FGV EMAp — Seminário — 29/04/21

Topological inference in Topological Data Analysis

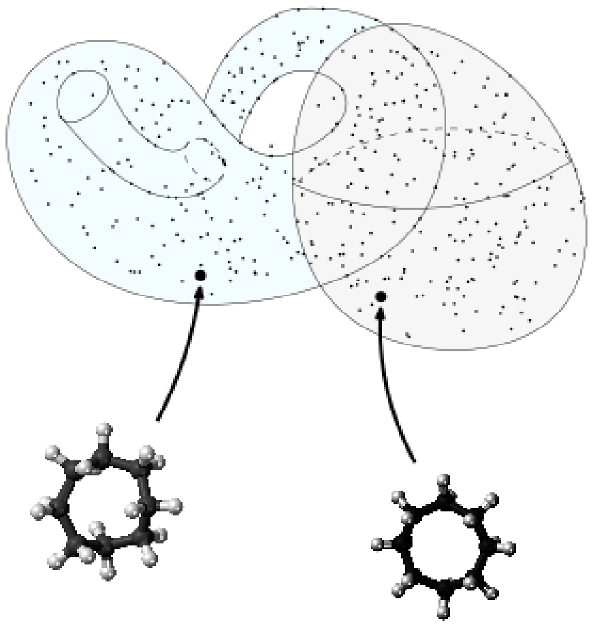
Talk II (/II): Persistence barcodes

https://raphaeltinarrage.github.io

Last update: April 28, 2021

Reminder of last talk

Some datasets contain topology

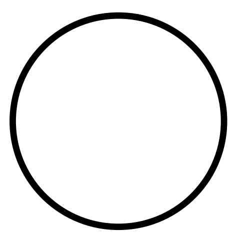
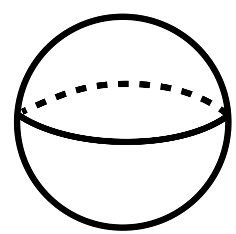
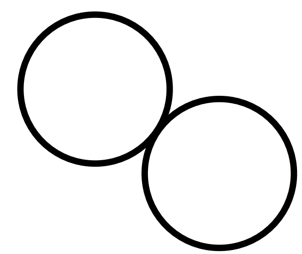
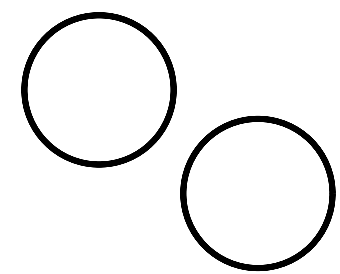


Invariants of homotopy classes allow to describe and understand topological spaces

Number of connected components

Euler characteristic χ

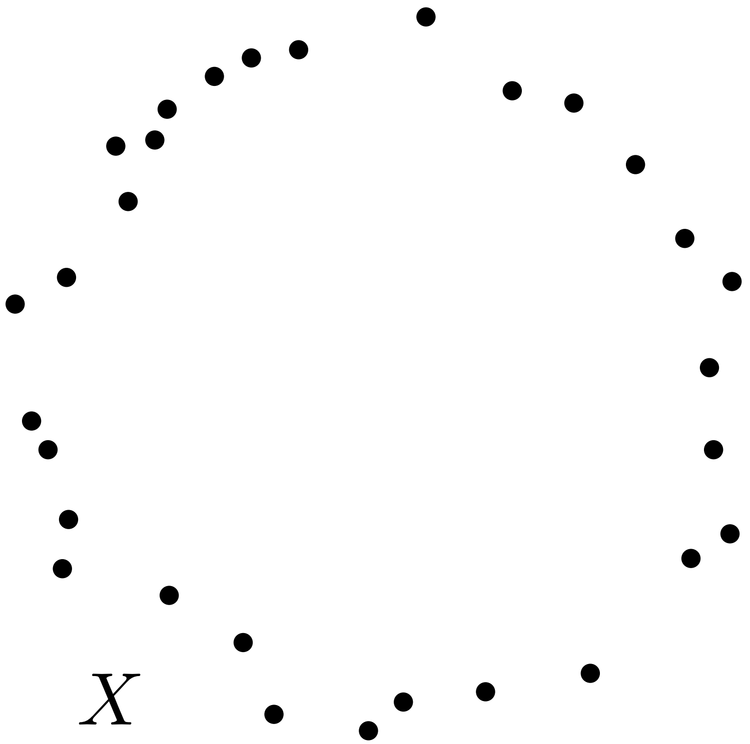
Betti numbers $\beta_0, \beta_1, \beta_2, \ldots$

X					
$\beta_0(X)$	1	1	1	1	2
$\beta_1(X)$	0	1	0	2	2
$\beta_2(X)$	0	0	1	0	0

Introduction

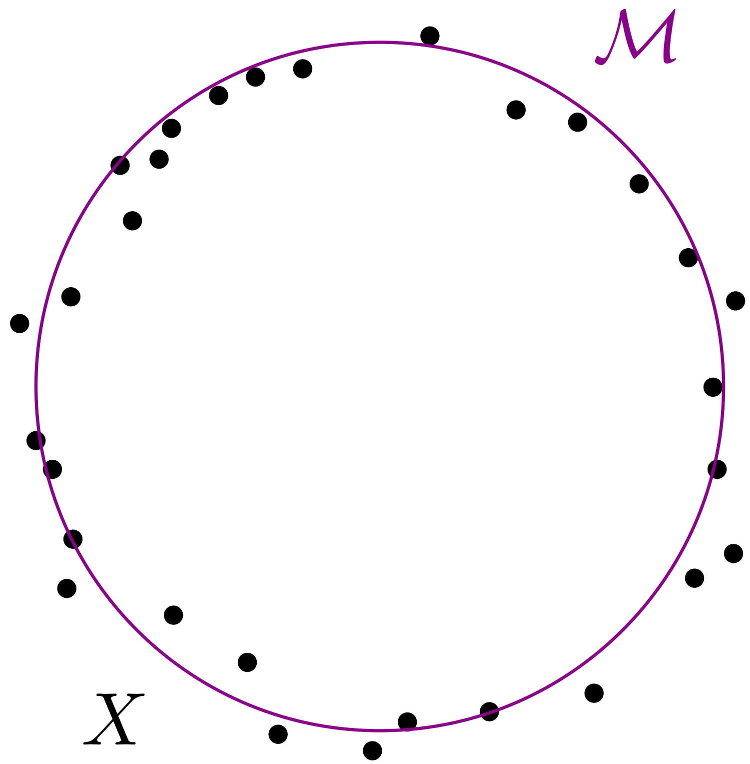
In real life, we are often given datasets that are subsets of the Euclidean space: $X \subset \mathbb{R}^n$.

Of course, X is finite.



In real life, we are often given datasets that are subsets of the Euclidean space: $X \subset \mathbb{R}^n$.

Of course, X is finite.



In Topological Data Analysis, we think of X as being a sample of **an underlying continuous object**, $\mathcal{M} \subset \mathbb{R}^n$.

Understanding the topology of $\mathcal{M}$ would give us interesting insights about our dataset.

I - Simplicial homology ≈ 15 min

1 - Homology groups

2 - Functoriality

II - Topological inference ≈ 10 min

1 - Parameter estimation

2 - Nerves

III - Persistent homology ≈ 20 min

1 - Persistence modules

2 - Decomposition

3 - Stability

IV - Applications ≈ 10 min

Overview

We will define **simplicial homology** over the field $\mathbb{Z}/2\mathbb{Z}$.

based on simplicial complexes

we will deal with linear algebra
over the field $\mathbb{Z}/2\mathbb{Z}$

We have to define:

- the chains,
- the boundary operators,
- the cycles and the boundaries,
- the homology groups.

Overview

The group $\mathbb{Z}/2\mathbb{Z}$ can be seen as the set $\{0, 1\}$ with the operation

$$\begin{aligned} 0 + 0 &= 0 \\ 0 + 1 &= 1 \\ 1 + 0 &= 1 \\ 1 + 1 &= 0 \end{aligned}$$

For any $n \geq 1$, the **product group** $(\mathbb{Z}/2\mathbb{Z})^n$ is the group whose underlying set is

$$(\mathbb{Z}/2\mathbb{Z})^n = \{(\epsilon_1, ..., \epsilon_n),\ \epsilon_1, ..., \epsilon_n \in \mathbb{Z}/2\mathbb{Z}\}$$

and whose operation is defined as

$$(\epsilon_1, ..., \epsilon_n) + (\epsilon'_1, ..., \epsilon'_n) = (\epsilon_1 + \epsilon'_1, ..., \epsilon_n + \epsilon'_n).$$

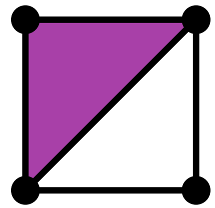
The group $\mathbb{Z}/2\mathbb{Z}$ can be given a **field** structure

$$\begin{aligned} 0 \times 0 &= 0 \\ 0 \times 1 &= 0 \\ 1 \times 0 &= 0 \\ 1 \times 1 &= 1 \end{aligned}$$

and $(\mathbb{Z}/2\mathbb{Z})^n$ can be seen as a $\mathbb{Z}/2\mathbb{Z}$-**vector space** over the field $\mathbb{Z}/2\mathbb{Z}$.

Chains

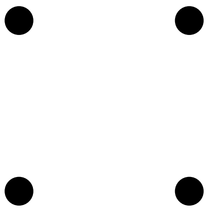
Definition (reminder): Let V be a set (called the set of *vertices*). A **simplicial complex** over V is a set K of subsets of V (called the *simplices*) such that, for every $\sigma \in K$ and every non-empty $\tau \subset \sigma$, we have $\tau \in K$.
The dimension of a simplex $\sigma \in K$ is $\dim(\sigma) = |\sigma| - 1$.



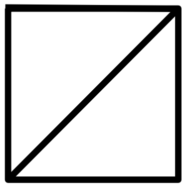
K

Let K be a simplicial complex. For any $n \geq 0$, define

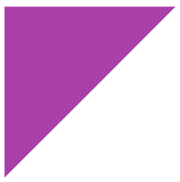
$$K_{(n)} = \{\sigma \in K,\ \dim(\sigma) = n\}.$$



$K_{(0)}$



$K_{(1)}$



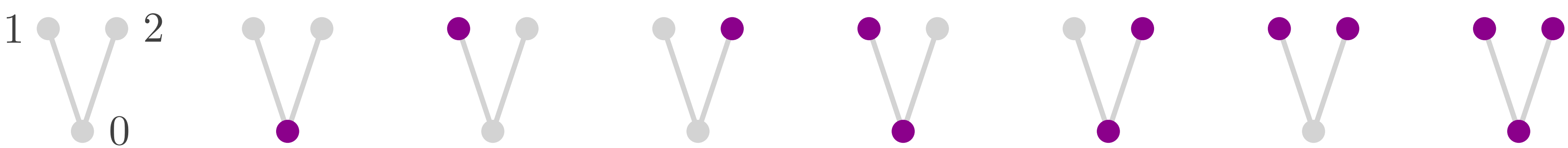
$K_{(2)}$

Chains

Let $n \geq 0$. The n-**chains** of K is the set $C_n(K)$ whose elements are the formal sums

$$\sum_{\sigma \in K_{(n)}} \epsilon_\sigma \cdot \sigma \quad \text{where} \quad \forall \sigma \in K_{(n)},\ \epsilon_\sigma \in \mathbb{Z}/2\mathbb{Z}.$$

Example: The 0-chains of $K = \{[0], [1], [2], [0,1], [0,2]\}$ are:



0 $\quad [0]$ $\quad [1]$ $\quad [2]$ $\quad [0]+[1]$ $\quad [0]+[2]$ $\quad [1]+[2]$ $\quad [0]+[1]+[2]$

and the 1-chains

0 $\quad [0,1]$ $\quad [0,2]$ $\quad [0,1]+[0,2]$

Chains

Let $n \geq 0$. The n-**chains** of K is the set $C_n(K)$ whose elements are the formal sums

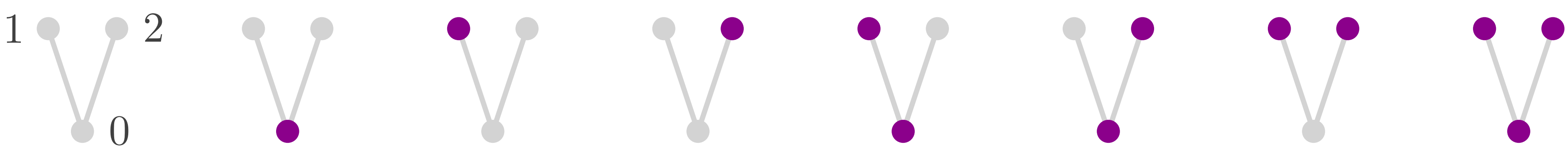
$$\sum_{\sigma \in K_{(n)}} \epsilon_\sigma \cdot \sigma \quad \text{where} \quad \forall \sigma \in K_{(n)},\ \epsilon_\sigma \in \mathbb{Z}/2\mathbb{Z}.$$

We can give $C_n(K)$ a **group structure** via

$$\sum_{\sigma \in K_{(n)}} \epsilon_\sigma \cdot \sigma + \sum_{\sigma \in K_{(n)}} \eta_\sigma \cdot \sigma = \sum_{\sigma \in K_{(n)}} (\epsilon_\sigma + \eta_\sigma) \cdot \sigma.$$

Moreover, $C_n(K)$ can be given a $\mathbb{Z}/2\mathbb{Z}$-vector space structure.

Example: The 0-chains of $K = \{[0], [1], [2], [0,1], [0,2]\}$ are:



0 $\quad [0]$ $\quad [1]$ $\quad [2]$ $\quad [0]+[1]$ $\quad [0]+[2]$ $\quad [1]+[2]$ $\quad [0]+[1]+[2]$

and the 1-chains

0 $\quad [0,1]$ $\quad [0,2]$ $\quad [0,1]+[0,2]$

Chains

Let $n \geq 0$. The n-**chains** of K is the set $C_n(K)$ whose elements are the formal sums

$$\sum_{\sigma \in K_{(n)}} \epsilon_\sigma \cdot \sigma \quad \text{where} \quad \forall \sigma \in K_{(n)},\ \epsilon_\sigma \in \mathbb{Z}/2\mathbb{Z}.$$

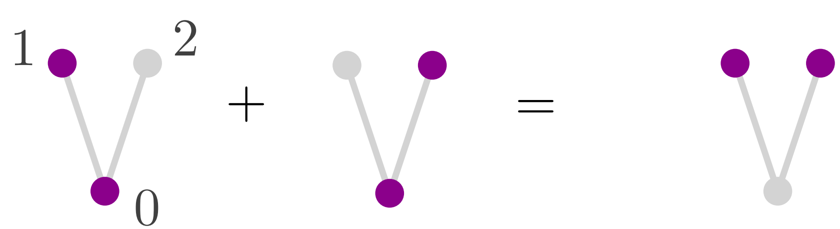
We can give $C_n(K)$ a **group structure** via

$$\sum_{\sigma \in K_{(n)}} \epsilon_\sigma \cdot \sigma + \sum_{\sigma \in K_{(n)}} \eta_\sigma \cdot \sigma = \sum_{\sigma \in K_{(n)}} (\epsilon_\sigma + \eta_\sigma) \cdot \sigma.$$

Moreover, $C_n(K)$ can be given a $\mathbb{Z}/2\mathbb{Z}$-vector space structure.

Example: In the simplicial complex $K = \{[0],[1],[2],[0,1],[0,2]\}$, the sum of the 0-chains $[0]+[1]$ and $[0]+[2]$ is $[1]+[2]$:

$$([0]+[1]) + ([0]+[2]) = [0]+[0]+[1]+[2] = [1]+[2].$$



Boundary operator

Let $n \geq 1$, and $\sigma = [x_0, ..., x_n] \in K_{(n)}$ a simplex of dimension n. We define its **boundary** as the following element of $C_{n-1}(K)$:

$$\partial_n \sigma = \sum_{\substack{\tau \subset \sigma \\ |\tau| = |\sigma| - 1}} \tau$$

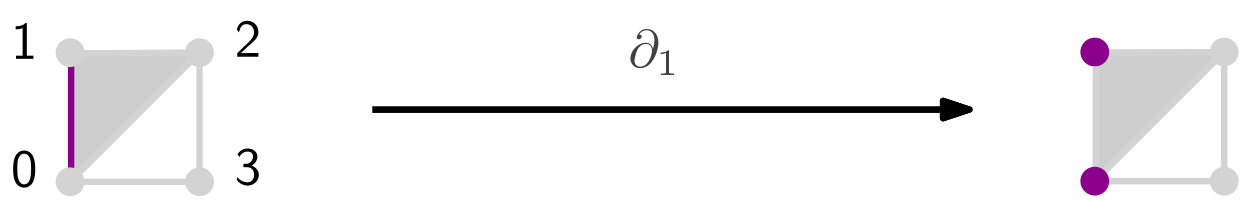
We can extend the operator ∂_n as a linear map $\partial_n \colon C_n(K) \to C_{n-1}(K)$.

Example: Consider the simplicial complex

$$K = \{[0], [1], [2], [3], [0,1], [0,2], [1,2], [1,3], [2,3], [0,1,2]\}.$$

The simplex $[0,1]$ has the faces $[0]$ and $[1]$. Hence

$$\partial_1 [0,1] = [0] + [1].$$



Boundary operator

Let $n \geq 1$, and $\sigma = [x_0, ..., x_n] \in K_{(n)}$ a simplex of dimension n. We define its **boundary** as the following element of $C_{n-1}(K)$:

$$\partial_n \sigma = \sum_{\substack{\tau \subset \sigma \\ |\tau| = |\sigma| - 1}} \tau$$

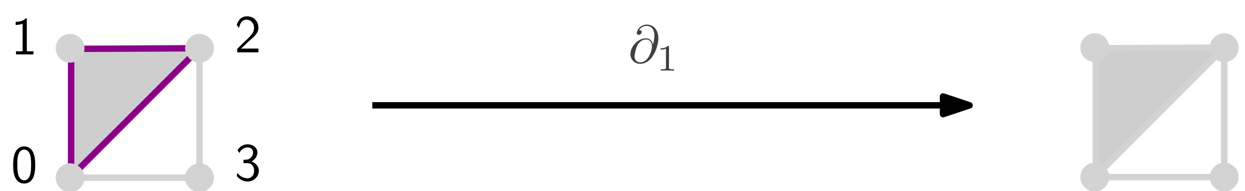
We can extend the operator ∂_n as a linear map $\partial_n \colon C_n(K) \to C_{n-1}(K)$.

Example: Consider the simplicial complex

$$K = \{[0], [1], [2], [3], [0, 1], [0, 2], [1, 2], [1, 3], [2, 3], [0, 1, 2]\}.$$

The boundary of the 1-chain $[0, 1] + [1, 2] + [2, 0]$ is

$$\begin{aligned} \partial_1 \big([0, 1] + [1, 2] + [2, 0]\big) &= \partial_1 [0, 1] + \partial_1 [1, 2] + \partial_1 [2, 0] \\ &= [0] + [1] + [1] + [2] + [2] + [0] = 0 \end{aligned}$$



Boundary operator

Let $n \geq 1$, and $\sigma = [x_0, ..., x_n] \in K_{(n)}$ a simplex of dimension n. We define its **boundary** as the following element of $C_{n-1}(K)$:

$$\partial_n \sigma = \sum_{\substack{\tau \subset \sigma \\ |\tau| = |\sigma| - 1}} \tau$$

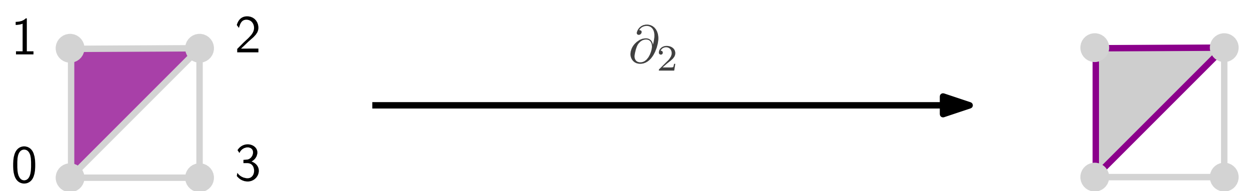
We can extend the operator ∂_n as a linear map $\partial_n : C_n(K) \to C_{n-1}(K)$.

Example: Consider the simplicial complex

$$K = \{[0], [1], [2], [3], [0,1], [0,2], [1,2], [1,3], [2,3], [0,1,2]\}.$$

The simplex $[0,1,2]$ has the faces $[0,1]$ and $[1,2]$ and $[2,0]$. Hence

$$\partial_2 [0,1,2] = [0,1] + [1,2] + [2,0].$$



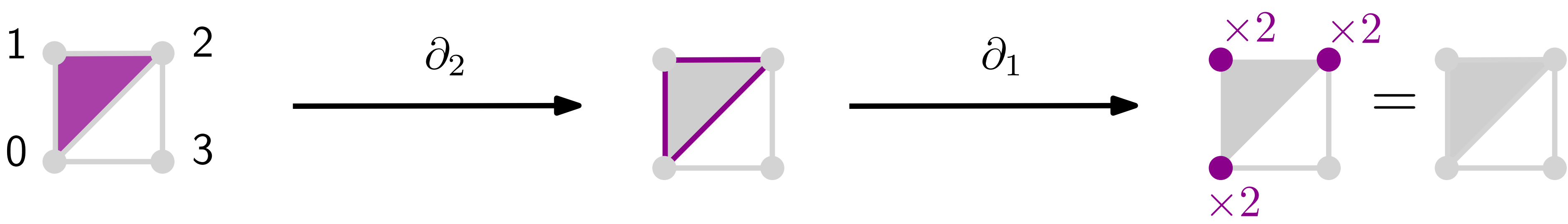
Boundary operator

Let $n \geq 1$, and $\sigma = [x_0, ..., x_n] \in K_{(n)}$ a simplex of dimension n. We define its **boundary** as the following element of $C_{n-1}(K)$:

$$\partial_n \sigma = \sum_{\substack{\tau \subset \sigma \\ |\tau| = |\sigma| - 1}} \tau$$

We can extend the operator ∂_n as a linear map $\partial_n : C_n(K) \to C_{n-1}(K)$.

Proposition: For any $n \geq 1$, for any $c \in C_n(K)$, we have $\partial_{n-1} \circ \partial_n(c) = 0$.



Cycles and boundaries

Let $n \geq 0$. We have a sequence of vector spaces

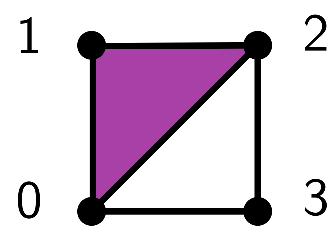
$$\cdots \longrightarrow C_{n+1}(K) \xrightarrow{\partial n+1} C_n(K) \xrightarrow{\partial n} C_{n-1}(K) \longrightarrow \ldots$$

The maps ∂_{n+1} and ∂_n are linear maps, and we can consider their kernel and image.

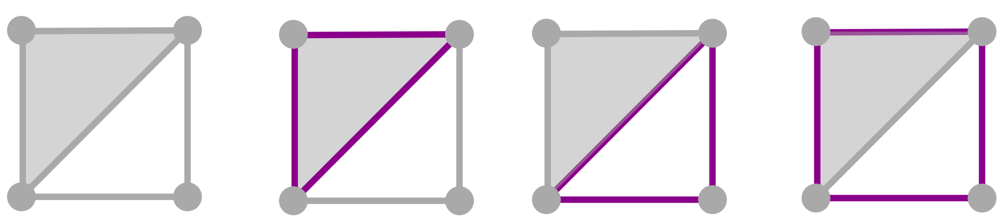
Definition: We define:

- The n-**cycles**: $Z_n(K) = \text{Ker}(\partial_n)$,
- The n-**boundaries**: $B_n(K) = \text{Im}(\partial_{n+1})$.

Example: Consider the simplicial complex



The 1-cycles are:



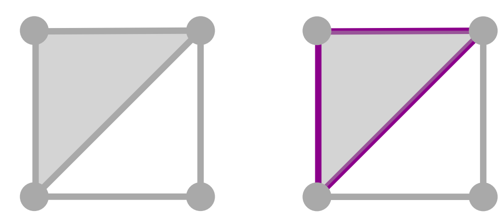
0

$[0,1]+[1,2]+[0,2]$

$[0,2]+[2,3]+[0,3]$

$[0,1]+[1,2]+[2,3]+[0,3].$

The 1-boundaries are:



0

$[0,1]+[1,2]+[0,2]$

Cycles and boundaries

Let $n \geq 0$. We have a sequence of vector spaces

$$\cdots \longrightarrow C_{n+1}(K) \xrightarrow{\partial n+1} C_n(K) \xrightarrow{\partial n} C_{n-1}(K) \longrightarrow \ldots$$

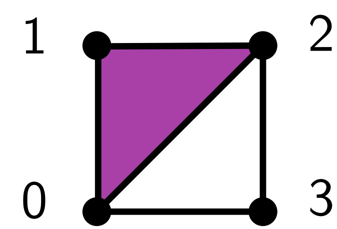
The maps ∂_{n+1} and ∂_n are linear maps, and we can consider their kernel and image.

Definition: We define:

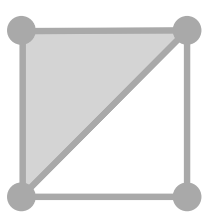
- The n-**cycles**: $Z_n(K) = \mathrm{Ker}(\partial_n)$,
- The n-**boundaries**: $B_n(K) = \mathrm{Im}(\partial_{n+1})$.

Proposition: We have $B_n(K) \subset Z_n(K)$.

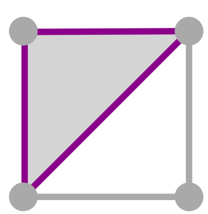
Example: Consider the simplicial complex



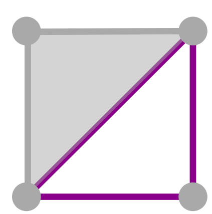
The 1-cycles are:



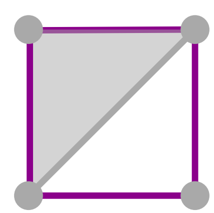
0



$[0,1] + [1,2] + [0,2]$

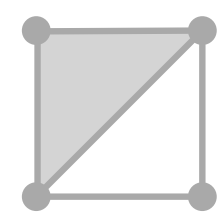


$[0,2] + [2,3] + [0,3]$

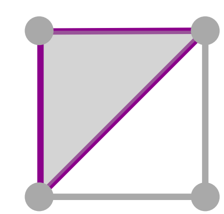


$[0,1] + [1,2] + [2,3] + [0,3]$.

The 1-boundaries are:



0



$[0,1] + [1,2] + [0,2]$

Cycles and boundaries

Let $n \geq 0$. We have a sequence of vector spaces

$$\cdots \longrightarrow C_{n+1}(K) \xrightarrow{\partial n+1} C_n(K) \xrightarrow{\partial n} C_{n-1}(K) \longrightarrow \ldots$$

The maps ∂_{n+1} and ∂_n are linear maps, and we can consider their kernel and image.

Definition: We define:

- The n-**cycles**: $Z_n(K) = \mathrm{Ker}(\partial_n)$,
- The n-**boundaries**: $B_n(K) = \mathrm{Im}(\partial_{n+1})$.

Proposition: We have $B_n(K) \subset Z_n(K)$.

Definition: We say that two chains $c, c' \in C_n(K)$ are **homologous** if there exists $b \in B_n(K)$ such that $c = c' + b$.

$\longrightarrow$ two chains are homologous if they are equal up to a boundary

Example:

$$[0,2] + [2,3] + [0,3] = [0,1] + [1,2] + [2,3] + [0,3] + [0,1] + [0,2] + [1,2].$$

Cycles and boundaries

Let $n \geq 0$. We have a sequence of vector spaces

$$\cdots \longrightarrow C_{n+1}(K) \xrightarrow{\partial n+1} C_n(K) \xrightarrow{\partial n} C_{n-1}(K) \longrightarrow \ldots$$

The maps ∂_{n+1} and ∂_n are linear maps, and we can consider their kernel and image.

Definition: We define:

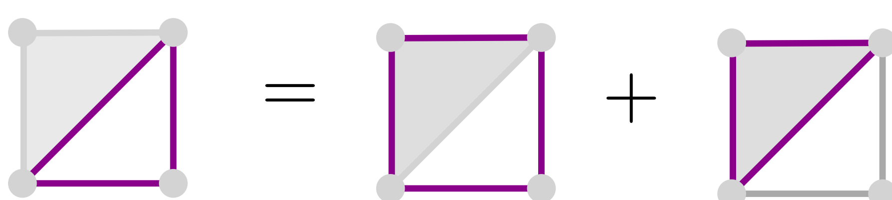
- The n-**cycles**: $Z_n(K) = \mathrm{Ker}(\partial_n)$,
- The n-**boundaries**: $B_n(K) = \mathrm{Im}(\partial_{n+1})$.

Proposition: We have $B_n(K) \subset Z_n(K)$.

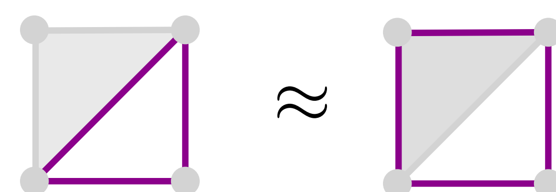
Definition: We say that two chains $c, c' \in C_n(K)$ are **homologous** if there exists $b \in B_n(K)$ such that $c = c' + b$.

$\longrightarrow$ interpretation: two cycles are homologous if they represent the same ‘hole’

Example:



hence



$$[0,2] + [2,3] + [0,3] = [0,1] + [1,2] + [2,3] + [0,3] + [0,1] + [0,2] + [1,2].$$

Homology groups

We have defined a sequence of vector spaces, connected by linear maps

$$\cdots \longrightarrow C_{n+1}(K) \longrightarrow C_n(K) \longrightarrow C_{n-1}(K) \longrightarrow \cdots$$

and for every $n \geq 0$, we have defined the cycles and the boundaries $Z_n(K)$ and $B_n(K)$.

Since $B_n(K) \subset Z_n(K)$, we can see $B_n(K)$ as a linear subspace of $Z_n(K)$.

Definition: The n^{th} **(simplicial) homology group** of K is the quotient vector space

$$H_n(K) = Z_n(K)/B_n(K).$$

We have defined a sequence of vector spaces, connected by linear maps

$$\cdots \longrightarrow C_{n+1}(K) \longrightarrow C_n(K) \longrightarrow C_{n-1}(K) \longrightarrow \cdots$$

and for every $n \geq 0$, we have defined the cycles and the boundaries $Z_n(K)$ and $B_n(K)$.

Since $B_n(K) \subset Z_n(K)$, we can see $B_n(K)$ as a linear subspace of $Z_n(K)$.

Definition: The n^{th} **(simplicial) homology group** of K is the quotient vector space

$$H_n(K) = Z_n(K)/B_n(K).$$

Remark: A finite $\mathbb{Z}/2\mathbb{Z}$-vector space must be isomorphic to $(\mathbb{Z}/2\mathbb{Z})^k$ for some k.

Definition: Let K be a simplicial complex and $n \geq 0$. Its n^{th} **Betti number** is the integer $\beta_n(K) = \dim H_n(K)$.

$$H_n(K) = (\mathbb{Z}/2\mathbb{Z})^k \quad \longrightarrow \quad \beta_n(K) = k$$

Homology groups

We have defined a sequence of vector spaces, connected by linear maps

$$\cdots \longrightarrow C_{n+1}(K) \longrightarrow C_n(K) \longrightarrow C_{n-1}(K) \longrightarrow \cdots$$

and for every $n \geq 0$, we have defined the cycles and the boundaries $Z_n(K)$ and $B_n(K)$.

Since $B_n(K) \subset Z_n(K)$, we can see $B_n(K)$ as a linear subspace of $Z_n(K)$.

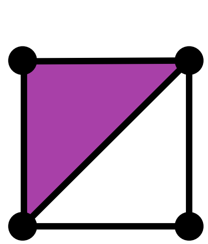
Definition: The n^{th} **(simplicial) homology group** of K is the quotient vector space

$$H_n(K) = Z_n(K)/B_n(K).$$

Remark: A finite $\mathbb{Z}/2\mathbb{Z}$-vector space must be isomorphic to $(\mathbb{Z}/2\mathbb{Z})^k$ for some k.

Definition: Let K be a simplicial complex and $n \geq 0$. Its n^{th} **Betti number** is the integer $\beta_n(K) = \dim H_n(K)$.

Example:



$$H_0(K) = \mathbb{Z}/2\mathbb{Z} \longrightarrow \beta_0(K) = 1$$

$$H_1(K) = \mathbb{Z}/2\mathbb{Z} \longrightarrow \beta_1(K) = 1$$

$$H_2(K) = 0 \longrightarrow \beta_2(K) = 0$$

Homology groups

We have defined a sequence of vector spaces, connected by linear maps

$$\cdots \longrightarrow C_{n+1}(K) \longrightarrow C_n(K) \longrightarrow C_{n-1}(K) \longrightarrow \cdots$$

and for every $n \geq 0$, we have defined the cycles and the boundaries $Z_n(K)$ and $B_n(K)$.

Since $B_n(K) \subset Z_n(K)$, we can see $B_n(K)$ as a linear subspace of $Z_n(K)$.

Definition: The n^{th} **(simplicial) homology group** of K is the quotient vector space

$$H_n(K) = Z_n(K)/B_n(K).$$

Remark: A finite $\mathbb{Z}/2\mathbb{Z}$-vector space must be isomorphic to $(\mathbb{Z}/2\mathbb{Z})^k$ for some k.

Definition: Let K be a simplicial complex and $n \geq 0$. Its n^{th} **Betti number** is the integer $\beta_n(K) = \dim H_n(K)$.

Proposition: If X and Y are two homotopy equivalent topological spaces, then for any $n \geq 0$ we have isomorphic homology groups $H_n(X) \simeq H_n(Y)$.
As a consequence, $\beta_n(X) = \beta_n(Y)$.

Homology groups

We have defined a sequence of vector spaces, connected by linear maps

$$\cdots \longrightarrow C_{n+1}(K) \longrightarrow C_n(K) \longrightarrow C_{n-1}(K) \longrightarrow \cdots$$

and for every $n \geq 0$, we have defined the cycles and the boundaries $Z_n(K)$ and $B_n(K)$.

Since $B_n(K) \subset Z_n(K)$, we can see $B_n(K)$ as a linear subspace of $Z_n(K)$.

Definition: The n^{th} **(simplicial) homology group** of K is the quotient vector space

$$H_n(K) = Z_n(K)/B_n(K).$$

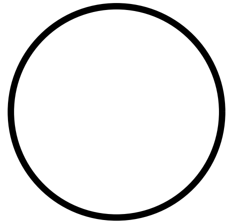
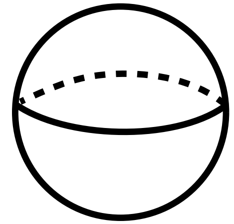
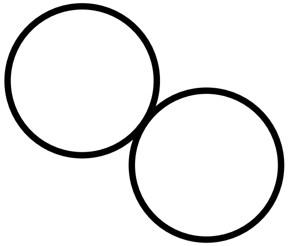
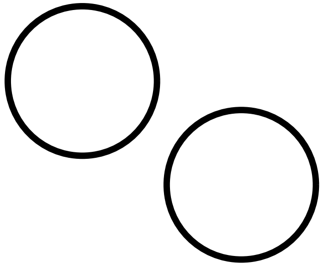
Remark: A finite $\mathbb{Z}/2\mathbb{Z}$-vector space must be isomorphic to $(\mathbb{Z}/2\mathbb{Z})^k$ for some k.

Definition: Let K be a simplicial complex and $n \geq 0$. Its n^{th} **Betti number** is the integer $\beta_n(K) = \dim H_n(K)$.

Proposition: If X and Y are two homotopy equivalent topological spaces, then for any $n \geq 0$ we have isomorphic homology groups $H_n(X) \simeq H_n(Y)$.
As a consequence, $\beta_n(X) = \beta_n(Y)$.

$\longrightarrow$ the theory works better with *singular homology*

Homology groups

X					
$H_0(X)$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$(\mathbb{Z}/2\mathbb{Z})^2$
$\beta_0(X)$	1	1	1	1	2
$H_1(X)$	0	$\mathbb{Z}/2\mathbb{Z}$	0	$(\mathbb{Z}/2\mathbb{Z})^2$	$(\mathbb{Z}/2\mathbb{Z})^2$
$\beta_1(X)$	0	1	0	2	2
$H_2(X)$	0	0	$\mathbb{Z}/2\mathbb{Z}$	0	0
$\beta_2(X)$	0	0	1	0	0

Homology is a functor

We have seen that homology transforms topological spaces into vector spaces

$$\begin{aligned} H_i \colon \mathrm{Top} &\longrightarrow \mathrm{Vect} \\ X &\longmapsto H_i(X) \end{aligned}$$

Actually, it also transforms **continous maps** into **linear maps**

$$X \xrightarrow{f} Y \qquad\qquad H_n(X) \xrightarrow{H_n(f)} H_n(Y)$$

This operation preserves **commutative diagrams**:

$$\begin{tikzcd} X \arrow[r, "f"'] \arrow[rr, bend left, "g\circ f"] & Y \arrow[r, "g"'] & Z, \end{tikzcd} \qquad \begin{tikzcd} H_n(X) \arrow[r, "H_n(f)"'] \arrow[rr, bend left, "H_n(g\circ f)"] & H_n(Y) \arrow[r, "H_n(g)"'] & H_n(Z). \end{tikzcd}$$

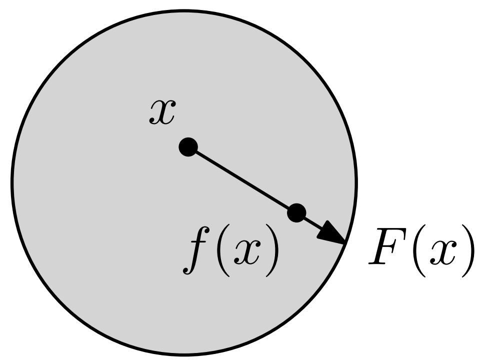
$$\boxed{H_n(g \circ f) = H_n(g) \circ H_n(f)}$$

Application in theory

Application: Brouwer's fixed point theorem

Let $f\colon \mathcal{B} \to \mathcal{B}$ be a continous map, where $\mathcal{B}$ is the unit closed ball of $\mathbb{R}^n$. Let us show that f has a fixed point ($f(x) = x$).

If not, we can define a map $F\colon \mathcal{B} \to \partial\mathcal{B}$ such that F restricted to $\partial\mathcal{B}$ is the identity. To do so, define $F(x)$ as the first intersection between the half-line $[x, f(x))$ and $\partial\mathcal{B}$.



Denote the inclusion $i\colon \partial\mathcal{B} \to \mathcal{B}$. Then $F \circ i\colon \partial\mathcal{B} \to \partial\mathcal{B}$ is the identity.
By functoriality, we have commutative diagrams

$$\begin{tikzcd} \partial\mathcal{B} \arrow[r, "i"'] \arrow[rr, bend left, "\mathrm{id}"] & \mathcal{B} \arrow[r, "F"'] & \partial\mathcal{B}, \end{tikzcd} \qquad \begin{tikzcd} H_i(\partial\mathcal{B}) \arrow[r, "H_i(i)"'] \arrow[rr, bend left, "H_i(\mathrm{id})"] & H_i(\mathcal{B}) \arrow[r, "H_i(F)"'] & H_i(\partial\mathcal{B}). \end{tikzcd}$$

But for $i = n - 1$, we have an absurdity:

$$\begin{tikzcd} \mathbb{Z}/2\mathbb{Z} \arrow[r] \arrow[rr, bend left, "id"] & 0 \arrow[r] & \mathbb{Z}/2\mathbb{Z}. \end{tikzcd}$$

I - Simplicial homology

1 - Homology groups

2 - Functoriality

II - Topological inference

1 - Parameter estimation

2 - Nerves

III - Persistent homology

1 - Persistence modules

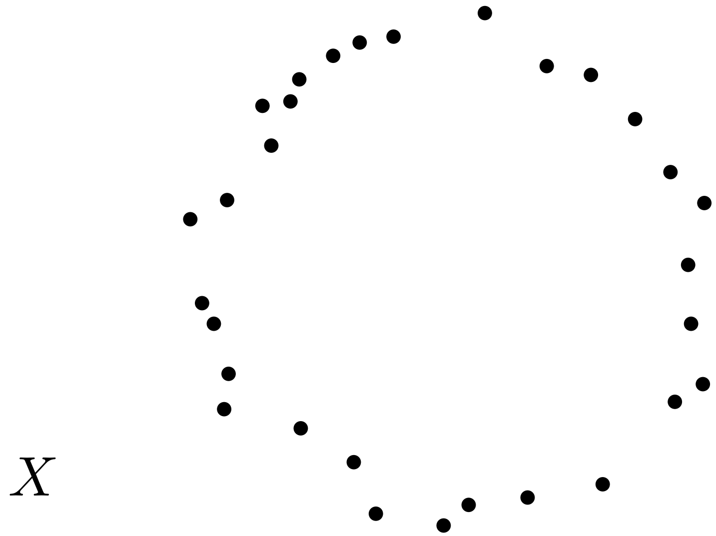
2 - Decomposition

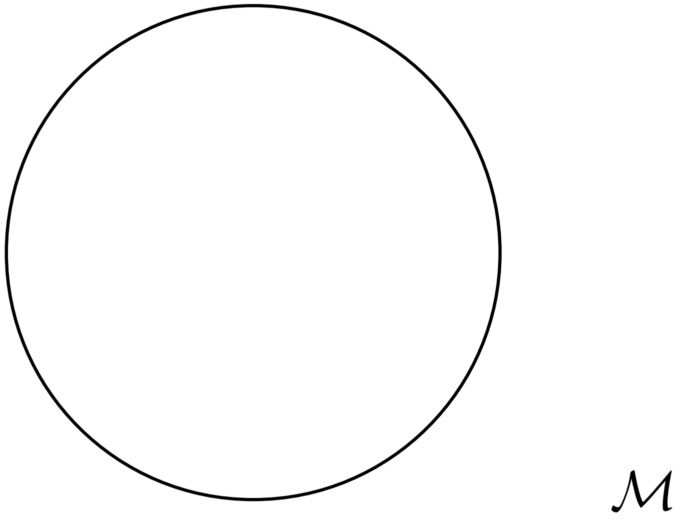
3 - Stability

IV - Applications

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.

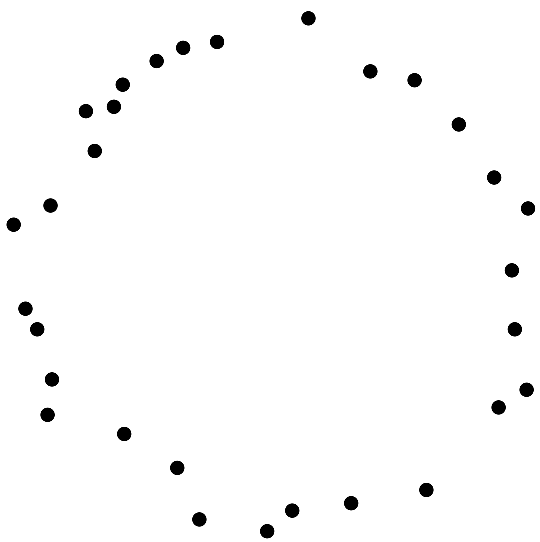


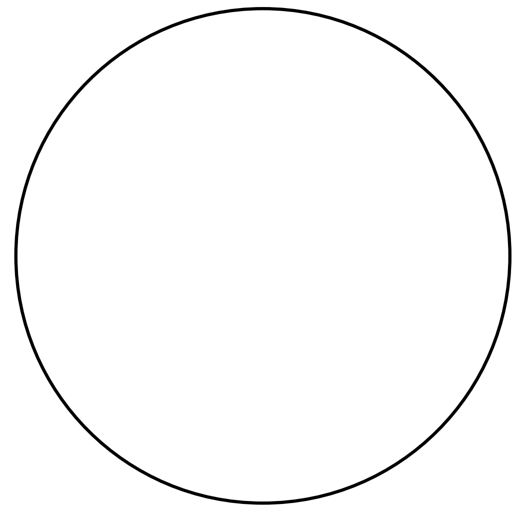


Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.

X





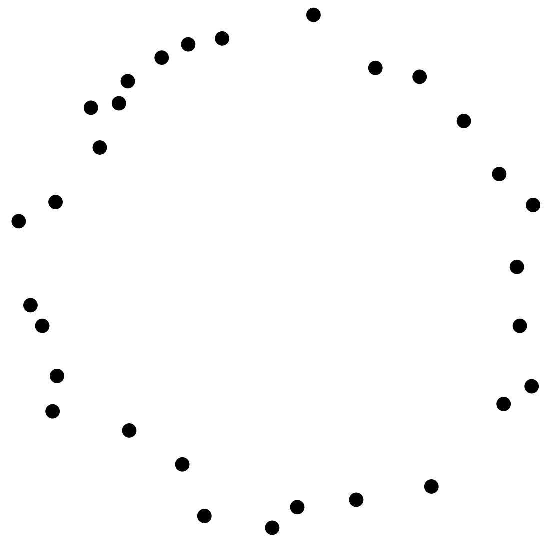
$\mathcal{M}$

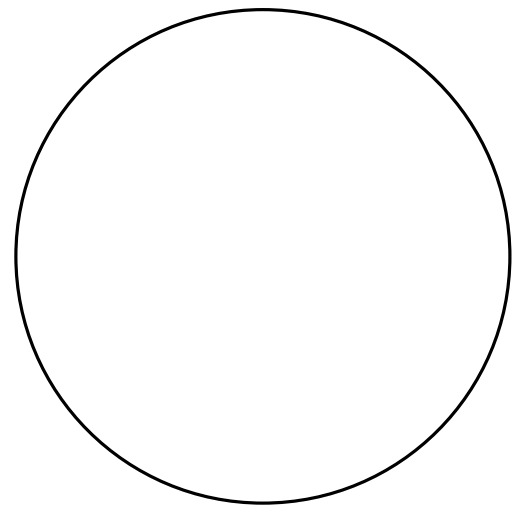
We cannot use X directly. Its homology is disapointing:

$$\beta_0(X) = 30 \quad \text{and} \quad \beta_i(X) = 0 \;\text{ for }\; i \geq 1$$

number of connected components
$=$ number of points of X

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.

X 

 $\mathcal{M}$

We cannot use X directly.

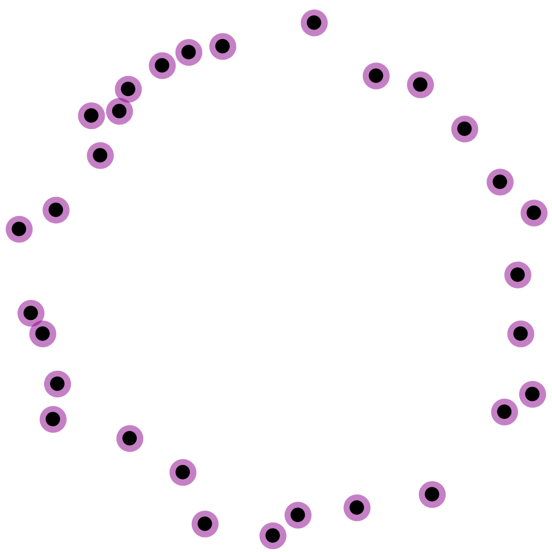
Idea: Thicken X.

Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

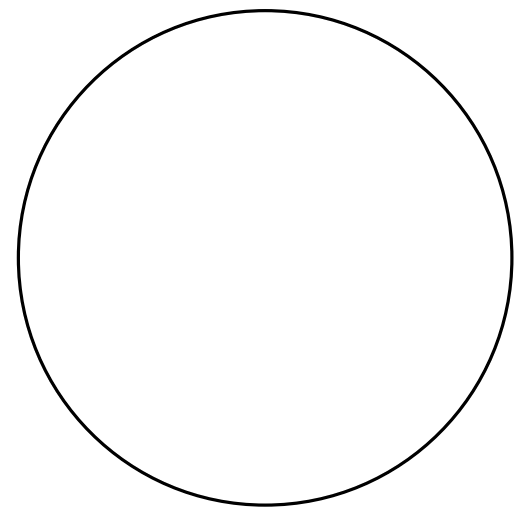
$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}.$$

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.



$X^{0.05}$



$\mathcal{M}$

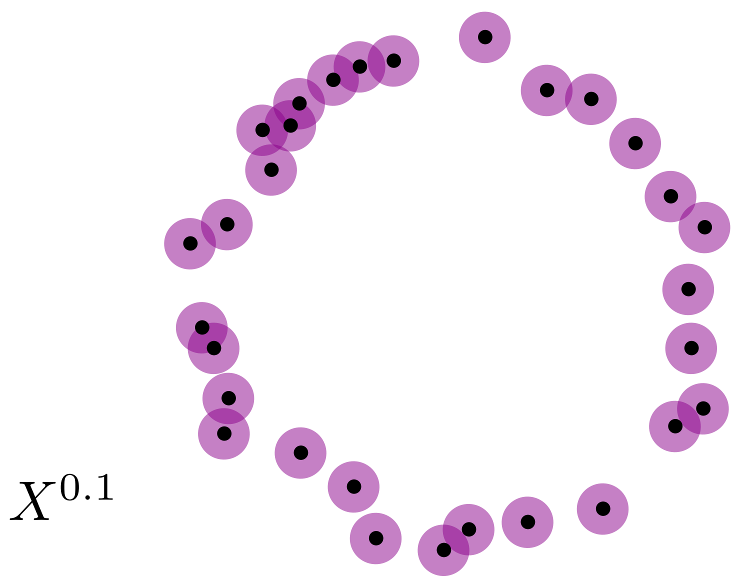
We cannot use X directly.

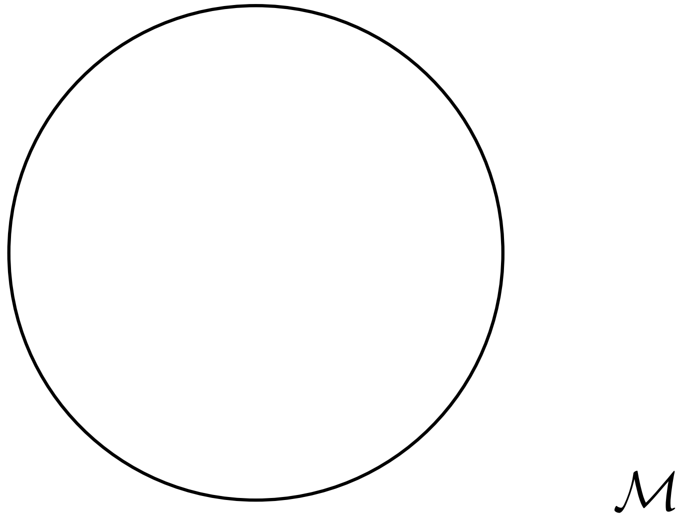
Idea: Thicken X.

Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}\,.$$

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.





We cannot use X directly.

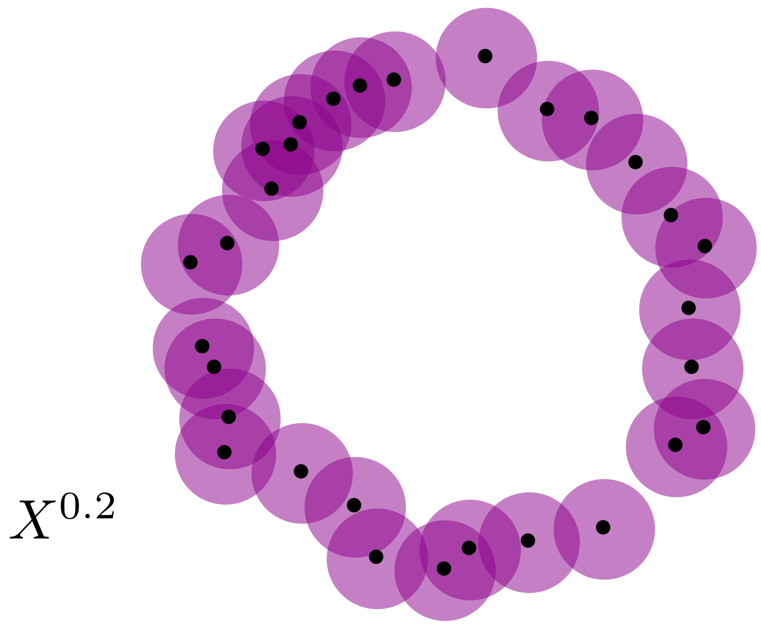
Idea: Thicken X.

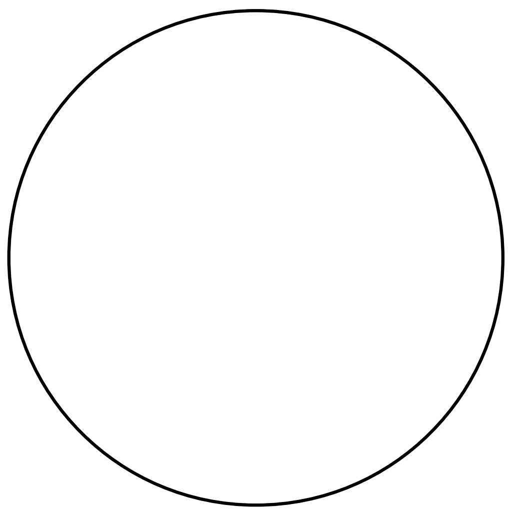
Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}.$$

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.





We cannot use X directly.

Idea: Thicken X.

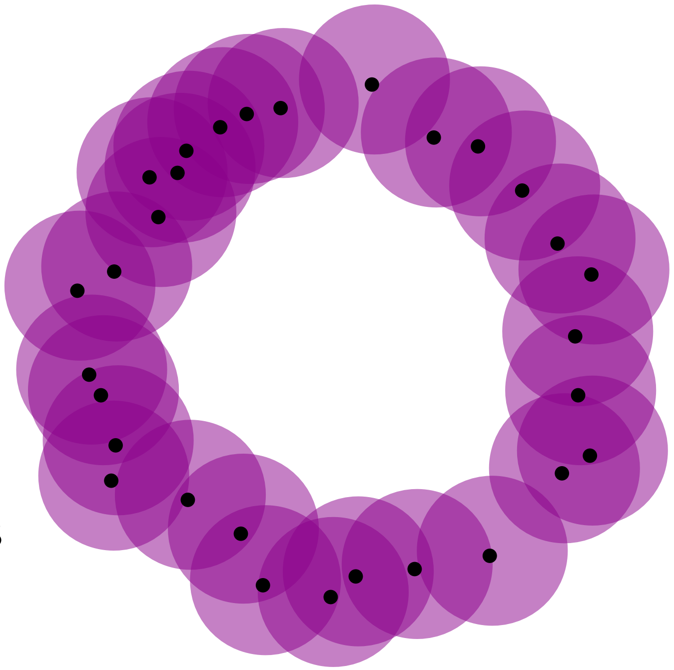
Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

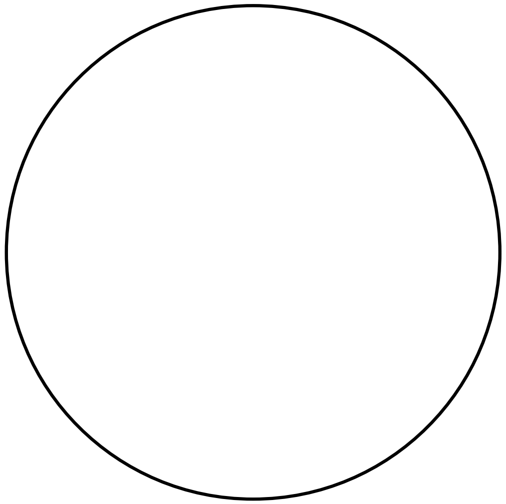
$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}\,.$$

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.

$X^{0.3}$





$\mathcal{M}$

We cannot use X directly.

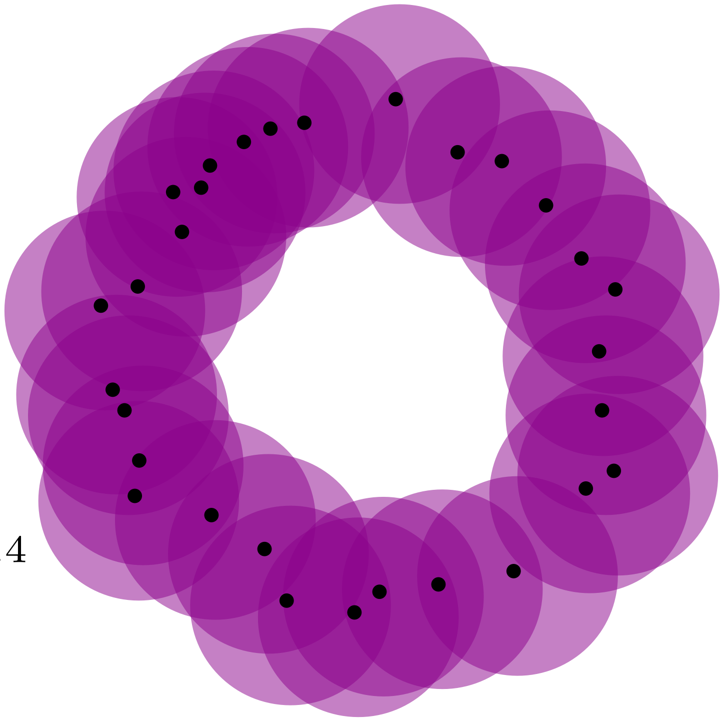
Idea: Thicken X.

Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

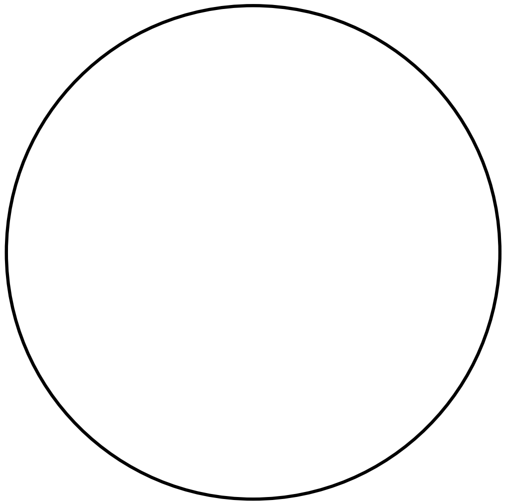
$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}.$$

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.



$X^{0.4}$



$\mathcal{M}$

We cannot use X directly.

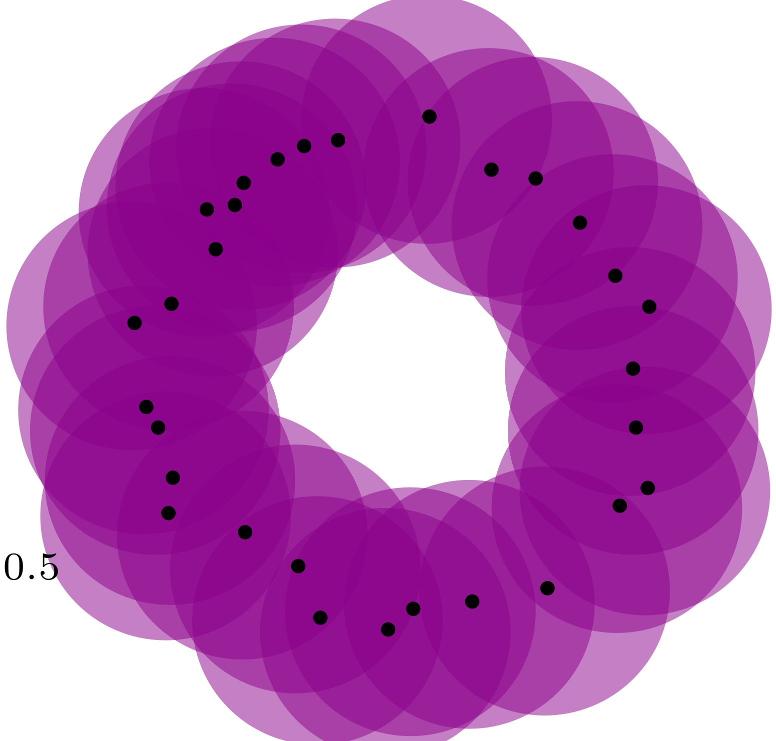
Idea: Thicken X.

Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

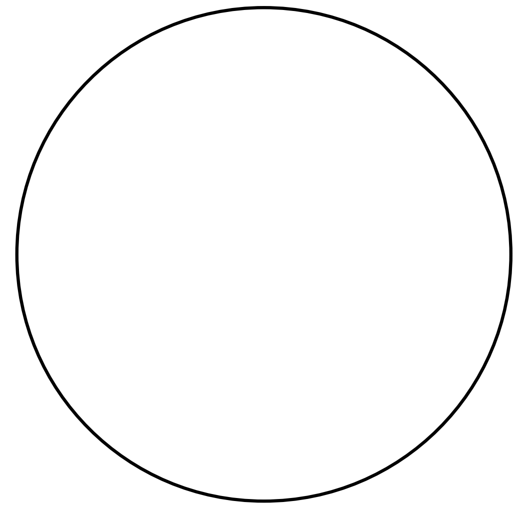
$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}.$$

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.



$X^{0.5}$



$\mathcal{M}$

We cannot use X directly.

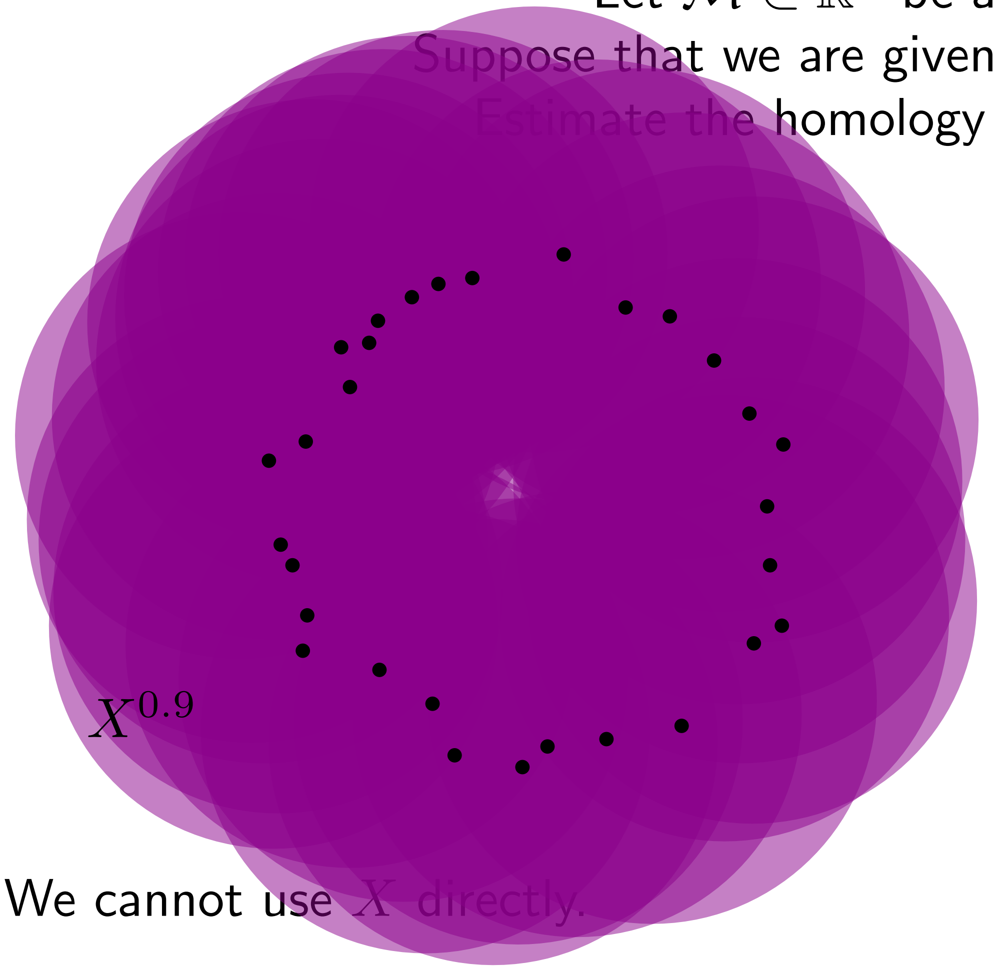
Idea: Thicken X.

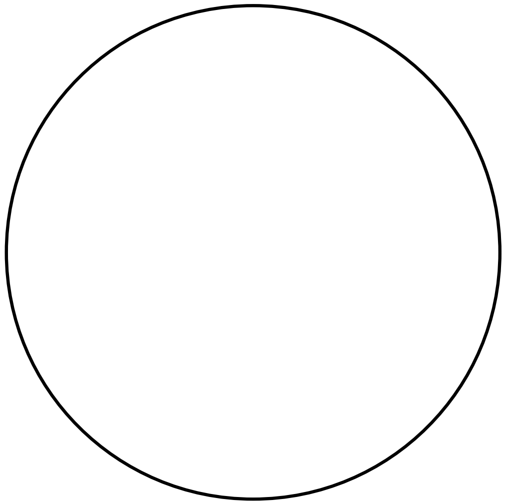
Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}.$$

Homological inference problem

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.





We cannot use X directly.

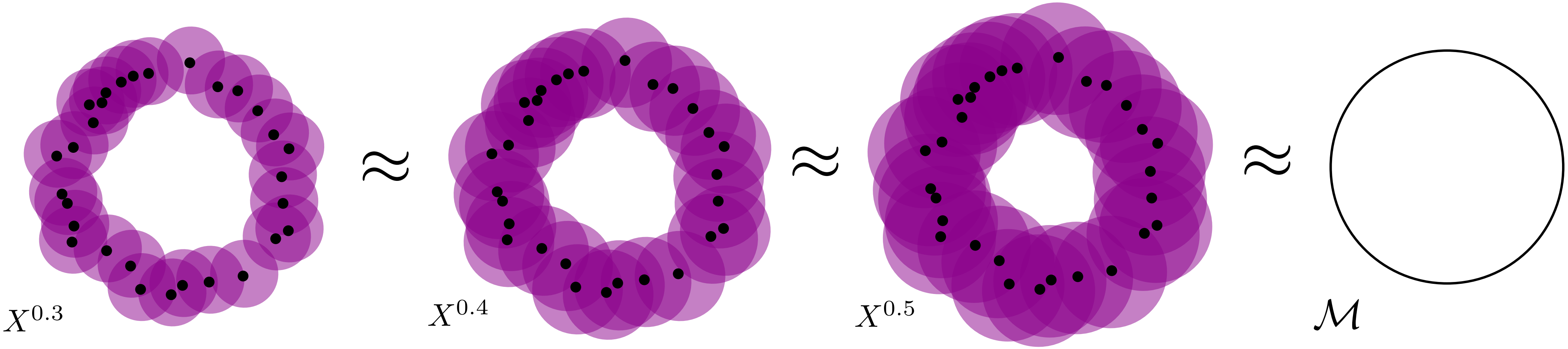
Idea: Thicken X.

Definition: For every $t \geq 0$, the *t-thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

$$X^t = \{y \in \mathbb{R}^n, \exists x \in X, \|x - y\| \leq t\}.$$

Homological inference problem

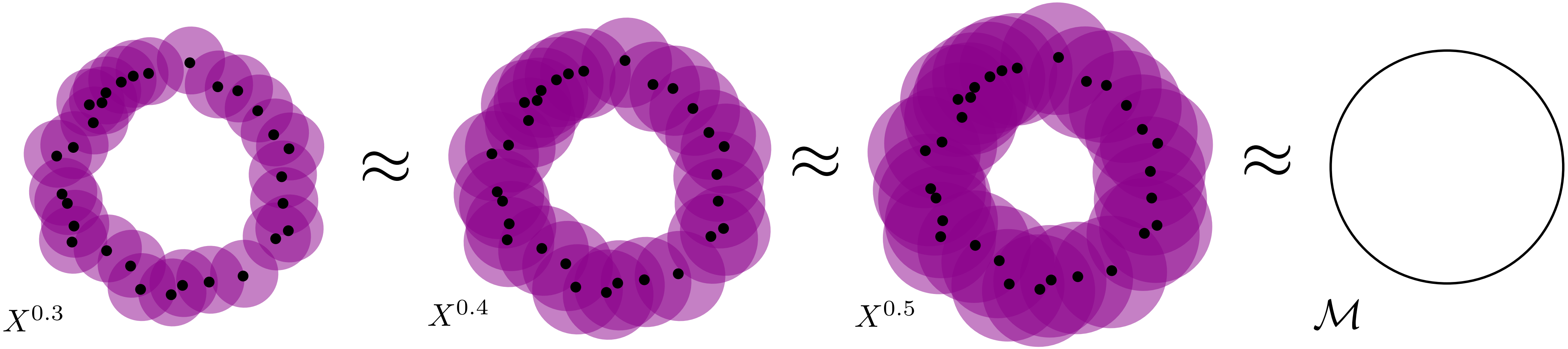
Some thickenings are homotopy equivalent to $\mathcal{M}$.



Hence we can recover the homology of $\mathcal{M}$:

$$\beta_0(\mathcal{M}) = \beta_0(X^{0.3})$$
$$\beta_1(\mathcal{M}) = \beta_1(X^{0.3})$$
$$\beta_2(\mathcal{M}) = \beta_2(X^{0.3})$$
$$\dots$$

Some thickenings are homotopy equivalent to $\mathcal{M}$.



Hence we can recover the homology of $\mathcal{M}$:

$$\begin{gathered}\beta_0(\mathcal{M}) = \beta_0(X^{0.3})\\ \beta_1(\mathcal{M}) = \beta_1(X^{0.3})\\ \beta_2(\mathcal{M}) = \beta_2(X^{0.3})\\ \dots\end{gathered}$$

Question 1: How to select a t such that $X^t \approx \mathcal{M}$?

Question 2: How to compute the homology groups of X^t?

I - Simplicial homology

- 1 - Homology groups
- 2 - Functoriality

II - Topological inference

- **1 - Parameter estimation**
- 2 - Nerves

III - Persistent homology

- 1 - Persistence modules
- 2 - Decomposition
- 3 - Stability

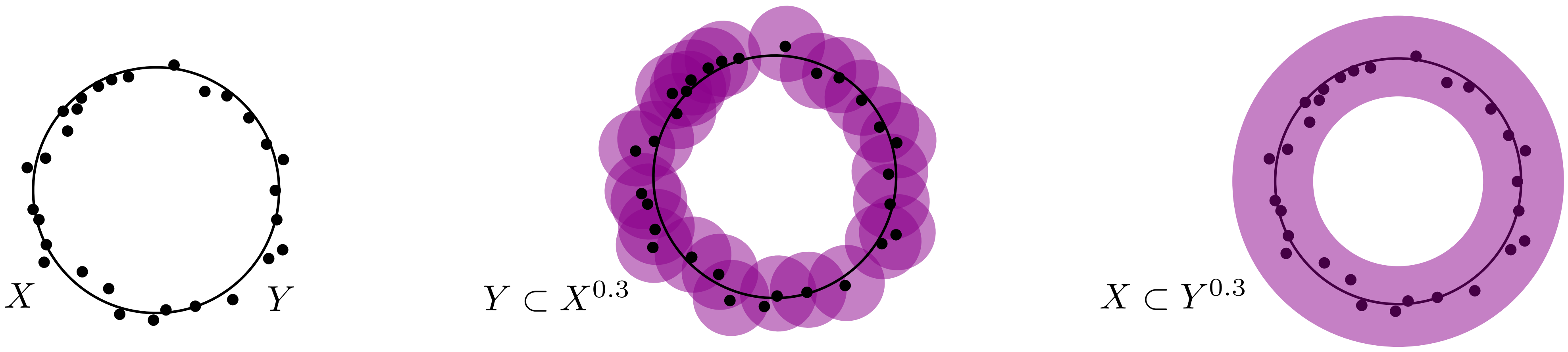
IV - Applications

Hausdorff distance

Definition: Let $X, Y \subset \mathbb{R}^n$ be (compact) subsets. The **Hausdorff distance** between X and Y is

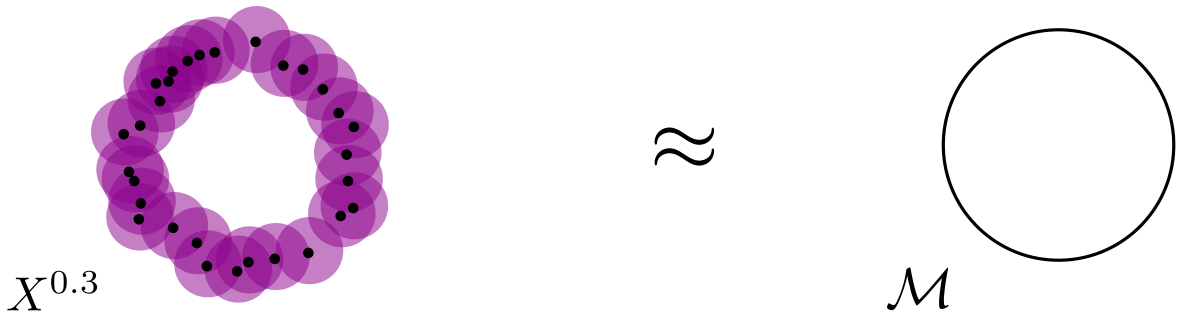
$$\mathrm{d_H}(X,Y) = \max\left\{ \sup_{y\in Y} \inf_{x\in X} \|x-y\|\,,\ \sup_{x\in X} \inf_{y\in Y} \|x-y\| \right\}.$$

Property: The Hausdorff distance is equal to $\inf\{t \geq 0, X \subset Y^t \text{ and } Y \subset X^t\}$.



Selection of the parameter t

Question 1: How to select a t such that $X^t \approx \mathcal{M}$?

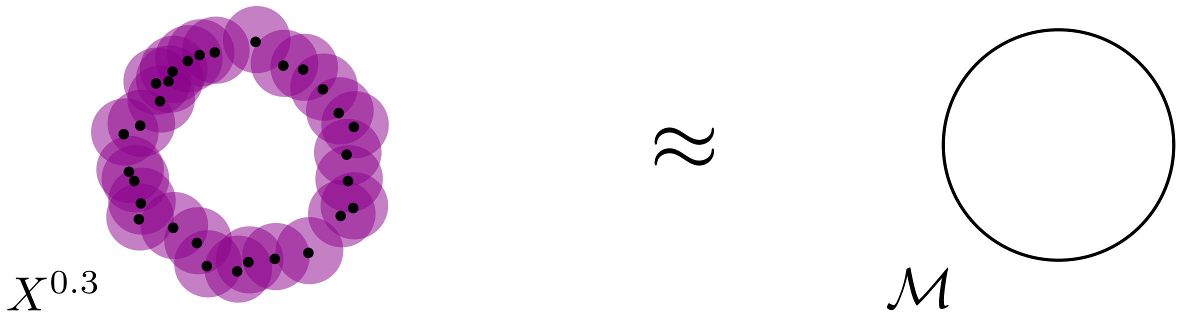


Theorem (Frédéric Chazal, David Cohen-Steiner, and André Lieutier, 2009):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$. Suppose that $\mathcal{M}$ has positive reach, and that $\mathrm{d_H}(X, \mathcal{M}) \leq \frac{1}{17}\mathrm{reach}(\mathcal{M})$.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

$$t \in \left[4\mathrm{d_H}(X, \mathcal{M}), \mathrm{reach}(\mathcal{M}) - 3\mathrm{d_H}(X, \mathcal{M})\right).$$

Selection of the parameter t

Question 1: How to select a t such that $X^t \approx \mathcal{M}$?



Theorem (Frédéric Chazal, David Cohen-Steiner, and André Lieutier, 2009):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$. Suppose that $\mathcal{M}$ has positive reach, and that $\mathrm{d_H}(X, \mathcal{M}) \le \frac{1}{17}\mathrm{reach}(\mathcal{M})$.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

$$t \in [4\mathrm{d_H}(X, \mathcal{M}), \mathrm{reach}(\mathcal{M}) - 3\mathrm{d_H}(X, \mathcal{M})).$$

Theorem (Partha Niyogi, Stephen Smale, and Shmuel Weinberger, 2008):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$, with $\mathcal{M}$ a submanifold, and X a finite subset of $\mathcal{M}$.
Suppose that $\mathcal{M}$ has positive reach.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

$$t \in \left[2\mathrm{d_H}(X, \mathcal{M}), \sqrt{\frac{3}{5}}\mathrm{reach}(\mathcal{M})\right).$$

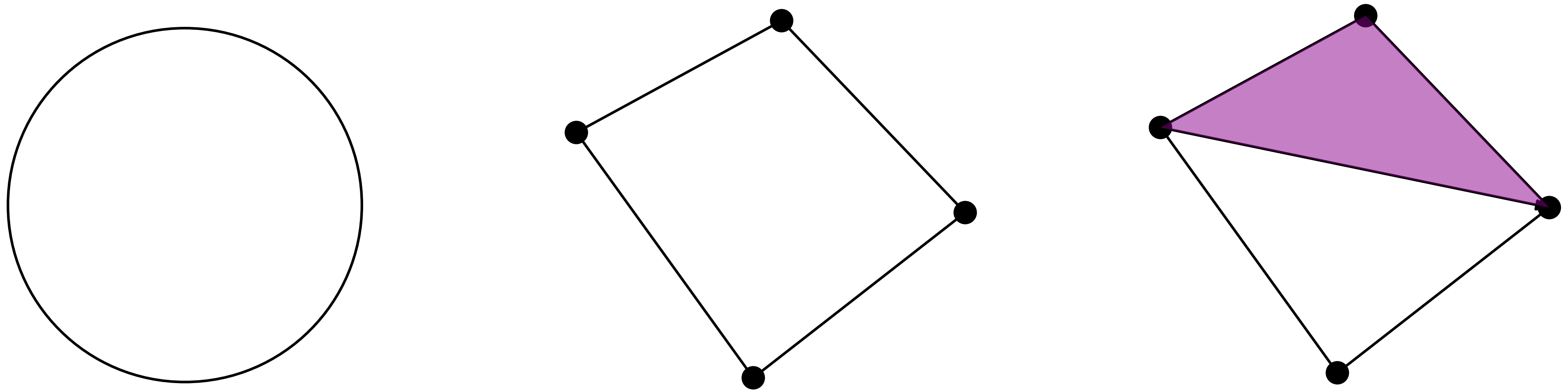
(Homotopy equivalent) triangulations

Question 2: How to compute the homology groups of X^t?

We need a triangulation of X^t, that is: a simplicial complex K homeomorphic to X^t.

Actually, we will define something weaker: a simplicial complex K that is homotopy equivalent to X^t.

Either case, we will have $\beta_i(X^t) = \beta_i(K)$ for all $i \geq 0$.

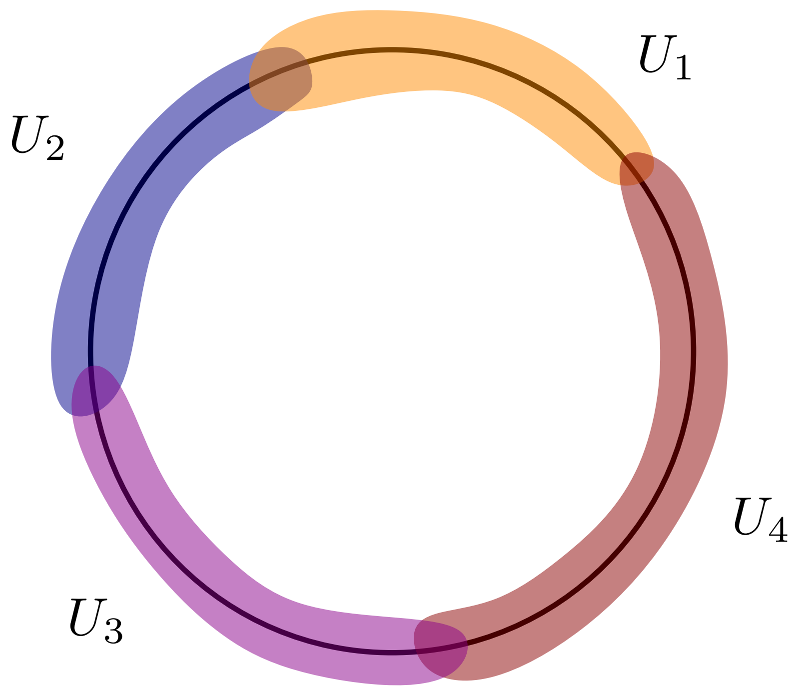


Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \leq i \leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \leq i \leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.

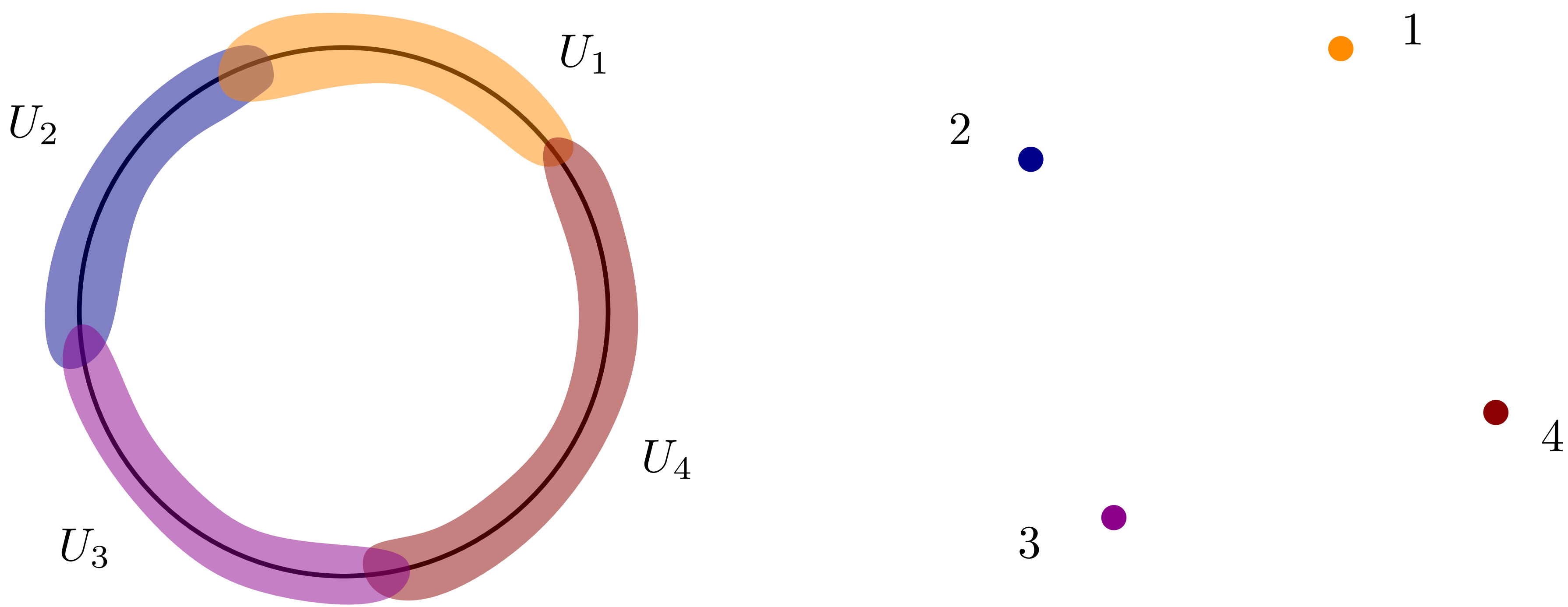


Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \leq i \leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \leq i \leq N} U_i = X.$$

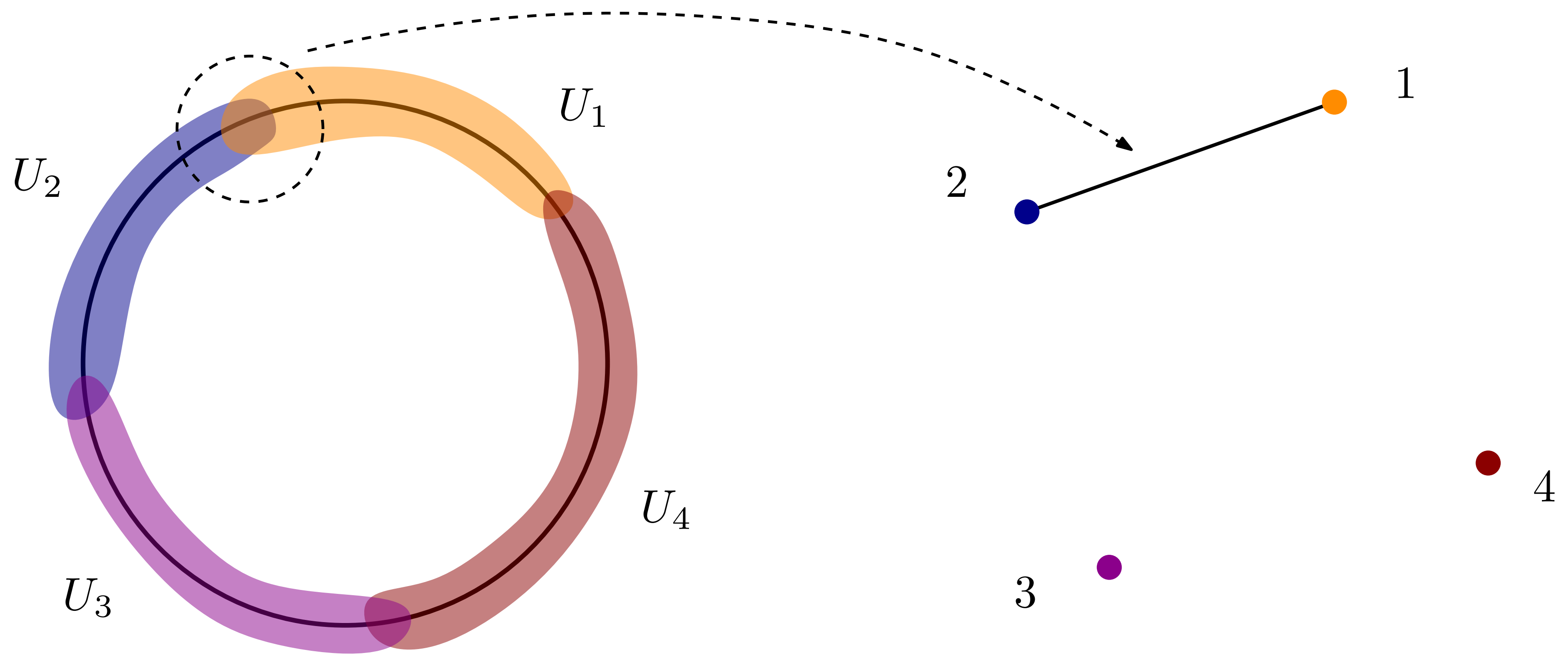
The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.



Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1\le i\le N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1\le i\le N} U_i = X.$$

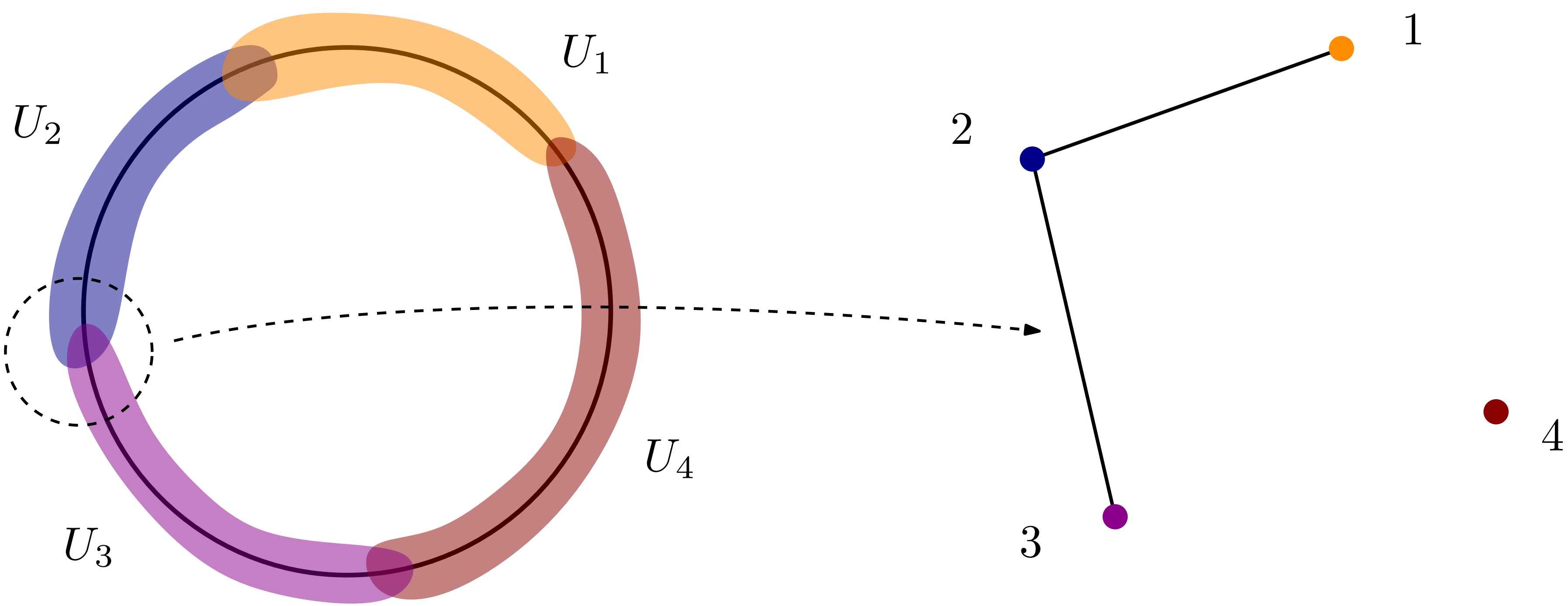
The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.



Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \le i \le N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \le i \le N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.

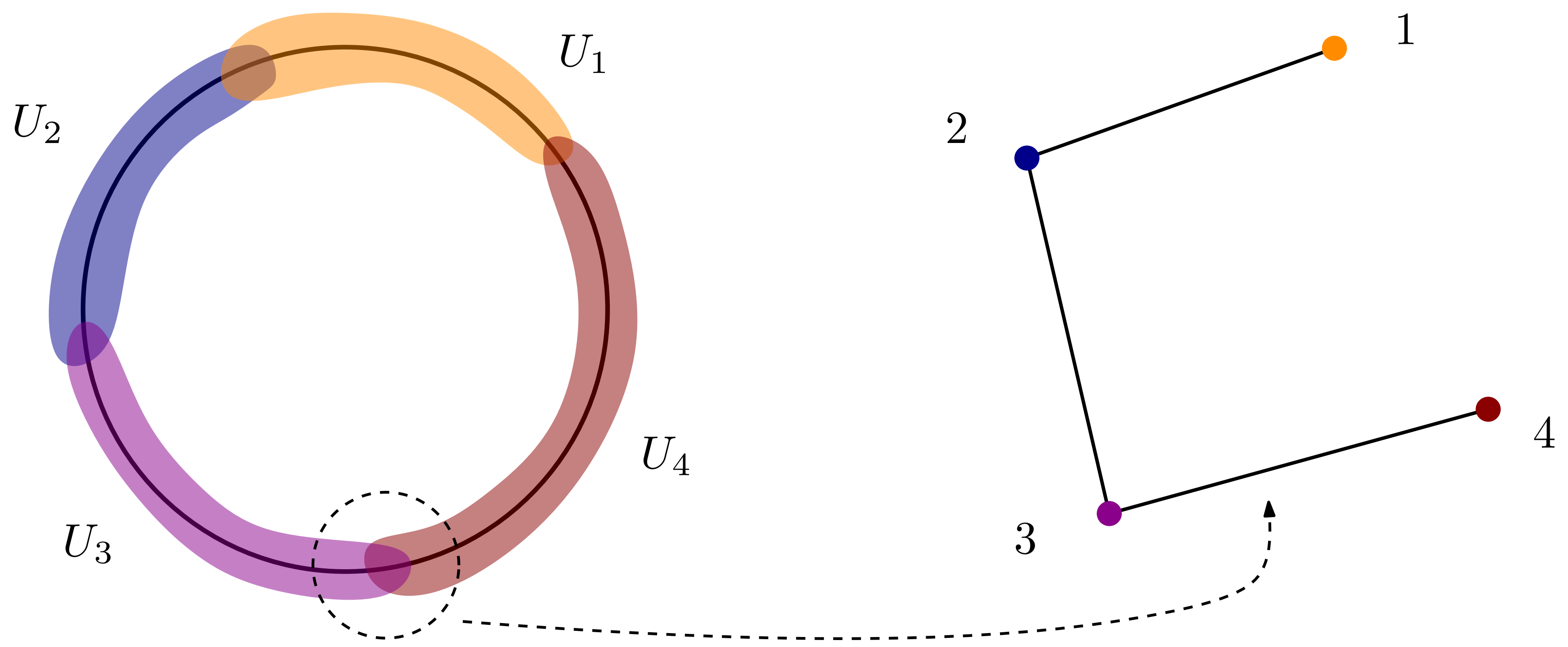


Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1\leq i\leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1\leq i\leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.

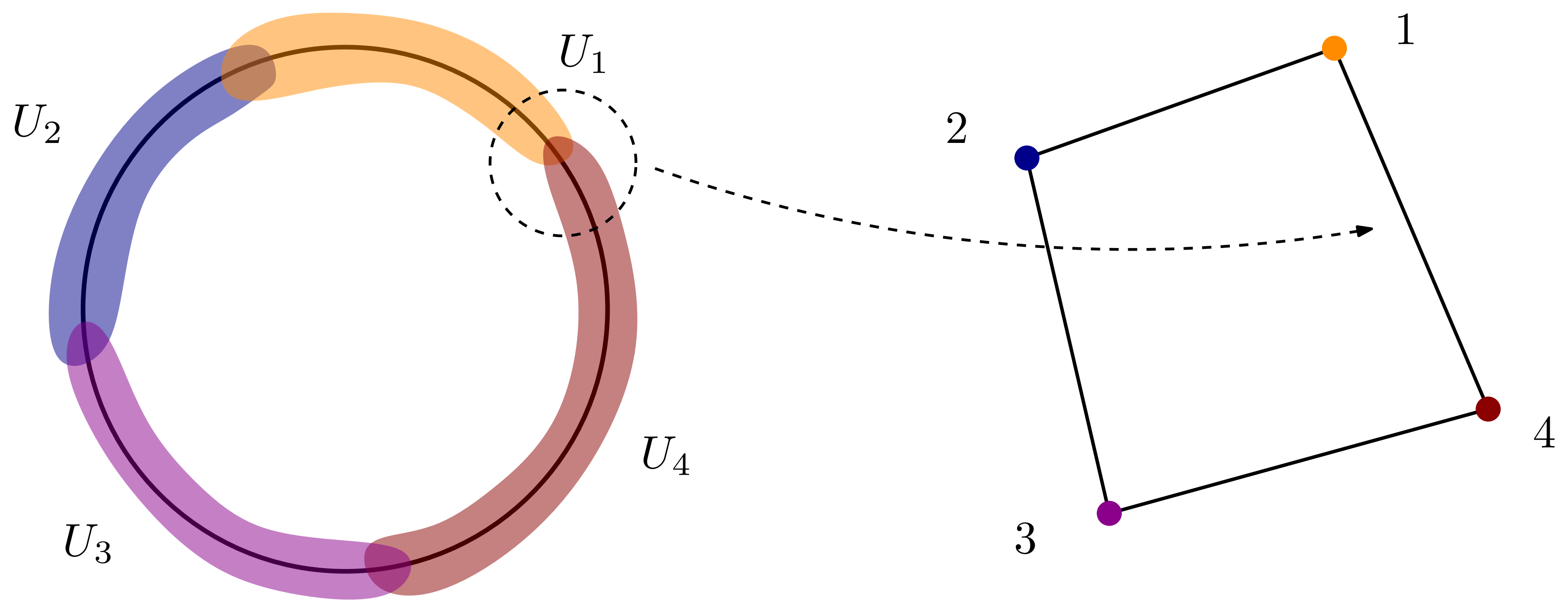


Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1\leq i\leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1\leq i\leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.

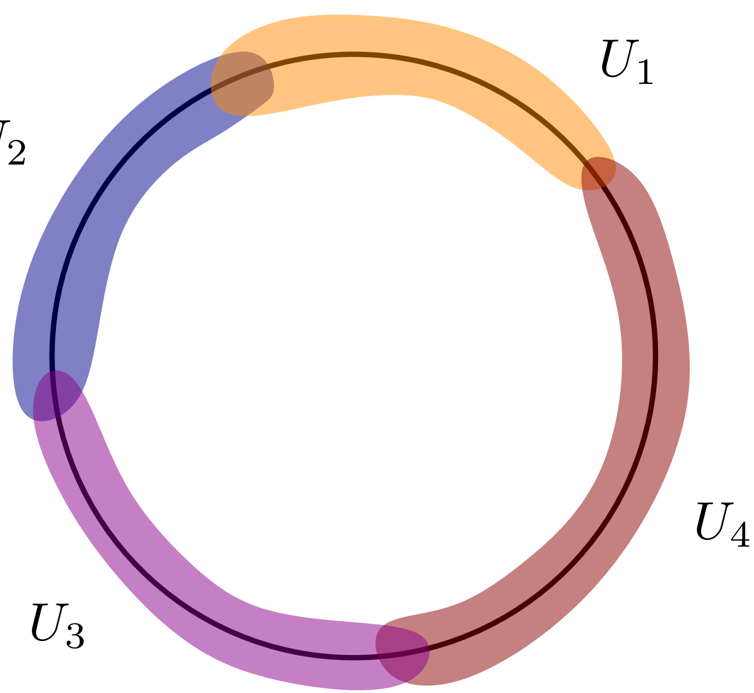


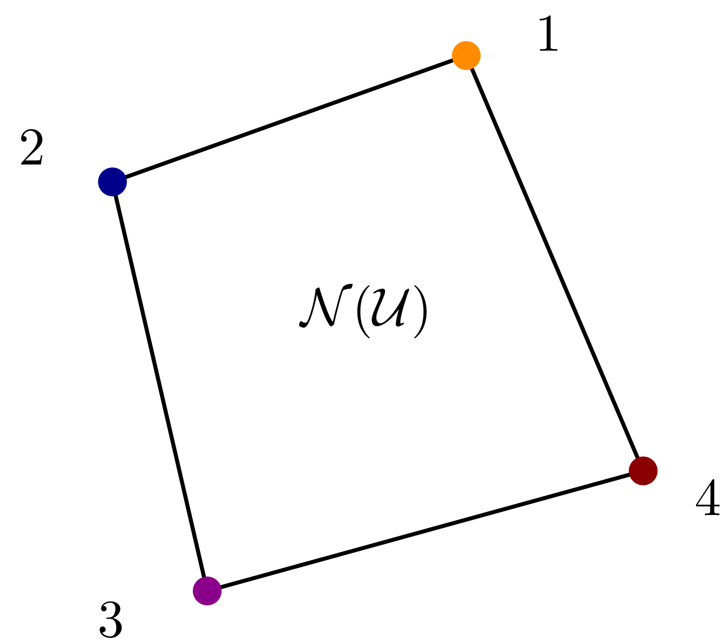
Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \leq i \leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \leq i \leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.



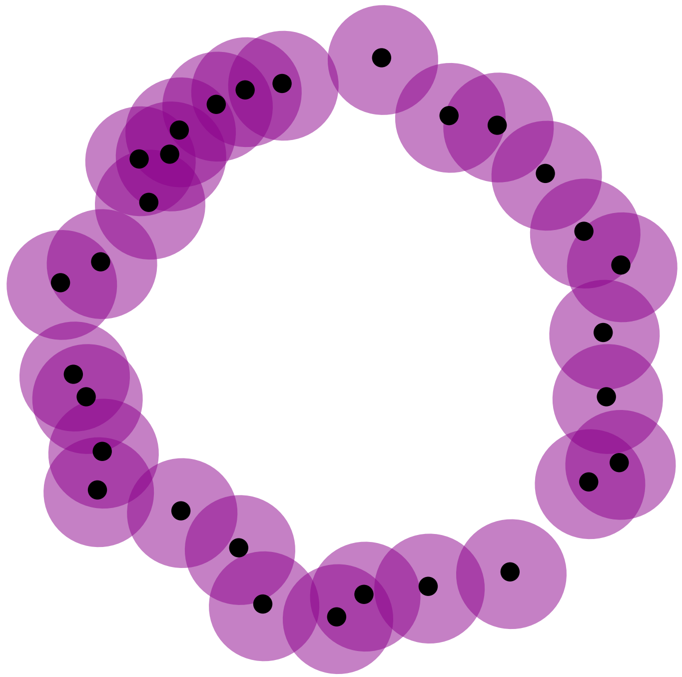


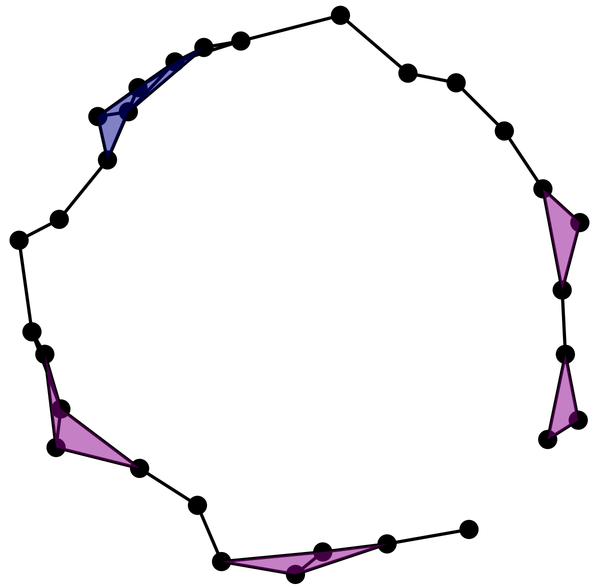
Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \leq i \leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \leq i \leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.



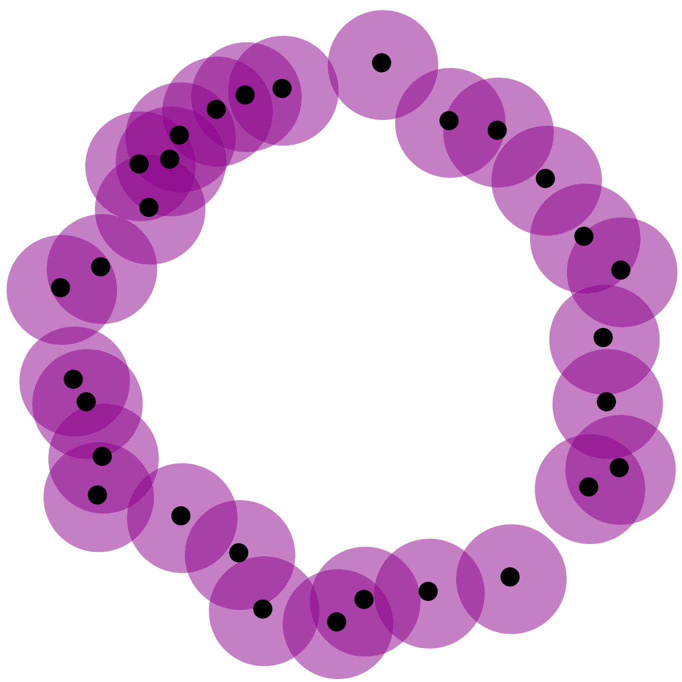


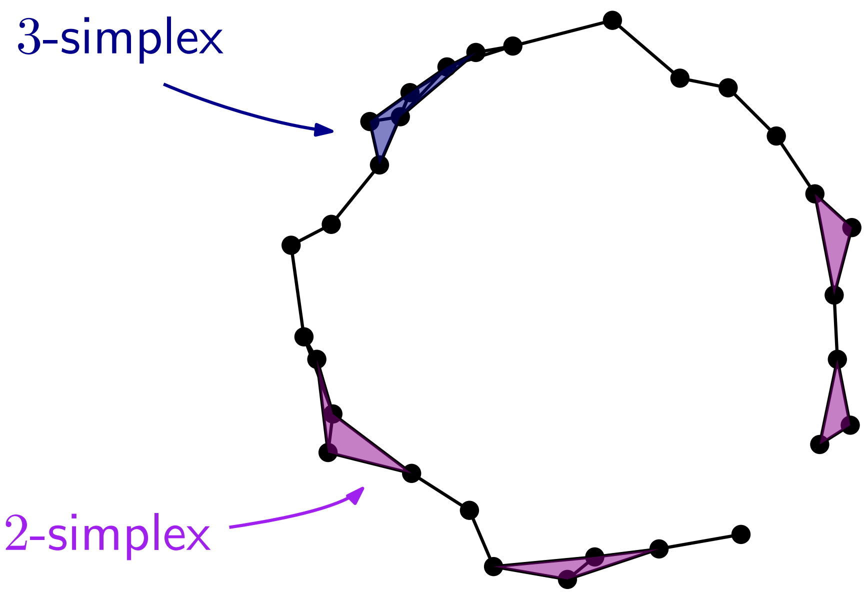
$X^{0.2} = \bigcup_{x \in X} \overline{\mathcal{B}}(x, 0.2)$ is covered by $\mathcal{U} = \{\overline{\mathcal{B}}(x, 0.2), x \in X\}$

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \leq i \leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \leq i \leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, \dots, N\}$ and whose m-simplices are the subsets $\{i_1, \dots, i_m\} \subset \{1, \dots, N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.





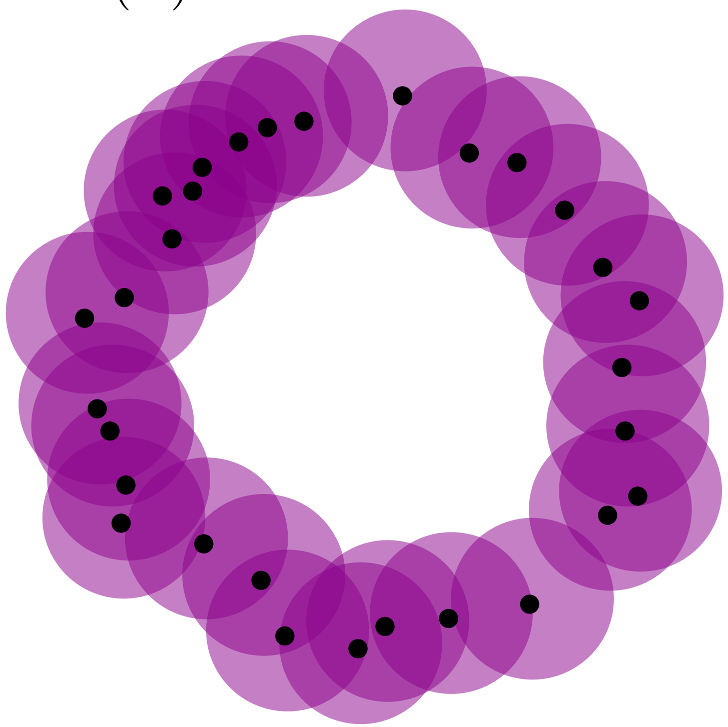
$X^{0.2} = \bigcup_{x \in X} \overline{\mathcal{B}}(x, 0.2)$ is covered by $\mathcal{U} = \{\overline{\mathcal{B}}(x, 0.2), x \in X\}$

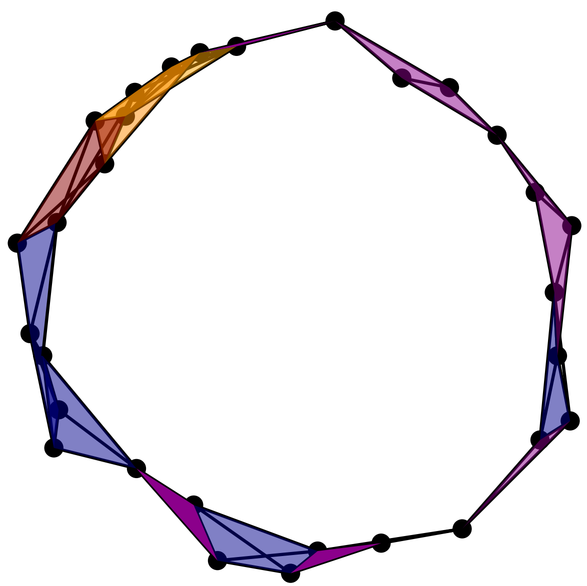
Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1\leq i\leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1\leq i\leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.





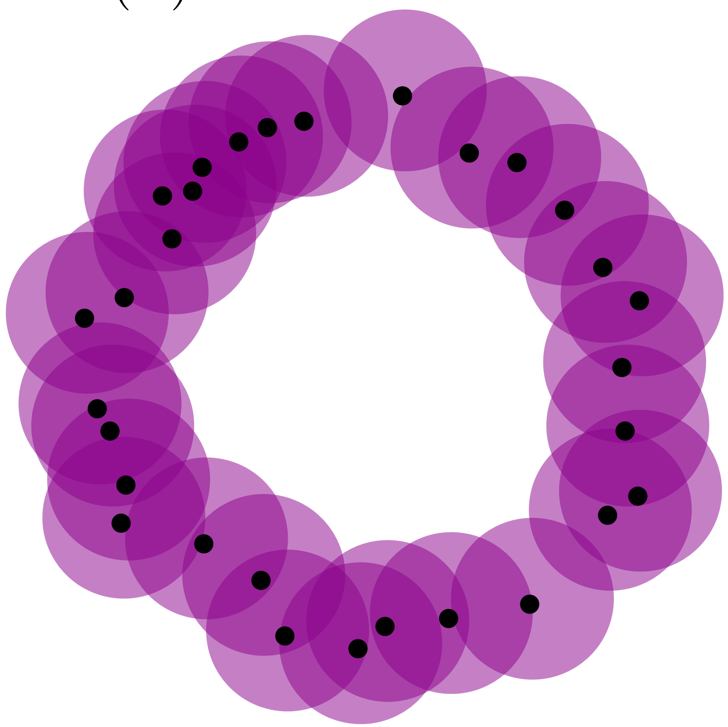
$X^{0.3} = \bigcup_{x\in X} \overline{\mathcal{B}}(x, 0.3)$ is covered by $\mathcal{U} = \{\overline{\mathcal{B}}(x, 0.3), x \in X\}$

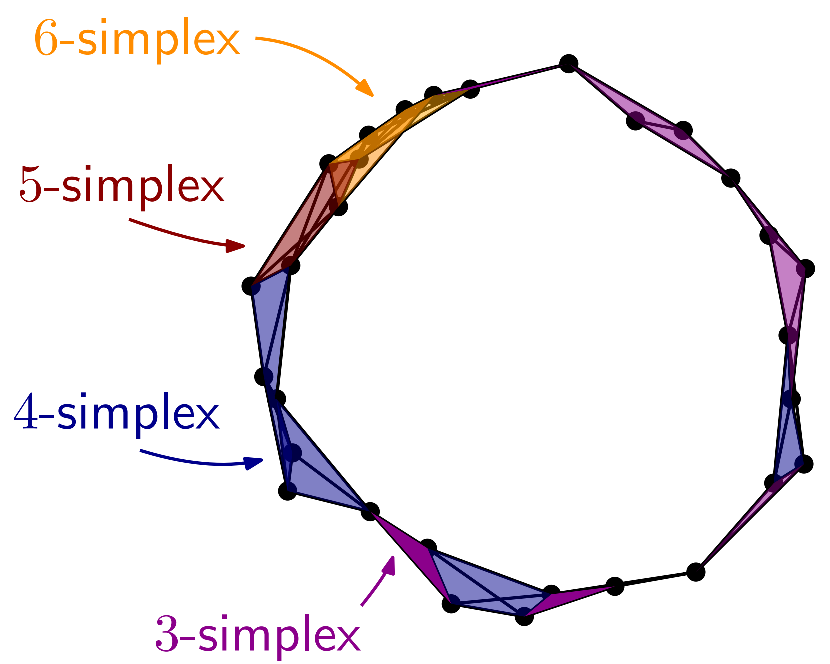
Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1\leq i\leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1\leq i\leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.





$X^{0.3} = \bigcup_{x\in X} \overline{\mathcal{B}}(x, 0.3)$ is covered by $\mathcal{U} = \left\{\overline{\mathcal{B}}(x, 0.3), x \in X\right\}$

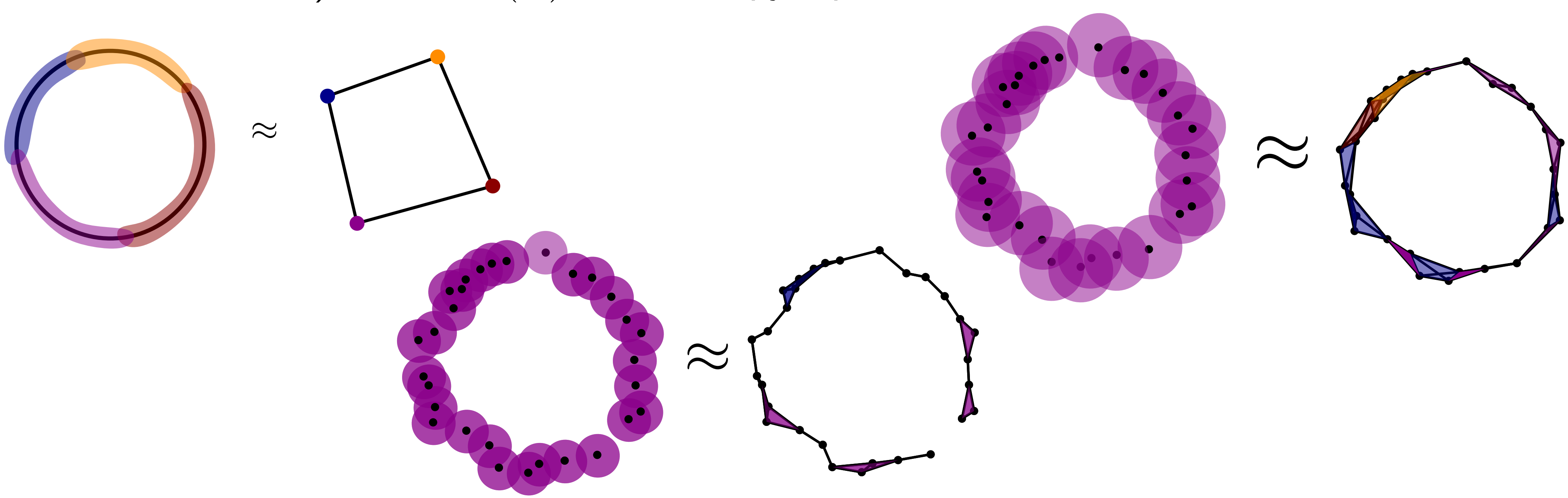
Nerves

Definition: Let X be a topological space, and $\mathcal{U} = \{U_i\}_{1 \leq i \leq N}$ a cover of X, that is, a collection of subsets $U_i \subset X$ such that

$$\bigcup_{1 \leq i \leq N} U_i = X.$$

The *nerve* of $\mathcal{U}$ is the simplicial complex with vertex set $\{1, ..., N\}$ and whose m-simplices are the subsets $\{i_1, ..., i_m\} \subset \{1, ..., N\}$ such that $\bigcap_{k=0}^{m} U_{i_k} \neq \emptyset$. It is denoted $\mathcal{N}(\mathcal{U})$.

Nerve theorem: Consider $X \subset \mathbb{R}^n$. Suppose that each U_i are balls (or more generally, closed and convex). Then $\mathcal{N}(\mathcal{U})$ is homotopy equivalent to X.



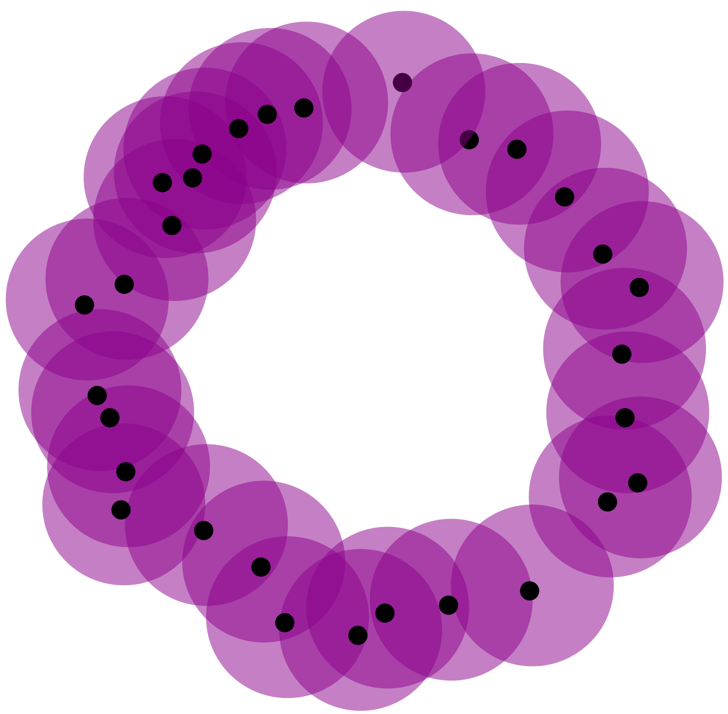
Čech complex

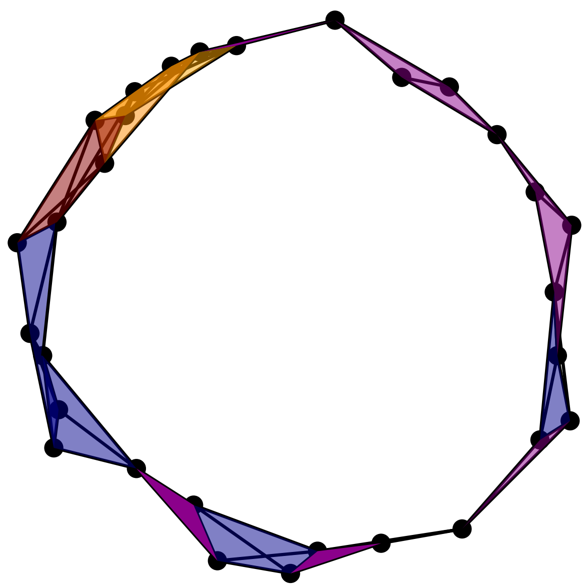
Let X be a finite subset of $\mathbb{R}^n$, and $t \geq 0$. Consider the collection

$$\mathcal{V}^t = \left\{\overline{\mathcal{B}}\left(x,t\right), x \in X\right\}.$$

This is a cover of the thickening X^t, and each components are closed balls.
By Nerve Theorem, its nerve $\mathcal{N}(\mathcal{V}^t)$ has the homotopy type of X^t.

Definition: This nerve is denoted $\check{\mathrm{C}}\mathrm{ech}^t(X)$ and is called the *Čech complex of X at time t*.





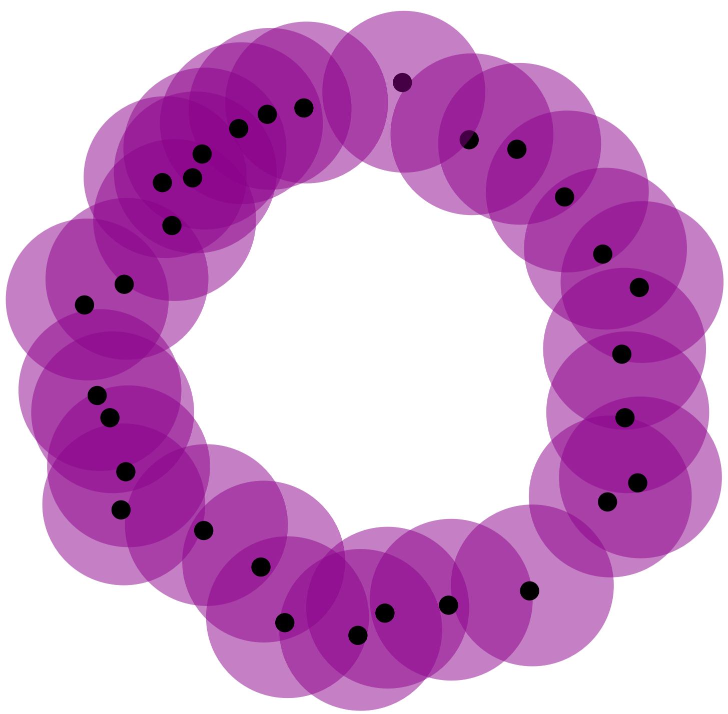
Čech complex

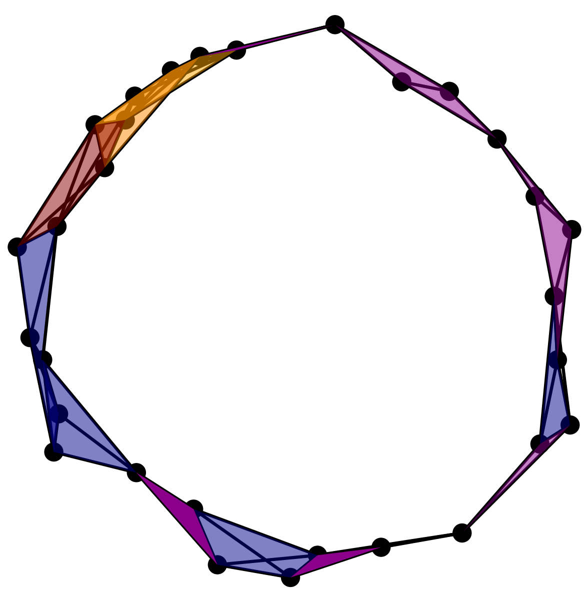
Let X be a finite subset of $\mathbb{R}^n$, and $t \geq 0$. Consider the collection

$$\mathcal{V}^t = \left\{ \overline{\mathcal{B}}\left(x, t\right), x \in X \right\}.$$

This is a cover of the thickening X^t, and each components are closed balls.
By Nerve Theorem, its nerve $\mathcal{N}(\mathcal{V}^t)$ has the homotopy type of X^t.

Definition: This nerve is denoted $\check{\mathrm{C}}\mathrm{ech}^t(X)$ and is called the *Čech complex of X at time t*.





$\longrightarrow$ The Question 2 (How to compute the homology groups of X^t?) is solved.

Problem of the scale

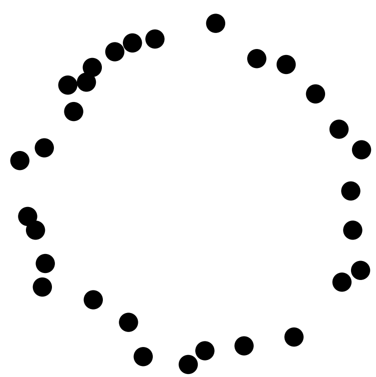
Question 1: How to select a t such that $X^t \approx \mathcal{M}$?

Theorem (Frédéric Chazal, David Cohen-Steiner, and André Lieutier, 2009):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$. Suppose that $\mathcal{M}$ has positive reach, and that $\mathrm{d_H}(X, \mathcal{M}) \leq \frac{1}{17}\mathrm{reach}(\mathcal{M})$.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

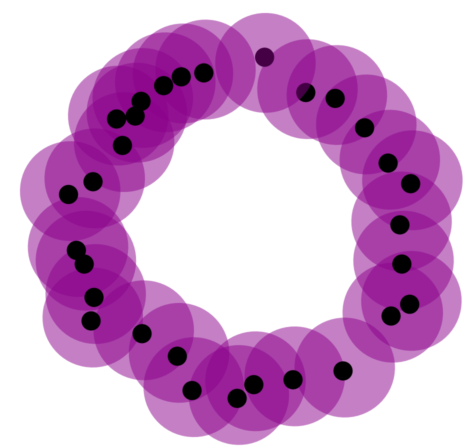
$$t \in [4\mathrm{d_H}(X,\mathcal{M}), \mathrm{reach}(\mathcal{M}) - 3\mathrm{d_H}(X,\mathcal{M})).$$

Theorem (Partha Niyogi, Stephen Smale, and Shmuel Weinberger, 2008):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$, with $\mathcal{M}$ a submanifold, and X a finite subset of $\mathcal{M}$. Suppose that $\mathcal{M}$ has positive reach.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

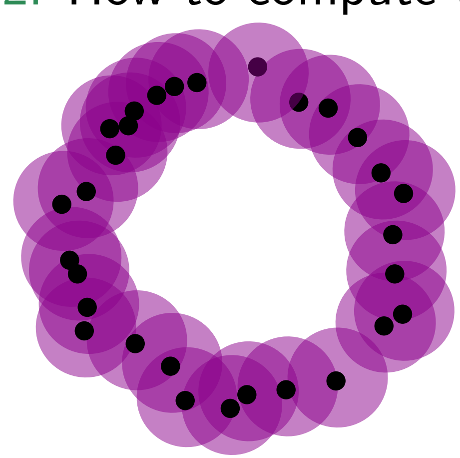
$$t \in \left[2\mathrm{d_H}(X,\mathcal{M}), \sqrt{\frac{3}{5}}\mathrm{reach}(\mathcal{M})\right).$$



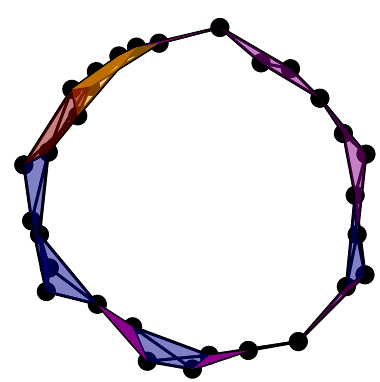
estimate t



Question 2: How to compute the homology groups of X^t?



compute the nerve



Problem of the scale

Question 1: How to select a t such that $X^t \approx \mathcal{M}$?

Theorem (Frédéric Chazal, David Cohen-Steiner, and André Lieutier, 2009):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$. Suppose that $\mathcal{M}$ has positive reach, and that $\mathrm{d_H}(X,\mathcal{M}) \leq \frac{1}{17}\mathrm{reach}(\mathcal{M})$.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

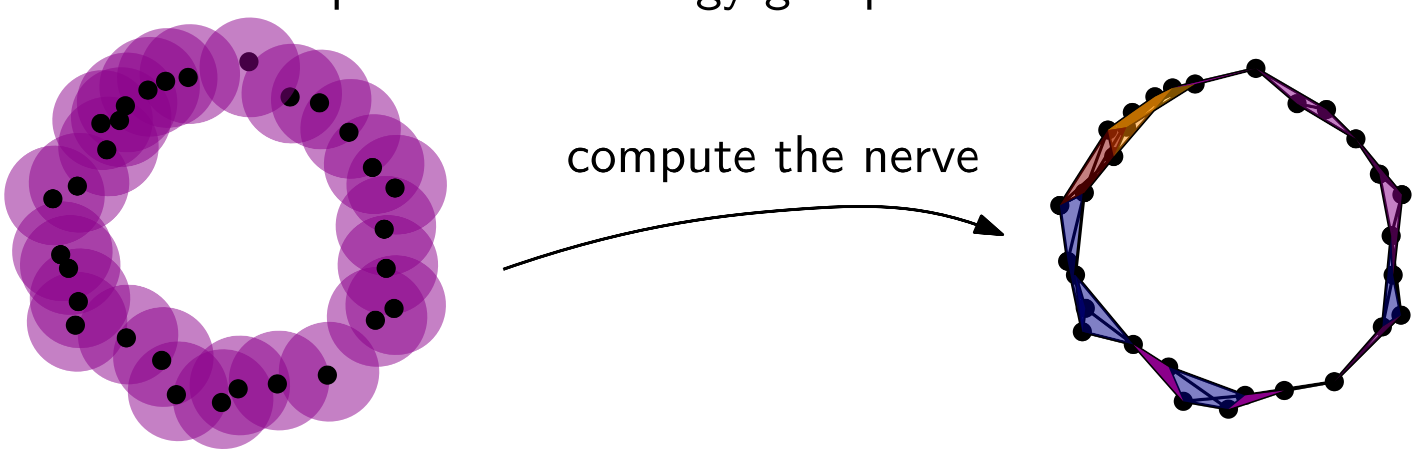
$$t \in \boxed{[4\mathrm{d_H}(X,\mathcal{M}), \mathrm{reach}(\mathcal{M}) - 3\mathrm{d_H}(X,\mathcal{M})).}$$

Theorem (Partha Niyogi, Stephen Smale, and Shmuel Weinberger, 2008):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$, with $\mathcal{M}$ a submanifold, and X a finite subset of $\mathcal{M}$. Suppose that $\mathcal{M}$ has positive reach.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

$$t \in \boxed{\left[2\mathrm{d_H}(X,\mathcal{M}), \sqrt{\frac{3}{5}}\mathrm{reach}(\mathcal{M})\right).}$$

these quantities are not known!

Question 2: How to compute the homology groups of X^t?



Problem of the scale

Question 1: How to select a t such that $X^t \approx \mathcal{M}$?

Theorem (Frédéric Chazal, David Cohen-Steiner, and André Lieutier, 2009):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$. Suppose that $\mathcal{M}$ has positive reach, and that $\mathrm{d_H}(X, \mathcal{M}) \leq \frac{1}{17}\mathrm{reach}(\mathcal{M})$.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

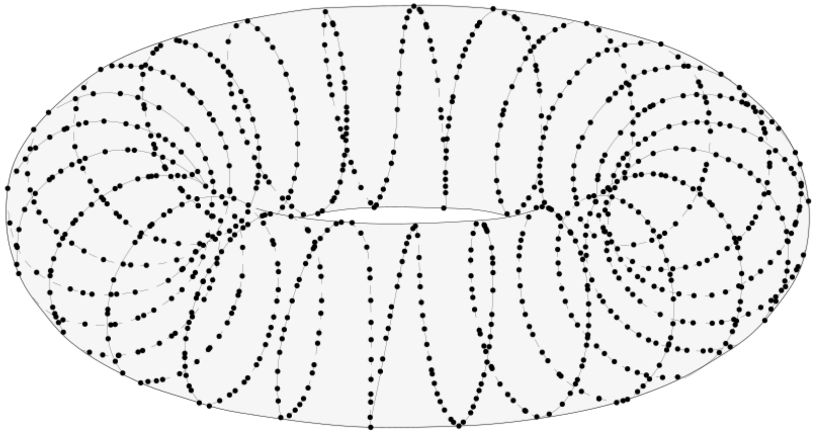
$$t \in \boxed{[4\mathrm{d_H}(X, \mathcal{M}), \mathrm{reach}(\mathcal{M}) - 3\mathrm{d_H}(X, \mathcal{M})).}$$

Theorem (Partha Niyogi, Stephen Smale, and Shmuel Weinberger, 2008):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$, with $\mathcal{M}$ a submanifold, and X a finite subset of $\mathcal{M}$. Suppose that $\mathcal{M}$ has positive reach.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

$$t \in \boxed{\left[2\mathrm{d_H}(X, \mathcal{M}), \sqrt{\frac{3}{5}}\mathrm{reach}(\mathcal{M})\right).}$$

these quantities are not known!

Is this object 1- or 2-dimensional?



Problem of the scale

Question 1: How to select a t such that $X^t \approx \mathcal{M}$?

Theorem (Frédéric Chazal, David Cohen-Steiner, and André Lieutier, 2009):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$. Suppose that $\mathcal{M}$ has positive reach, and that $\mathrm{d_H}(X, \mathcal{M}) \leq \frac{1}{17}\mathrm{reach}(\mathcal{M})$.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

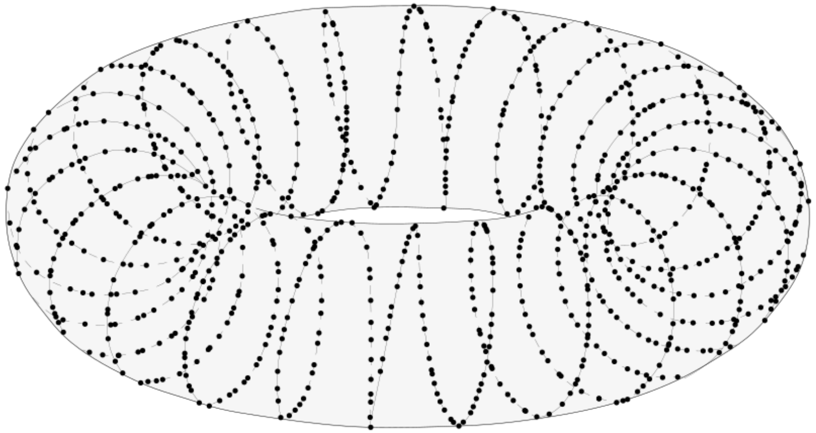
$$t \in \boxed{[4\mathrm{d_H}(X, \mathcal{M}), \mathrm{reach}(\mathcal{M}) - 3\mathrm{d_H}(X, \mathcal{M})).}$$

Theorem (Partha Niyogi, Stephen Smale, and Shmuel Weinberger, 2008):
Let X and $\mathcal{M}$ be subsets of $\mathbb{R}^n$, with $\mathcal{M}$ a submanifold, and X a finite subset of $\mathcal{M}$. Suppose that $\mathcal{M}$ has positive reach.
Then X^t and $\mathcal{M}$ are homotopic equivalent, provided that

$$t \in \boxed{\left[2\mathrm{d_H}(X, \mathcal{M}), \sqrt{\frac{3}{5}}\mathrm{reach}(\mathcal{M})\right).}$$

these quantities are not known!

Is this object 1- or 2-dimensional?

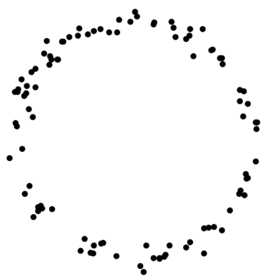
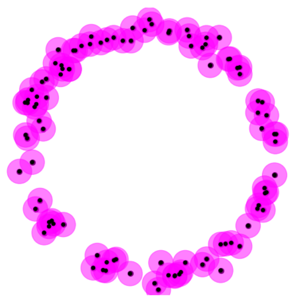
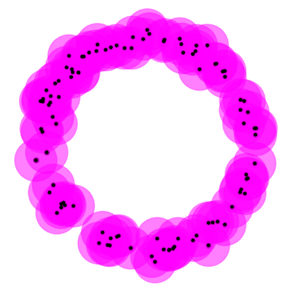
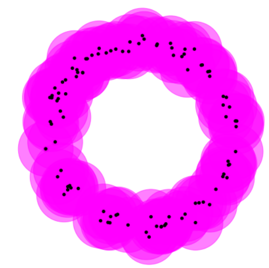
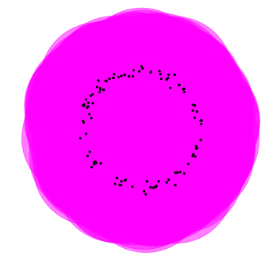


Idea (multiscale analysis): Instead of estimating a value of t, we will choose all of them.

Definition: The **Čech filtration** of X is the collection $V[X] = (X^t)_{t \geq 0}$.

Homology of the Čech filtration

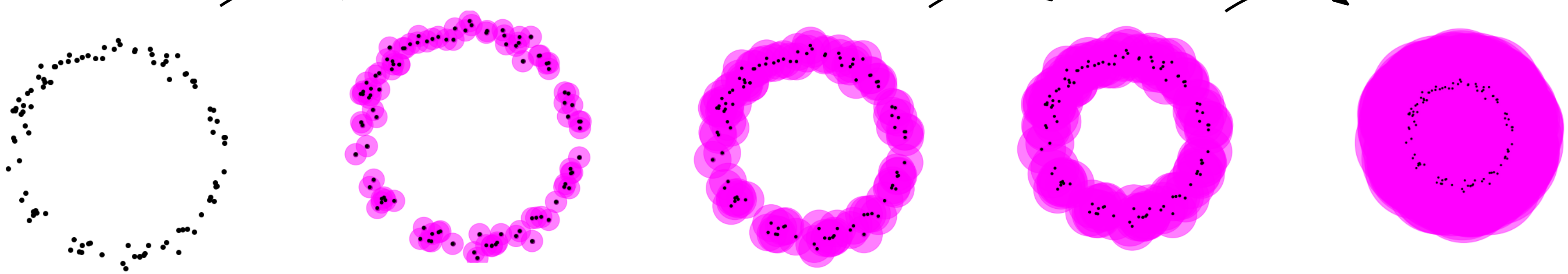
Let us compute the homology of each thickening:

X^t	 $X^0 = X$	 $X^{0.1}$	 $X^{0.2}$	 $X^{0.3}$	 X^1
$H_0(X^t)$	$(\mathbb{Z}/2\mathbb{Z})^{100}$	$(\mathbb{Z}/2\mathbb{Z})^5$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$
$H_1(X^t)$	0	0	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	0

Homology of the Čech filtration

Let us compute the homology of each thickening:

inclusions i_s^t: $X^0 \xrightarrow{i_0^{0.1}} X^{0.1} \xrightarrow{i_{0.1}^{0.2}} X^{0.2} \xrightarrow{i_{0.2}^{0.3}} X^{0.3} \xrightarrow{i_{0.3}^{1}} X^1$



X^t	$X^0 = X$	$X^{0.1}$	$X^{0.2}$	$X^{0.3}$	X^1
$H_0(X^t)$	$(\mathbb{Z}/2\mathbb{Z})^{100}$	$(\mathbb{Z}/2\mathbb{Z})^{5}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$
$H_1(X^t)$	0	0	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/2\mathbb{Z}$	0

$$0 \xrightarrow{(i_0^{0.1})_*} 0 \xrightarrow{(i_{0.1}^{0.2})_*} \mathbb{Z}/2\mathbb{Z} \xrightarrow{(i_{0.2}^{0.3})_*} \mathbb{Z}/2\mathbb{Z} \xrightarrow{(i_{0.3}^{1})_*} 0$$

The data of $(H_i(X^t))_{t\geq 0}$ and $((i_s^t)_*)_{s\leq t}$ is called a **persistence module**.

Persistence modules

Definition: A **persistence module $\mathbb{V}$ over $\mathbb{R}^+$ with coefficients in $\mathbb{Z}/2\mathbb{Z}$** is a pair $(\mathbb{V}, v)$ where $\mathbb{V} = (V^t)_{t\in\mathbb{R}^+}$ is a family of $\mathbb{Z}/2\mathbb{Z}$-vector spaces, and $v = (v_s^t \colon V^s \to V^t)_{s\leq t\in\mathbb{R}^+}$ a family of linear maps such that:

- for every $t \in \mathbb{R}^+$, $v_t^t \colon V^t \to V^t$ is the identity map,
- for every $r, s, t \in \mathbb{R}^+$ such that $r \leq s \leq t$, we have $v_s^t \circ v_r^s = v_r^t$.

$$\begin{tikzcd} V^r \arrow[r, "v_r^s"] \arrow[rr, bend right, "v_r^t"'] & V^s \arrow[r, "v_s^t"] & V^t \end{tikzcd}$$

Persistence module associated to the Čech filtration:

$$\begin{tikzcd} \cdots \arrow[r, dashed] & X^{t_1} \arrow[r, hook, "i_{t_1}^{t_2}"] & X^{t_2} \arrow[r, hook, "i_{t_2}^{t_3}"] & X^{t_3} \arrow[r, hook, "i_{t_3}^{t_4}"] & X^{t_4} \arrow[r, hook, dashed] & \cdots \\ \cdots \arrow[r, dashed] & H_i(X^{t_1}) \arrow[r, "(i_{t_1}^{t_2})_*"] & H_i(X^{t_2}) \arrow[r, "(i_{t_2}^{t_3})_*"] & H_i(X^{t_3}) \arrow[r, "(i_{t_3}^{t_4})_*"] & H_i(X^{t_4}) \arrow[r, dash, dashed] & \cdots \end{tikzcd}$$

Tracking cycles over time

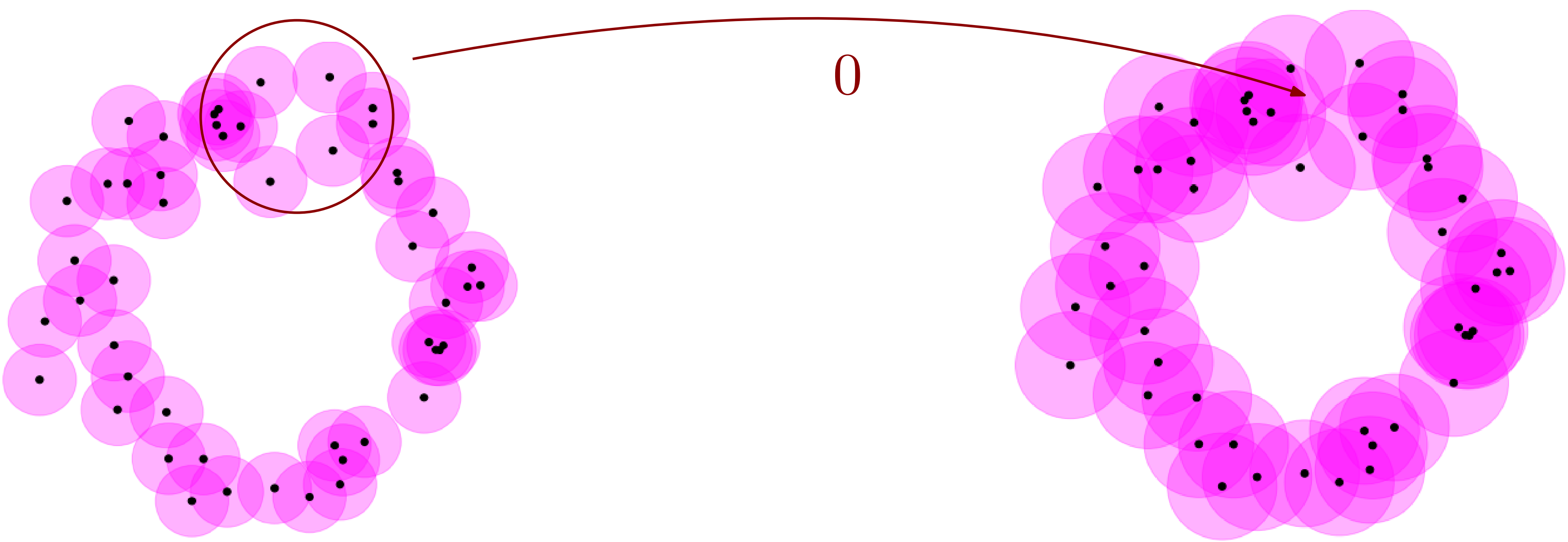
$$\dashrightarrow H_i(X^{t_1}) \xrightarrow{\left(i_{t_1}^{t_2}\right)_*} H_i(X^{t_2}) \xrightarrow{\left(i_{t_2}^{t_3}\right)_*} H_i(X^{t_3}) \xrightarrow{\left(i_{t_2}^{t_3}\right)_*} H_i(X^{t_4}) \dashrightarrow$$

Let $i \geq 0$, $t_0 \geq 0$ and consider a cycle $c \in H_i(X^{t_0})$.

Its **death time** is: $\sup\left\{t \geq t_0, \left(i_{t_0}^{t}\right)_* (c) \neq 0\right\}$,

its **birth time** is: $\inf\left\{t \geq t_0, \left(i_{t}^{t_0}\right)^{-1} (\{c\})_* \neq \emptyset\right\}$,

its **persistence** is the difference.



Tracking cycles over time

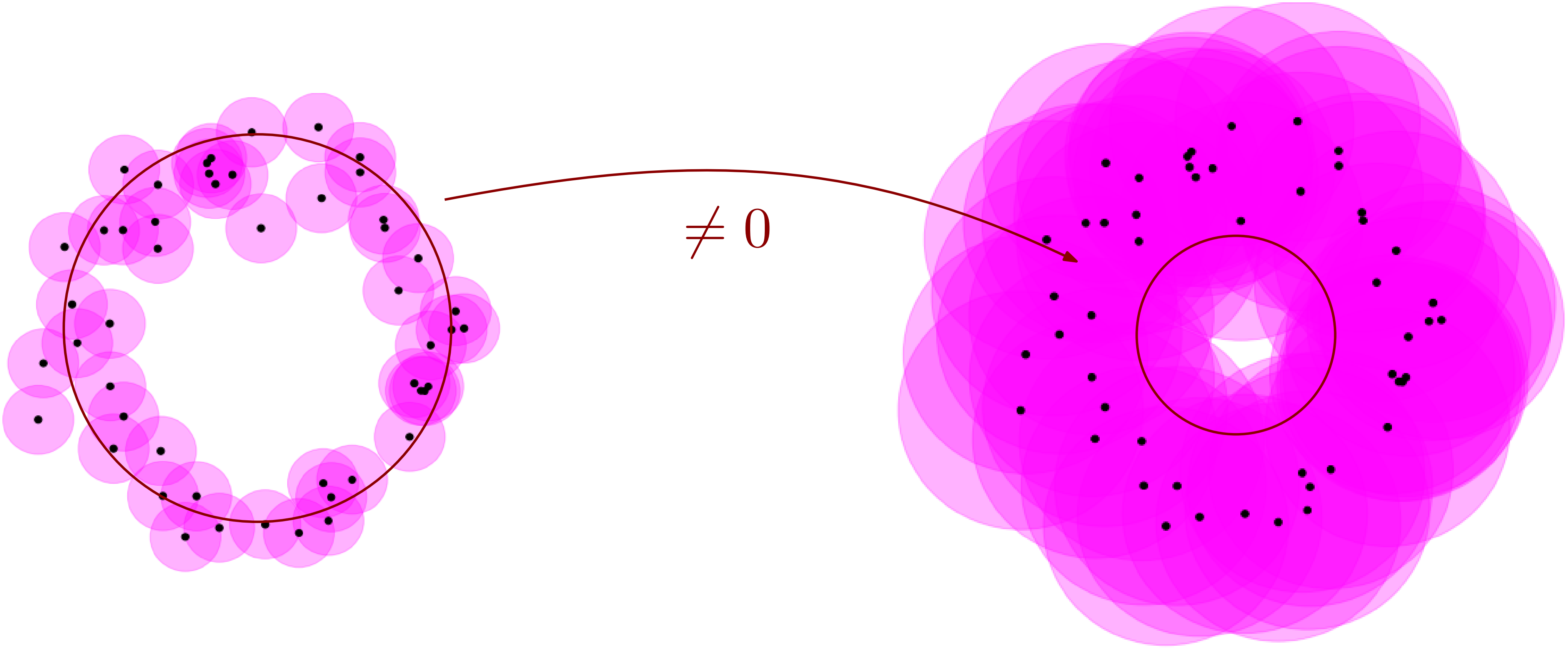
$$\dashrightarrow H_i(X^{t_1}) \xrightarrow{\left(i_{t_1}^{t_2}\right)_*} H_i(X^{t_2}) \xrightarrow{\left(i_{t_2}^{t_3}\right)_*} H_i(X^{t_3}) \xrightarrow{\left(i_{t_2}^{t_3}\right)_*} H_i(X^{t_4}) \dashrightarrow$$

Let $i \geq 0$, $t_0 \geq 0$ and consider a cycle $c \in H_i(X^{t_0})$.

Its **death time** is: $\sup\left\{t \geq t_0, \left(i_{t_0}^{t}\right)_* (c) \neq 0\right\}$,

its **birth time** is: $\inf\left\{t \geq t_0, \left(i_t^{t_0}\right)^{-1} (\{c\})_* \neq \emptyset\right\}$,

its **persistence** is the difference.



Tracking cycles over time

$$\dashrightarrow H_i(X^{t_1}) \xrightarrow{\left(i_{t_1}^{t_2}\right)_*} H_i(X^{t_2}) \xrightarrow{\left(i_{t_2}^{t_3}\right)_*} H_i(X^{t_3}) \xrightarrow{\left(i_{t_2}^{t_3}\right)_*} H_i(X^{t_4}) \dashrightarrow$$

Let $i \geq 0$, $t_0 \geq 0$ and consider a cycle $c \in H_i(X^{t_0})$.

Its **death time** is: $\sup \left\{t \geq t_0, \left(i_{t_0}^{t}\right)_* (c) \neq 0\right\}$,

its **birth time** is: $\inf \left\{t \geq t_0, \left(i_{t}^{t_0}\right)^{-1} (\{c\})_* \neq \emptyset\right\}$,

its **persistence** is the difference.

As a rule of thumb:

- cycles with large persistence correspond to important topological features of the dataset,
- cycles with short persistence corresponds to topological noise.

Persistence barcode

Theorem (Crawley-Boevey, 2015):

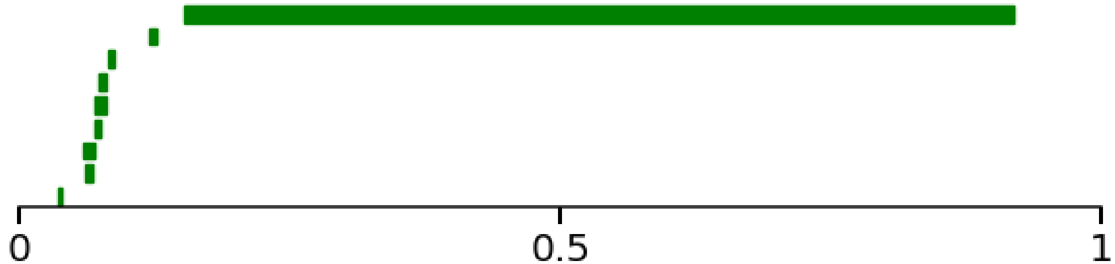
A (regular) persistence module can be decomposed as a sum of interval-modules.

This multiset of intervals is called **barcode**. It is a complete invariant of persistence modules.

Persistence module: $\mathbb{V}$

Barcode: $\{$ [0.171, 0.897), [0.035, 0.049), [0.037, 0.046), [0.072, 0.078), [0.077, 0.083), [0.046, 0.050), [0.050, 0.054), [0.036, 0.040), [0.089, 0.092) $\}$

Graphical representation:



Persistence barcode

Barcodes of the persistence module associated to the Čech filtration: H_0 in red and H_1 in green.

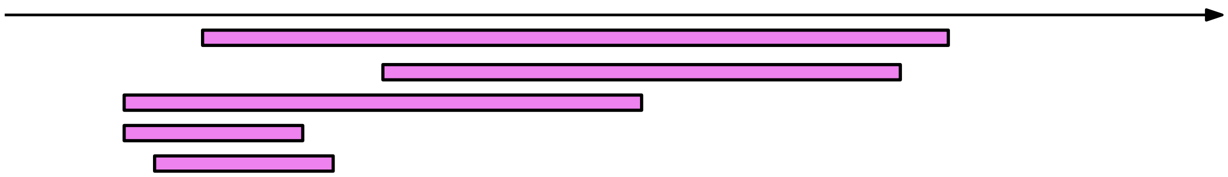
Persistence barcode

Barcodes of the persistence module associated to the Čech filtration: H_0 in red and H_1 in green.

On a barcode, on reads homology **at each step**, and sees how it **evolves**.

Persistence diagrams

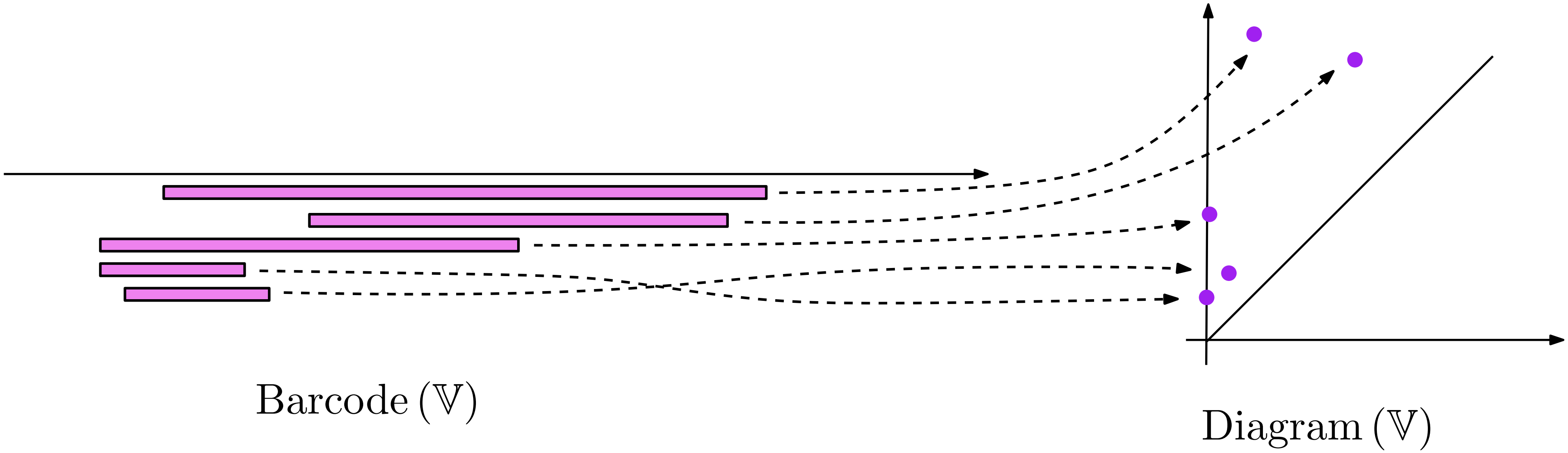
Associated to a persistence module $\mathbb{V}$ is its persistence barcode.



$\text{Barcode}\,(\mathbb{V})$

Persistence diagrams

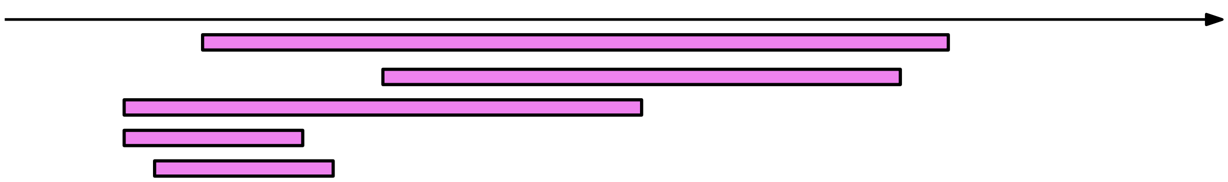
Associated to a persistence module $\mathbb{V}$ is its persistence barcode.



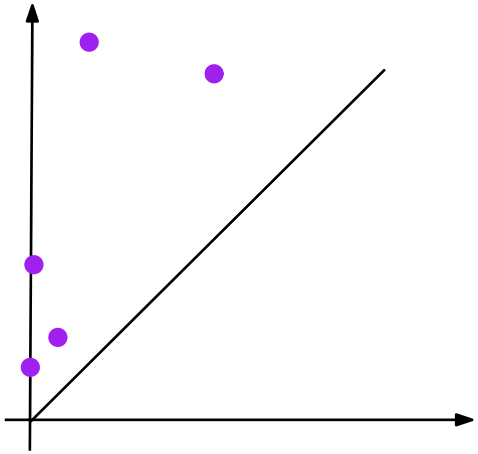
For every $[a, b]$, $(a, b]$, $[a, b)$ or (a, b) in $\mathrm{Barcode}\,(\mathbb{V})$, consider the point (a, b) of $\mathbb{R}^2$.
The collection of all such points is the **persistence diagram** of $\mathbb{V}$.

Persistence diagrams

Associated to a persistence module $\mathbb{V}$ is its persistence barcode.



$\mathrm{Barcode}\,(\mathbb{V})$



$\mathrm{Diagram}\,(\mathbb{V})$

For every $[a,b]$, $(a,b]$, $[a,b)$ or (a,b) in $\mathrm{Barcode}\,(\mathbb{V})$, consider the point (a,b) of $\mathbb{R}^2$.
The collection of all such points is the **persistence diagram** of $\mathbb{V}$.

I - Simplicial homology

1 - Homology groups

2 - Functoriality

II - Topological inference

1 - Parameter estimation

2 - Nerves

III - Persistent homology

1 - Persistence modules

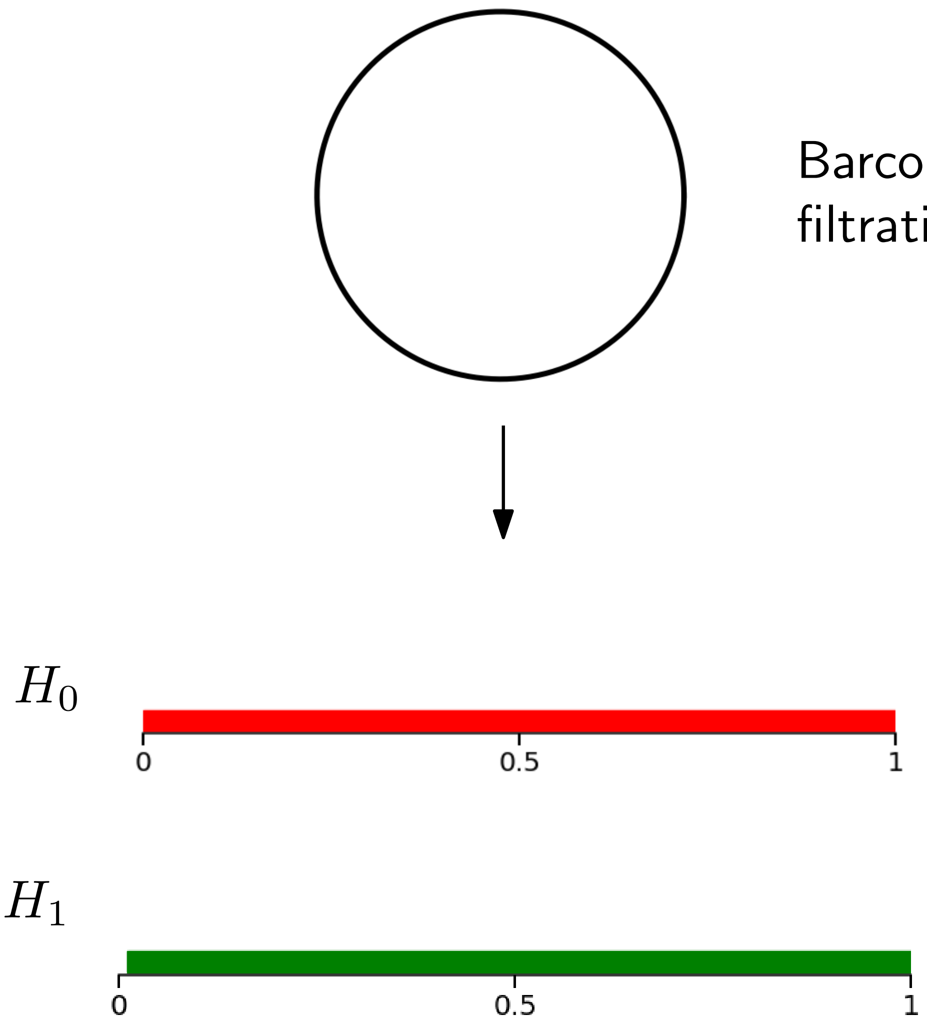
2 - Decomposition

3 - Stability

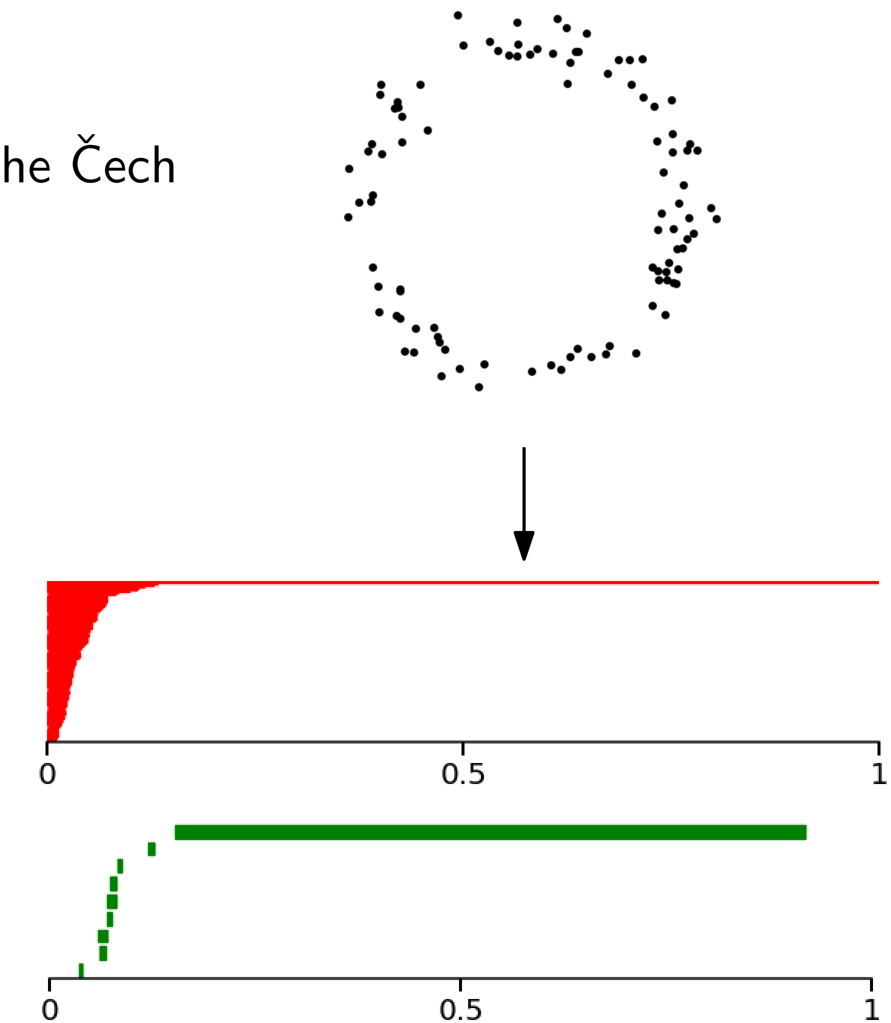
IV - Applications

Stability of persistence barcodes

Let $X \subset \mathbb{R}^n$ finite, seen as a sample of $\mathcal{M}$.

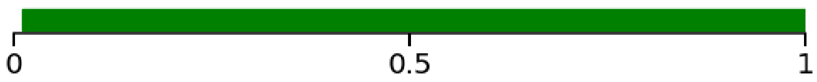


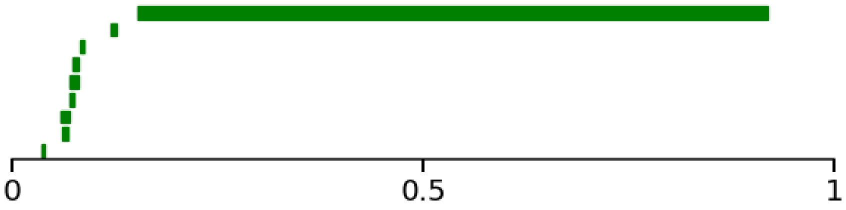
Barcodes of the Čech filtration



Bottleneck distance

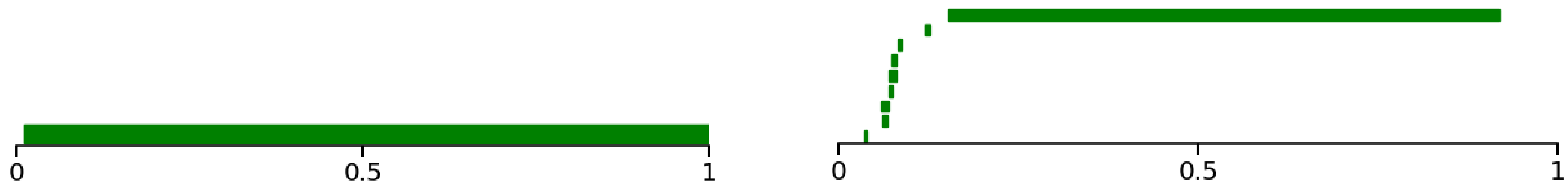
Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.





Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



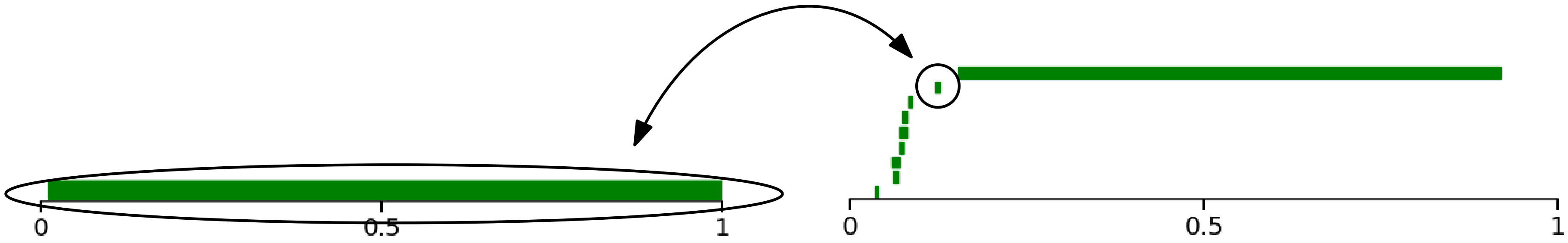
A **partial matching** between the barcodes is a subset $M \subset P \times Q$ such that

- for every $p \in P$, there exists at most one $q \in Q$ such that $(p, q) \in M$,
- for every $q \in Q$, there exists at most one $p \in P$ such that $(p, q) \in M$.

The points $p \in P$ (resp. $q \in Q$) such that there exists $q \in Q$ (resp. $p \in P$) with $(p, q) \in M$ are said **matched** by M.

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



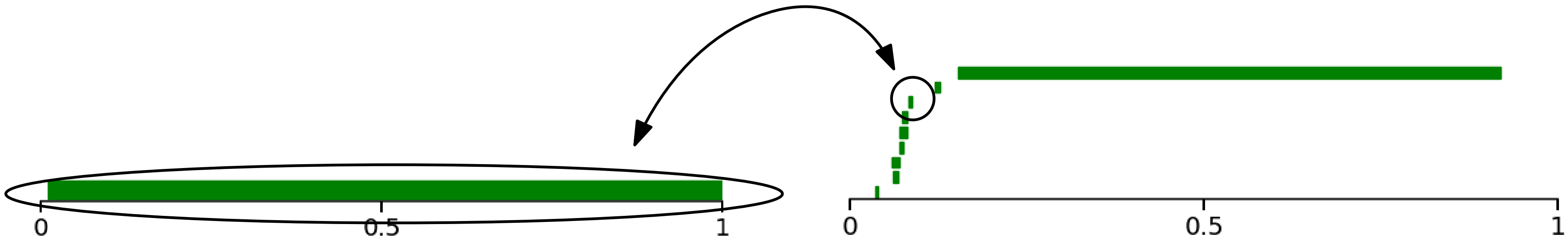
A **partial matching** between the barcodes is a subset $M \subset P \times Q$ such that

- for every $p \in P$, there exists at most one $q \in Q$ such that $(p, q) \in M$,
- for every $q \in Q$, there exists at most one $p \in P$ such that $(p, q) \in M$.

The points $p \in P$ (resp. $q \in Q$) such that there exists $q \in Q$ (resp. $p \in P$) with $(p, q) \in M$ are said **matched** by M.

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \le b_i$ for all $i \in \mathcal{I}$.



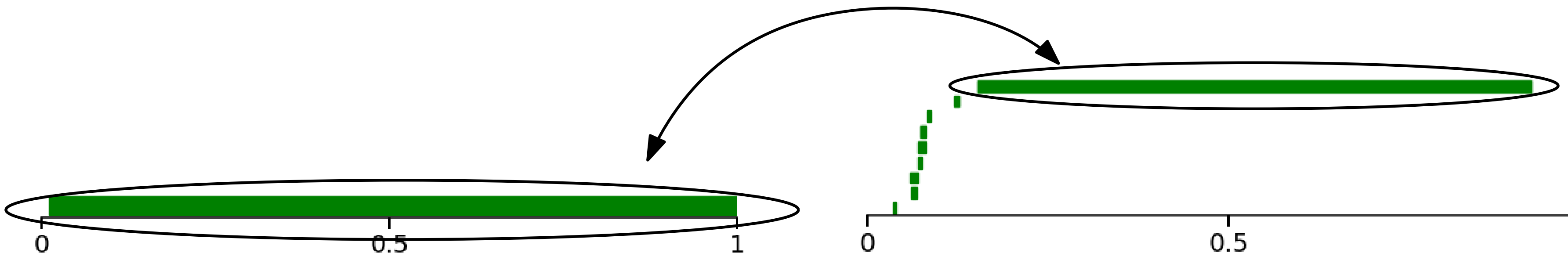
A **partial matching** between the barcodes is a subset $M \subset P \times Q$ such that

- for every $p \in P$, there exists at most one $q \in Q$ such that $(p, q) \in M$,
- for every $q \in Q$, there exists at most one $p \in P$ such that $(p, q) \in M$.

The points $p \in P$ (resp. $q \in Q$) such that there exists $q \in Q$ (resp. $p \in P$) with $(p, q) \in M$ are said **matched** by M.

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



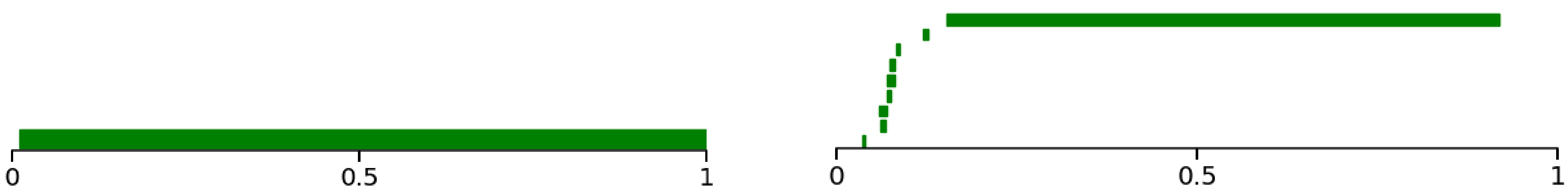
A **partial matching** between the barcodes is a subset $M \subset P \times Q$ such that

- for every $p \in P$, there exists at most one $q \in Q$ such that $(p, q) \in M$,
- for every $q \in Q$, there exists at most one $p \in P$ such that $(p, q) \in M$.

The points $p \in P$ (resp. $q \in Q$) such that there exists $q \in Q$ (resp. $p \in P$) with $(p, q) \in M$ are said **matched** by M.

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



A **partial matching** between the barcodes is a subset $M \subset P \times Q$ such that

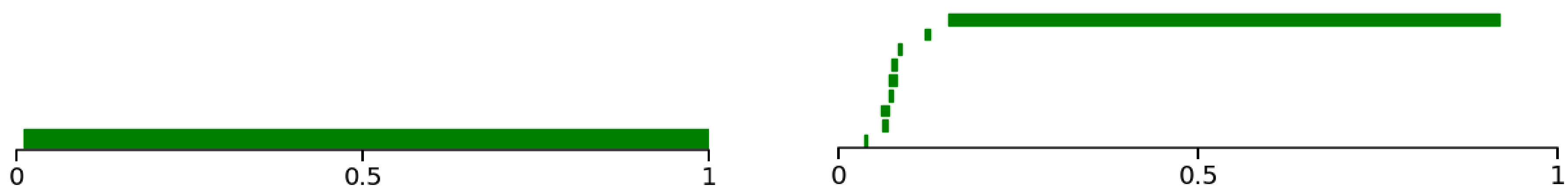
- for every $p \in P$, there exists at most one $q \in Q$ such that $(p, q) \in M$,
- for every $q \in Q$, there exists at most one $p \in P$ such that $(p, q) \in M$.

The points $p \in P$ (resp. $q \in Q$) such that there exists $q \in Q$ (resp. $p \in P$) with $(p, q) \in M$ are said **matched** by M.

If a point $p \in P$ (resp. $q \in Q$) is not matched by M, we consider that it is matched with the singleton $\overline{p} = \left[\frac{p_1+p_2}{2}, \frac{p_1+p_2}{2}\right]$ (resp. $\overline{q} = \left[\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}\right]$).

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



A **partial matching** between the barcodes is a subset $M \subset P \times Q$ such that

- for every $p \in P$, there exists at most one $q \in Q$ such that $(p, q) \in M$,
- for every $q \in Q$, there exists at most one $p \in P$ such that $(p, q) \in M$.

The points $p \in P$ (resp. $q \in Q$) such that there exists $q \in Q$ (resp. $p \in P$) with $(p, q) \in M$ are said **matched** by M.

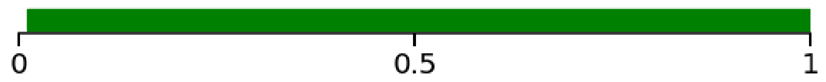
If a point $p \in P$ (resp. $q \in Q$) is not matched by M, we consider that it is matched with the singleton $\overline{p} = \left[\frac{p_1+p_2}{2}, \frac{p_1+p_2}{2}\right]$ (resp. $\overline{q} = \left[\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}\right]$).

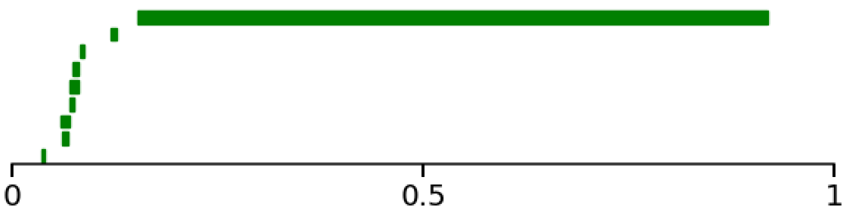
The **cost** of a matched pair (p, q) (resp. $(p, \overline{p})$, resp. $(\overline{q}, q)$) is the sup norm $\|p - q\|_\infty = \sup\{|p_1 - q_1|, |p_2 - q_2|\}$ (resp. $\|p - \overline{p}\|_\infty$, resp. $\|\overline{q} - q\|_\infty$).

The **cost** of the partial matching M, denoted $\mathrm{cost}(M)$, is the supremum of all costs.

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



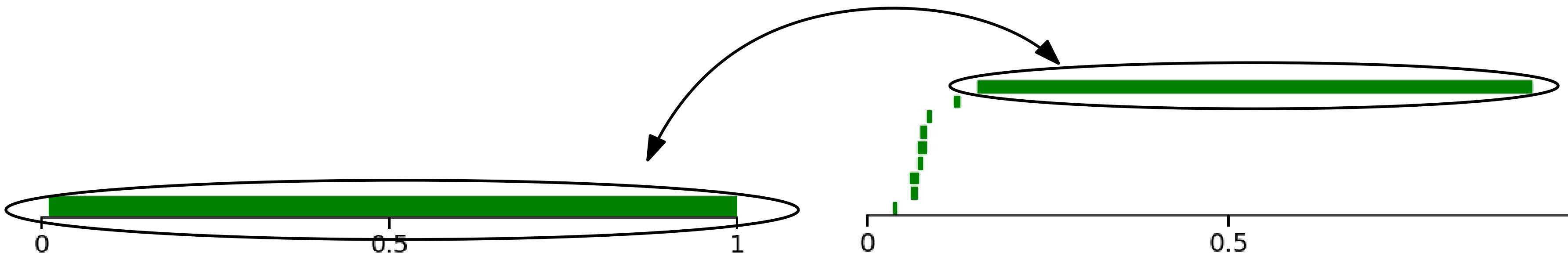


Definition: The **bottleck distance** between P and Q is defined as the infimum of costs over all the partial matchings:

$$\mathrm{d_b}\,(P,Q) = \inf\{\mathrm{cost}(M),\ M \text{ is a partial matching between } P \text{ and } Q\}.$$

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.

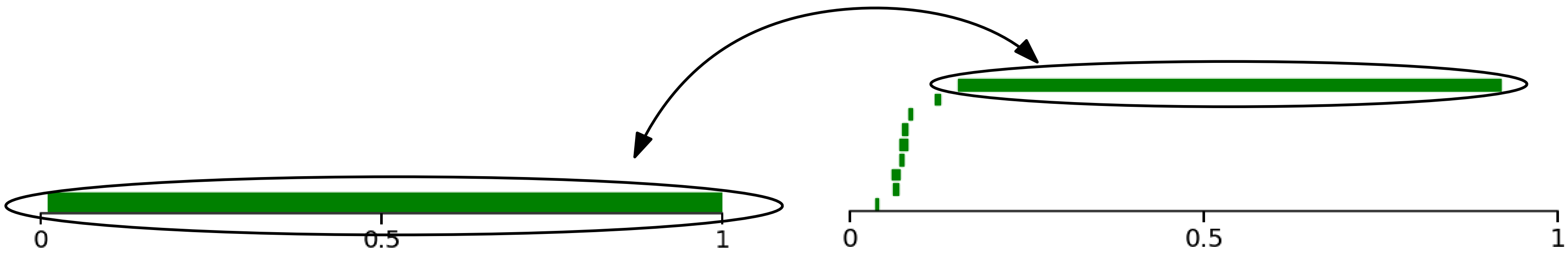


Definition: The **bottleck distance** between P and Q is defined as the infimum of costs over all the partial matchings:

$$\mathrm{d_b}\,(P, Q) = \inf\{\mathrm{cost}(M),\ M \text{ is a partial matching between } P \text{ and } Q\}.$$

Bottleneck distance

Consider two barcodes P and Q, that is, multisets of intervals $\{(a_i, b_i), i \in \mathcal{I}\}$ of $(\overline{\mathbb{R}^+})^2$ such that $a_i \leq b_i$ for all $i \in \mathcal{I}$.



Definition: The **bottleck distance** between P and Q is defined as the infimum of costs over all the partial matchings:

$$\mathrm{d_b}(P, Q) = \inf\{\mathrm{cost}(M),\ M \text{ is a partial matching between } P \text{ and } Q\}.$$

If $\mathbb{U}$ and $\mathbb{V}$ are two decomposable persistence modules, we define their **bottleneck distance** as

$$\mathrm{d_b}(\mathbb{U}, \mathbb{V}) = \mathrm{d_b}(\mathrm{Barcode}(\mathbb{U}), \mathrm{Barcode}(\mathbb{V})).$$

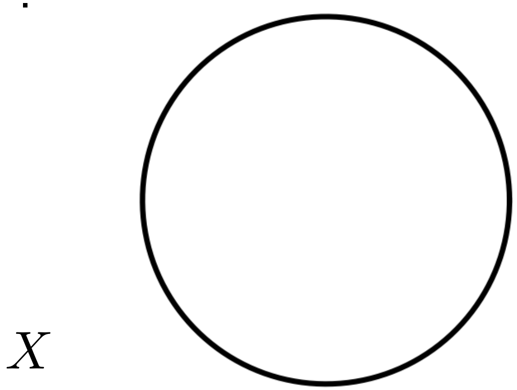
Stability theorem

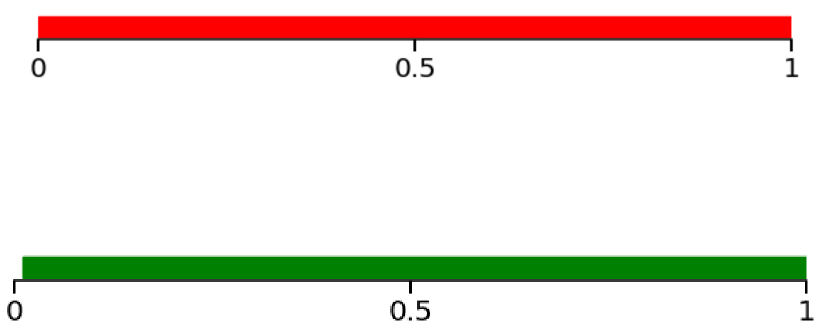
Theorem (Cohen-Steiner, Edelsbrunner, Harer, 2005):
Let X and Y be two subsets of $\mathbb{R}^n$.
Consider their Čech (resp. Rips) filtrations, and the corresponding i^{th} homology persistence modules, $\mathbb{U}$ and $\mathbb{V}$.
Then
$$\mathrm{d_b}\left(\mathrm{Barcode}\left(\mathbb{U}\right), \mathrm{Barcode}\left(\mathbb{V}\right)\right) \leq \mathrm{d_H}\left(X, Y\right)$$

bottleneck distance

Hausdorff distance

.





Stability theorem

Theorem (Cohen-Steiner, Edelsbrunner, Harer, 2005):
Let X and Y be two subsets of $\mathbb{R}^n$.
Consider their Čech (resp. Rips) filtrations, and the corresponding i^{th} homology persistence modules, $\mathbb{U}$ and $\mathbb{V}$.
Then
$$\mathrm{d_b}\left(\mathrm{Barcode}\left(\mathbb{U}\right), \mathrm{Barcode}\left(\mathbb{V}\right)\right) \leq \mathrm{d_H}\left(X, Y\right)$$

.

bottleneck distance

Hausdorff distance

Theorem (Chazal, de Silva, Glisse, Oudot, 2009):
We have
$$\mathrm{d_i}\left(\mathbb{U}, \mathbb{V}\right) = \mathrm{d_b}\left(\mathbb{U}, \mathbb{V}\right),$$
where $\mathrm{d_i}$ is the **interleaving distance**.

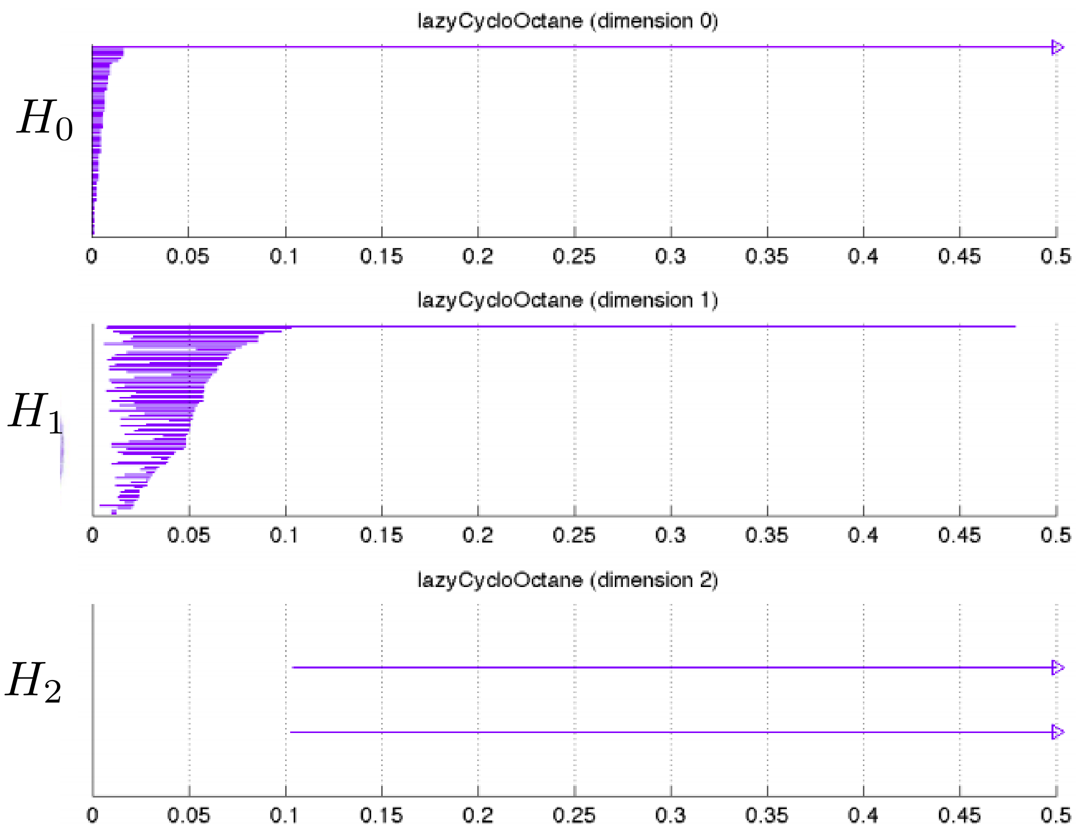
Topological inference I

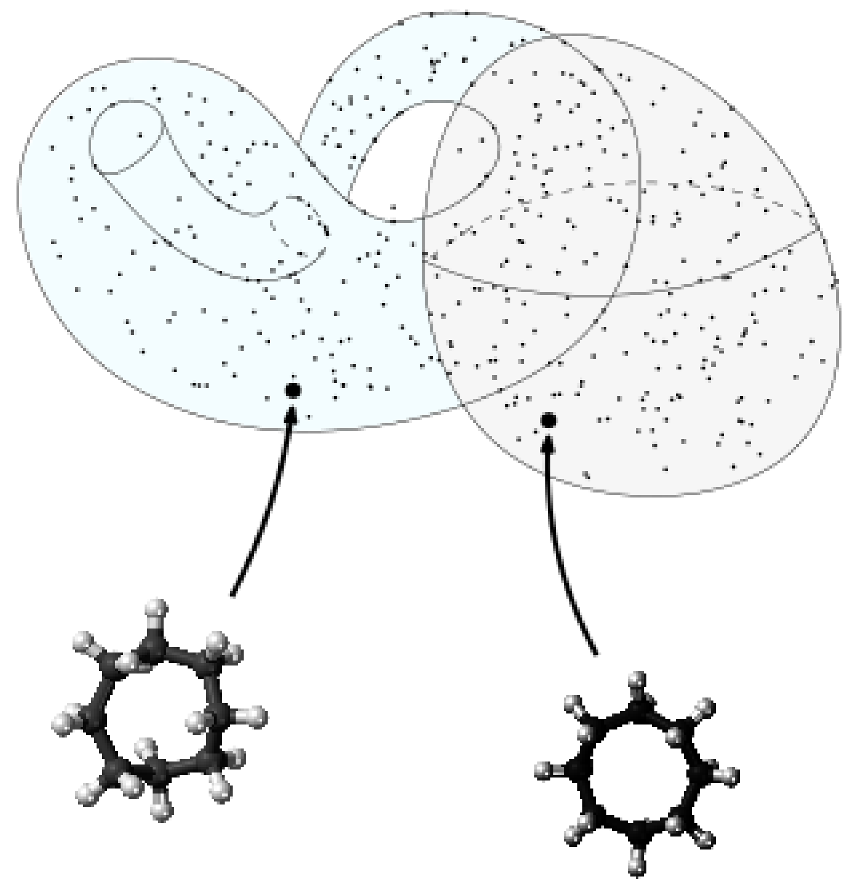
S. Martin, A. Thompson, E. A. Coutsias, and J-P. Watson, Topology of cyclo-octane energy landscape , 2010

https://www.researchgate.net/publication/44697030_Topology_of_Cyclooctane_Energy_Landscape

The cyclo-octane molecule $\mathrm{C_8H_{16}}$ contains 24 atoms.
By generating many of these molecules, we obtain a point cloud in $\mathbb{R}^{72}$ $(3 \times 24 = 72)$.

We obtain the barcodes:





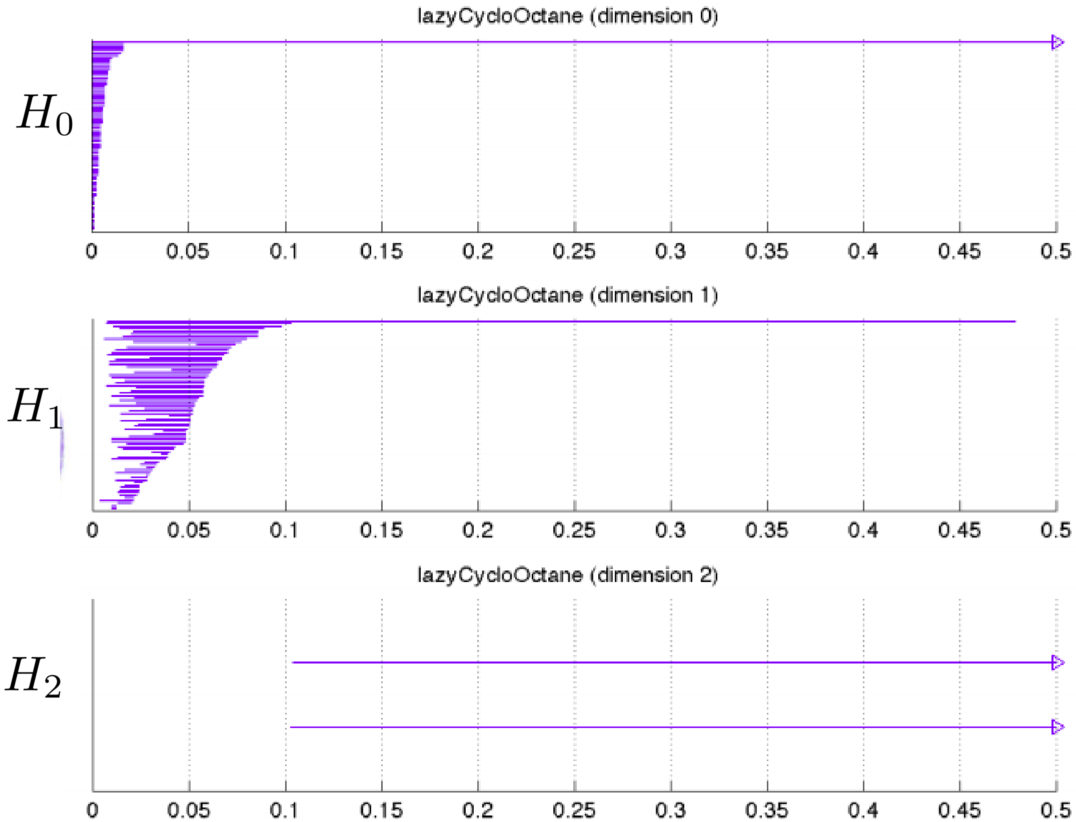
Topological inference I

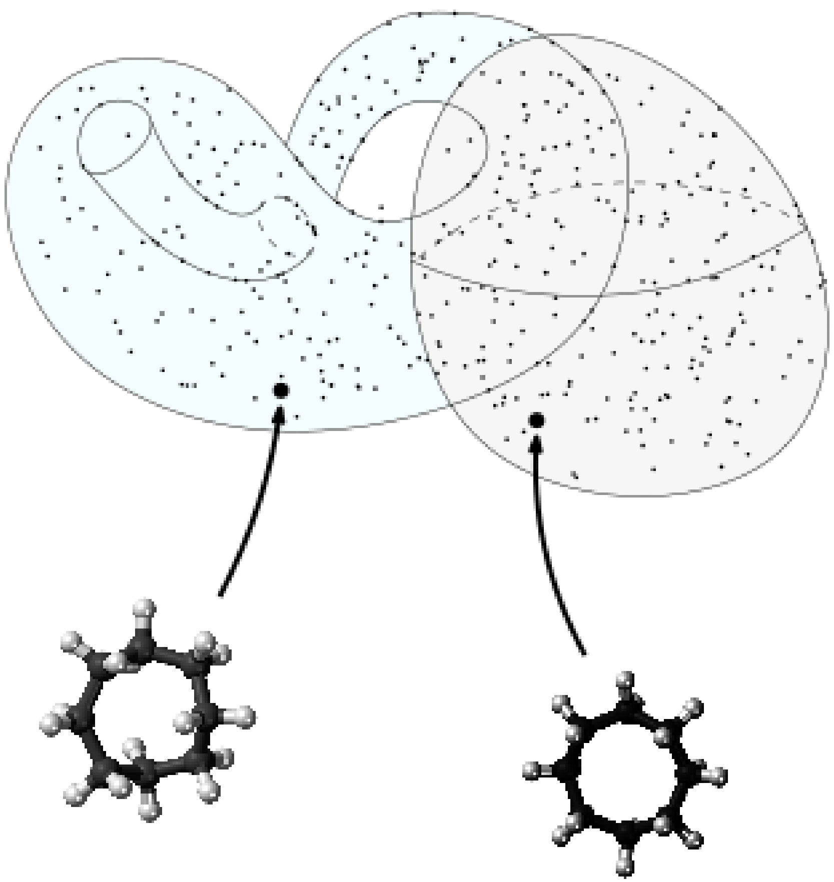
S. Martin, A. Thompson, E. A. Coutsias, and J-P. Watson, Topology of cyclo-octane energy landscape , 2010

https://www.researchgate.net/publication/44697030_Topology_of_Cyclooctane_Energy_Landscape

The cyclo-octane molecule $\mathrm{C_8H_{16}}$ contains 24 atoms.
By generating many of these molecules, we obtain a point cloud in $\mathbb{R}^{72}$ $(3 \times 24 = 72)$.

We obtain the barcodes:





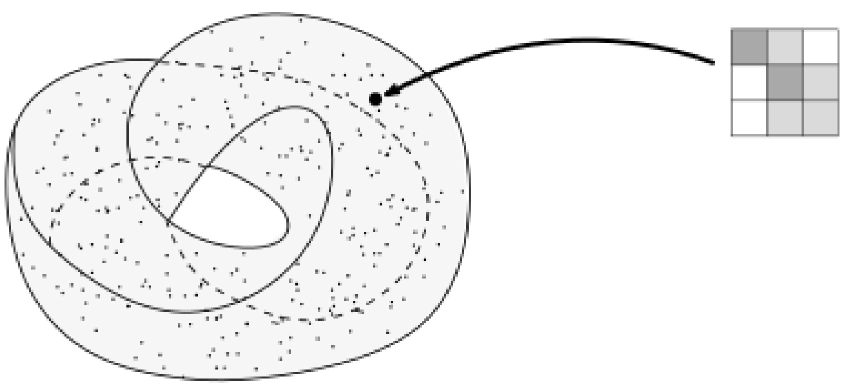
We deduce : $H_0 = \mathbb{Z}/2\mathbb{Z}, \quad H_1 = \mathbb{Z}/2\mathbb{Z}, \quad H_2 = (\mathbb{Z}/2\mathbb{Z})^2$

Topological inference II

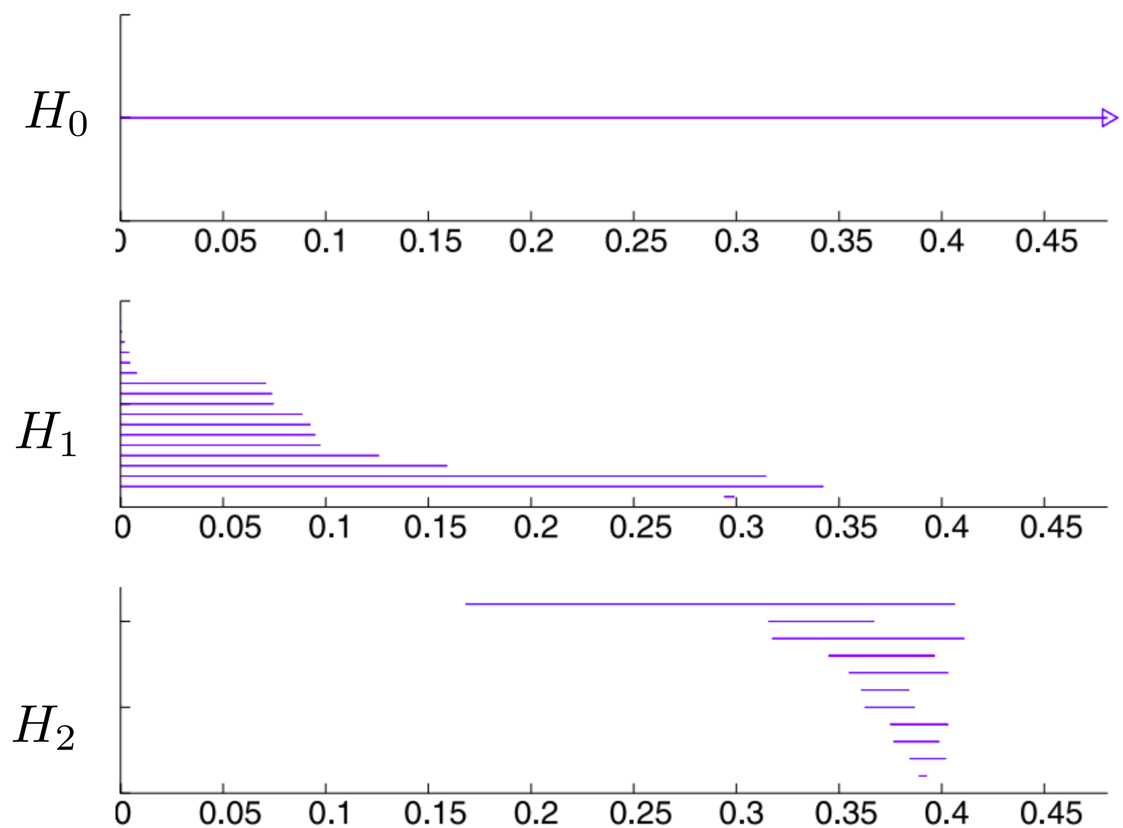
G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, On the Local Behavior of Spaces of Natural Images, 2008

https://link.springer.com/article/10.1007/s11263-007-0056-x

From a large collection of natural images, the authors extract 3×3 patches. Since it consists of 9 pixels, each of these patches can be seen as a 9-dimensional vector, and the whole set as a point cloud in $\mathbb{R}^9$.



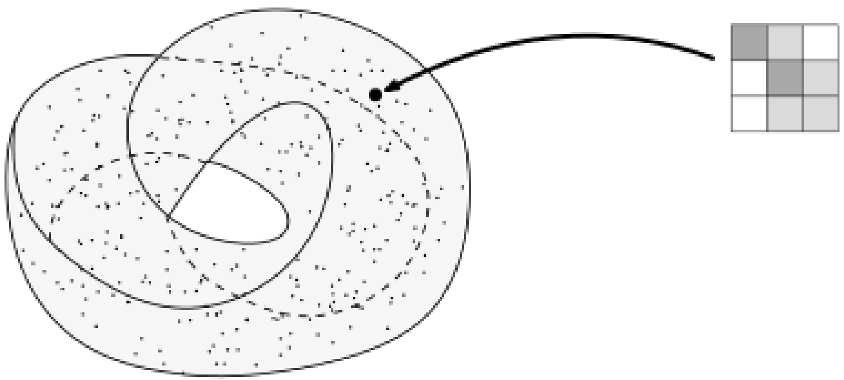
We get the barcodes:



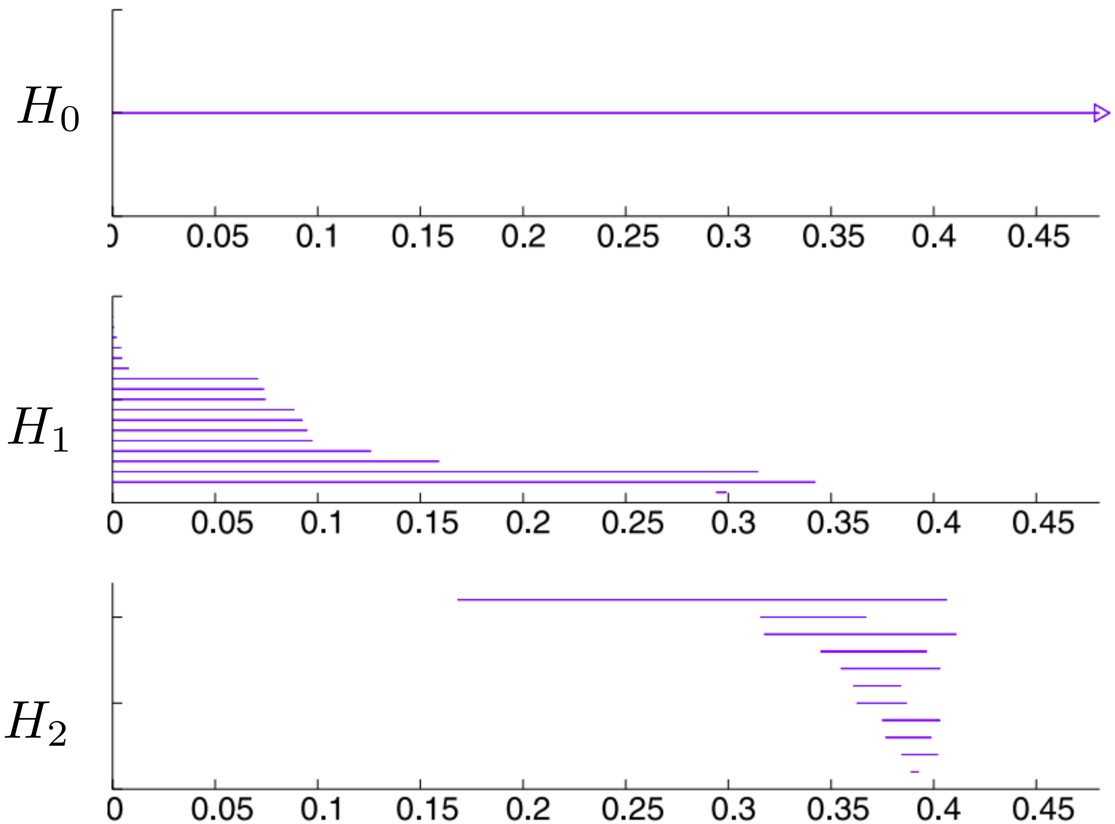
G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, On the Local Behavior of Spaces of Natural Images, 2008

https://link.springer.com/article/10.1007/s11263-007-0056-x

From a large collection of natural images, the authors extract 3×3 patches.
Since it consists of 9 pixels, each of these patches can be seen as a 9-dimensional vector, and the whole set as a point cloud in $\mathbb{R}^9$.



We get the barcodes:



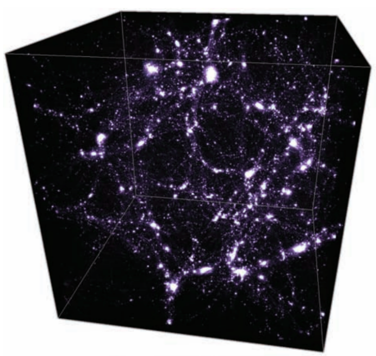
We deduce:

$H_0 = \mathbb{Z}/2\mathbb{Z}$,
$H_1 = (\mathbb{Z}/2\mathbb{Z})^2$,
$H_2 = \mathbb{Z}/2\mathbb{Z}$.

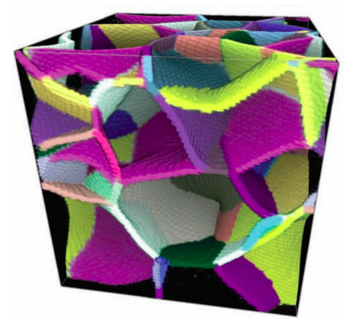
Multiscale analysis I

T. Sousbie, The persistent cosmic web and its filamentary structure, 2011

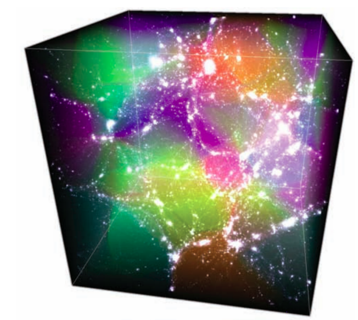
https://www.giss.nasa.gov/staff/mway/cluster/sousbie2011mnras.pdf



seen as an object of dimension 1



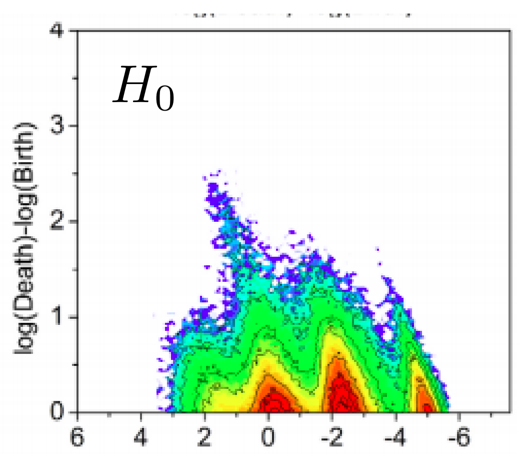
of dimension 2

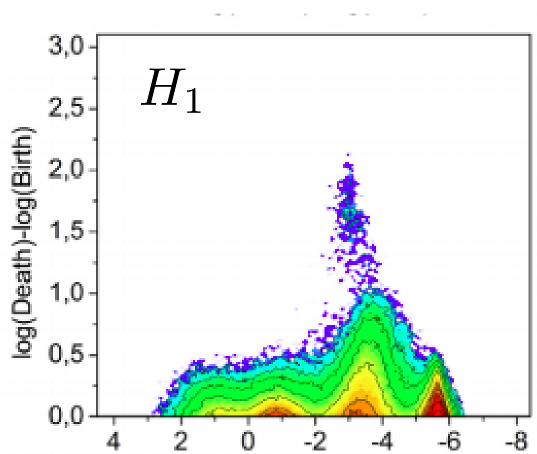


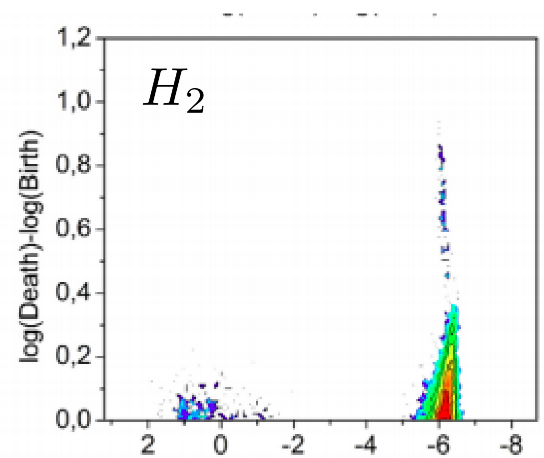
of dimension 3

P. Pranav, H. Edelsbrunner, R. de Weygaert, G. Vegter, M. Kerber, B. Jones and M. Wintraecken, The topology of the cosmic web in terms of persistent Betti numbers, 2016

https://arxiv.org/pdf/1608.04519.pdff





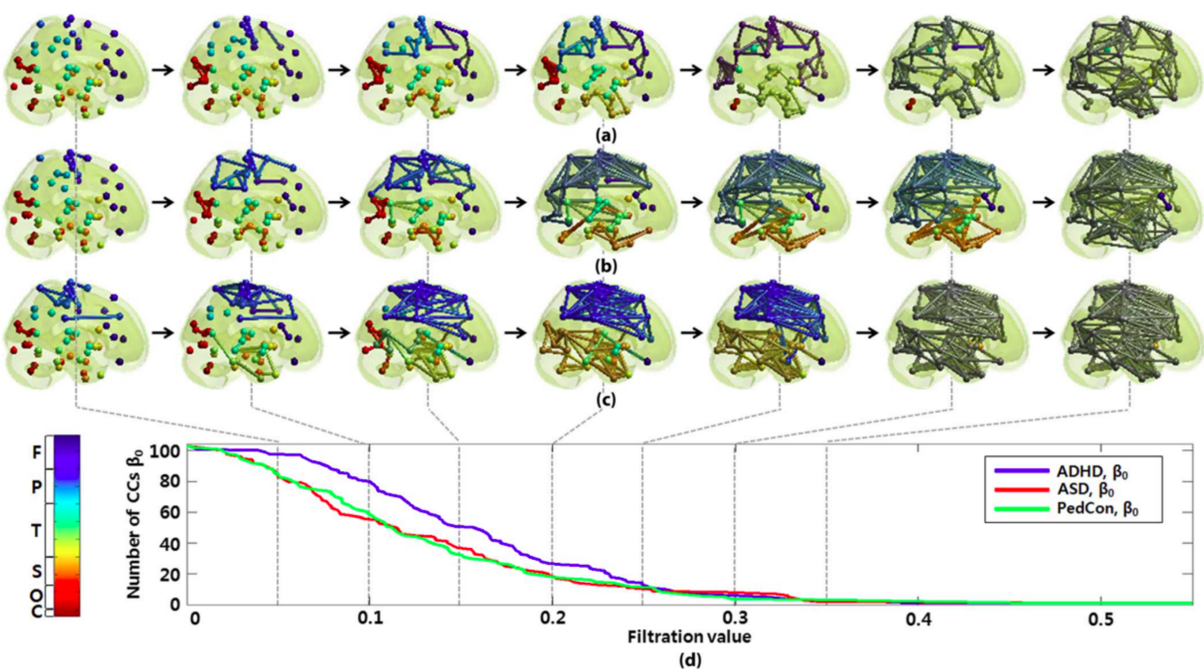


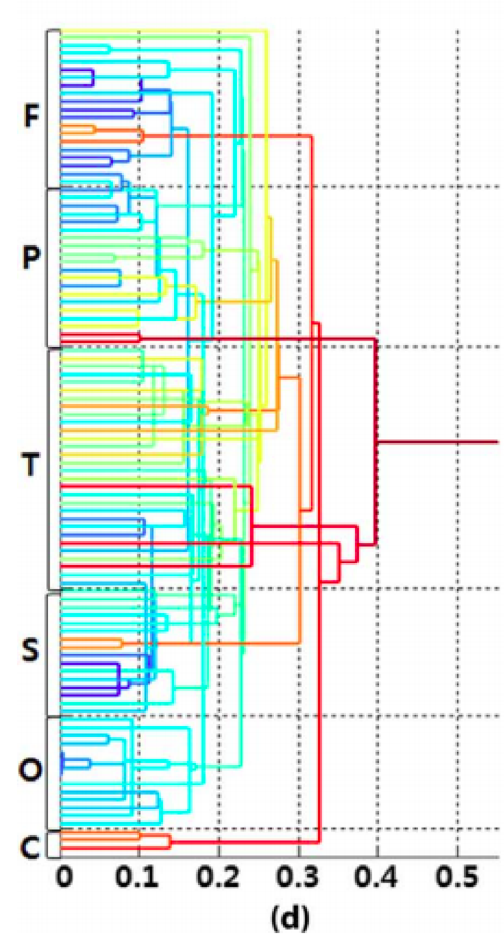
Average persistence diagrams (log-scale) for a Voronoï evolution model

Hyekyoung Lee, Hyejin Kang, Moo K Chung, Bung-Nyun Kim, Dong Soo Lee, Persistent brain network homology from the perspective of dendrogram, 2012

http://pages.stat.wisc.edu/~mchung/papers/lee.2012.TMI.pdf

$\longrightarrow$ H_0-persistent homology induces a hierarchical clustering





In machine learning I

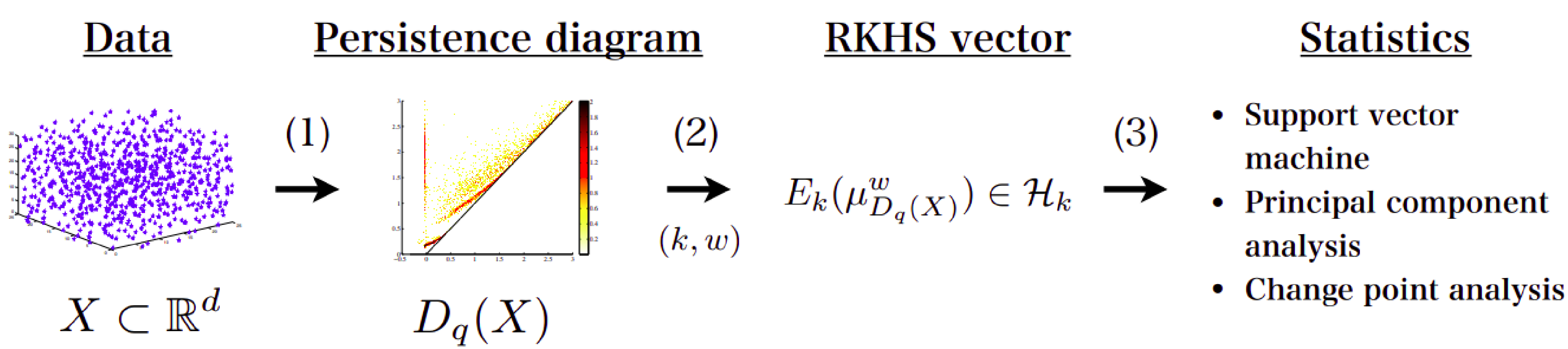
Mathieu Carrière, Marco Cuturi, Steve Oudot, Sliced Wasserstein Kernel for Persistence Diagrams, 2017

https://arxiv.org/abs/1706.03358

Genki Kusano, Kenji Fukumizu, Yasuaki Hiraoka, Kernel Method for Persistence Diagrams via Kernel Embedding and Weight Factor, 2018

https://www.jmlr.org/papers/volume18/17-317/17-317.pdf

⟶ Barcodes are not subsets of some Euclidean space, hence usual machine learning methods cannot be used directly



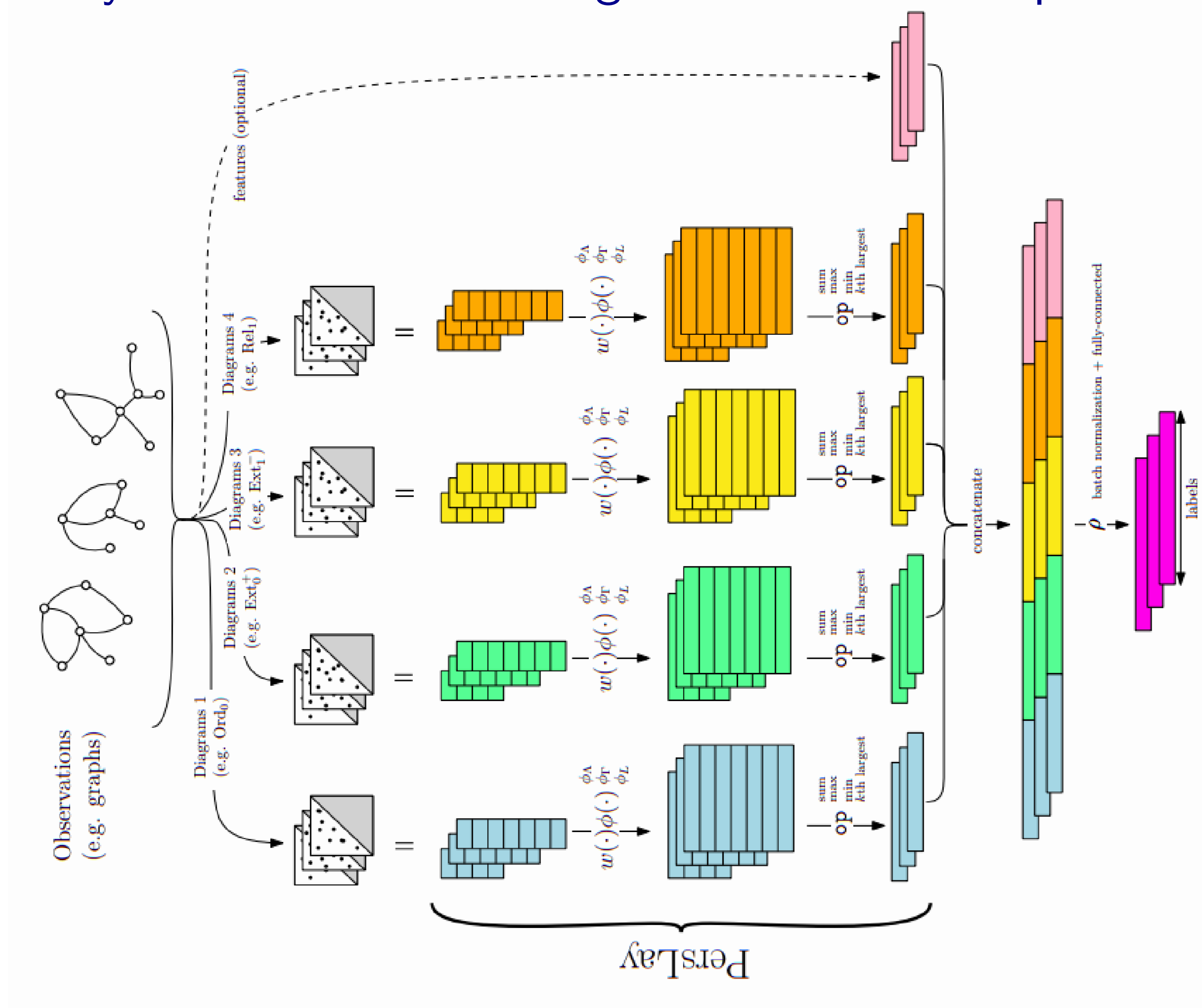
In machine learning II

Rickard Brüel-Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, Primoz Skraba, Leonidas J. Guibas, Gunnar Carlsson, A Topology Layer for Machine Learning, 2019

https://arxiv.org/abs/1905.12200

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, Yuhei Umeda, PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures, 2019

https://arxiv.org/abs/1904.09378



Classification

Frédéric Chazal, Steve Oudot, Primoz Skraba, Leonidas J. Guibas, Persistence-Based Clustering in Riemannian Manifolds, 2011
https://geometrica.saclay.inria.fr/team/Fred.Chazal/papers/cgos-pbc-09/cgos-pbcrm-11.pdf

Chunyuan Li, Maks Ovsjanikov, Frederic Chazal, Persistence-based Structural Recognition, 2014
https://geometrica.saclay.inria.fr/team/Fred.Chazal/papers/loc-pbsr-14/CVPR2014.pdf

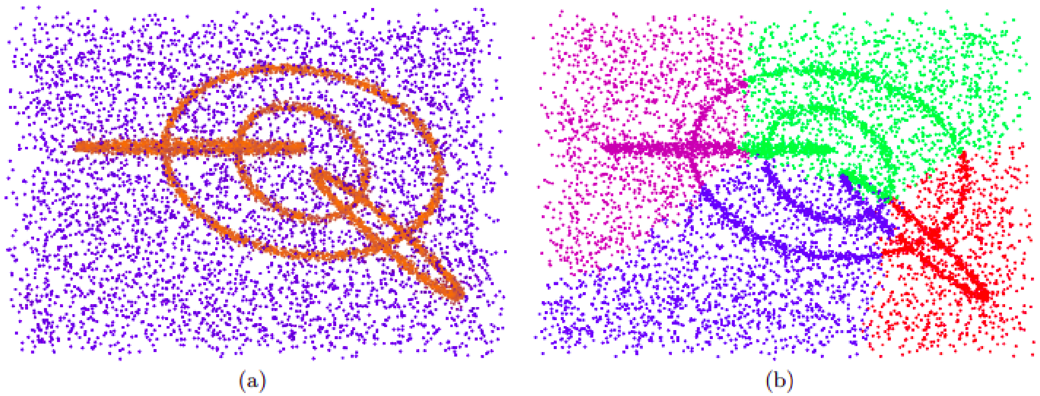
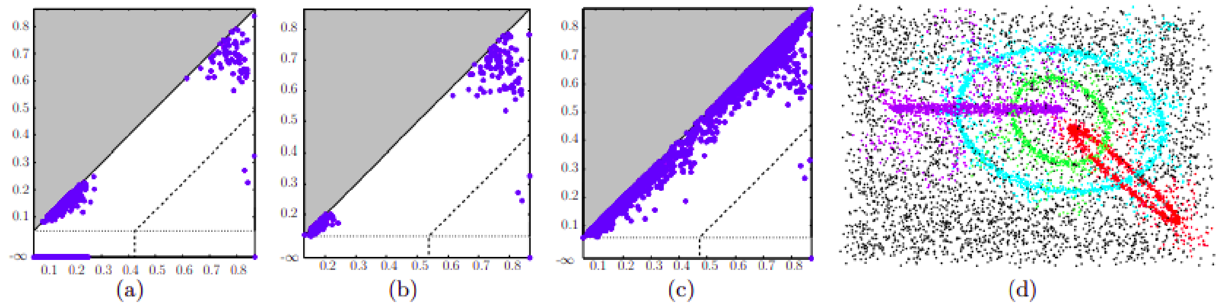
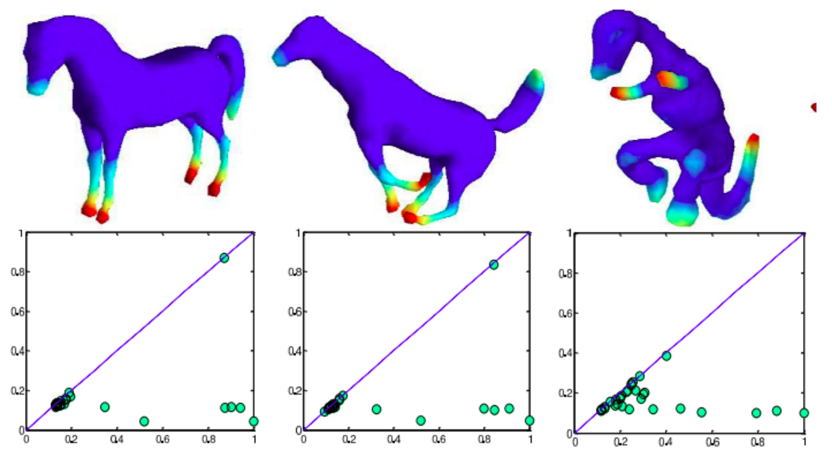


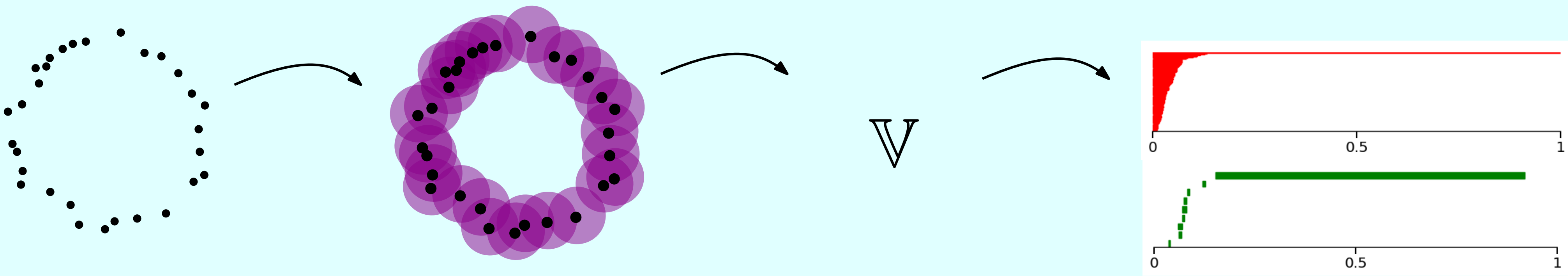
Figure 7: (a) The rings data set with the estimated density function. (b) The result obtained using spectral clustering.





Conclusion

Persistent homology allows a **multiscale** and **stable** estimation of the homology of the dataset.



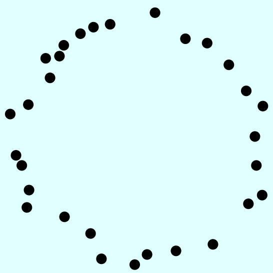
Point cloud Čech filtration Persistence module Barcodes

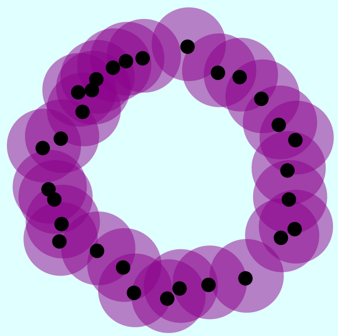
Illuminates data analysis from a different angle than the usual methods.

A course about TDA: `https://raphaeltinarrage.github.io/EMAp.html`

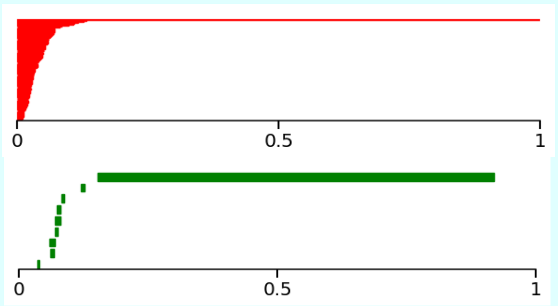
Conclusion

Persistent homology allows a **multiscale** and **stable** estimation of the homology of the dataset.





$\mathbb{V}$



Point cloud — Čech filtration — Persistence module — Barcodes

Illuminates data analysis from a different angle than the usual methods.

A course about TDA: `https://raphaeltinarrage.github.io/EMAp.html`

Valeu!