July 9, 2009

TGDA Workshop

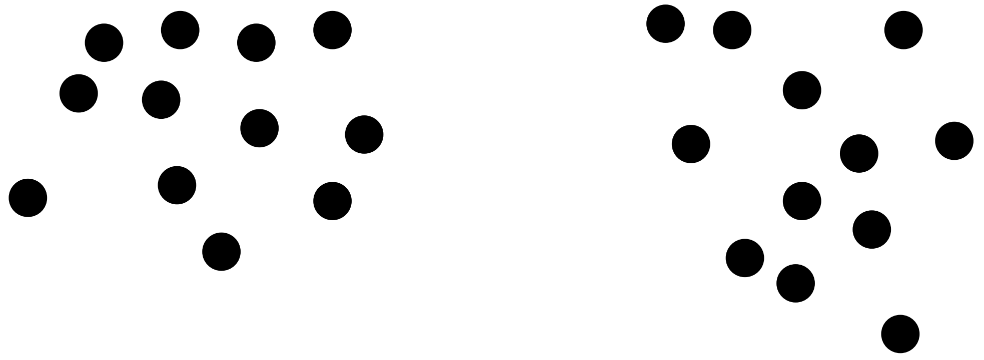
Persistence based Clustering

Primoz Skraba

joint work with
Frédéric Chazal, Steve Y. Oudot, Leonidas J. Guibas

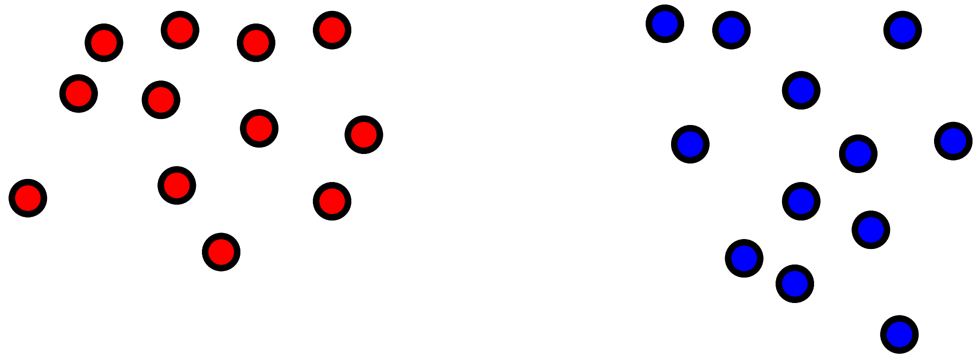
Clustering

- Input samples



Clustering

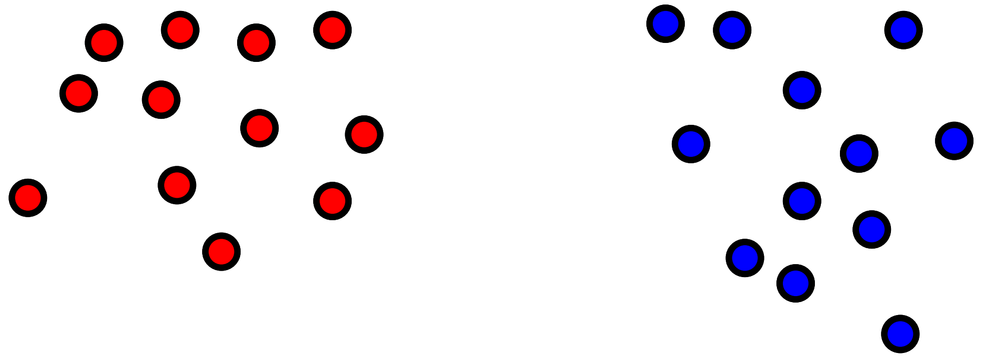
- Input samples
- "Important" segments/clusters



Clustering

- Input samples
- "Important" segments/clusters

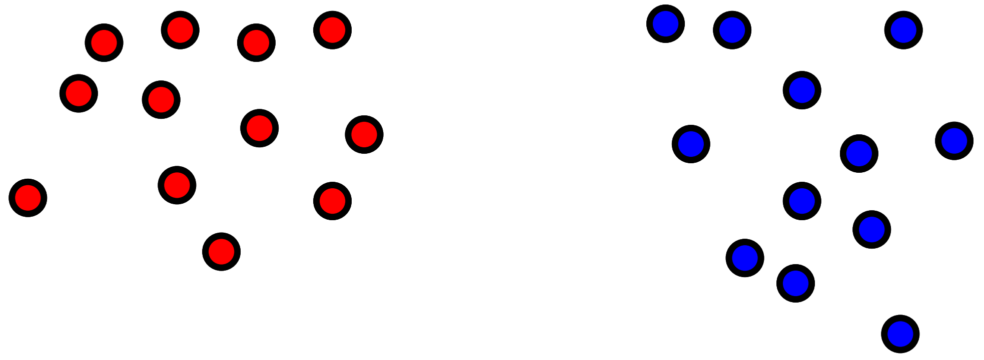
  ill-posed problem



Clustering

- Input samples
- "Important" segments/clusters

  ill-posed problem



- Extensive previous work

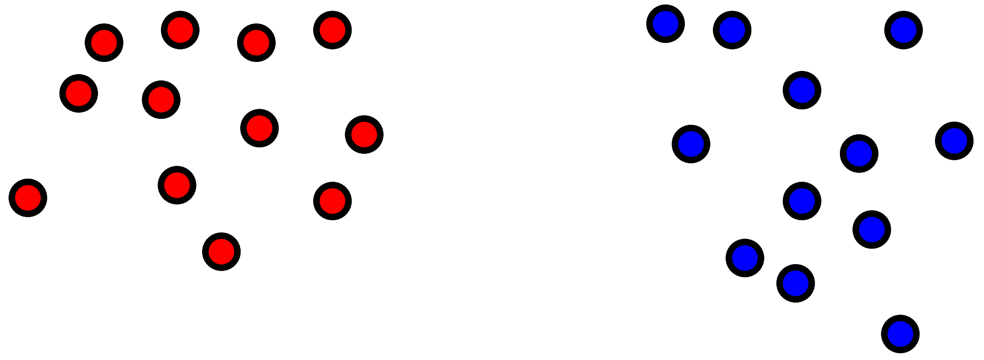
  - k-means
  - spectral clustering
  - mode-seeking (mean-shift)

Clustering

- Input samples

- "Important" segments/clusters

  ill-posed problem



- Extensive previous work

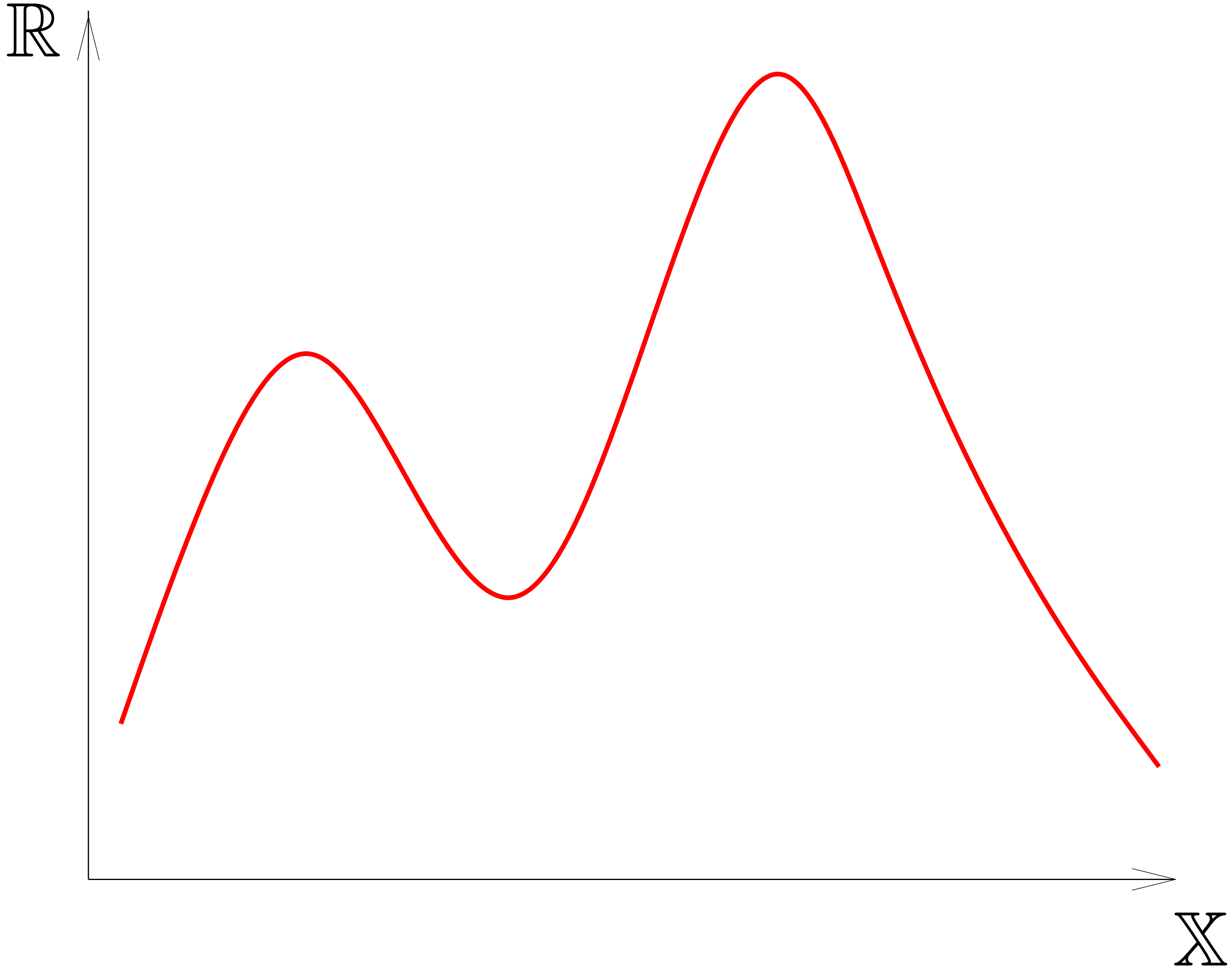
  - k-means
  - spectral clustering
  - mode-seeking (mean-shift)

- Our viewpoint:

  data points drawn at random from some unknown density distribution f

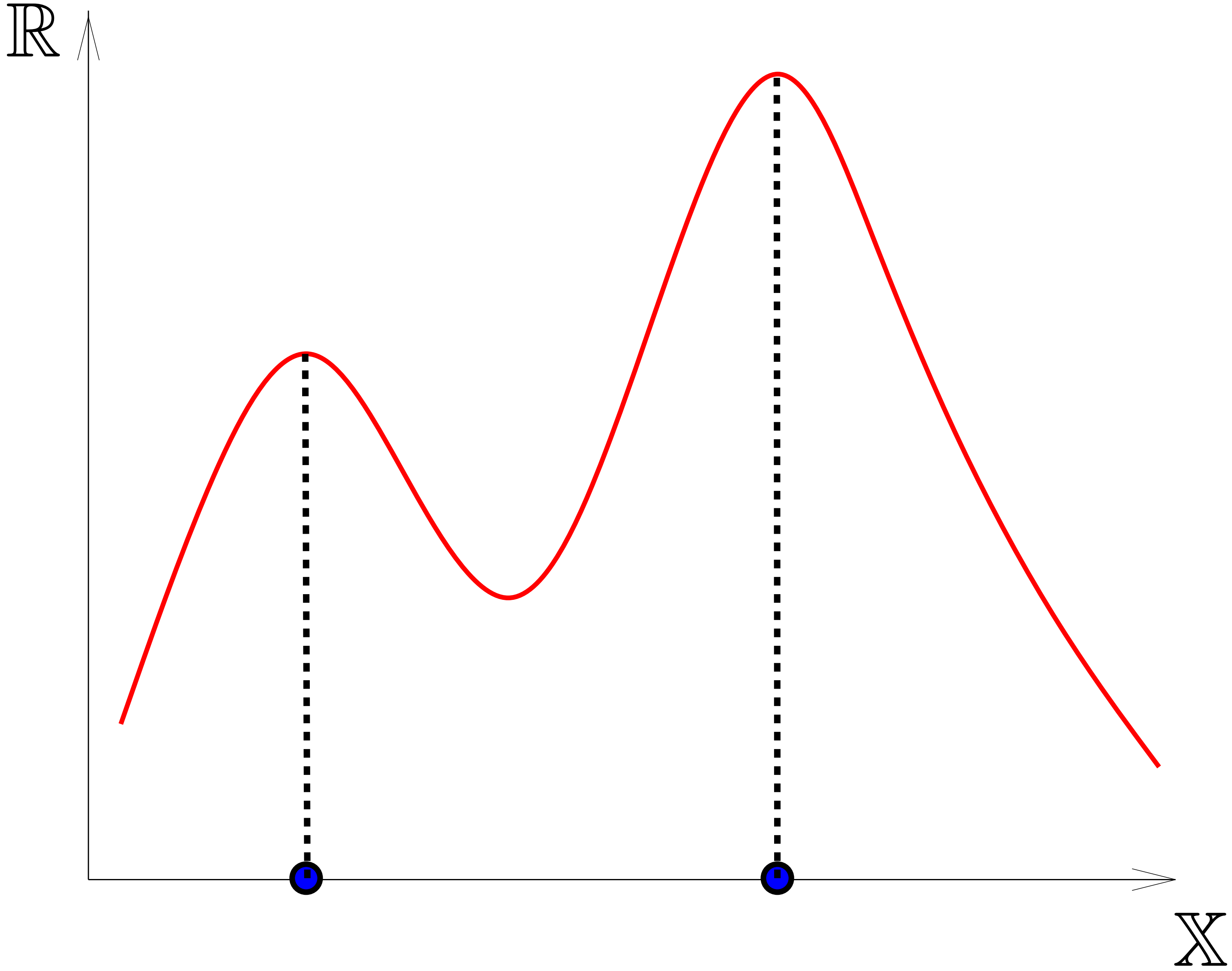
Definition of a Cluster

- Basins of attraction of "significant" peaks of f



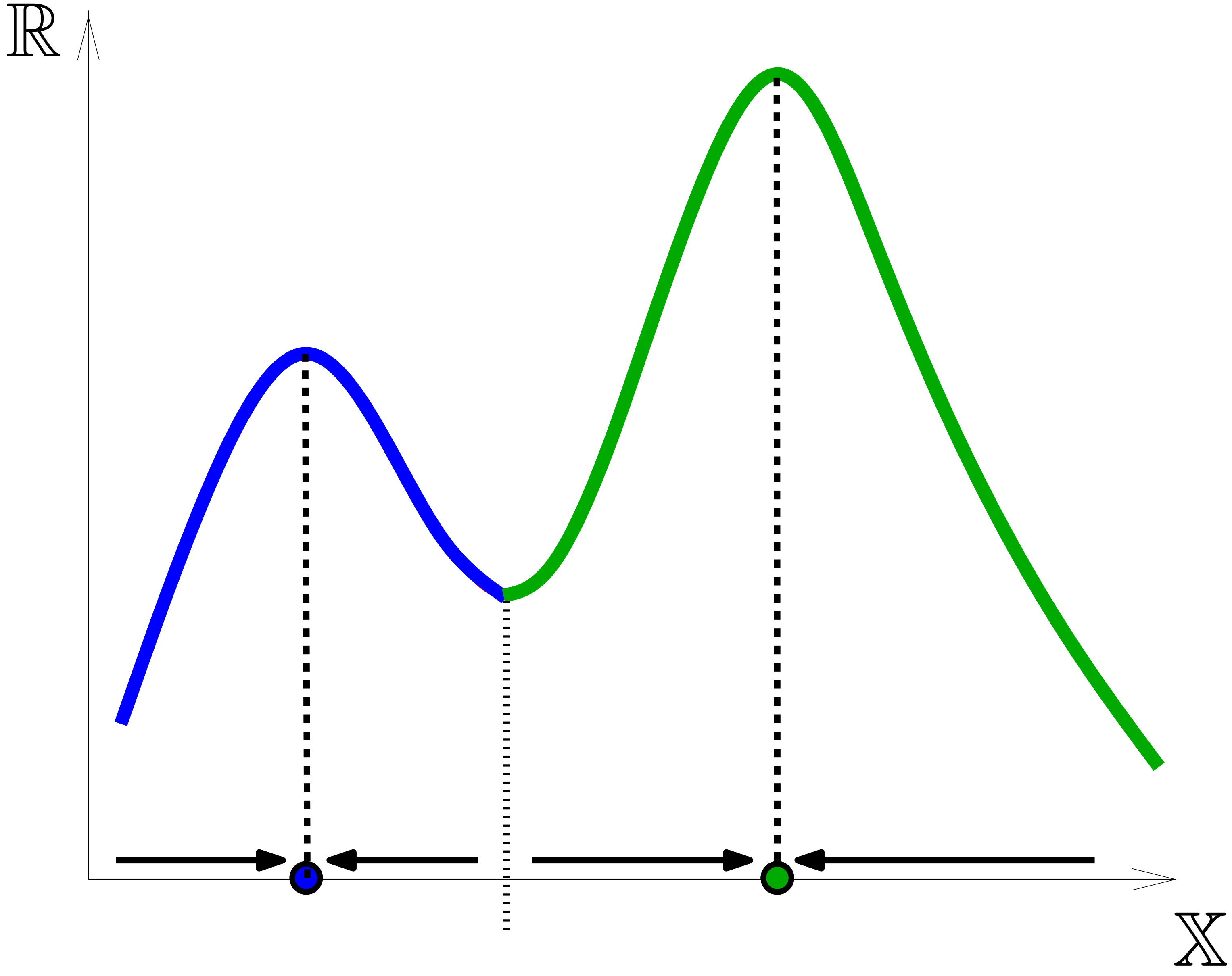
Definition of a Cluster

- Basins of attraction of “significant” peaks of f



Definition of a Cluster

- Basins of attraction of "significant" peaks of f



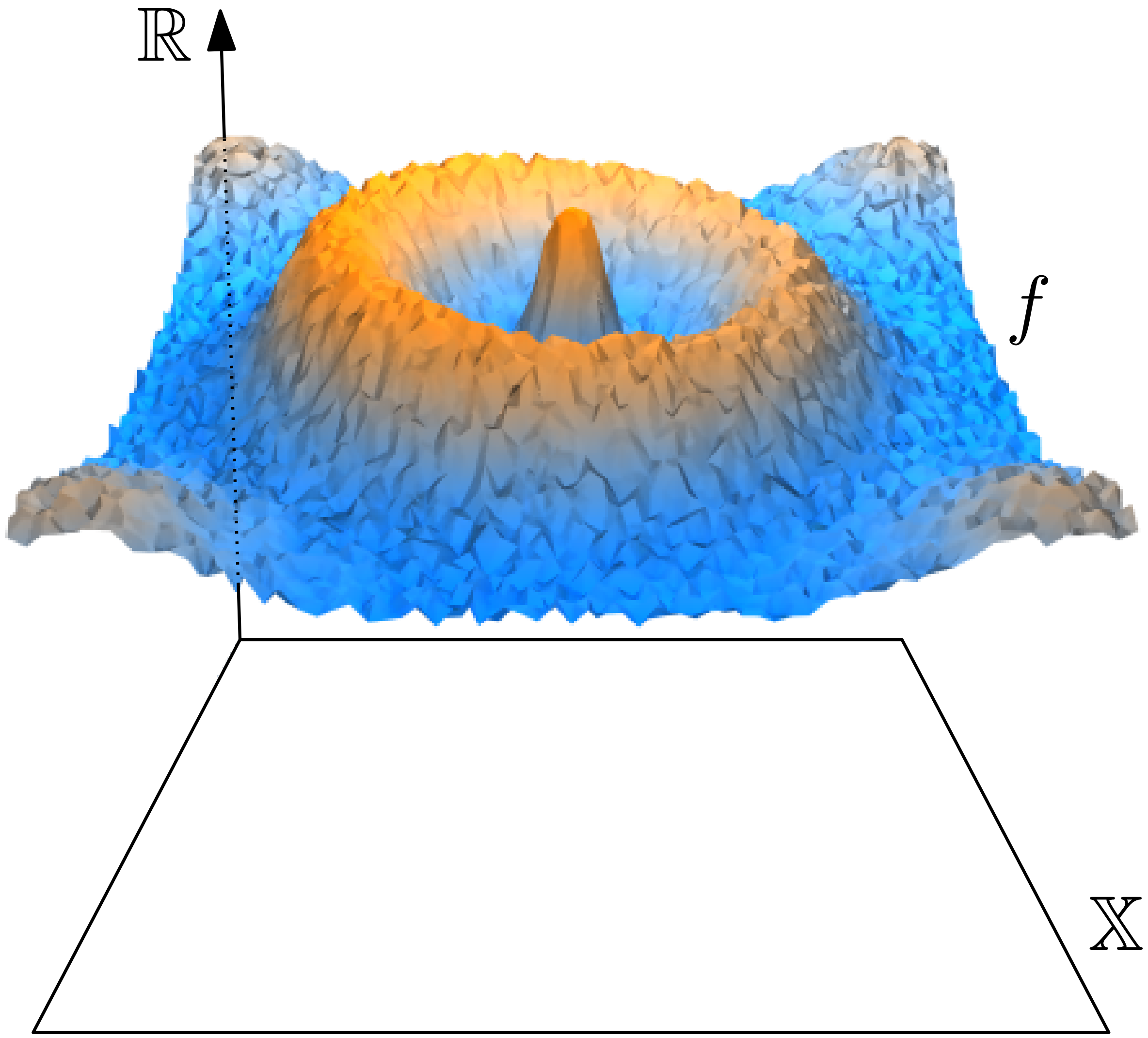
Outline

- Background: scalar field analysis
- Algorithm
- Number of clusters
- Results (Interpretation of persistence diagrams)
- Spatial stability
- Conclusions

Scalar Field Analysis*

Setting: $\mathbb{X}$ topological space, $f : \mathbb{X} \to \mathbb{R}$

Input: A finite sampling L of $\mathbb{X}$,
the values of f at the sample points

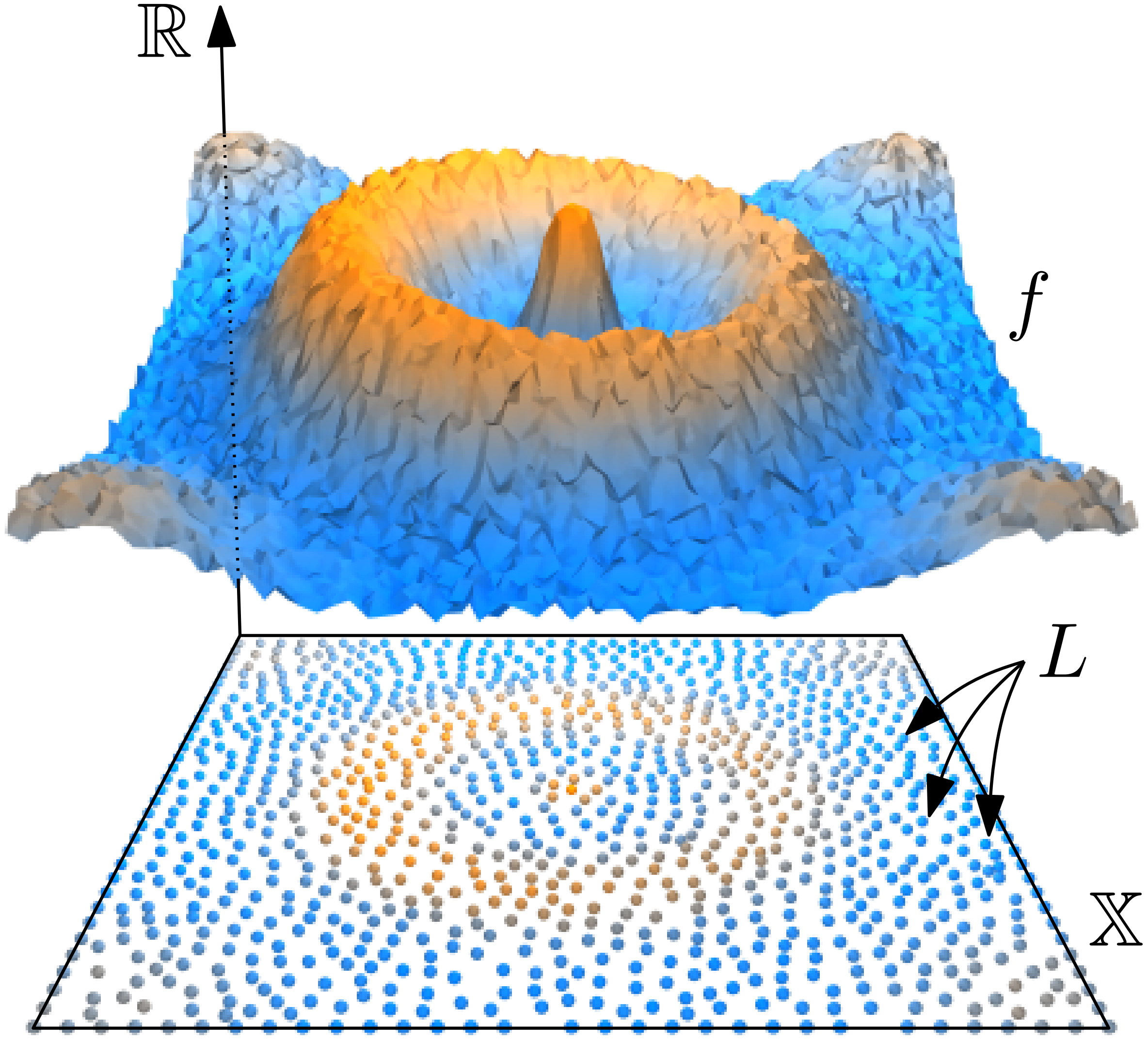


*[Chazal, Guibas, Oudot, Skraba '09]

Scalar Field Analysis*

Setting: $\mathbb{X}$ topological space, $f : \mathbb{X} \to \mathbb{R}$

Input: A finite sampling L of $\mathbb{X}$,
the values of f at the sample points



*[Chazal, Guibas, Oudot, Skraba '09]

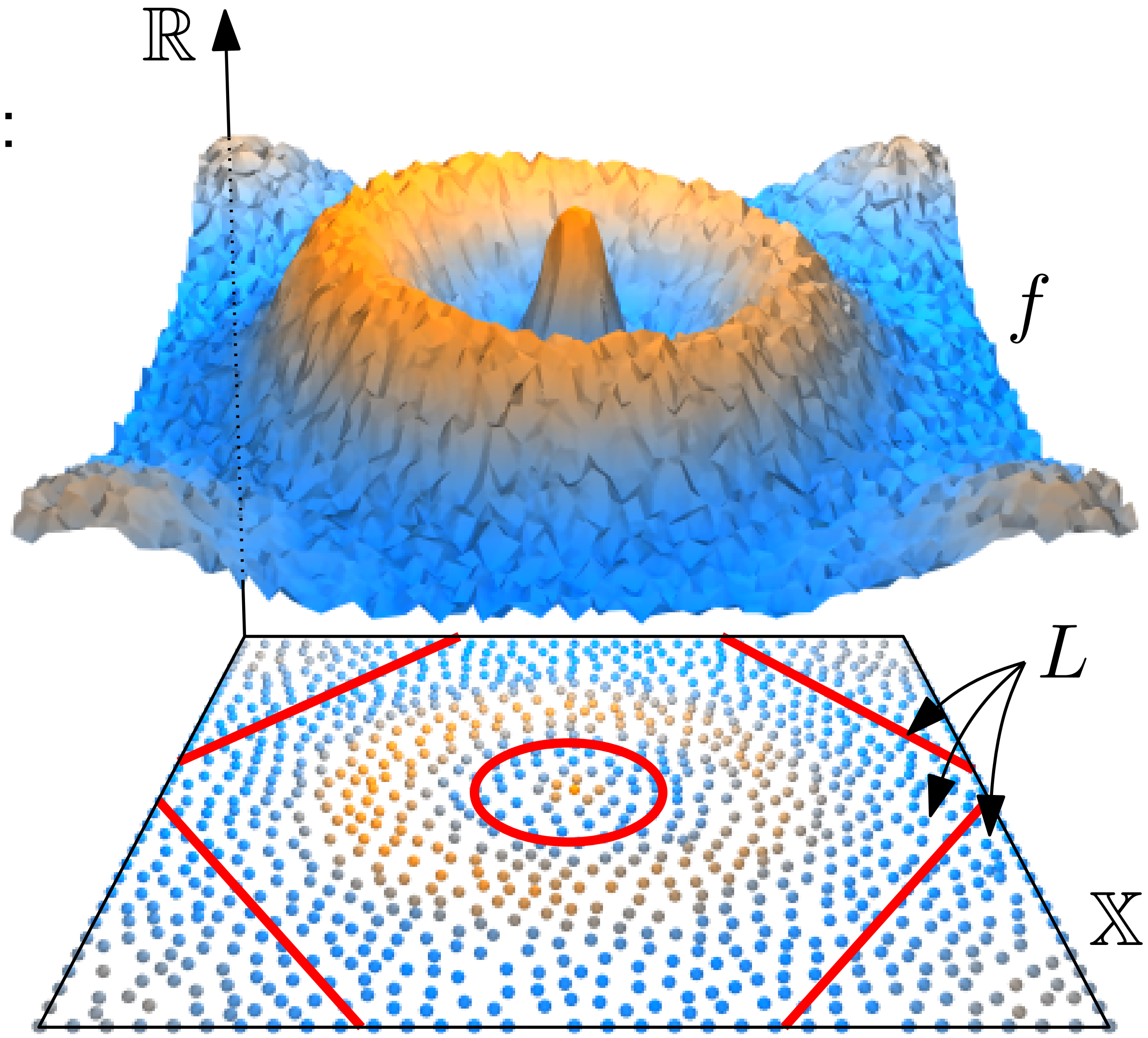
Scalar Field Analysis*

Setting: $\mathbb{X}$ topological space, $f : \mathbb{X} \to \mathbb{R}$

Input: A finite sampling L of $\mathbb{X}$,
the values of f at the sample points

Goal: Analyze landscape of graph(f):

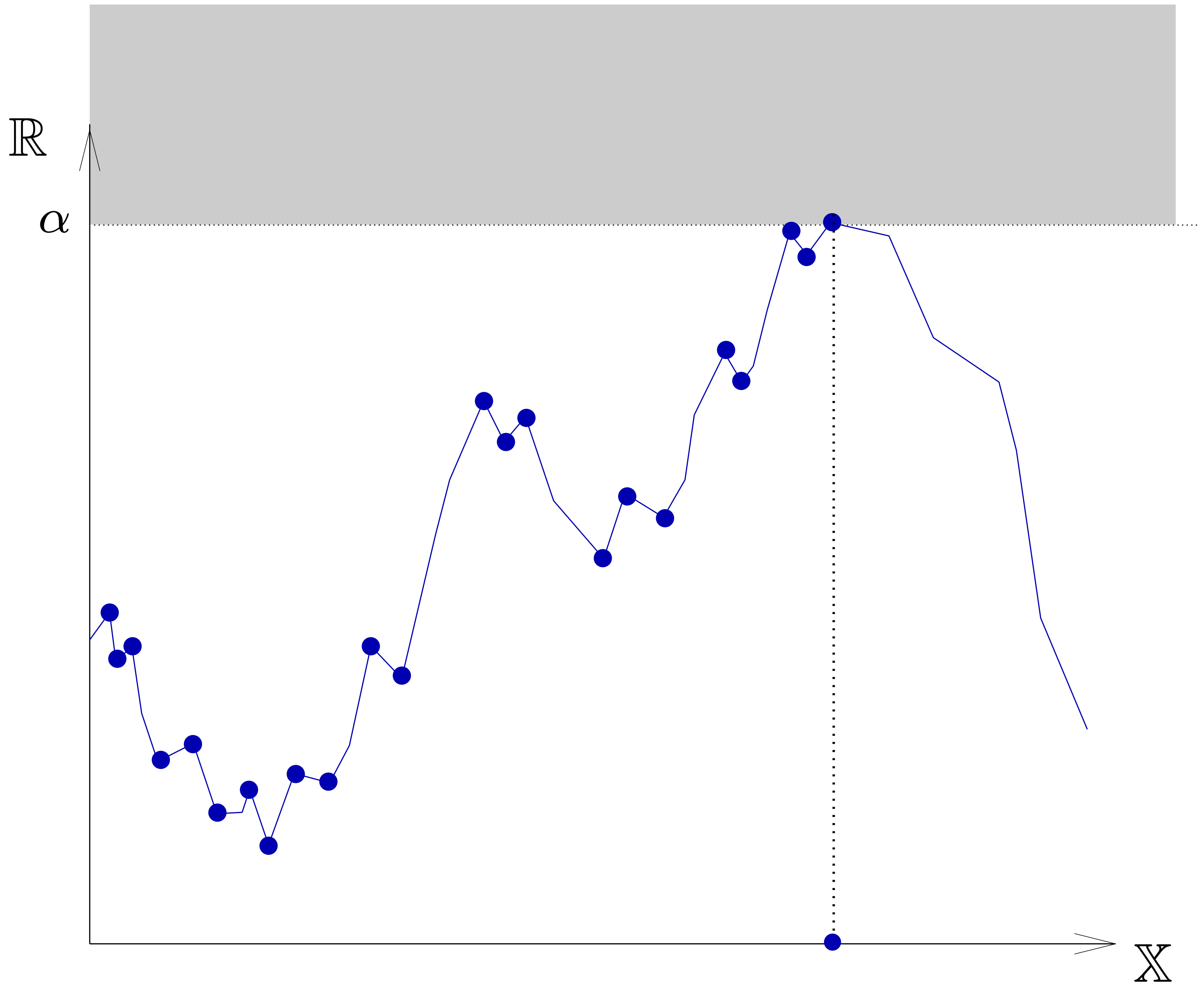
- *prominent* peaks/valleys
- basins of attraction



*[Chazal, Guibas, Oudot, Skraba '09]

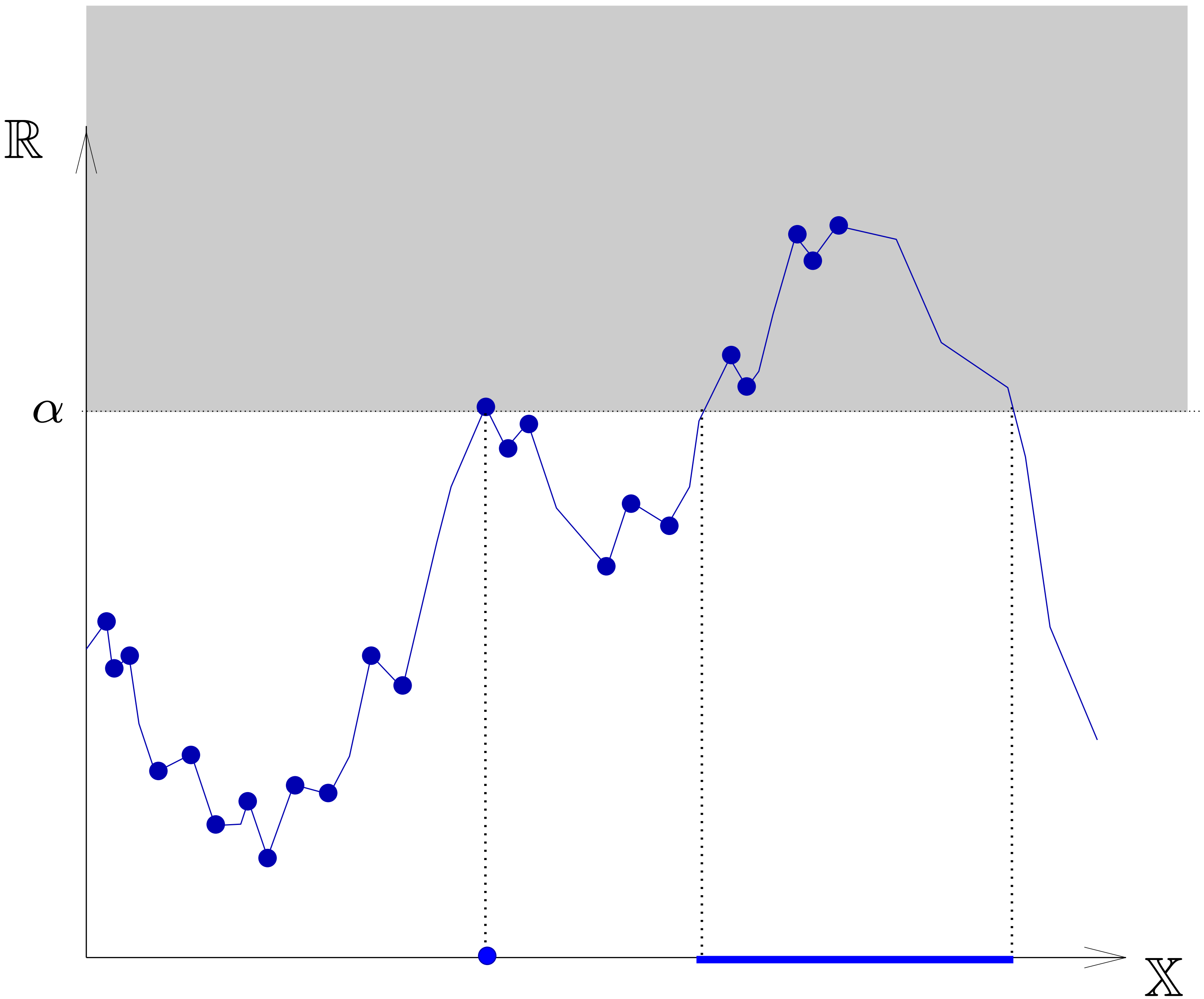
Persistence-Based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as α spans $\mathbb{R}$.



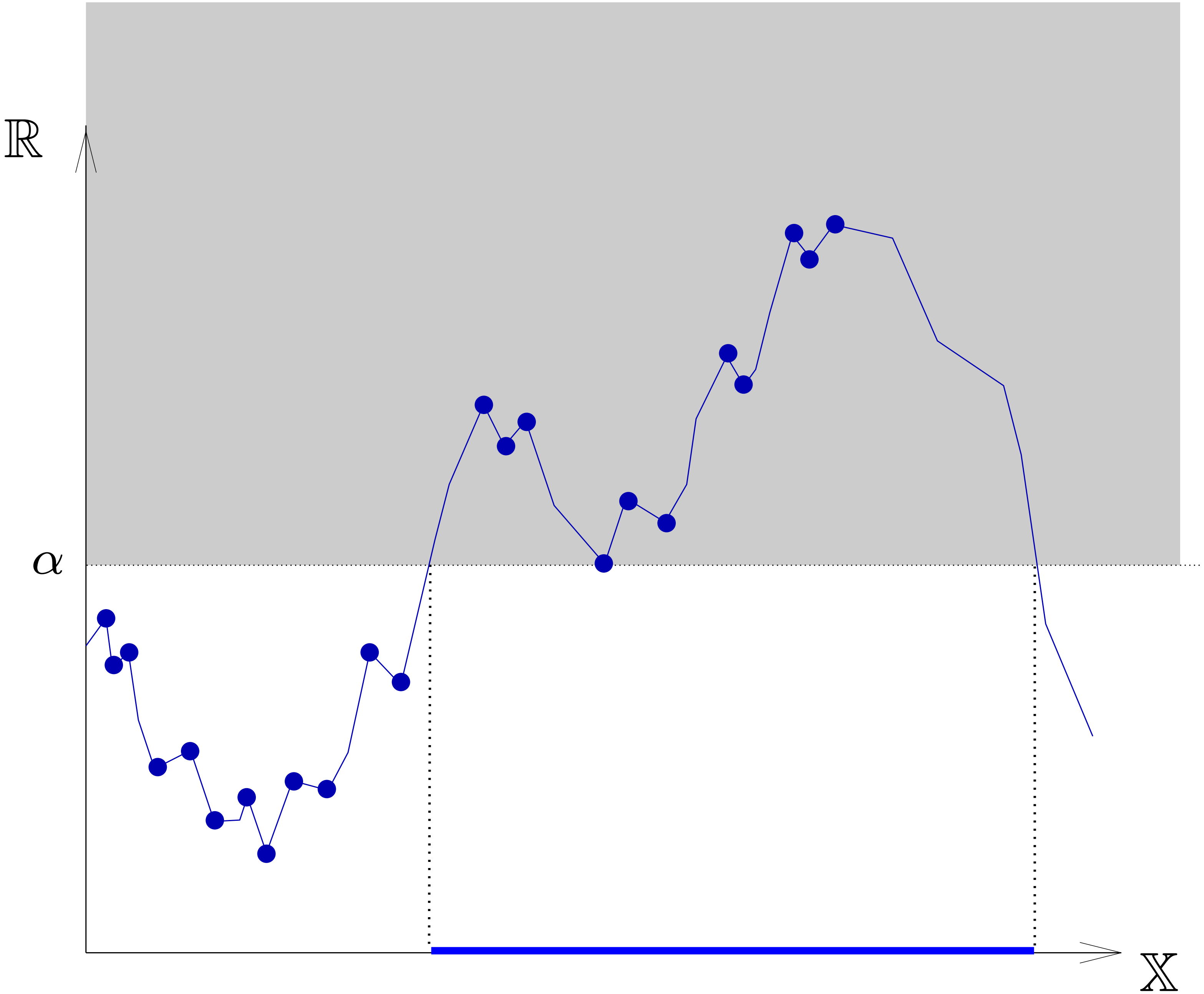
Persistence-Based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as α spans $\mathbb{R}$.



Persistence-Based Approach in a nutshell...

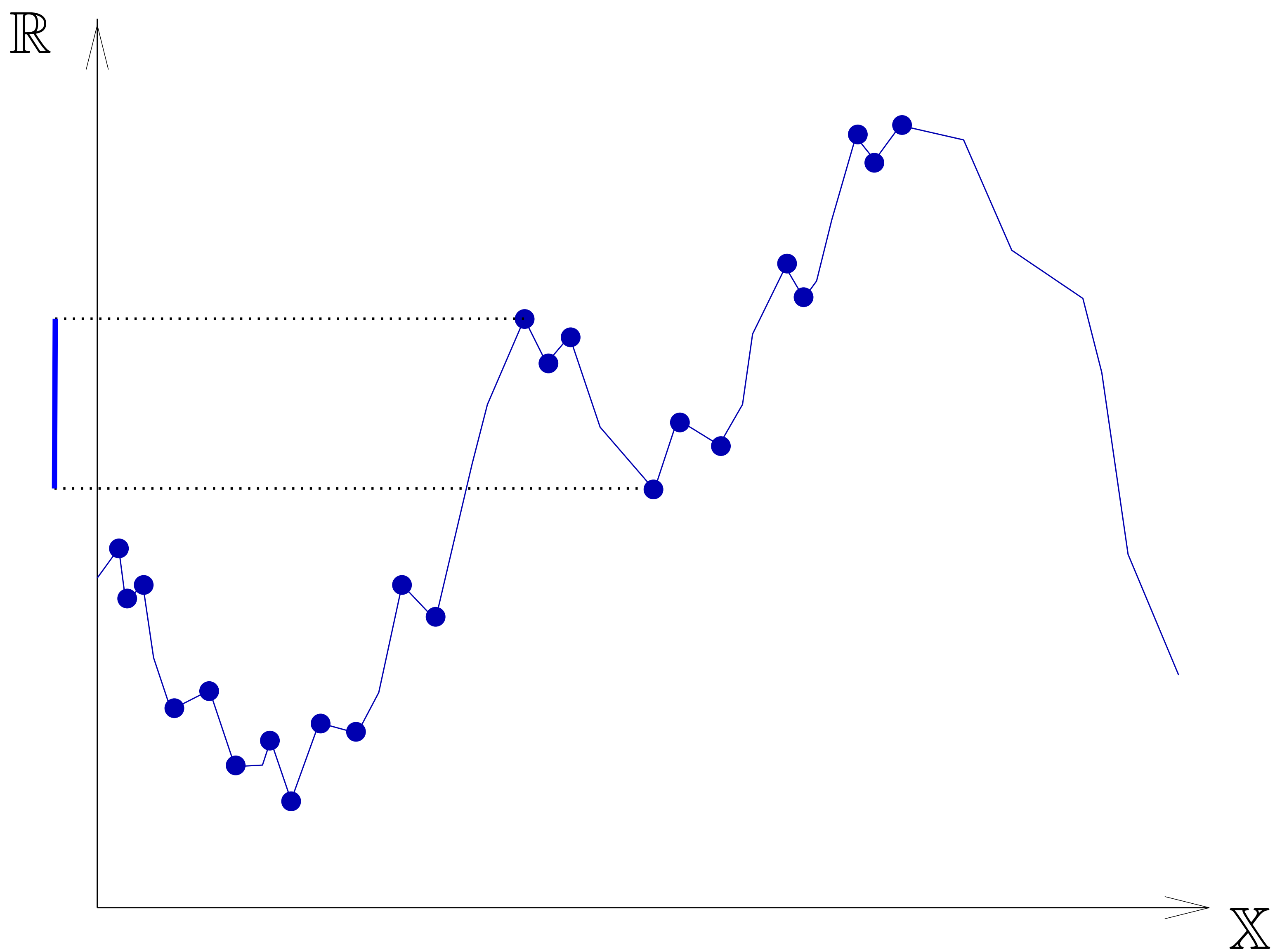
- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as α spans $\mathbb{R}$.



Persistence-Based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as α spans $\mathbb{R}$.

- finite set of intervals (barcode) encode birth/death of homological features.



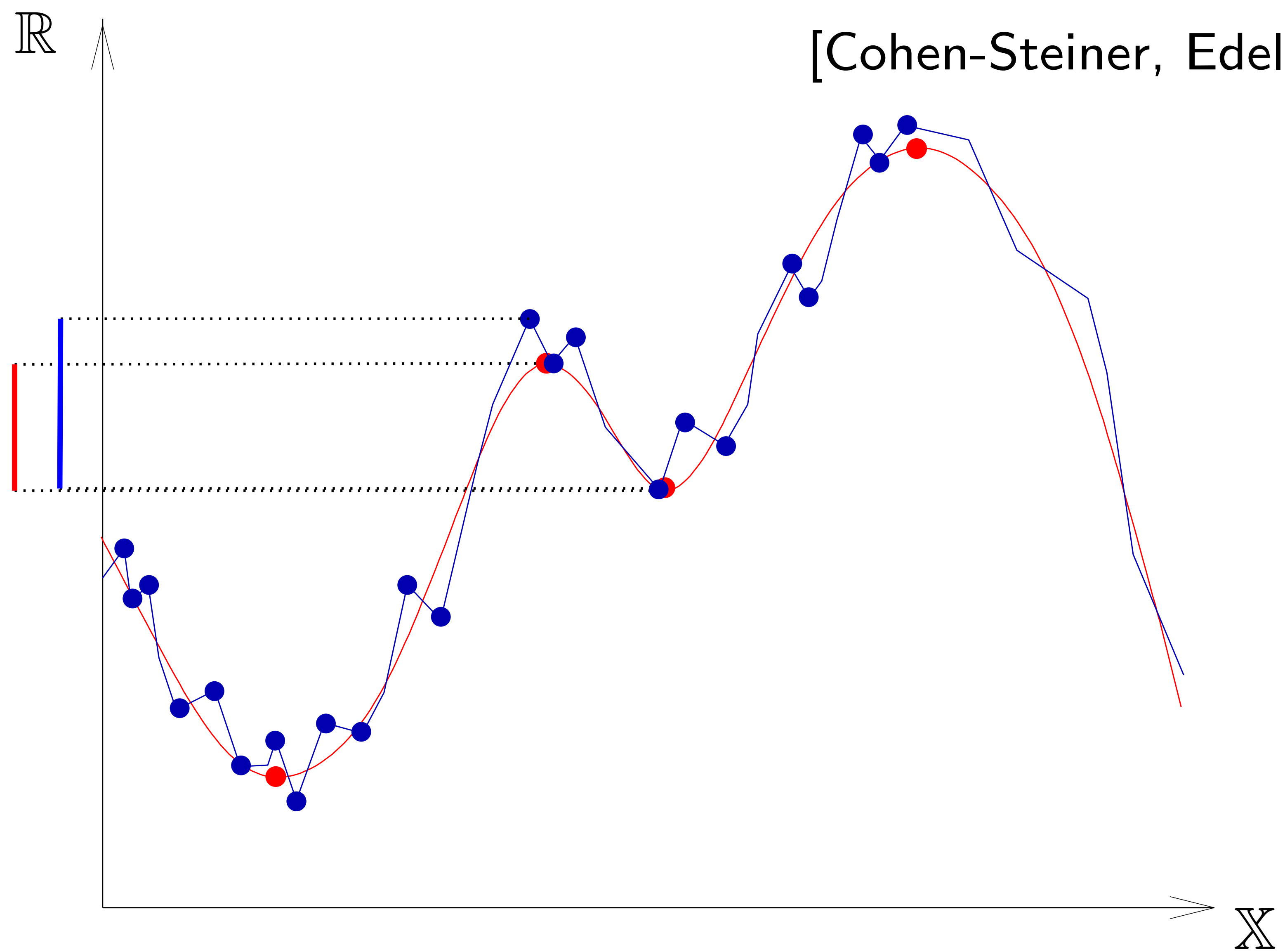
Persistence-Based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as α spans $\mathbb{R}$.

- finite set of intervals (barcode) encode birth/death of homological features.

- barcode of $\hat{f}$ is close to barcode of f provided that $\|\hat{f} - f\|_\infty$ is small.

[Cohen-Steiner, Edelsbrunner, Harer '05]

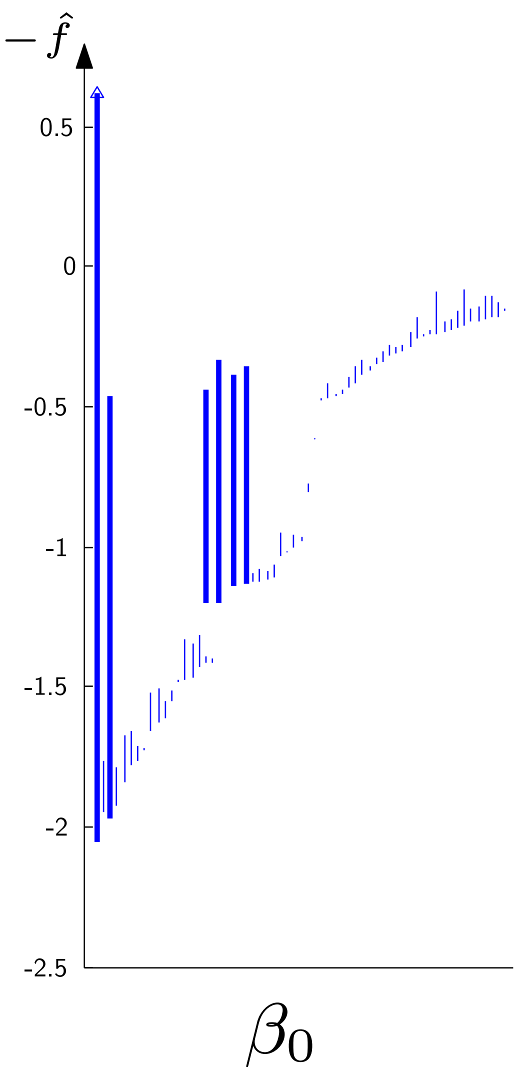


Persistence-Based Approach

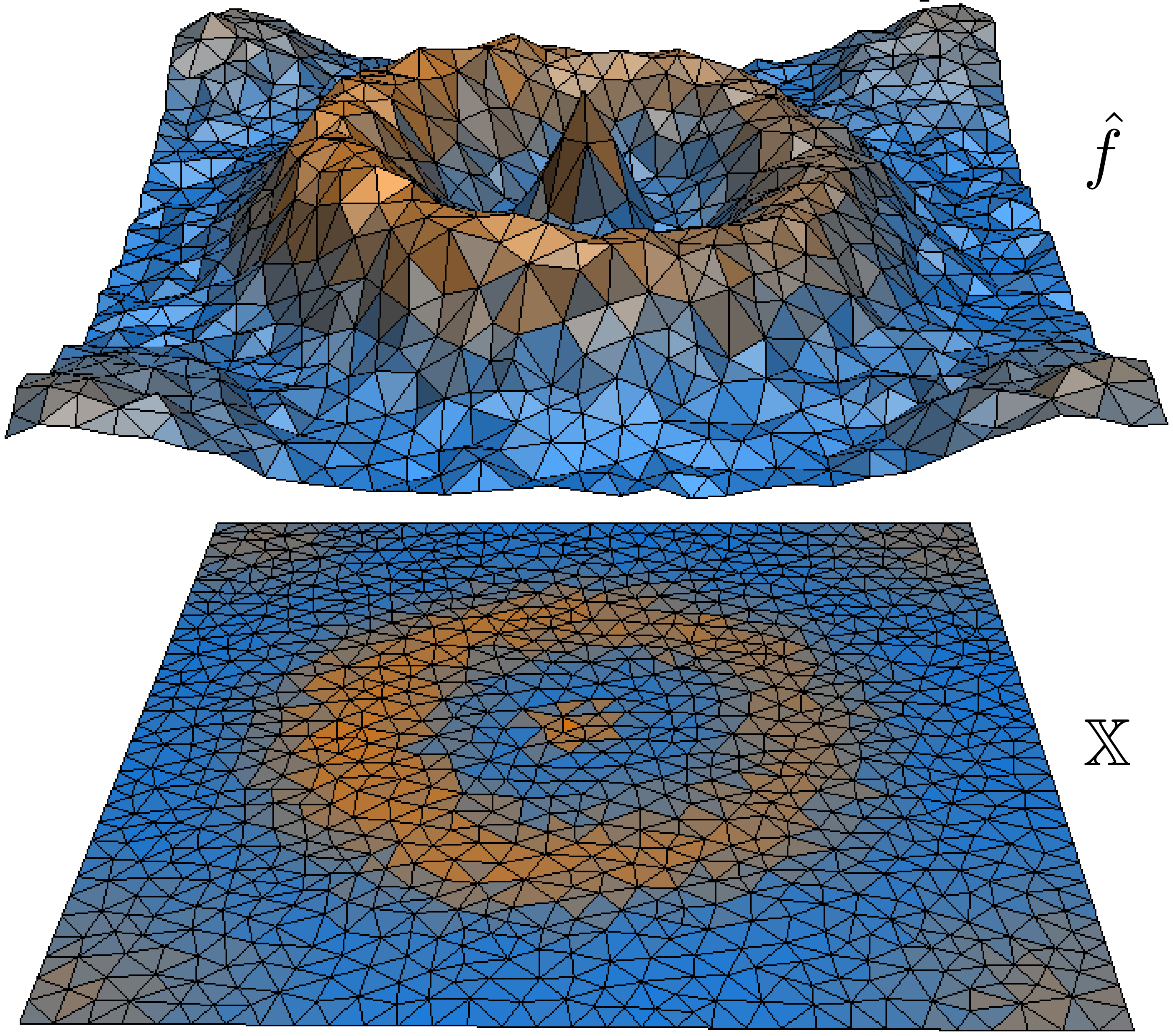
Assumptions: $\mathbb{X}$ triangulated space, $f : \mathbb{X} \to \mathbb{R}$ Lipschitz continuous

$\rightarrow$ build PL approximation $\hat{f}$ of f

$\rightarrow$ apply persistence algo. to $\pm\hat{f}$ [Edelsbrunner, Letscher, Zomorodian '00]



(6 prominent peaks)

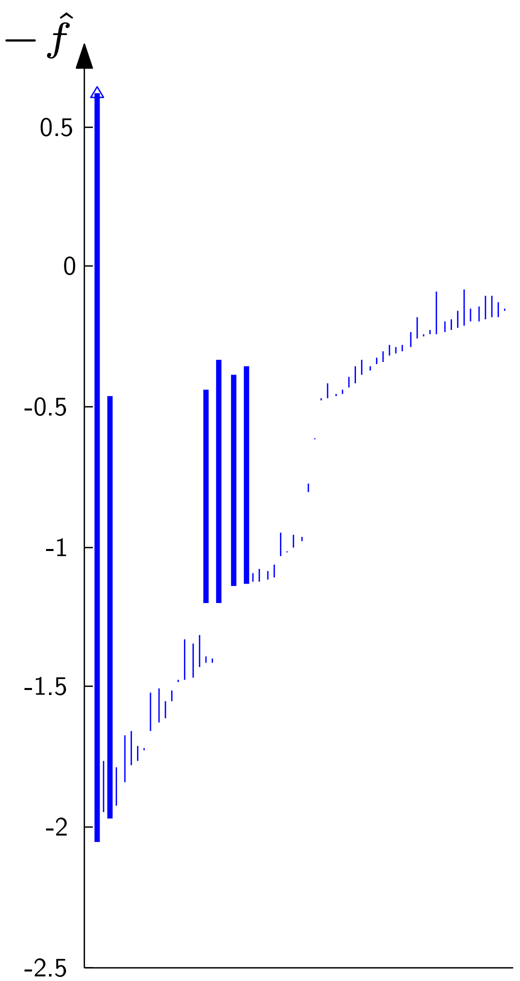


Persistence-Based Approach

Assumptions: $\mathbb{X}$ triangulated space, $f : \mathbb{X} \to \mathbb{R}$ Lipschitz continuous

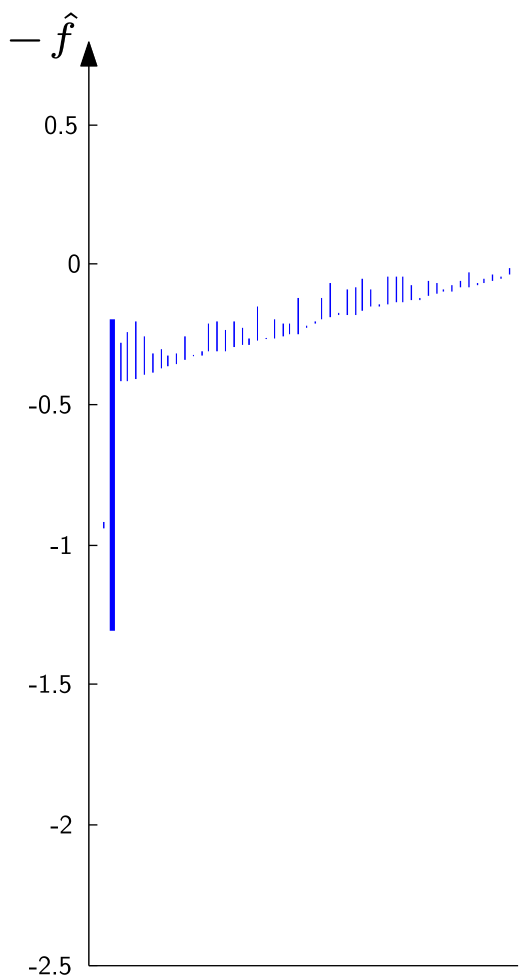
$\longrightarrow$ build PL approximation $\hat{f}$ of f

$\longrightarrow$ apply persistence algo. to $\pm\hat{f}$ [Edelsbrunner, Letscher, Zomorodian '00]



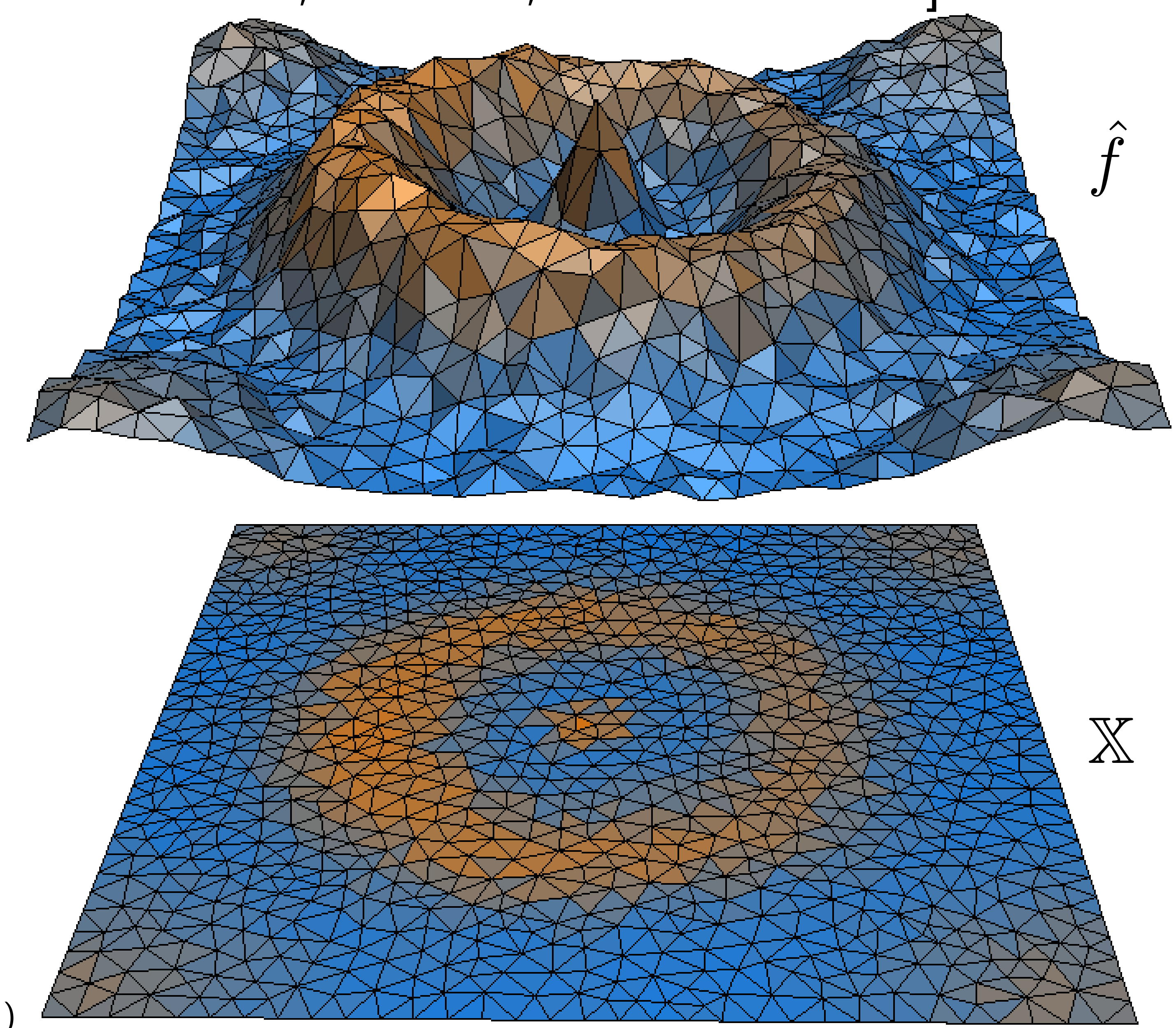
β_0

(6 prominent peaks)



β_1

(ring-shaped basin of attraction)



Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

- Access to L **not** $\mathbb{X}$

Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

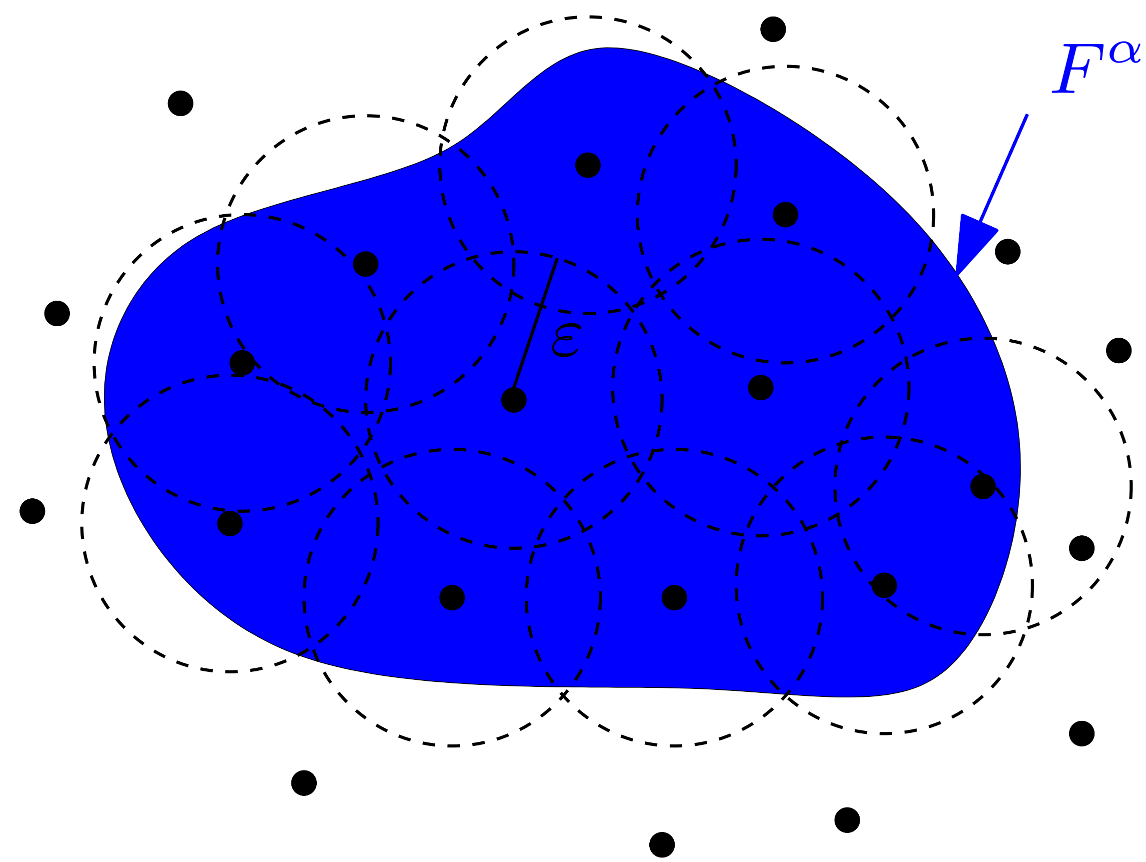
- Access to L **not** $\mathbb{X}$

$$F^\alpha := f^{-1}([\alpha, \infty))$$

$$L_\alpha := L \cap F^\alpha$$

$$L_\alpha^\varepsilon := \bigcup_{p \in L_\alpha} B_{\mathbb{X}}(p, \varepsilon)$$

$$\boxed{\forall \alpha \in \mathbb{R},\ L_{\alpha + c\varepsilon}^\varepsilon \subseteq F^\alpha \subseteq L_{\alpha - c\varepsilon}^\varepsilon}$$



Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

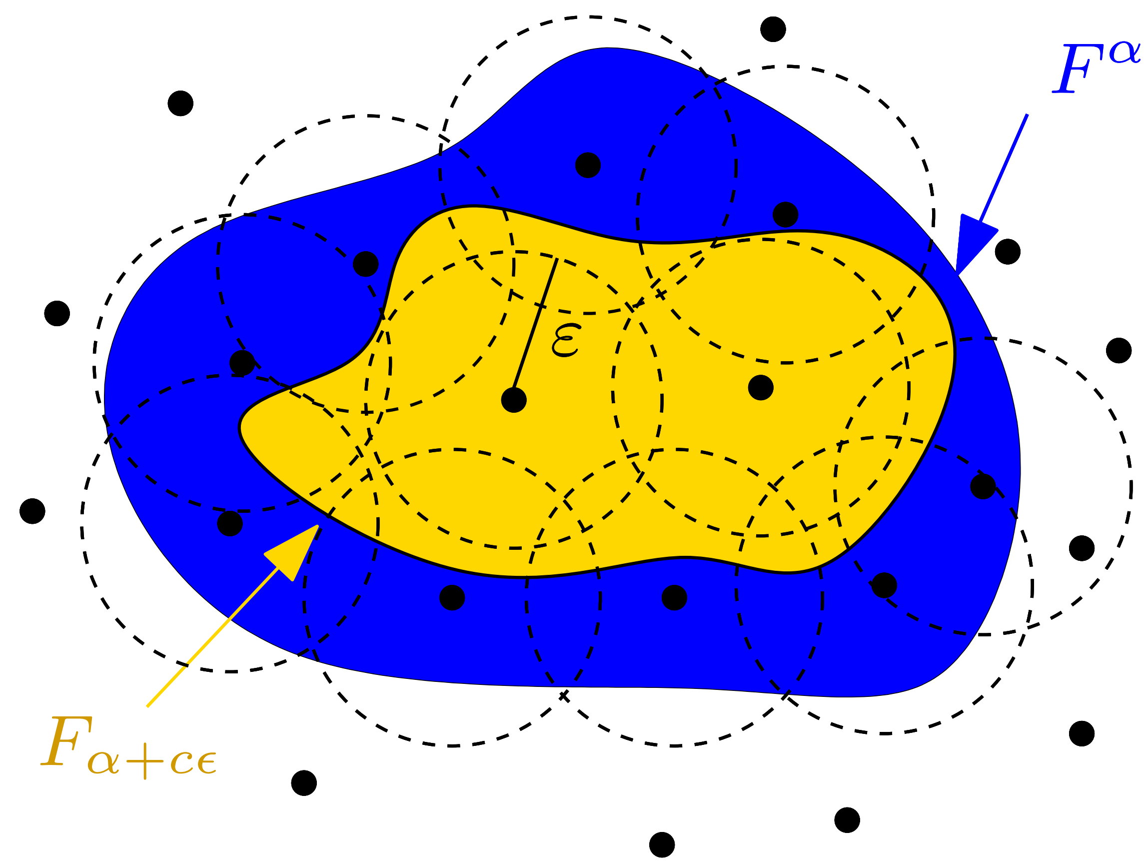
- Access to L **not** $\mathbb{X}$

$$F^\alpha := f^{-1}([\alpha, \infty))$$

$$L_\alpha := L \cap F^\alpha$$

$$L_\alpha^\varepsilon := \bigcup_{p \in L_\alpha} B_{\mathbb{X}}(p, \varepsilon)$$

$$\boxed{\forall \alpha \in \mathbb{R},\ L_{\alpha+c\varepsilon}^\varepsilon \subseteq F^\alpha \subseteq L_{\alpha-c\varepsilon}^\varepsilon}$$



Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

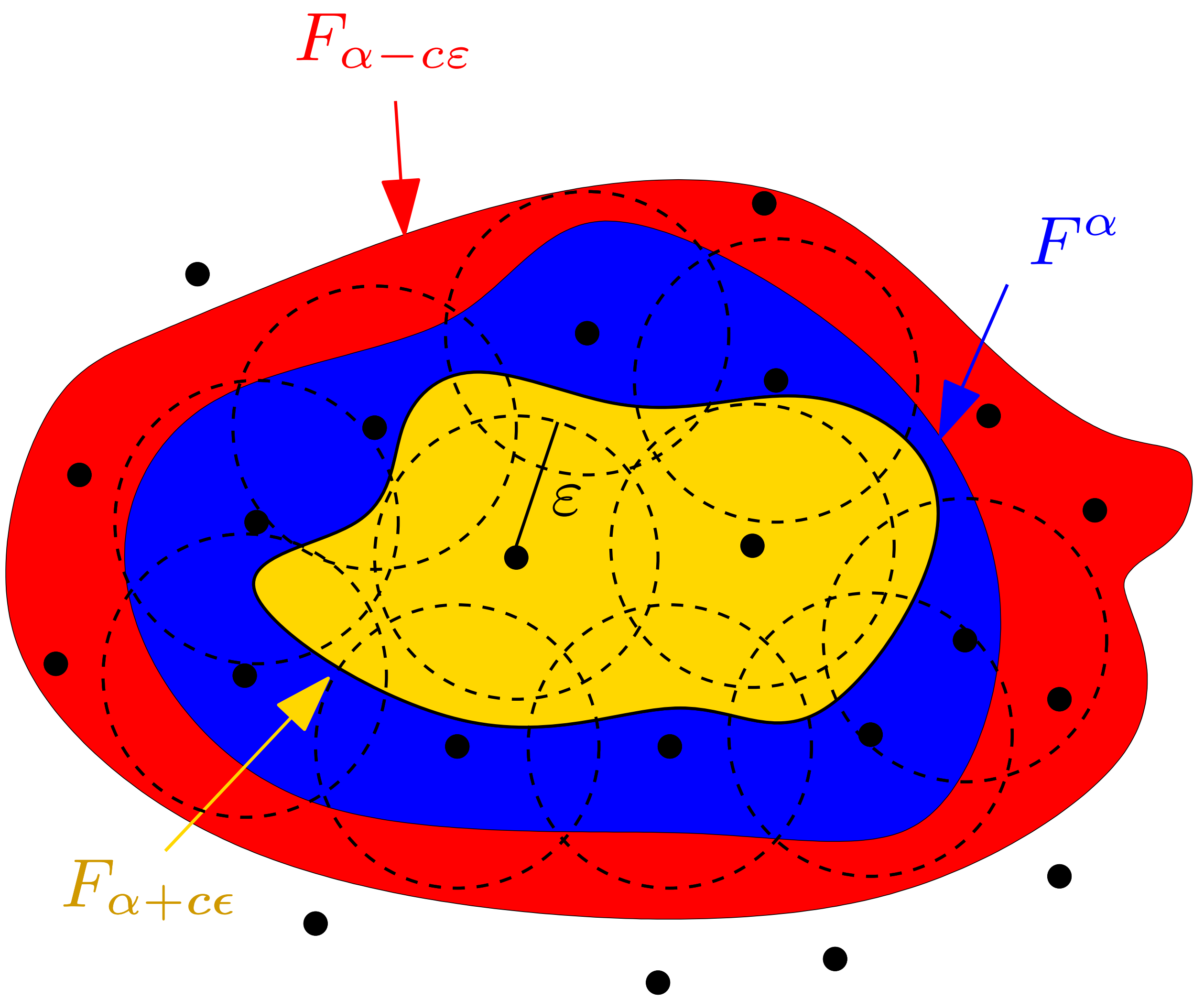
- Access to L **not** $\mathbb{X}$

$$F^\alpha := f^{-1}([\alpha, \infty))$$

$$L_\alpha := L \cap F^\alpha$$

$$L_\alpha^\varepsilon := \bigcup_{p \in L_\alpha} B_{\mathbb{X}}(p, \varepsilon)$$

$$\boxed{\forall \alpha \in \mathbb{R},\ L^\varepsilon_{\alpha+c\varepsilon} \subseteq F^\alpha \subseteq L^\varepsilon_{\alpha-c\varepsilon}}$$



Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

- Access to L **not** $\mathbb{X}$

$$F^\alpha := f^{-1}([\alpha, \infty))$$

$$L_\alpha := L \cap F^\alpha$$

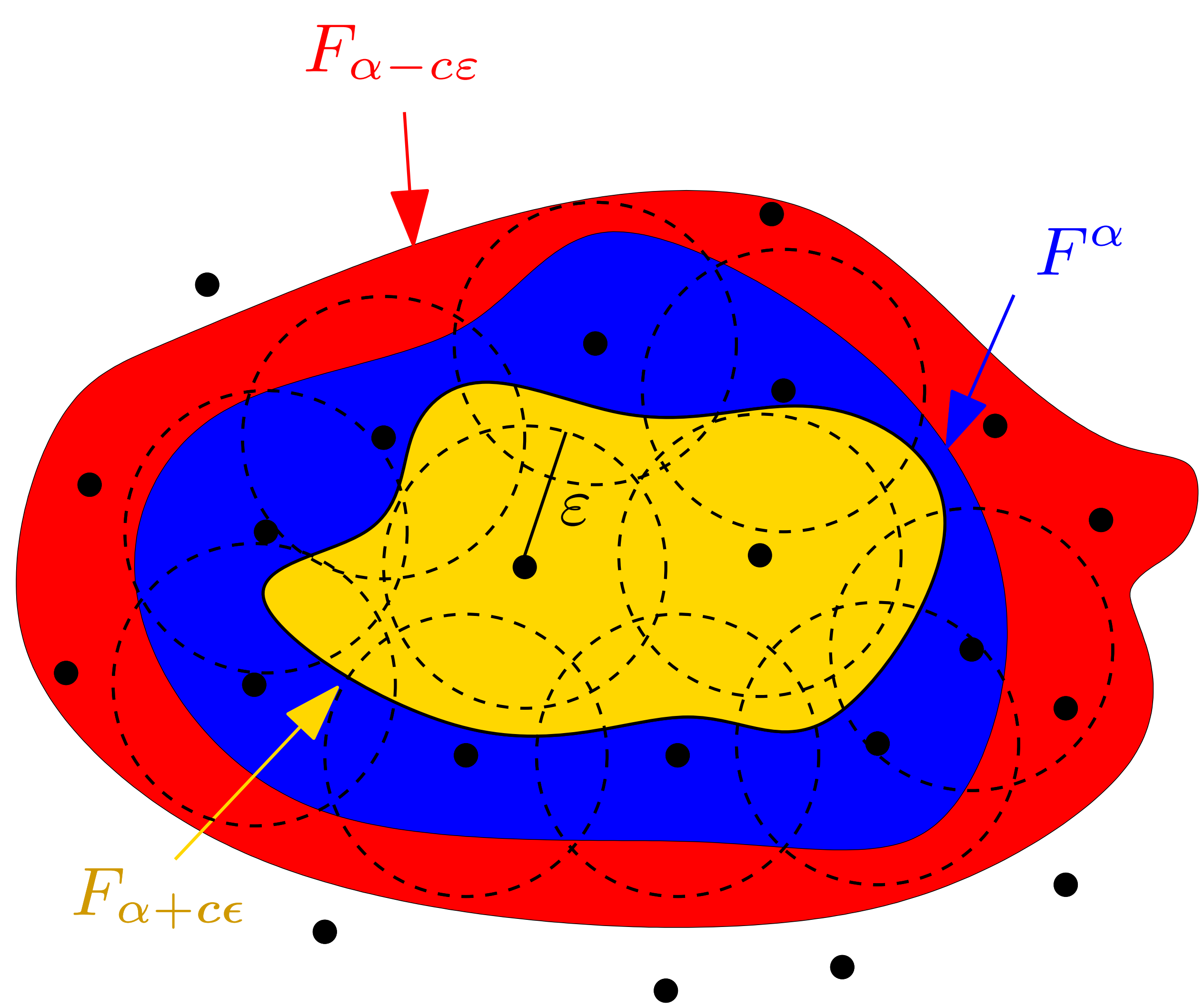
$$L_\alpha^\varepsilon := \bigcup_{p \in L_\alpha} B_{\mathbb{X}}(p, \varepsilon)$$

$$\boxed{\forall \alpha \in \mathbb{R},\ L_{\alpha+c\varepsilon}^\varepsilon \subseteq F^\alpha \subseteq L_{\alpha-c\varepsilon}^\varepsilon}$$

the filtrations $\{F^\alpha\}_{\alpha\in\mathbb{R}}$ and $\{L_\alpha^\varepsilon\}_{\alpha\in\mathbb{R}}$ are $c\varepsilon$-interleaved

$$\Downarrow$$

their barcodes are $c\varepsilon$-close.



[Chazal, Cohen-Steiner, Glisse, Guibas, Oudot '09]

Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

Guarantee:

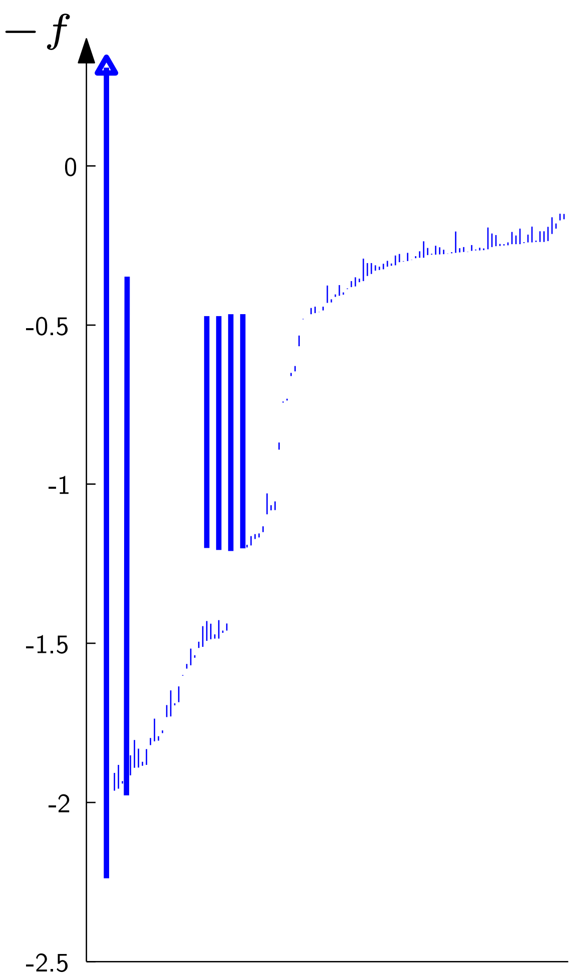
$\forall \delta \geq \varepsilon$, $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ and $\{\mathcal{R}^\delta(L_\alpha) \hookrightarrow \mathcal{R}^{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$ are $2c\delta$-interleaved

$\Downarrow$ [Chazal, Cohen-Steiner, Glisse, Guibas, Oudot '09]

their barcodes are $2c\delta$-close.

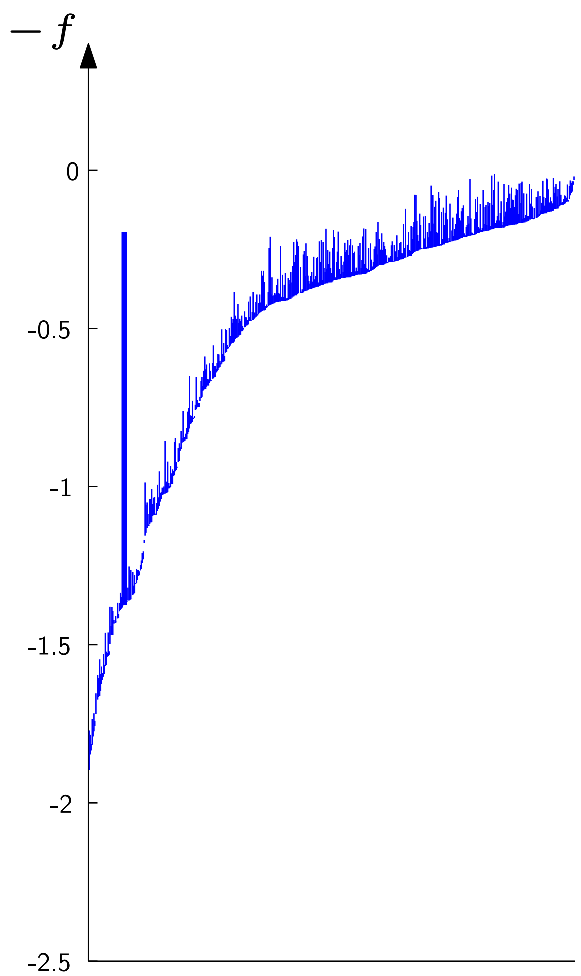
Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.



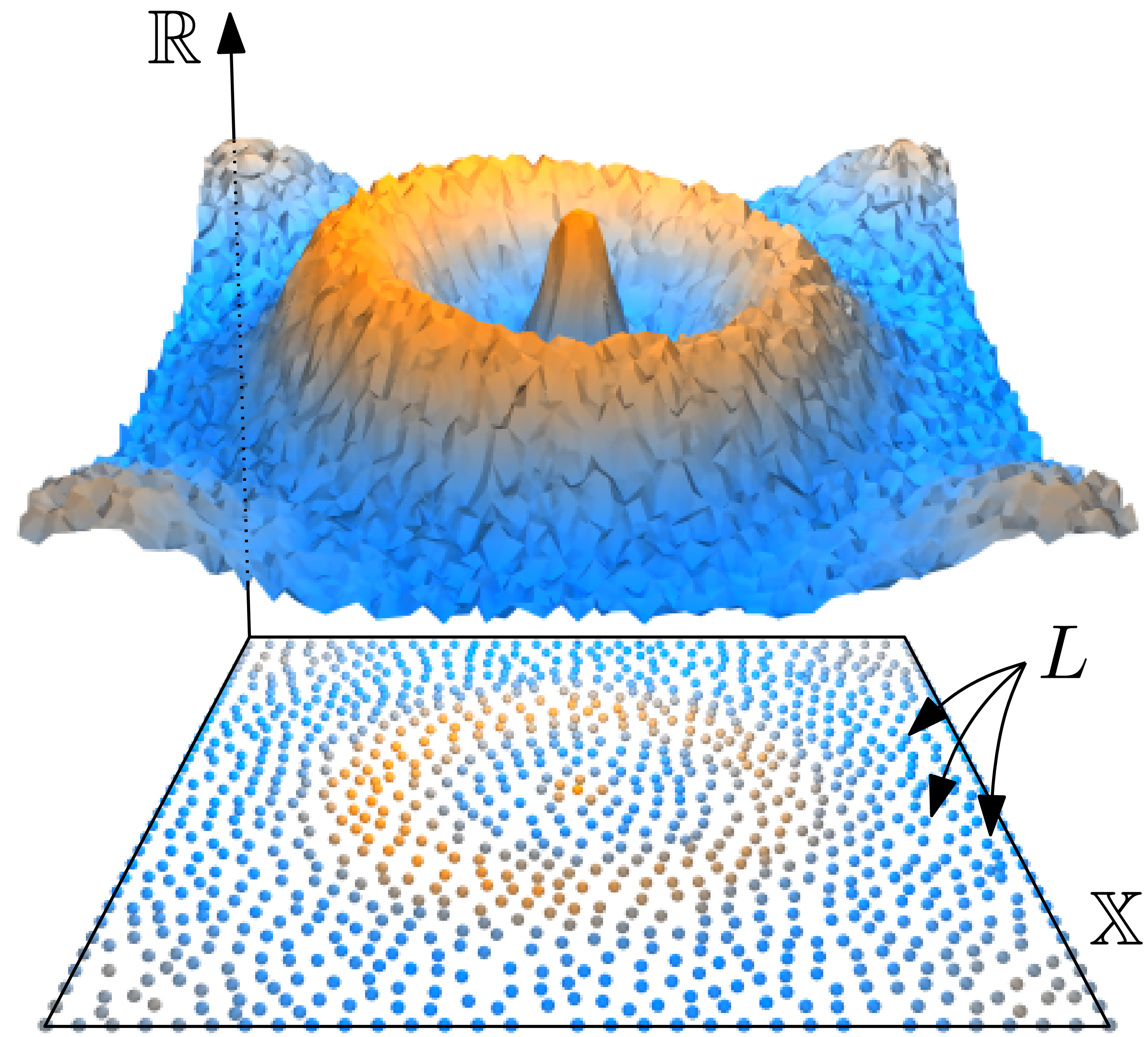
β_0

(6 prominent peaks)



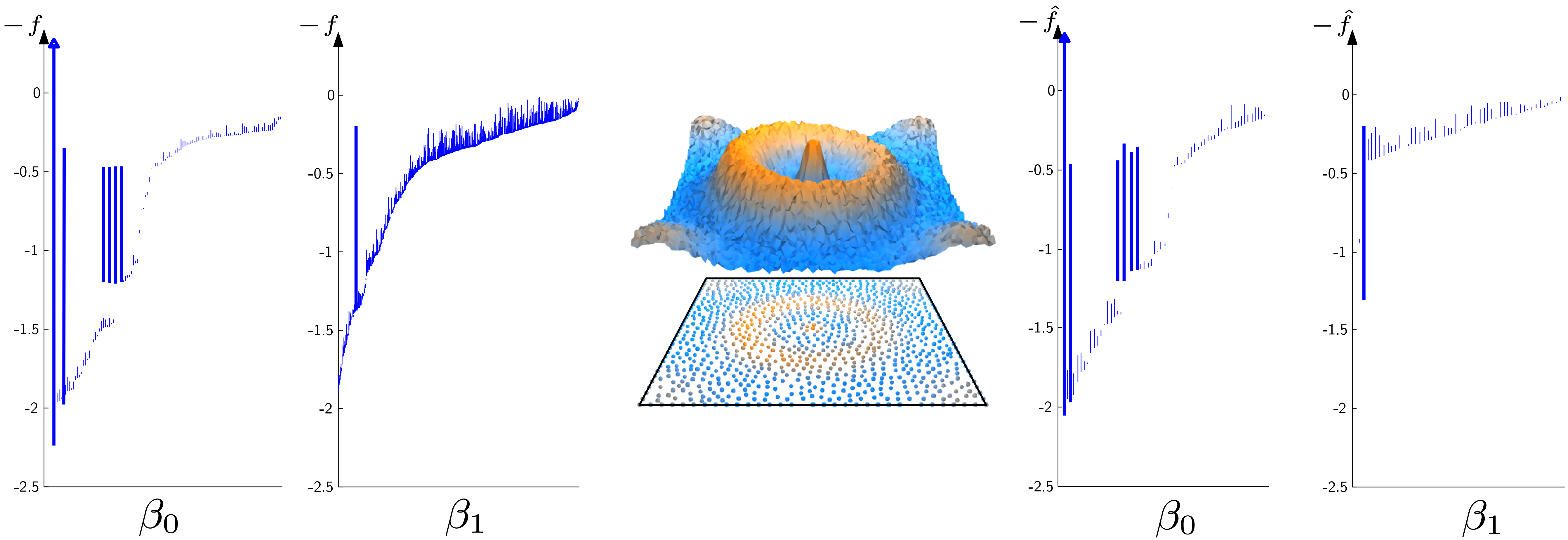
β_1

(ring-shaped basin of attraction)



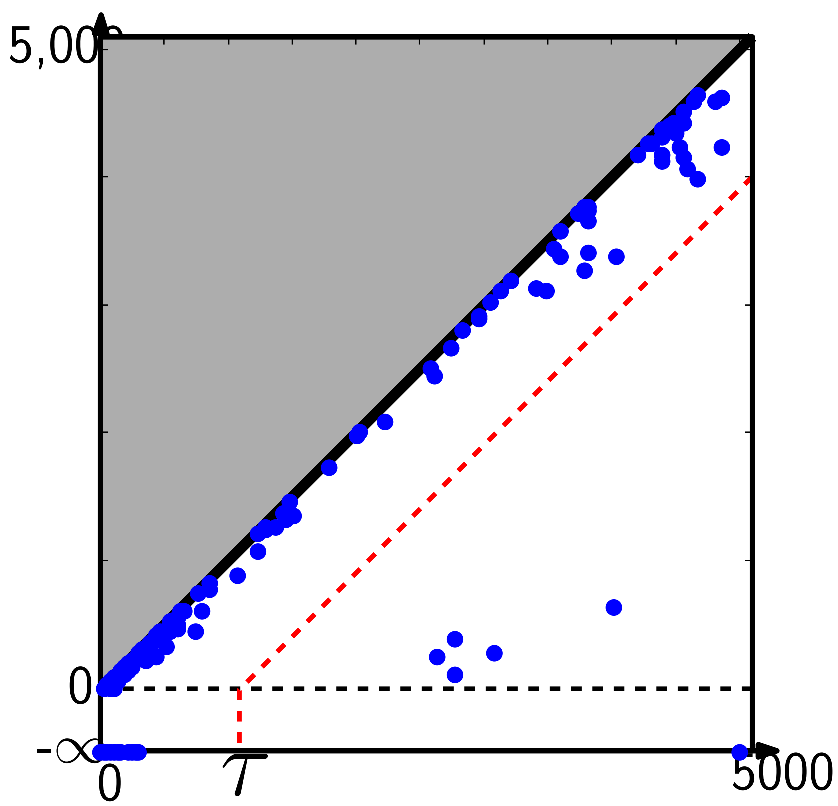
Approximation of Super-Level Sets

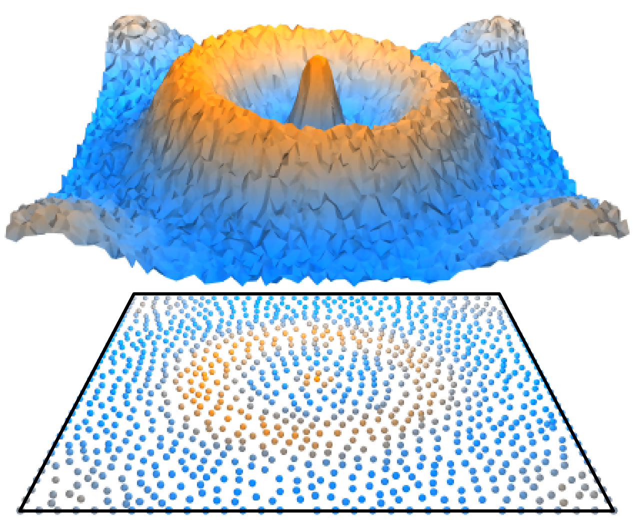
Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

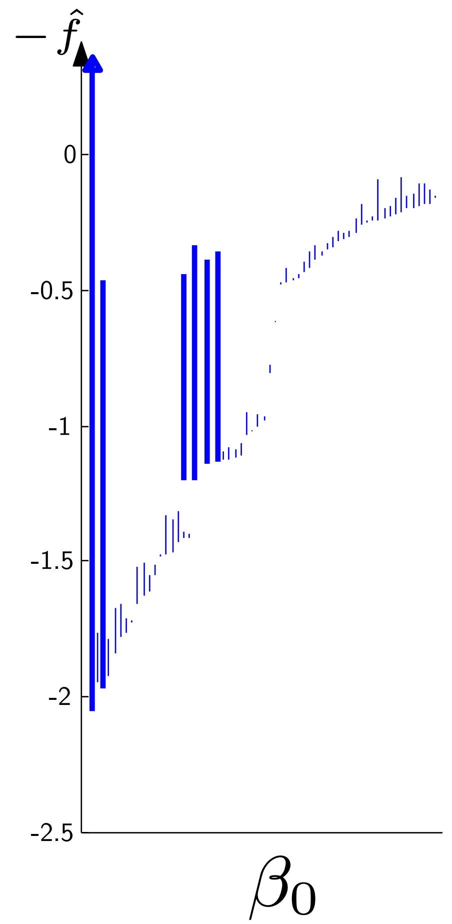


Approximation of Super-Level Sets

Assumptions: $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ c-Lipschitz,
L geodesic ε-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

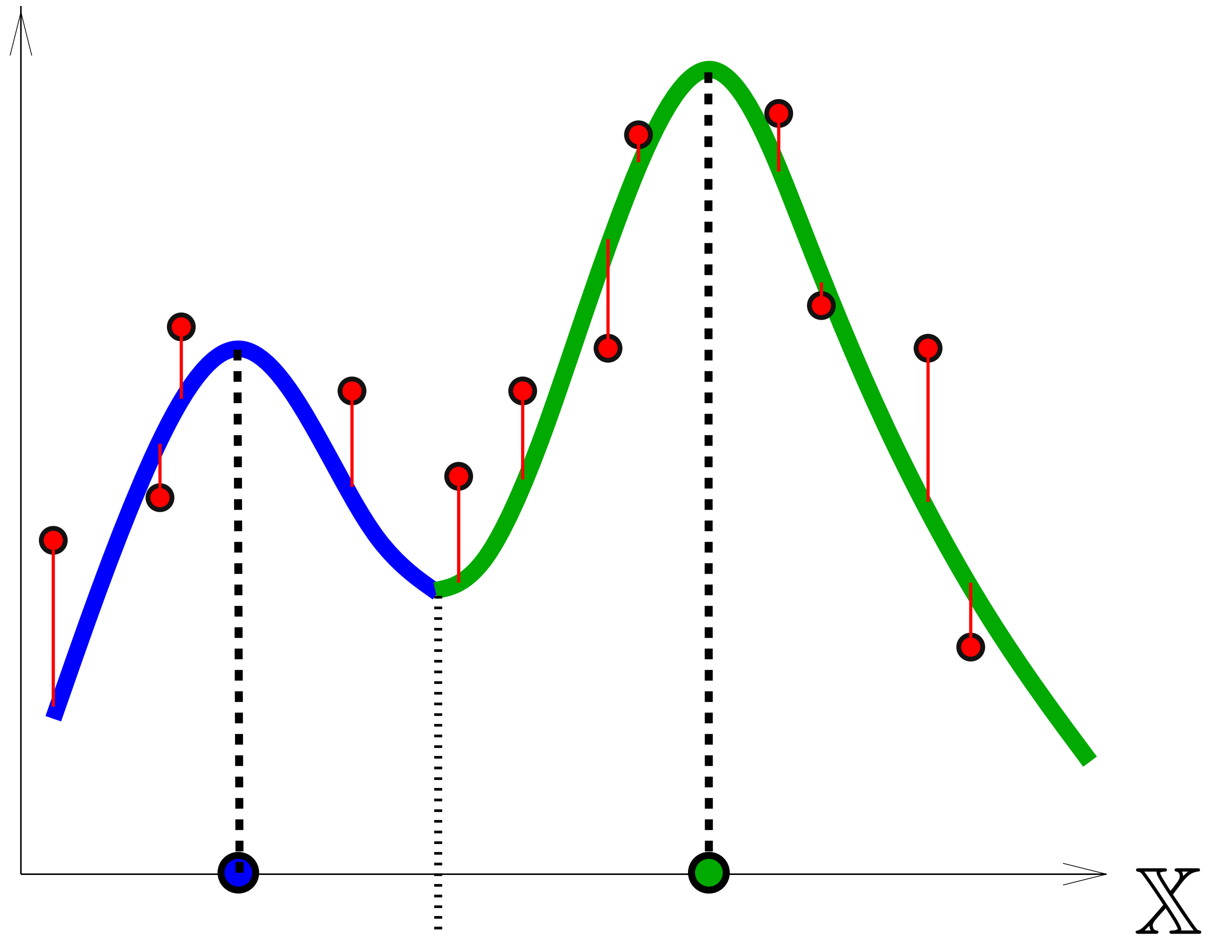






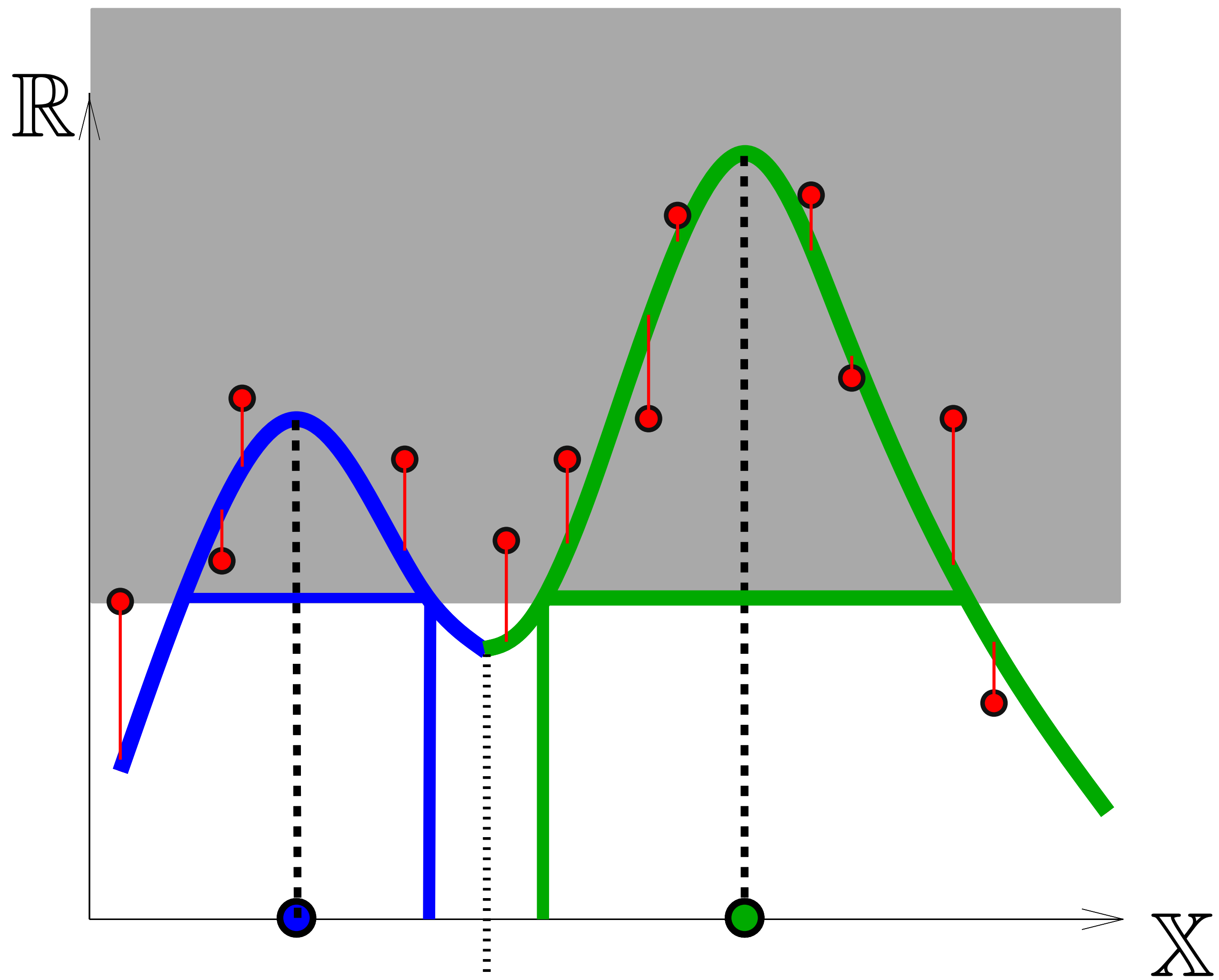
Homological Features and Clusters

- Samples drawn from f
- Estimate $\hat{f}$ from samples



Homological Features and Clusters

- Samples drawn from f
- Estimate $\hat{f}$ from samples



Clusters: Prominent peaks correspond to persistent connected components of the super-level set filtration of f

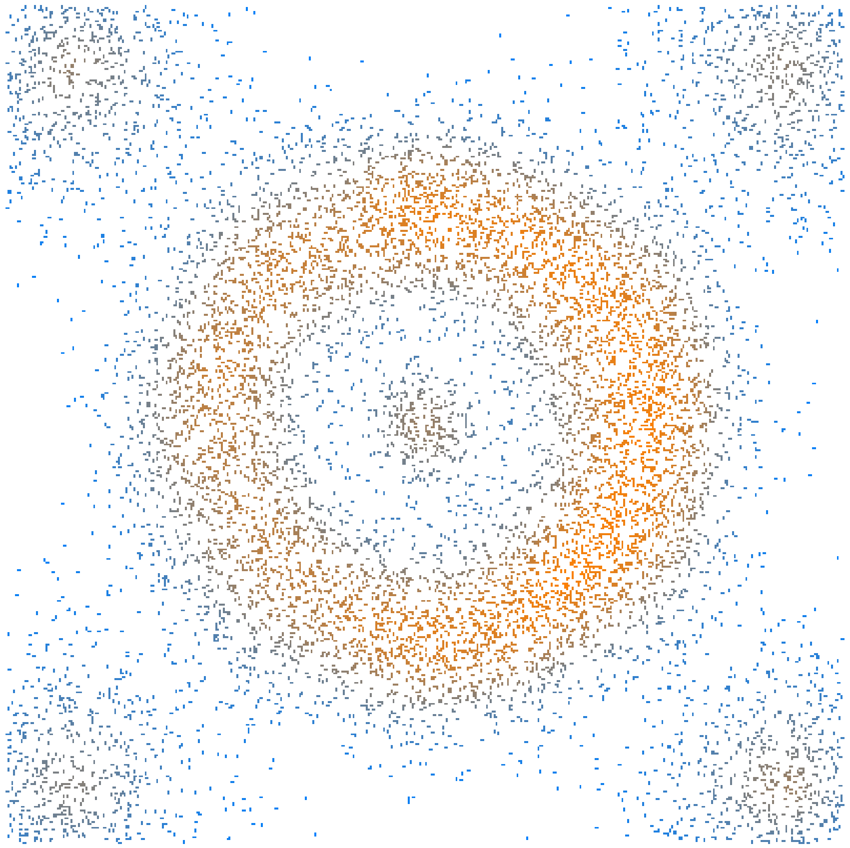
Computing Clusters

How do we compute clusters from a barcode?

Computing Clusters

How do we compute clusters from a barcode?

Input: Samples with estimated density $\hat{f}$



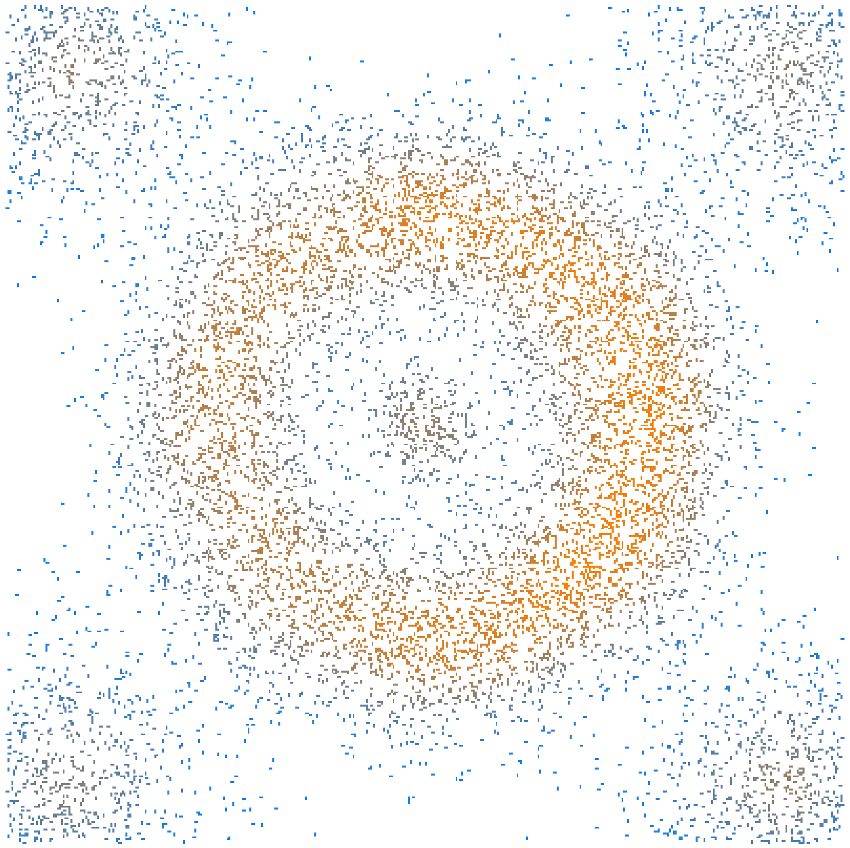
Computing Clusters

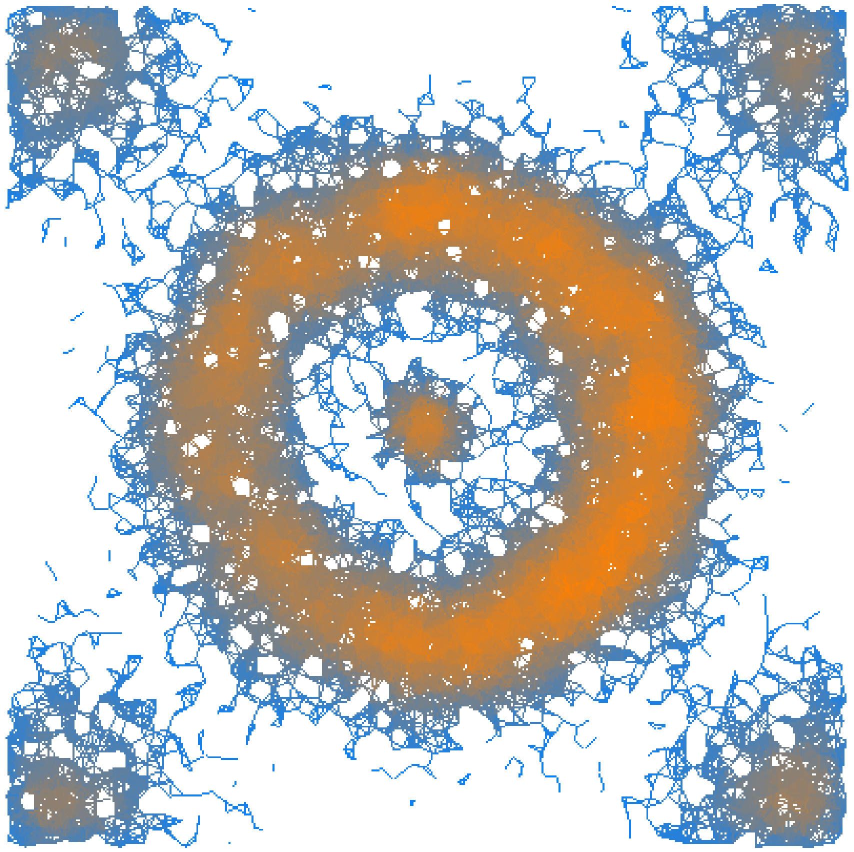
How do we compute clusters from a barcode?

Input: Samples with estimated density $\hat{f}$

Two steps:

1. Mode-seeking step [Koontz et. al. '76]





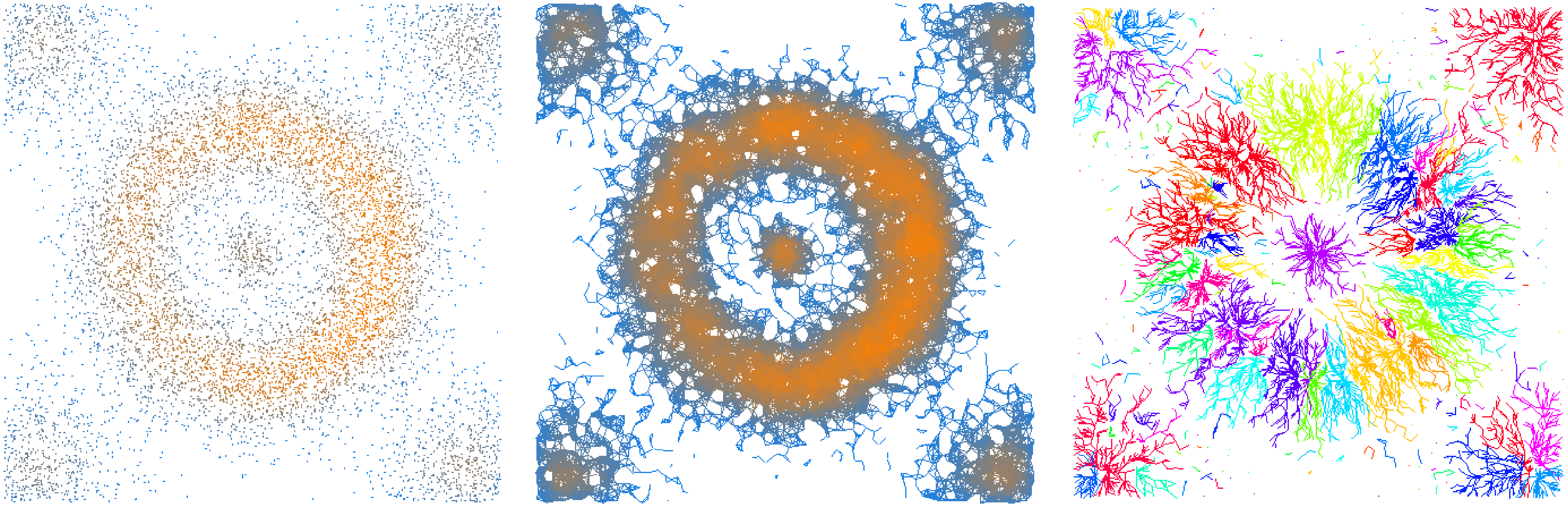
Computing Clusters

How do we compute clusters from a barcode?

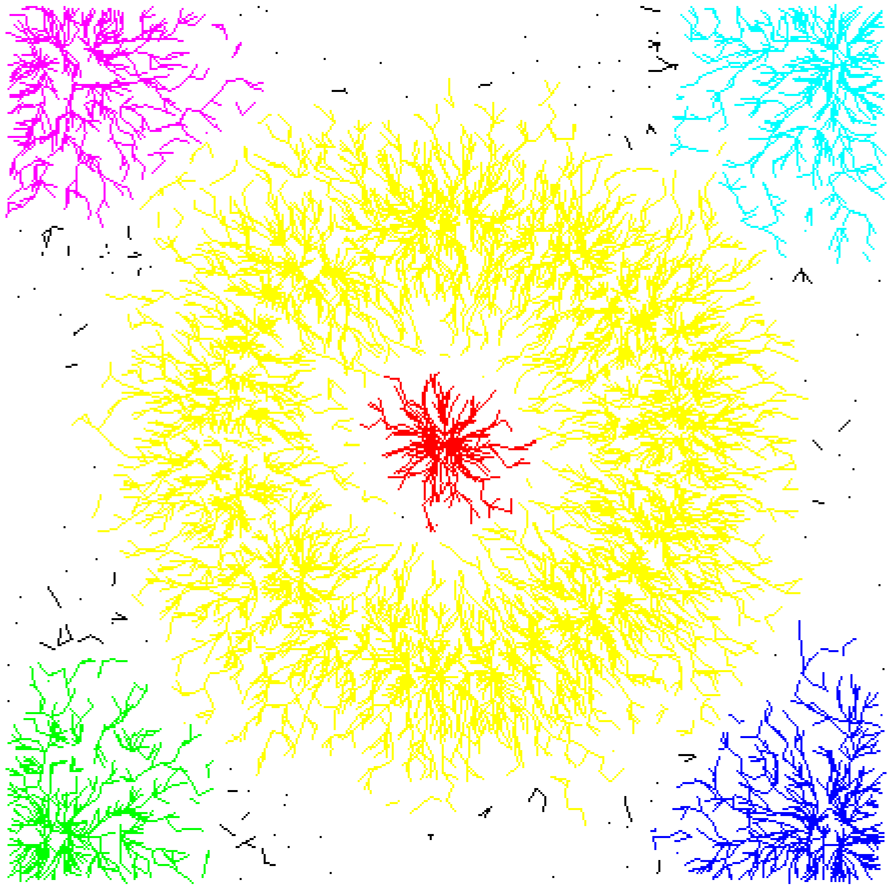
Input: Samples with estimated density $\hat{f}$

Two steps:

1. Mode-seeking step [Koontz et. al. '76]



Computing Clusters

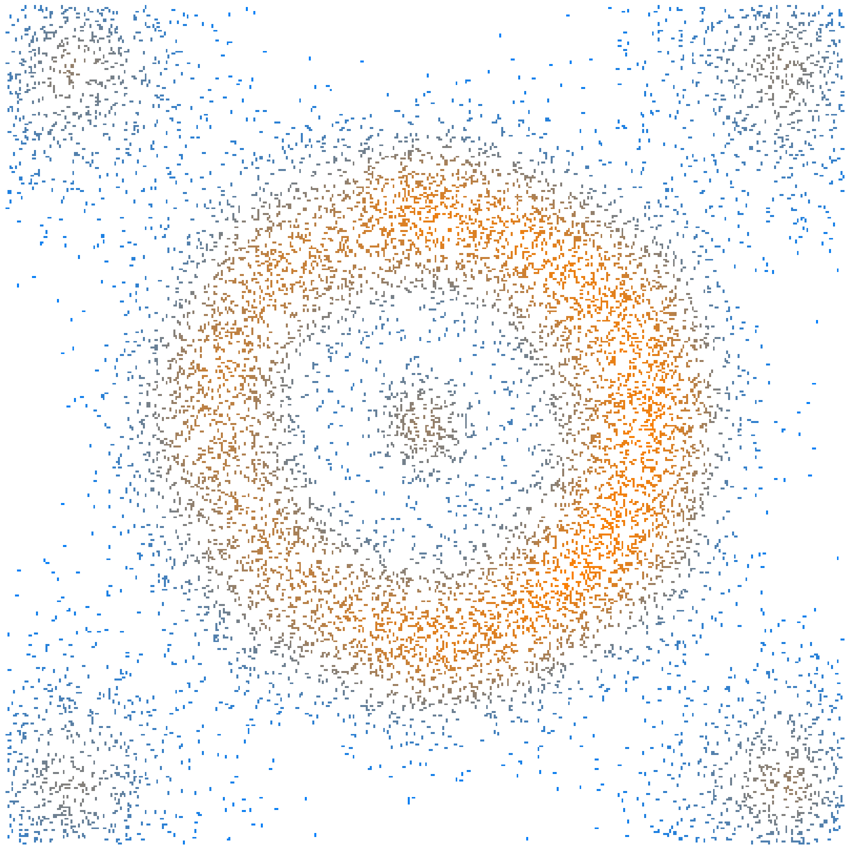


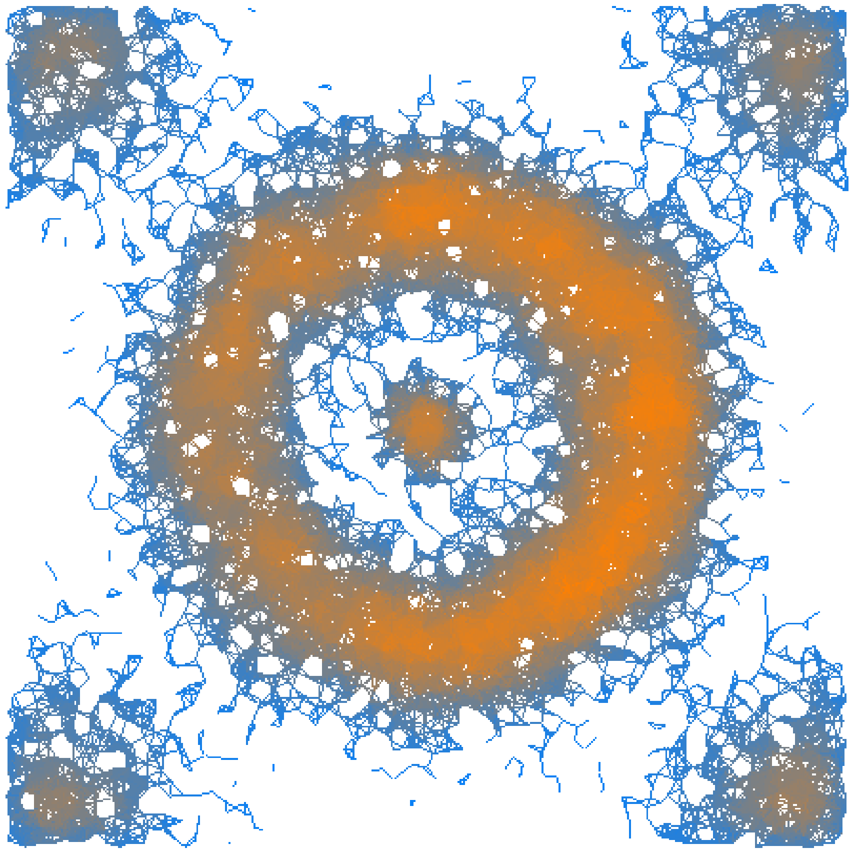
How do we compute clusters from a barcode?

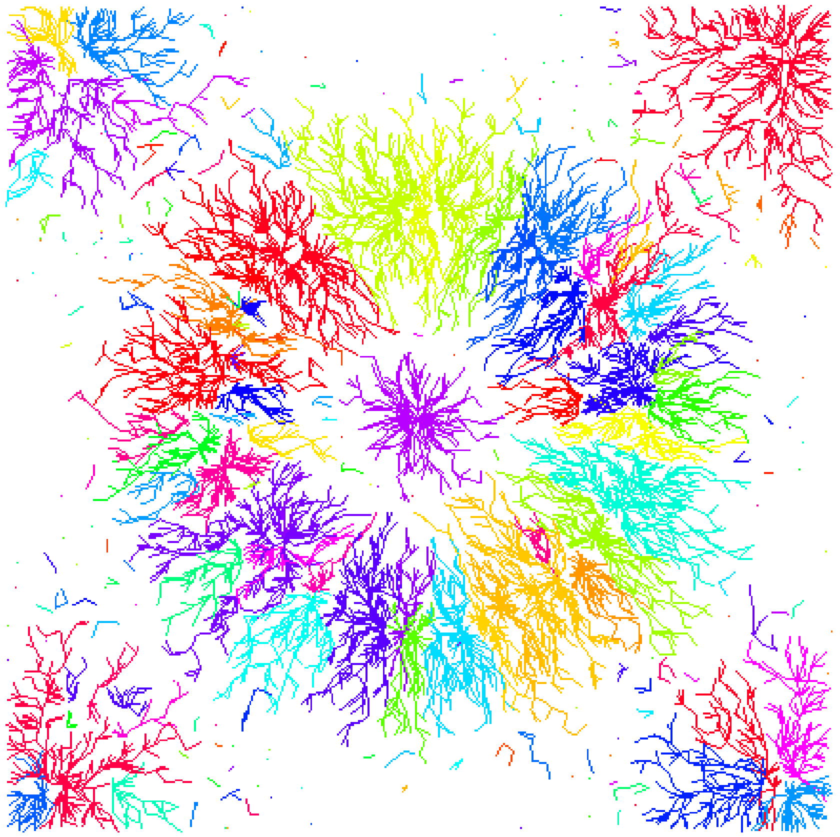
Input: Samples with estimated density $\hat{f}$

Two steps:

1. Mode-seeking step [Koontz et. al. '76]
2. Merge clusters according to persistence







Algorithm

- Input: $f(x), \mathcal{R}_\delta, \alpha$

Algorithm

- Input: $f(x), \mathcal{R}_\delta, \alpha$

1. Sort x according to f
2. For $x \in L$

   2a. For neighbors of x in $\mathcal{R}_\delta$

   If no higher neighbors $\Rightarrow$ new cluster

   else assign x to ∇f

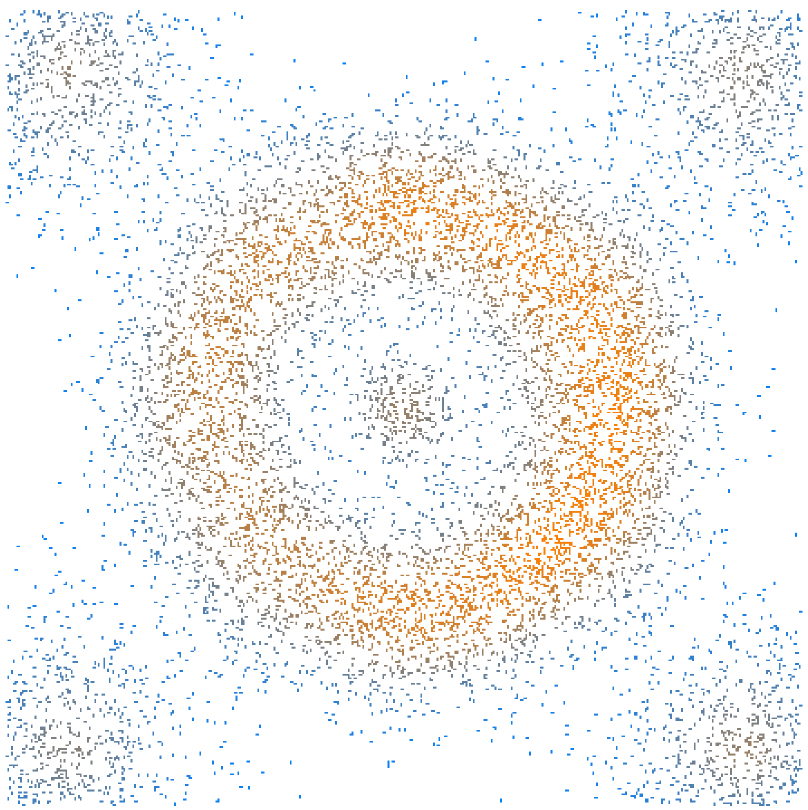
   2b. For adjacent clusters y to x

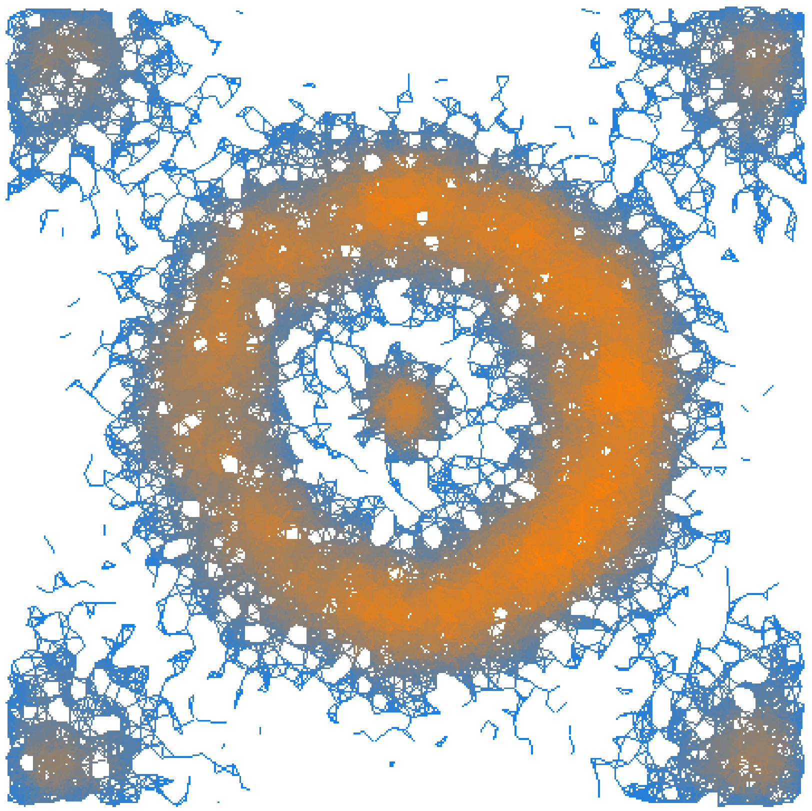
   if $|f(y) - f(x)| \leq \alpha$

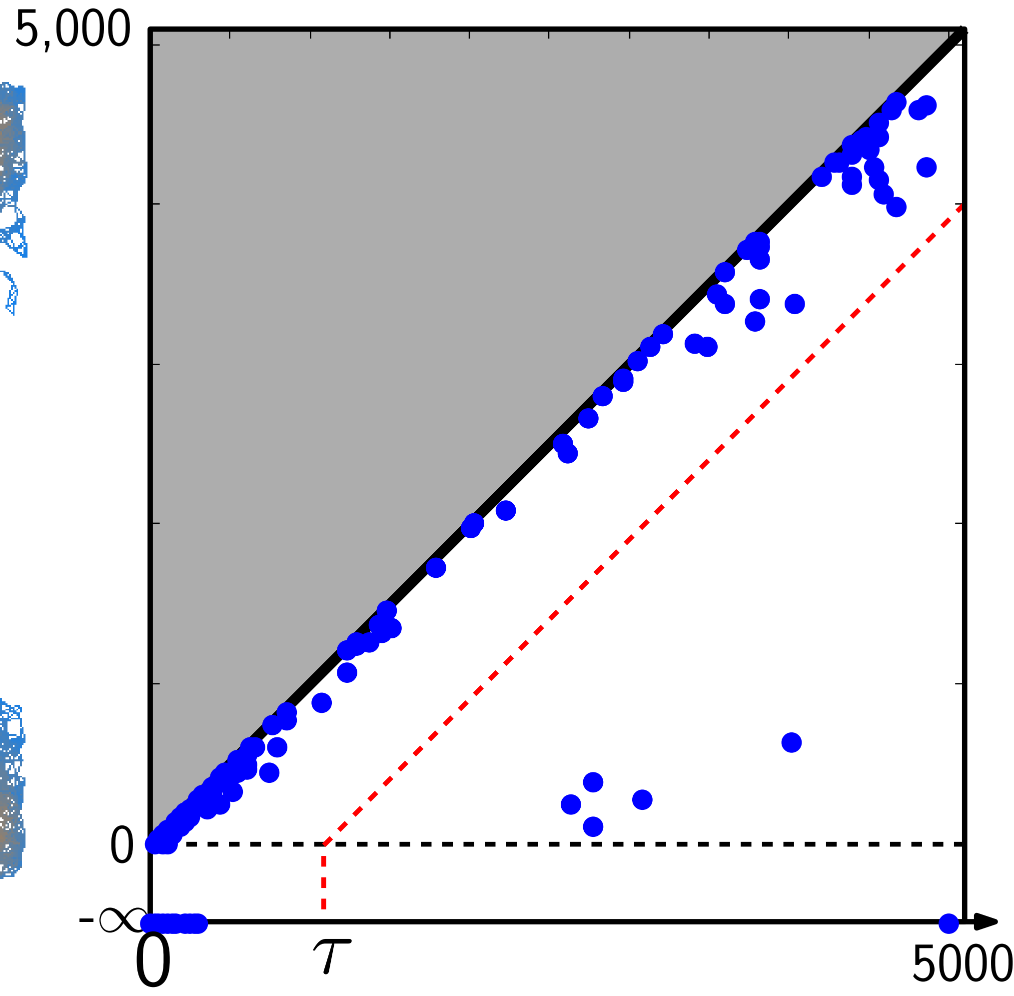
   merge into oldest adjacent cluster

Putting it together

- Estimate density
- Run algorithm with $\alpha = \infty$
  - Standard persistence algorithm
- Use persistence diagram to choose threshold
- Re-run algorithm

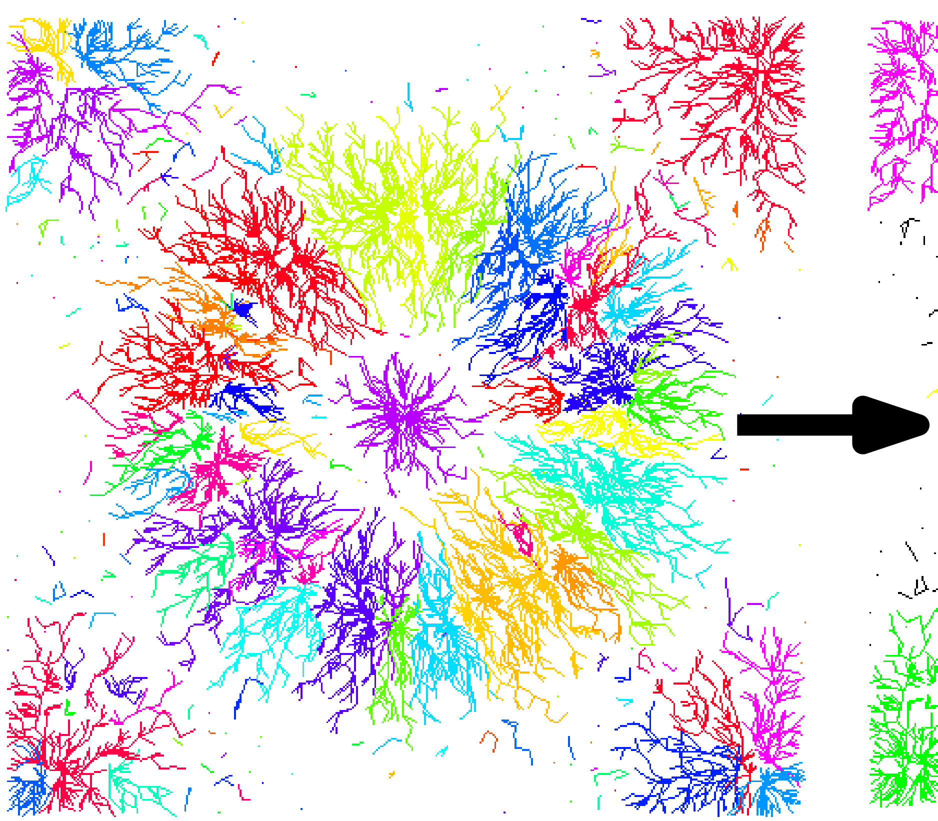


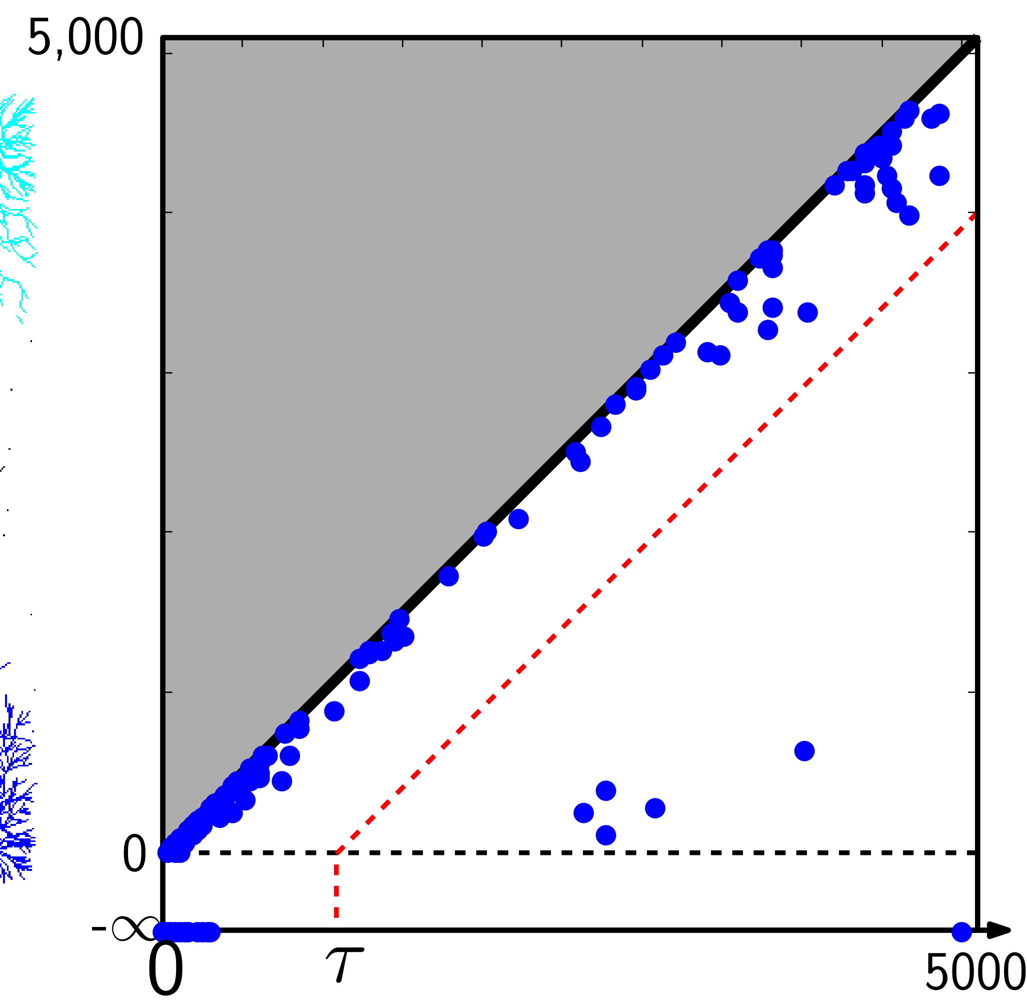




Putting it together

- Estimate density
- Run algorithm with $\alpha = \infty$
  - Standard persistence algorithm
- Use persistence diagram to choose threshold
- Re-run algorithm





Theoretical Guarantees

- Applying the result from scalar field work

Theoretical Guarantees

- Applying the result from scalar field work

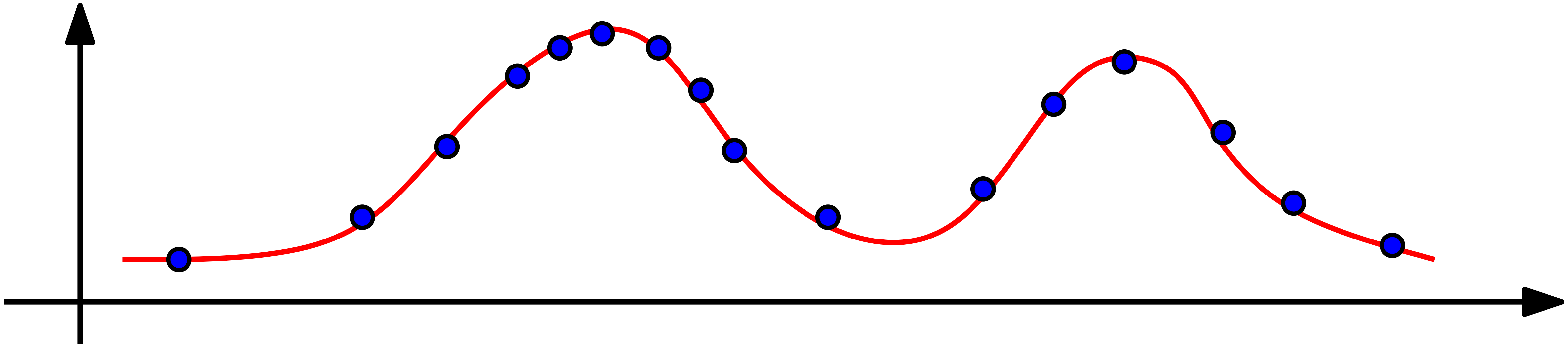
  Approximation depends on $c\delta$

Theoretical Guarantees

- Applying the result from scalar field work

Approximation depends on $c\delta$

Whole space is **not** uniformly sampled

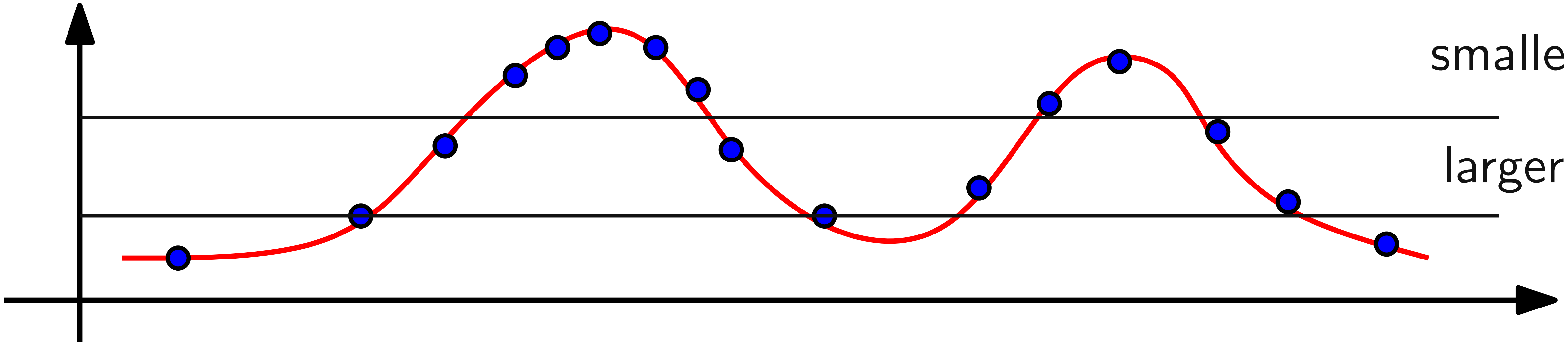


Theoretical Guarantees

- Applying the result from scalar field work

Approximation depends on $c\delta$

Whole space is **not** uniformly sampled



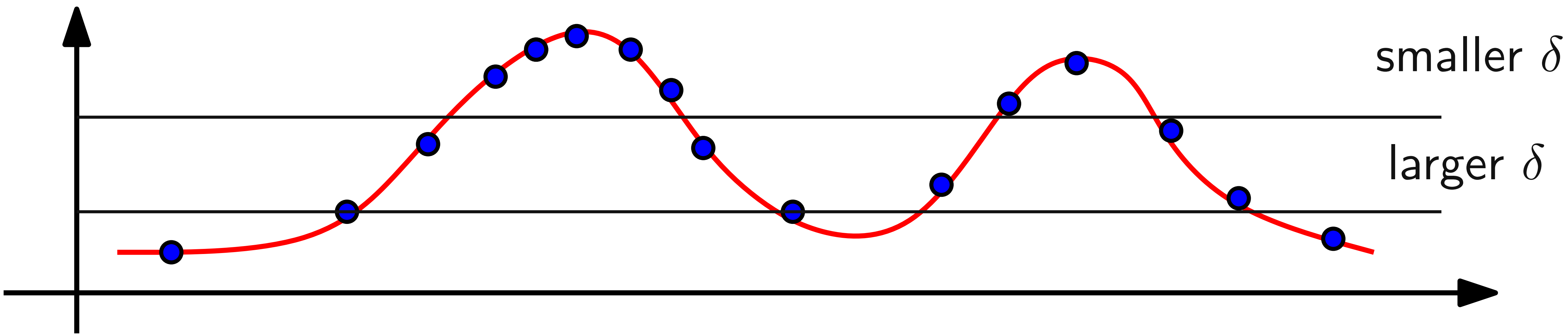
- Approximation result holds in well-sampled regions w.h.p.

Theoretical Guarantees

- Applying the result from scalar field work

Approximation depends on $c\delta$

Whole space is **not** uniformly sampled



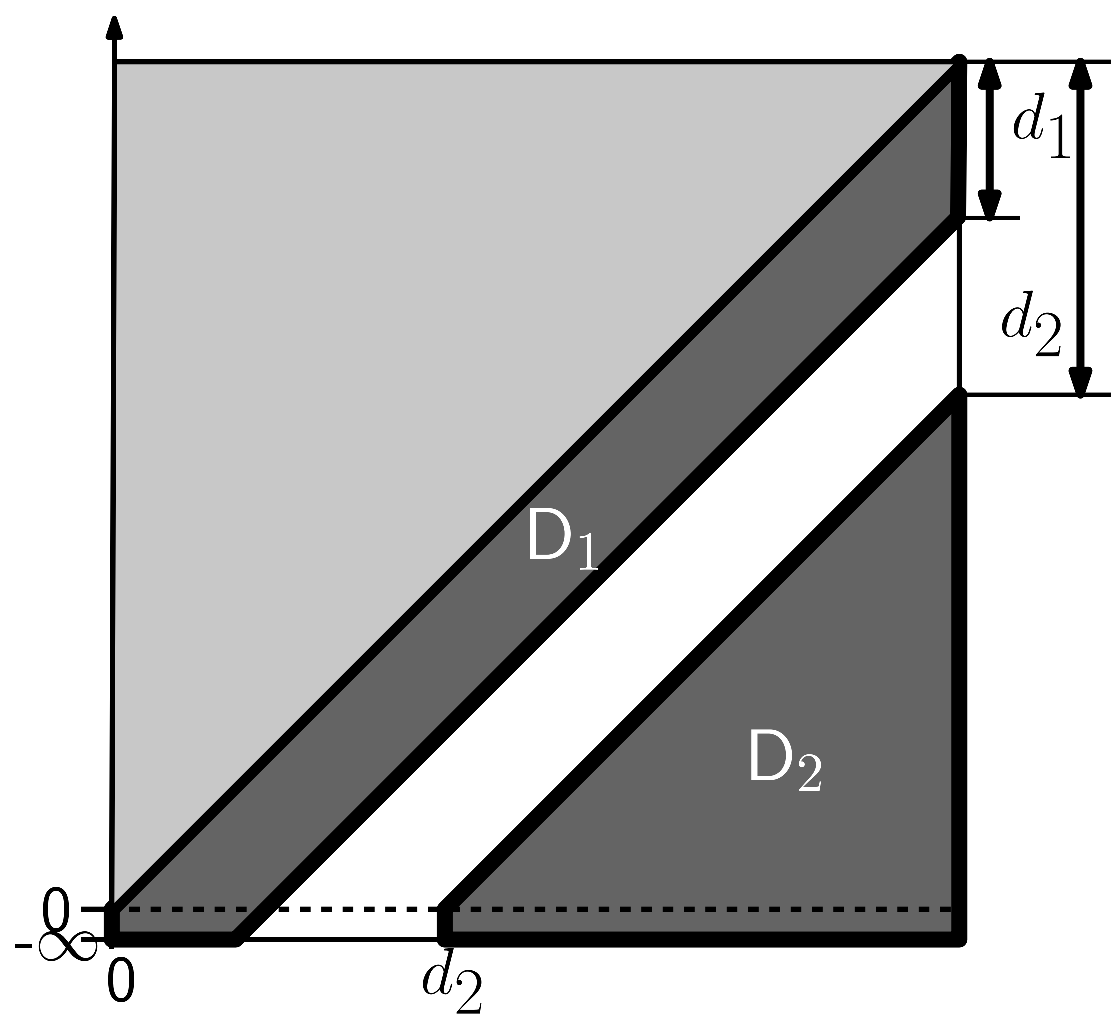
- Approximation result holds in well-sampled regions w.h.p.

- More points $\Rightarrow$ more of the space

Number of Clusters

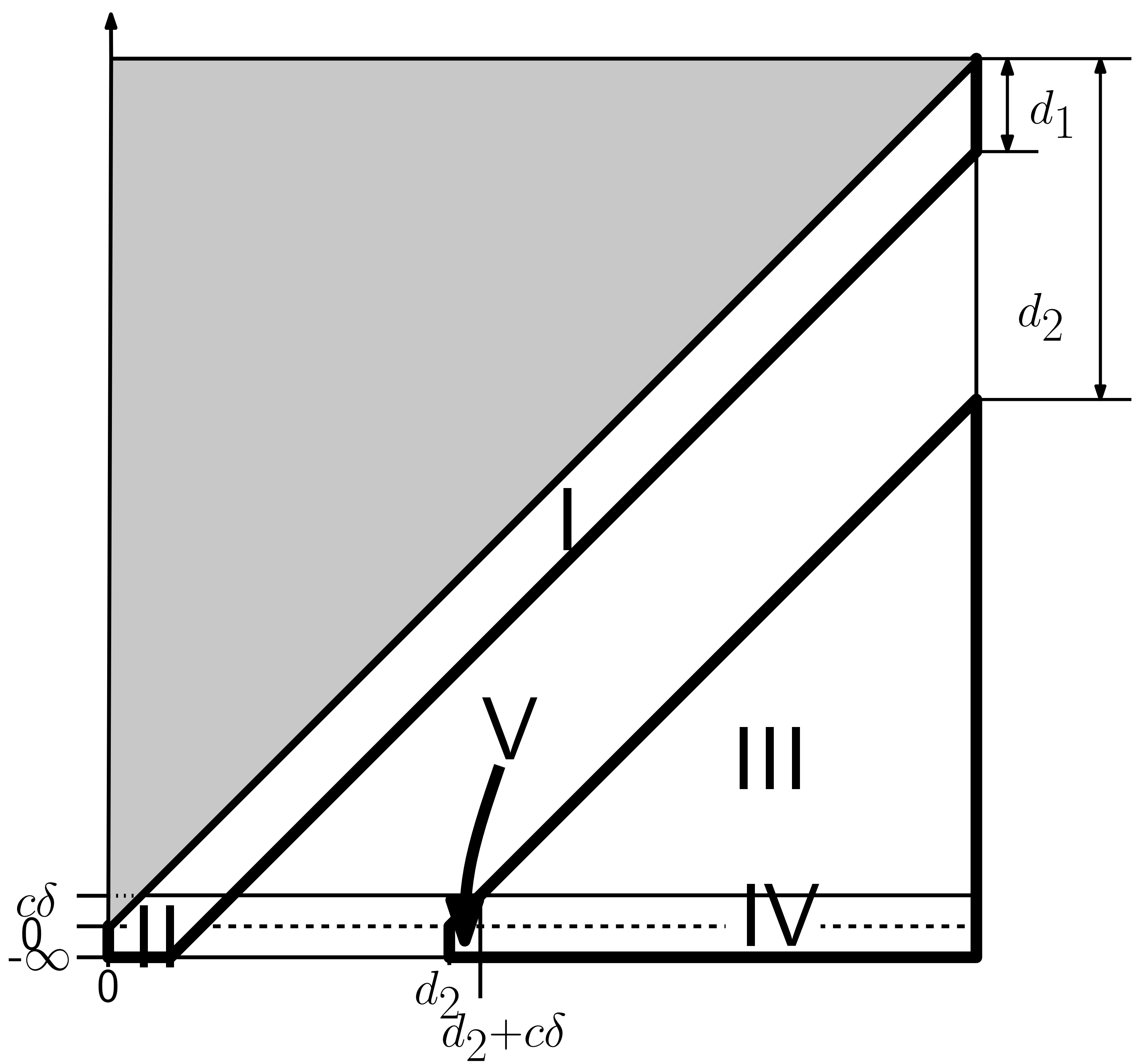
- Define a *signal-to-noise* ratio

Definition: Given two values $d_2 > d_1 \geq 0$, the persistence diagram $\mathrm{D}_0 f$ is called (d_1, d_2)-*separated* if every point of $\mathrm{D}_0 f$ lies either in the region D_1 above the diagonal line $y = x - d_1$ or in the region D_2 below the diagonal $y = x - d_2$ and to the right of the vertical line $x = d_2$.



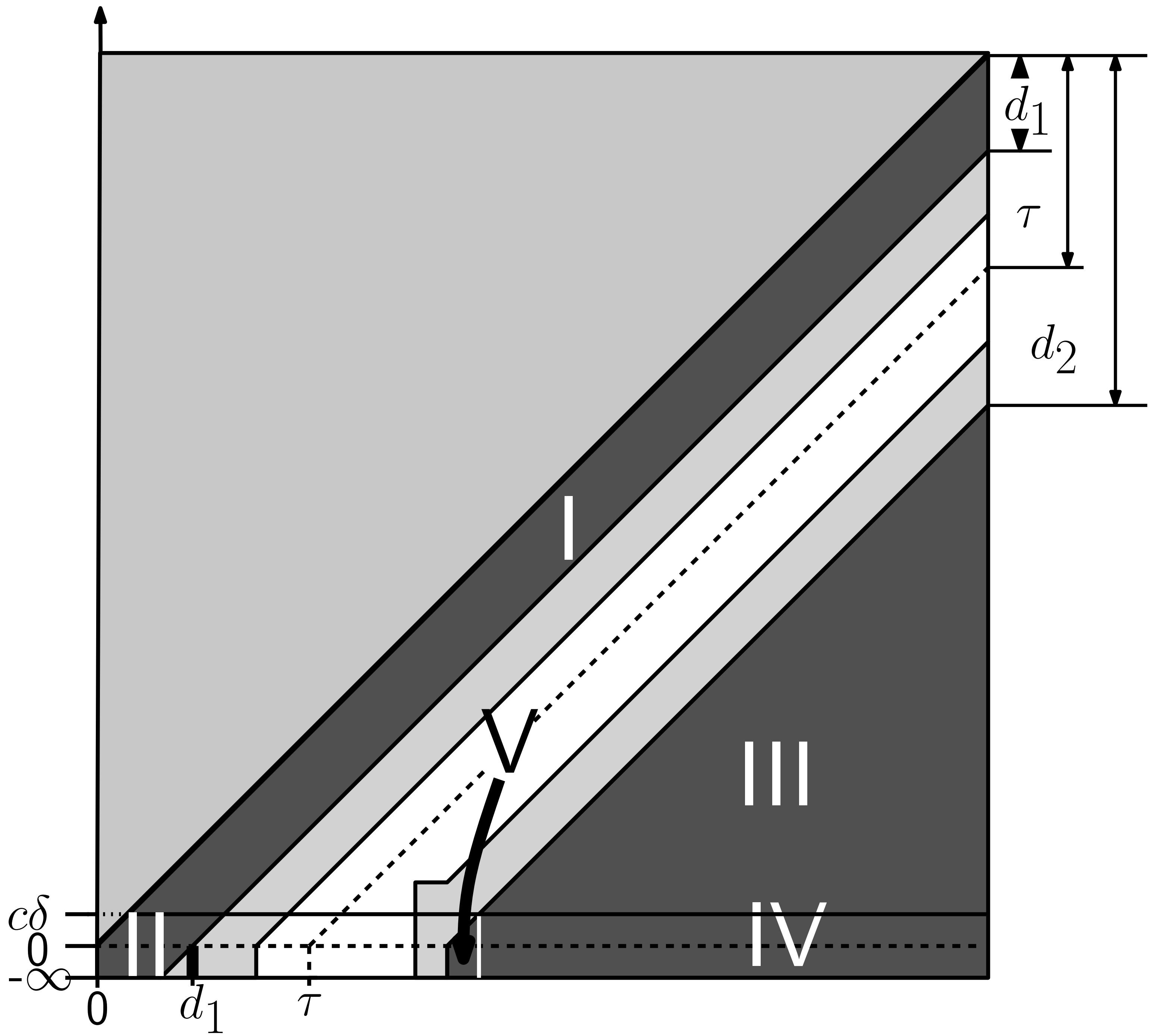
Approximation

- Assume enough points that up to $c\delta$ is well-sampled w.h.p.



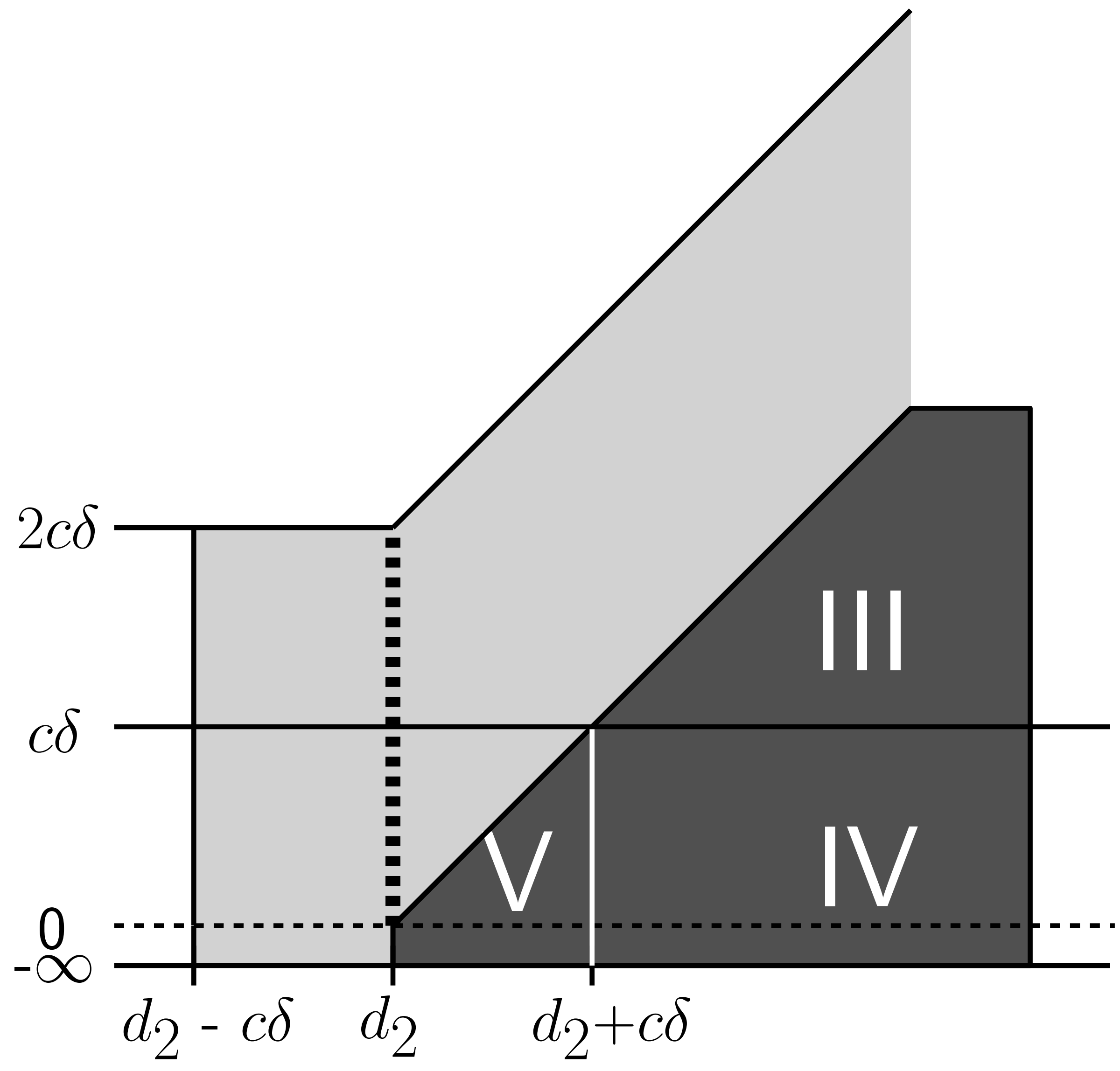
Approximation

- Assume enough points that up to $c\delta$ is well-sampled w.h.p.



Approximation

- Assume enough points that up to $c\delta$ is well-sampled w.h.p.



Feedback and Interpreting Diagrams

- If peaks are prominent enough, we will get the “right” number of clusters
- Practically,
  - Gives a sense of stability of the number of clusters
  - Choice of threshold transparent w.r.t. number of clusters

- Rips parameter $\delta =$ spatial scale
  - Trade-off

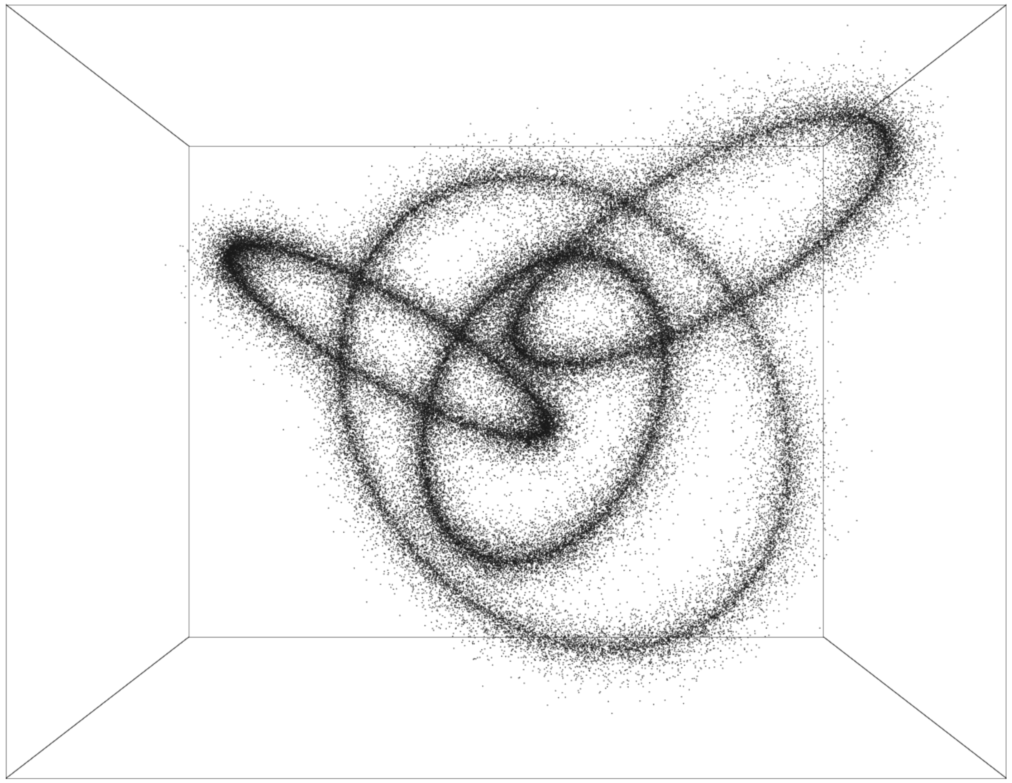
$$\text{Small } \delta = \text{good approximation}$$
$$\text{Large } \delta = \text{holds over a larger part of the space}$$

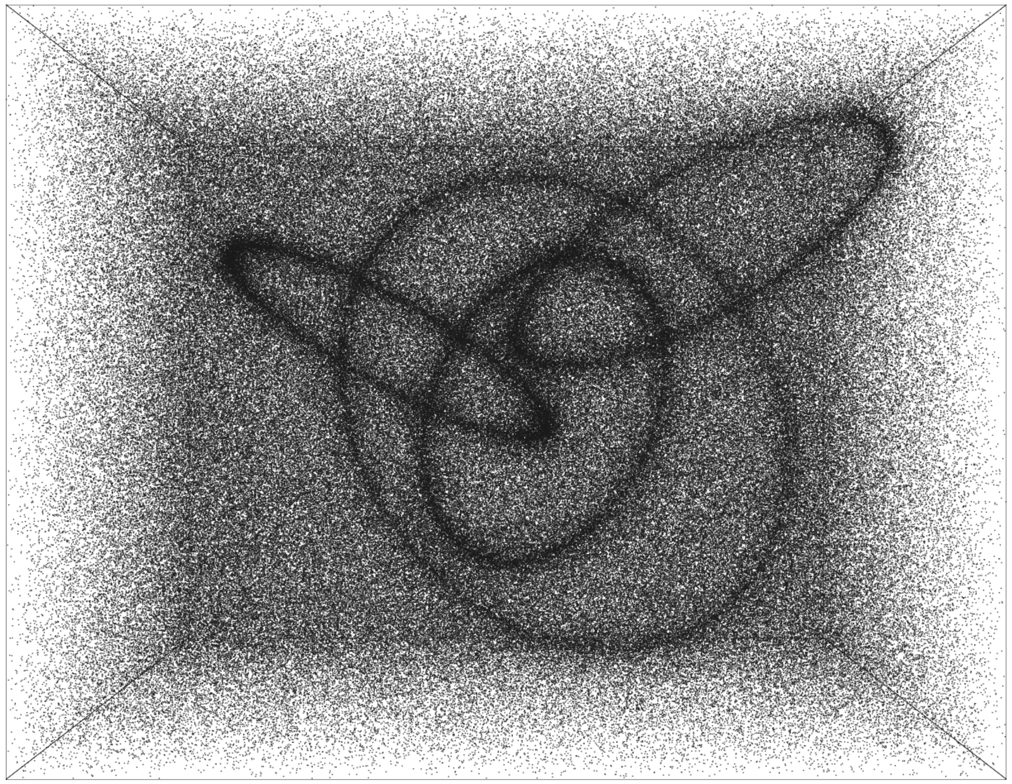
Experiments

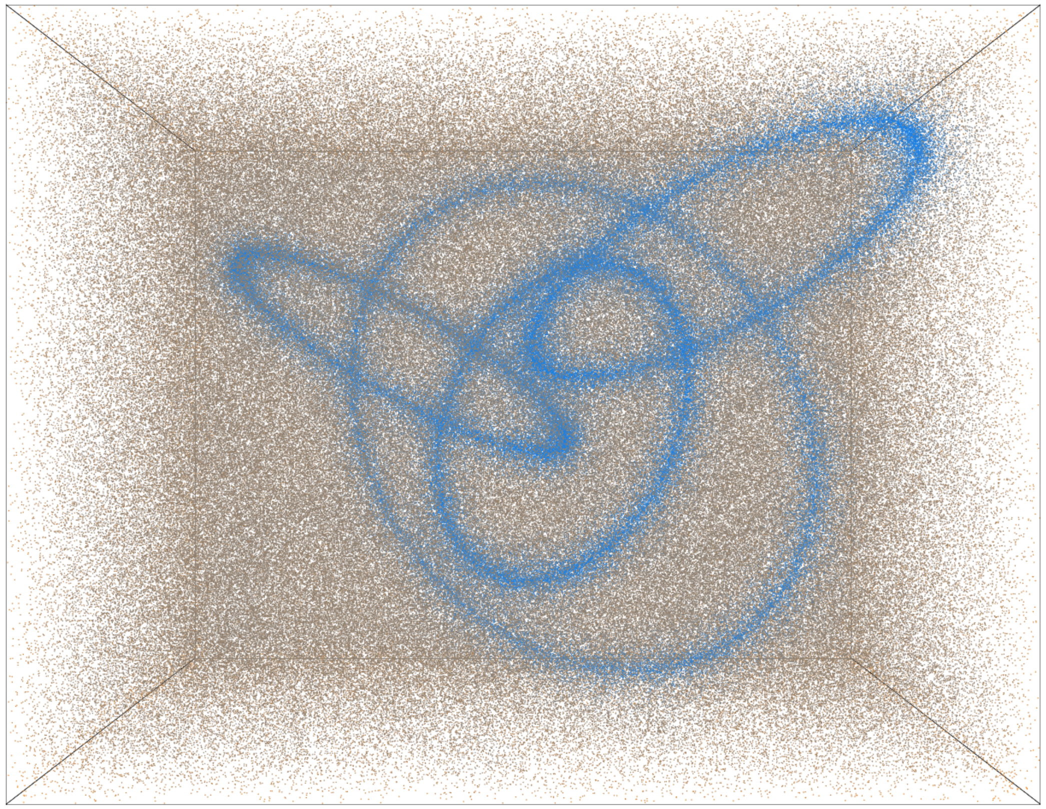
- Synthetic dataset
- Image segmentation
- Alanine-dipeptide conformations

4 Rings

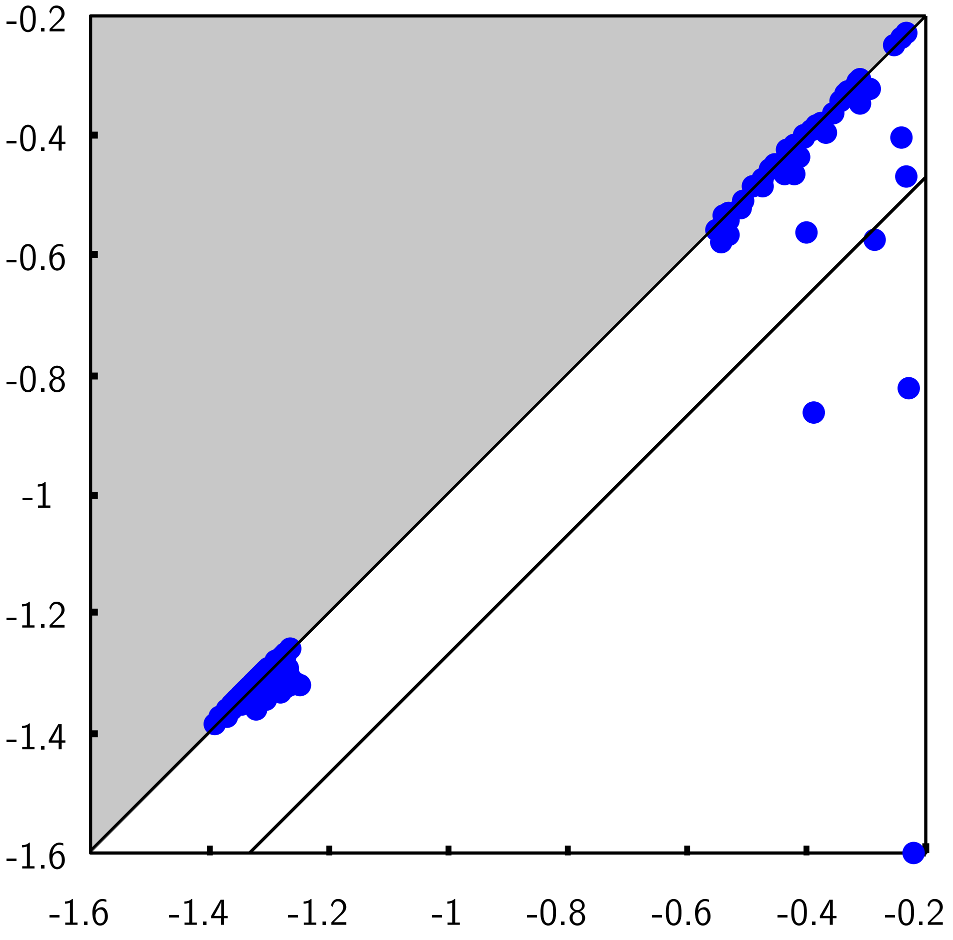
- Interlocking rings in $\mathbb{R}^3$
- 600k (100k + 500k) points total



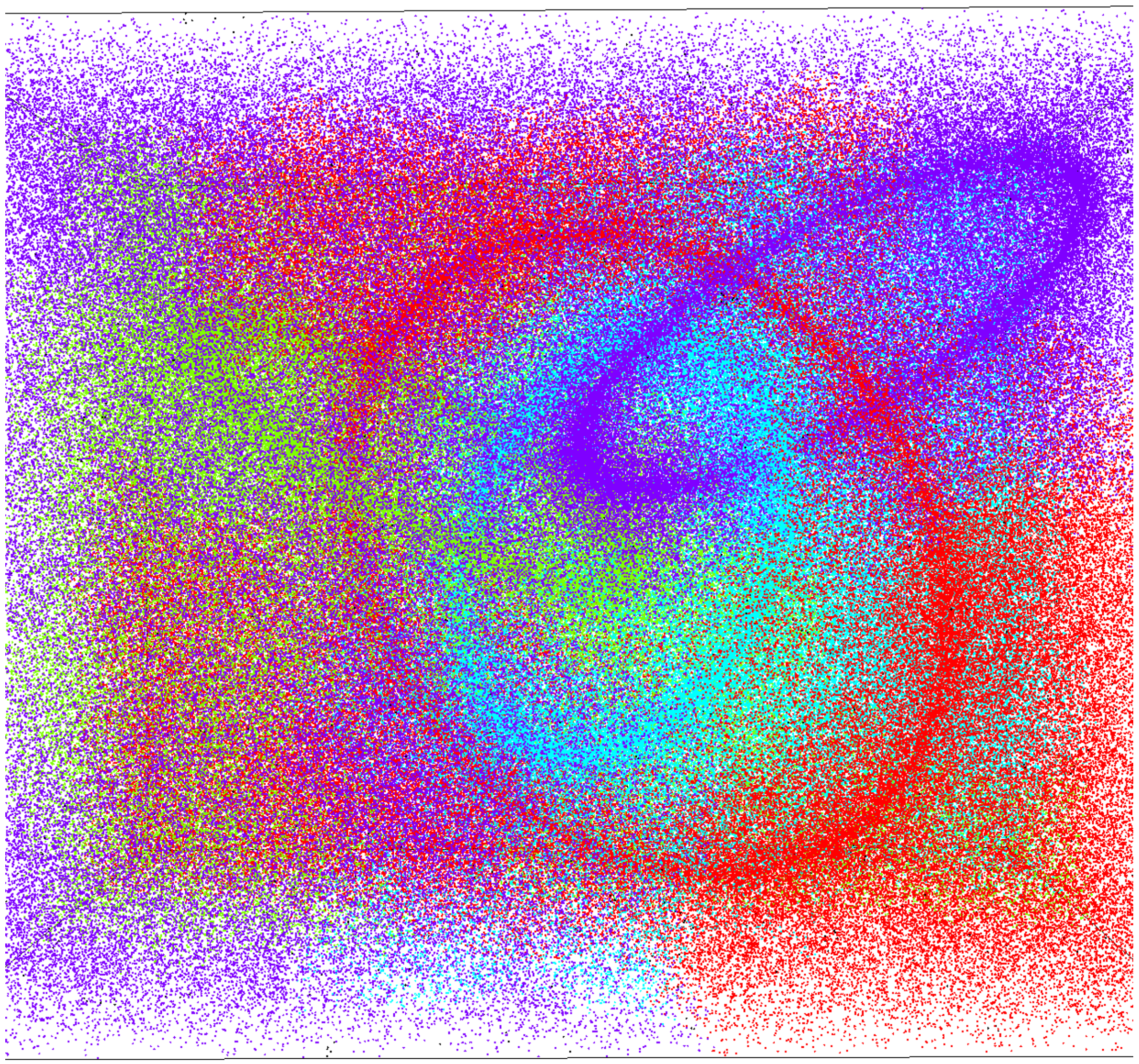


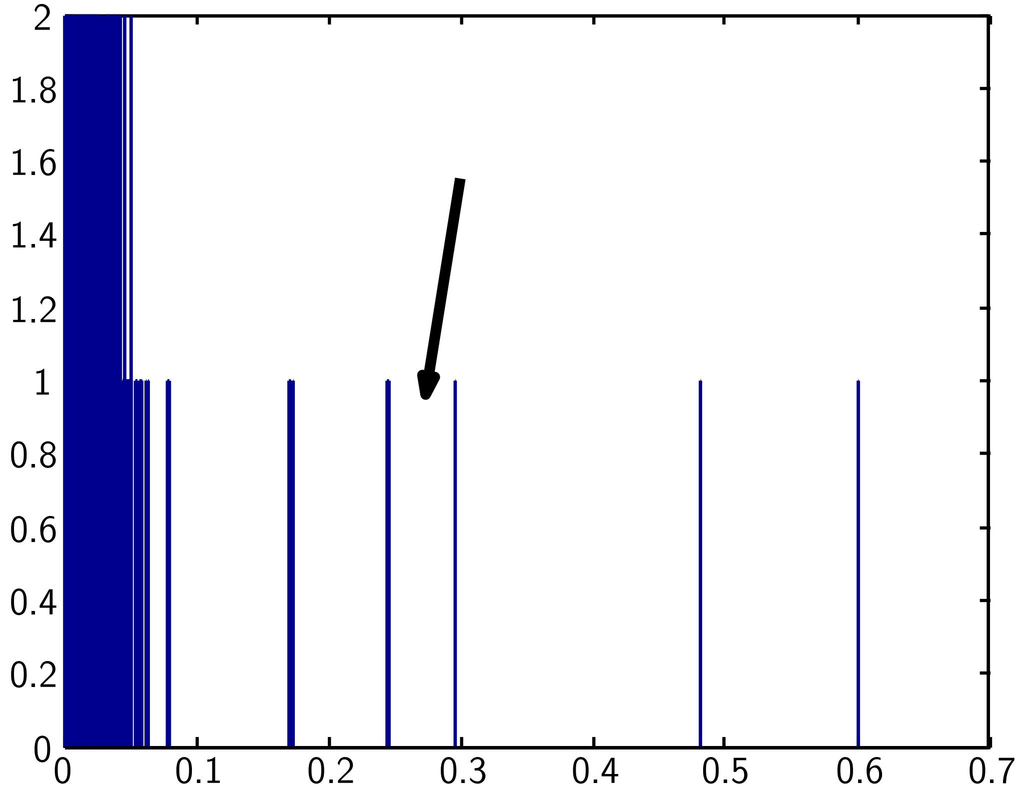


4 Rings



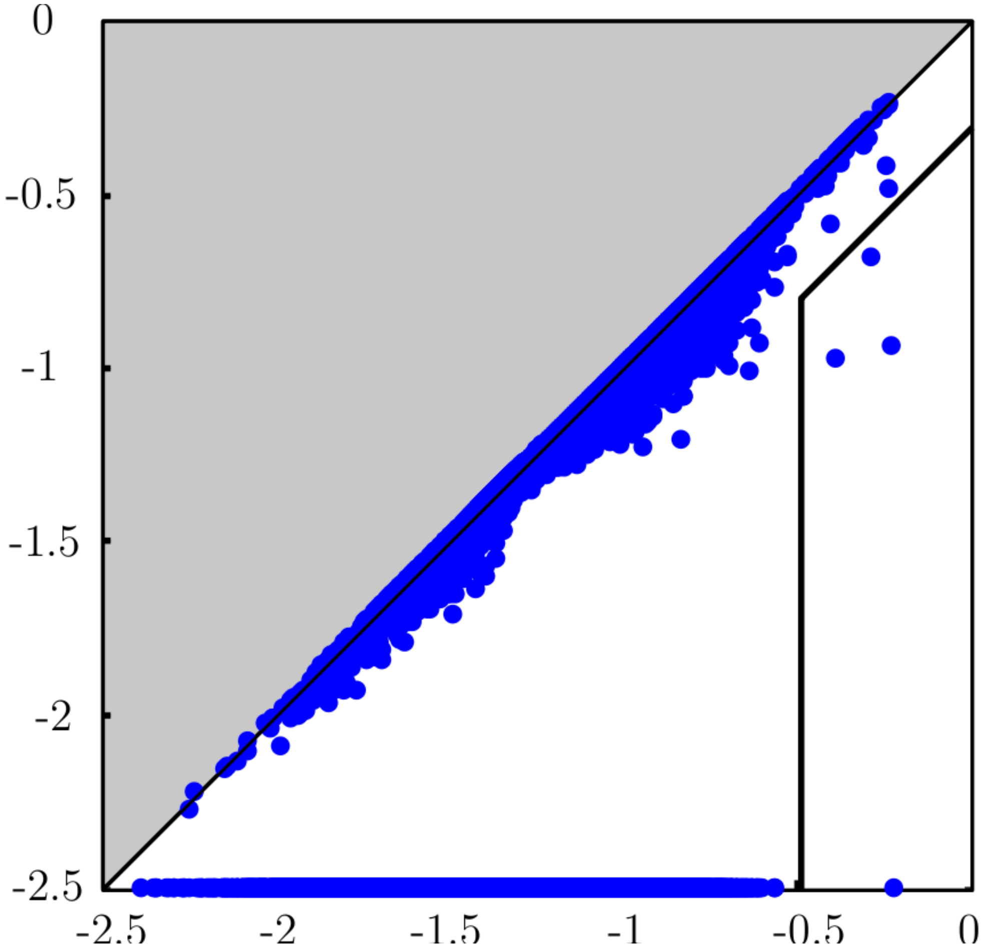
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.6
-1.4
-1.2
-1
-0.8
-0.6
-0.4
-0.2



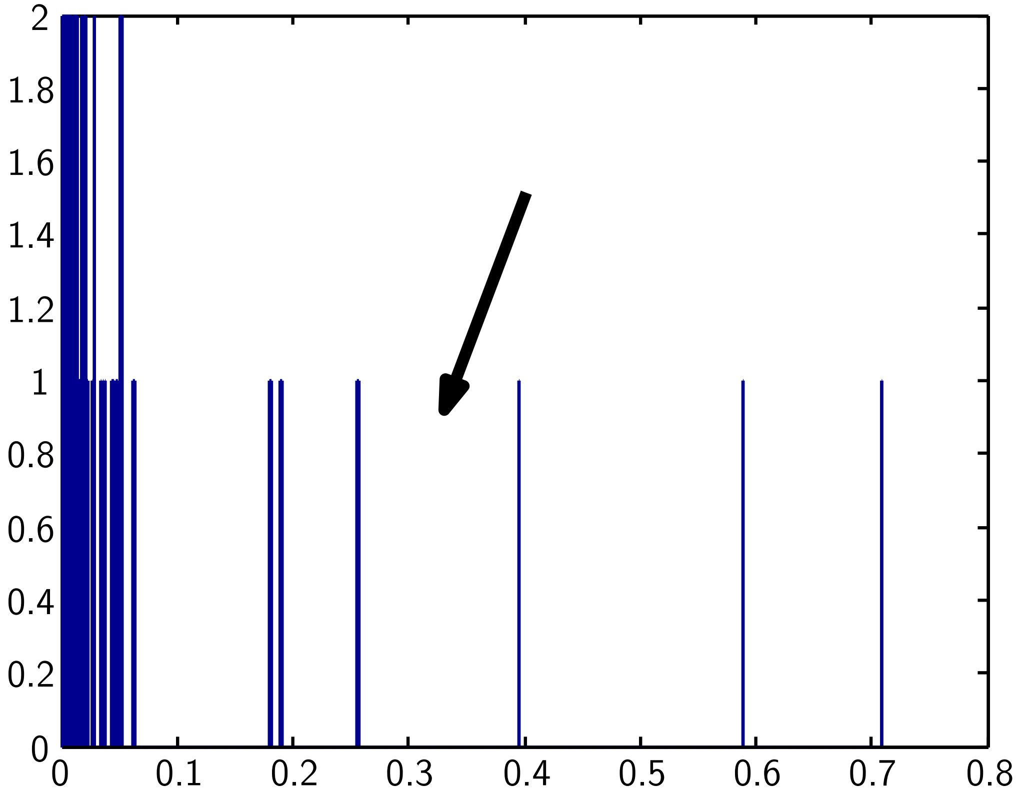


2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

4 Rings



0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0



2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

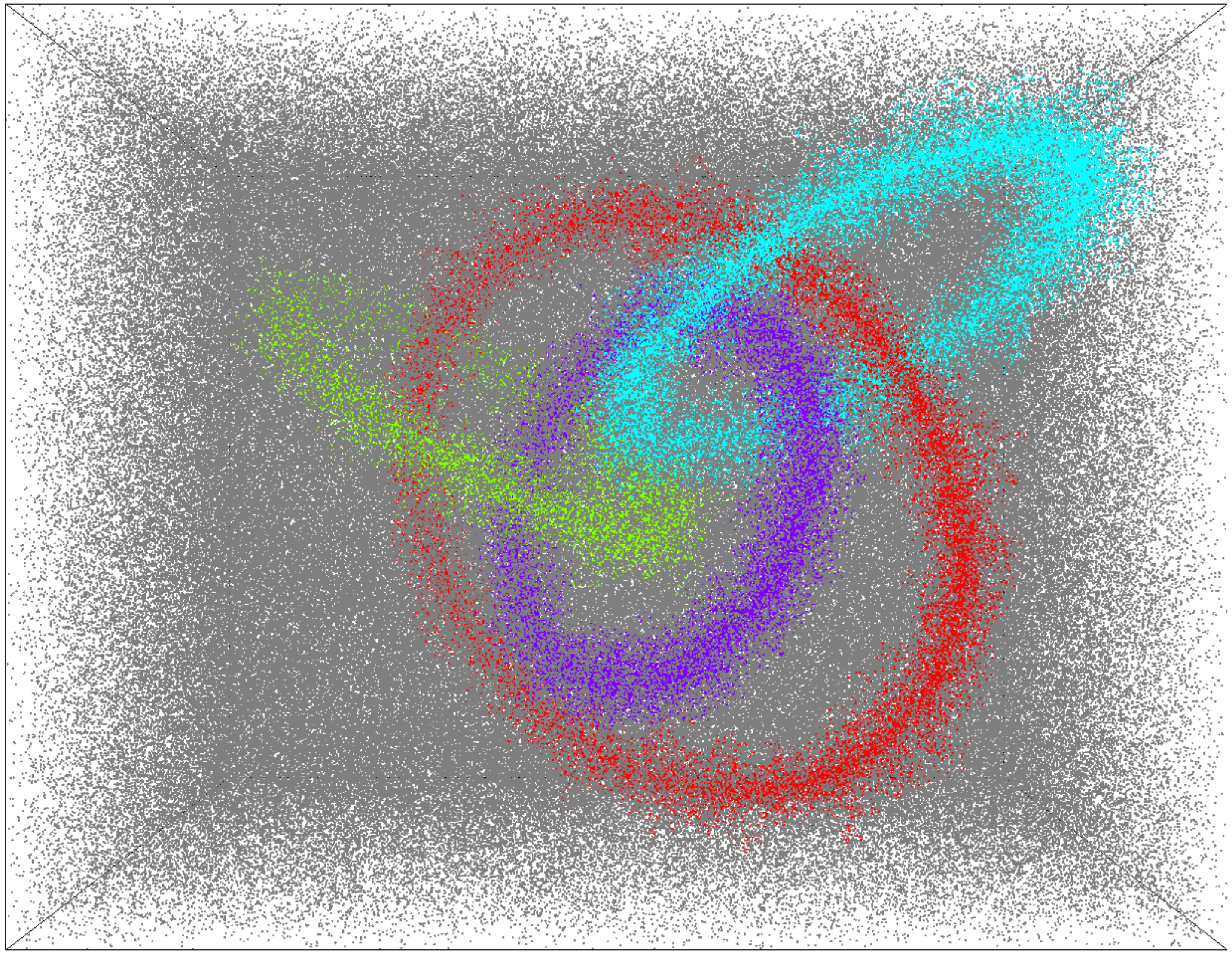
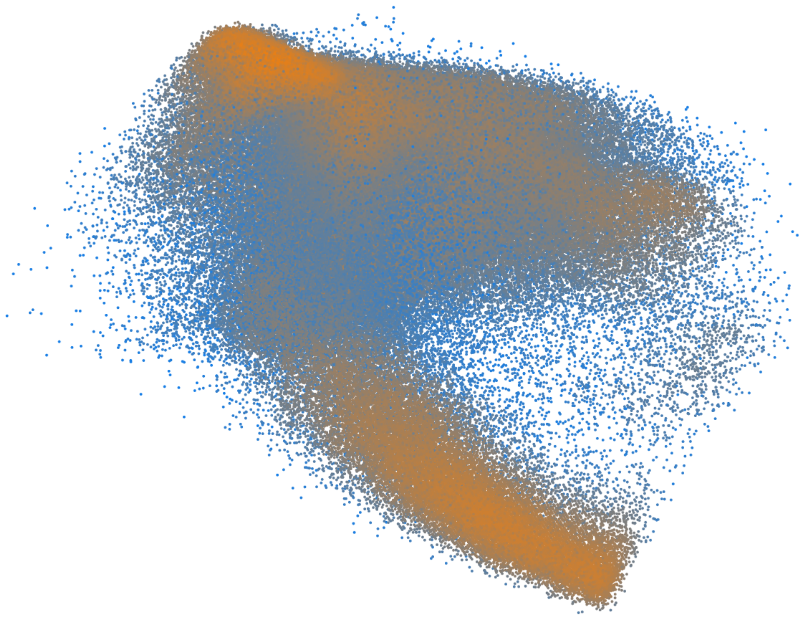
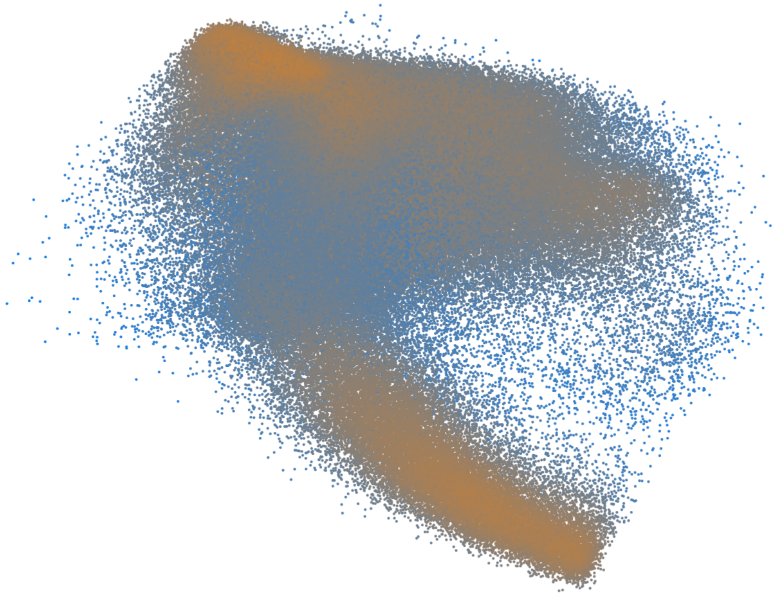


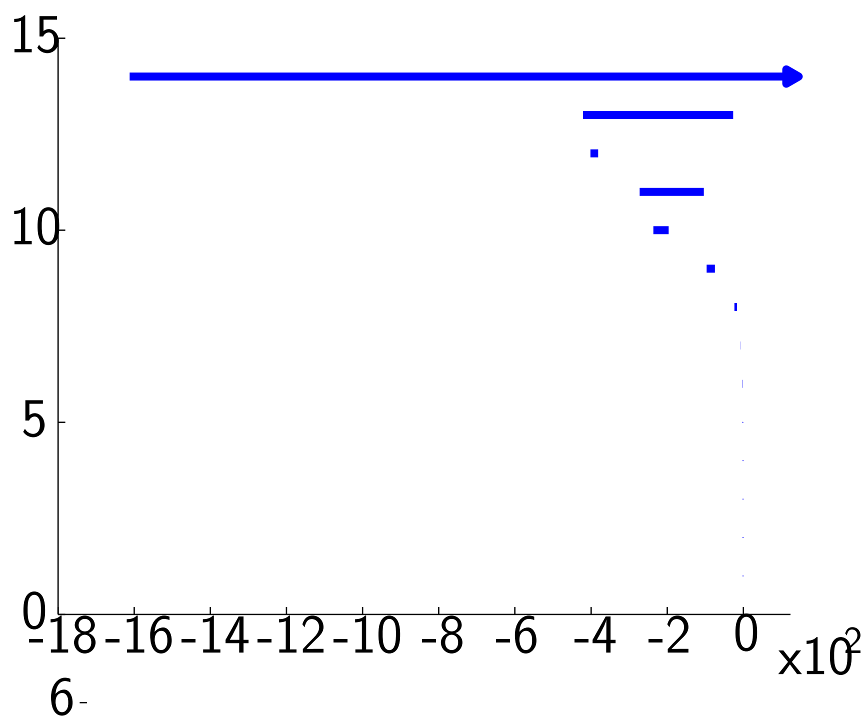
Image Segmentation

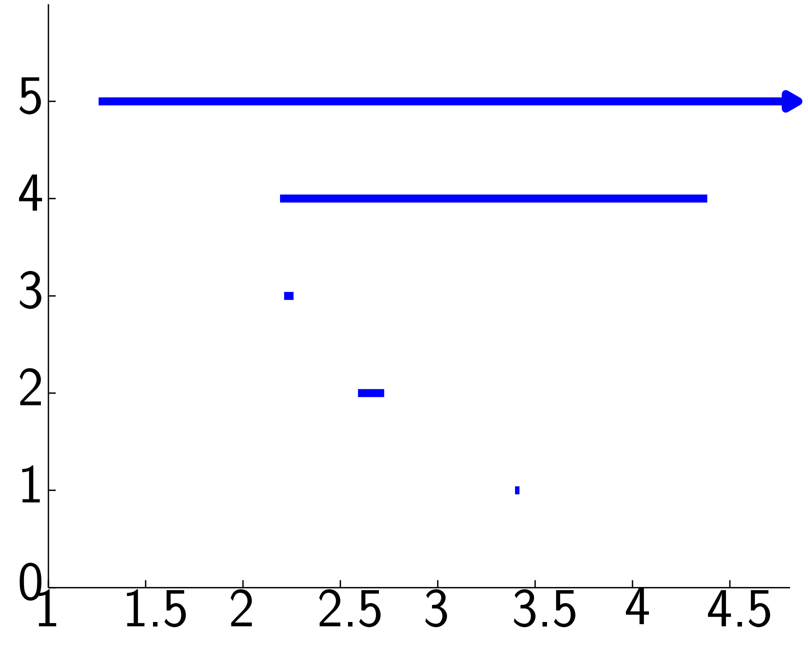
- Each pixel is assigned color coordinates in LUV space

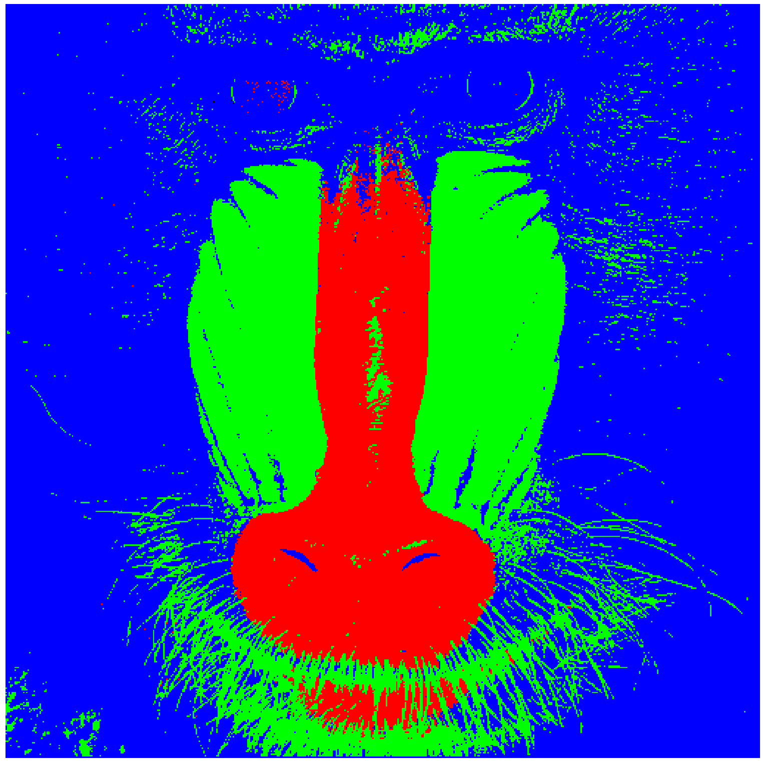




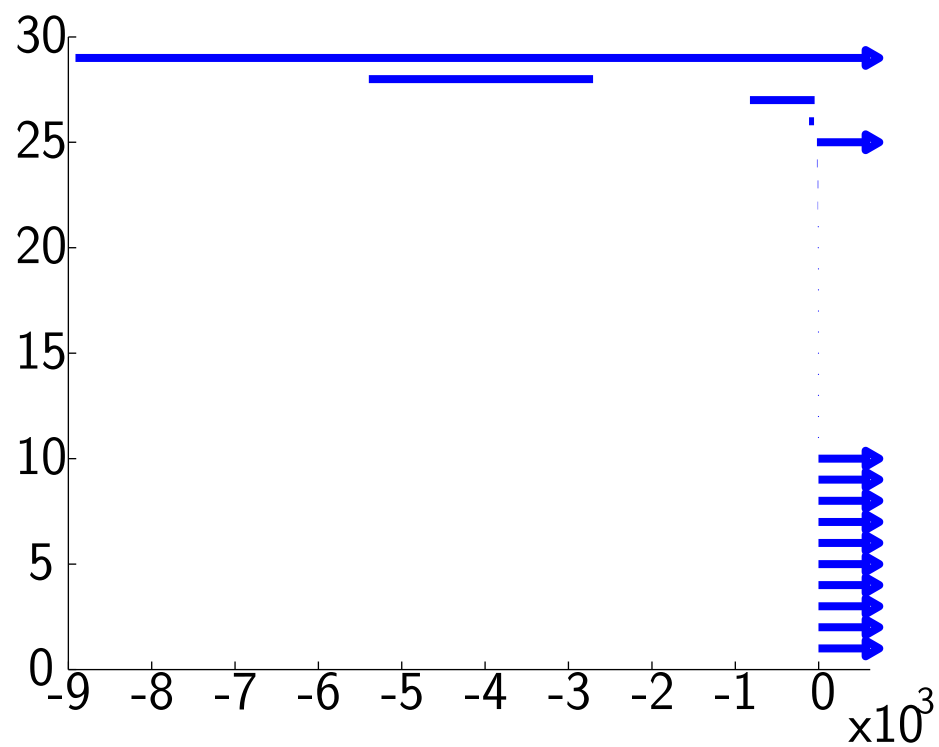
Mandrill

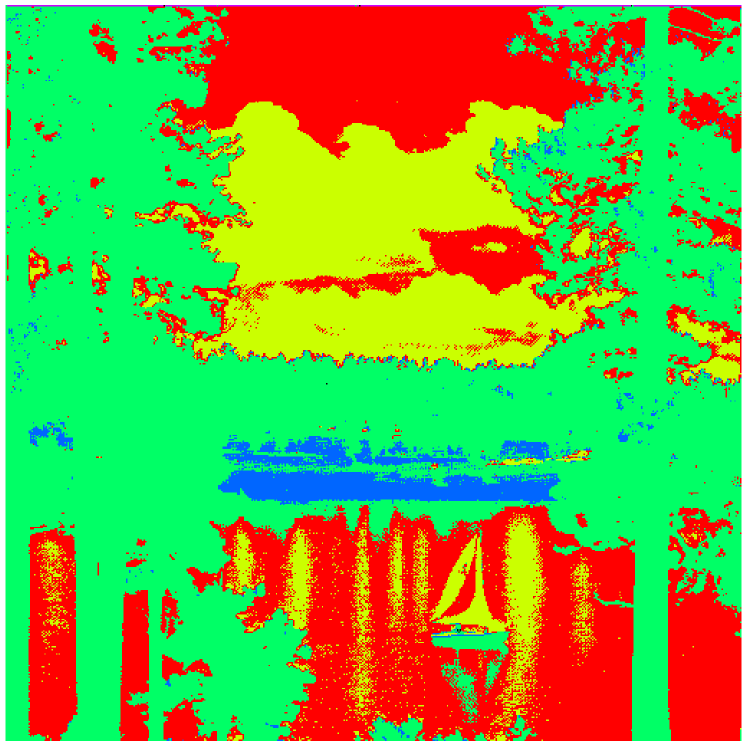


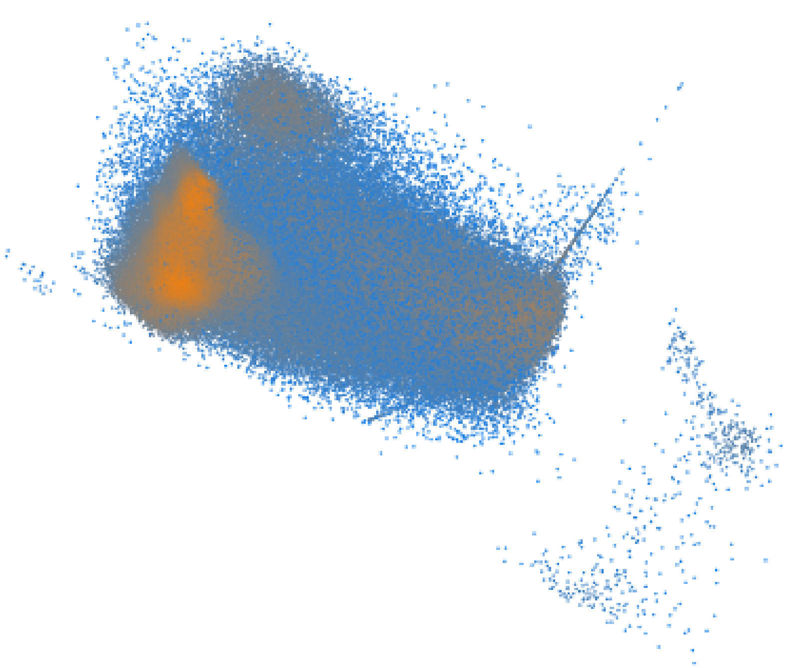


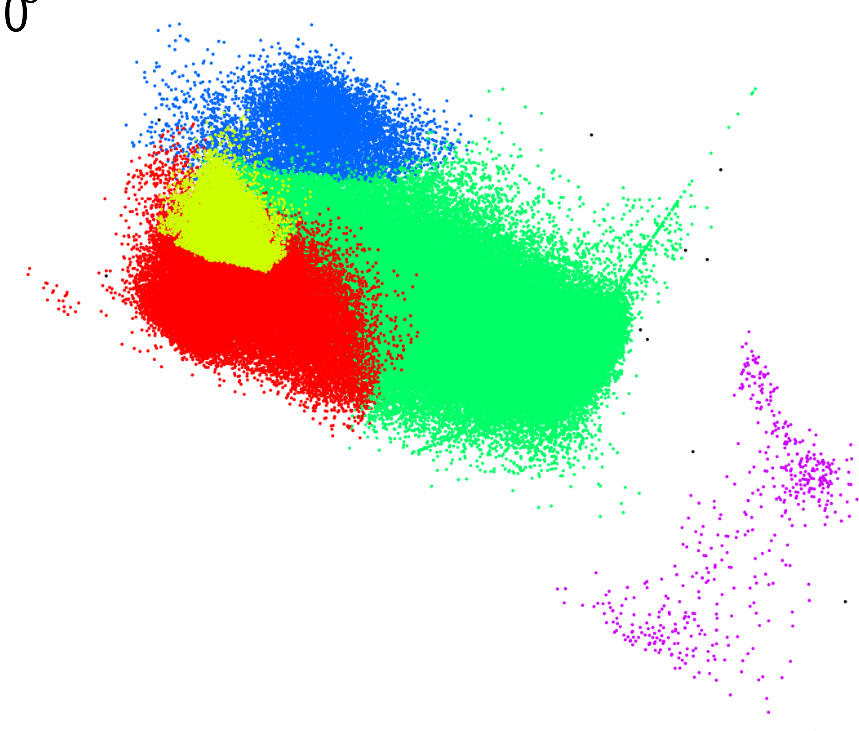


Landscape

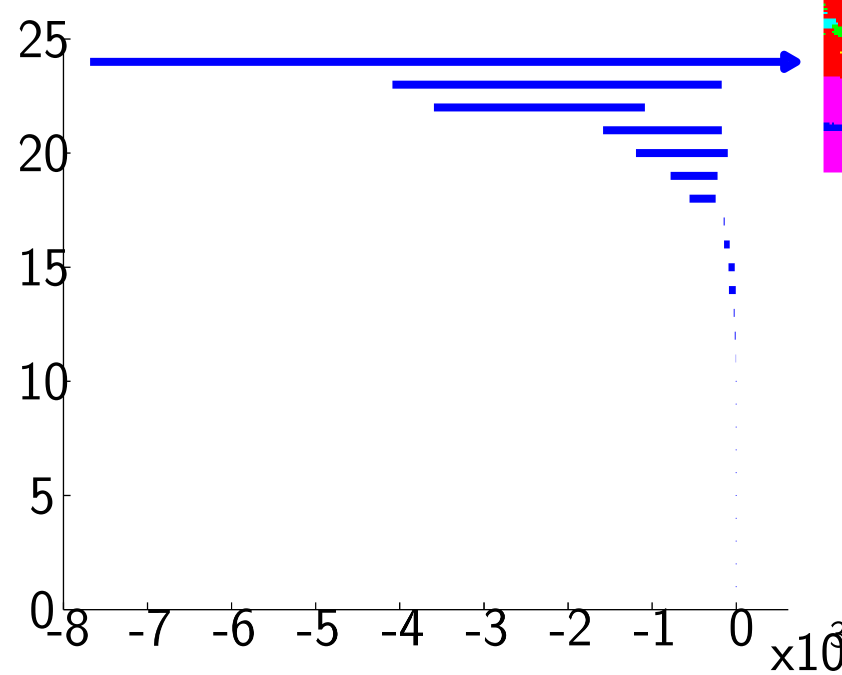


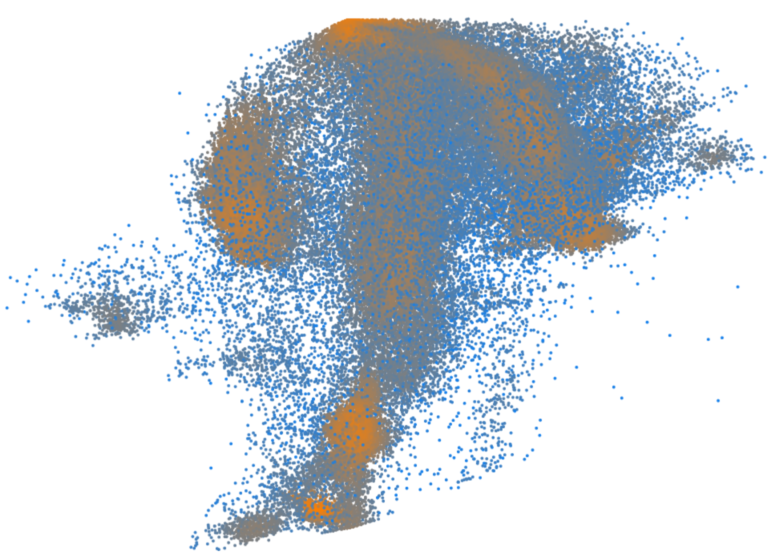


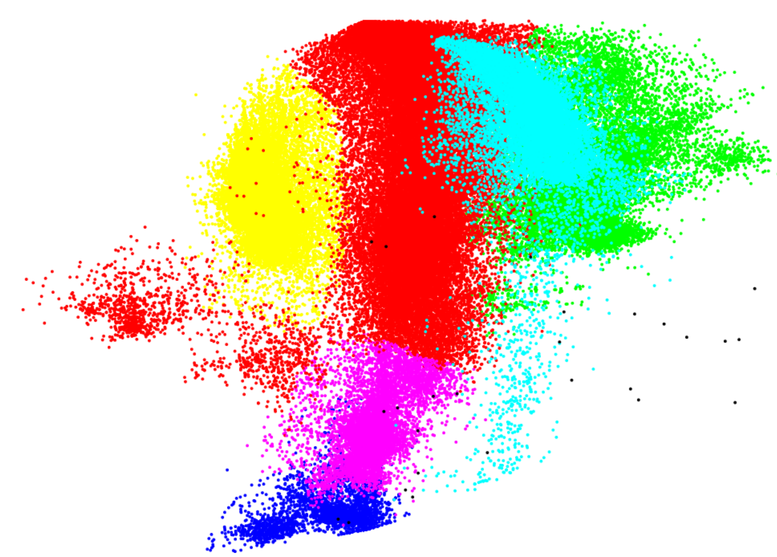




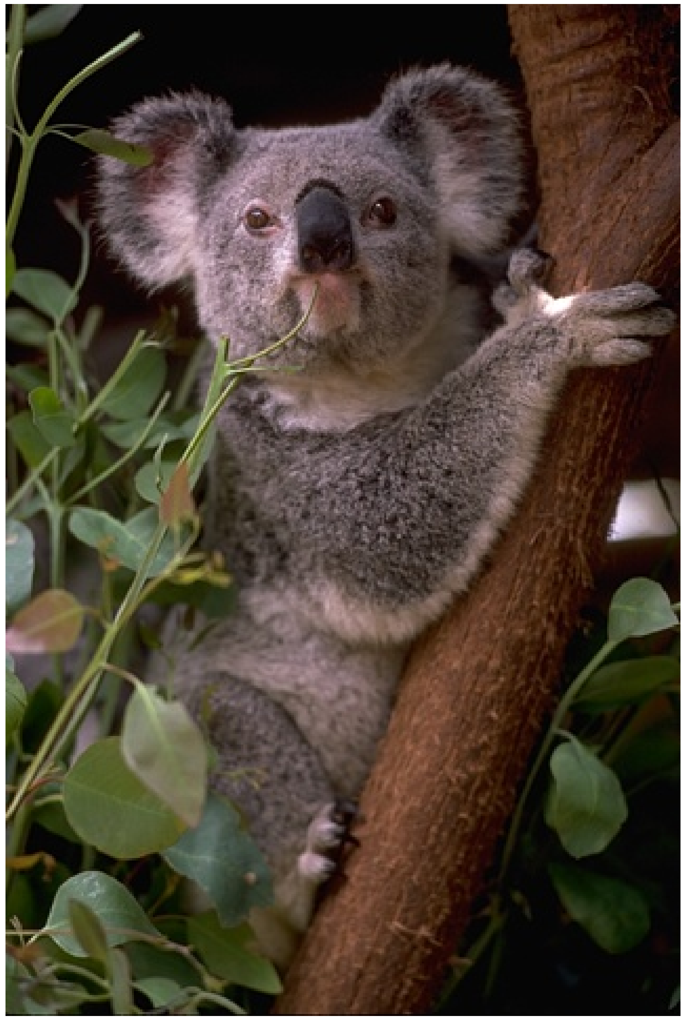
Street

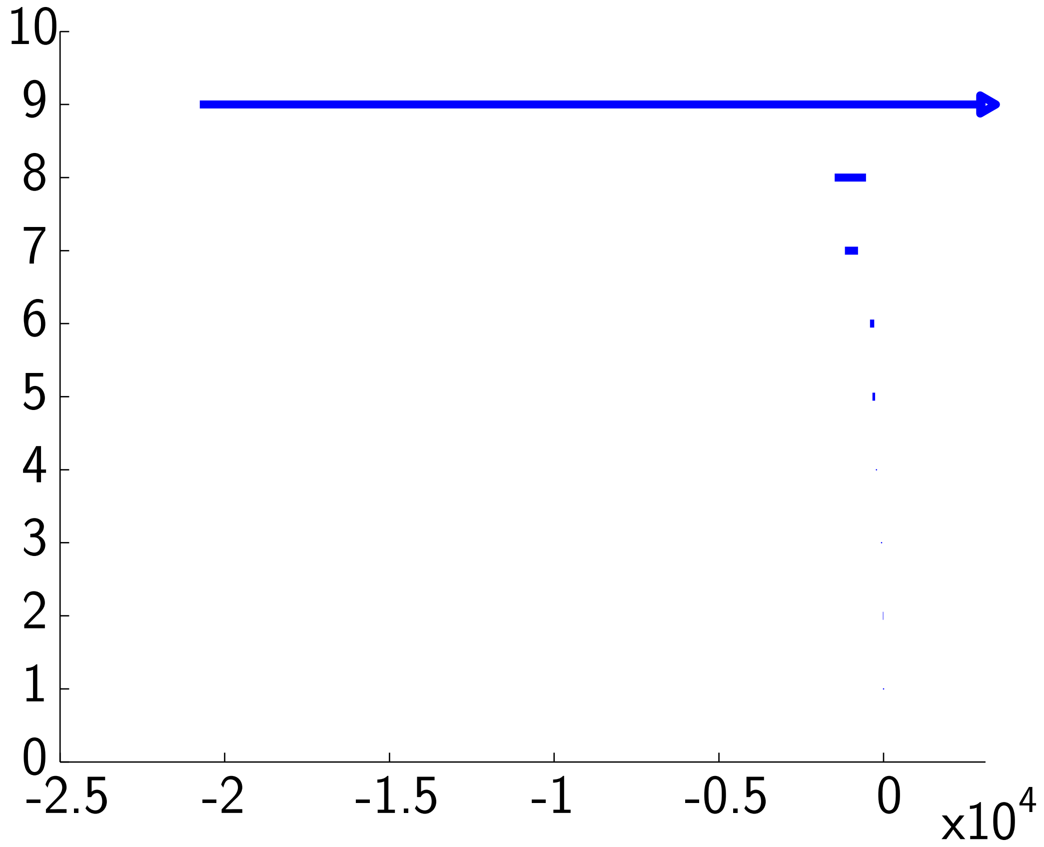






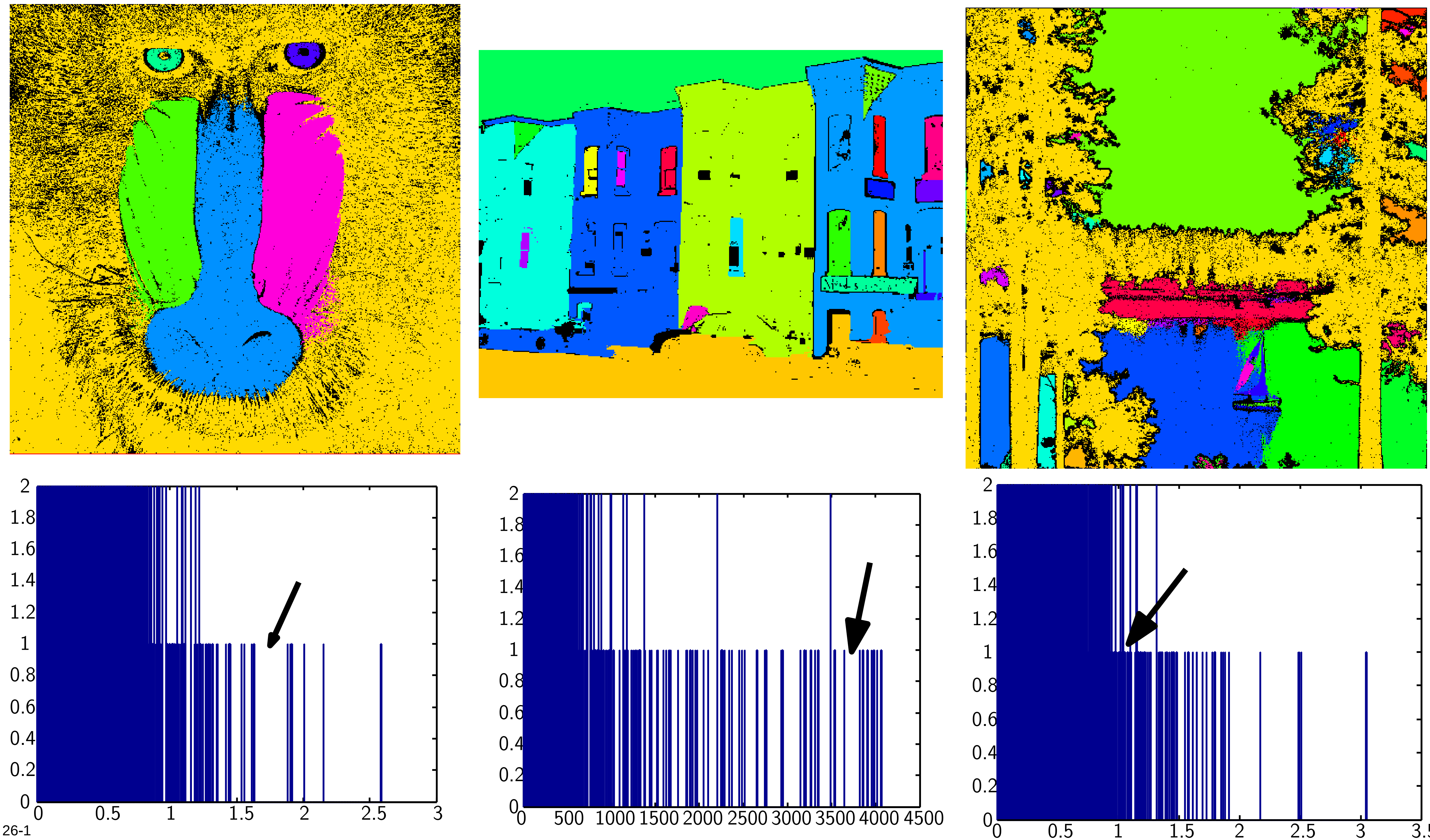
Koala





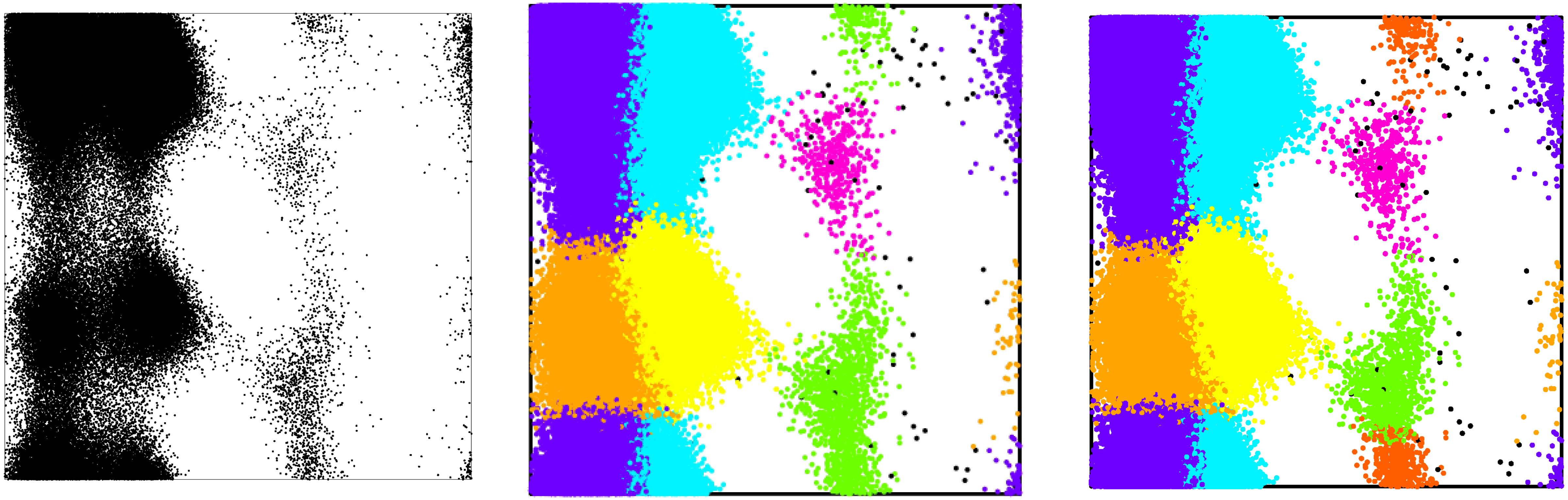
Incorporating Spatial Information

- Neighborhood graph: proximity in LUV space and image

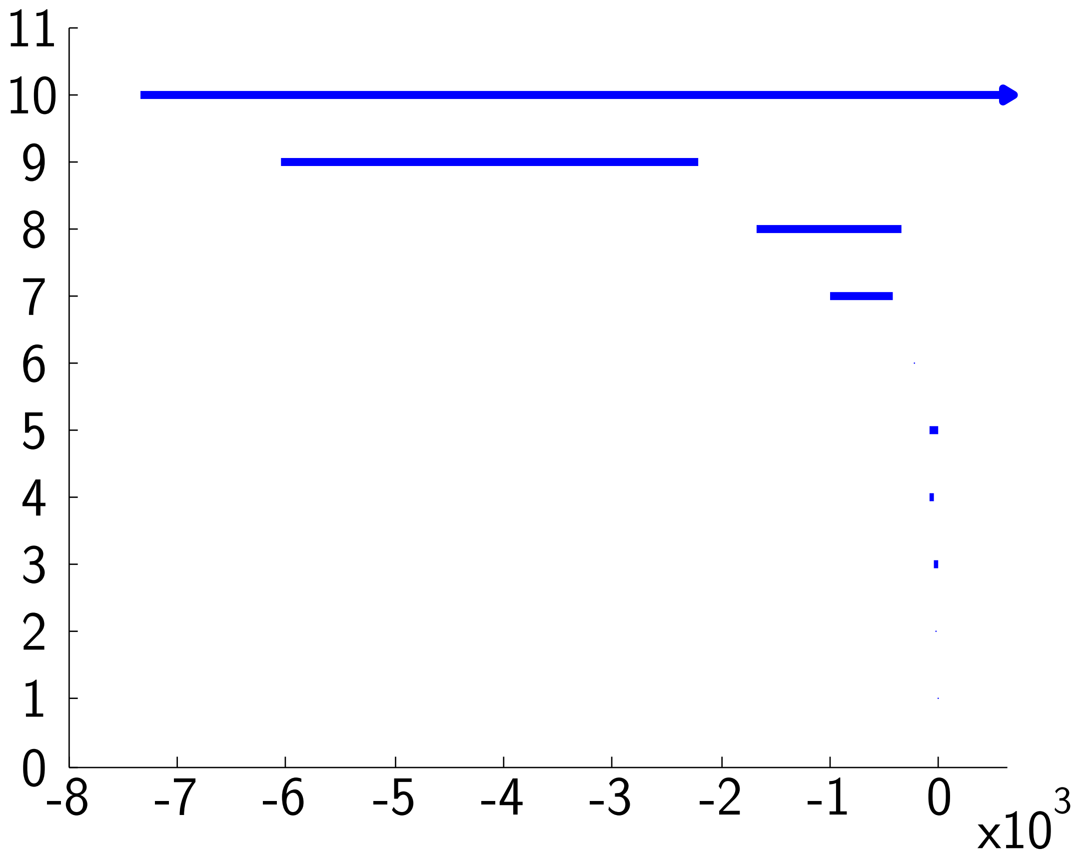


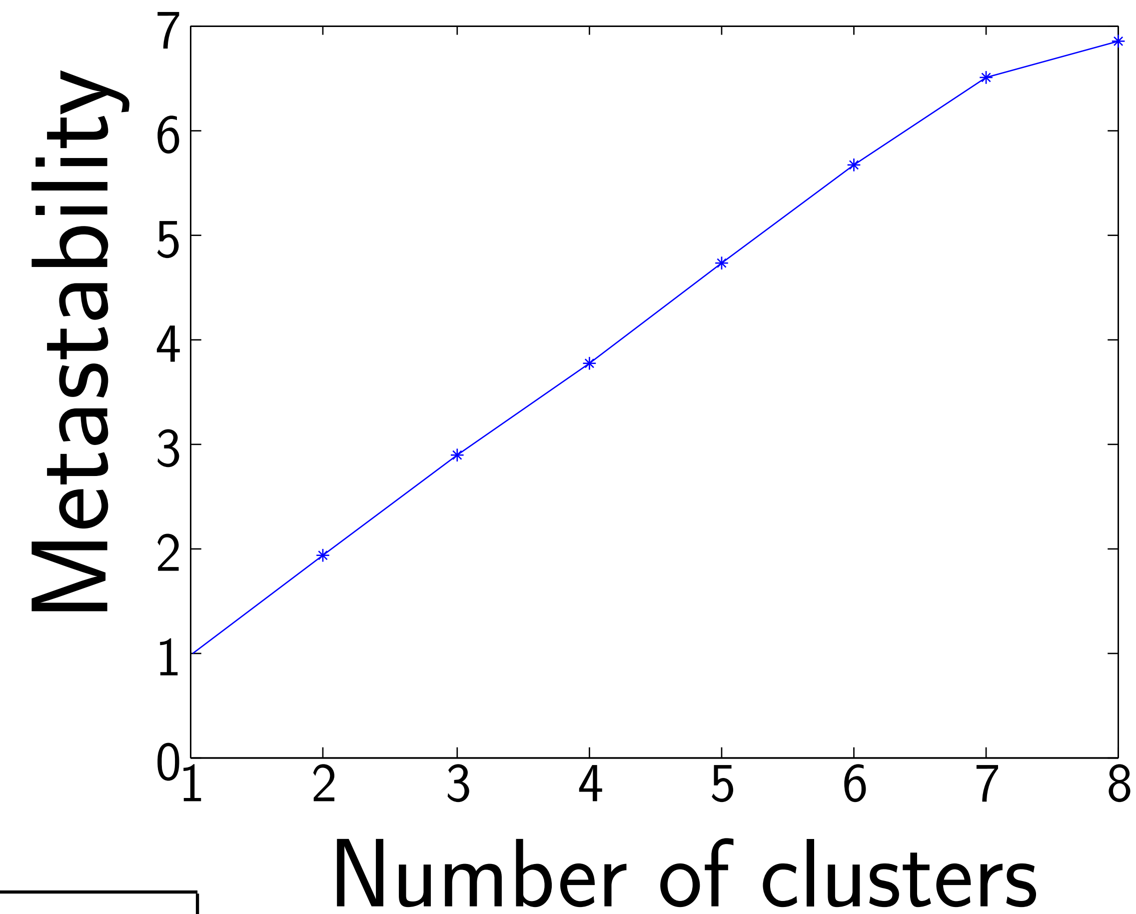
Alanine-dipeptide Conformations

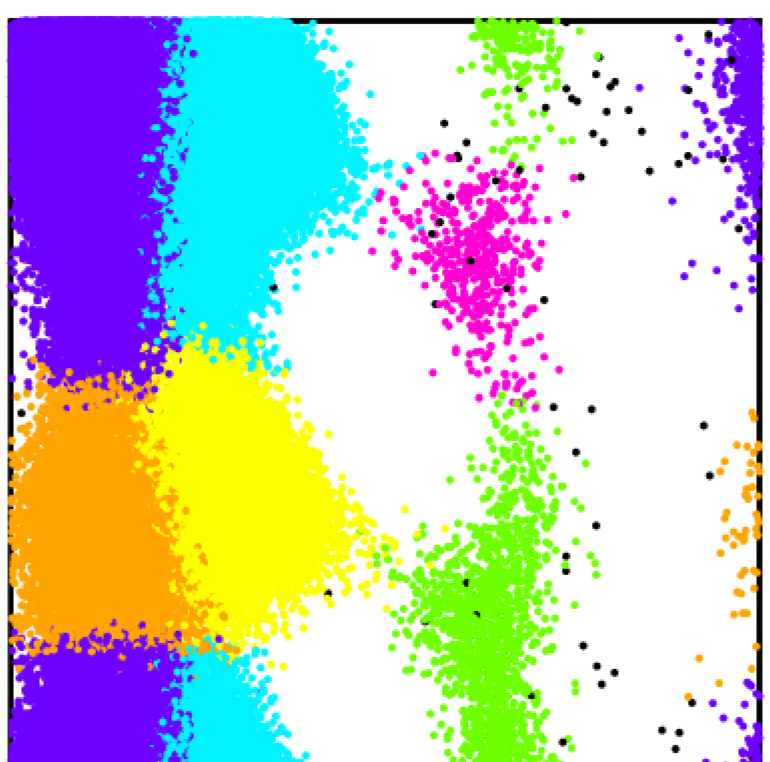
- Clustering in 22-dim space
- 192k points



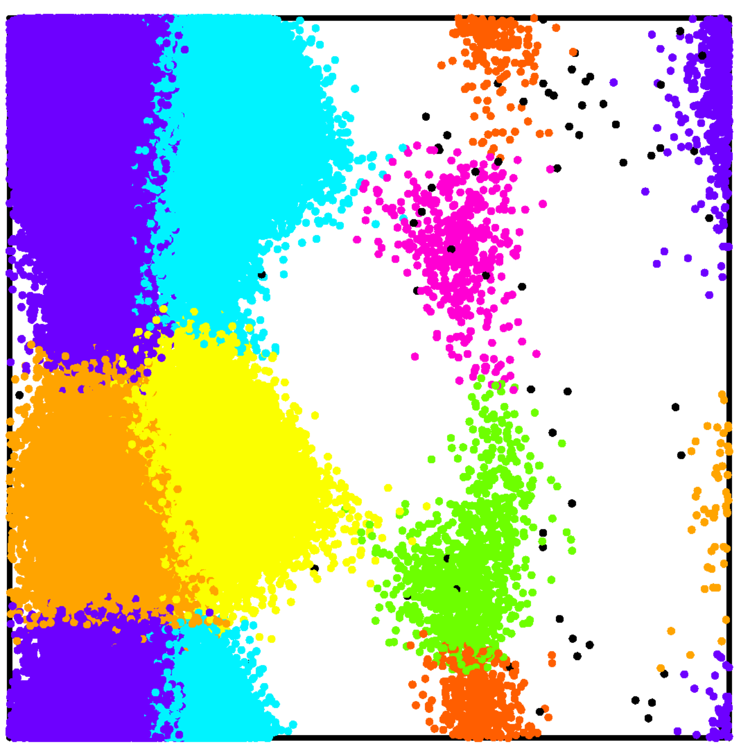
Alanine-dipeptide Conformations







Rank	Prominence
1	∞
2	5677
3	3828
4	1335
5	850
6	316
7	258
8	72
9	30
10	22



Spatial stability

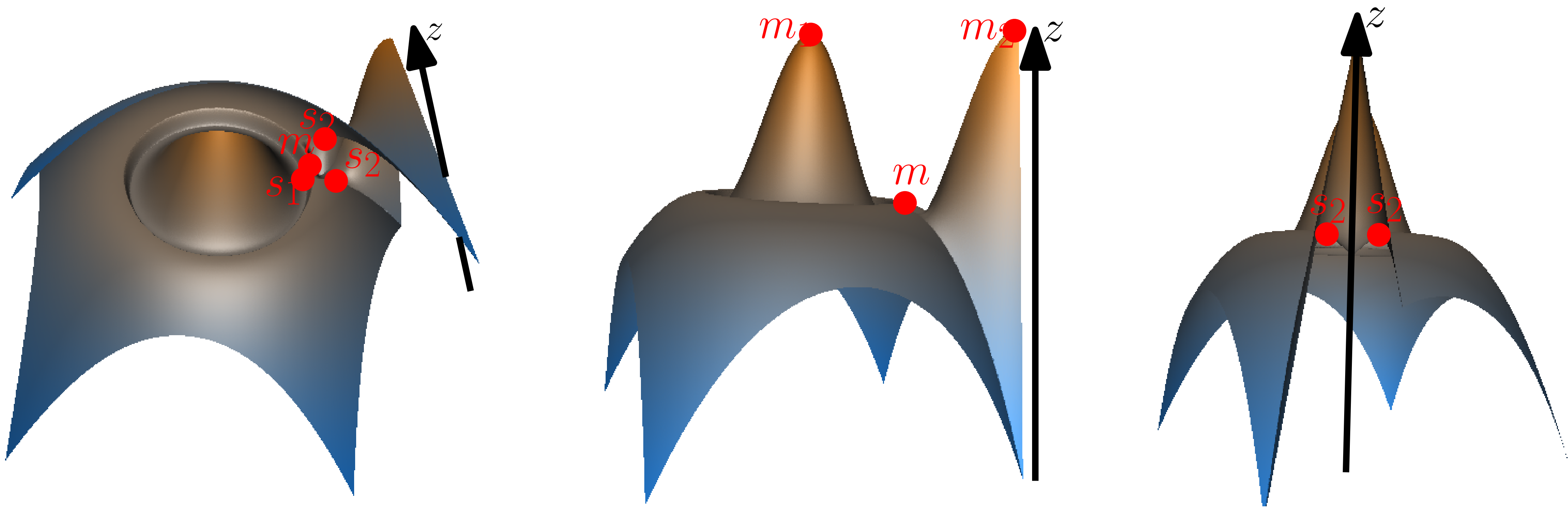
- Number of clusters are correct
- Can we say anything about the clusters themselves?

    1. Each prominent cluster has a stable part
    2. Unstable part can be very large

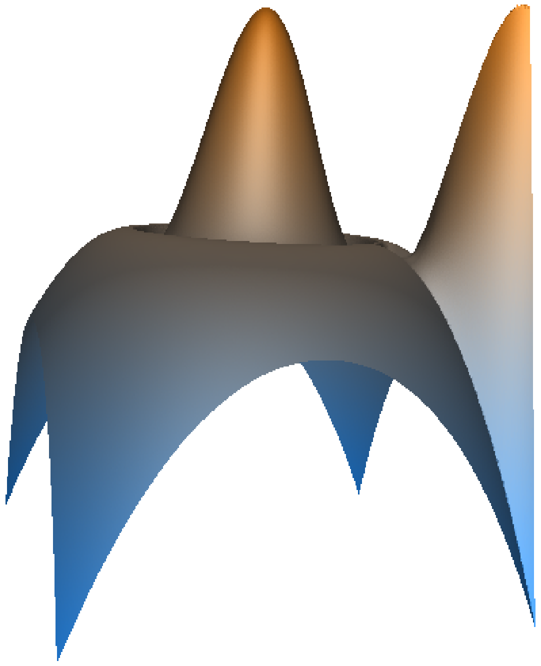
Spatial stability

- Number of clusters are correct
- Can we say anything about the clusters themselves?

    1. Each prominent cluster has a stable part
    2. Unstable part can be very large

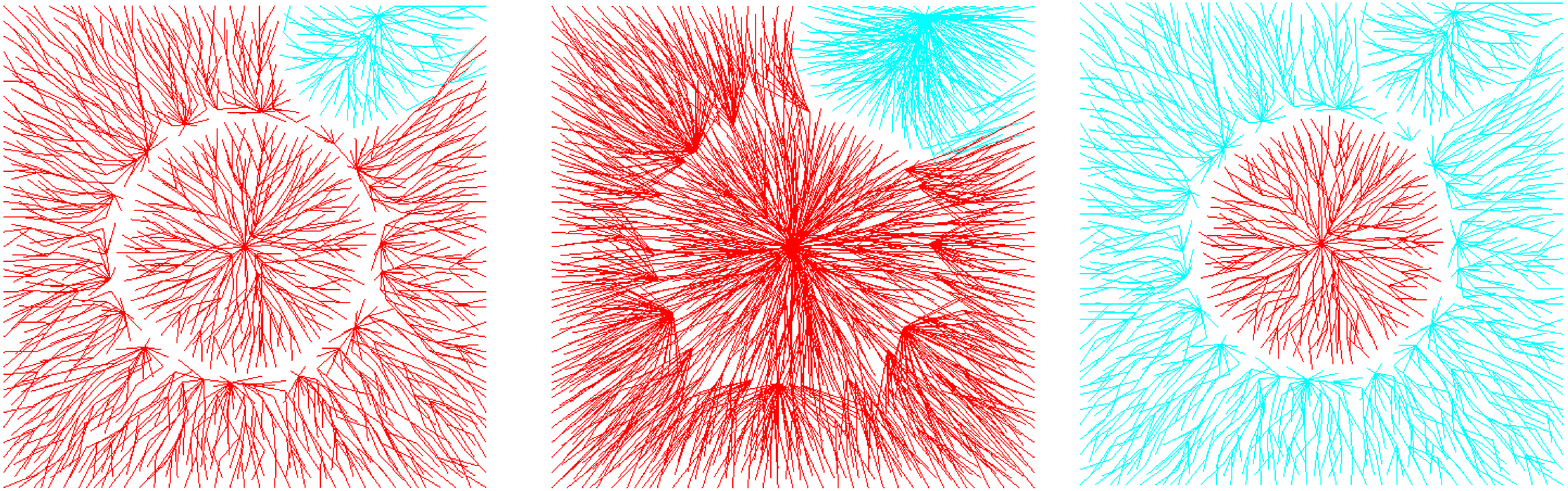


Spatial stability



- Number of clusters are correct
- Can we say anything about the clusters themselves?

    1. Each prominent cluster has a stable part
    2. Unstable part can be very large



Stable Part

Idea: Prominent clusters have a minimum size under c-Lipschitz assumption

- Under small pertubations, prominent peak part of the "same" cluster

- Soft clustering

    1. Run the algorithm multiple times, with small pertubations
    2. Find one-to-one correspondance between clusters
    3. Find stable and unstable parts

Conclusions

- Practical clustering algorithm (efficient in space and time)

- General framework
  - Use your favorite density estimator
  - Choice of neighborhood graph

- Easily-interpreted feedback
  - No "black box" effect

- Theoretical guarantees
  - Number of clusters
  - Spatial stability

- Soft-clustering

- Higher-dimensional features