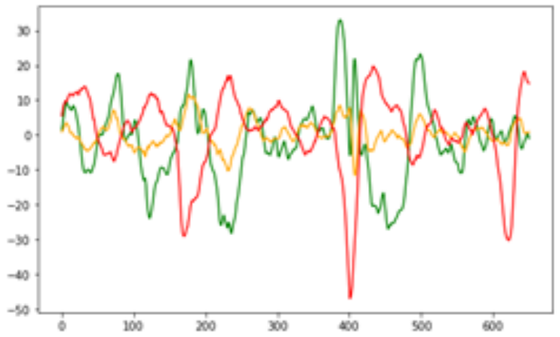
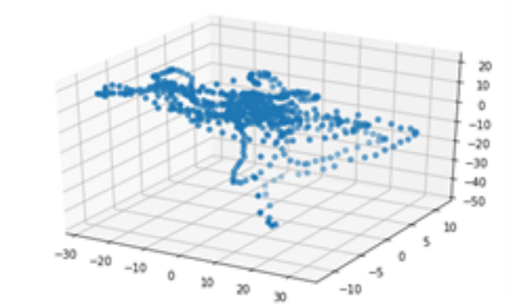
Introduction to Topological Data Analysis – I

Marc Glisse

Inria, France

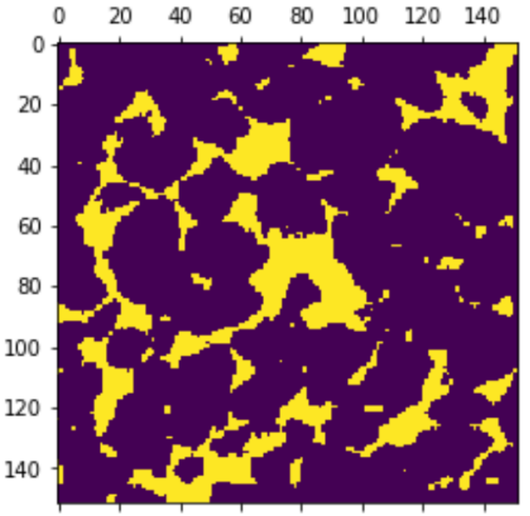
Motivation

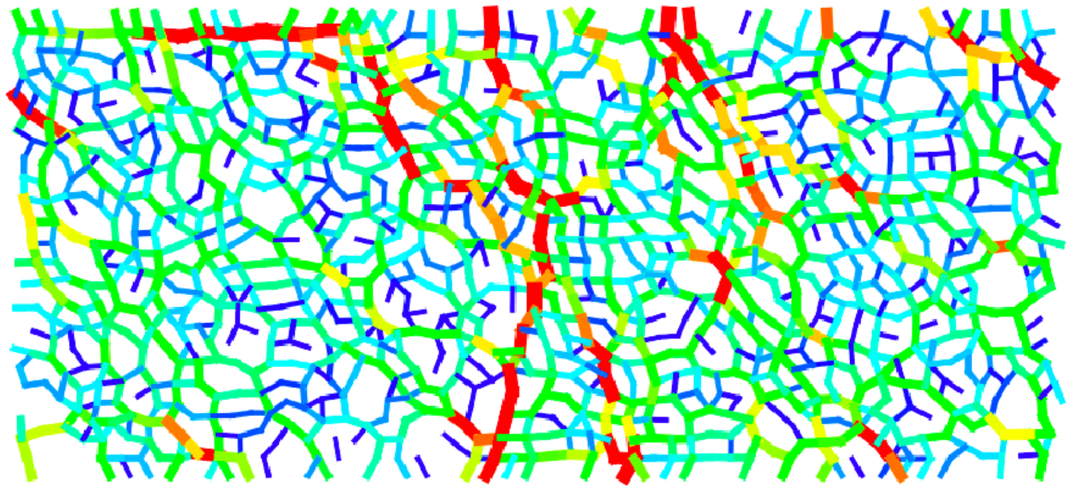




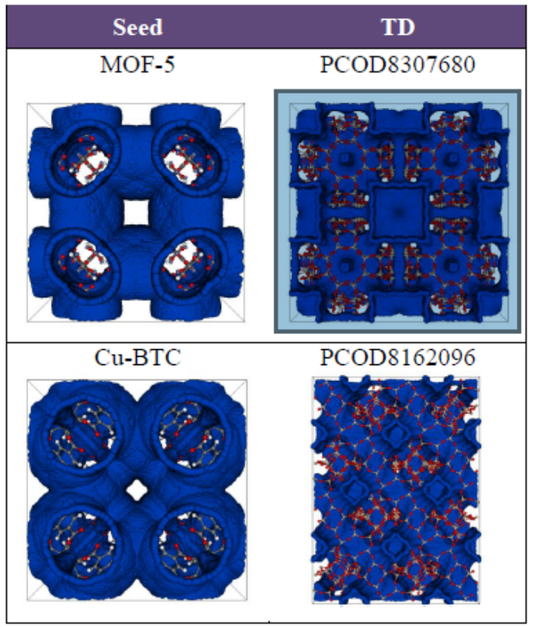
[Sensors]

[3D images
(porous rocks)]





[Force fields in granular media]

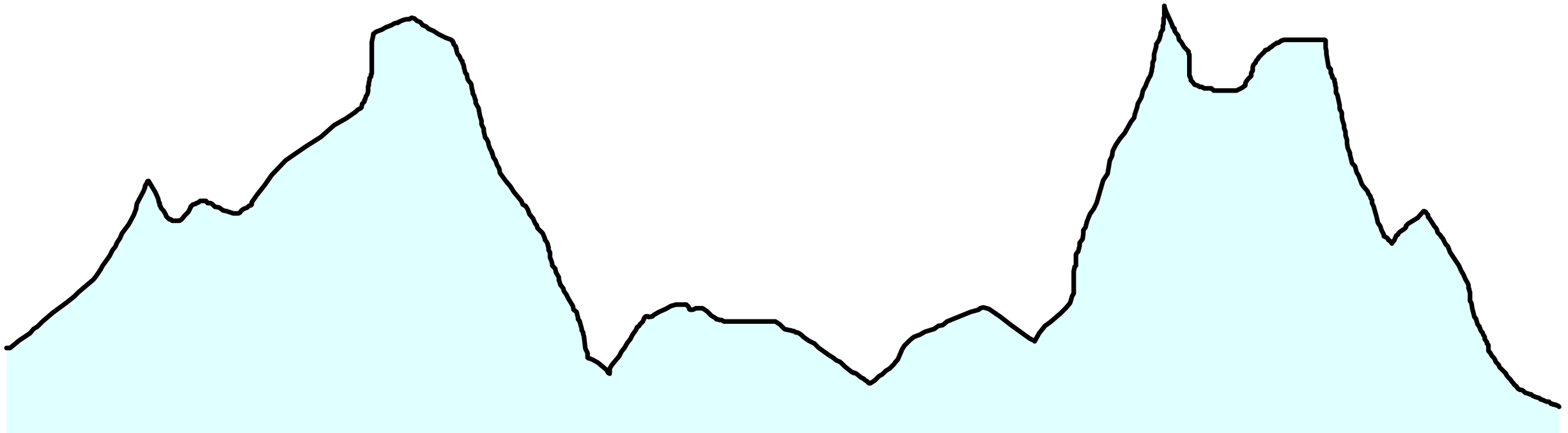


[Nano-materials -
Li et al 2017]

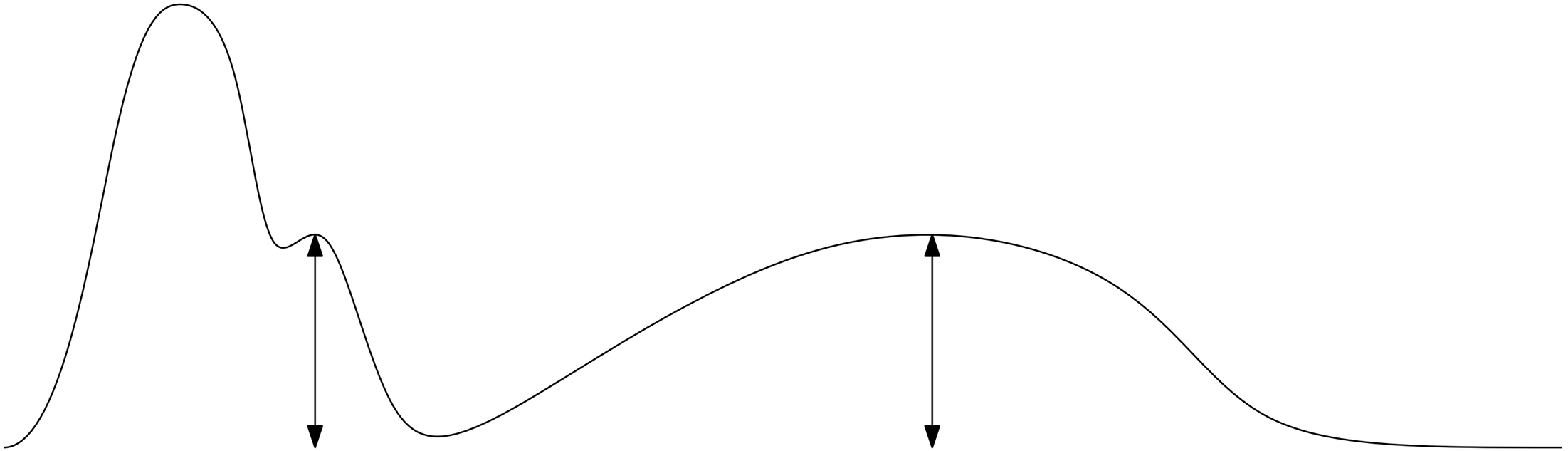
Data often have topological / geometric structure

How many mountains?

How many mountains?



Height is insufficient

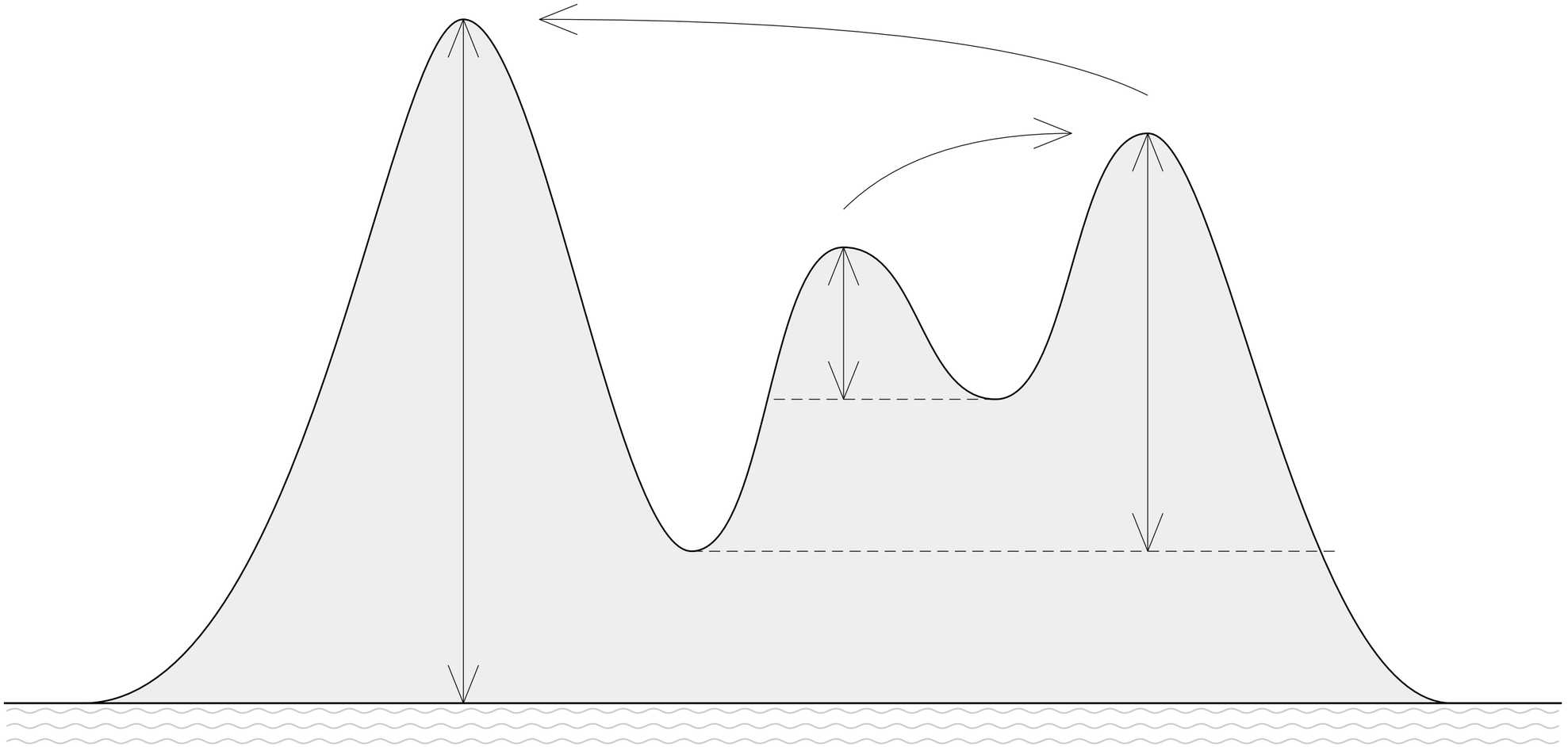


Small bump on the side of a bigger mountain

Independent mountain

Prominence (Topography)

Local maximum: how low do you need to go before you can reach a higher maximum?



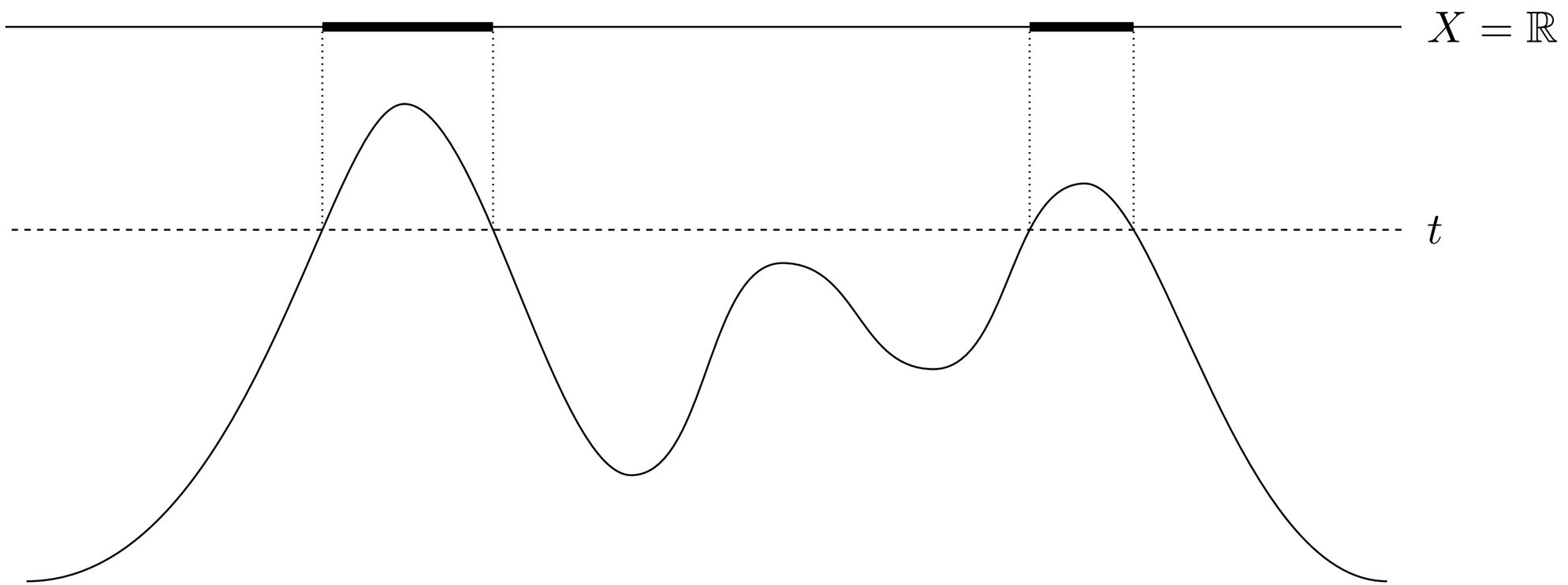
[wikipedia]

Superlevelsets

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

$F_t = f^{-1}([t, +\infty)) = \{x \in X, f(x) \geq t\}$

$F_{+\infty} = \emptyset,$ $F_t \subseteq F_{t'}$ when $t \geq t',$ $F_{-\infty} = X$

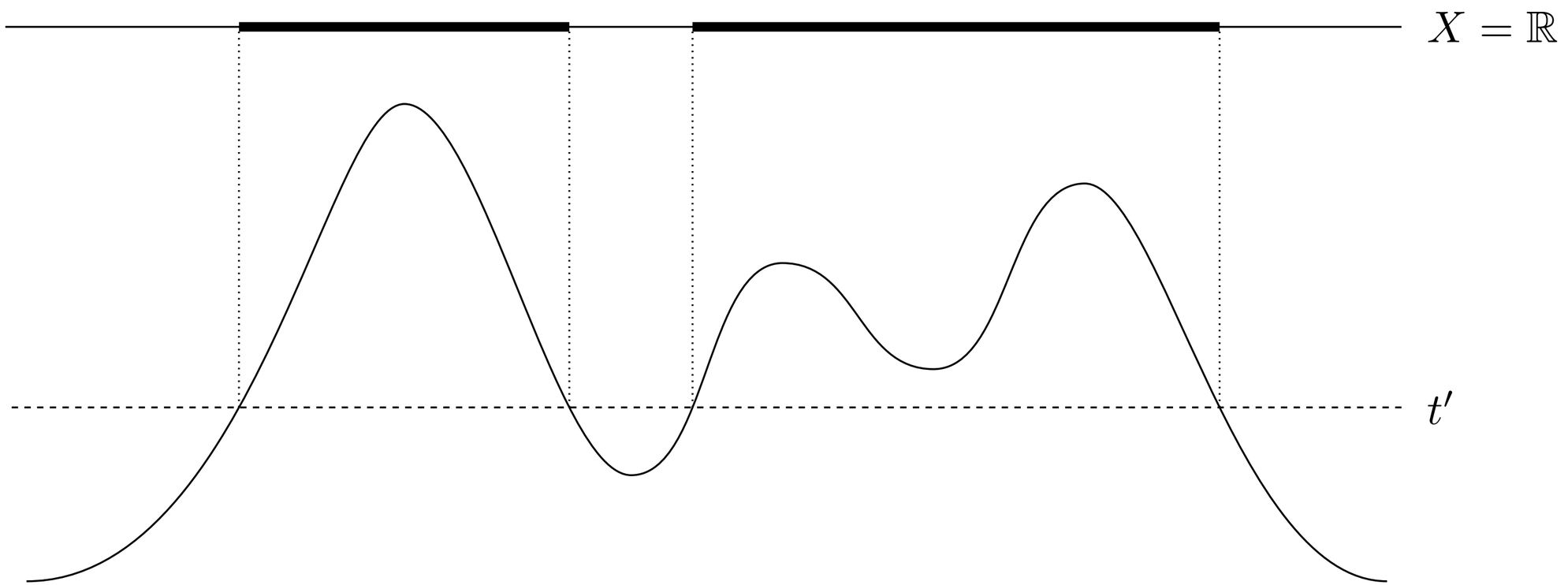


Superlevelsets

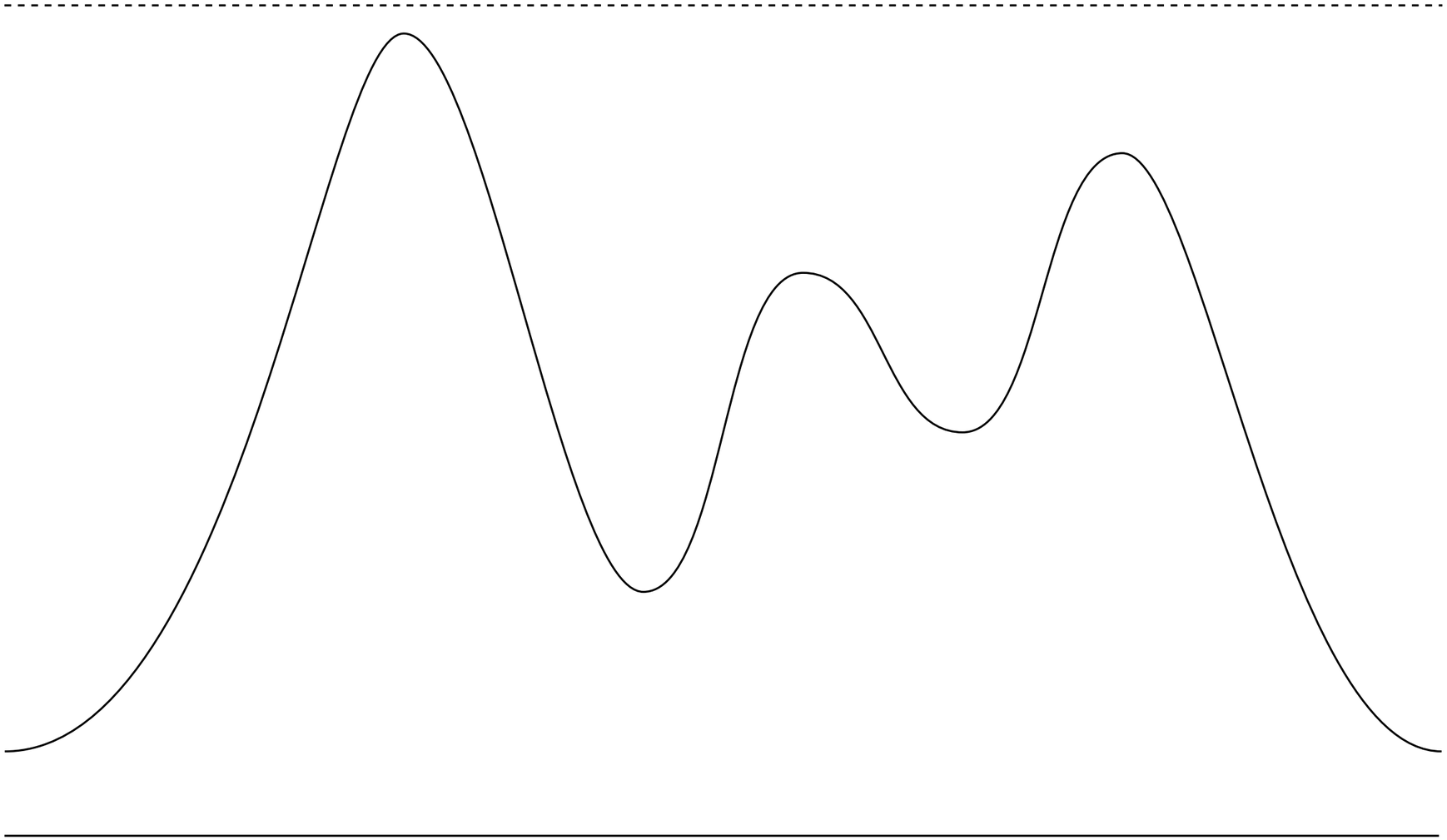
$f : X \to \mathbb{R}$

$F_t = f^{-1}([t, +\infty)) = \{x \in X, f(x) \geq t\}$

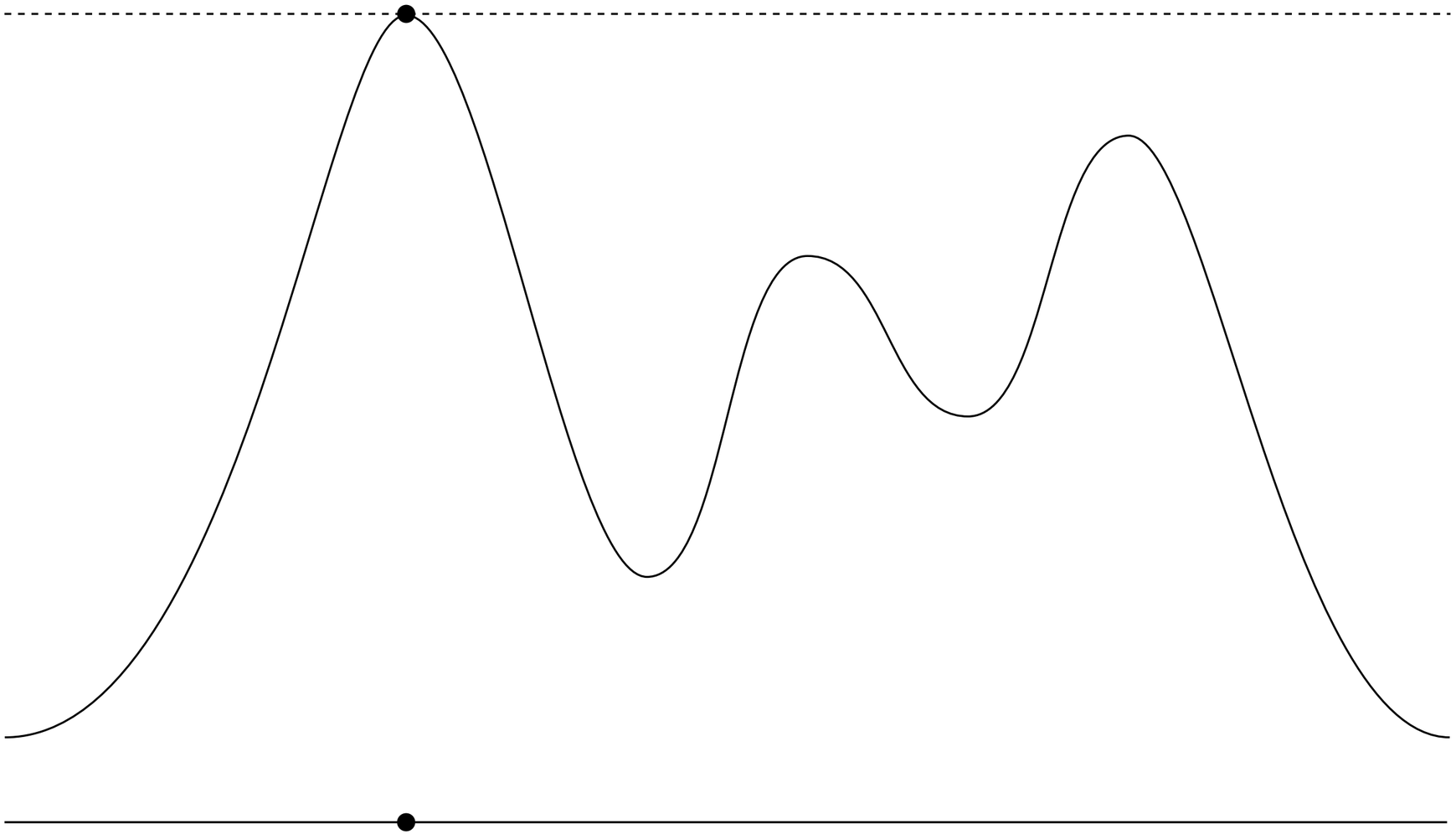
$F_{+\infty} = \emptyset,$ $F_t \subseteq F_{t'}$ when $t \geq t',$ $F_{-\infty} = X$



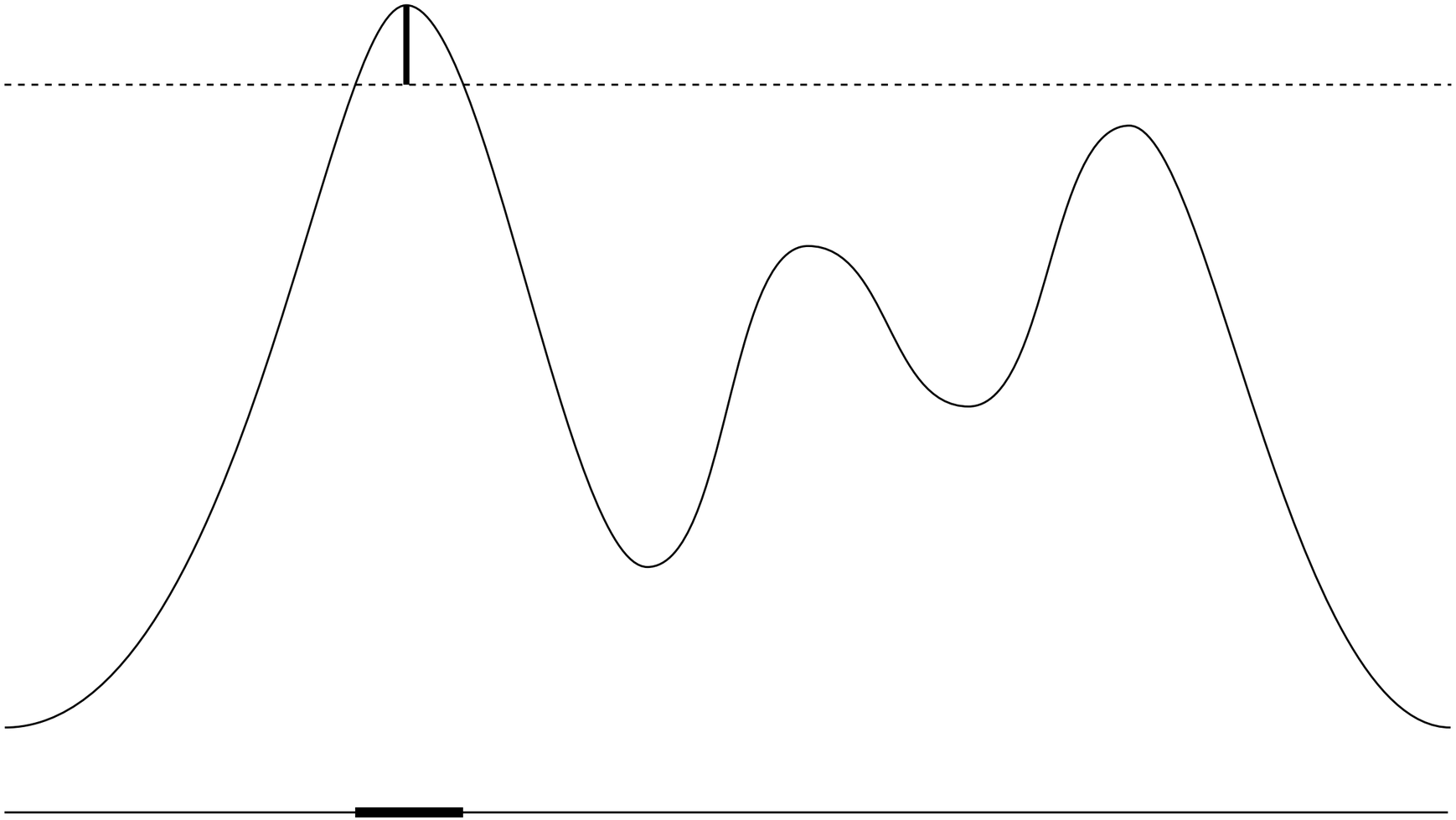
Tracking connected components in superlevelsets



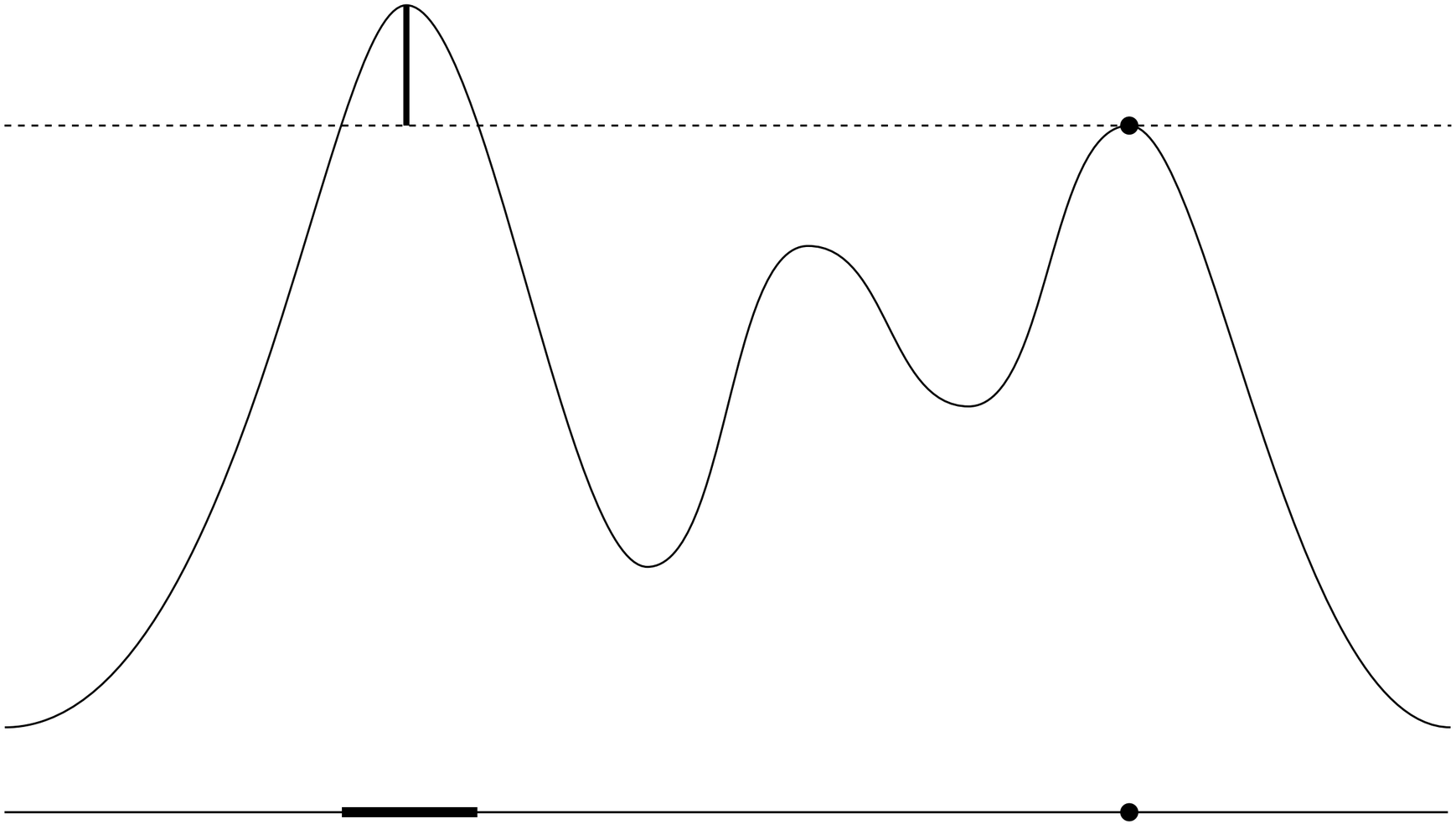
Tracking connected components in superlevelsets



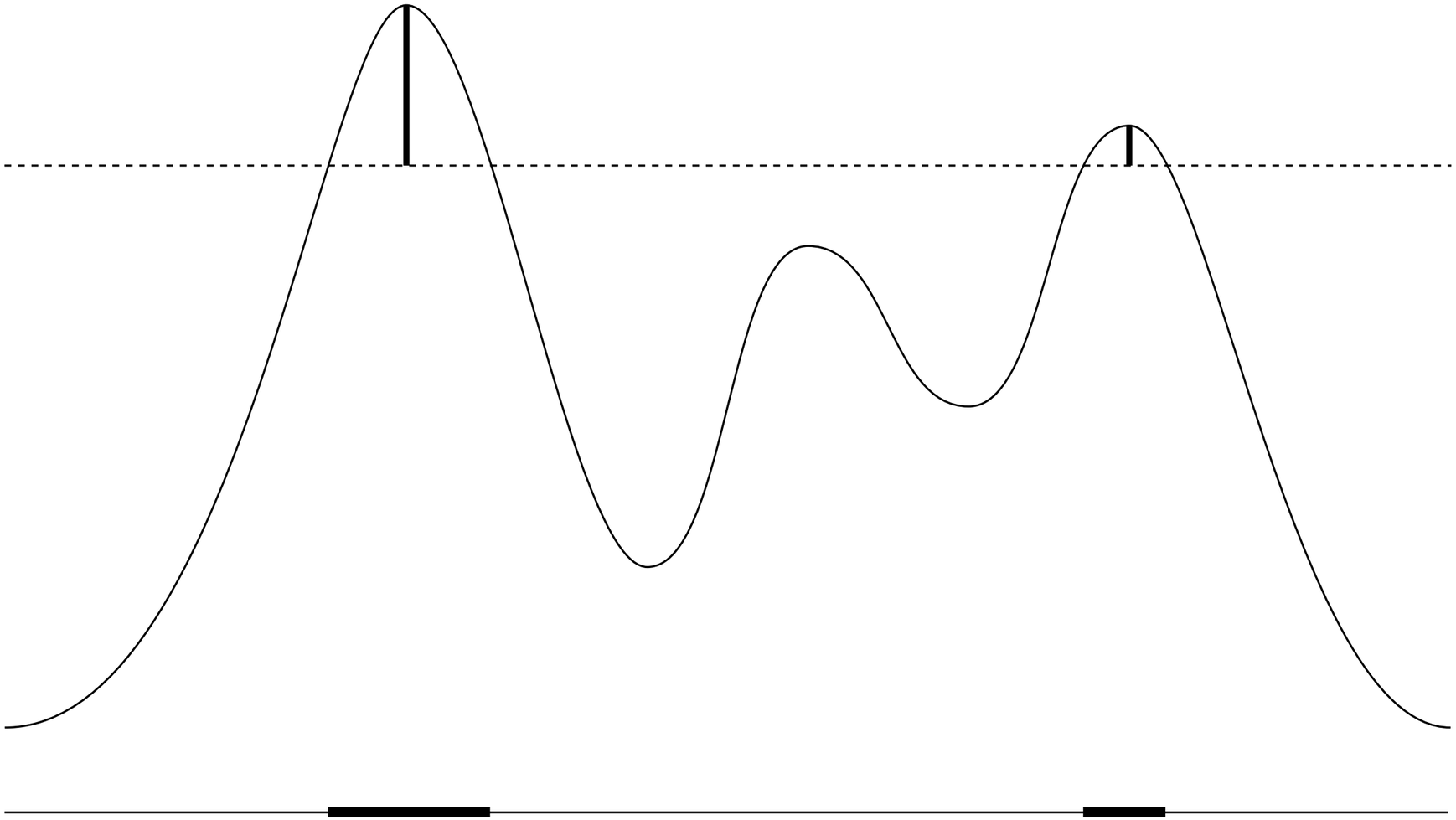
Tracking connected components in superlevelsets



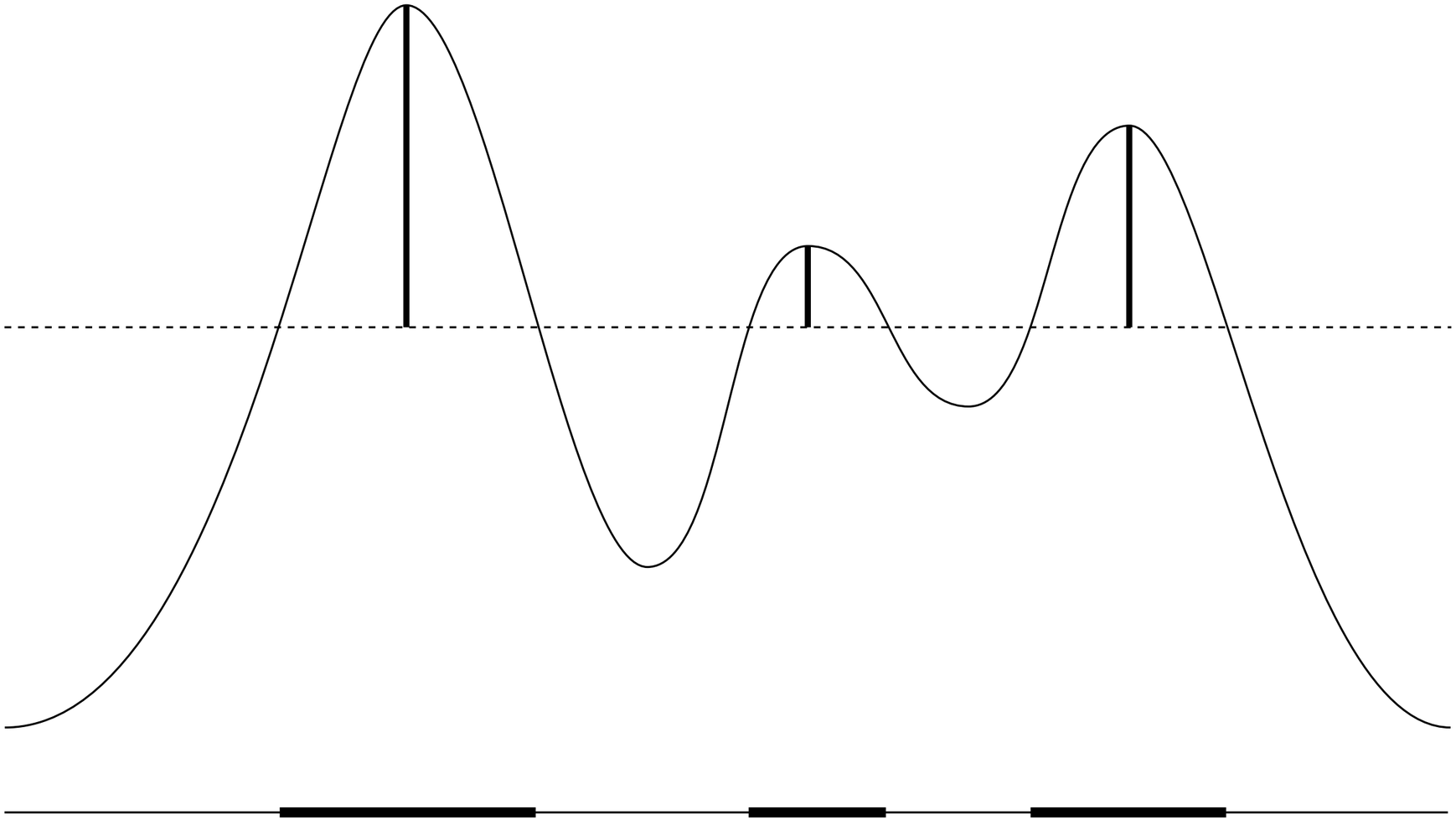
Tracking connected components in superlevelsets



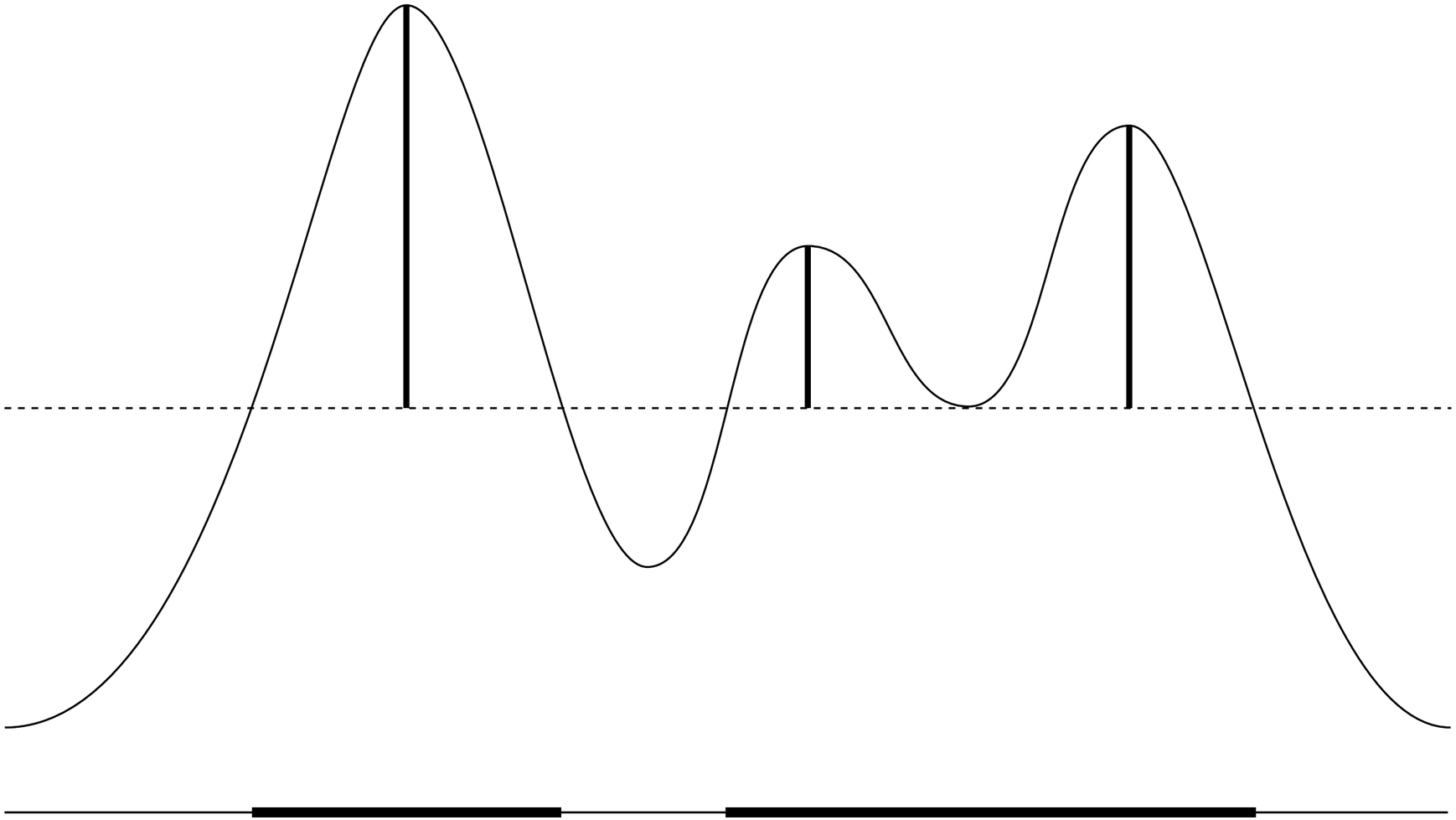
Tracking connected components in superlevelsets



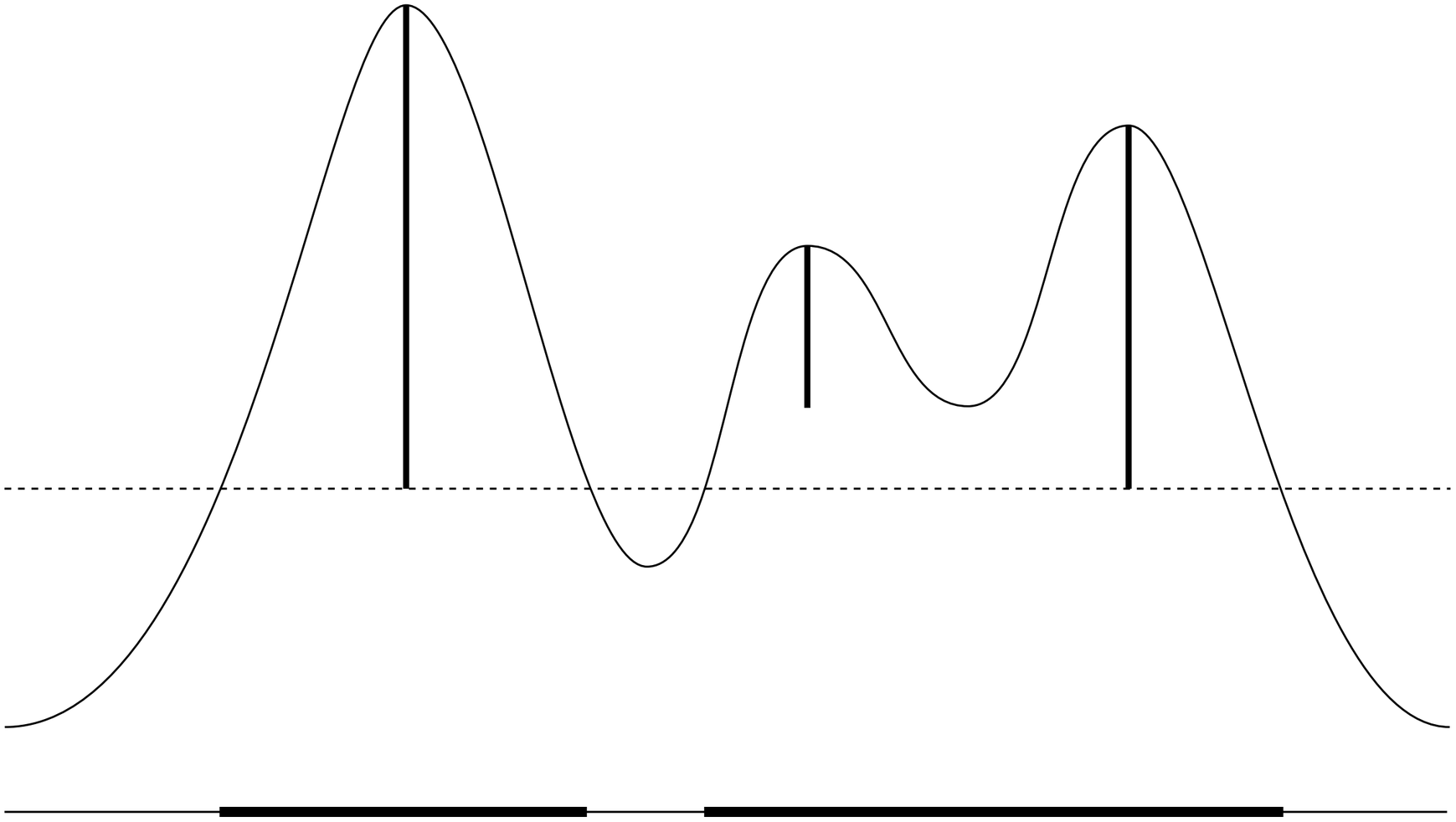
Tracking connected components in superlevelsets



Tracking connected components in superlevelsets

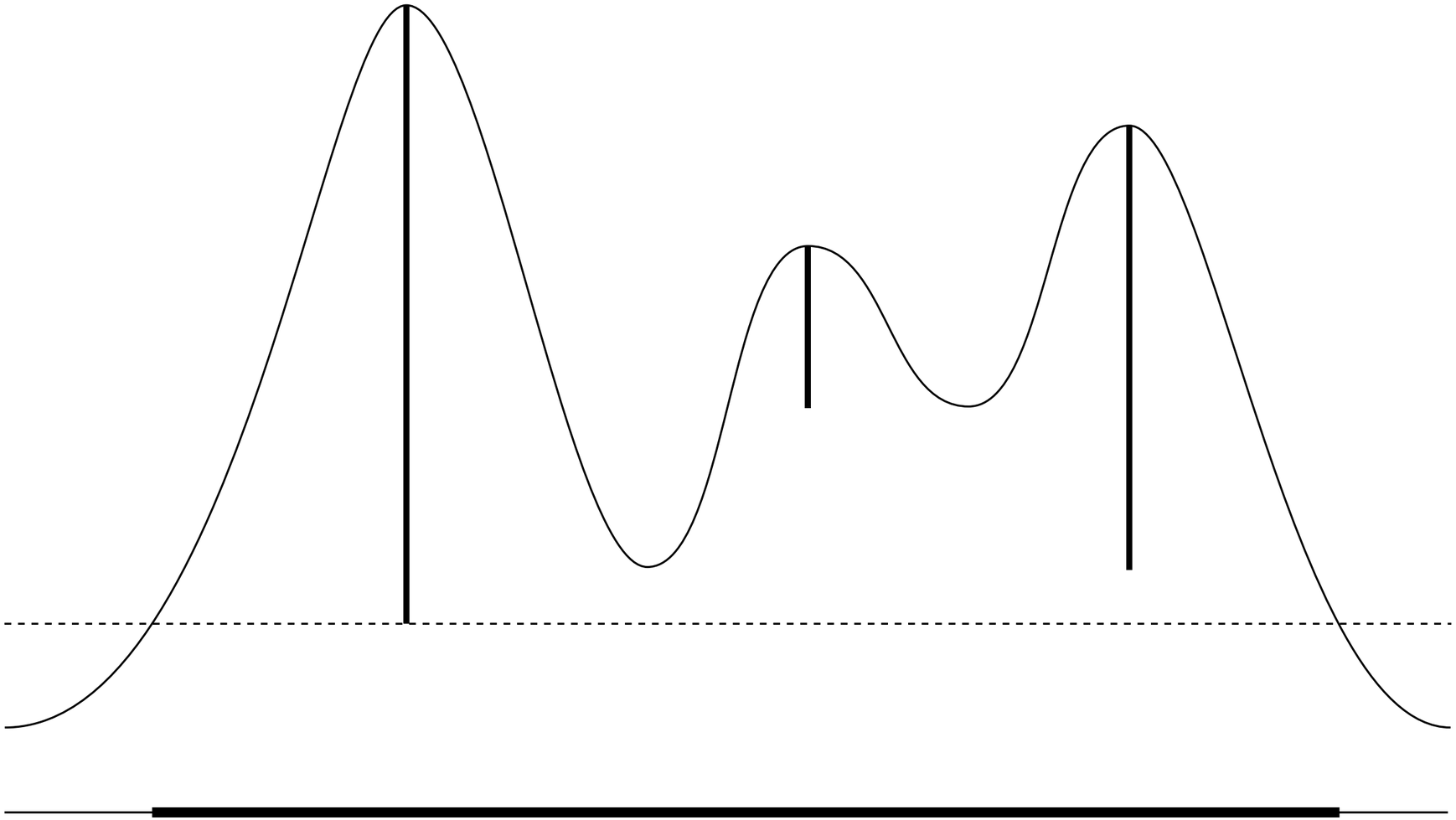


Tracking connected components in superlevelsets



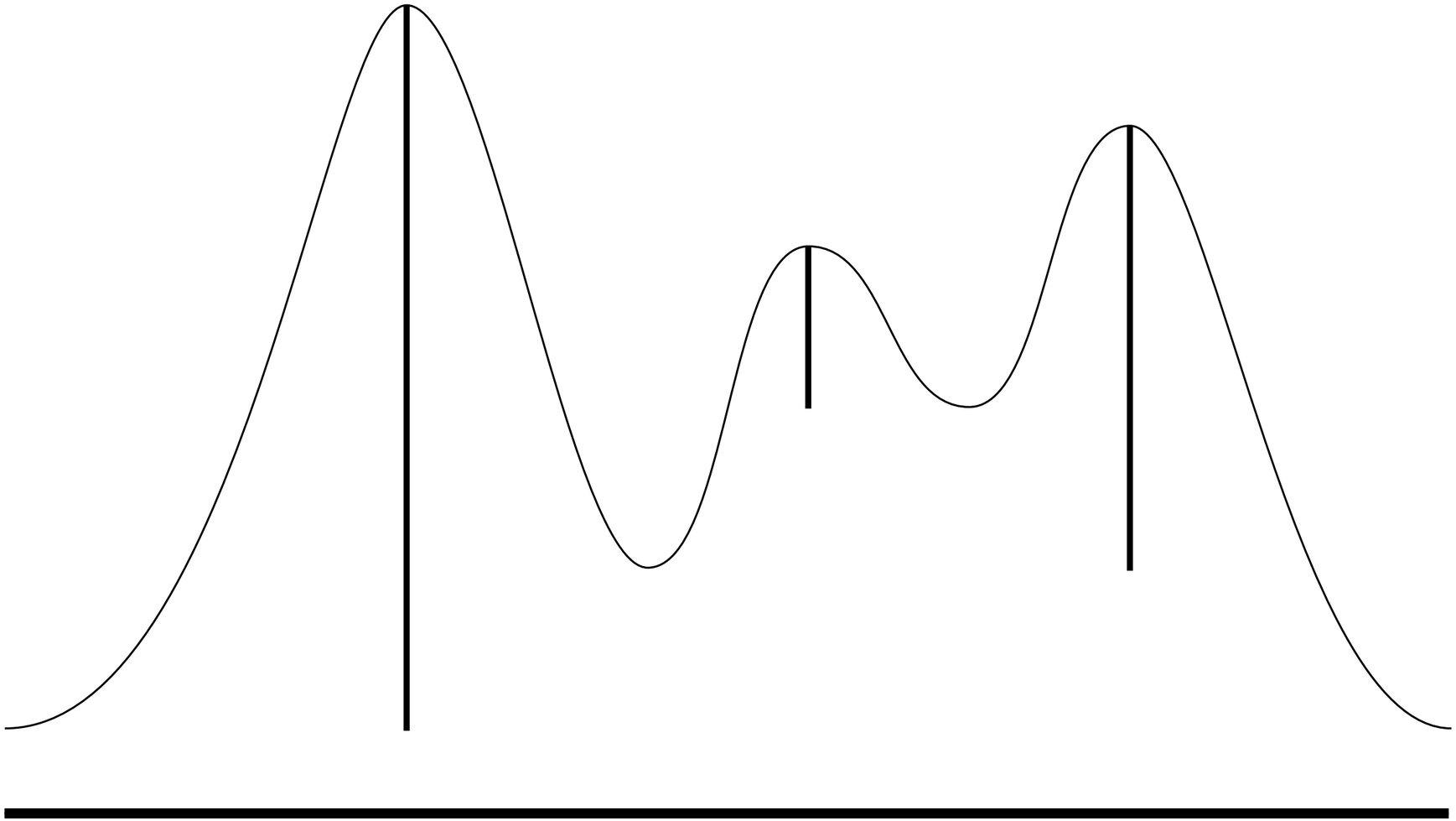
Smaller mountain merged into a bigger mountain: end the bar

Tracking connected components in superlevelsets

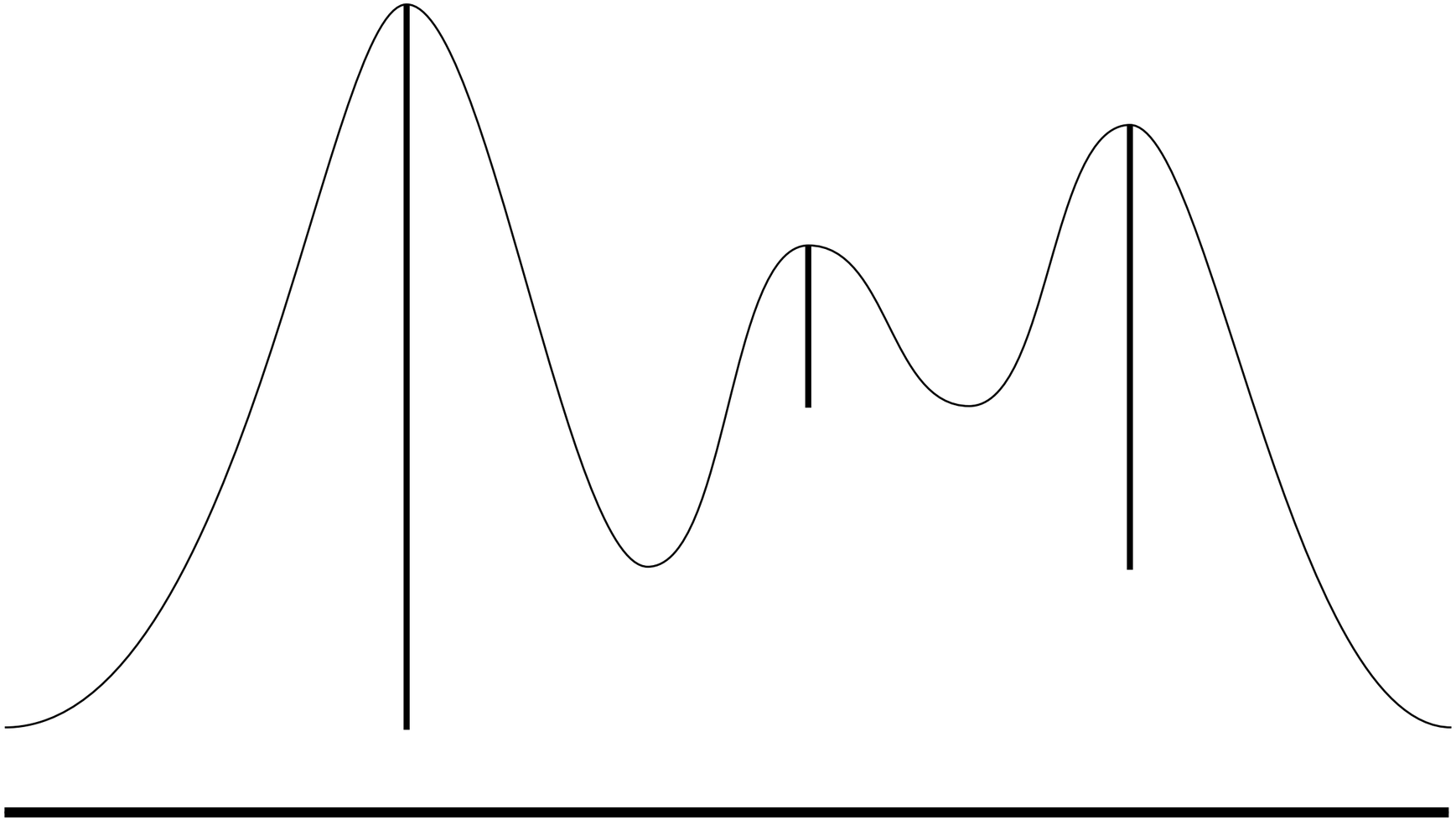


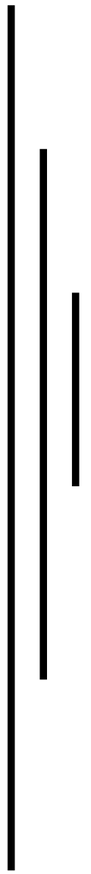
Smaller mountain merged into a bigger mountain: end the bar

Tracking connected components in superlevelsets



Tracking connected components in superlevelsets





Barcode

Barcode

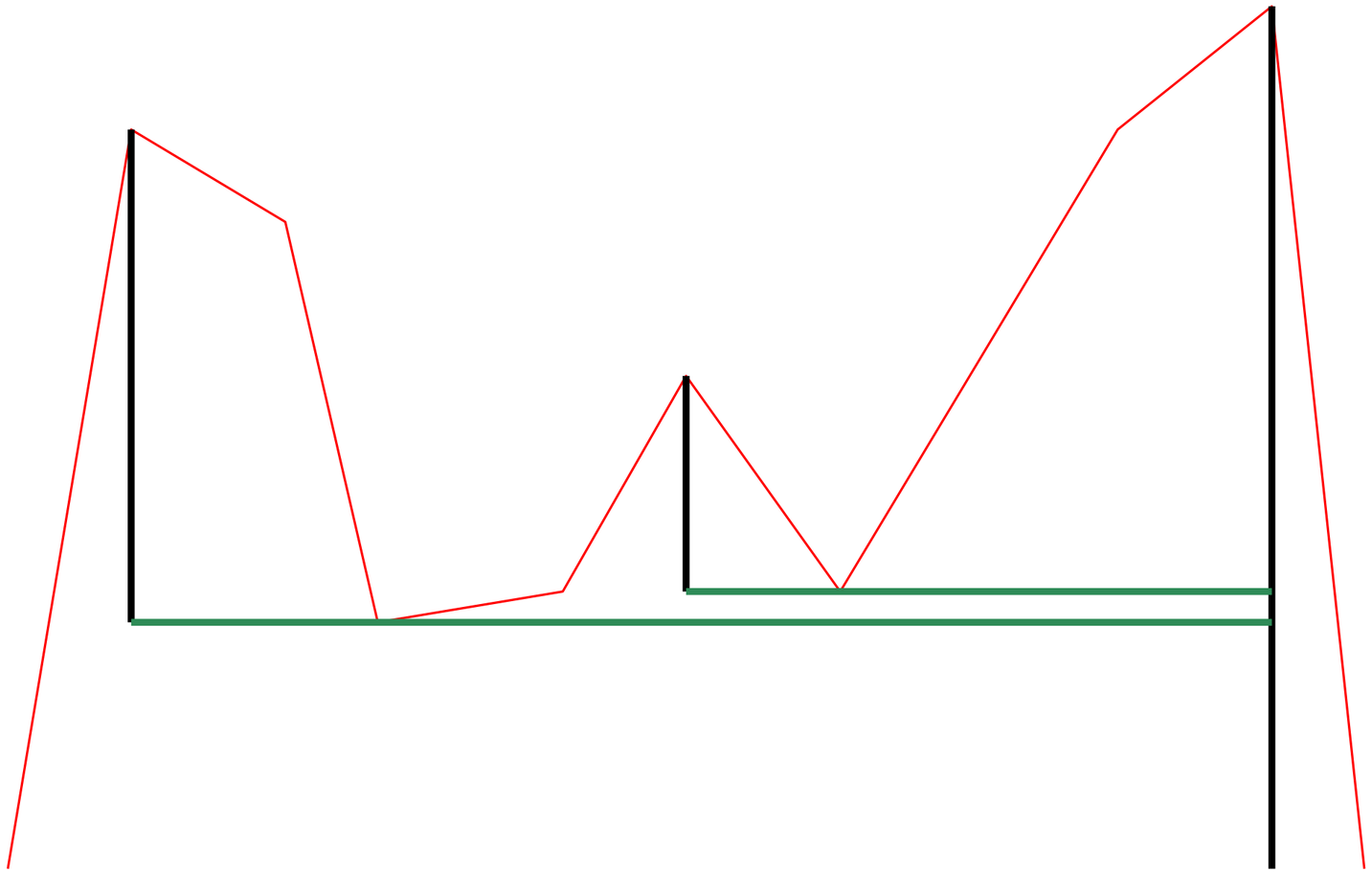
What it contains, for each local maximum

- height
- prominence

Things it ignores

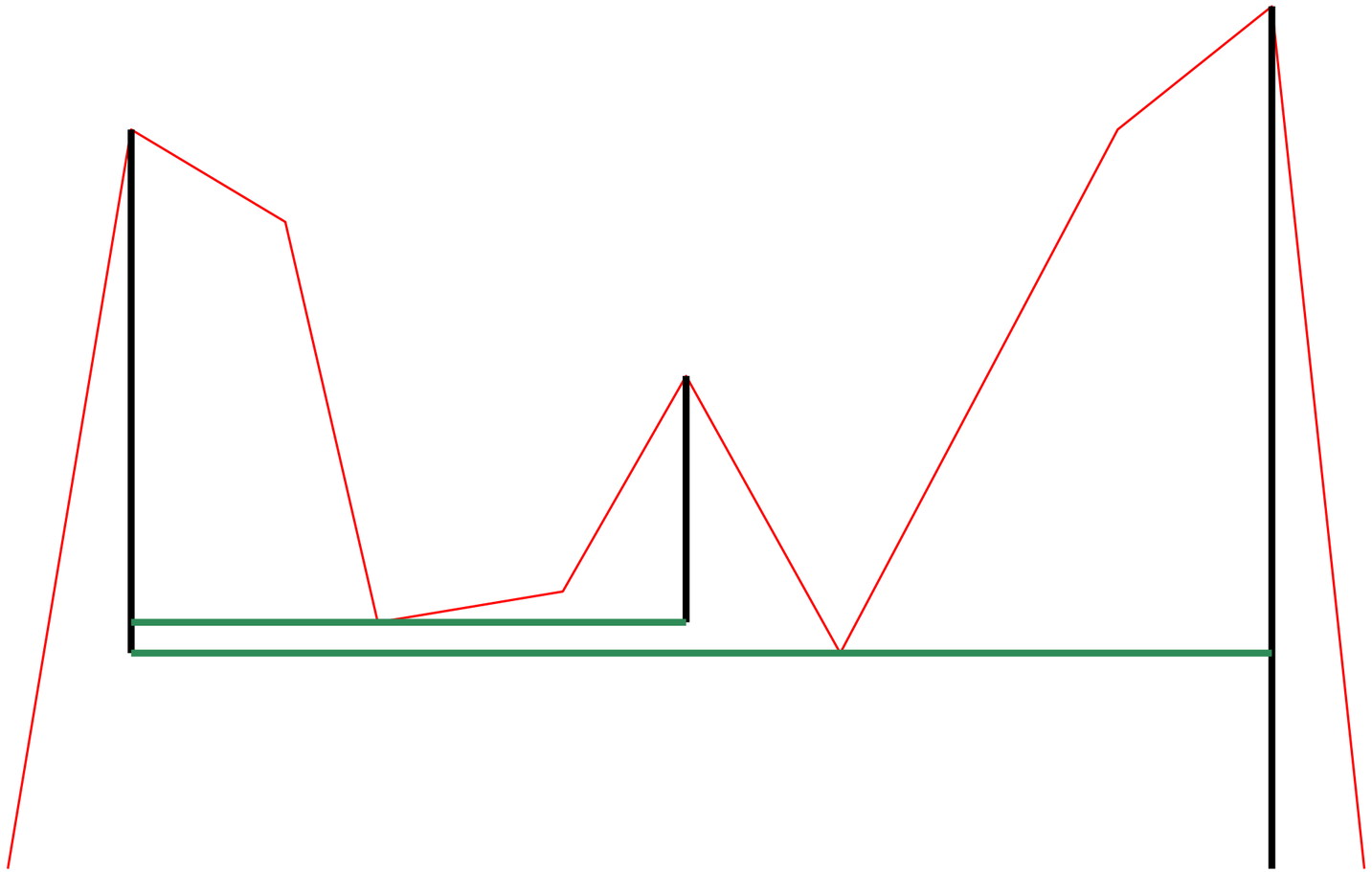
- position of the local maximum
- when a bar stops, which other bar it merges with
- what mountains are adjacent
- width of the mountains
- invariant by reparametrization

Why not the merge tree?



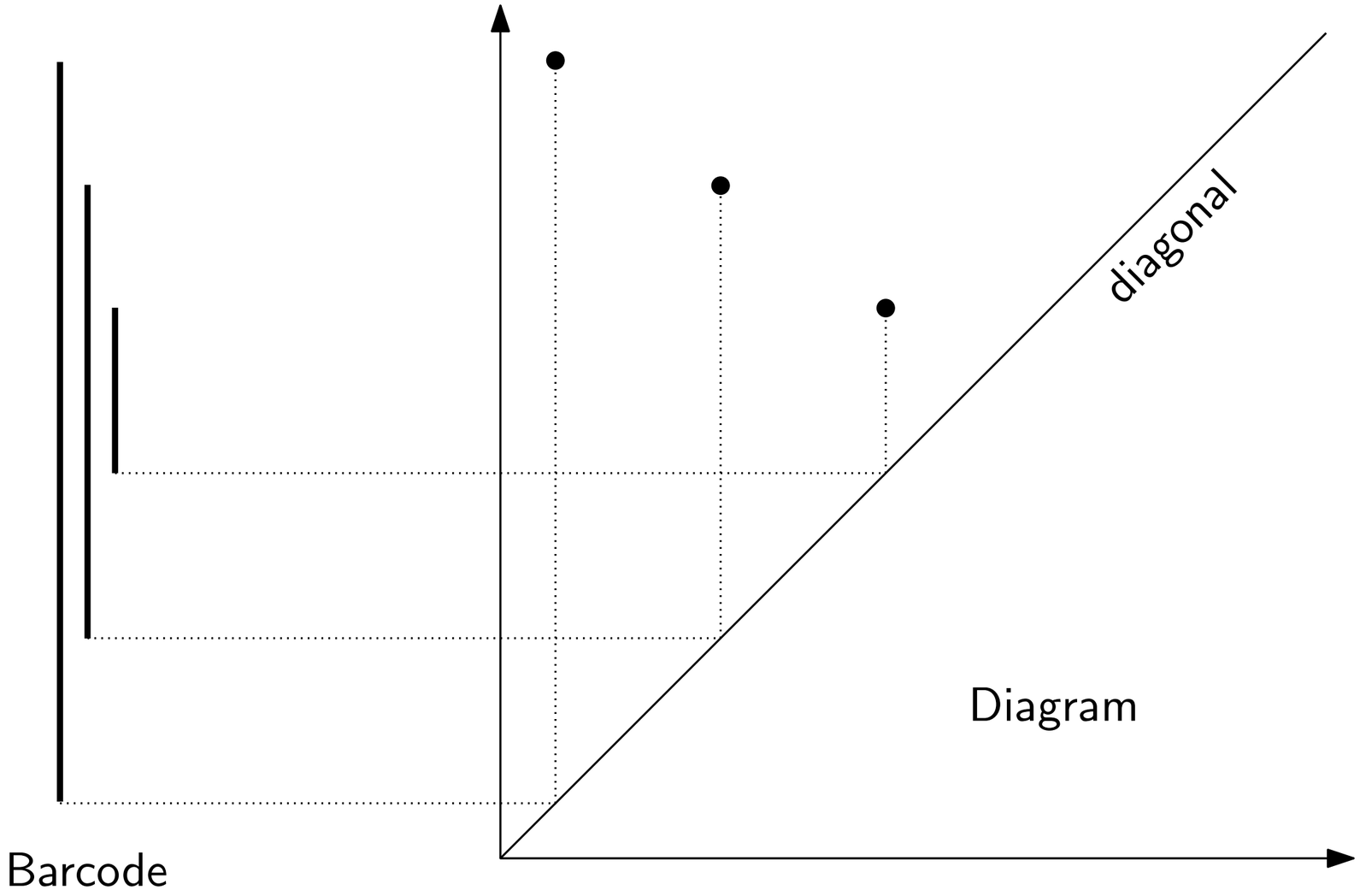
Unstable

Why not the merge tree?



Unstable

Persistence diagram



Interval (a, b)
$=$
Point (a, b)

Short bar
$=$
Point close to
the diagonal

(points with multiplicity)

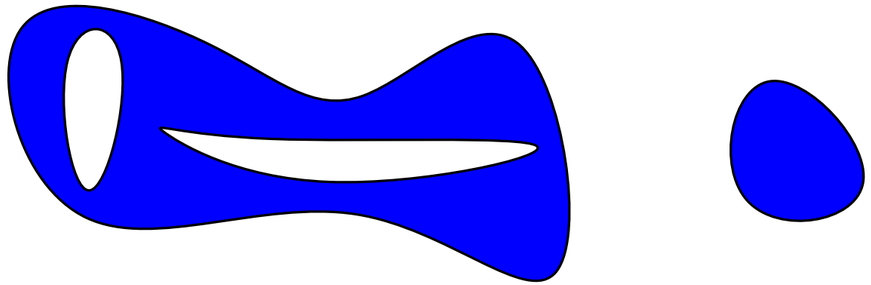
Higher dimension

Higher dimension

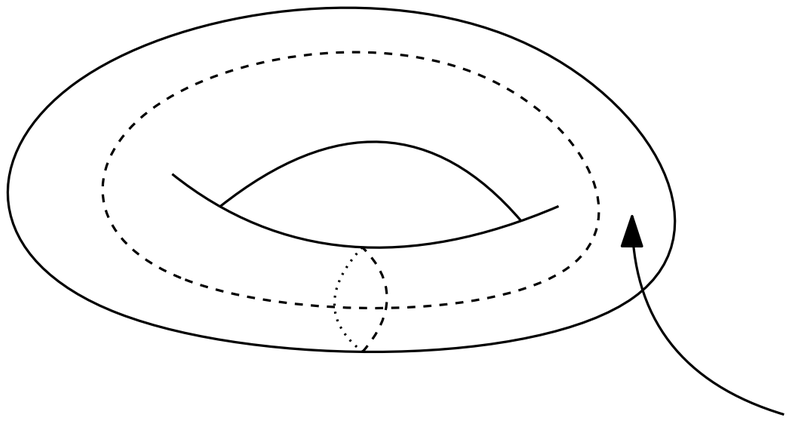
Higher dimension

Mathematical tool: homology

Defines "holes" of all dimensions. dim 0: connected components, dim 1: loops, etc.



2 connected components, 2 loops



1 connected component, 2 loops, 1 cavity

Higher dimension

$f : \mathbb{R}^2 \to \mathbb{R}$

Superlevelsets: sweep a horizontal plane

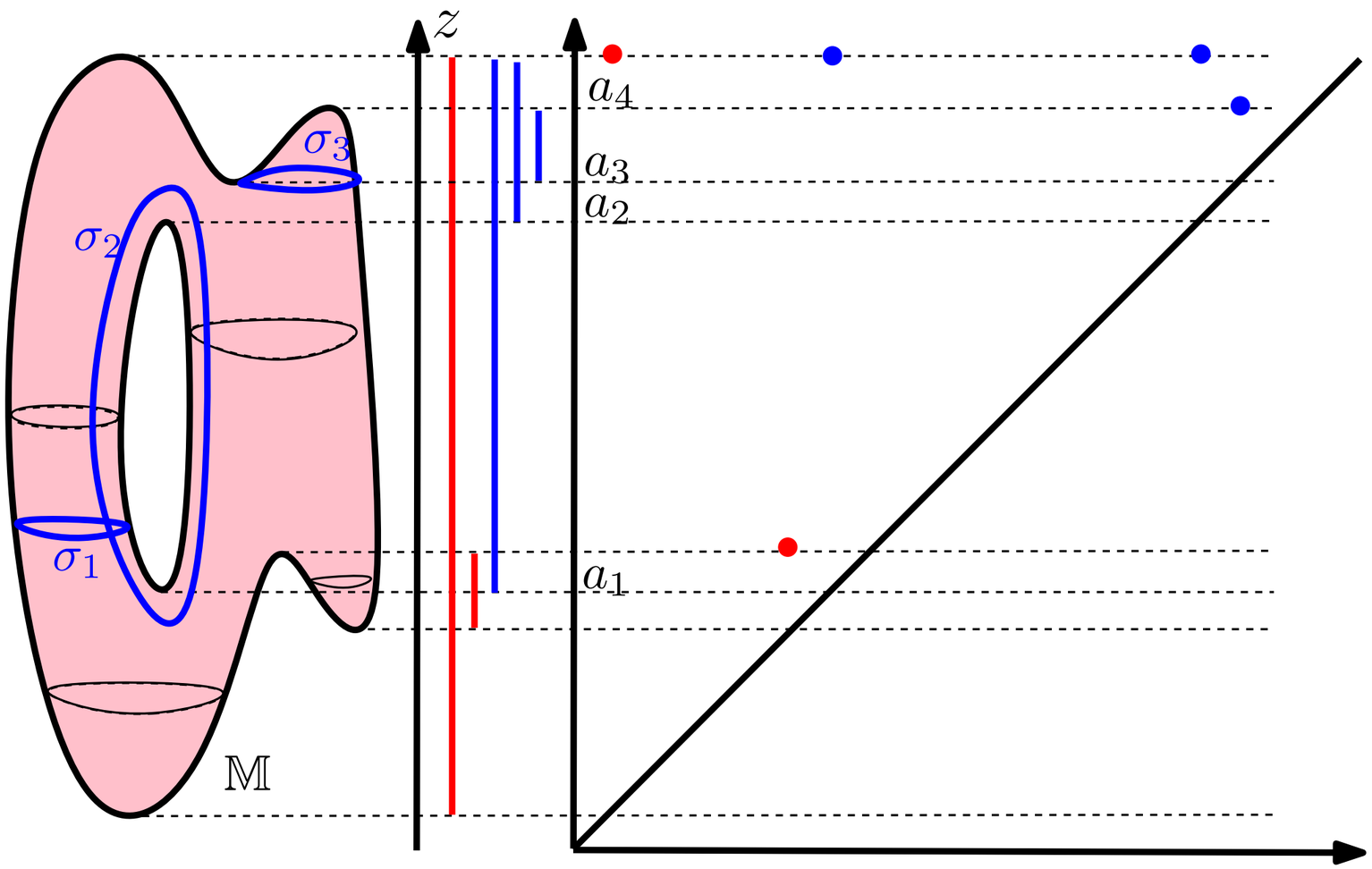
local maximum: new connected component

saddle point: merge 2 components, or create loop

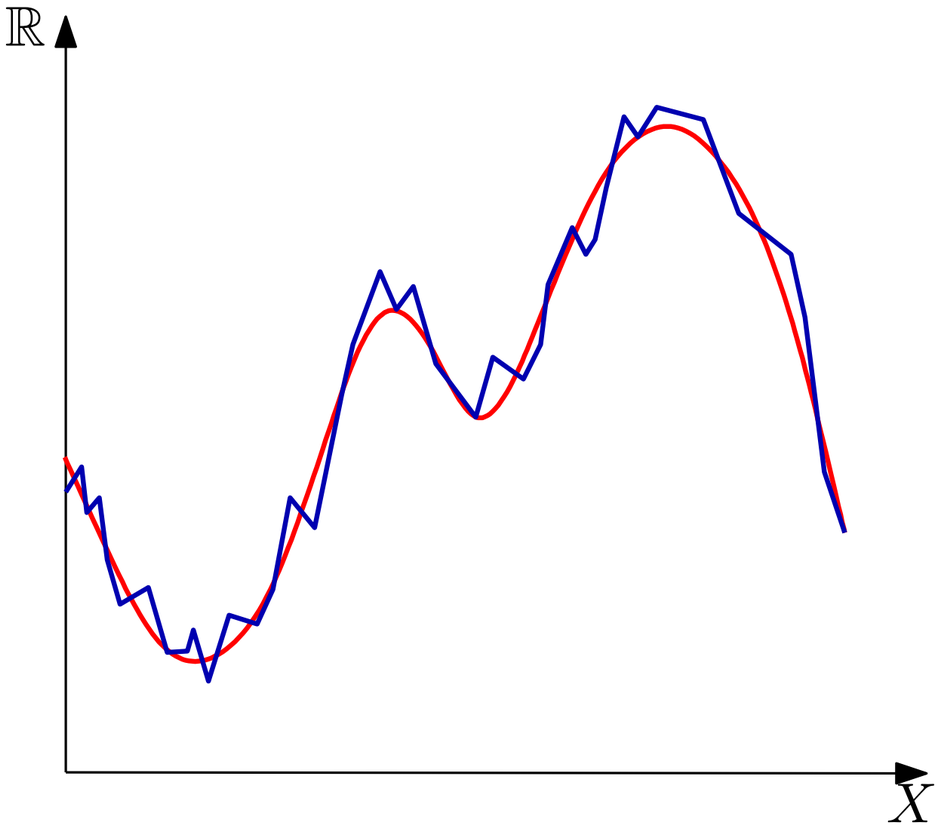
local minimum: kill (fill) loop

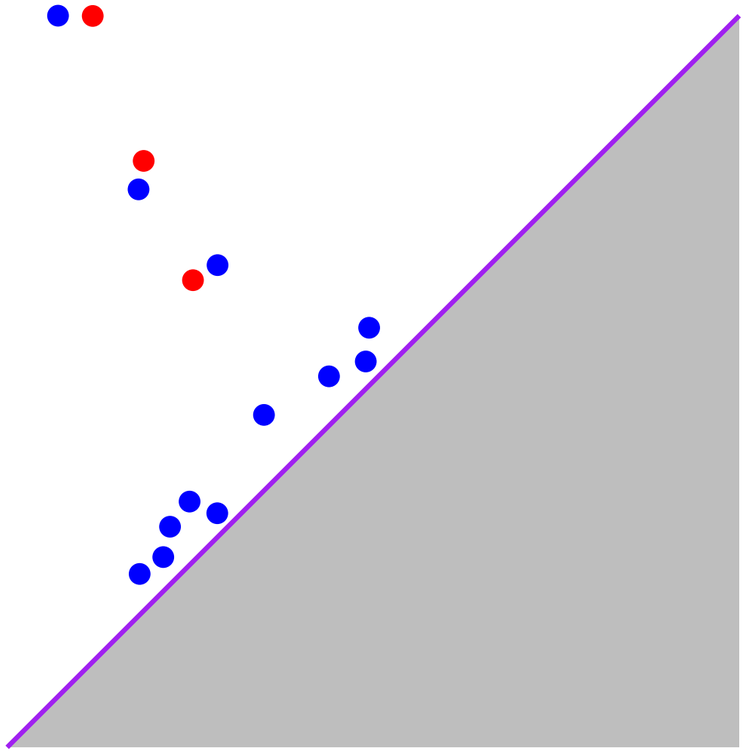
One barcode / diagram per dimension (often drawn together in different colors)

Higher dimension

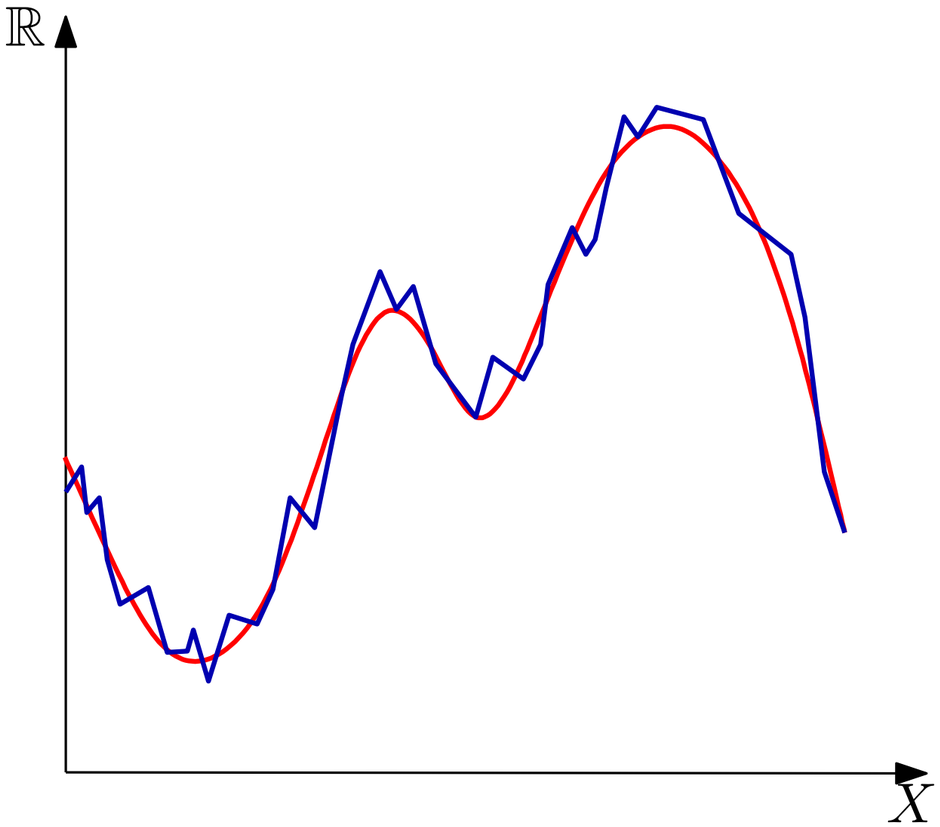


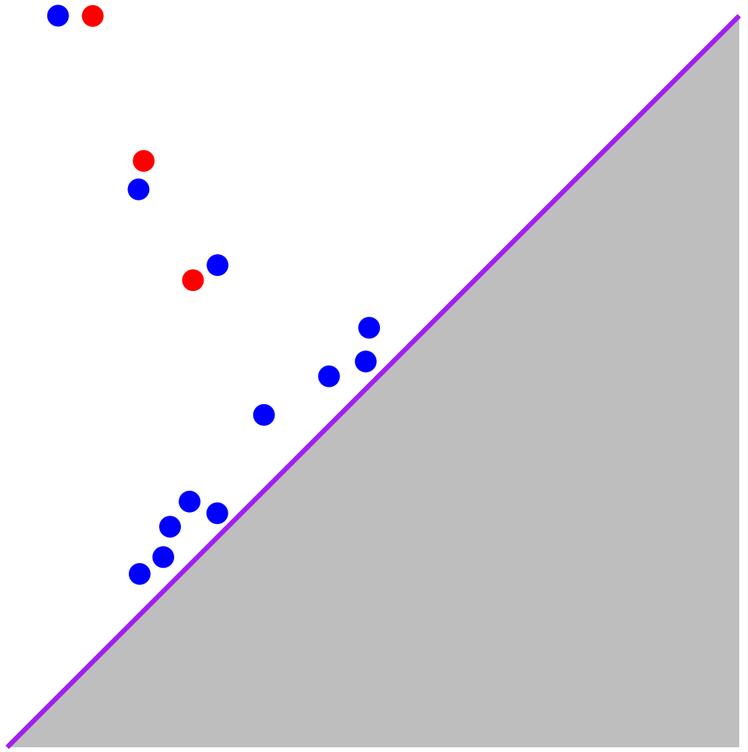
Stability





Stability





$$||f - g||_\infty = \sup_{x \in X} \{|f(x) - g(x)|\}$$

Bottleneck distance

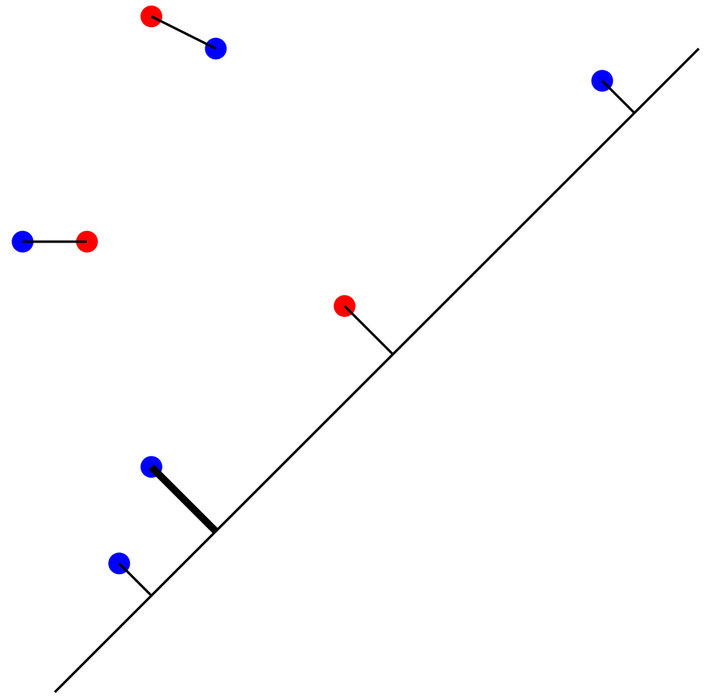
Partial matching, the rest matched with the diagonal

The worst pair defines the cost

Sup norm between points: $\max(|x - x'|, |y - y'|)$

Minimum over all matchings ($\sim$ Wasserstein W_∞)

Can also define other distances W_p



Stability Theorem

$$d_B(Dgm(f), Dgm(g)) \leq ||f - g||_\infty$$

Dgm is 1-Lipschitz

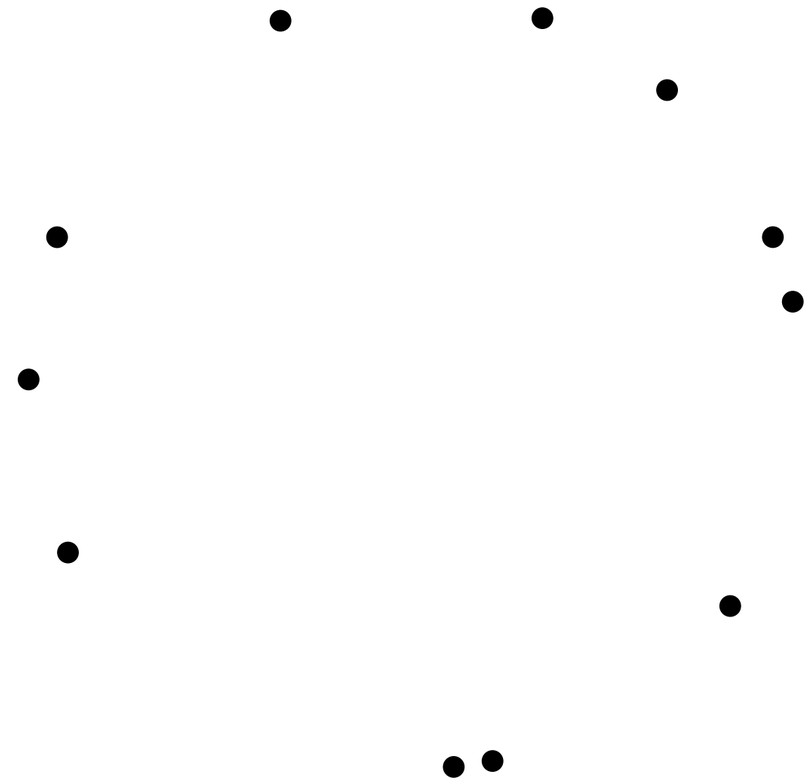
Independent (?) problem: point clouds

Input: point set P

Assumption: P approaches some unknown ideal object

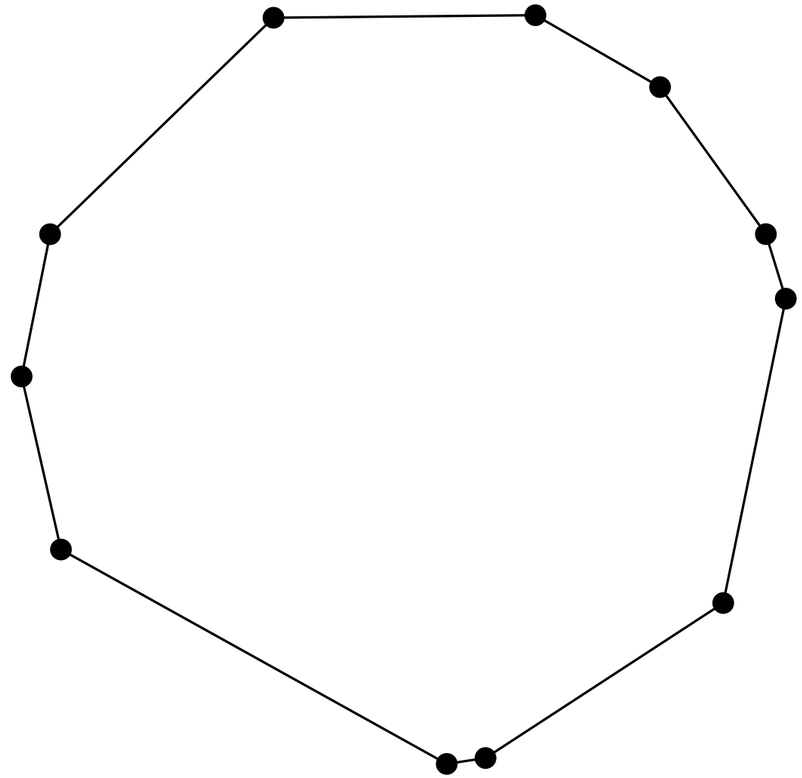
What can we do?

Strong reconstruction



Strong reconstruction

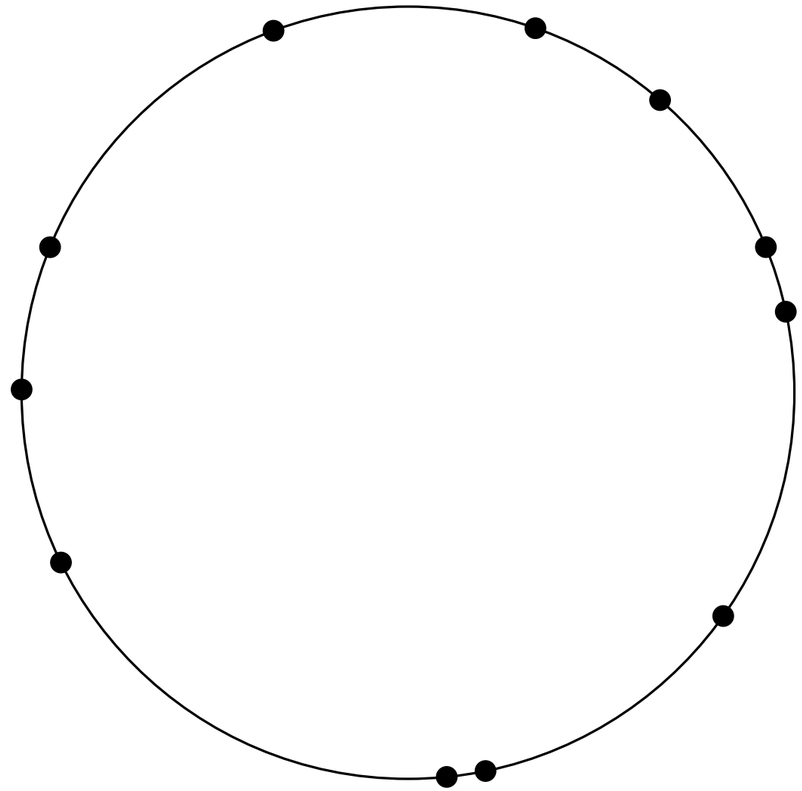
"Connect the dots" homeomorphic reconstruction



Strong reconstruction

"Connect the dots" homeomorphic reconstruction

Diffeomorphism?

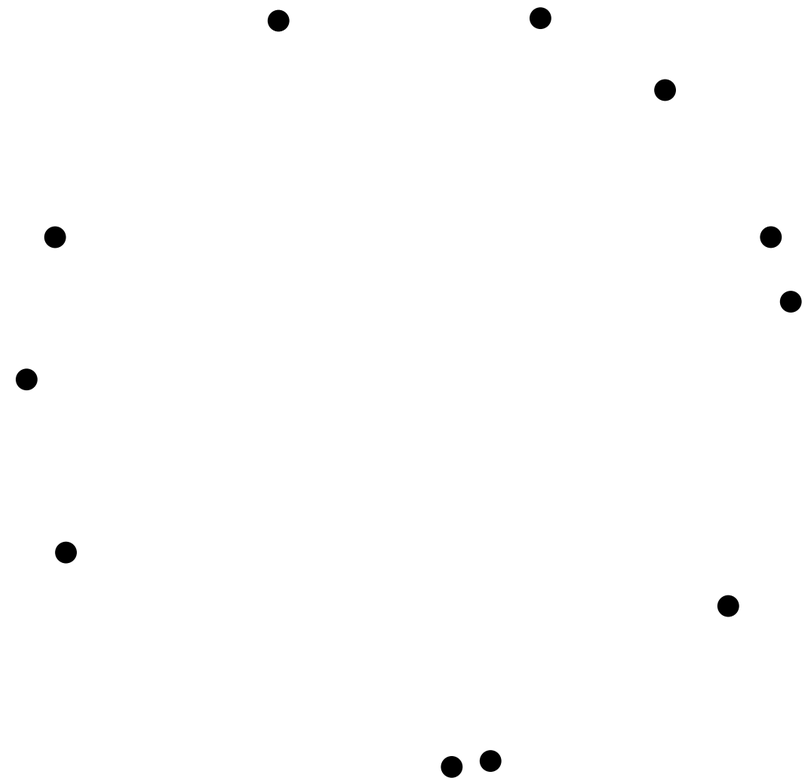


Strong reconstruction

"Connect the dots" homeomorphic reconstruction

Diffeomorphism?

Requires a nice sampling



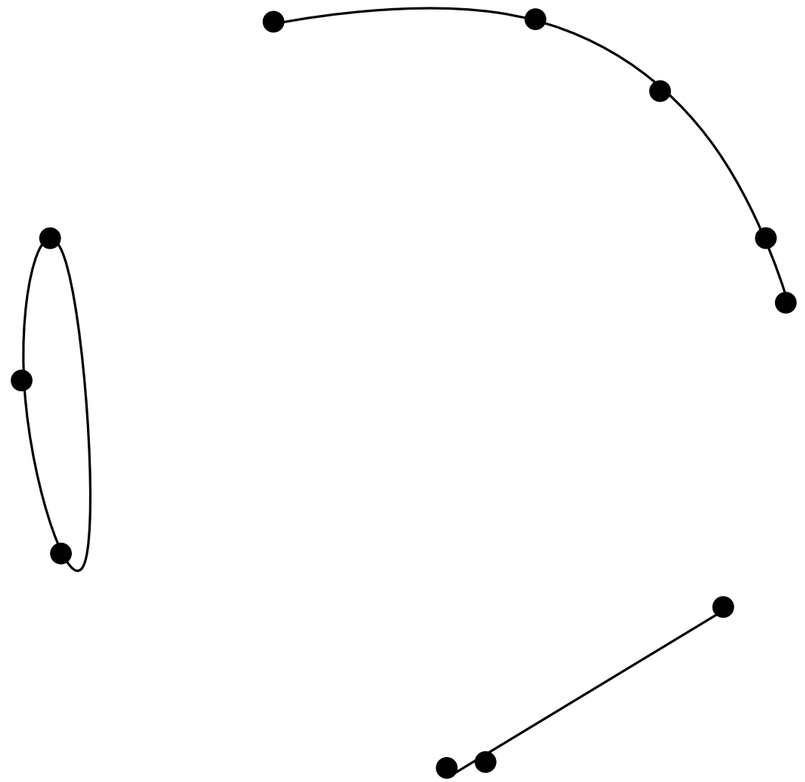
Strong reconstruction

"Connect the dots" homeomorphic reconstruction

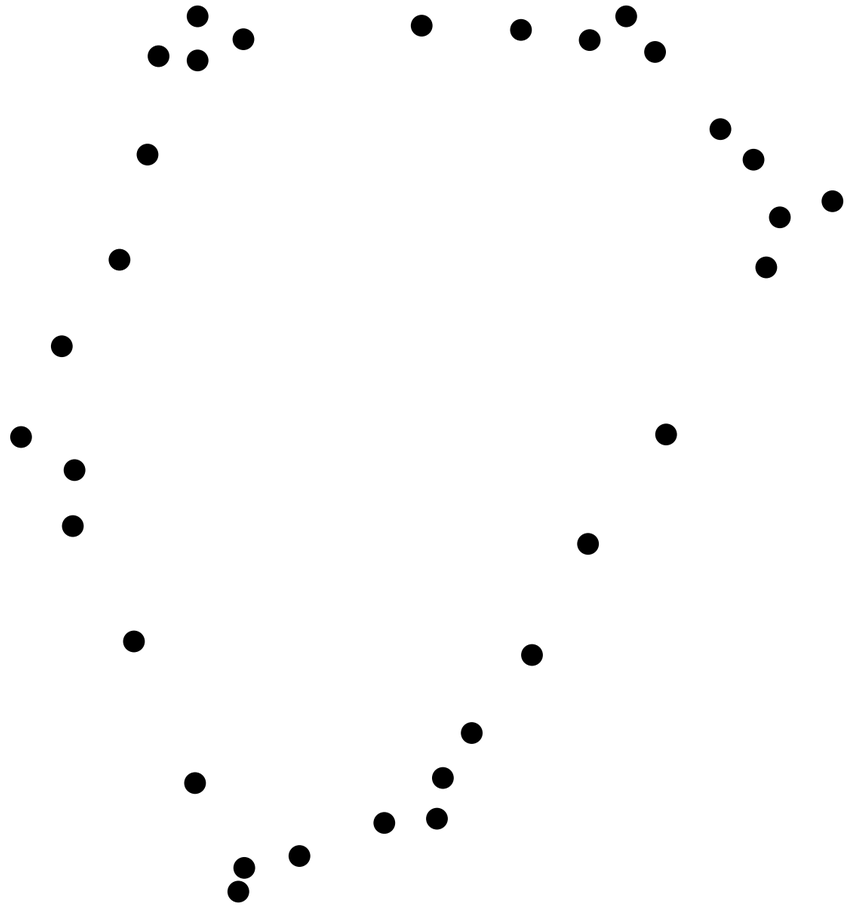
Diffeomorphism?

Requires a nice sampling

Requires hypotheses on the model



Weaker reconstruction

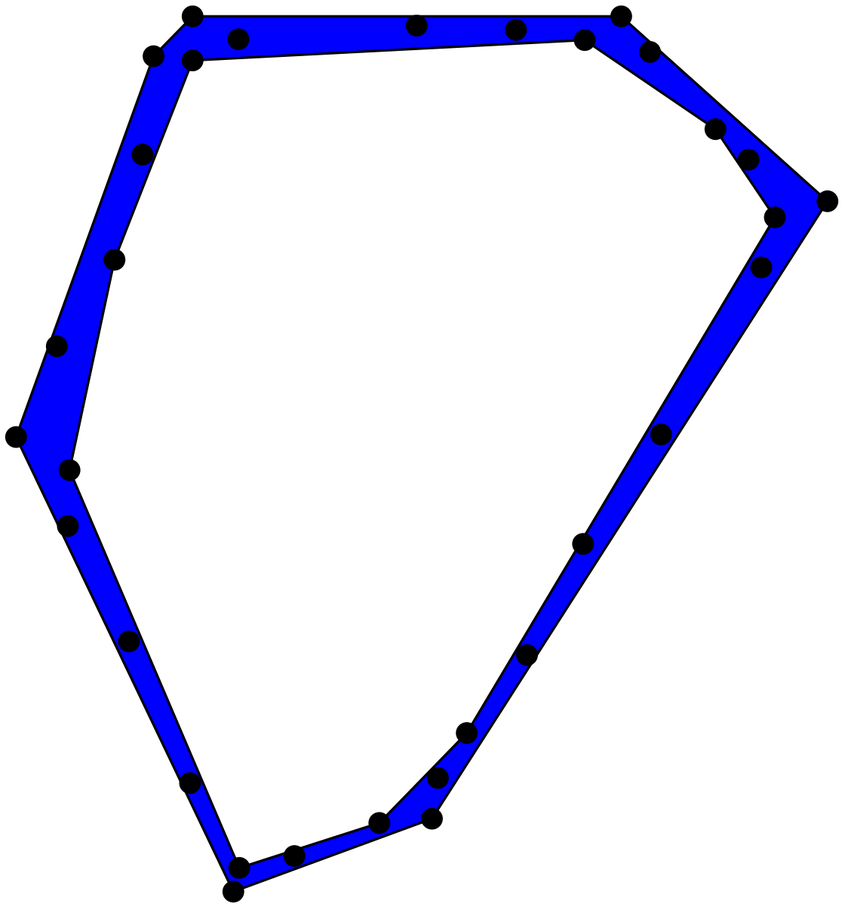


Weaker reconstruction

Thickened version of the object

Not homeomorphic to a circle

Same homotopy type



Even weaker

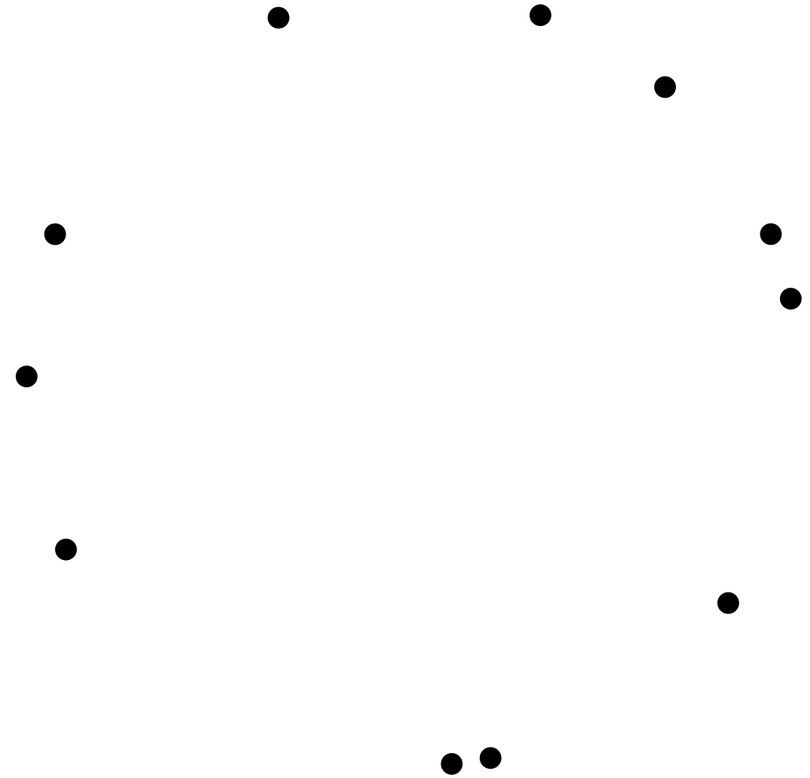
Clustering

Mapper (graph) – next class

Persistent homology

Topology of points?

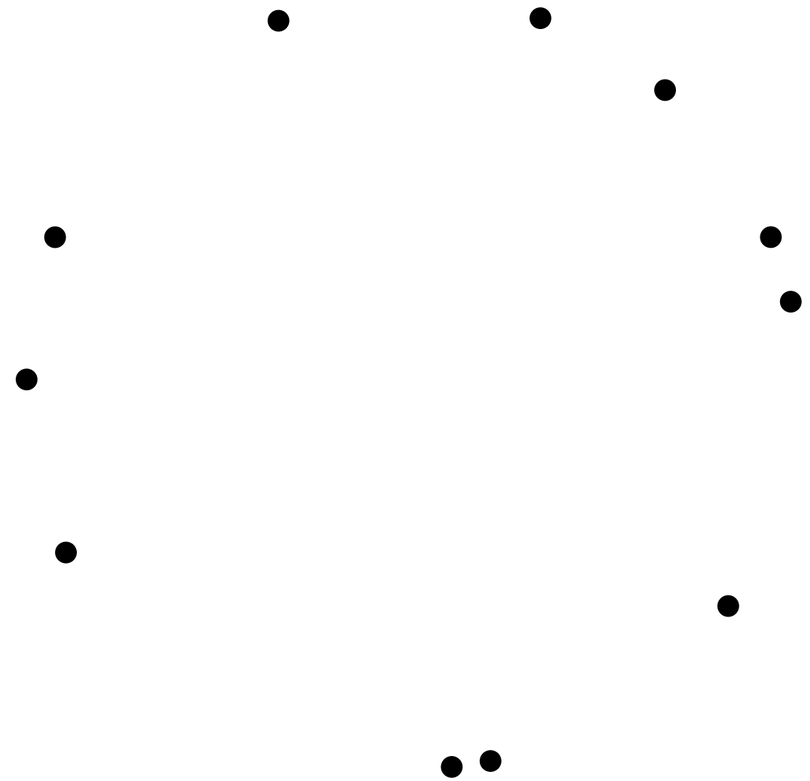
Just n connected components...



Topology of points?

Just n connected components...

Walk back, blur

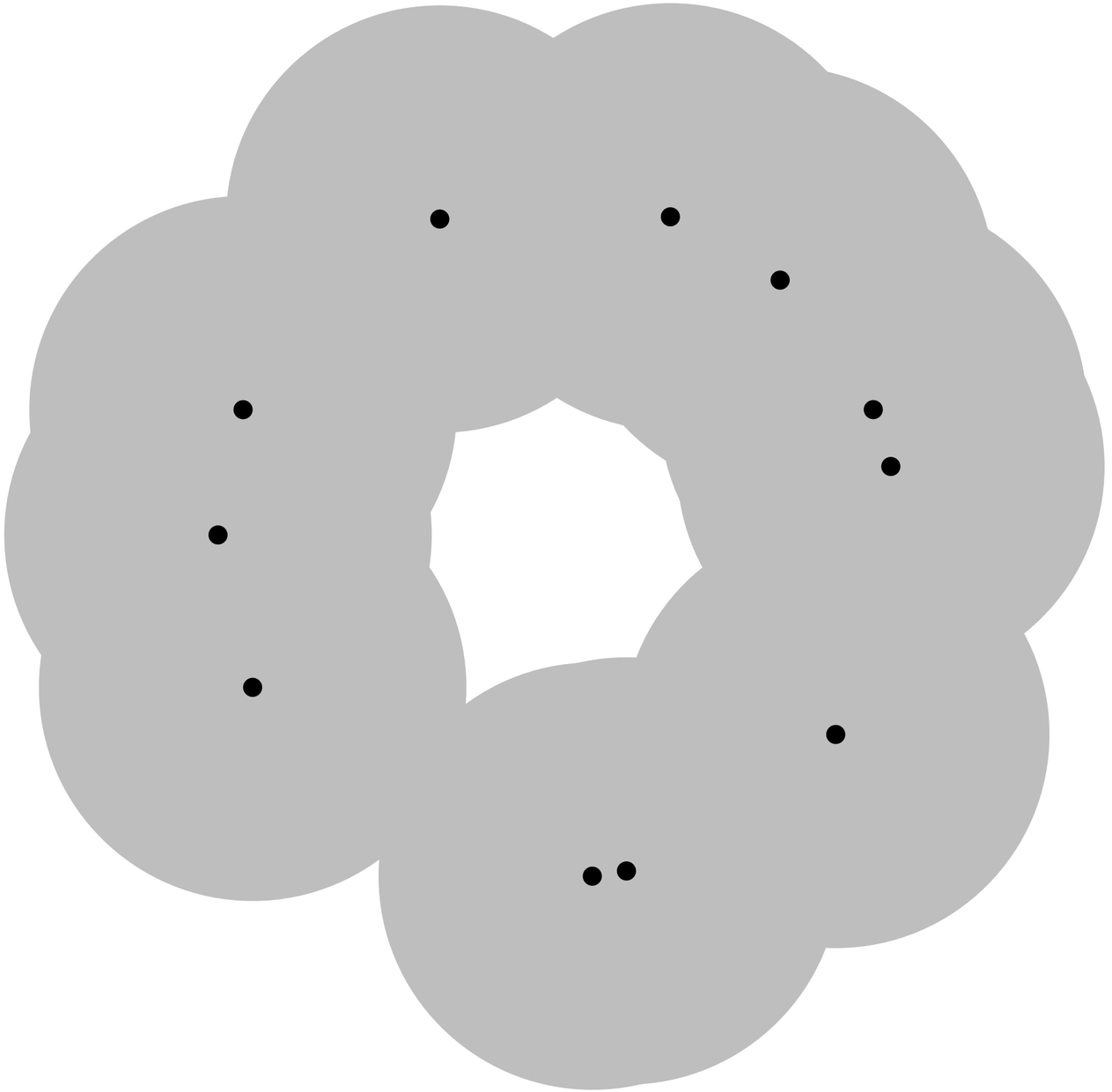


Topology of points?

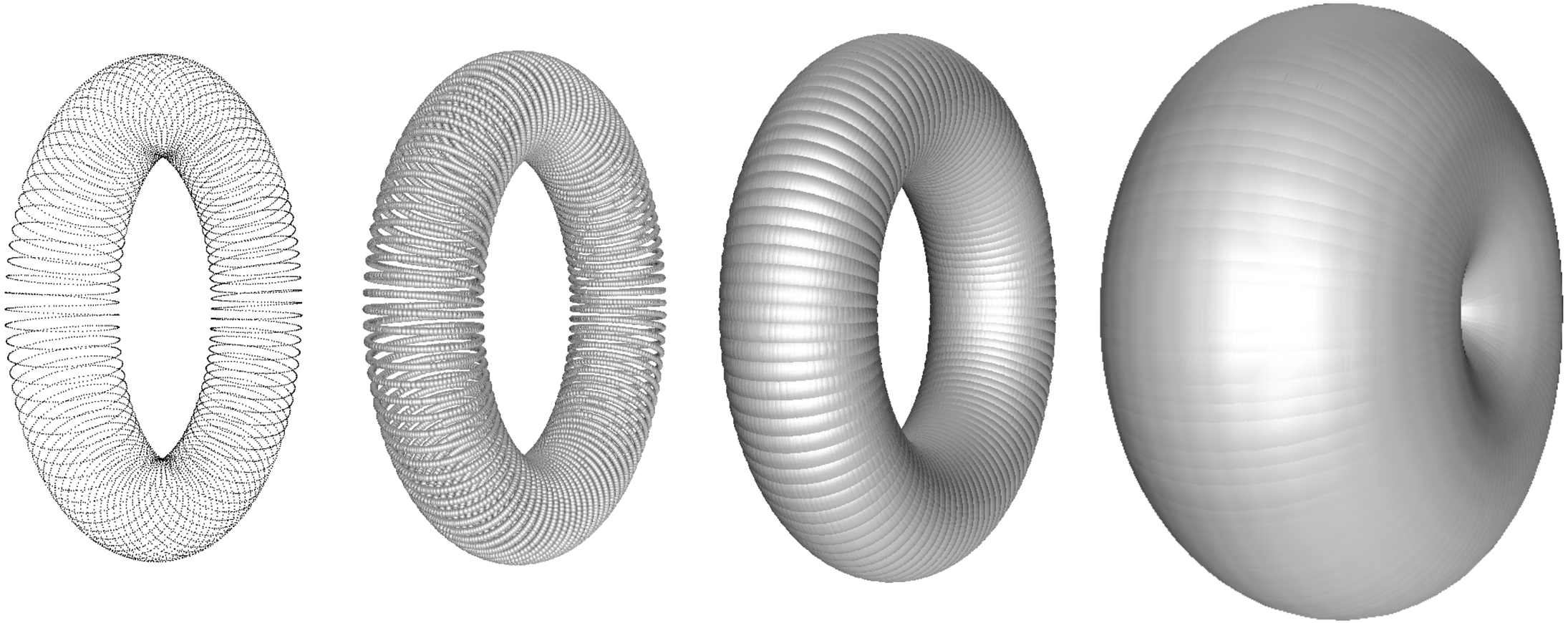
Just n connected components...

Walk back, blur

1 connected component, 1 loop



Choosing the scale



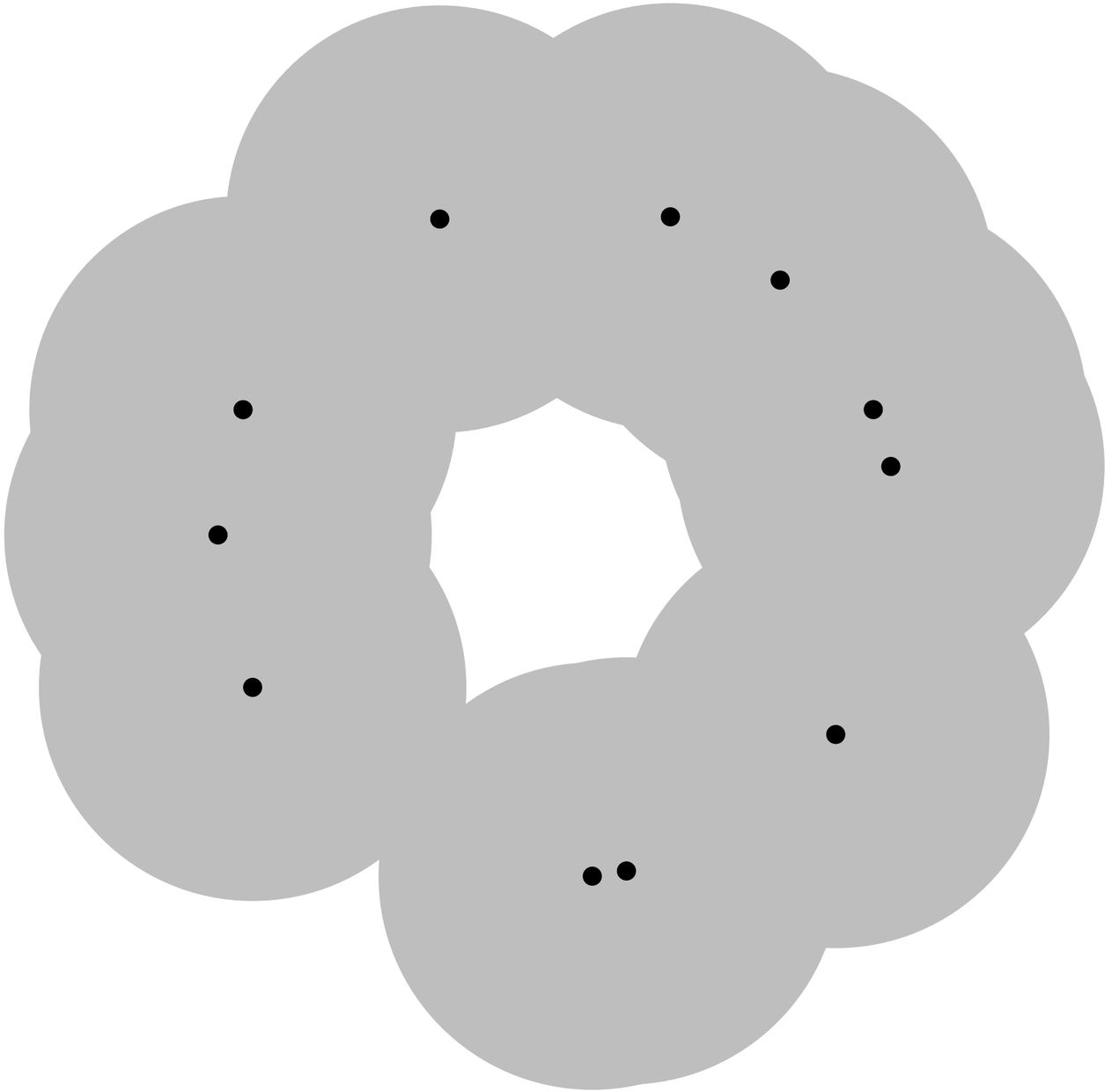
Ill defined problem

$\Longrightarrow$ Look at all scales!

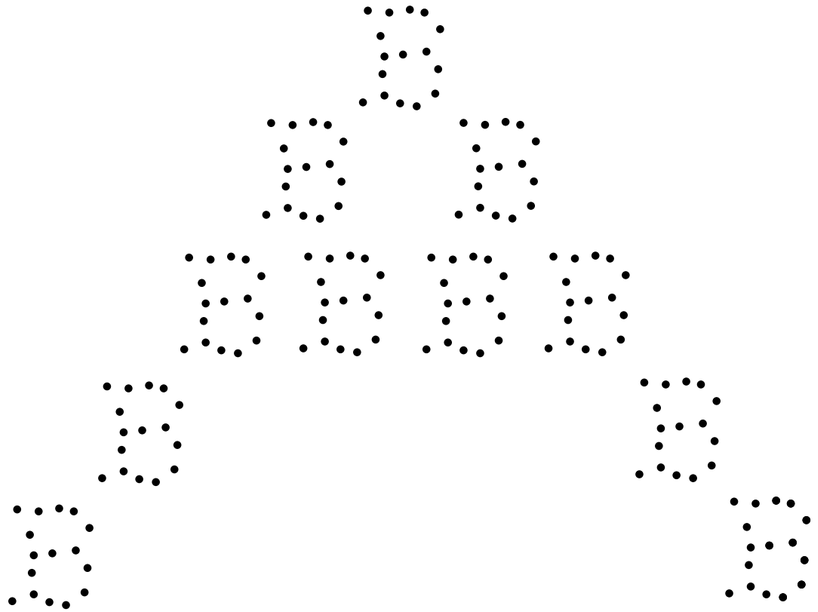
Link with functions

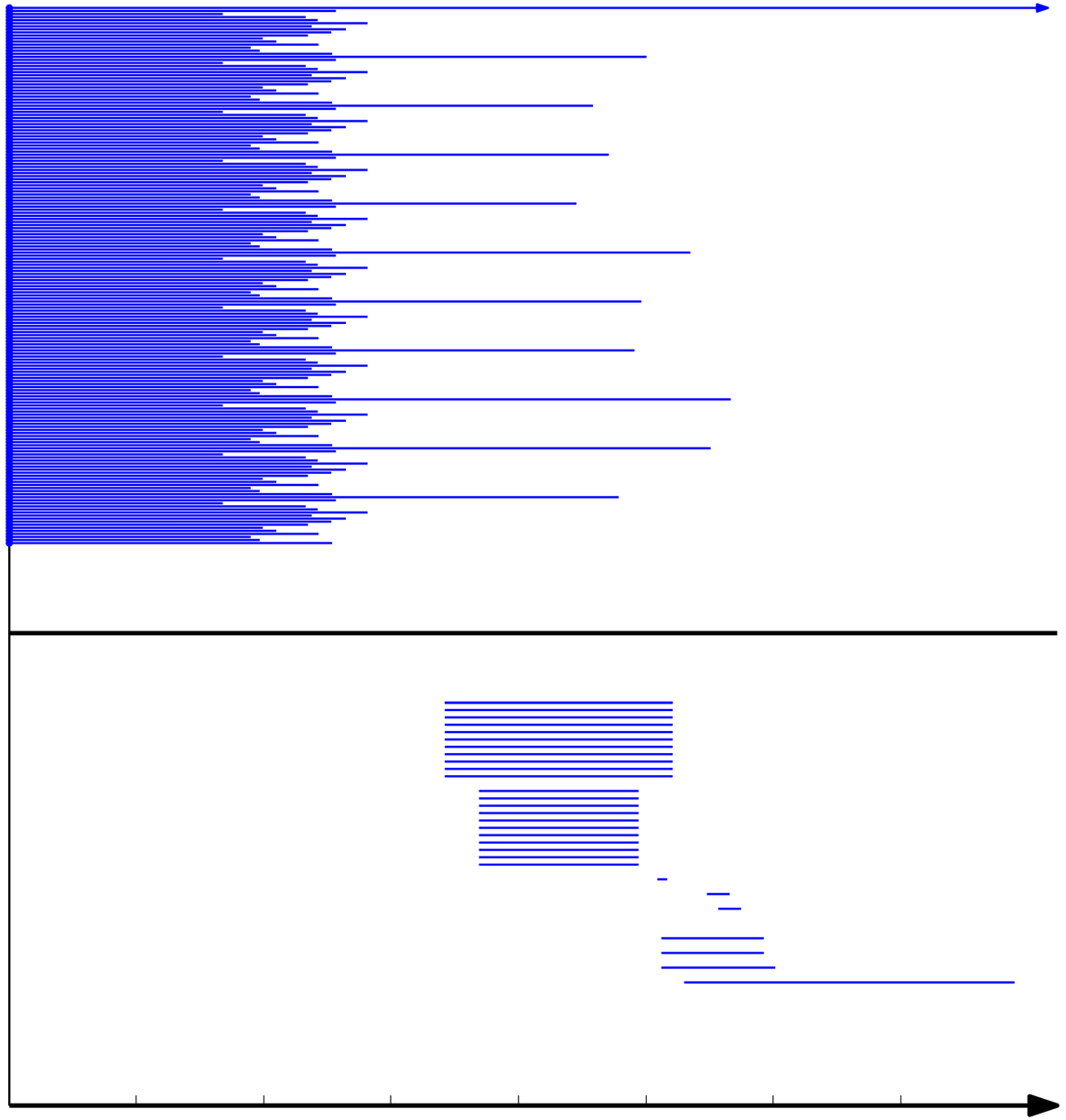
$f : \mathbb{R}^d \to \mathbb{R}$
$f(x) = \min_{p \in P} ||x - p||_2$

Union of balls = **sub**levelset of f

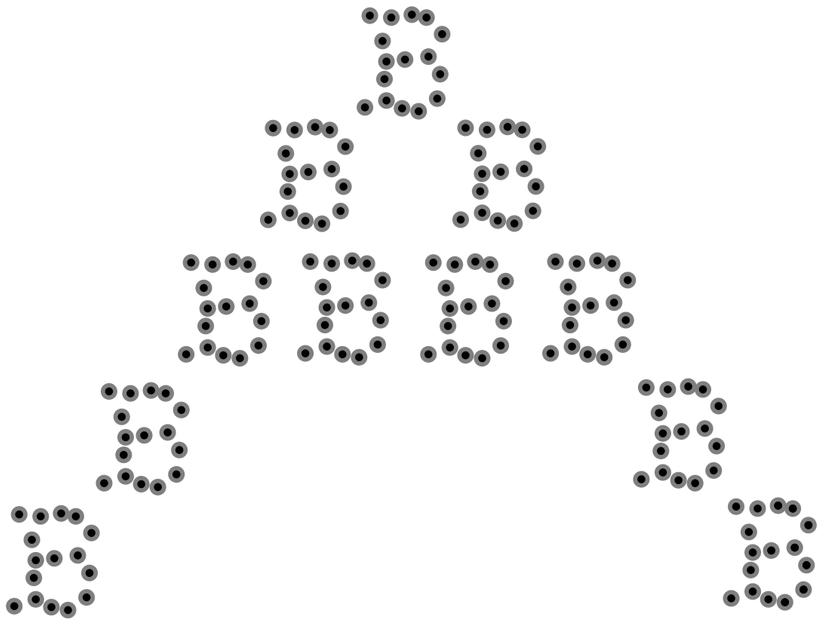


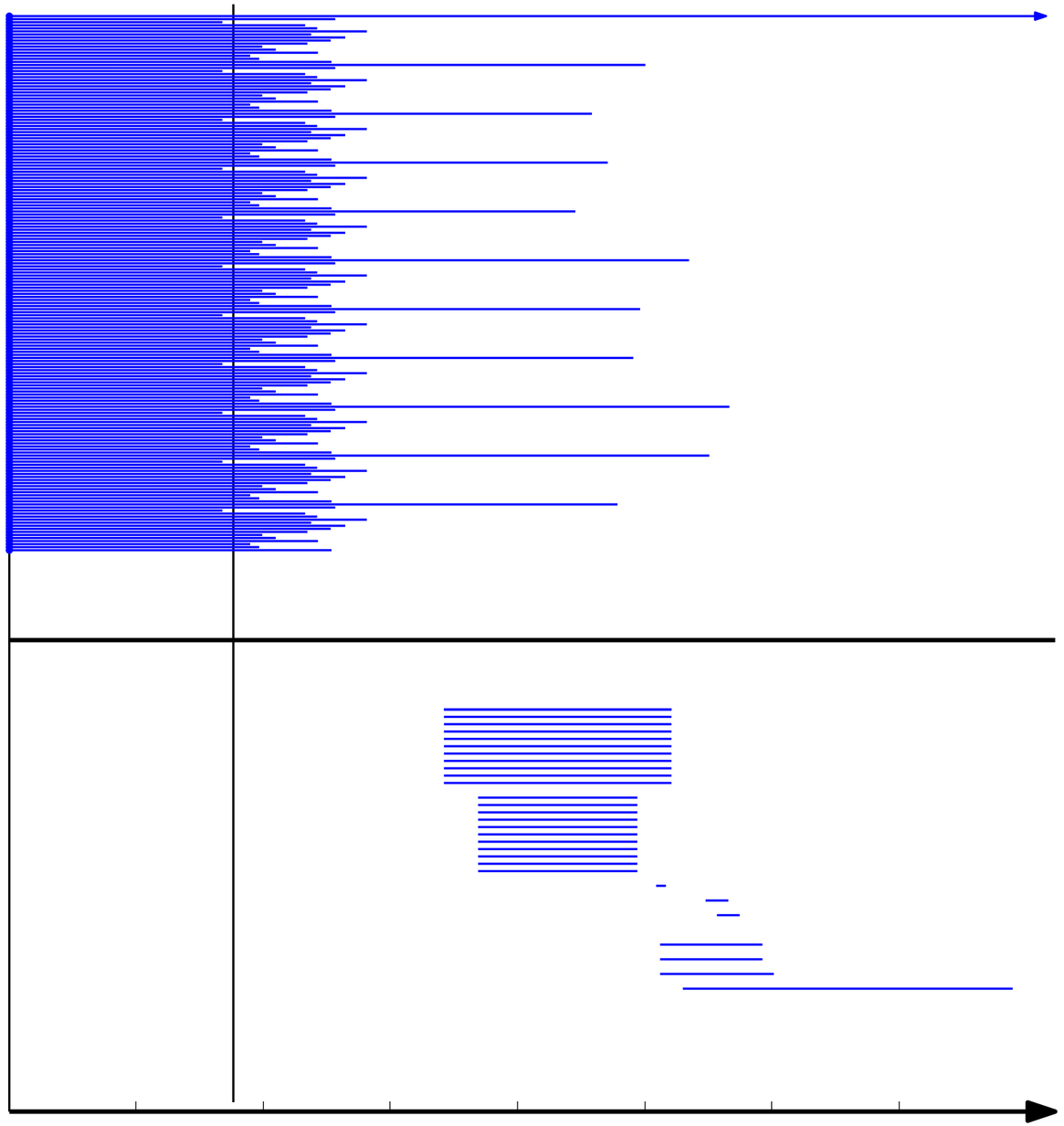
Persistence of offset filtration



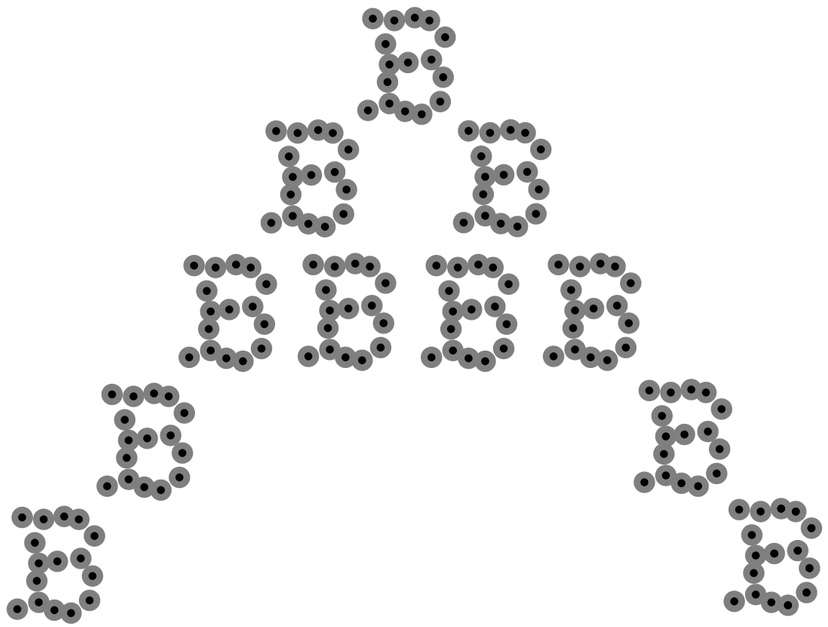


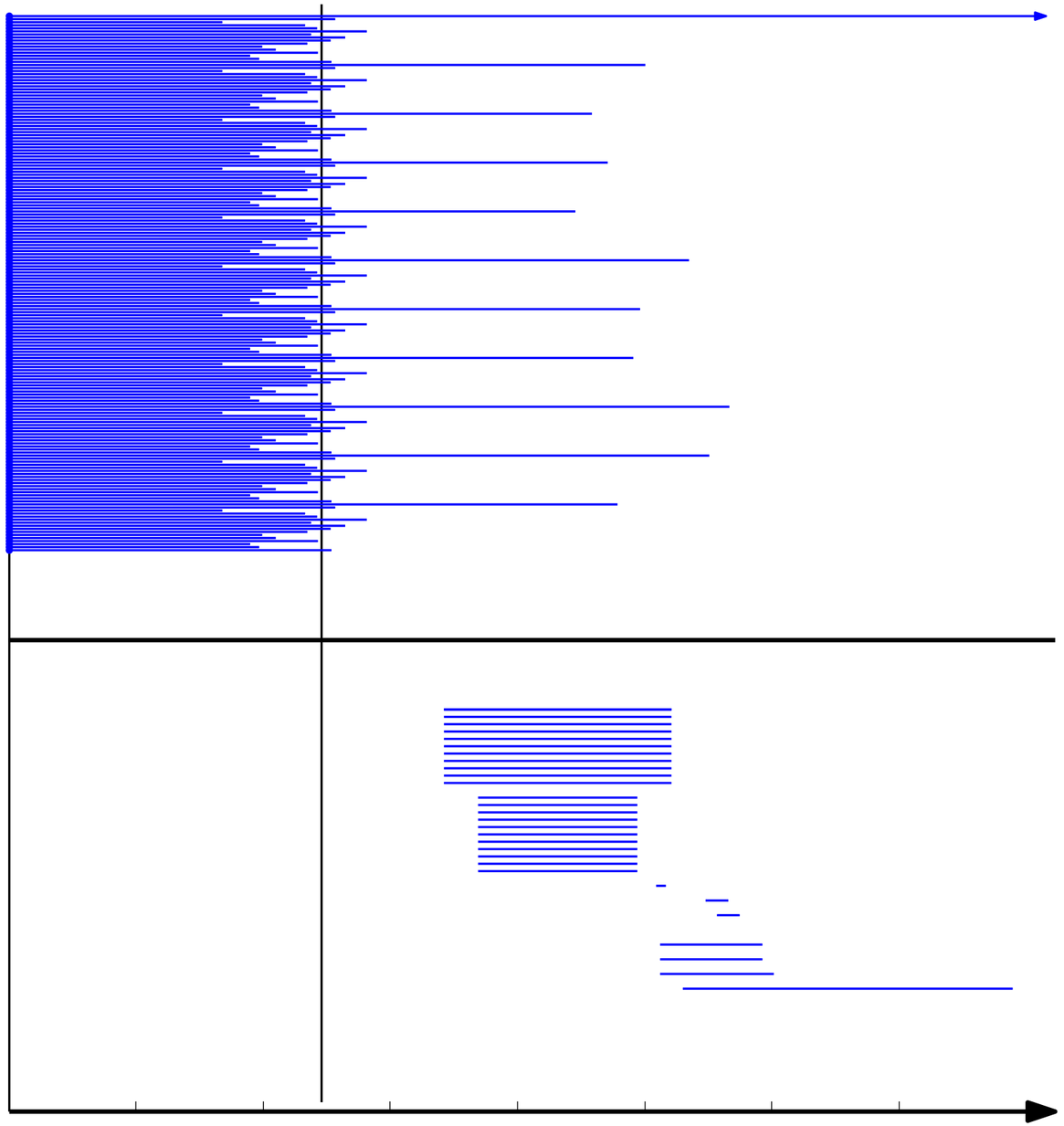
Persistence of offset filtration



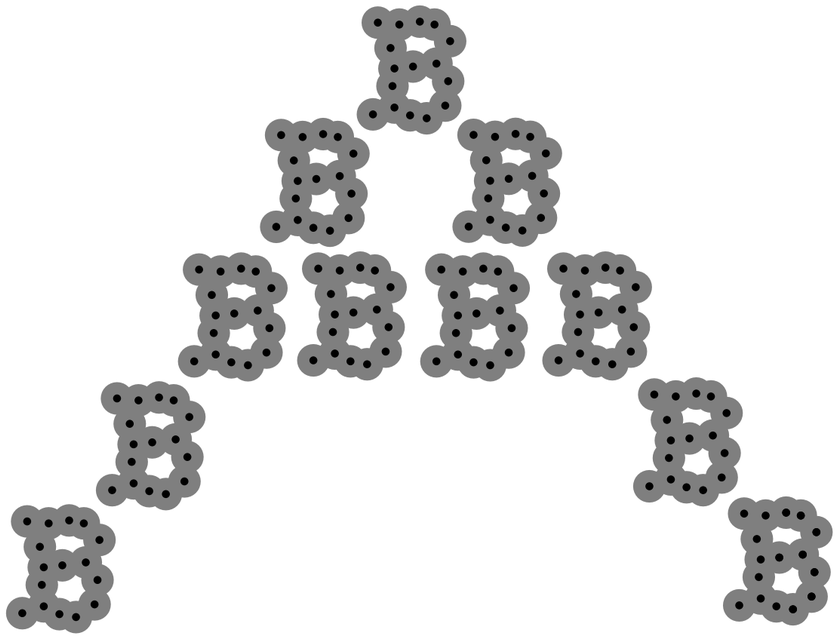


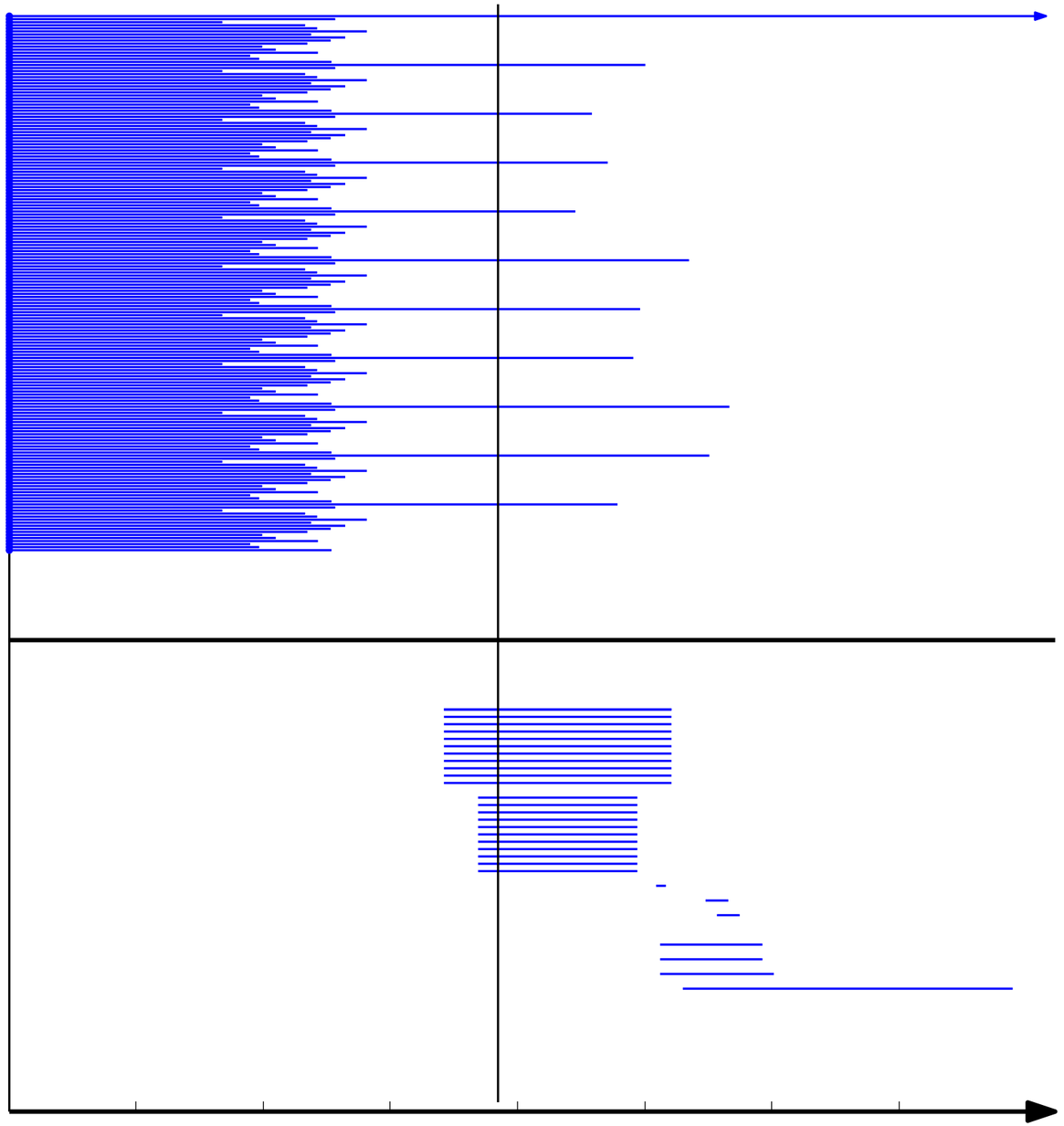
Persistence of offset filtration



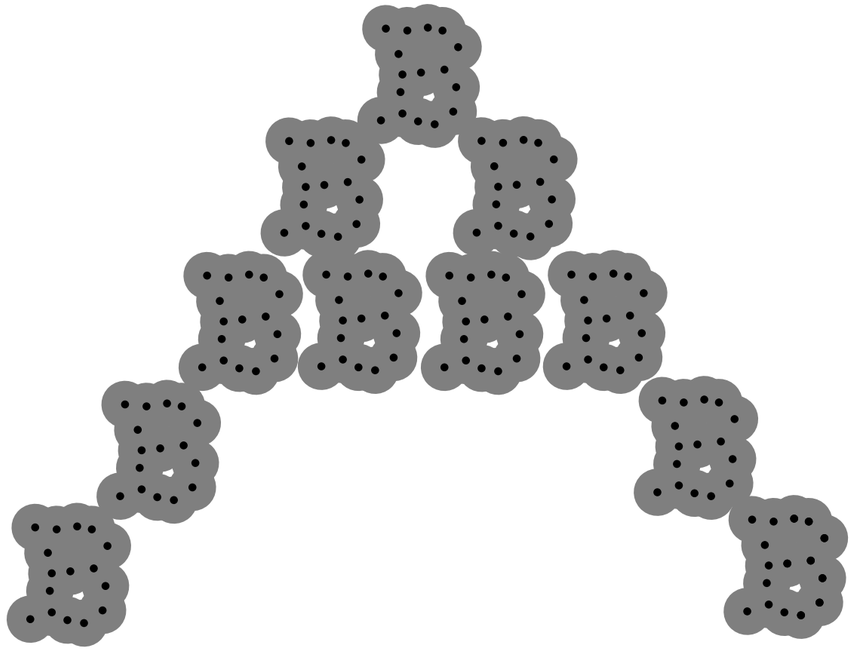


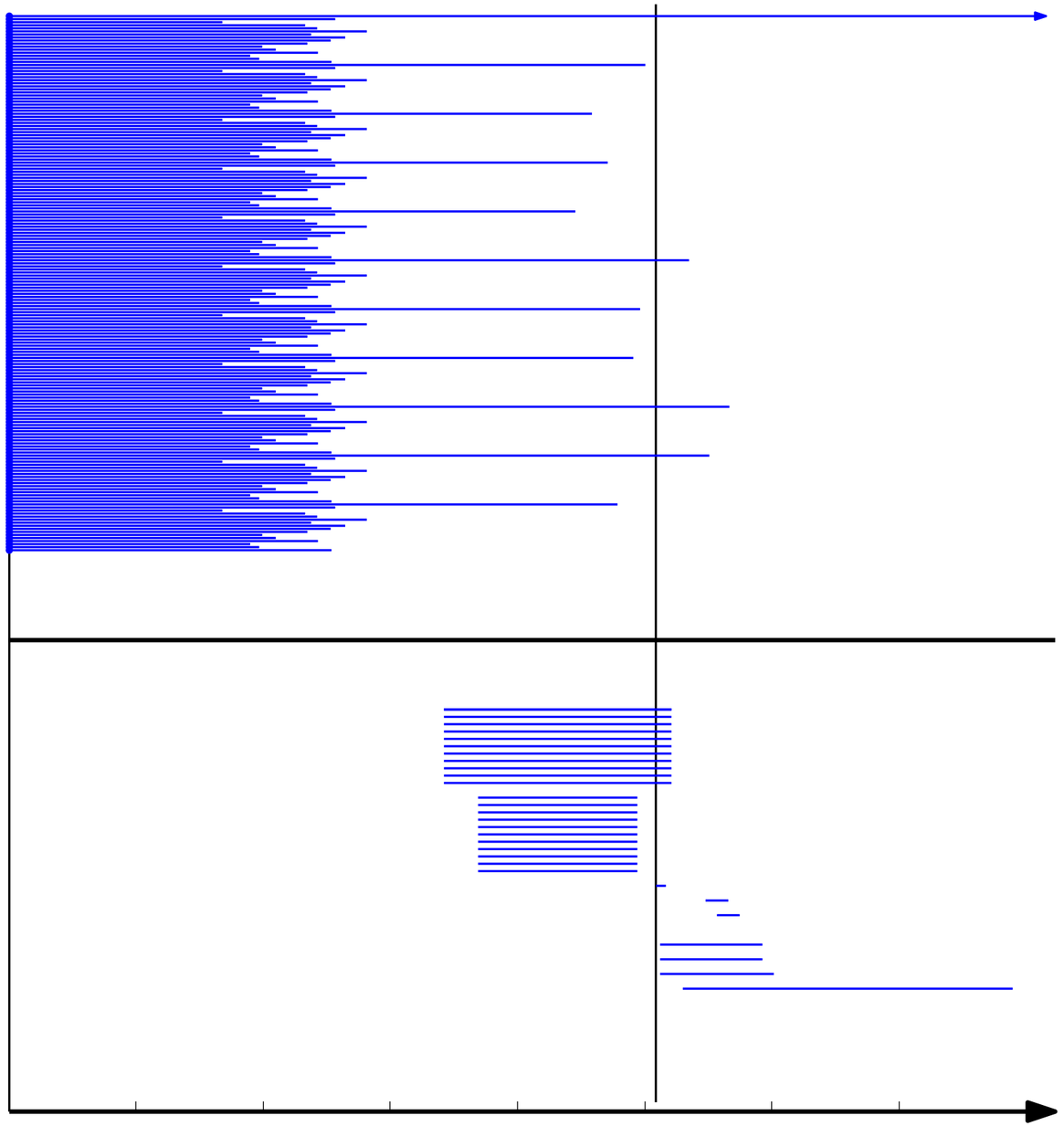
Persistence of offset filtration



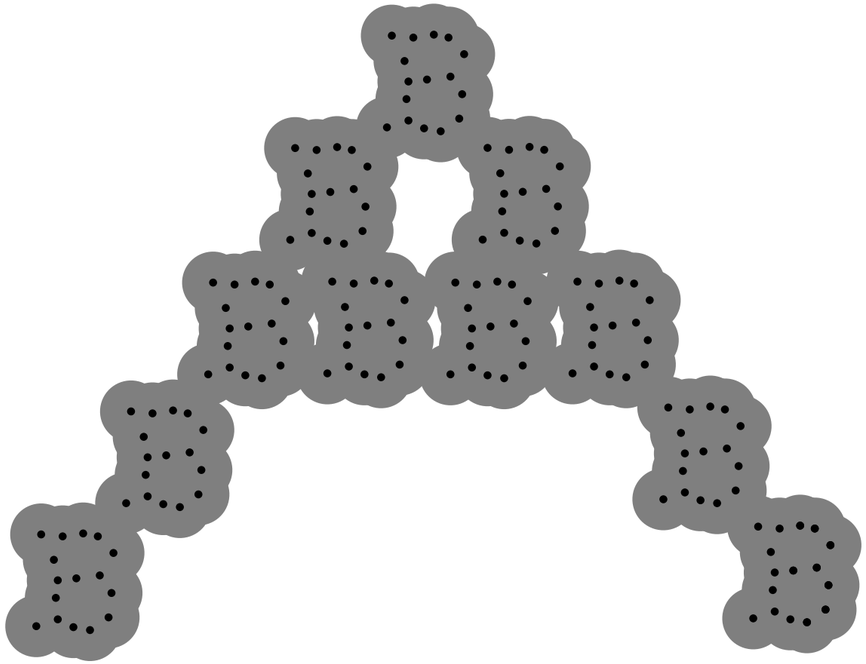


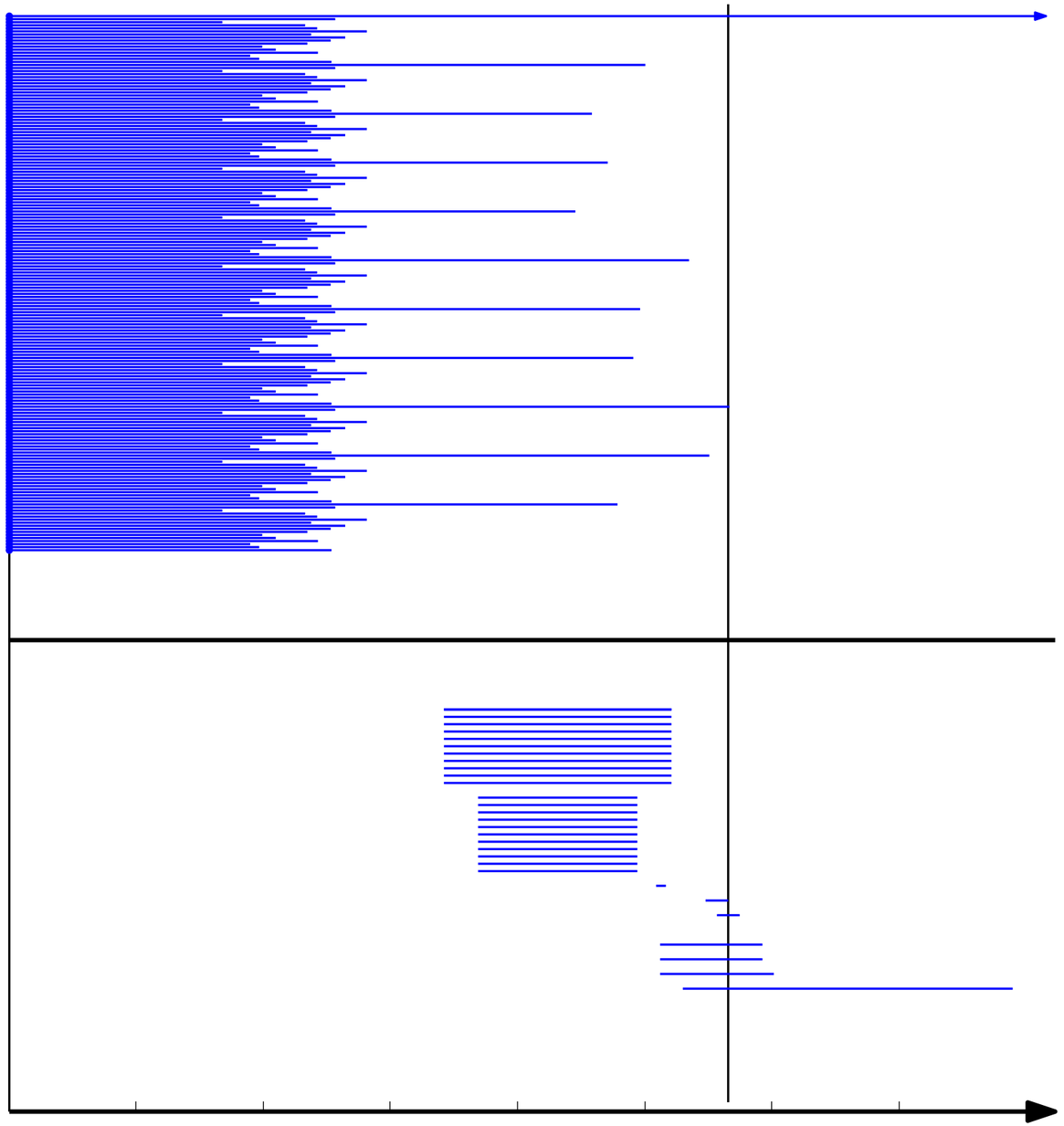
Persistence of offset filtration



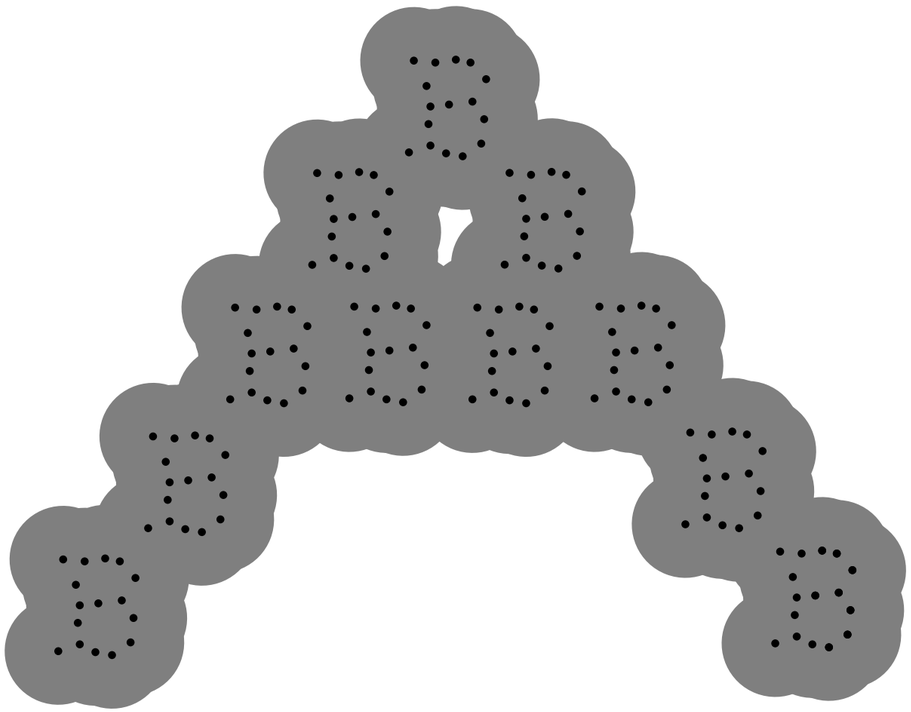


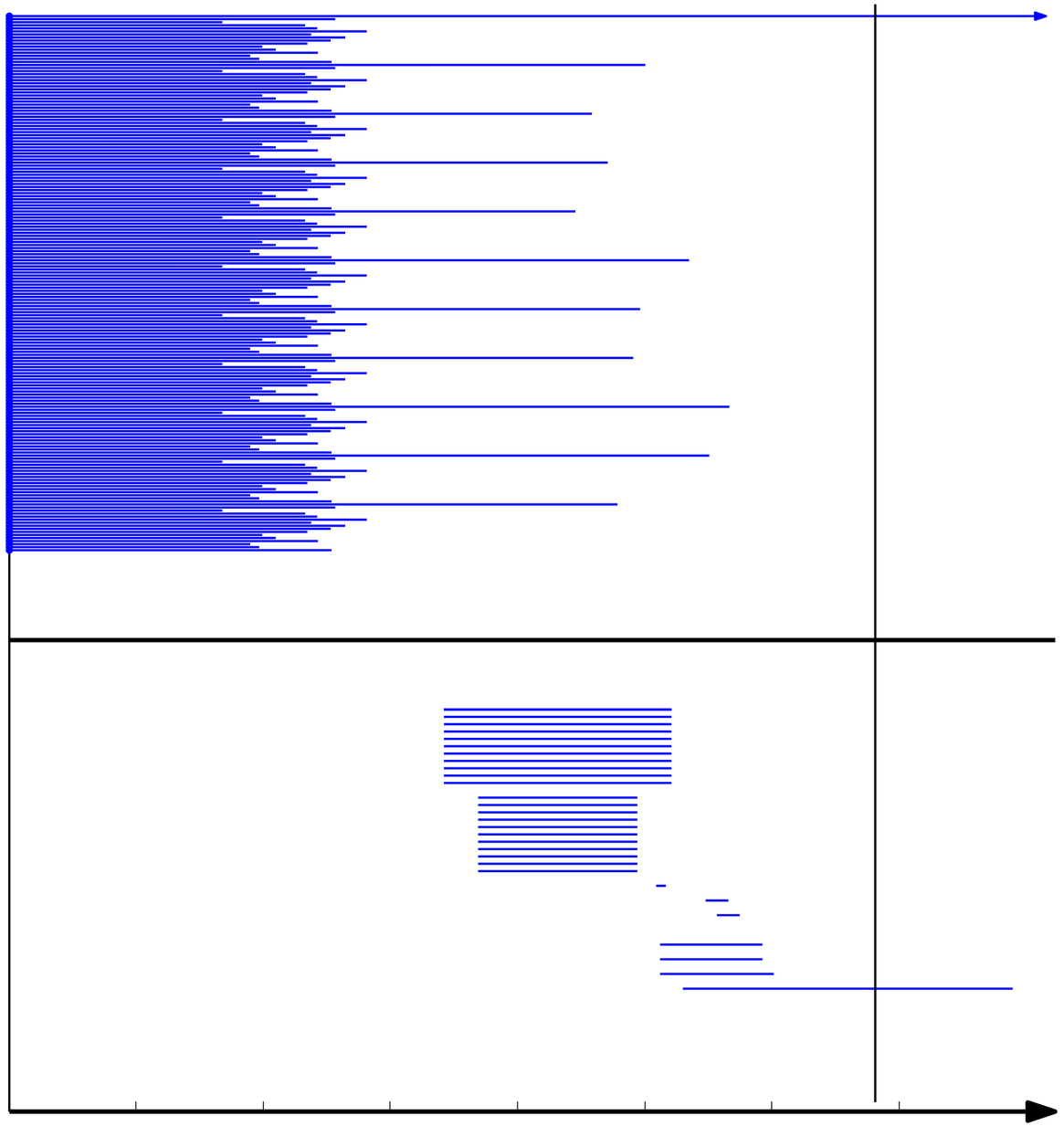
Persistence of offset filtration



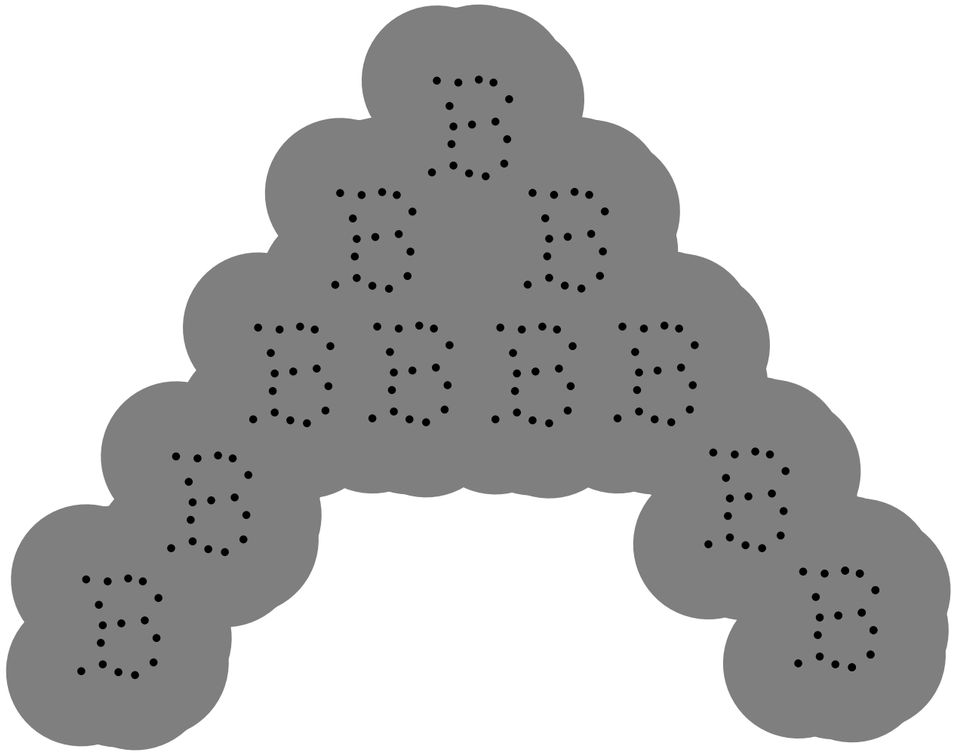


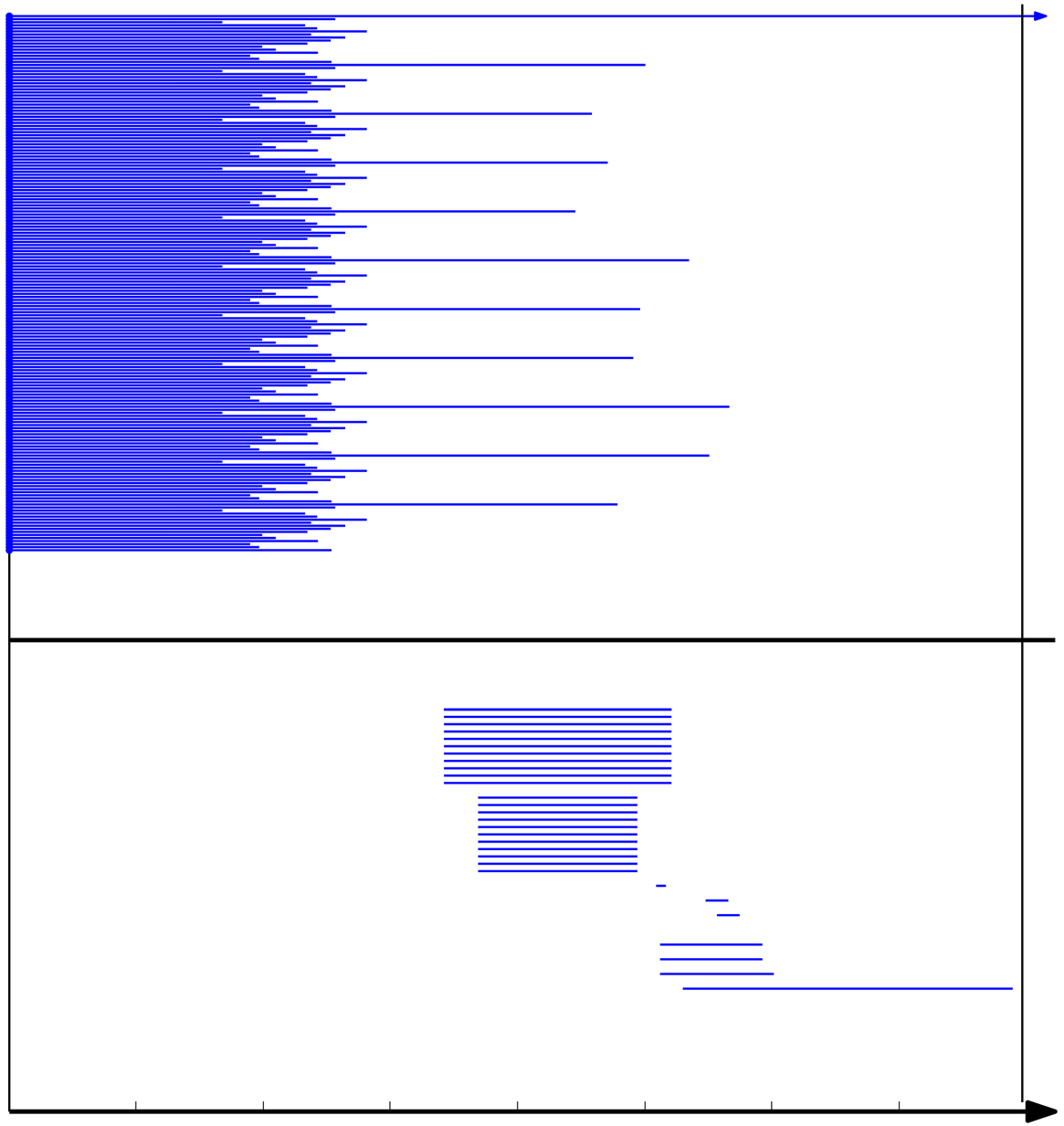
Persistence of offset filtration



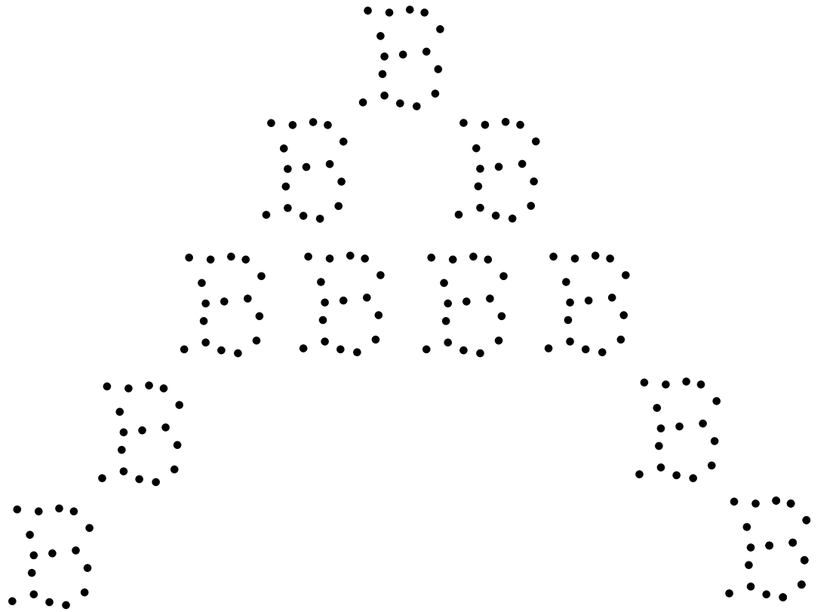


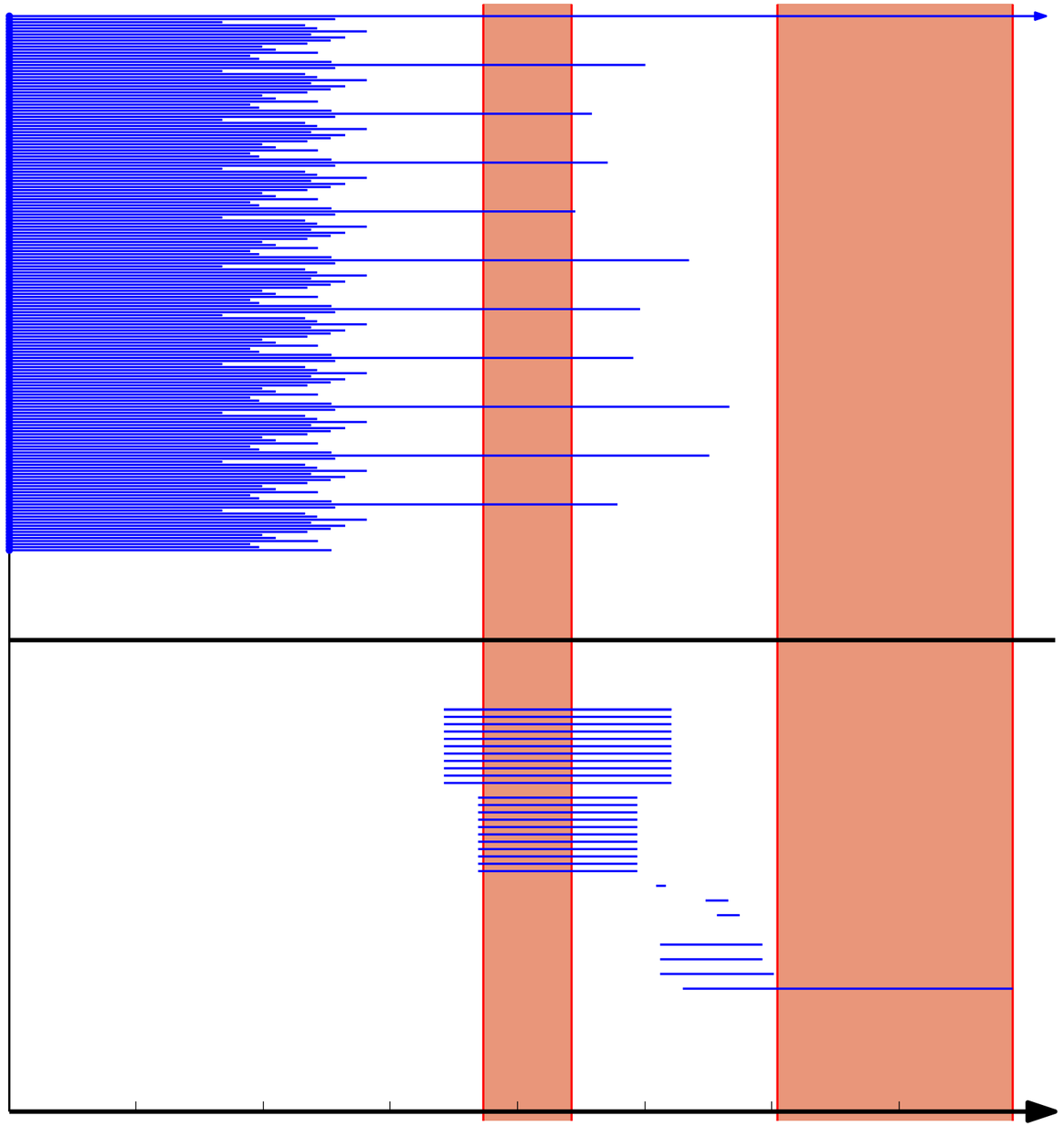
Persistence of offset filtration





Persistence of offset filtration



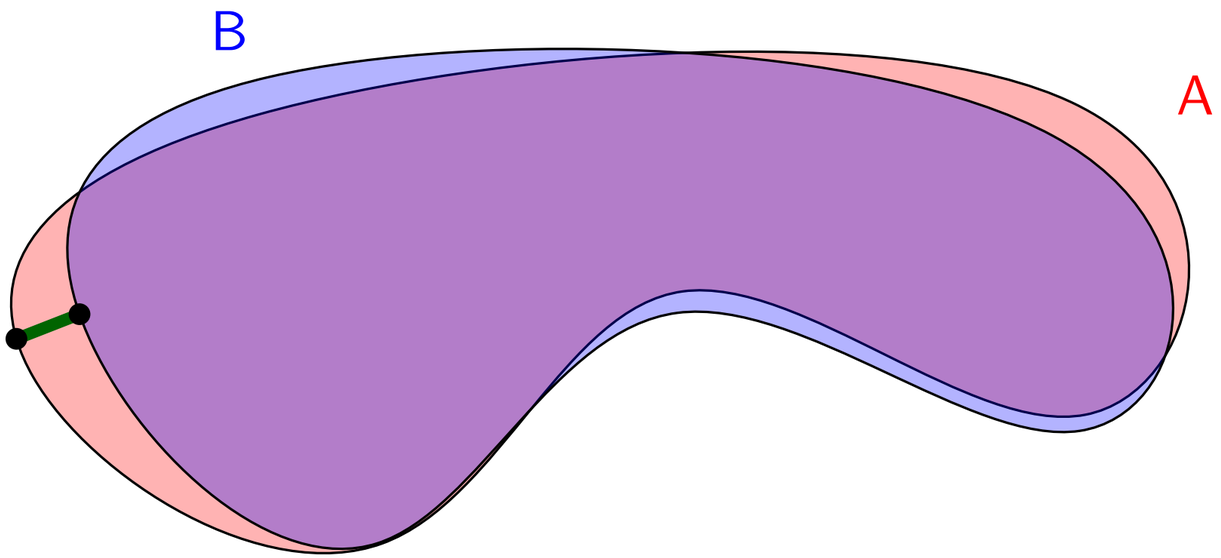


Stability Theorem

$$d_B(Dgm(f_P), Dgm(f_Q)) \leq ||f_P - f_Q||_\infty = d_H(P, Q)$$

d_H: Hausdorff distance

$d_H(A, B) = \max\{\sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B)\}$ where $d(b, A) = \inf_{a \in A} d(b, a)$



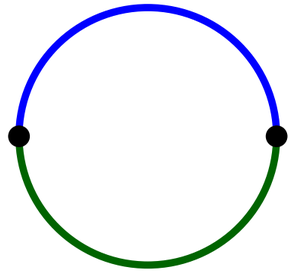
How can we compute all that?

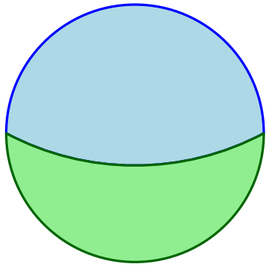
Computers need finite representations

Homology needs a notion of *boundary*

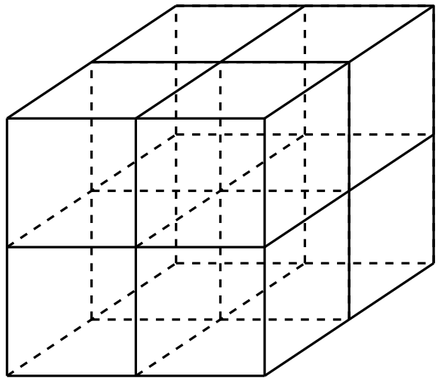
Hole $\sim$ gluing 2 regions with the same boundary

$\Longrightarrow$ Cell complexes (generalization of graphs)

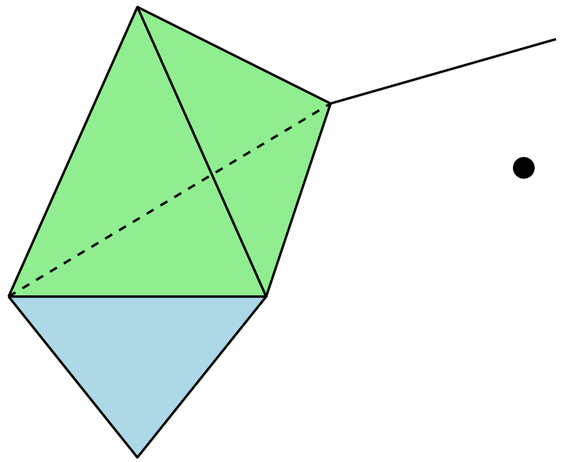




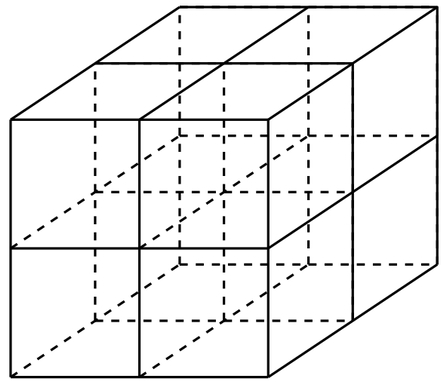
Cubical



Simplicial



Filtered cubical complex



The cells are cubes of all dimensions (including vertices, edges, squares)

Represent a region of $\mathbb{R}^d$ by a subset of cells (*subcomplex*)

Growing region $=$ sequence of subcomplexes $K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n$

Each cell σ has a *filtration value* $f(\sigma)$: time of appearance

Subcomplex $K_t = \{\sigma, f(\sigma) <= t\}$

Homology needs boundaries: the faces of a cell of K_i are also cells of K_i

$\sigma \subset \tau \implies f(\sigma) \leq f(\tau)$

Can compute the persistence diagram of the sequence of subcomplexes

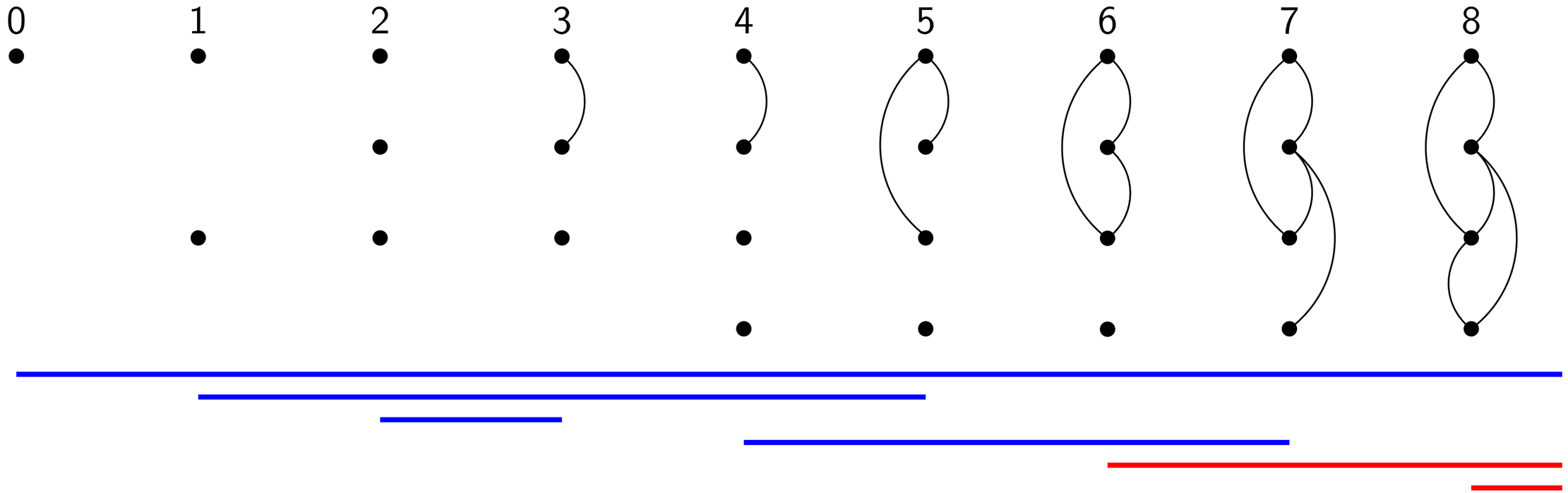
Persistence algorithm in dim 0

Only uses a (filtered) graph

Insert vertices and edges one by one (filtration order)

Vertex: new connected component

Edge ab: if a and b in separate components, kill the youngest, otherwise new loop (ignored for dim 0)



Persistence algorithm in dim 0

Only uses a (filtered) graph

Insert vertices and edges one by one (filtration order)

Vertex: new connected component

Edge ab: if a and b in separate components, kill the youngest, otherwise new loop (ignored for dim 0)

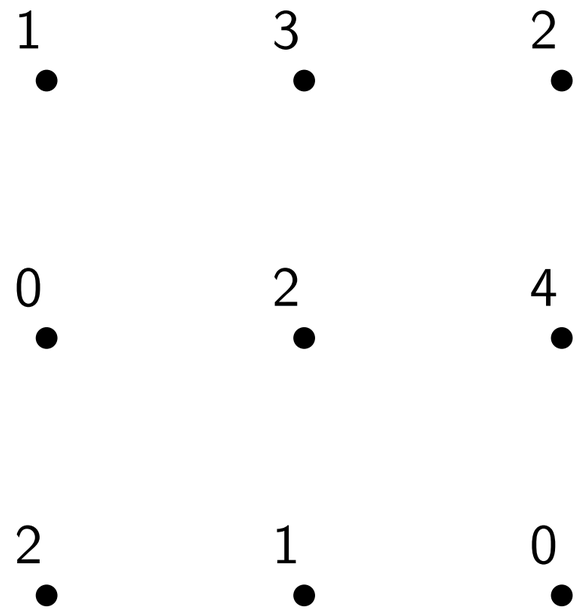
Disjoint-set data structure (aka union-find): very fast, running time dominated by sorting edges

Dimension p:

- Use cells of dimension p and $p+1$
- Replace "separate components" with algebra (boundary not in the vector space generated by previous boundaries)
- Worst case $\Theta(n^3)$, in practice $O(n)$ (n number of cells)
- **Co**homology? Same diagram

Functions

Discretize the function: grid points



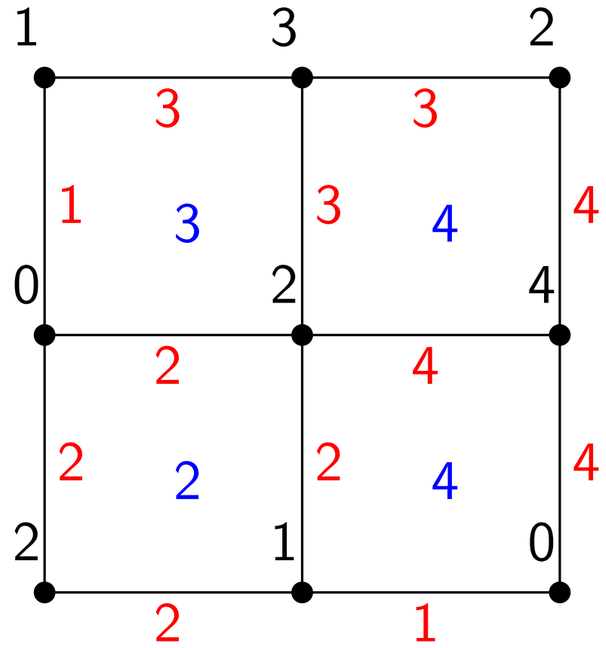
Functions

Discretize the function: grid points

Extend to other cells: lower-star filtration

$f(\sigma) =$ maximum value at its vertices

Same persistence diagram as a piecewise linear interpolation



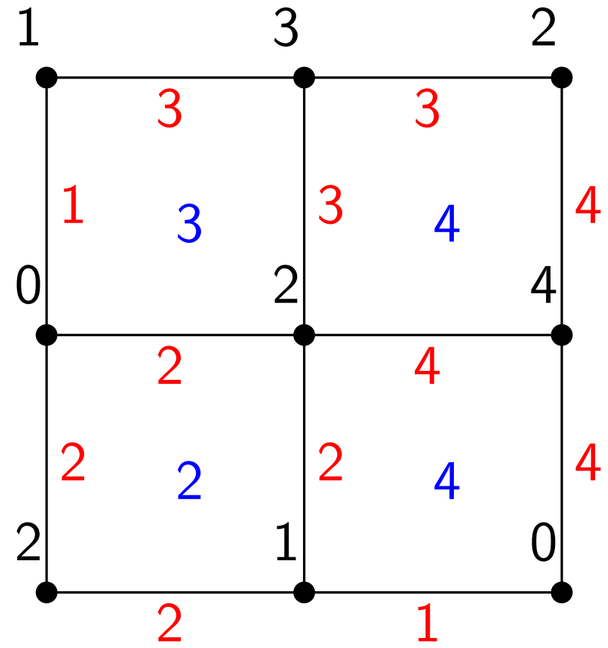
Functions

Discretize the function: grid points

Extend to other cells: lower-star filtration

$f(\sigma) =$ maximum value at its vertices

Same persistence diagram as a piecewise linear interpolation



level 0

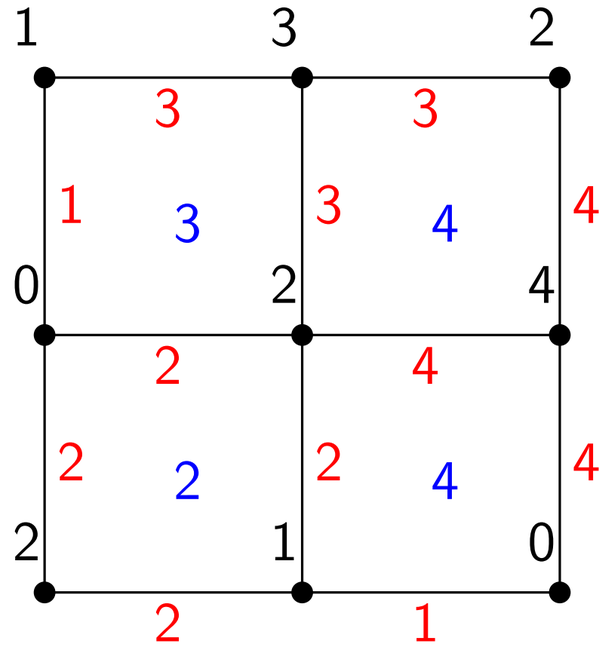
Functions

Discretize the function: grid points

Extend to other cells: lower-star filtration

$f(\sigma) =$ maximum value at its vertices

Same persistence diagram as a piecewise linear interpolation



level 1

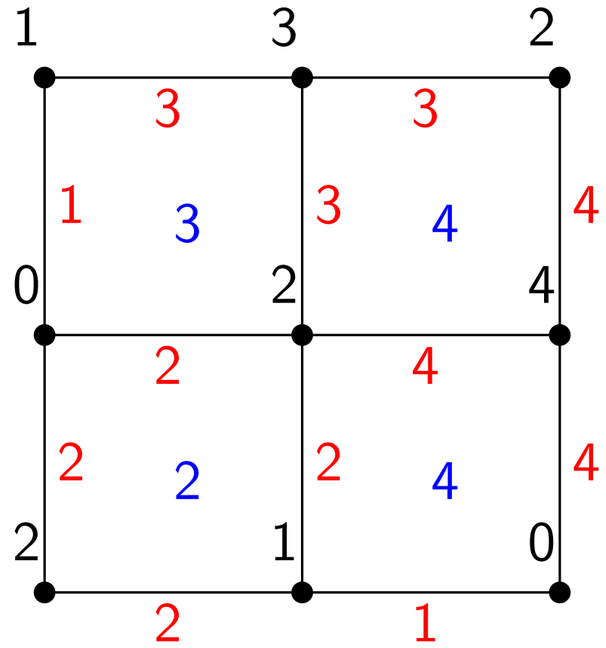
Functions

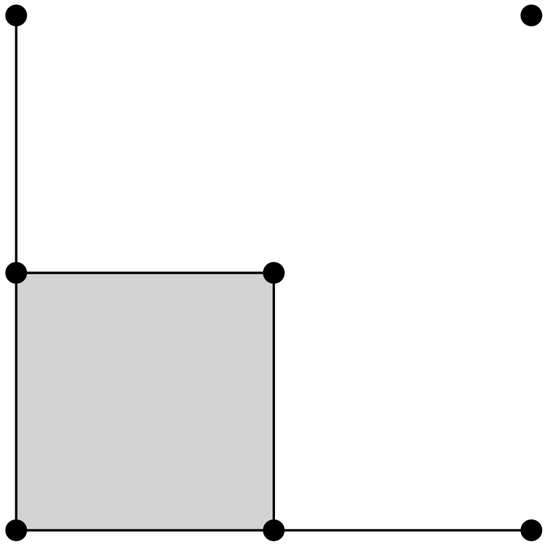
Discretize the function: grid points

Extend to other cells: lower-star filtration

$f(\sigma) =$ maximum value at its vertices

Same persistence diagram as a piecewise linear interpolation





level 2

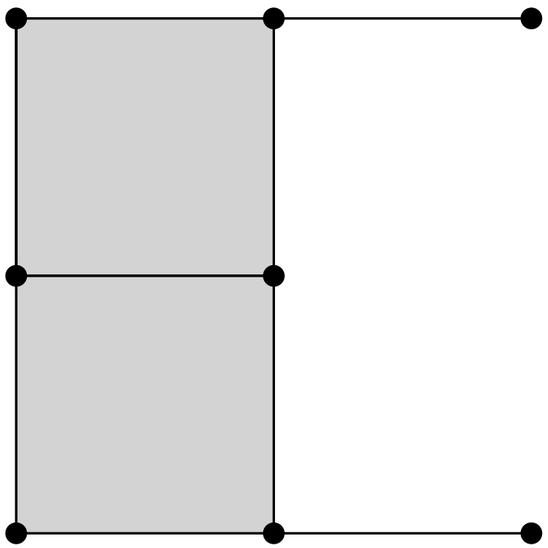
Functions

Discretize the function: grid points

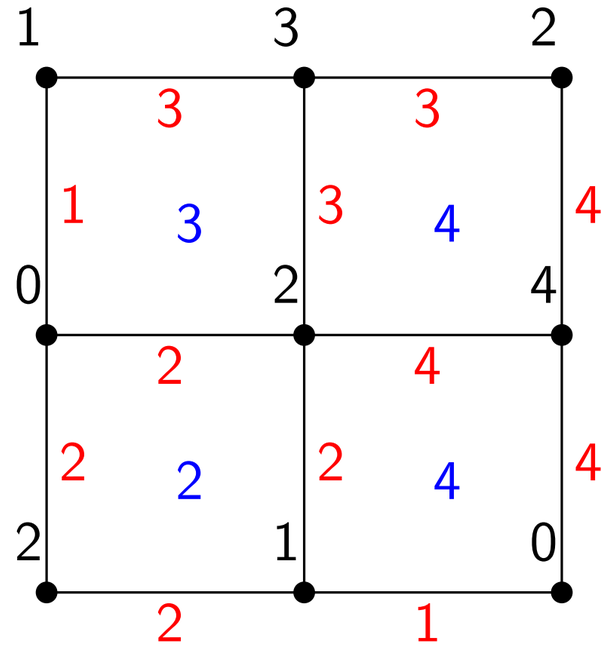
Extend to other cells: lower-star filtration

$f(\sigma) =$ maximum value at its vertices

Same persistence diagram as a piecewise linear interpolation



level 3



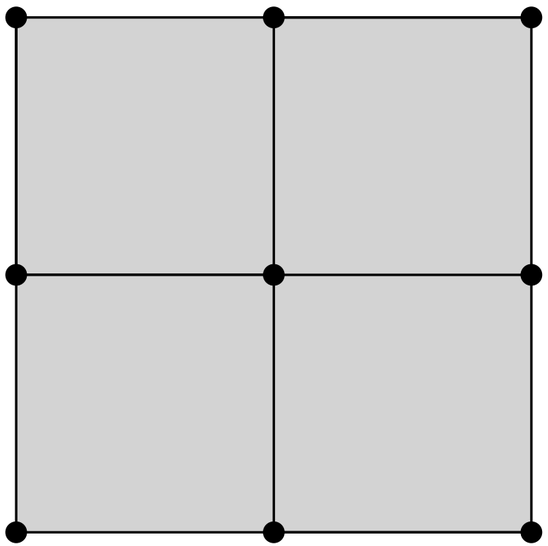
Functions

Discretize the function: grid points

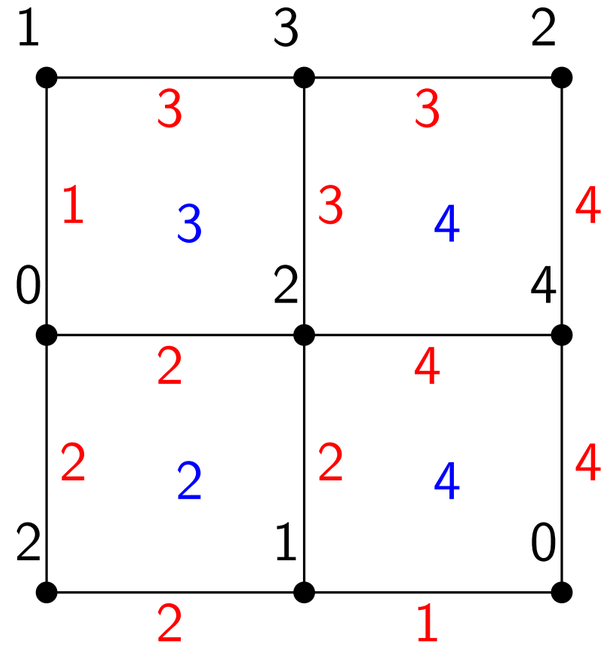
Extend to other cells: lower-star filtration

$f(\sigma) =$ maximum value at its vertices

Same persistence diagram as a piecewise linear interpolation

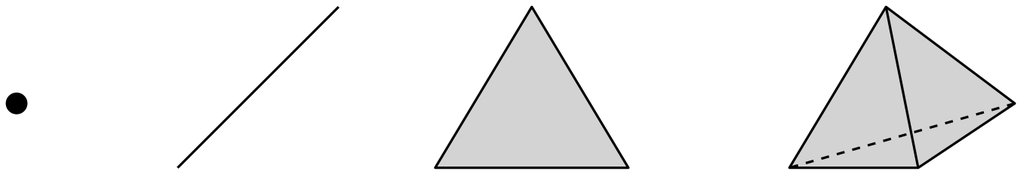


level 4



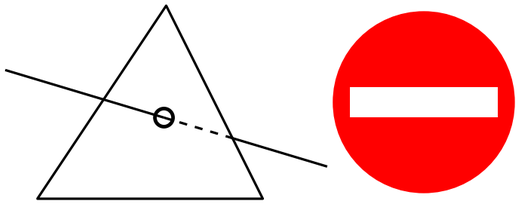
Simplicial complex

Simplex: vertex, edge, triangle, tetrahedron, etc



Simplices have the *minimal* number of vertices for their dimension (no squares or pentagons)

Nice combinatorial intersection: $\sigma \cap \tau$ is a common face of σ and τ or empty.

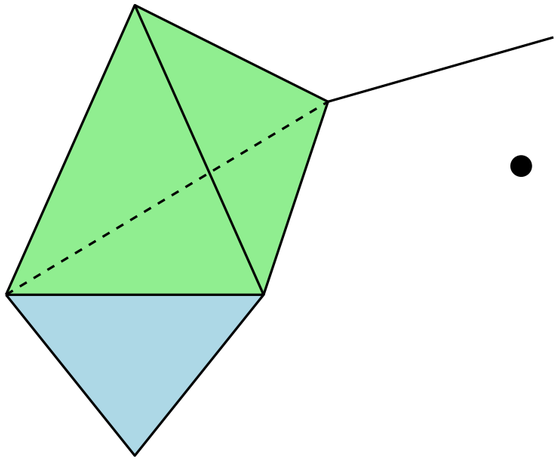


Represent a region by a subset of cells (*subcomplex*)

Growing region = sequence of subcomplexes $K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n$

Each cell σ has a *filtration value* $f(\sigma)$: time of appearance

Subcomplex $K_t = \{\sigma, f(\sigma) <= t\}$



Homology needs boundaries: the faces of a simplex of K_i are also in K_i

$\sigma \subset \tau \implies f(\sigma) \leq f(\tau)$

Can compute the persistence diagram of the sequence of subcomplexes

Nerve of a cover $\{U_1, \ldots, U_n\}$

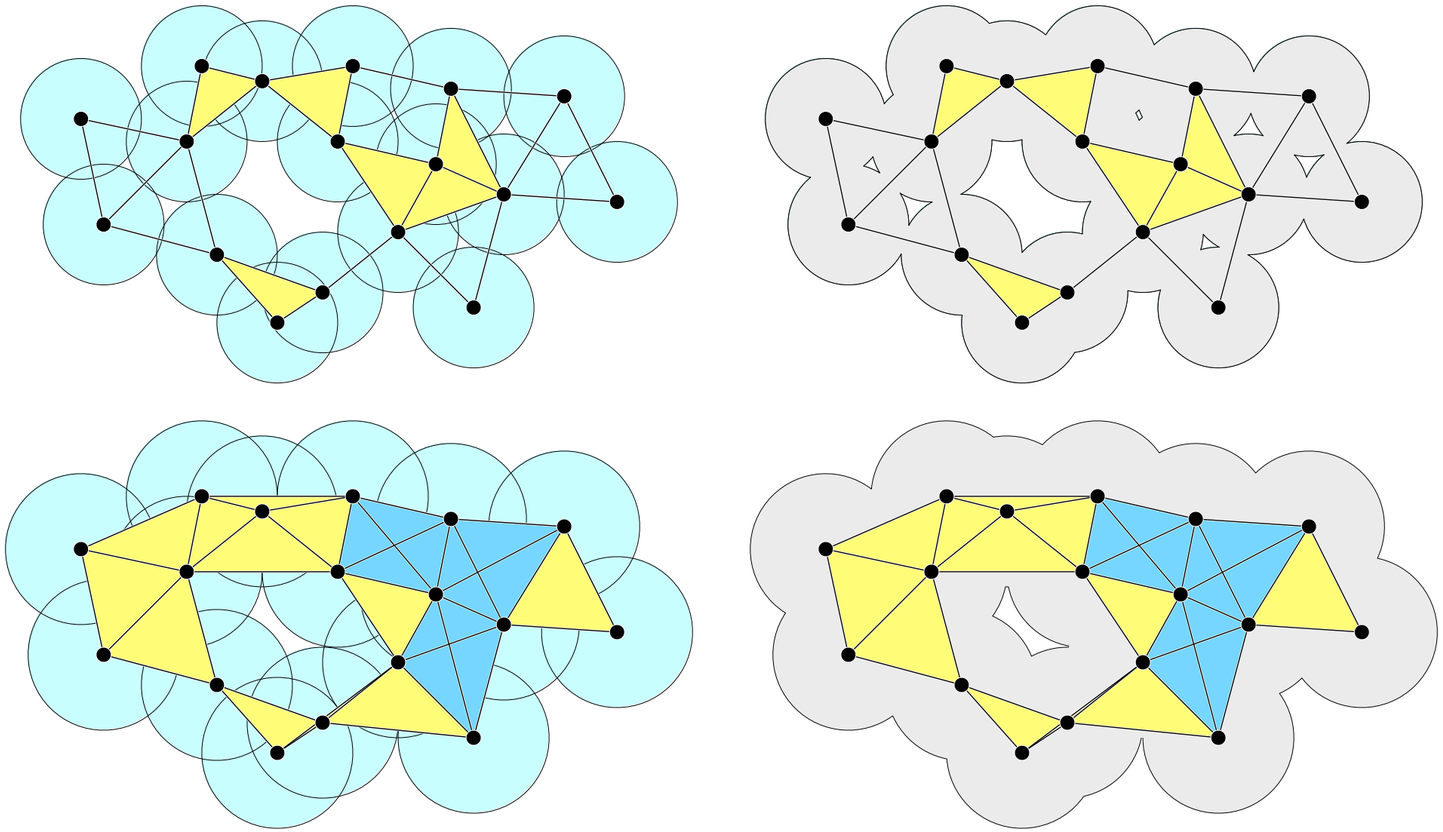
One vertex v_i per U_i

$\sigma = [v_{i_0}, \ldots v_{i_k}] \in K \iff \bigcap_{j=0}^{k} U_{i_j}$

Abstract (not embedded)

Nerve Theorem: If all intersections $\bigcap_{i \in I \subseteq [1,n]} U_i$ are either empty or contractible, then the union $\bigcup_{i=1}^{n} U_i$ has the same *homotopy type* as the Nerve (simplicial complex).

Always true for convex objects (balls)



Persistent nerve of a growing cover $\{U_1^t, \ldots, U_n^t\}$

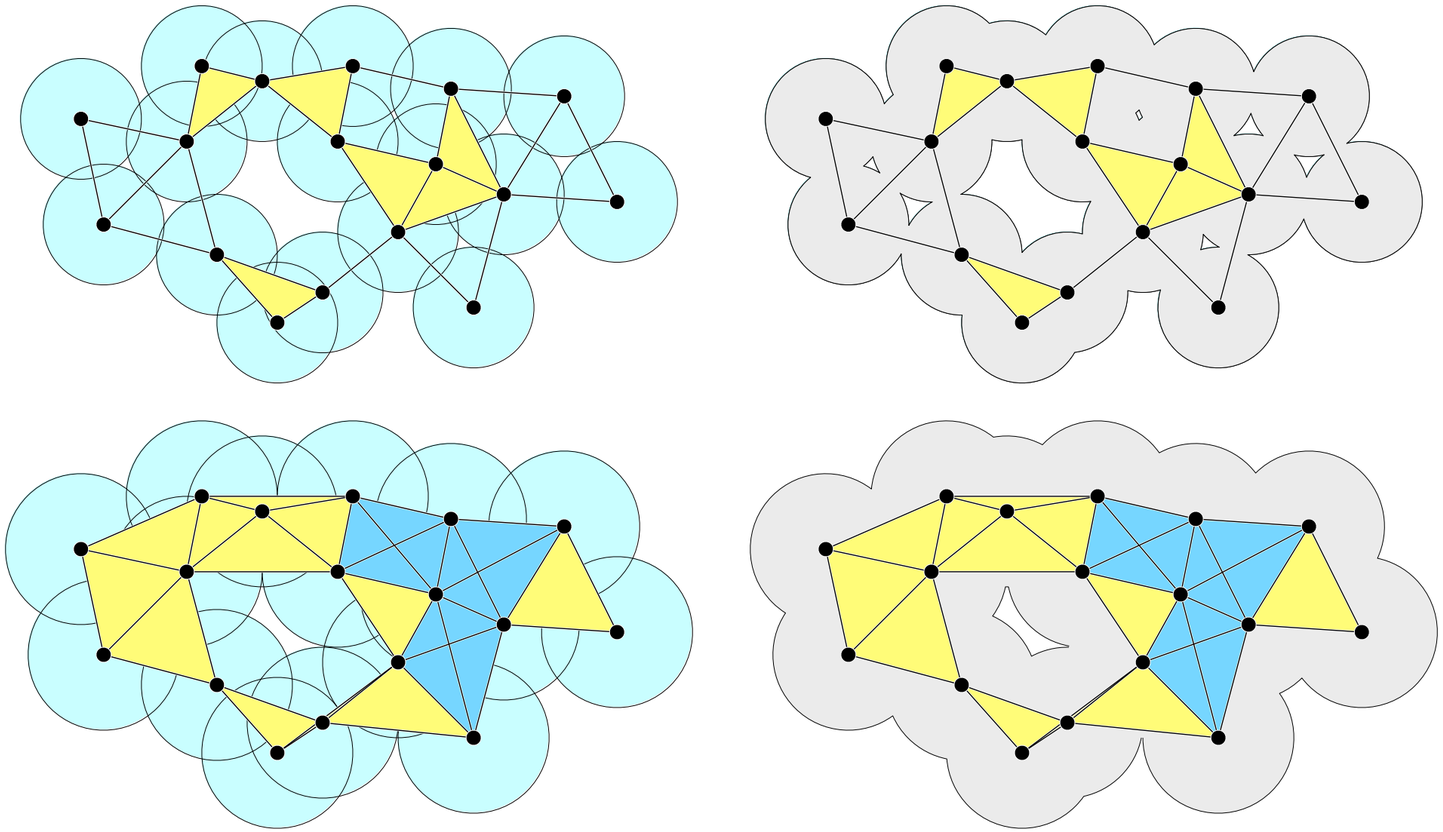
$U_i^t \subseteq U_i^{t'}$ when $t \leq t'$

K^t: Nerve of $\{U_1^t, \ldots, U_n^t\}$

$K^t \subseteq K^{t'}$ when $t \leq t'$

Persistence diagram of $U^t = \bigcup_{i=1}^n U_i^t$

Persistence diagram of K^t



Persistent nerve theorem: If all intersections $\bigcap_{i \in I \subseteq [1,n]} U_i^t$ are either empty or contractible, then the two diagrams are the same.

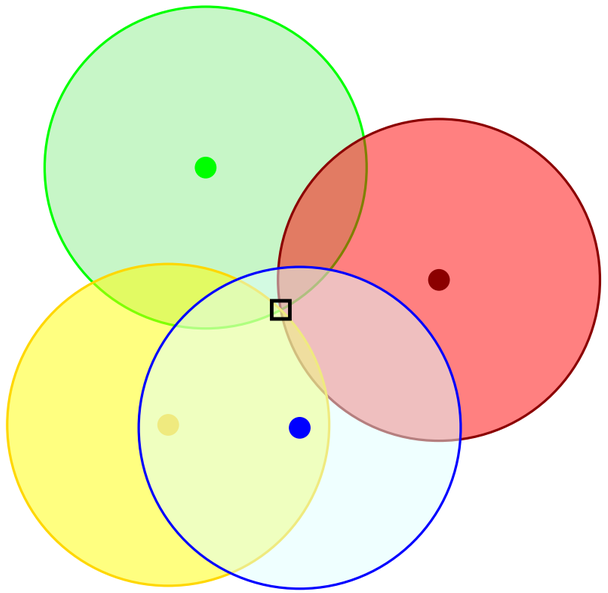
Čech filtration

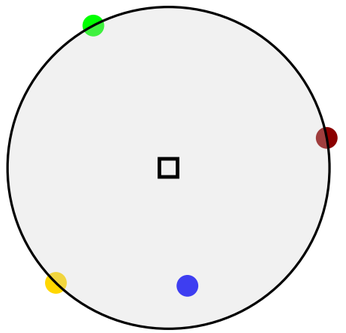
Finite point set P, parameter r

$C_r(P)$ is the nerve of the union of the balls of radius r centered on the points of P

The sequence of $C_r(P)$ when r increases defines a filtered simplicial complex

$f(\sigma) =$ radius of Minimal Enclosing Ball of the vertices of σ





Čech filtration

Finite point set P, parameter r

$C_r(P)$ is the nerve of the union of the balls of radius r centered on the points of P

The sequence of $C_r(P)$ when r increases defines a filtered simplicial complex

$f(\sigma) =$ radius of Minimal Enclosing Ball of the vertices of σ

1 persistence diagram for the growing union of balls when r increases

1 persistence diagram for the sequence of Čech complexes when r increases

Persistent nerve theorem: those 2 diagrams are the same

Čech filtration

Finite point set P, parameter r

$C_r(P)$ is the nerve of the union of the balls of radius r centered on the points of P

The sequence of $C_r(P)$ when r increases defines a filtered simplicial complex

$f(\sigma) =$ radius of Minimal Enclosing Ball of the vertices of σ

1 persistence diagram for the growing union of balls when r increases

1 persistence diagram for the sequence of Čech complexes when r increases

Persistent nerve theorem: those 2 diagrams are the same

Weaknesses:

- Big, $2^{|P|}$ simplices if we do not limit it
- Numerical, geometric computation to test the intersection of k balls in $\mathbb{R}^d$

Rips filtration

Finite point set P, parameter r

$\sigma \in R_r(P) \iff \text{diam}(\sigma) \leq 2r$

Same graph as the Čech complex $C_r(P)$

Add simplices for cliques (complete subgraphs): purely combinatorial

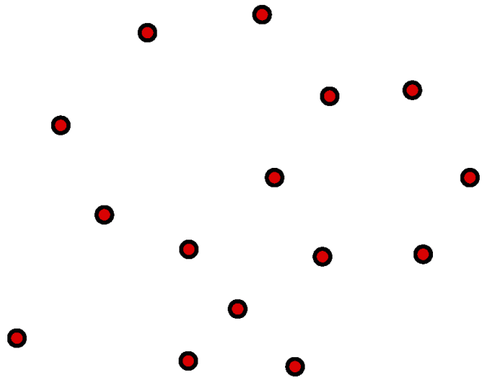
$C_r(P) \subseteq R_r(P) \subseteq C_{2r}(P)$

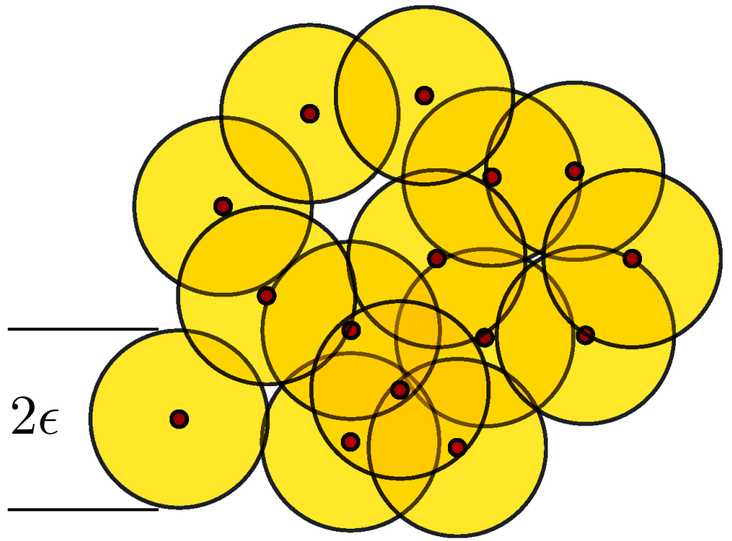
$\implies$ persistence diagrams of Rips and Čech are at distance less than 2 in log-scale

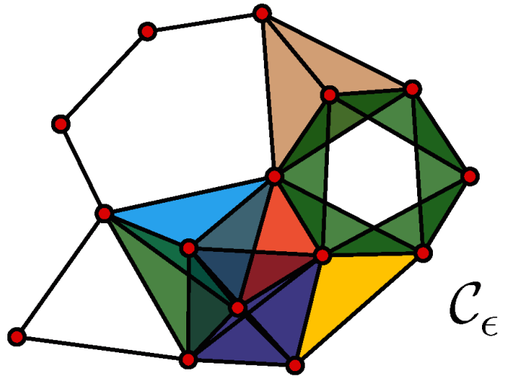
Very flexible (arbitrary value on edges)

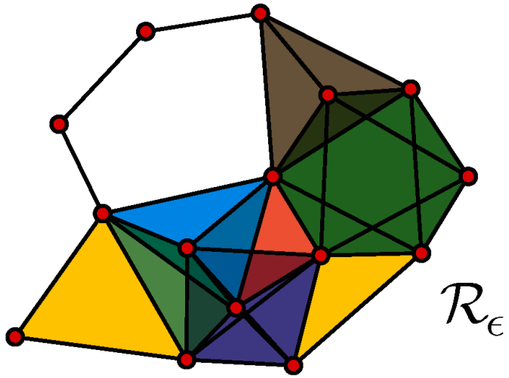
Drawbacks: not exact topology, and even bigger than Čech

Stability theorem still applies





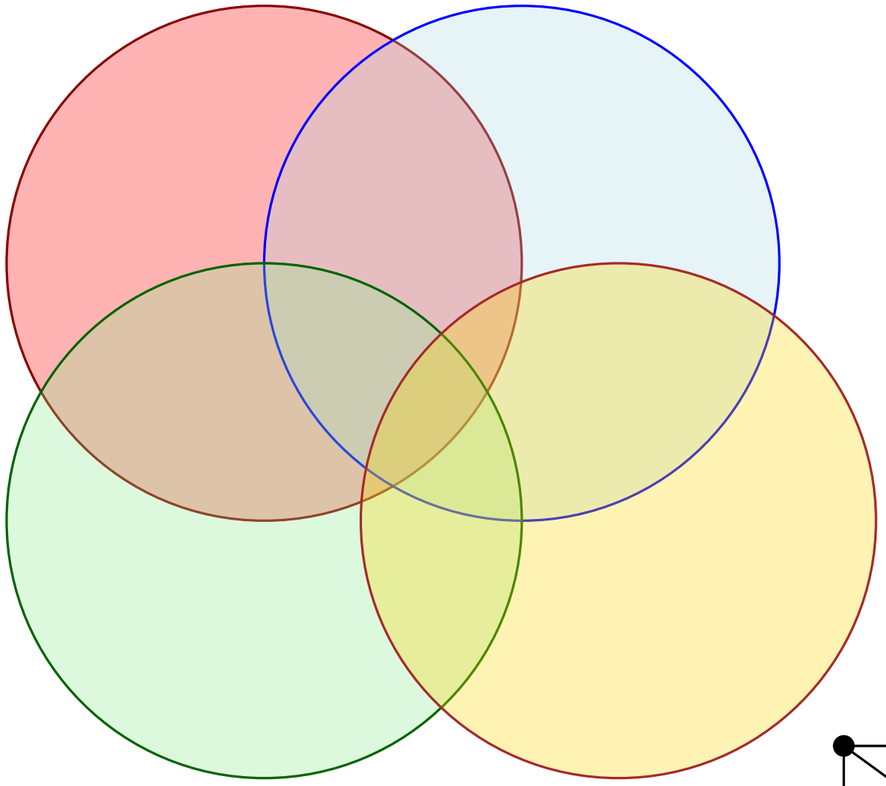




α-complex

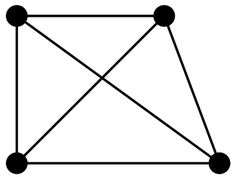
Čech: lot of redundancy in the cover when r is large.

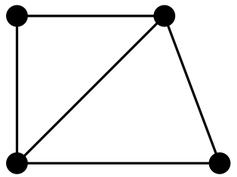
Idea: no need to keep growing in places already covered by other balls

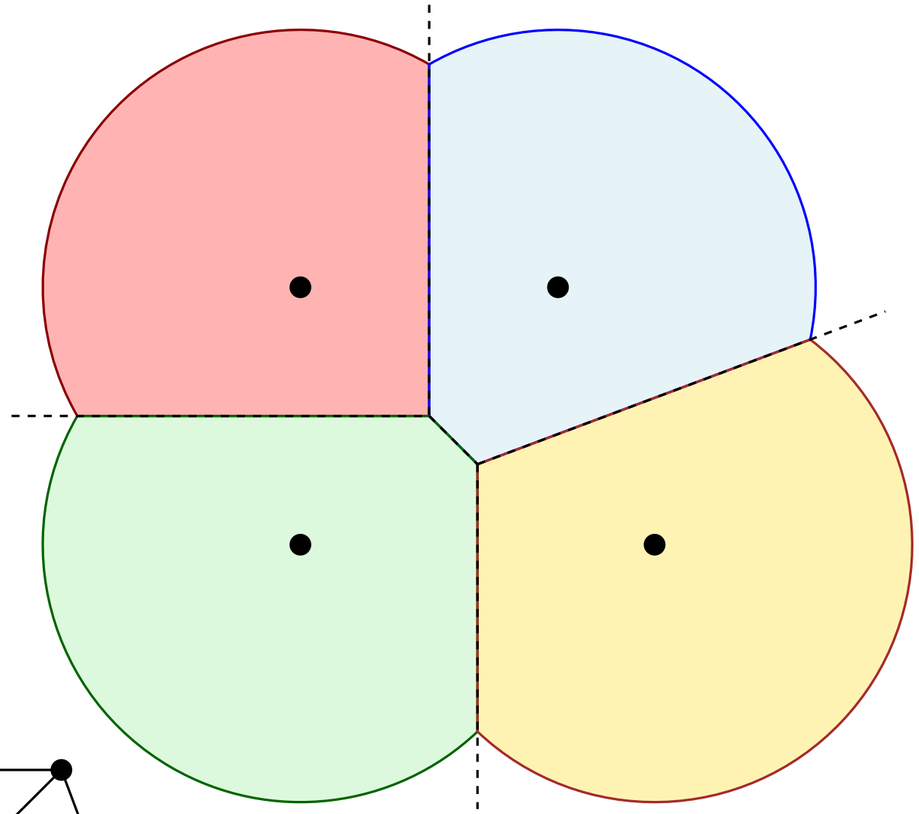


Čech complex

- Restrict balls to Voronoi cells
- Same union
- Still convex $\Longrightarrow$ Nerve theorem
- Same persistence diagram
- Smaller complex
- Embedded







α-complex

Choice: Rips or α-complex?

Seldom compute the true Čech complex

Rips: easy, very flexible. Independent of any embedding. Have to limit dimension and edge length

α-complex: only defined in Euclidean $\mathbb{R}^d$, currently only efficient for small d, but super efficient there

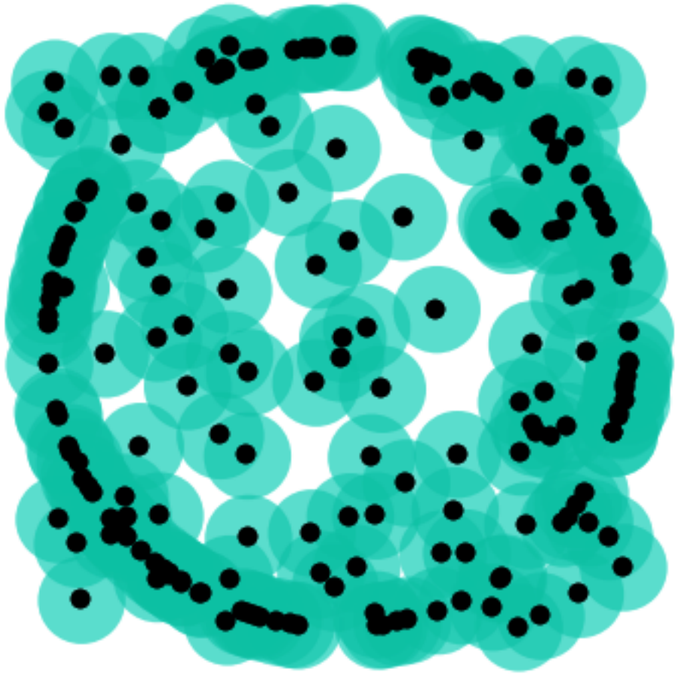
Several other alternatives, including sparse Rips, etc

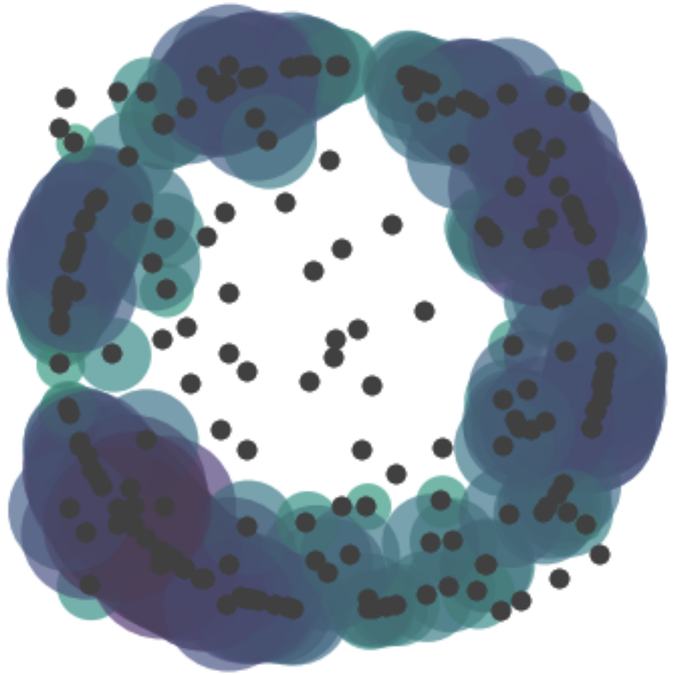
Noise (outliers)

Stability theorem: only handles small perturbations, one outlier can break everything

Idea: superlevelsets of a density estimator

Computable version: weighted version of Čech, Rips; penalize points in low density regions



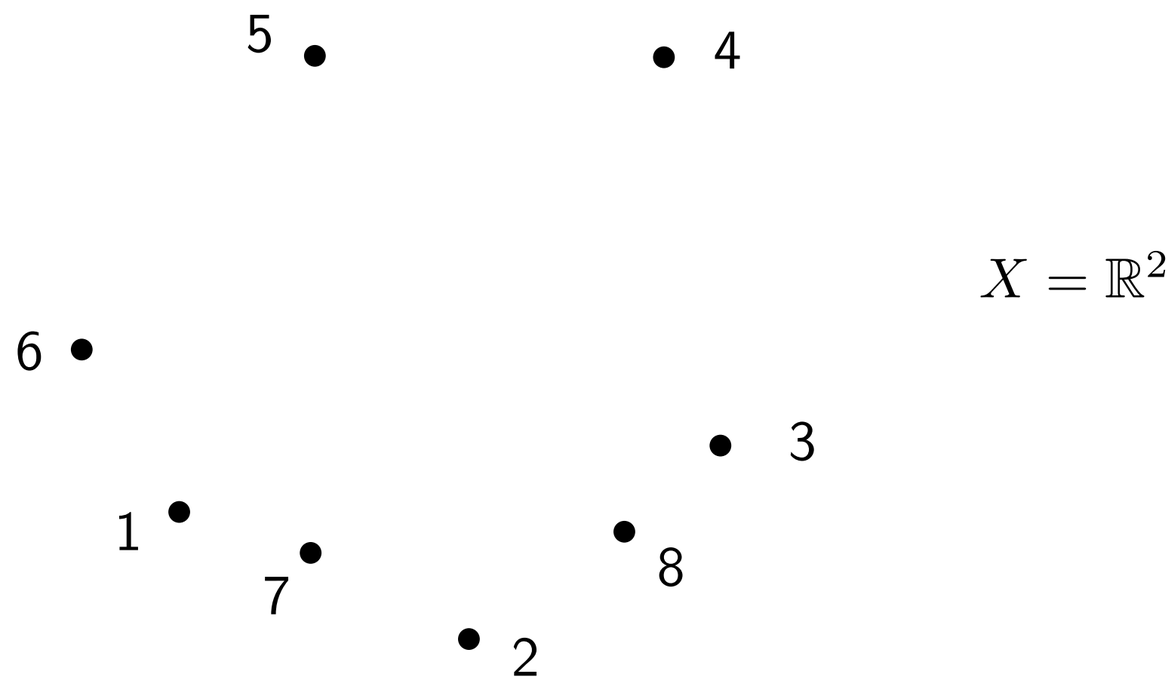


Crossover: time series as a point cloud

Time series: function $f : \mathbb{R} \to X$

If $X = \mathbb{R}$ use sublevelsets?

Alternative idea: forget time, see $f(\mathbb{R}) \subseteq X$ as a point set, compute Čech complex

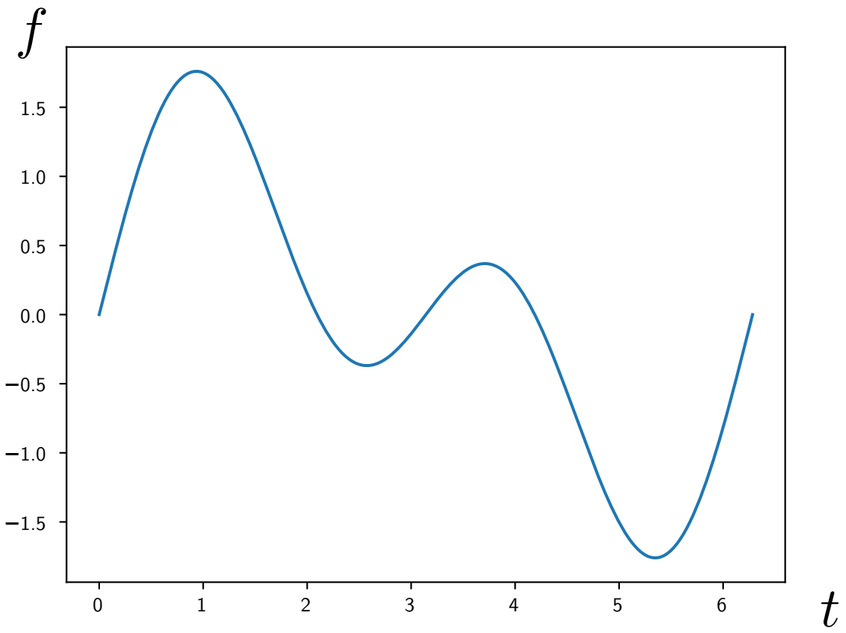


Crossover: time series as a point cloud

Time series: function $f : \mathbb{R} \to X$

If $X = \mathbb{R}$ use sublevelsets?

Alternative idea: forget time, see $f(\mathbb{R}) \subseteq X$ as a point set, compute Čech complex



$f(\mathbb{R})$

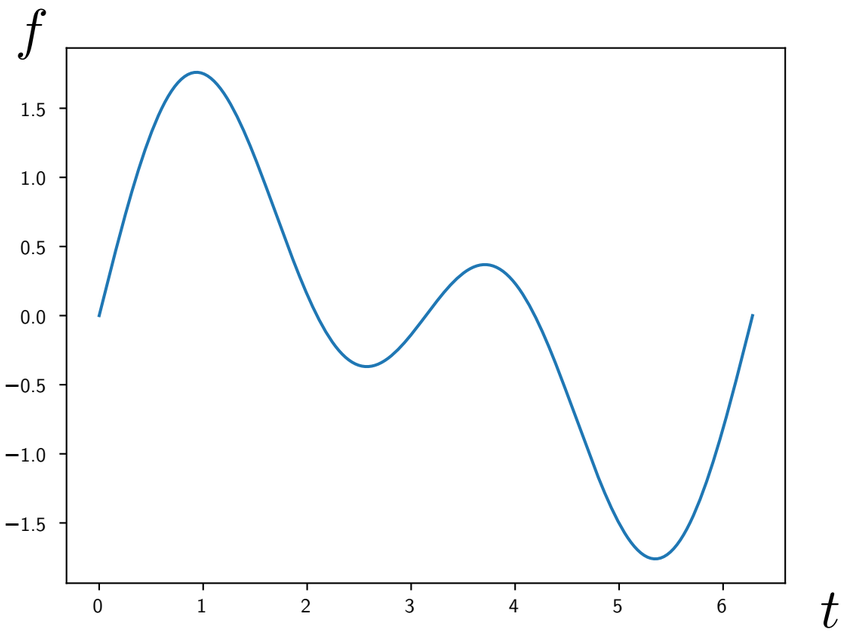
Crossover: time series as a point cloud

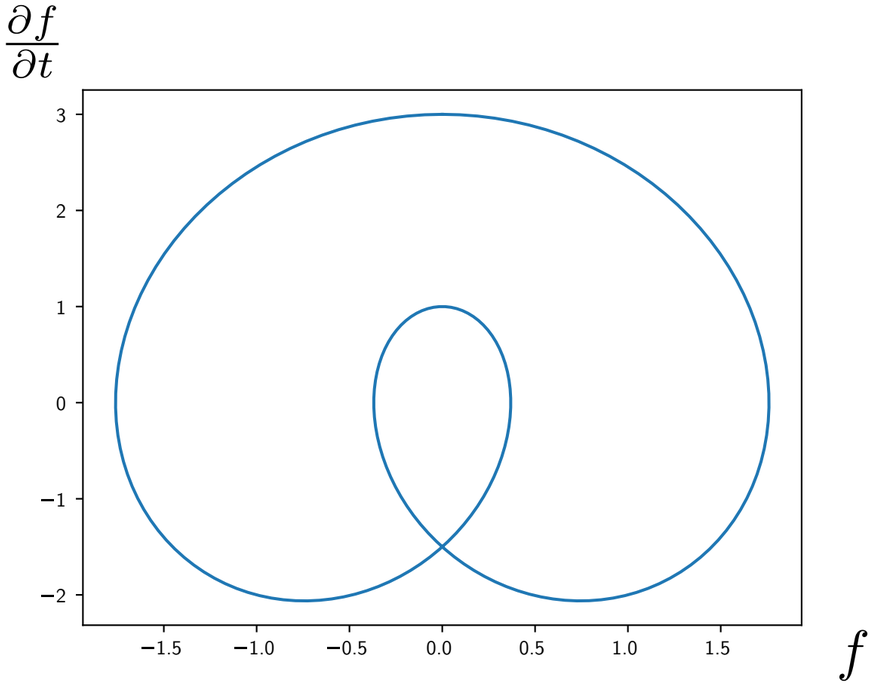
Time series: function $f : \mathbb{R} \to X$

If $X = \mathbb{R}$ use sublevelsets?

Alternative idea: forget time, see $f(\mathbb{R}) \subseteq X$ as a point set, compute Čech complex

Lose too much information? Enrich f first, define $g = (f, \frac{\partial f}{\partial t})$





Crossover: time series as a point cloud

Time series: function $f : \mathbb{R} \to X$

If $X = \mathbb{R}$ use sublevelsets?

Alternative idea: forget time, see $f(\mathbb{R}) \subseteq X$ as a point set, compute Čech complex

Lose too much information? Enrich f first, define $g = (f, \frac{\partial f}{\partial t})$

In practice, from sequence u_n, define e.g. $v_n = (u_n, u_{n+2}, u_{n+7})$

Inspired by Taken's theorem in dynamical systems

Conclusion

Can be expensive to compute: aim for a smaller complex

Hardest part: deciding on what function to compute persistence

Next classes: how to use this in stat / ML