# 1 k-Means Clustering

- **Complexity of Proximity Matrix** - Hierarchical clustering requires a distance or similarity matrix between all pairs of cases. That's an extremely large matrix if you have tens of thousands of cases in your data file.

- A clustering method that doesn't require computation of all possible distances is ***k-means clustering***. It differs from hierarchical clustering in several ways. You have to know in advance the number of clusters you want. You can't get solutions for a range of cluster numbers unless you rerun the analysis for each different number of clusters.

- The algorithm repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during the analysis. In agglomerative hierarchical clustering, on the other hand, cases are added only to existing clusters. They are forever captive in their cluster, with a widening circle of "neighbours".

- The algorithm is called **k-means**, where **k** is the number of clusters you want, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest.

- **Computational Differences**: The k-means algorithm follows an entirely different concept than the hierarchical methods discussed before. This algorithm is not based on distance measures such as Euclidean distance or city-block distance, but uses the ***within-cluster variation*** as a measure to form homogenous clusters. Specifically, the procedure aims at segmenting the data in such away that the within-cluster variation is minimized. Consequently, we do not need to decide on a distance measure in the first step of the analysis.

- The action in the algorithm centers around finding the k-means. You start out with an initial set of means and classify cases based on their distances to the centers.

- Next, you compute the cluster means again, using the cases that are assigned to the cluster; then, you reclassify all cases based on the new set of means. You keep repeating this step until cluster means don't change much between successive steps.

- Finally, you calculate the means of the clusters once again and assign the cases to their permanent clusters.