Université du Luxembourg
March 12 - 16, 2018

Topological Data Analysis

Steve Oudot – Nicolas Berkouk

Resources:

http://geometrica.saclay.inria.fr/team/Steve.Oudot/courses/Luxembourg/

Outline

Monday: introduction to TDA and singular homology (Steve)

Tuesday: 1-d persistence theory (Nicolas + Steve)

Wednesday: applications to data analysis (Steve)

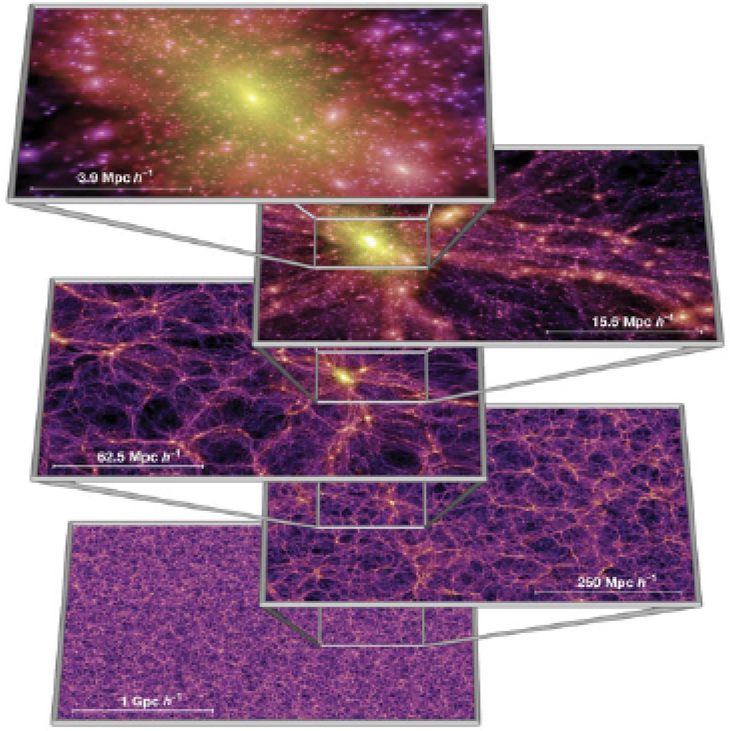
Thursday: multi-dimensional persistence (Nicolas)

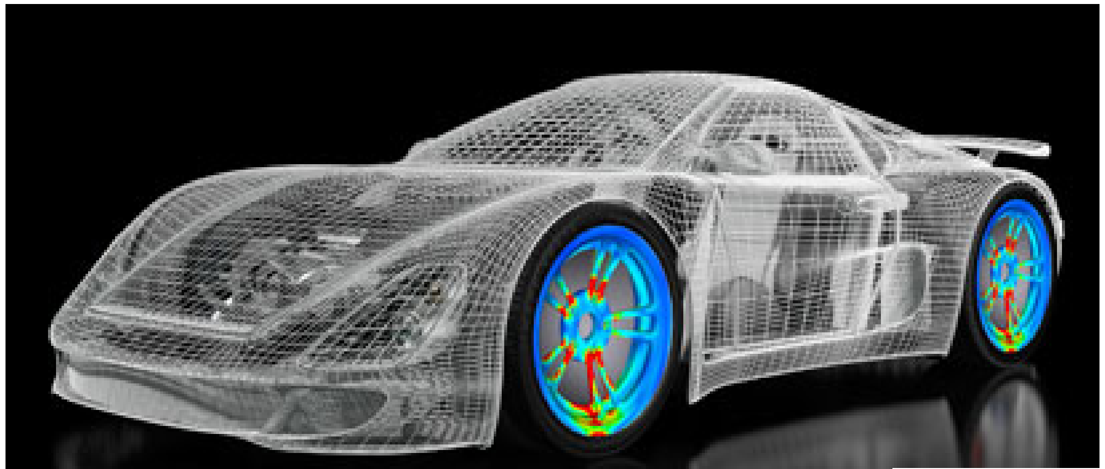
Friday: Persistence and Sheaf theory (Nicolas)

Context: the data deluge

Data are becoming ever more massive and **complex**:

- academia
- industry
- general public

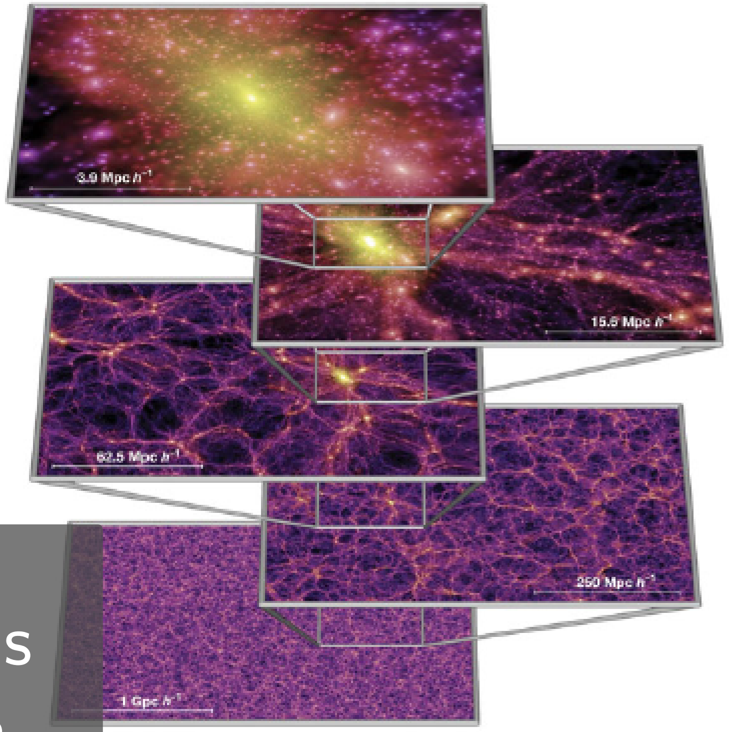




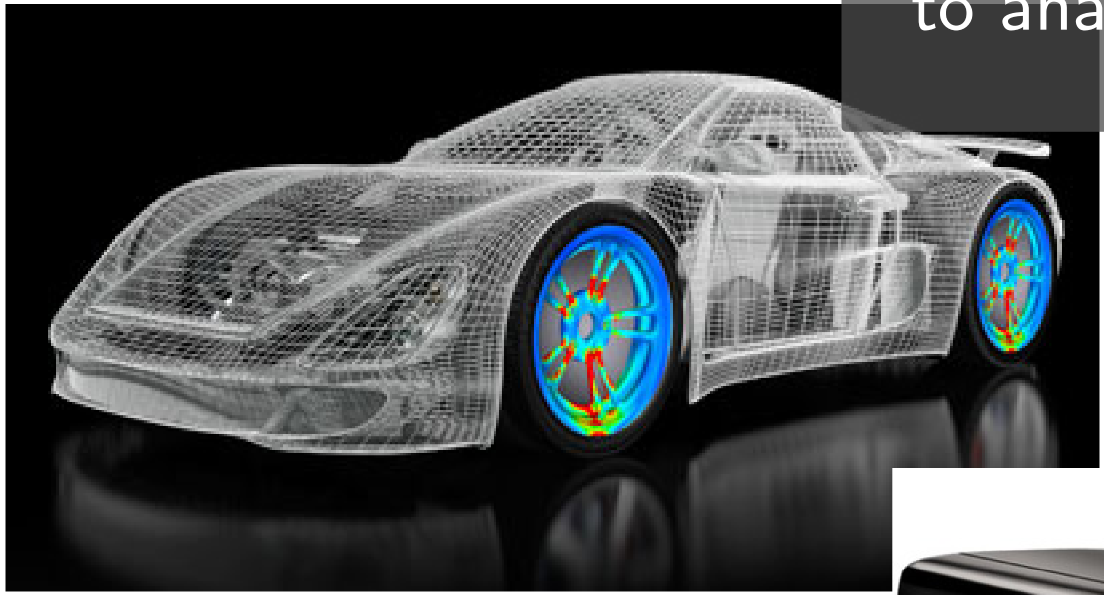
Context: the data deluge

Data are becoming ever more massive and **complex**:

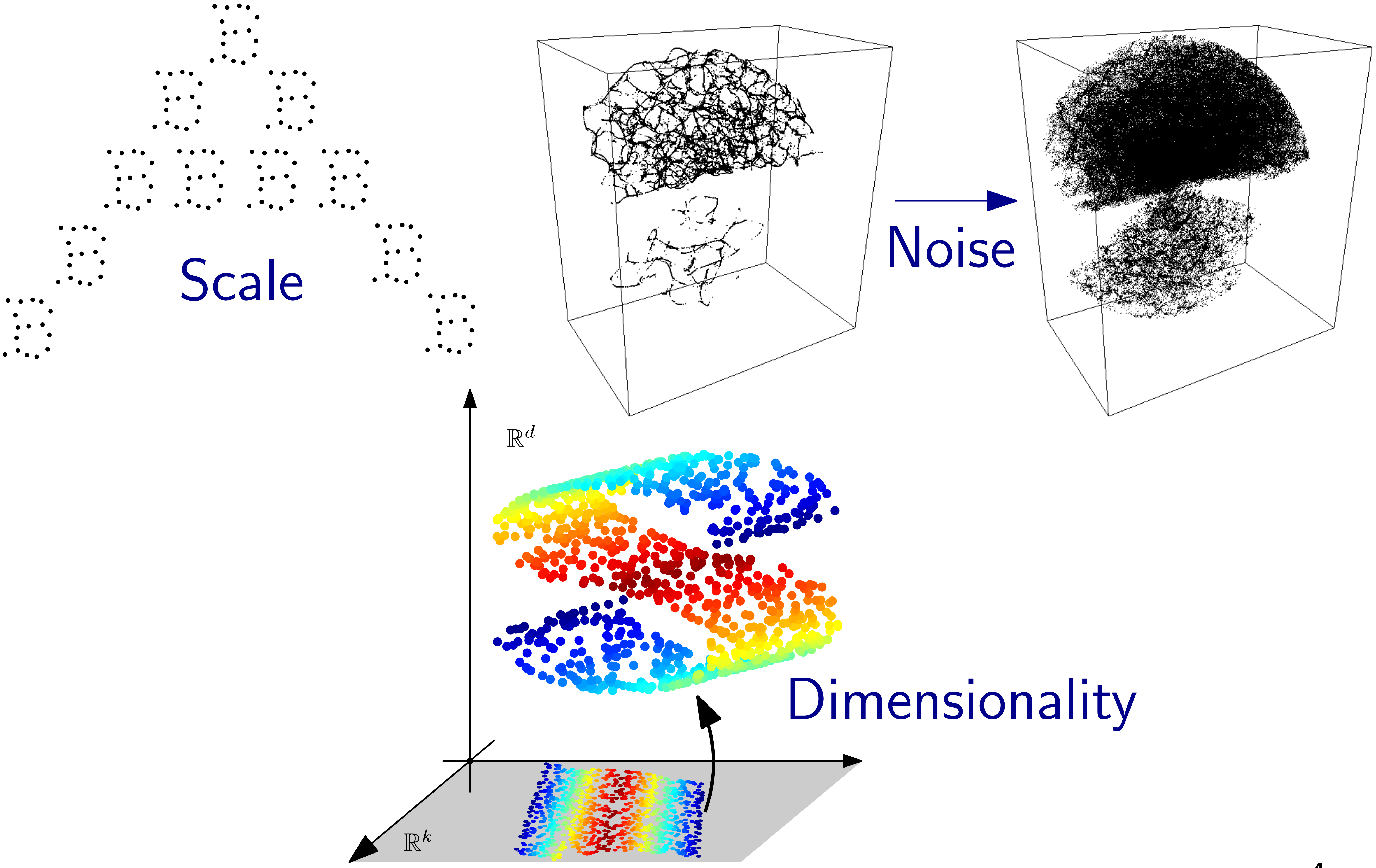
- academia
- industry
- general public



Need scalable and robust methods to analyze and classify these data



Challenges

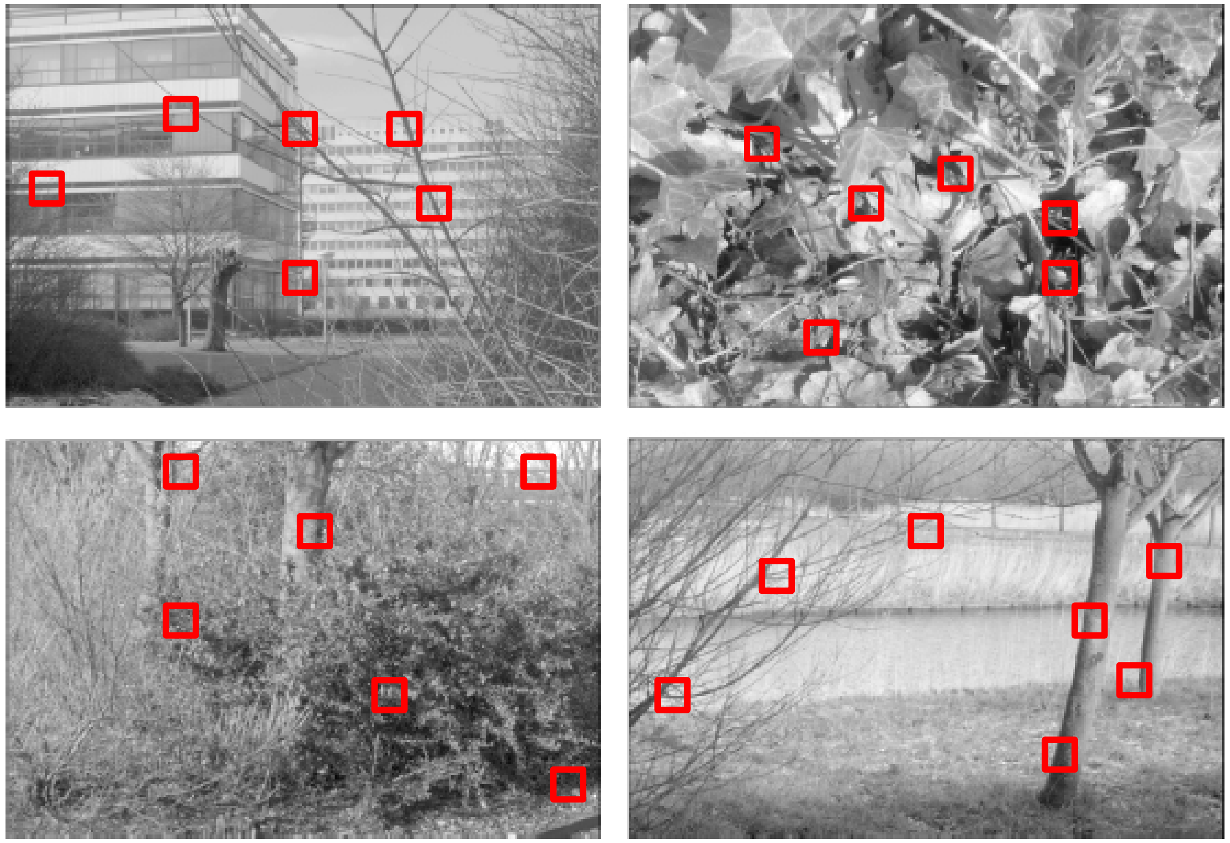


Scale

Noise

Dimensionality

Challenges

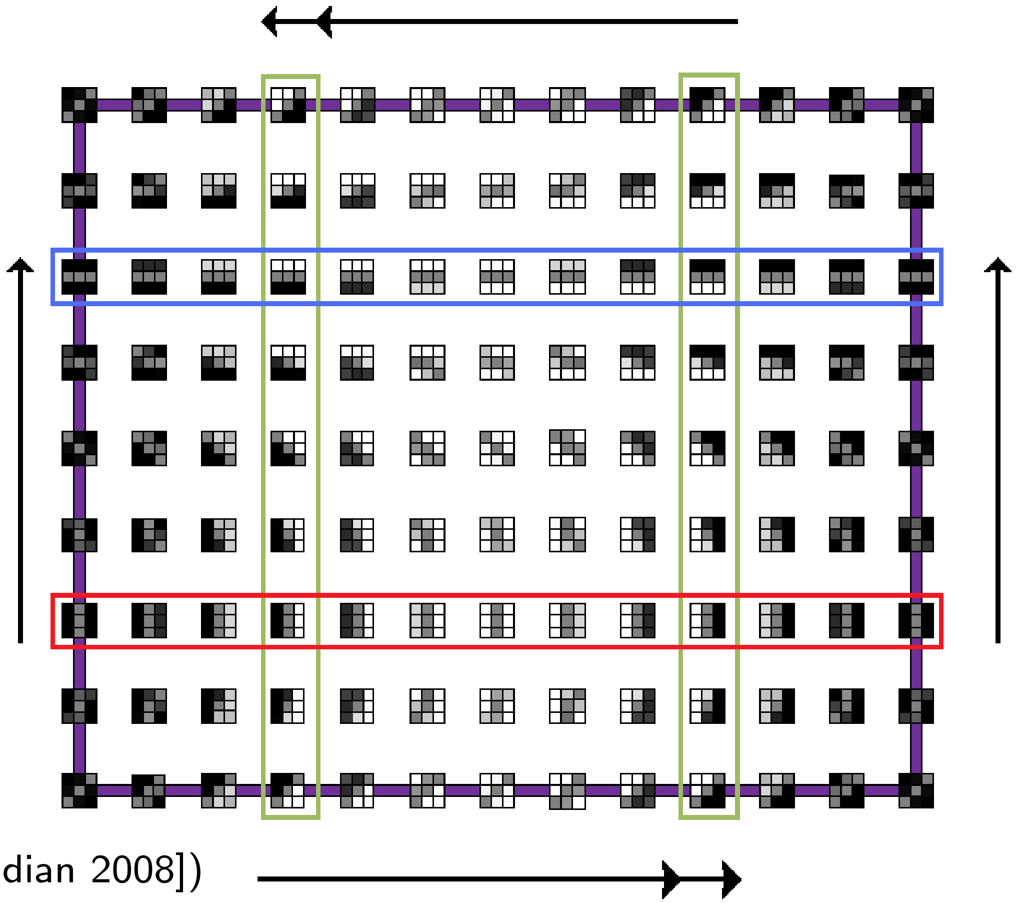


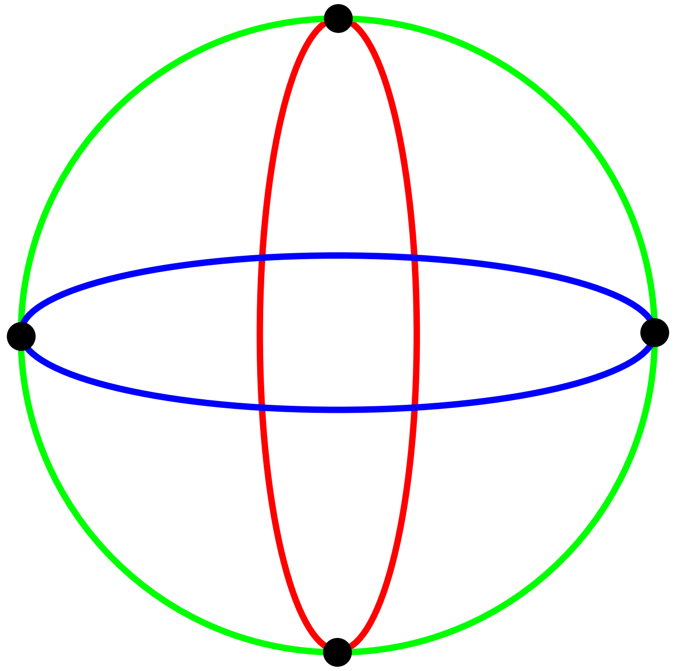
4 million data points in $\mathbb{R}^9$

(source: [Lee, Pederson, Mumford 2003])

Motivation: study cognitive representation of space of images

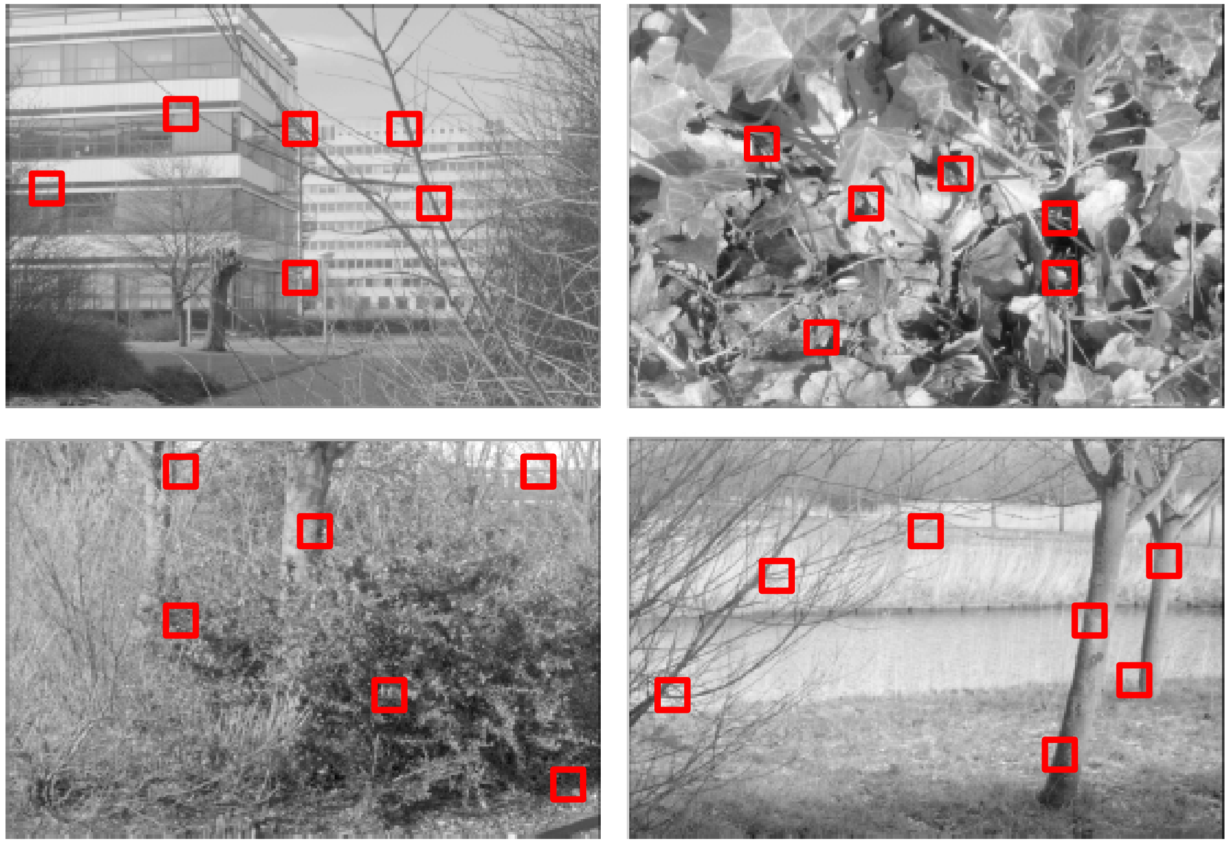
Topology





(source: [Carlsson, Ishkhanov, de Silva, Zomorodian 2008])

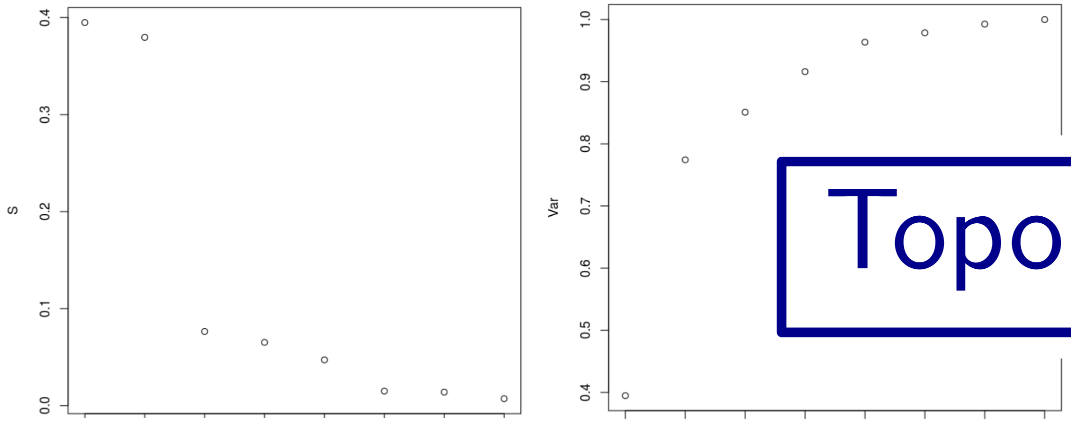
Challenges



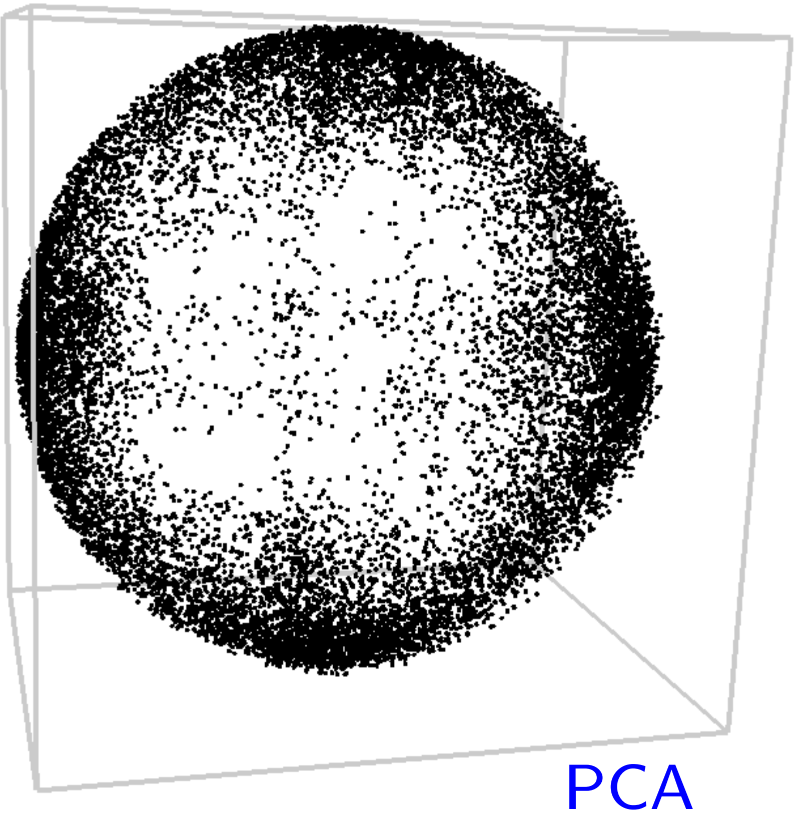
4 million data points in $\mathbb{R}^9$

(source: [Lee, Pederson, Mumford 2003])

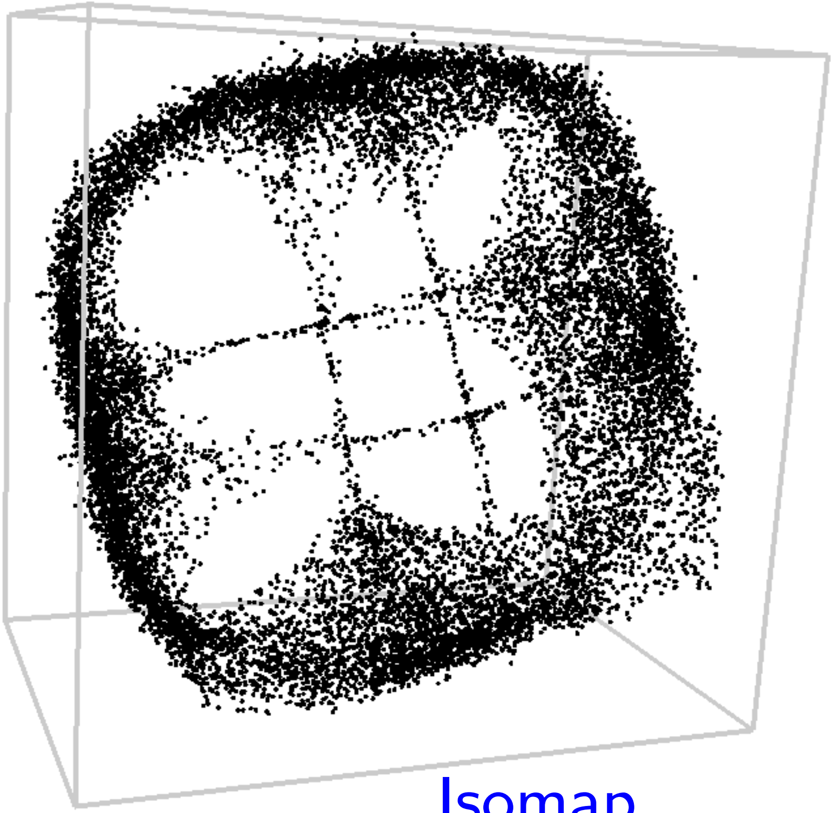
Motivation: study cognitive representation of space of images



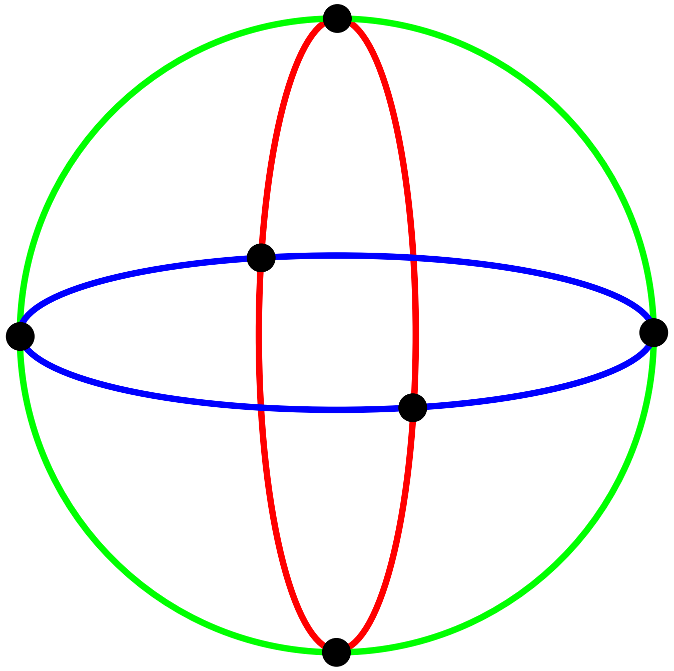
Topology



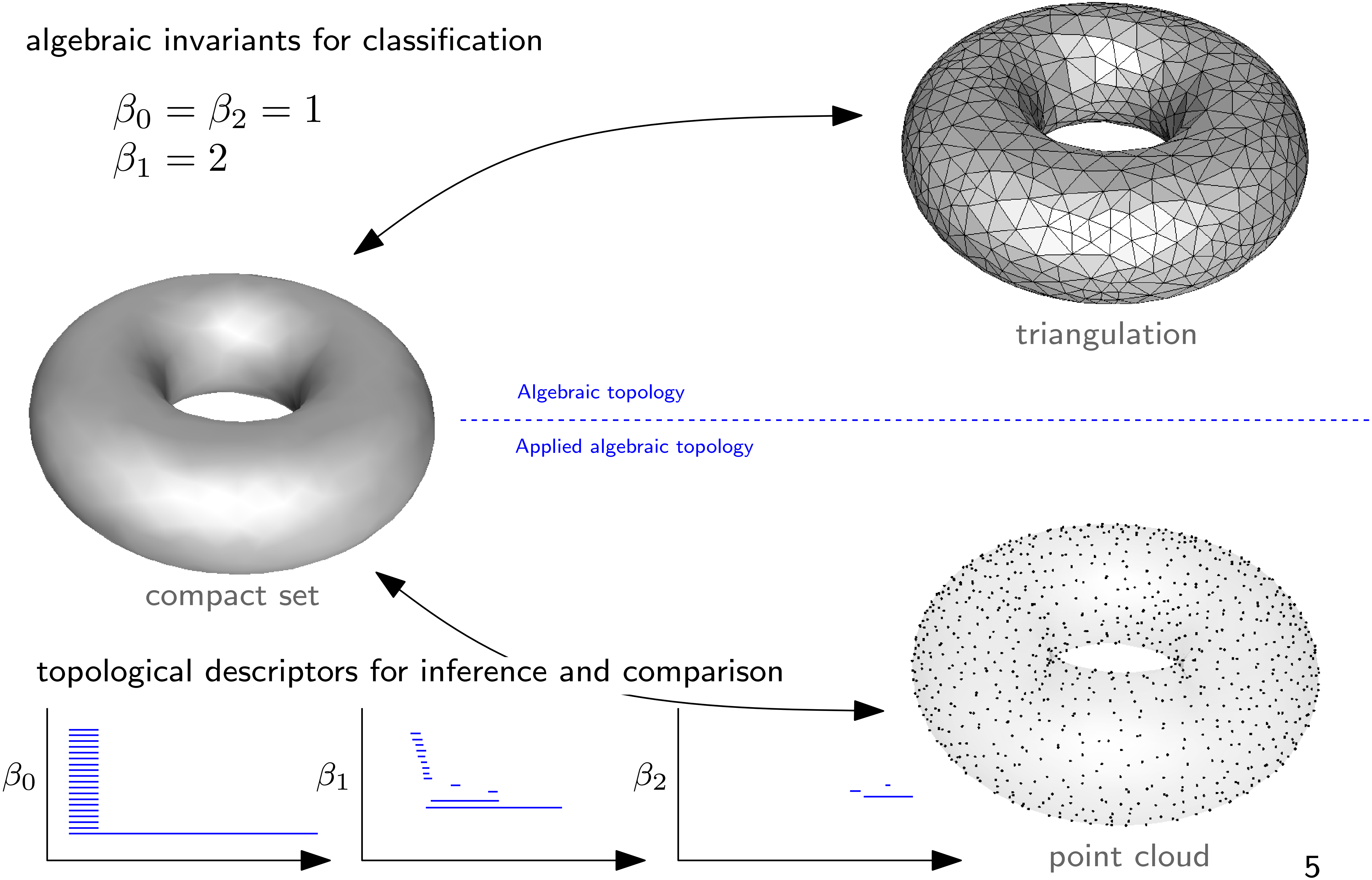
PCA



Isomap



Topological Data Analysis (TDA)



algebraic invariants for classification

$$\beta_0 = \beta_2 = 1$$
$$\beta_1 = 2$$

triangulation

compact set

Algebraic topology

Applied algebraic topology

topological descriptors for inference and comparison

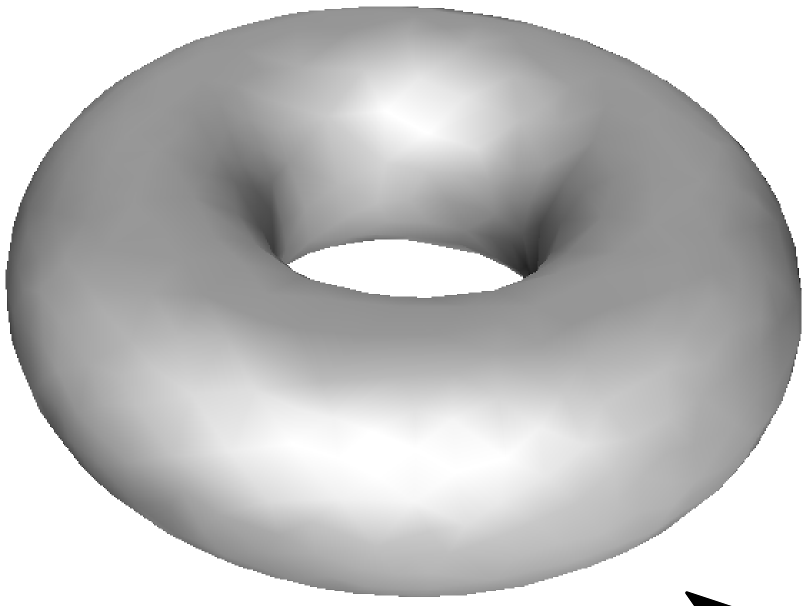
point cloud

Topological Data Analysis (TDA)

Properties of topological descriptors:

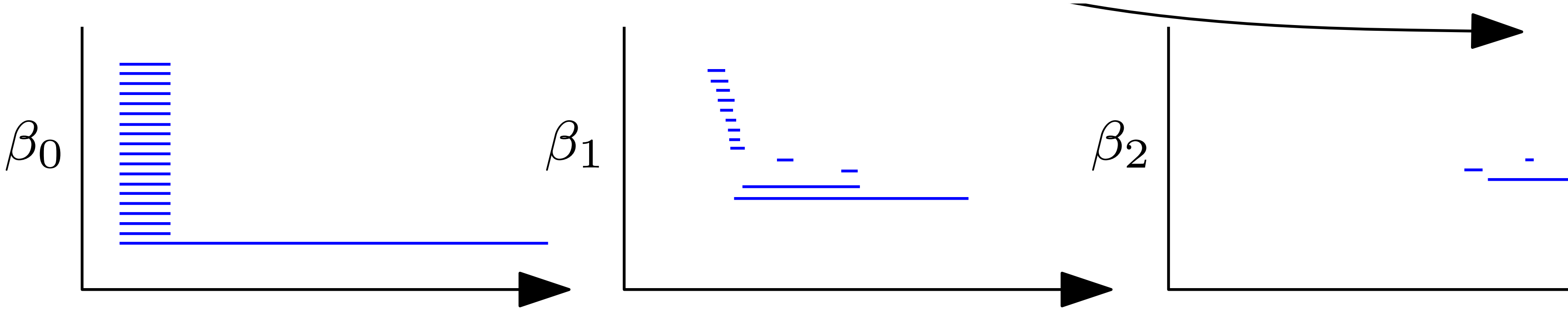
- invariance under reparametrizations
- stability under perturbations
- discrimination power

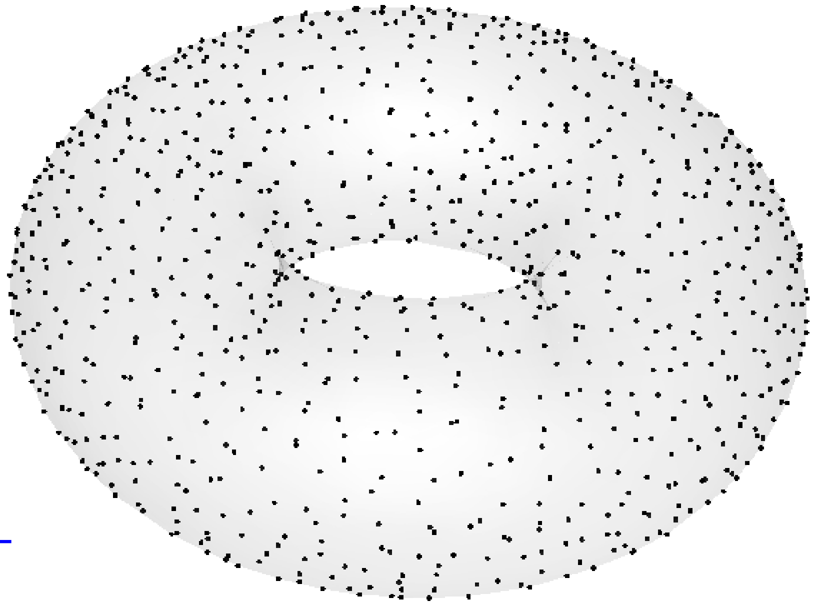
Applied algebraic topology



compact set

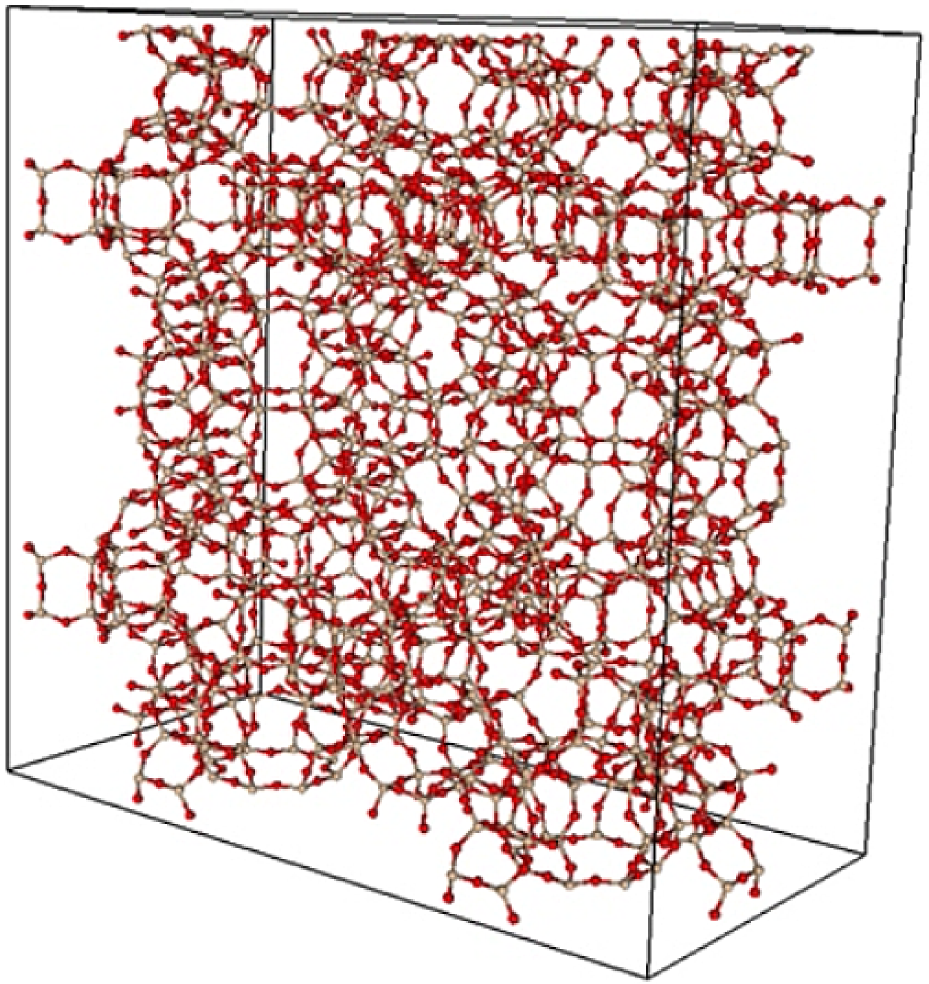
topological descriptors for inference and comparison





point cloud

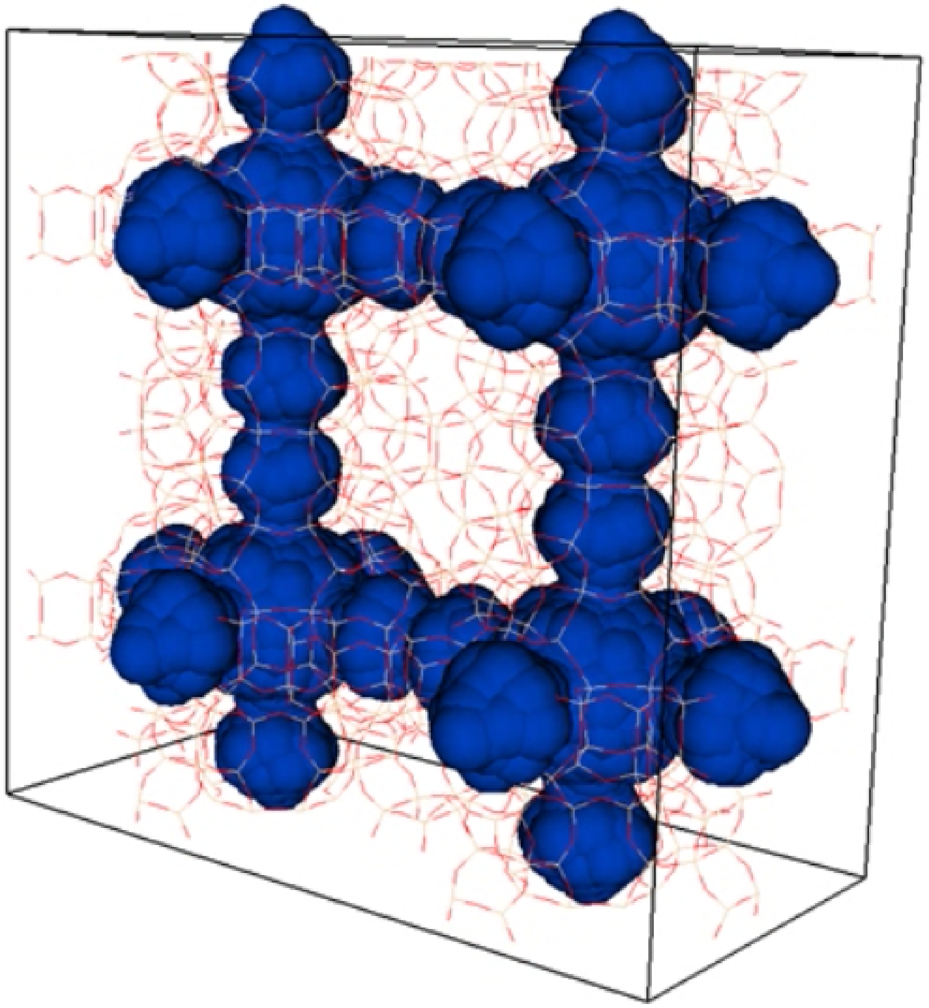
Example: Materials Sciences



Nanoporous material (zeolite ⇝ nanofilter)

[Y. Lee et al.: Quantifying similarity of pore-geometry in nanoporous materials, 2017]

Example: Materials Sciences

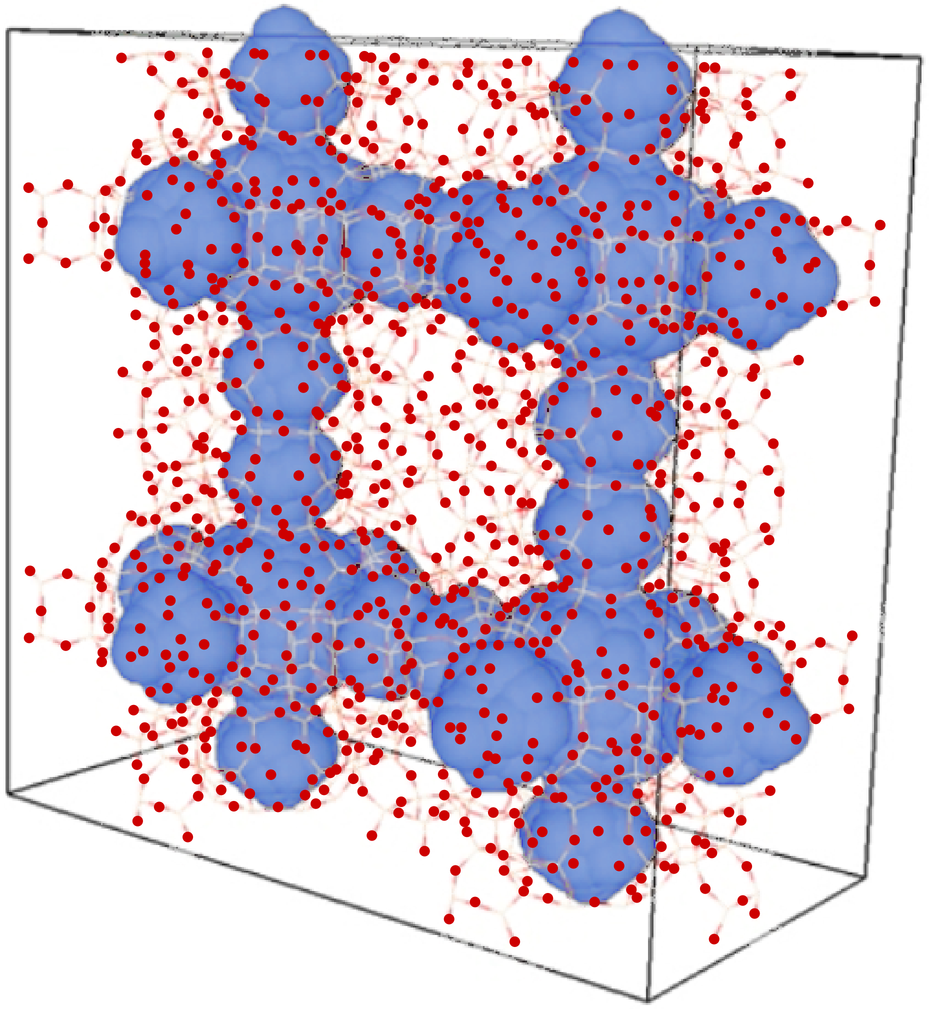


Nanoporous material (zeolite ⇝ nanofilter)

cavities (size, shape, network) determine the material's physical properties

[Y. Lee et al.: Quantifying similarity of pore-geometry in nanoporous materials, 2017]

Example: Materials Sciences



Nanoporous material (zeolite ⇝ nanofilter)

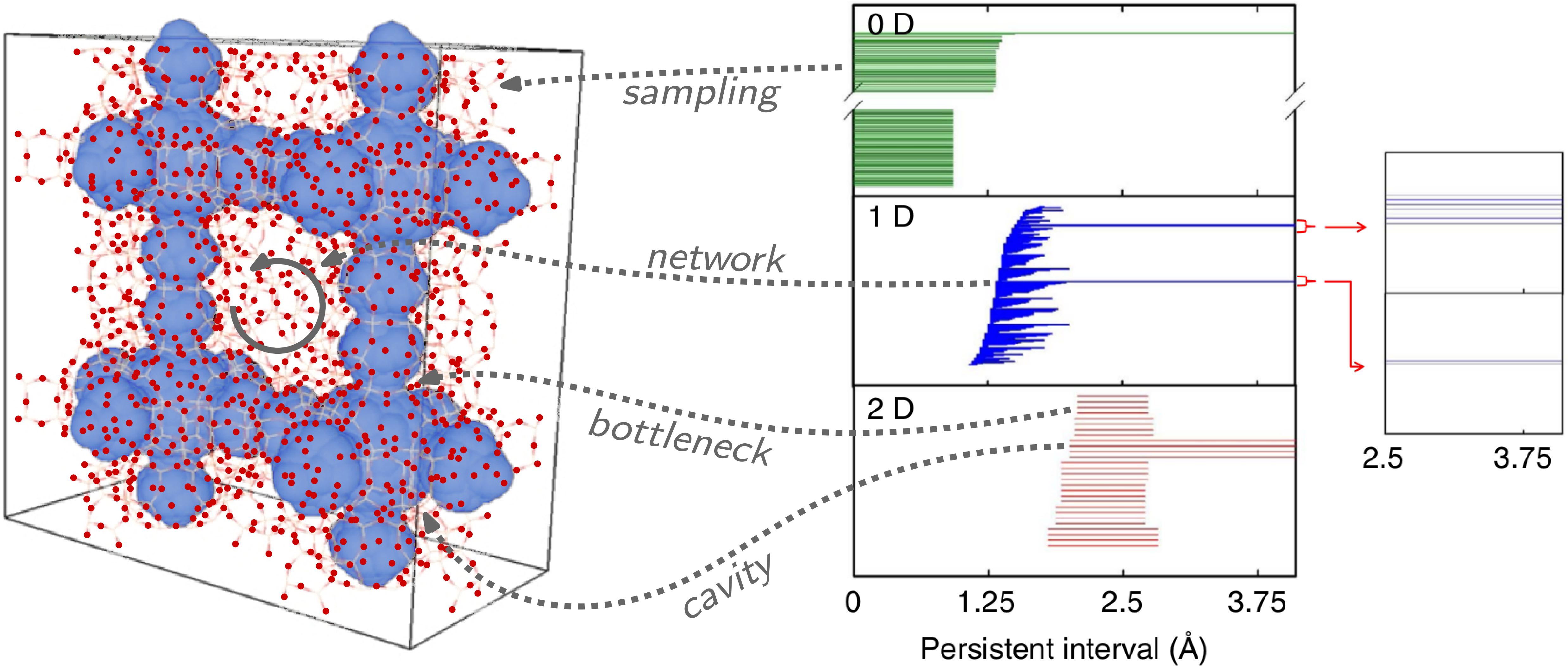
cavities (size, shape, network) determine the material's physical properties

→ need for descriptors that can:

> **capture** shapes hidden in data
>
> **reveal** these shapes to users
>
> **be used as features** for learning

[Y. Lee et al.: Quantifying similarity of pore-geometry in nanoporous materials, 2017]

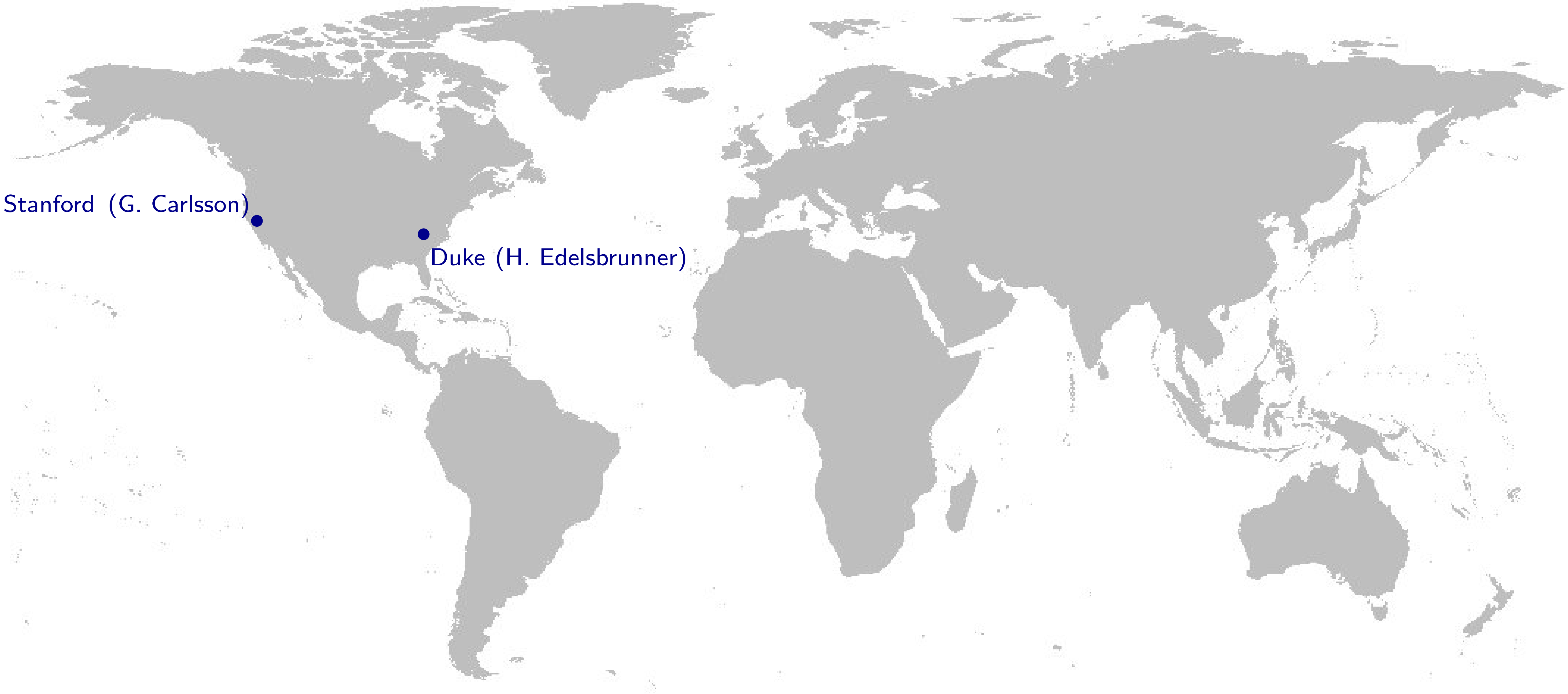
Example: Materials Sciences



→ effective **classification** and **retrieval**, **explainable** properties

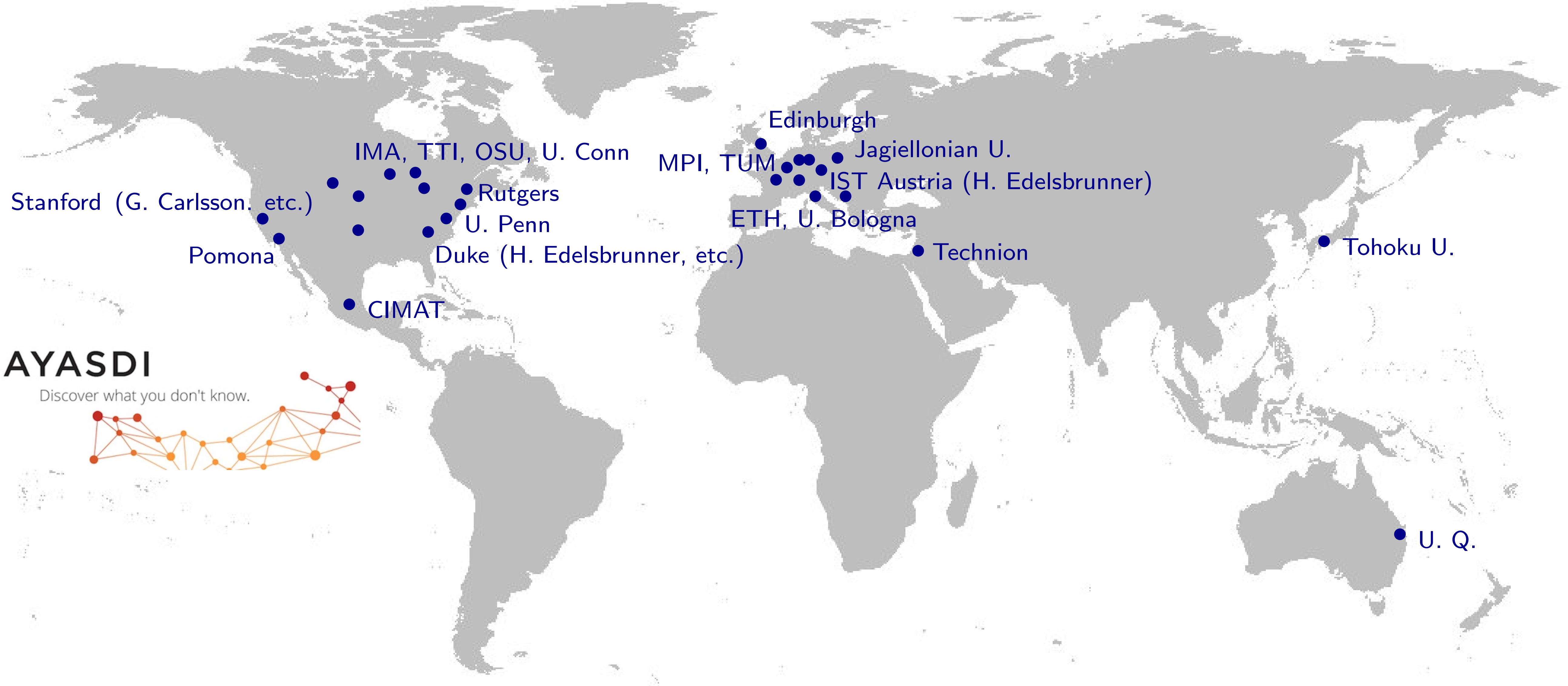
[Y. Lee et al.: Quantifying similarity of pore-geometry in nanoporous materials, 2017]

The TDA community (as of 2002)



- 2 research groups (5-10 researchers)

The TDA community (as of 2016)



- 50-100 researchers working on theoretical foundations
- 200-300 researchers at the interface with applications
- very successful applications and company (Ayasdi)

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,

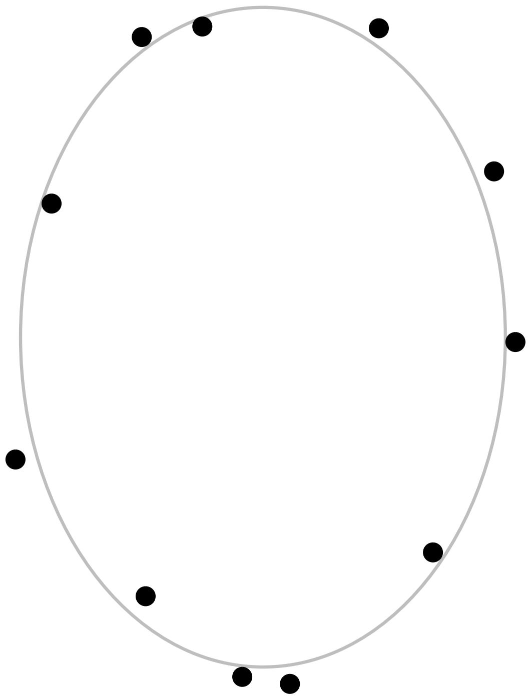
Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,
- analysis of 2d cortical thickness data,
- time series analysis,

Some applications

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution,
- analysis of social and spatial networks like neurons, genes, online messages, air passengers, Twitter, face-to-face contact, etc.,
- coverage and hole detection in wireless sensor fields,
- multiple hypothesis tracking on urban vehicular data,
- analysis of the statistics of high-contrast image patches,
- image segmentation,
- 1d signal denoising,
- 3d shape classification/segmentation/matching,
- clustering of protein conformations,
- measurement of protein compressibility,
- identification of breast cancer subtypes,
- analysis of activity patterns in the primary visual cortex,
- identification of hidden networks in the U.S. house of representatives,
- analysis of 2d cortical thickness data,
- time series analysis,
- refinement of the classification of NBA players,
- discrimination of electroencephalogram signals recorded before and during epileptic seizures,
- statistical analysis of orthodontic data,
- measurement of structural changes during lipid vesicle fusion,
- characterization of the frequency and scale of lateral gene transfer in pathogenic bacteria,
- pattern detection in gene expression data,
- study of the cosmic web and its filamentary structure,

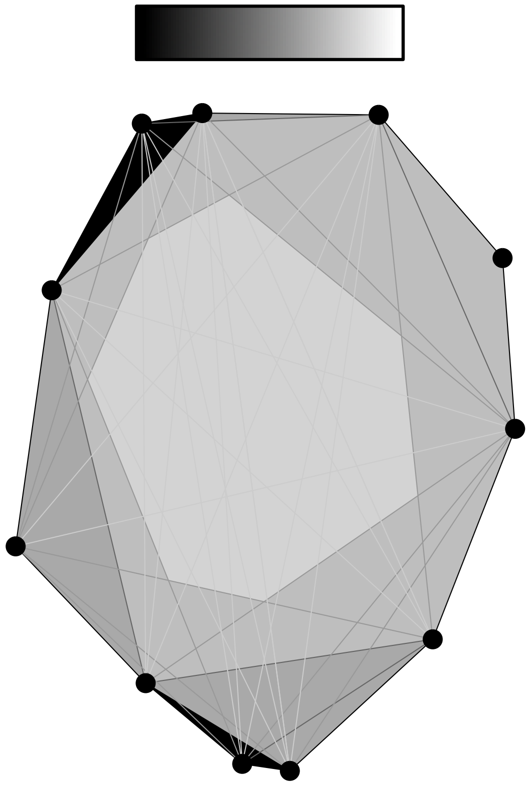
The TDA pipeline in a nutshell



point cloud P

proximity rule

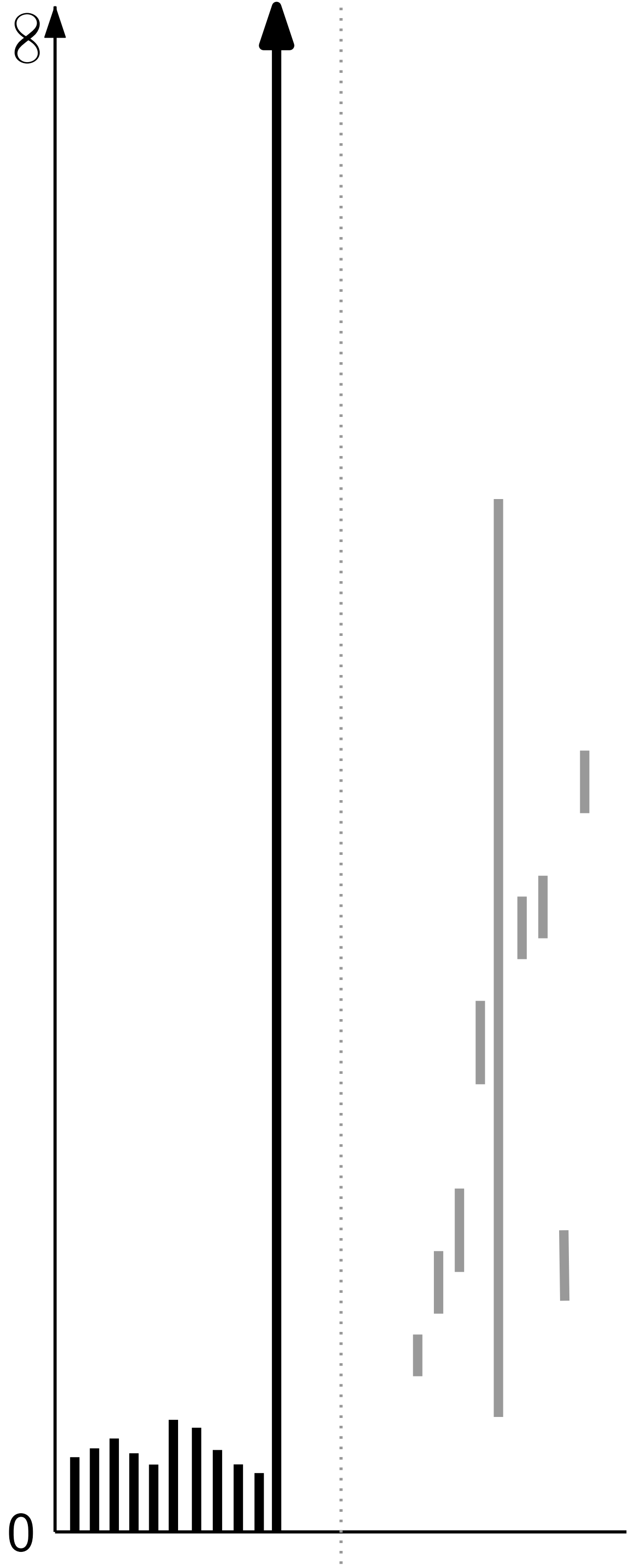
filter: $X \to \mathbb{R}$



filtration $\mathcal{F}$

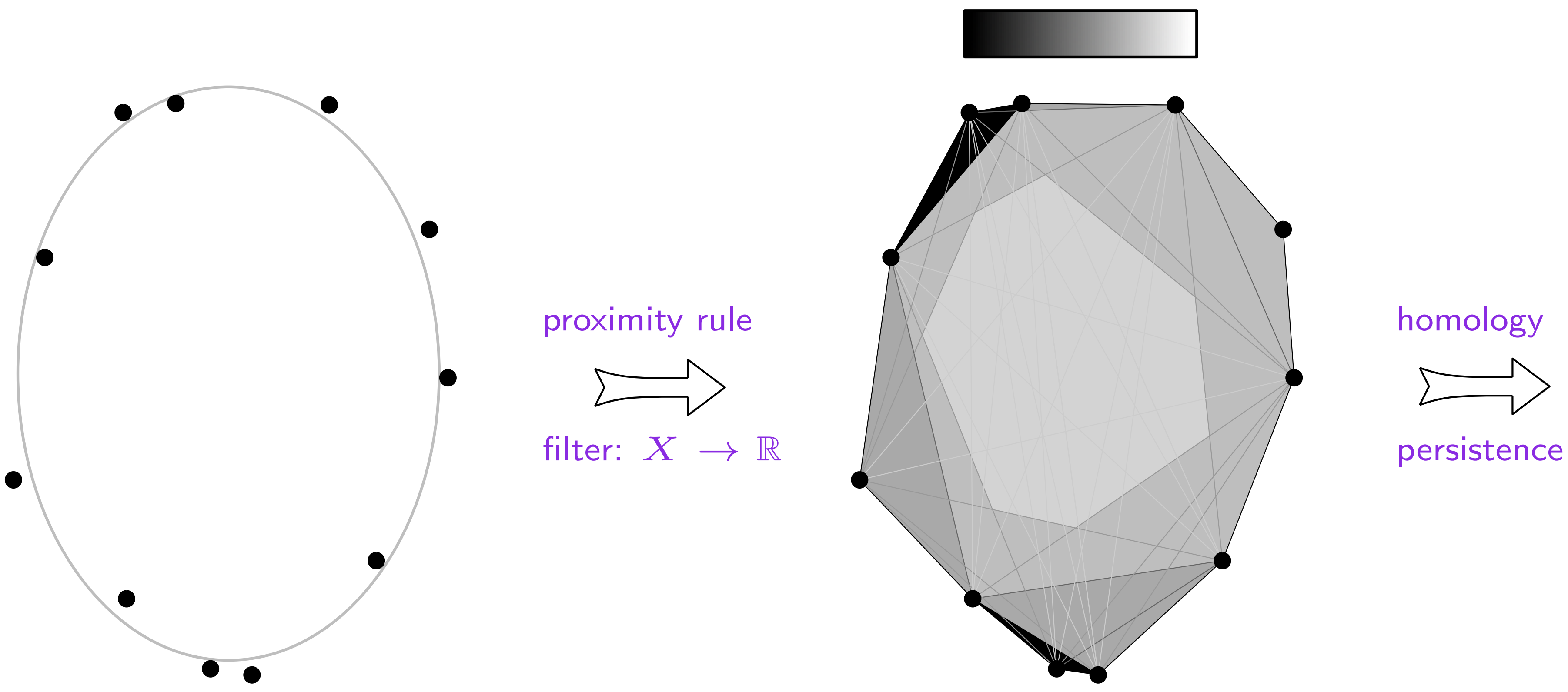
homology

persistence



barcode (descriptor)

The TDA pipeline in a nutshell



proximity rule

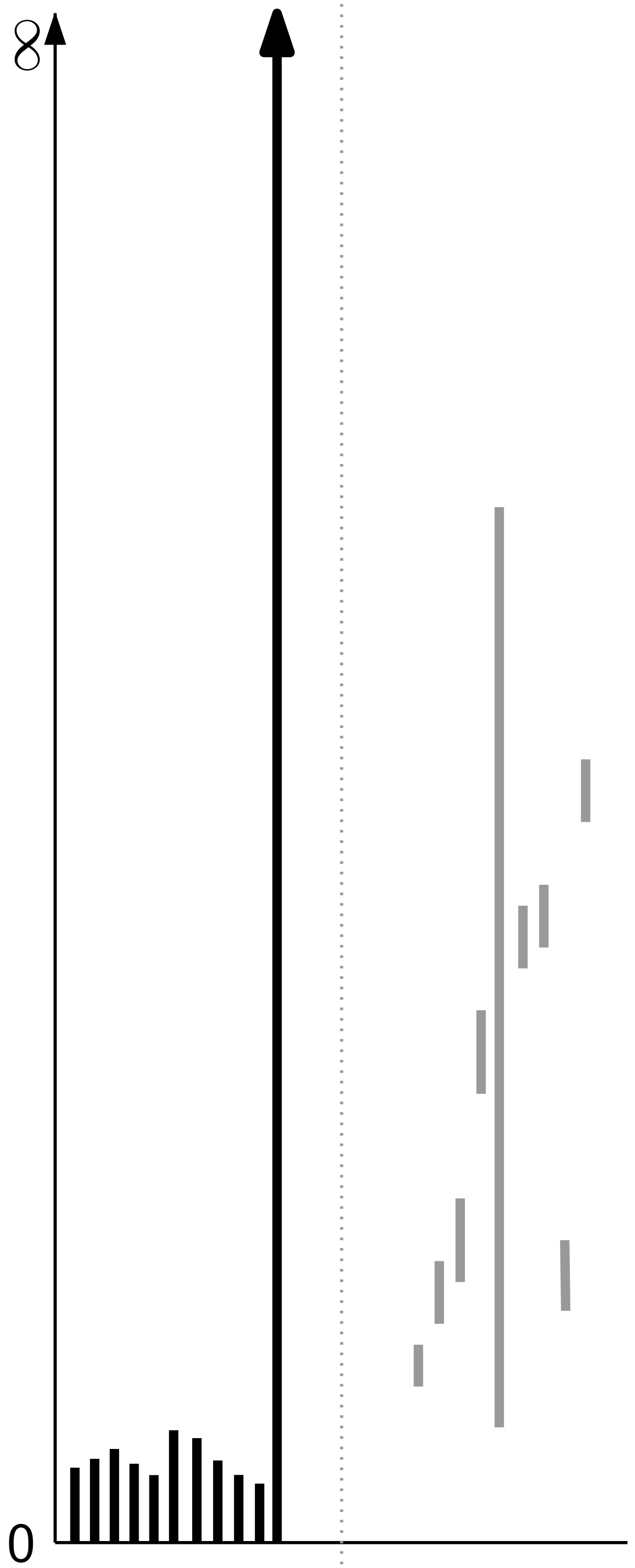
filter: $X \to \mathbb{R}$

homology

persistence

point cloud P

filtration $\mathcal{F}$

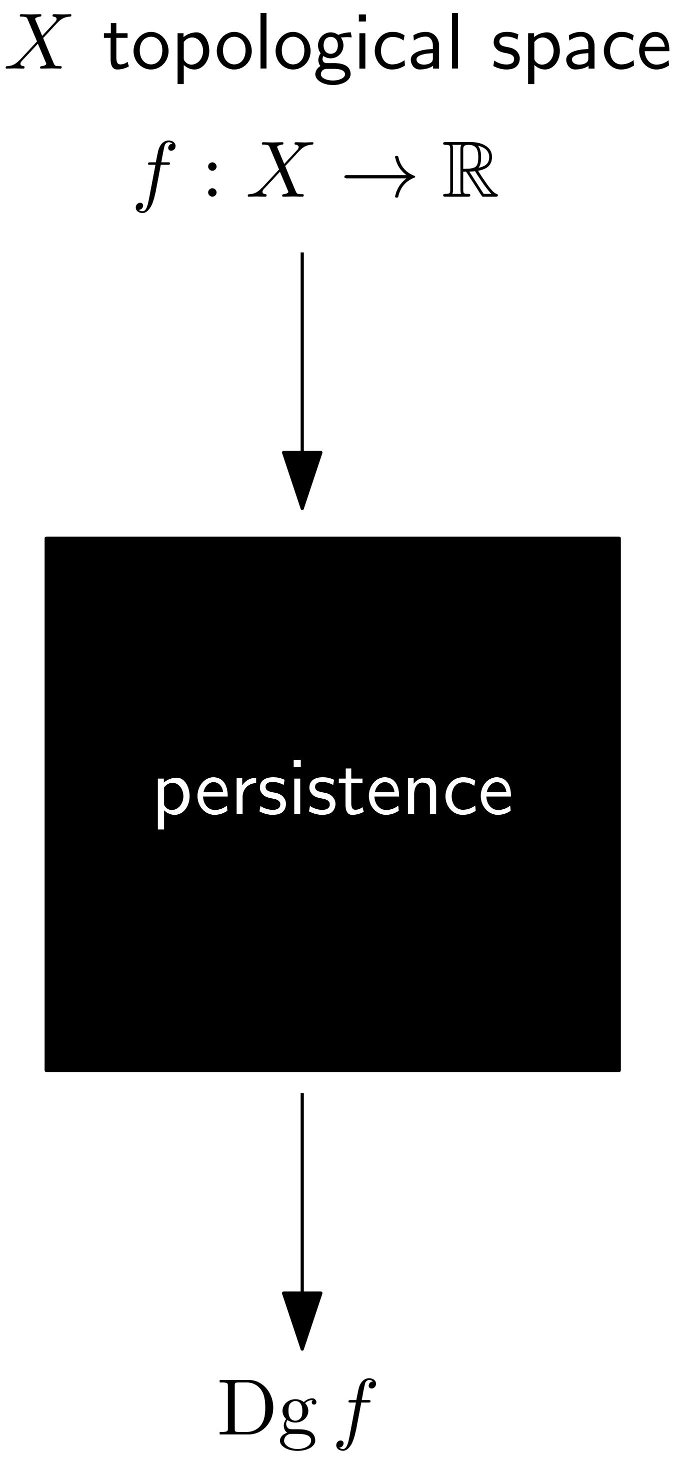


barcode (descriptor)

The 5 pillars of the theory (persistence theory):

- decomposition theorems (existence of barcodes)
- algorithms (computation of barcodes)
- stability theorems (barcodes as stable descriptors)
- statistical frameworks for barcodes
- vectorizations and kernels on barcodes for learning

What are barcodes?

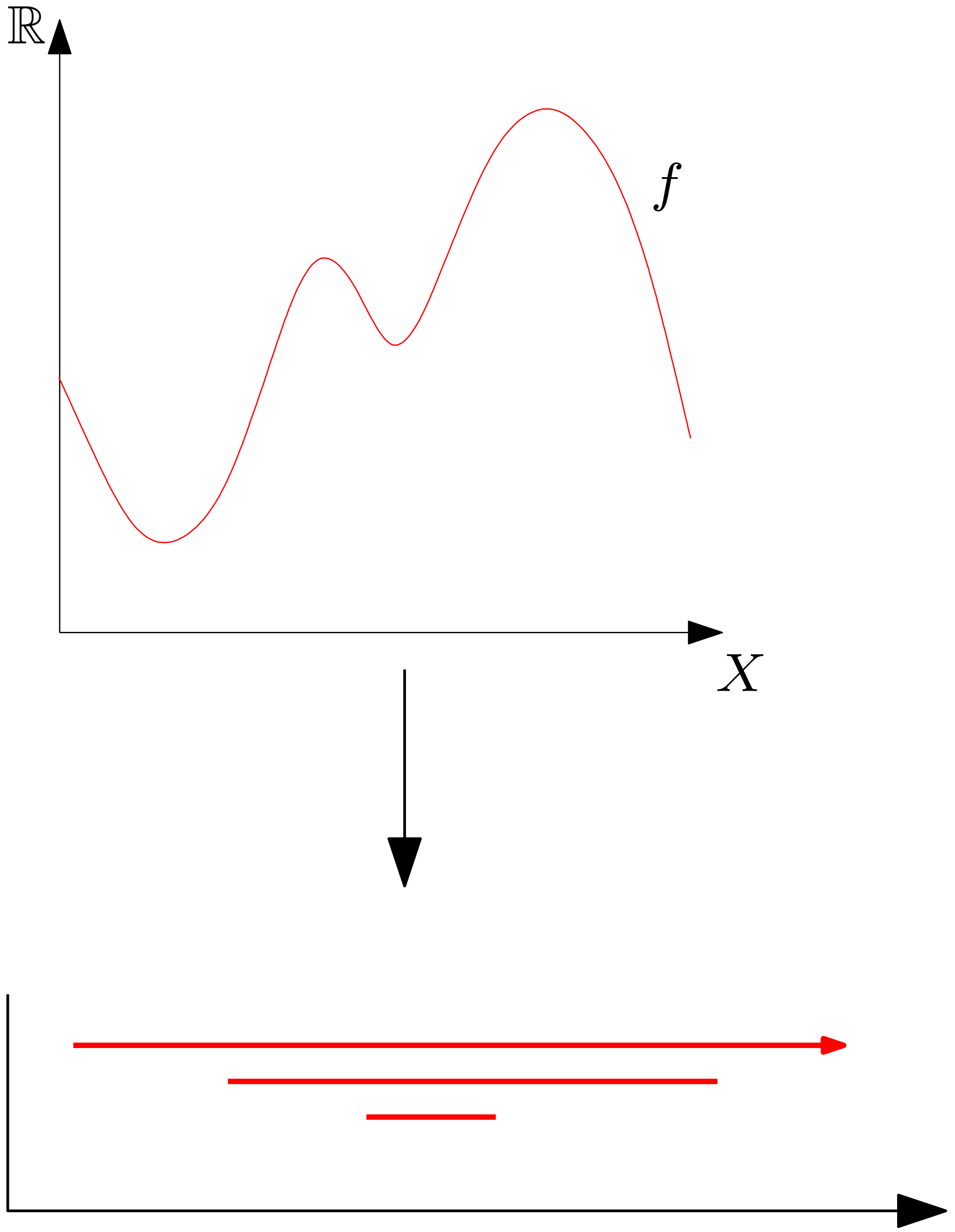


X topological space

$f : X \to \mathbb{R}$

persistence

$\mathrm{Dg}\, f$

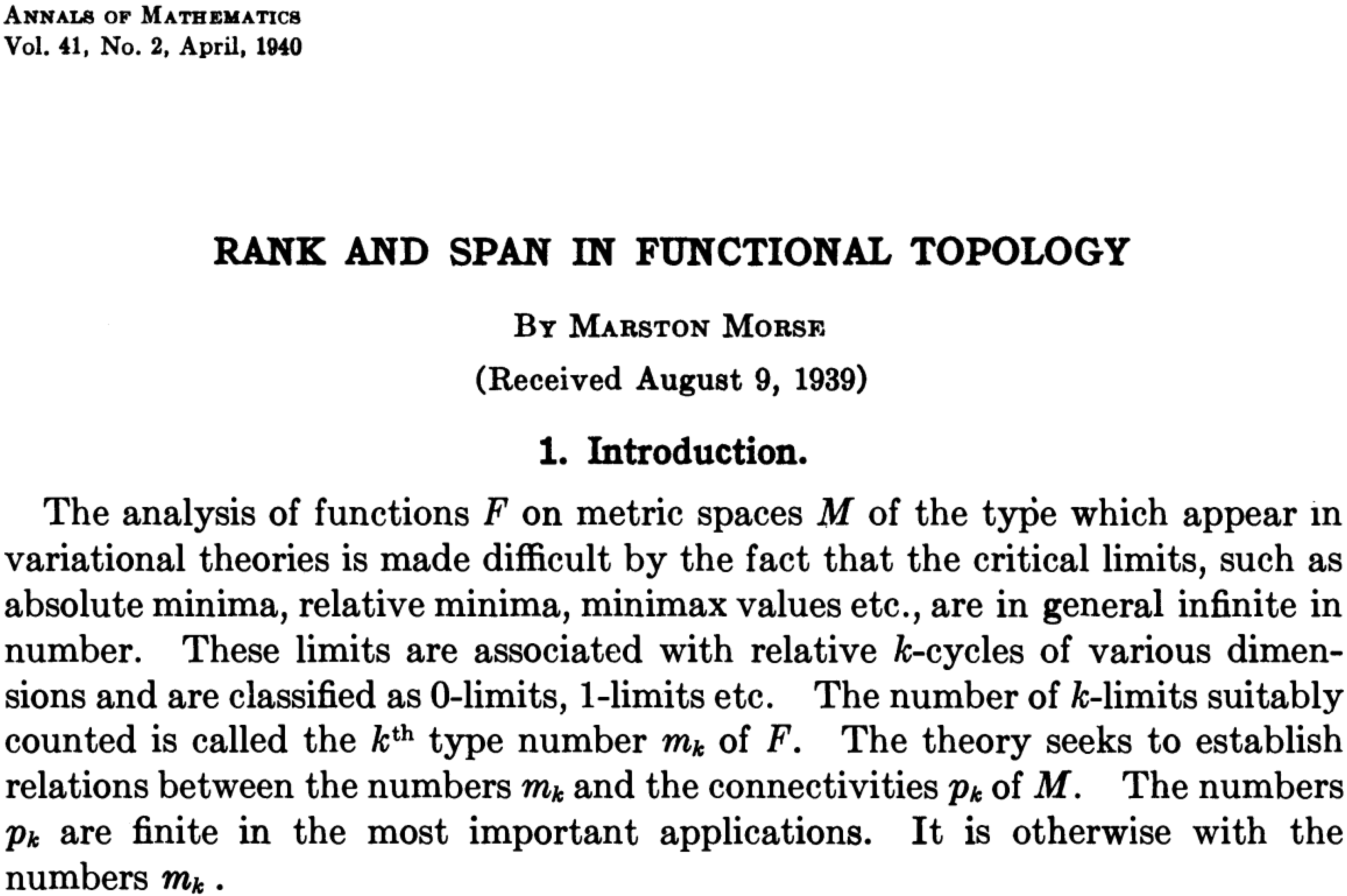


descriptor: *persistence barcode* / *diagram*

encodes the topological structure of the pair (X, f)

Inside the black box

Generalized Morse theory:



ANNALS OF MATHEMATICS
Vol. 41, No. 2, April, 1940

RANK AND SPAN IN FUNCTIONAL TOPOLOGY

BY MARSTON MORSE

(Received August 9, 1939)

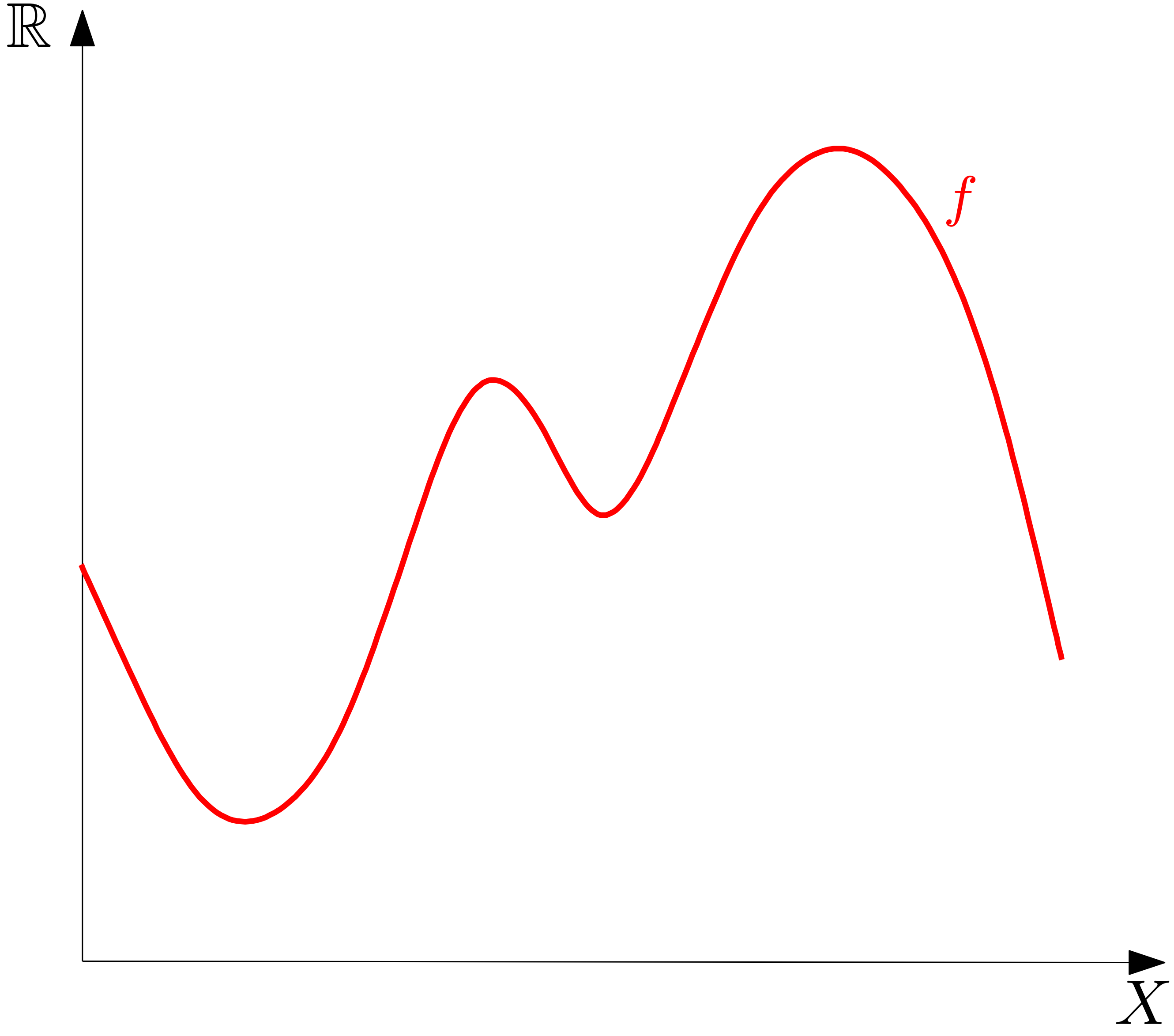
1. Introduction.

The analysis of functions F on metric spaces M of the type which appear in variational theories is made difficult by the fact that the critical limits, such as absolute minima, relative minima, minimax values etc., are in general infinite in number. These limits are associated with relative k-cycles of various dimensions and are classified as 0-limits, 1-limits etc. The number of k-limits suitably counted is called the k^{th} type number m_k of F. The theory seeks to establish relations between the numbers m_k and the connectivities p_k of M. The numbers p_k are finite in the most important applications. It is otherwise with the numbers m_k .

Inside the black box

Generalized Morse theory:

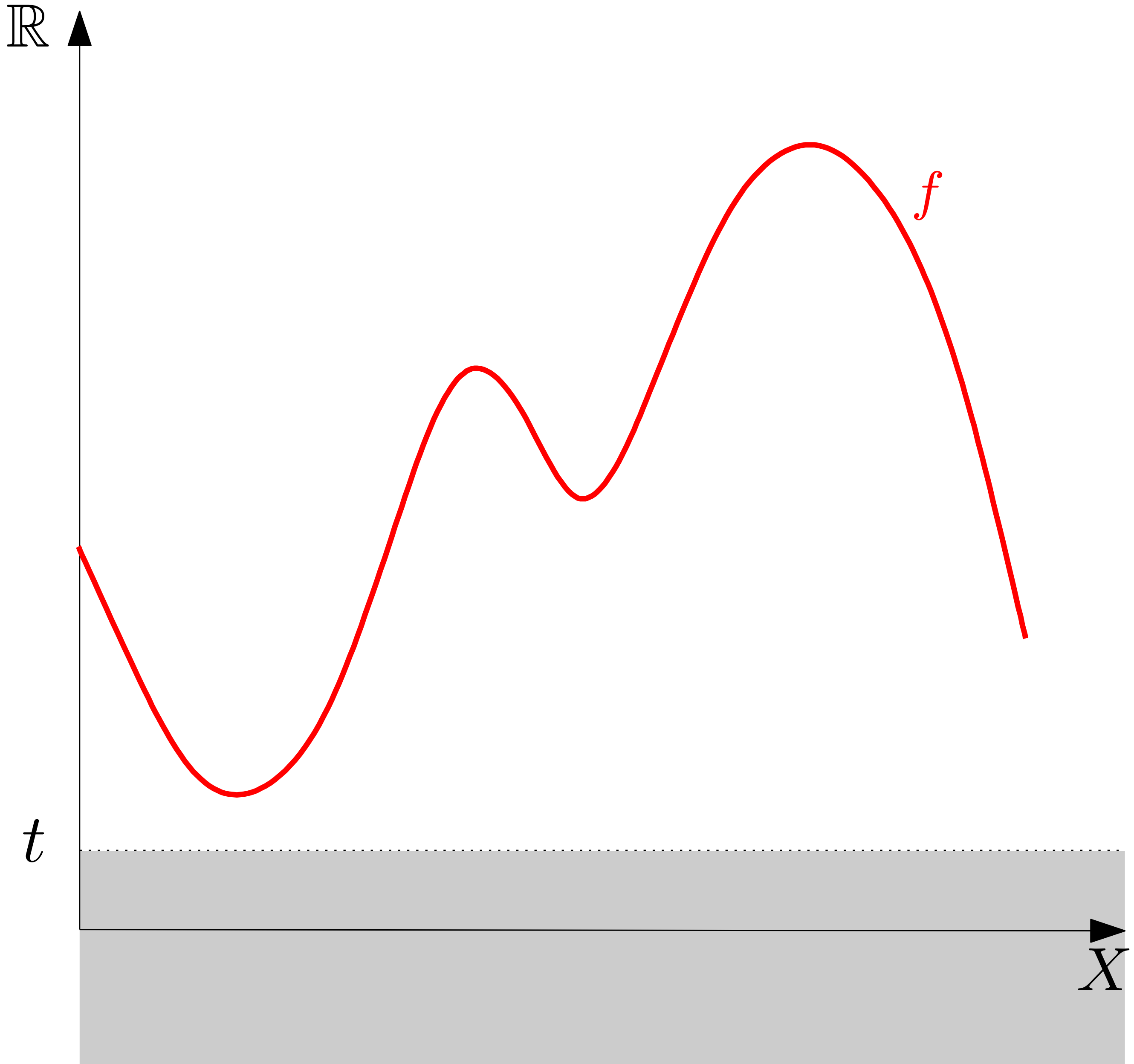
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

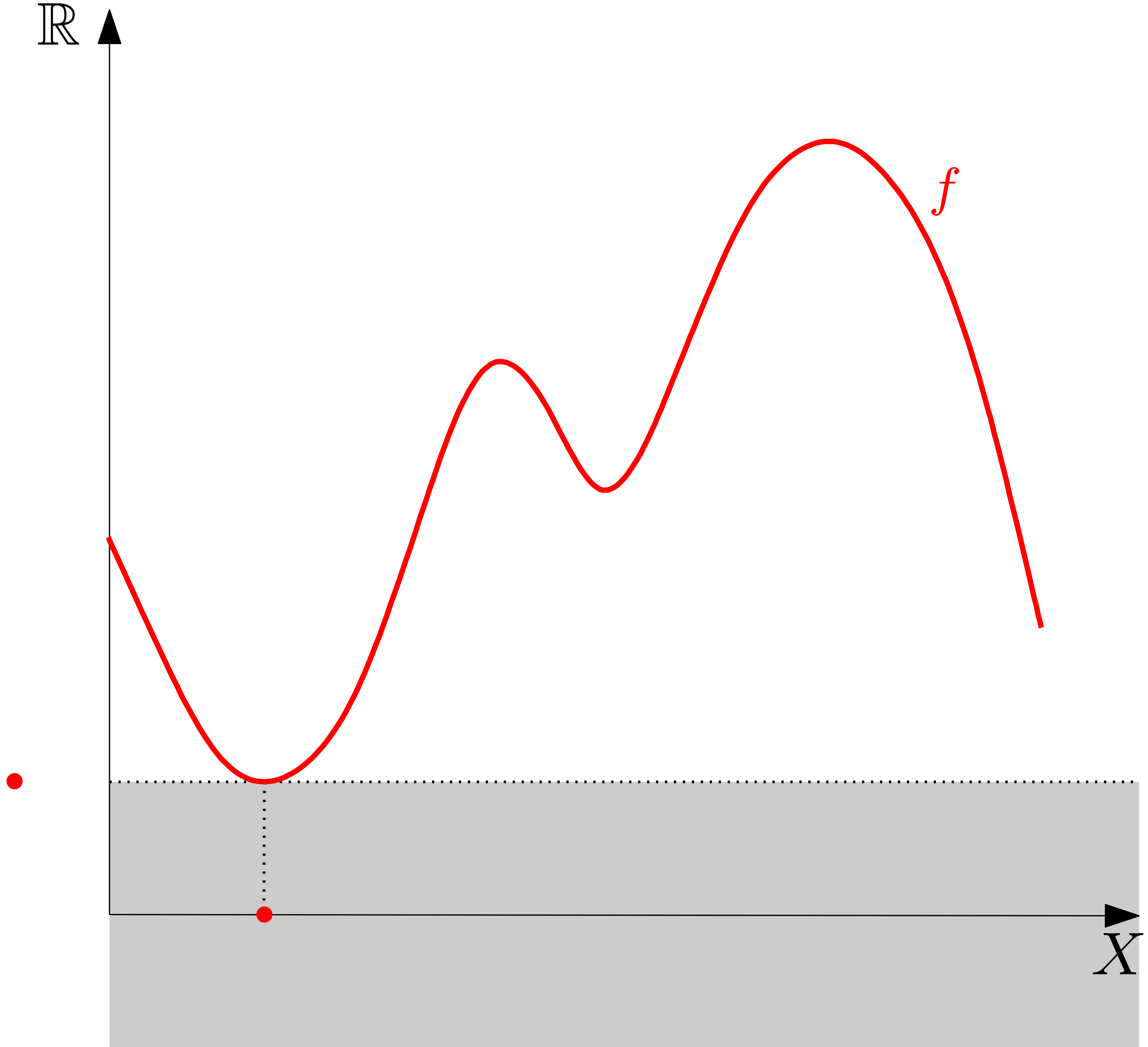
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

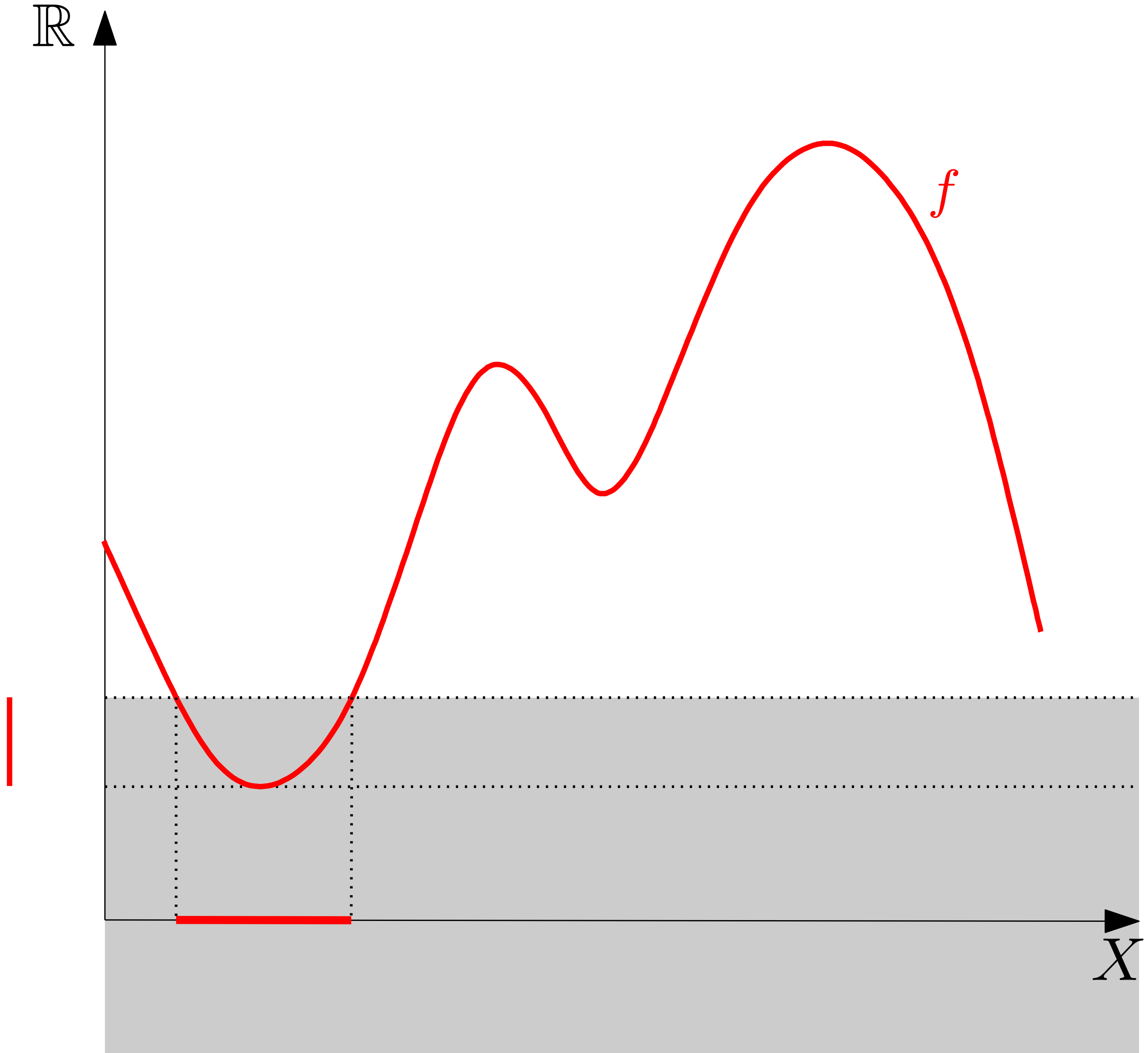
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

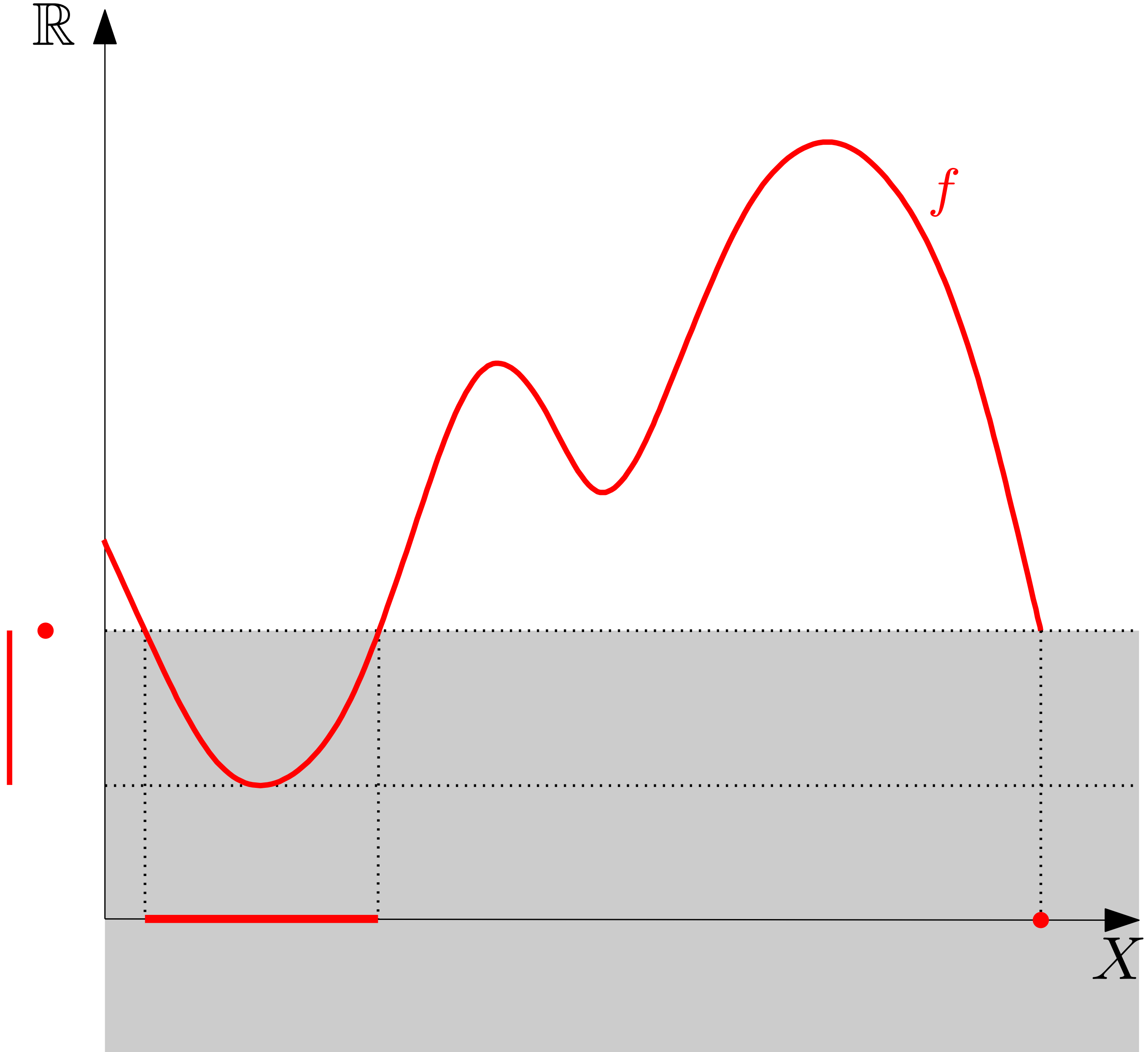
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

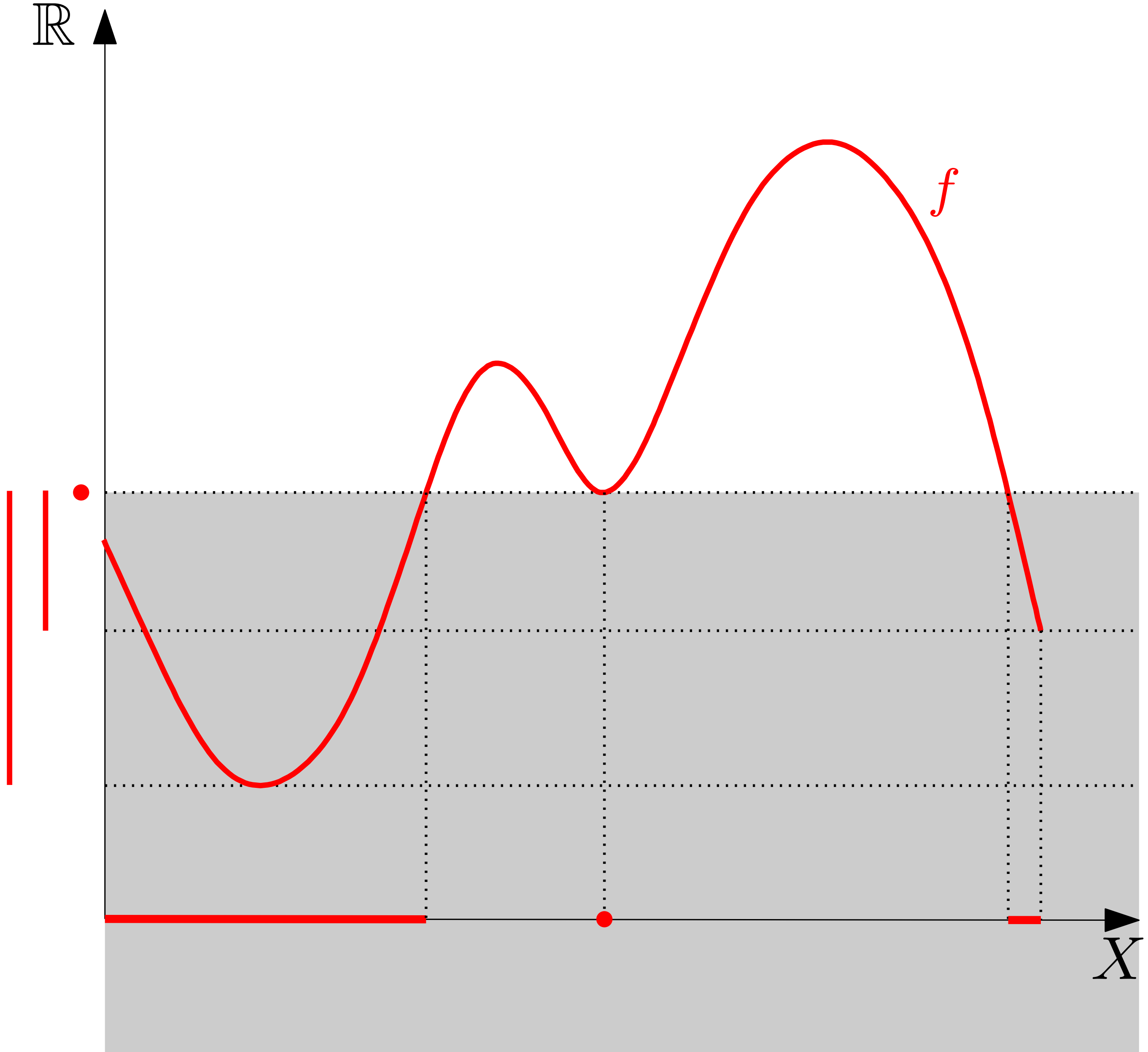
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

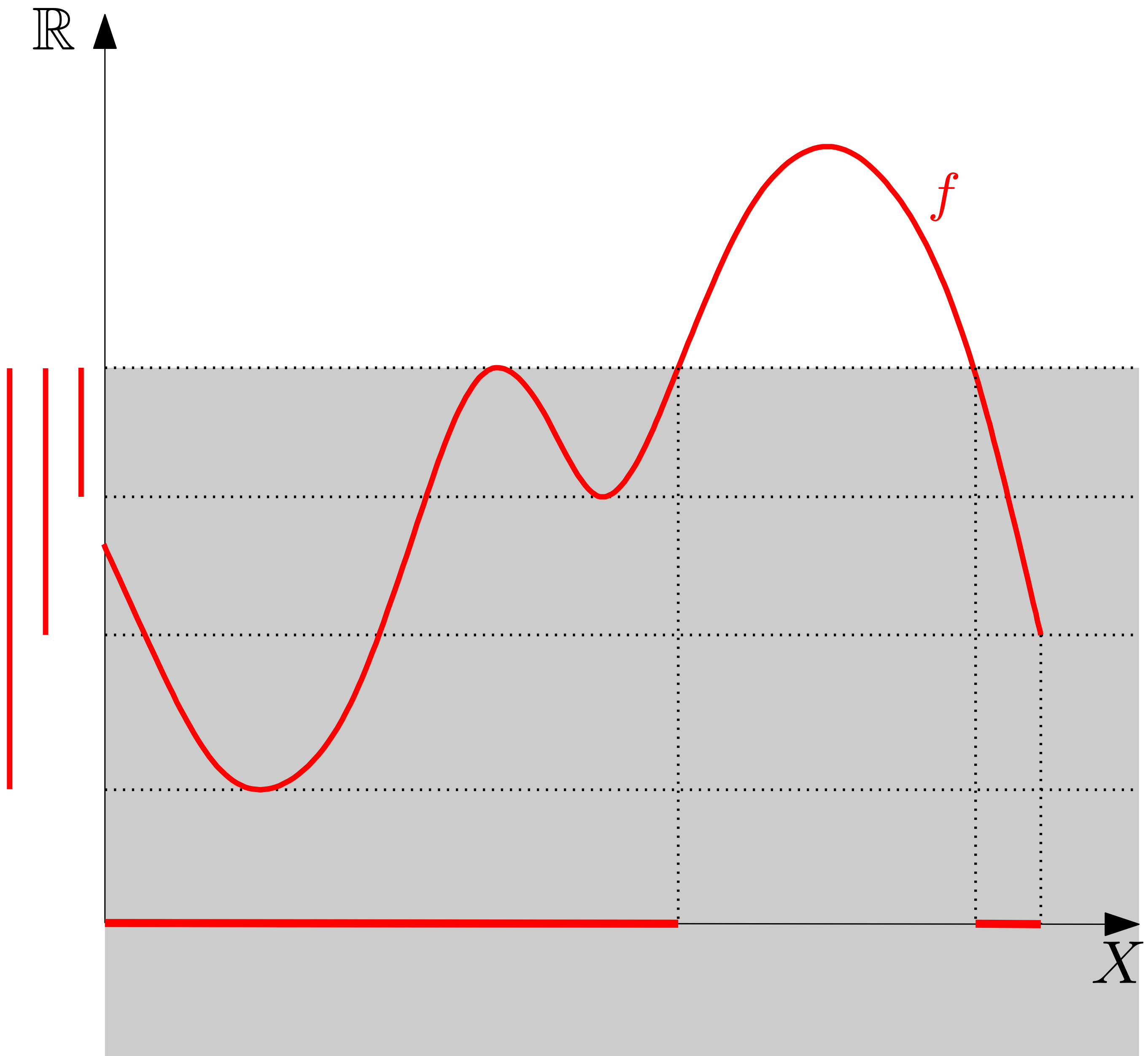
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

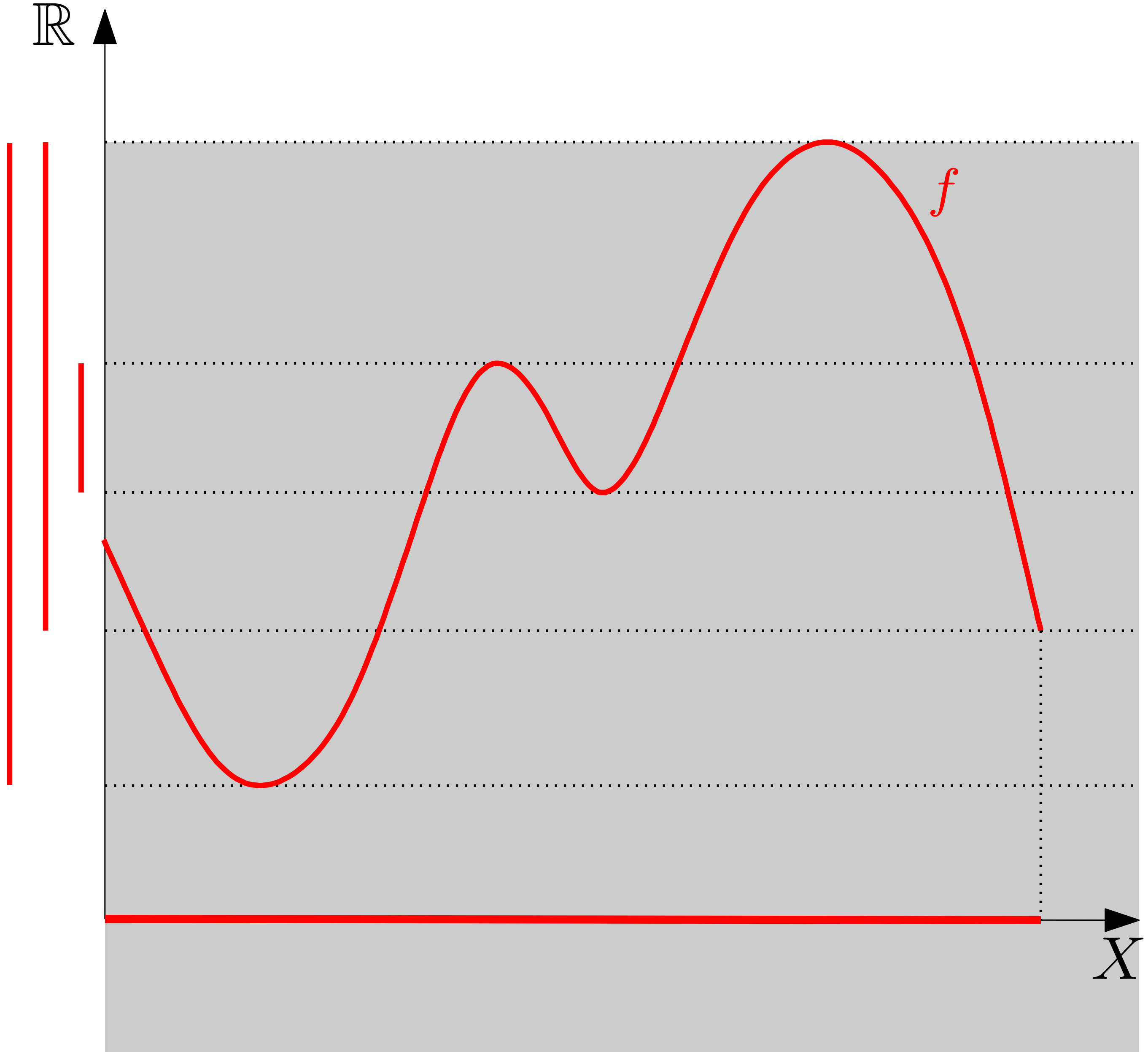
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

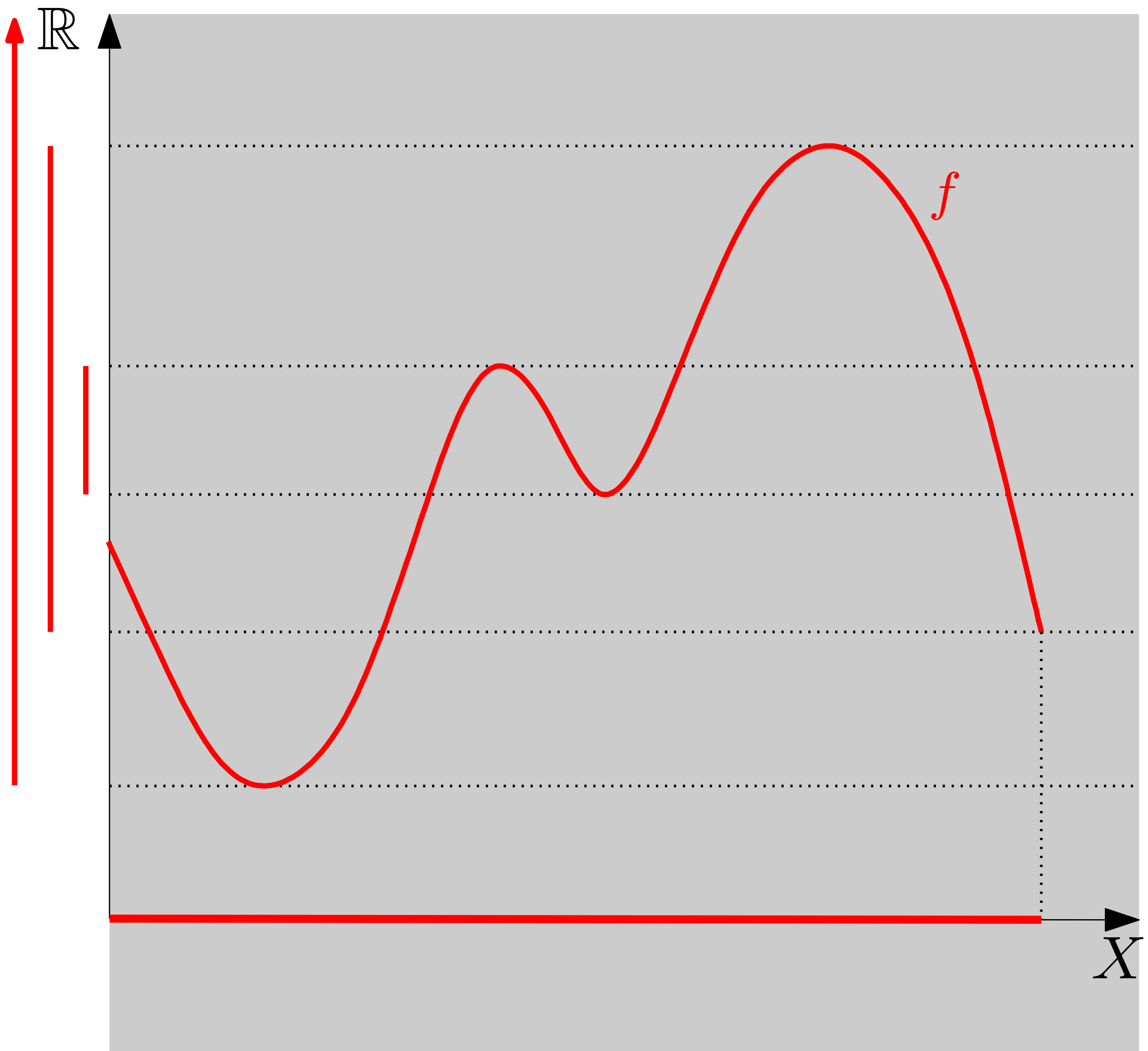
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family



Inside the black box

Generalized Morse theory:

- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family
- Finite set of intervals (barcode) encodes births/deaths of *topological features*

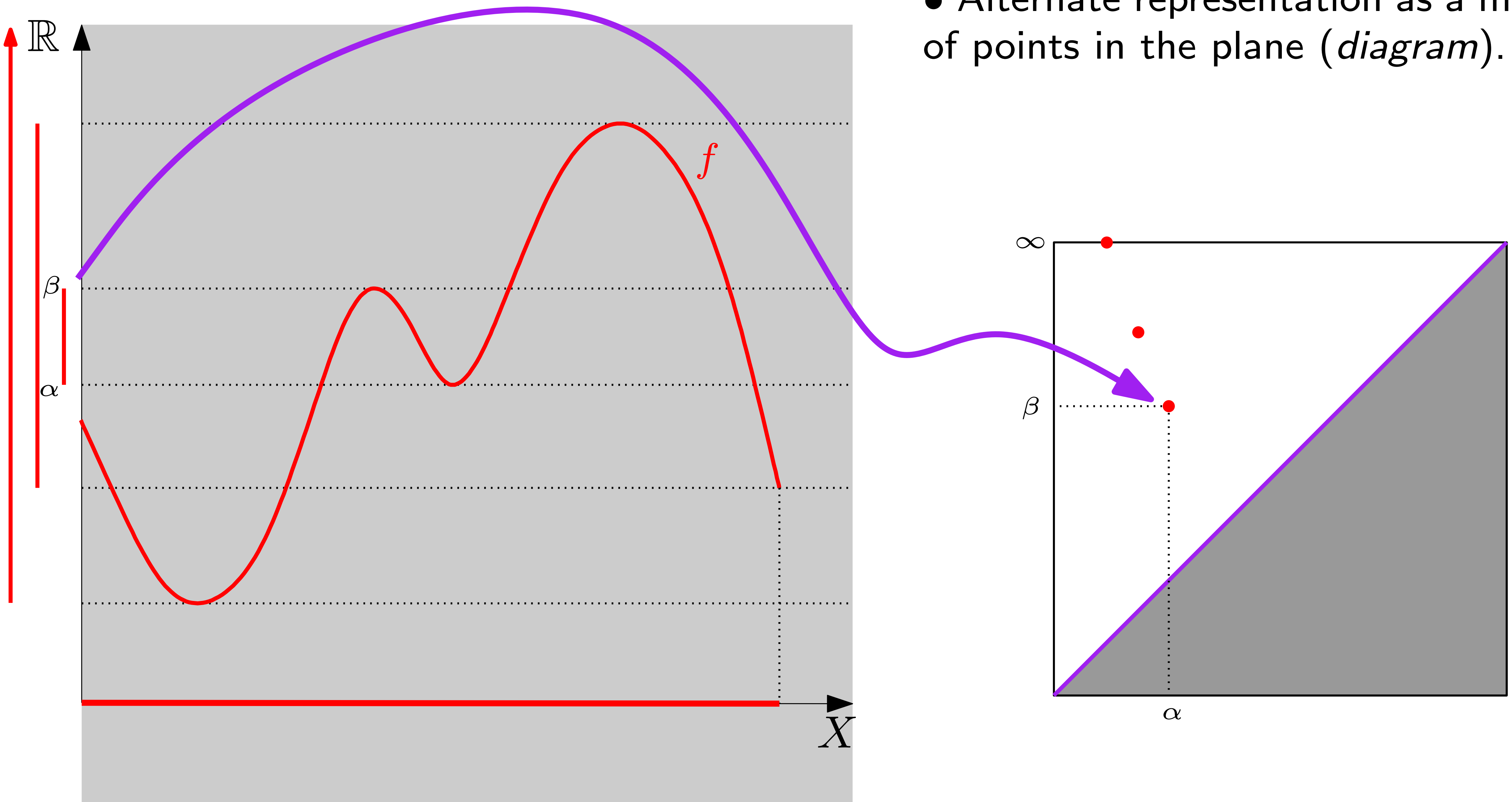


Inside the black box

Generalized Morse theory:

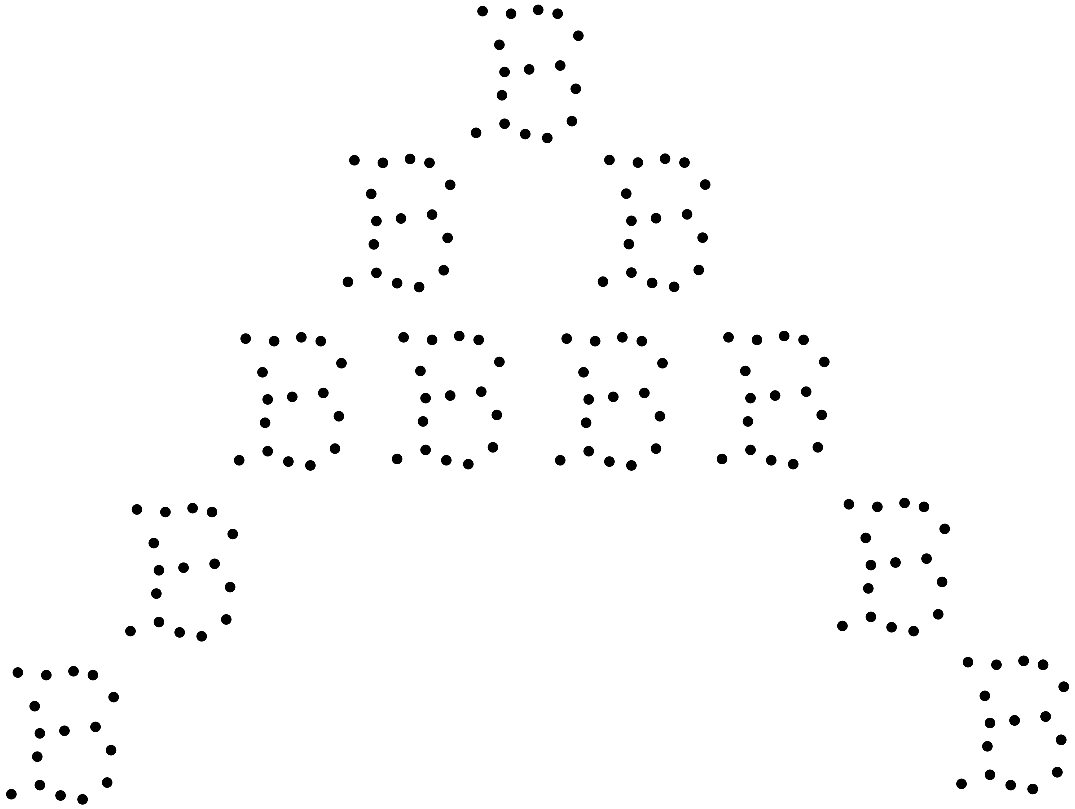
- Nested family (*filtration*) of sublevel-sets $F_t = f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$
- Track the evolution of the topology throughout the family
- Finite set of intervals (barcode) encodes births/deaths of *topological features*

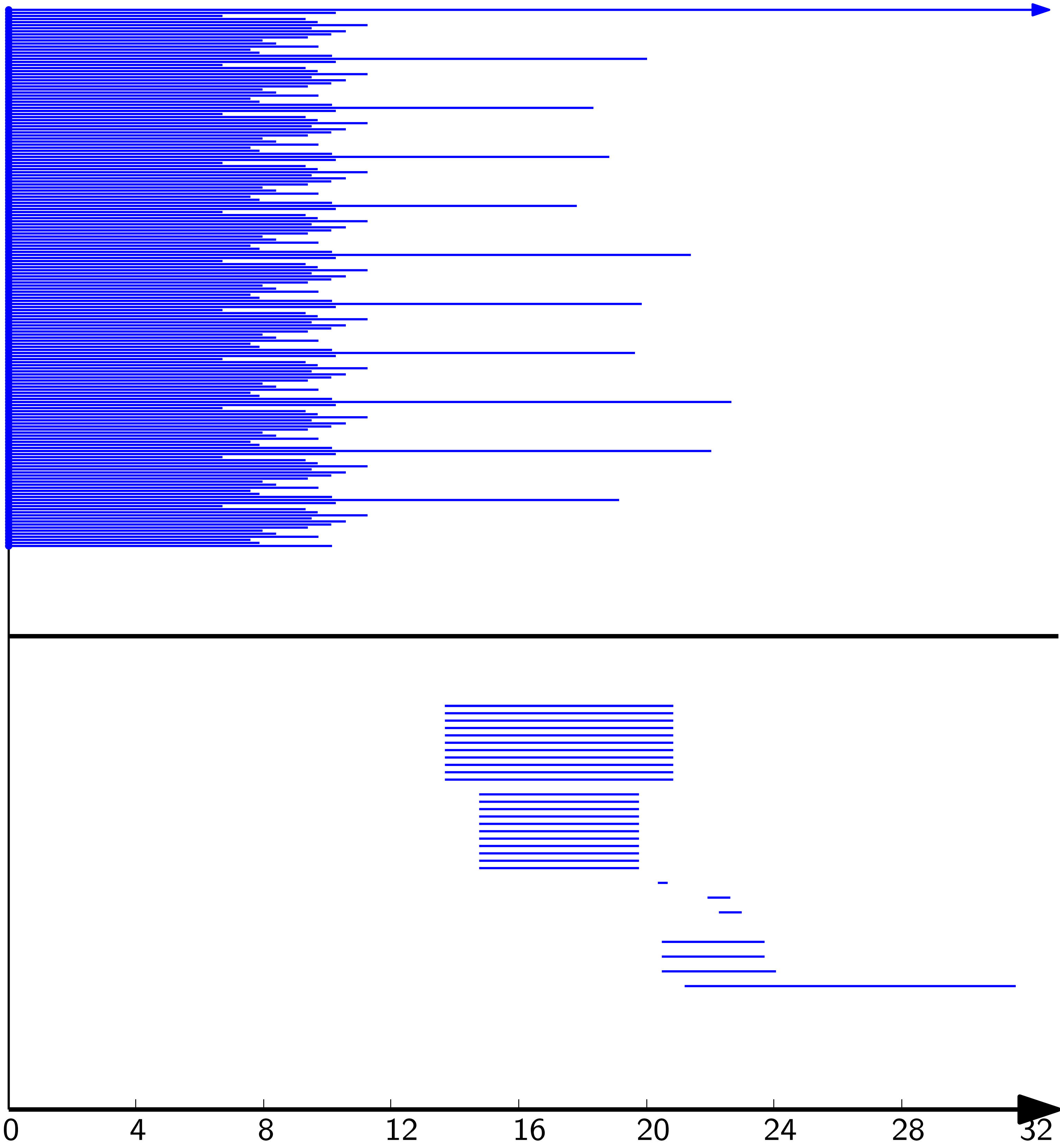
- Alternate representation as a multiset of points in the plane (*diagram*).



Example: distance function

$$f_P: \quad \mathbb{R}^2 \to \mathbb{R}$$
$$x \mapsto \min_{p\in P} \|x-p\|_2$$



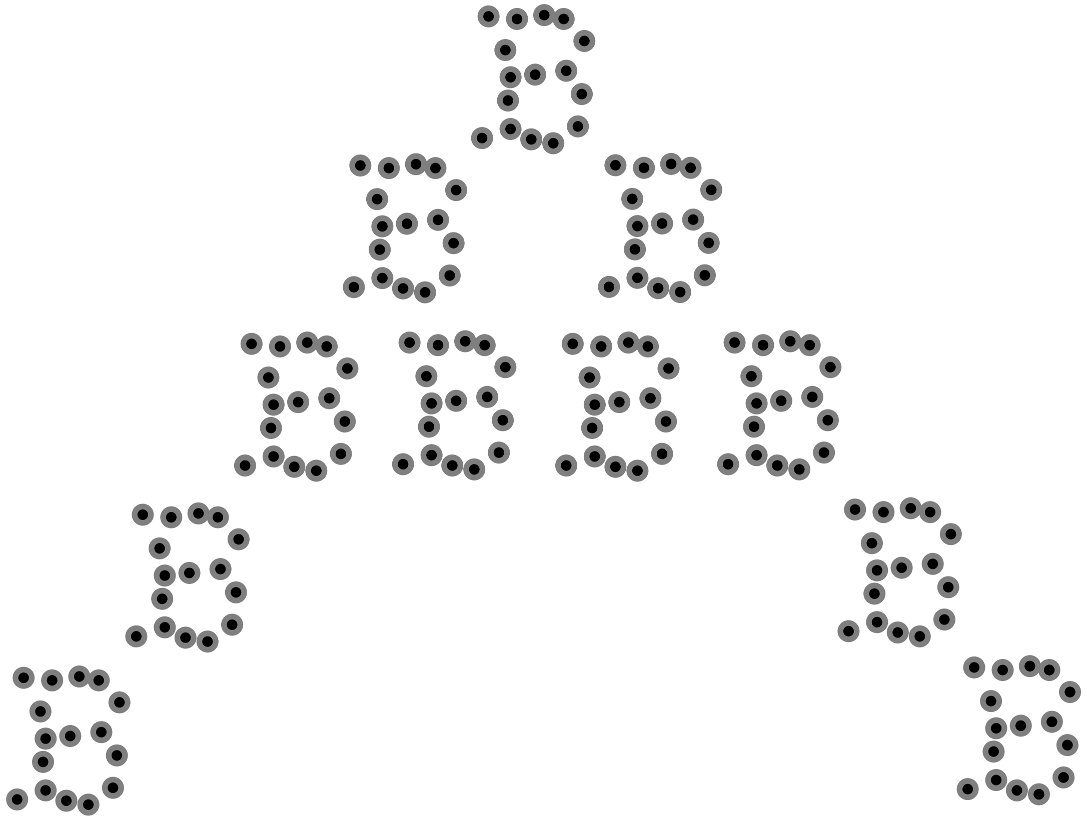


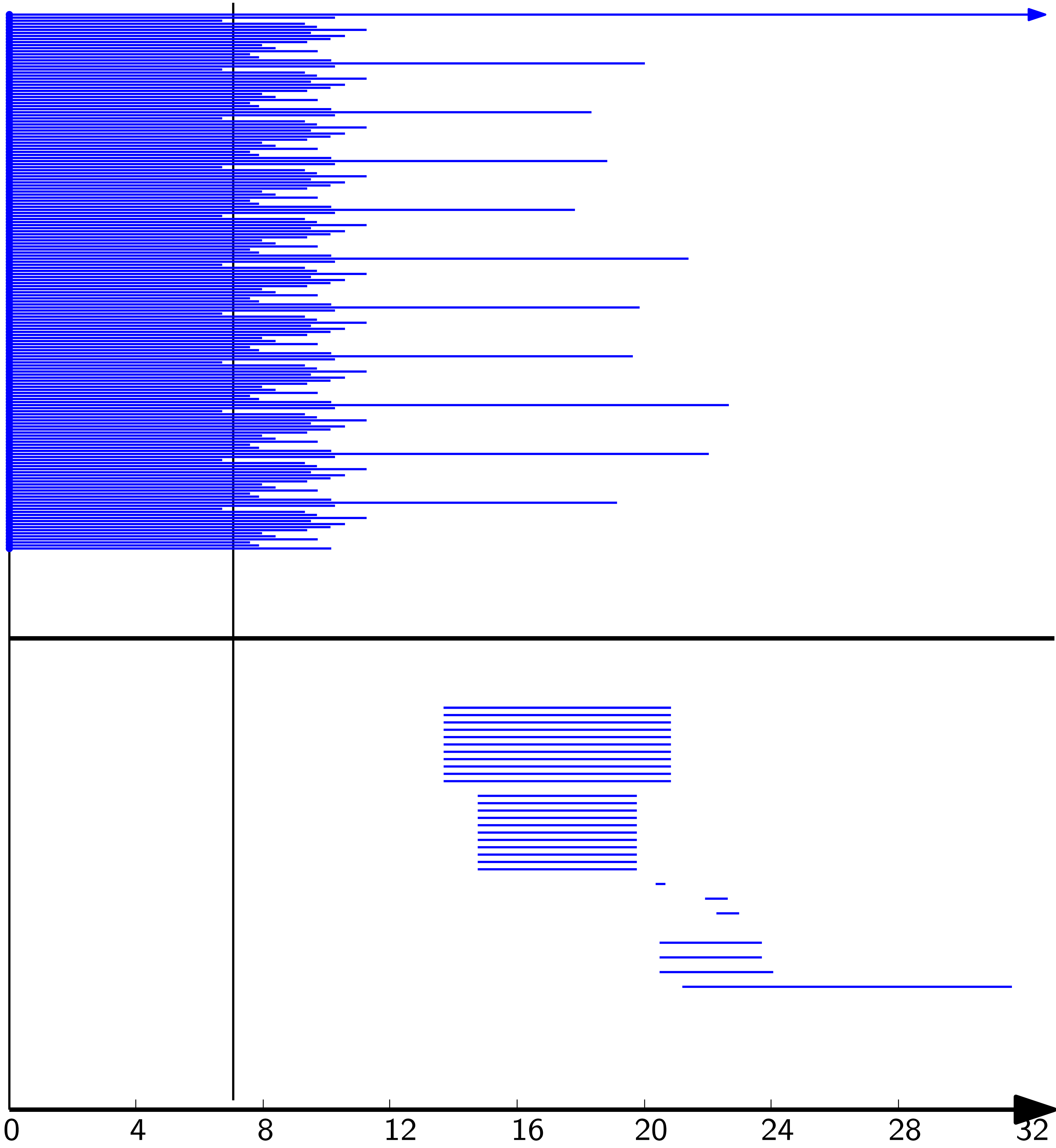
topology:

- connected components
- holes
- voids ...

Example: distance function

$$f_P : \quad \mathbb{R}^2 \to \mathbb{R}$$
$$x \mapsto \min_{p \in P} \|x - p\|_2$$



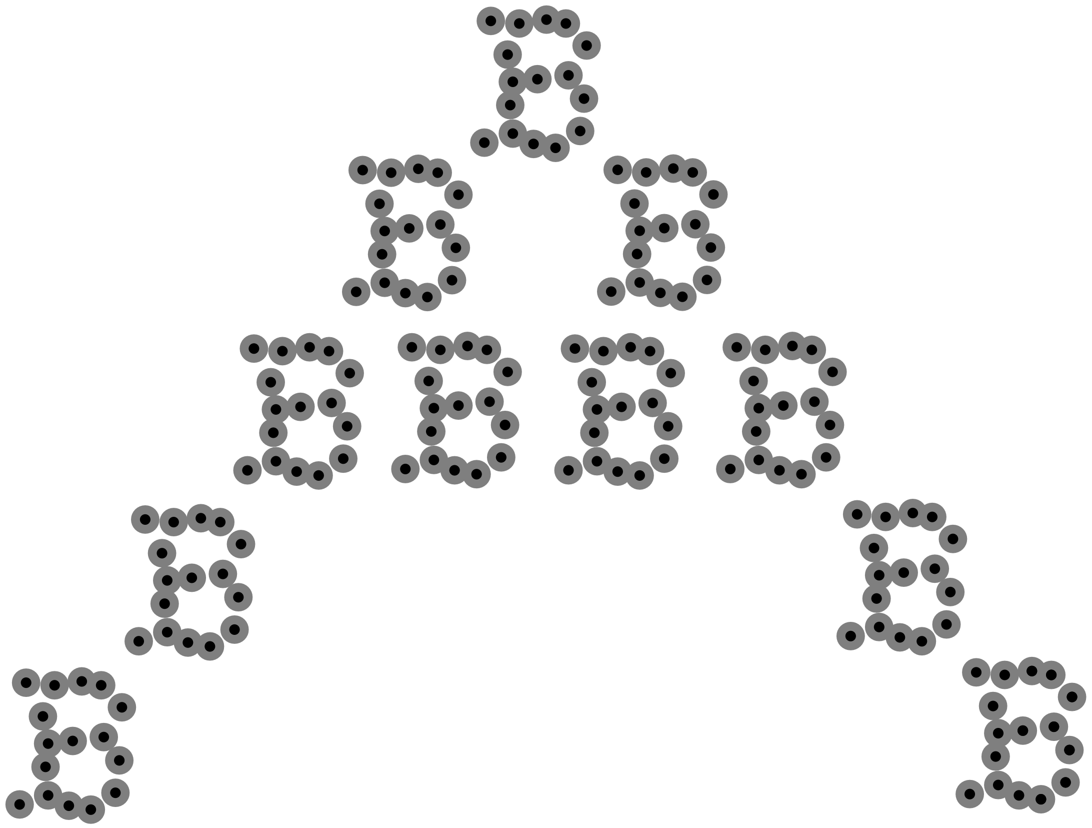


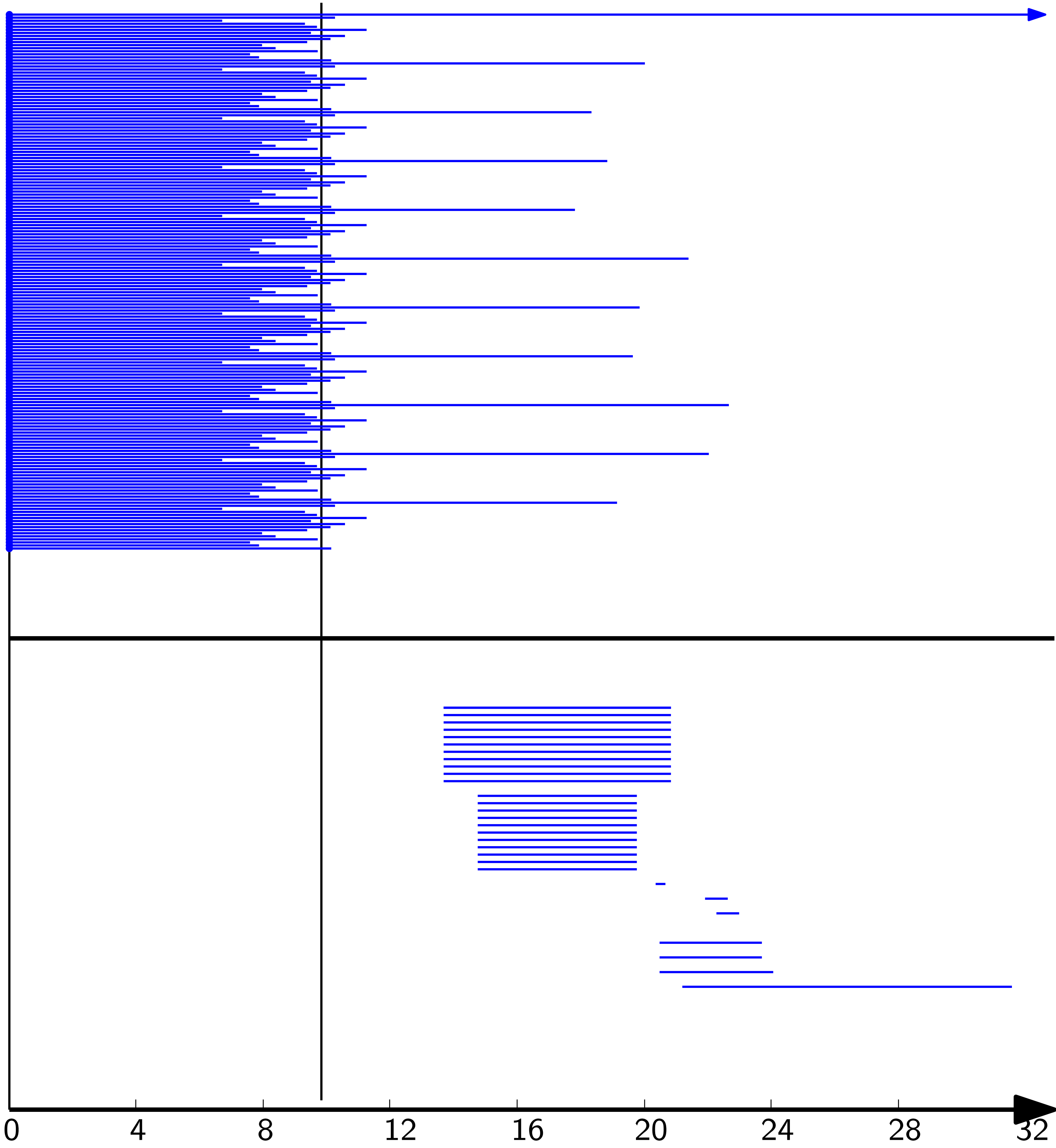
topology:

- connected components
- holes
- voids ...

Example: distance function

$$\begin{aligned} f_P : \quad & \mathbb{R}^2 \to \mathbb{R} \\ & x \mapsto \min_{p \in P} \|x - p\|_2 \end{aligned}$$



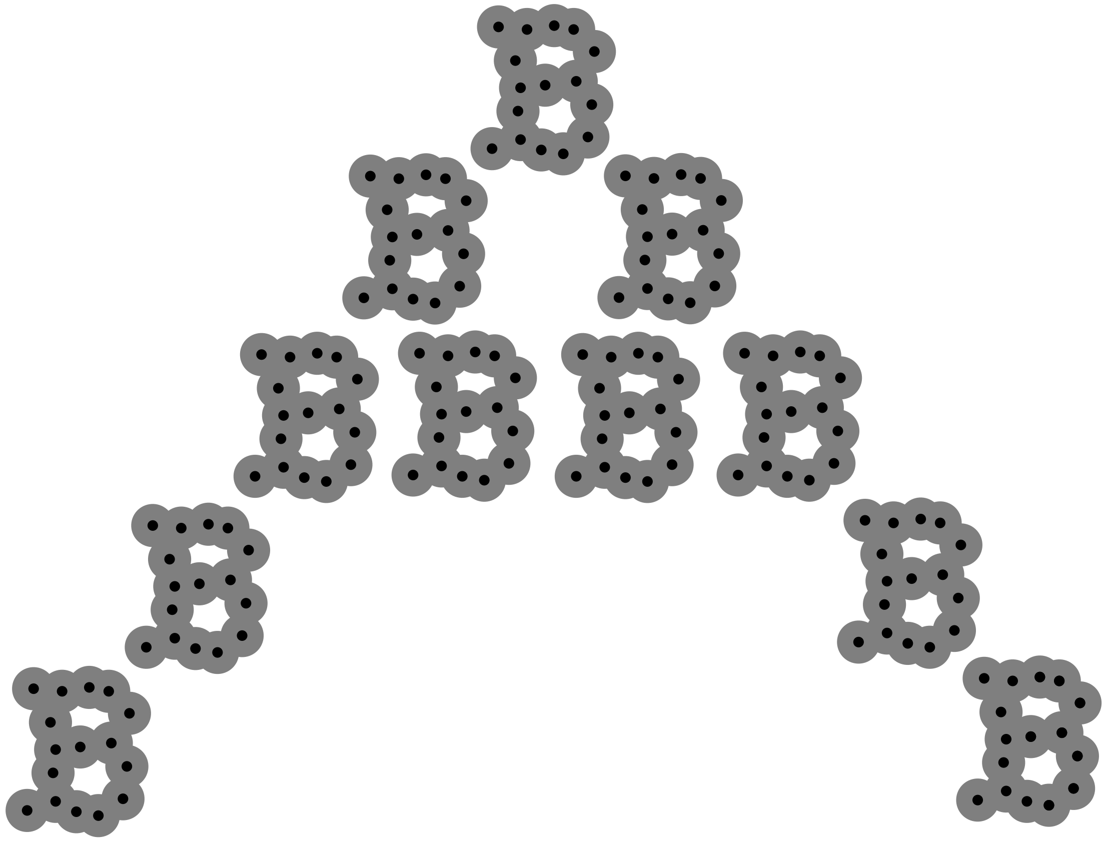


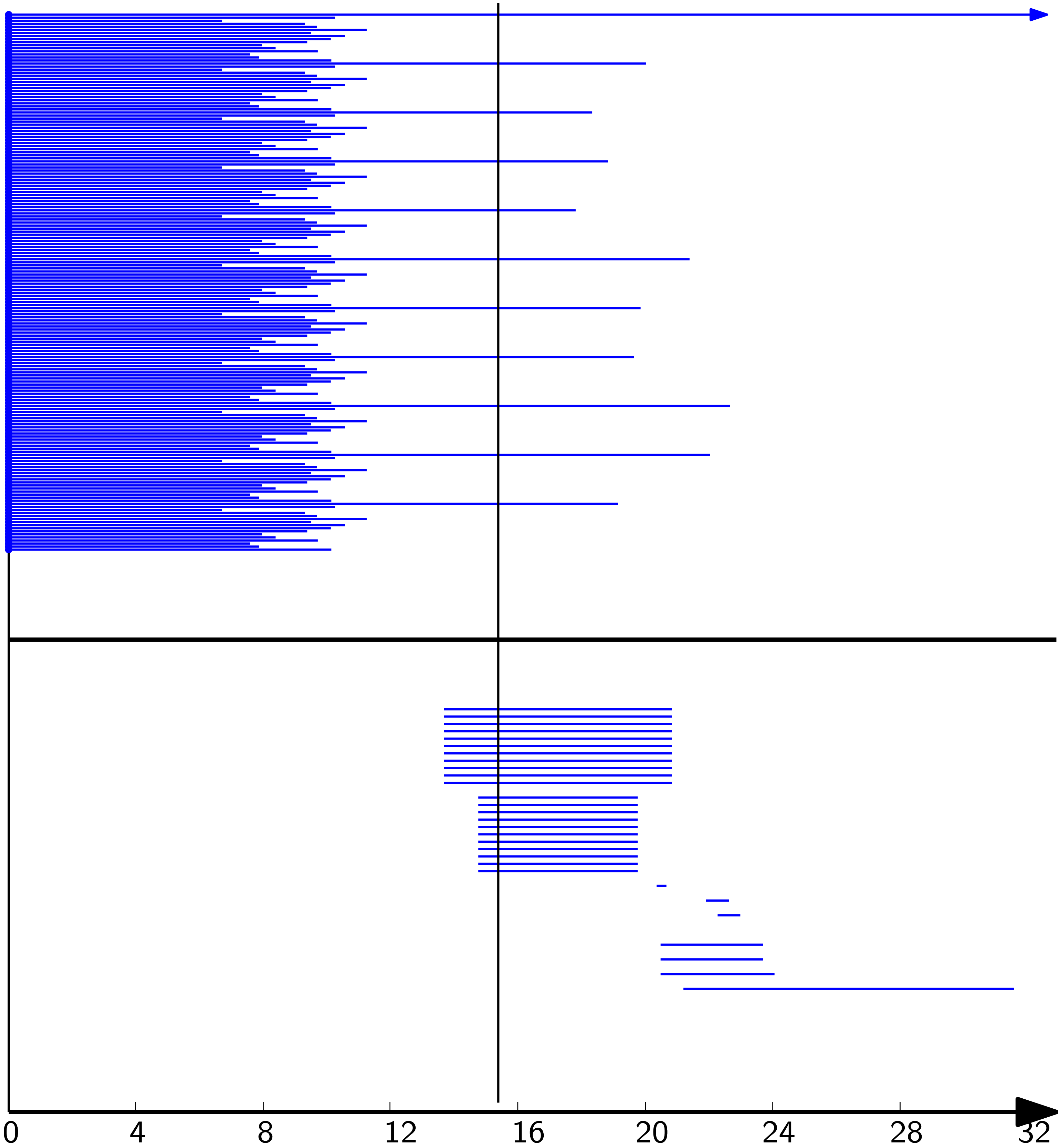
topology:

- connected components
- holes
- voids ...

Example: distance function

$$f_P : \quad \mathbb{R}^2 \to \mathbb{R}$$
$$x \mapsto \min_{p\in P} \|x-p\|_2$$



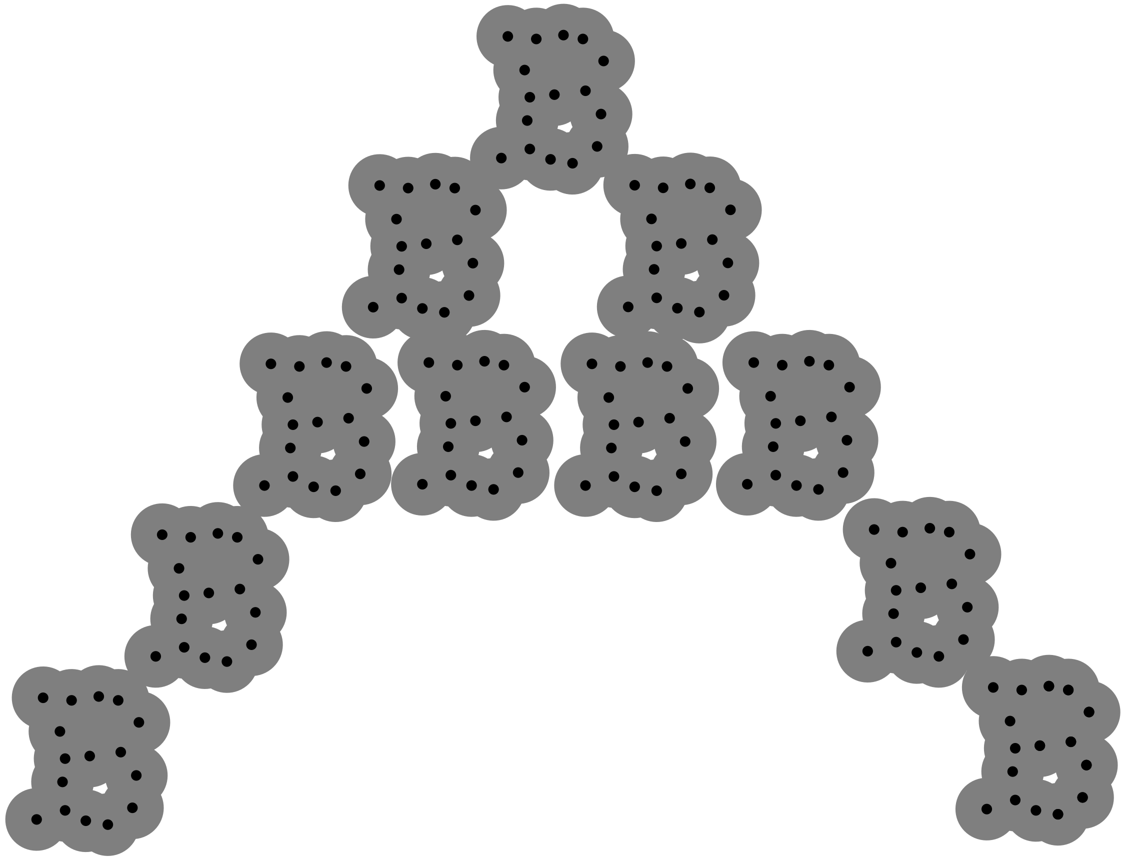


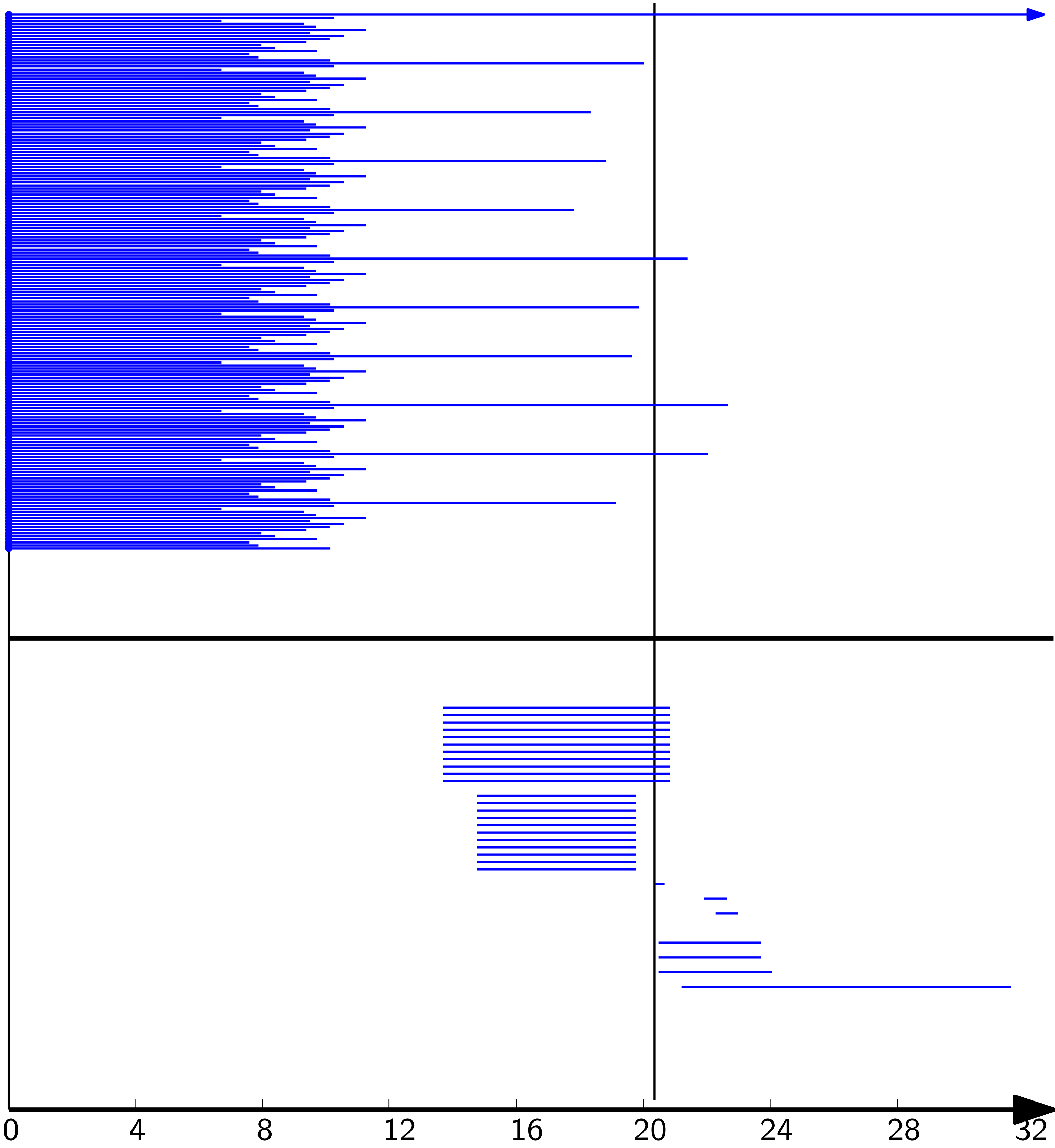
topology:

- connected components
- holes
- voids ...

Example: distance function

$$\begin{aligned} f_P : \quad & \mathbb{R}^2 \to \mathbb{R} \\ & x \mapsto \min_{p \in P} \|x - p\|_2 \end{aligned}$$



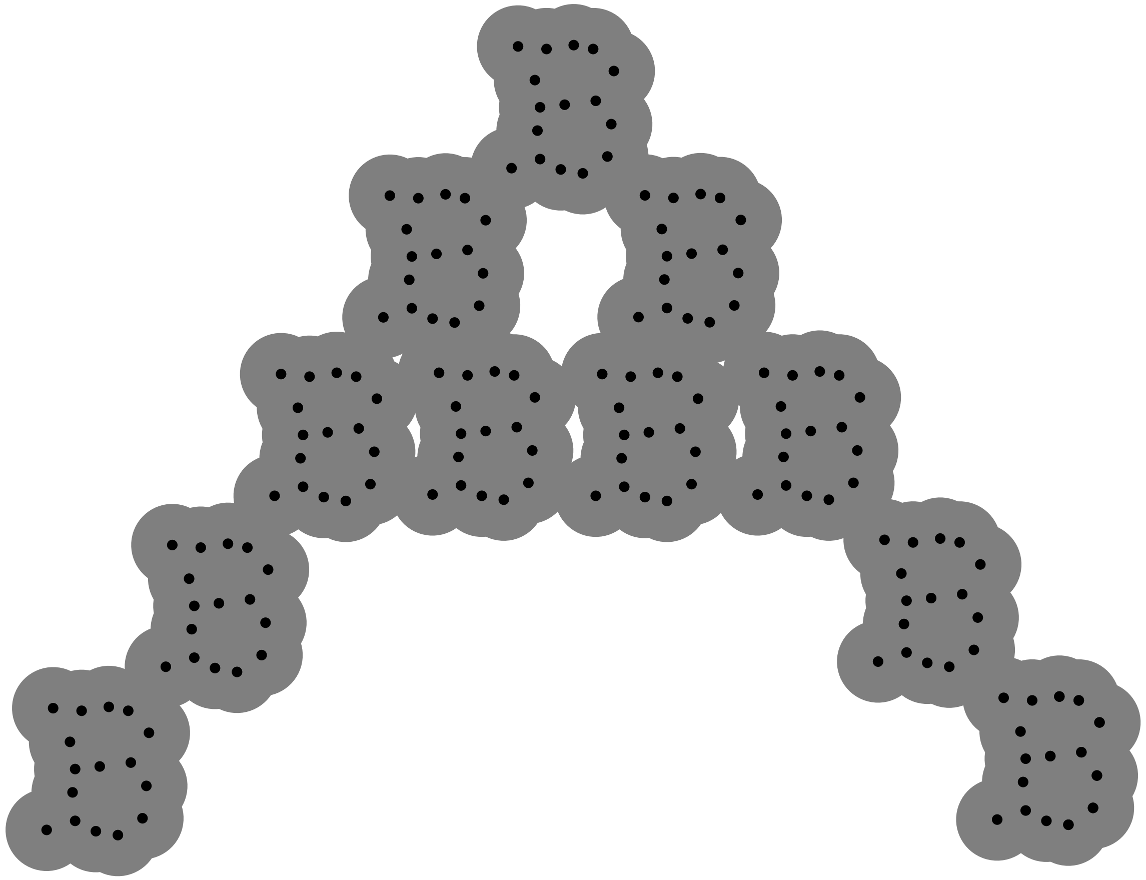


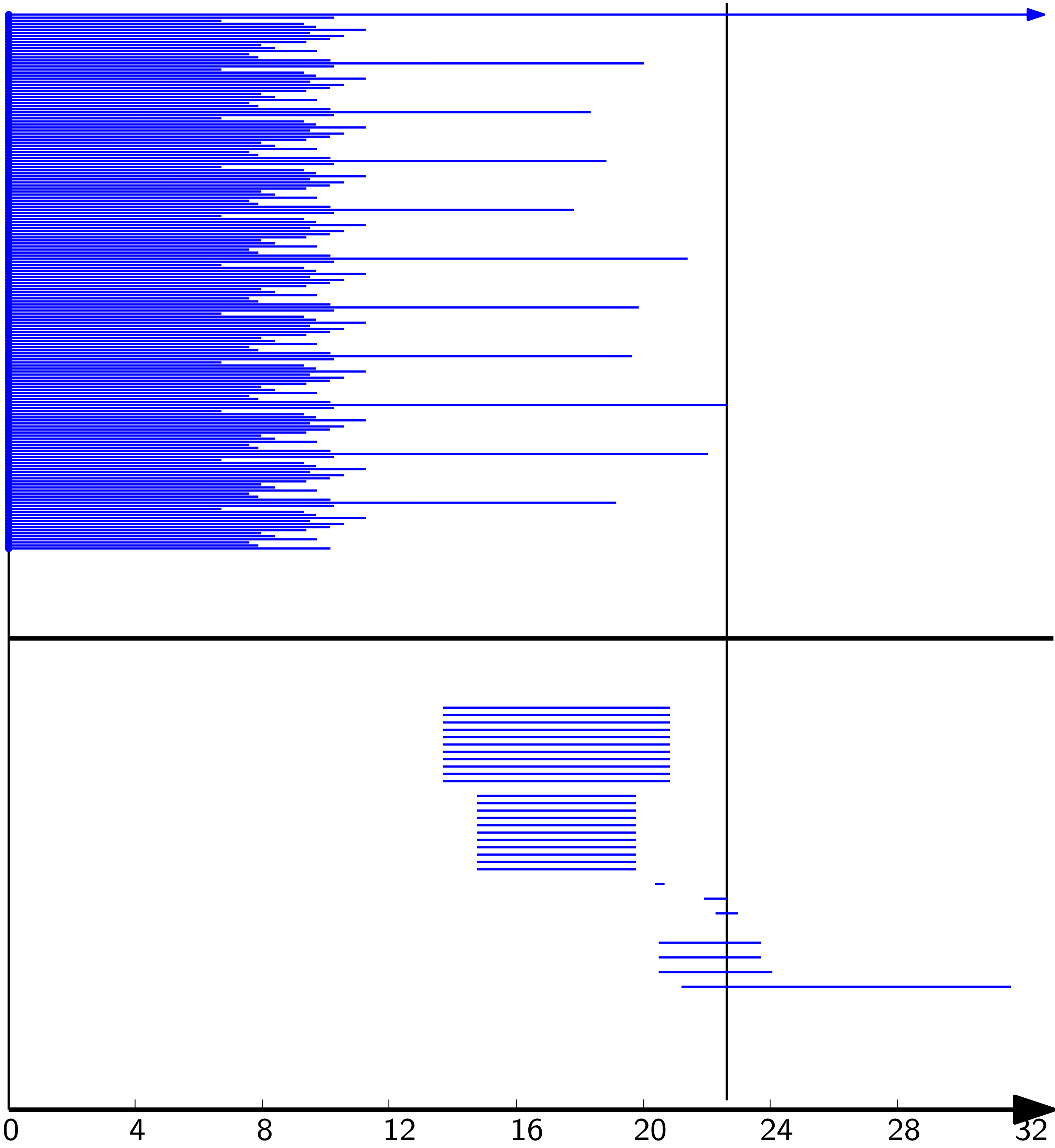
topology:

- connected components
- holes
- voids ...

Example: distance function

$$f_P: \quad \mathbb{R}^2 \to \mathbb{R}$$
$$x \mapsto \min_{p \in P} \|x - p\|_2$$



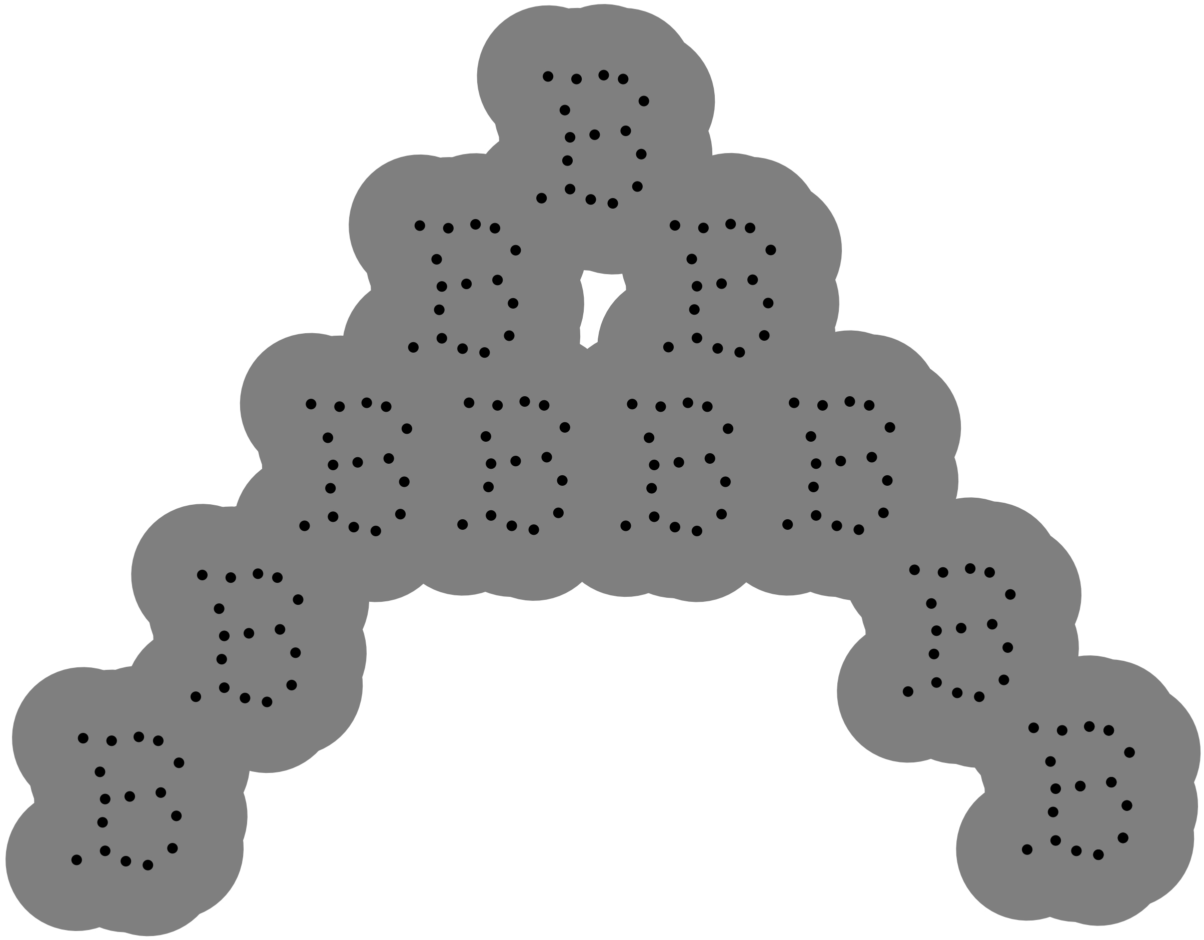


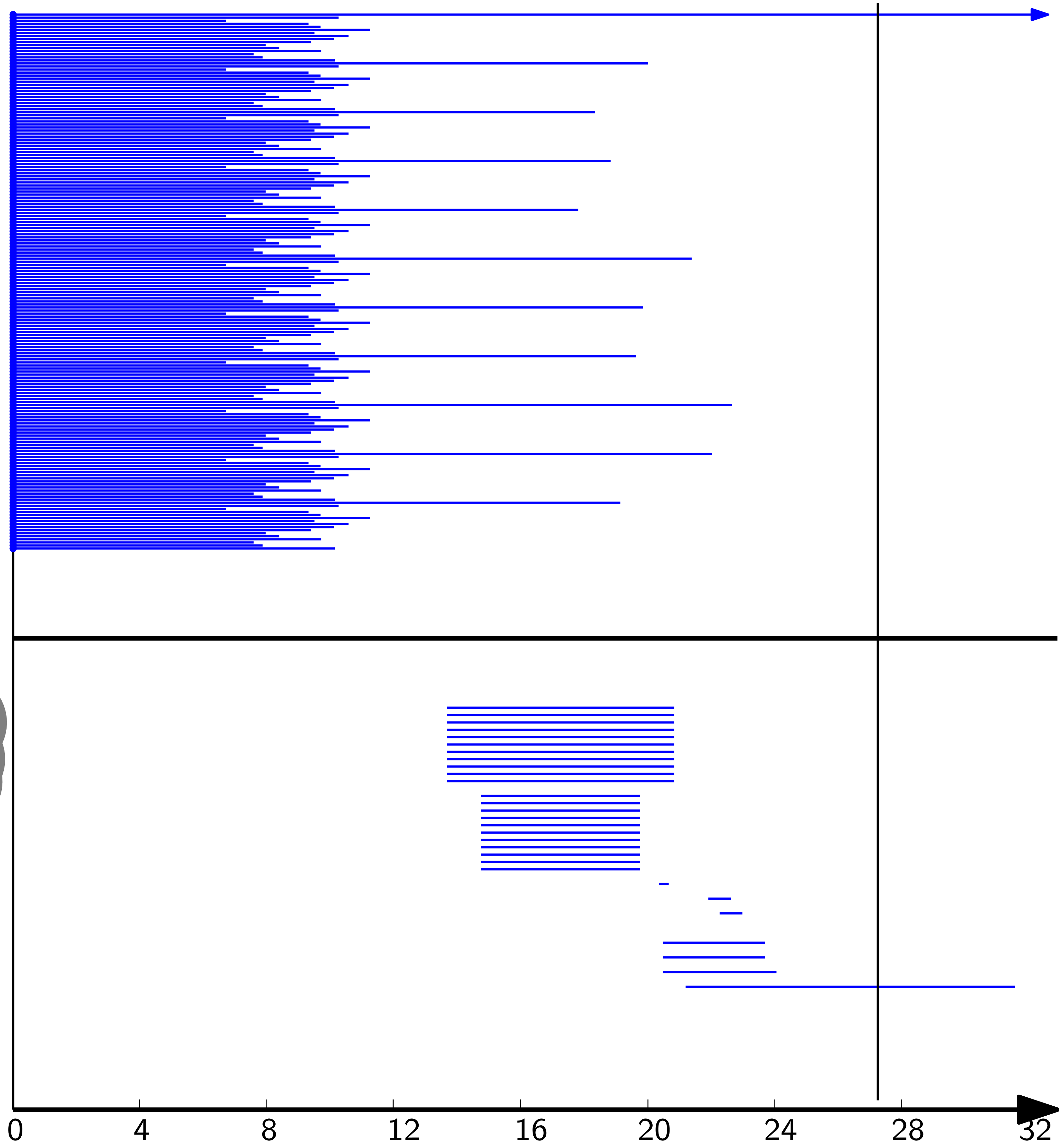
topology:

- connected components
- holes
- voids ...

Example: distance function

$$f_P : \quad \mathbb{R}^2 \to \mathbb{R}$$
$$x \mapsto \min_{p \in P} \|x - p\|_2$$



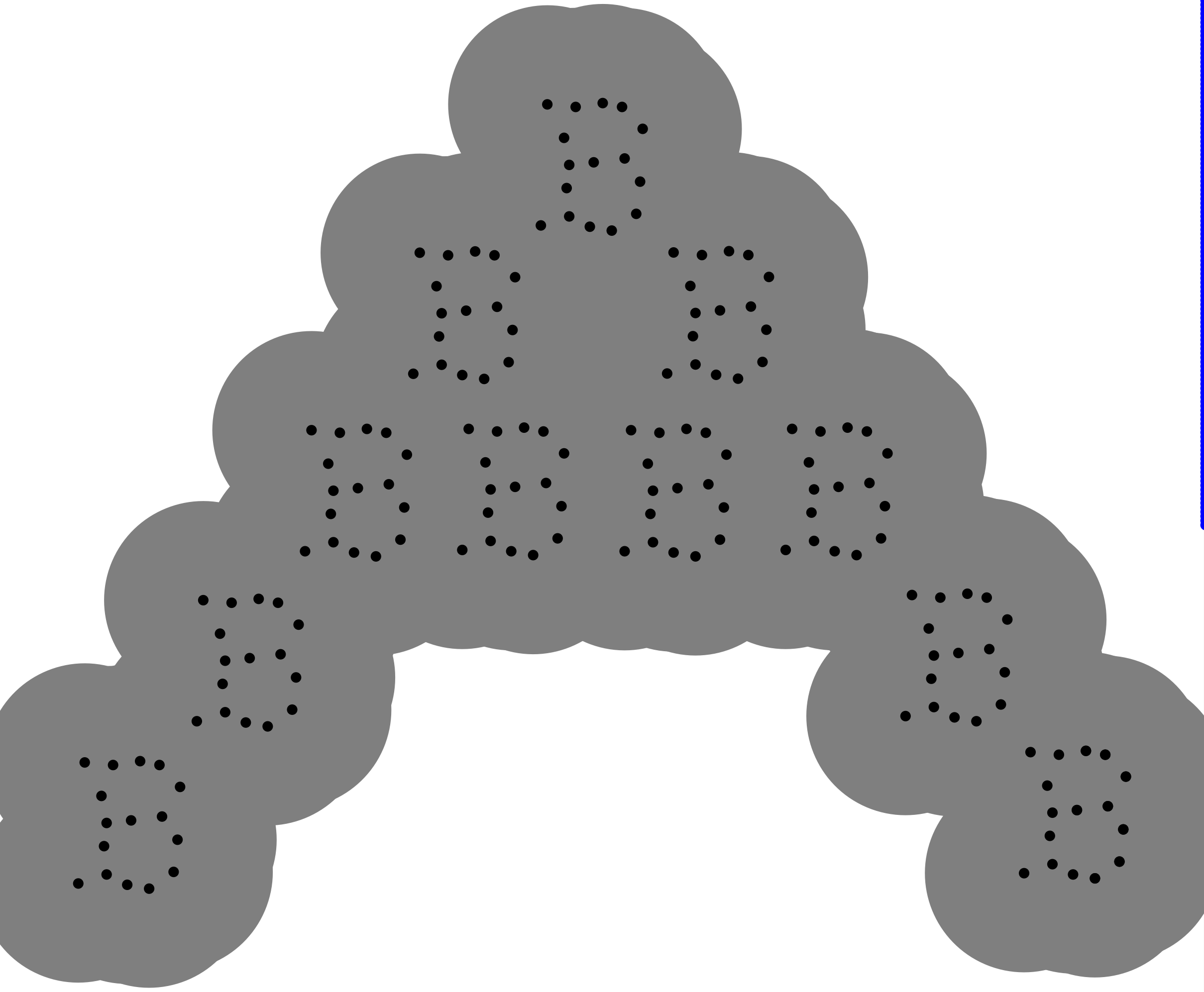


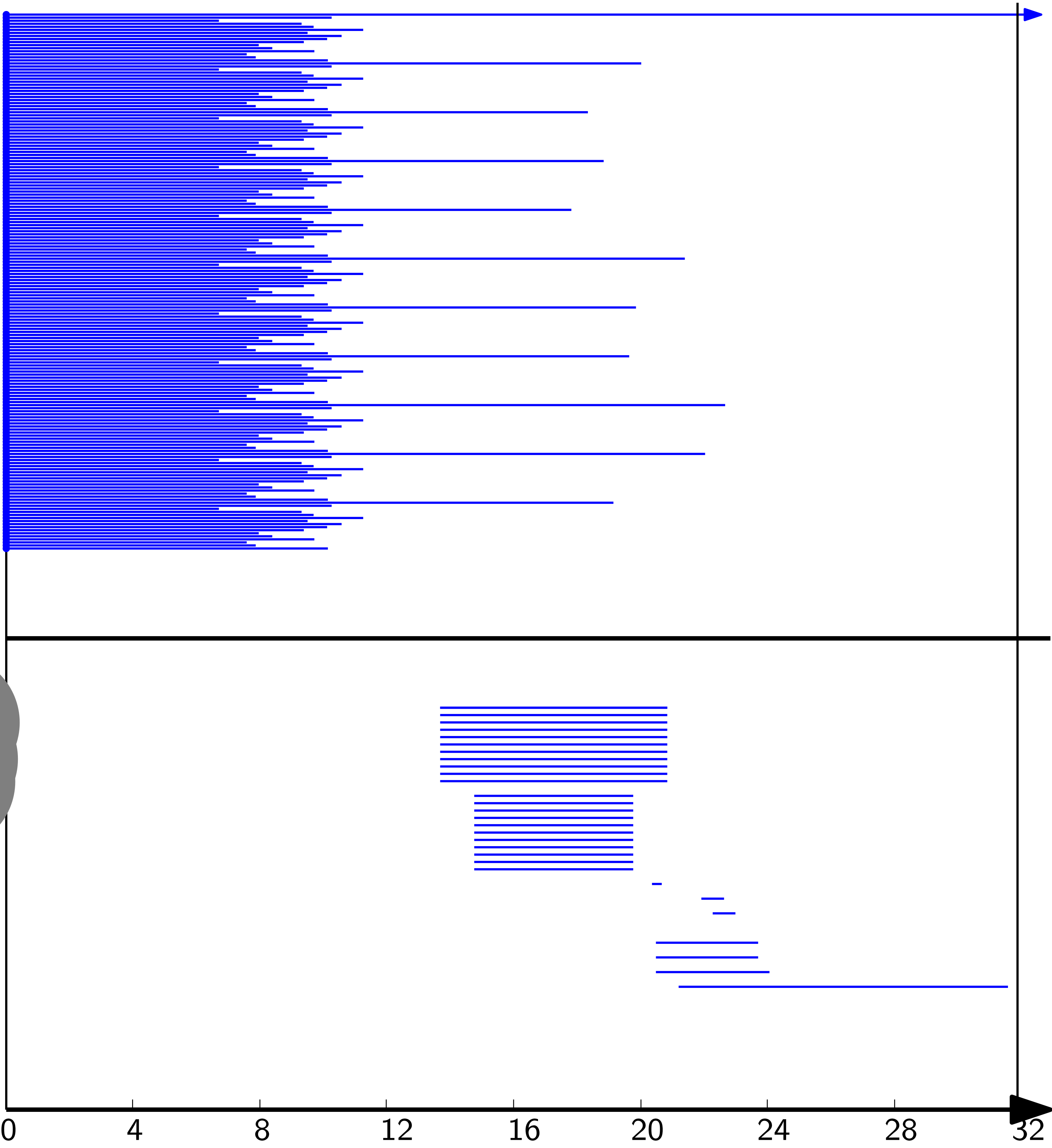
topology:

- connected components
- holes
- voids ...

Example: distance function

$$f_P : \quad \mathbb{R}^2 \to \mathbb{R}$$
$$\qquad x \mapsto \min_{p \in P} \|x - p\|_2$$



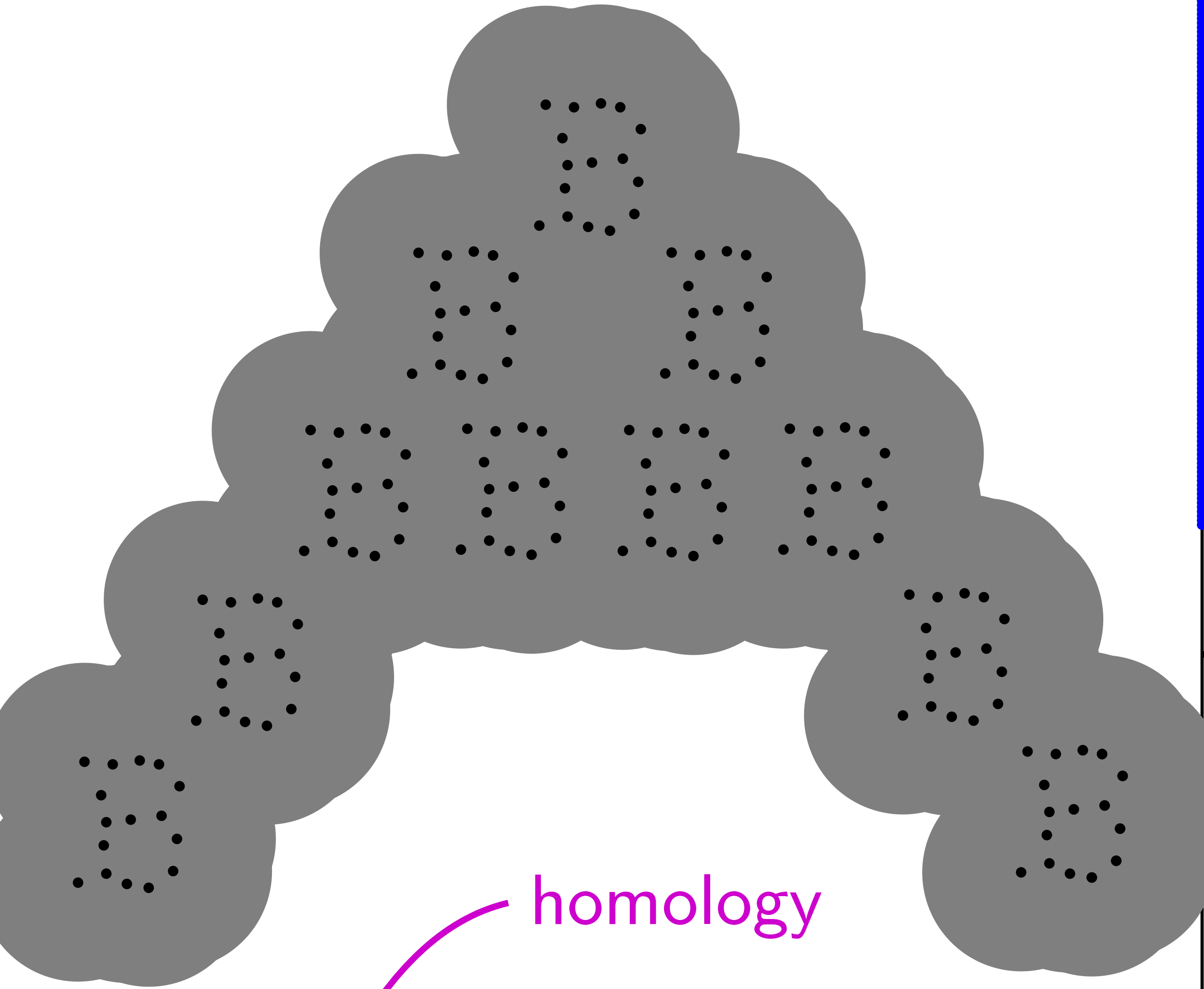


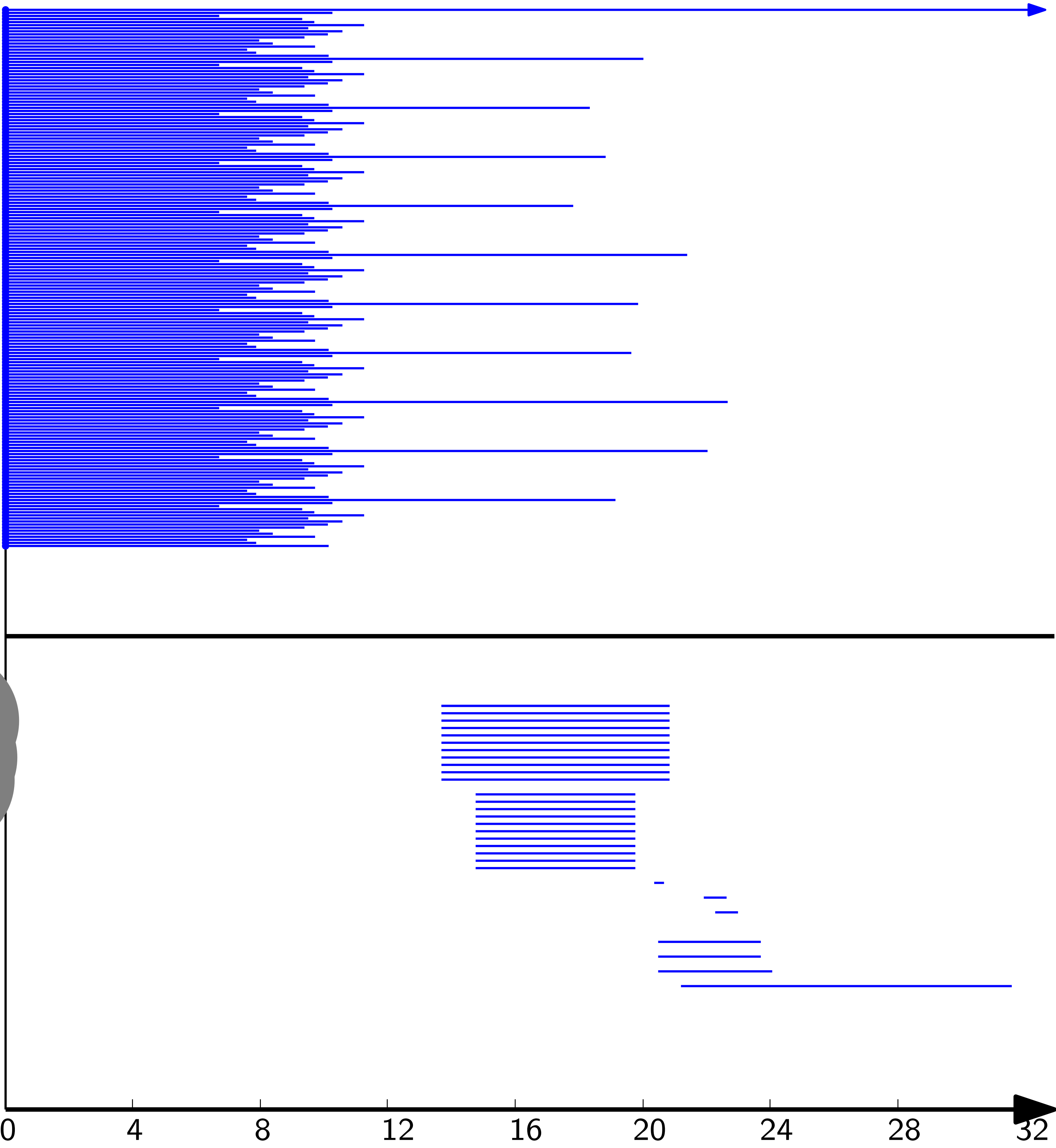
topology:

- connected components
- holes
- voids ...

Example: distance function

$$f_P: \quad \mathbb{R}^2 \to \mathbb{R}$$
$$x \mapsto \min_{p \in P} \|x - p\|_2$$



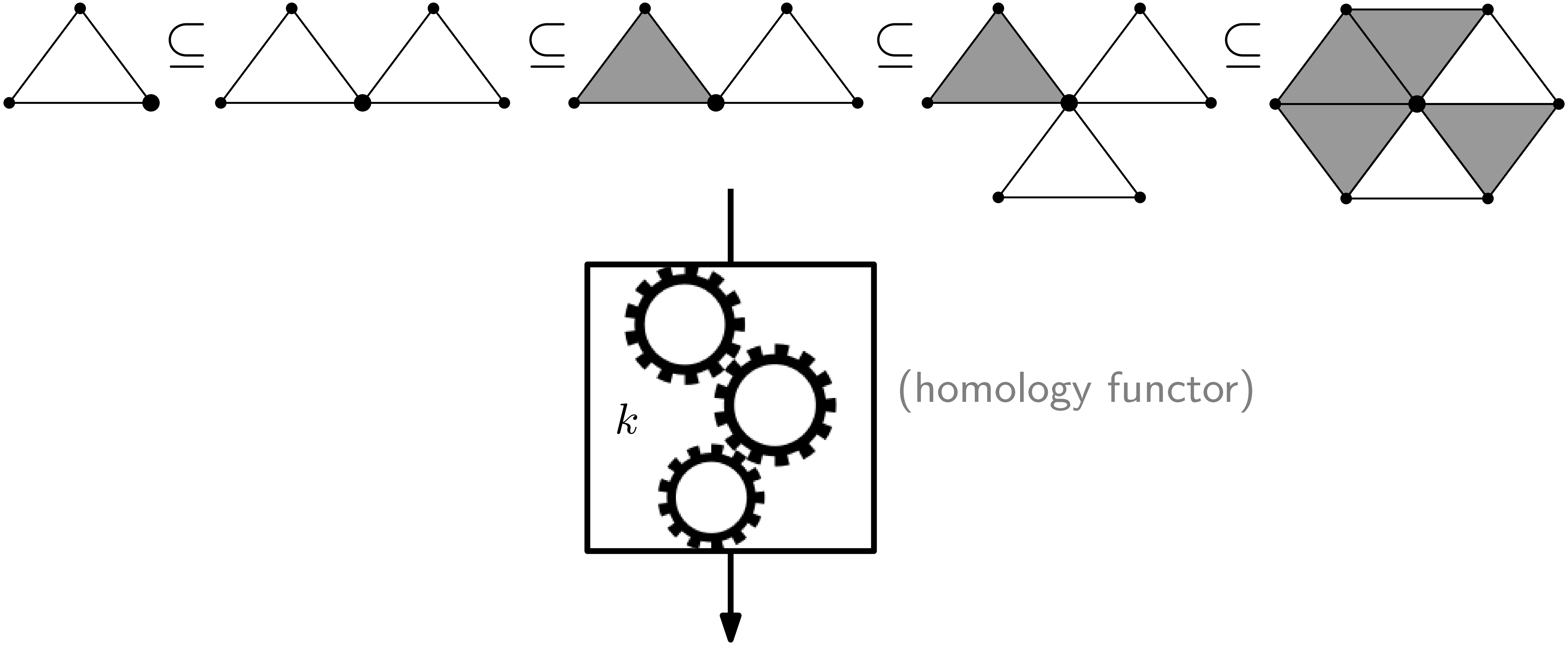


homology

topology:

- connected components
- holes
- voids ...

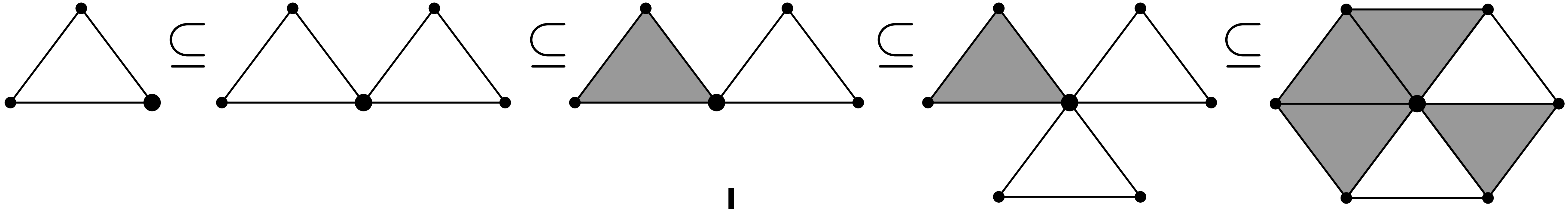
Homology of sublevel sets

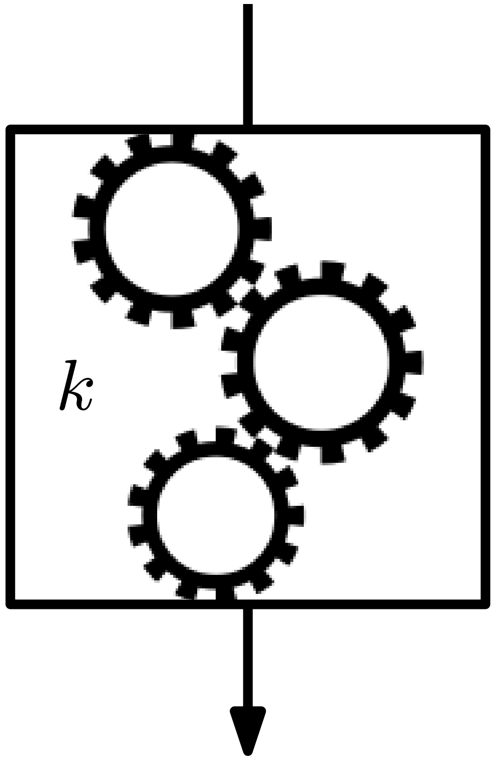


(homology functor)

$$k \xrightarrow{\begin{pmatrix}1\\0\end{pmatrix}} k^2 \xrightarrow{\begin{pmatrix}0 & 1\end{pmatrix}} k \xrightarrow{\begin{pmatrix}0\\1\end{pmatrix}} k^2 \xrightarrow{\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}} k^2 \quad \dots$$

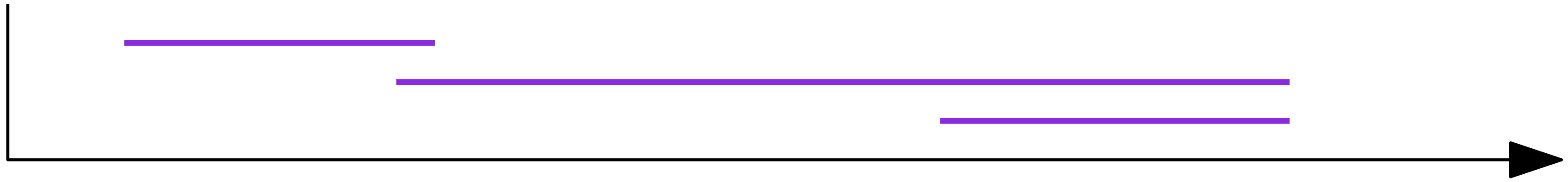
Homology of sublevel sets



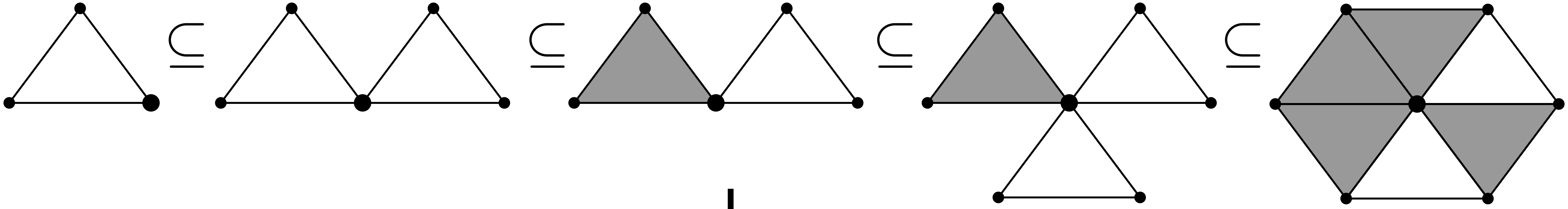


(homology functor)

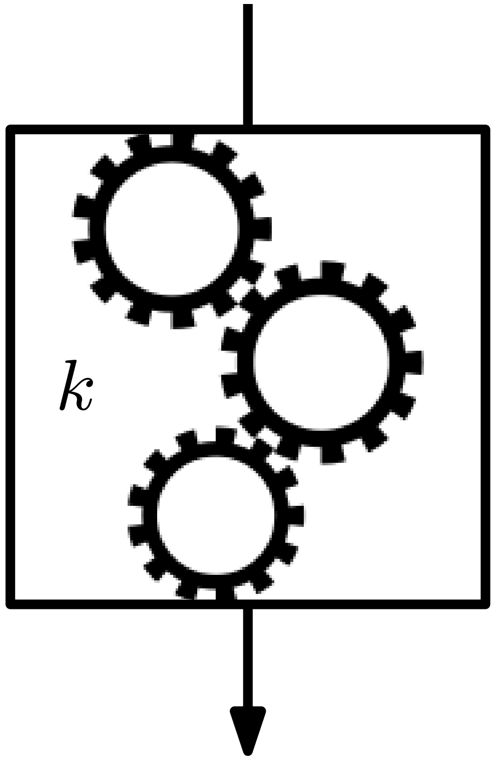
$$k \xrightarrow{\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)} k^2 \xrightarrow{\left(\begin{smallmatrix}0 & 1\end{smallmatrix}\right)} k \xrightarrow{\left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)} k^2 \xrightarrow{\left(\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right)} k^2 \quad \cdots$$



Homology of sublevel sets



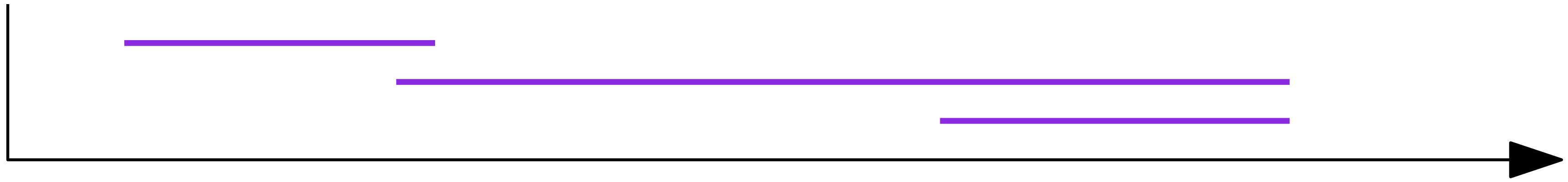
today



(homology functor)

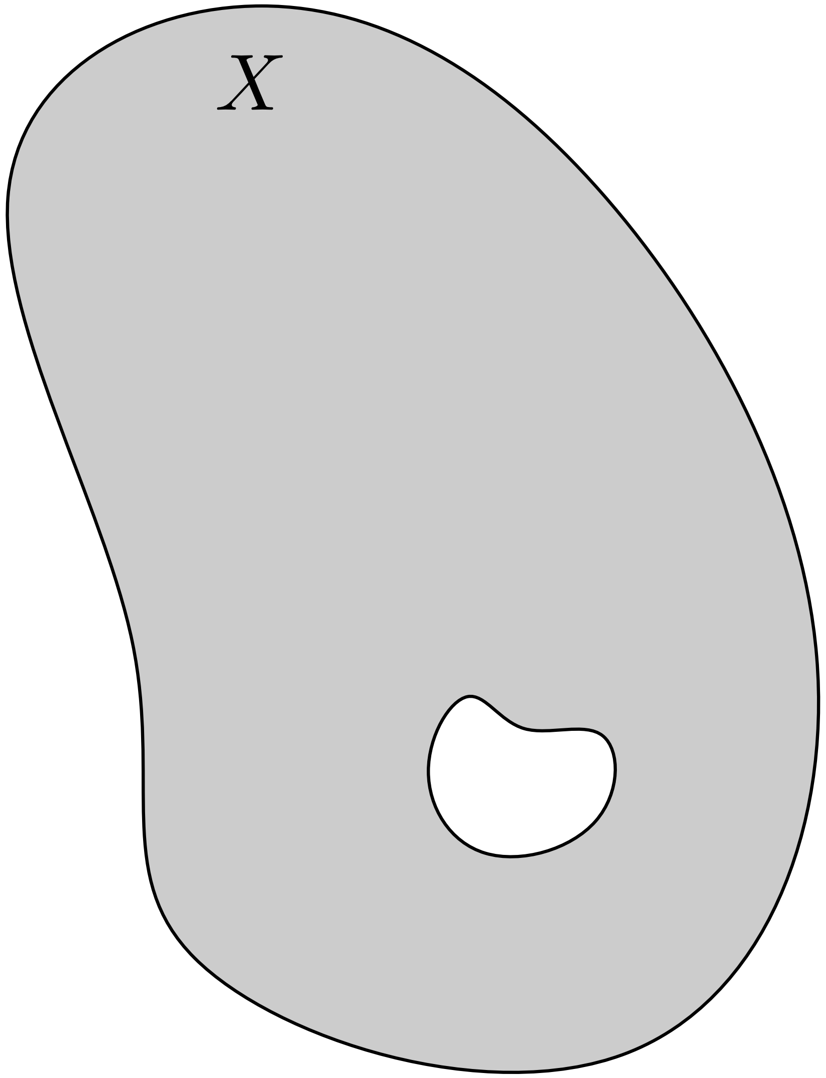
tomorrow

$$k \xrightarrow{\binom{1}{0}} k^2 \xrightarrow{(\,0\ \ 1\,)} k \xrightarrow{\binom{0}{1}} k^2 \xrightarrow{\begin{pmatrix}1&0\\0&1\end{pmatrix}} k^2 \quad \cdots$$



Intuition behind homology

An invariant that captures ”holes” of all dimensions in a topological space



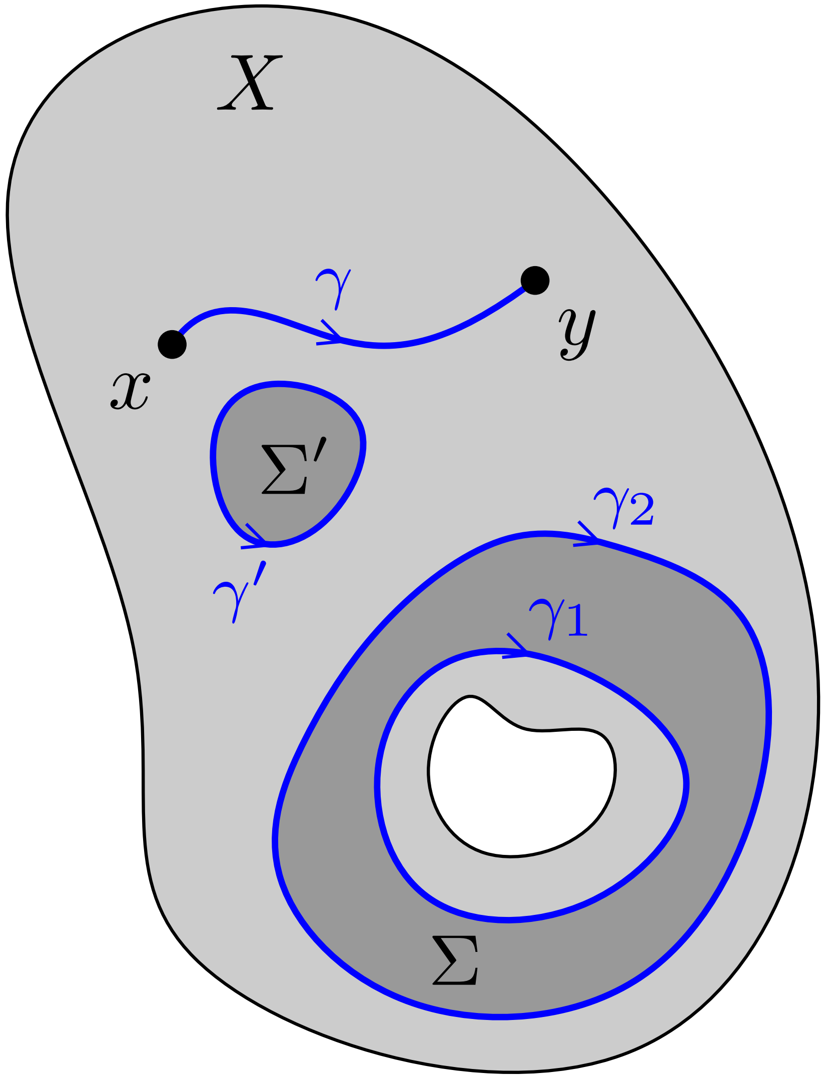
Intuition behind homology

An invariant that captures ”holes” of all dimensions in a topological space

$$x \equiv y \quad \Longleftrightarrow \quad \exists \gamma \text{ s.t. } \{x, y\} = \partial \gamma$$

$$\gamma' \equiv 0 \quad \Longleftrightarrow \quad \exists \Sigma' \text{ s.t. } \gamma' = \partial \Sigma'$$

$$\gamma_1 \equiv \gamma_2 \quad \Longleftrightarrow \quad \exists \Sigma \text{ s.t. } \gamma_1 \cup \gamma_2 = \partial \Sigma$$



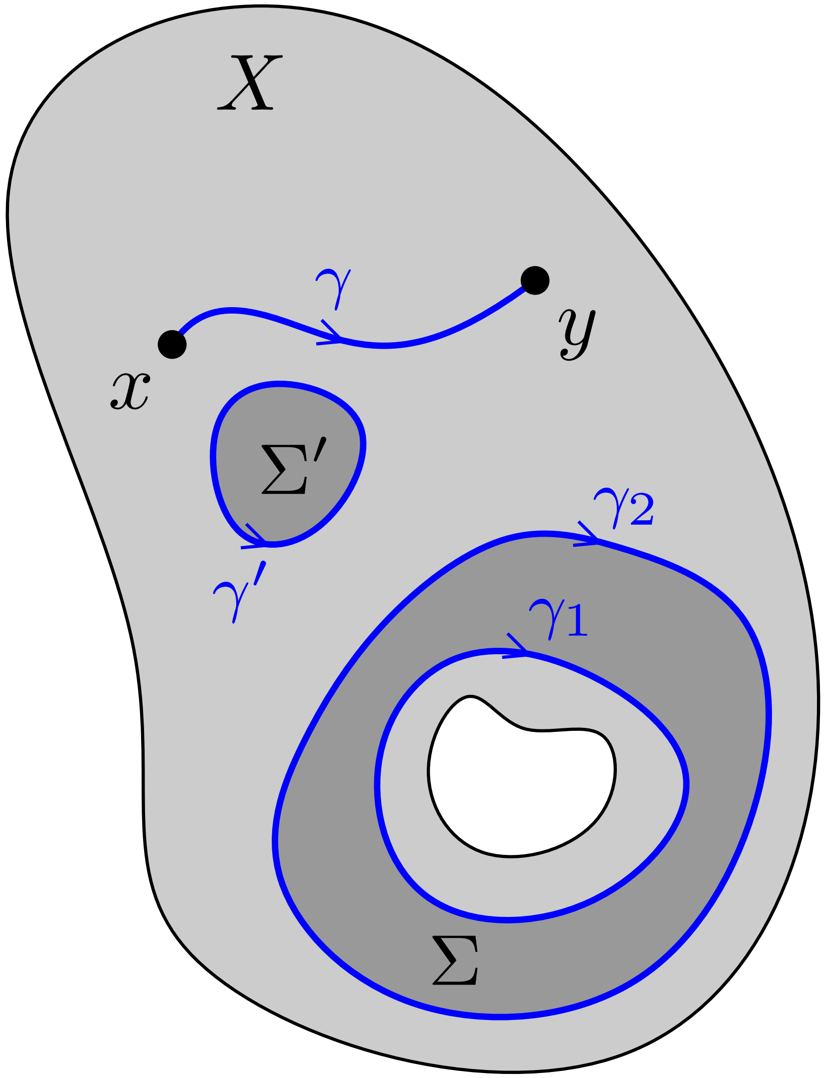
Intuition behind homology

An invariant that captures ”holes” of all dimensions in a topological space

$$x \equiv y \quad \Longleftrightarrow \quad \exists\gamma \text{ s.t. } \underset{x-y}{\{x,y\}} = \partial\gamma$$

$$\gamma' \equiv 0 \quad \Longleftrightarrow \quad \exists\Sigma' \text{ s.t. } \gamma' = \partial\Sigma'$$

$$\gamma_1 \equiv \gamma_2 \quad \Longleftrightarrow \quad \exists\Sigma \text{ s.t. } \underset{\gamma_1-\gamma_2}{\gamma_1 \cup \gamma_2} = \partial\Sigma$$



$x, y, \gamma, \gamma'\gamma_1, \gamma_2, \Sigma, \Sigma'$: elements of some module

∂ : linear operator

cycles $\leftarrow \ker\partial$ boundaries $\leftarrow \operatorname{im}\partial$ homology $\leftarrow \ker\partial/\operatorname{im}\partial$