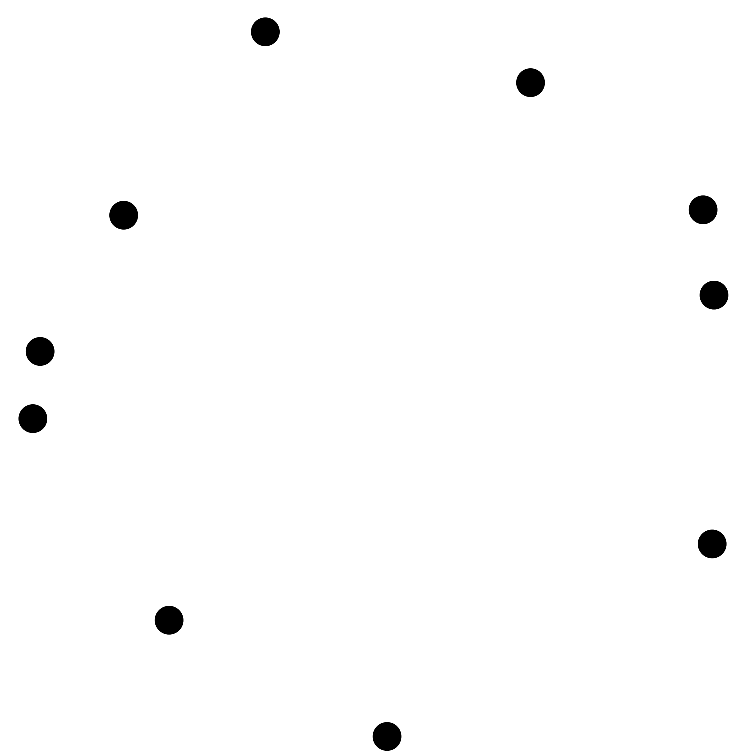
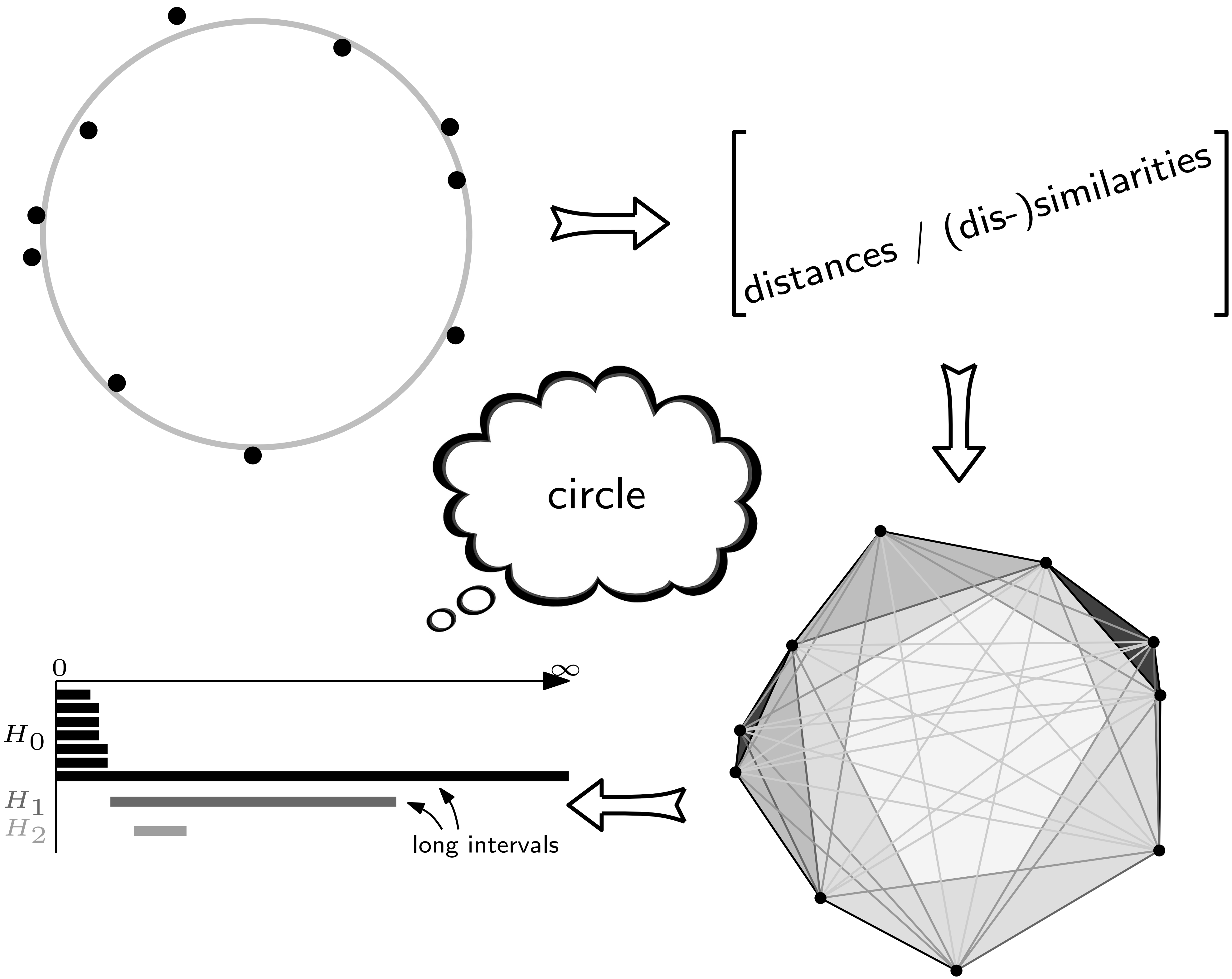


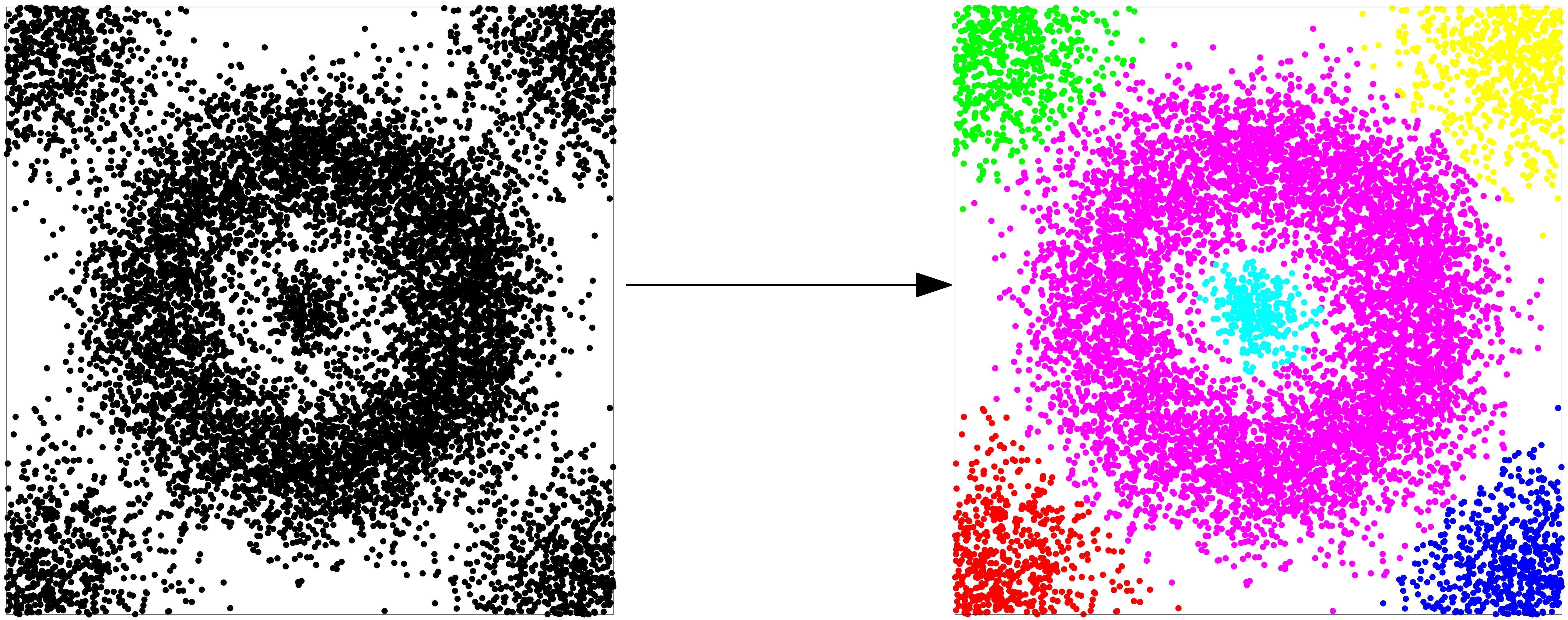
Topological Data Analysis



Cluster Analysis

Input: a finite set of observations: - point cloud with coordinates

- distance / (dis-)similarity matrix



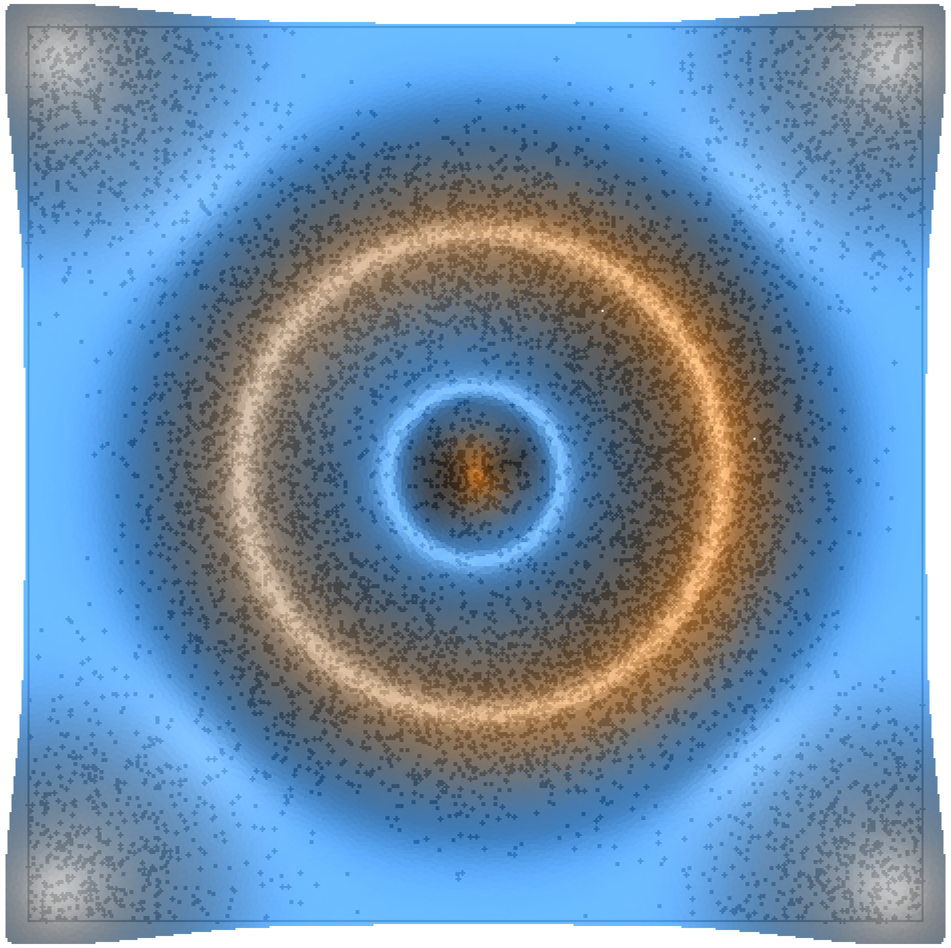
Task:

partition the data points into a collection of *relevant* subsets called clusters

Clustering using Topological Persistence

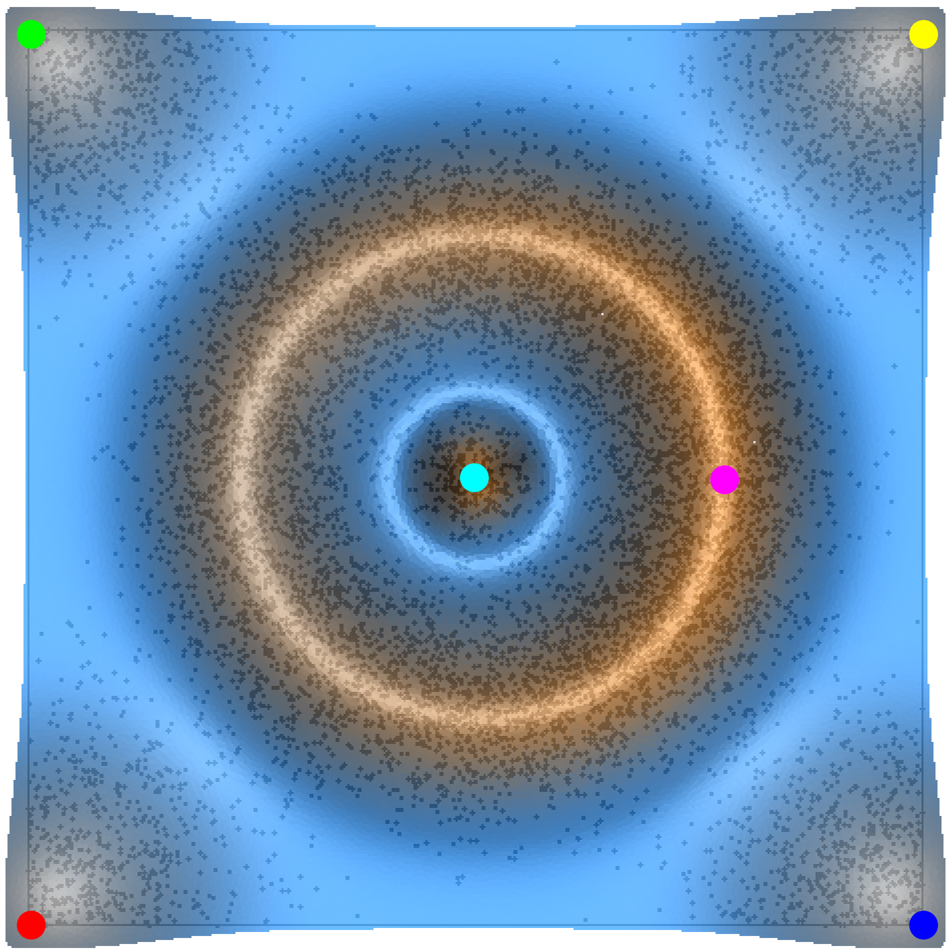
Mode-Seeking Paradigm

- Assume the data points are sampled from some unknown probability distribution
- Partition the data according to the basins of attraction of the peaks of the density



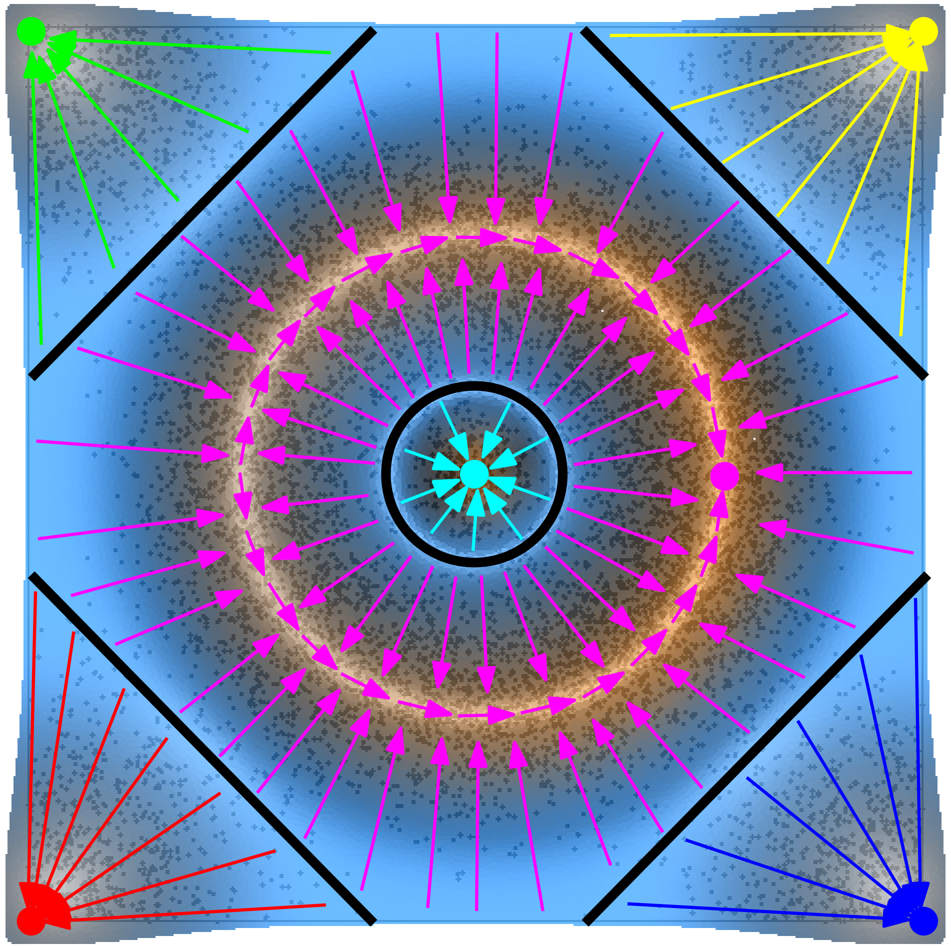
Mode-Seeking Paradigm

- Assume the data points are sampled from some unknown probability distribution
- Partition the data according to the basins of attraction of the peaks of the density



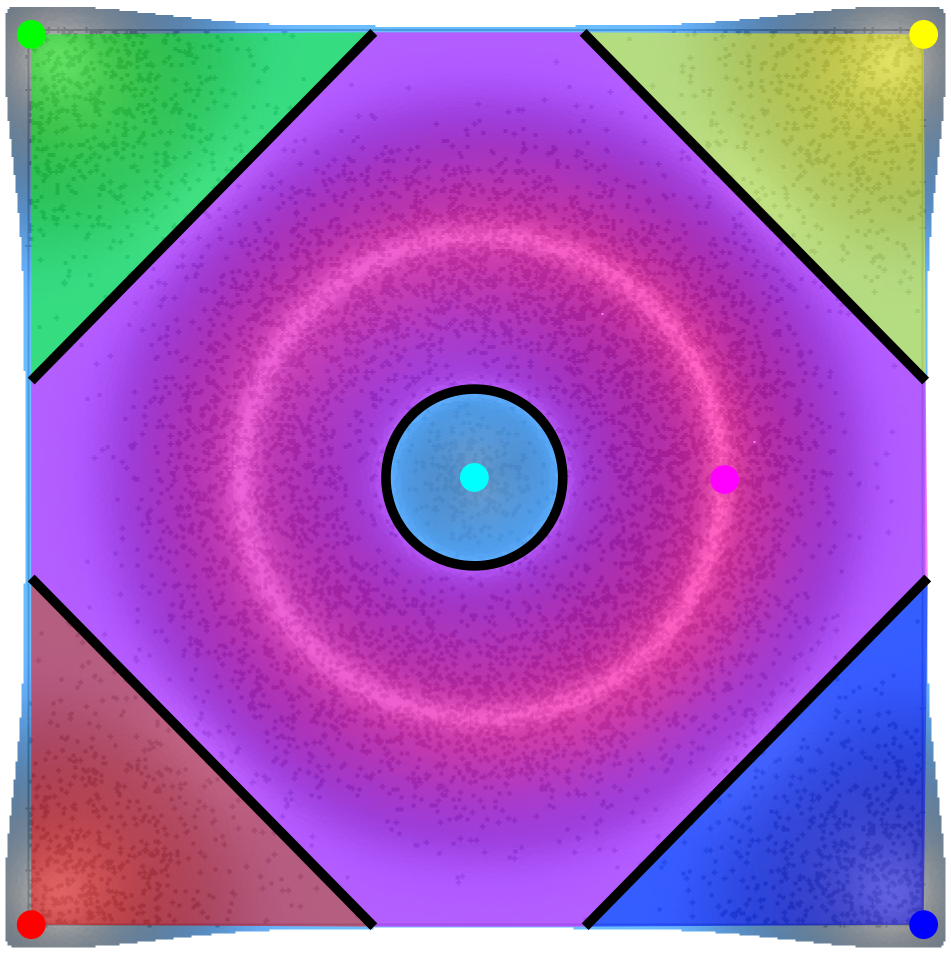
Mode-Seeking Paradigm

- Assume the data points are sampled from some unknown probability distribution
- Partition the data according to the basins of attraction of the peaks of the density



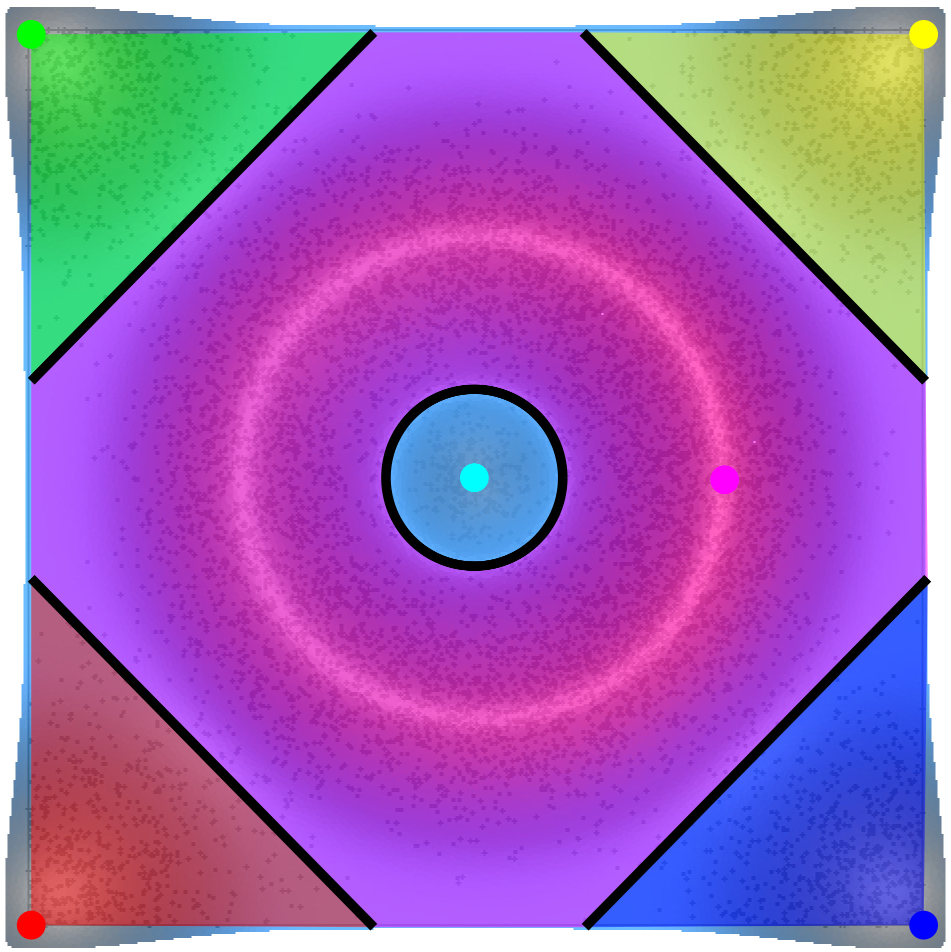
Mode-Seeking Paradigm

- Assume the data points are sampled from some unknown probability distribution
- Partition the data according to the basins of attraction of the peaks of the density

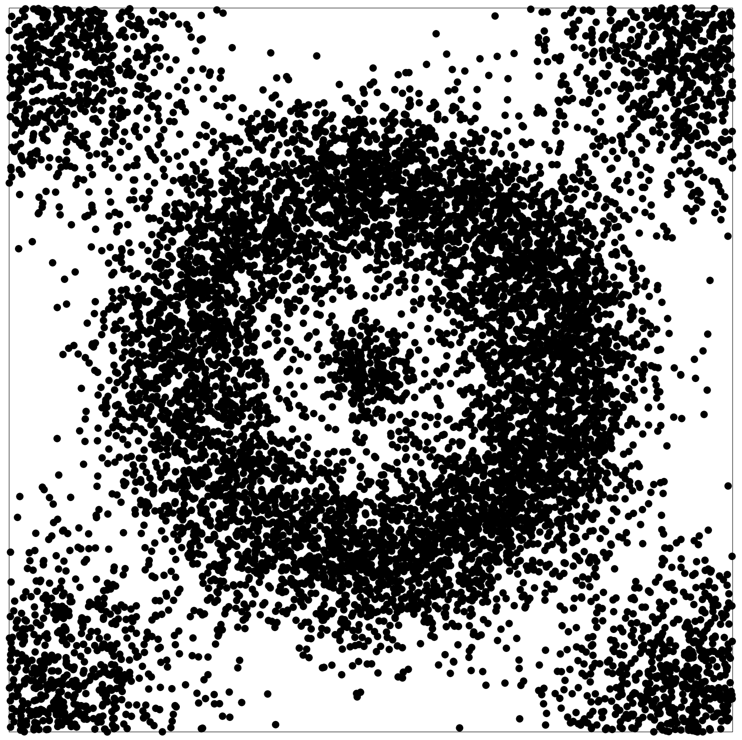


Mode-Seeking Paradigm

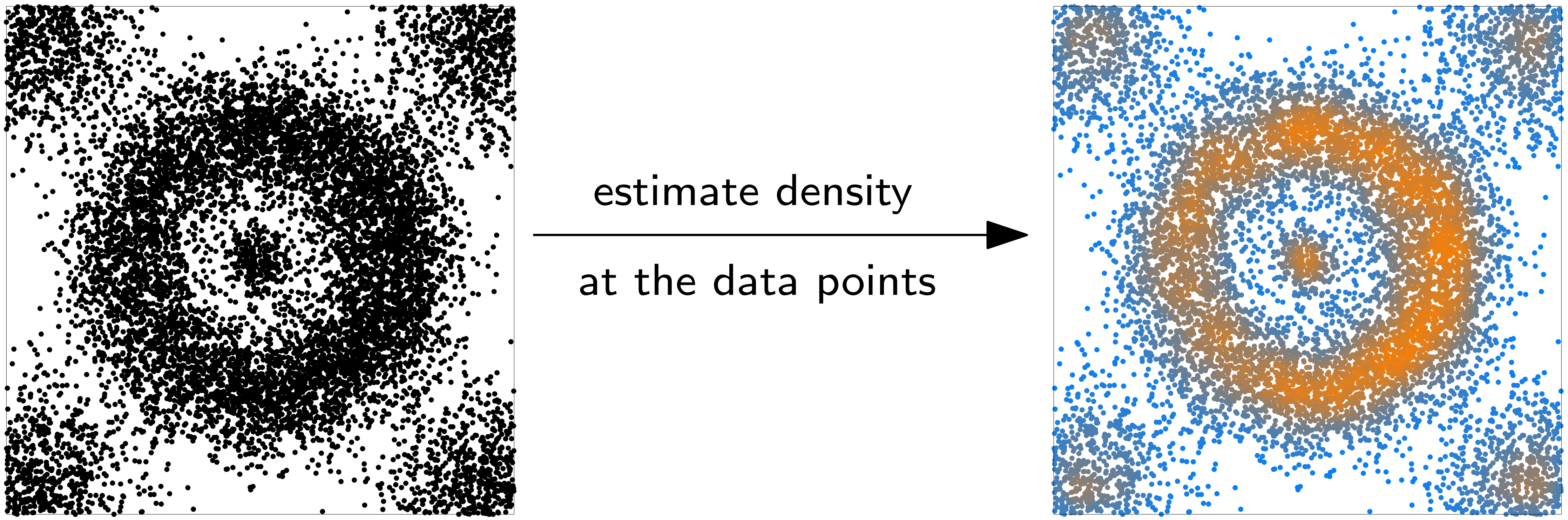
- Assume the data points are sampled from some $\boxed{\text{unknown}}$ probability distribution
- Partition the data according to the basins of attraction of the peaks of the density



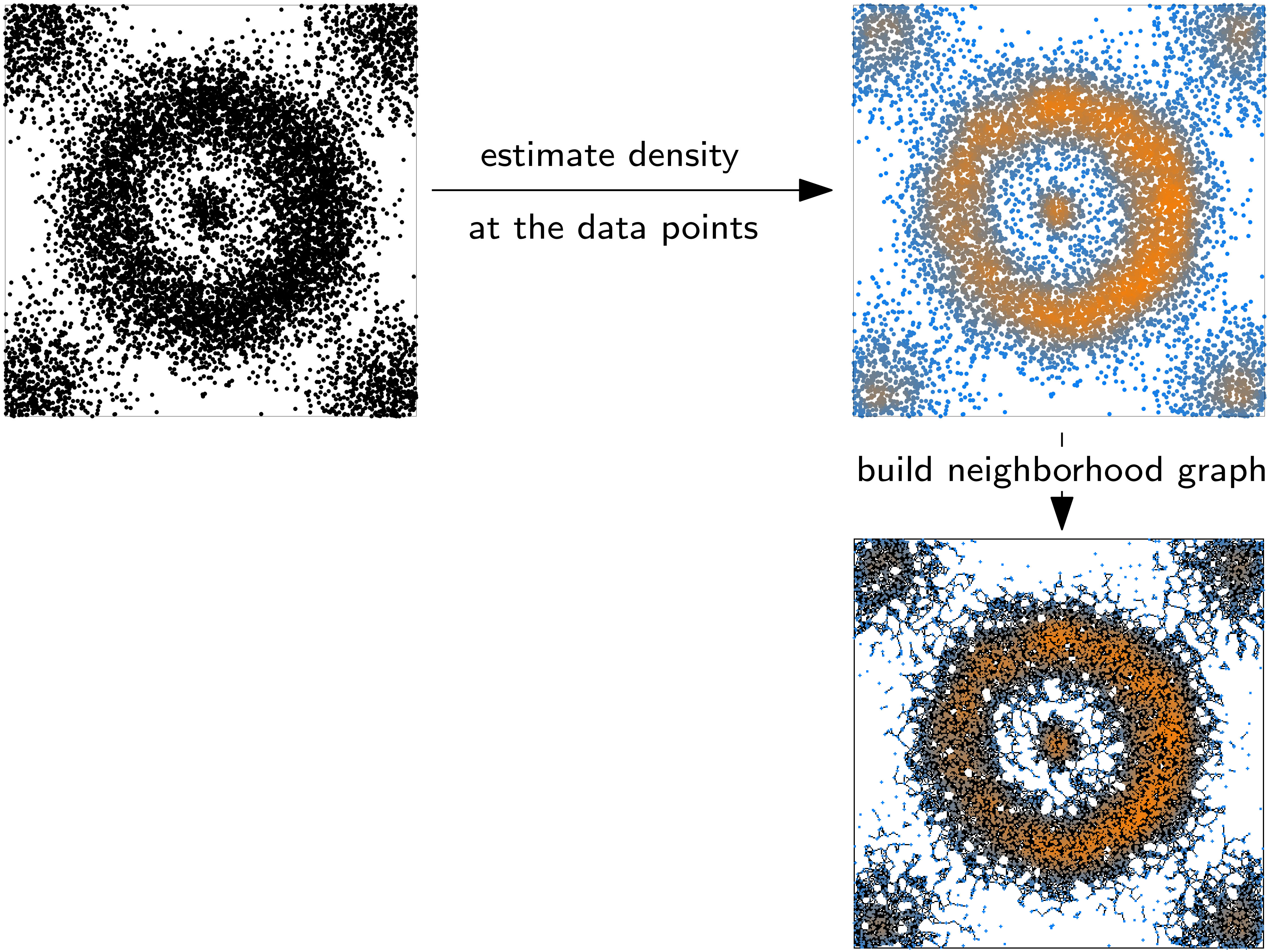
[Koontz, Narendra, Fukunaga'76] in a Nutshell



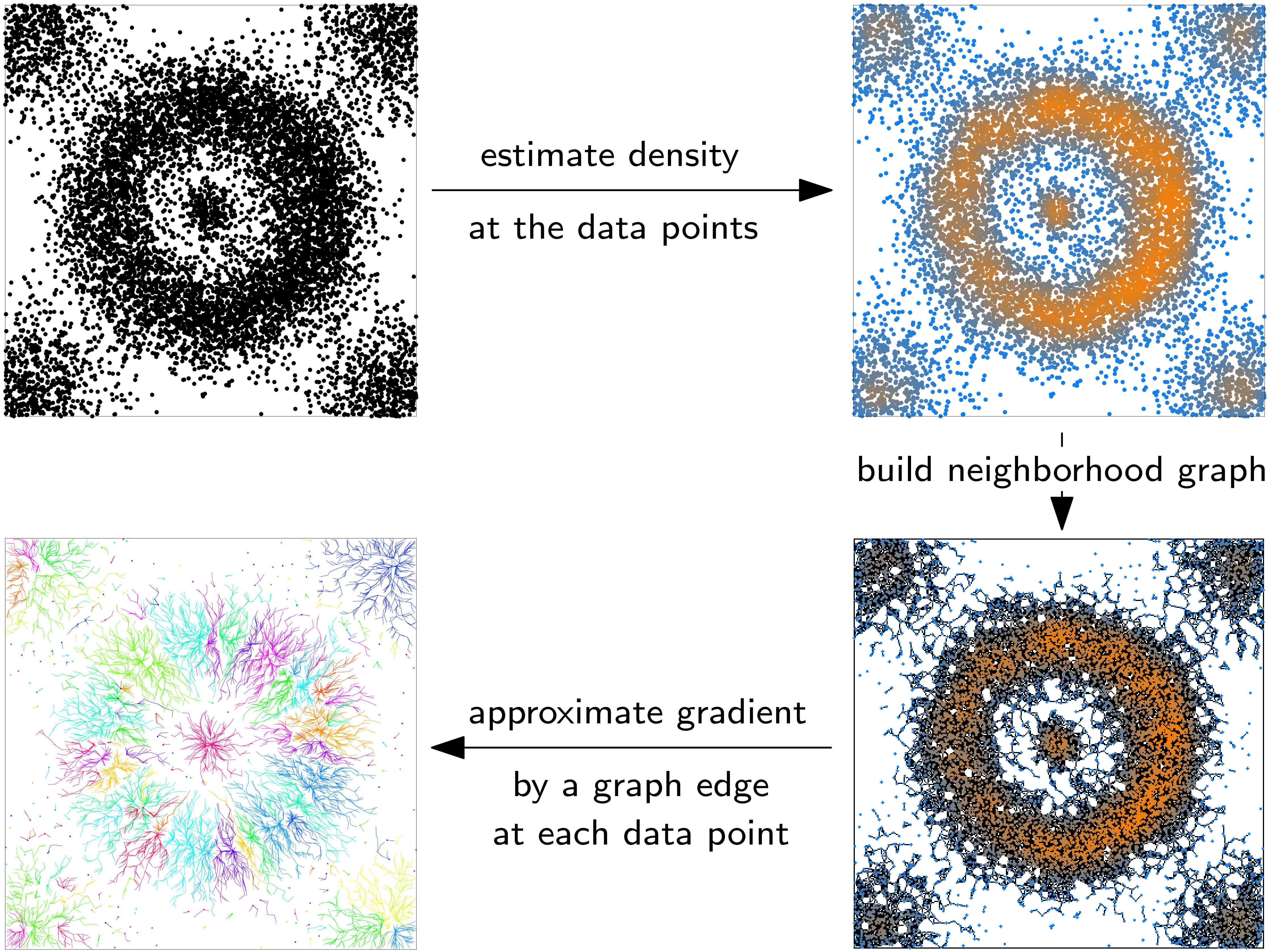
[Koontz, Narendra, Fukunaga'76] in a Nutshell



[Koontz, Narendra, Fukunaga'76] in a Nutshell

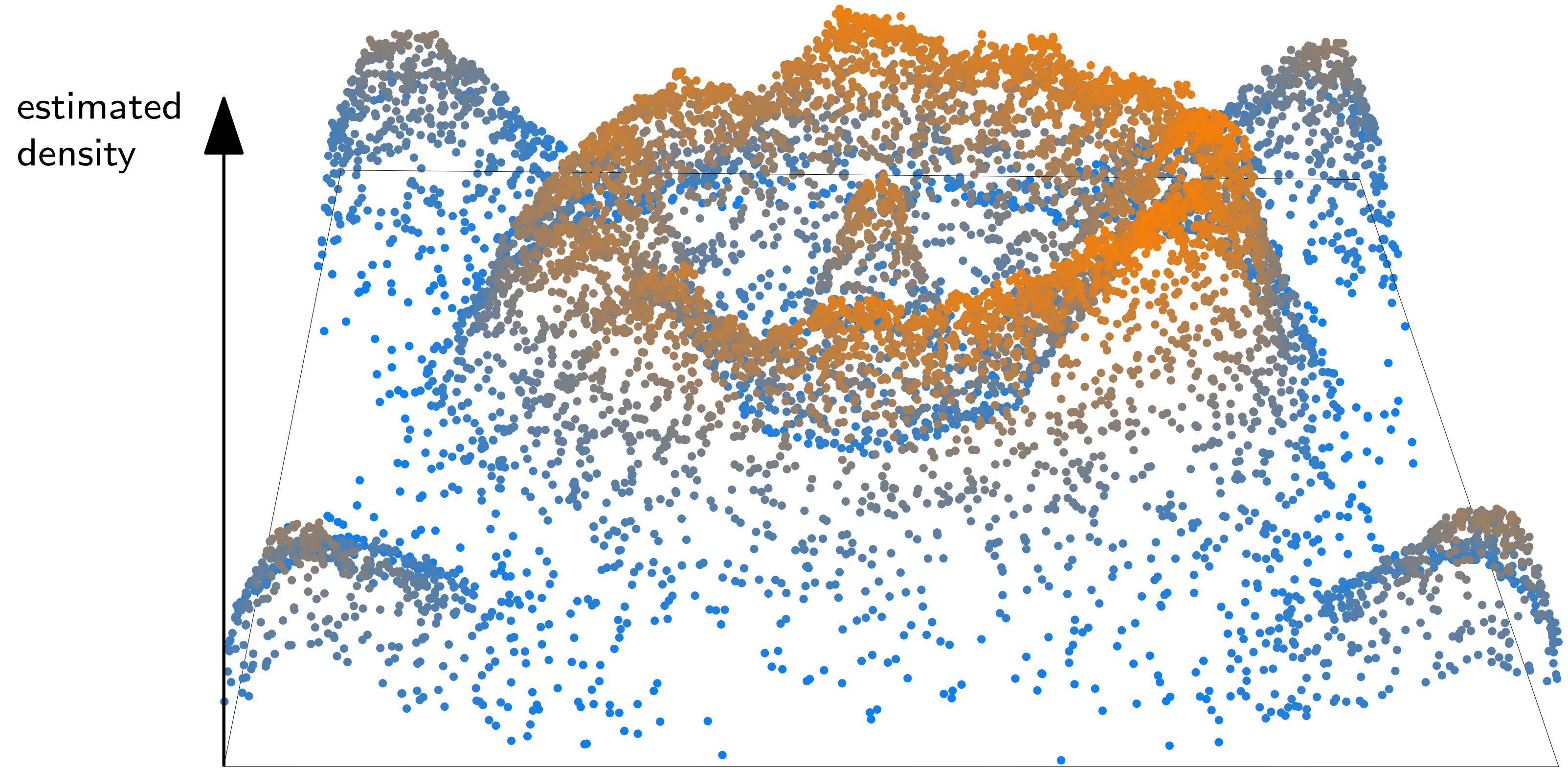


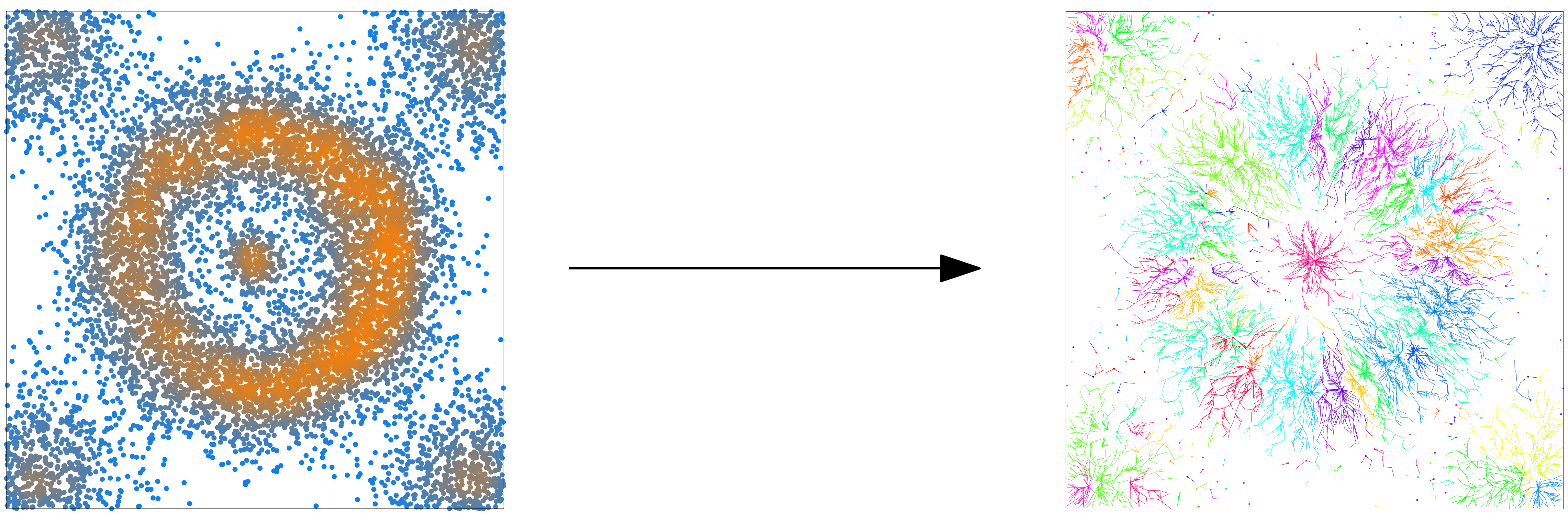
[Koontz, Narendra, Fukunaga'76] in a Nutshell



Why things are likely to go ill

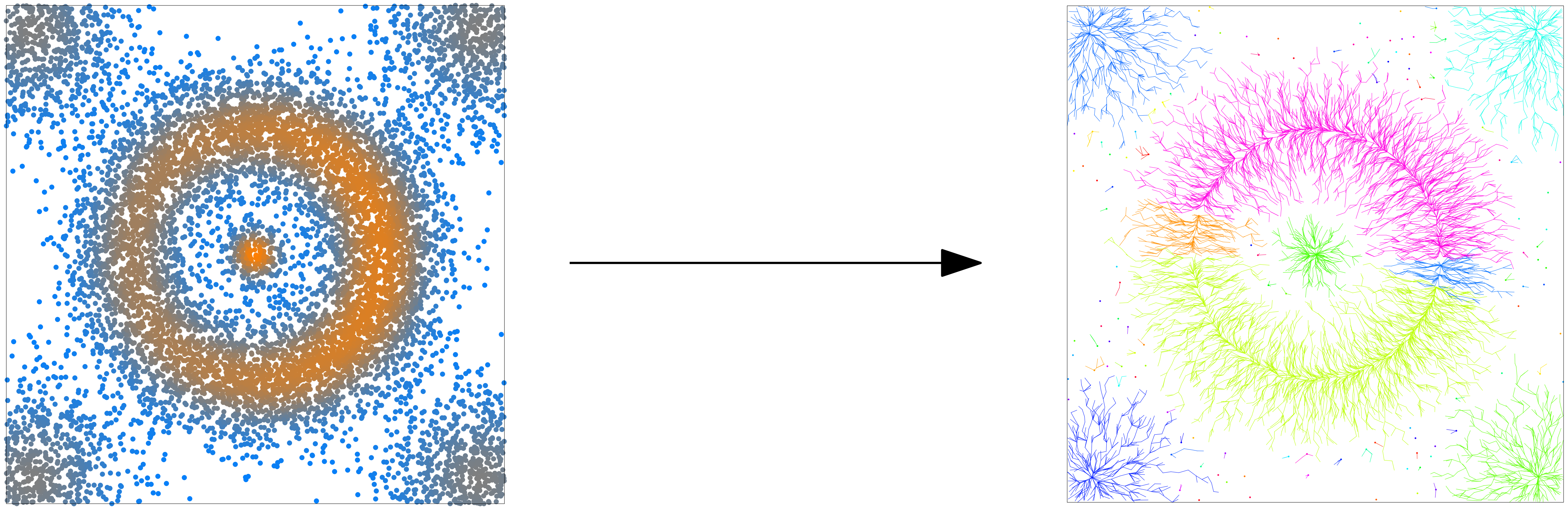
- Density estimator





Why things are likely to go ill

- Density estimator
- Neighborhood graph



Enter Topological Persistence

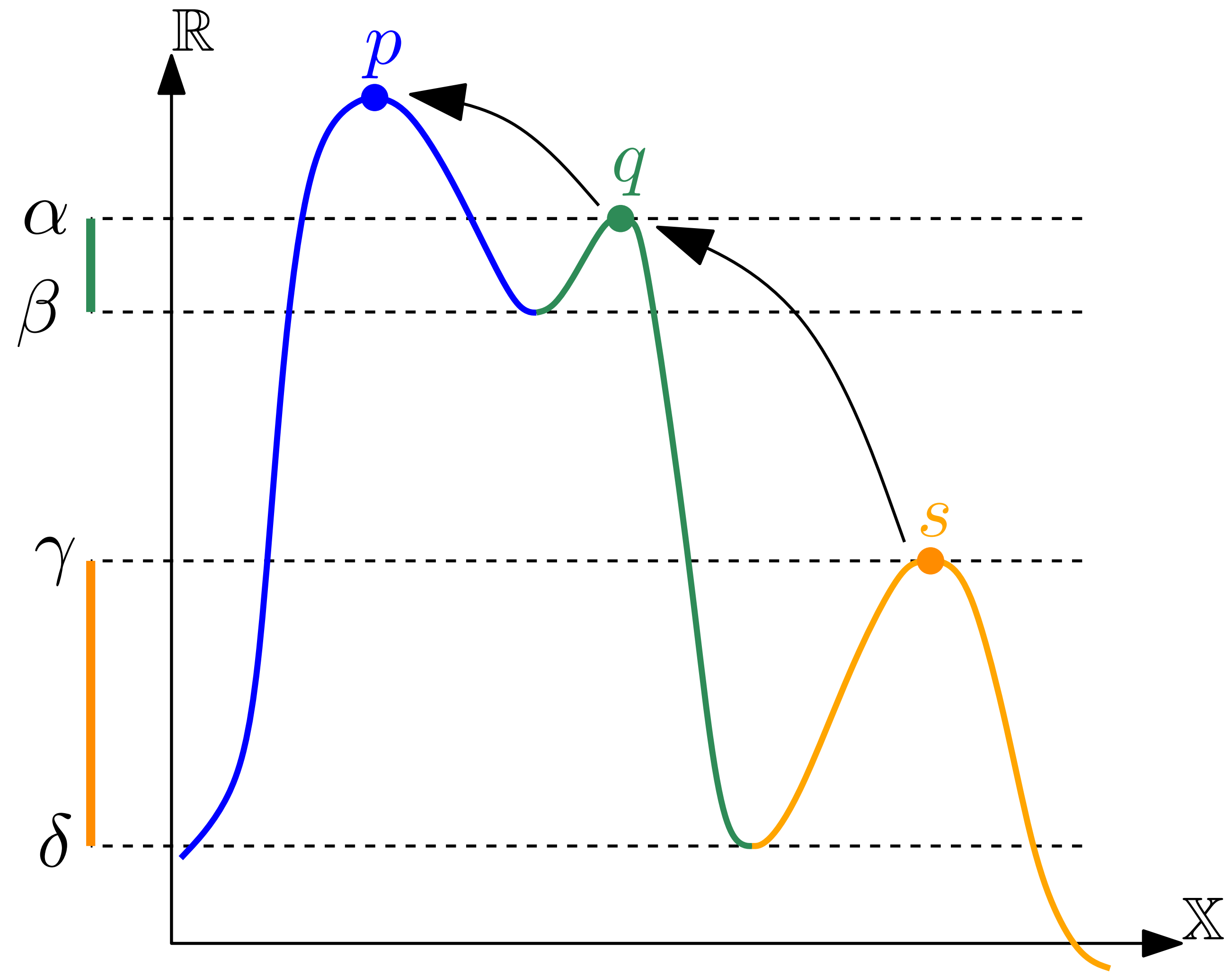
Given $f : \mathbb{X} \to \mathbb{R}$ ($\mathbb{X}$: ambient space / graph , f: density / estimator),

- quantifies the prominence of each peak of f,
- builds a hierarchy of the peaks of f.

prominence(p)$= +\infty$

prominence(q)$= \alpha - \beta$

prominence(s)$= \gamma - \delta$

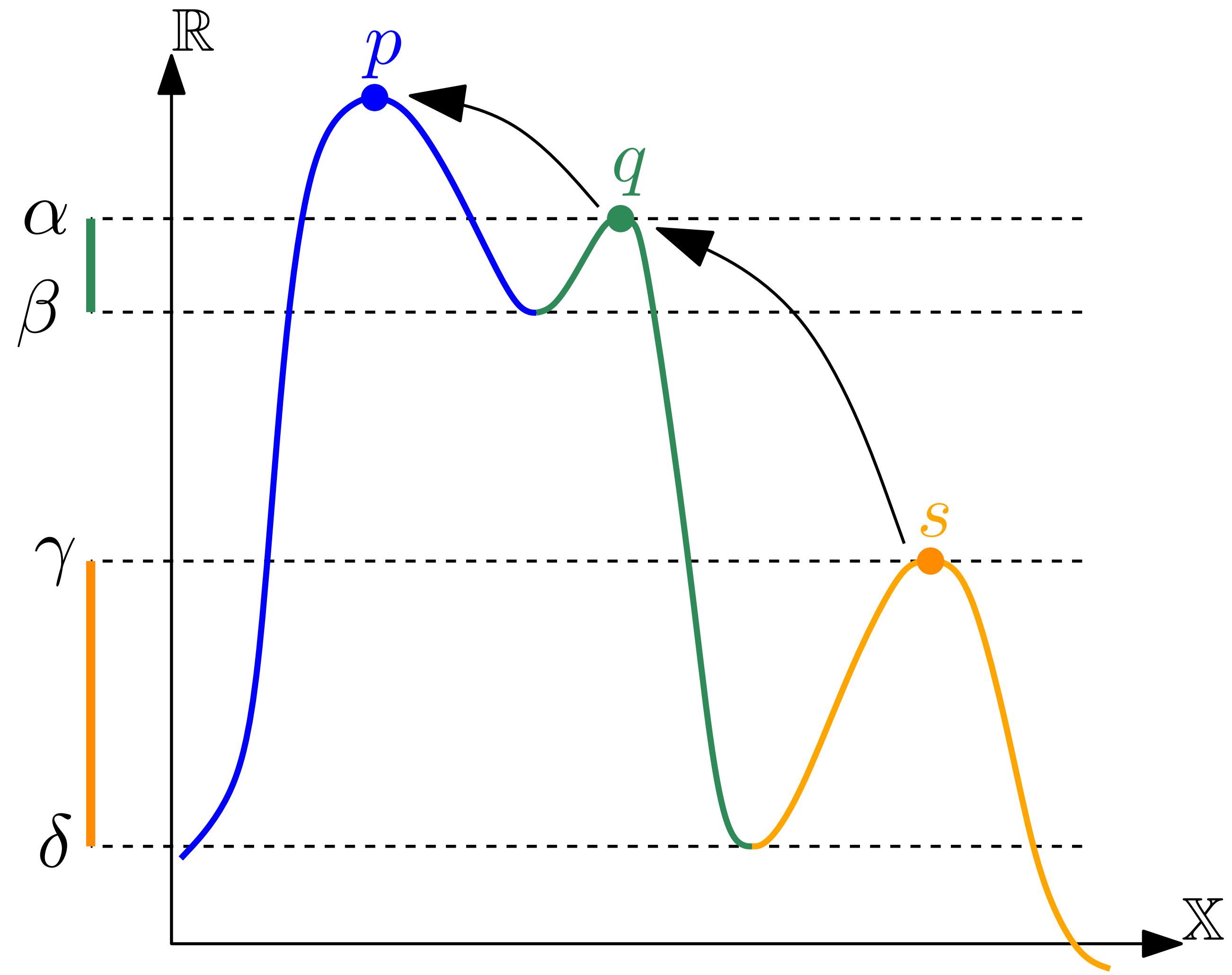


Merging Clusters using Persistence (in a nutshell)

Given the initial clustering:

- Choose a threshold $\tau \geq 0$ and merge those clusters of prominence $< \tau$
- merge clusters according to the hierarchy (merge each cluster into its parent)

$$0 \leq \tau \leq \alpha - \beta$$



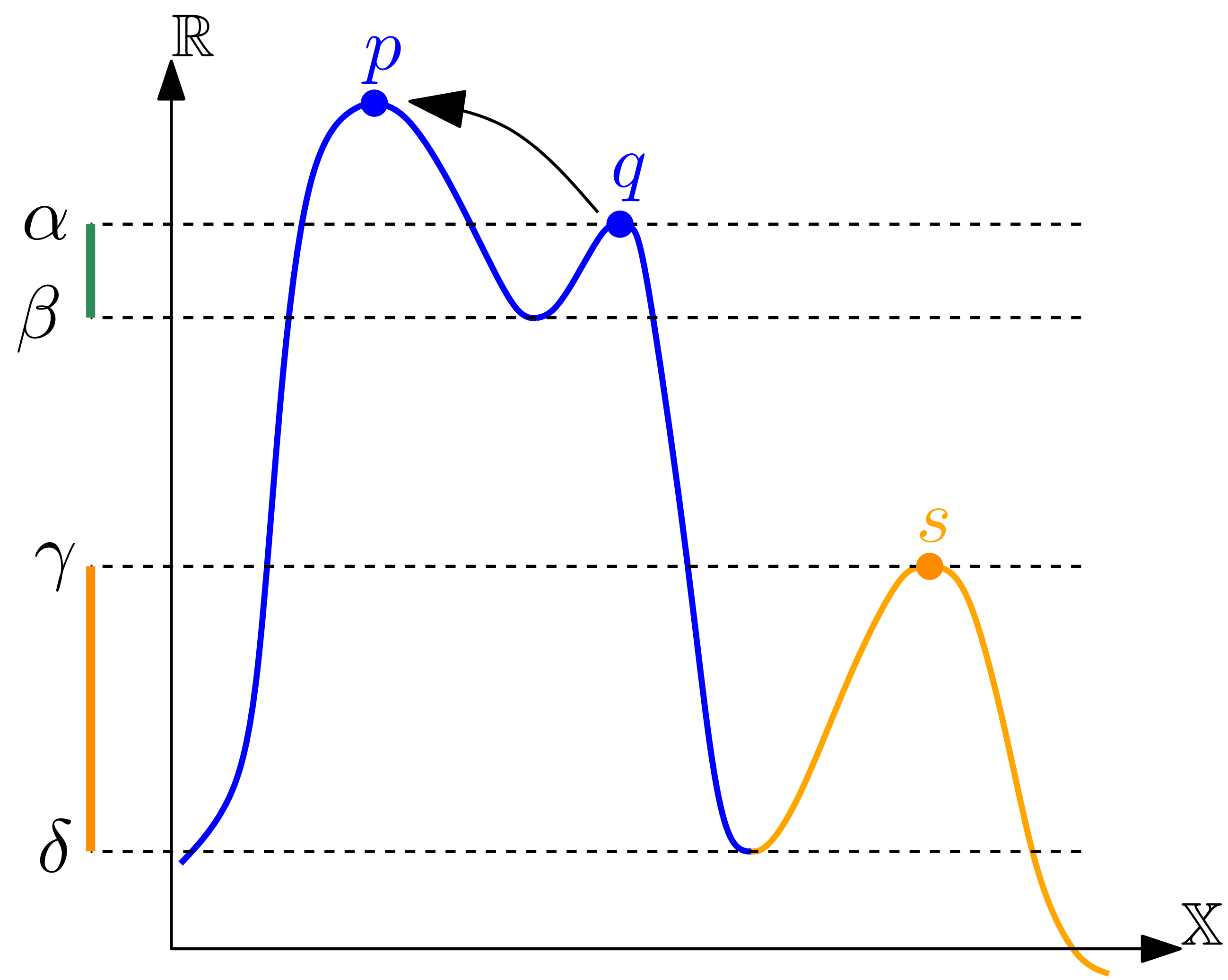
Merging Clusters using Persistence (in a nutshell)

Given the initial clustering:

- Choose a threshold $\tau \geq 0$ and merge those clusters of prominence $< \tau$

- merge clusters according to the hierarchy (merge each cluster into its parent)

$$\alpha - \beta < \tau \leq \gamma - \delta$$



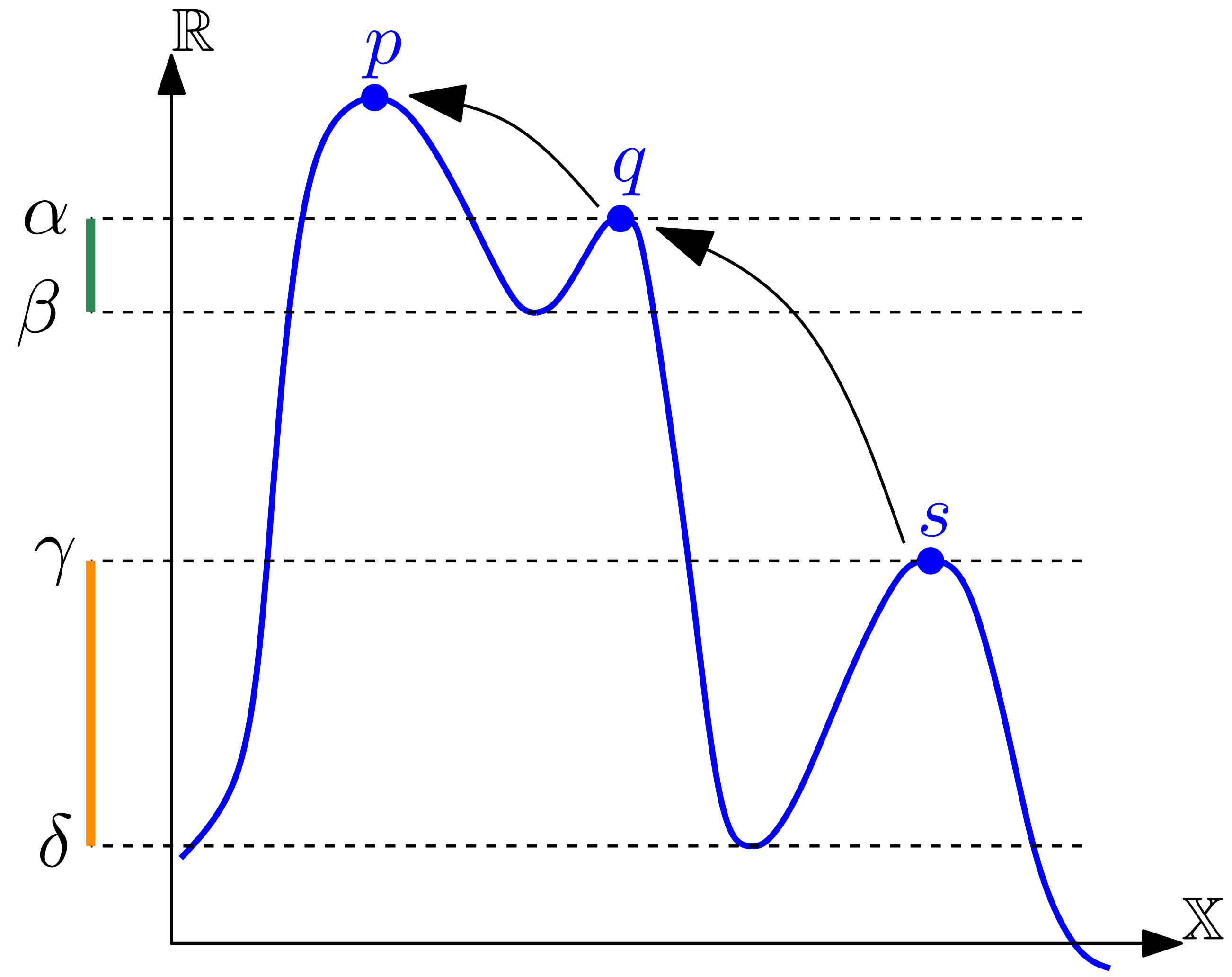
Merging Clusters using Persistence (in a nutshell)

Given the initial clustering:

- Choose a threshold $\tau \geq 0$ and merge those clusters of prominence $< \tau$

- merge clusters according to the hierarchy (merge each cluster into its parent)

$$\gamma - \delta < \tau \leq +\infty$$

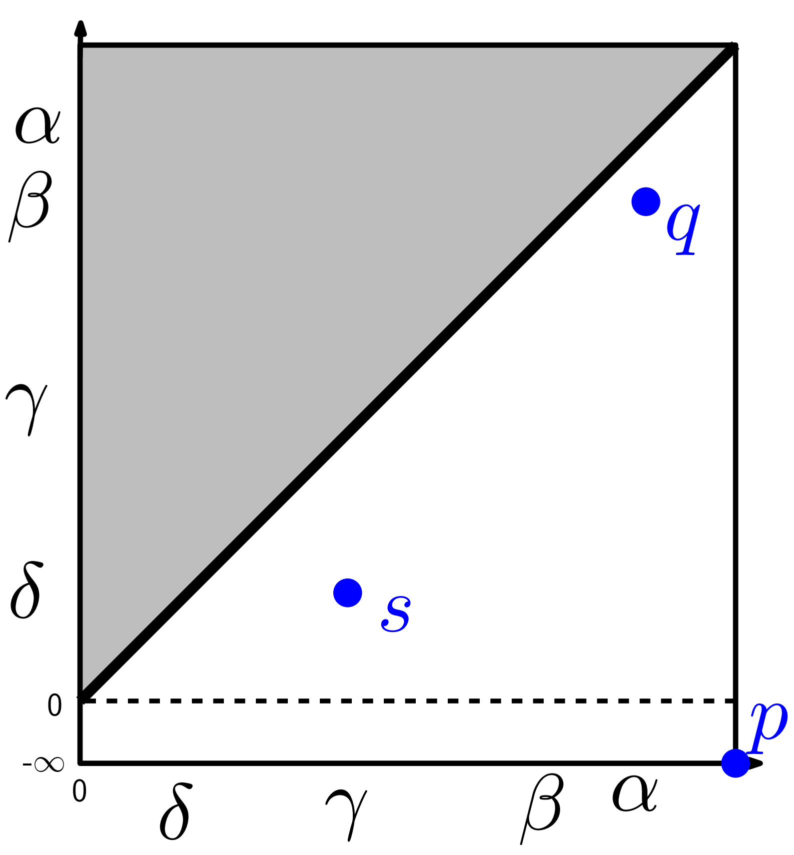


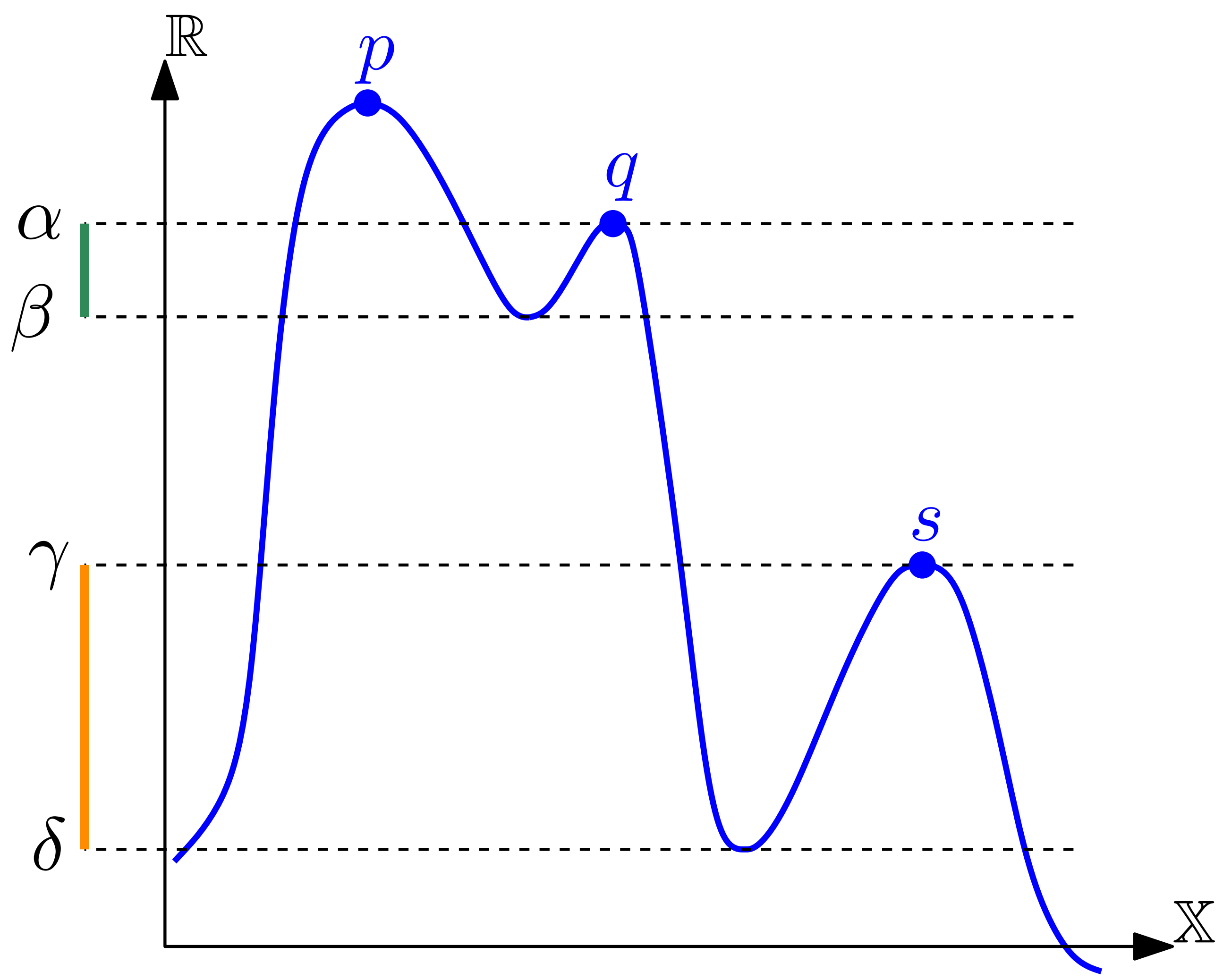
Merging Clusters using Persistence (in a nutshell)

Choice of merging parameter τ:

- report prominences in a 2-d *persistence diagram*

- **promise:** under suitable conditions, diagrams of density and estimator are *close*





Merging Clusters using Persistence (in a nutshell)

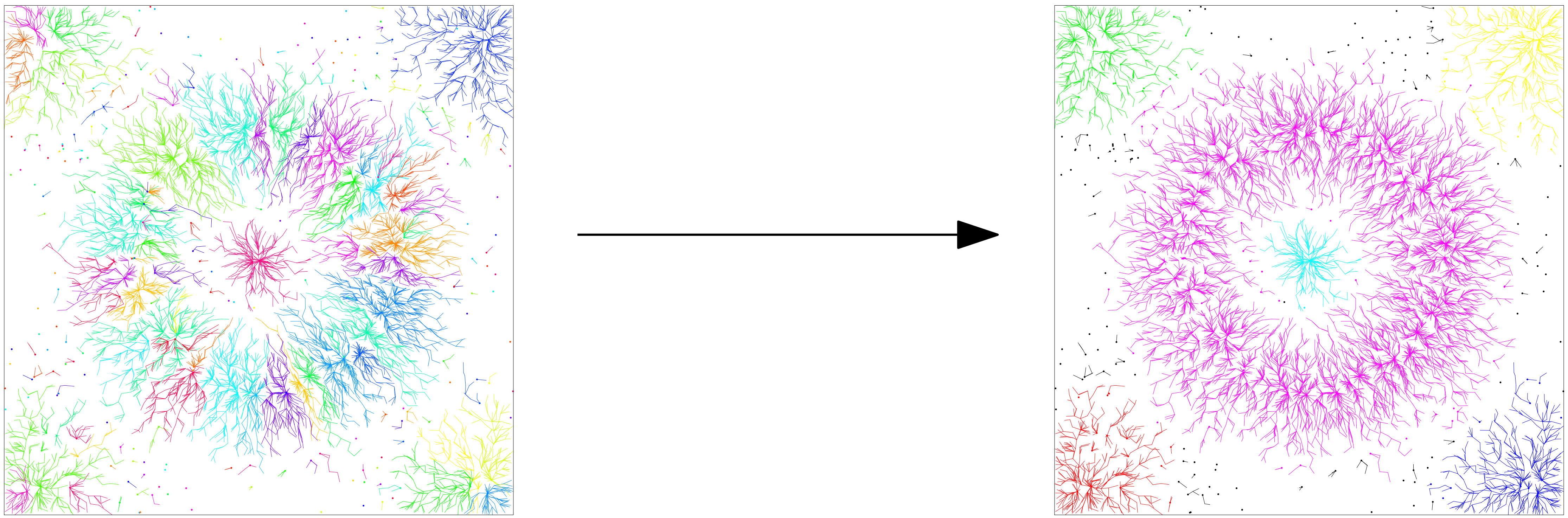
Choice of merging parameter τ:

- report prominences in a 2-d *persistence diagram*

- **promise:** under suitable conditions, diagrams of density and estimator are *close*

  $\rightarrow$ signal can be distinguished from noise

  $\rightarrow$ output after merge has the correct number of clusters

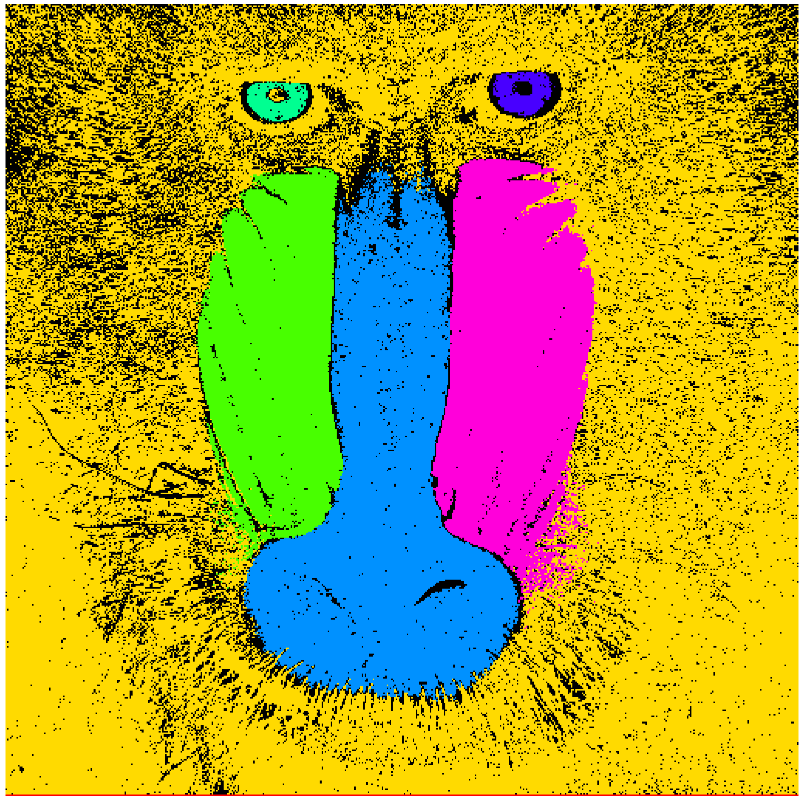


Experimental Results

Image Segmentation

Density is estimated in 3D color space (Luv)

Neighborhood graph is built in image domain



Distribution of prominences does not usually exhibit a clear unique gap

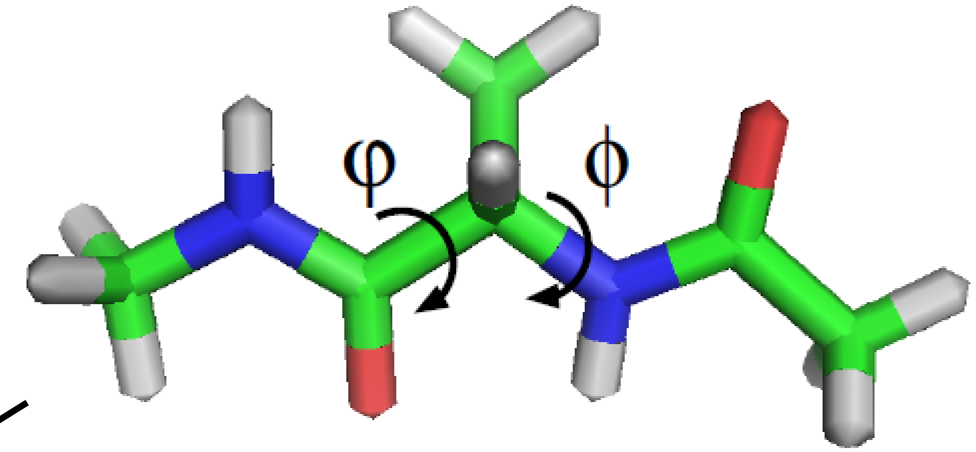
Still, relationship between choice of τ and number of obtained clusters remains explicit

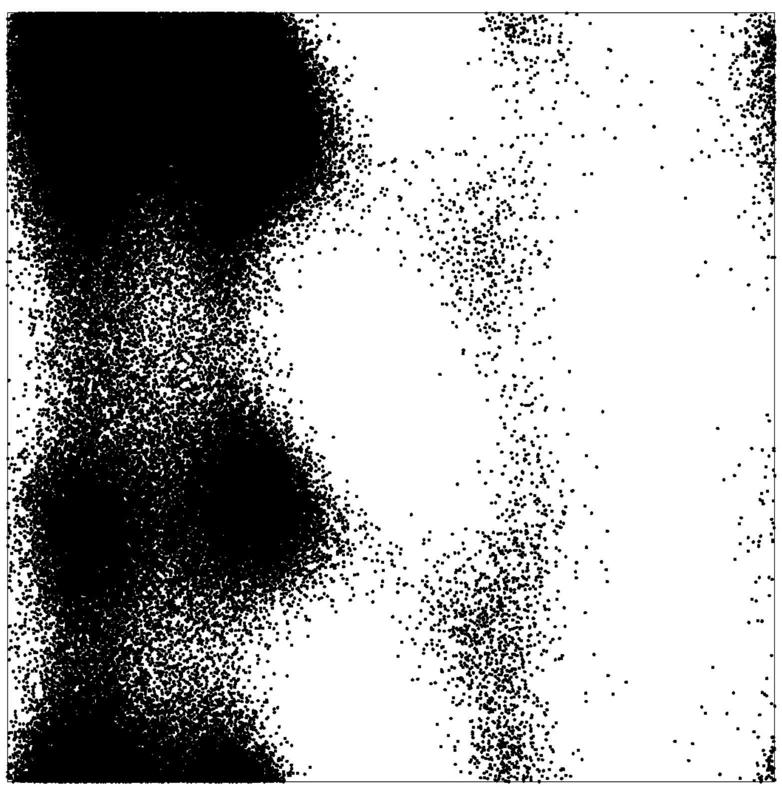
Experimental Results

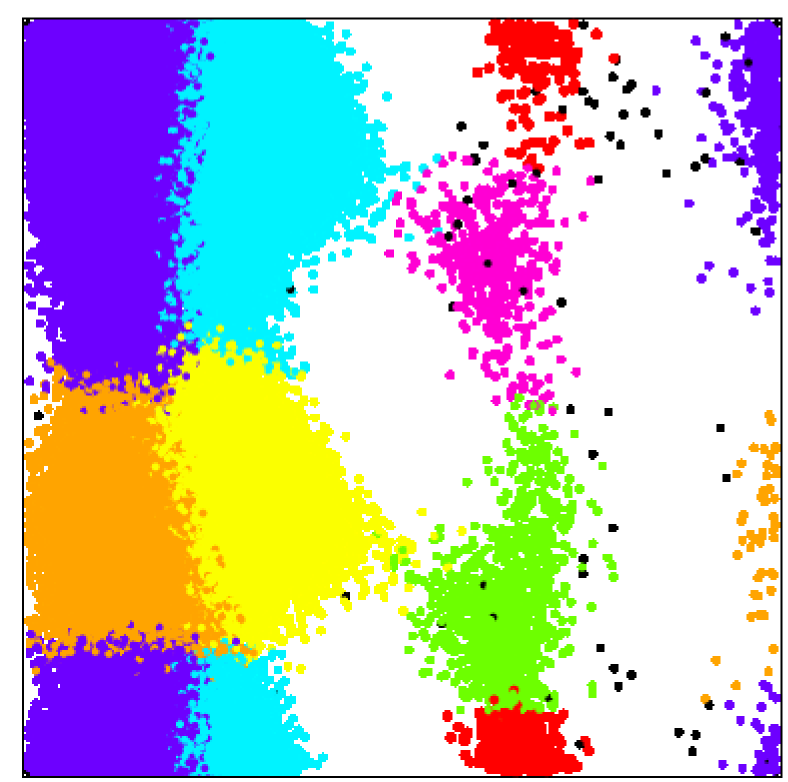
Biological Data

Alanine-Dipeptide conformations ($\mathbb{R}^{21}$)

RMSD distance (non-Euclidean)







Common belief: 6 metastable states

PD shows anywhere between 4 and 7 clusters

To Do

Theory:

- soft clustering:

  → repeat experiment with perturbed estimator

  → track clusters across runs

  → get stochastic vectors

  → theoretical guarantees?

$$\begin{array}{cc} \text{clusters} & \left[\begin{array}{c} \qquad\qquad \\ \\ \end{array} \right] \\ & \text{data points} \end{array}$$

To Do

Theory:

- soft clustering:

Experimentations:

- larger proteins
- 3d shapes processing (segmentation, symmetry detection)

To Do

Theory:

- soft clustering:

Experimentations:

- larger proteins
- 3d shapes processing (segmentation, symmetry detection)

Code:

experimental $\dashrightarrow$ released

experimental $\dashrightarrow$ online clustering service

Environment

Geometrica group:

- 4 perm. research. (geom. / topo.)
- 1 invited prof. (stats)
- 3 PhD (topo. / apprentissage)
- 3 post-docs

Inria Saclay

Digiteo Bldg.

Polytechnique Campus

Contact information:

Steve Oudot (`steve.oudot@inria.fr`)

Frédéric Chazal (`frederic.chazal@inria.fr`)