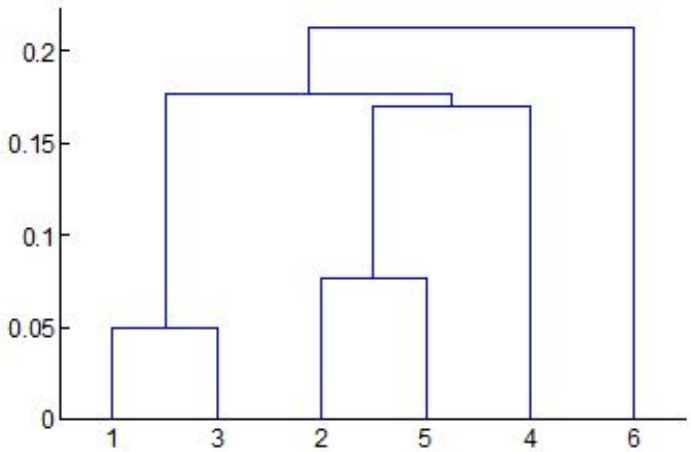
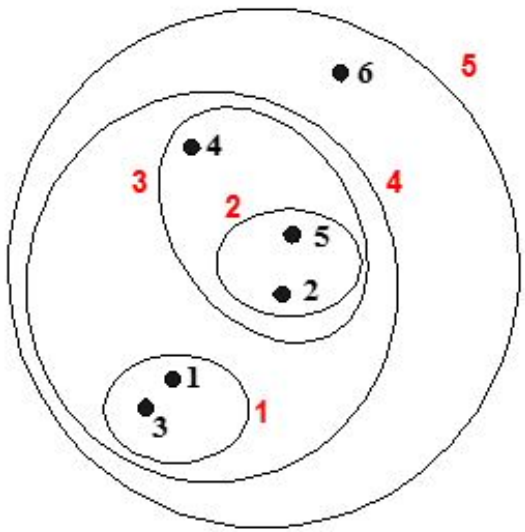
Machine learning

Hierarchical clustering (or hierarchical cluster analysis (HCA) is an alternative approach to partitioning clustering for finding groups of objects based on their similarity.



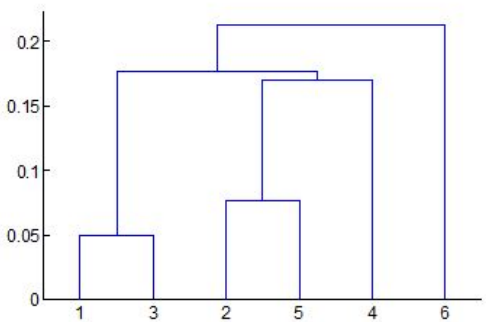


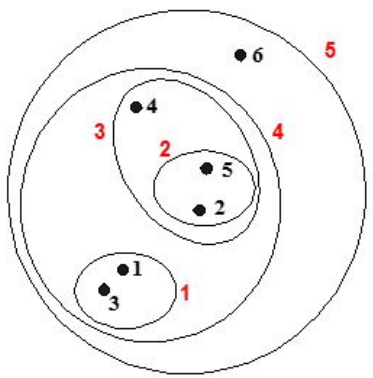
The result of hierarchical clustering is a tree-based representation of the objects, which is also known as **dendrogram**.

Hierarchical clustering

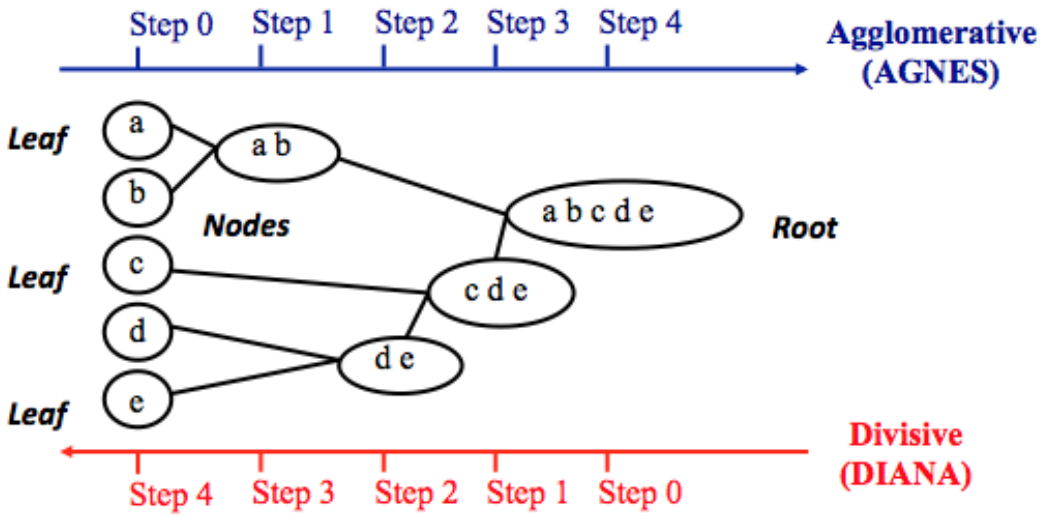
Do not have to assume any particular number of clusters.

Any desired number of clusters can be obtained by cutting the dendrogram at the proper level.





Hierarchical clustering

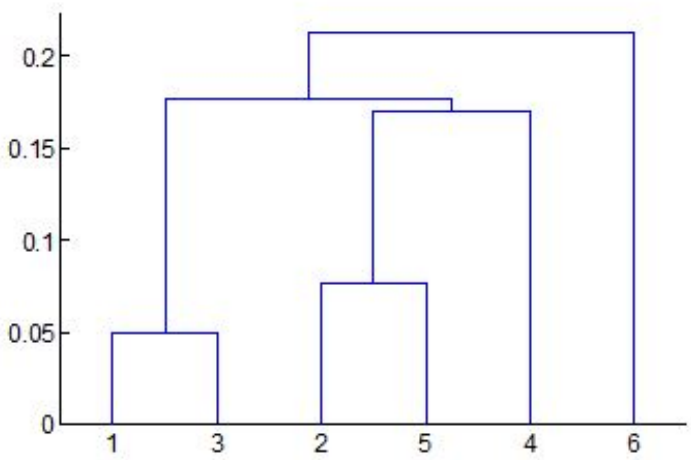


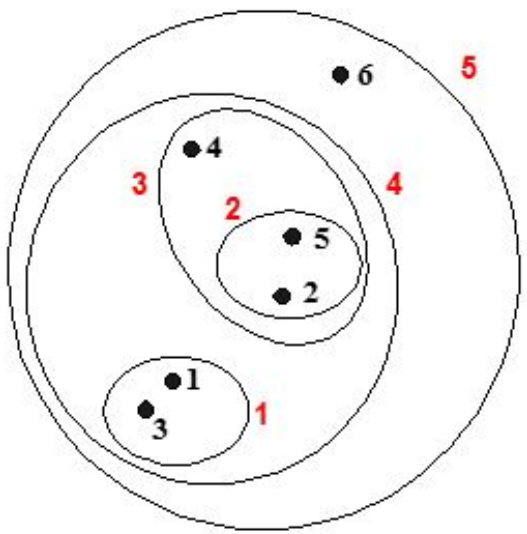
Two main types of hierarchical clustering

- **Agglomerative clustering** - considering each data point as its own cluster and merging them together into larger groups from the bottom up into a single cluster.
- **Divisive clustering** - start with one, all-inclusive cluster. At each step, split a cluster until each cluster contains an individual point.

Agglomerative hierarchical clustering

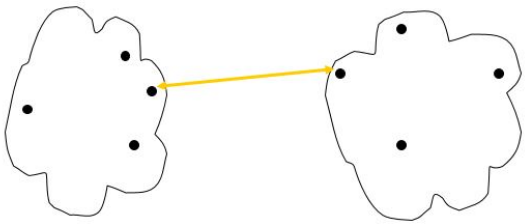
- Preparing the data
- Computing dissimilarity information between every pair of objects in the data set.
- Using linkage function to group objects into hierarchical cluster tree, based on the distance information generated at step 1. Objects/clusters that are in close proximity are linked together using the linkage function.
- Determining where to cut the hierarchical tree into clusters. This creates a partition of the data.

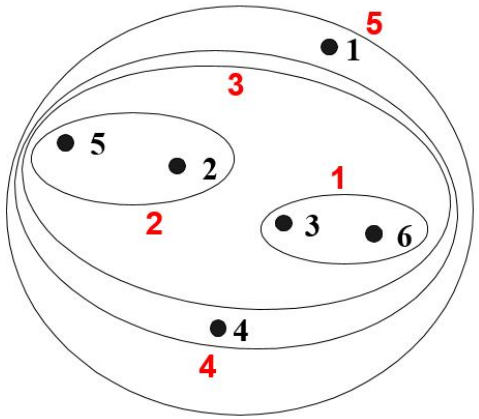


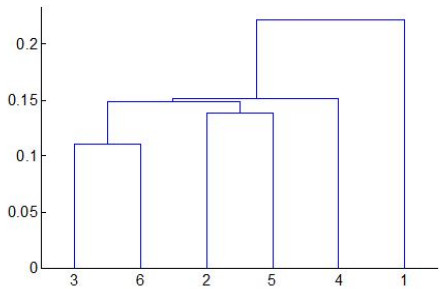


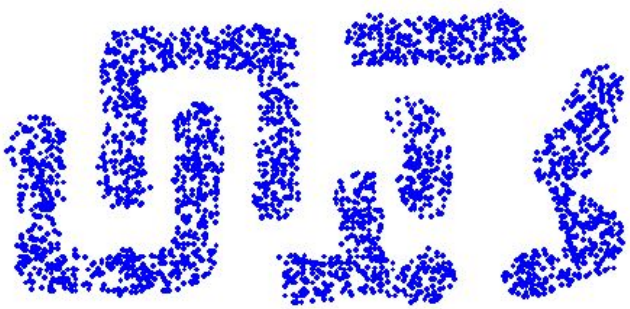
Agglomerative hierarchical clustering

- **minimum or single link** – The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, "loose" clusters.

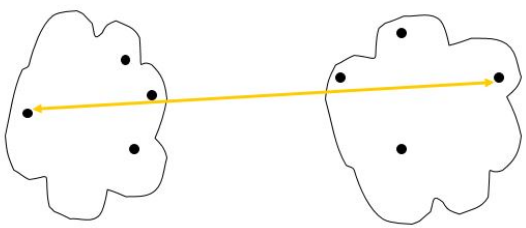


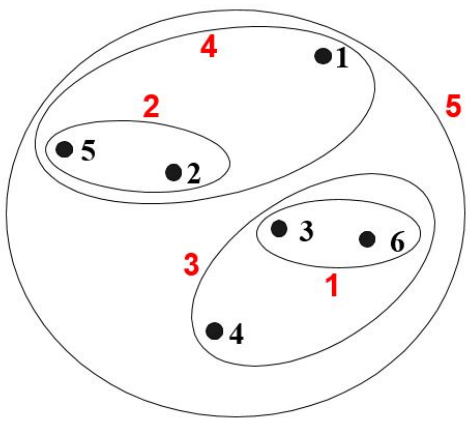


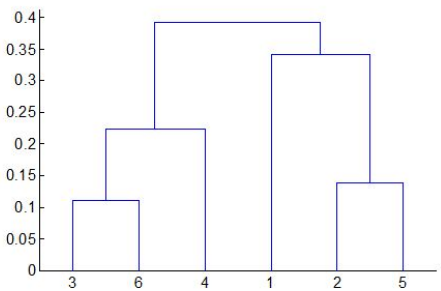




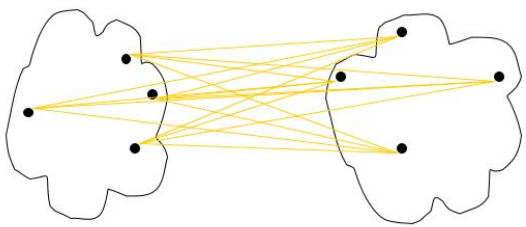
- **maximum or complete linkage** – The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.

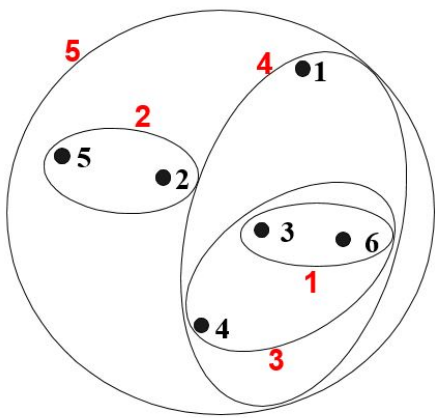


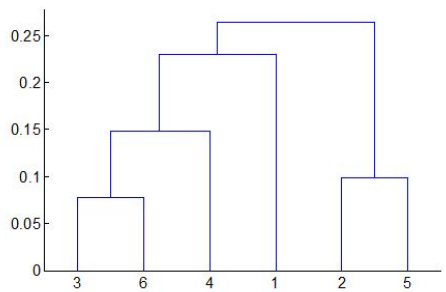




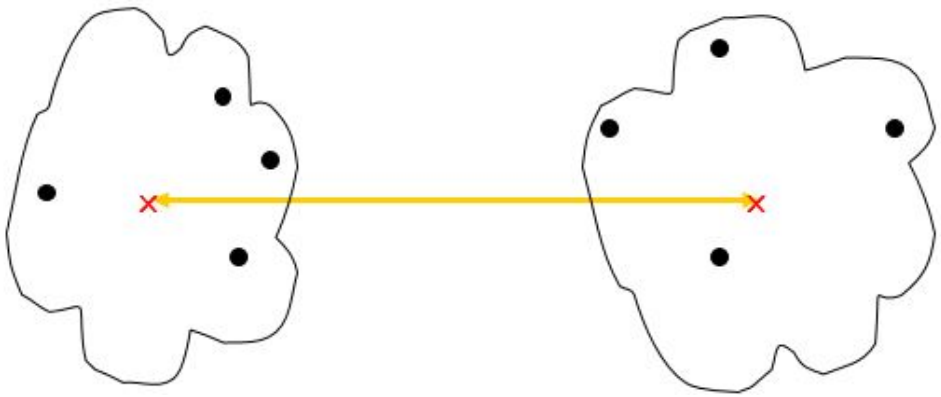
- **Mean or average linkage** – The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2.







- **Centroid linkage** – The distance between two clusters is defined as the distance between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2.



Ward's minimum variance method It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

$$\frac{1}{m_i} \sum_{x \in C_i} (x - c_i)^2 \rightarrow min$$

Assessing Clustering Tendency

Does your dataset contain meaningful clusters? If yes, then how many clusters are there.

A big issue, in cluster analysis, is that clustering methods will return clusters even if the data does not contain any clusters.

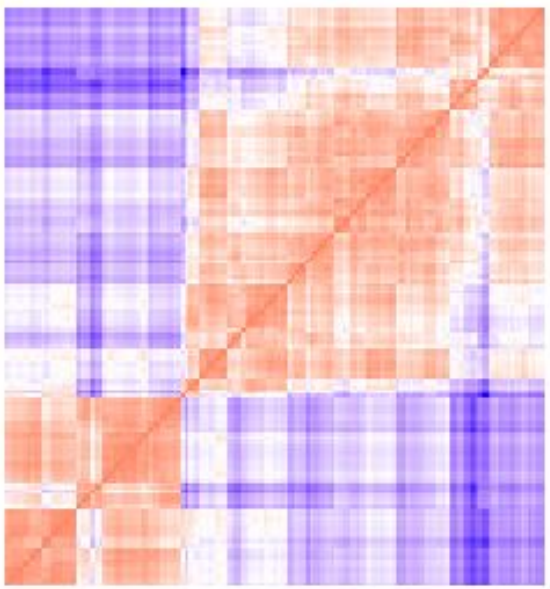
Methods for assessing clustering tendency

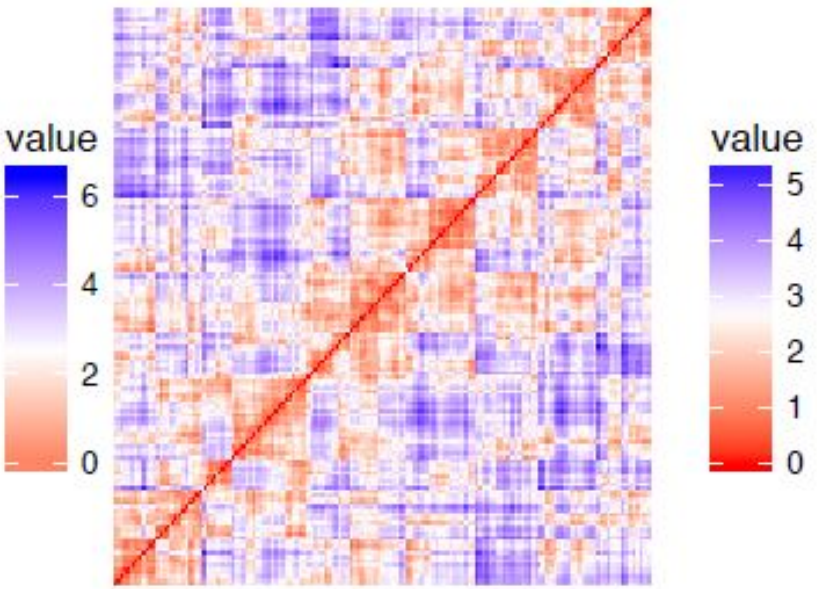
- statistical (Hopkins statistic)
- visual methods

Visual assessment of cluster tendency

The algorithm of the visual assessment of cluster tendency is as follow:

- compute the dissimilarity (DM) matrix between the objects in the dataset
- reorder the DM so that similar objects are close to one another. This process create an ordered dissimilarity matrix (ODM)
- the ODM is displayed as an ordered dissimilarity image (ODI), which is the visual output of algorithm





Hopkins statistic

The Hopkins statistic is used to assess the clustering tendency of a data set by measuring the probability that a given data set is generated by a uniform.

It tests the spatial randomness of the data.

Hopkins statistic

- sample uniformly n points $(p_1, ..., p_n)$ from D.
- for each point $p_i \in D$, find it's nearest neighbor p_j then compute the distance between p_i and p_j and denote it as $x_i = dist(p_i, p_j)$.
- generate a simulated dataset $random_D$ drawn from a random uniform distribution with n points $(q_1, ..., q_n)$ and the same variation as the original real data set D.
- for each point $q_i \in random_D$, find it's nearest neighbor q_j in D then compute the distance between q_i and q_j and denote it $y_i = dist(q_i, q_j)$.
- calculate the Hopkins statistic (H) as the mean nearest neighbor distance in the random dataset divided by the sum of the mean nearest neighbor distances in the real and across the simulated dataset.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

A value of H about 0.5 means that $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} y_i$ are close to each other, and thus the data D is uniformly distributed.

Hopkins statistic

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

A value of H about 0.5 means that $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} y_i$ are close to each other, and thus the data D is uniformly distributed.

The null and the alternative hypotheses are defined as follow

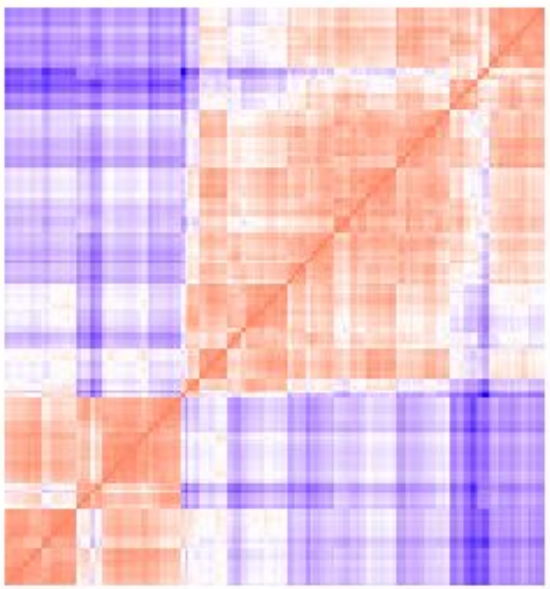
- **null hypothesis**: the dataset D is uniformly distributed (i.e., no meaningful clusters)
- **alternative hypothesis**: the dataset D is not uniformly distributed (i.e., contains meaningful clusters)

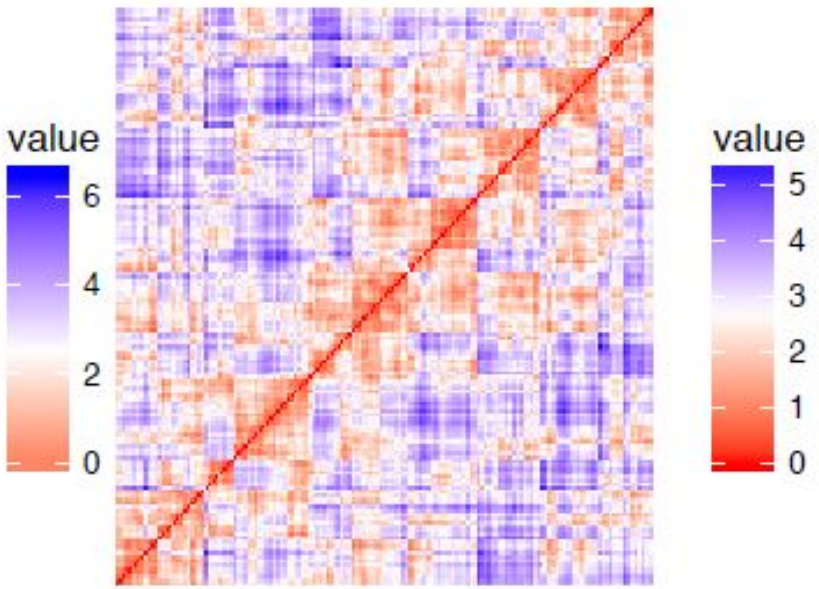
If the value of Hopkins statistic is close to one, then we can reject the null hypothesis and conclude that **the dataset D is significantly a clusterable data**.

Visual assessment of cluster tendency

The algorithm of the visual assessment of cluster tendency is as follow:

- compute the dissimilarity (DM) matrix between the objects in the dataset
- reorder the DM so that similar objects are close to one another. This process create an ordered dissimilarity matrix (ODM)
- the ODM is displayed as an ordered dissimilarity image (ODI), which is the visual output of algorithm





Thank you for your attention!!!