Clustering algorithms: agglomerative clustering

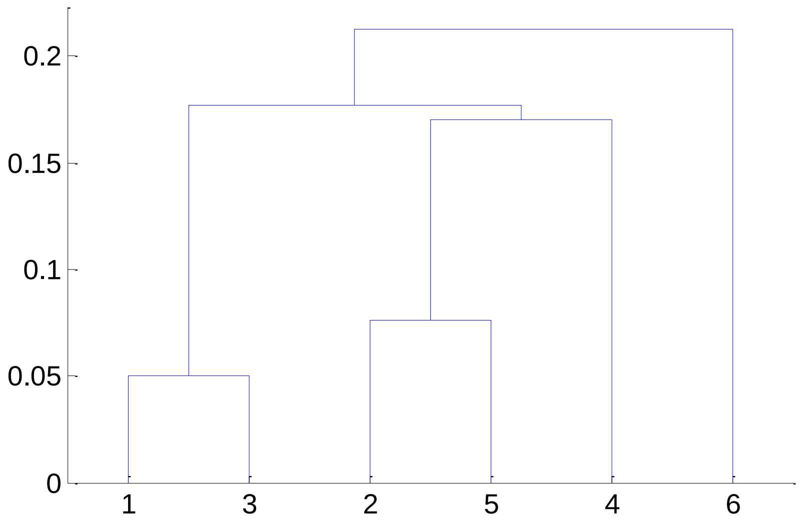
Lecture 20

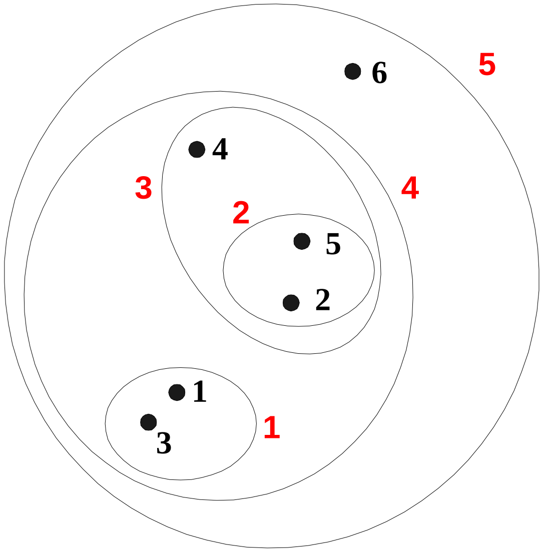
Clustering algorithms

- *K*-means clustering
- Agglomerative hierarchical clustering
- Density-based clustering

Hierarchical Clustering

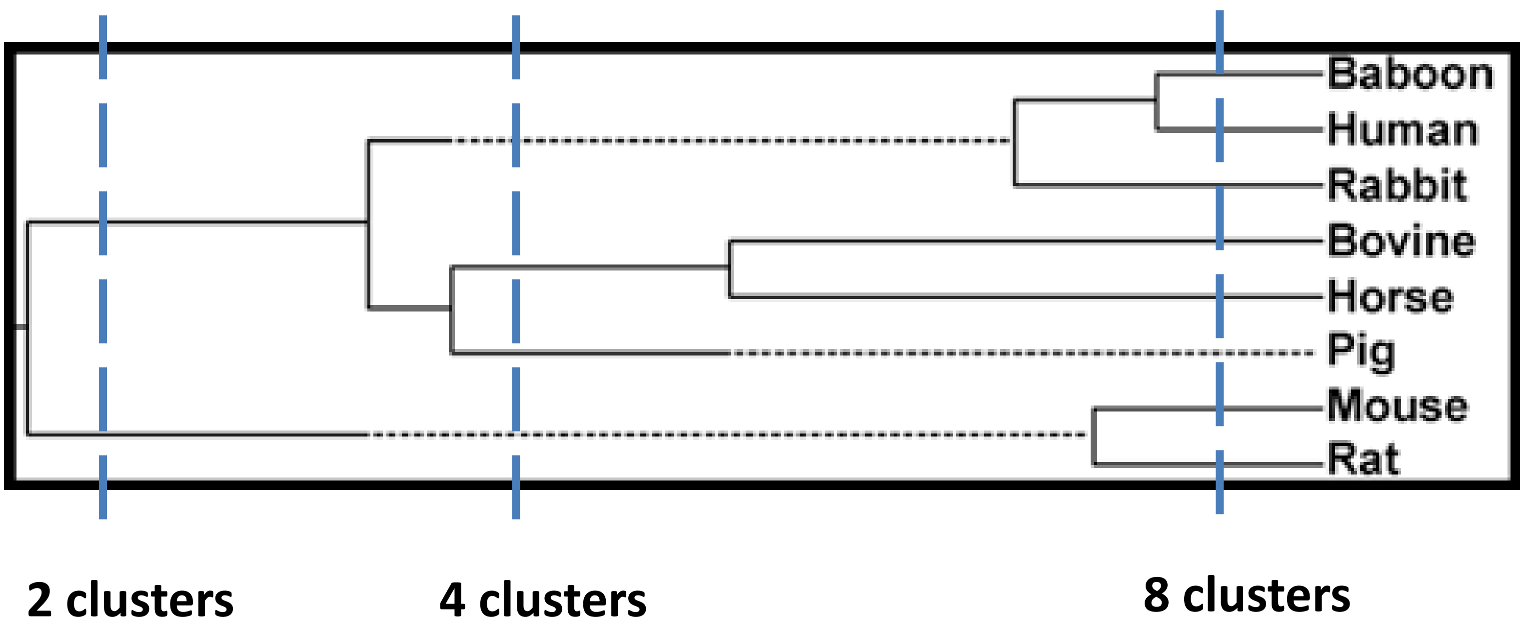
- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a *dendrogram*
  - A tree like diagram that records the sequences of merges or splits





Strengths of hierarchical clustering

- Do not have to assume any particular number of clusters
  - ‘cut’ the dendogram at the proper level

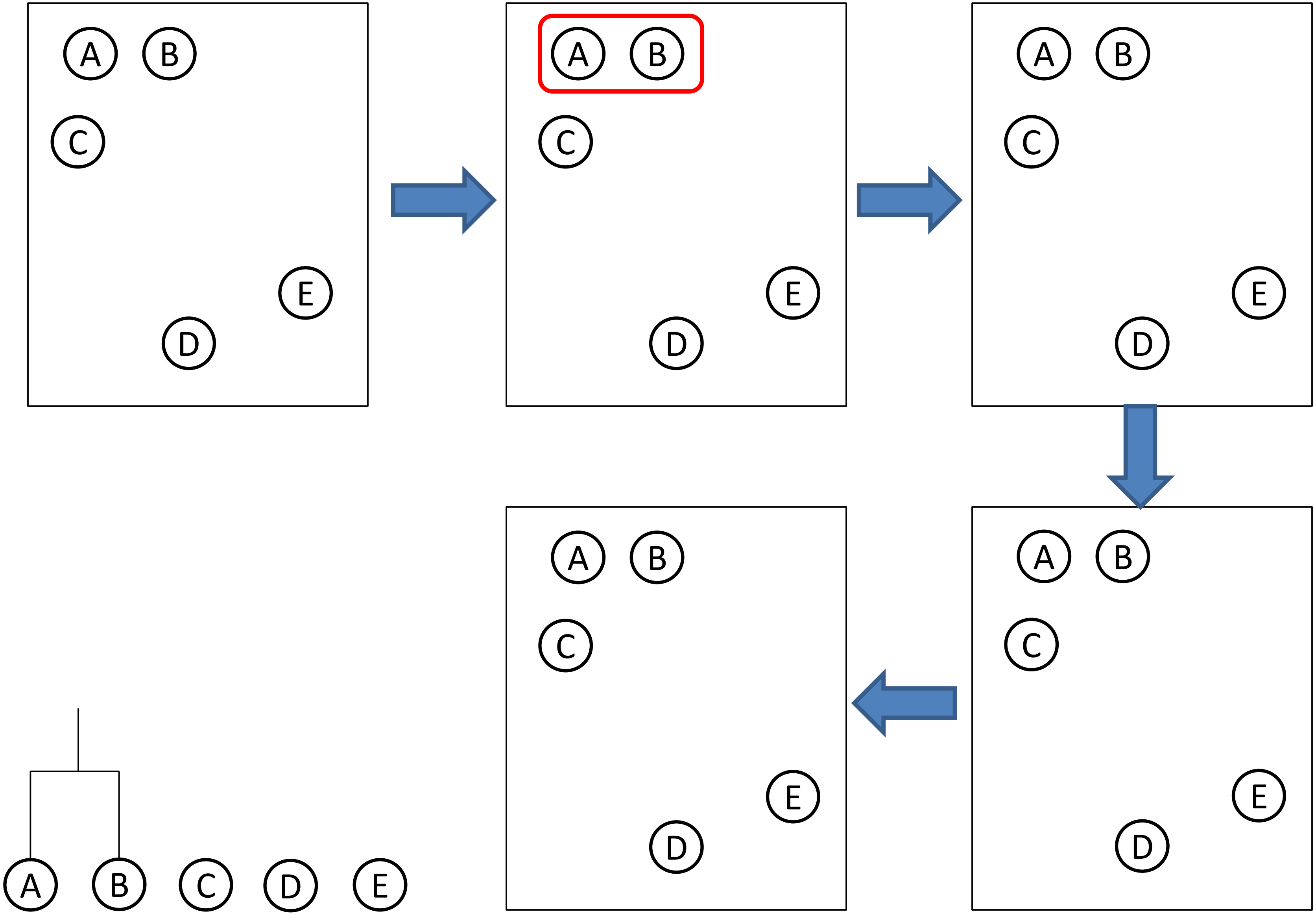


Types of hierarchical clustering

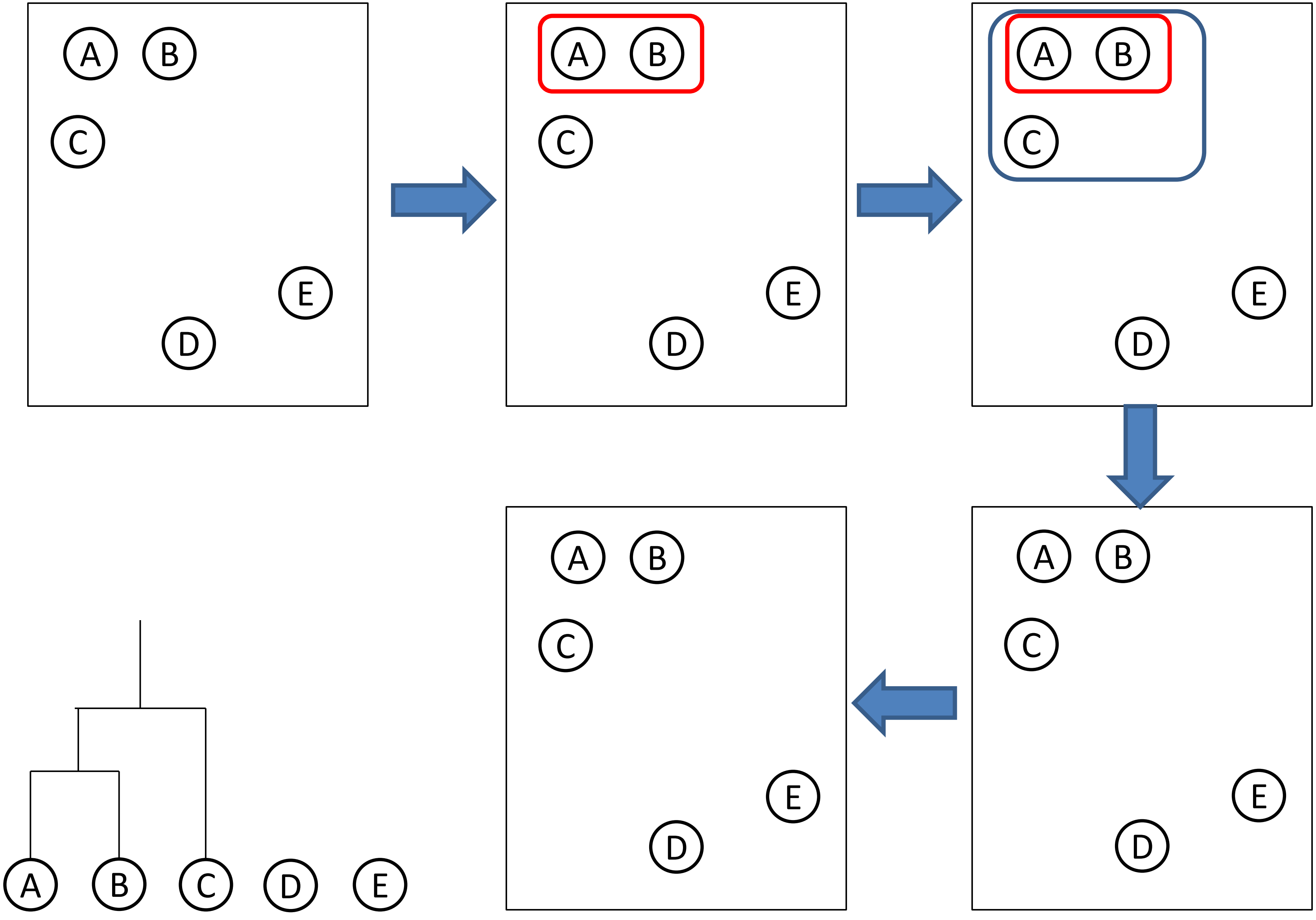
- *Agglomerative* – starts with each point as a cluster, and performs successive merges

- *Divisive* – starts with all points as a cluster and performs successive splits

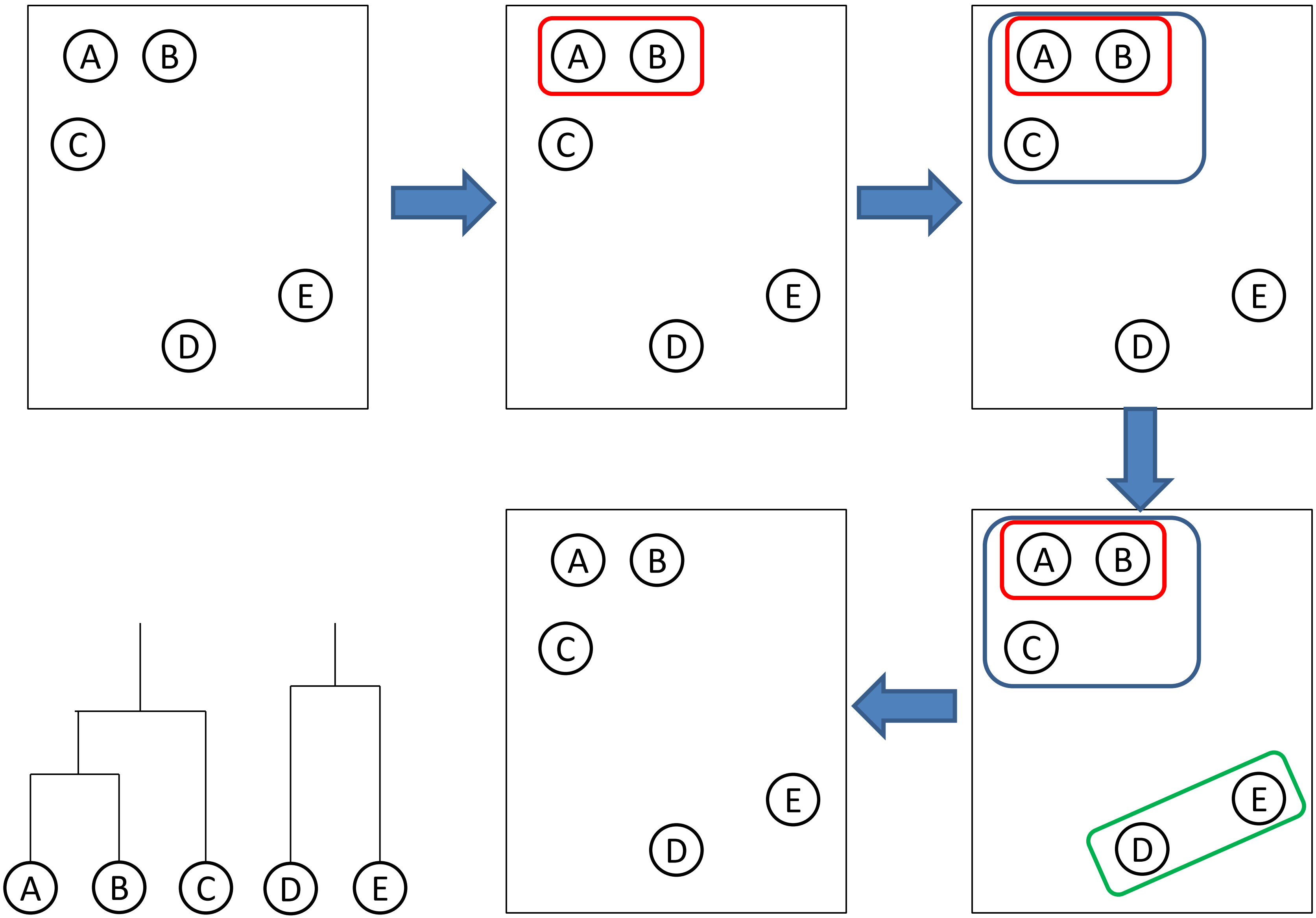
Hierarchical clustering example



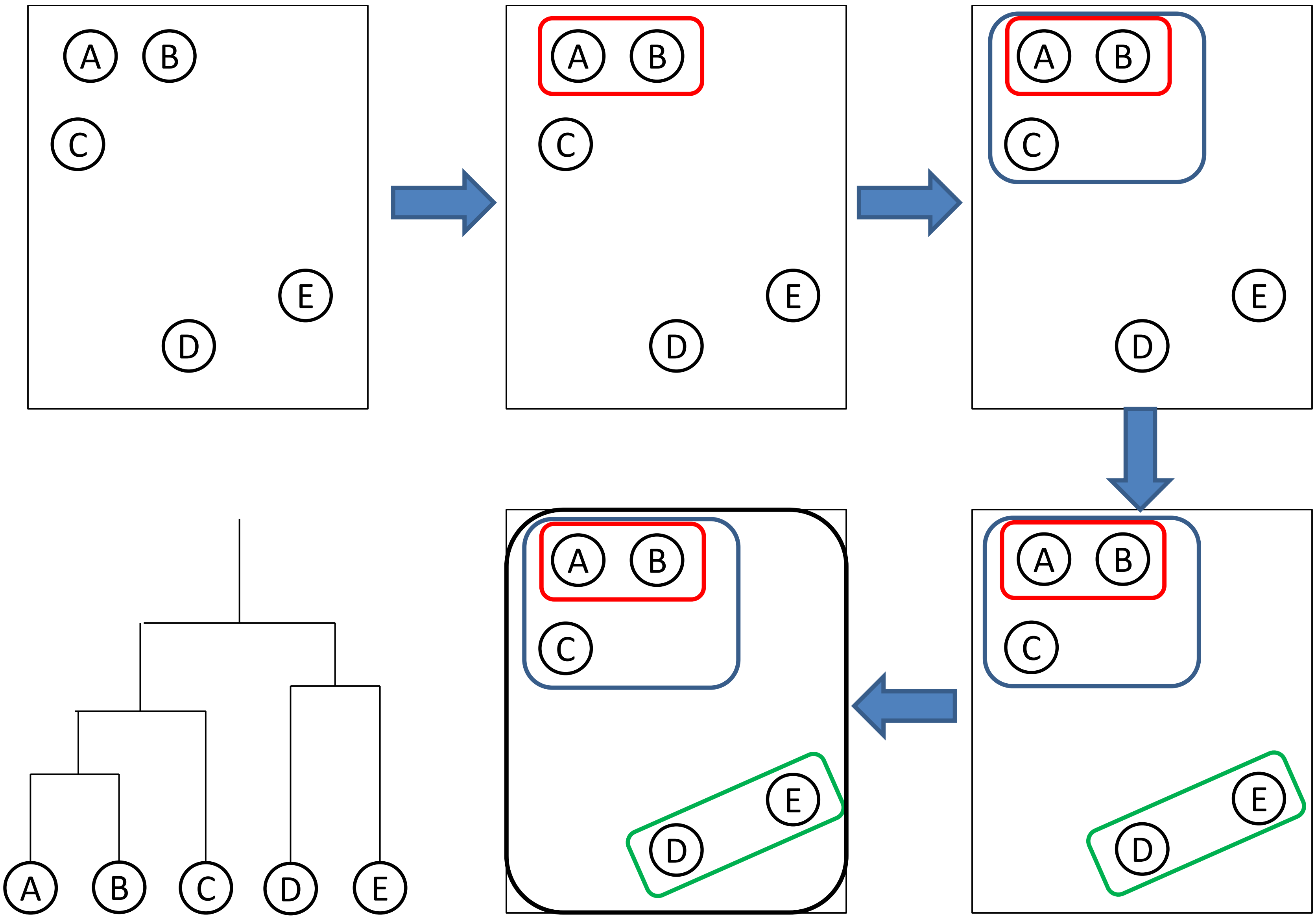
Hierarchical clustering example



Hierarchical clustering example



Hierarchical clustering example



Hierarchical Clustering Algorithm

- Start with the points as individual clusters
- At each step, merge the closest pair of clusters until only one cluster left.

Hierarchical Clustering Algorithm

Let each data point be a cluster

Compute the proximity matrix

Repeat

  Merge the two closest clusters

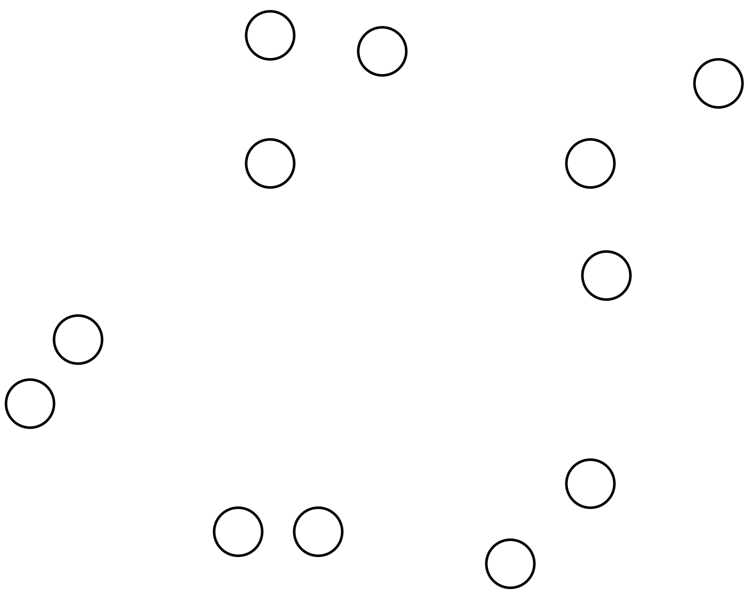
  Update the proximity matrix

Until only a single cluster remains

- Key operation is the computation of the proximity of two clusters.

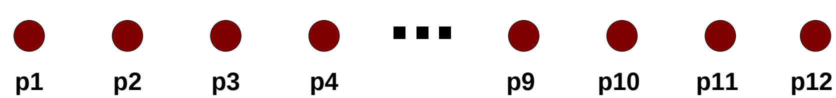
Starting Situation

- Start with clusters of individual points and a proximity matrix



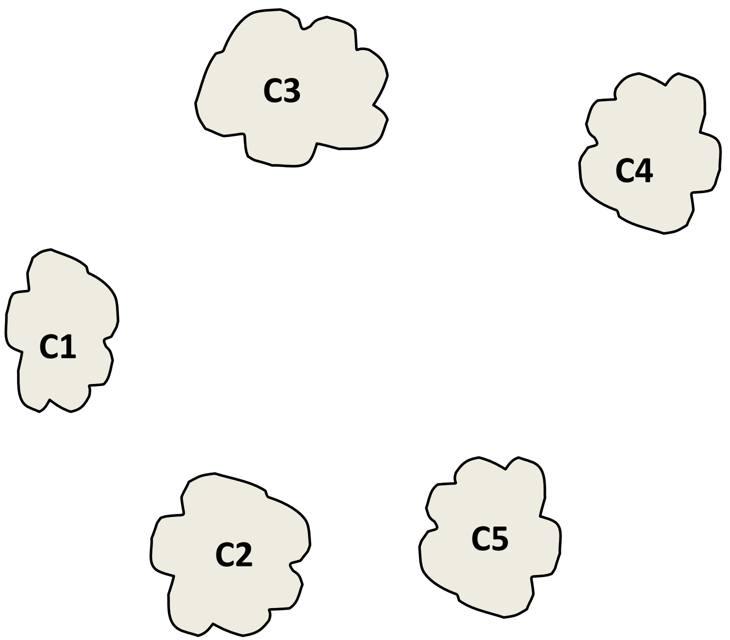
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



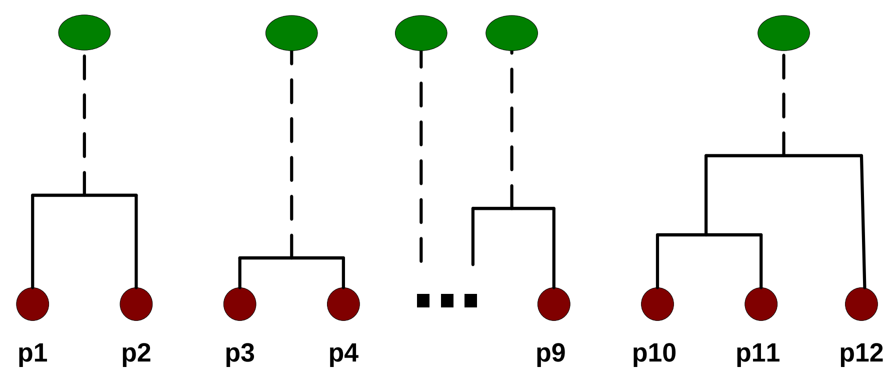
Intermediate Situation

- After some merging steps, we have some clusters



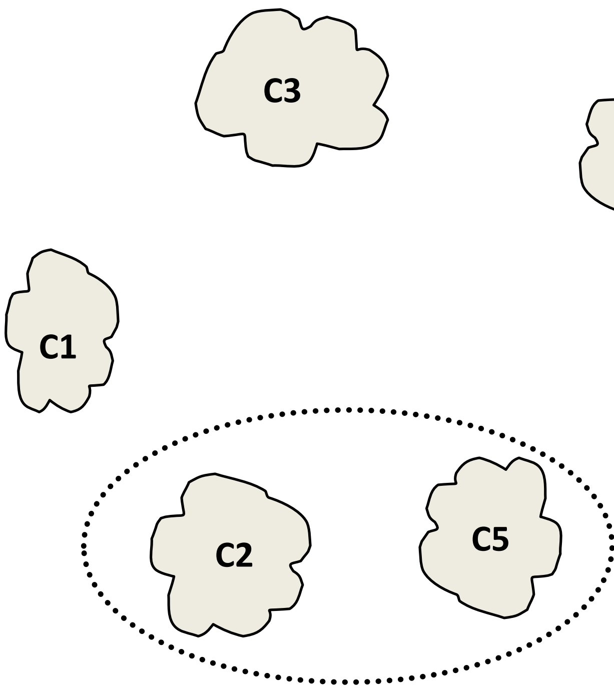
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

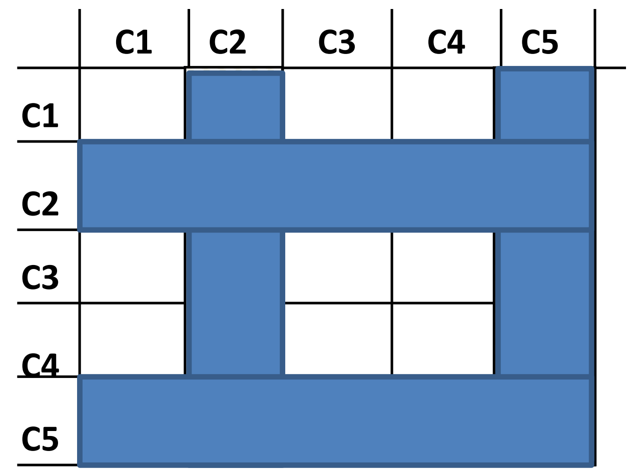
Proximity Matrix



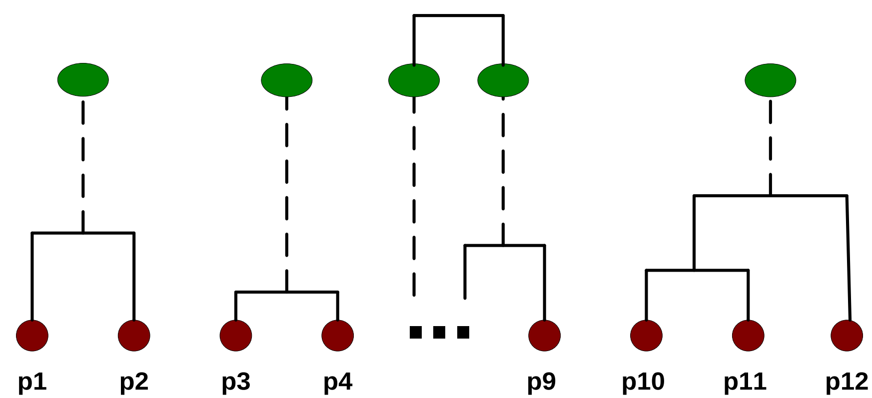
Intermediate Situation

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



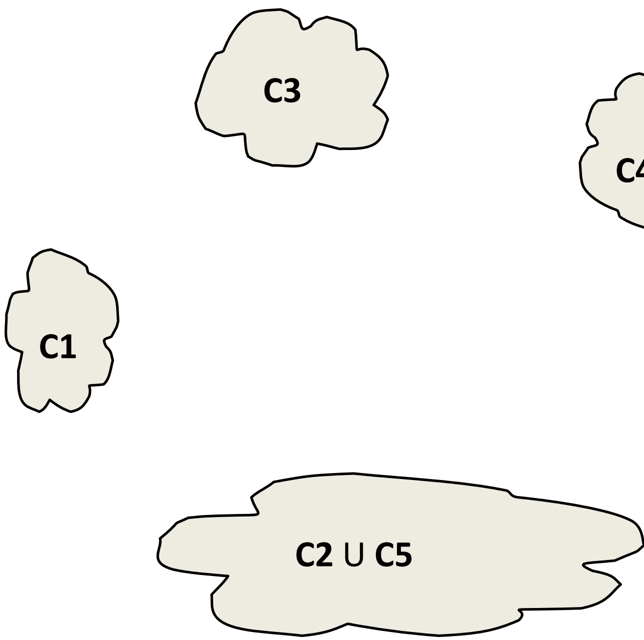


Proximity Matrix



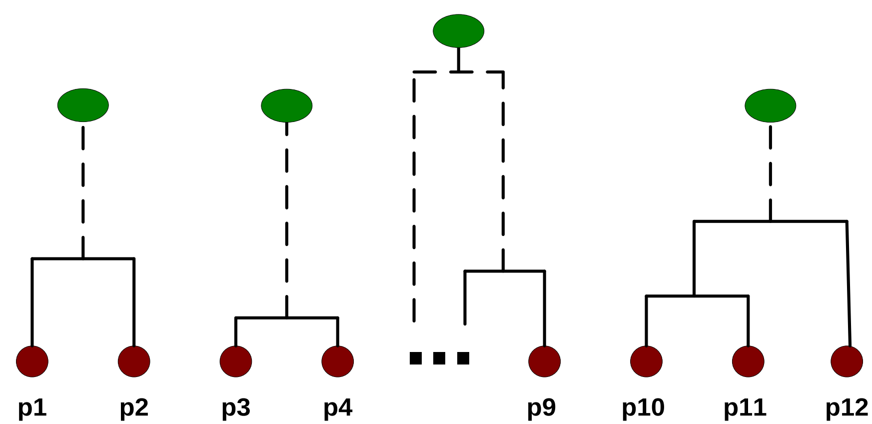
After Merging

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

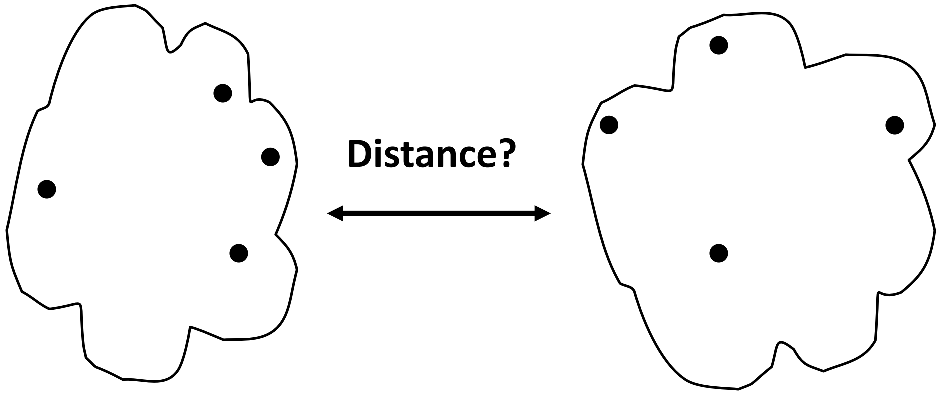


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?		?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Distance

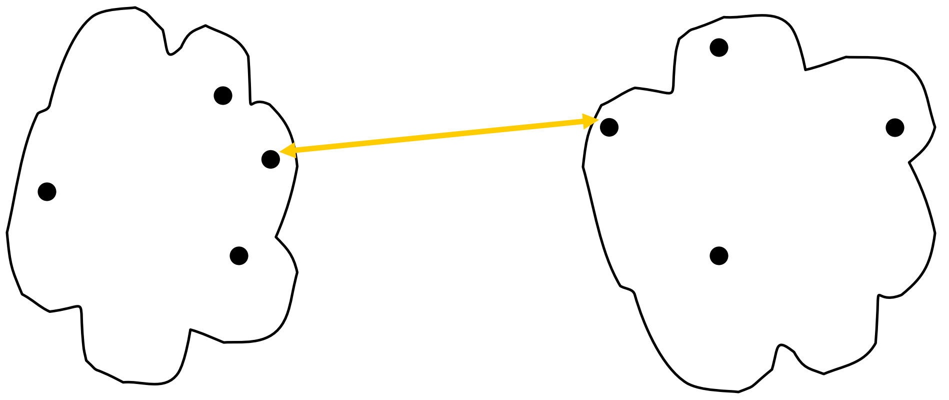


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Centroids Distance
- Group Average

Inter-Cluster Distance: MIN

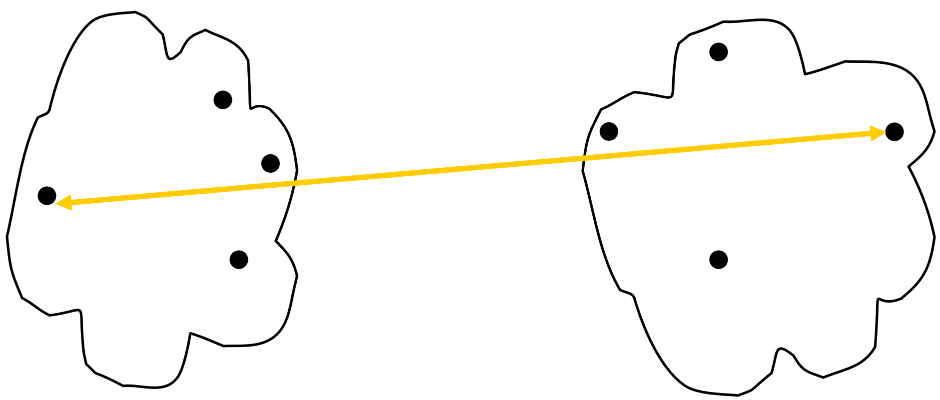


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Problem: sensitive to outliers

Inter-Cluster Distance: MAX

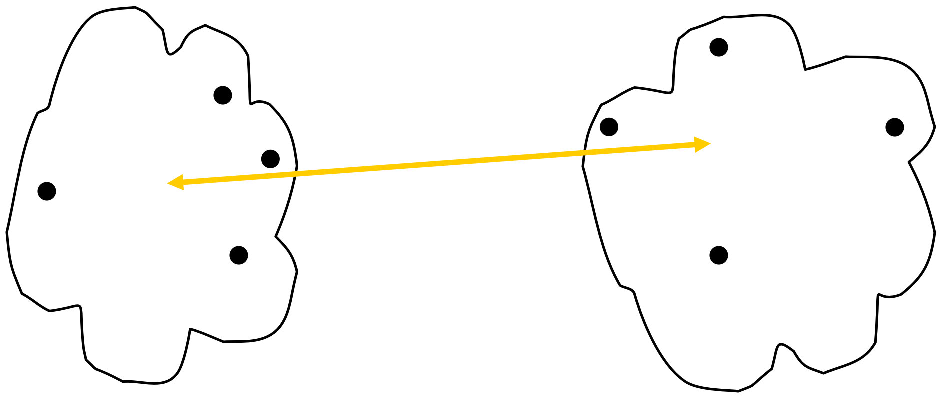


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Problem: tends to break large clusters

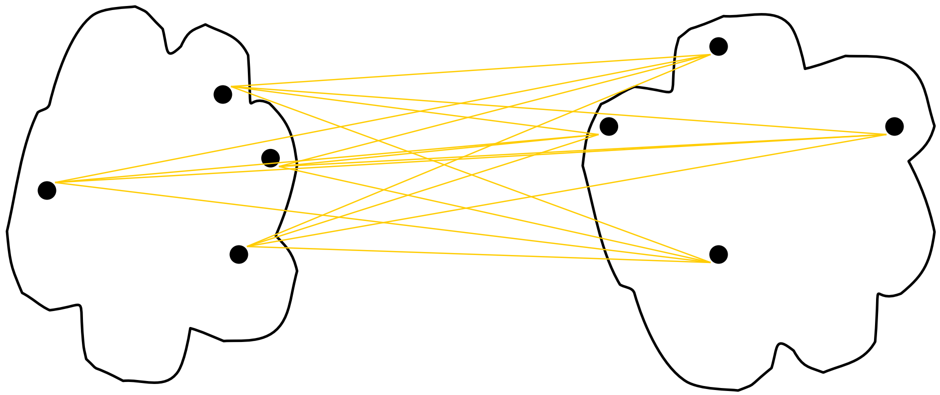
Inter-Cluster Distance: Centroid distance



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

Proximity Matrix

Inter-Cluster Distance: Group Average



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

.

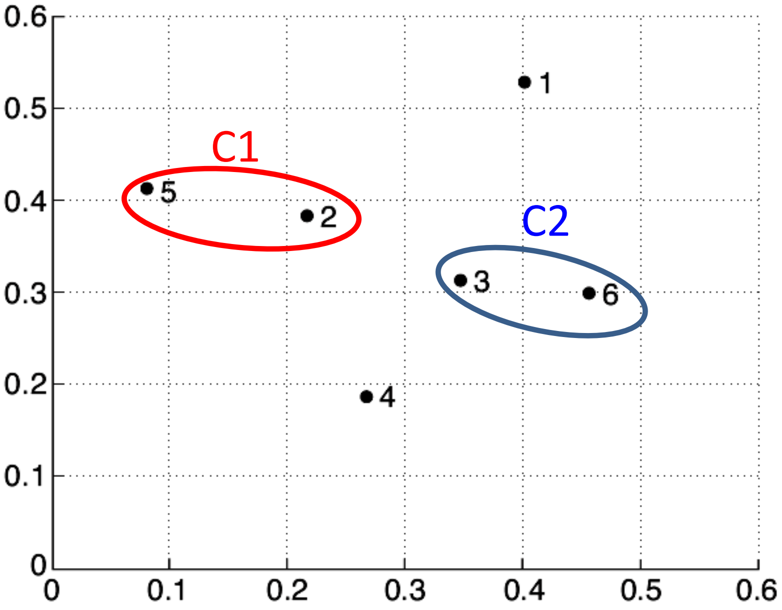
.

.

Proximity Matrix

Cluster Distance: MIN (single link)

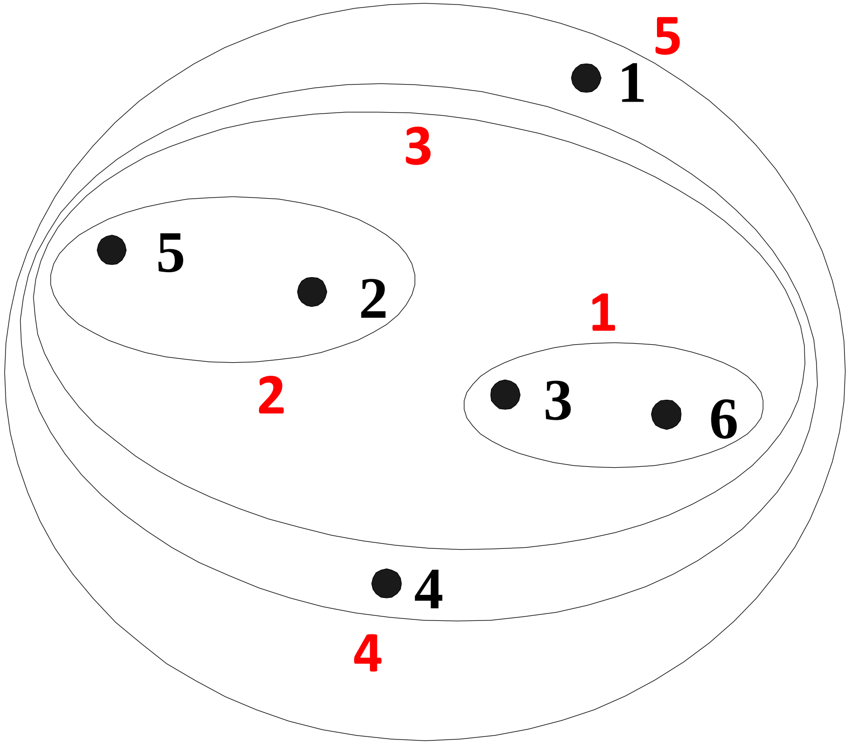
- Distance between two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points



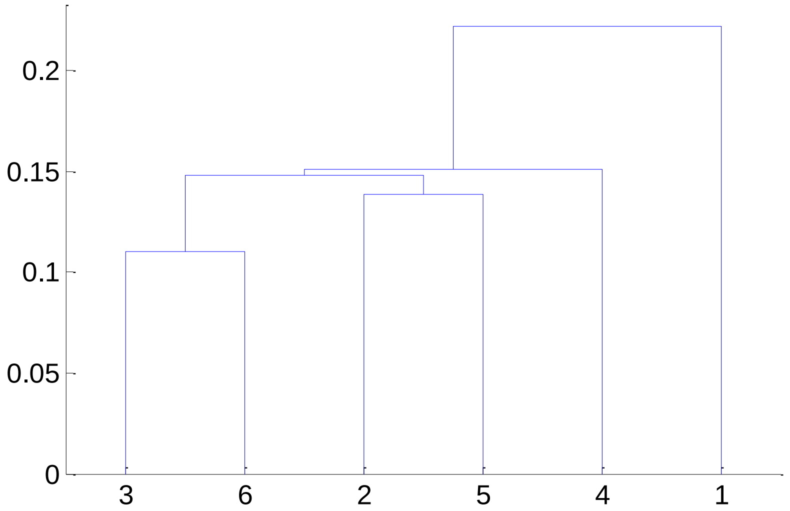
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

$d(\text{C1,C2})=0.15$

Hierarchical Clustering: MIN



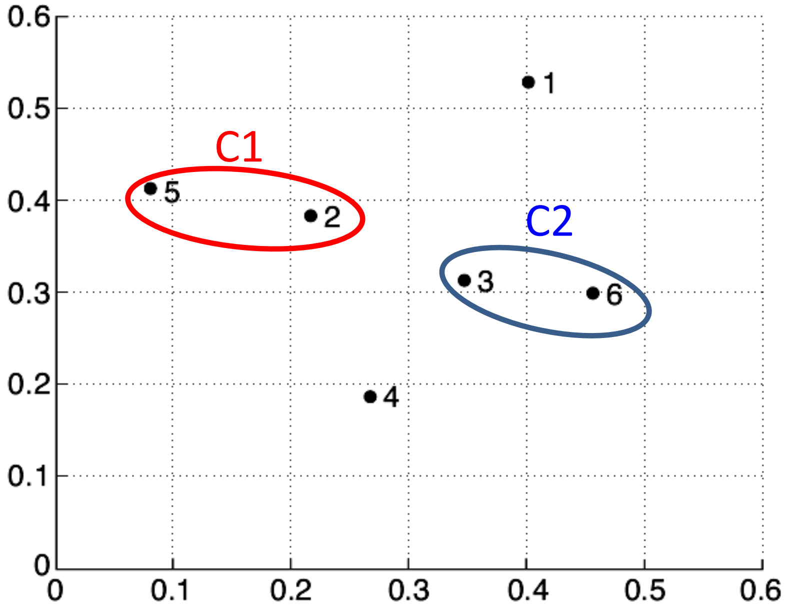
Nested Clusters



Dendrogram

Cluster Distance: MAX

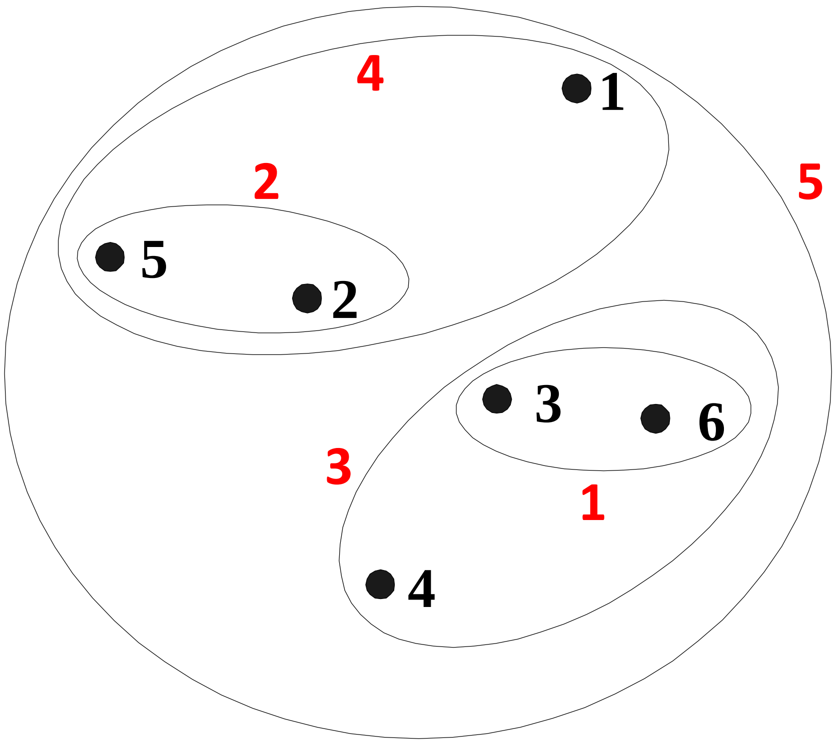
- Distance between two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by one pair of points



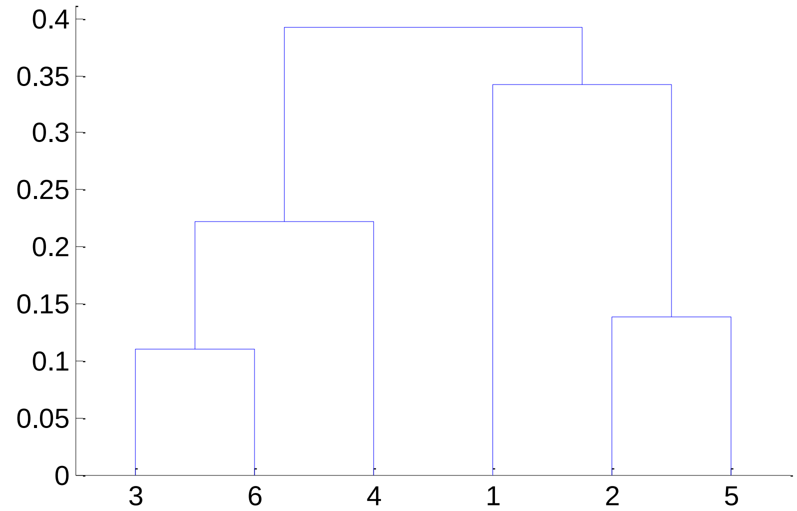
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

$d(\mathrm{C1},\mathrm{C2})=0.39$

Hierarchical Clustering: MAX



Nested Clusters



Dendrogram

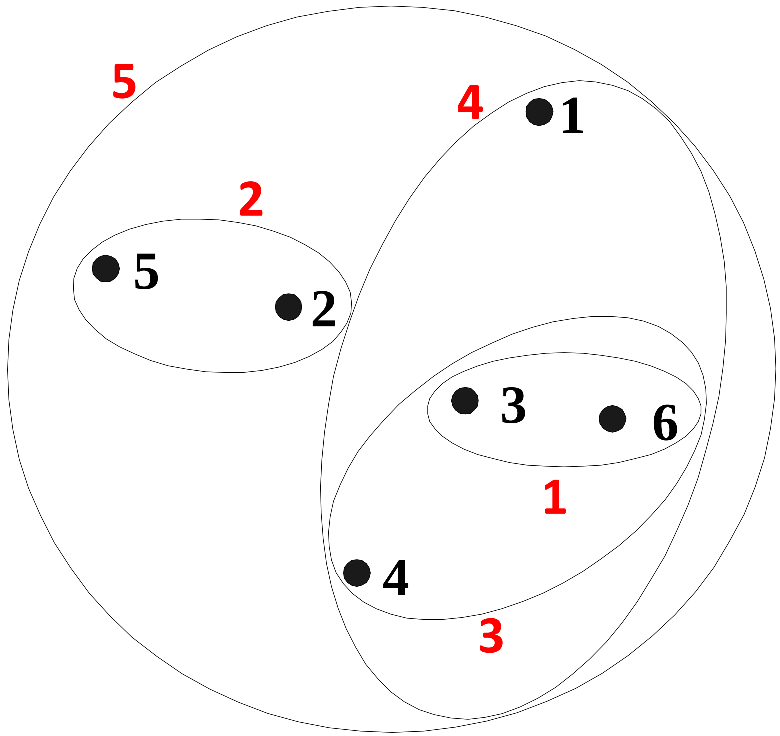
Hierarchical clustering: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

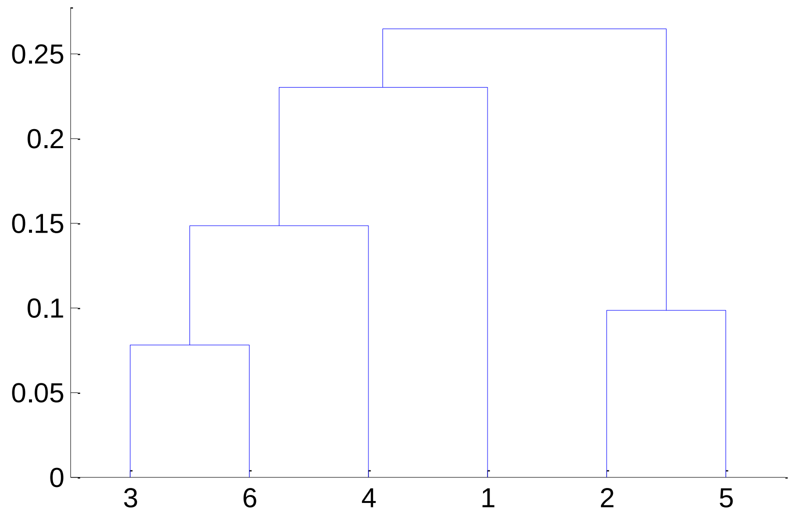
$$\mathbf{proximity}(\mathbf{Cluster_i}, \mathbf{Cluster_j}) = \frac{\sum\limits_{\substack{p_i \in \mathbf{Cluster_i} \\ p_j \in \mathbf{Cluster_j}}} \mathbf{proximity}(p_i, p_j)}{|\mathbf{Cluster_i}| * |\mathbf{Cluster_j}|}$$

  – uses all pairs of points from two clusters

Cluster distance: Group Average



Nested Clusters

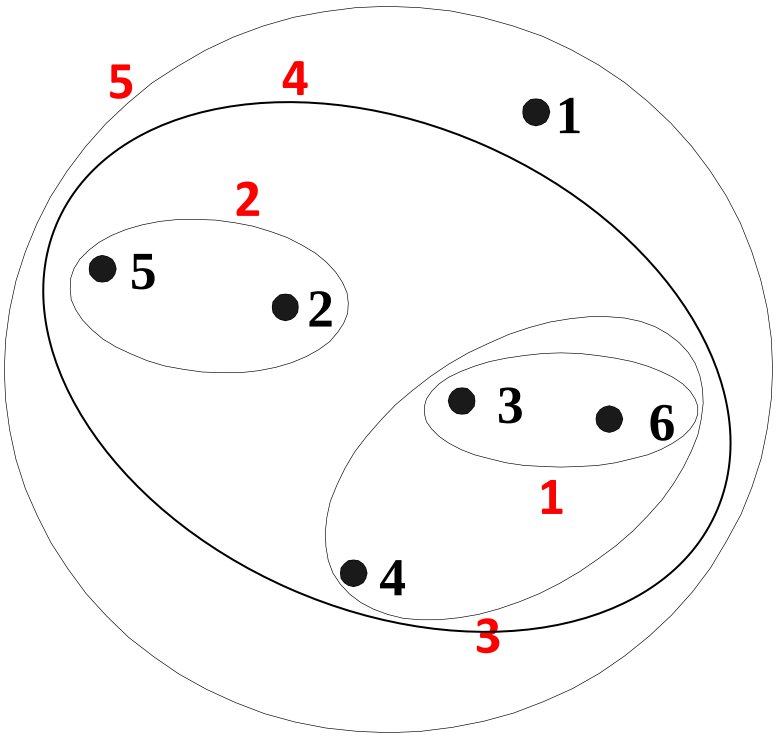


Dendrogram

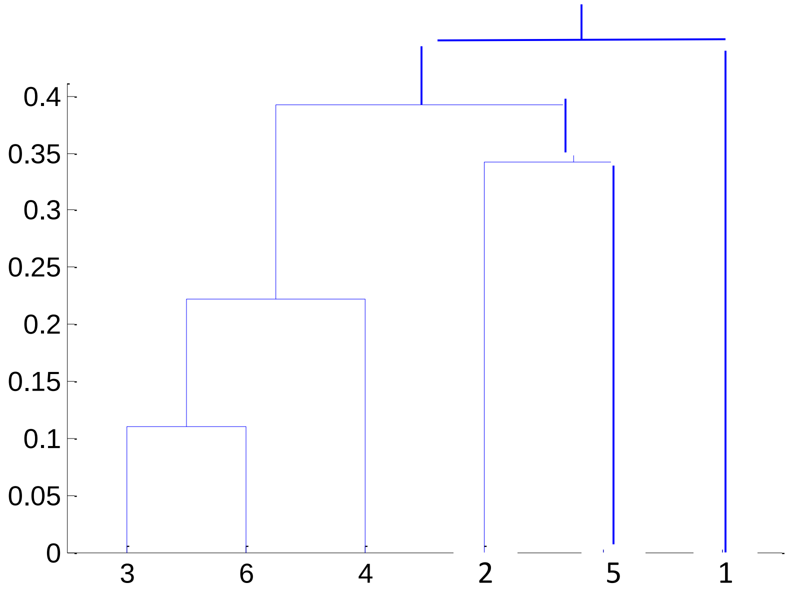
Cluster Distance: Centroid distance

- Distance between two clusters is based on the distance between their centroids
  - Determined by all points in each cluster

Cluster distance: Centroid distance



Nested Clusters



Dendrogram

Hierarchical Clustering: Time and Space

- $O(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.

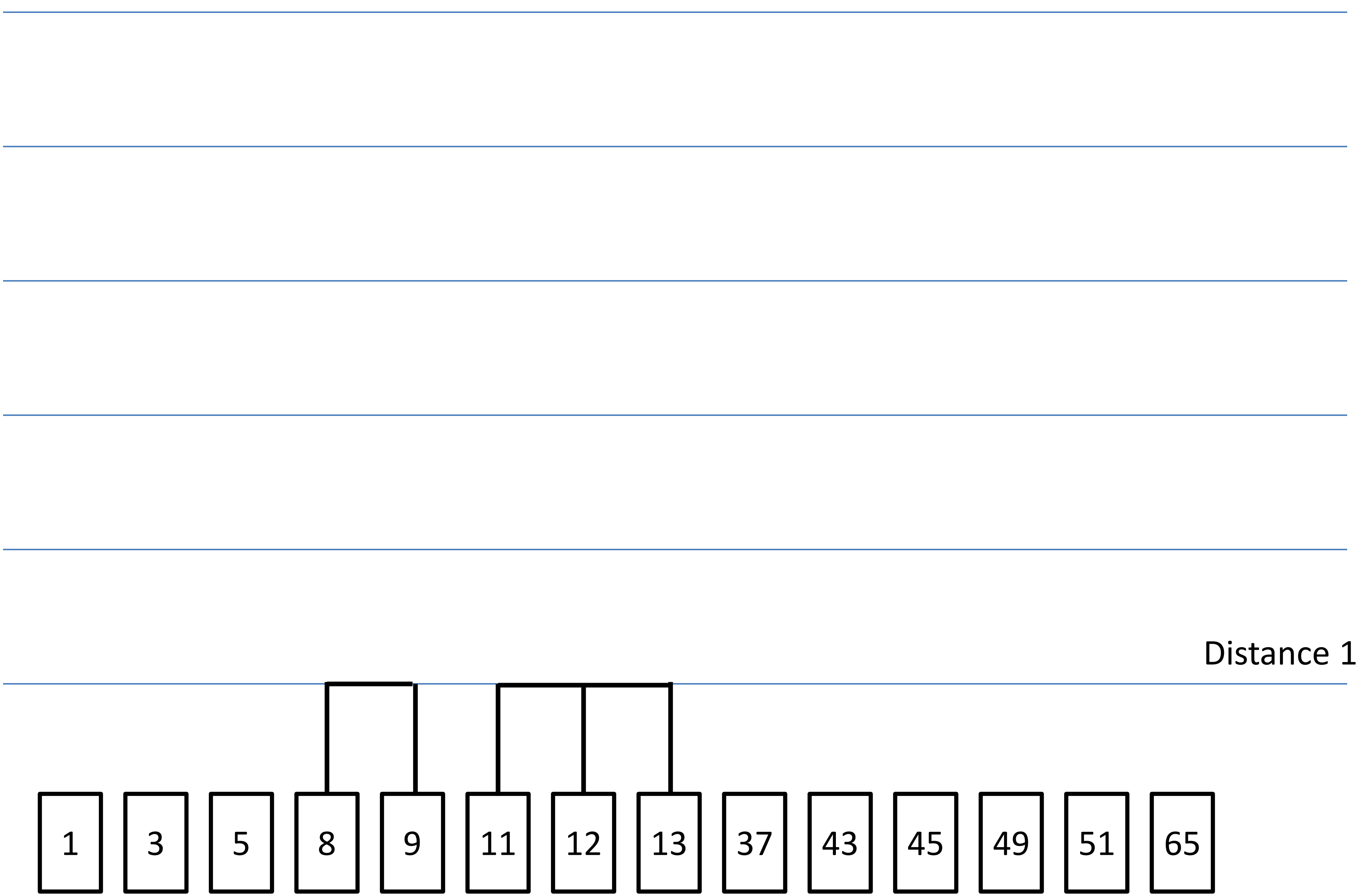
- $O(N^3)$ time in many cases
  - There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
  - Complexity can be reduced to $O(N^2 \log(N))$ time using more advanced data structures

Hierarchical clustering is expensive !

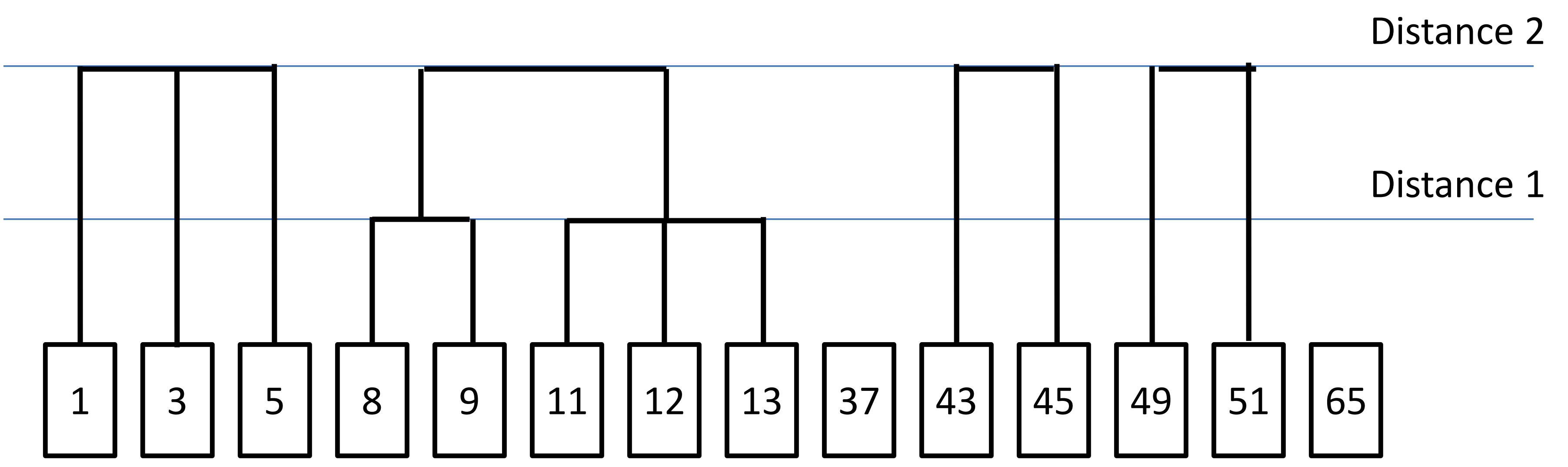
Example: clustering people by age

- Example in one dimension (to skip proximity matrix computation)
- The data consists of the ages of people at a family gathering.
- The goal is to cluster participants by age
- The distance between people is the difference in their ages.
- The procedure: sort participants by age, then begin clustering the closest groups

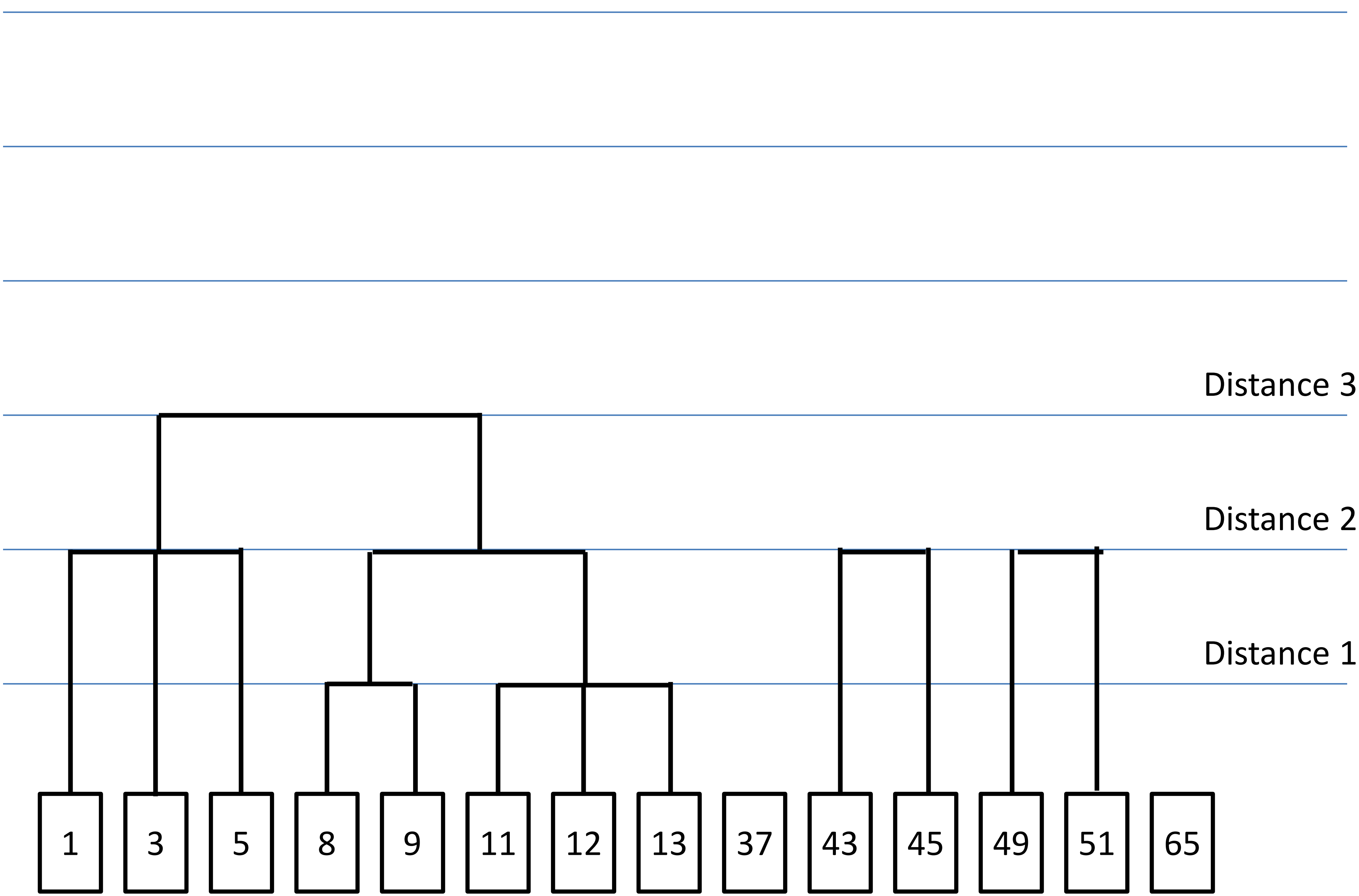
Distance between clusters: MIN



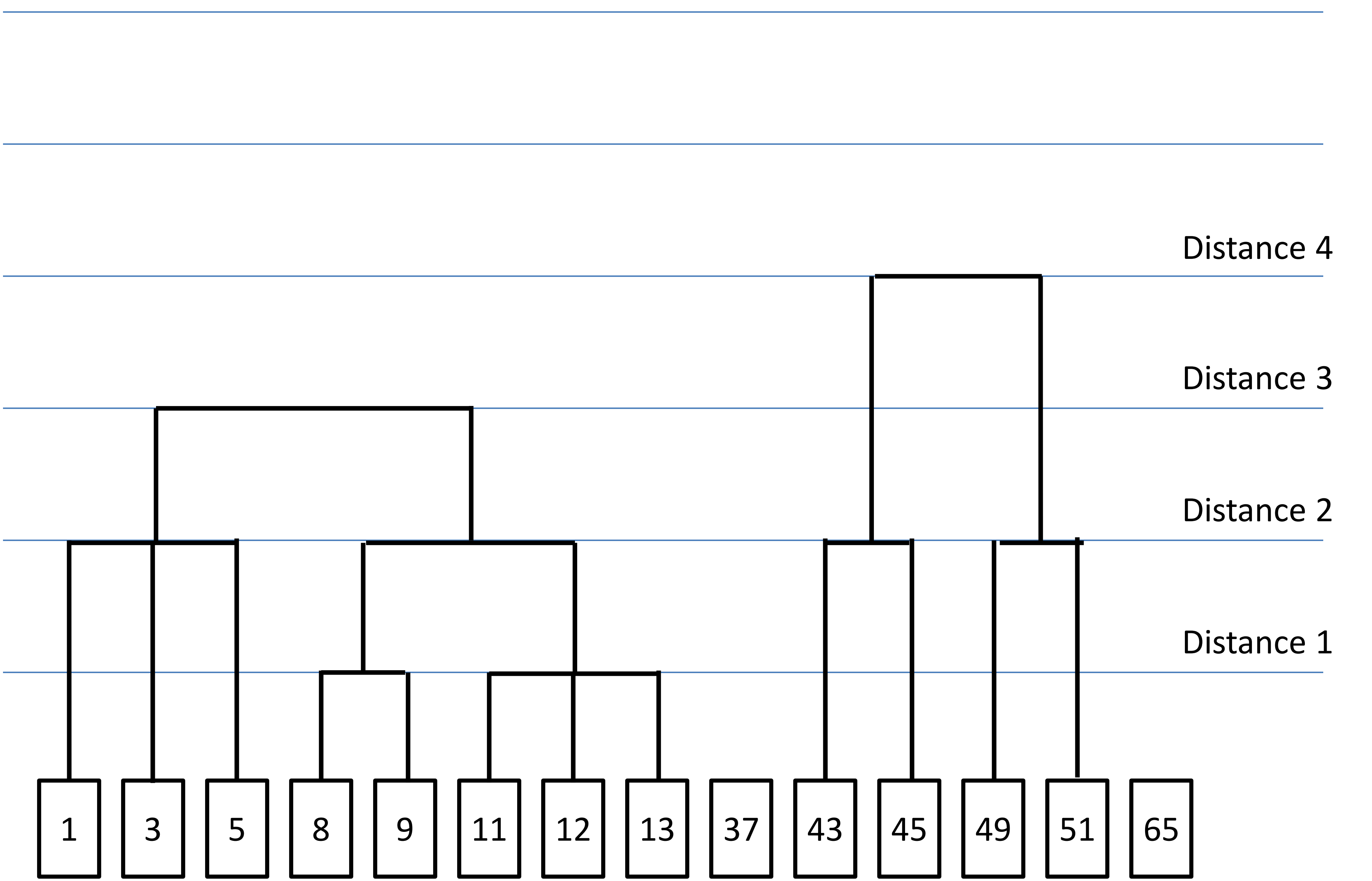
Distance between clusters: MIN



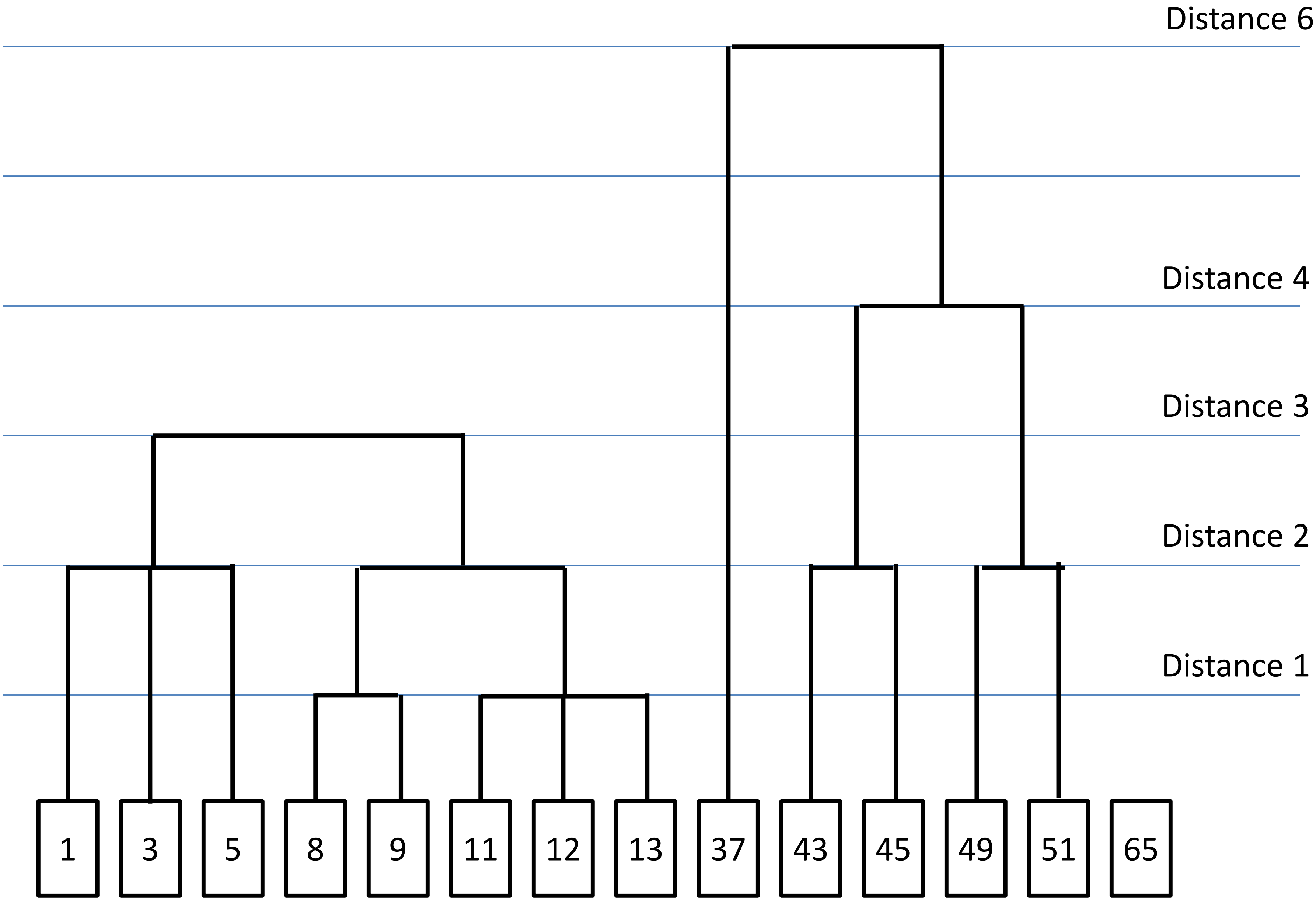
Distance between clusters: MIN



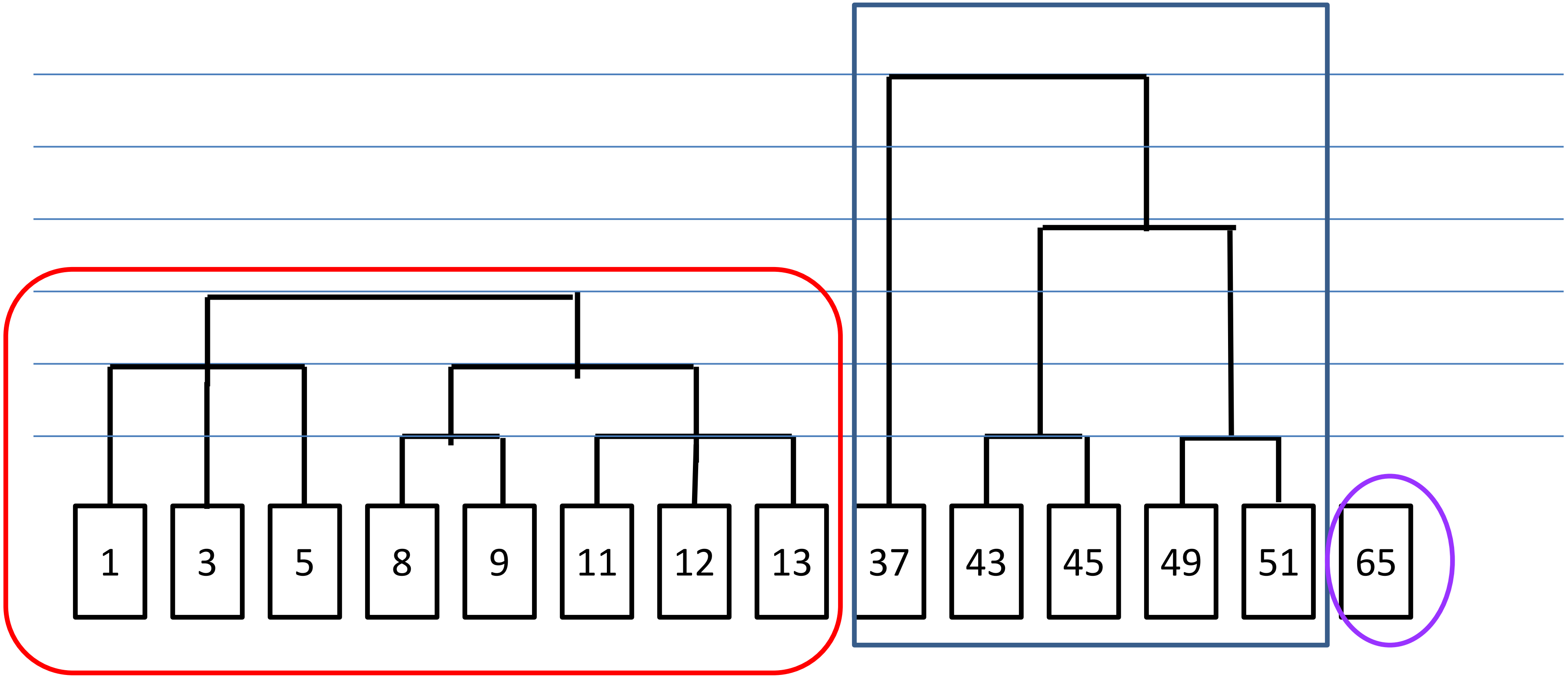
Distance between clusters: MIN



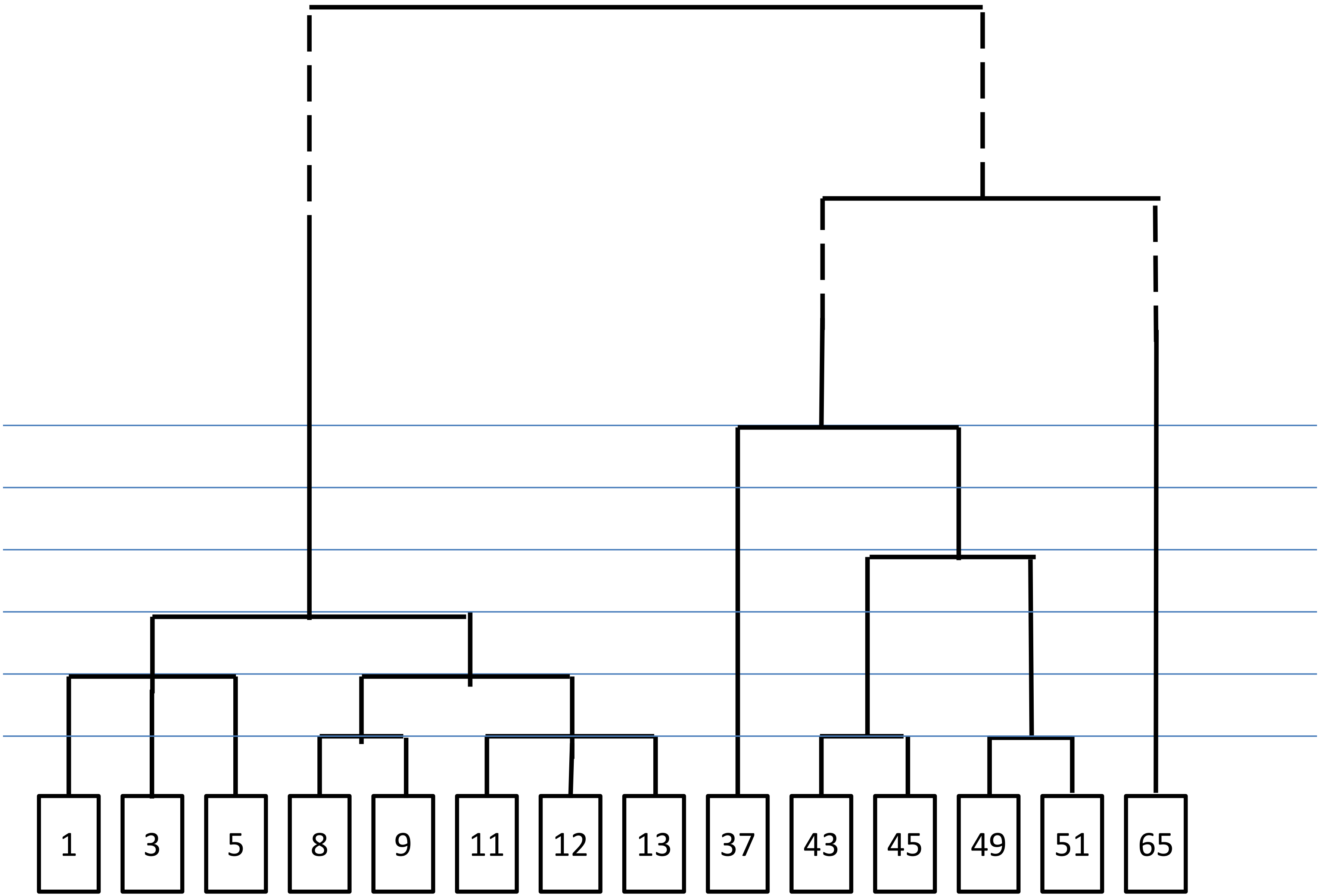
Distance between clusters: MIN



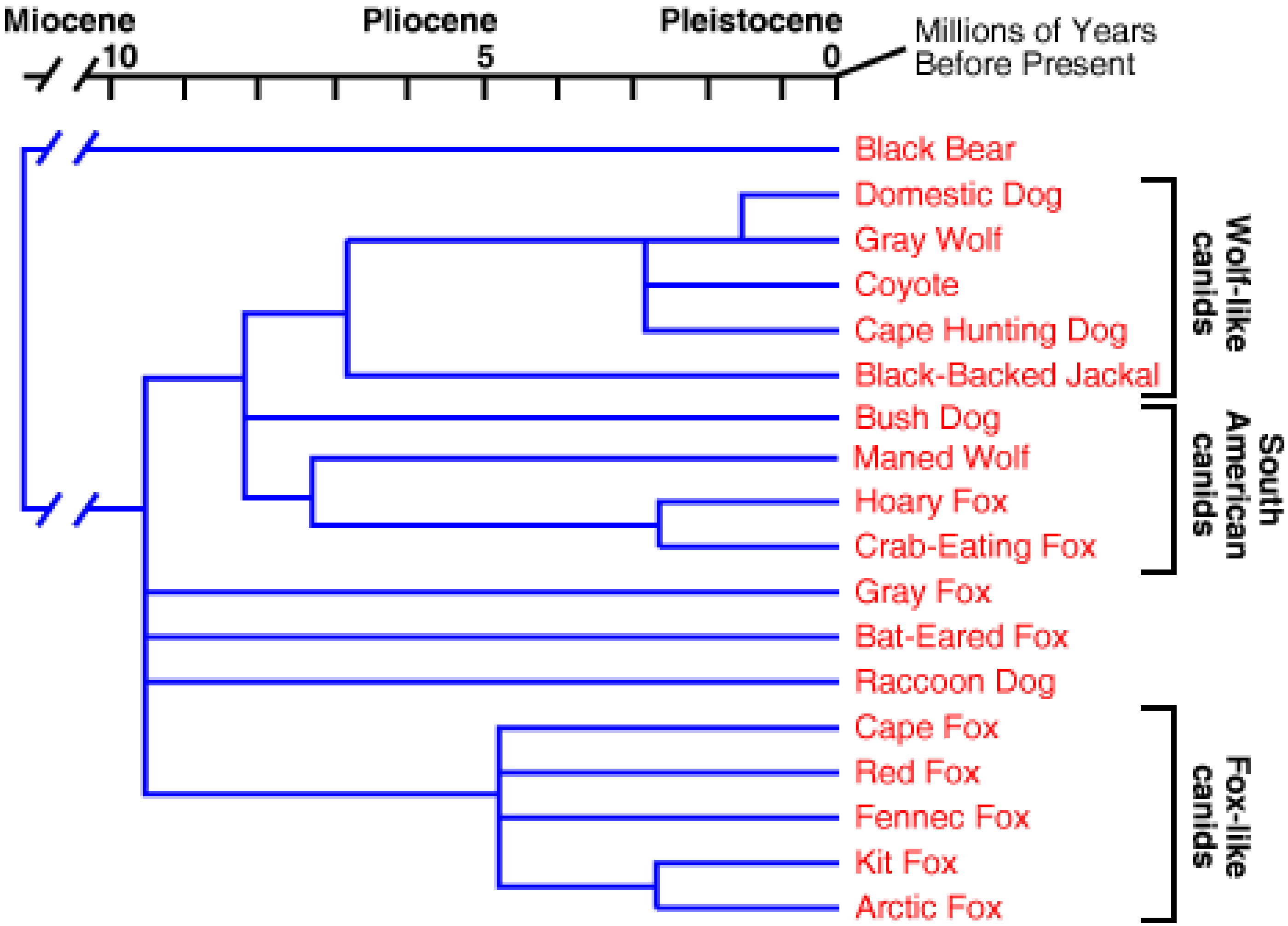
3 groups detected



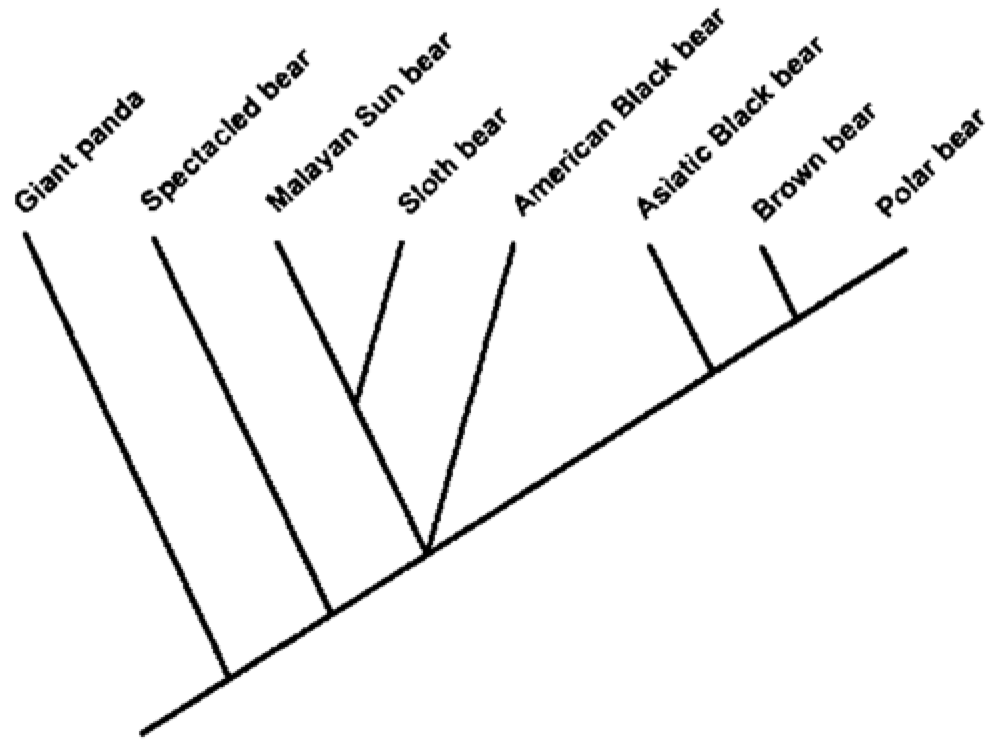
Final dendrogram



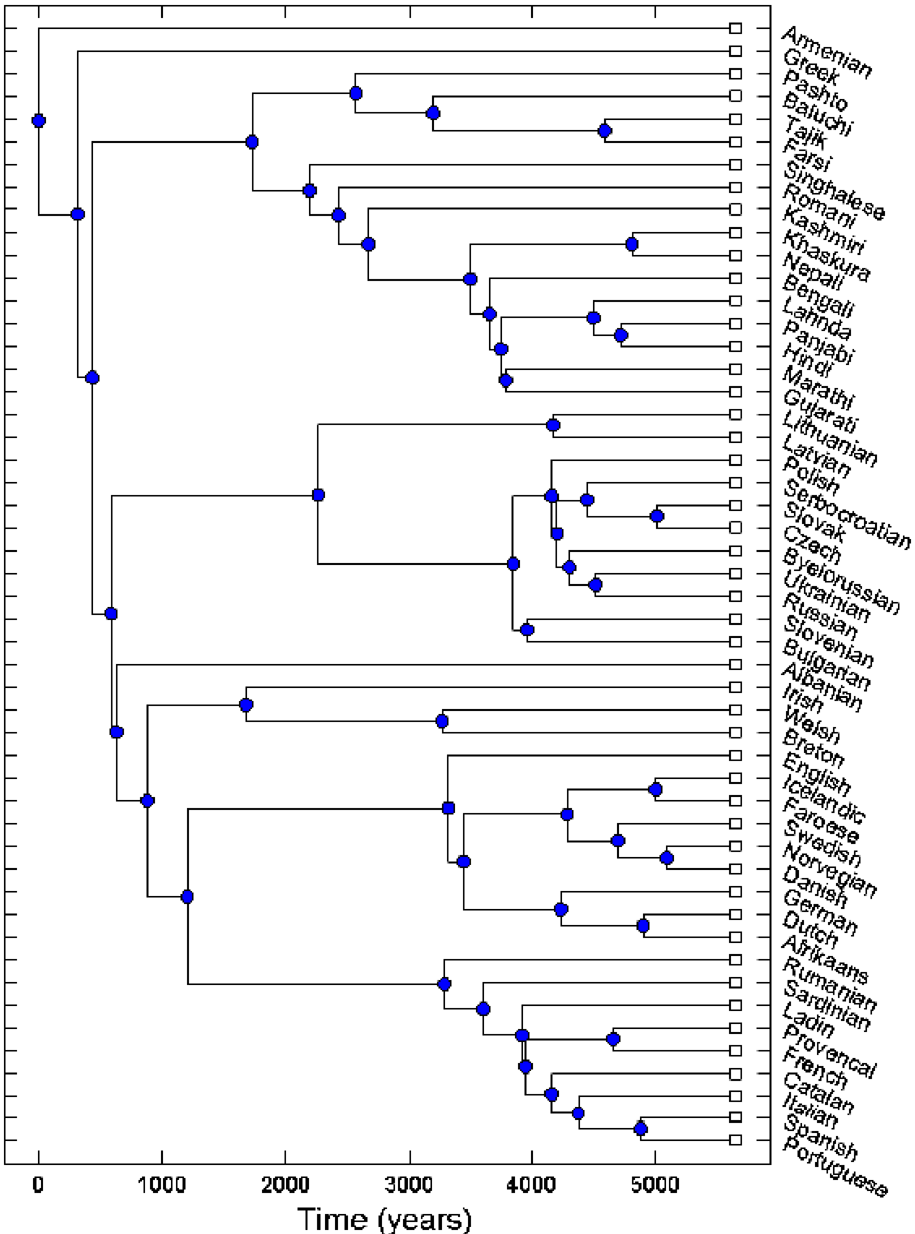
Hierarchical clustering application: evolution of Canidae



Giant Panda is a bear



Hierarchical clustering application: languages evolution



From
“Indo-European languages tree by Levenshtein distance”
by M. Serval and F. Petroni

Hierarchical clustering application: languages evolution