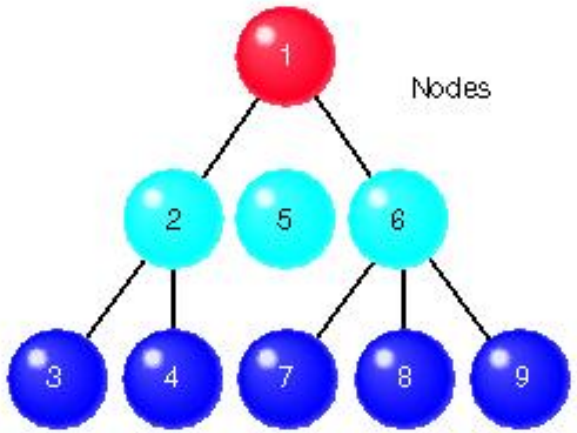
Hierarchical Clustering

Basics

- hierarchical clustering works by grouping data objects into a tree of clusters



  - clusters $\rightarrow$ subclusters $\rightarrow$ subsubclusters $\rightarrow \cdots$

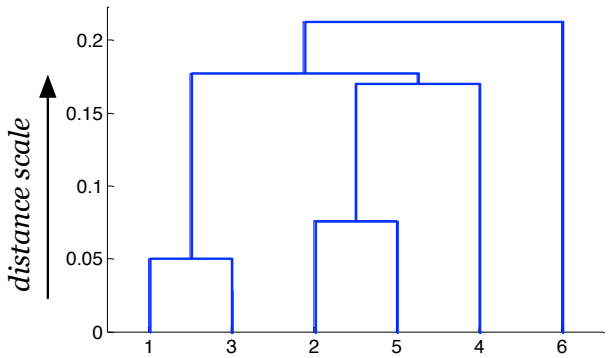
Why do we need hierarchies?

Example

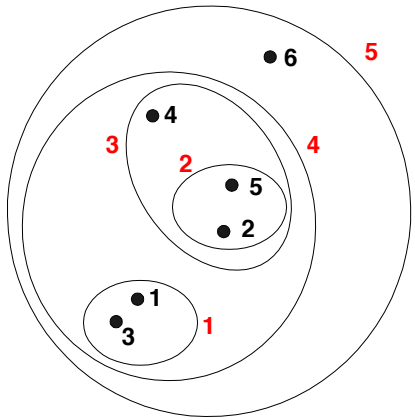
Biology, library book categorization, web directories

Basics

- use a similarity or distance (dissimilarity) matrix
- merge/split one cluster at a time
- can be graphically displayed by
  - dendrograms
  - nested cluster diagrams



Dendrogram

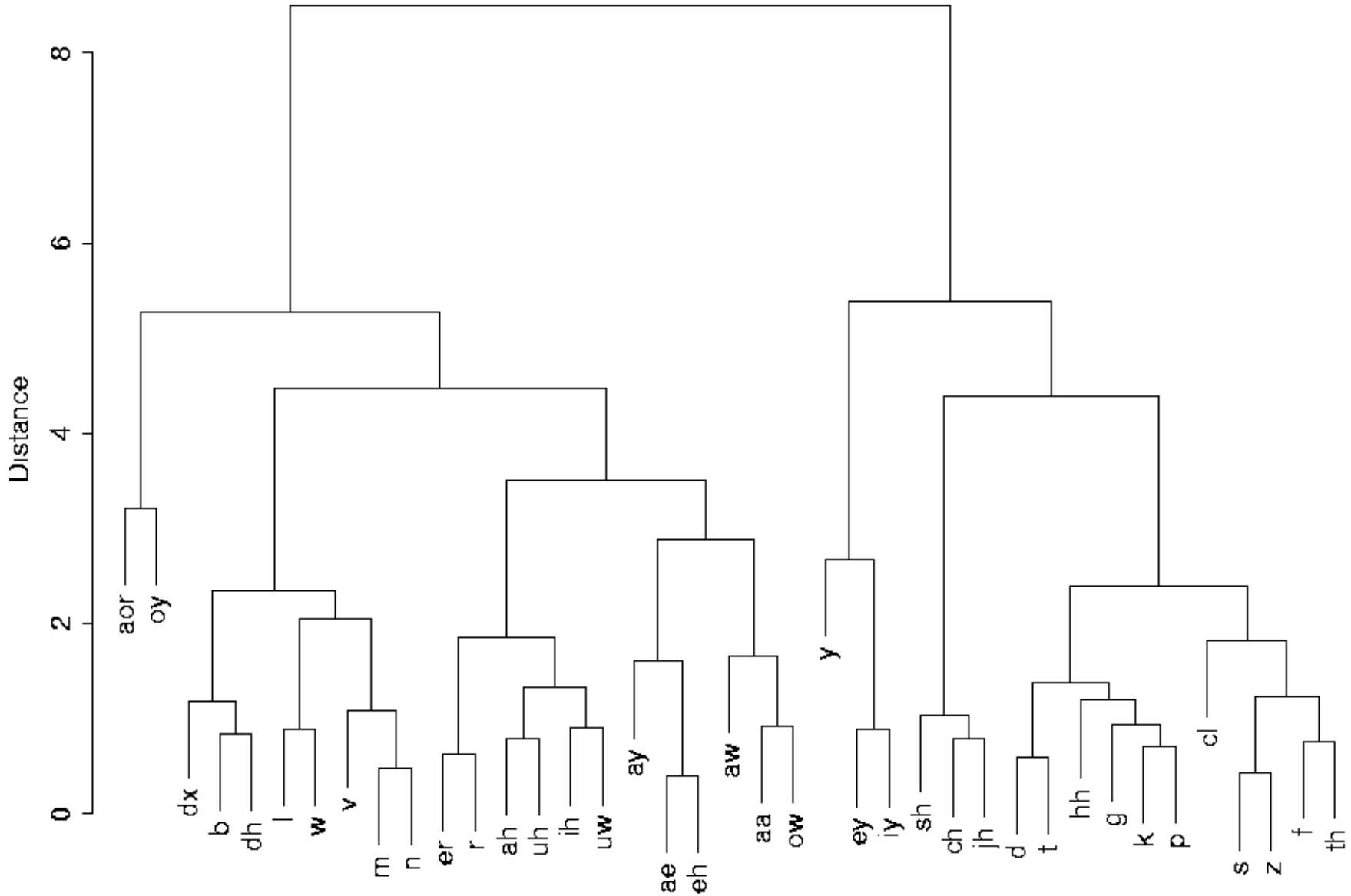


Nested cluster diagram

Dendrogram

- given any two samples x and x', at some level they will be grouped together in the same cluster
- if two samples are in the same cluster at some level $c \rightarrow$ they remain together at all higher levels ($> c$)

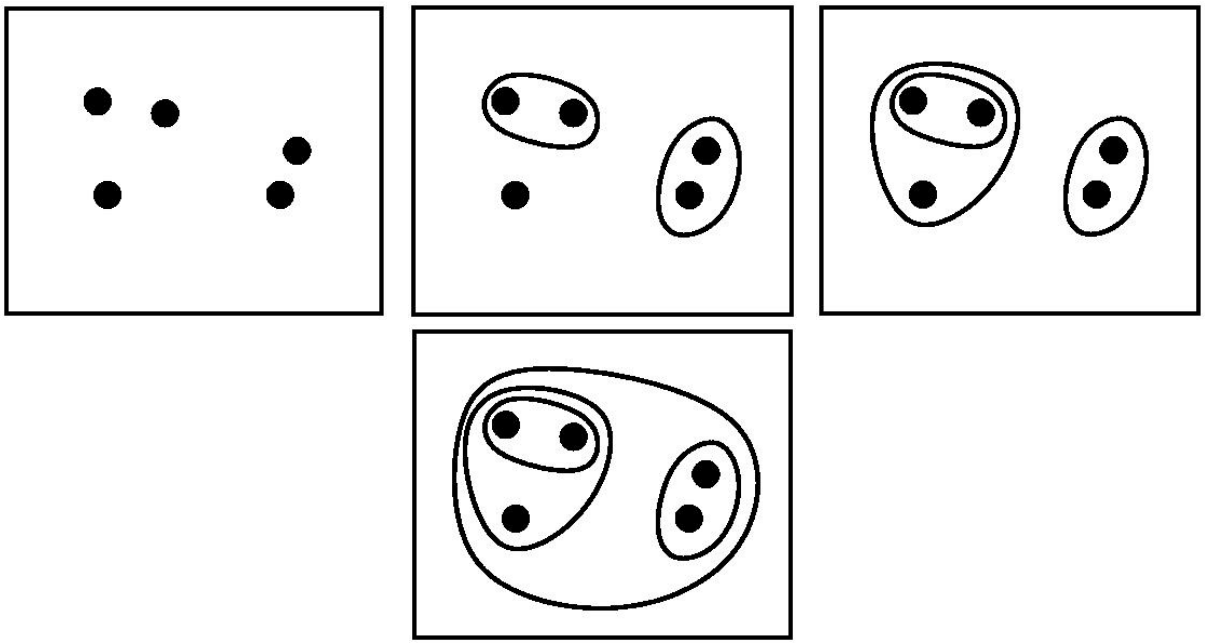
Example: Dendrogram of 39 English Sounds



Agglomerative vs Divisive

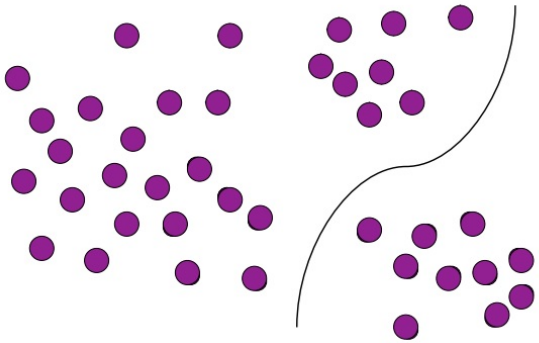
Two types of hierarchical clustering methods:

1. agglomerative (bottom-up)
   - start with the points as individual clusters
   - at each step, merge the closest pair of clusters until only one cluster (or k clusters) left

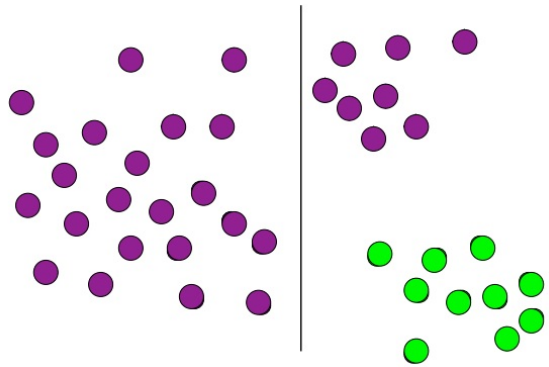


2. divisive (top-down)
   - start with one, all-inclusive cluster
   - at each step, split a cluster until each cluster contains a point (or there are k clusters)

Divisive Hierarchical Clustering: Example

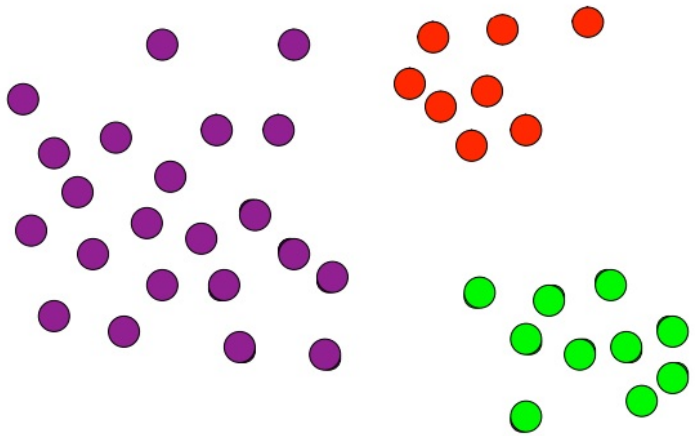


(a) 1st cut



(b) 2nd cut

final result



Agglomerative Hierarchical Clustering

input: dataset D of points
output: tree of clusters

1. compute the distance matrix over D
2. initialize each data point as a different cluster
3. **repeat**
4. merge the two closest clusters
5. update the distance matrix
6. **until** only one cluster remains

Example

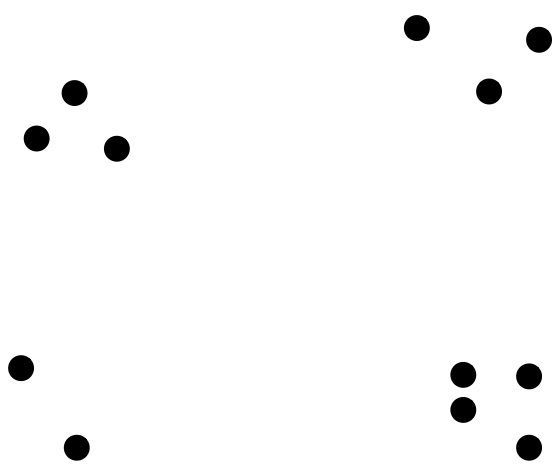
Initial setting

- start with the initial dataset D and compute the distance matrix that records distances between data points

Distance matrix

	p_1	p_2	p_3	...	p_N
p_1					
p_2					
p_3					
...					
p_N					

Dataset

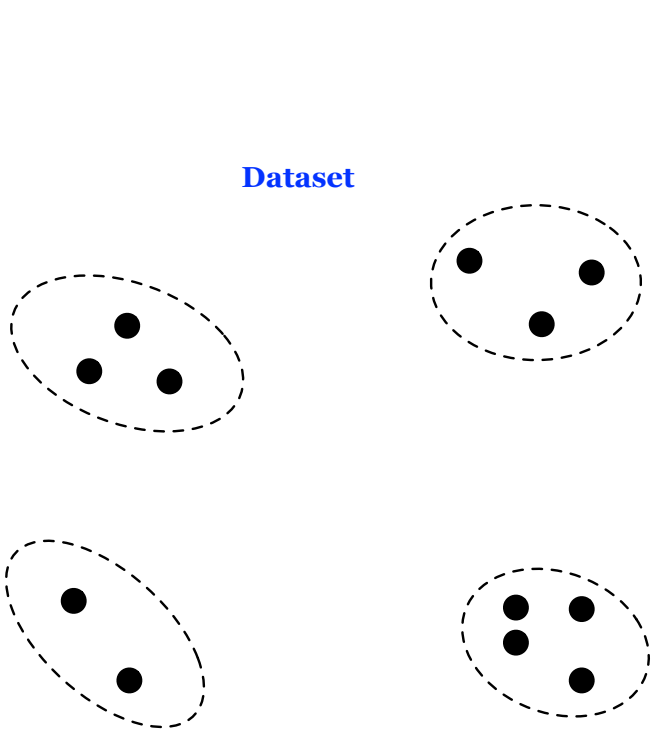


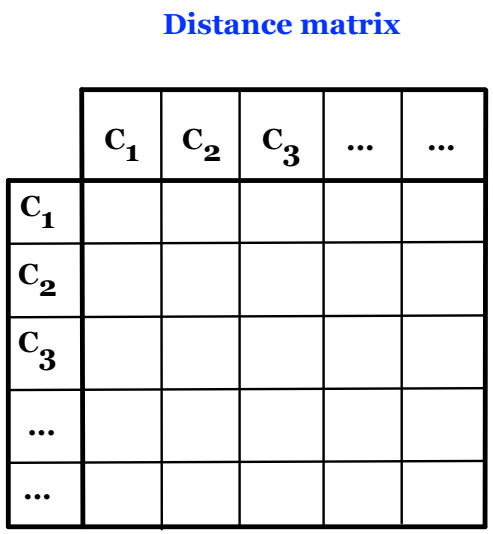
Dendrogram

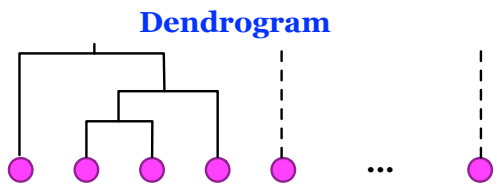
Example...

After some steps

- several clusters have been formed
- also a new distance matrix has been computed, which records distances between clusters



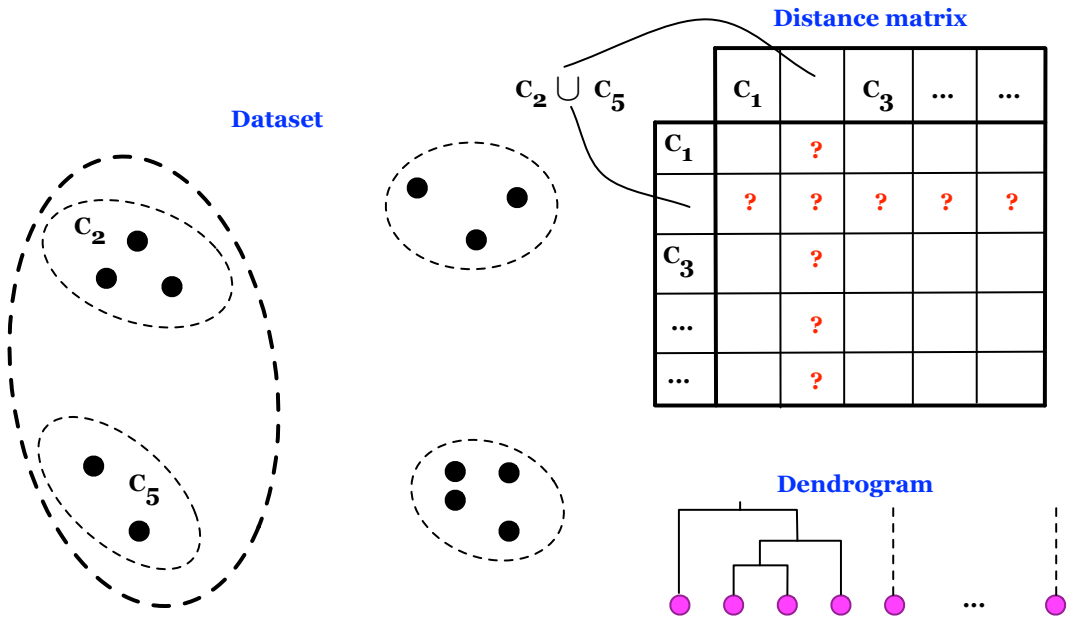




Example...

Merging two closest clusters

- suppose we merge clusters C_2 with C_5 in the figure below



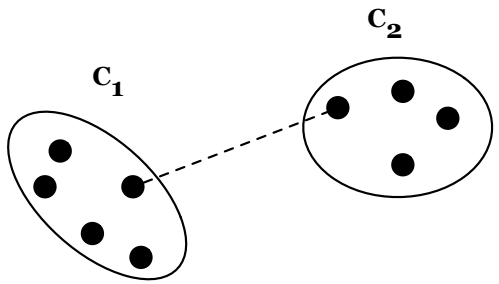
How to update the distance matrix?

Distance between Clusters

- involve computing distances between clusters
  - each cluster is a set of points
- different definitions of the distance between clusters leads to different clustering behavior
- we will explore the following distances:
  - single-link distance
  - complete-link distance
  - group average distance

Single-Link Distance

- let C_1 and C_2 be two clusters



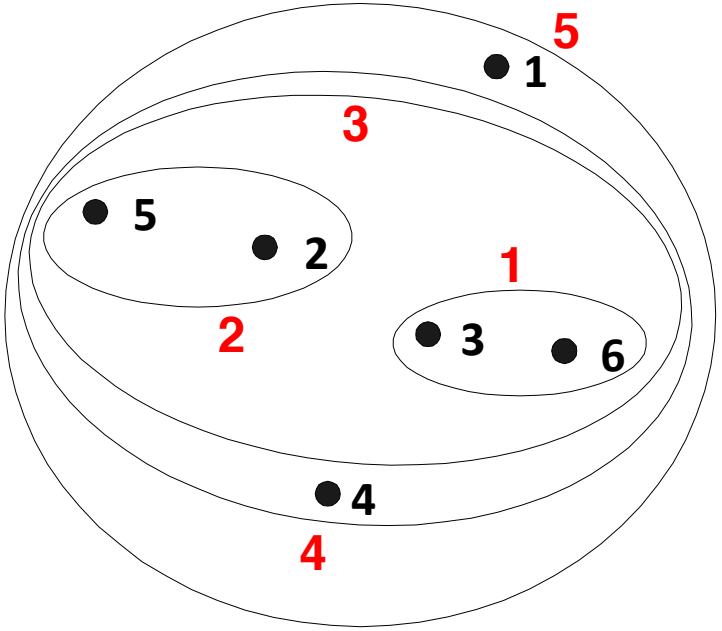
- single-link distance: the minimum distance between any object in C_1 and any object in C_2

$$dist_{single}(C_1, C_2) = \min_{x_1, x_2}\{dist(x_1, x_2) \mid x_1 \in C_1, x_2 \in C_2\}$$

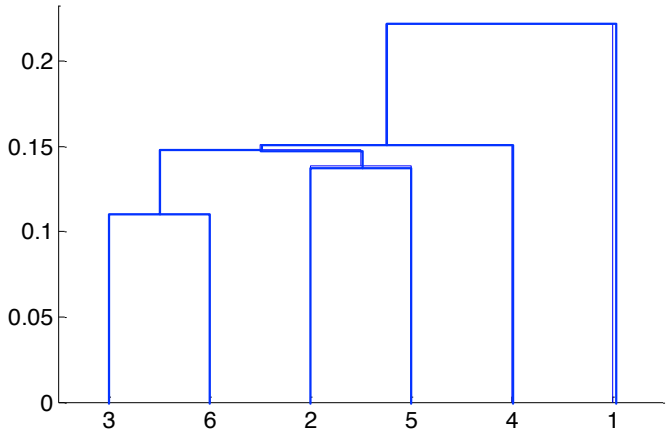
  - i.e., defined by the most similar pair of objects
- depends on a distance metric, such as Euclidean distance

Example

Example



Nested cluster diagram



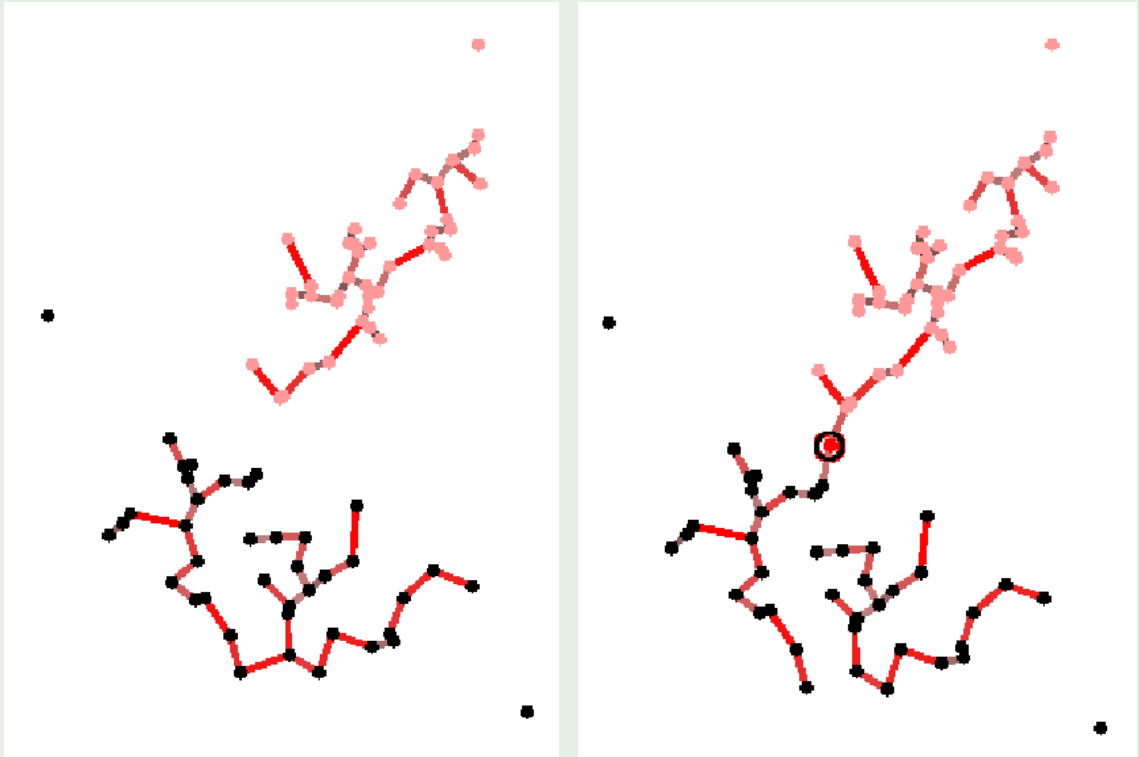
Dendrogram

- tends to produce long clusters

Limitation

Sensitive to noise or slight changes in positions of the data points

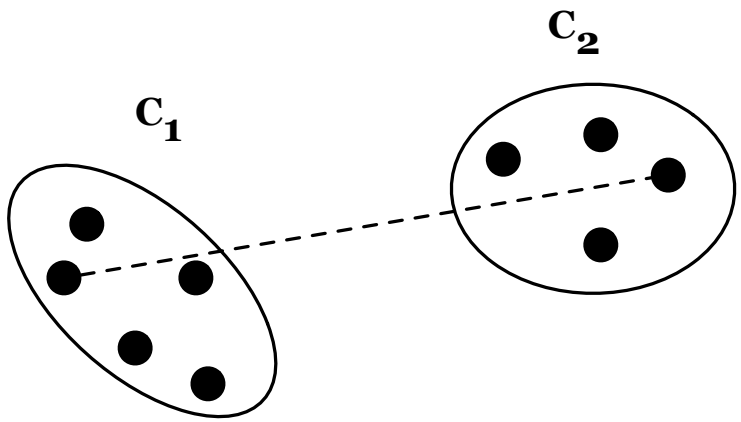
Example



- different result due to addition of a new point

- single-link

Complete-Link Distance



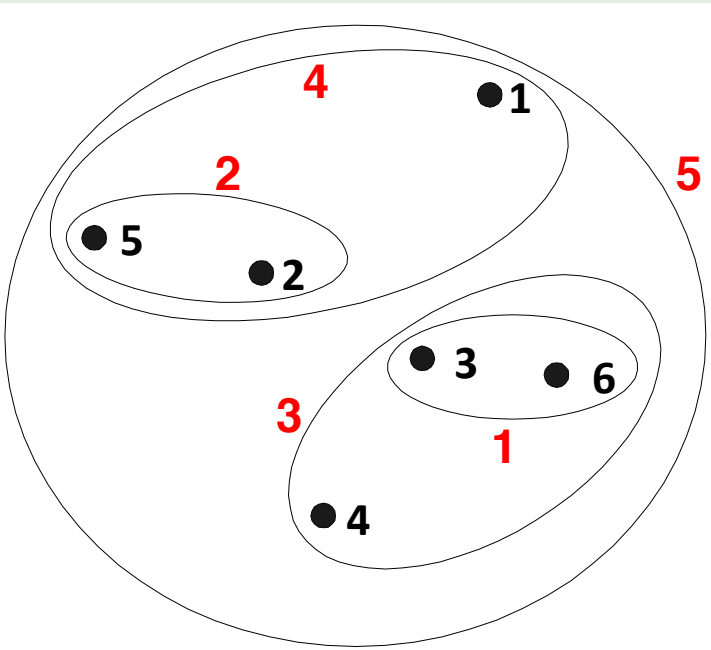
- complete-link distance: the maximum distance between any object in C_1 and any object in C_2

$$dist_{complete}(C_1, C_2) = \max_{x_1, x_2}\{dist(x_1, x_2) \mid x_1 \in C_1, x_2 \in C_2\}$$

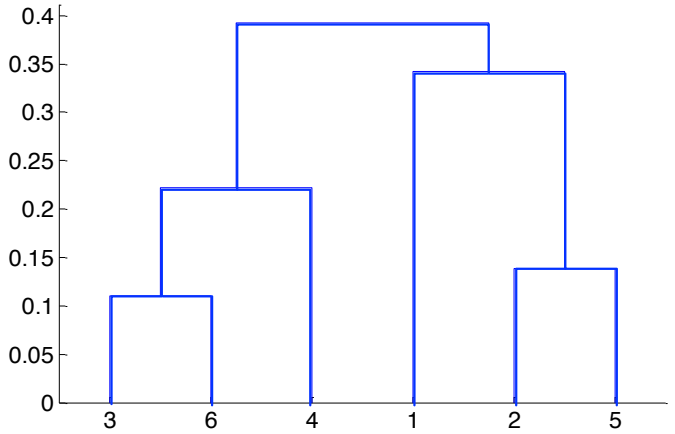
  - i.e., defined by the most dissimilar pair of objects
- again depends on a distance metric

Example

Example



Nested cluster diagram

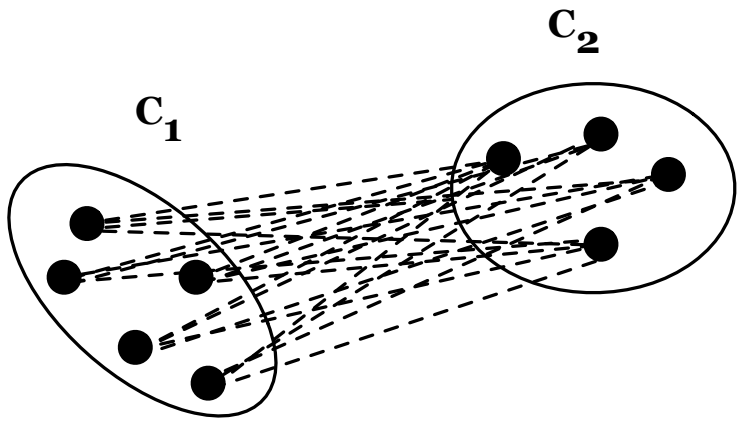


Dendrogram

- at each iteration, the size (largest diameter) of the partition is increased as little as possible
  - tends to produce very tight clusters
  - problematic if the true clusters are elongated

Group Average Distance

- let C_1 and C_2 be two clusters, with cardinality N_1 and N_2, respectively

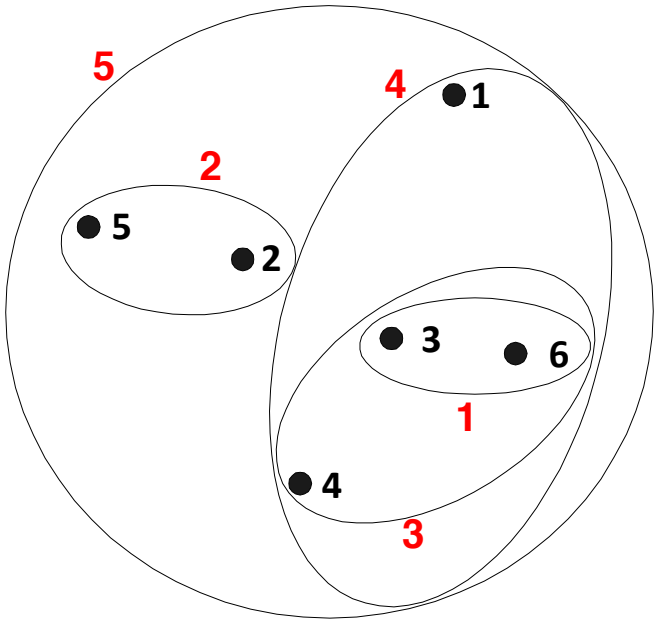


- group average distance: the average distance between any object in C_1 and any object in C_2

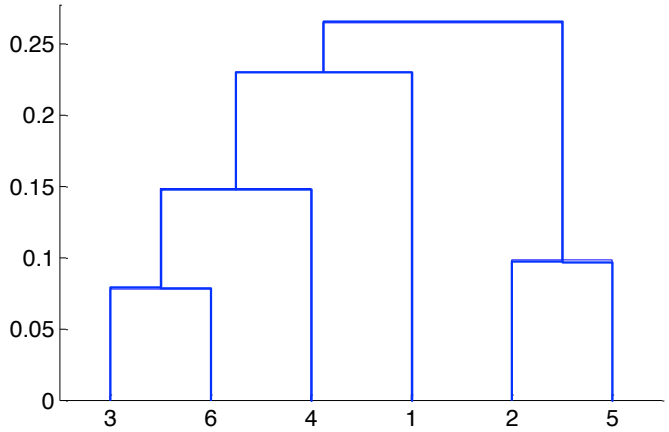
$$dist_{avg}(C_1, C_2) = \frac{1}{N_1 \cdot N_2} \sum_{x_1 \in C_1, x_2 \in C_2} dist(x_1, x_2)$$

  - i.e., defined by all the objects in the union of the two clusters

Group Average Distance...

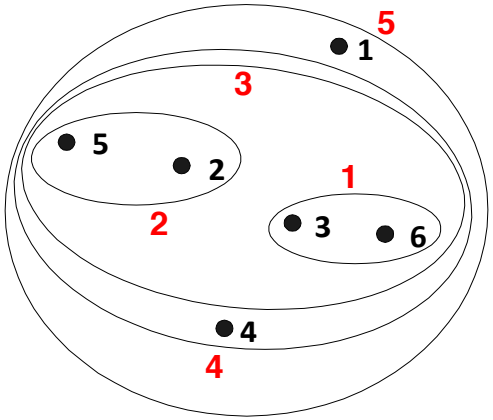


Nested cluster diagram

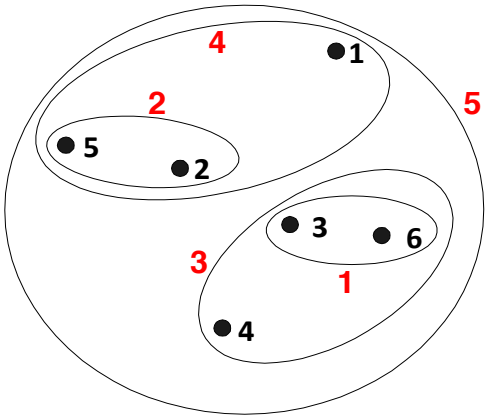


Dendrogram

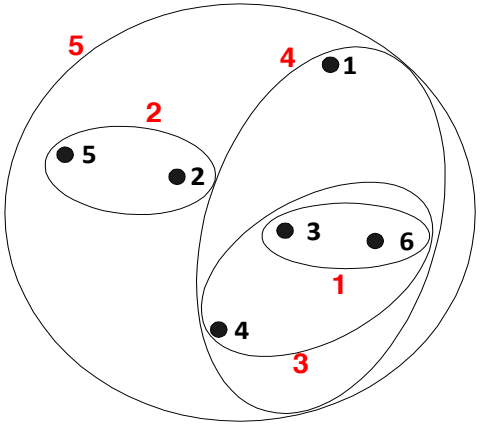
Comparison of Different Results



Single-link



Complete-link



Group average