Clustering by fast search-and-find of density peaks

Alessandro Laio, Maria d'Errico and Alex Rodriguez

SISSA (Trieste)

What is a cluster?

clus·ter [kluhs-ter] , *noun*

1.a number of things of the same kind, growing or held together; a bunch: *a cluster of grapes.*

2.a group of things or persons close together: *There was a cluster of tourists at the gate.*

3.*U.S. Army.* a small metal design placed on a ribbon representing an awarded medal to ind icate that the same medal has been awarded again: *oak-leaf cluster.*

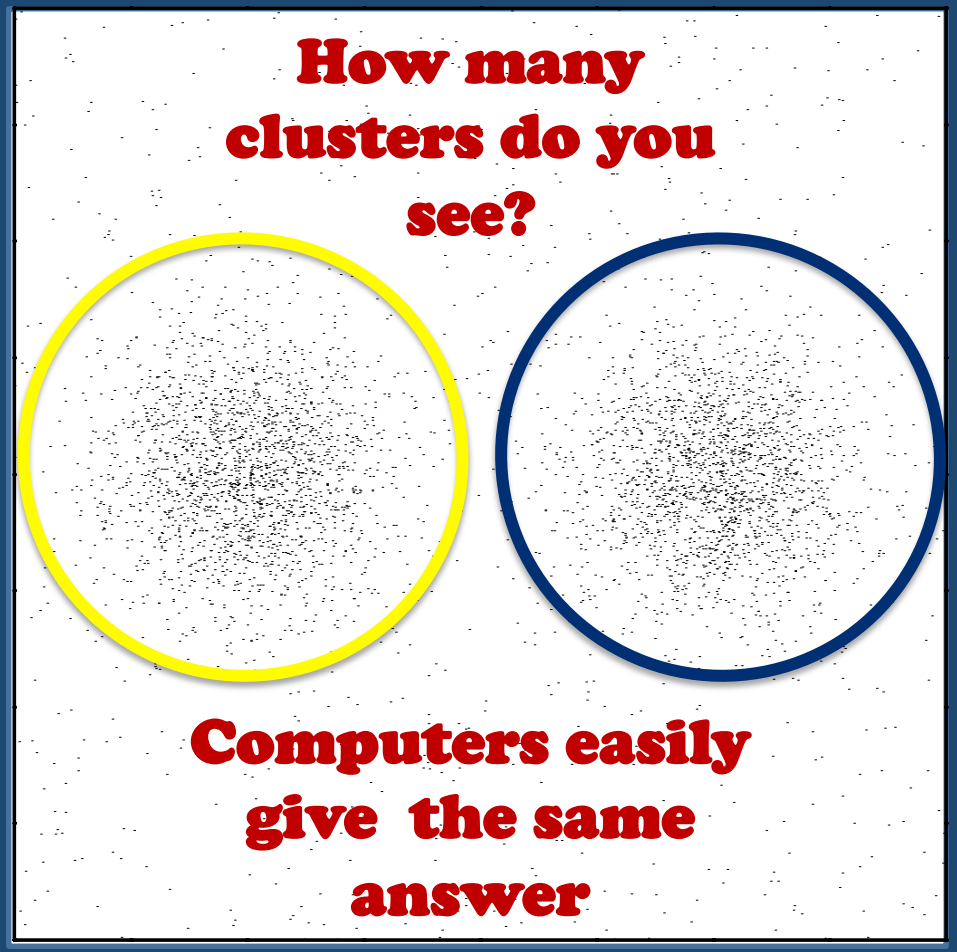
4.*Phonetics* . a succession of two or more contiguous consonants in an utterance: cluster of *strap.*

5.*Astronomy* . a group of neighboring stars, held together by mutual gravitation, that have essentially the same age and composition and thus supposedly a common origin.

What is a cluster?

Visually it is a region with a **high density** of points.

separated from other dense regions

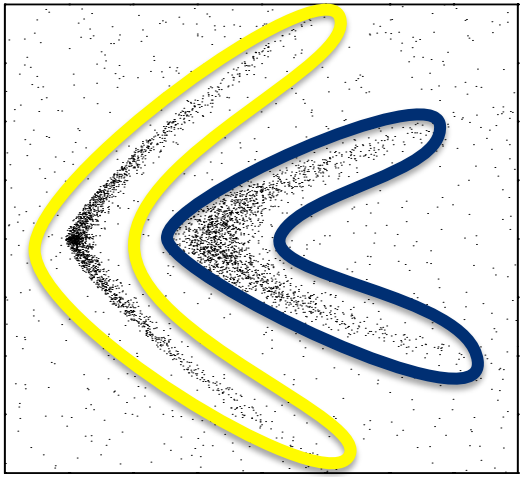


What is a cluster?

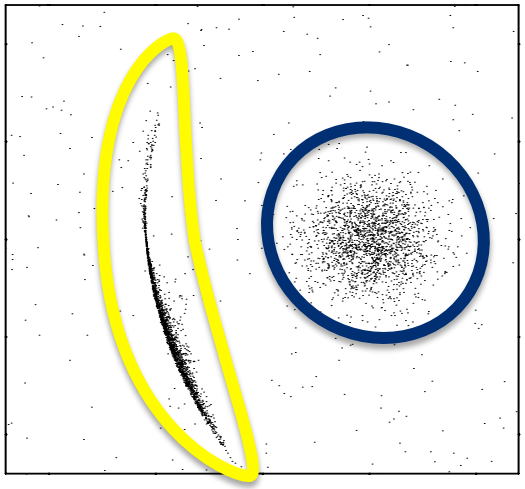
And now? How many clusters?

Computers have big problems here!

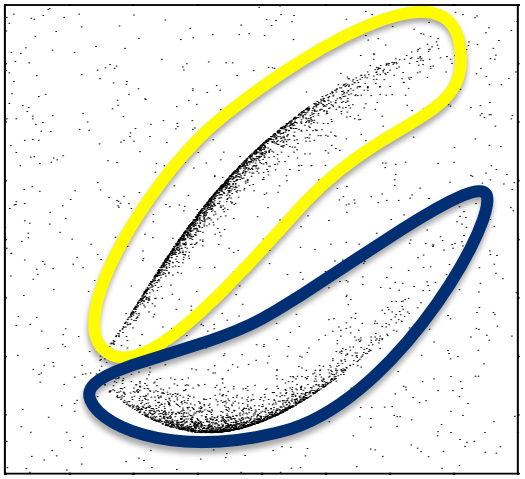
B



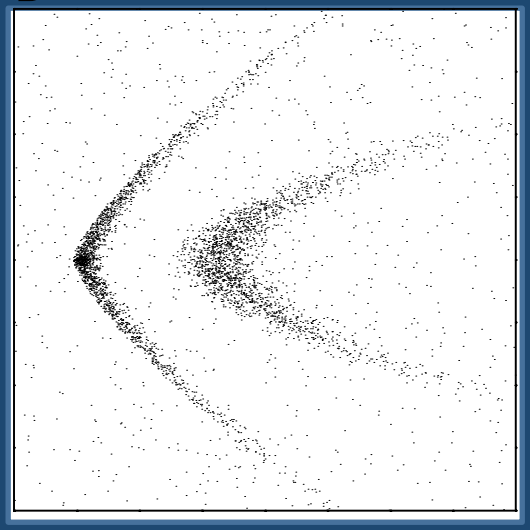
C



D



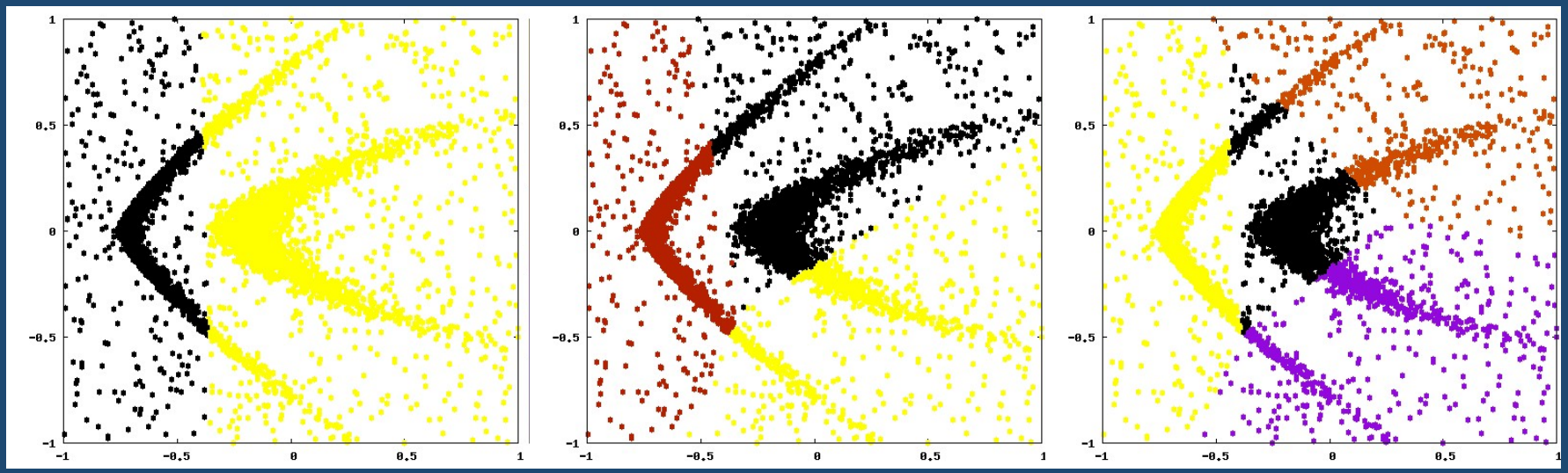
Inefficiency of standard algorithms



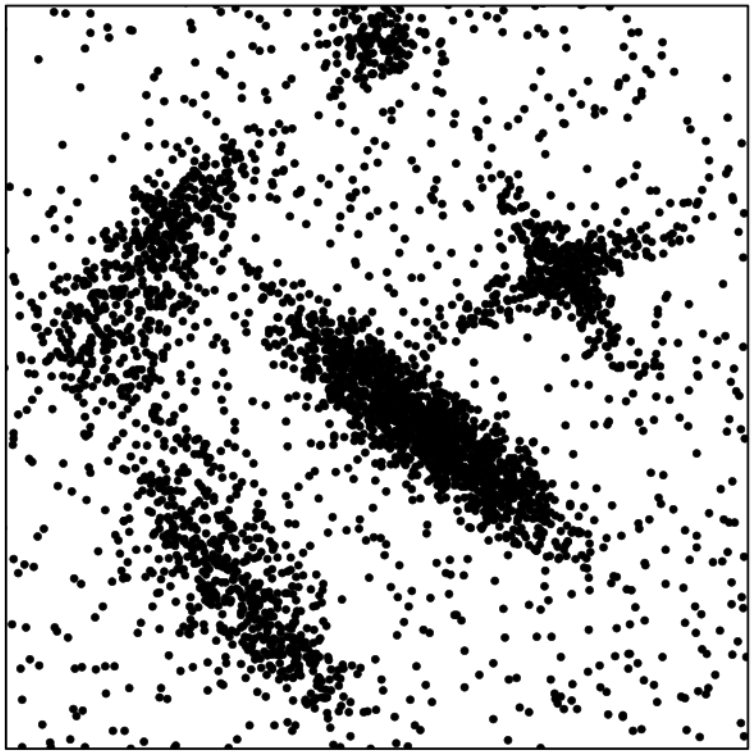
K-means algorithm (11748 citations!!!!)

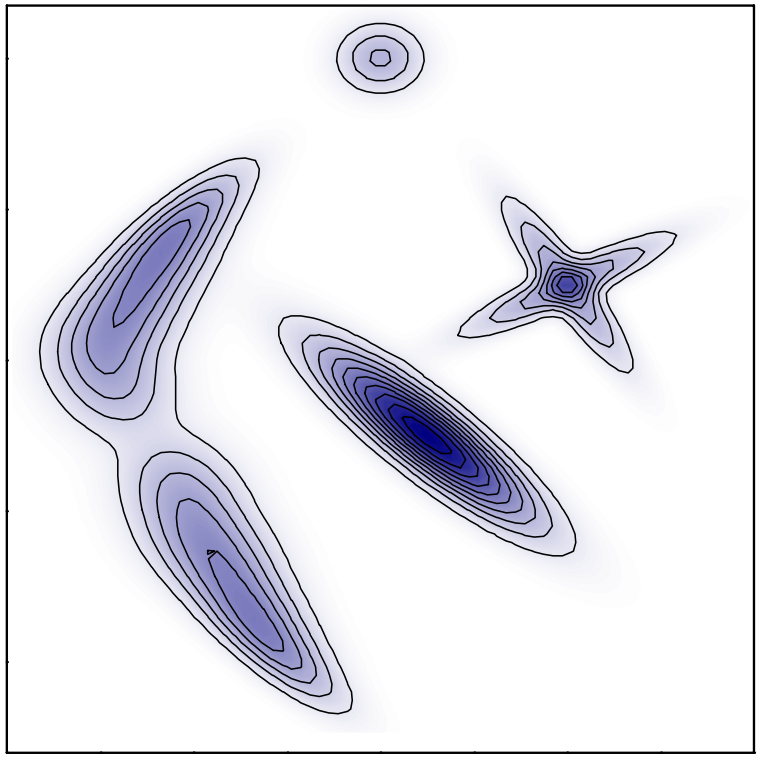
It assigns each point to the closest cluster center. Variables: the number of centers and their location

By construction it is unable to recognize non-spherical clusters



What is a cluster?



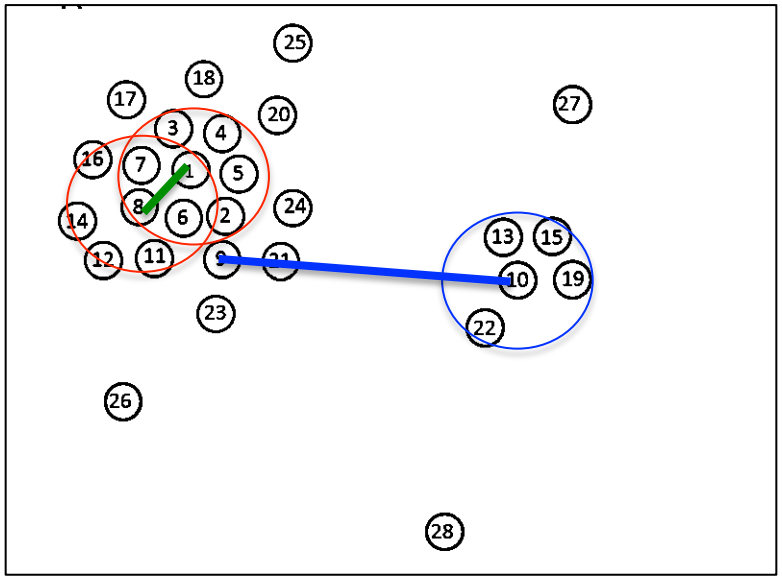


Clusters = peaks in the density of points
= peaks in the “mother” probability distribution

Our approach: fast search-and-find of density peaks

SCIENCE, 1492, vol 322 (2014)

EXAMPLE: Clustering in a 2-dimensional space



1) Compute the local density around each point

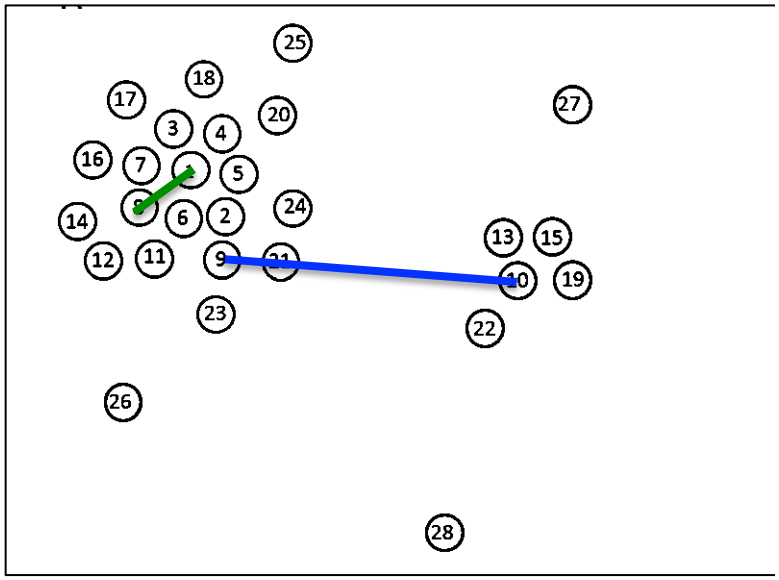
$\rho(1)=7$
$\rho(8)=5$
$\rho(10)=4$

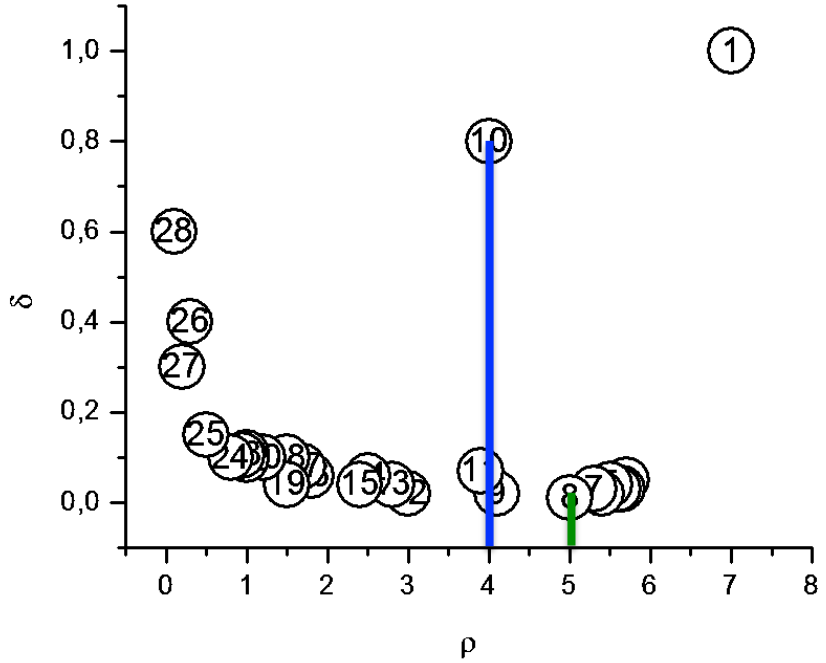
2) For each point compute the distance with all the points with higher density. Take the minimum value.

Our approach: fast search-and-find of density peaks

SCIENCE, 1492, vol 322 (2014)

3) For each point, plot the minimum distance as a function of the density.

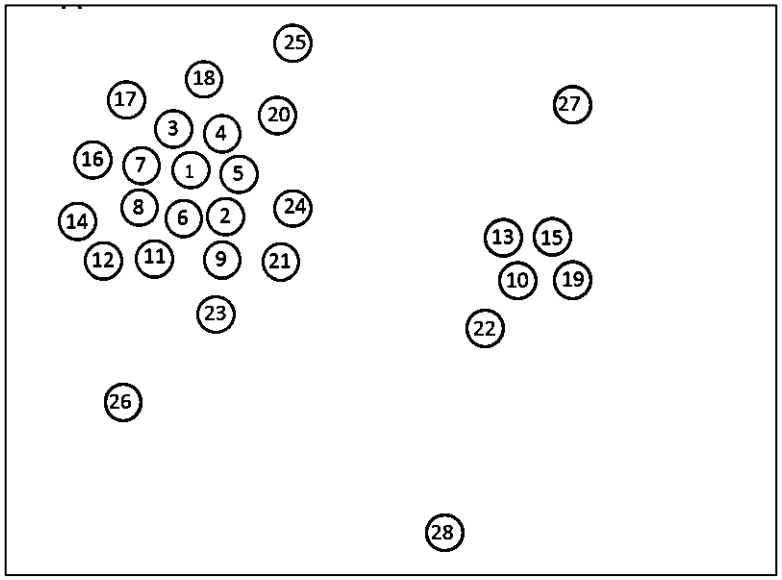


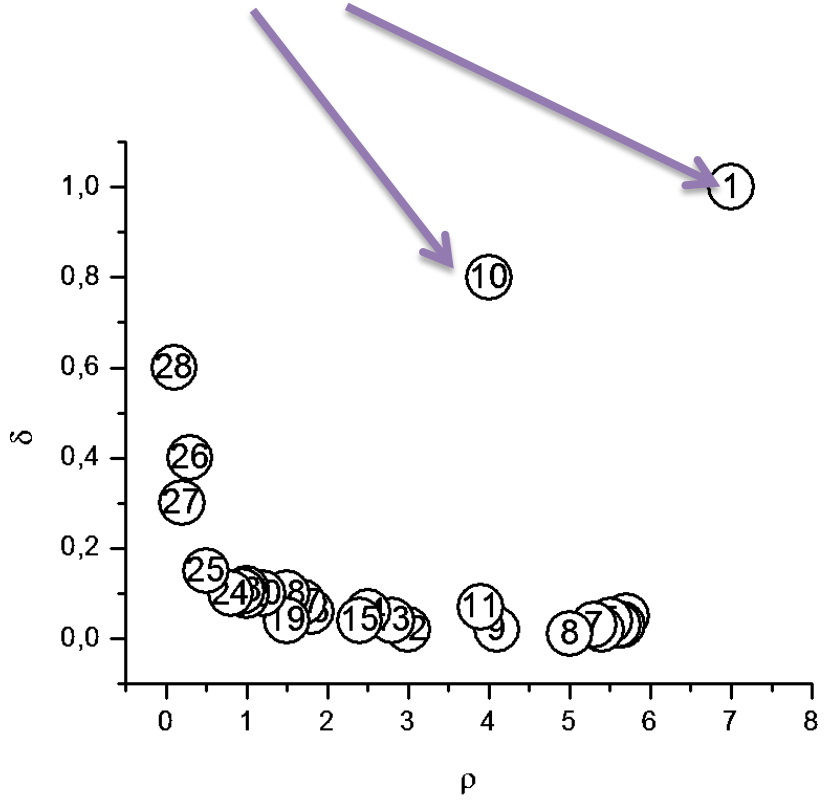


Our approach: fast search-and-find of density peaks

SCIENCE, 1492, vol 322 (2014)

4) the “outliers” in this graph are the cluster centers

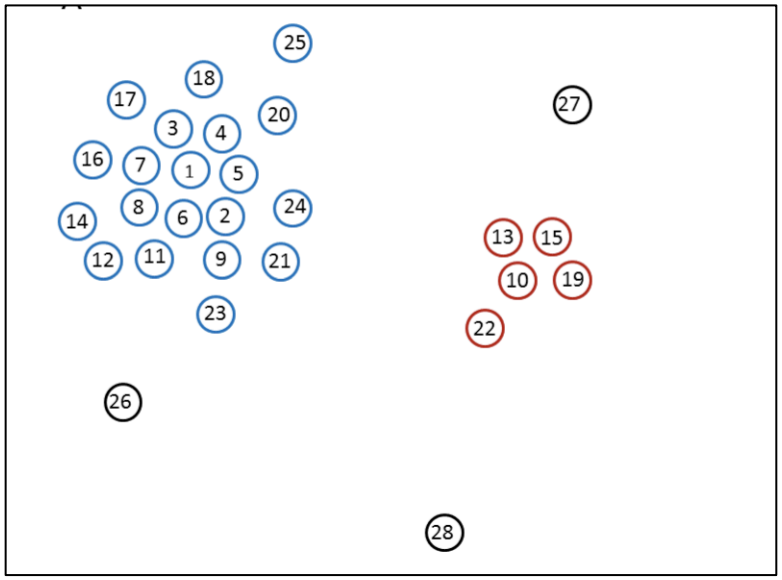


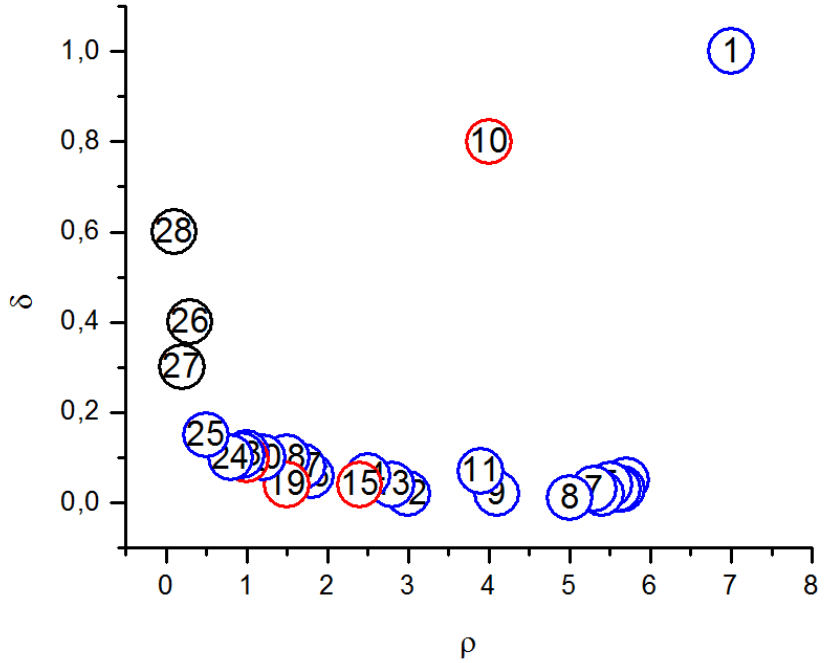


Our approach: fast search-and-find of density peaks

4)) the “outliers” in this graph are the cluster centers

5) Assign each point to the same cluster of its nearest neighbor of higher density





Not even an algorithm…

Given a distance matrix d_{ij}, for each data point *i* compute:

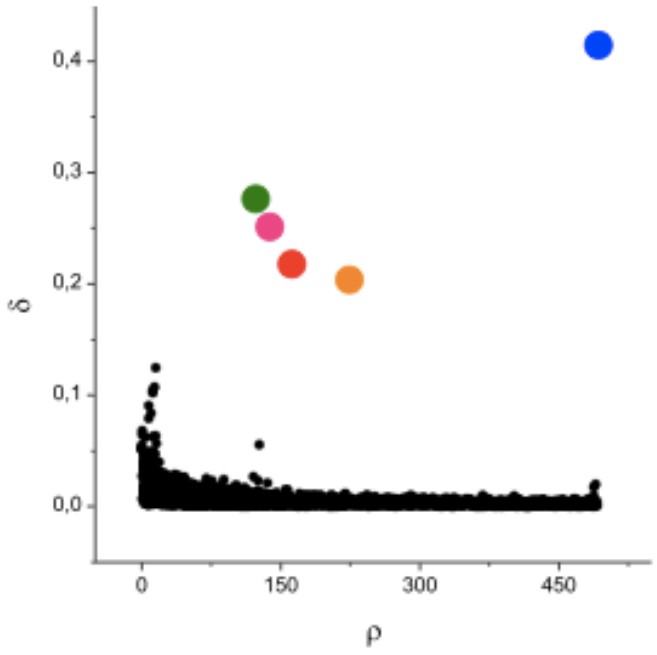
$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right)$$ (number of data points within a distance d_c)

$$\delta_i = \min_{j:\rho_j > \rho_i} \left(d_{ij}\right)$$ (distance of the closest data point of higher density)

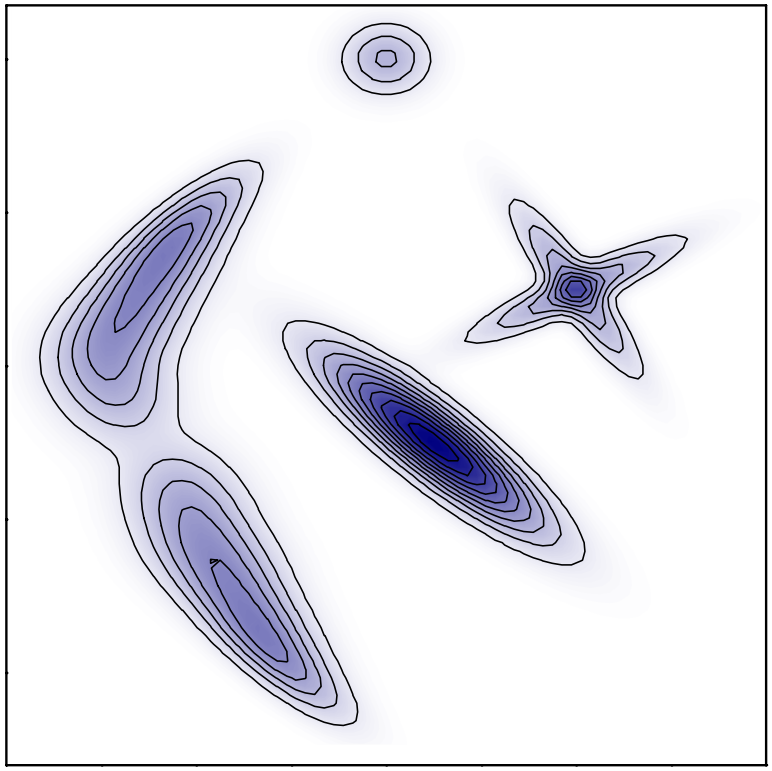
Plot δ as a function of ρ

One free parameter: the cutoff distance d_c

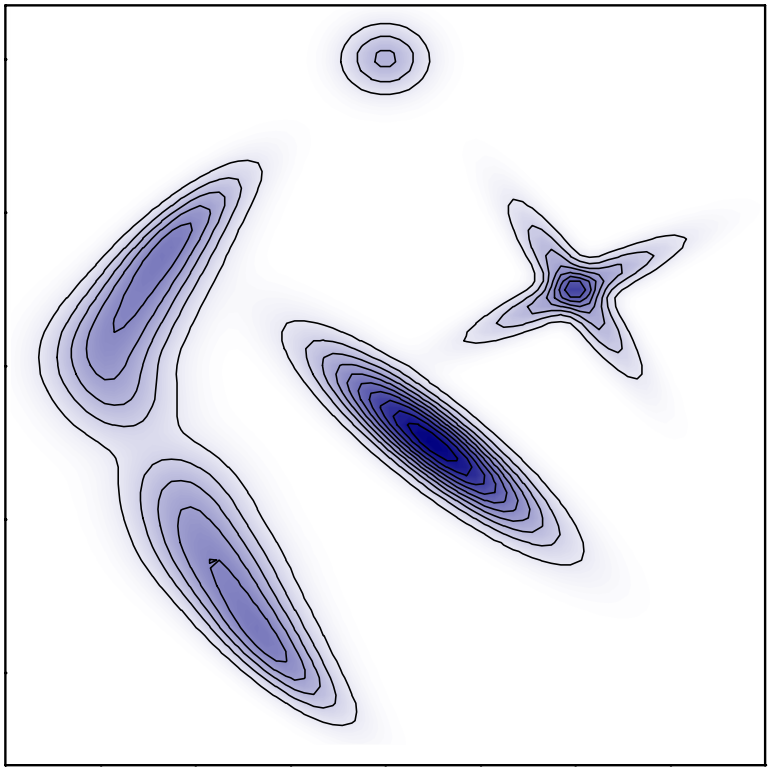
However: the method is only sensitive to the relative density of two data points, not to the absolute value of ρ

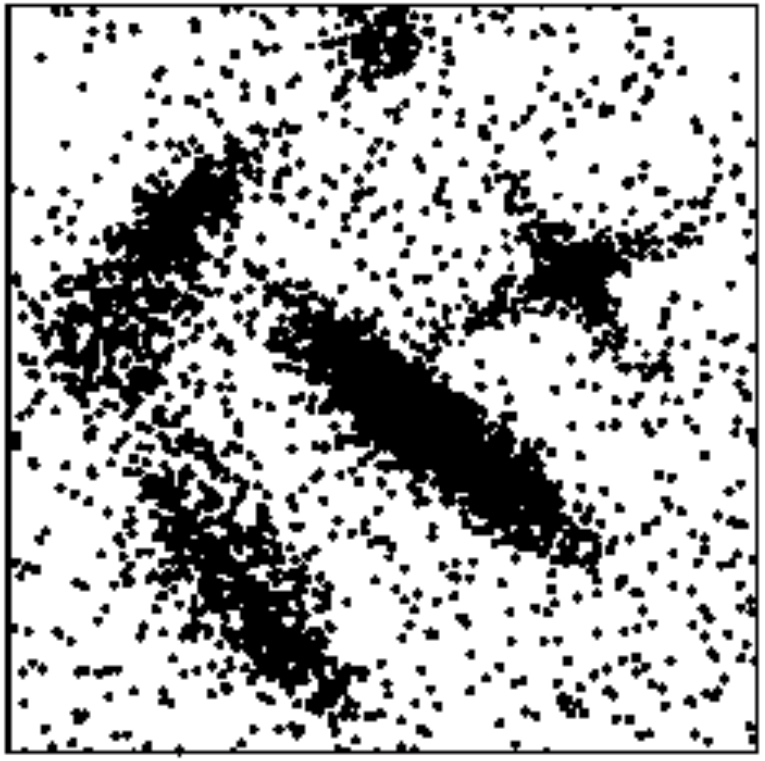


The clustering approach at work

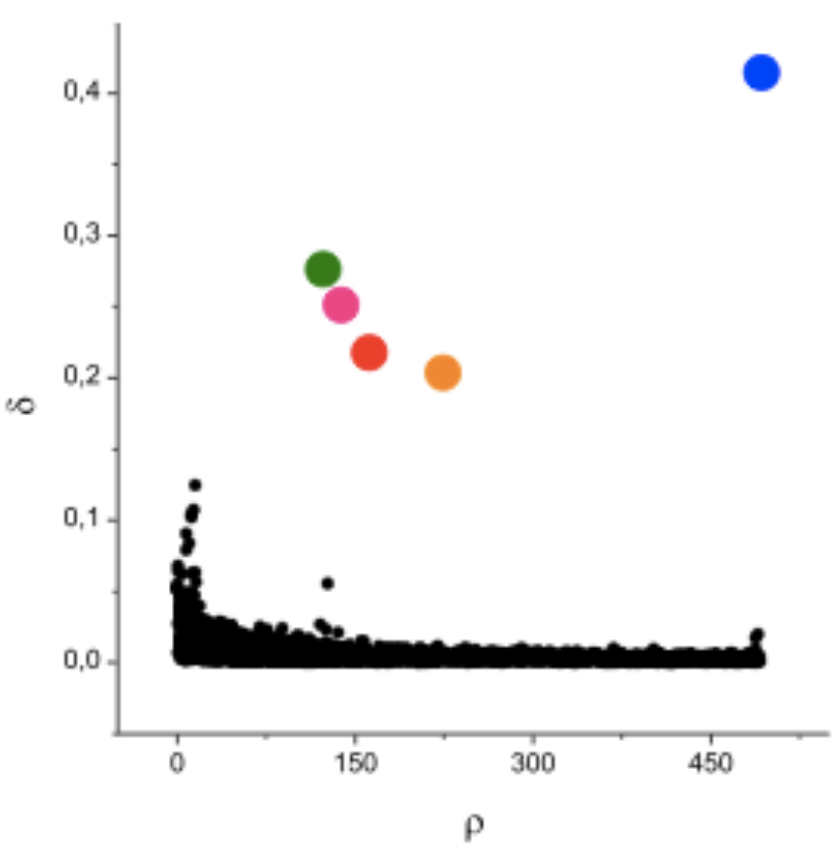


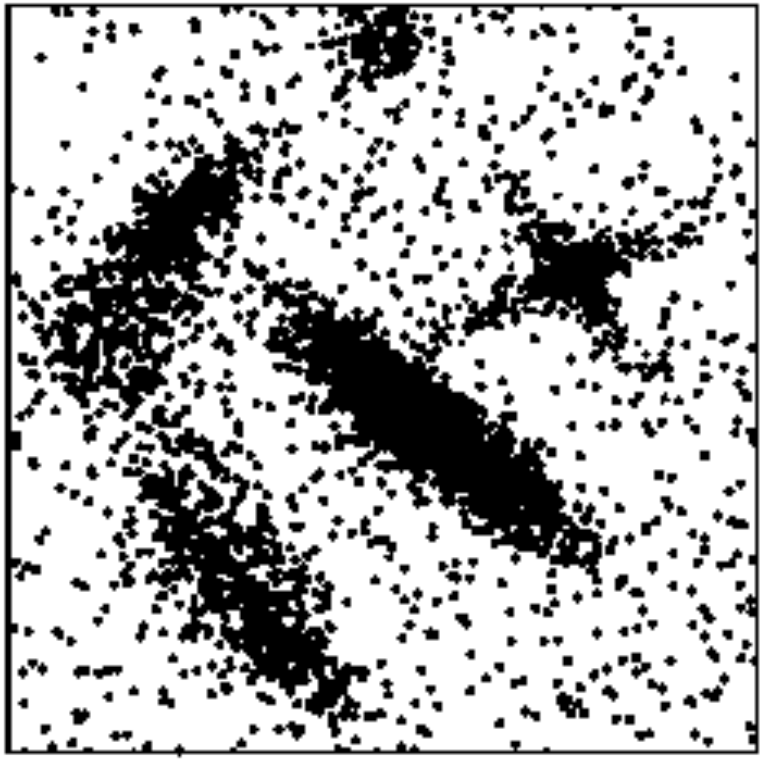
The clustering approach at work



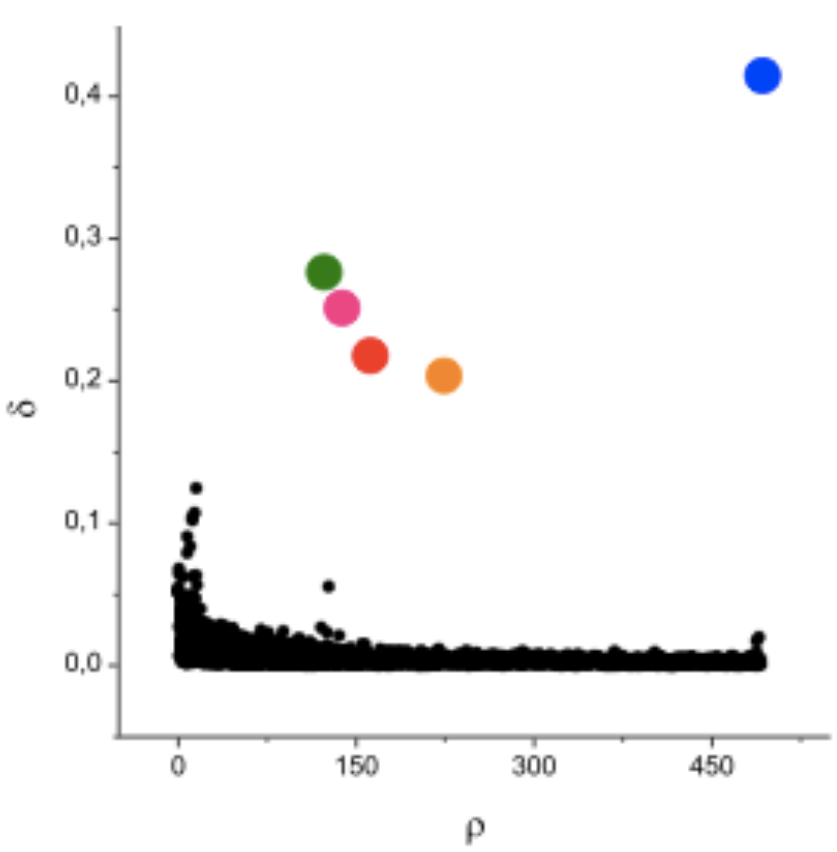


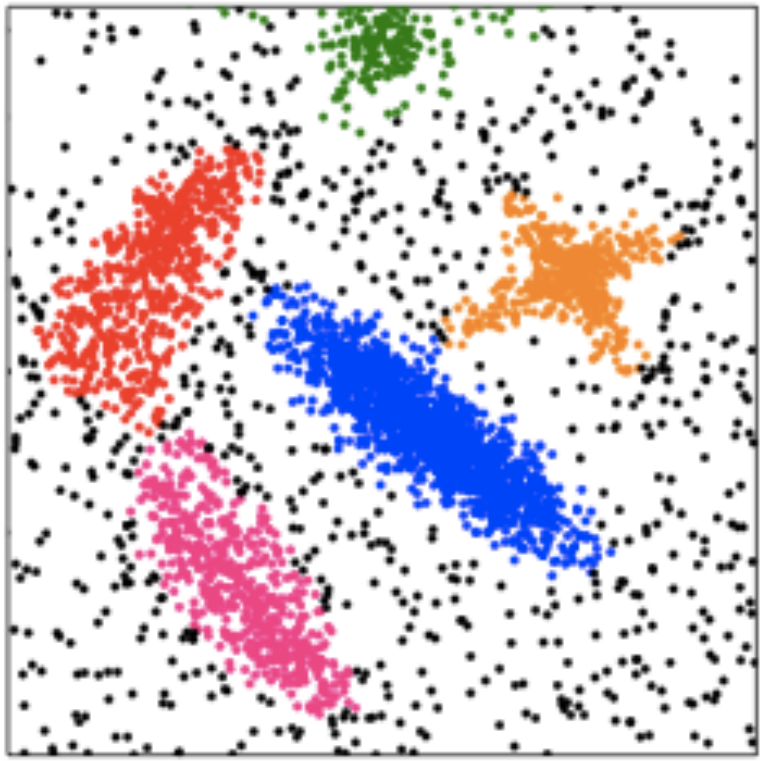
The clustering approach at work



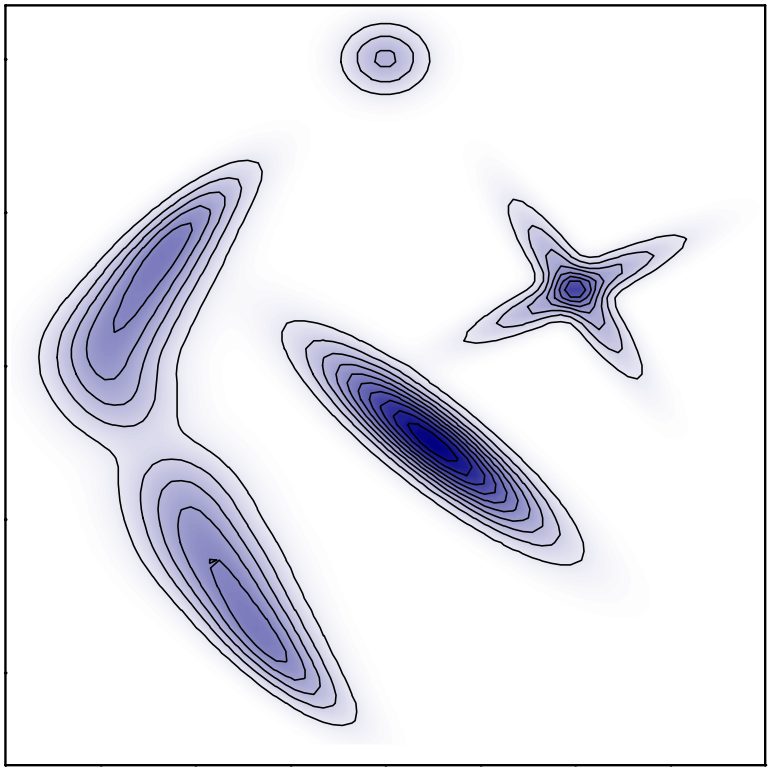


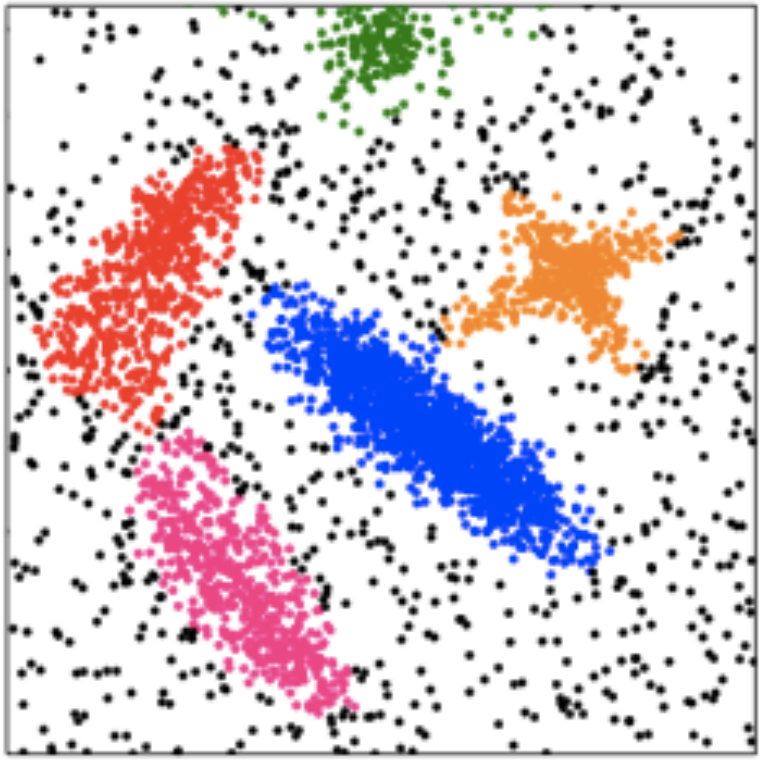
The clustering approach at work



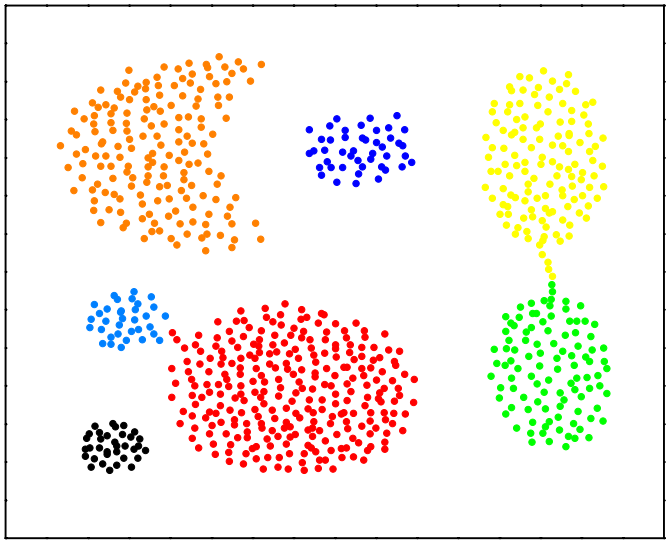


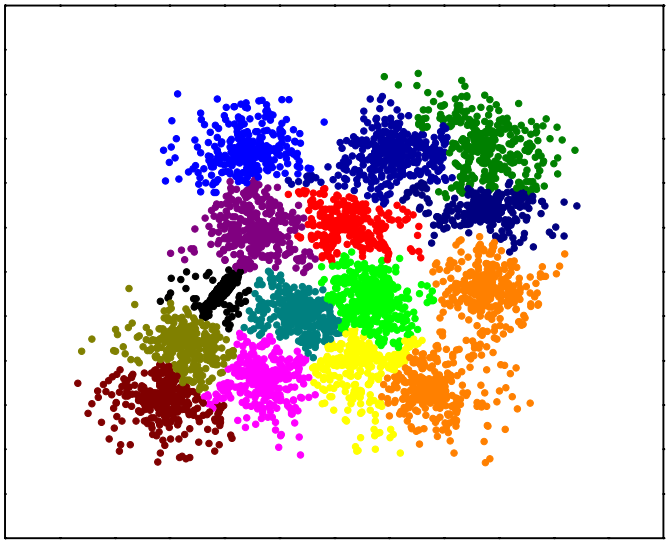
The clustering approach at work



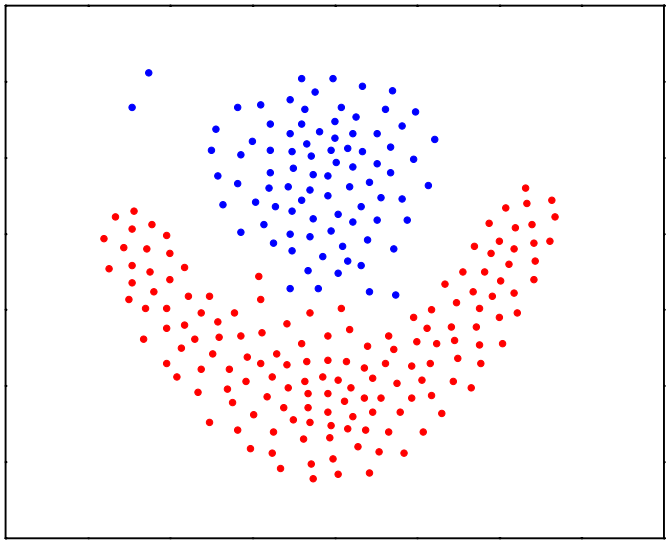


The clustering approach at work

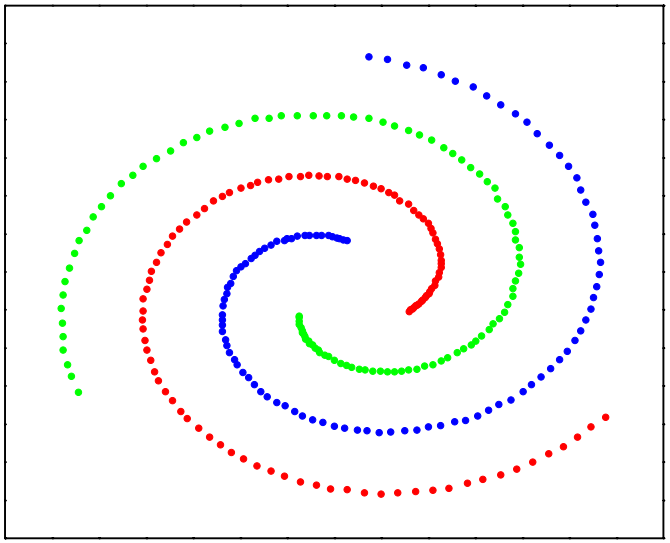




C

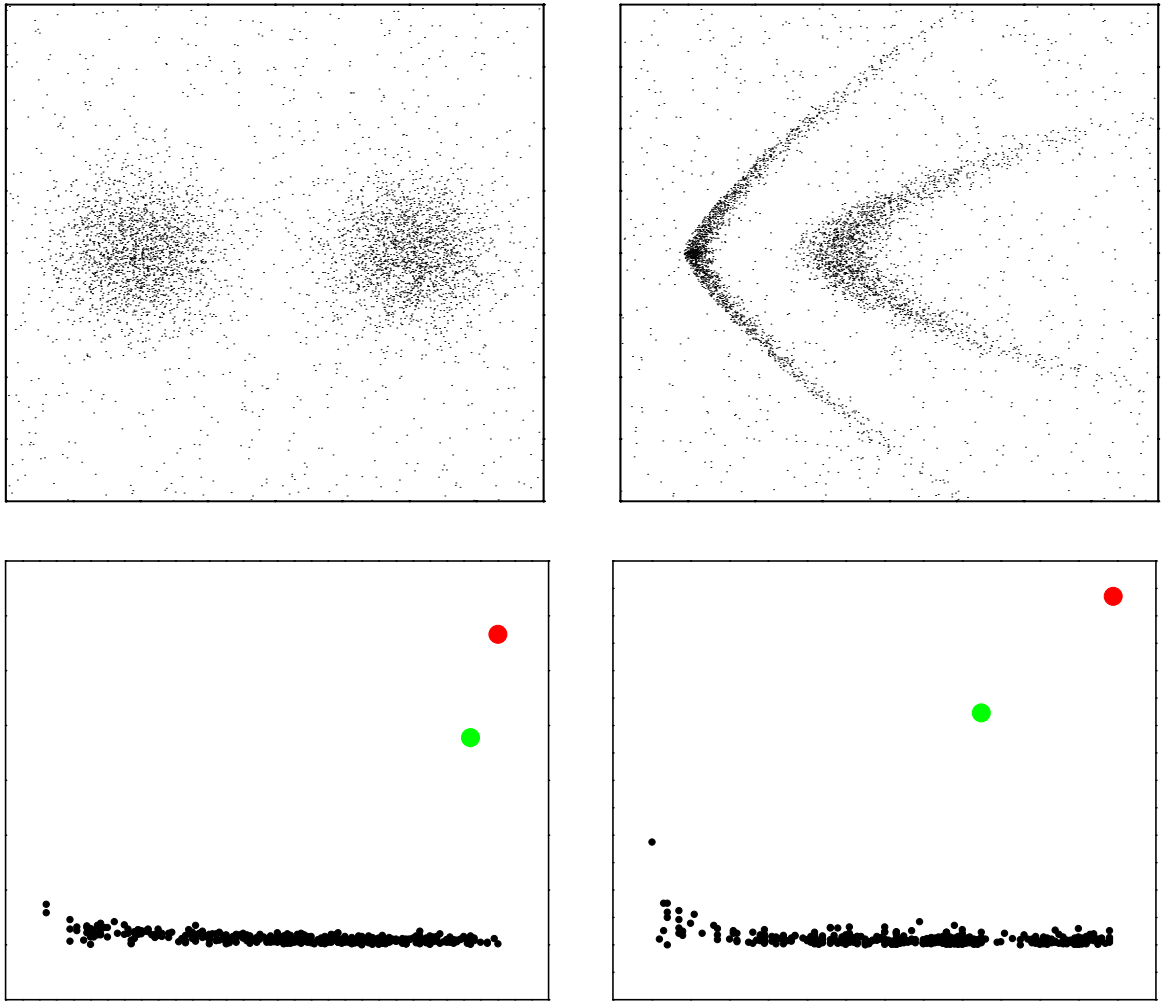


D



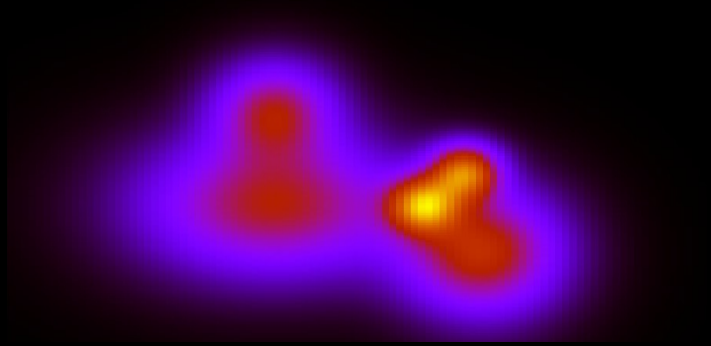
SCIENCE, 1492, vol 322 (2014)

Robust with respect to changes in the metric

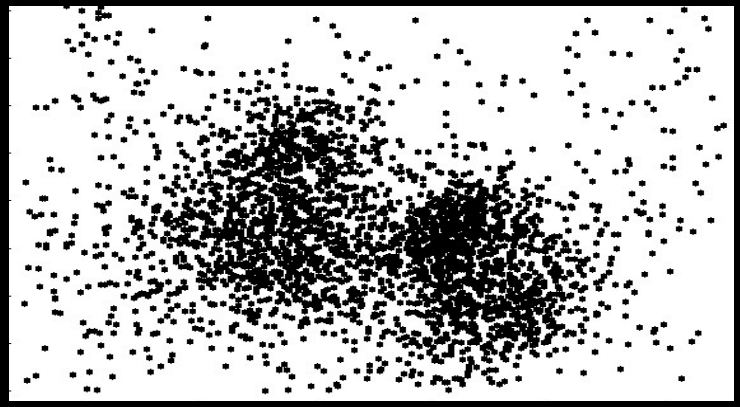


SCIENCE, 1492, vol 322 (2014)

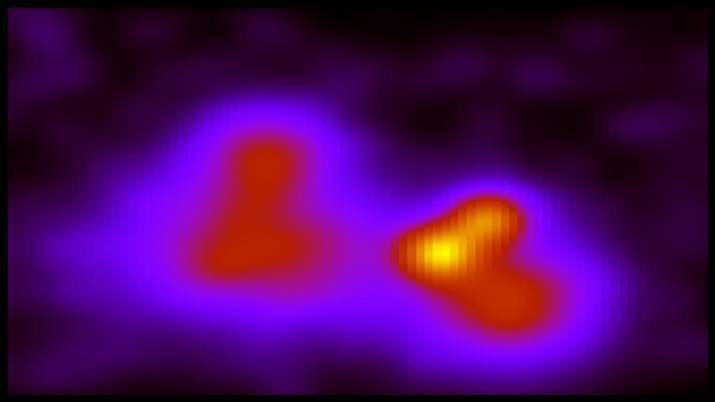
What about noise?



The true density

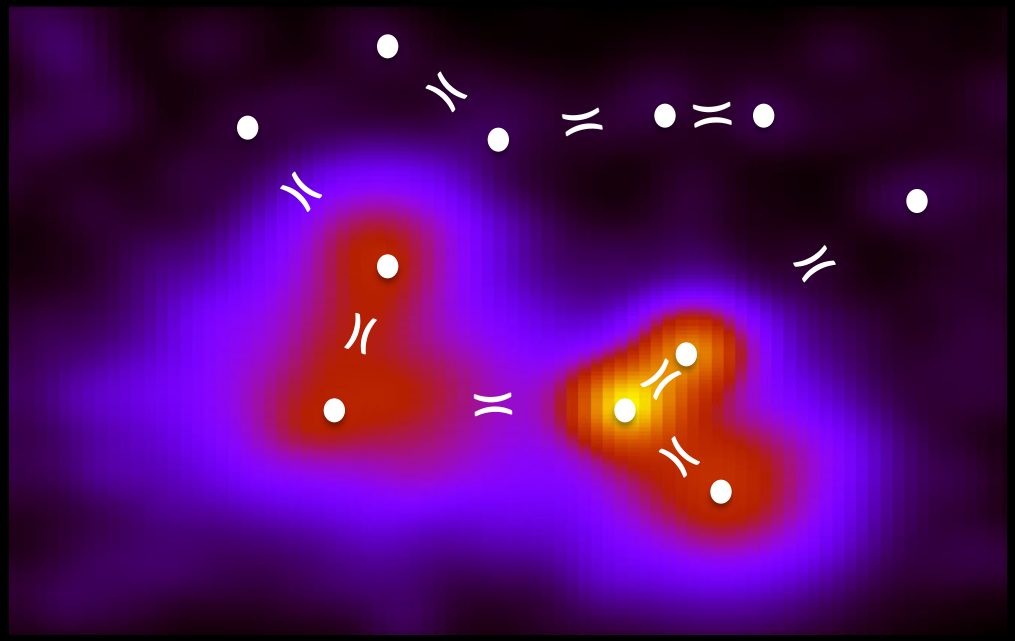


3000 points sampled
from this density



The density reconstructed
from the 3000 points
(Gaussian estimator)

What about noise?



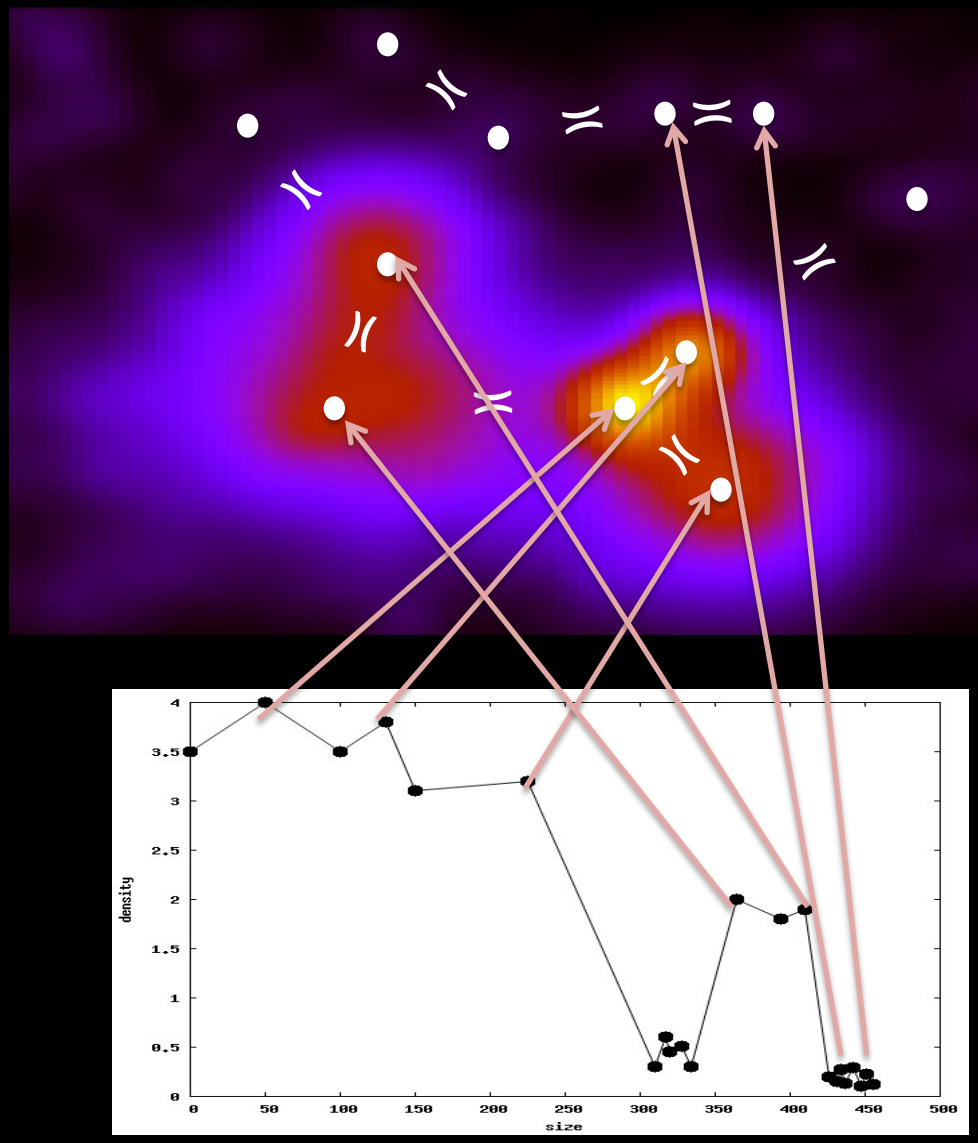
Density maxima

For each maximum, find the highest saddle point towards a higher peak

Build a tree-like structure, where each peak is a leaf, and nodes are assigned according to the density values at the saddles

How can we distinguish genuine density peaks from noise?

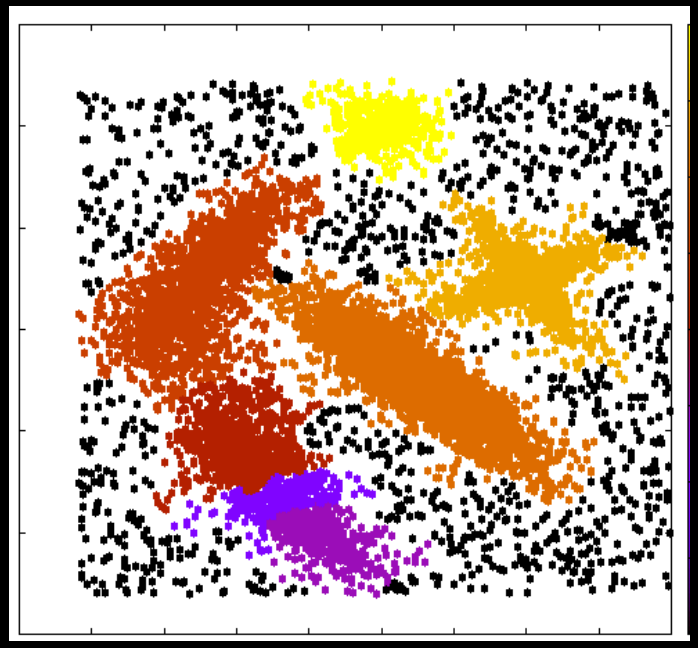
What about noise?

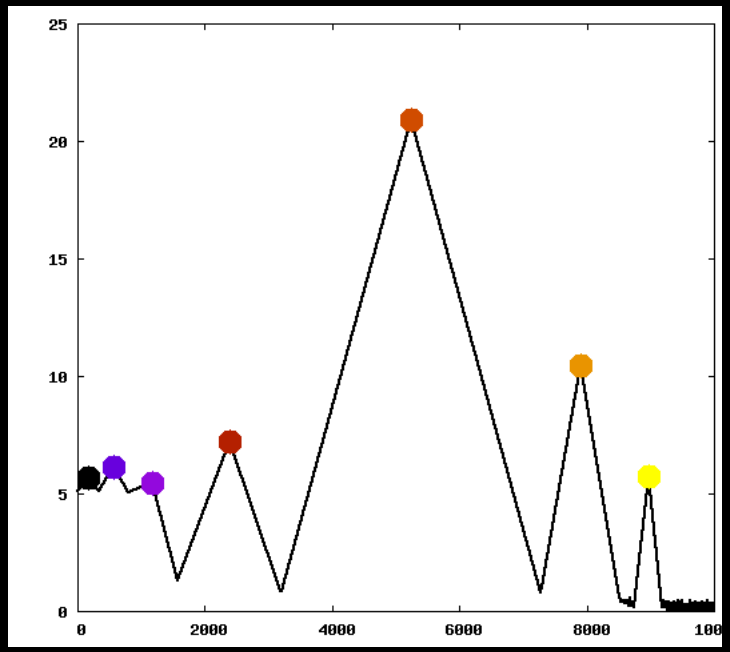


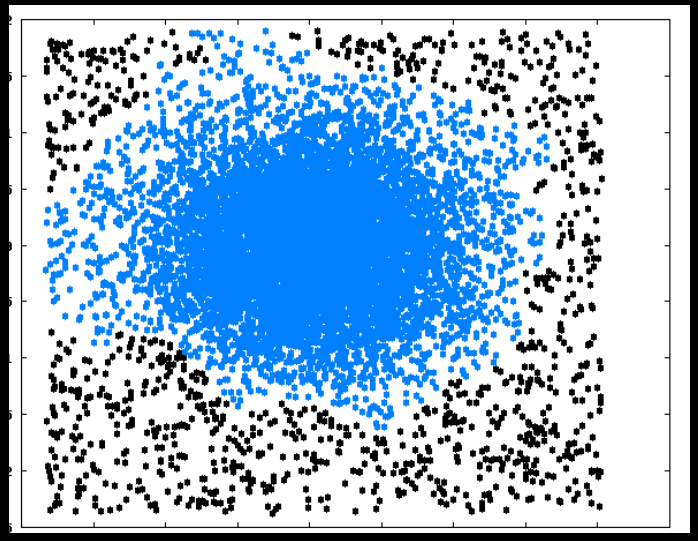
Density maxima

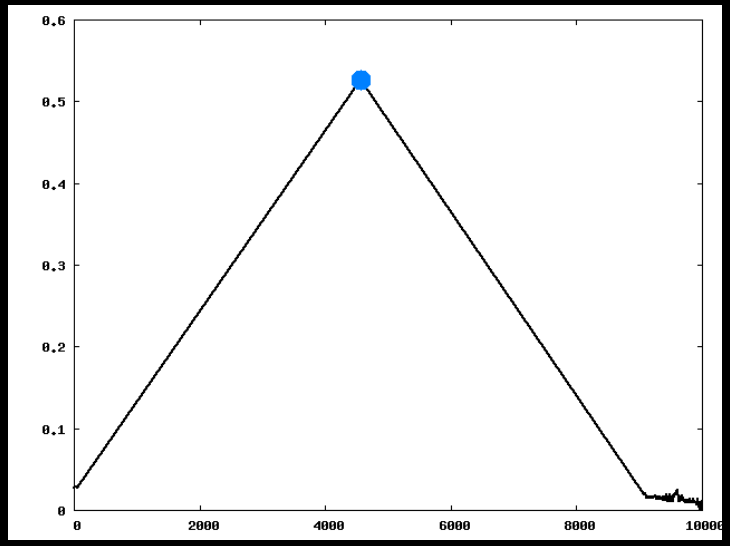
For each maximum, find the highest saddle point towards a highest peak ç

Build a tree-like structure, where each peak is a leaf, and nodes are assigned according to the density values at the saddles




25
20
15
10
5
0
0
2000
4000
6000
8000




0.6
0.5
0.4
0.3
0.2
0.1
0
0
2000
4000
6000
8000

Possible applications

This algorithm can be applied any time it is possible to define a DISTANCE function between each pair of elements in a set:

Possible applications

This algorithm can be applied any time it is possible to define a DISTANCE function between each pair of elements in a set:

- **classification of living organisms**
- marketing strategies
- libraries (book sorting)
- google search
- …
- even FACE recognition!!!

Possible applications

This algorithm can be applied any time it is possible to define a DISTANCE function between each pair of elements in a set:

- classification of living organisms
- **marketing strategies**
- libraries (book sorting)
- google search
- …
- even FACE recognition!!!

Possible applications

This algorithm can be applied any time it is possible to define a DISTANCE function between each pair of elements in a set:

- classification of living organisms
- marketing strategies
- **libraries (book sorting)**
- google search
- …
- even FACE recognition!!!

Possible applications

This algorithm can be applied any time it is possible to define a DISTANCE function between each pair of elements in a set:

- classification of living organisms
- marketing strategies
- libraries (book sorting)
- **google search**
- …
- even FACE recognition!!!

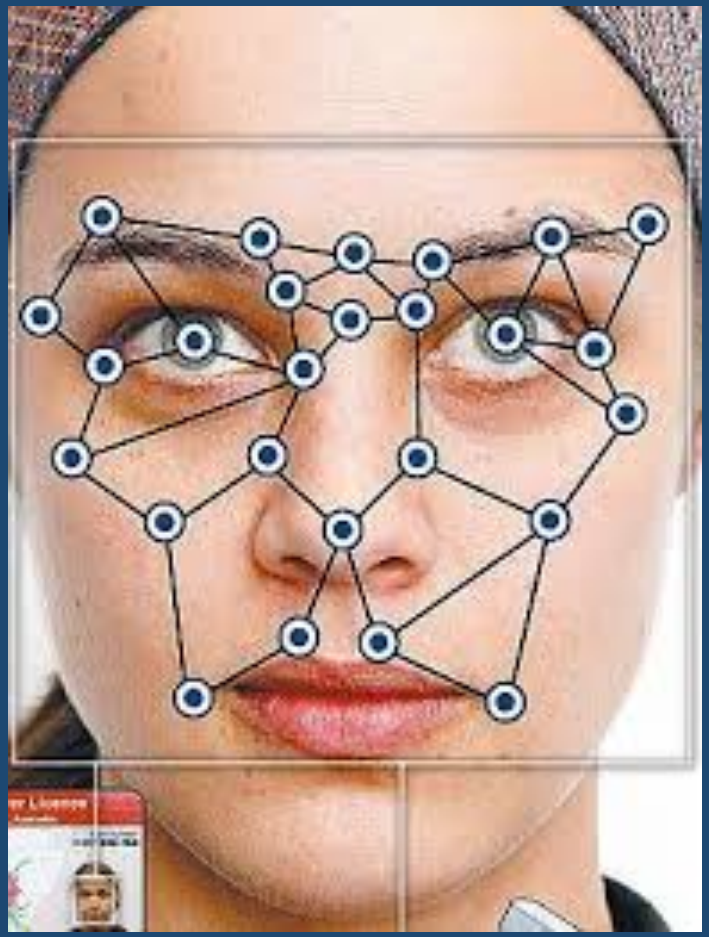
Possible applications

This algorithm can be applied any time it is possible to define a DISTANCE function between each pair of elements in a set:

- classification of living organisms
- marketing strategies
- libraries (book sorting)
- google search
- …
- **even FACE recognition!!!**

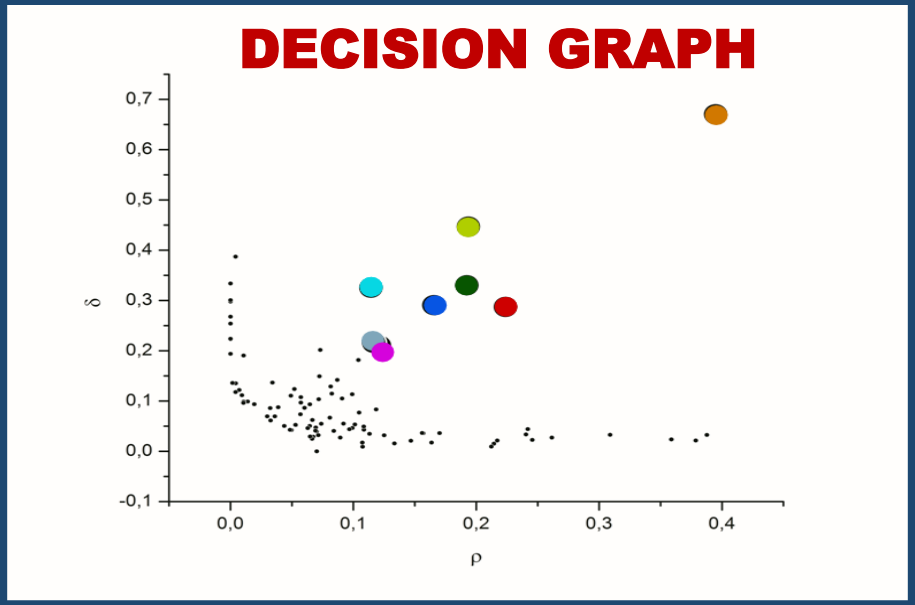
Clustering Algorithm applied to faces

Clustering Algorithm applied to faces

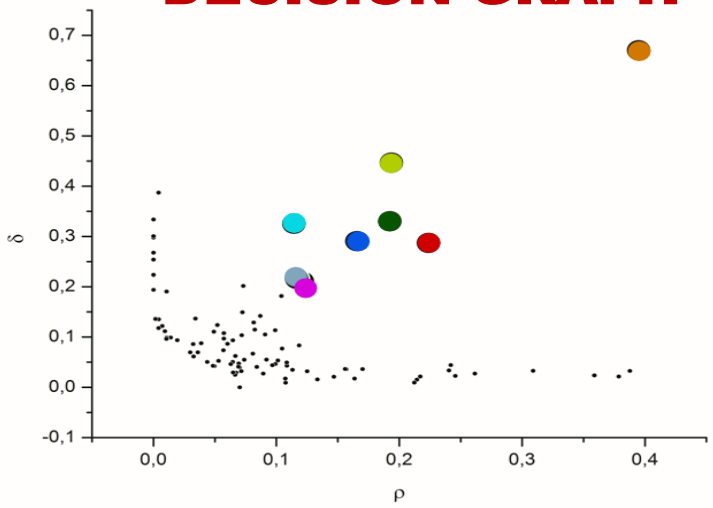


It is possible to define a "distance" between faces, based on some stable features.

*Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *Image Processing, IEEE Transactions on*, *18*(11), 2385-2401.

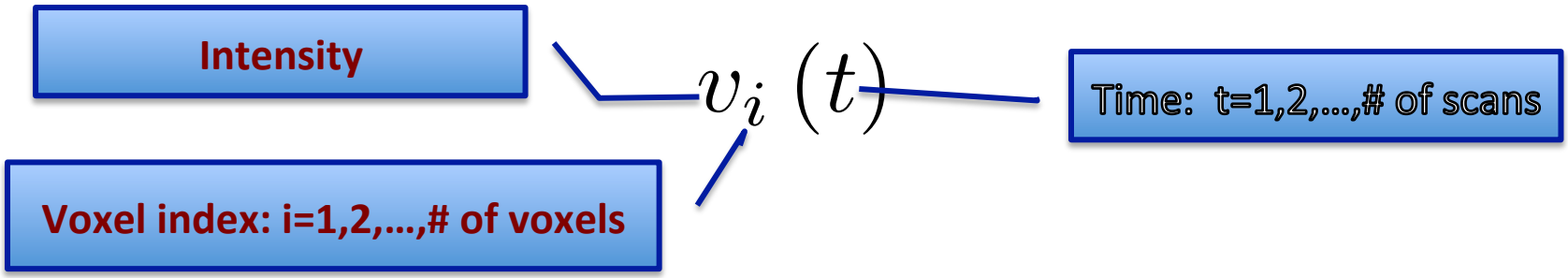

DECISION GRAPH
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
δ
0,0
0,1
0,2
0,3
0,4
ρ

DECISION GRAPH



The clustering approach at work: Analysis of a fMRI experiment (D. Amati, M. Maieron, F. Pizzagalli)

Outcome of a fMRI experiment: signal intensity for ~100,000 voxels covering densely the brain. The signal is measured every ~2 seconds for a total time of a few minutes.



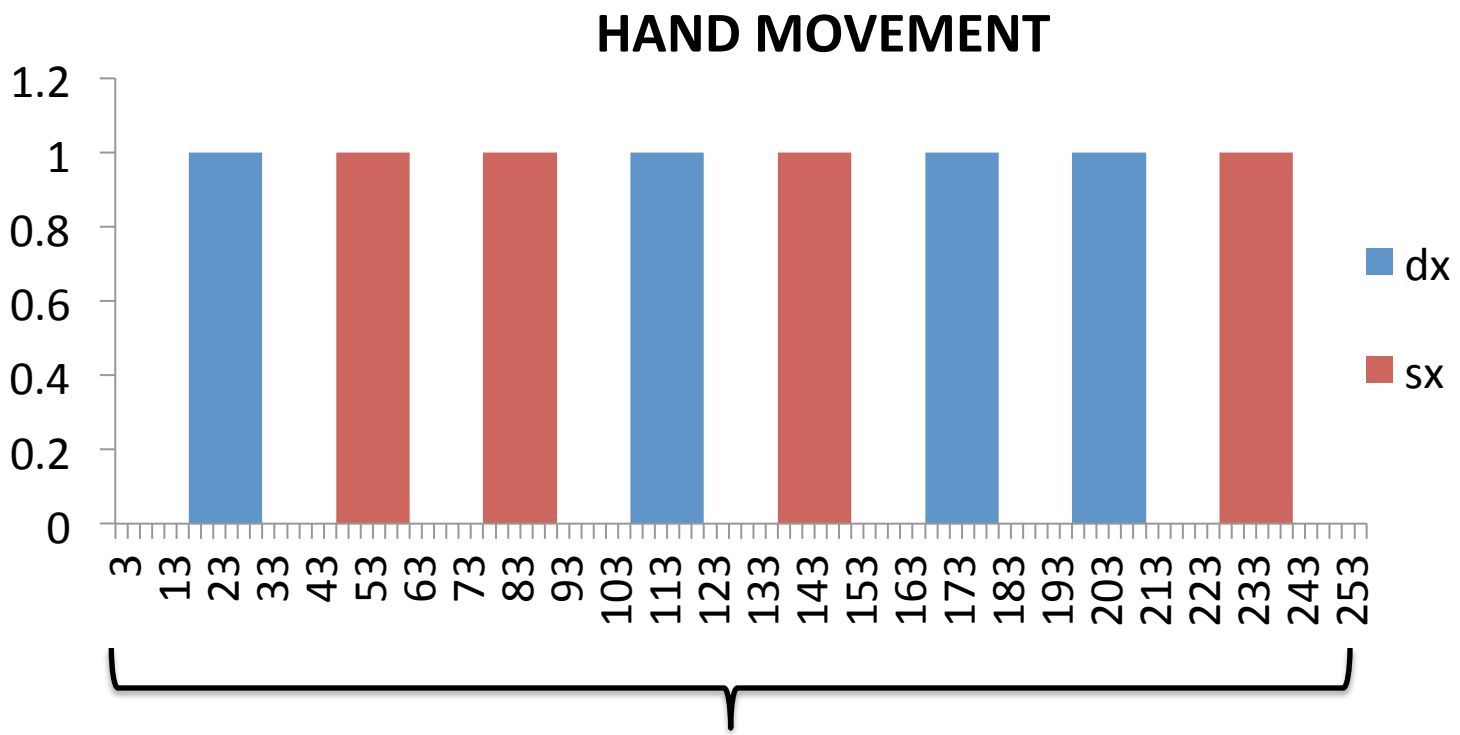
General idea: if the subject is performing a task, the voxels in the brain region involved in this task must have a similar v(t).

We look for large and connected regions with voxels with a similar v(t), namely with a similar time evolution.

Similarity measure: $$d_{ij} = \sqrt{\sum_{t=1}^{T} \left(v_i(t) - v_j(t)\right)^2}$$

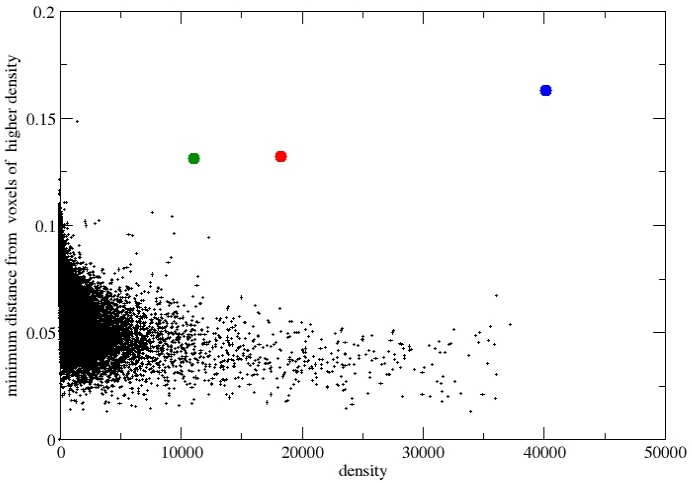
The clustering approach at work: Analysis of a fMRI experiment (D. Amati, M. Maieron, F. Pizzagalli)

The subject was scanned while moving the right or left hand. They saw the words "move left", "move right" or "stop" in a random fashion through the glasses.

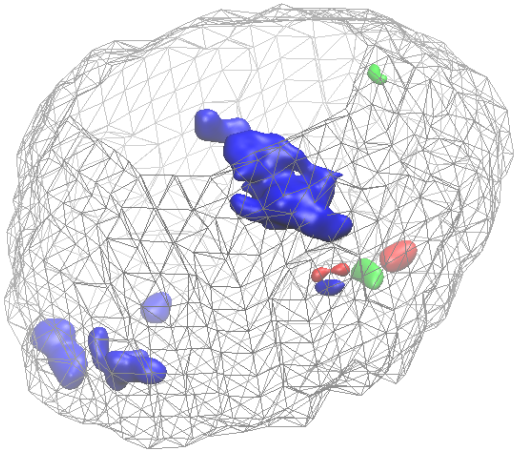


3T Achieva Philips
T2* BOLD–sensitive gradient-recalled EPI sequence
standard Head Coil 8 channels
TR/TE = 2500/32 ms
matrix 128X128 , in-plane resolution 1.8 X 1.8
#slices 34, thickness = 3mm, no gap

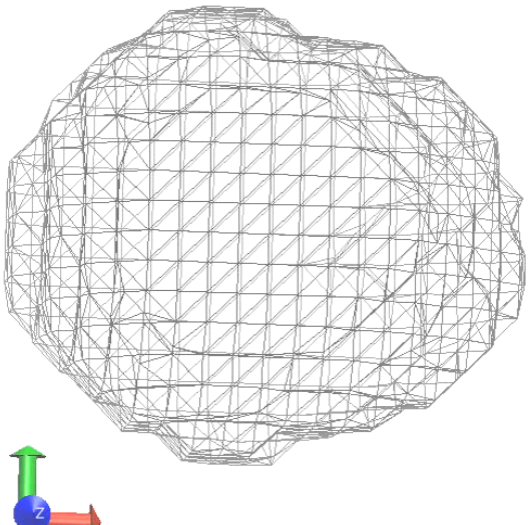
The clustering approach at work: Analysis of a fMRI experiment (D. Amati, M. Maieron, F. Pizzagalli)

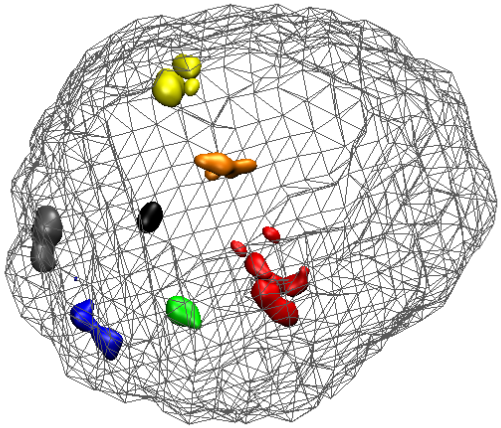


Time window 24-36: decision graph



Time window 24-36: clusters

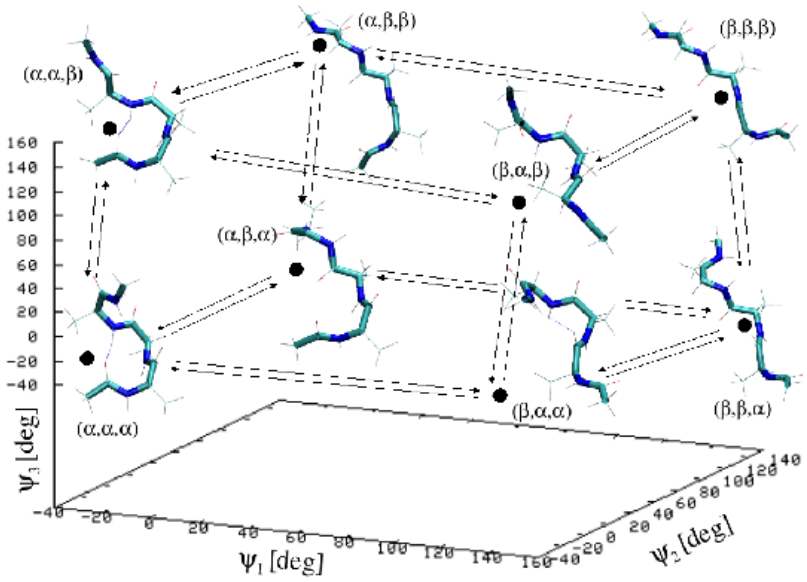


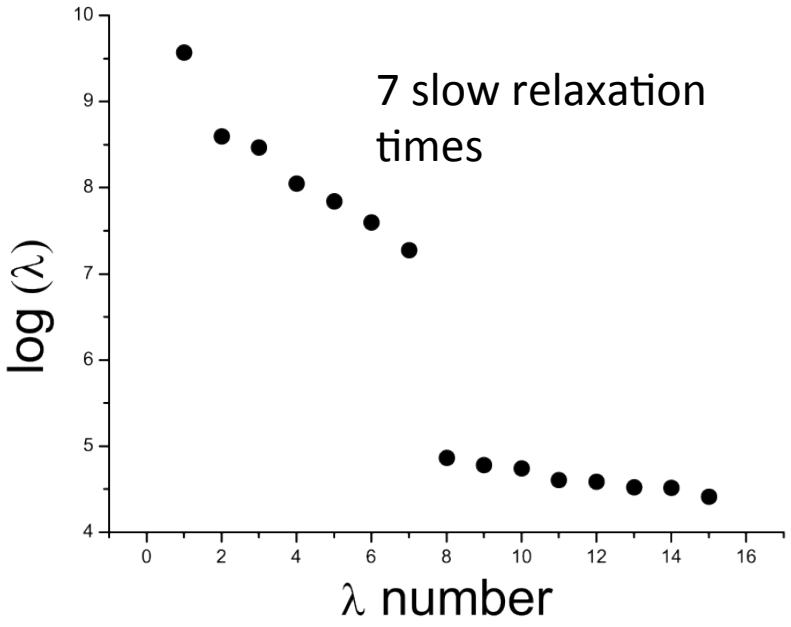


Overlap between the cluster of all the time windows

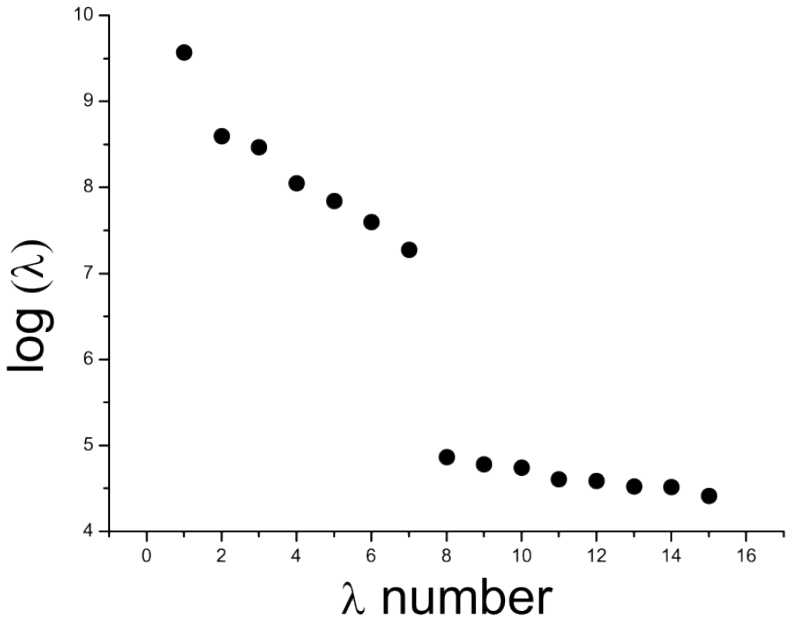
The clustering approach at work: molecular dynamics

3000 ns of molecular dynamics of 3-Ala in water solution, at 300 K

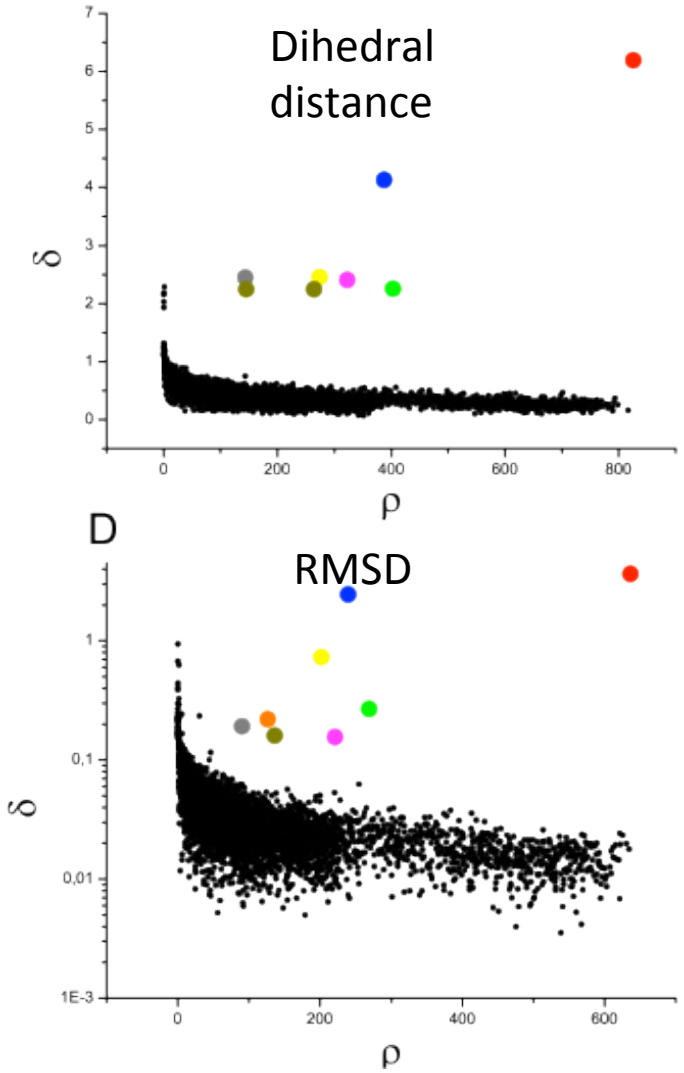


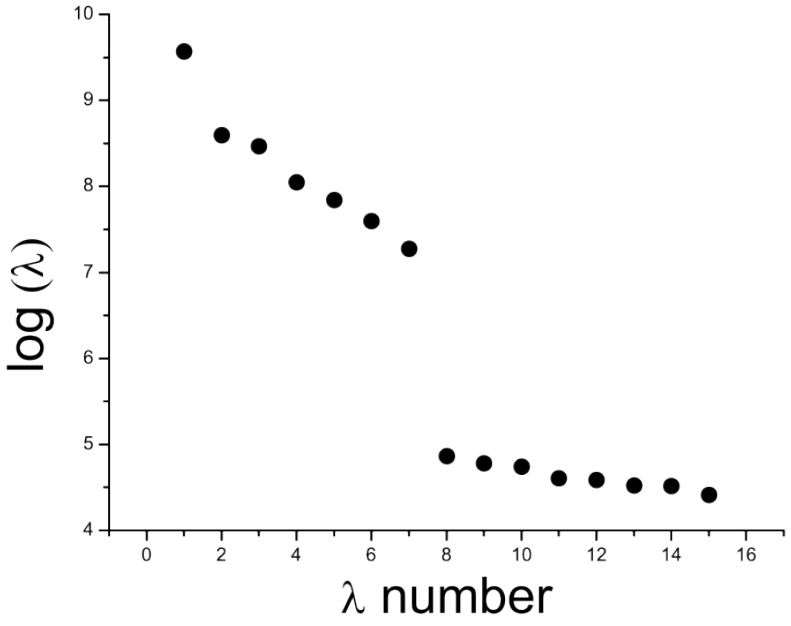


The clustering approach at work: molecular dynamics



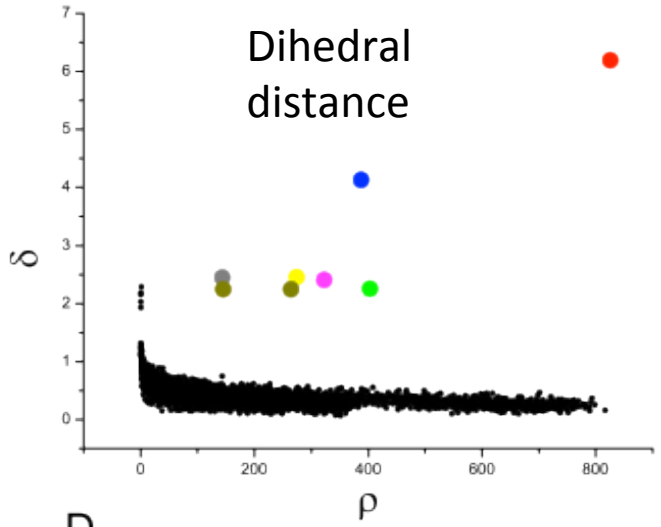
The clustering approach at work: molecular dynamics

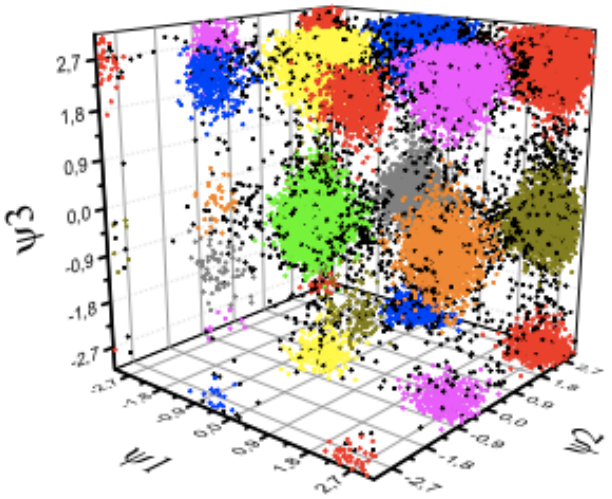




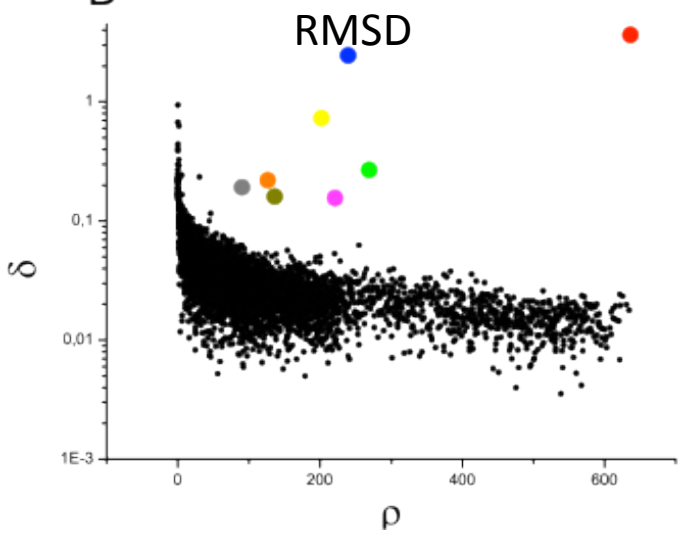
SCIENCE, 1492, vol 322 (2014)

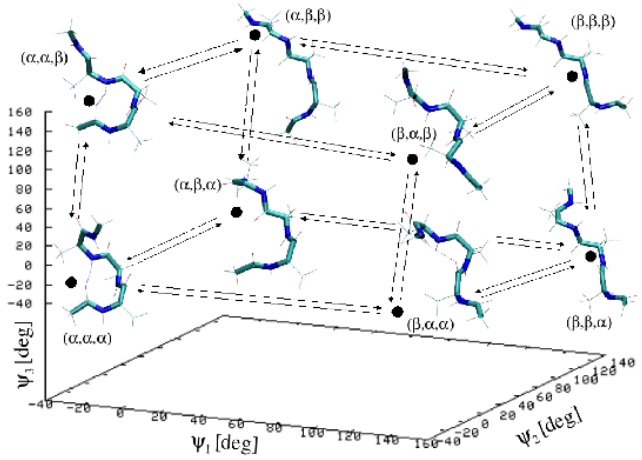
The clustering approach at work: Test on molecular dynamics





D





SCIENCE, 1492, vol 322 (2014)

Conclusions

- The approach allows detecting non-spherical clusters
- It allows detecting clusters with different densities
- Robust with respect to changes in the metric
- No optimization, no variational parameters: clusters are found deterministically from the data.
- The number of clusters is found automatically
- Outliers and background noise are automatically recognized and excluded from the analysis.

Alex Rodriguez
Maria d'Errico

Thank-you:
Daniele Amati, Erio Tosatti,
Jessica Nasica,
Francesca Rizzato