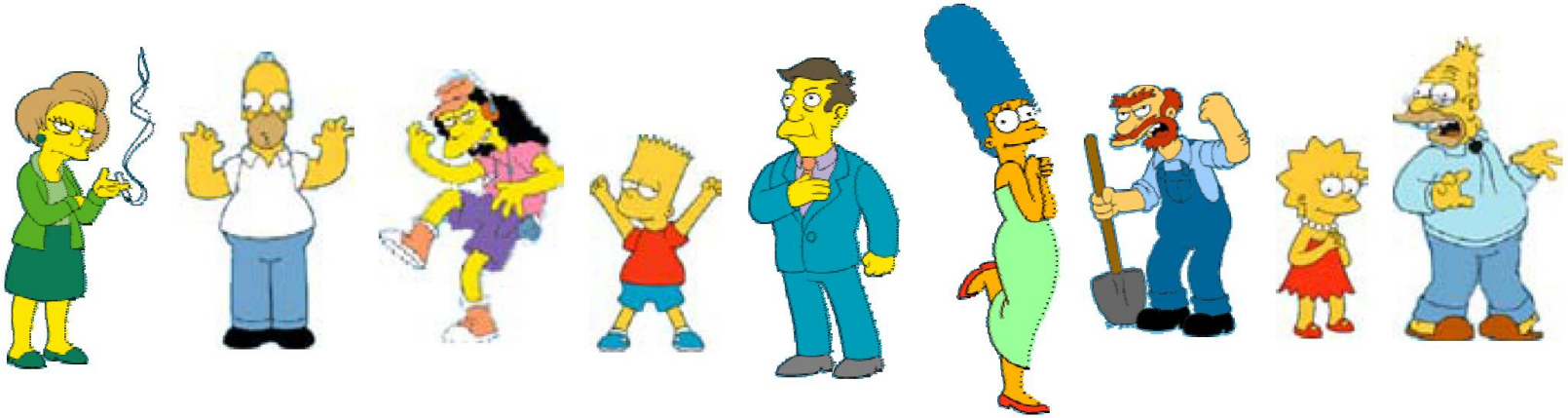
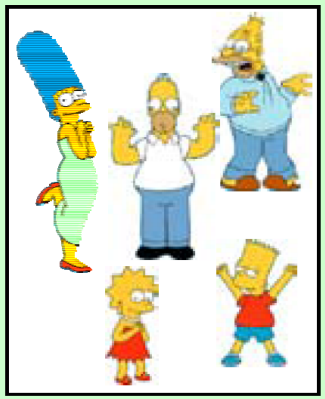
DATA, MSML, BIOI 602
Principles of Data Science

Heng Huang, Ph.D.
Department of Computer Science

What is a natural grouping among these objects?



Clustering is subjective



Simpson's Family

School Employees

Females

Males

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

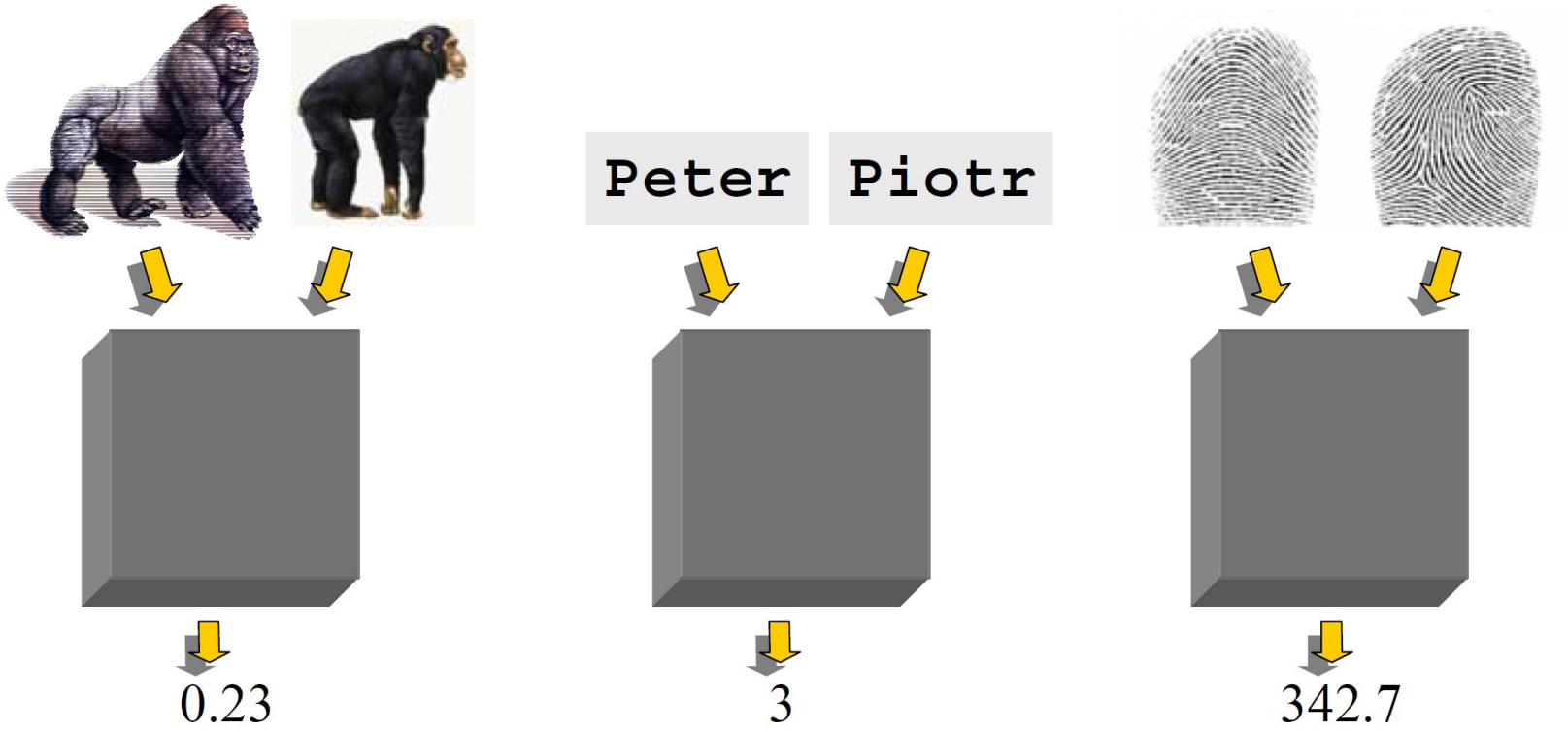
What is Similarity?

- The quality or state of being similar

  likeness, resemblance - Webster's Dictionary

Defining Distance Measures

Definition Let $O1$ and $O2$ be two objects from the universe of possible objects. The distance dissimilarity between $O1$ and $O2$ is a real number denoted by $D(O1,O2)$



What properties should a distance measure have?

- $D(A,B) = D(B,A)$ *Symmetry*

  Otherwise you could claim: Alex looks like Bob, but Bob looks nothing like Alex.
- $D(A,A) = 0$ *Constancy of Self-Similarity*

  Otherwise you could claim: Alex looks more like Bob, than Bob does.
- $D(A,B) = 0$ If A= B *Positivity Separation*

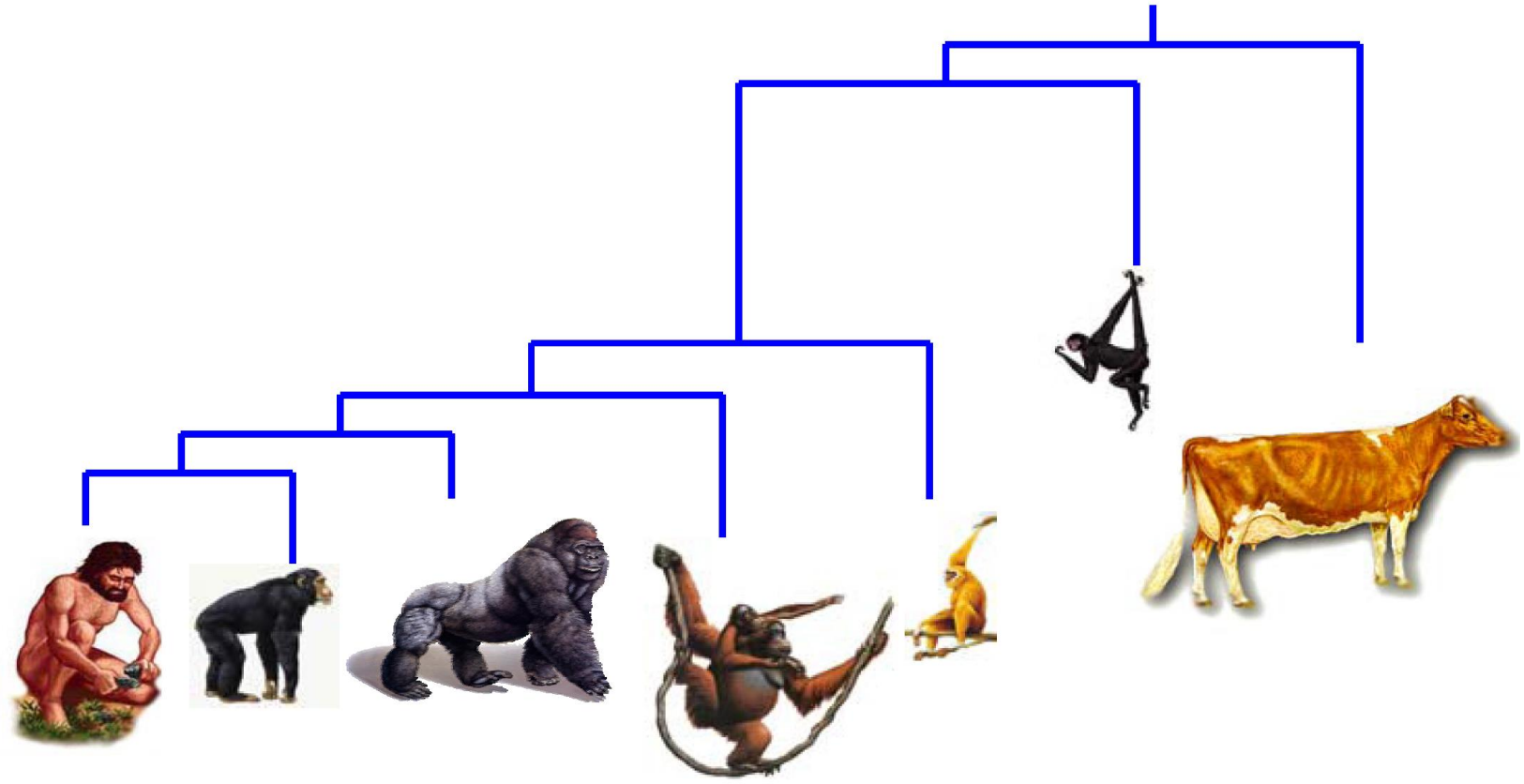
  Otherwise there are objects in your world that are different, but you cannot tell apart.
- $D(A,B) <= D(A,C) + D(B,C)$ *Triangular Inequality*

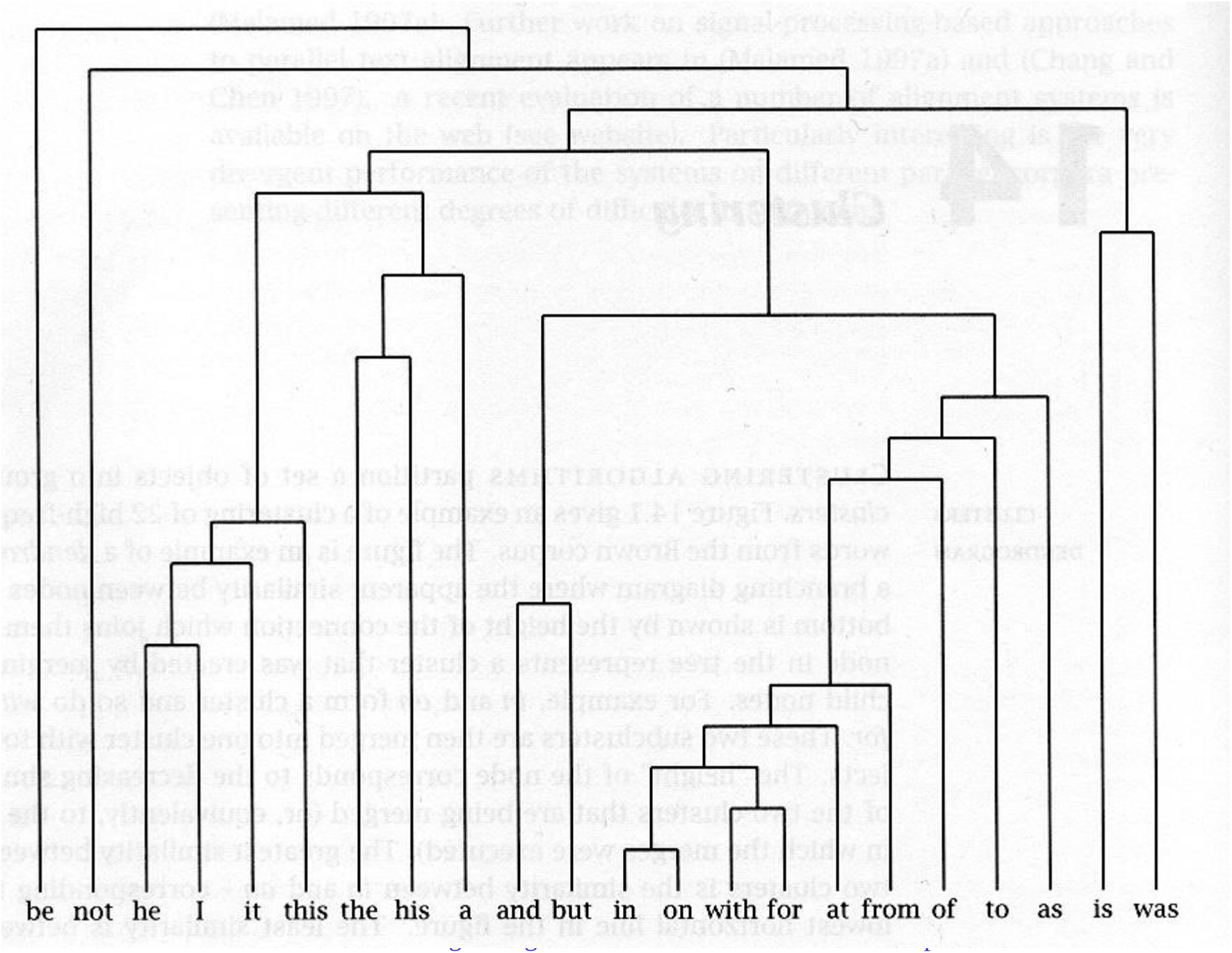
  Otherwise you could claim: Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl

Types of Clustering

- Hierarchical
  - Bottom Up:
    - Start with objects and group most similar ones.
  - Top down:
    - Start with all objects and divide into groups so as to maximize within-group similarity.
  - Single-link, complete-link, group-average
- Non-hierarchical
  - K-means
  - EM-algorithm
- Graph based model
  - Spectral clustering

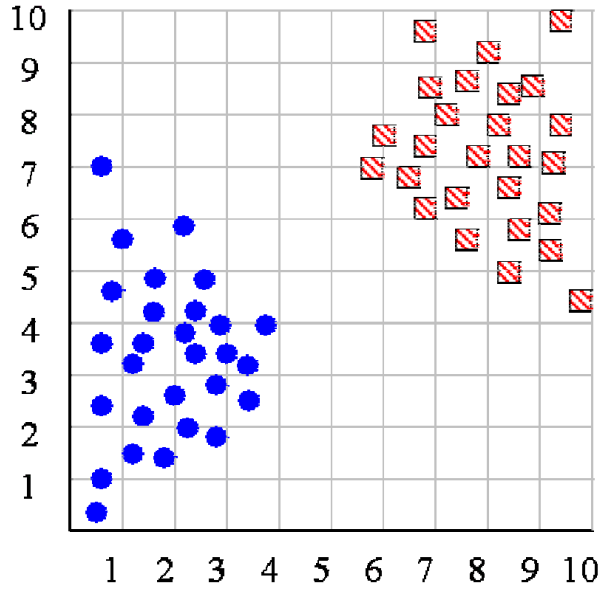
Perfectly clustered using a hierarchy




be not he I it this the his a and but in on with for at from of to as is was

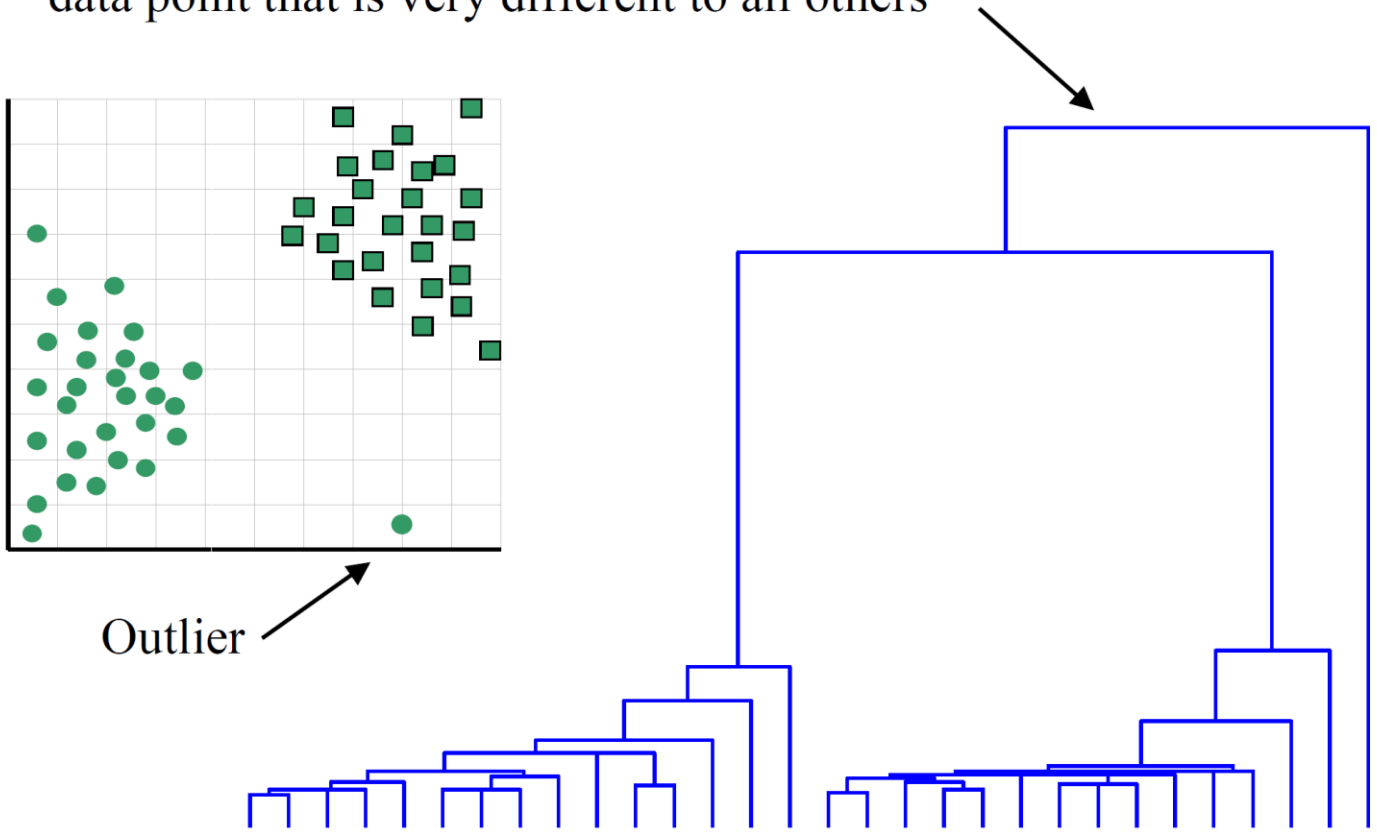
Cluster Number

We can look at the dendrogram to determine the correct number of clusters. In this case, the two highly separated subtrees are highly suggestive of two clusters.



Use of a dendrogram to detect outliers

The single isolated branch is suggestive of a data point that is very different to all others



Hierarchical Clustering

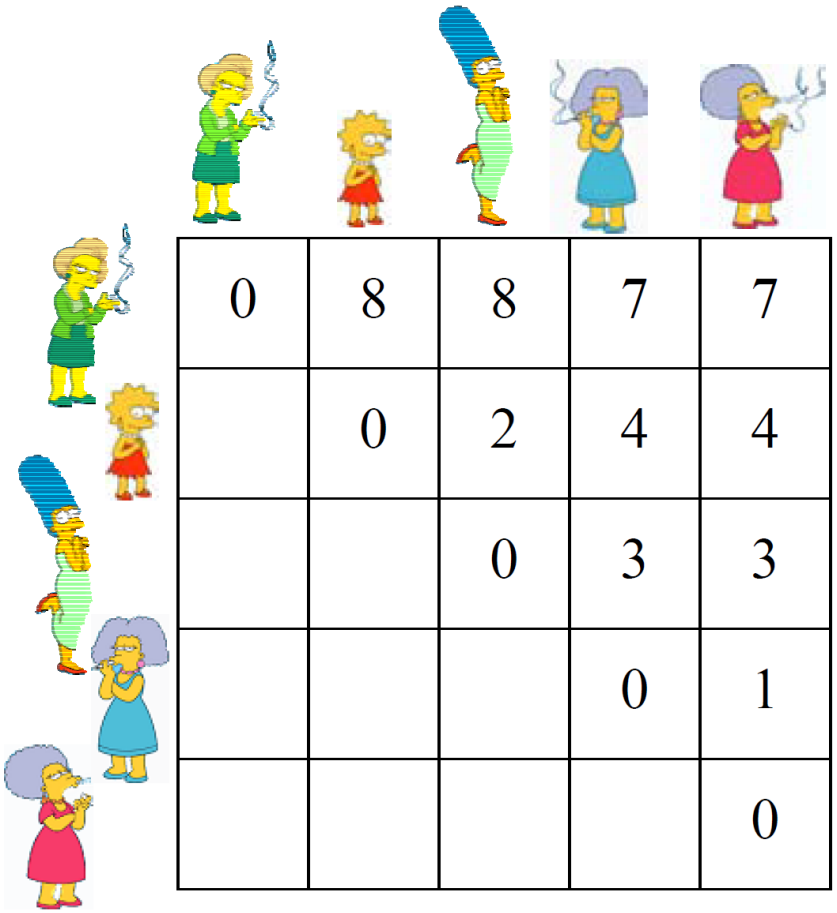
- Bottom-up:
  1. Start with a separate cluster for each object
  2. Determine the two most similar clusters and merge into a new cluster. Repeat on the new clusters that have been formed.
  3. Terminate when one large cluster containing all objects has been formed

Example of a similarity measure:

$$d_{ij} = \sum_{K=1}^{L} (x_{ik} - x_{jk})^2$$

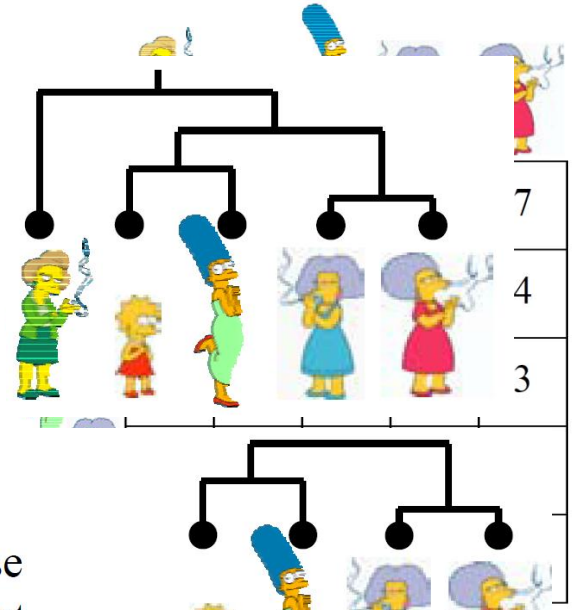
- Top-down
  1. Start from a cluster of all objects
  2. Iteratively determine the cluster that is least coherent and split it.
  3. Repeat until all clusters have one object.

Distance Matrix

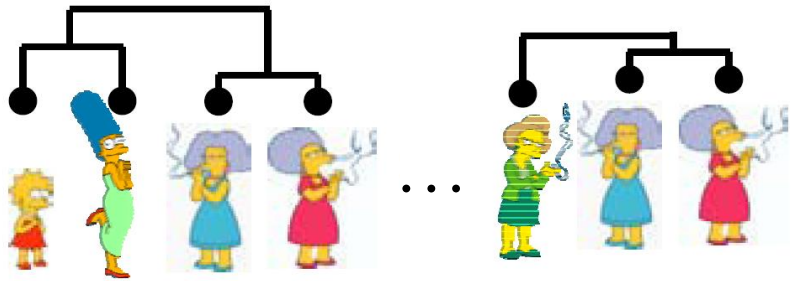


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

Bottom-Up Agglomerative

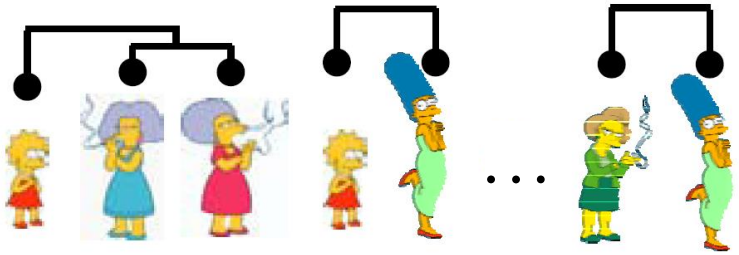


Consider all possible merges

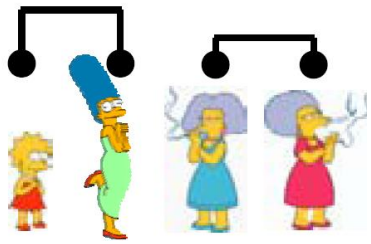


Choose the best

Consider all possible merges

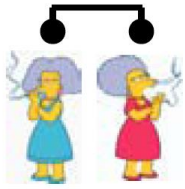


Choose the best



Consider all possible merges

Choose the best

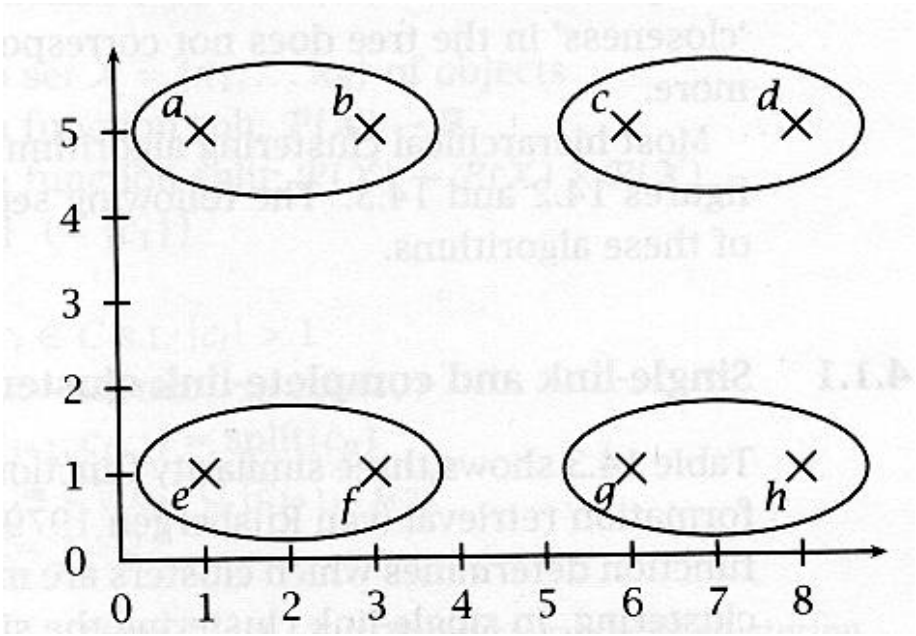


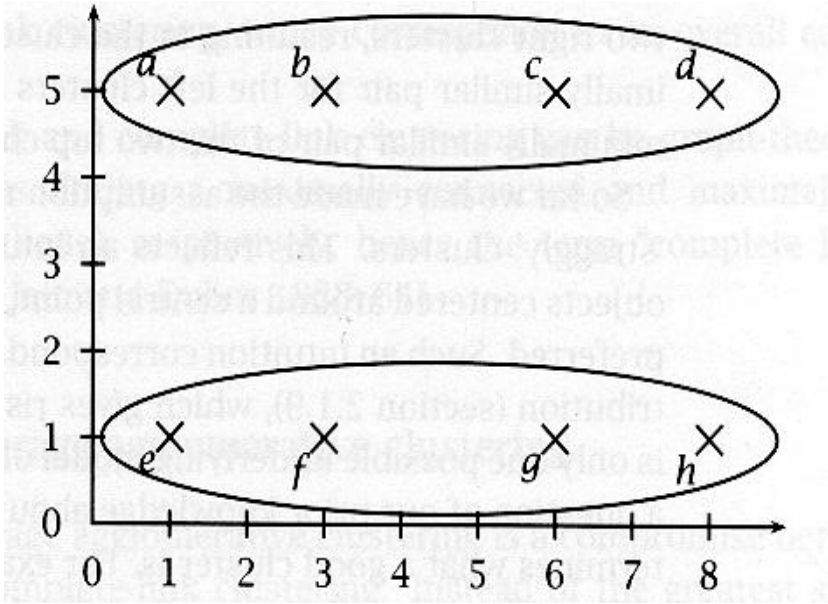
Similarity Measures for Hierarchical Clustering

- Single-link
  - Similarity of two most similar members
- Complete-link
  - Similarity of two least similar members
- Group-average
  - Average similarity between members

Single-Link

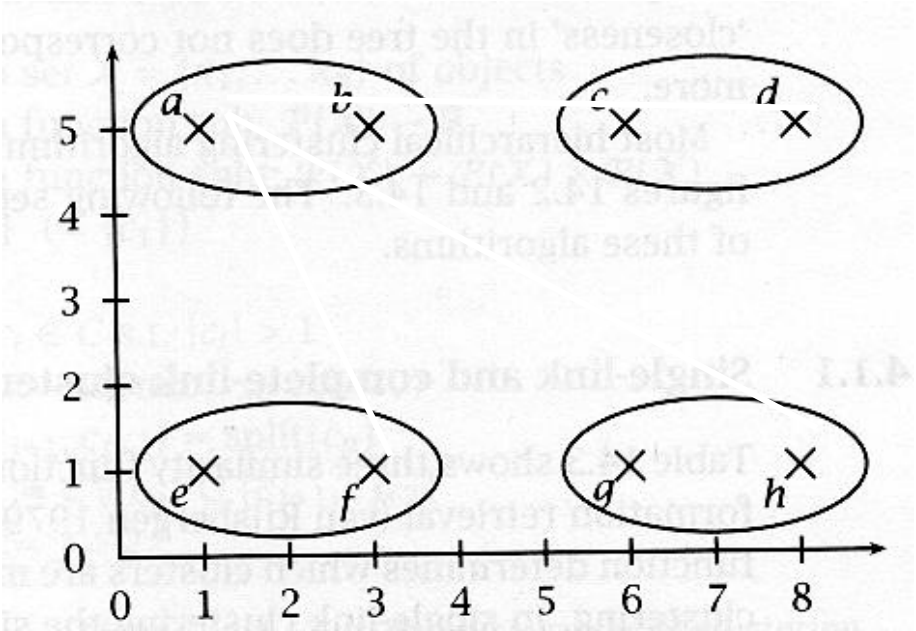
- Similarity function focuses on local coherence

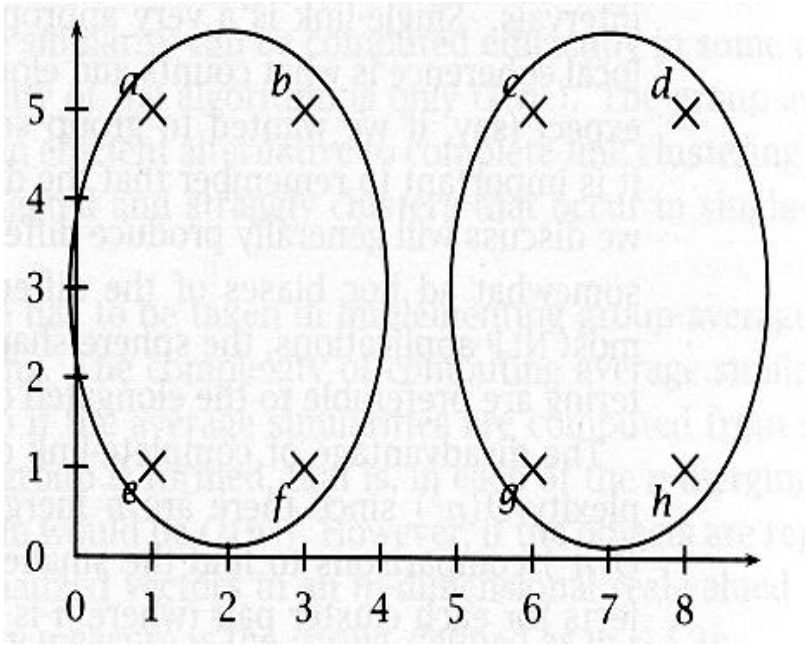




Complete-Link

- Similarity function focuses on global cluster quality





Group-Average

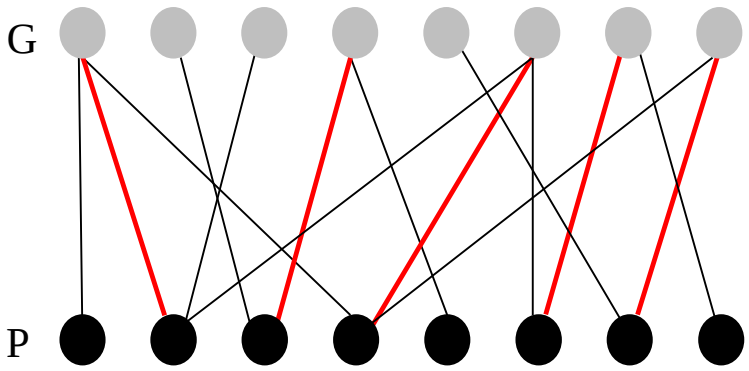
- Instead of greatest similarity between elements of clusters or the least similarity the merge criterion is average similarity.
- Compromise between single-link and complete-link clustering

Cluster Validation

Algorithm 21.4: Algorithm for matching partitions and clusters

MatchPartitionCluster **(**P,C,$match$**)**:

1 **foreach** $p \in P$ **do**
2 $match(p) \leftarrow \emptyset$
3 **foreach** $c \in C$ **do**
4 $overlap(p, c) \leftarrow \frac{|p \cap c|}{|p|}$
5 **while** $overlap \neq \emptyset$ **do**
6 $(p_{max}, c_{max}) \leftarrow GetMaxOverlap(overlap)$
7 $match(p_{max}) \leftarrow c_{max}$
8 $overlap \leftarrow overlap - \{overlap(p_{max}, *), overlap(*, c_{max})\}$

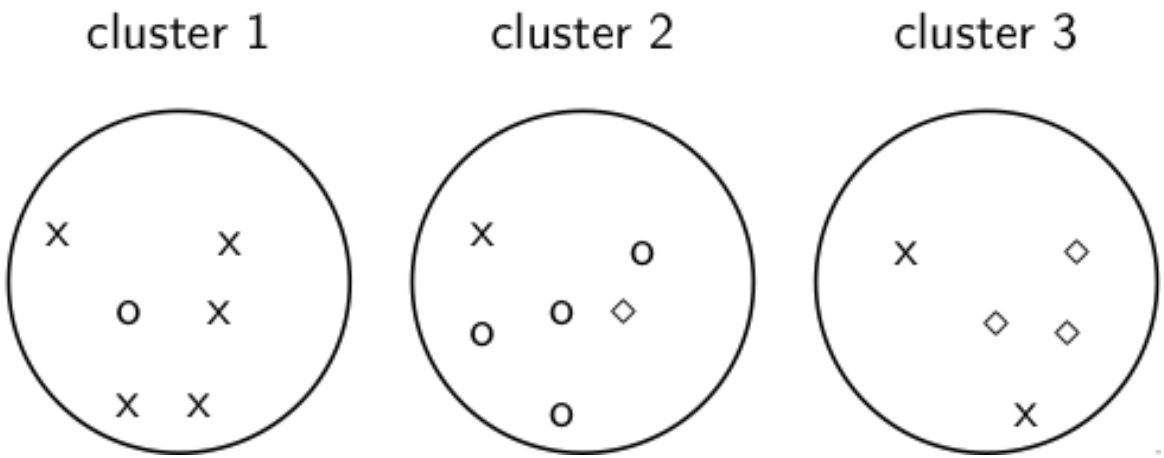


External Criterion: Purity

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $C = \{c_1, c_2, \ldots, c_J\}$ is the set of classes.
- For each cluster ω_k : find class c_j with most members n_{kj} in ω_k
- Sum all n_{kj} and divide by total number of points

External Criterion: Purity



To compute purity: $5 = \max_j |\omega_1 \cap c_j|$ (class x, cluster 1); $4 = \max_j |\omega_2 \cap c_j|$ (class o, cluster 2); and $3 = \max_j |\omega_3 \cap c_j|$ (class ◇, cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

Rand Index

- Definition: $\mathrm{RI} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$
- Based on 2x2 contingency table of all pairs of documents:

	same cluster	different clusters
same class	true positives (TP)	false negatives (FN)
different classes	false positives (FP)	true negatives (TN)

- TP+FN+FP+TN is the total number of pairs.
- There are $\binom{N}{2}$ pairs for N documents.
- Example: $\binom{17}{2}$ = 136 in o/◊/x example
- Each pair is either positive or negative (the clustering puts the two documents in the same or in different clusters) . . .
- . . . and either "true" (correct) or "false" (incorrect): the clustering decision is correct or incorrect.

Rand Index Example

- As an example, we compute RI for the o/◊/x example. We first compute TP + FP. The three clusters contain 6, 6, and 5 points, respectively, so the total number of "positives" or pairs of documents that are in the same cluster is:

$$\mathrm{TP} + \mathrm{FP} = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

- Of these, the x pairs in cluster 1, the o pairs in cluster 2, the ◊ pairs in cluster 3, and the x pair in cluster 3 are true positives:

$$\mathrm{TP} = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

- Thus, FP = 40 − 20 = 20. FN and TN are computed similarly.

Rand measure for the o/⋄/x example

	same cluster	different clusters
same class	TP = 20	FN = 24
different classes	FP = 20	TN = 72

RI is then

- $(20 + 72)/(20 + 20 + 24 + 72) \approx 0.68$.