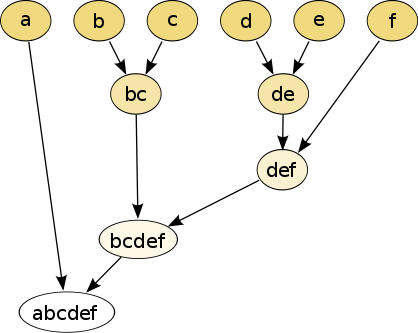
**Introduction – cluster analysis**

The task of cluster analysis is to divide the set into groups of cases in which the elements forming the group are similar to each other (most often - close, in the sense of distance), and the elements belonging to other groups - different.  
Clustering methods:  
a) hierarchical - the division of the set into clusters is obtained on the basis of the previous division (either more general or detailed) in such a way that solutions form a hierarchical structure  
b) partitioning - clusters are obtained by generating partitions of the data set

**Hierarchical Clustering**

[](http://upload.wikimedia.org/wikipedia/commons/a/ad/Hierarchical_clustering_simple_diagram.svg)

**Partition clustering**

The k-means algorithm works as follows:

• Select the number of clusters k.

• Randomly specify clusters or randomly select cases and treat them as clusters

• Assign points to clusters - taking into account their shortest distance from the center of the cluster

• Calculate cluster centers - as the center of gravity of the points forming them.

• Repeat the last two points until the stop condition is met (e.g. no change to assigning objects to clusters)

How to check if the solution of the clustering process is "good"? For example, verify that it is compatible with known labels (if they are given). You can do this by counting the Rand index (n - is the number of elements of the whole set):

 R = \frac{a+b}{a+b+c+d} = \frac{a+b}{{n \choose 2 }}

a - number of pairs that belong to the same cluster in division I and II  
b - number of pairs that belong to these different clusters in division I and II  
c - number of pairs that belong to the same cluster in division I and to various in II  
d - the number of pairs that belong to different clusters in division I and to the same in II

**Experimental part of the lab**

* Please import the set of seeds (note: the last column is the class labels - we skip them in the clustering - but they can be used to verify the applied method) and perform clustering using the hierarchical method (Matlab: linkage function, Python’s sklearn: AgglomerativeClustering). Please, experiment with the methods of calculating the distance between clusters and make an illustration of the clustering result (Matlab: *dendrogram* function, Python: *dendrogram*, <https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>).
* Find the clustering solution for the set of seeds using k-means (Matlab – *kmeans*, Python’s sklearn KMeans: <https://scikit-learn.org/stable/modules/clustering.html#k-means>) on a set with 7 attributes and on a set reduced using PCA (up to 2 dimensions). Is the result better (in the sense of the value of the Rand index closer to 1 - when we count it against the class labels?)

Hint (calculating the Rand index): compare the cluster labels' vector obtained as a result of k-means (division I) with the last, omitted, column of seeds set (division II) - iteratively checking individual elements according to the above defined criteria and counting a and b.

Python: <https://stackoverflow.com/questions/49586742/rand-index-function-clustering-performance-evaluation>

Matlab: <https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/13916/versions/3/previews/valid_RandIndex.m/index.html>