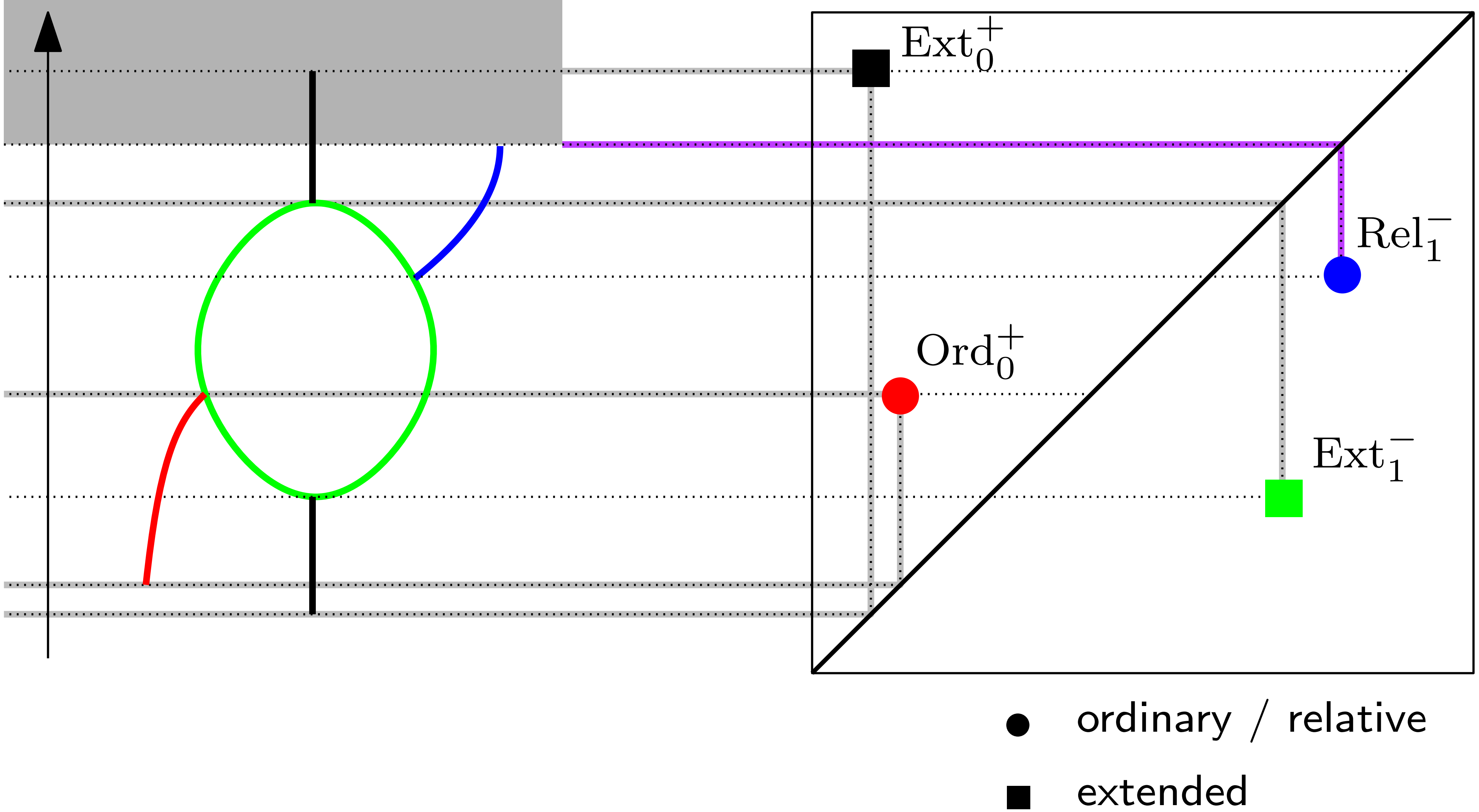
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

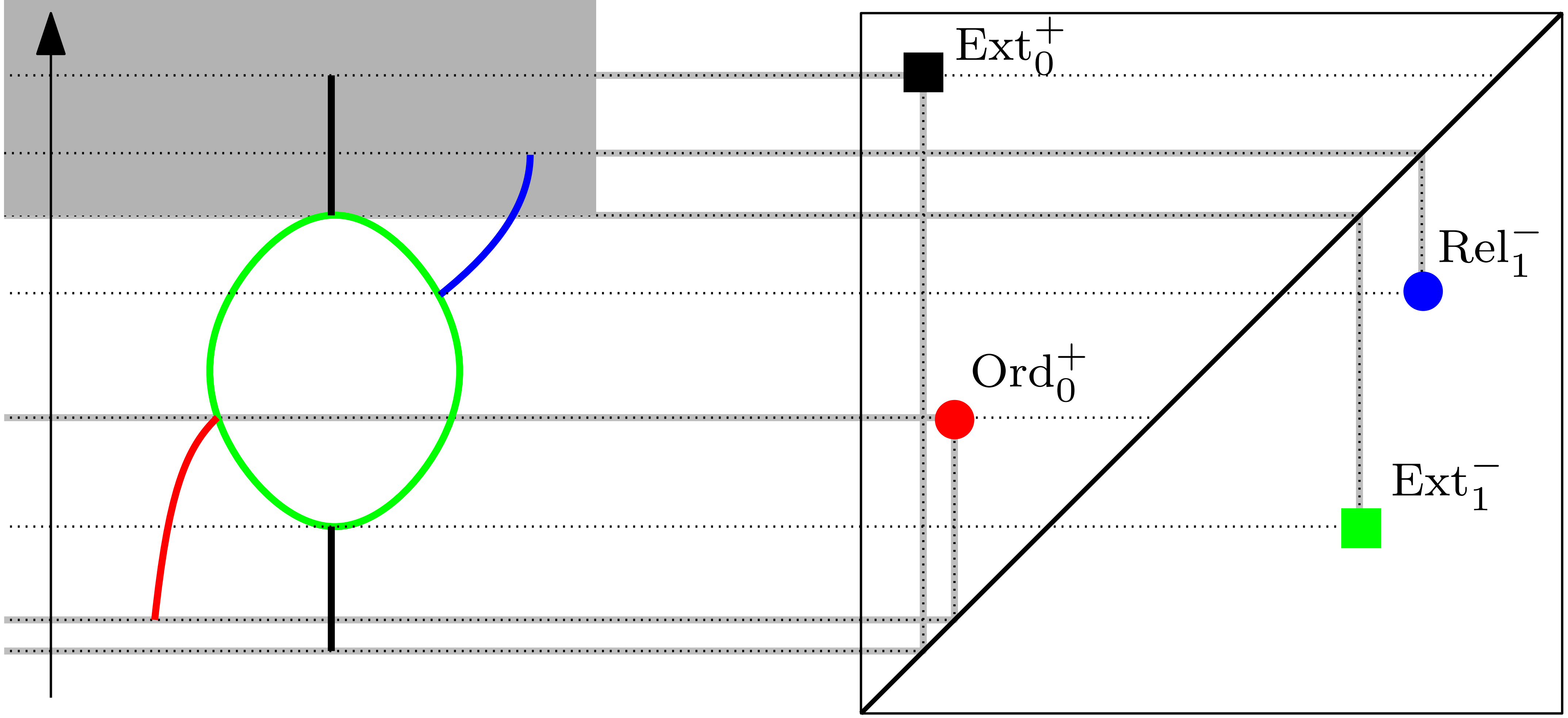
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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



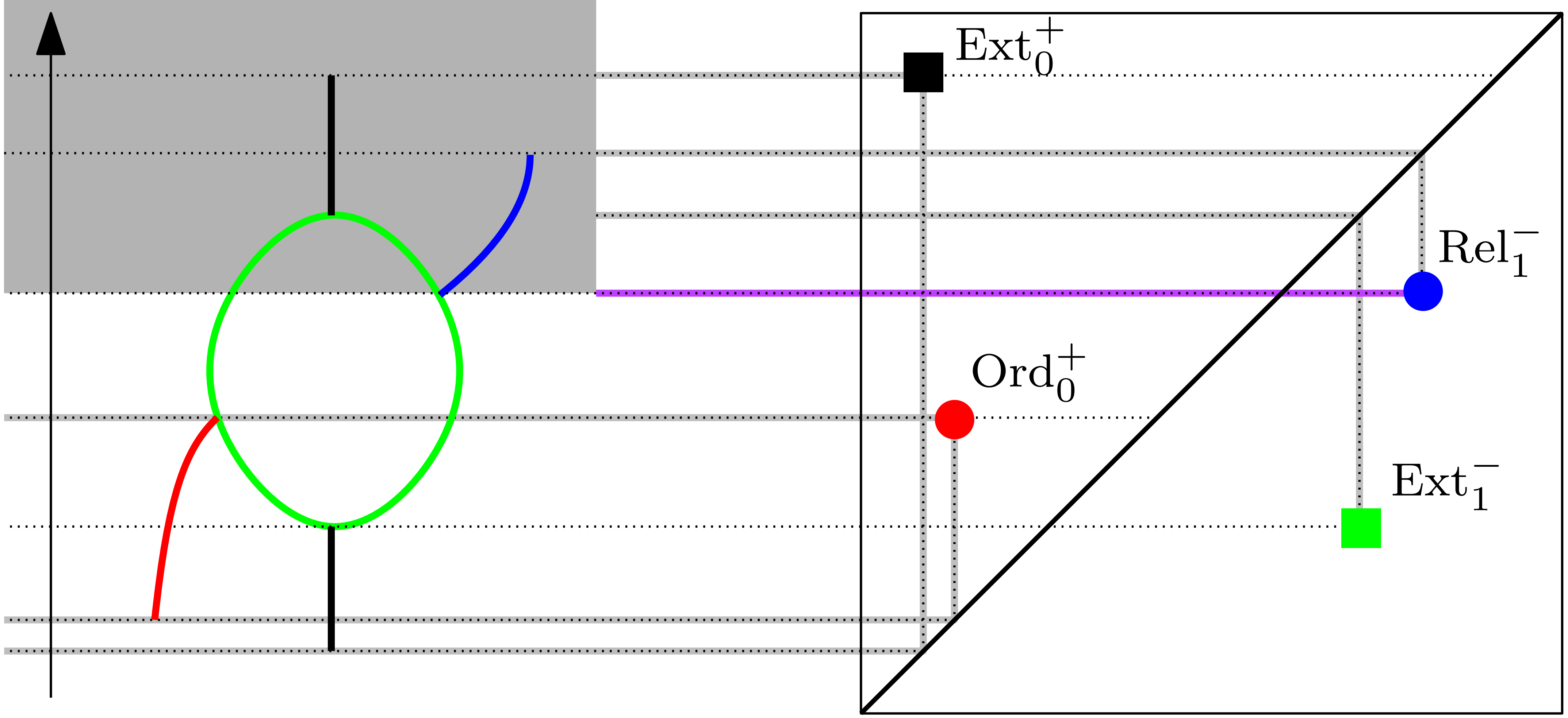
● ordinary / relative

■ extended

Descriptor for Reeb graph

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

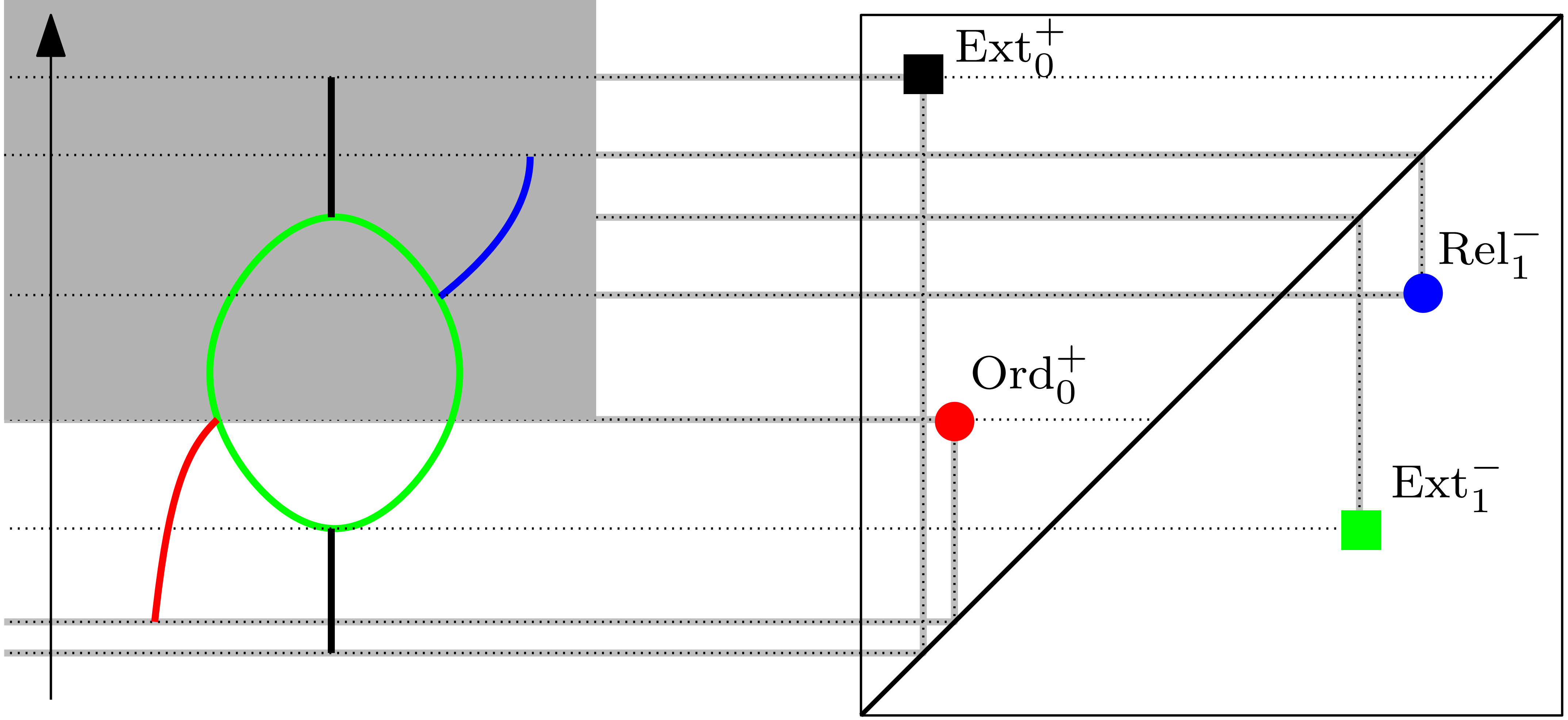


- ● ordinary / relative
- ■ extended

Descriptor for Reeb graph

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



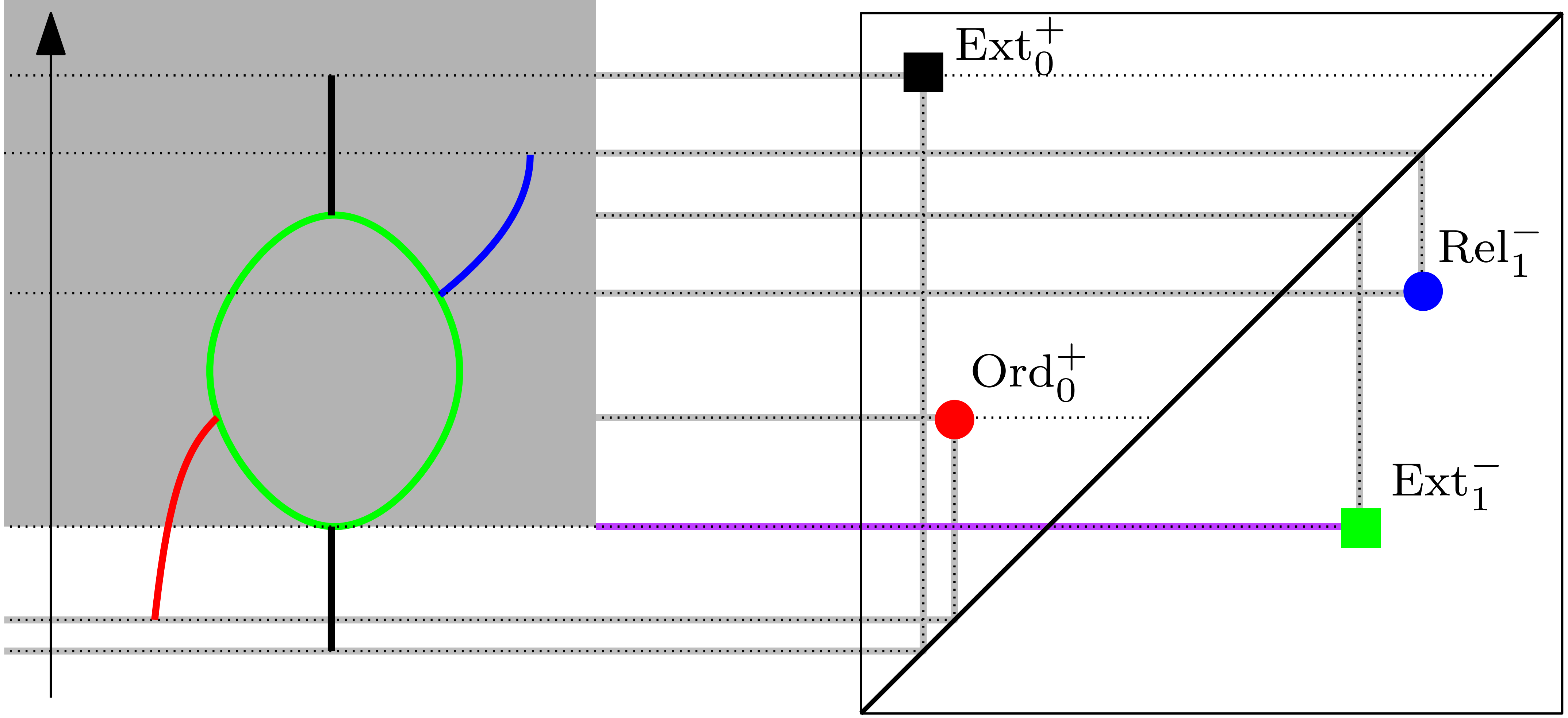
● ordinary / relative

■ extended

Descriptor for Reeb graph

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

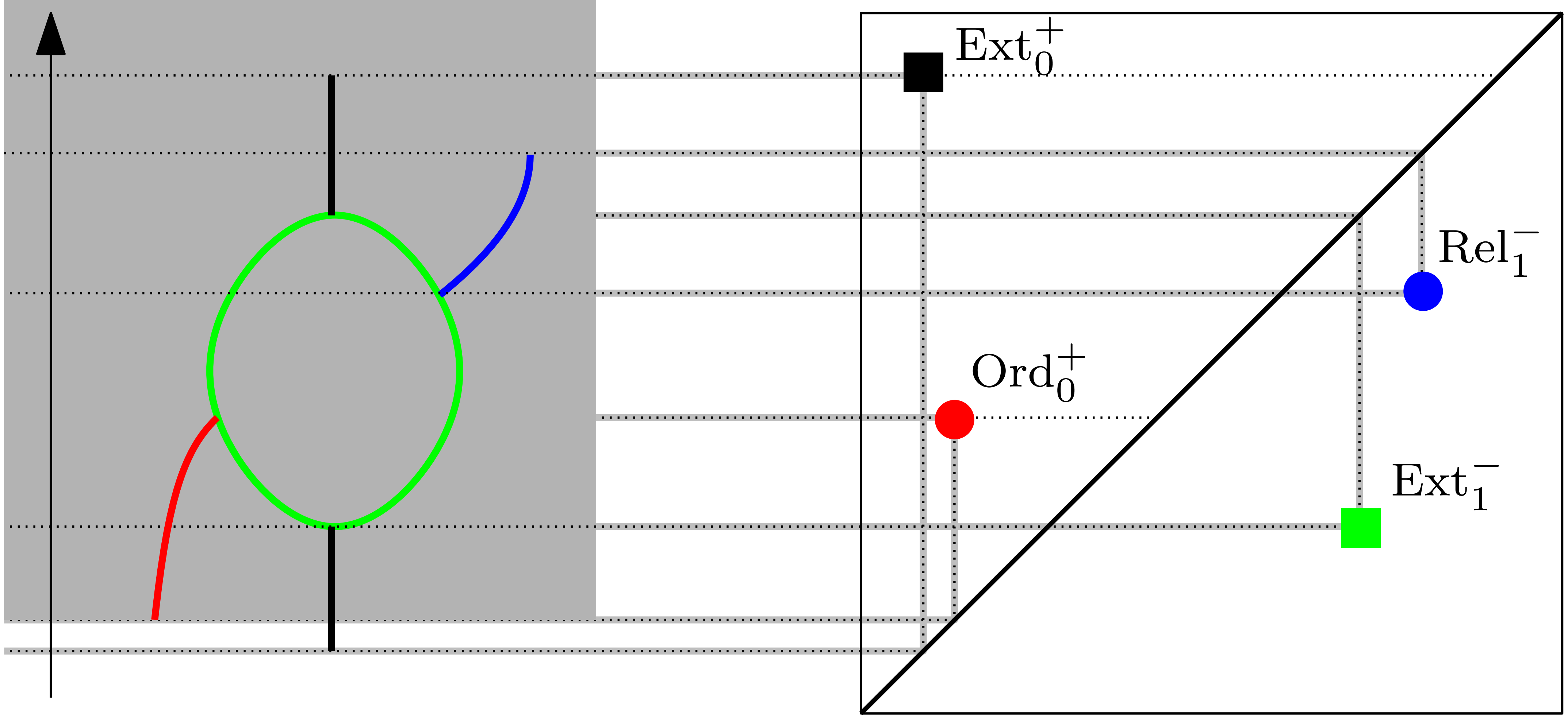


- ● ordinary / relative
- ■ extended

Descriptor for Reeb graph

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



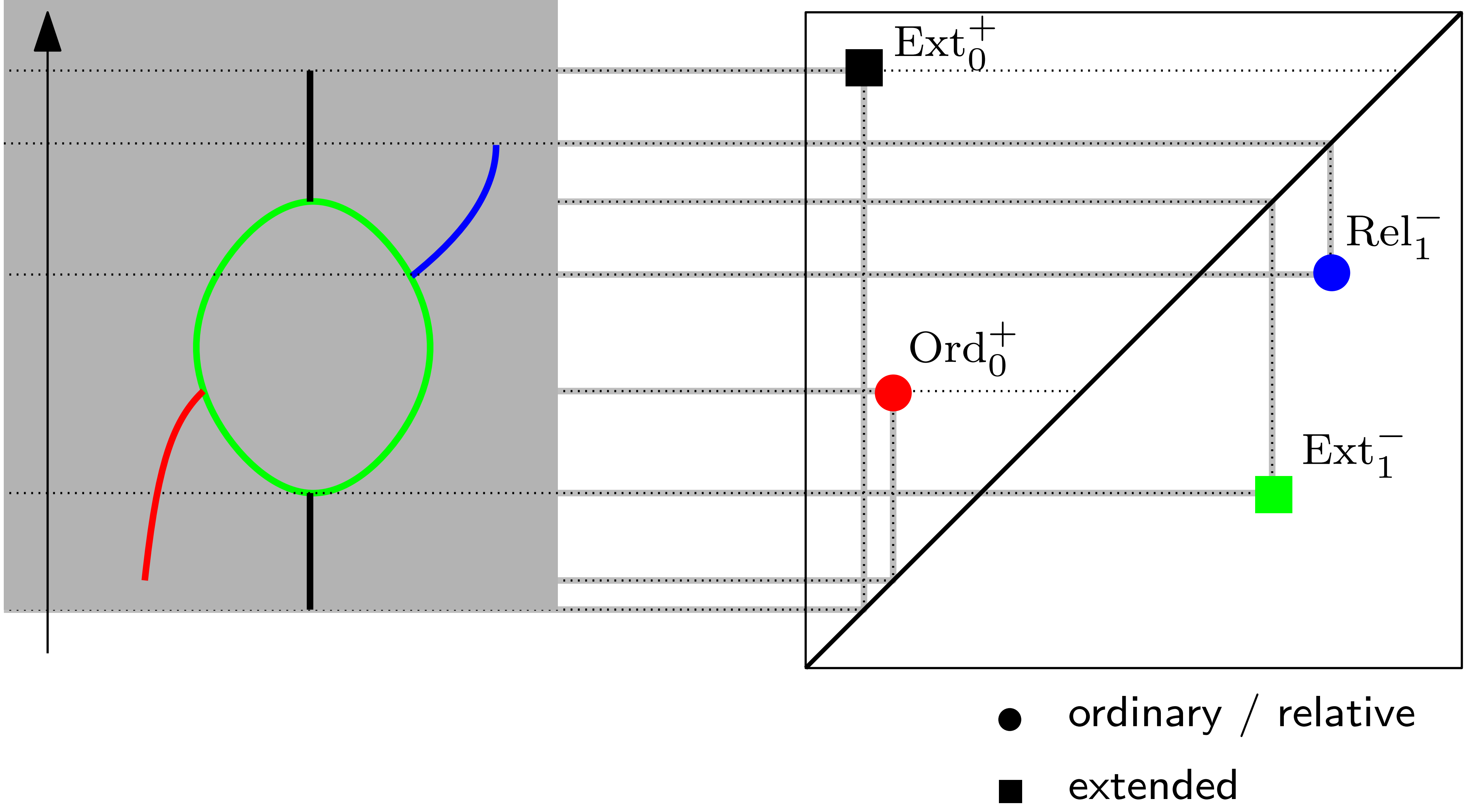
● ordinary / relative

■ extended

Descriptor for Reeb graph

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

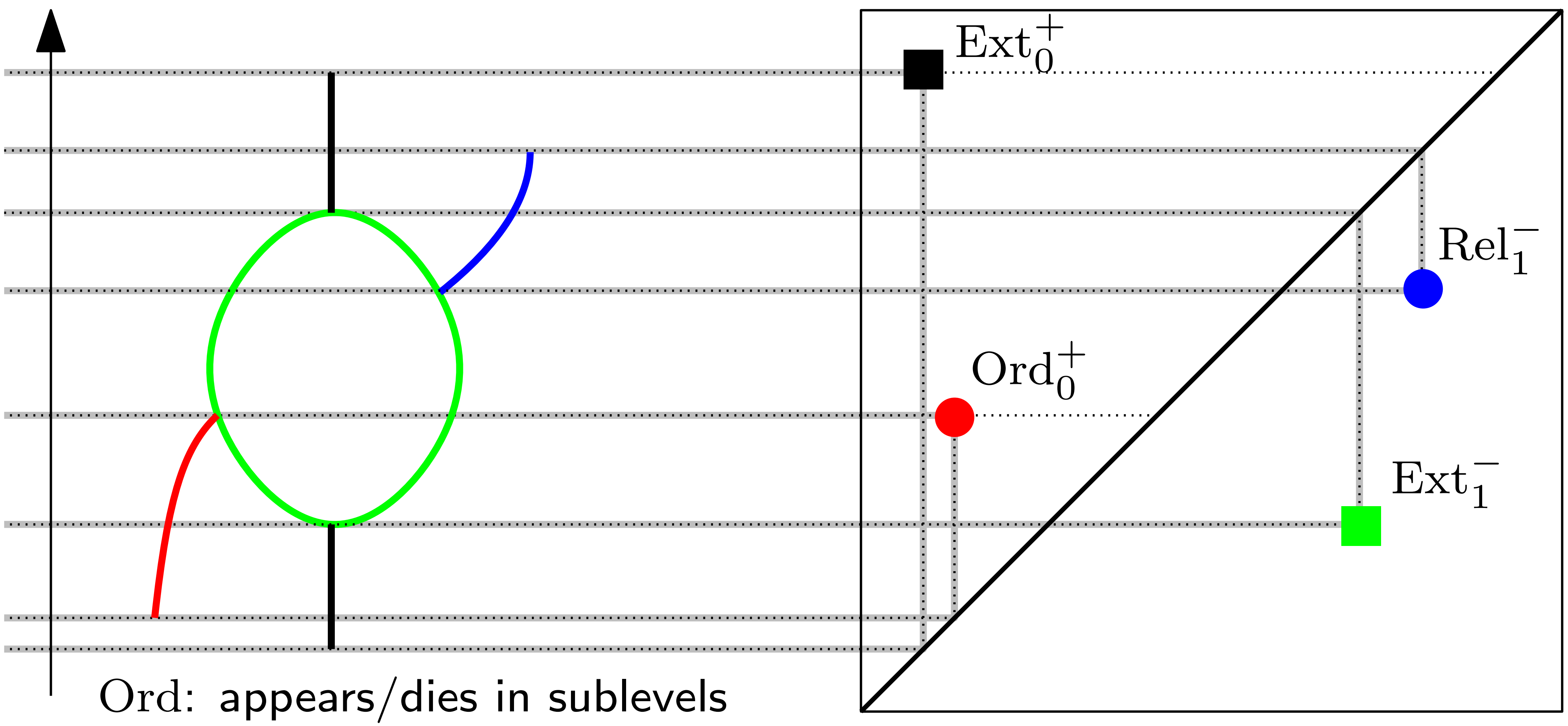
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Descriptor for Reeb graph

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

Descriptor for Reeb graph

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

Theorem (decomposition): [Crawley-Boevey'12] $< \cdots <$ [Gabriel'72]
Every extended persistence module M decomposes as a direct sum:

$$M \cong \bigoplus_{I \in \mathcal{I}} \mathbf{k}_I$$

where each summand $\mathbf{k}_I$ is an *interval module*, i.e. $\mathbf{k}_I :=$

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

Moreover, the decomposition is essentially unique [Azumaya'51].

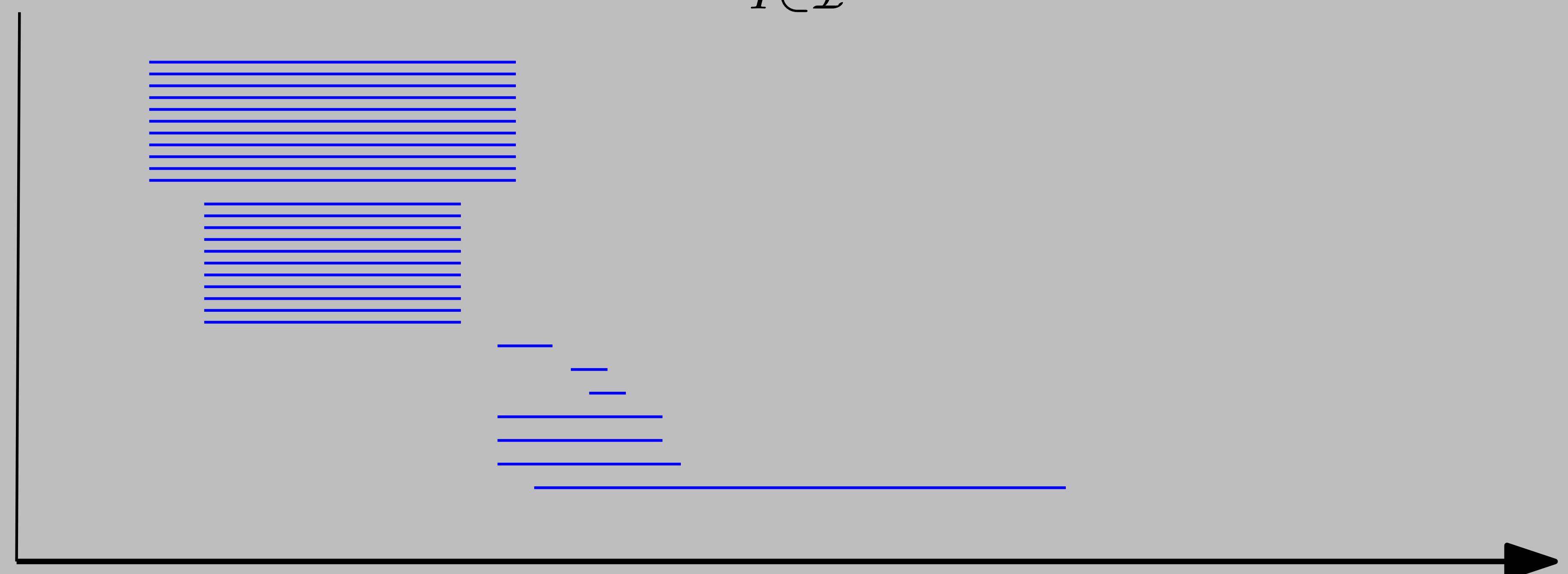
Descriptor for Reeb graph

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

Theorem (decomposition): [Crawley-Boevey'12] $<\cdots<$ [Gabriel'72]
Every extended persistence module M decomposes as a direct sum:

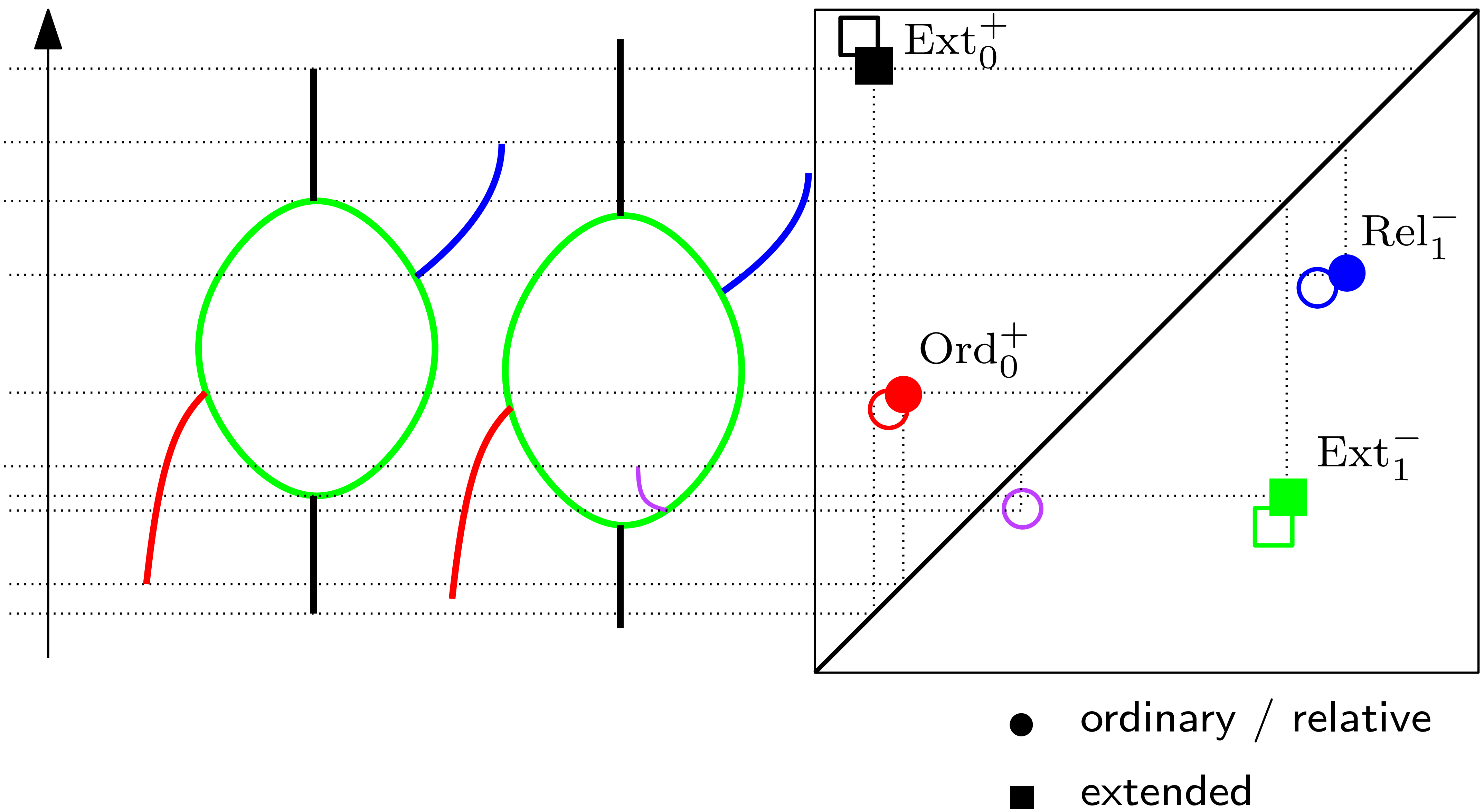
$$M \cong \bigoplus_{I \in \mathcal{I}} \mathbf{k}_I$$



Descriptor for Reeb graph

Theorem (stability): [Bauer, Ge, Wang 2013]

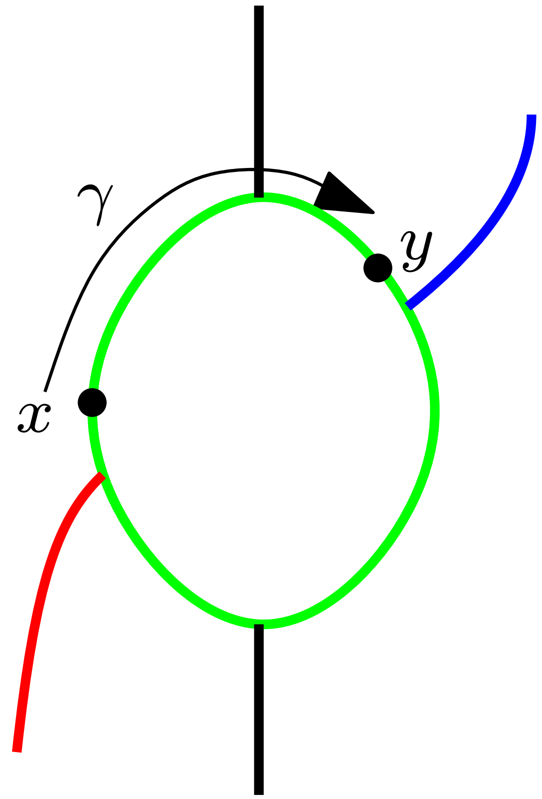
$$\mathrm{d}_B(\mathrm{Dg}\,\mathrm{R}_f, \mathrm{Dg}\,\mathrm{R}_g) \leq 6\,\mathrm{d}_{\mathrm{GH}}(\mathrm{R}_f, \mathrm{R}_g)$$



Descriptor for Reeb graph

Theorem (stability): [Bauer, Ge, Wang 2013]

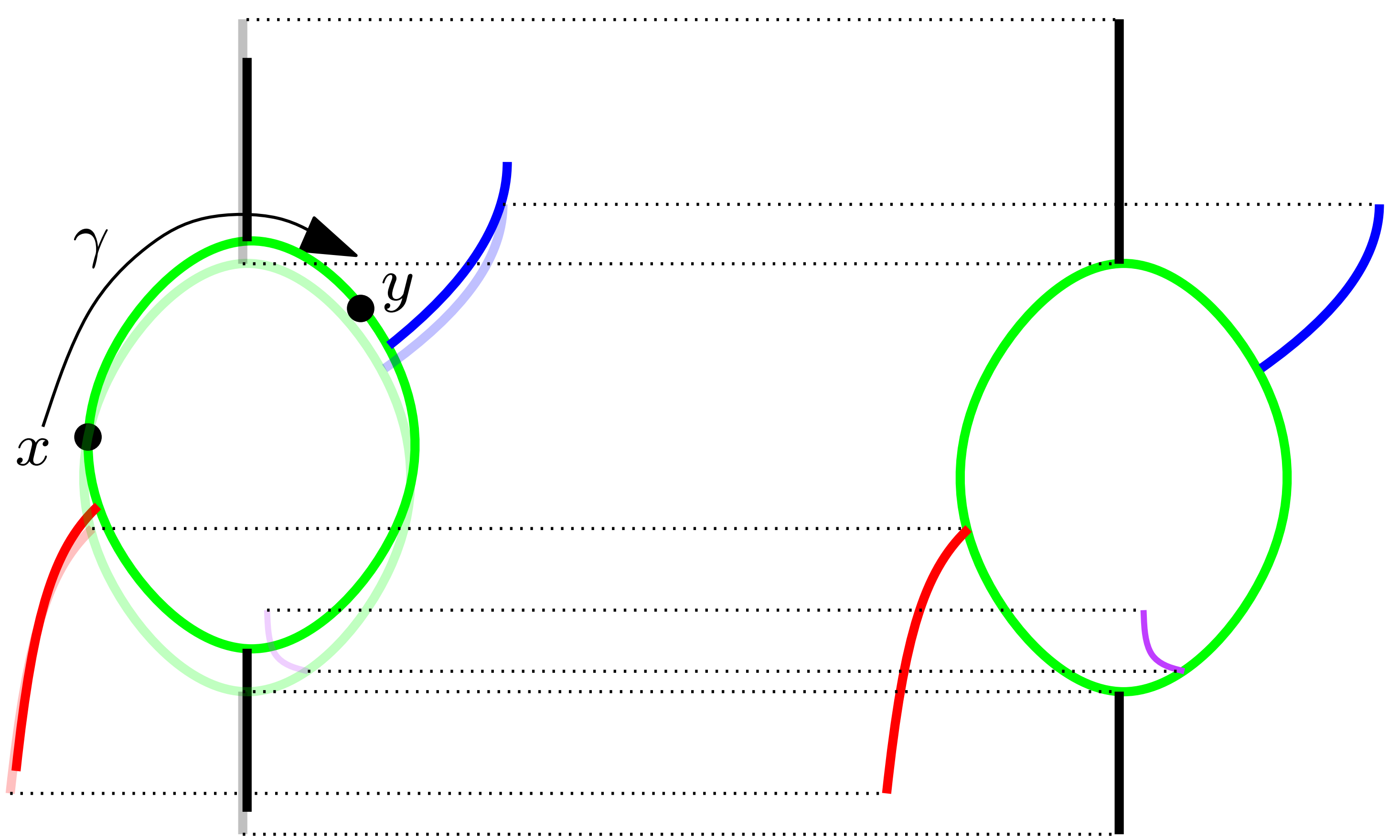
$$\mathrm{d}_B(\mathrm{Dg}\,\mathrm{R}_f, \mathrm{Dg}\,\mathrm{R}_g) \leq 6\,\mathrm{d}_{\mathrm{GH}}(\mathrm{R}_f, \mathrm{R}_g)$$



Descriptor for Reeb graph

Theorem (stability): [Bauer, Ge, Wang 2013]

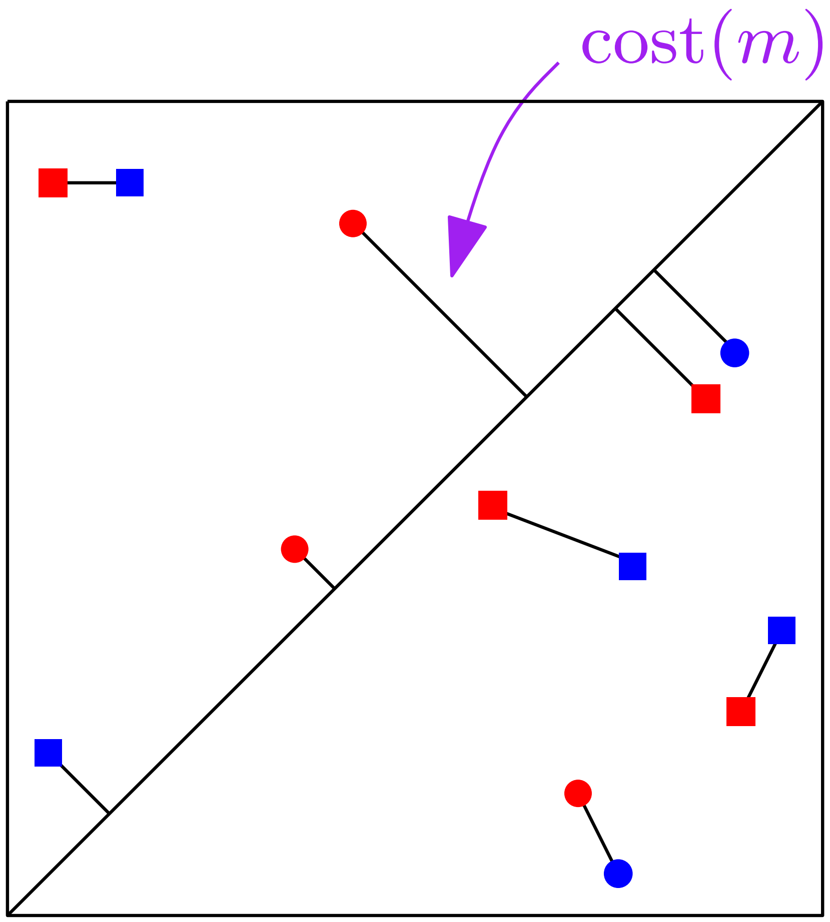
$$\mathrm{d}_B(\mathrm{Dg}\,\mathrm{R}_f, \mathrm{Dg}\,\mathrm{R}_g) \leq 6\,\mathrm{d}_{\mathrm{GH}}(\mathrm{R}_f, \mathrm{R}_g)$$



Descriptor for Reeb graph

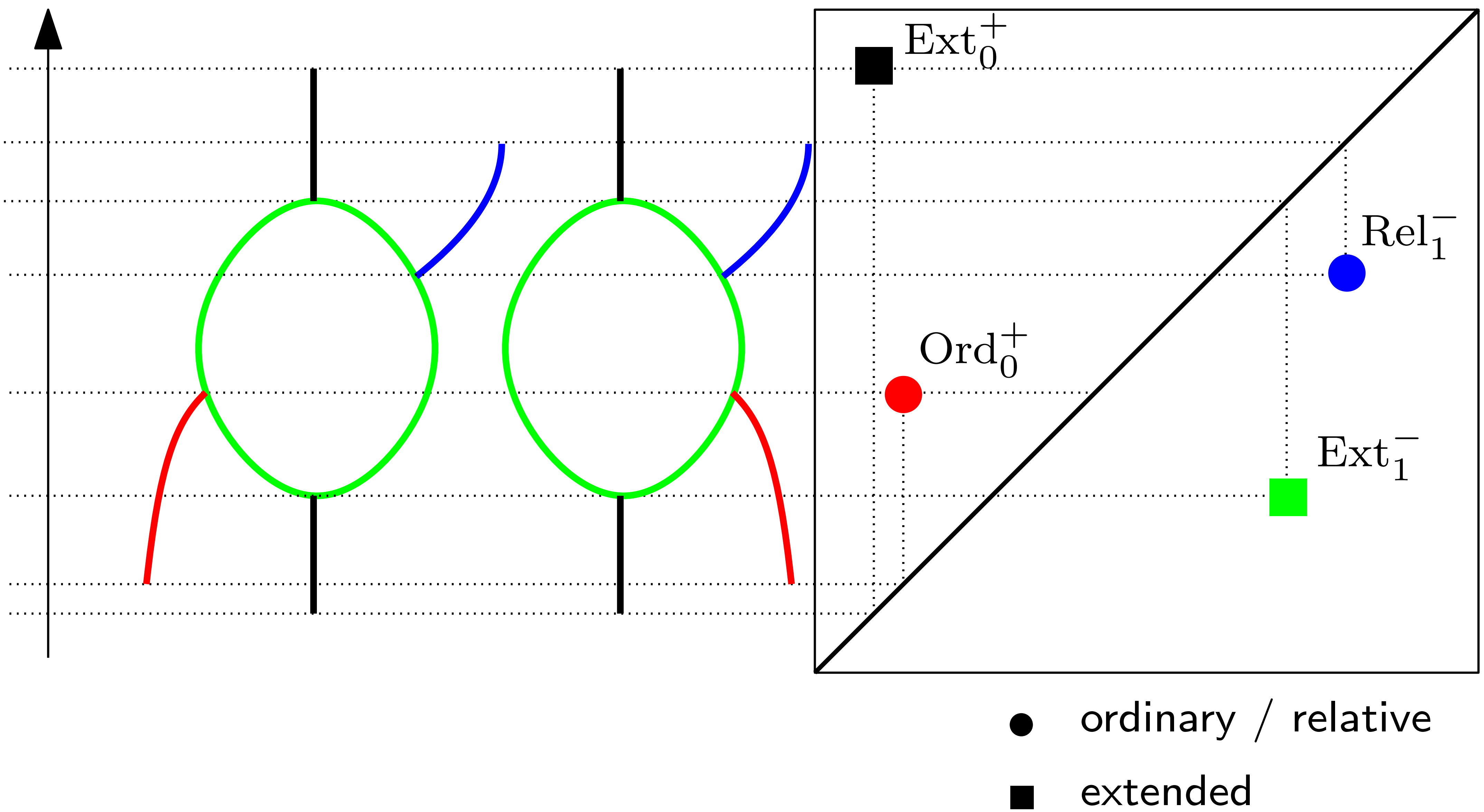
Theorem (stability): [Bauer, Ge, Wang 2013]

$$\mathrm{d}_B(\mathrm{Dg}\,\mathrm{R}_f, \mathrm{Dg}\,\mathrm{R}_g) \leq 6\,\mathrm{d}_{\mathrm{GH}}(\mathrm{R}_f, \mathrm{R}_g)$$



Descriptor for Reeb graph

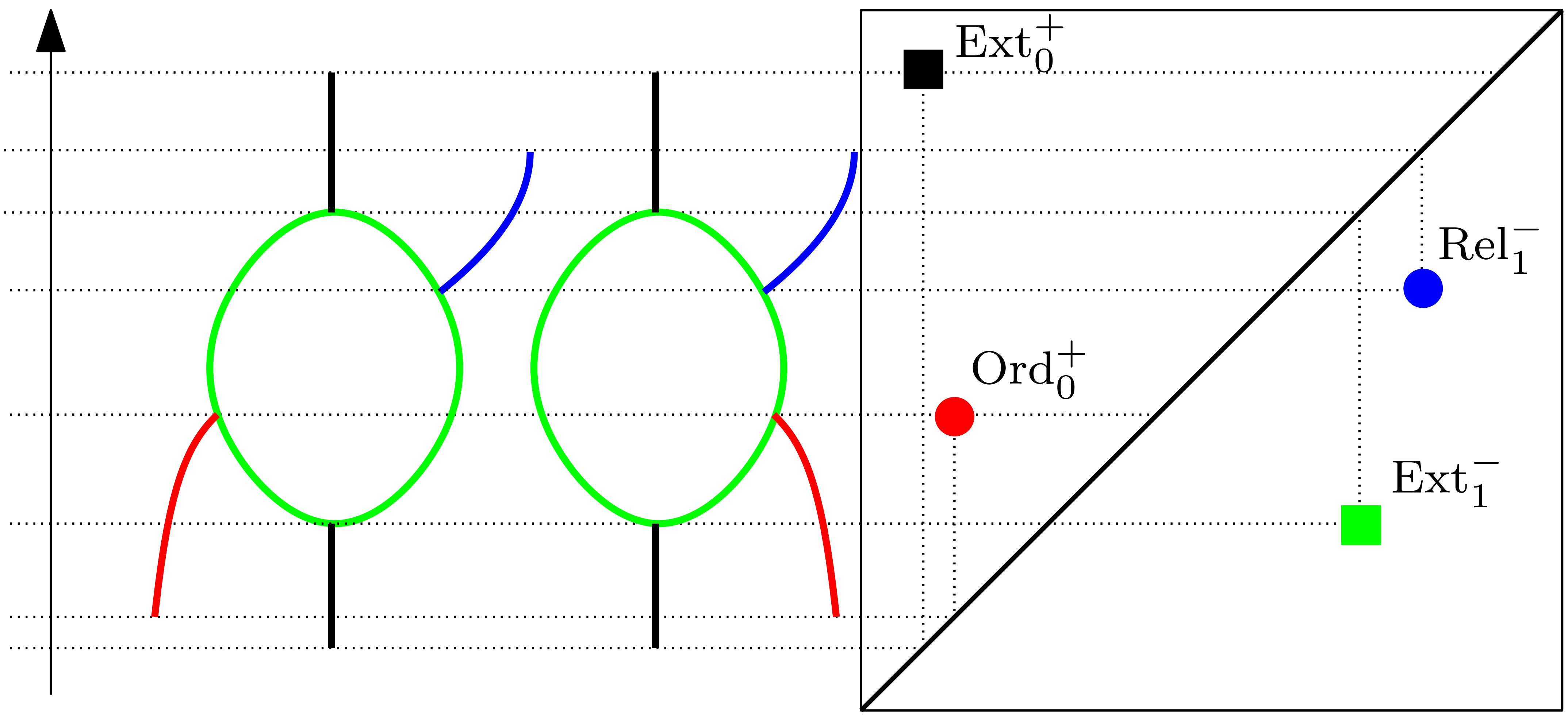
Note: $\mathrm{d}_B(\mathrm{Dg}\,\cdot, \mathrm{Dg}\,\cdot)$ is only a pseudometric on Reeb graphs



Descriptor for Reeb graph

Note: $\mathrm{d}_B(\mathrm{Dg}\,\cdot, \mathrm{Dg}\,\cdot)$ is only a pseudometric on Reeb graphs

Thm: [Carrière, O. 2017]
$\mathrm{d}_B(\mathrm{Dg}\,\cdot, \mathrm{Dg}\,\cdot)$ is *locally* a metric equivalent to d_{GH}

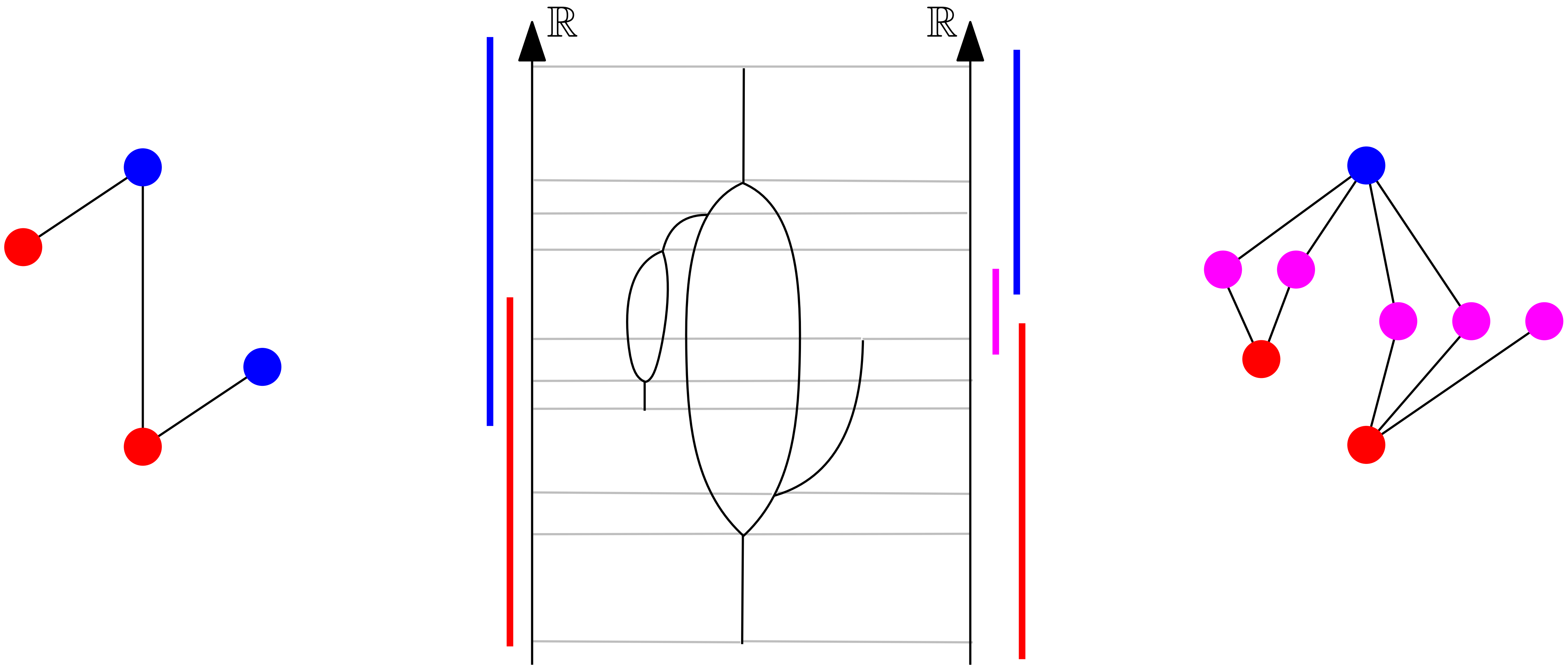


● ordinary / relative

■ extended

Descriptor for Mapper

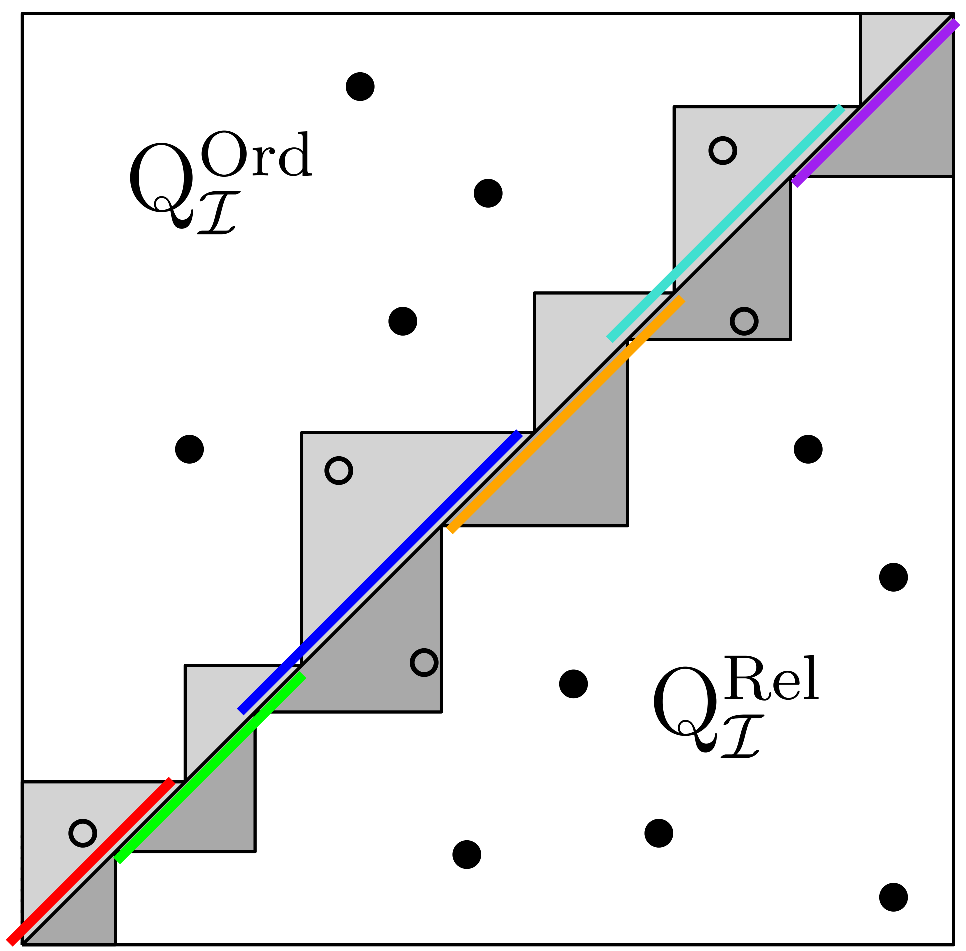
Reminder: mapper $\equiv$ *pixelized* Reeb graph

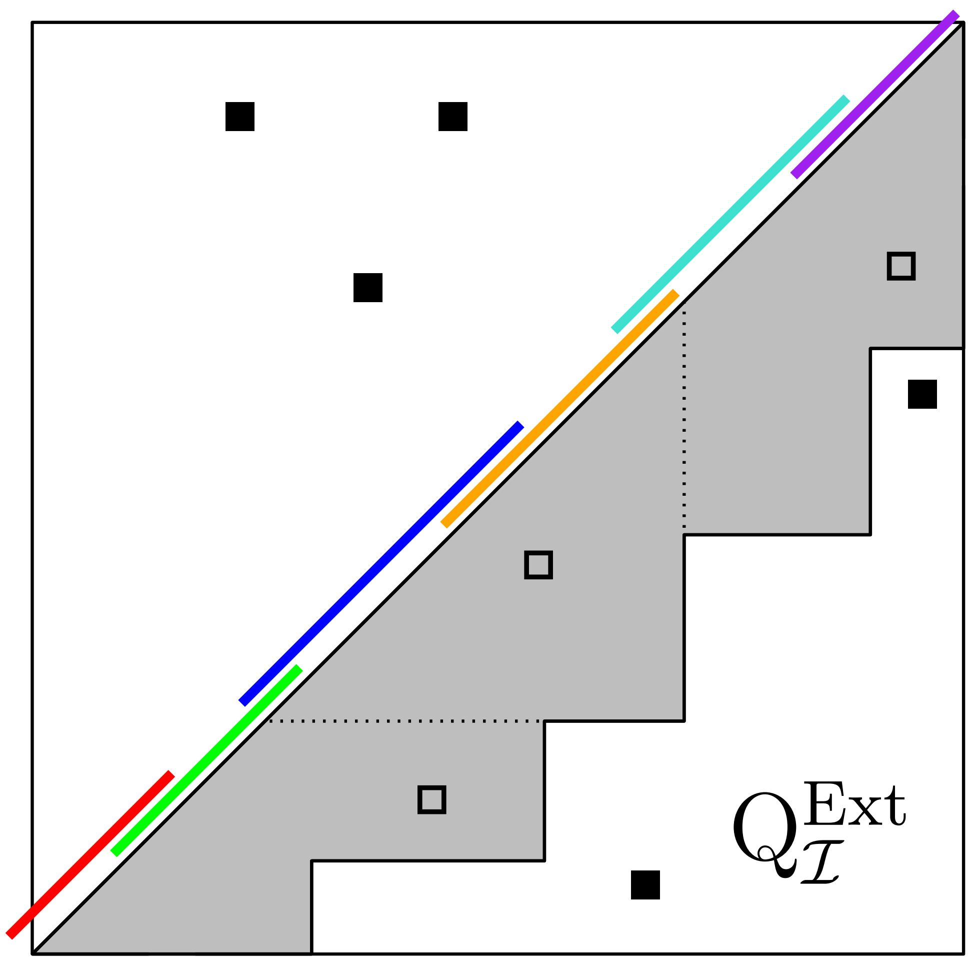


Descriptor for Mapper

Def: Given $X, f, \mathcal{I}$:

$$\mathrm{Dg}\,\mathrm{M}_f := \left(\mathrm{Ord}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ord}}\right) \cup \left(\mathrm{Rel}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Rel}}\right) \cup \left(\mathrm{Ext}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}\right)$$



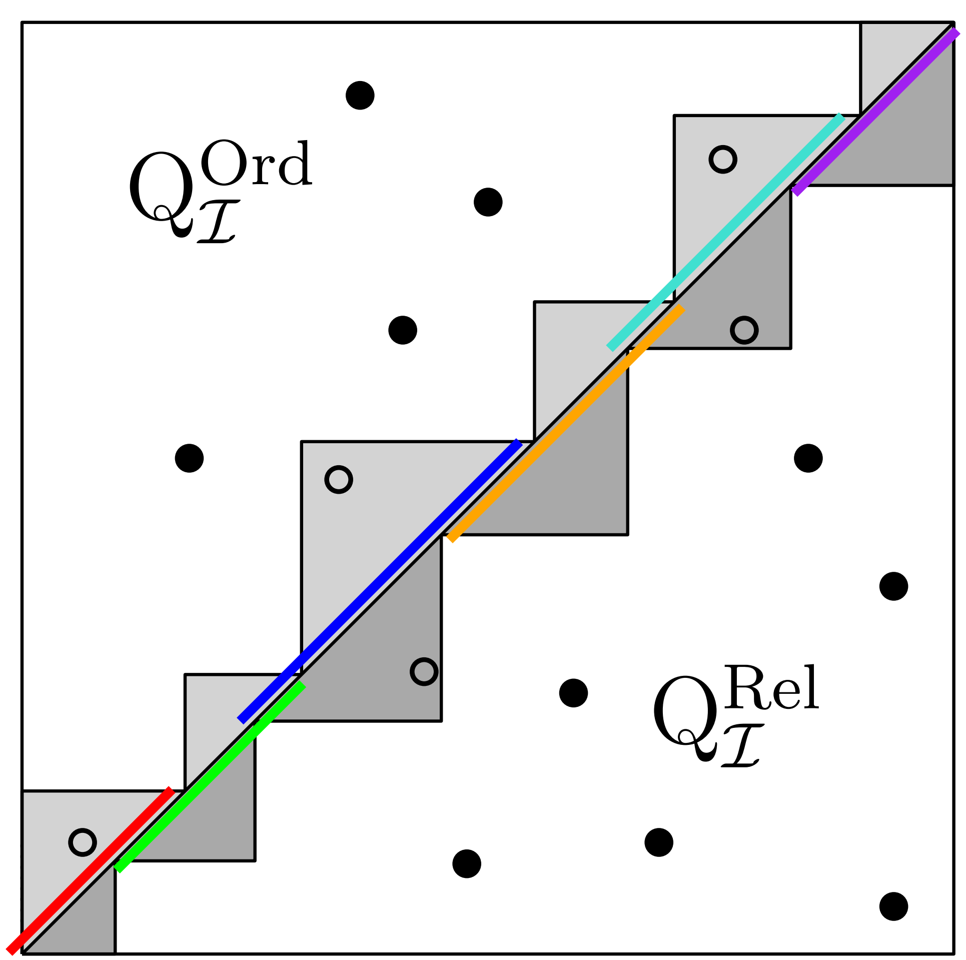


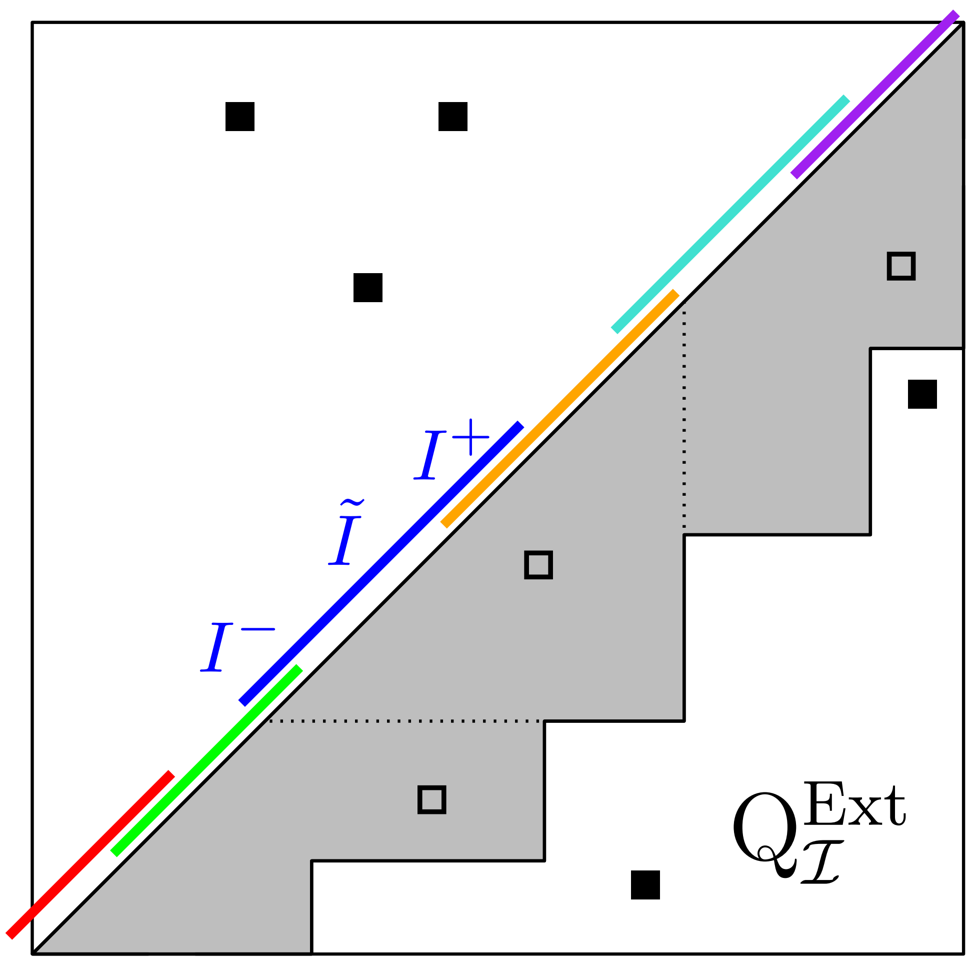
Descriptor for Mapper

Def: Given $X, f, \mathcal{I}$:

$$\mathrm{Dg}\,\mathrm{M}_f := \left(\mathrm{Ord}\,\mathrm{R}_f \setminus Q^{\mathrm{Ord}}_{\mathcal{I}}\right) \cup \left(\mathrm{Rel}\,\mathrm{R}_f \setminus Q^{\mathrm{Rel}}_{\mathcal{I}}\right) \cup \left(\mathrm{Ext}\,\mathrm{R}_f \setminus Q^{\mathrm{Ext}}_{\mathcal{I}}\right)$$

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

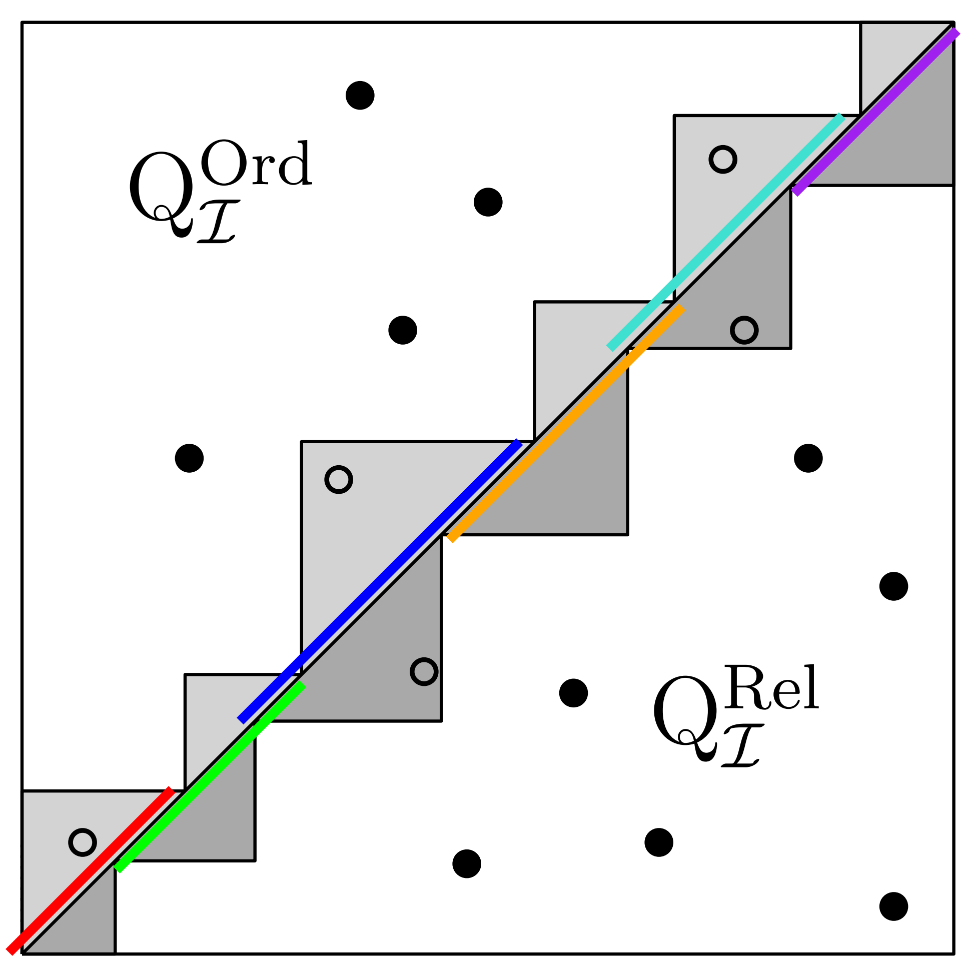


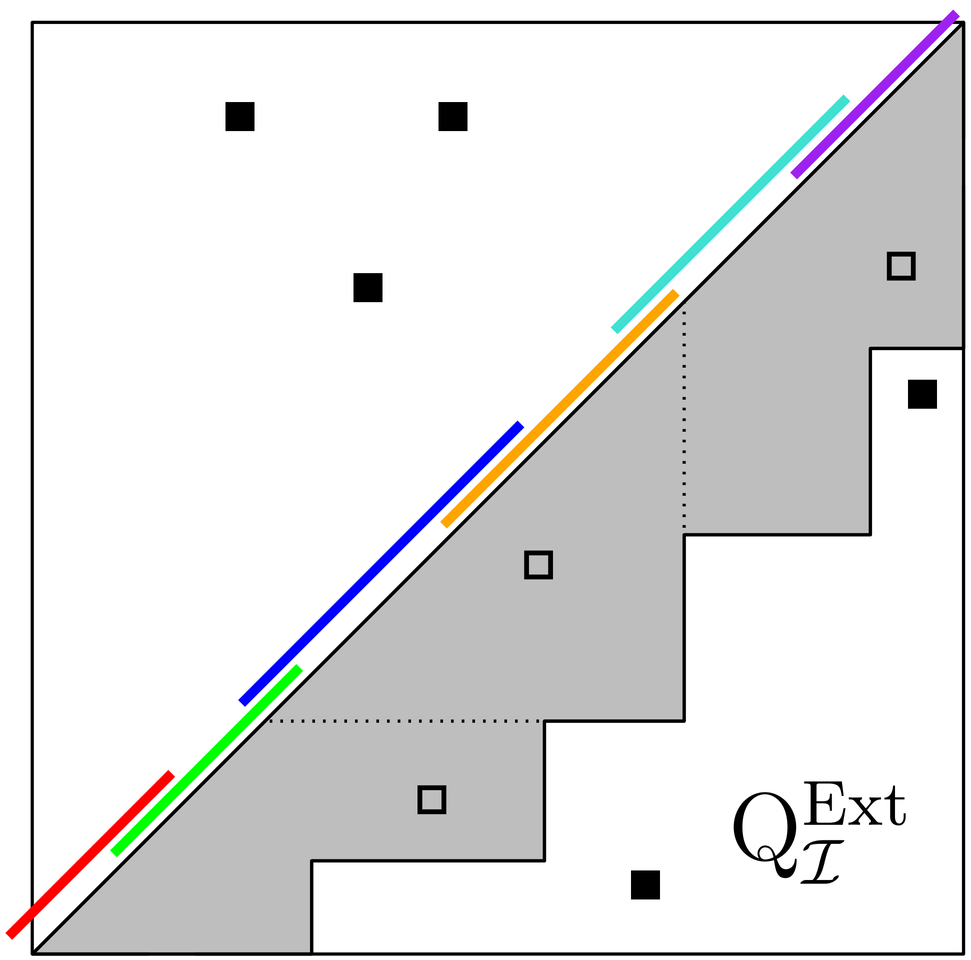


Descriptor for Mapper

Thm: [Carrière, O. 2016]
$\mathrm{Dg}\ \mathrm{M}_f$ 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



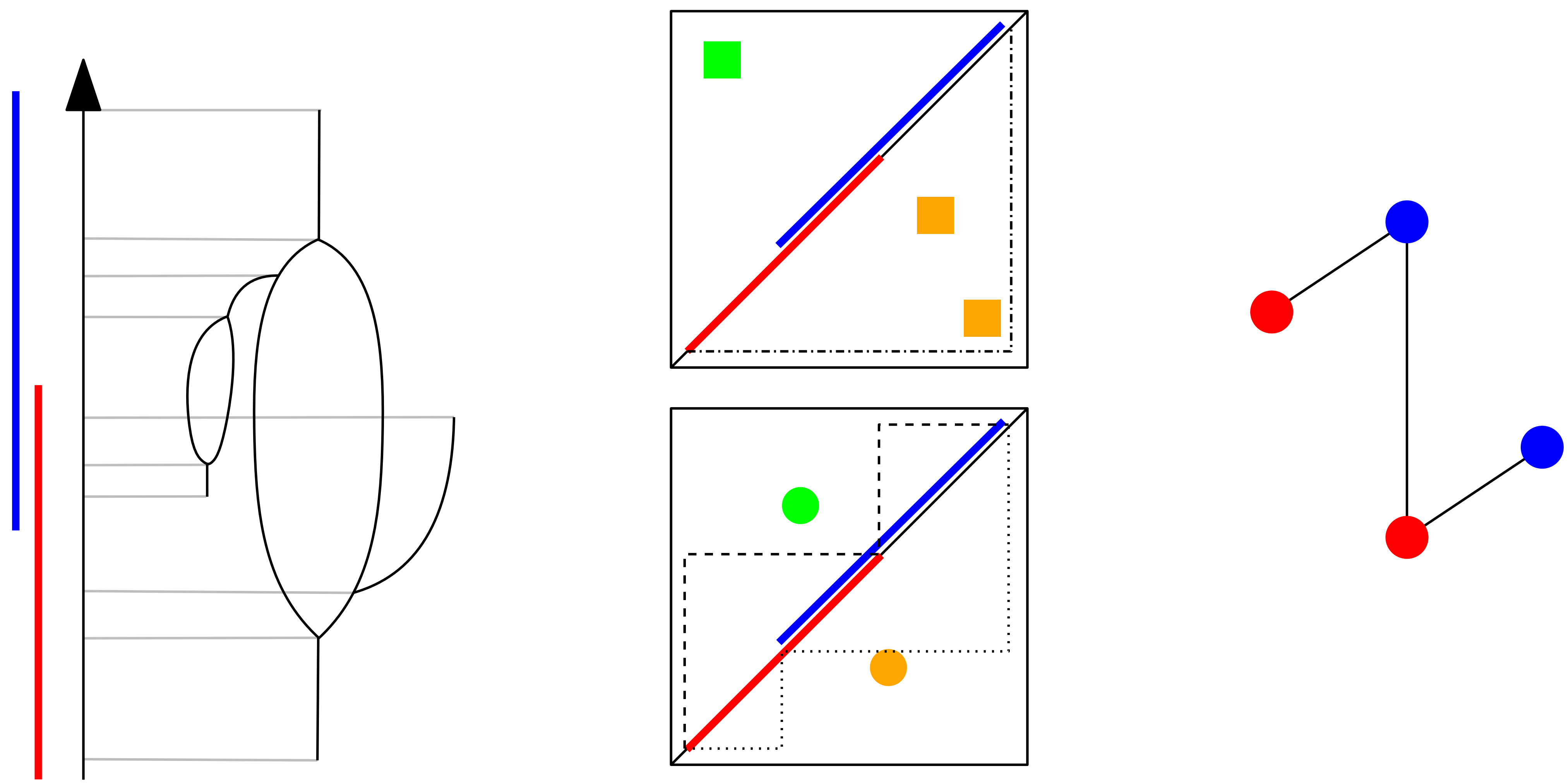


Descriptor for Mapper

Thm: [Carrière, O. 2016]

$\mathrm{Dg}\ \mathrm{M}_f$ 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

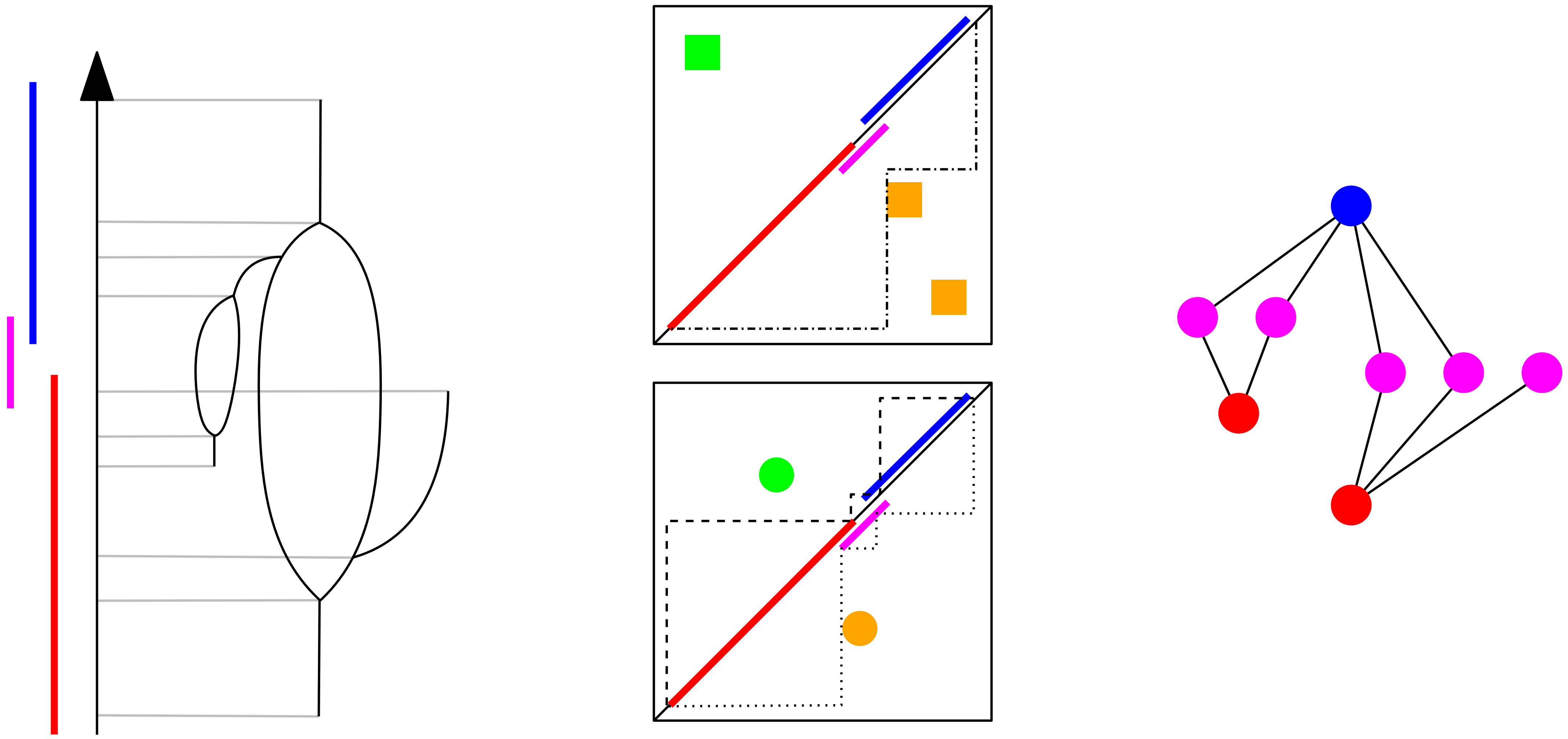


Descriptor for Mapper

Thm: [Carrière, O. 2016]

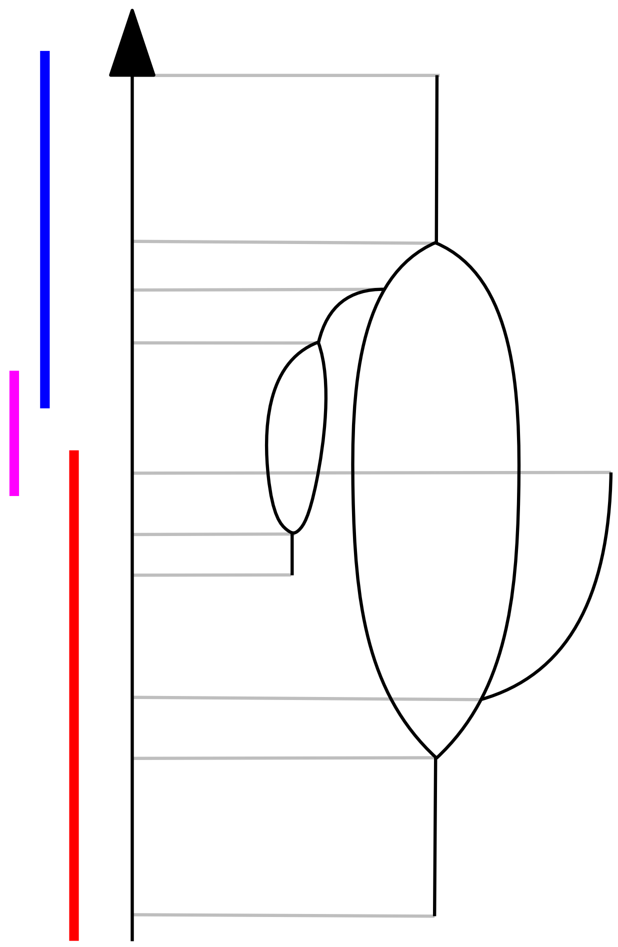
$\mathrm{Dg}\ \mathrm{M}_f$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X, \mathcal{I})$:

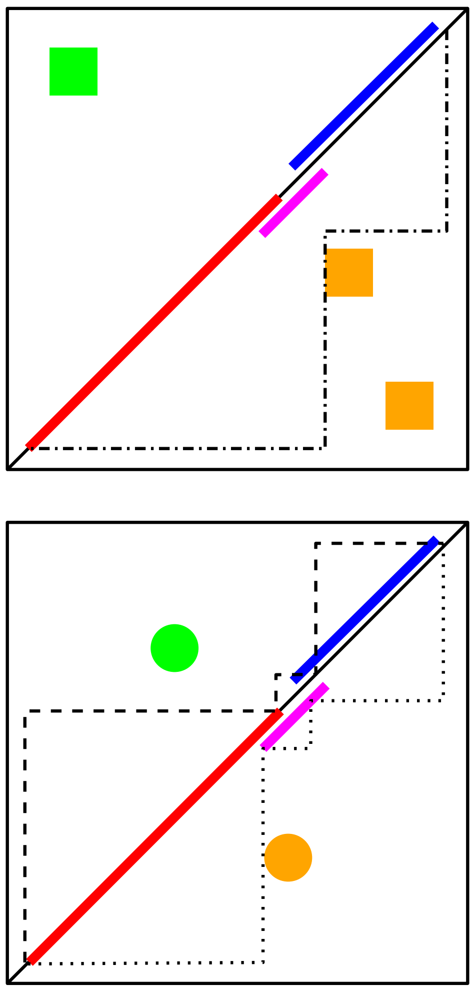
$\mathrm{Ord}_0 \longleftrightarrow$ downward branches	$\mathrm{Ext}_0 \longleftrightarrow$ trunks (cc)
$\mathrm{Rel}_1 \longleftrightarrow$ upward branches	$\mathrm{Ext}_1 \longleftrightarrow$ loops

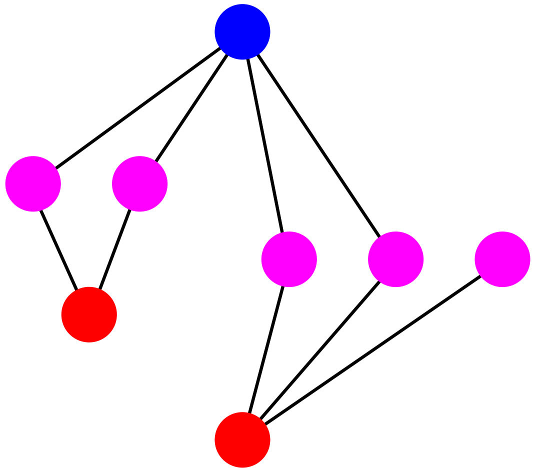


Descriptor for Mapper

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



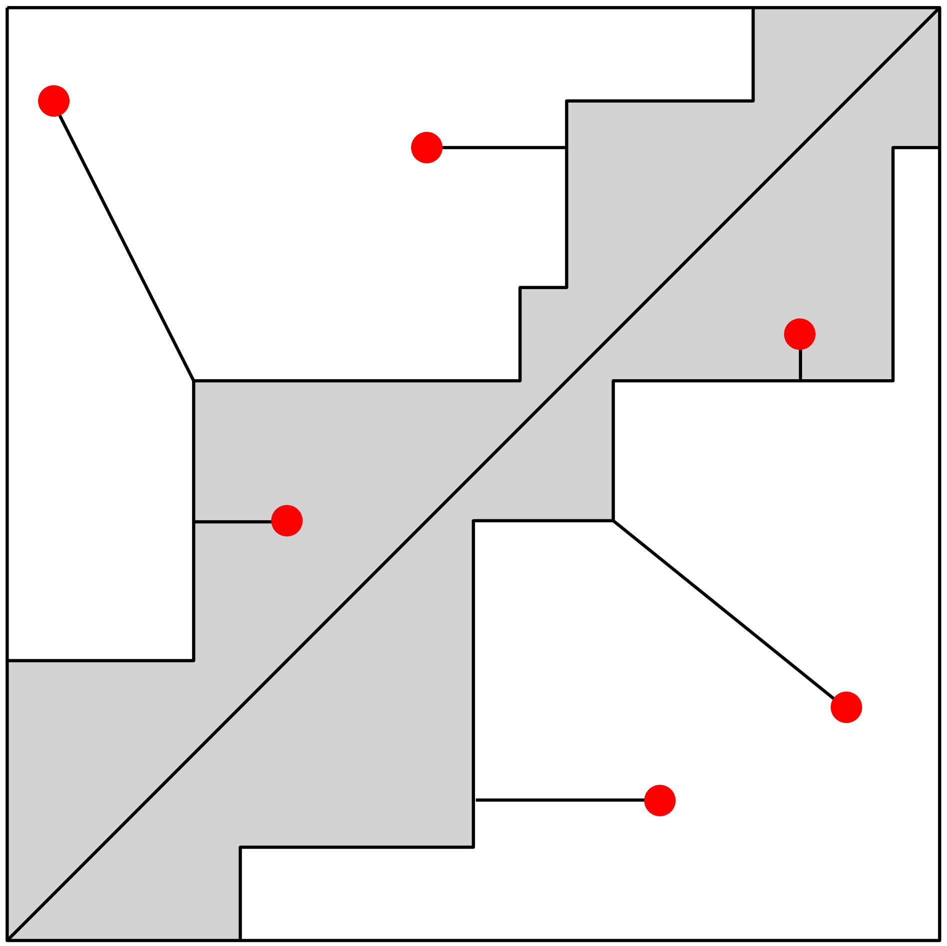




Stability of Mapper

Definition: $\mathrm{Dg}\ \mathrm{M}_f := \left(\mathrm{Ord}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ord}}\right) \cup \left(\mathrm{Rel}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Rel}}\right) \cup \left(\mathrm{Ext}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}\right)$

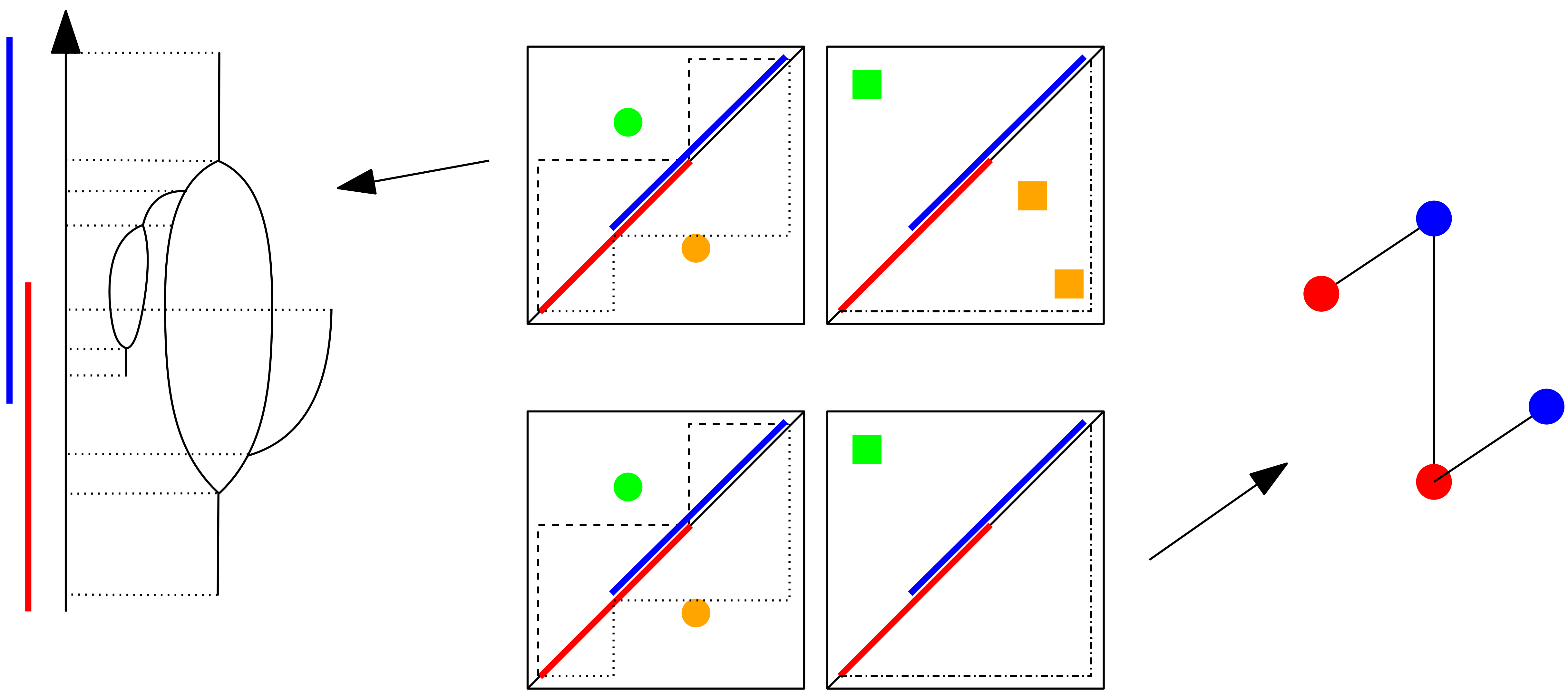
Observation: distance to staircase boundary measures (in-)stability of each feature of $\mathrm{M}_f(X, \mathcal{I})$ w.r.t. perturbations of $(X, f, \mathcal{I})$



Stability of Mapper

Definition: $\mathrm{Dg}\ \mathrm{M}_f := \left(\mathrm{Ord}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ord}}\right) \cup \left(\mathrm{Rel}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Rel}}\right) \cup \left(\mathrm{Ext}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}\right)$

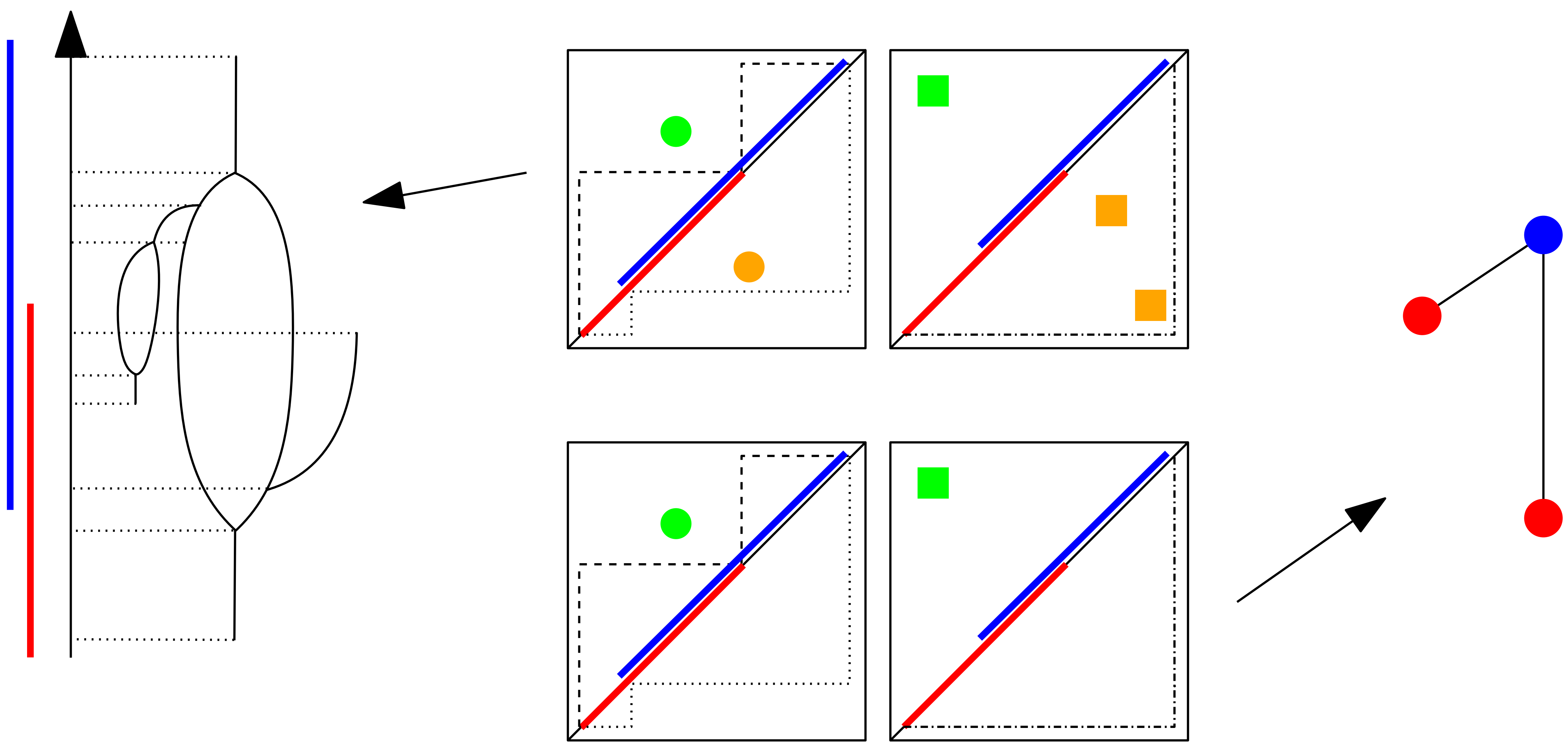
Observation: distance to staircase boundary measures (in-)stability of each feature of $\mathrm{M}_f(X, \mathcal{I})$ w.r.t. perturbations of $(X, f, \mathcal{I})$



Stability of Mapper

Definition: $\mathrm{Dg}\ \mathrm{M}_f := \left(\mathrm{Ord}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ord}}\right) \cup \left(\mathrm{Rel}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Rel}}\right) \cup \left(\mathrm{Ext}\,\mathrm{R}_f \setminus Q_{\mathcal{I}}^{\mathrm{Ext}}\right)$

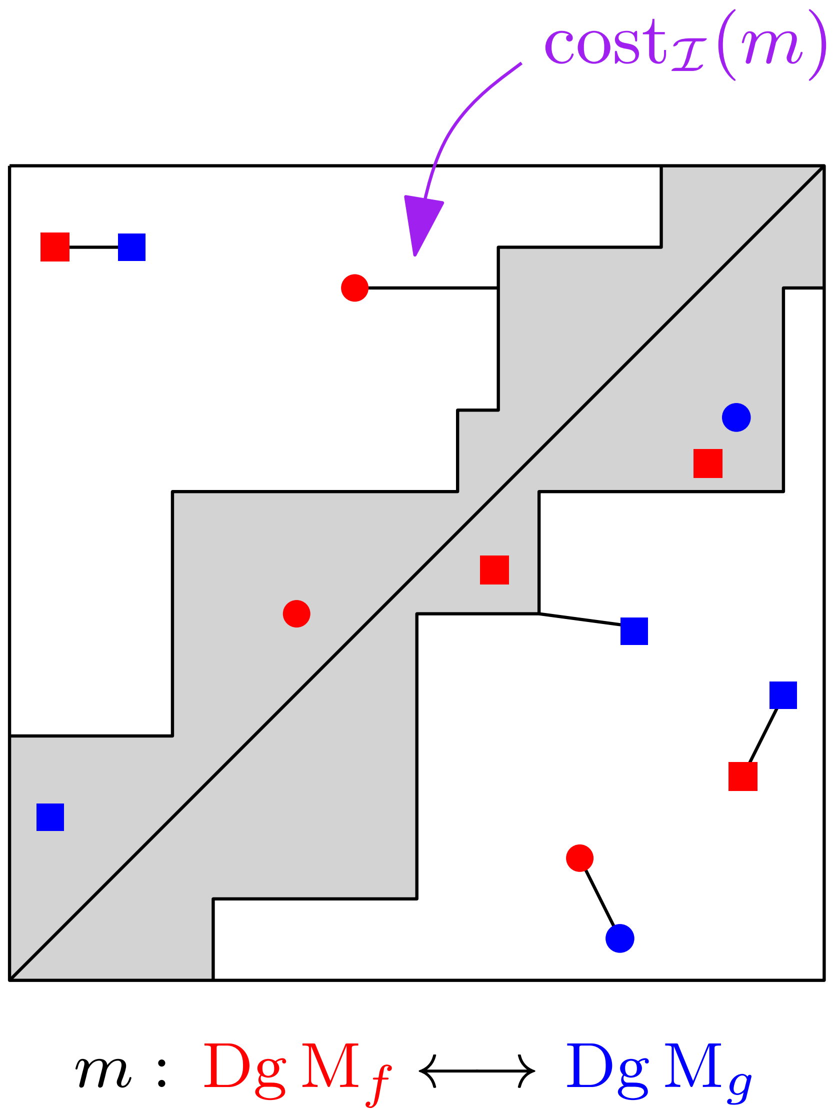
Observation: distance to staircase boundary measures (in-)stability of each feature of $\mathrm{M}_f(X, \mathcal{I})$ w.r.t. perturbations of $(X, f, \mathcal{I})$



Stability of Mapper

Definition: Given $X, \mathcal{I}$:

$$\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f,\ \mathrm{Dg}\,\mathrm{M}_g) := \inf_{m} \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$$



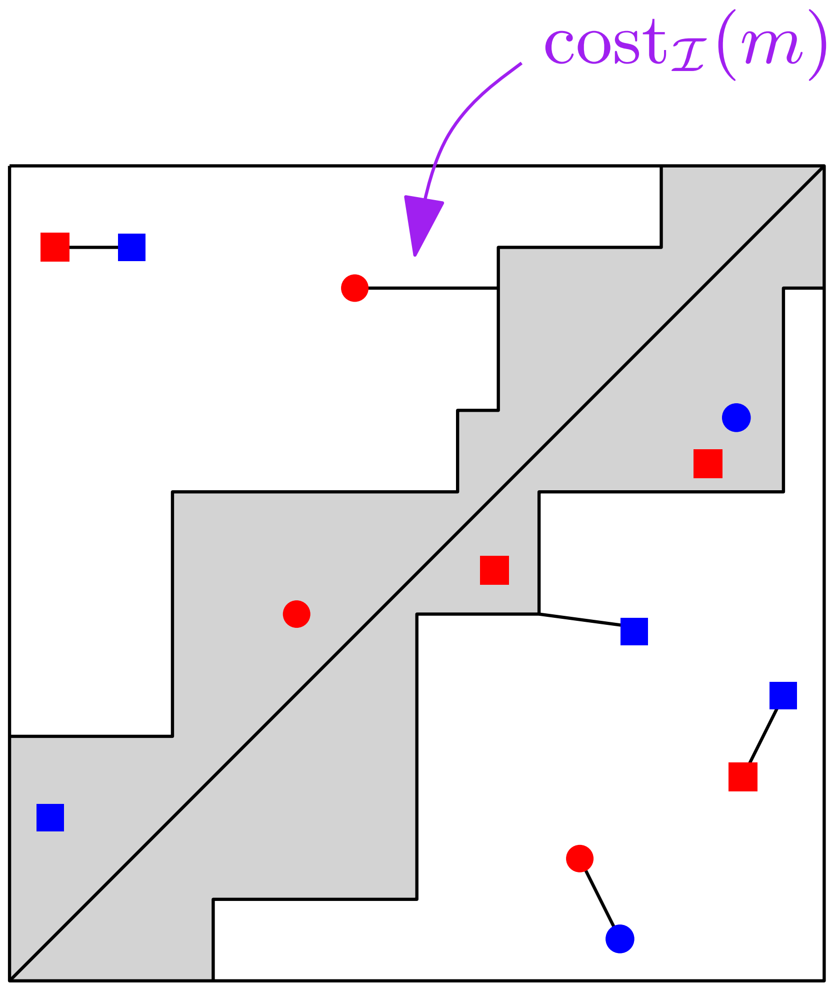
Stability of Mapper

Definition: Given $X, \mathcal{I}$:

$$\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f,\ \mathrm{Dg}\,\mathrm{M}_g) := \inf_{m} \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$$

Thm: [Carrière, O. 2016]
For any Morse-type functions $f, g : X \to \mathbb{R}$:

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



$m : \mathrm{Dg}\,\mathrm{M}_f \longleftrightarrow \mathrm{Dg}\,\mathrm{M}_g$

Stability of Mapper

Definition: Given $X, \mathcal{I}$:

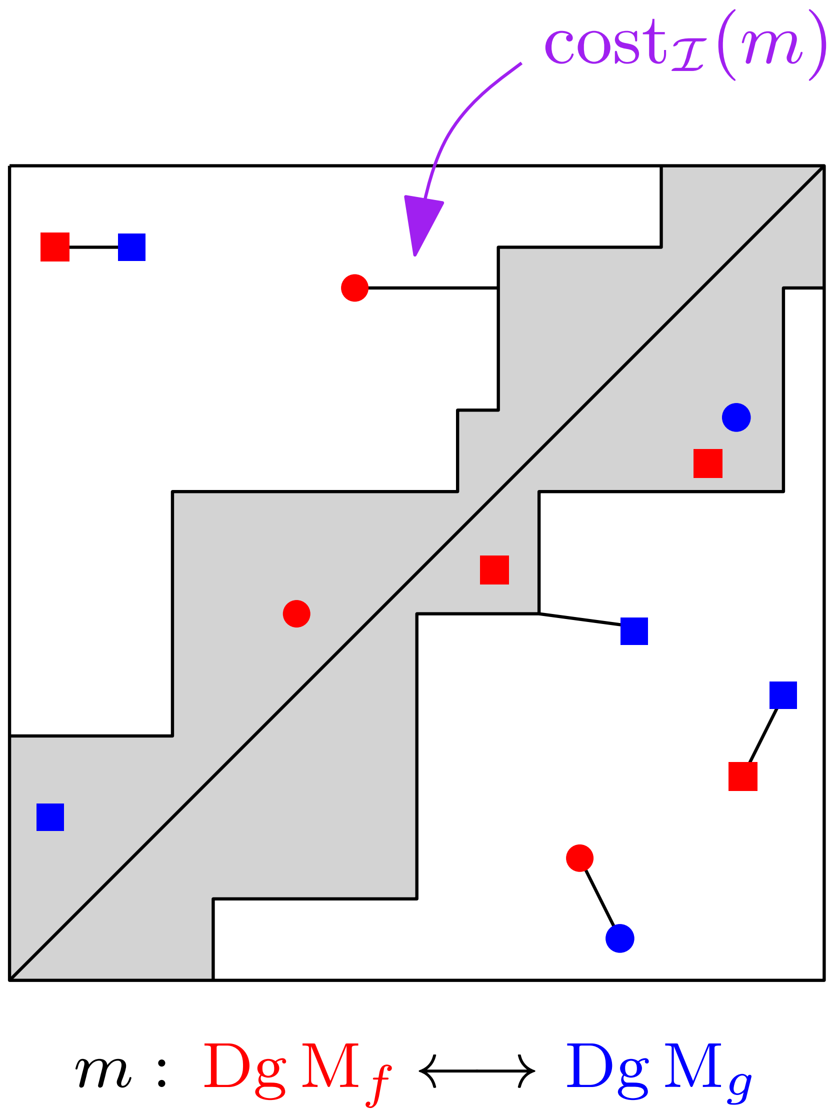
$$\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f,\ \mathrm{Dg}\,\mathrm{M}_g) := \inf_m \mathrm{cost}_{\mathcal{I}}(\mathrm{m})$$

Thm: [Carrière, O. 2016]
For any Morse-type functions $f, g : X \to \mathbb{R}$:

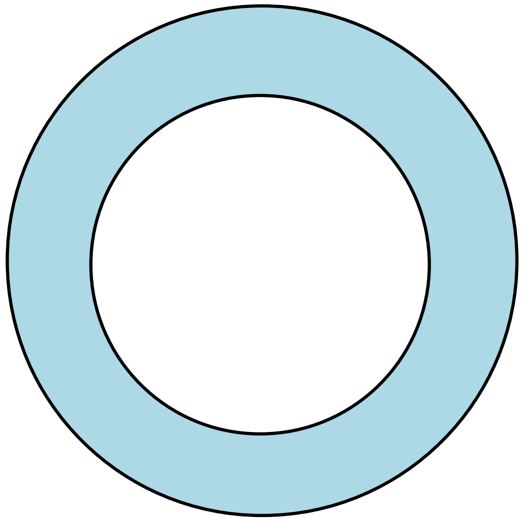
$$\mathrm{d}_{\mathcal{I}}(\mathrm{Dg}\,\mathrm{M}_f(X, \mathcal{I}),\ \mathrm{Dg}\,\mathrm{M}_g(X, \mathcal{I})) \leq \|f - g\|_\infty$$

Extensions to:

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

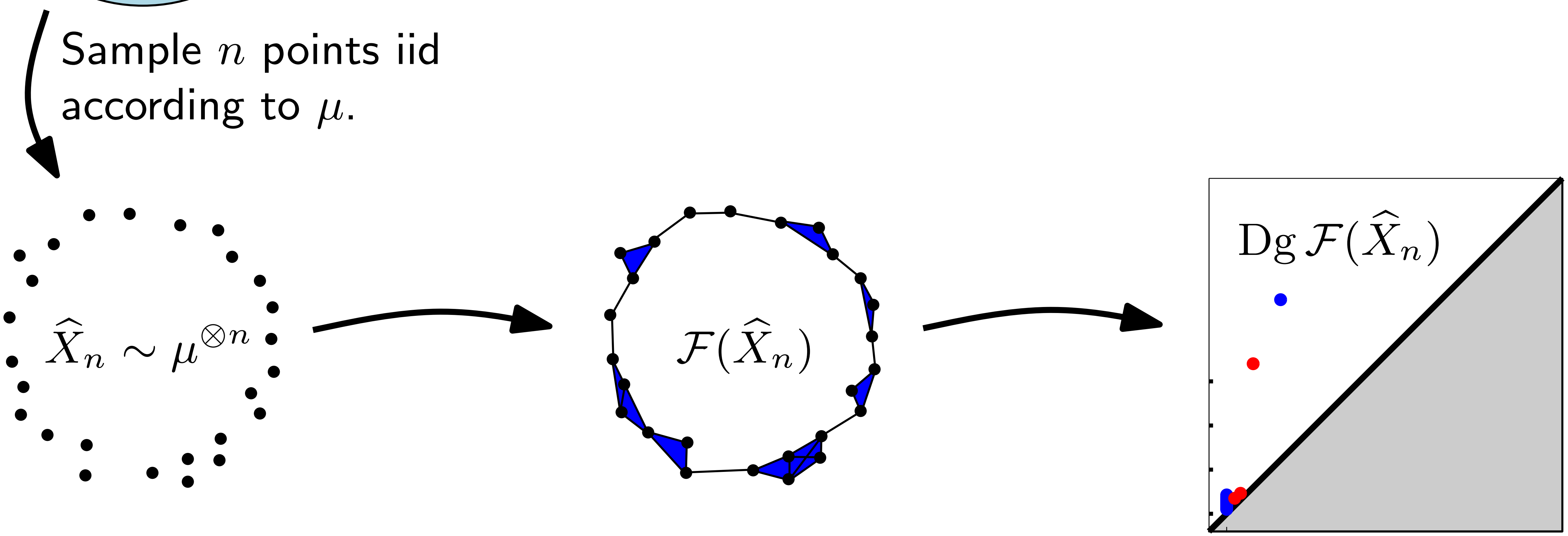


Statistics via push-forwards

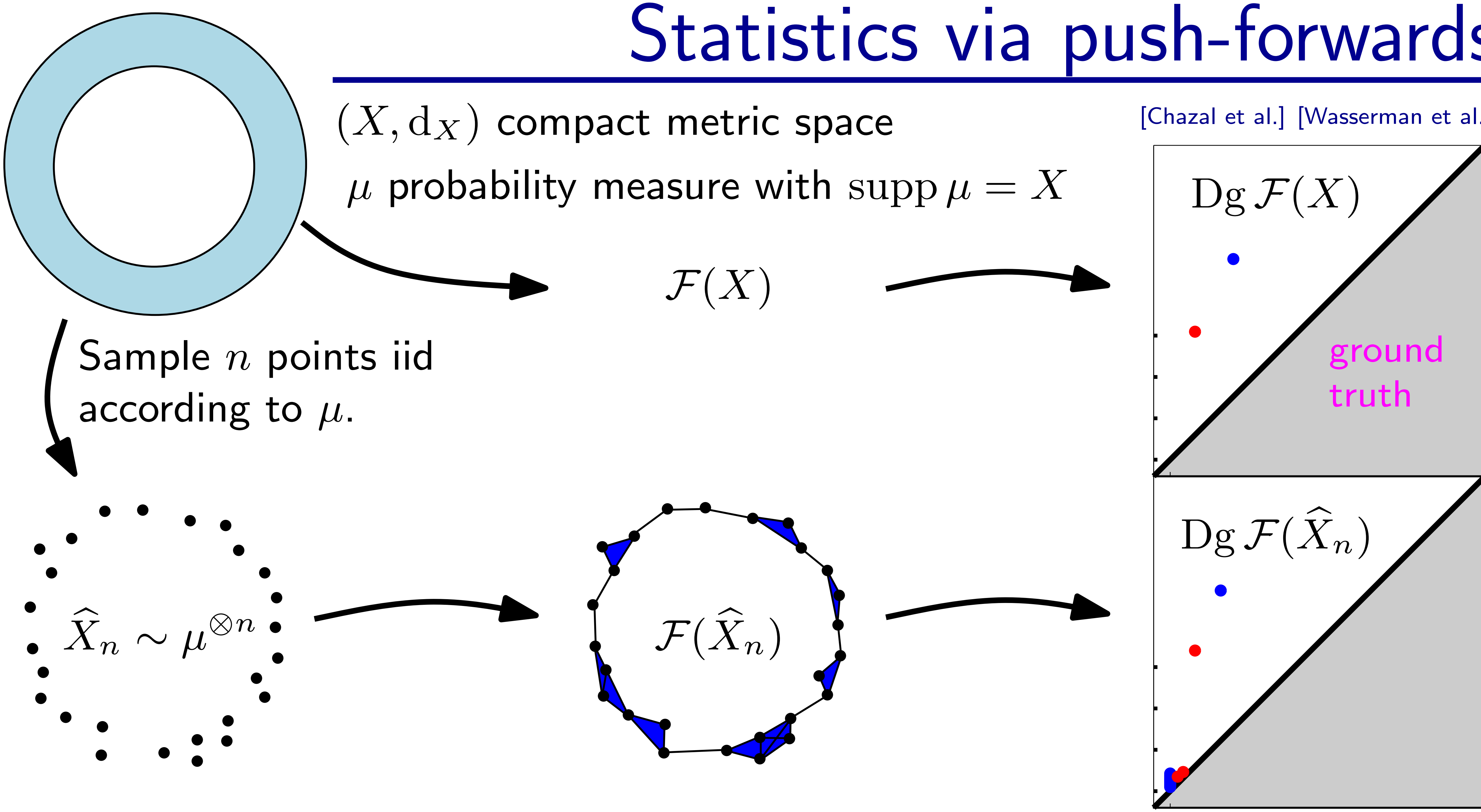


(X, d_X) compact metric space [Chazal et al.] [Wasserman et al.]

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



Statistics via push-forwards



Questions:

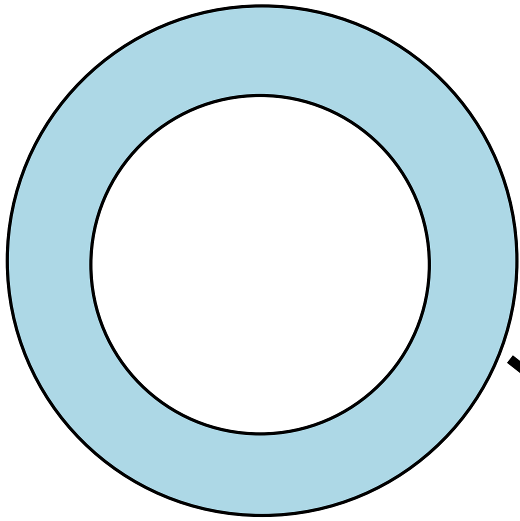
- Statistical properties of the estimator $\operatorname{Dg} \mathcal{F}(\widehat{X}_n)$?
- Convergence to the ground truth $\operatorname{Dg} \mathcal{F}(X)$? Deviation bounds?

Statistics via push-forwards

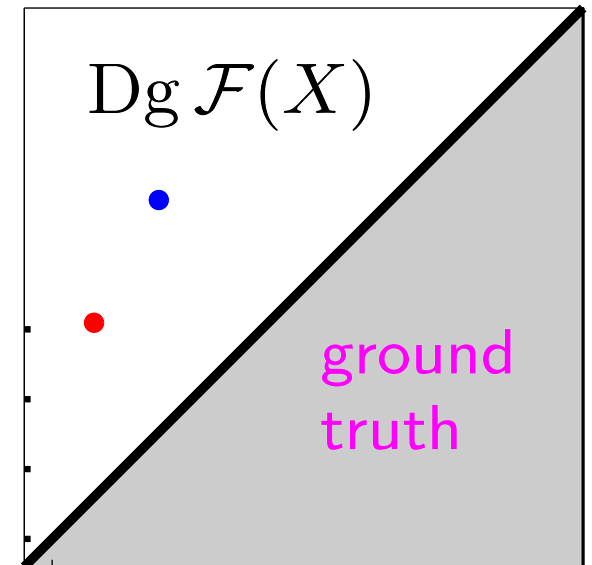
[Chazal et al.] [Wasserman et al.]

(X, d_X) compact metric space

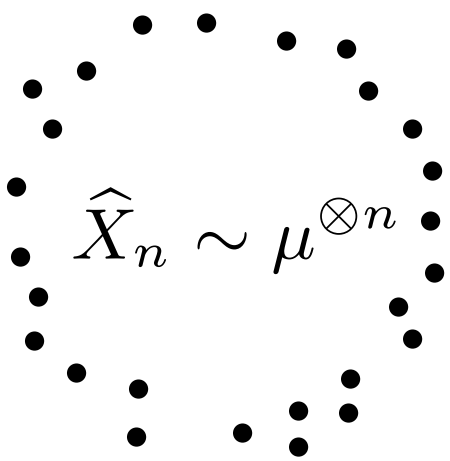
μ probability measure with $\operatorname{supp} \mu = X$

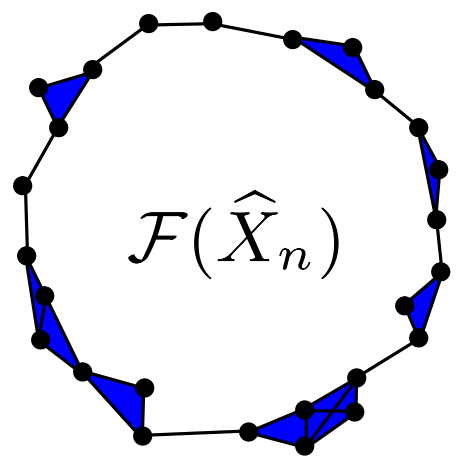


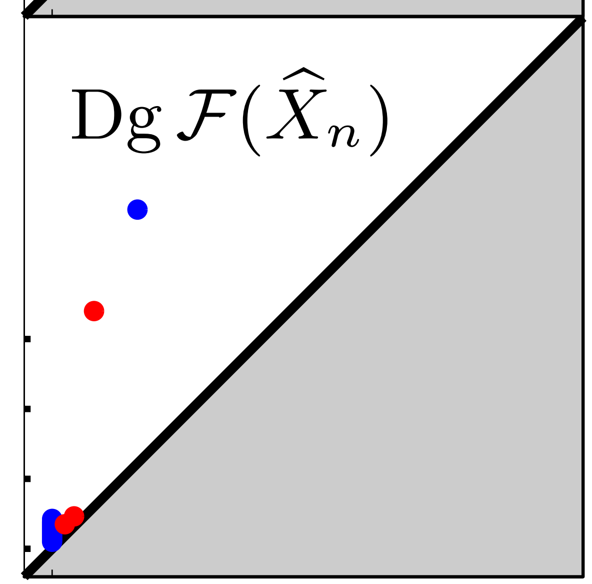
$\mathcal{F}(X)$



Sample n points iid according to μ.





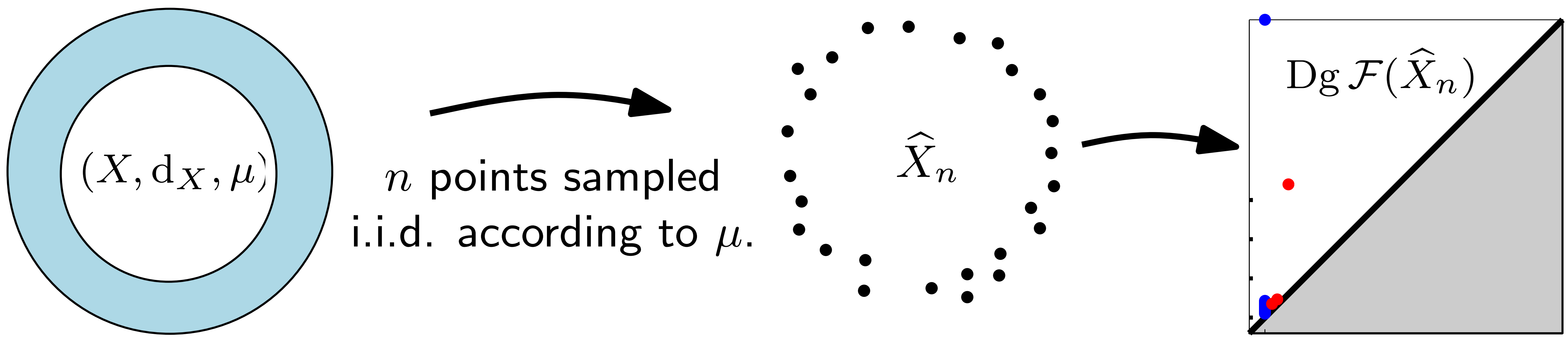


Stability thm: $\mathrm{d}_B(\mathrm{Dg}\, \mathcal{F}(\widehat{X}_n), \mathrm{Dg}\, \mathcal{F}(X)) \leq 2\mathrm{d}_{\mathrm{H}}(\widehat{X}_n, X)$ [Chazal et al. 2009/13]

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

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

Deviation inequality / rate of convergence



Hyp: μ is (a, b)-**standard**:

$$\forall x \in X,\ \forall r > 0,\ \mu(B(x,r)) \geq \min(ar^b, 1)$$

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

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

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

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

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

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

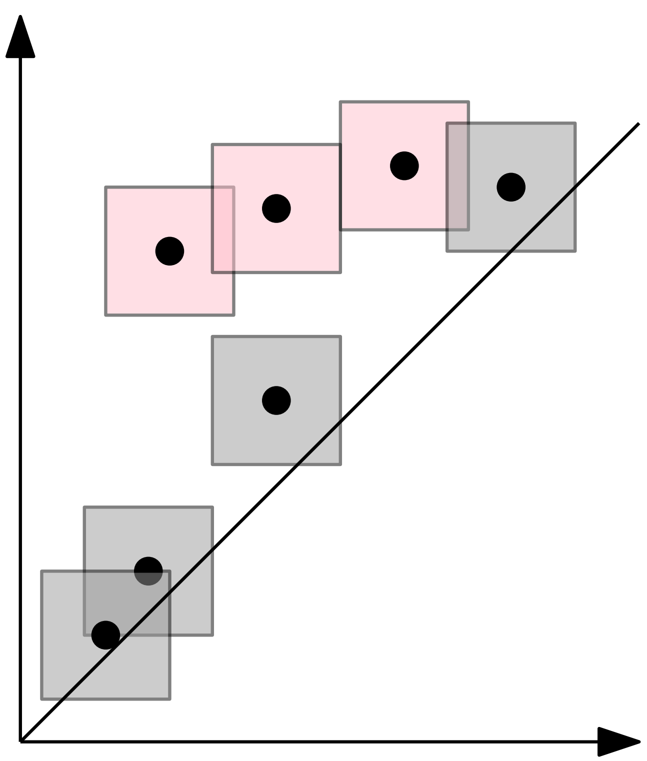
Confidence regions

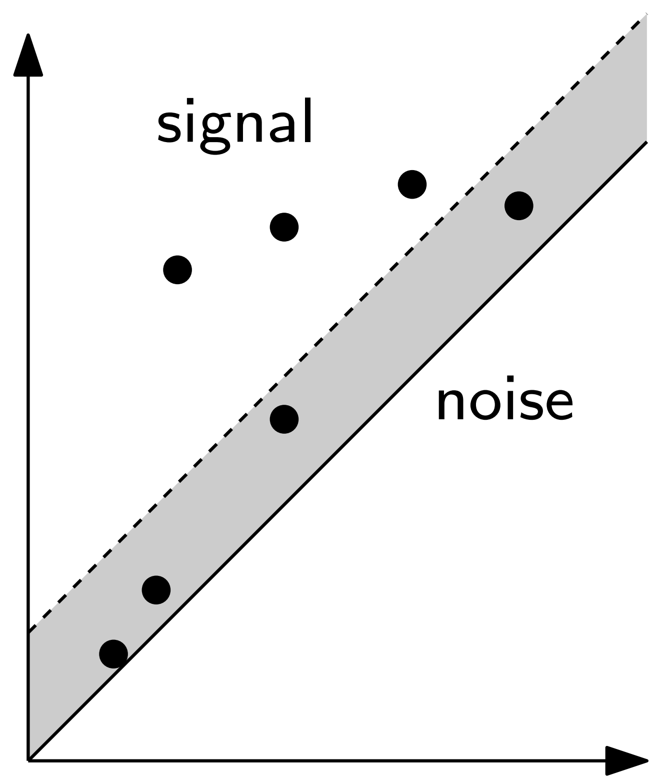
Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

$\to$ confidence region: d_B-ball of radius $c_n(\alpha)$ around $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$





Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

Note: we already have an inequality of this kind but...

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

unknown

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

Bootstrap: (ideally)

- draw $X^* = X_1^*, \cdots, X_n^*$ iid from $\mu_{\widehat{X}_n}$ (empirical measure on $\widehat{X}_n$)
- compute $d^* = \mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(X^*), \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)\right)$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^N I(\sqrt{n}\, d_i^* \geq t)$

Principle [Efron 1979]: variations of $\mathrm{Dg}\,\mathcal{F}(X^*)$ around $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ are same as variations of $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ around $\mathrm{Dg}\,\mathcal{F}(X)$.

Note: requires some conditions on (X, d_X, μ), hence the $\sqrt{n}$.

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

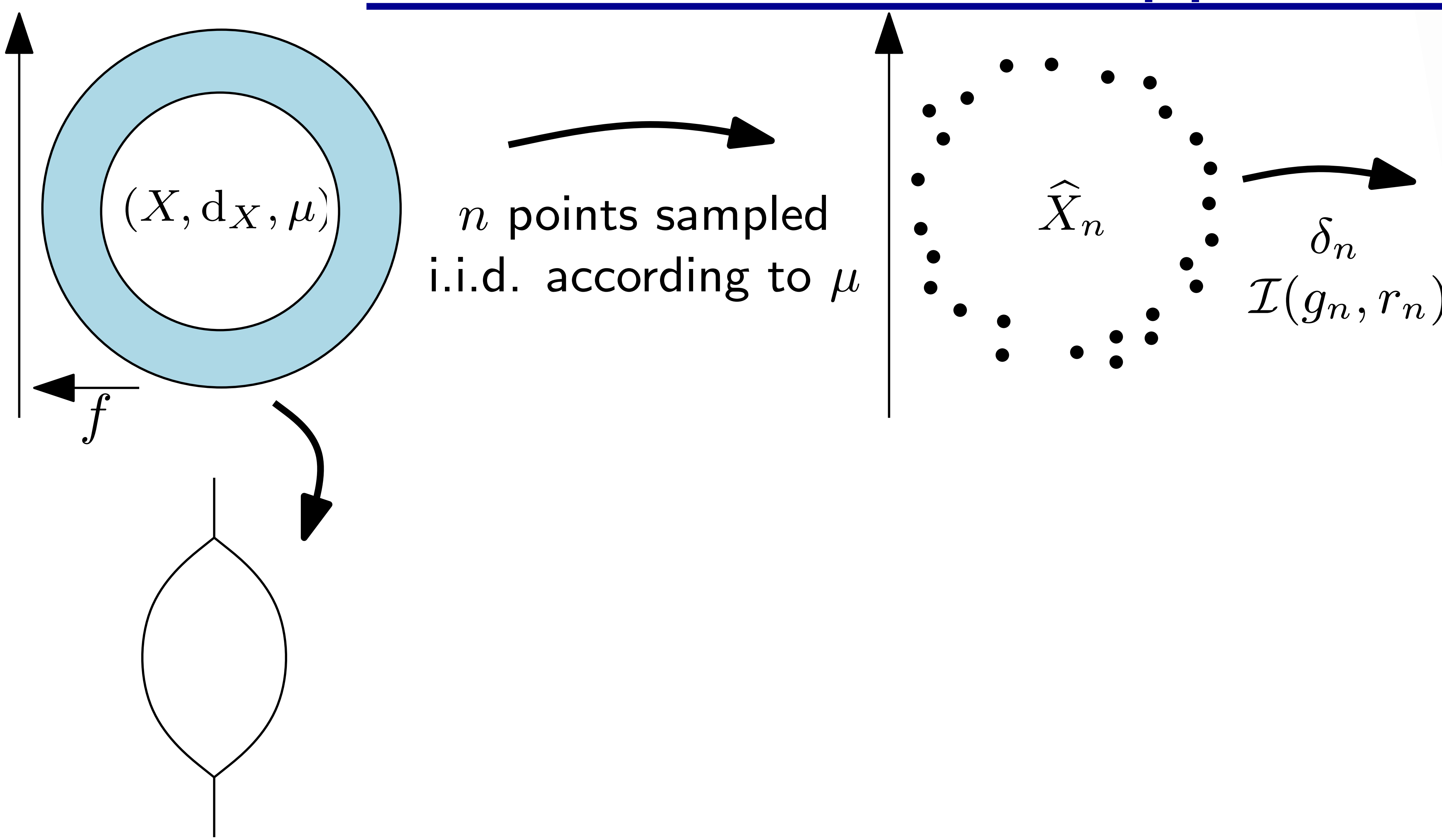
Bootstrap: (in fact)

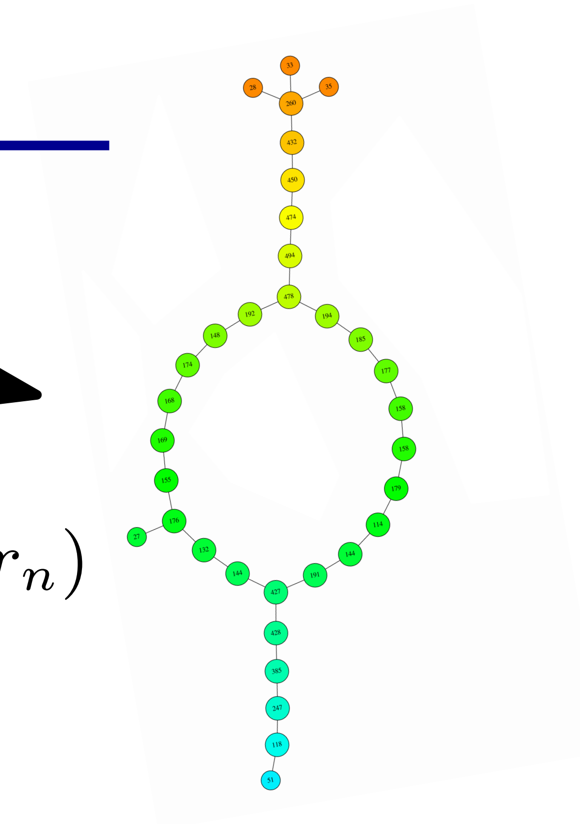
- draw $X^* = X_1^*, \cdots, X_n^*$ iid from $\mu_{\widehat{X}_n}$ (empirical measure on $\widehat{X}_n$)
- compute $d^* = \cancel{\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(X^*), \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)\right)}$ $\mathrm{d}_{\mathrm{H}}(X^*, \hat{X}_n)$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^N I(\sqrt{n}\, d_i^* \geq t)$

Theorem [Balakrishnan et al. 2013] [Chazal et al. 2014]:

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_B\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > \frac{q_\alpha}{\sqrt{n}}\right) \leq \alpha.$$

Statistics for Mapper

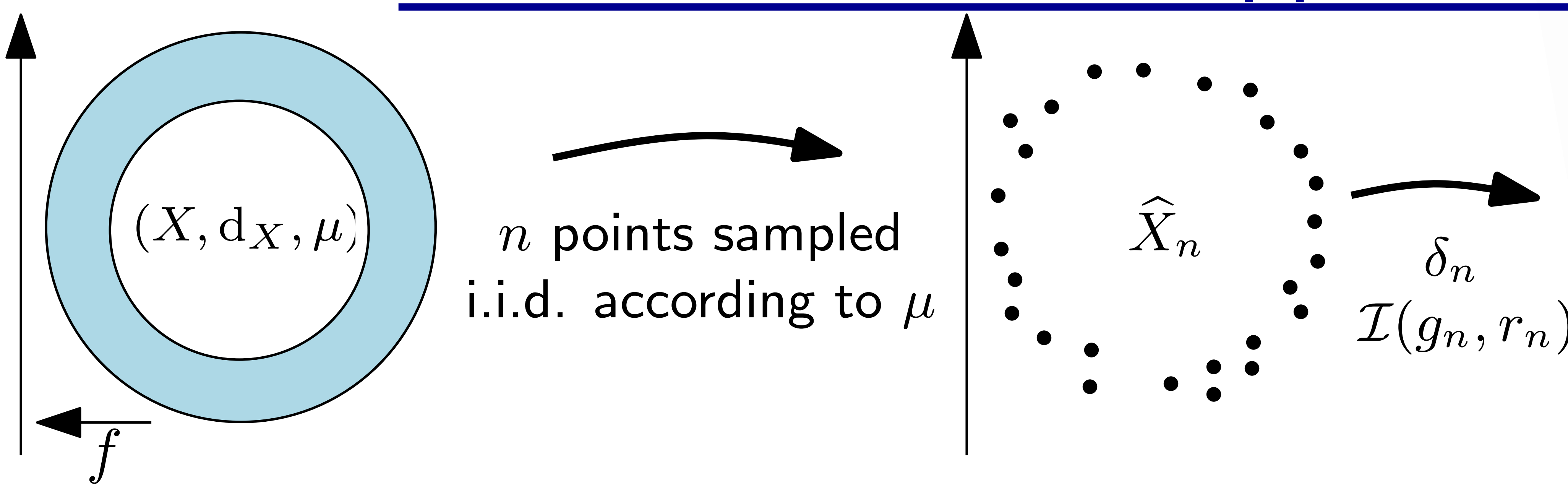


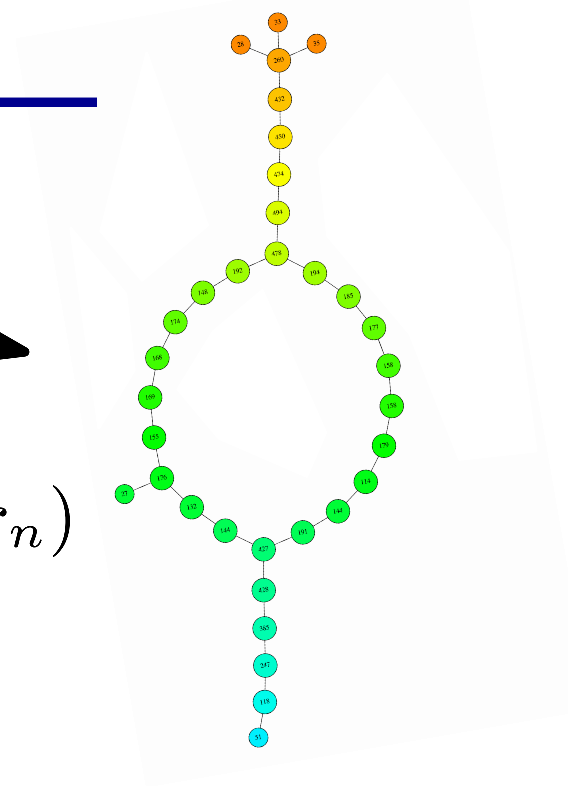


Questions:

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

Statistics for Mapper





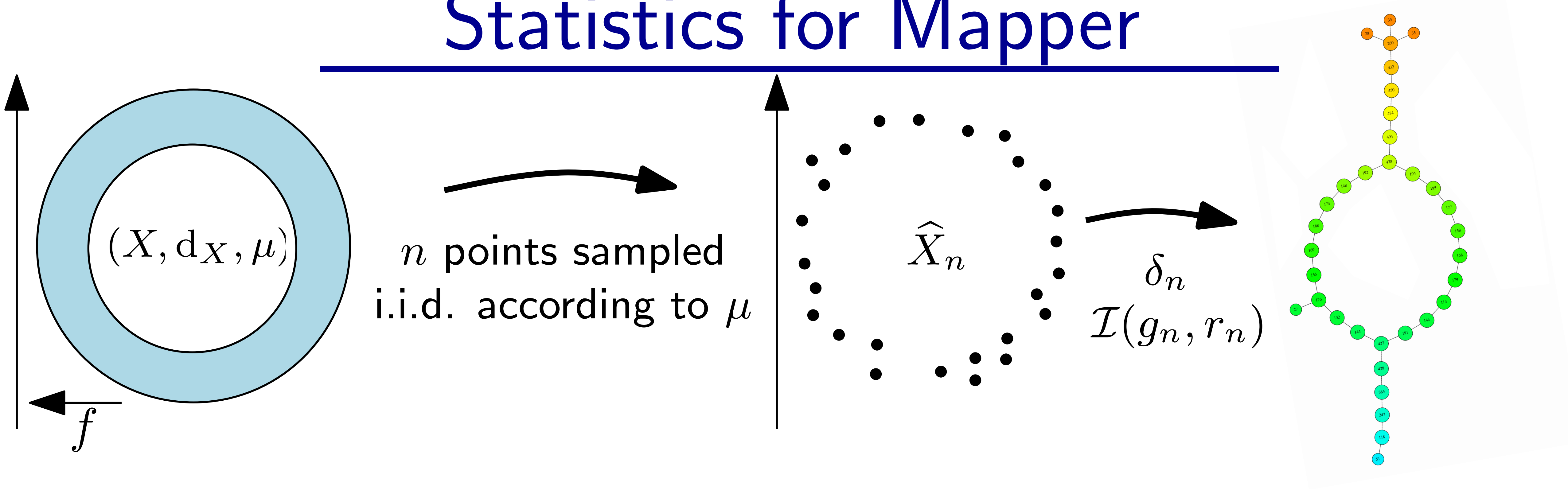
Theorem [Carrière, Michel, O. 2017]:

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

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

where C depends only on a, b, c. Moreover, the estimator $\mathrm{Dg}\,\mathrm{M}_f(\widehat{X}_n, \delta_n, \mathcal{I}(g_n, r_n))$ is **minimax optimal** (up to $\log n$ factors) on the space $\mathcal{P}$ of (a, b)-standard probability measures on X.

Statistics for Mapper



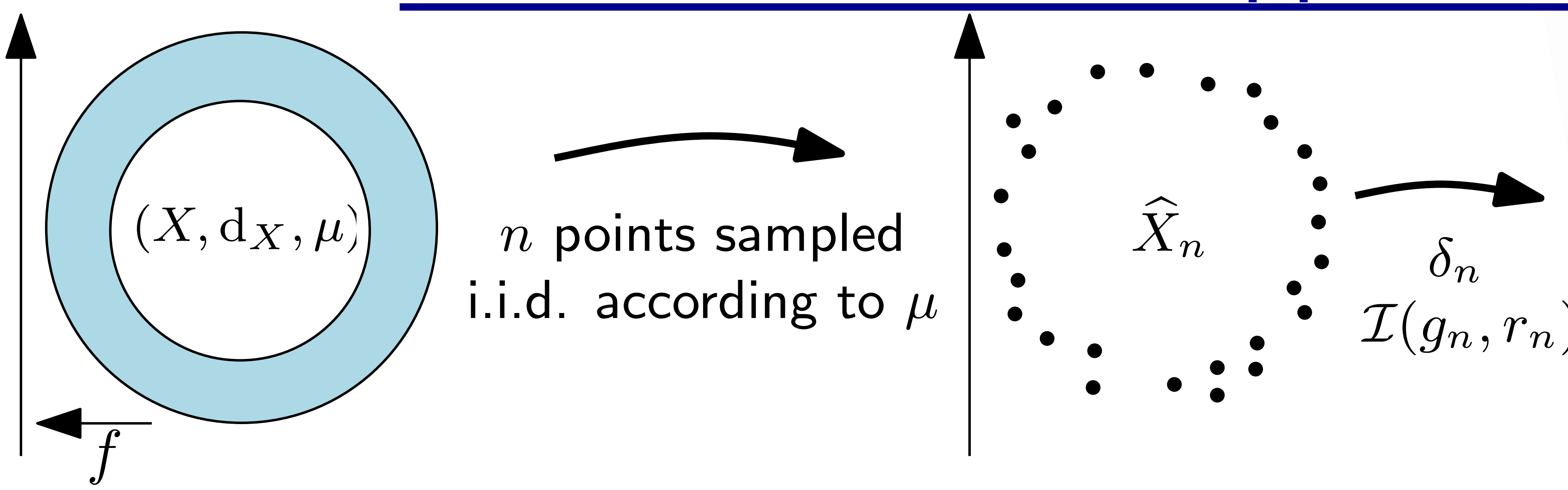
Theorem [Carrière, Michel, O. 2017]:

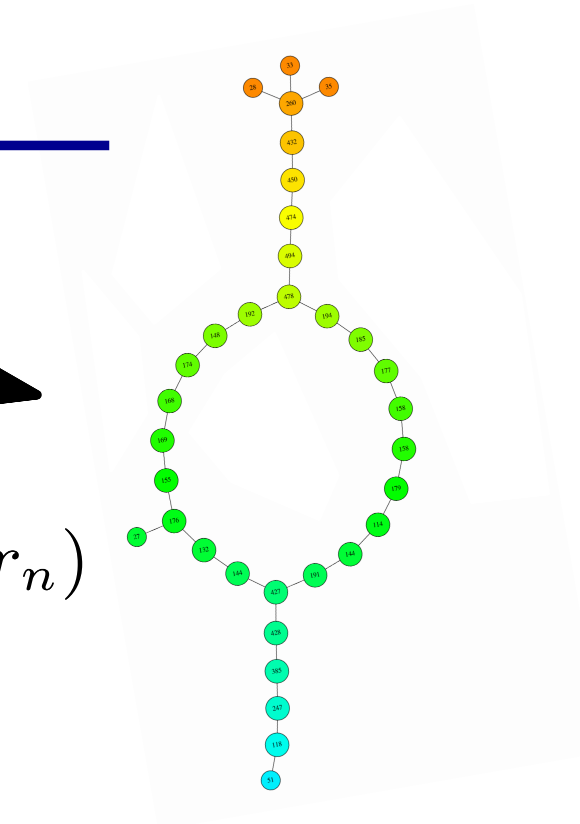
If μ is (a, b)-standard, f is c-Lipschitz, $\delta_n = 4\left(\frac{2\log n}{\boxed{a}n}\right)^{1/\boxed{b}}$, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, then $\forall \varepsilon > 0$:

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

where C depends only on a, b, c. Moreover, the estimator $\mathrm{Dg}\,\mathrm{M}_f(\widehat{X}_n, \delta_n, \mathcal{I}(g_n, r_n))$ is **minimax optimal** (up to $\log n$ factors) on the space $\mathcal{P}$ of (a, b)-standard probability measures on X.

Statistics for Mapper

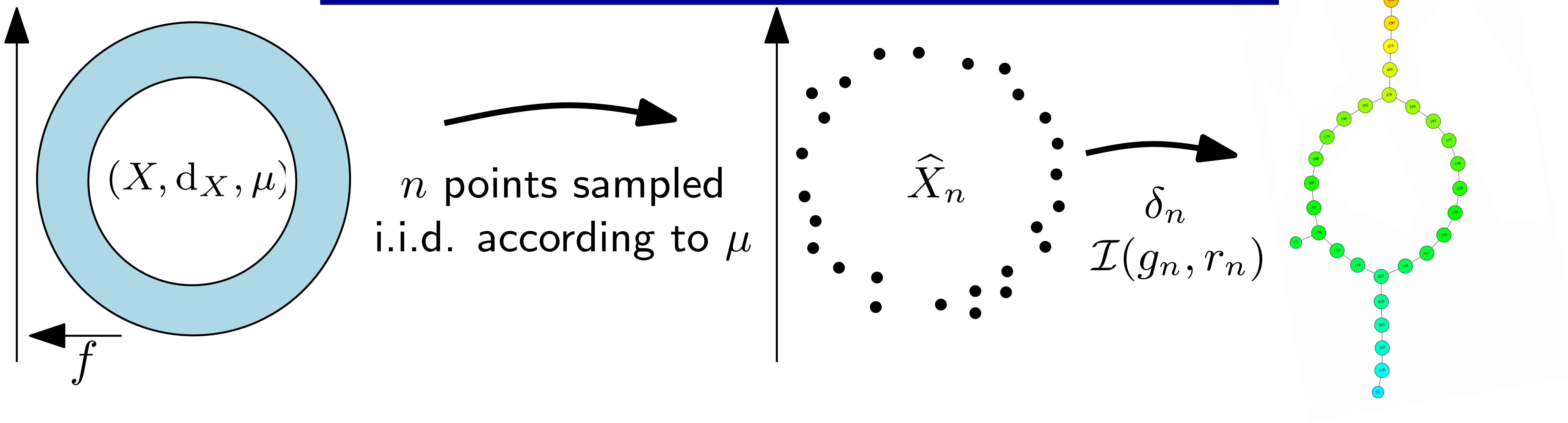




$\rightarrow$ subsampling to tune δ_n: let $\beta > 0$ and take $(s_n) = \frac{n}{\log(n)^{1+\beta}}$

$\delta_n := \mathrm{d_H}(\hat{X}_n^{s(n)}, \hat{X}_n)$ 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 $(s_n) = \frac{n}{\log(n)^{1+\beta}}$

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

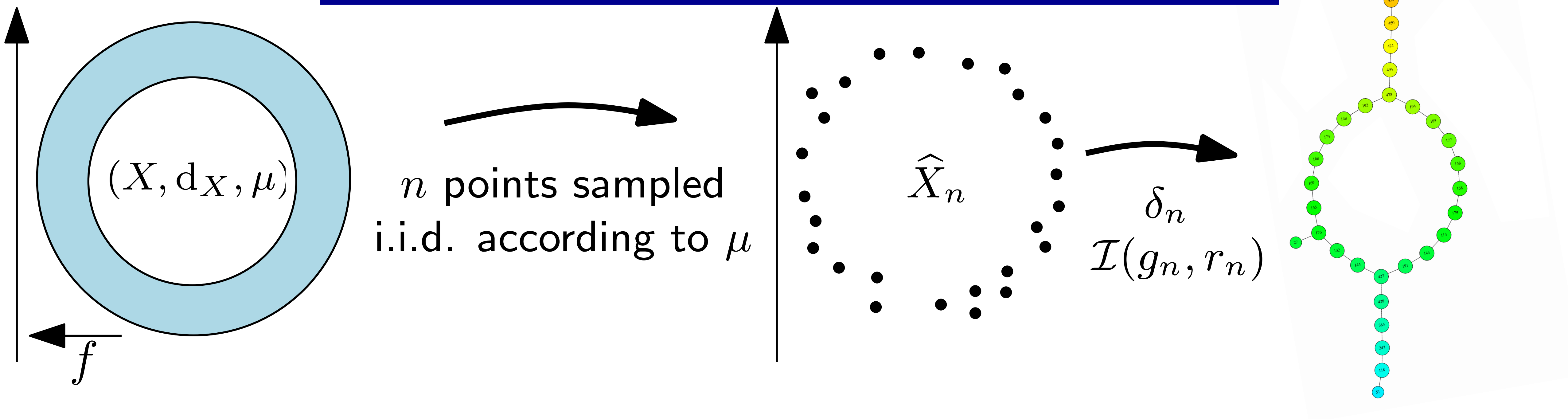
Theorem [Carrière, Michel, O. 2016]:

If μ is (a,b)-standard, f is c-Lipschitz, δ_n as above, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, then $\forall \varepsilon > 0$:

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

where C depends only on a, b, c.

Statistics for Mapper



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

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

Theorem [Carrière, Michel, O. 2016]:

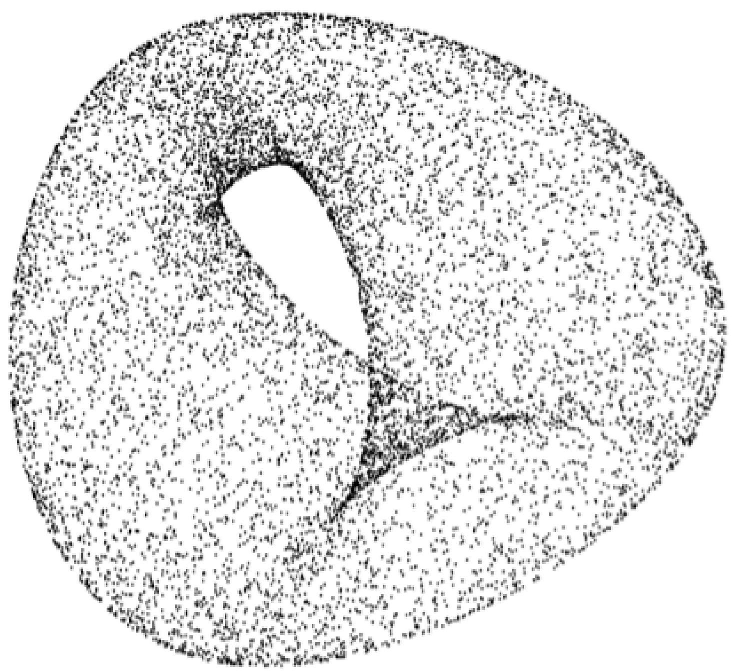
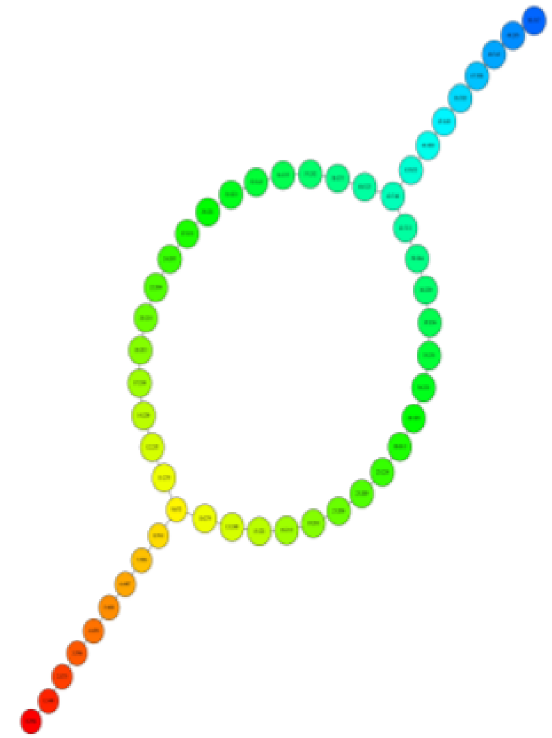
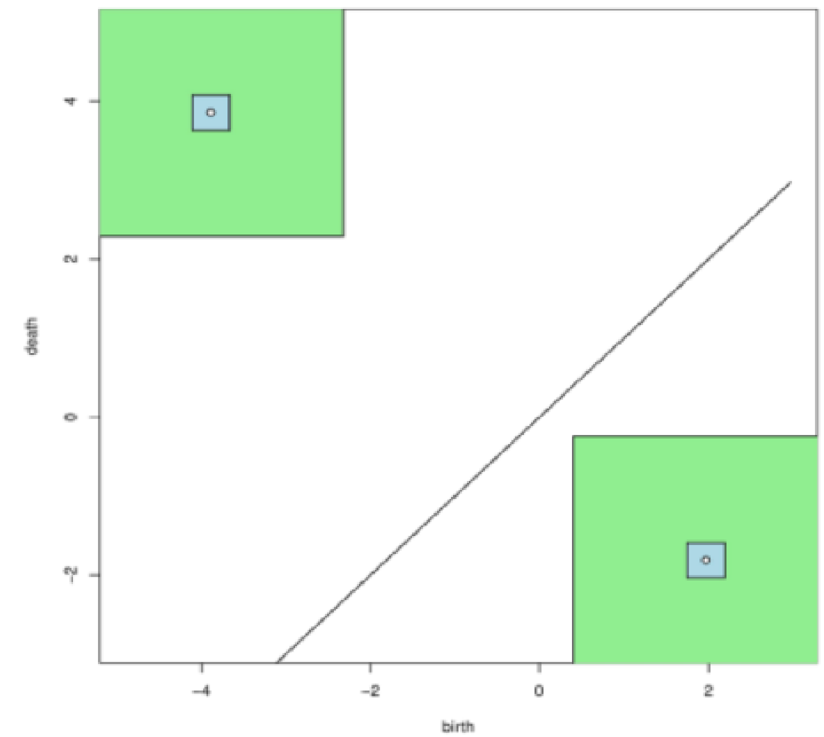
If μ is (a, b)-standard, f is c-Lipschitz, δ_n as above, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, then $\forall \varepsilon > 0$:

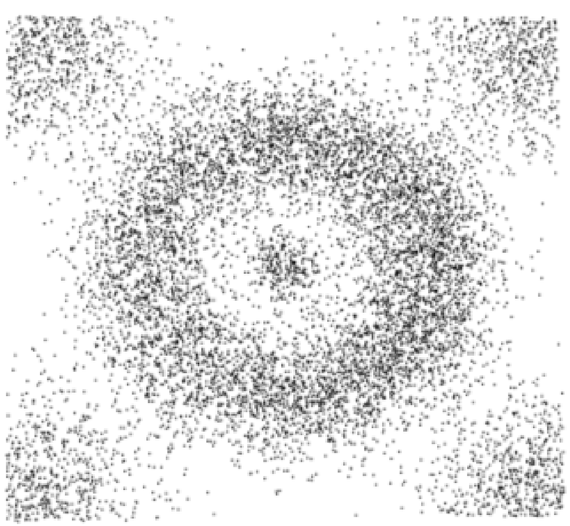
$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{Dg}\,\mathrm{M}_f(\widehat{X}_n, \delta_n, \mathcal{I}(g_n, r_n)),\ \mathrm{Dg}\,\mathrm{R}_f(X)\right)\right] \leq C\,\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b},$$

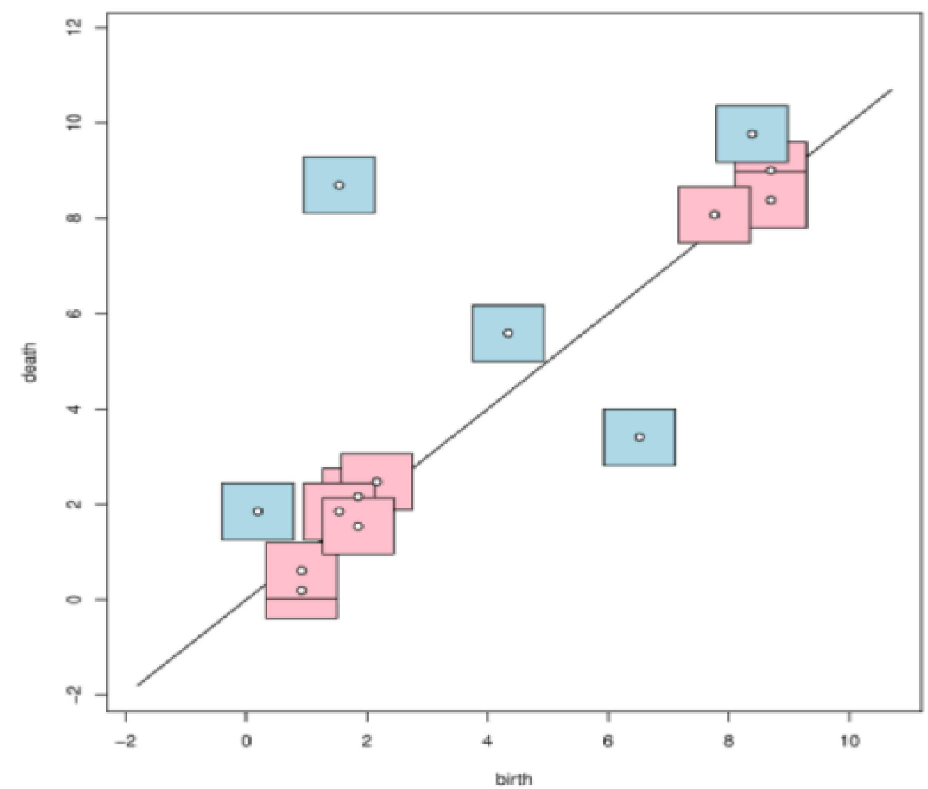
where C depends only on a, b, c. $\rightarrow$ iterate subsampling to get confidence regions

Experiments

confidence level: 85%

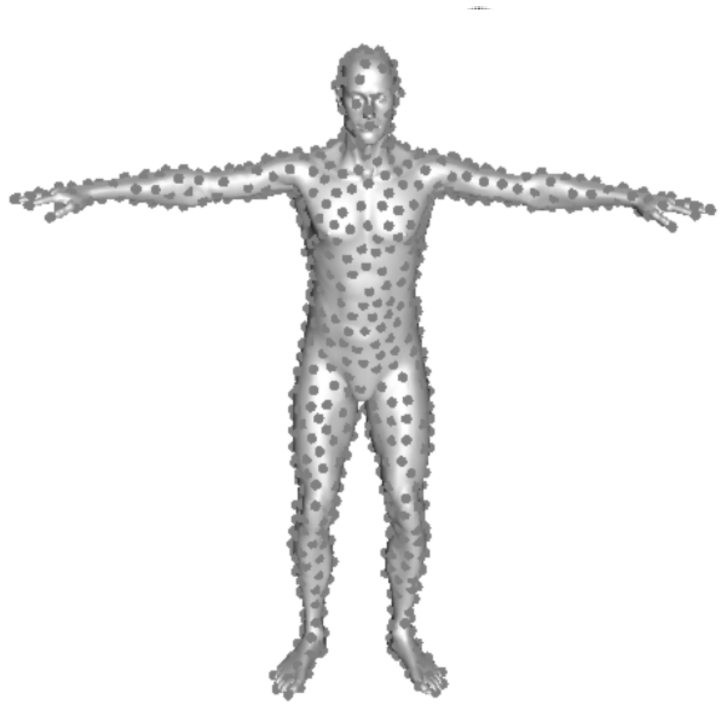




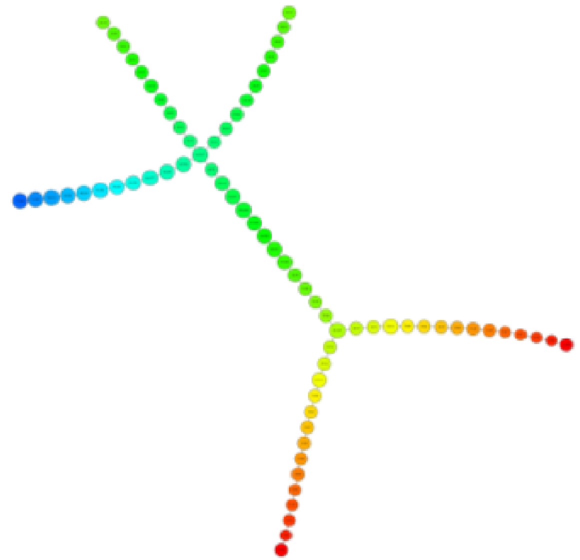


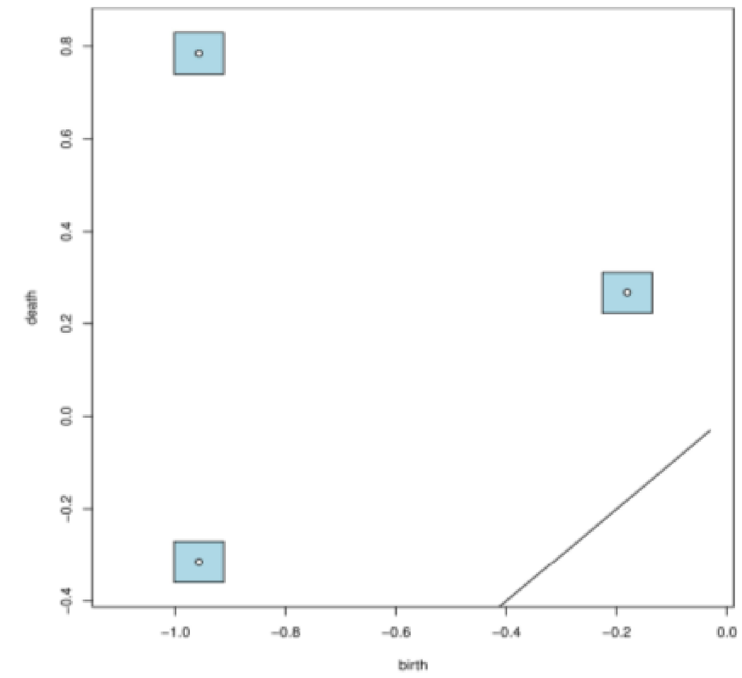


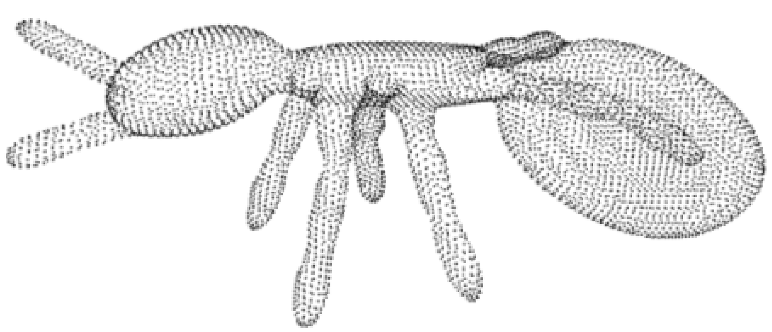
Experiments

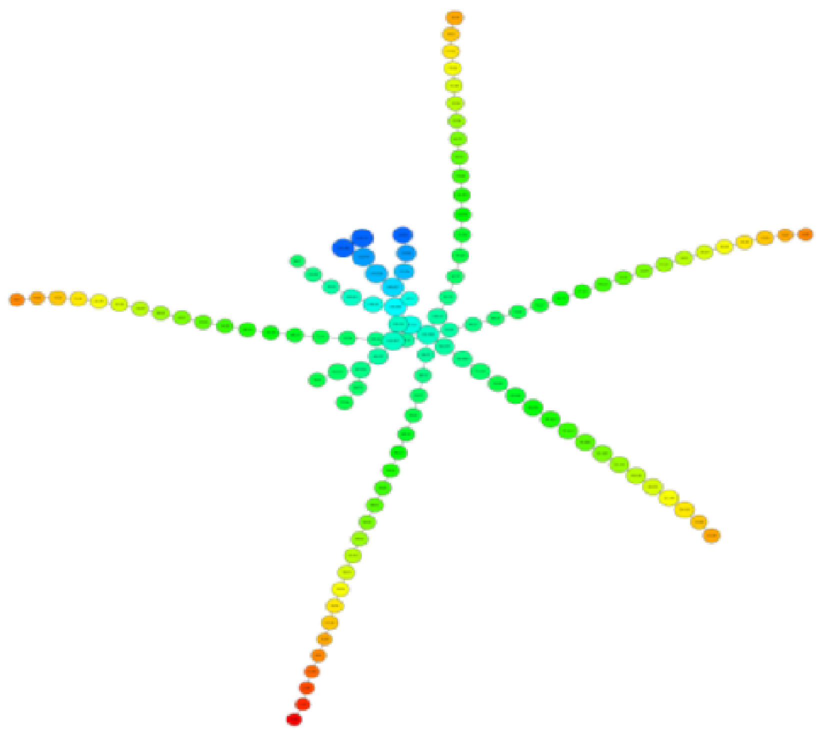
confidence level: 85%

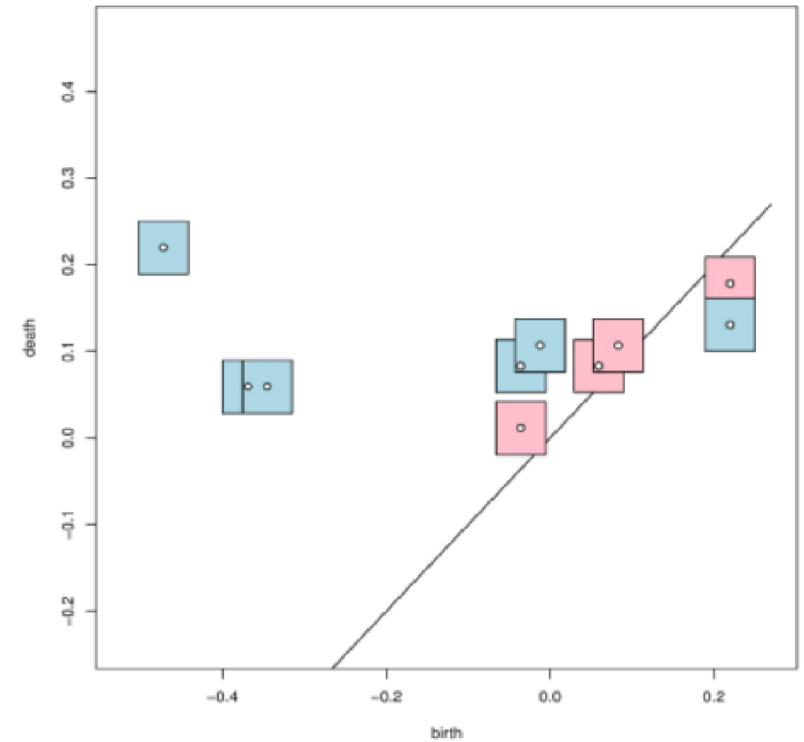






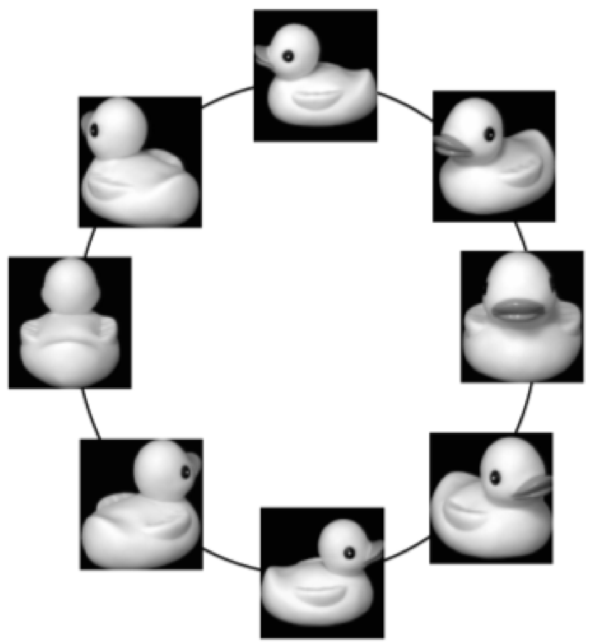


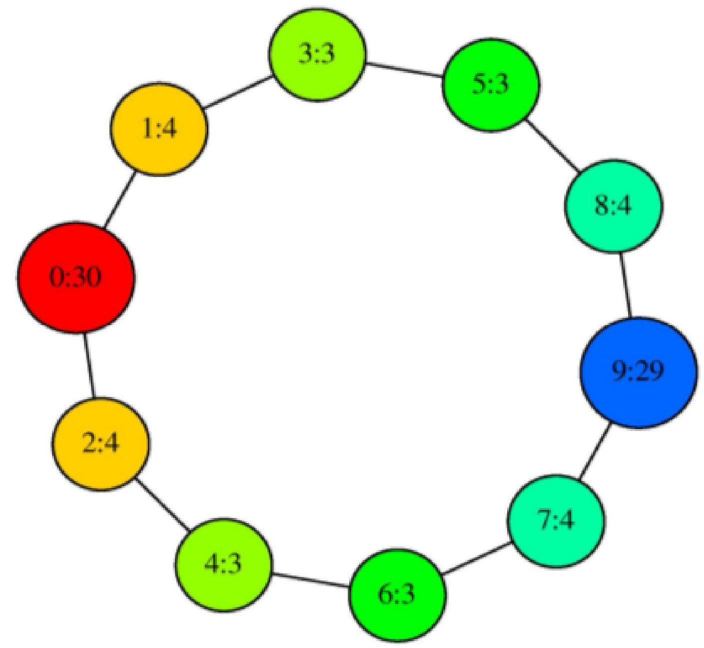


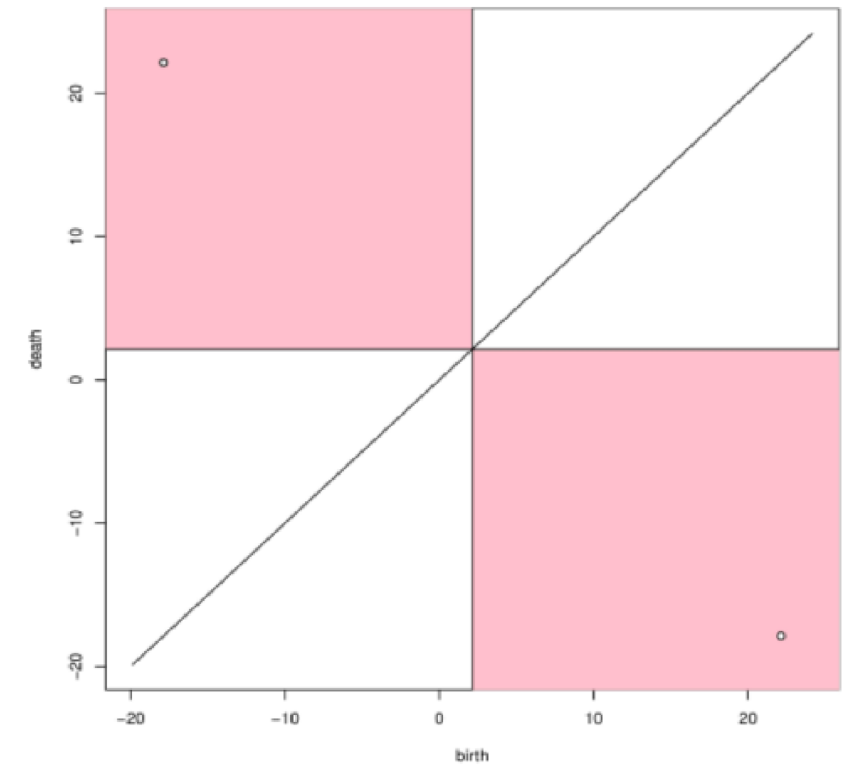


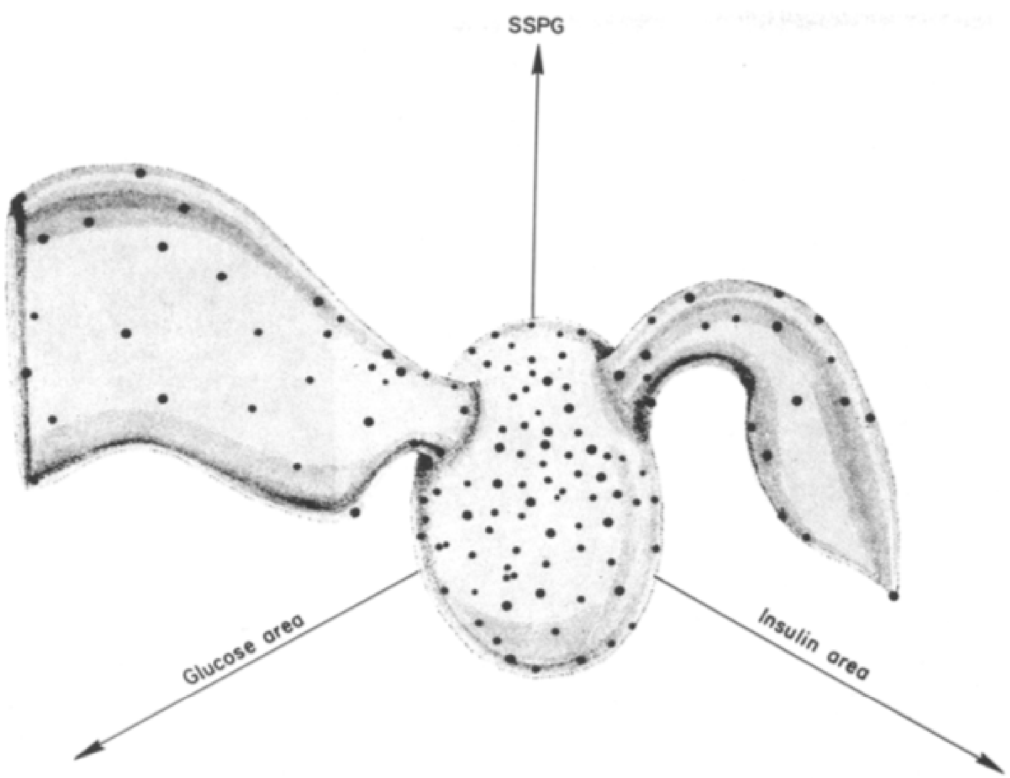
Experiments

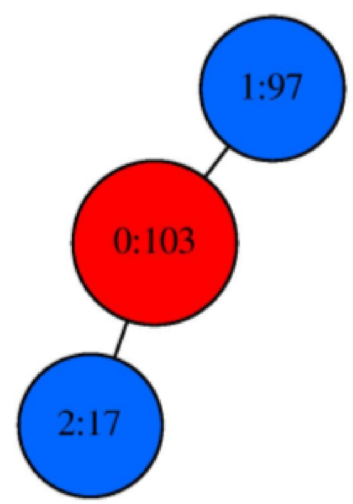
confidence level: 85%

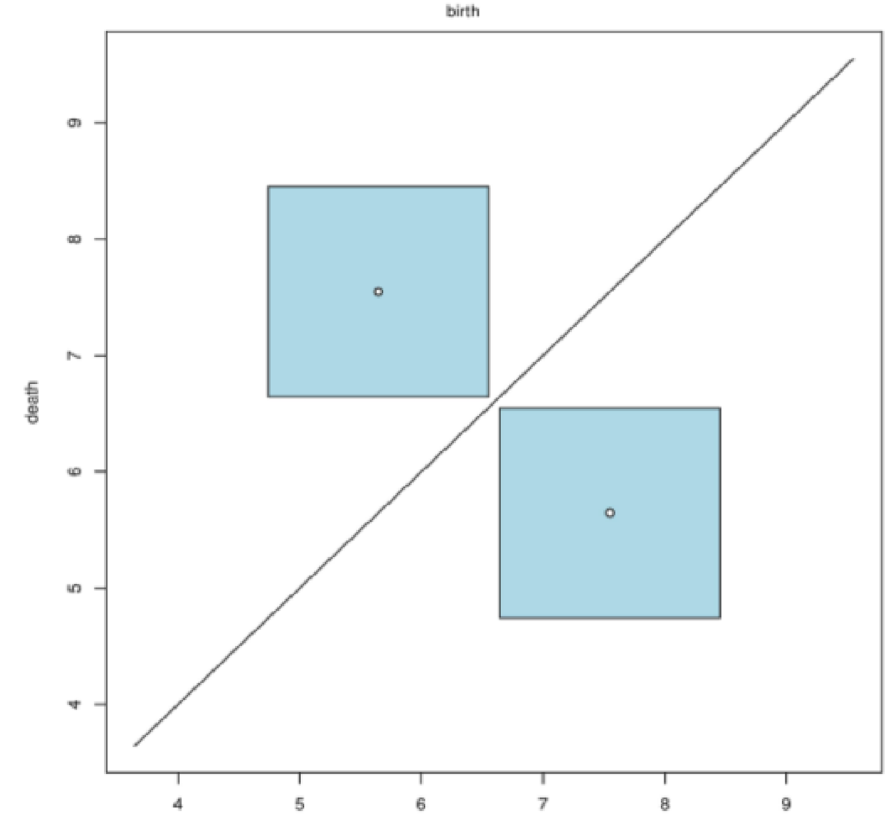












References

Metrics on Reeb graphs:

- M. Carrière, S. O. (2017): Proc. Sympos. Comput. Geom.
- de Silva, Munch, Patel (2016): Discrete Comput. Geometry, 55(4):854–906
- di Fabio, Landi (2016): Discrete Comput. Geometry, 55(2):423–461
- F. Chazal, R. Huang, J. Sun (2015): Discrete Comput. Geometry, 53(3):621–649
- U. Bauer, E. Munch, Y. Wang (2015): Proc. Sympos. Comput. Geom.
- U. Bauer, X. Ge, Y. Wang (2014): Proc. Sympos. Comput. Geom.
- Morozov, Weber (2014): Proc. TopoinVis

Mapper: structure, stability, statistics:

- M. Carrière, B. Michel, S. O. (2017): arXiv 1511.05823 [math.AT]
- M. Carrière, S. O. (2016): Proc. Sympos. Comput. Geom.
- E. Munch, B. Wang (2016): Proc. Sympos. Comput. Geom.
- G.Singh, F. Mémoli, G. Carlsson (2007): Proc. Sympos. Point-Based Graphics

Miscellaneous:

- Agarwal, Fox, Nath, Sidiropoulos, Wang (2015): Proc. Sympos. Algorithms Comput.
- Carr, Duke (2014): IEEE Trans Vis Comput Graph.
- Cohen-Steiner, Edelsbrunner, Harer (2009): J. Found. Comput. Mathematics
- M. Gromov (1981):*Structures métriques pour les variétés riemanniennes*. CEDIC