Hierarchical Data Analysis

Agglomerative Clustering

Ira A. Fulton Schools of Engineering
Arizona State University

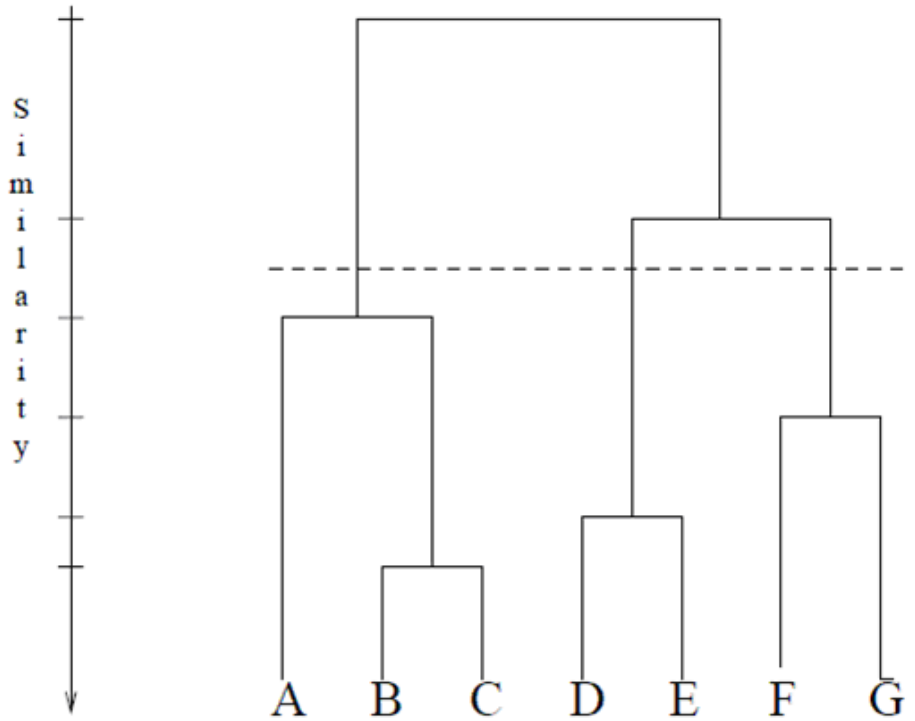
Objective

Objective

Apply methods of
hierarchical data analysis

Agglomerative Clustering Algorithm

- Most popular hierarchical clustering technique



Agglomerative Clustering Algorithm

Basic algorithm

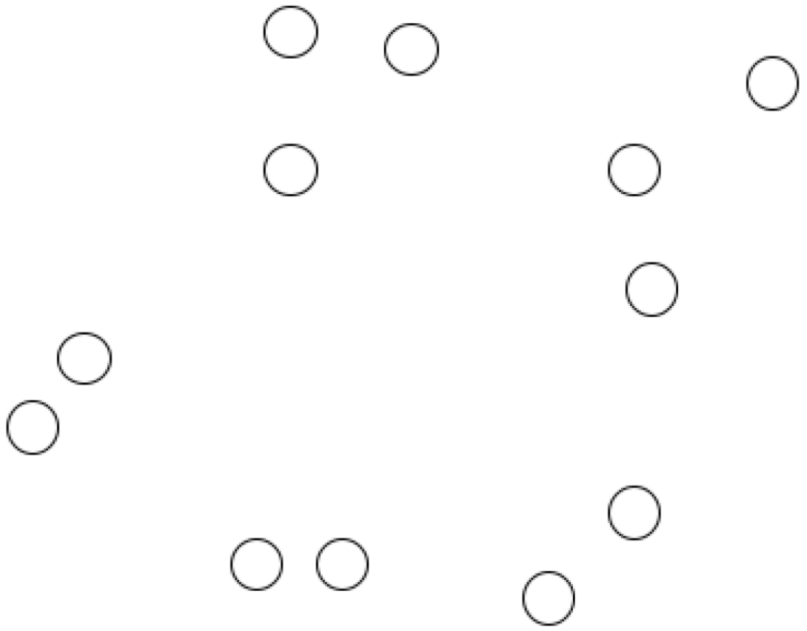
1. Compute the distance matrix between the input data points
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the distance matrix
6. **Until** only a single cluster remains

Agglomerative Clustering Algorithm

- Key operation is the computation of distance between two clusters
  - Different definitions of the distance between clusters lead to different algorithms

Hierarchical Clustering: Input/Initial Setting

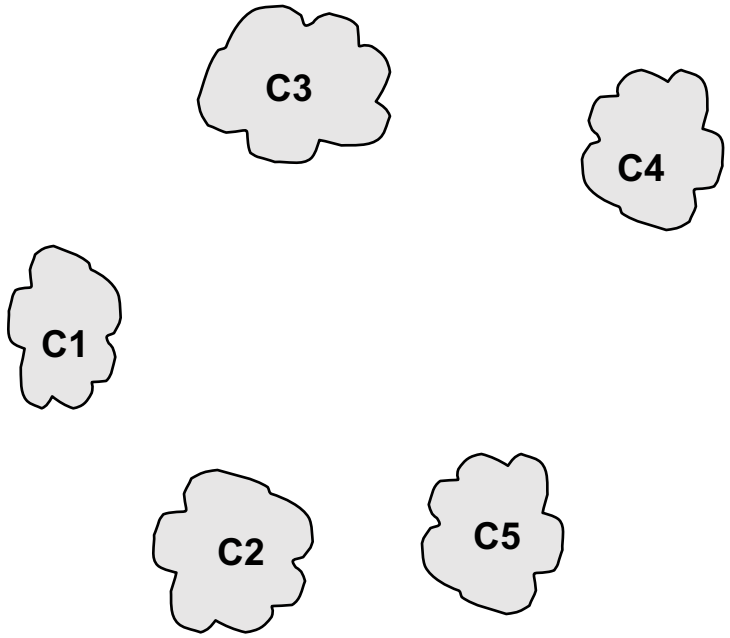
Start with clusters of individual points and a distance/proximity matrix



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

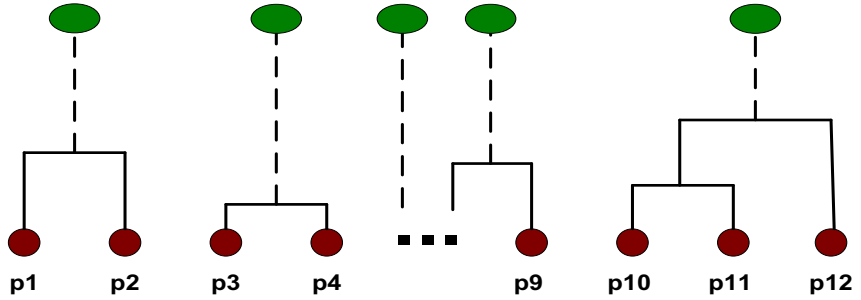
Intermediate State

After some merging steps, we have some clusters



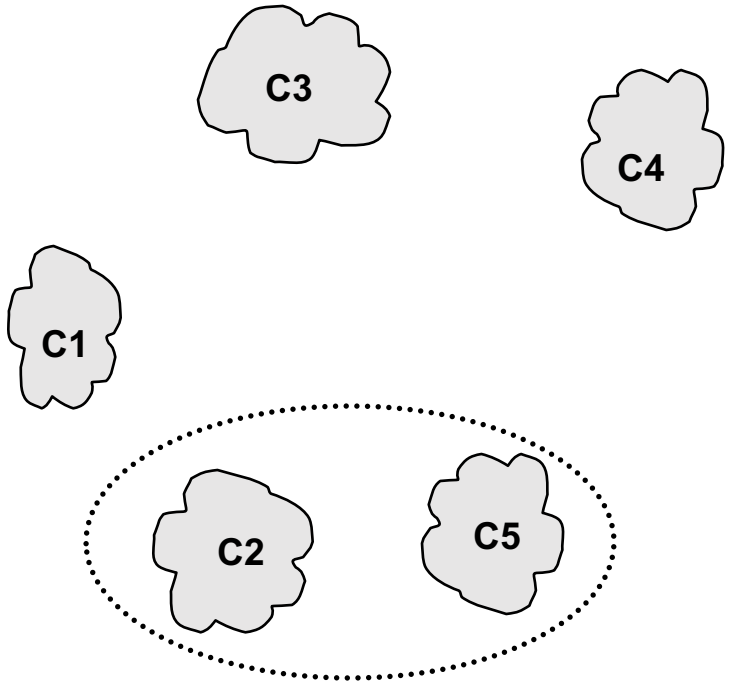
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/Proximity Matrix



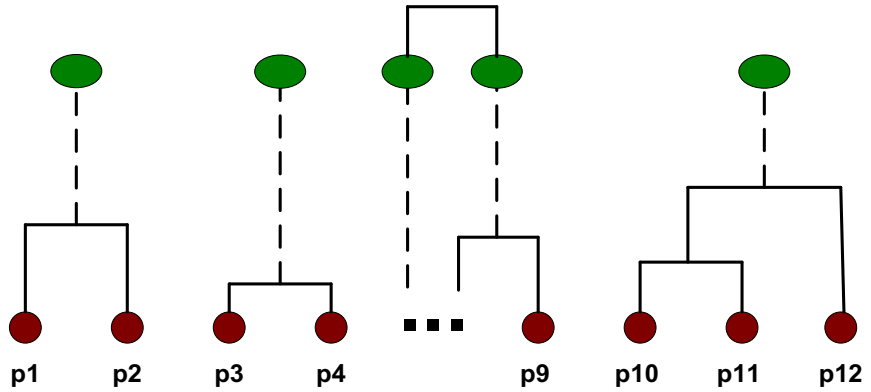
Intermediate State

Merge two closest clusters (C2 and C5) and update distance matrix



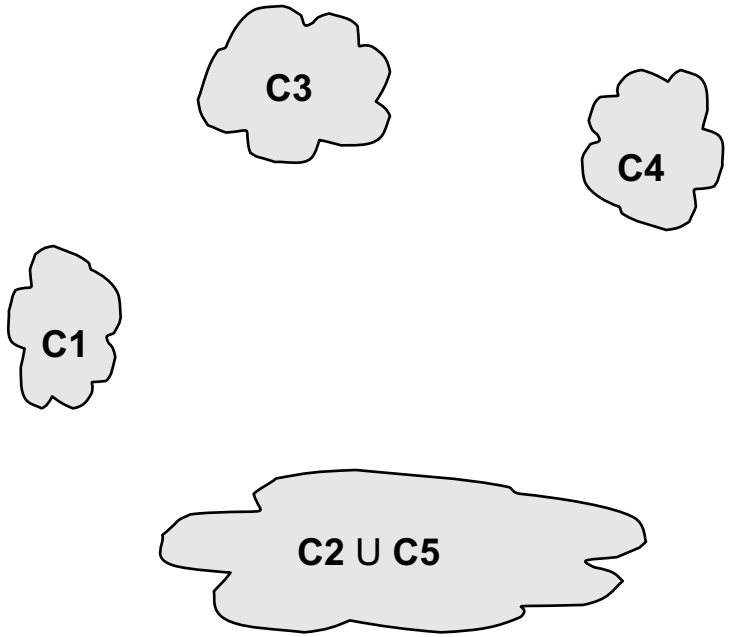
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/Proximity Matrix

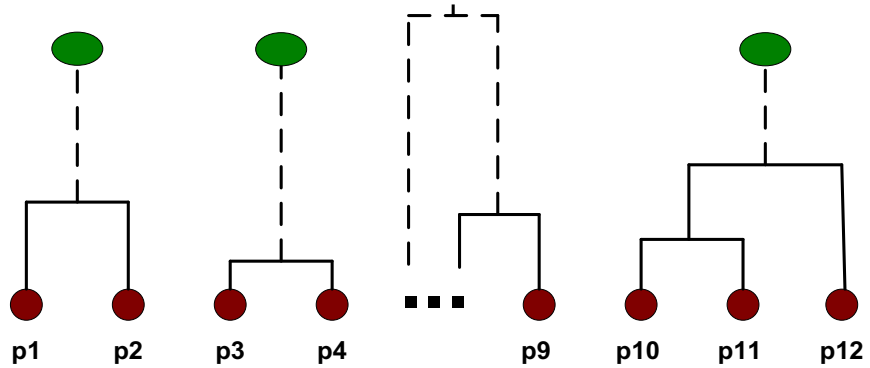


After Merging

"How do we update the distance matrix?"



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



Distance between two clusters

Each cluster is a set of points

How do we define distance between two sets of points?

- Lots of alternatives
- Not an easy task

Distance between two clusters

- Single-link distance between clusters C_i and C_j is the *minimum distance* between any object in C_i and any object in C_j
- The distance is defined by the two most similar objects

$$D_{sl}(C_i, C_j) = \min_{x,y} \{ d(x, y) | x \in C_i, y \in C_j \}$$