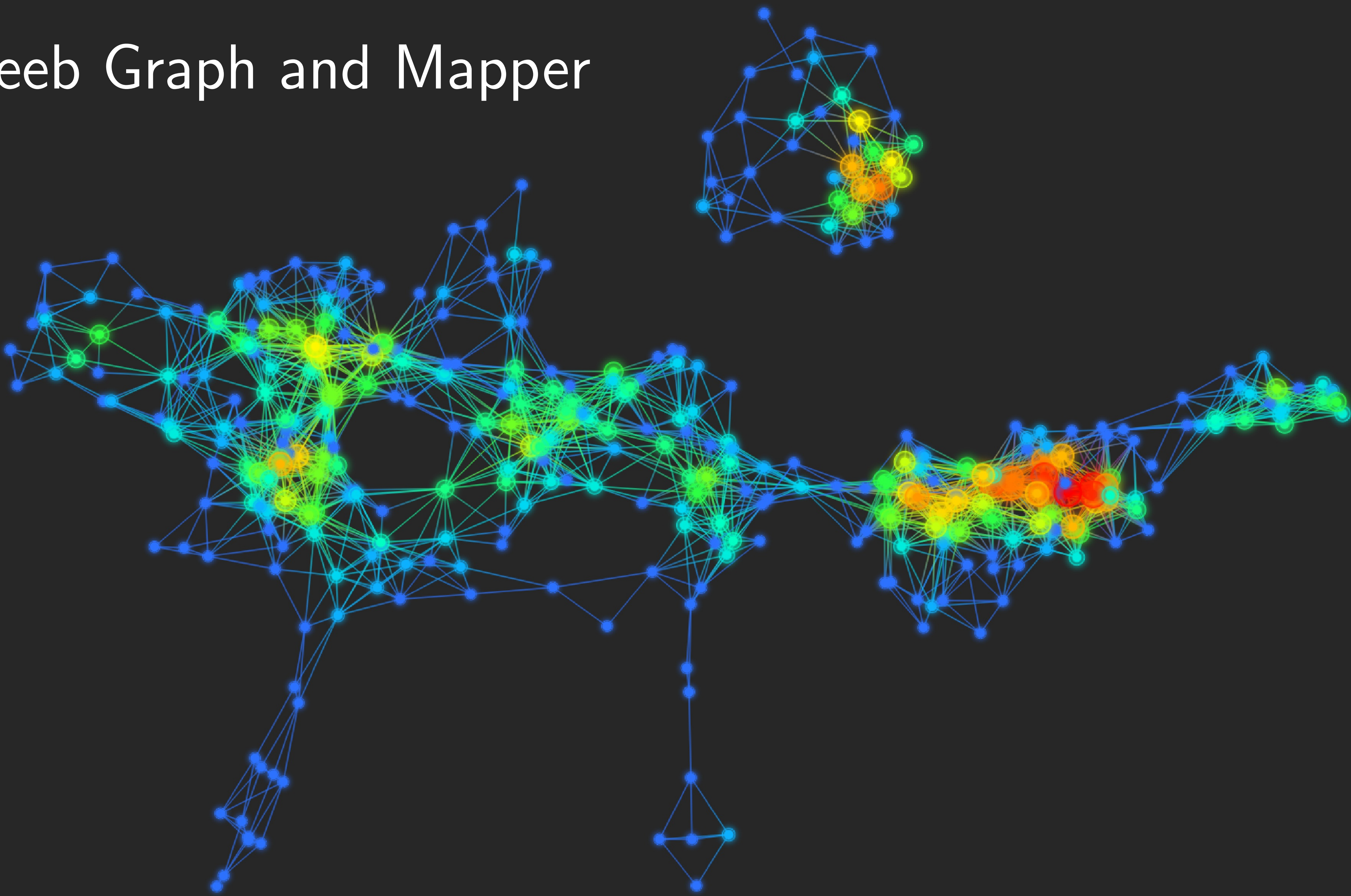
Reeb Graph and Mapper



Motivations

get a higher-level understanding of the structure of data

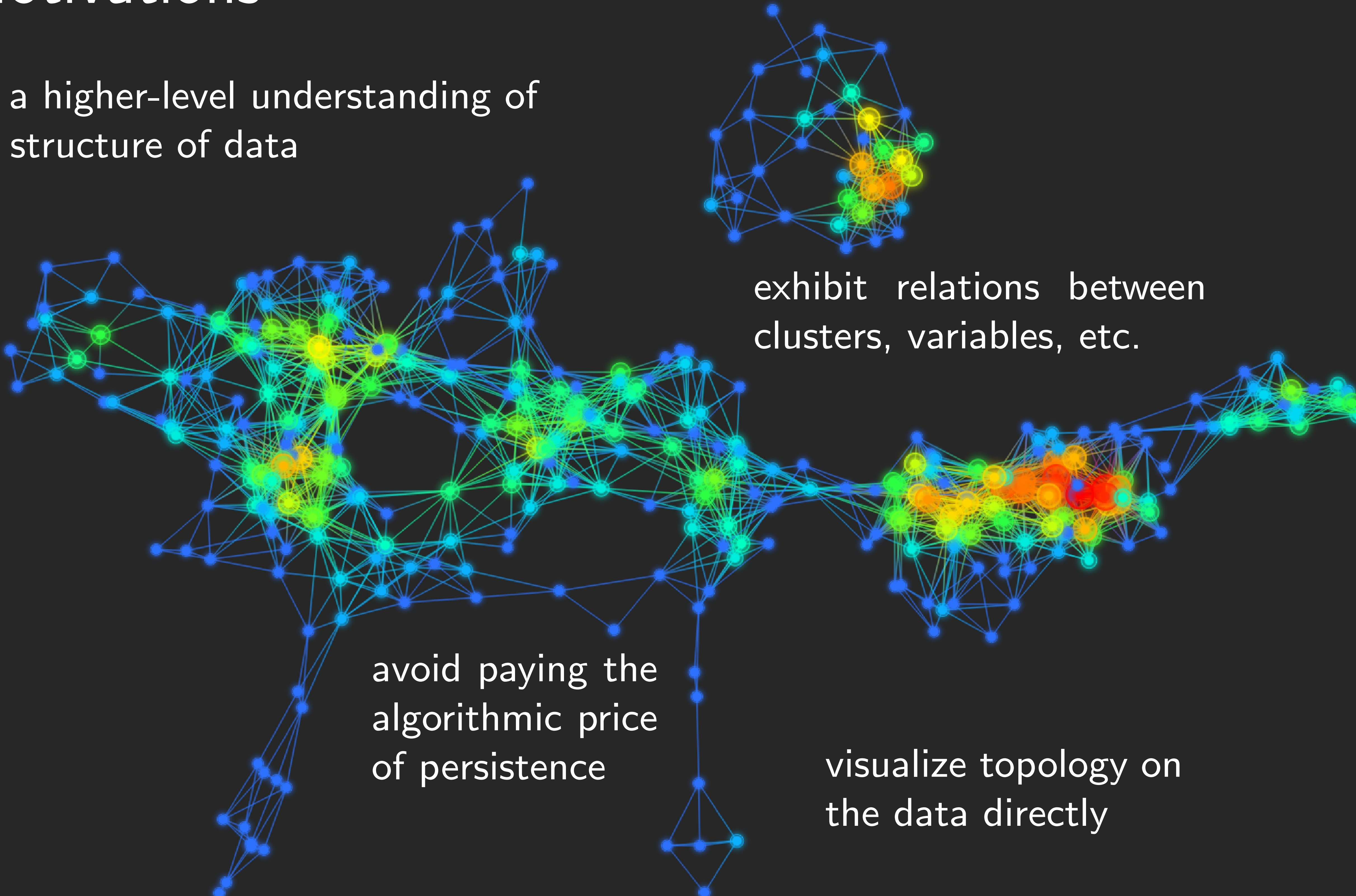


exhibit relations between clusters, variables, etc.

avoid paying the algorithmic price of persistence

visualize topology on the data directly

Motivations

get a higher-level understanding of
the structure of data

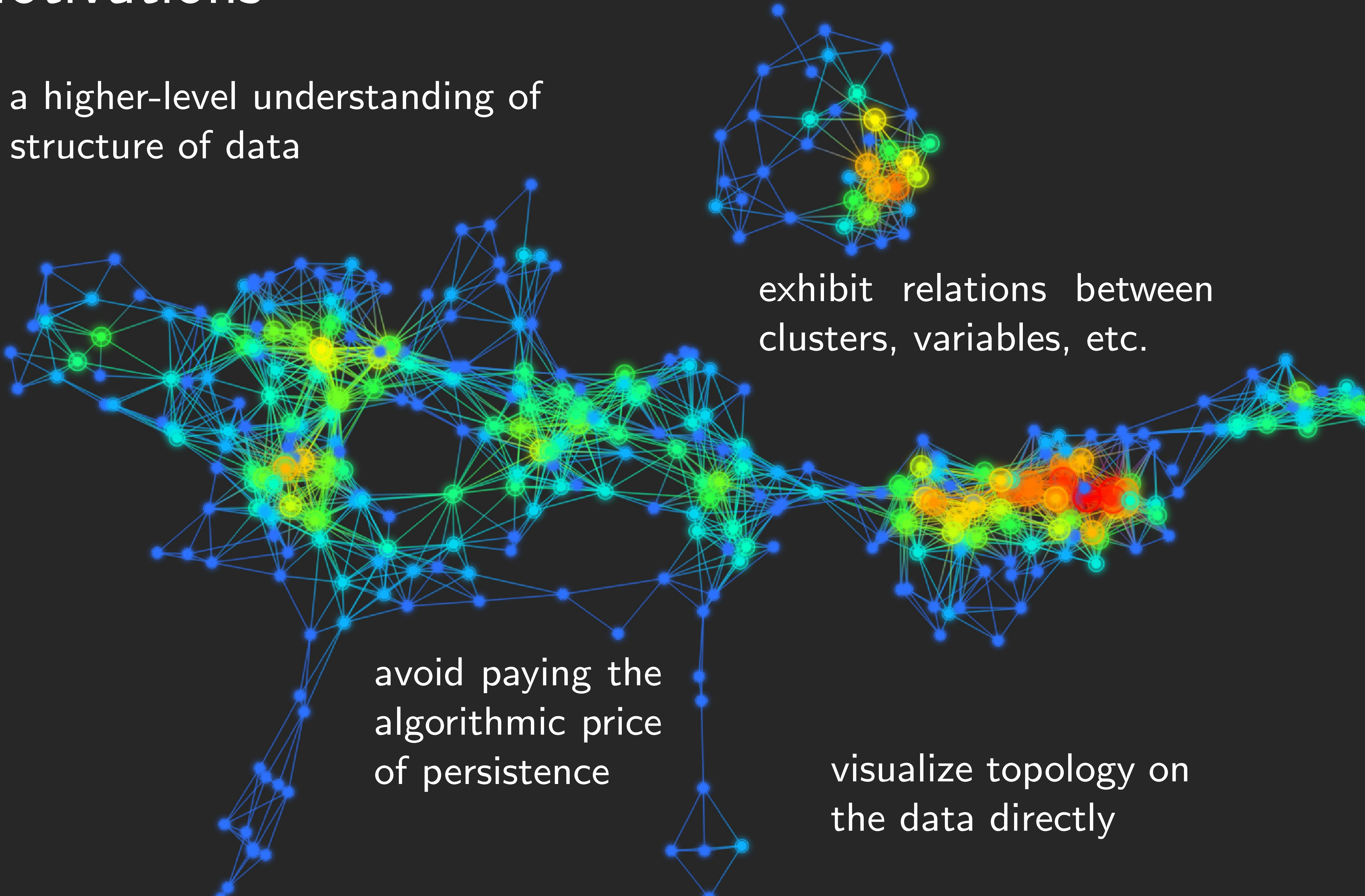


exhibit relations between
clusters, variables, etc.

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

Input:

- continuous function ('filter', 'lens'...) $f : X \to \mathbb{R}$

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

Method:

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

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

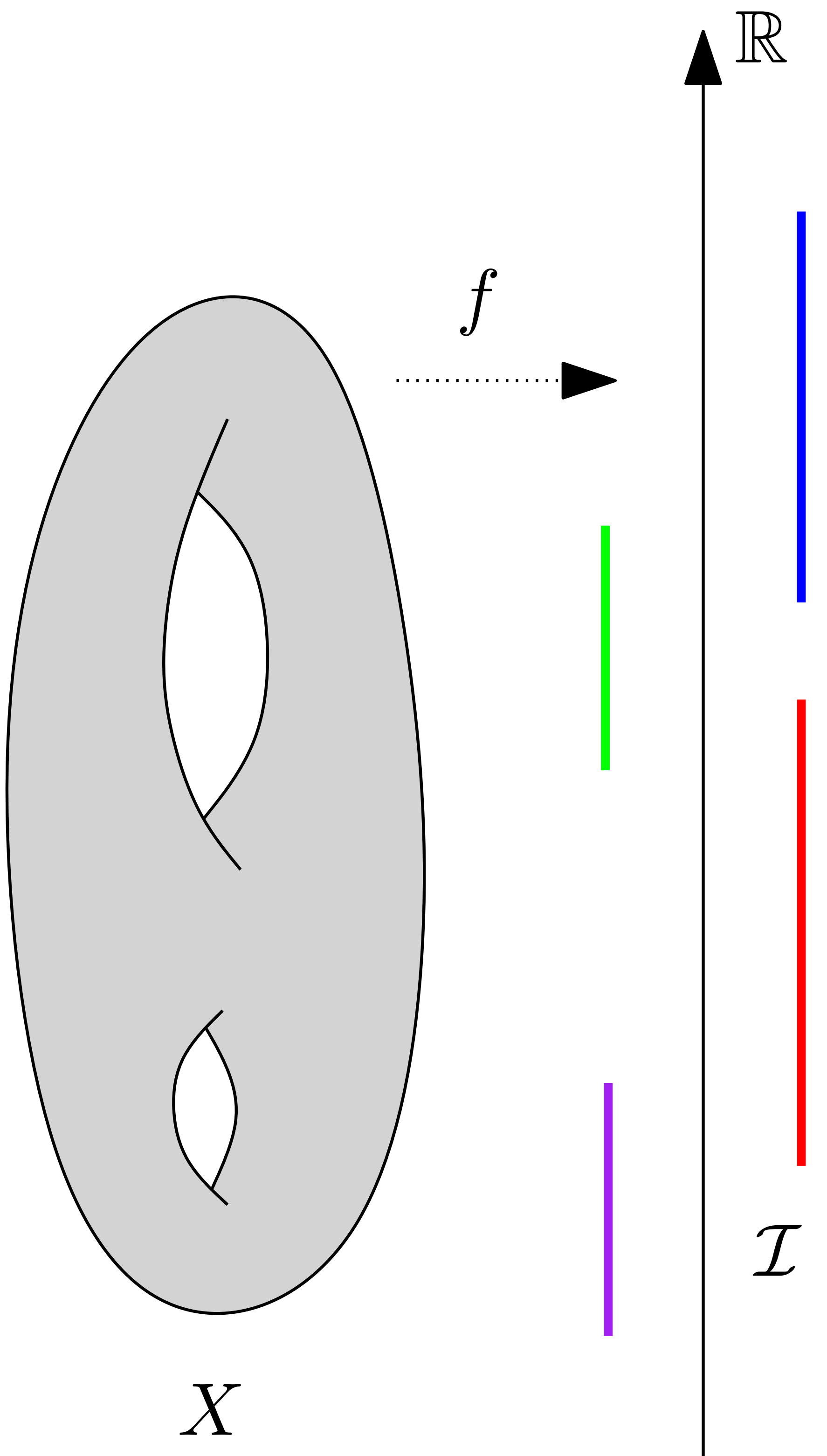
• The Mapper is the *nerve* of $\mathcal{V}$:

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

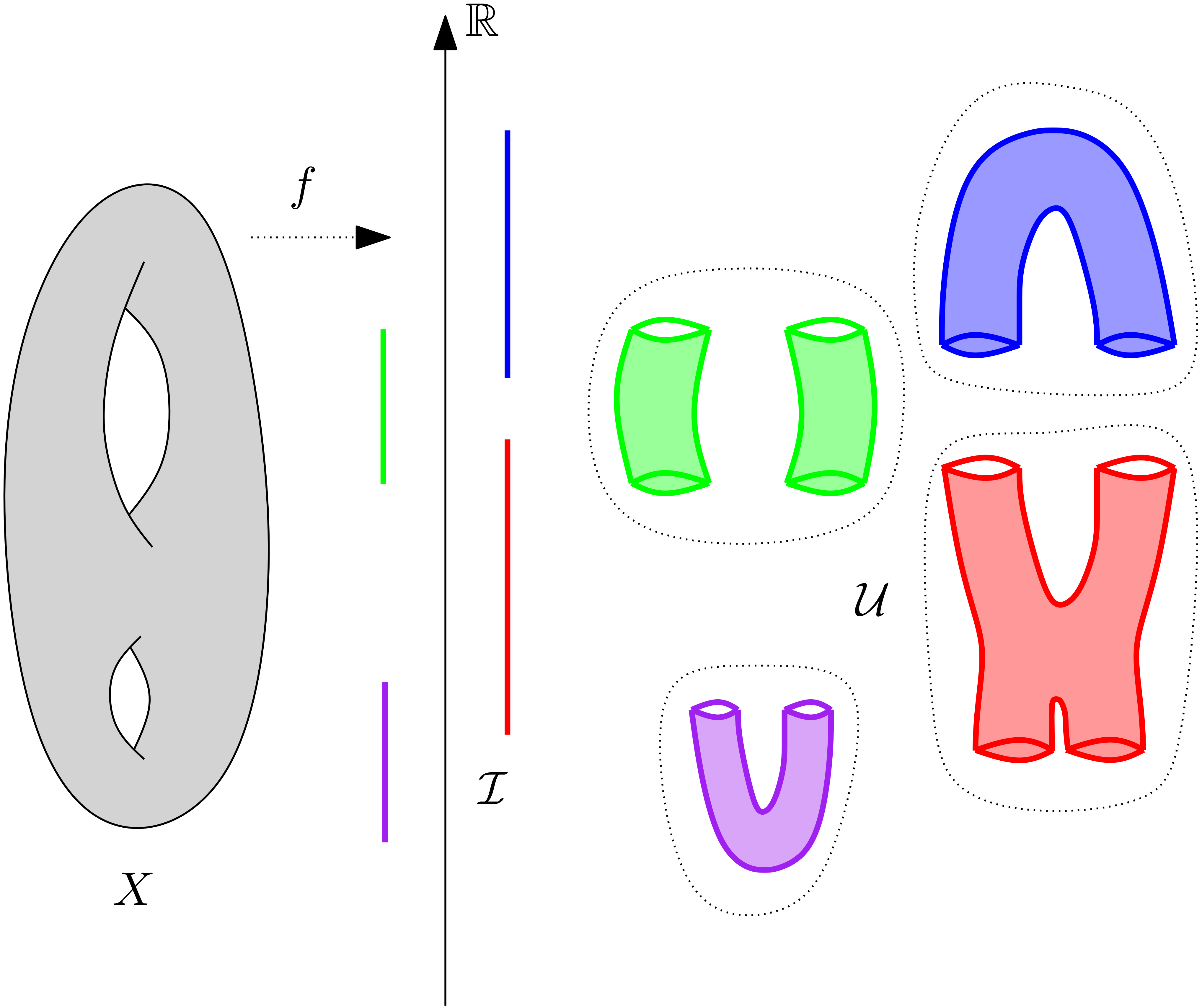
- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$

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

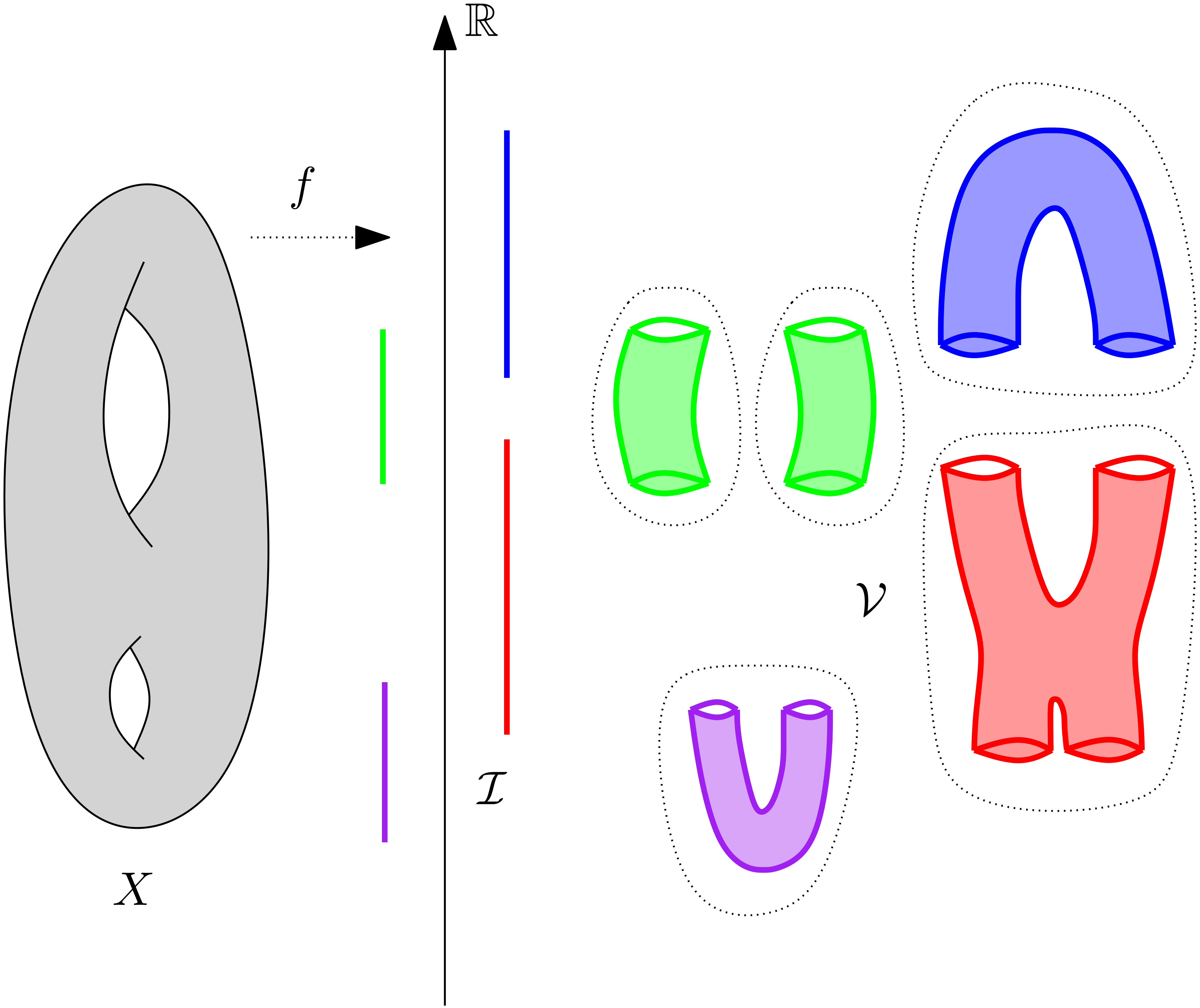
Mapper in the continuous setting



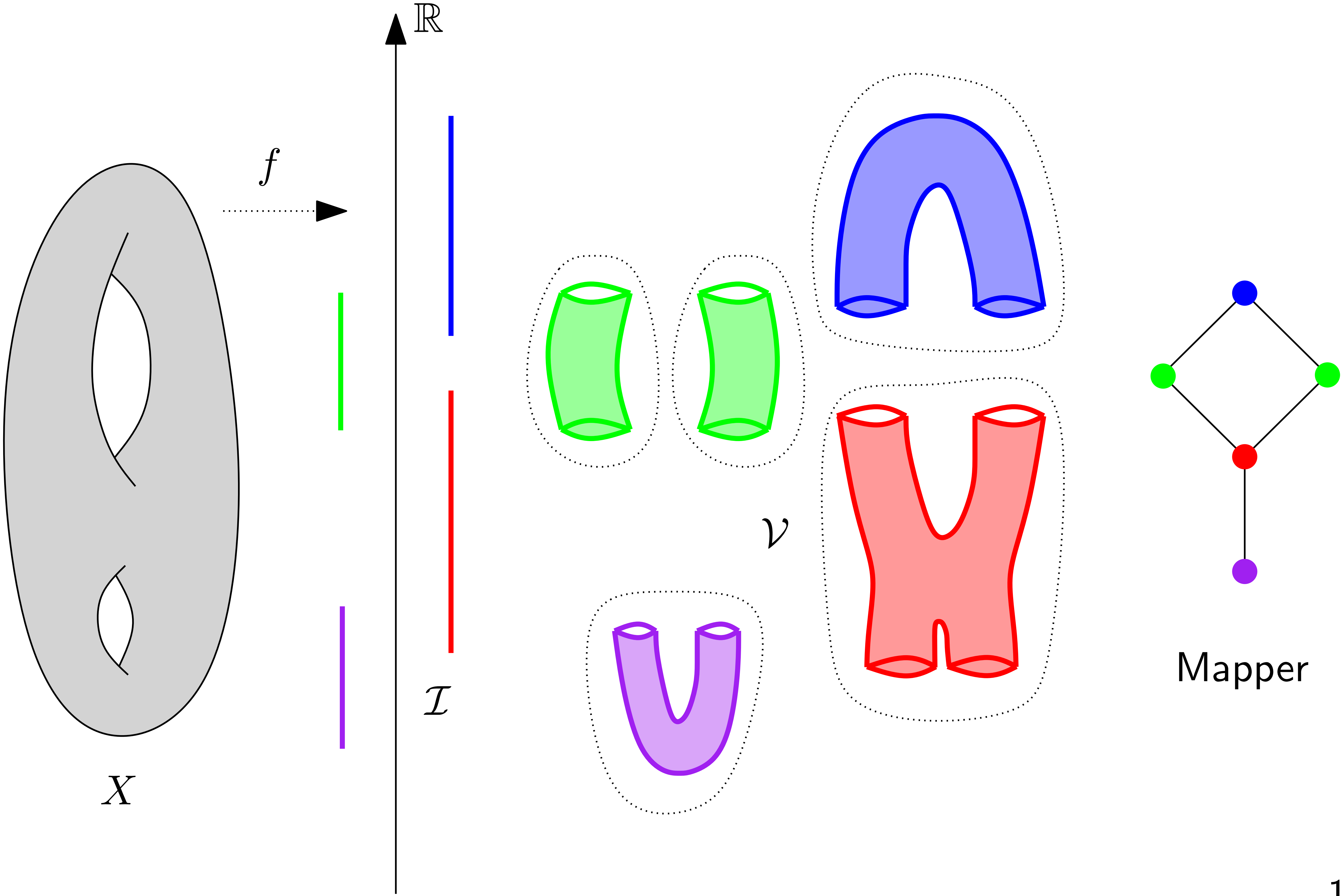
Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting



Mapper in practice

Input: - point cloud P with metric d_P

- continuous function ('filter', 'lens'...) $f : P \to \mathbb{R}$

- 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 $\mathcal{V}$:

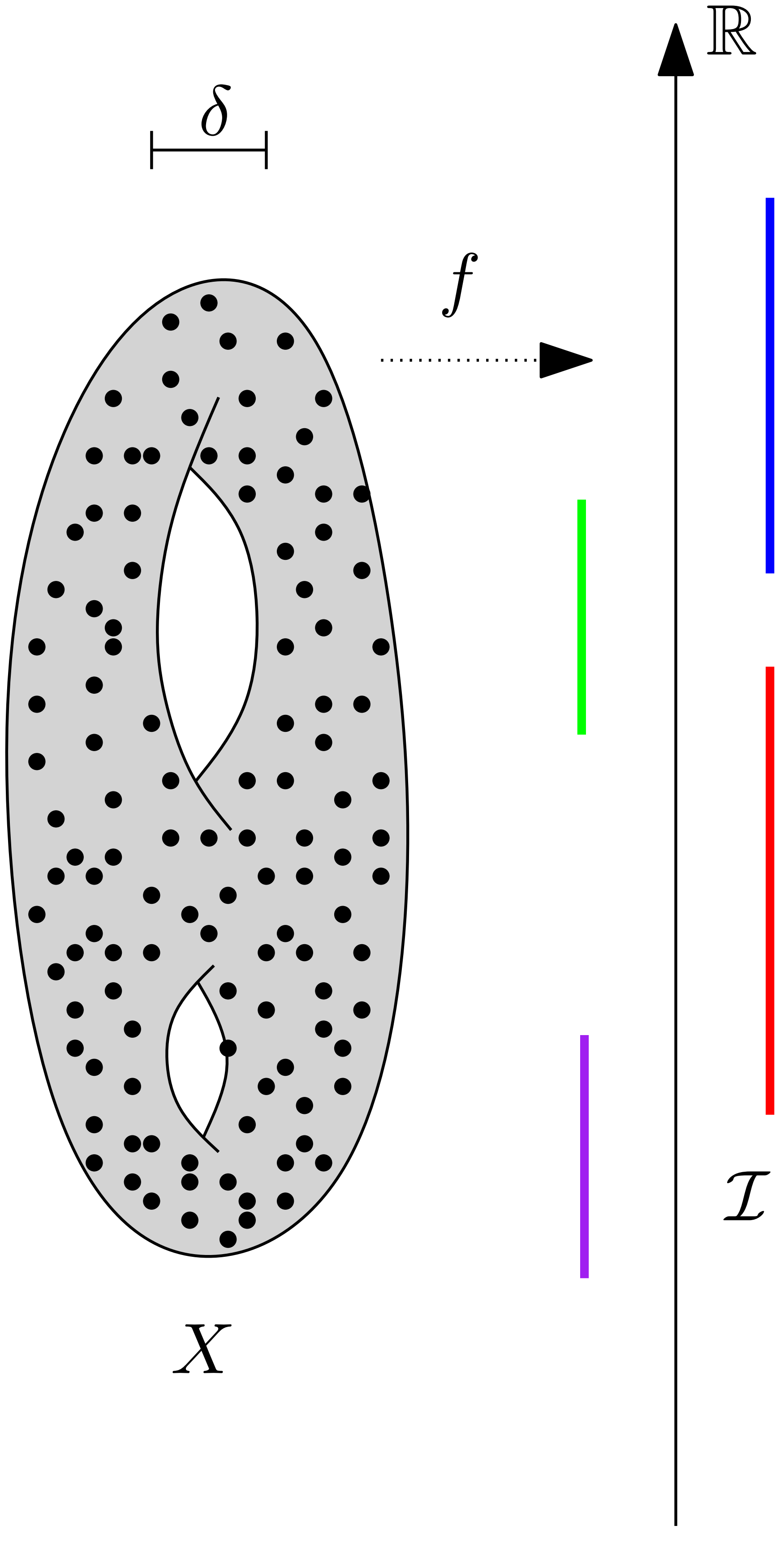
- 1 vertex per element $V \in \mathcal{V}$

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

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

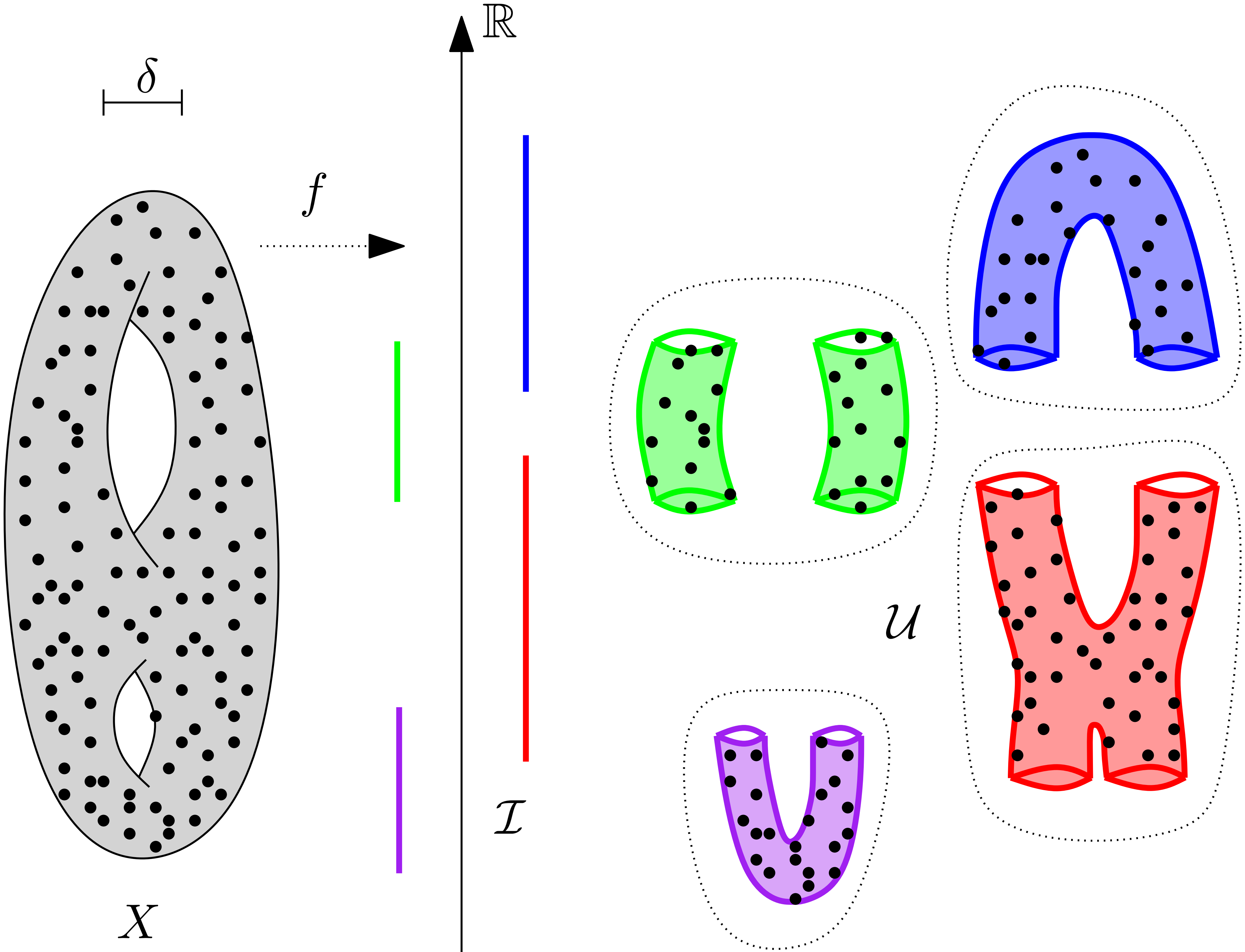
(intersections materialized by data points)

Mapper in practice



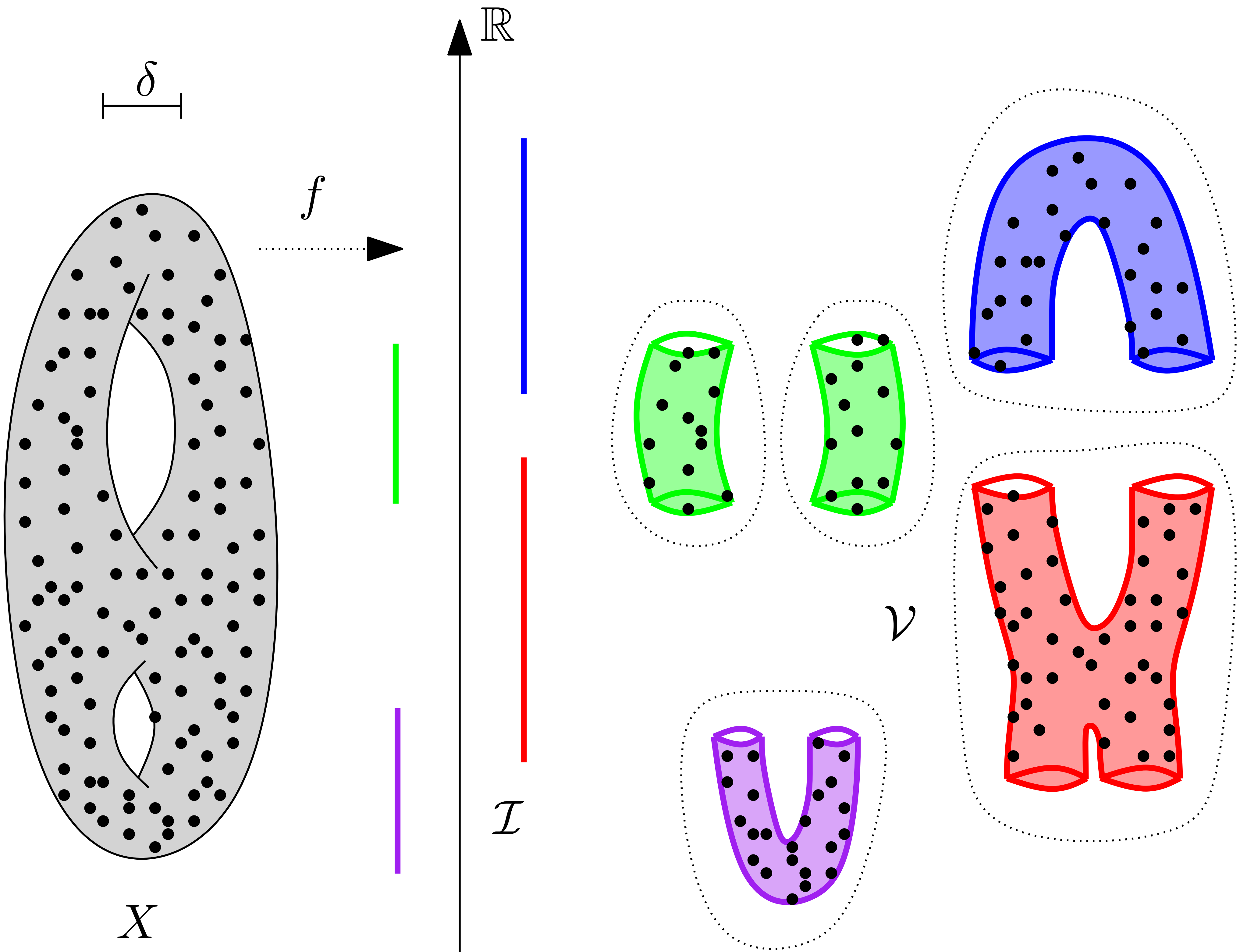
$G = \delta$-neighborhood graph

Mapper in practice



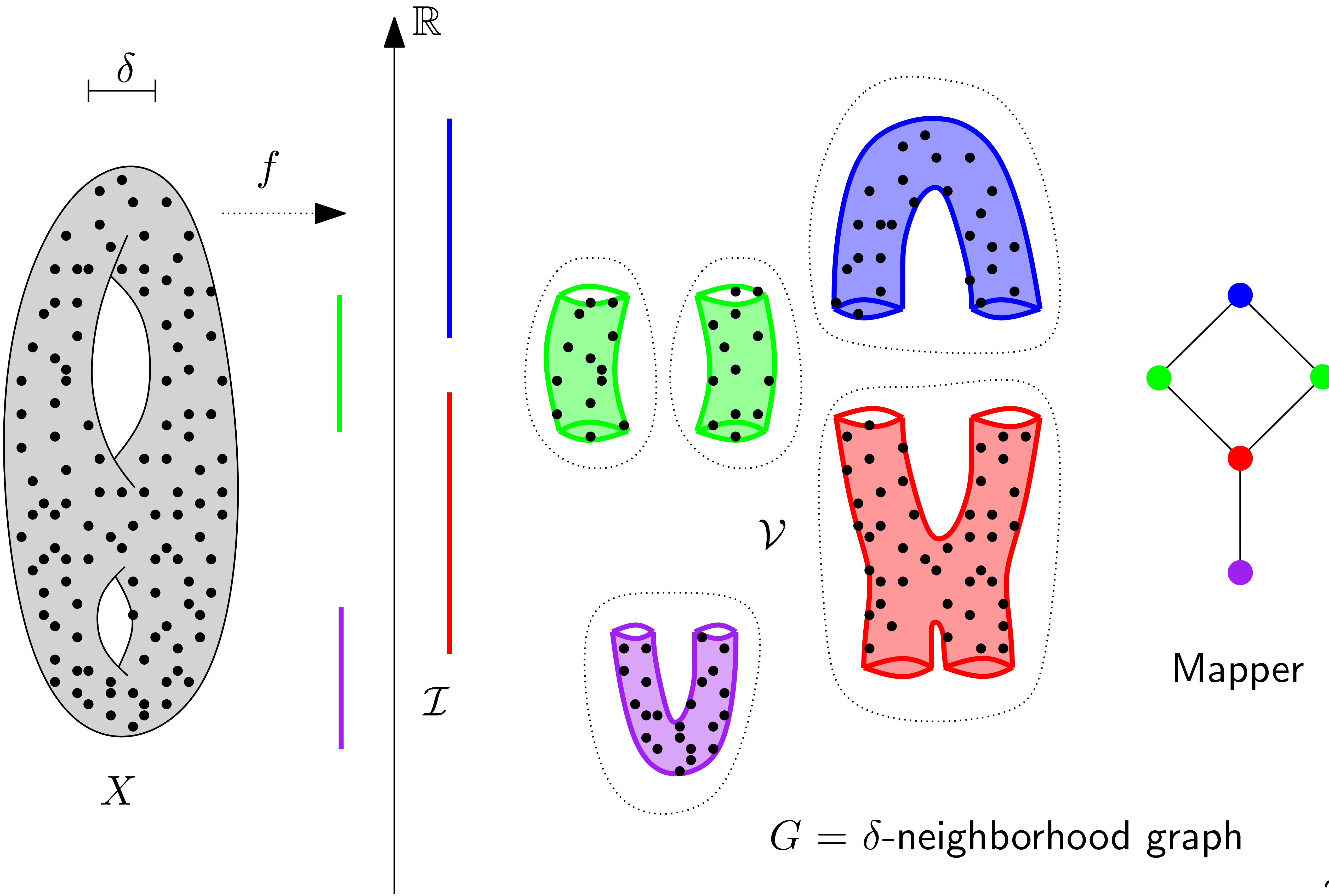
$G = \delta$-neighborhood graph

Mapper in practice



$G = \delta$-neighborhood graph

Mapper in practice



Mapper in practice

Parameters:

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

Mapper in practice

Parameters:

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

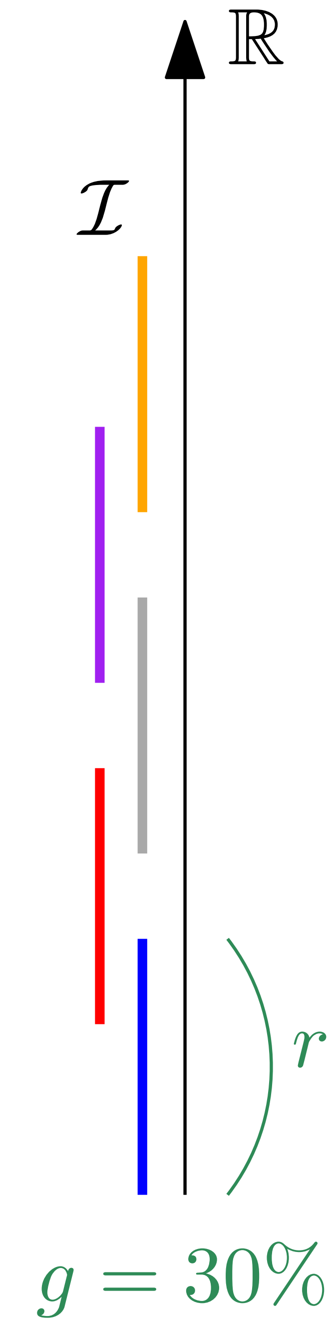
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals

- neighborhood size δ

range scale

geometric scale

Mapper in practice



Parameters:

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

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

- neighborhood size δ

range scale

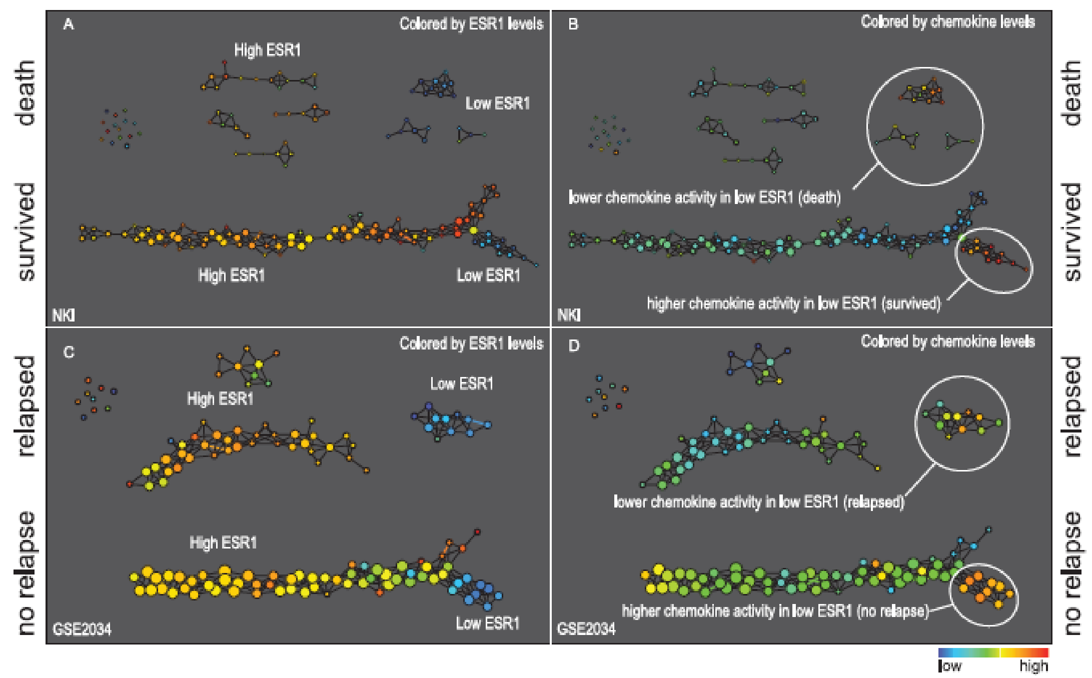
geometric scale

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

- resolution / granularity: r (diameter of intervals)

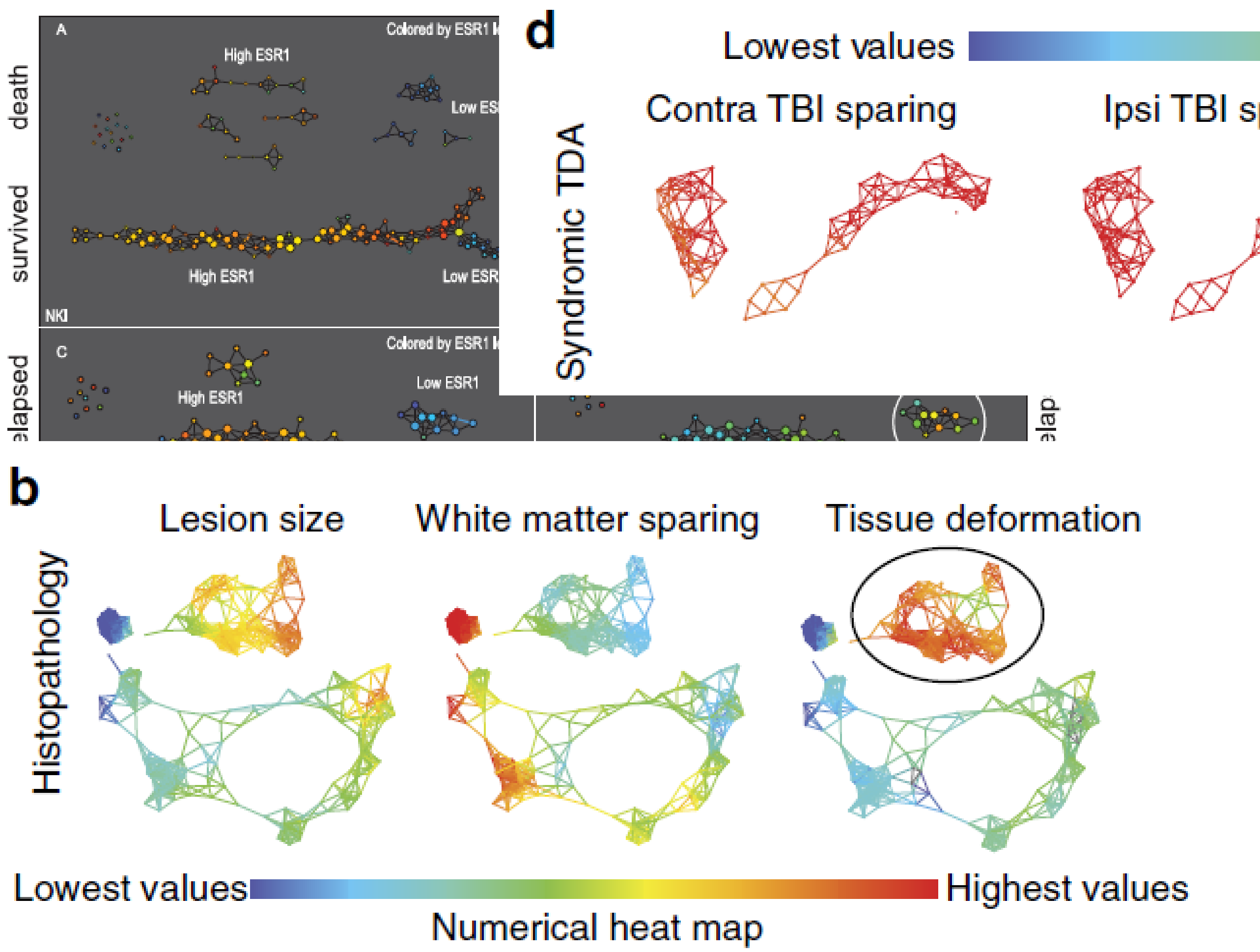
- gain: g (percentage of overlap)

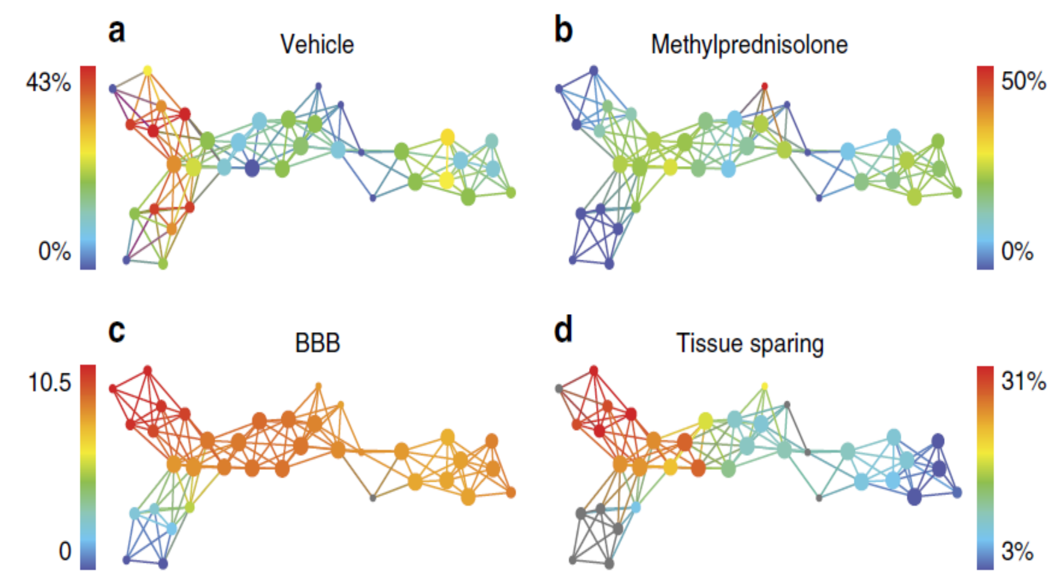
Mapper in applications



breast cancer subtype

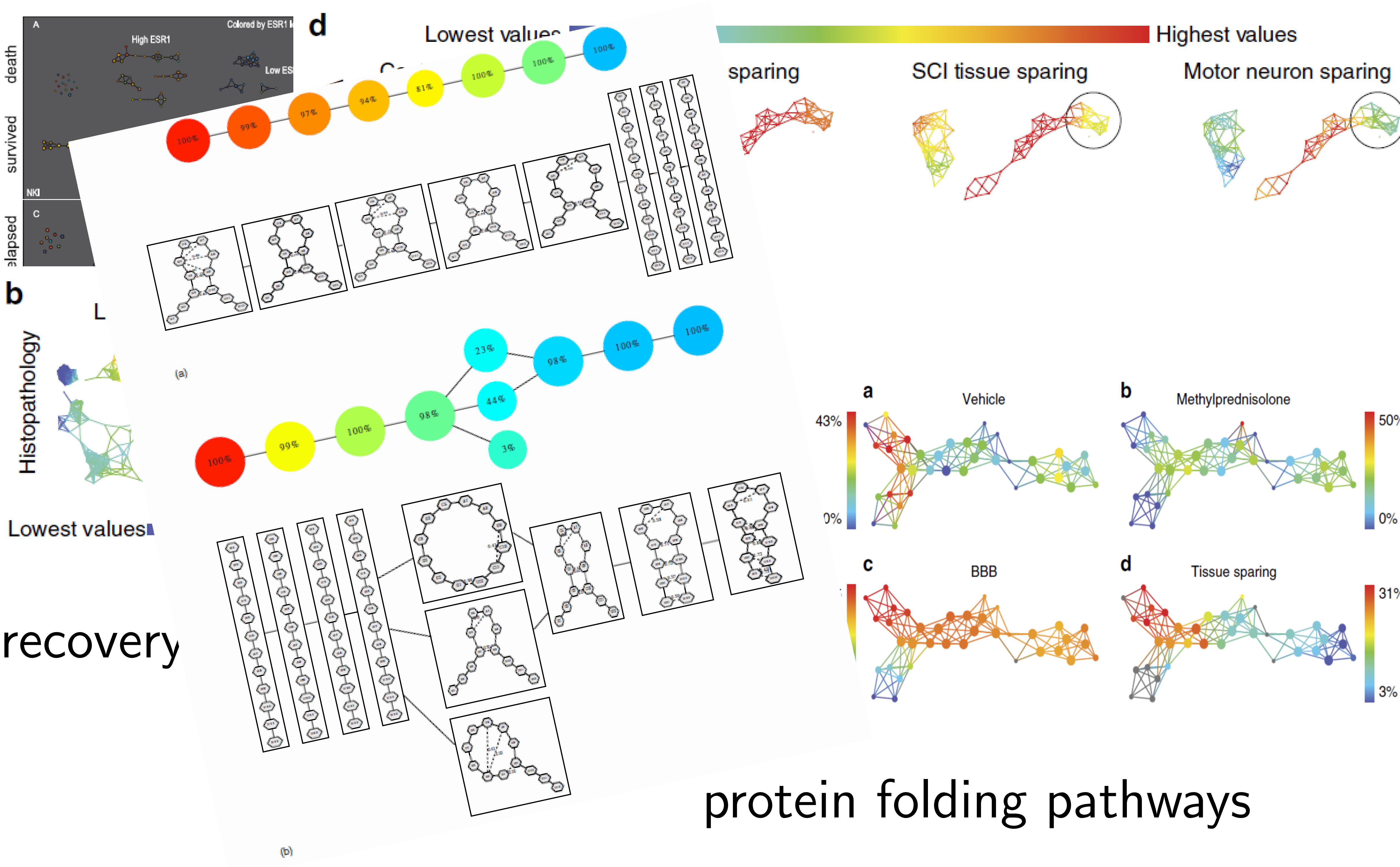
Mapper in applications





recovery from spinal cord injuries

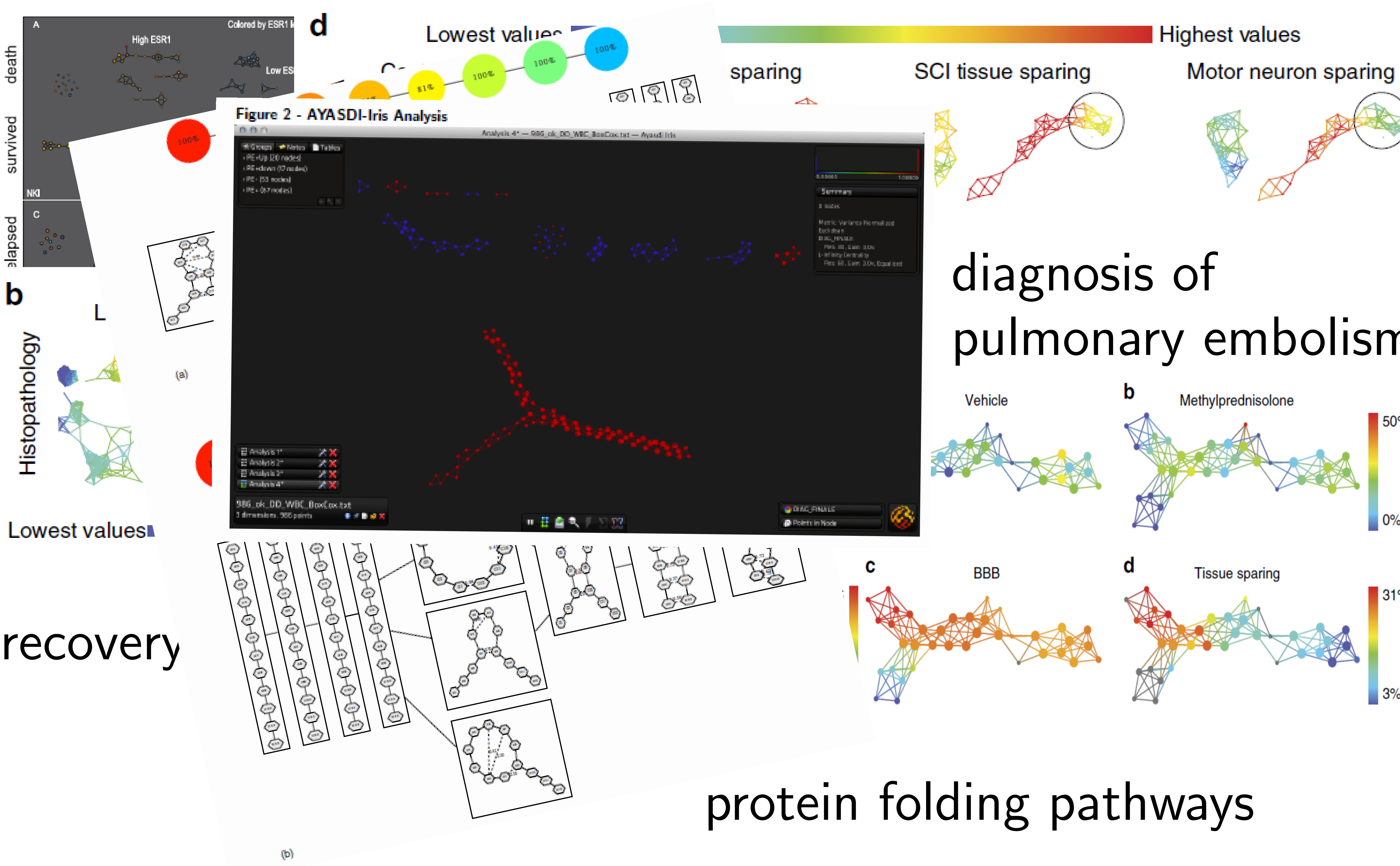
Mapper in applications



recovery

protein folding pathways

Mapper in applications



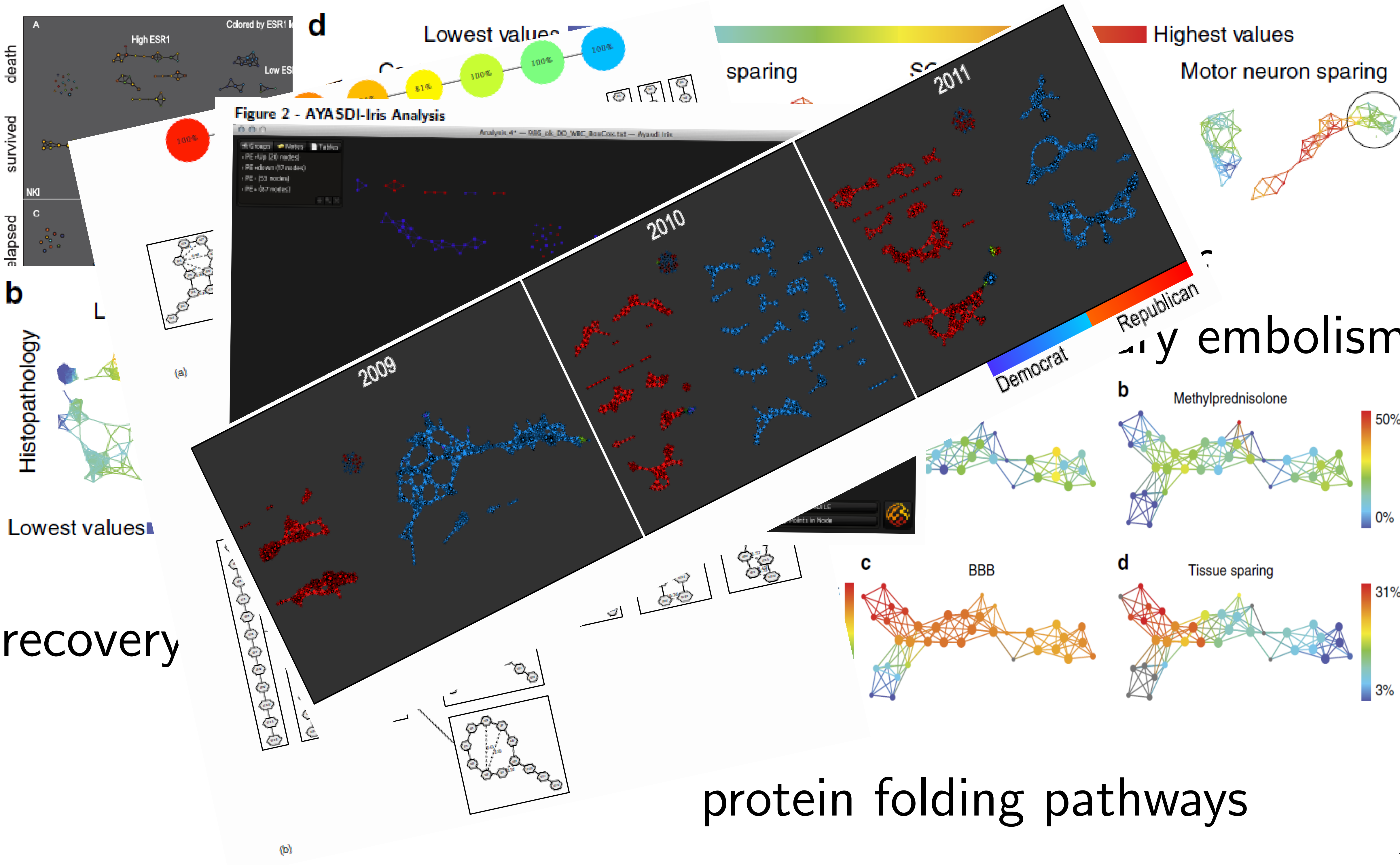
diagnosis of pulmonary embolism

recovery

protein folding pathways

Mapper in applications

implicit networks in the US house of representatives



...ry embolism

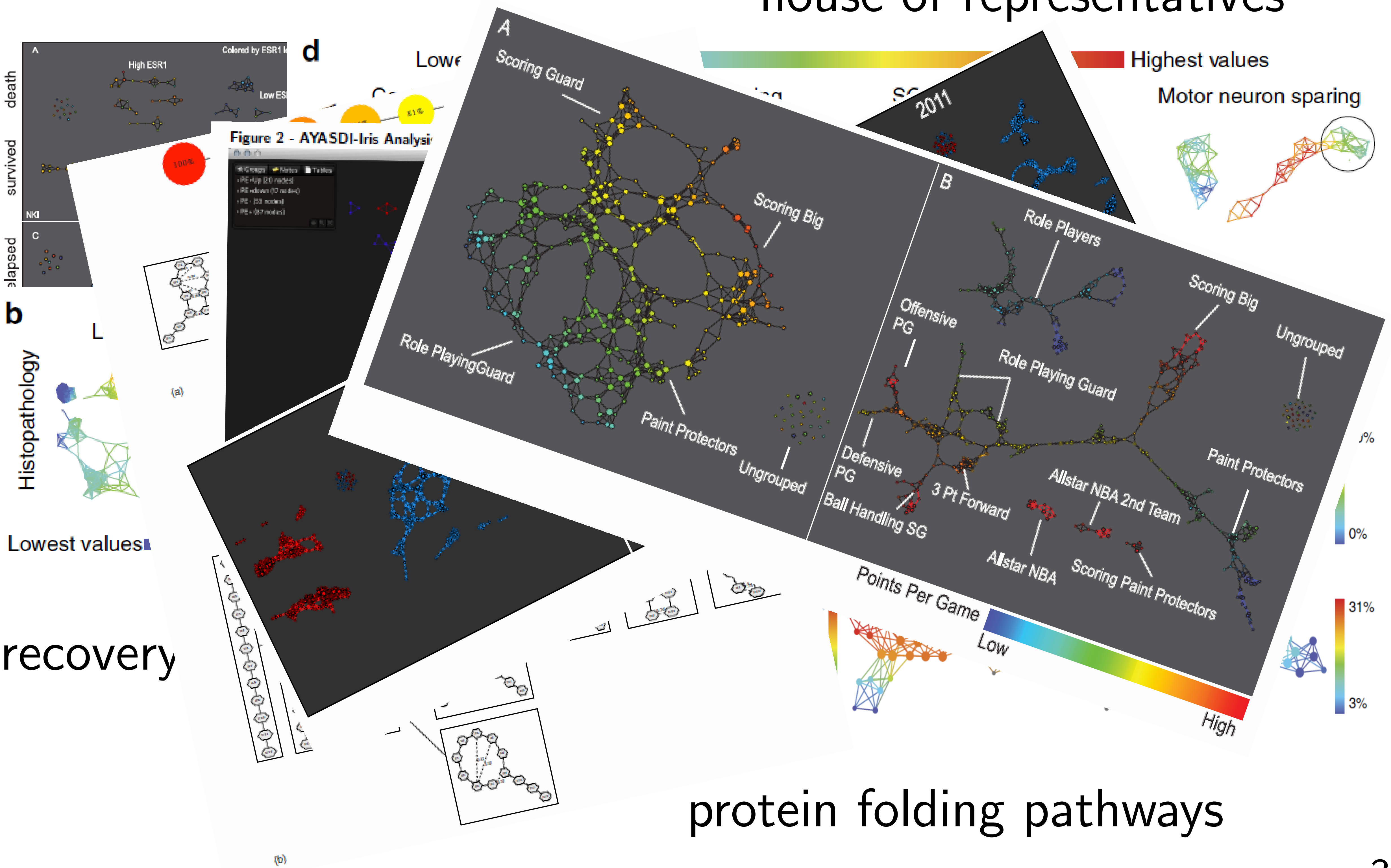
recovery

protein folding pathways

Mapper in applications

classification of NBA players

implicit networks in the US house of representatives



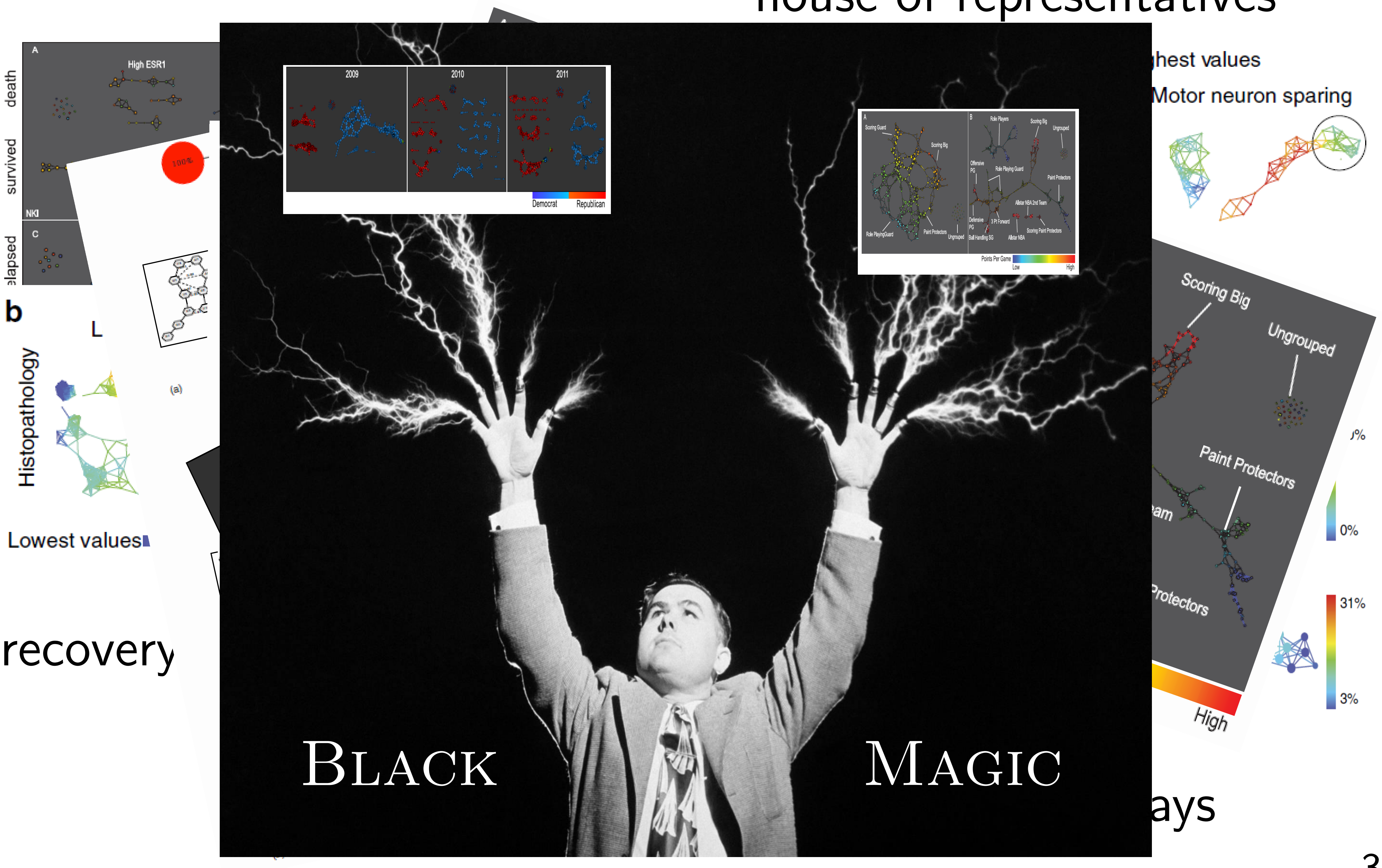
recovery

protein folding pathways

Mapper in applications

classification of NBA players

implicit networks in the US house of representatives



recovery

ays

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

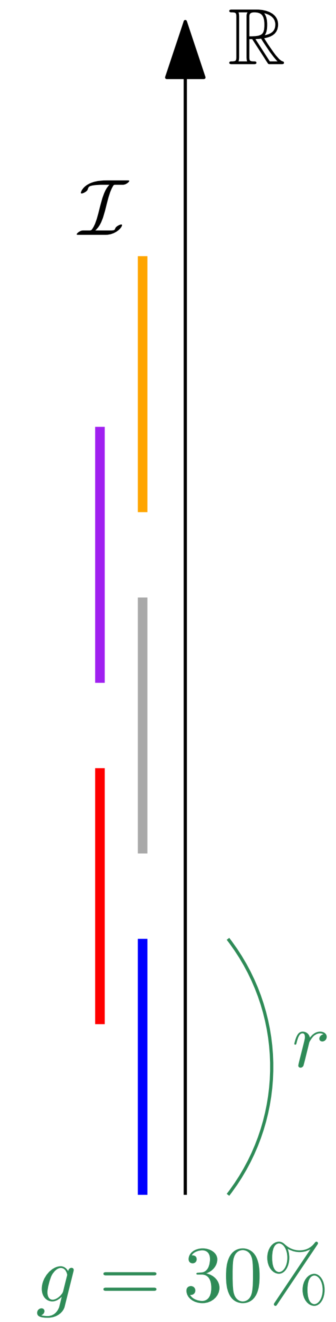
Using Topological Data Analysis for Diagnosis Pulmonary Embolism, Rucco et al., arXiv preprint, 2014

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

Choice of parameters



Parameters:

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

range scale

geometric scale

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

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

Choice of parameters

How to choose r, g and δ?

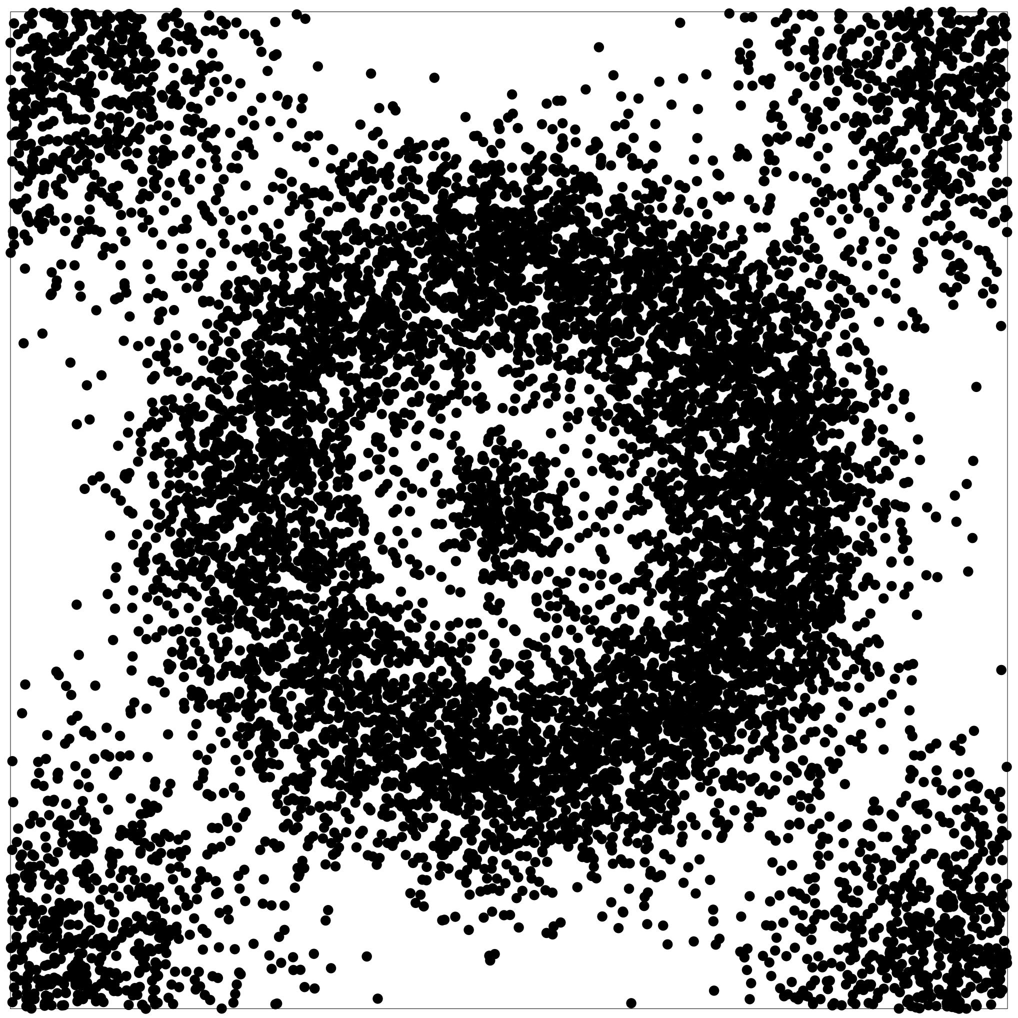
$\rightarrow$ in practice: trial-and-error (or some vague procedure)

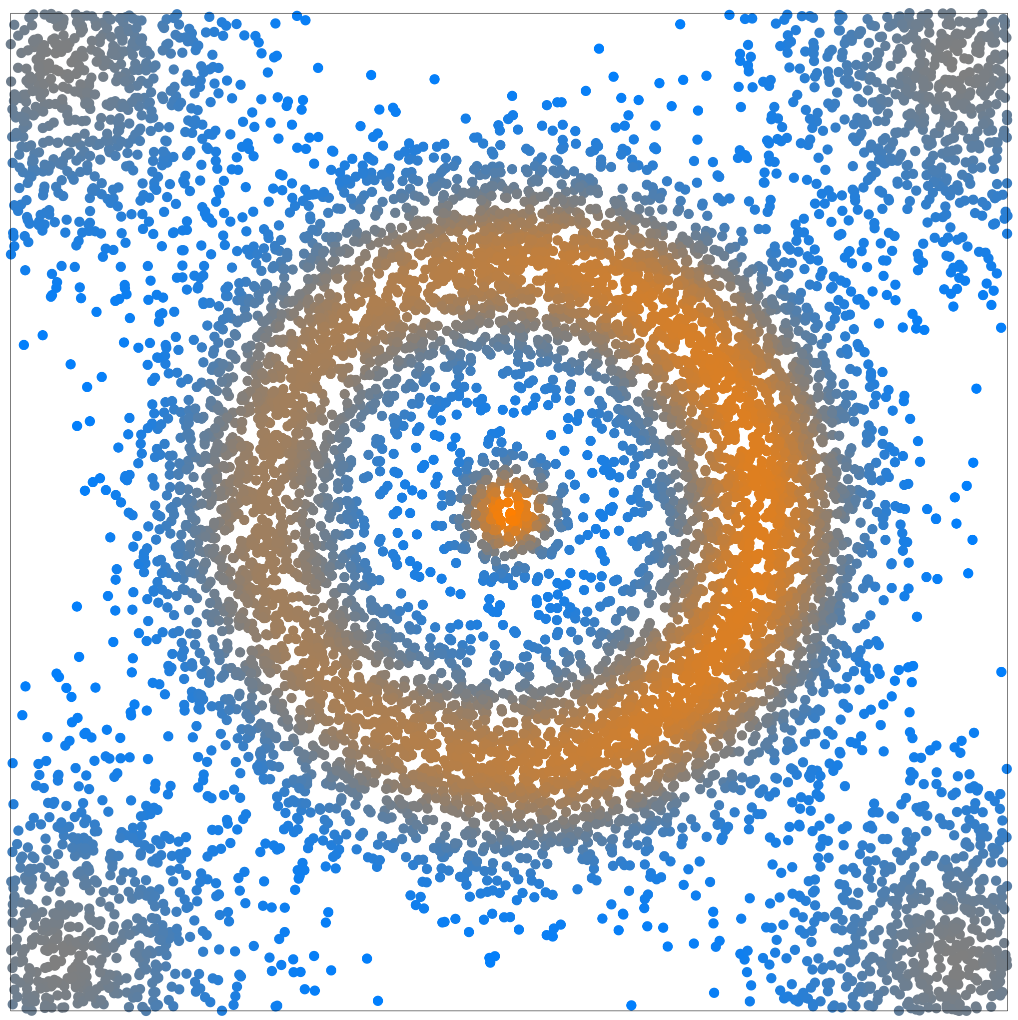
> high-dimensional data sets[40,48]. This is performed automatically within the software, by deploying an ensemble machine learning algorithm that iterates through overlapping subject bins of different sizes that resample the metric space (with replacement), thereby using a combination of the metric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

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





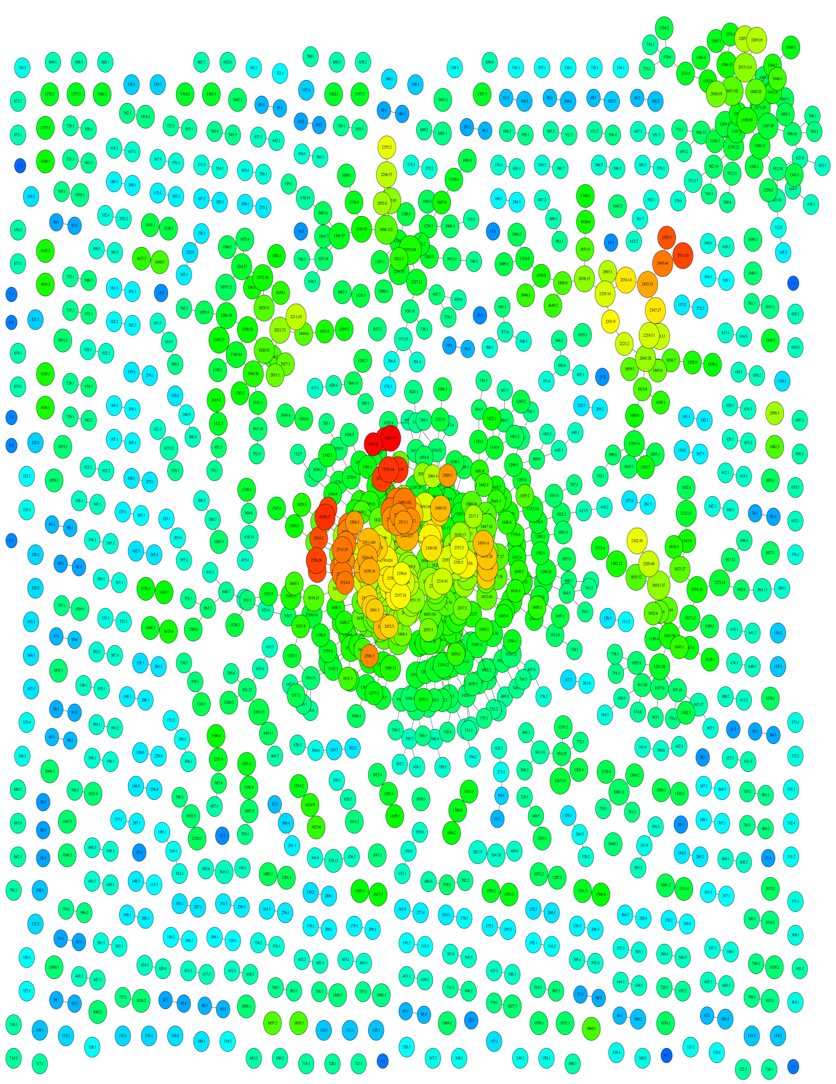
Choice of parameters

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

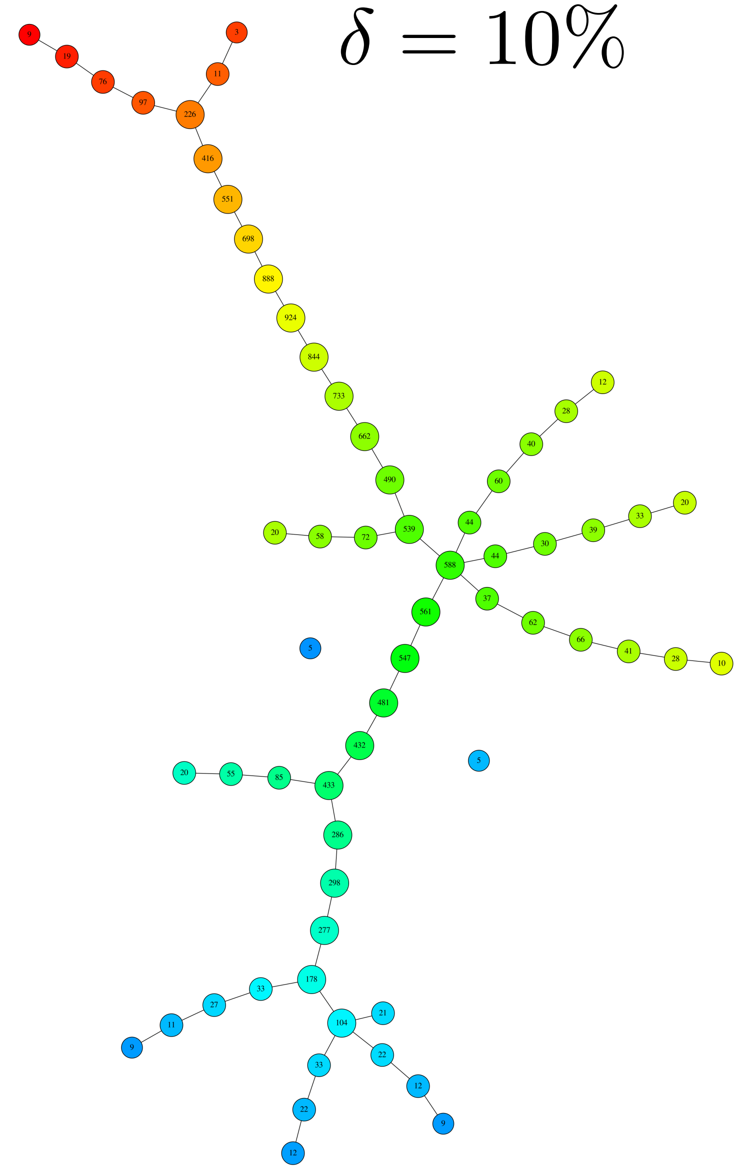
$f =$ density estimator, $r = 1/30, g = 20\%$

$\delta =$ percentage of the diameter of X

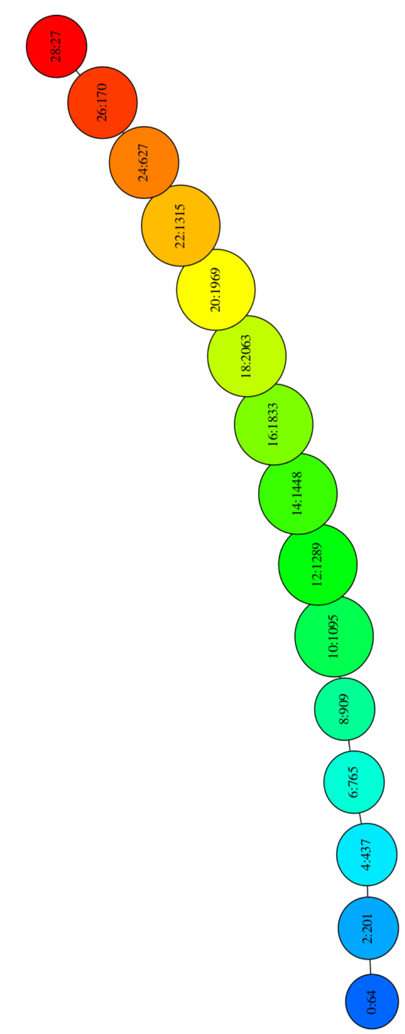
$\delta = 1\%$



$\delta = 10\%$



$\delta = 30\%$



Choice of parameters

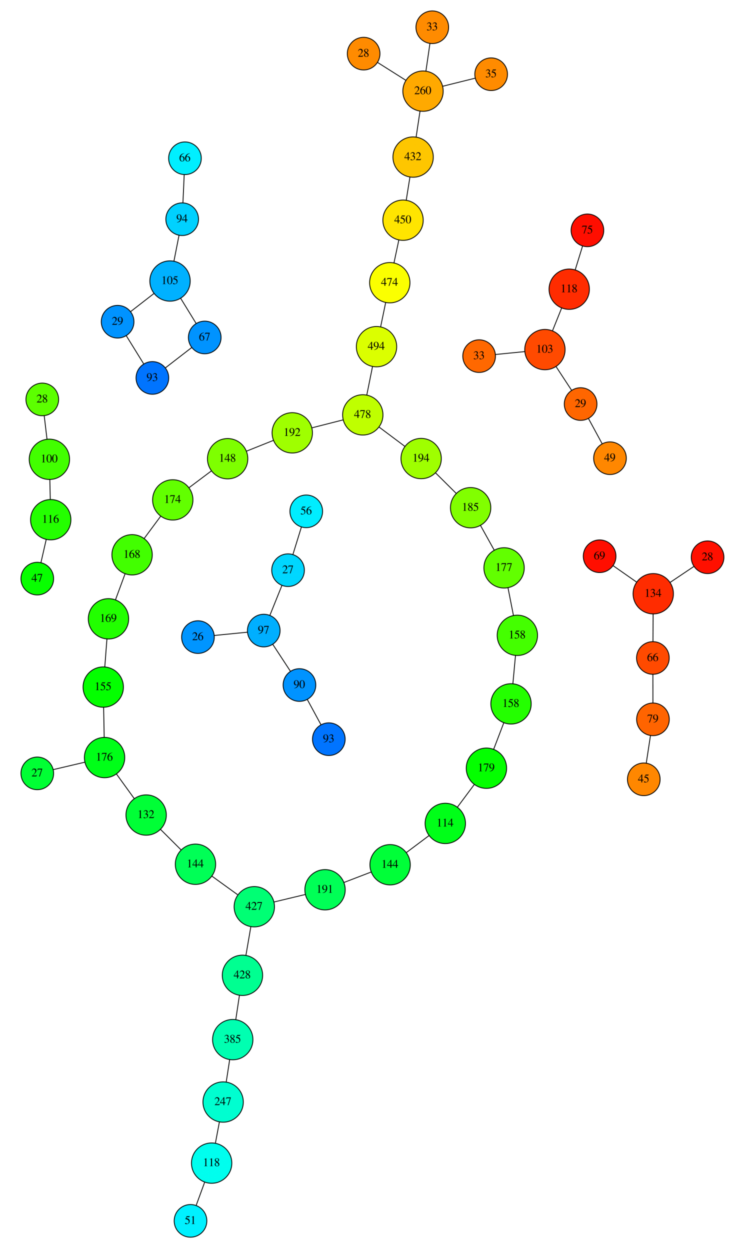
Illustration: $P \subset \mathbb{R}^2$ sampled from known probability distribution

$f =$ abscissa, $r = 1/30, g = 10\%$

$\delta =$ percentage of the diameter of X

$$\delta = 1\%$$

$+$ small cluster removal



Choice of parameters

Structure and Stability of the 1-Dimensional Mapper.
Carrière, O. 2016

$\rightarrow$ clarifies formally the roles of r and g in the continous setting

$\rightarrow$ gives sufficient conditions on δ to get approximation results

$\rightarrow$ also gives a notion of distance and stability for Mappers...

Choice of parameters

Structure and Stability of the 1-Dimensional Mapper.
Carrière, O. 2016

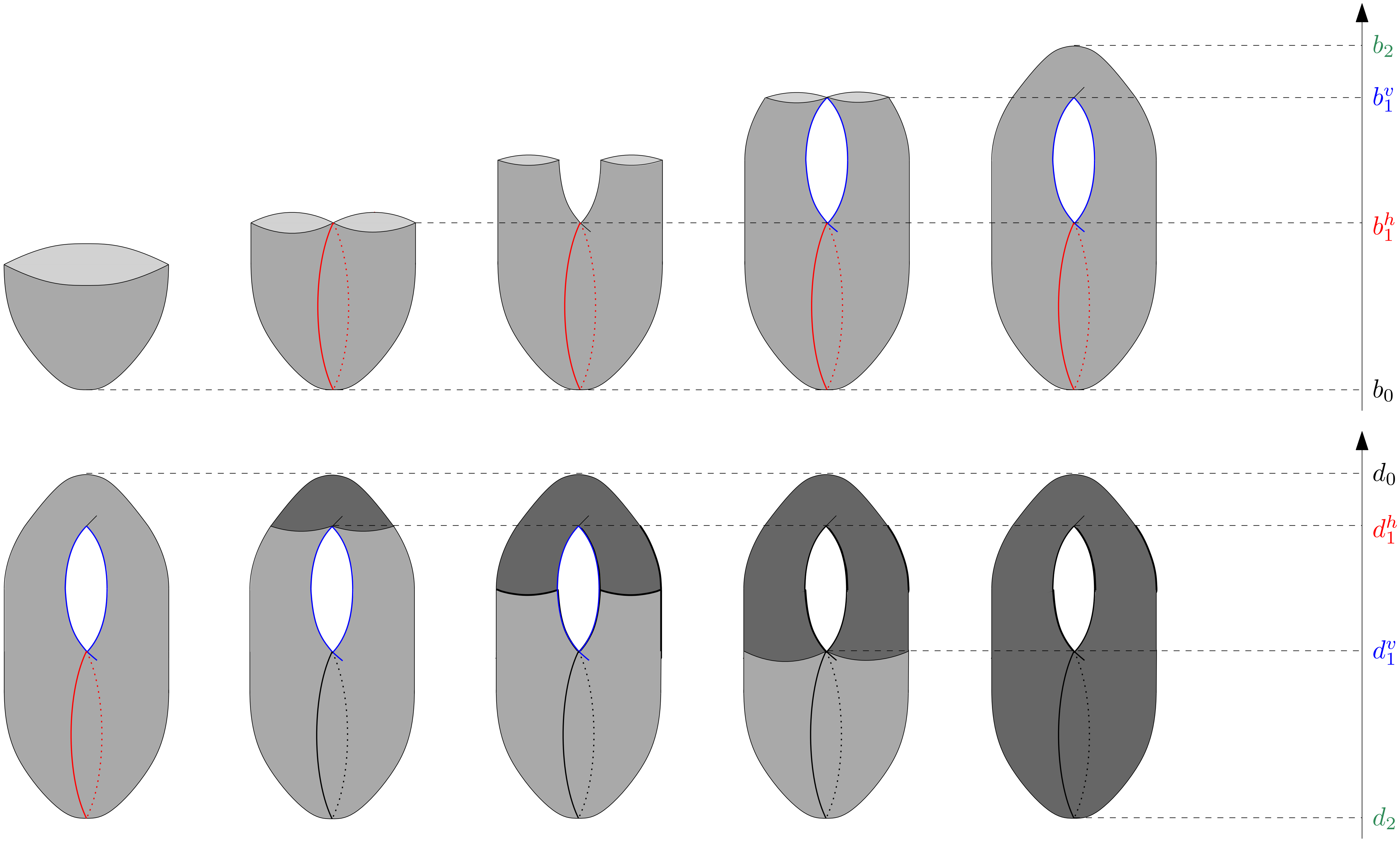
$\rightarrow$ clarifies formally the roles of r and g in the continous setting

$\rightarrow$ gives sufficient conditions on δ to get approximation results

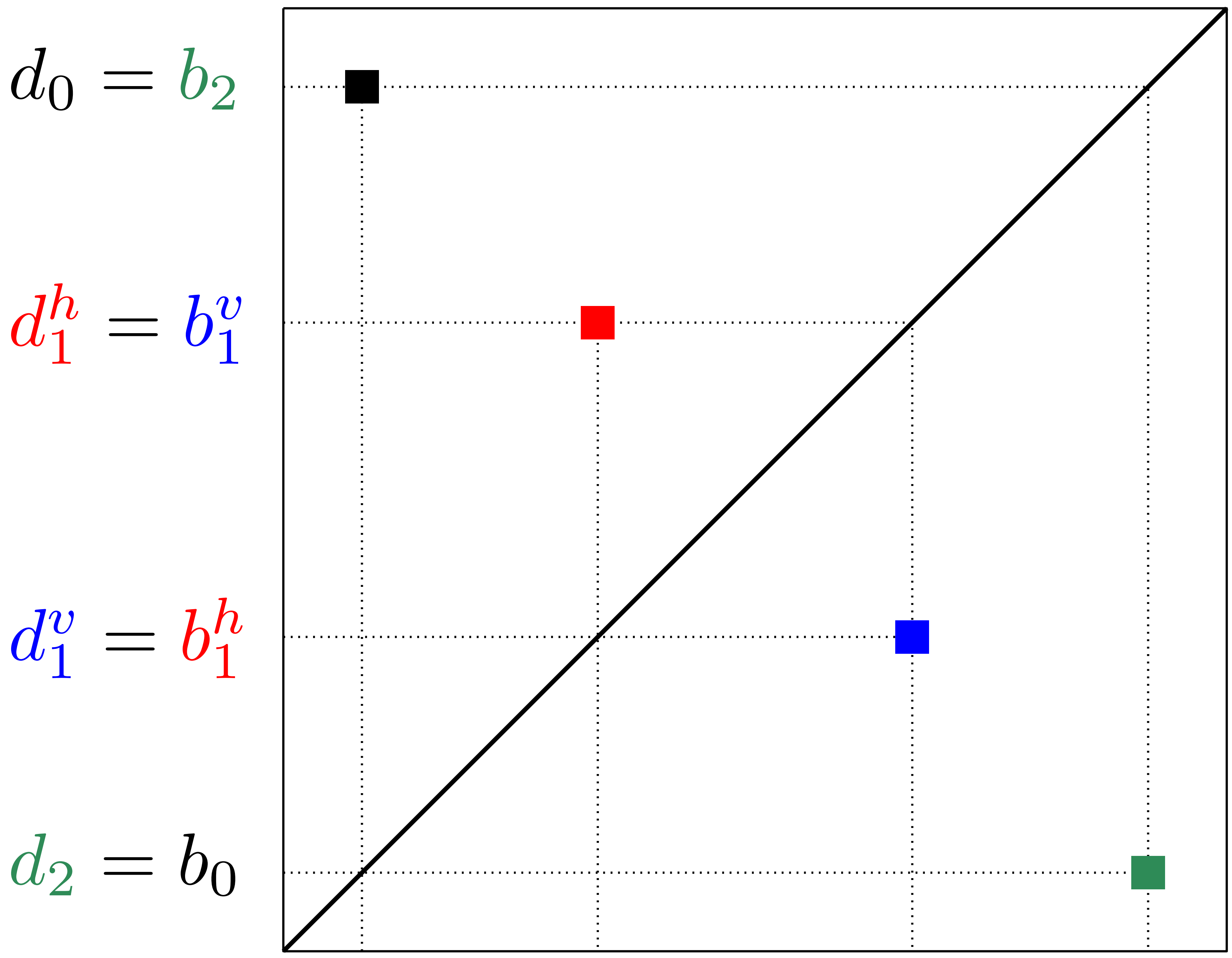
$\rightarrow$ also gives a notion of distance and stability for Mappers...

Other publications:
- de Silva, Munch, Patel. *Categorified Reeb Graphs*. 2015
- Munch, Wang. *Convergence between Categorical Representations of Reeb Space and Mapper*. 2016

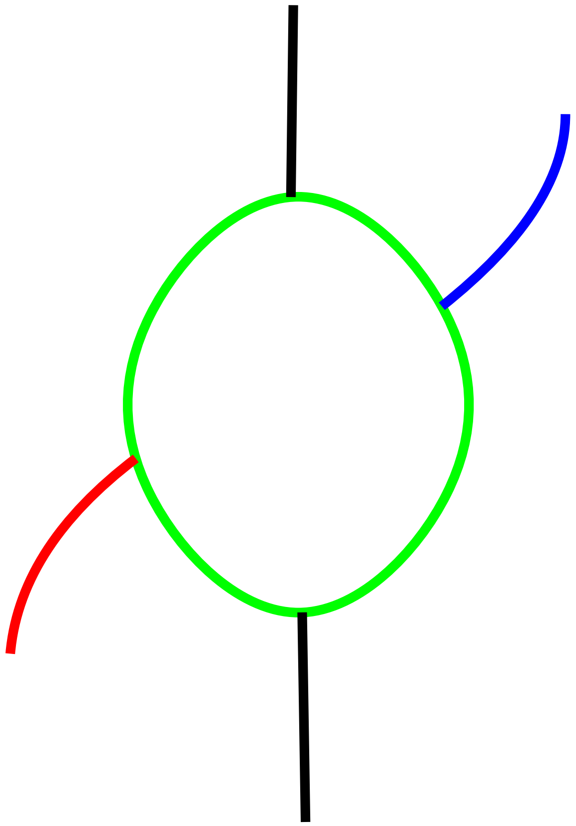
Extended Persistence

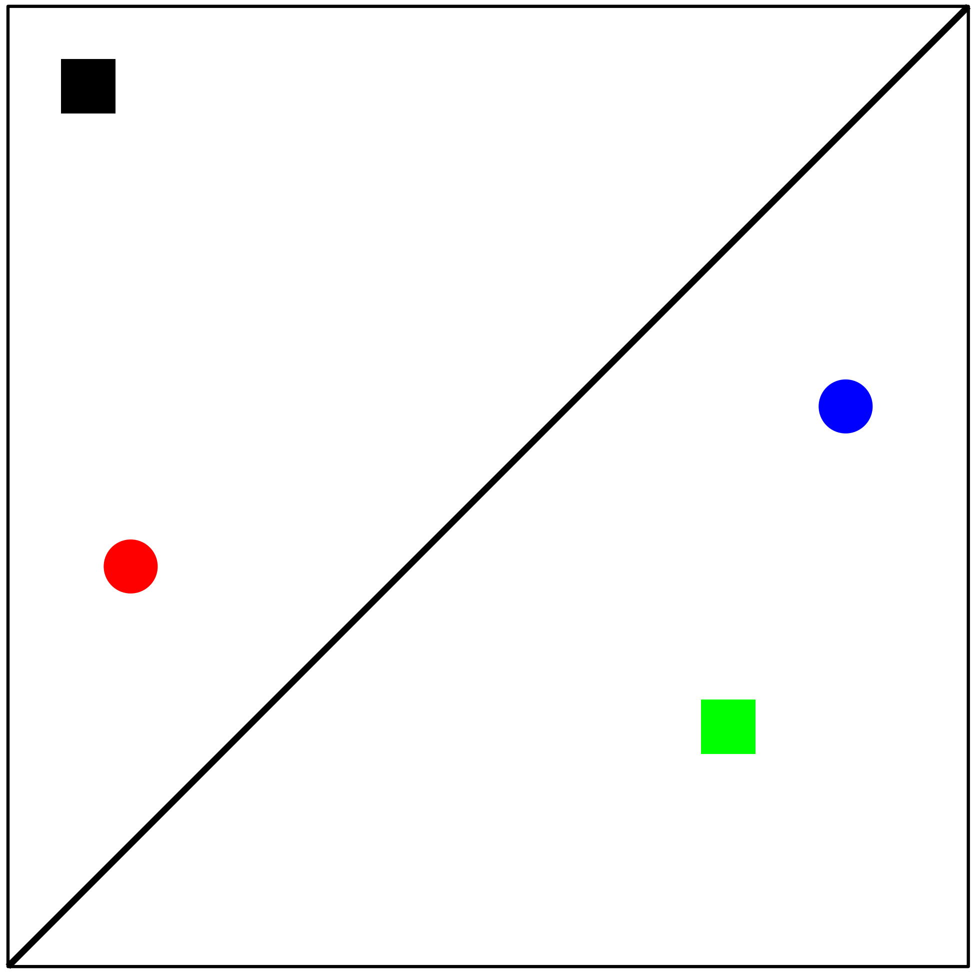


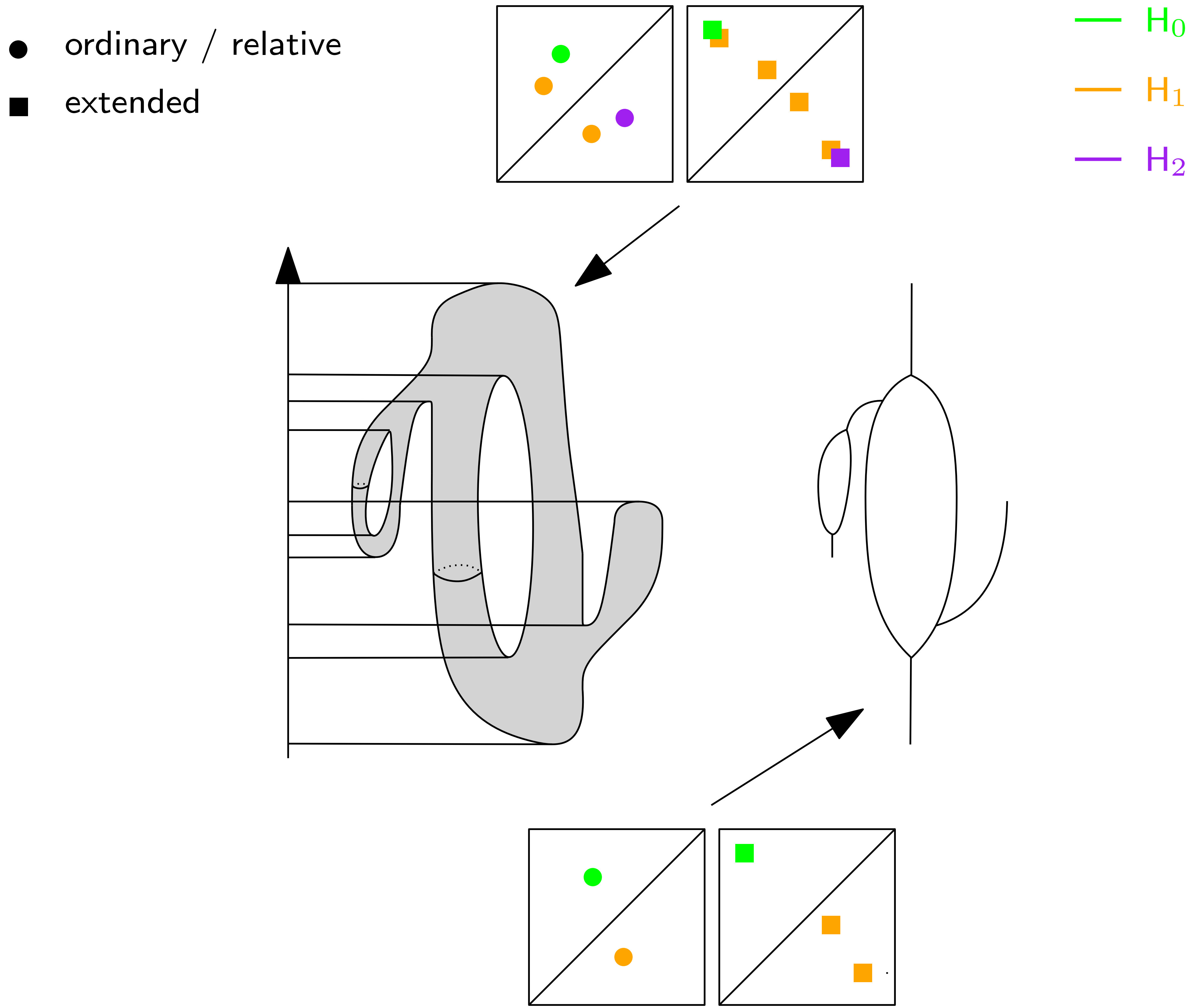
Extended Persistence

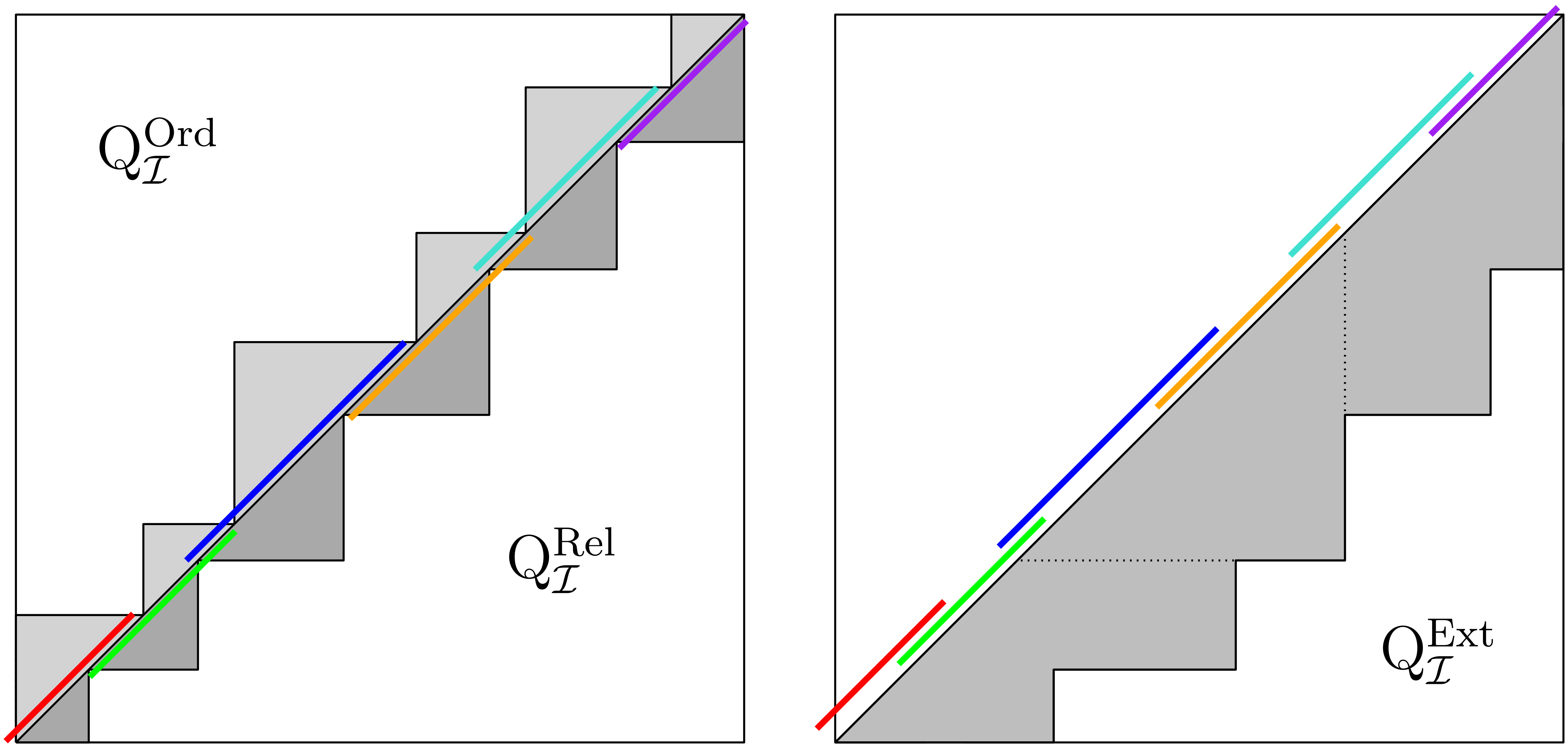


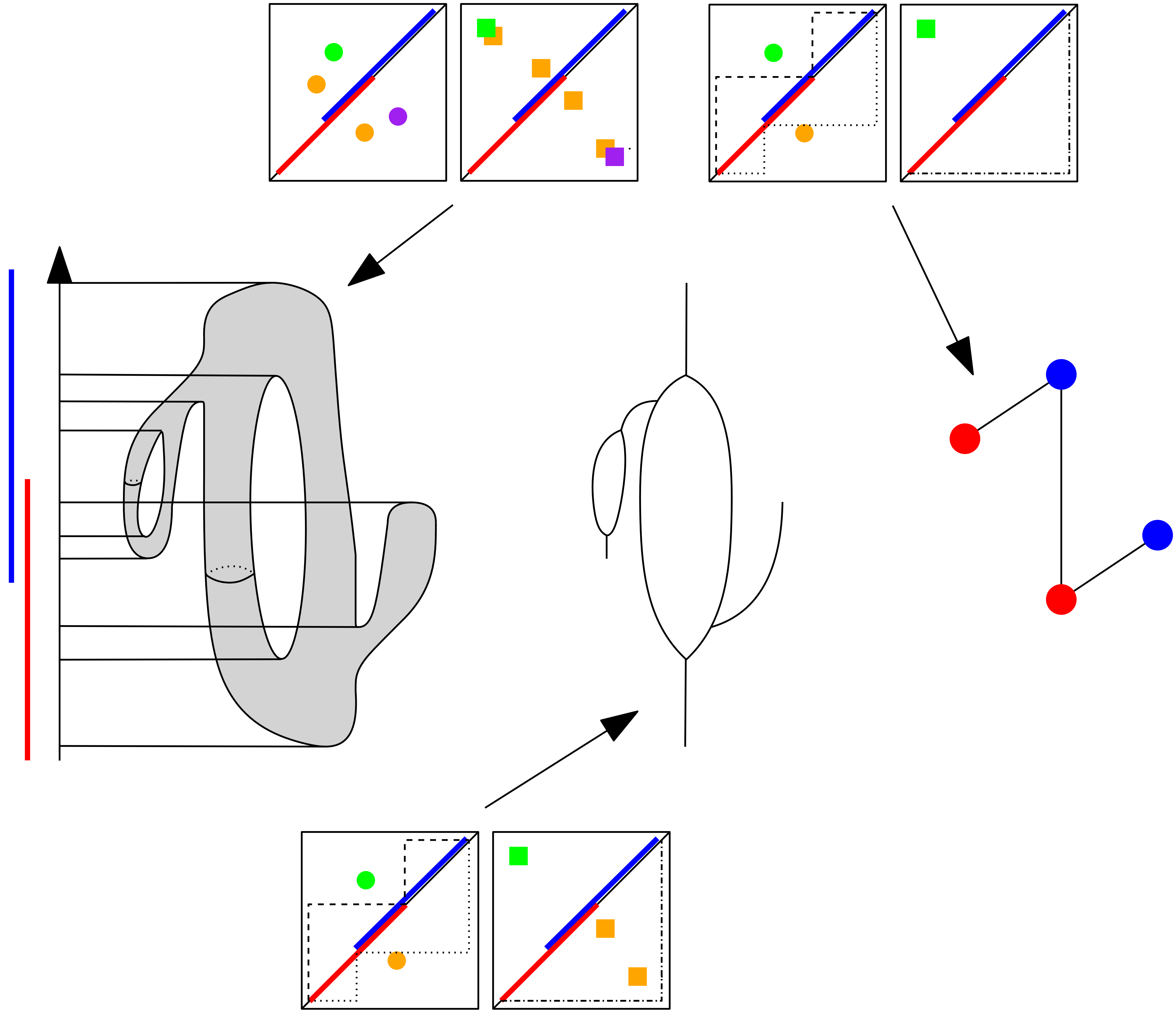
Extended Persistence

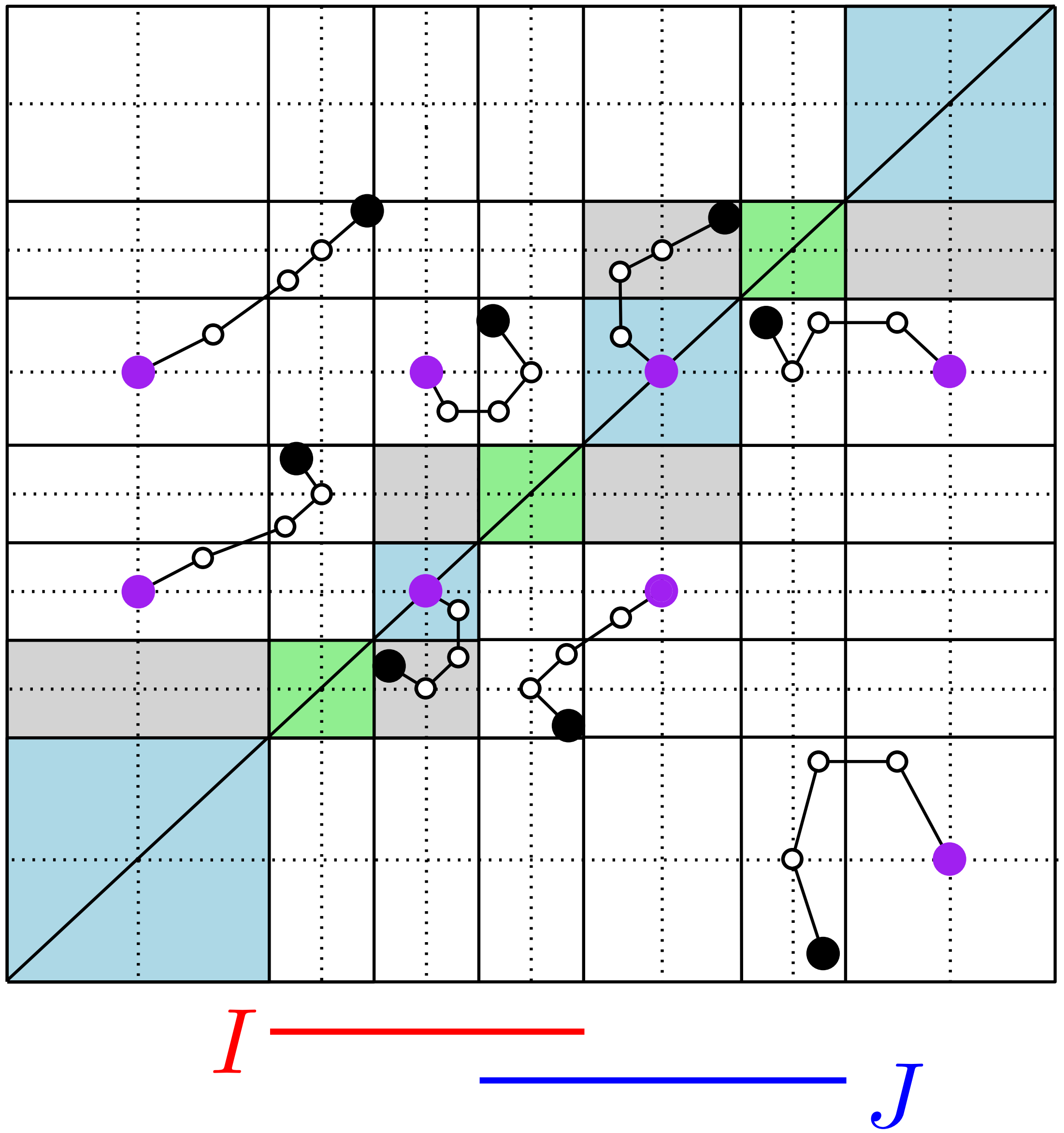




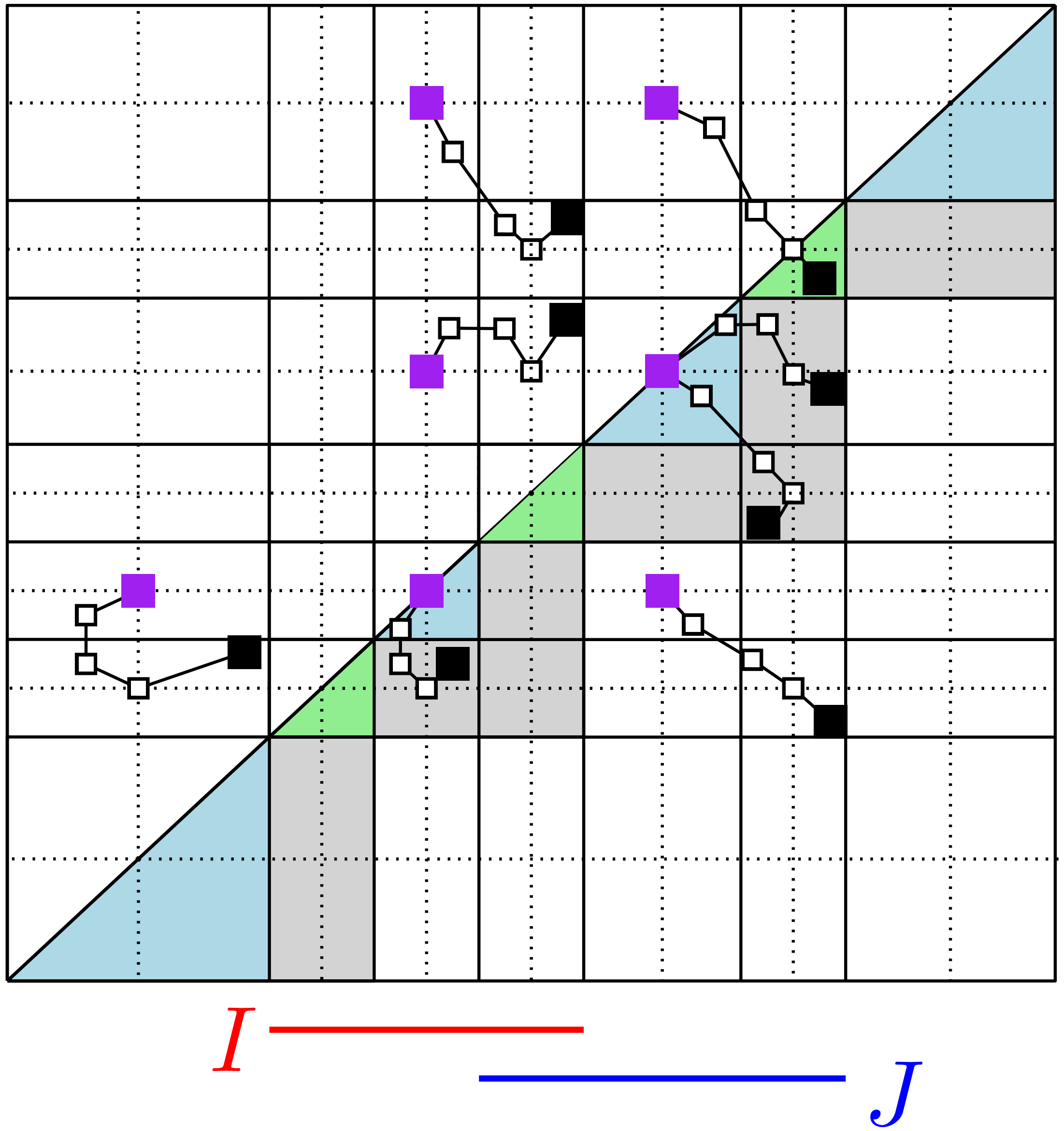

ordinary / relative
extended
H_0
H_1
H_2


$Q_{\mathcal{I}}^{\mathrm{Ord}}$
$Q_{\mathcal{I}}^{\mathrm{Rel}}$
$Q_{\mathcal{I}}^{\mathrm{Ext}}$





I
J



I
J