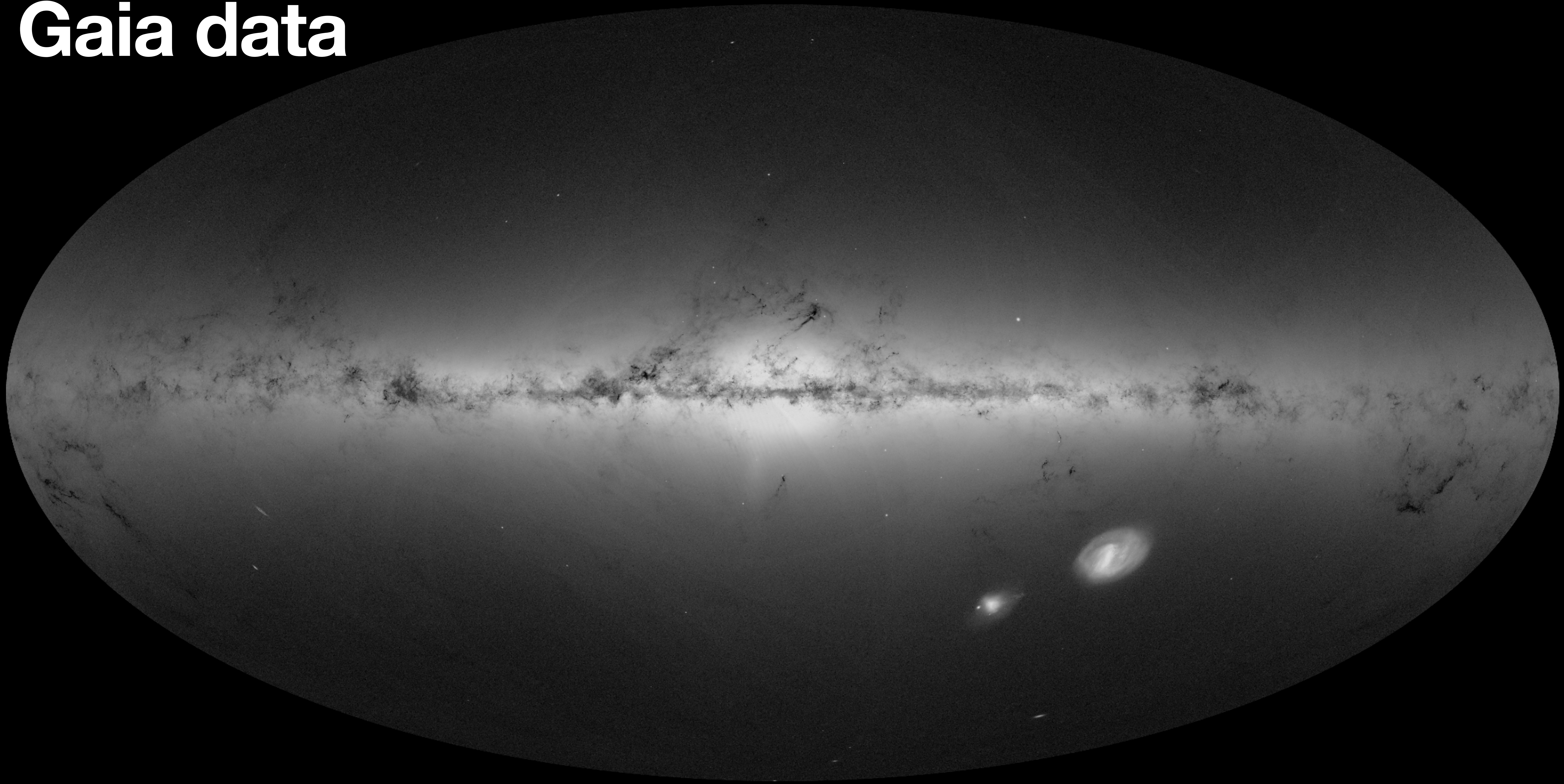
Significance Mode Analysis for hierarchical structures

Extracting stellar populations from large-scale surveys

Sebastian Ratzenböck @ Uni Vienna

ML-IAP/CCA-2023

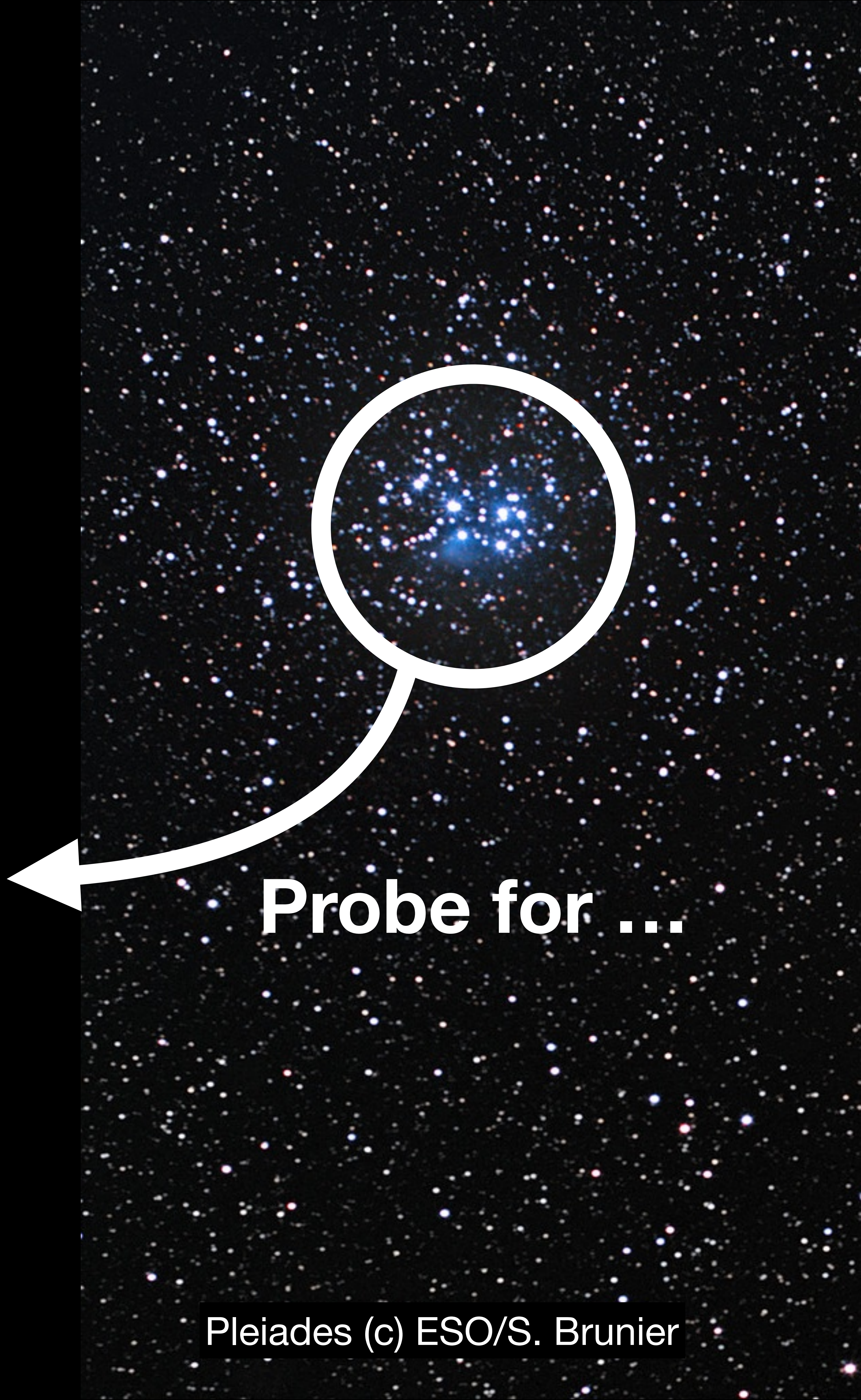
Gaia data



Stellar populations

Born from same molecular cloud

- Thought to be birthplace of most stars
  (Lada & Lada 2003; Parker & Goodwin 2007)
- Structure formation and evolution
- Chemical composition of Milky Way
- Exoplanet formation and evolution
- Stellar initial mass function



Pleiades (c) ESO/S. Brunier


M80 (c) Hubble Heritage Team


Pleiades (c) ESO/S. Brunier


ESA Gaia, DPAC - Moitinho+2017

Identifying stellar populations

Problem definition

- Low dimensional feature space

3 positional axes + 2 tangential velocities

Stars that move together were born together
(Kamdar+2019)

Identifying stellar populations

Problem definition

- Low dimensional feature space
- Projection effects in velocities

Identifying stellar populations

Problem definition

- Low dimensional feature space
- Projection effects in velocities
- Millions to billions of data points

Identifying stellar populations
Problem definition

- Low dimensional feature space
- Projection effects in velocities
- Millions to billions of data points
- 95 — 99% noise

Identifying stellar populations

Problem definition

- Low dimensional feature space
- Projection effects in velocities
- Millions to billions of data points
- 95 — 99% noise
- Wide variety of (non-convex) cluster morphologies

Tidal tails (Meingast+2019a), **Streams** (Meingast+2019b), **Strings** (Kounkel+2019), **Rings** (Cantat-Gaudin+2019), **Snakes** (Wang+2021), **Pearls** (Coronado+2021), ...

Identifying stellar populations

Problem definition

- Low dimensional feature space
- Projection effects in velocities
- Millions to billions of data points
- 95 — 99% noise
- Wide variety of (non-convex) cluster morphologies
- No accurate simulations / forward models

Identifying stellar populations
Problem definition

- Low dimensional feature space
- Projection effects in velocities
- Millions to billions of data points
- 95 — 99% noise
- Wide variety of (non-convex) cluster morphologies
- No accurate simulations / forward models

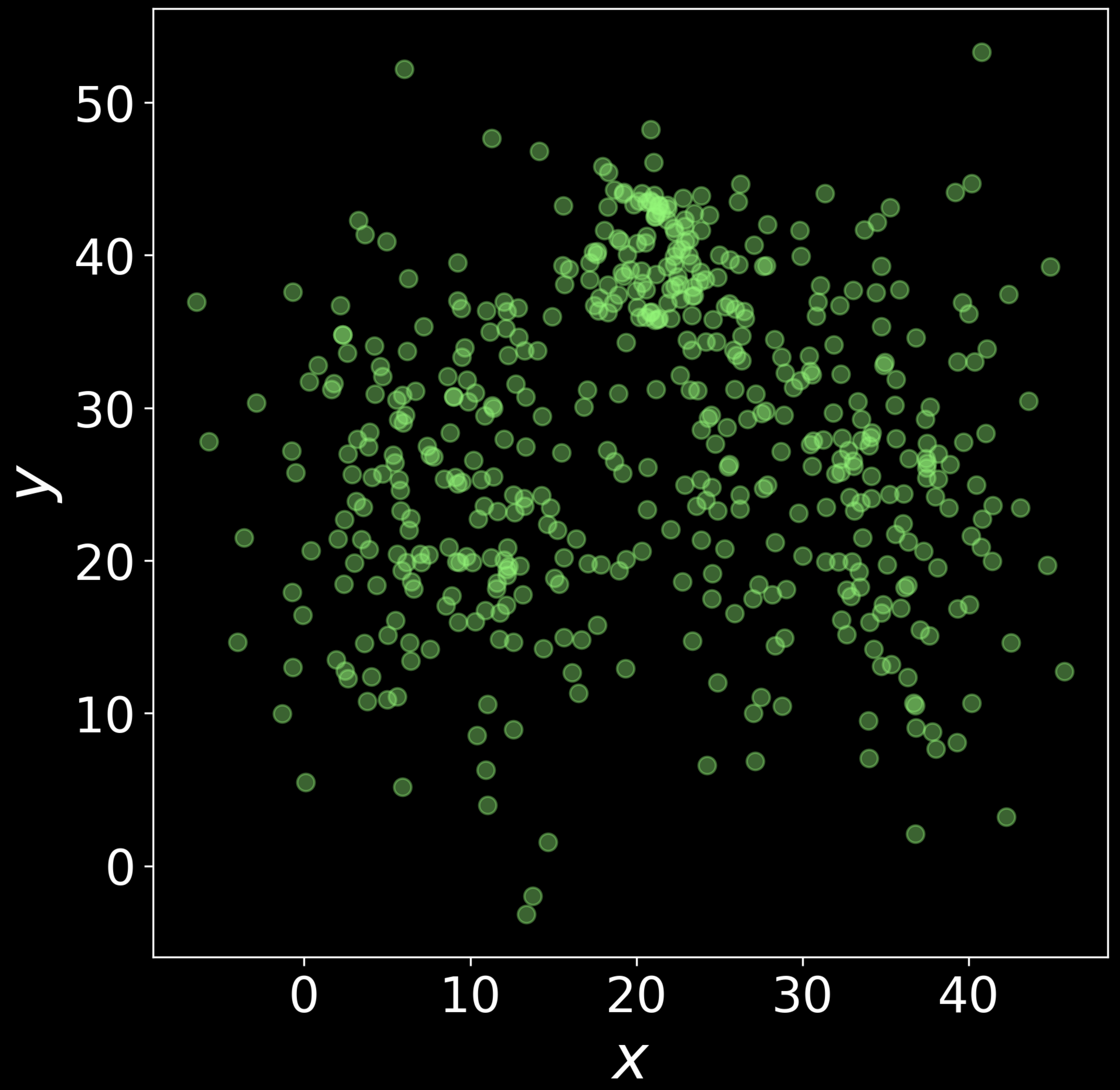
➡ **Nonparametric, density based clustering**

Recap: Density based clustering

Nonparametric, density-based clustering

Problem definition

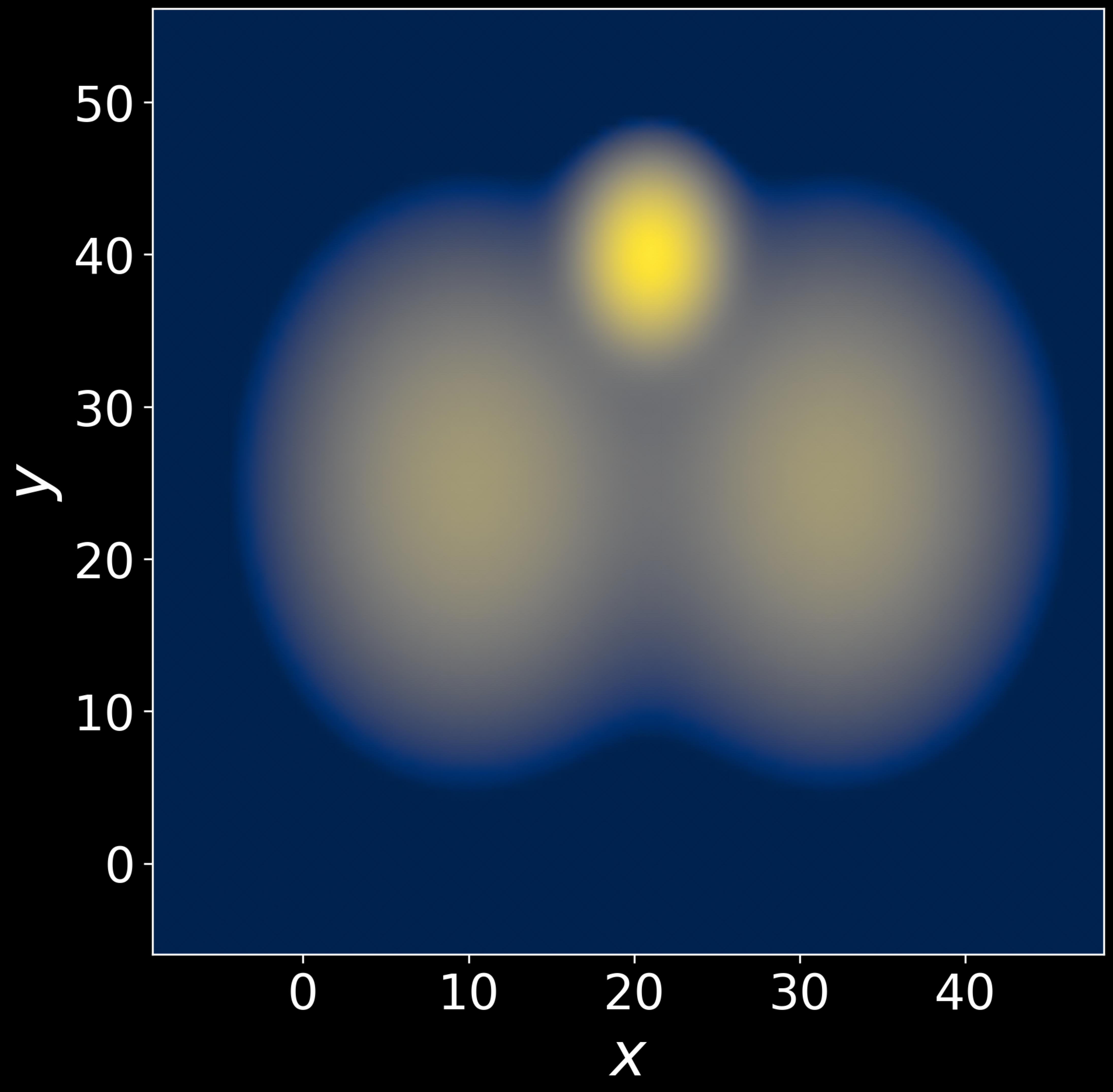
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$



Nonparametric, density-based clustering

Problem definition

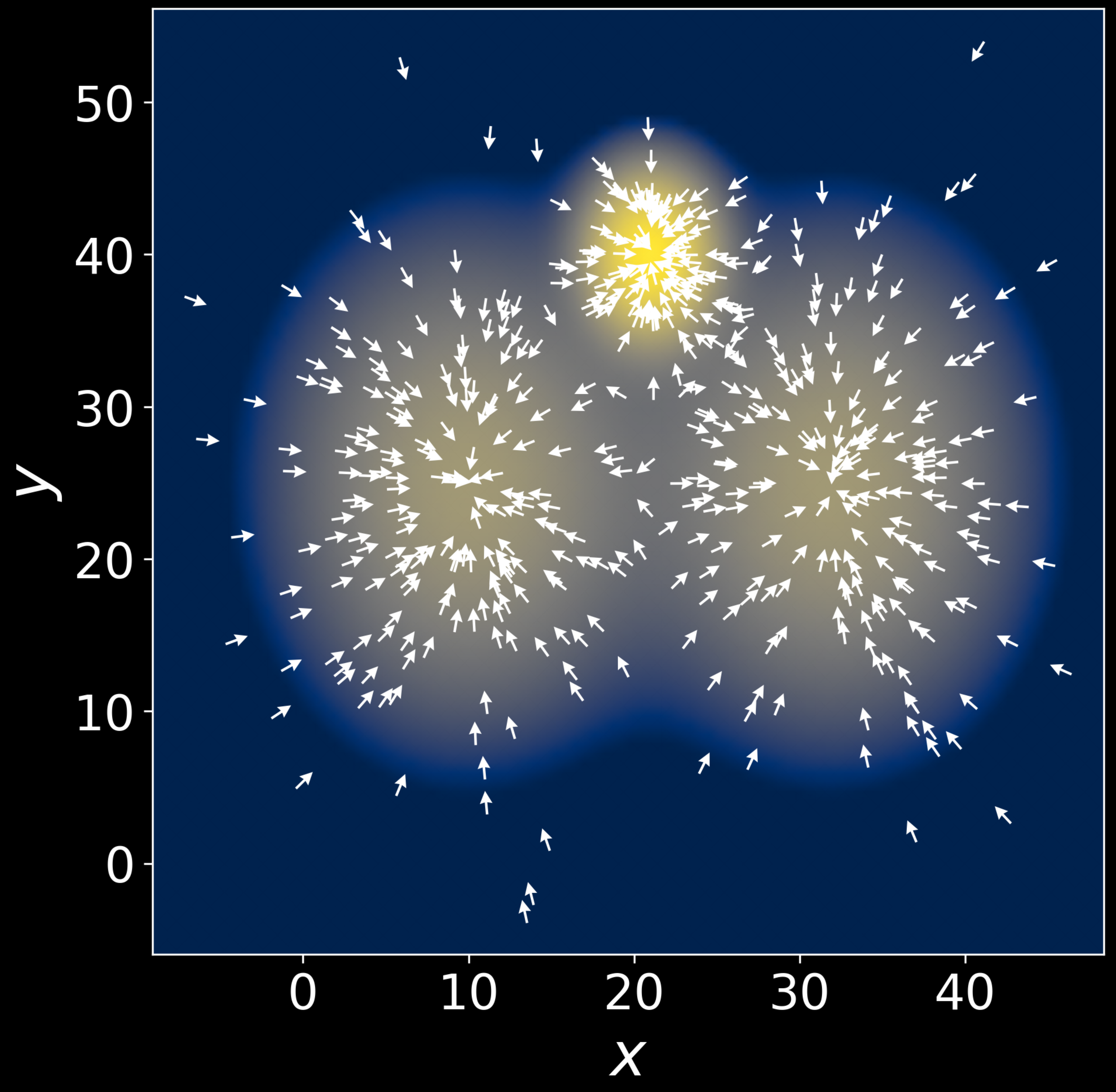
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$



Nonparametric, density-based clustering

Problem definition

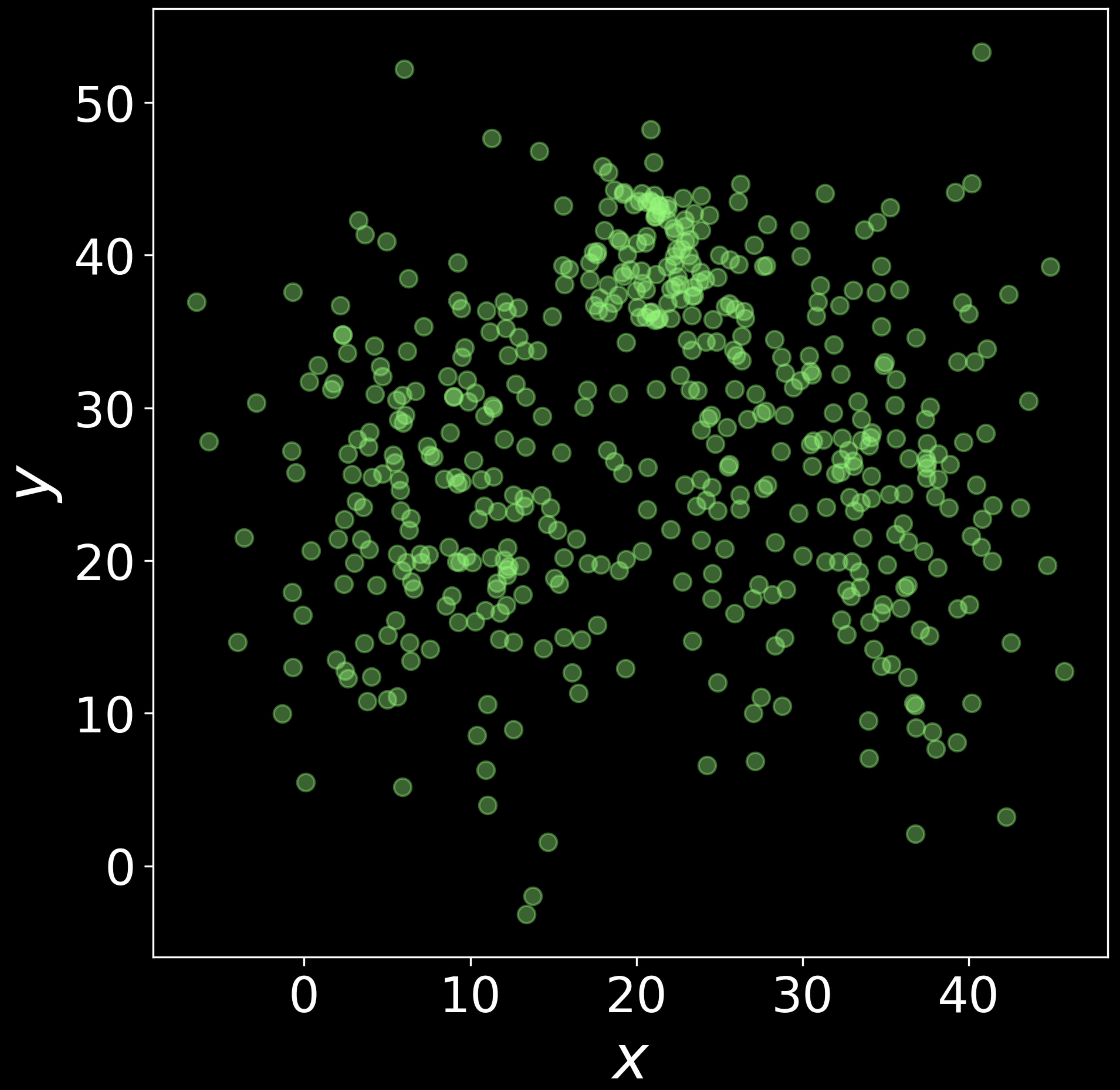
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

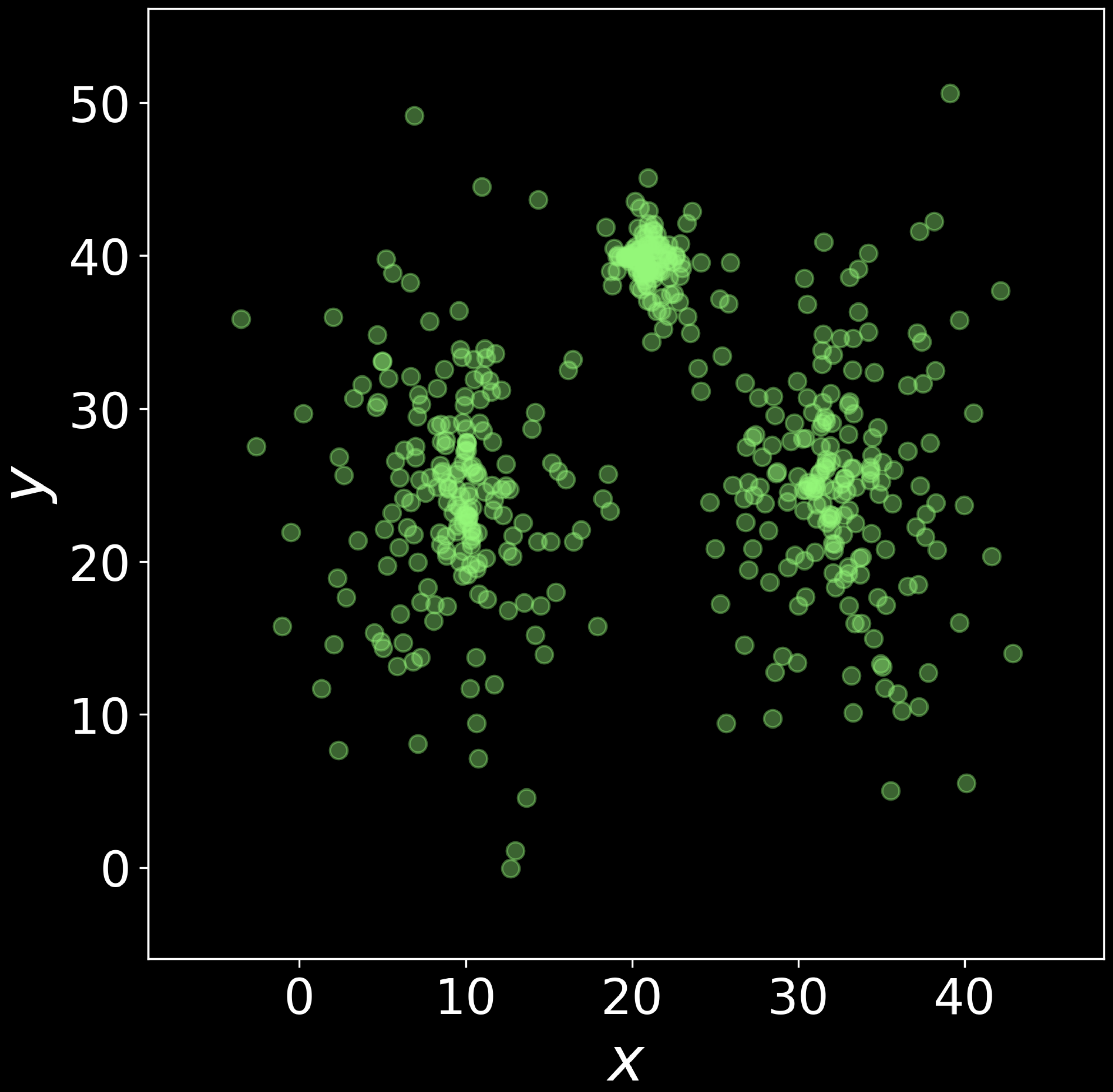
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

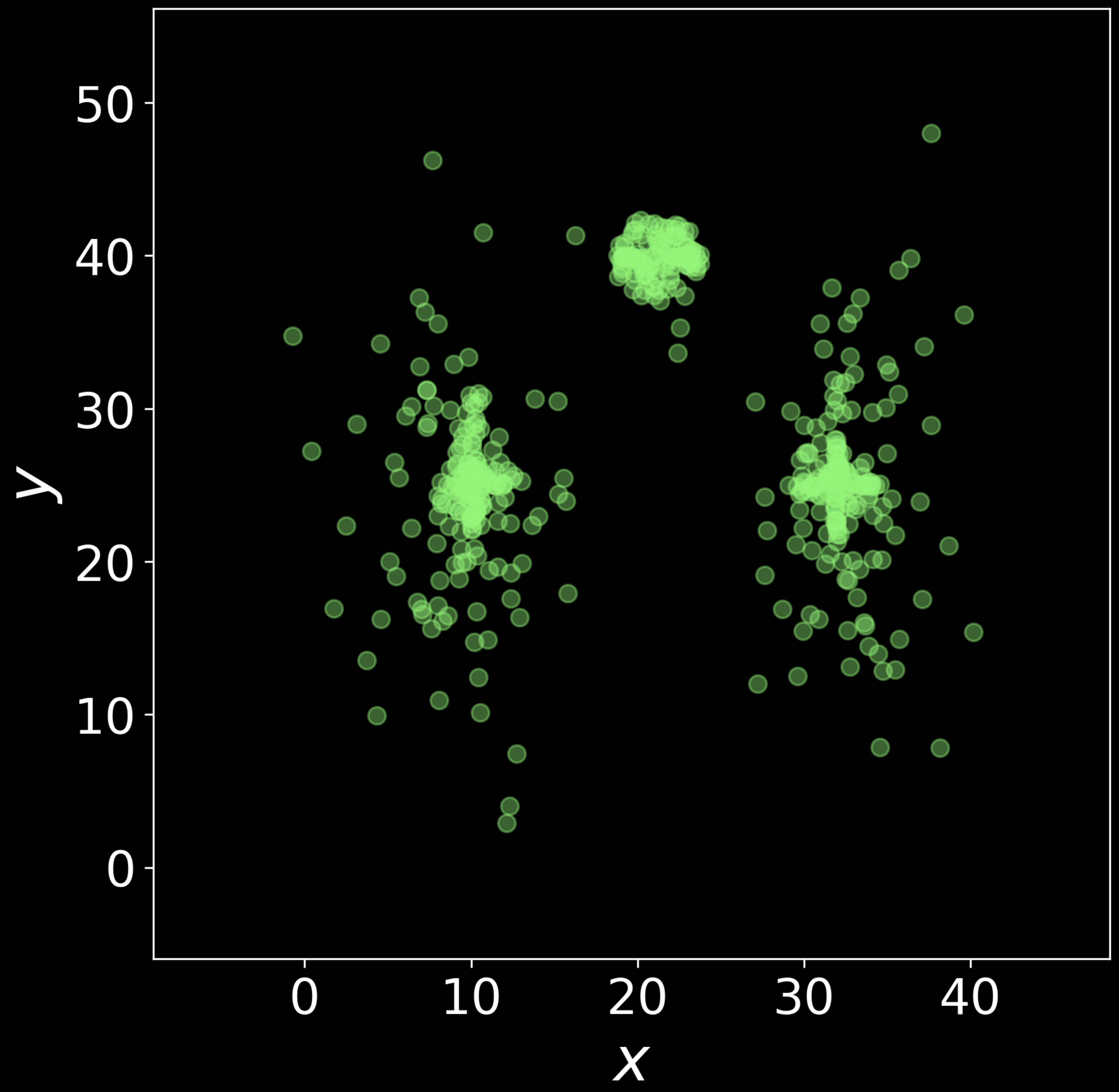
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

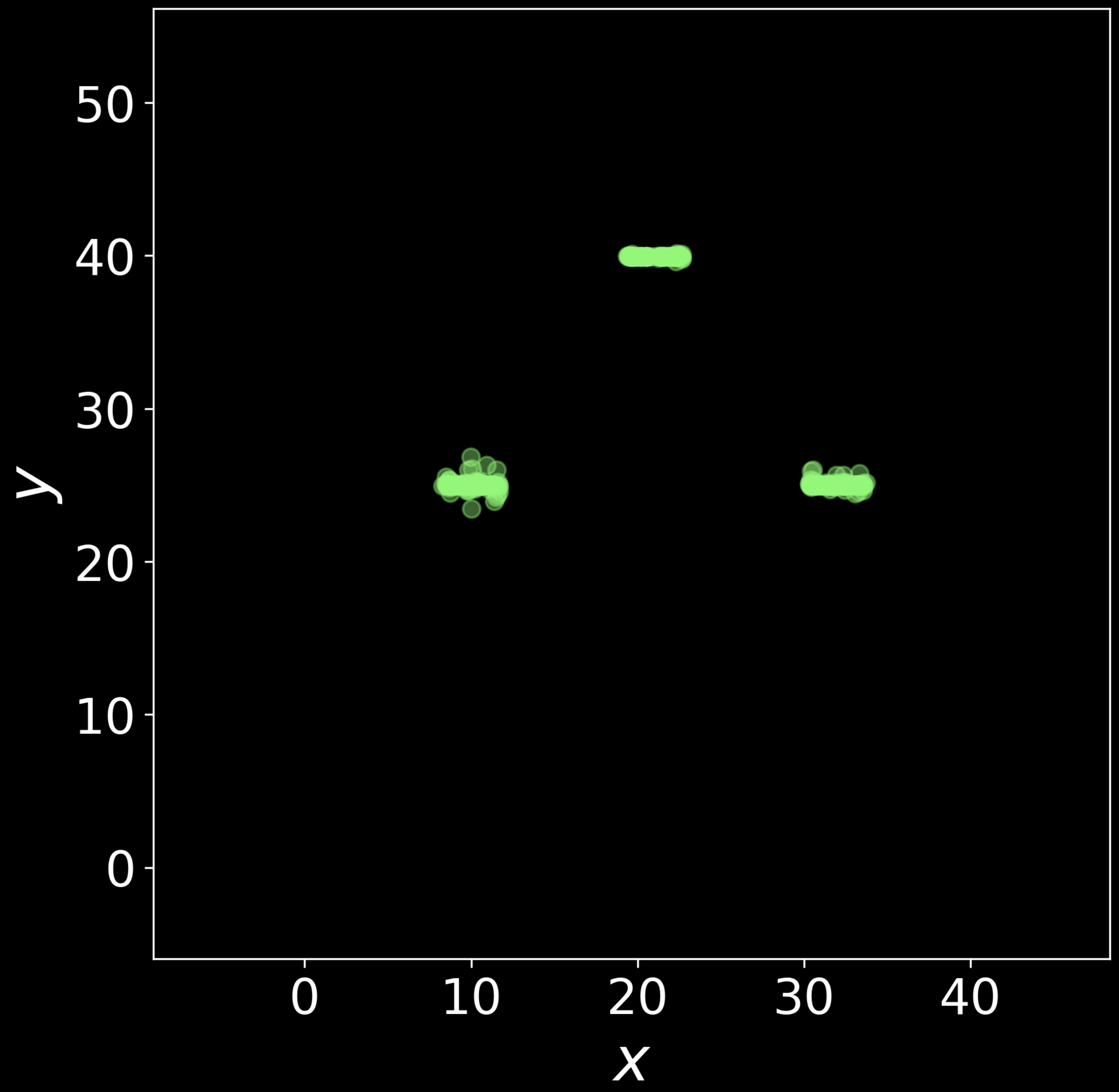
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

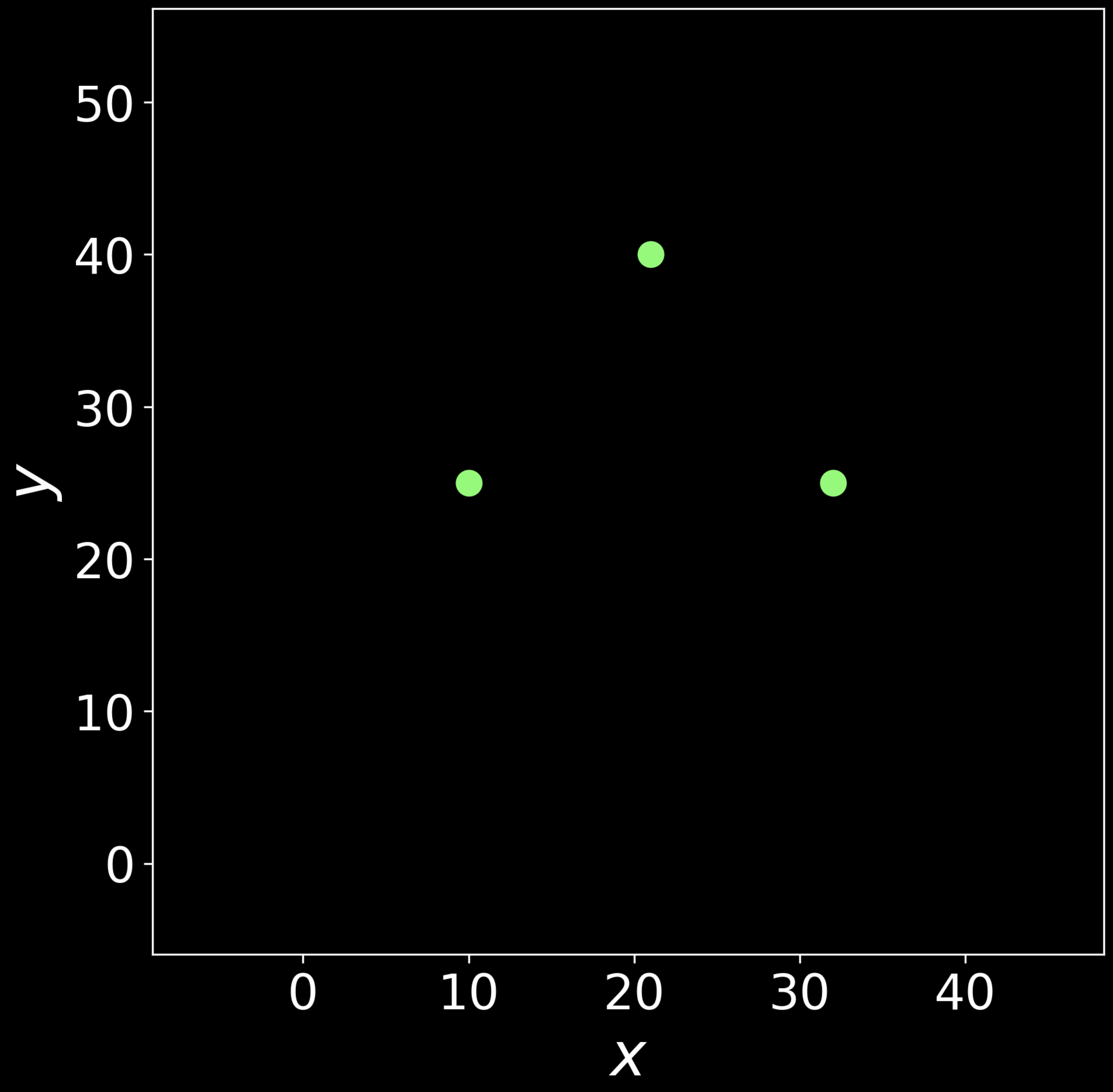
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

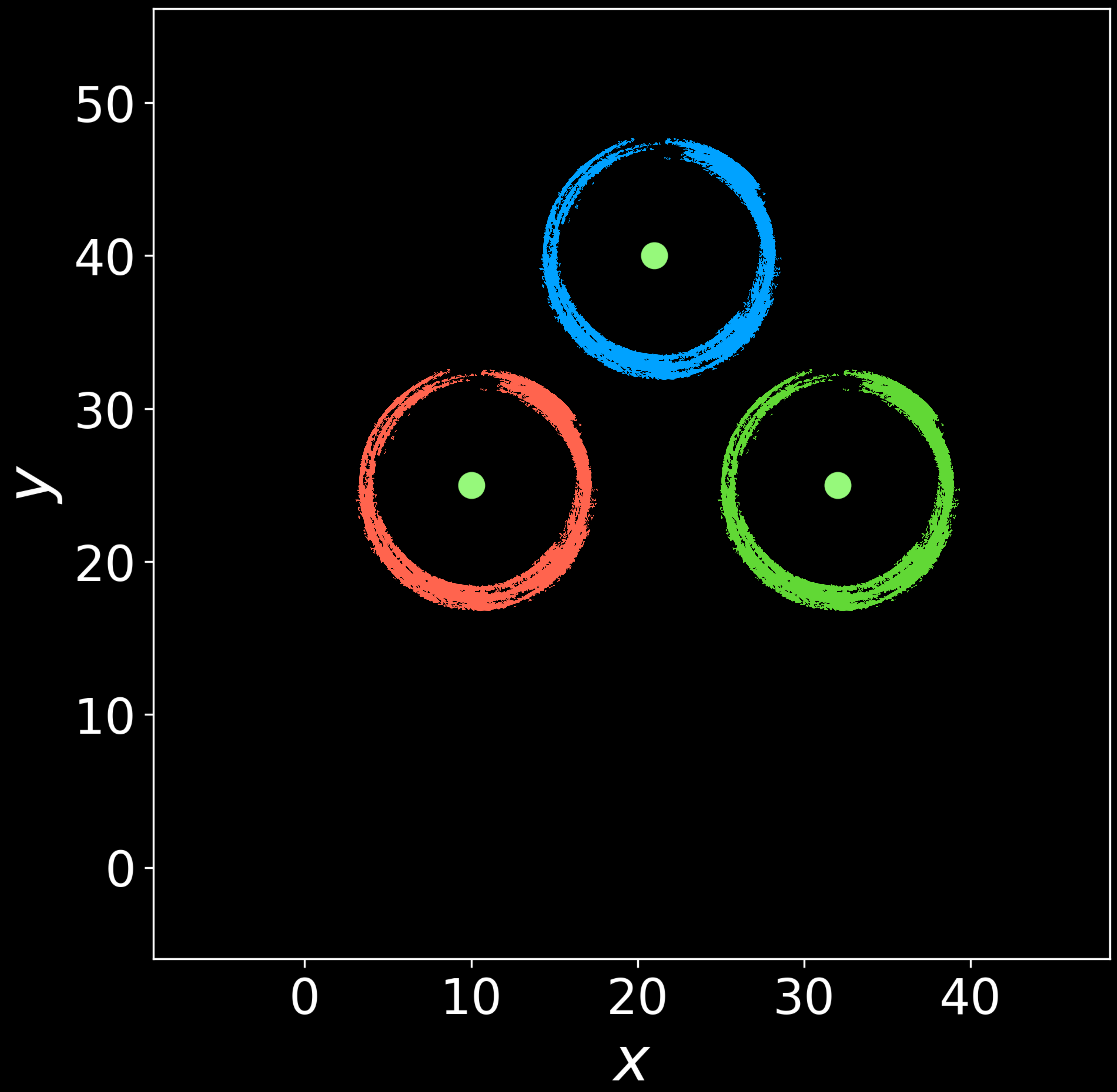
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

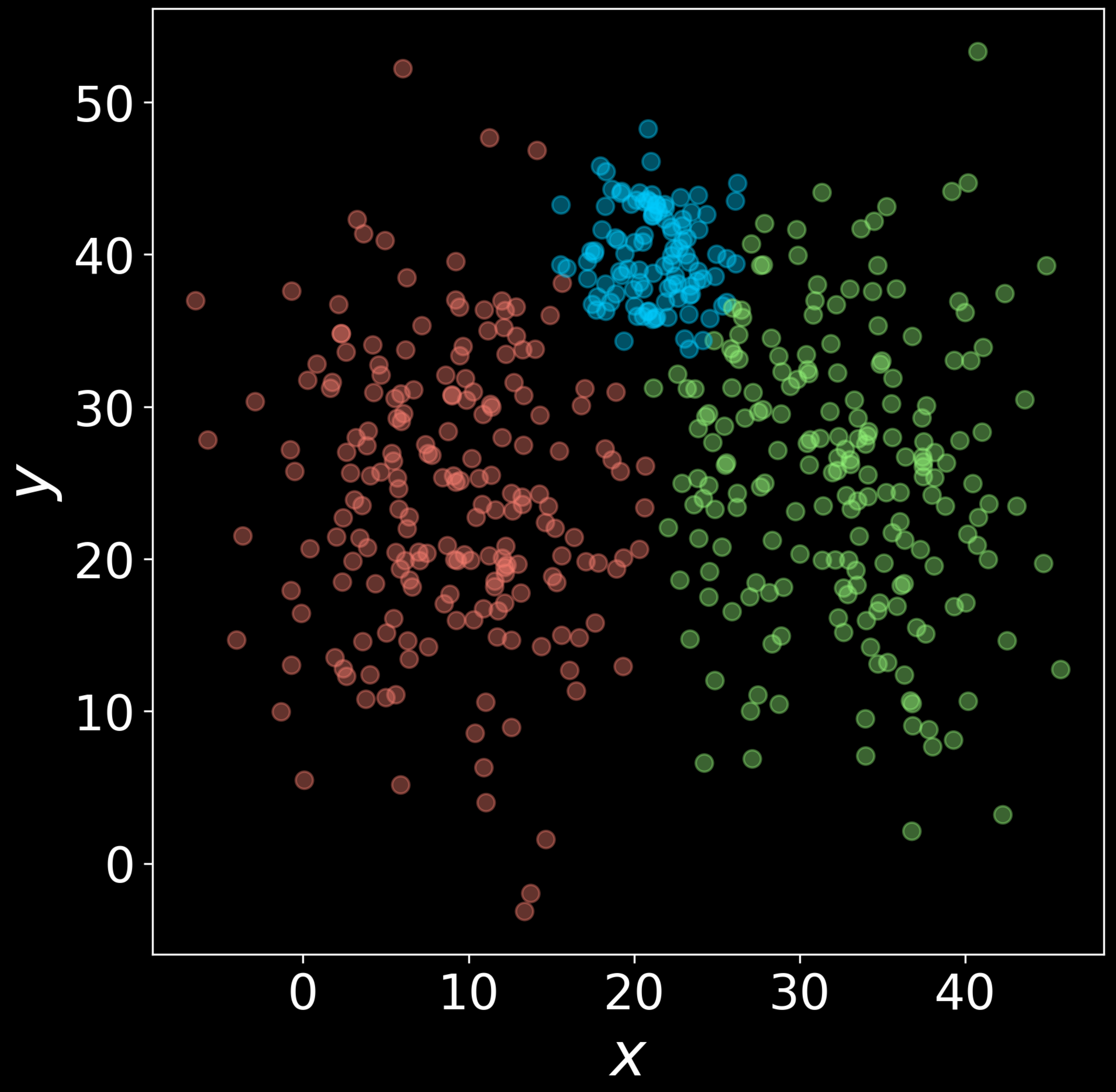
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

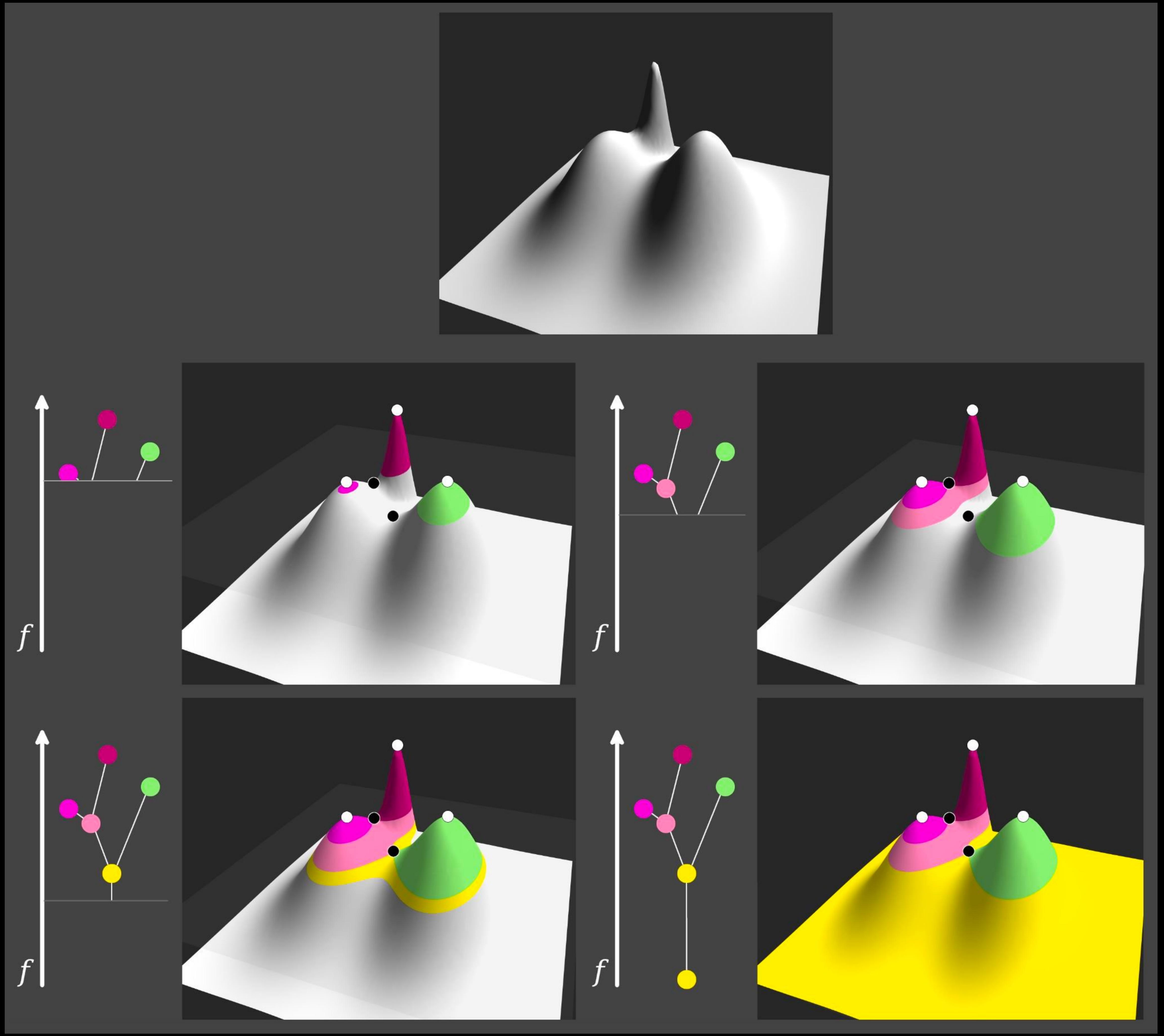
- Data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, x_i \in \mathbb{R}^p$
- Data generated from density: $X \sim f$
- **Wishart (1969) cluster definition**
  - $\mathbf{x}_i$ associated with modes of f
  - Propagate $\mathbf{x}_i$ along ∇f



Nonparametric, density-based clustering

Problem definition

- Level set: $L(\lambda) = \{f(\mathbf{x}) \geq \lambda\}$
- **Hartigan (1975) cluster definition**
  - Connected components of $L(\lambda)$
  - Cluster tree: vary λ: $\infty \rightarrow -\infty$

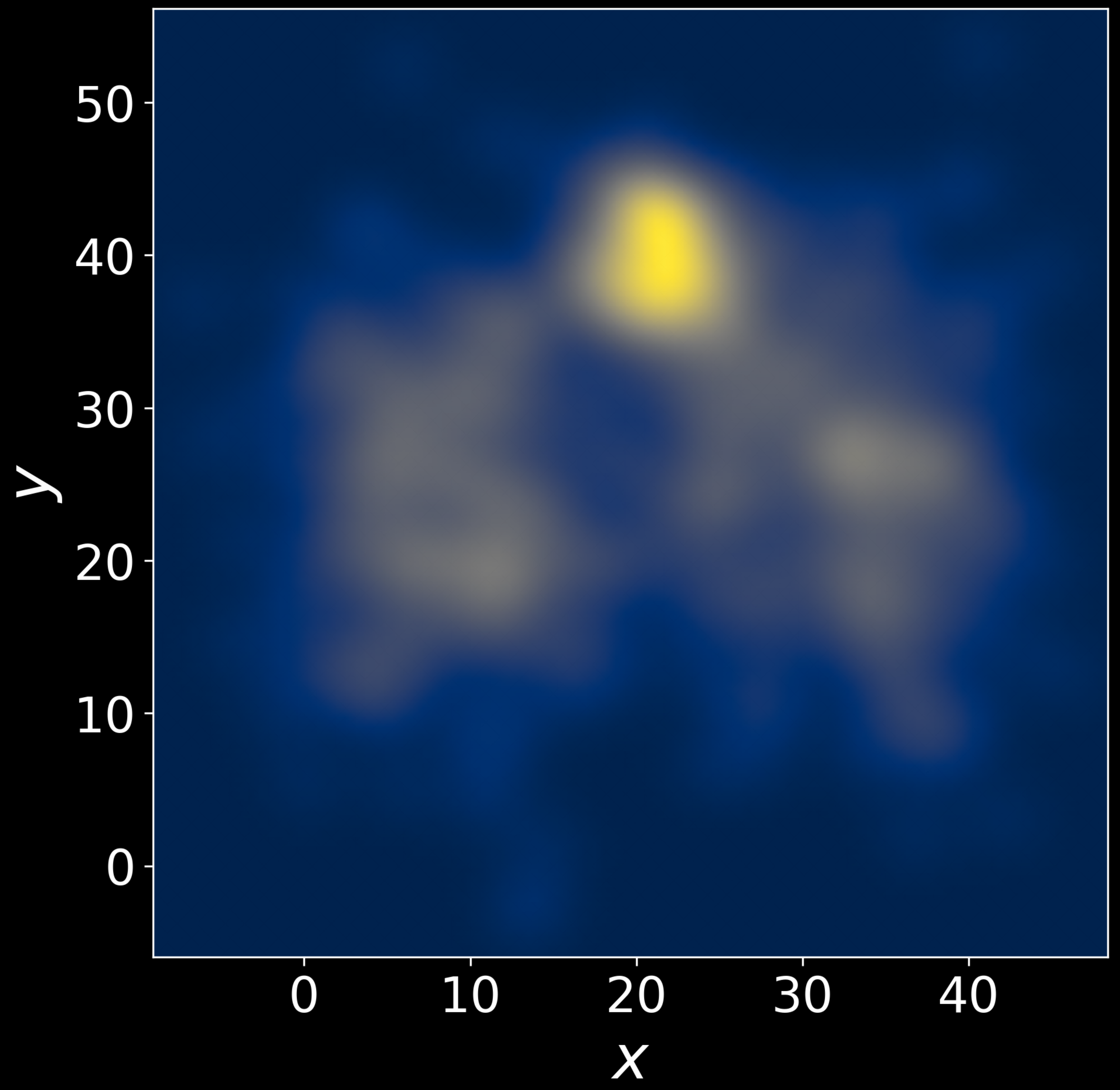


Reality:
Estimate density from data

Reality: Estimate density from data

- Estimate density $\hat{f}$ from data X

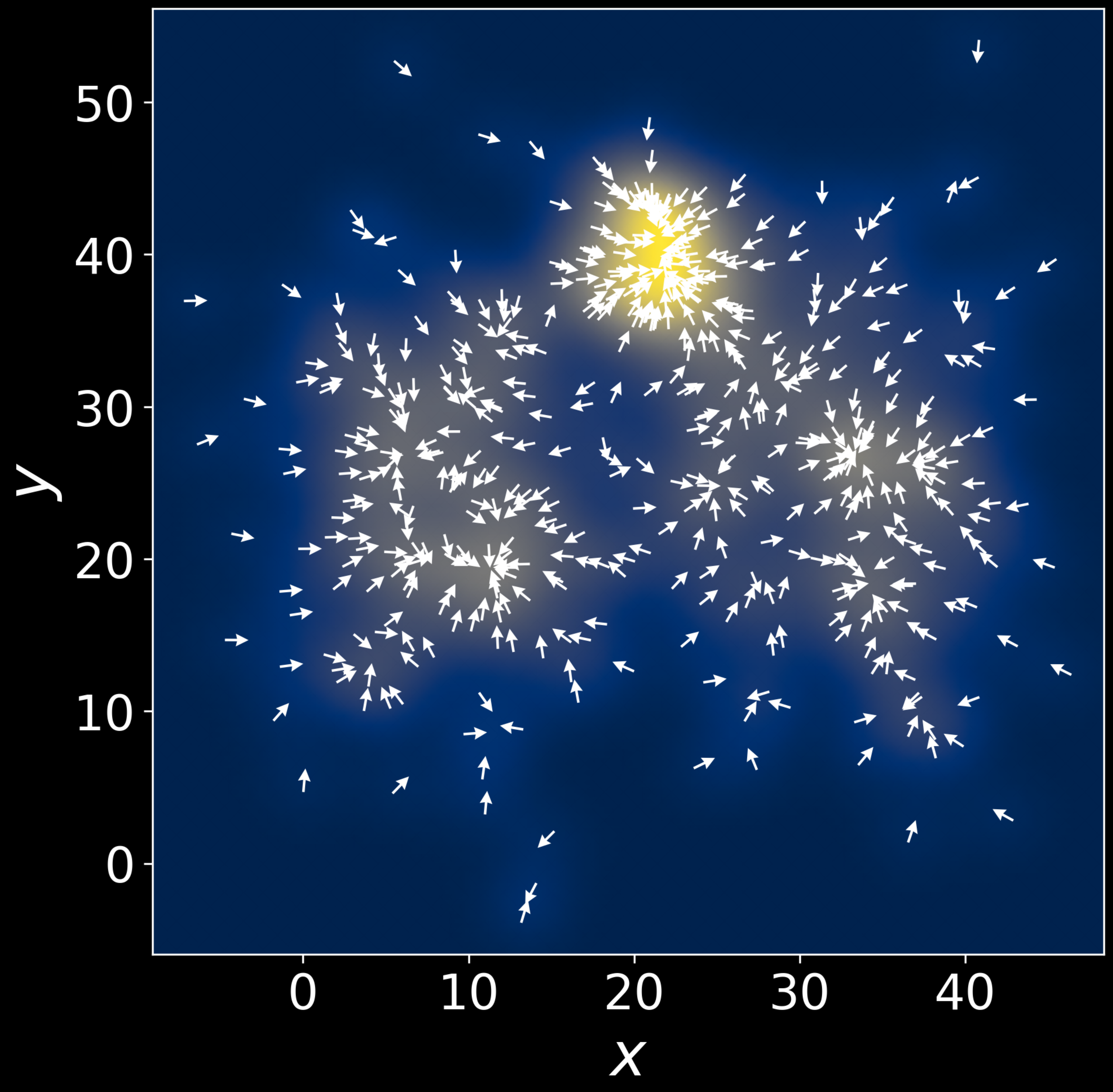
➡ produces spurious clusters



Reality: Estimate density from data

- Estimate density $\hat{f}$ from data X

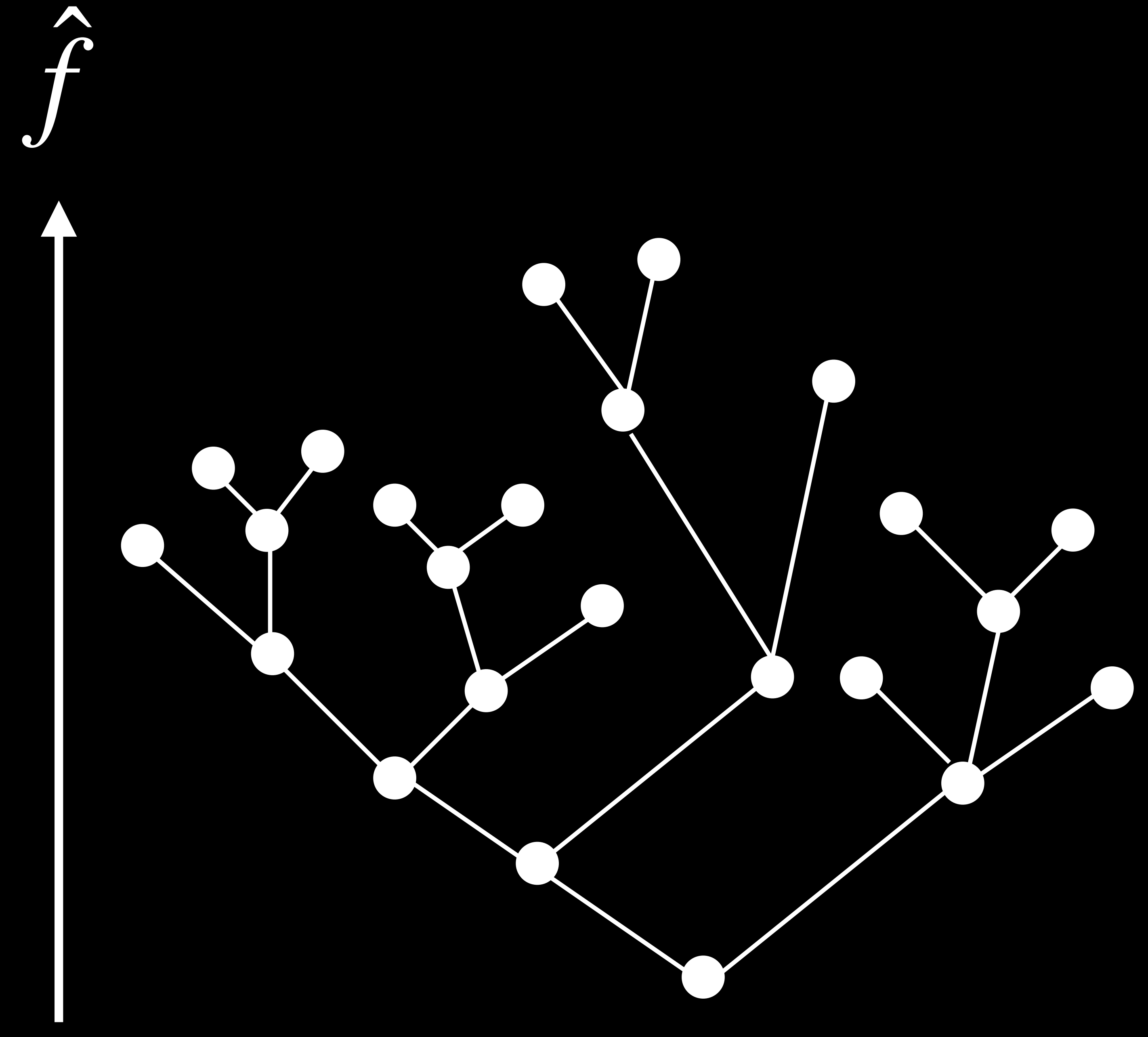
➡ produces spurious clusters

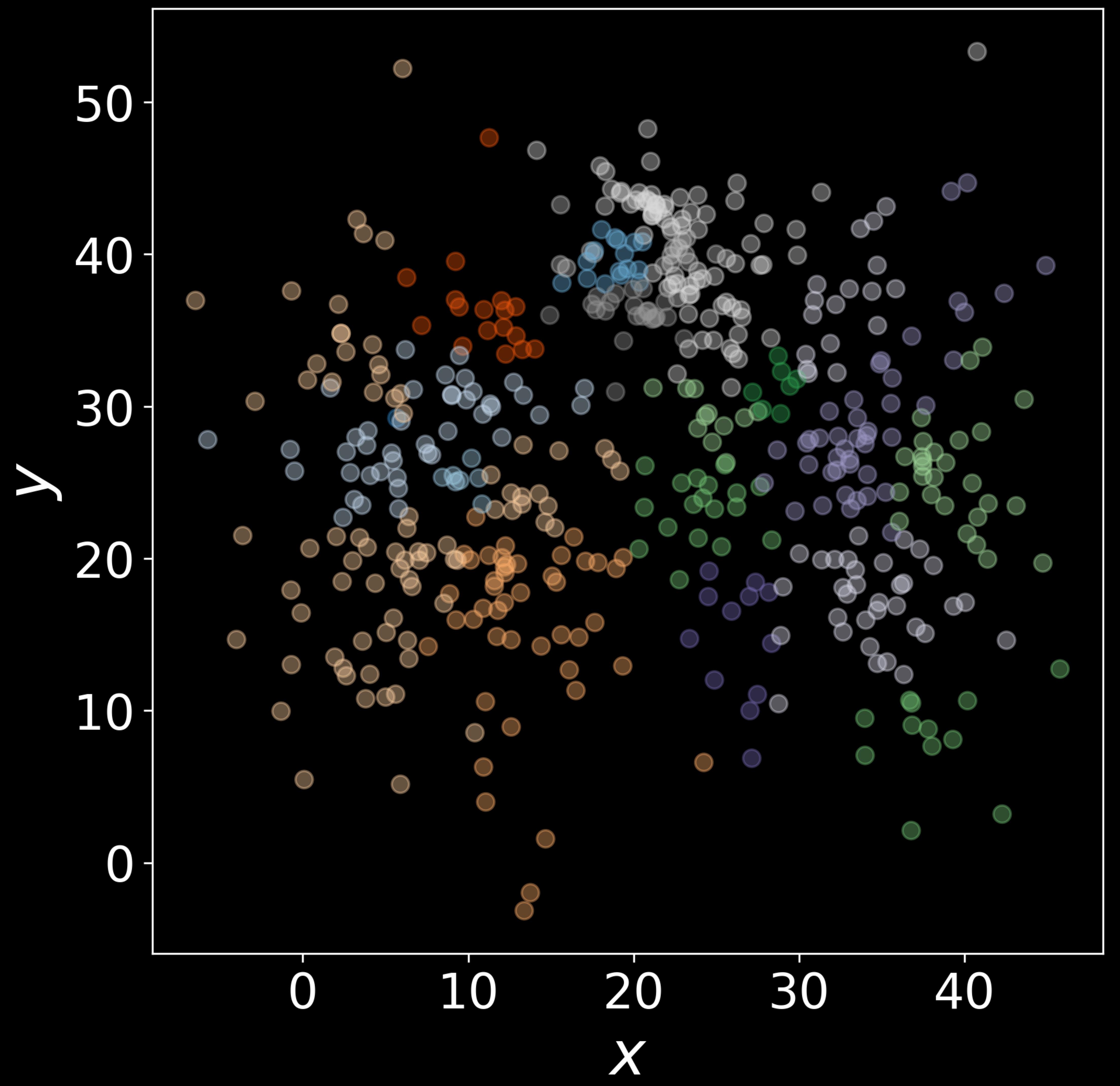


Reality: Estimate density from data

- Estimate density $\hat{f}$ from data X

➡ produces spurious clusters

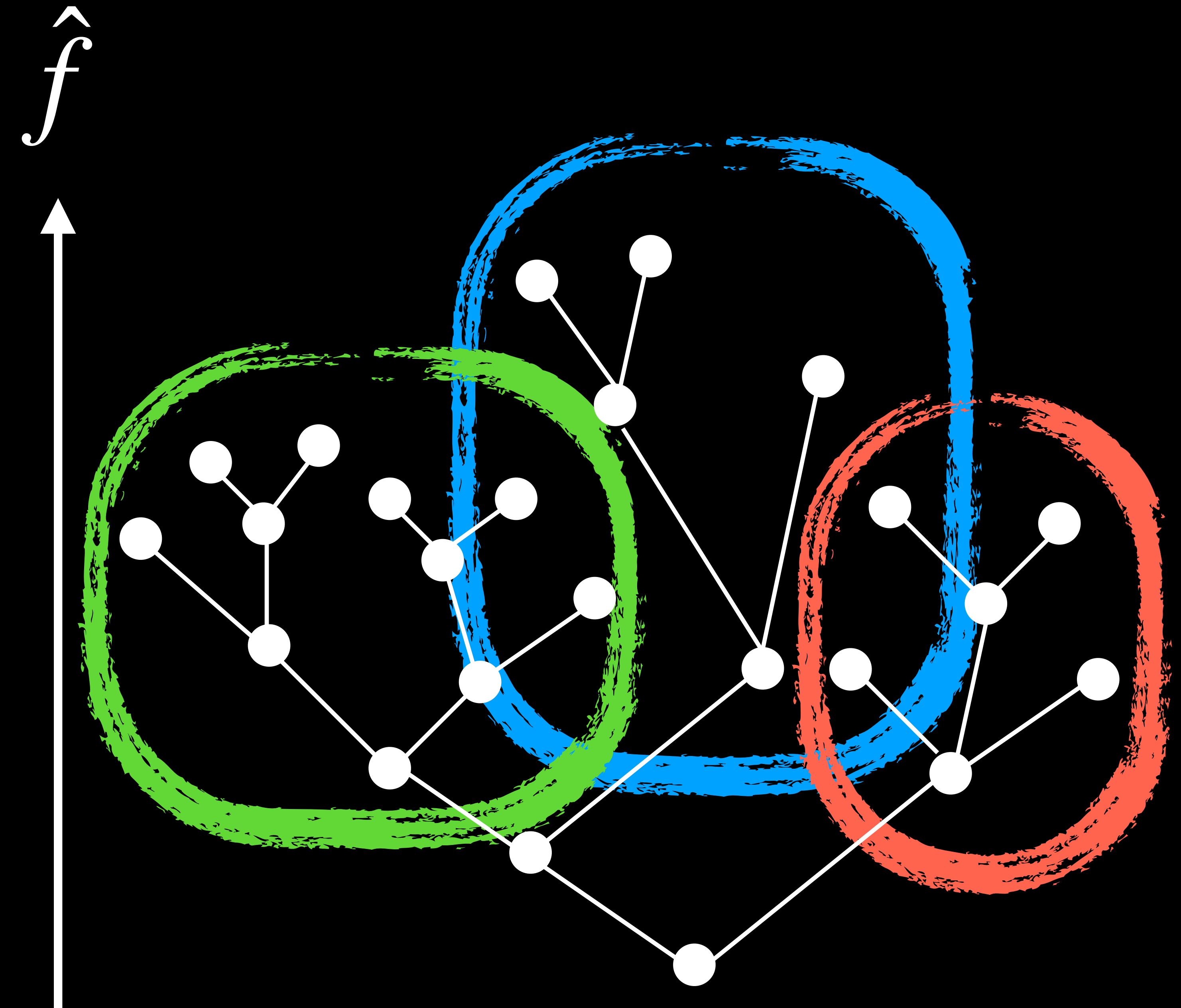


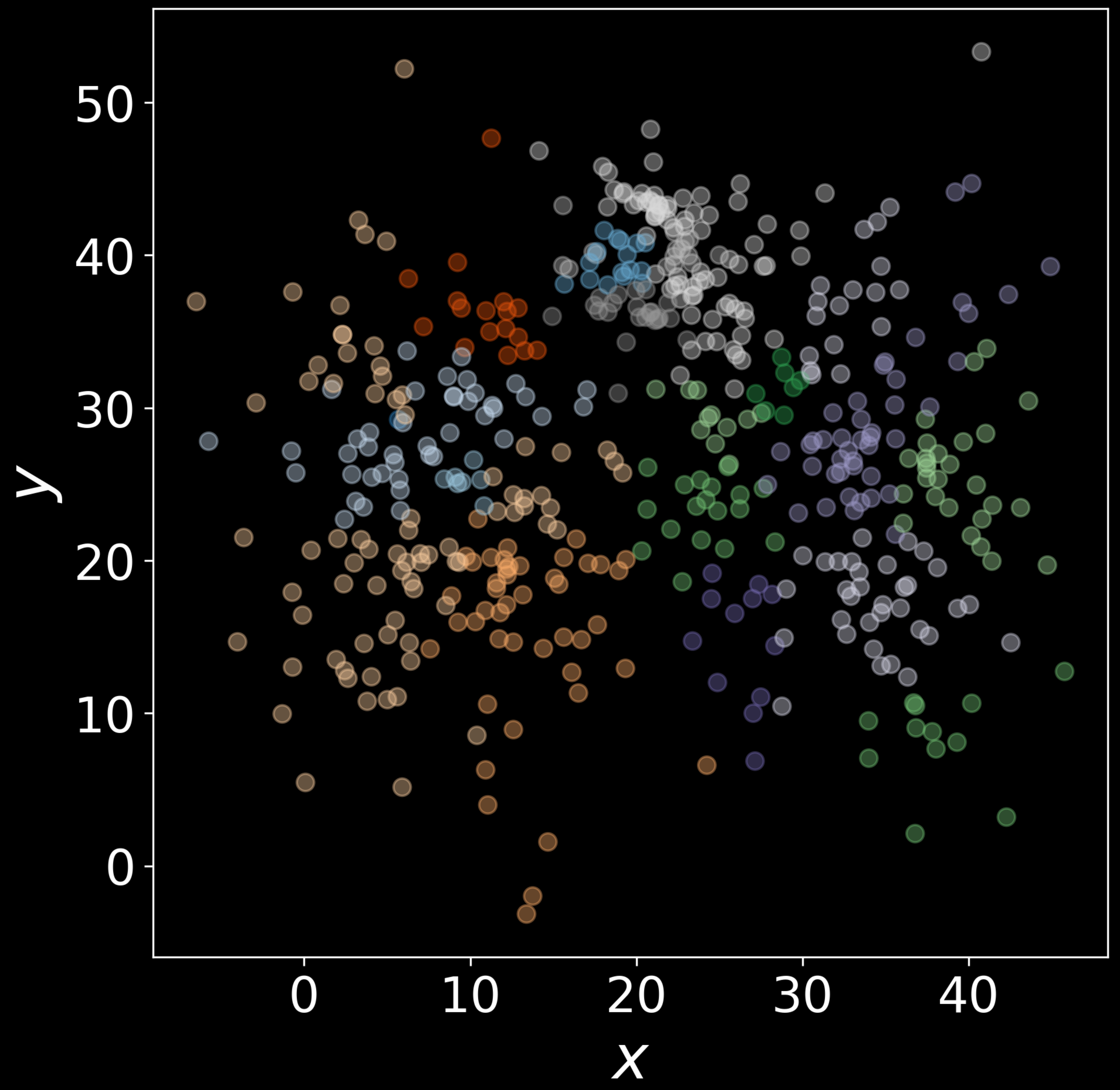


Reality: Estimate density from data

- Estimate density $\hat{f}$ from data X

➡ produces spurious clusters

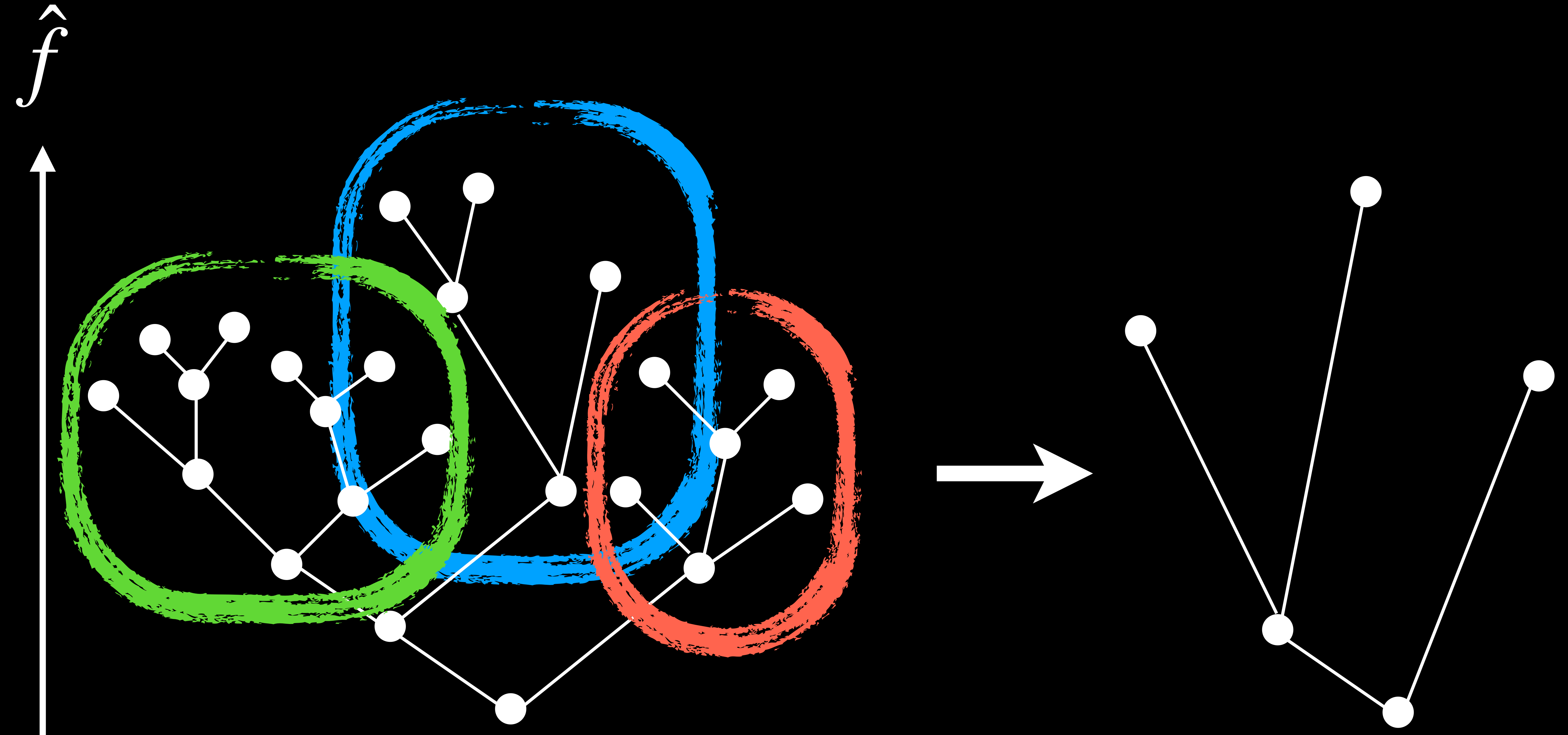


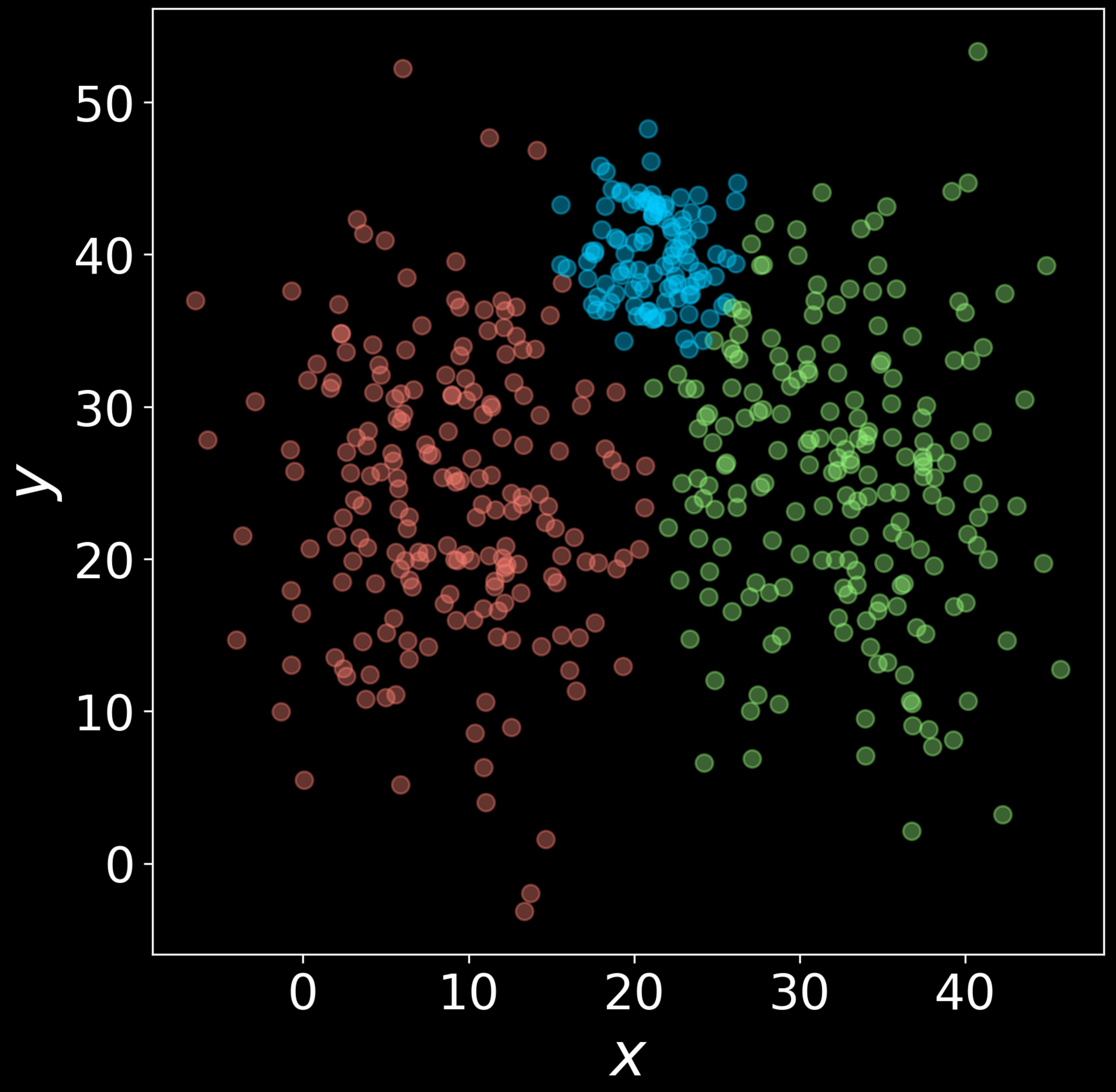


Reality: Estimate density from data

- Estimate density $\hat{f}$ from data X

➡ produces spurious clusters

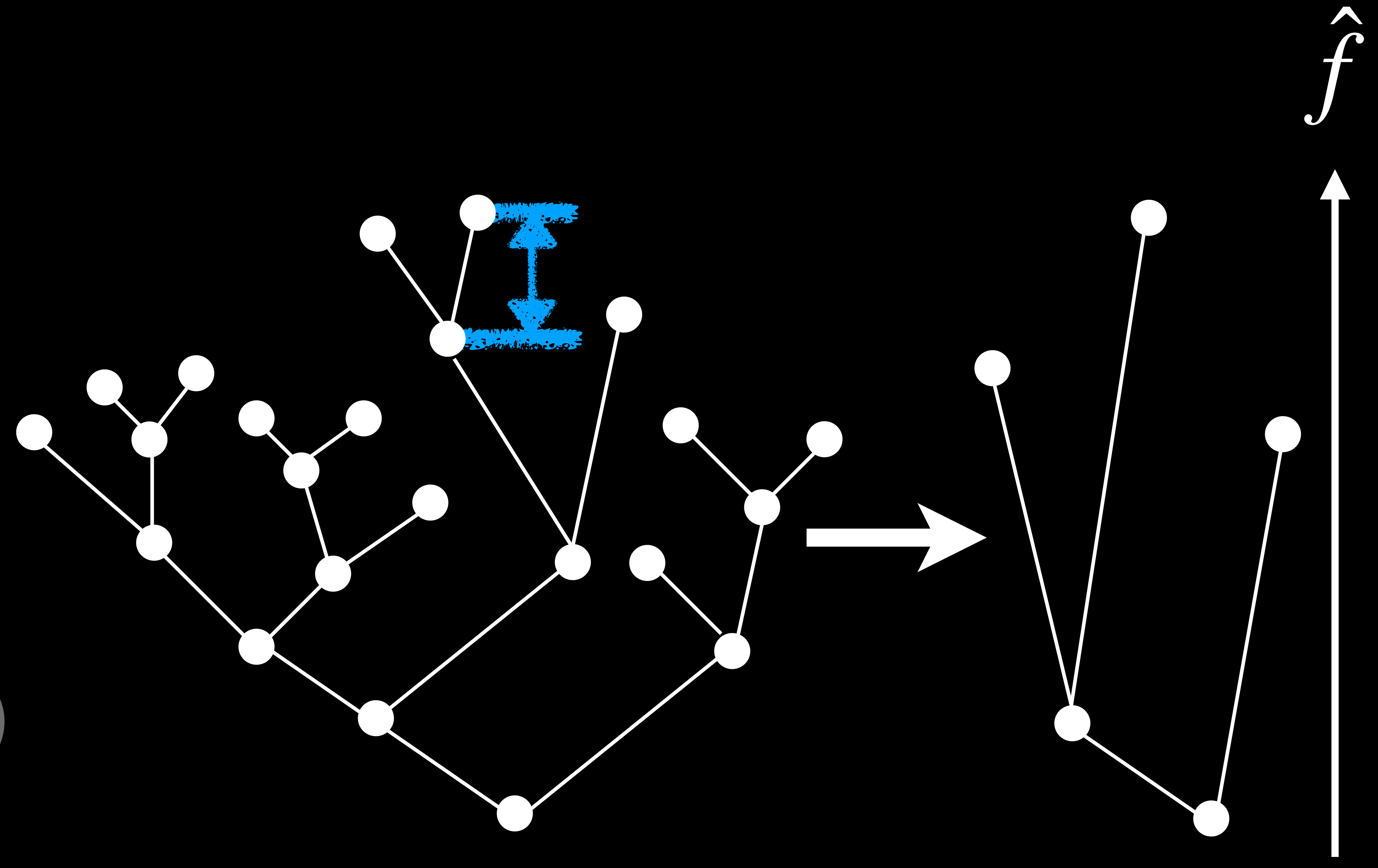




Pruning cluster tree

Current strategies

- Density difference $\Delta\hat{f}$
  (Chazal+2013)
- Normalised $\Delta\hat{f}$
  (Ding+2016)
- Distance based
  (Stuetzle+2010; Kpotufe+2011; Chaudhuri+2014)
- Relative excess of mass
  (HDBSCAN; Campello+2013)



Pruning cluster tree

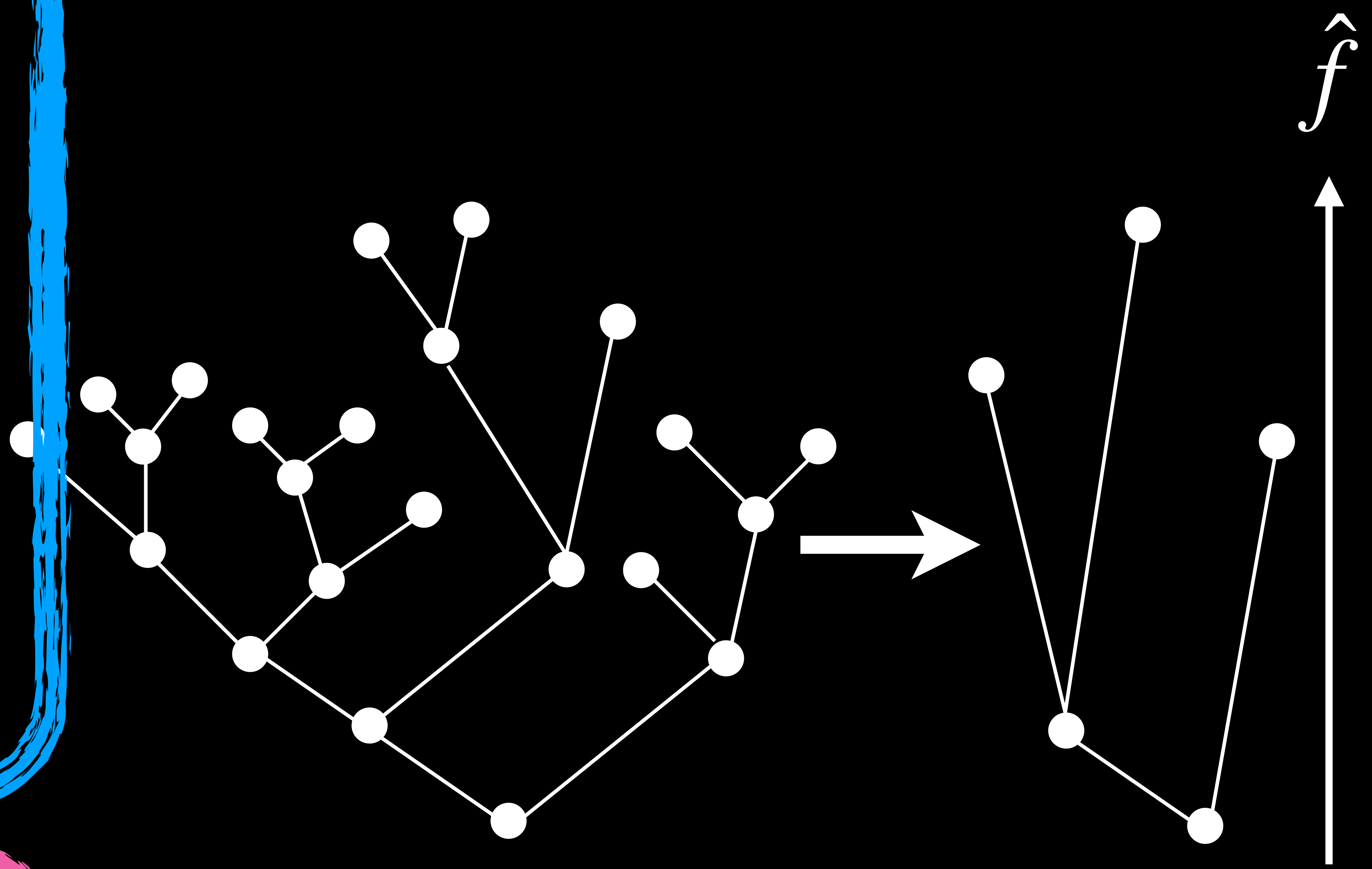
Current strategies

- Density difference $\Delta\hat{f}$
  (Chazal+2013)
- Normalised $\Delta\hat{f}$
  (Ding+2016)
- Distance based
  (Stuetzle+2010; Kpotufe+2011; Chaudhuri+2014)

Hard to determine threshold for $N \gg 1$

- Relative excess of mass
  (HDBSCAN; Campello+2013)

Typically over-merges



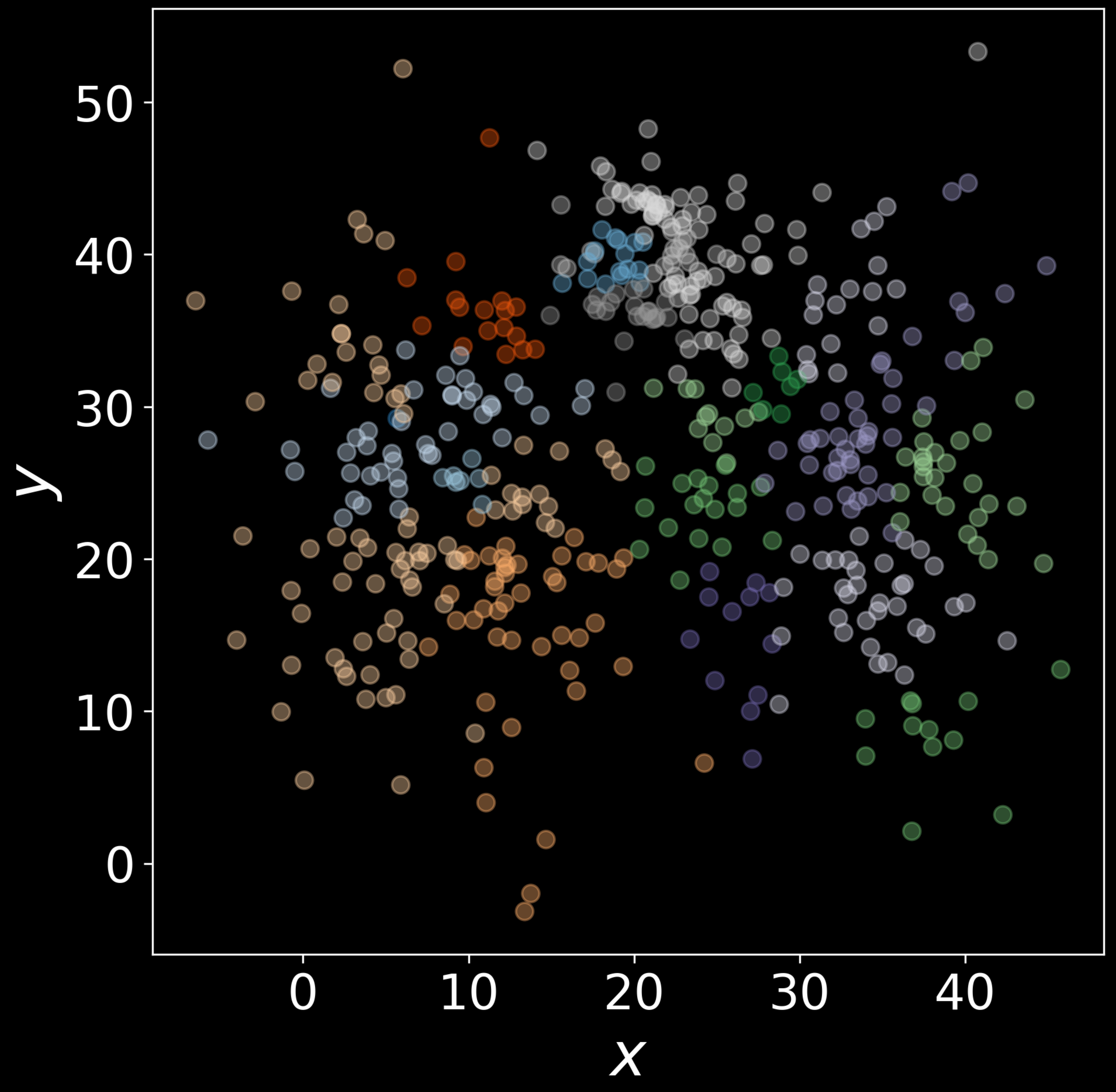
Going back to Wishart (1969)

Clusters are modes of f

What constitutes a cluster?

Clusters are **modal regions** of f

➡ **Test for multimodality**



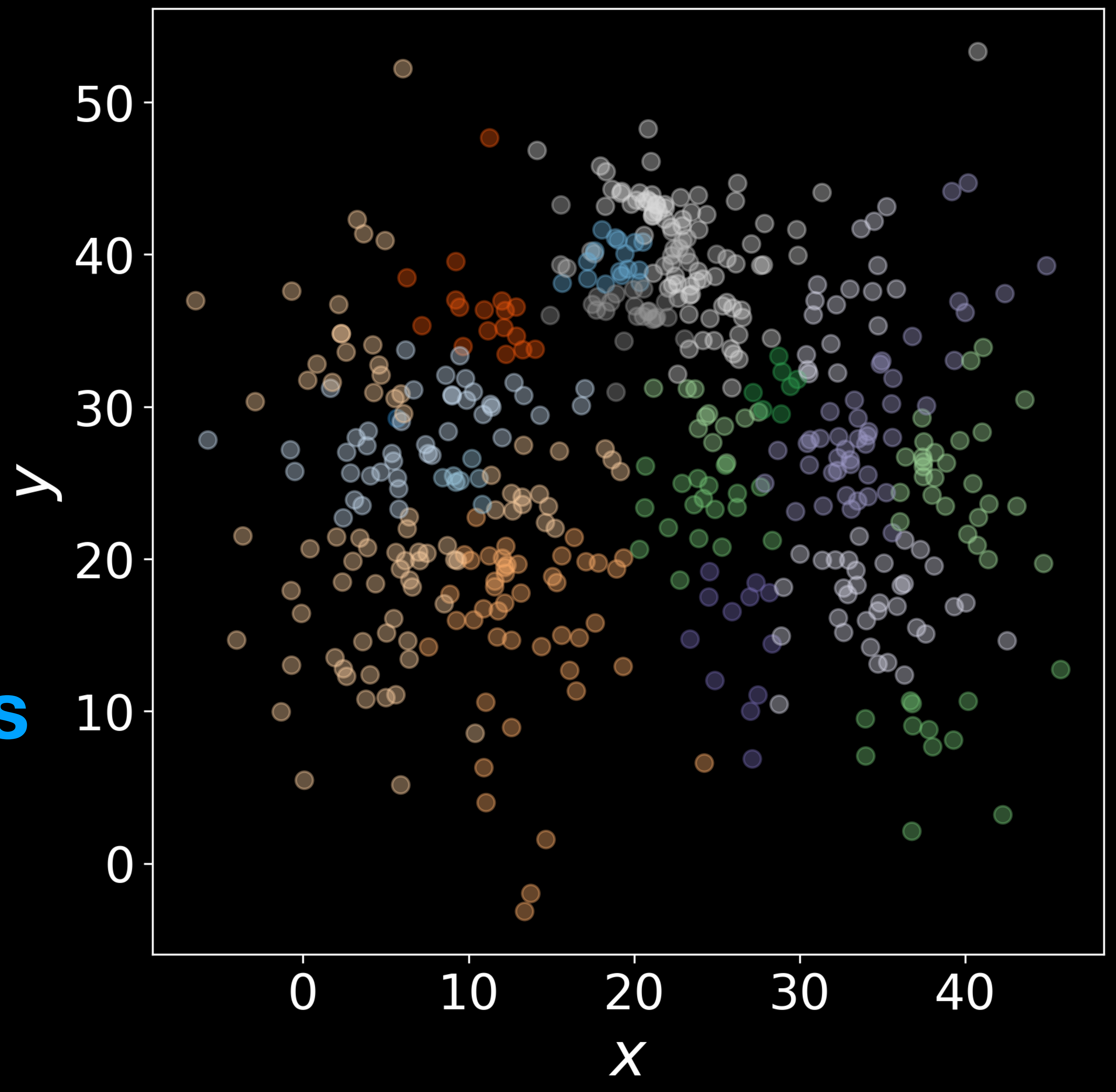
What constitutes a cluster?

Clusters are **modal regions** of f

➡ **Test for multimodality**

H_0: Points belong to **single mode**

H_1: Points belong to **multiple modes**



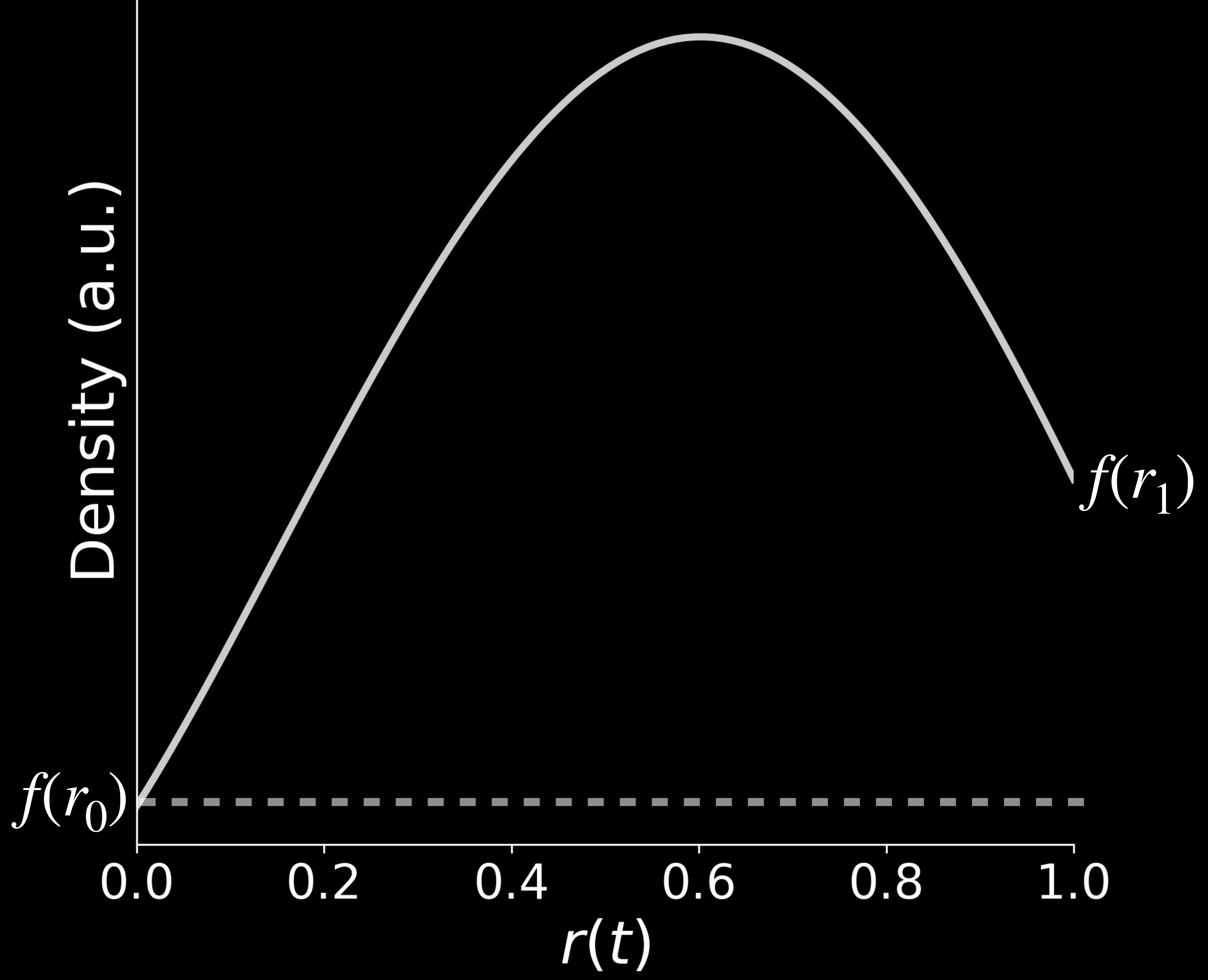
Modality along paths

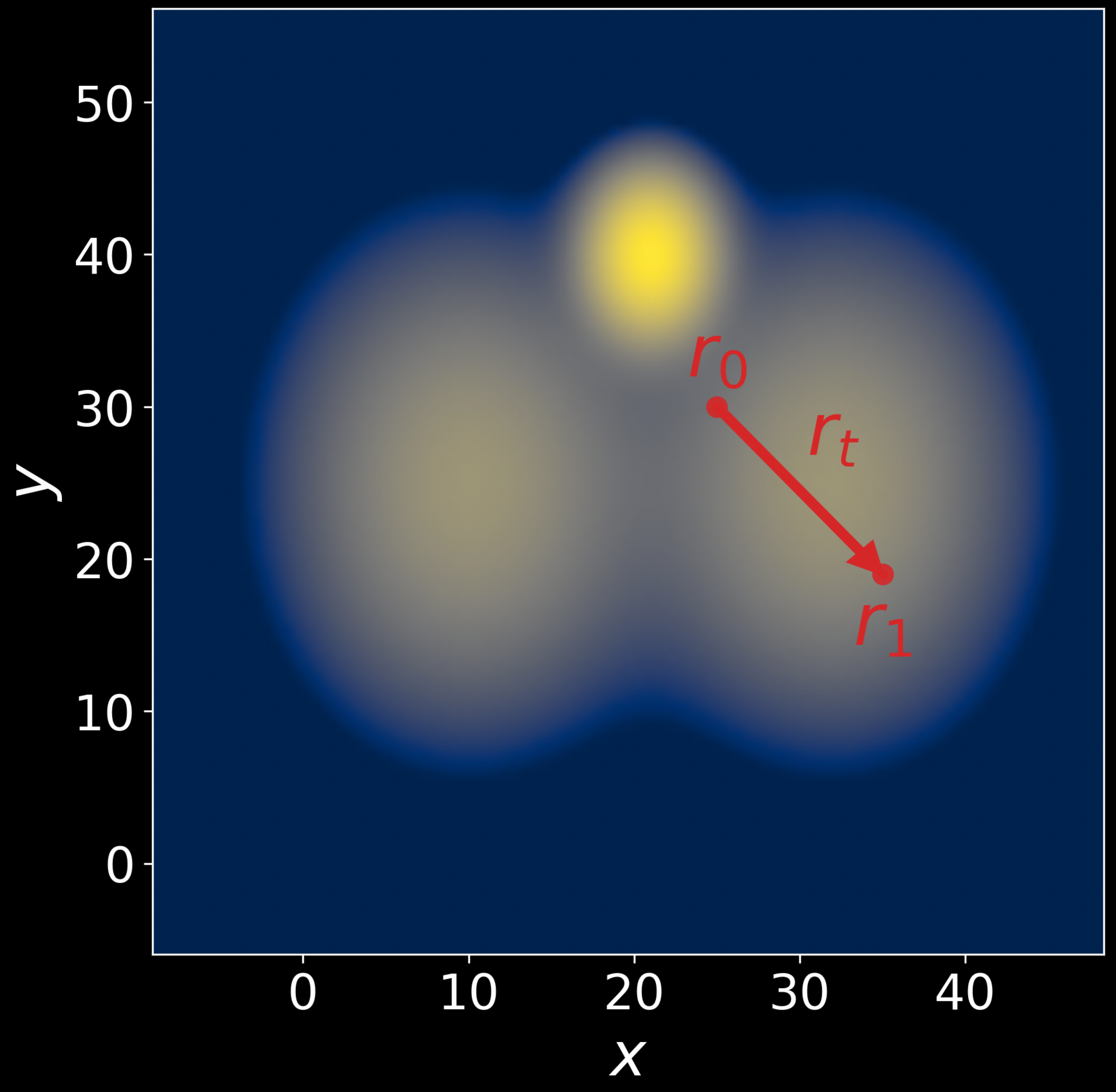
Multiple modes: density dip along path

$$H_1 : \exists t \mid f(r_t) < \min\{f(r_0), f(r_1)\}$$

Single mode: no density dip

$$H_0 : \nexists t \mid f(r_t) < \min\{f(r_0), f(r_1)\}$$





Multimodality test statistic

$$H_0 : \nexists t \mid f(r_t) < \min\{f(r_0), f(r_1)\}$$

$$T(t) := \min\{\log f(r_0), \log f(r_1)\} - \log f(r_t)$$

Multimodality test statistic

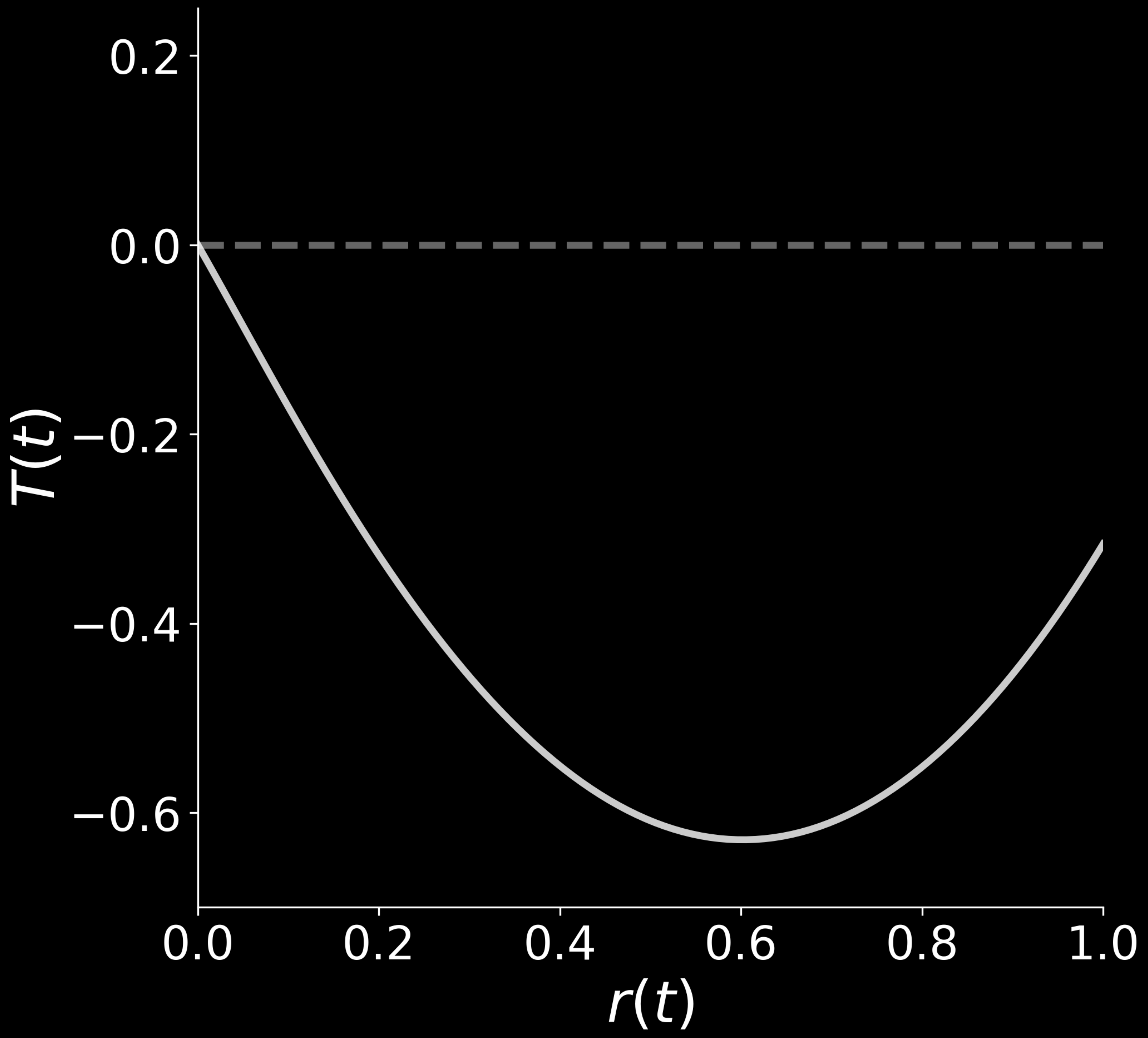
$$H_0 : \nexists t \mid f(r_t) < \min\{f(r_0), f(r_1)\}$$

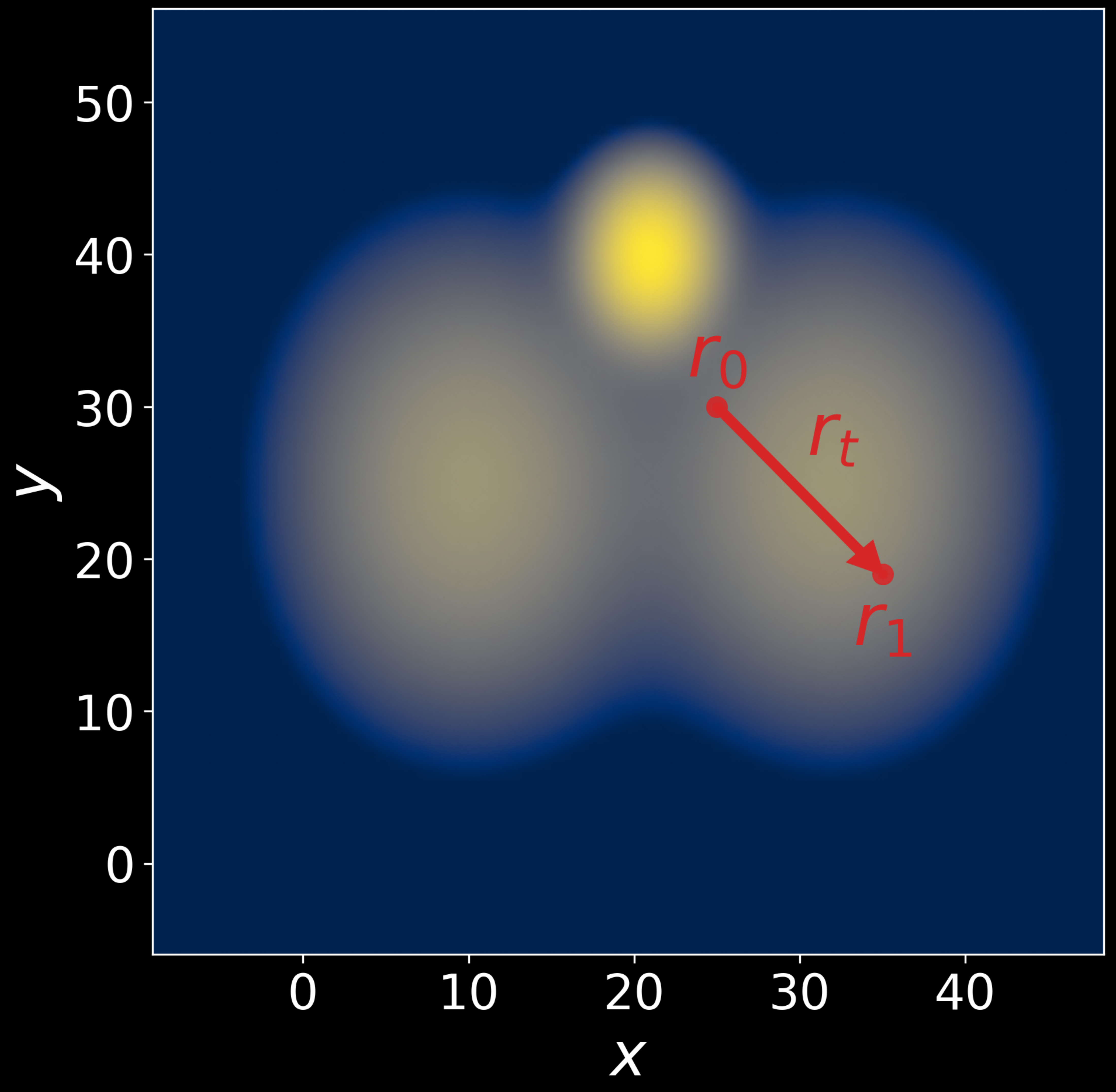
$$T(t) := \min\{\log f(r_0), \log f(r_1)\} - \log f(r_t)$$

$$H_0 : T(t) \leq 0 \quad \forall t \in (0,1)$$

Let's apply: f

$$H_0 : T(t) \leq 0 \quad \forall t \in (0,1)$$ ✓

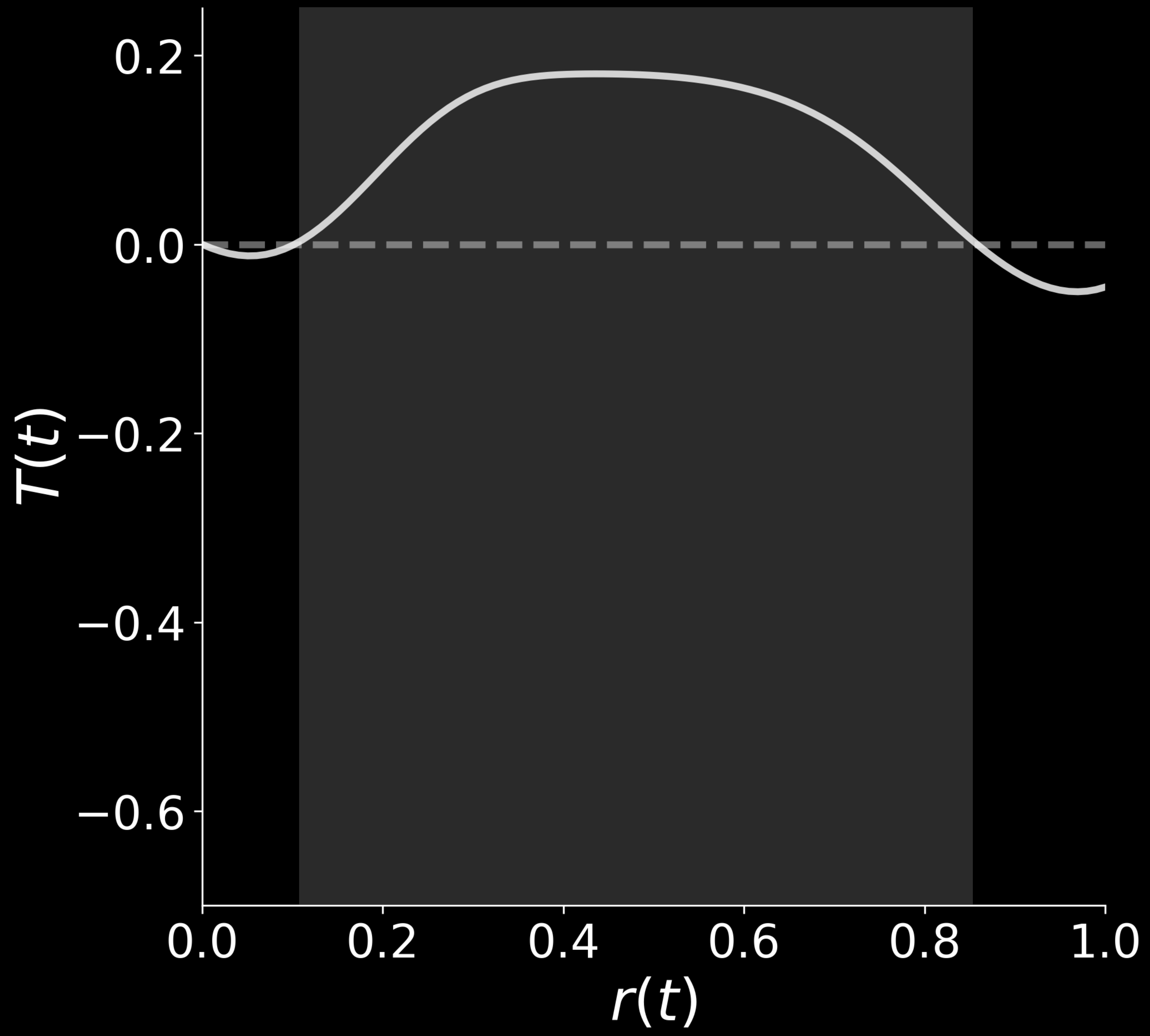


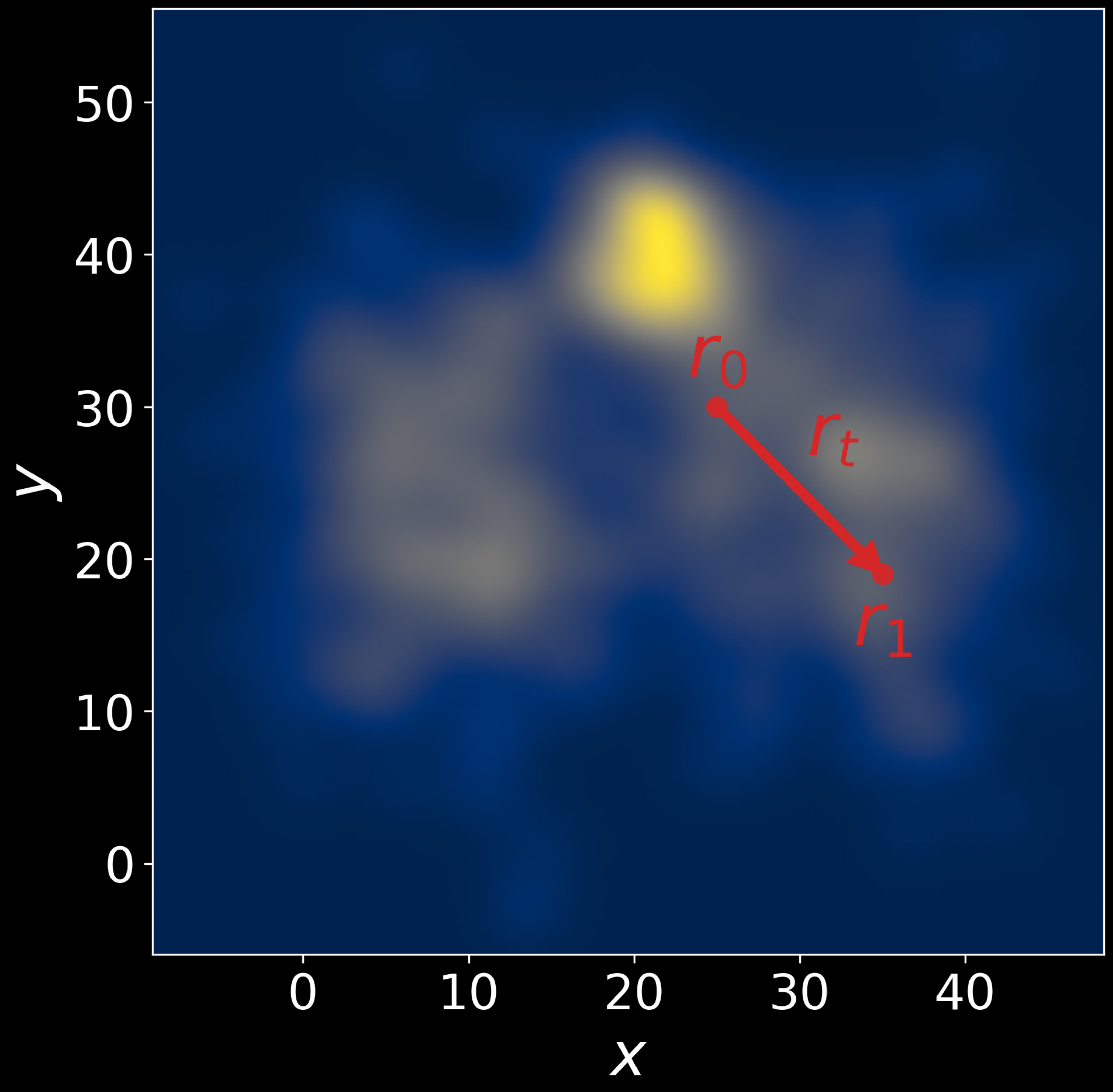


On estimated density?

Let's apply: $\hat{f}$

$$H_0 : T(t) \leq 0 \ \ \forall t \in (0,1)$$ ✗





$$f \to \hat{f} : T(t) \to \hat{T}(t)$$

Multimodality test statistic: $\hat{T}(t)$

$$T(t) := \min\{\log f(r_0), \log f(r_1)\} - \log f(r_t)$$

$$\hat{f}(x) \propto \frac{1}{d_k^p(x)}$$

← **k-NN density estimator**

Multimodality test statistic: $\hat{T}(t)$

$$T(t) := \min\{\log f(r_0), \log f(r_1)\} - \log f(r_t)$$

$$\hat{f}(x) \propto \frac{1}{d_k^p(x)}$$

← **k-NN density estimator**

$$\hat{T}(t) := -p \max\{\log d_k(r_0), \log d_k(r_1)\} + p \log d_k(r_t)$$

Multimodality test statistic: $\hat{T}(t)$

$$T(t) := \min\{\log f(r_0), \log f(r_1)\} - \log f(r_t)$$

$$\hat{f}(x) \propto \frac{1}{d_k^p(x)}$$

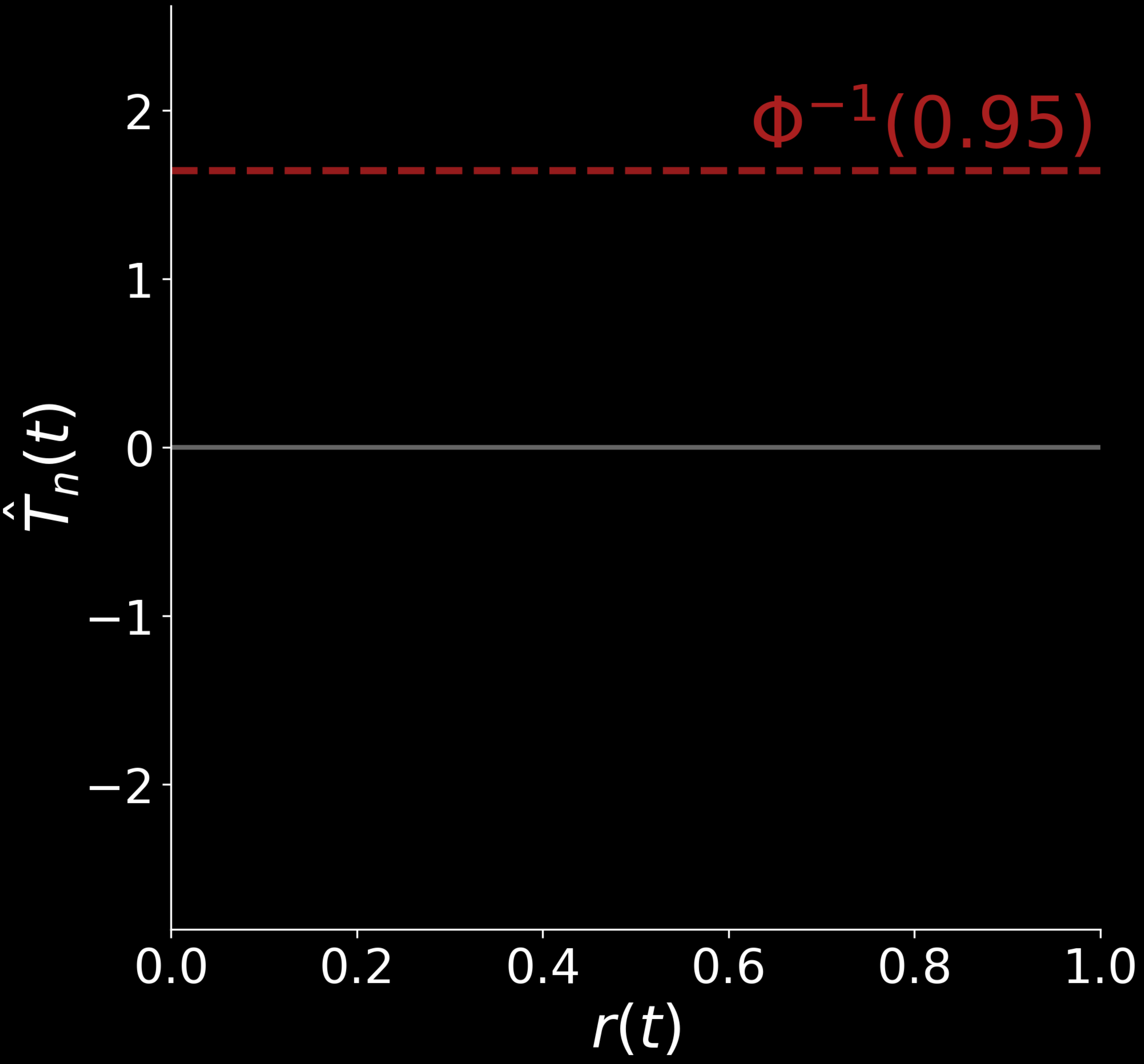
← **k-NN density estimator**

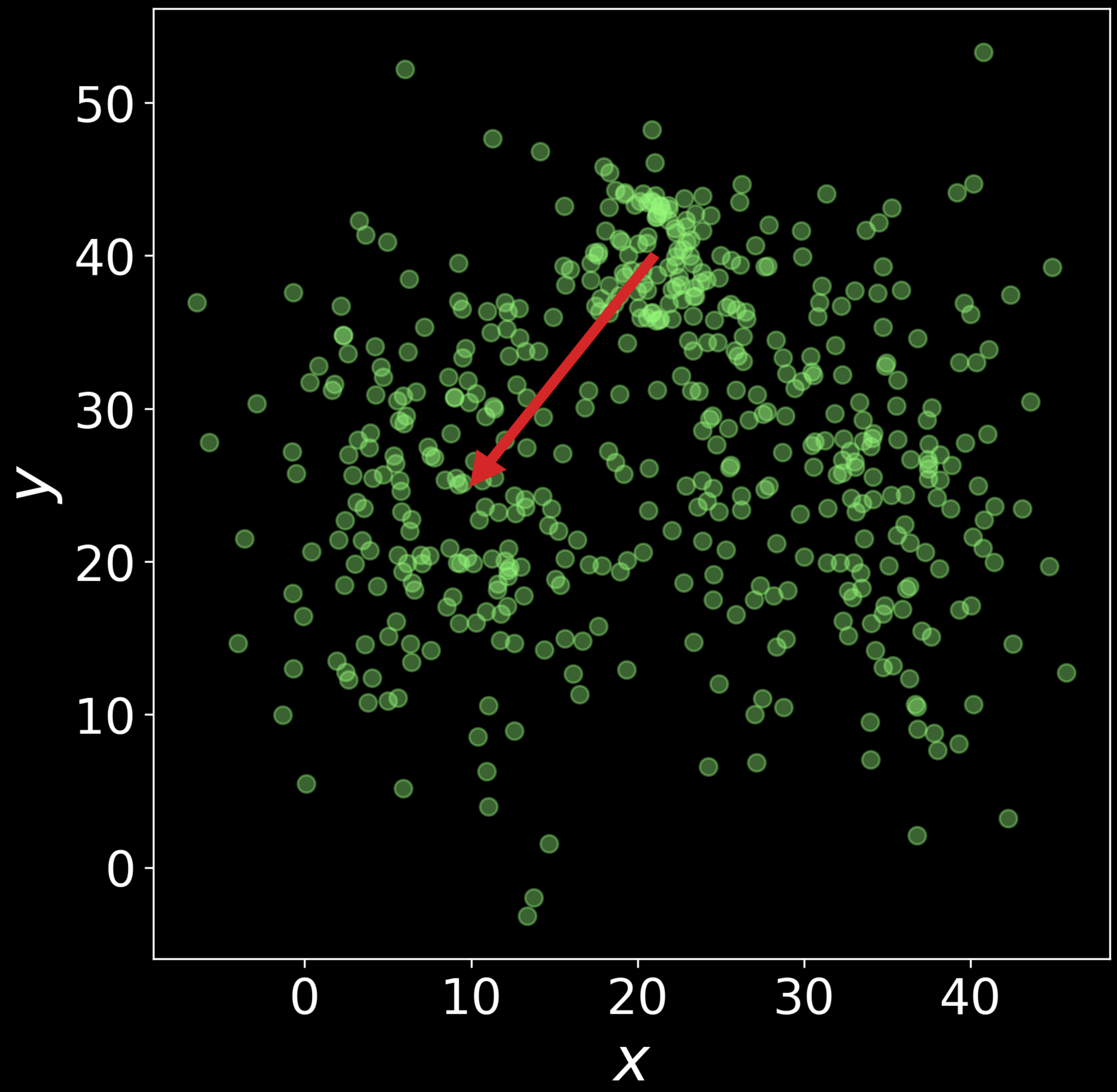
$$\hat{T}(t) := -p \max\{\log d_k(r_0), \log d_k(r_1)\} + p \log d_k(r_t)$$

Burman & Polonik (2009) show $H_0 : \hat{T}(t) \sim \mathcal{N}(0,1) \times c$

Let’s test it!

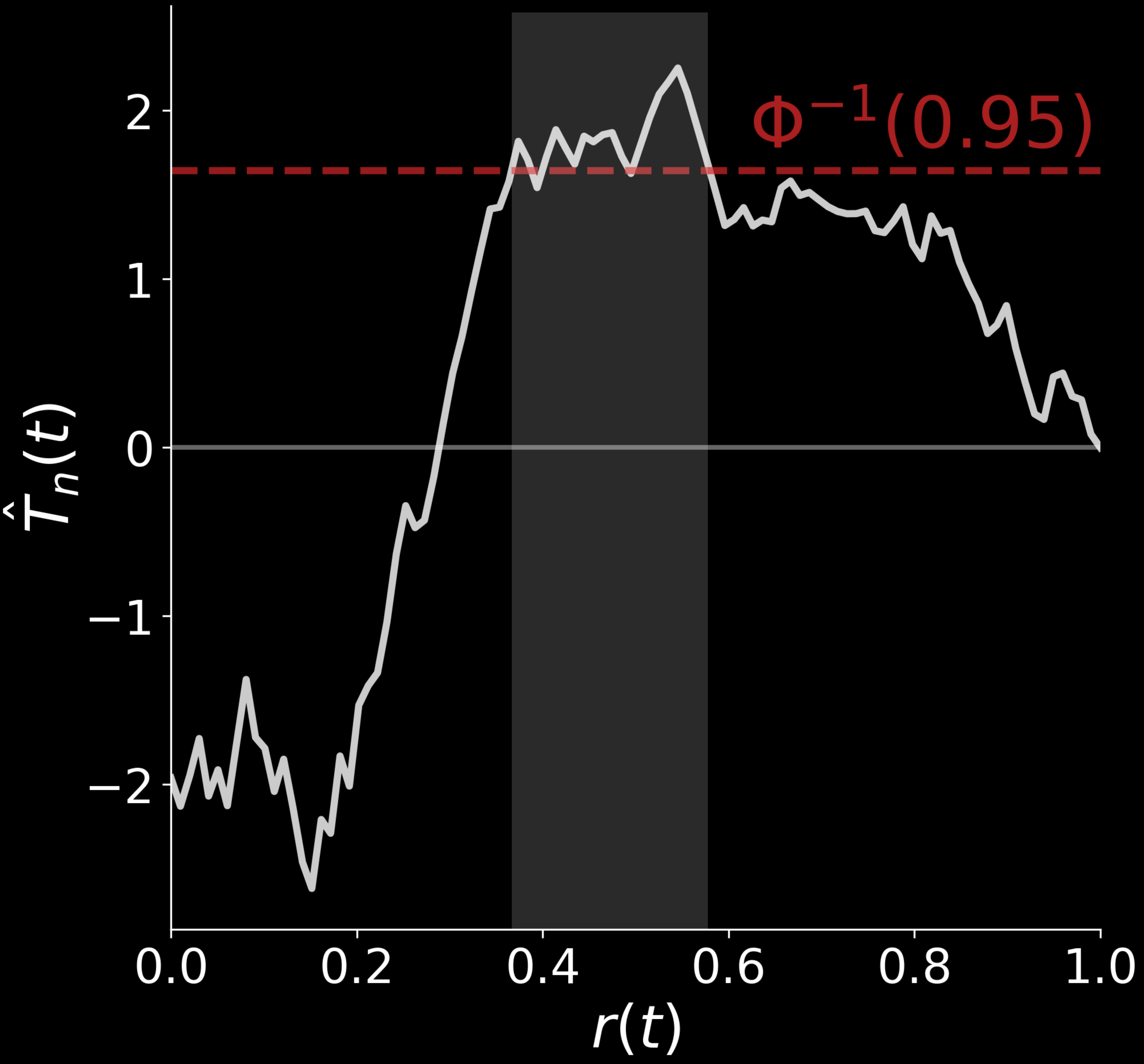
Let's apply: $\hat{f}$ $\quad H_1 : \hat{T}_n(t) > \Phi^{-1}(1-\alpha)$?

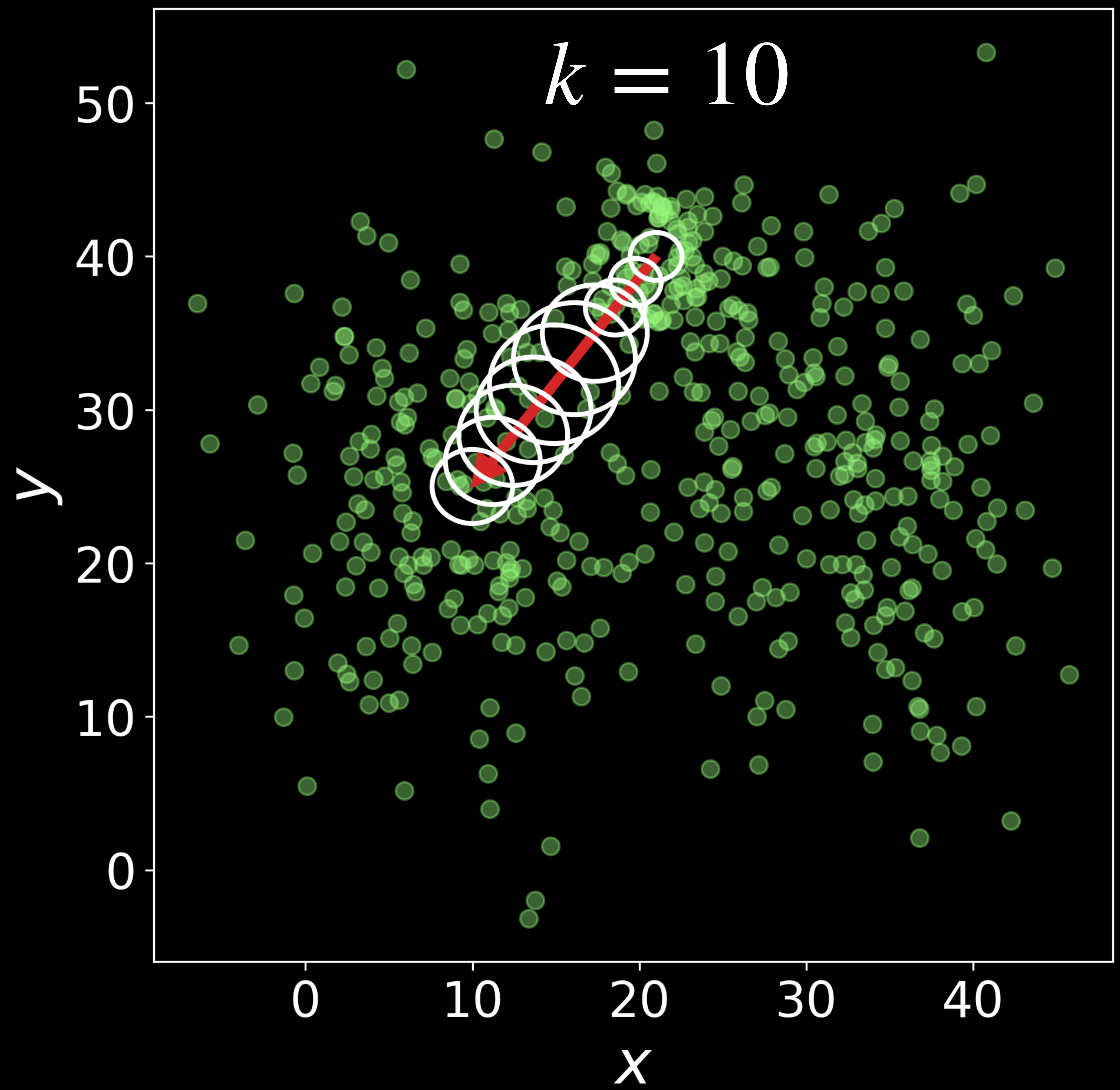




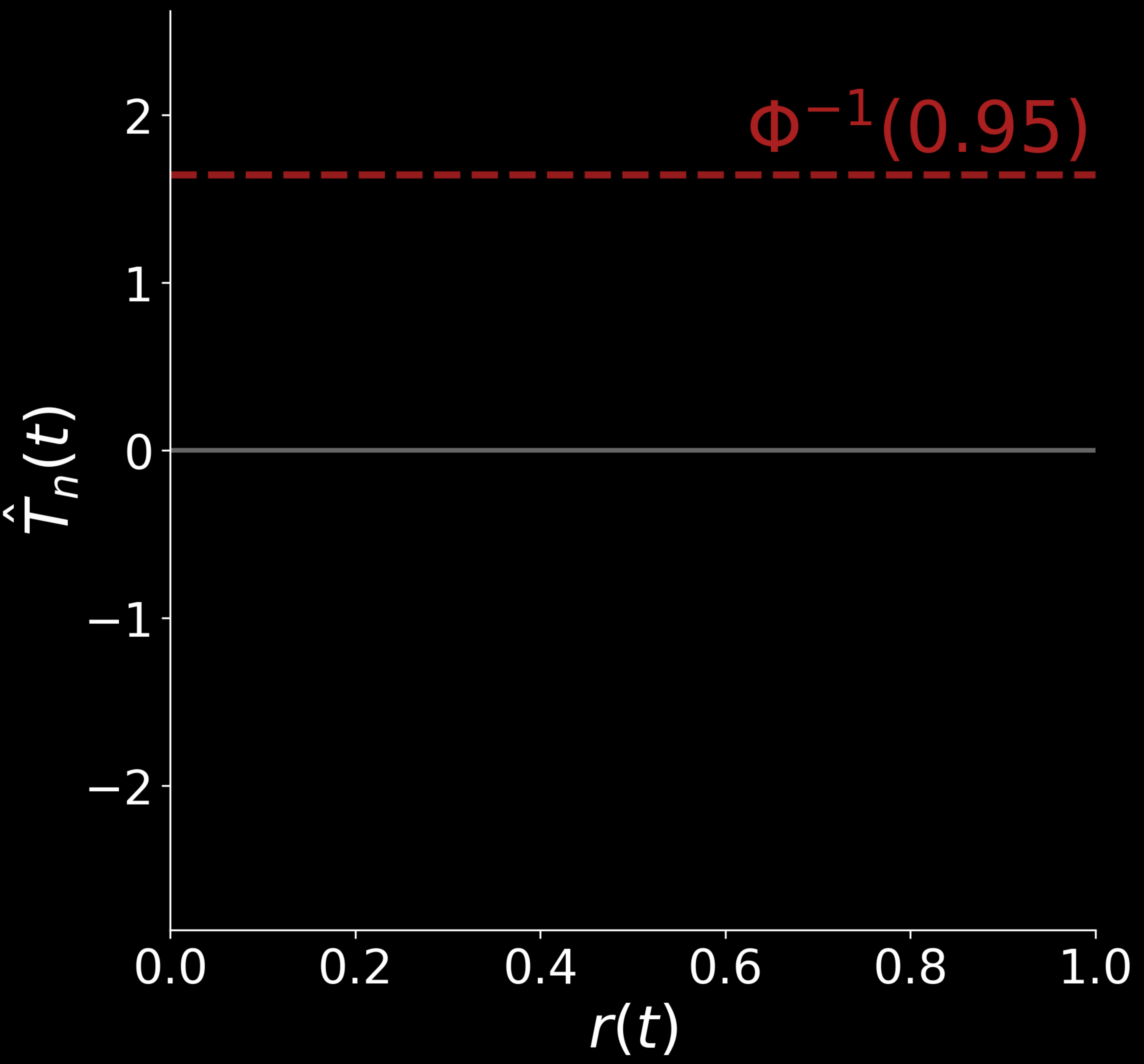
Let's apply: $\hat{f}$

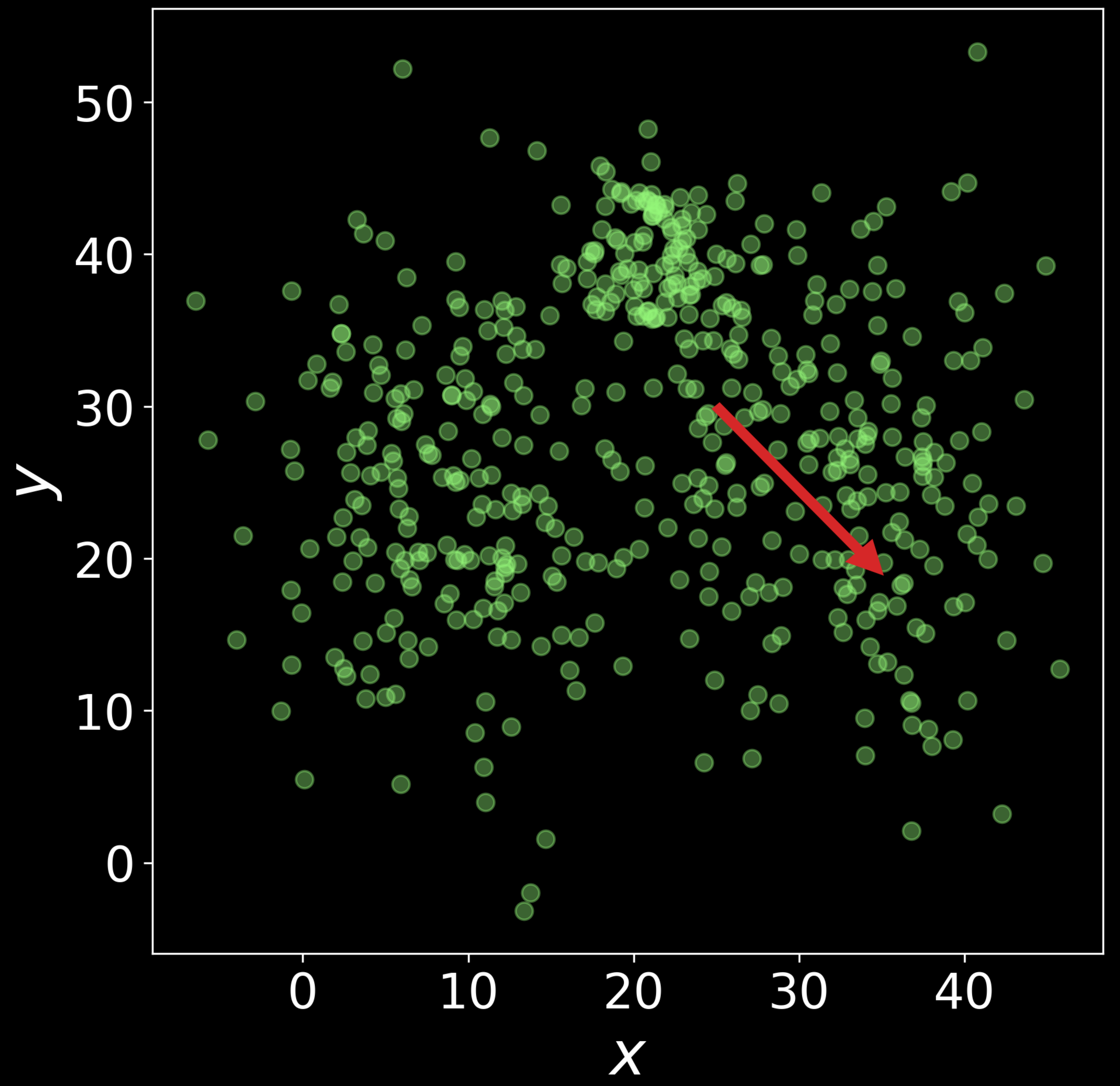
$$H_1 : \hat{T}_n(t) > \Phi^{-1}(1 - \alpha)$$





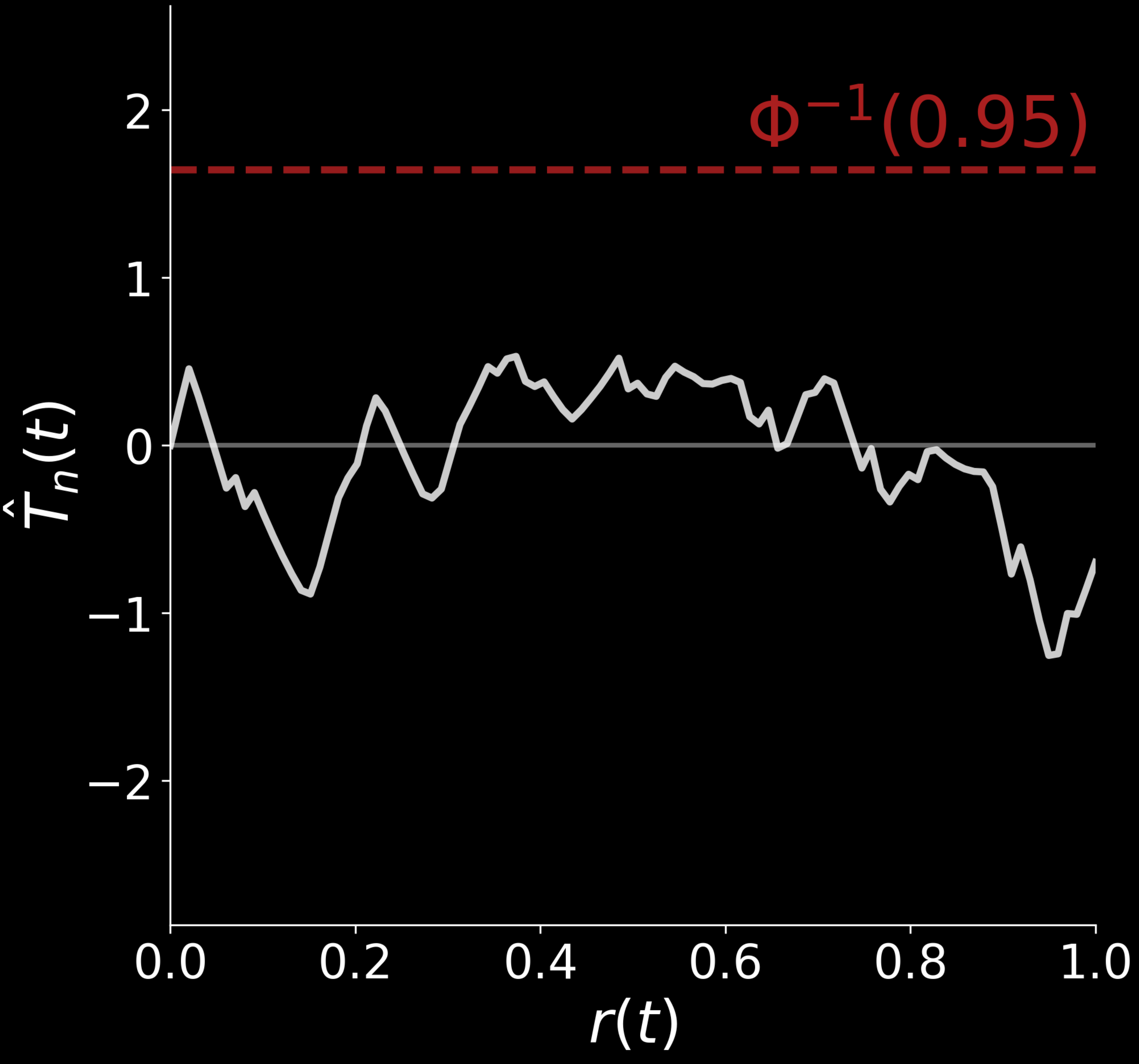
Let's apply: $\hat{f}$ $H_0 : \hat{T}_n(t) \leq \Phi^{-1}(1-\alpha)$?

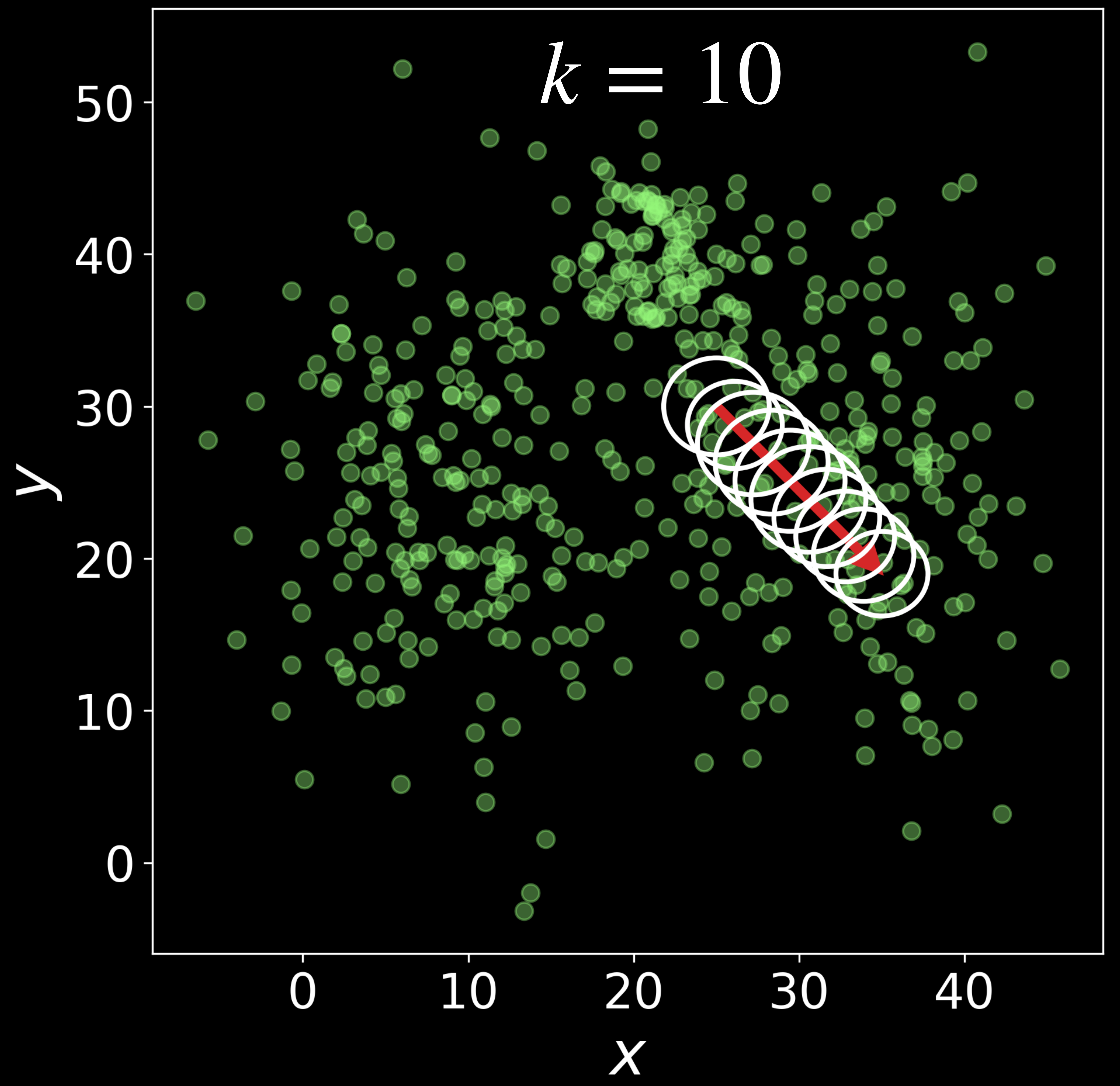




Let's apply: $\hat{f}$

$$H_0 : \hat{T}_n(t) \leq \Phi^{-1}(1 - \alpha)$$ ✓

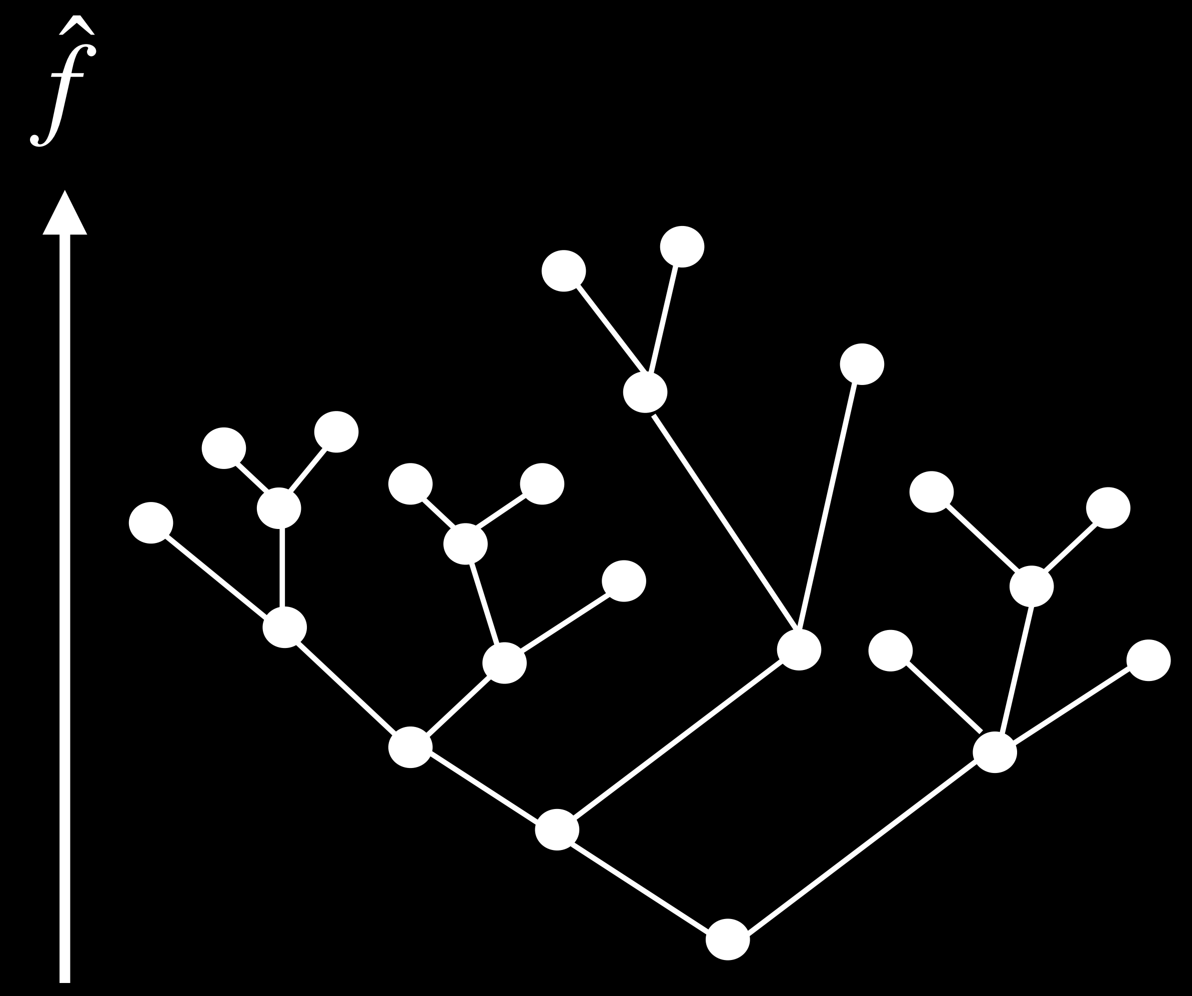


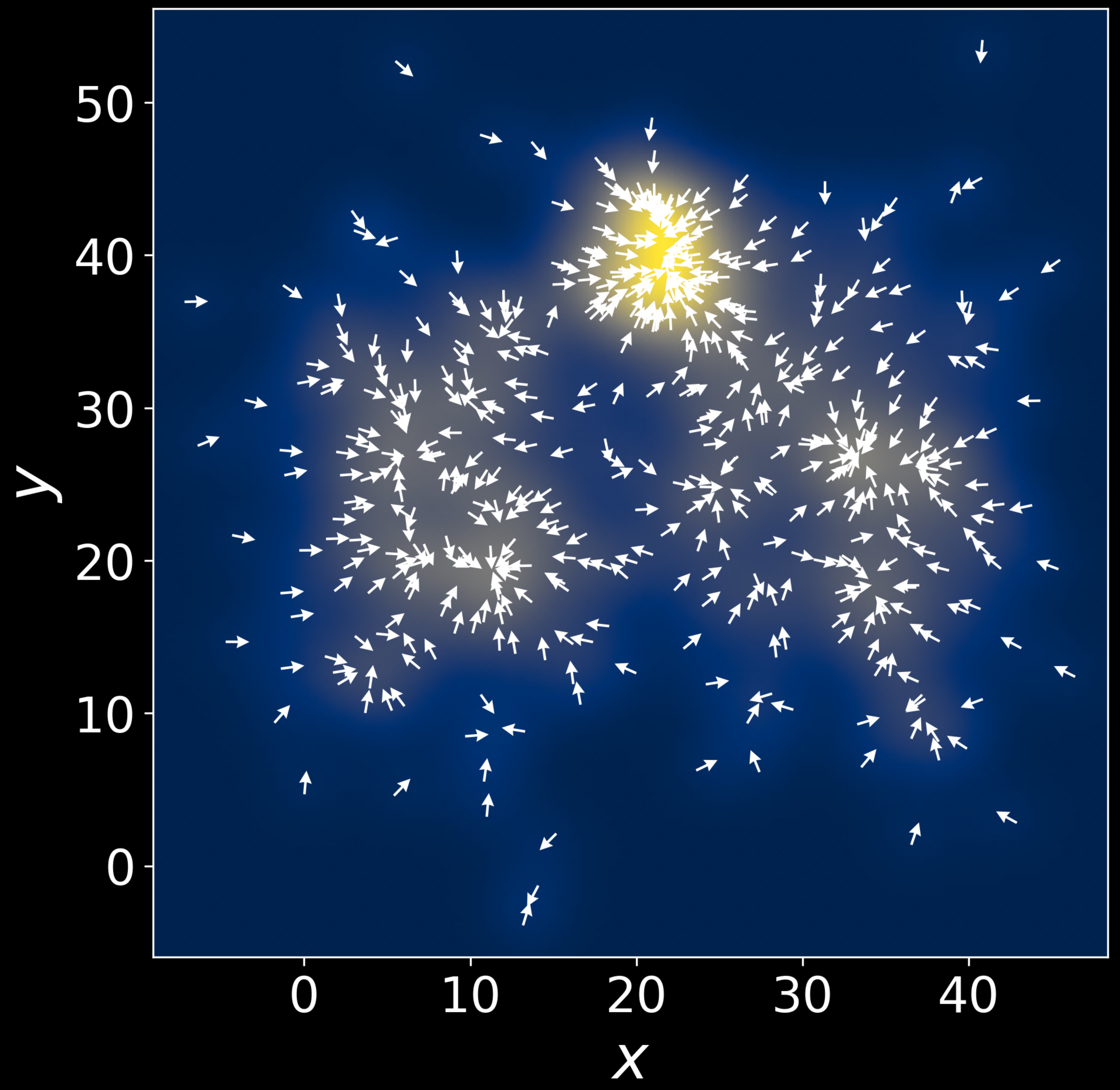


Putting it all together

Clustering pipeline: SigMA

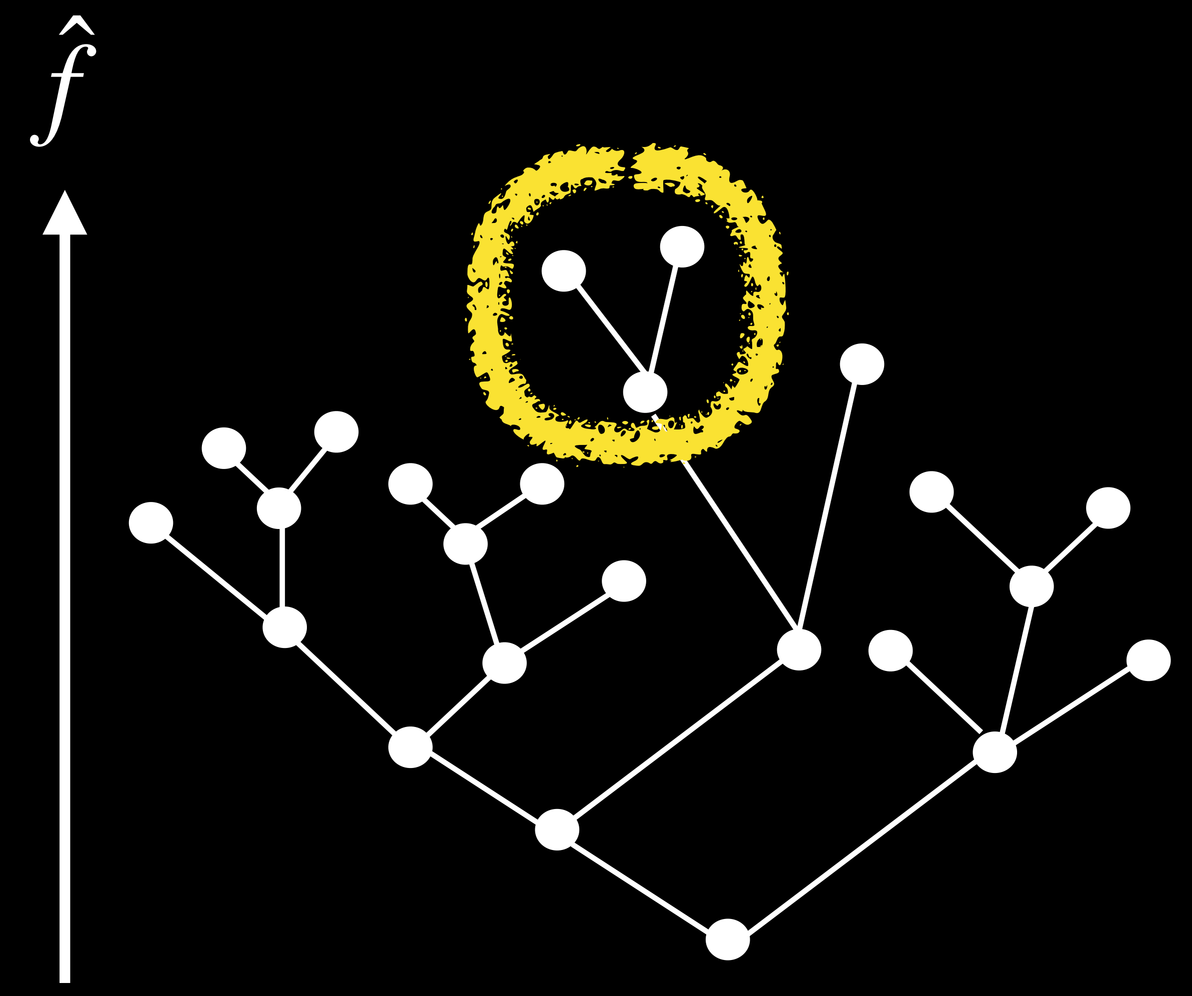
1. Gradient ascent step — cluster tree

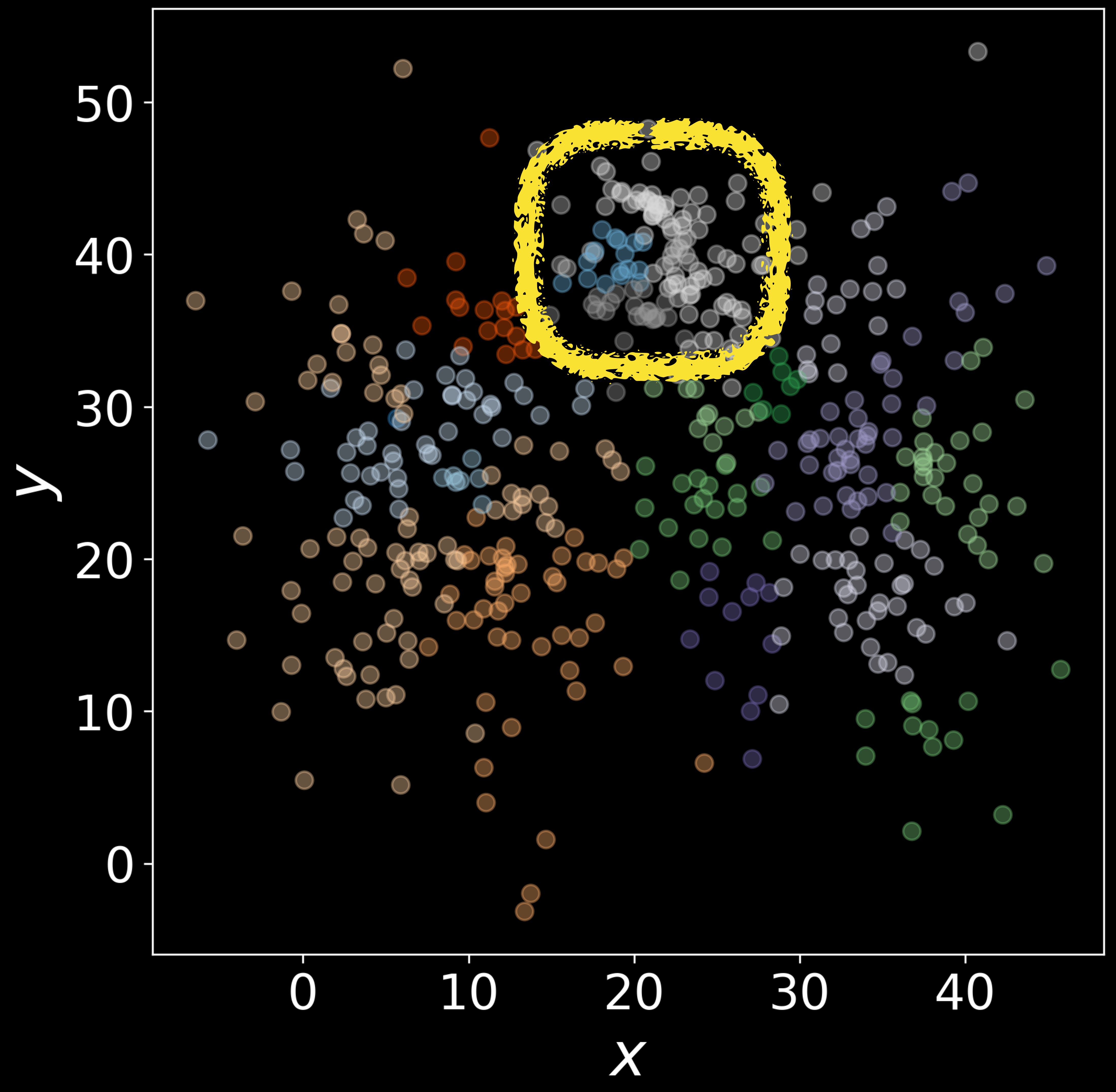




Clustering pipeline: SigMA

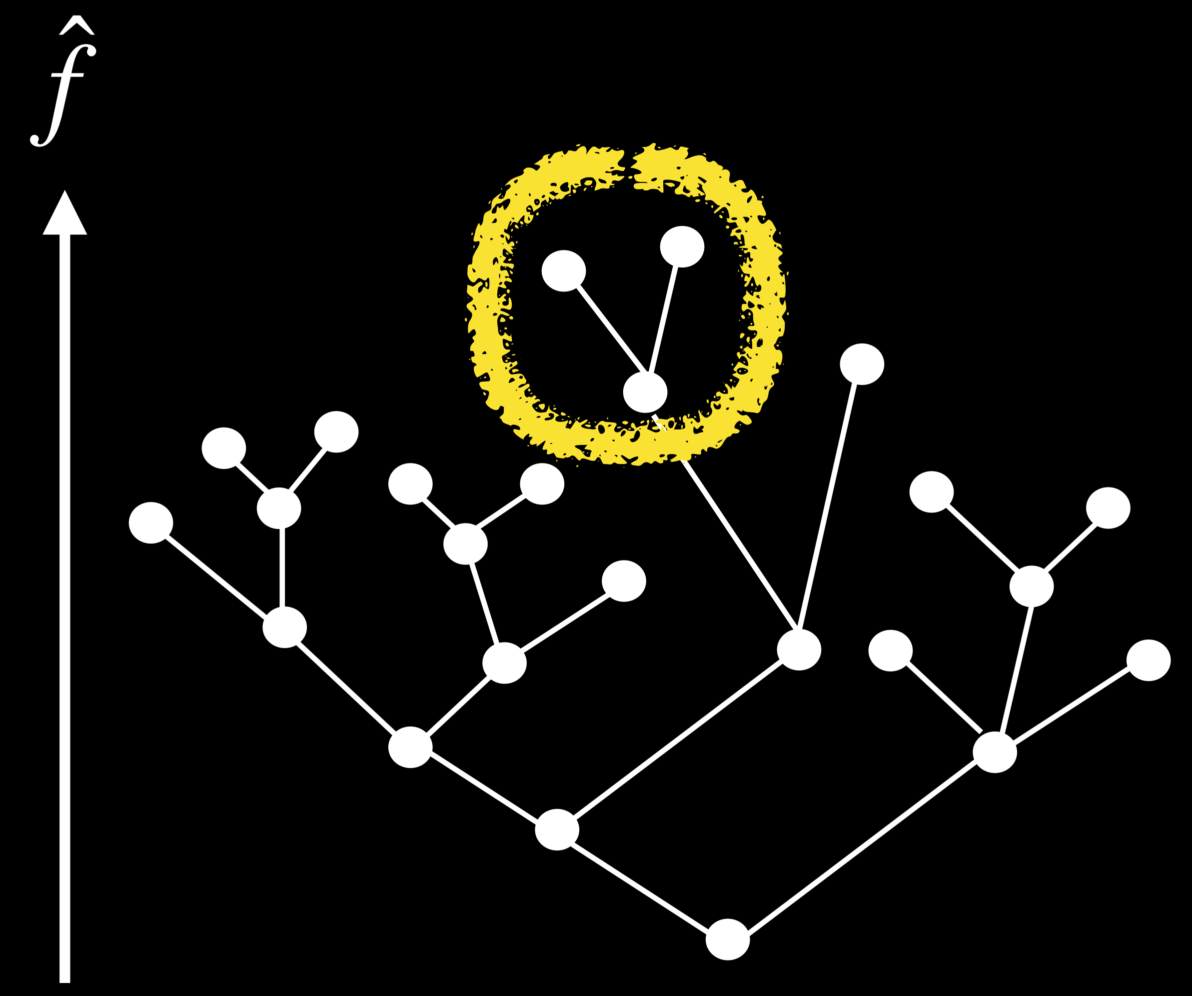
1. Gradient ascent step
2. Scan saddle points: $\max \hat{f} \rightarrow \min \hat{f}$

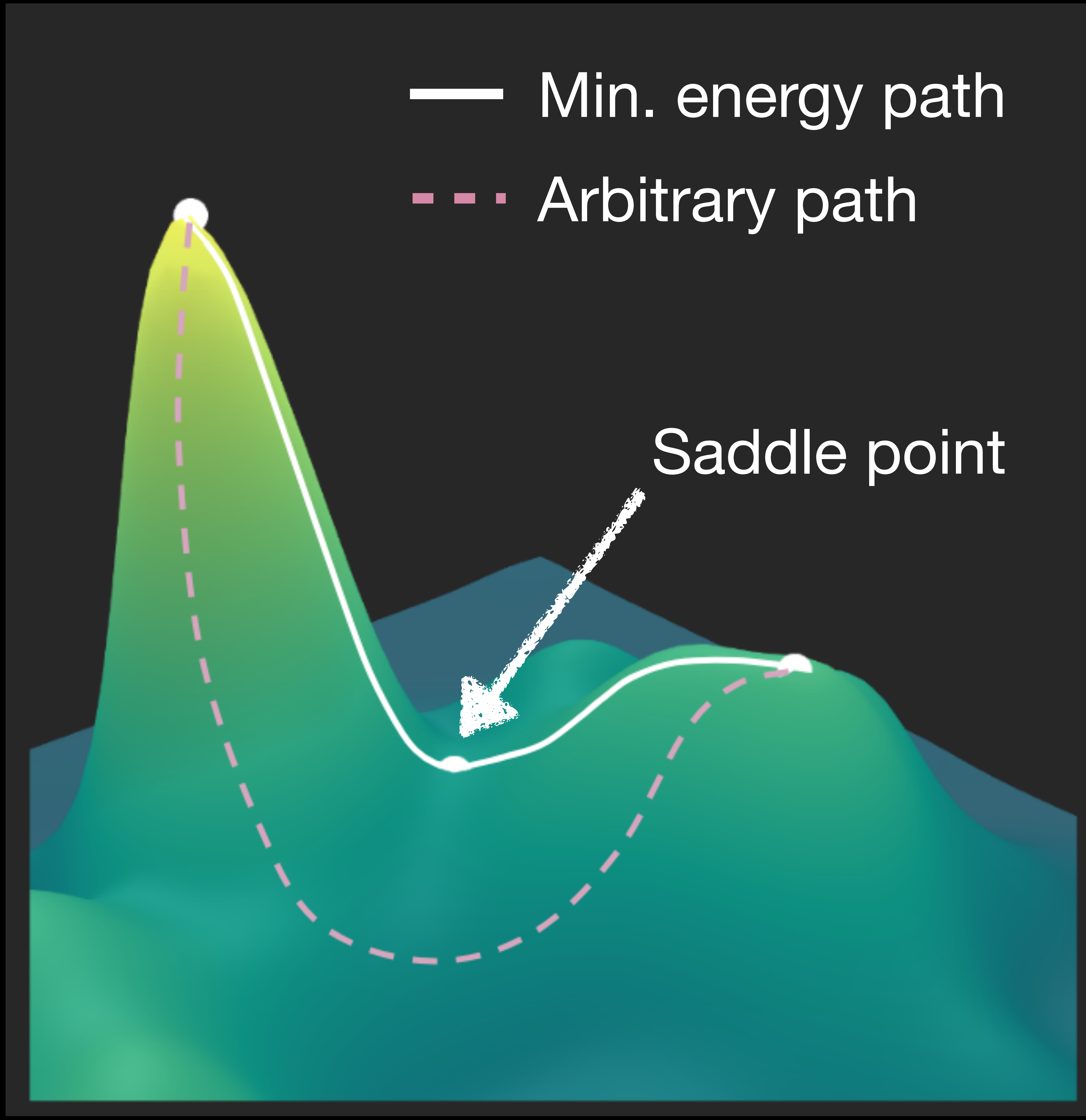




Clustering pipeline: SigMA

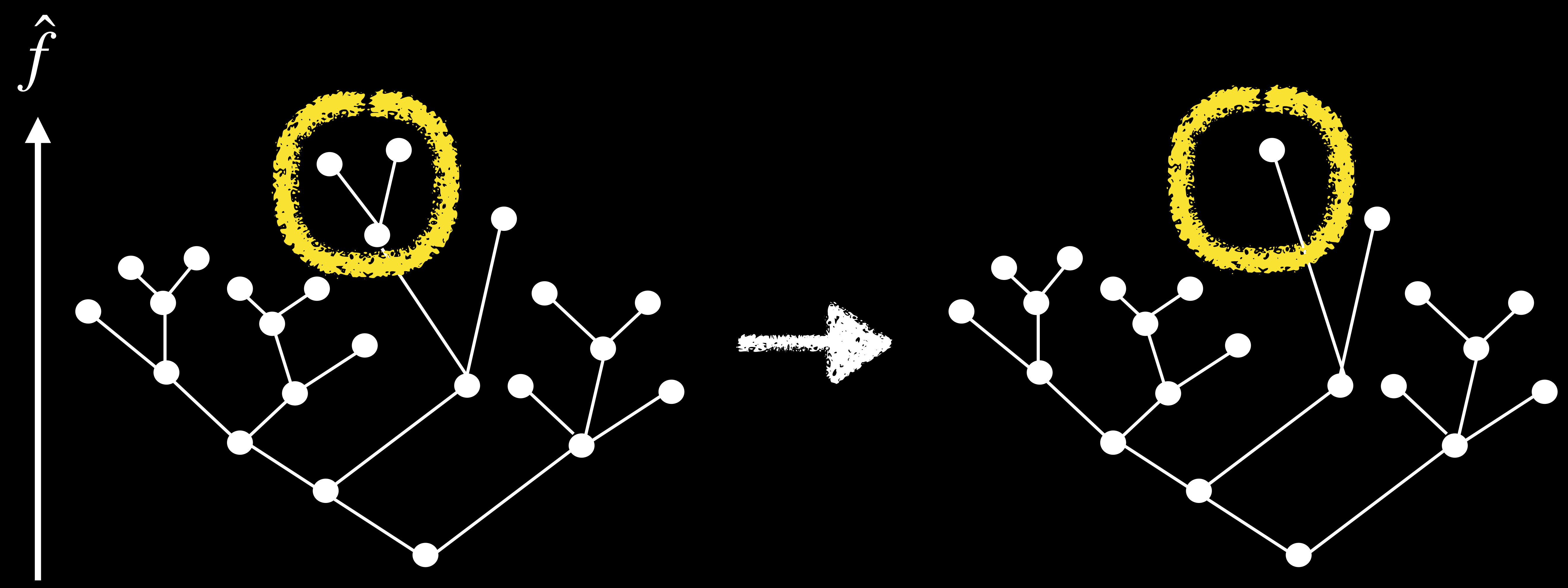
1. Gradient ascent step
2. Scan saddle points: max $\hat{f} \rightarrow$ min $\hat{f}$
   A. Test modality between modes

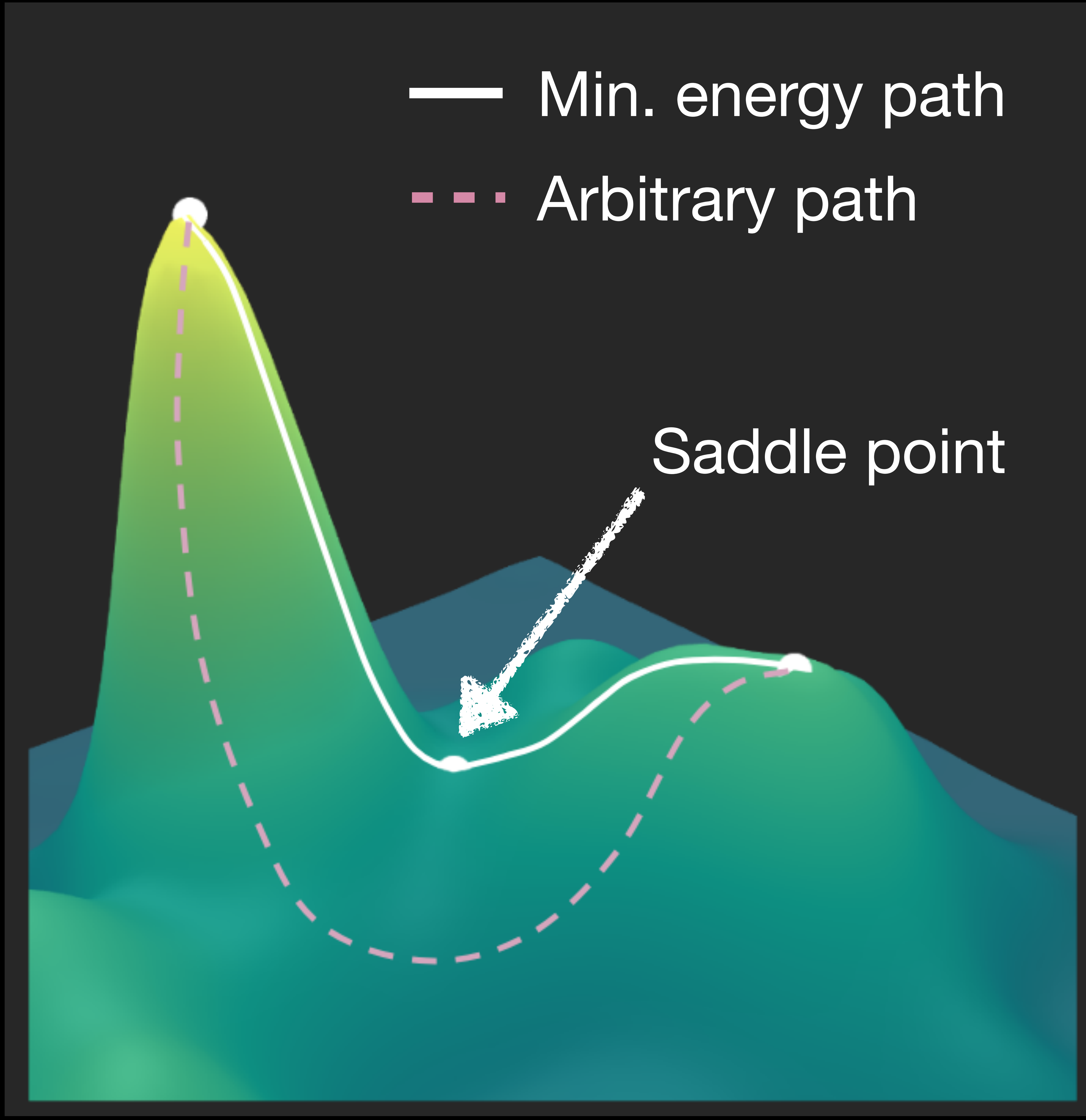




Clustering pipeline: SigMA

1. Gradient ascent step
2. Scan saddle points: max $\hat{f} \rightarrow$ min $\hat{f}$
   A. Test modality between modes
   B. If H_0 cannot be rejected — merge

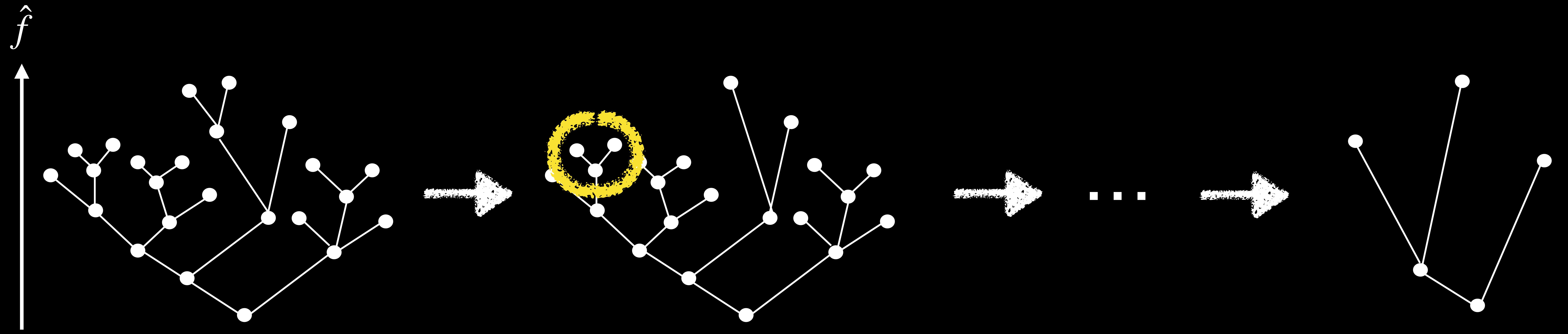




Clustering pipeline: SigMA

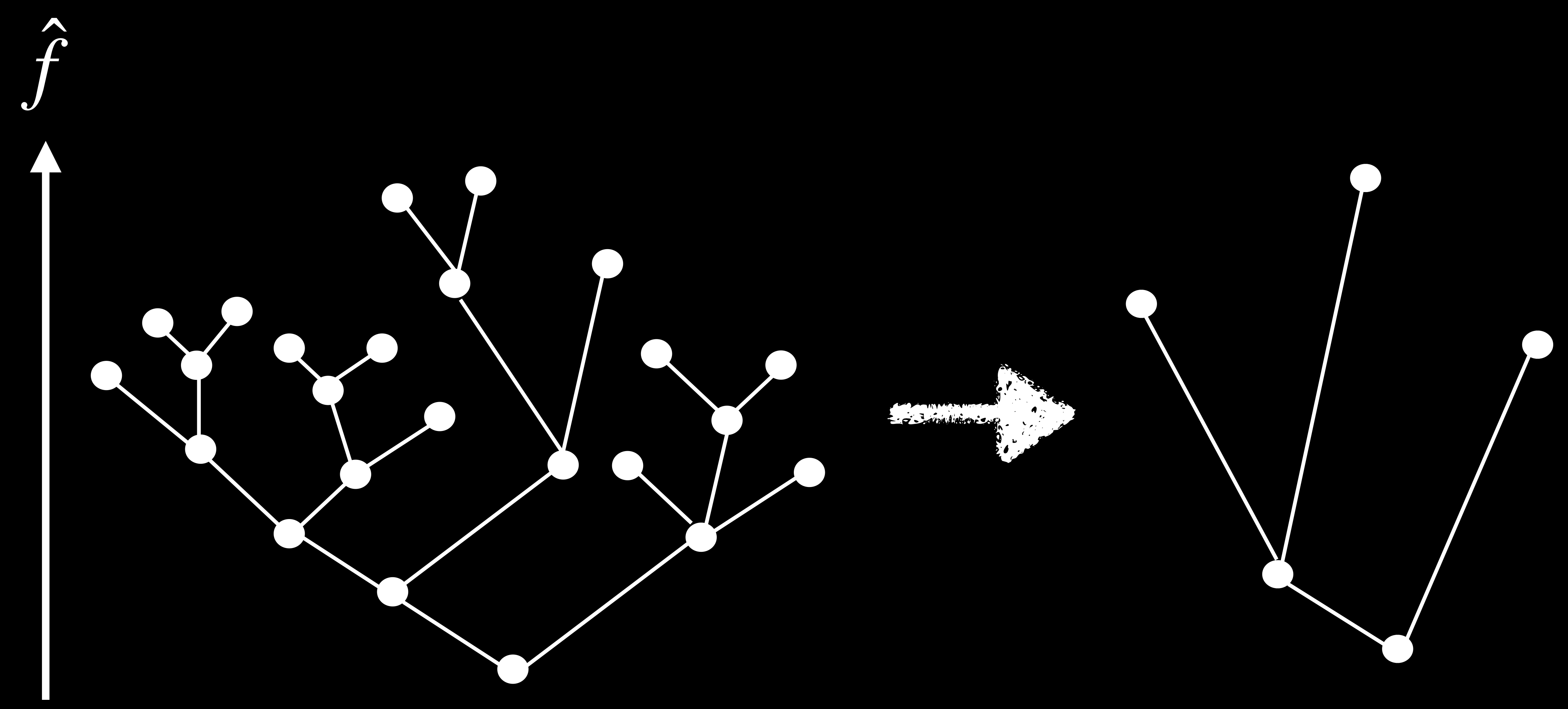
1. Gradient ascent step
2. Scan saddle points: max $\hat{f} \rightarrow$ min $\hat{f}$
   A. Test modality between modes
   B. If H_0 cannot be rejected — merge

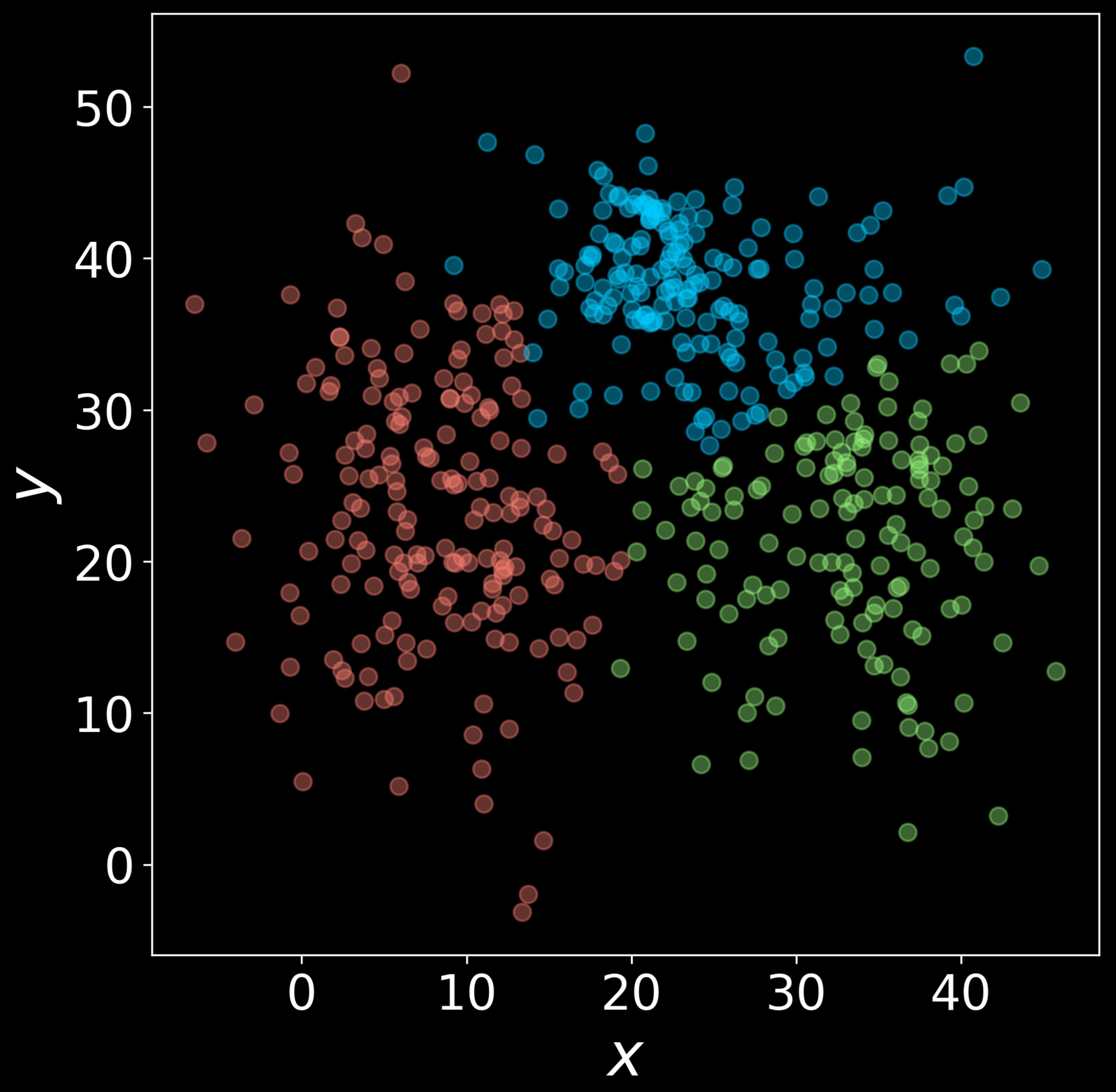
Next saddle point



Clustering pipeline: SigMA

1. Gradient ascent step
2. Scan saddle points: max $\hat{f} \rightarrow$ min $\hat{f}$
   A. Test modality between modes
   B. If H_0 cannot be rejected — merge



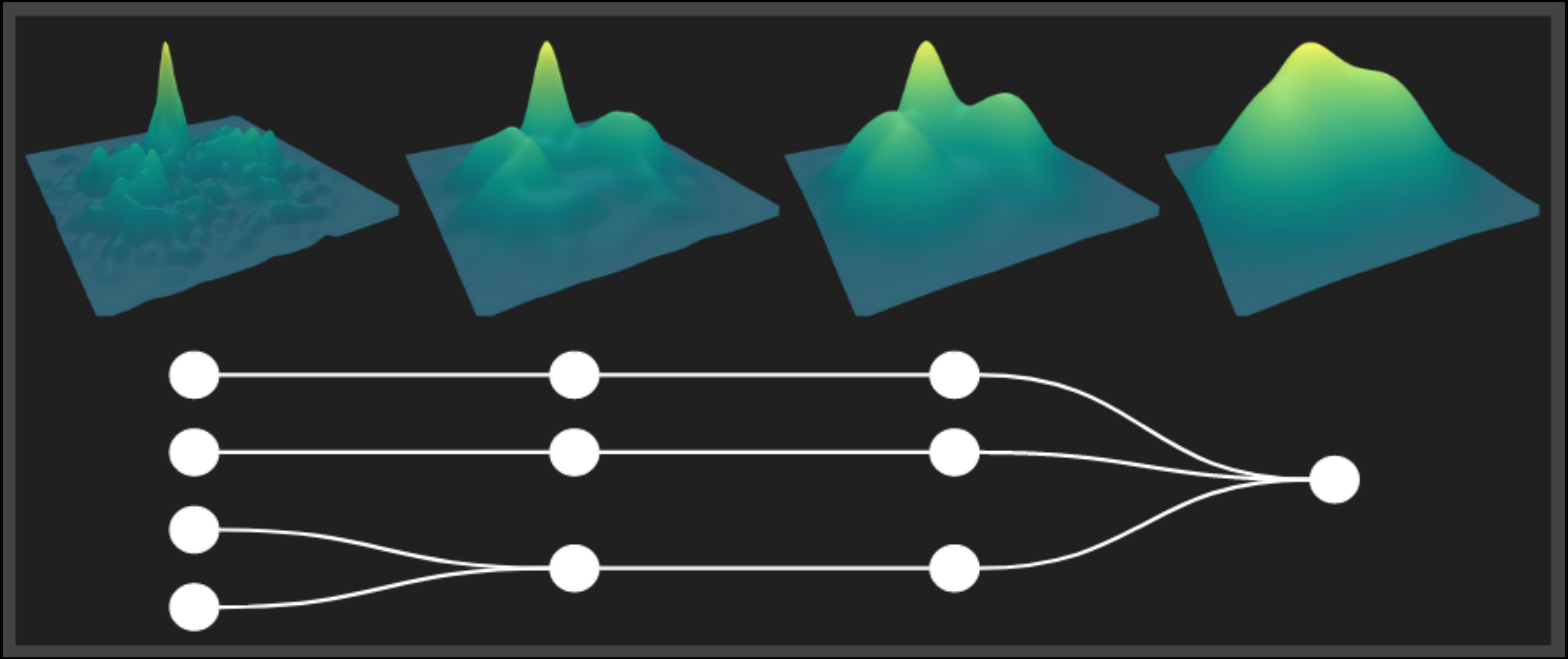


How to set parameters?

`SigMA(k,`α`)`

Choosing k

$$\hat{T}_n(t) \sim \mathcal{N}(0,1) \iff \log N < k < N^{4/(4+p)}$$



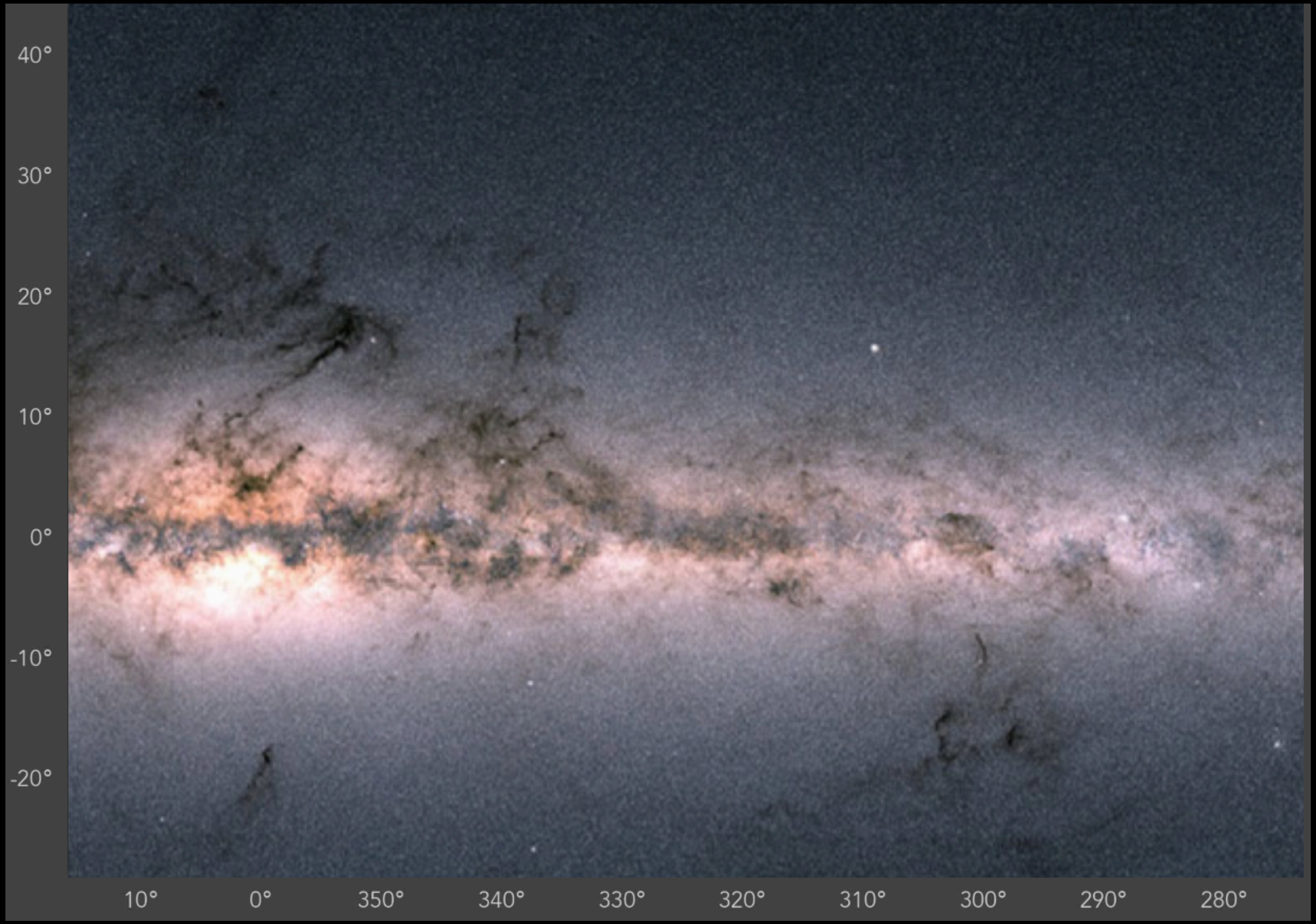
Choosing α

- Many hypotheses tests increases chance of false positives
- Limit proportion of **false positives** among all positives
  - Apply **Benjamini & Hochberg procedure**

➡ Data driven way of choosing significance α

Results on Sco-Cen

Application to Sco-Cen OB association



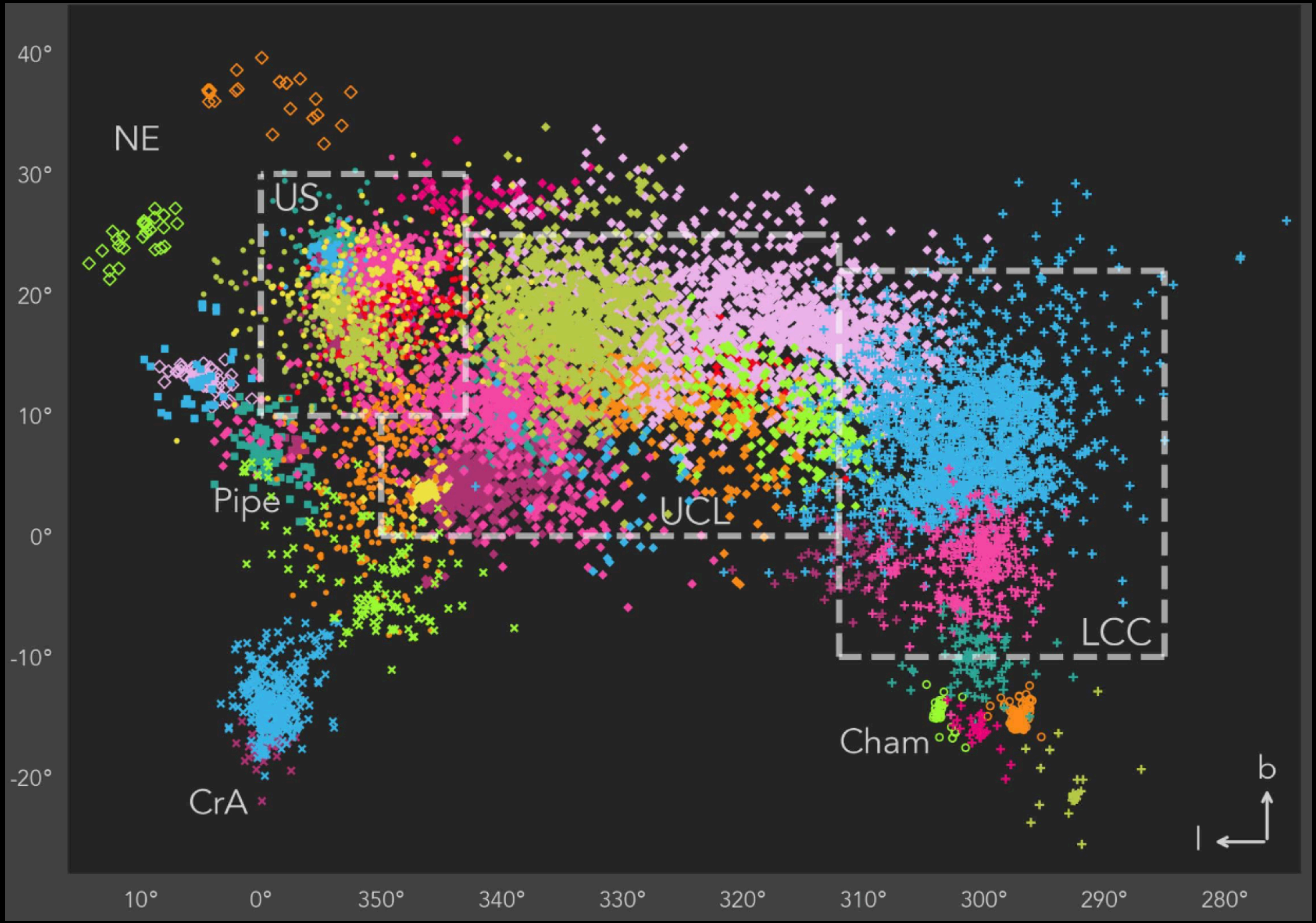
Application to Sco-Cen OB association

Consists of 37 groups

- Unseen substructure

Validated via

- narrow HRD
- B stars in center



Thank you!

Backup

Time complexity

Density computation (k-d tree)

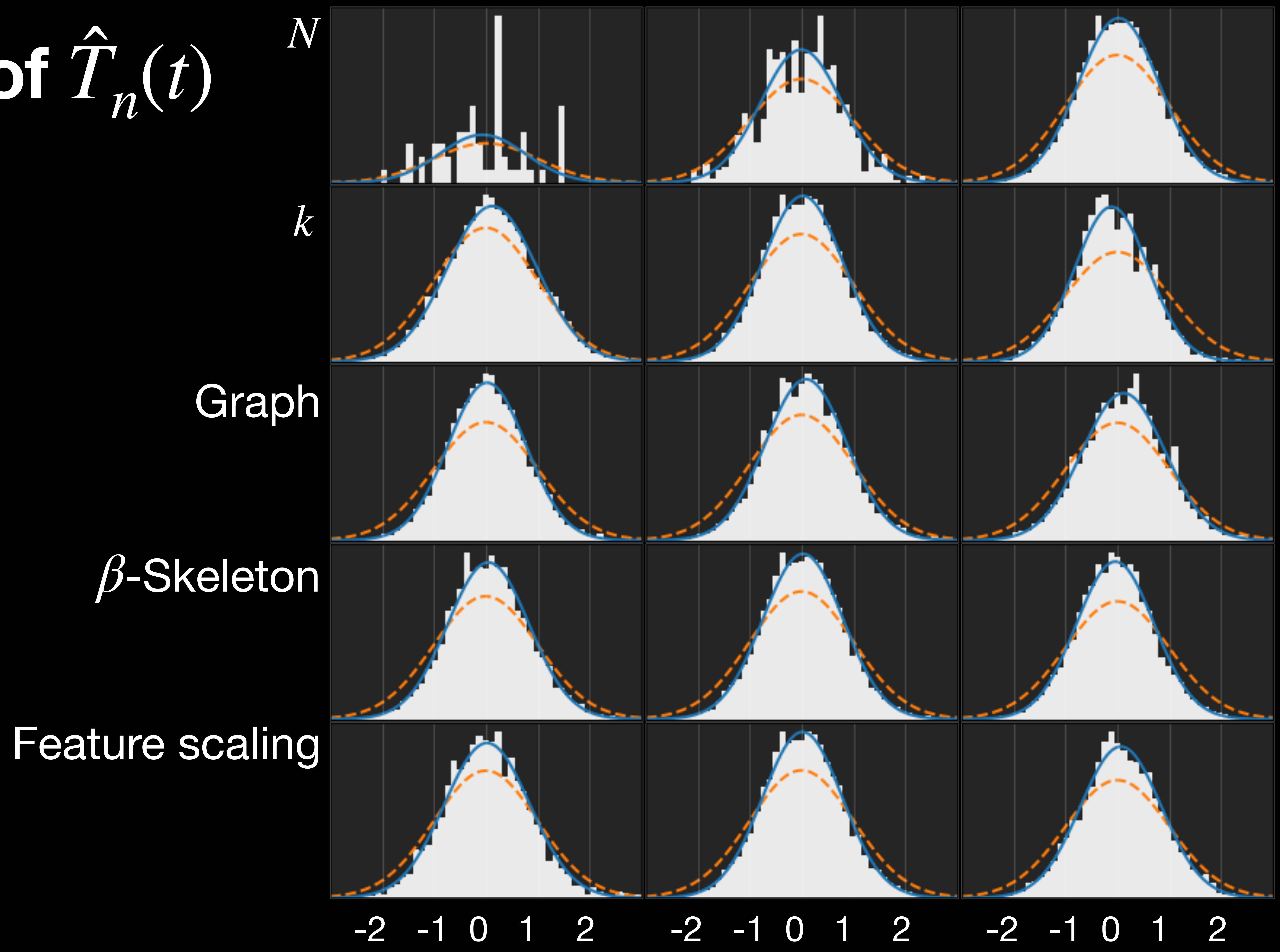
mode & saddle search (union find)

$$\mathcal{O}(pN\log N) + \mathcal{O}(pN\log N) + \mathcal{O}(Nk) + \mathcal{O}(|\mathcal{S}|)$$

Graph construction

Cluster tree pruning

Robustness of $\hat{T}_n(t)$



Application to Sco-Cen OB association

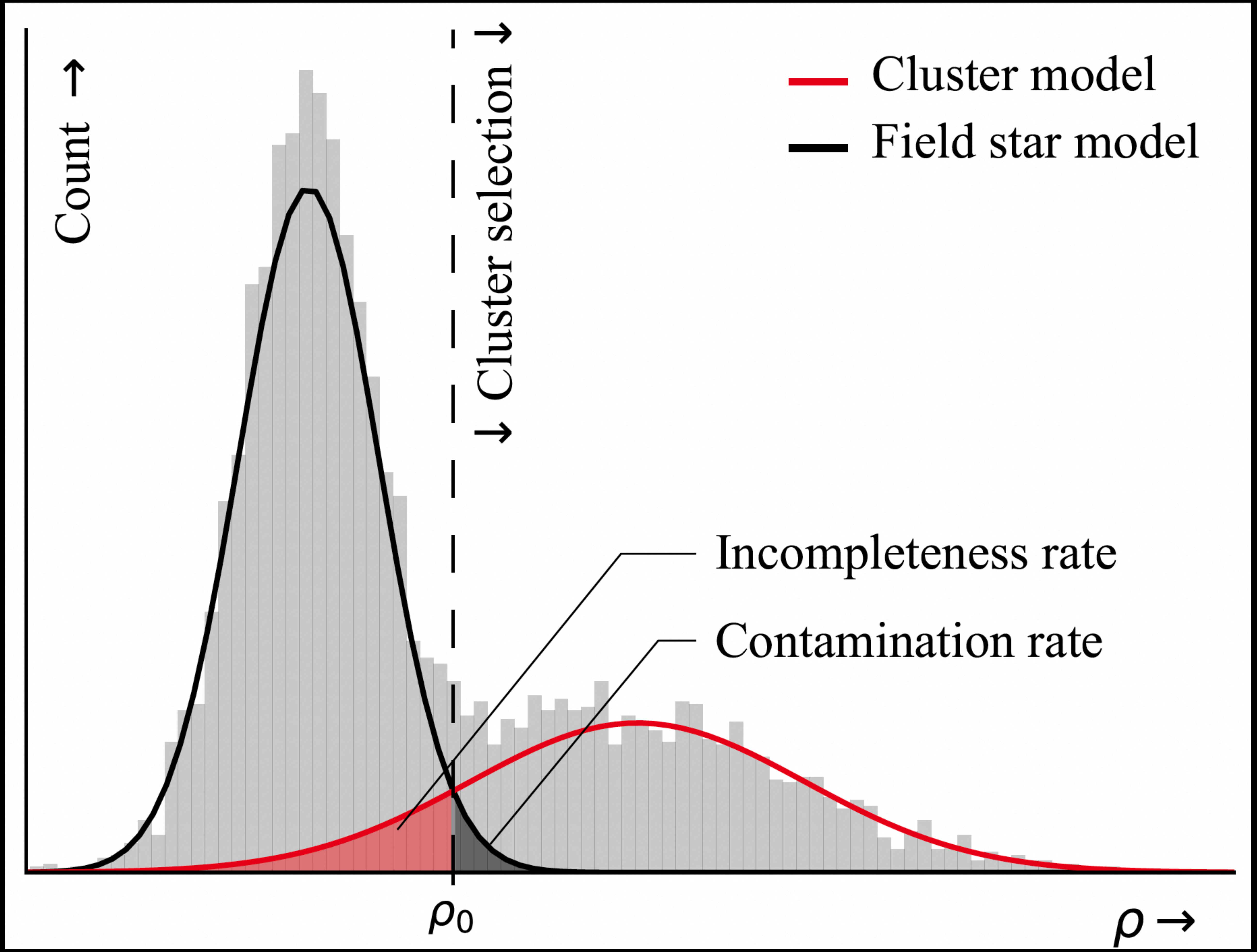
Consists of 37 groups

- Unseen substructure

Validated via

- narrow HRD
- B stars in center
- Age gradients

Background reduction



Background reduction