

10601

Machine Learning

Hierarchical clustering

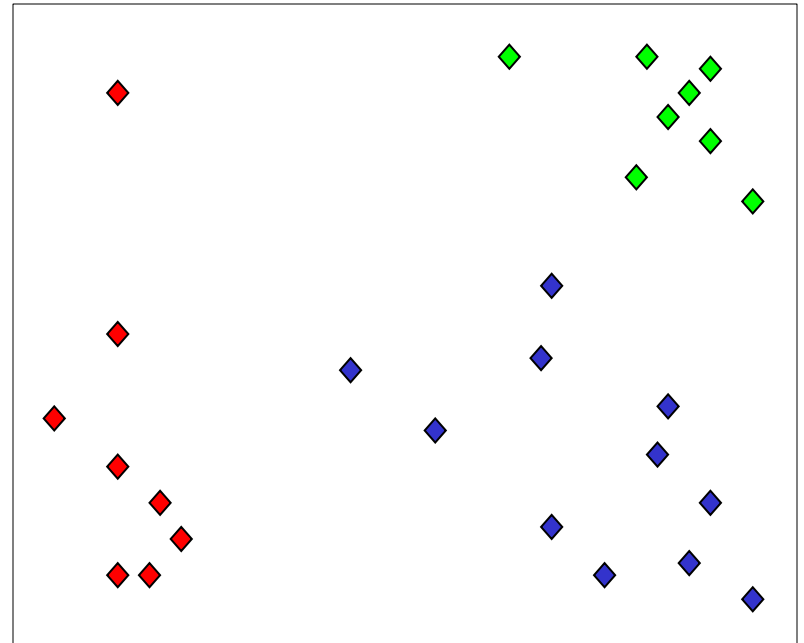
Reading: Bishop: 9-9.2

# Second half: Overview

- Clustering
  - Hierarchical, semi-supervised learning
- Graphical models
  - Bayesian networks, HMMs, Reasoning under uncertainty
- Putting it together
  - Model / feature selection, Boosting, dimensionality reduction
- Advanced classification
  - SVM

# What is Clustering?

- Organizing data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.
- Why do we want to do that?
- Any REAL application?



# Example: clusty

Clusty Search » simpsons - Mozilla Firefox

File Edit View History Bookmarks Tools Help Most Visited @yahoo @cs @andrew gmail sb compbio BBC

http://clusty.com/search?v%3afile=viv\_1023%4019%3akiZm1v&v%3aframe=tree&v%3astate= Google

web news images wikipedia blogs jobs more »

simpsons Search advanced preferences

clusters sources sites remix

All Results (224)

- Pictures (62)
- Games (21)
- Movie (18)
- Collectibles (14)
- Downloads (15)

• **Witness, Trial** (10)

- Bruce Fromong (4)
- Jurors Hear (3)
- Alleged robbery (3)
- Murder, Las Vegas (2)
- Other Topics (1)

• FOX, Broadcasting Company (7)

• Quotes (12)











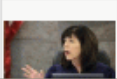
• Episode Guides (6)

• Simpson College (10)

more | all clusters

Cluster **Witness, Trial** contains 10 documents.

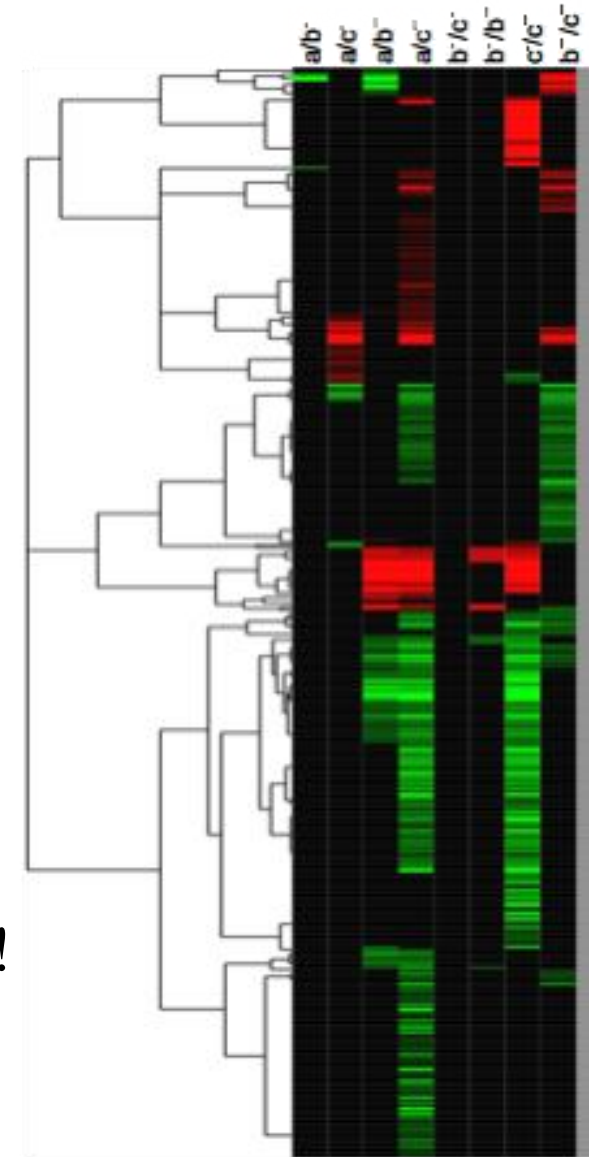
Search Results

- Witness contradicts self in O.J. Simpson trial**     
 Sep 17, 2008 - A key **witness** in the O.J. **Simpson** robbery **trial** was confronted with contradictions in his **testimony** Tuesday, including his claim that he didn't try to profit from the casino hotel room confrontation that led to charges against the former football star. Memorabilia dealer Bruce Fromong, who returned to the stand after becoming ill Monday, told defense attorney Gabriel Grasso he didn't have money on his mind while allegedly being robbed of sports collectibles by **Simpson** and a group of other men. "You ...  
[news.yahoo.com/s/ap/20080917/ap\\_on\\_re\\_us/oj\\_simpson](http://news.yahoo.com/s/ap/20080917/ap_on_re_us/oj_simpson) - [cache] - Yahoo! News
- Witness in Simpson trial says gun brandished in incident**     
 Sep 16, 2008 - A **witness** who says he was robbed by O.J. **Simpson** testified that a gun was brandished during the incident as the former football star's robbery and kidnapping **trial** opened. Bruce Fromong, 54, one of the two collectibles dealers at the center of the case, told the jury on Monday that someone in the room during the alleged robbery shouted, "Put the gun down," contradicting **Simpson's** claim he did not know firearms were present. The **witness** said he could not recall which of the six men who burst into the ...  
[news.yahoo.com/s/afp/20080916/en\\_afp/entertainmentuscrimetrialsimpson](http://news.yahoo.com/s/afp/20080916/en_afp/entertainmentuscrimetrialsimpson) - [cache] - Yahoo! News
- Key OJ Simpson witness clutches chest in court**     
 Sep 16, 2008 - A key **witness** in O.J. **Simpson's** kidnap and robbery **trial** became ill on Monday while testifying about a hotel room confrontation at the heart of the case -- clutching his chest before bailiffs helped him from the **witness** stand.

Done

# Example: clustering genes

- Microarrays measures the activities of all genes in different conditions
- Clustering genes can help determine new functions for unknown genes
- An early “killer application” in this area
  - The most cited (12,309) paper in PNAS!



# Unsupervised learning

- Clustering methods are unsupervised learning techniques
  - We do not have a teacher that provides examples with their labels
- We will also discuss dimensionality reduction, another unsupervised learning method later in the course

# Outline

- Distance functions
- Hierarchical clustering
- Number of clusters

# What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

**Webster's Dictionary**



Similarity is hard to define, but...

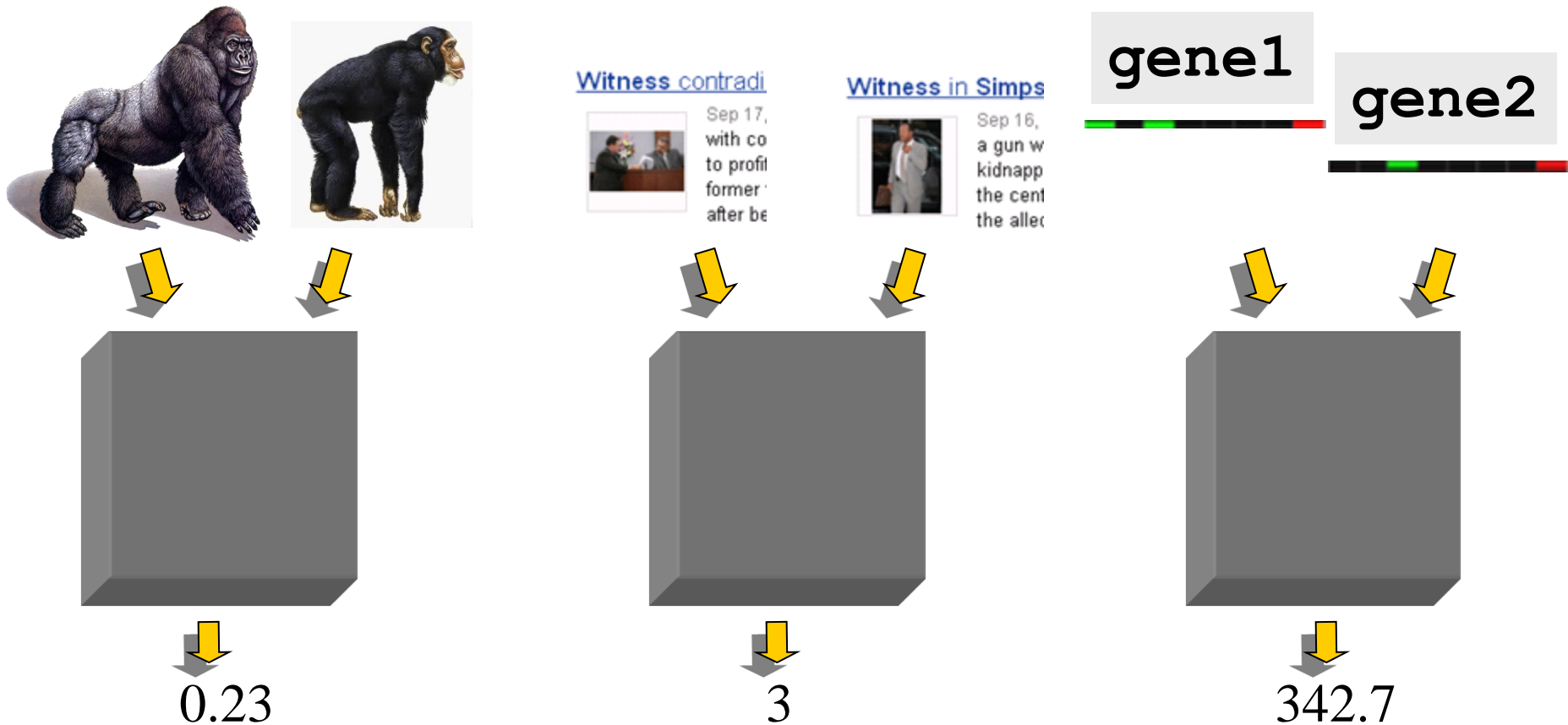
*“We know it when we see it”*

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.



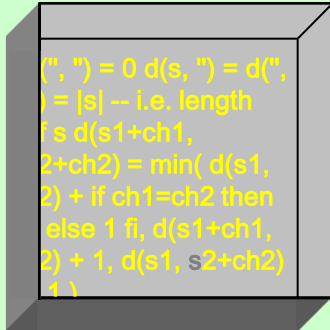
# Defining Distance Measures

**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  $D(O_1, O_2)$



gene1

gene2



Inside these black boxes:  
some function on two variables  
(might be simple or very  
complex)

3

A few examples:

- Euclidian distance

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Correlation coefficient

$$s(x, y) = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$$

- Similarity rather than distance
- Can determine similar trends

# Outline

- Distance measure
- Hierarchical clustering
- Number of clusters

# Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Interpretability and usability

## Optional

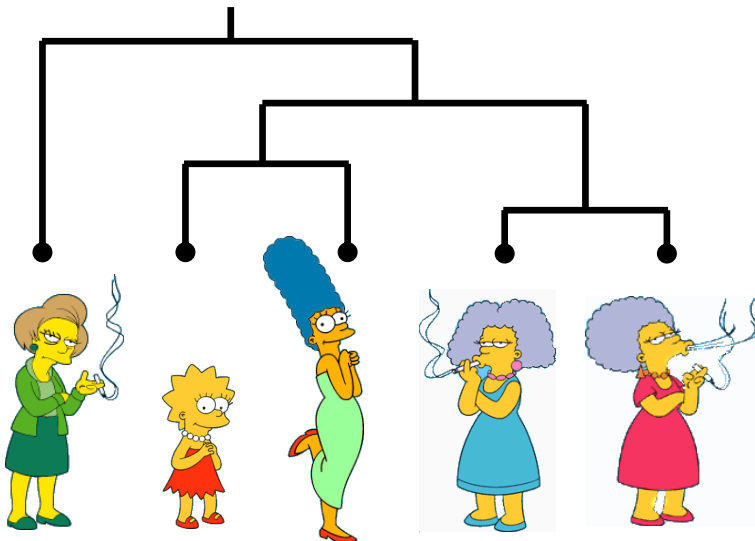
- Incorporation of user-specified constraints

# Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion (focus of this class)

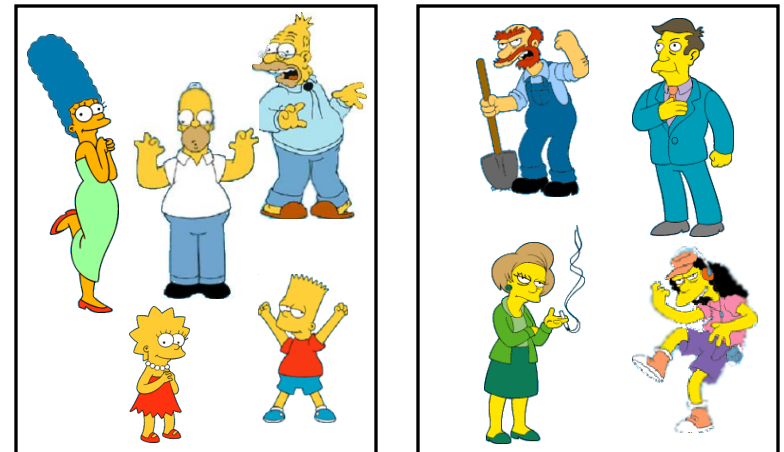
Bottom up or top down

**Hierarchical**



Top down

**Partitional**

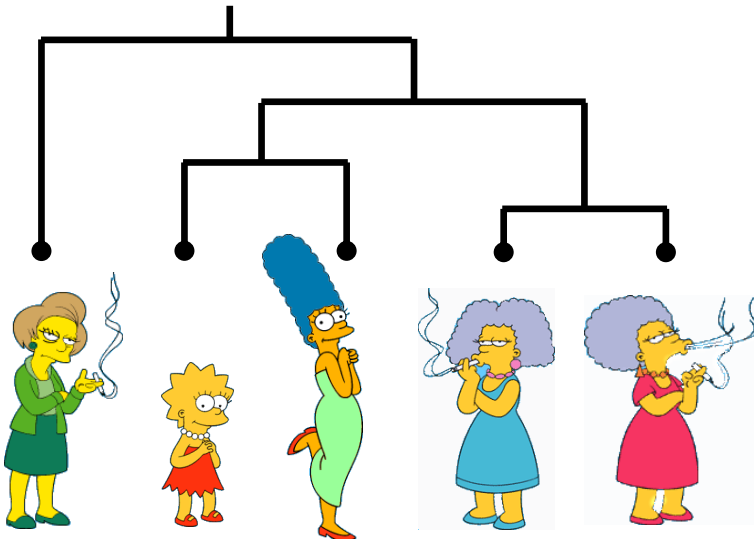


# (How-to) Hierarchical Clustering

The number of dendrograms with  $n$  leafs =  $(2n - 3)! / [(2^{(n-2)}) (n - 2)!]$

Number of Leafs	Number of Possible Dendrograms
2	1
3	3
4	15
5	105
...	...
10	34,459,425

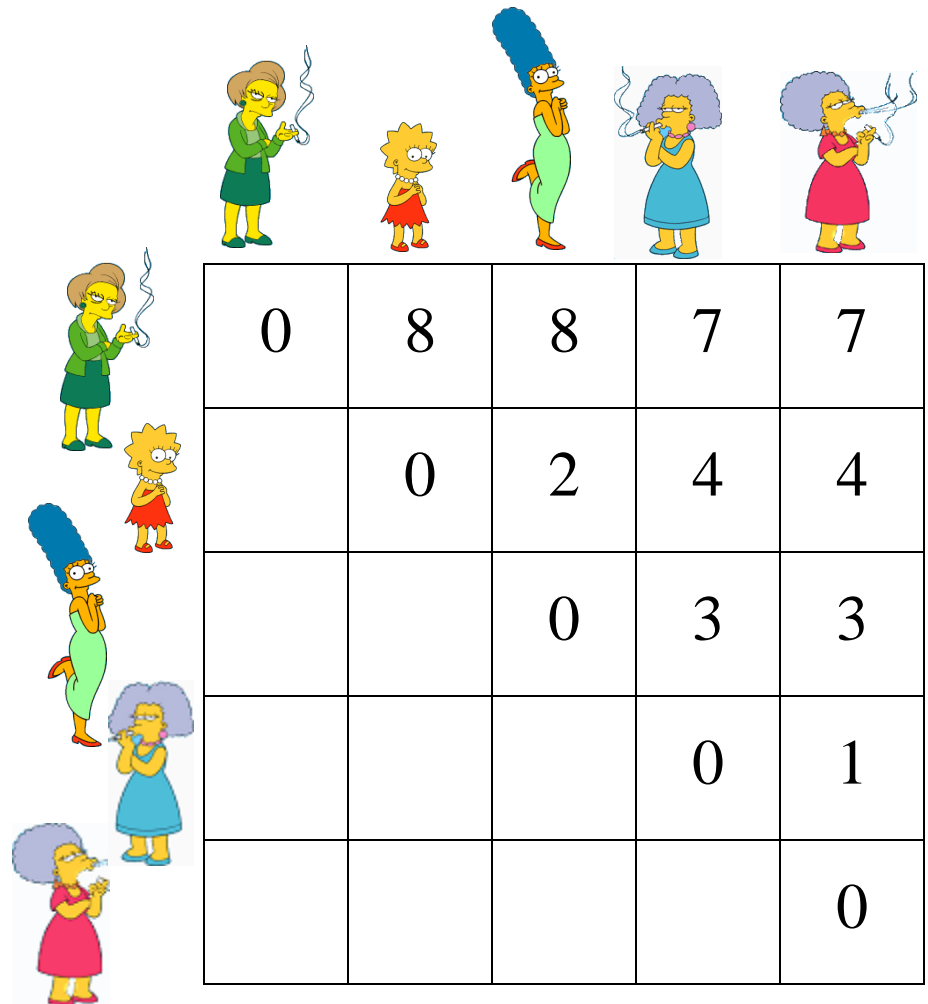
**Bottom-Up (agglomerative):** Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.










We begin with a distance matrix which contains the distances between every pair of objects in our database.

$$D(\text{Mrs. Simpson}, \text{Lisa Simpson}) = 8$$

$$D(\text{Marge Simpson}, \text{Bart Simpson}) = 1$$

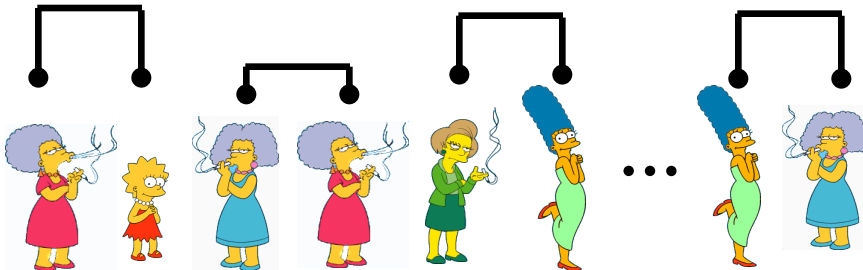


				
	0	8	8	7
		0	2	4
			0	3
				0

# Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

Consider all possible merges...



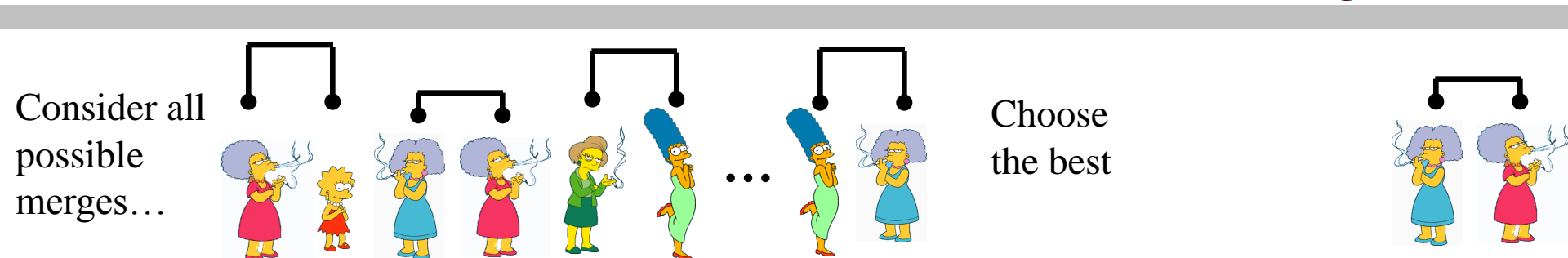
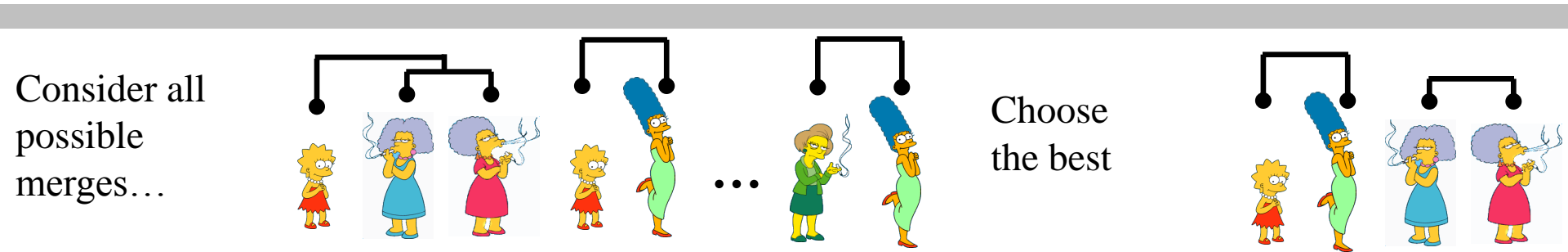
Choose the best





# Bottom-Up (agglomerative):

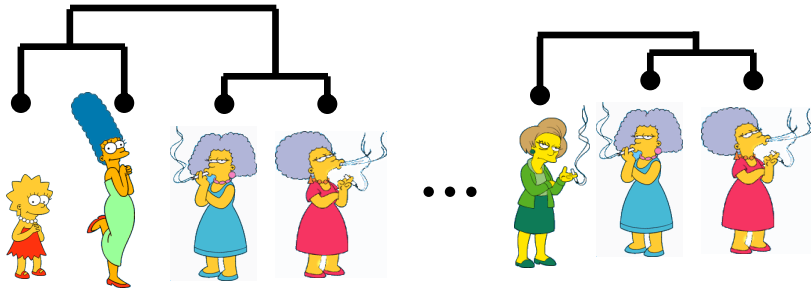
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



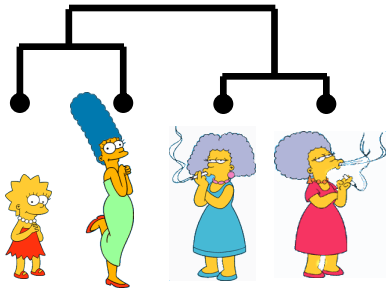
# Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

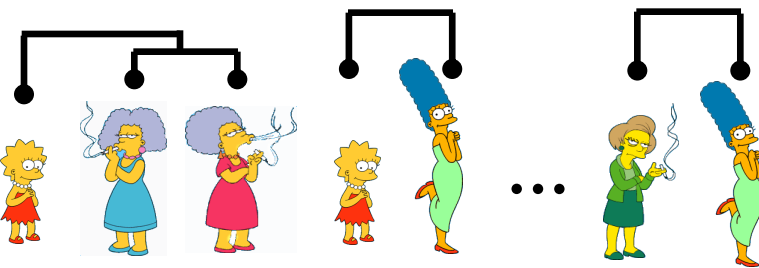
Consider all possible merges...



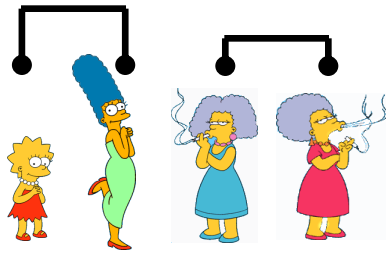
Choose the best



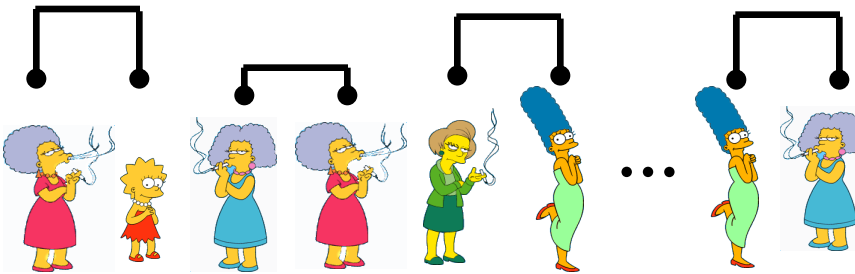
Consider all possible merges...



Choose the best



Consider all possible merges...

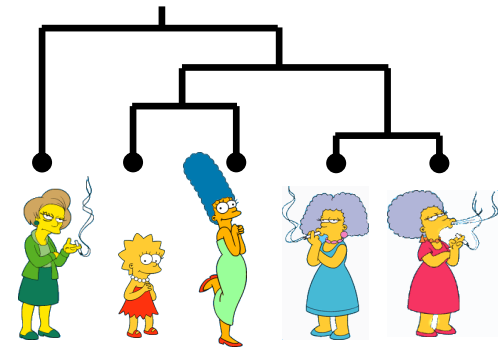


Choose the best

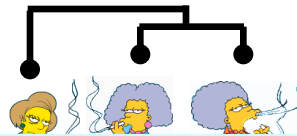
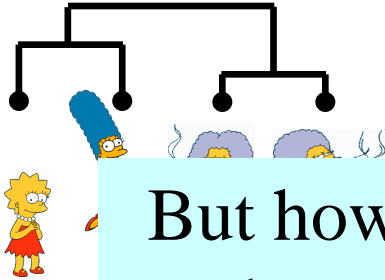


# Bottom-Up (agglomerative):

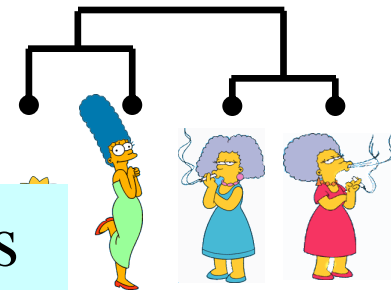
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



Consider all possible merges...

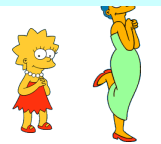
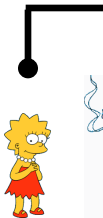


Choose



But how do we compute distances between clusters rather than objects?

Consider all possible merges...



...



the best



Consider all possible merges...



...

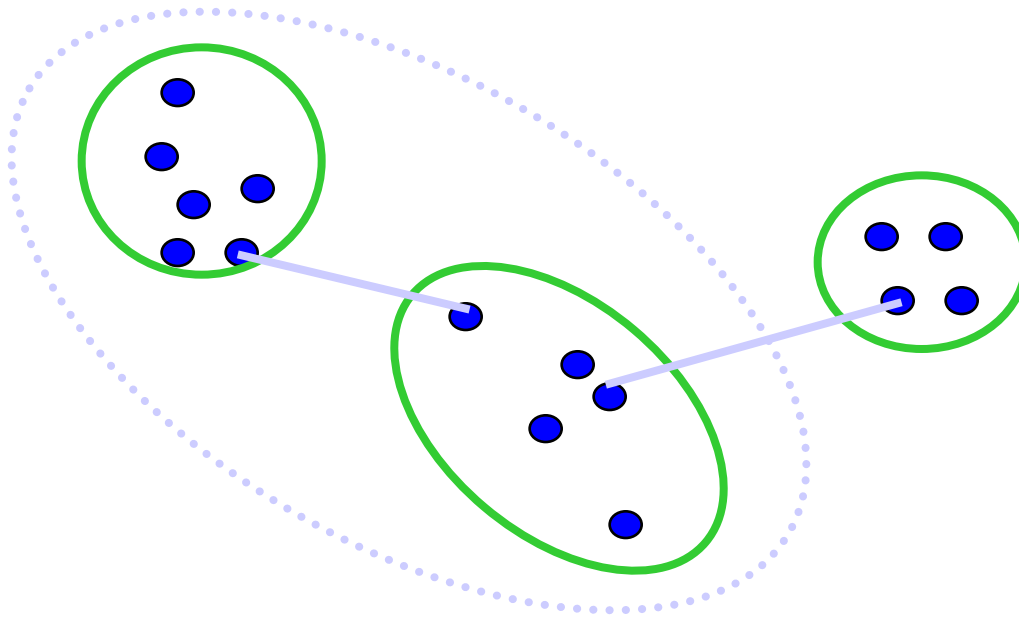


Choose the best



# Computing distance between clusters: Single Link

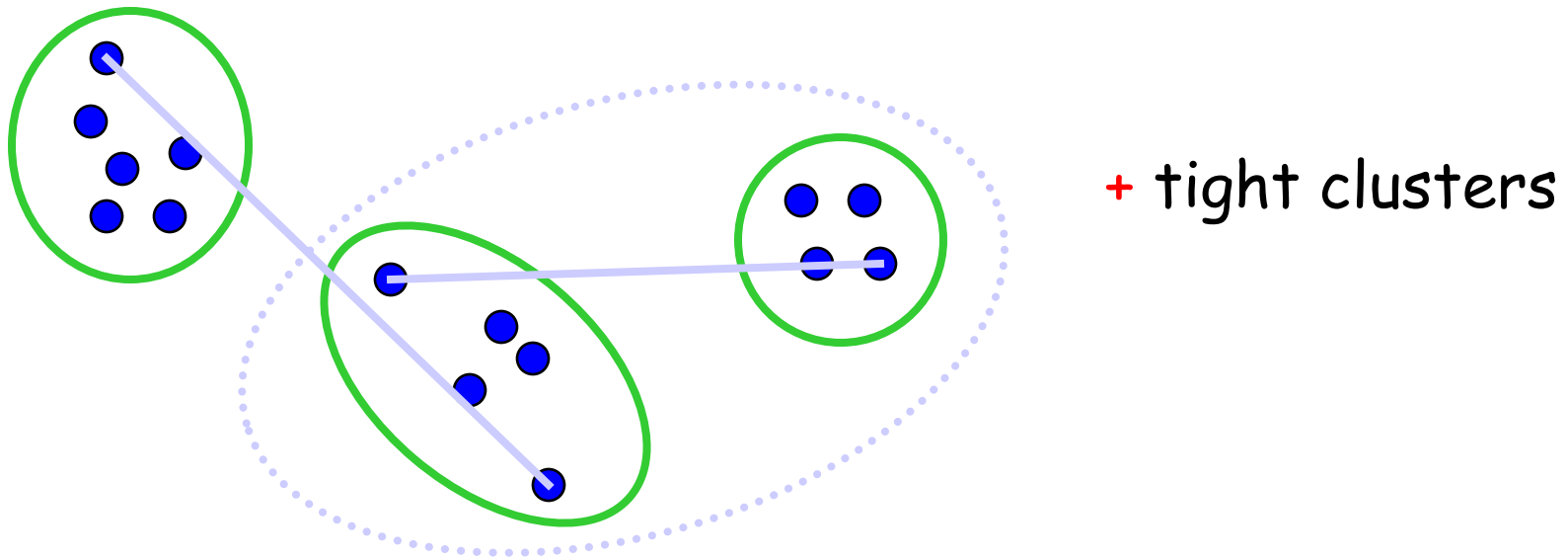
- cluster distance = distance of two **closest** members in each class



- Potentially long and skinny clusters

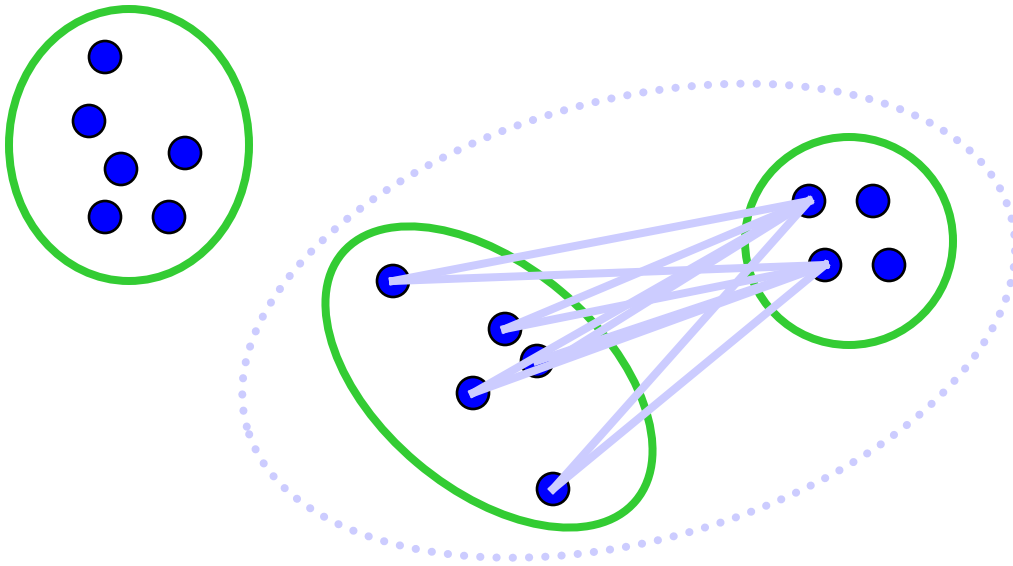
# Computing distance between clusters: : Complete Link

- cluster distance = distance of two farthest members



# Computing distance between clusters: Average Link

- cluster distance = average distance of all pairs

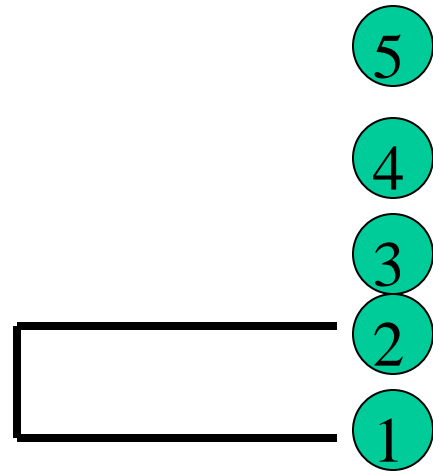


**the most widely  
used measure**

**Robust against  
noise**

# Example: single link

$$\begin{array}{c} 1 \quad 2 \quad 3 \quad 4 \quad 5 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \left[ \begin{array}{ccccc} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{array} \right] \end{array}$$



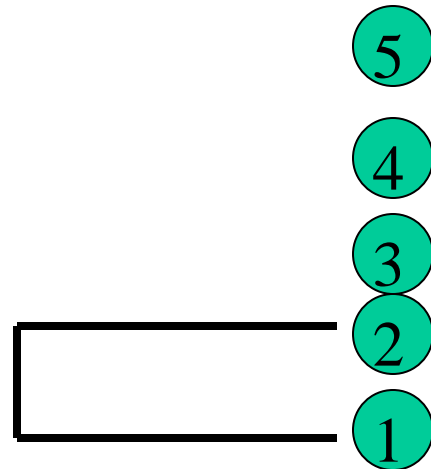
# Example: single link

$$\begin{array}{c}
 \begin{array}{ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
 \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix}
 \end{array}
 \quad \rightarrow \quad
 \begin{array}{c}
 \begin{array}{cccc}
 & (1,2) & 3 & 4 & 5 \\
 \begin{array}{c} (1,2) \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix}
 \end{array}
 \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6, 3\} = 3$$

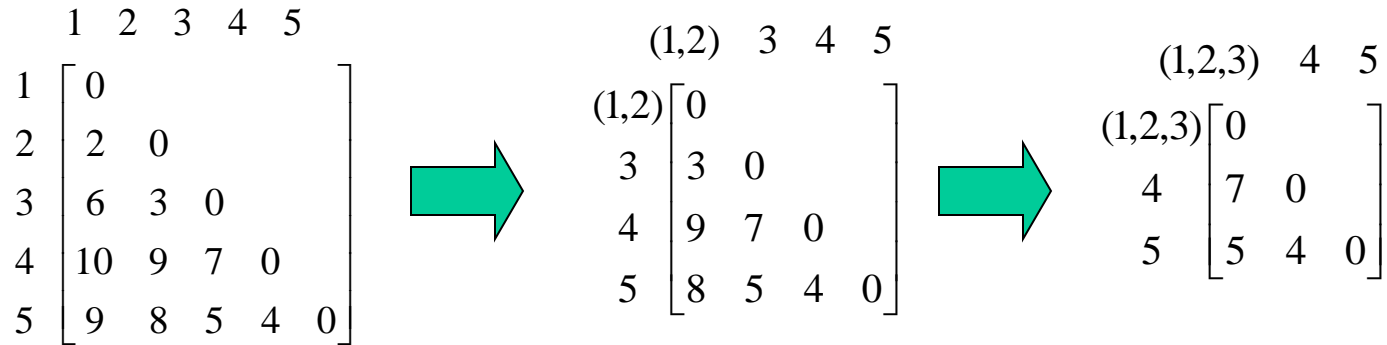
$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10, 9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9, 8\} = 8$$



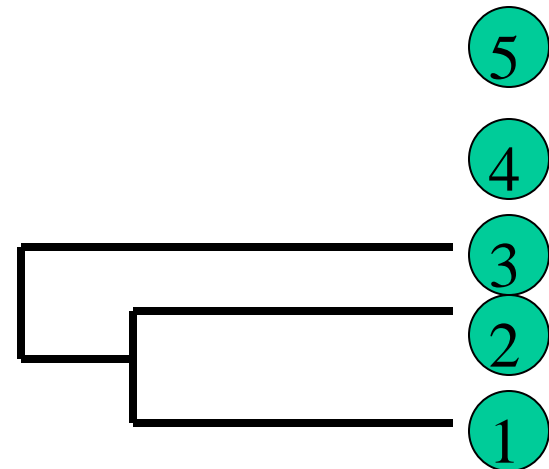


# Example: single link

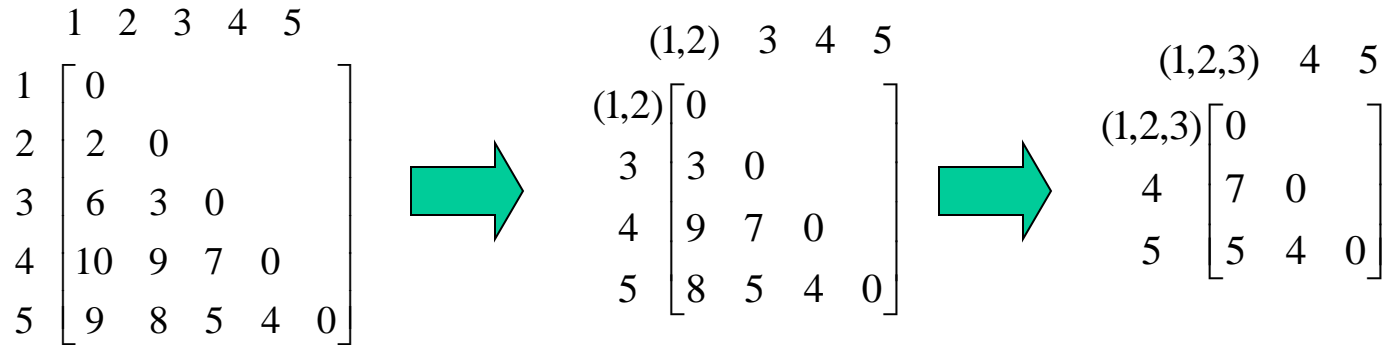


$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9, 7\} = 7$$

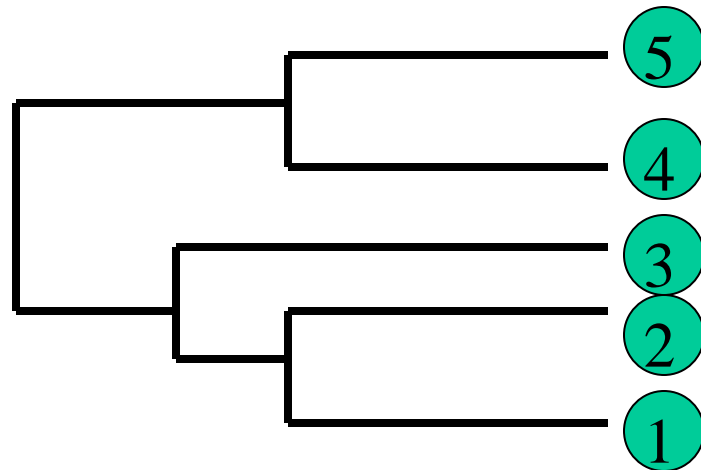
$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8, 5\} = 5$$

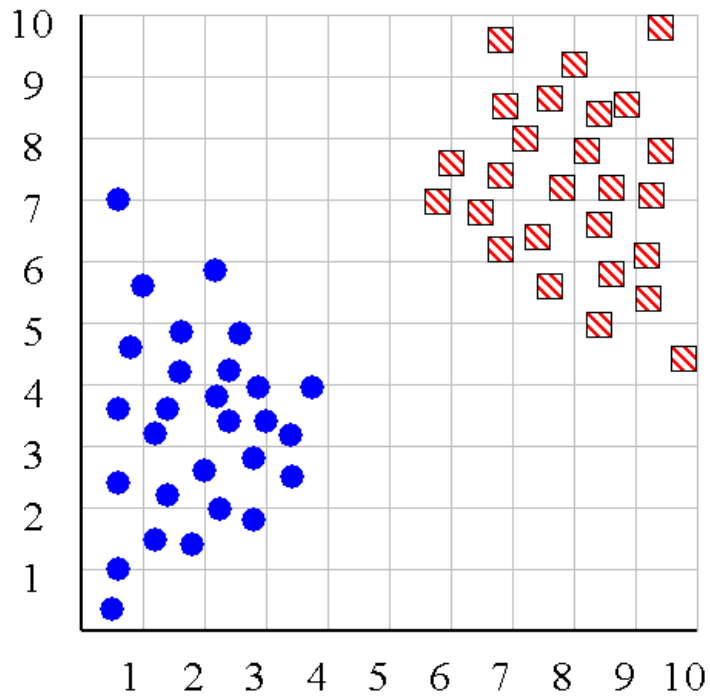


# Example: single link

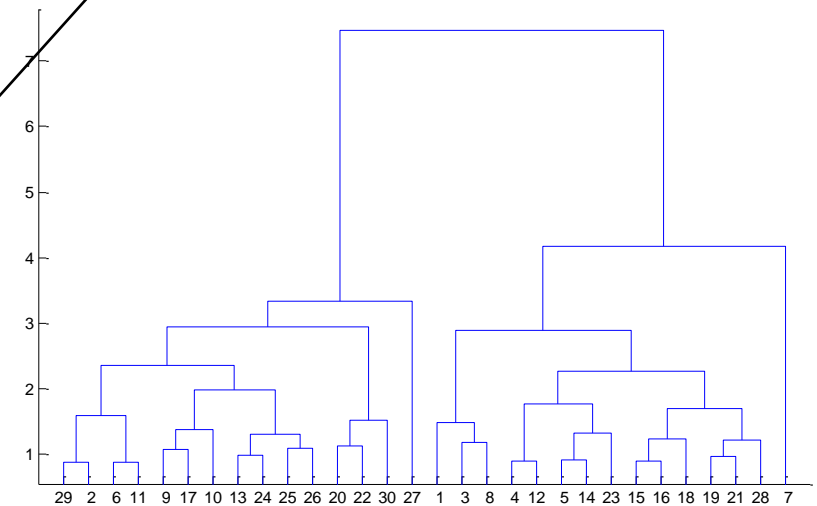
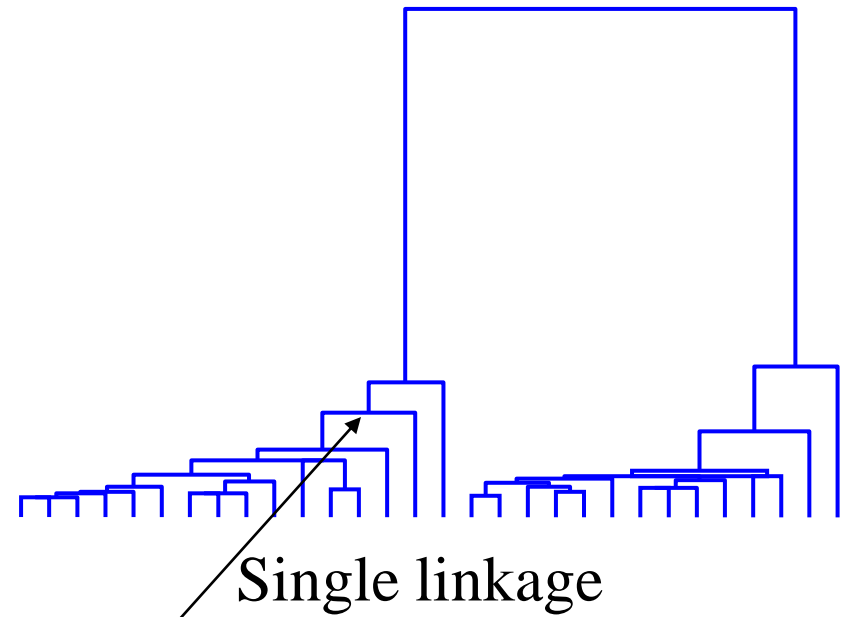


$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$





Height represents  
distance between objects  
/ clusters



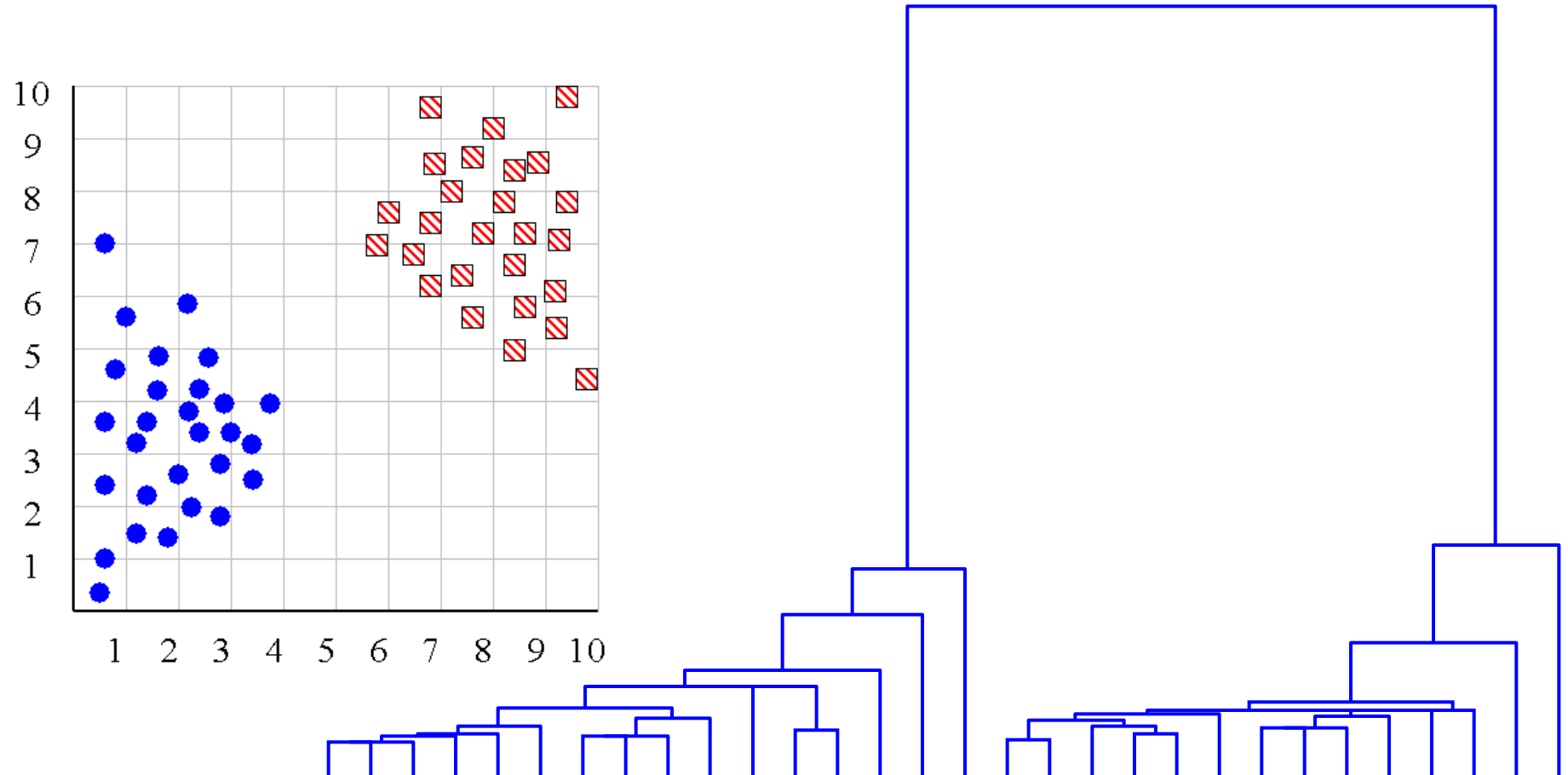
Average linkage

# Summary of Hierarchical Clustering Methods

- No need to specify the number of clusters in advance.
- Hierarchical structure maps nicely onto human intuition for some domains
- They do not scale well: time complexity of at least  $O(n^2)$ , where  $n$  is the number of total objects.
- Like any heuristic search algorithms, local optima are a problem.
- Interpretation of results is (very) subjective.

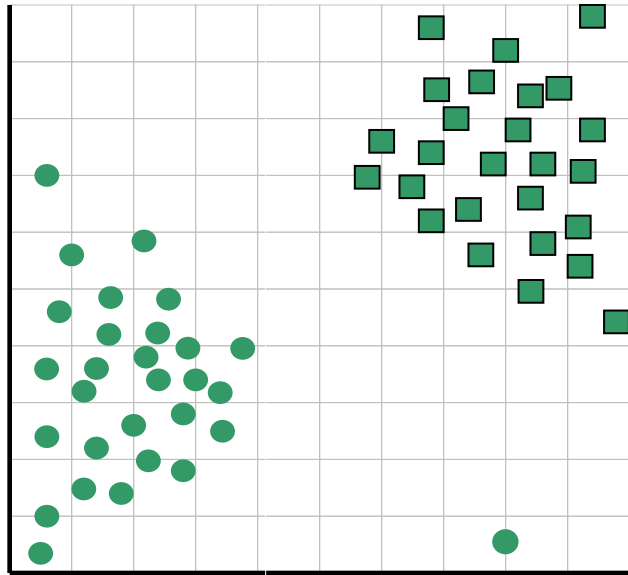
# But what are the clusters?

In some cases we can determine the “correct” number of clusters. However, things are rarely this clear cut, unfortunately.

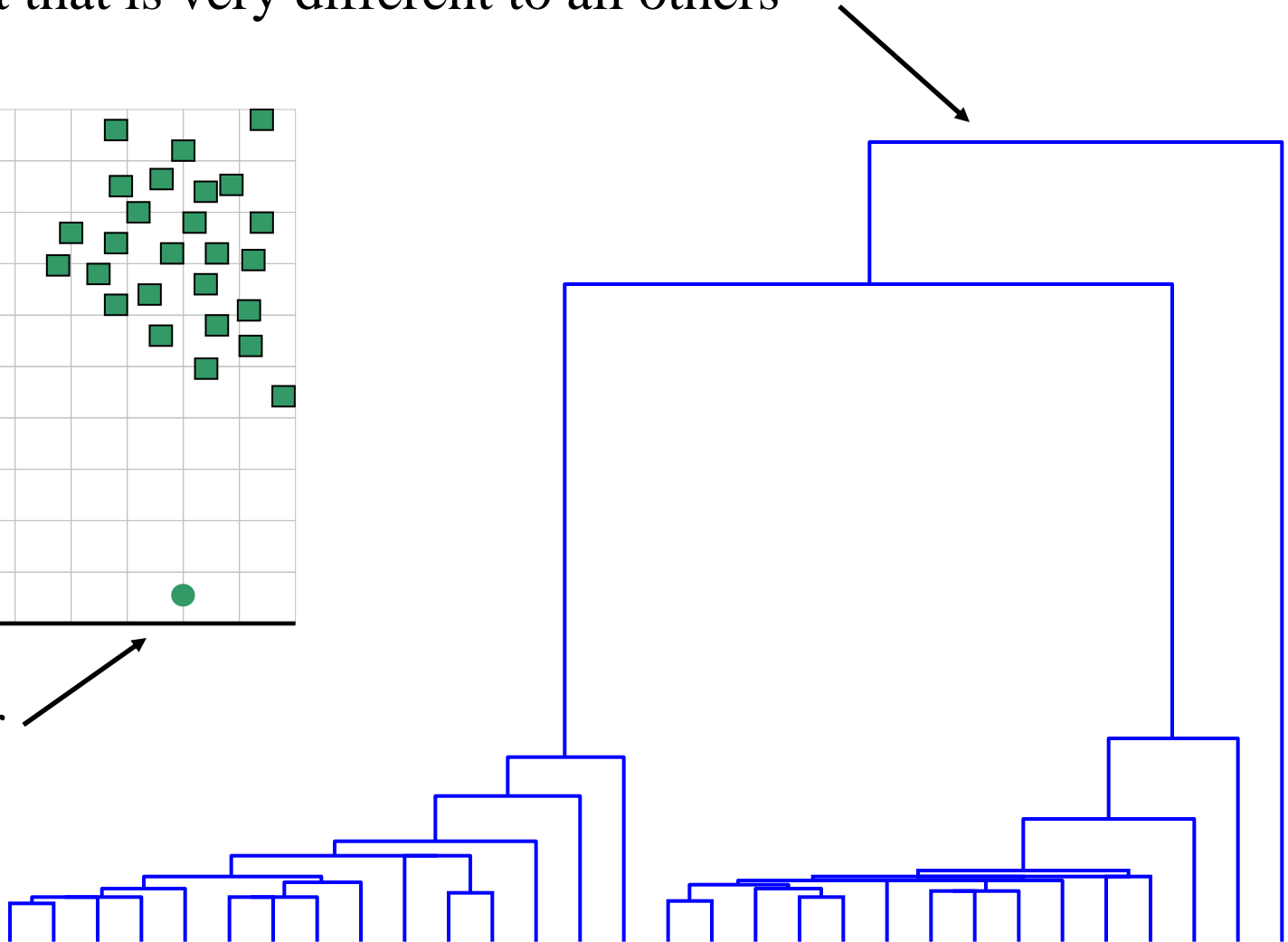


# One potential use of a dendrogram is to detect outliers

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

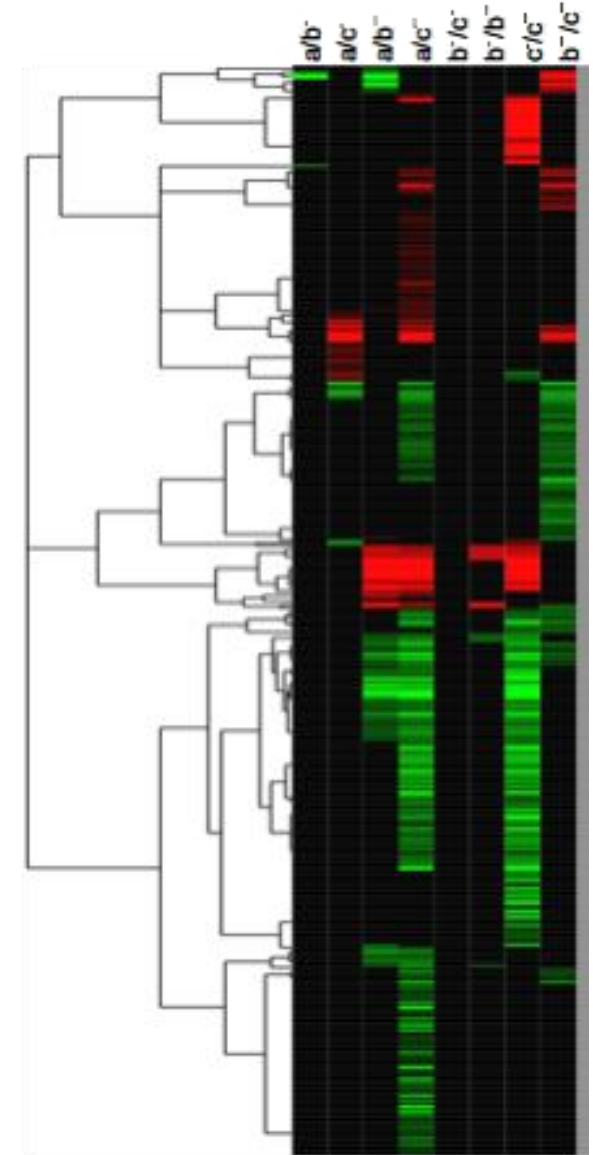


Outlier



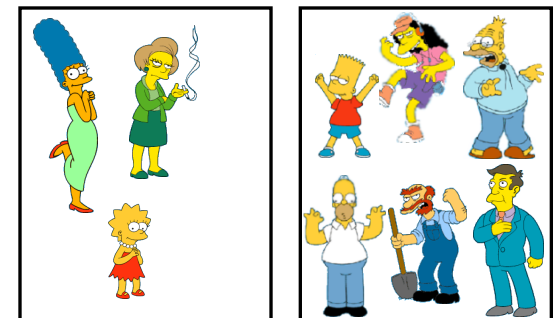
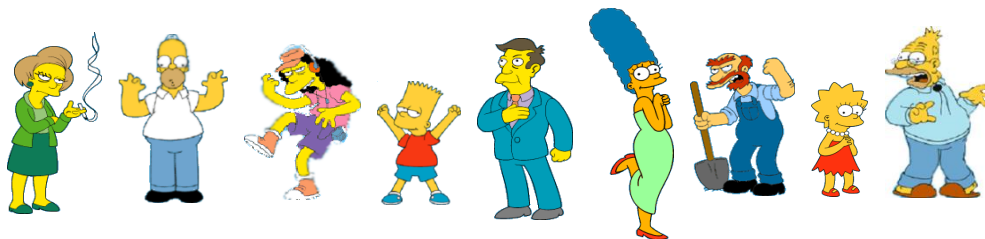
# Example: clustering genes

- Microarrays measures the activities of all genes in different conditions
- Clustering genes can help determine new functions for unknown genes



# Partitional Clustering

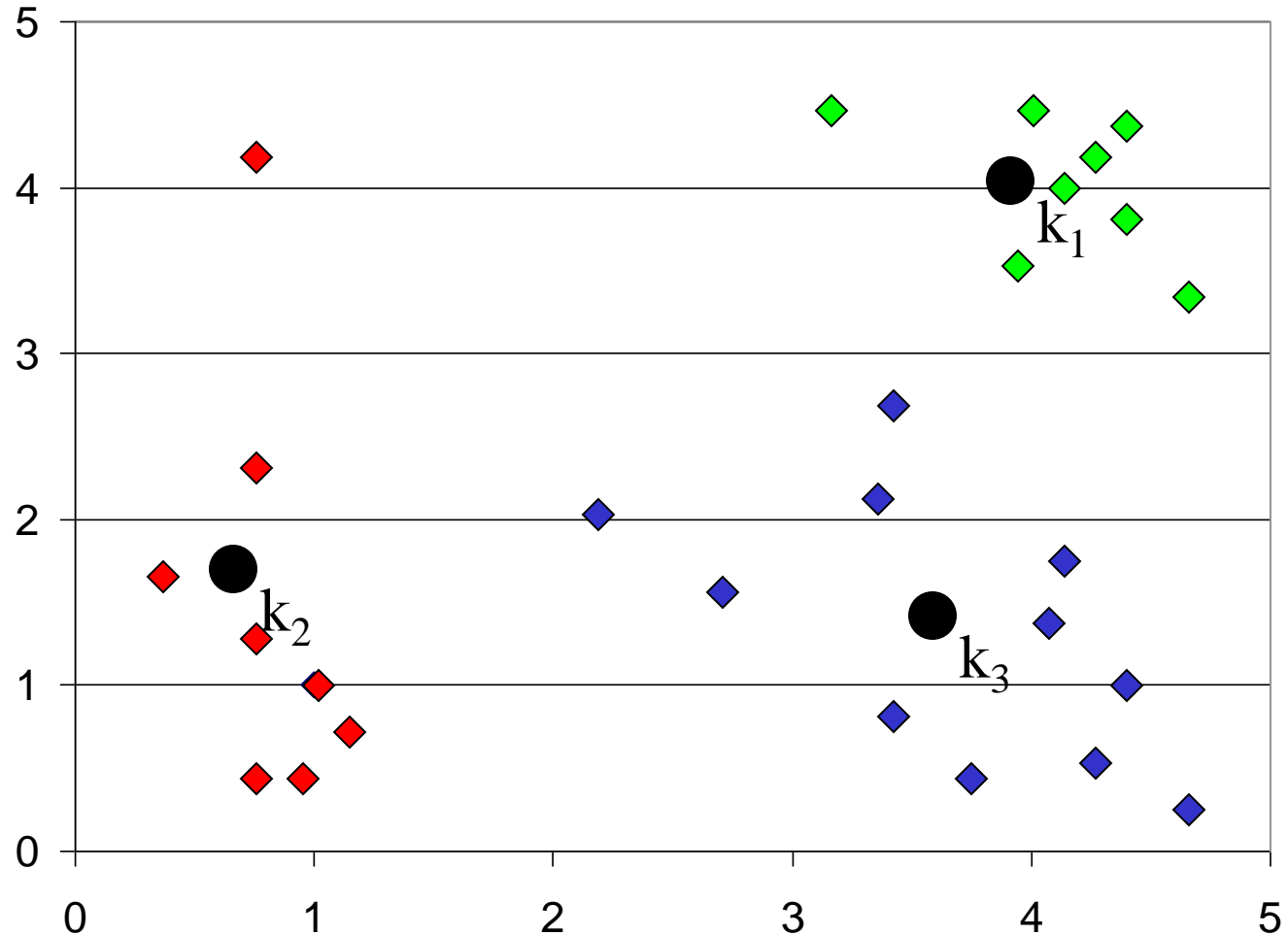
- Nonhierarchical, each instance is placed in exactly one of  $K$  non-overlapping clusters.
- Since the output is only one set of clusters the user has to specify the desired number of clusters  $K$ .



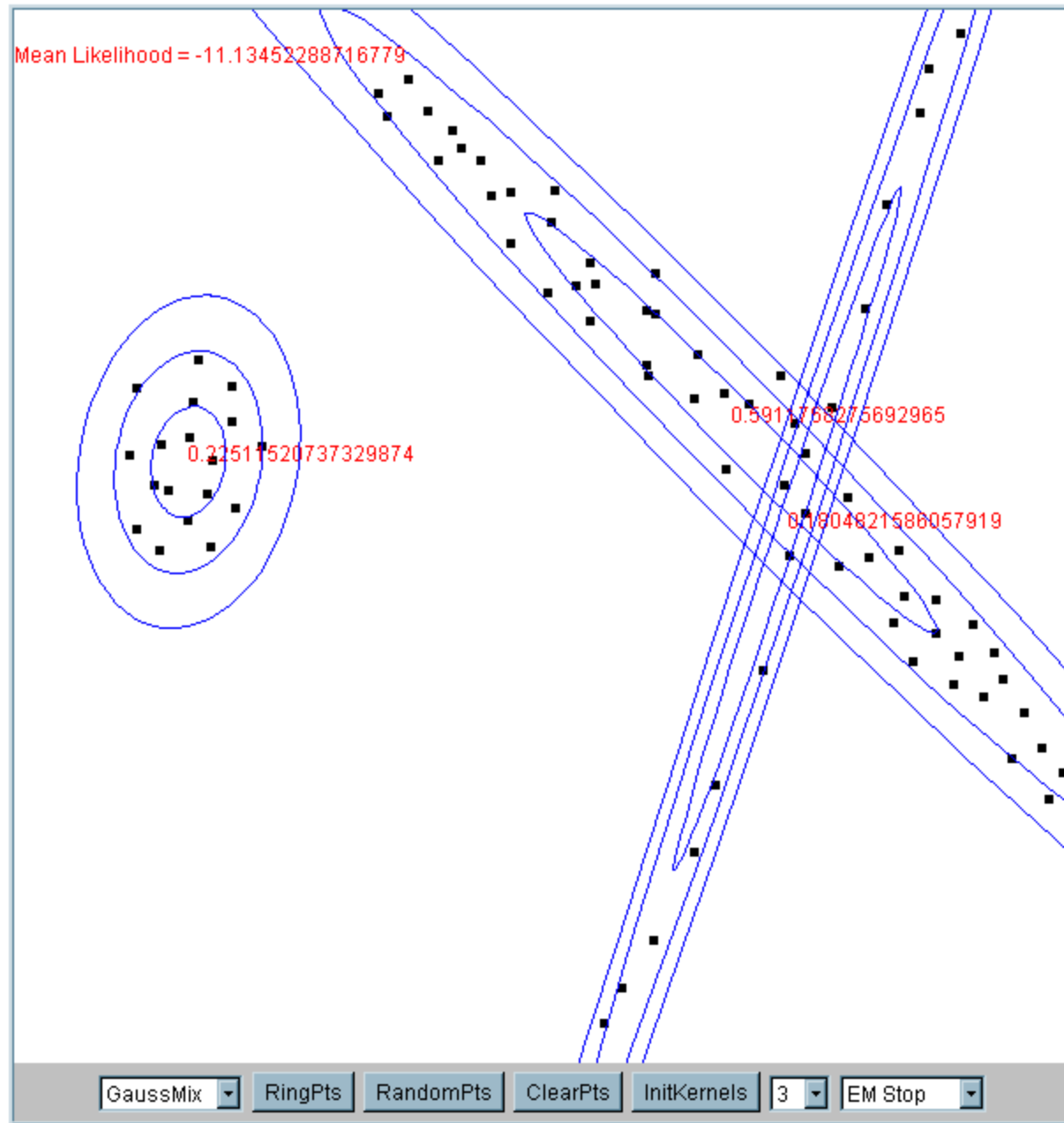


# K-means Clustering: Finished!

Re-assign and move centers, until ...  
no objects changed membership.



# Gaussian mixture clustering



# Clustering methods: Comparison

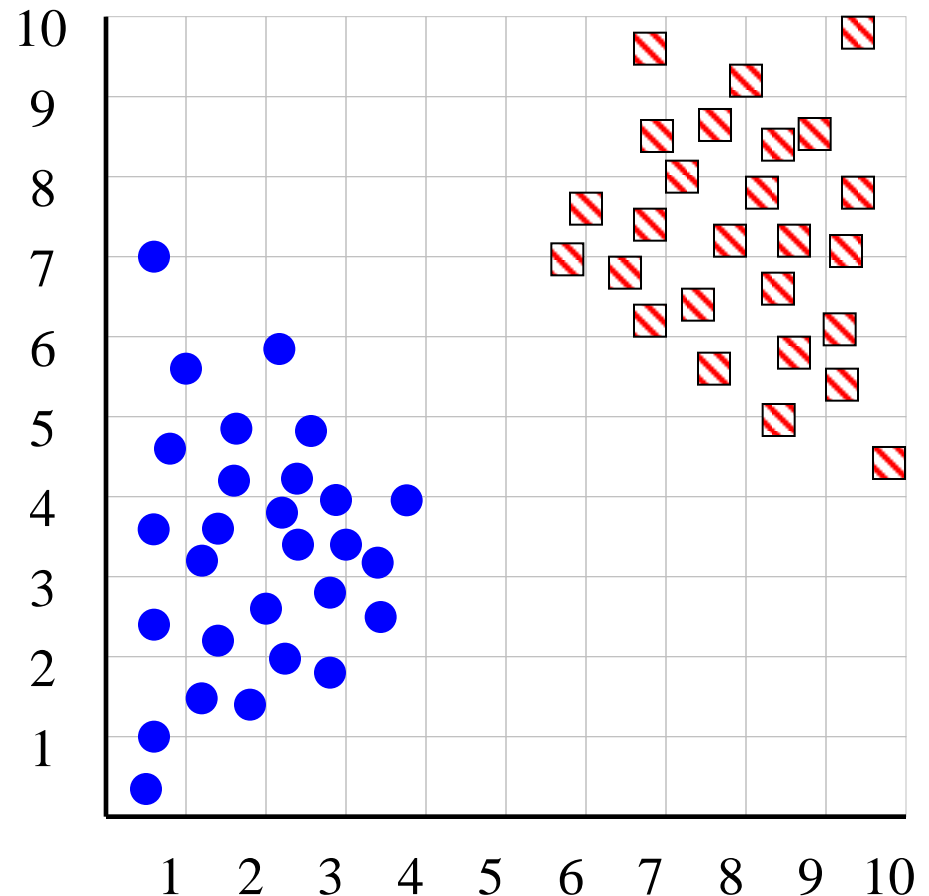
	<b>Hierarchical</b>	<b>K-means</b>	<b>GMM</b>
<b>Running time</b>	naively, $O(N^3)$	fastest (each iteration is linear)	fast (each iteration is linear)
<b>Assumptions</b>	requires a similarity / distance measure	strong assumptions	strongest assumptions
<b>Input parameters</b>	none	$K$ (number of clusters)	$K$ (number of clusters)
<b>Clusters</b>	subjective (only a tree is returned)	exactly $K$ clusters	exactly $K$ clusters

# Outline

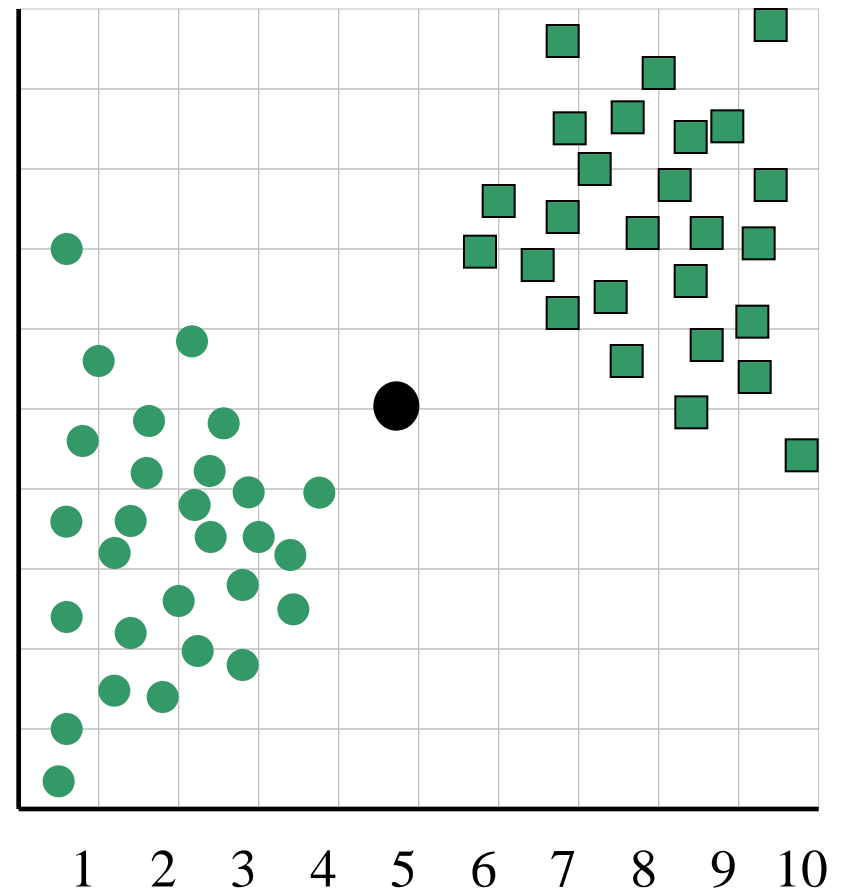
- Distance measure
- Hierarchical clustering
- Number of clusters

# How can we tell the *right* number of clusters?

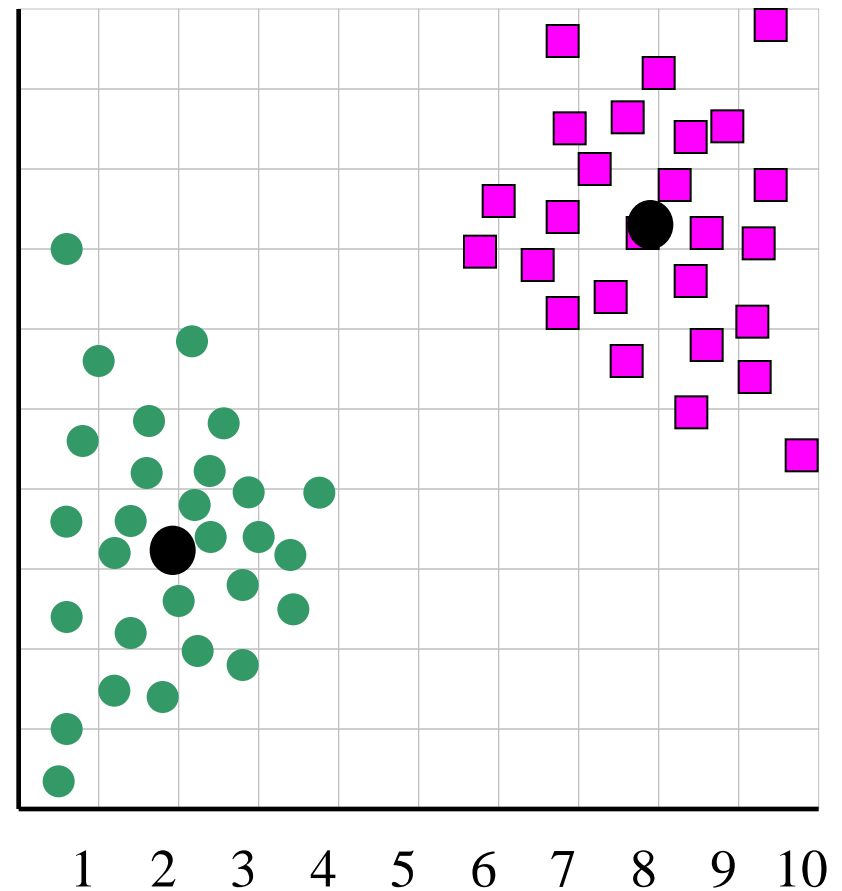
In general, this is an unsolved problem. However there are many approximate methods. In the next few slides we will see an example.



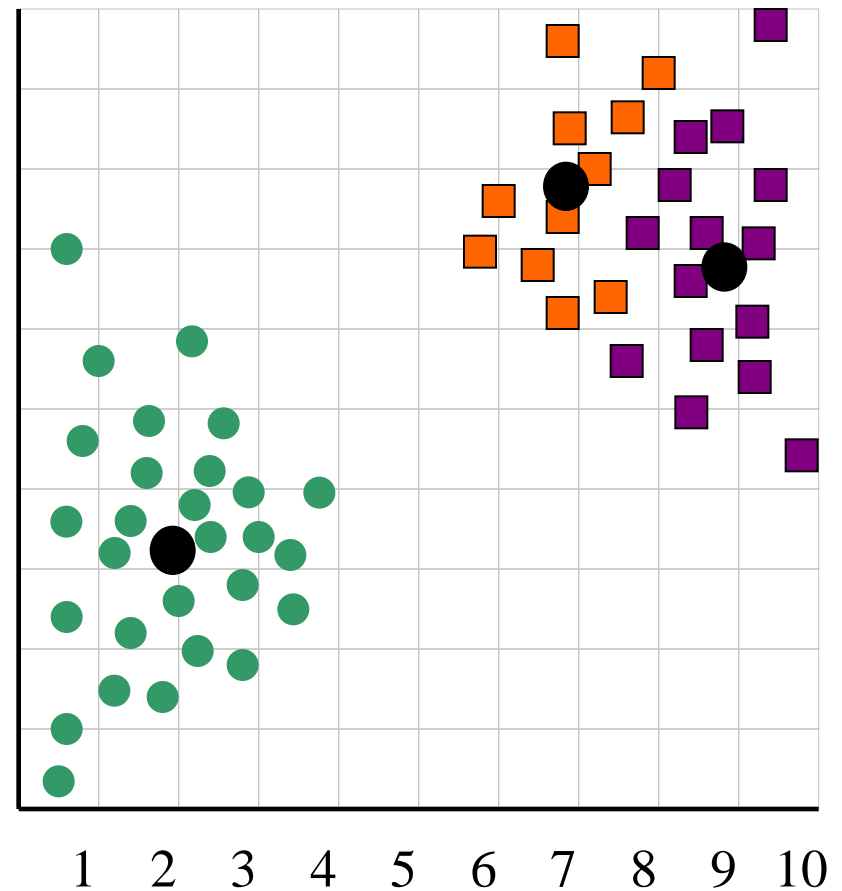
When  $k = 1$ , the objective function is 873.0



When  $k = 2$ , the objective function is 173.1



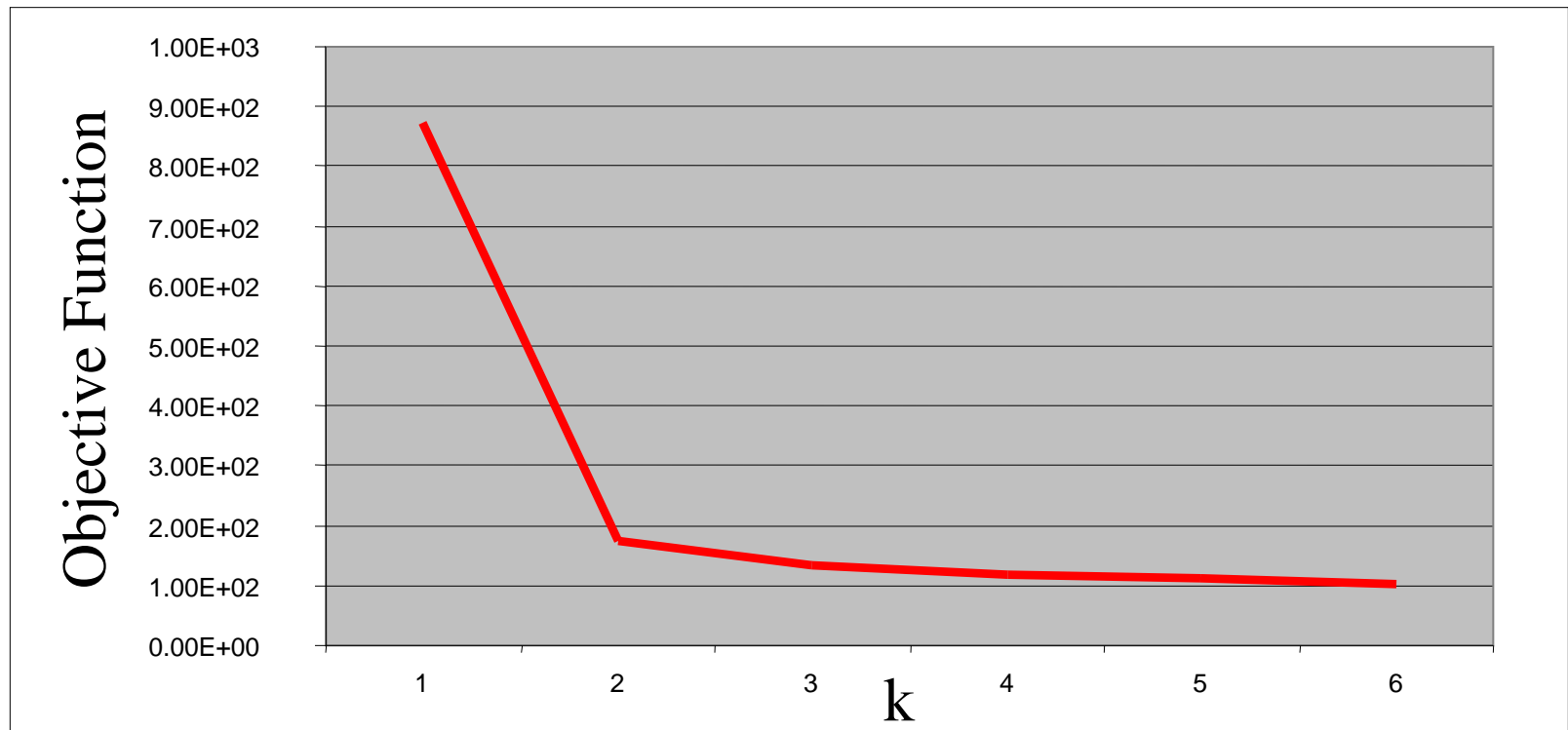
When  $k = 3$ , the objective function is 133.6





We can plot the objective function values for  $k$  equals 1 to 6...

The abrupt change at  $k = 2$ , is highly suggestive of two clusters in the data. This technique for determining the number of clusters is known as “knee finding” or “elbow finding”.

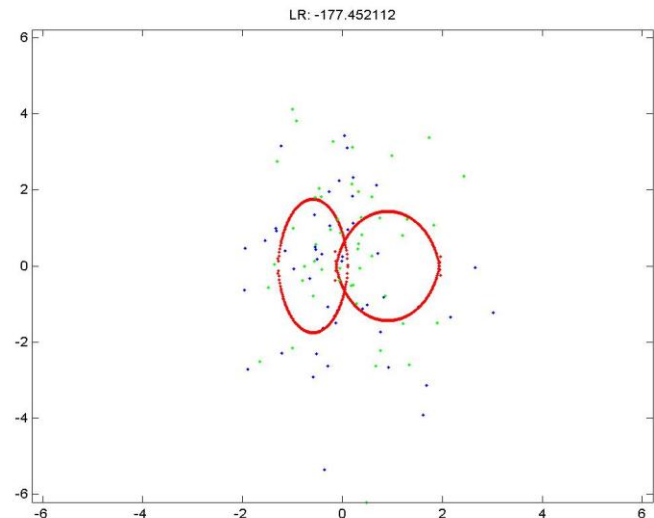
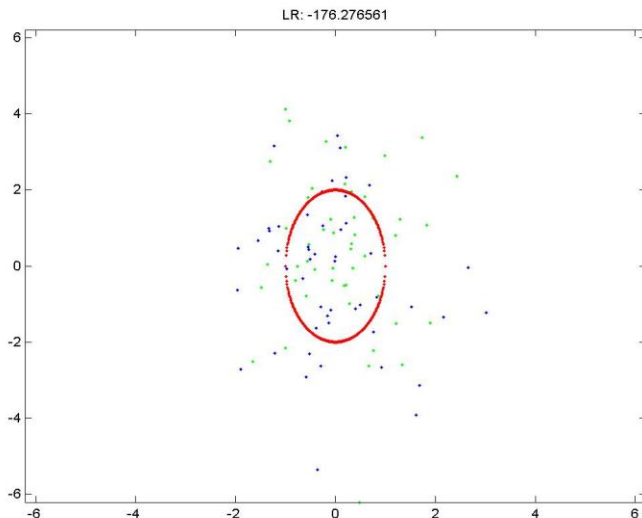


Note that the results are not always as clear cut as in this toy example

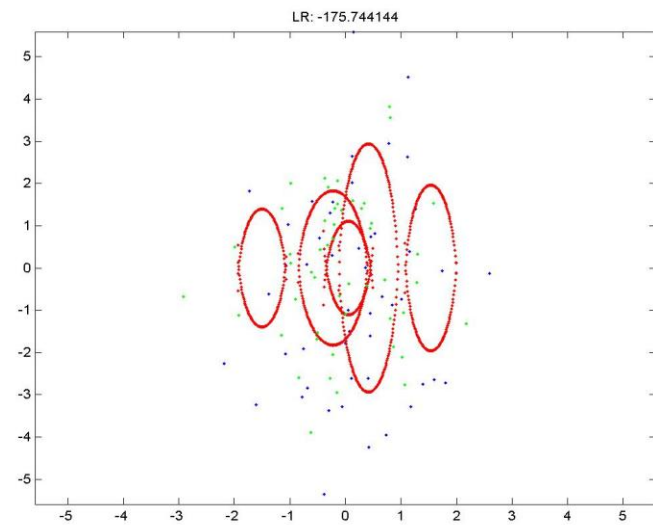
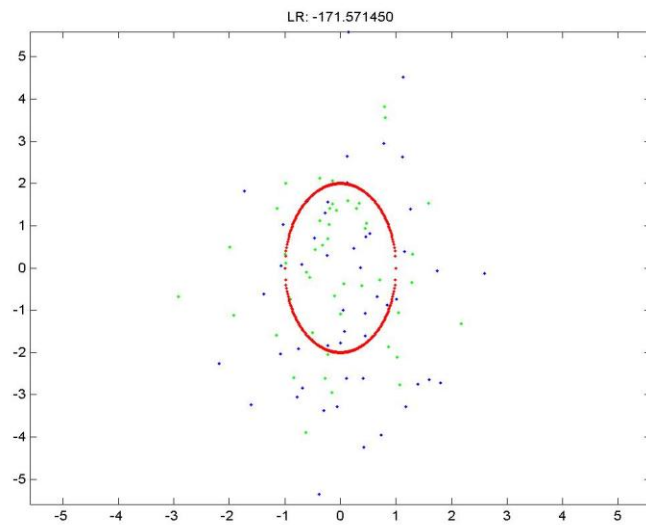
# Cross validation

- We can also use cross validation to determine the correct number of classes
- Recall that GMMs is a generative model. We can compute the likelihood of the left out data to determine which model (number of clusters) is more accurate

$$p(x_1 \cdots x_n \mid \theta) = \prod_{j=1}^n \left( \sum_{i=1}^k p(x_j \mid C = i) w_i \right)$$



# Cross validation



# What you should know

- Why is clustering useful
- What are the different types of clustering algorithms
- What are the assumptions we are making for each, and what can we get from them
- Unsolved issues: number of clusters, initialization, etc.