Hierarchical Data Analysis

What is Hierarchical Clustering?

Ira A. Fulton Schools of Engineering
Arizona State University

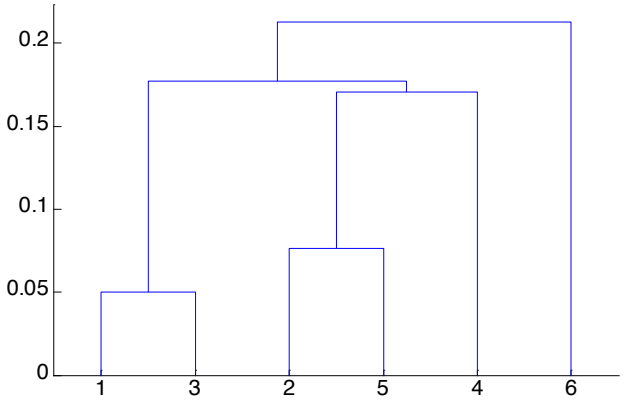
Objective

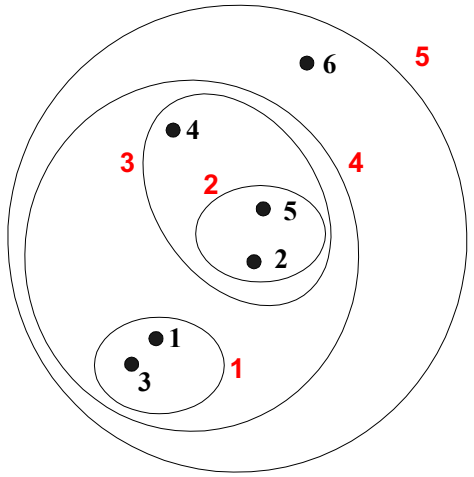
Objective

Apply methods of
hierarchical data analysis

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram (along with other options)
  - A tree-like diagram that records the sequences of merges or splits





Strengths of Hierarchical Clustering

- No assumptions on the number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendogram at the proper level
- Hierarchical clusterings may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., phylogeny reconstruction, etc), web (e.g., product catalogs) etc

Hierarchical Clustering

Agglomerative:

- Start with the points as individual clusters
- At each step, merge the closest pair of clusters until only one cluster (or **k** clusters) left

Divisive:

- Start with one, all-inclusive cluster
- At each step, split a cluster until each cluster contains a point (or there are **k** clusters)

Traditional hierarchical algorithms use a similarity or distance matrix

- Merge or split one cluster at a time

Complexity of Hierarchical Clustering

- Distance matrix is used for deciding which clusters to merge/split
- Not usable for large datasets
- At least quadratic in the number of data points