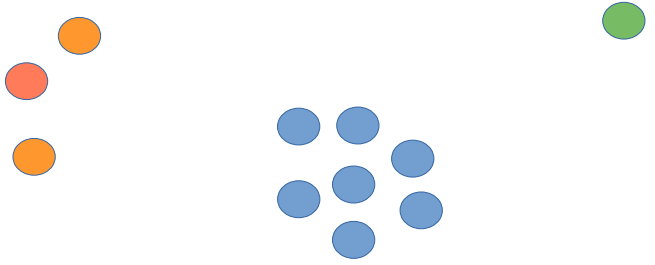
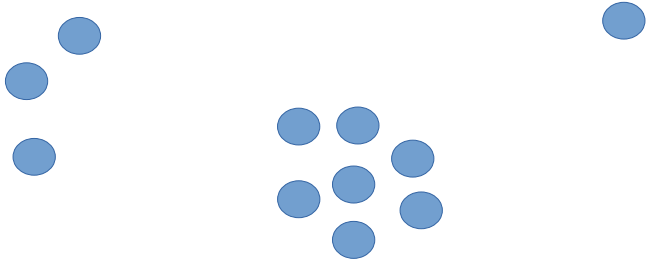
Clustering



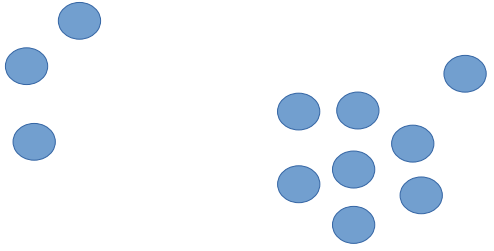
Clustering

Group similar data points together, no labels
"unsupervised"

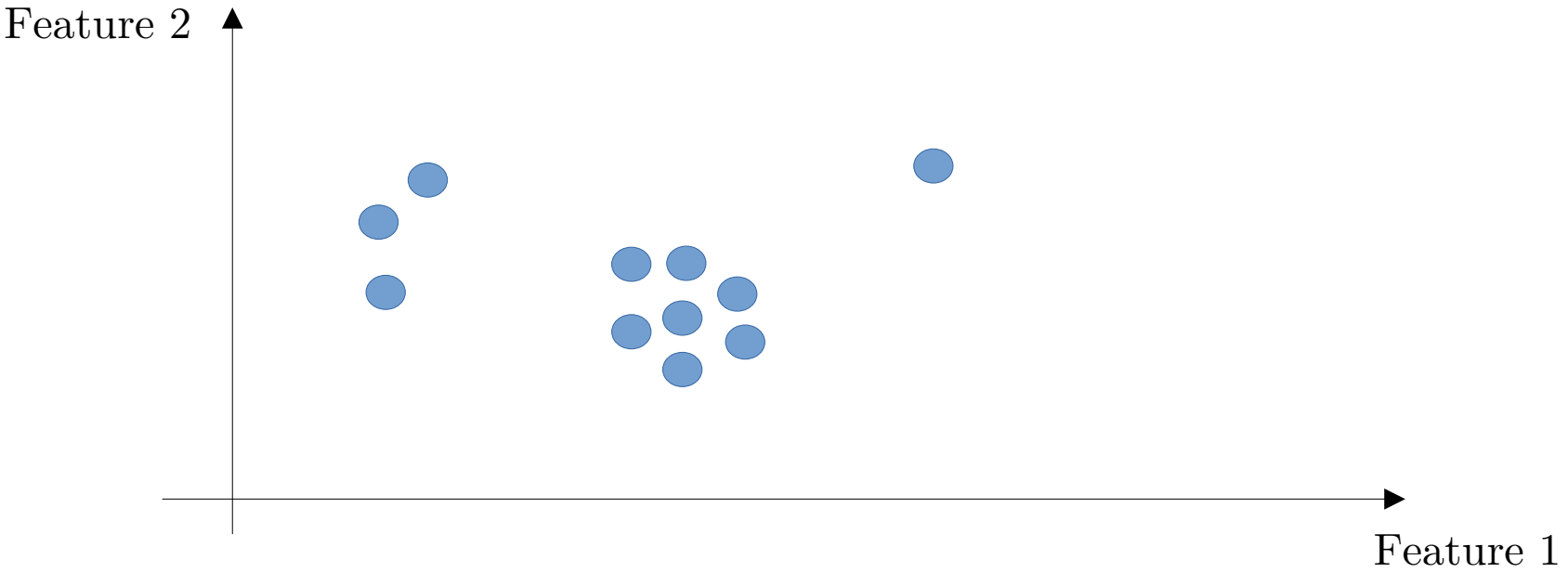


Clustering

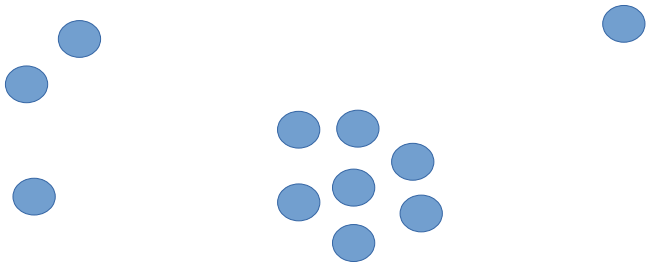
Group similar data points together, no labels
"unsupervised"



What do we need for clustering?



Group similar data points together, no labels
"unsupervised"



In addition to the data, we need some additional information such as:

- A definition of distance or dissimilarity (how close?)
- Minimal inter-cluster distance (Should I merge?)
- Other measures of separability?
- Number of clusters (find the merging distance automatically)
- ...

Different clustering methods

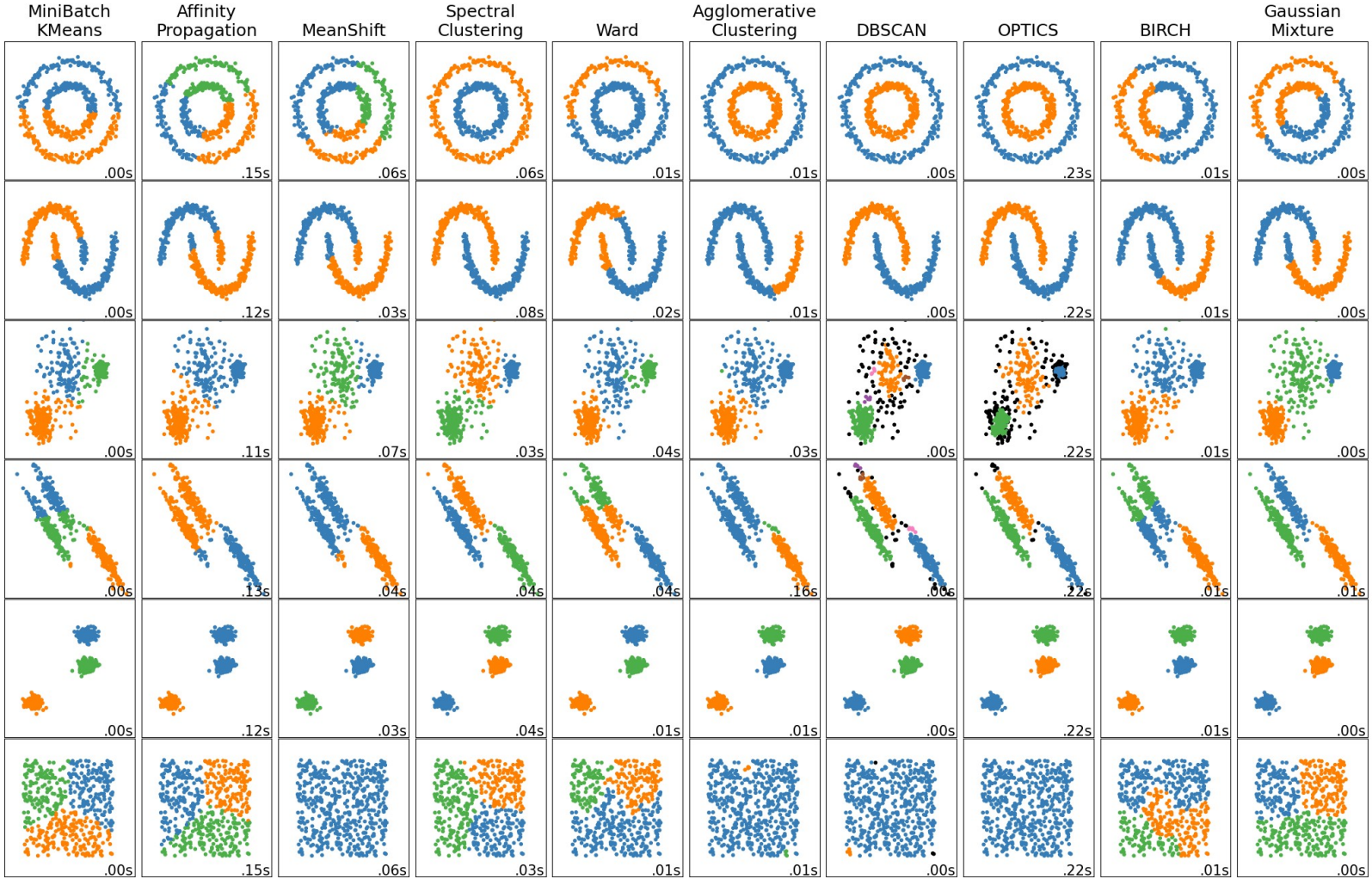
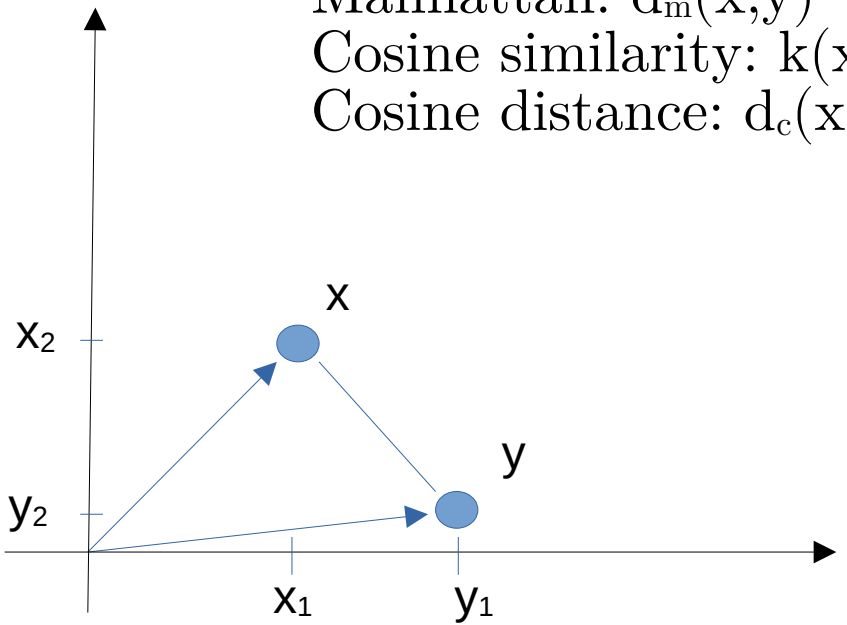


Figure from scikit-learn, clustering

Distance or similarity

Some common distances in dim *n*:
Euclidean: $d_e(x,y)$
Manhattan: $d_m(x,y)$
Cosine similarity: $k(x,y)$
Cosine distance: $d_c(x,y) = 1 - k(x,y)$



$$d_e(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d_m(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

$$k(x,y) = \frac{(x,y)}{\|x\|\,\|y\|} = \cos\theta$$

Sequential clustering

- Datapoints in space
- Distance (or dissimilarity measure) **between a point and a cluster** $d(x_i, C_k)$
- Threshold θ
- Max number of clusters q

Basic Sequential Algorithmic Sequence (BSAS)

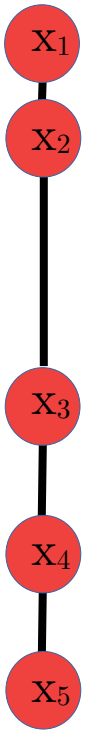
Let m=1
$C_m = x_1$
For i=2 to N,
  Find $C_k : d(x_i, C_k) = \min_{1 \le j \le m} d(x_i, C_j)$
  If $(d(x_i, C_k) > \Theta)$ AND $(m < q)$,
    m=m+1
    $C_m = x_i$
  Else,
    $C_k = C_k \cup x_i$
    Where necessary, update representatives
  end
end

Iterate over the samples

Create a new cluster

Add to closest cluster

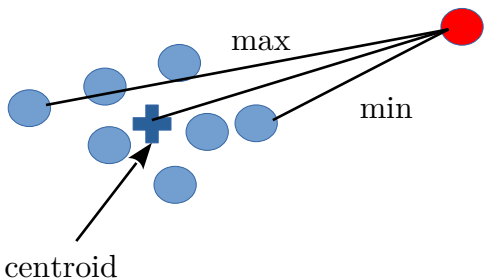
See next slide

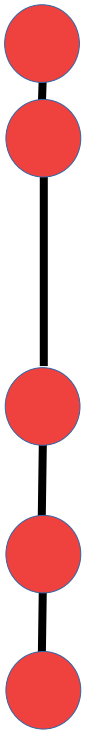


Sequential clustering

Distance between point x_i and cluster C_k: $d(x_i,C_k)$

"update representative": mean point, centroid
Or
Min/max distance between x_i and all the points in C_k

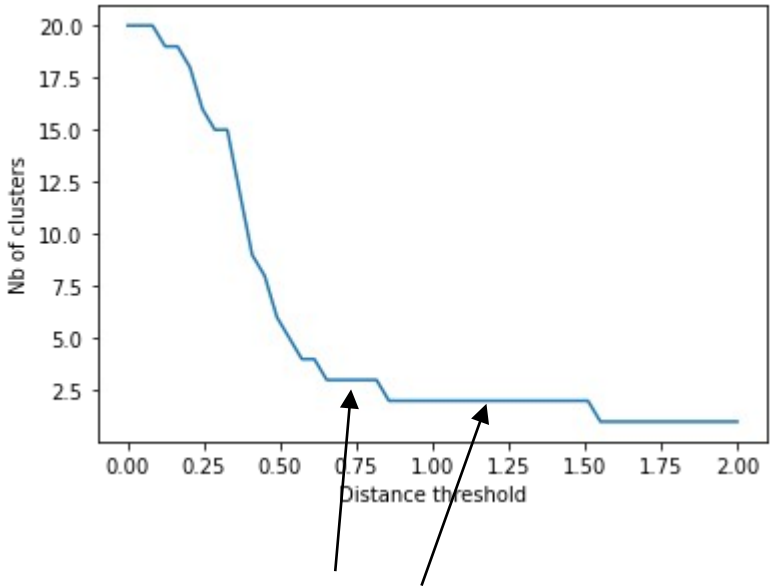




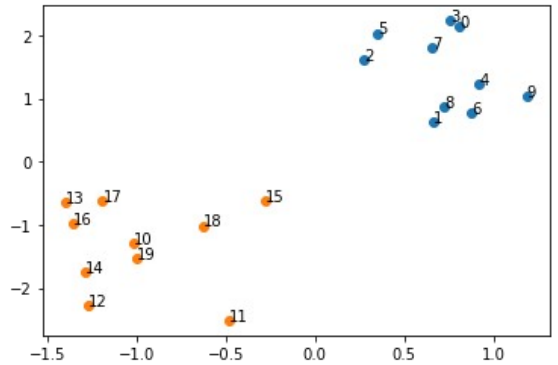
Number of clusters?

Simple way to estimate the best number of clusters

- Run the clustering for different θ



Flat areas where changing the threshold does not affect the nb of clusters → gaps



Select the number associated to the largest flat area

Hierarchical clustering

Keep track of the clustering steps (which sample merged when)
Agglomerative or divisive

https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

We focus on agglomerative clustering (the most popular)
- Divisive have a better global view but requires more computation.

This is an introduction, you will get more details in the course Pattern Recognition FYS-3012

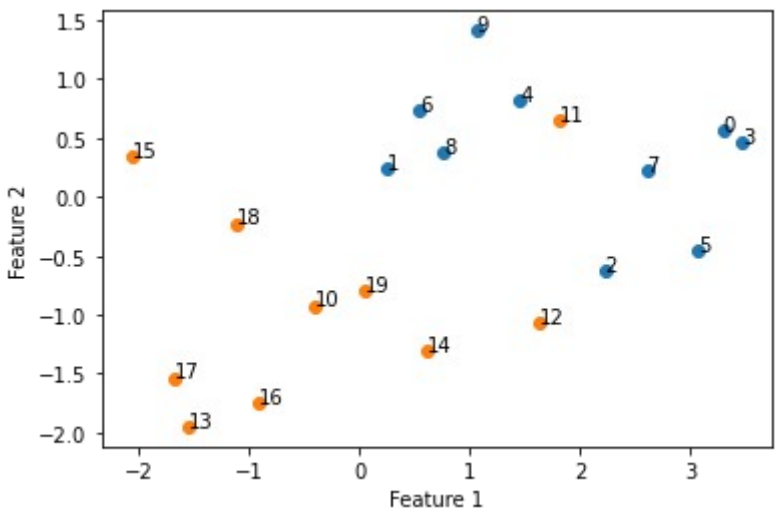
Hierarchical clustering
Dendrogram

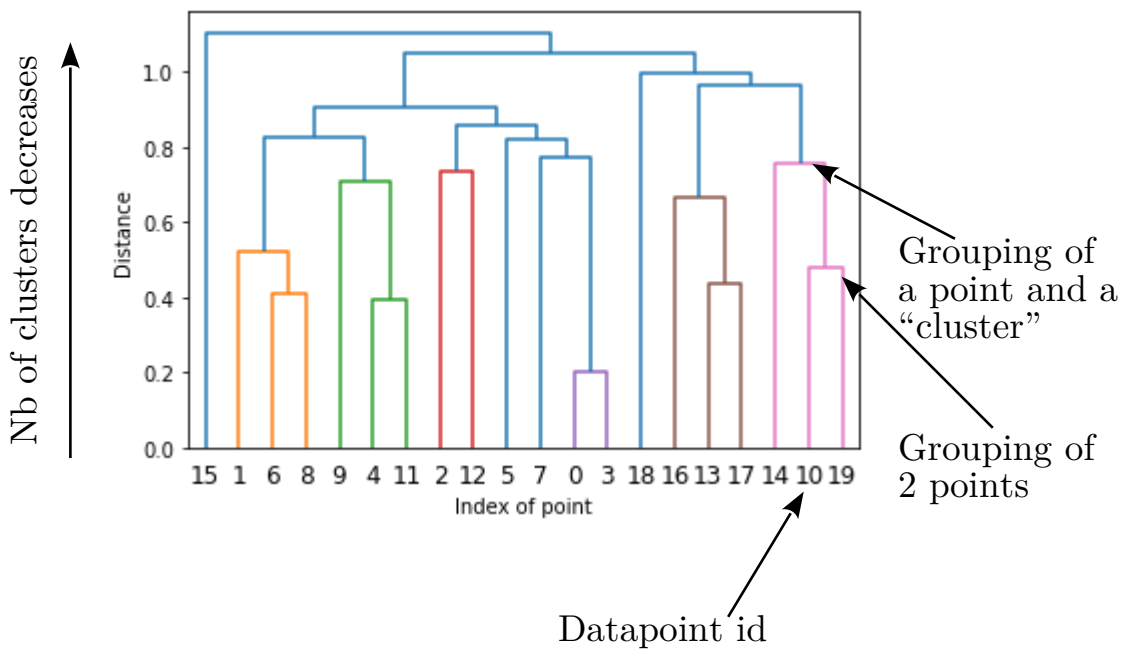
Group the closest datapoints, one by one

Agglomerative scheme:
$R_0 = \{C_i=\{x_i\}, i=1,...,N\}$, each sample is a cluster
$t= 0$
Repeat:
- $t = t+1$
- Find the "closest" 2 clusters C_i and C_j
- Merge them $C_q=C_i \cup C_j$ and update $R_t=(R_{t-1}-\{C_i ,C_j\}) \cup \{C_q\}$
- Until all samples are in a single cluster

Hierarchical clustering
Dendrogram

Group the closest datapoints, one by one



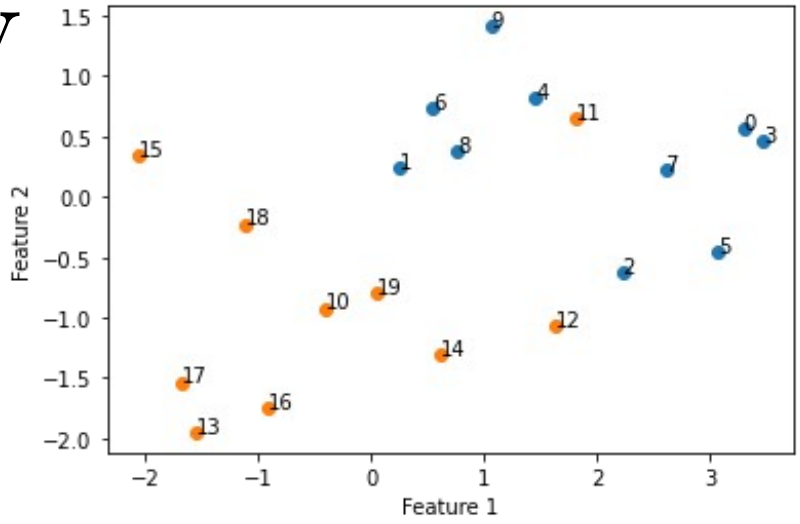


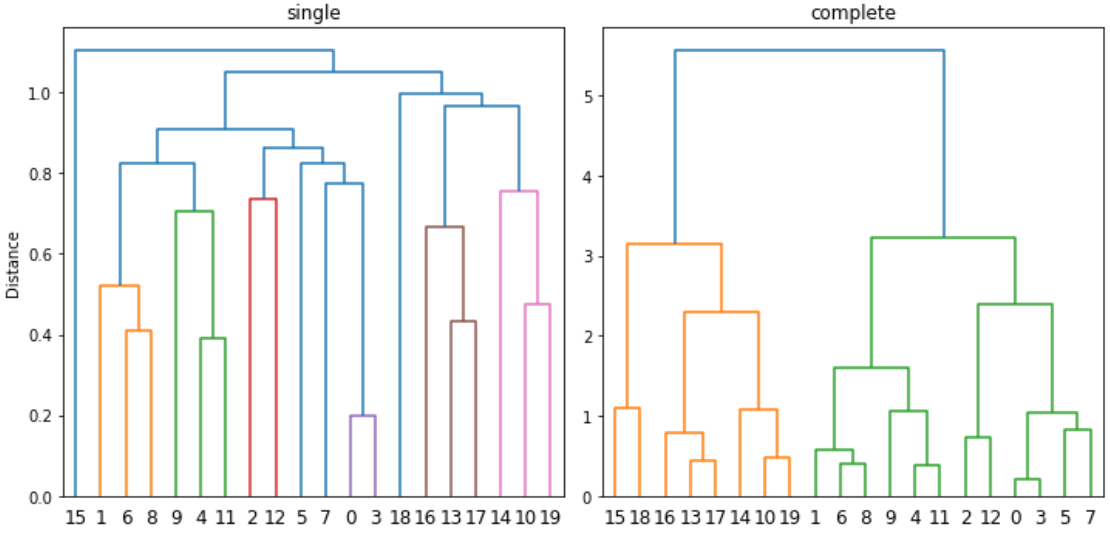
What is the definition of "closest" between a point and a cluster?

Distance or dissimilarity

Cluster C_i and C_j are merged into one (C_i,C_j).
Distance from any C_s to (C_i,C_j) is:

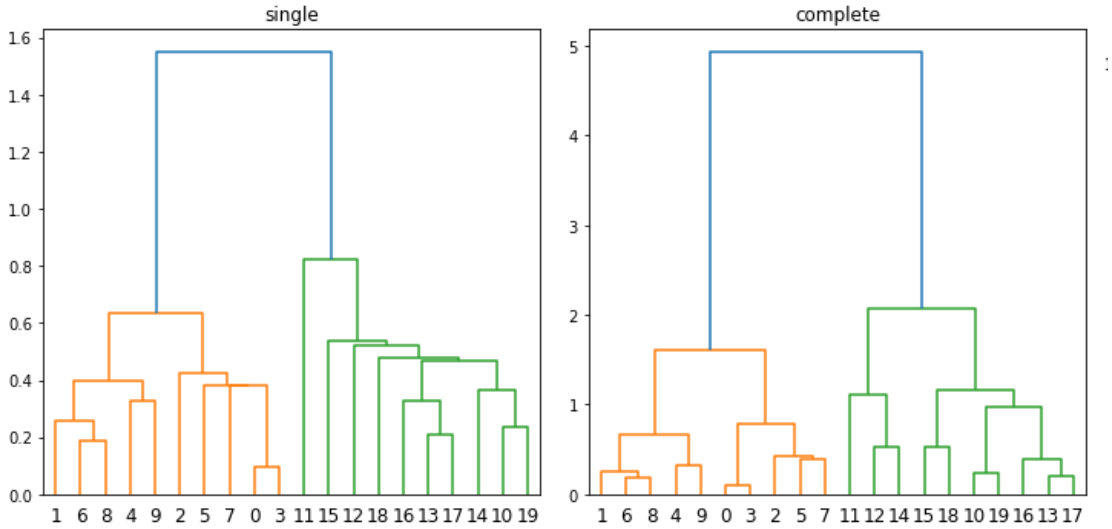
- Single linkage: $\min(d(C_s,C_i), d(C_s,C_j))$
- Average: $[d(C_s,C_i)+d(C_s,C_j)]/2$
- Complete linkage: $\max(d(C_s,C_i), d(C_s,C_j))$

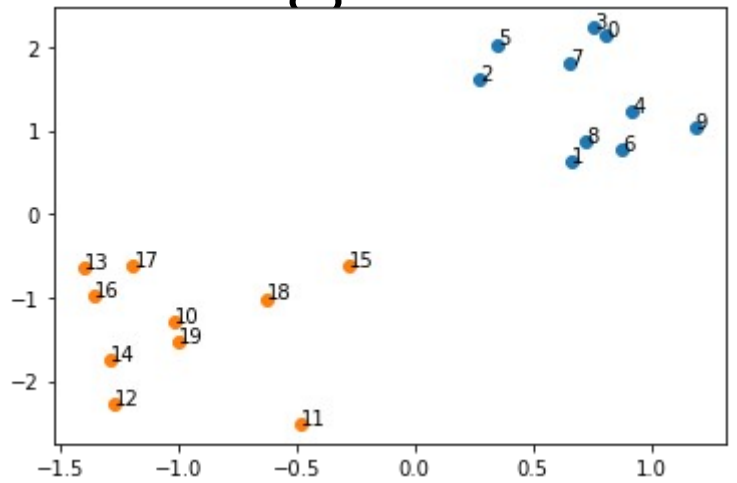


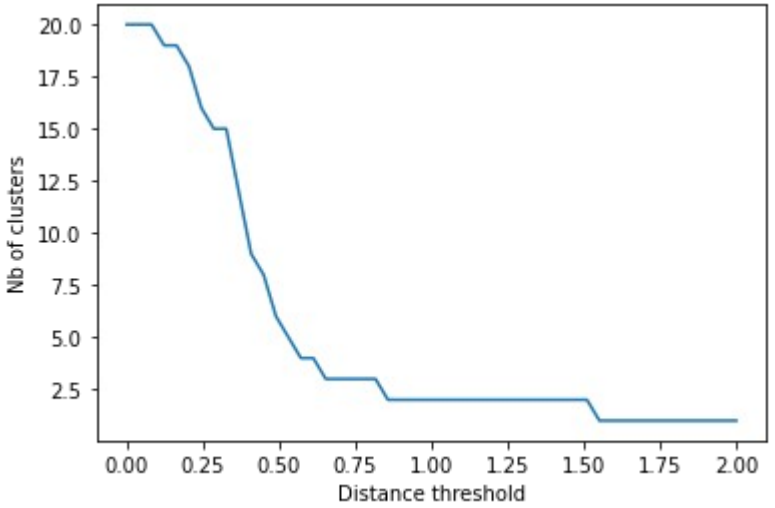


Hierarchical clustering

Best number of clusters?







Conclusion

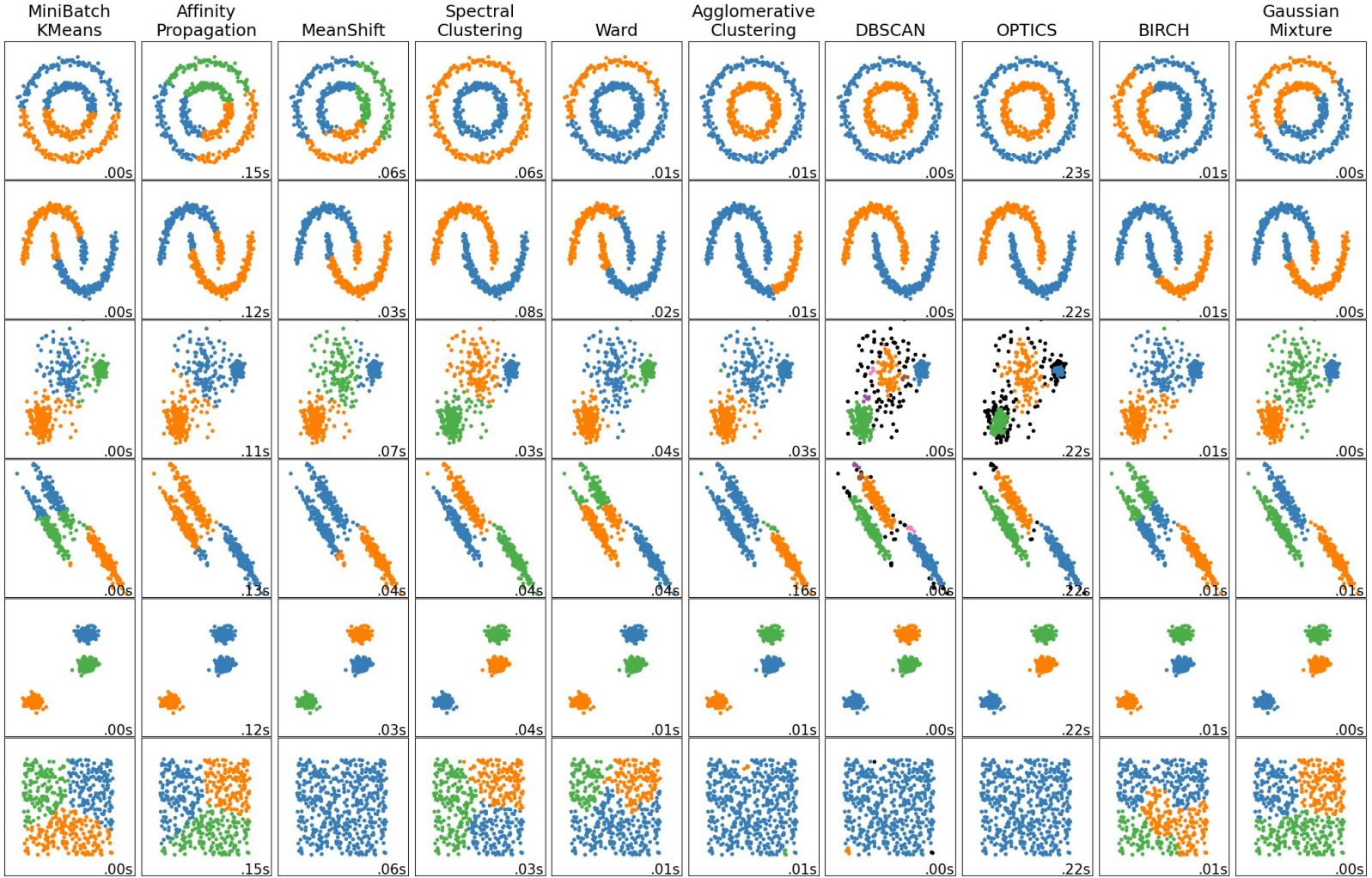


Figure from scikit-learn, clustering