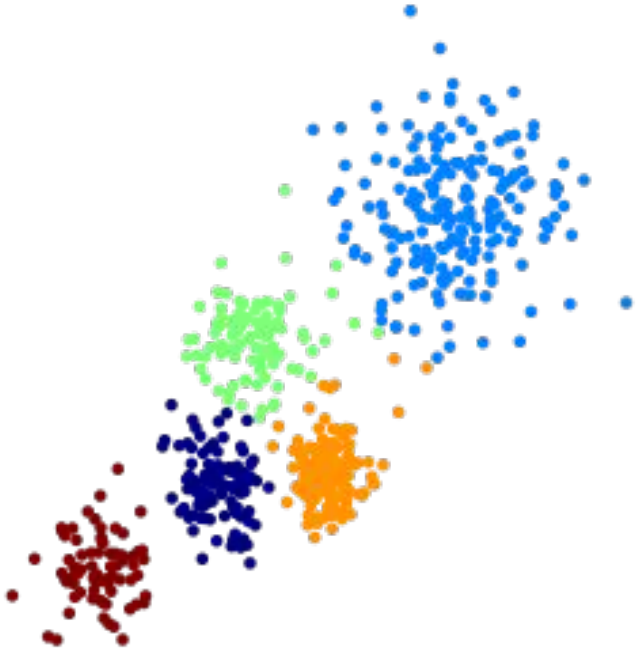
Clustering

Hierarchical clustering

Cary!!

DSI SEA, jf.omhover



Clustering

Hierarchical clustering

OBJECTIVES

- **Describe** and **implement** the HAC algorithm
- **Compare** purpose and utility of k-means and HAC
- **Discuss** the role of metrics for applying clustering to different problems
- **Analyze** how the (high) dimensionality of data impacts metrics based clustering techniques

Hierarchical Clustering

- Type of agglomerative clustering - meaning that we will iteratively group observations together based on their distance from one another.

- As we continue to group observations together we form a hierarchy of their similarity to one another.

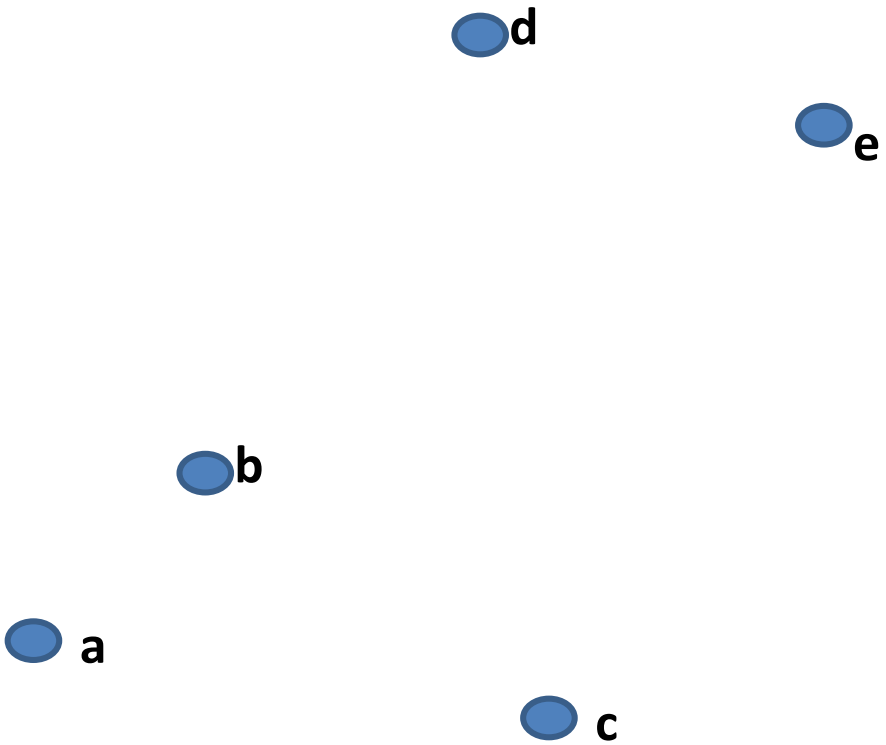
- This will force us to answer different questions then we did in k-Means.

- No longer do we have to choose the number of clusters up front.

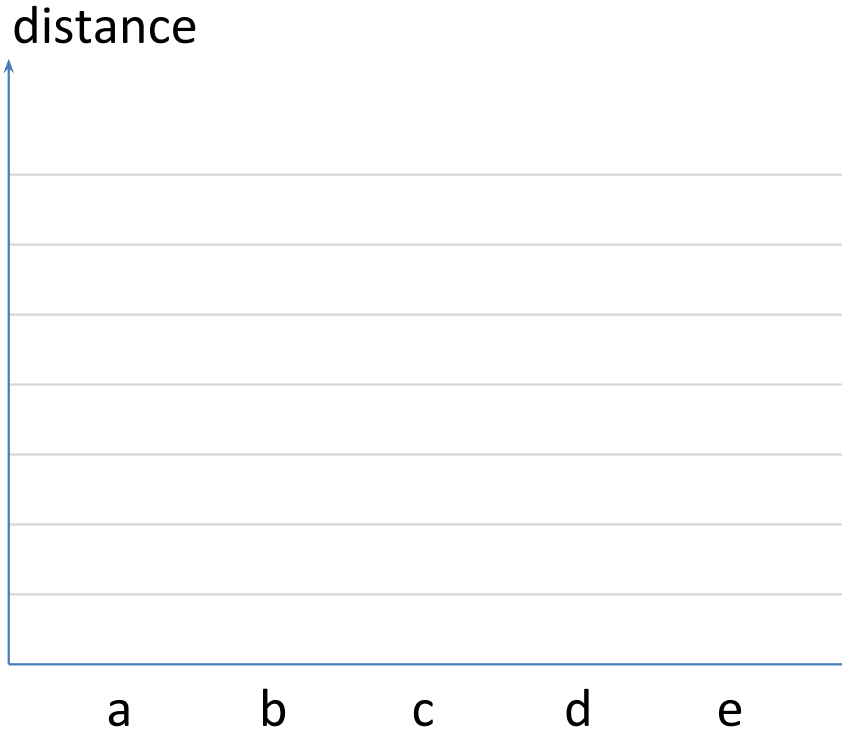
- Instead we'll have to define the nature of successive grouping of observations → define linkages.

Hierarchical Clustering (step by step)

1 - Computing distances between observations
2 – Identification / choose a minimum
3 – Fusion of observations



Observations

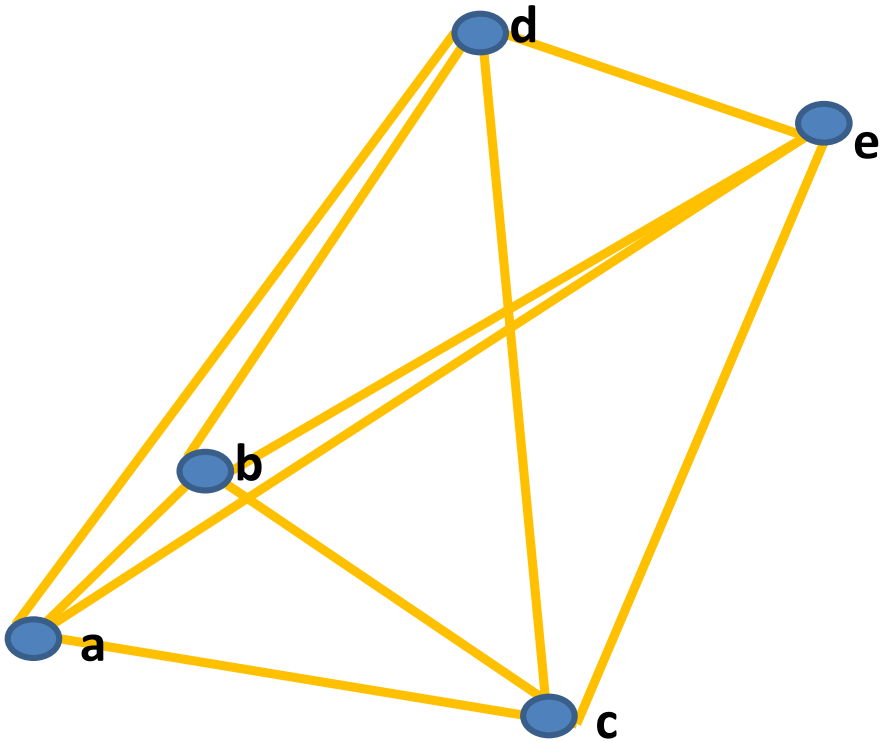


Dendrogram

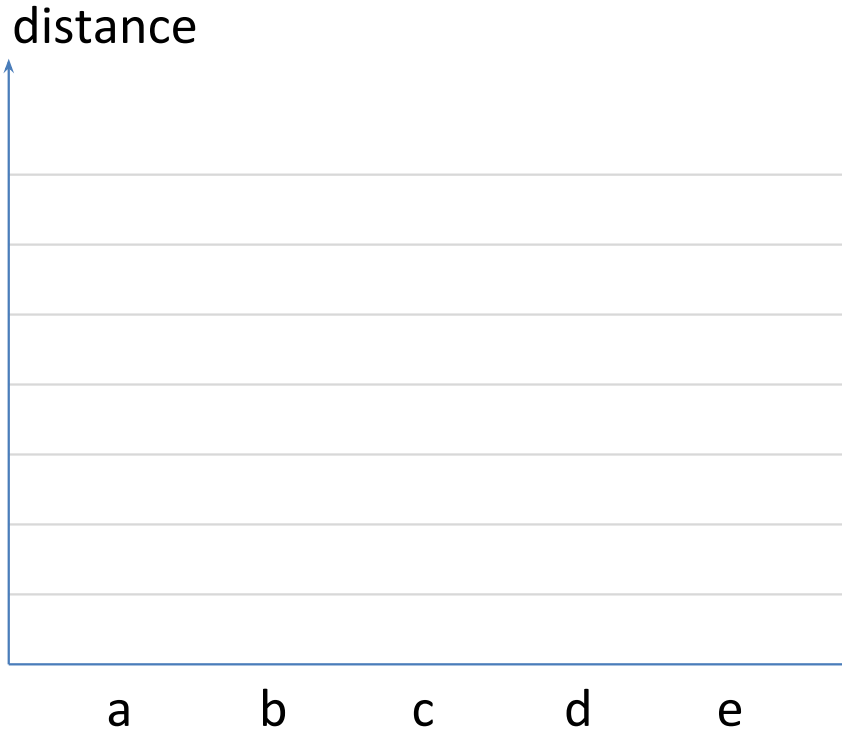
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

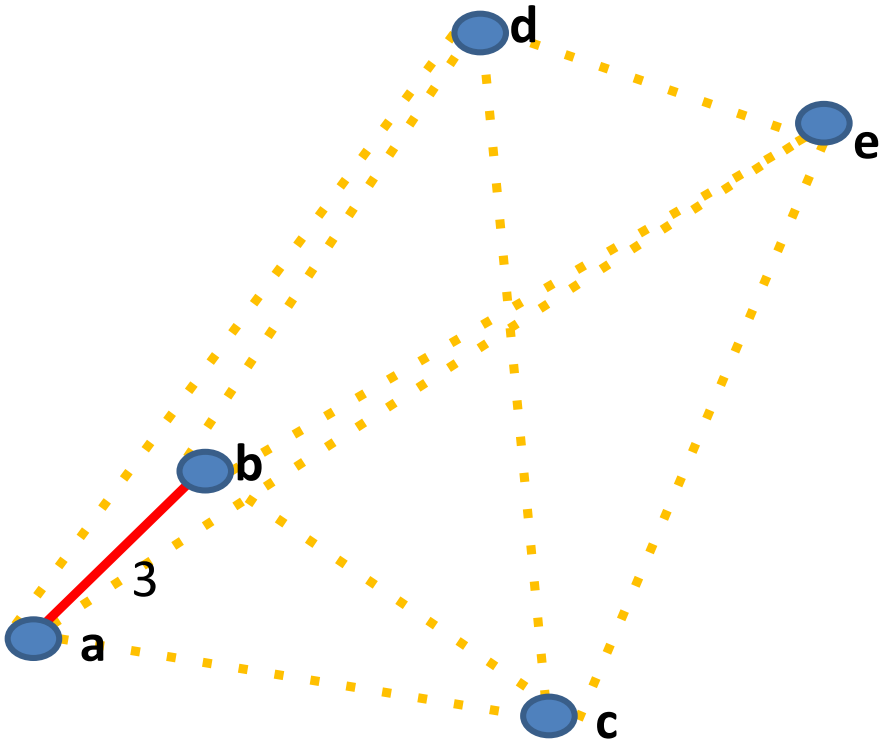


Dendrogram

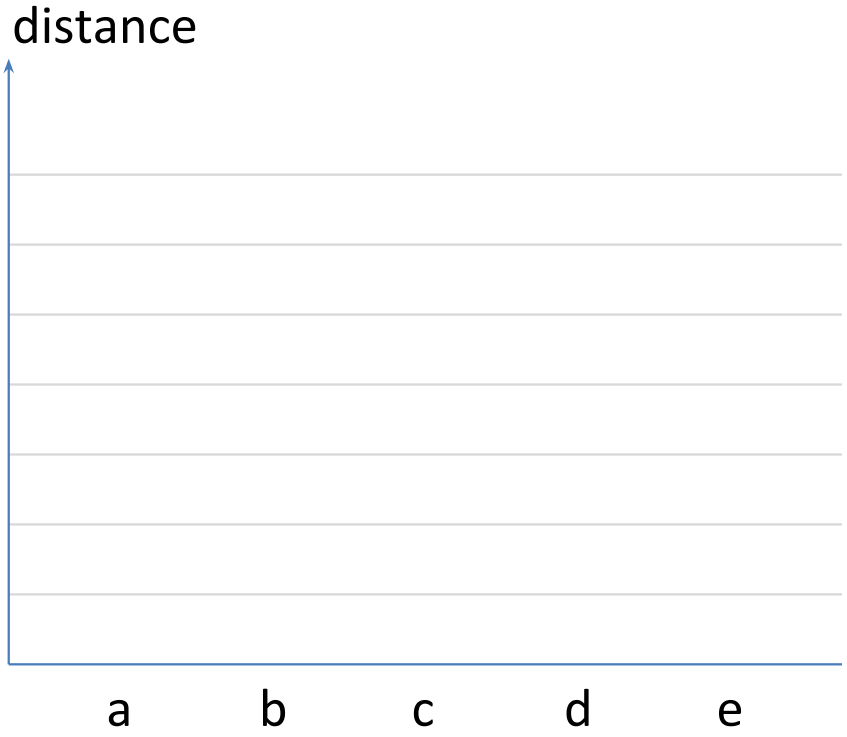
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations

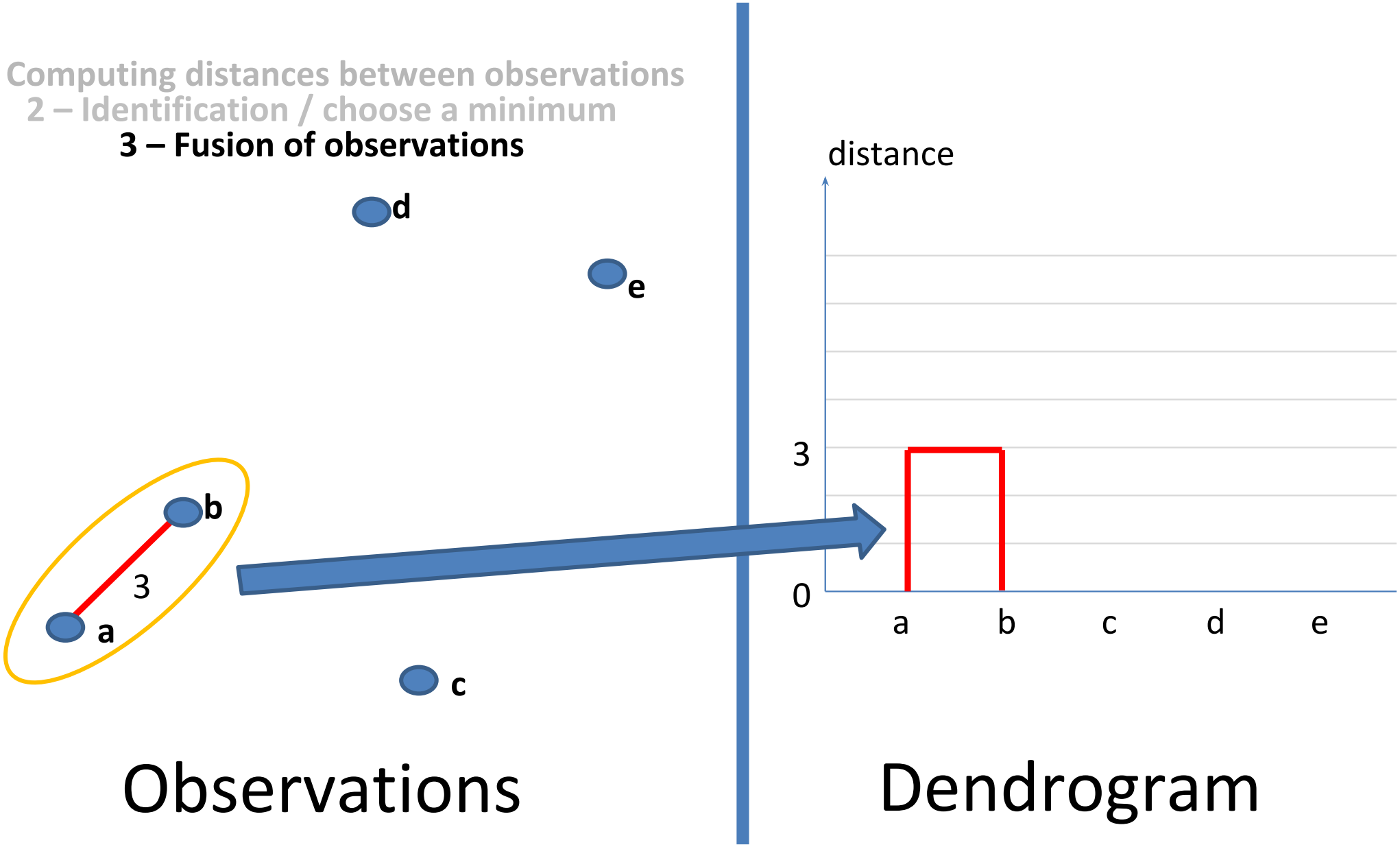


Observations



Dendrogram

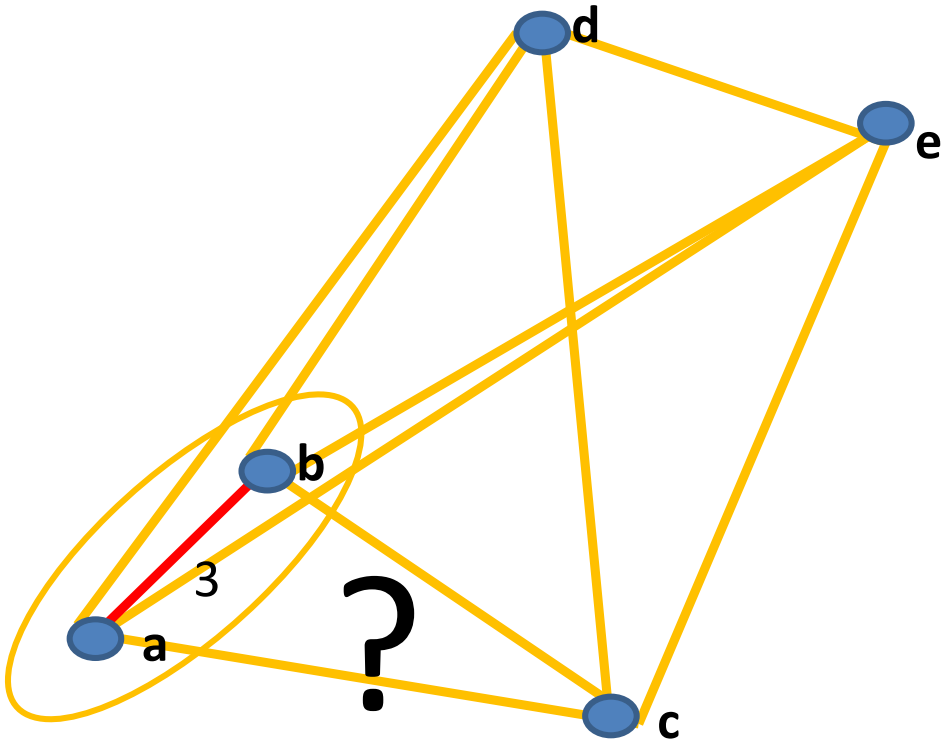
1 - Computing distances between observations
2 – Identification / choose a minimum
3 – Fusion of observations



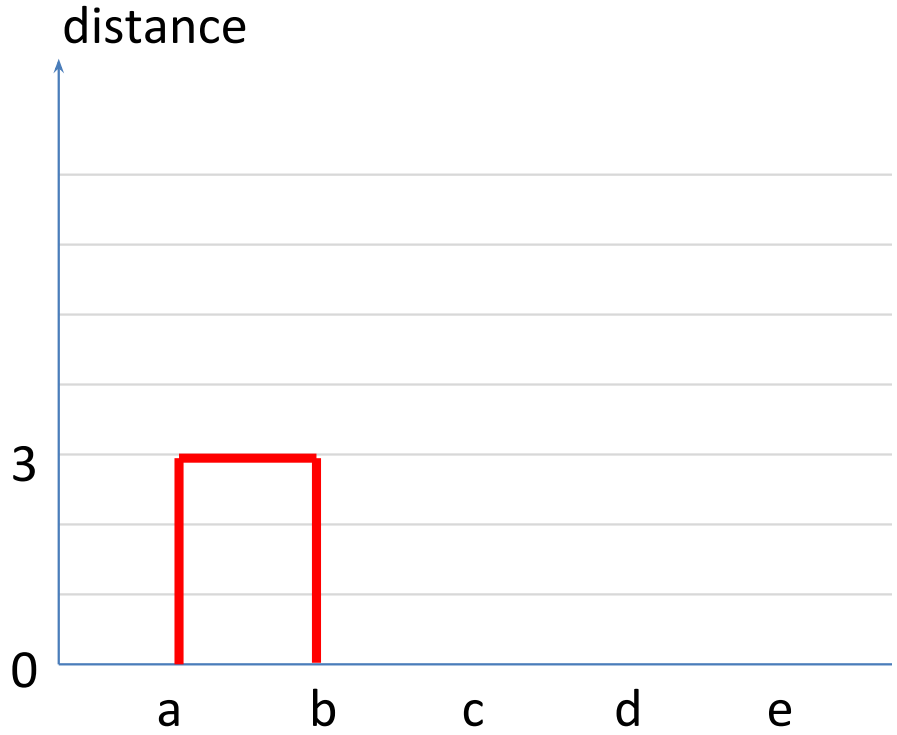
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

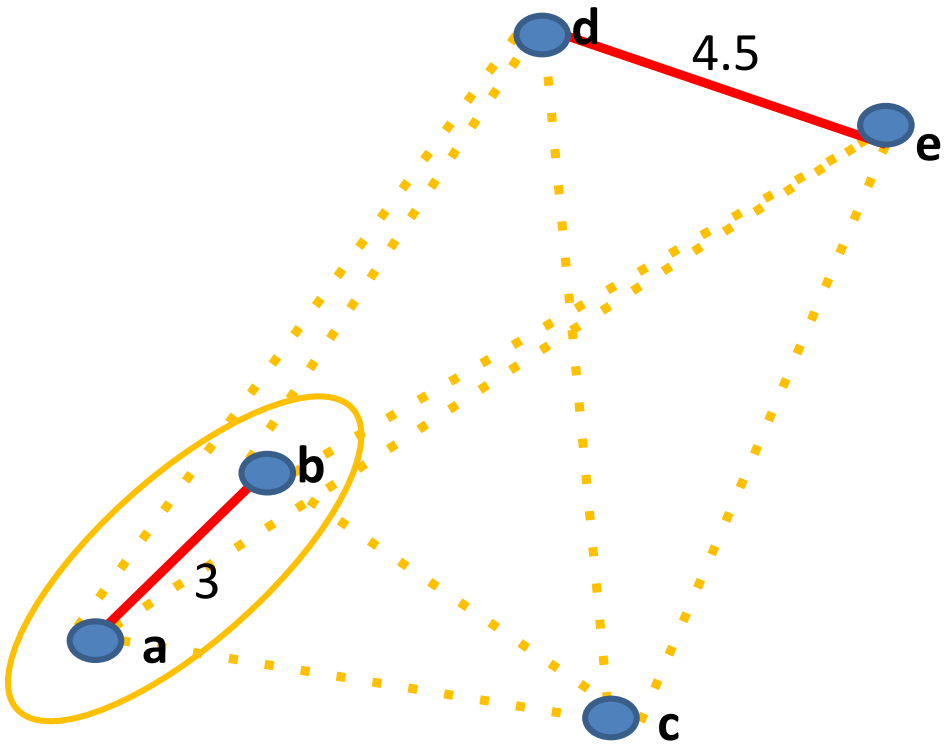


Dendrogram

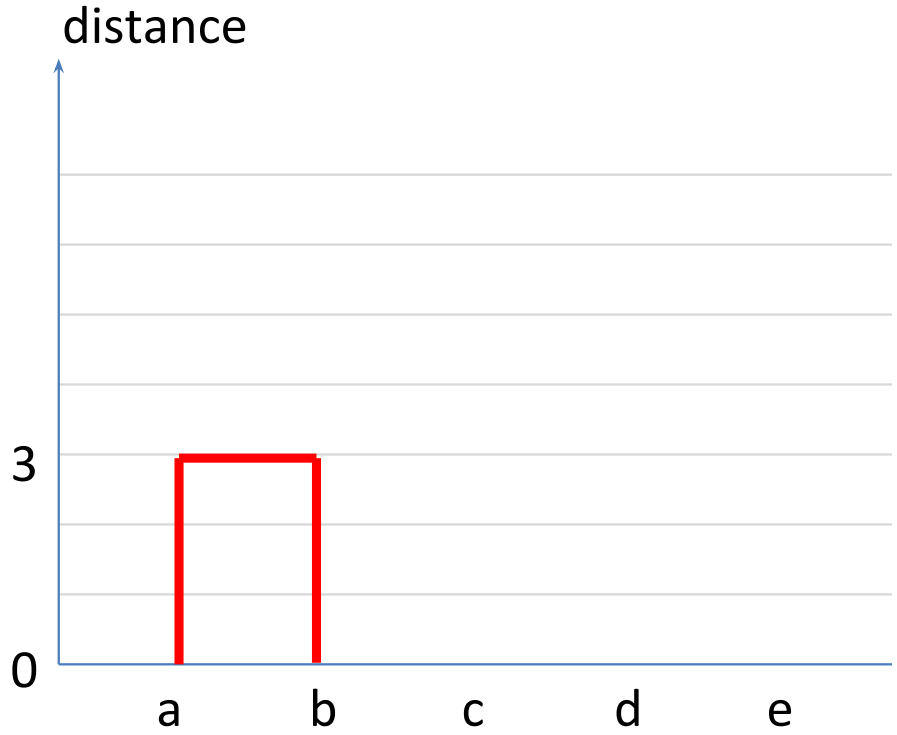
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

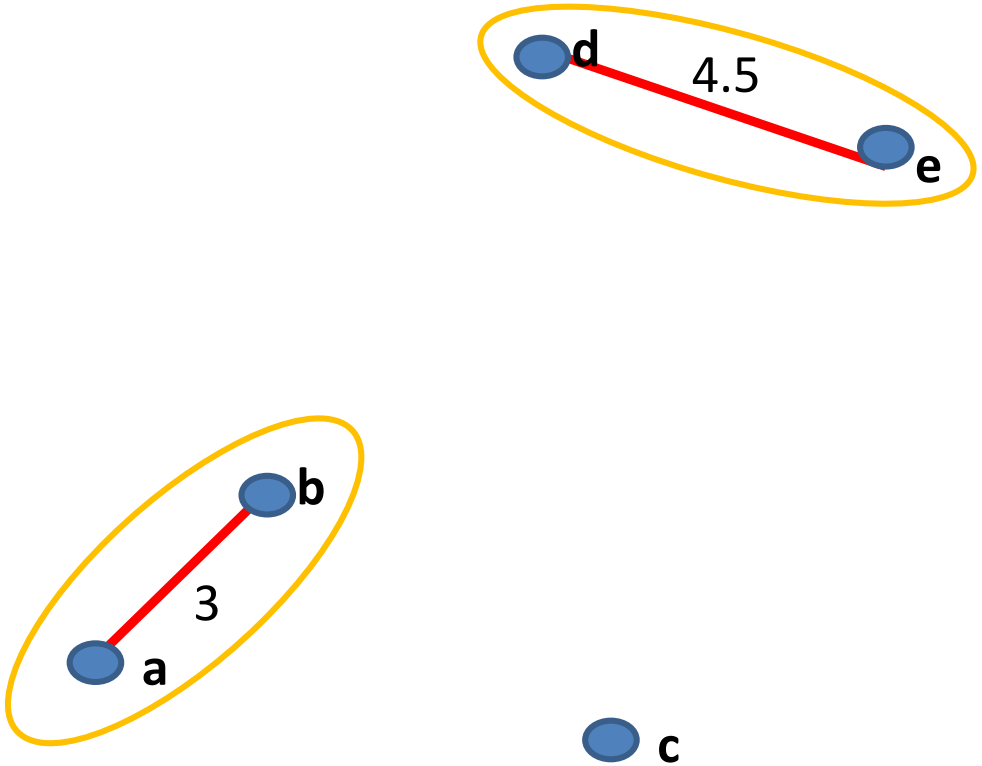


Dendrogram

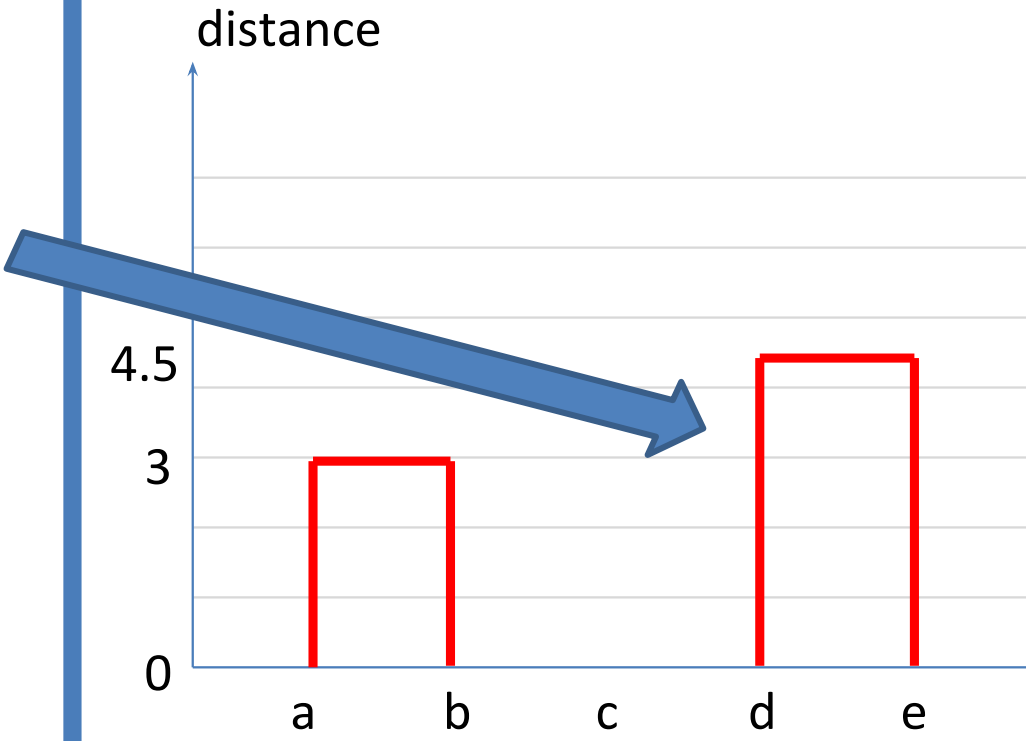
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

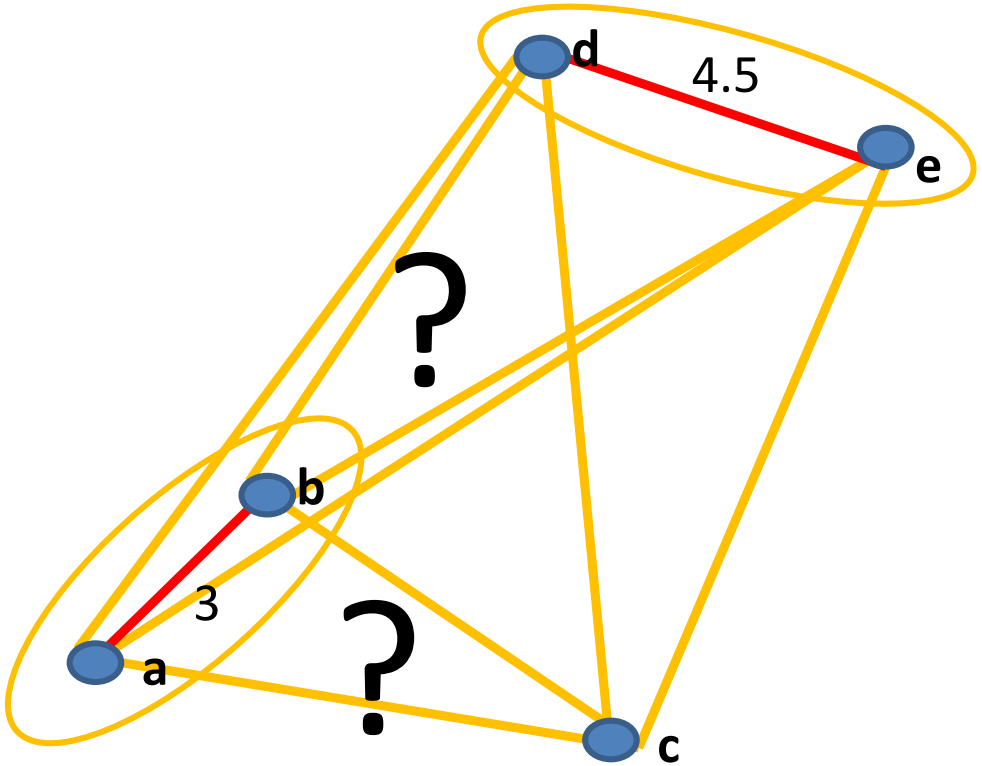


Dendrogram

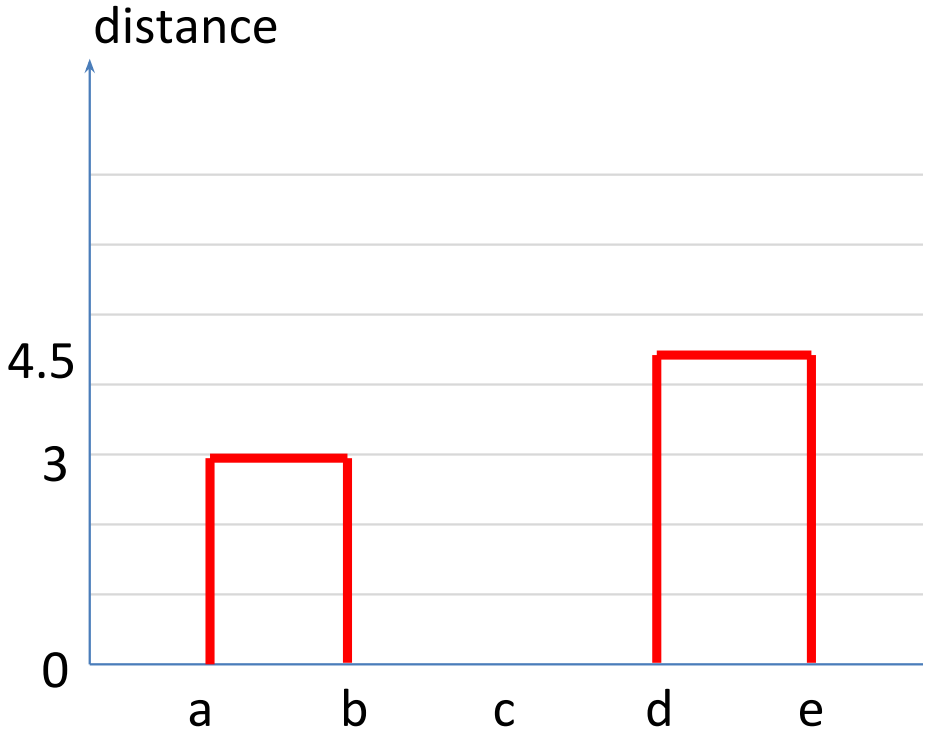
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

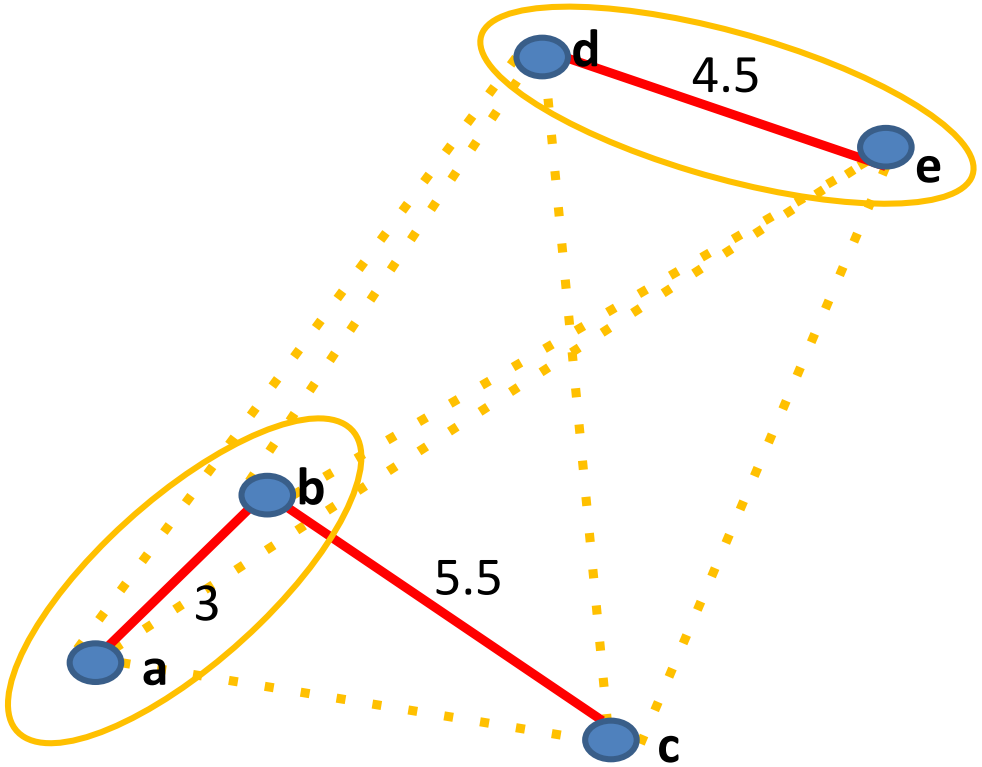


Dendrogram

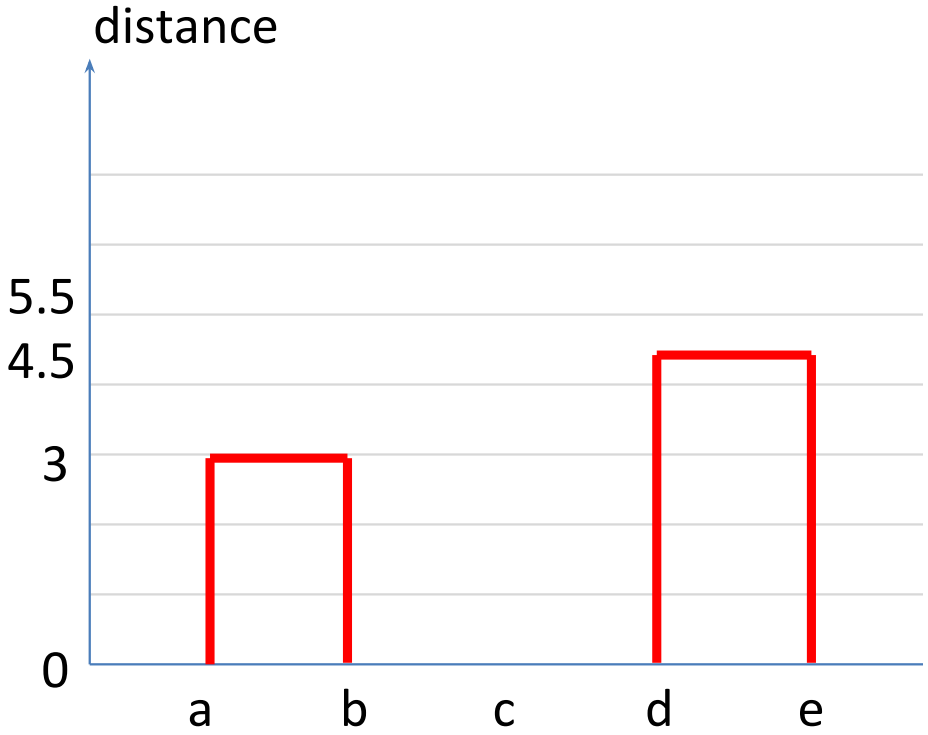
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations

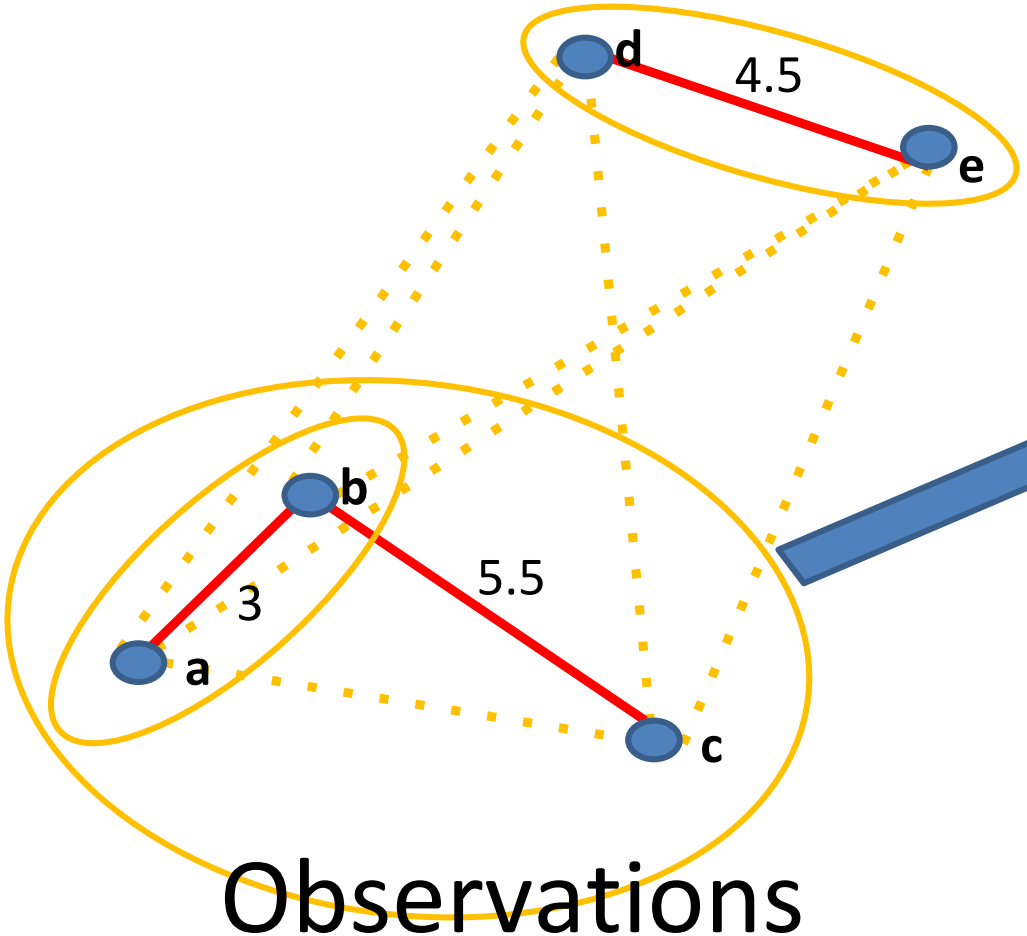


Observations



Dendrogram

1 - Computing distances between observations
2 – Identification / choose a minimum
3 – Fusion of observations



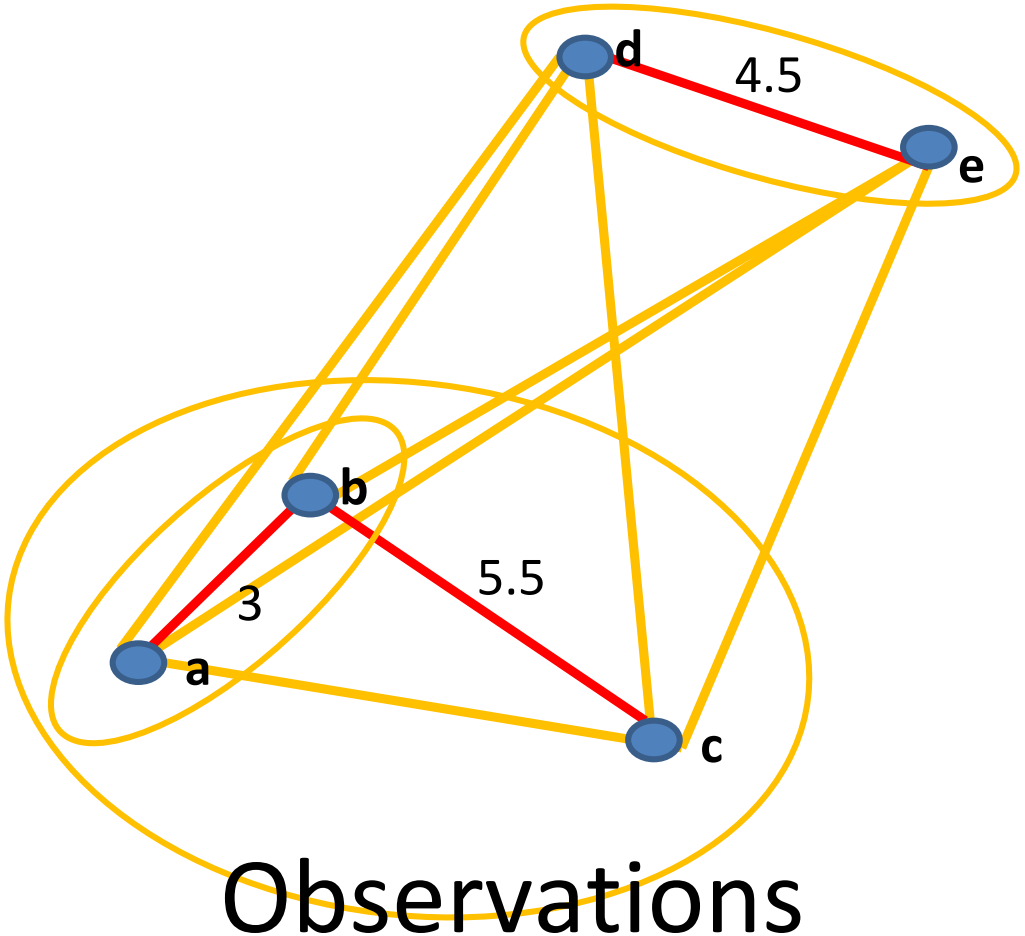
Observations

Dendrogram

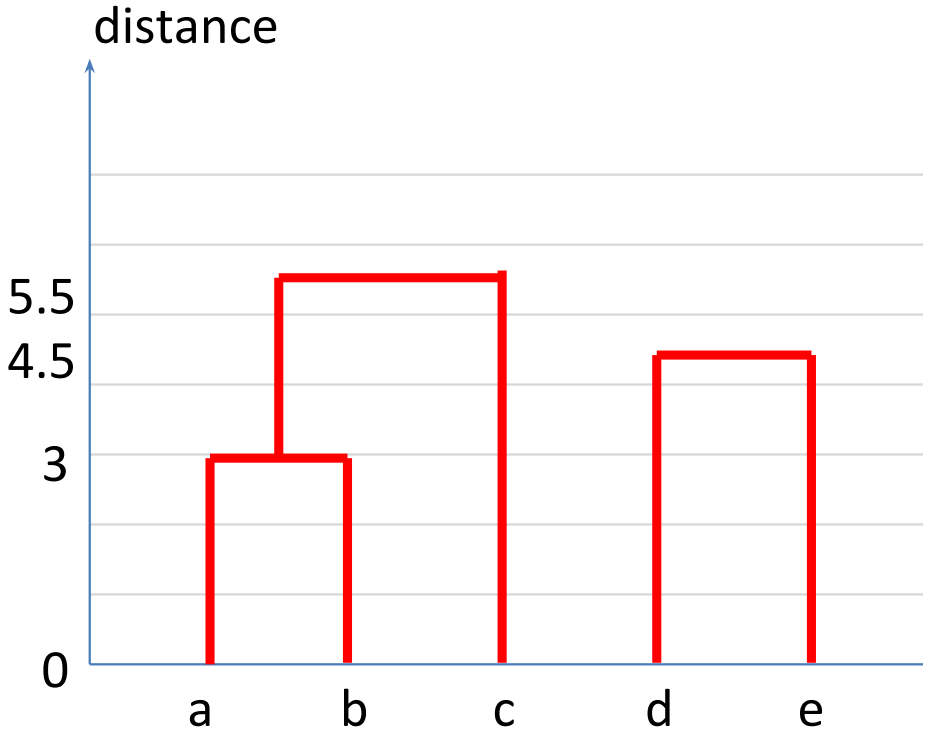
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

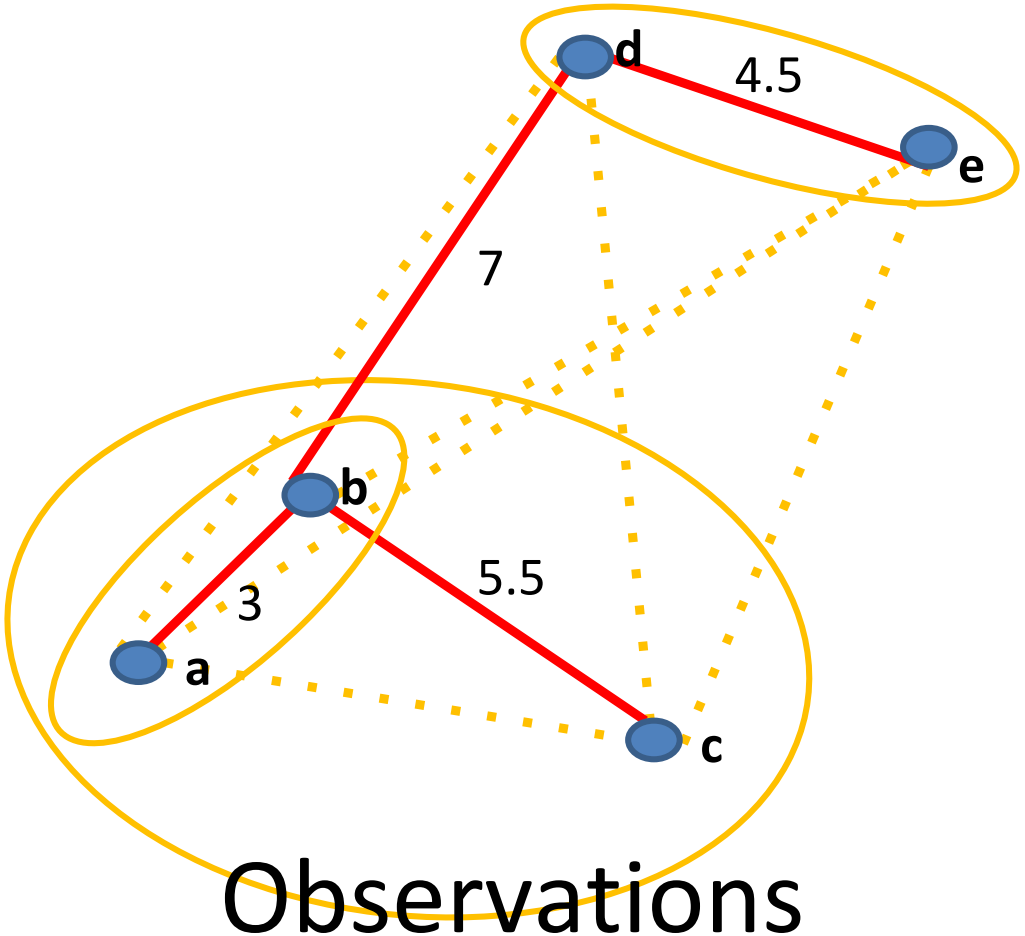


Dendrogram

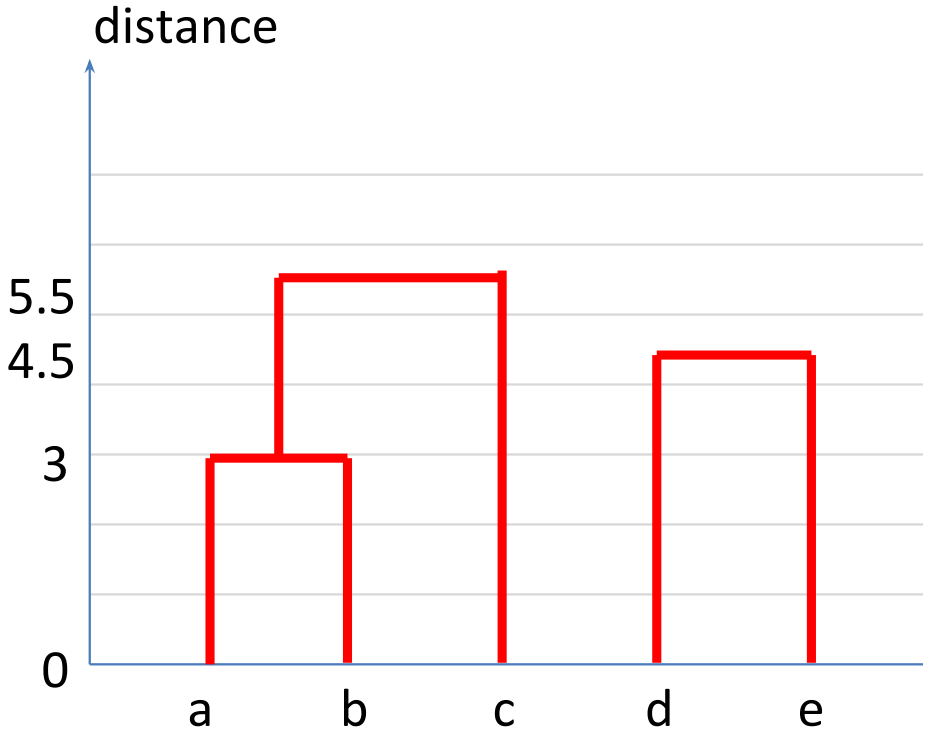
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

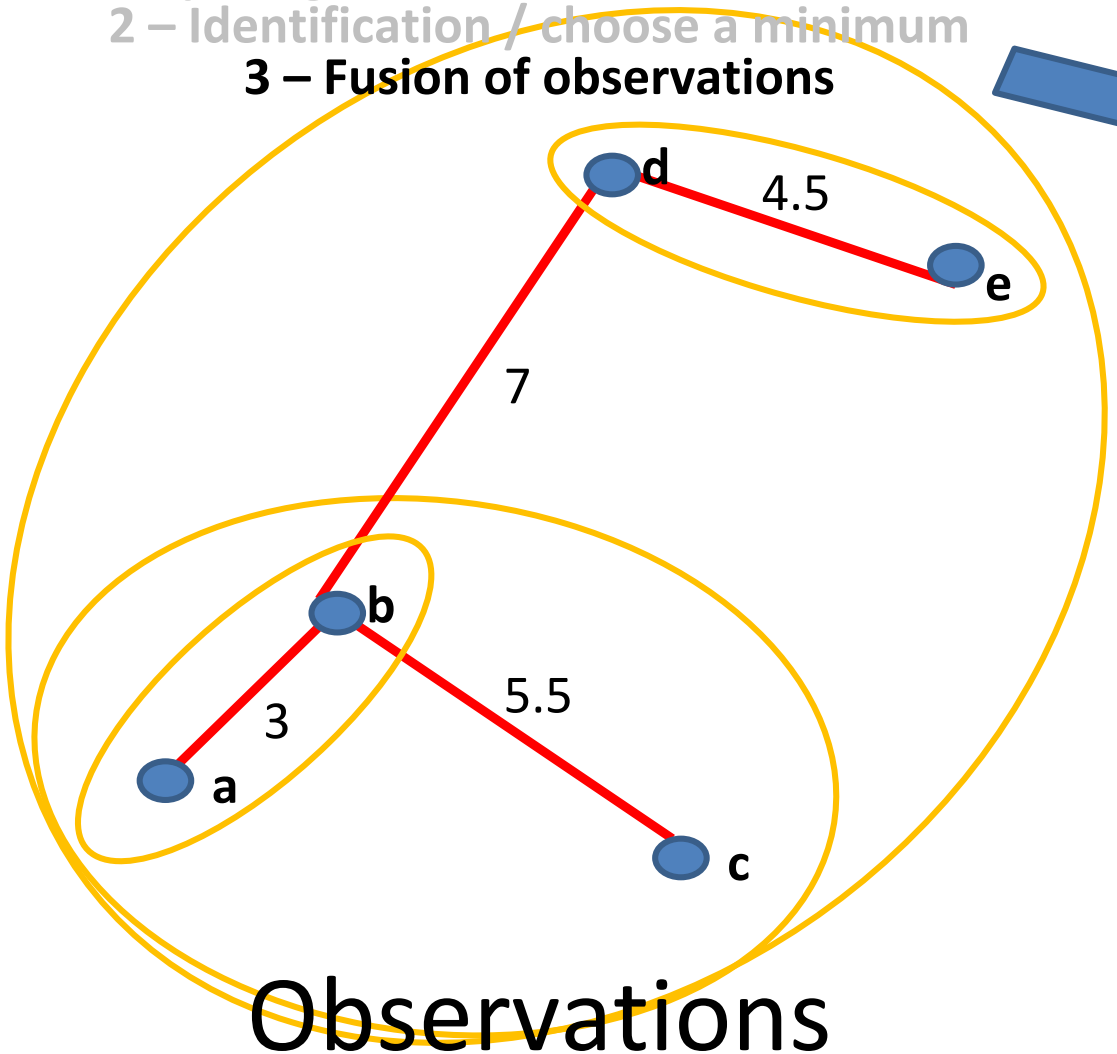


Dendrogram

1 - Computing distances between observations

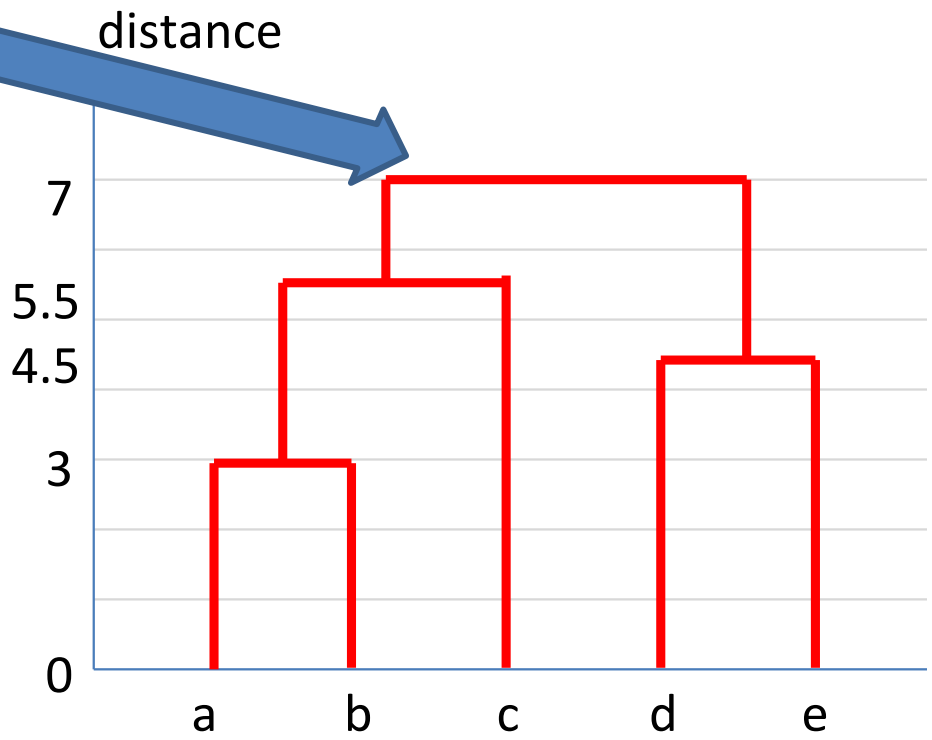
2 – Identification / choose a minimum

3 – Fusion of observations



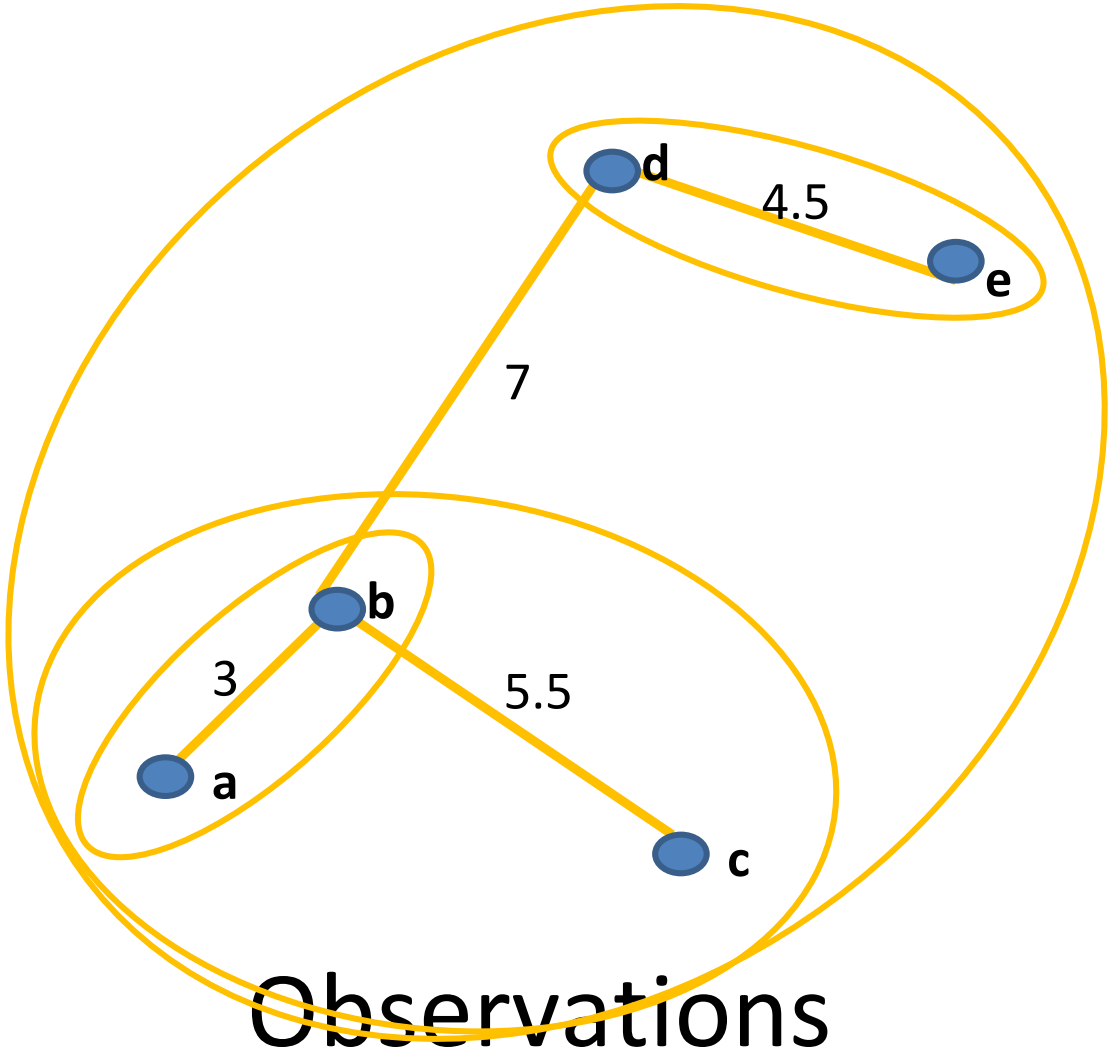
Observations

distance

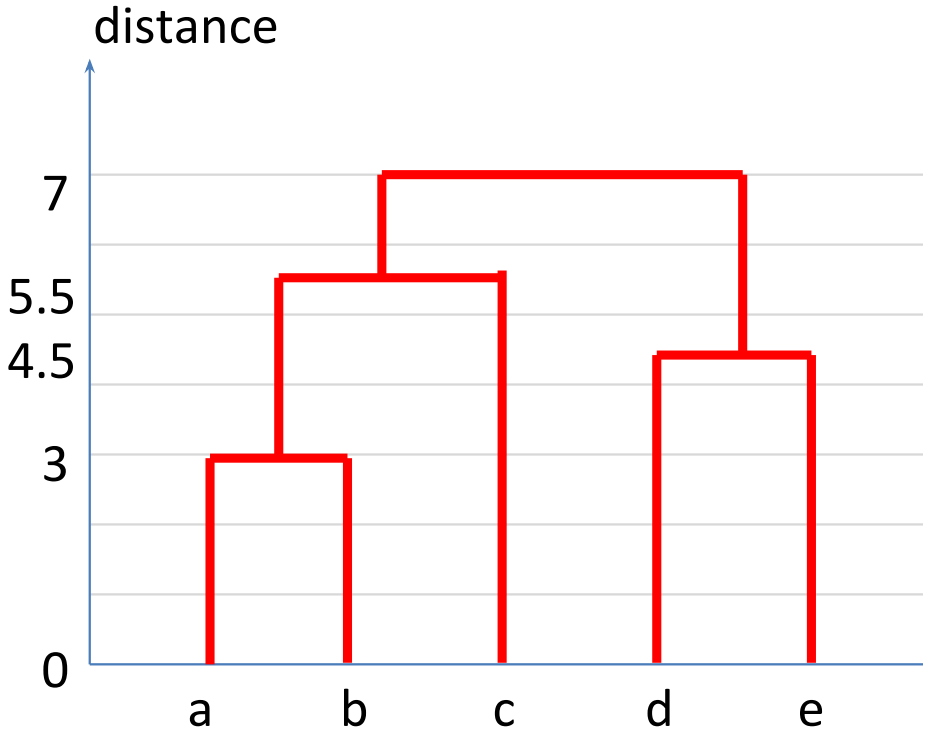


STOP !

Dendrogram



Observations



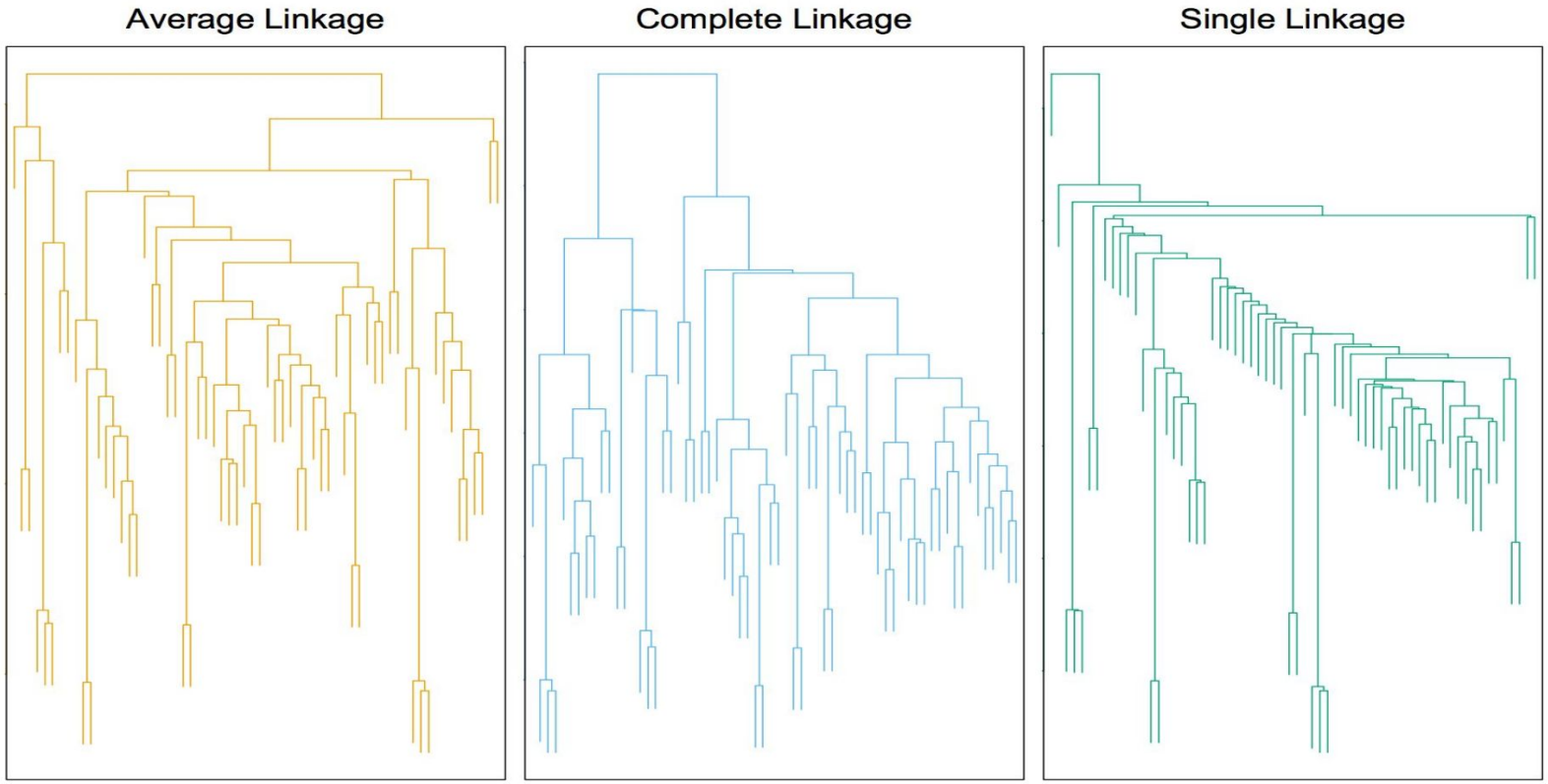
Dendrogram

HAC Linkage

How do we define dissimilarity between clusters?

- **Complete:** Maximum pairwise dissimilarity between points in clusters – good
- **Average:** Average of pairwise dissimilarity between points in clusters – also good
- **Single:** Minimum pairwise dissimilarity between points in clusters – not as good; can lead to long narrow clusters

Linkage on Dendrograms



- Not too sensitive to outliers
- Compromise between complete linkage and single

- More sensitive to outliers
- May violate “closeness”

- Less sensitive to outliers
- Handles irregular shapes fairly naturally

Metrics / Distances / Similarities

Metrics

Distance

$$d : X \times X \to [0, \infty),$$

1. $d(x, y) \geq 0$ non-negativity or separation axiom
2. $d(x, y) = 0 \Leftrightarrow x = y$ identity of indiscernibles
3. $d(x, y) = d(y, x)$ symmetry
4. $d(x, z) \leq d(x, y) + d(y, z)$ subadditivity or triangle inequality

Similarity Measure [Tversky]

Increases with the quantity of common features between A and B
Decreases with the quantity of features that are specific to A, specific to B

How would you measure the similarity between...

- Vectors in a data array
- TFIDF vectors
- Sets (Bags / Transactions)
- Strings

Similarity between… TFIDF vectors

- Occurrences / tfidf
- Only positive values

- Cosine Similarity

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Similarity between… sets

- Jaccard Similarity

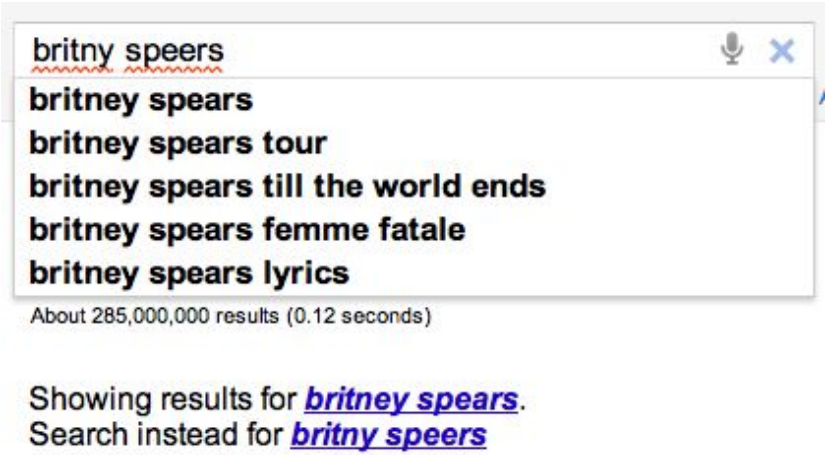
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Similarity between… strings

488941	britney spears
40134	brittany spears
36315	brittney spears
24342	britany spears
7331	britny spears
6633	briteny spears
2696	britteny spears
1807	briney spears
1635	brittny spears

...

[source]



[source]

=> EDIT DISTANCE
How many editions (add/sub/switch) are needed at the least to transform one string into another ?

! Can be applied to sequences of clicks

Pair Assignment