Foundations of Geometric Methods in Data Analysis
2020-2021

Topological exploratory data analysis: nerves, Mappers and robust inference

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

Class outline: taking a step back...

- 1. Nearest neighbors in Euclidean and metric spaces: data structures and algorithms
- 2. Nearest neighbors in Euclidean and metric spaces: analysis
- 3. Topological exploratory data analysis: nerves, Mappers and robust inference
- 4. Topological clustering: algorithms and introduction to persistent homology
- 5. Comparing samplings, distributions, clusterings
- 6. Dimensionality reduction algorithms
- 7. Topological machine learning: descriptors and stability
- 8. Topological optimization: gradient descent and regularization

Class outline: taking a step back...

Classes 3, 4, 7, 8 are about

Topological Data Analysis (TDA)

Class outline: taking a step back...

Classes 3, 4, 7, 8 are about

Topological Data Analysis (TDA)

Goal: Study geometric data sets with techniques coming from *topology*.

Class outline: taking a step back...

Classes 3, 4, 7, 8 are about

Topological Data Analysis (TDA)

Goal: Study geometric data sets with techniques coming from *topology*.

Question: What is topology?

Class outline: taking a step back...

Classes 3, 4, 7, 8 are about

Topological Data Analysis (TDA)

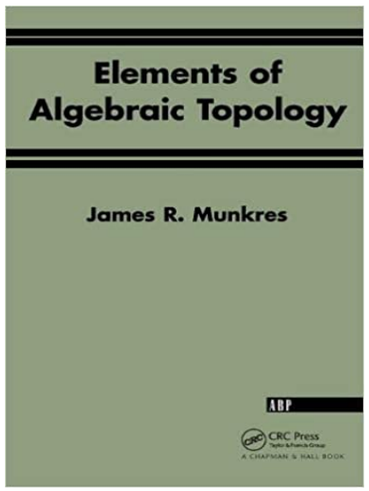
Goal: Study geometric data sets with techniques coming from *topology*.

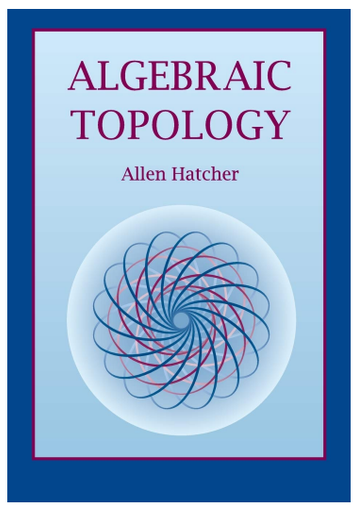
Question: What is topology?

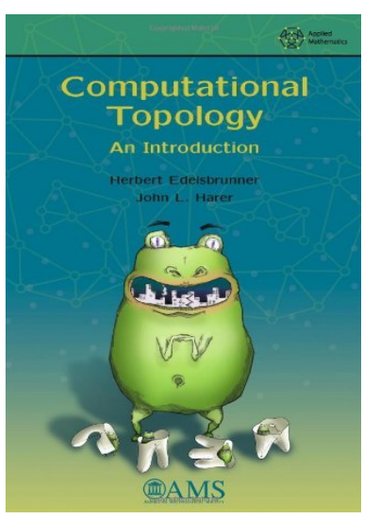
[*Elements of Algebraic Topology*, Munkres, CRC Press, 1984]

[*Algebraic Topology*, Hatcher, Cambridge University Press, 2002]

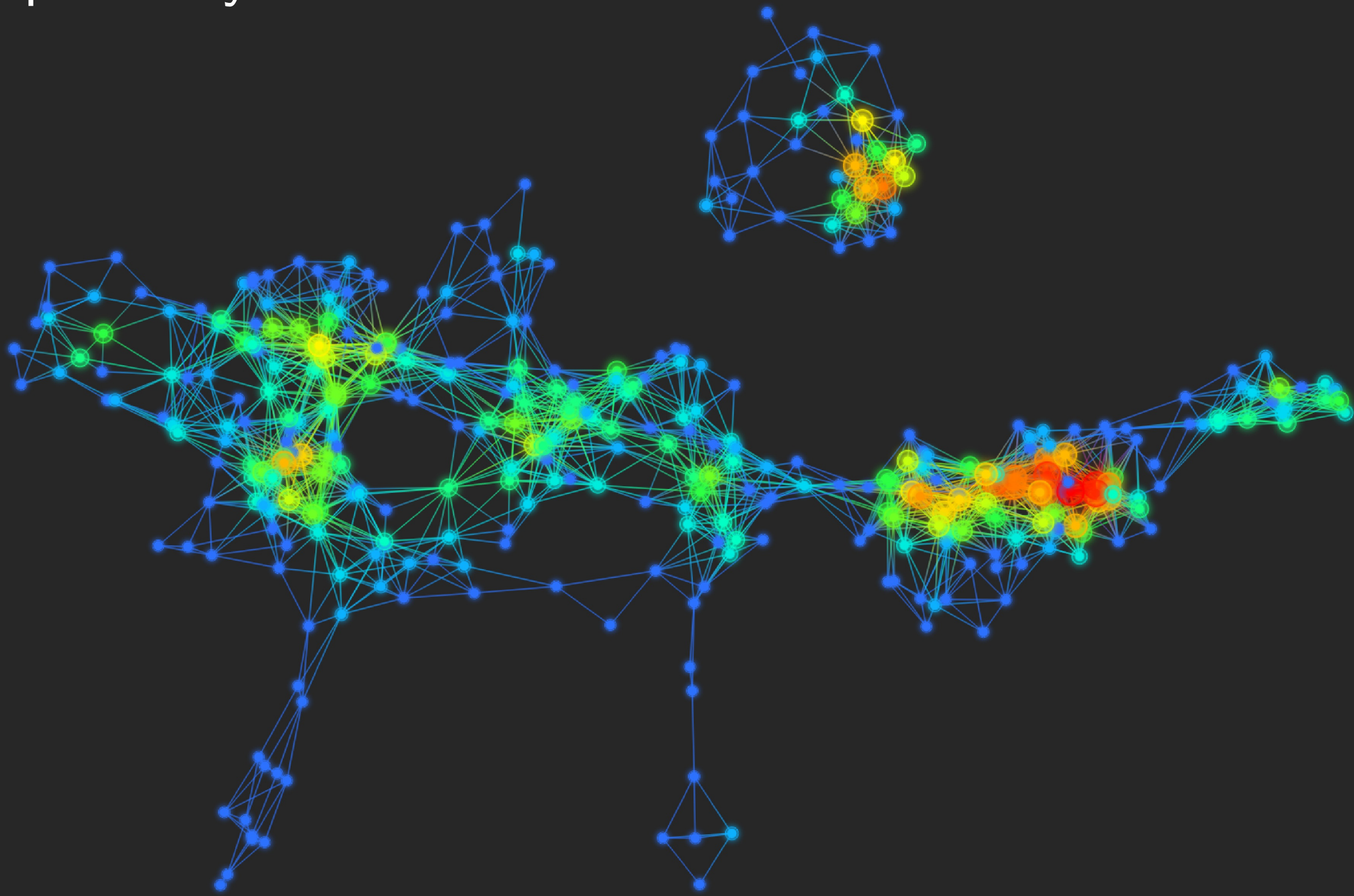
[*Computational Topology: an introduction*, Edelsbrunner, Harer, AMS, 2010]



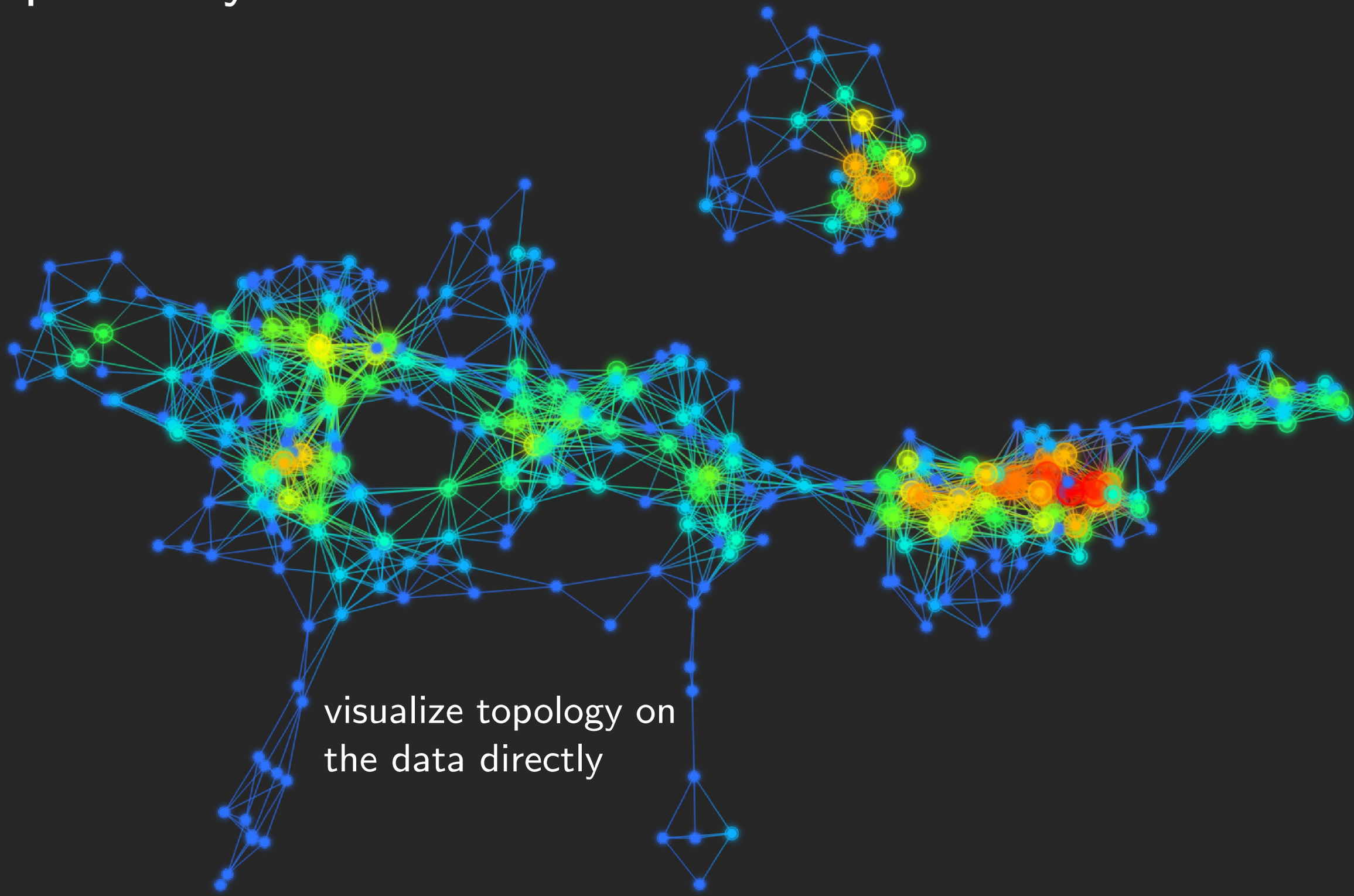




Exploratory TDA



Exploratory TDA



visualize topology on
the data directly

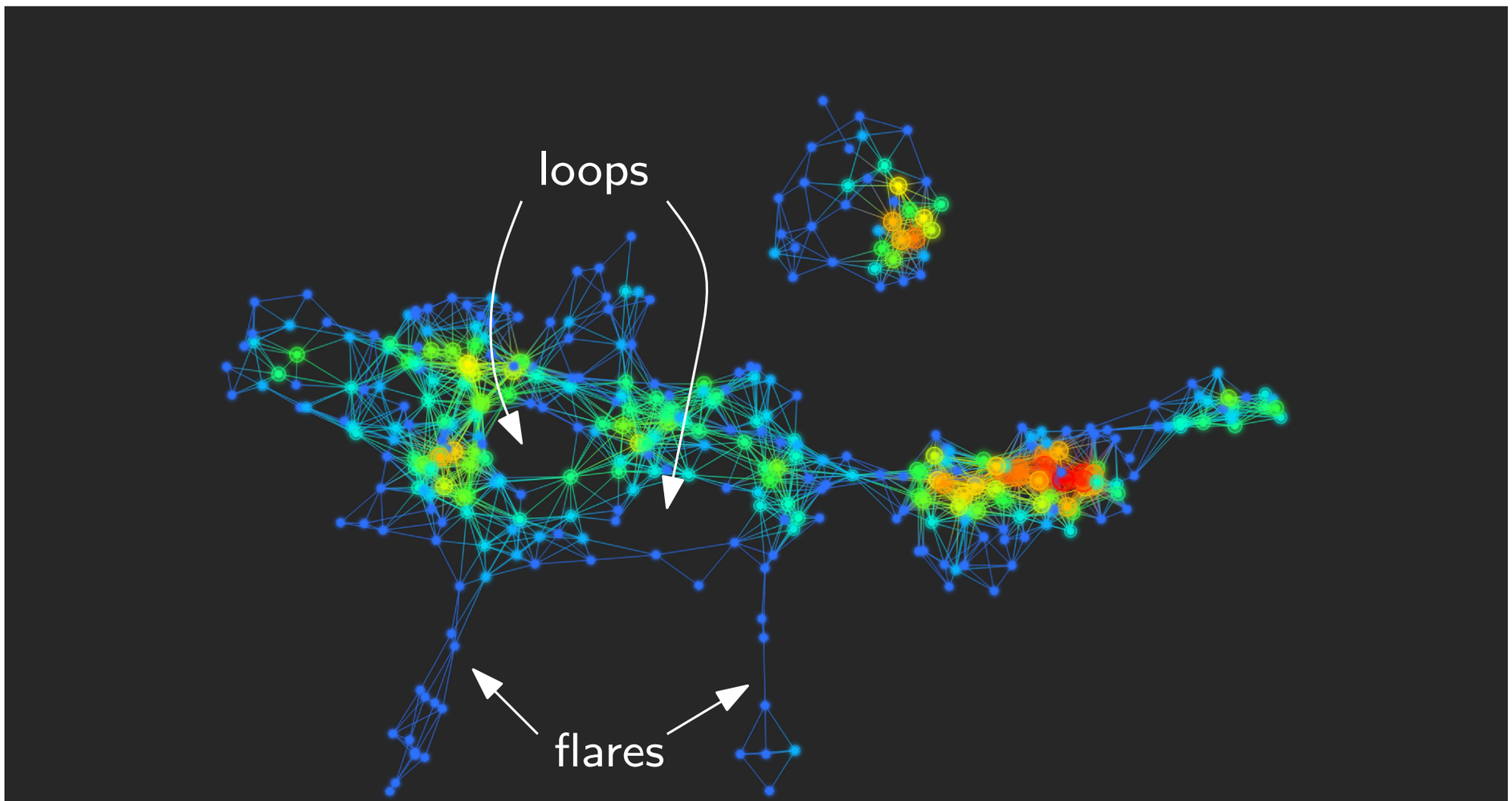
Applications of exploratory TDA

Two types of applications:

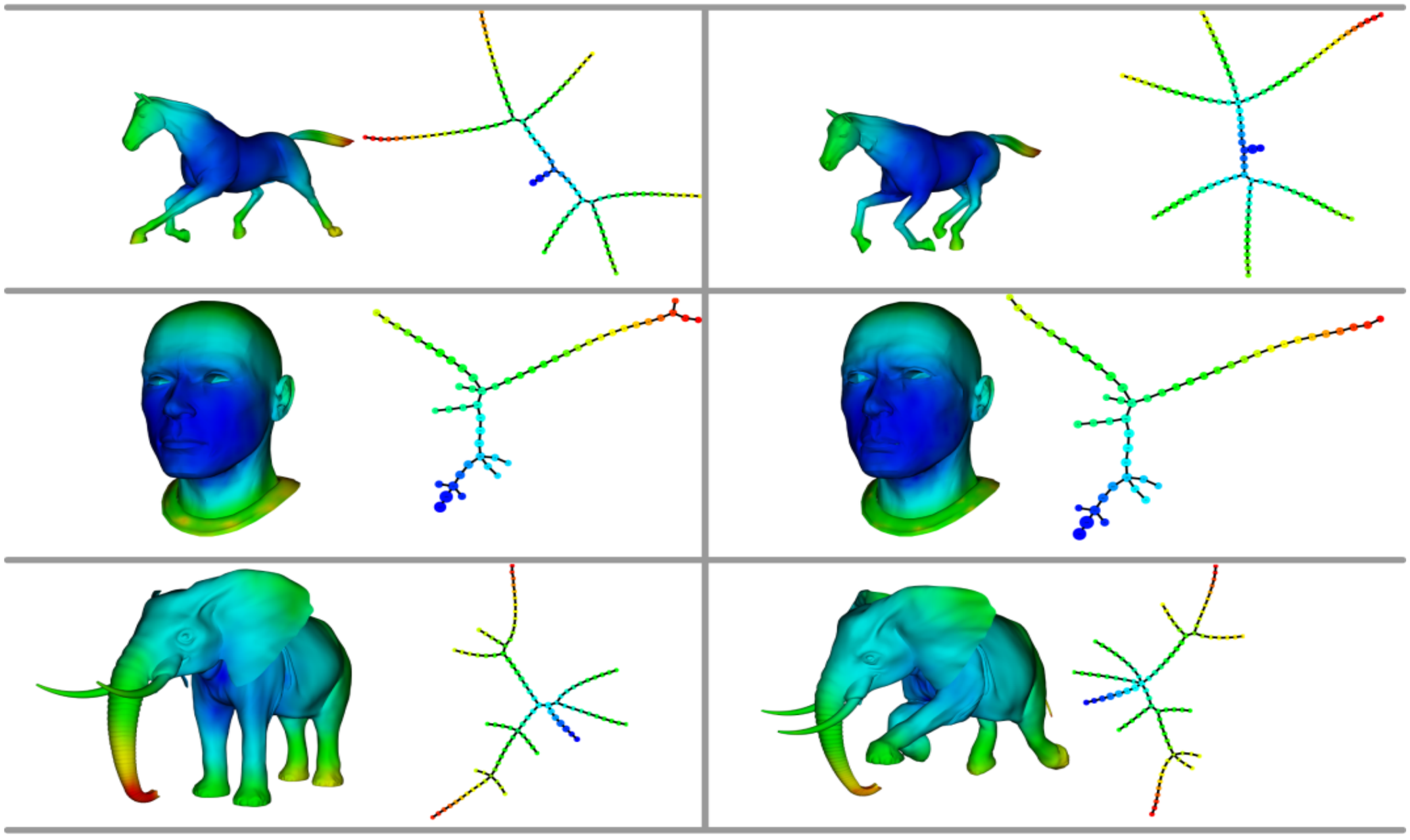
- clustering
- feature selection

Principle: identify statistically relevant subpopulations through topological patterns (flares, loops).

(usually with Kolmogorov-Smirnov 2-sample tests—see class 5)



Applications of exploratory TDA



3d shapes classification

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

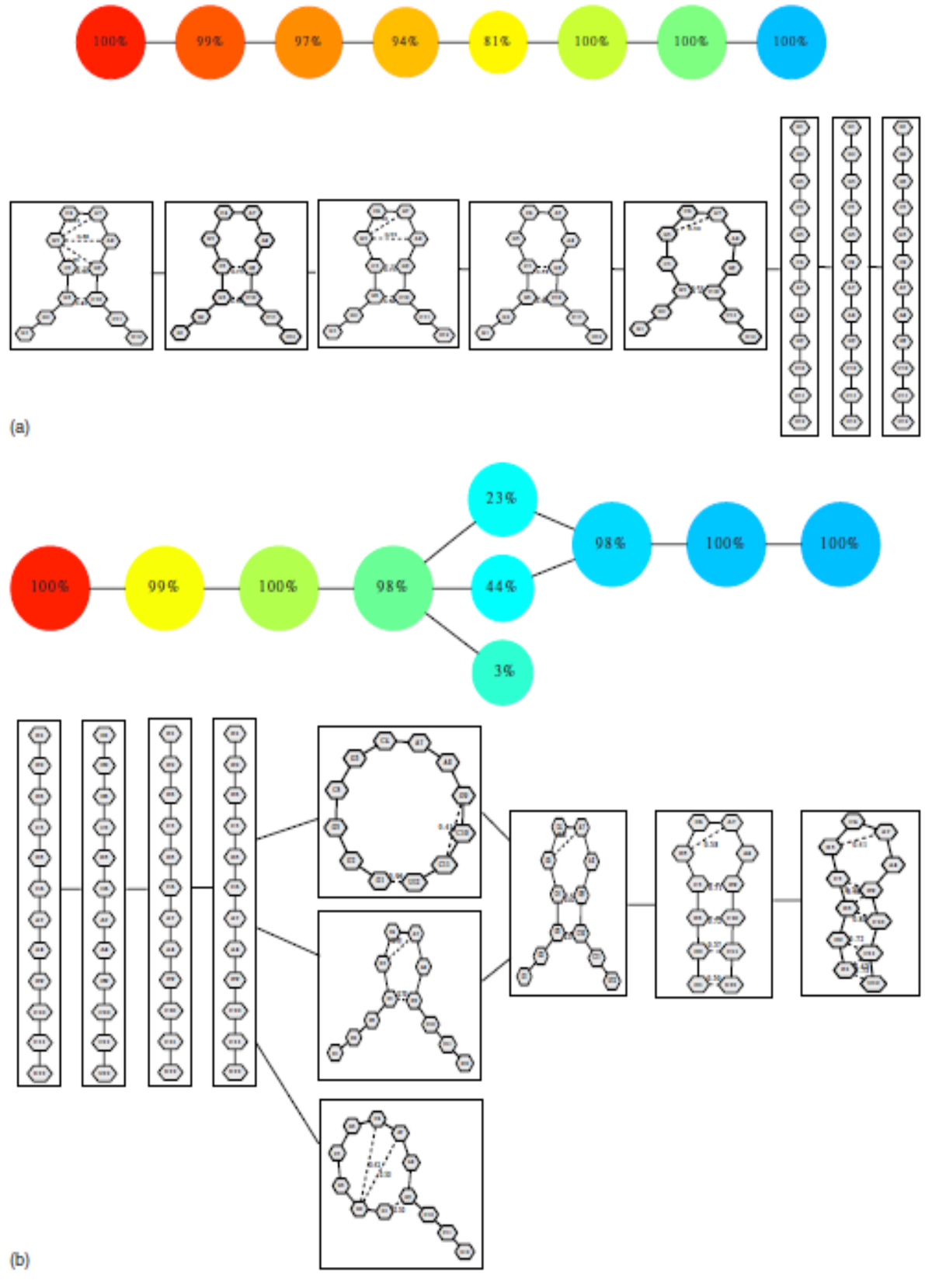
Applications of exploratory TDA

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

Data: conformations of molecules.

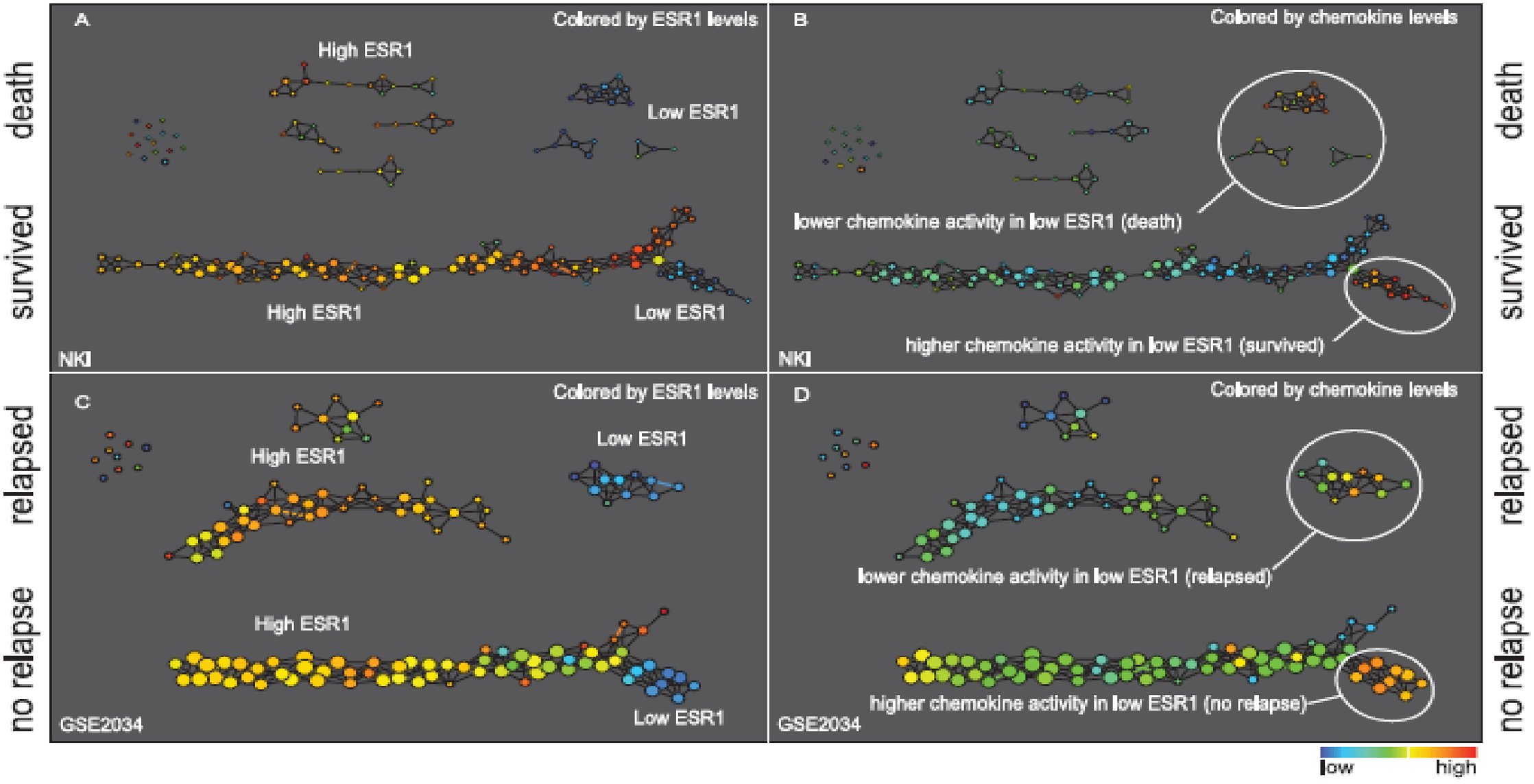
Goal: detect folding pathways.

Idea: 1 loop = 2 pathways.



Applications of exploratory TDA

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



Data: breast cancer patients that went through specific therapy.

Goal: detect variables that influence survival after therapy in breast cancer.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

Q: What is the most basic brick (space) topology can work on?

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

Q: What is the most basic brick (space) topology can work on?

A: The so-called *topological spaces*.

Def: A *topological space* is a set X equipped with a *topology*, i.e., a family $\mathcal{O}$ of subsets of X, called the *open sets* of X, such that:
(i) the empty set $\emptyset$ and X are elements of $\mathcal{O}$,
(ii) any union of elements of $\mathcal{O}$ is an element of $\mathcal{O}$,
(iii) any finite intersection of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

Open sets are the tools that allow to define *continuity*, which is the primary notion that allow to compare spaces in topology.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

Q: What is the most basic brick (space) topology can work on?

A: The so-called *topological spaces*.

Def: A *topological space* is a set X equipped with a *topology*, i.e., a family $\mathcal{O}$ of subsets of X, called the *open sets* of X, such that:
(i) the empty set $\emptyset$ and X are elements of $\mathcal{O}$,
(ii) any union of elements of $\mathcal{O}$ is an element of $\mathcal{O}$,
(iii) any finite intersection of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

Open sets are the tools that allow to define *continuity*, which is the primary notion that allow to compare spaces in topology.

Def: a map $f : X \to Y$ is *continuous* if and only if the pre-image $f^{-1}(O_Y) = \{x \in X : f(x) \in O_Y\}$ of any open set $O_Y \subseteq Y$ is an open set of X.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

A very common family of topological spaces is comprised of the *metric spaces*.

Def: A metric (or distance) on X is a map $d : X \times X \to [0, +\infty)$ such that:
(i) for any $x, y \in X$, $d(x, y) = d(y, x)$,
(ii) for any $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$,
(iii) for any $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.
The set X together with d is a metric space.

The smallest topology containing all the open balls $B(x, r) = \{y \in X : d(x, y) < r\}$ is called the metric topology on X induced by d.

Ex: the standard topology in an Euclidean space is the one induced by the metric defined by the norm: $d(x, y) = \|x - y\|$.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Def: Here are the main comparison tools of topology:

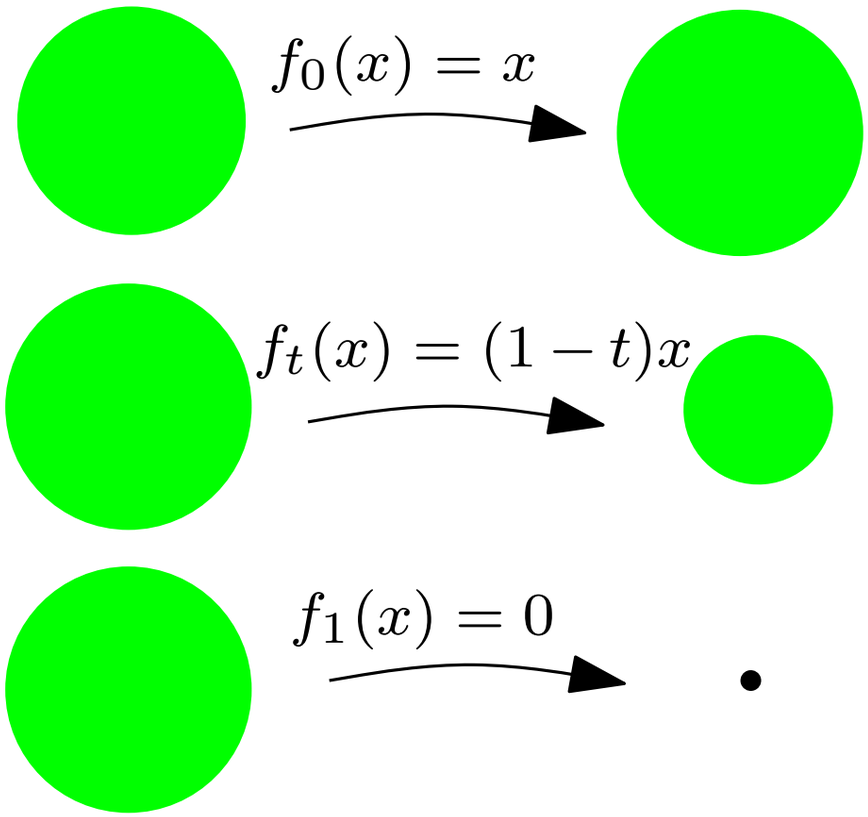
- Two maps $f_0 : X \to Y$ and $f_1 : X \to Y$ are *homotopic* if $\exists$ a continuous map $F : [0,1] \times X \to Y$ s.t. $\forall x \in X$, $F(0,x) = f_0(x)$ and $F_1(1,x) = f_1(x)$. X and Y are homotopy equivalent if $\exists$ continuous maps $f : X \to Y$ and $g : Y \to X$ s.t. $g \circ f$ is homotopic to id_X and $f \circ g$ is homotopic to id_Y.
- X and Y are homeomorphic if $\exists$ a bijection (homeomorphism) $h : X \to Y$ s.t. h and h^{-1} are continuous.
- X and Y are isotopic if $\exists$ a continuous map (isotopy) $F : X \times [0,1] \to Y$ s.t. $F(.,0) = \mathrm{id}_X$, $F(X,1) = Y$ and $\forall t \in [0,1]$, $F(.,t)$ is an homeomorphism.

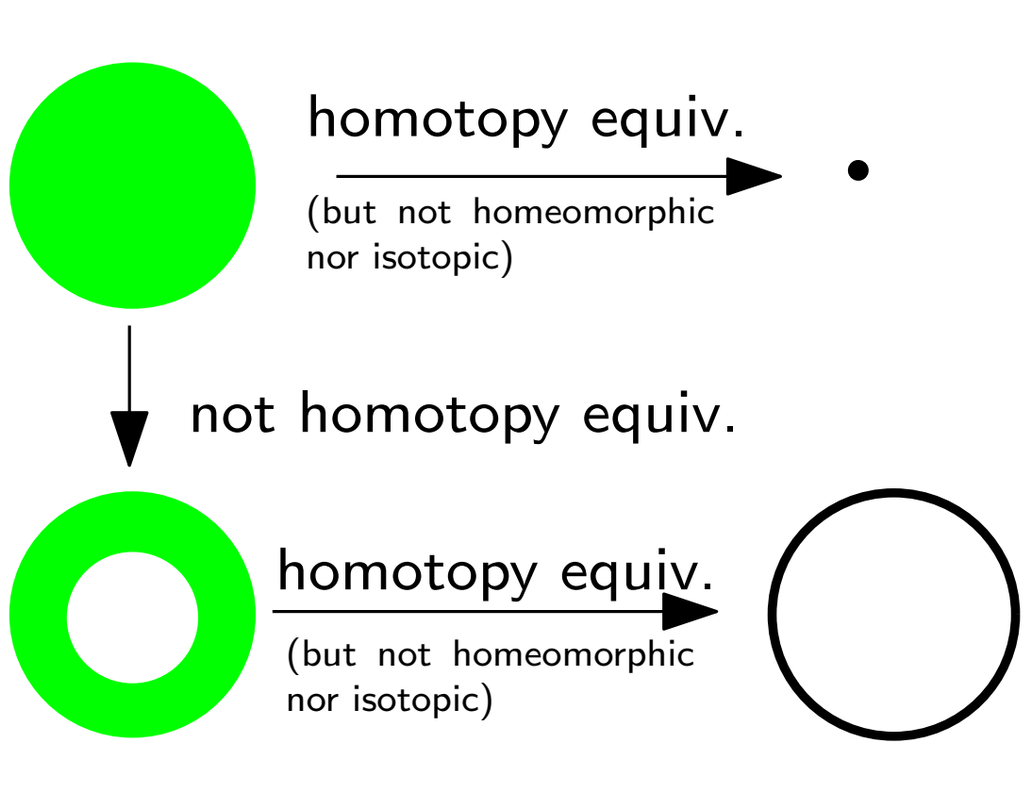
Q: Which notion is stronger/weaker?

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.





A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Previous examples are particular homotopy equivalences called *deformation retracts*.

Def: If $Y \subseteq X$ and if there exists a continuous map $F : [0,1] \times X \to X$ s.t.:
(i) $\forall x \in X,\ F(0,x) = x$
(ii) $\forall x \in X,\ F(1,x) \in Y$
(iii) $\forall y \in Y,\ \forall t \in [0,1],\ F(t,y) \in Y$
then X and Y are homotopy equivalent. If one replaces condition *(iii)* by $\forall y \in Y,\ \forall t \in [0,1],\ H(t,y) = y$ then H is a deformation retract of X onto Y.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Q: Can you find two spaces that are homeomorphic but not isotopic?

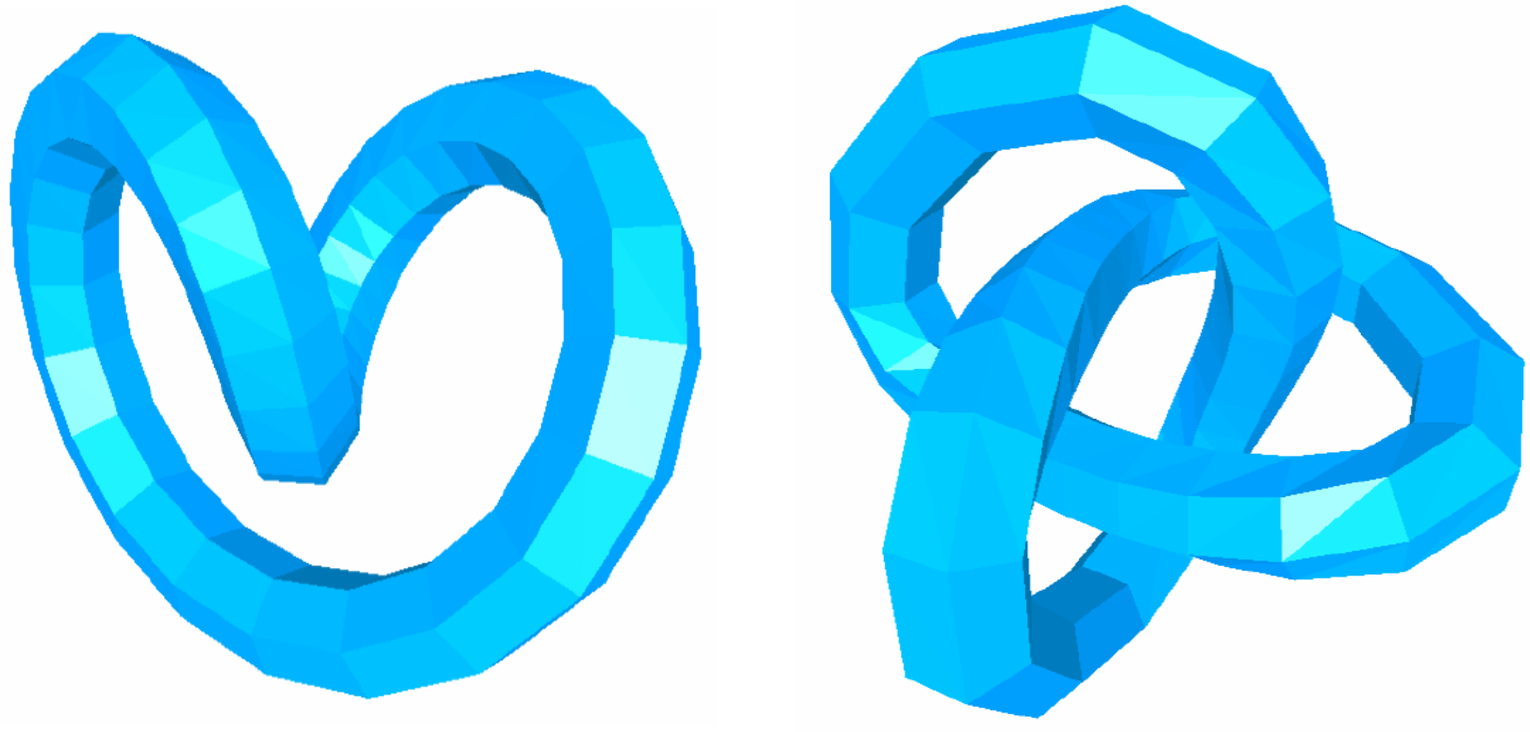
A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Q: Can you find two spaces that are homeomorphic but not isotopic?

A: Torus and trefoil knot.



A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Q: Can you find an isotopy between these guys?

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Pb 1: How to encode topological spaces for computational purposes?

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces*.

In topology, two spaces are the same (i.e., belong to the same class) if one 'continuously deforms' onto the other.

Pb 1: How to encode topological spaces for computational purposes?

Pb 2: Looking for homotopy equivalences/homeomorphisms/isotopies is extremely difficult. Are there mathematical quantities that are invariant to homotopy equivalences **and** easy to compute?

A topological space fit for computation

Pb 1: How to encode topological spaces for computational purposes?

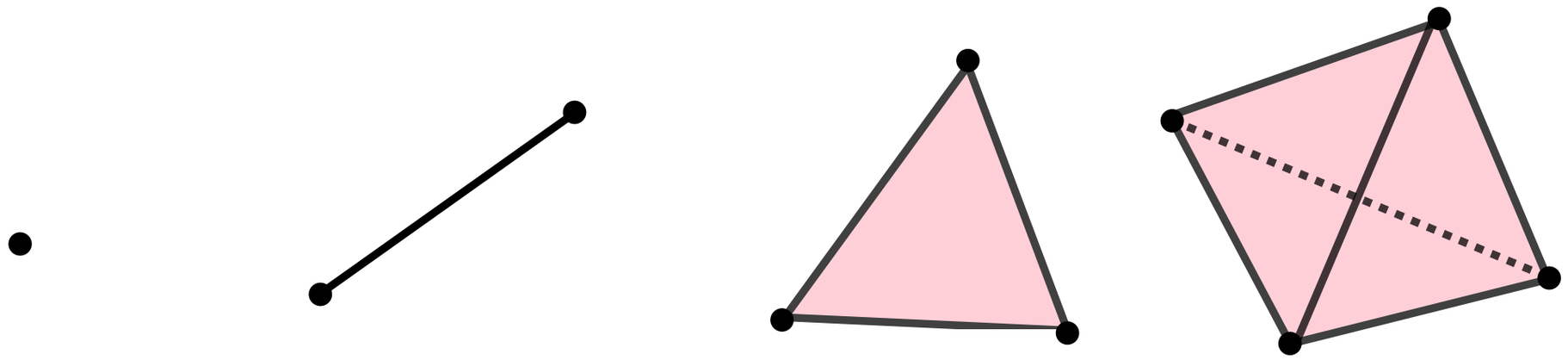
A topological space fit for computation

Pb 1: How to encode topological spaces for computational purposes?

A: Using spaces made of small convex bricks, namely the *simplicial complexes* made of *simplices*.

Simplex and simplicial complex

Simplex and simplicial complex



0-simplex:
vertex

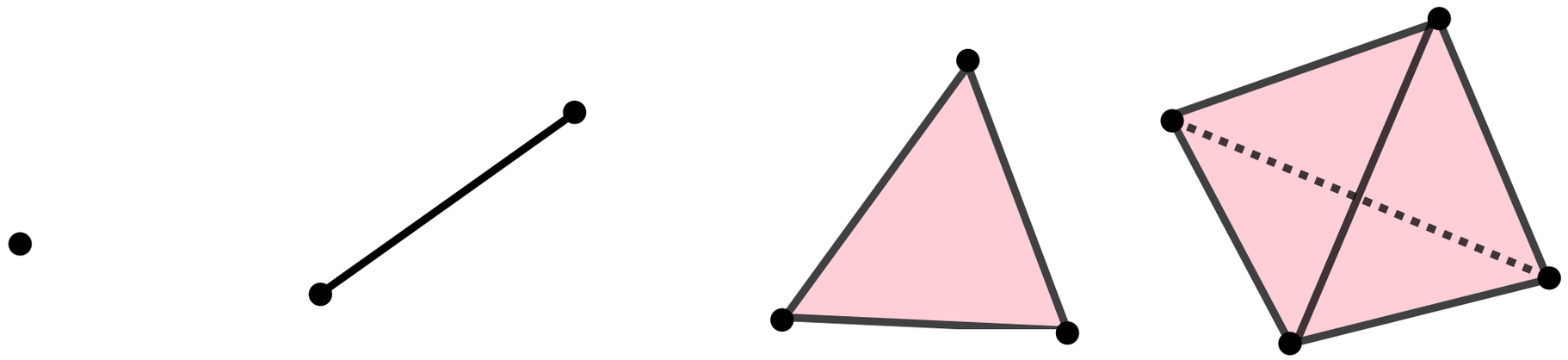
1-simplex:
edge

2-simplex:
triangle

3-simplex:
tetrahedron

etc...

Simplex and simplicial complex



0-simplex: vertex

1-simplex: edge

2-simplex: triangle

3-simplex: tetrahedron

etc...

Def: Given a set $P = \{p_0, \ldots, p_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the k-dimensional simplex σ (or k-simplex for short) spanned by P is the set of convex combinations

$$\sum_{i=0}^{k} \lambda_i \, p_i, \quad \text{with} \quad \sum_{i=0}^{k} \lambda_i = 1 \quad \text{and} \quad \lambda_i \geq 0.$$

The points $p_0, \ldots, p_k$ are called the vertices of σ.

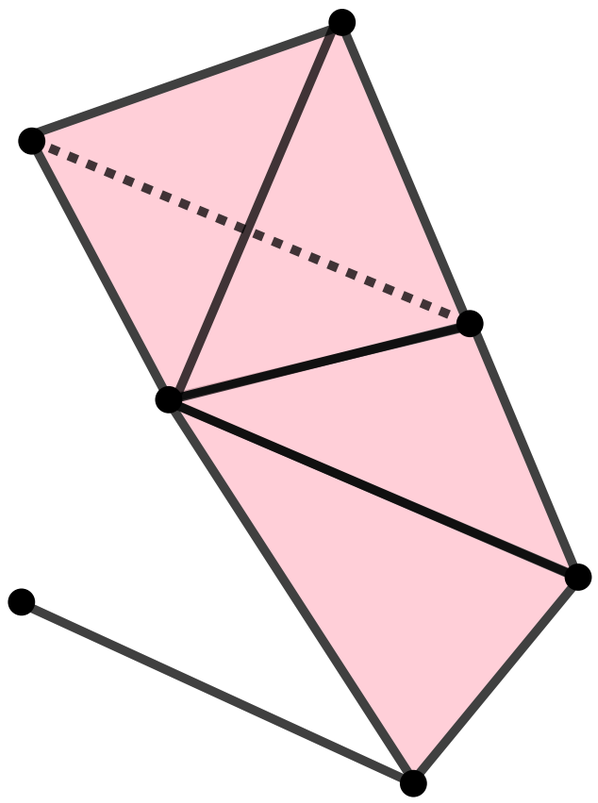
Simplex and simplicial complex

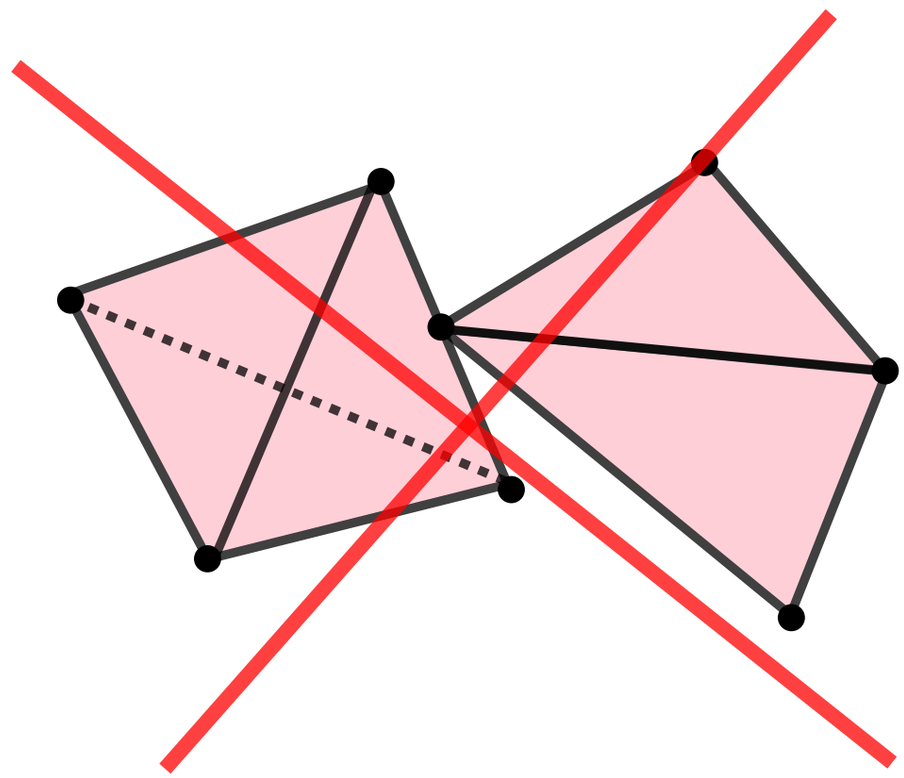
Def: A simplicial complex K in $\mathbb{R}^d$ is a collection of simplices s.t.:

- *(i)* any face of a simplex of K is a simplex of K,
- *(ii)* the intersection of any two simplices of K is either empty or a common face of both.

The underlying space of K (written $|K| \subseteq \mathbb{R}^d$) is the union of its simplices.

Simplex and simplicial complex





Def: A simplicial complex K in $\mathbb{R}^d$ is a collection of simplices s.t.:

- *(i)* any face of a simplex of K is a simplex of K,
- *(ii)* the intersection of any two simplices of K is either empty or a common face of both.

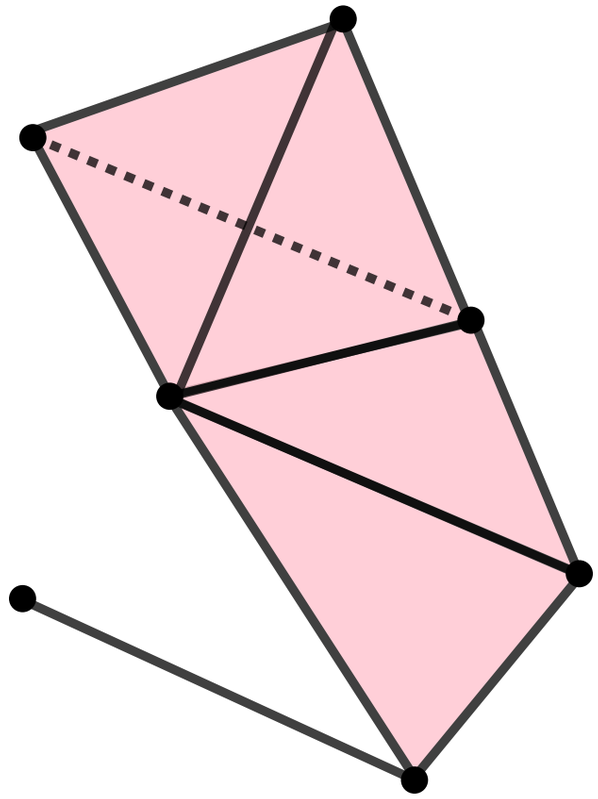
The underlying space of K (written $|K| \subseteq \mathbb{R}^d$) is the union of its simplices.

Abstract simplex and simplicial complex

Def: Let $P = \{p_1, \cdots p_n\}$ be a (finite) set. An abstract simplicial complex K with vertex set P is a set of subsets of P satisfying the two conditions:

- *(i)* the elements of P belong to K,
- *(ii)* if $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$.

The elements of K are the simplices.

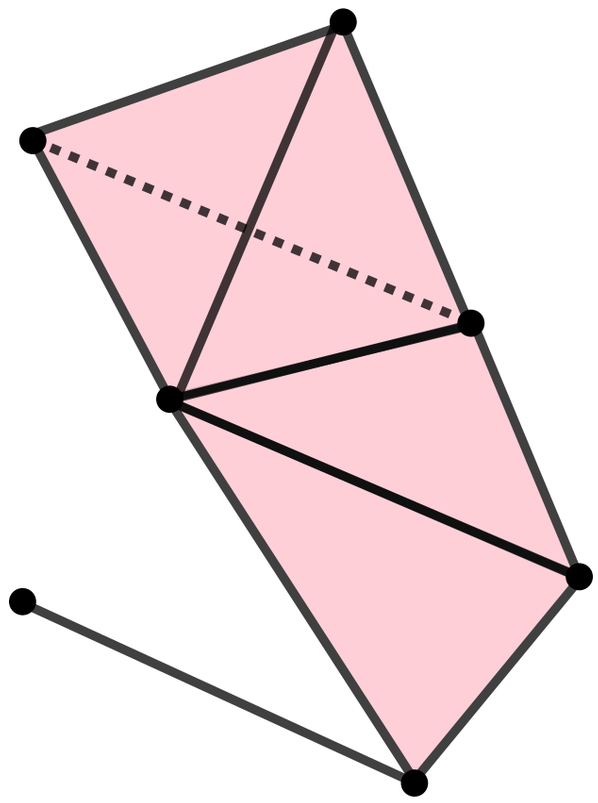


Abstract simplex and simplicial complex

Def: Let $P = \{p_1, \cdots p_n\}$ be a (finite) set. An abstract simplicial complex K with vertex set P is a set of subsets of P satisfying the two conditions:

- *(i)* the elements of P belong to K,
- *(ii)* if $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$.

The elements of K are the simplices.



IMPORTANT

Simplicial complexes can be seen at the same time as geometric/topological spaces (good for topological/geometrical inference) and as combinatorial objects (abstract simplicial complexes, good for computations).

An invariant fit for computation

Pb 2: Looking for homotopy equivalences/homeomorphisms/isotopies is extremely difficult. Are there mathematical quantities that are invariant to homotopy equivalences **and** easy to compute?

An invariant fit for computation

Pb 2: Looking for homotopy equivalences/homeomorphisms/isotopies is extremely difficult. Are there mathematical quantities that are invariant to homotopy equivalences **and** easy to compute?

A: The *holes*, encoded in the *homology groups* H_k, $k \in \mathbb{N}$

The homology groups

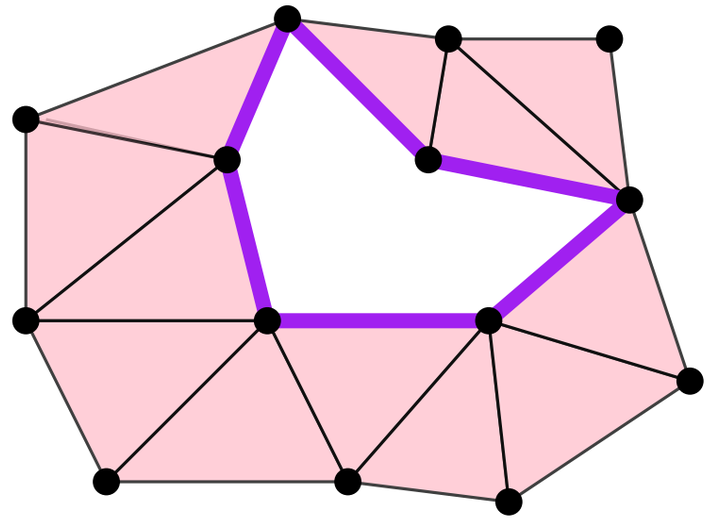
The homology groups

Q: How to characterize a hole in a simplicial complex?

The homology groups

Q: How to characterize a hole in a simplicial complex?

A: A hole (in 1D) is a path whose first and end points are the same, a loop.

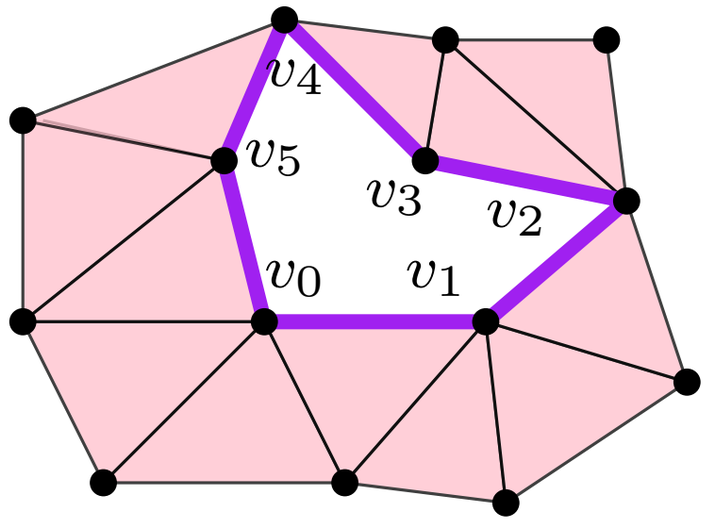


The homology groups

Q: How to characterize a hole in a simplicial complex?

A: A hole (in 1D) is a path whose first and end points are the same, a loop.

The sequence of 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$, $[v_4, v_5]$, $[v_5, v_0]$ is a hole

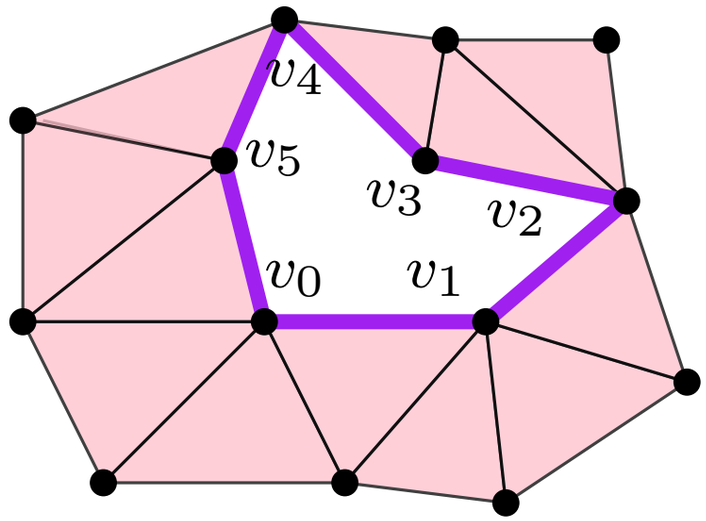


The homology groups

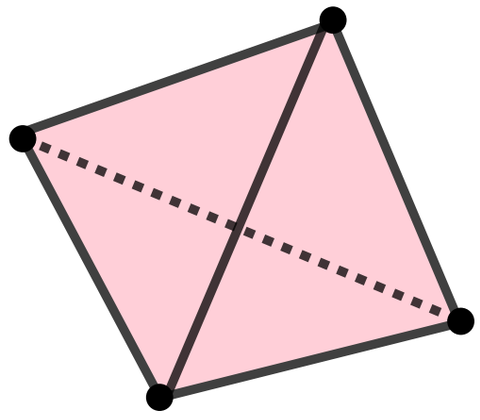
Q: How to characterize a hole in a simplicial complex?

A: A hole (in 1D) is a path whose first and end points are the same, a loop.

The sequence of 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$, $[v_4, v_5]$, $[v_5, v_0]$ is a hole



But what about higher dimensional holes (like the inside of a tetrahedron)?

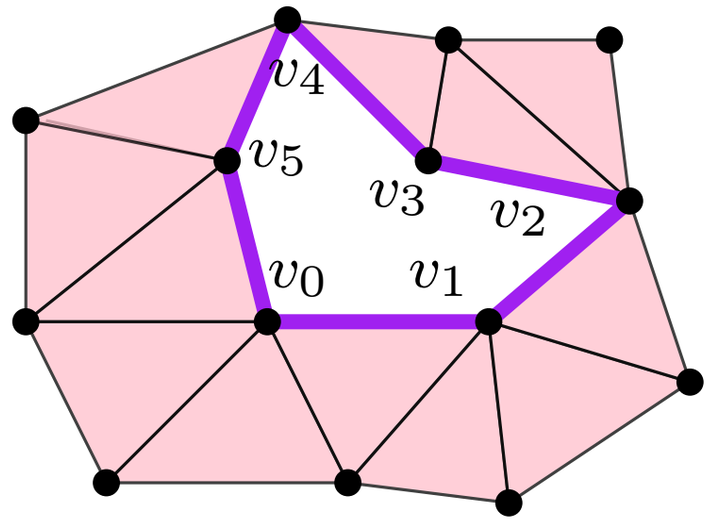


The homology groups

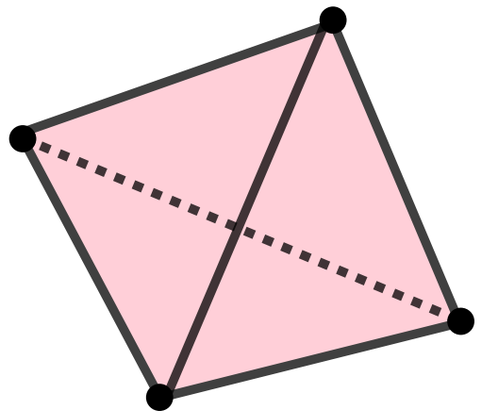
Q: How to characterize a hole in a simplicial complex?

A: A hole (in 1D) is a path whose first and end points are the same, a loop.

The sequence of 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$, $[v_4, v_5]$, $[v_5, v_0]$ is a hole



But what about higher dimensional holes (like the inside of a tetrahedron)?



A: A hole in dimension d is a simplicial complex in which each $(d-1)$-simplex appears an even number of times.

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

Def: The *boundary* of a d-simplex is the chain made of its $(d-1)$-simplices.

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

Def: The *boundary* of a d-simplex is the chain made of its $(d-1)$-simplices.

$$\partial_n [v_1, \ldots, v_{n+1}] = \sum_{i=1}^{n+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}]$$

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

Def: The *boundary* of a d-simplex is the chain made of its $(d-1)$-simplices.

$$\partial_n [v_1, \ldots, v_{n+1}] = \sum_{i=1}^{n+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

Def: The *boundary* of a d-simplex is the chain made of its $(d-1)$-simplices.

$$\partial_n [v_1, \ldots, v_{n+1}] = \sum_{i=1}^{n+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

$$= [v_0] + [v_1] + [v_1] + [v_2] + [v_2] + [v_3] + [v_3] + [v_4] + [v_4] + [v_5] + [v_5] + [v_0]$$

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

Def: The *boundary* of a d-simplex is the chain made of its $(d-1)$-simplices.

$$\partial_n [v_1, \ldots, v_{n+1}] = \sum_{i=1}^{n+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

$$= [v_0] + [v_1] + [v_1] + [v_2] + [v_2] + [v_3] + [v_3] + [v_4] + [v_4] + [v_5] + [v_5] + [v_0]$$

$$= [v_0] + [v_0] = 0.$$

Def: A *d-cycle* is a d-chain C s.t. $\partial C = 0$.

The homology groups

Def: A *d-chain* is a formal sum of d-simplices with coefficients in $\mathbb{Z}/2\mathbb{Z}$.

$$C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0].$$

Def: The *boundary* of a d-simplex is the chain made of its $(d-1)$-simplices.

$$\partial_n [v_1, \ldots, v_{n+1}] = \sum_{i=1}^{n+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}]$$

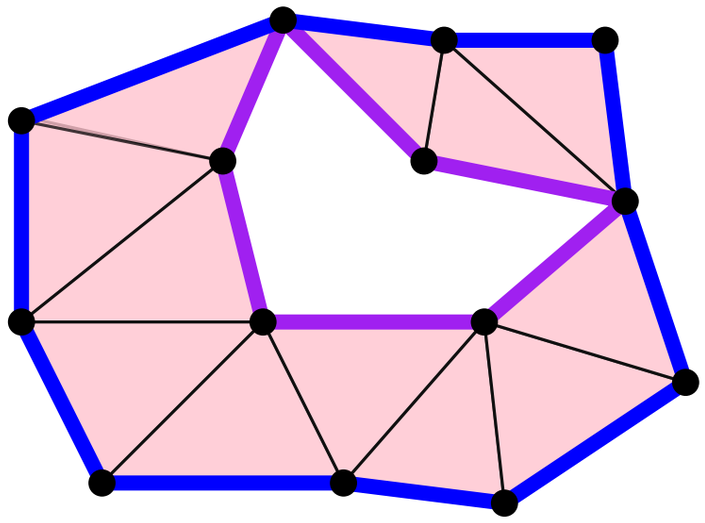
$$\partial_1 C = \partial_1[v_0, v_1] + \partial_1[v_1, v_2] + \partial_1[v_2, v_3] + \partial_1[v_3, v_4] + \partial_1[v_4, v_5] + \partial_1[v_5, v_0]$$

$$= [v_0] + [v_1] + [v_1] + [v_2] + [v_2] + [v_3] + [v_3] + [v_4] + [v_4] + [v_5] + [v_5] + [v_0]$$

$$= [v_0] + [v_0] = 0.$$

Def: A *d-cycle* is a d-chain C s.t. $\partial C = 0$.

Pb: Cycles are not holes!!



The homology groups

Lemma: $\partial_{n-1} \circ \partial_n = 0$.

Q: Prove it.

The homology groups

Lemma: $\partial_{n-1} \circ \partial_n = 0$.

Q: Prove it.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$':

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$

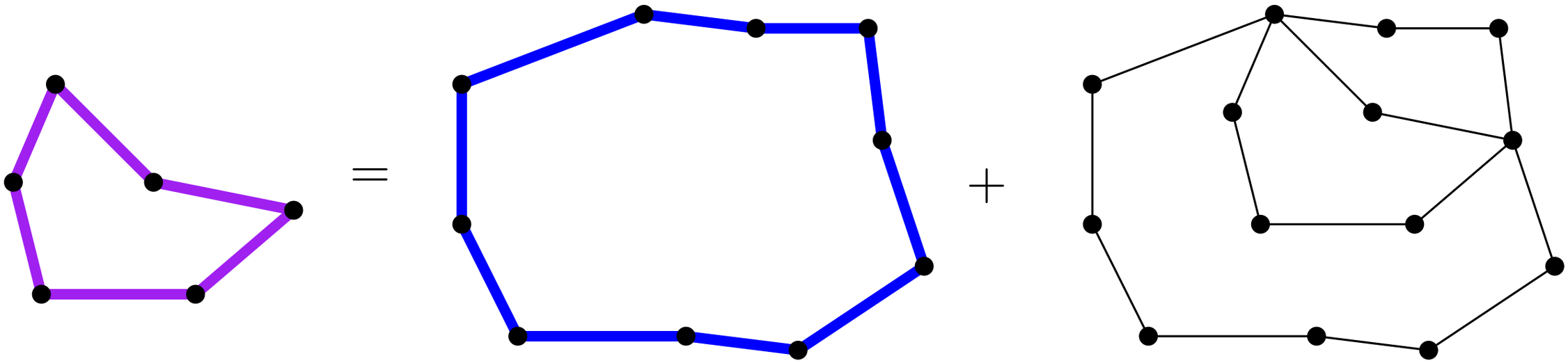
The homology groups

Lemma: $\partial_{n-1} \circ \partial_n = 0$.

Q: Prove it.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$':

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$



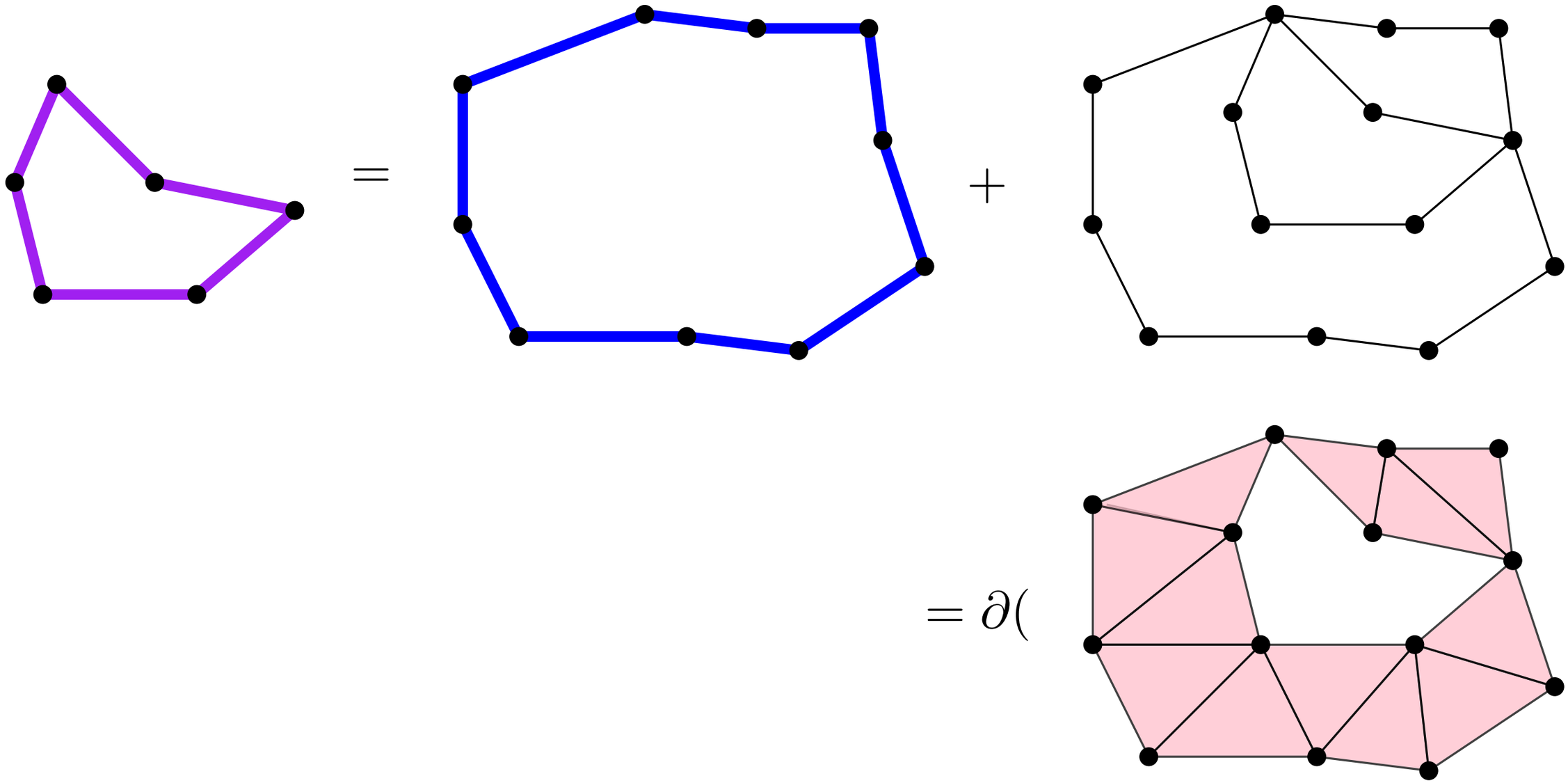
The homology groups

Lemma: $\partial_{n-1} \circ \partial_n = 0$.

Q: Prove it.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$':

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$



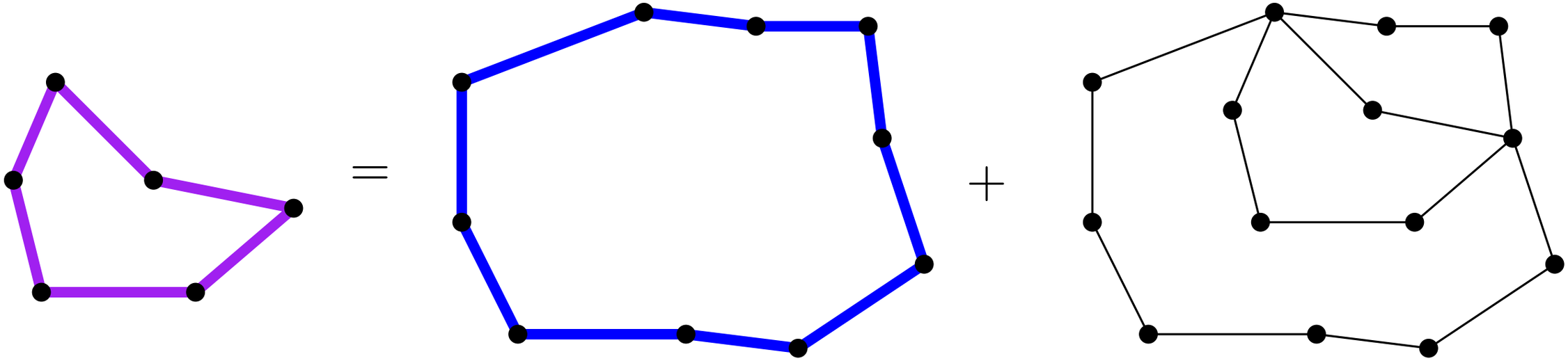
The homology groups

Lemma: $\partial_{n-1} \circ \partial_n = 0.$

Q: Prove it.

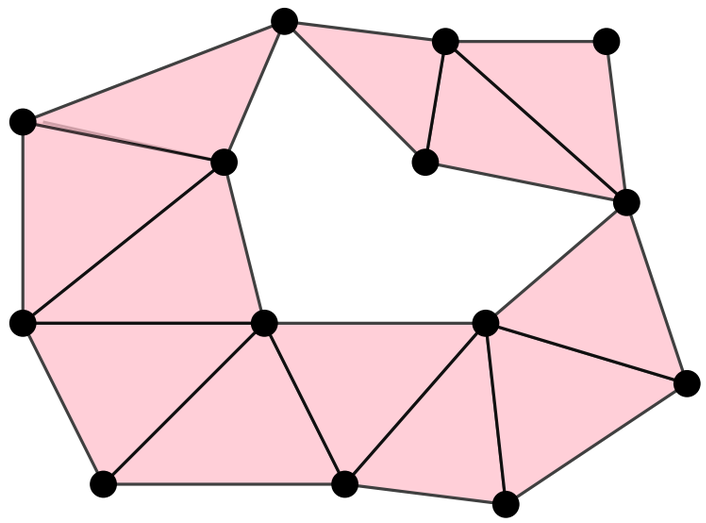
Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$':

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$



$$H_k = Z_k / B_k$$

$$= \partial($$



$$)$$

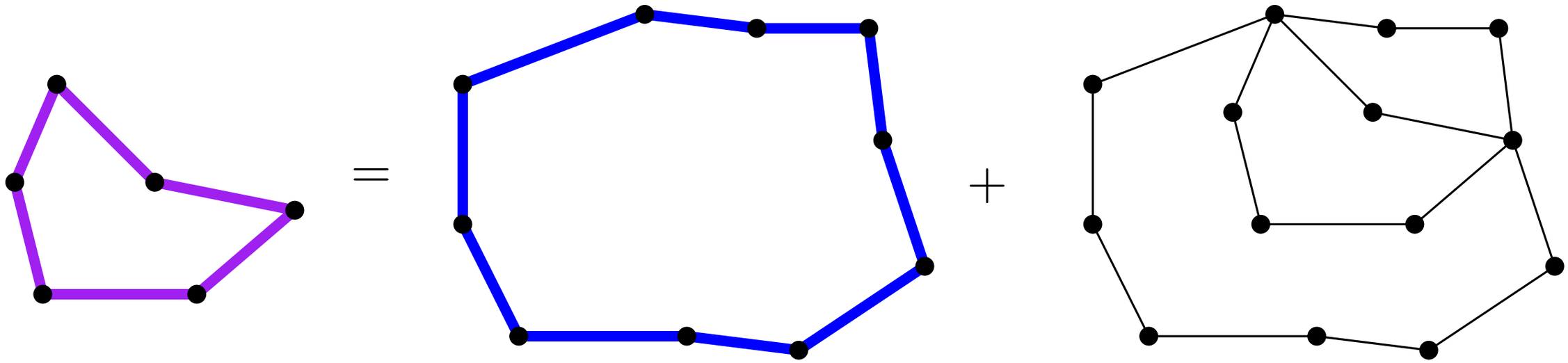
The homology groups

Lemma: $\partial_{n-1} \circ \partial_n = 0$.

Q: Prove it.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$':

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$

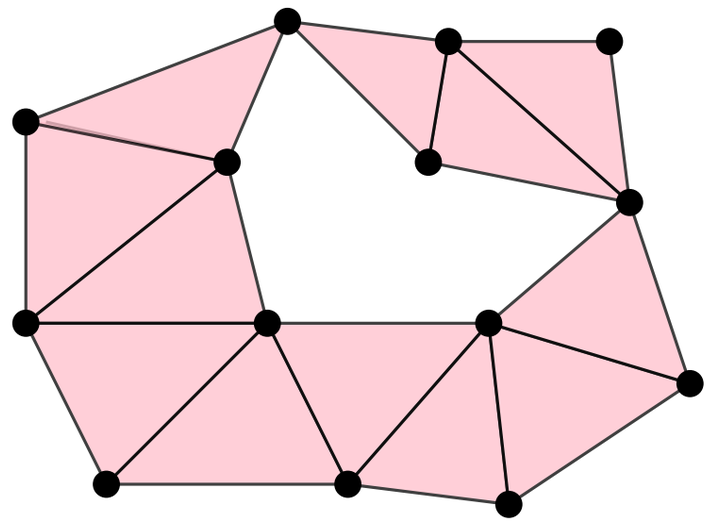


group of k-cycles

$$H_k = Z_k / B_k$$

group of 'cycles minus boundaries'

$\mathrm{im}(\partial_{k+1})$

$= \partial($  $)$

The homology groups

H_k is a group (vector space) in which each element is an equivalence class of cycles associated to the same hole.

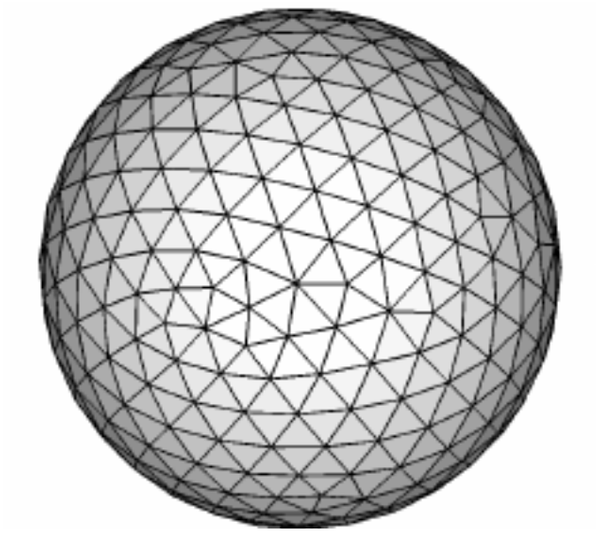
Def: The dimension of H_k is called the *Betti number* β_k.

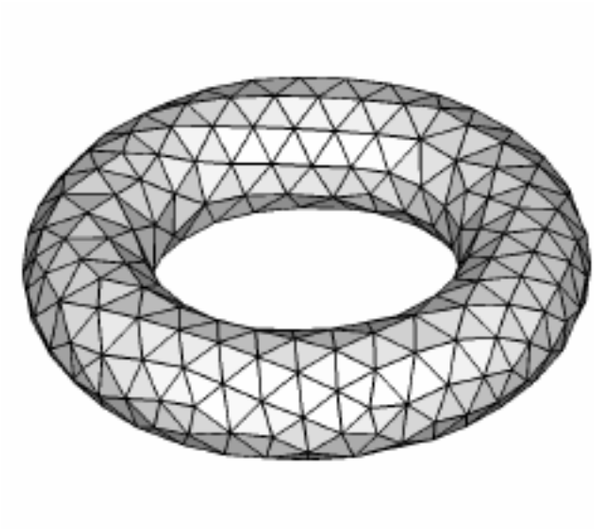
The homology groups

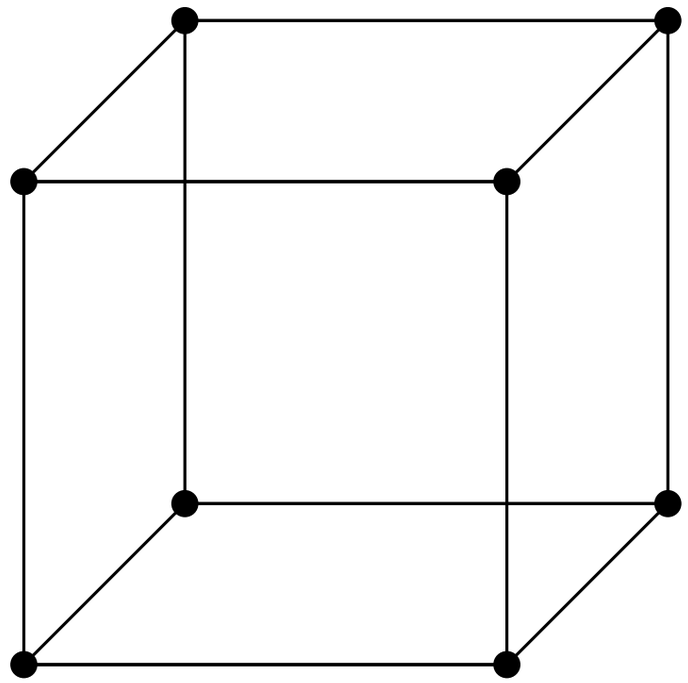
H_k is a group (vector space) in which each element is an equivalence class of cycles associated to the same hole.

Def: The dimension of H_k is called the *Betti number* β_k.

Q: What are the Betti numbers of:





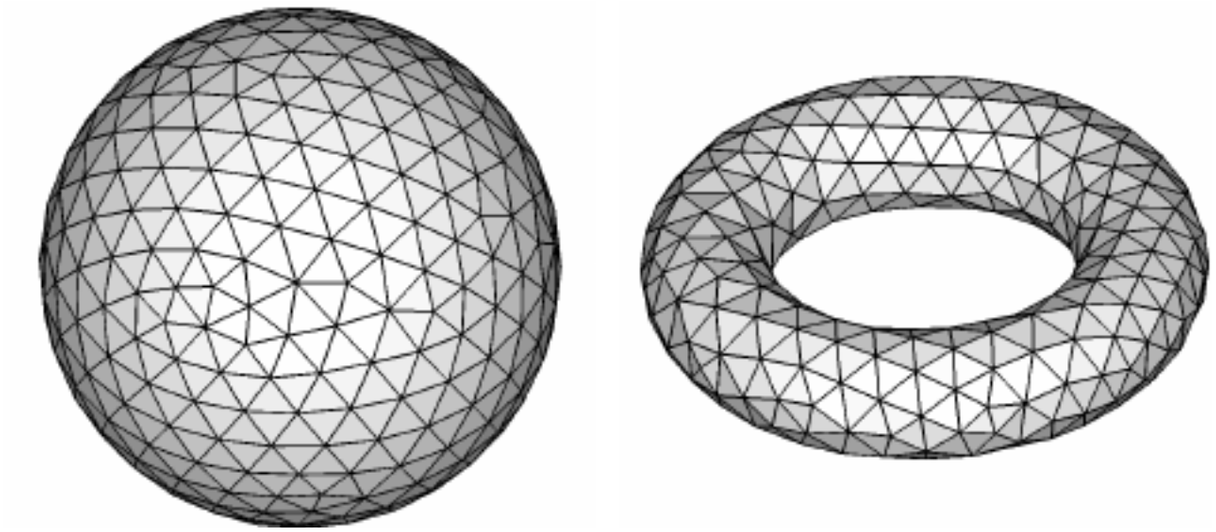


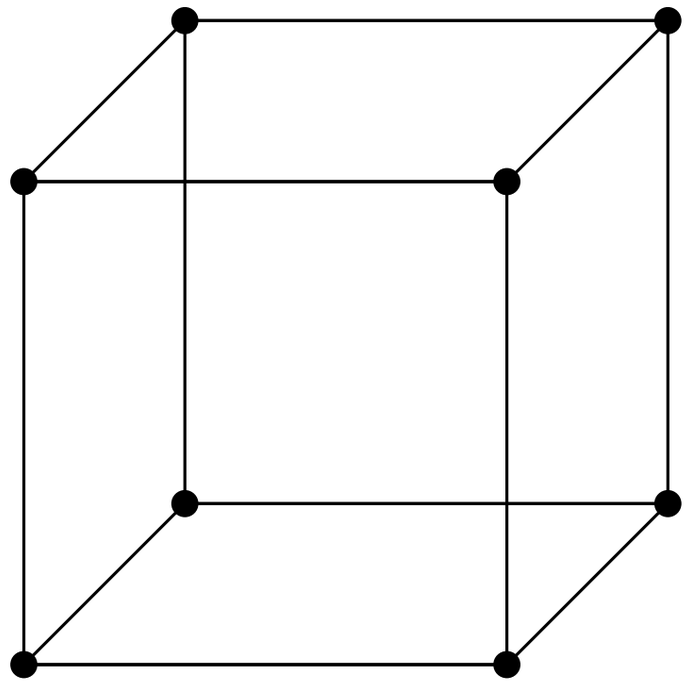
The homology groups

H_k is a group (vector space) in which each element is an equivalence class of cycles associated to the same hole.

Def: The dimension of H_k is called the *Betti number* β_k.

Q: What are the Betti numbers of:





The whole point of homology groups and Betti numbers is that they satisfy:

$$H_k(X) \not\sim H_k(Y) \Longrightarrow X \not\sim Y$$

Topological exploratory data analysis

Topological exploratory data analysis

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

Topological exploratory data analysis

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

Idea:

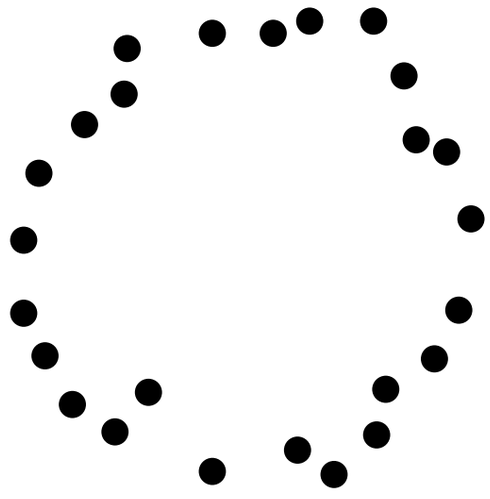
- Group data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of 'clusters'.

Topological exploratory data analysis

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

Idea:

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

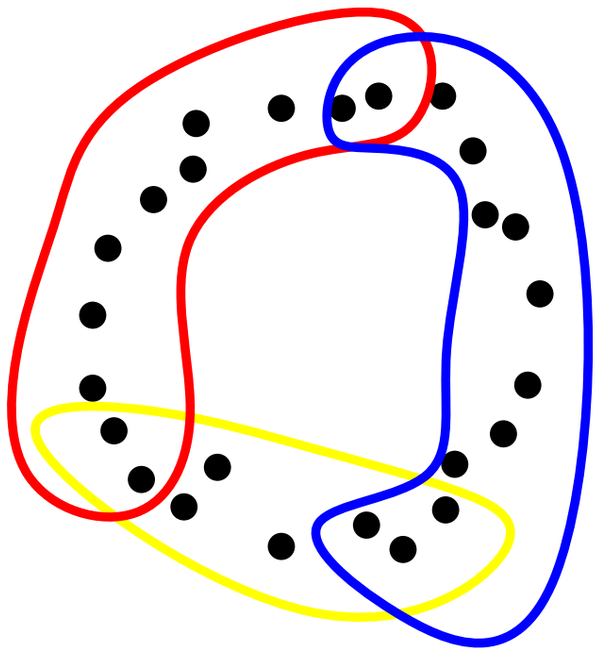


Topological exploratory data analysis

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

Idea:

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

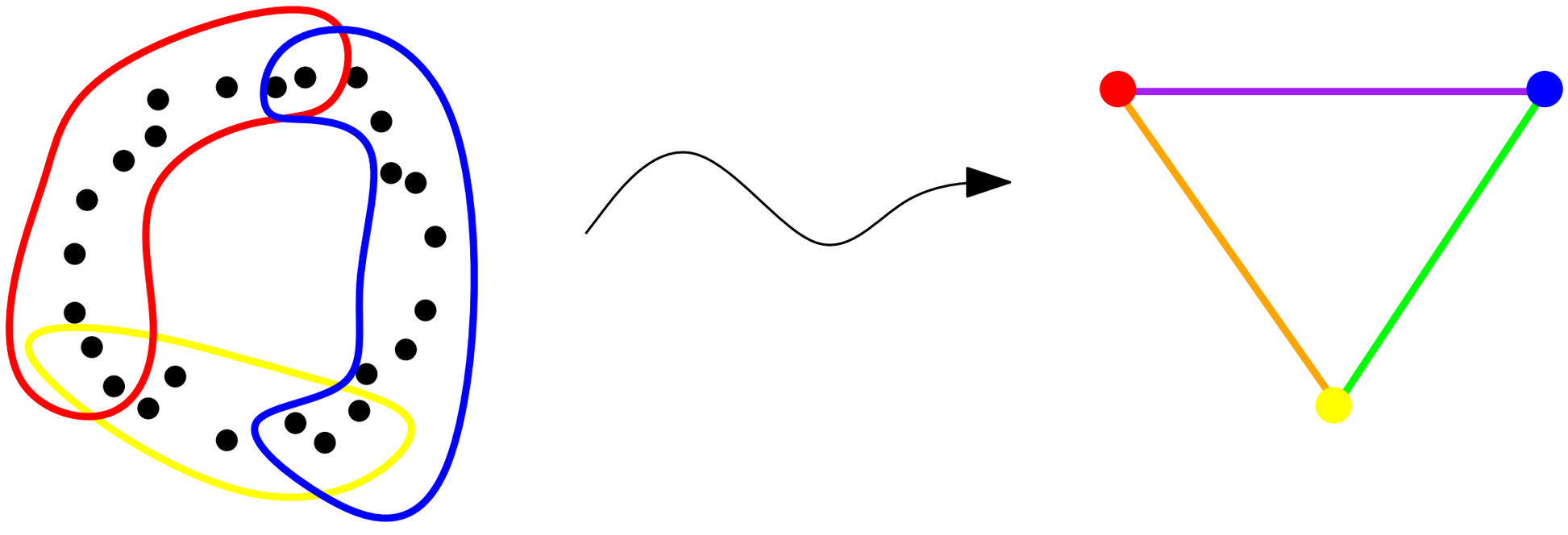


Topological exploratory data analysis

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

Idea:

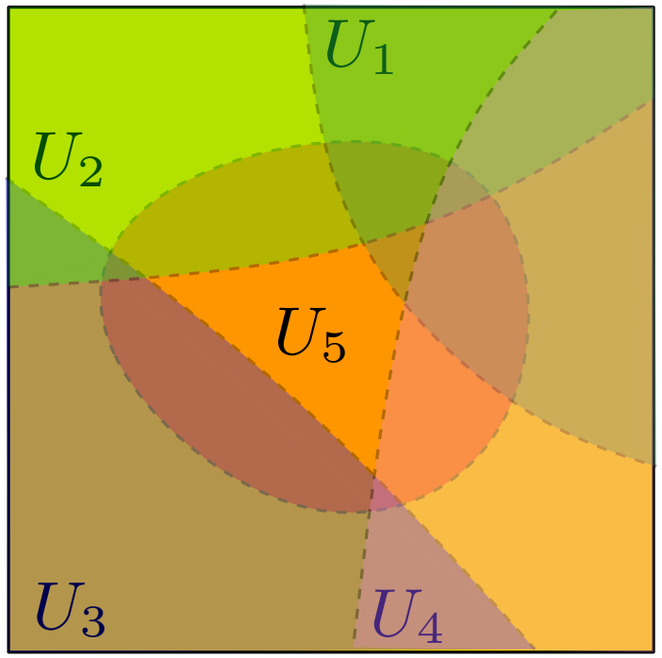
- Group data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of 'clusters'.



Topological exploratory data analysis

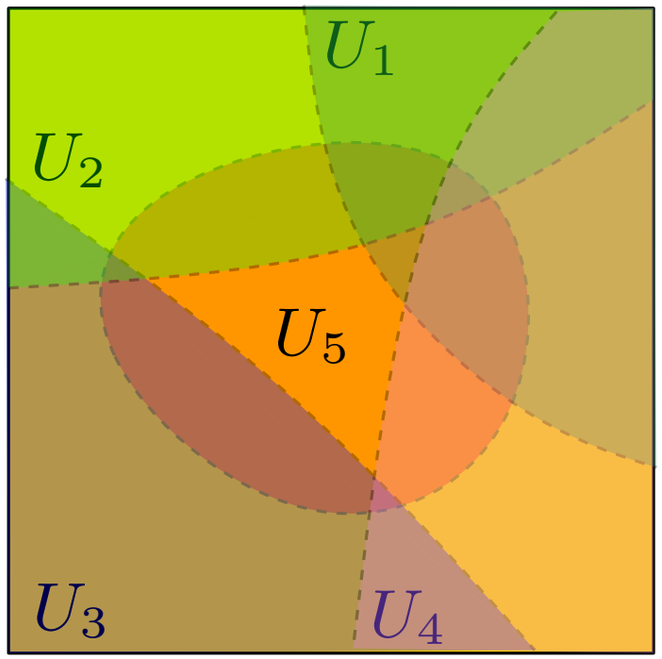
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

Topological exploratory data analysis



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

Topological exploratory data analysis

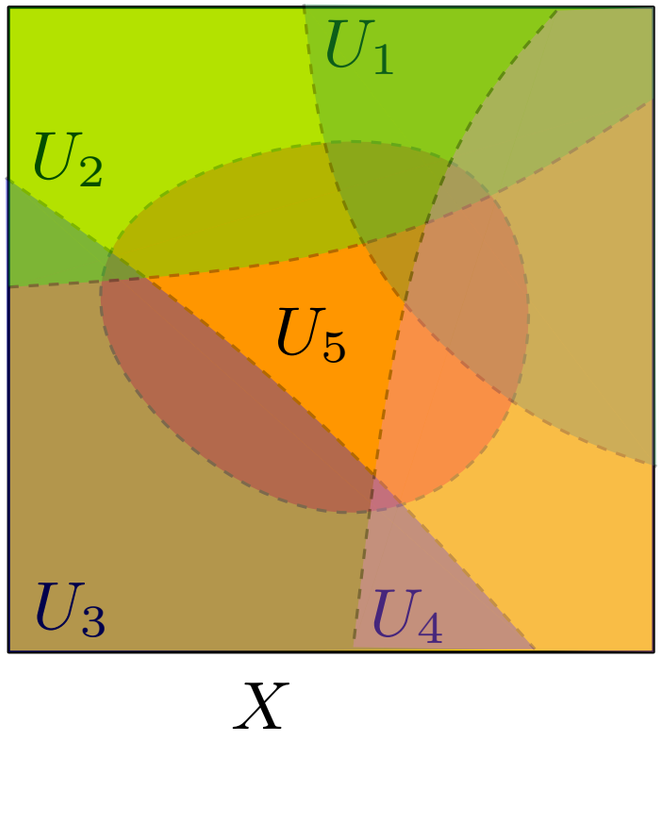


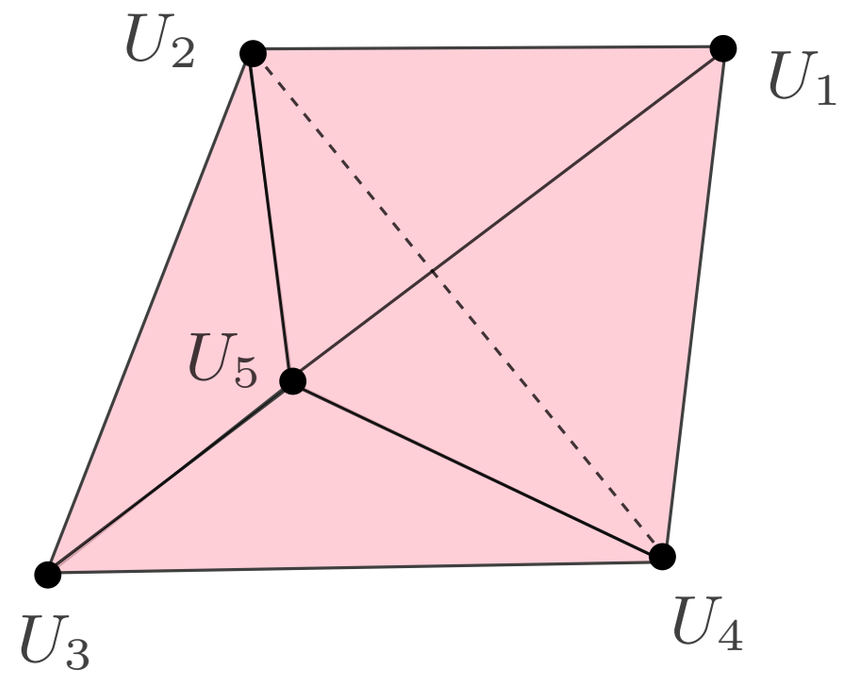
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

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

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

Topological exploratory data analysis





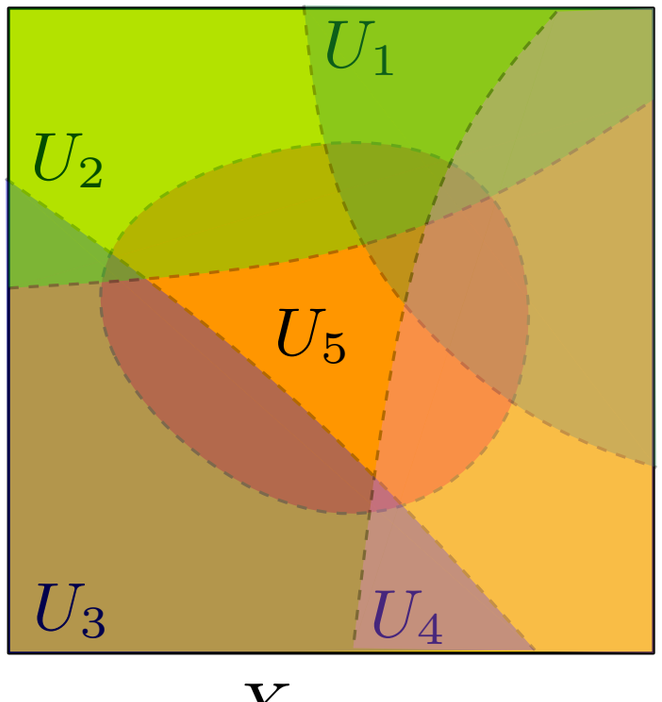
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

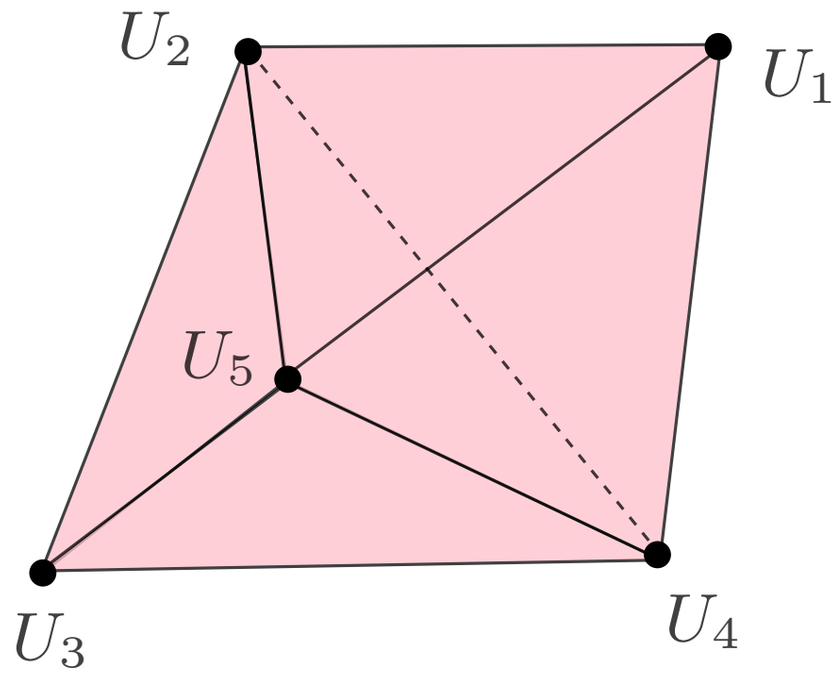
Def: Given a cover of a topological space X, $\mathcal{U} = (U_i)_{i \in I}$, its nerve is the abstract simplicial complex $C(\mathcal{U})$ whose vertex set is $\mathcal{U}$ and s.t.

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

Topological exploratory data analysis

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





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

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

Topological exploratory data analysis

Q: How to build meaningful covers?

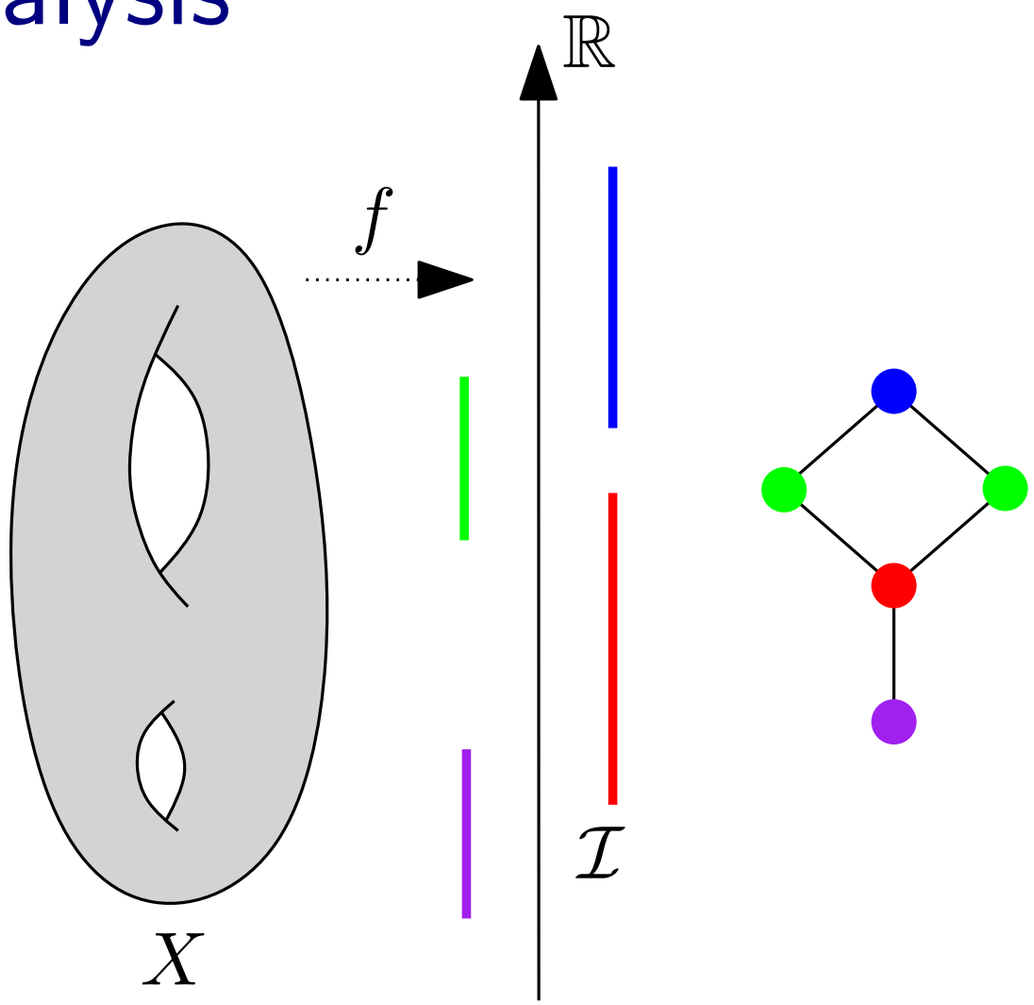
Two directions:

Topological exploratory data analysis

Q: How to build meaningful covers?

Two directions:

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

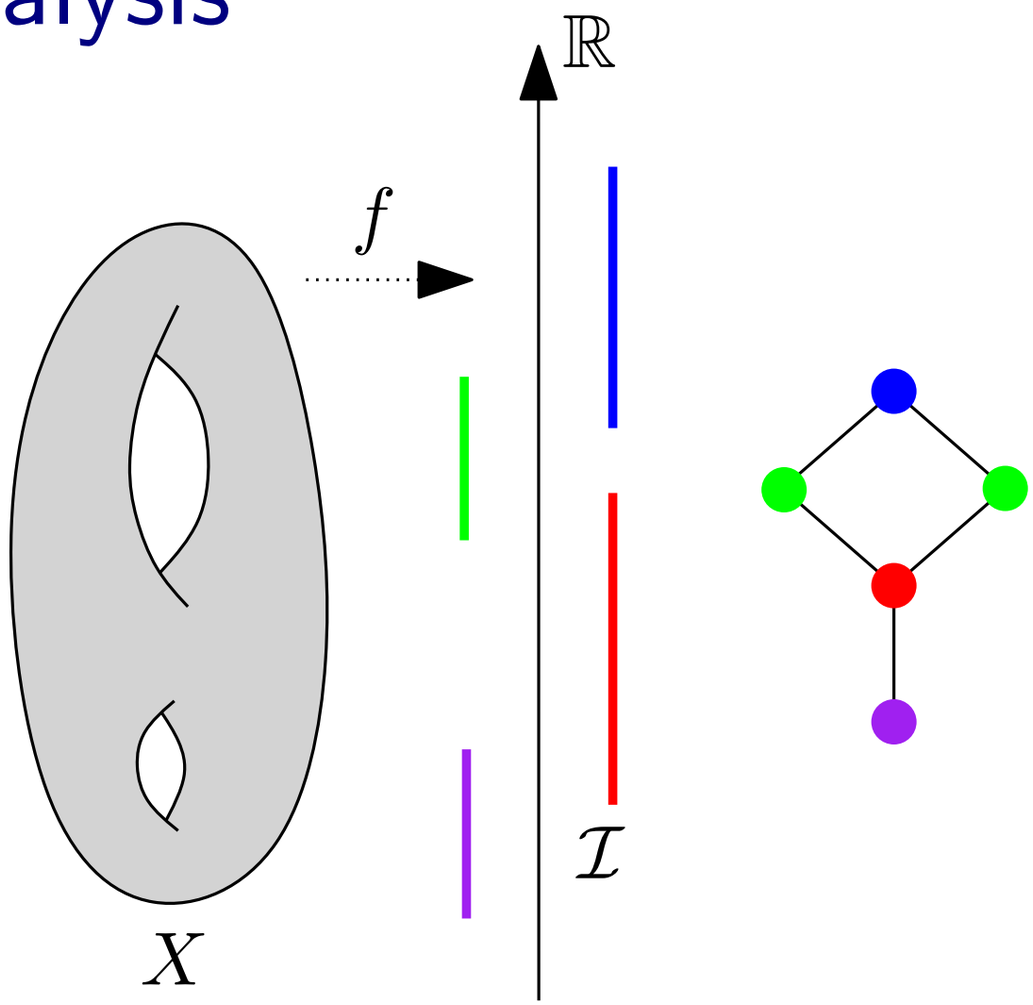


Topological exploratory data analysis

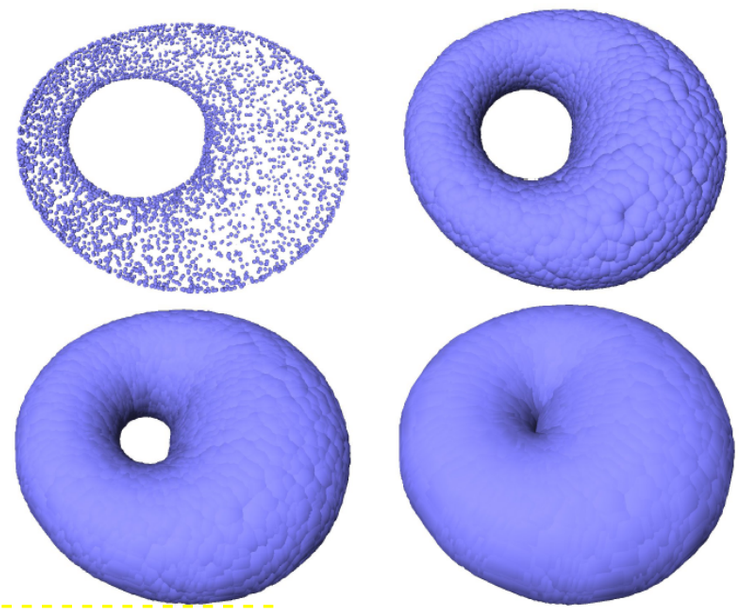
Q: How to build meaningful covers?

Two directions:

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



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

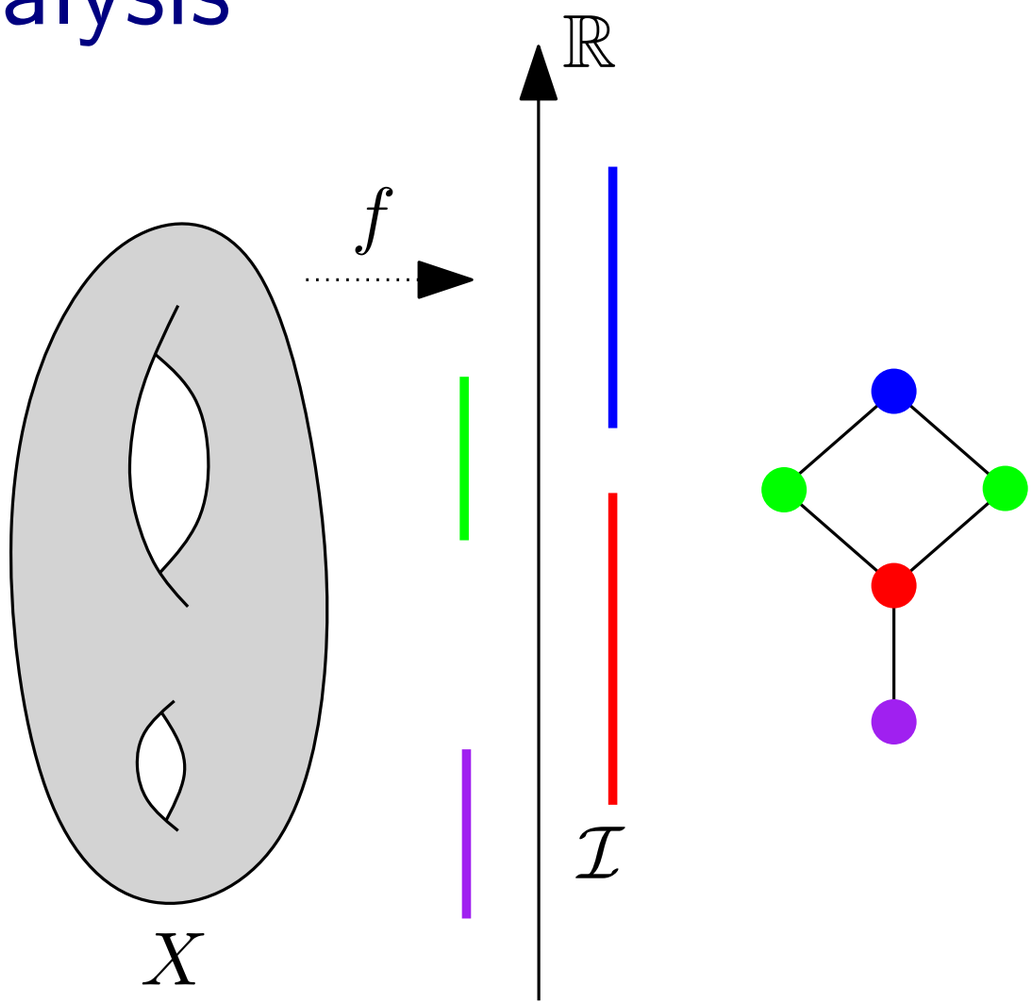


Topological exploratory data analysis

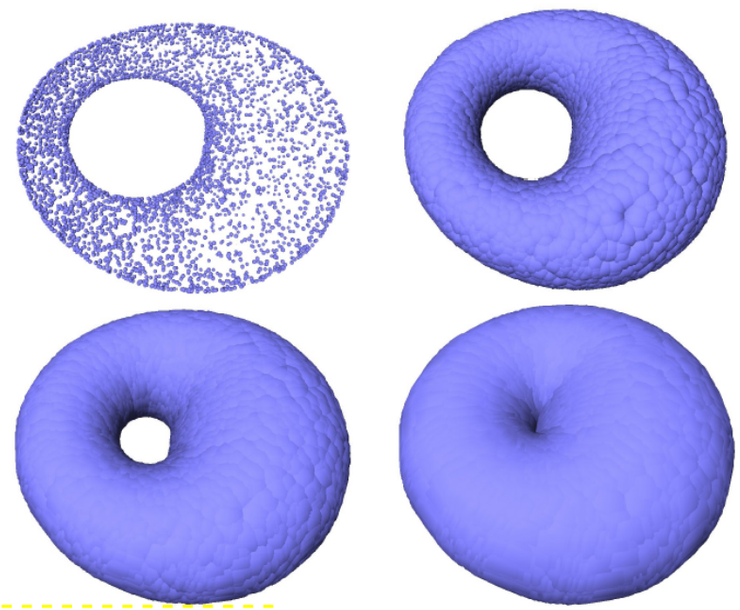
Q: How to build meaningful covers?

Two directions:

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



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



Topological exploratory data analysis: statistical aspects of Mapper

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

[*Probabilistic Convergence and Stability of Random Mapper Graphs*, Brown, Bobrowski, Munch, Wang, J. Applied Comput. Topo., 2020]

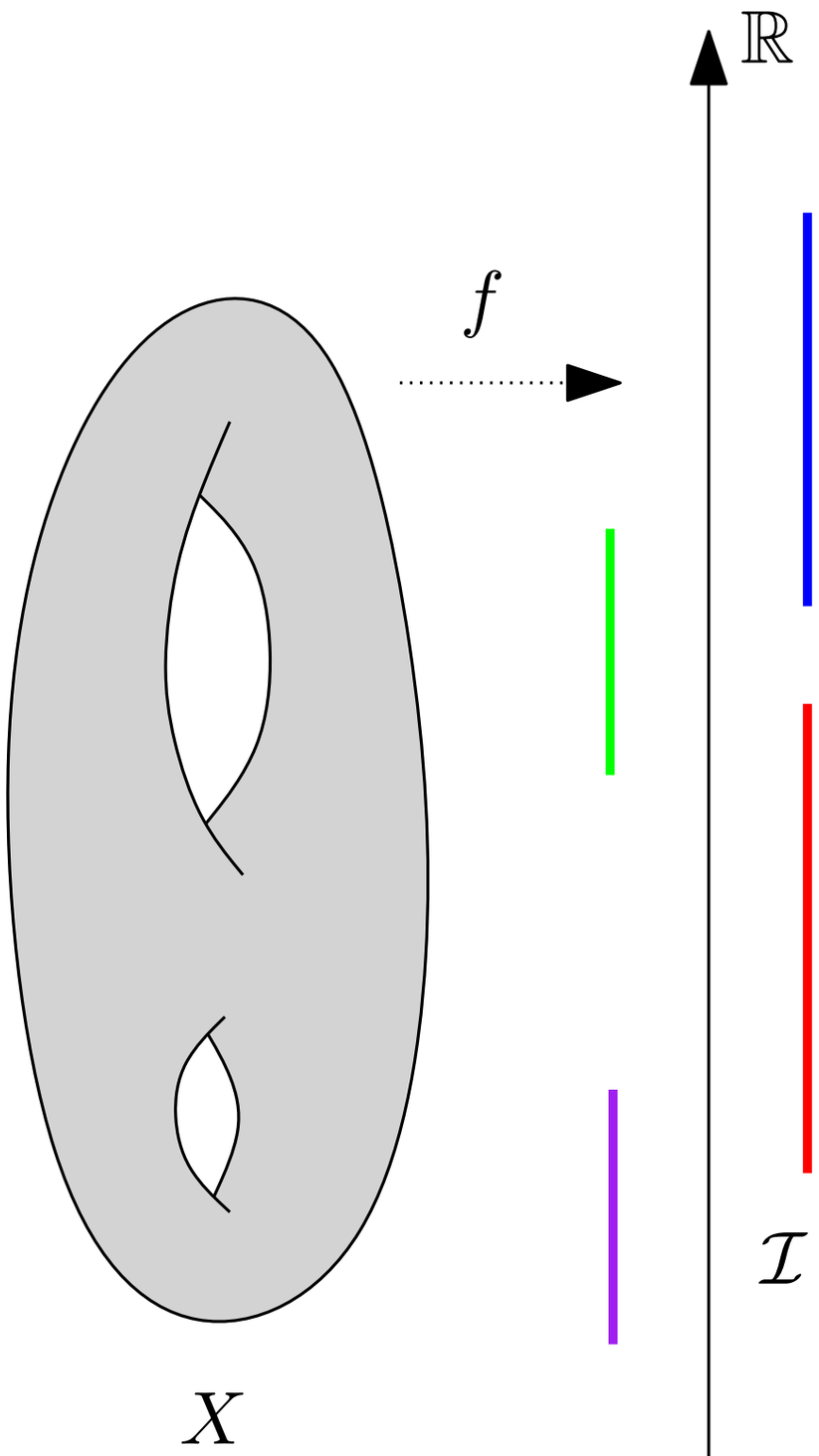
[*Statistical Analysis and Parameter Selection for Mapper*, Carrière, Michel, Oudot, J. Machine Learning Research, 2018]

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

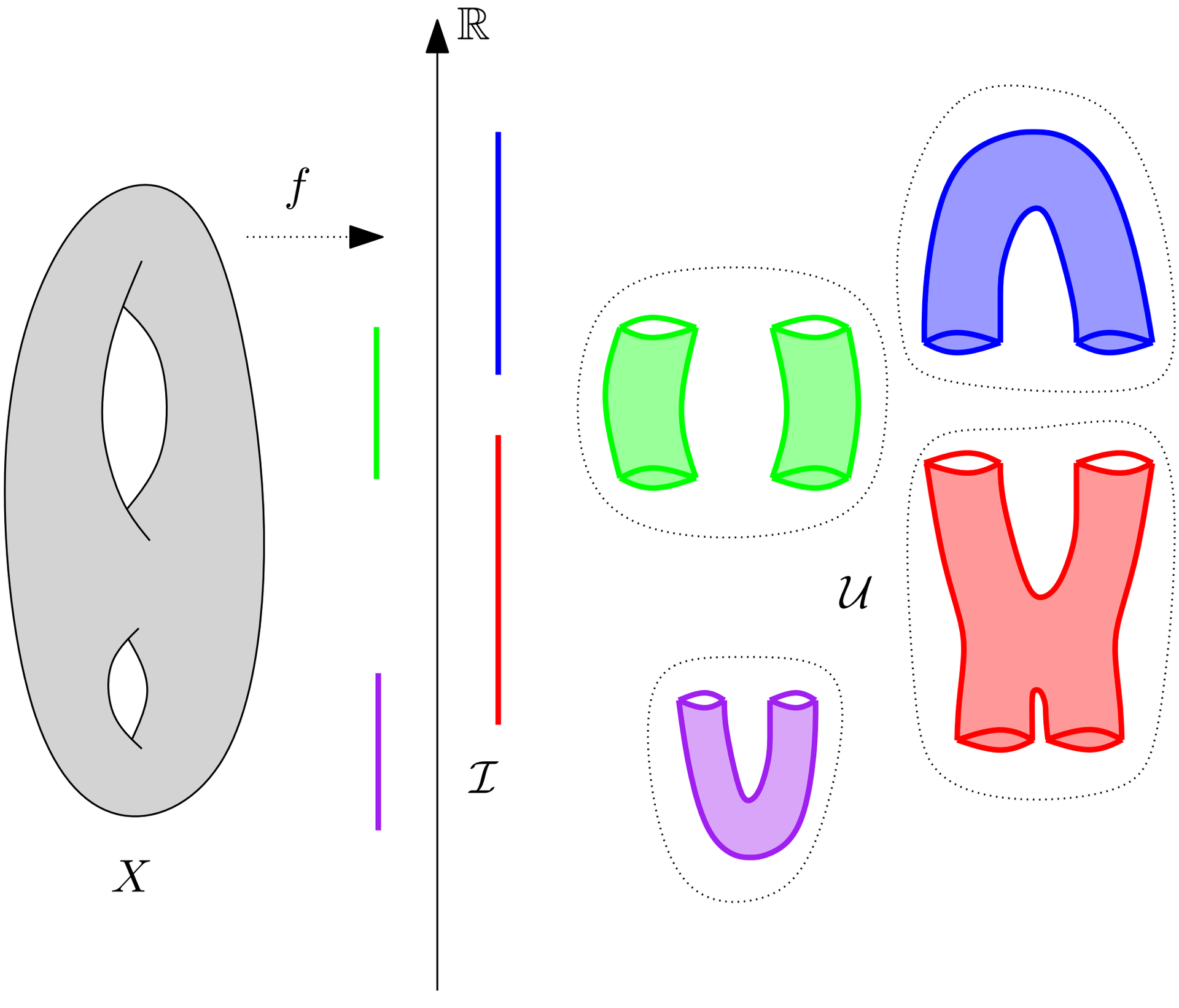
[*Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique*, Reeb, C. R. Acad. Sci. Paris, 1946]

[*Reeb Graphs: Approximation and Persistence*, Dey, Wang, Discr. Comput. Geom., 2013]

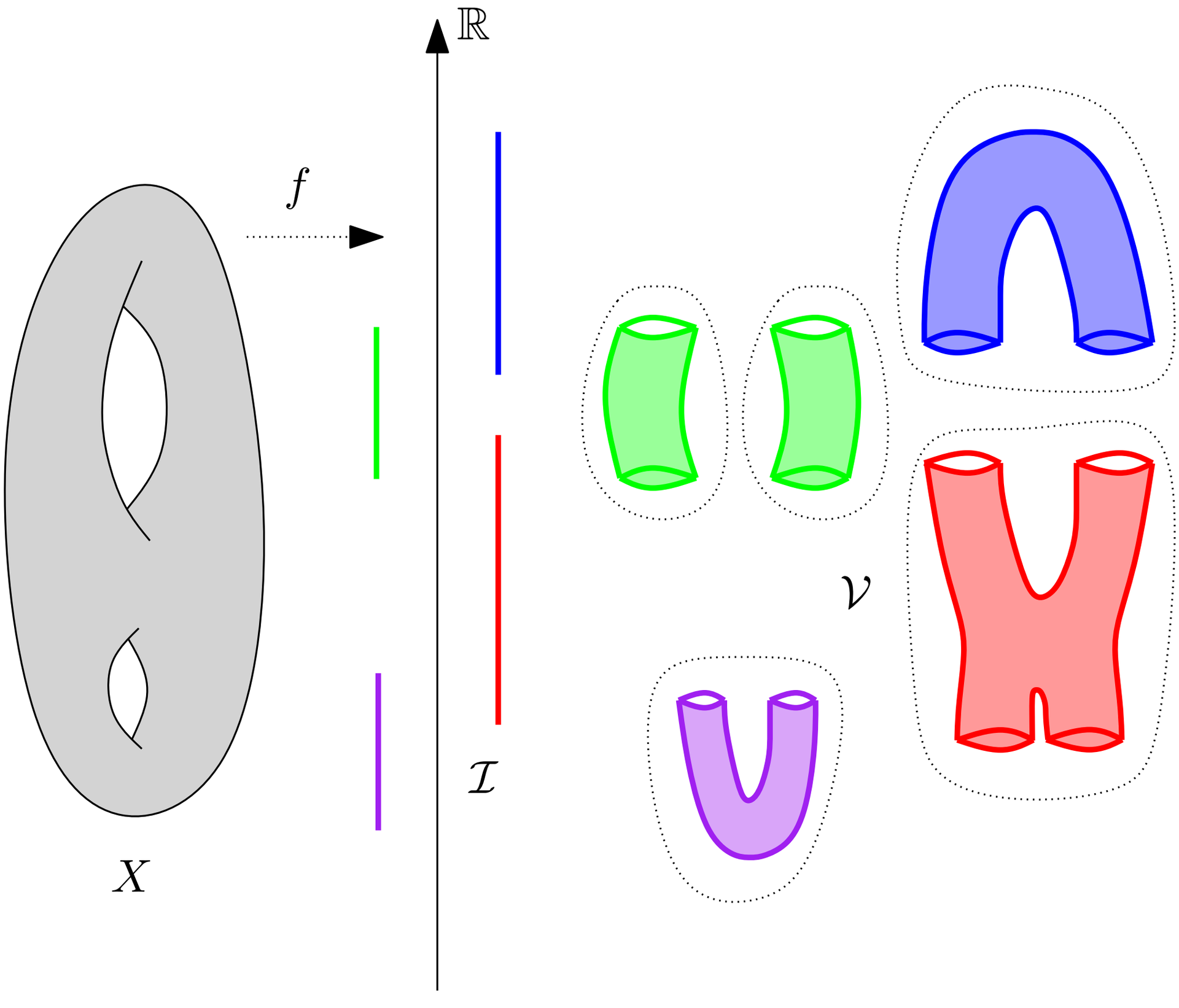
Mapper in the continuous setting



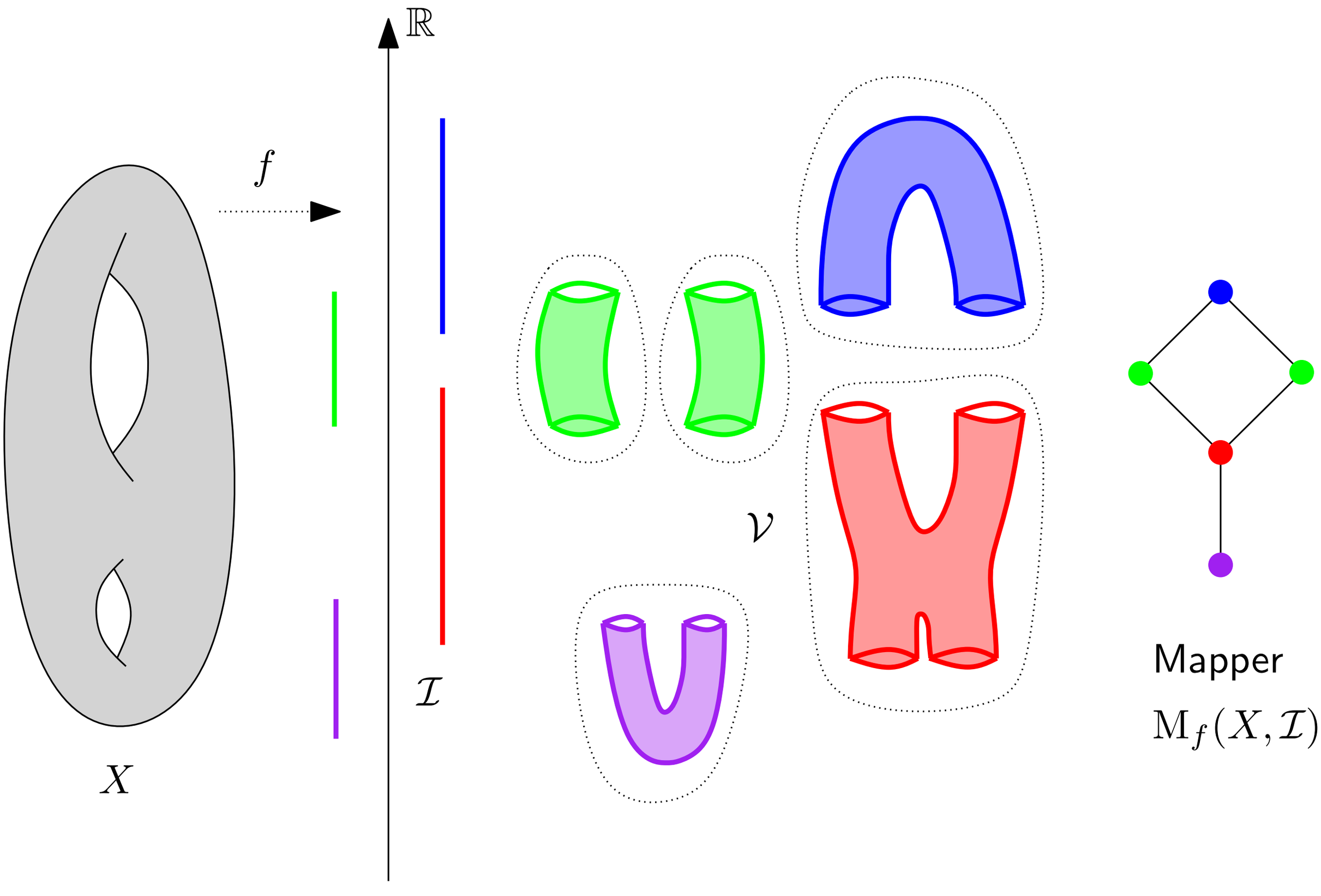
Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting

Input:

- topological space X
- continuous function $f : X \to \mathbb{R}$
- 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 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 the 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 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 $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

Warning: The nerve theorem does not apply in general!

Mapper in practice

Input:

- point cloud $P \subseteq X$ with metric d_P
- continuous function $f : P \to \mathbb{R}$
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals: $\mathrm{im} f \subseteq \bigcup_{I \in \mathcal{I}} I$

Method:

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

• Refine $\mathcal{U}$ by separating each of its elements into its various clusters, as identified by a clustering algorithm $\to$ connected cover $\mathcal{V}$

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

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

intersections are assessed by the presence of common data points

Mapper in practice

Parameters:

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

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

- clustering algorithm $\mathcal{C}$

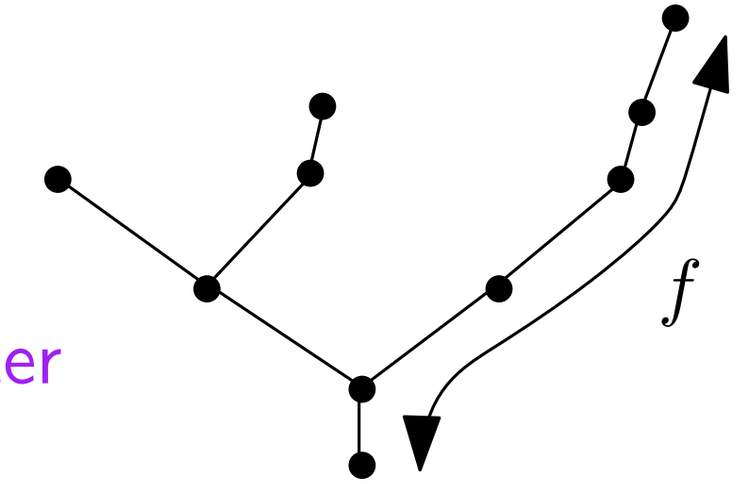
Mapper in practice

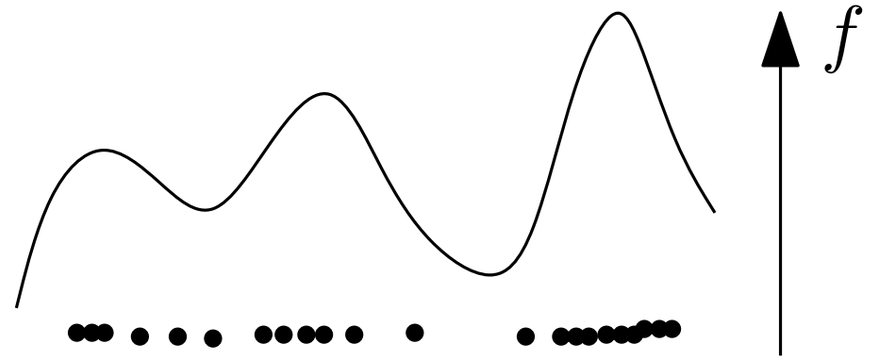
Parameters:

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

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

- clustering algorithm $\mathcal{C}$





Classical choices:

- density estimates
- centrality $f(x) = \sum_{y \in X} d(x, y)$
- eccentricity $f(x) = \max_{y \in X} d(x, y)$
- PCA coordinates
- Eigenfunctions of graph laplacians.
- Functions detecting outliers.
- Distance to a root point.
- Prior knowledge

Mapper in practice

Parameters:

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

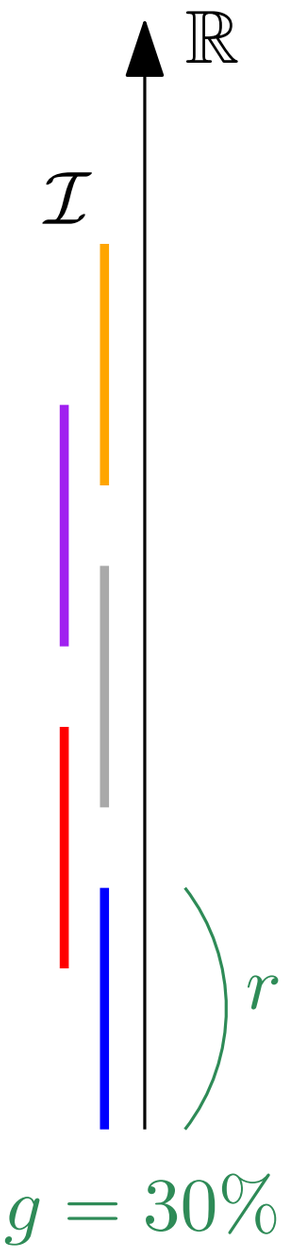
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals

- clustering algorithm $\mathcal{C}$

range scale

Uniform cover:

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



Mapper in practice

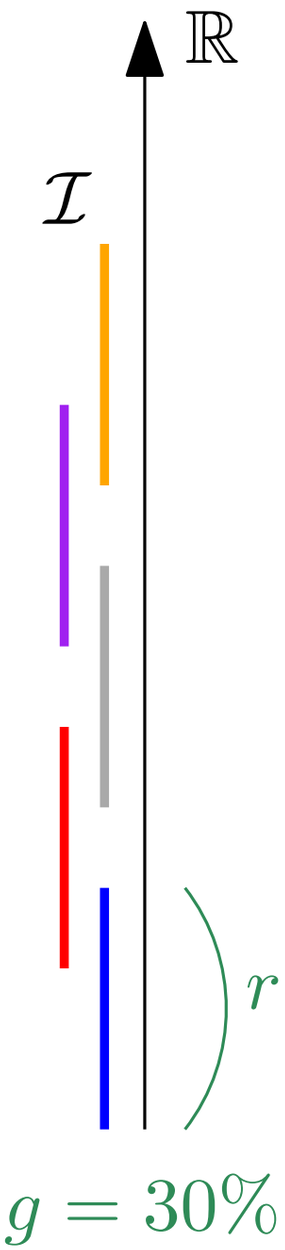
Parameters:

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

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

- clustering algorithm $\mathcal{C}$

range scale



Uniform cover:

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

Intuition:

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

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

Mapper in practice

Parameters:

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

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

- clustering algorithm $\mathcal{C}$

Classical choices:

- any clustering algorithm works

- different clustering algorithms/parameters for each preimage

- for theoretical reasons, we prefer to work with

  hierarchical clustering with (predefined) neighborhood size δ

geometric scale

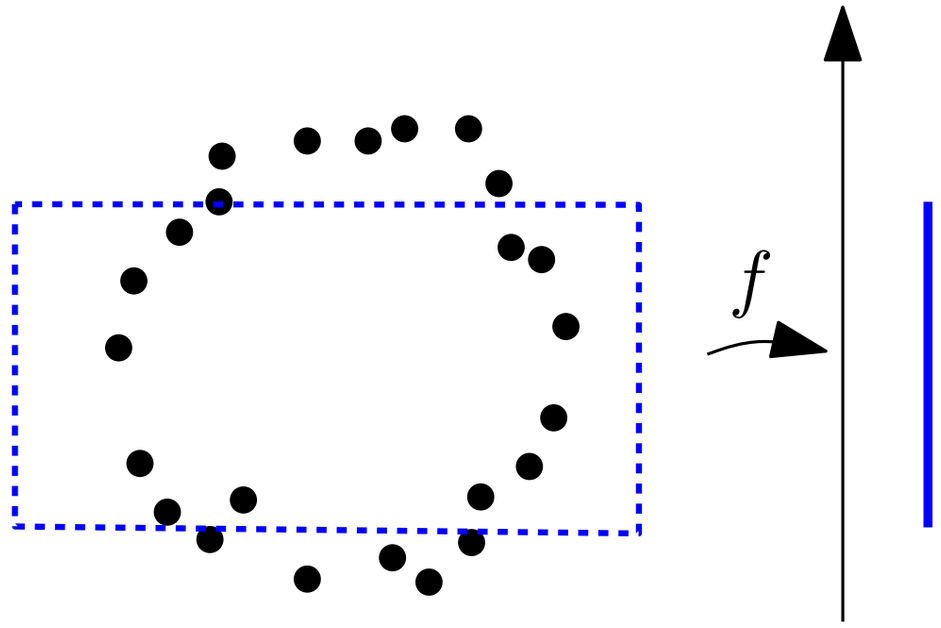
Mapper in practice

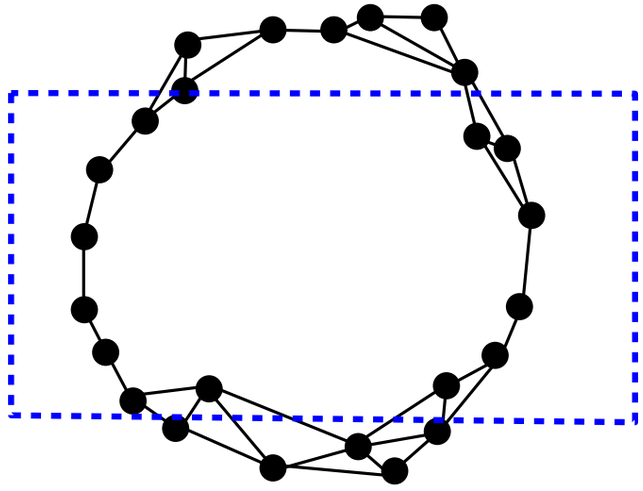
Parameters:

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

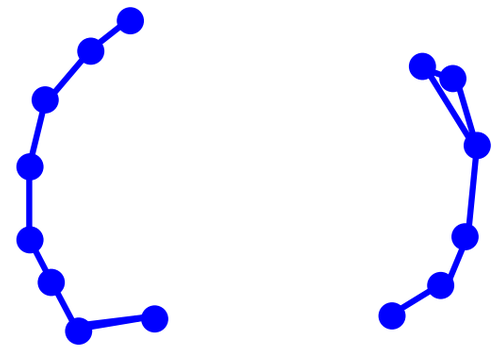
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals

- clustering algorithm $\mathcal{C}$



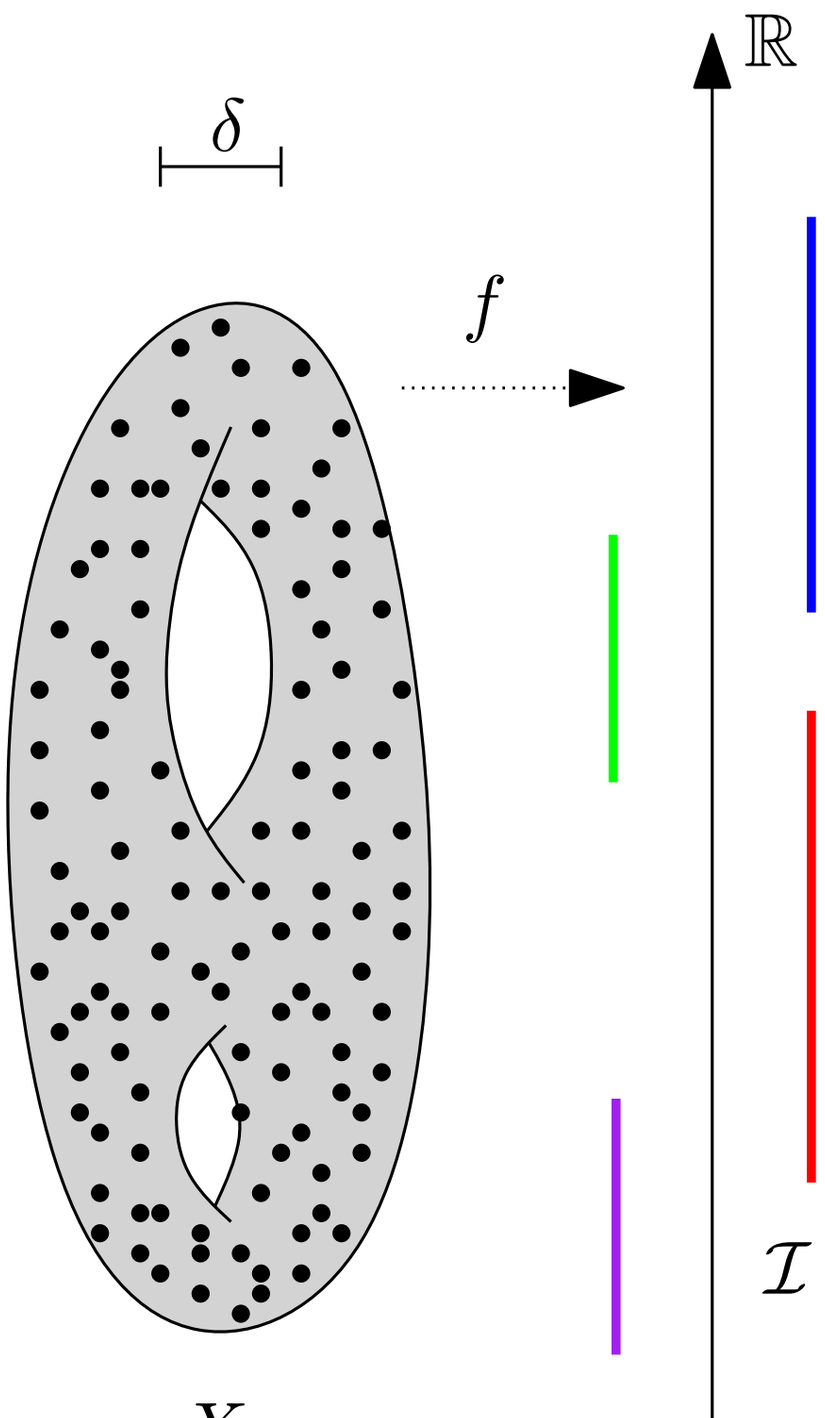


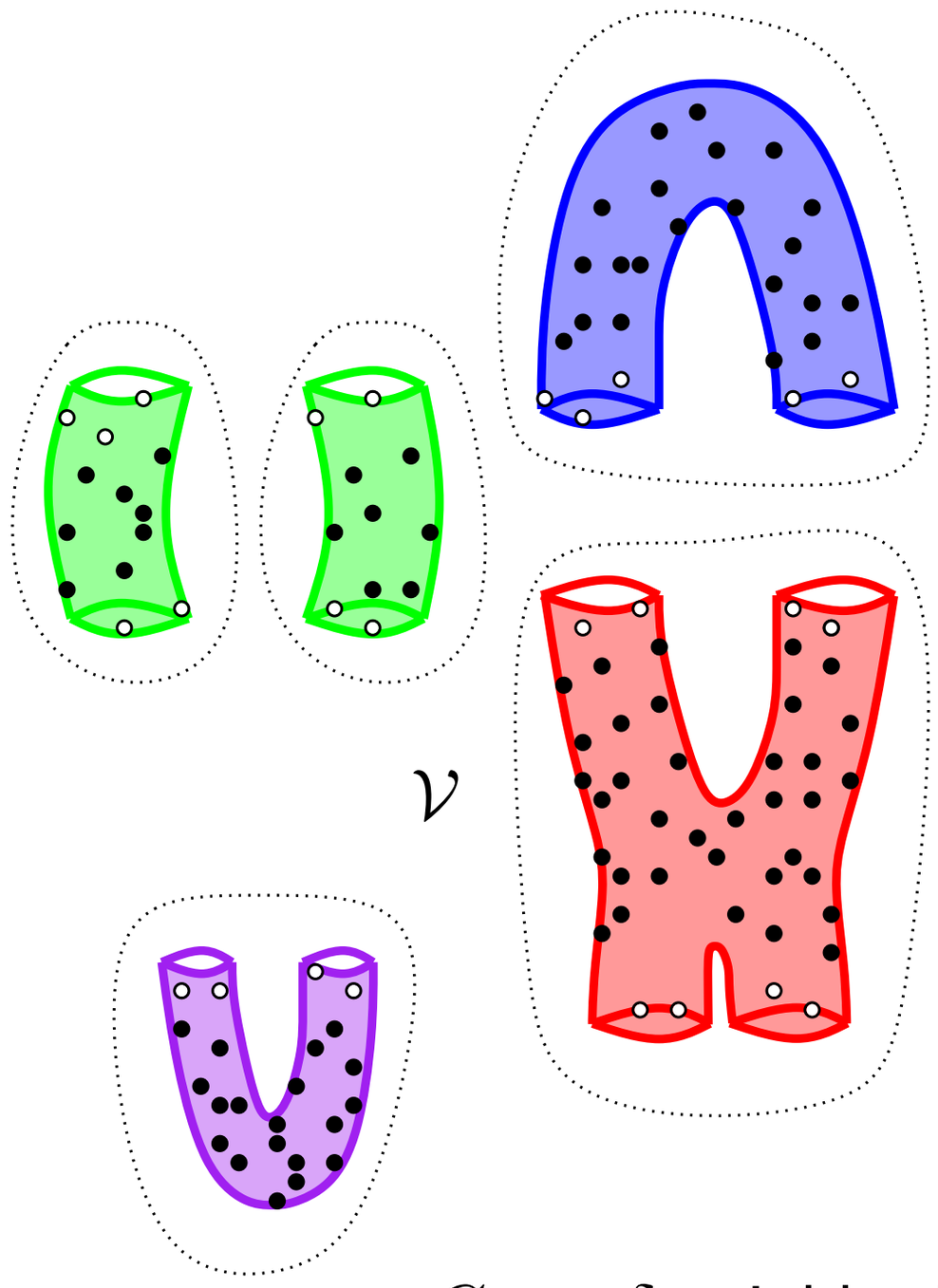
Build a neighboring graph (kNN,...)

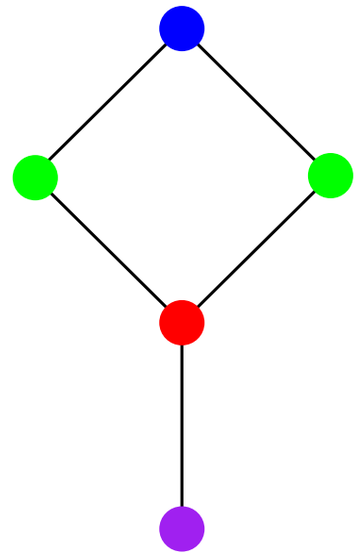


Take the connected components of the subgraph spanned by the vertices in the preimage $f^{-1}(U)$.

Mapper in practice







Mapper

$\mathrm{M}^{\bullet}_{f,\delta}(P,\mathcal{I})$

$G_\delta = \delta$-neighborhood graph

The Mapper instability

Pbm: Mapper is known to be very sensitive to noise/parameters!

The Mapper instability

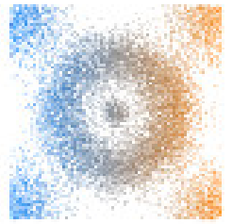
Pbm: Mapper is known to be very sensitive to noise/parameters!

In practice: trial-and-error

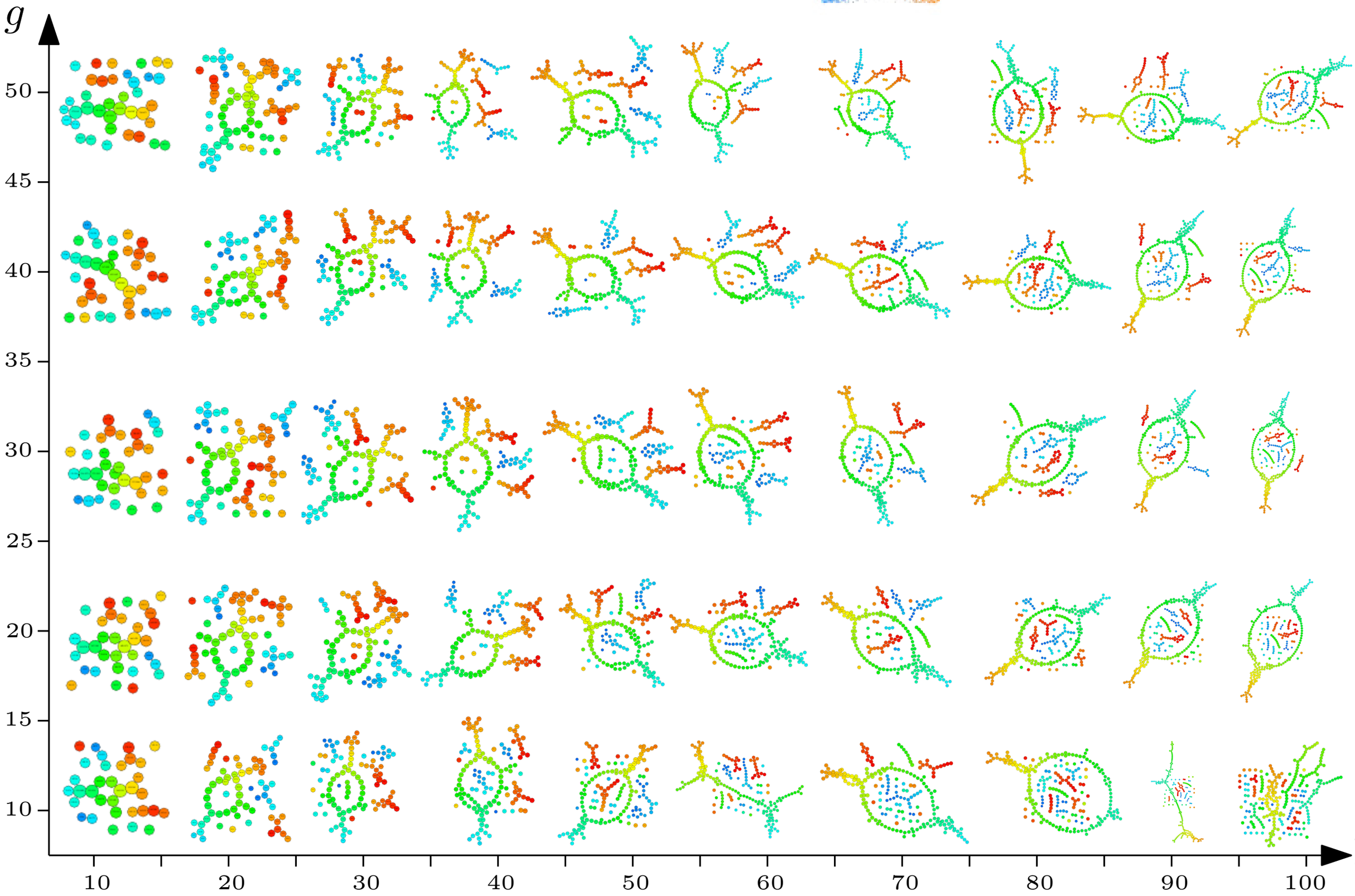
high-dimensional data sets[40,48]. This is performed automatically within the software, by deploying an ensemble machine learning algorithm that iterates through overlapping subject bins of different sizes that resample the metric space (with replacement), thereby using a combination of the metric location and similarity of subjects in the network topology. After performing millions of iterations, the algorithm returns the most stable, consensus vote for the resulting 'golden network' (Reeb graph), representing the multidimensional data shape[12,40].

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

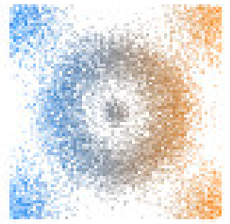
The Mapper instability



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



The Mapper instability



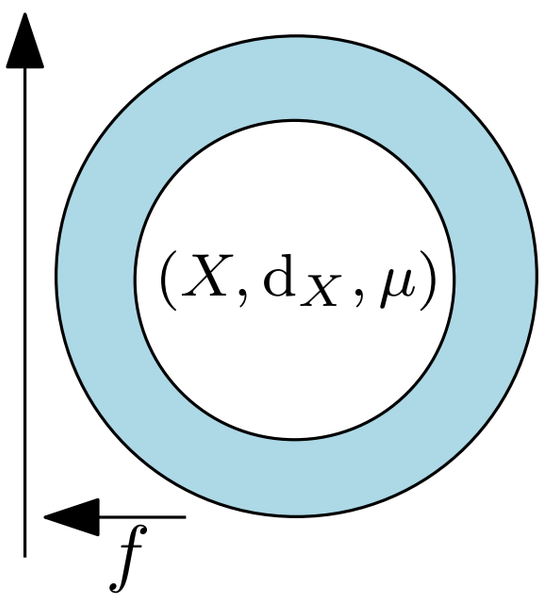
$f = f_x,\ \delta = 1\%$

g

50
45
40
35
30
25
20
15
10

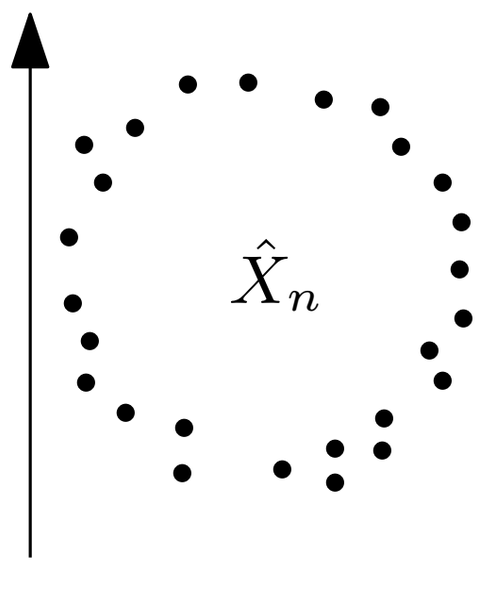
10 20 30 40 50 60 70 80 90 100 $1/r$

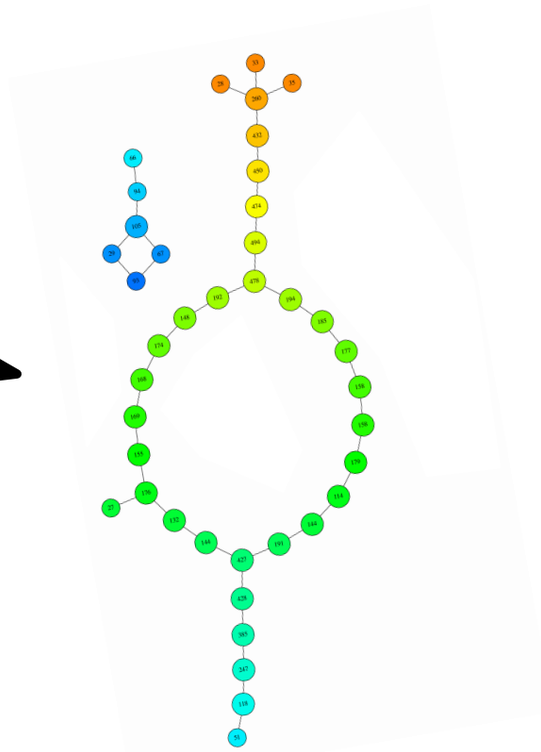
Understanding the Mapper structure



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

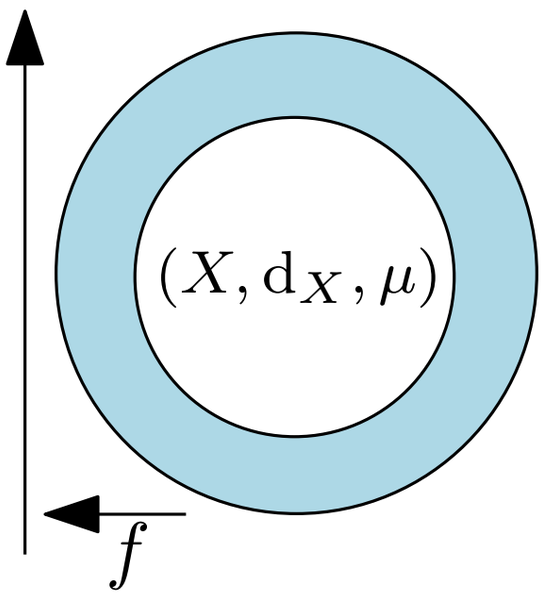
$+$ cover $\mathcal{I}$





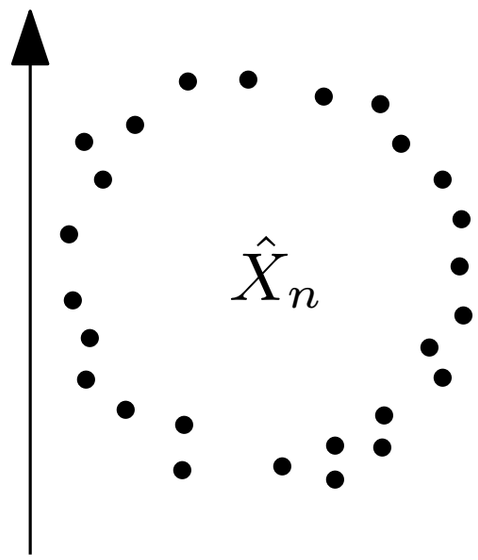
$\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$

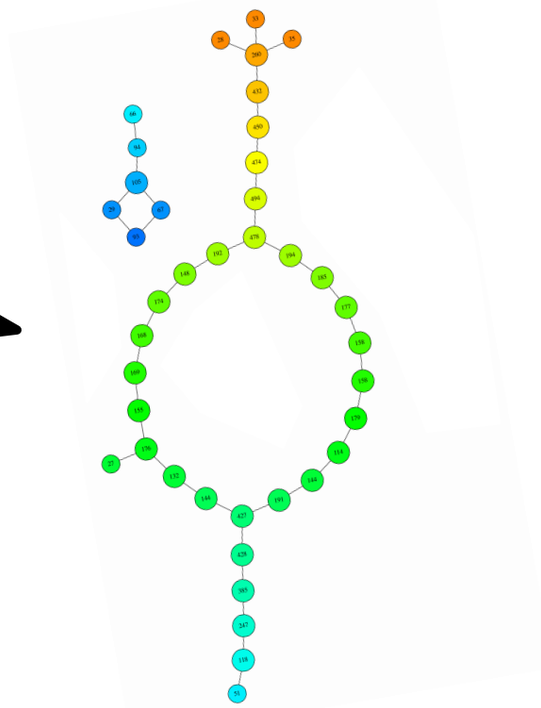
Understanding the Mapper structure



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

$+$ cover $\mathcal{I}$



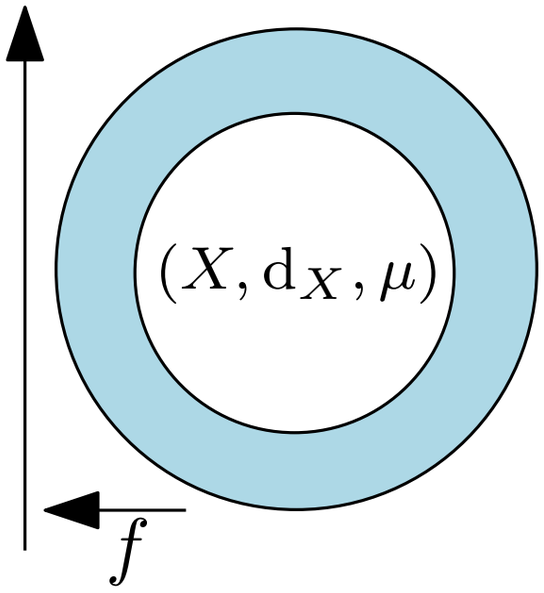


$\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$

We can decompose the relation between Mapper and X with

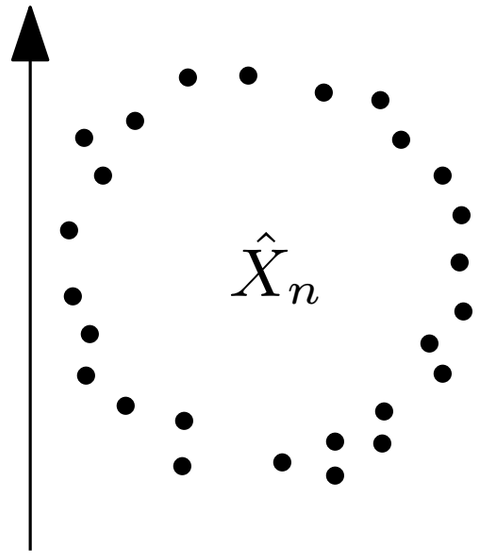
$$X \longleftrightarrow \mathrm{M}_f(X, \mathcal{I}) \longleftrightarrow \mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$$

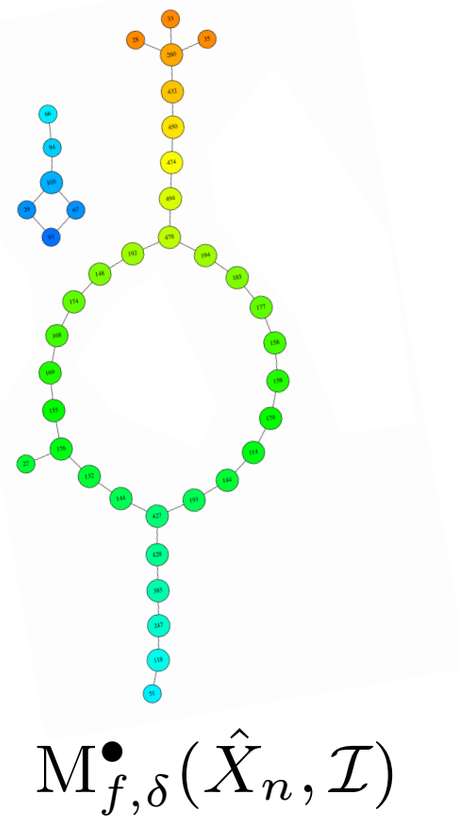
Understanding the Mapper structure



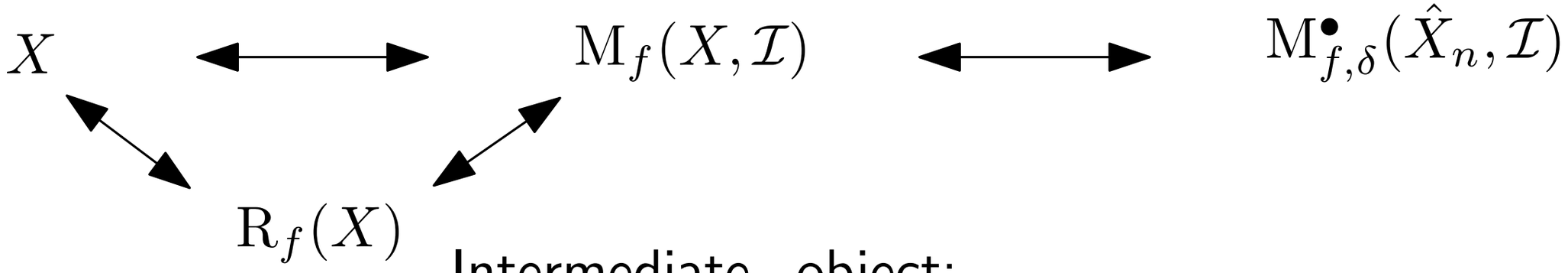
n points sampled
i.i.d. according to μ.

$+$ cover $\mathcal{I}$





We can decompose the relation between Mapper and X with

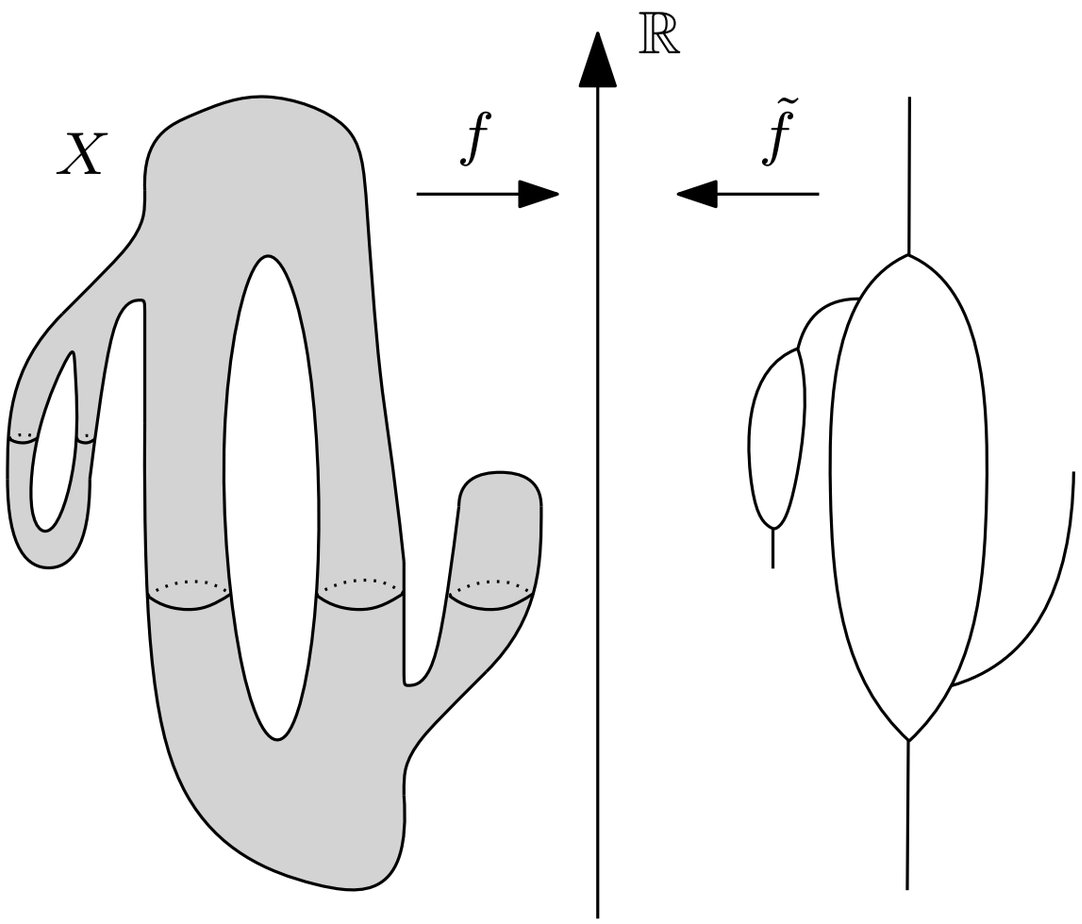


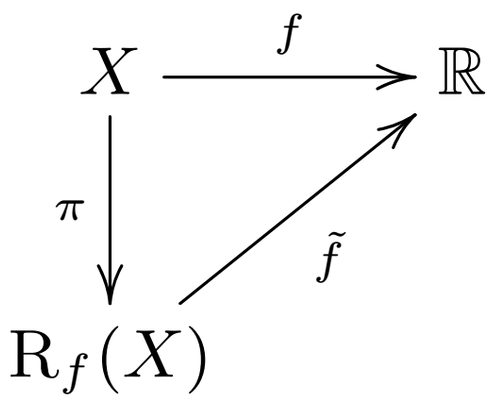
Intermediate object:
the Reeb graph

The Reeb graph and the space

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

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

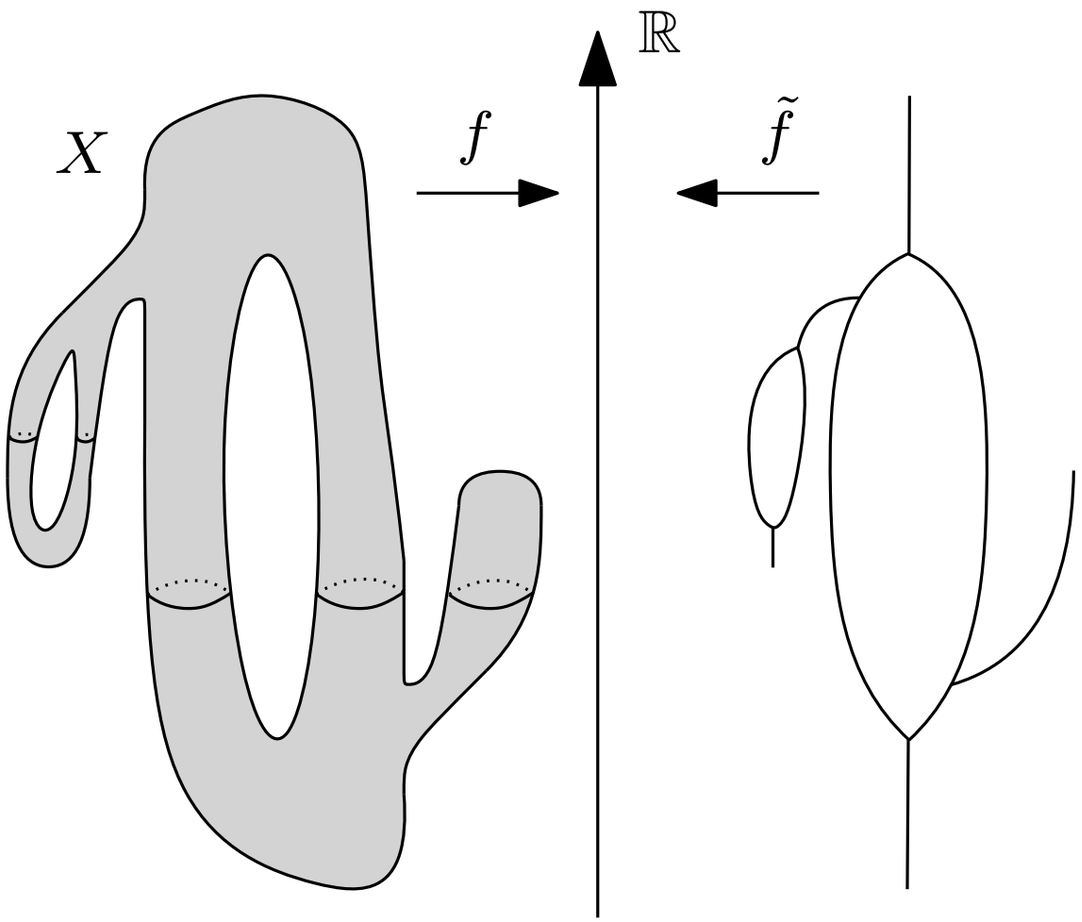




The Reeb graph and the space

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

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

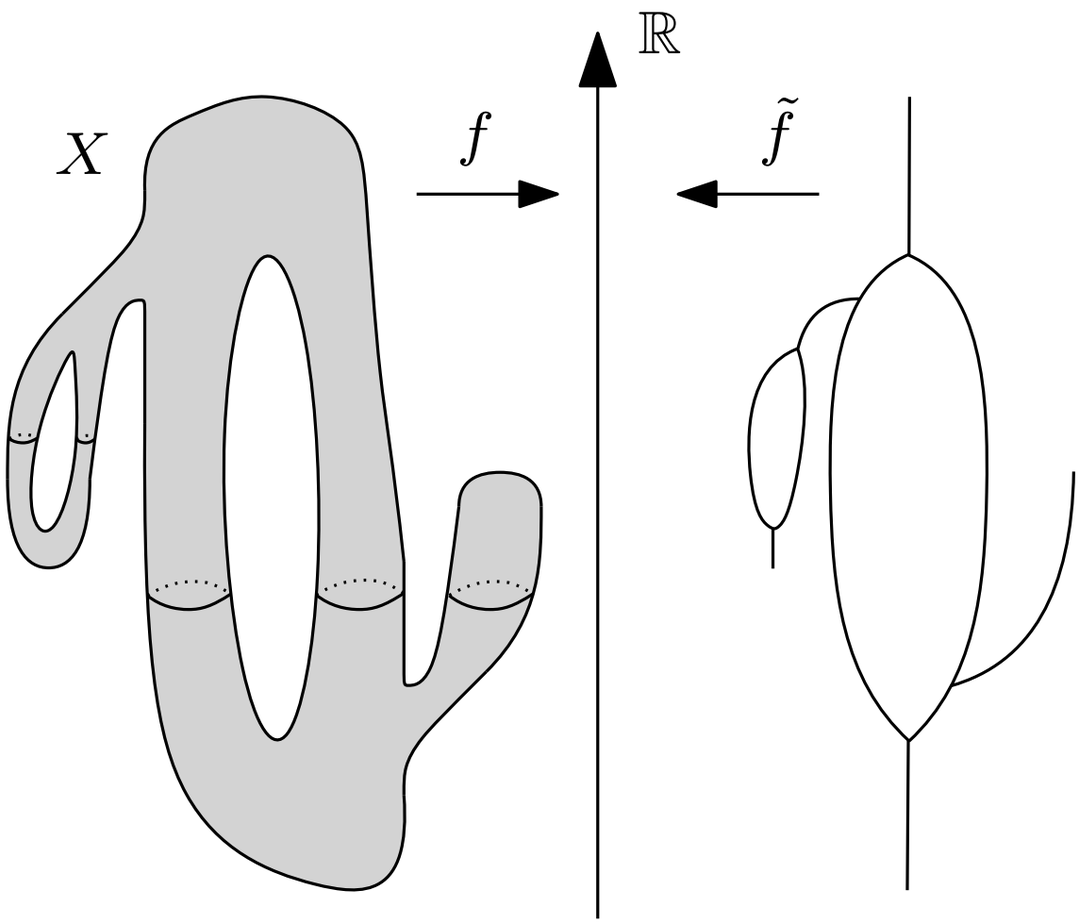


Def: A homology class $\omega \in H_k(X)$ is *horizontal w.r.t.* f, written $\omega \in \bar{H}_k(X)$, if it contains a representative cycle on which the filter is constant.

The Reeb graph and the space

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

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



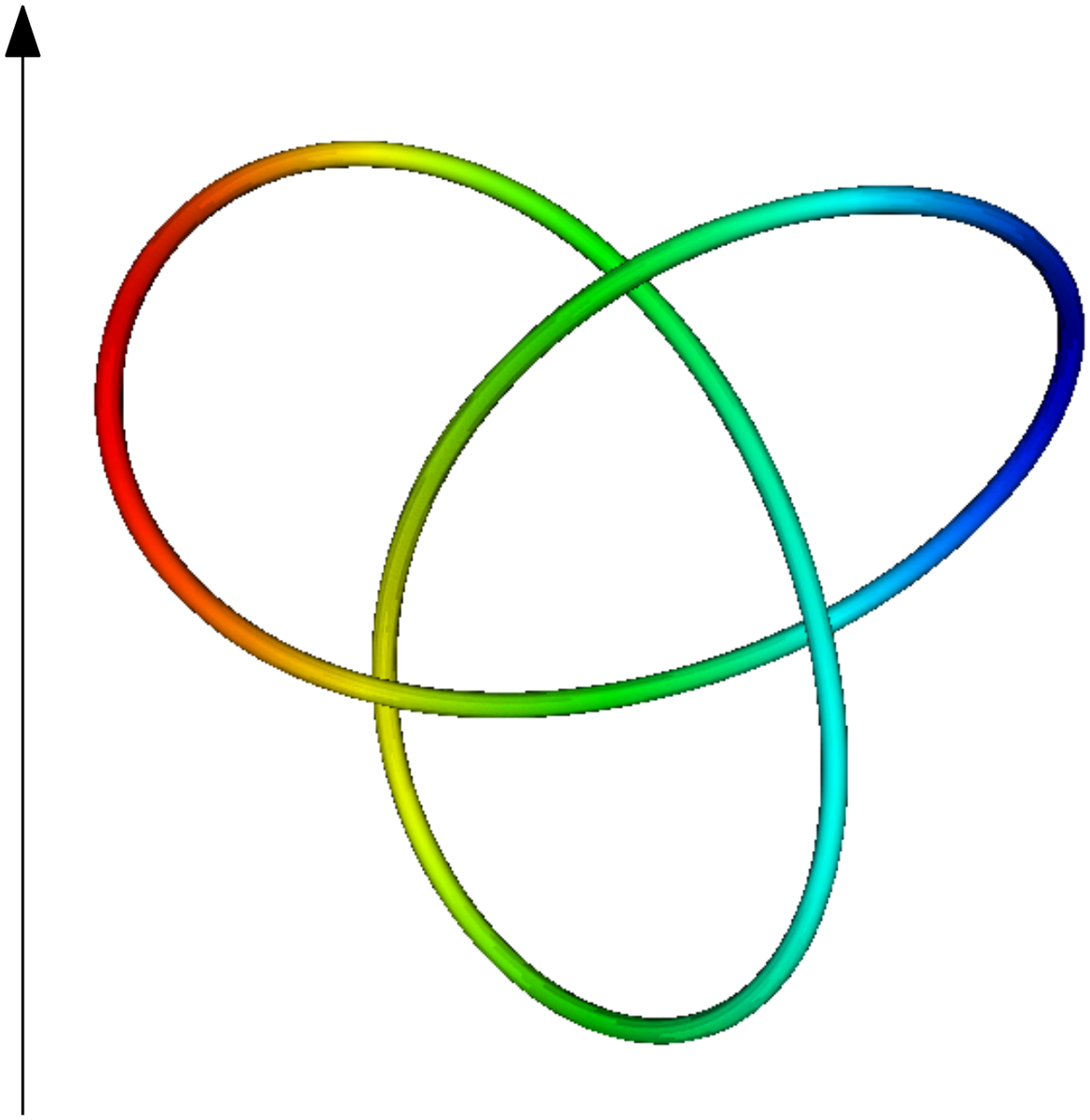
Def: A homology class $\omega \in H_k(X)$ is *horizontal w.r.t.* f, written $\omega \in \bar{H}_k(X)$, if it contains a representative cycle on which the filter is constant.

Th: $H_k(\mathrm{R}_f(X)) \sim H_k(X)/\bar{H}_k(X)$

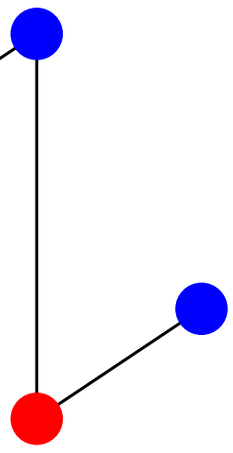
The homology groups of the Reeb graph are the ones of X minus the horizontal groups.

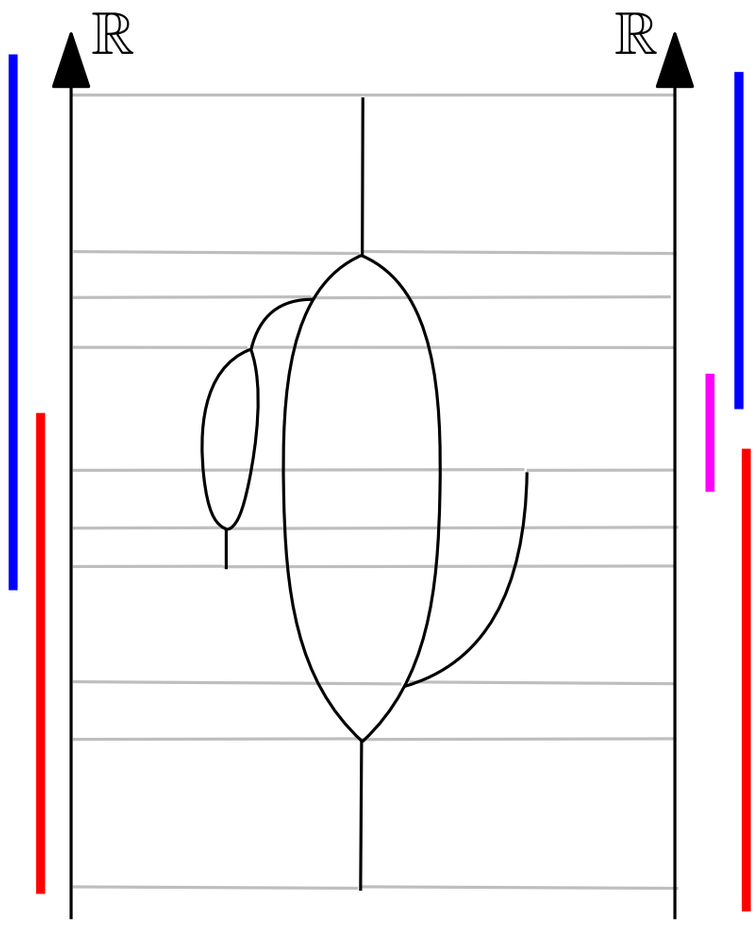
The Reeb graph and the space

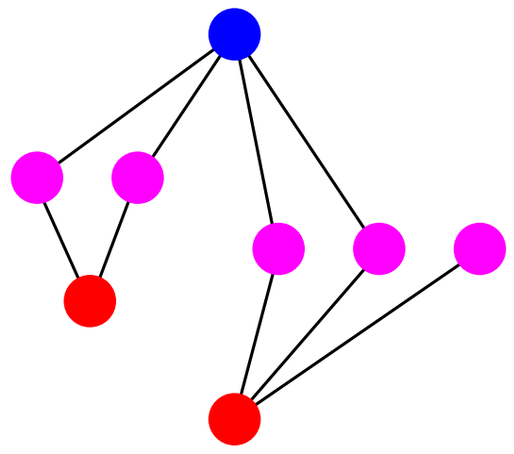
Q: What is the Reeb graph of the height function on the trefoil knot?



The Reeb graph and Mapper



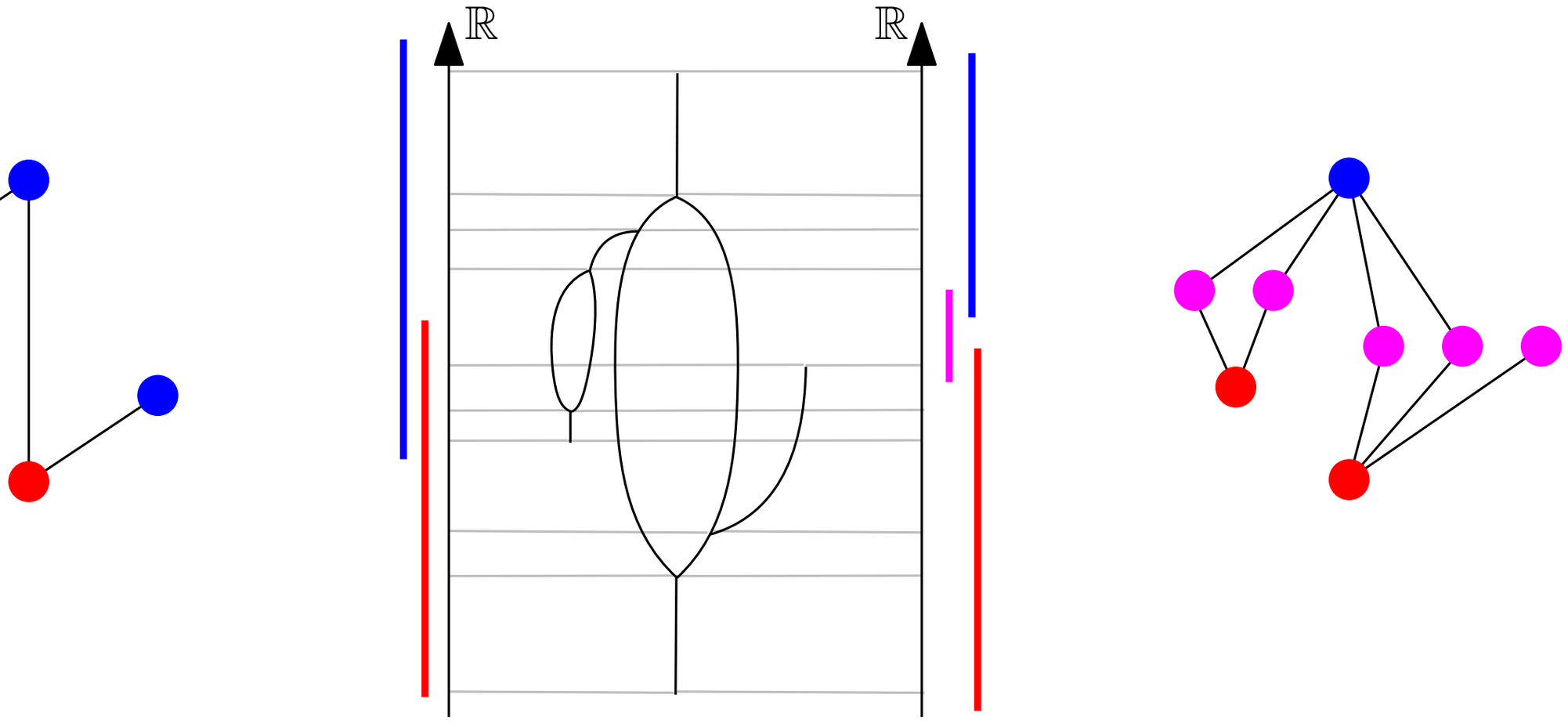




(continuous) Mapper $\equiv$ *pixelized* Reeb graph

The Reeb graph and Mapper

Homology of Mapper $\sim$ Homology of Reeb graph minus ???

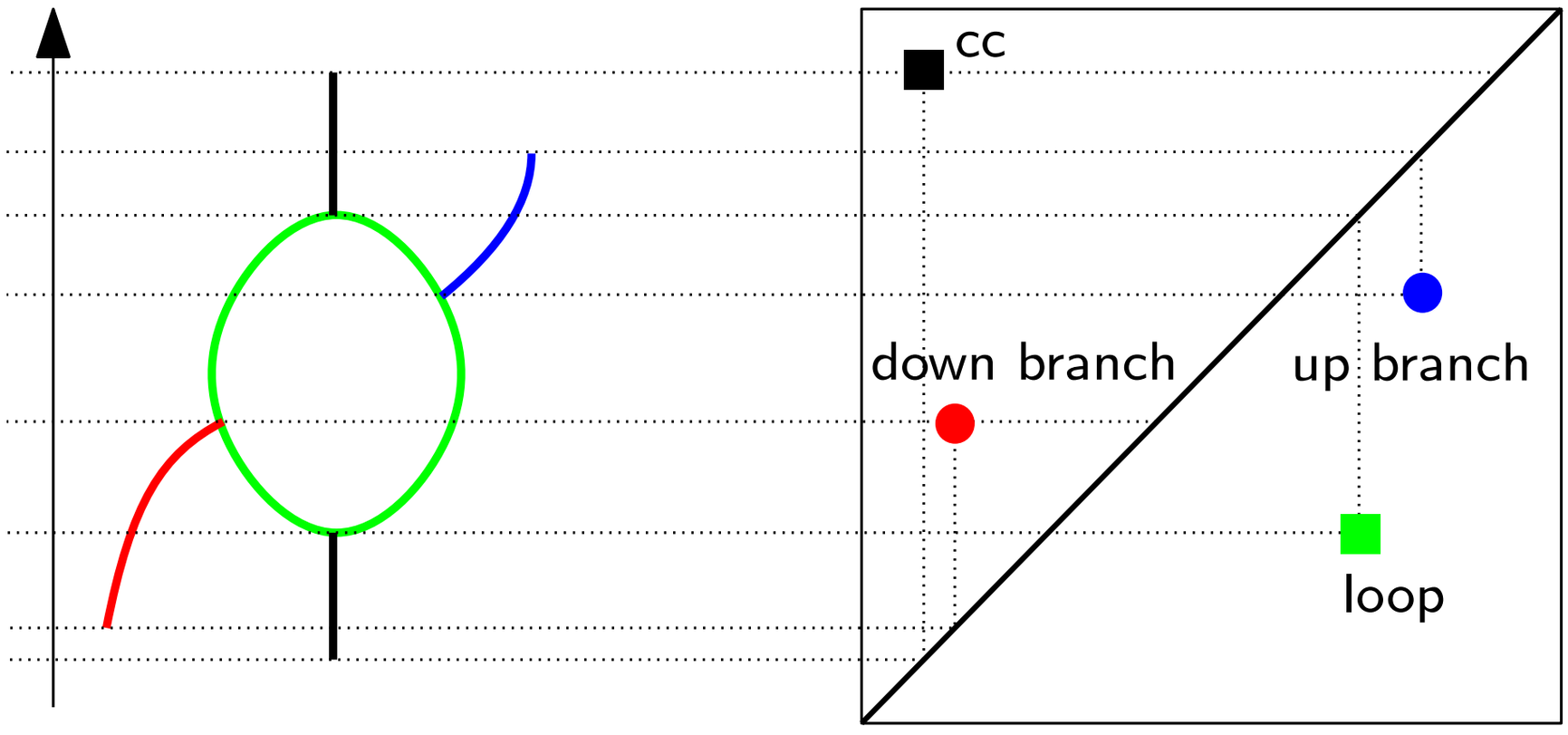


(continuous) Mapper $\equiv$ *pixelized* Reeb graph

The Reeb graph and Mapper

Homology of Mapper $\sim$ Homology of Reeb graph minus ???

Characterization through *topological bag-of-word* (TBOW).

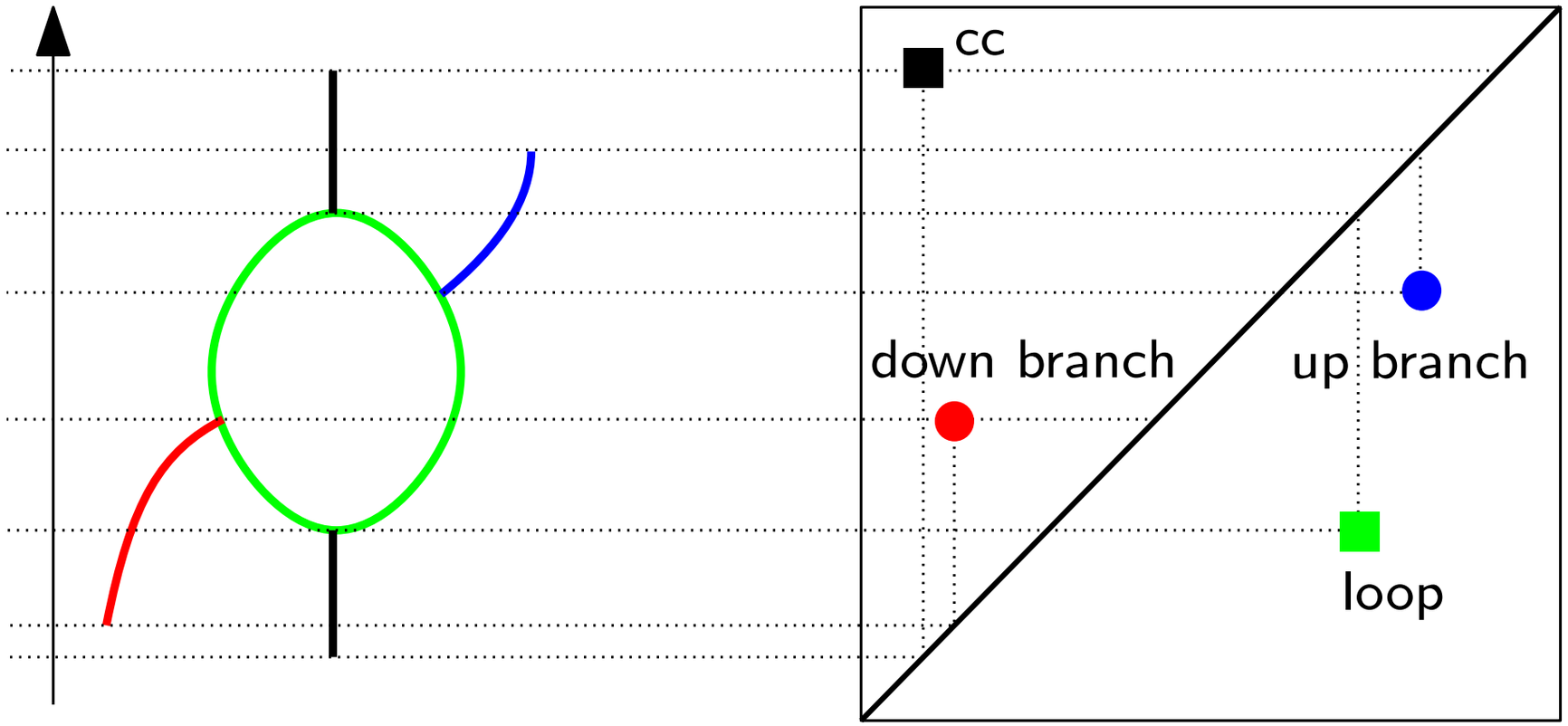


Points are homology classes, coordinates given by optimal representative cycles.

The Reeb graph and Mapper

Homology of Mapper $\sim$ Homology of Reeb graph minus ???

Characterization through *topological bag-of-word* (TBOW).

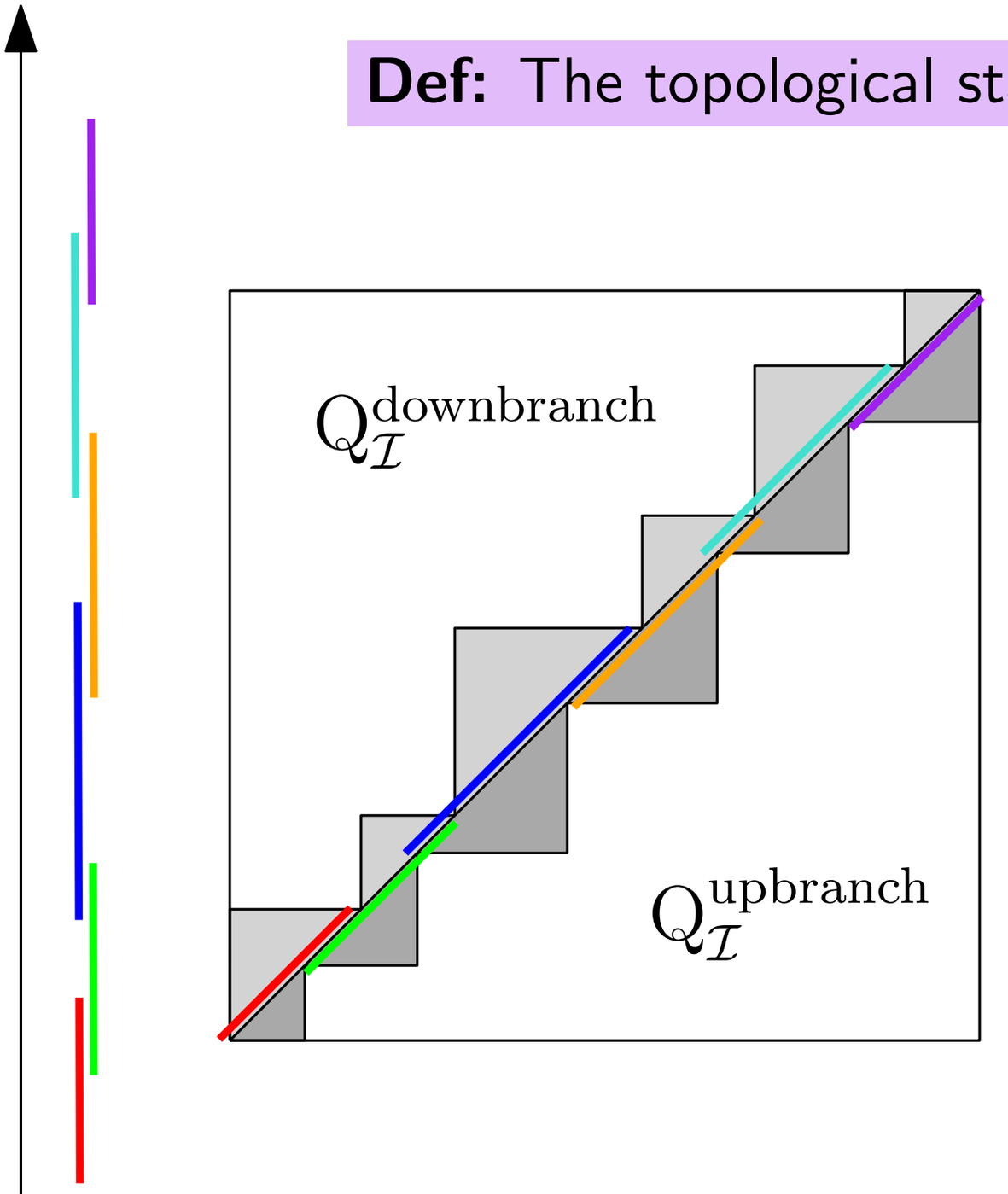


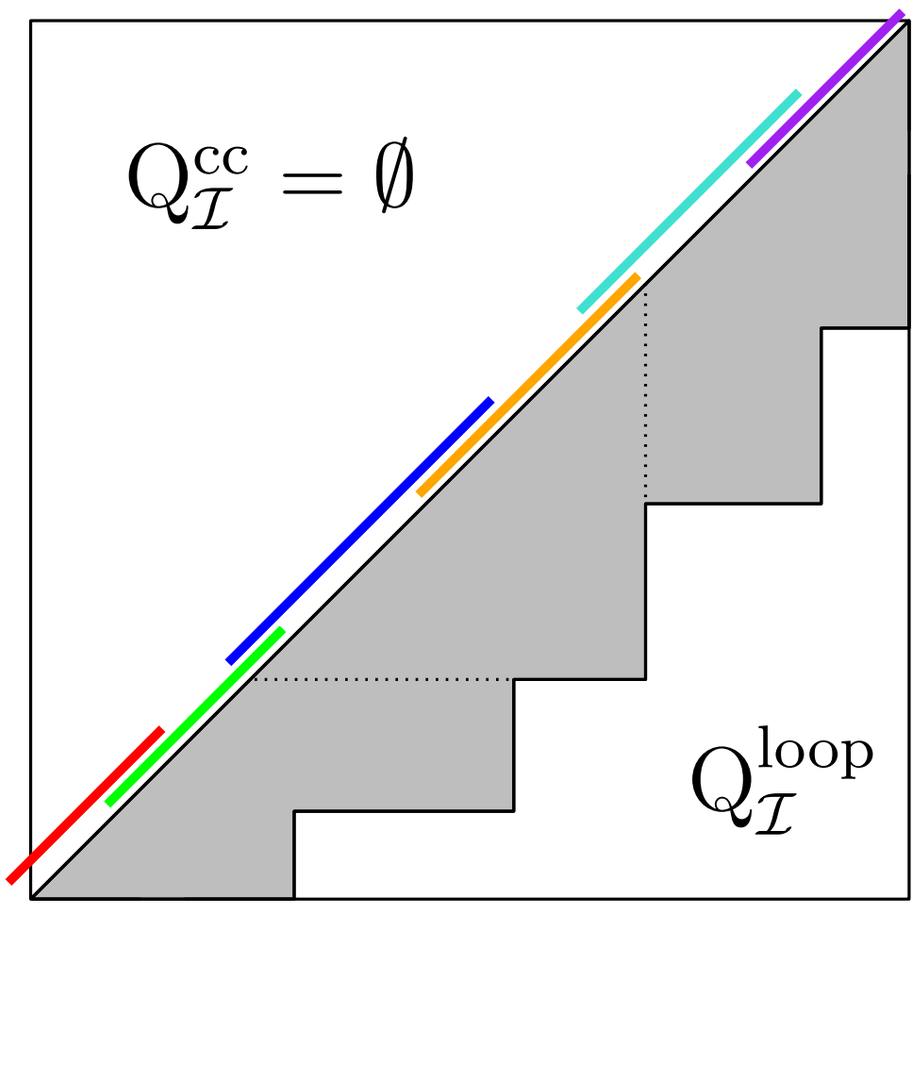
TBOW of Mapper = TBOW of Reeb graph *minus some points.*

Points are homology classes, coordinates given by optimal representative cycles.

The Reeb graph and Mapper

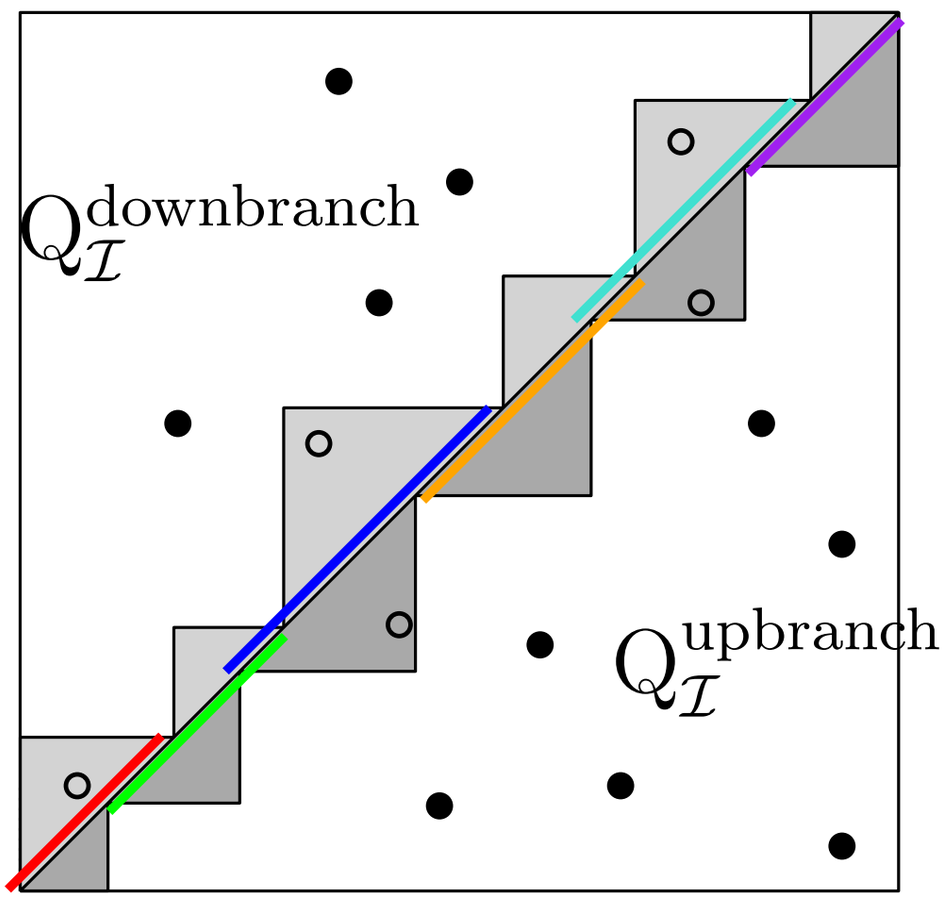
Def: The topological staircases of a cover $\mathcal{I}$ are:

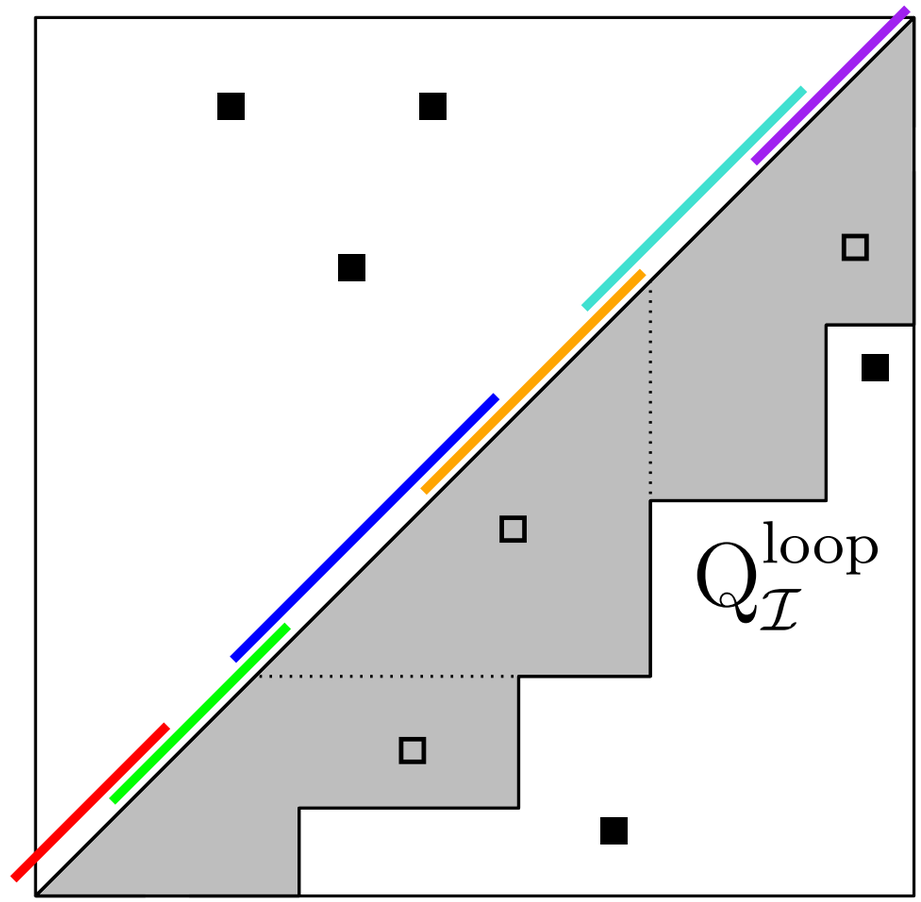




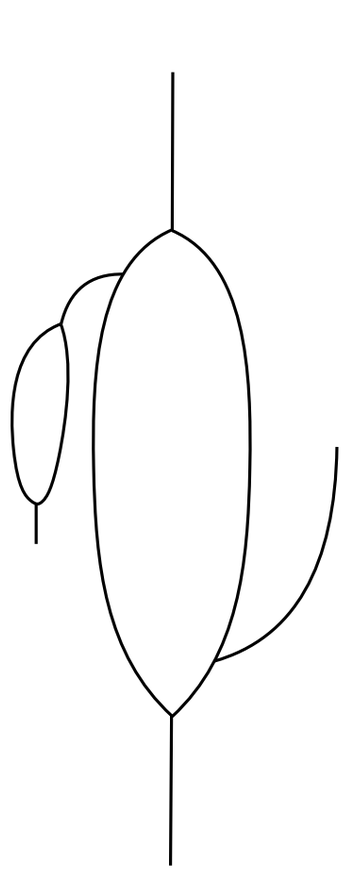
The Reeb graph and Mapper

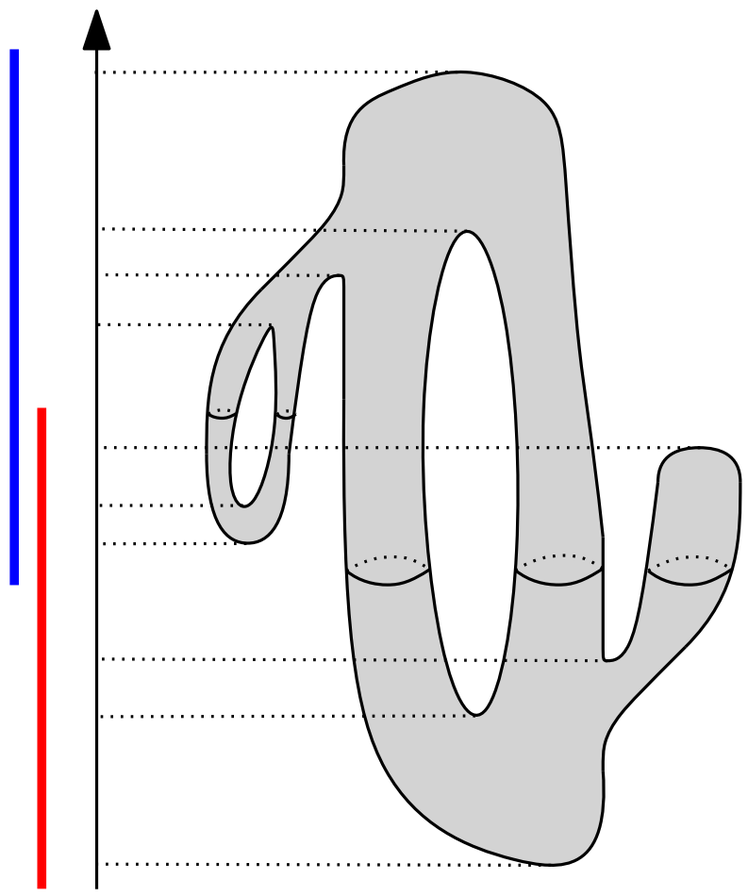
Thm: TBOW of (cont.) Mapper $\sim$ TBOW of Reeb graph minus the points inside the topological staircases.

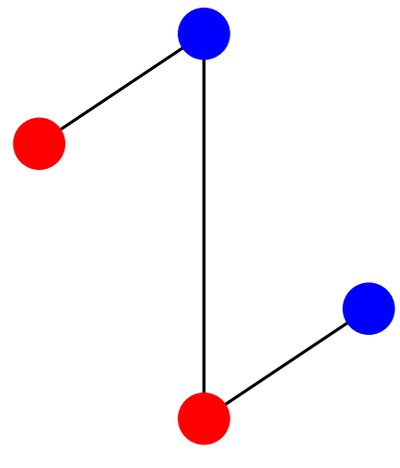


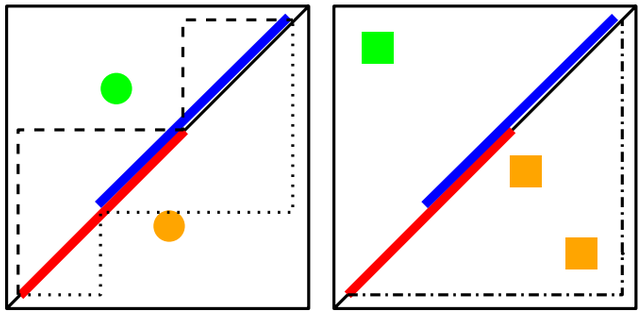


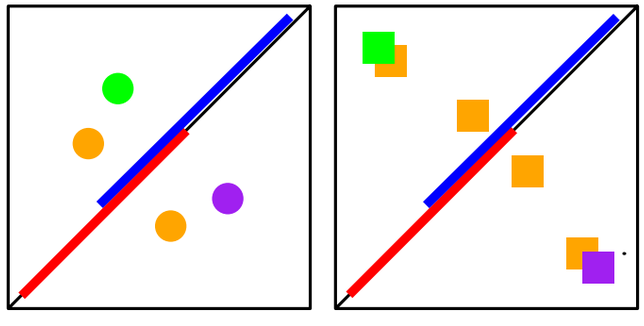
The Reeb graph and Mapper

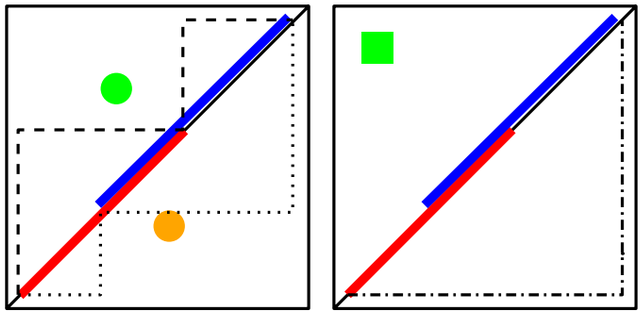






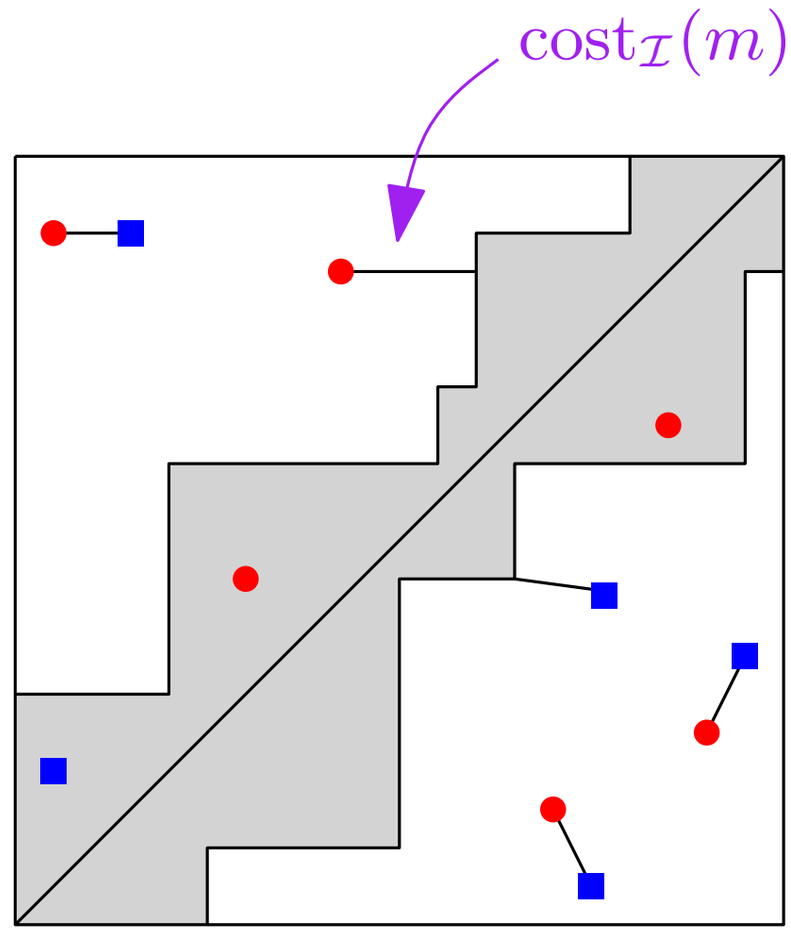






Fixing a first source of instability

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



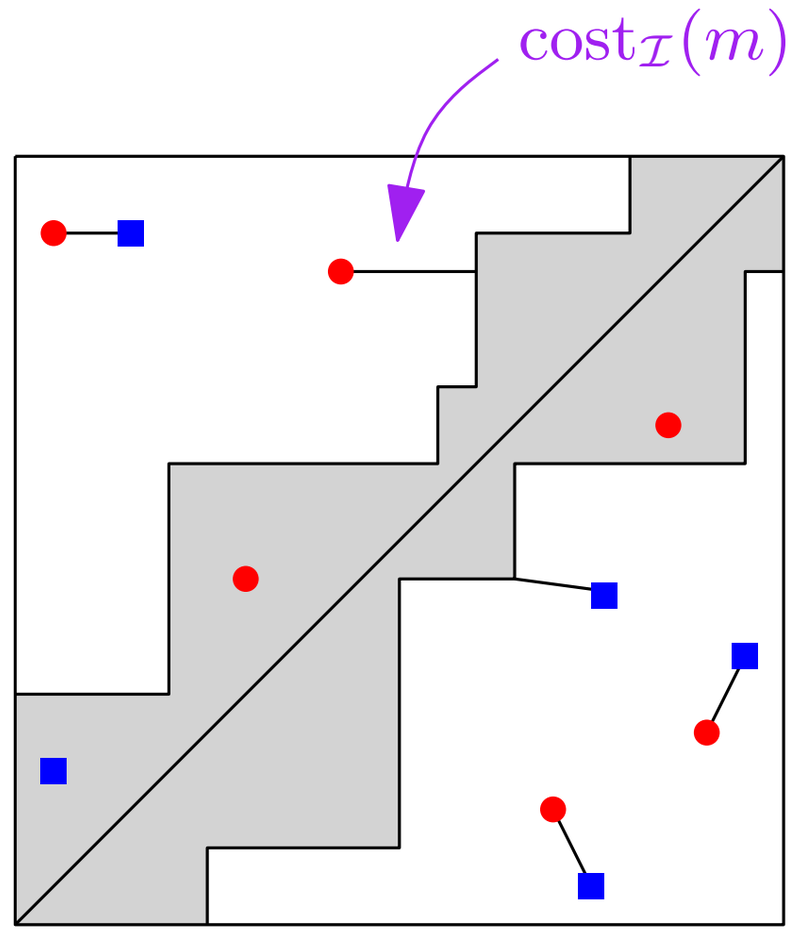
$m : \mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{R}_f(X)$

Fixing a first source of instability

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

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

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



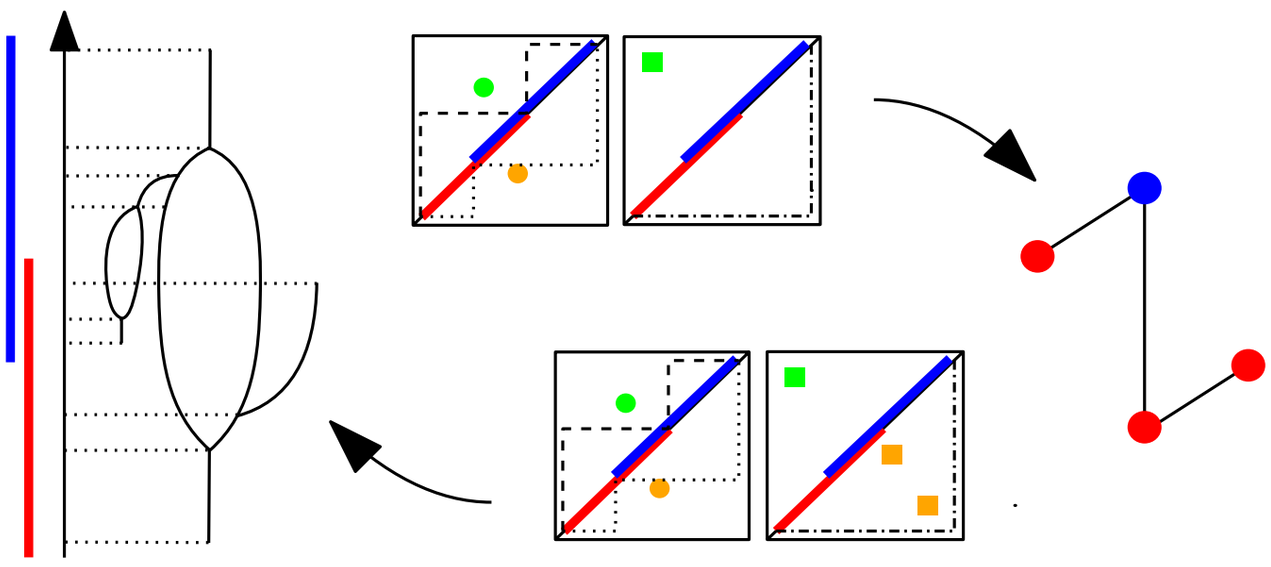
$m : \mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{R}_f(X)$

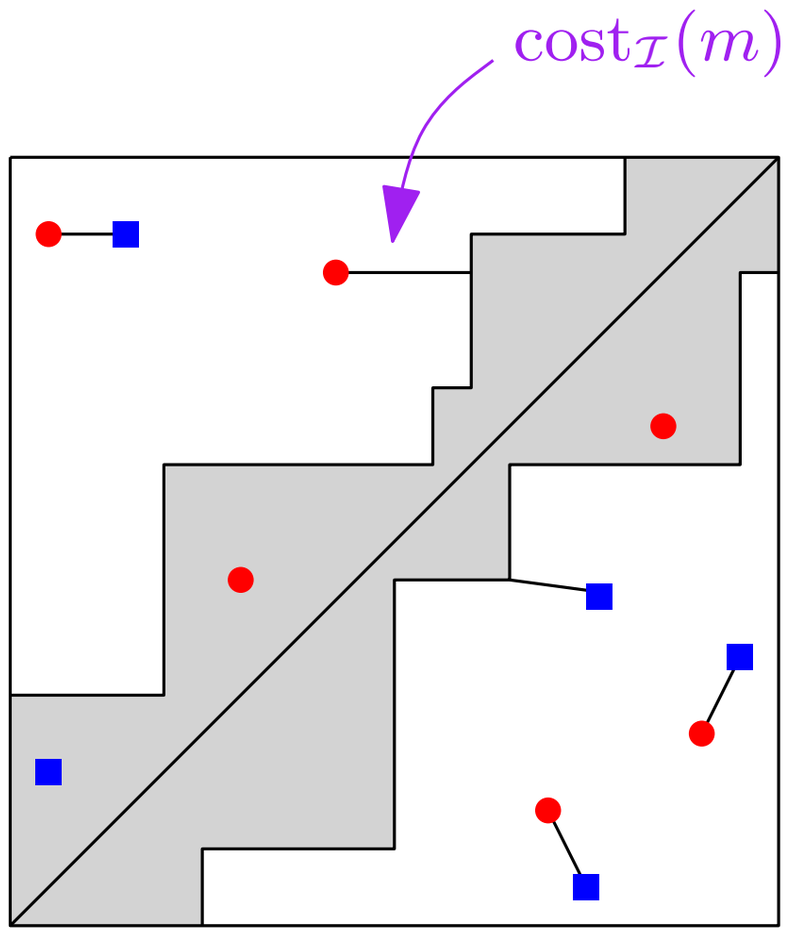
Fixing a first source of instability

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

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

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





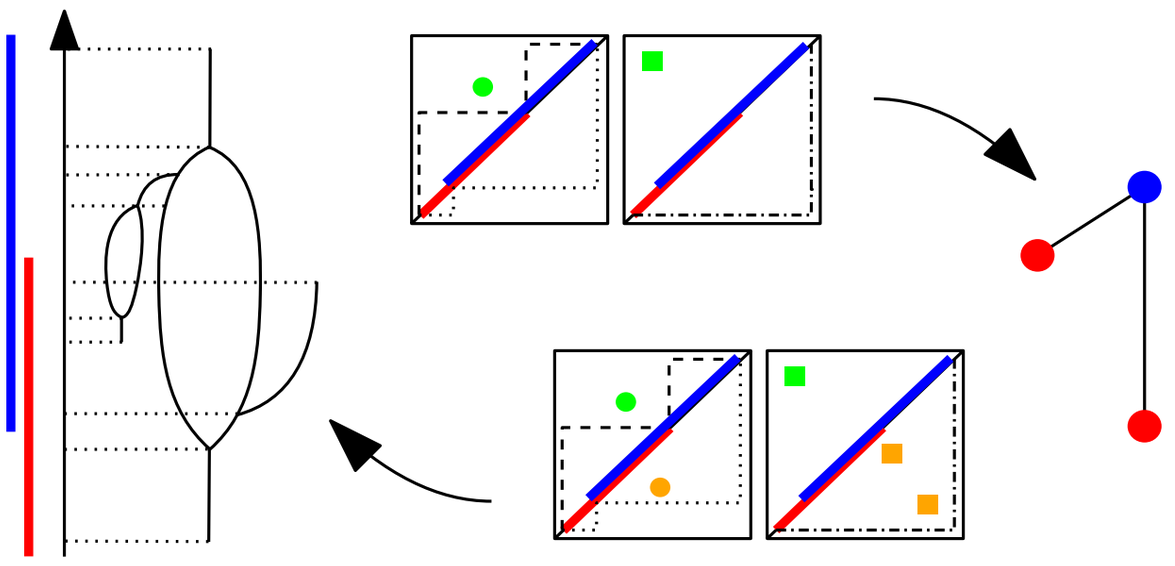
$m : \mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{R}_f(X)$

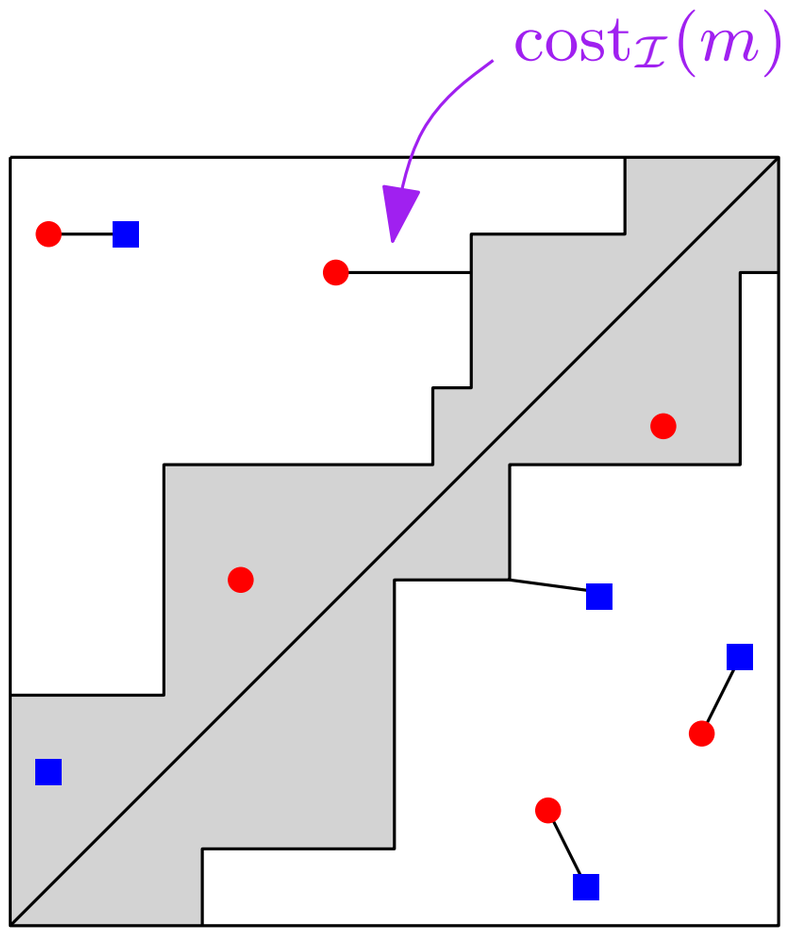
Fixing a first source of instability

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

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

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





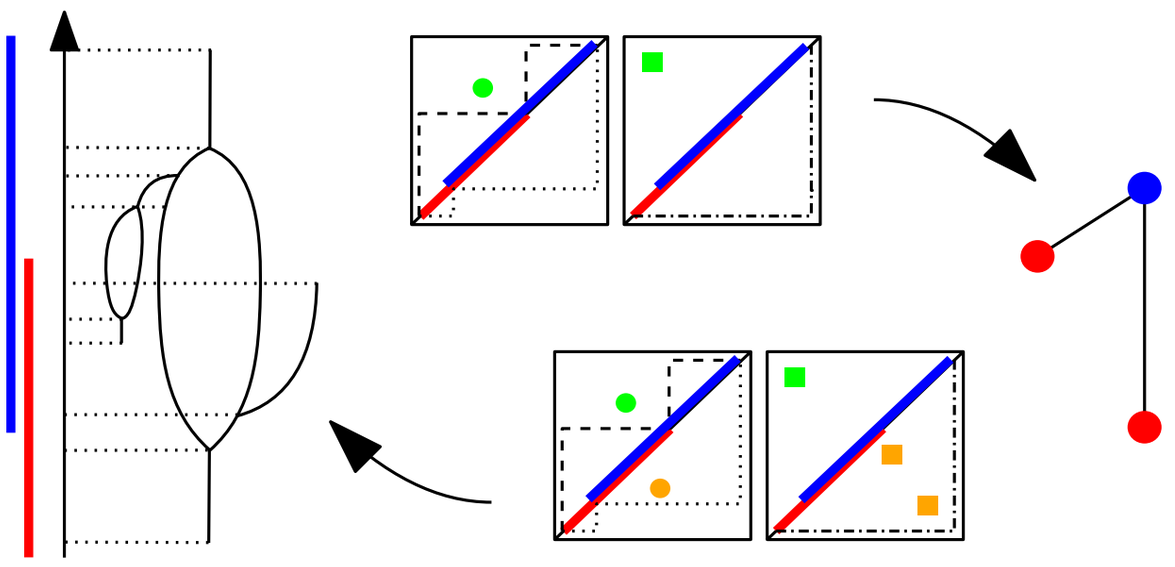
$m : \mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{R}_f(X)$

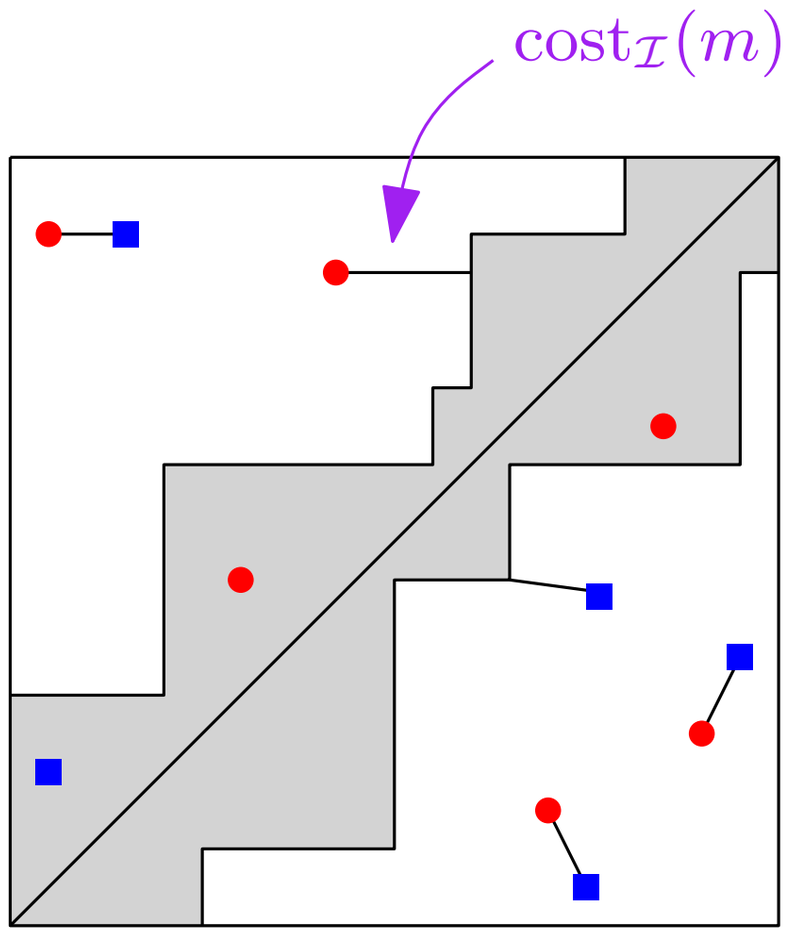
Fixing a first source of instability

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

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

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





$m : \mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{R}_f(X)$

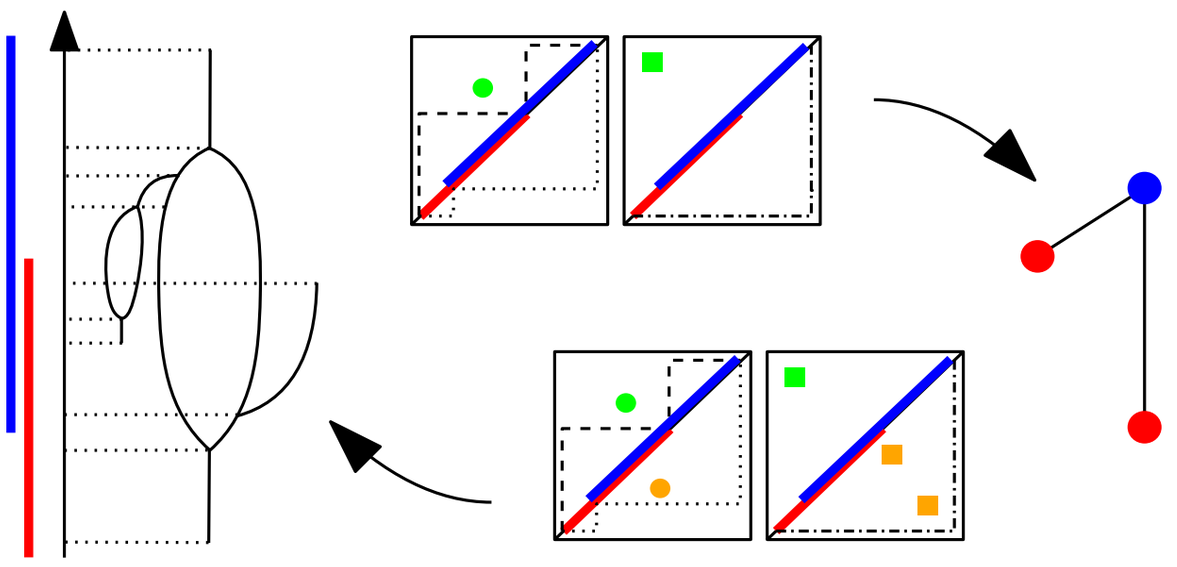
Cor: $\mathrm{M}_f(X,\mathcal{I})$ and $\mathrm{R}_f(X)$ have the same homology groups whenever the resolution r of $\mathcal{I}$ is small enough

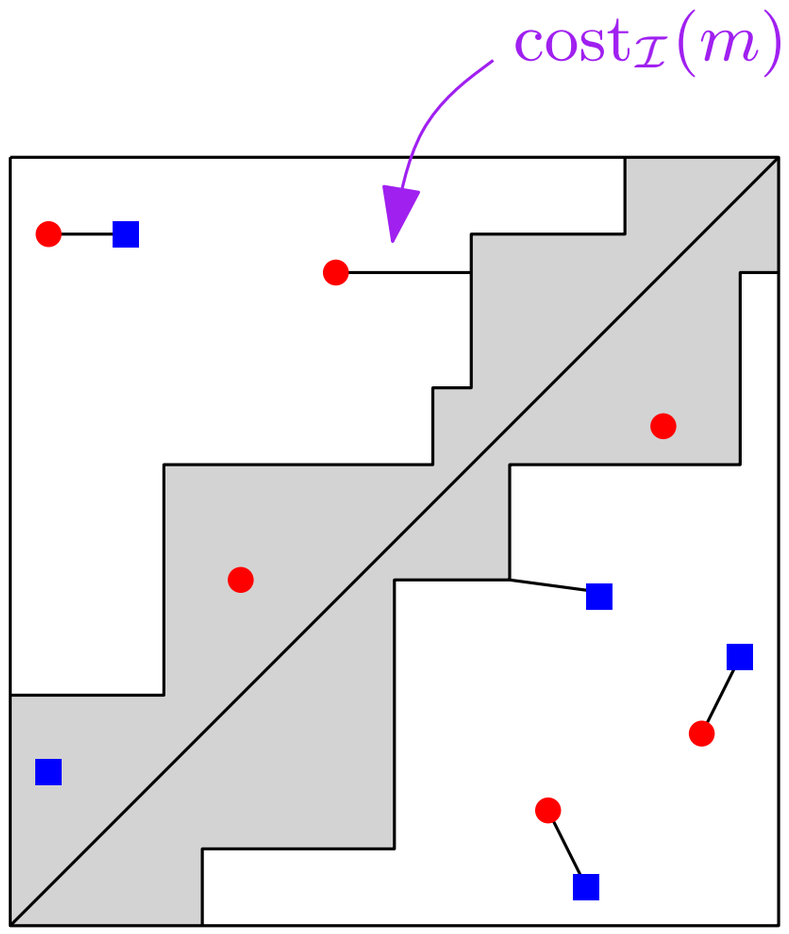
Fixing a first source of instability

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

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

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

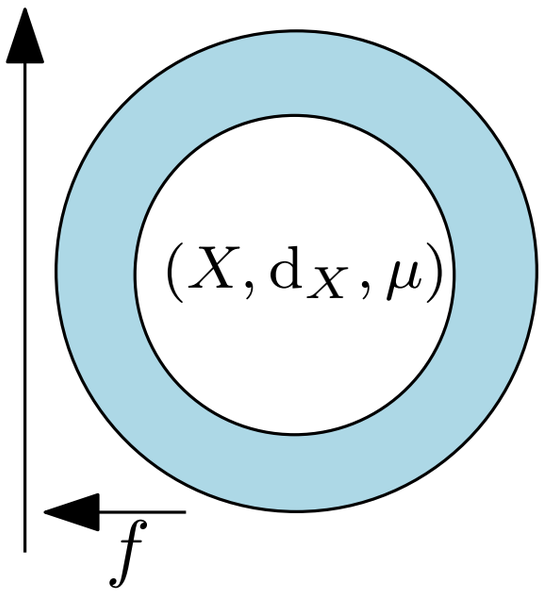




$m : \mathrm{M}_f(X,\mathcal{I}) \longleftrightarrow \mathrm{R}_f(X)$

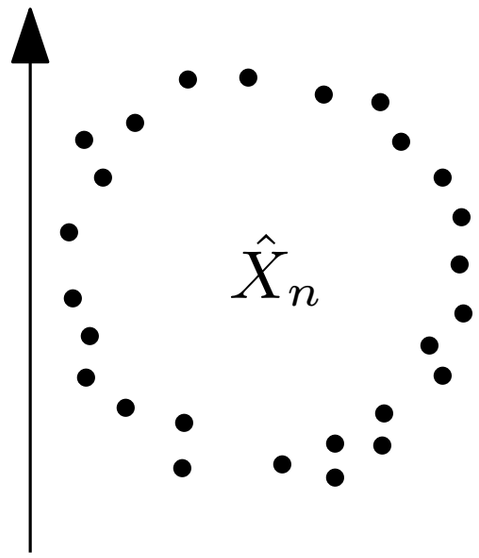
Cor: $\mathrm{M}_f(X,\mathcal{I})$ and $\mathrm{R}_f(X)$ have the same homology groups whenever the resolution r of $\mathcal{I}$ is small enough (smaller than the smallest distance from the TBOW of $\mathrm{R}_f(X)$ to the diagonal Δ).

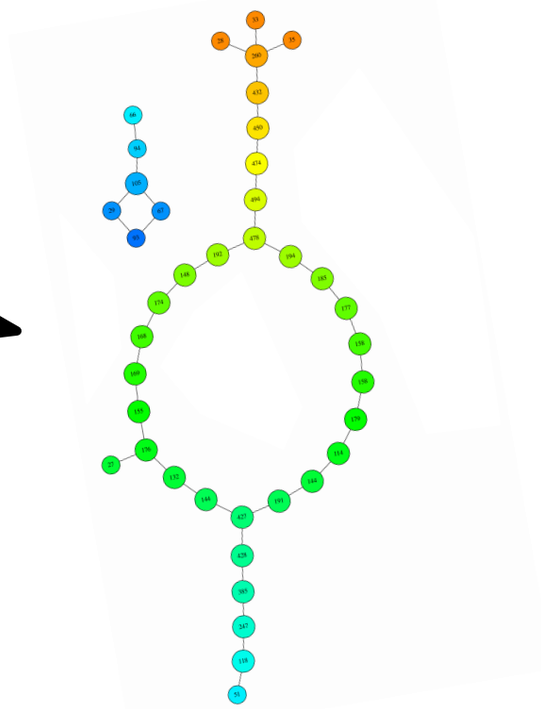
Understanding the Mapper structure



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

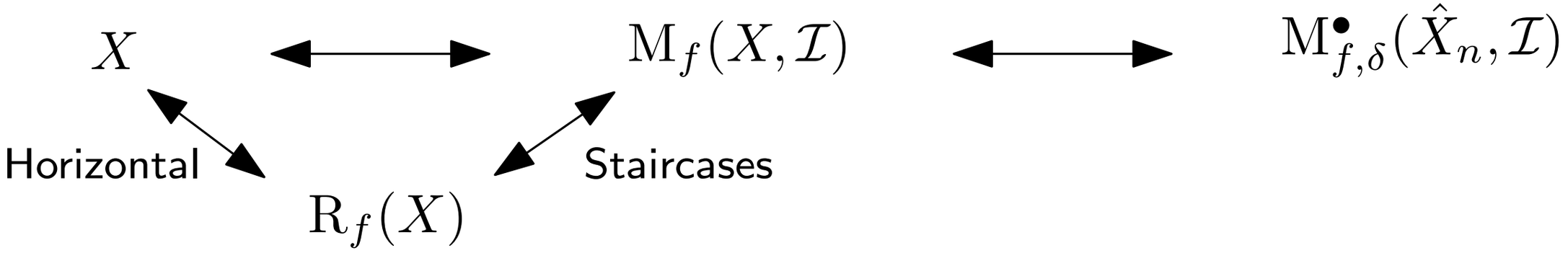
$+$ cover $\mathcal{I}$



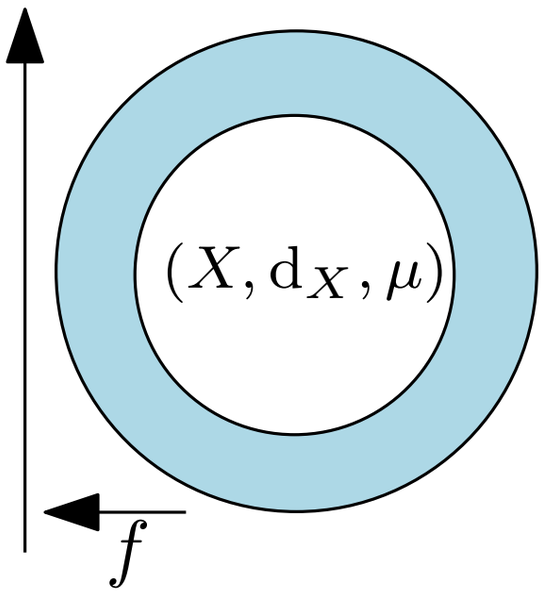


$\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$

We can decompose the relation between Mapper and X with

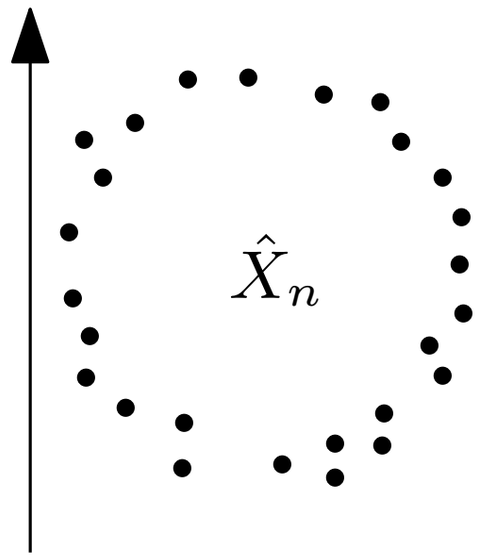


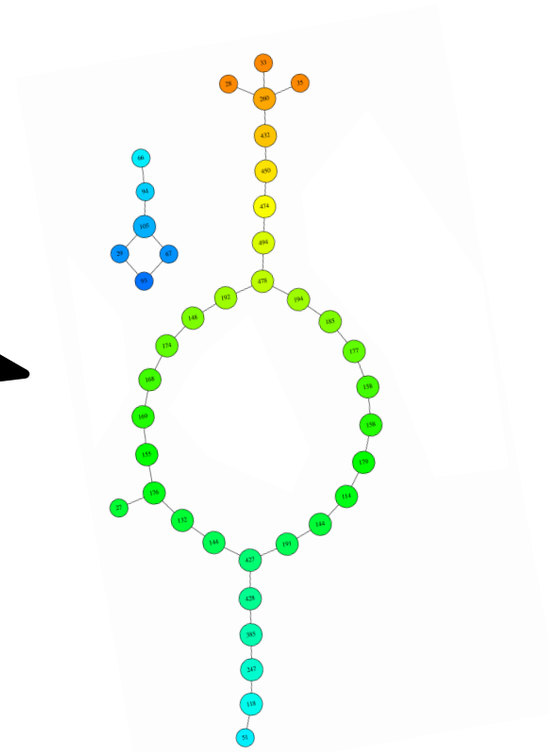
Understanding the Mapper structure



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

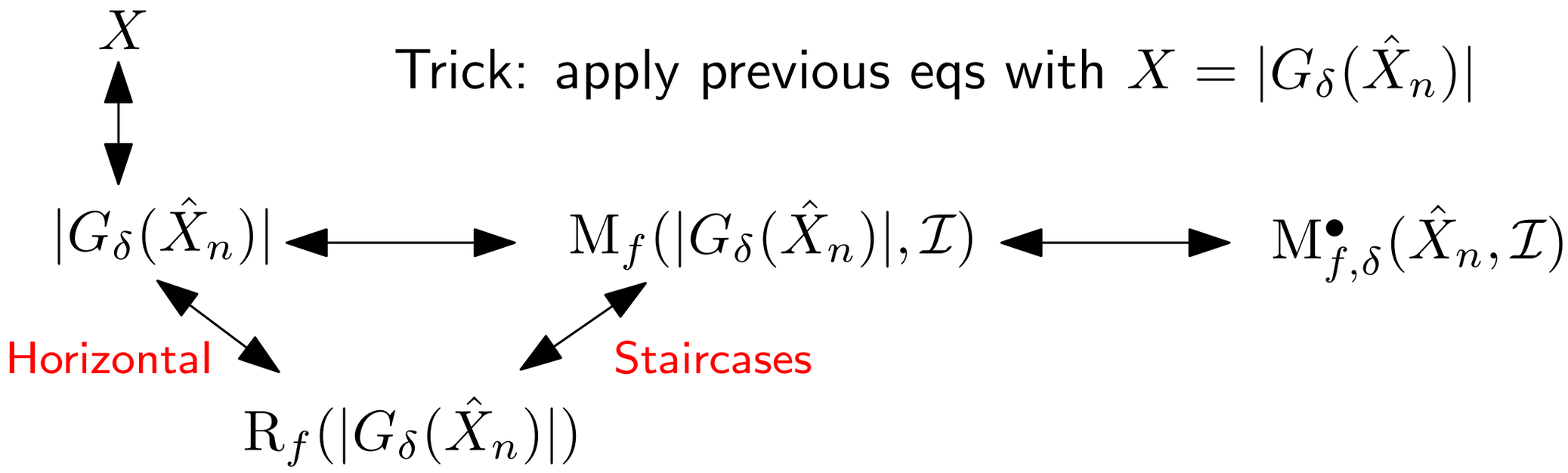
$+$ cover $\mathcal{I}$



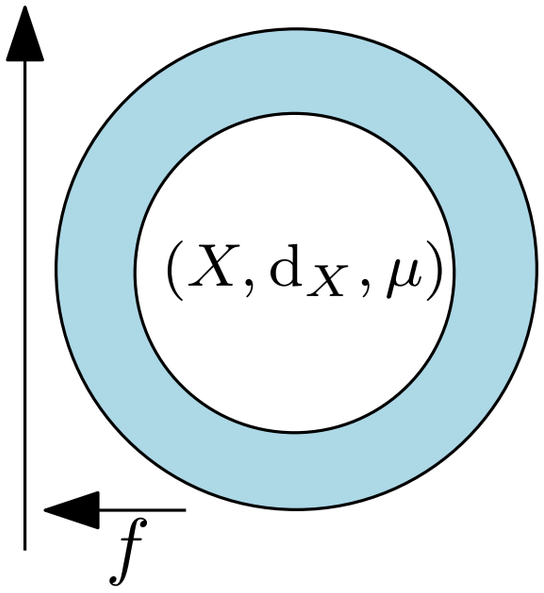


$\mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I})$

We can decompose the relation between Mapper and X with

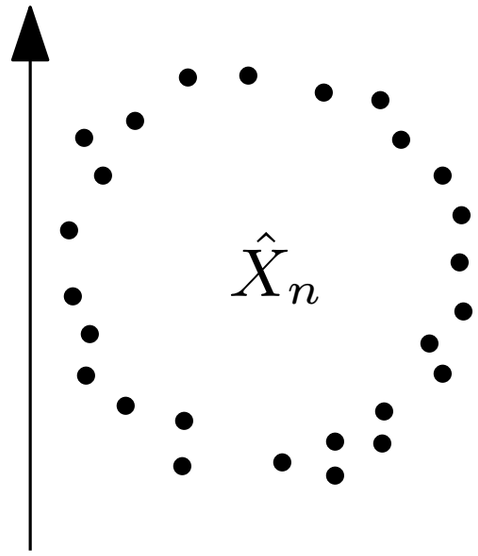


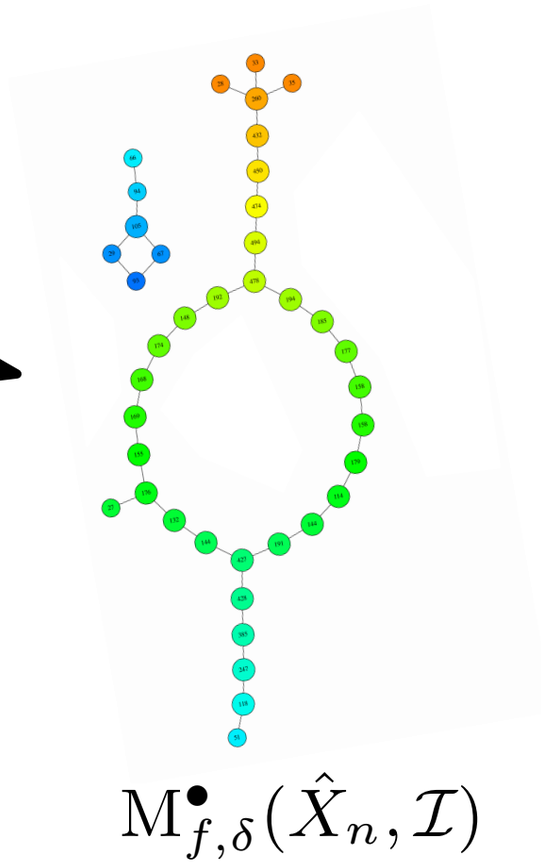
Understanding the Mapper structure



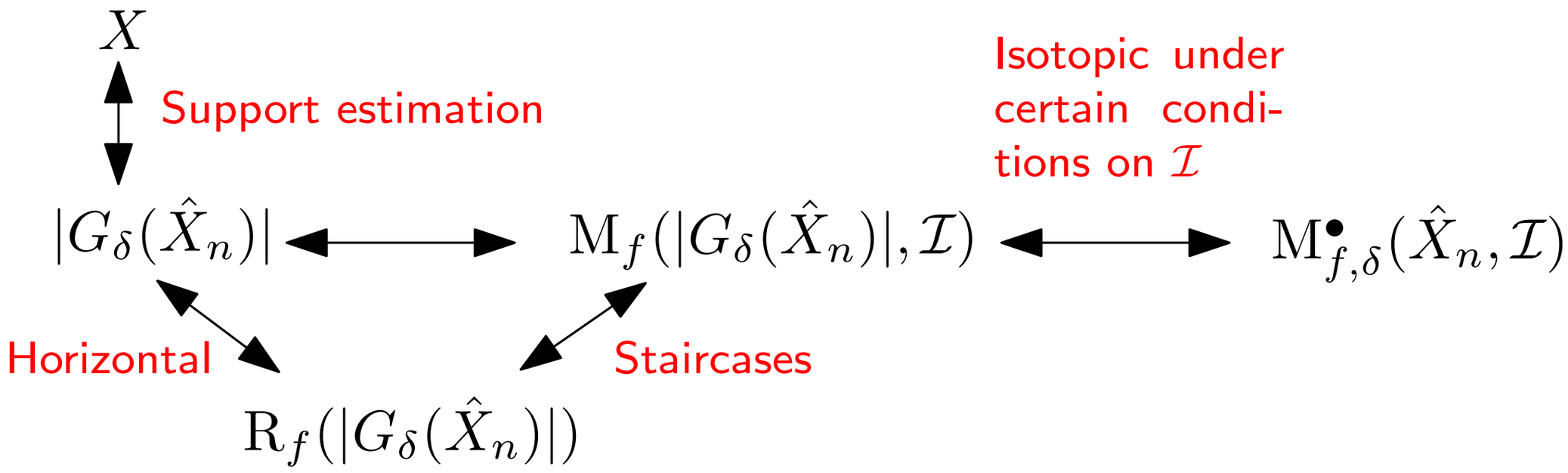
n points sampled i.i.d. according to μ.

$+$ cover $\mathcal{I}$

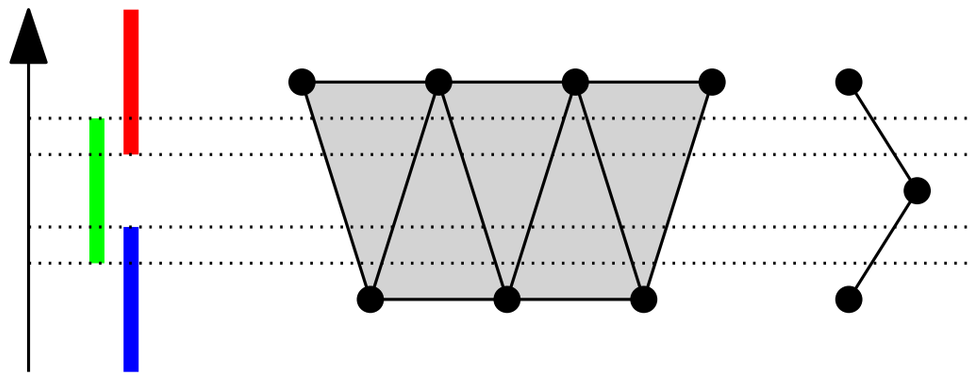




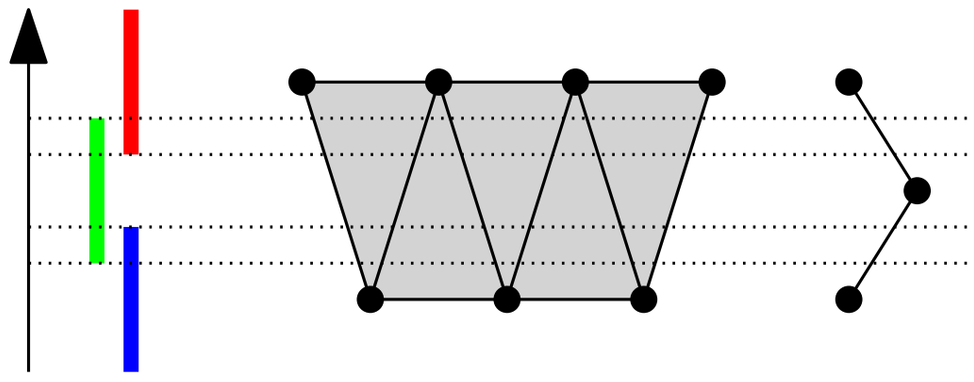
We can decompose the relation between Mapper and X with

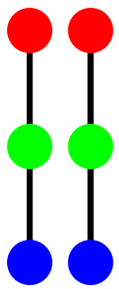


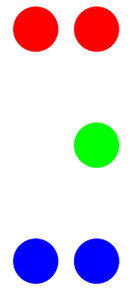
Intersection-crossing edges



Intersection-crossing edges





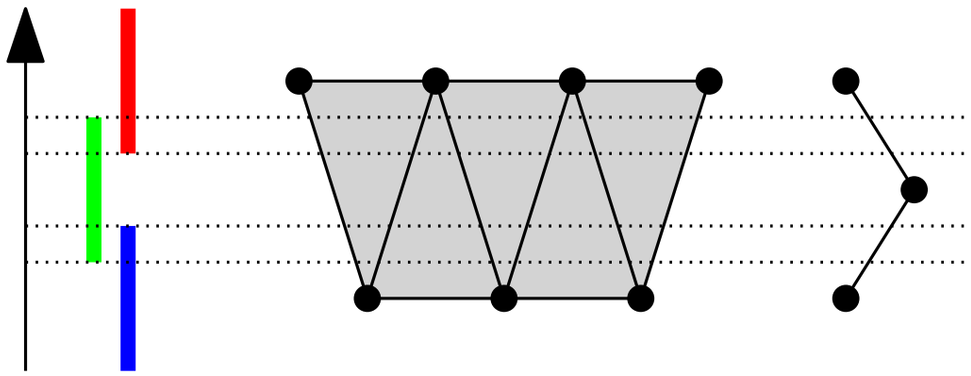


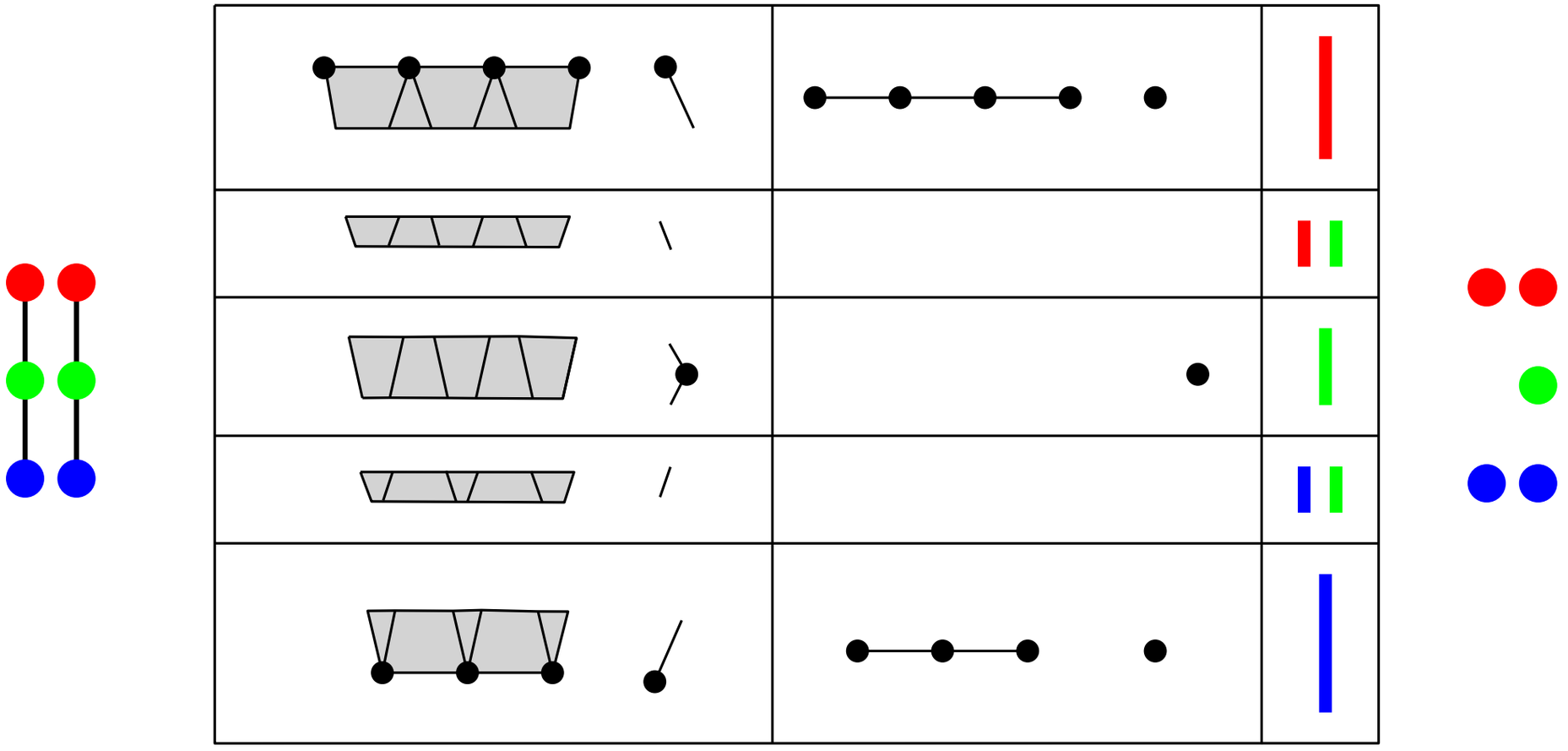
Intersection-crossing edges

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

$$\mathrm{M}_f(|G_\delta(\hat{X}_n)|, \mathcal{I}) \sim \mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I}).$$

Q: Prove it.





Support estimation

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

$$\mathrm{M}_f(|G_\delta(\hat{X}_n)|, \mathcal{I}) \sim \mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I}).$$

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

the homology groups of $\mathrm{R}_f(X)$ and $\mathrm{R}_f(|G_\delta(\hat{X}_n)|)$
only differ by classes of size less than $2c\delta$.

Support estimation

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

$$\mathrm{M}_f(|G_\delta(\hat{X}_n)|, \mathcal{I}) \sim \mathrm{M}^{\bullet}_{f,\delta}(\hat{X}_n, \mathcal{I}).$$

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

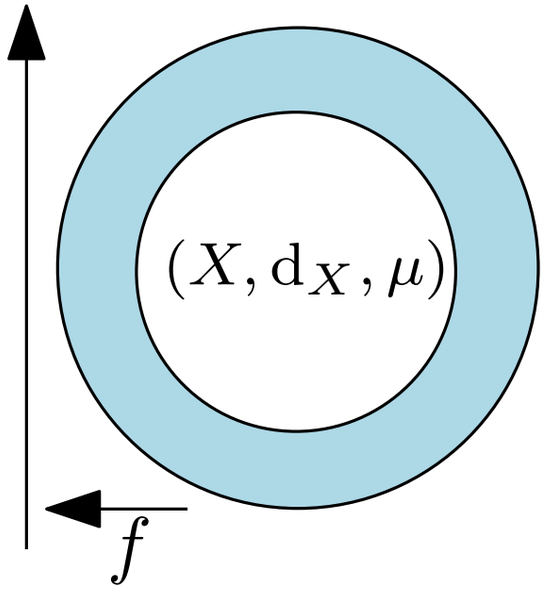
the homology groups of $\mathrm{R}_f(X)$ and $\mathrm{R}_f(|G_\delta(\hat{X}_n)|)$
only differ by classes of size less than $2c\delta$.

rch: reach of X — see later

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

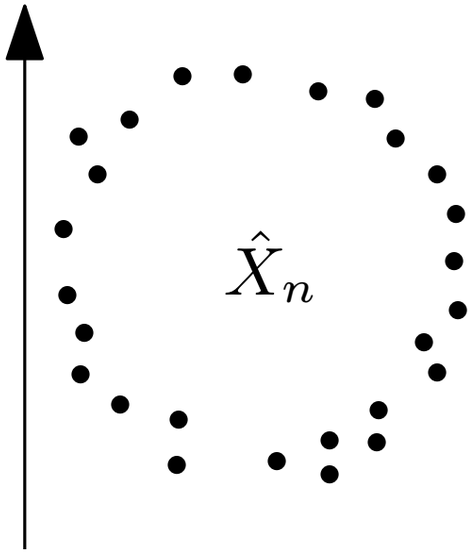
d_H: Hausdorff distance — see later

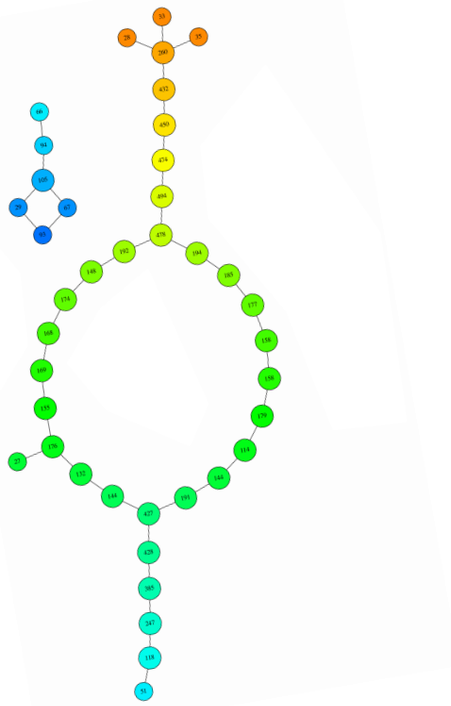
Understanding the Mapper structure



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

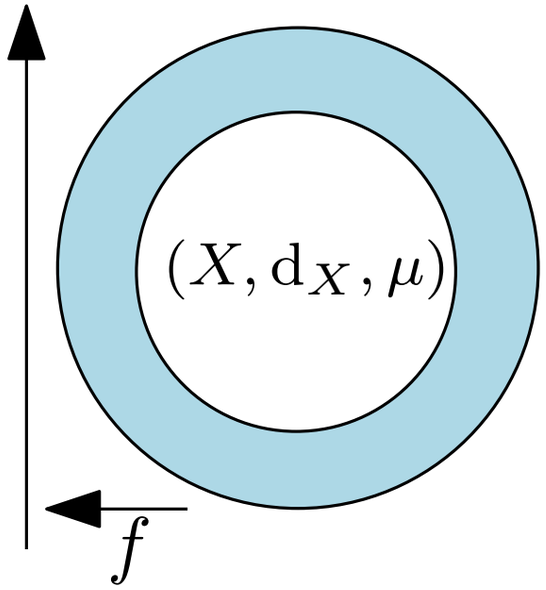
$+$ cover $\mathcal{I}$





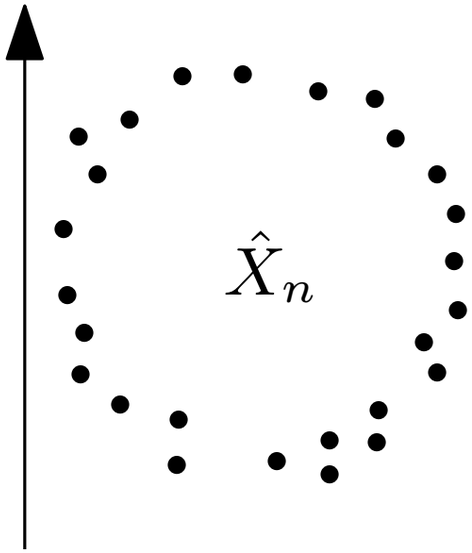
$\mathrm{M}^{\bullet}_{f,\delta}(X_n, \mathcal{I})$

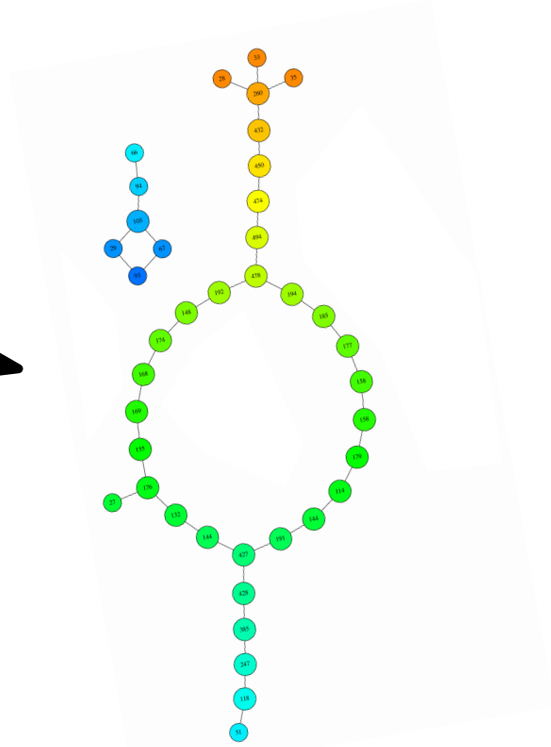
Understanding the Mapper structure



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

$+$ cover $\mathcal{I}$





$\mathrm{M}^{\bullet}_{f,\delta}(X_n, \mathcal{I})$

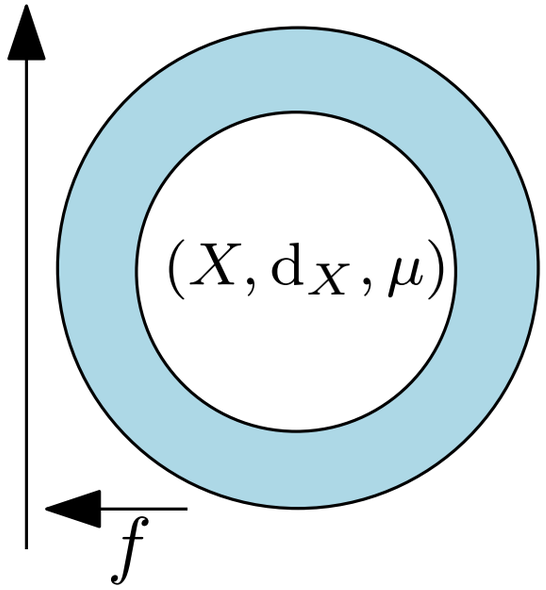
Let $\beta > 0$ and take $s(n) = \frac{n}{\log(n)^{1+\beta}}$.

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

Thm: If f is c-Lipschitz and $X \subseteq \mathbb{R}^d$, then for:

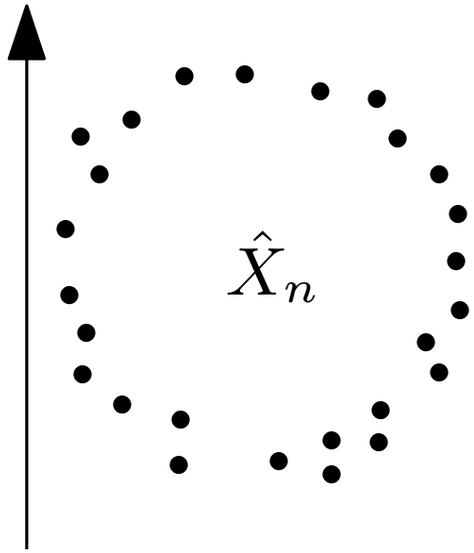
$$\delta_n = d_H(\hat{X}_n^{s(n)}, \hat{X}_n),\ g_n \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n},$$

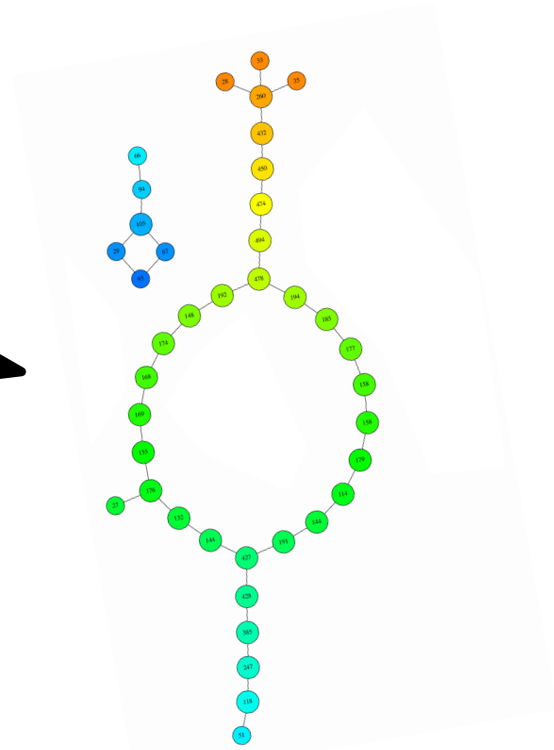
Understanding the Mapper structure



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

$+$ cover $\mathcal{I}$





$\mathrm{M}^{\bullet}_{f,\delta}(X_n, \mathcal{I})$

Let $\beta > 0$ and take $s(n) = \frac{n}{\log(n)^{1+\beta}}$.

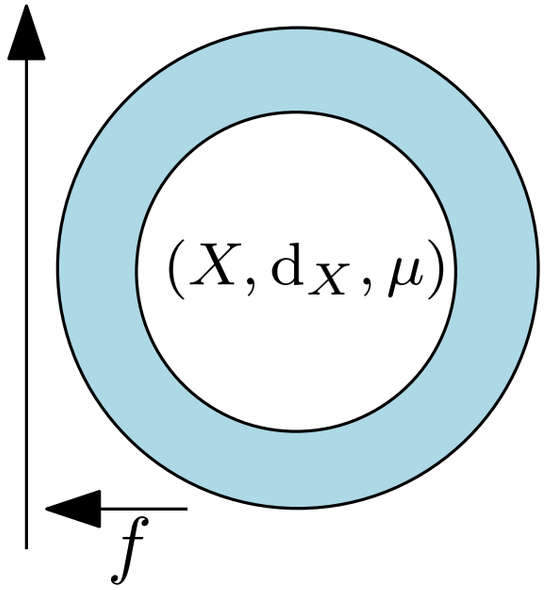
Let $\delta_n := d_H(\hat{X}_n^{s(n)}, \hat{X}_n)$ where $\hat{X}_n^{s(n)}$ is a subset of $\hat{X}_n$ of size $s(n)$.

Thm: If f is c-Lipschitz and $X \subseteq \mathbb{R}^d$, then for:

$$\delta_n = d_H(\hat{X}_n^{s(n)}, \hat{X}_n),\ g_n \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right),\ r_n = \frac{c\delta_n}{g_n},$$

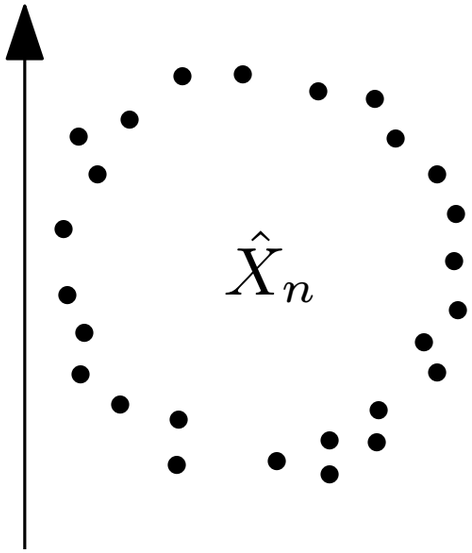
the homology groups of $\mathrm{M}^{\bullet}_{f,\delta_n}(\hat{X}_n, \mathcal{I}(g_n, r_n))$ and $\mathrm{R}_f(X)$ are isomorphic, with expected size difference less than $O\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/d}$.

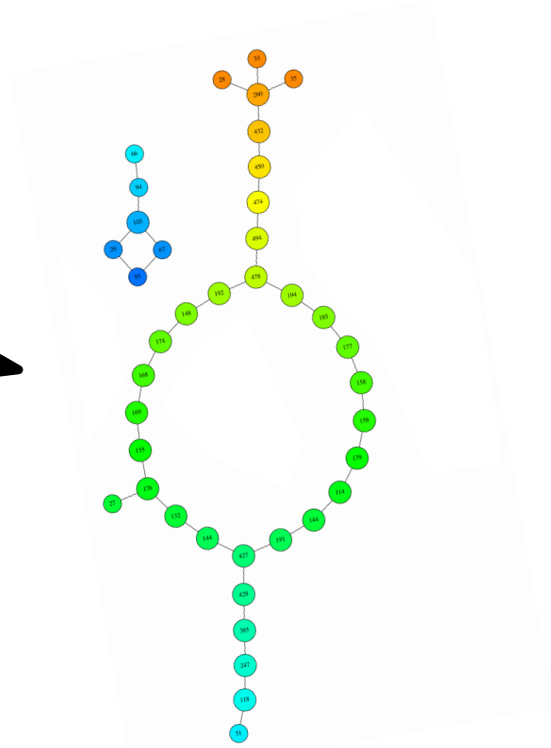
Understanding the Mapper structure



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

$+$ cover $\mathcal{I}$





$\mathrm{M}^{\bullet}_{f,\delta}(X_n, \mathcal{I})$

Let $\beta > 0$ and take $s(n) = \frac{n}{\log(n)^{1+\beta}}$.

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

Thm: If f is c-Lipschitz and $X \subseteq \mathbb{R}^d$, then for:

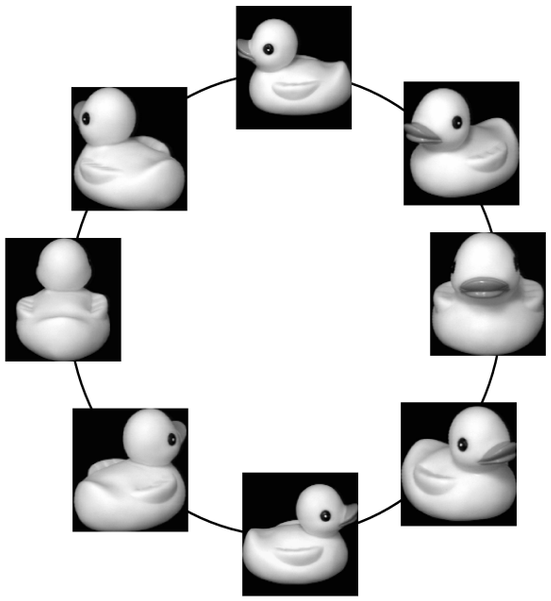
Can be used for confidence regions!

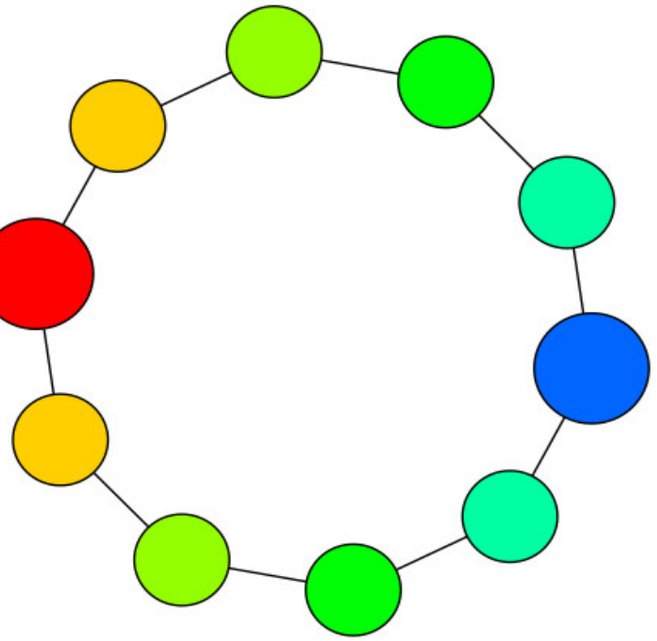
$$\delta_n = d_H(\hat{X}_n^{s(n)}, \hat{X}_n),\ g_n \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right),\ r_n = \tfrac{c\delta_n}{g_n},$$

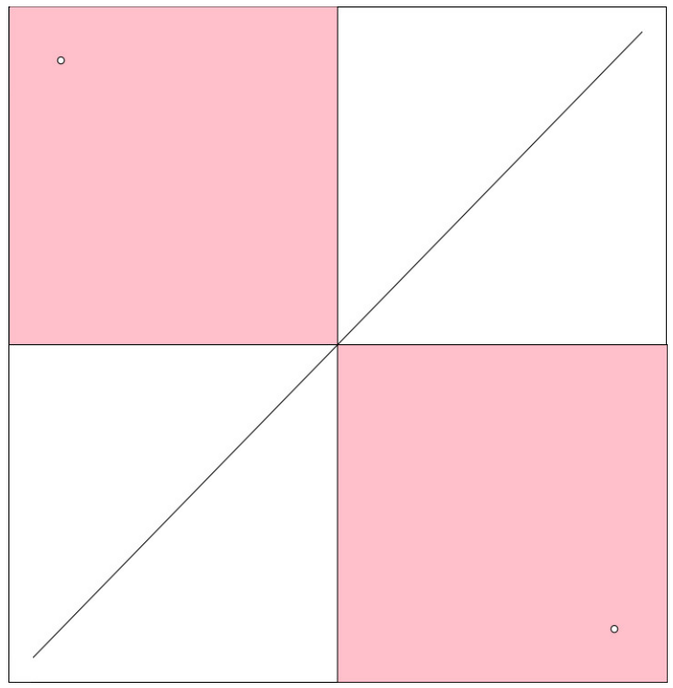
the homology groups of $\mathrm{M}^{\bullet}_{f,\delta_n}(\hat{X}_n, \mathcal{I}(g_n, r_n))$ and $\mathrm{R}_f(X)$ are isomorphic, with expected size difference less than $O\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/d}$.

Examples: Images

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

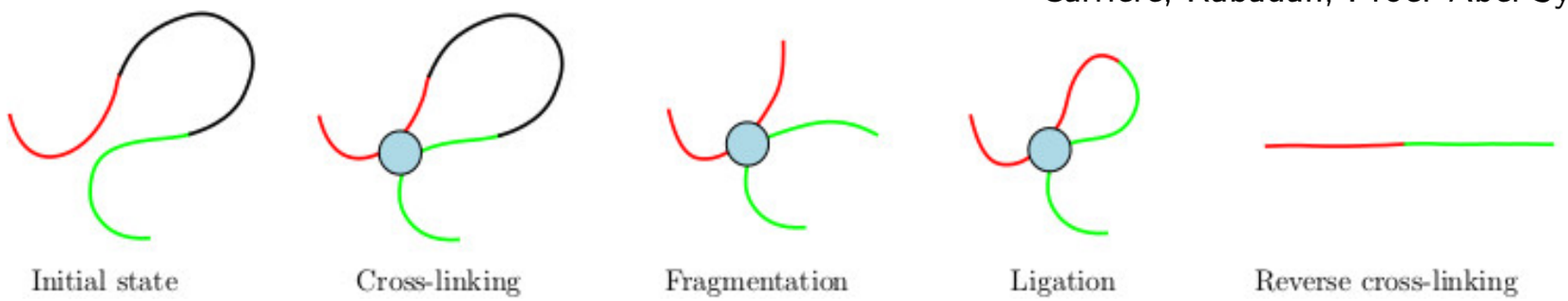


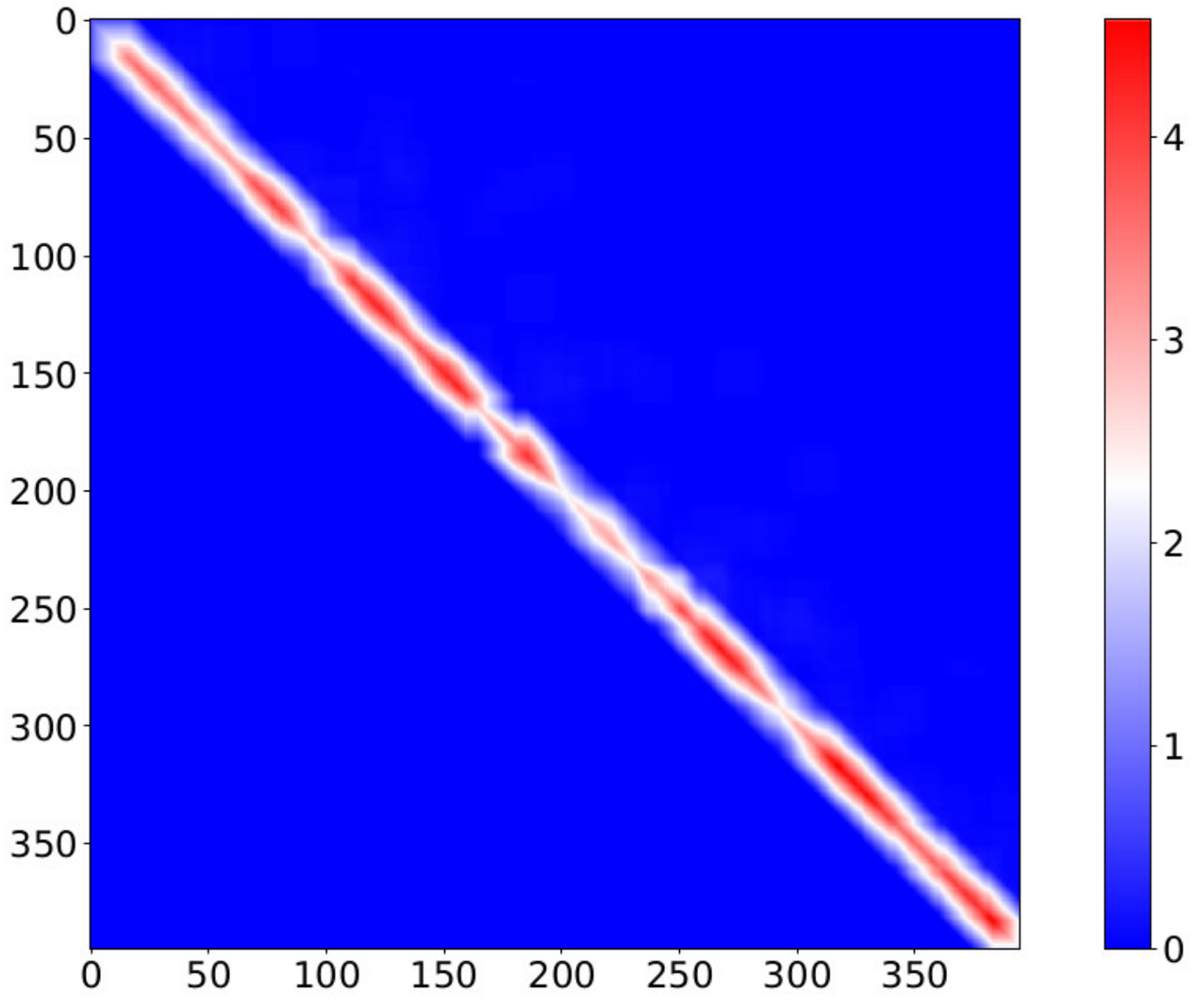




Examples: Chromosome conformation capture

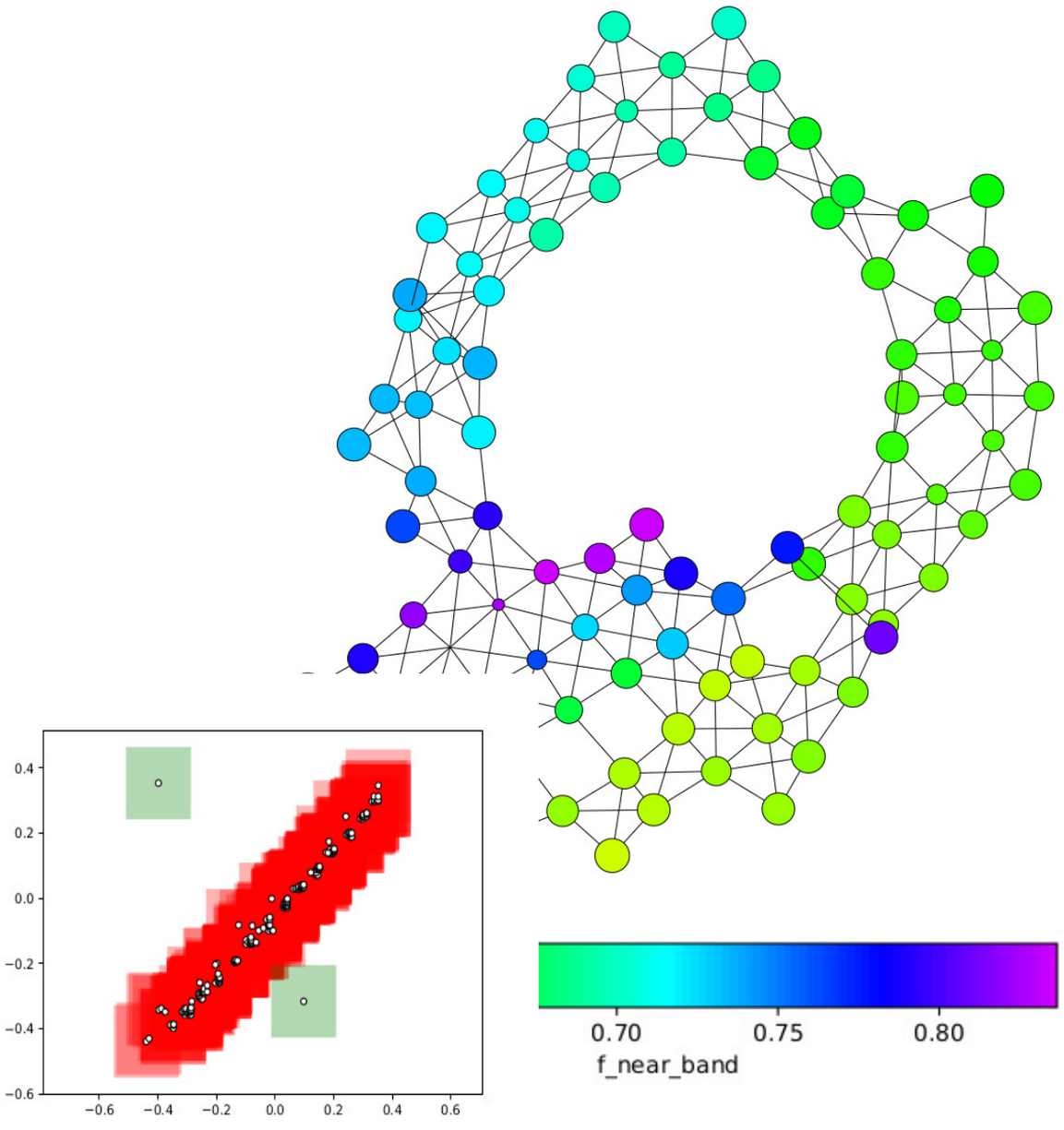
[*Topological data analysis of single-cell Hi-C contact maps*, Carrière, Rabadán, Proc. Abel Symp., 2018]

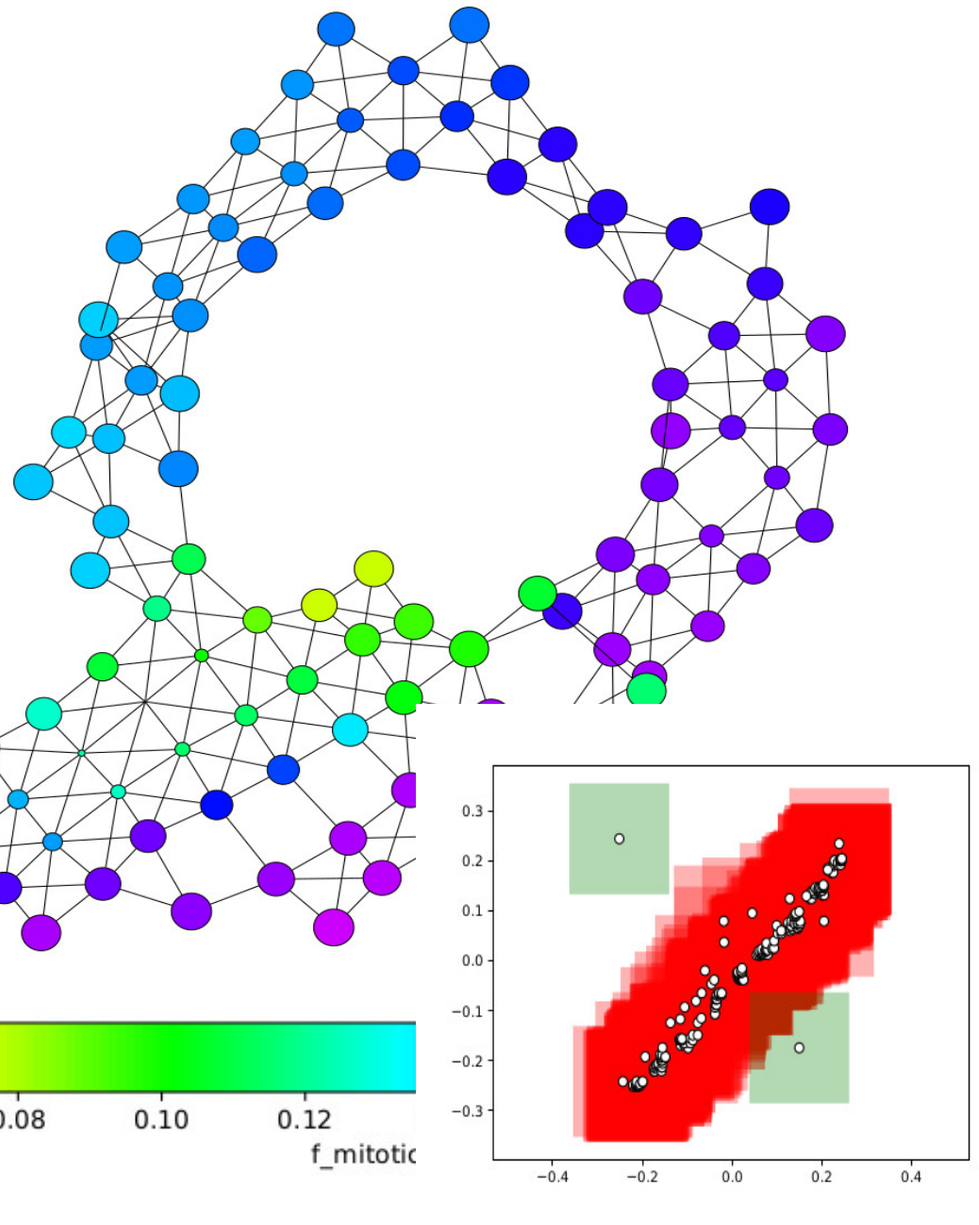




Examples: Chromosome conformation capture

[*Topological data analysis of single-cell Hi-C contact maps*, Carrière, Rabadán, Proc. Abel Symp., 2018]

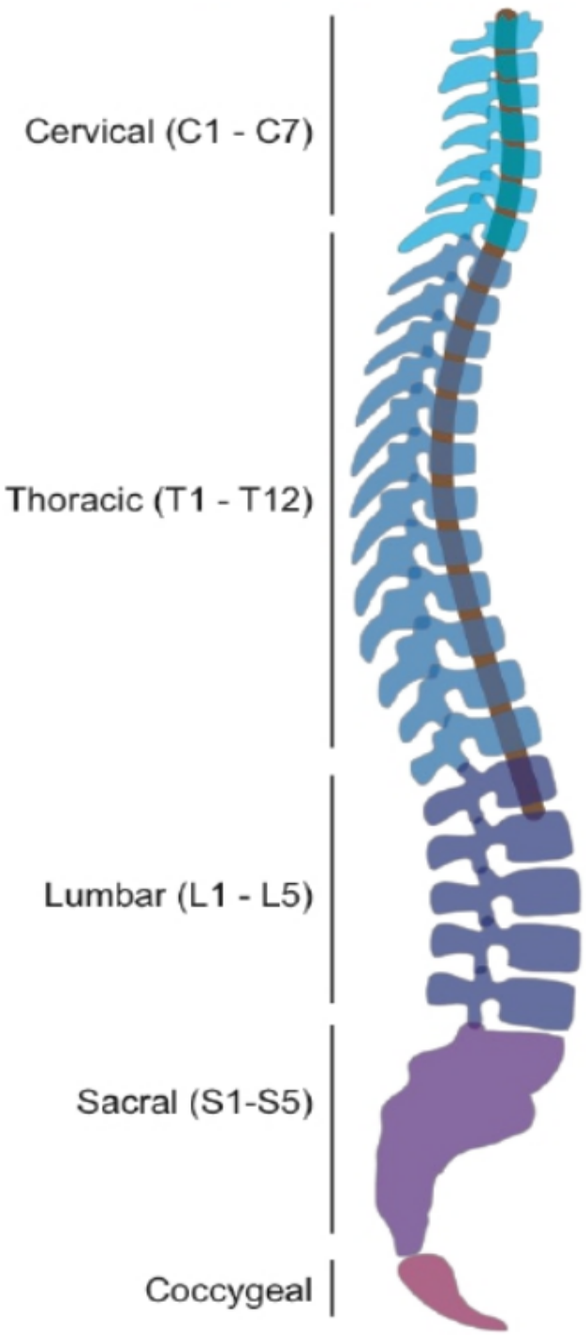


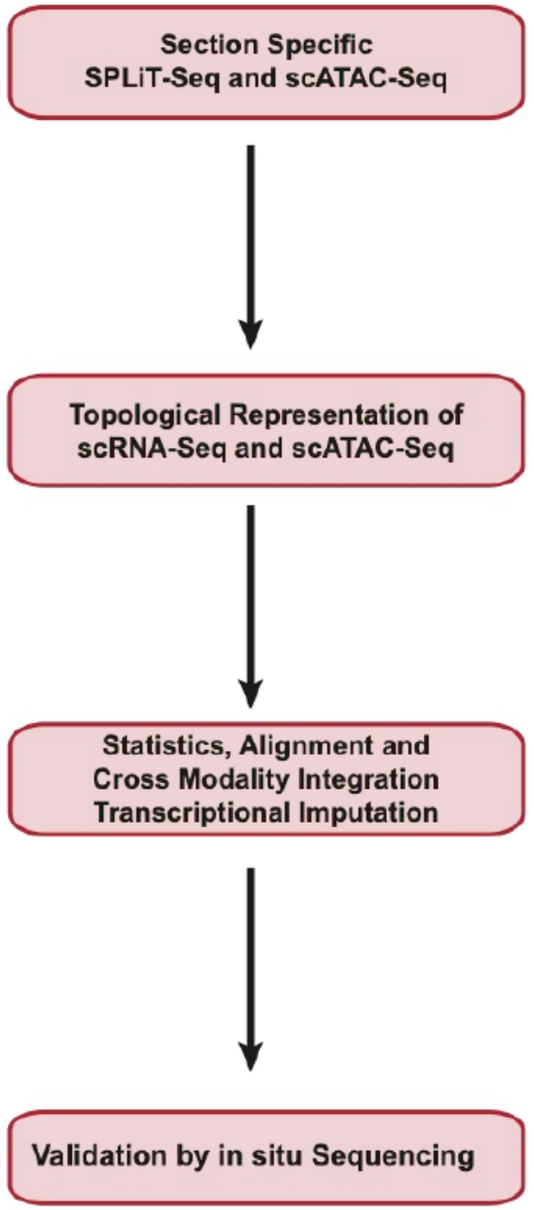


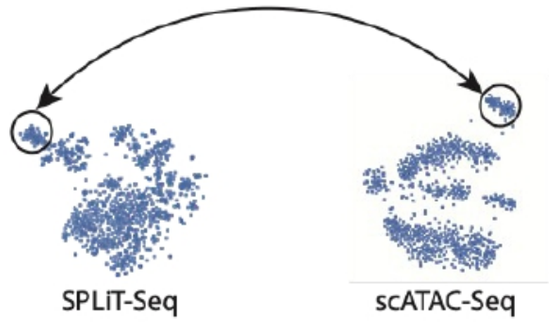
Formal identification of cell cycle with 95% confidence

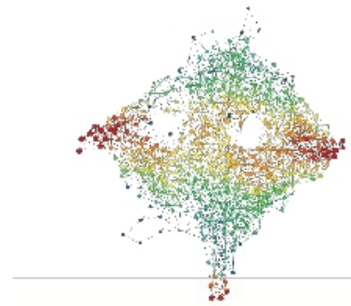
Examples: Spinal cord data

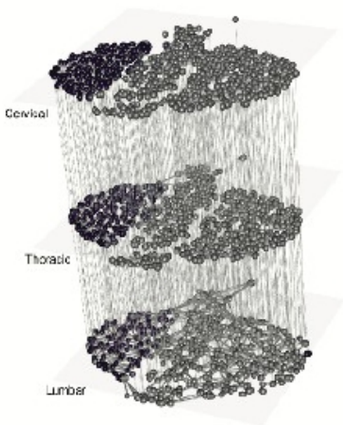
[Work in progress with Rizvi, Kandror, Loper, Wang, Chen, Maniatis, Paninski, Rabadán]

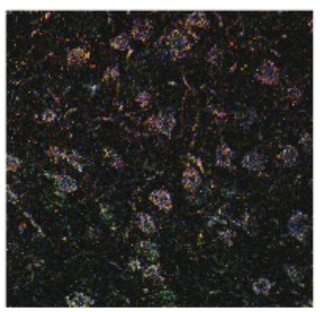






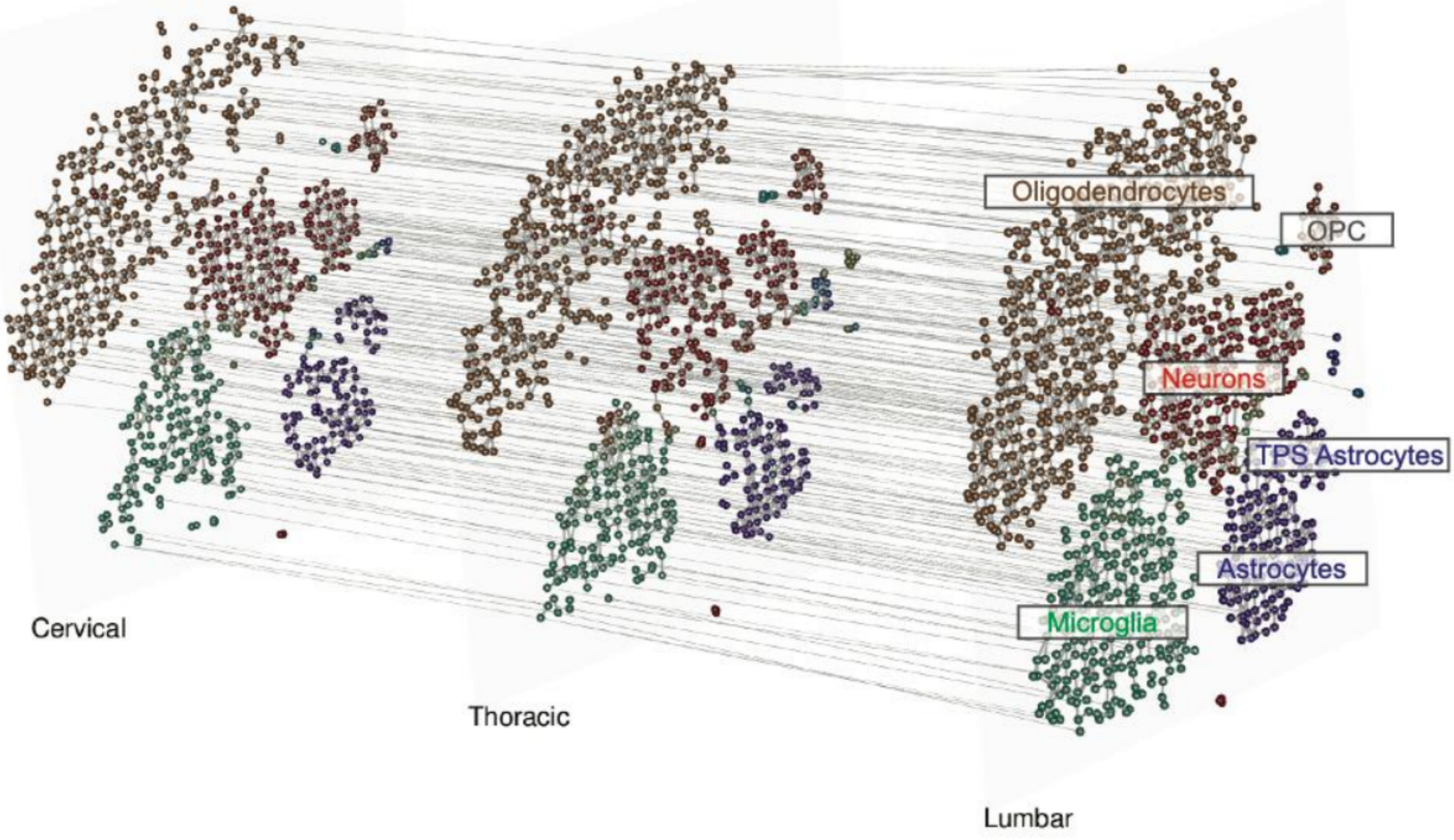






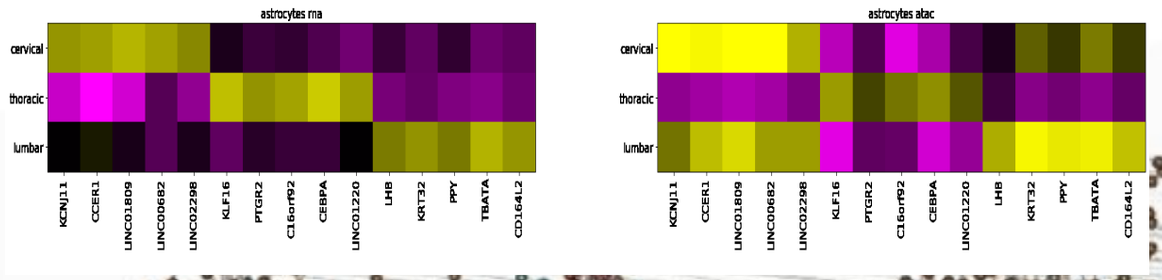
Examples: Spinal cord data

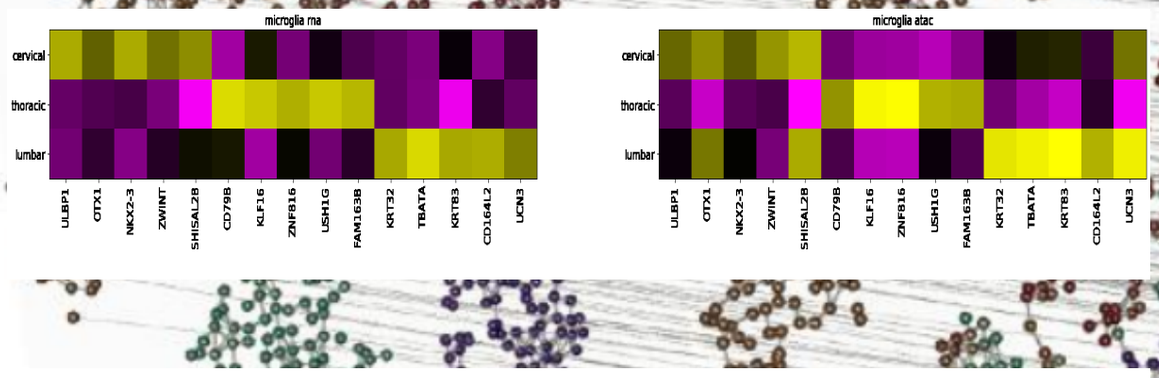
[Work in progress with Rizvi, Kandror, Loper, Wang, Chen, Maniatis, Paninski, Rabadán]

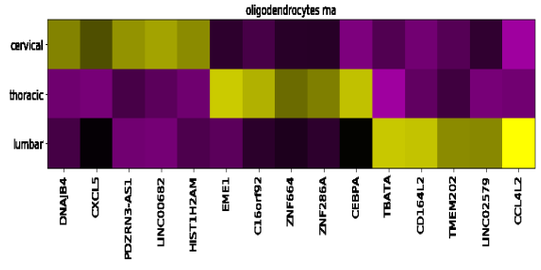


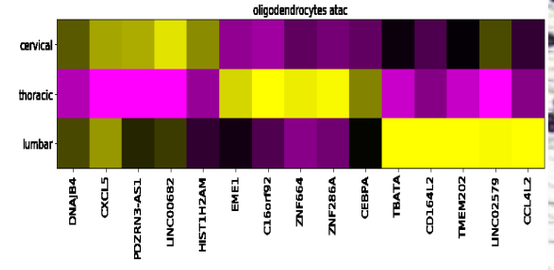
Examples: Spinal cord data

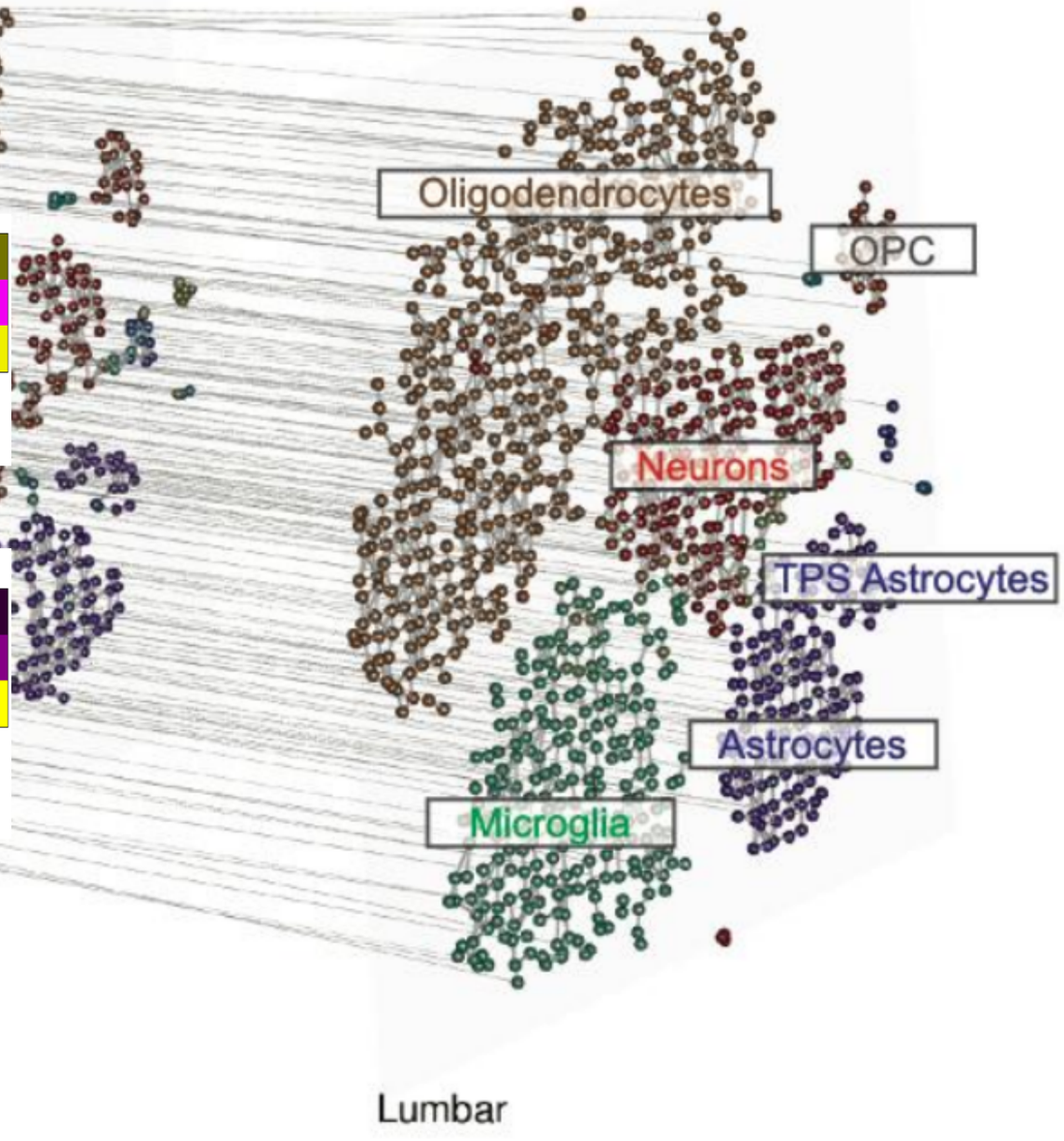
[Work in progress with Rizvi, Kandror, Loper, Wang, Chen, Maniatis, Paninski, Rabadán]







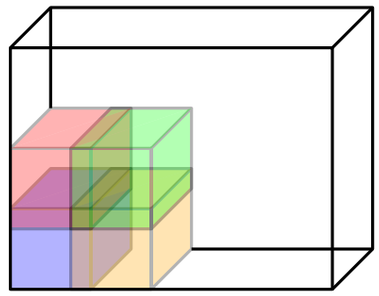




Examples: Interpreting machine learning classifier

[*Approximation of Reeb spaces with Mappers and applications to stochastic filters*, Carrière, Michel, 2021]

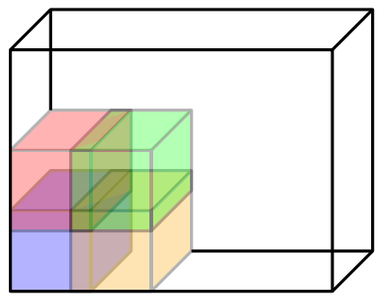
All previous results can be extended with filters $X \to \mathbb{R}^p$ and covers with hypercubes.



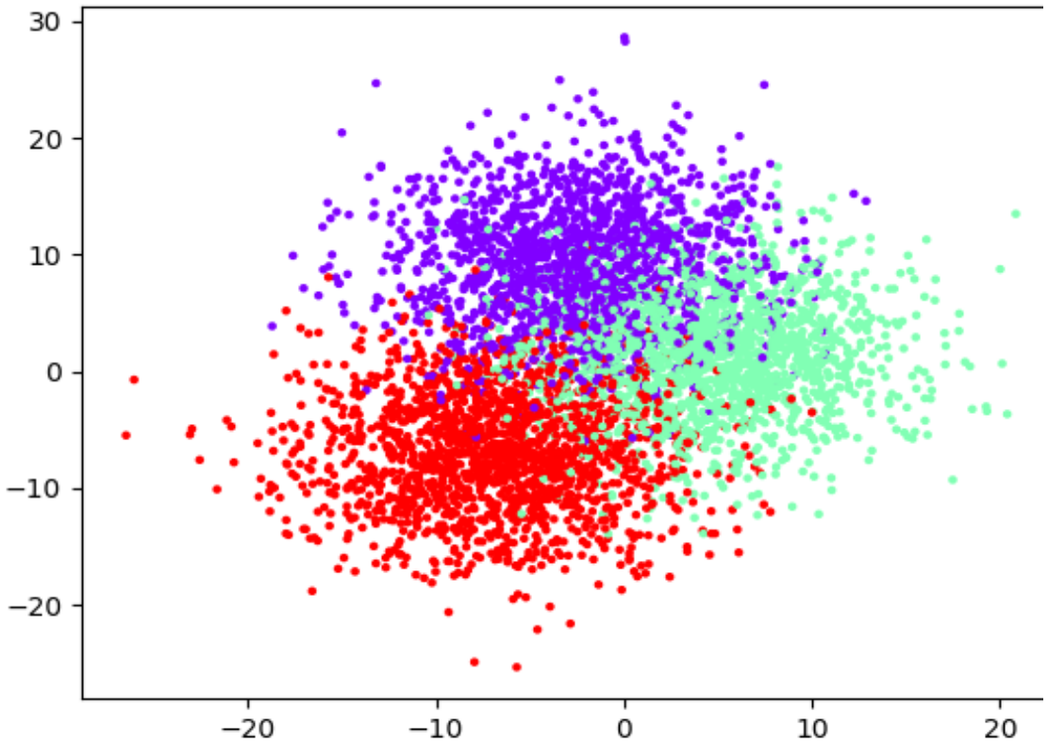
Examples: Interpreting machine learning classifier

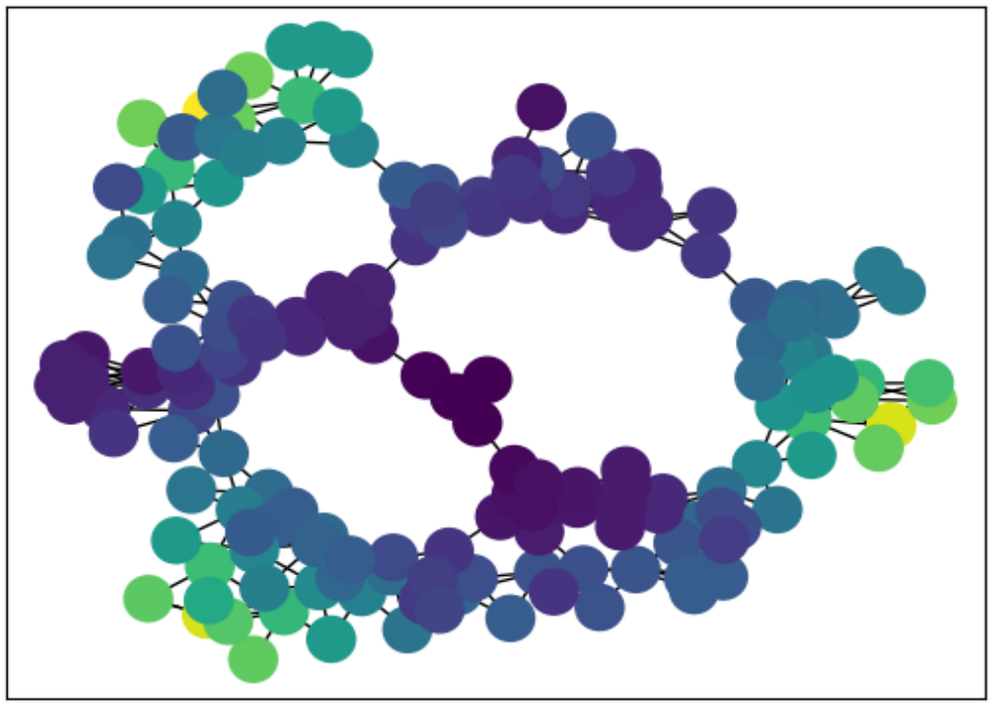
[*Approximation of Reeb spaces with Mappers and applications to stochastic filters*, Carrière, Michel, 2021]

All previous results can be extended with filters $X \to \mathbb{R}^p$ and covers with hypercubes.



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

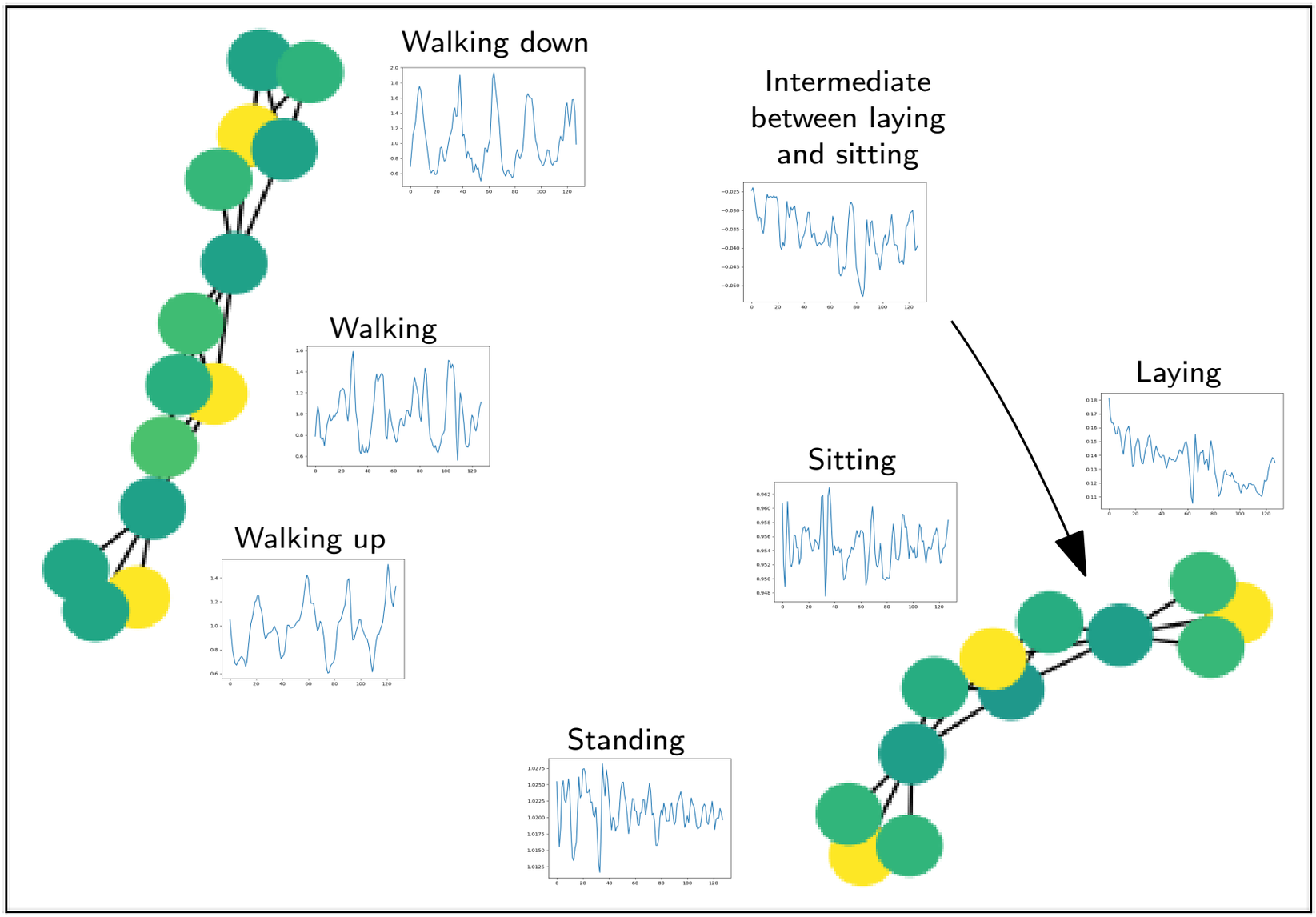




Examples: Interpreting machine learning classifier

[*Approximation of Reeb spaces with Mappers and applications to stochastic filters*, Carrière, Michel, 2021]

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

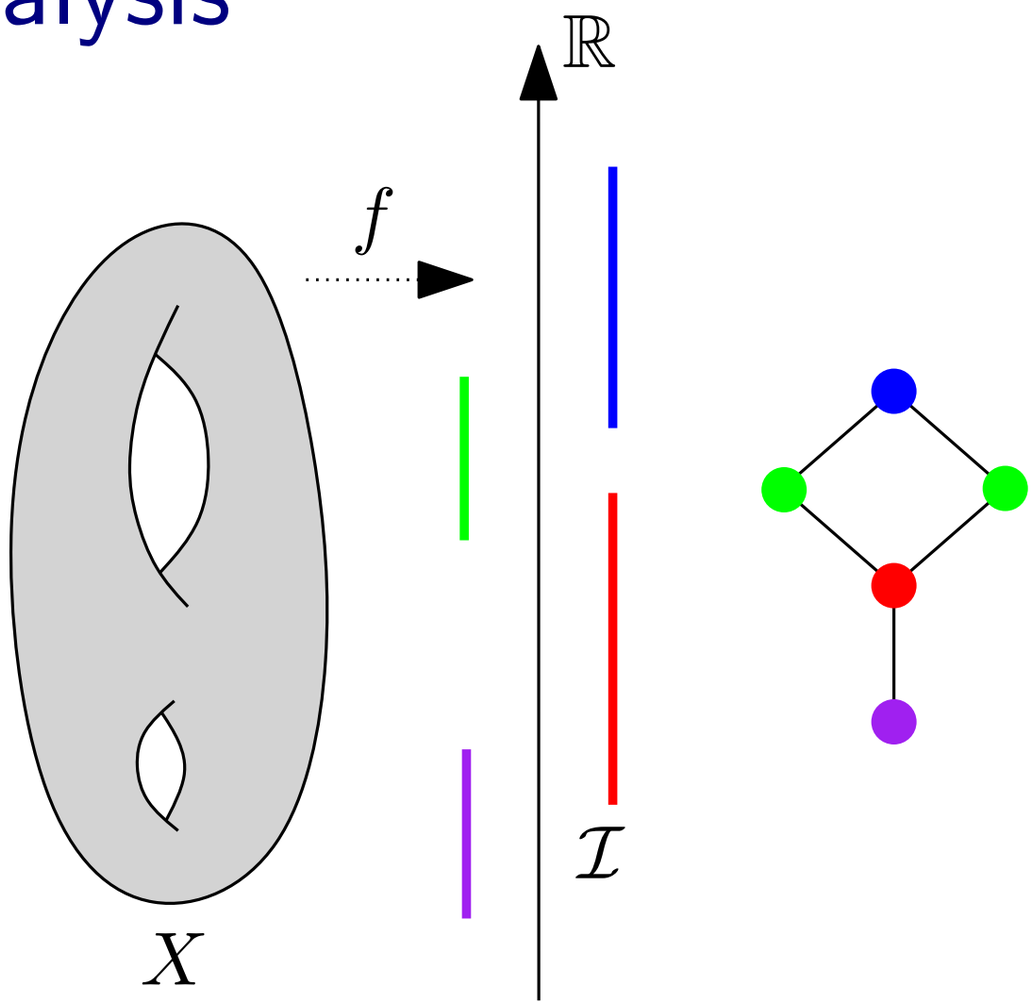


Topological exploratory data analysis

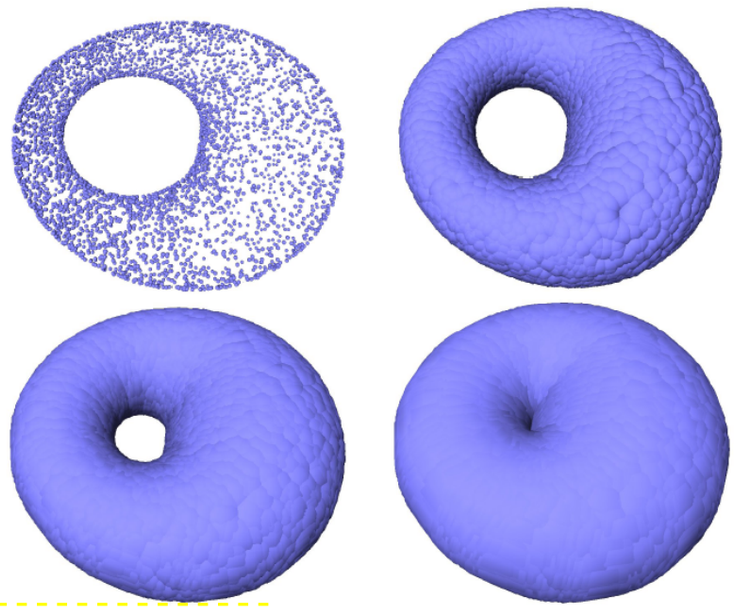
Q: How to build meaningful covers?

Two directions:

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



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



A few basic ideas about geometric inference: union of balls and distance functions

[*Minimax rates for homology inference*, Balakrishnan, Rinaldo, Sheehy, Singh, Wasserman, AISTATS 2012]

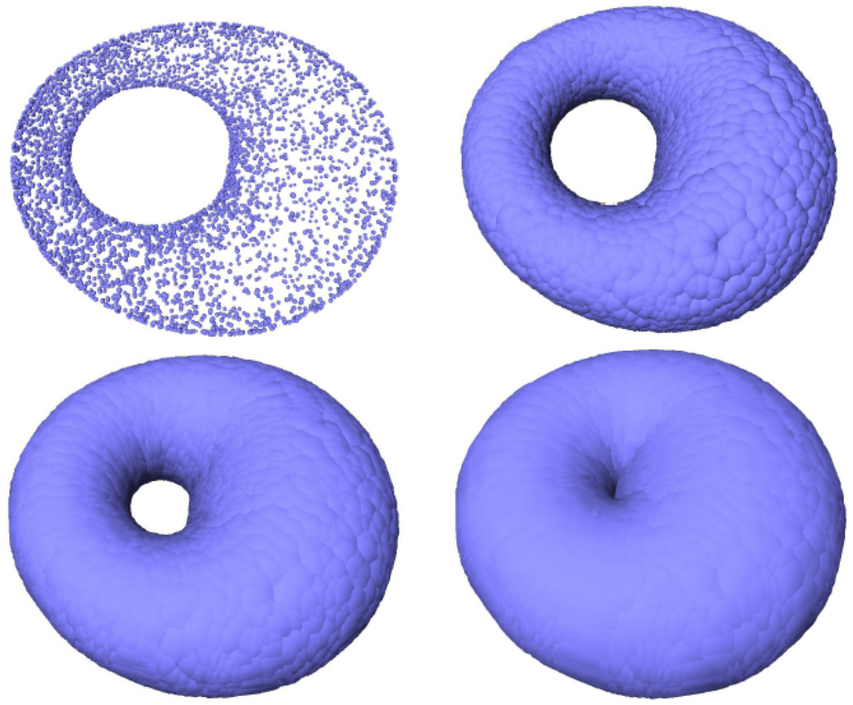
[*Finding the homology if submanifolds with high confidence from radom samples*, Niyogi, Smale, Weinberger, Discr. Comput. Geom., 2008]

[*Smooth manifold reconstruction from noisy and non uniform approximation with guarantees*, Chazal, Lieutier, Comp. Geom: theory and Applications, 2008]

[*Manifold estimation and singular deconvolution under Hausdorff loss*, Genovese, Perone-Pacifico, Verdinelli, Wasserman, Ann. Stat., 2012]

Union of balls and distance functions

Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



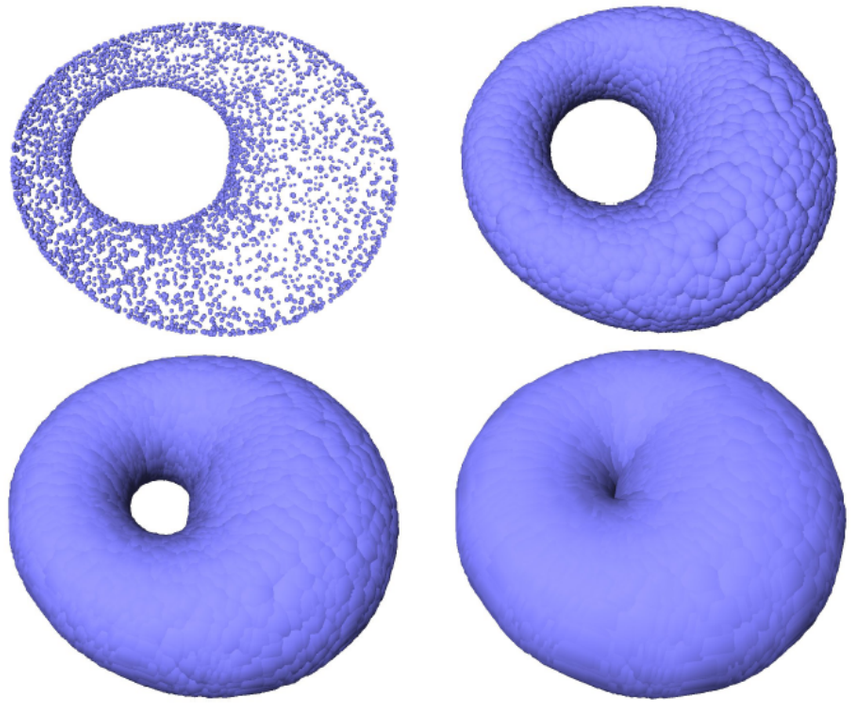
General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.

Union of balls and distance functions

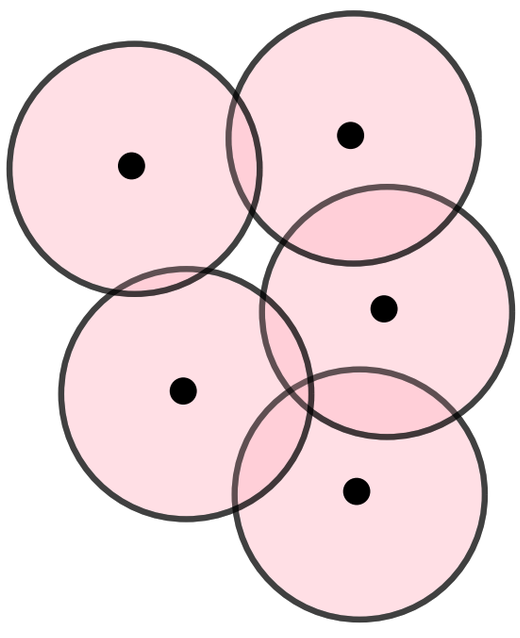
Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



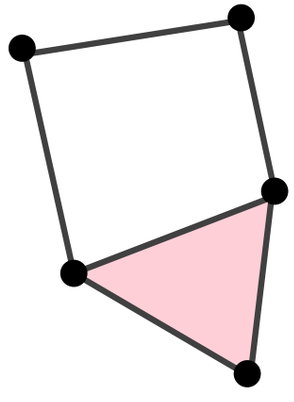
General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.

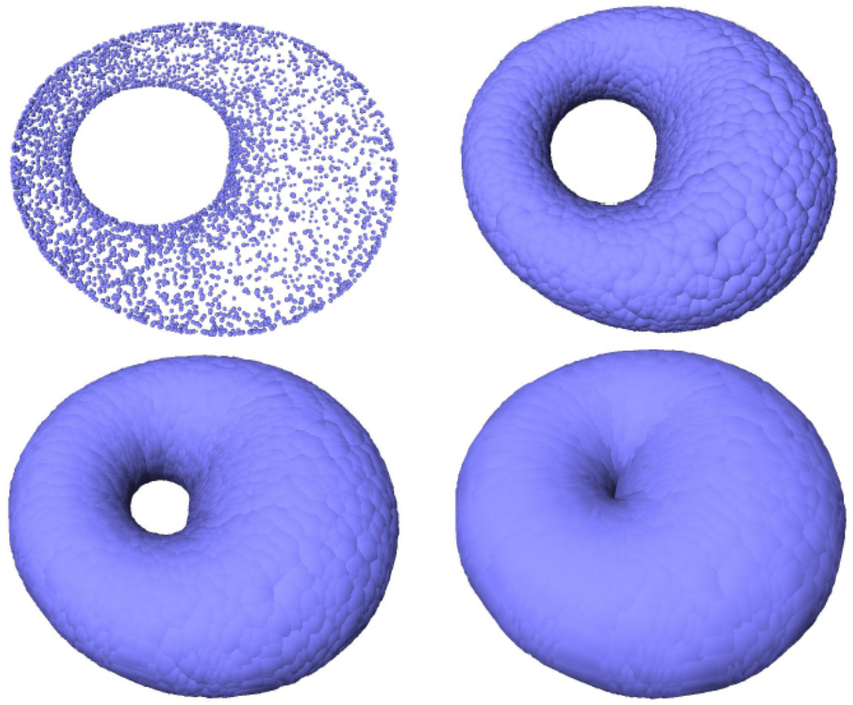


Nerve theorem



Union of balls and distance functions

Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.

Sublevel set of the distance function $d_P : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

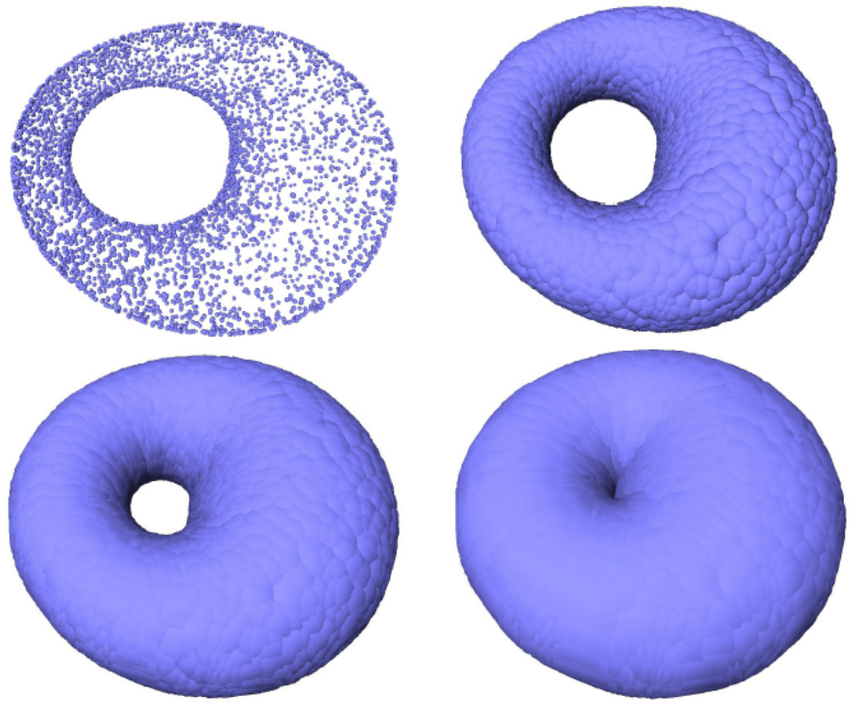
$$d_P(x) = \inf_{p \in P} \|x - p\|$$

Compare the topology of the offsets

$$M^r = d_M^{-1}([0, r]) \text{ and } P^r = d_P^{-1}([0, r])$$

Union of balls and distance functions

Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.

Sublevel set of the distance function $d_P : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

$$d_P(x) = \inf_{p \in P} \|x - p\|$$

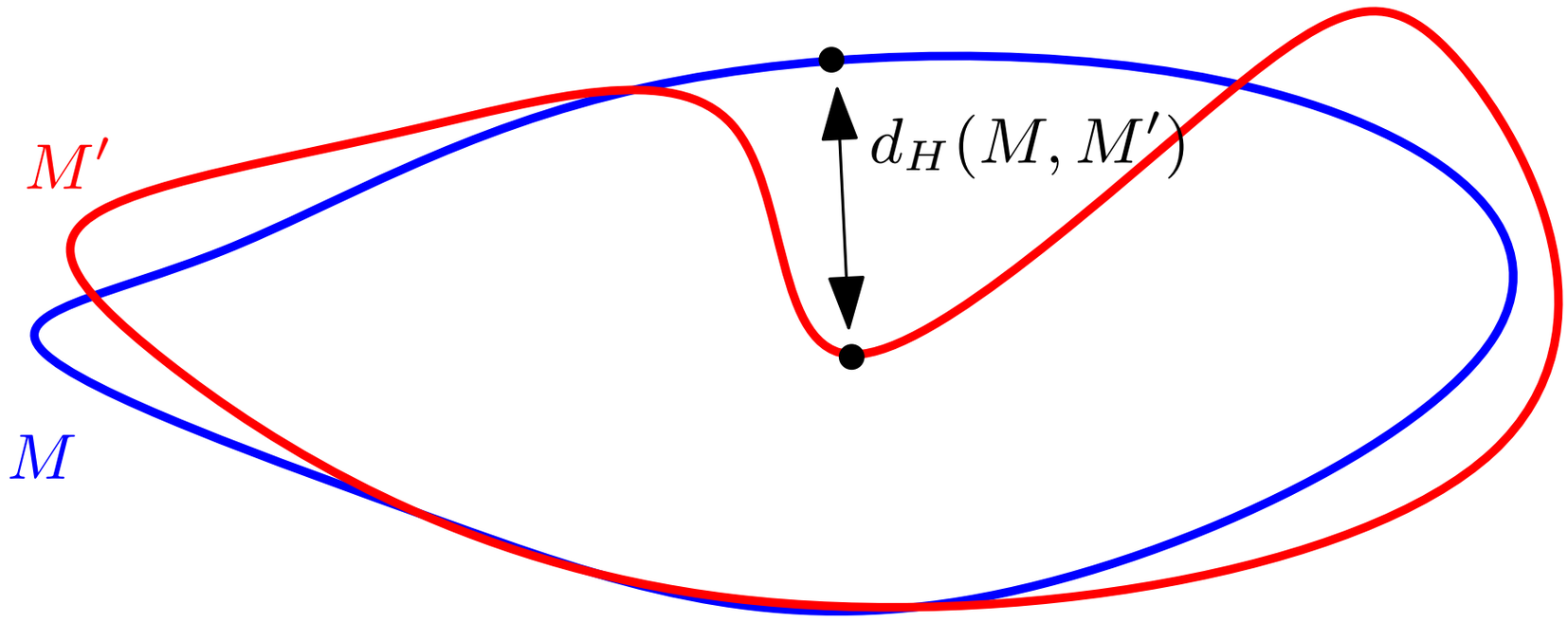
Regularity conditions?

Sampling conditions?

Compare the topology of the offsets

$$M^r = d_M^{-1}([0, r]) \text{ and } P^r = d_P^{-1}([0, r])$$

The Hausdorff distance



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

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

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

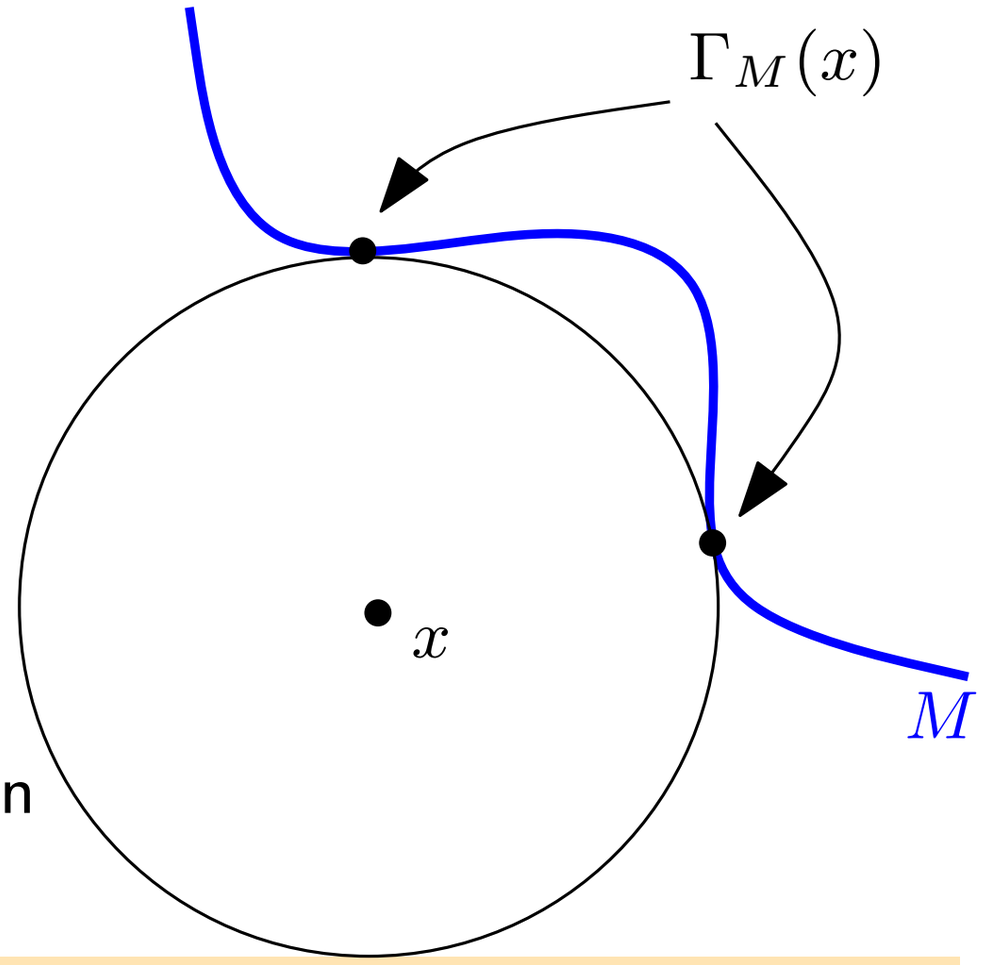
$$d_H(M, M') = \sup_{x \in \mathbb{R}^d} |d_M(x) - d_{M'}(x)|$$

Medial axis and critical points

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

Def: The medial axis of M:

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



$x \in \mathbb{R}^d$ is a critical point of d_M iff x is contained in the convex hull of $\Gamma_M(x)$.

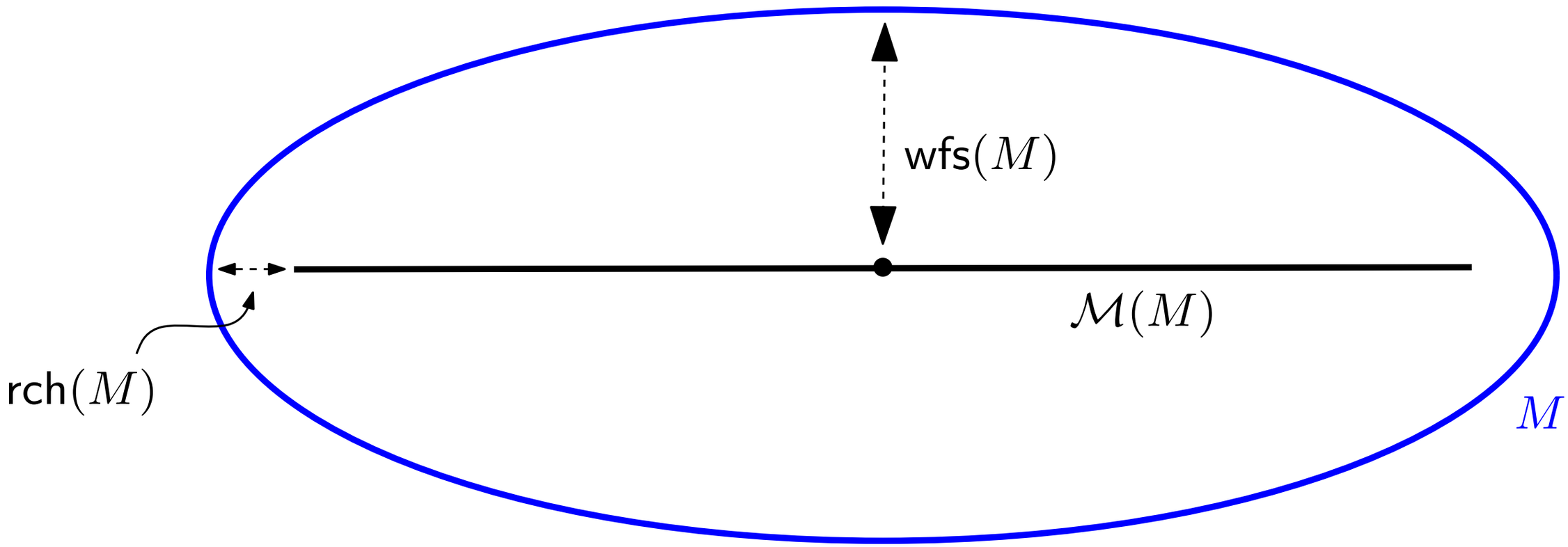
Thm: Let $M \subset \mathbb{R}^d$ be a compact set.

- if r is a regular value of d_M, then $d_M^{-1}(r)$ is a topological submanifold of $\mathbb{R}^d$ of codim 1.
- Let $0 < r_1 < r_2$ be such that $[r_1, r_2]$ does not contain any critical value of d_M. Then all the level sets $d_M^{-1}(r)$, $r \in [r_1, r_2]$ are isotopic and

$$M^{r_2} \setminus M^{r_1} = \{x \in \mathbb{R}^d : r_1 < d_M(x) \leq r_2\}$$

is homeomorphic to $d_M^{-1}(r_1) \times (r_1, r_2]$.

Reach and weak feature size



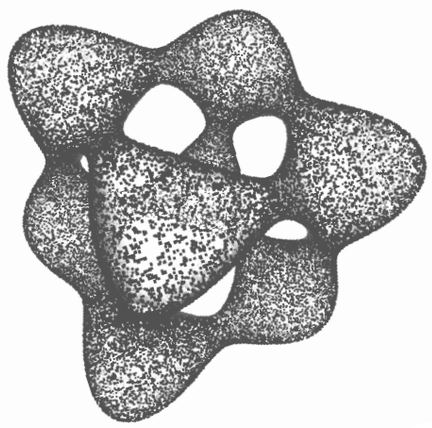
Def: The reach of M, rch(M) is the smallest distance from $\mathcal{M}(M)$ to M:

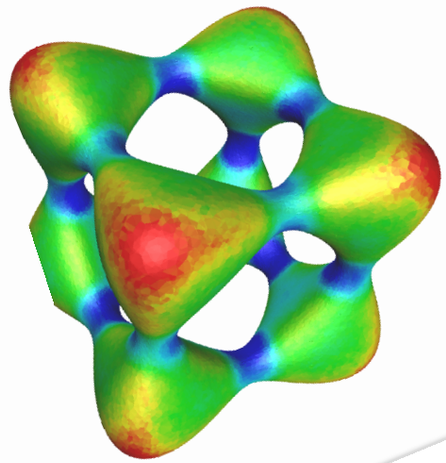
$$\text{rch}(M) = \inf_{y \in \mathcal{M}(M)} d_M(y)$$

Def: The weak feature size of M, wfs(M), is the smallest distance from the set of critical points of d_M to M:

$$\text{wfs}(M) = \inf\{d_M(y) : y \in \mathbb{R}^d \setminus M \text{ and } y \text{ crit. point of } d_M\}$$

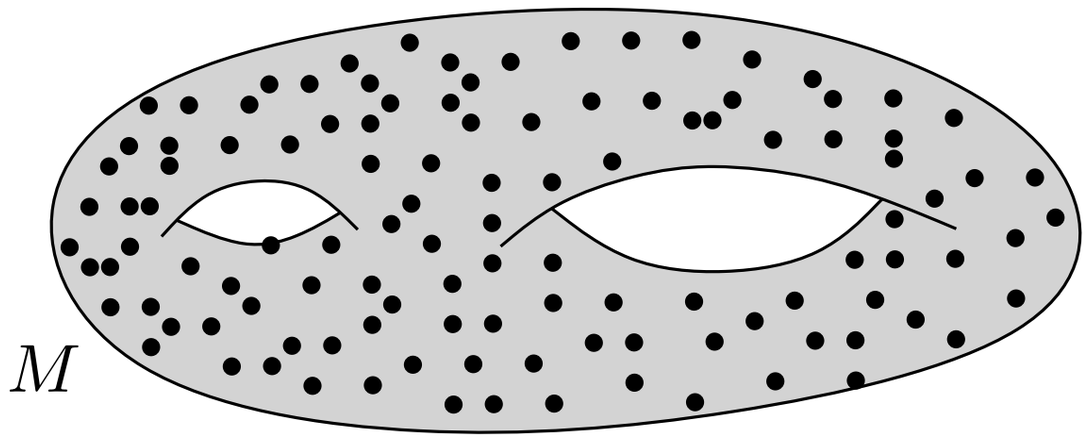
Geometric inference





Thm: Let $M \subset \mathbb{R}^d$ be such that $\tau = \mathrm{rch}(M) > 0$ and let $P \subset \mathbb{R}^d$ be such that $d_H(M, P) < c\tau$ for some (explicit) constant c. Then, for well-chosen (and explicit) r, P^r, and thus its nerve, is homotopy equivalent to M.

The probabilistic setting



Let $M \subset \mathbb{R}^d$ be a k-dim compact submanifold with positive reach $\mathrm{rch}(M) \geq \tau > 0$.

Let P be a probability measure such that $\mathrm{Supp}(P) = M$ which is (a, k)-*standard*: there exists $r_0 \geq \tau/8 > 0$ s.t. for any $x \in M$, $r \leq r_0$, $P(B(x, r)) \geq ar^k$.

Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be n points i.i.d. sampled according to P.

Goal: Upper bound $P(X^r \not\sim M)$ where $\sim$ denotes the homotopy equivalence.

↳ Connection to support estimation problems: it is enough to bound $P(d_H(X, M) > \varepsilon)$.

Minimax risk

Let $\mathcal{Q} = \mathcal{Q}(d, k, \tau, a)$ be the family of probability measures on $\mathbb{R}^d$ s.t. $\forall \mu \in \mathcal{Q}$:
- Supp(μ) is a compact k-dimensional manifold with reach larger than $\tau > 0$;
- μ is (a, k)-standard.

Given $\mu \in \mathcal{Q}$, Supp$(\mu) = M$, denote by $\hat{M}$ any homotopy type estimator of M that takes as input n-uples of points from M and outputs a set whose homotopy type "estimates" the homotopy type of M (e.g. a union of balls).

Def: The minimax risk is $R_n = \inf_{\hat{M}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{M} \not\sim M)$.

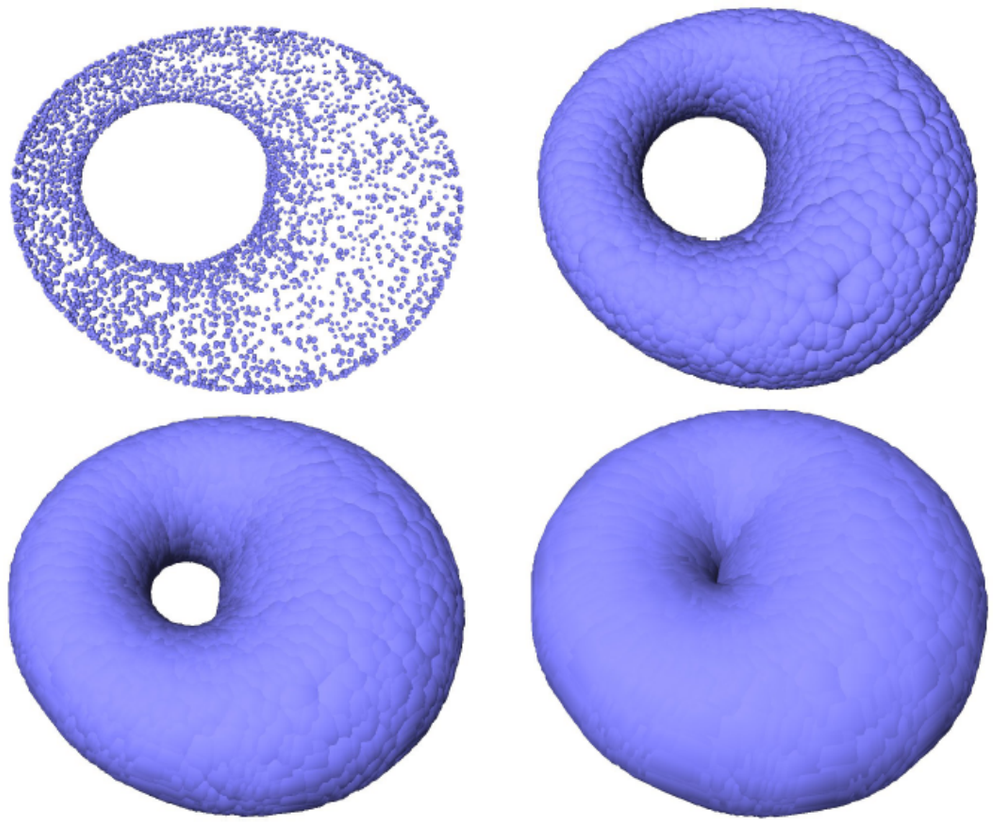
Thm: There exist constants $C_a, C'_a, C''_a > 0$ s.t.

$$\frac{1}{8} \exp(-nC_a\tau^k) \leq R_n \leq C'_a \frac{1}{\tau^k} \exp(-nC''_a\tau^k)$$

More details on geometric inference and minimax convergence rates

Reconstruction theorem: smooth case

Let $M \subset \mathbb{R}^d$ be a k-dimensional compact submanifold with positive reach $\text{rch}(M) \geq \tau > 0$.



Lem: for any $0 < r < \tau$, the offset M^r deformation retracts on M. In particular, M^r and M are homotopy equivalent.

Reconstruction theorem:
Let $X \subset \mathbb{R}^d$ be a compact set such that $d_H(M, X) = \varepsilon < \frac{1}{17}\tau$. Then for any r s.t.

$$4\varepsilon \leq r < \tau - 3\varepsilon$$

$\frac{7\varepsilon}{2}$ $\tau - \frac{9}{2}\varepsilon$

the offset X^r and M are homotopy equivalent.

$\varepsilon < \frac{\tau}{8}$

Rem: A more careful analysis leads to slightly better bounds.

The probabilistic setting: a basic lemma

Lem: Let $\{A_i\}_{i=1,\dots,l}$ be a finite collection of measurable sets such that $M \subset \cup_{i=1}^{l} A_i$ and $P(A_i) > \alpha$ for some $\alpha > 0$.
Let $X = \{x_1, \dots, x_n\}$ a set of n i.i.d. points sampled according to P.
Then, we have

$$P(X \cap A_i = \emptyset, \text{ for some } i = 1, \dots, l) \leq 1 - e^{-n\alpha}.$$

Proof:

- Let E_i be the event $X \cap A_i = \emptyset$
- $P(E_i) = (1 - P(A_i))^n \leq (1 - \alpha)^n$
- $P(\cup E_i) \leq \sum P(E_i) \leq l(1 - \alpha)^n$
- Use that $(1 - \alpha) \leq e^{-\alpha}$

Idea: Take $A_i = B(x_i, r)$ and bound l.

Covering and packing numbers

$C_M(r) =$ minimum number of balls of radius r needed to cover M

$P_M(r) =$ maximum number of balls of radius r and center on M that may be packed into M without overlap in M
$= \max\{k : \exists y_1, \ldots, y_k \in M \text{ s.t. } \forall i \neq j,\ B(y_i, r) \cap B(y_j, r) \cap M = \emptyset\}$.

Covering and packing numbers

$C_M(r) =$ minimum number of balls of radius r needed to cover M

$P_M(r) =$ maximum number of balls of radius r and center on M that may be packed into M without overlap in M
$= \max\{k : \exists y_1, \ldots, y_k \in M$ s.t. $\forall i \neq j,\ B(y_i, r) \cap B(y_j, r) \cap M = \emptyset\}$.

Lem: $C_M(2r) \leq P_M(r)$.

Cor: for any $r \leq 2r_0$, $C_M(r) \leq P_M(\frac{r}{2}) \leq P(M)/P(B(x, r/2)) \leq \frac{2^k}{a} r^{-k}$.

An upper bound

1. As soon as $d_H(X, M) < \frac{\tau}{8}$, one can recover the homotopy type of M from $X^{r'}$ for well chosen r'.
2. Let $r = \frac{\tau}{8}$. Then
$$C_M(r) \leq \frac{2^{4k}}{a} \left(\frac{1}{\tau}\right)^k$$
3. Let $B_1, \dots B_l$ a covering of M by balls of radius r with $l \leq \frac{2^{4k}}{a} \left(\frac{1}{\tau}\right)^k$. For any $i = 1, \dots, l$, $P(B_i) \geq \alpha = ar^k = \frac{a}{2^{3k}} \tau^k$.
4. Then
$$P(d_H(X, M) > \frac{\tau}{8}) \leq le^{-n\alpha} \leq \frac{2^{4k}}{a} \frac{1}{\tau^k} e^{-n \frac{a}{2^{3k}} \tau^k}$$

Cor: Let $a' = \frac{a}{2^{3k}}$ and let $r' \in (\frac{7\tau}{16}, \frac{5\tau}{8})$. then

$$P(X^{r'} \not\sim M) \leq \frac{2^k}{a'} \frac{1}{\tau^k} e^{-na'\tau^k}$$

Rem: This bound only depends on a, k and τ.

A lower bound with Le Cam lemma

Lem: Let $\mathcal{Q}$ be a set of probability distributions and let $\theta(Q)$ taking values in a metric space with metric ρ. Let $Q_1, Q_2 \in \mathcal{Q}$ be any pair of distributions. Let $x_1, \ldots, x_n$ be n points i.i.d sampled from some $Q \in \mathcal{Q}$. Then

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \left[\rho(\hat{\theta}(x_1, \ldots, x_n), \theta(Q)) \right] \geq \frac{1}{4} \rho(\theta(Q_1), \theta(Q_2)) (1 - TV(Q_1, Q_2))^{2n}$$

where Q^n is the product measure and $TV(.,.)$ is the total variation distance.

A lower bound with Le Cam lemma

Lem: Let $\mathcal{Q}$ be a set of probability distributions and let $\theta(Q)$ taking values in a metric space with metric ρ. Let $Q_1, Q_2 \in \mathcal{Q}$ be any pair of distributions. Let $x_1, \ldots, x_n$ be n points i.i.d sampled from some $Q \in \mathcal{Q}$. Then

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \left[\rho(\hat{\theta}(x_1, \ldots, x_n), \theta(Q)) \right] \geq \frac{1}{4} \rho(\theta(Q_1), \theta(Q_2)) (1 - TV(Q_1, Q_2))^{2n}$$

where Q^n is the product measure and $TV(.,.)$ is the total variation distance.

In our case:

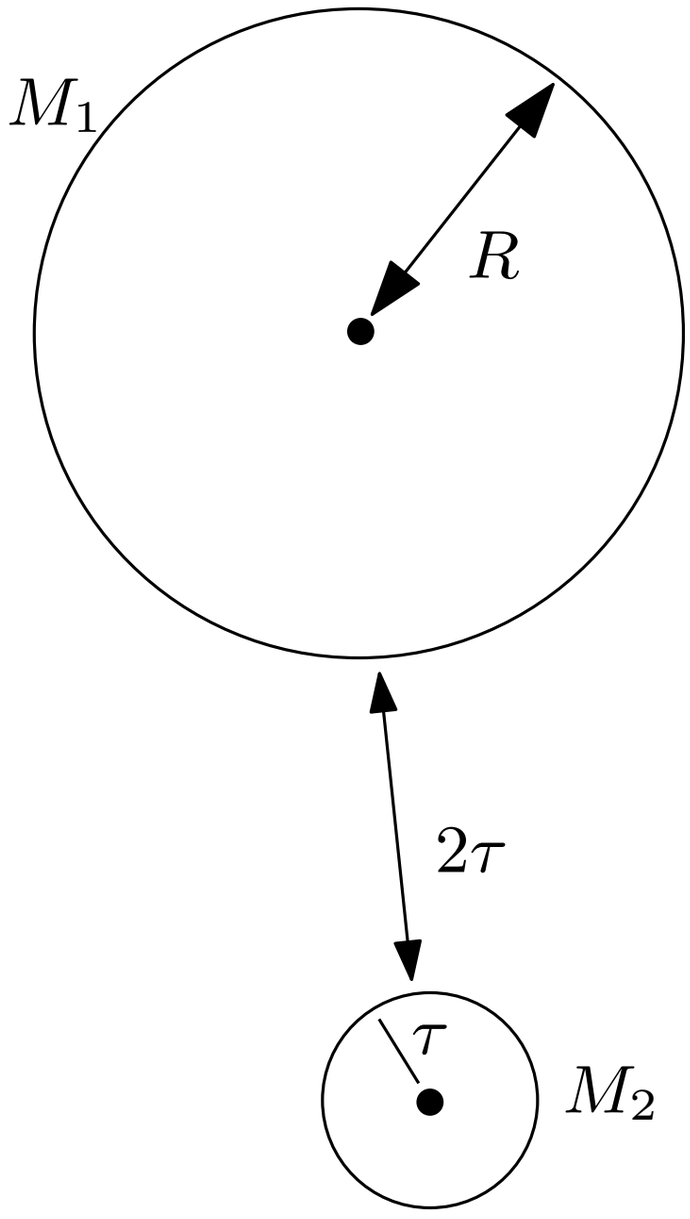
- $\mathcal{Q}$ as before
- metric space: the set of homotopy equivalent classes of compact subsets of $\mathbb{R}^d$ with $\rho(K, K') = 1$ if K and K' are not homotopy equivalent, and 0 otherwise.

A lower bound with Le Cam lemma

$M_1 = S^k(0, R)$: a k dim. sphere of radius $R > \tau$.

$M_2 = M_1 \cup S^k(\tau)$ where $S^k(\tau)$ is at distance at least 2τ from M_1.

$v_k = \text{vol}(S^k(0, 1))$



Q_1: unif density (w.r.t. k-vol) $\rightarrow f_1 = \frac{1}{v_k R^k}$

Q_2: unif density (w.r.t. k-vol) $\rightarrow f_2 = \frac{1}{v_k (R^k + \tau^k)}$

$$\begin{aligned} TV(Q_1, Q_2) &= Q_2(M_2 \setminus M_1) - Q_1(M_2 \setminus M_1) \\ &= f_2 v_k \tau^k - 0 = \frac{\tau^k}{R^k + \tau^k} \leq C_R \tau^k \end{aligned}$$

So, $(1 - TV(Q_1, Q_2))^{2n} \geq (1 - C_a \tau^k)^{2n}$

$$R_n \geq \frac{1}{8}(1 - C_a \tau^k)^{2n} \geq \frac{1}{8} \exp(-4 C_a n \tau^k)$$

Affinity, total variation and Hellinger distances

Let P and Q be two (σ-finite, Borel) probability measures with density p and q with respect to any third measure that dominates both P and Q.

Def: Let $p \wedge q(x) = \min(p(x), q(x))$. The *affinity* between P and Q is

$$\|P \wedge Q\| = \int p \wedge q = 1 - \frac{1}{2} \int |p - q|$$

Def: The *total variation* distance between P and Q si defined as

$$\begin{aligned} TV(P,Q) &= \sup_{A \ borel \ set} |P(A) - Q(A)| \\ &= P(G) - Q(G) \text{ where } G = \{x : p(x) \geq q(x)\} \\ &= 1 - \int p \wedge q = 1 - \|P \wedge Q\| \end{aligned}$$

Def: The *Hellinger distance* between P and Q is defined by

$$h^2(P,Q) = \int (\sqrt{p} - \sqrt{q})^2 = 2(1 - \int \sqrt{pq})$$

Proof of Le Cam lemma

Let $\hat{\theta}$ and n be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}(\rho(\hat{\theta}, \theta(Q)) \geq \frac{1}{2} \underbrace{\left[\mathbb{E}_{Q_1^n}(\rho(\hat{\theta}, \theta(Q_1)) + \mathbb{E}_{Q_2^n}(\rho(\hat{\theta}, \theta(Q_2))\right]}_{A}$$

Proof of Le Cam lemma

Let $\hat{\theta}$ and n be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}(\rho(\hat{\theta}, \theta(Q)) \geq \frac{1}{2} \underbrace{\left[\mathbb{E}_{Q_1^n}(\rho(\hat{\theta}, \theta(Q_1)) + \mathbb{E}_{Q_2^n}(\rho(\hat{\theta}, \theta(Q_2))\right]}_{A}$$

Let μ be a measure dominating Q_1 and Q_2.

$$A = \int \rho(\hat{\theta}, \theta(Q_1)) q_{1,n} d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_{2,n} d\mu^n$$

Proof of Le Cam lemma

Let $\hat{\theta}$ and n be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}(\rho(\hat{\theta}, \theta(Q)) \geq \frac{1}{2} \underbrace{\left[\mathbb{E}_{Q_1^n}(\rho(\hat{\theta}, \theta(Q_1)) + \mathbb{E}_{Q_2^n}(\rho(\hat{\theta}, \theta(Q_2))\right]}_{A}$$

Let μ be a measure dominating Q_1 and Q_2.

$$A = \int \rho(\hat{\theta}, \theta(Q_1)) q_{1,n} d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_{2,n} d\mu^n$$

But $\rho(\hat{\theta}, \theta(Q_1)) + \rho(\hat{\theta}, \theta(Q_2)) \geq \rho(\theta(Q_1), \theta(Q_2))$

Proof of Le Cam lemma

Let $\hat{\theta}$ and n be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}(\rho(\hat{\theta}, \theta(Q)) \geq \frac{1}{2} \underbrace{\left[\mathbb{E}_{Q_1^n}(\rho(\hat{\theta}, \theta(Q_1)) + \mathbb{E}_{Q_2^n}(\rho(\hat{\theta}, \theta(Q_2))\right]}_{A}$$

Let μ be a measure dominating Q_1 and Q_2.

$$A = \int \rho(\hat{\theta}, \theta(Q_1)) q_{1,n} d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_{2,n} d\mu^n$$

But $\rho(\hat{\theta}, \theta(Q_1)) + \rho(\hat{\theta}, \theta(Q_2)) \geq \rho(\theta(Q_1), \theta(Q_2))$

$$A \geq \rho(\theta(Q_1), \theta(Q_2)) \int q_{1,n} \wedge q_{2,n} d\mu^n = \rho(\theta(Q_1), \theta(Q_2)) \| Q_1^n \wedge Q_2^n \|$$

Proof of Le Cam lemma

Let $\hat{\theta}$ and n be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}(\rho(\hat{\theta}, \theta(Q)) \geq \frac{1}{2} \underbrace{\left[\mathbb{E}_{Q_1^n}(\rho(\hat{\theta}, \theta(Q_1)) + \mathbb{E}_{Q_2^n}(\rho(\hat{\theta}, \theta(Q_2))\right]}_{A}$$

Let μ be a measure dominating Q_1 and Q_2.

$$A = \int \rho(\hat{\theta}, \theta(Q_1)) q_{1,n} d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_{2,n} d\mu^n$$

But $\rho(\hat{\theta}, \theta(Q_1)) + \rho(\hat{\theta}, \theta(Q_2)) \geq \rho(\theta(Q_1), \theta(Q_2))$

$$A \geq \rho(\theta(Q_1), \theta(Q_2)) \int q_{1,n} \wedge q_{2,n} d\mu^n = \rho(\theta(Q_1), \theta(Q_2)) \|Q_1^n \wedge Q_2^n\|$$

Lem: $\|Q_1^n \wedge Q_2^n\| \geq \frac{1}{2}\left(1 - \frac{1}{2}\int |q_1 - q_2|\right)^{2n} = \frac{1}{2}\|Q_1 \wedge Q_2\|^{2n}$

Proof of Le Cam lemma

Proof:

Claim A: $h^2(P,Q) \leq \int |p-q| = l_1(P,Q)$

Claim B: $h^2(P^n,Q^n) = 2\left(1-[1-\frac{h^2(P,Q)}{2}]^n\right)$

Claim C: $\left(1-\frac{h^2(P,Q)}{2}\right)^2 \leq 2\|P \wedge Q\|$

$$
\begin{aligned}
\|Q_1^n \wedge Q_2^n\| &\geq \frac{1}{2}\left(1-\frac{h^2(Q_1^n,Q_2^n)}{2}\right)^2 && (C)\\
&= \frac{1}{2}\left(1-\frac{h^2(Q_1,Q_2)}{2}\right)^{2n} && (B)\\
&\geq \frac{1}{2}\left(1-\frac{l_1(Q_1,Q_2)}{2}\right)^{2n} = \frac{1}{2}\|Q_1 \wedge Q_2\|^{2n} && (A)
\end{aligned}
$$