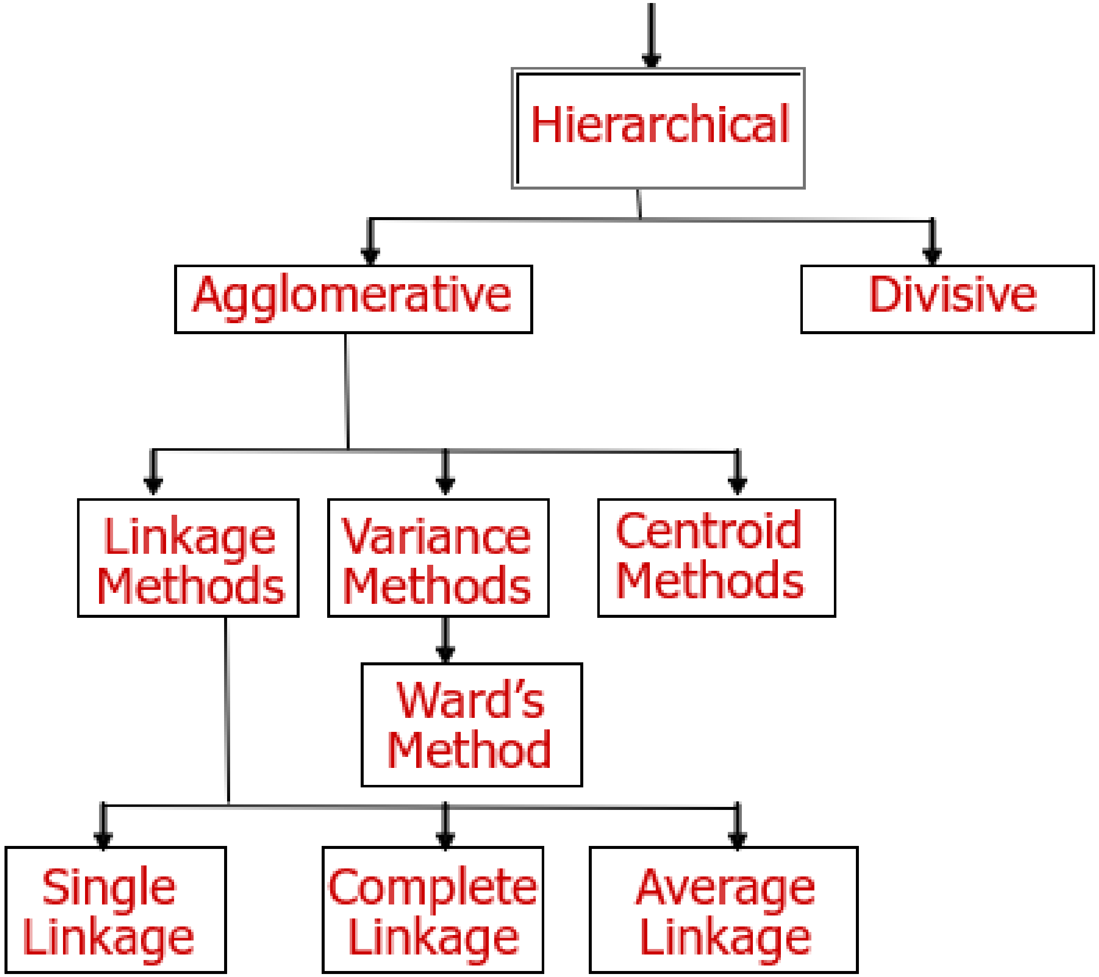
Hierarchical Clustering

Prof.Dr. Songül Varlı
Yıldız Technical University
Computer Engineering Department
svarli@yildiz.edu.tr

Hierarchical Clustering

- A hierarchical clustering method works by grouping objects into a tree of clusters.
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

Hierarchical Clustering



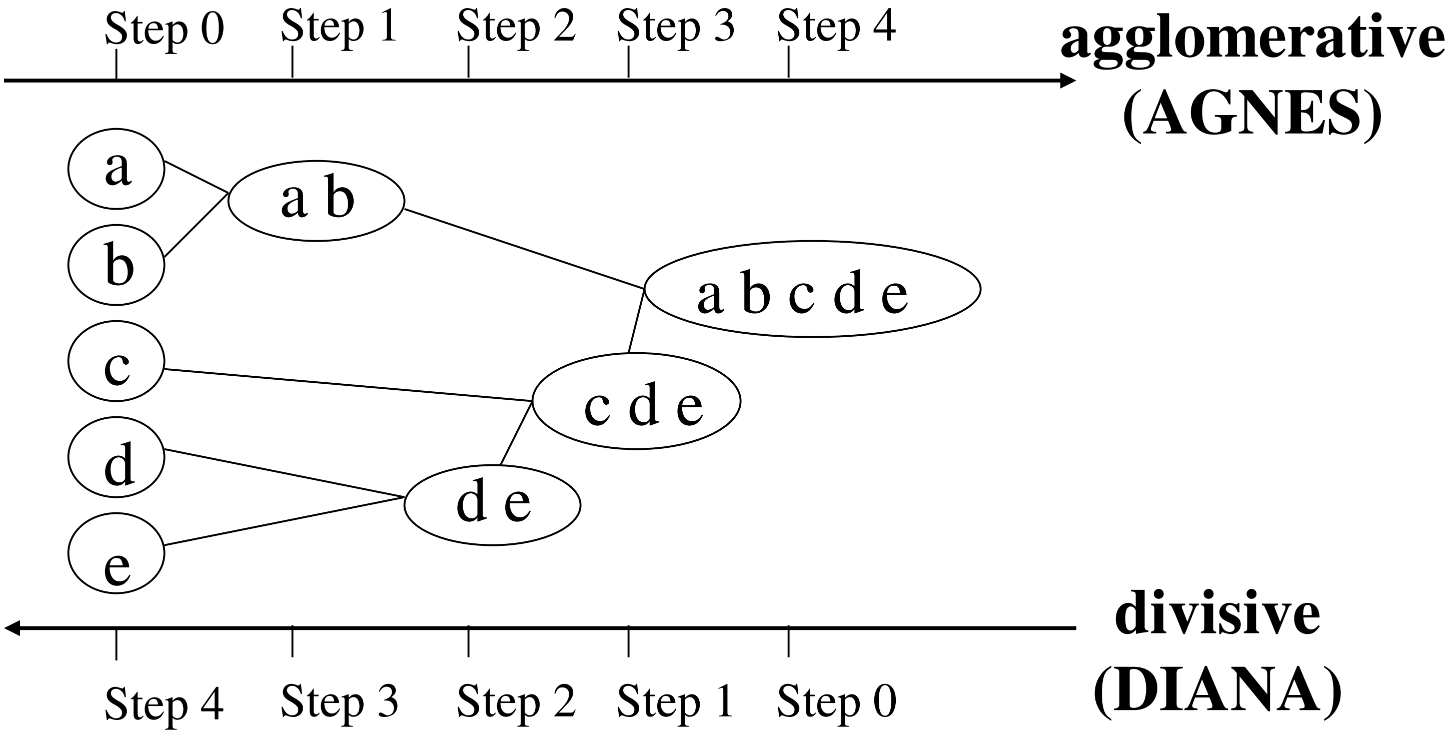
Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category.

Divisive Hierarchical Clustering

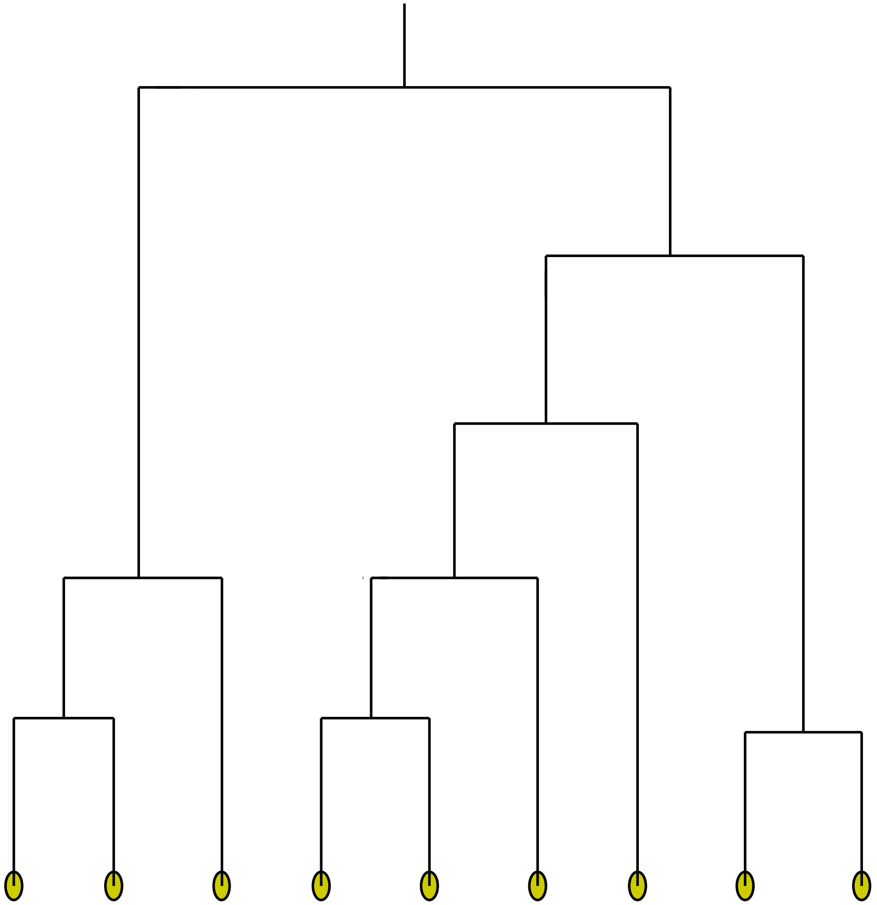
- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces, until each object form a cluster on its own or until it satisfies certain termination conditions, such as a desired number of cluster or the diameter of each cluster is within a certain threshold.

- Example: A data-set has five objects {a,b,c,d,e}
- AGNES (Agglomerative Nesting)
- DIANA (Divisive Analysis)



Dendrogram: Hierarchical Clustering

- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.



AGNES (Agglomerative Nesting)

- Initially, AGNES places each objects into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, cluster C_1 and C_2 may be merged if an object in C_1 and object in C_2 form the minimum Euclidean distance between any two objects from different clusters.
- This is single-linkage approach in that each cluster is represented by all of the objects in the cluster, and the similarity between two clusters is measured by similarity of the closest pair of data points belonging to different clusters.

Distance between clusters

- Four widely used measure for distance between clusters are as follows, where $|p - p'|$ is the distance between two objects or points, p and p' ;
  - m_i is the mean for clusters, C_i
  - n_i is the number of objects C_i

1. Minimum Distance:
2. Maximum Distance:
3. Mean Distance:
4. Average Distance:

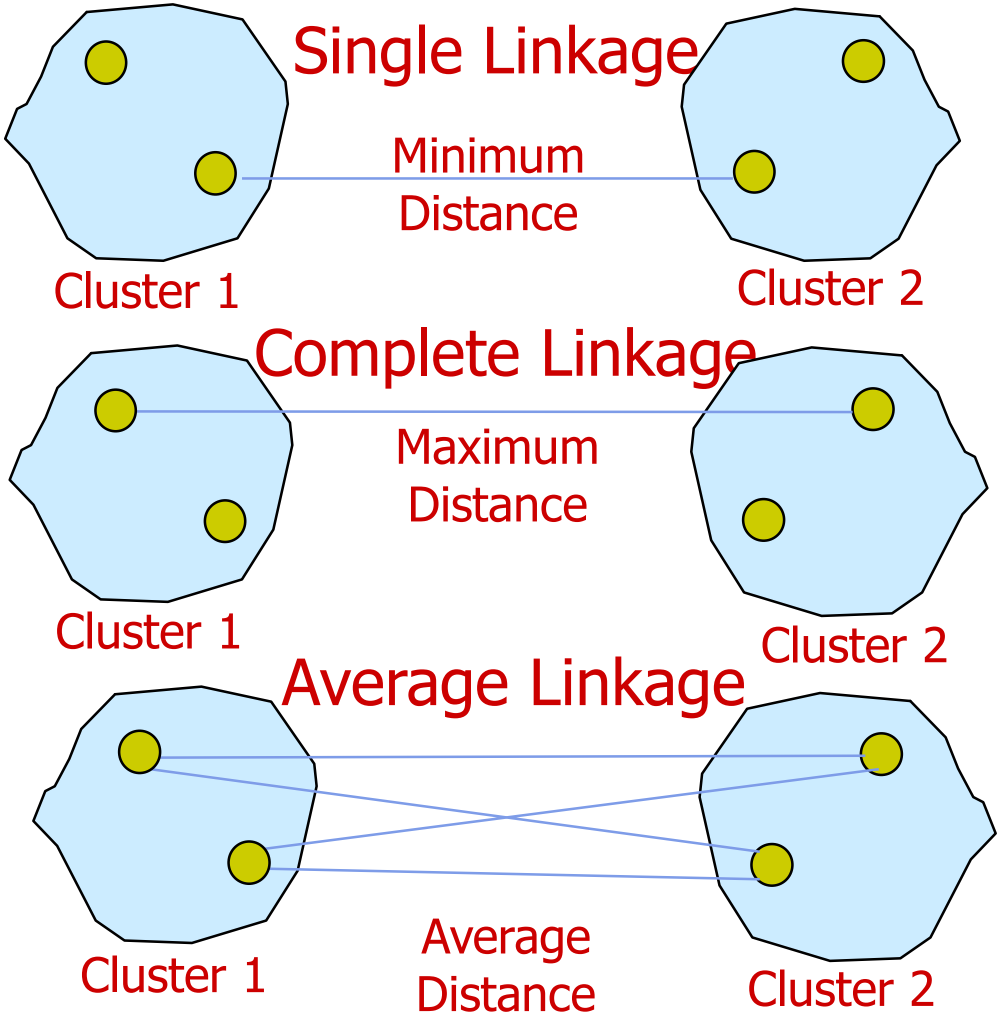
Single Linkage Algorithm:

- When an algorithm uses the minimum-distance $d_{min}(C_i,C_j)$, to measure the distance between clusters, it is sometimes called nearest-neighbor clustering algorithm. Moreover, if the clustering process is terminated when the distance between nearest clusters exceed an arbitrary threshold, it is called a single-linkage algorithm.

Complete Linkage Algorithm:

- When an algorithm uses the maximum-distance $d_{max}(C_i,C_j)$, to measure the distance between clusters, it is sometimes called a farthest-neighbor clustering algorithm. If the clustering process is terminated when the maximum distance between nearest clusters exceed an arbitrary threshold, it is called a complete-linkage algorithm.
- The distance between two clusters is determined by the most distant nodes in two clusters.

Linkage methods of Clustering



Single Linkage

Minimum Distance

Cluster 1

Cluster 2

Complete Linkage

Maximum Distance

Cluster 1

Cluster 2

Average Linkage

Average Distance

Cluster 1

Cluster 2

- The above minimum and maximum measures represent two extremes in measuring the distance between clusters.
- They tend to be overly sensitive to outliers or noisy data.
- The use of mean or average distance is compromise between min. and max. distance and overcomes the outlier sensitivity problem.

Hierarchical Agglomerative Clustering-Variance and Centroid Method

- **Variance methods** generate clusters to minimize the within-cluster variance.

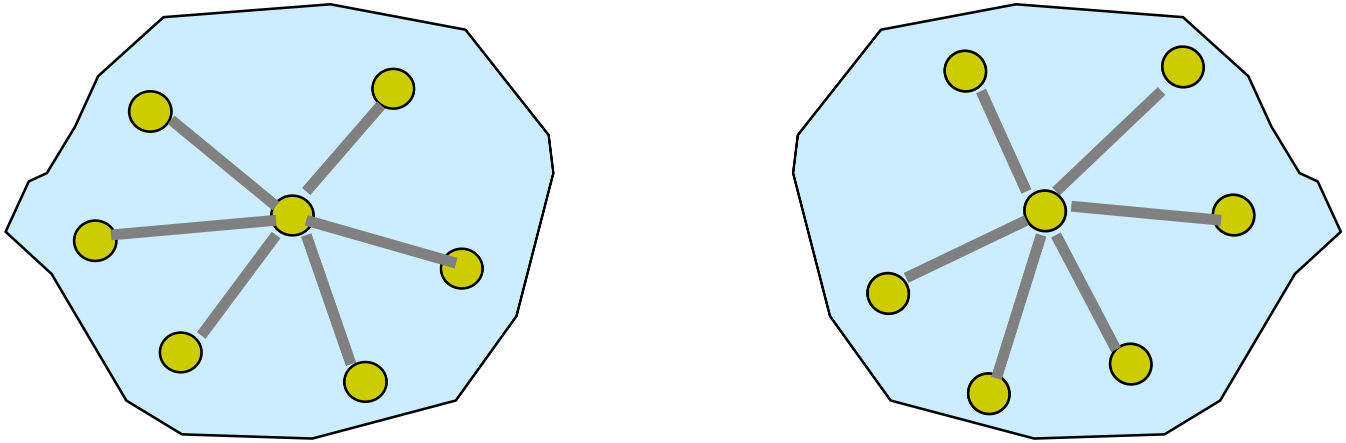
- **Ward's procedure** is commonly used. For each cluster, the sum of squares is calculated. The two clusters with the smallest increase in the overall sum of squares within cluster distances are combined.

- In the **centroid methods,** the distance between two clusters is the distance between their centroids (means for all the variables),

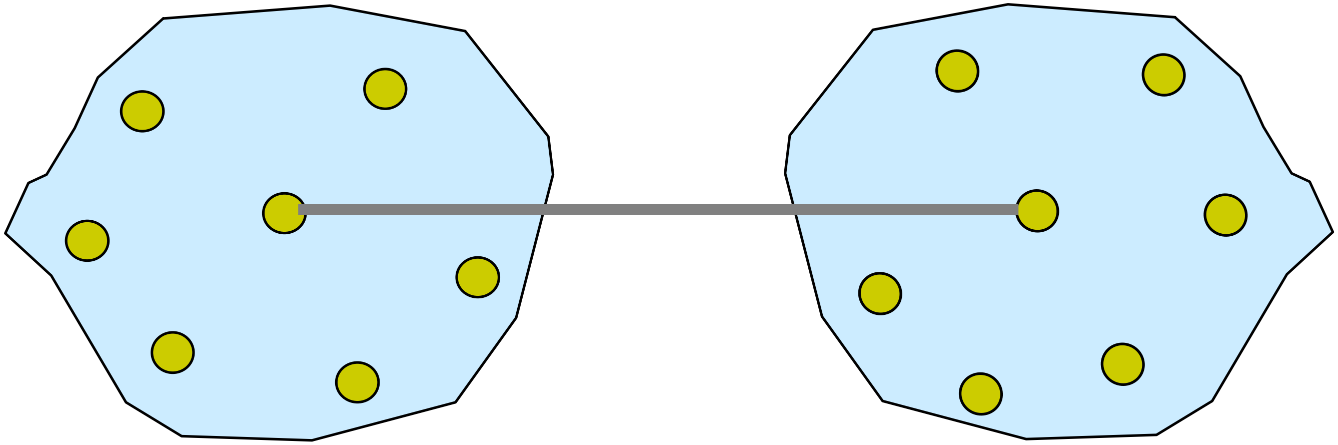
- Of the hierarchical methods, average linkage and Ward's methods have been shown to perform better than the other procedures.

Other Agglomerative Clustering Methods

Ward's Procedure



Centroid Method



What Is A Good Clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used

External criteria for clustering quality

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to ground truth … requires *labeled data*
- Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, ω_1, ω_2, …, ω_K with n_i members.

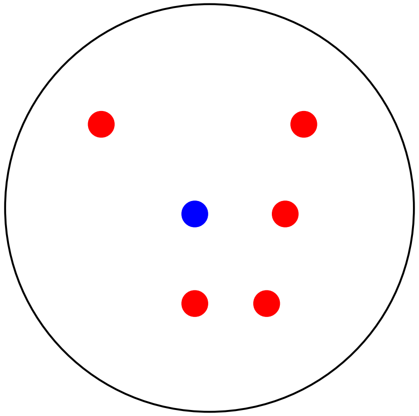
External Evaluation of Cluster Quality

- Simple measure: purity, the ratio between the dominant class in the cluster π_i and the size of cluster ω_i

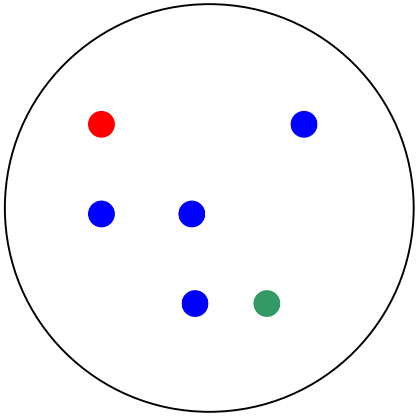
$$Purity(\omega_i) = \frac{1}{n_i} \max_j (n_{ij}) \qquad j \in C$$

- Biased because having *n* clusters maximizes purity
- Others are entropy of classes in clusters (or mutual information between classes and clusters)

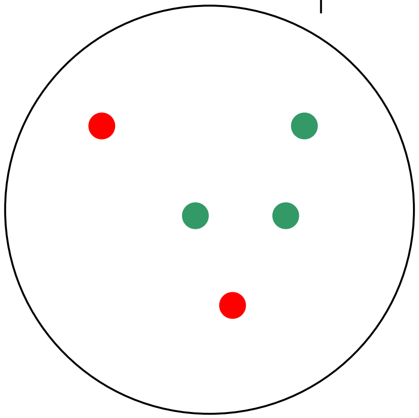
Purity Example



Cluster I



Cluster II



Cluster III

Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5