Clustering the data

In the field of *Data Mining*, *Cluster Analysis* or *Clustering* is a process of grouping data objects from a dataset into multiple groups or clusters. The essential criterion for the quality of the clustering is a certain *similarity*, such that data objects are similar to other objects in the same cluster and dissimilar to objects from other clusters.
In the scope of this work, we decided to use well-known partitioning methods such as k-means or k-medoids. In general, given n data objects, partitioning methods distribute the data object into k clusters with $k < n$, using a distance measure to evaluate the repsective similarity. Partitioning methods form exclusive clusters by ensuring that each cluster contains at least one object. Note that the number k of clusters has to be chosen manually prior to the partitioning process.

0.1 k-Means

As pointed above, k-Means is one of the most used clustering algorithms. k-means is a *centroid based technique*, which means that each cluster is represented by a data point that also is the center of the cluster. A distance measure is then used to assigns every remaining data object from the data set to the best fitting cluster according to it's similarity to the center of this cluster and it's dissimilarity to the centers of any other cluster.
The data objects within a dataset are considered to reside in a euclidean space. Thus, the euclidean distance is used to calculate a score for the similarity of two data points. When using k-Means, the quality of a cluster C_i can b1e evaluated by computing the sum of squared erros between all data points and the centroid in C_i. This method is known as *within-cluster variation* and defined as follows:
[TODO: LEAST SQUARES]
Given k, the first step of k-Means is to select k random objects in the data set as initial representatives for the cluster centers. After that, each remaining data object is assigned to the best fitting cluster in terms of similarity, based on a similarity score that is determined by the euclidean distance between the data point and every cluster center. The overall goal of k-Means clustering is to iteratively optimize the within-cluster variation. In order to achieve this, for each cluster centroid is reassigned as the mean of all objects within that cluster. By considering the updated centroids, every data point is redistributed to the now best fitting cluster. This iterative process will continue, until no better clustering can be found, which means that the latest clustering equals the previous distribution.
[TODO: ALGORITHMUS]

0.2 k-Medoids

By the choic1e of using euclidean distance as a similarity meaasure, k-Means is sensitive to outliers in a wây, that they can heavily distort the mean of a cluster and, as a result, worsen the within-cluster variation.
k-medoids is a modification of k-means,