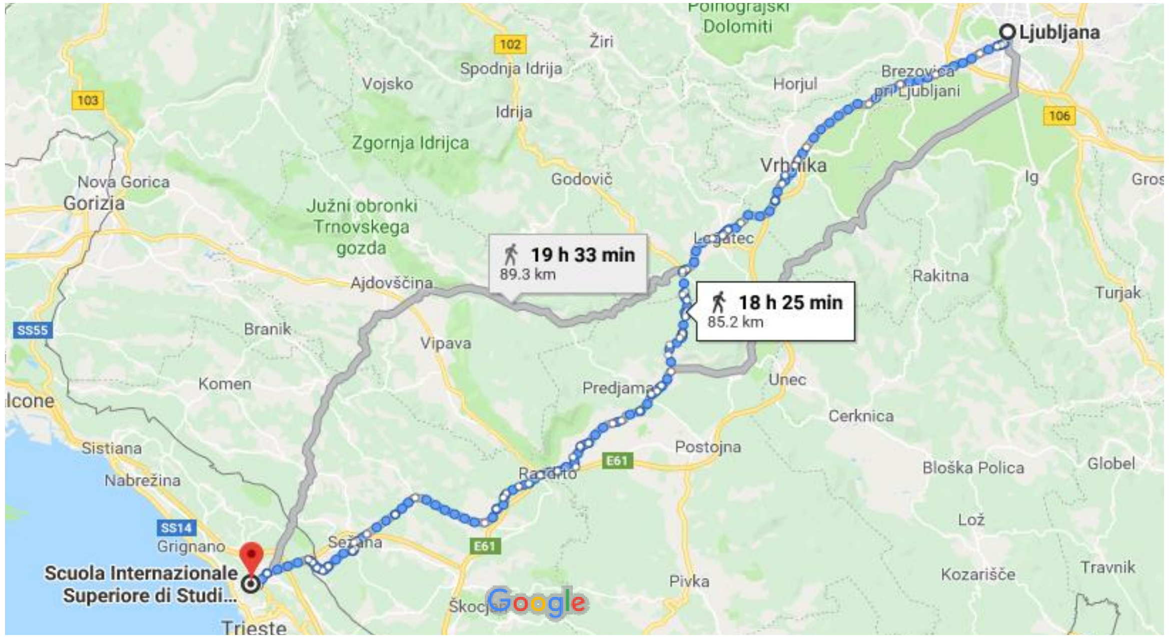
SYNTHETIC MAPS FOR NAVIGATING HIGH-DIMENSIONAL DATA SPACES

Alessandro Laio

Molecular and statistical biophysics, SISSA (Trieste)

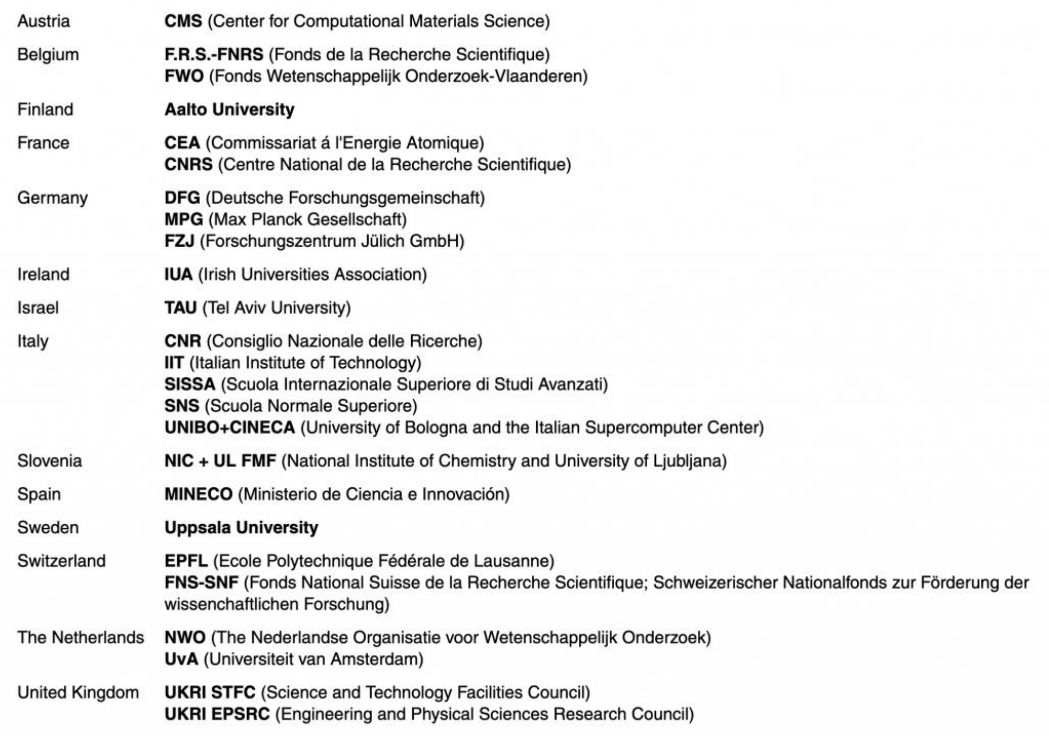
MA PER SEGUIR VIRTUTE E CANOSCENZA

- Graduate school
- ~250 PhD students in 10 PhD courses
- ~ 90 PIs, ~200 non-permanent scientific staff
- Mathematics, numerical simulations, statistical phys., cognitive neurosciences, condensed matter phys., physical chemistry, astrophysics, cosmology, data science.

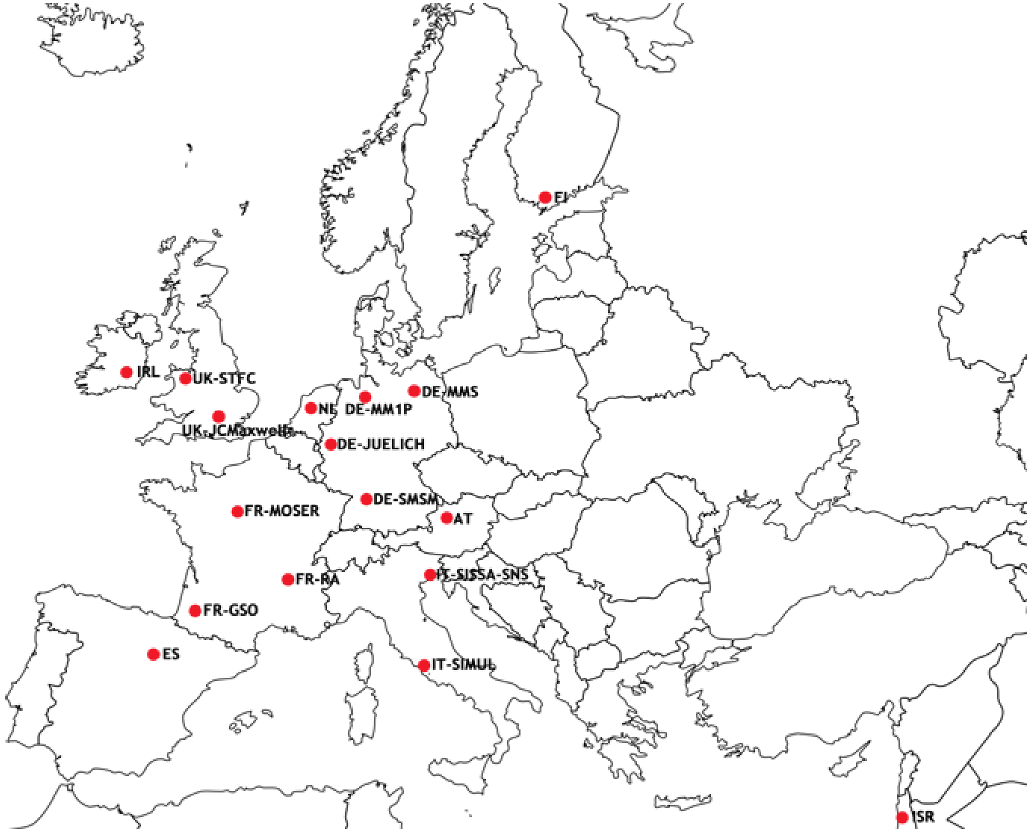


cecam

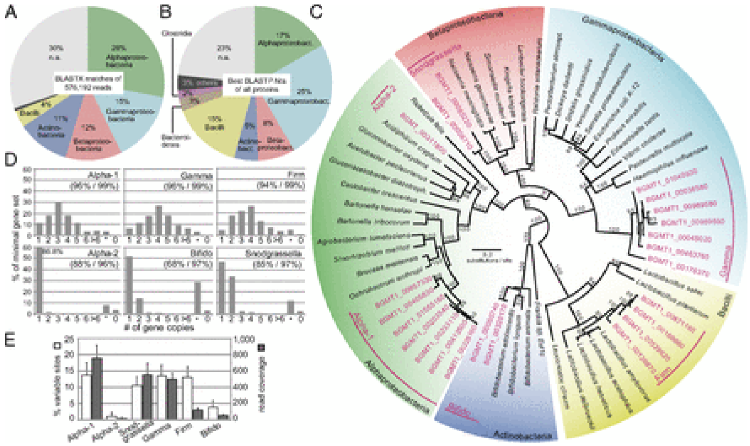
Centre Européen de Calcul Atomique et Moléculaire



Austria	**CMS** (Center for Computational Materials Science)
Belgium	**F.R.S.-FNRS** (Fonds de la Recherche Scientifique) **FWO** (Fonds Wetenschappelijk Onderzoek-Vlaanderen)
Finland	**Aalto University**
France	**CEA** (Commissariat á l'Energie Atomique) **CNRS** (Centre National de la Recherche Scientifique)
Germany	**DFG** (Deutsche Forschungsgemeinschaft) **MPG** (Max Planck Gesellschaft) **FZJ** (Forschungszentrum Jülich GmbH)
Ireland	**IUA** (Irish Universities Association)
Israel	**TAU** (Tel Aviv University)
Italy	**CNR** (Consiglio Nazionale delle Ricerche) **IIT** (Italian Institute of Technology) **SISSA** (Scuola Internazionale Superiore di Studi Avanzati) **SNS** (Scuola Normale Superiore) **UNIBO+CINECA** (University of Bologna and the Italian Supercomputer Center)
Slovenia	**NIC + UL FMF** (National Institute of Chemistry and University of Ljubljana)
Spain	**MINECO** (Ministerio de Ciencia e Innovación)
Sweden	**Uppsala University**
Switzerland	**EPFL** (Ecole Polytechnique Fédérale de Lausanne) **FNS-SNF** (Fonds National Suisse de la Recherche Scientifique; Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung)
The Netherlands	**NWO** (The Nederlandse Organisatie voor Wetenschappelijk Onderzoek) **UvA** (Universiteit van Amsterdam)
United Kingdom	**UKRI STFC** (Science and Technology Facilities Council) **UKRI EPSRC** (Engineering and Physical Sciences Research Council)

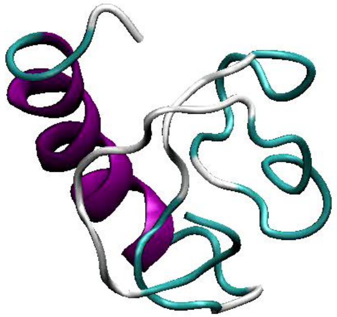


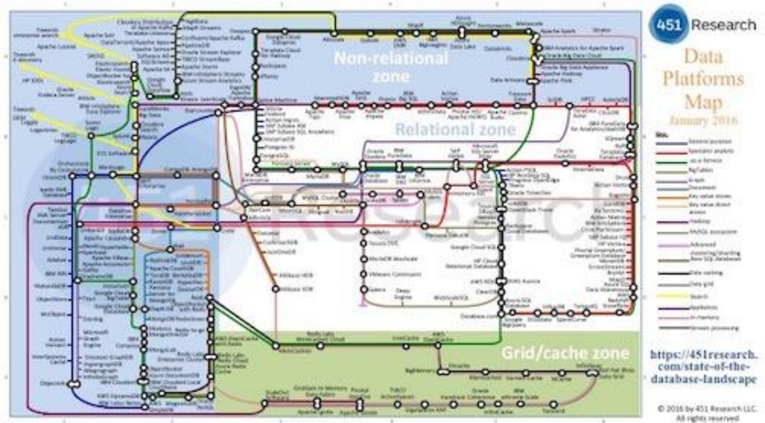
Complex data landscapes are everywhere



Microbiota of the honey bee (P. Engel et al, PNAS, 109, 11002 (2012)

The configuration space of a 60-residue polypeptide explored by atomistic simulations (P. Cossio et al, Plos Comp. Biol. (2011)

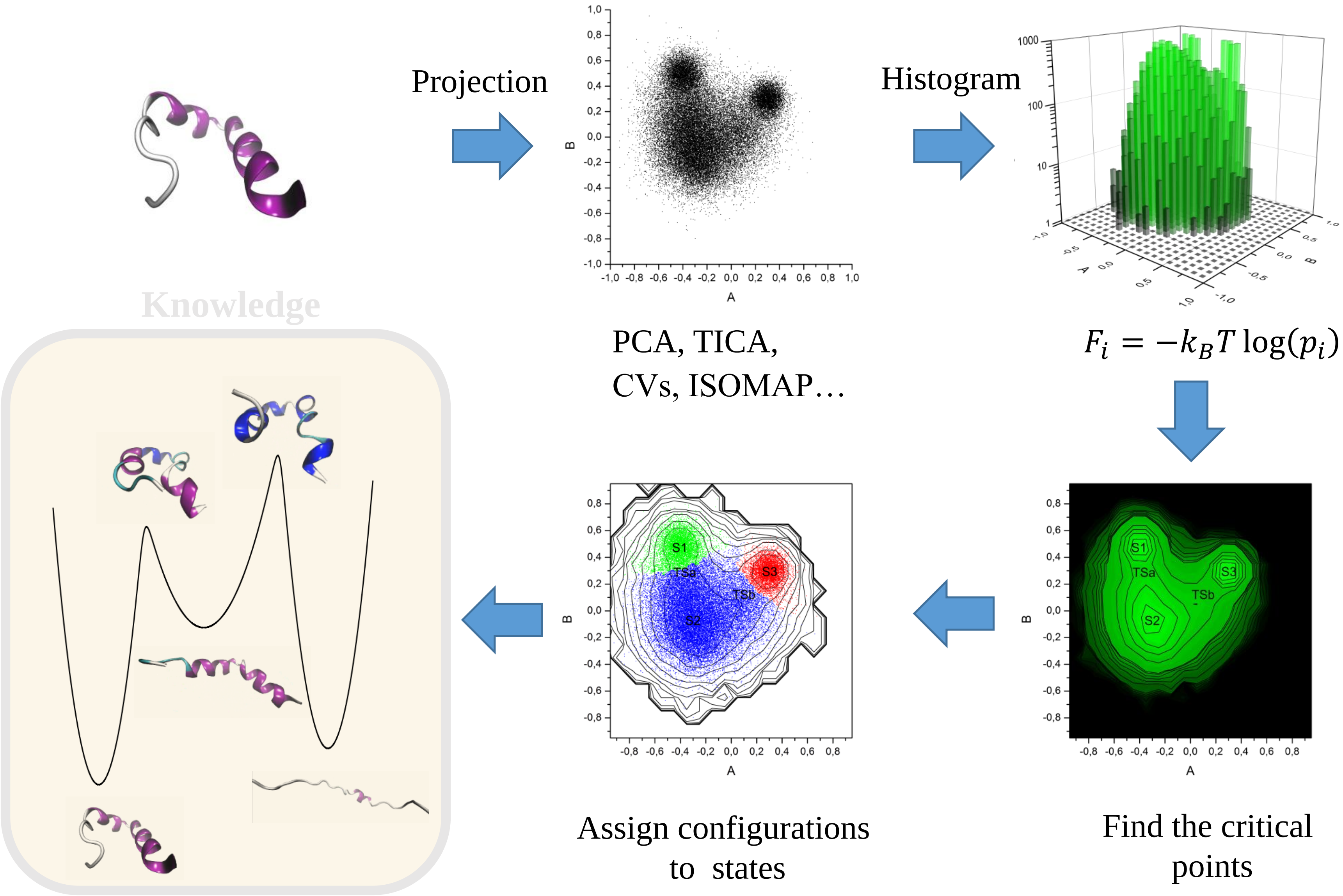



451 Research
Data Platforms Map
January 2016
Non-relational zone
Relational zone
Grid/cache zone
https://451research.com/state-of-the-database-landscape
© 2016 by 451 Research LLC. All rights reserved

Database landscape map: https://451research.com/state-of-the-database-landscape

How can I get a low-dimensional map from my data?

How can I get a low-dimensional map from my data?



Projection

PCA, TICA,
CVs, ISOMAP…

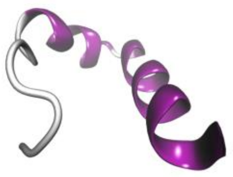
Histogram

$$F_i = -k_B T \log(p_i)$$

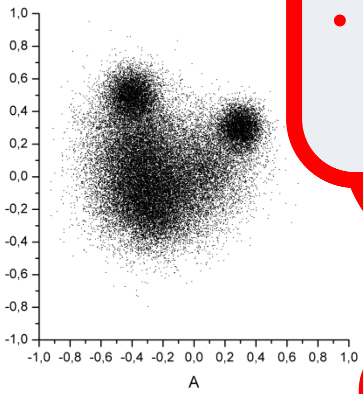
Find the critical points

Assign configurations to states

PROBLEMS



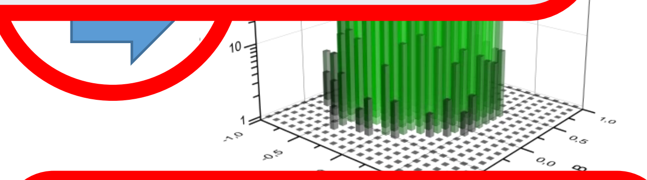
Projection



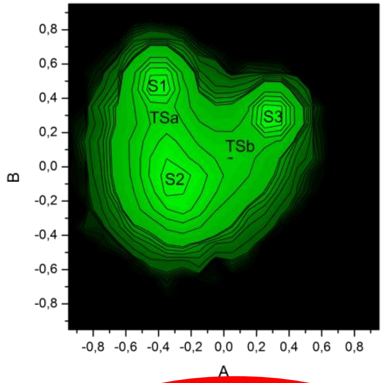
Choose the collect

variables…

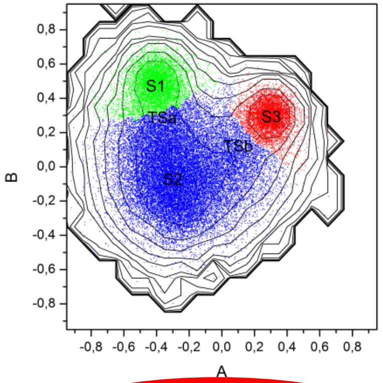
- *Which dimension?*
- *Which variables?*



- *Impossible in high dimension*
- *Binning parameters*

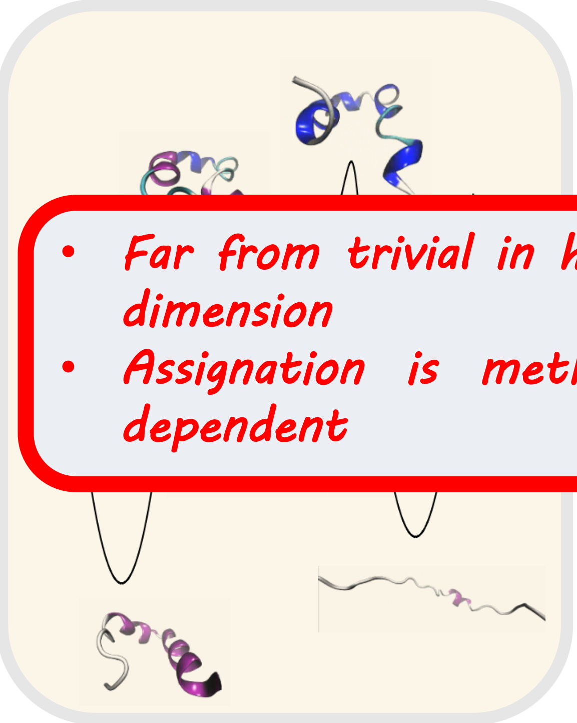


Find the critical points



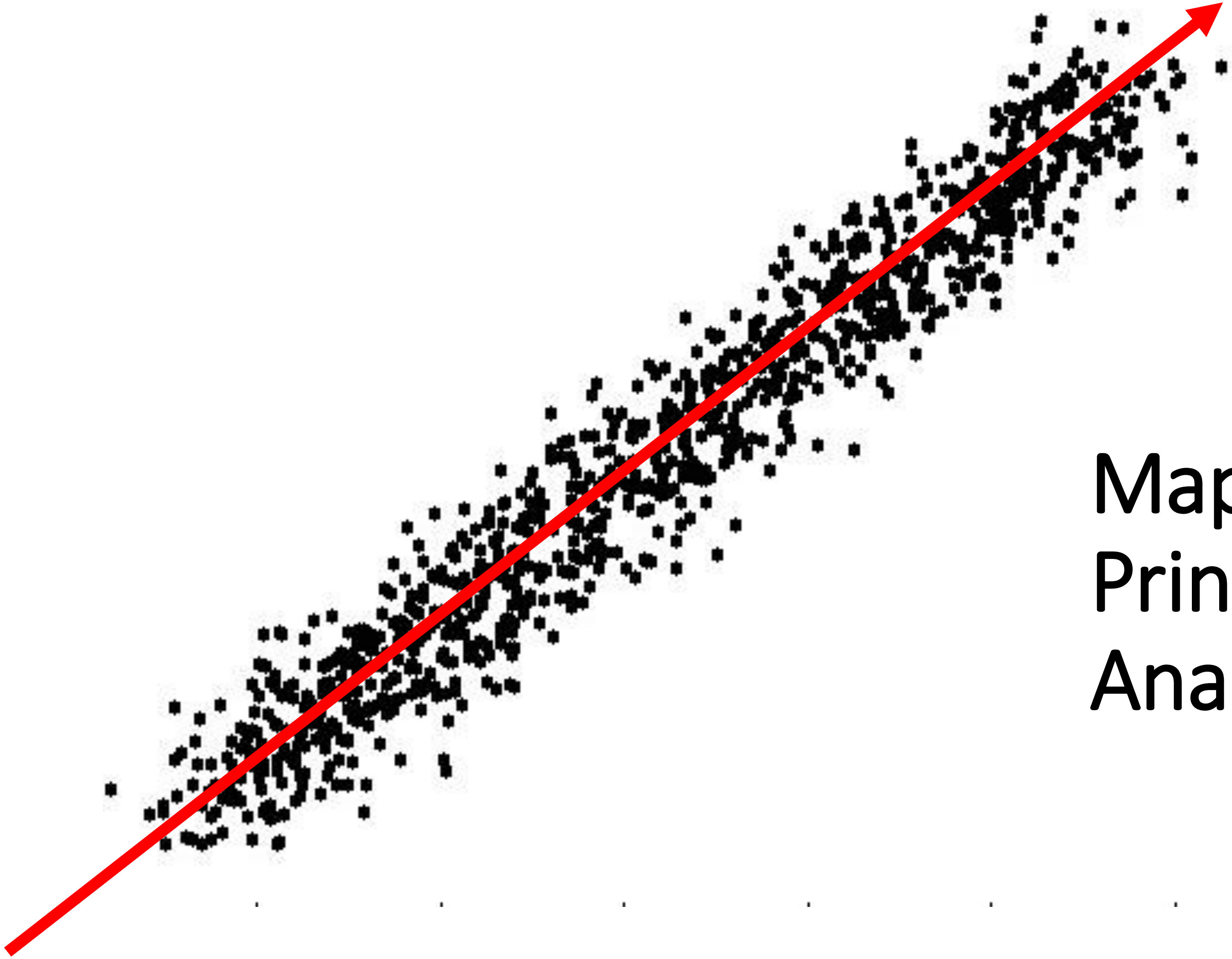
Assign configurations to states

Knowledge



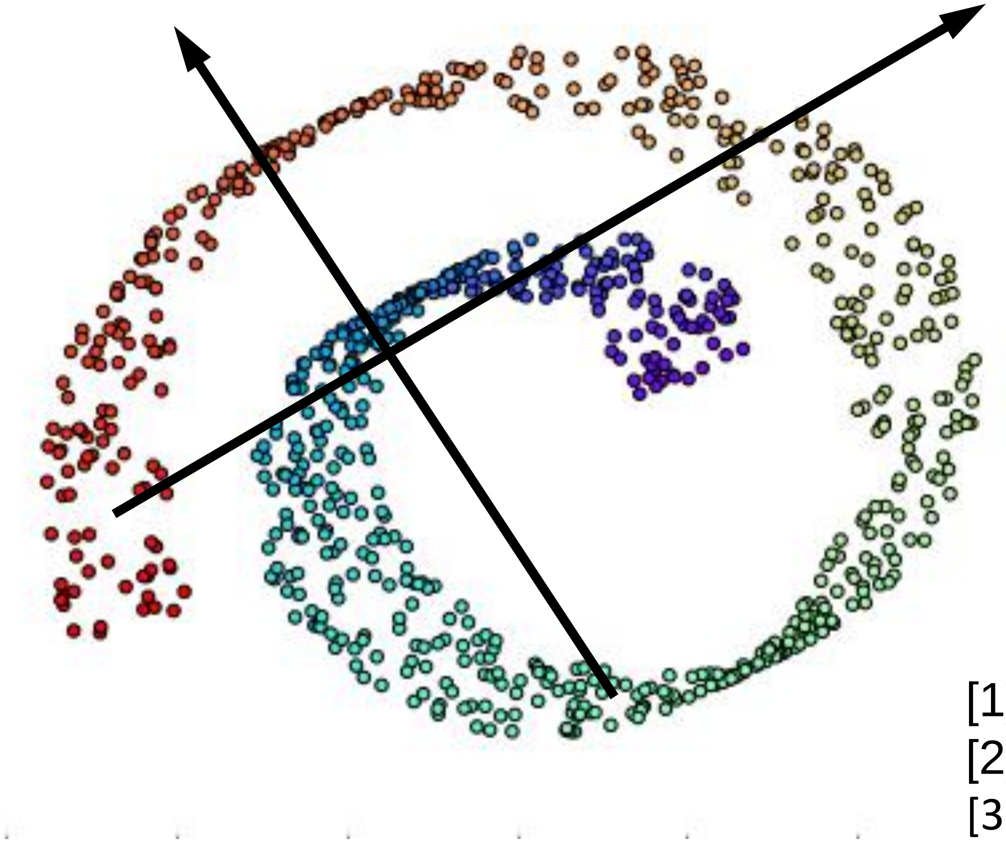
- *Far from trivial in high dimension*
- *Assignation is method dependent*

A critical difficulty: projection (without choosing the collective variables)



Map it to d=1 by Principal Component Analysis (PCA)

What happens if the manifold containing the data is curved?



Here PCA cannot work

Non-linear projection:

- Kernel PCA [1]
- Diffusion map [2]
- Local Linear Embedding [3]
- Isomap [4]
- Sketch map [5]

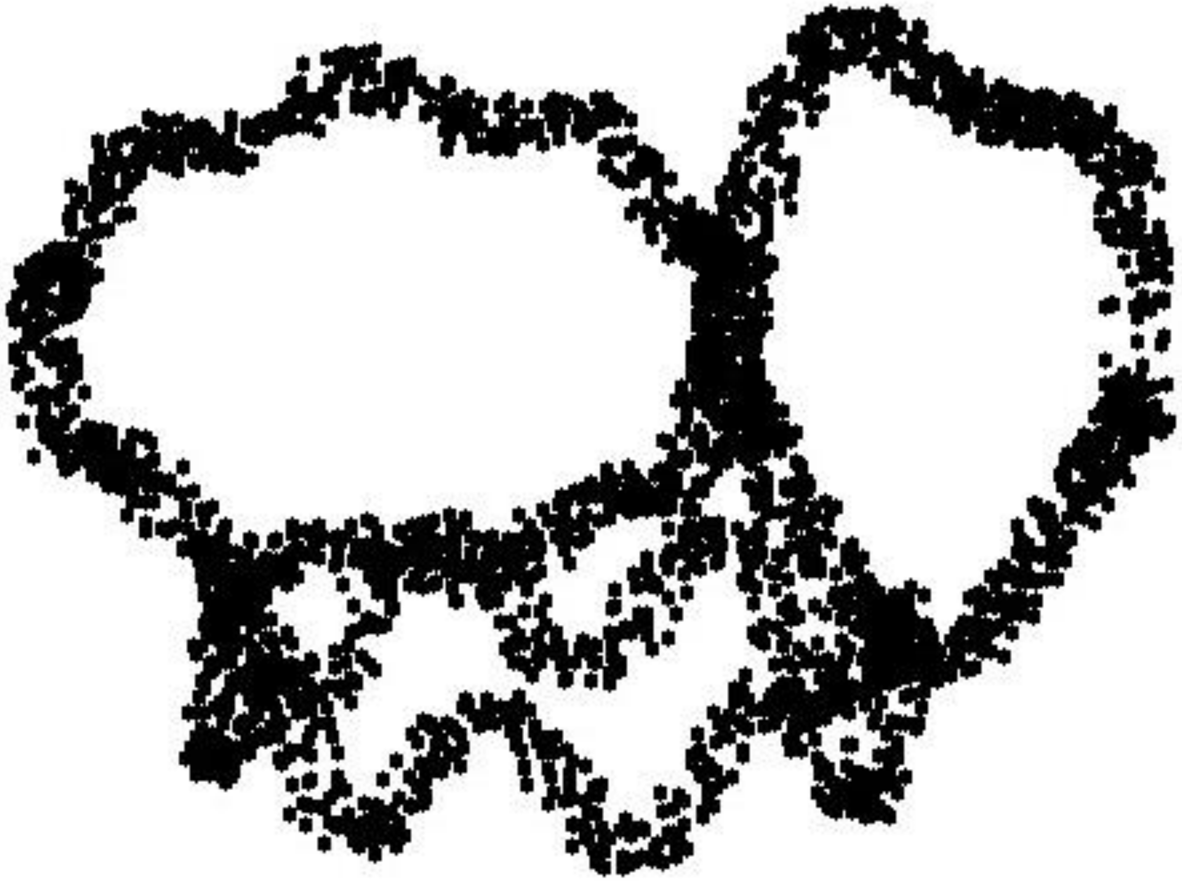
[1] Nat. Biotechnol. 2008, 26, 303– 304.

[2] Proc. Natl. Acad. Sci. USA 2005, 102, 7426–7431.

[3] Science 2000, 290, 2323–2326.

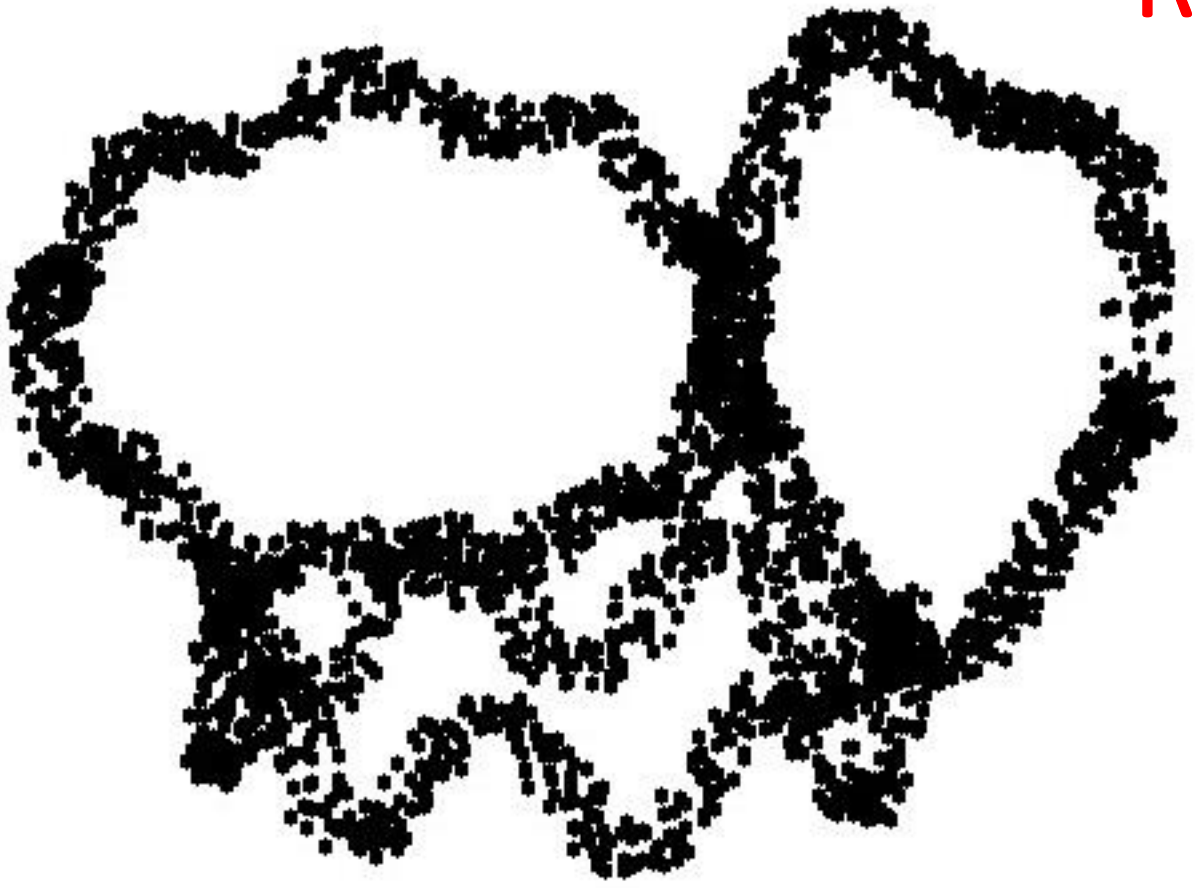
[4] Science 2000, 290, 2319–2323.

[5] Proc. Natl. Acad. Sci. USA 2011, 108, 16916–16921



What now? Approximately one dimensional, but can we map this to a line?

Real-world data:



Curved and twisted hypersurfaces

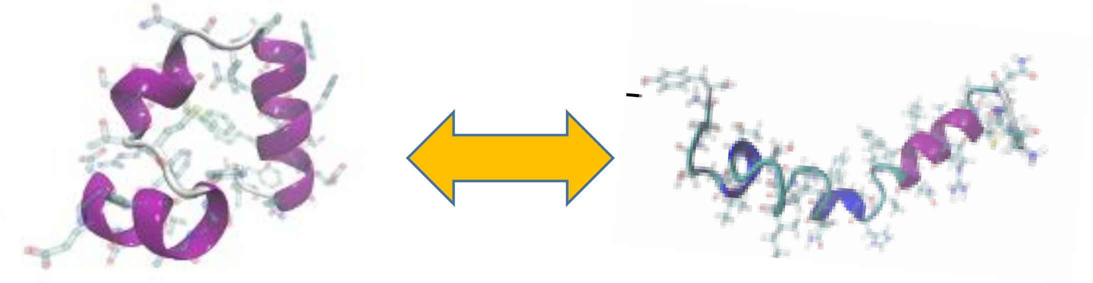
Complex topologies (no hyperplanes)

Local dimension of the embedding manifold of ~10 or more

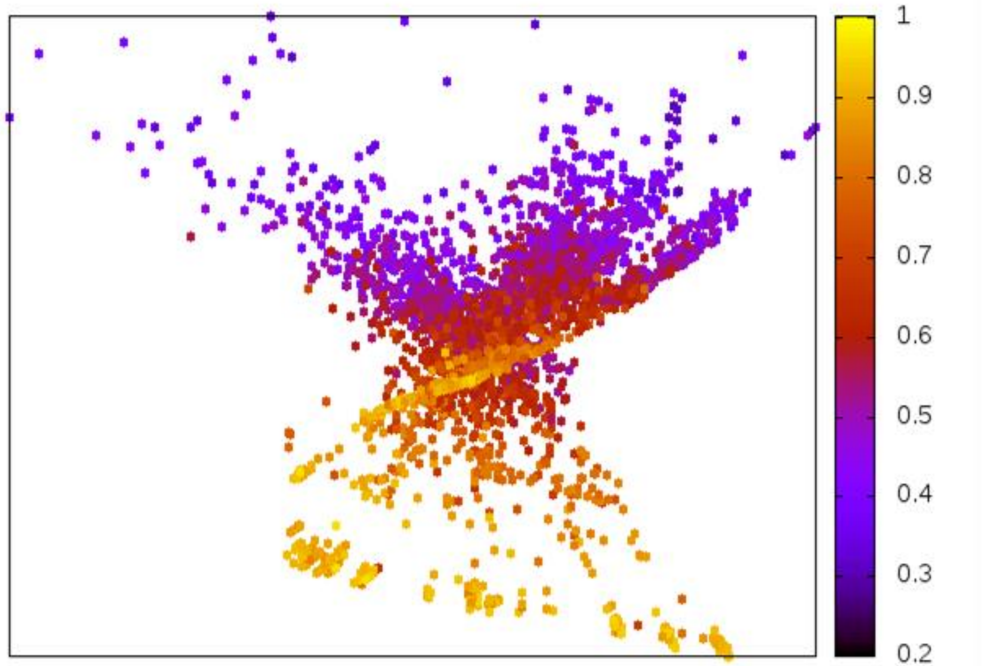
Mapping to d=2 or 3 is normally meaningless

… still, let’s do it!!!!!!!!!!!

Folding of a 32-residue protein (Villin headpiece)

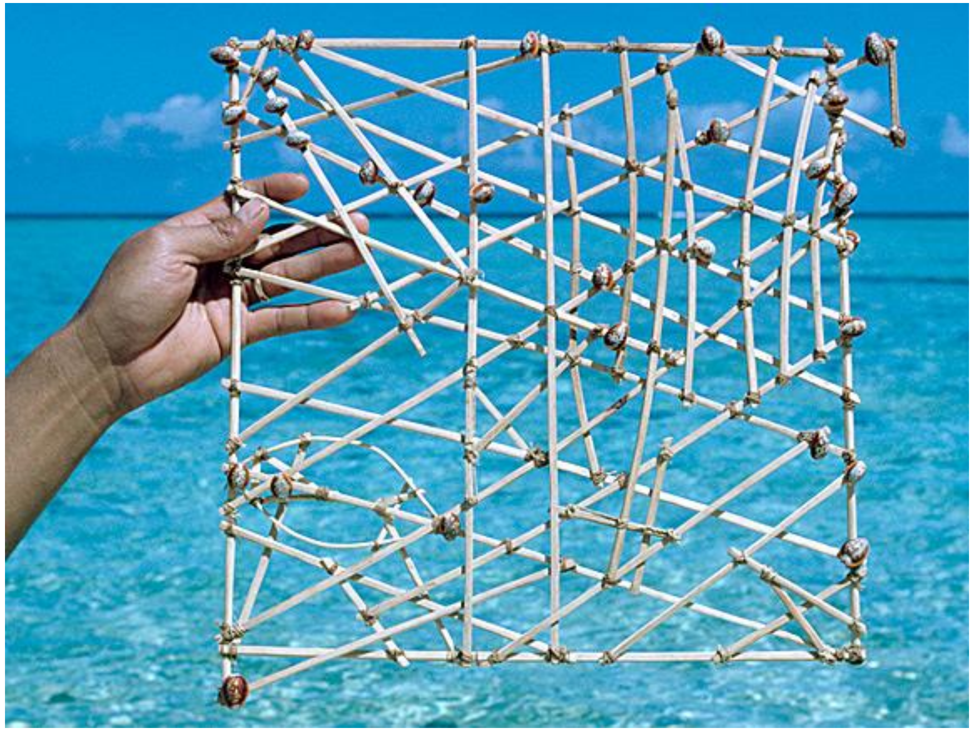


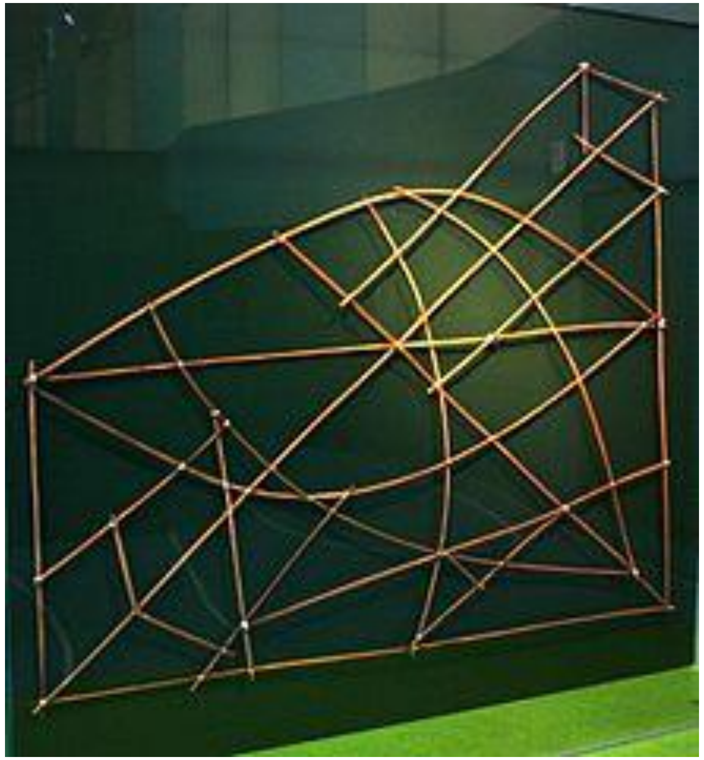
- 0.4 ms of molecular dynamics
- ~32000 configurations
- ~1000 atoms+ solvent
- Project to two dimensions by ISOMAP [Science 2000, 290, 2319–2323]



Mapping to d=2 or 3 is normally meaningless

Learn from Marshallese sailors

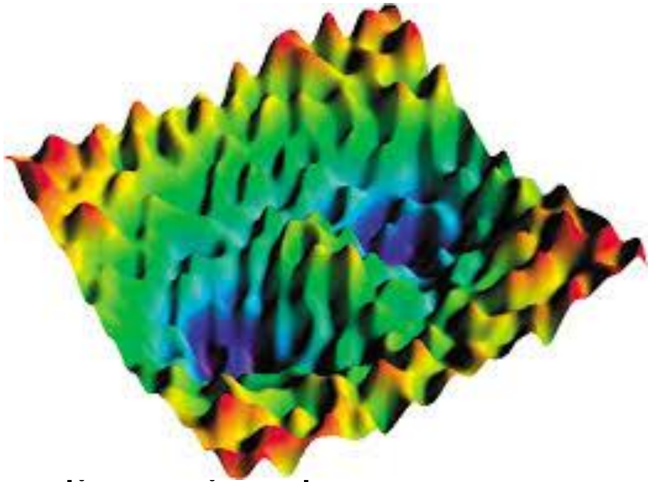




It is not a proper map: it conveys at the same time information about islands positions and sea swell directions

Our perspective on data landscapes

- Data are generated from a high-dimensional probability distribution.



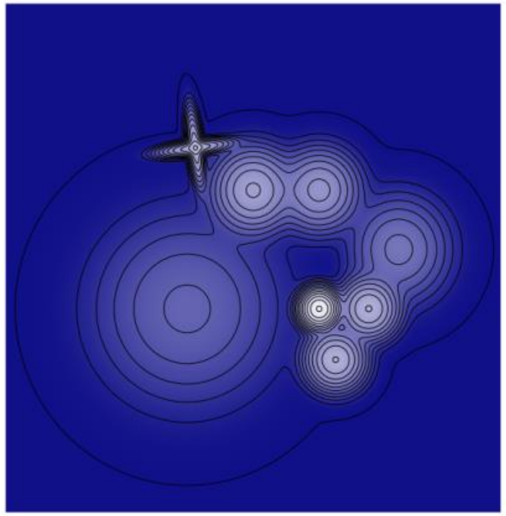
We do not attempt projecting the data on a low-dimensional manifold (like in PCA, etc)

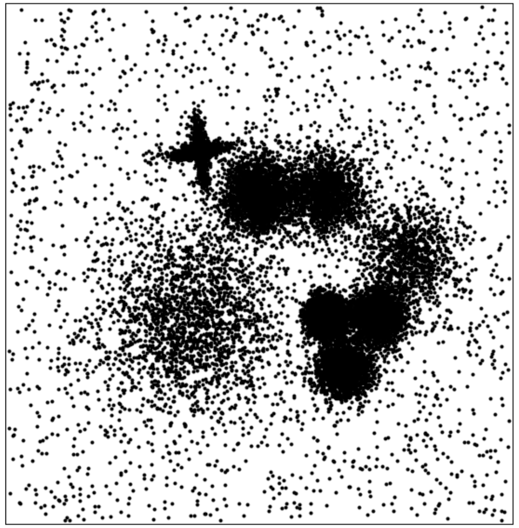
- We build a topography of the landscape, namely a list of probability peaks, and of the saddle points connecting them

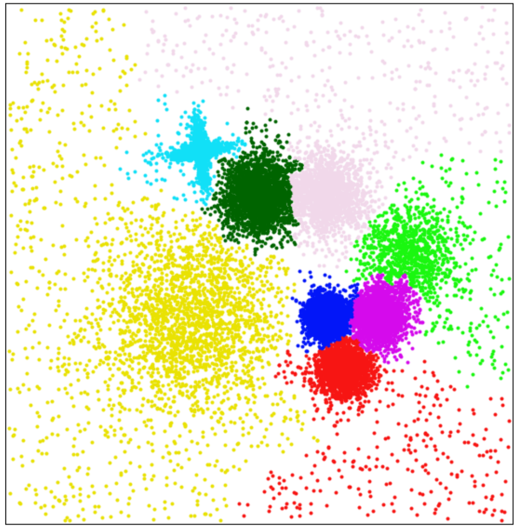
A compact representation of this topography is possible even if the data are embedded in a high-dimensional manifold

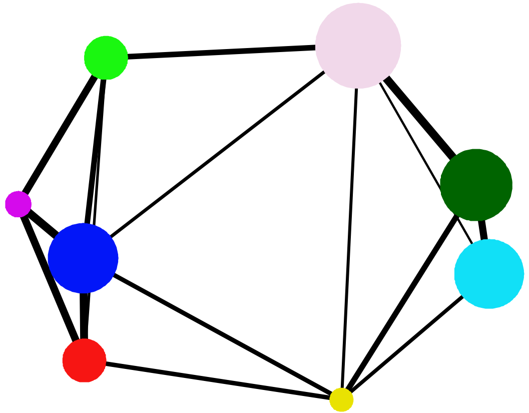
The topography of a data landscape

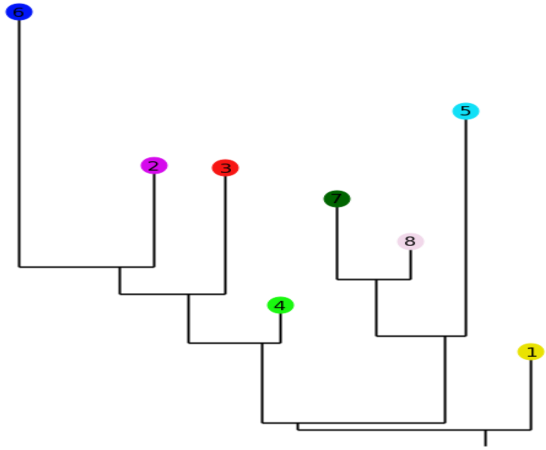
a list of properties of all the probability peaks, and of the saddle points connecting them





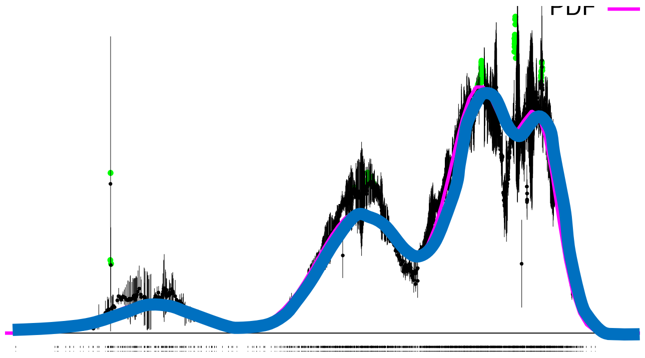




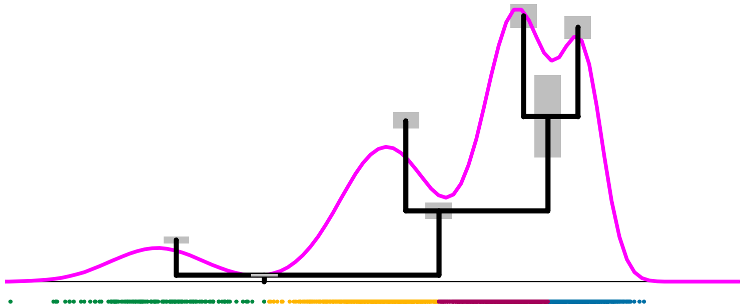


Building a topography of a data landscape

- We first compute the dimension of the manifold containing the data [Sci Rep. 12140, vol 7 (2017)]
  - We estimate the probability density at each data point. [JCTC in press (2018)]

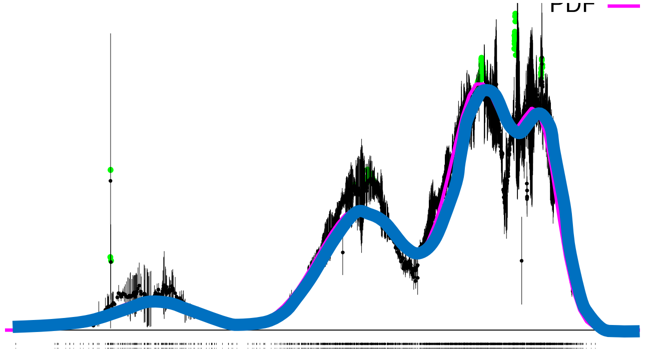


- We then find the probability maxima by Density Peak clustering [Science, 1492, vol 322 (2014)]
  - We compute the probability at the boundary between each pair of maxima.
- Graphical representation of the topography

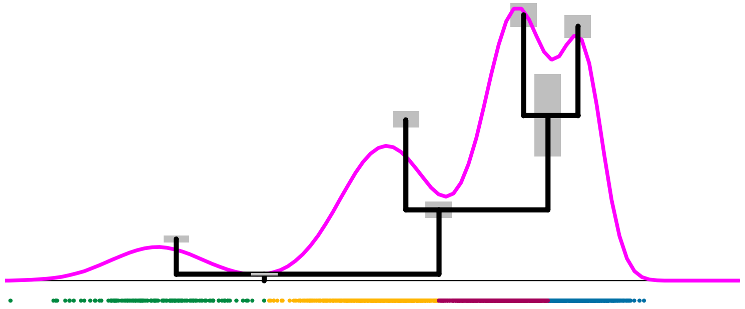


Building a topography of a data landscape

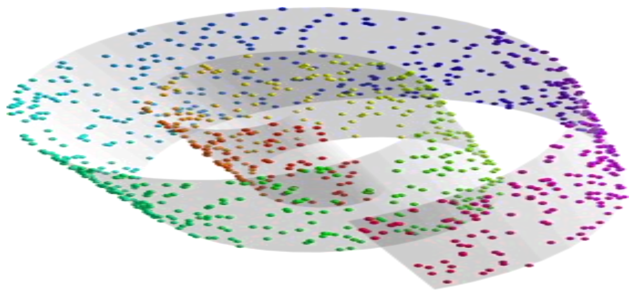
- We first compute the dimension of the manifold containing the data [Sci Rep. 12140, vol 7 (2017)]
    - We estimate the probability density at each data point. [JCTC in press (2018)]



- We then find the probability maxima by Density Peak clustering [Science, 1492, vol 322 (2014)]
    - We compute the probability at the boundary between each pair of maxima.
- Graphical representation of the topography



Estimate the intrinsic dimension of the data set



- ID: Minimum number of parameters required to describe the data while minimizing the information loss.
- Many methods for estimating it are based on the scaling of the number of neighbors with the distance [1]
- We developed an estimator based only in the distance from the two first nearest neighbors (Facco *et al*, *Sci. Rep*. **2017**).

$$n_i(r) \approx \rho_i r^d$$

$$\mu_i = \frac{r_{i,2}}{r_{i,1}} \qquad P(\mu|\rho) = P(\mu) = \frac{d}{\mu^{1+d}}$$

THE ESTIMATE OF d IS DECOUPLED FROM THE ESTIMATE OF ρ

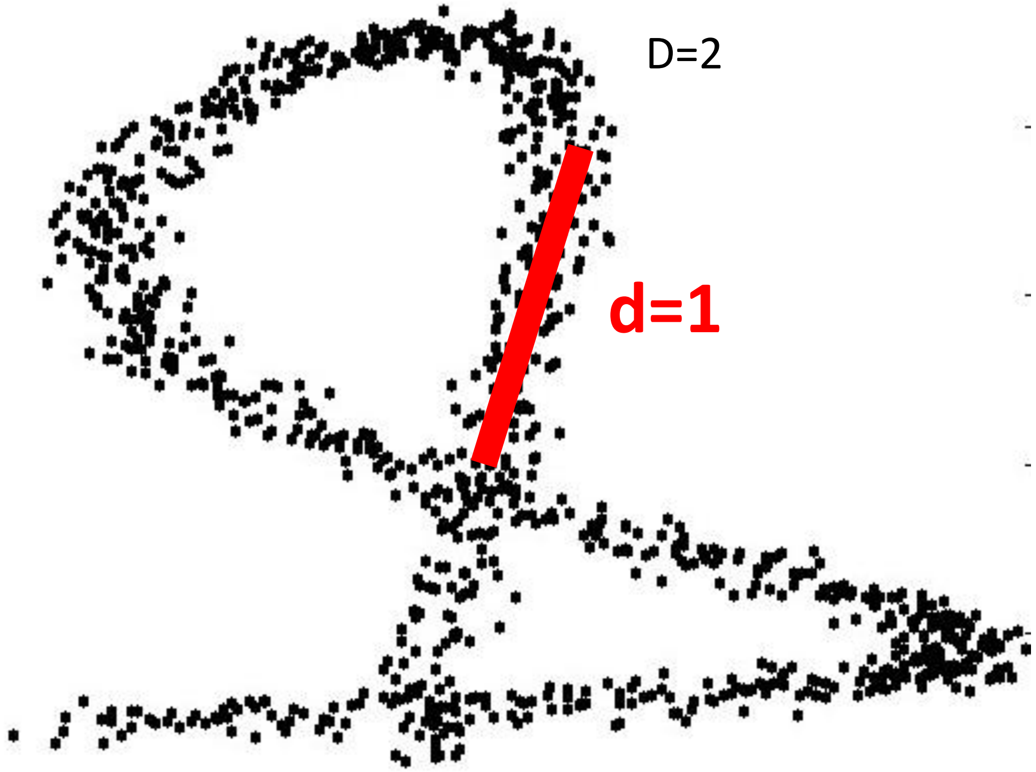
[1] PRL, 50, 346 (1983)
[2] *Proc. Machine Vision Conf.*, 27.1–27.10 (2003)
[3] Sci Rep. 6, 31377 (2016)
[4] Math. Prob. In Eng. Art. 759567 (2015)
[5] Patt. Recog. 42, 780 (2009)

The intrinsic dimension:
a matter of scale

Example: a sample of configurations in a MD run of a biomolecule with N atoms. No constraints on the bonds.

An exact estimator should give **d=3N** if the sample is large enough

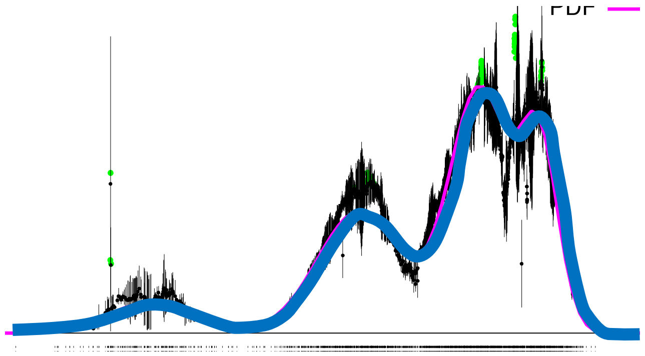
This estimate is irrelevant!
A good estimate of the ID should provide the number of directions in which the system can move significantly



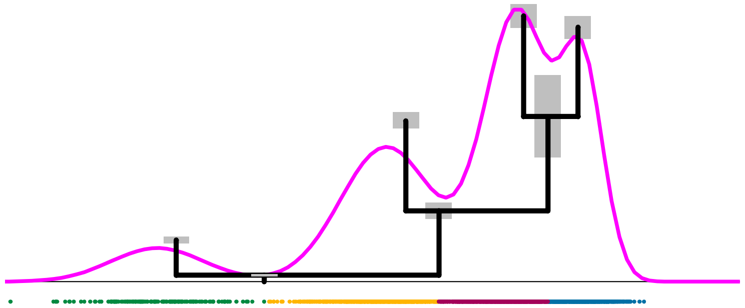
“soft” constraints: the system *practically* can move in only a few directions

Building a topography of a data landscape

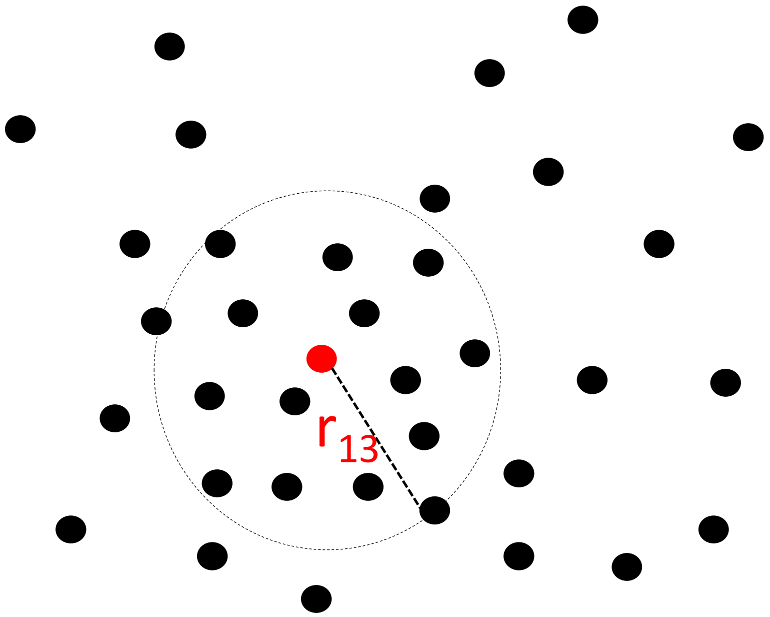
- We first compute the dimension of the manifold containing the data [Sci Rep. 12140, vol 7 (2017)]
  - We estimate the probability density at each data point. [JCTC in press (2018)]



- We then find the probability maxima by Density Peak clustering [Science, 1492, vol 322 (2014)]
  - We compute the probability at the boundary between each pair of maxima.
- Graphical representation of the topography



The k-nearest neighbor estimator



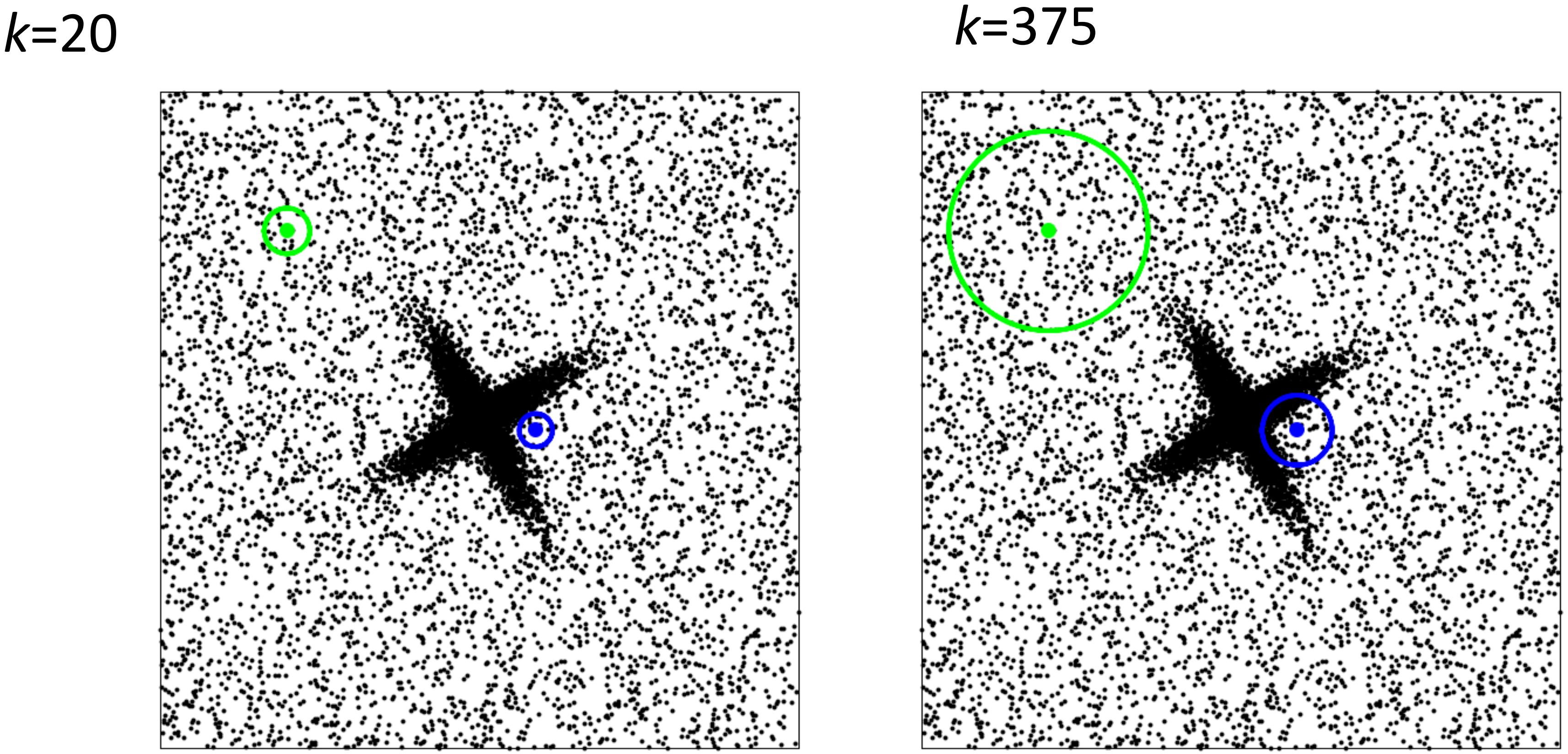
k=13

density: $\rho \approx 13/\pi r_{13}{}^{2}$

error$\approx \frac{r}{\sqrt{13}}$

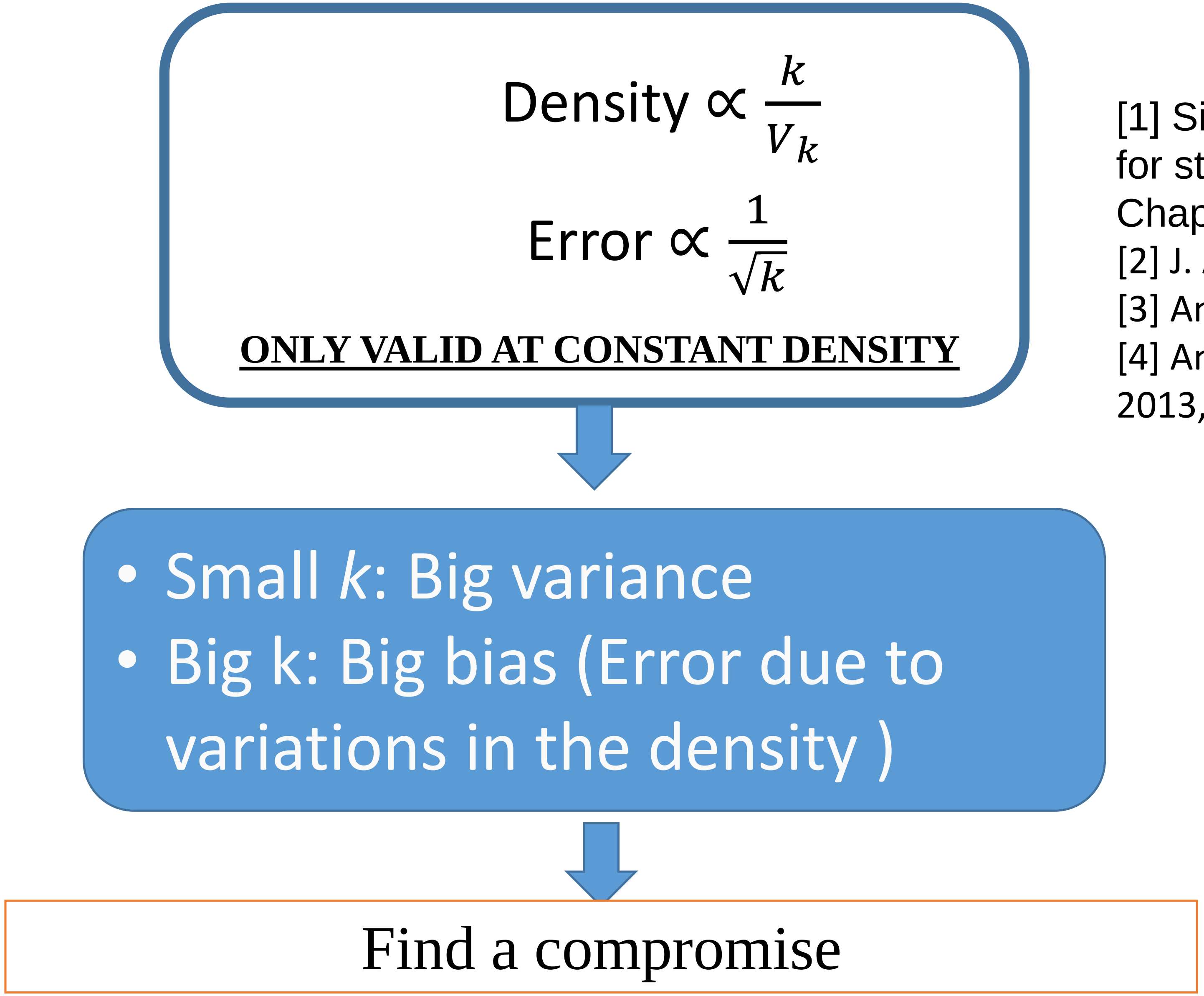
The key problem: highly non-uniform densities

density: $\rho = k/\pi r_k^2$



We need a different *k* for each point

Adaptive density estimate



$$\text{Density} \propto \frac{k}{V_k}$$

$$\text{Error} \propto \frac{1}{\sqrt{k}}$$

<u>ONLY VALID AT CONSTANT DENSITY</u>

- Small *k*: Big variance
- Big k: Big bias (Error due to variations in the density)

Find a compromise

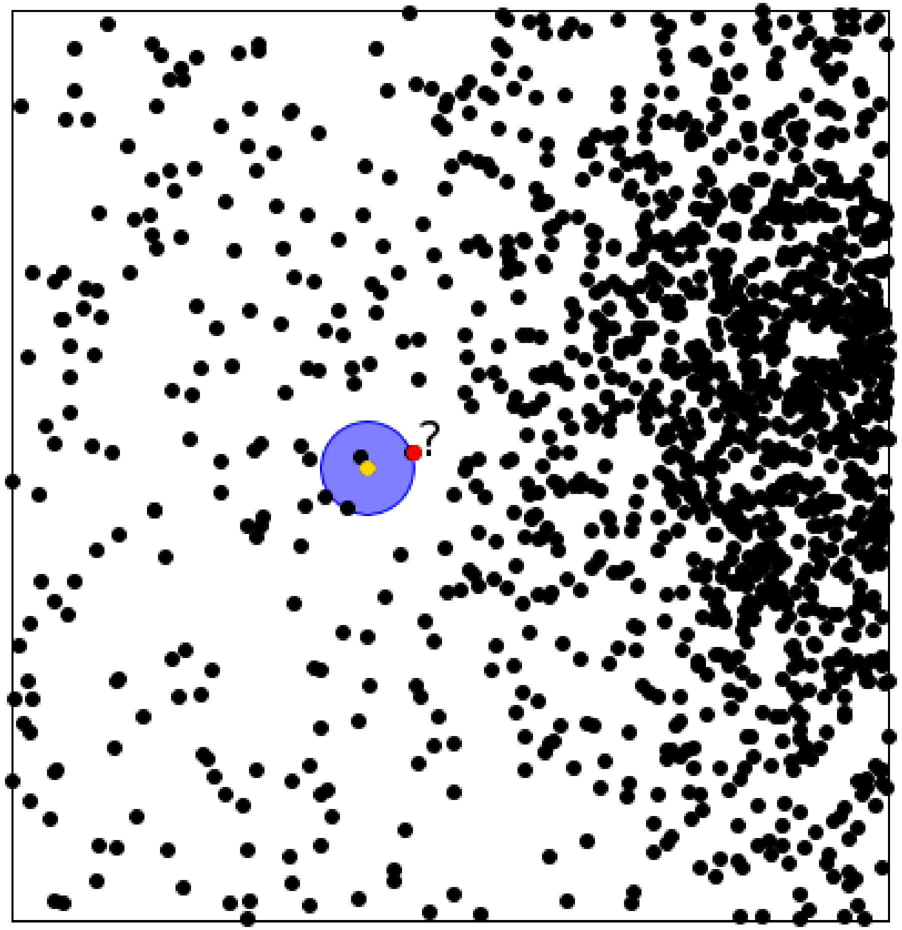
[1] Silverman, B. W. Density estimation for statistics and data analysis; Chapman and Hall, 1986

[2] J. Am. Stat. Assoc. 1996, 91, 401–407.

[3] Ann. Statist. 1997, 25, 929–947.

[4] Ann. Inst. H. Poincar Probab. Statist. 2013, 49, 900–914.

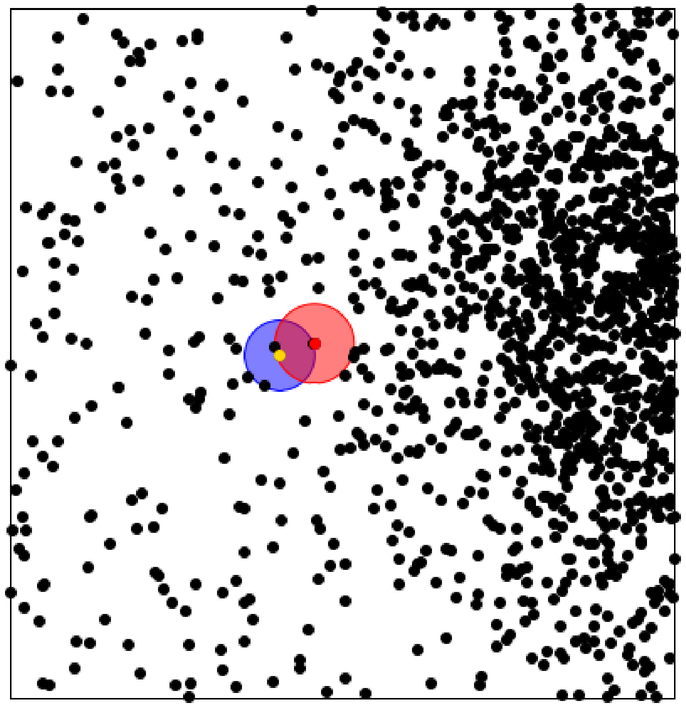
Obtaining a position dependent *k*



Starting with a very small *k*:
Should we include the next neighbor in the density estimate?

Two different hypothesis

The two points have the different densities

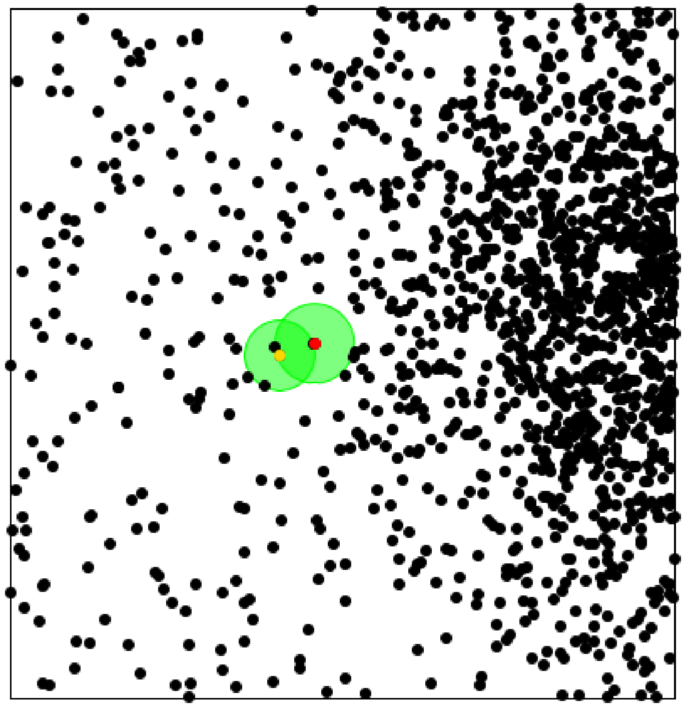


$$\mathcal{L}_{M1} = \max_{\rho,\rho'} \mathcal{L}_{i,k}(\rho) + \mathcal{L}_{j,k}(\rho') =$$

$$k \cdot \log \frac{k^2}{V_{i,k}V_{j,k}} - 2 \cdot k.$$

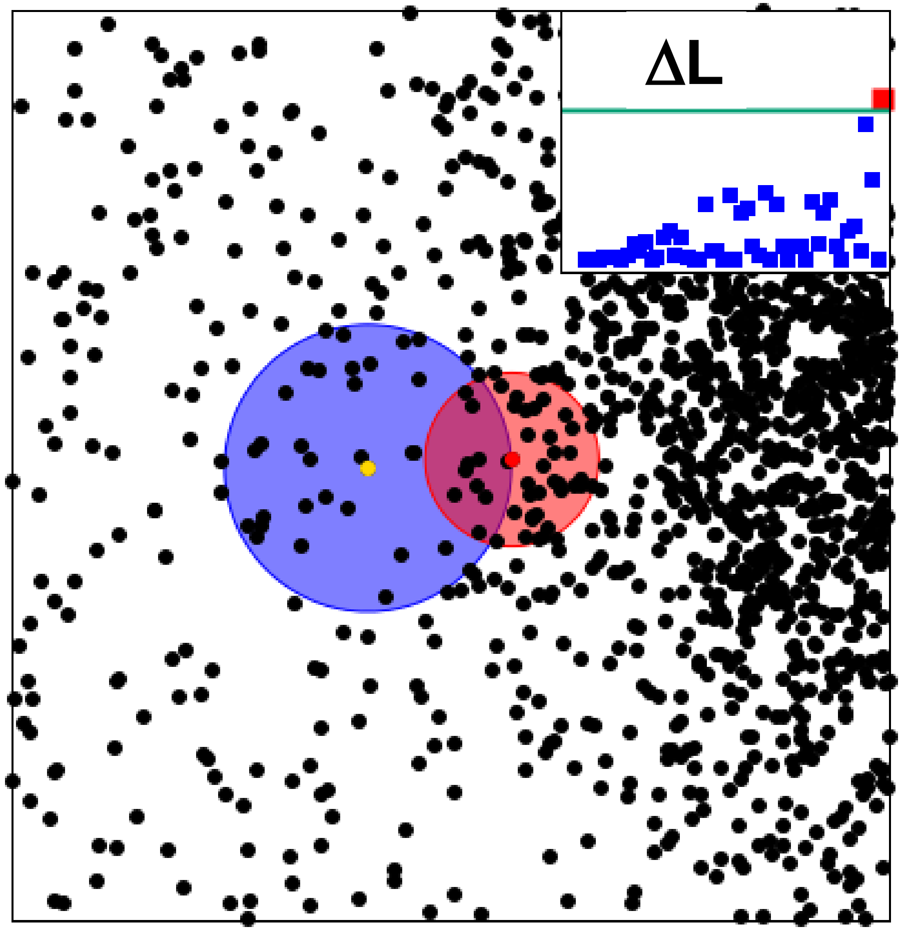
$$V_{ik} = \omega_d r_{ik}^d$$

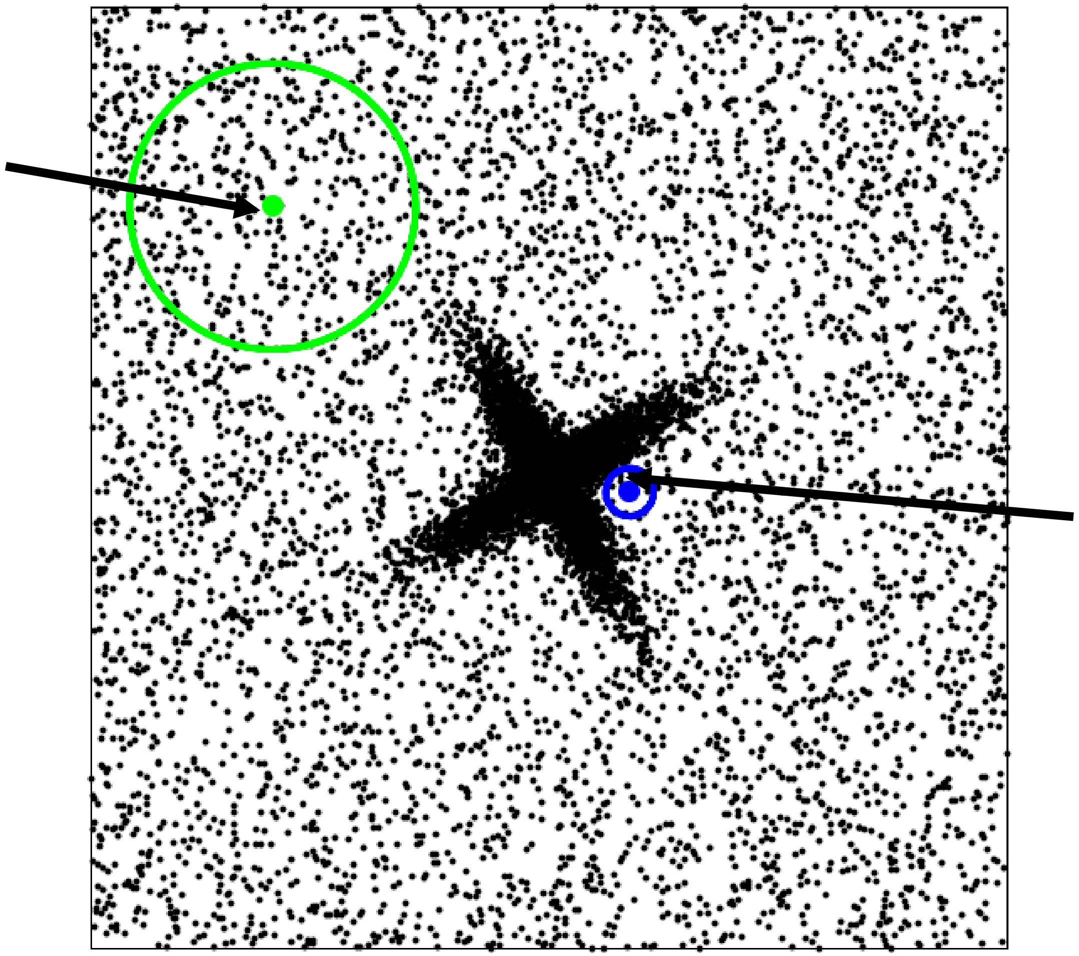
The two points have the same density



$$\mathcal{L}_{M2} = \max_{\rho} \mathcal{L}_{i,k}(\rho) + \mathcal{L}_{j,k}(\rho) =$$

$$2 \cdot k \cdot \log \frac{2 \cdot k}{V_{i,k} + V_{j,k}} - 2 \cdot k.$$

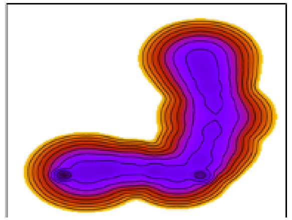

ΔL



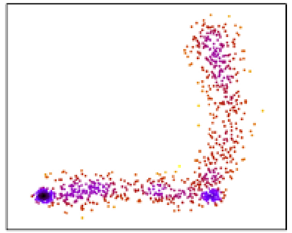
k=270

k=13

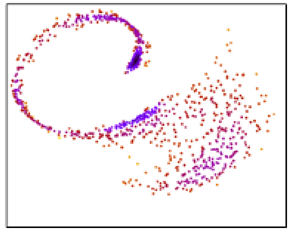
Benchmarks on realistic densities



Analytical P in d between 2 and 7

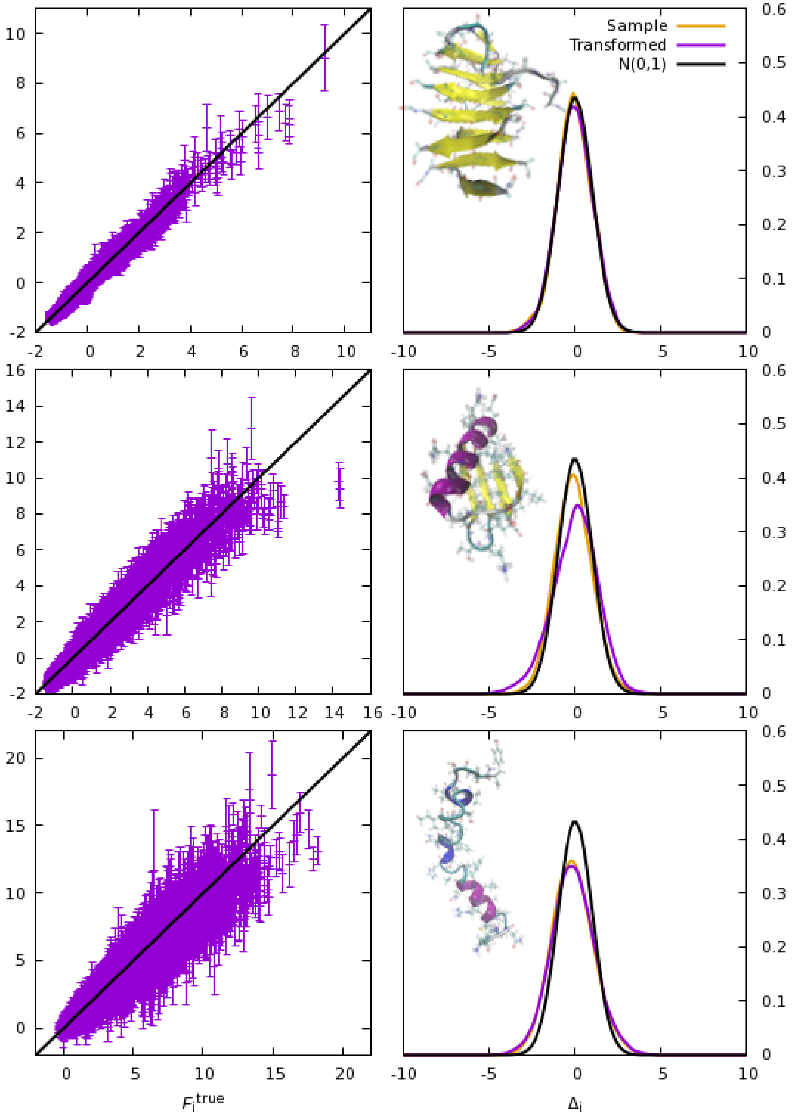


Sample 10000 points



Embed on a curved hypersurface

Embed on 20 dimensional space; rotate.



Amyloid-β
d=2

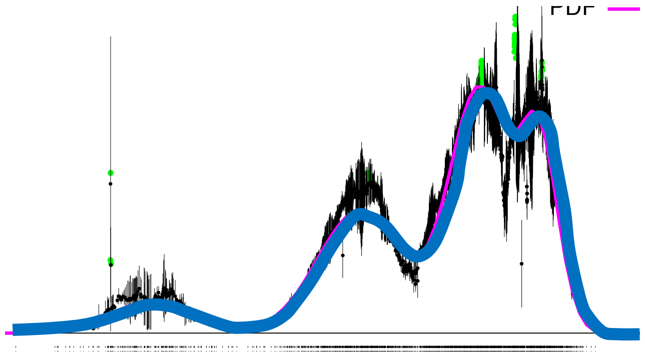
GB3
d=4

hIAPP
d=7

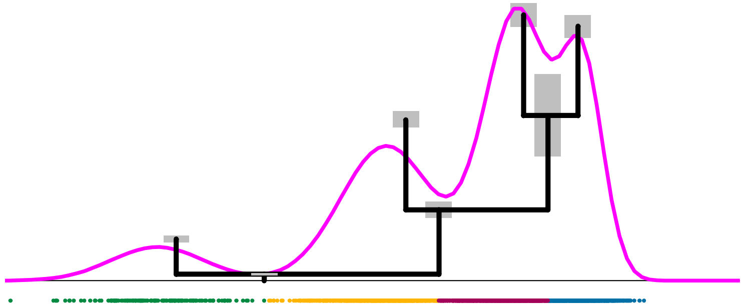
- We correctly estimate the density **on the manifold** containing the data
- We correctly predict the error

Building a topography of a data landscape

- We first compute the dimension of the manifold containing the data [Sci Rep. 12140, vol 7 (2017)]
    - We estimate the probability density at each data point. [JCTC in press (2018)]

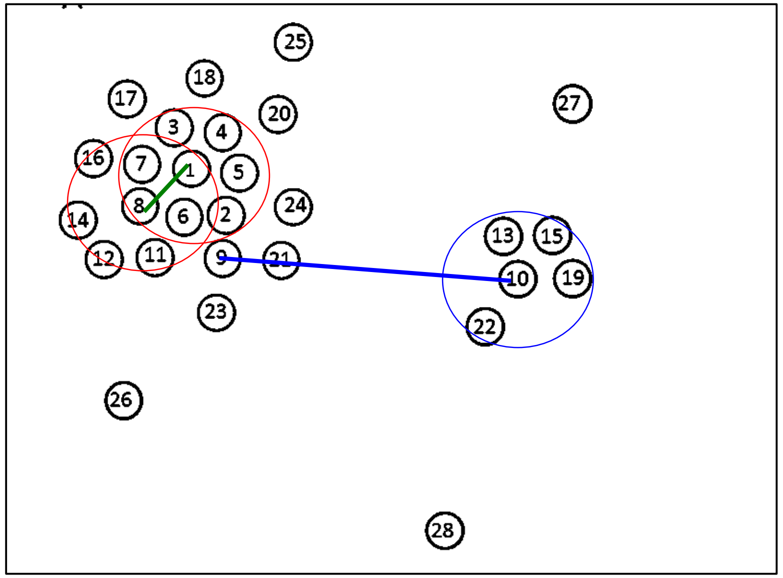


- We then find the probability maxima by Density Peak clustering [Science, 1492, vol 322 (2014)]
    - We compute the probability at the boundary between each pair of maxima.
    - Graphical representation of the topography



Finding the density peaks

The idea: the point at the top of a density peak is far from any other point with higher density



1) Compute the local density around each point

$\rho(1)=7$

$\rho(8)=5$

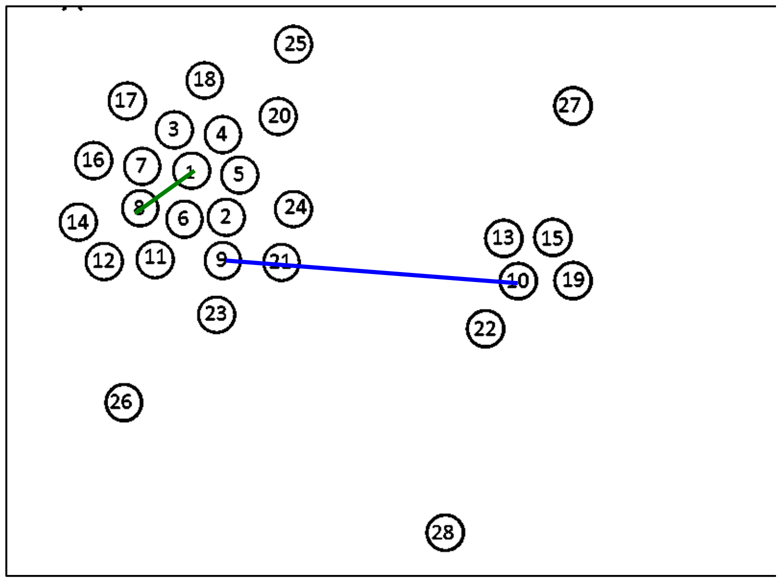
$\rho(10)=4$

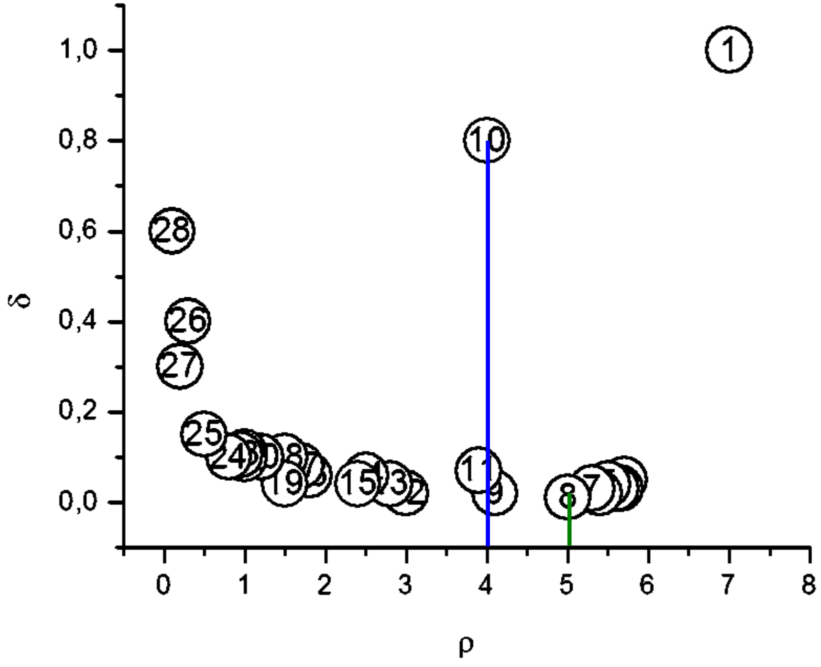
2) For each point compute the distance with all the points with higher density. Take the minimum value.

Finding the density peaks

The idea: the point at the top of a density peak is far from any other point with higher density

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

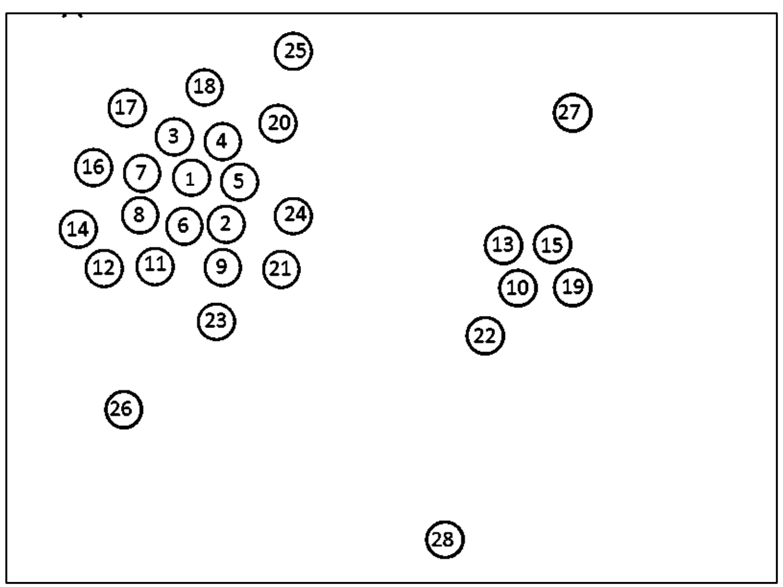


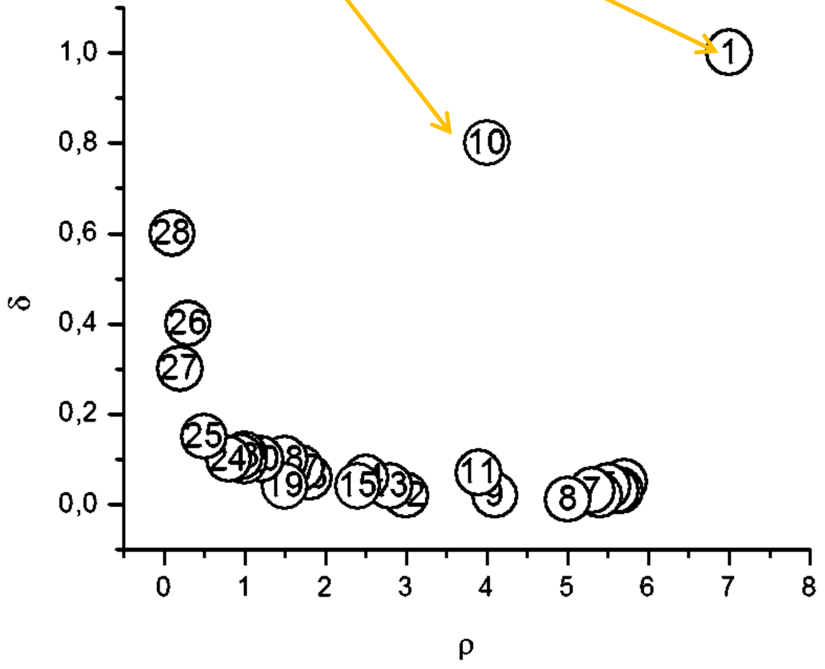


Finding the density peaks

The idea: the point at the top of a density peak is far from any other point with higher density

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



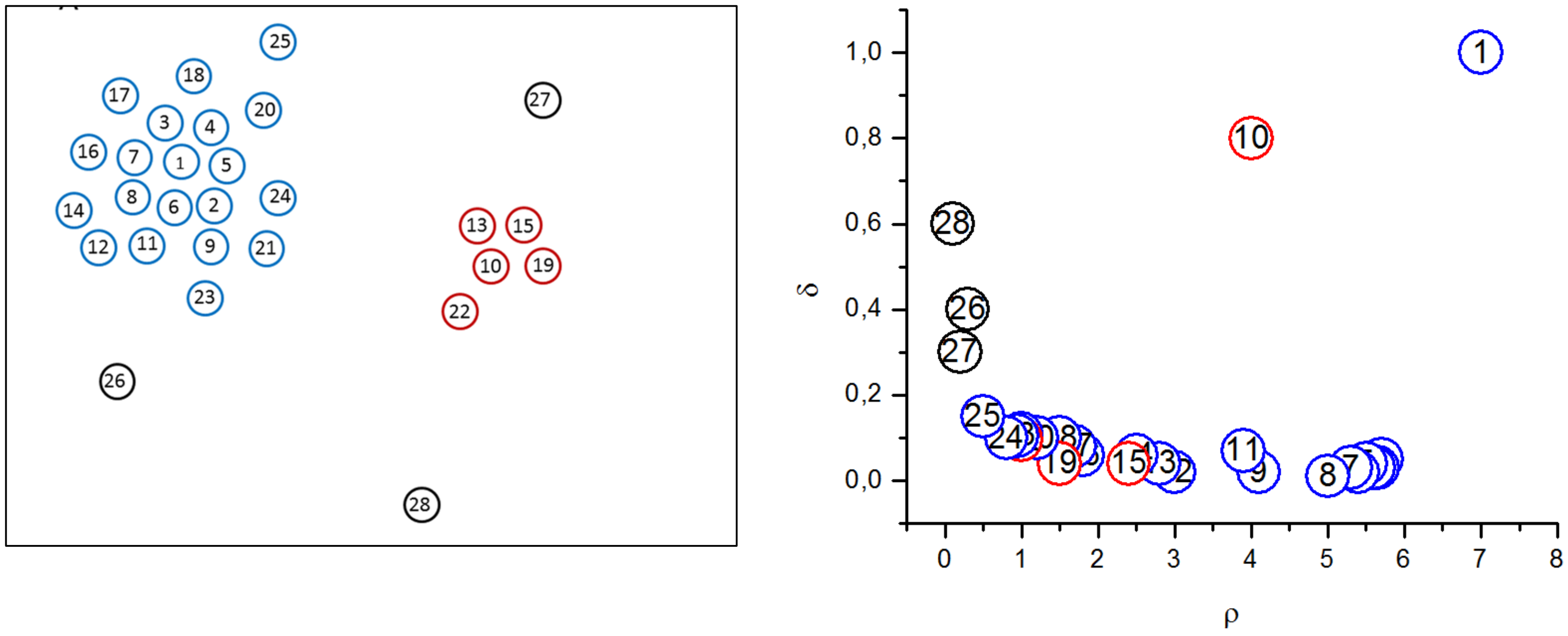


SCIENCE, 1492, vol 322 (2014)

Finding the 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



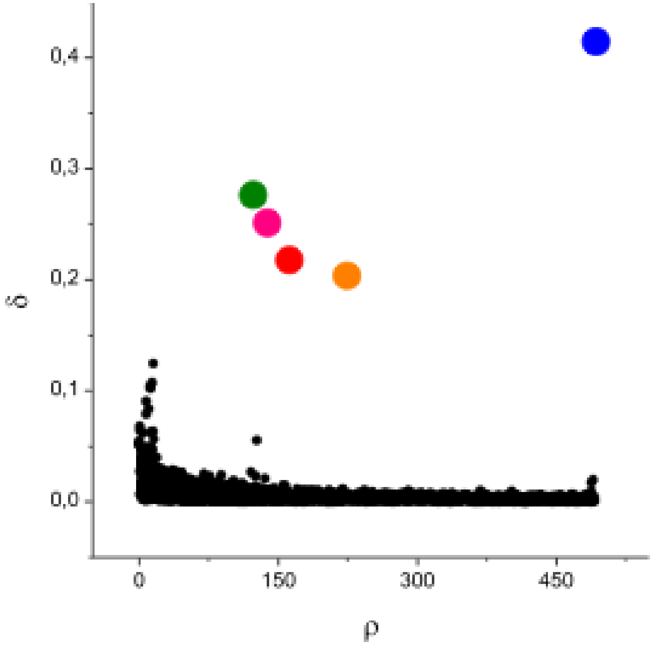
SCIENCE, 1492, vol 322 (2014)

No optimization required...

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

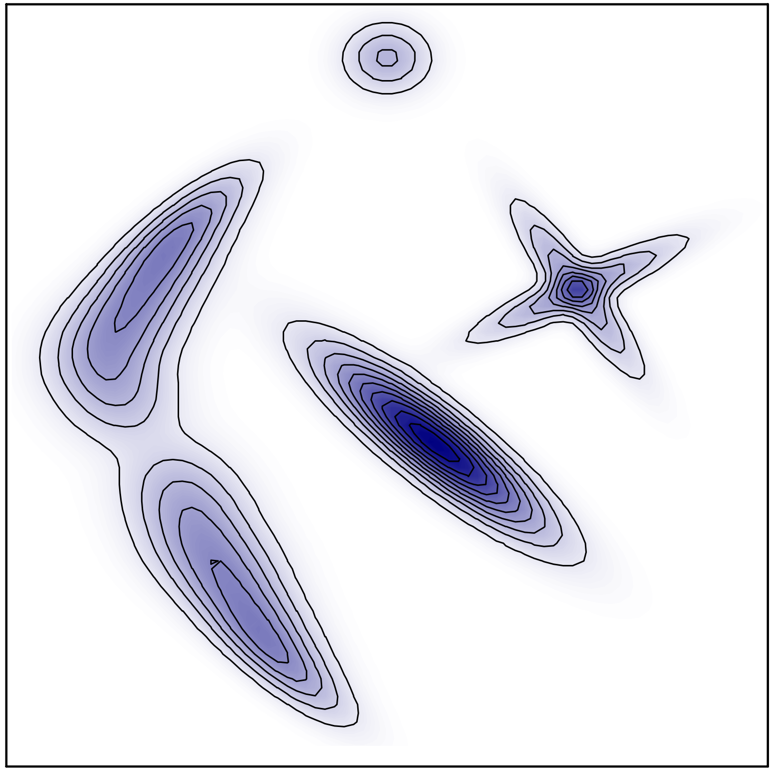
Cluster centers: the points whose δ_i is larger than the radius of the neighborhood used to estimate its density

Plot δ as a function of ρ

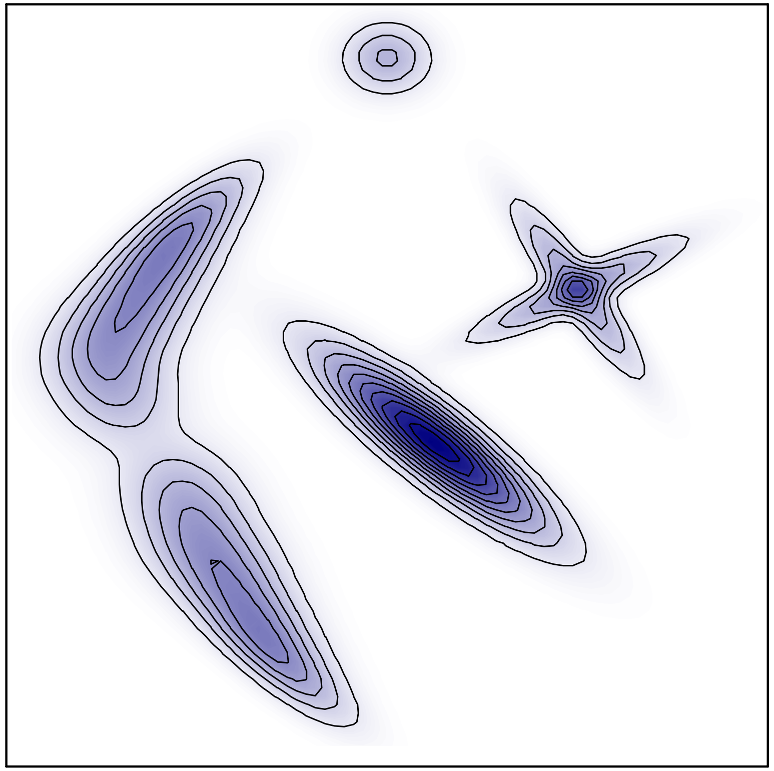


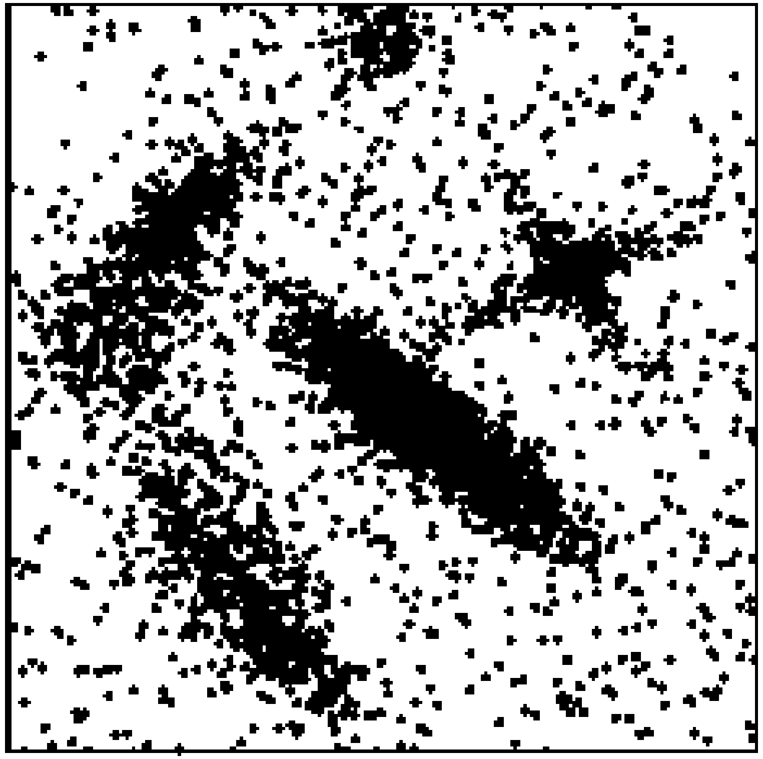
UNSUPERVISED: the density estimate is non-parametric. The number of clusters is determined automatically

The clustering approach at work

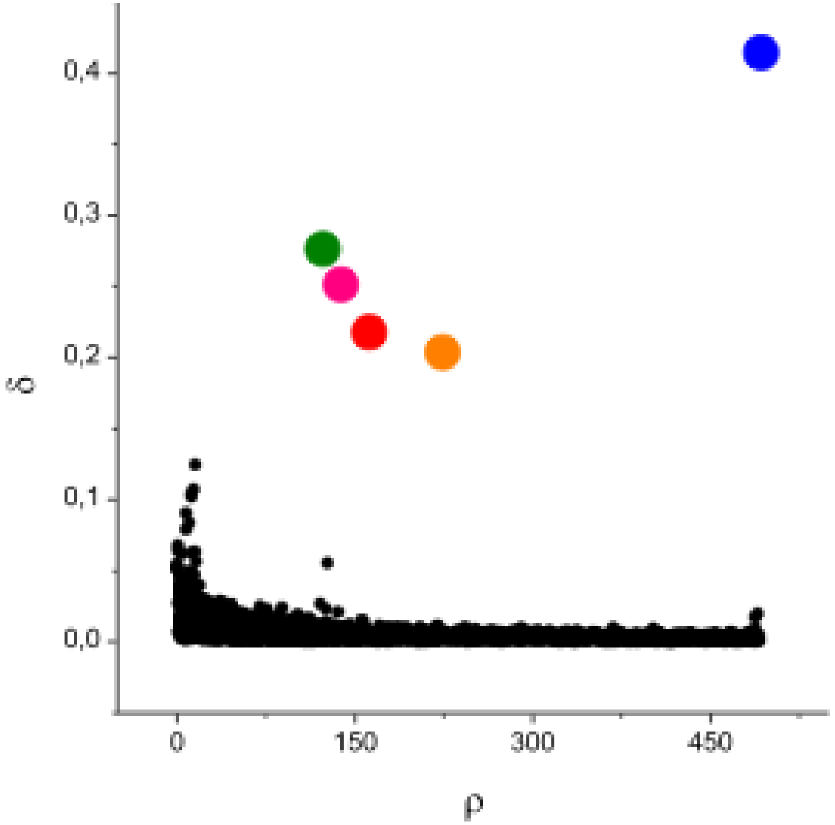


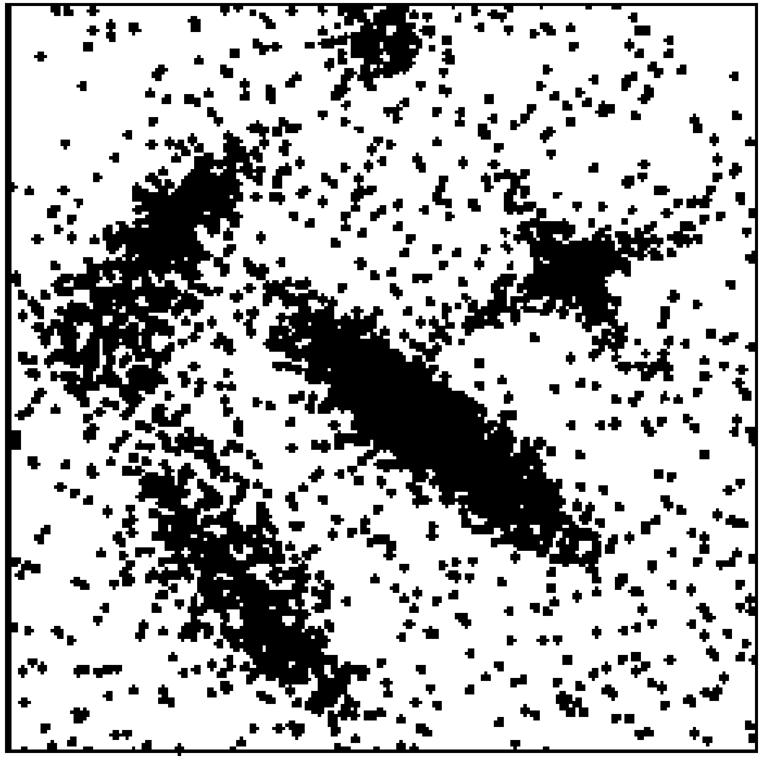
The clustering approach at work



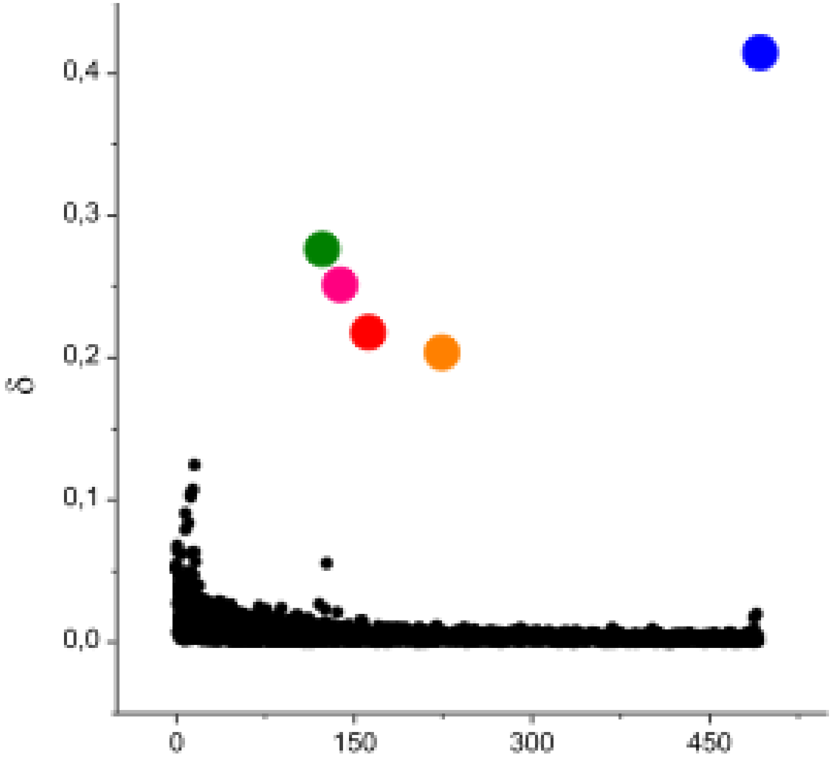


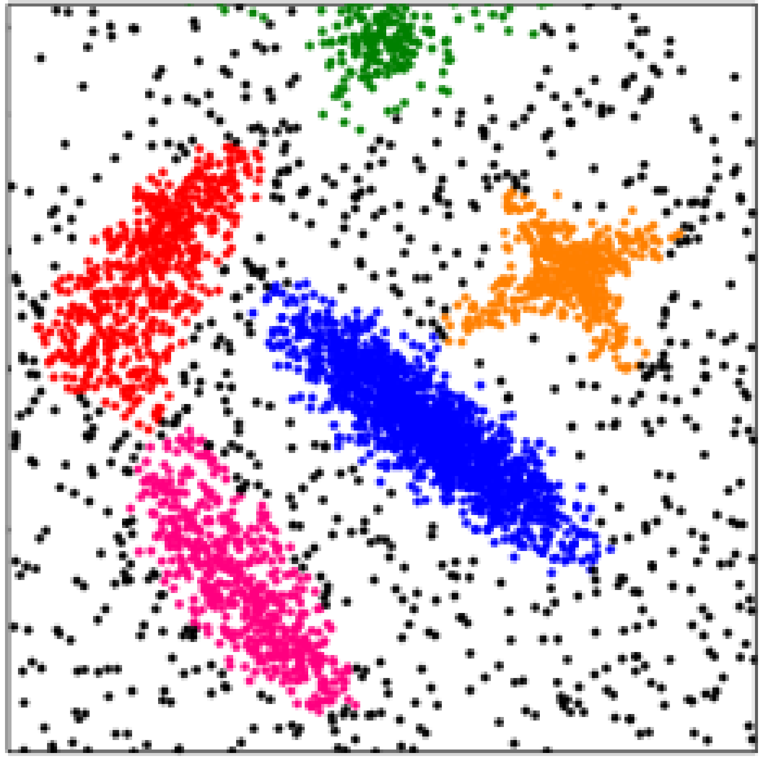
The clustering approach at work





The clustering approach at work

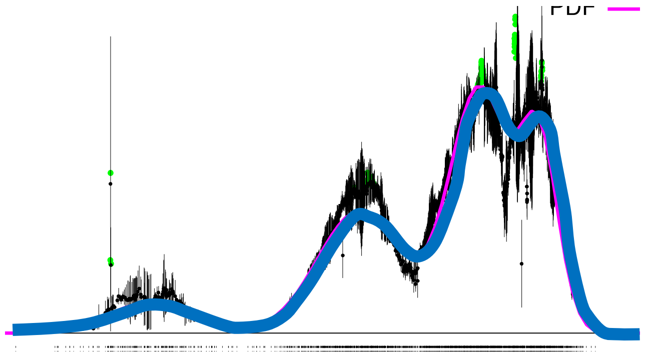




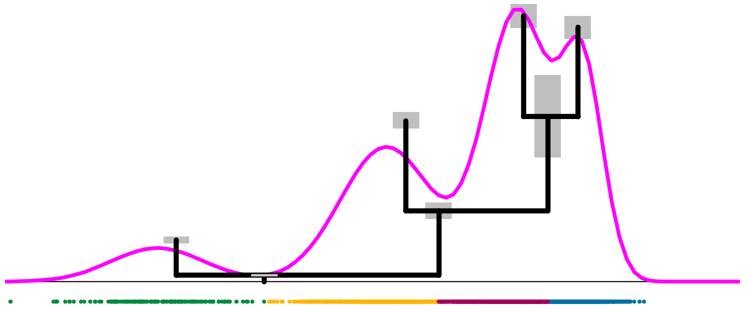
- The approach allows detecting non-spherical clusters
- It allows detecting clusters with different densities

Building a topography of a data landscape

- We first compute the dimension of the manifold containing the data [Sci Rep. 12140, vol 7 (2017)]
  - We estimate the probability density at each data point. [JCTC in press (2018)]

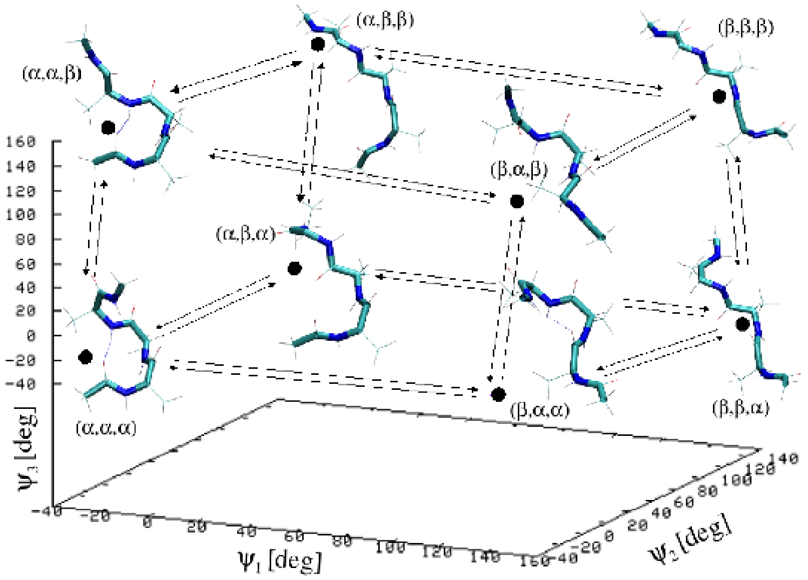


- We then find the probability maxima by Density Peak clustering [Science, 1492, vol 322 (2014)]
  - We compute the probability at the boundary between each pair of maxima.
- Graphical representation of the topography



Clustering a MD trajectory

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

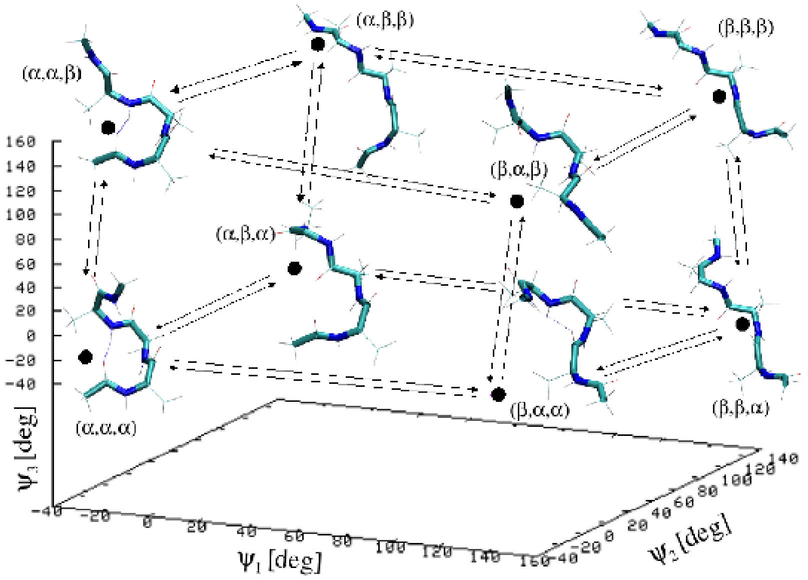


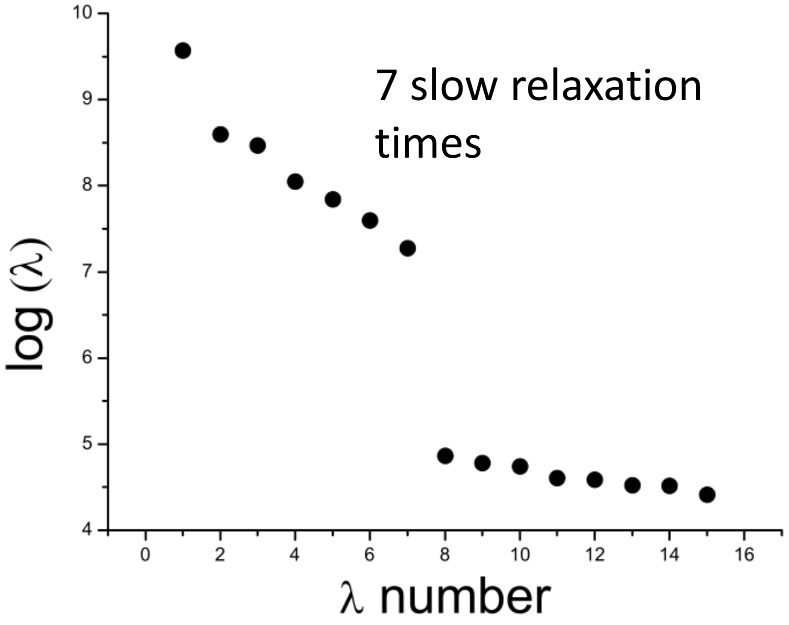
Building a MARCKOV STATE MODEL

- Find the microstates (set of very similar configurations). Typically 1000
- Compute the transition probability between the microstates at a time lag τ: P(a,τ|b)
- Diagonalize P. Its eigenvalues are the relaxation times of the system. The sign of eigenvector allow distinguishing the conformers

Clustering a MD trajectory

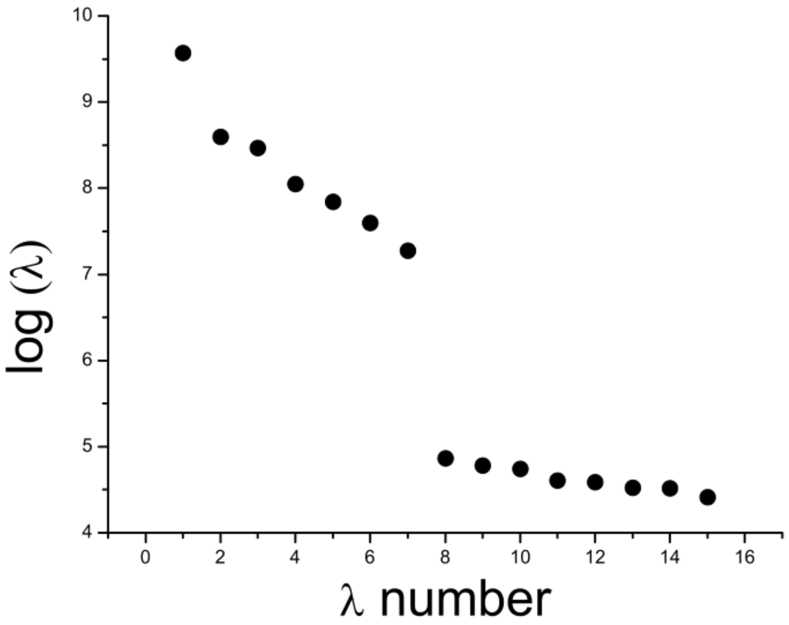
3000 ns of molecular dynamics of 3-Ala in water solution, at 300 K



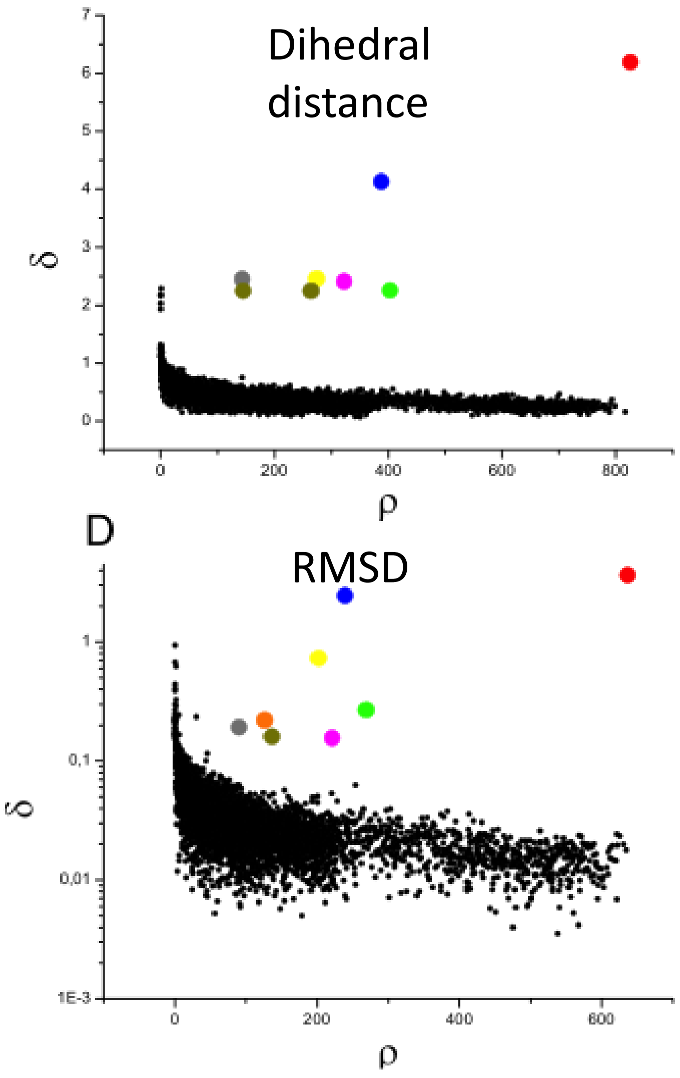


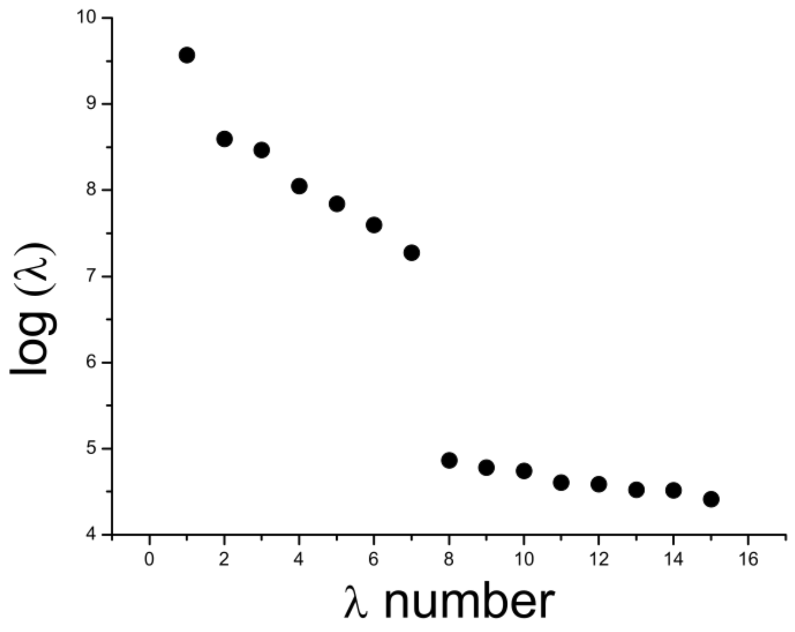
7 slow relaxation times

Clustering a MD trajectory

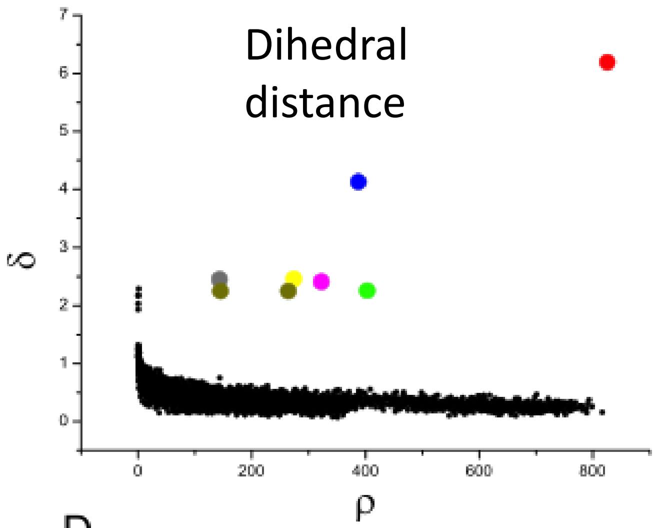


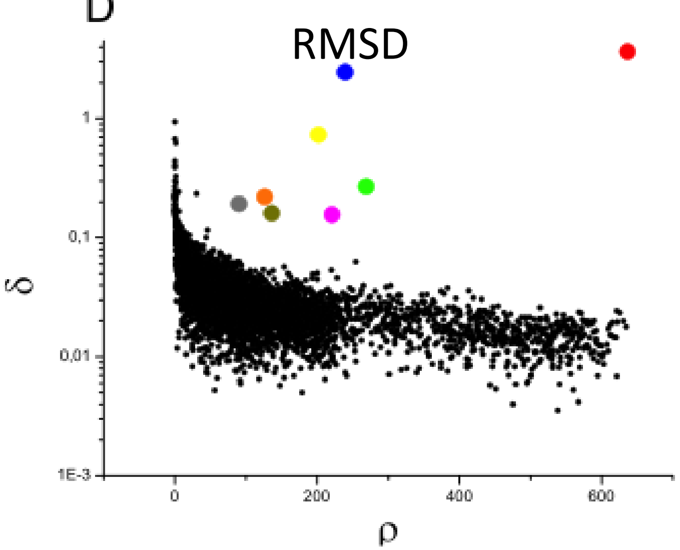
Clustering a MD trajectory

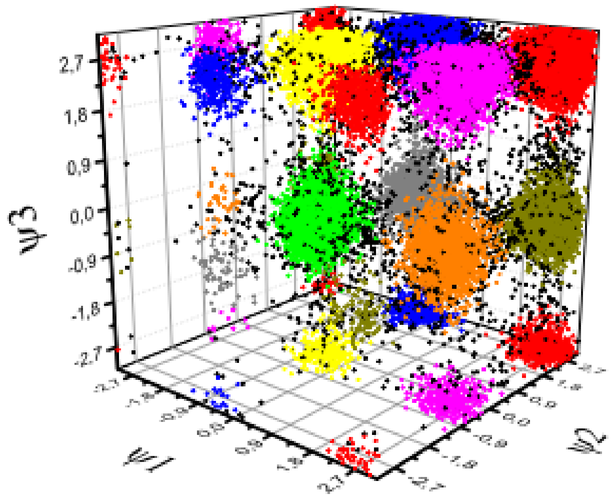


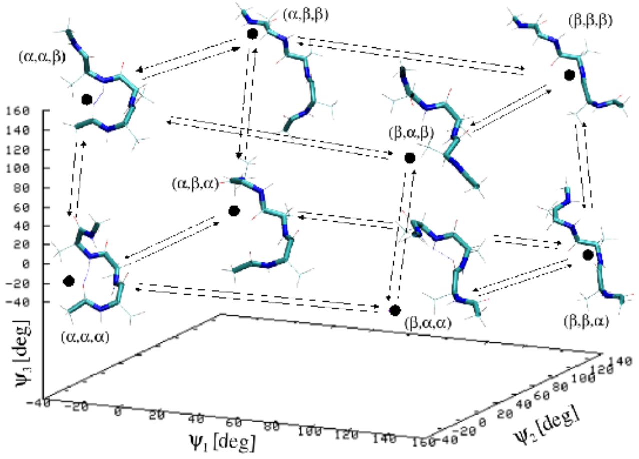


Clustering a MD trajectory



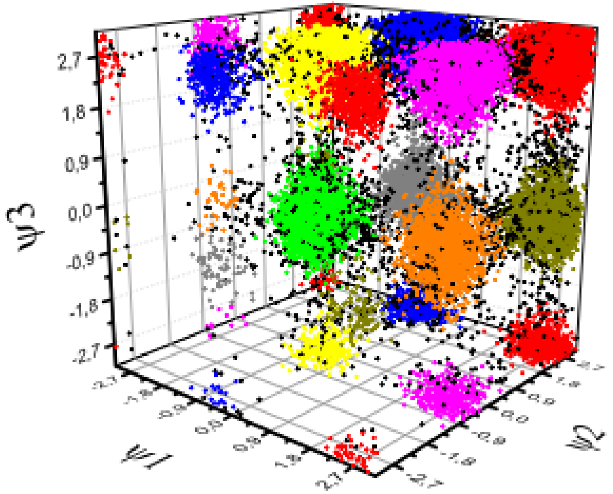


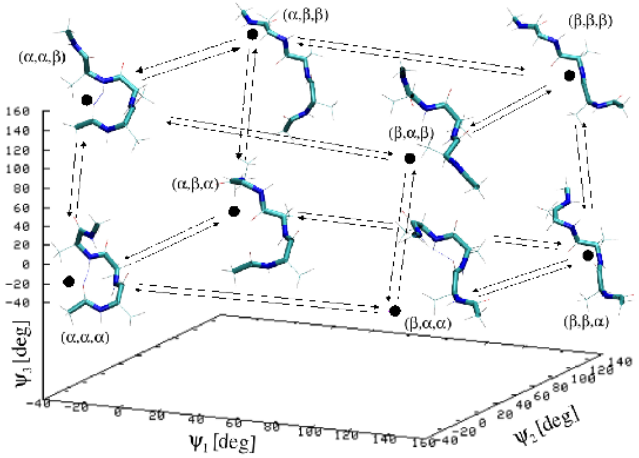




SCIENCE, 1492, vol 322 (2014)

Clustering a MD trajectory





Density-Peak clusters ≈ Inherent states of a Markov State Model

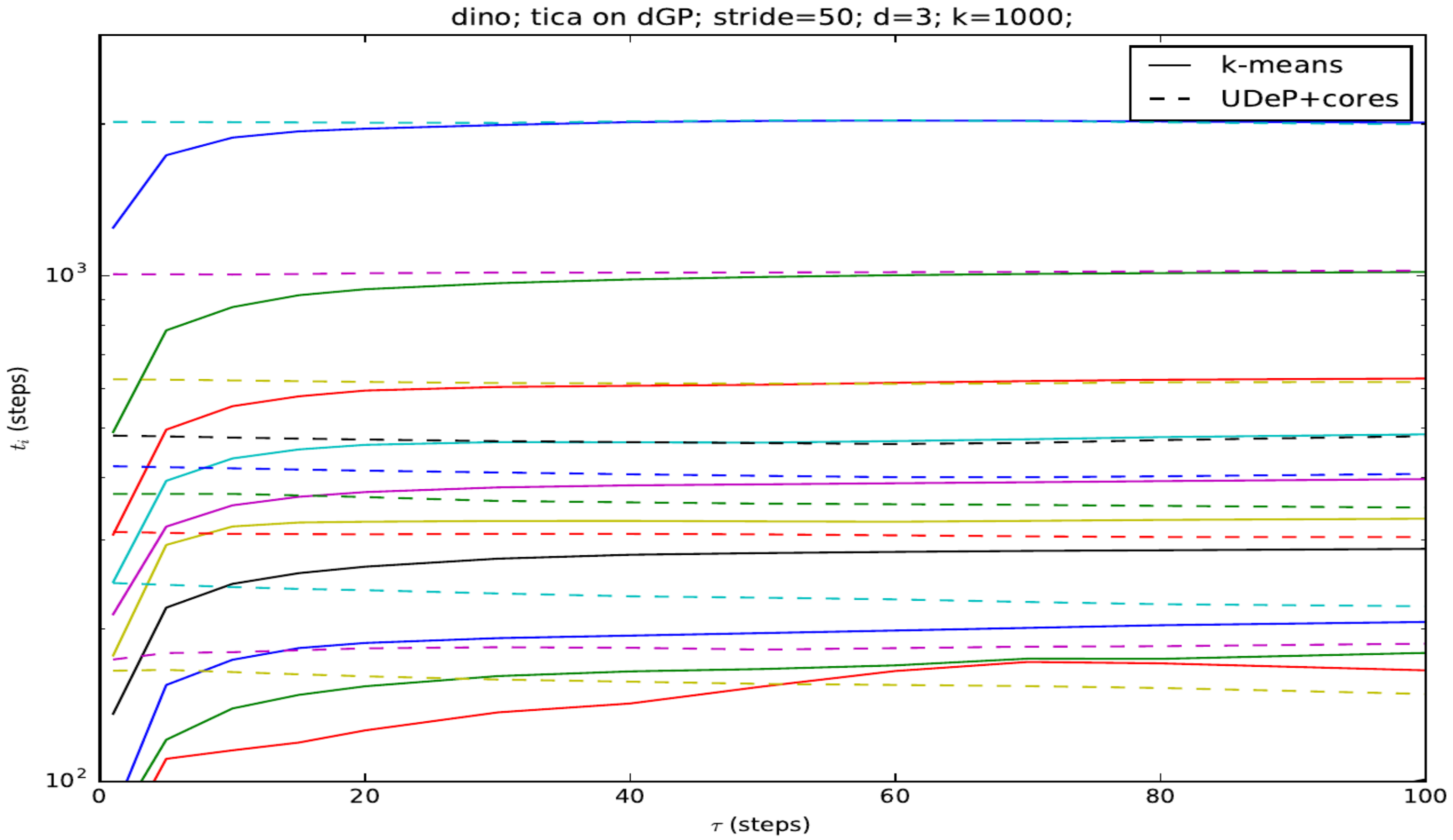
Density Peak clustering + MSM
(analysis from G. Pinamonti and G. Bussi)

Adenine dinucleotide

Coordinates: TICA-projected with kinetic map rescaling (F. Noe and C. Clementi). Core set approach.

Tot simulation time = 8 μs

Method	N. clusters
K-means	1000
Dens.Peak	18



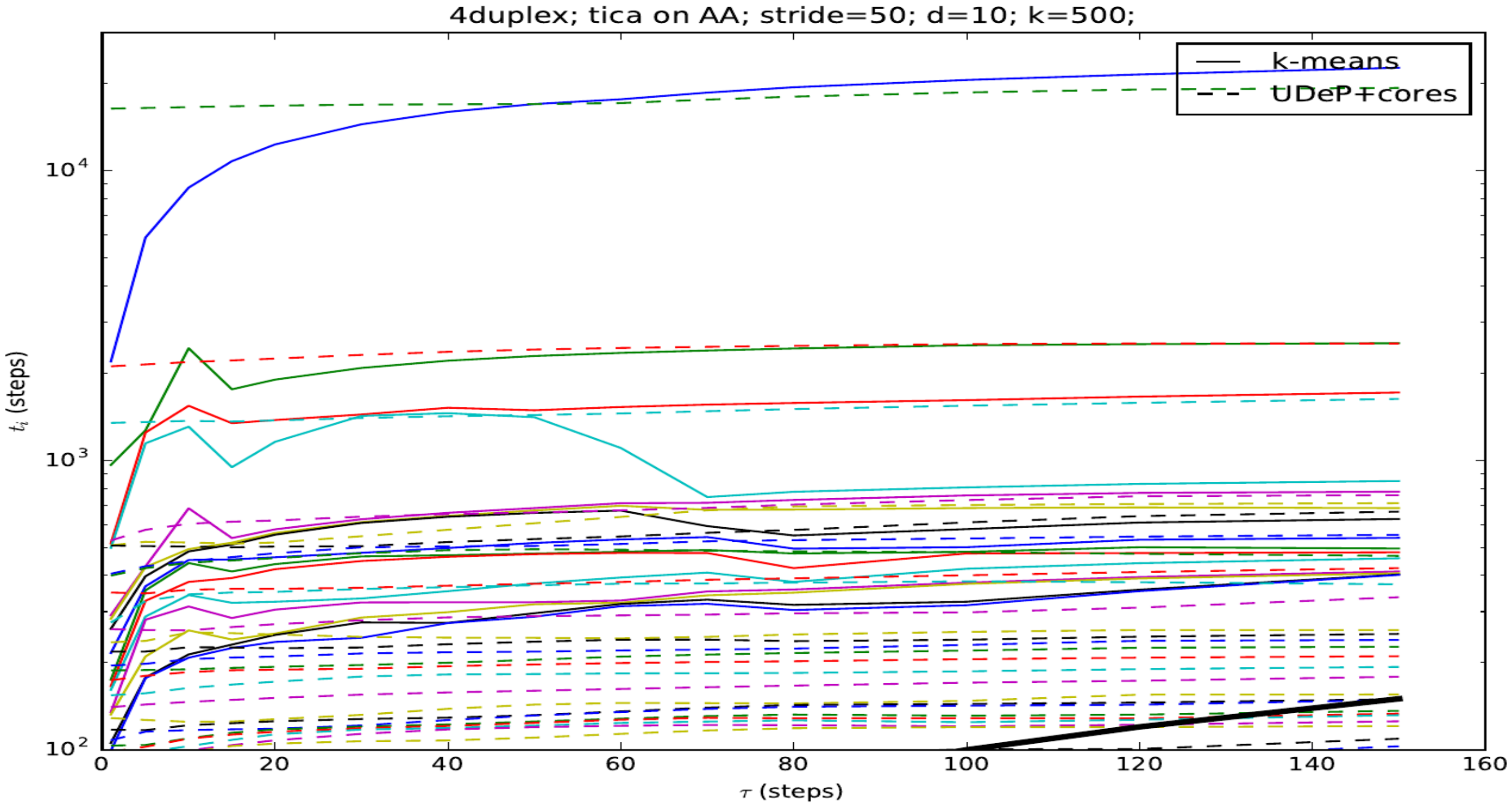
Density Peak clustering + MSM
(analysis from G. Pinamonti and G. Bussi)

4-nt duplex

Coordinates: TICA-projected with kinetic map rescaling (F. Noe and C. Clementi) Core set approach

Tot simulation time = 84 μs

Method	N. clusters
K-means	500
Dens.Peak	46



Density Peak clustering + MSM

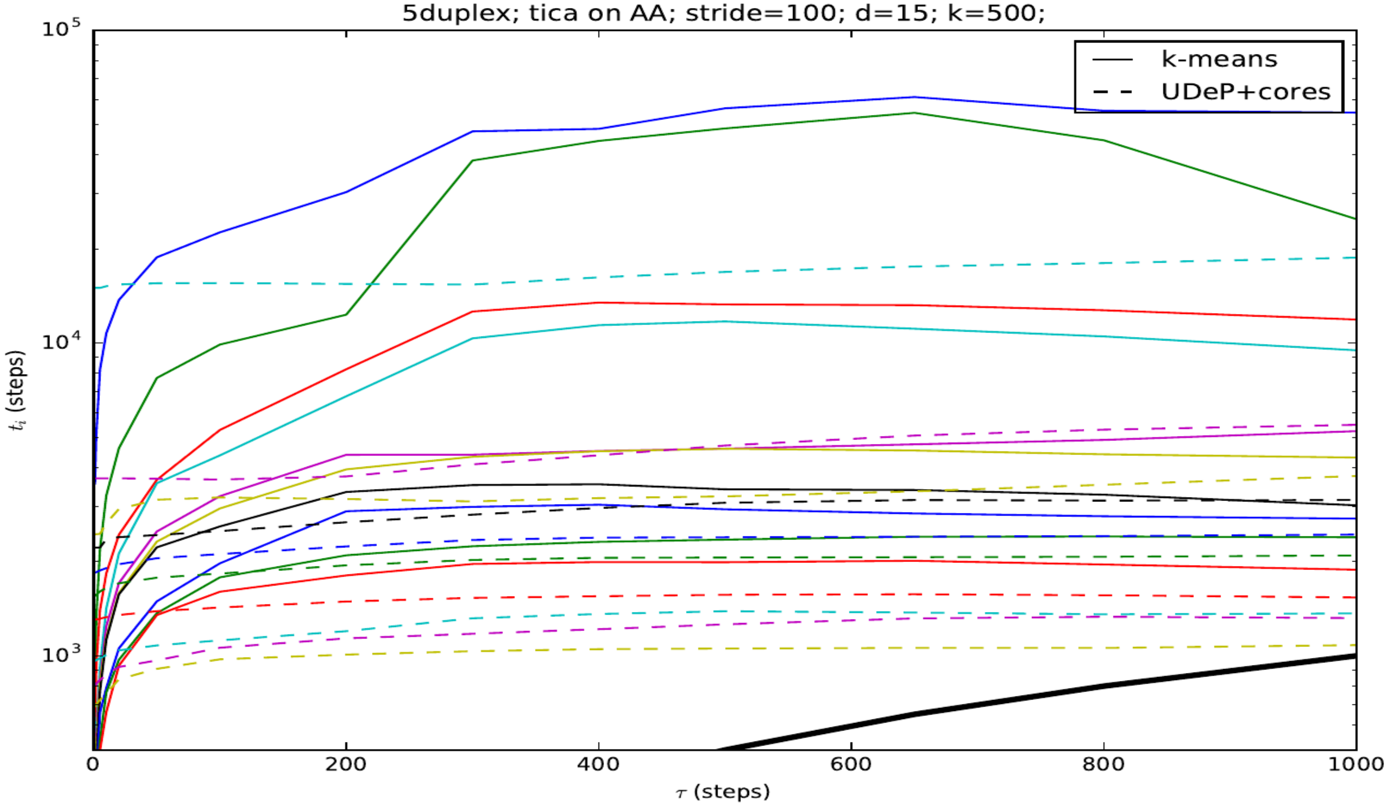
(analysis from G. Pinamonti and G. Bussi)

5-nt duplex

Coordinates: TICA-projected with kinetic map rescaling (F. Noe and C. Clementi). Core set approach.

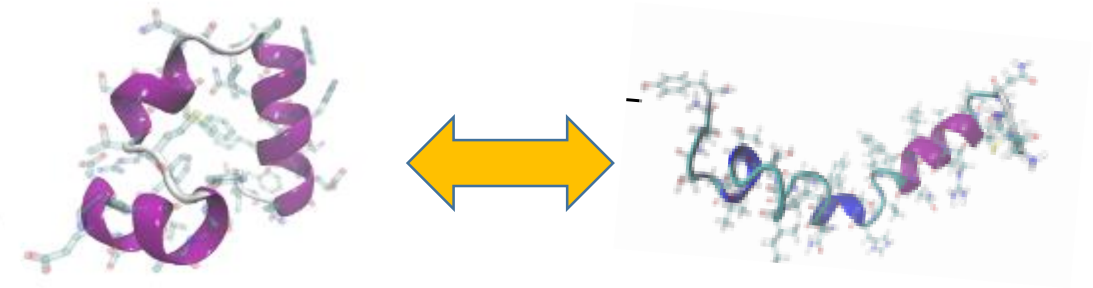
Tot simulation time = 134 μs

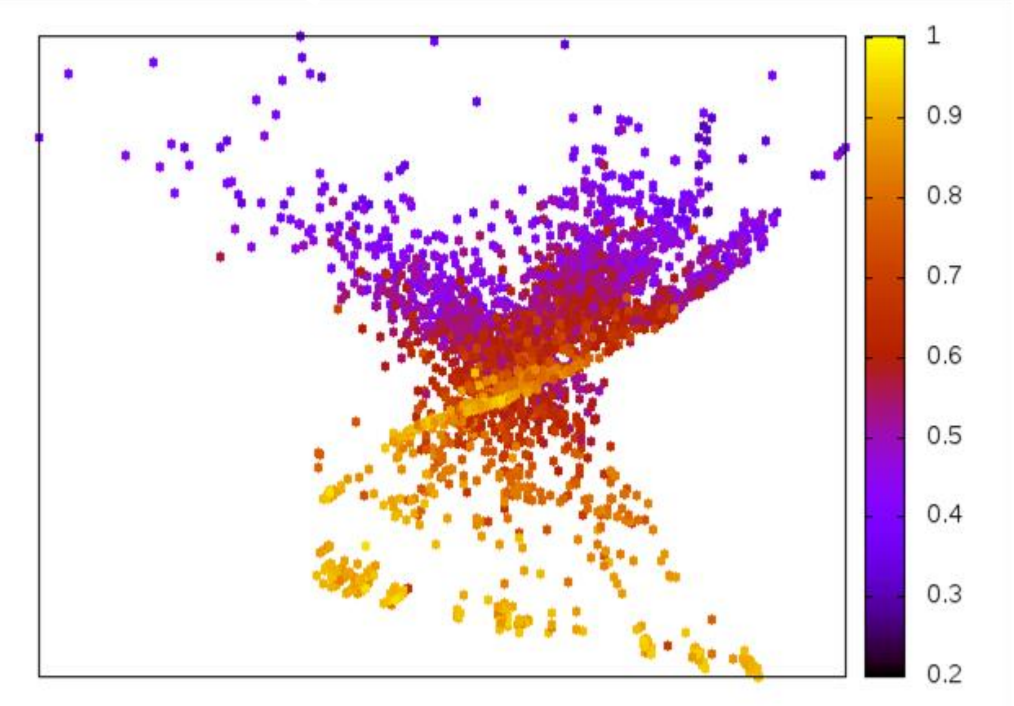
Method	N. clusters
K-means	500
Dens.Peak	39



Folding of a 32-residue protein (Villin headpiece)

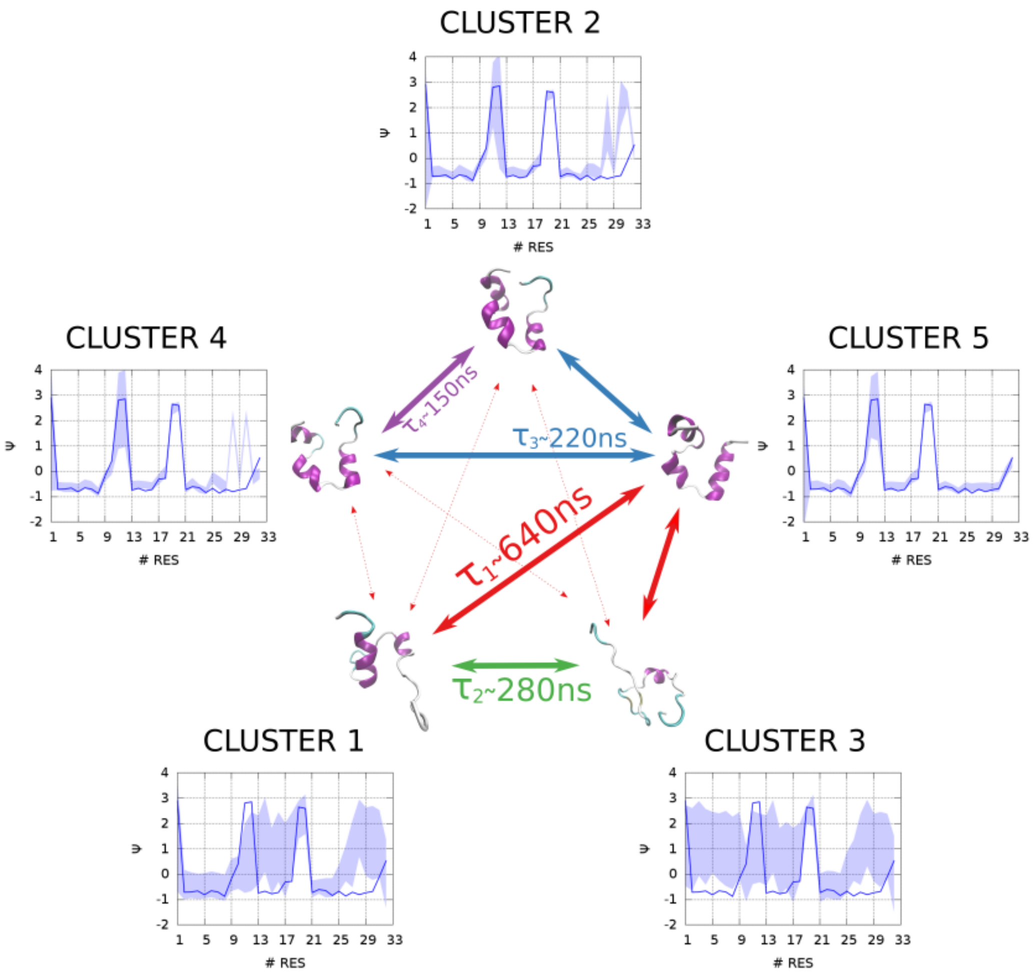
- 0.4 ms of molecular dynamics
- ~32000 configurations
- ~1000 atoms+ solvent
- Project to two dimensions by ISOMAP [Science 2000, 290, 2319–2323]





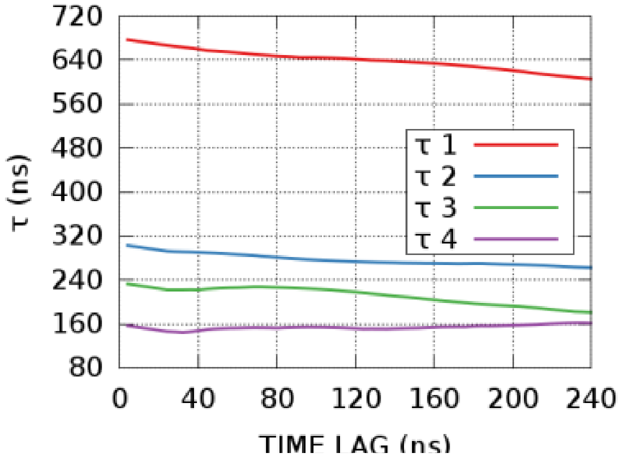
Folding of a 32-residue protein (Villin headpiece)

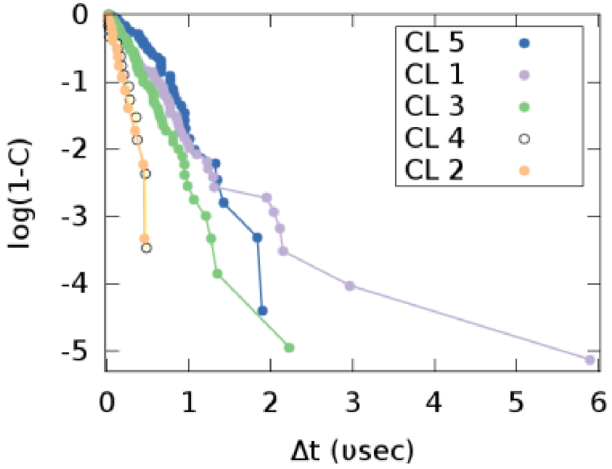
- 0.4 ms of molecular dynamics
- ~32000 configurations
- ~1000 atoms+ solvent
- Intrinsic dimension d~12
- ~5 statistically meaningful probability peaks (clusters)
- **The most populated cluster is the folded state**
- **Two unfolded states**



Folding of a 32-residue protein (Villin headpiece)

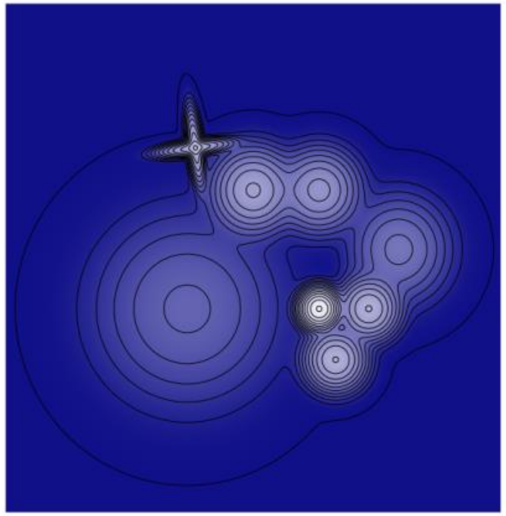
- 0.4 ms of molecular dynamics
- ~32000 configurations
- ~1000 atoms+ solvent
- Intrinsic dimension d~12
- ~5 statistically meaningful probability peaks (clusters)
- **Clean kinetics: exponential distribution of residence time.**

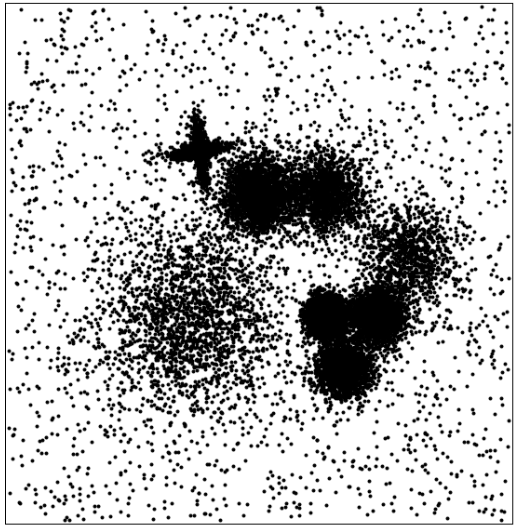


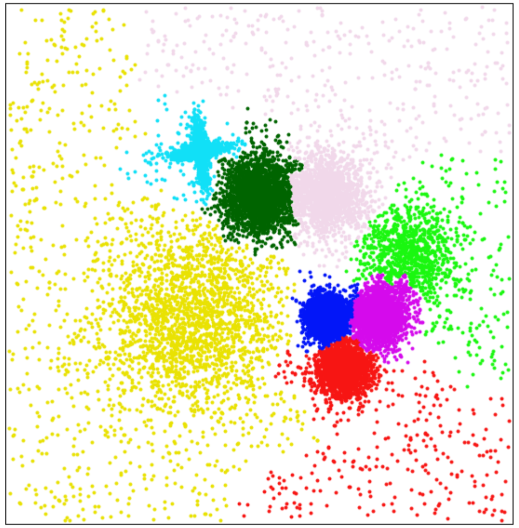


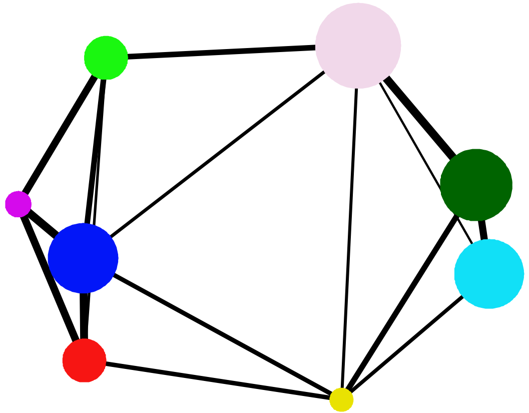
The topography of a data landscape

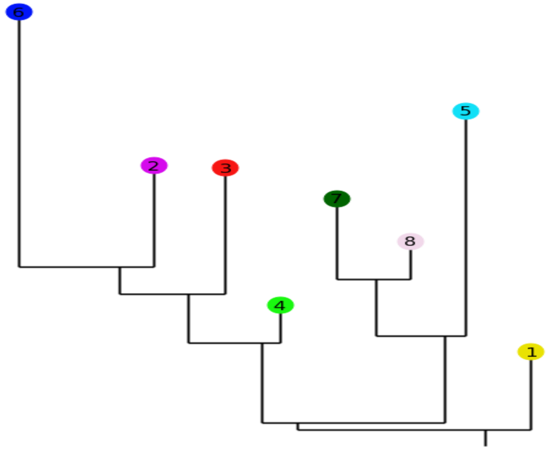
a list of properties of all the probability peaks, and of the saddle points connecting them





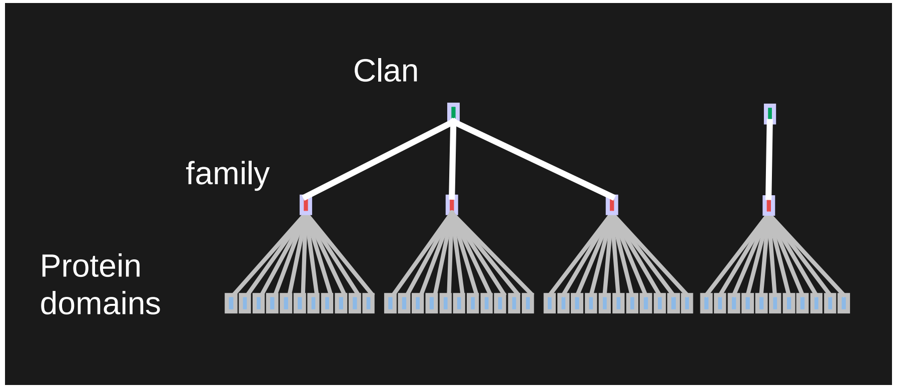






Automatic recognition of protein families

•Benchmark: **PFAM**. A widely used database of curated protein families, containing over 14800 families



Families that are supposed to share **the same evolutionary history** are grouped into **clans**

A family is defined by a profile hidden Markov model (HMM)

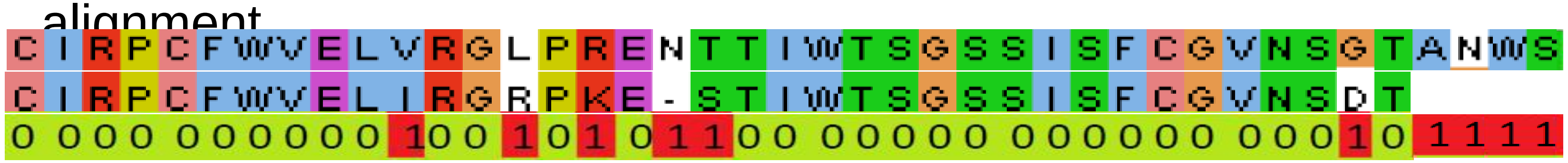
Profile HMMs are built from an aligned set of curator-defined family-representative sequences

A high-quality seed for alignment is essential.

A lot of handwork

Automatic recognition of protein families

Distance between two sequences: Hamming distance after pairwise alignment



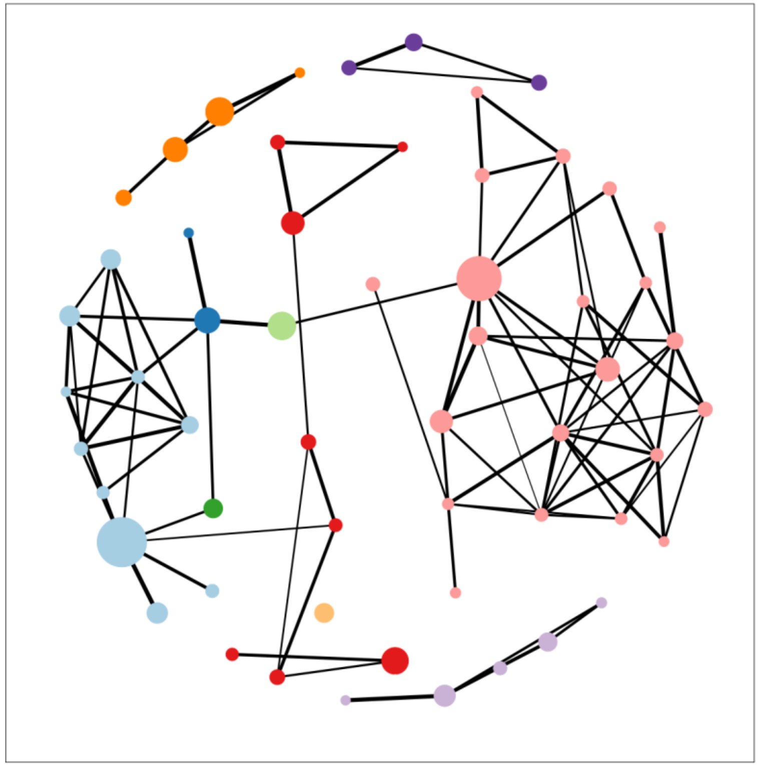
```
C I R P C F W V E L V R G L P R E N T T I W T S G S S I S F C G V N S G T A N W S
C I R P C F W V E L I R G R P K E - S T I W T S G S S I S F C G V N S D T
0 000 0000000 100 101 0 1100 00000 0000000 000 10 1111
```

- Triangular inequality satisfied

We analyze the PUA clan (~20000 sequences, 8 families)

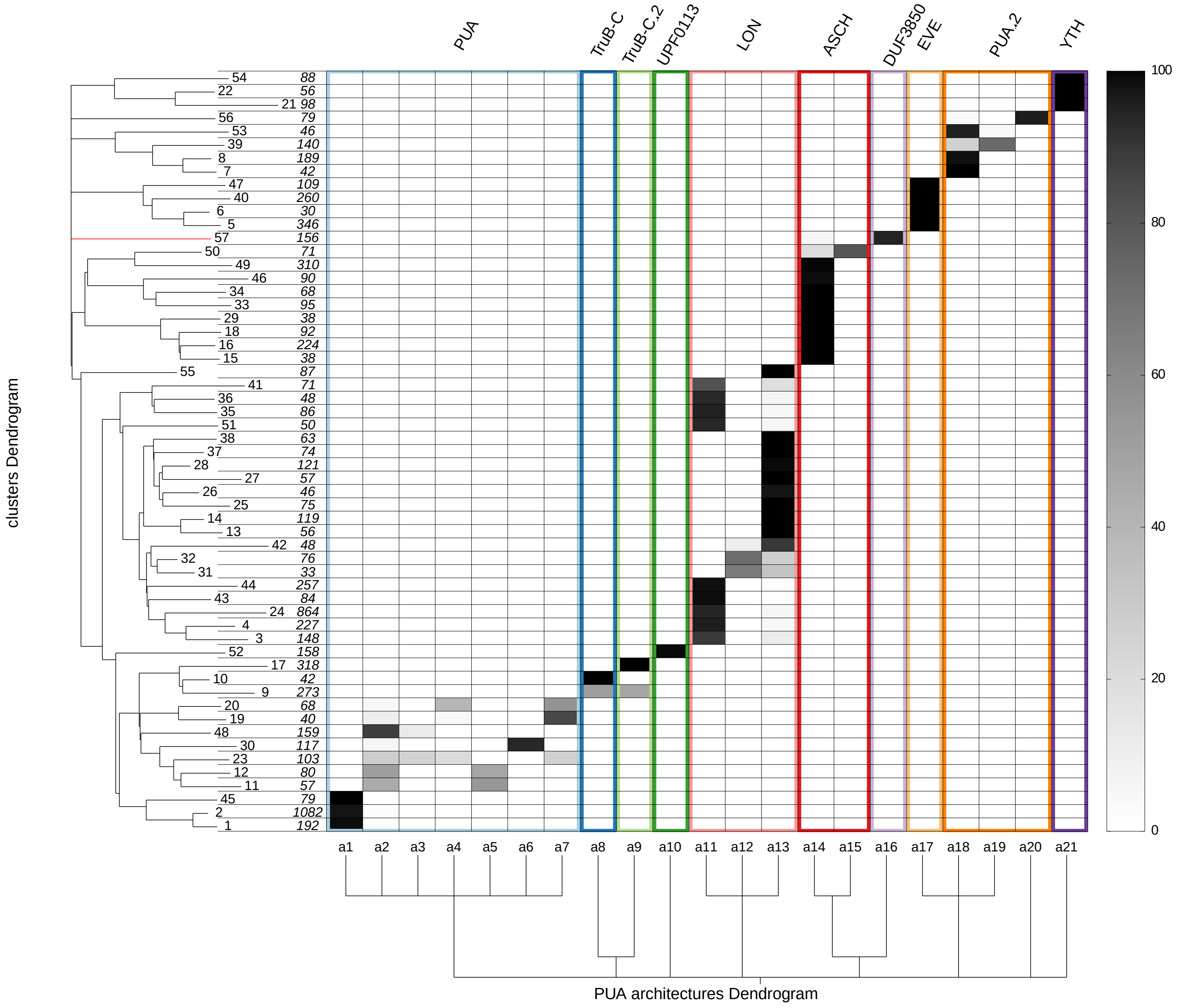
We find ~40 density peaks

- Clusters are pure (contain only proteins from the same family)
- Clusters belonging to the same family are linked together



Automatic recognition of protein families

Comparison between the topography and the PFAM classifiation



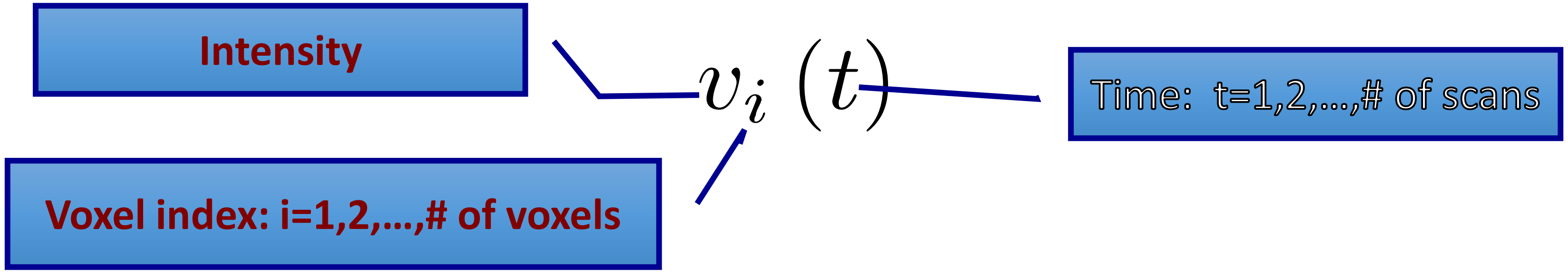
- Results are **consistent with Pfam classification** on the coarse grain scale

- However, families show **a statistically robust inner structure** : ARCHITECTURES!!!!!

- **A automatic and parameter-free approach for classifying protein sequences in families.**

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.



Intensity

$$v_i\,(t)$$

Time: t=1,2,...,# of scans

Voxel index: i=1,2,...,# of voxels

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).

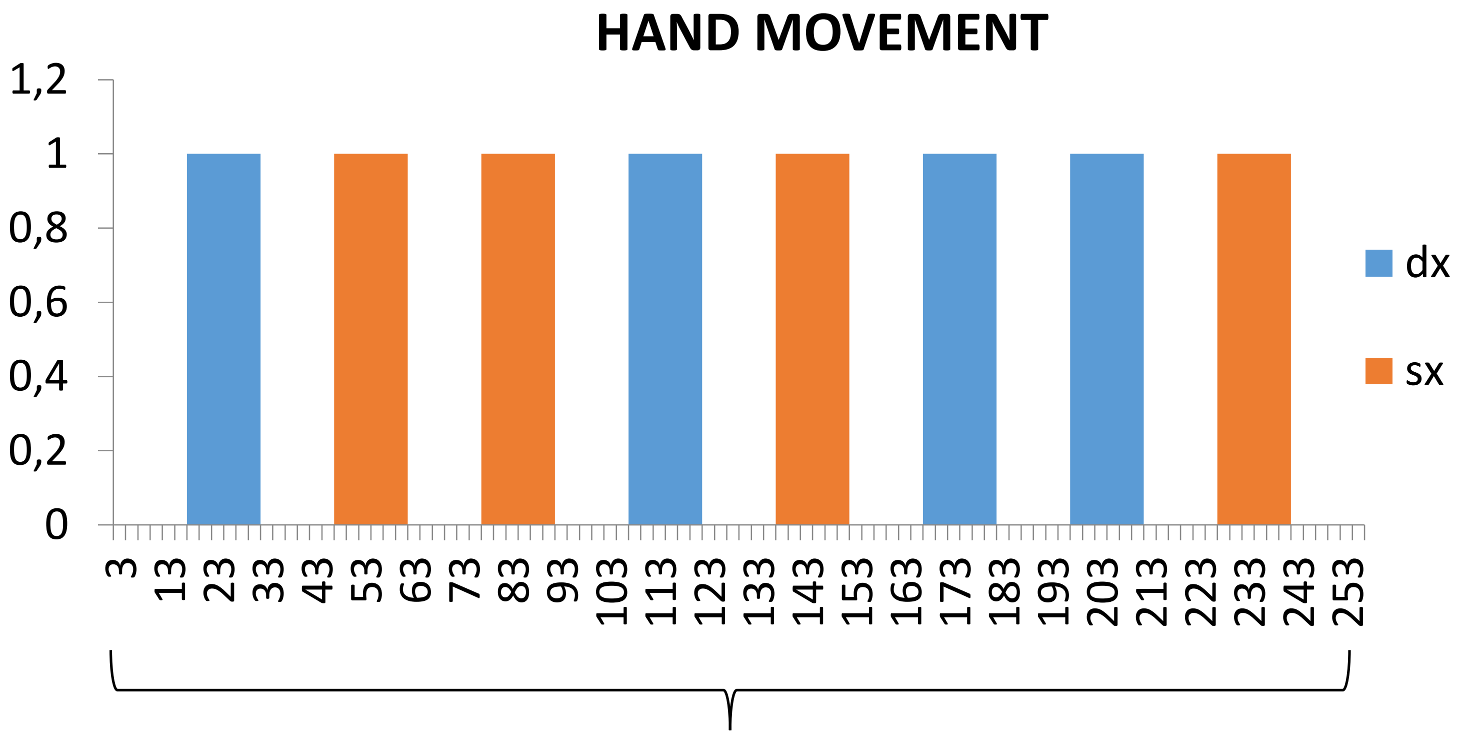
We look for large and connected regions with voxels with a similar v(t), namely with a similar time evolution.

Similarity measure:

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

Analysis of a fMRI experiment (D. Amati, M. Maieron, F. Pizzagalli)

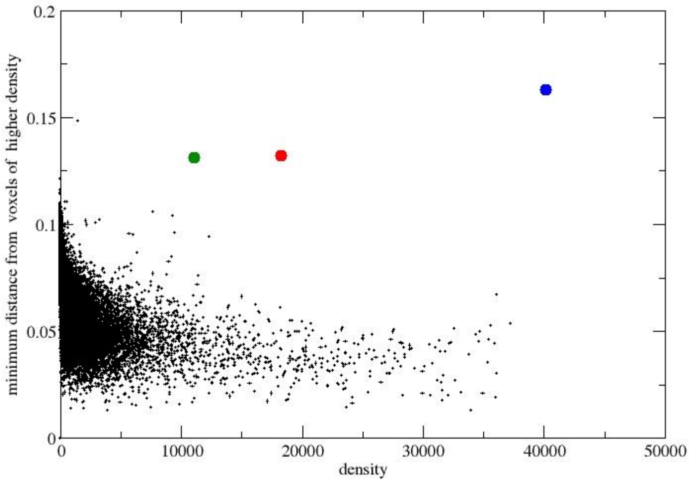
The subject was scanned while moving the right or left hand. They saw the words "move left", "move right" or "stop" in a random fashion through the glasses.



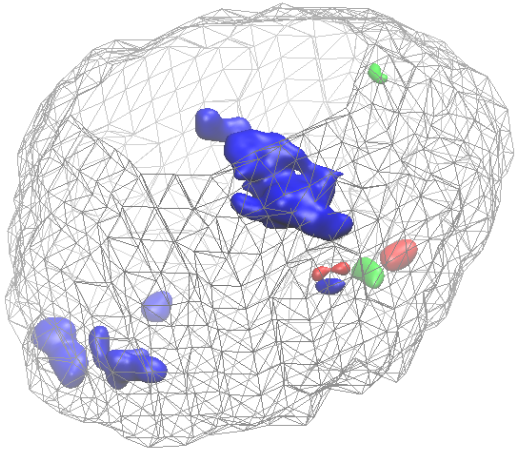
102 scans

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

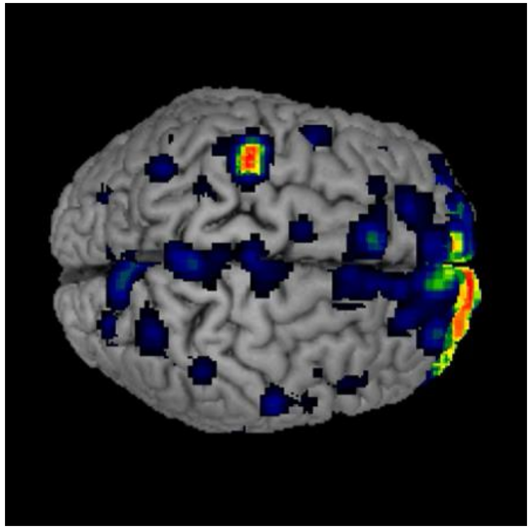
The clustering approach at work:
Analysis of a fMRI experiment (D. Amati, M. Maieron, F. Pizzagalli)

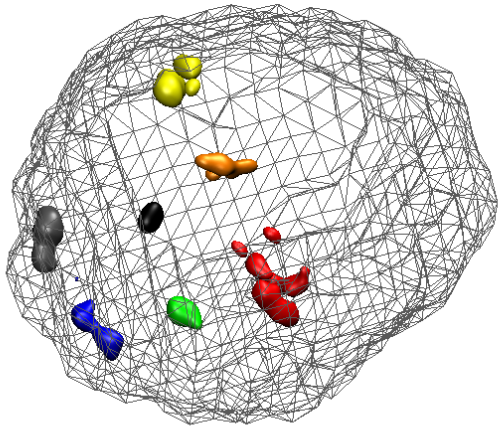


Time window 24-36: decision graph



Time window 24-36: clusters





Overlap between the cluster
of all the time windows

Conclusions

An unsupervised method able to map the topography of a multidimensional probability distributions, providing a measure of the position and height of density peaks and of the saddle points between them.

Key ingredients:

- A robust algorithm for determining the intrinsic dimension of the manifold containing the data [Sci. Rep. (2017)
- A density estimator, capable of providing also an estimate of the error [JCTC (2018)]
- A procedure for finding automatically the probability peaks, regardless of their shape and of the dimensionality [SCIENCE, 1492, vol 322 (2014)]

Coworkers:

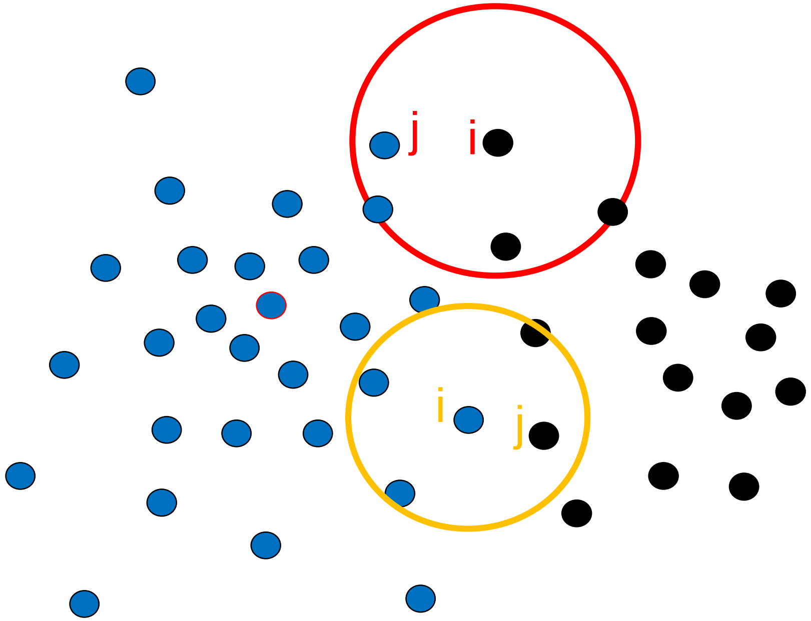
Alex Rodriguez
Maria d'Errico
Elena Facco
Michele Allegra
Giulia Sormani
Elena Russo

Thank-you:

Daniele Amati, Erio Tosatti, Francesca Rizzato, Marco Borrelli

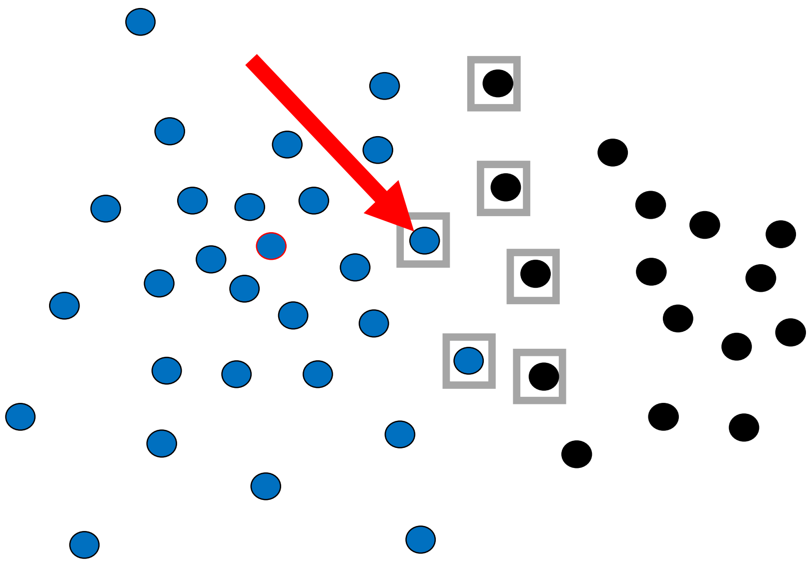
Finding the saddle points

- A point i, belonging to cluster c, is assumed to be at the border between cluster c and c’, if its closest point j belonging to c’, is within a distance $d_{c\hat{}i}$ and if i is the closest point to j among those belonging to c.



Finding the saddle points

- A point i, belonging to cluster c, is assumed to be at the border between cluster c and c’, if its closest point j belonging to c’, is within a distance d_{c^i} and if i is the closest point to j among those belonging to c.
- Saddle point: the point with the highest density among the border points between cluster c and c’,

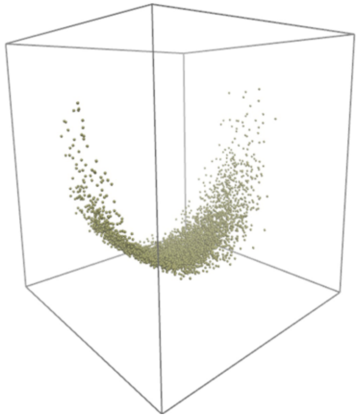


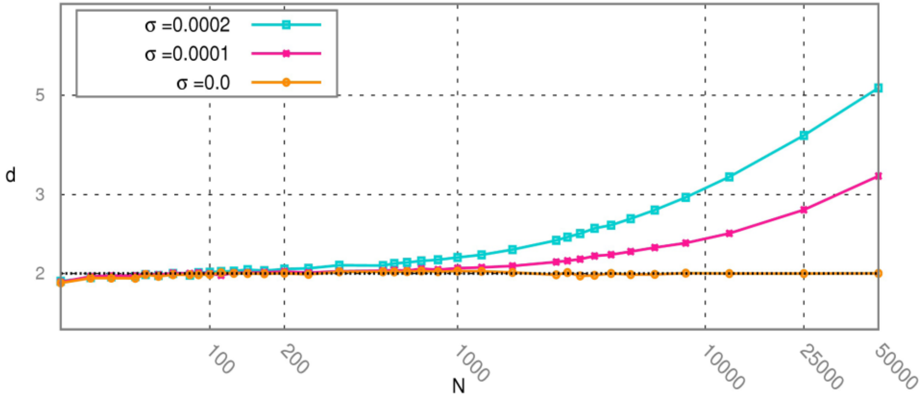
A scale-dependent estimation of the ID:

We randomly extract subsamples in the dataset. The smaller its size, the larger the typical nearest neighbor distance

We compute the ID as a function of the size of the subsample

Example: 2d gaussian wrapped around a swissroll and embedded in a 30 dimensional space+ 30 dimensional noise.





A plateau in the plot of d vs N indicates the number of “soft” directions