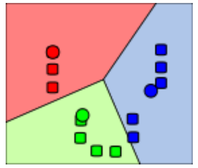
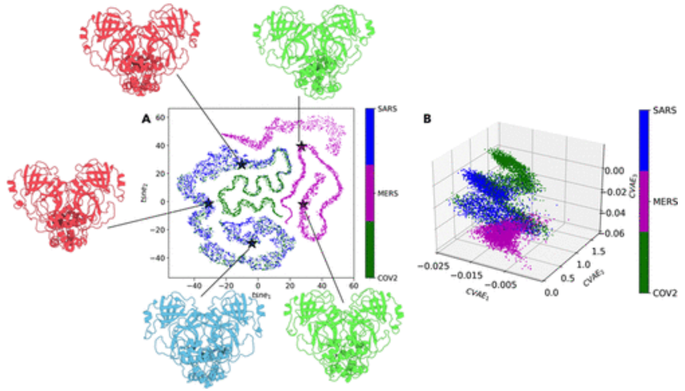
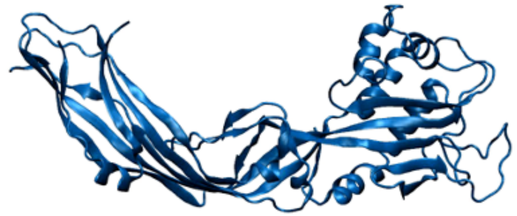
From biomolecular data to information

Antonia Mey

Matteo Degiacomi







Durham University

matteo.t.degiacomi@dur.ac.uk

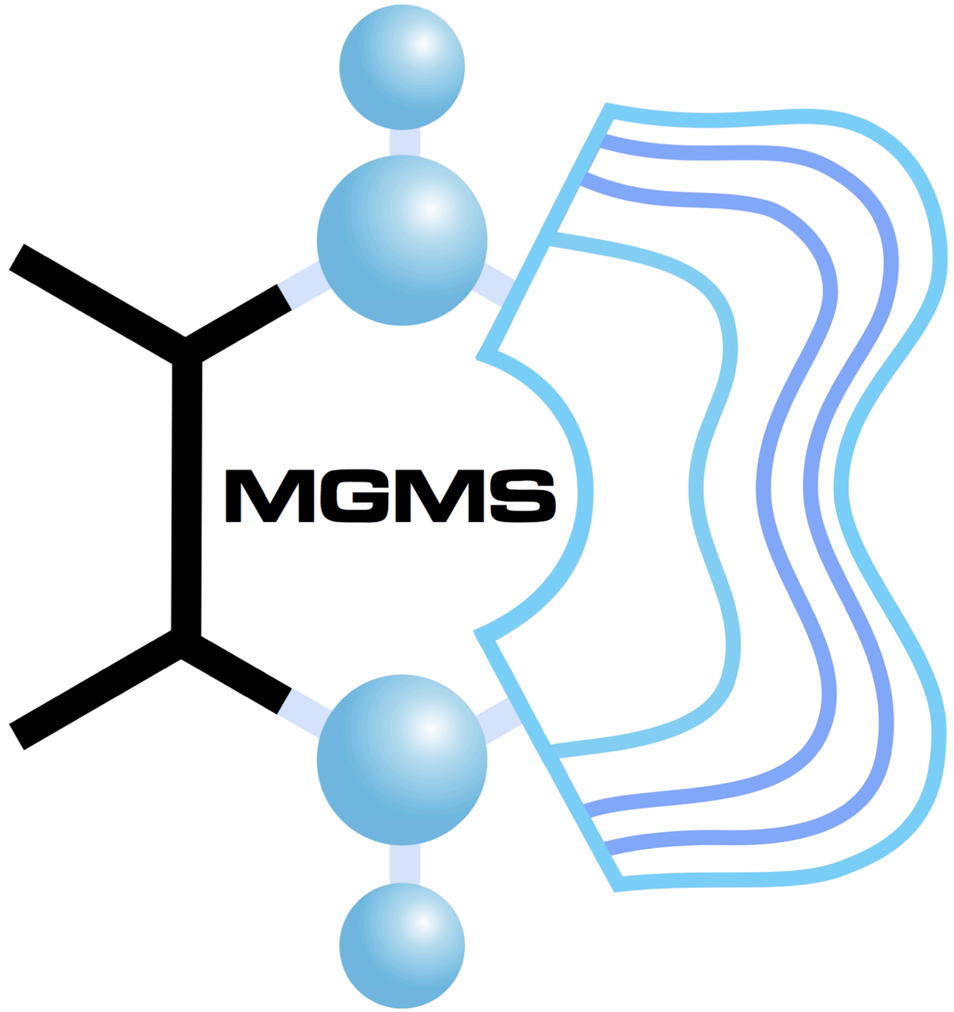
@MatteoDegiacomi

antonia.mey@ed.ac.uk

@ppxasjsm

THE UNIVERSITY of EDINBURGH

Thanks to our sponsors



Schedule

Thursday — Antonia Mey

13:30-14:45	ML Clustering
14:45-15:00	☕ break ☕
15:00-16:20	ML Dimensionality Reduction
16:20-16:30	Closing remarks
18:00-onwards	Informal social event

Friday — Matteo Degiacomi

09:30-10:45	ML Classification
10:45-11:00	☕ break ☕
11:00-12:30	ML Regressions and Neural Networks
12:30-13:30	Lunch
13:30-onwards	Bring your own problem

What is machine learning?

Artificial intelligence

Design an intelligent agent that perceives its environment and makes decisions to maximise chances of achieving its goal.

Machine learning

Gives computers the ability to learn without specifically being programmed (Arthur Samuel 1959)

Supervised learning

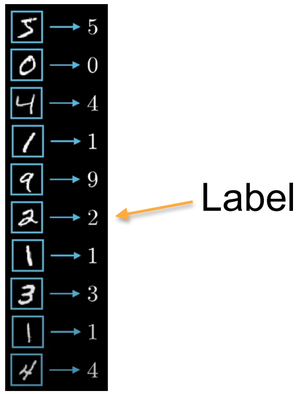
Unsupervised learning

reinforcement learning

https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12

Some machine learning terminology

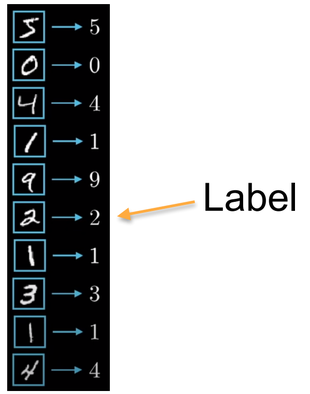
Data with labels

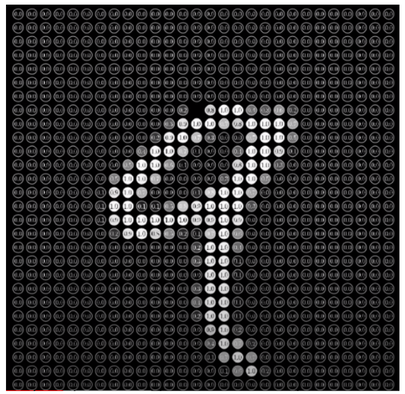


Some machine learning terminology

Data with labels

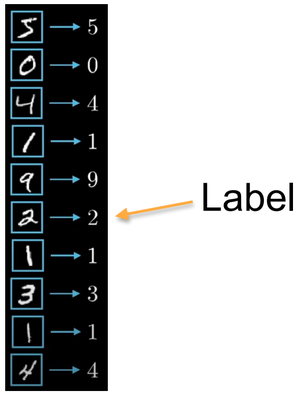
Feature





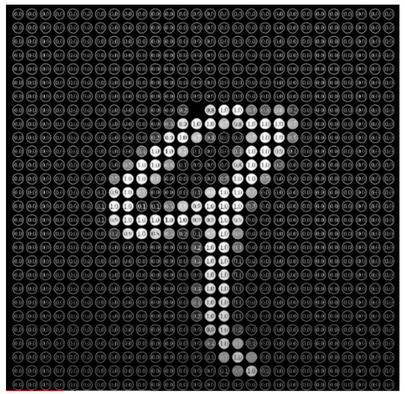
Some machine learning terminology

Data with labels



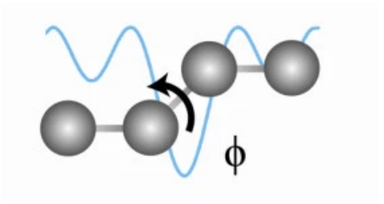
Feature

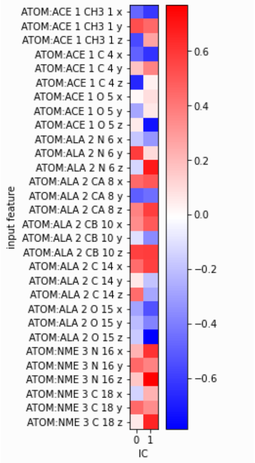
28x28

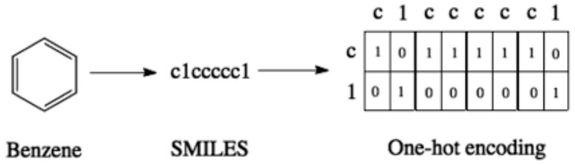


vector with 784 entries

Features in MD

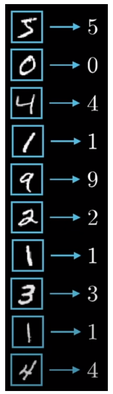






Some machine learning terminology

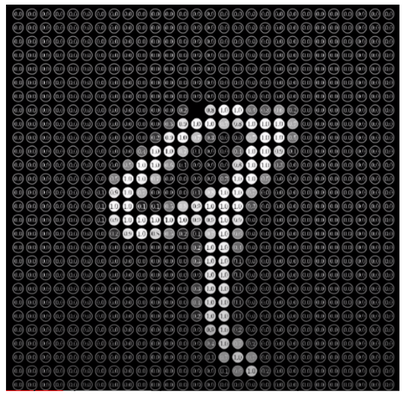
Data with labels



Label

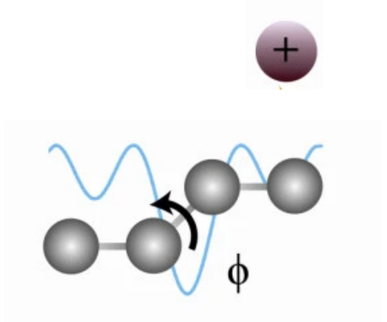
Feature

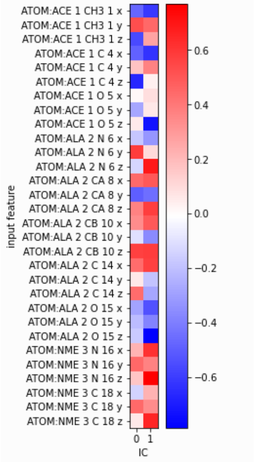
28x28

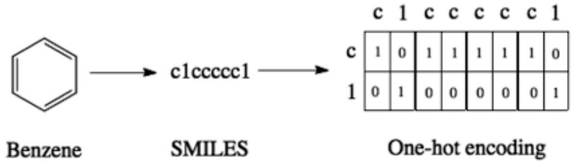


vector with 784 entries

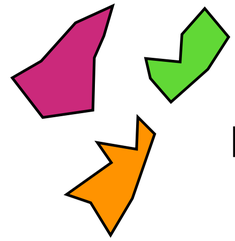
Features in MD







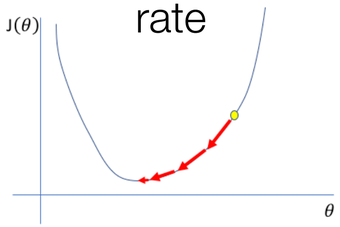
Hyper parameter



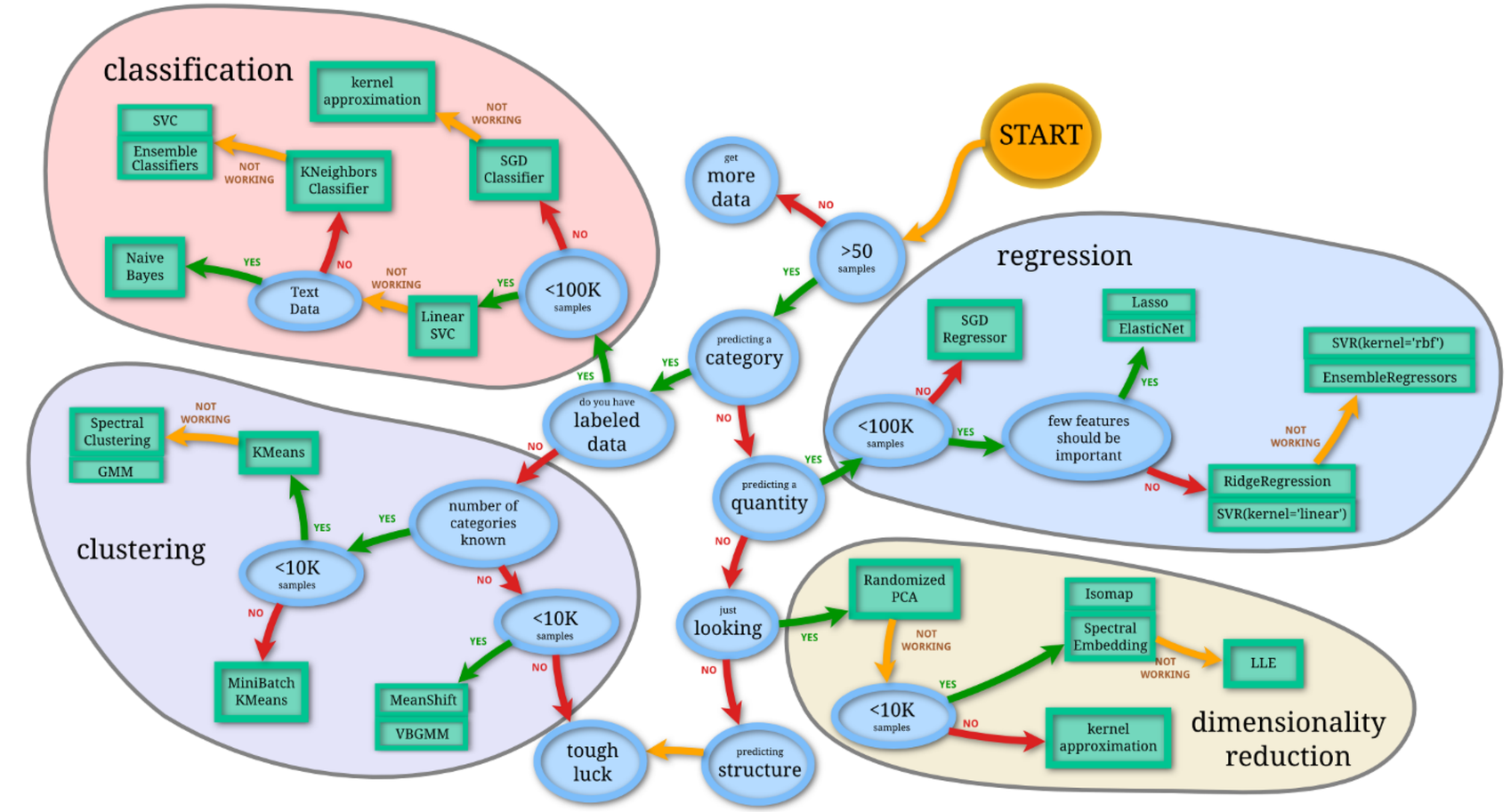
N-clusters

Dimensions for reduction

Learning rate

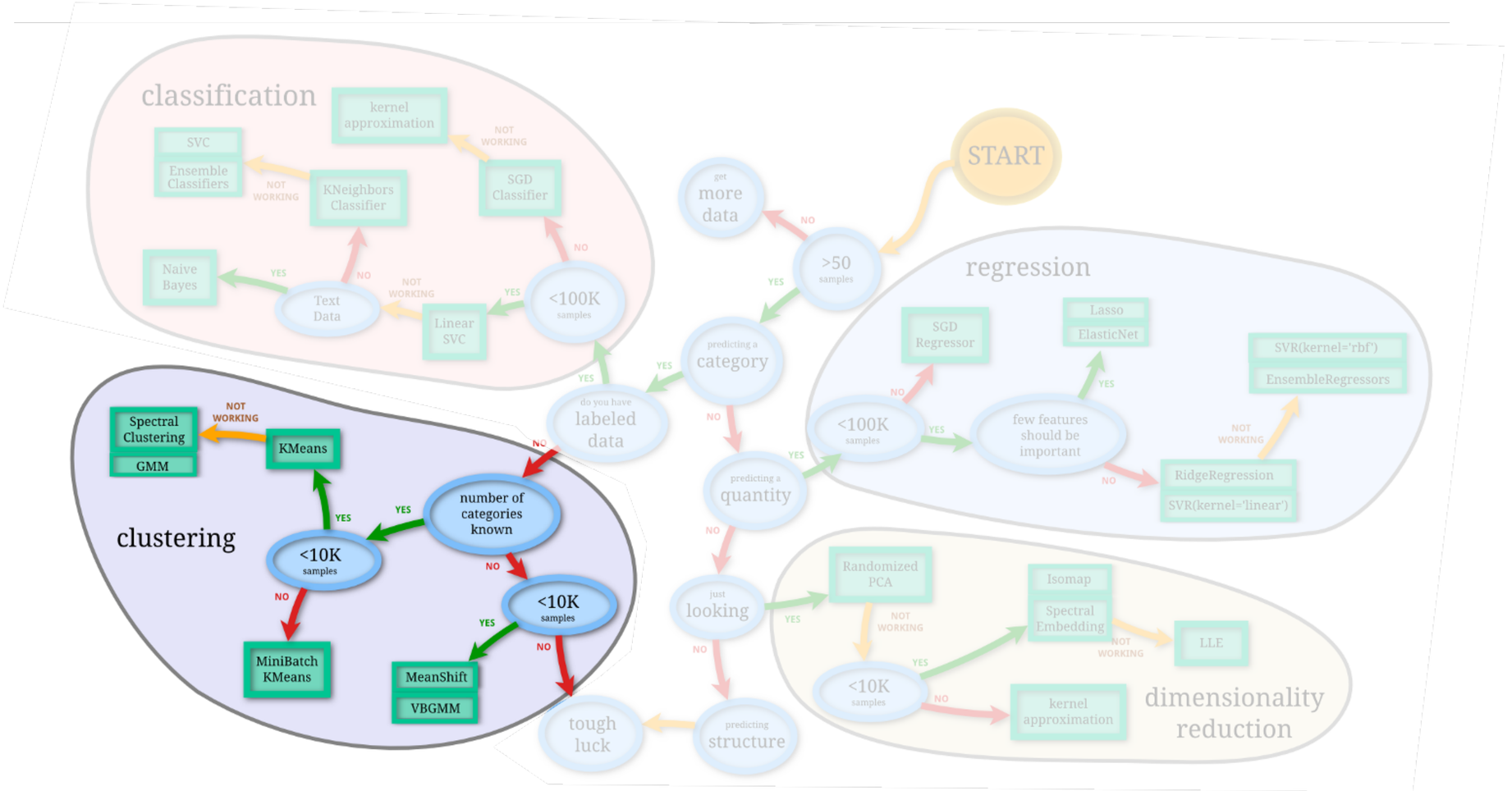


The Data Mining World



From **scikit-learn.org**

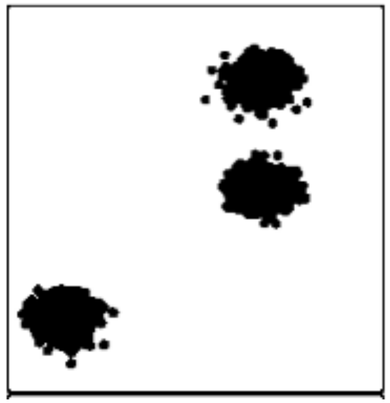
The Data Mining World — Clustering



From **scikit-learn.org**

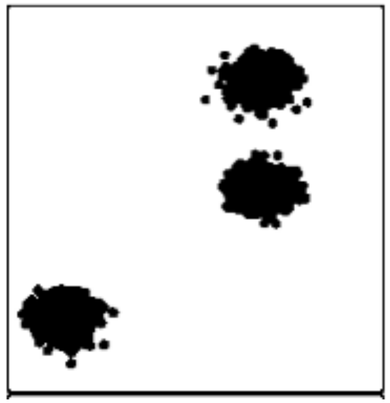
Clustering is an unsupervised learning technique

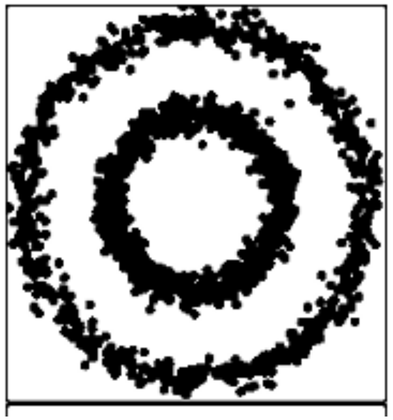
How many clusters are there?



Clustering is an unsupervised learning technique

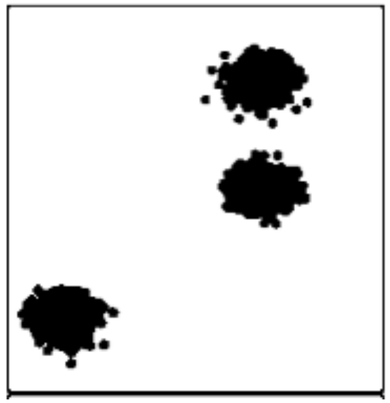
How many clusters are there?

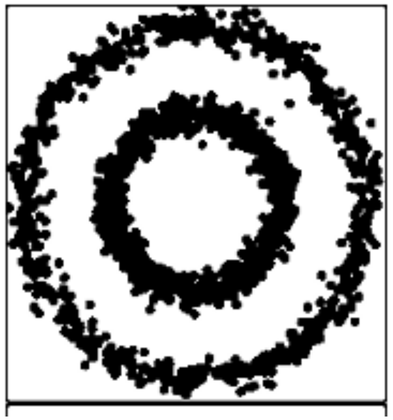




Clustering is an unsupervised learning technique

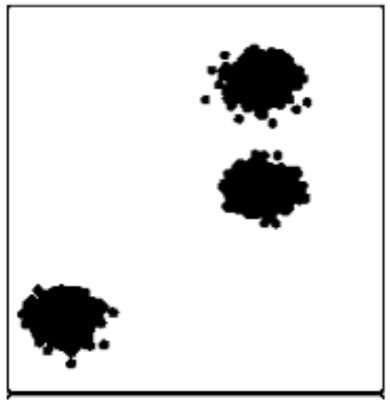
How many clusters are there?

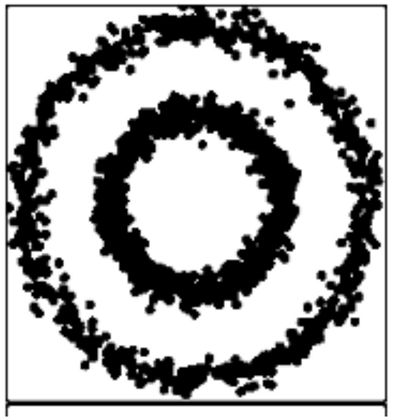


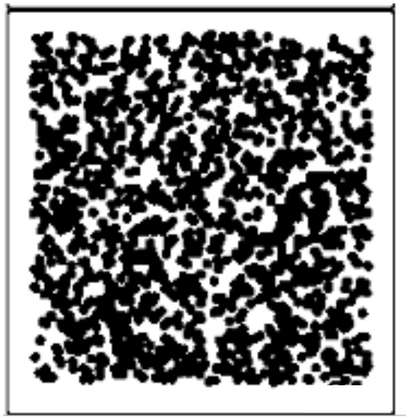


Clustering is an unsupervised learning technique

How many clusters are there?

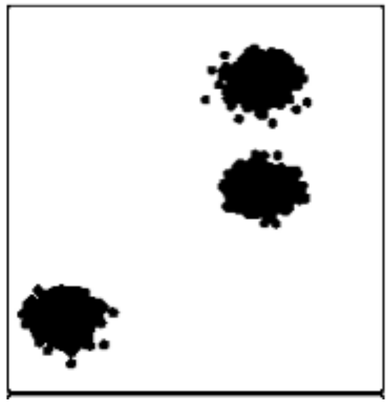
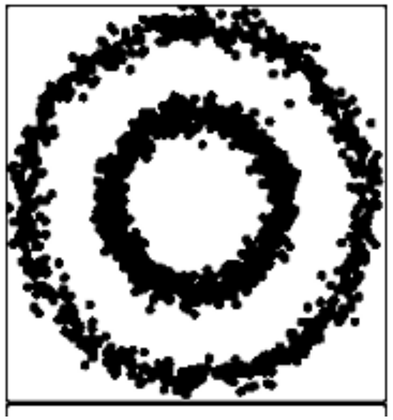


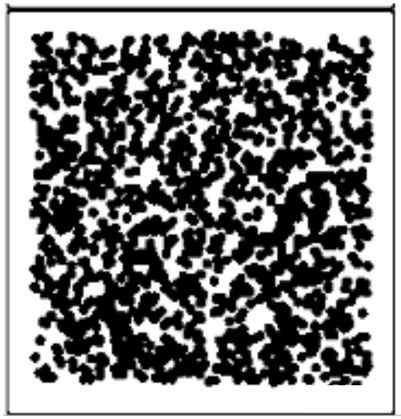




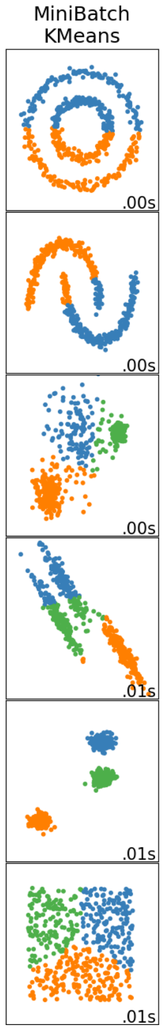
Clustering is an unsupervised learning technique

How many clusters are there?

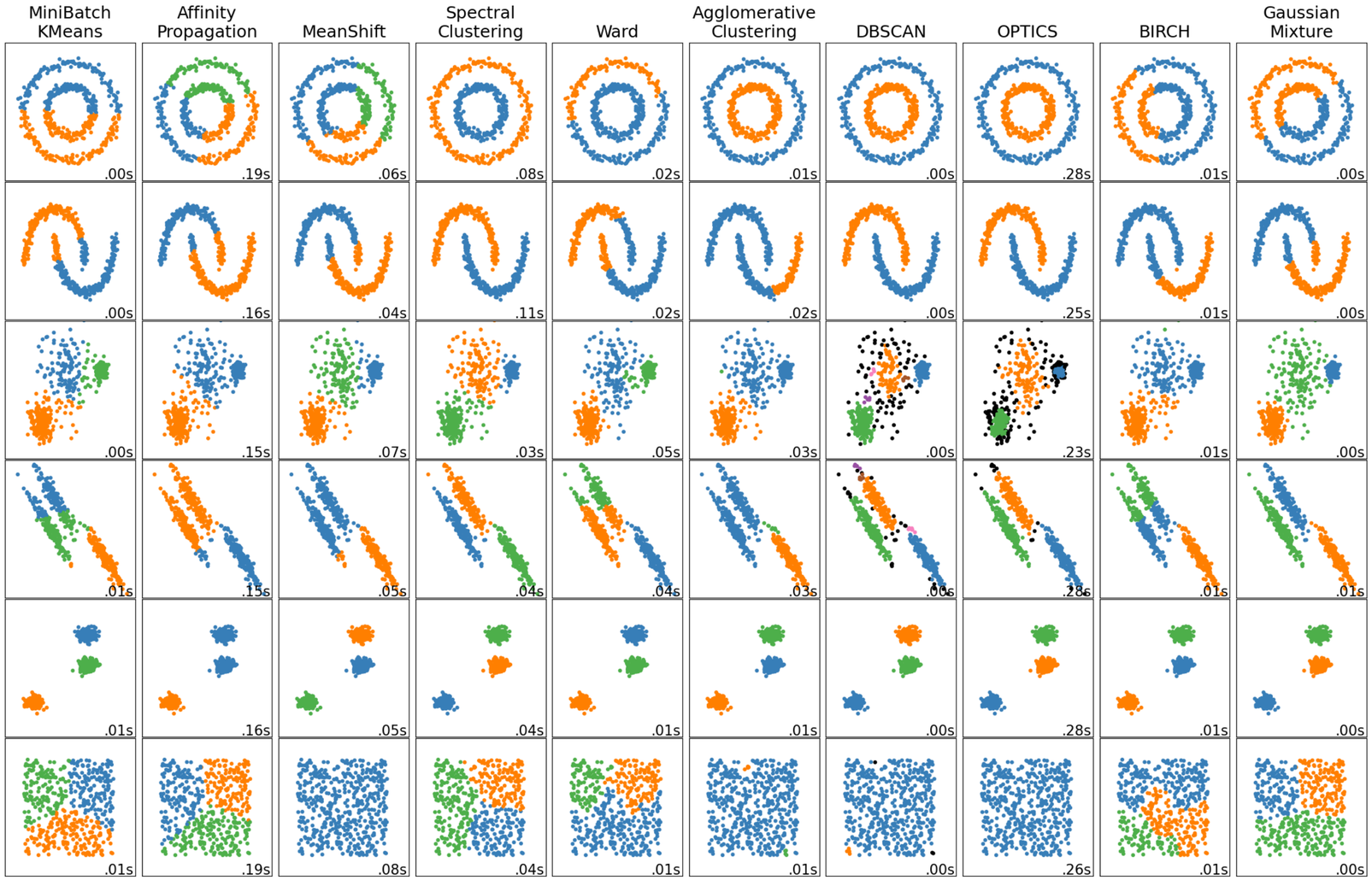






There are many different clustering algorithms

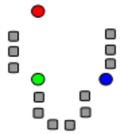


There are many different clustering algorithms

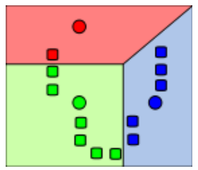


K-means, DBSCAN and spectral clustering

K-means

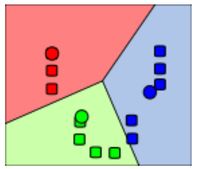


Initial guess



K-clusters are generated with the nearest mean

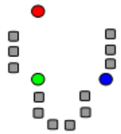
Centroid of the k-lcusters becomes the new mean



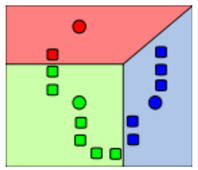
Iterate until convergence

K-means, DBSCAN and spectral clustering

K-means

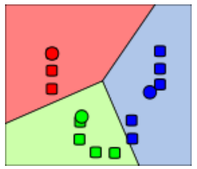


Initial guess



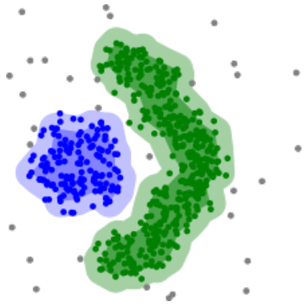
K-clusters are generated with the nearest mean

Centroid of the k-lcusters becomes the new mean



Iterate until convergence

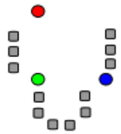
DBSCAN



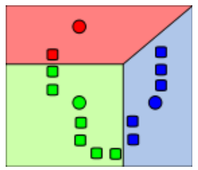
1. Find the points in the ε (eps) neighbourhood of every point, and identify the core points with more than minPts neighbours.
2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise assign it to noise.

K-means, DBSCAN and spectral clustering

K-means

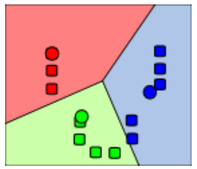


Initial guess



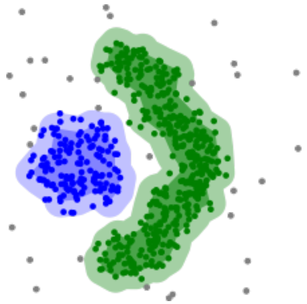
K-clusters are generated with the nearest mean

Centroid of the k-lcusters becomes the new mean



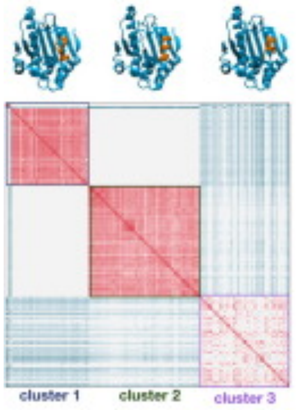
Iterate until convergence

DBSCAN



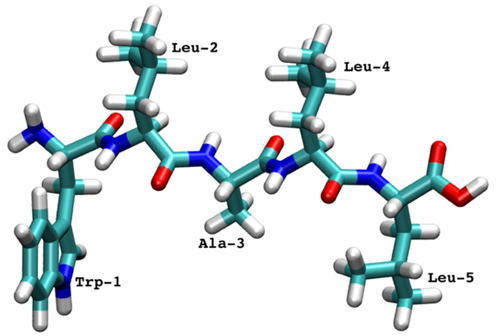
1. Find the points in the ε (eps) neighbourhood of every point, and identify the core points with more than minPts neighbours.
2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise assign it to noise.

Spectral clustering

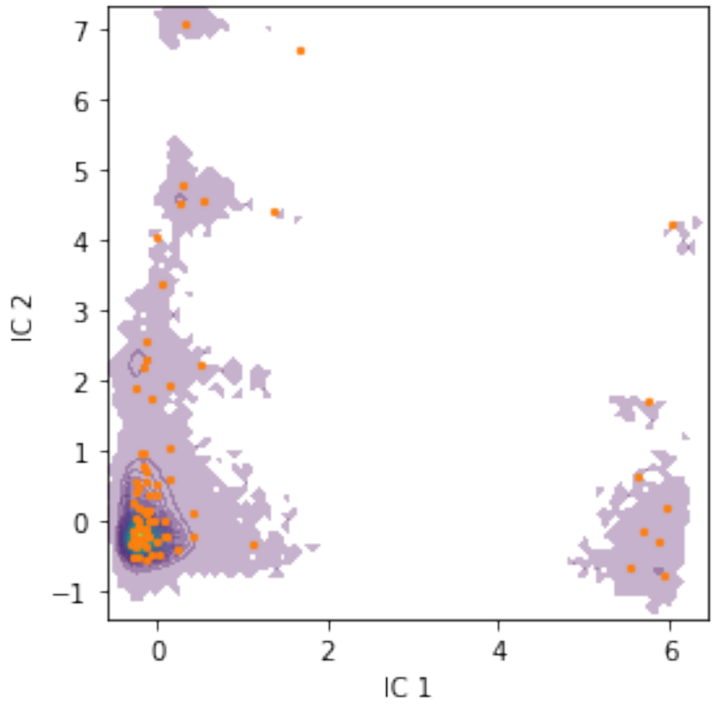


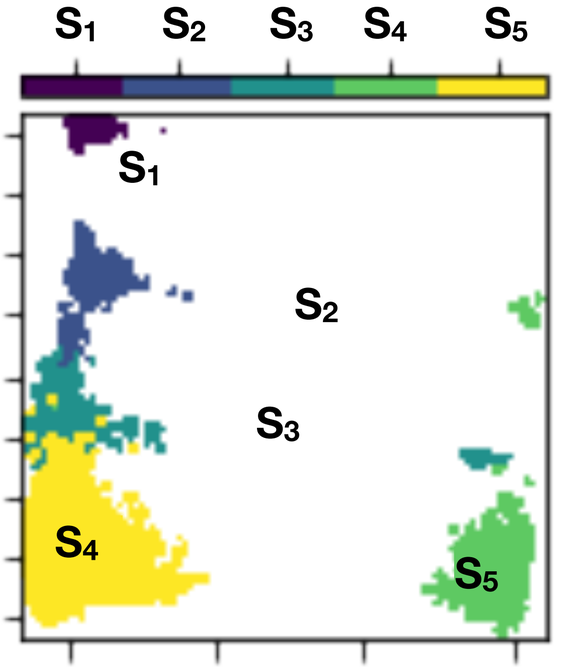
In spectral clustering clusters are found by doing an eigenvalue decomposition of the Laplacian

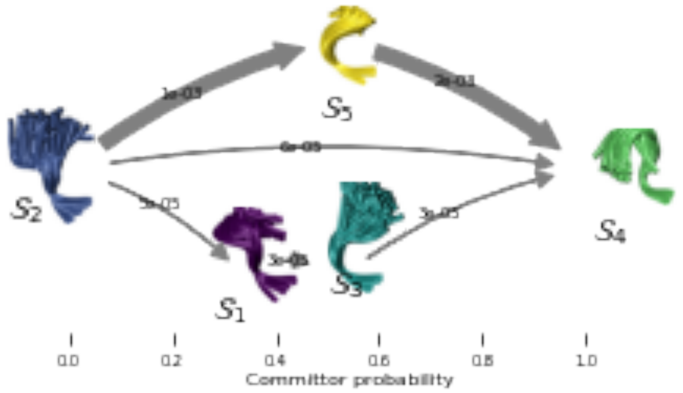
K-means example



Clustering is one of the first steps in building a Markov State Model








Break Time!