Density peak clustering: an invitation

Michele Allegra, Institut de Neurosciences de la Timone

Outline

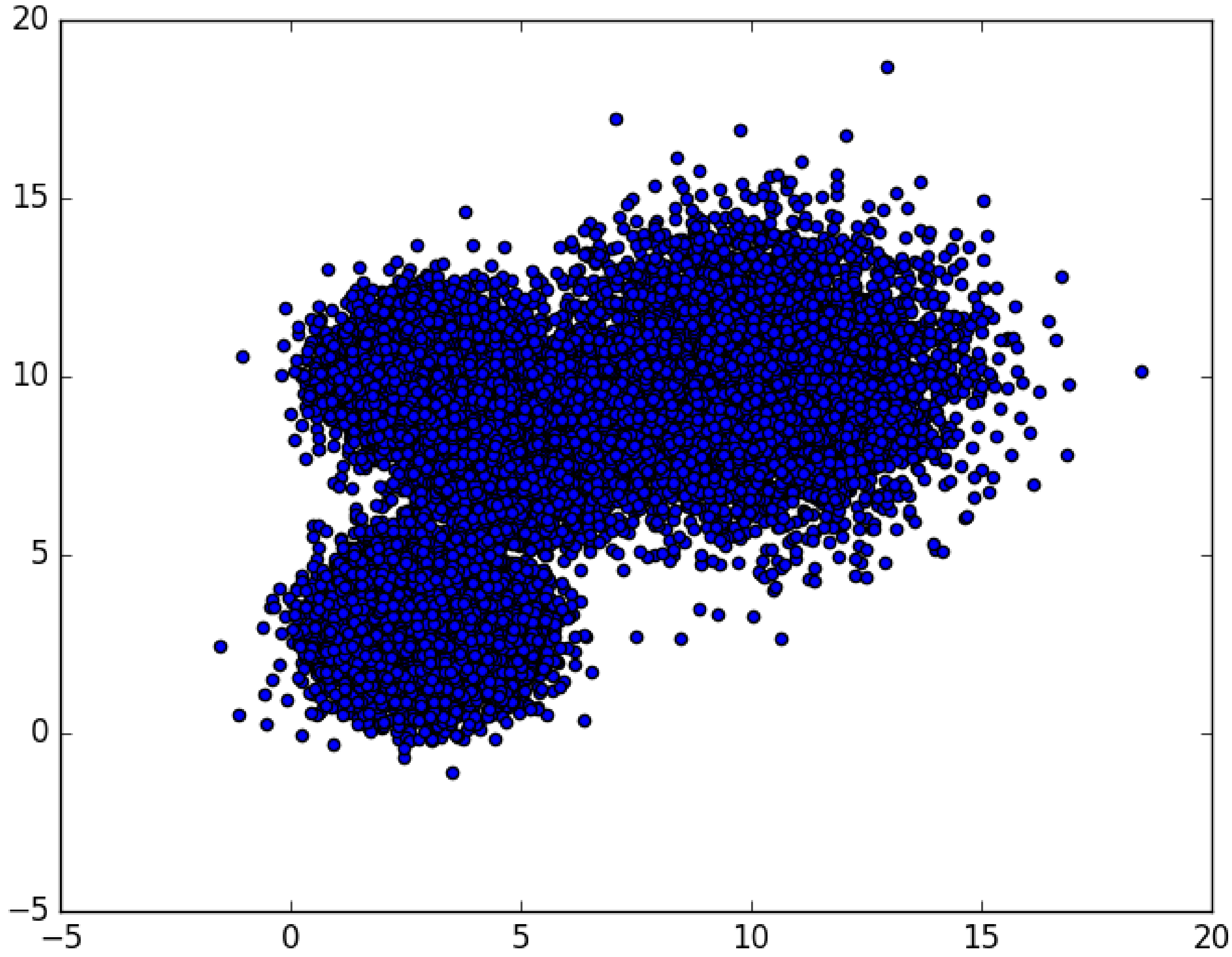
- Common approaches to clustering
  - Partitioning: k-means clustering
  - Density clustering: dbscan

- Density peak clustering
  - The basic algorithm
  - The improved algorithm: topography of a data landscape

- Applications
  - A blind test: states of firing network
  - (Application of DPC to fMRI data)

A toy example

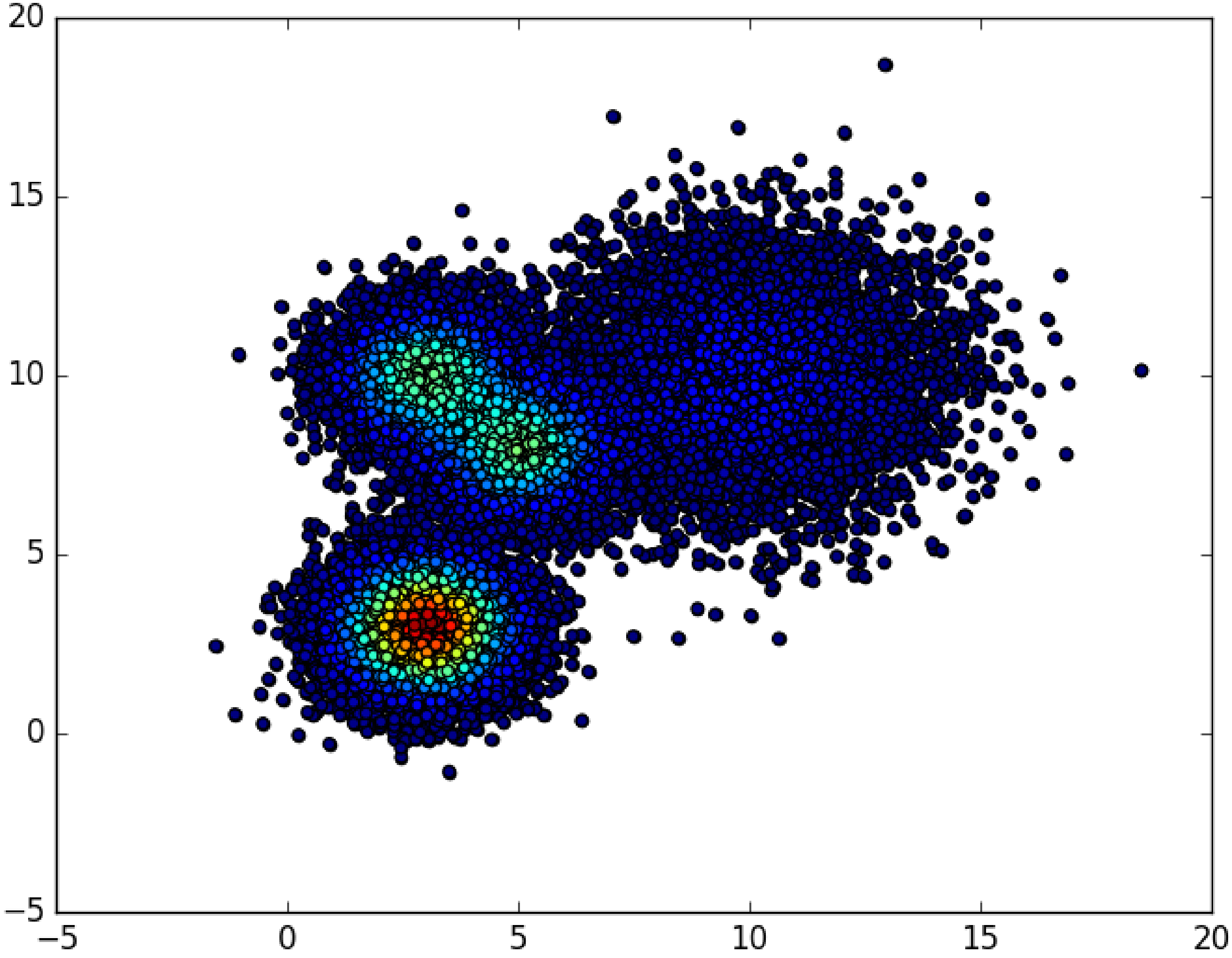
How many clusters are there?



Random points from four Gaussians with different locations and variances

A toy example

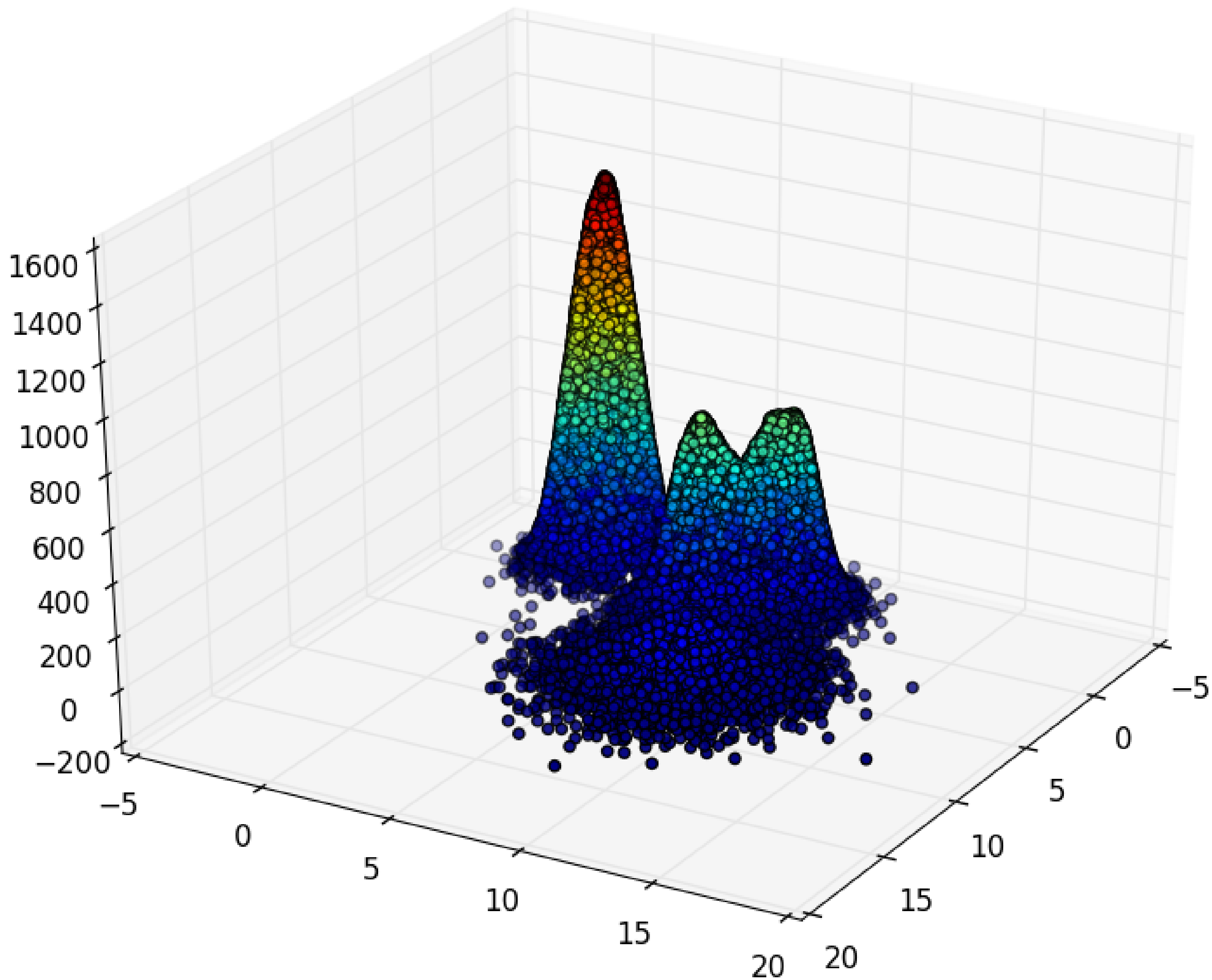
Random points from four Gaussians with different locations and variances



There are four clusters

A toy example

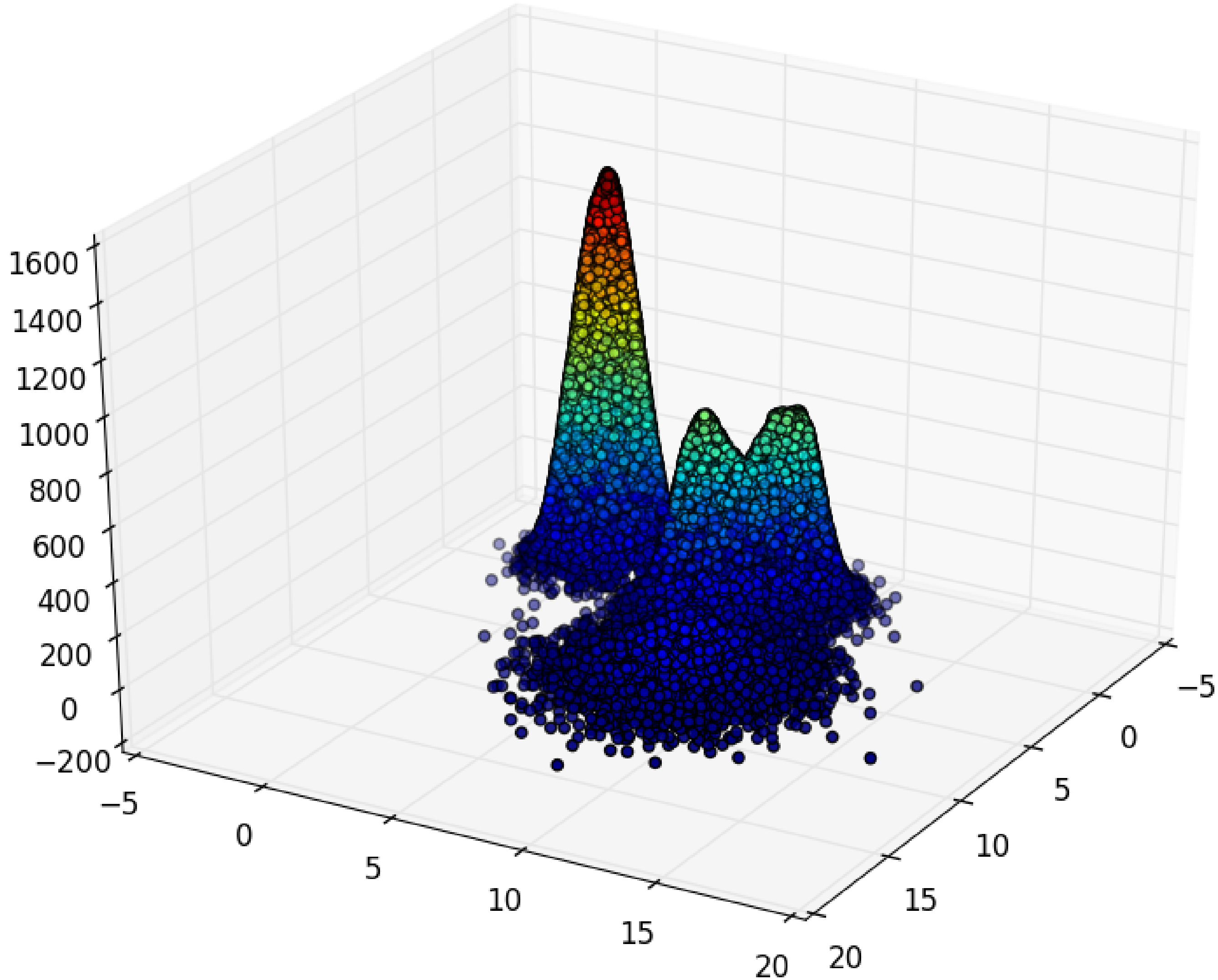
How many clusters are there?



There are four clusters (two are sub-clusters of larger cluster)

A toy example

How many clusters are there?



The “correct” number of clusters depends on the scale

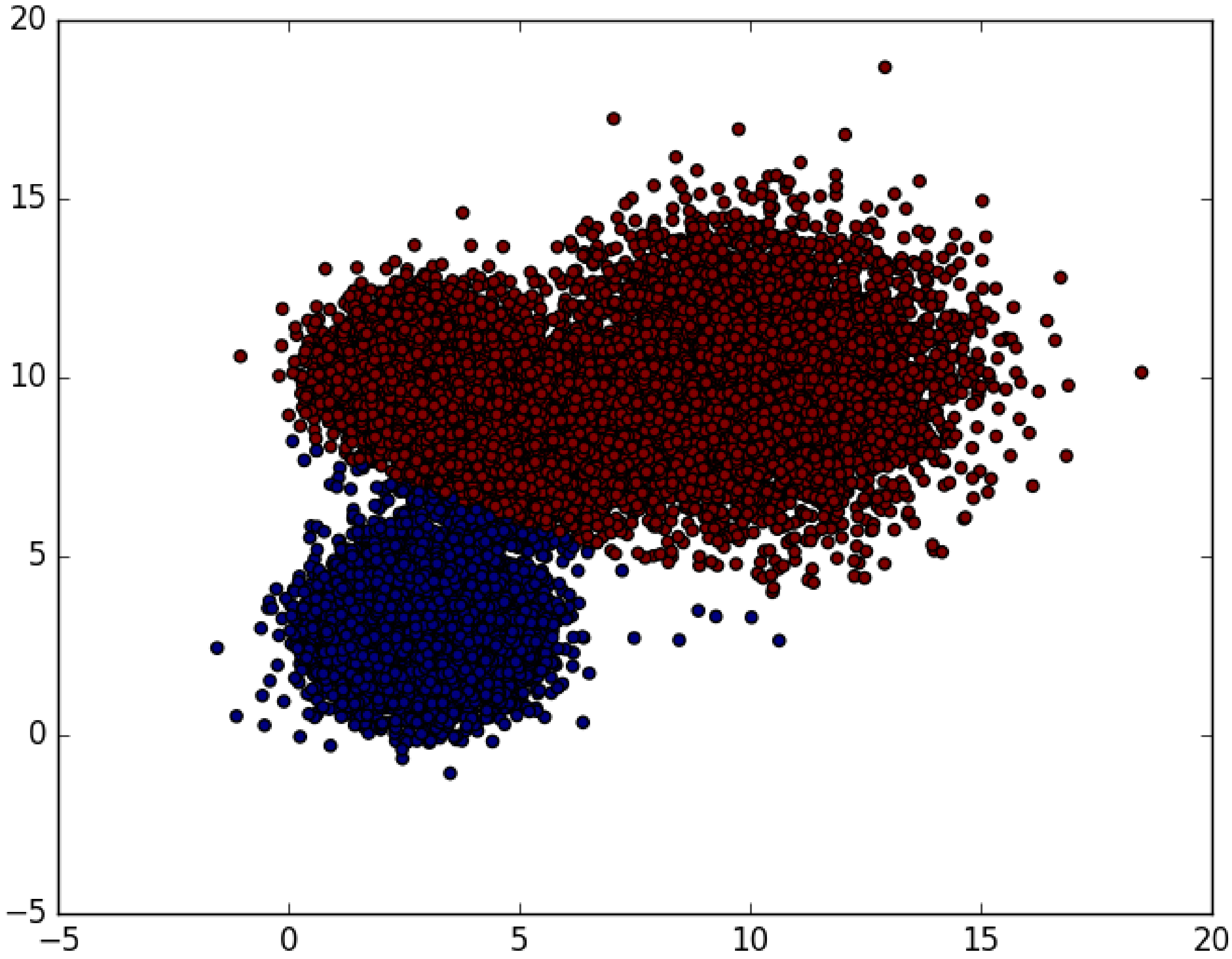
K-means clustering

- The most popular clustering algorithm
- Dates back to more than 50 years ago
  (J. B. MacQueen (1967), Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability)
- A partition algorithm: divides all points into K sets such that points in each set are close to the set mean
- In brief:
  - initialize randomly all points in K clusters, then iterate:
  - assignment step, reassign a point to cluster whith closest mean
  - update step, recompute set mean
- Simple and effective

K-means clustering

- Three main limitations:

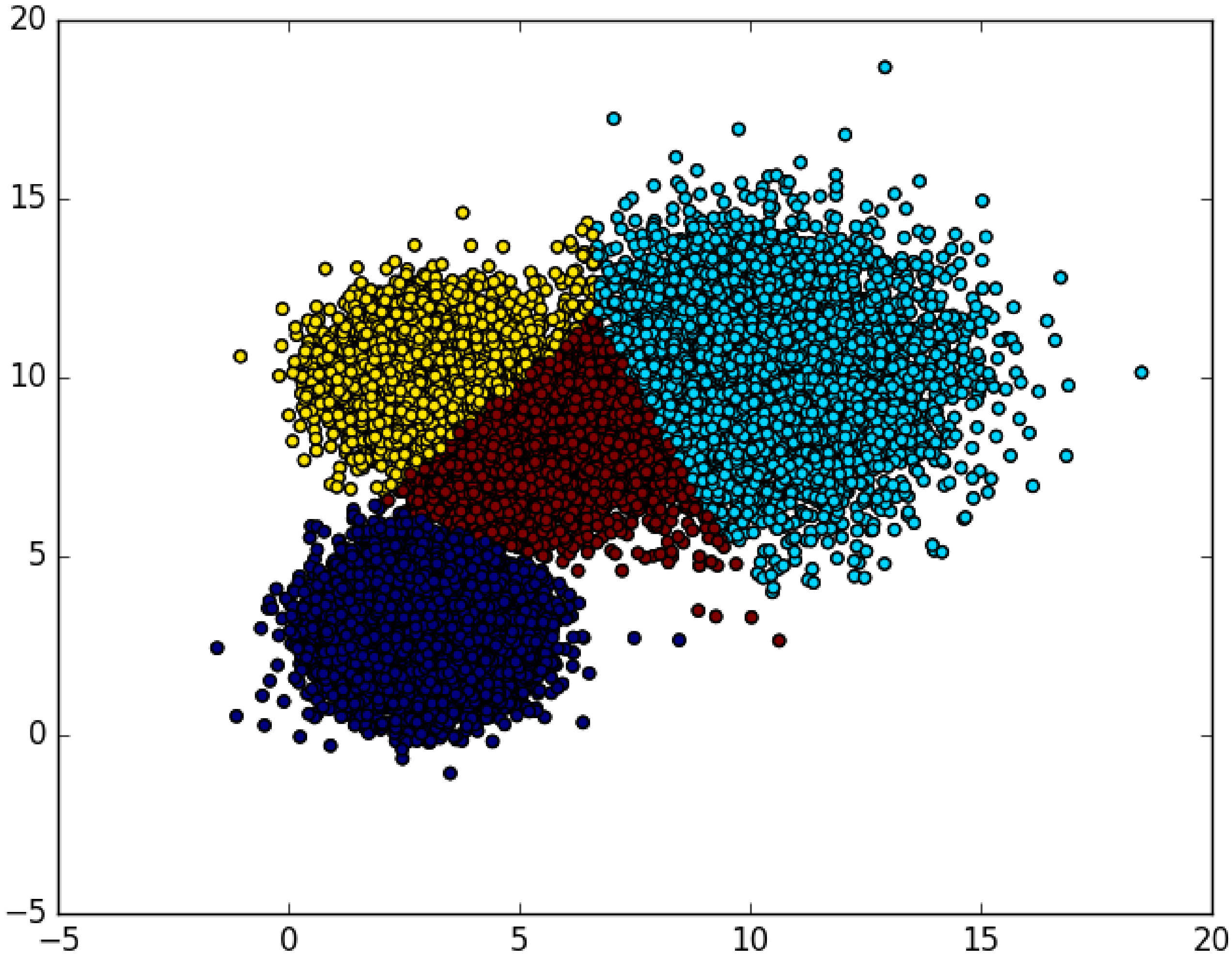
- 1) What is the right K? Free parameter selected by the user



K=2

K-means clustering

- Three main limitations:
- 1) What is the right K? Free parameter selected by the user

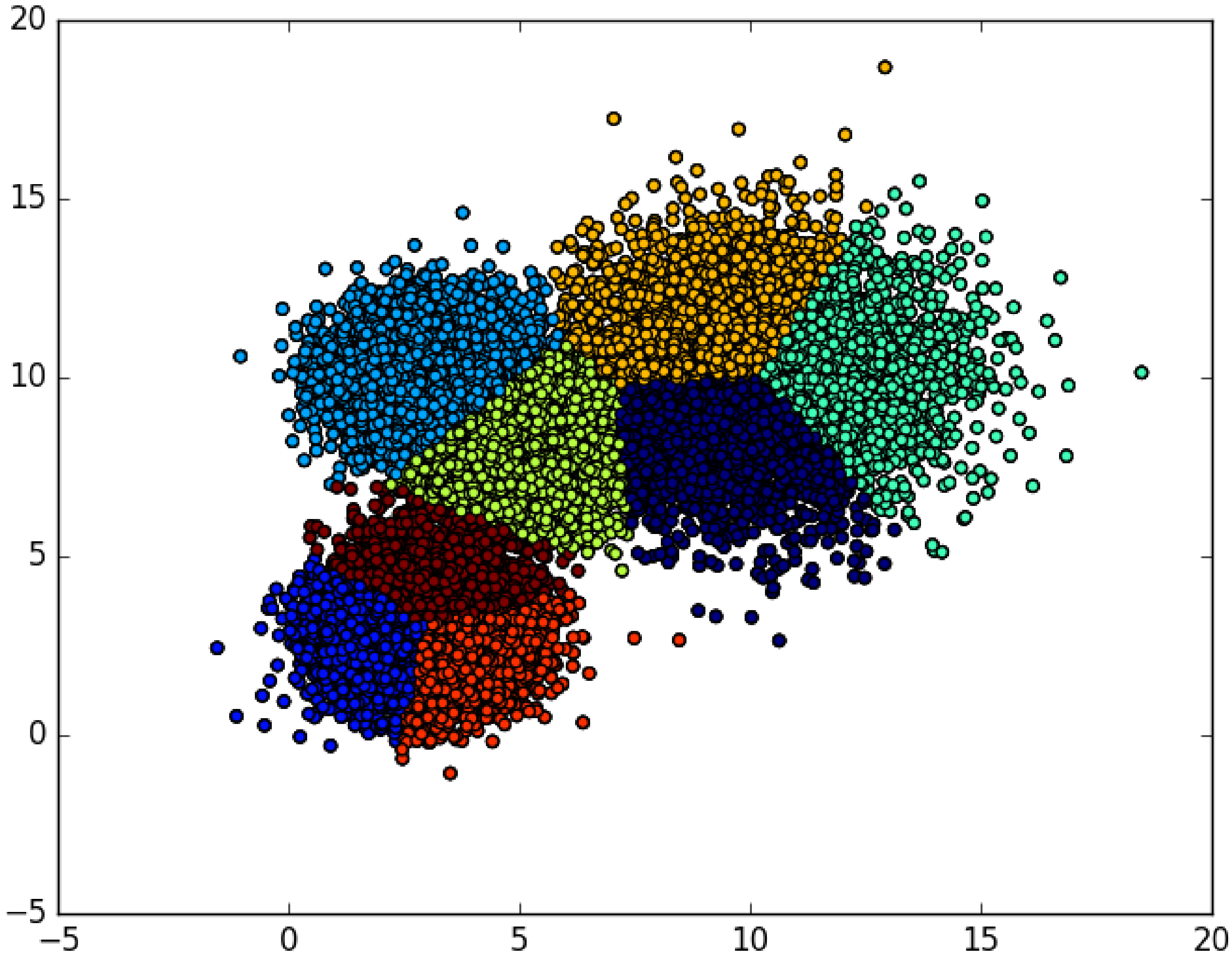


K=4

K-means clustering

- Three main limitations:

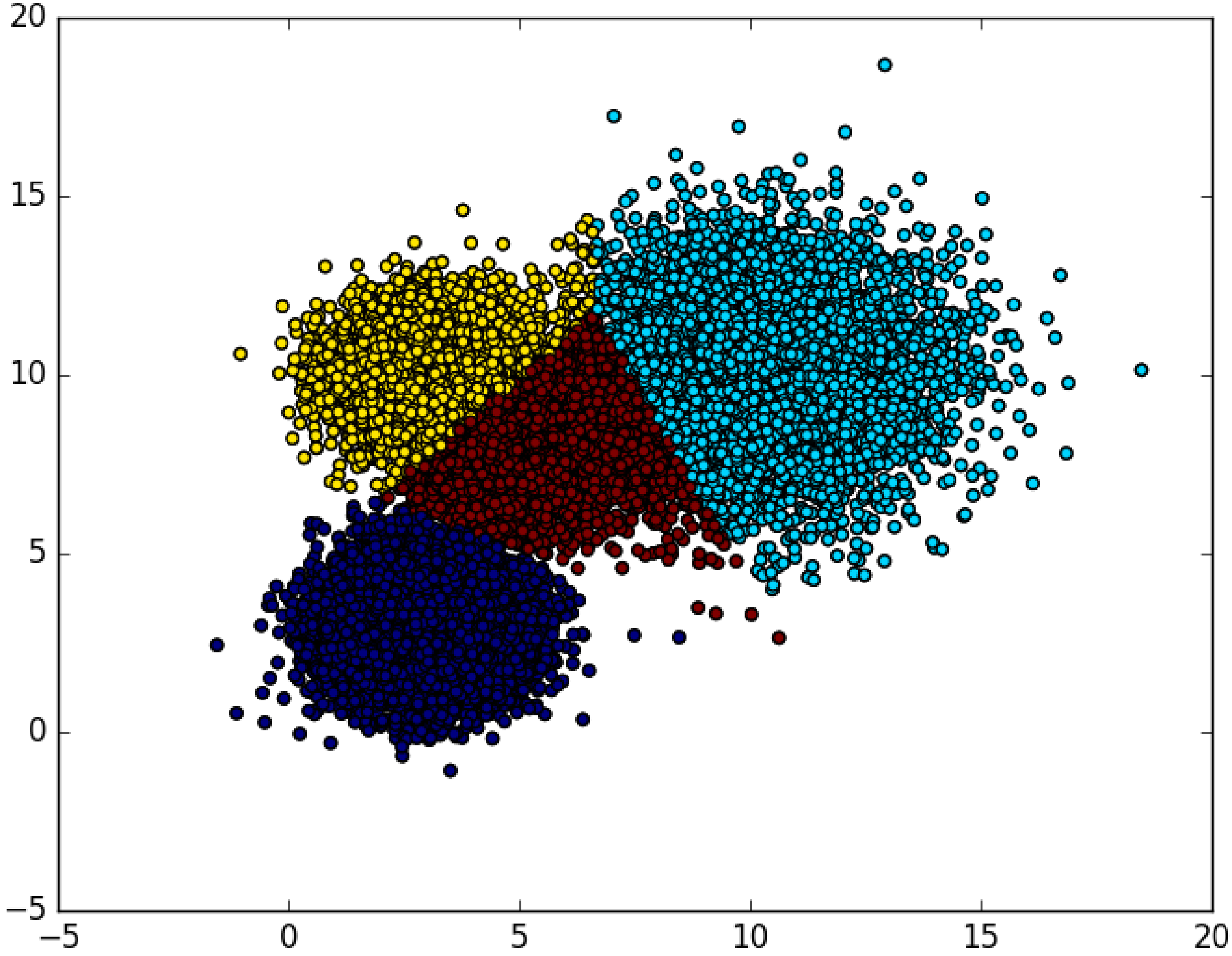
- 1) What is the right K? *Free parameter* selected by the user



K=8

K-means clustering

- Three main limitations:
- 2) By construction, it can only find convex (ideally spherical) clusters



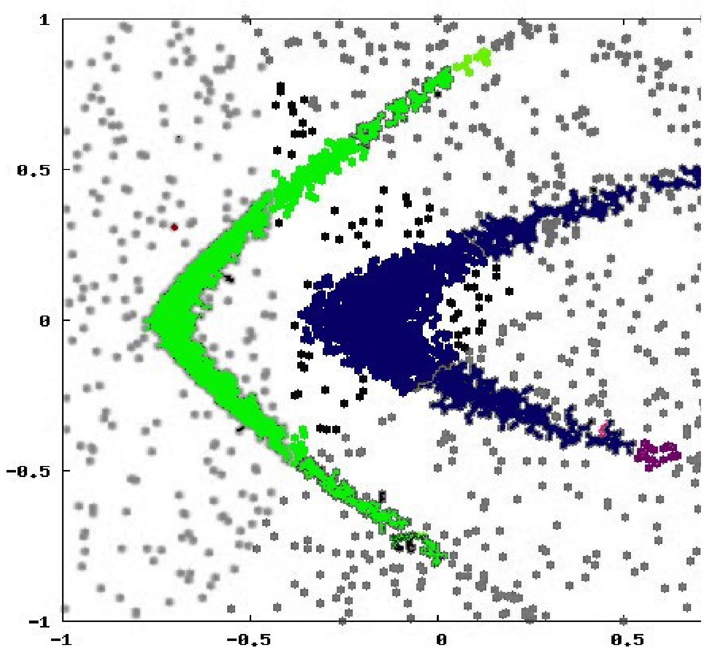
Here OK...

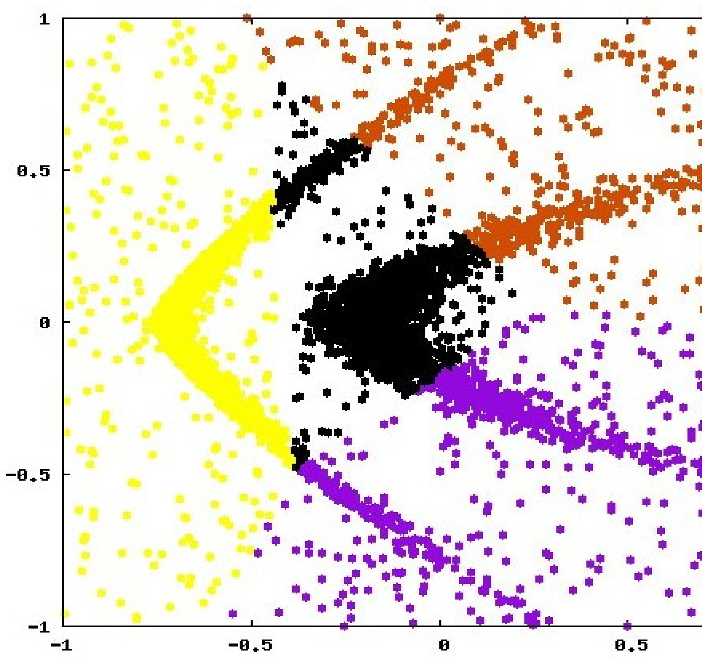
K-means clustering

- Three main limitations:

- 2) By construction, it can only find convex (ideally spherical) clusters

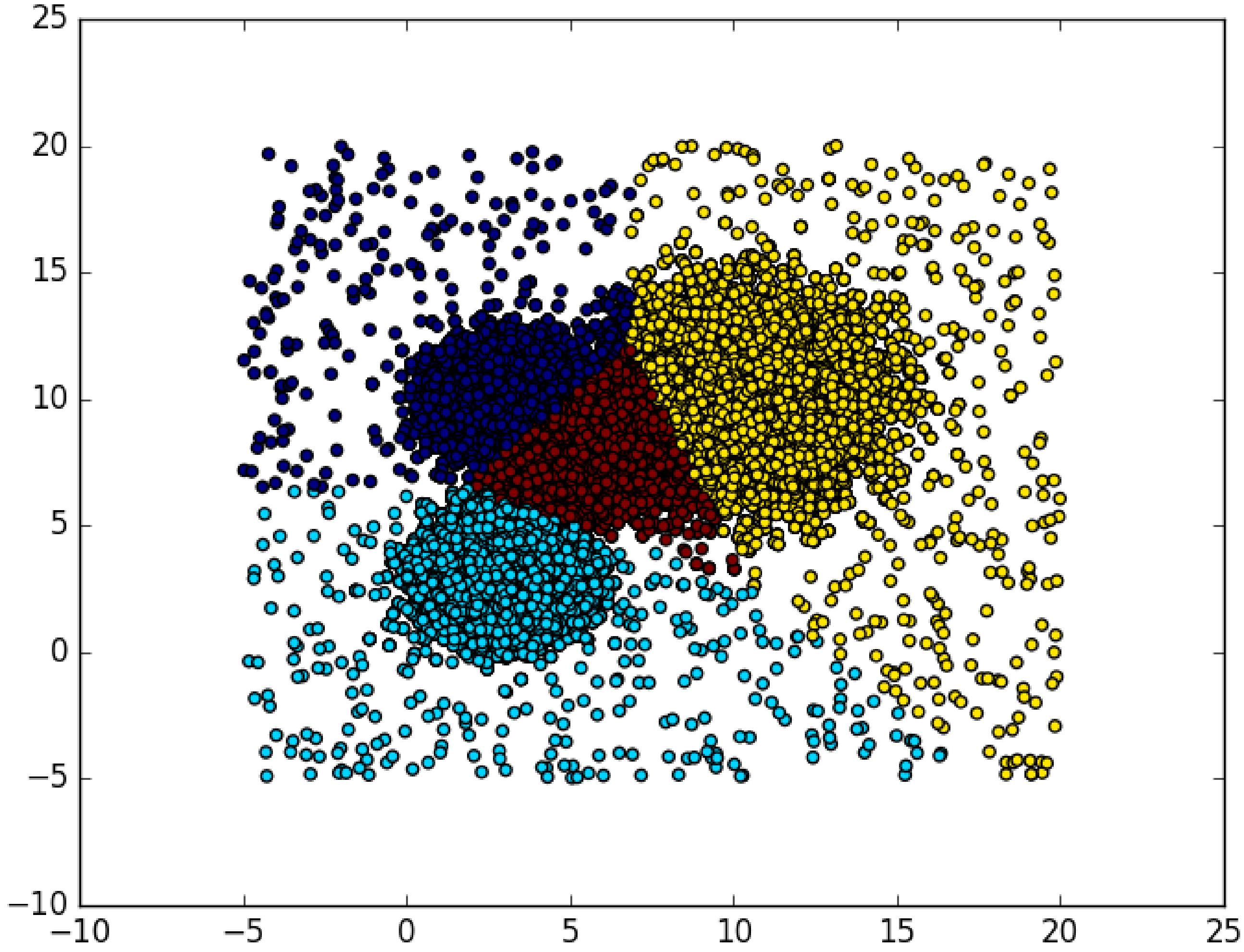
Here
KO...





K-means clustering

- Three main limitations:
- 3) no filtering, has trouble with noise and outliers



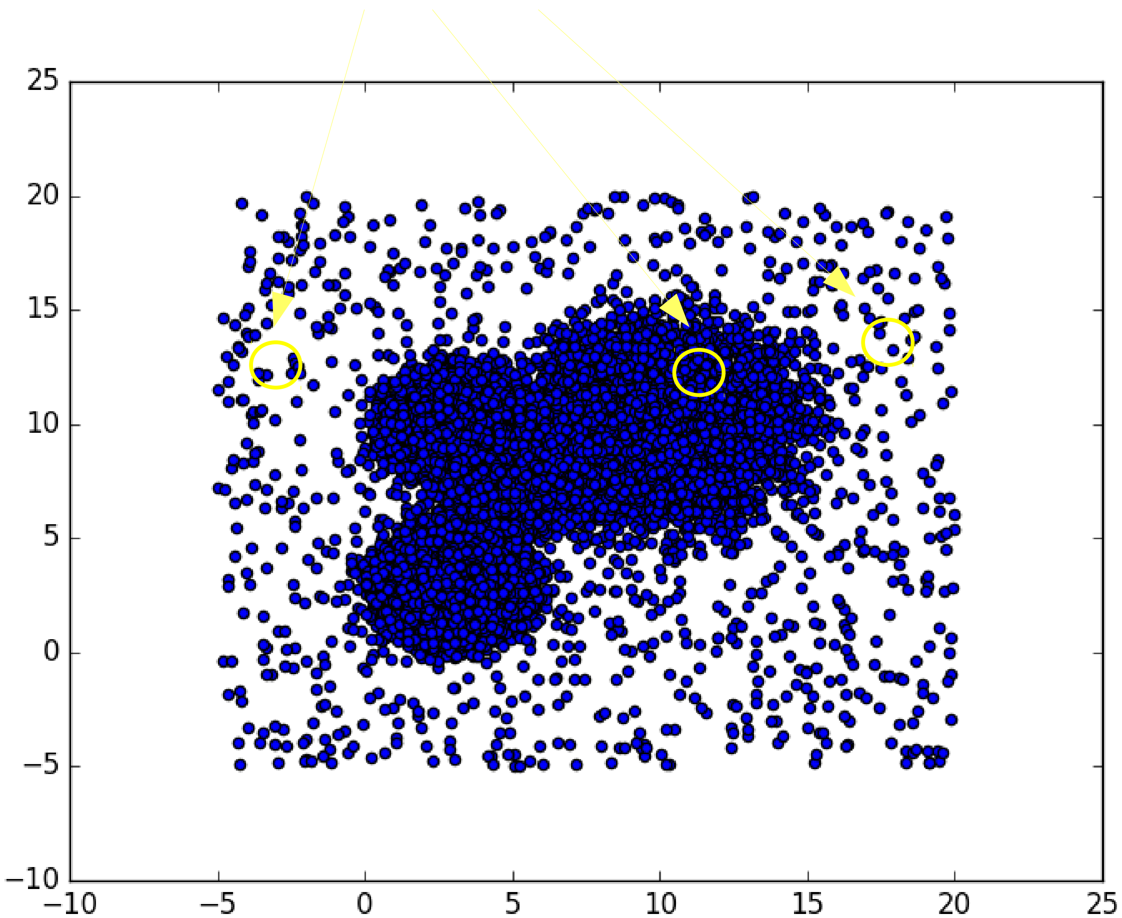
Also noise is clustered

Density clustering: dbscan

- A widely employed alternative to Dbscan
  (M. Ester et al. (1996) Proc. of the II International Conference on Knowledge Discovery and Data Mining)
- Density-based algorithm: identifies clusters as regions of high density
- Offers solution to limits of K-means:
  - No need to specify K
  - Finds clusters of arbitrary shape
  - Filters out noise

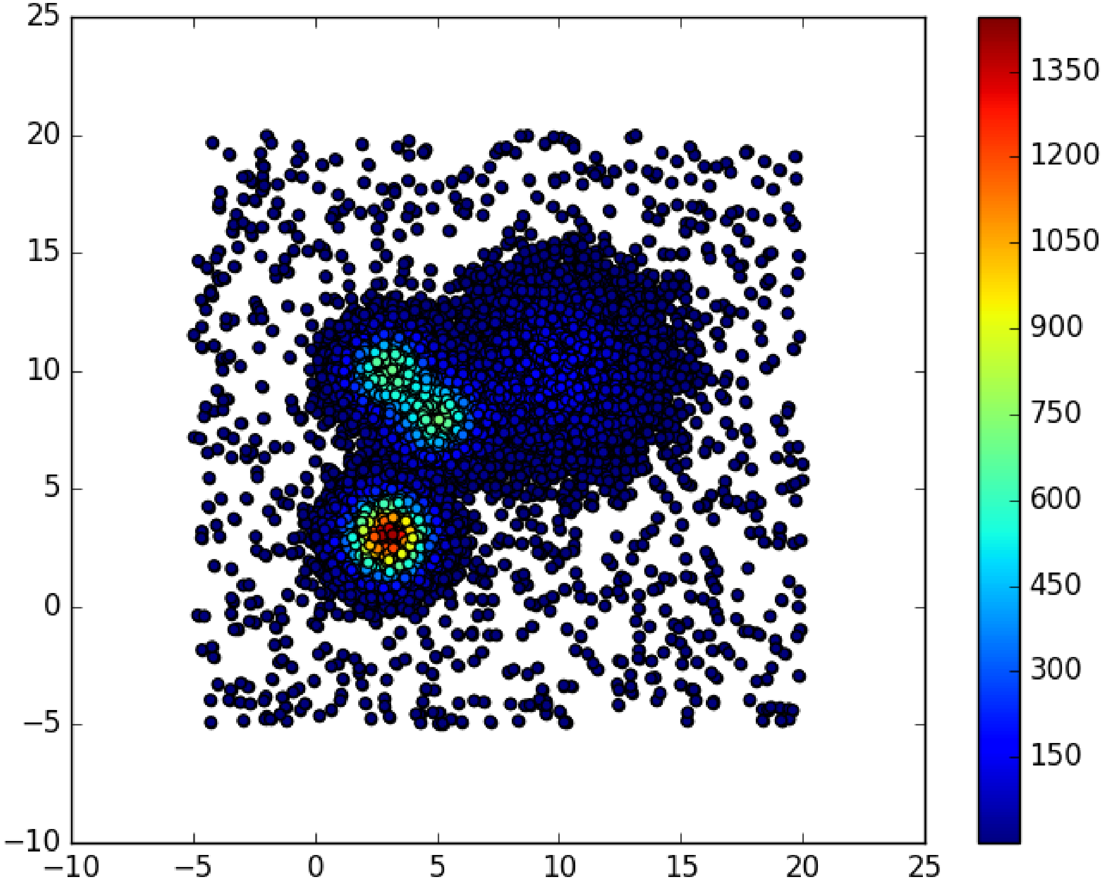
Density clustering: dbscan

- In brief:
  - estimate **density ρ** at each point:
  - count number of points within a ball of radius ε centered on the point (ε-neighbors)



Density clustering: dbscan

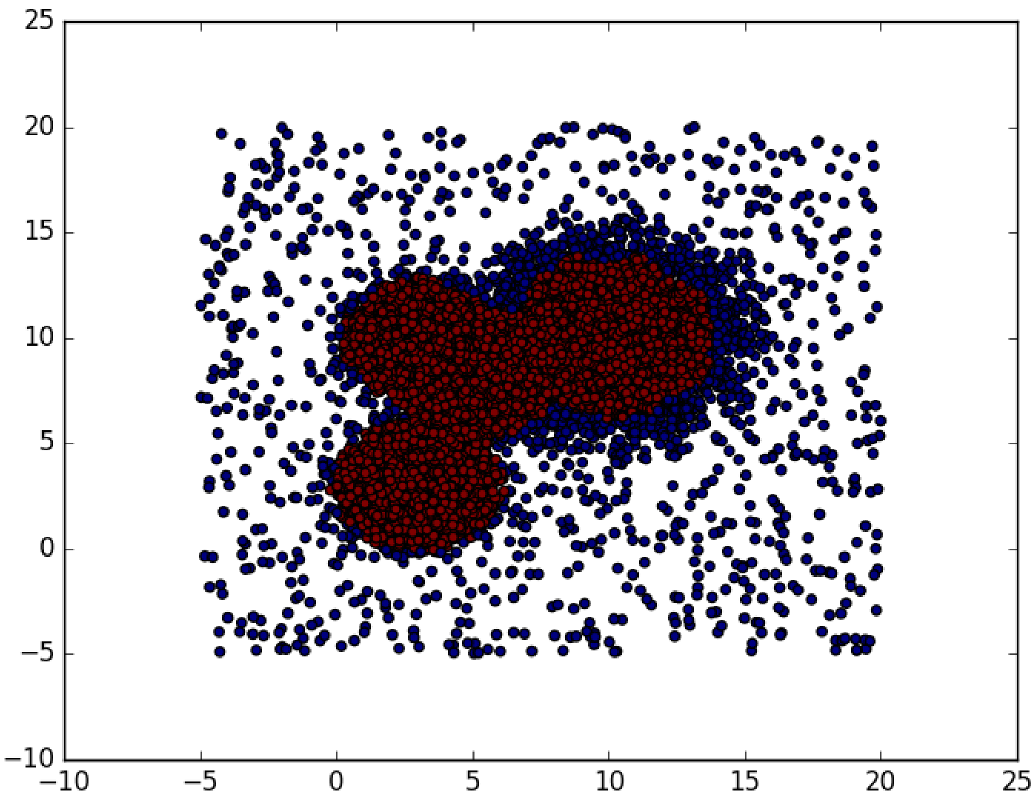
- In brief:
  - estimate **density ρ** at each point:
  - count number of points within a ball of radius ε centered on the point (ε-neighbors)



ε=0.5

Density clustering: dbscan

- In brief:
  - Identify clusters as connected regions **above density theshold, $\rho > \rho_0$**
  - Remaining points are discarded as noise



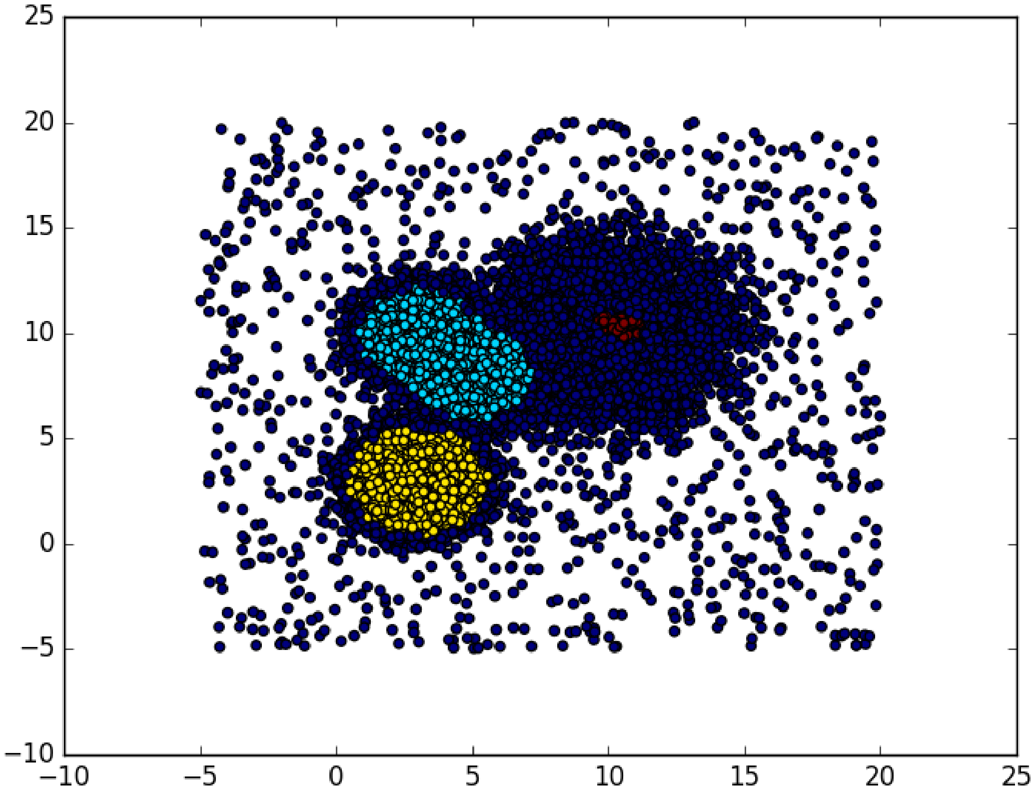
ε=0.5

ρ_0=50

1 cluster

Density clustering: dbscan

- In brief:
  - Identify clusters as connected regions with $\rho > \rho_0$
  - Remaining points are discarded as noise



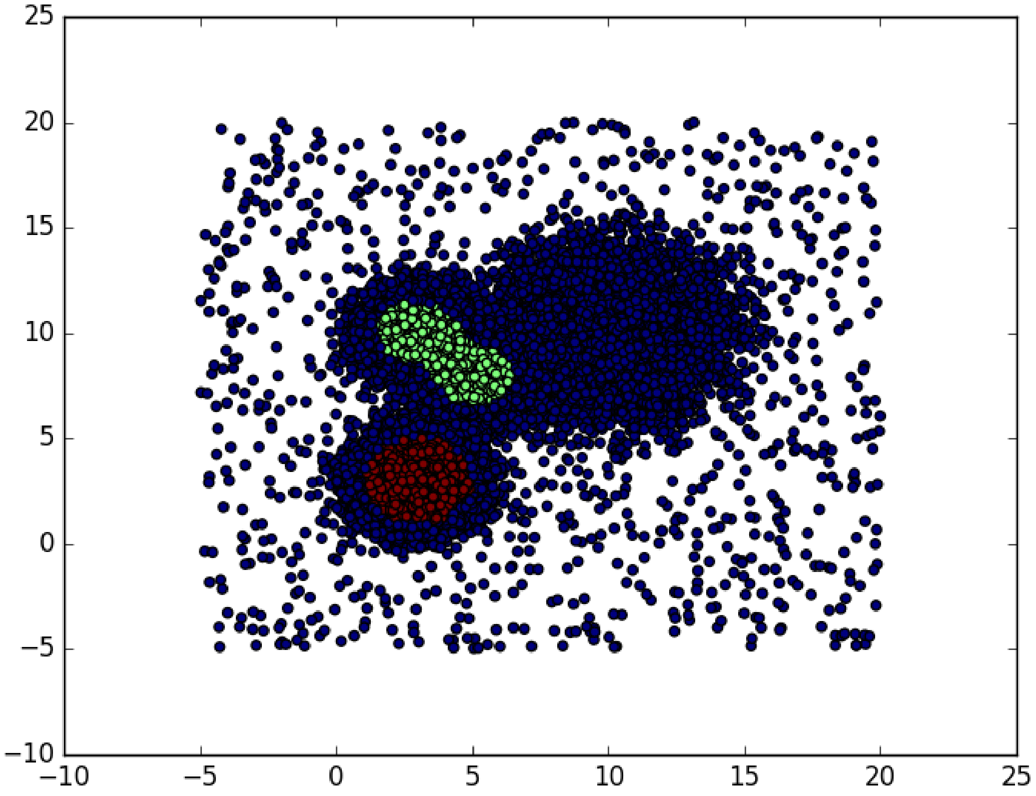
ε=0.5

ρ_0=200

3 clusters

Density clustering: dbscan

- In brief:
  - Identify clusters as connected regions with $\rho > \rho_0$
  - Remaining points are discarded as noise



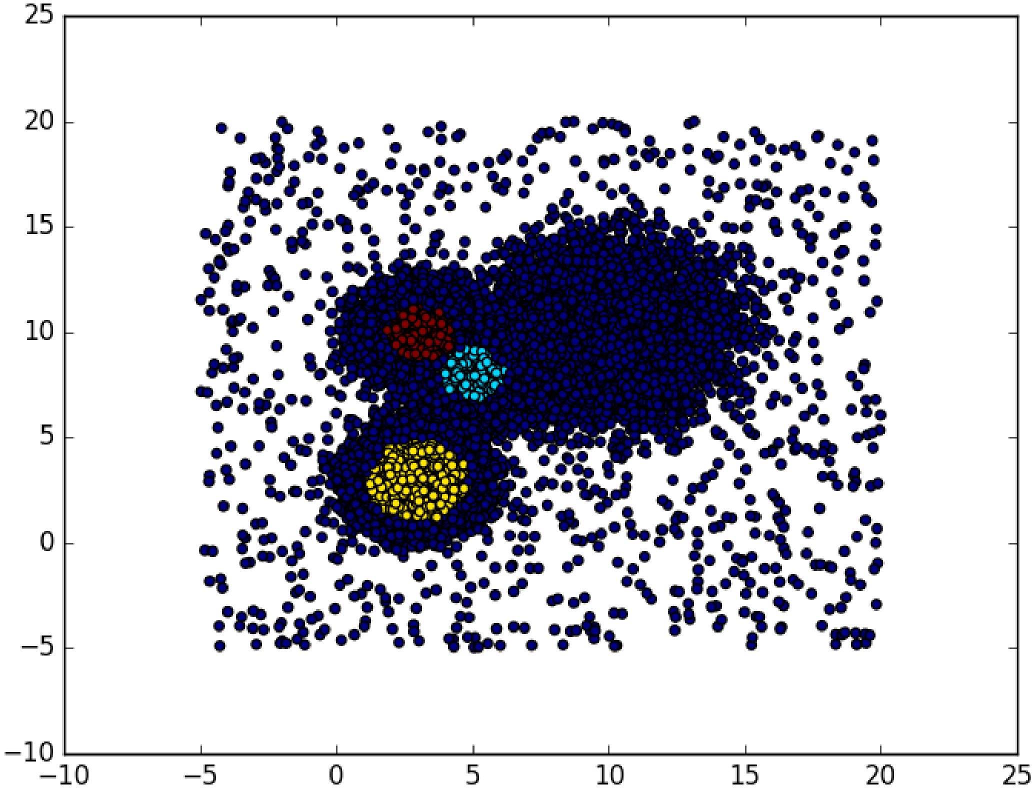
ε=0.5

ρ_0=500

2 clusters

Density clustering: dbscan

- Identify clusters as connected regions with $\rho > \rho_0$
- Remaining points are discarded as noise



ε=0.5

ρ_0=600

3 clusters

No threshold allows to somultaneously find all 4 clusters

Density clustering: dbscan

- Density-based algorithm: identifies clusters as regions of high density

- In brief:
  - estimate density ρ at each point as number of ε-neighbors
  - identify clusters as connected regions with $\rho > \rho_0$
  - remaining points are discarded as noise

- Offers solution to limits of K-means:
  - No need to specify K
  - Finds clusters of arbitrary shape
  - Filters out noise
  - Computationally light

- **A major limitation:**
  - **What is right density threshold ρ_0? Results strongly depend on the chosen threshold (and also on ε...)**
  - **Cannot resolve significant clusters at different density scales**

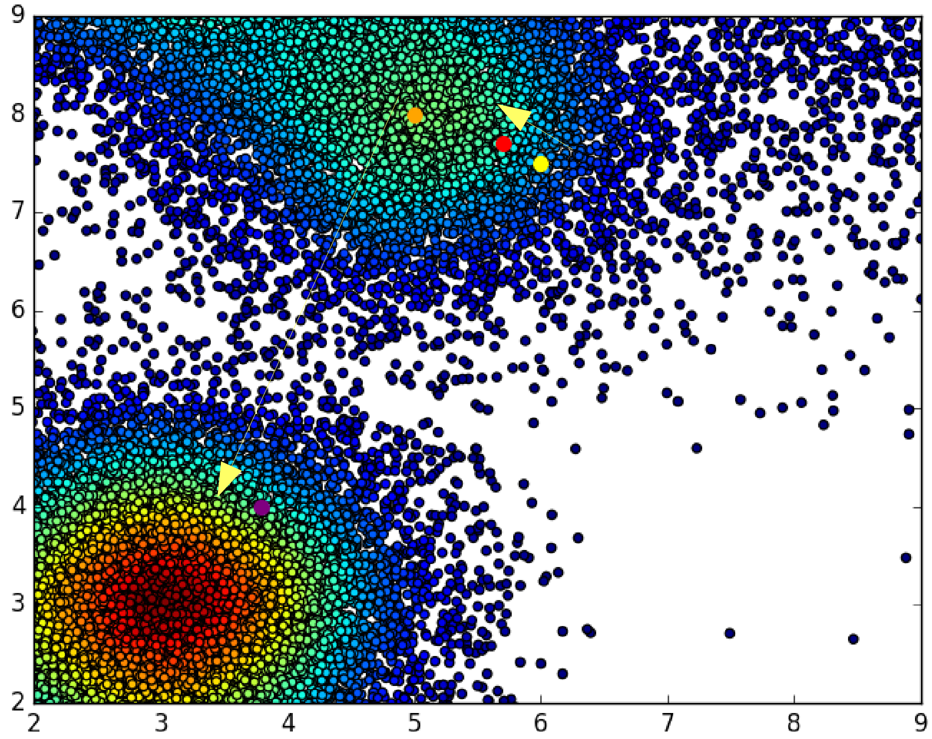
Density peak clustering

- Density-based algorithm: identifies clusters as regions around local maxima of the density
  (A. Rodriguez and A. Laio, Science, 2014, vol. 344, no 6191, p. 1492-1496).

- Offers solution to limitations of db-scan
  - **Does not fix threshold,** can resolve significant clusters at different density scales

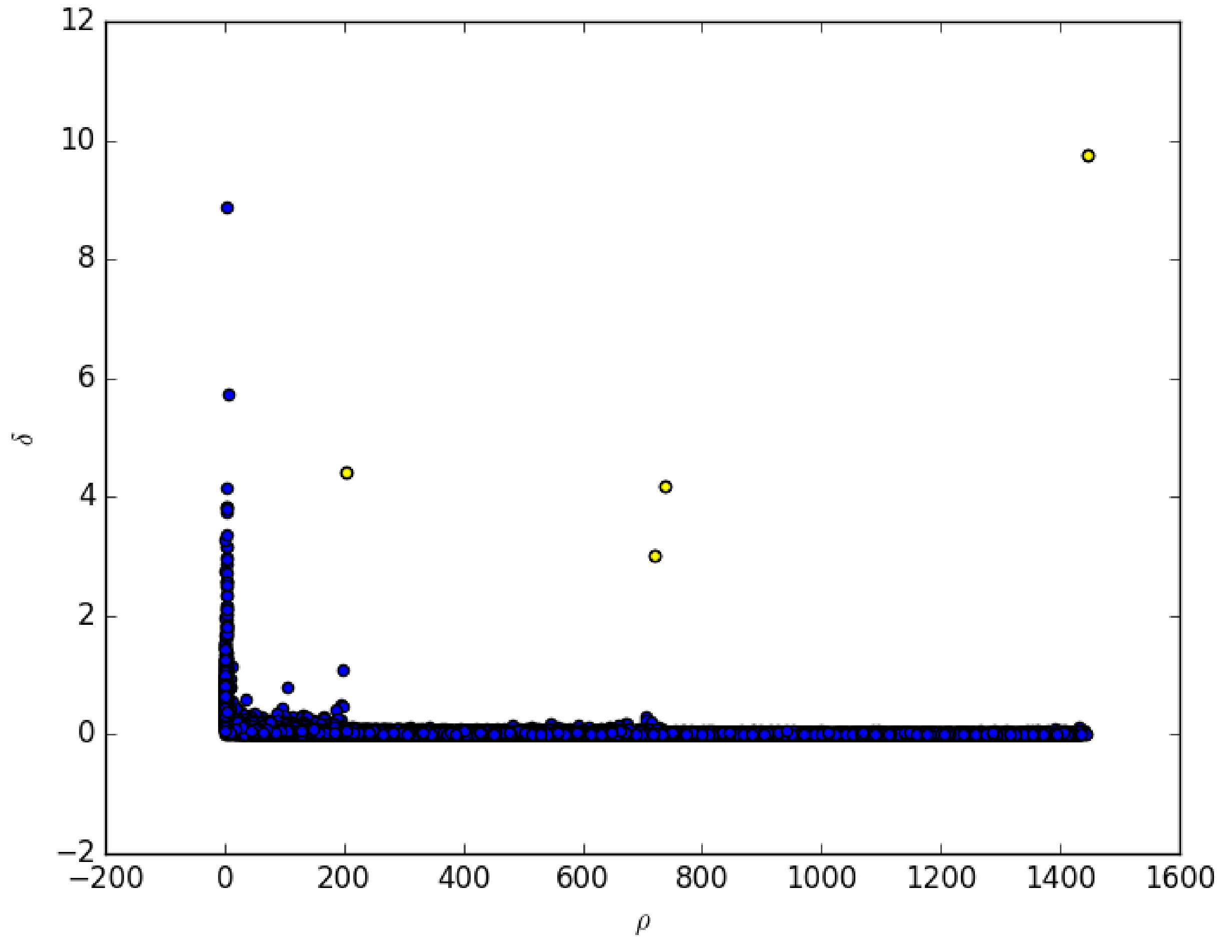
Density peak clustering

- Identify **local maxima** of ρ
- Local maxima are far from any other point with higher ρ
- For each point, compute *minimum distance from point of higher density*
  $\boldsymbol{\delta_i = \min(d_{ij} : \rho_j > \rho_i)}$



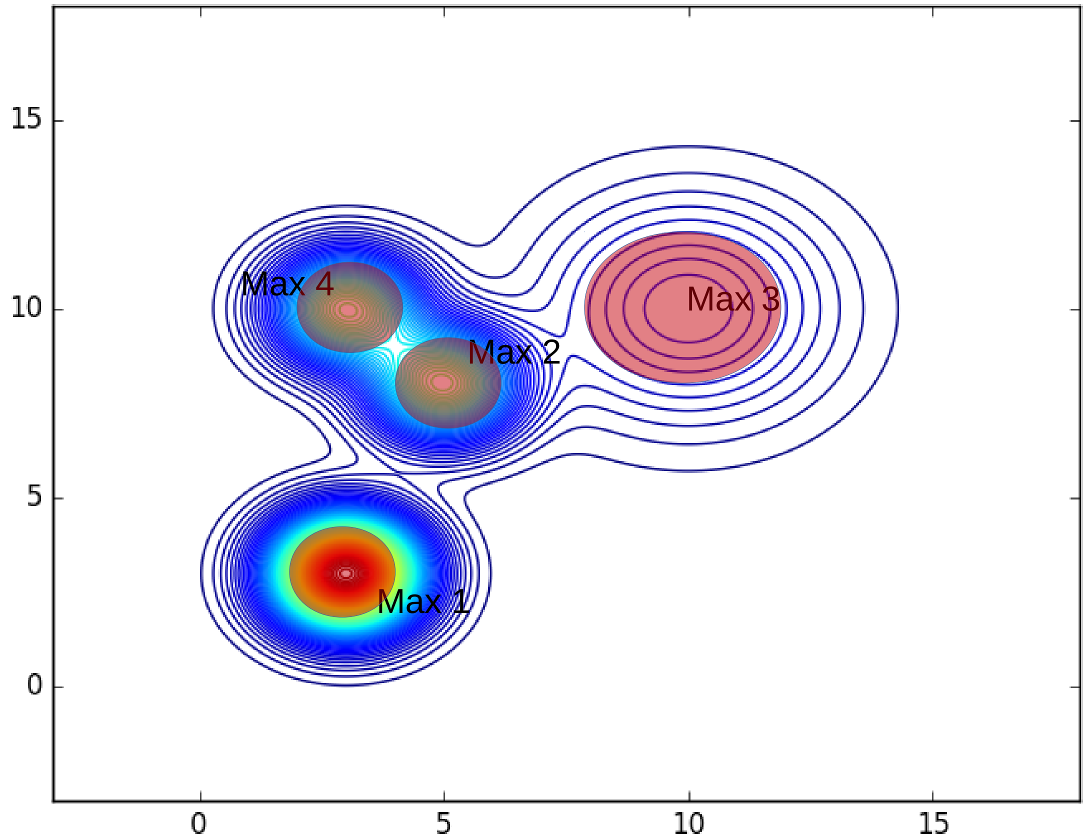
Density peak clustering

- δ_i is high only for local maxima
- identify density peaks as points with high δ
- they appear from **decision graph** ρ vs δ



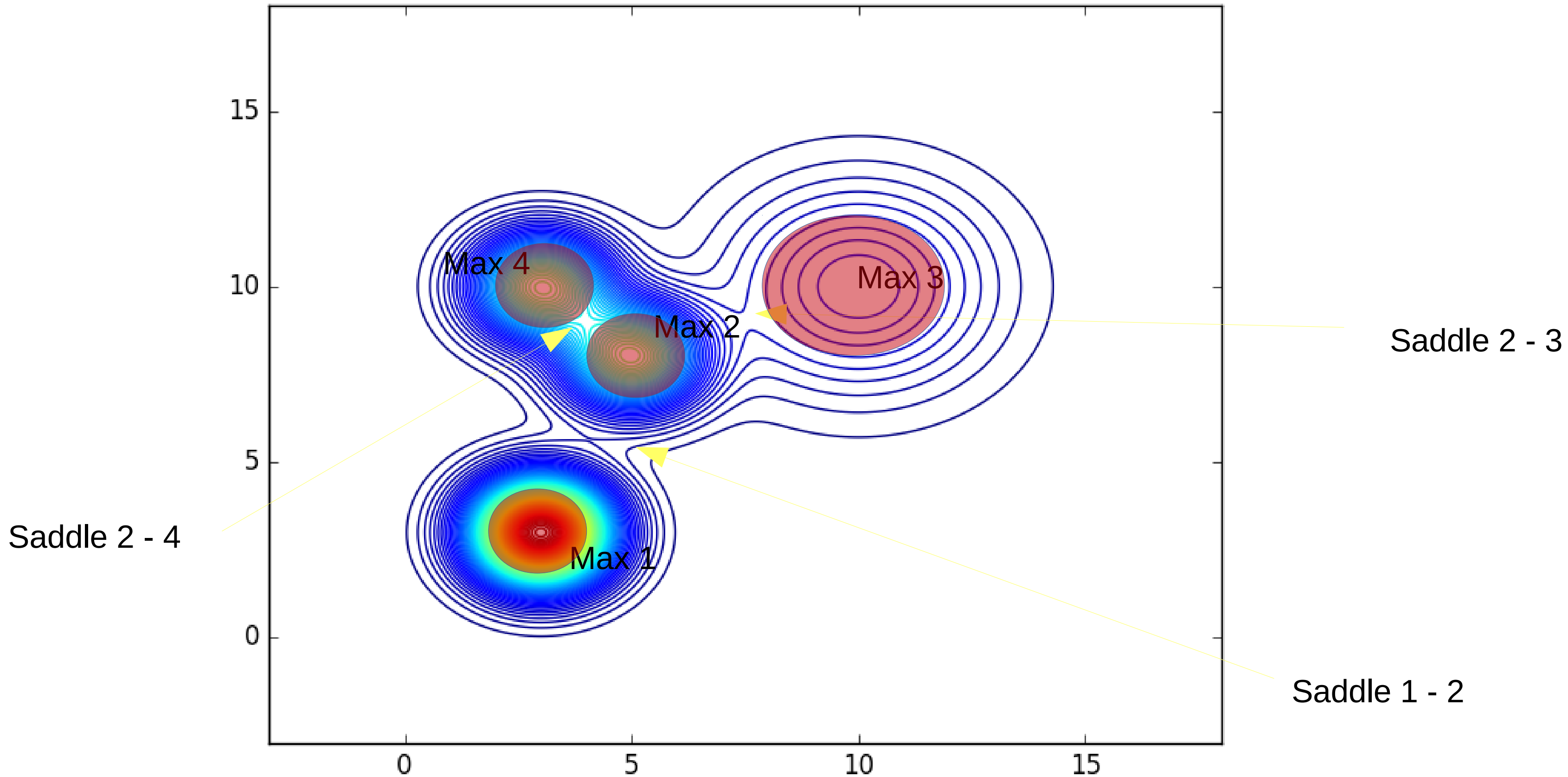
Density peak clustering

- **A topographic approach to define clusters**
- Clusters are defined as peaks correspoding to each local maximum
- Each peak corresponds to closed contour lines



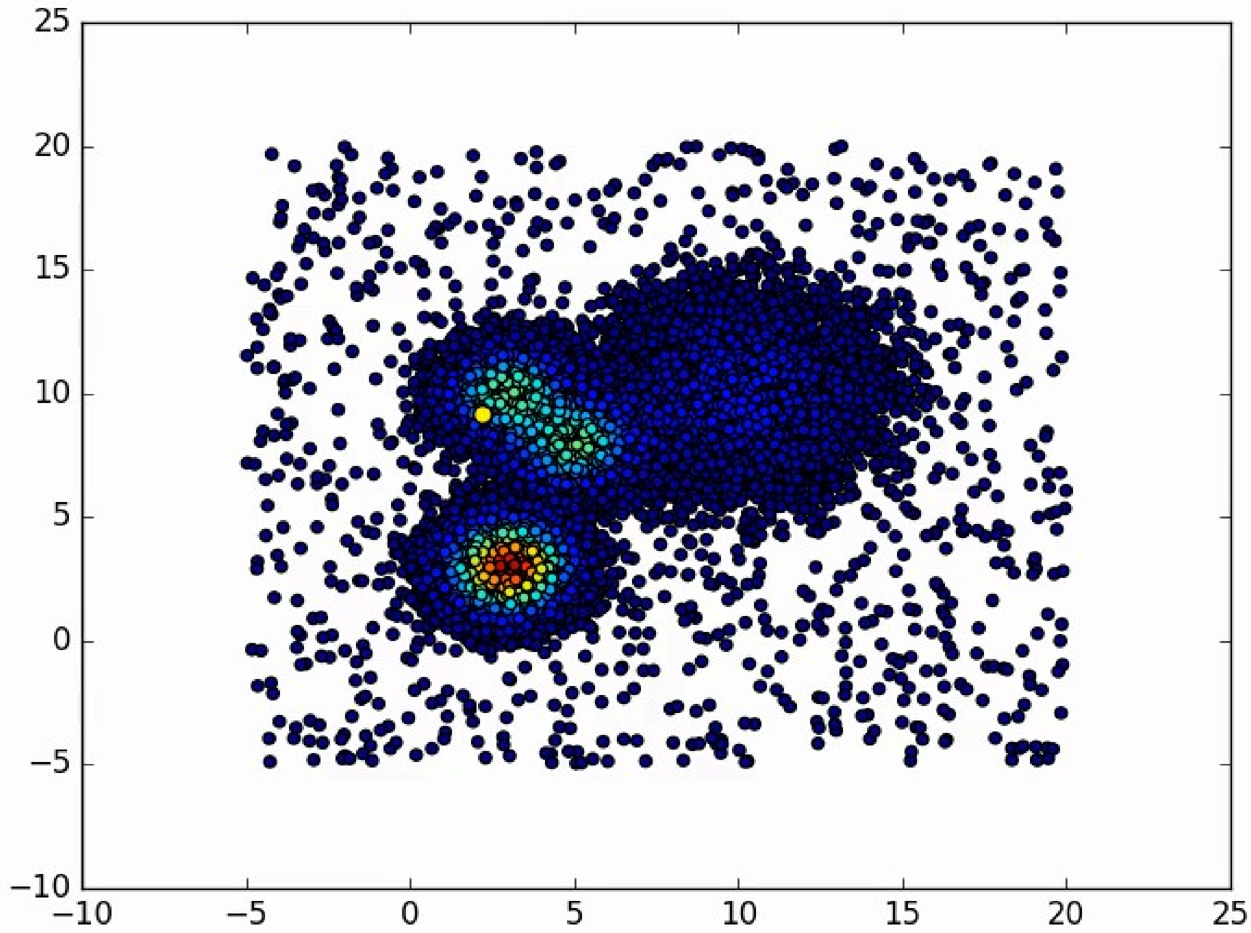
Density peak clustering

- First find **saddle points** between each pair of maxima
- For each maximum, the peak should include only points with ρ higher that the highest saddle point between the maximum and other maxima



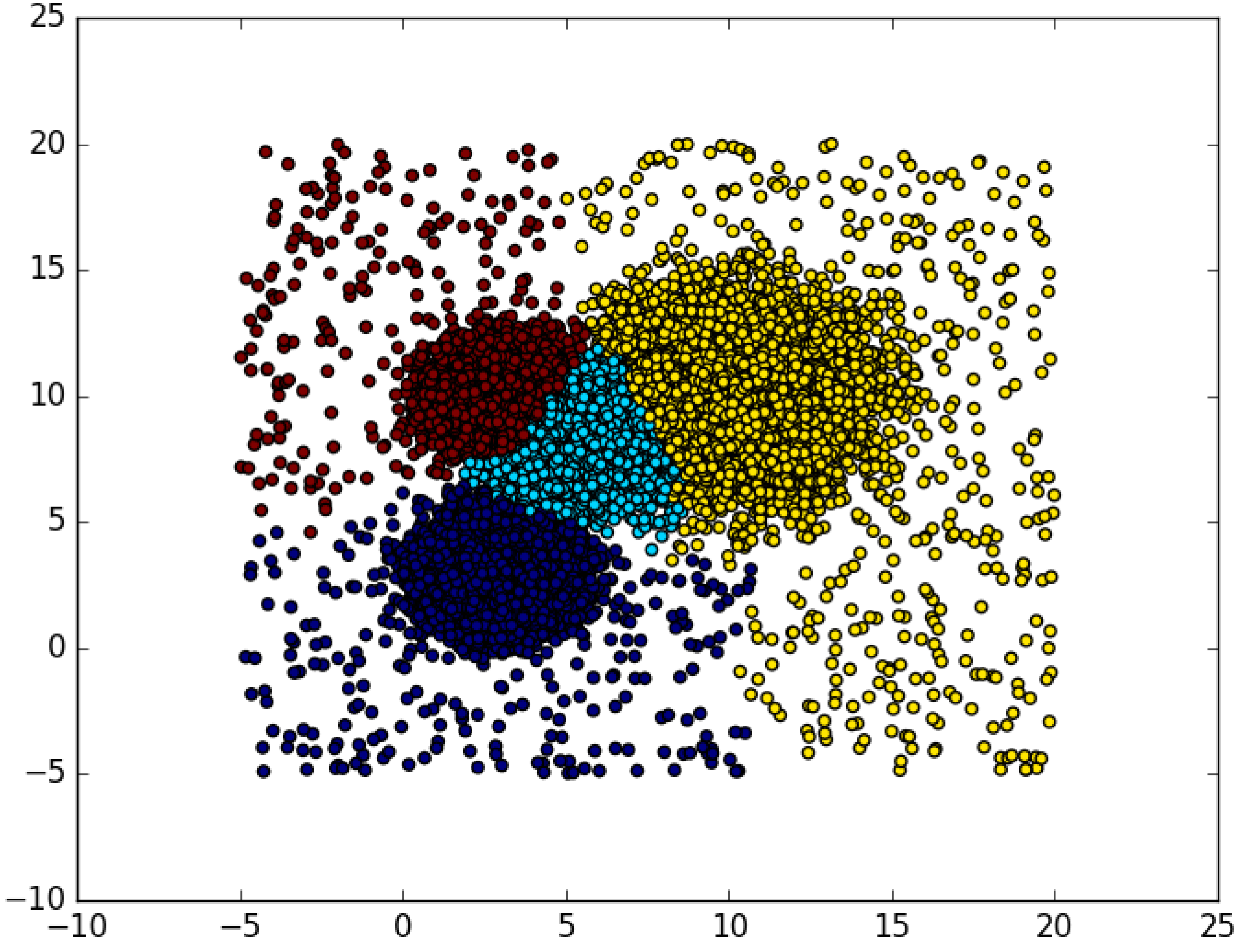
Density peak clustering

- Primary assignment: points assigned to a maximum by following a path of increasing density



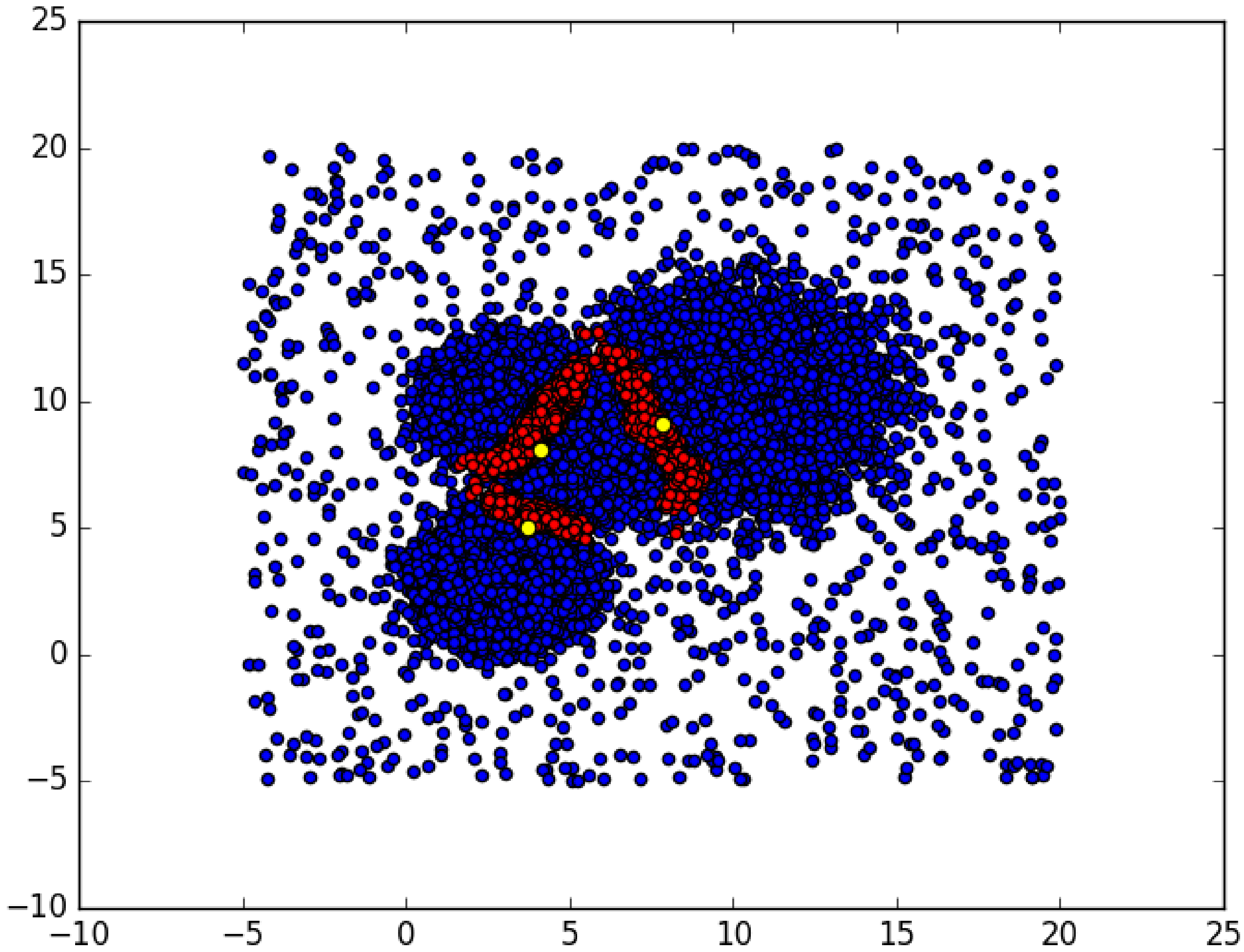
Density peak clustering

- Primary assignment: points assigned to a maximum by following a path of increasing density



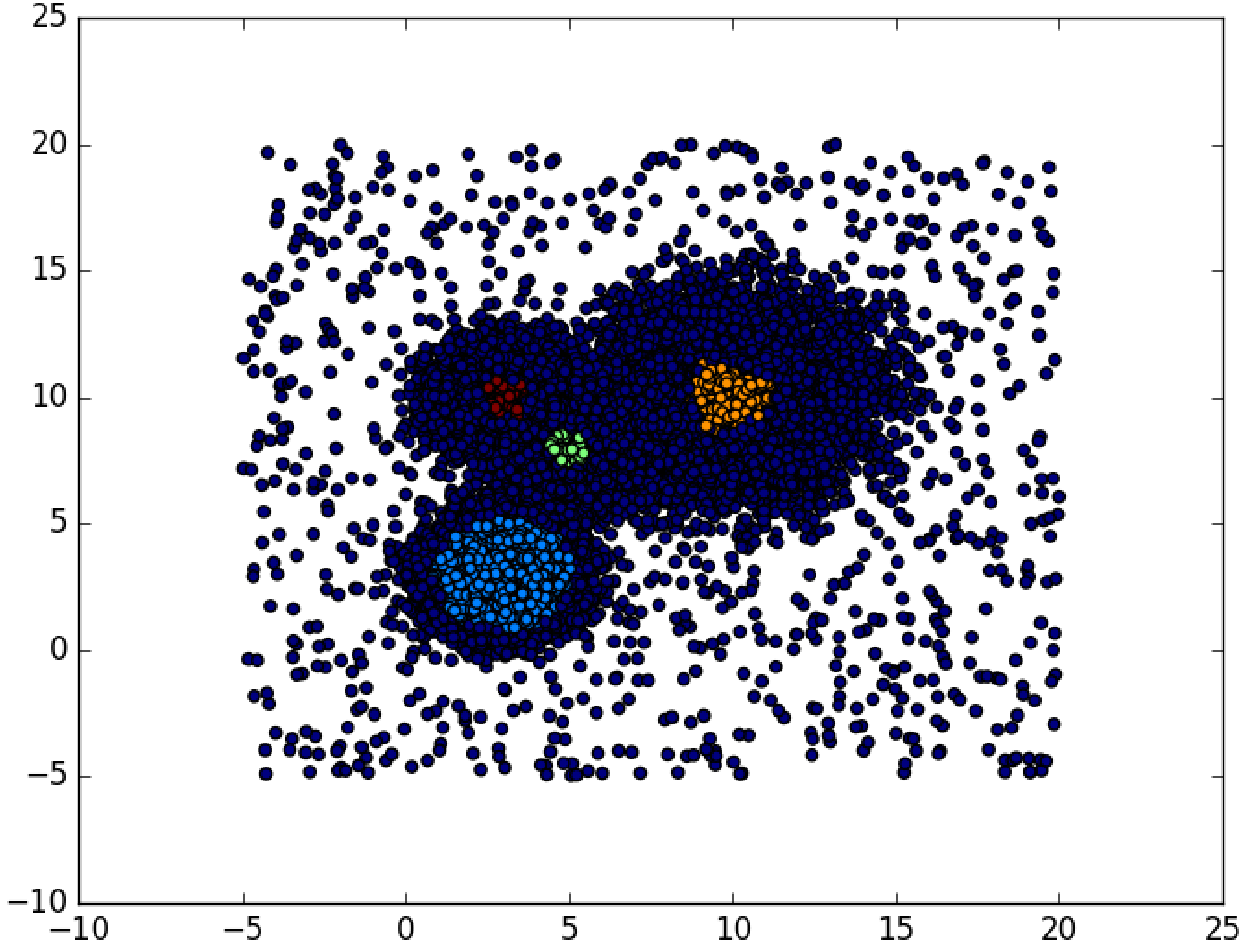
Density peak clustering

- Identify “borders” between clusters
- Find maximum density of border points: saddle point



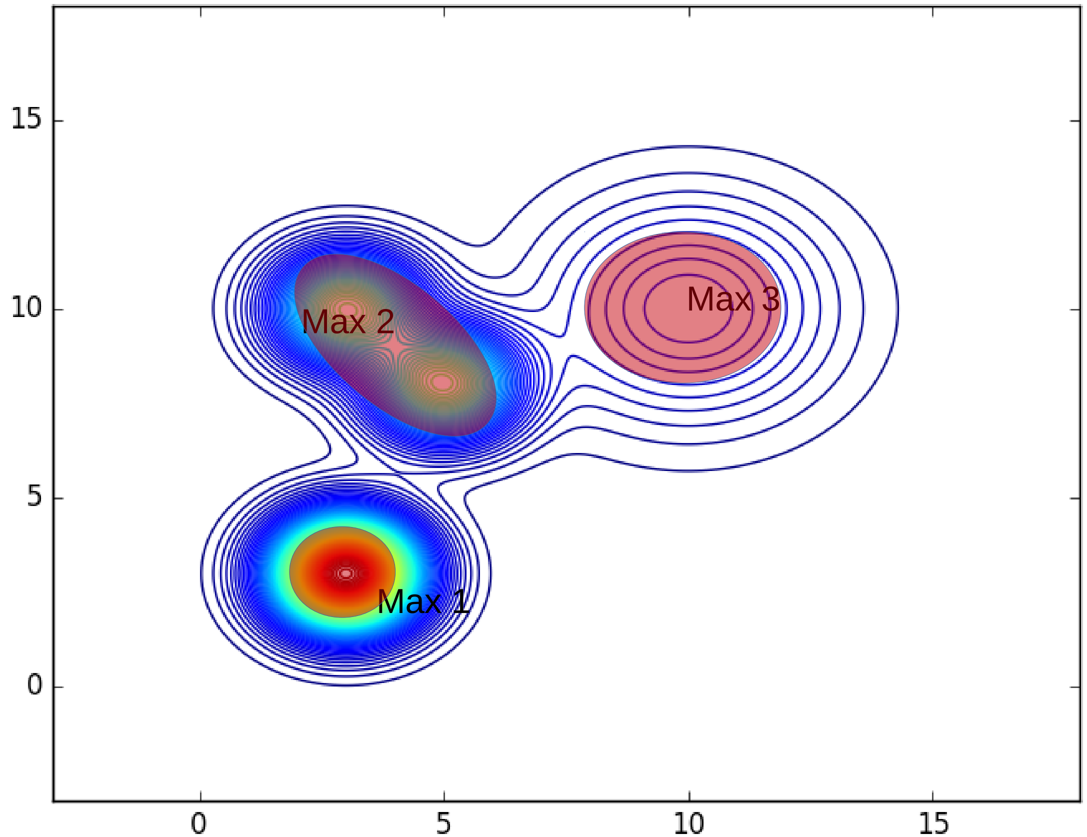
Density peak clustering

- "cut" clusters at border density
- retain **cluster cores**



Density peak clustering

- This gives topography at fine scale
- Does not allow to appreciate topography at coarse scale



Density peak clustering

- Density-based algorithm: identifies clusters as regions around local maxima of the density

- In brief:
  - estimate density ρ at each point as number of ε-neighbors
  - identify local maxima of ρ (points far from other points with higher ρ)
  - assign remaining points to one of the density peaks, keeping cluster cores

- Offers solution to limits of db-scan
  - Does not fix threshold and can find clusters at different density scales

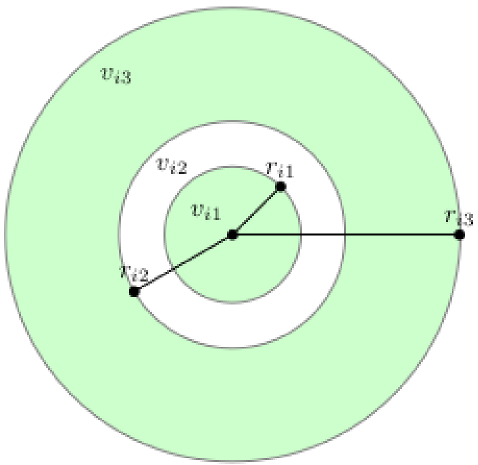
Main limitations:
- **Density estimation depends on free parameter ε**
- **Method depends on visual heuristics to identify number of clusters**
- **Yields topographic at very fine scale**

Improving density peak clustering

1) Make density estimation parameter free

- K-nearest-neighbor: Assume $\rho \approx$ const in small region ar
- For each point *i*, consider its *k* nearest neighbors at distances $r_{i1}, r_{i2}, r_{i3}, \ldots$
- density= *k*/volume of sphere containing the *k* points

$$\rho = \frac{k}{V_{ik}} \qquad \delta\rho = \frac{\sqrt{k}}{V_{ik}} \qquad V_{ik} = \omega_d r_{ik}^d$$



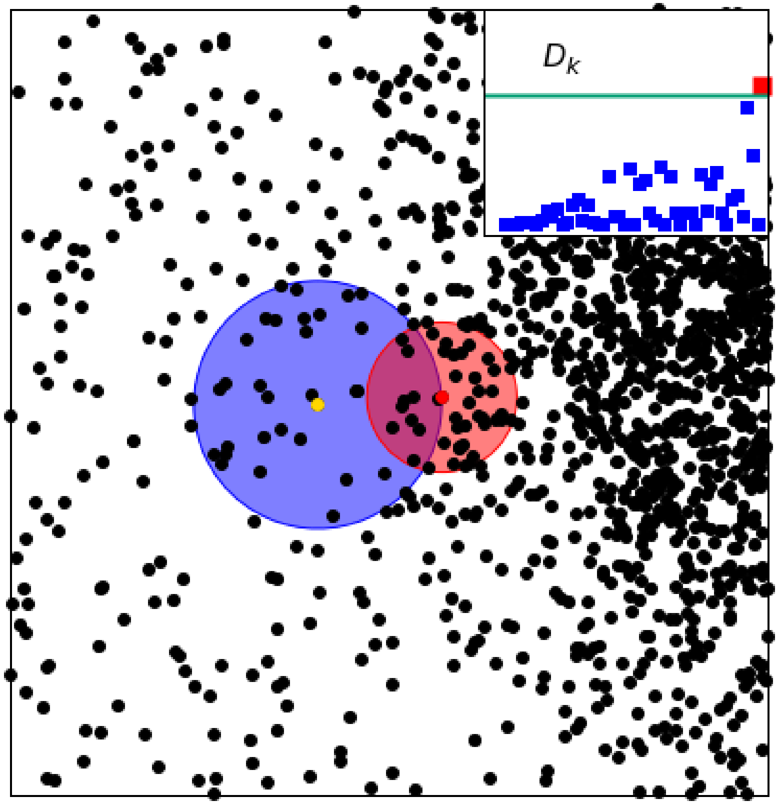
- Two problems:
- *what is right k*?
- *what is right d*?

Improving density peak clustering

1) Make density estimation parameter free

(A. Rodriguez et al., 2018, J. Chem. Th. and Comp. 14 (3), 1206)

- *what is right k?*



optimally adjust k for each point

(k should be as large as possible, but only include points with approximately equal density)

Improving density peak clustering

- *what is right d***?**

- TWO-NN idea: **finding suitable function of the distances that depends only on intrinsic dimension *d* and not on density ρ**

- Then if d_{i1}, d_{i2} are distances from 1st and 2nd neighbor of point i,

  their ratio $\mu_i = \frac{d_{i2}}{d_{i1}}$ follows a Pareto distribution: $f(\mu_i) = d\mu_i^{-(d+1)}$

- depends only on *d*!

- **Collect the μ for each point. Fit their emprical distribution and estimate *d***

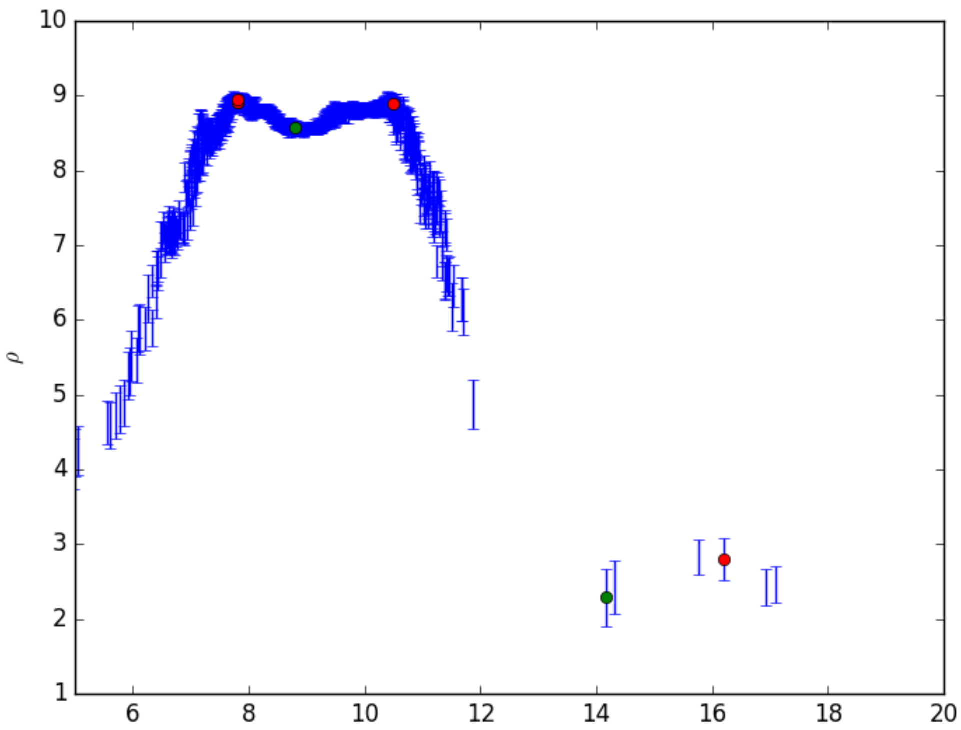
E Facco, M D'Errico, A Rodriguez, A Laio, Scientific Reports 7, 12140 (2017)
M. Allegra, E. Facco, A. Laio and A. Mira, arxiv:1902.10459 (2019)

Improving density peak clustering

2) Automatically identify candidate maxima

(M. D'Errico et al. , arXiv:1802.10549)

- All points with ρ higher than their neighbors are candidate maxima



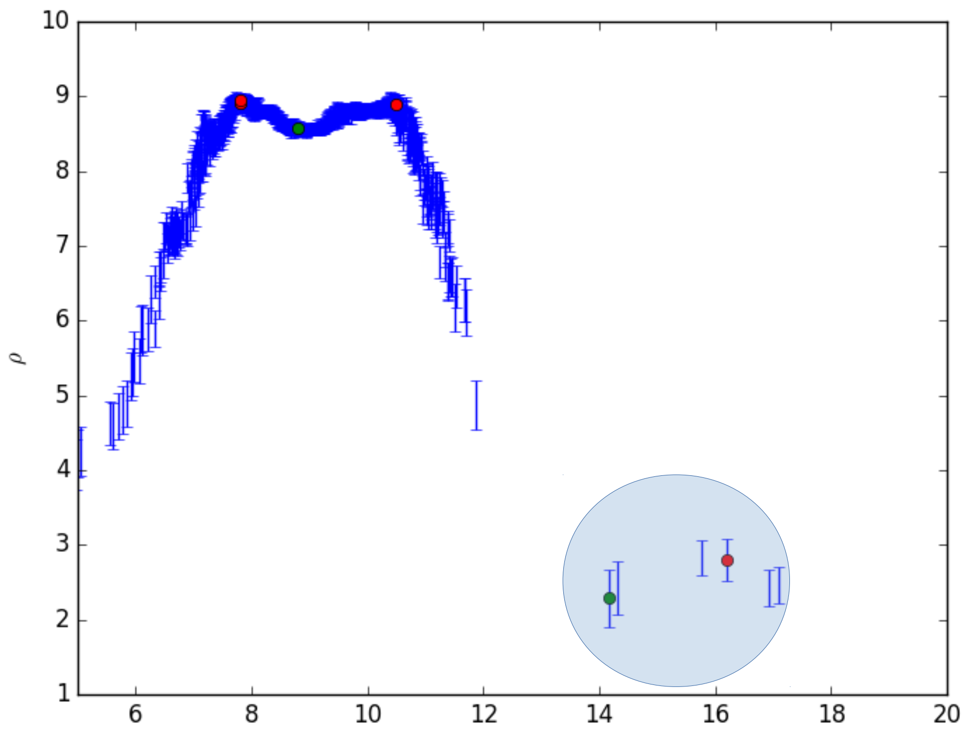
Improving density peak clustering

3) Keep only significant peaks (within given confidence level Z)

compare ρ of each maximum with that highest saddle points

$$z = \frac{\rho_{peak} - \rho_{saddle}}{\sqrt{(\Delta\rho_{peak})^2 + (\Delta\rho_{saddle})^2}}$$

Keep only point with “significant” Z (e.g., Z > 2)



not significant

Improving density peak clustering

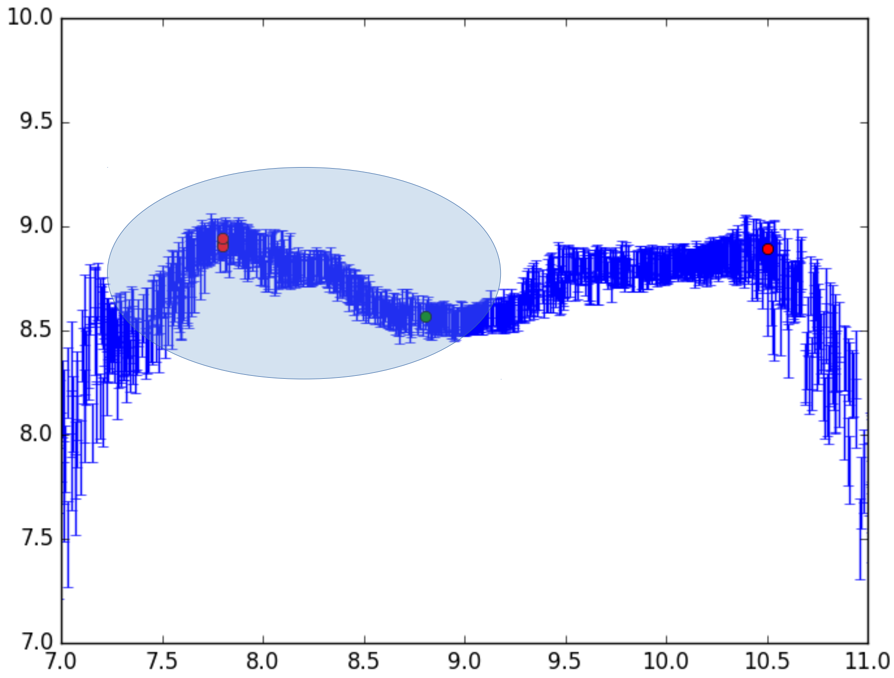
3) Keep only significant peaks (within given confidence level Z)

compare ρ of each maximum with that highest saddle points

$$z = \frac{\rho_{peak} - \rho_{saddle}}{\sqrt{(\Delta\rho_{peak})^2 + (\Delta\rho_{saddle})^2}}$$

Keep only point with “significant” Z (e.g., Z > 2)

Z controls the resolution of the method

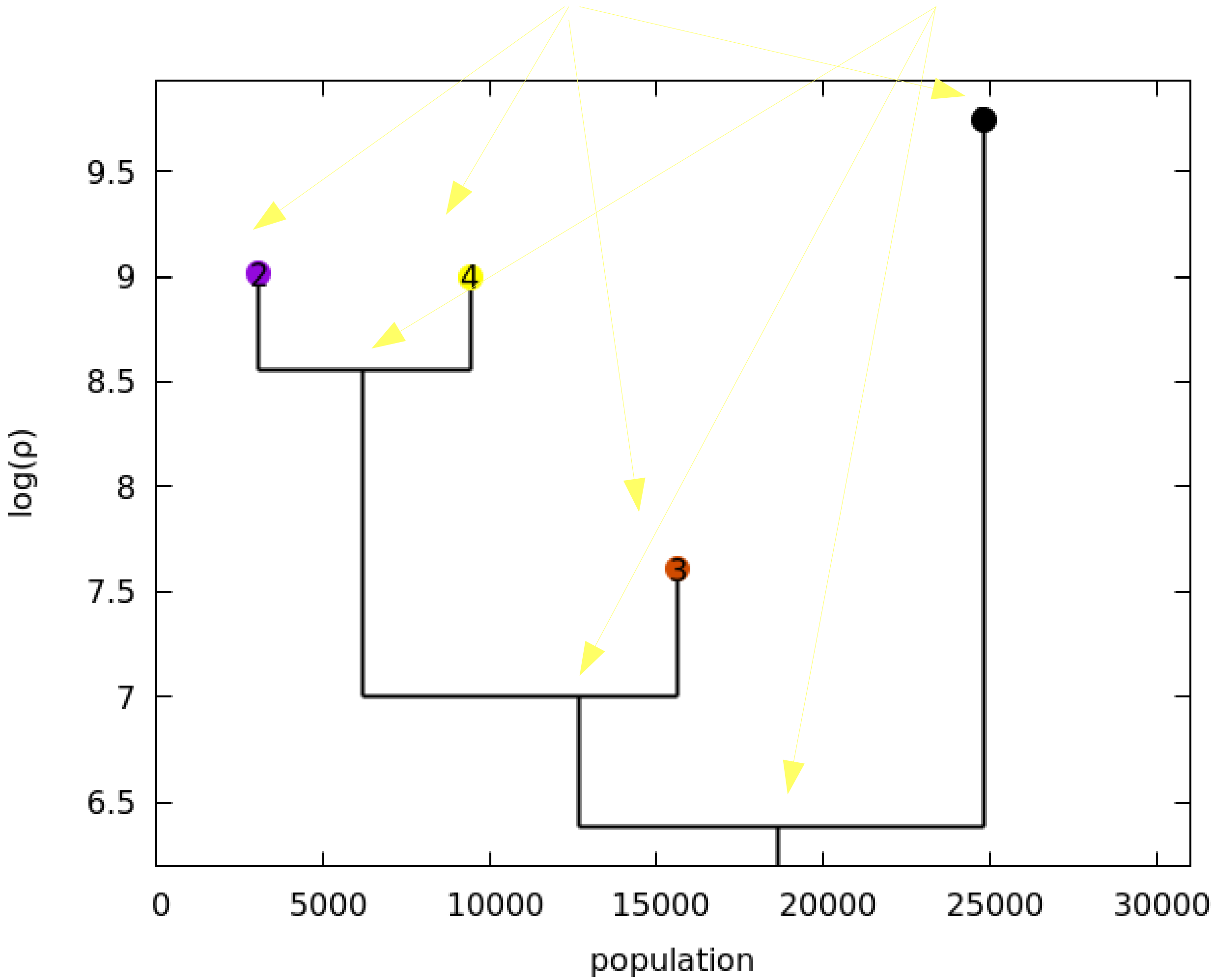


significant

Improving density peak clustering

4) Compact *multiscale* representation of the topography

Use information about *maxima* and *saddle points*



Z=2

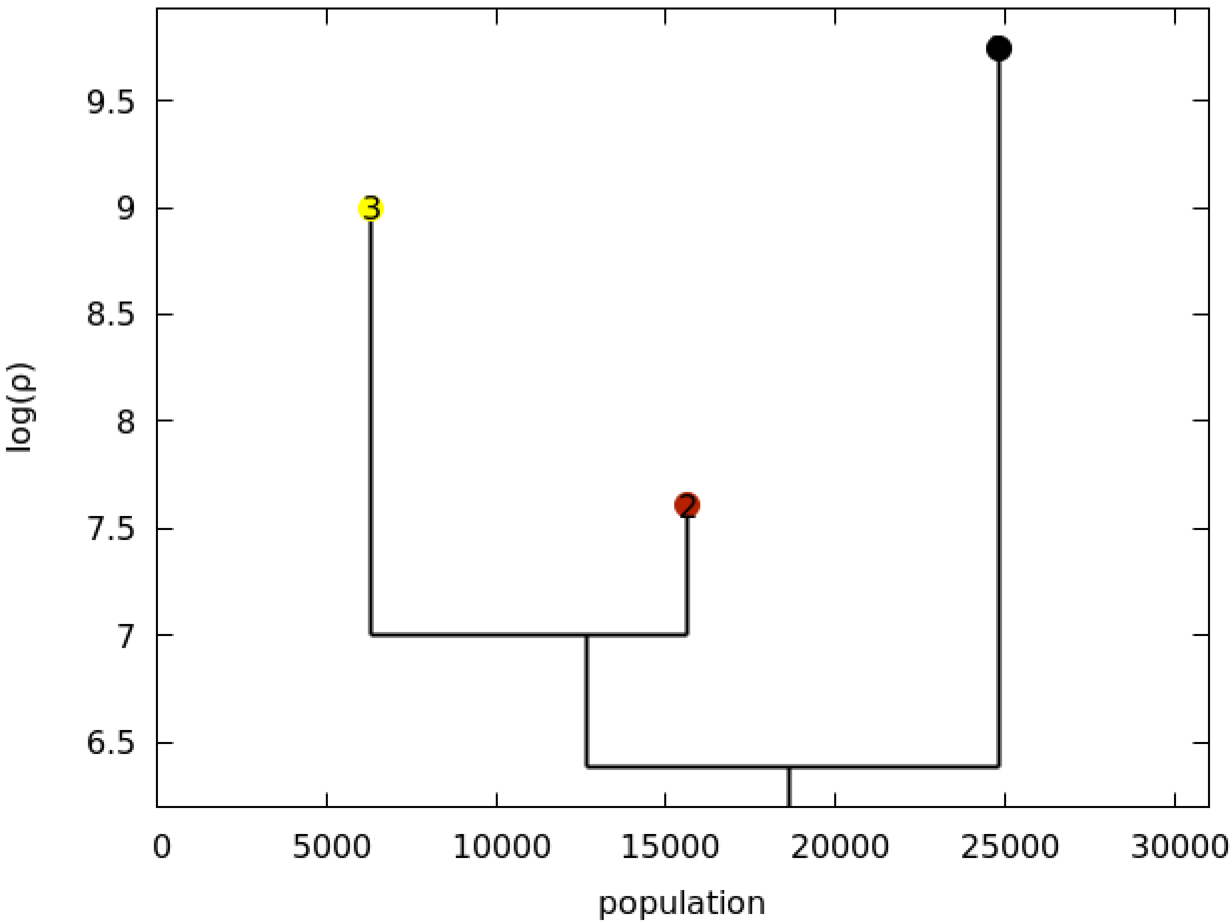
link together pair of peaks with highest saddle point between them

density dendrogram

Improving density peak clustering

4) Compact representation of the topography

For different Z, different histograms (different resolution)



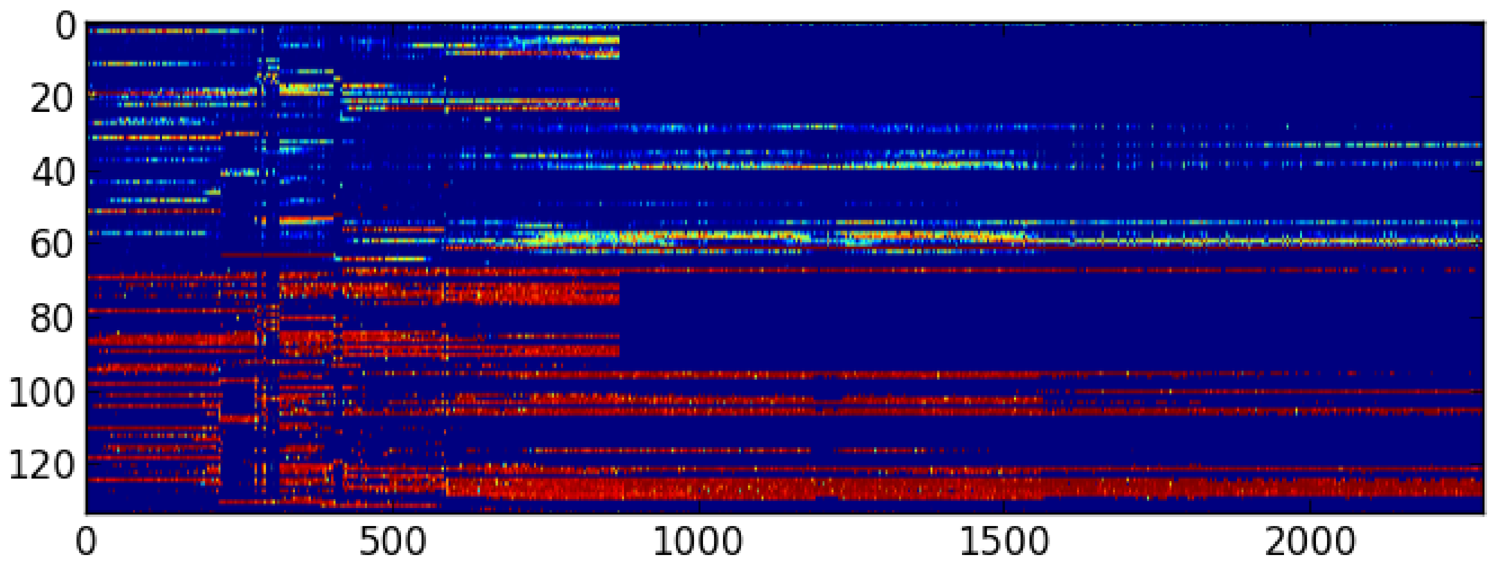
Z=3

A (blind) application: state detection

Firing network (65 channels) *(courtesy of Nicola Pedreschi)*

For each channel, two “topological” features representing channel’s role in the network were evaluated in sliding windows:

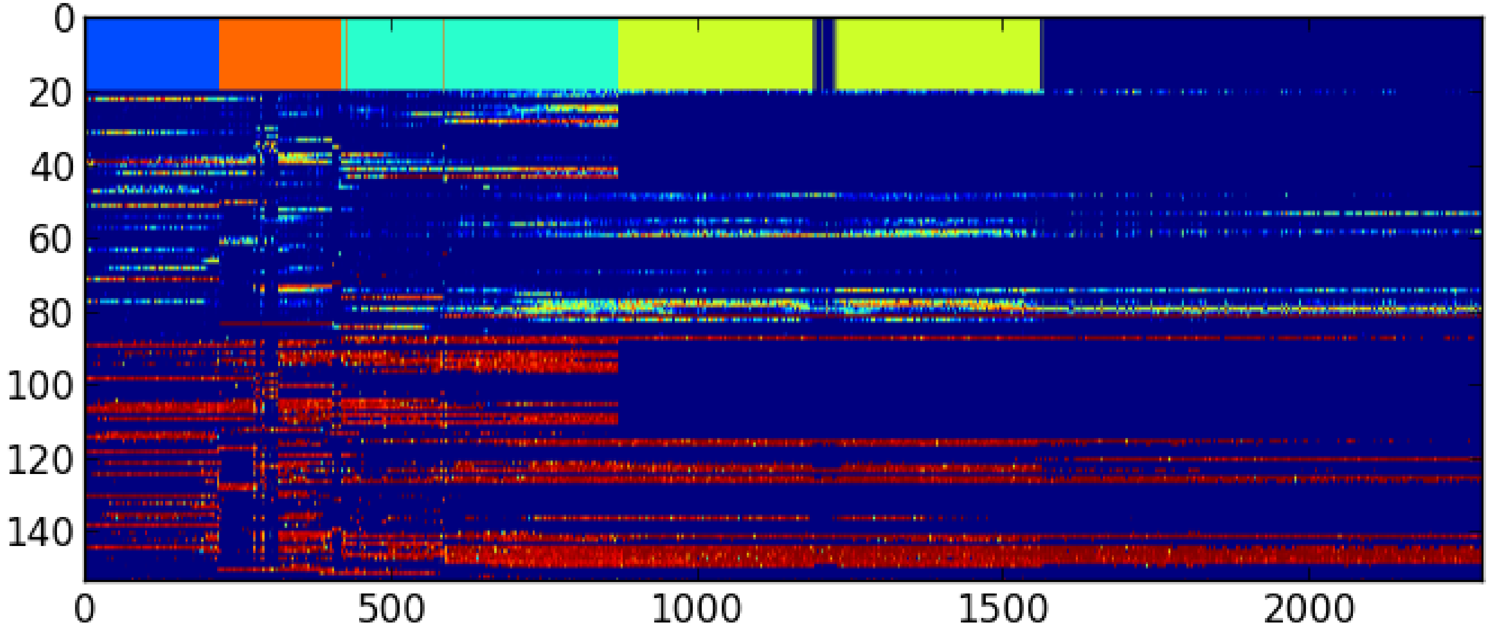
- “liquididy” (variability of a node)
- “coreness” (whether node belongs to core/periphery of the network)



Clearly, a few dyamic regimes (“states”)

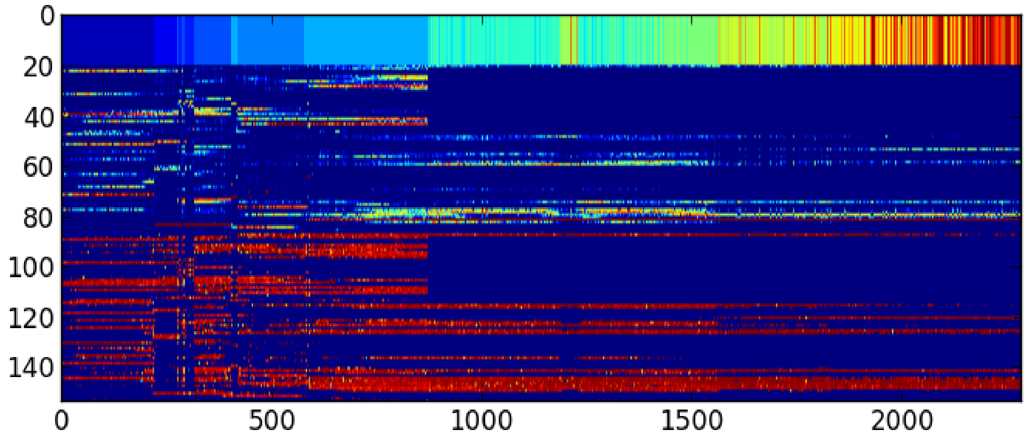
A blind application

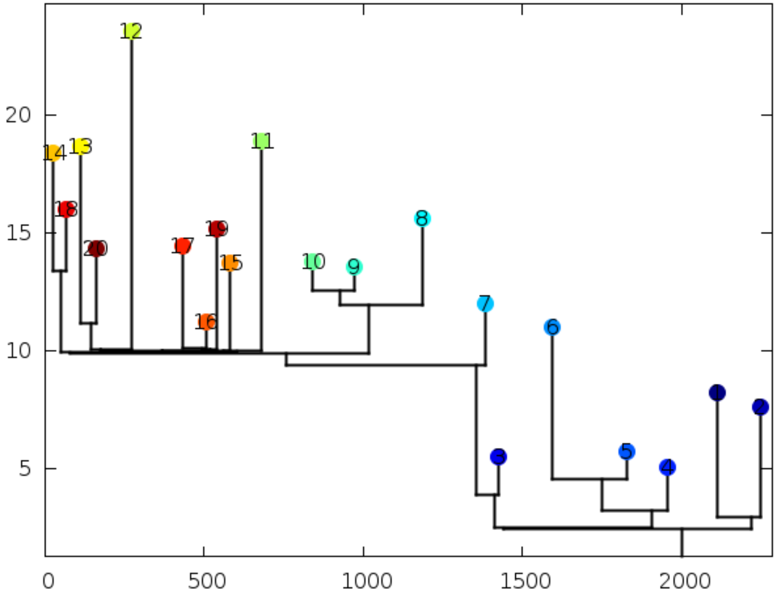
K-means with K=5



A blind application

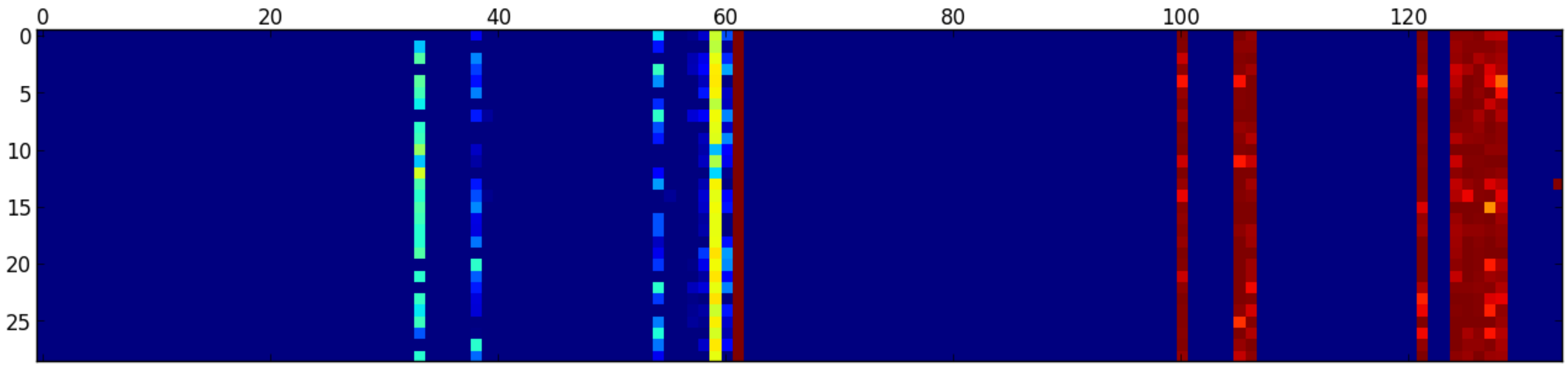
DPC with Z=2



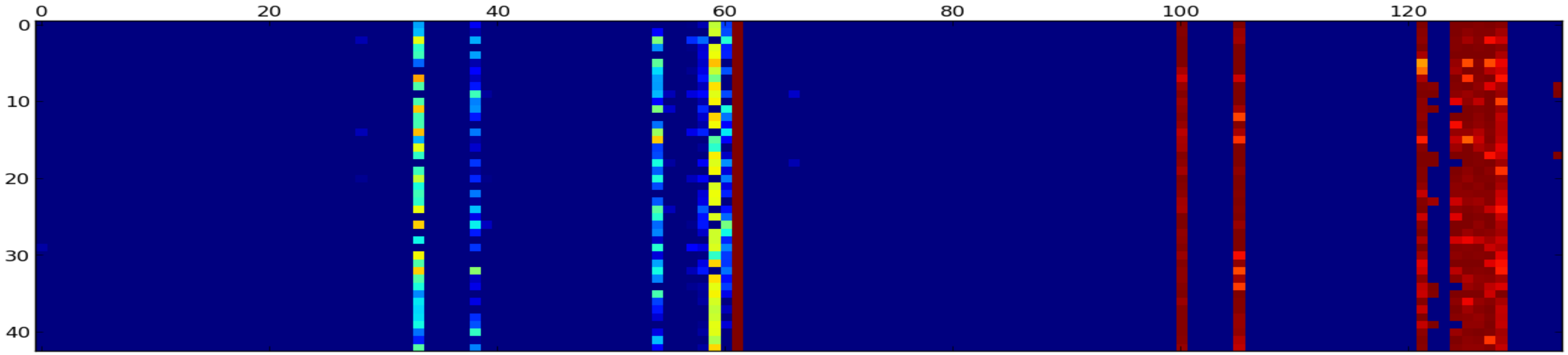


A blind application

Subtle differences...



Cluster 18

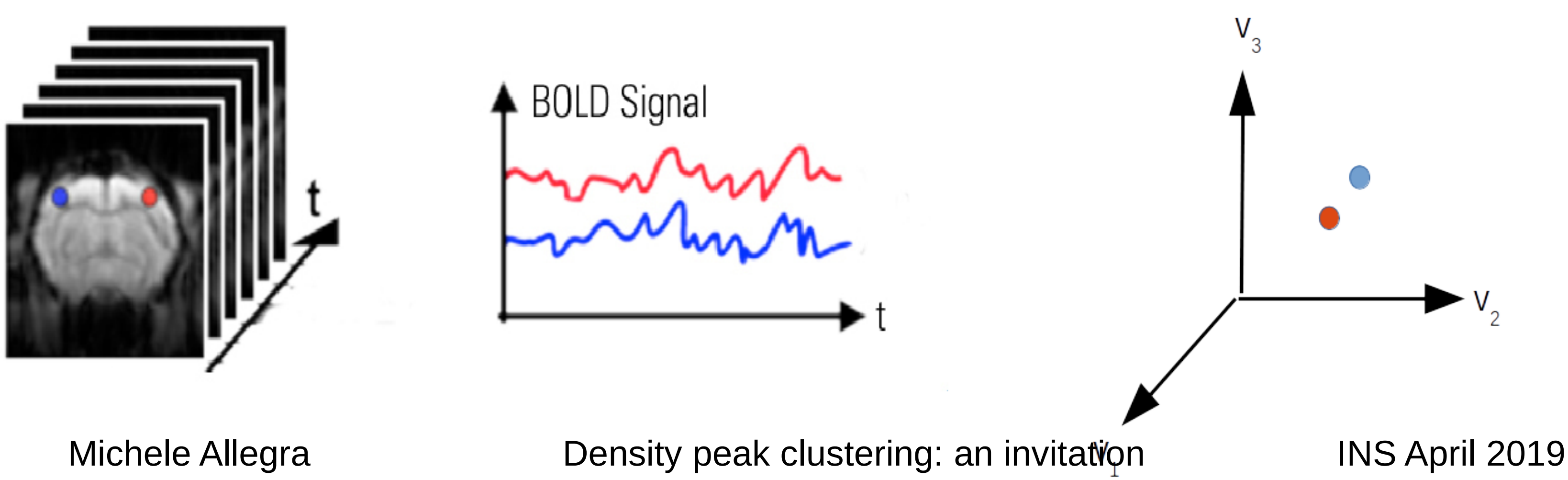


Cluster 19

Applying DPC to fMRI

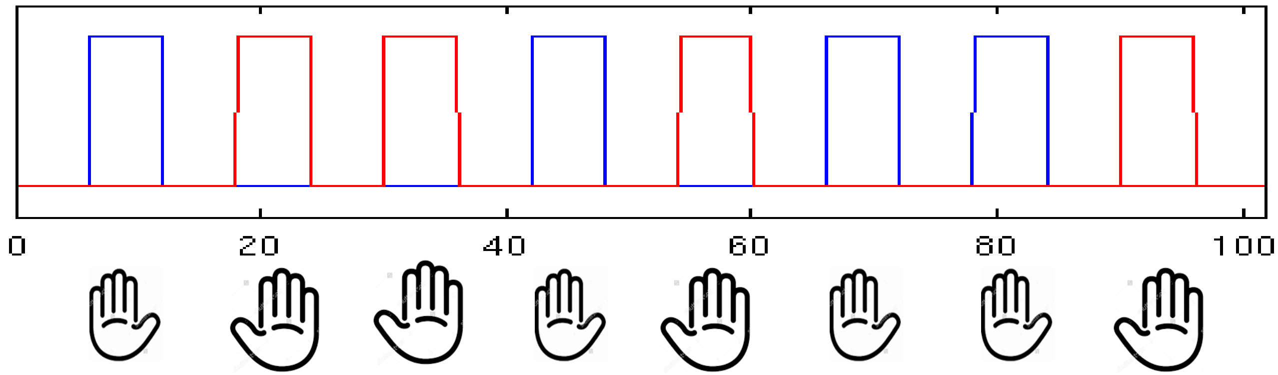
Allegra et al., Hum Brain Mapp 2017

- Apply DPC in the space of BOLD time series
- consider window of *T* frames
- to each voxel corresponds a BOLD time series of *T* values, $v_1, v_2, \ldots, v_T$
- consider *T*-dimensional space of time-series
- each voxel time series is a point in this space
- cluster in this space is group of voxels with coherent BOLD
- We call such clustering **Coherence Density Peak Clustering (CDPC)**



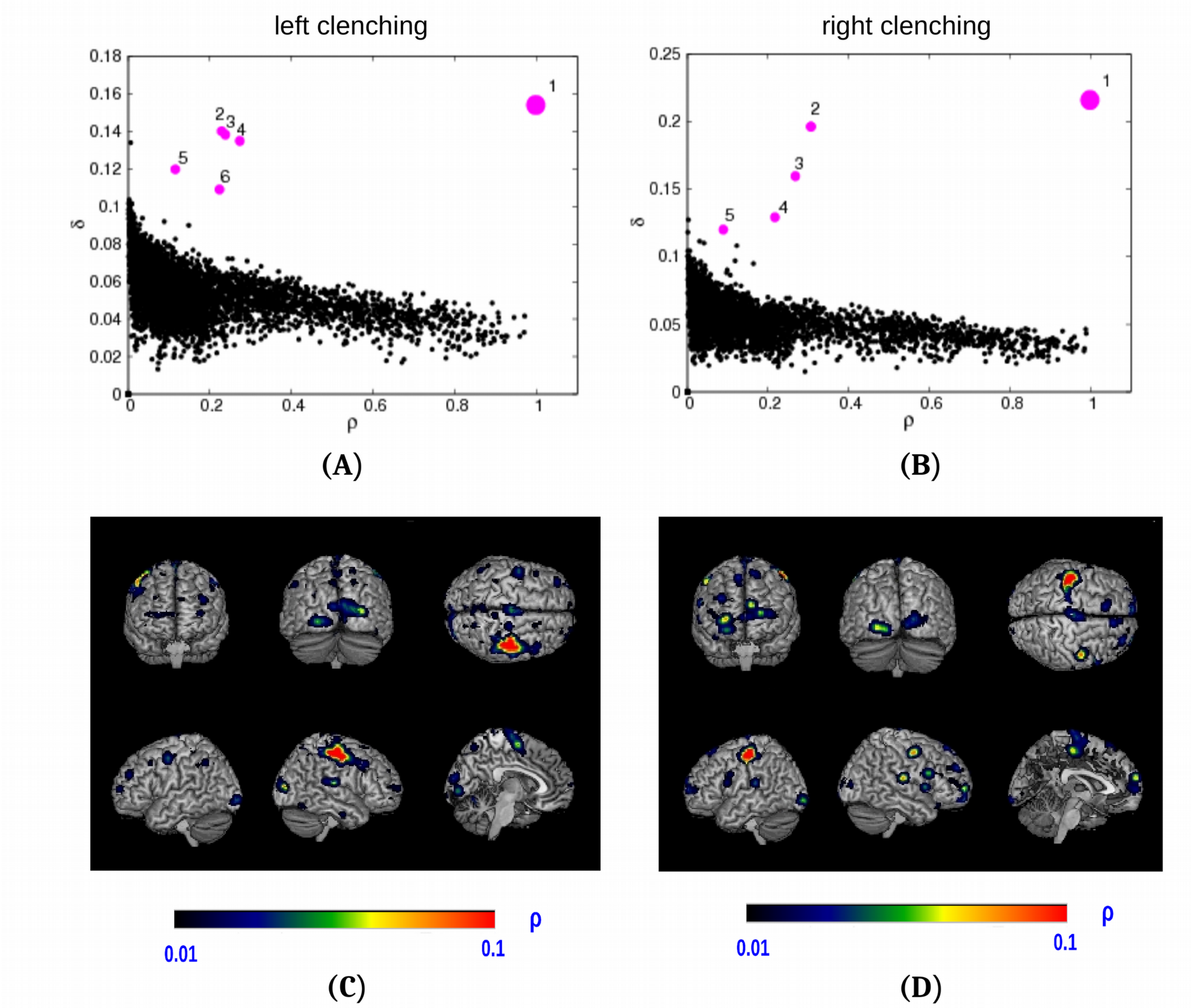
Simple validation: motor experiment

- First test in motor experiment (alternative trials left/right clenching, visually cued)



- can we reconstruct activity patterns in single trials?
- Apply to short time windows (~12 volumes, ~20 s) corresponding to single clenching trials

Simple validation: motor experiment



In window corresponding to left/right clenching trial we find main cluster including right/left motor cortex

The cluster also includes part of the visual cortex (clenching was visually cued)

Conclusions

- Traditional clustering methods have a few limitations
- K-means: difficult to identify number of clusters, remove noise, deal with nonspherical clusters
- DB-scan: difficult to adjust the free parameters, cannot resolve structures at different scales
- Density peak clustering is a parameter-free clustering method that allows to reconstruct the complex topography of a data space
- The method rests on an intricate density estimation, also accounting for the intrinsic dimension of the data
- Density peak clustering allows to observe the system at different resolution levels (often non-trivial)

Acknowledgments

Alessandro Laio

Maria D’Errico

Elena Facco

Alex Rodriguez

int institut de neurosciences de la timone

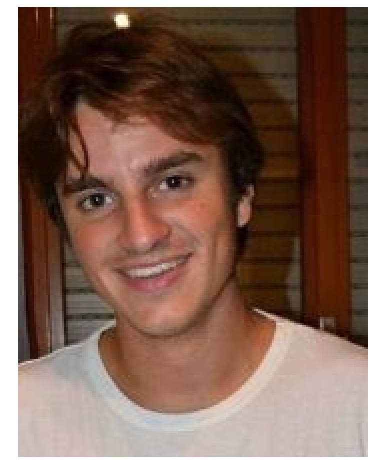
Thank you for the invitation!!

Institut de Neurosciences des Systèmes

INS

Spase Petkoski

Giovanni Rabuffo



Nicola Pedreschi

Thank you for your attention!!