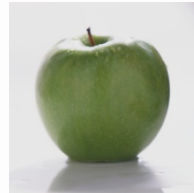


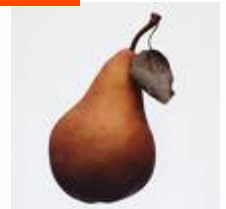


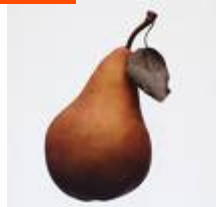
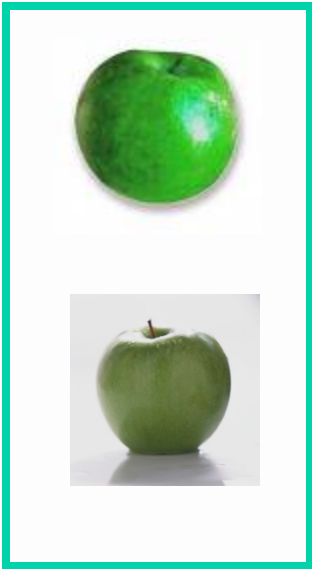
Clustering: Hierarchical Methods

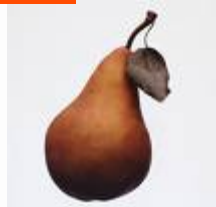
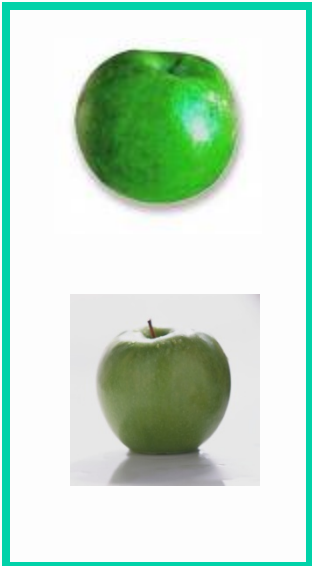
Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

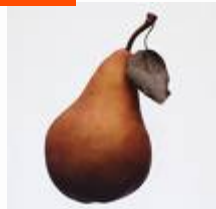
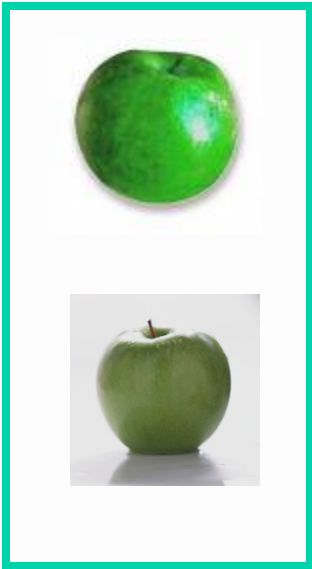


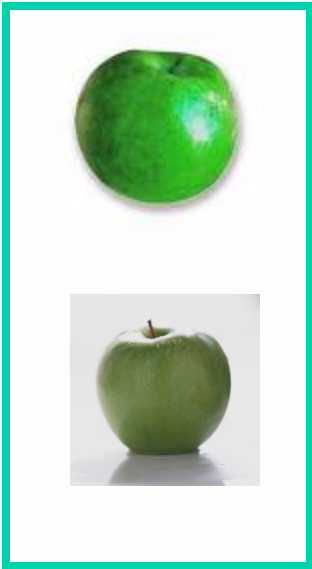
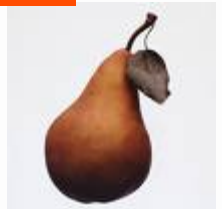


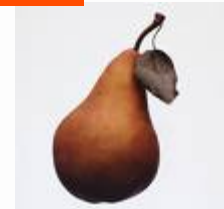
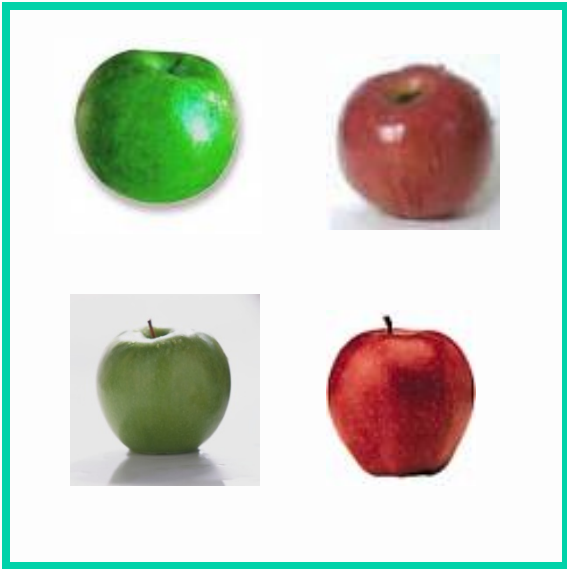


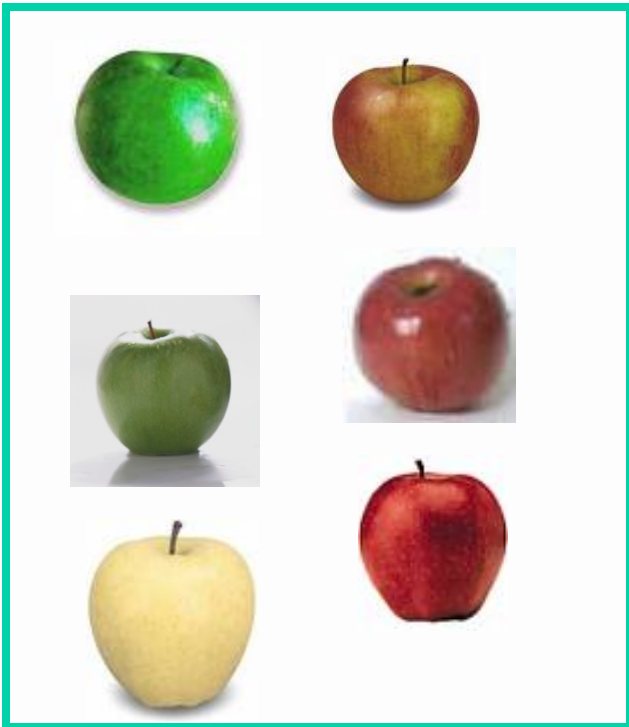
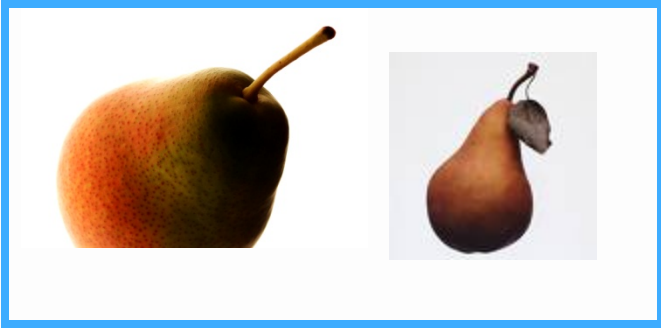








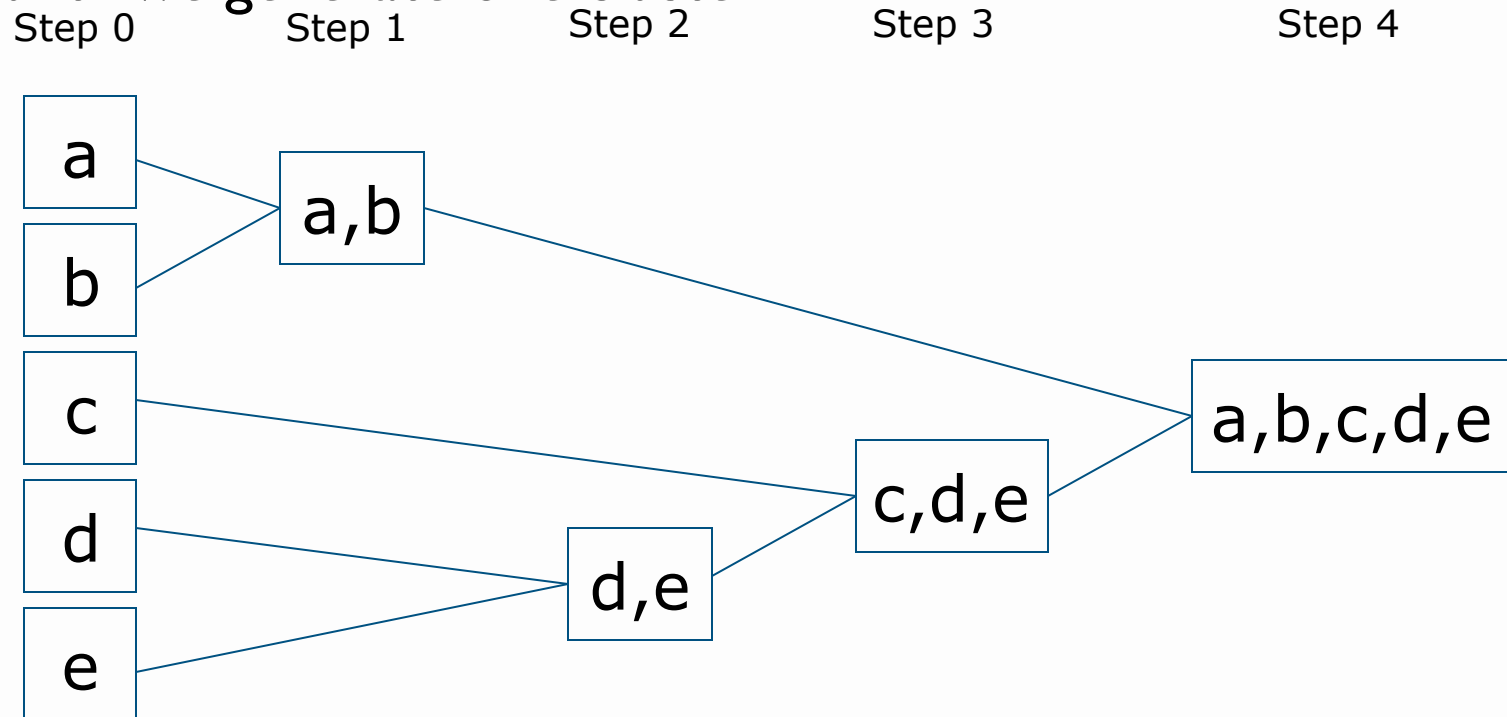




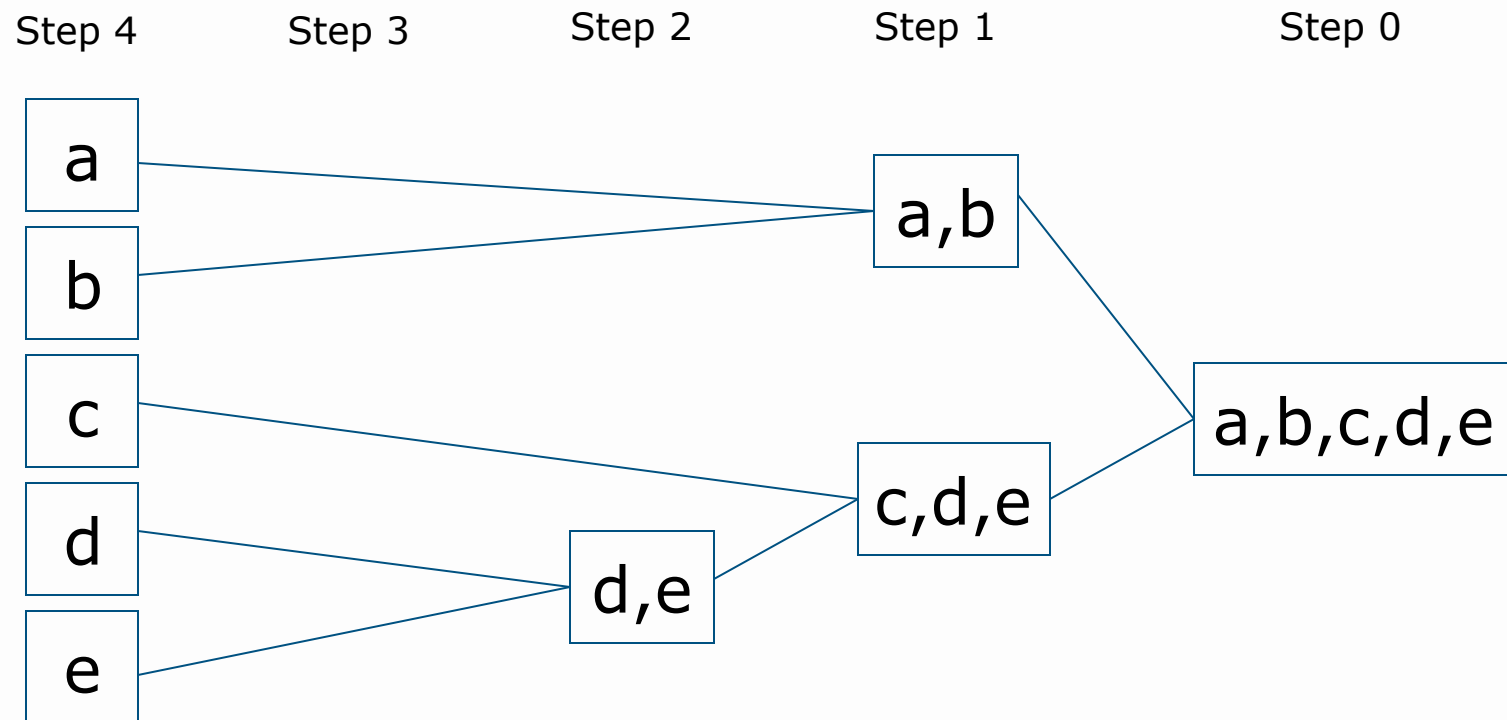
What is Hierarchical Clustering?

- Suppose we have five items, a, b, c, d, and e.
- Initially, we consider one cluster for each item
- Then, at each step we merge together the most similar clusters,

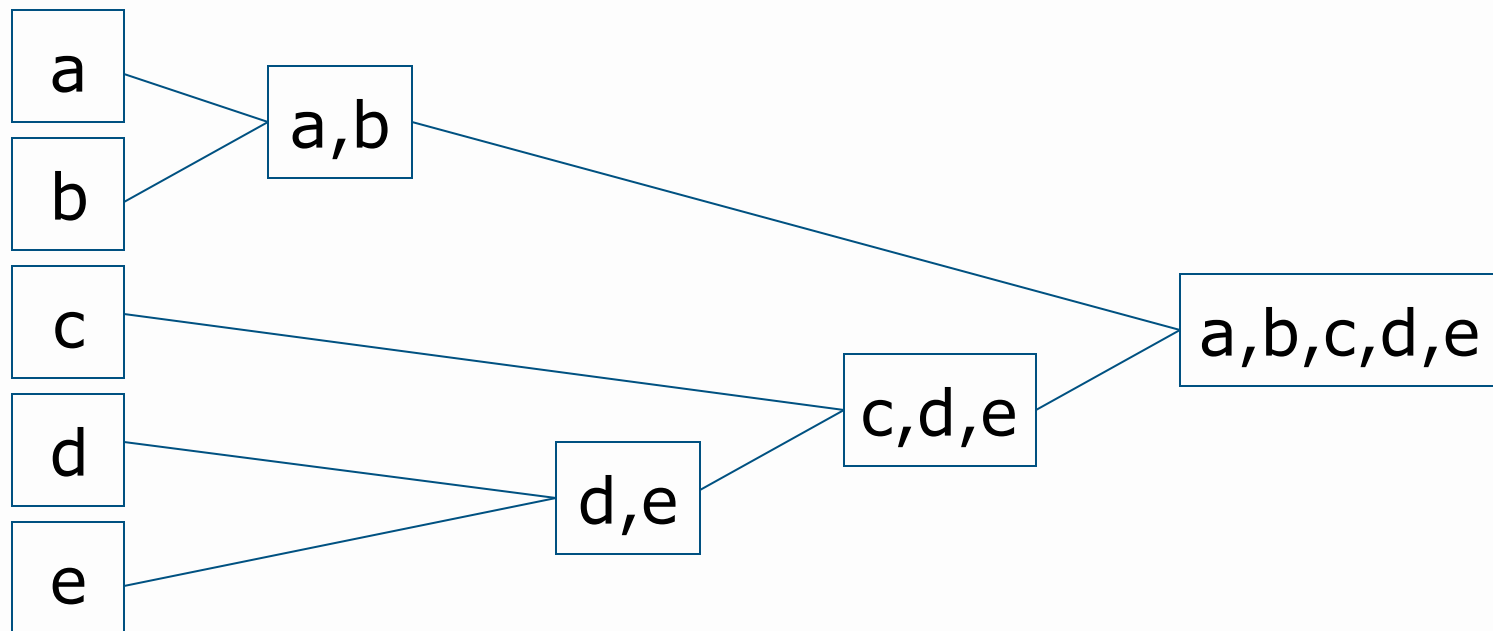
until we generate one cluster



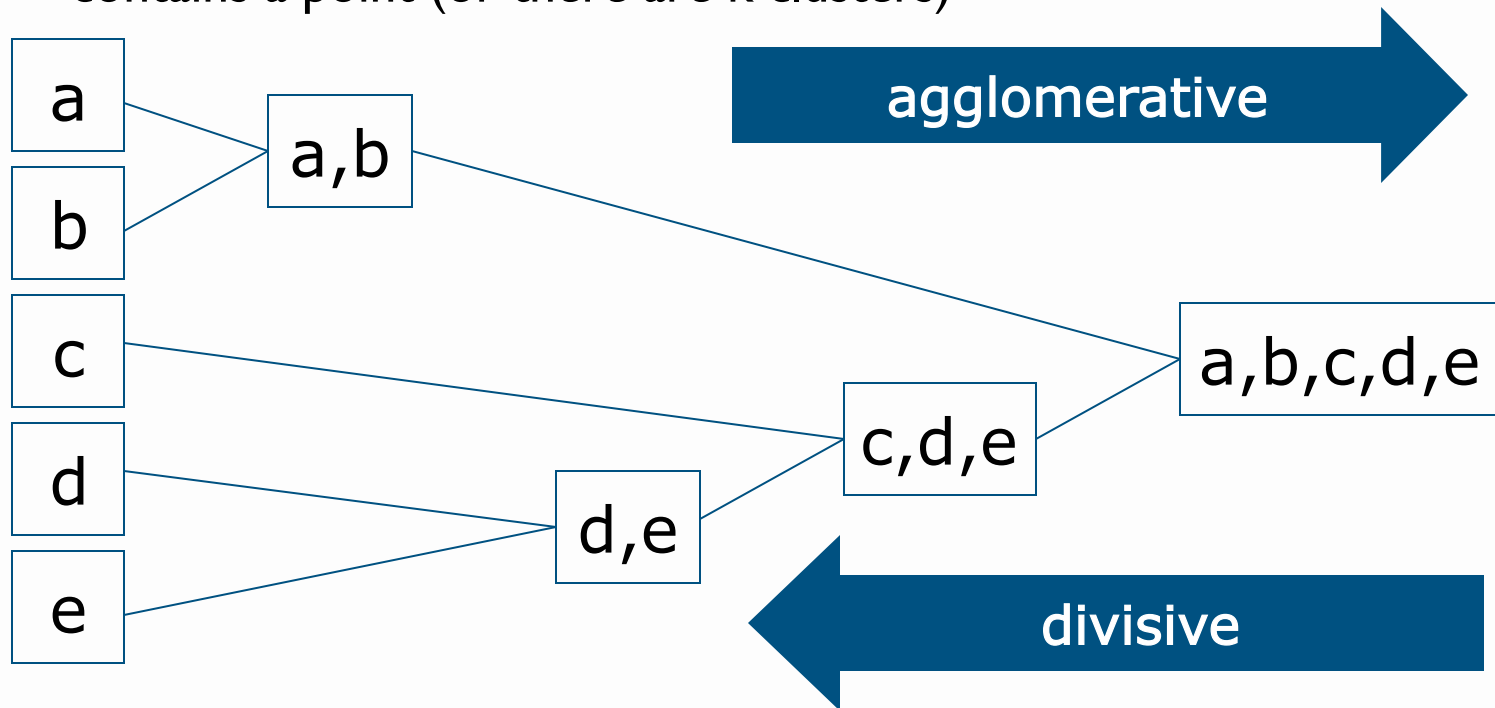
- Alternatively, we start from one cluster containing the five elements
- Then, at each step we split one cluster to improve intracluster similarity, until all the elements are contained in one cluster



- By far, it is the most common clustering technique
- Produces a hierarchy of nested clusters
- The hierarchy be visualized as a dendrogram: a tree like diagram that records the sequences of merges or splits



- Agglomerative
 - Start individual clusters, at each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive
 - Start with one cluster, at each step, split a cluster until each cluster contains a point (or there are k clusters)

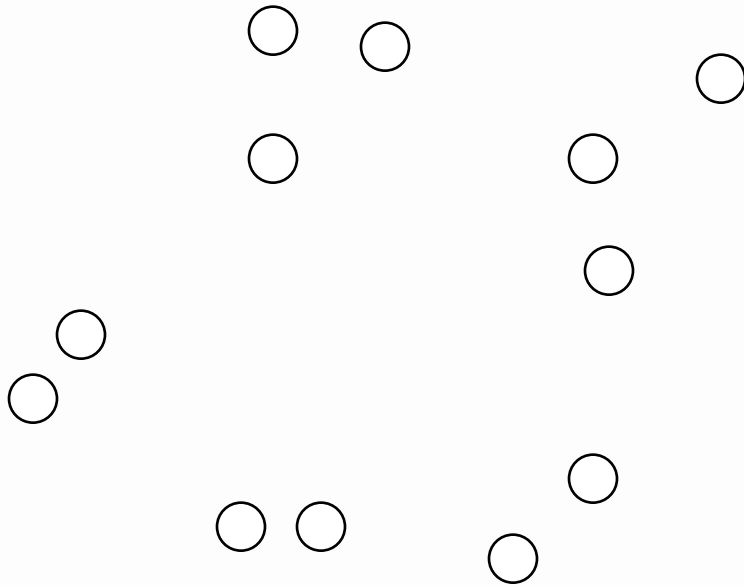


- No need to assume any particular number of clusters
- Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level
- They may correspond to meaningful taxonomies
- Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, ...)
- Traditional hierarchical algorithms use a similarity or distance matrix to merge or split one cluster at a time

- More popular hierarchical clustering technique
- Compute the proximity matrix
- Let each data point be a cluster
- Repeat
 - Merge the two closest clusters
 - Update the proximity matrix
- Until only a single cluster remains

- Key operation is the computation of the proximity of two clusters
- Different approaches to defining the distance between clusters distinguish the different algorithms

- Start with clusters of individual points and a proximity matrix

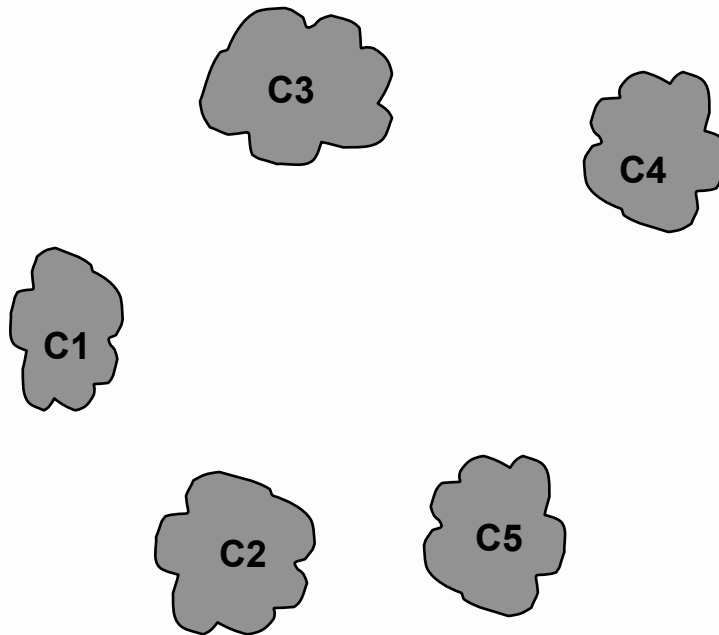


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

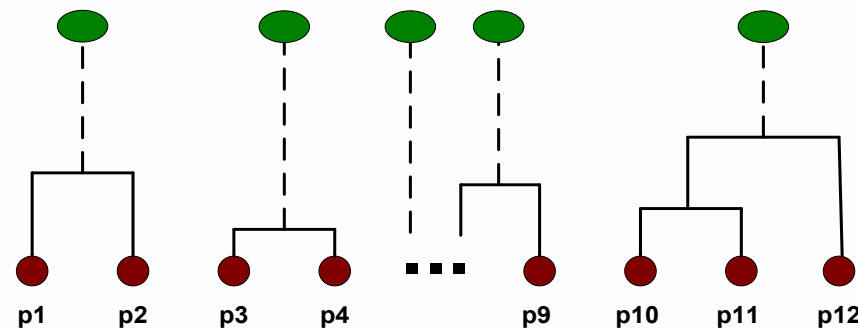


- After some merging steps, we have some clusters

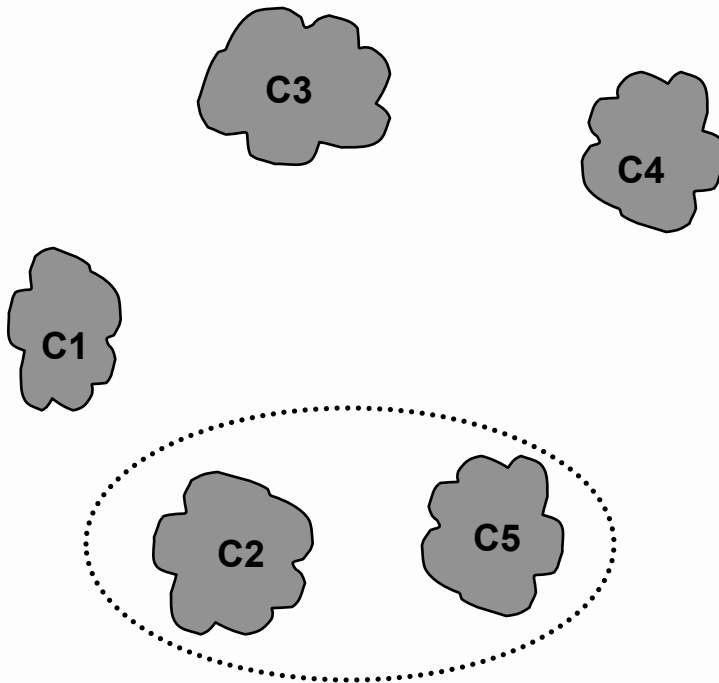


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

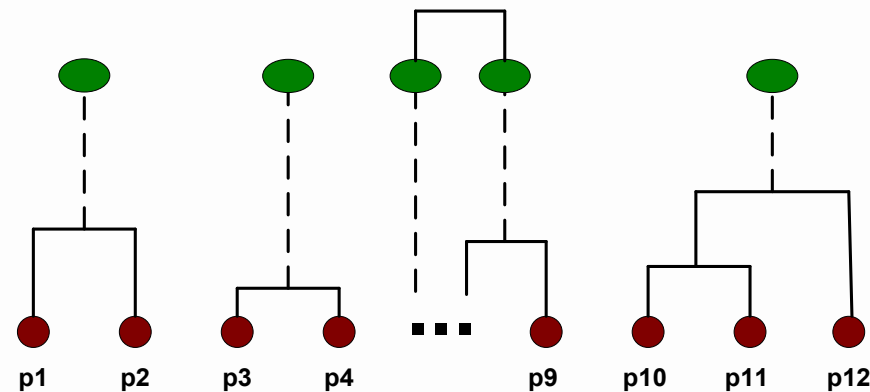


- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

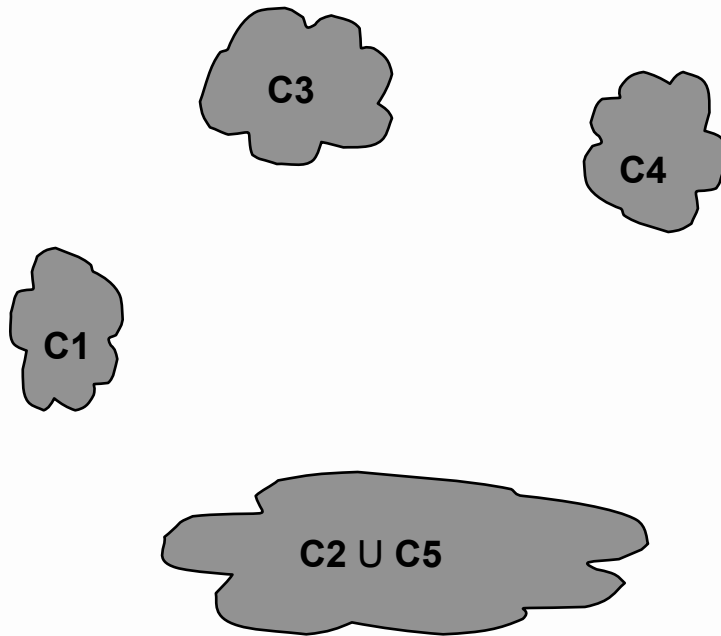


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

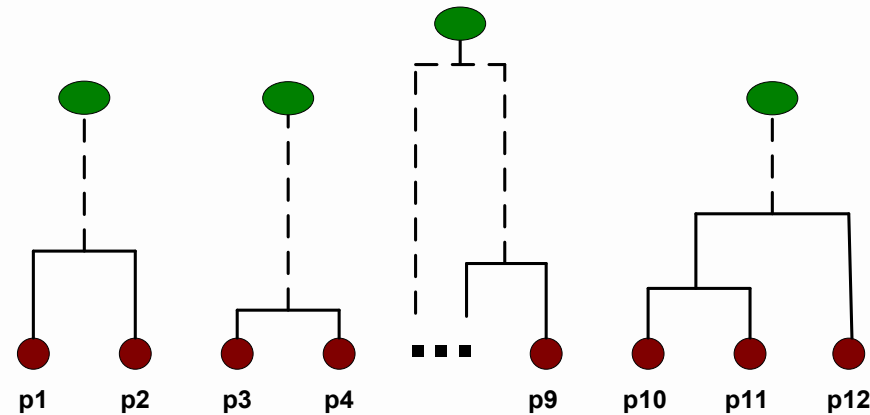


- The question is “How do we update the proximity matrix?”

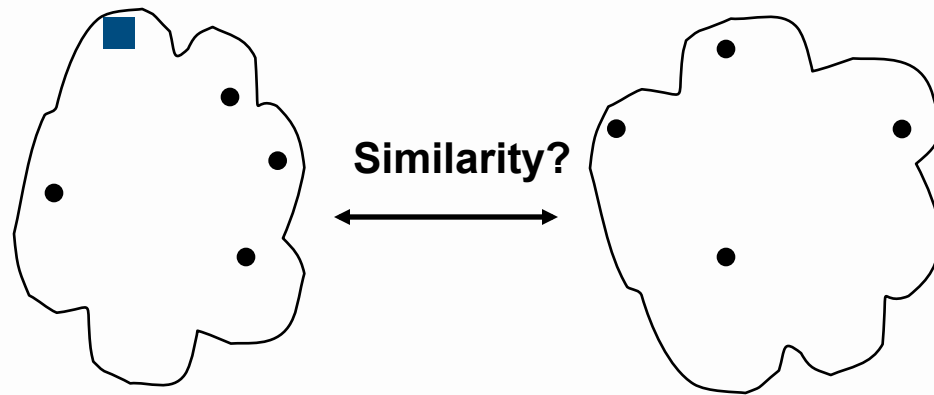


		C2 U C5			
		C1	C5	C3	C4
C1			?		
C2 U C5		?	?	?	?
C3			?		
C4			?		

Proximity Matrix



How to Define Inter-Cluster Similarity

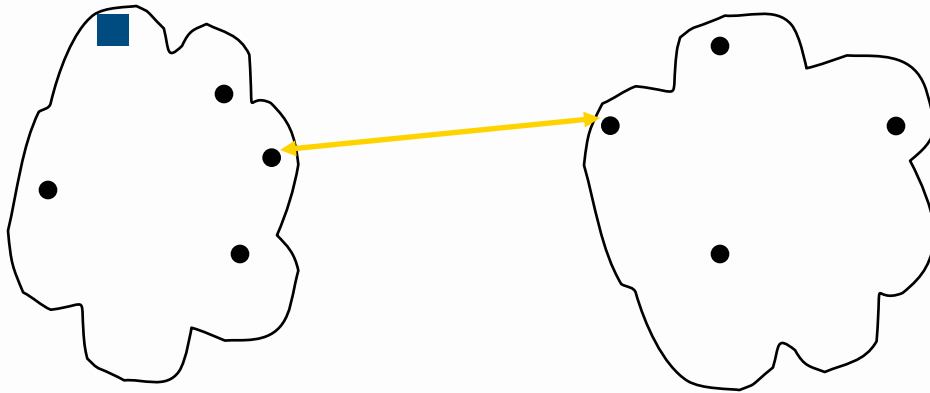


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - ▶ Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

· **Proximity Matrix**

How to Define Inter-Cluster Similarity

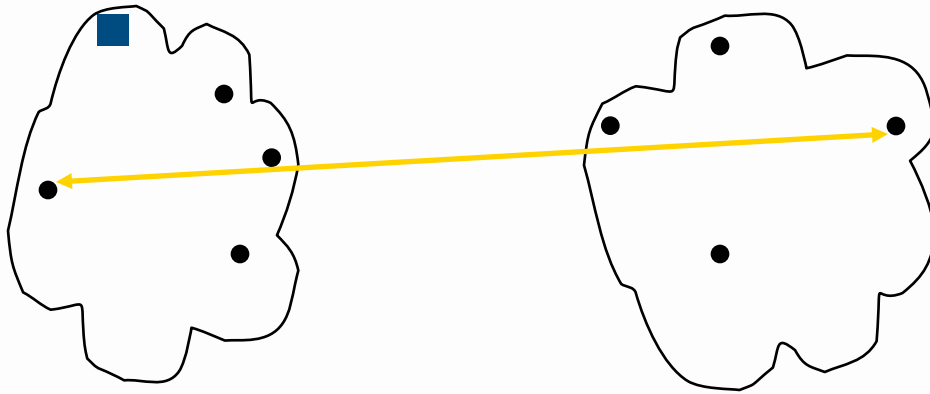


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - ▶ Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

· **Proximity Matrix**

How to Define Inter-Cluster Similarity

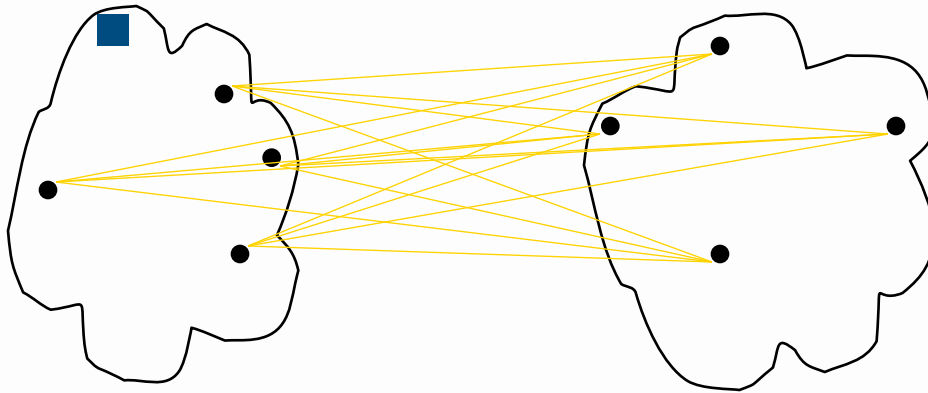


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - ▶ Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

· **Proximity Matrix**

How to Define Inter-Cluster Similarity

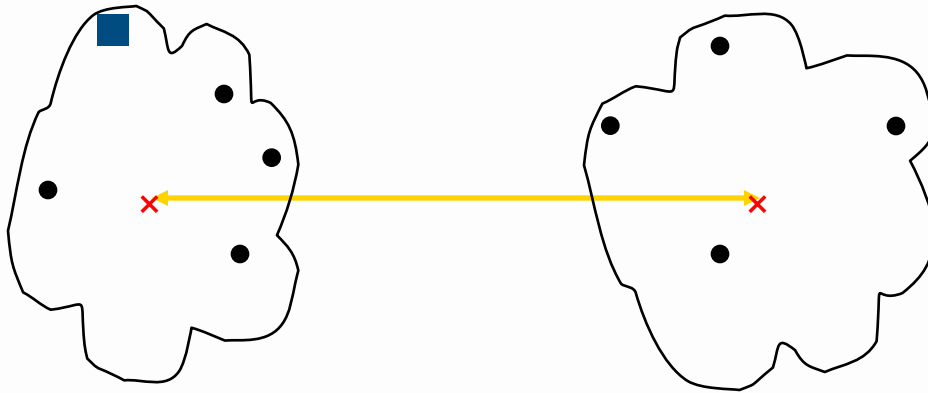


- MIN
- MAX
- Group Average**
- Distance Between Centroids
- Other methods driven by an objective function
 - ▶ Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

· **Proximity Matrix**

How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - ▶ Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

· **Proximity Matrix**

Typical Alternatives to Calculate the Distance between Clusters

- Single link
 - smallest distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- Complete link
 - largest distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- Average
 - average distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- Centroid
 - distance between the centroids of two clusters, i.e., $\text{dis}(K_i, K_j) = \text{dis}(C_i, C_j)$
- ...

Hierarchical Clustering: Time and Space Requirements

- $O(N^2)$ space since it uses the proximity matrix.
 - N is the number of points.
- $O(N^3)$ time in many cases
 - There are N steps and at each step the size, N^2 , proximity matrix must be updated and searched
 - Complexity can be reduced to $O(N^2 \log(N))$ time for some approaches

- Suppose we have five items, a, b, c, d, and e.
- We want to perform hierarchical clustering on five instances following an agglomerative approach
- First: we compute the distance or similarity matrix
- D_{ij} is the distance between instance “i” and “j”

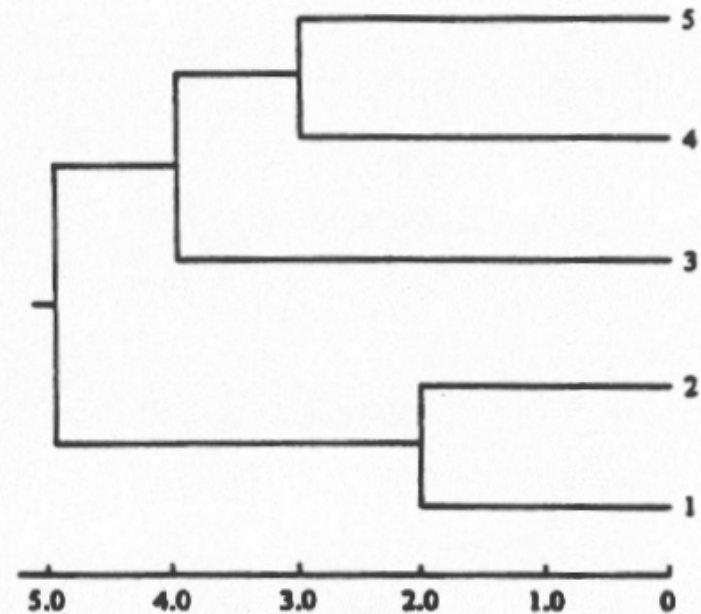
$$D = \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}$$

- Group the two instances that are closer
- In this case, a and b are the closest items ($D_2, l=2$)
- Compute again the distance matrix, and start again.
- Suppose we apply single-linkage (MIN), we need to compute the distance between the new cluster $\{1,2\}$ and the others
 - $d(12)3 = \min[d13, d23] = d23 = 5.0$
 - $d(12)4 = \min[d14, d24] = d24 = 9.0$
 - $d(12)5 = \min[d15, d25] = d25 = 8.0$

- The new distance matrix is,

$$D = \begin{pmatrix} 0.0 & & & & \\ 5.0 & 0.0 & & & \\ 9.0 & 4.0 & 0.0 & & \\ 8.0 & 5.0 & 3.0 & 0.0 & \\ & & & & \end{pmatrix}$$

- At the end, we obtain the following dendrogram



Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
 - Sensitivity to noise and outliers
 - Difficulty handling different sized clusters and convex shapes
 - Breaking large clusters

- Major weakness of agglomerative clustering methods
 - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
 - can never undo what was done previously
- Integration of hierarchical with distance-based clustering
 - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
 - ROCK (1999): clustering categorical data by neighbor and link analysis
 - CHAMELEON (1999): hierarchical clustering using dynamic modeling