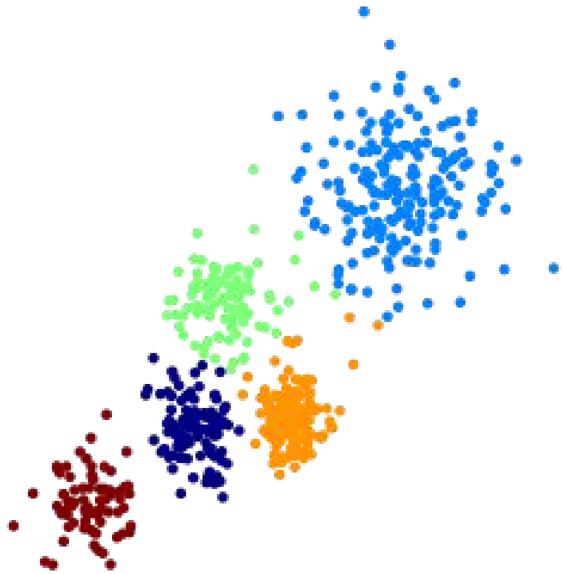
Clustering

K-Means
& Hierarchical Clustering

Natalie Hunt



K-Means

Objective: minimize WCSS
“within cluster sum of squares”

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

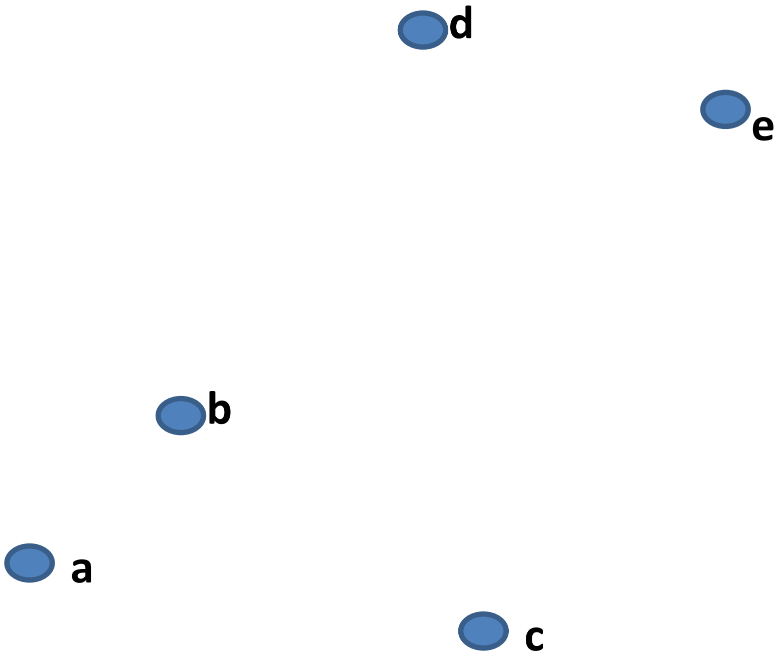
1. Randomly assign a number, from 1 to K, to each of the observations.
2. **Iterate** until the cluster assignments stop changing:
   a. For each of the K clusters, compute the cluster ***centroid***: the vector of the *p* features **means** for the observations in the k-th cluster
   b. **Assign** each observation to the cluster whose centroid is **closest** (defined using Euclidian distance)

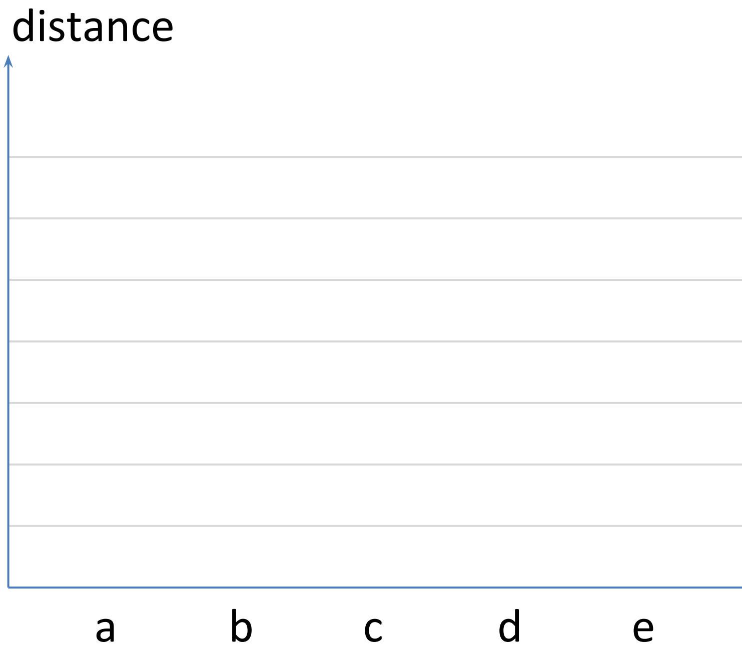
K-Means in a nutshell :

- **Computing distances**
- **Computing means**

Hierarchical Clustering (step by step)

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





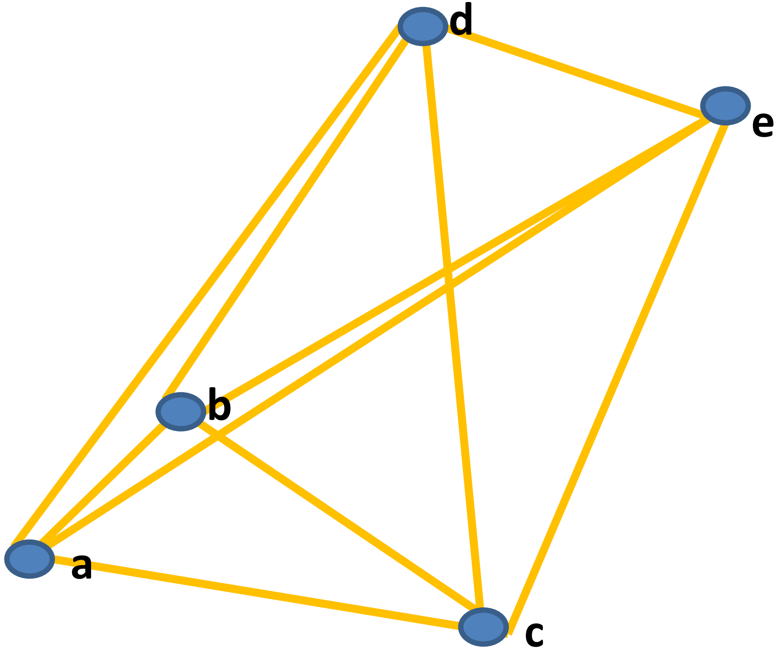
Observations

Dendrogram

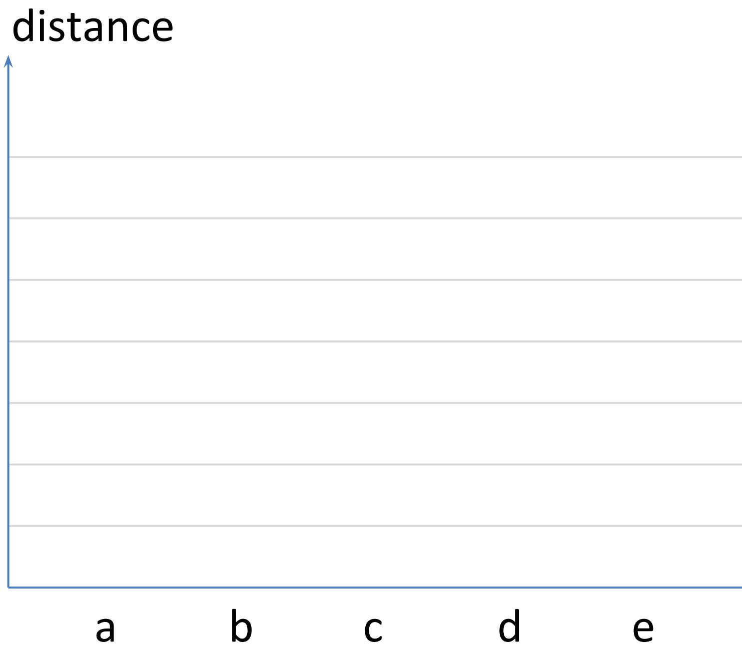
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations

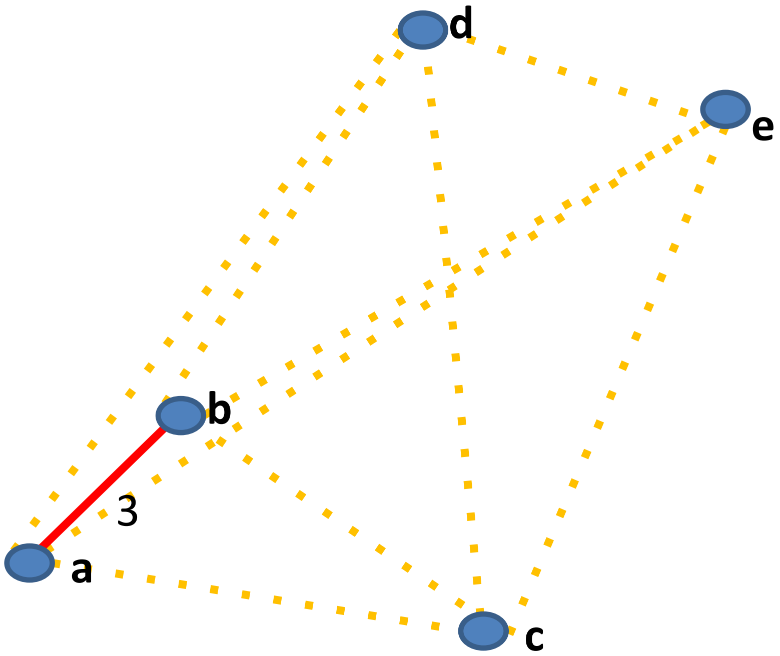


Observations

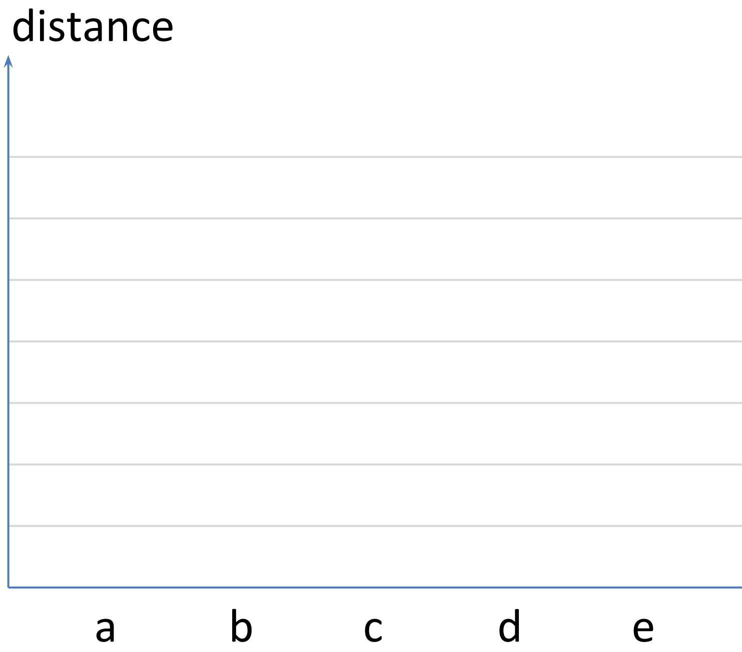


Dendrogram

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



Observations

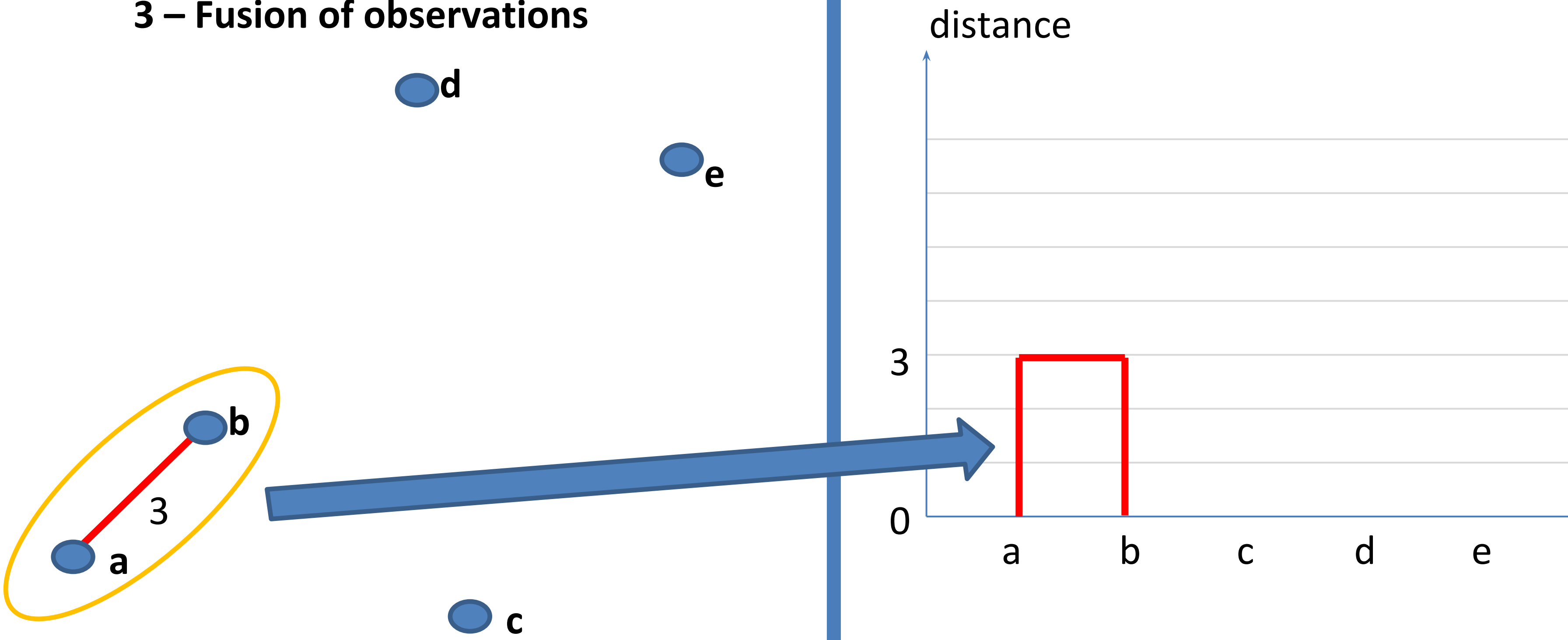


Dendrogram

1 - Computing distances between observations

2 – Identification / choose a minimum

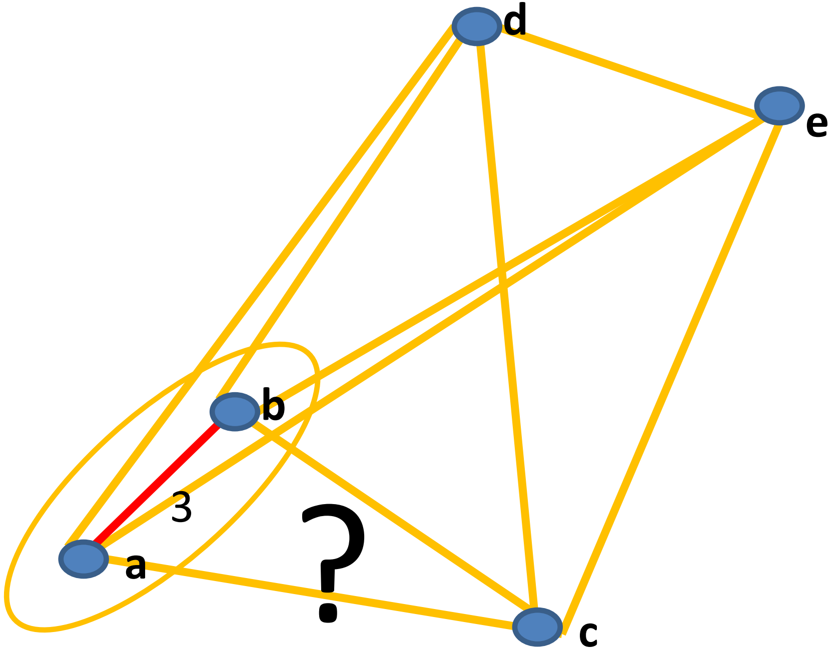
3 – Fusion of observations

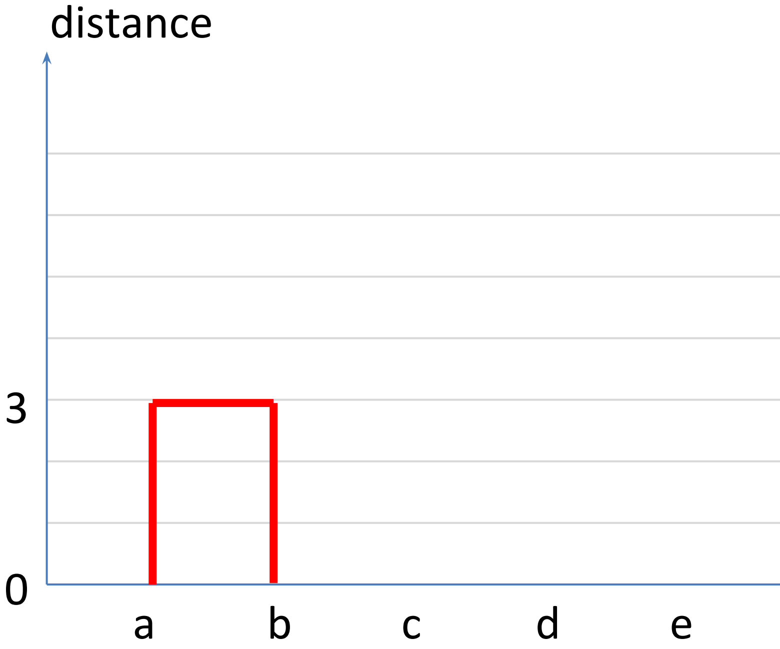


Observations

Dendrogram

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





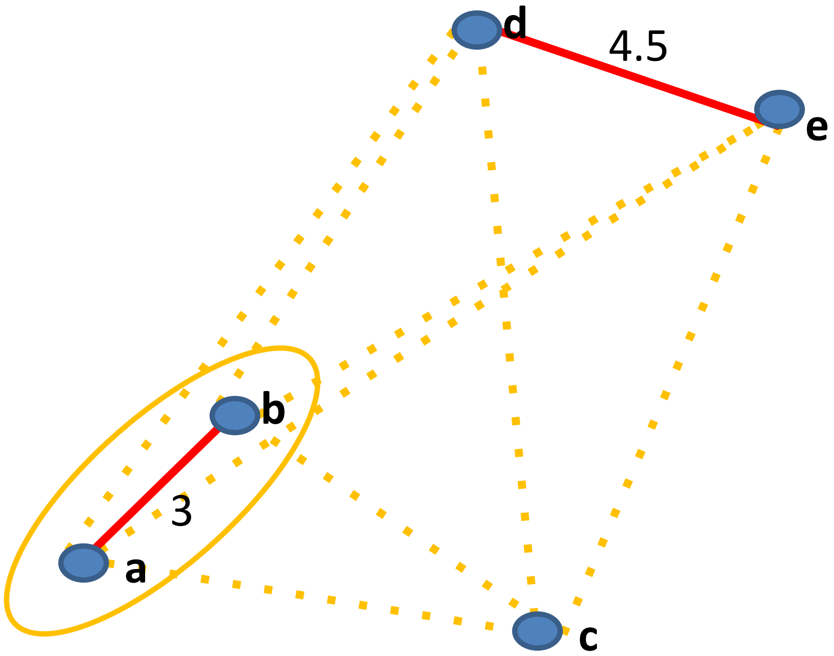
Observations

Dendrogram

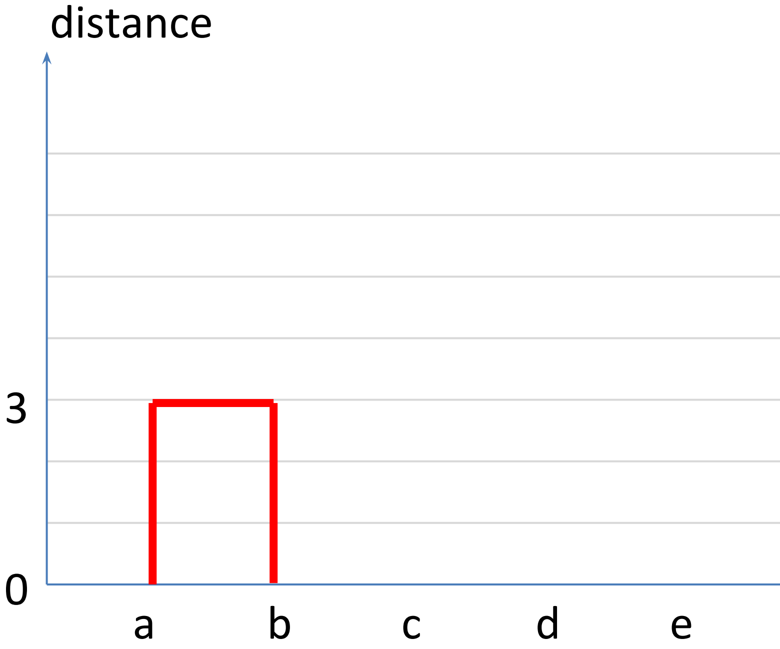
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

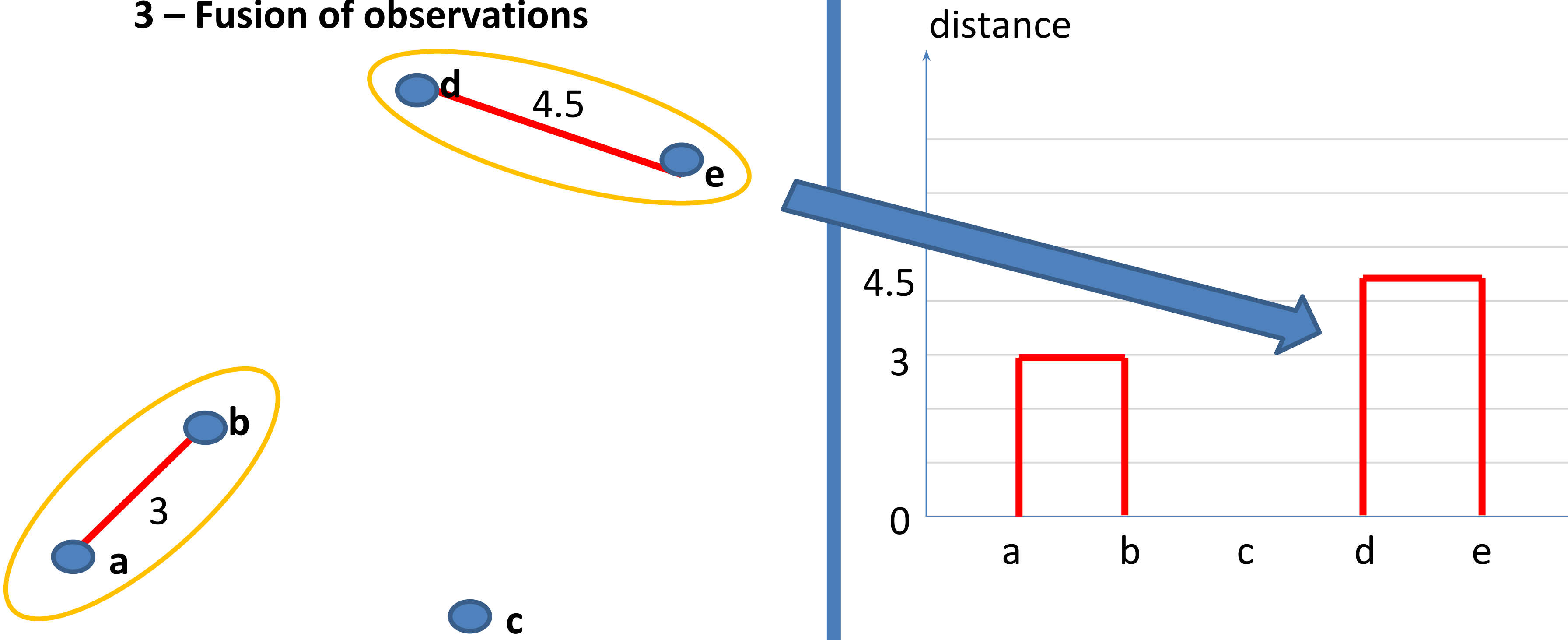


Dendrogram

1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



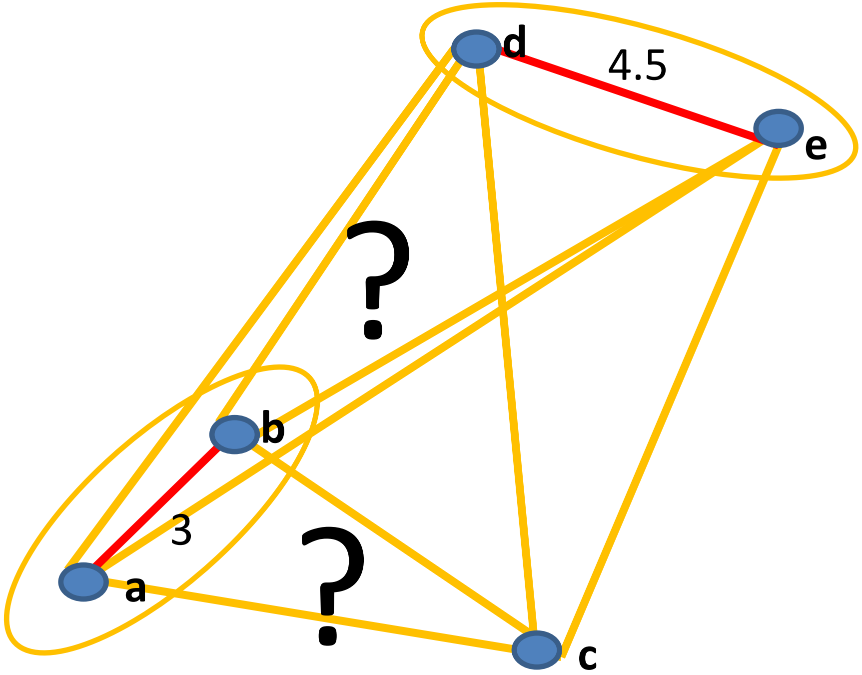
Observations

Dendrogram

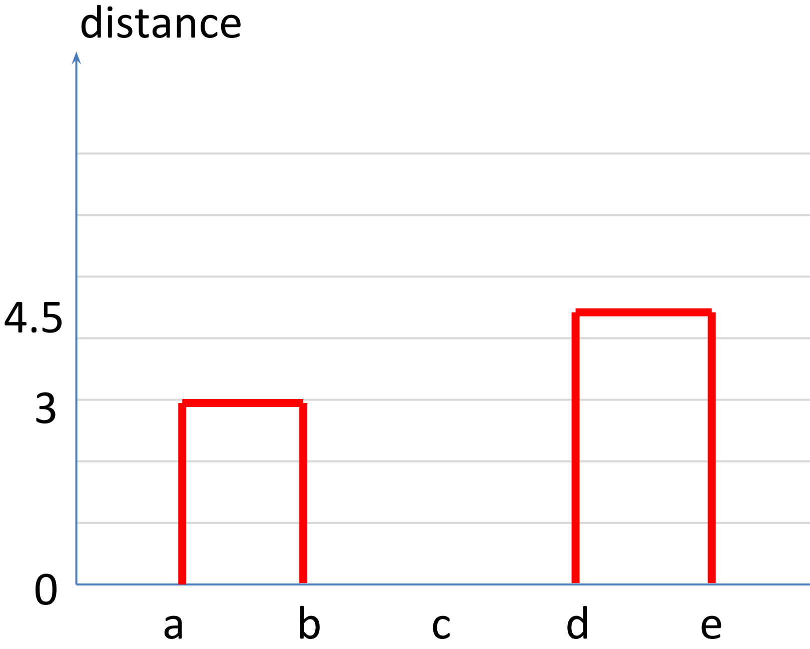
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations



Observations

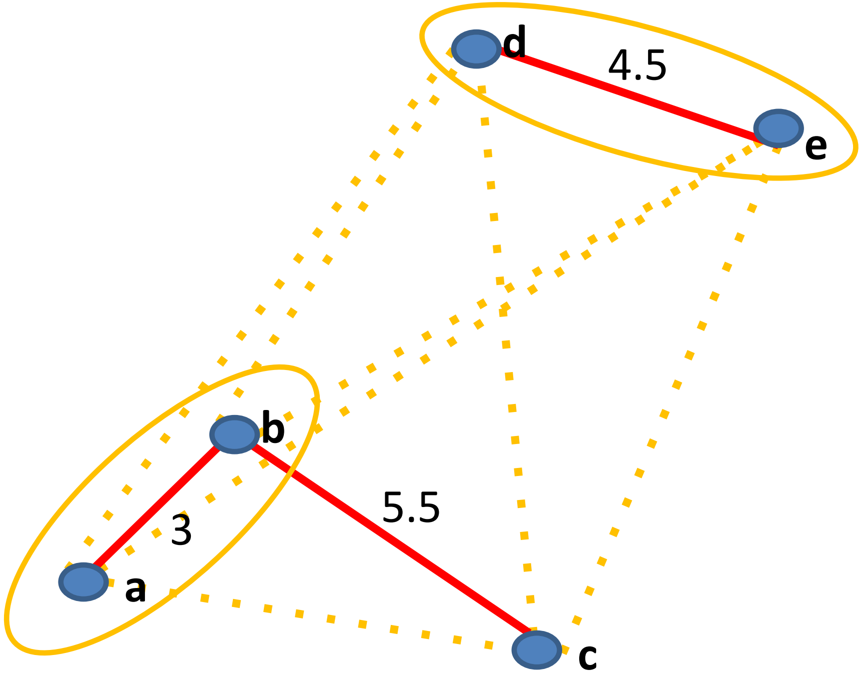


Dendrogram

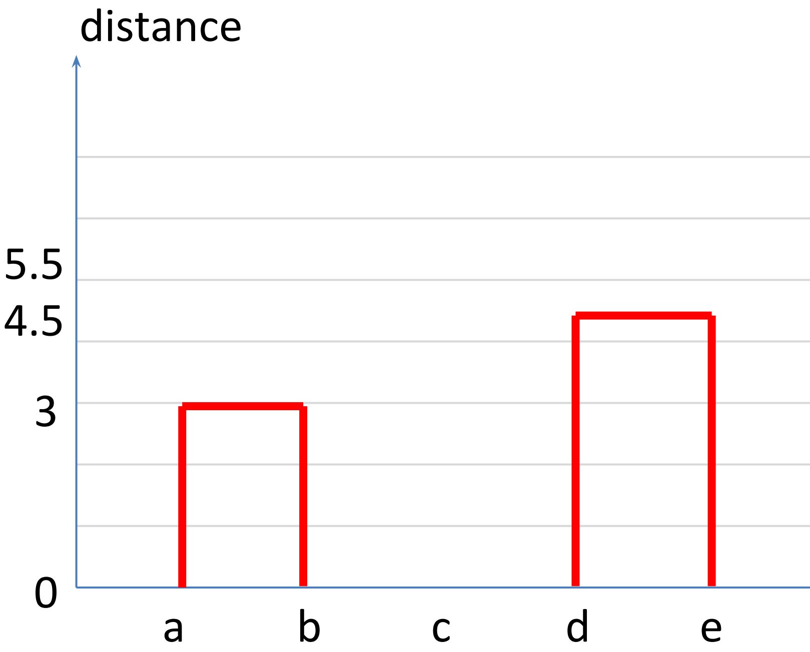
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations

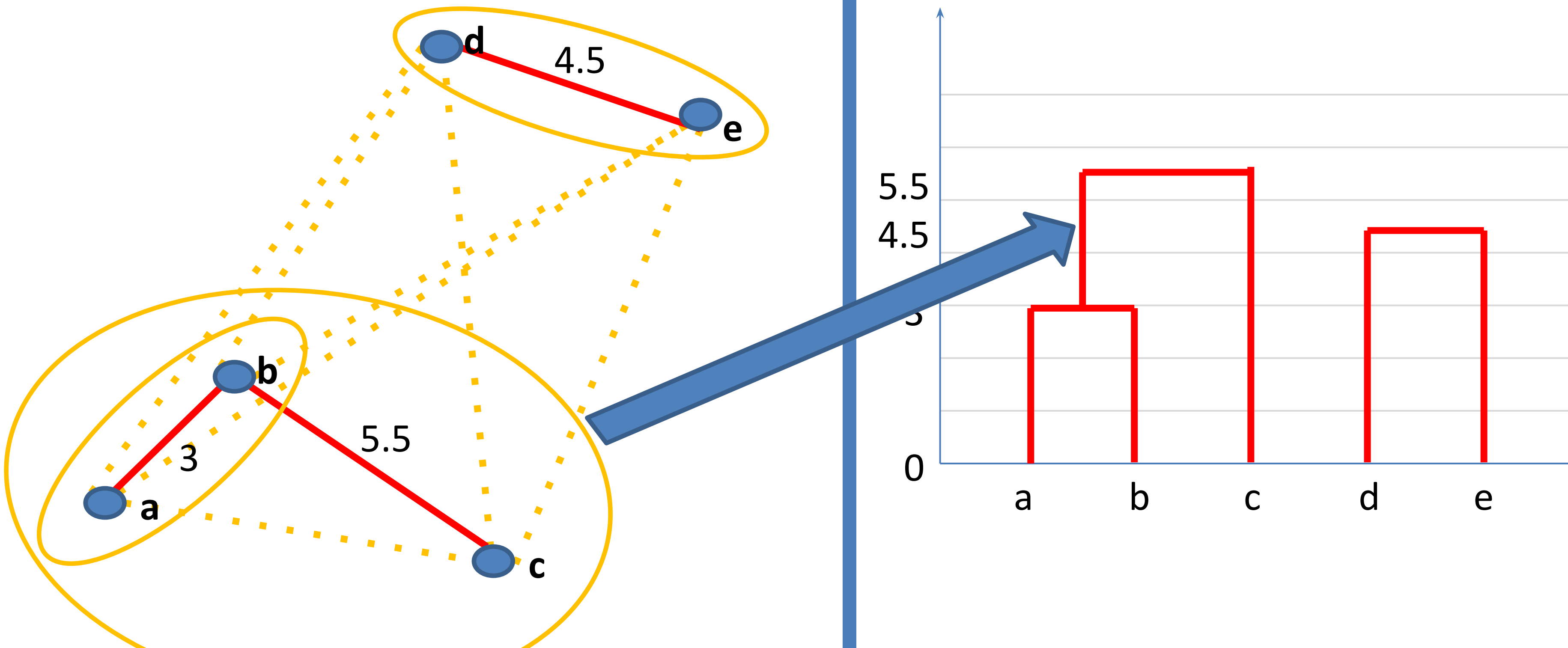


Observations



Dendrogram

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



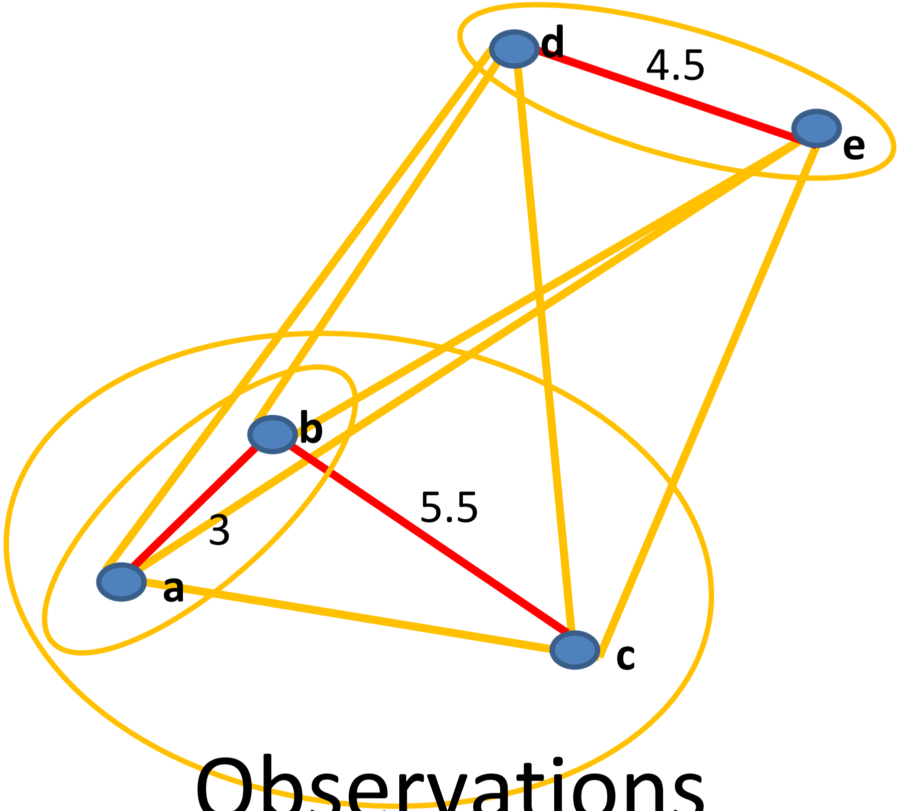
Observations

Dendrogram

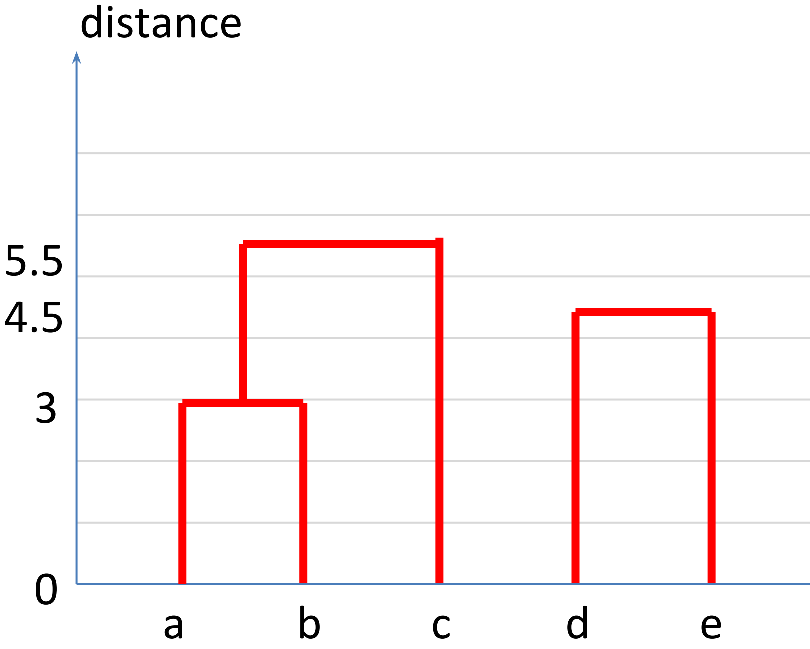
1 - Computing distances between observations

2 – Identification / choose a minimum

3 – Fusion of observations

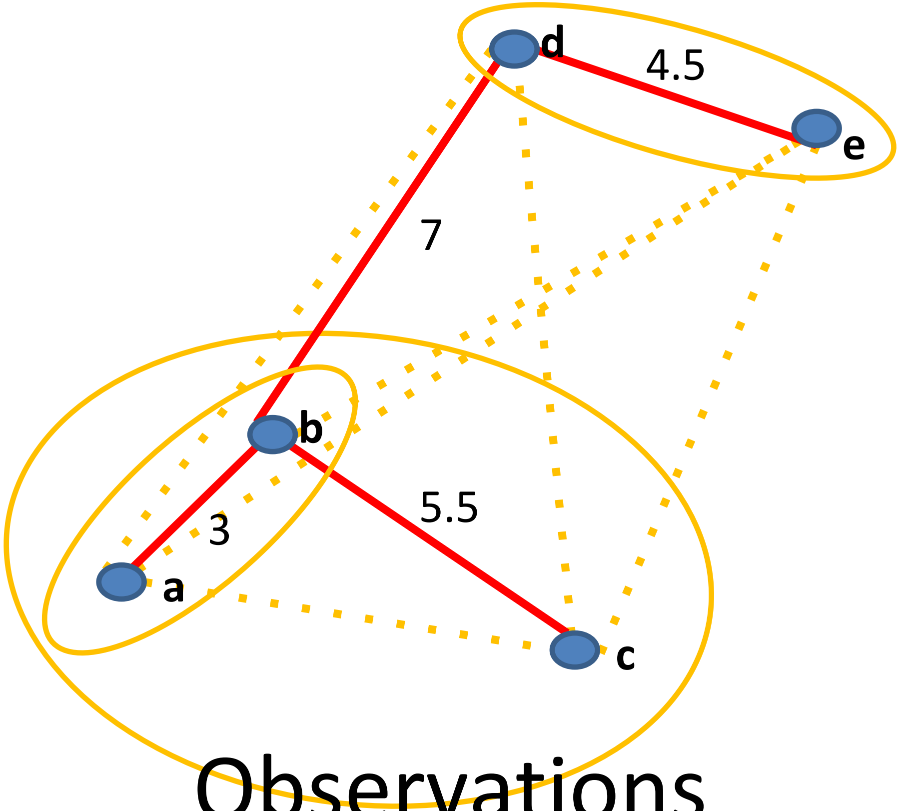


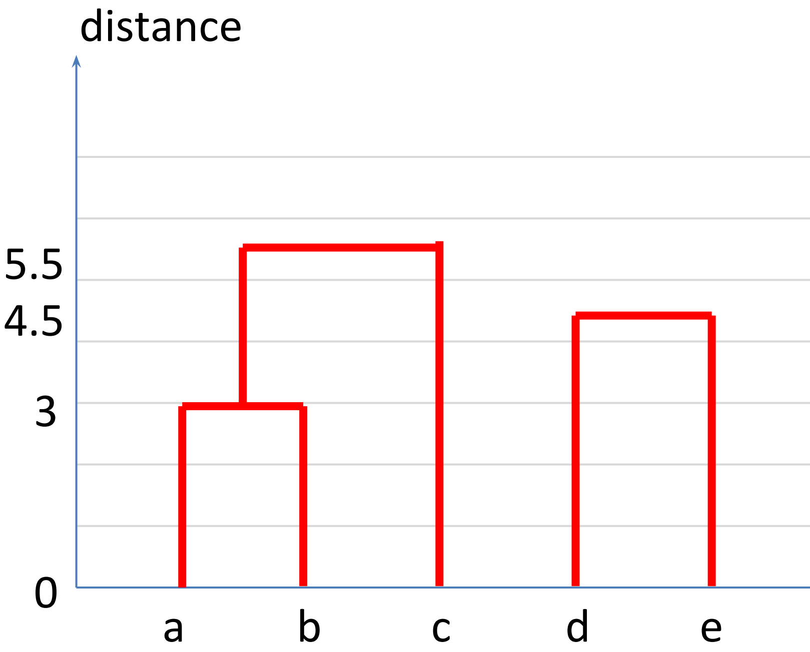
Observations



Dendrogram

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

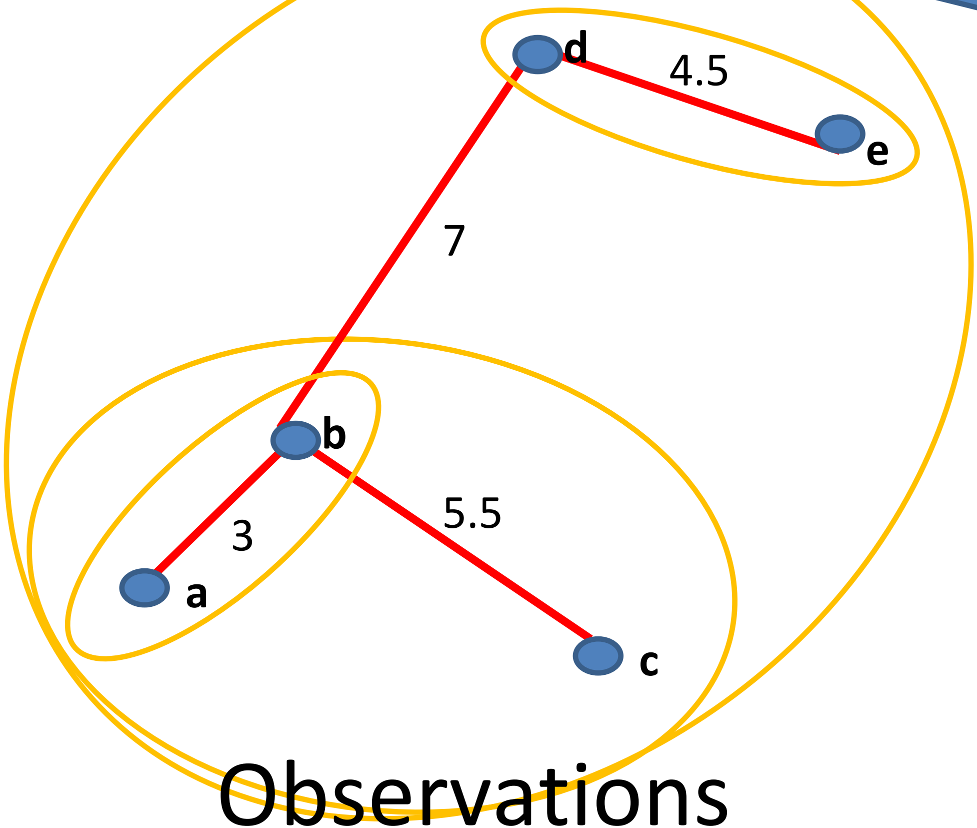




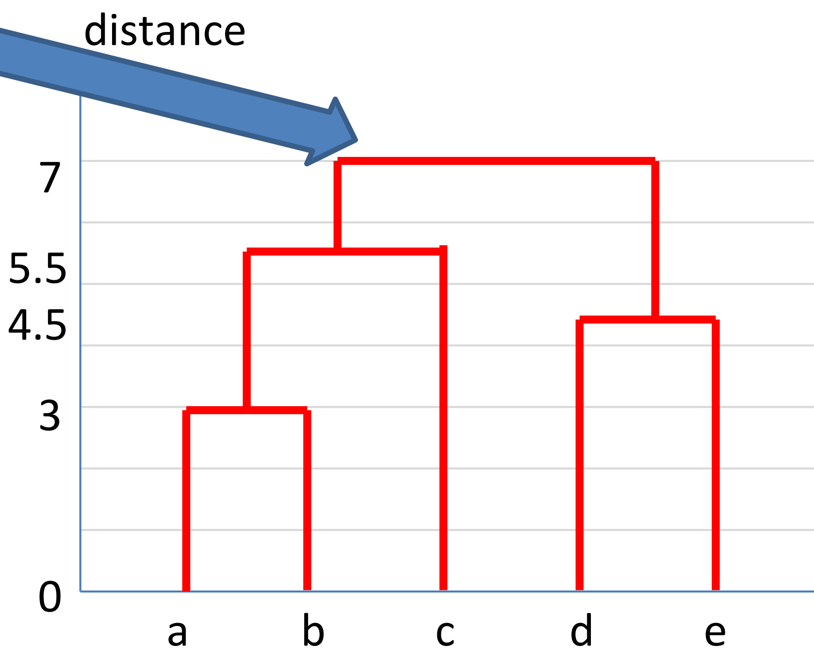
Observations

Dendrogram

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

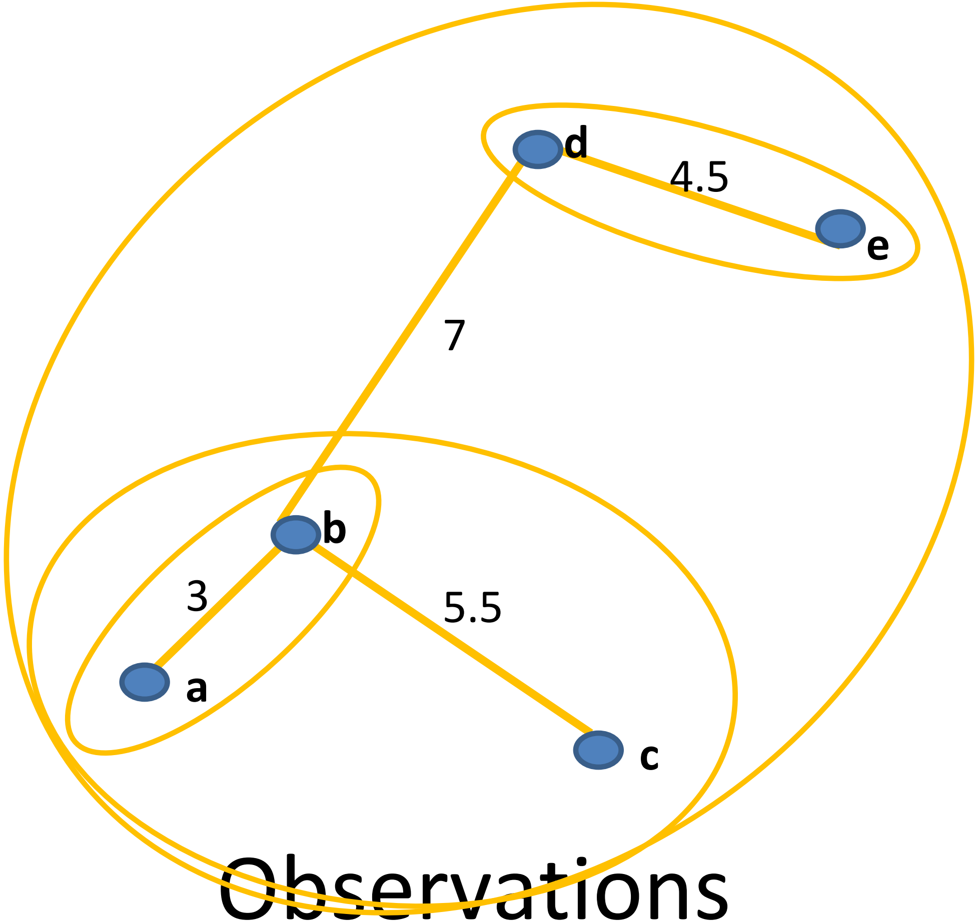


Observations

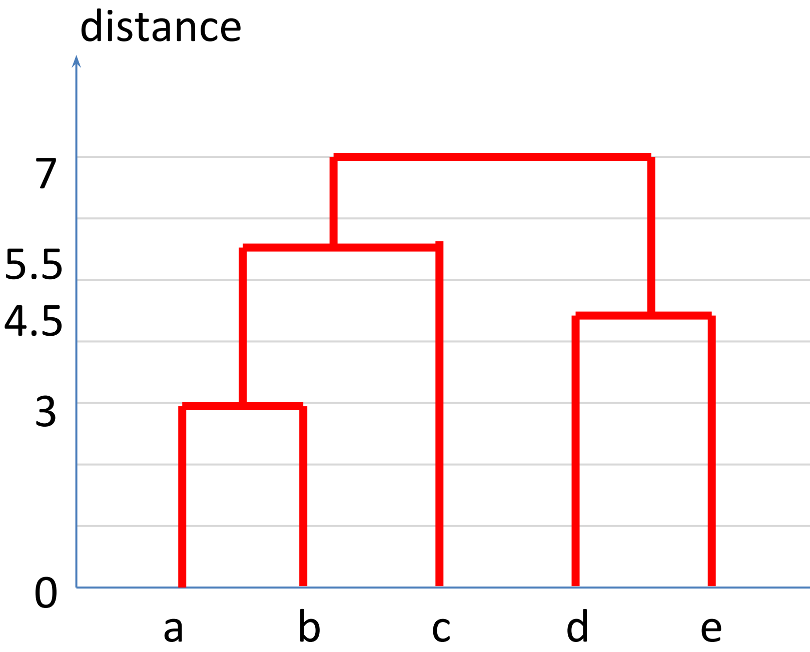


STOP !

Dendrogram



Observations



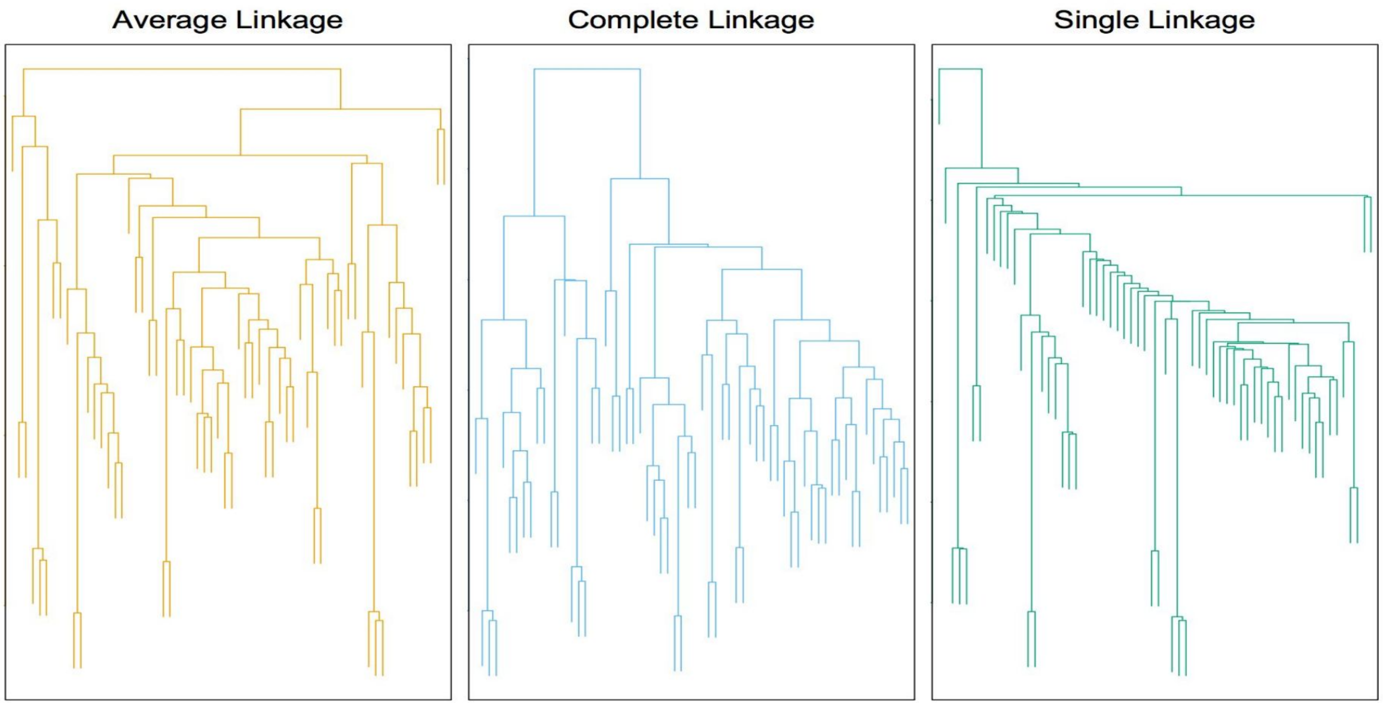
Dendrogram

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
- TF IDF vectors
- Sets (Bags / Transactions)
- Time series
- Strings
- Images
- ...

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

- Tversky Index

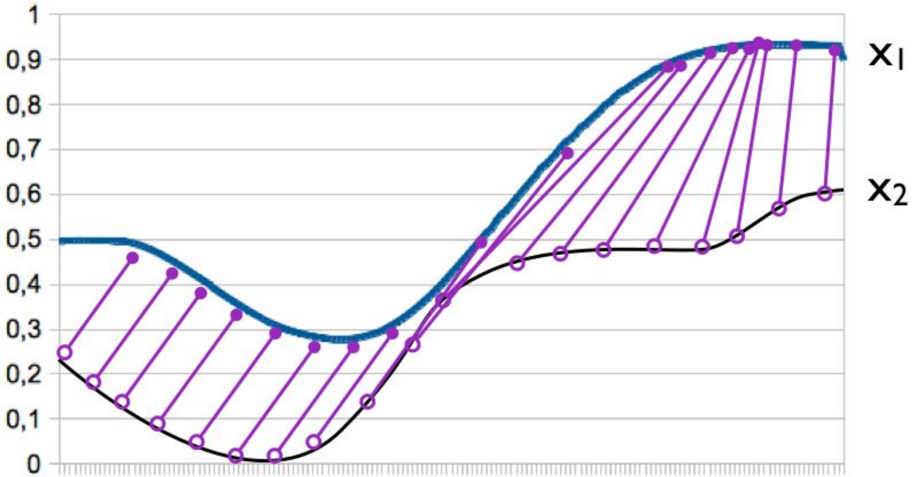
$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|}$$

- Jaccard Measure

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

Similarity between… time series

- Dynamic Time Warp

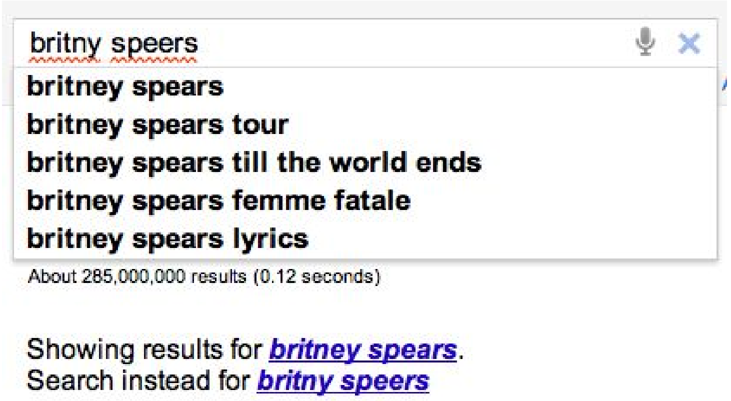


[source]

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

Similarity between… images

Create image signatures / feature vectors: color / texture / shape features

Pair Assignment