Unsupervised Learning

Find groups in the data

- No labels nor response -> unsupervised
- Define groups based on similarity

Group customers, target ads

- A priori, you can’t really put labels on customers
- Group similar customers

You can then try to interpret the grouping, and send targeted ads to the groups

Note: you may want to assign labels to groups a posteriori

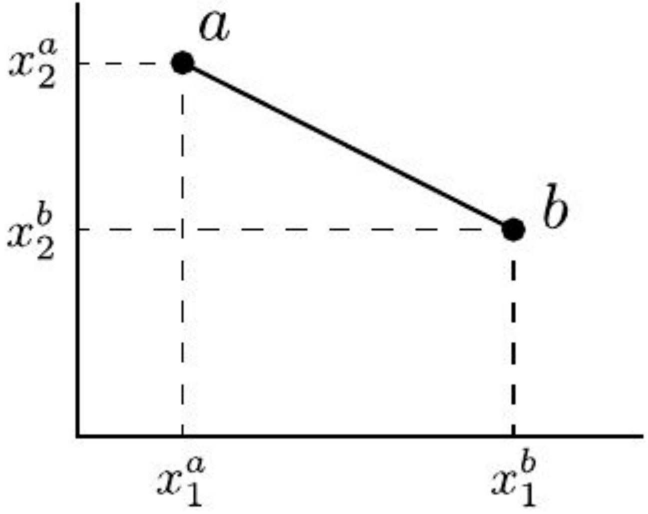
CAMBRIDGE SPARK

Defining similarity

After a pre-processing step, you have a data matrix with **n** rows (observations) and **p** columns (features). Each row is a “point”.

- How to define similarity between points?
- If the features are numerical, we can use the Euclidean distance
- What if some features are categorical?
    - Ignore
    - Embed into numerical

Euclidean distance



$$d(a, b)^2 = \sum_{i=1:2} (x_i^a - x_i^b)^2$$

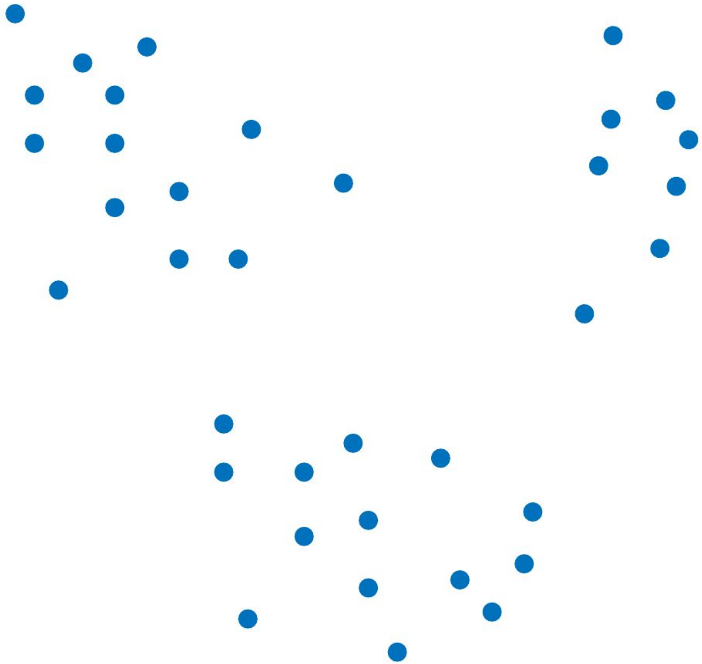
Can be generalised from 2-D to n-D

CAMBRIDGE SPARK

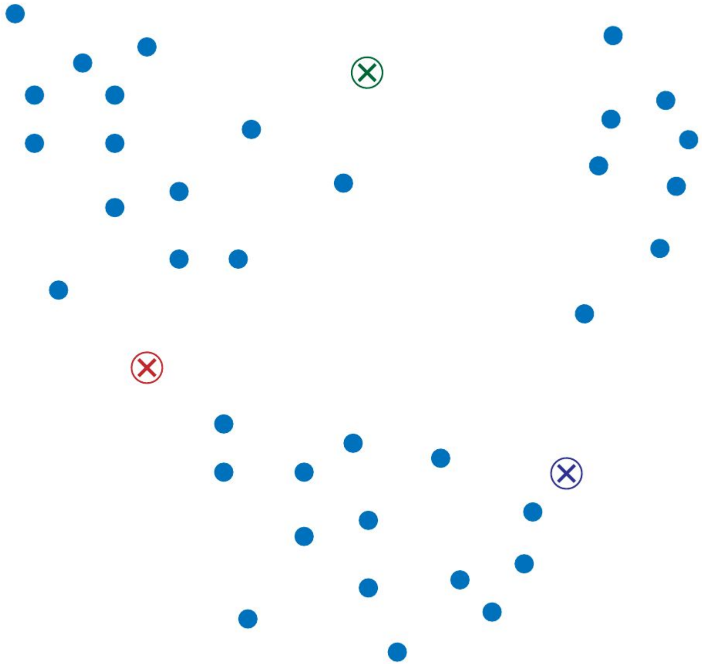
KMeans

CAMBRIDGE SPARK

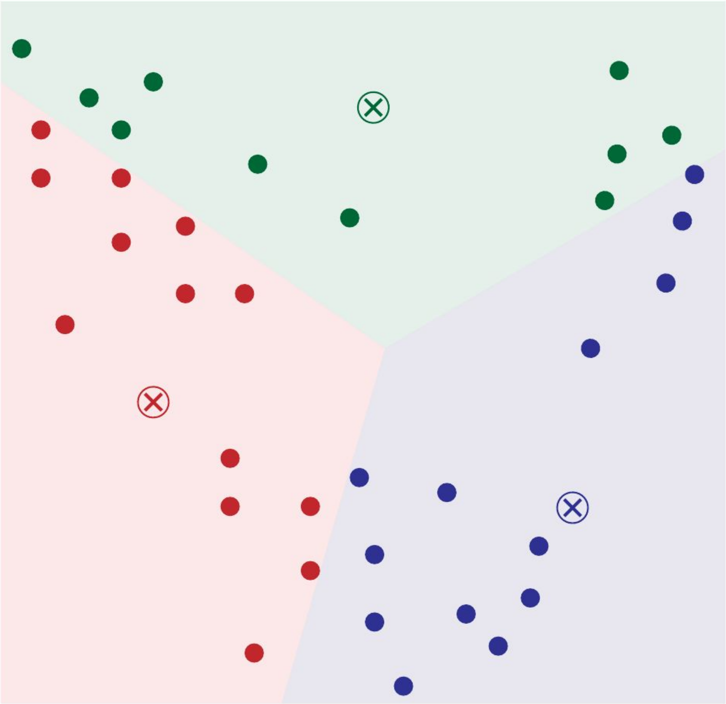
K-means



K-means

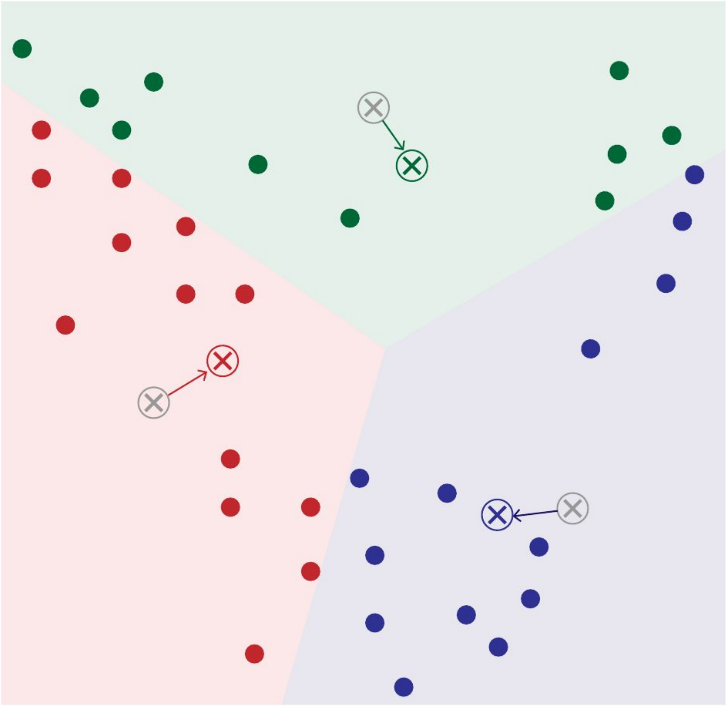


K-means

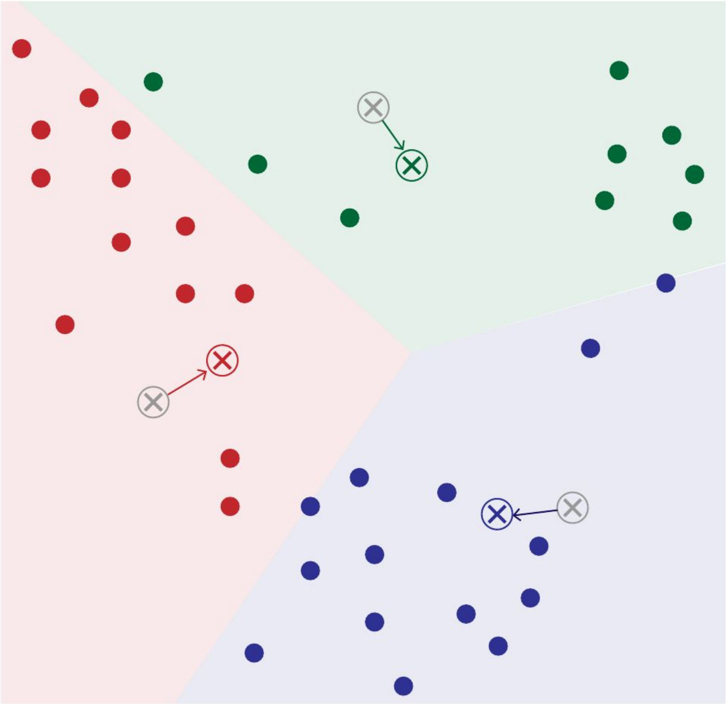


CAMBRIDGE SPARK

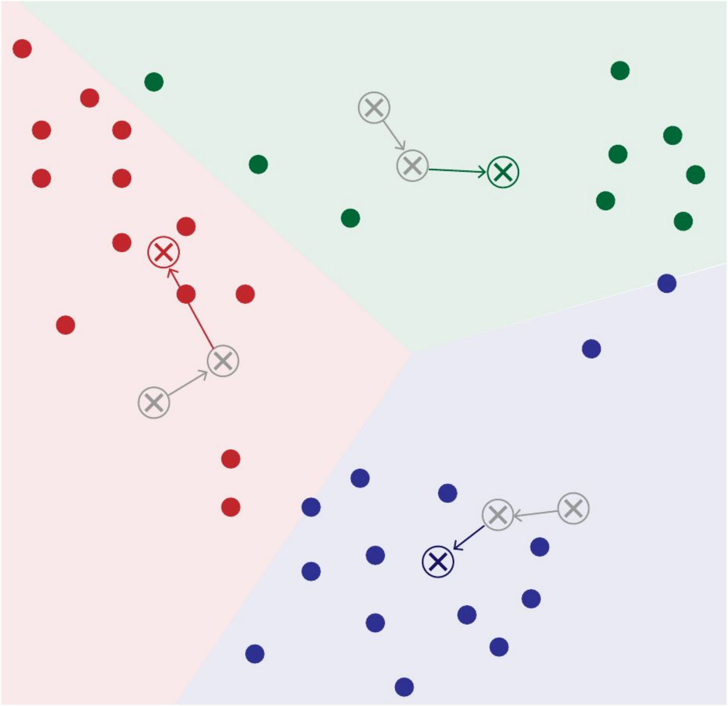
K-means



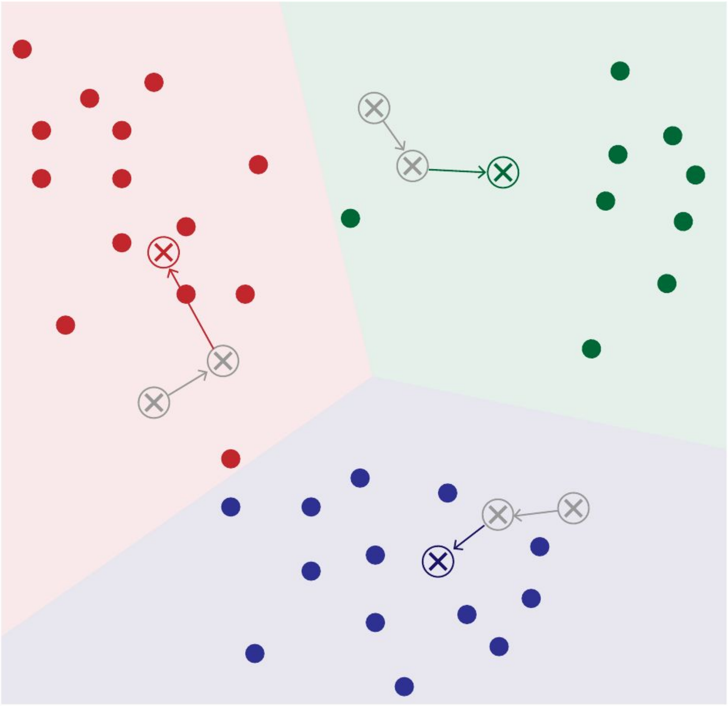
K-means



K-means

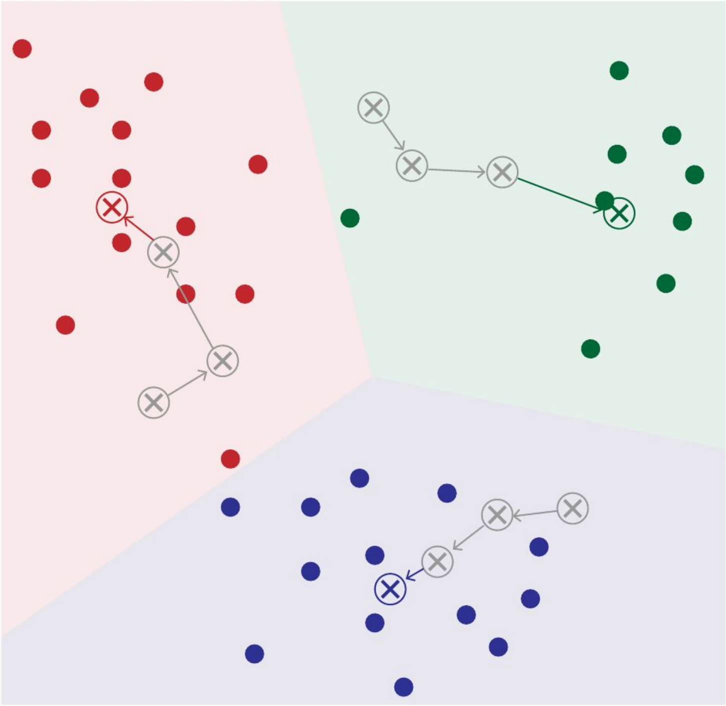


K-means

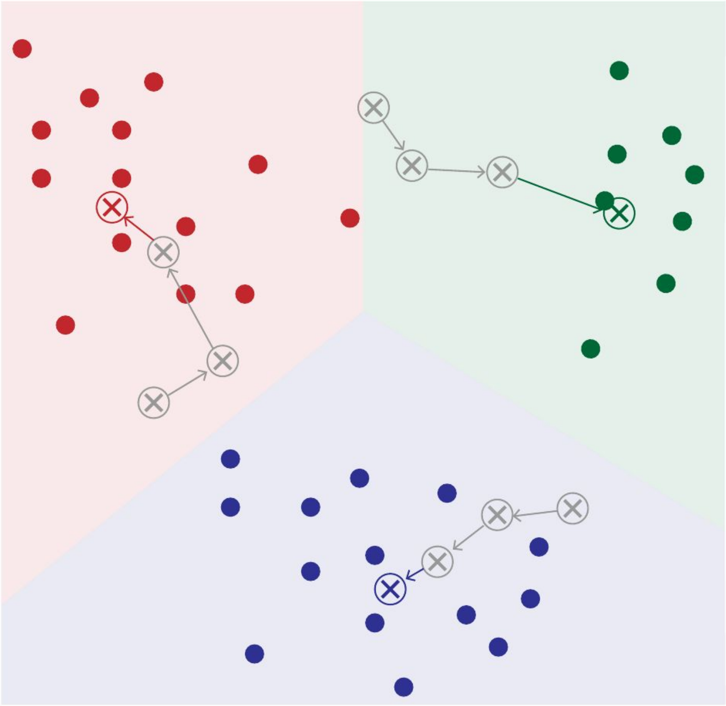


CAMBRIDGE SPARK

K-means



K-means



K-means - Summary

- Start with K “means” drawn at random
- Assign data points to the nearest one
- Update the position of the means to correspond to the mean of those points
- Repeat...

CAMBRIDGE SPARK

K-means - Pros and cons

Pros

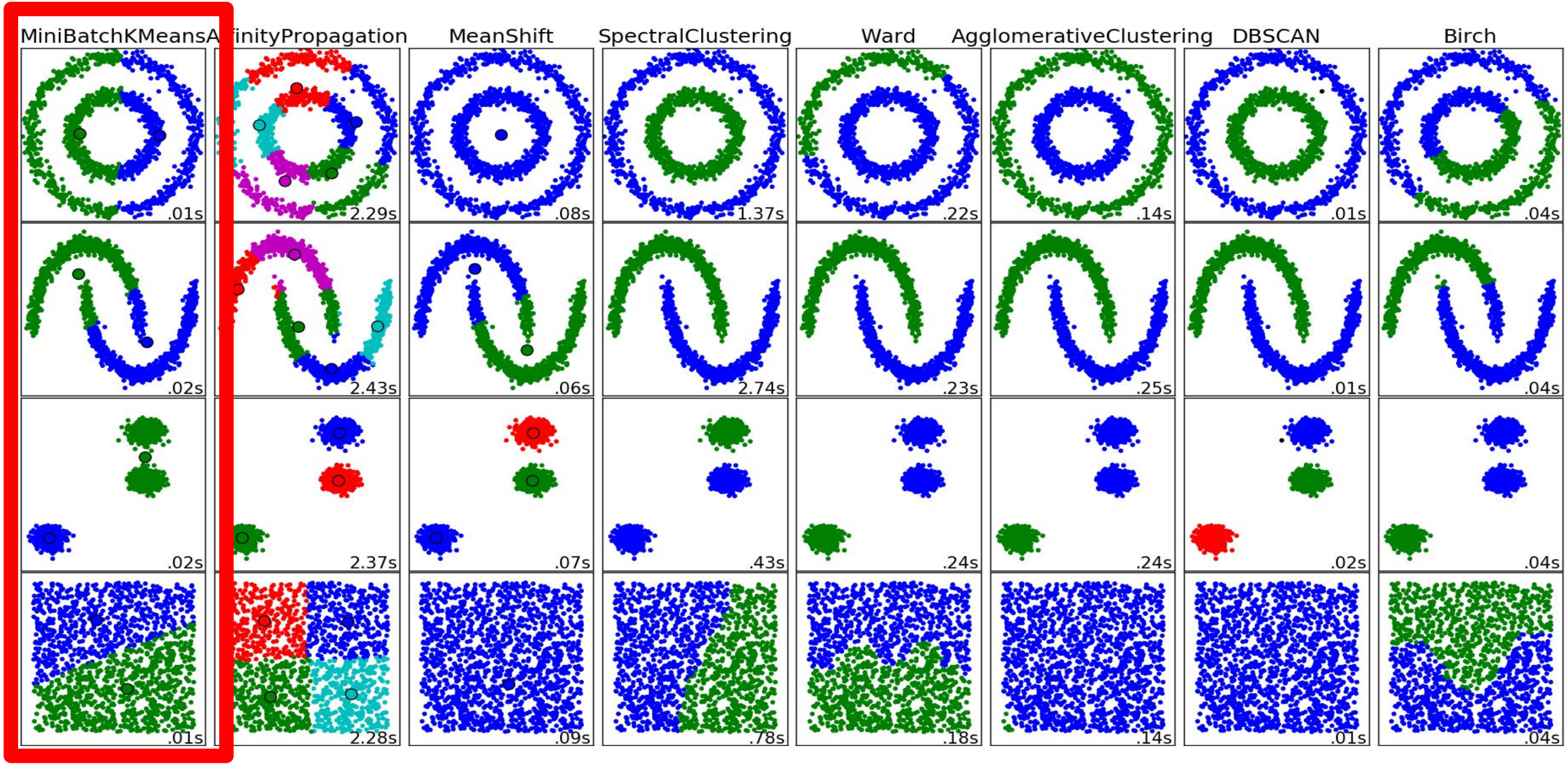
- Cheap to compute
- Easy to interpret
- Efficient implementations available
- Assigning a new point is straightforward

Cons

- Need to guess K
- Clusters are globular
- Sensitive to initialisation
- Sensitive to noise

CAMBRIDGE SPARK

Comparisons of clustering algorithms



RK

Hands-on session

01-unsupervised_learning.ipynb

Hierarchical Clustering

CAMBRIDGE SPARK

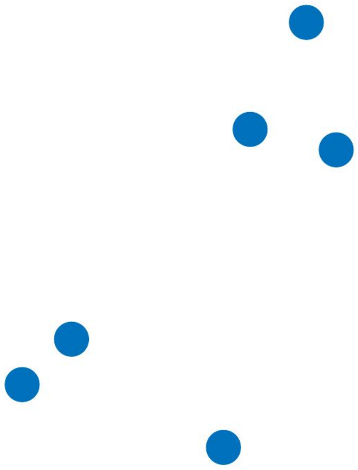
Hierarchical clustering

Building a hierarchy of clusters sequentially.

- **Agglomerative** (bottom-up): Start considering each point as a cluster then merge the closest ones and repeat
- **Divisive** (top-down): Start with one single cluster and divide to have groups with reduced variance

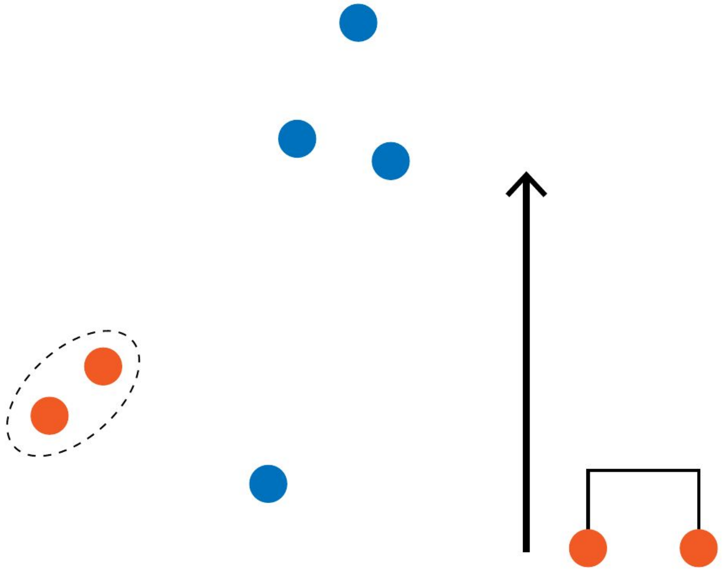
Let’s take a look at agglomerative hierarchical clustering (a.k.a. linkage)

Linkage algorithm

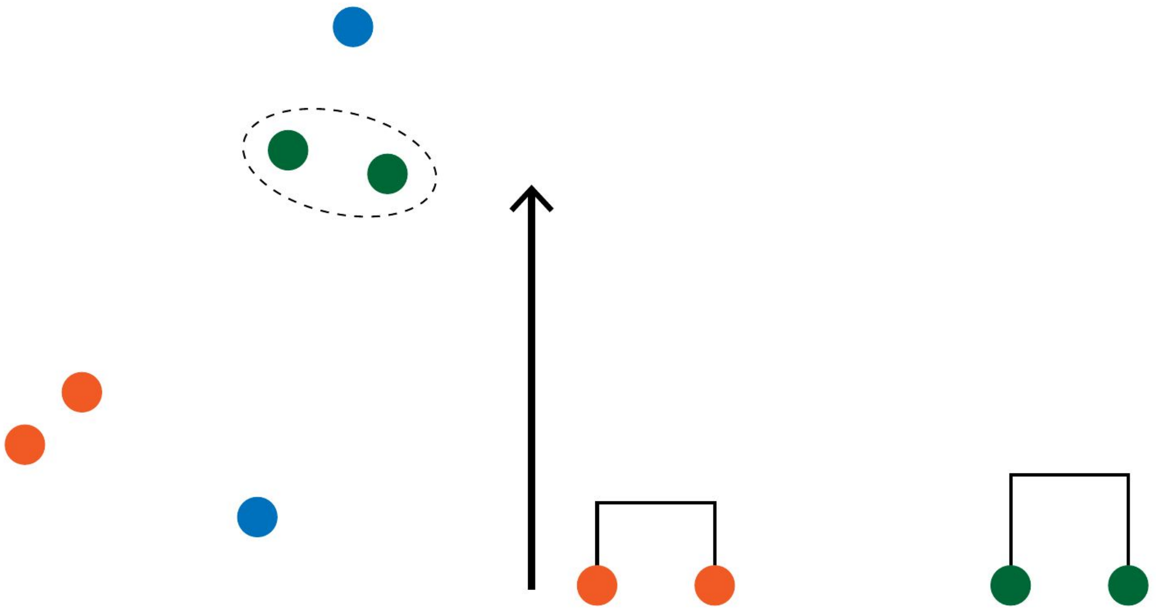


CAMBRIDGE SPARK

Linkage algorithm

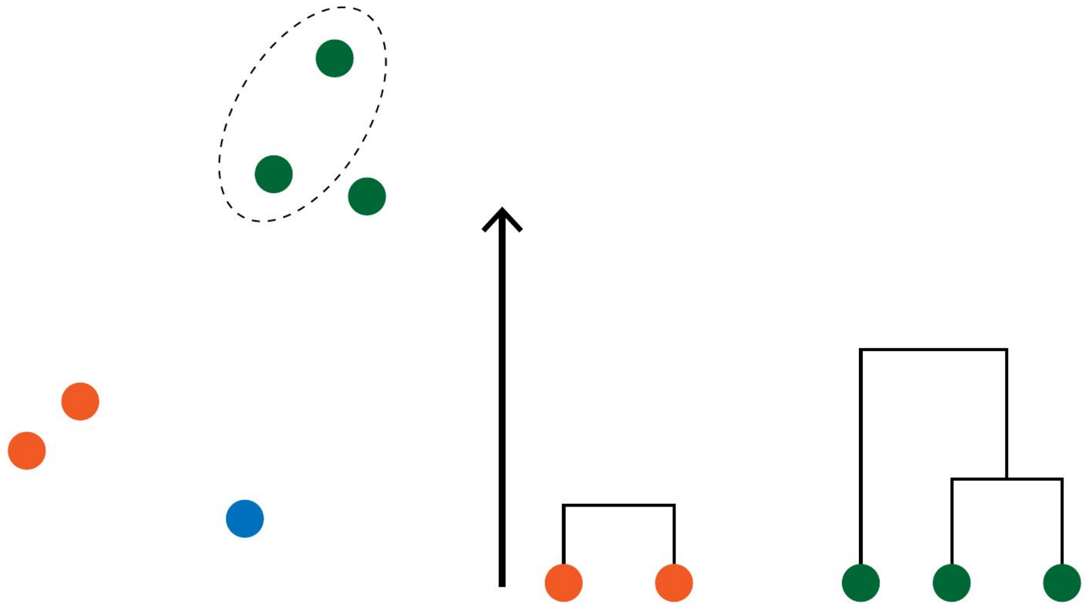


Linkage algorithm



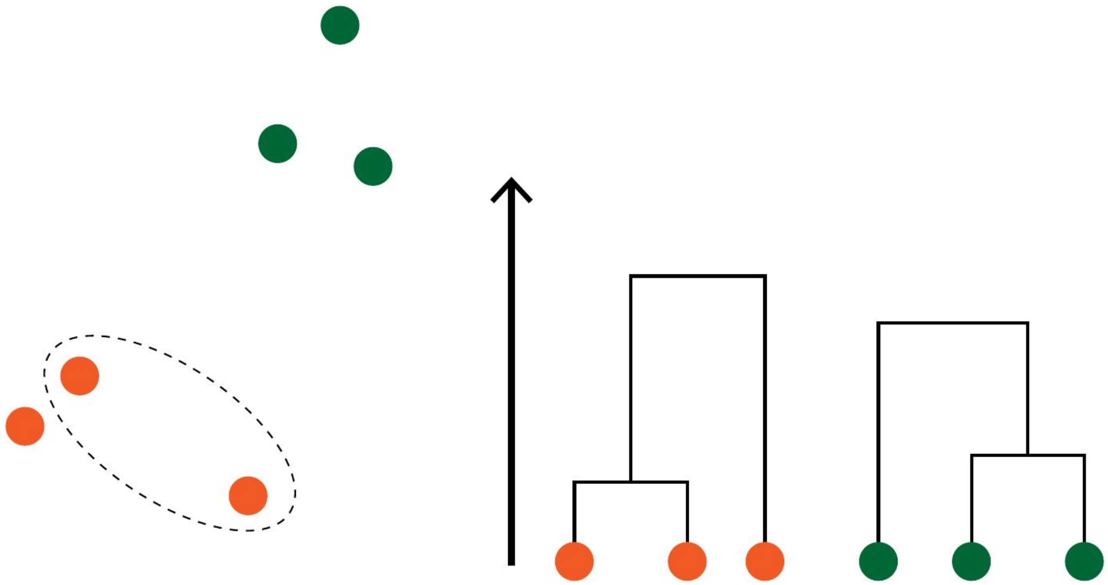
CAMBRIDGE SPARK

Linkage algorithm



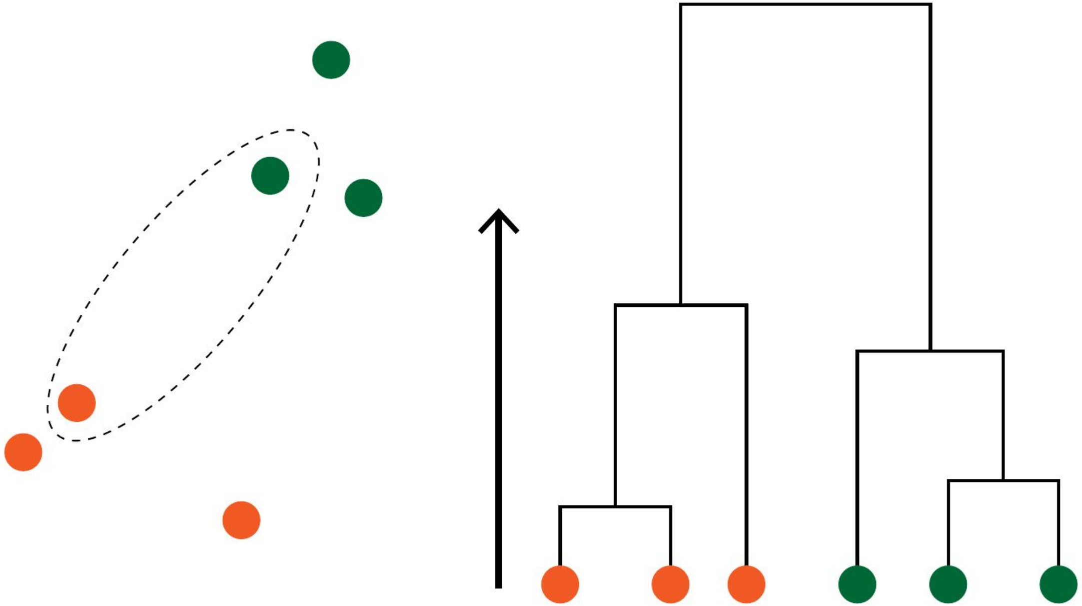
CAMBRIDGE SPARK

Linkage algorithm



CAMBRIDGE SPARK

Linkage algorithm

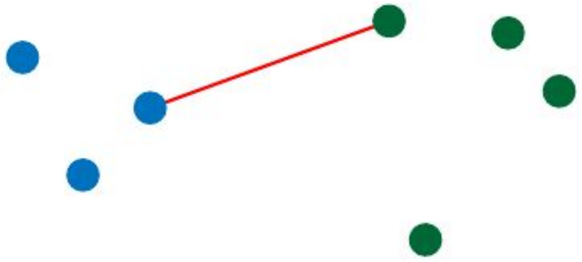


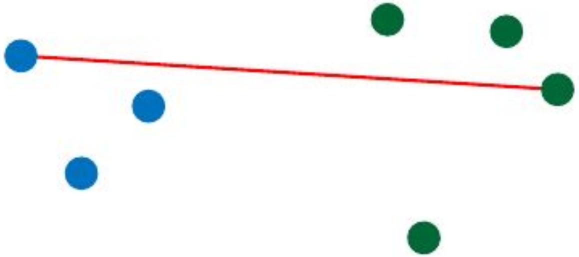
CAMBRIDGE SPARK

Linkage algorithm

Two strategies to merge clusters:

- **Single linkage**: closest point distance (build spanning trees)
- **Complete linkage**: furthest point distance (to avoid elongated clusters)





CAMBRIDGE SPARK

Linkage: pros and cons

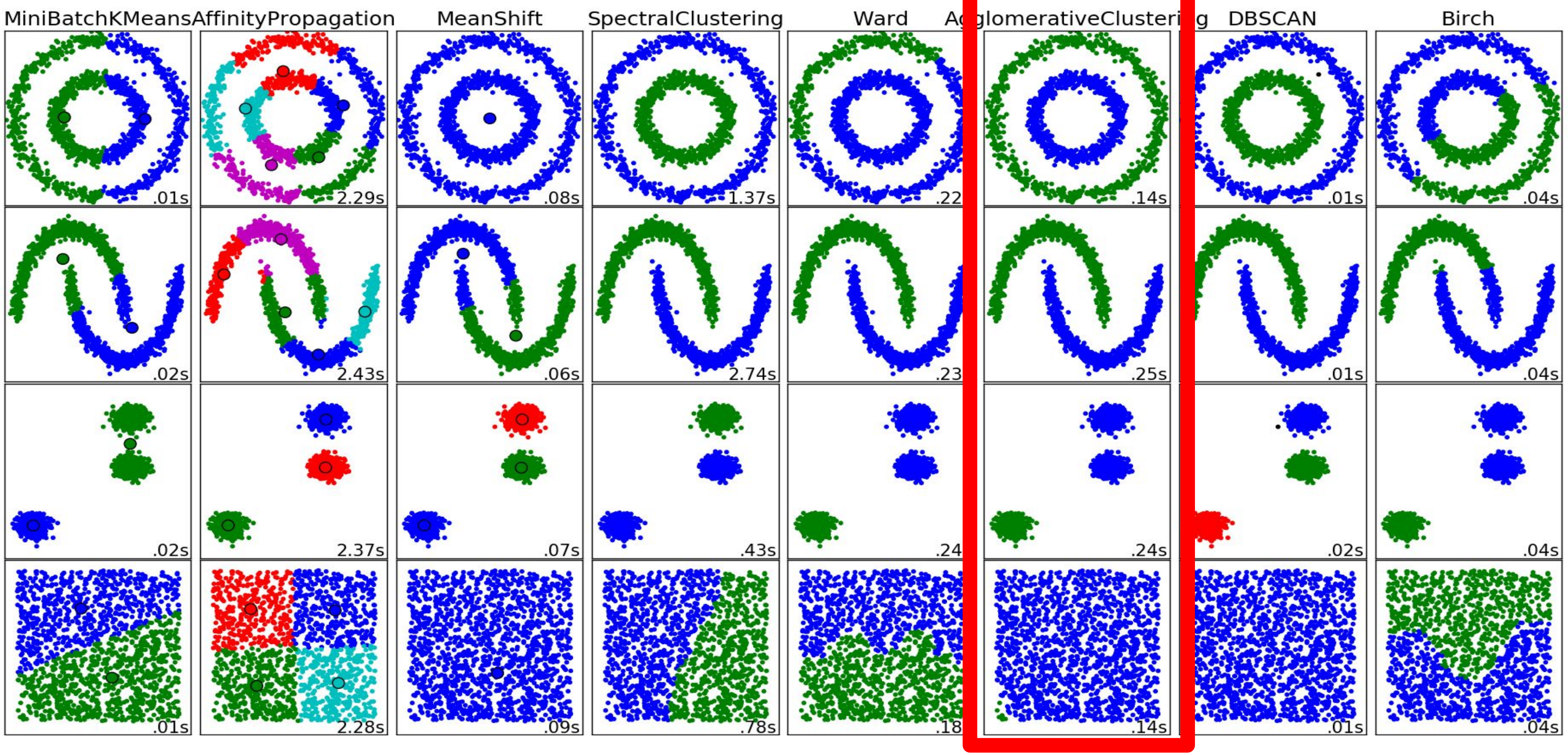
Pros

- Clusters are not necessarily globular
- No dependence upon initialisation
- Dendogram shows a good summary

Cons

- Slower than K-means
- Still need to pick a number of clusters
- Assigning a new point is not straightforward
- Sensitive to noise

Comparisons of clustering algorithms



Hands-on session

01-unsupervised_learning.ipynb

DBSCAN

DBSCAN

Clusters = Zones of **high-density**
Two parameters: **min_samples** and **eps**

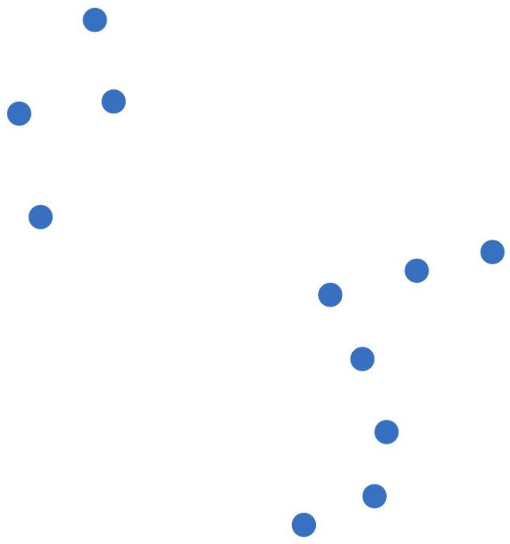
Algorithm:

- Start at a random point, consider all points within radius **eps**
- If that covers **min_samples**, keep that ball
    - Expend by considering esp-balls around every point of the current ball and iterate
- Otherwise mark the point as **noise**

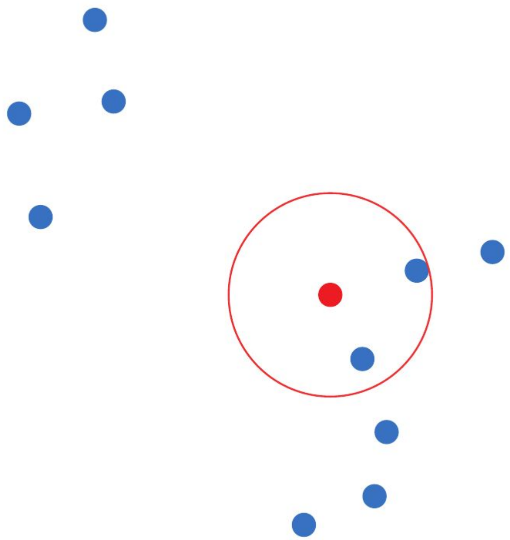
… let's see this in action

CAMBRIDGE SPARK

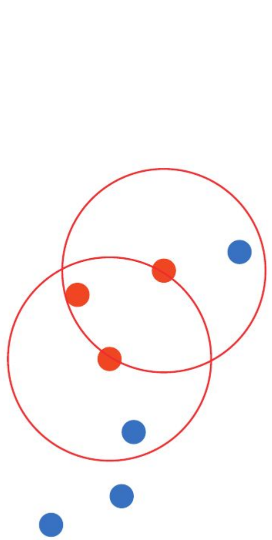
DBSCAN



DBSCAN

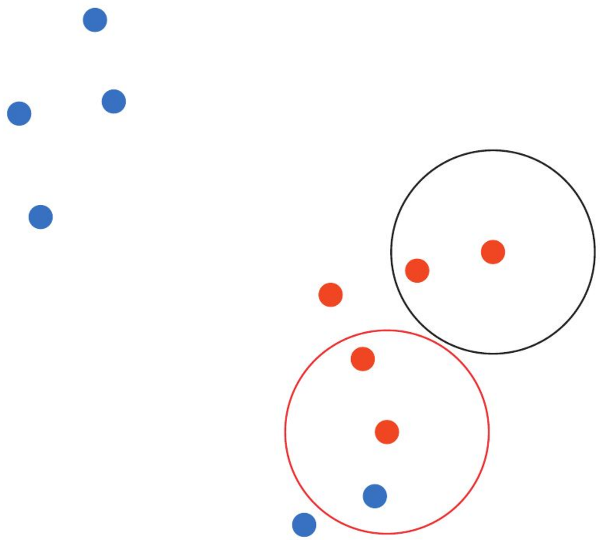


DBSCAN

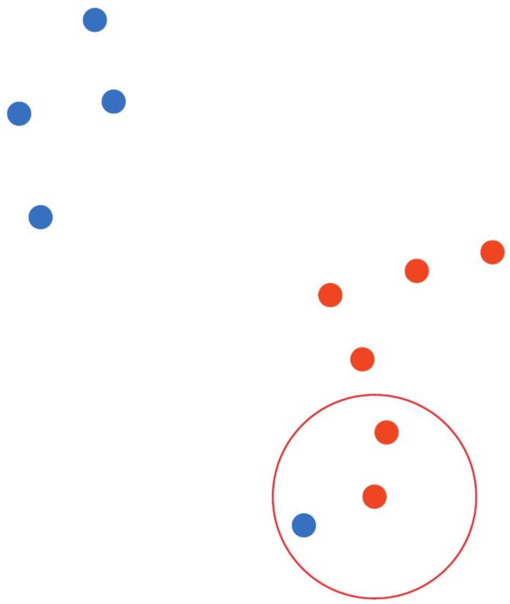


CAMBRIDGE SPARK

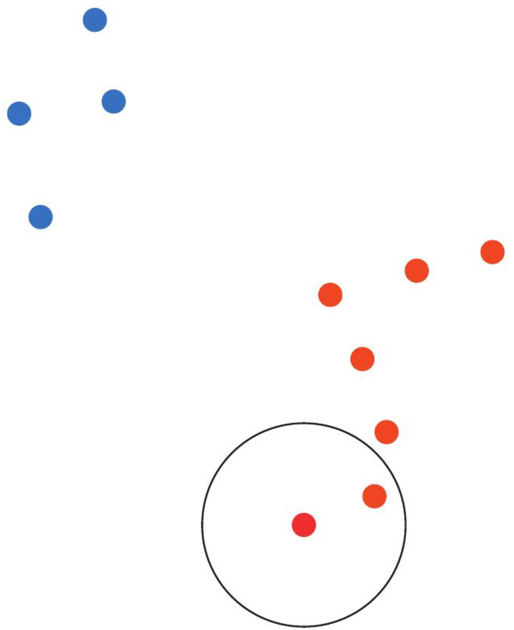
DBSCAN



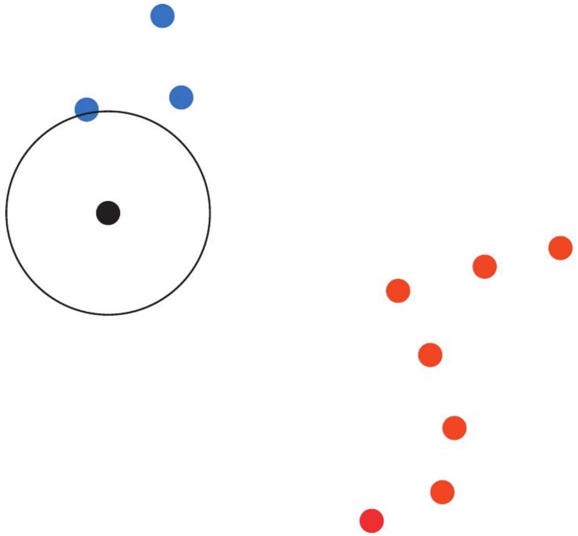
DBSCAN



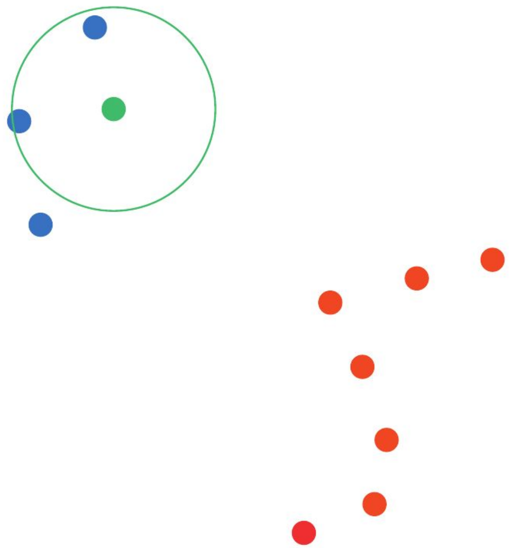
DBSCAN



DBSCAN

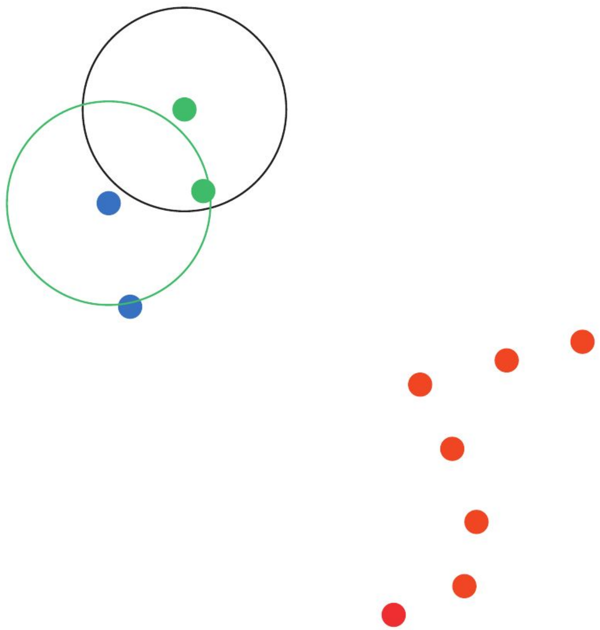


DBSCAN

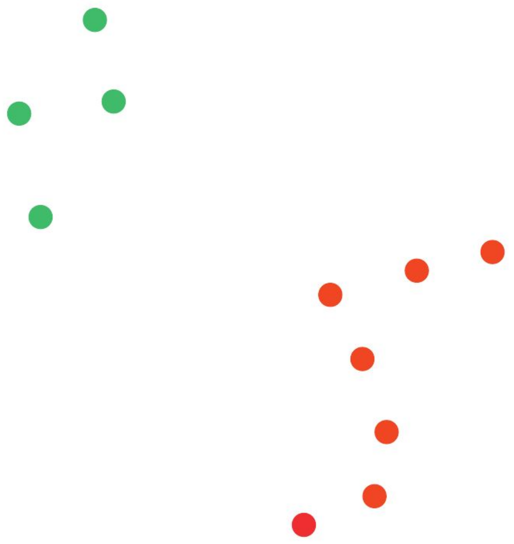


CAMBRIDGE SPARK

DBSCAN



DBSCAN



CAMBRIDGE SPARK

DBSCAN: pros and cons

Pros

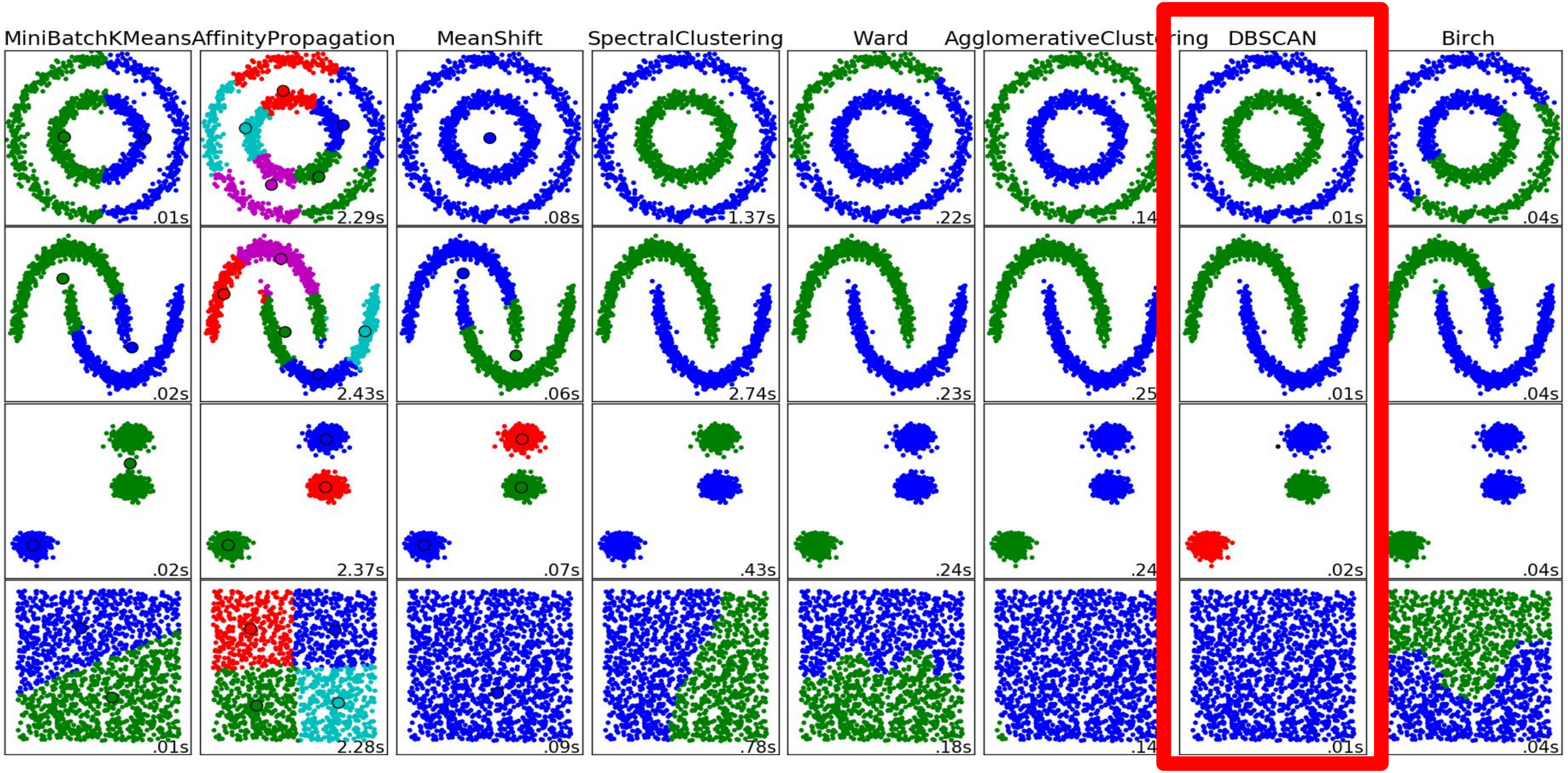
- Clusters are not necessarily globular
- No choice of number of clusters
- Very efficient implementations exist
- Robust to noise

Cons

- The **eps** and **min_samples** can be hard to tune
- If clusters have significantly different densities it is hard to find a meaningful **eps**, **min_samples**

CAMBRIDGE SPARK

Comparisons of clustering algorithms



Hands-on session

01-unsupervised_learning.ipynb