Machine Learning

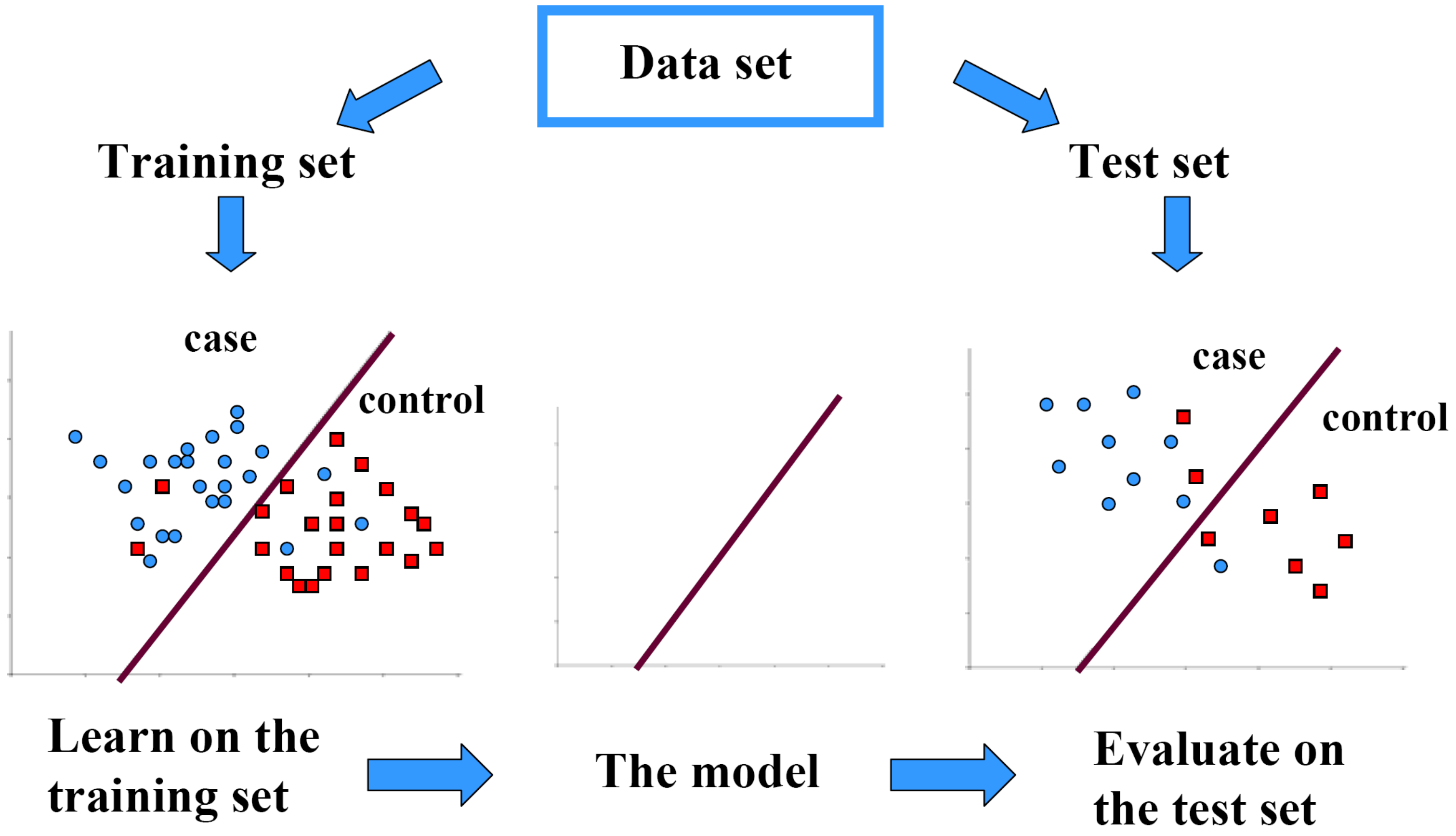
CSE 6363 (Fall 2016)

Lecture 15 Validation, KNN, Clustering

Heng Huang, Ph.D.

Department of Computer Science and Engineering

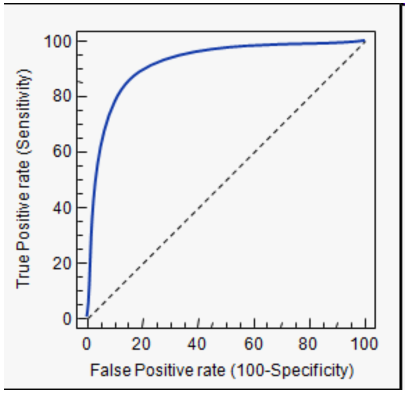
Evaluation for Classification



Evaluation Metrics

Confusion matrix: Records the percentages of examples in the testing set that fall into each group

		Actual	
		Case	**Control**
Prediction	**Case**	**TP** 0.3	**FP** 0.1
	Control	**FN** 0.2	**TN** 0.4



Misclassification error:

$$E = FP + FN$$

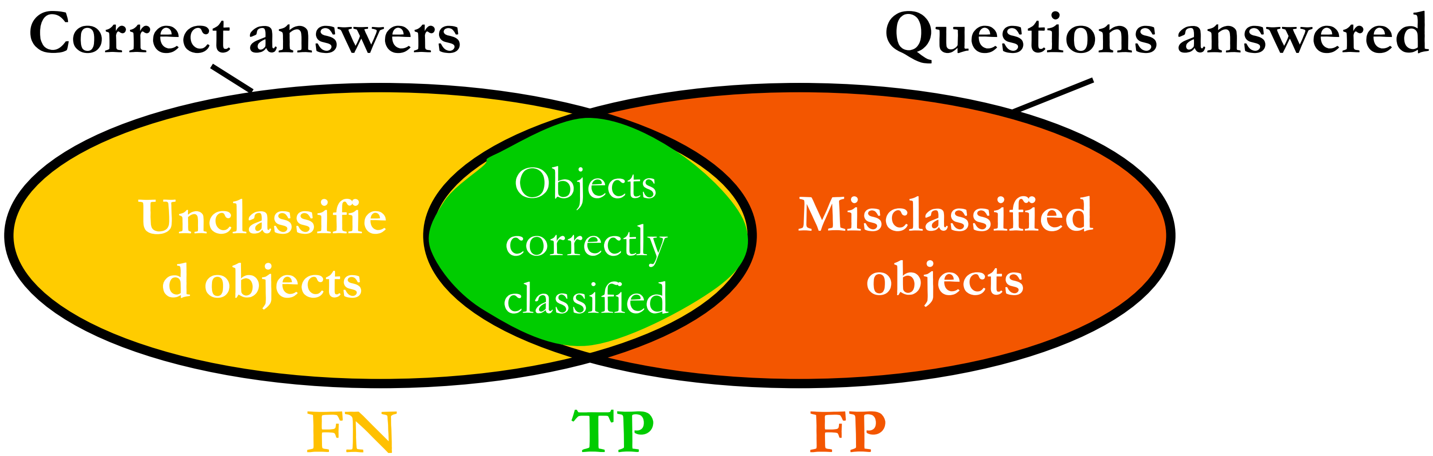
Sensitivity:

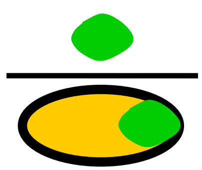
$$SN = \frac{TP}{TP + FN}$$

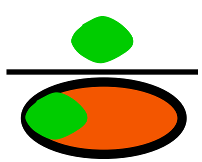
Specificity:

$$SP = \frac{TN}{TN + FP}$$

Precision-Recall



Recall=  = fraction of all objects correctly classified

Precision=  = fraction of all questions correctly answered

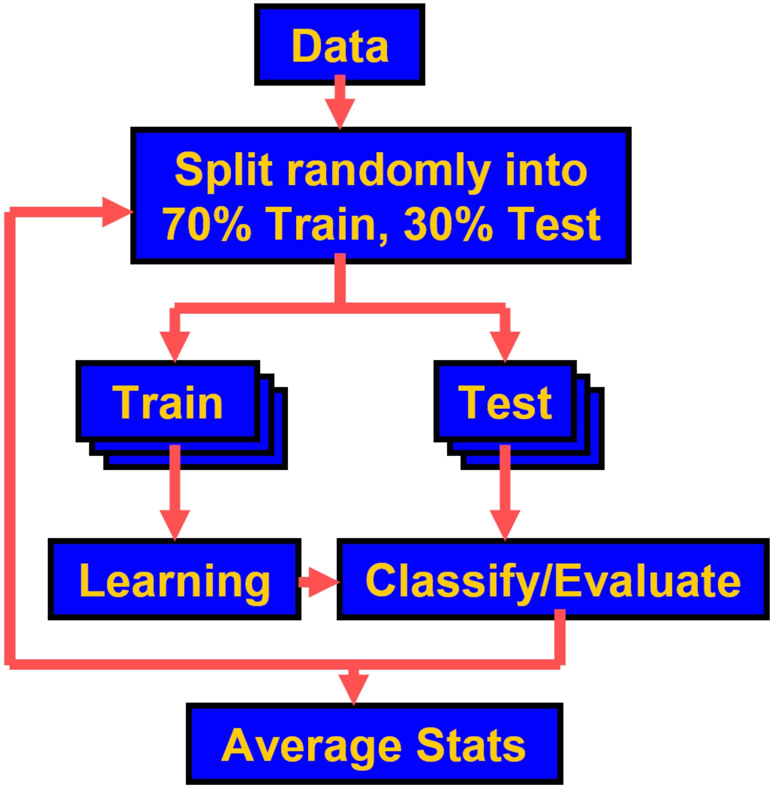
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Evaluation

- Problem: if the sample size is relatively small one split may be lucky or unlucky hence biasing the statistics
- Solution: use multiple train/test splits and average their results
- Random resampling validation techniques:
  - random sub-sampling
  - k-fold cross-validation
  - bootstrap-based validation

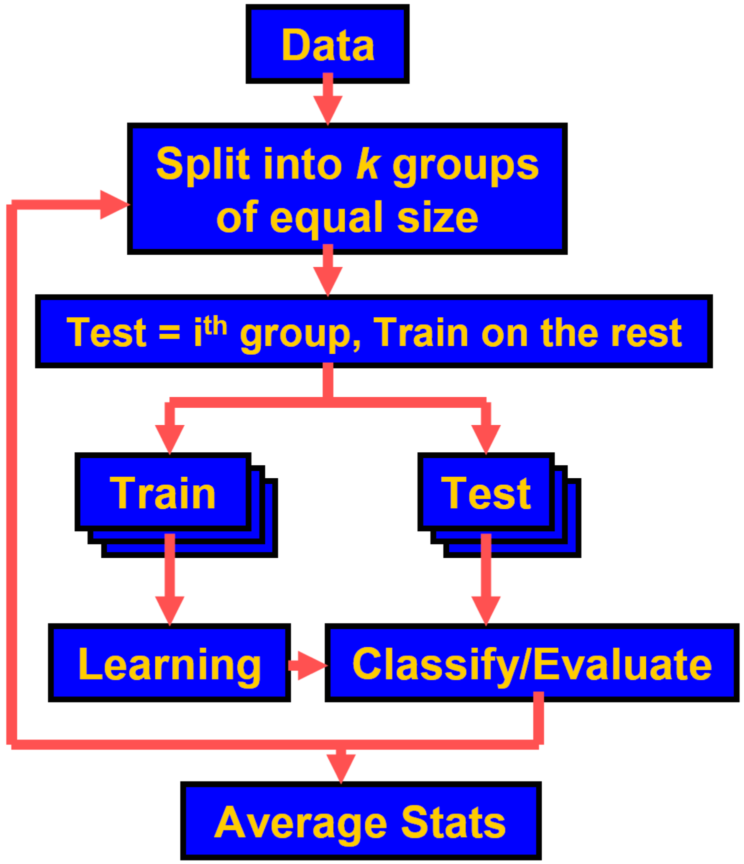
Random Sub-sampling

- Split the data into train and test set with some split ratio (typically 70:30)
- Repeat this k times for different random splits
- Average the results of statistics



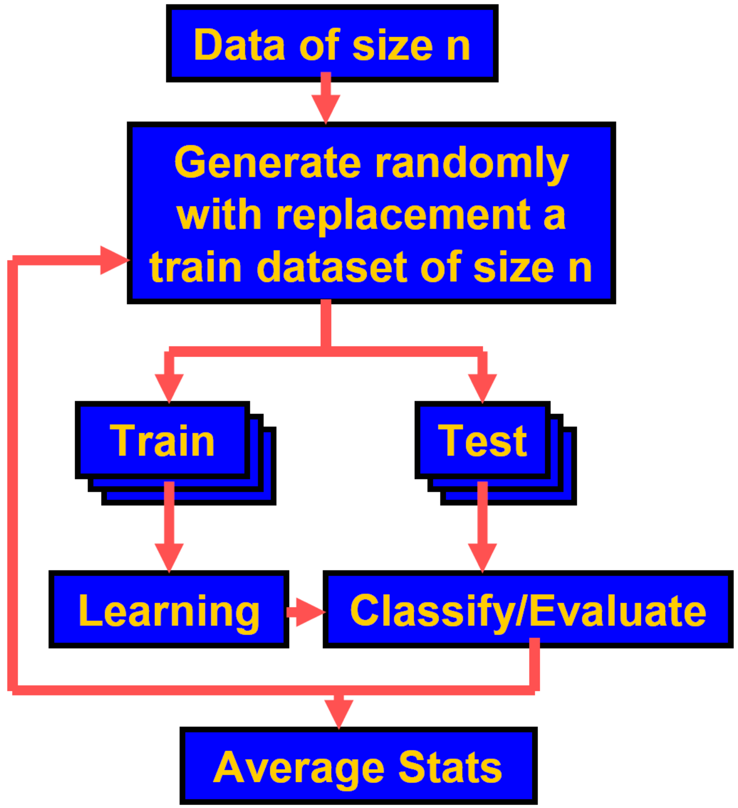
K-fold Cross-validation

- Split the data into k equal size groups
- Use each group once as a test set, and the remaining groups as the training set
- Repeat this k times for k groups
- Average the results of statistics



Bootstrap-based Validation

- Bootstrap technique - used primarily to estimate the sampling distribution of an estimator
- Generate randomly with replacement a training dataset of size n that equals the original data size
- Some examples are repeated in the training set, some are missing
- Build a test set from examples not used in the training set.



Parametric Models

A **parametric** model implements a very restricted family of functions **f(x; w)**, leaving only a few parameters **w** to be learned. It thus expresses a strong presupposition (= prior) about the structure of the data.

Example: parametric density estimation

Assume density is isotropic Gaussian: $f(\mathbf{x}_k; \mathbf{w}) = N(\mathbf{x}_k; \boldsymbol{\mu}, \sigma^2\mathbf{I})$

=> need only determine optimal mean $\boldsymbol{\mu}*$ and variance σ^2*

ML quickly gives

$$\boldsymbol{\mu}* = \frac{1}{n}\sum_k \mathbf{x}_k\,, \qquad \sigma^2* = \frac{1}{n}\sum_k \|\mathbf{x}_k - \boldsymbol{\mu}\|^2.$$

Non-Parametric Models

Non-parametric models make only weak, general prior assumptions about the data, such as smoothness. **f(x; w)** is constructed directly over the memorized training data **X**; the construction involves no or few parameters **w** to be learned.

Example: *k*-nearest neighbor methods

The model's output **f(x; w)** for some new datum **x** is calculated by combining (in some fixed way) the memorized responses for the *k* nearest neighbors of **x** in the training data. Example:

(regression) interpolate between nearest neighbor responses

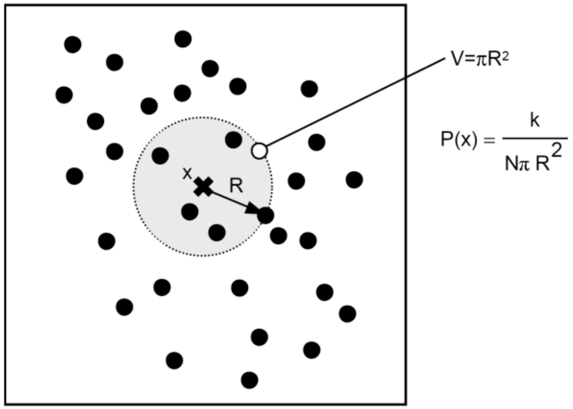
(classification) take majority vote of nearest neighbor classes

K Nearest Neighbor Classifier

- **The kNN classifier is based on non-parametric density estimation techniques**
  - Let us assume we seek to estimate the density function P(x) from a dataset of examples
  - P(x) can be approximated by the expression

$$P(x) \cong \frac{k}{NV} \quad \text{where} \quad \begin{cases} V \text{ is the volume surrounding } x \\ N \text{ is the total number of examples} \\ k \text{ is the number of examples inside } V \end{cases}$$

  - The volume V is determined by the D-dim distance $R_k^D(x)$ between x and its k nearest neighbor



$$P(x) \cong \frac{k}{NV} = \frac{k}{N \cdot c_D \cdot R_k^D(x)}$$

    - Where c_D is the volume of the unit sphere in D dimensions

K Nearest Neighbor Classifier

- **We use the previous result to estimate the posterior probability**
  - The unconditional density is, again, estimated with

$$P(x|\omega_i) = \frac{k_i}{N_i V}$$

  - And the priors can be estimated by

$$P(\omega_i) = \frac{N_i}{N}$$

  - The posterior probability then becomes

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)} = \frac{\frac{k_i}{N_i V} \cdot \frac{N_i}{N}}{\frac{k}{NV}} = \frac{k_i}{k}$$

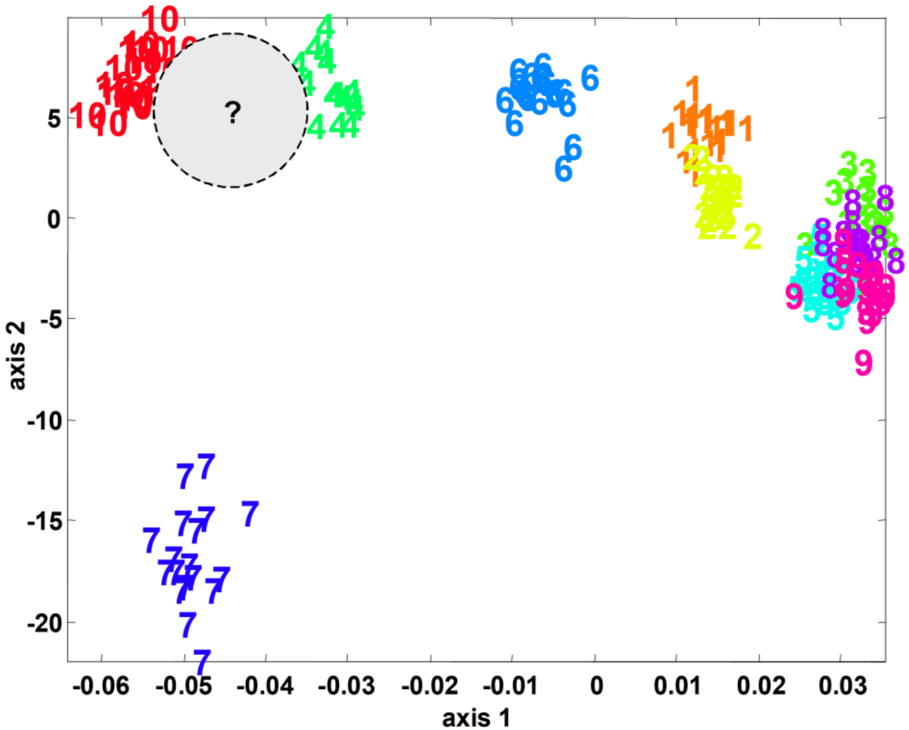
  - Yielding discriminant functions

$$\boxed{g_i(x) = \frac{k_i}{k}}$$

    - This is known as the k Nearest Neighbor classifier

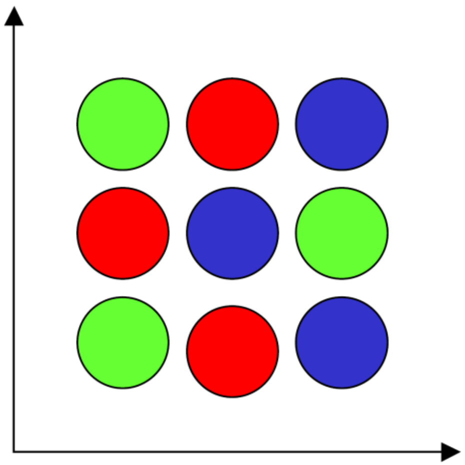
K Nearest Neighbor Classifier

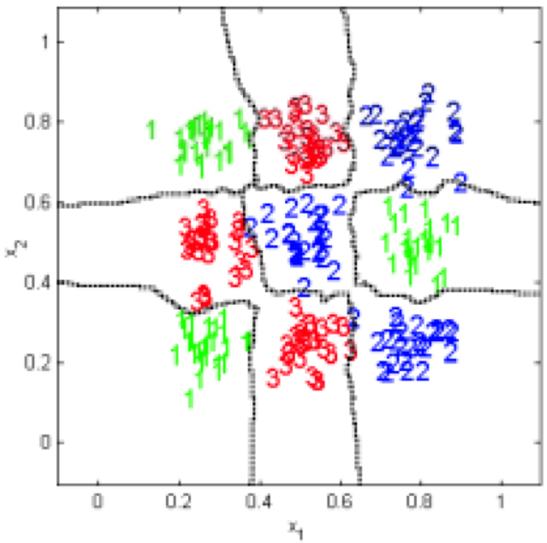
- **The kNN classifier is a very intuitive method**
  - Examples are classified based on their similarity with training data
    - For a given unlabeled example $x_u \in \Re^D$, find the k "closest" labeled examples in the training data set and assign x_u to the class that appears most frequently within the k-subset
- **The kNN only requires**
  - An integer k
  - A set of labeled examples
  - A measure of "closeness"

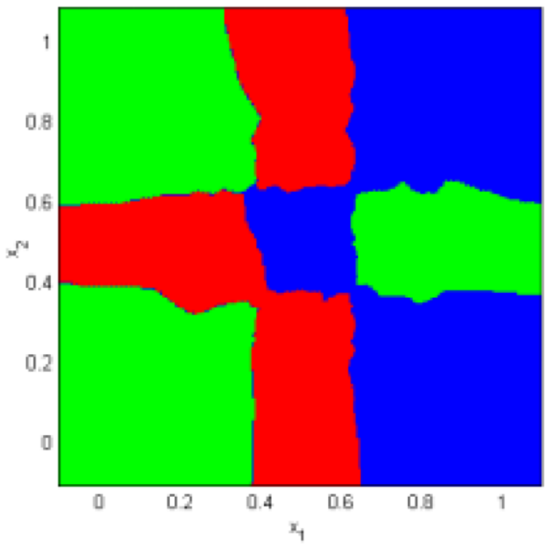


kNN in Action: Example 1

- **We generate data for a 2-dimensional 3-class problem, where the class-conditional densities are multi-modal, and non-linearly separable**
- **We used kNN with**
  - k = five
  - Metric = Euclidean distance

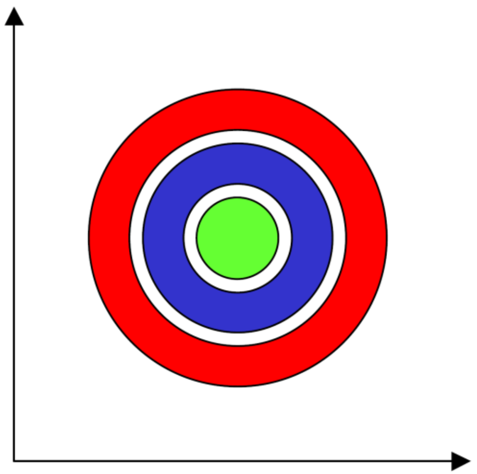


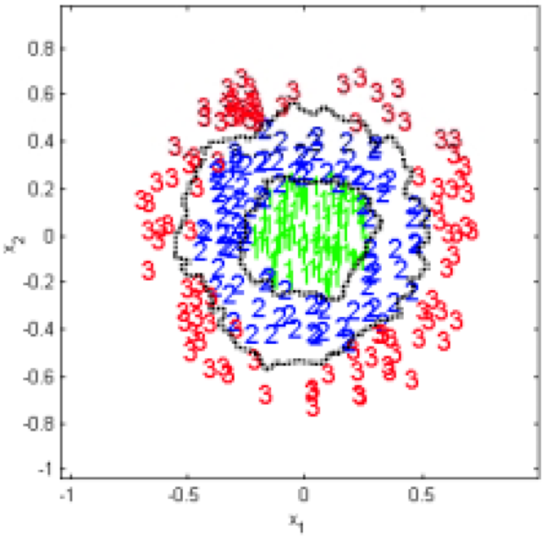


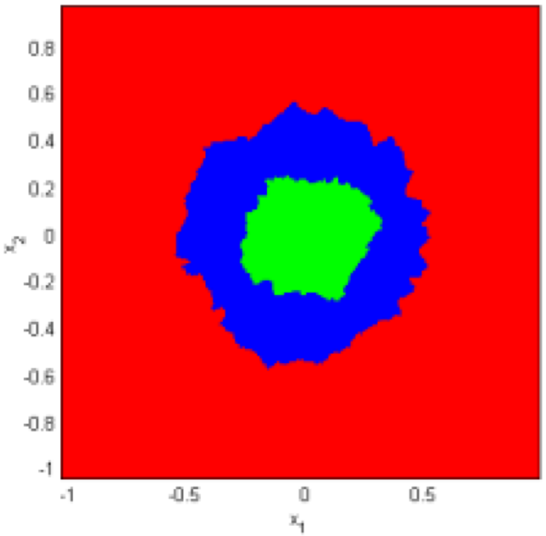


kNN in Action: Example 2

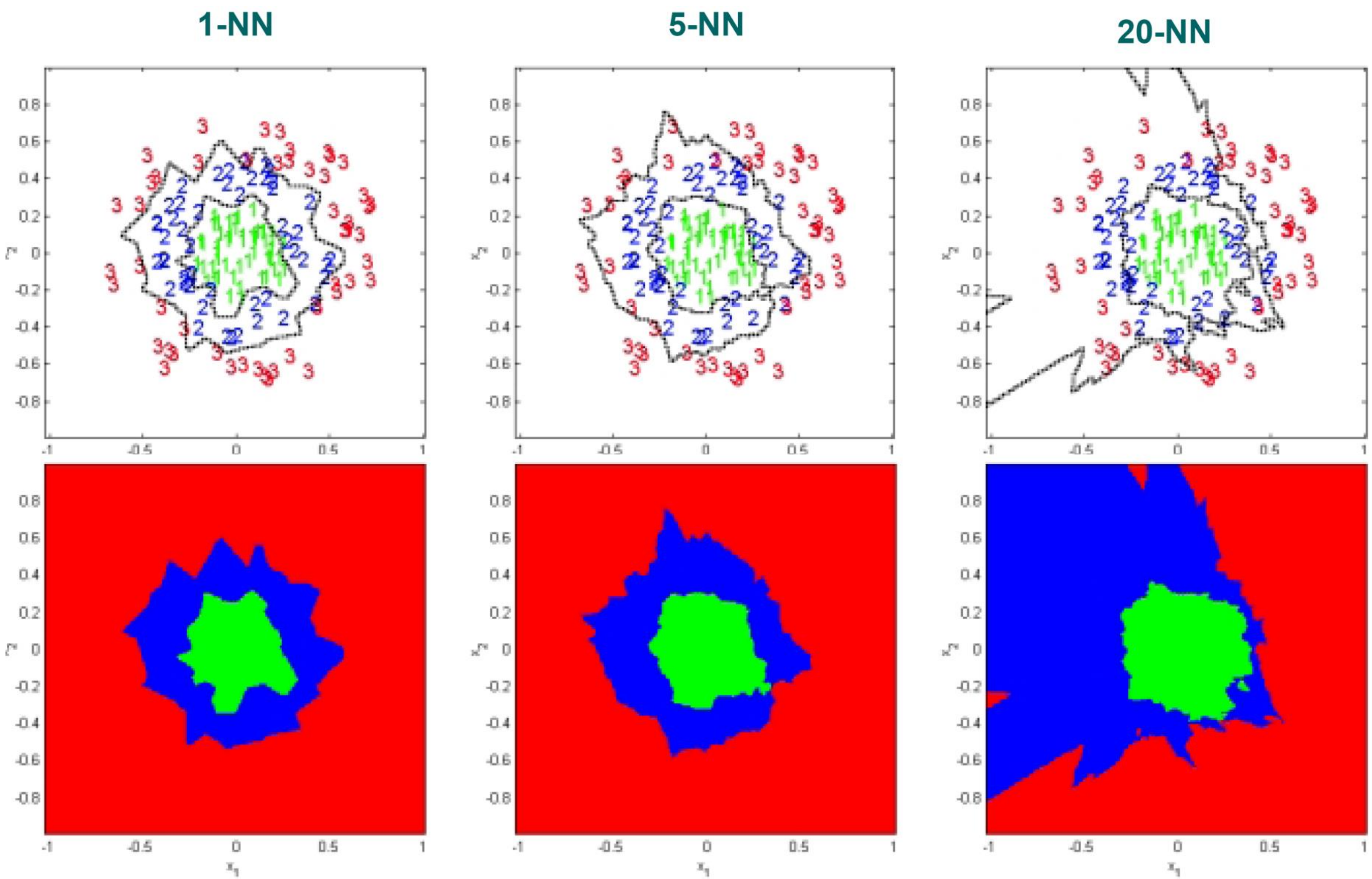
- **We generate data for a 2-dim 3-class problem, where the likelihoods are unimodal, and are distributed in rings around a common mean**
  - These classes are also non-linearly separable
- **We used kNN with**
  - k = five
  - Metric = Euclidean distance





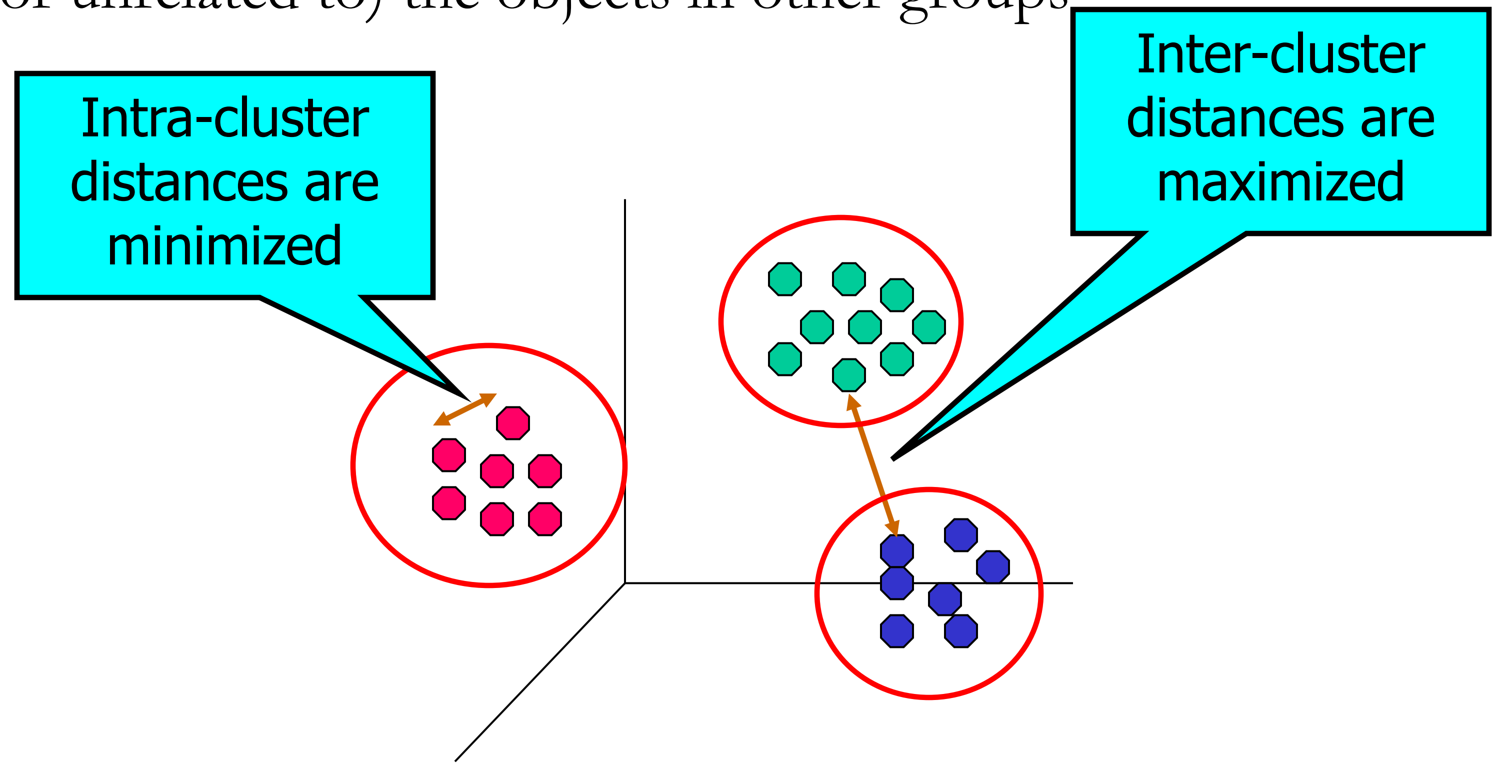


kNN versus 1NN

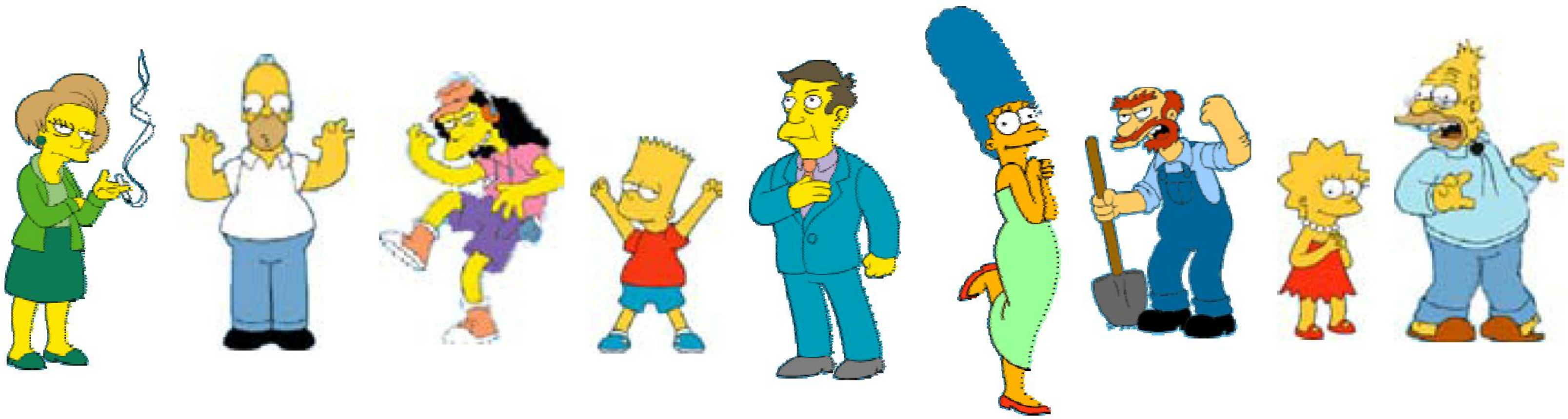


What is Cluster Analysis?

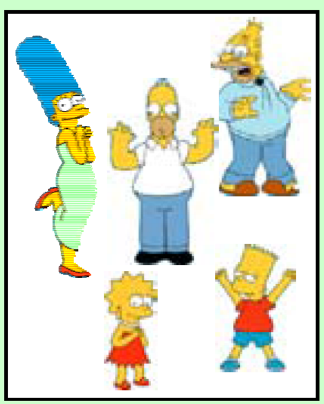
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



What is a natural grouping among these objects?

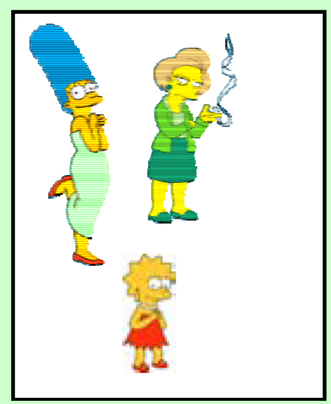


Clustering is subjective



Simpson's Family

School Employees



Females

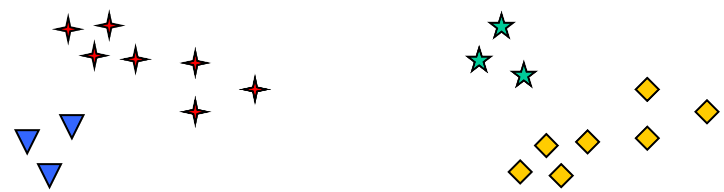
Males

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters



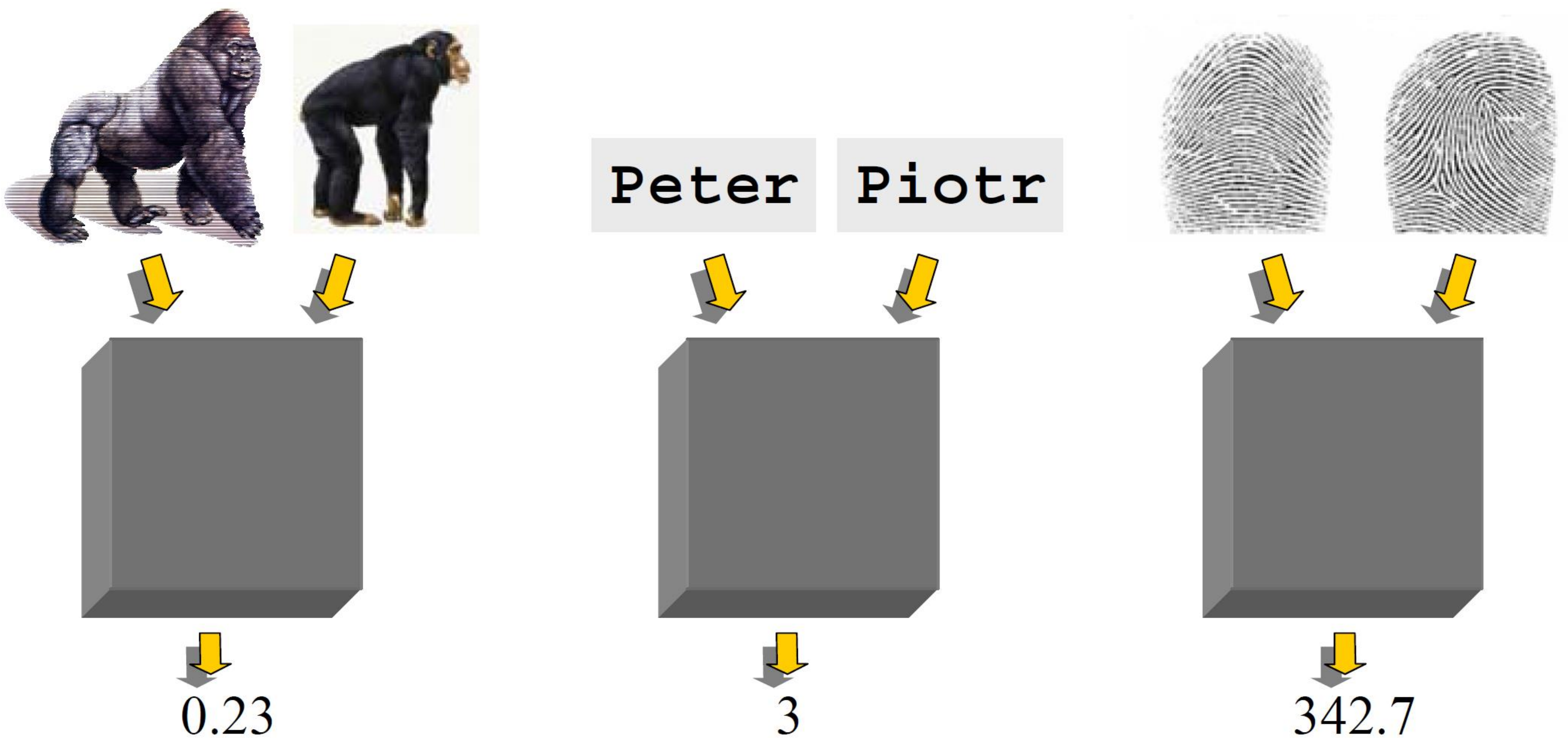
Four Clusters

What is Similarity?

- The quality or state of being similar
  likeness, resemblance - Webster's Dictionary

Defining Distance Measures

Definition Let *O*1 and *O*2 be two objects from the universe of possible objects. The distance dissimilarity between *O*1 and *O*2 is a real number denoted by *D*(*O*1,*O*2)



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*

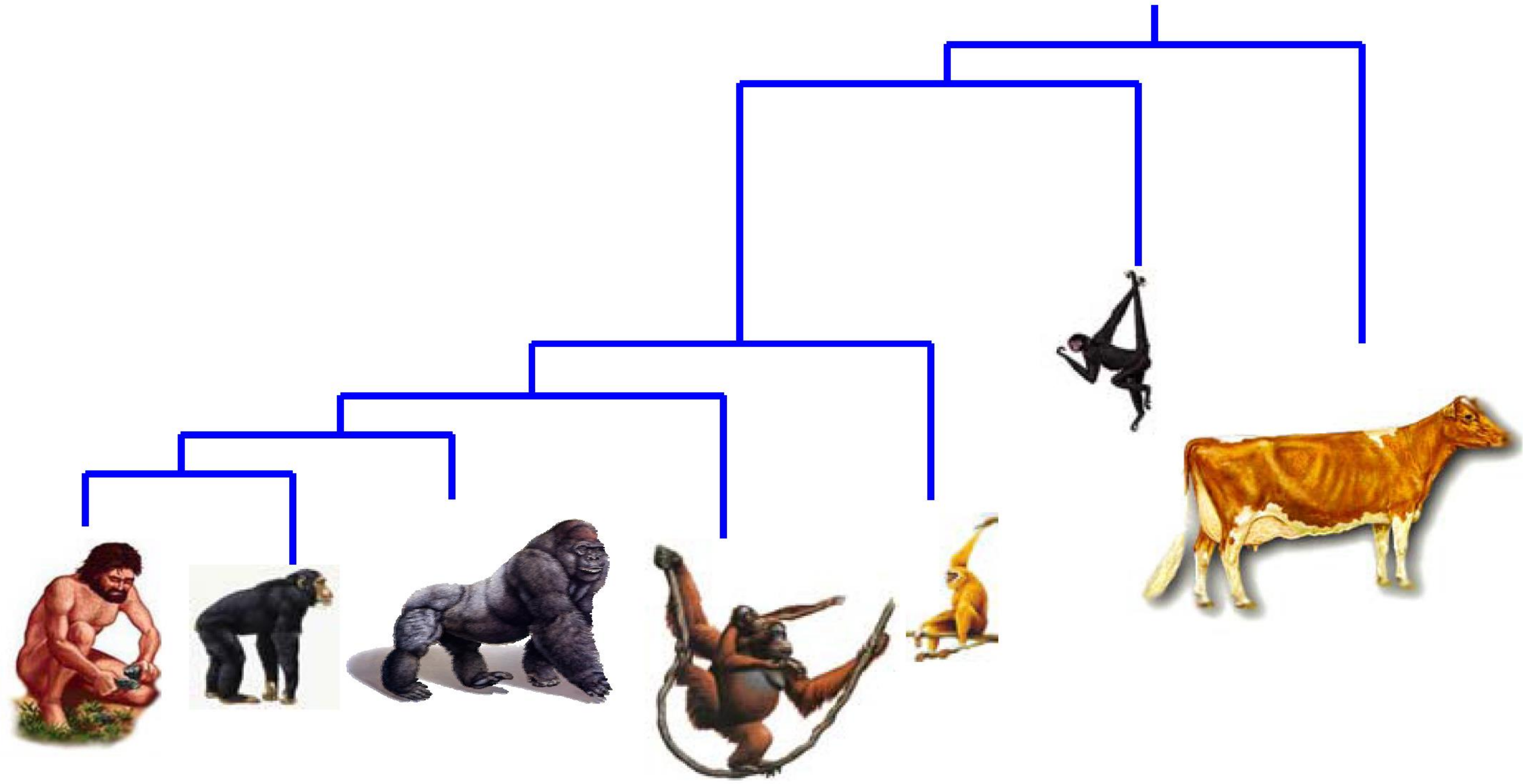
  Otherwise there are objects in your world that are different, but you cannot tell apart.
- *D*(A,B) <= *D*(A,C) + *D*(B,C) *Triangular Inequality*

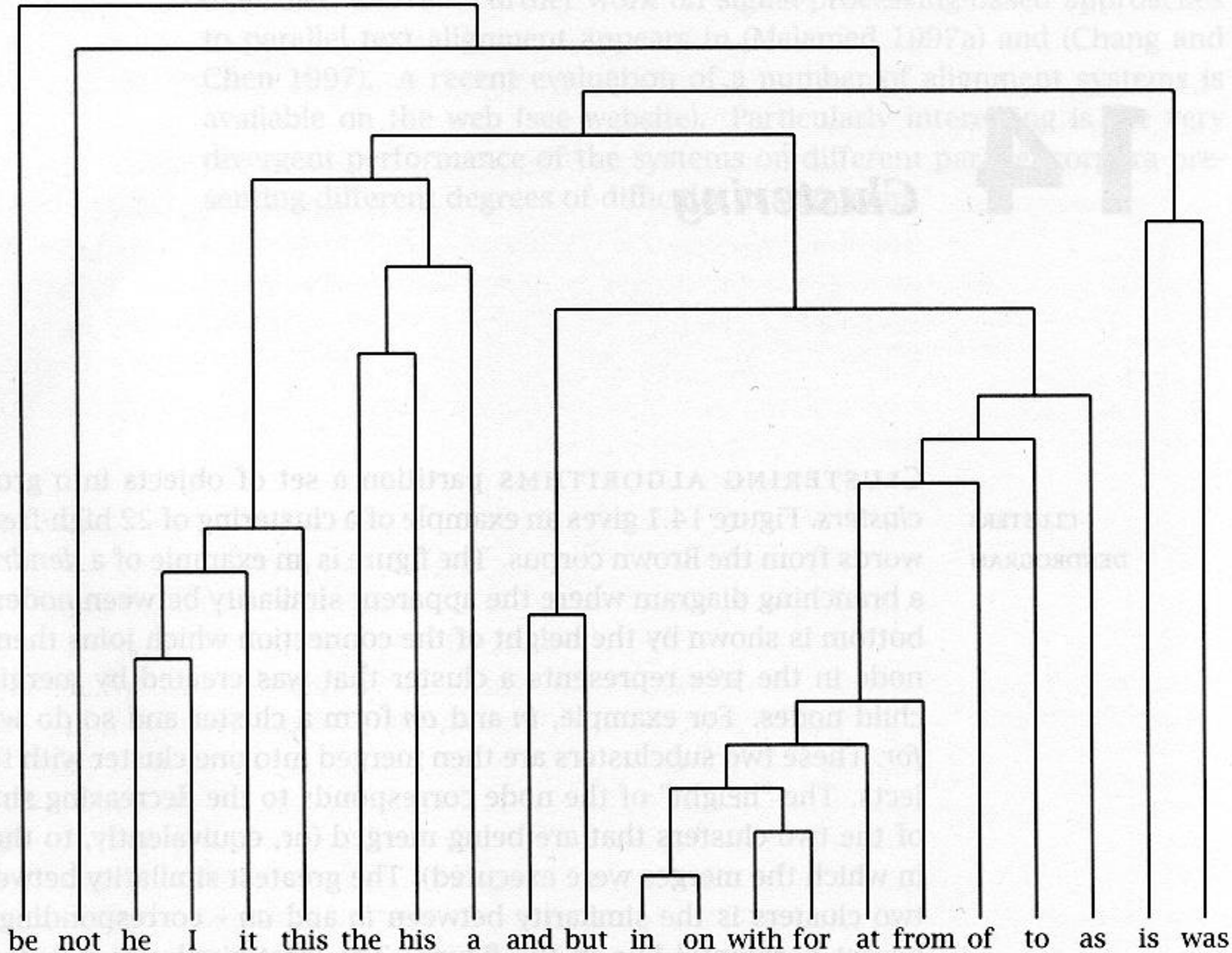
  Otherwise you could claim: Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl

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

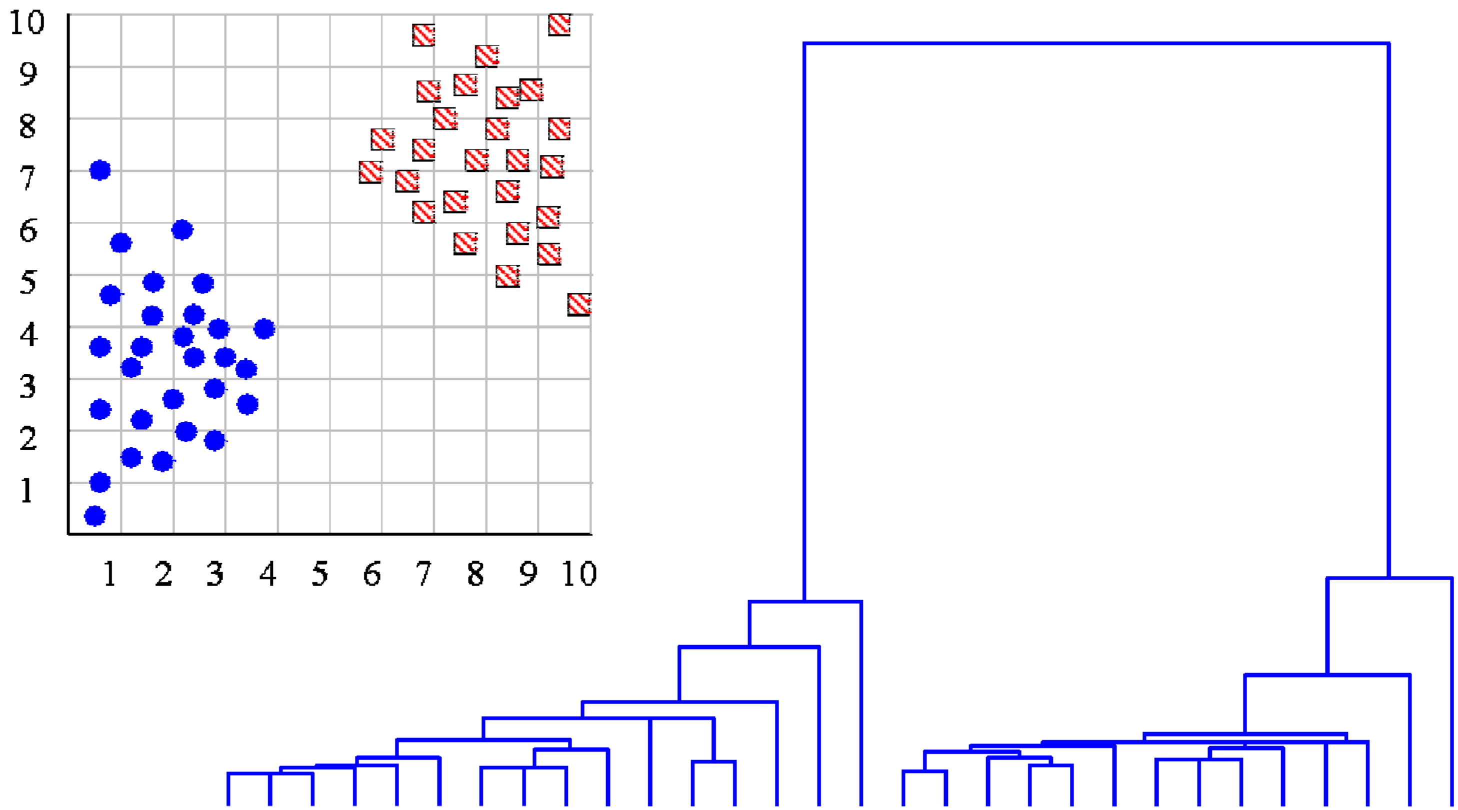
Perfectly clustered using a hierarchy



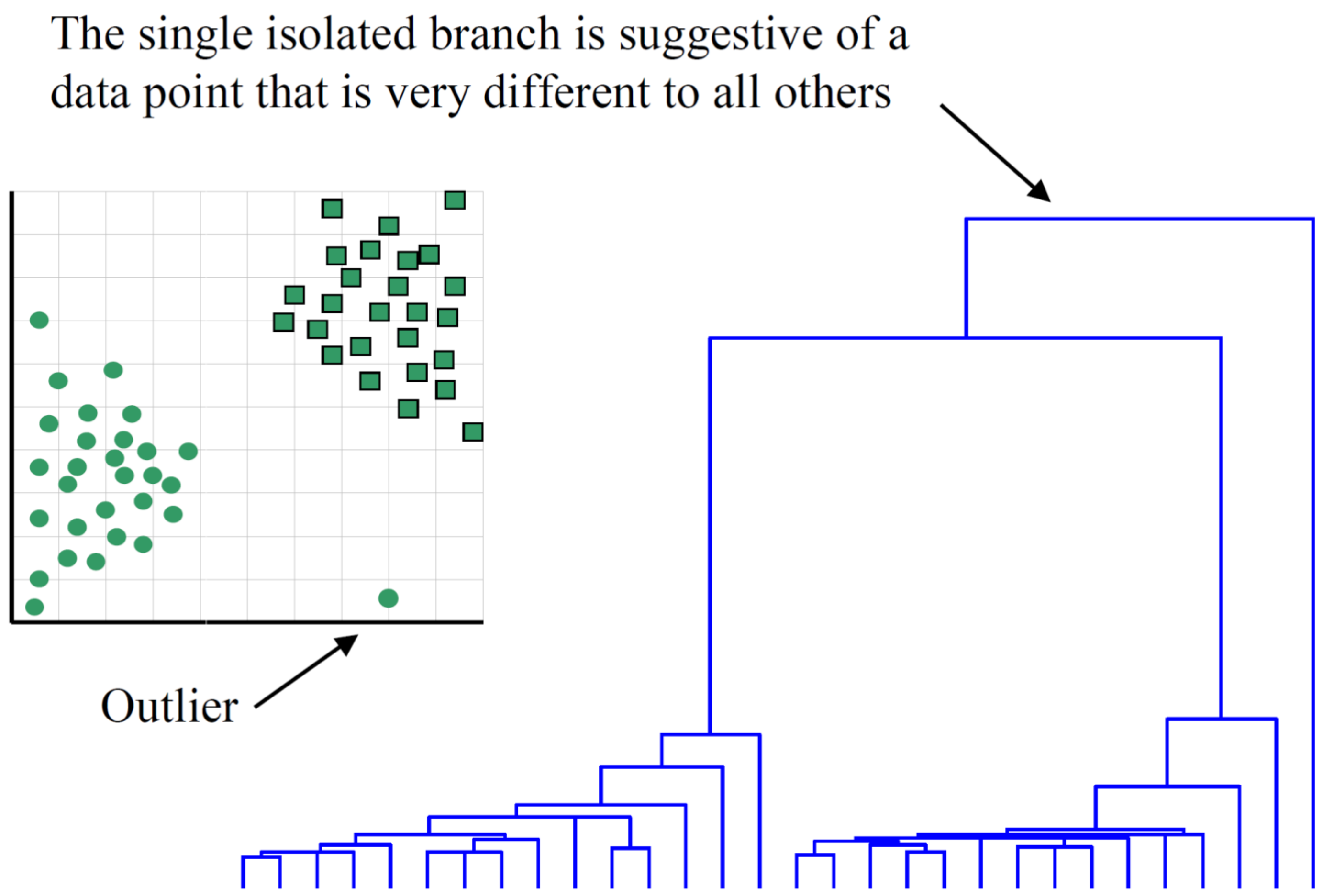

be not he I it this the his a and but in on with for at from of to as is was

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



Hierarchical Clustering

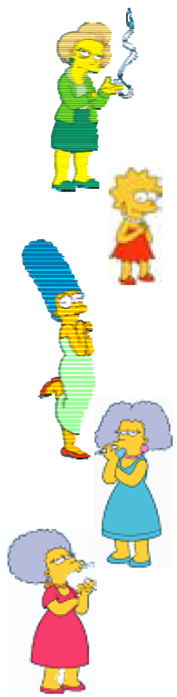
- Bottom-up:
  1. Start with a separate cluster for each object
  2. Determine the two most similar clusters and merge into a new cluster. Repeat on the new clusters that have been formed.
  3. Terminate when one large cluster containing all objects has been formed

Example of a similarity measure:

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

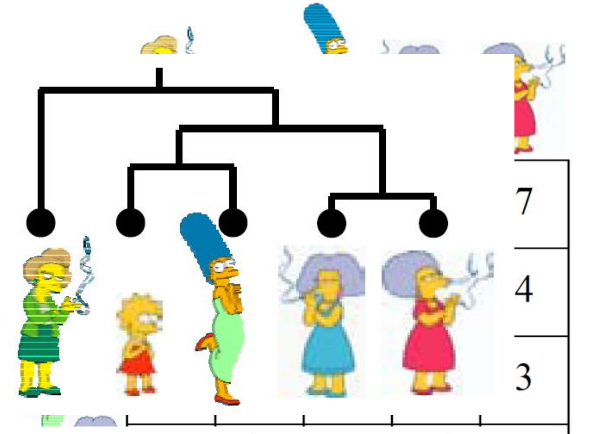
- Top-down
  1. Start from a cluster of all objects
  2. Iteratively determine the cluster that is least coherent and split it.
  3. Repeat until all clusters have one object.

Distance Matrix

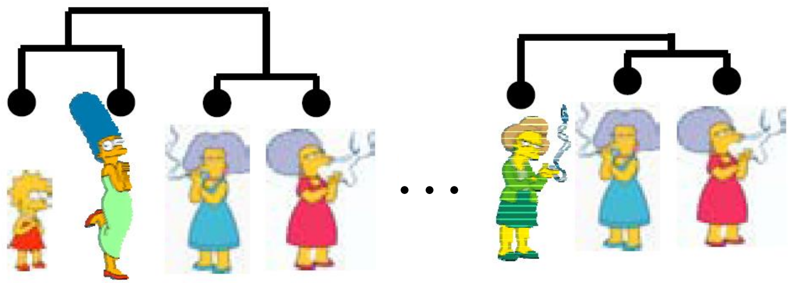


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

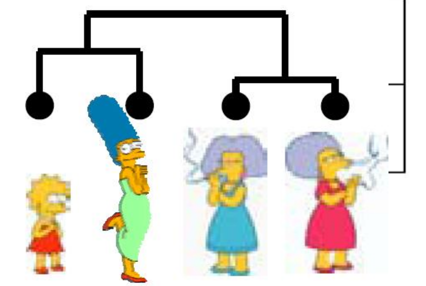
Bottom-Up Agglomerative



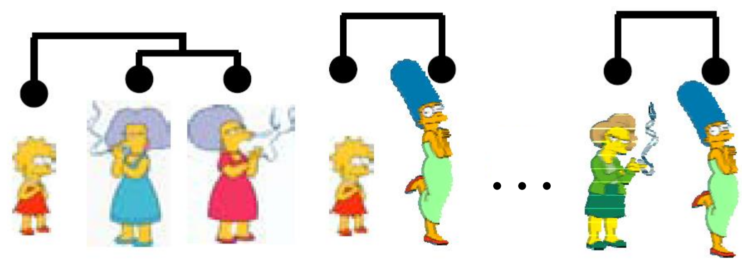
Consider all possible merges



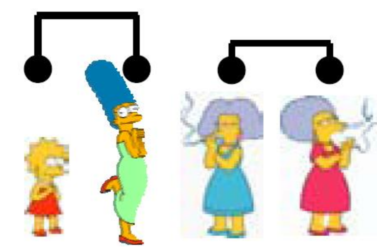
Choose the best



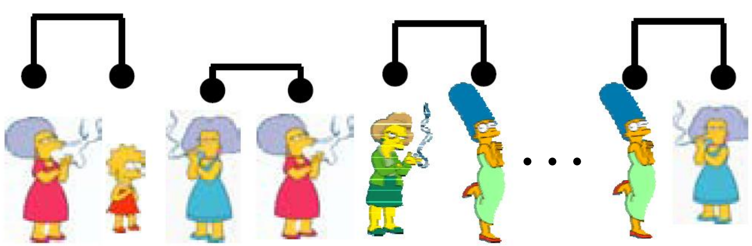
Consider all possible merges



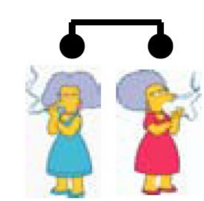
Choose the best



Consider all possible merges



Choose the best

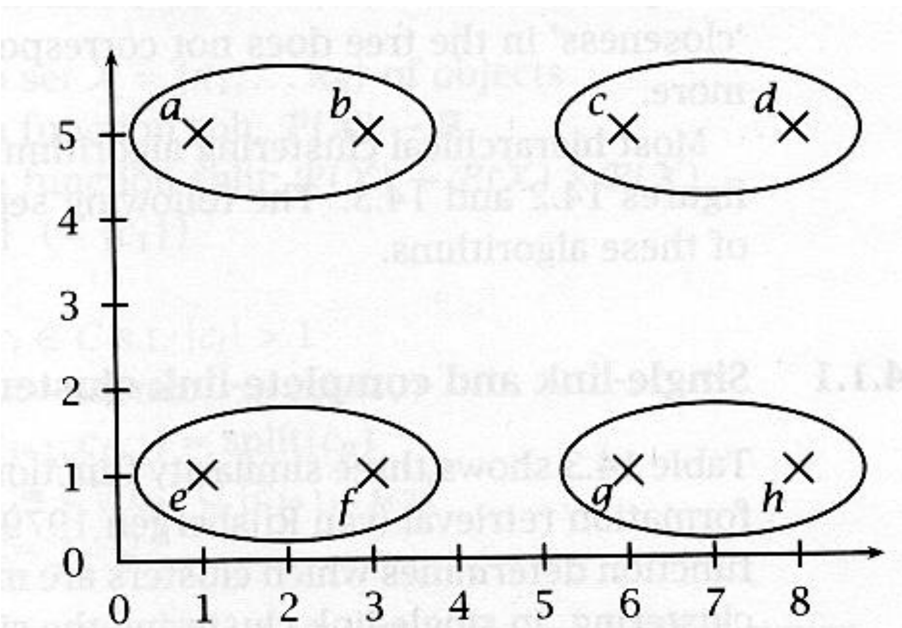


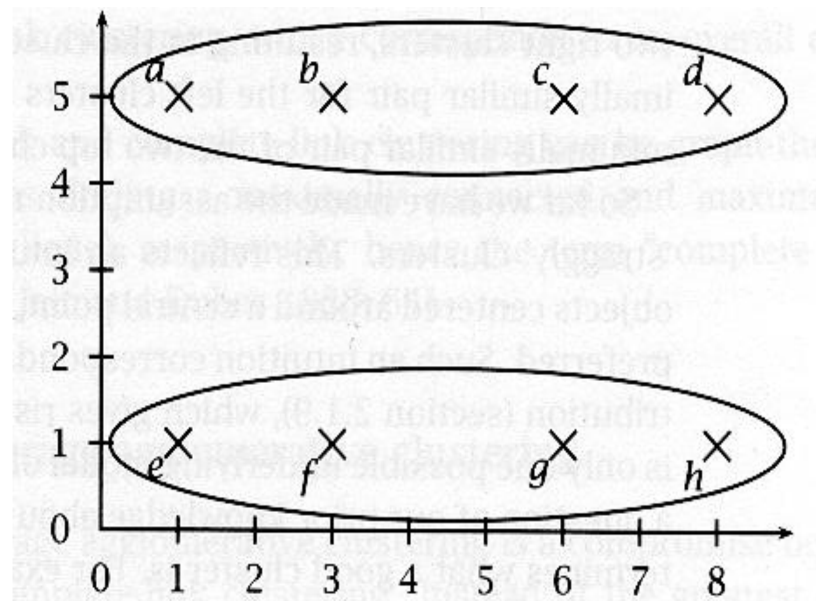
Similarity Measures for Hierarchical Clustering

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

Single-Link

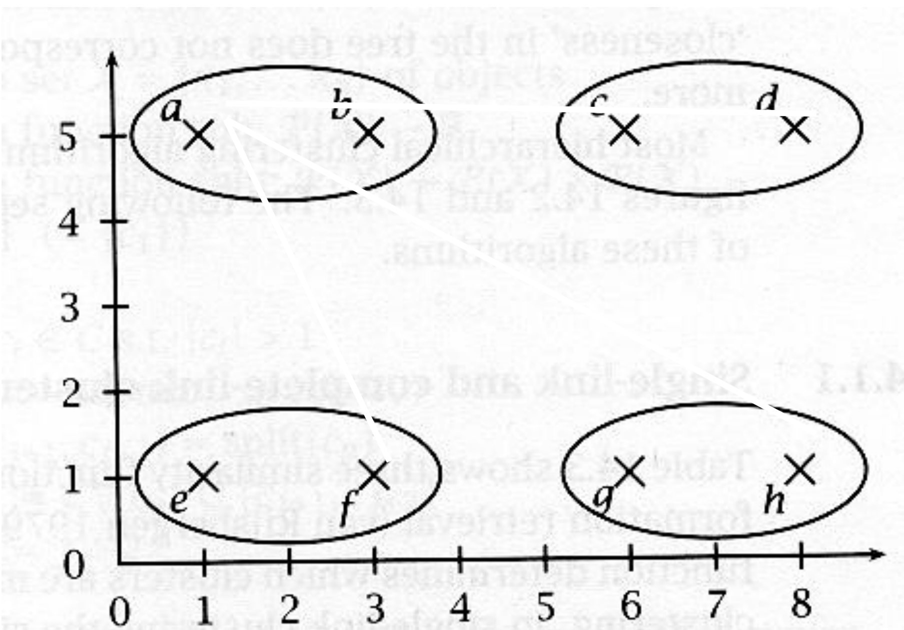
- Similarity function focuses on local coherence

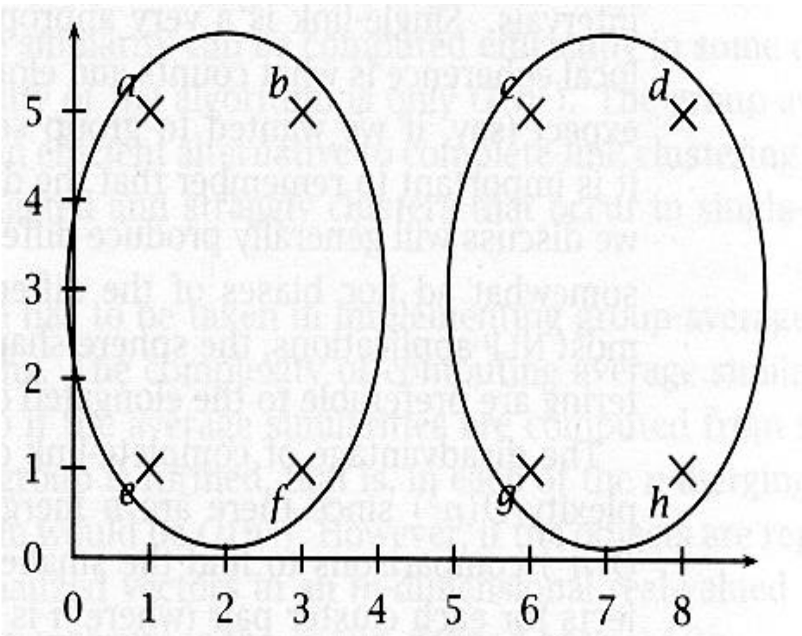




Complete-Link

- 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