# EDAMI laboratory

Clustering

# Basic notions

- Object – an entity described by a set of attributes
  – Nominal attributes
  – Numerical attributes
- Data base (DB) – a set of objects.

# Clustering

The purpose of clustering is to divide a set of objects into groups including similar objects (objects having similar values of attributes).

In some methods the number of groups has to be given as an input parameter.

A good clustering must have the following property:

- high similarity between objects within groups,
- low similarity between objects belonging to different groups.

Similarity is often defined as a certain distance measure between two objects.

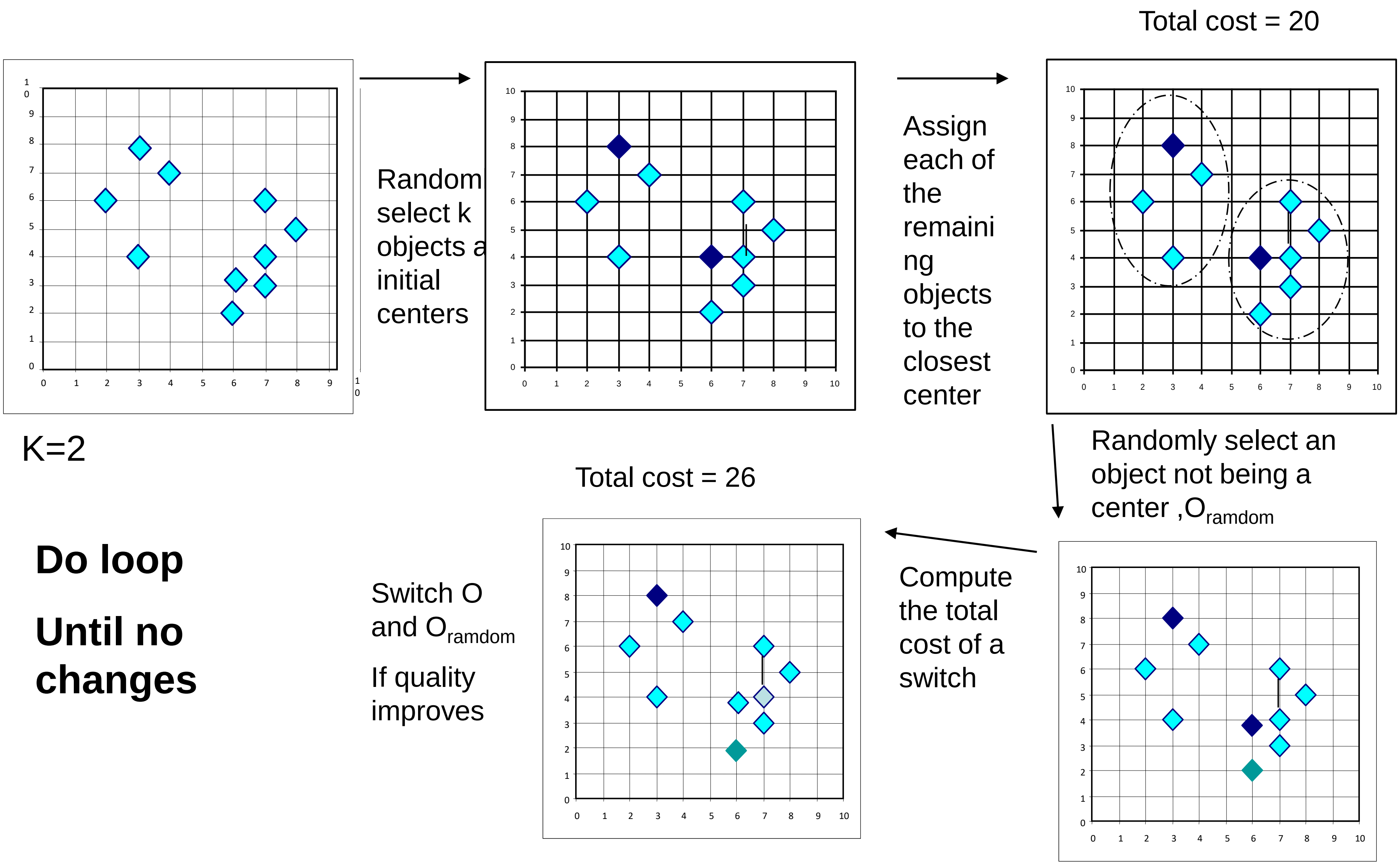
# Partitioning algorithms: basic notions

- Partitioning method: create a division of the database *D* composed of *n* objects into *k* clusters, minimize the sum of distances squared

$$\sum_{i=1}^{k}\sum_{p\in C_i} dist(p, c_i)^2$$

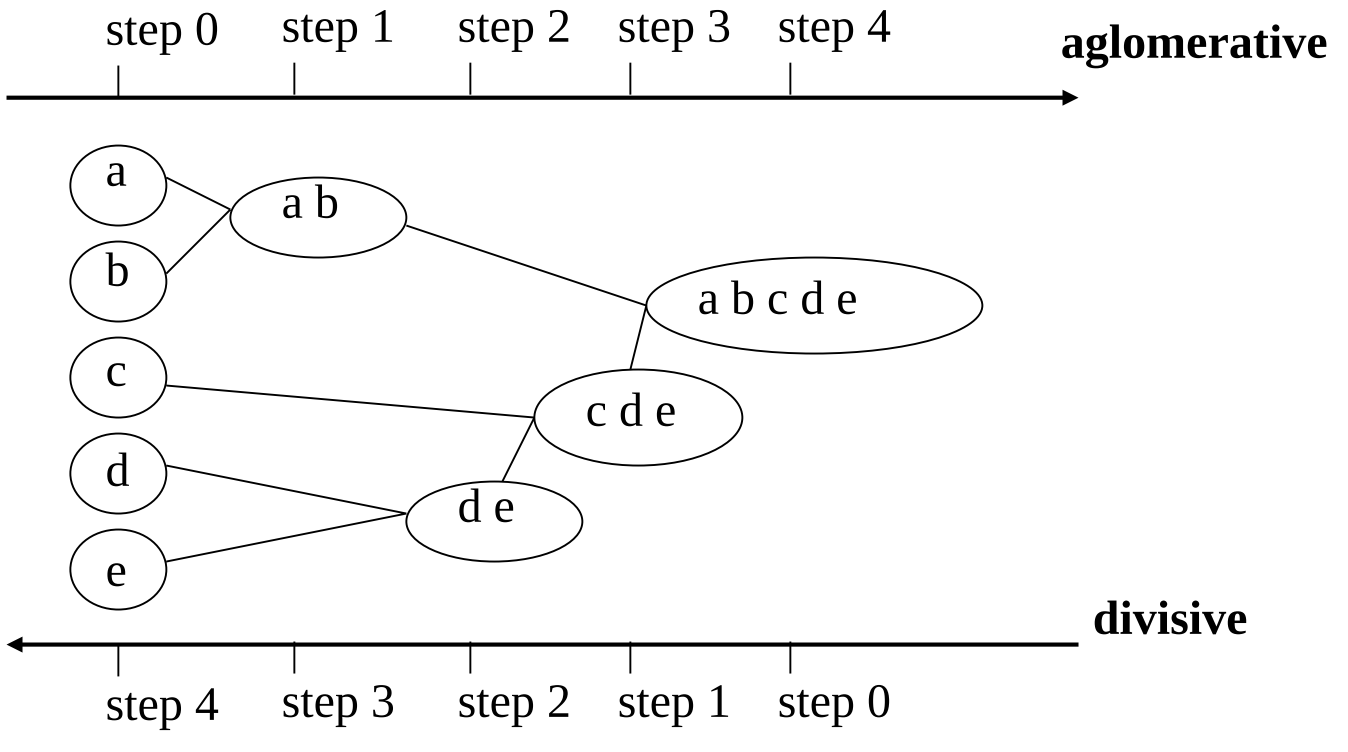
- Given *k* find such partitioning into *k* clusters that optimizes the selected partitioning criterion
  - heuristic methods: k-medoids and k-means algorithms
    - *k-means* (MacQueen'67): each cluster is represented by its center
    - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): each cluster is represented by one of the objects in the cluster

# *Typical k-means algorithm*



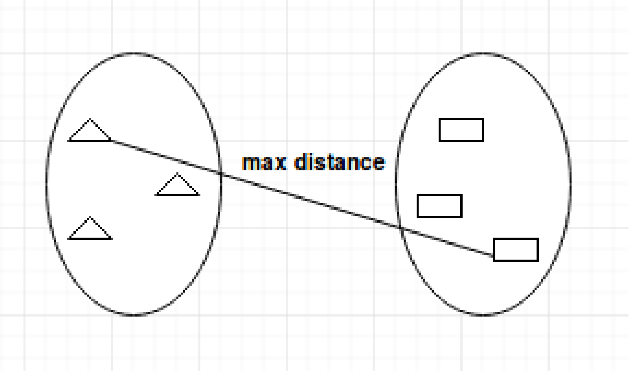
# Hierarchical clustering

- Aglomerative (most methods belong to this category)
- Divisive (finish after reaching the stop criterion, e.g. fixed numer of clusters, fixed diameter of a cluster)



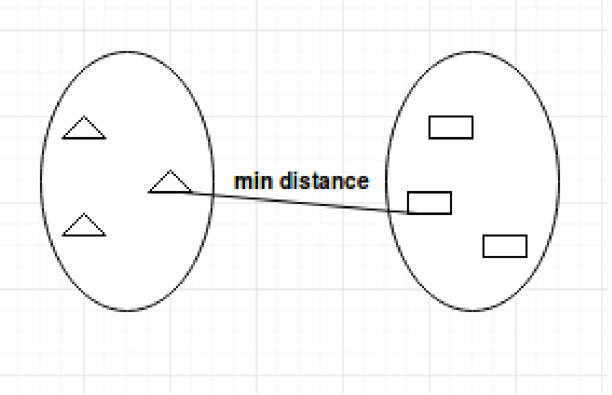
# Hierarchical clustering: cluster linkage methods

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.



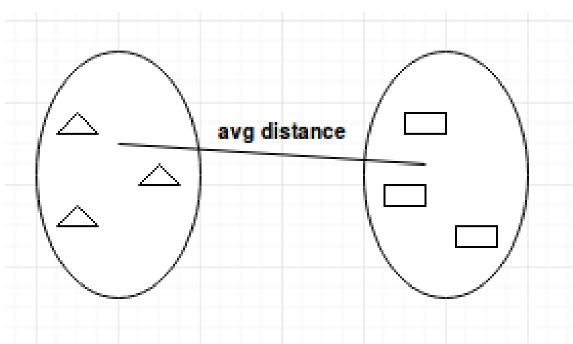
# Hierarchical clustering: cluster linkage methods

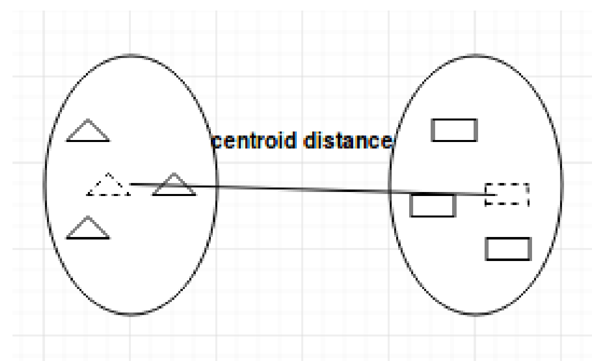
Minimum or single linkage: the distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, "loose" clusters.



# Hierarchical clustering: cluster linkage methods

- 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.



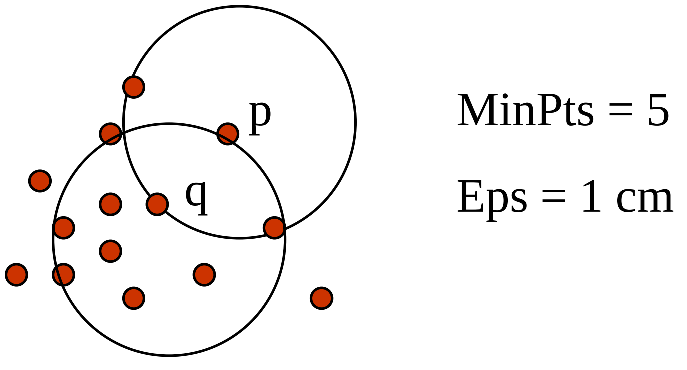




At each stage of the clustering process the two clusters, that have the smallest linkage distance, are linked together.

# Density-based clustering: basic notions

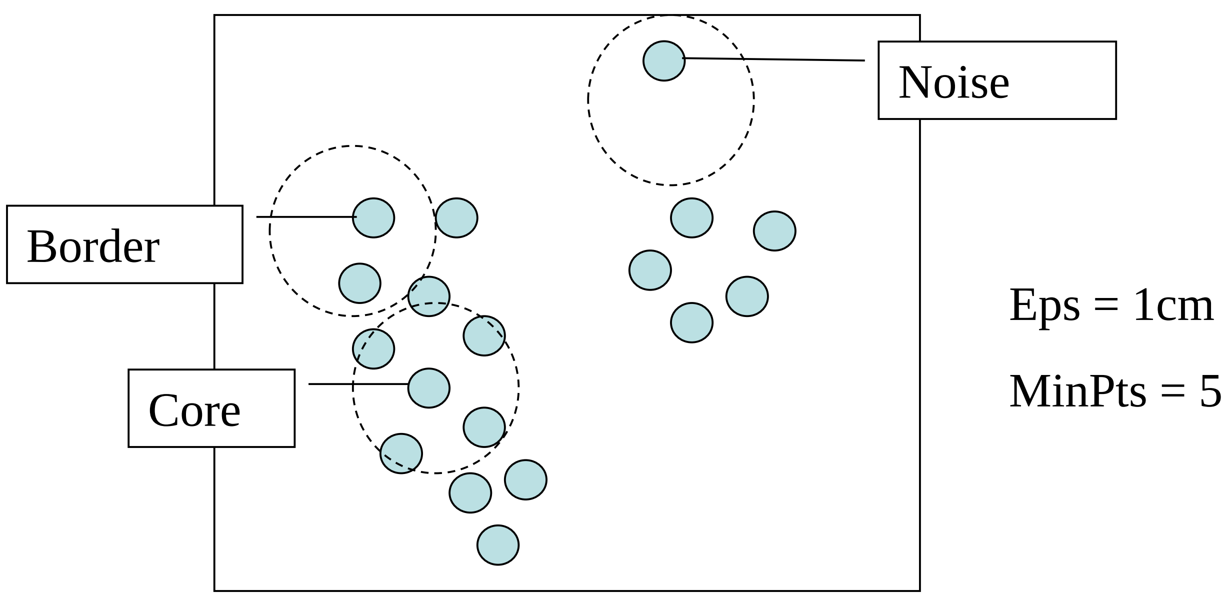
- Two parameters*:*
  - *Eps*: maximum neighborhood diameter
  - *MinPts*: minimum nb of points in the neighborhood Eps of the point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid dist(p,q) <= Eps\}$
- Directly density-reachable: Point *p* is directly density-reachable from point *q* with respect to *Eps* and *MinPts* if
  - *p* belongs to $N_{Eps}(q)$
  - conditio of the core point:

$$|N_{Eps}(q)| >= MinPts$$



# DBSCAN

Based on the notion of a density-based cluster: a cluster is defined as the maximum set of density-reachable points.



# Evaluation of clustering quality (1)

## Index Silhouette

$$Silhouette(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

where:

a(x) – the average distance between x and other objects in a group including x

b(x) – the minimum average distance between x and the nearest group not including x.

The index has a value from the range <-1, 1>, where 1 means that the object is assigned to the best possible group, 0 - the object is located between two groups, and -1 - wrong assignment of the object.

$$GSilhouette = \frac{1}{N} \sum_{i=1}^{N} Silhouette(x_i)$$

where: N – number of objects

# Evaluation of clustering quality (2)

## Rand index

W – reference clustering,  G – obtained clustering

A – a number of pairs of objects belonging to the same group in W and G

B – a number of pairs of objects belonging to the different groups in W and G

a – a number of pairs of objects belonging to the same group in W but not in G

b – a number of pairs of objects belonging to the different group in W but in the same in G

$$\mathrm{R} = \frac{A+B}{A+B+a+b} = \frac{A+B}{n(n-1)/2}$$

# Clustering in R (1)

Standard packages

- scale() – for centering and/or scaling the columns of a numeric matrix.
- kmeans()  - k-Means algorithm, returns a kmeans object with a description of clusters.
- hclust ()  - hierarchical clustering.
- cutree() – for cutting trees obtained by the means of hclust() function.
- plot() – for visualization of the clusters.

# Clustering in R(2)

- Package fpc
  - dbscan () - an implementation of the DBScan algorithm
  - plotcluster() – a function for plotting clusters.
- Package cluster
  - Includes implementations of several clustering algorithms