



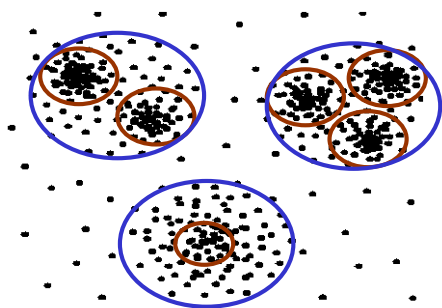
Big Data Analytics

– Week 3: Hierarchical Clustering –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

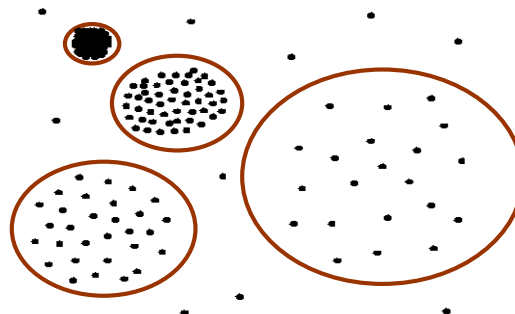
From Partitioning to Hierarchical Clustering

- Global parameters to separate all clusters with a partitioning clustering method may not exist



*hierarchical
cluster structure*

and/or



*largely differing
densities and sizes*

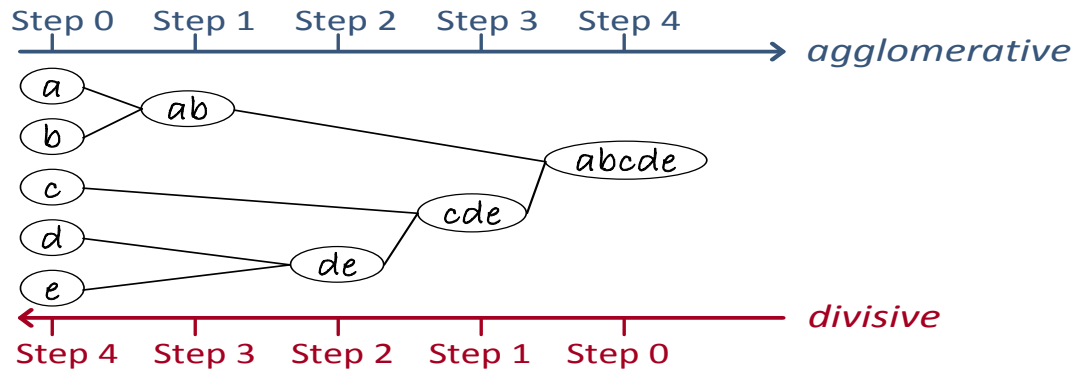
- Need a hierarchical clustering algorithm in these situations

Hierarchical Clustering: Basic Notions

Hierarchical decomposition of the data set (with respect to a given similarity measure) into a set of nested clusters

Result represented by a so called *dendrogram* (greek $\delta\epsilon\nu\delta\rho\omicron$ = tree)

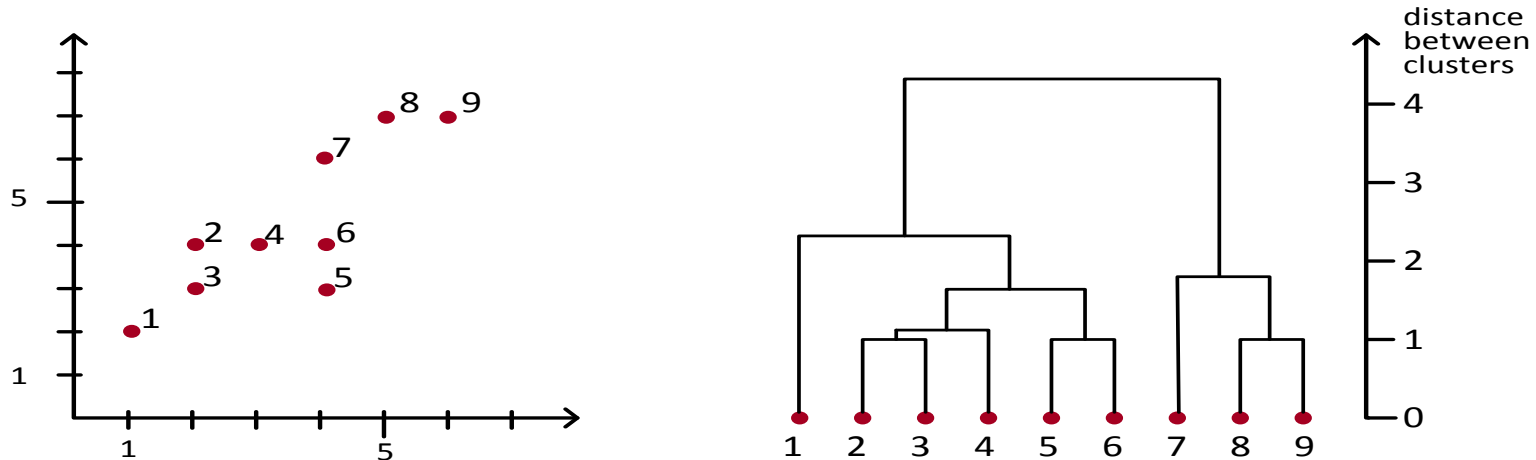
- Nodes in the dendrogram represent possible clusters
- can be constructed bottom-up (agglomerative approach) or top down (divisive approach)



Hierarchical Clustering: Example

Interpretation of the dendrogram

- The root represents the whole data set
- A leaf represents a single object in the data set
- An internal node represents the union of all objects in its sub-tree
- The height of an internal node represents the distance between its two child nodes



Agglomerative Hierarchical Clustering

1. Initially, each object forms its own cluster
 2. Compute all pairwise distances between the initial clusters (objects)
 3. Merge the closest pair (A, B) in the set of the current clusters into a new cluster $C = A \cup B$
 4. Remove A and B from the set of current clusters; insert C into the set of current clusters
 5. If the set of current clusters contains only C (i.e., if C represents all objects from the database): STOP
 6. Else: determine the distance between the new cluster C and all other clusters in the set of current clusters; go to step 3.
- Requires a distance function for clusters (sets of objects)

Single Link Method and Variants

Given: a distance function $dist(p, q)$ for database objects

The following distance functions for clusters (i.e., sets of objects) X and Y are commonly used for hierarchical clustering:

Single-Link: $dist_sl(X, Y) = \min_{x \in X, y \in Y} dist(x, y)$

Complete-Link: $dist_cl(X, Y) = \max_{x \in X, y \in Y} dist(x, y)$

Average-Link: $dist_al(X, Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} dist(x, y)$