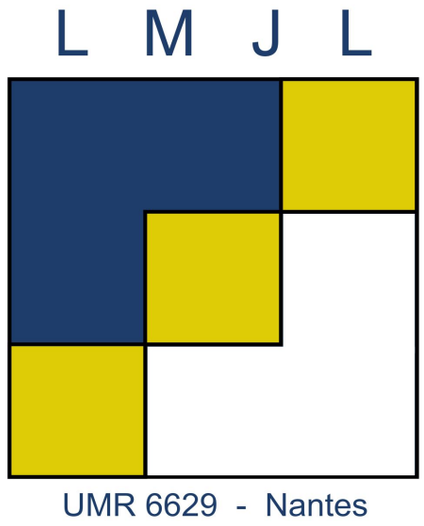
Journées de Géométrie Algorithmique
Aussois, Décembre 2017

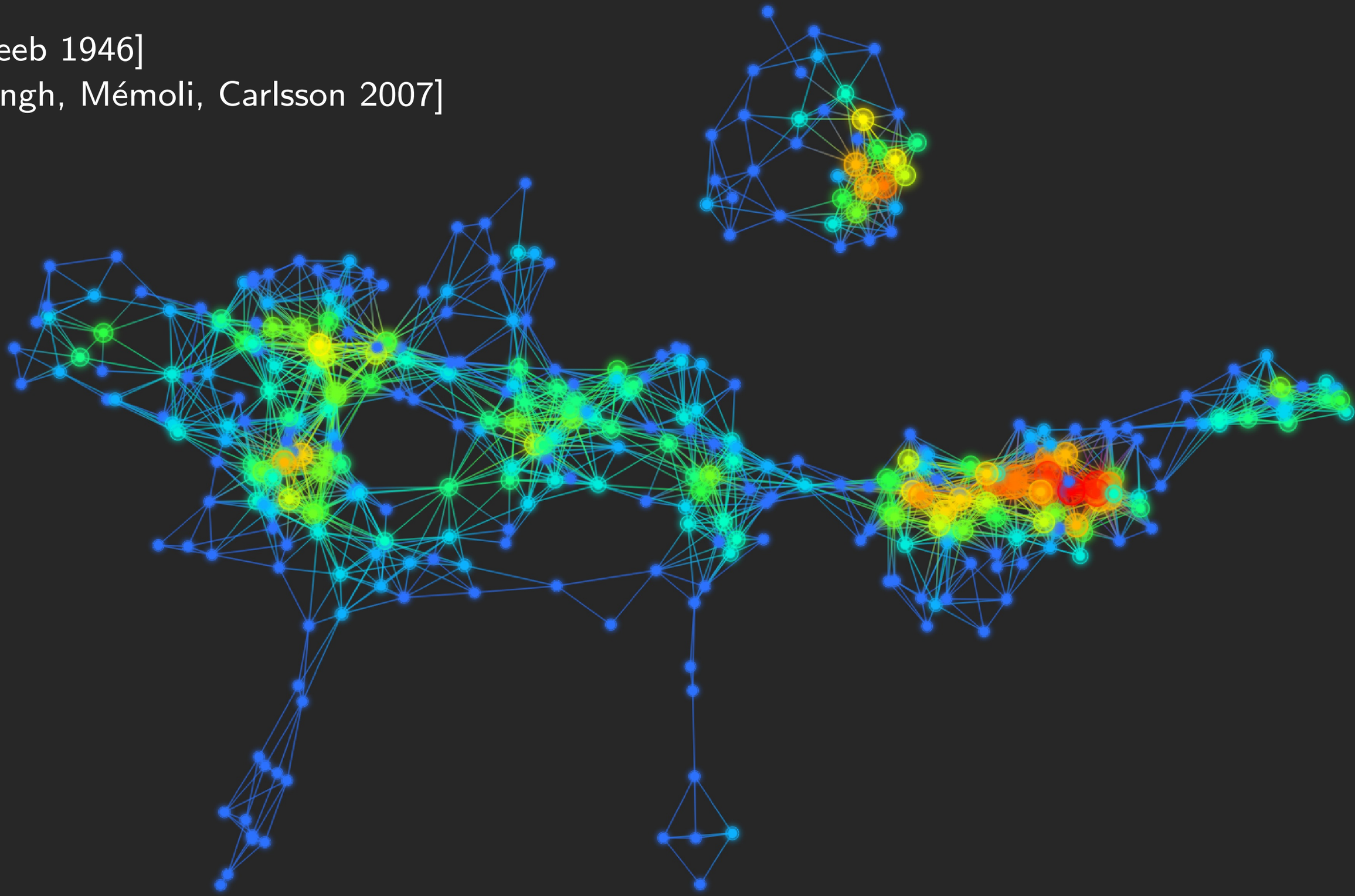
A theoretical framework for the analysis of Reeb graph and Mapper

Mathieu Carrière — joint work with B. Michel and S. Oudot



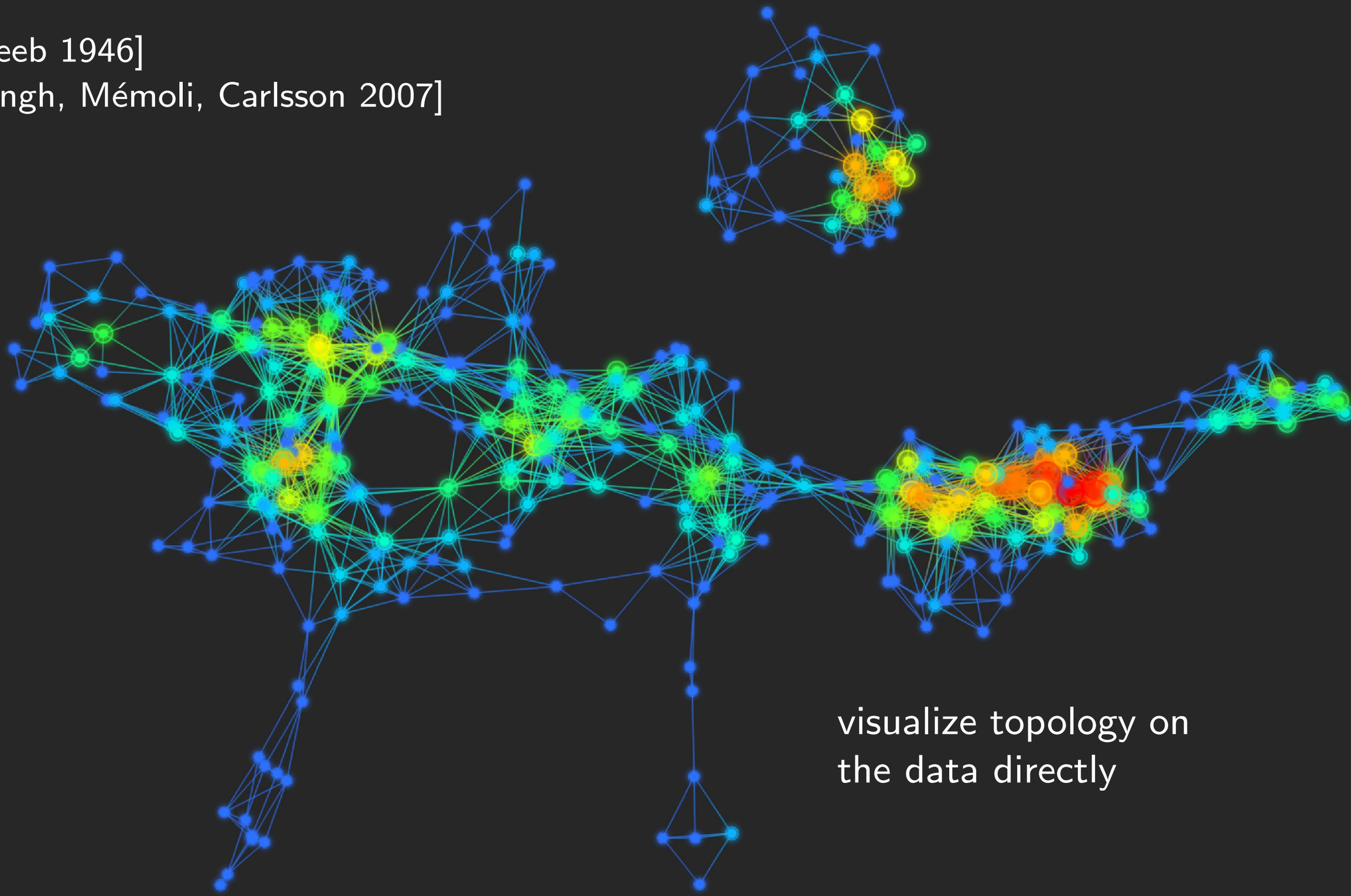
Reeb Graphs and Mapper

[Reeb 1946]
[Singh, Mémoli, Carlsson 2007]



Reeb Graphs and Mapper

[Reeb 1946]
[Singh, Mémoli, Carlsson 2007]



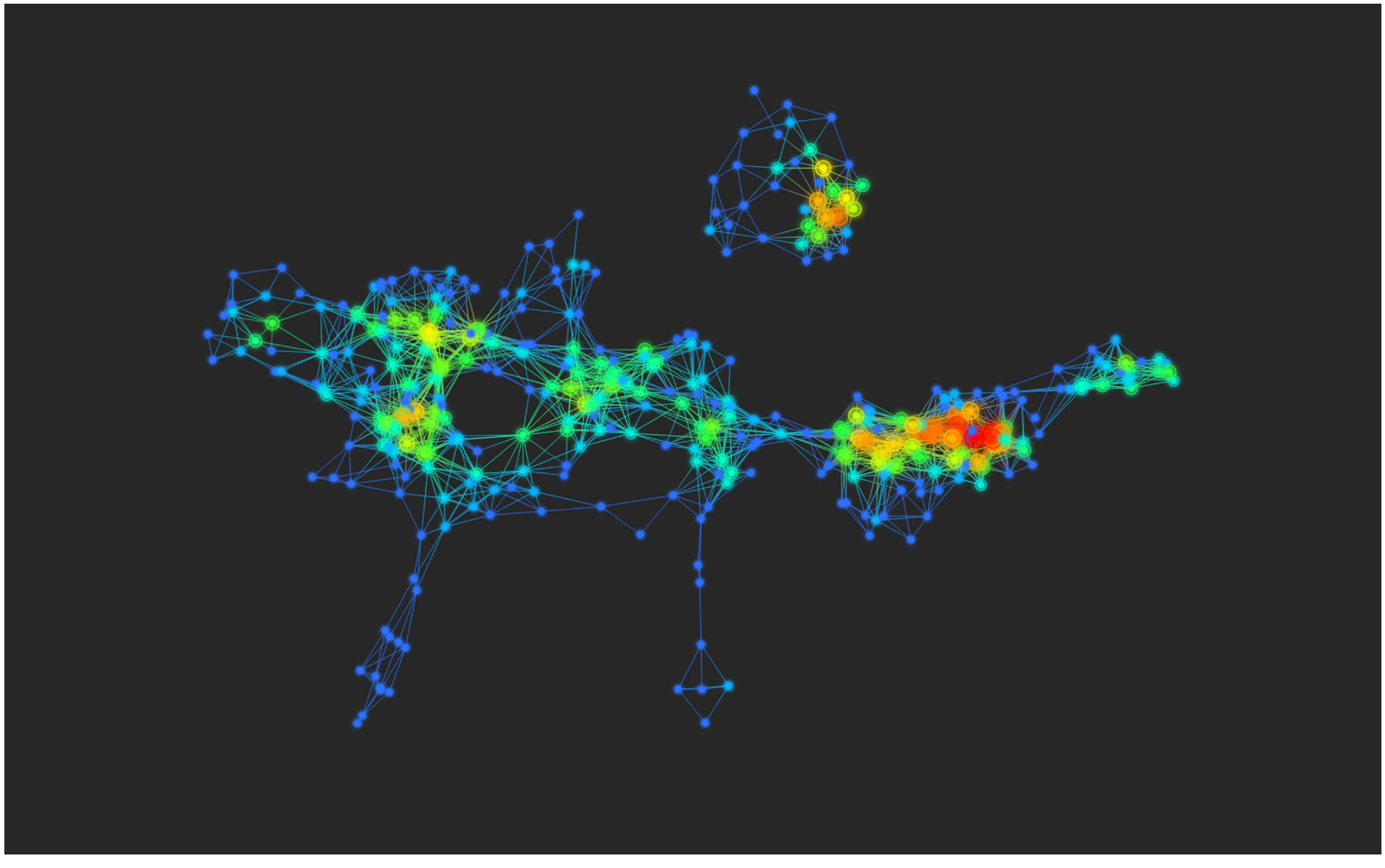
visualize topology on
the data directly

Applications

$\longrightarrow$ visualization

$\longrightarrow$ clustering

$\longrightarrow$ feature selection



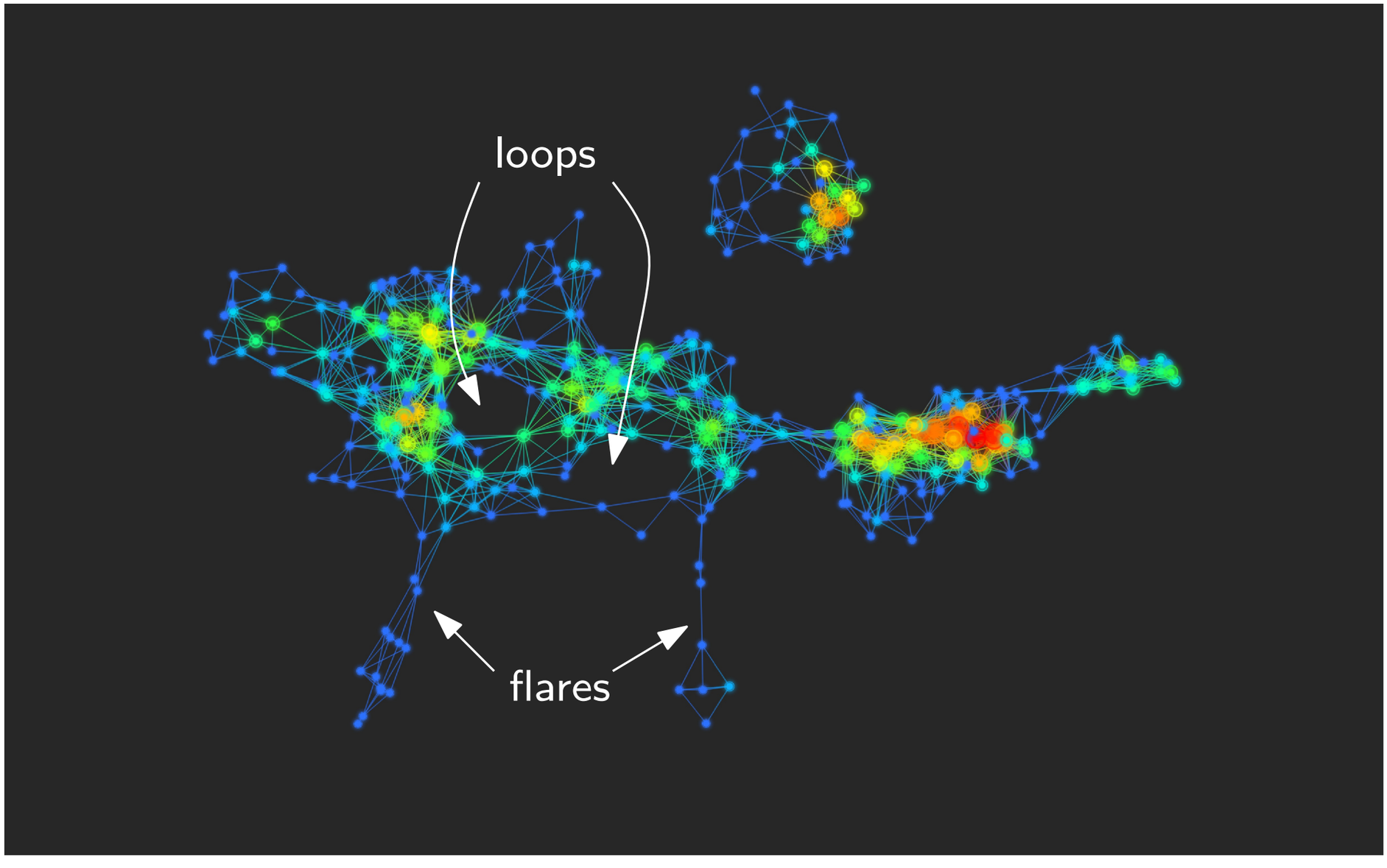
Applications

$\longrightarrow$ visualization

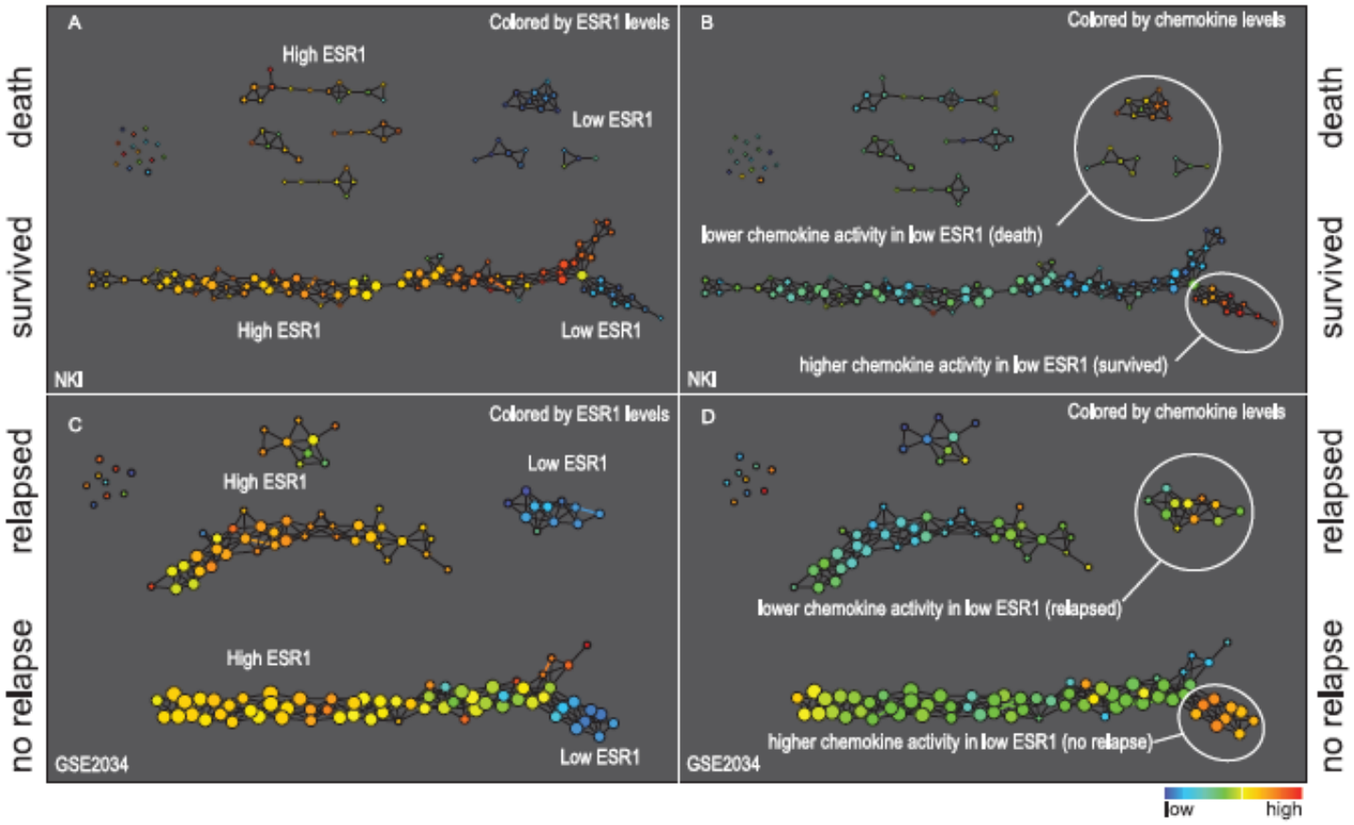
$\longrightarrow$ clustering

$\longrightarrow$ feature selection

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

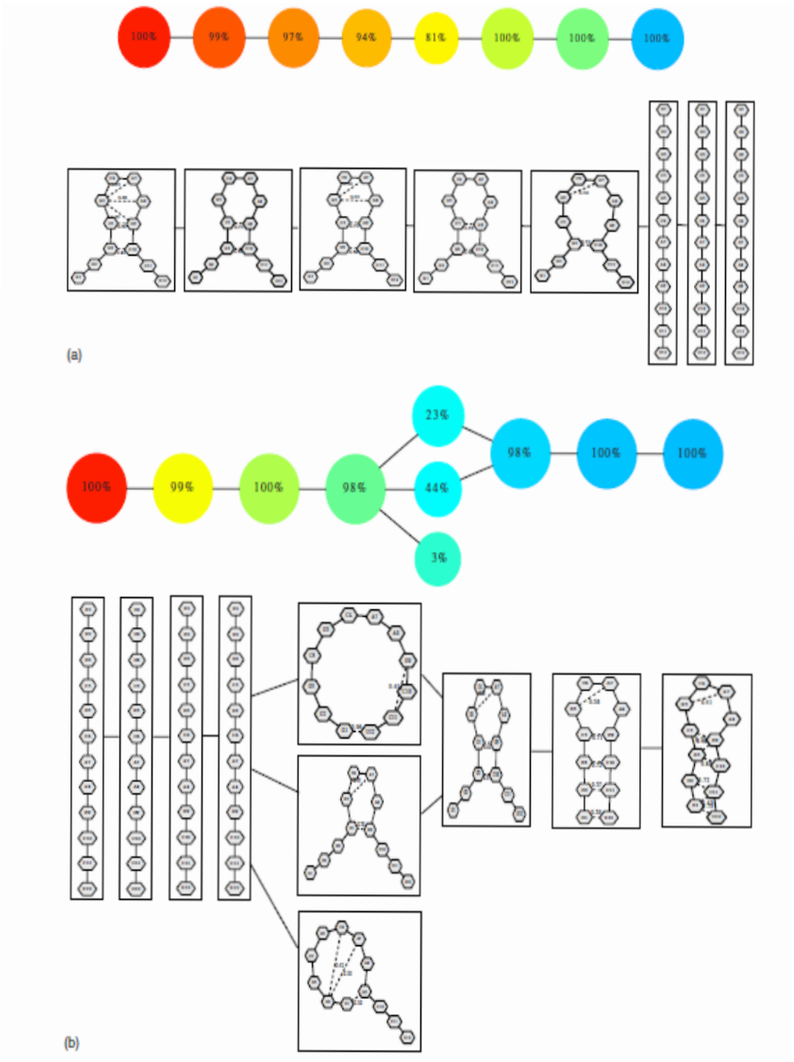


Applications



breast cancer subtype identification

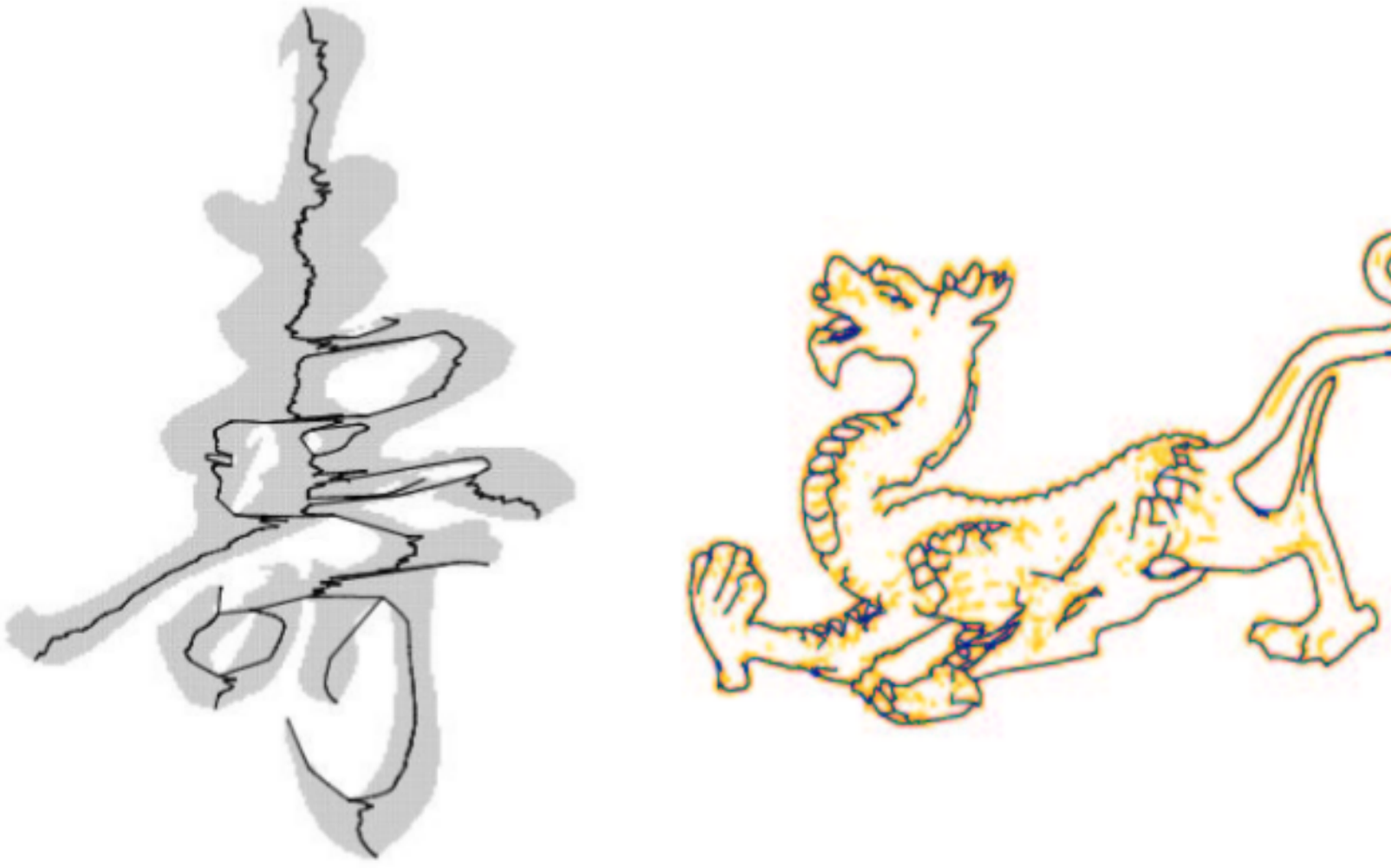
[Nicolau et al. 2011]



protein folding pathways

[Yao et al. 2009]

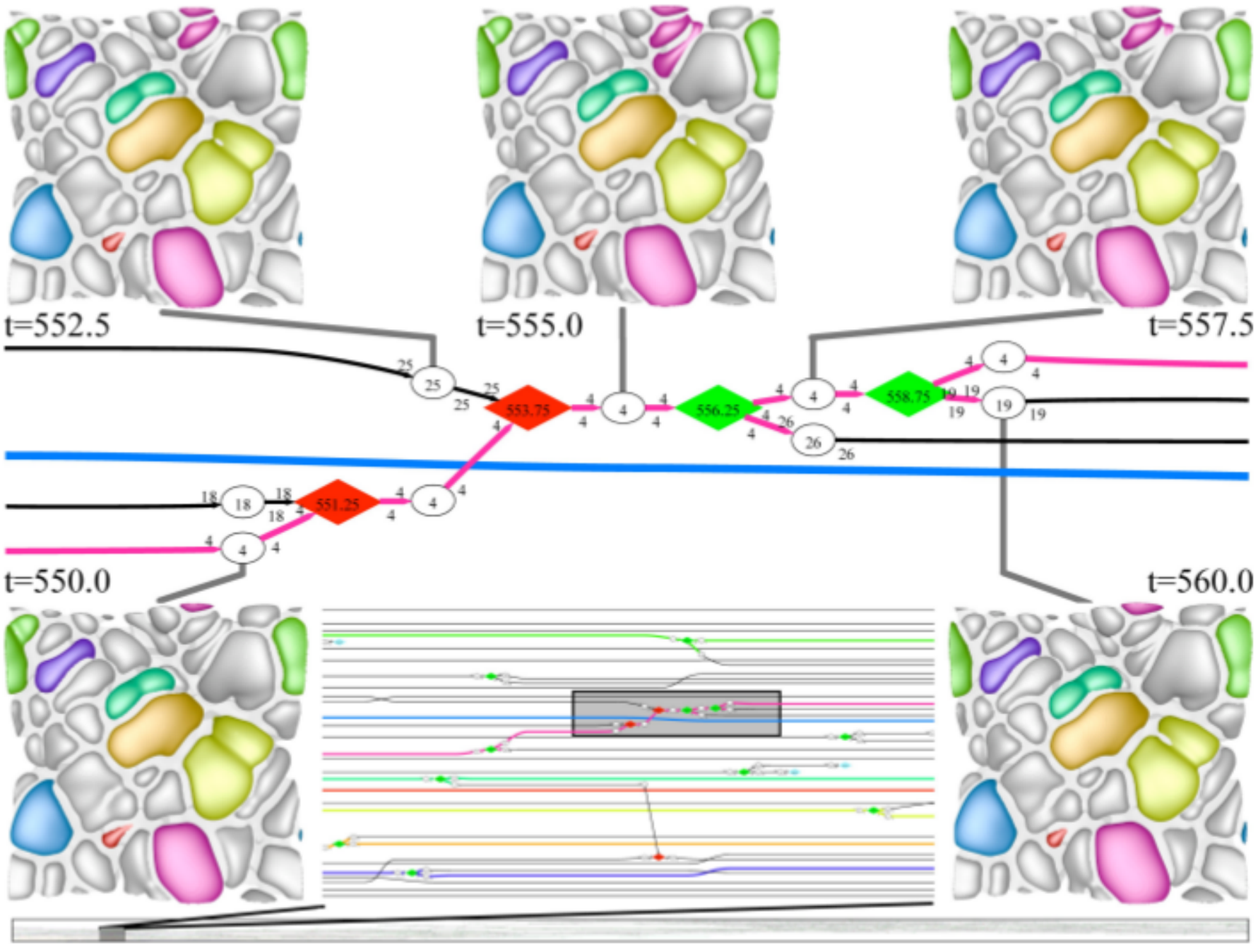
Applications



Data Skeletonization

[Ge et al. 2011]

Applications



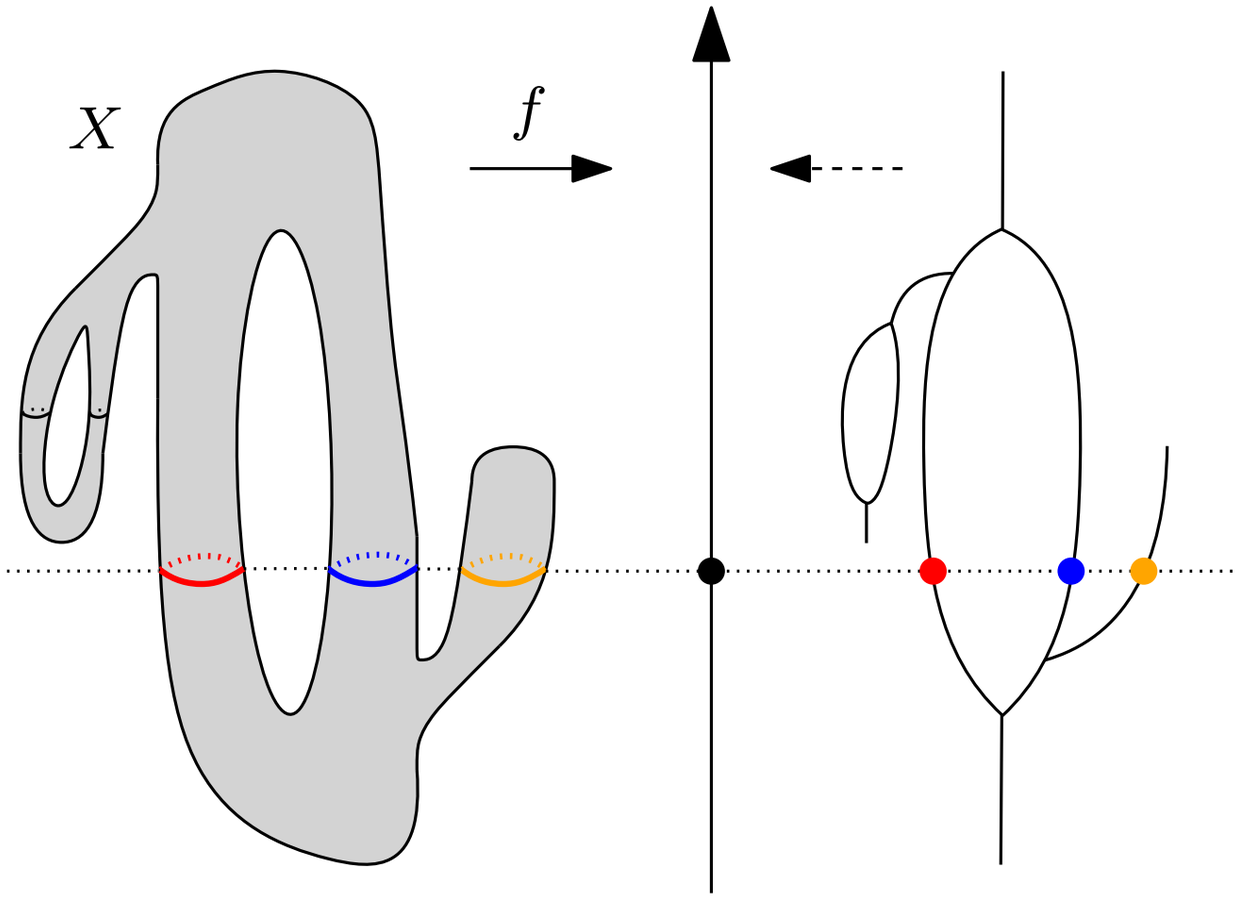
Burning regions of lean hydrogen flame over time

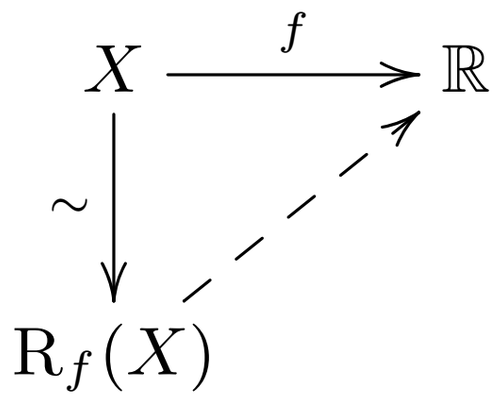
[Weber et al. 2011]

Reeb Graphs

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

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





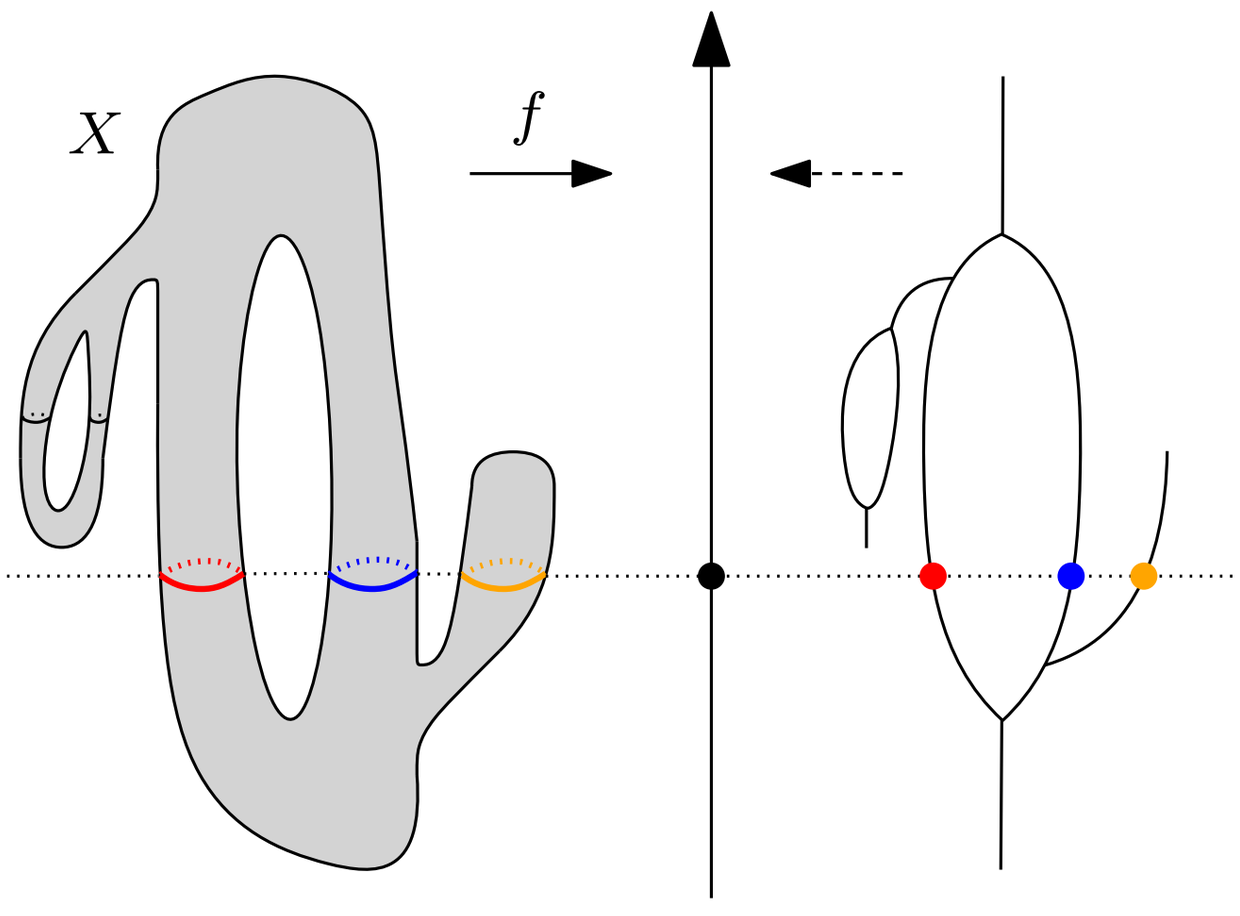
Prop: $\mathrm{R}_f(X)$ is a graph if (X, f) is of **Morse type**

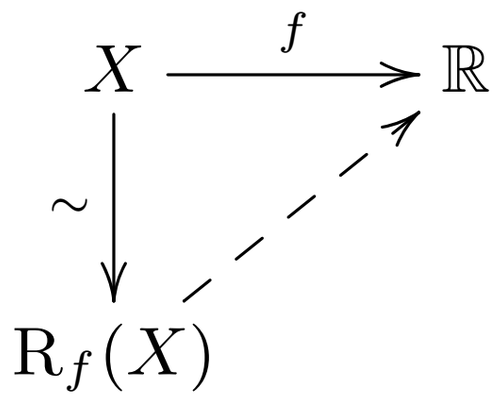
Reeb Graphs

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

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

Caveat: computation from point cloud is difficult





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

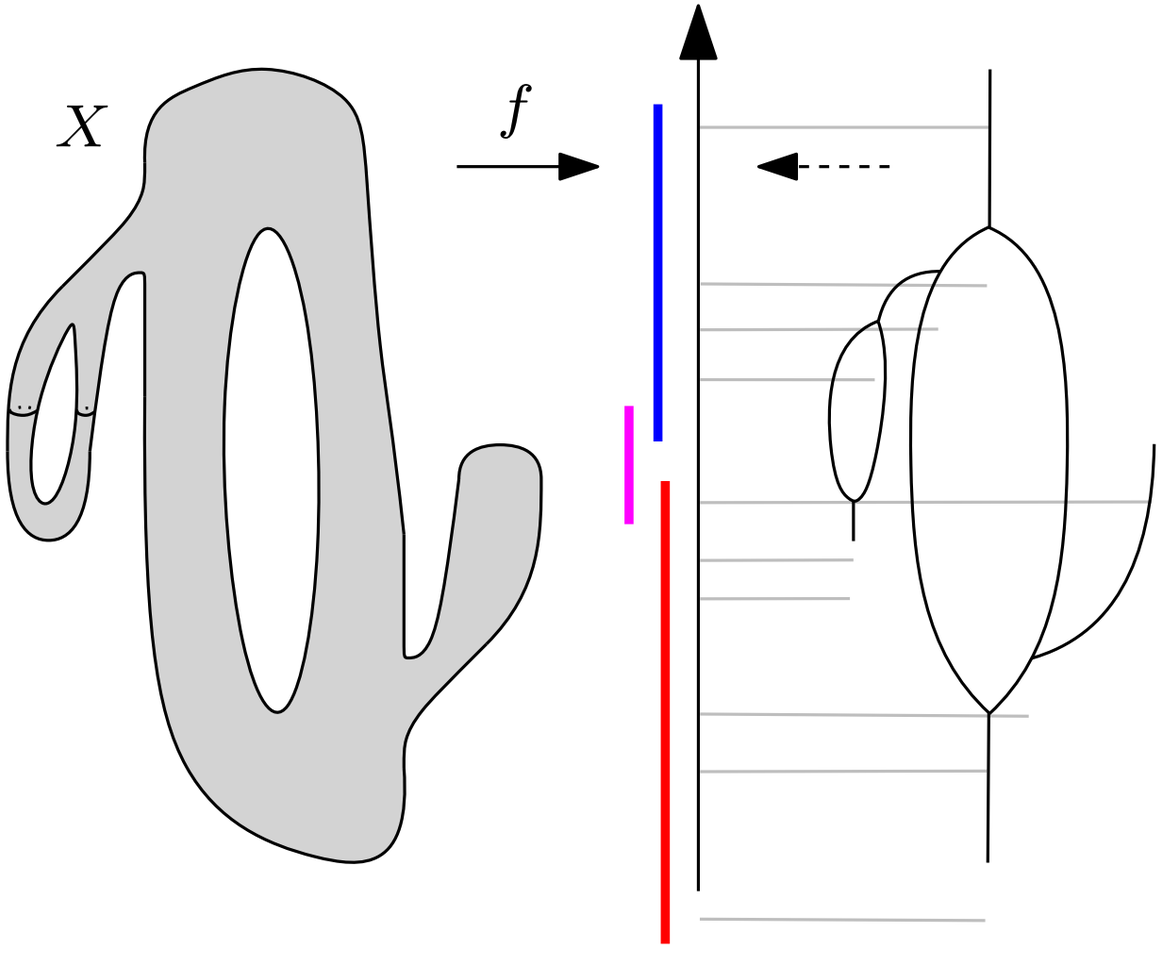
Reeb Graphs

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

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

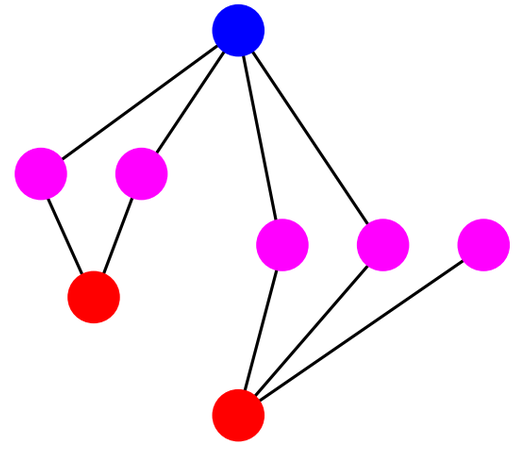
Caveat: computation from point cloud is difficult

$\rightarrow$ Proxy: **Mapper**

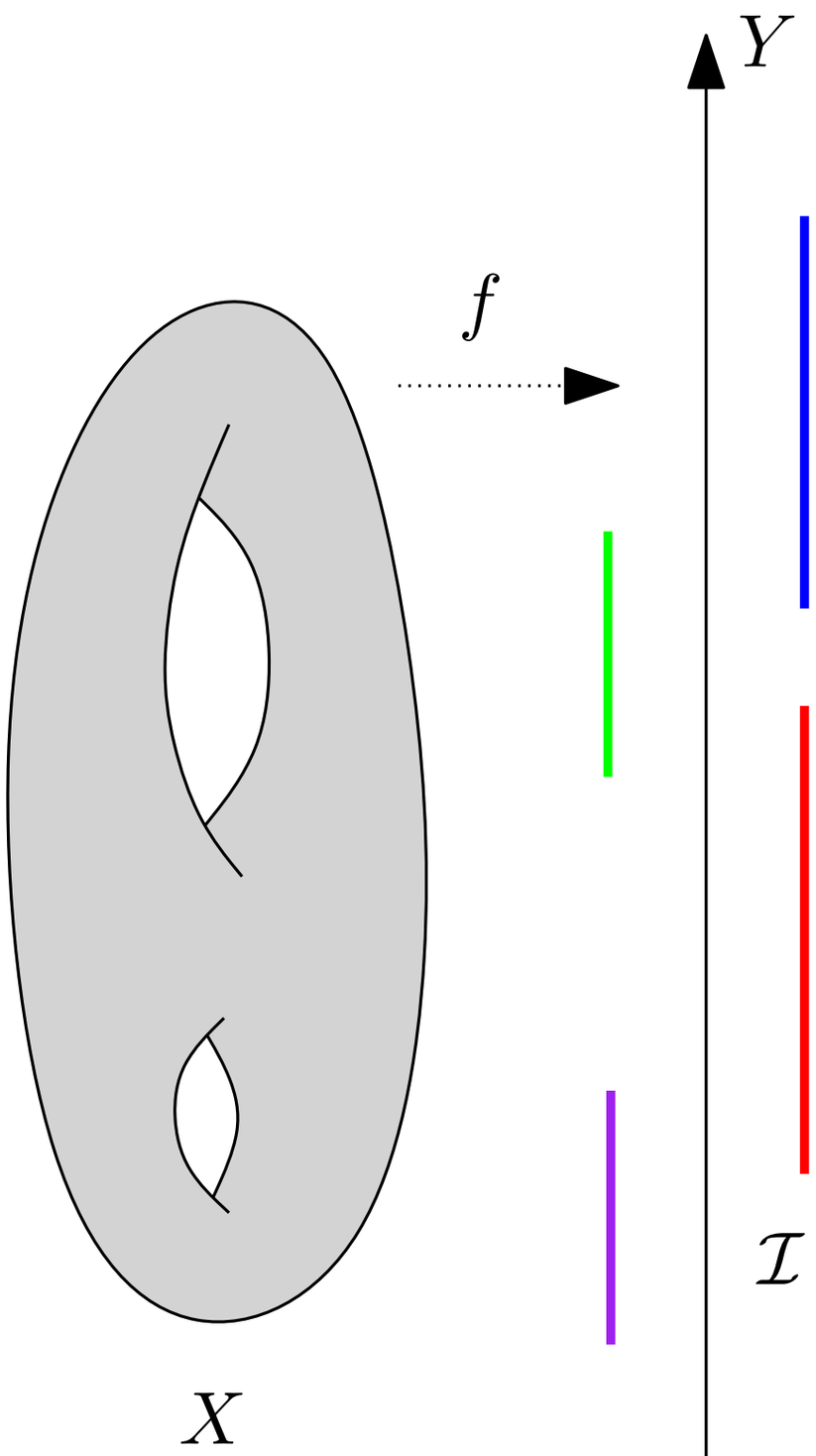


Pixelized Reeb graph

Singleton $\rightarrow$ interval: $f^{-1}(I)$



Mapper (continuous setting)



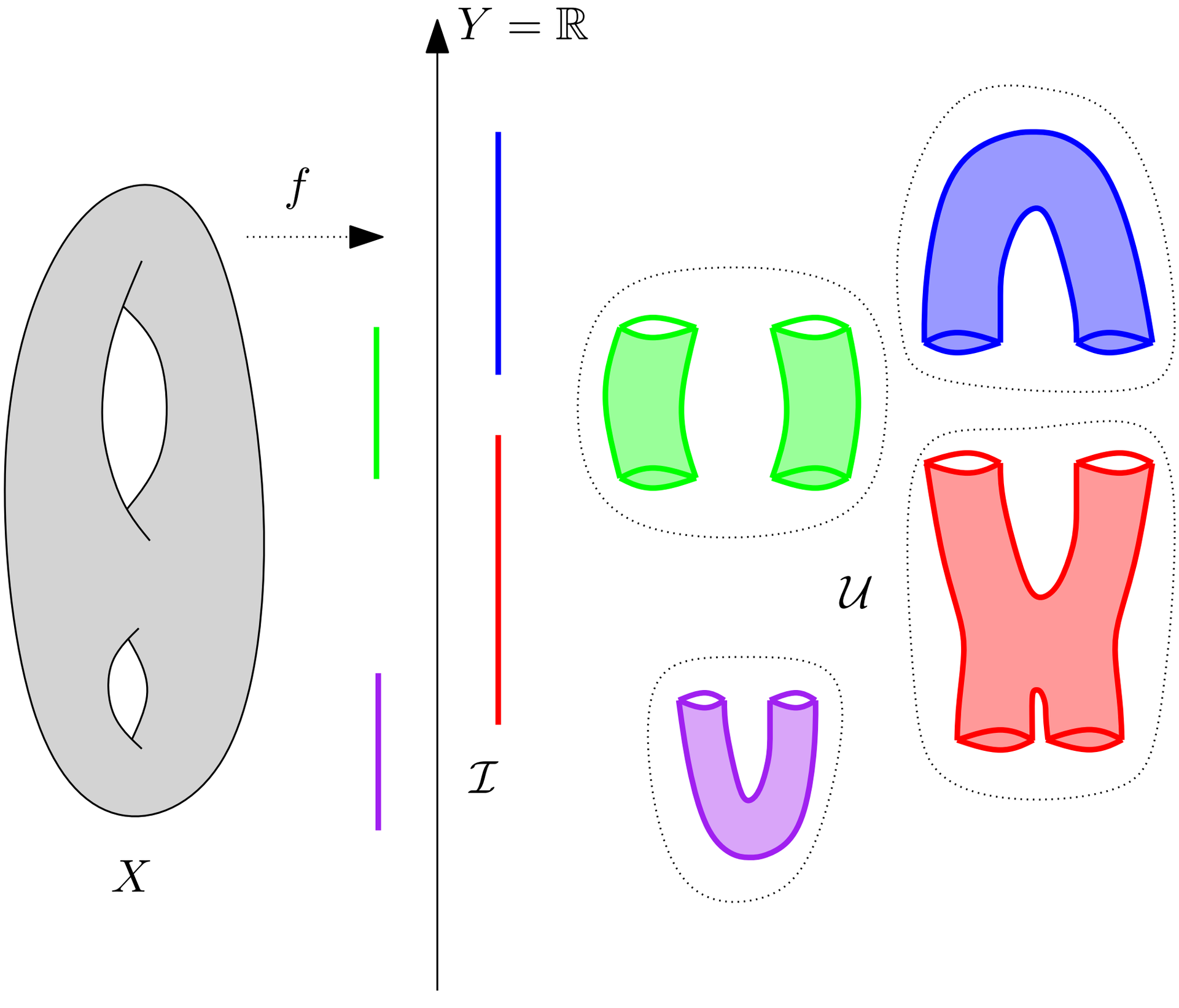
X

f

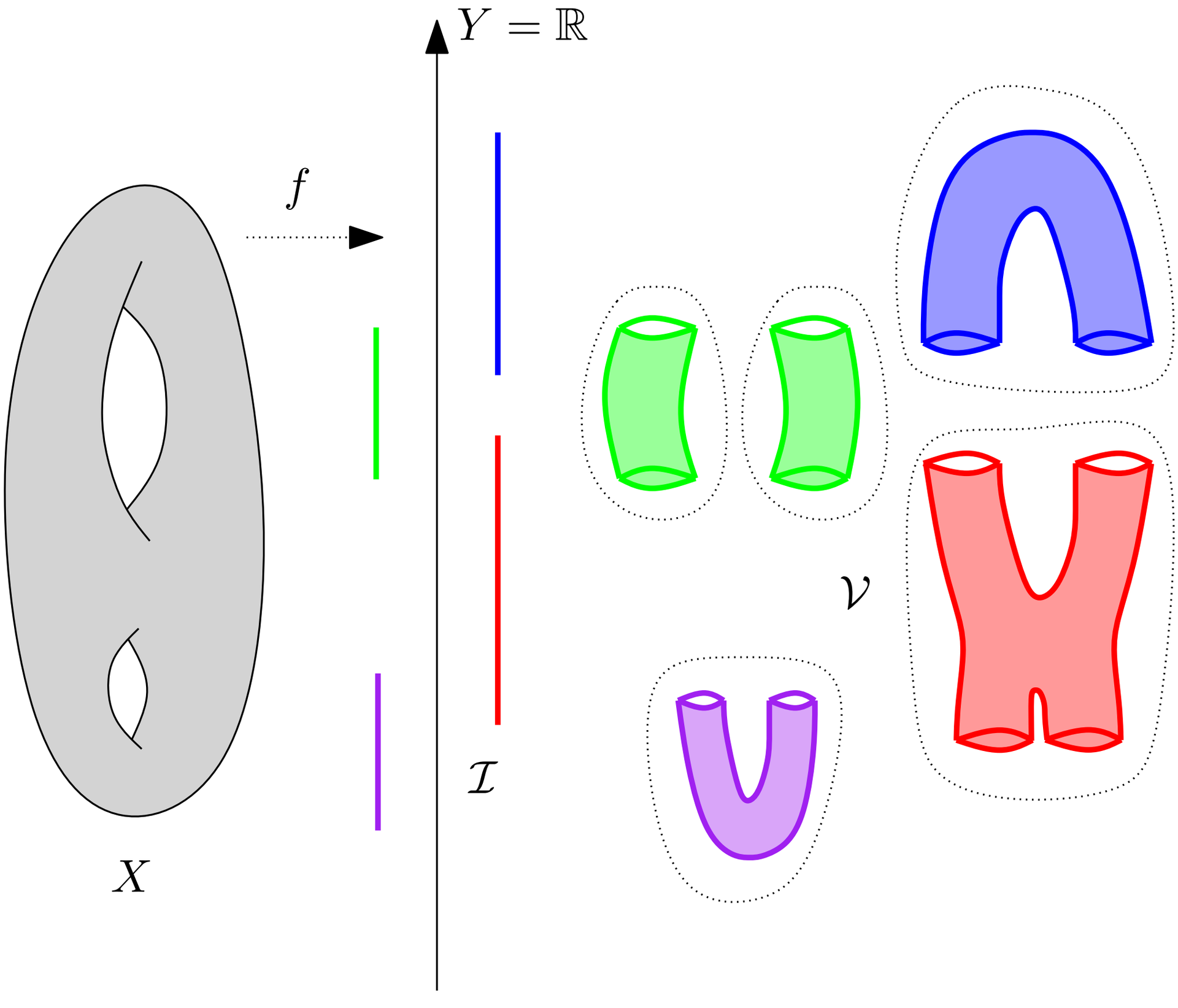
$Y = \mathbb{R}$

$\mathcal{I}$

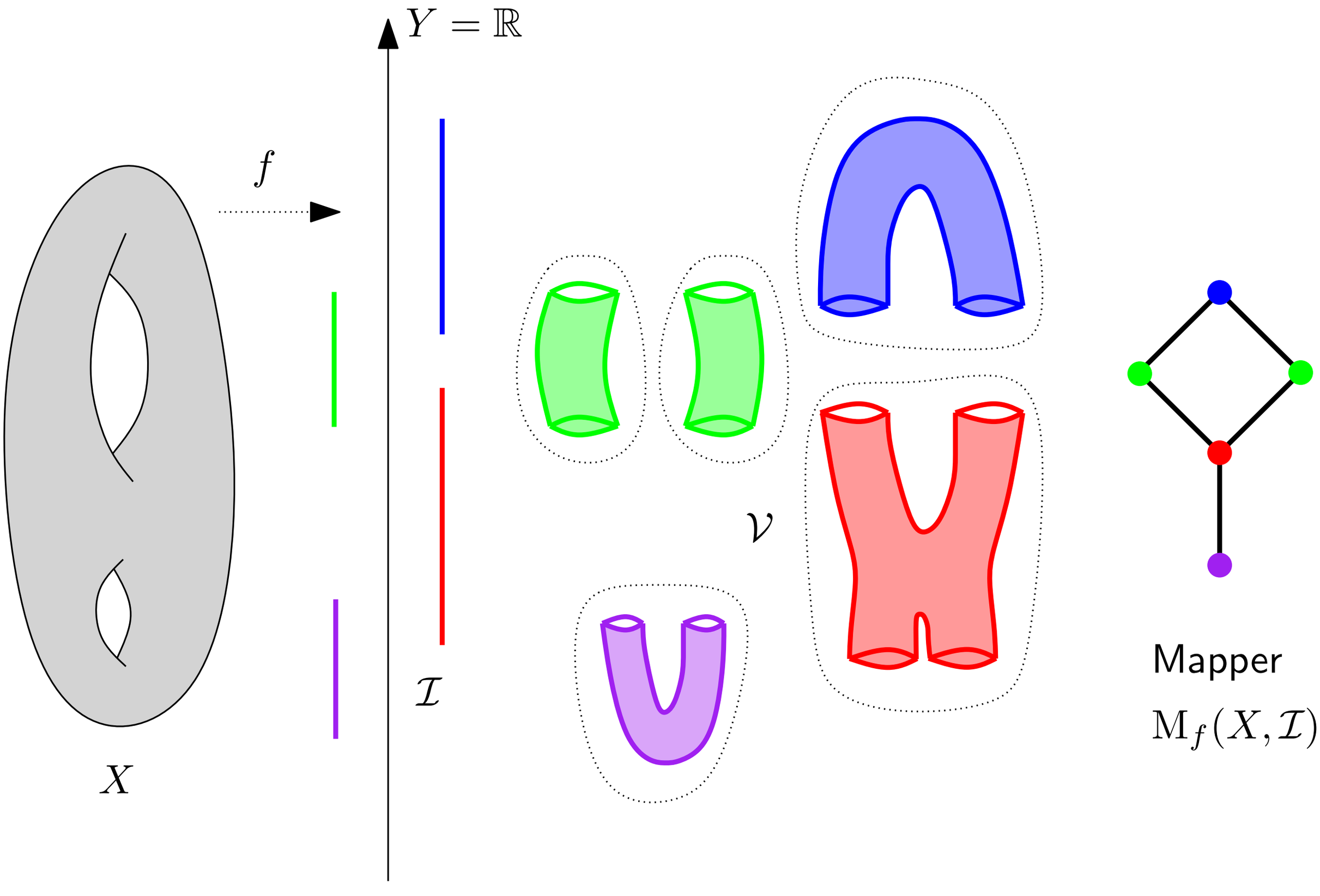
Mapper (continuous setting)



Mapper (continuous setting)



Mapper (continuous setting)



Mapper (continuous setting)

Input:

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

Method:

- Compute *pullback cover* $\mathcal{U}$ of X: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$
- Refine $\mathcal{U}$ by separating the connected components
- 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 (discrete setting)

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

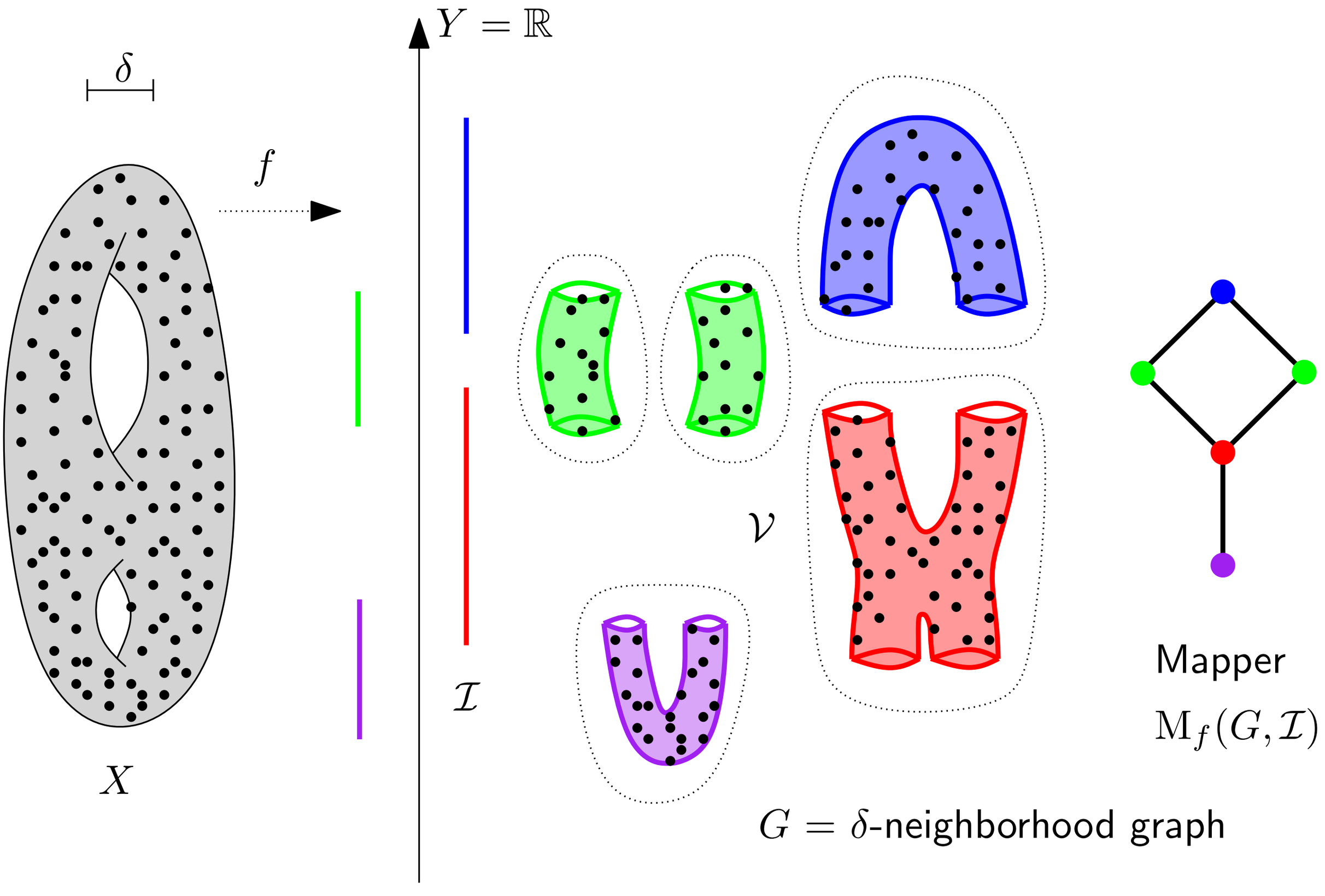
• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in $G \to$ connected cover $\mathcal{V}$

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

(intersections materialized by data points)

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

Mapper (discrete setting)



Parameters

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

lens or filter

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

range scale

- neighborhood size δ

geometric scale

Parameters

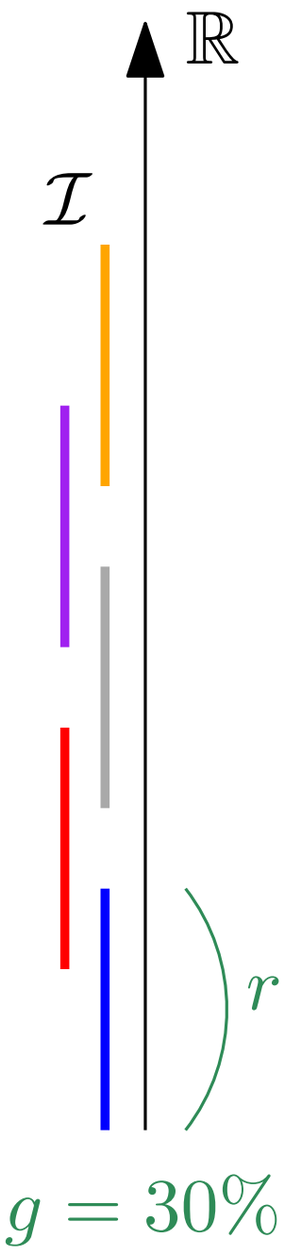
- function $f : P \to \mathbb{R}$ — lens or filter
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals
- neighborhood size δ

geometric scale

range scale

uniform cover $\mathcal{I}$:

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



Parameters

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

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

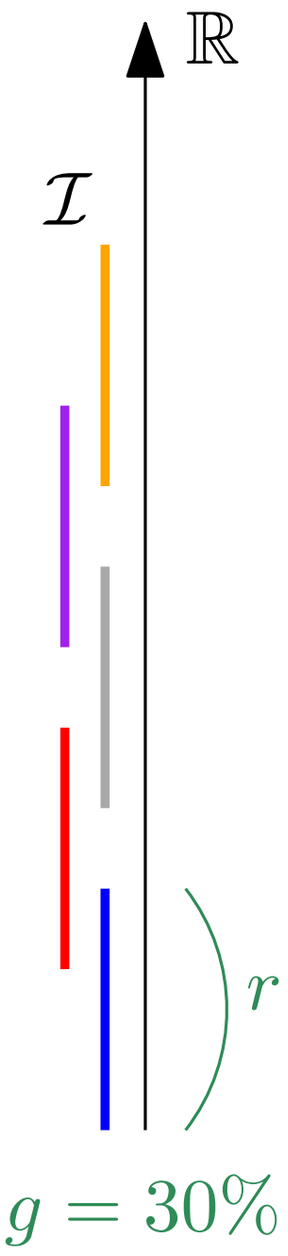
- neighborhood size δ ← geometric scale

uniform cover $\mathcal{I}$:

- resolution / granularity: r (diameter of intervals)

- gain: g (percentage of overlap)

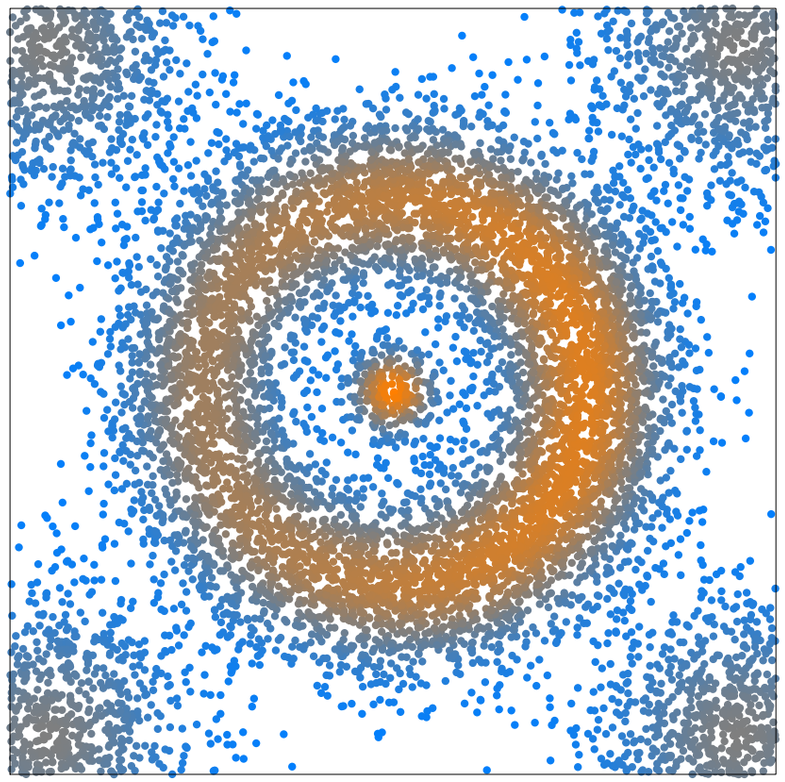
In practice: trial-and-error

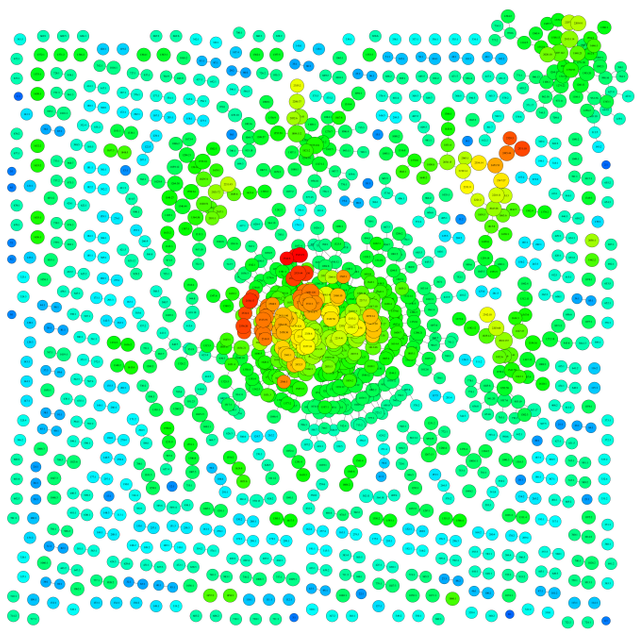


Examples

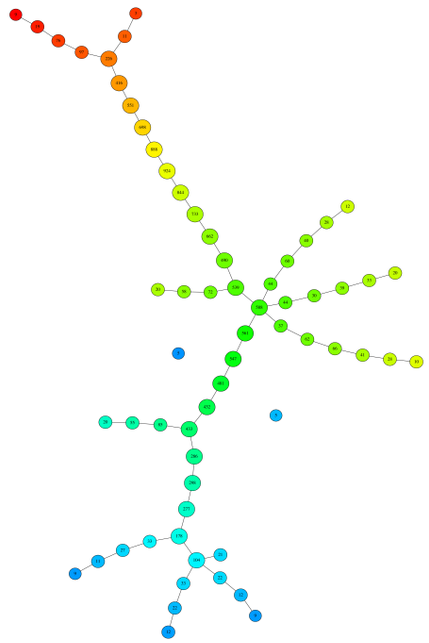
$\hat{f}$ = density estimator

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

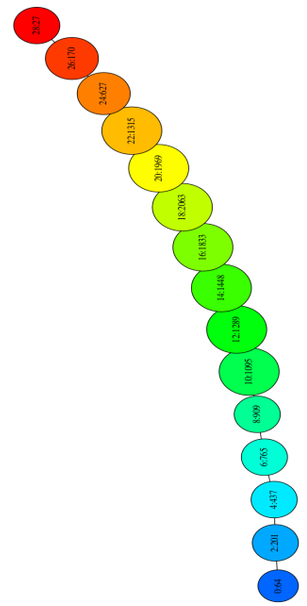




$\delta = 0.1\%$

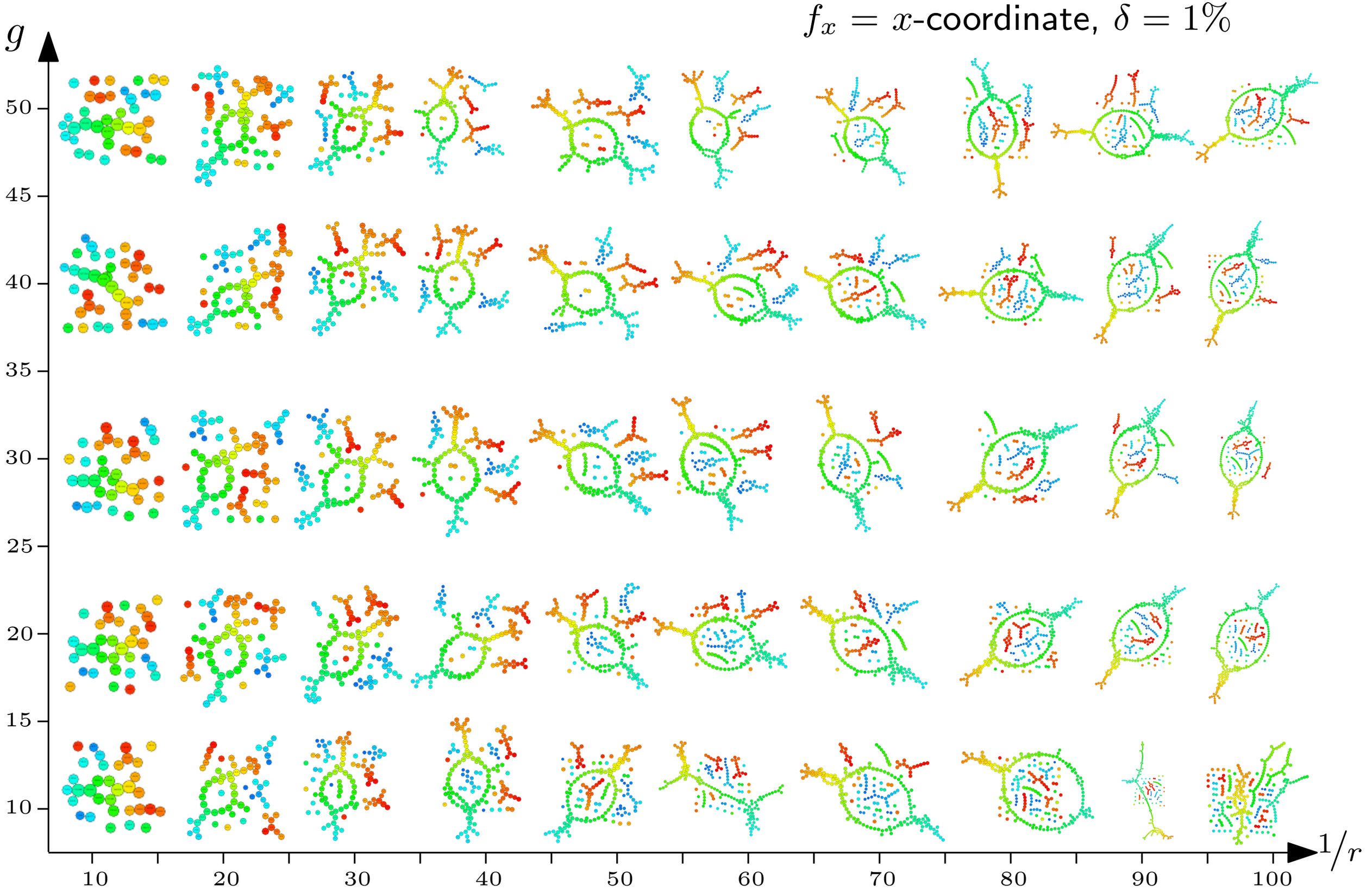


$\delta = 1\%$



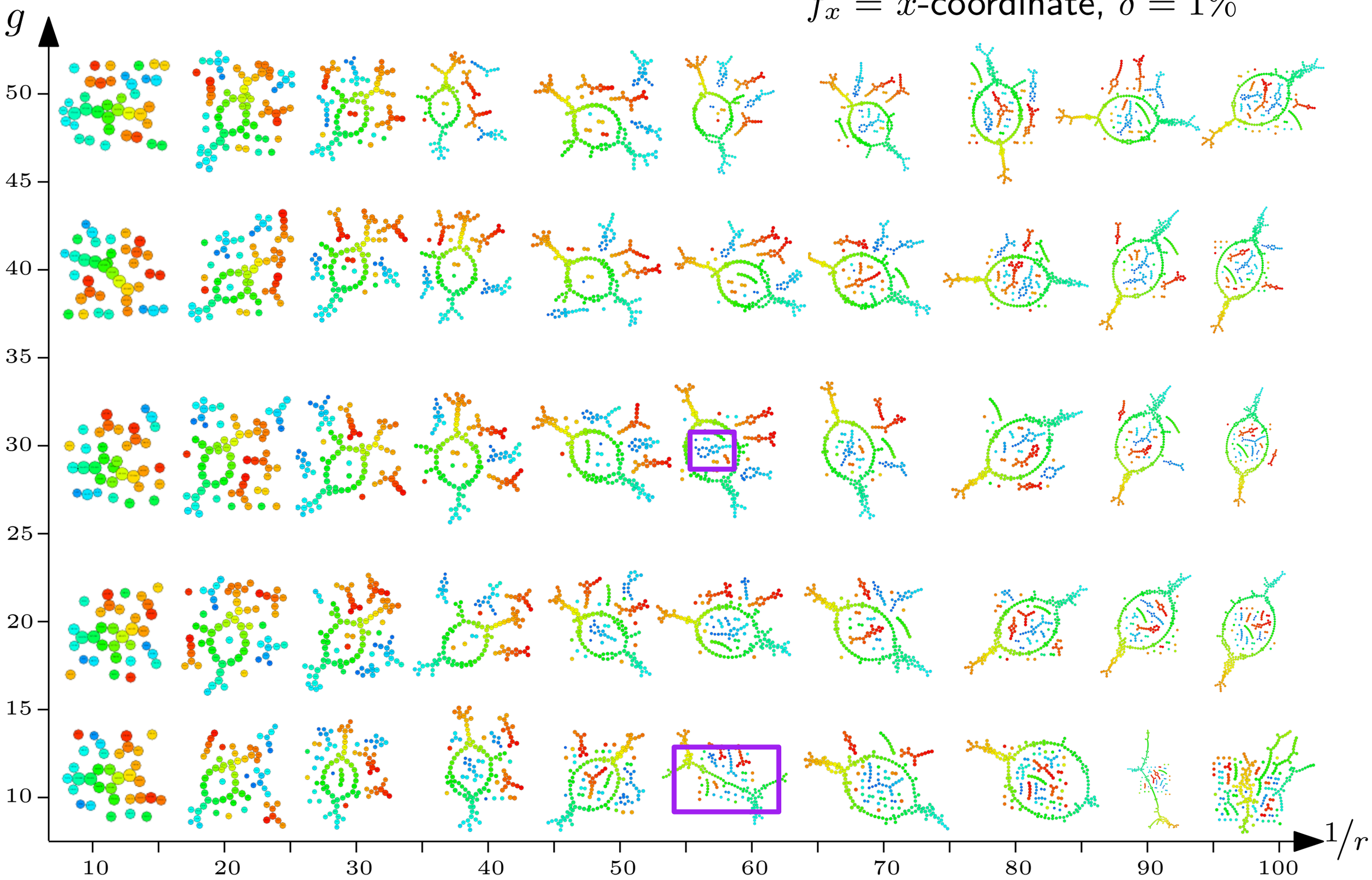
$\delta = 10\%$

Examples


$f_x = x$-coordinate, $\delta = 1\%$
g
50
45
40
35
30
25
20
15
10
$1/r$
10
20
30
40
50
60
70
80
90
100

Examples

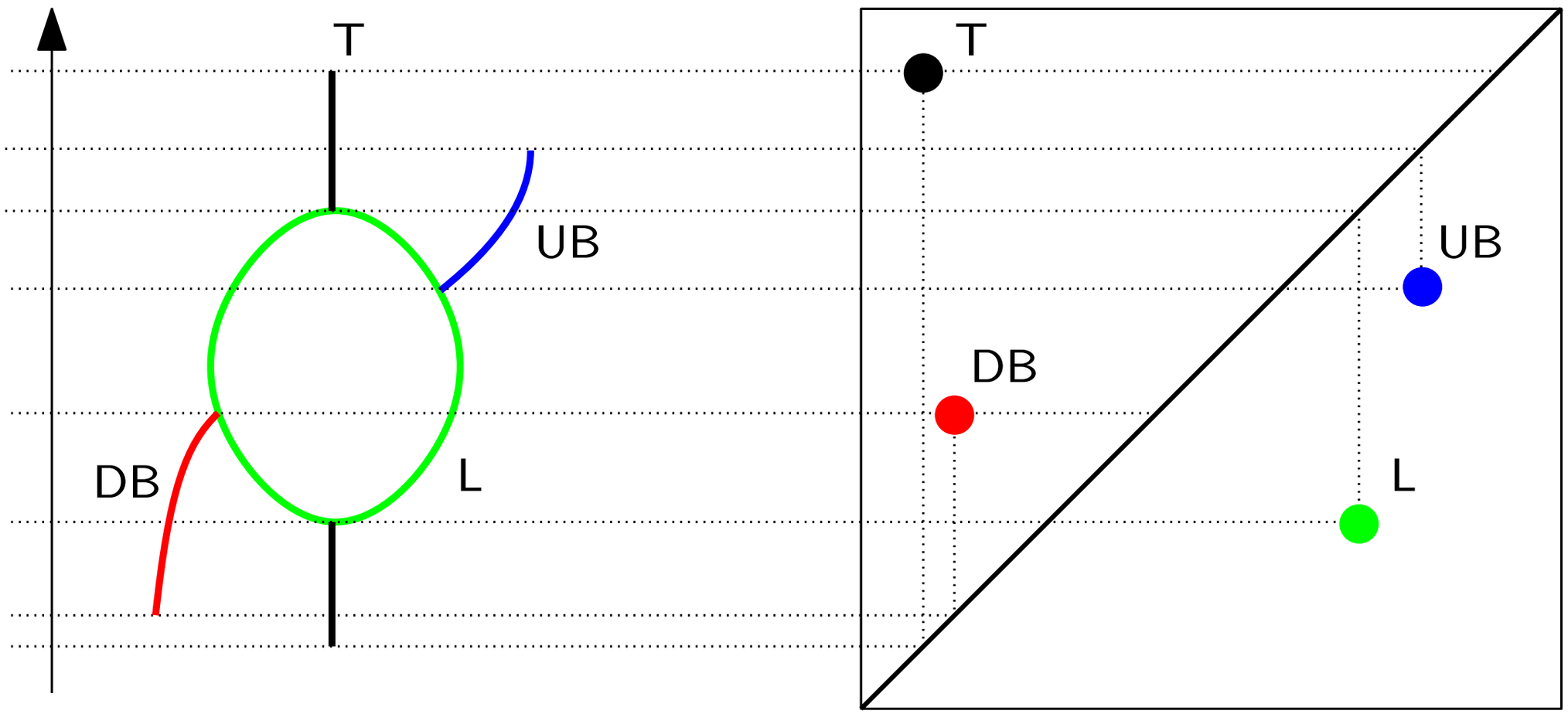
$f_x = x$-coordinate, $\delta = 1\%$



Extended Persistence Diagram

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

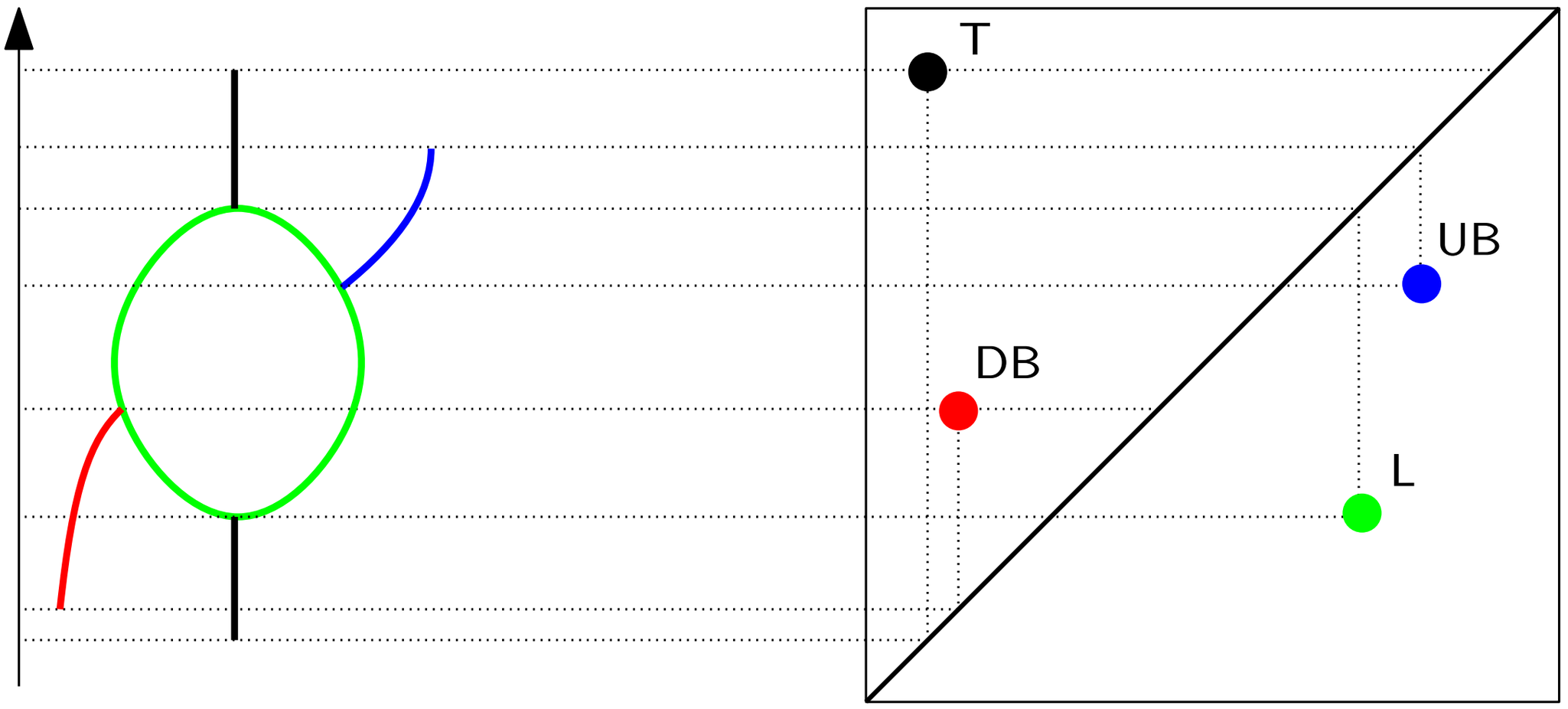
$\mathrm{DB}(\mathrm{R}_f) \longleftrightarrow$ downward branches	$\mathrm{T}(\mathrm{R}_f) \longleftrightarrow$ trunks (cc)
$\mathrm{UB}(\mathrm{R}_f) \longleftrightarrow$ upward branches	$\mathrm{L}(\mathrm{R}_f) \longleftrightarrow$ loops



Extended Persistence Diagram

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

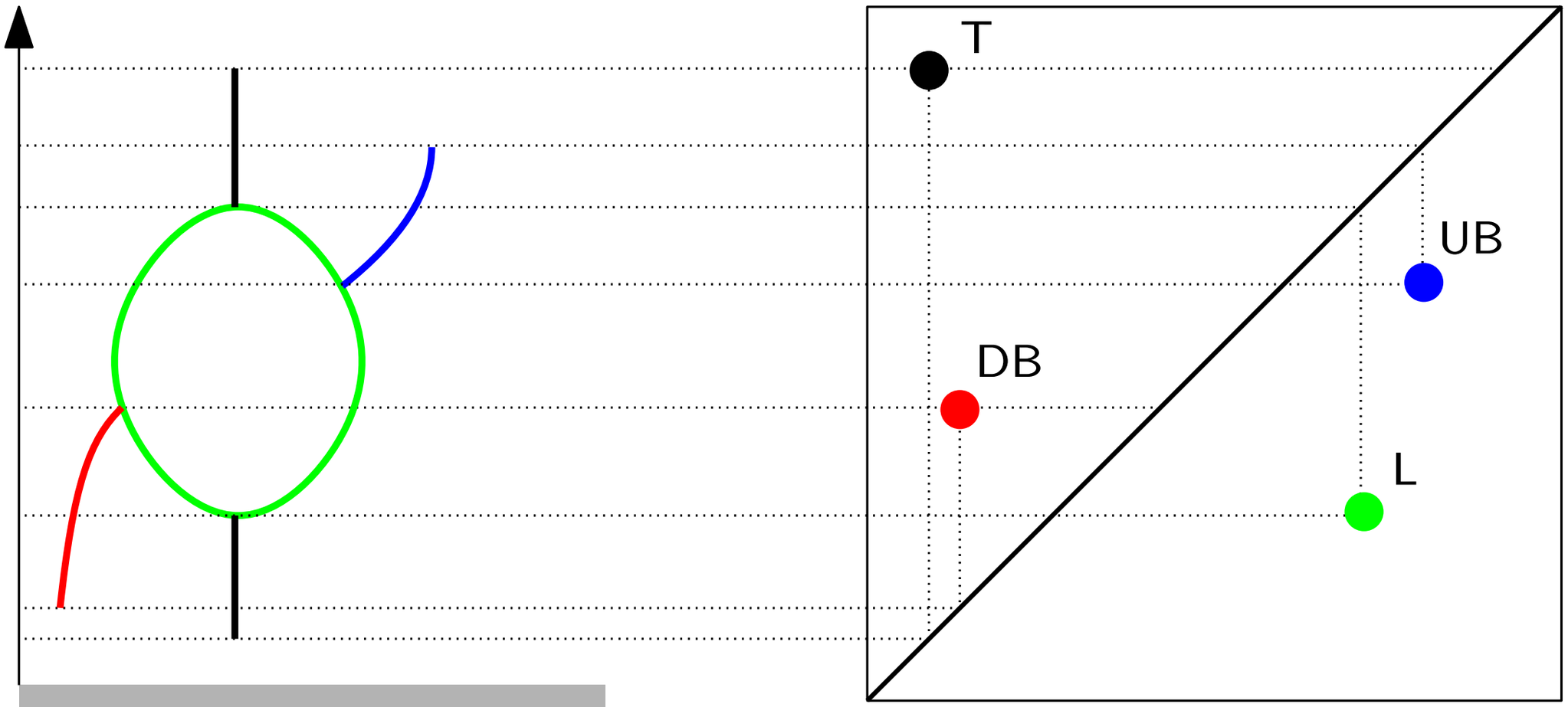
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

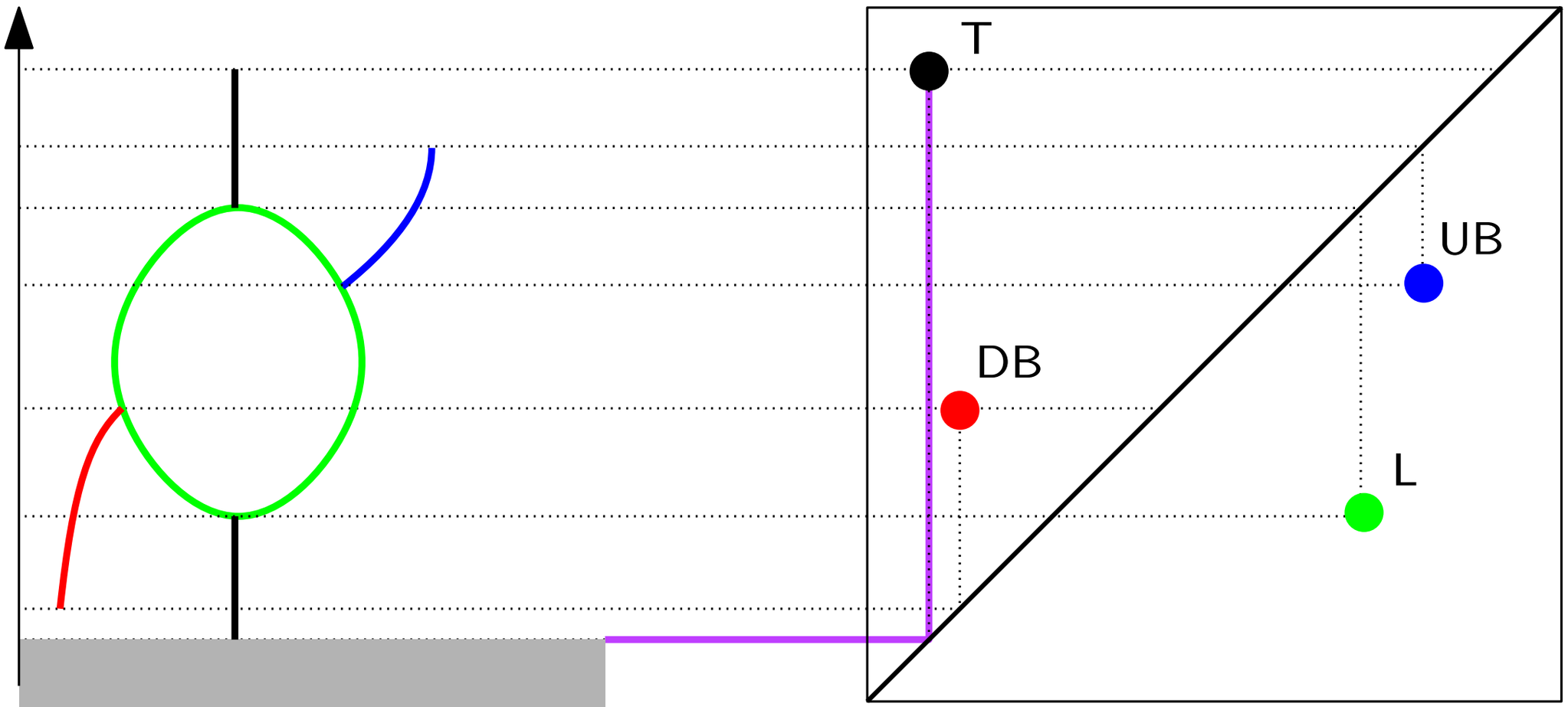
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

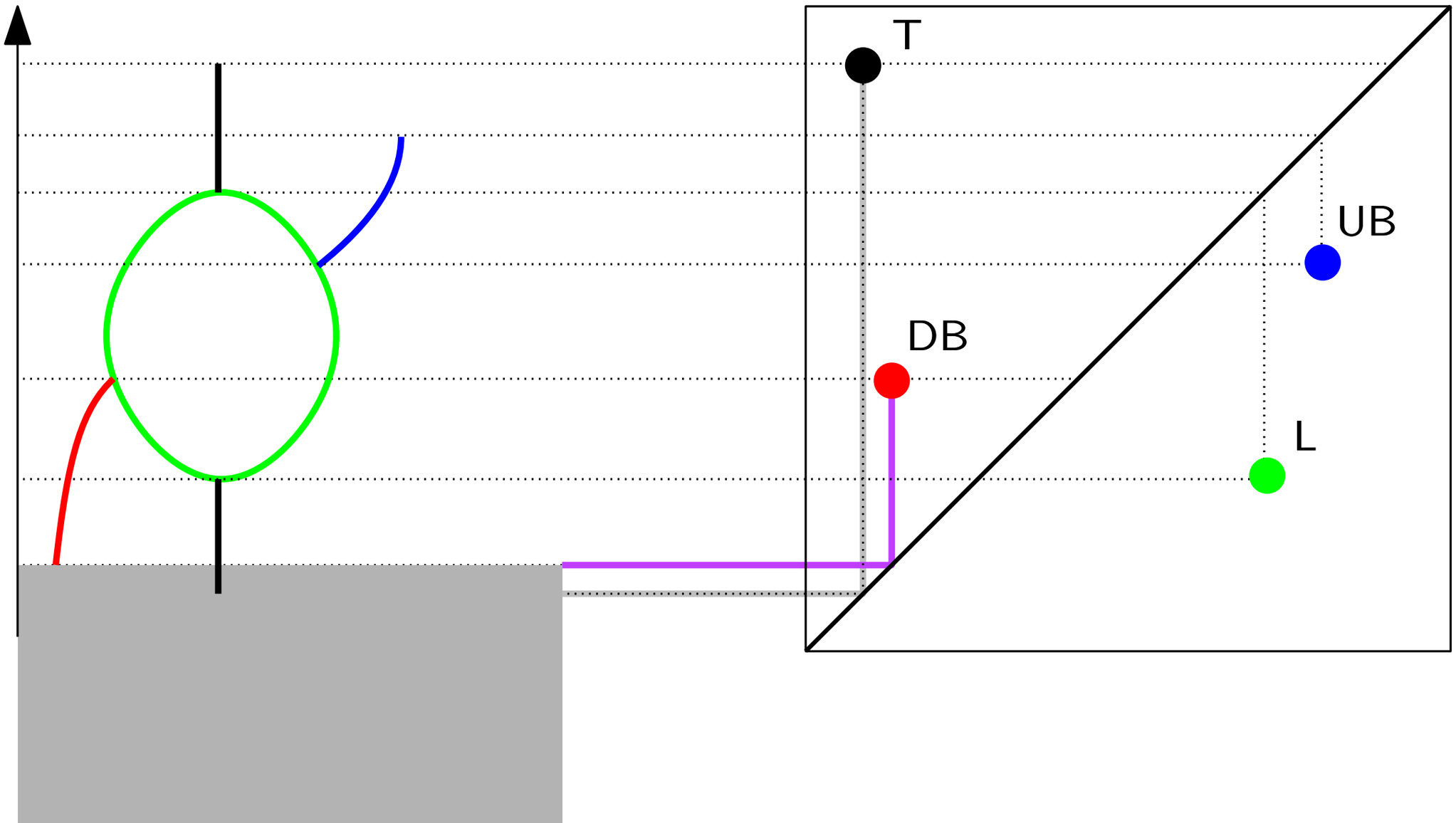
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

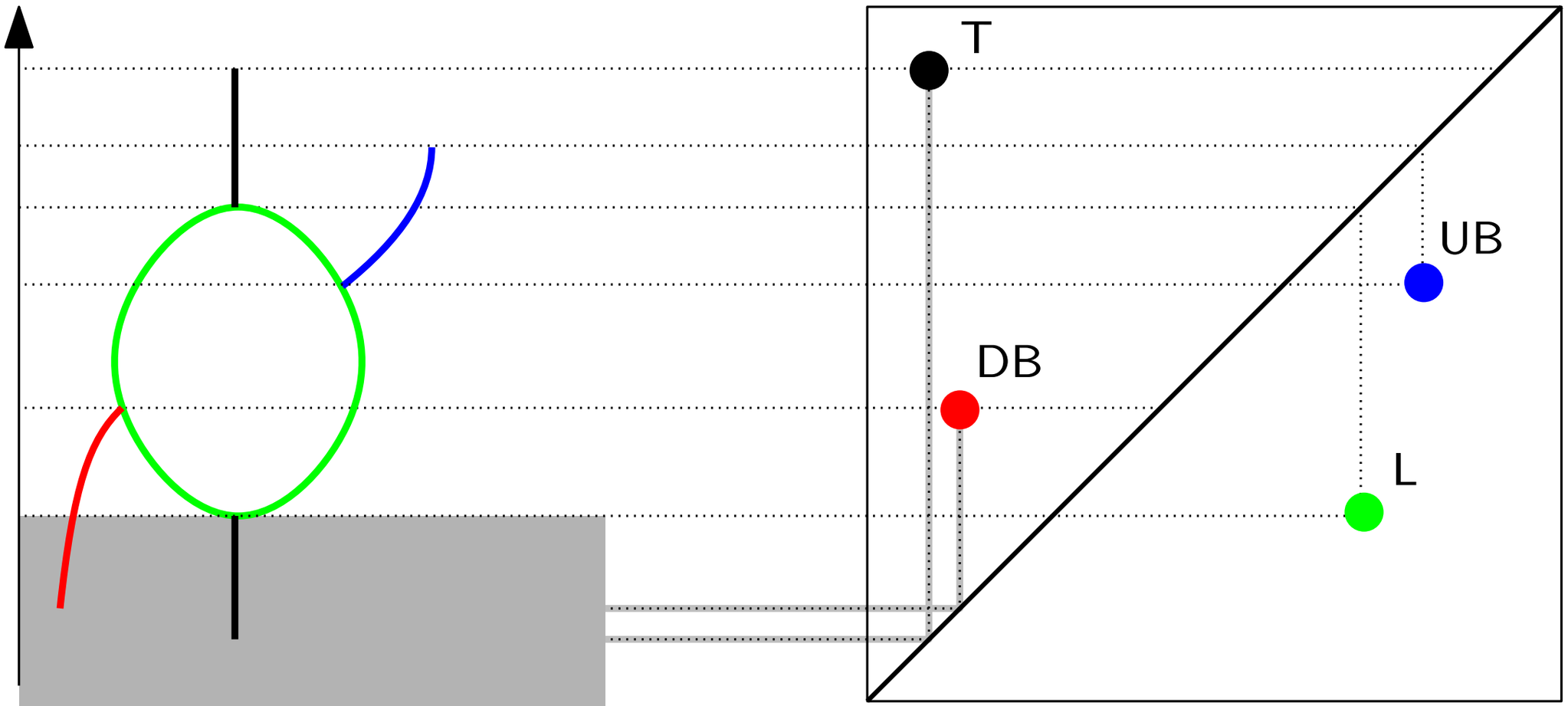
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

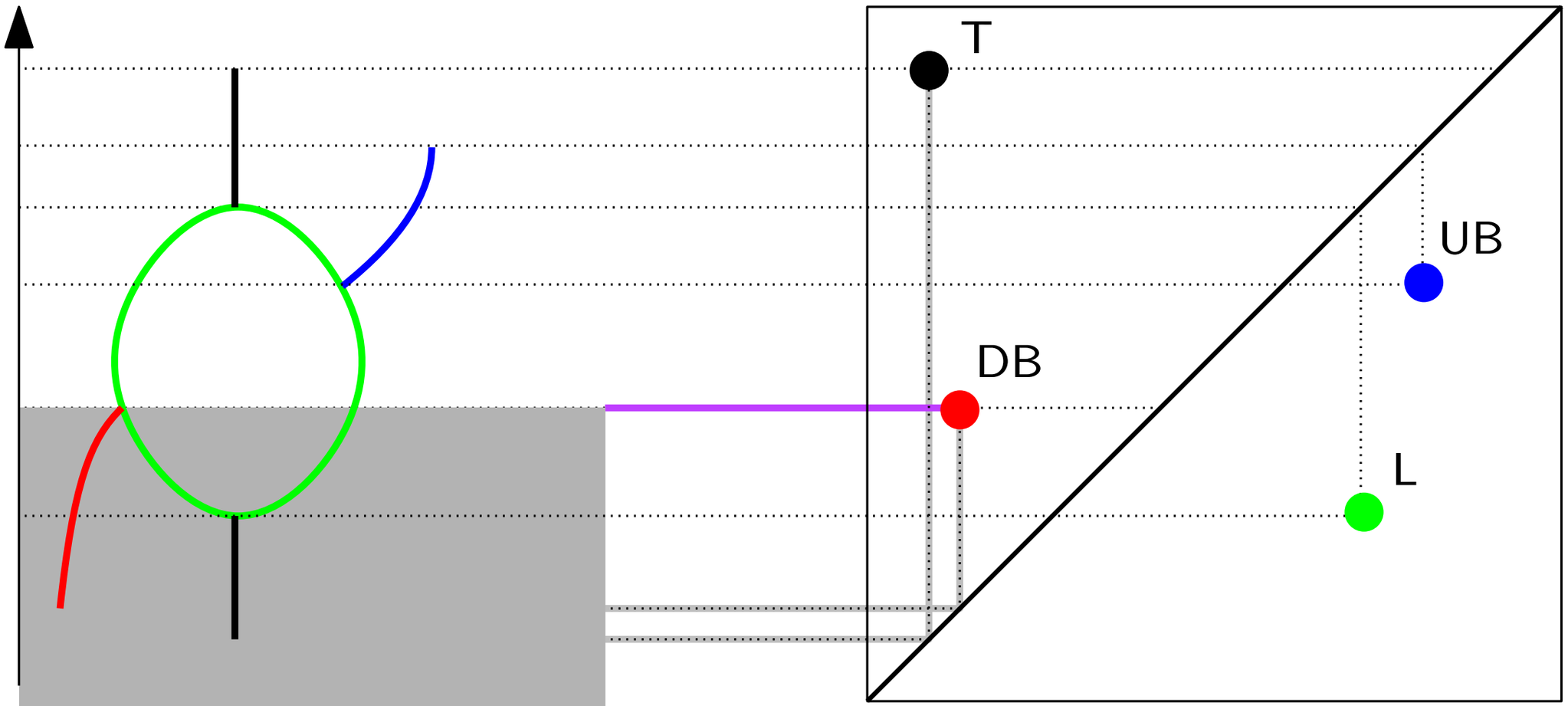
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

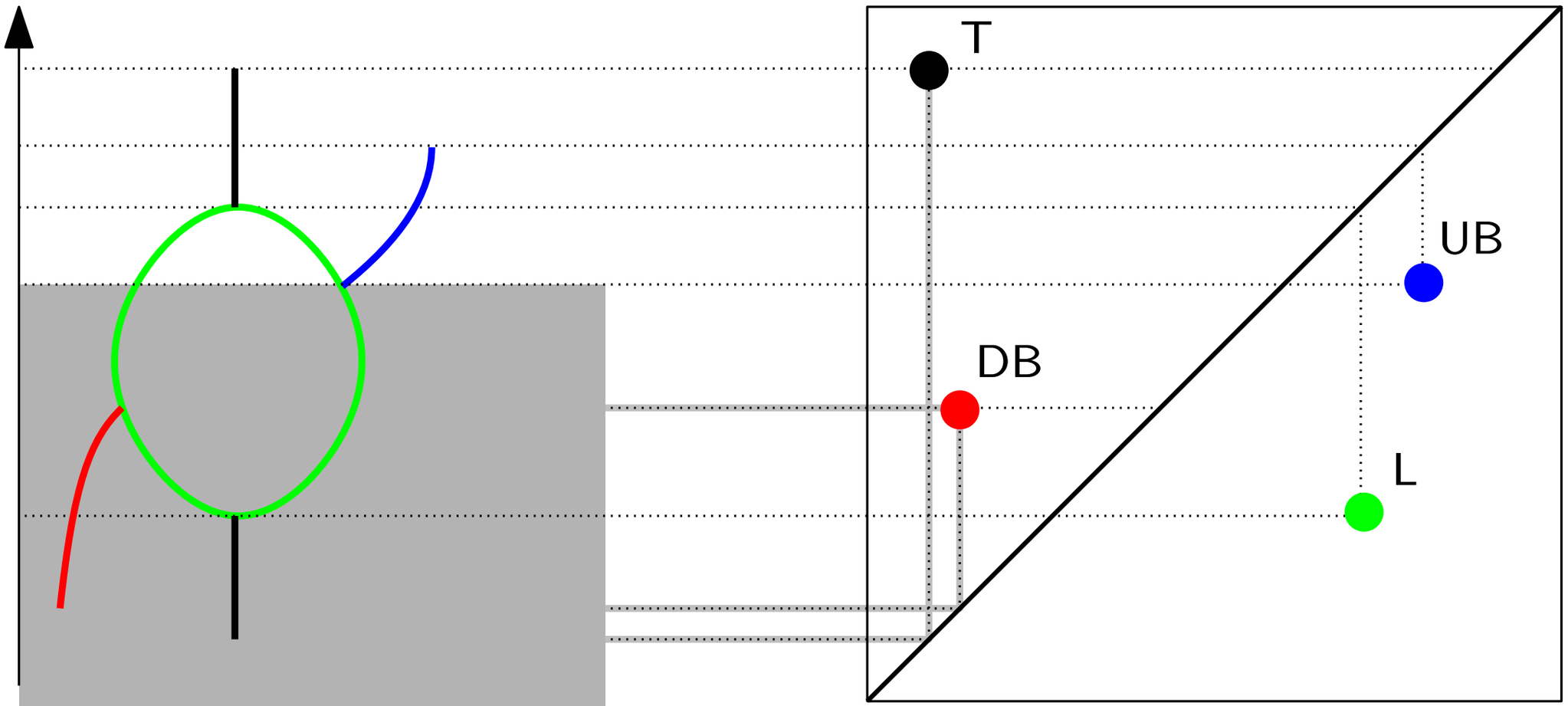
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

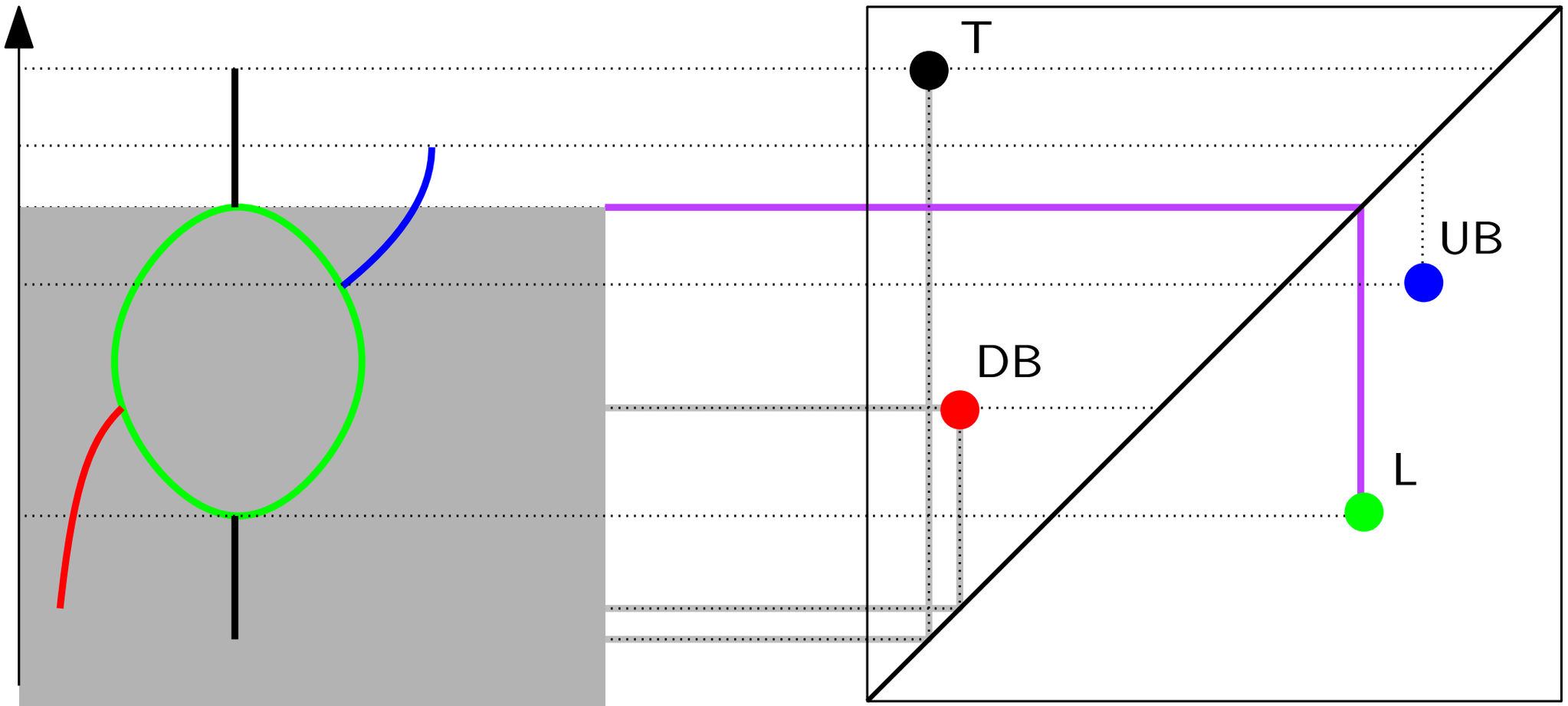
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

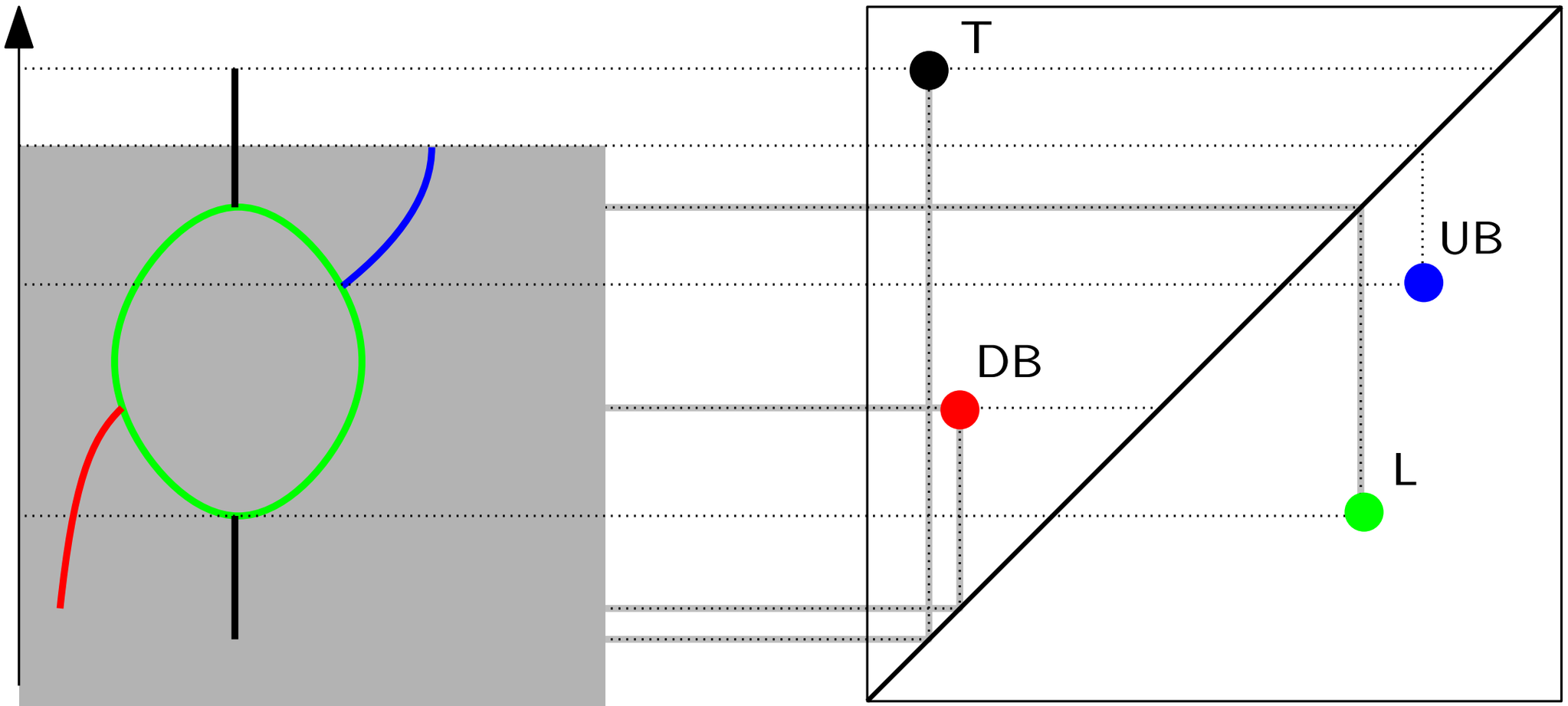
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

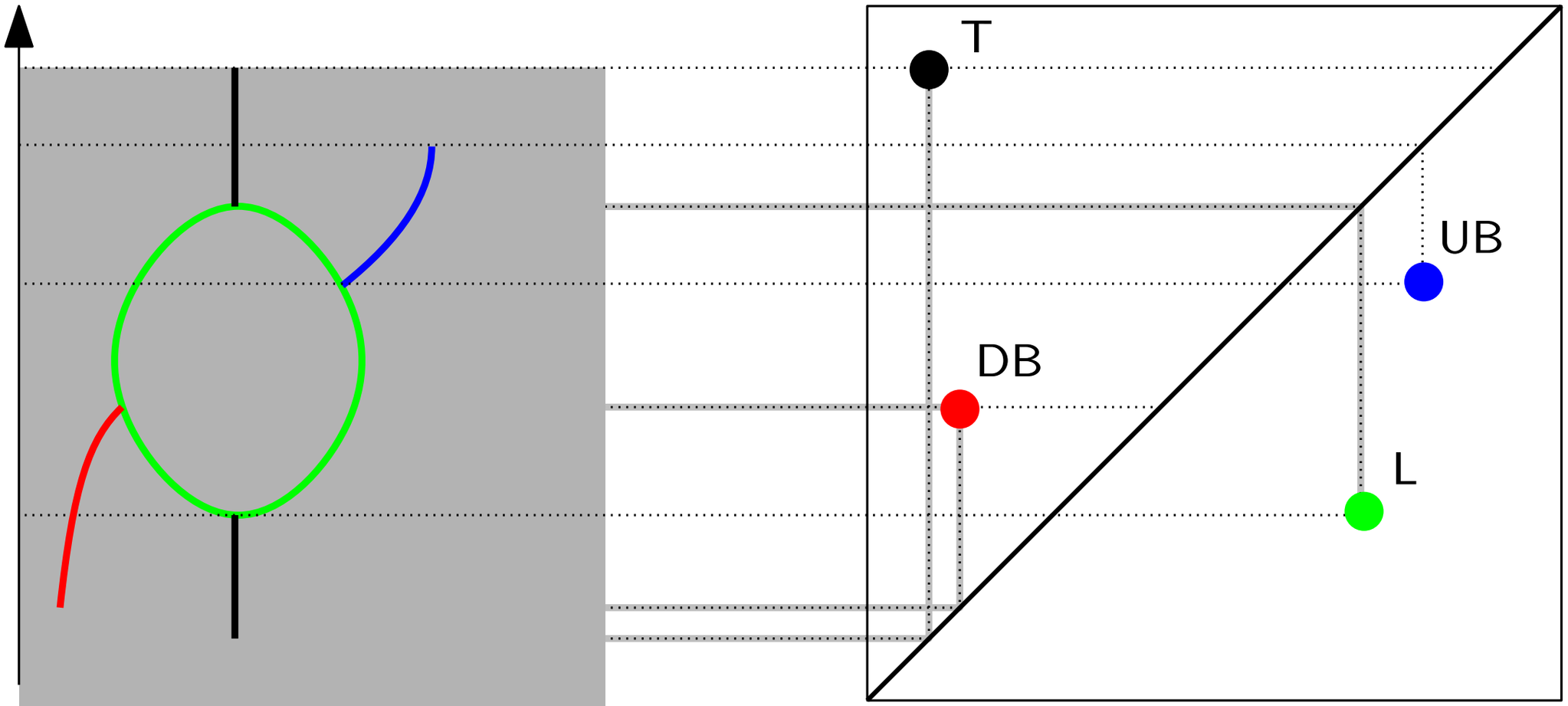
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

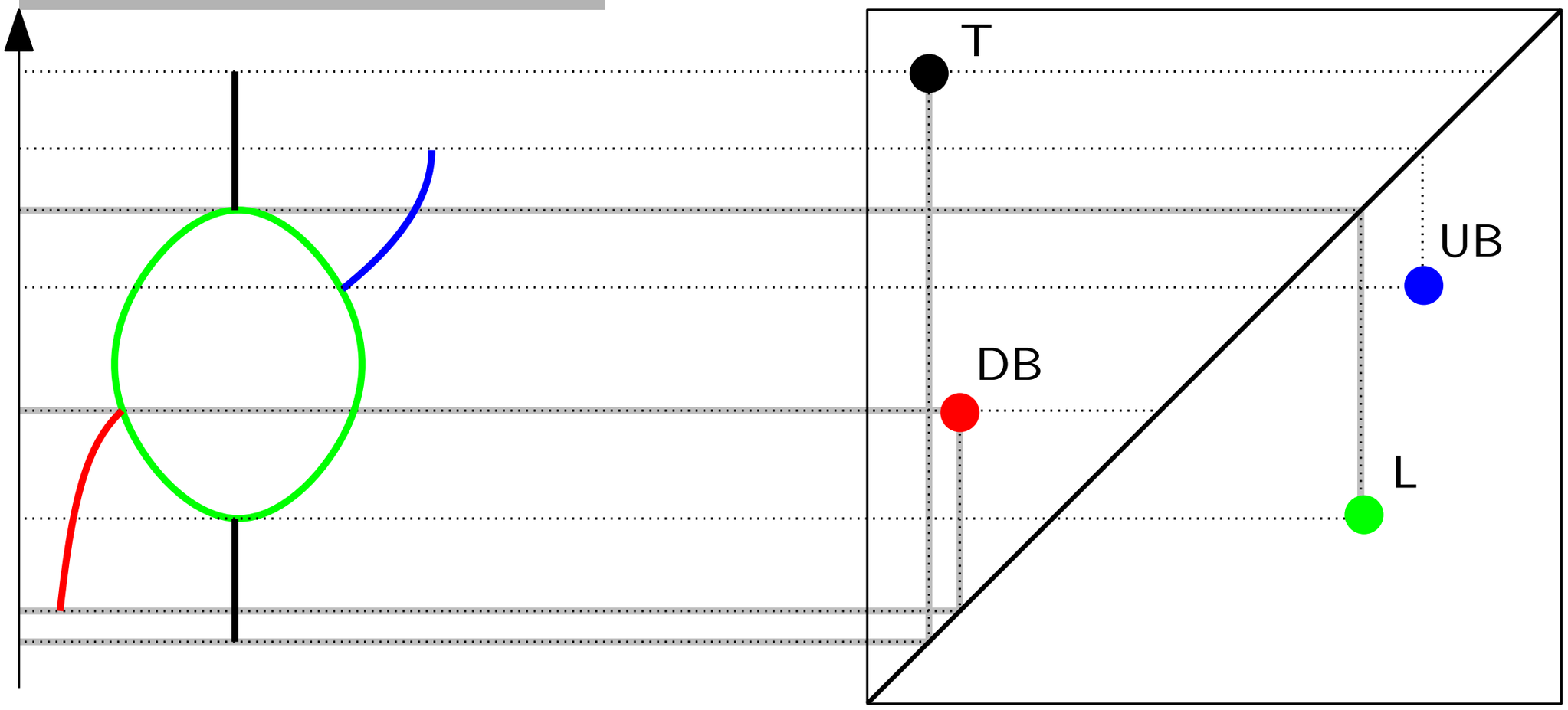
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

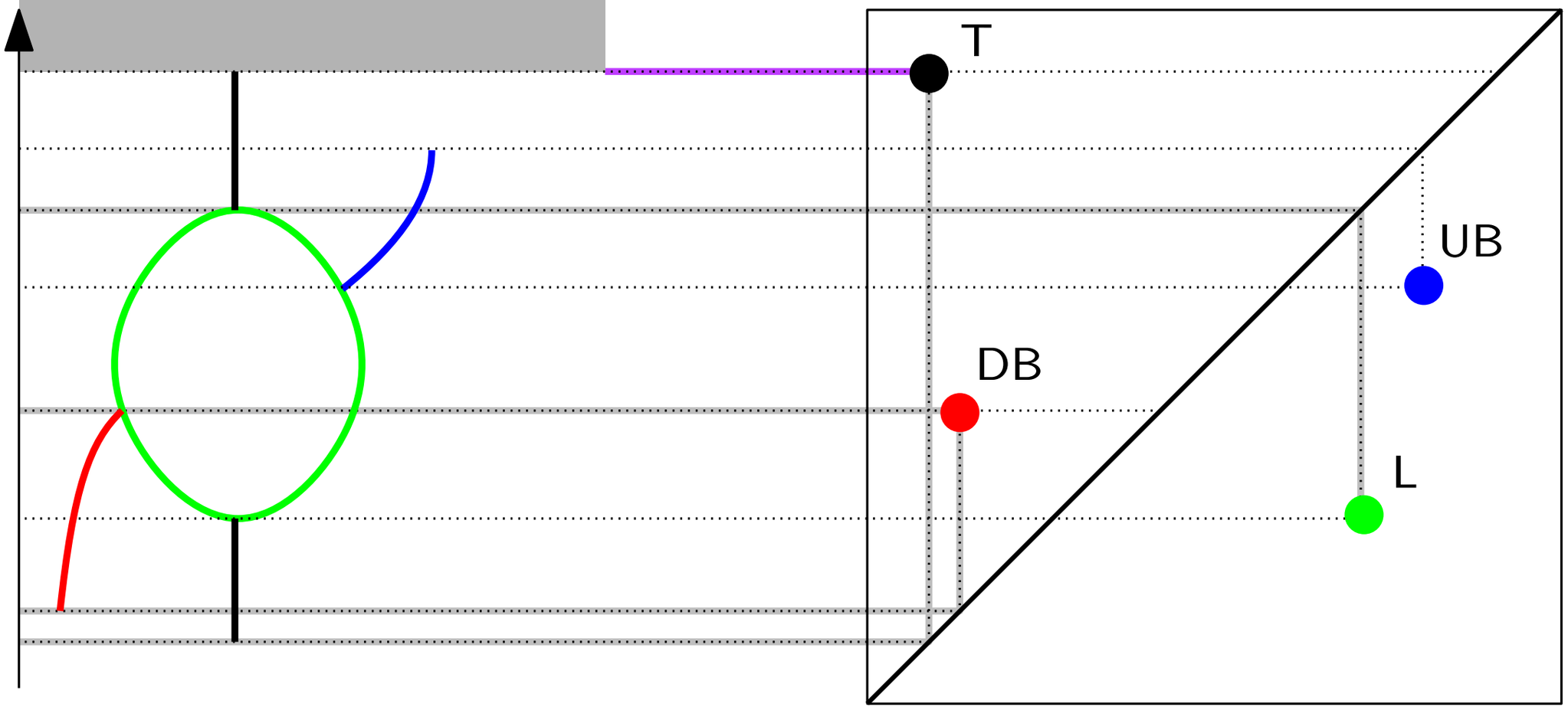
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

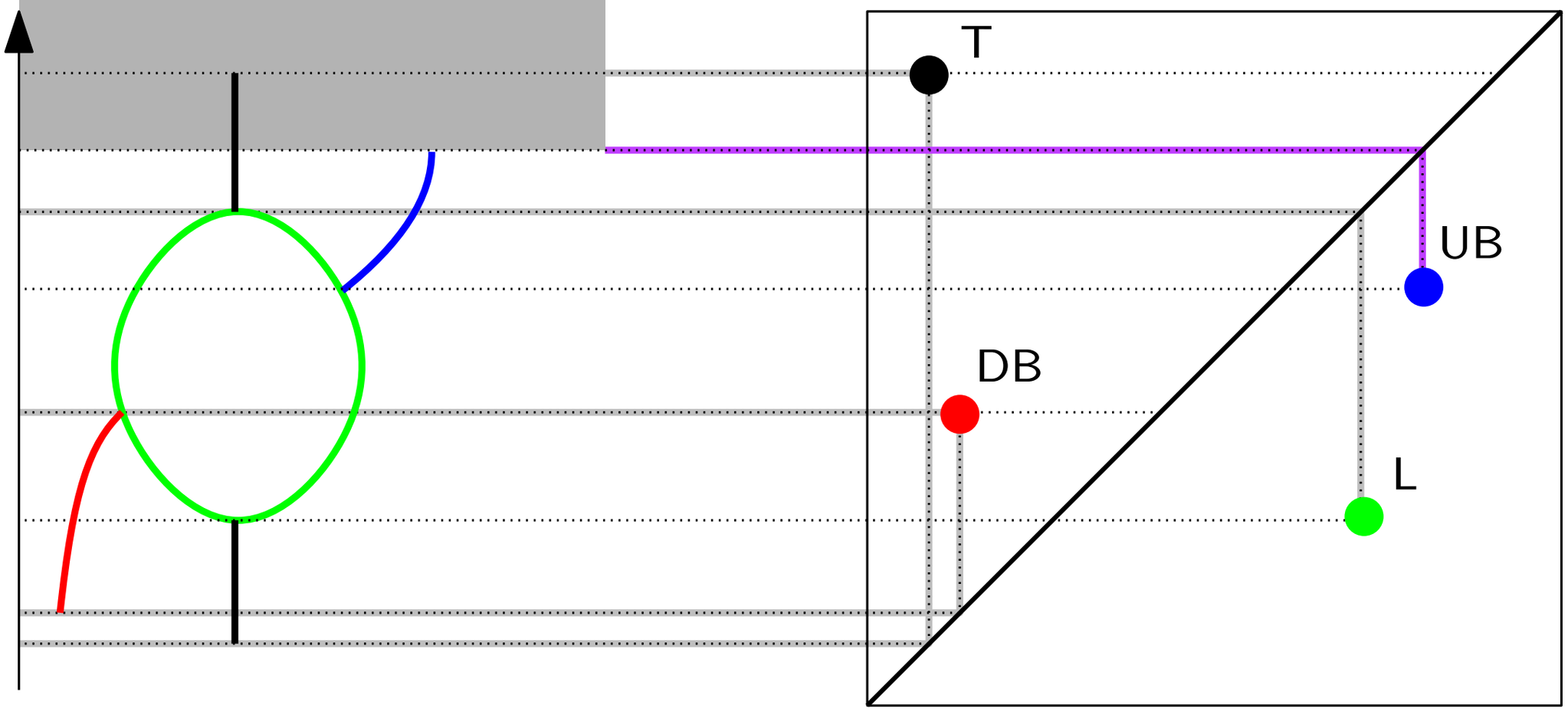
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

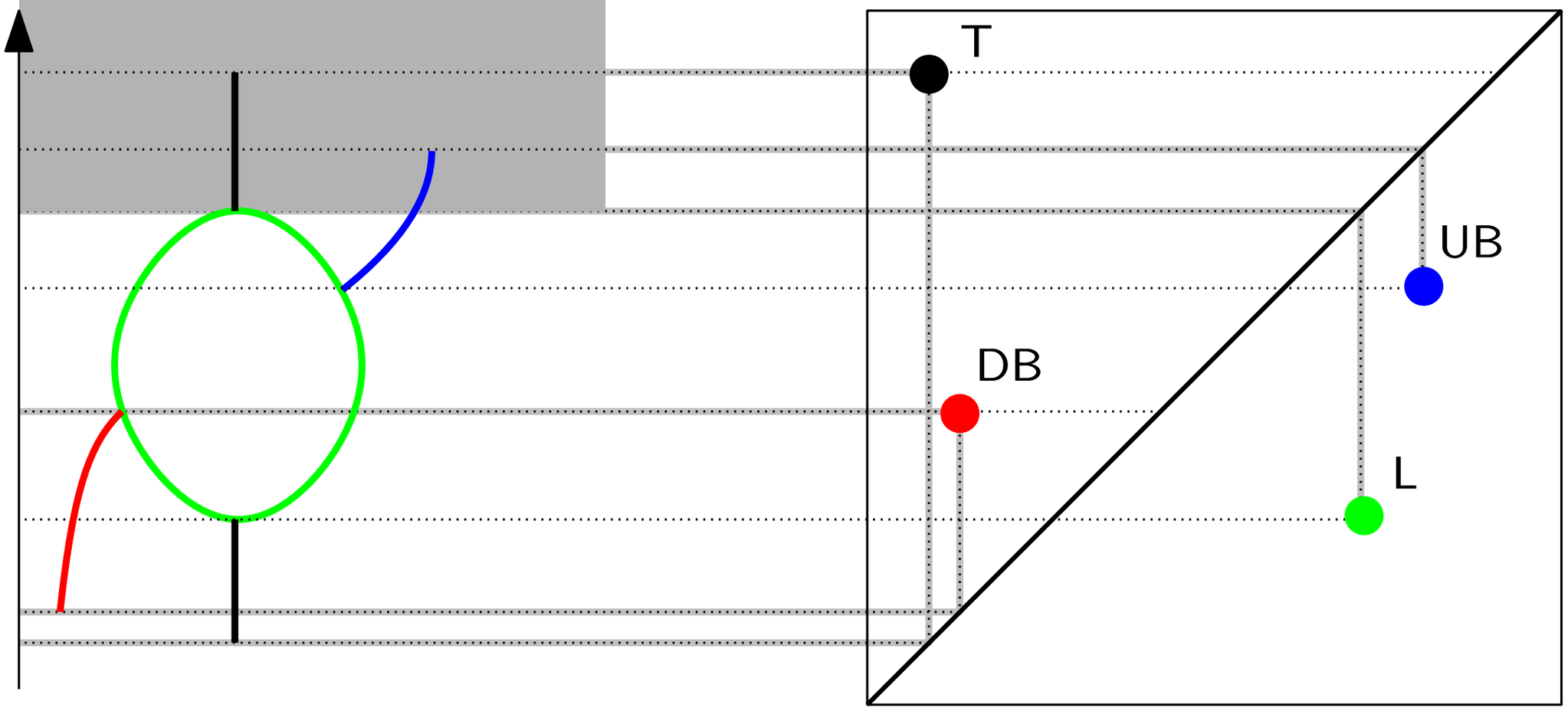
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

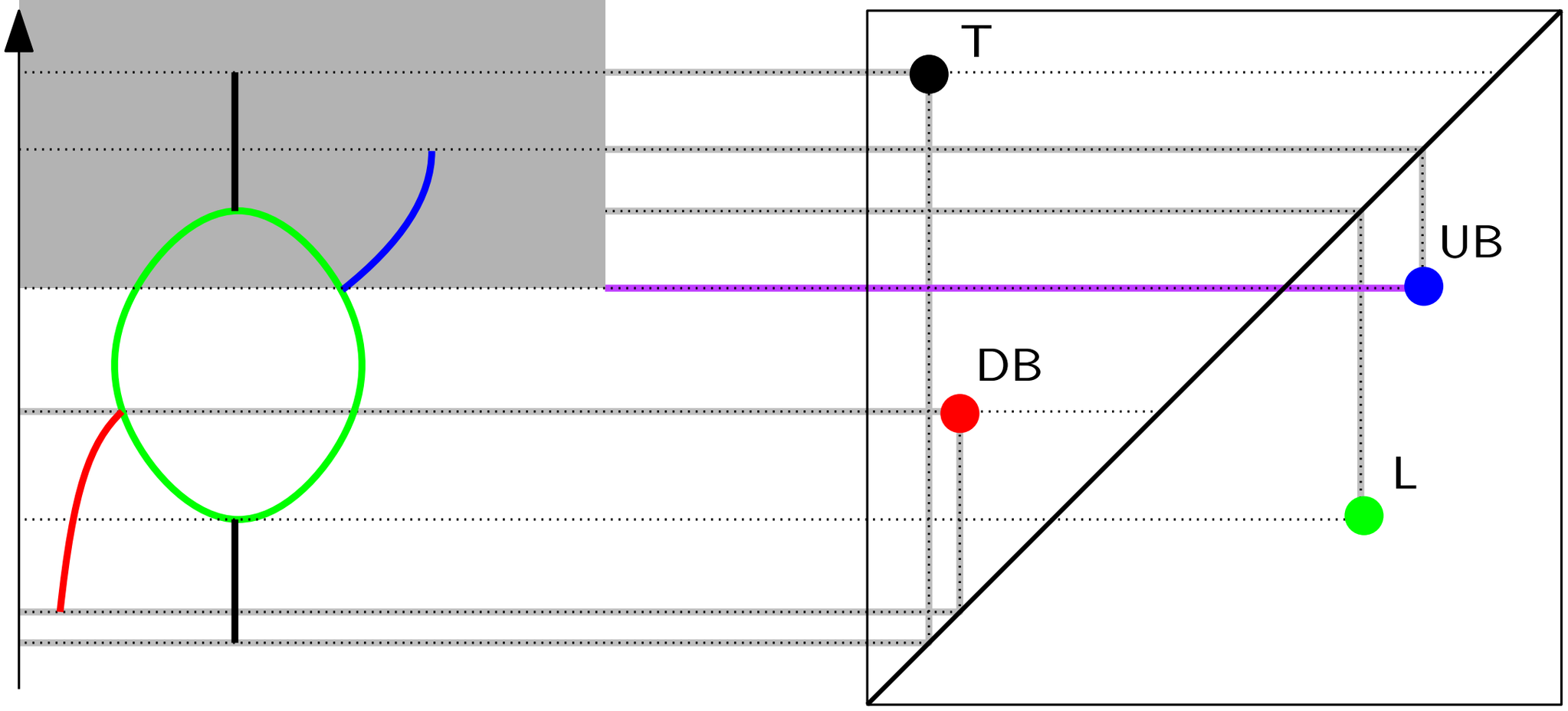
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

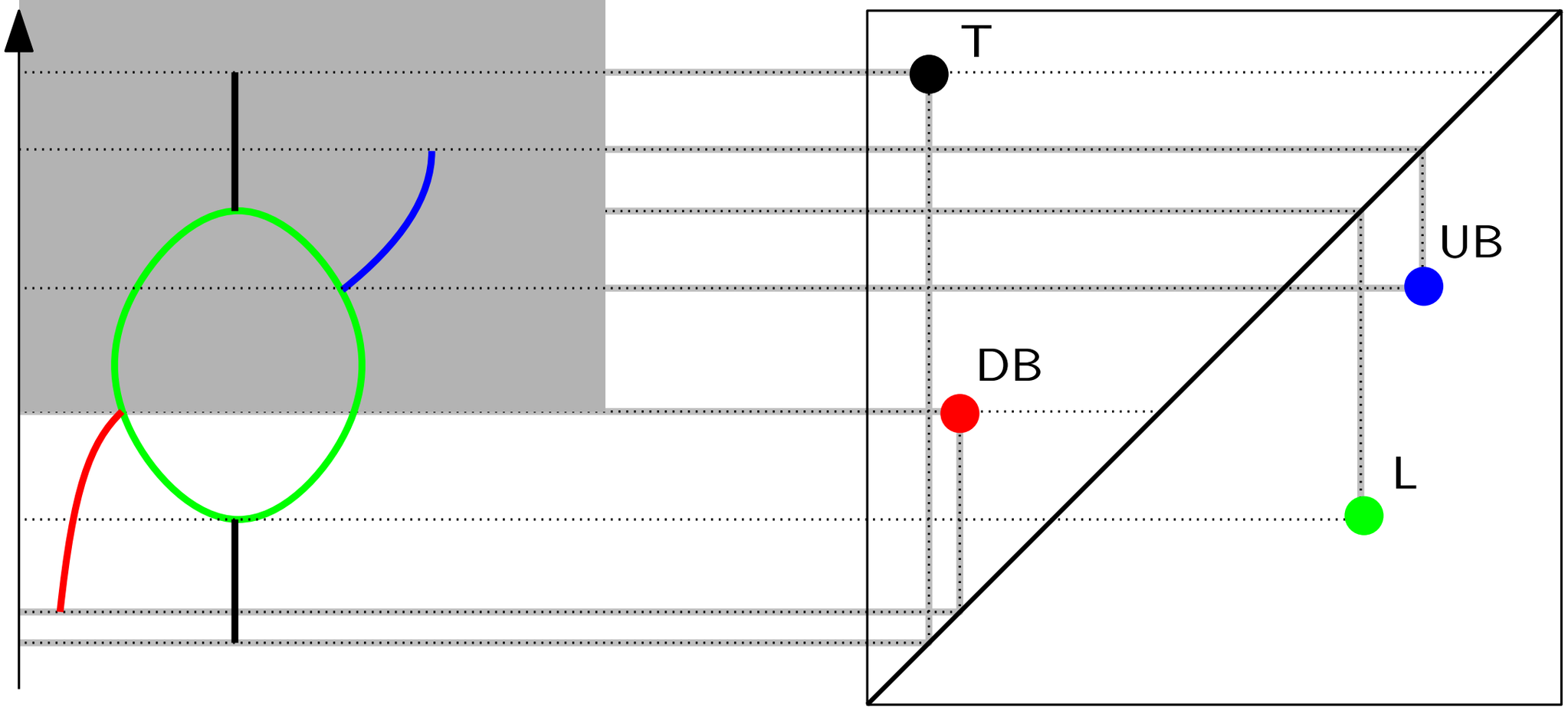
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

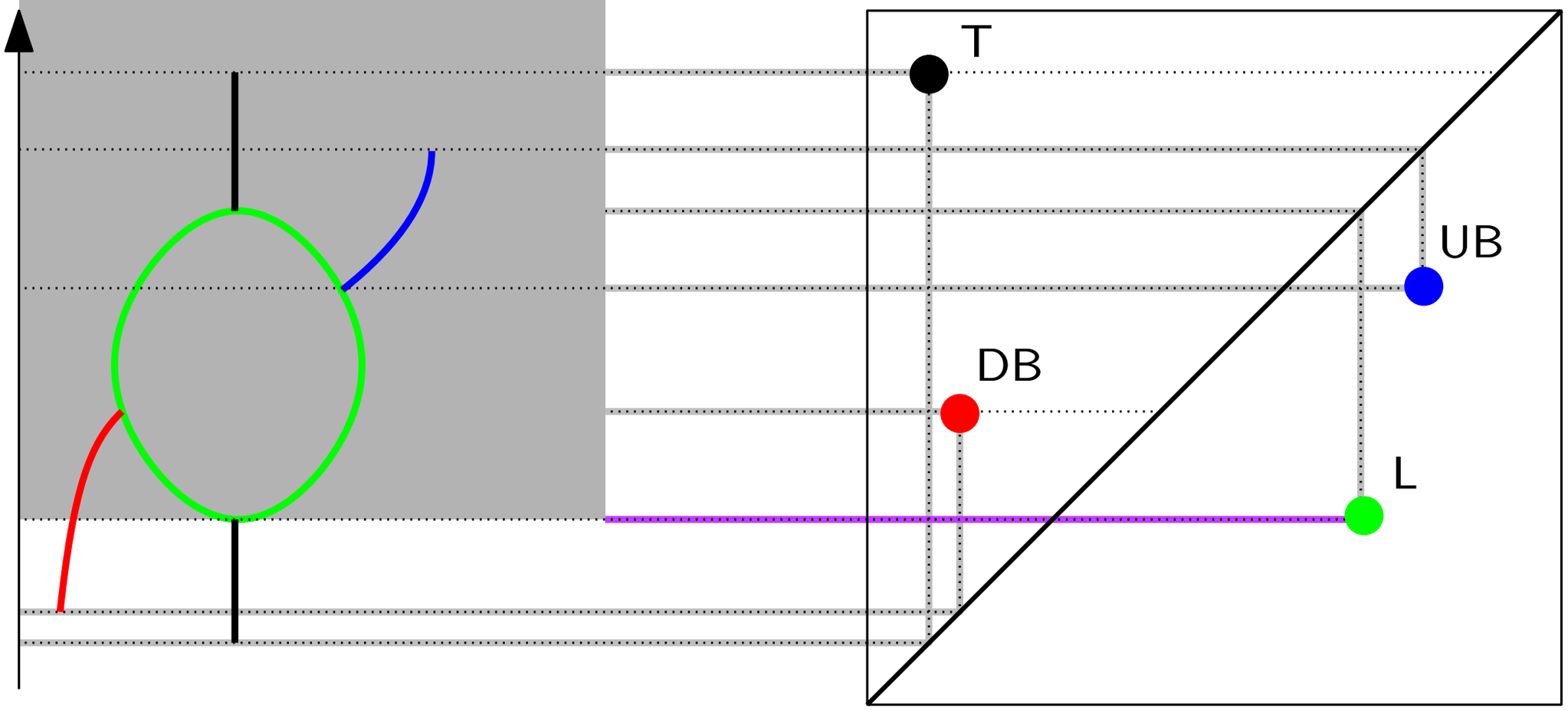
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

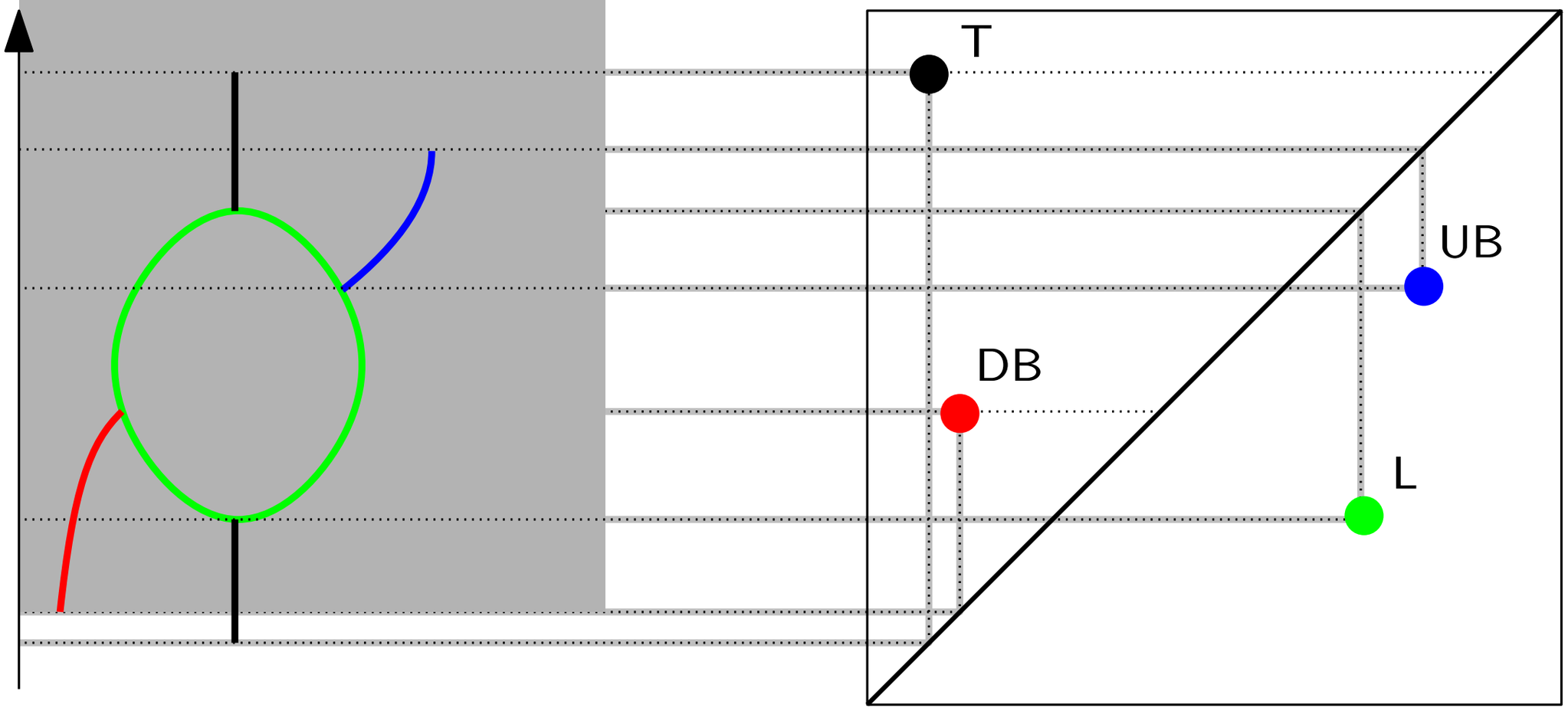
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

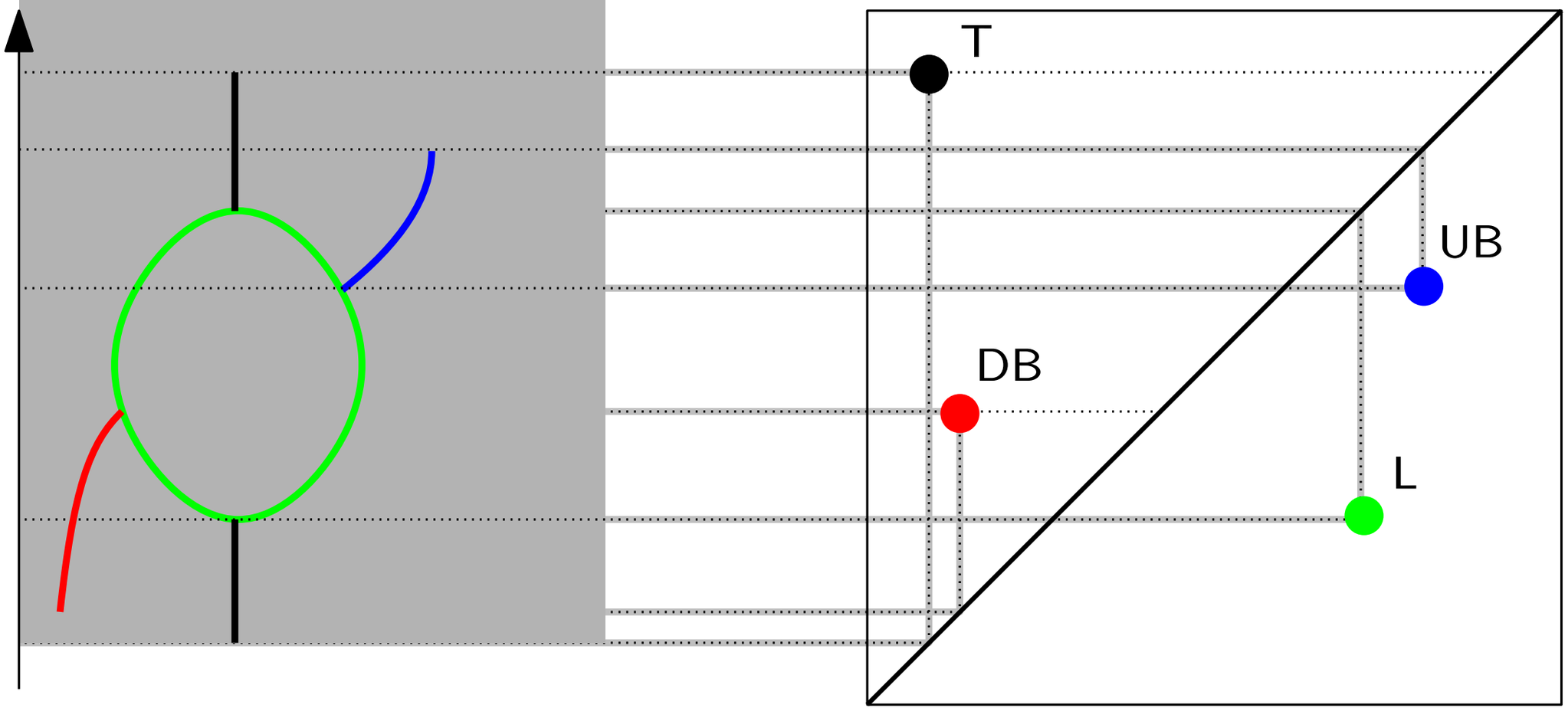
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

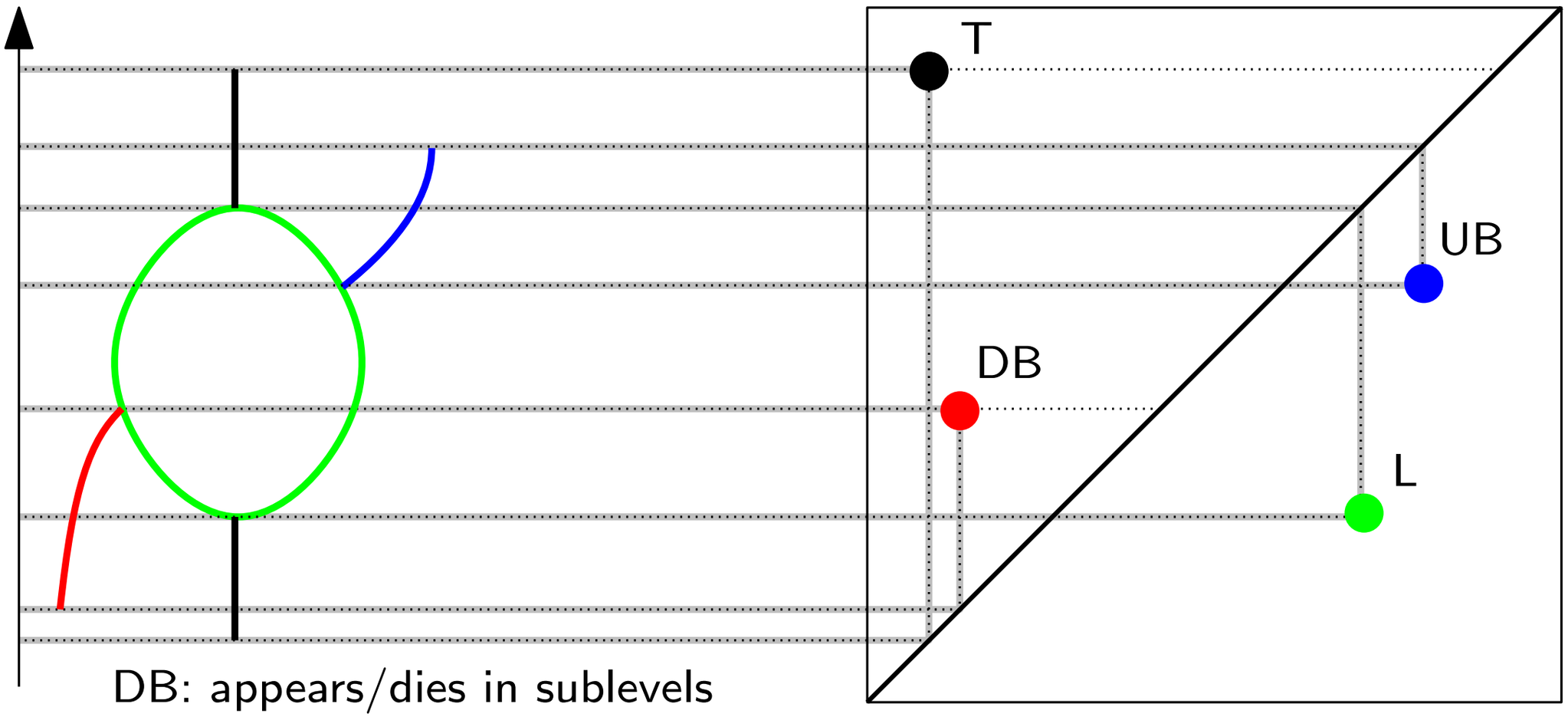
- family of *excursion sets* (sublevel then superlevel sets) of Reeb graph
- use *homological algebra* to encode the evolution of the topology of the family



Extended Persistence Diagram

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

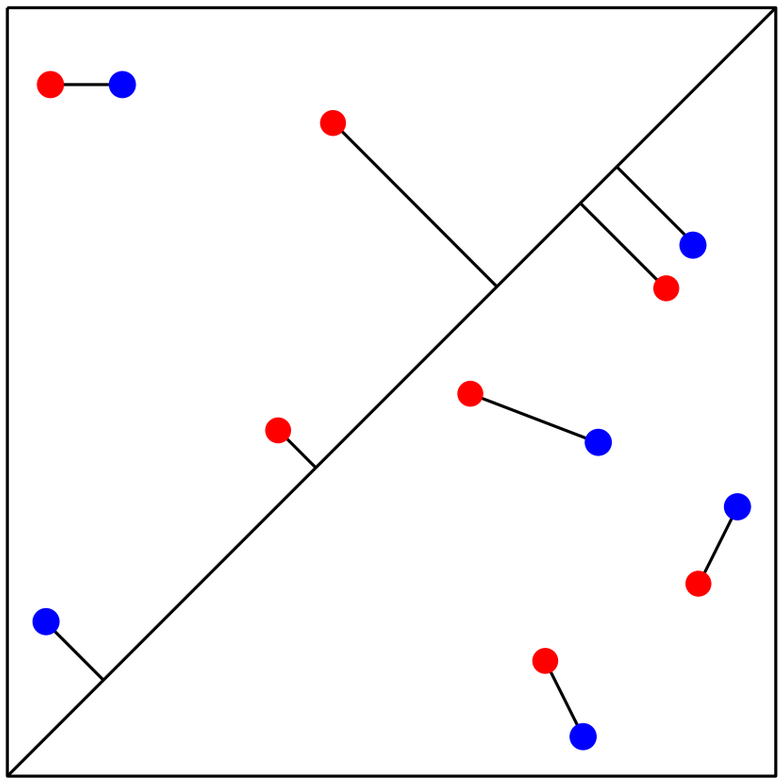


DB: appears/dies in sublevels

UB: appears/dies in superlevels

T/L: appears in sublevels, dies in superlevels

Metric for Extended Persistence Diagrams



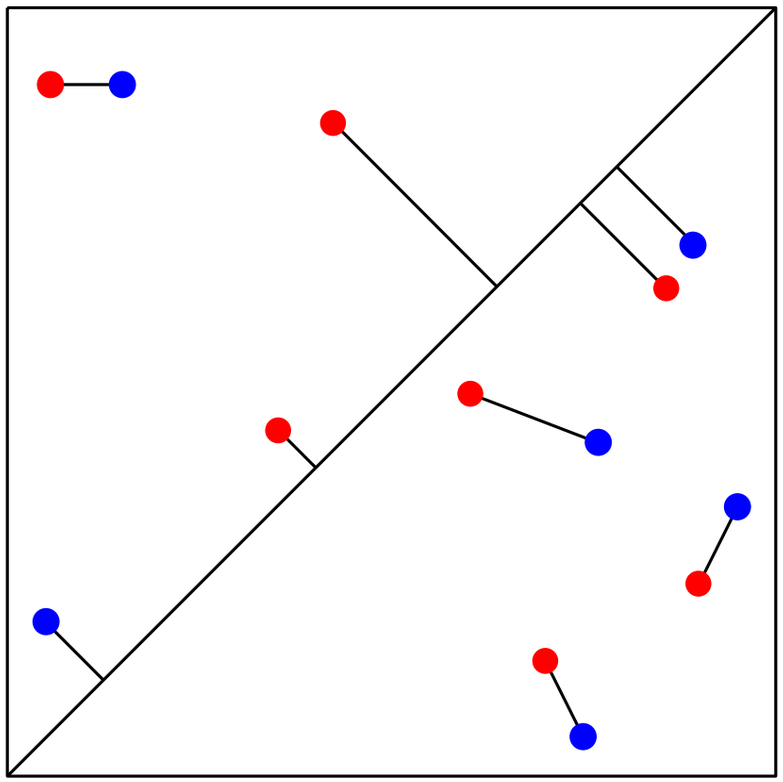
Metric for Extended Persistence Diagrams

Partial matching $M : \mathrm{Dg} \leftrightarrow \mathrm{Dg}'$

Matched pair $(x, y) \in M$: $c(x, y) = \|x - y\|_\infty$

Unmatched point $z \in X \sqcup Y$: $c(z) = \|z - \bar{z}\|_\infty$

$c(M) = \max\{\ \max_{(x,\,y)} c(x, y),\ \max_z c(z)\ \}$



Metric for Extended Persistence Diagrams

Partial matching $M : \mathrm{Dg} \leftrightarrow \mathrm{Dg}'$

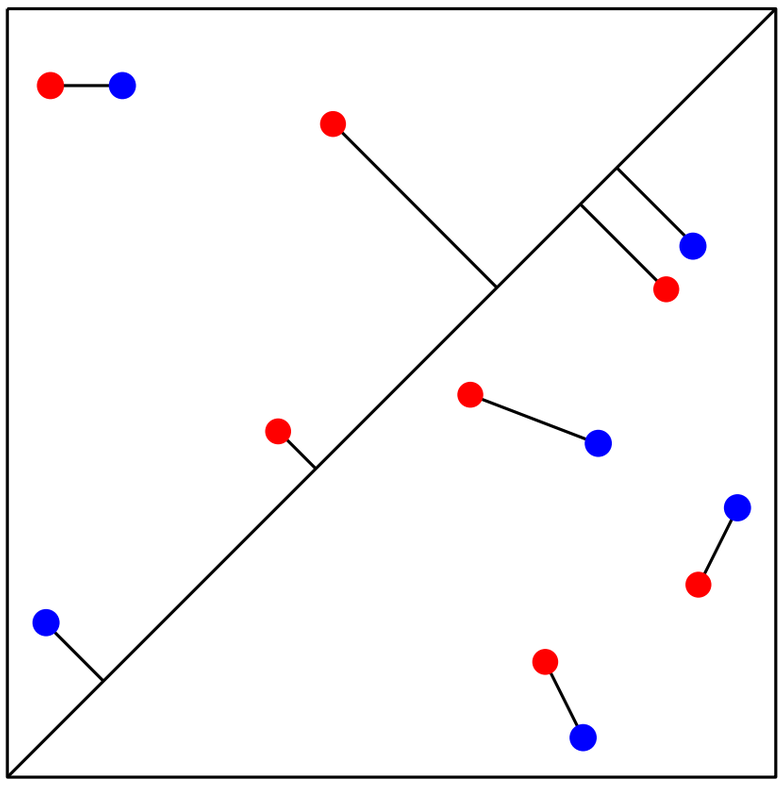
Matched pair $(x, y) \in M$: $c(x, y) = \|x - y\|_\infty$

Unmatched point $z \in X \sqcup Y$: $c(z) = \|z - \bar{z}\|_\infty$

$c(M) = \max\{ \max_{(x,\, y)} c(x, y),\ \max_z c(z) \}$

Def: bottleneck distance:

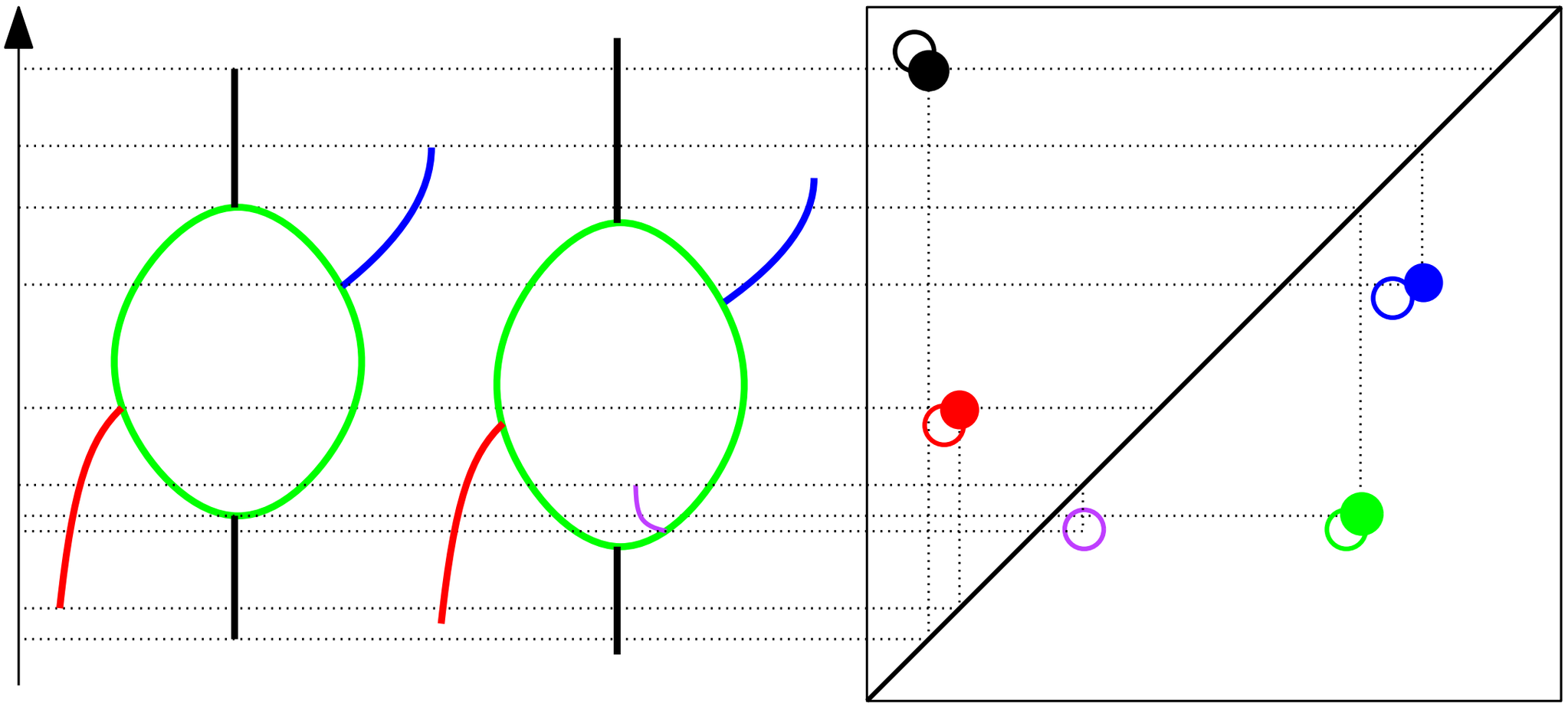
$$\mathrm{d}_B(\mathrm{Dg}\,, \mathrm{Dg}') = \inf_{M:\mathrm{Dg} \leftrightarrow \mathrm{Dg}'} c(M)$$



Metric Properties

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

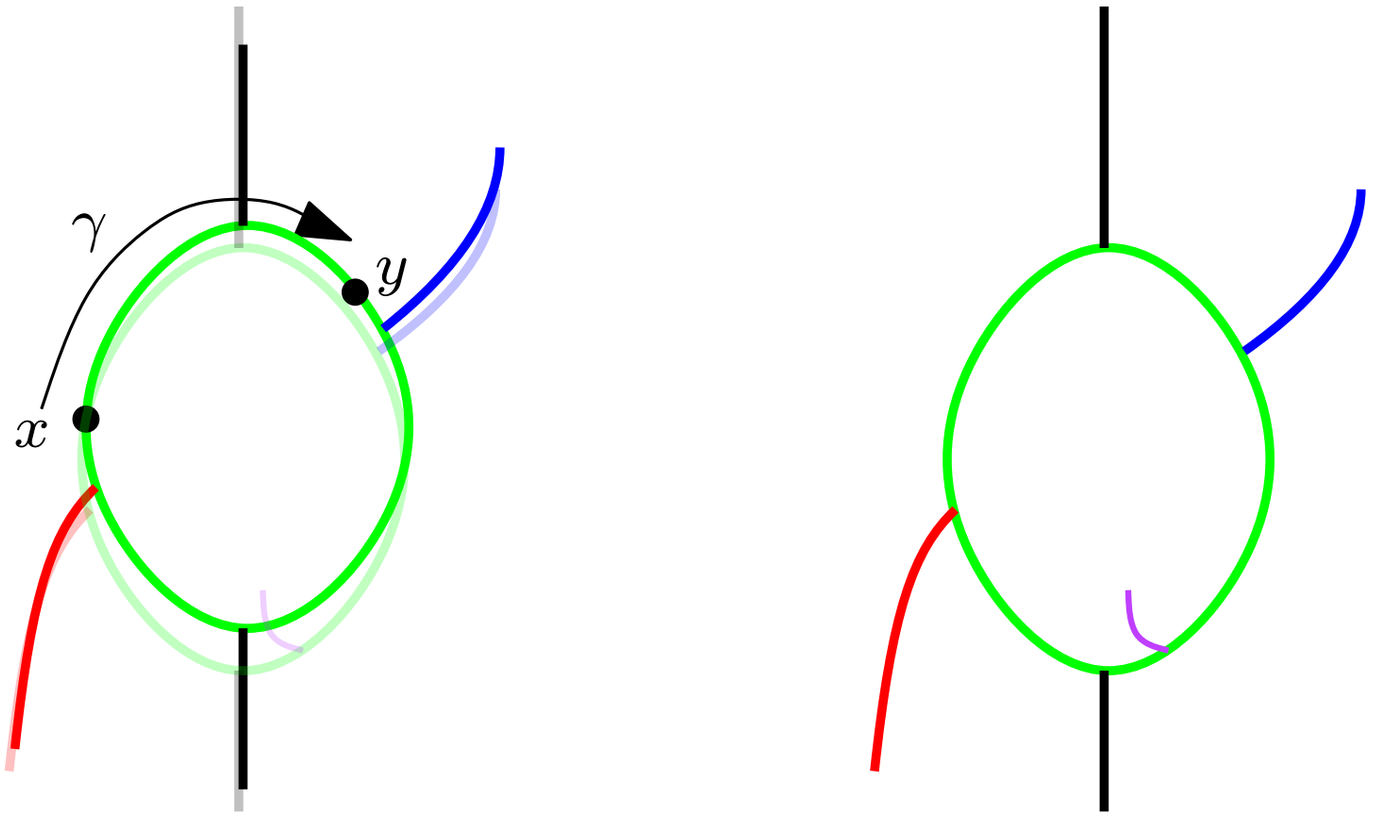
$$\mathrm{d}_B(\mathrm{Dg}\,(\mathrm{R}_f), \mathrm{Dg}\,(\mathrm{R}_g)) \leq 6\,\mathrm{d}_{\mathrm{GH}}(\mathrm{R}_f, \mathrm{R}_g)$$



Metric Properties

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

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

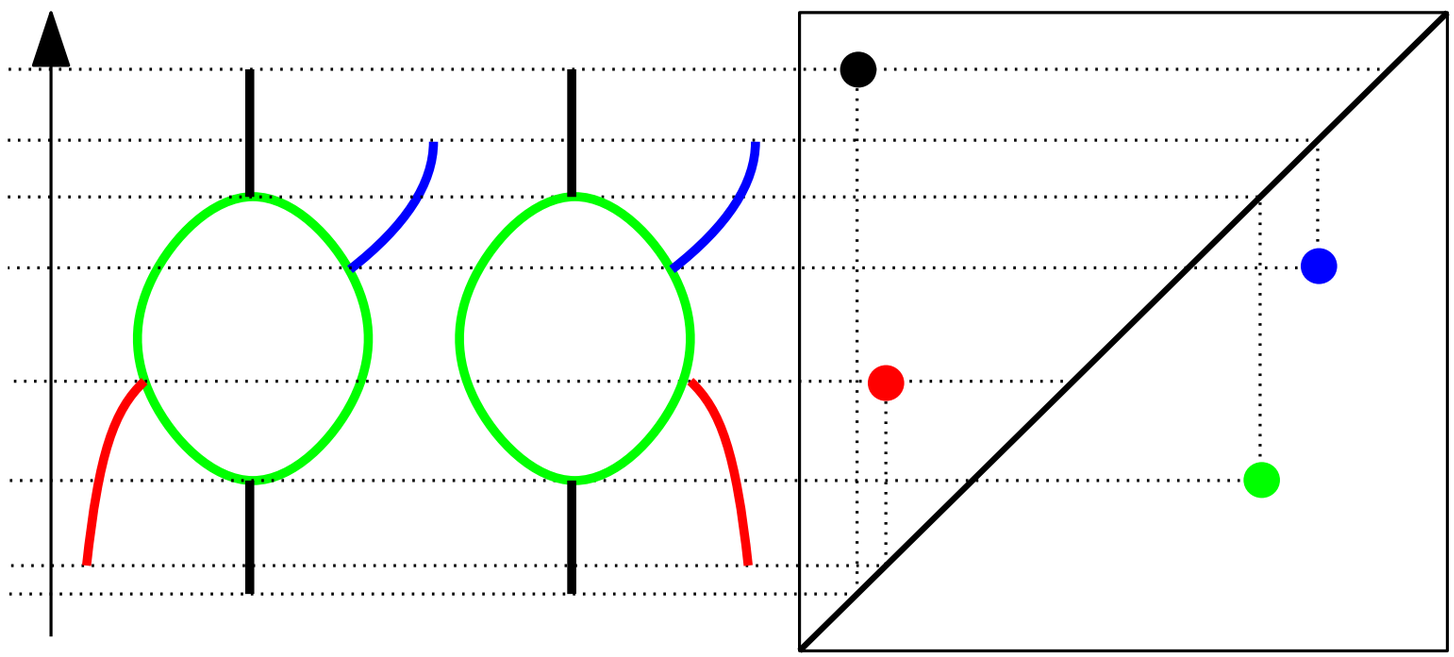


Metric Properties

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

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

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



Metric Properties

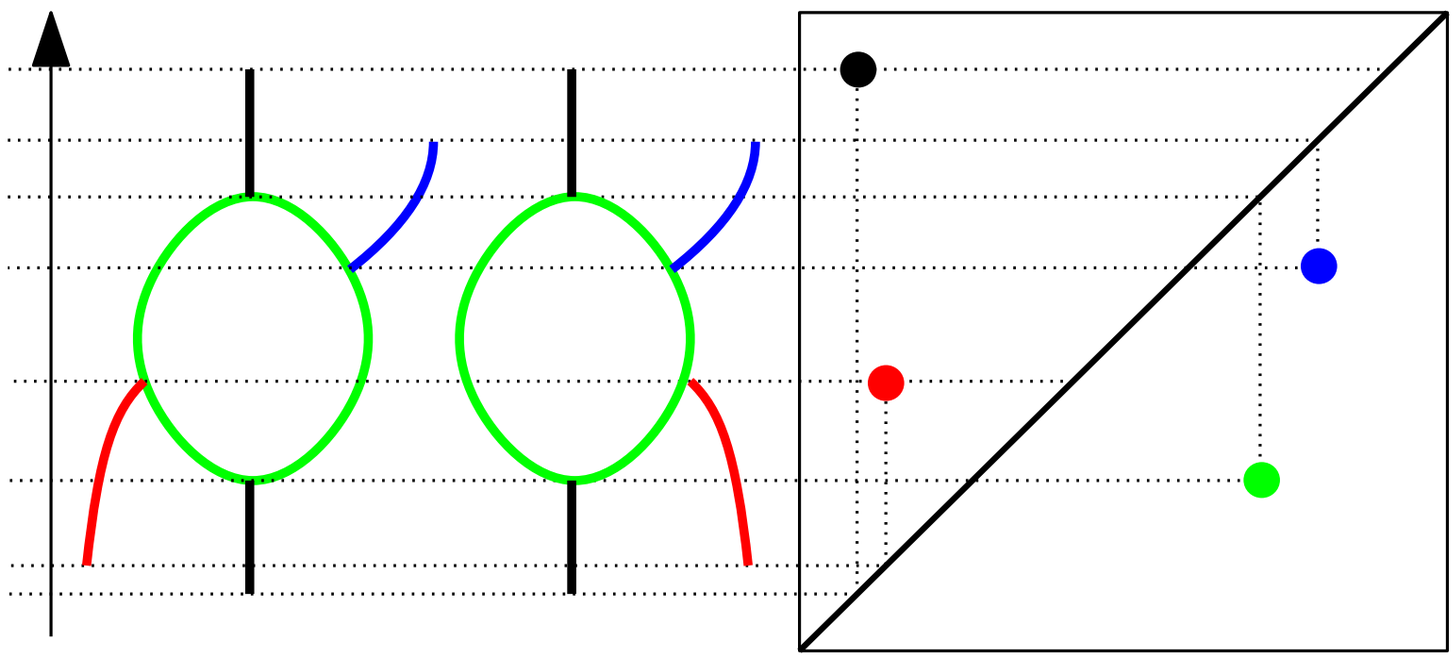
Thm (stability): [Bauer, Ge, Wang 2013]

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

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

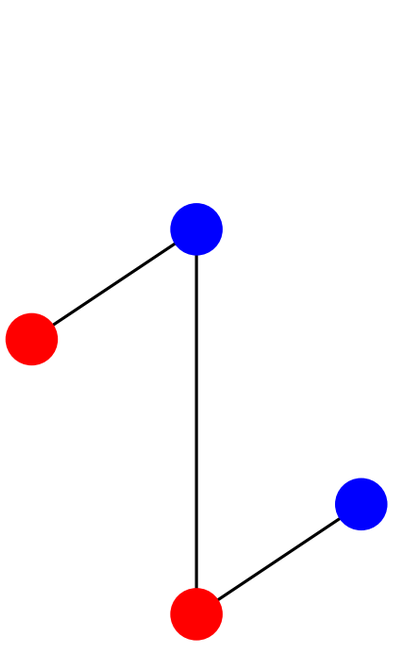
Thm: [C., Oudot 2016]

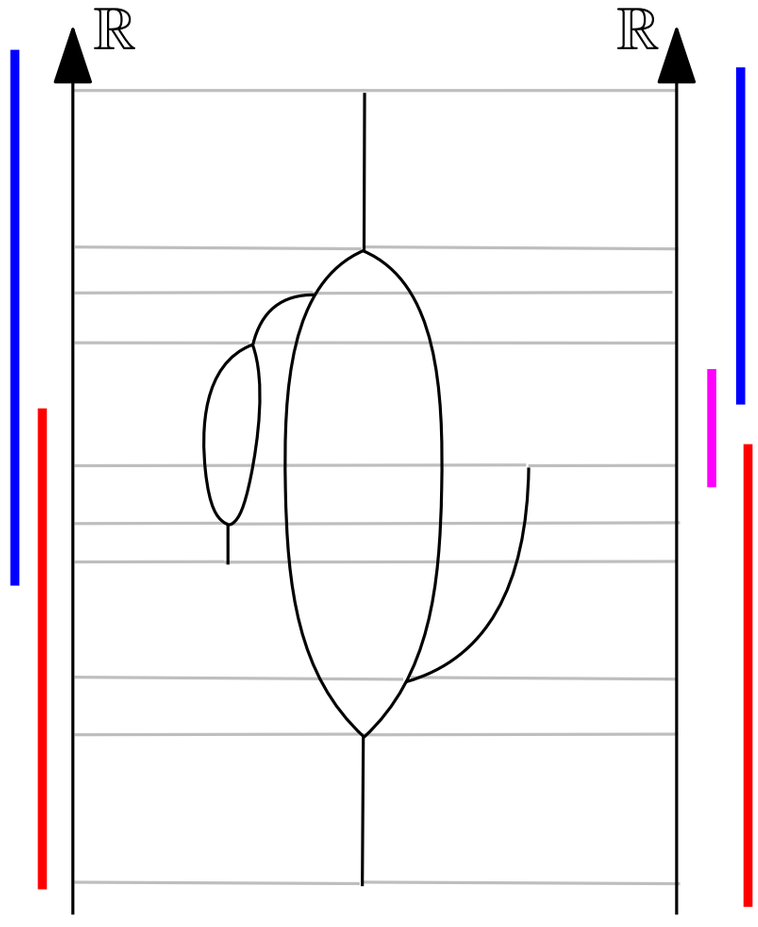
$\mathrm{d}_B(\mathrm{Dg}\,(\cdot), \mathrm{Dg}\,(\cdot))$ is *locally* a metric equivalent to d_{GH}

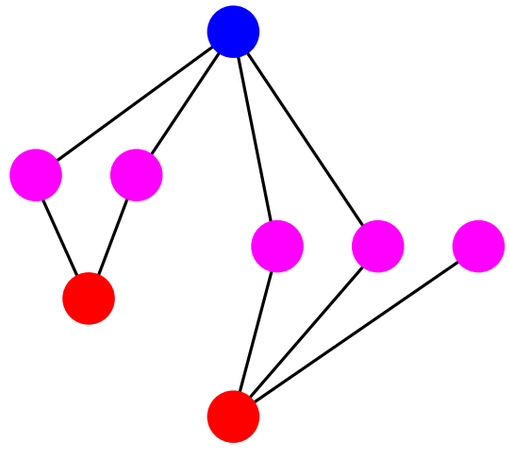


Structure of Mapper

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







Structure of Mapper

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

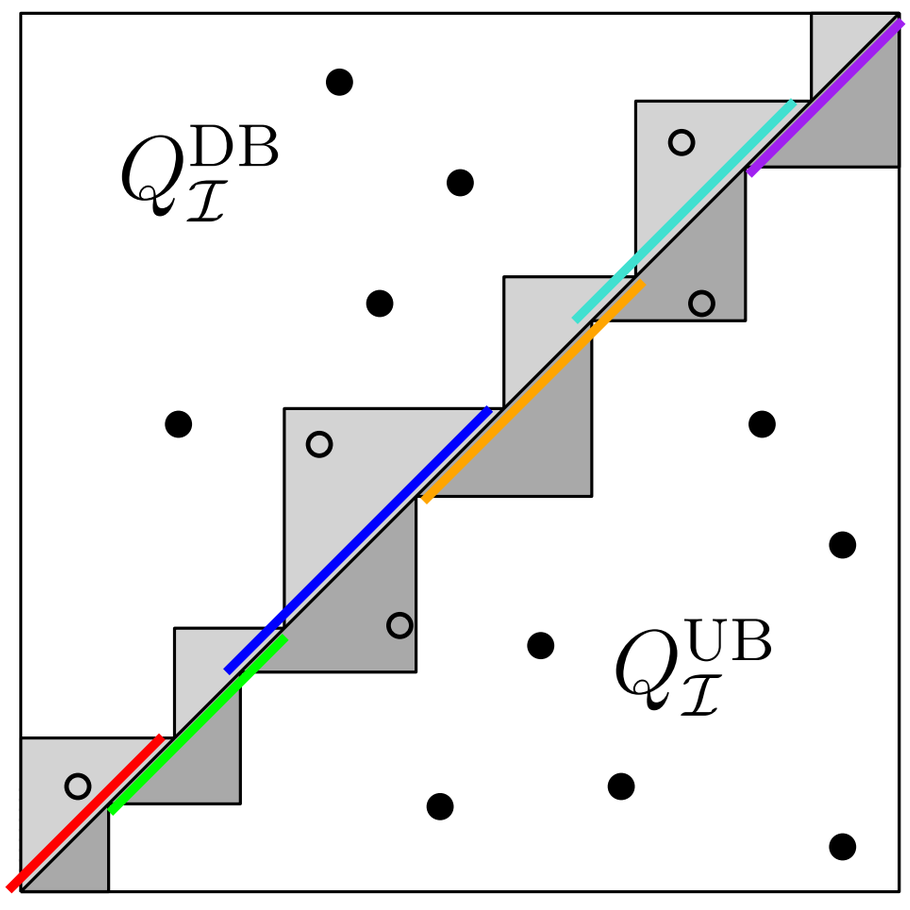
$$\mathrm{Dg}\left(\mathrm{M}_f\right) = \left(\mathrm{DB}(\mathrm{R}_f) \setminus Q_{\mathcal{I}}^{\mathrm{DB}}\right) \cup \left(\mathrm{UB}(\mathrm{R}_f) \setminus Q_{\mathcal{I}}^{\mathrm{UB}}\right) \cup \left(\mathrm{L}(\mathrm{R}_f) \setminus Q_{\mathcal{I}}^{\mathrm{L}}\right)$$

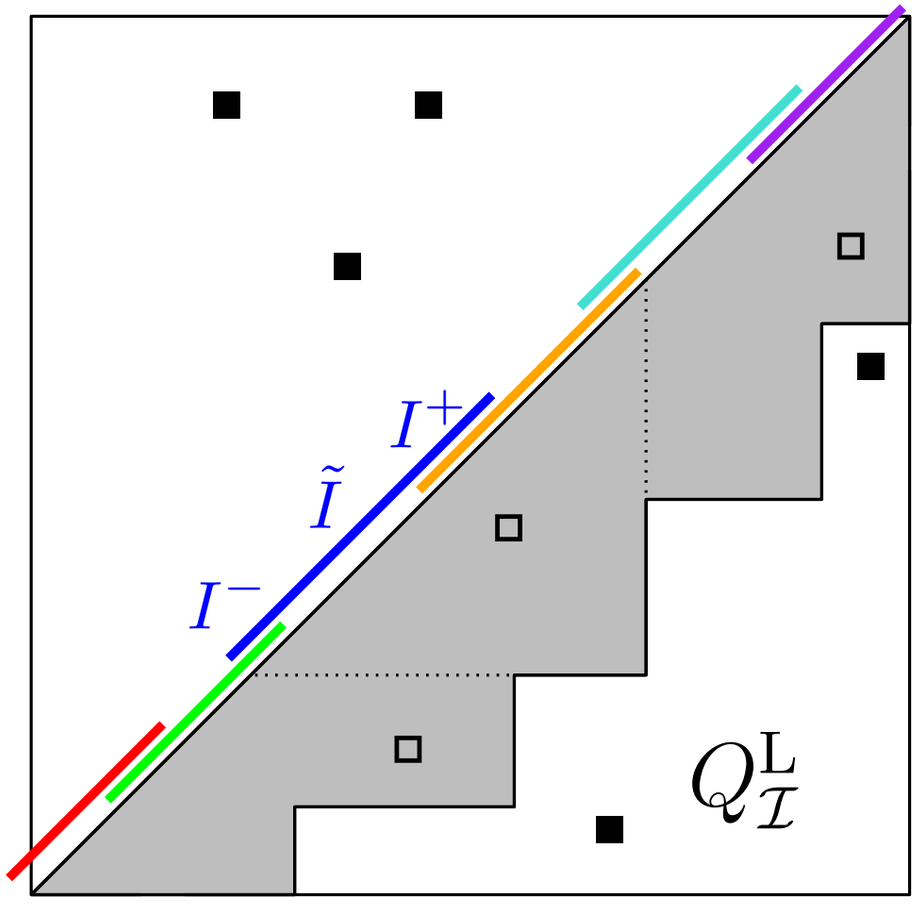
Structure of Mapper

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

$$\mathrm{Dg}\left(\mathrm{M}_f\right) = \left(\mathrm{DB}(\mathrm{R}_f) \setminus Q_{\mathcal{I}}^{\mathrm{DB}}\right) \cup \left(\mathrm{UB}(\mathrm{R}_f) \setminus Q_{\mathcal{I}}^{\mathrm{UB}}\right) \cup \left(\mathrm{L}(\mathrm{R}_f) \setminus Q_{\mathcal{I}}^{\mathrm{L}}\right)$$

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





Structure of Mapper

Thm: [C., Oudot 2016]

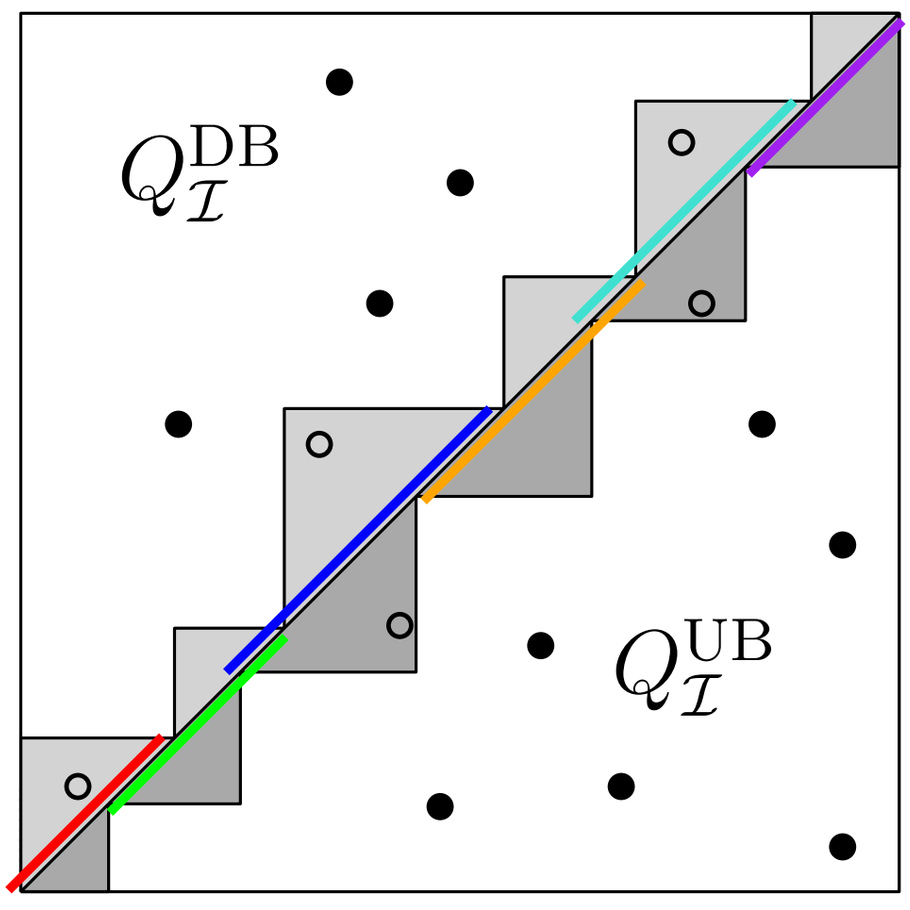
$\mathrm{Dg}\,(\mathrm{M}_f)$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X, \mathcal{I})$:

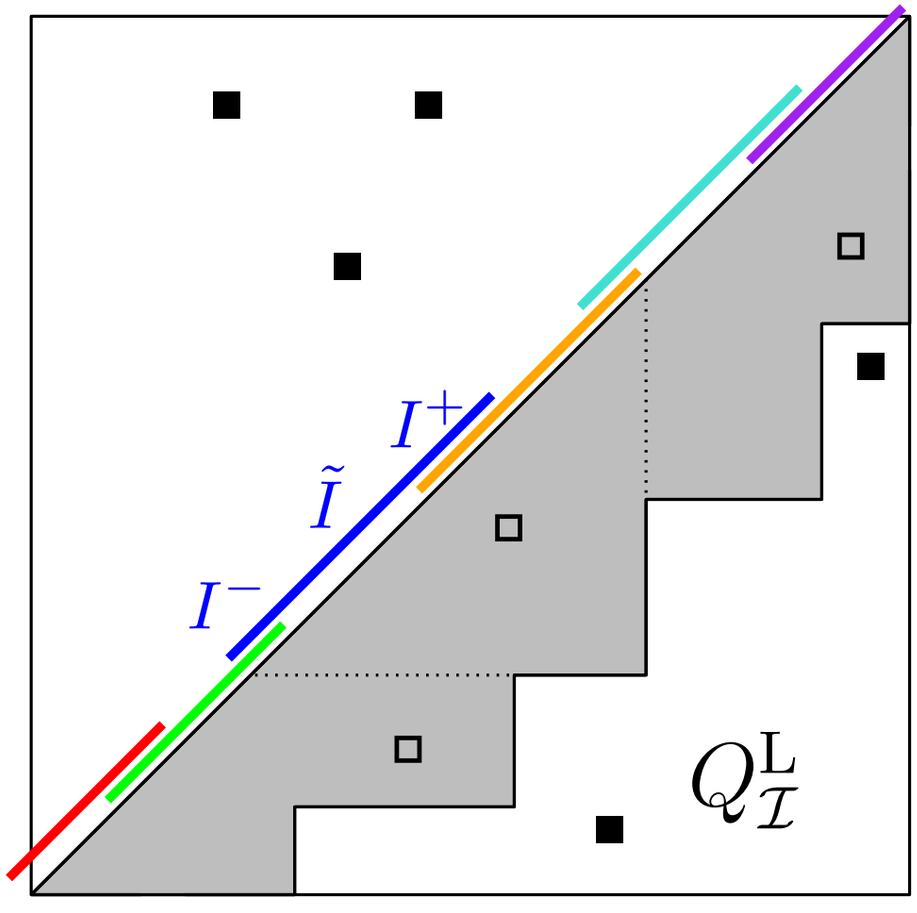
DB $\longleftrightarrow$ downward branches

UB $\longleftrightarrow$ upward branches

T $\longleftrightarrow$ trunks (cc)

L $\longleftrightarrow$ loops





Structure of Mapper

Thm: [C., Oudot 2016]

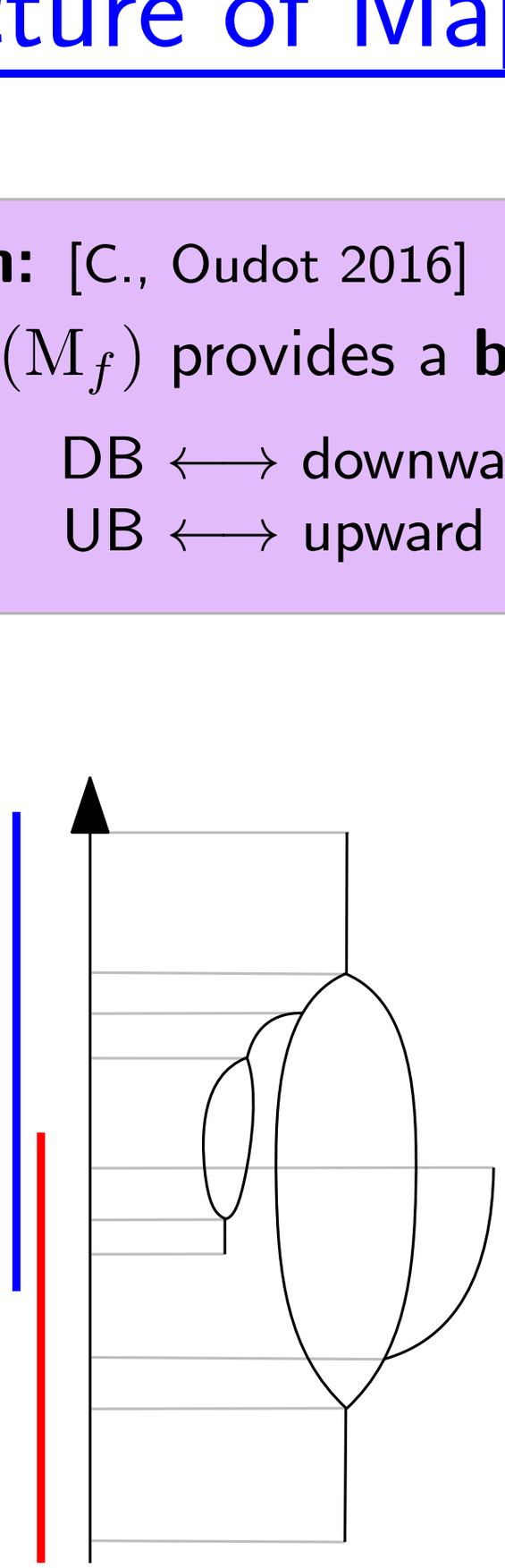
$\mathrm{Dg}\,(\mathrm{M}_f)$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X, \mathcal{I})$:

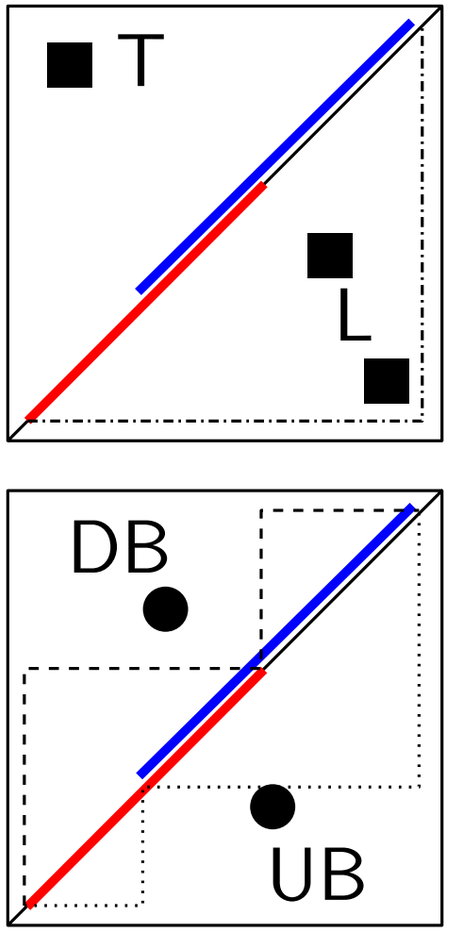
DB $\longleftrightarrow$ downward branches

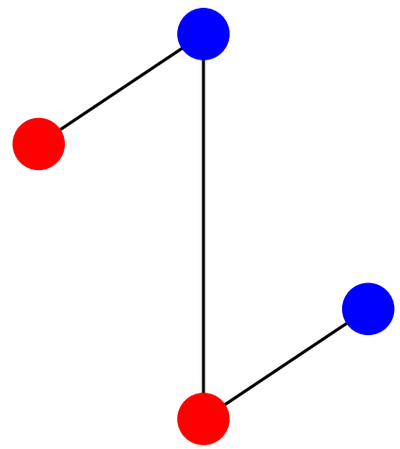
UB $\longleftrightarrow$ upward branches

T $\longleftrightarrow$ trunks (cc)

L $\longleftrightarrow$ loops







Structure of Mapper

Thm: [C., Oudot 2016]

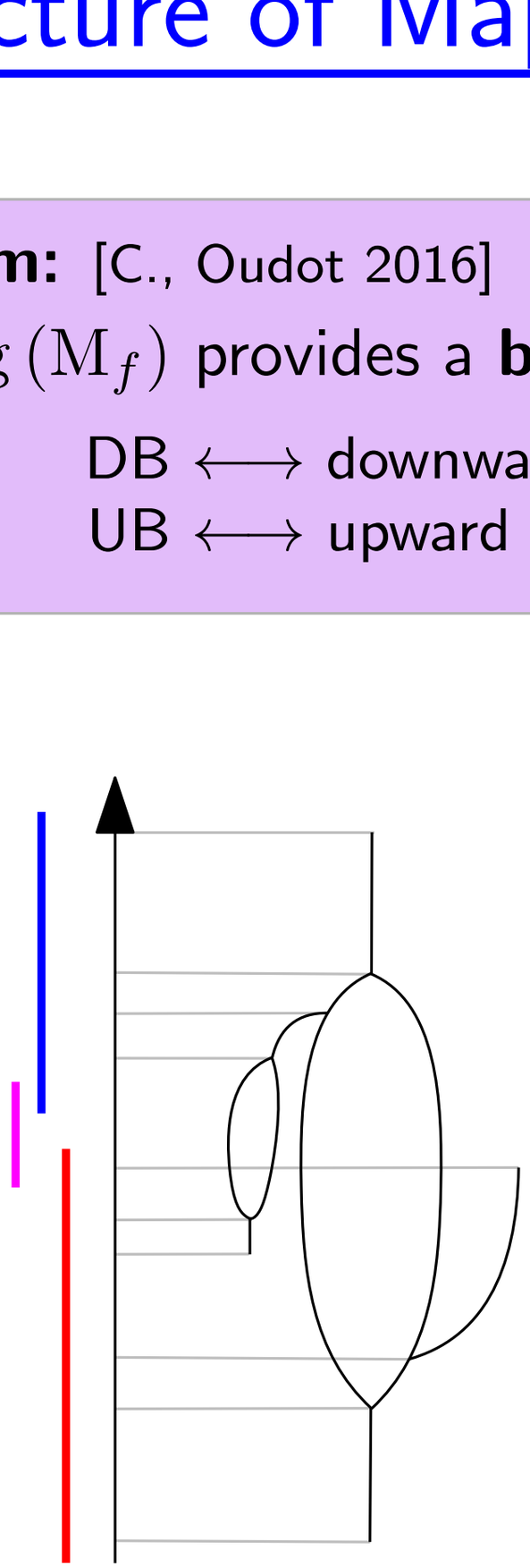
$\mathrm{Dg}\,(\mathrm{M}_f)$ provides a **bag-of-features** descriptor for $\mathrm{M}_f(X, \mathcal{I})$:

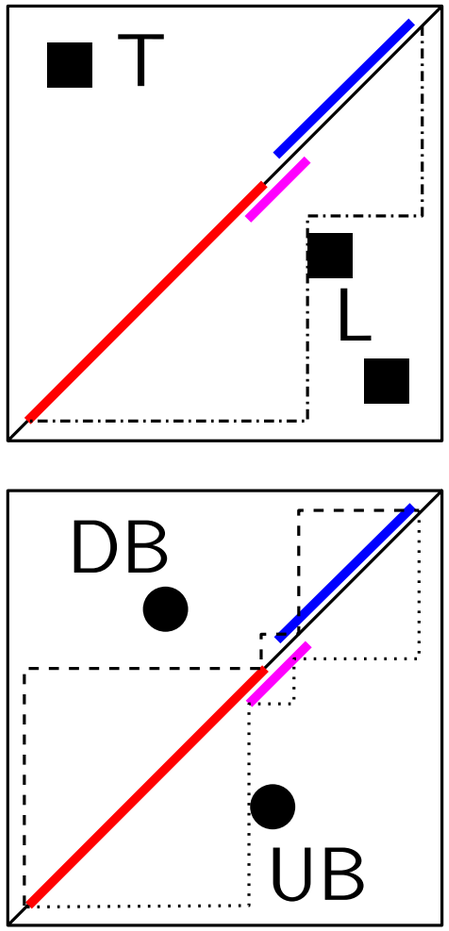
DB $\longleftrightarrow$ downward branches

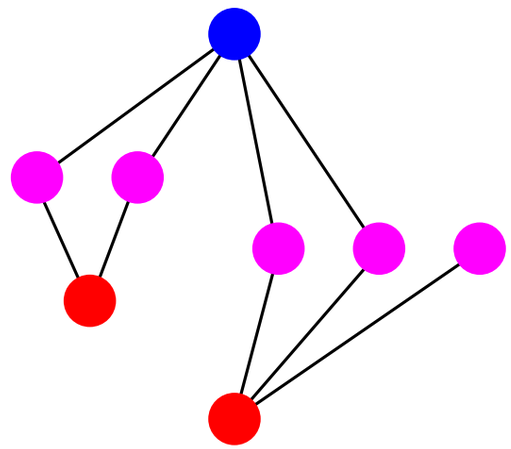
UB $\longleftrightarrow$ upward branches

T $\longleftrightarrow$ trunks (cc)

L $\longleftrightarrow$ loops

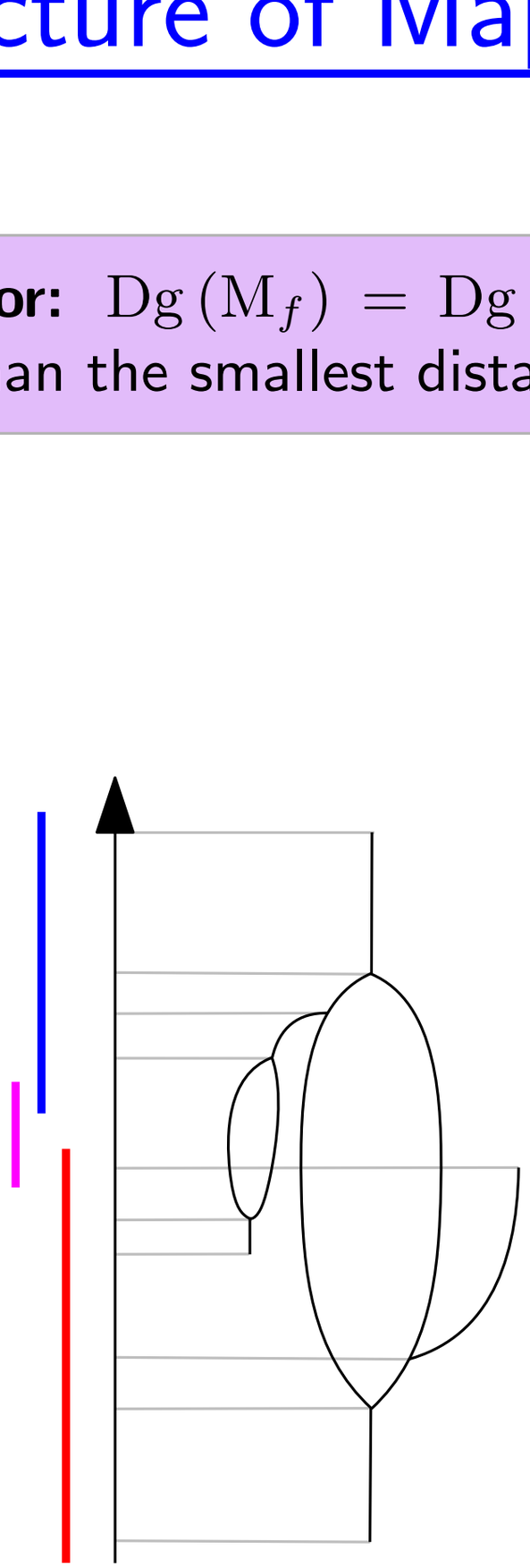


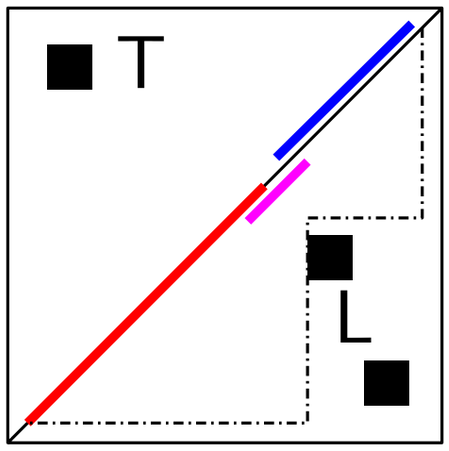


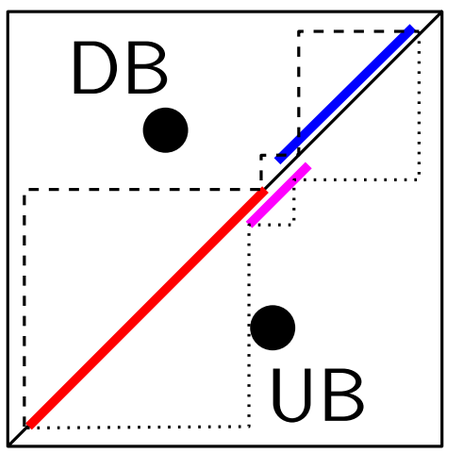


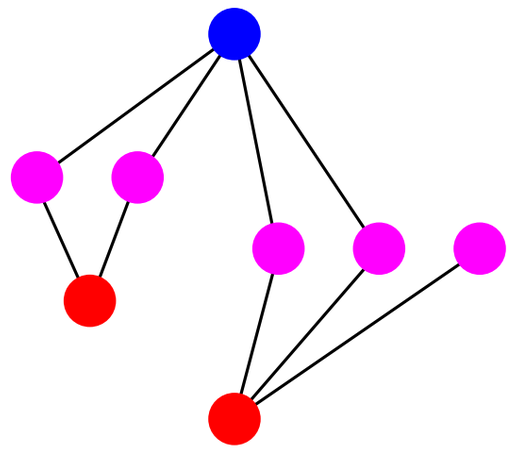
Structure of Mapper

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





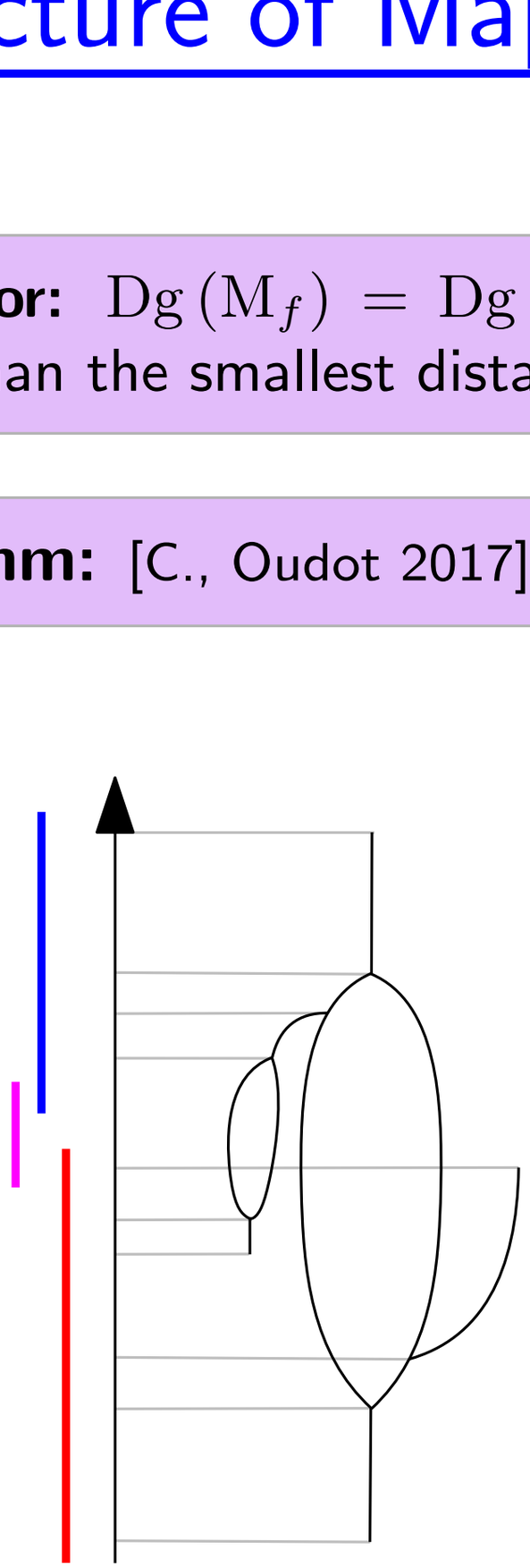


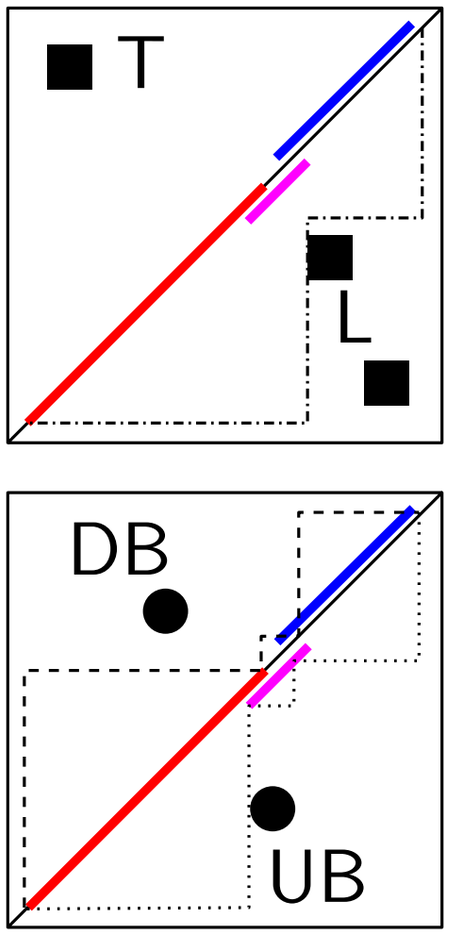


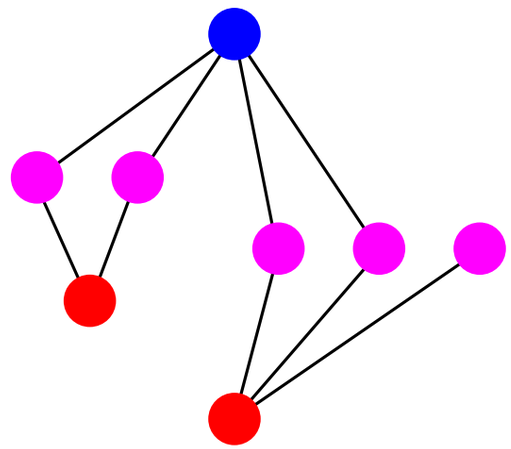
Structure of Mapper

Cor: $\mathrm{Dg}\,(\mathrm{M}_f) = \mathrm{Dg}\,(\mathrm{R}_f)$ 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

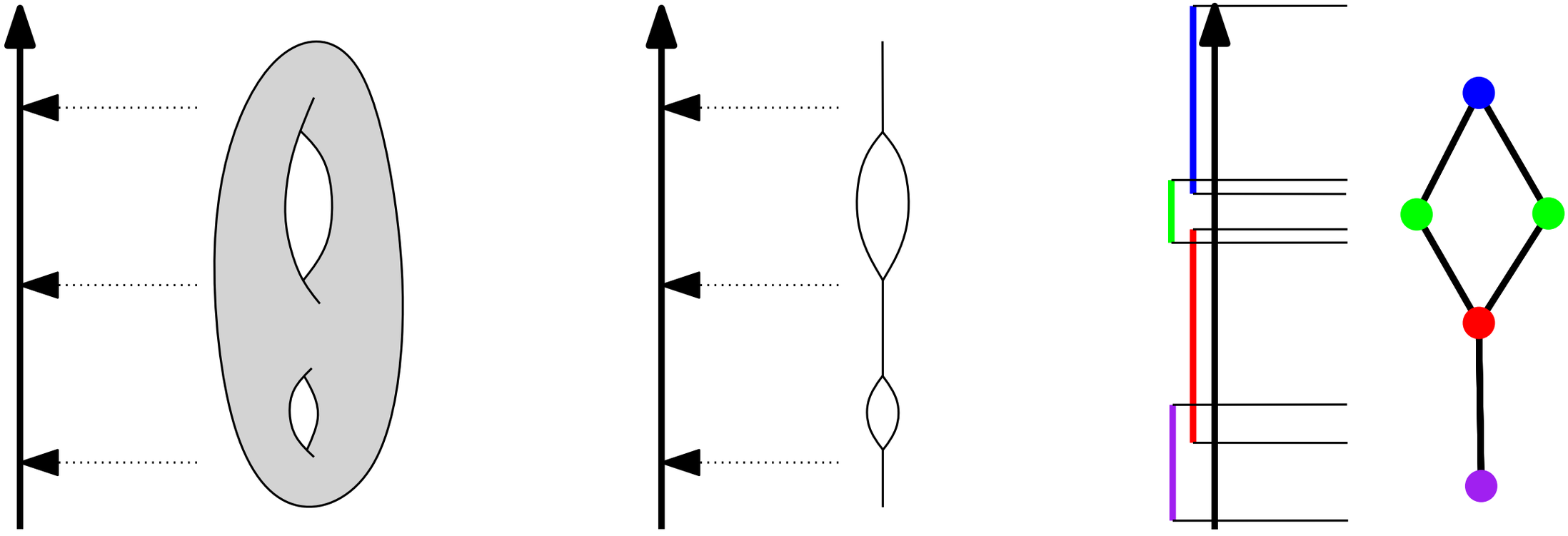
Thm: [C., Oudot 2017] $\mathrm{d_{GH}}(\mathrm{M}_f(X, \mathcal{I}), \mathrm{R}_f(X)) \leq 3r$

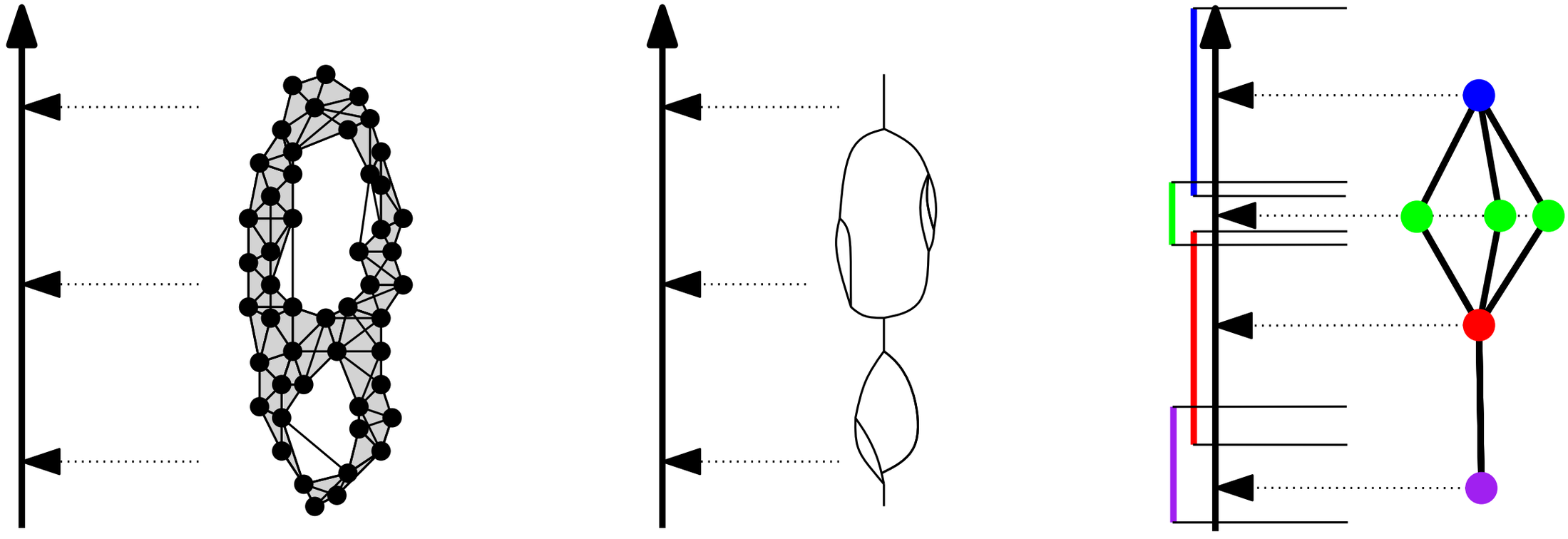




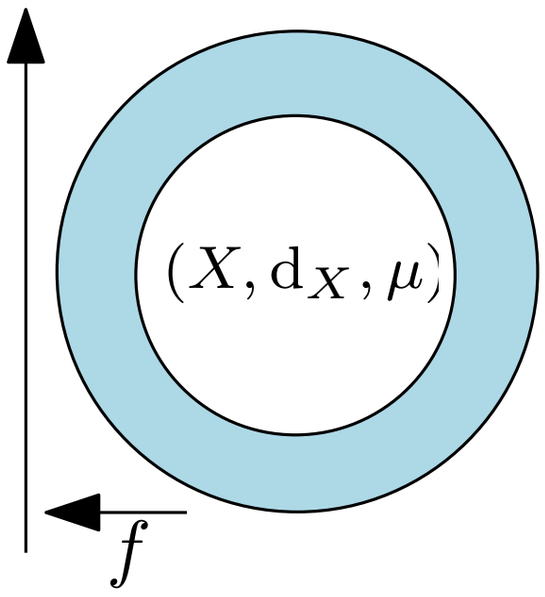


Statistics for Mapper

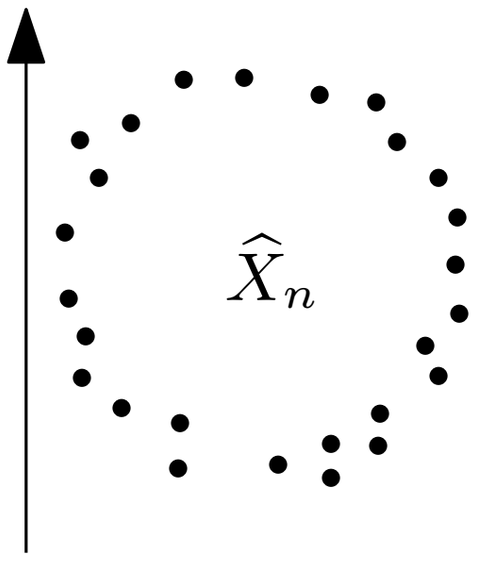




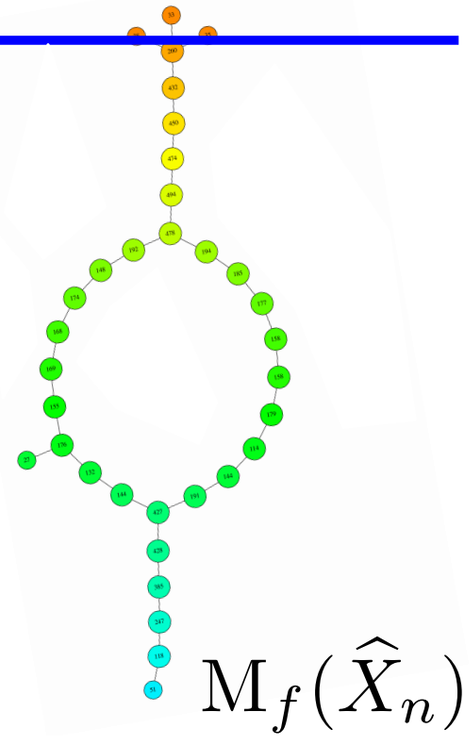
Setup



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

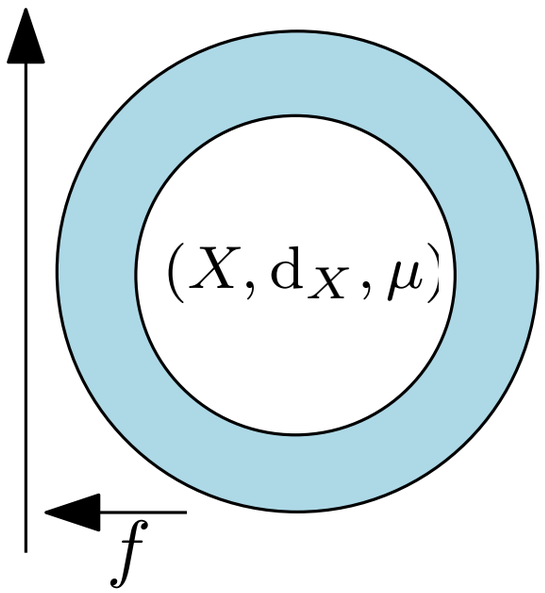


δ_n
g_n, r_n

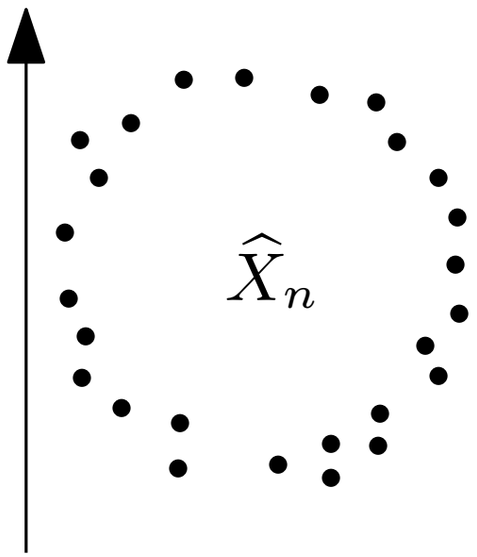


$\mathrm{M}_f(\widehat{X}_n)$

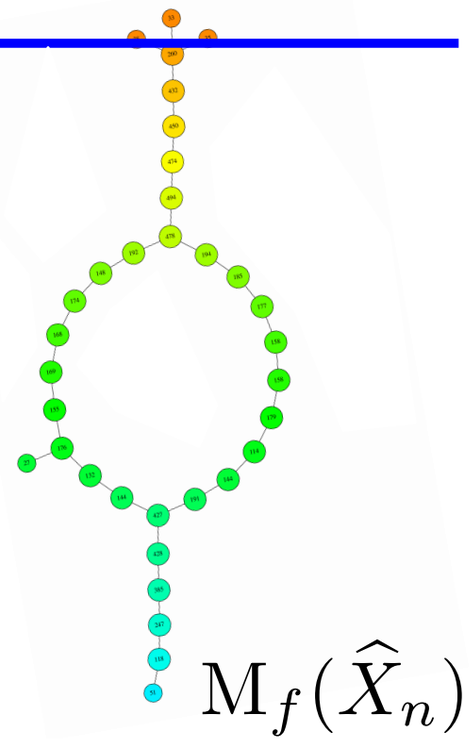
Setup



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



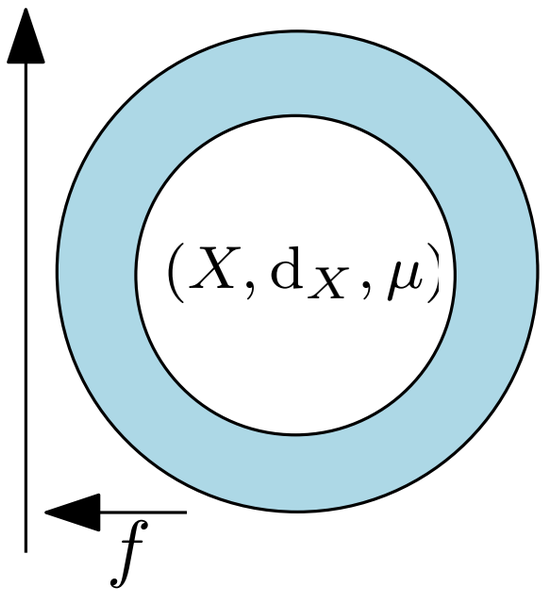
δ_n
g_n, r_n



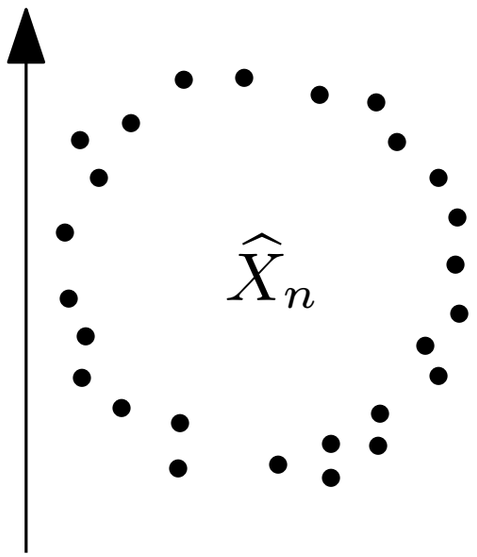
$\mathrm{M}_f(\widehat{X}_n)$

Prop: [C., Michel, Oudot 2017] $\widehat{\mathrm{M}}_n = \mathrm{M}_f(\widehat{X}_n)$ is measurable

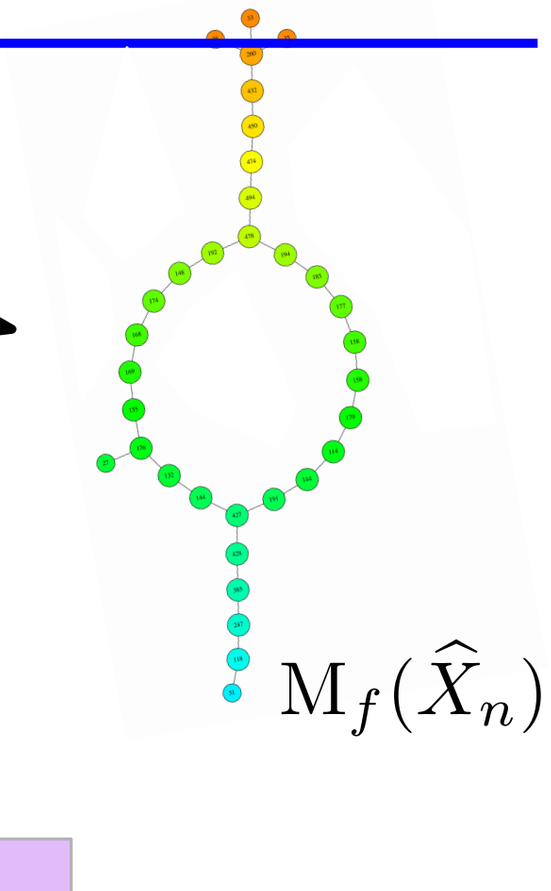
Setup



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



δ_n
g_n, r_n

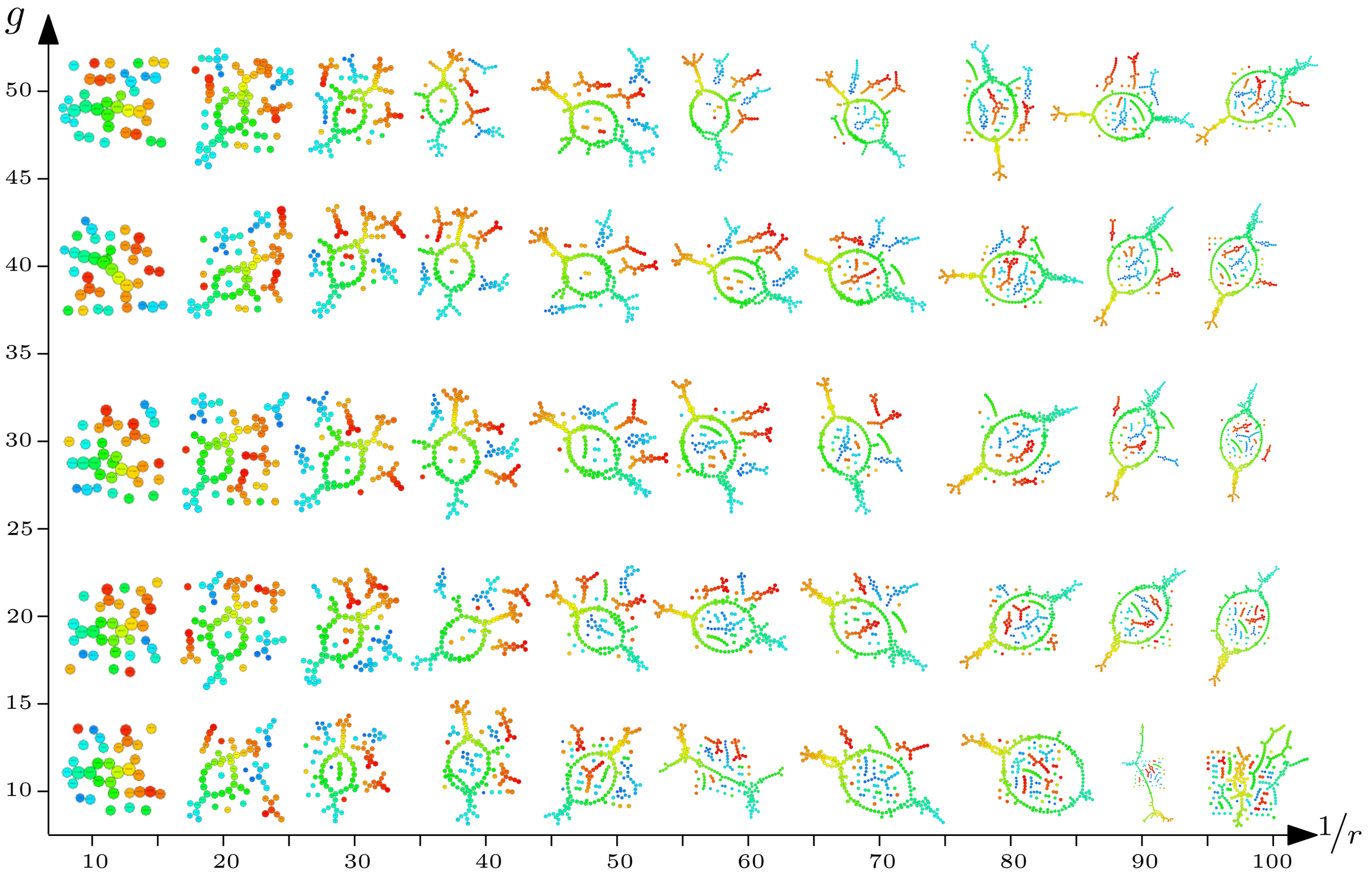


Prop: [C., Michel, Oudot 2017] $\widehat{\mathrm{M}}_n = \mathrm{M}_f(\widehat{X}_n)$ is measurable

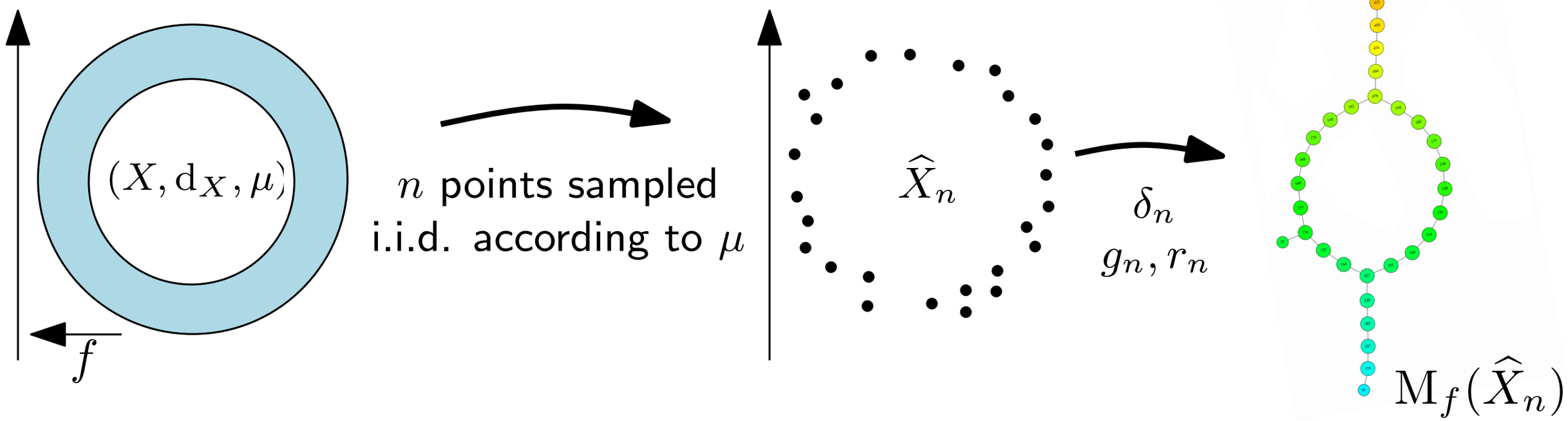
Goal: Find heuristics to compute "good" δ_n, g_n, r_n

Assess quality through confidence regions and convergence rates

Setup



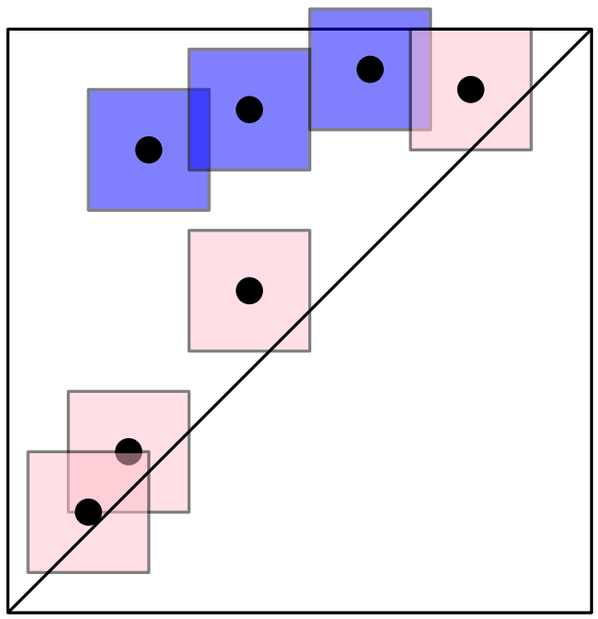
Setup

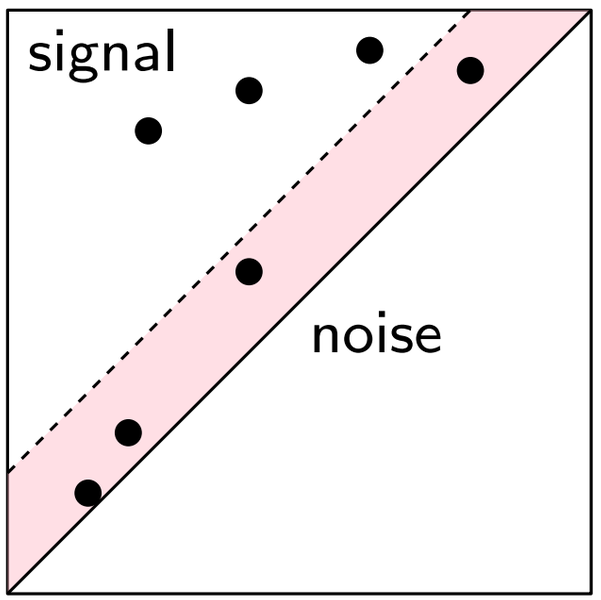


Confidence regions: given $\alpha \in (0,1)$, find $c_n(\alpha) \geq 0$ s.t.:

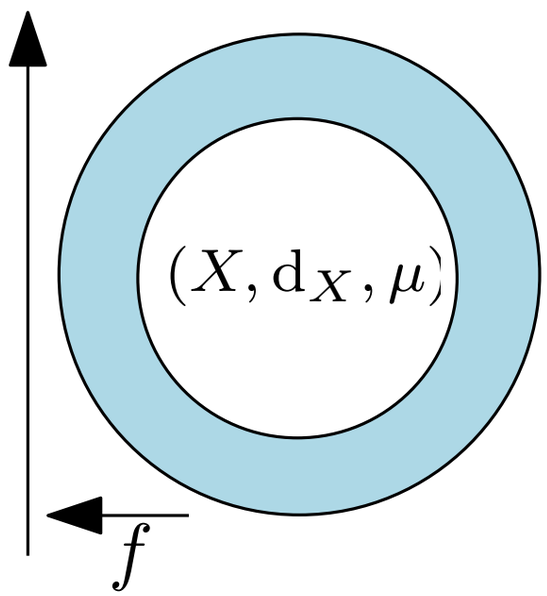
$$\limsup_{n\to\infty} \quad \mathbb{P}\left(\mathrm{d}_B\left(\widehat{\mathrm{M}}_n, \mathrm{R}_f(X)\right) > c_n(\alpha)\right) \leq \alpha$$

$\rightarrow$ d_B-ball of radius $c_n(\alpha)$ around $\mathrm{Dg}\left(\widehat{\mathrm{M}}_n\right)$

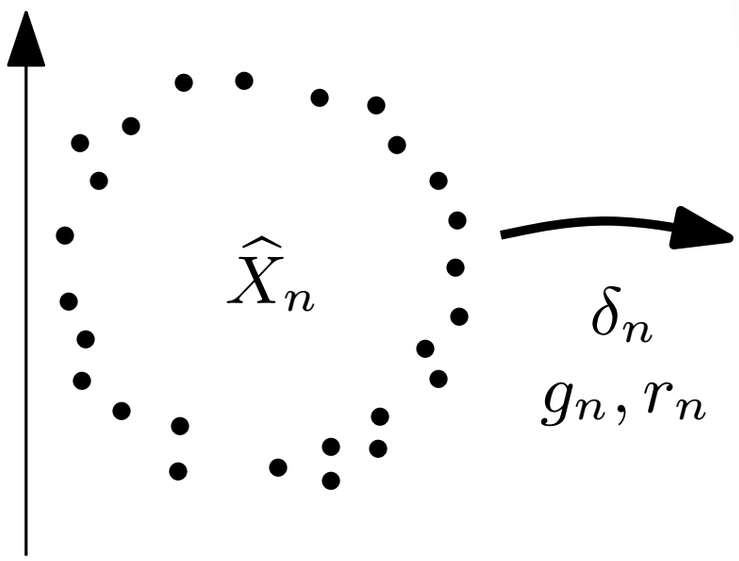


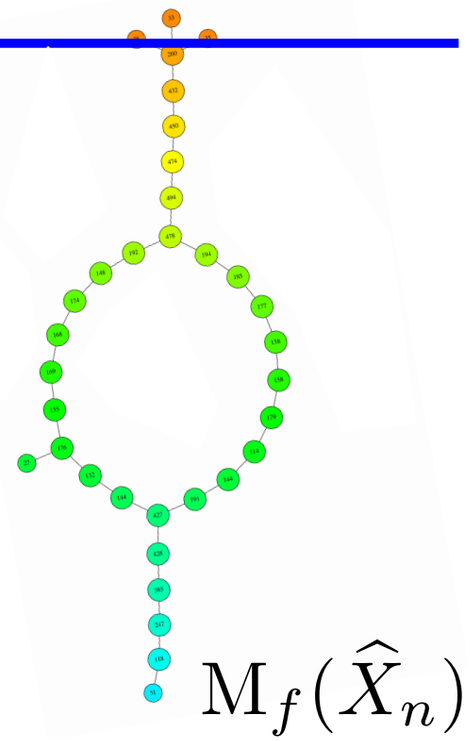


Setup



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





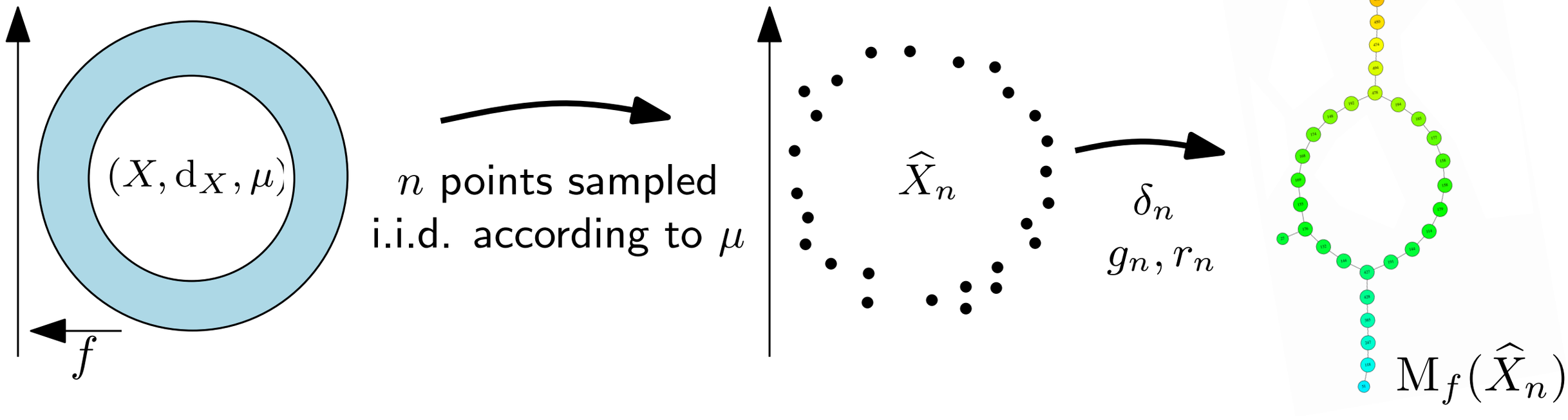
Confidence regions: given $\alpha \in (0,1)$, find $c_n(\alpha) \geq 0$ s.t.:

$$\limsup_{n\to\infty} \quad \mathbb{P}\left(\mathrm{d}_B\left(\widehat{\mathrm{M}}_n, \mathrm{R}_f(X)\right) > c_n(\alpha)\right) \leq \alpha$$

$\rightarrow$ d_B-ball of radius $c_n(\alpha)$ around $\mathrm{Dg}\left(\widehat{\mathrm{M}}_n\right)$

Convergence Rate: estimate $\mathbb{E}\left[\mathrm{d}_B(\widehat{\mathrm{M}}_n, \mathrm{R}_f(X))\right]$ w.r.t. n

Rate of Convergence



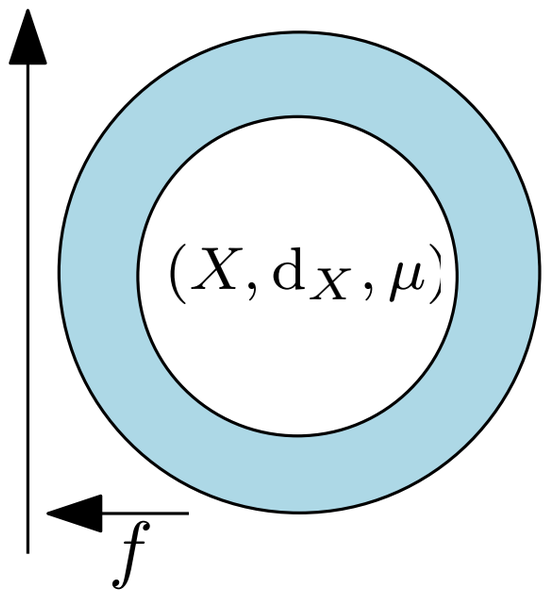
Regularity of the filter function: (exact) modulus of continuity of f

$$\omega(\delta) = \sup_{\|x-x'\| \leq \delta} |f(x) - f(x')|$$

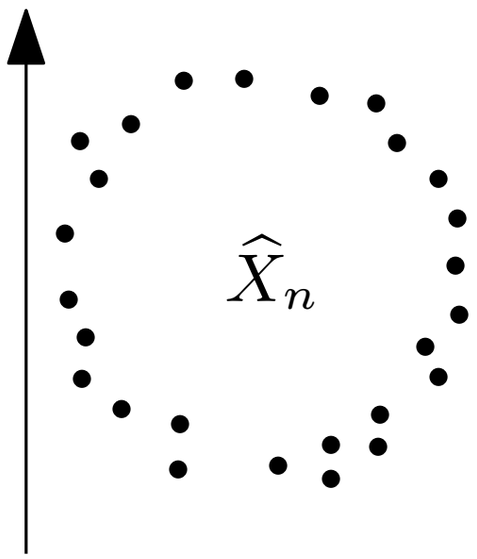
Approximation inequality: [C., Michel, Oudot 2017]
Let $\widehat{X}_n \subset X$. Under some *regularity assumptions* on X, f, δ, r, g, one has:

$$\mathrm{d}_B \left(\mathrm{R}_f(X), \mathrm{M}_f(\widehat{X}_n) \right) \leq r + 2\omega(\delta)$$

Rate of Convergence

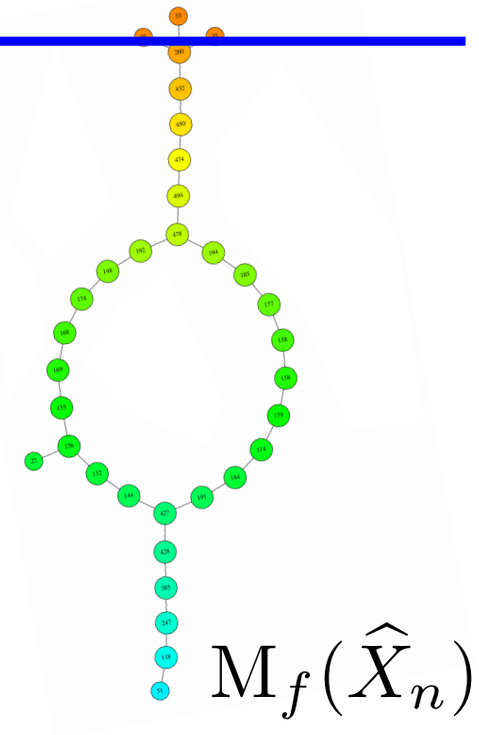


n points sampled
i.i.d. according to μ



δ_n

g_n, r_n



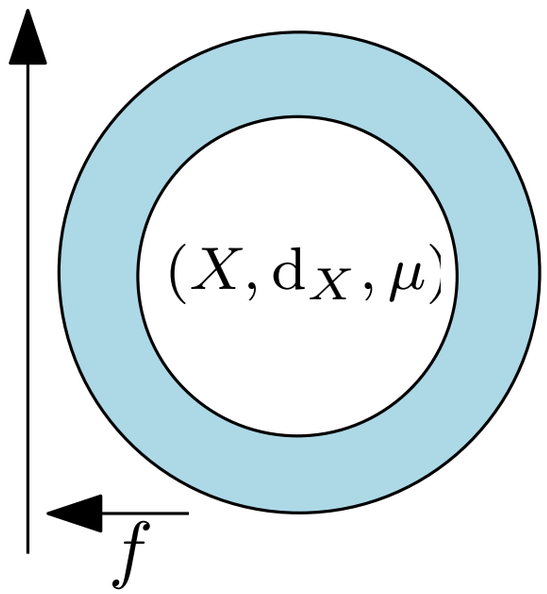
$$4\mathrm{d_H}(\widehat{X}_n, X) \leq \delta \leq C(X)$$

$$\max\{|f(X_i) - f(X_j)| \ : \ \|X_i - X_j\| \leq \delta\} < gr$$

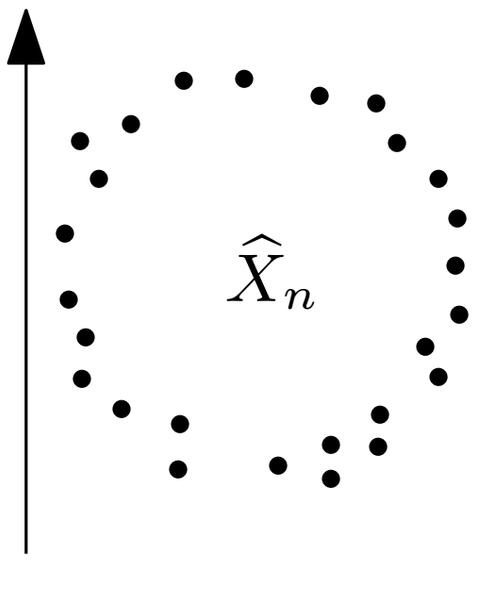
Approximation inequality: [C., Michel, Oudot 2017]
Let $\widehat{X}_n \subset X$. Under some *regularity assumptions* on X, f, δ, r, g, one has:

$$\mathrm{d}_B\left(\mathrm{R}_f(X), \mathrm{M}_f(\widehat{X}_n)\right) \leq r + 2\omega(\delta)$$

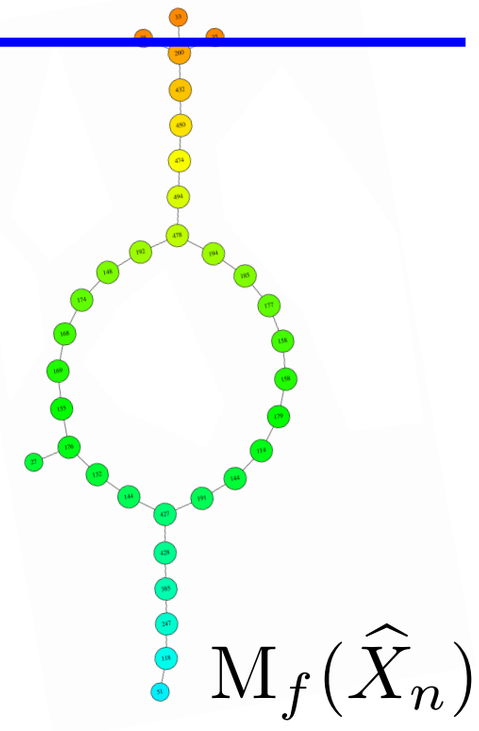
Rate of Convergence



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



δ_n
g_n, r_n



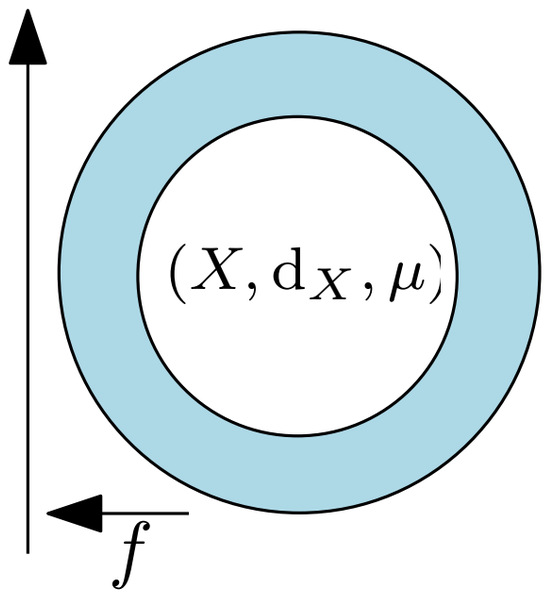
$$4\mathrm{d_H}(X_n, X) \leq \delta \leq C(X)$$

$$\max\{\max\{|f(X_i) - f(X_j)|, |\widehat{f}(X_i) - \widehat{f}(X_j)|\} \ : \ \|X_i - X_j\| \leq \delta\} \leq rg$$

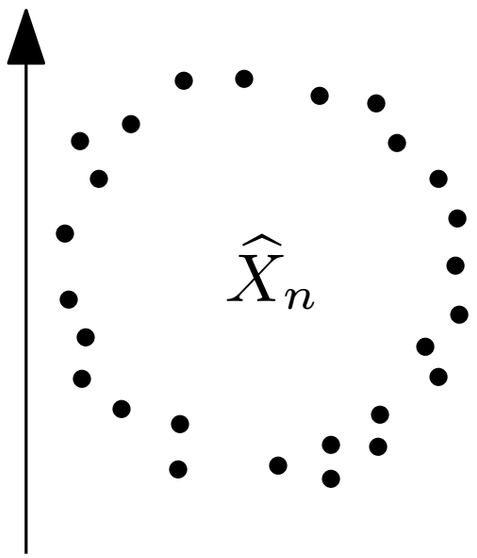
Approximation inequality: [C., Michel, Oudot 2017]
Let $\widehat{X}_n \subset X$. Under some *regularity assumptions* on X, f, δ, r, g, one has:

$$\mathrm{d}_B\left(\mathrm{R}_f(X), \mathrm{M}_{\widehat{f}}(\widehat{X}_n)\right) \leq 2r + 2\omega(\delta) + \max\{|f(X_i) - \widehat{f}(X_i)|\}$$

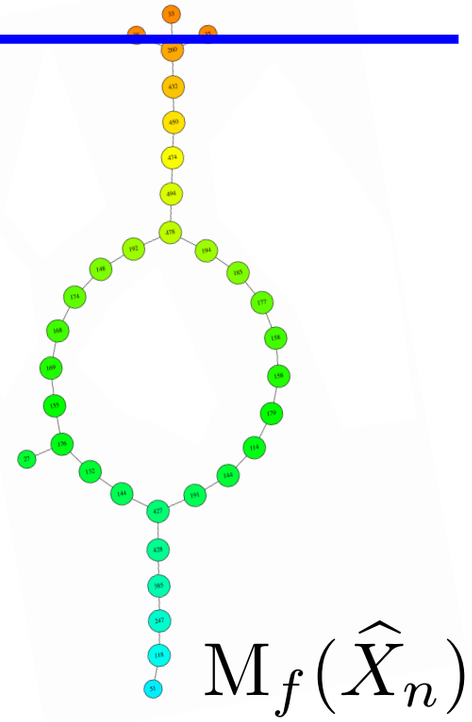
Rate of Convergence



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



δ_n
g_n, r_n



$\mathrm{M}_f(\widehat{X}_n)$

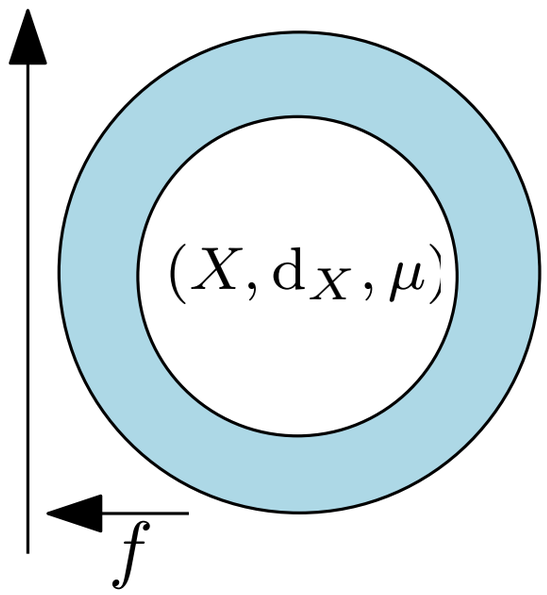
$$V_n = \max\{f(X_i) - f(X_j) \ : \ \|X_i - X_j\| \leq \delta_n\}$$

Thm: [C., Michel, Oudot 2017]

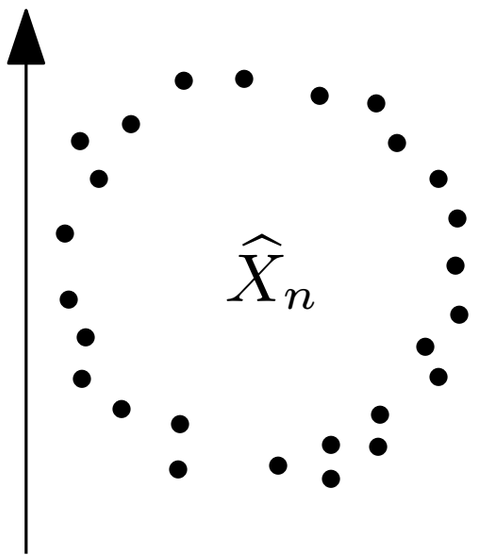
If μ is (a, b)-standard, then for $\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b}$, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{V_n}{g_n}$, one has:

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

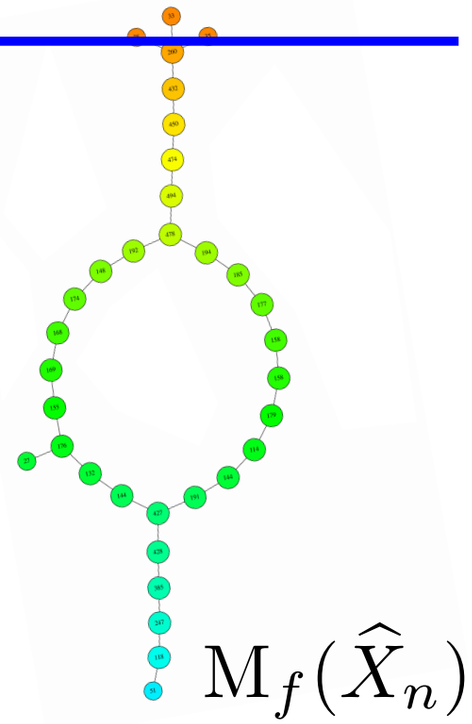
Rate of Convergence



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



δ_n
g_n, r_n



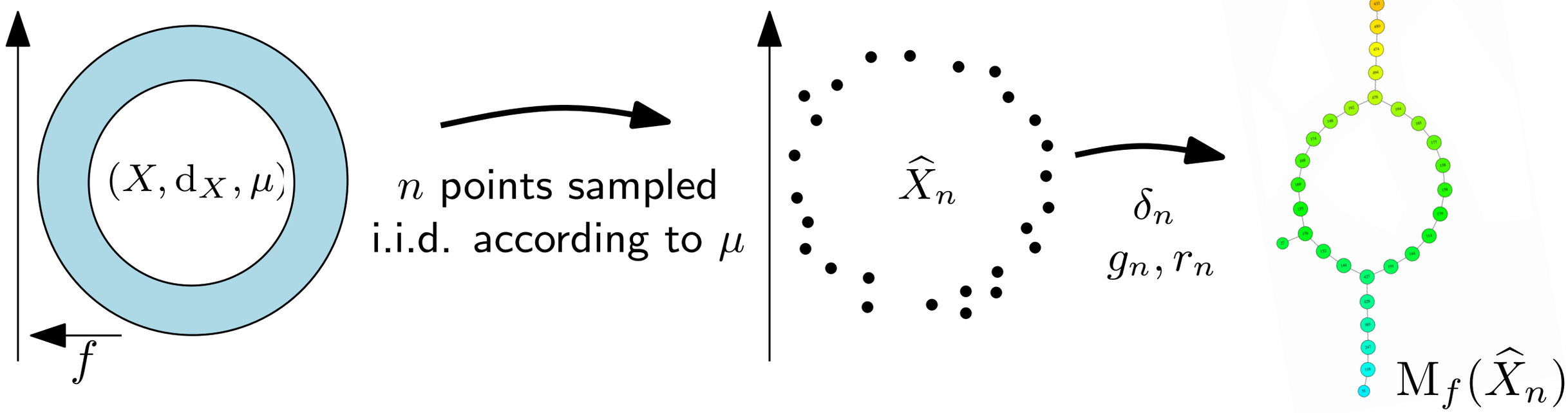
$$V_n = \max\{f(X_i) - f(X_j) \ : \ \|X_i - X_j\| \leq \delta_n\}$$

Thm: [C., Michel, Oudot 2017]

If μ is (a, b)-standard, then for $\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b}$, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{V_n}{g_n}$, one has:

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

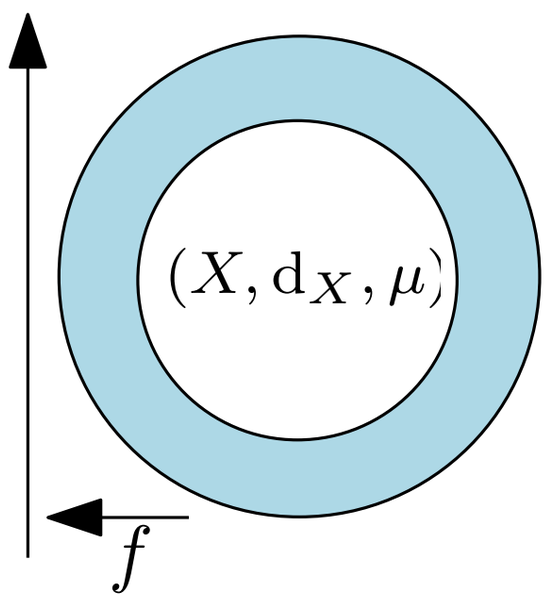
Rate of Convergence



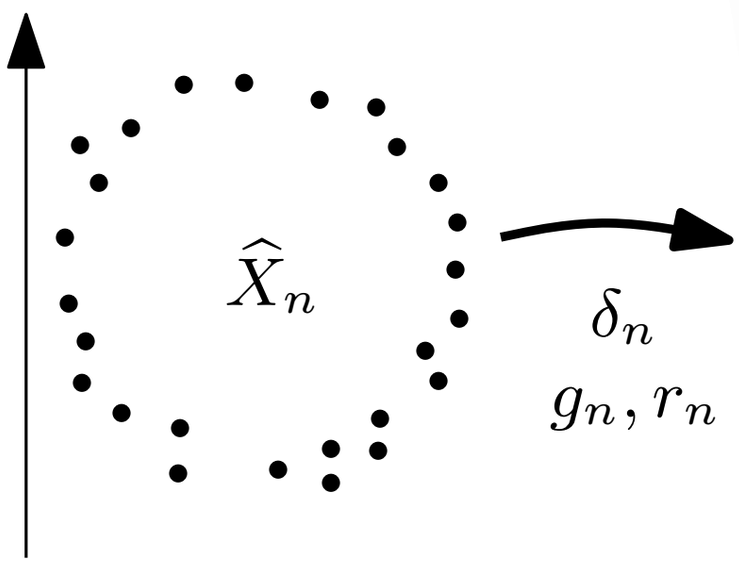
Subsampling to tune δ_n: let $\beta > 0$ and take $s(n) = n\log(n)^{-(1+\beta)}$

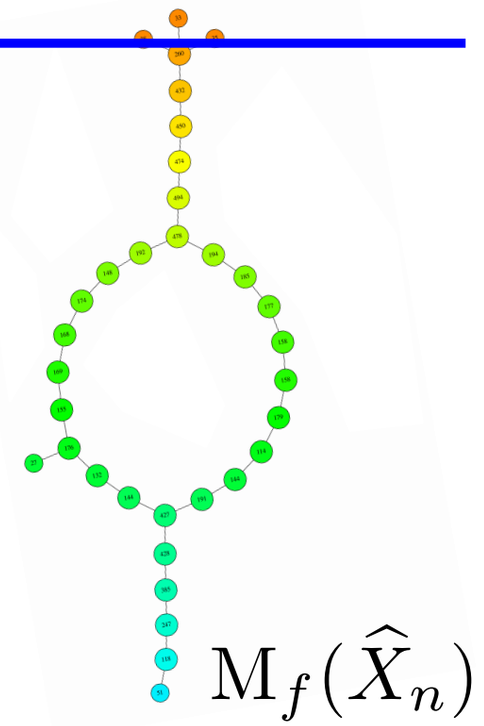
$\delta_n = \mathrm{d_H}(\hat{X}_n^{s(n)}, \hat{X}_n)$ where $\hat{X}_n^{s(n)} \subset \hat{X}_n$ of size $s(n)$

Rate of Convergence



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





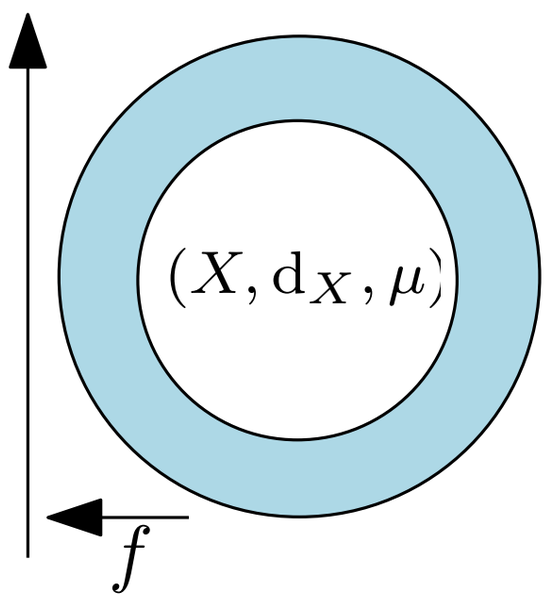
Subsampling to tune δ_n: let $\beta > 0$ and take $s(n) = n\log(n)^{-(1+\beta)}$

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

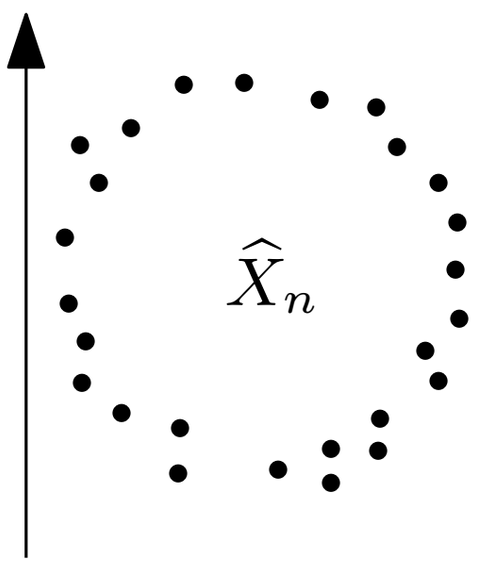
Thm: [C., Michel, Oudot 2017]
If μ is (a, b)-standard, then for δ_n, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{V_n}{g_n}$, one has

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

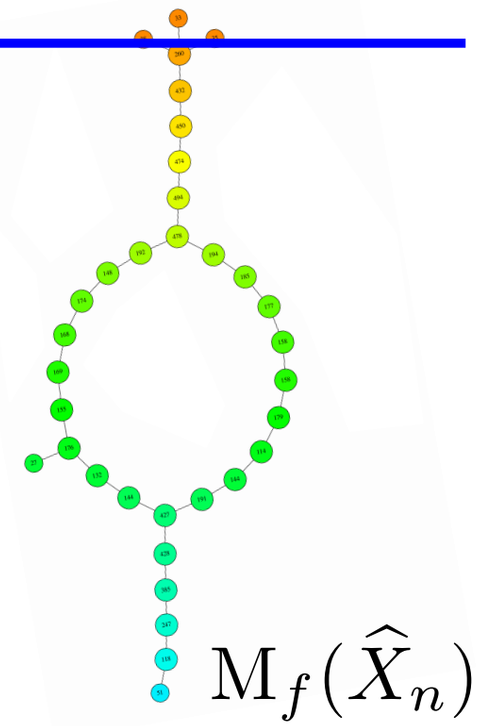
Rate of Convergence



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



δ_n
g_n, r_n



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

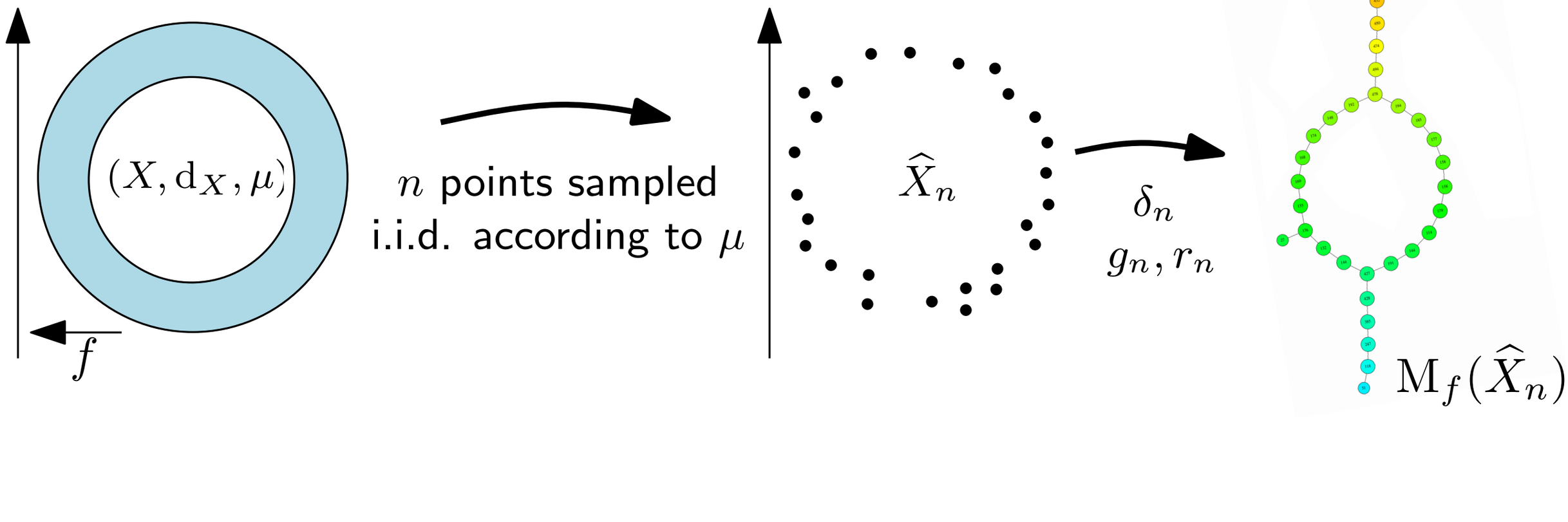
$\delta_n = \mathrm{d_H}(\hat{X}_n^{s(n)}, \hat{X}_n)$ where $\hat{X}_n^{s(n)} \subset \hat{X}_n$ of size $s(n)$

Thm: [C., Michel, Oudot 2017]

If μ is (a, b)-standard, then for δ_n, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{\max\{V_n, \widehat{V}_n\}}{g_n}$, one has

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_B\left(\mathrm{M}_{\widehat{f}}(\widehat{X}_n), \mathrm{R}_f(X)\right)\right] \lesssim \omega\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b} + \mathbb{E}\left[\max|f(X) - \widehat{f}(X)|\right]$$

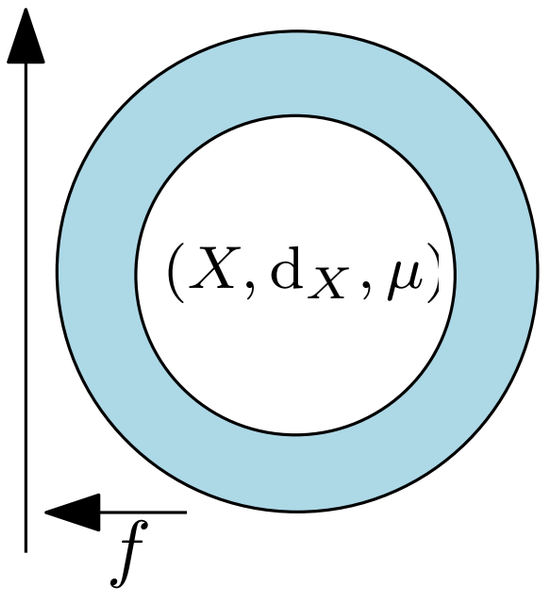
Rate of Convergence



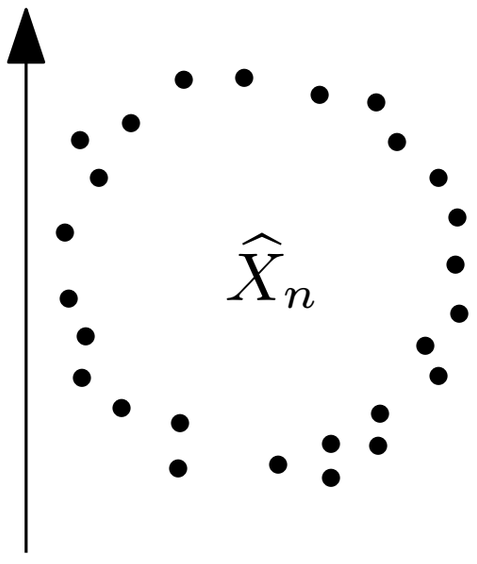
Minimax Optimality: [C., Michel, Oudot 2017] for any estimator $\widehat{\mathrm{R}}_n$,

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

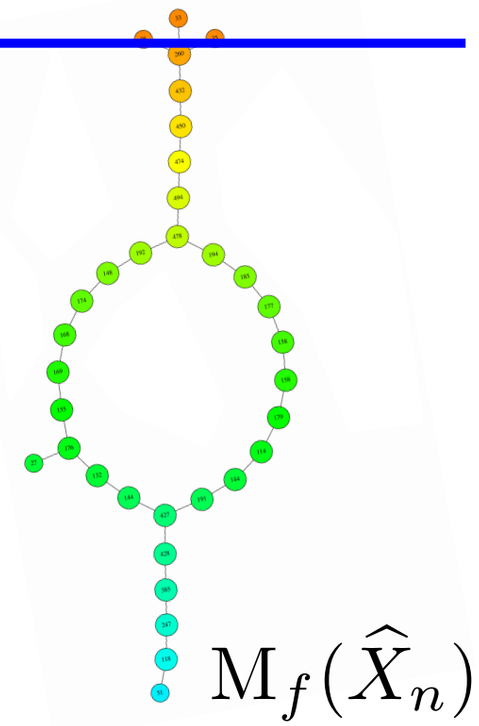
Rate of Convergence



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



δ_n
g_n, r_n



Ex : PCA filter

Π_1: 1st principal direction of covariance operator

$\widehat{\Pi}_1$: 1st principal direction of empirical covariance operator

Using [Biau et. al. 2012]:

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

Confidence Regions

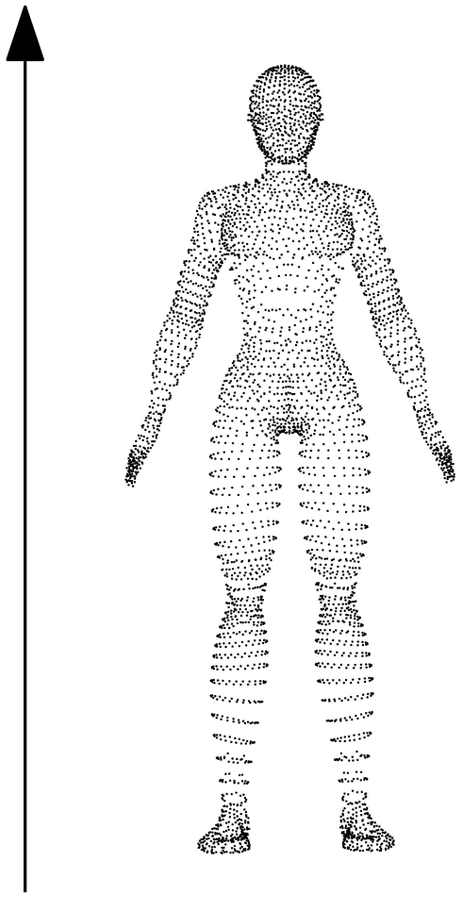
Either from proof of previous result with:

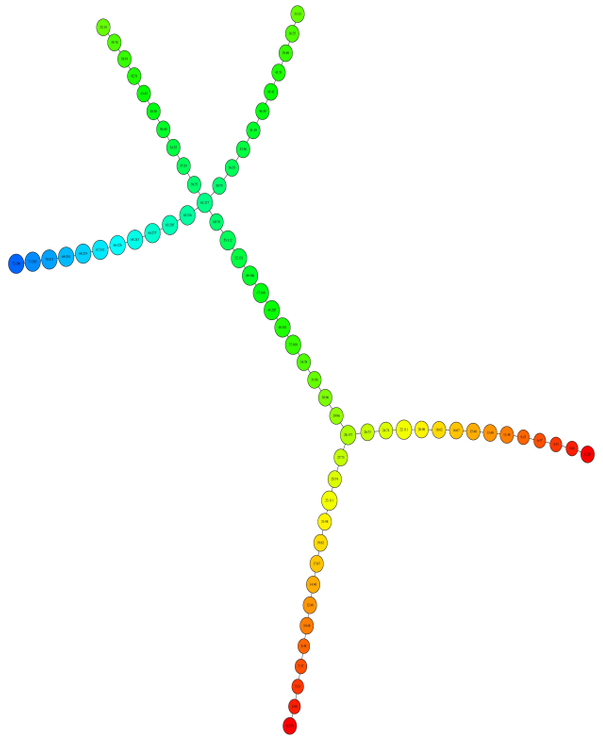
$$\mathbb{E}\left[\mathrm{d}_B\left(\mathrm{M}_f(\widehat{X}_n), \mathrm{R}_f(X)\right)\right] = \int_\alpha \mathbb{P}\left(\mathrm{d}_B\left(\mathrm{M}_f(\widehat{X}_n), \mathrm{R}_f(X)\right) \geq \alpha\right) \mathrm{d}\alpha$$

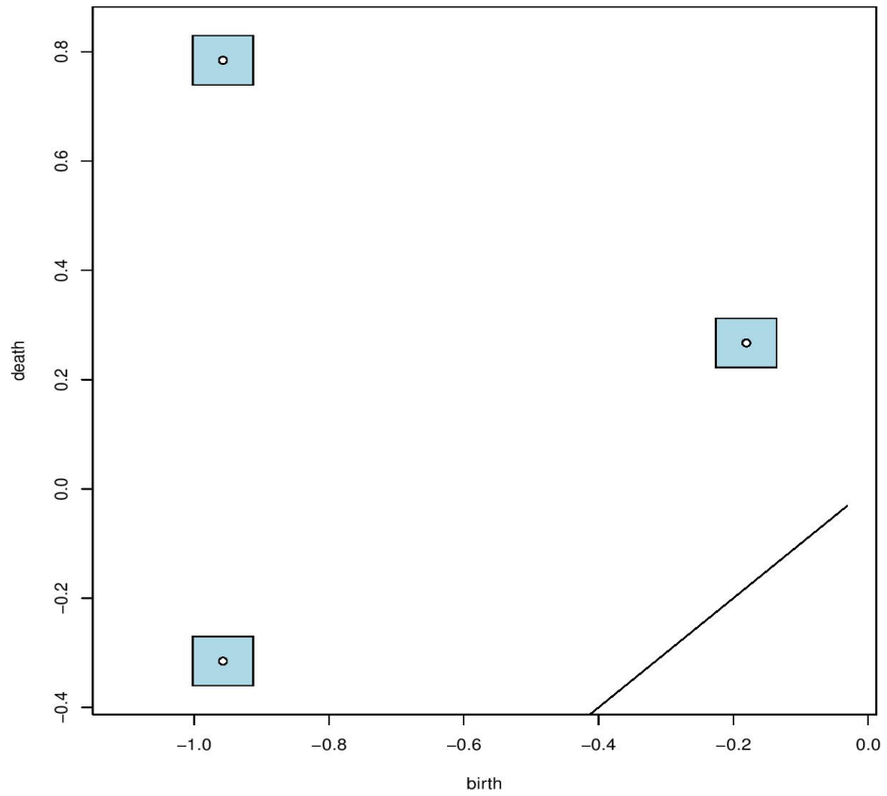
Or bootstrap (only empirical):

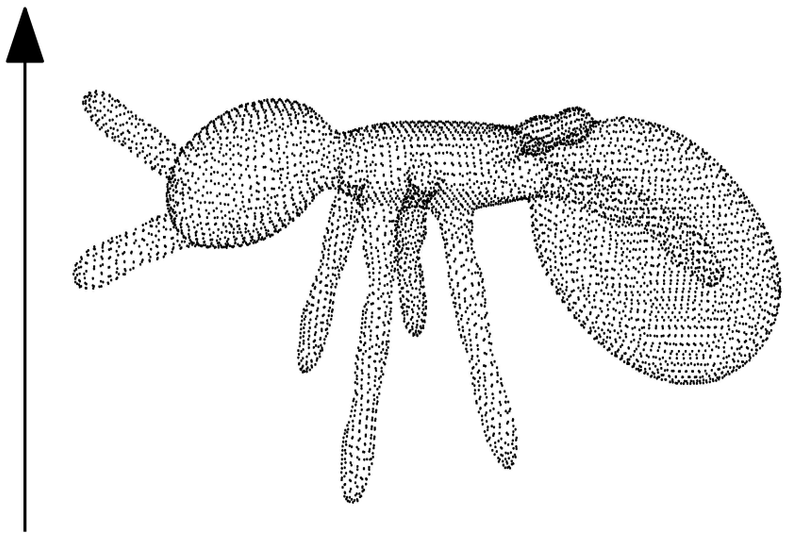
- draw $\widehat{X}_n^* = X_1^*, \cdots, X_n^*$ iid from $\mu_{\widehat{X}_n}$ (empirical measure on $\widehat{X}_n$)
- compute $\mathrm{d}^* = \mathrm{d}_B\left(\mathrm{M}_f(\widehat{X}_n^*), \mathrm{M}_f(\widehat{X}_n)\right)$
- repeat N times to get $\mathrm{d}_1^*, \cdots, \mathrm{d}_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^N \mathbb{I}(\sqrt{n}\,\mathrm{d}_i^* \geq t)$
- take $c_n(\alpha) = \frac{q_\alpha}{\sqrt{n}}$

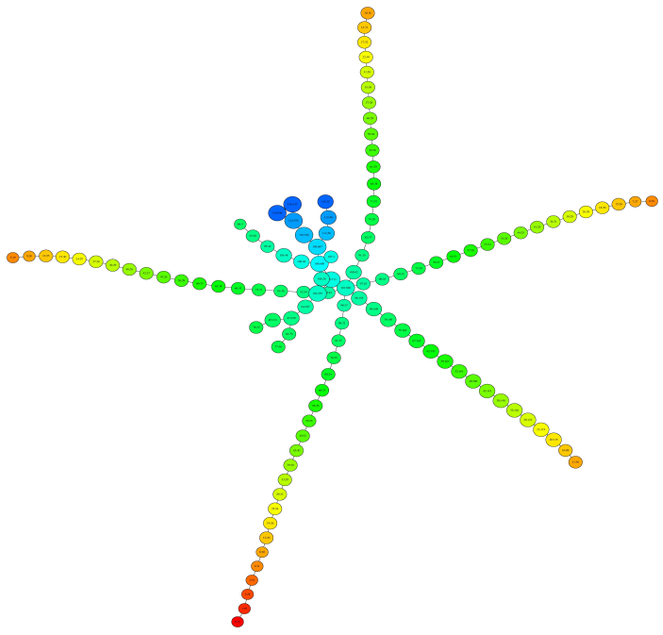
Experiments

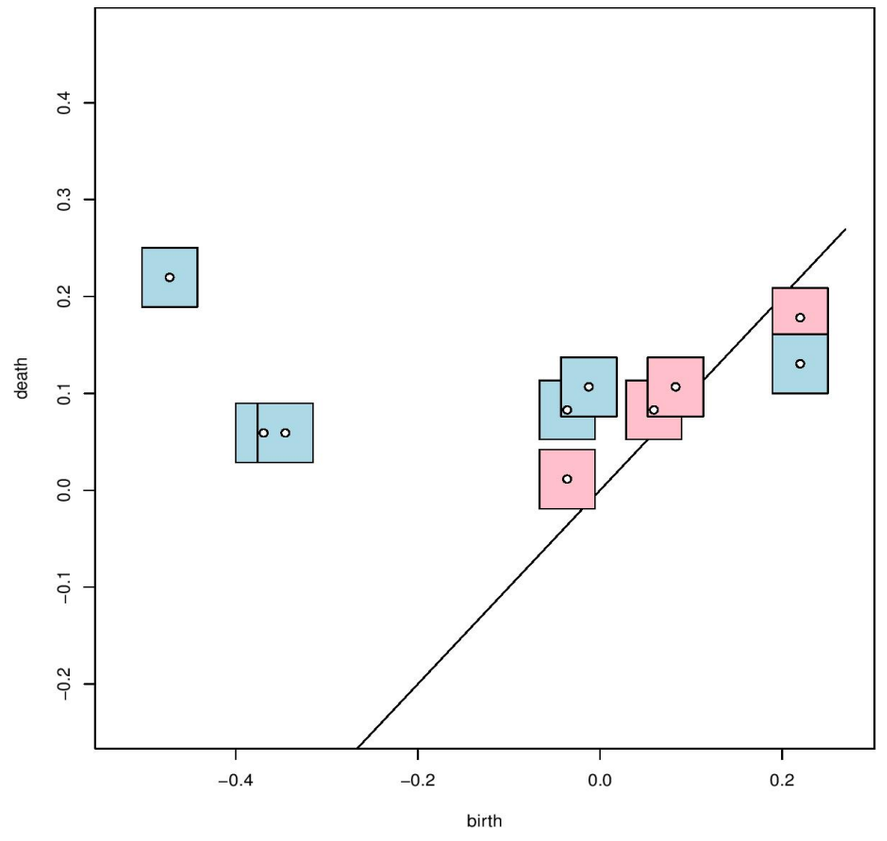




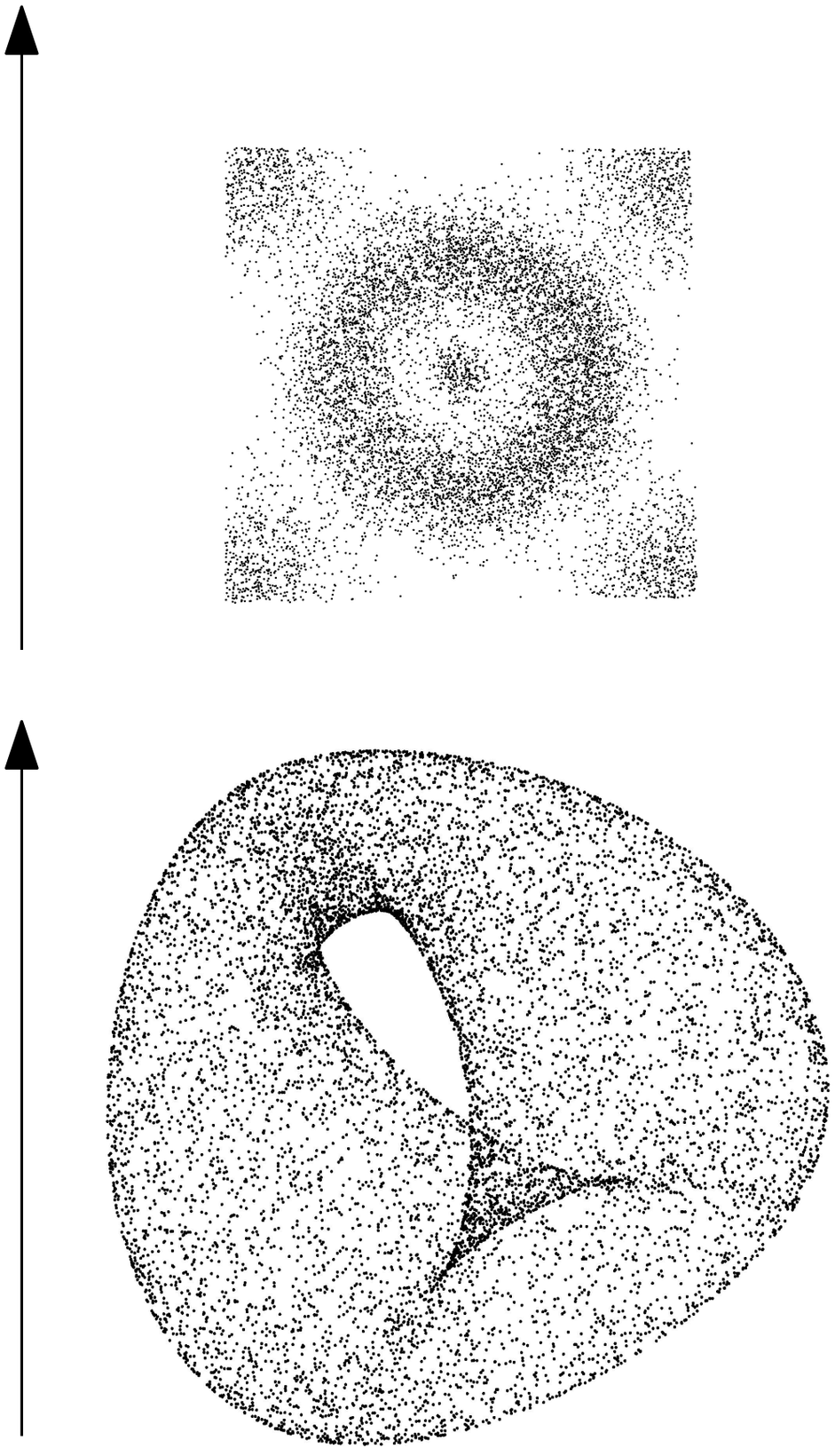


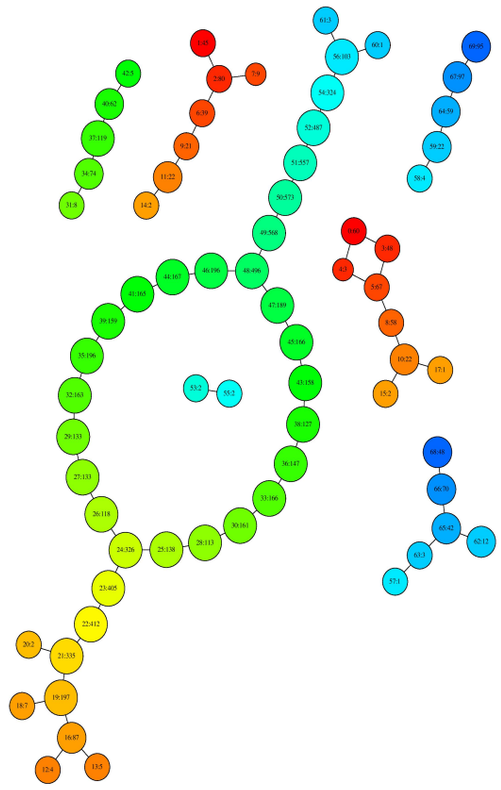


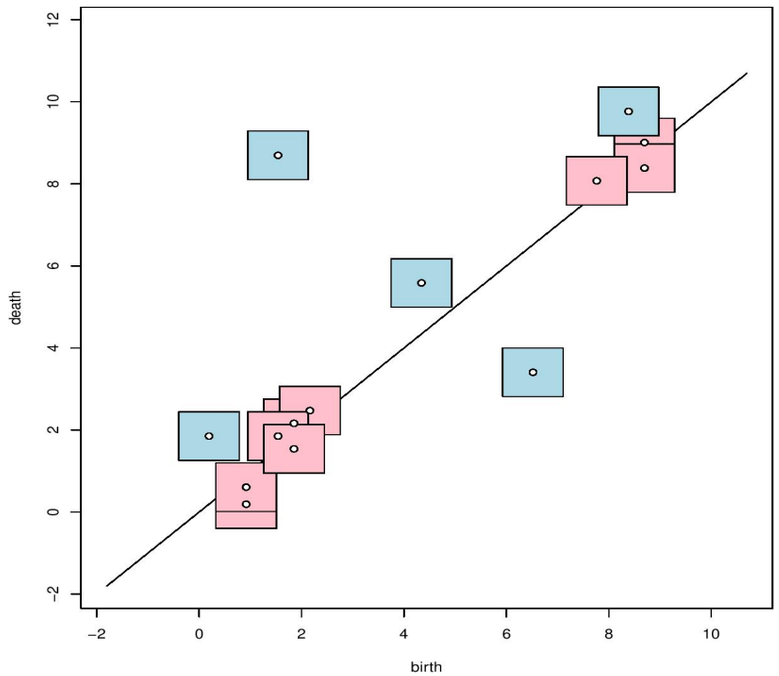


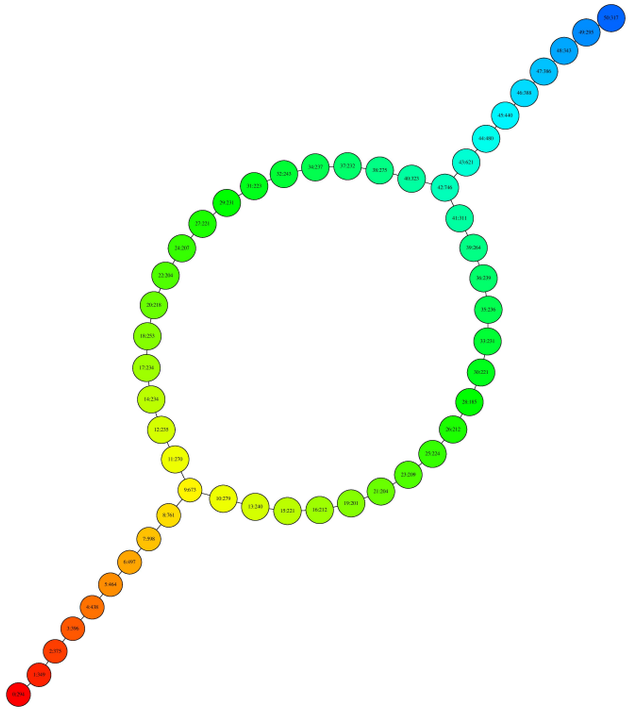


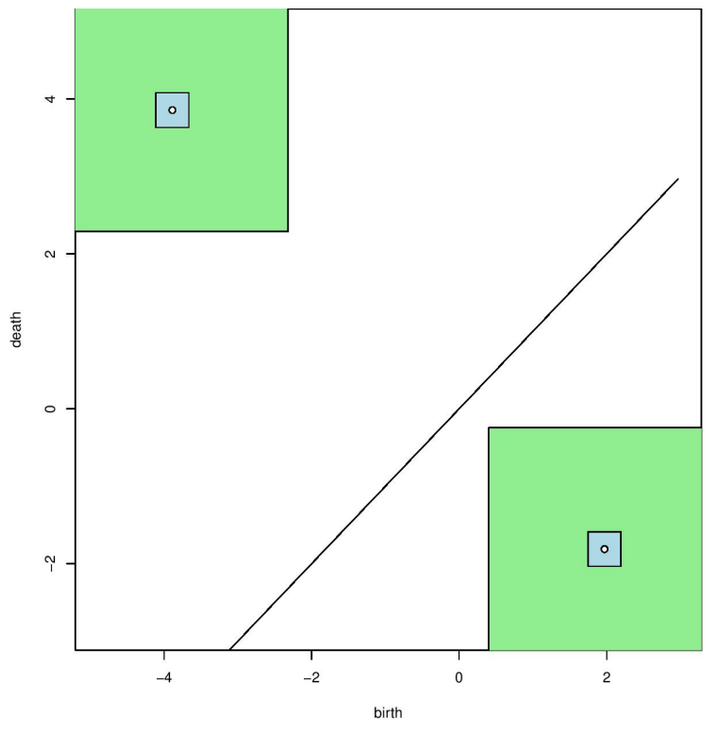
Experiments



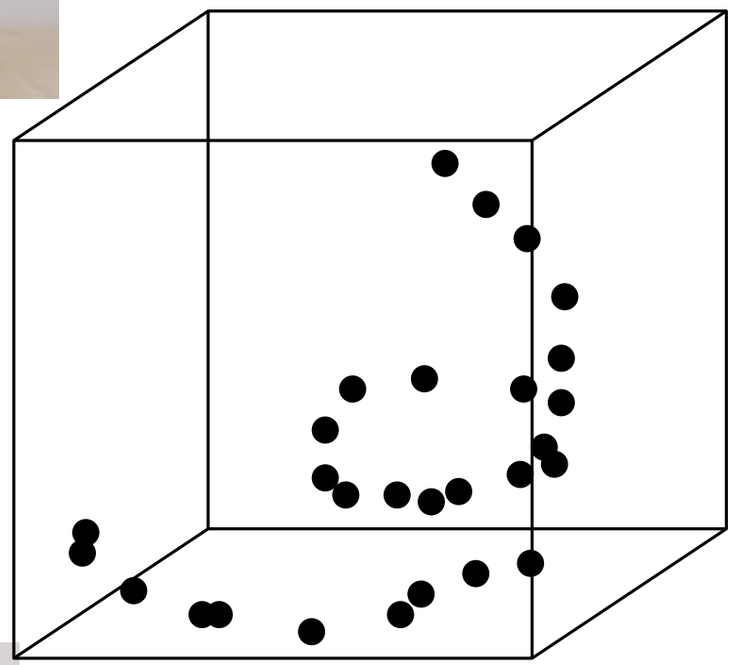


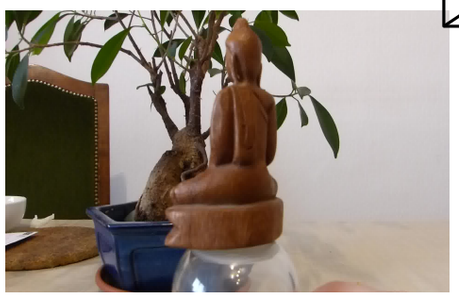




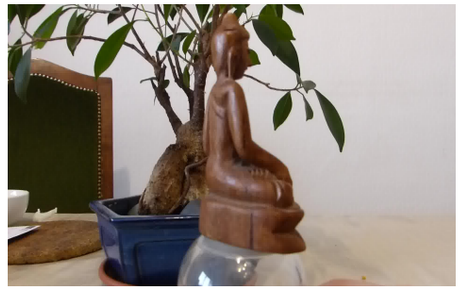


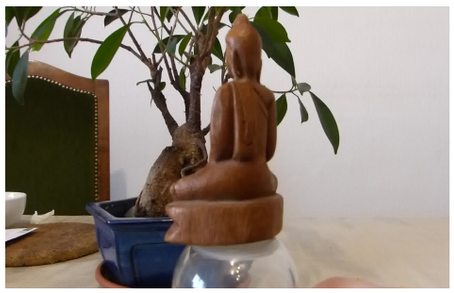
Experiments



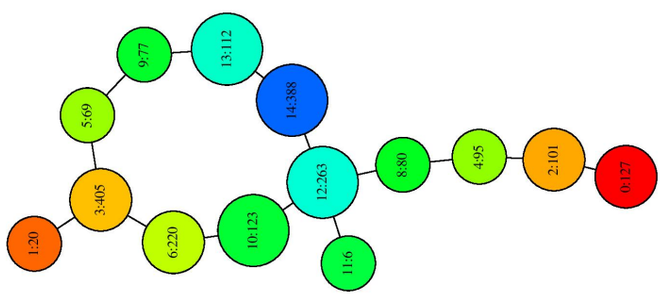


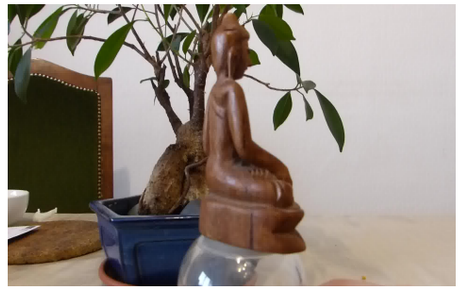
Experiments

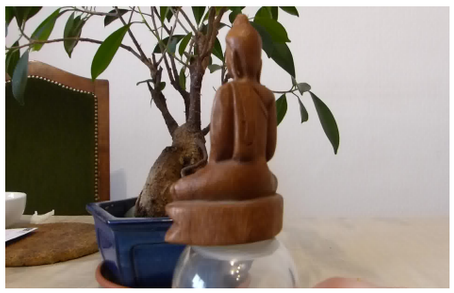


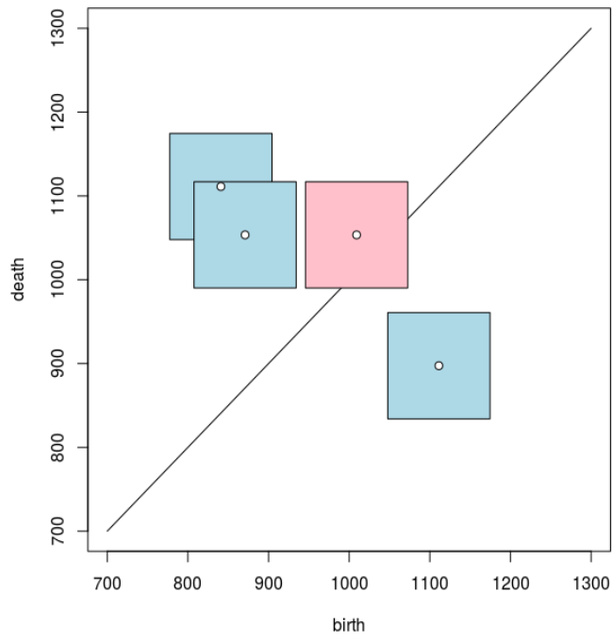


Experiments









Conclusion

Structure and Stability of Mapper

Parameter Selection for Mapper

Extensions:

Multivariate function $f : X \to \mathbb{R}^n$

Space of Reeb graphs (curvature, barycenters, interpolation...)

Conclusion

Structure and Stability of Mapper

Parameter Selection for Mapper

Extensions:

Multivariate function $f : X \to \mathbb{R}^n$

Space of Reeb graphs (curvature, barycenters, interpolation...)

Thank you!