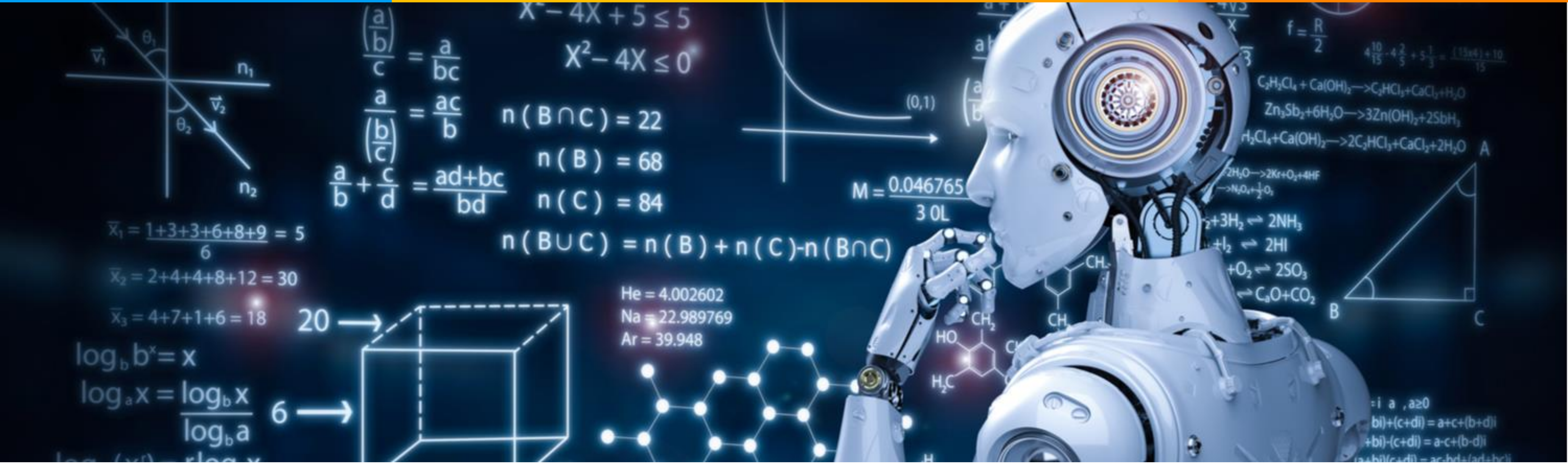
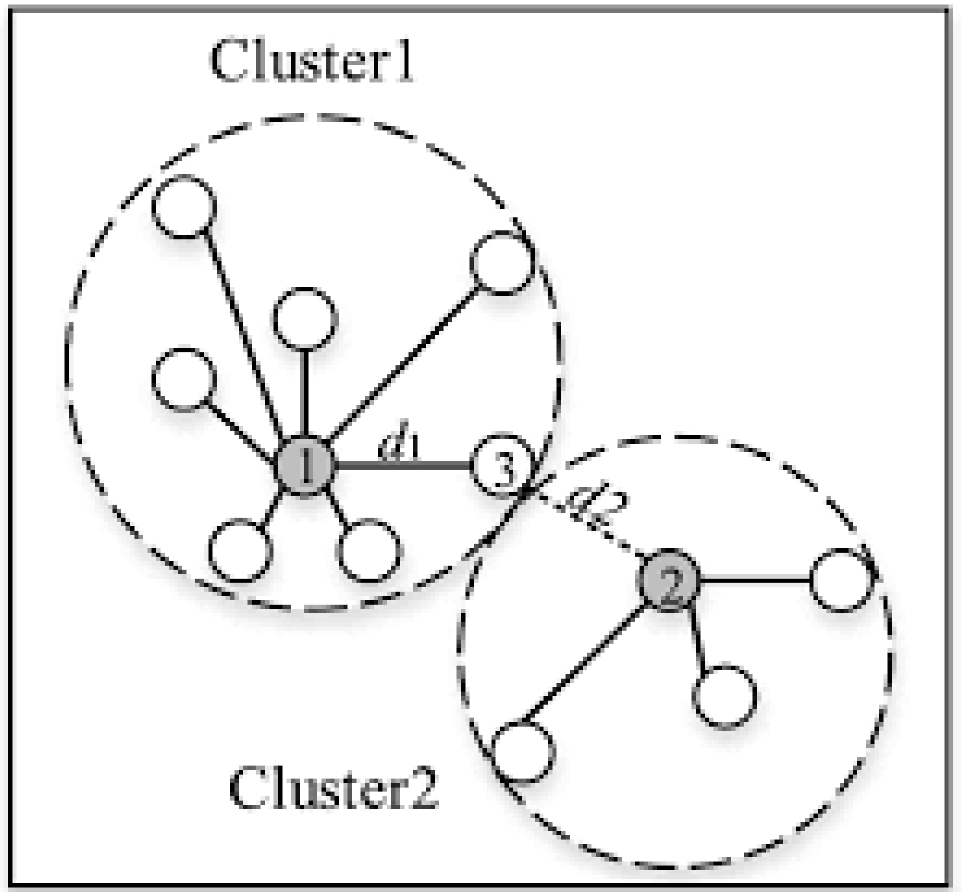
Duration: 60 min | Srikanth Shenoy | AML 5102 | M.E. AI & ML

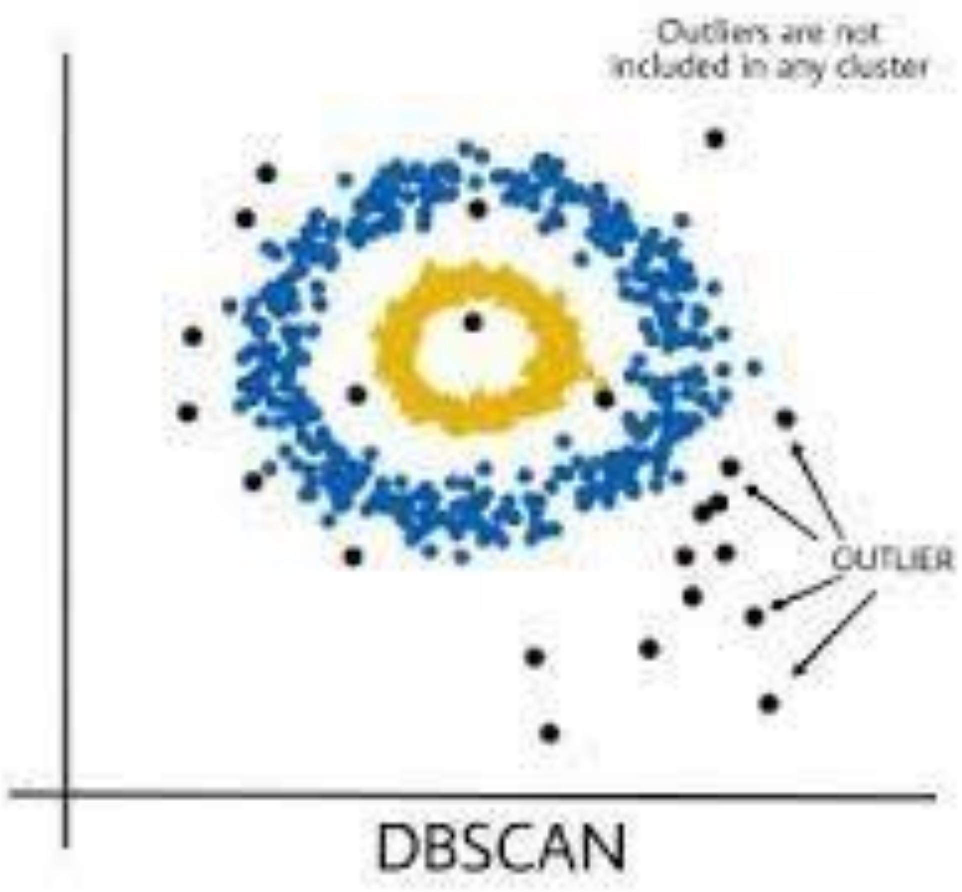


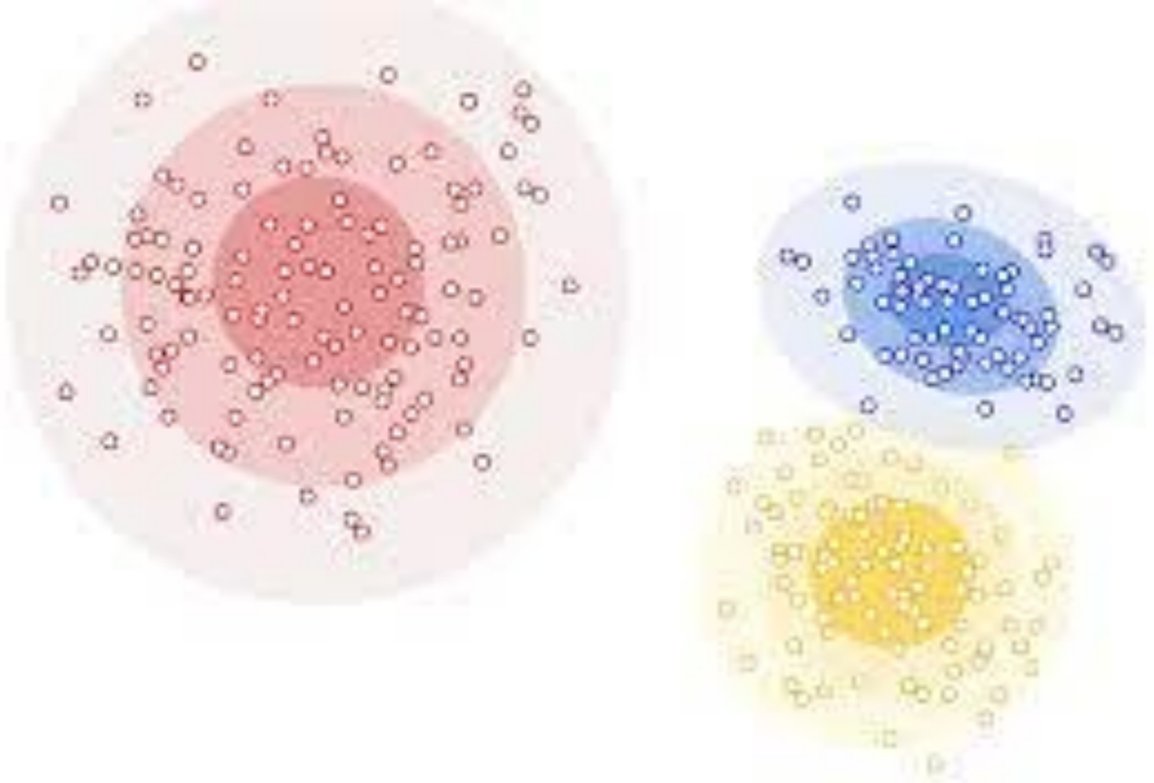
Lecture 16: Hierarchial Clustering

Cluster Analysis - Recap

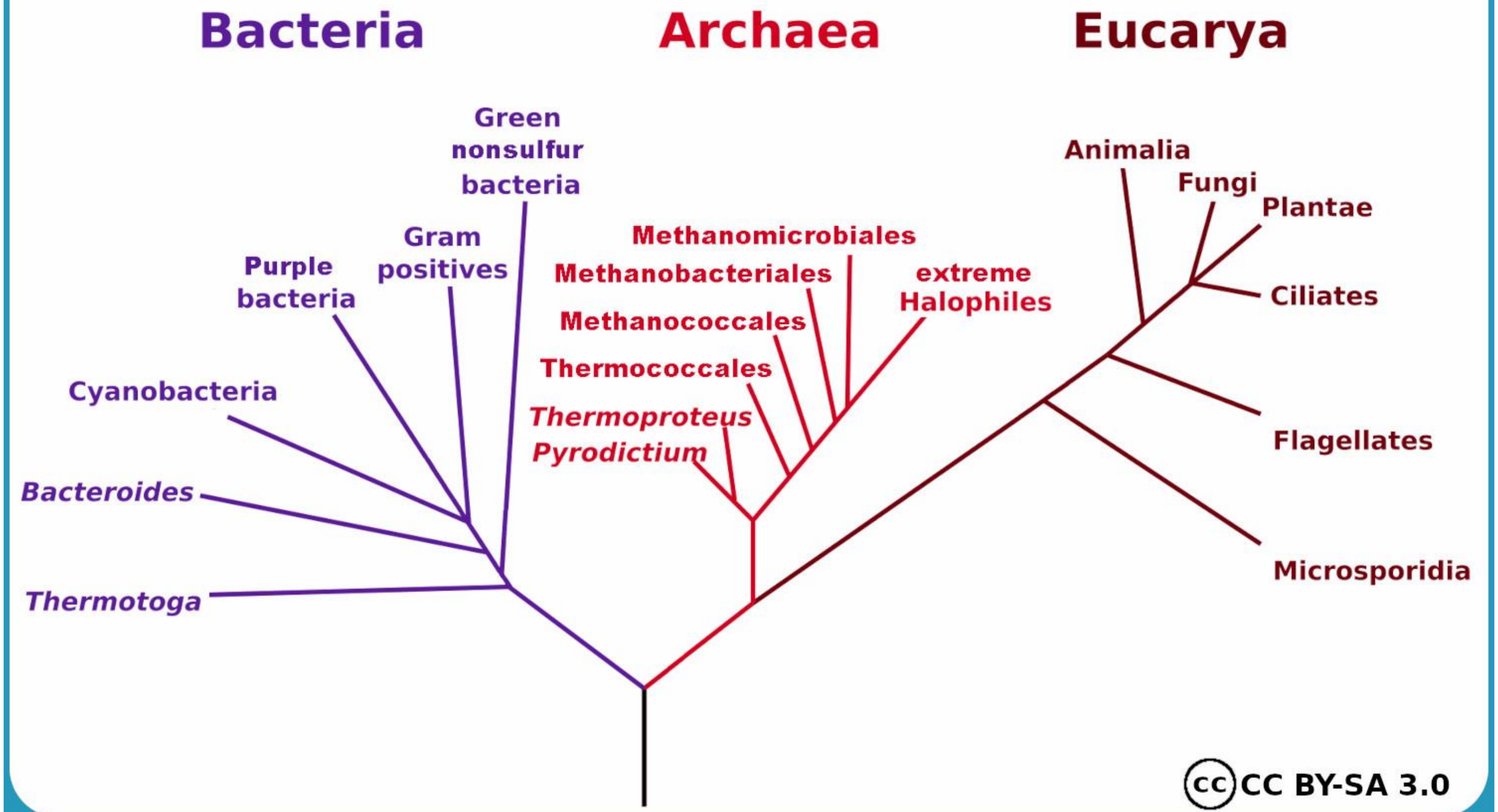
- Centroid based clustering
- Distribution based clustering
- Density based clustering
- Connectivity based clustering



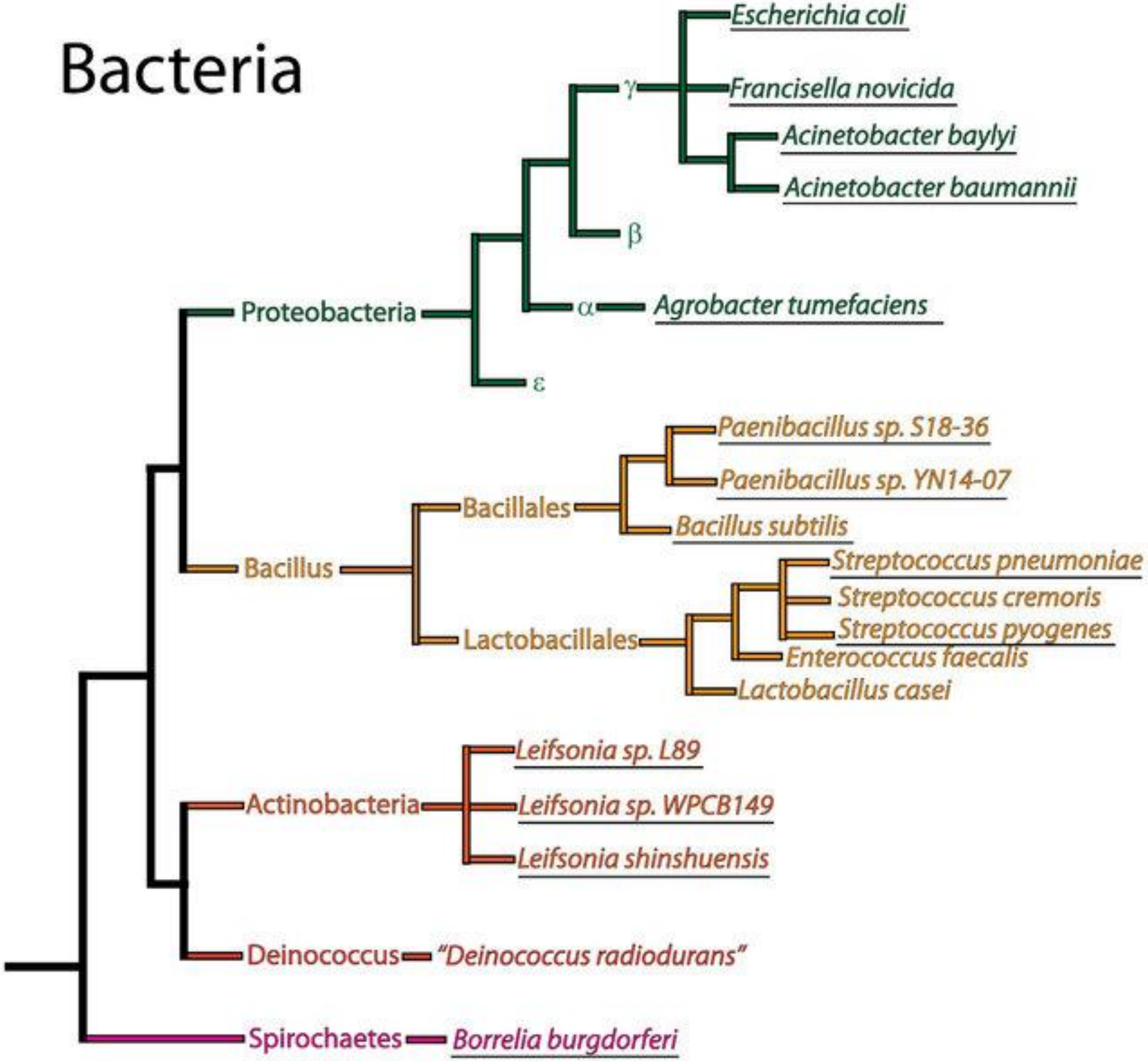




Phylogenetic Tree of Life



Bacteria



Escherichia coli
γ
Francisella novicida
Acinetobacter baylyi
Acinetobacter baumannii
β
Proteobacteria
α
Agrobacter tumefaciens
ε
Paenibacillus sp. S18-36
Paenibacillus sp. YN14-07
Bacillales
Bacillus subtilis
Bacillus
Streptococcus pneumoniae
Streptococcus cremoris
Lactobacillales
Streptococcus pyogenes
Enterococcus faecalis
Lactobacillus casei
Leifsonia sp. L89
Actinobacteria
Leifsonia sp. WPCB149
Leifsonia shinshuensis
Deinococcus
"Deinococcus radiodurans"
Spirochaetes
Borrelia burgdorferi

Types of Hierarchial Clustering

Intro to hierarchical clustering

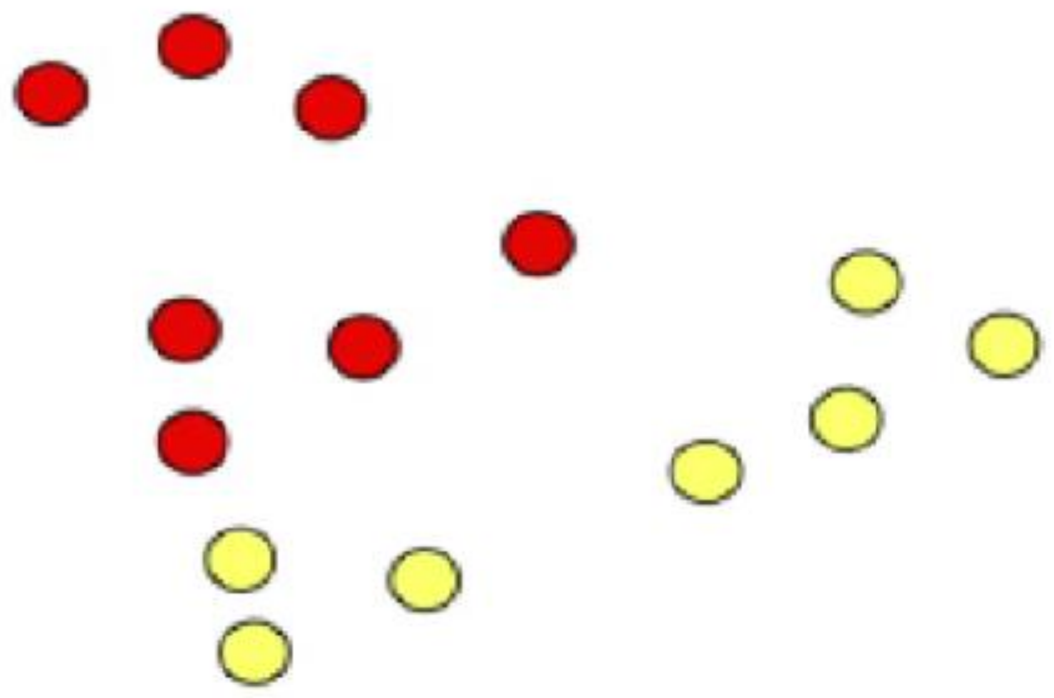
- How many clusters in the data is
  - Technical question
  - Centroid, Distribution & Density answers this question
- Business questions
  - How are clusters related?
  - How to say 2 clusters are different but are also similar?
- We need a hierarchical organization of clusters
  - Based on high level effects / granular sub groupings
- Very natural for some problems
  - Taxonomy

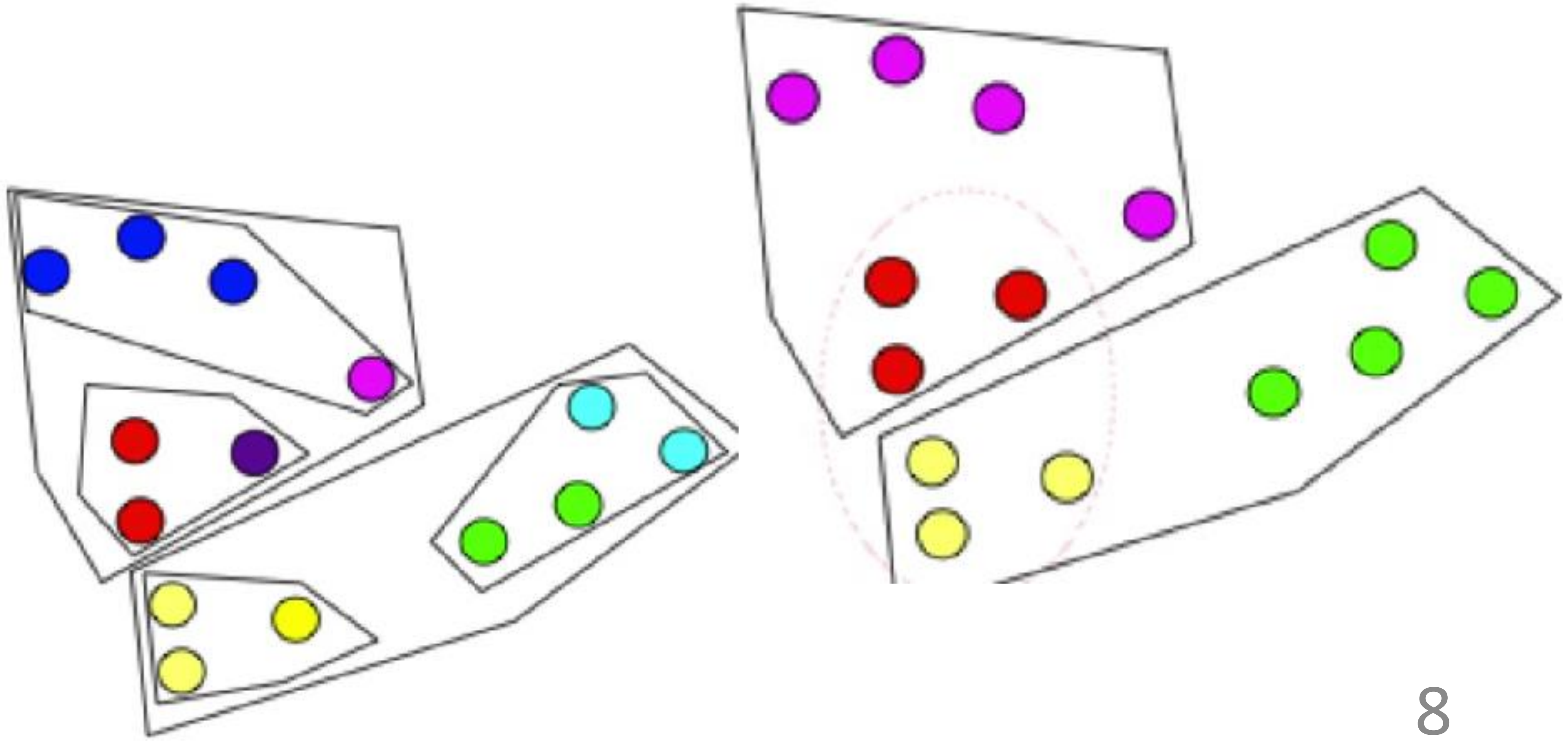
Hierarchical clustering

- Top-down
  - Start with one cluster, split recursively
  - Divisive clustering
- Bottom-up
  - Each point is a cluster, merge clusters
  - Agglomerative clustering
- Tree in both cases
  - All data in one cluster at top
  - Each data in its own cluster at bottom

Divisive clustering

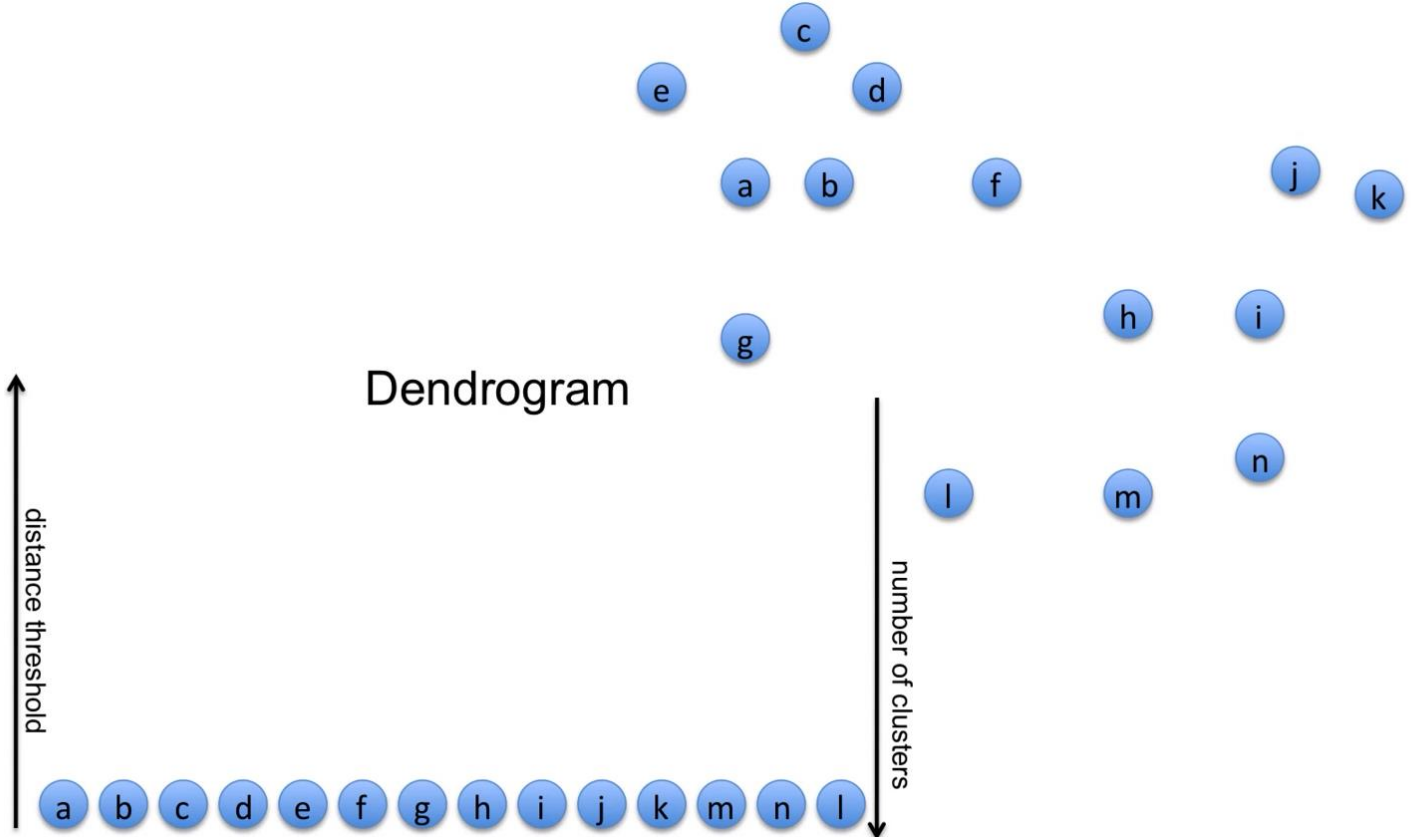
- Recursively run K means (say K = 2)
- Partition into two may not be reasonable
- Split into 2. Further split each into 2
- No going back and changing the splits
  - Greedy
  - Sometimes problematic
- Computationally efficient



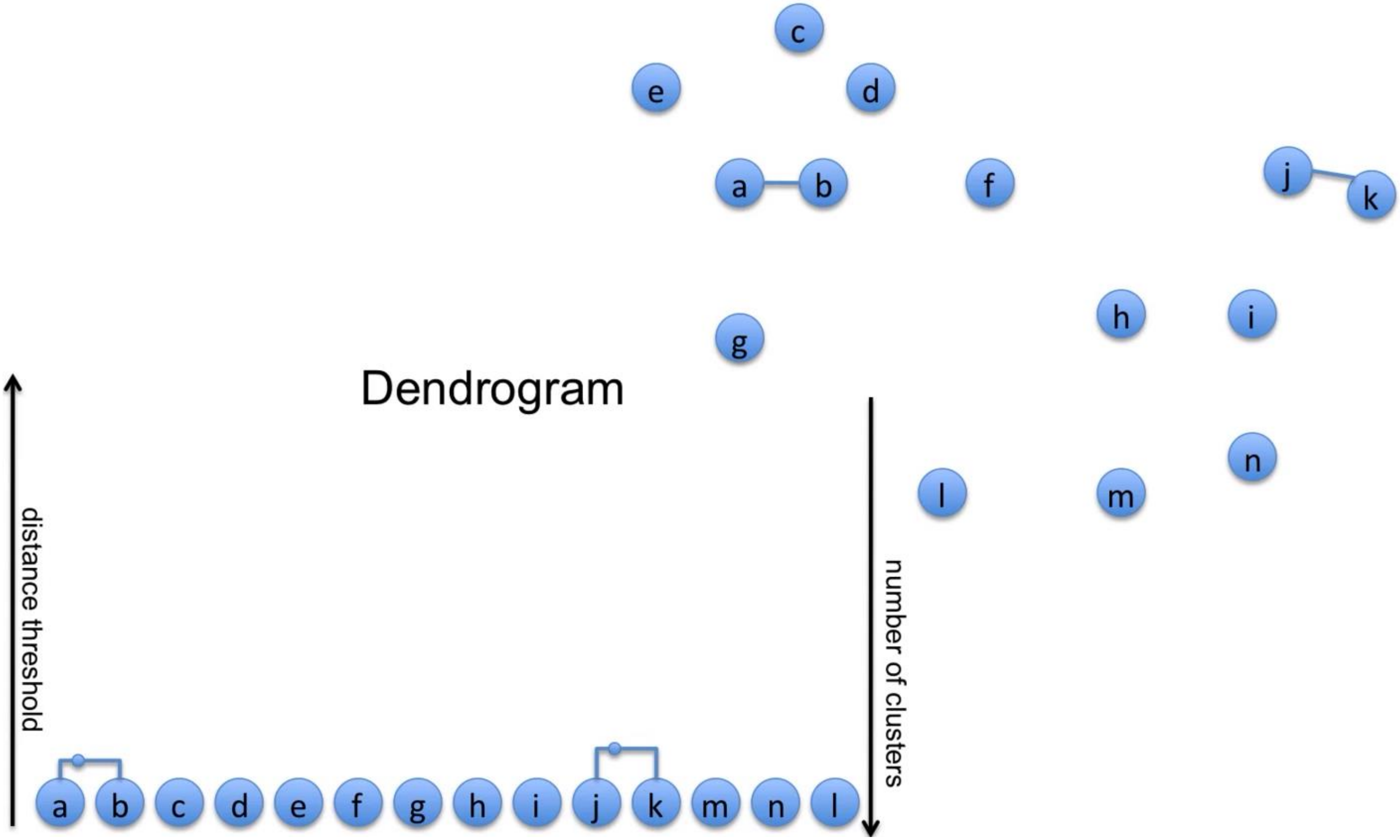


Agglomerative Clustering

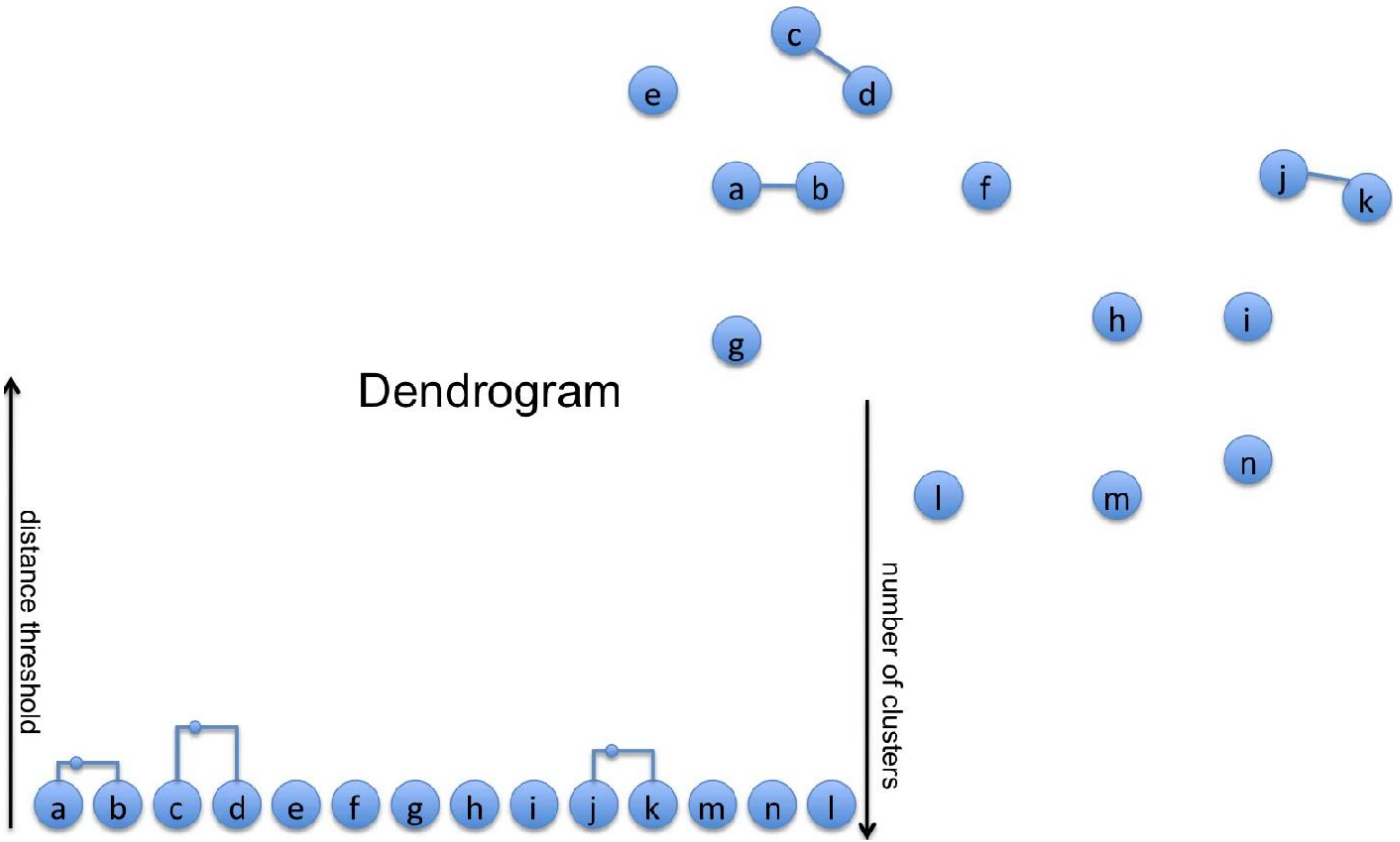
Dendrogram

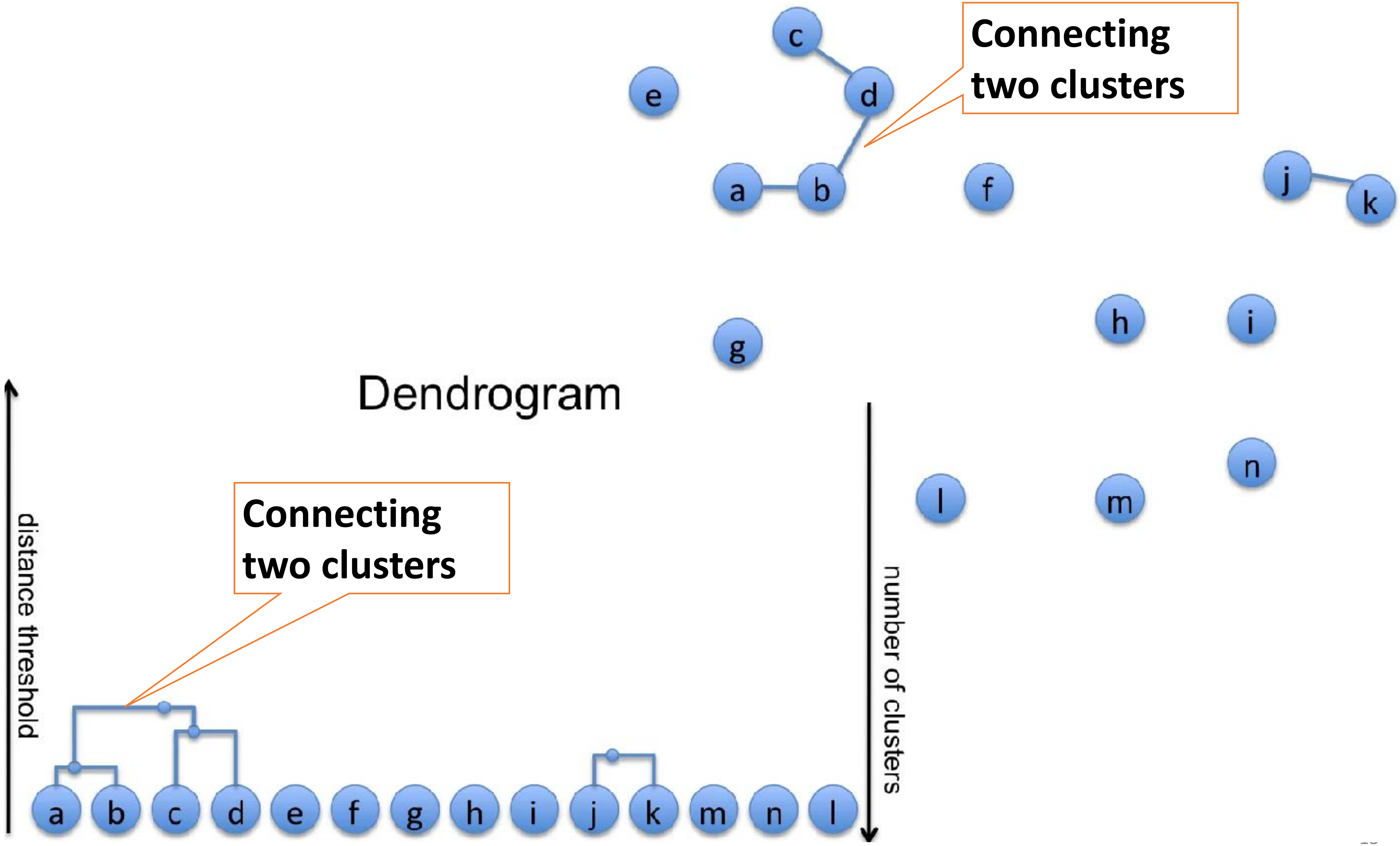


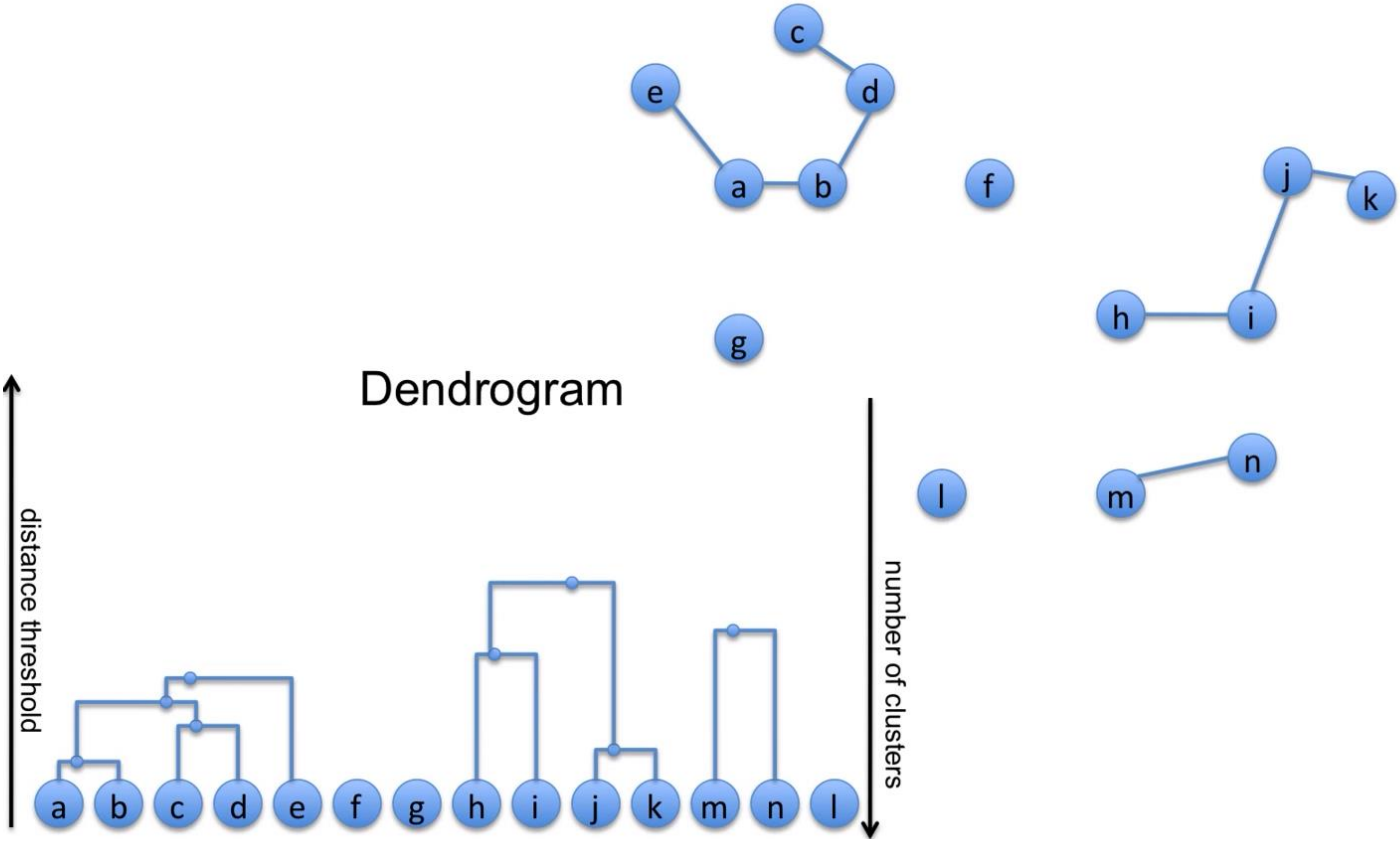
c
e
d
a
b
f
j
k
h
i
g
distance threshold
number of clusters
n
l
m
a b c d e f g h i j k m n l



c
e
d
a
b
f
j
k
h
i
g
Dendrogram
distance threshold
number of clusters
n
l
m
a b c d e f g h i j k m n l

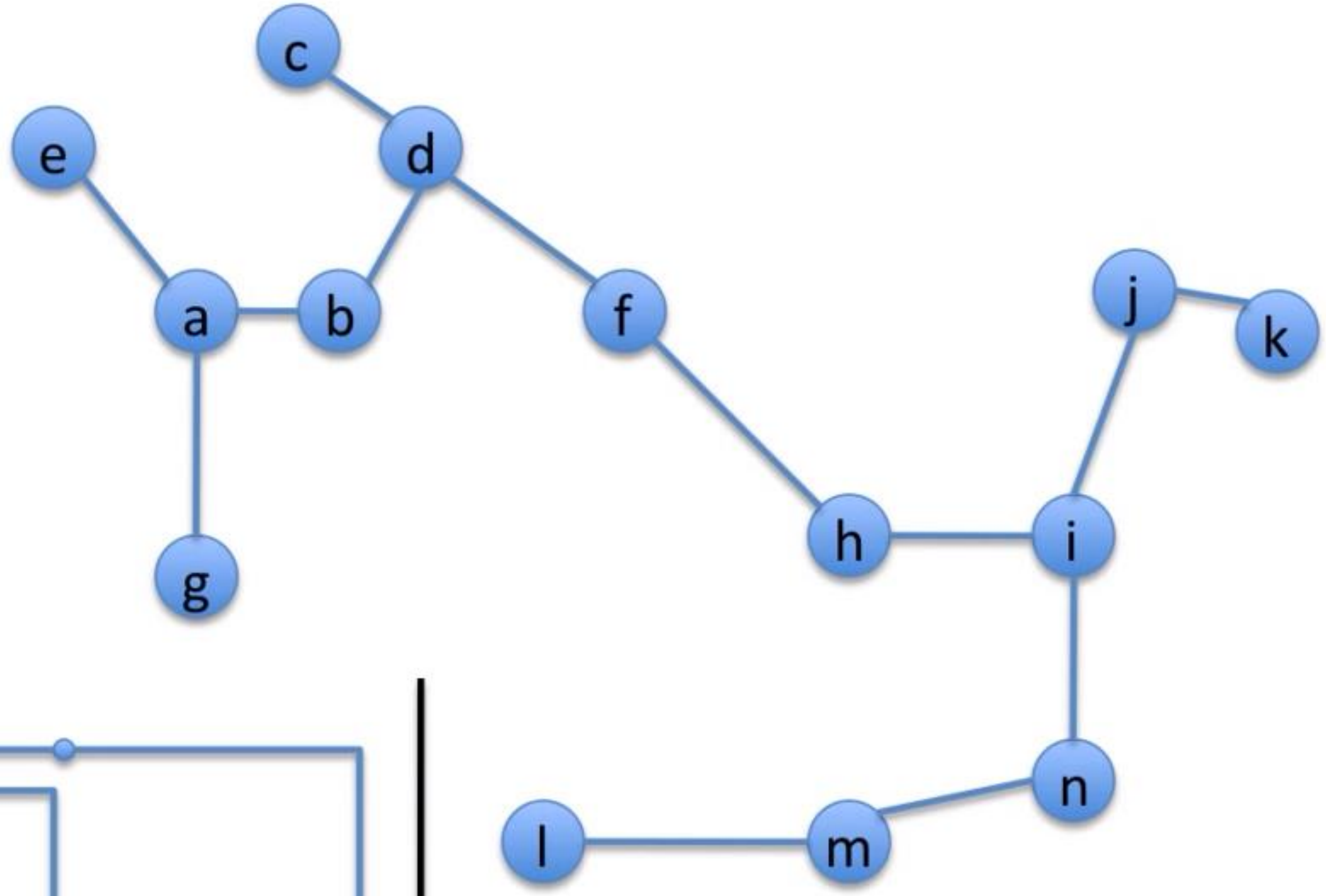

c
d
e
a
b
f
j
k
h
i
g
Dendrogram
distance threshold
number of clusters
n
l
m
a b c d e f g h i j k m n l

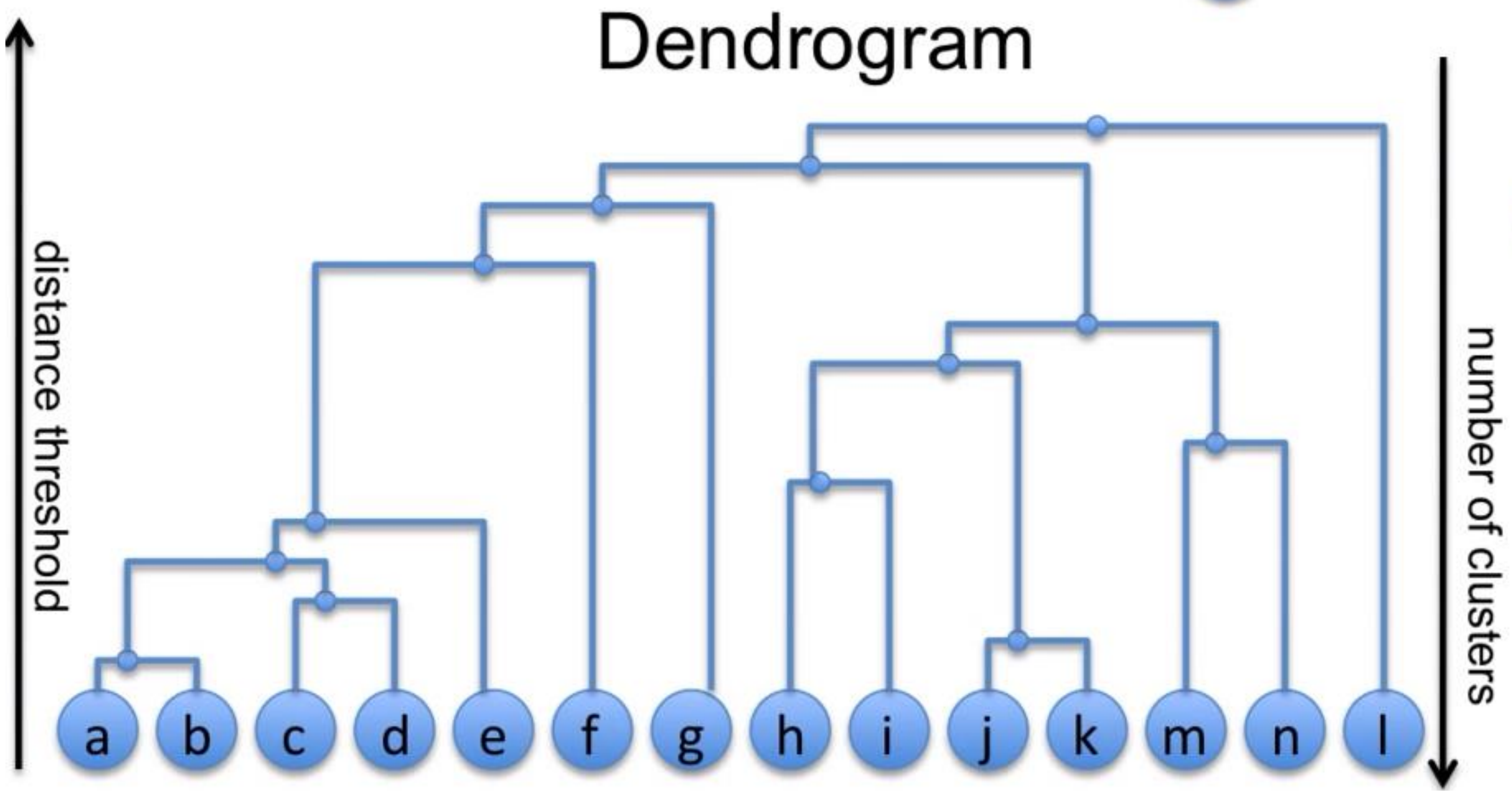

Connecting two clusters
Dendrogram
distance threshold
number of clusters
Connecting two clusters
a b c d e f g h i j k l m n

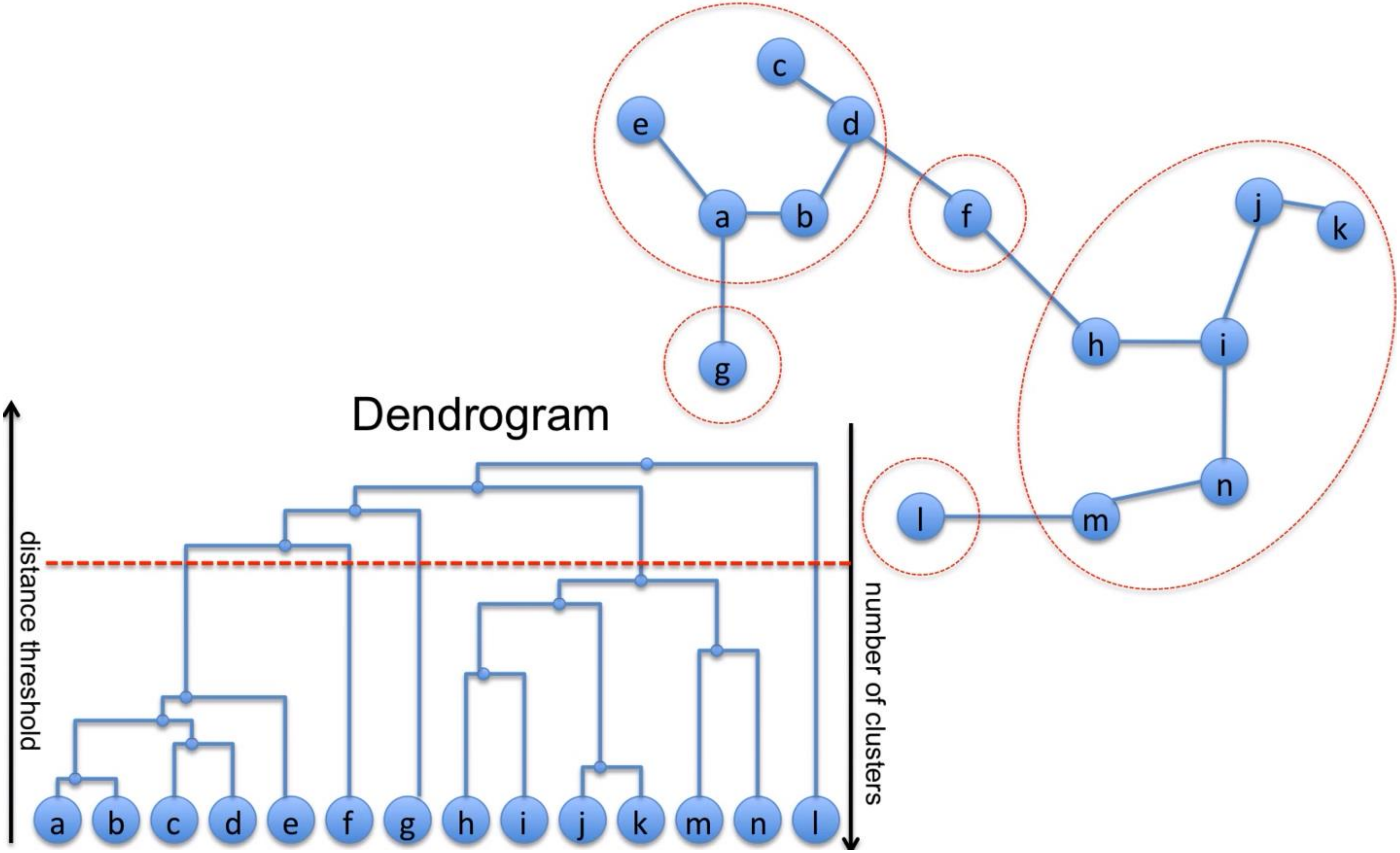


Dendrogram
distance threshold
number of clusters
a
b
c
d
e
f
g
h
i
j
k
l
m
n

How many clusters is good?







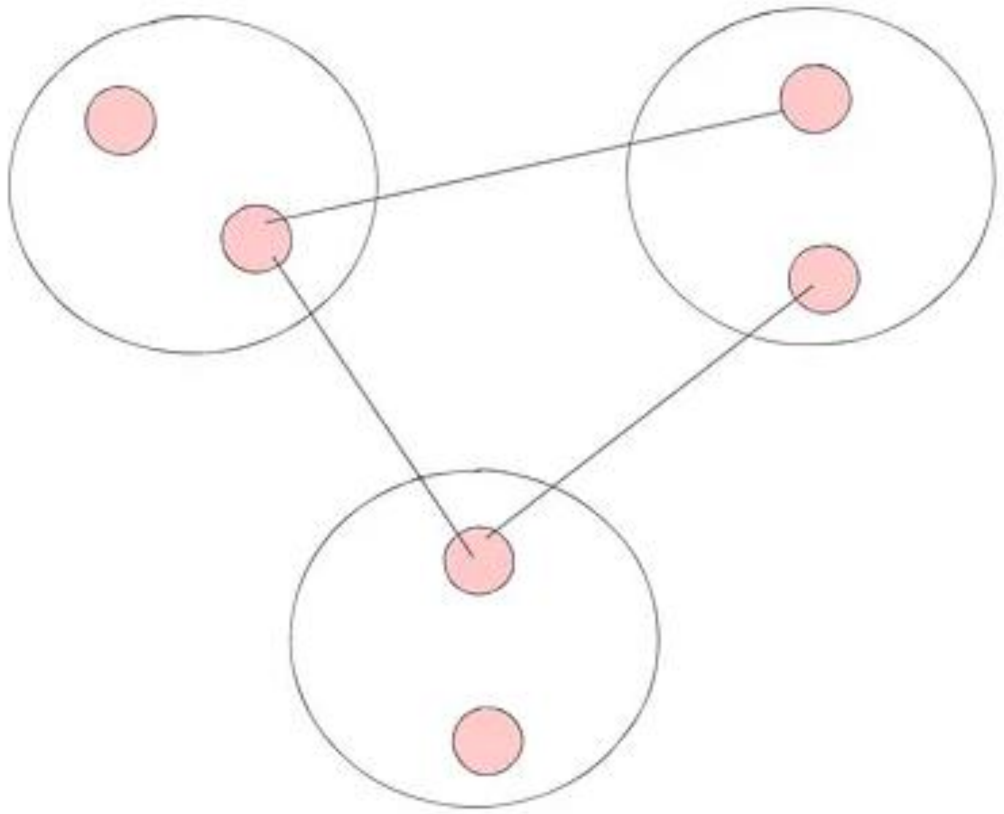
Dendrogram
distance threshold
number of clusters
a b c d e f g h i j k m n l

Agglomerative clustering

- Idea: Nearby points end up in same cluster
- Algorithm
  - Start with collection C of N singleton clusters
  - cj = {xj}
  - Find another cluster that is closest min ij (ci, cj)
  - Merge ci and cj into new cluster ci+j
  - Remove ci and cj crom C. Add ci+j
  - Repeat
- Slower algorithm
- Understand with dendrogram (connectivity tree)

Cluster distance metric for hierarchical clustering

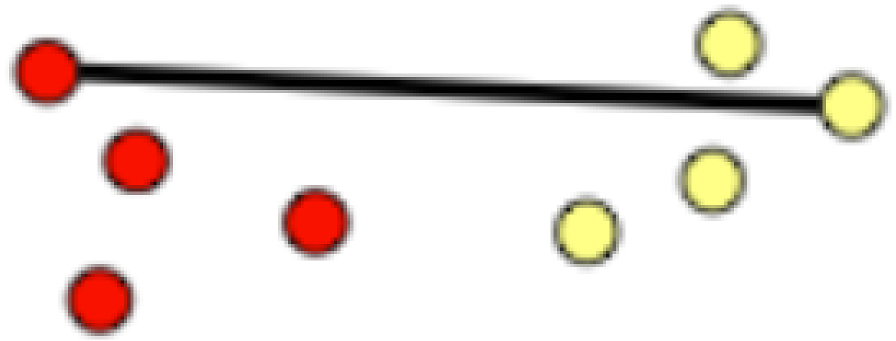
- Distance between clusters
- Not individual points

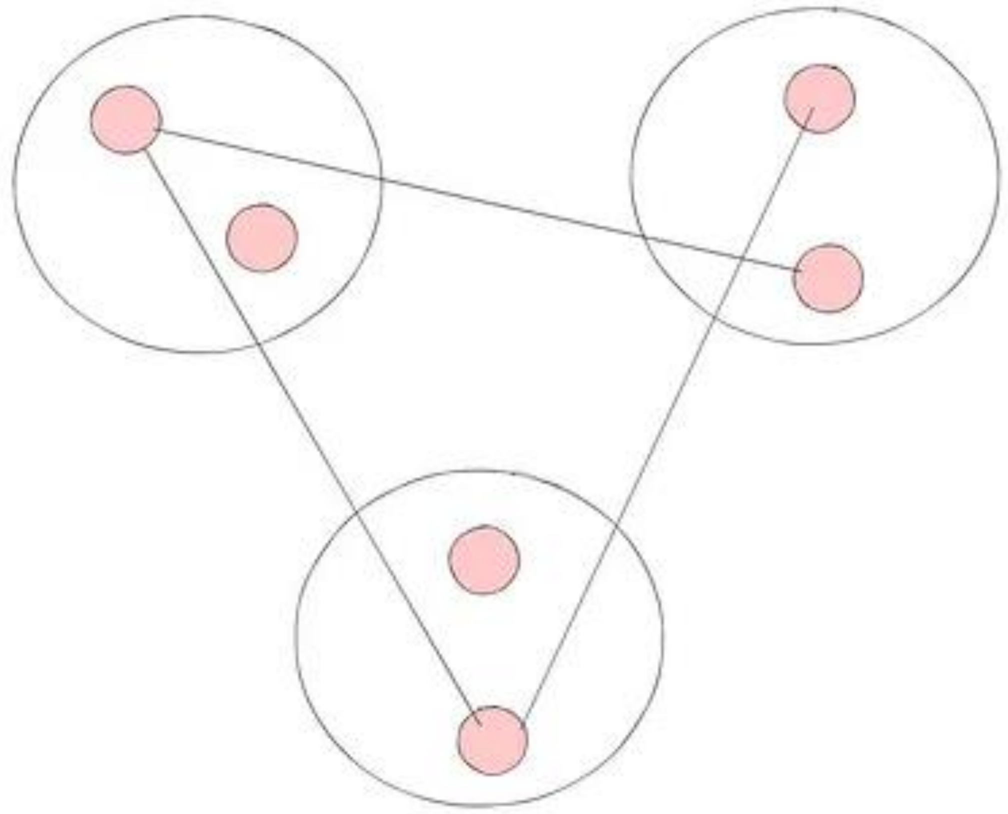


- Single Link
  - Distance b/w closest elements of clusters

$$\mathcal{D}(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} \mathcal{D}(x_1, x_2)$$

  - Produces long chains
- Complete Link





  - Distance b/w farthest elements of clusters

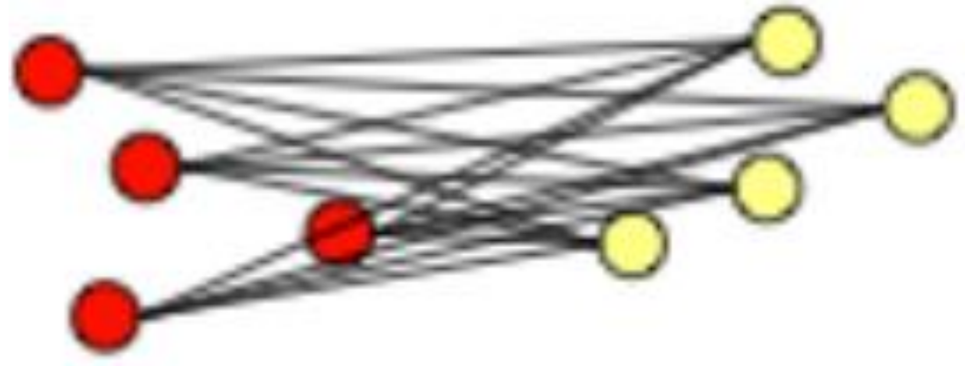
$$\mathcal{D}(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} \mathcal{D}(x_1, x_2)$$

  - Produces spherical clusters

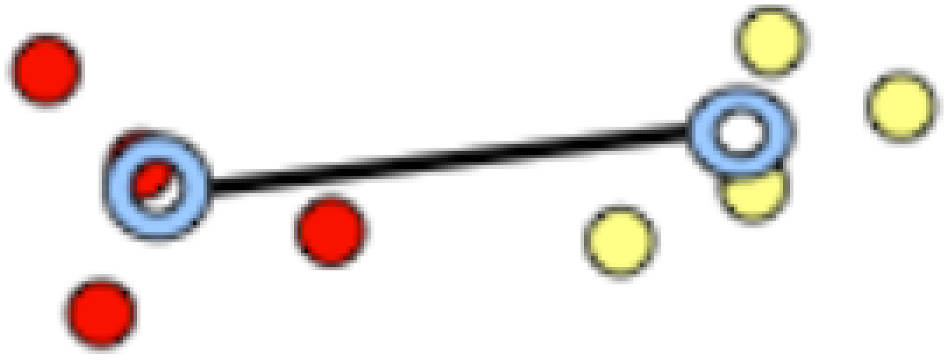
Cluster distance metric for hierarchical clustering

- Average Link

$$\mathcal{D}(c_1, c_2) = \frac{1}{|c_1||c_2|} \sum_{x_1 \in c_1} \sum_{x_2 \in c_2} \mathcal{D}(x_1, x_2)$$



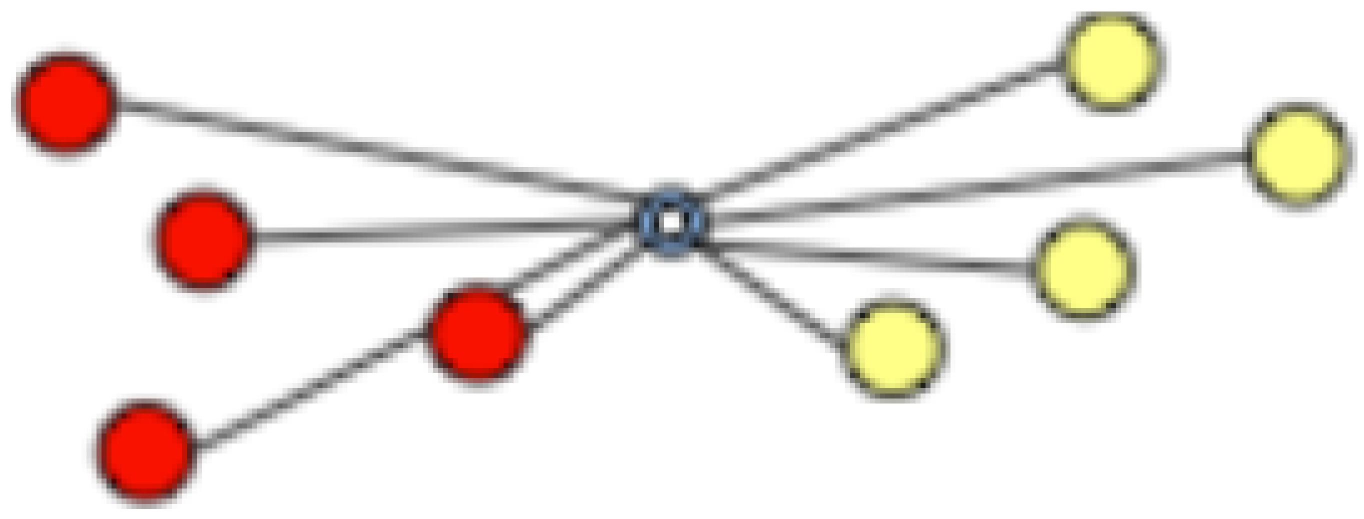
  - Avg of pairwise distance b/w cluster elements
  - Less affected by outliers
- Centroid Link
  - Distance b/w cluster centroids

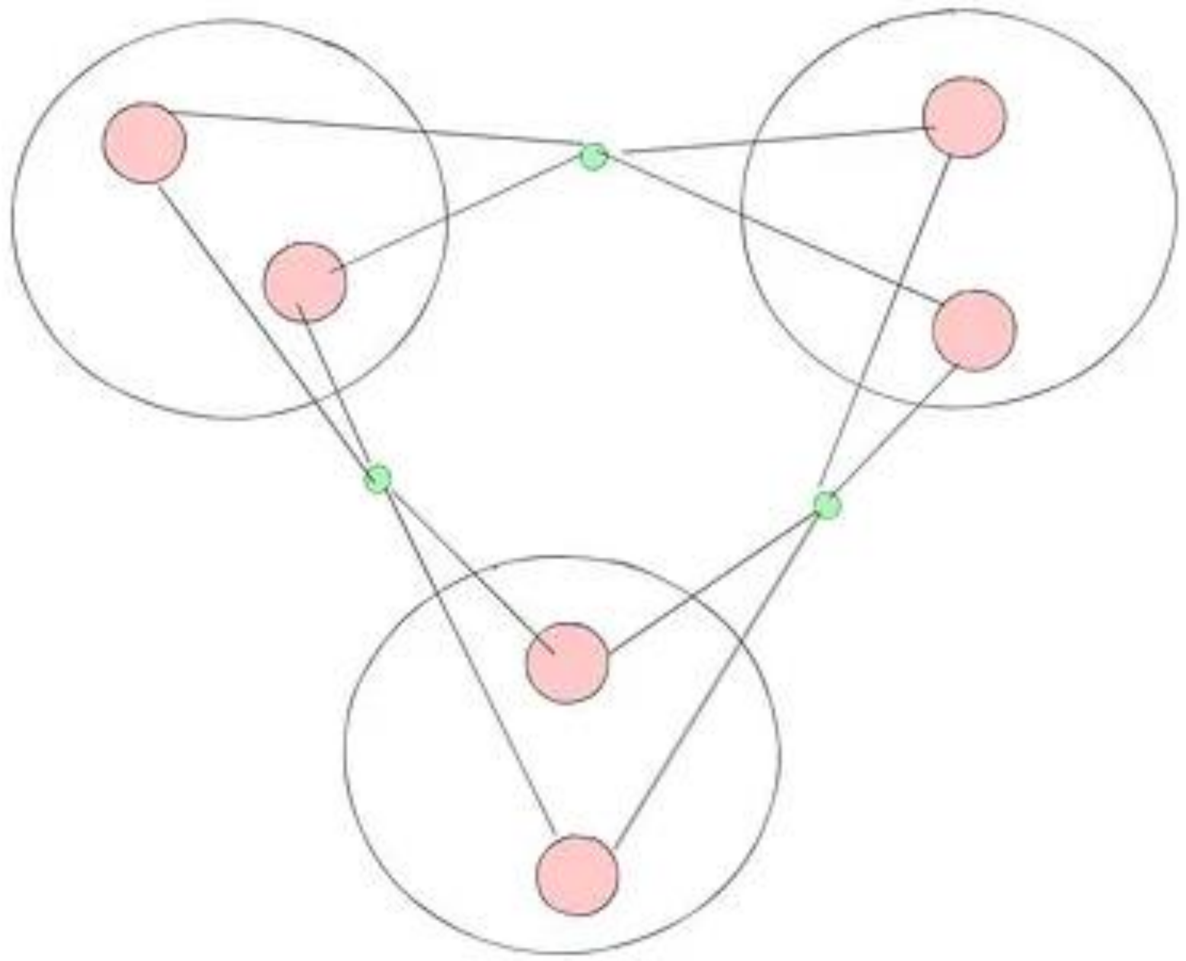


$$\mathcal{D}(c_1, c_2) = \mathcal{D}\left(\left(\frac{1}{|c_1|} \sum_{x_1 \in c_1} \right) \ , \ \left(\frac{1}{|c_2|} \sum_{x_2 \in c_2} \right) \right)$$

Cluster distance metric for hierarchical clustering

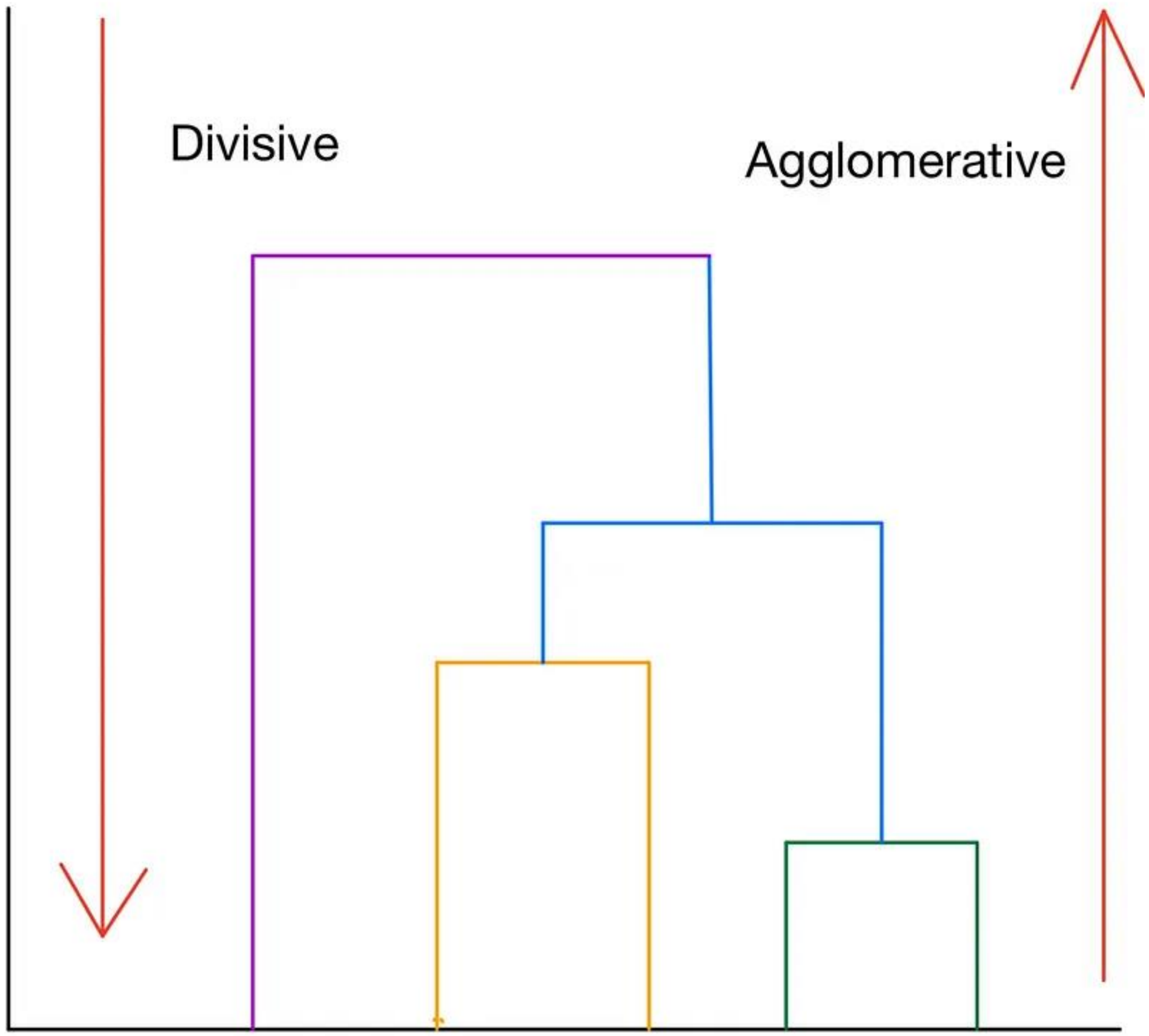
- Ward Link
  - If the two clusters were to join
  - What will be their mean and standard deviation?
  - The two clusters with minimum SD will be merged

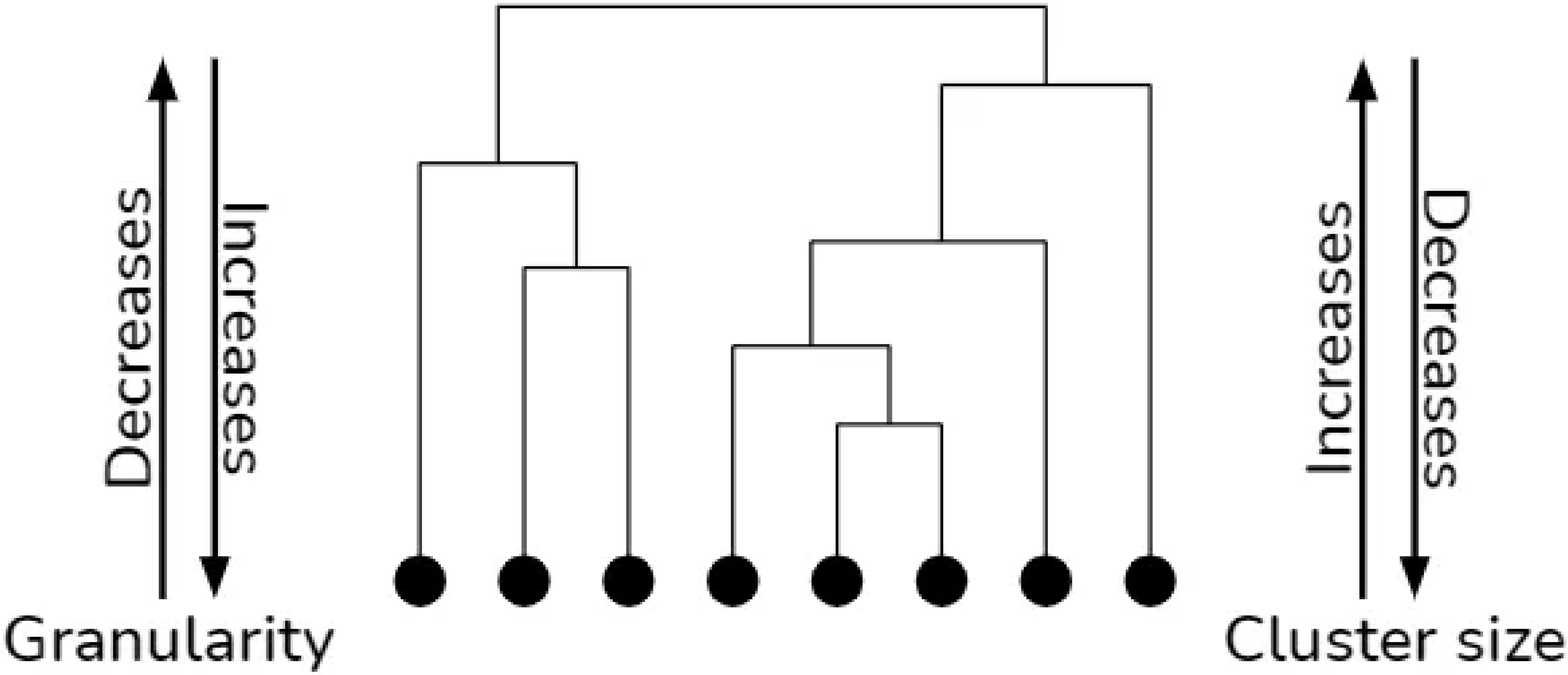


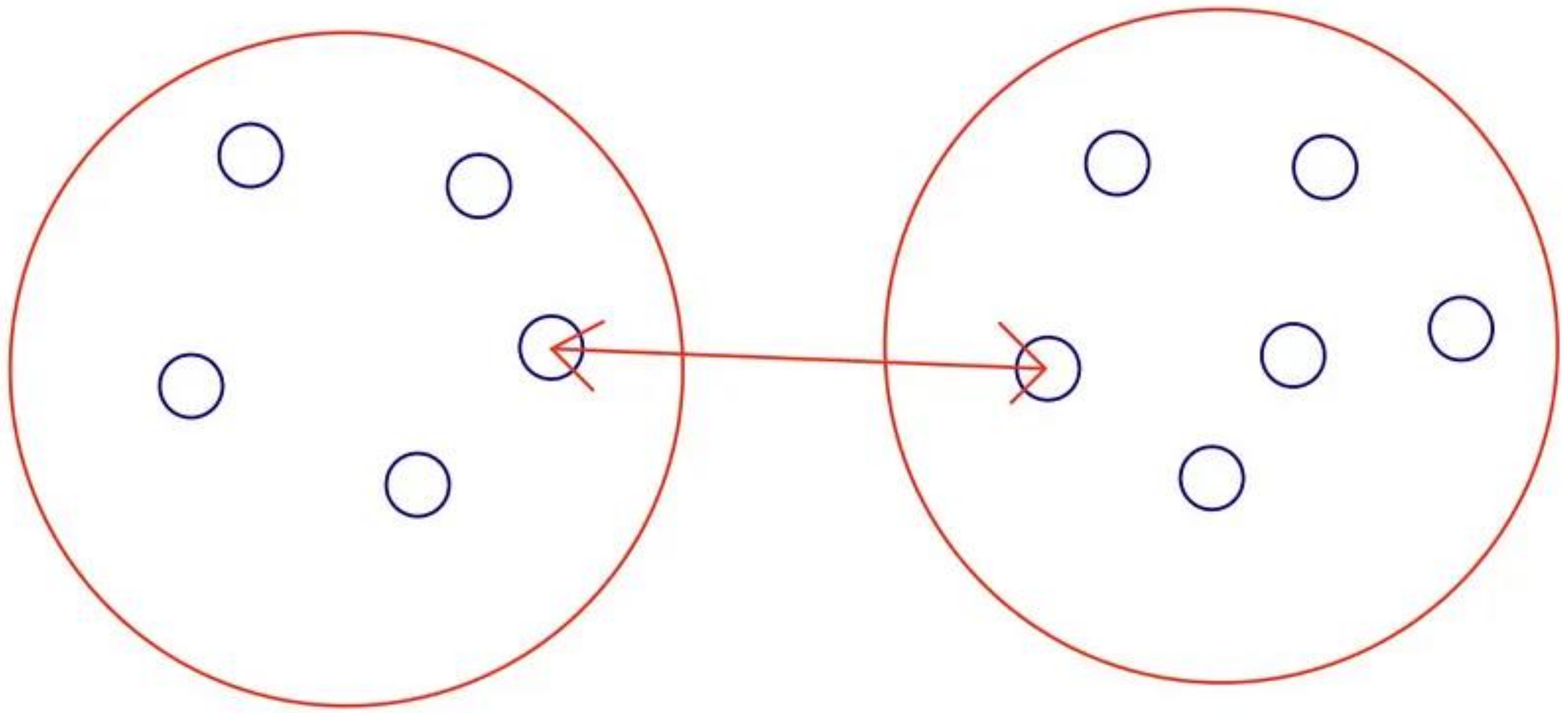


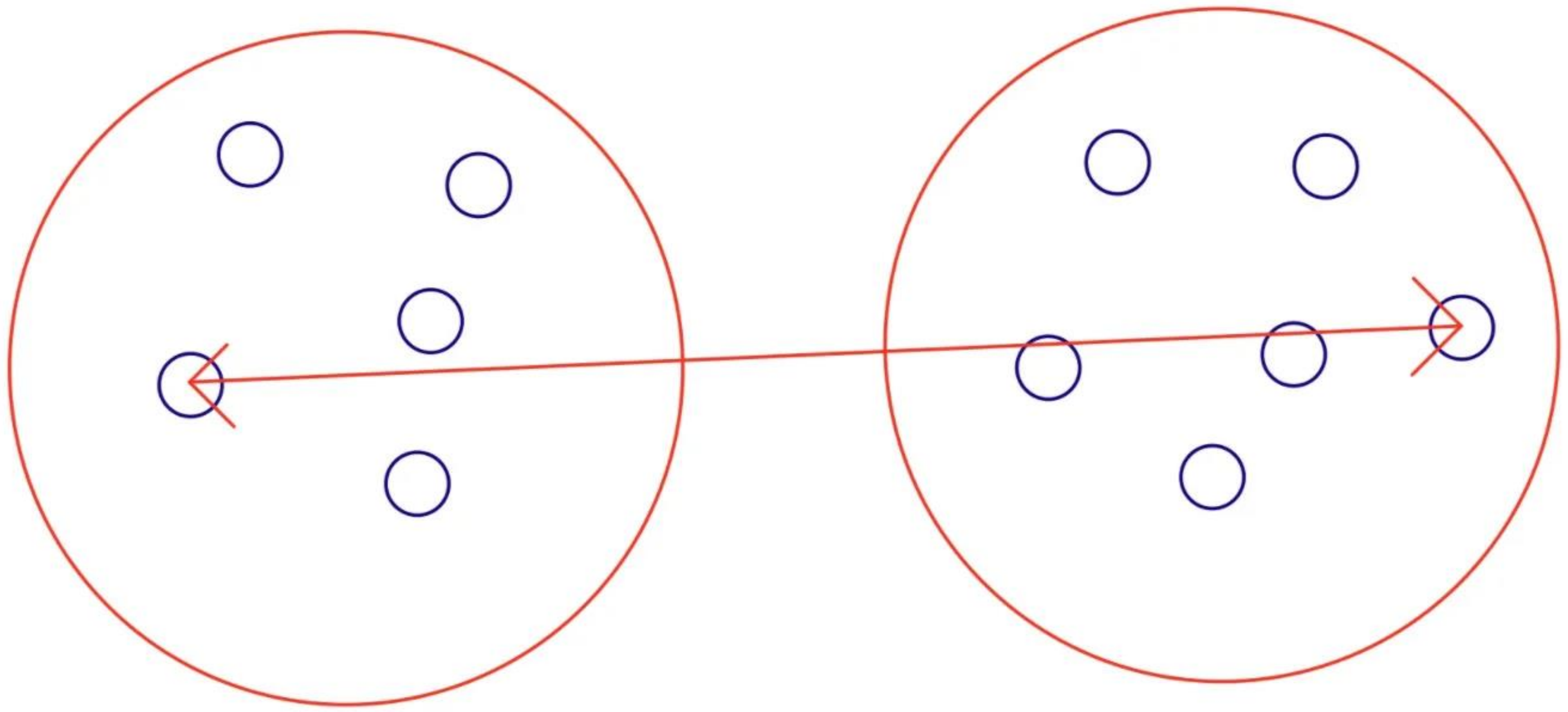

$X^2-4X\leq 0$
$n(B\cap C)=22$
$n(B)=68$
$n(C)=84$
$n(B\cup C)=n(B)+n(C)-n(B\cap C)$
$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}$
$\bar{x}_1=\frac{1+3+3+6+8+9}{6}=5$
$\bar{x}_2=2+4+4+8+12=30$
$\bar{x}_3=4+7+1+6=18$
$\log_b b^x=x$
$\log_a x=\frac{\log_b x}{\log_b a}$
20
6
He = 4.002602
Na = 22.989769
Ar = 39.948
$M=\frac{0.046765}{3.0L}$
(0,1)
$f=\frac{R}{2}$
$Zn_3Sb_2+6H_2O\rightarrow 3Zn(OH)_2+2SbH_3$
A
B
C

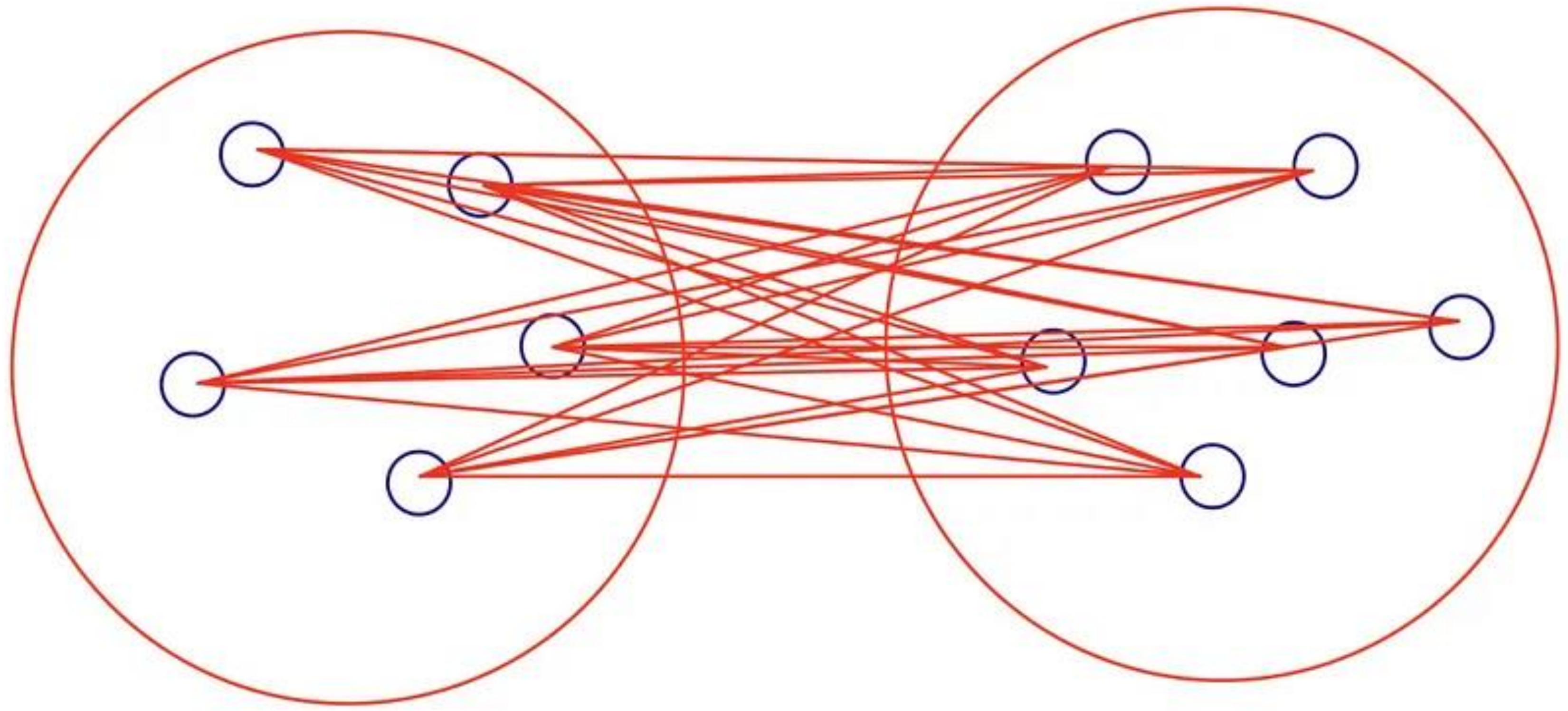
Recap

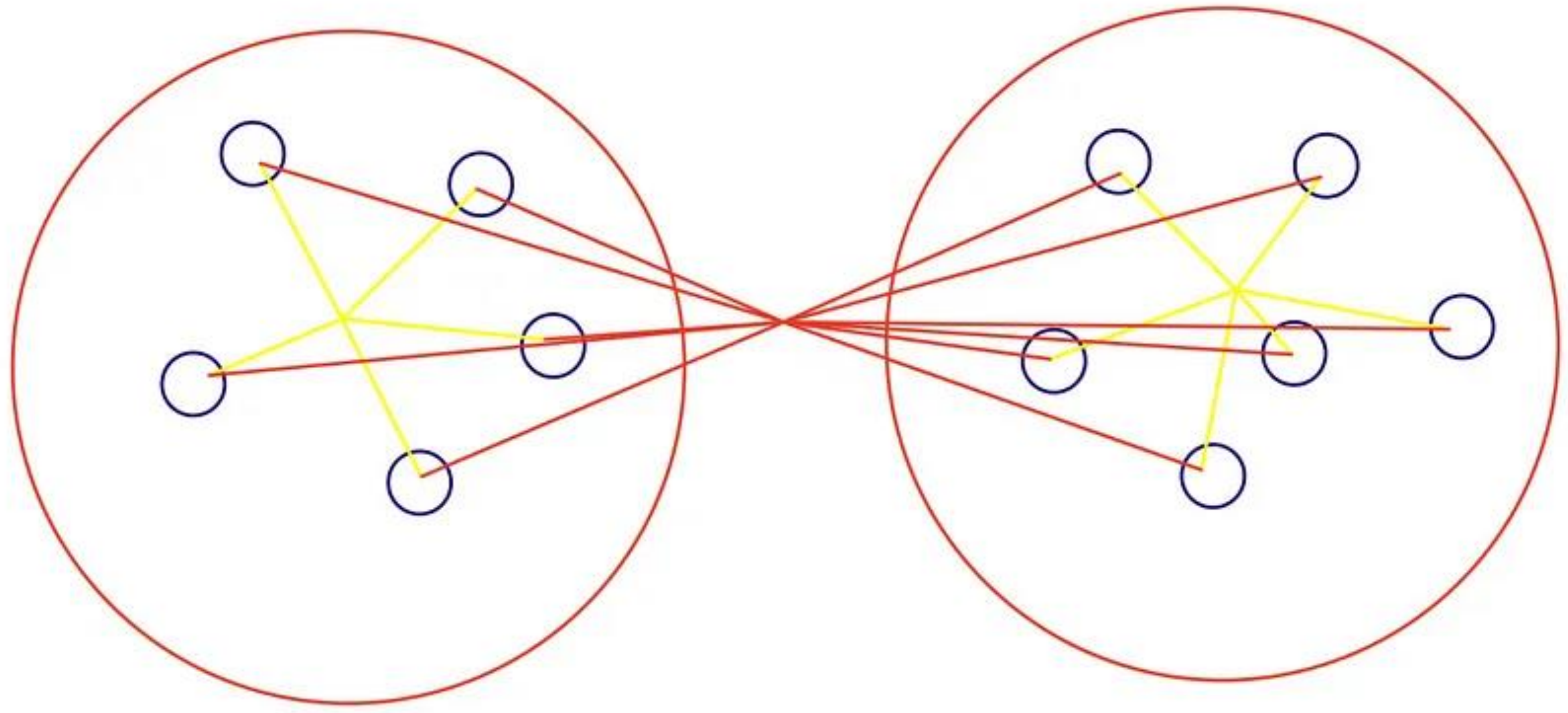

Divisive
Agglomerative


Decreases
Increases
Granularity
Increases
Decreases
Cluster size









Connectivity v/s other clustering algorithms

- Compactness (Cluster cohesion)
- Separation

$$Index = \frac{(\alpha \times Separation)}{(\beta \times Compactness)}$$

- Density – looks for a point within a epsilon distance
- Connectivity – No upper limit on nearest point distance
  - No good evaluation metric
  - Empirical or distance metric should capture domain knowledge


QUESTIONS