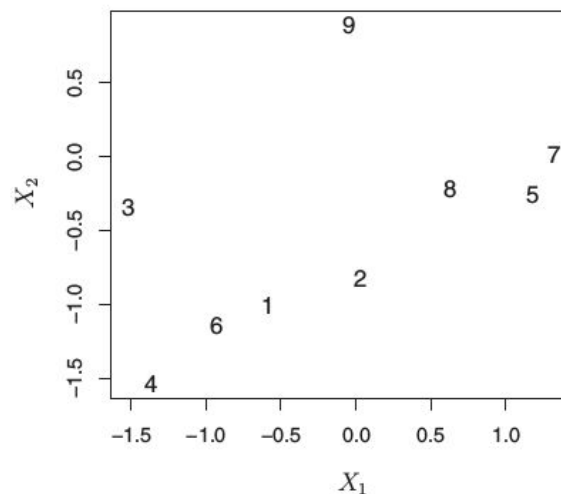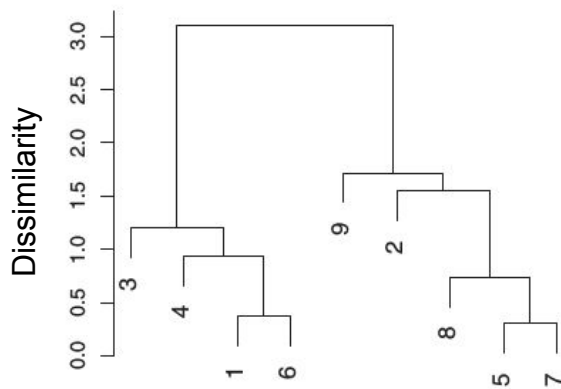# Hierarchical Clustering

Hierarchical clustering

- Another clustering method (creating groups through *hierarchies*)
- Don't have to commit to a value of k beforehand
- Results don't depend on initialization
- Not limited to euclidean distance as the similarity metric
- Easy visualized through dendrograms
 - "Height of fusion" on dendrogram quantifies the separation of clusters

Hierarchical clustering algorithm

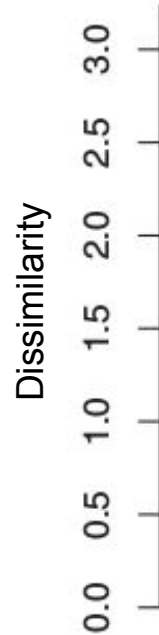**Algorithm 10.2** *Hierarchical Clustering* *(Agglomerative, bottom up)*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed. **In units of your dissimilarity metric**

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i - 1$ remaining clusters.

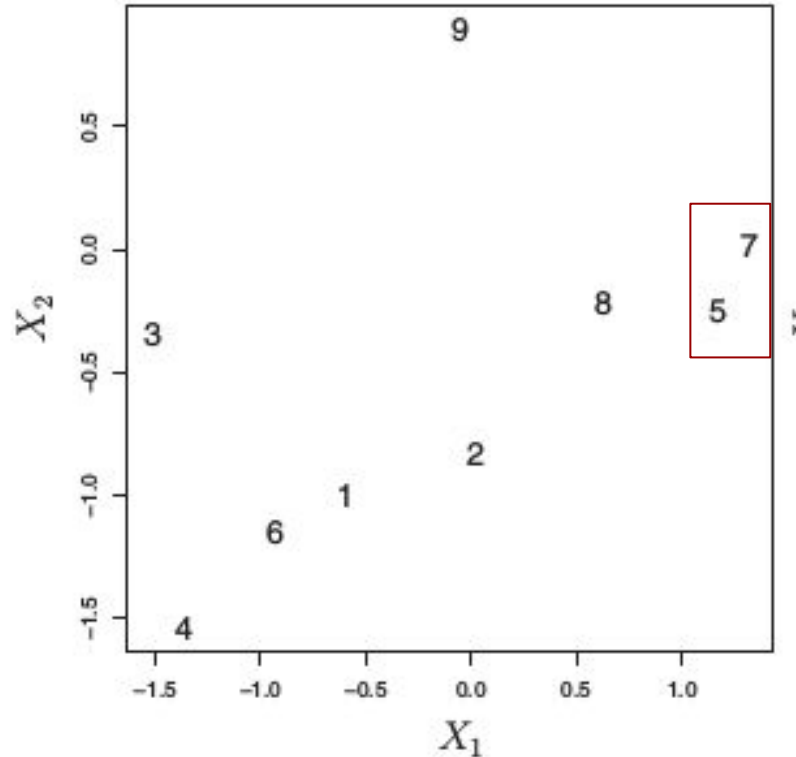# Hierarchical clustering algorithm (visualized)
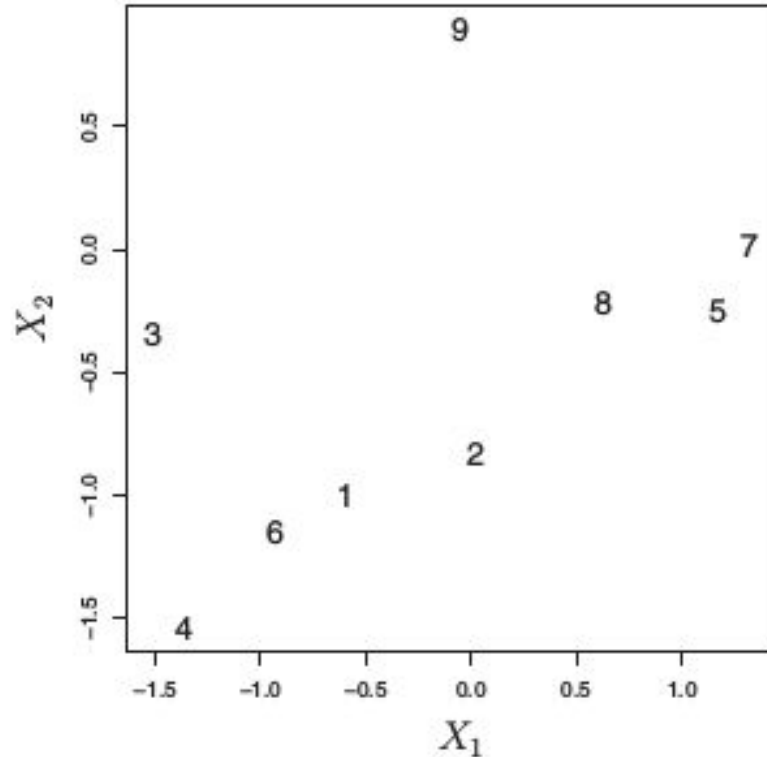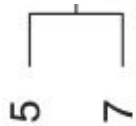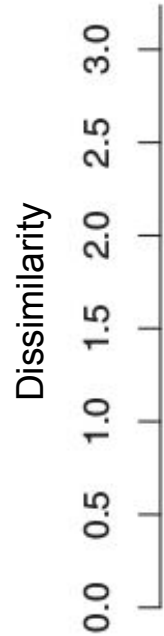
Data

Dendrogram

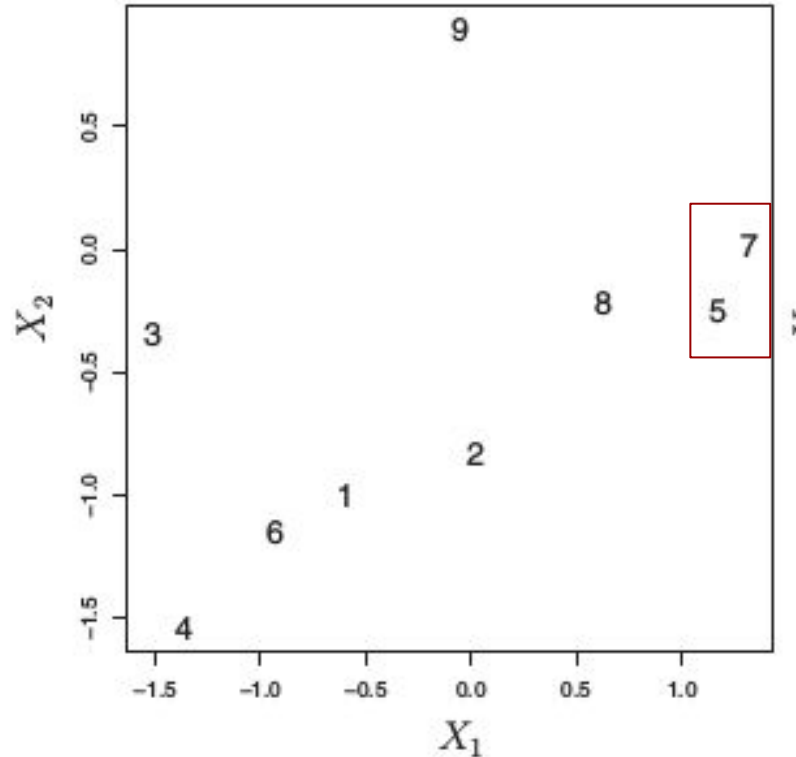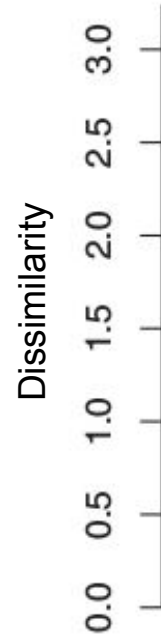# Hierarchical clustering algorithm (visualized)

Data

Dendrogram

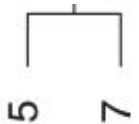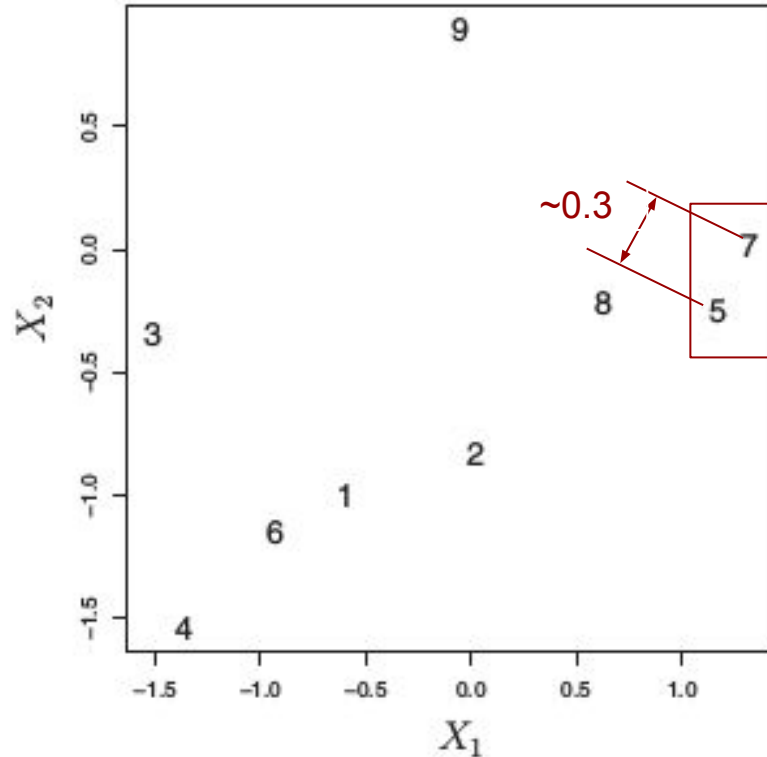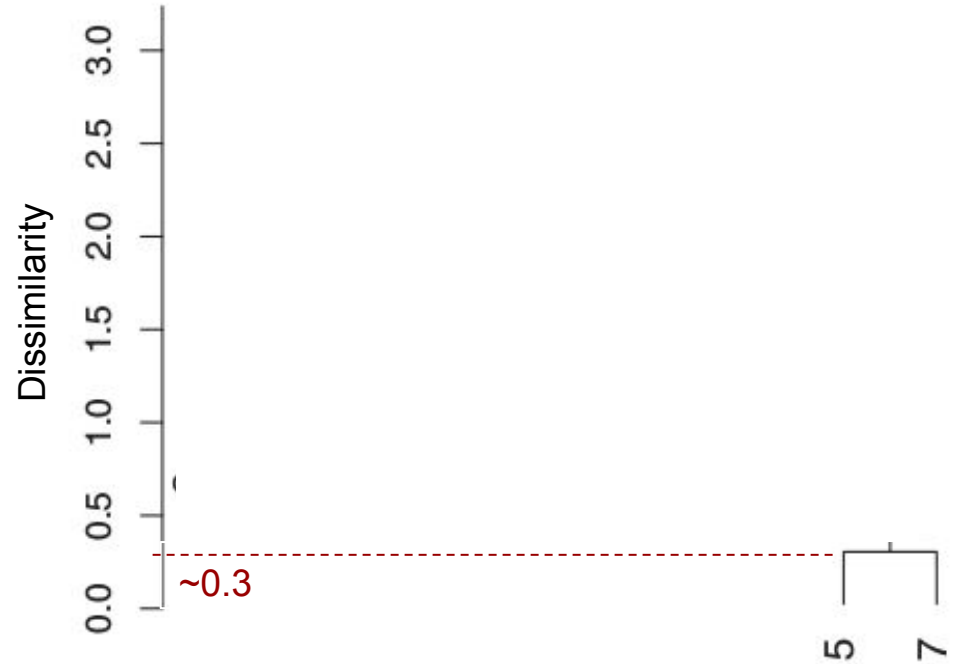# Hierarchical clustering algorithm (visualized)

Data

Dendrogram

Go through the
p_dist_square.ipynb example

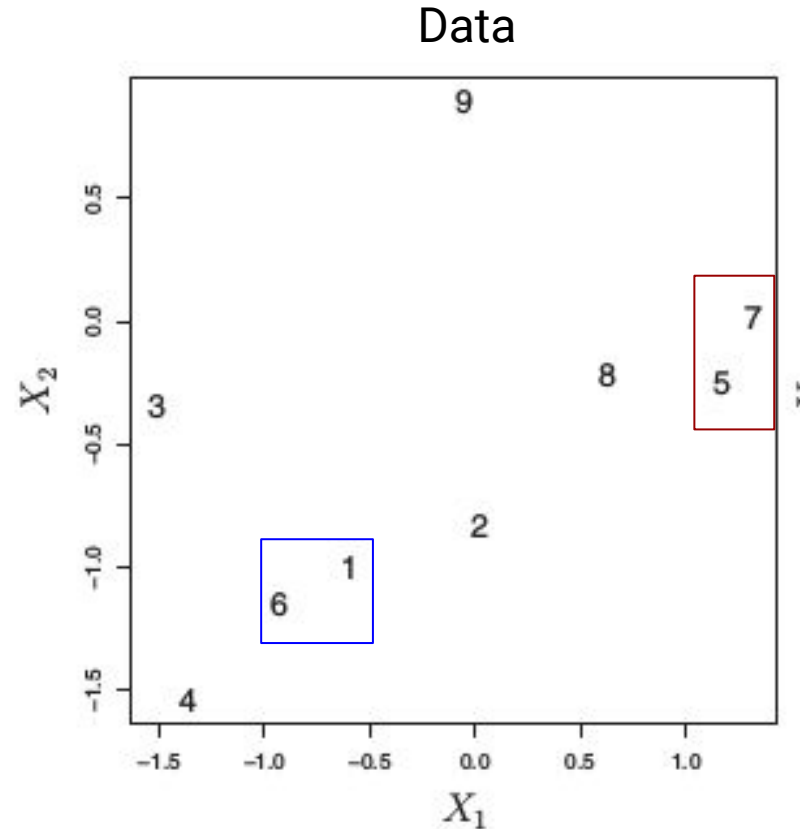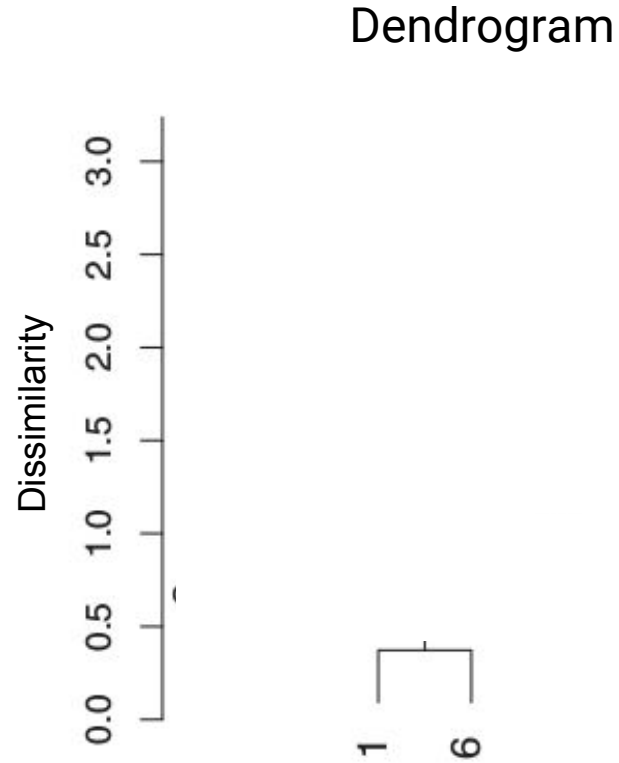# Hierarchical clustering algorithm (visualized)

Data

Dendrogram

Hierarchical clustering algorithm (visualized)

Data

Dendrogram

Hierarchical clustering algorithm (visualized)

Data

Dendrogram
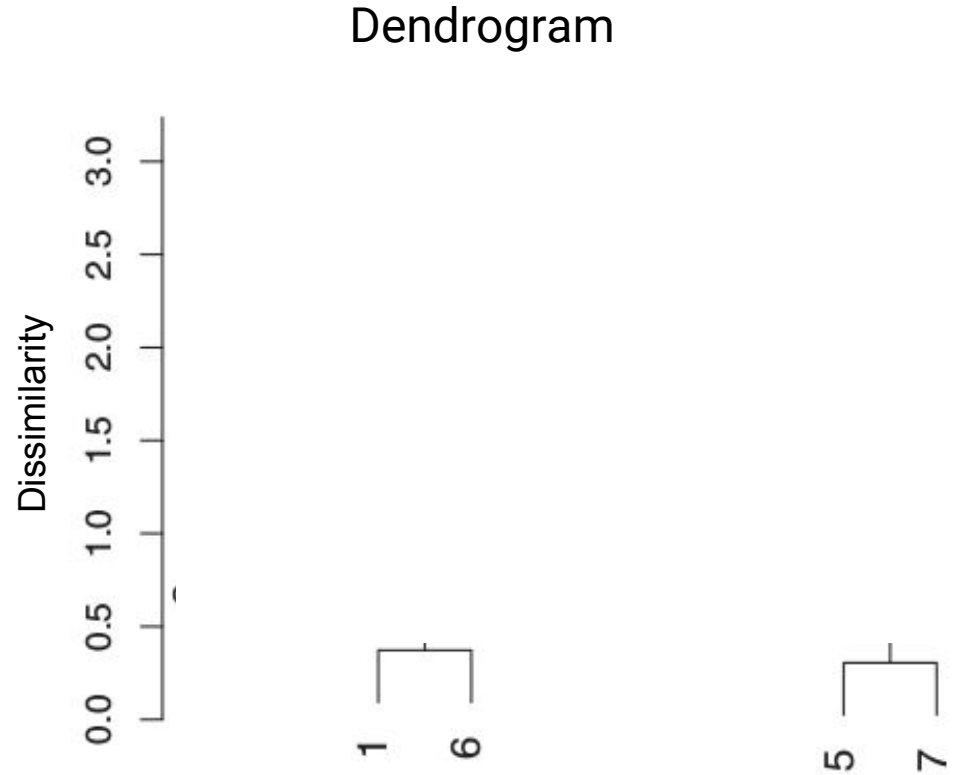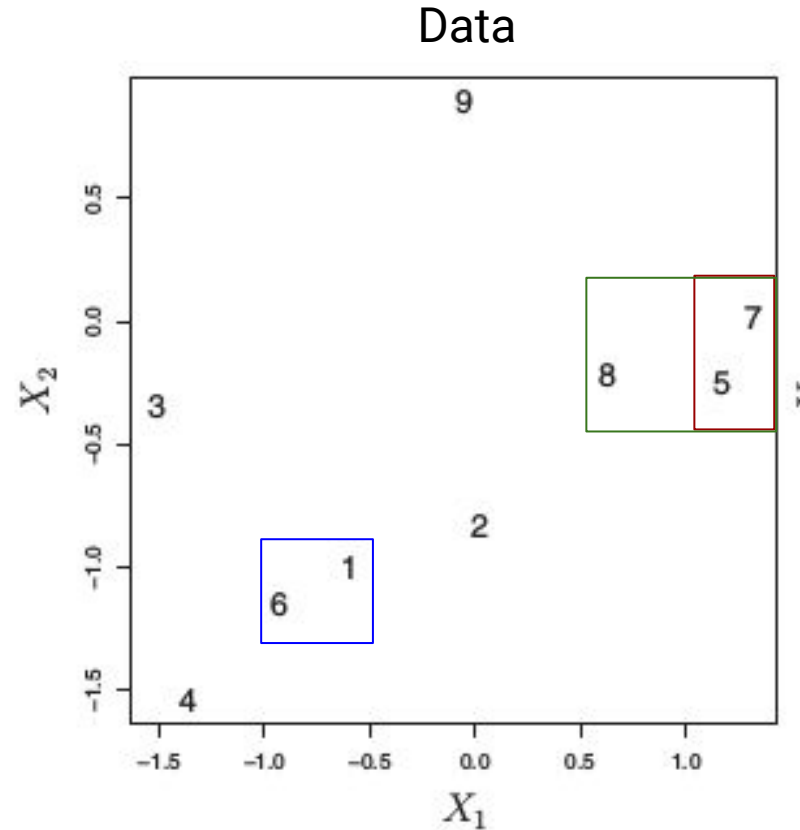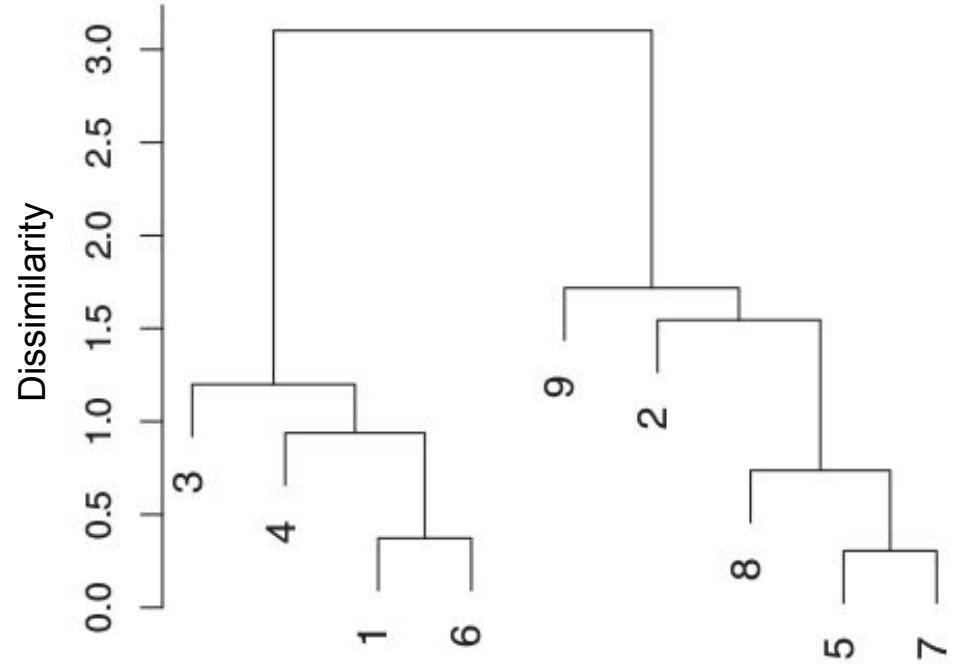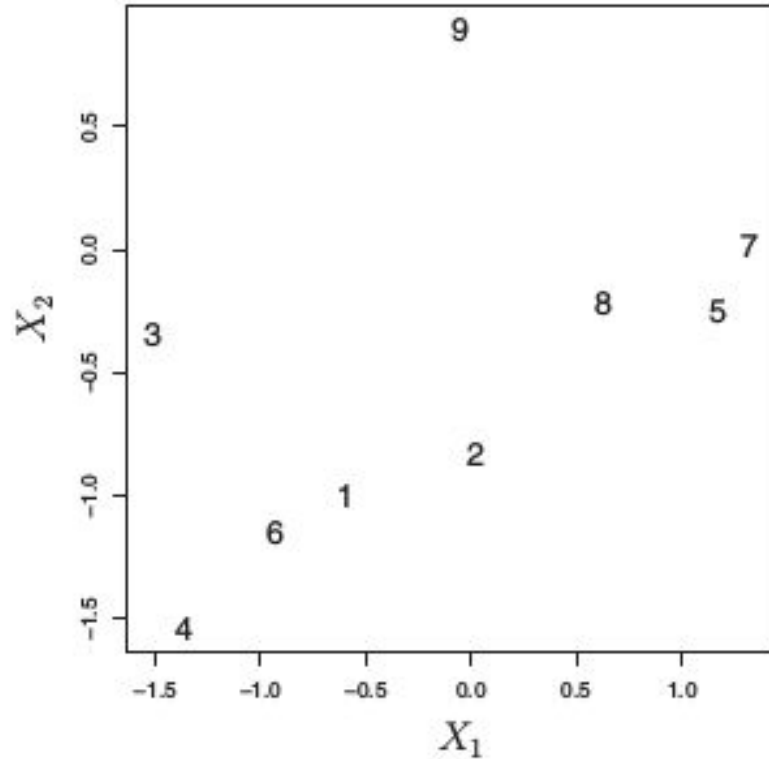
Hierarchical clustering algorithm (on your own)
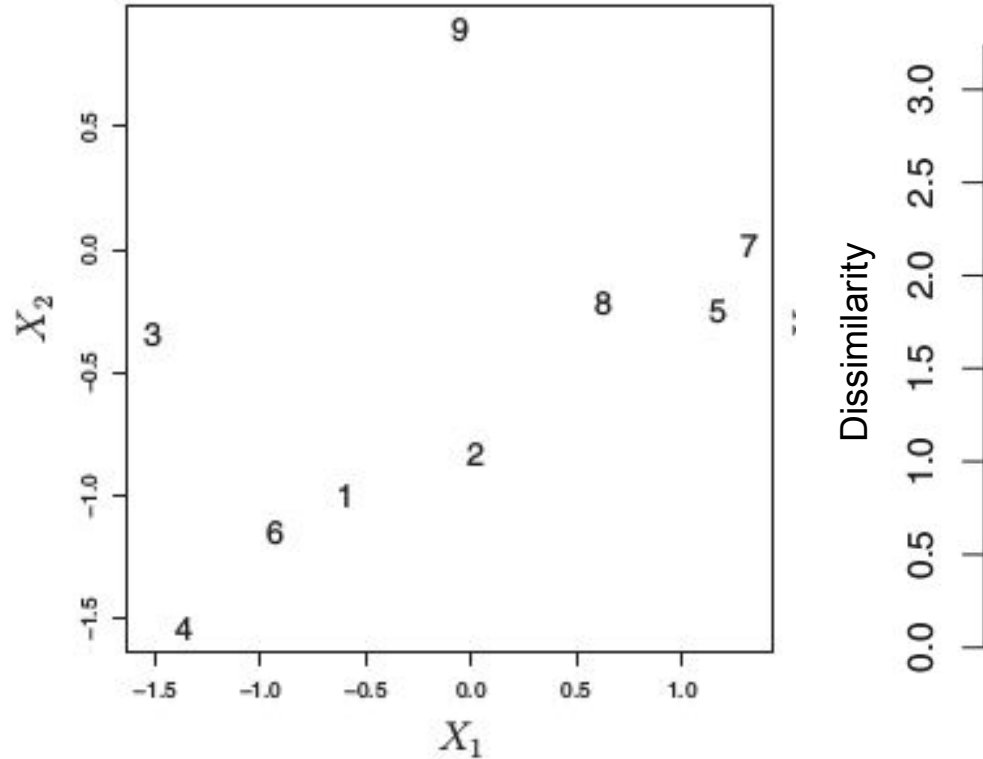
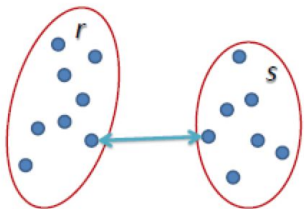# Hierarchical clustering algorithm (on your own)

Measures of (dis)similarity between groups

Single Linkage In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster.

"Nearest neighbor"
Drawback: Chaining - several clusters may be joined to just because of a few close cases
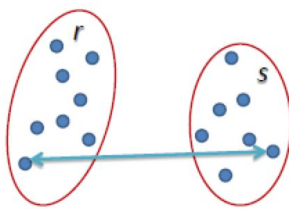
Complete Linkage In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster.
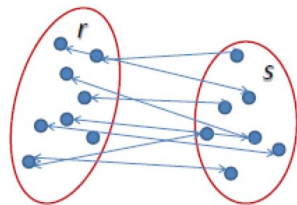
"Farthest neighbor"
Drawback: Cluster outliers prevent otherwise close clusters from merging.

Average Linkage In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

$$L(r,s) = \min(D(x_{r_i}, x_{s_j}))$$

$$L(r,s) = \max(D(x_{r_i}, x_{s_j}))$$

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{r_i}, x_{s_j})$$

Average and complete are most common

http://www.saedsayad.com/clustering_hierarchical.htm

Choice of
linkage matters

(clusters based on same data)

Single
3 clusters
at chosen
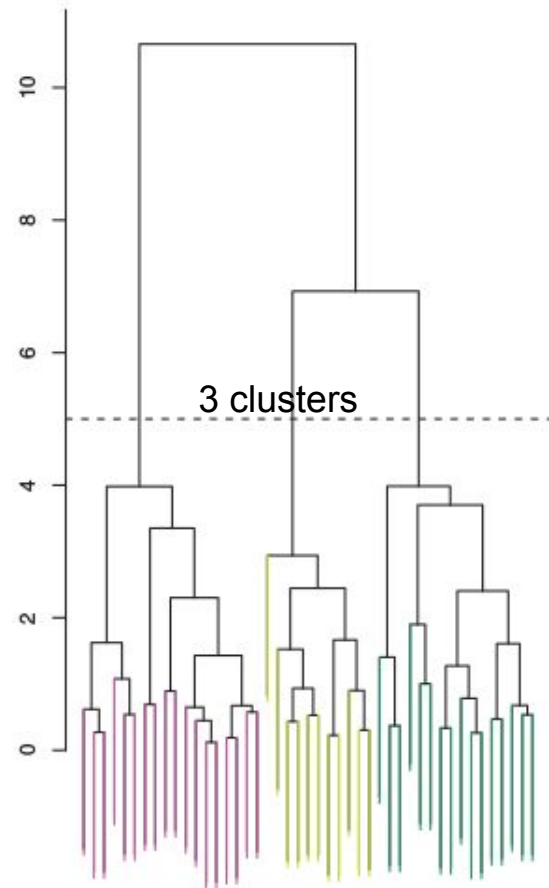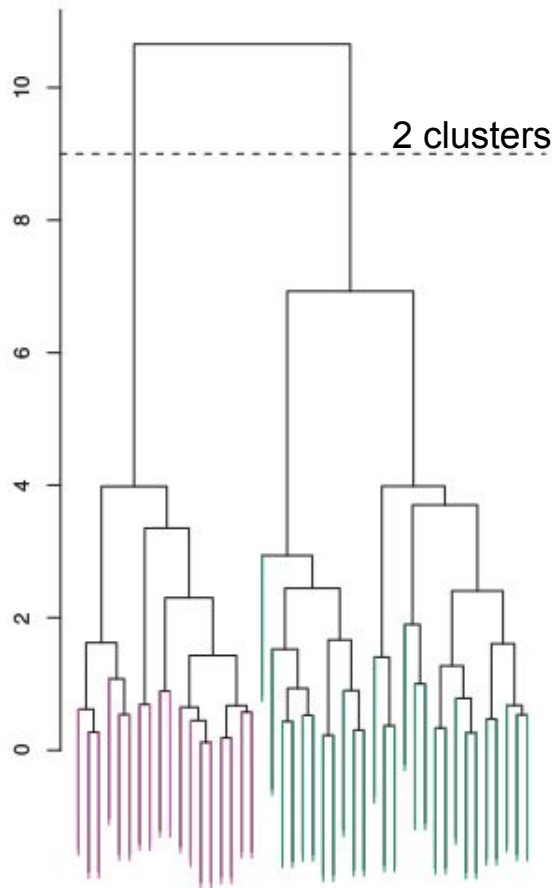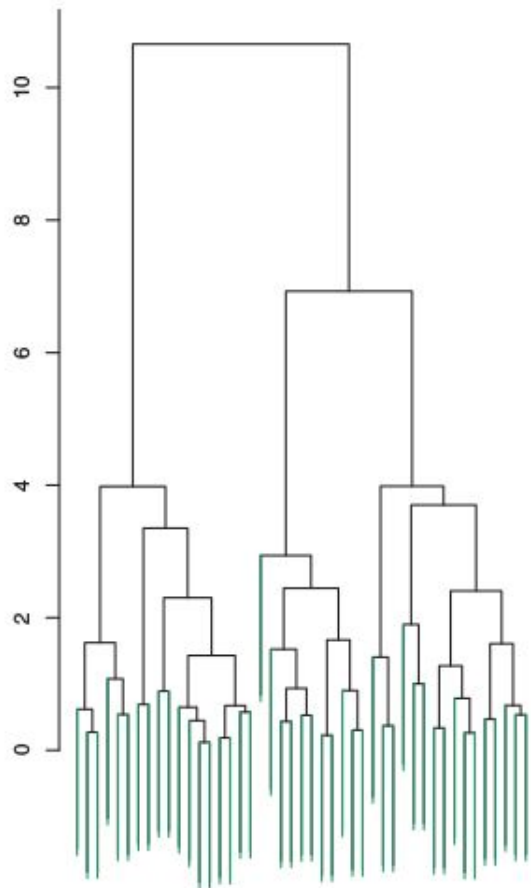distance

Complete
5 clusters
at chosen
distance

Average
4 clusters
at chosen
distance

Chart from :

Yim, O. and K.T. Ramdeen, "Hierarchical Cluster Analysis:
Comparison of Three Linkage Measures and
Application to Psychological Data" The Quantitative Methods for
Psychology, vol. 11, no. 1, 2015.
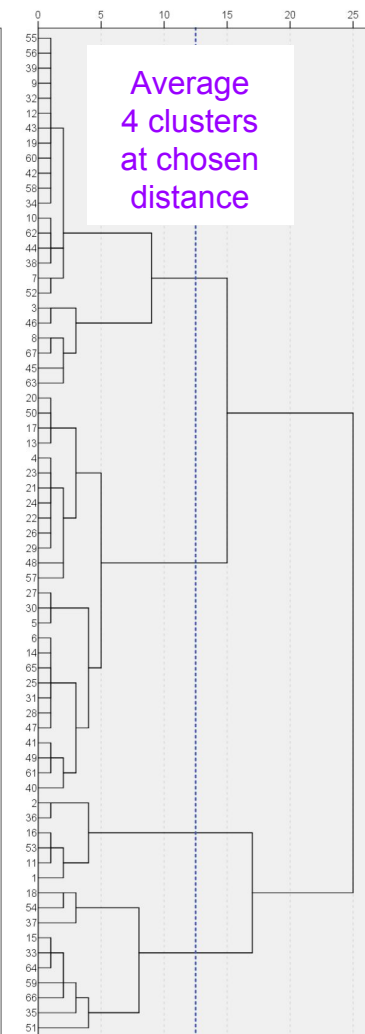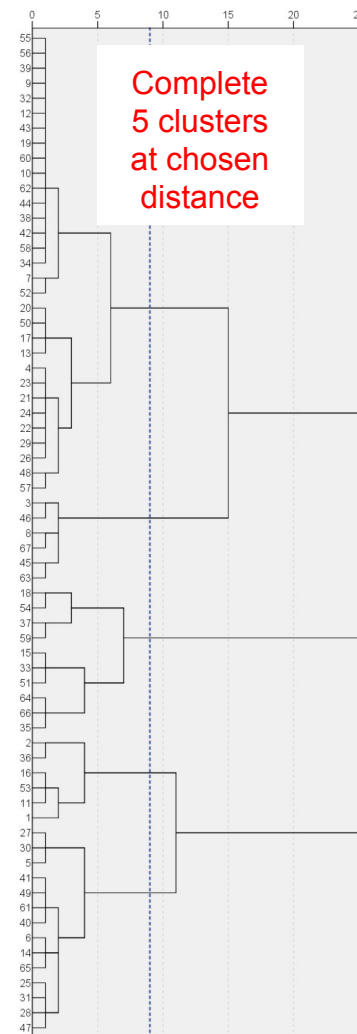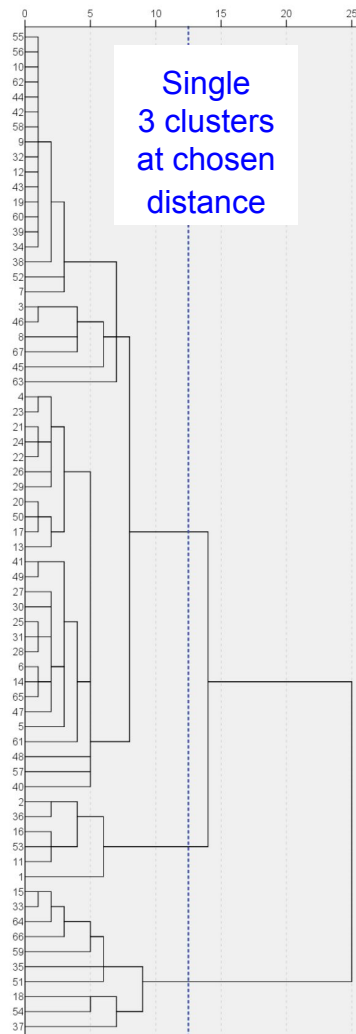
Choice of clusters - where you make the cut

Go through Hierarchical jupyter notebook