

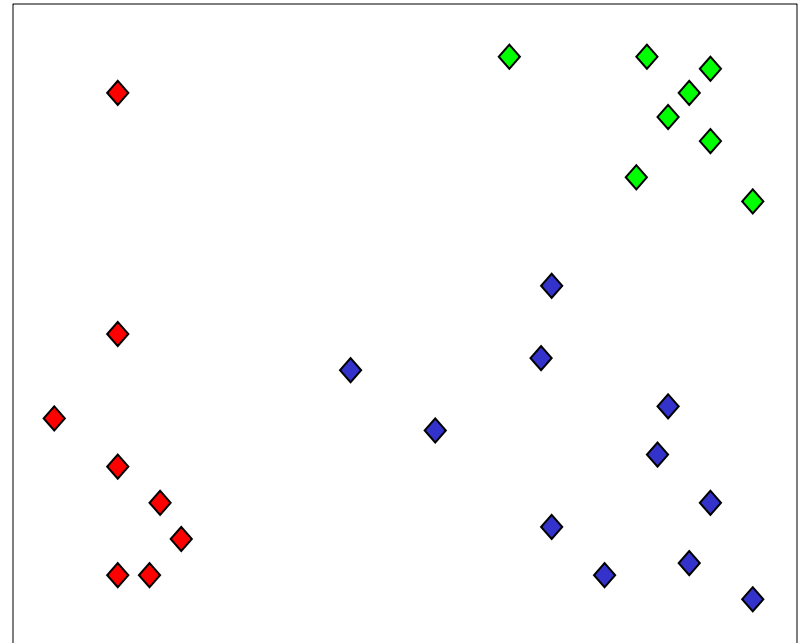
10701

Machine Learning

Hierarchical clustering and  
k-means

# 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)



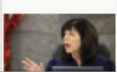
• Episode Guides (6)

• Simpson College (10)

more | all clusters

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

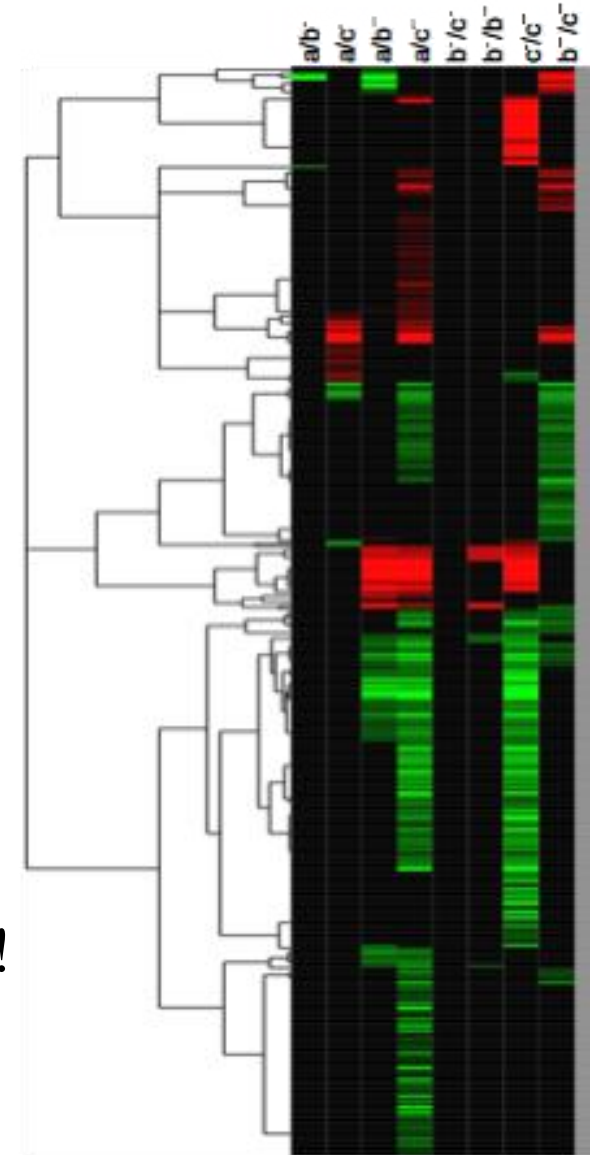
Search Results

- 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/entertainmentuscrimetrialssimpson](http://news.yahoo.com/s/afp/20080916/en_afp/entertainmentuscrimetrialssimpson) - [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 (11,591) paper in PNAS!



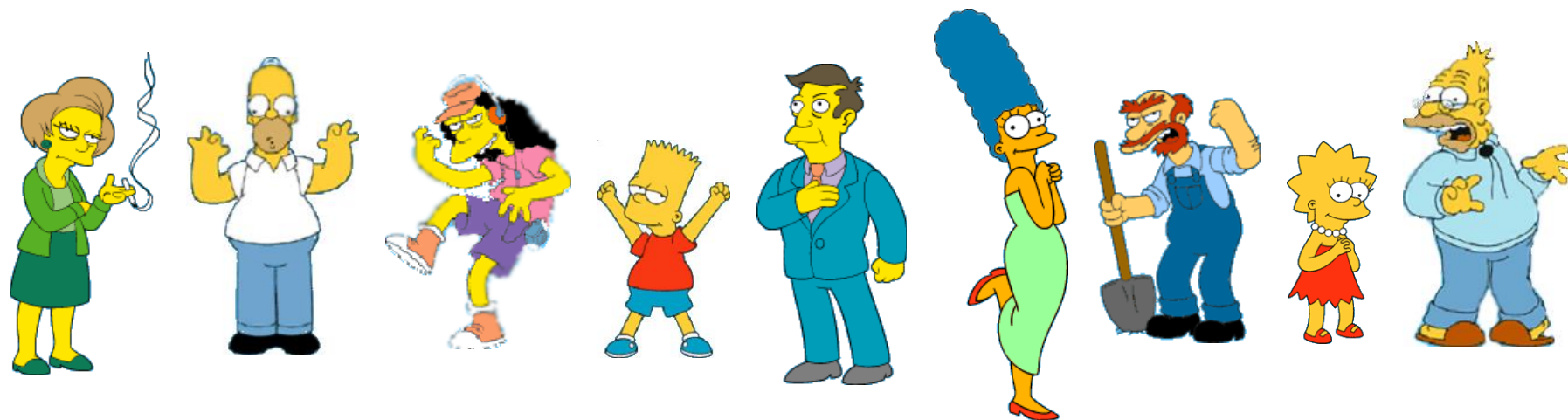
# Why clustering?

- Organizing data into clusters provides information about the internal structure of the data
  - Ex. Clusty and clustering genes above
- Sometimes the partitioning is the goal
  - Ex. Image segmentation
- Knowledge discovery in data
  - Ex. Underlying rules, reoccurring patterns, topics, etc.

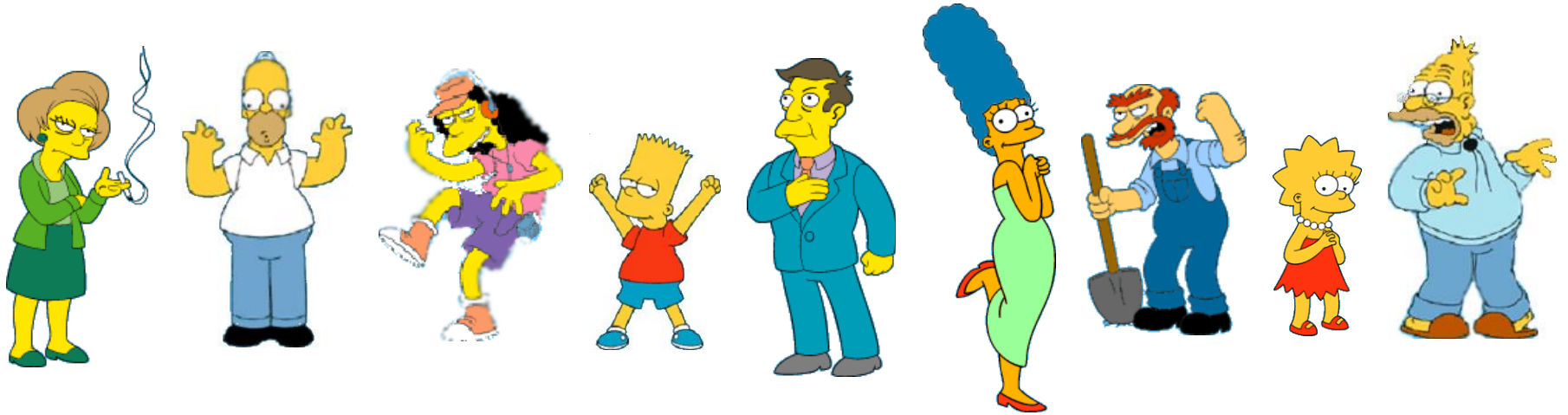
# Unsupervised learning

- 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

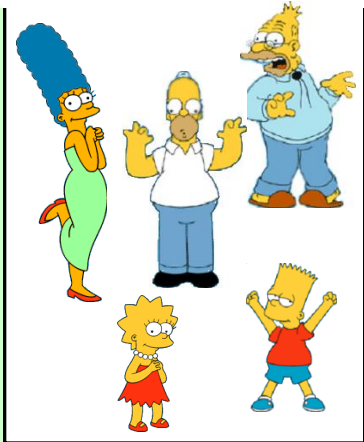
# What is a natural grouping among these objects?



# What is a natural grouping among these objects?



## Clustering is subjective



Simpson's Family



School Employees



Females



Males



# What is Similarity?

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

**Webster's Dictionary**



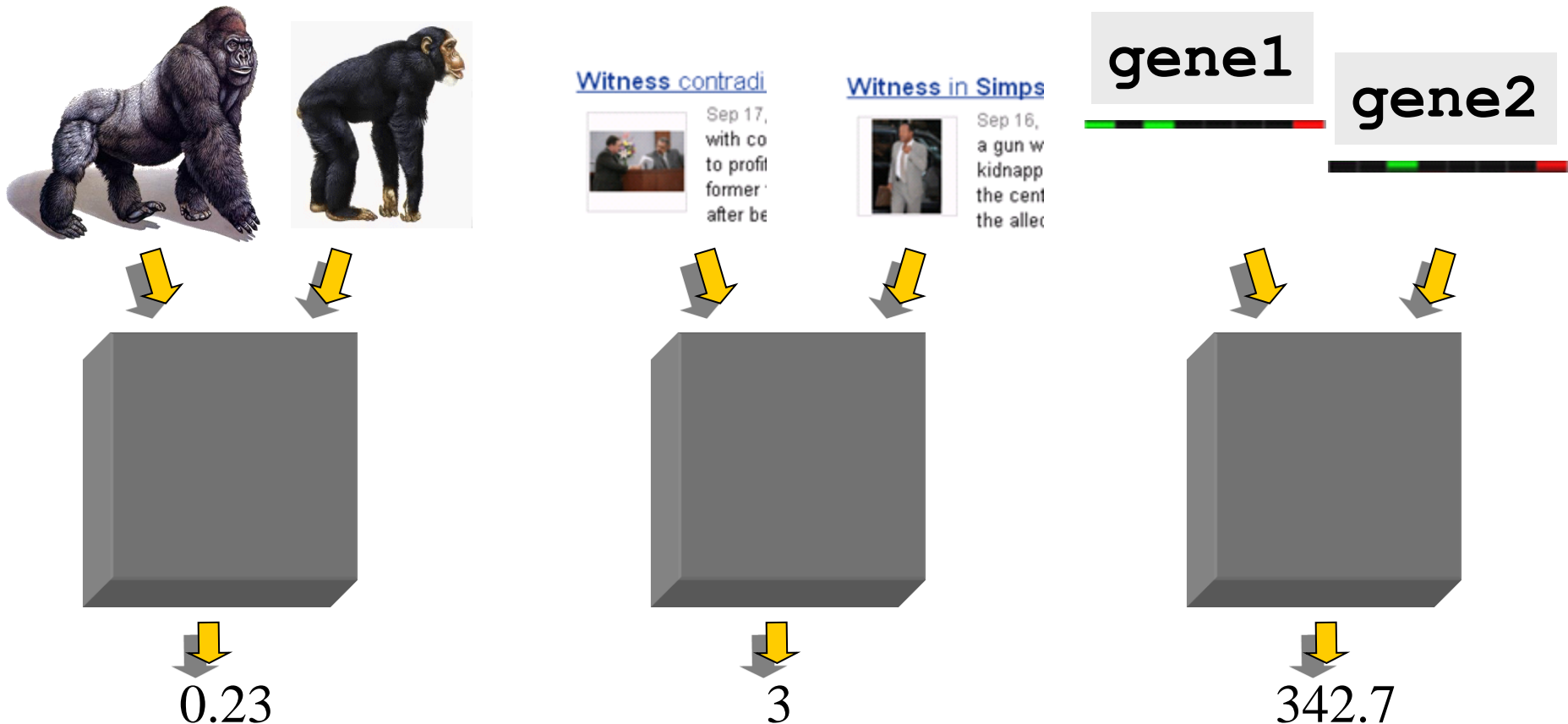
Similarity is hard to define, but...

*“We know it when we see it”*

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

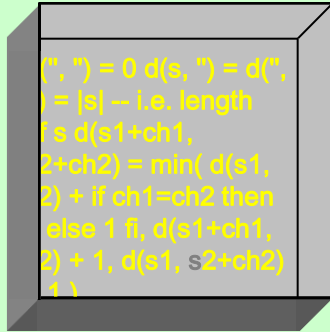
# Defining Distance Measures

**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

# 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

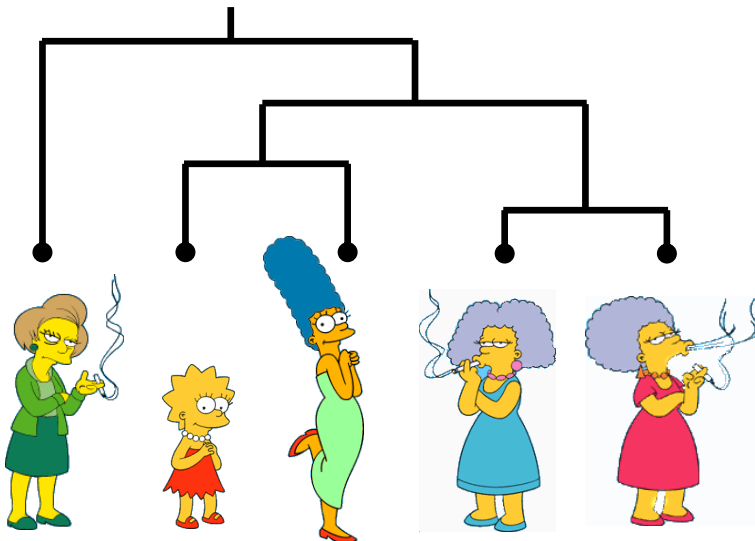
- Incorporation of user-specified constraints

# 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)

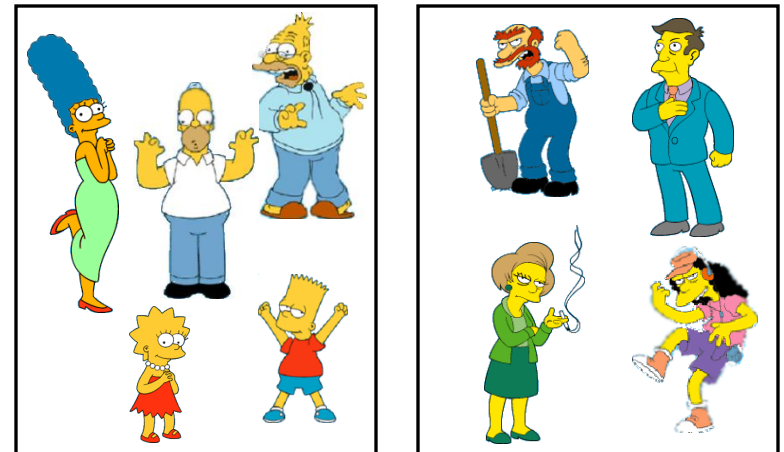
Bottom up or top down

**Hierarchical**



Top down

**Partitional**

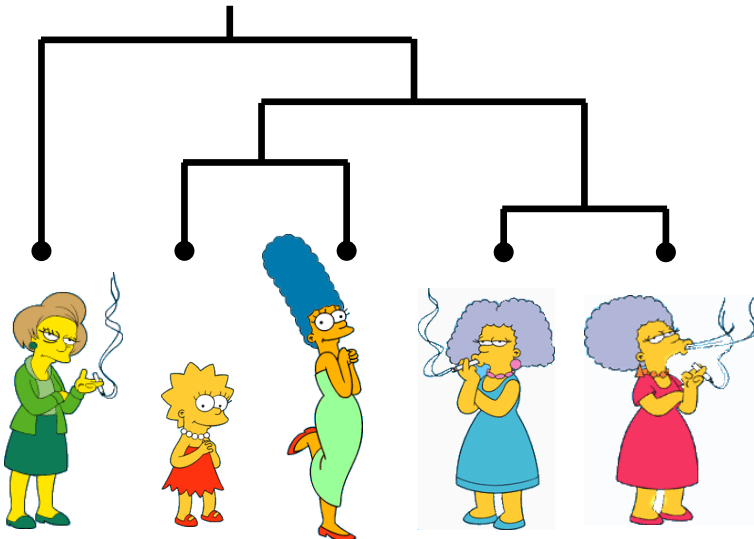


# (How-to) Hierarchical Clustering


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

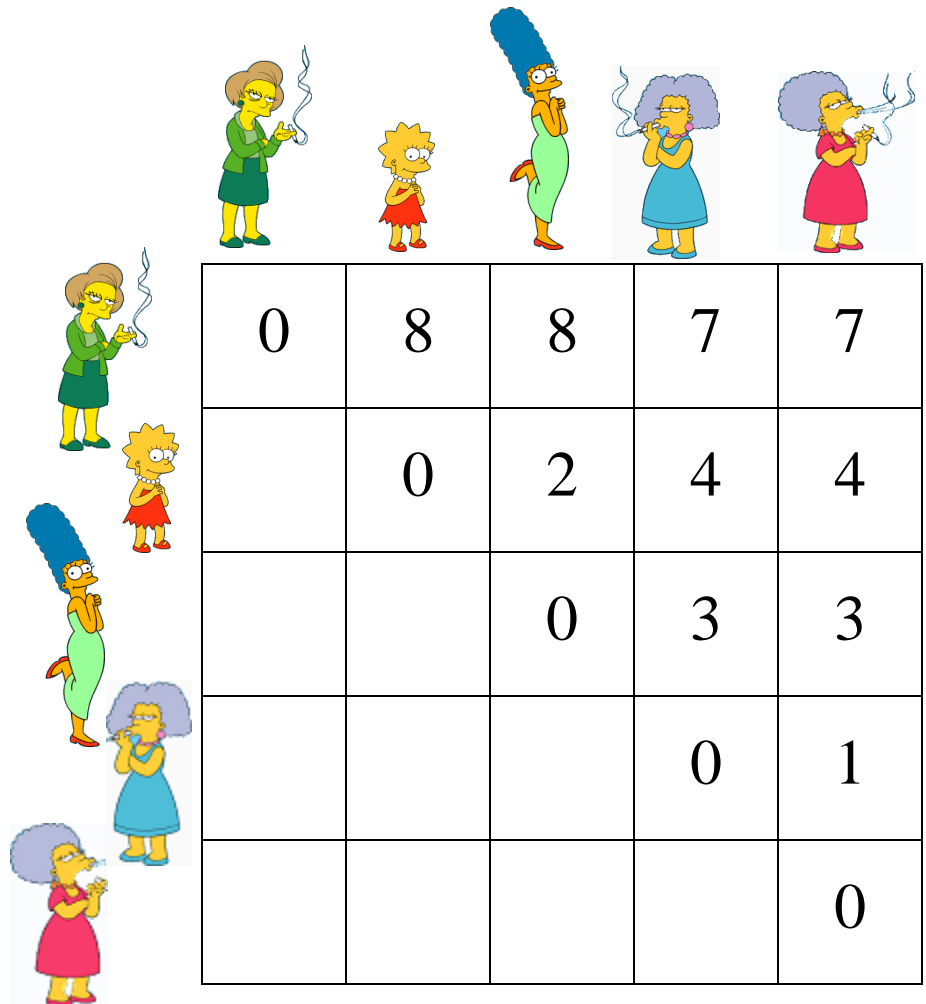
**Bottom-Up (agglomerative):** Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.













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


$$D(\text{Marge}, \text{Lisa}) = 8$$


$$D(\text{Barbara}, \text{Edna}) = 1$$

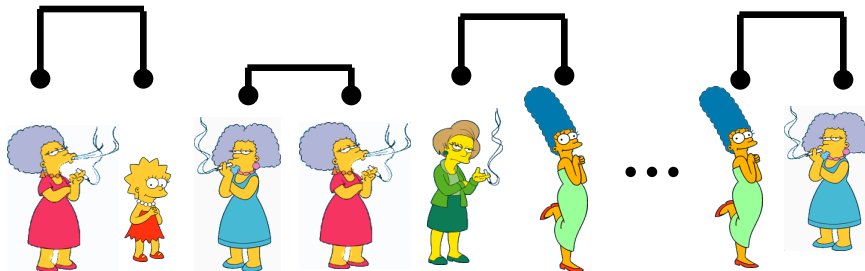


					
	0	8	8	7	7
		0	2	4	4
			0	3	3
				0	1
					0

# 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...



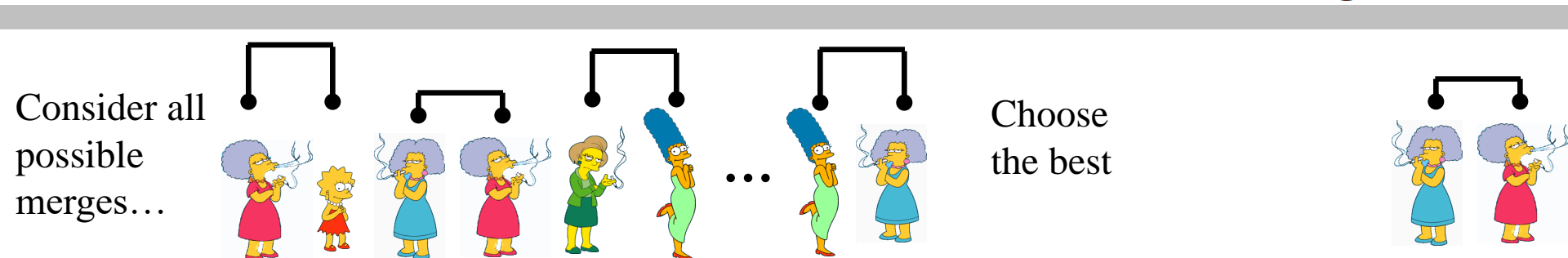
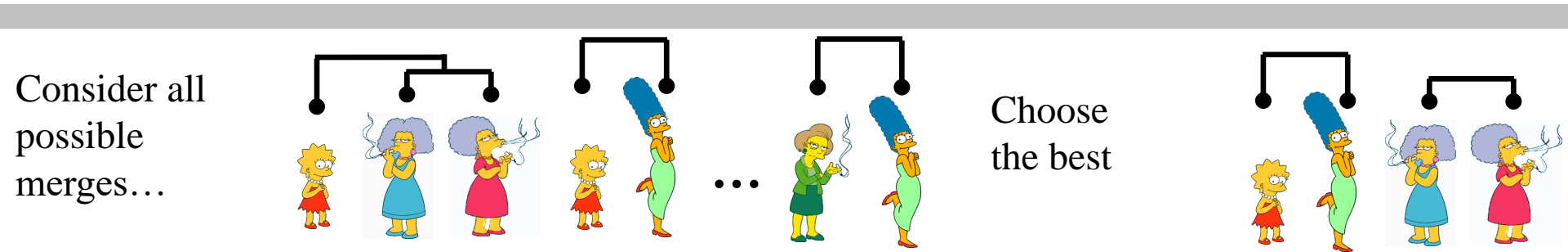
Choose the best





# Bottom-Up (agglomerative):

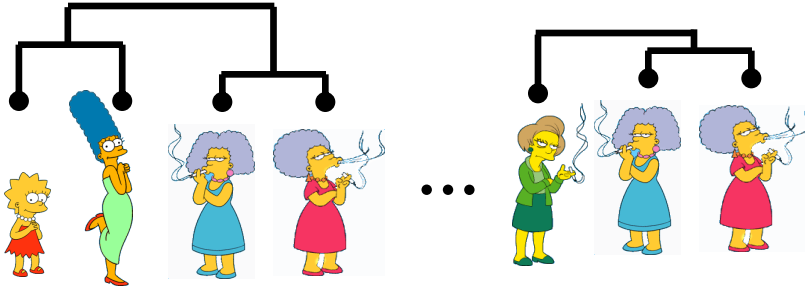
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



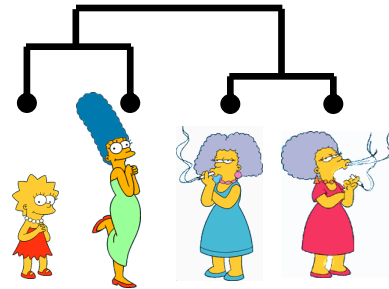
# Bottom-Up (agglomerative):

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.

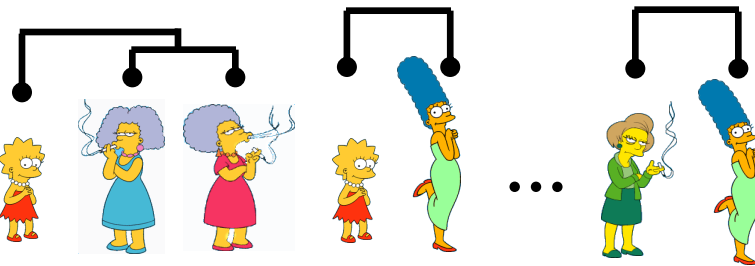
Consider all possible merges...



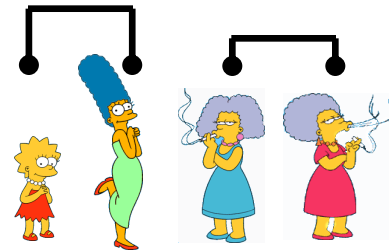
Choose the best



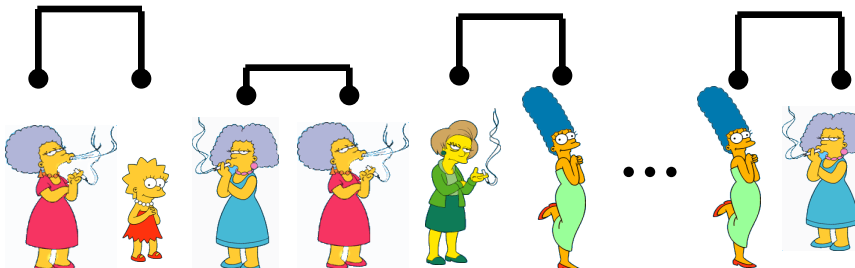
Consider all possible merges...



Choose the best



Consider all possible merges...

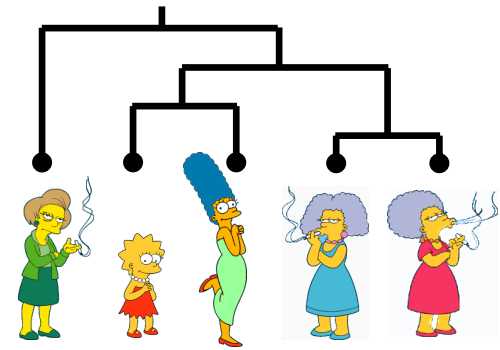


Choose the best

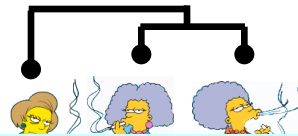
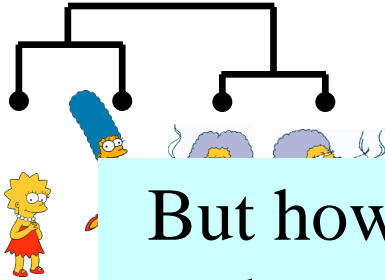


# Bottom-Up (agglomerative):

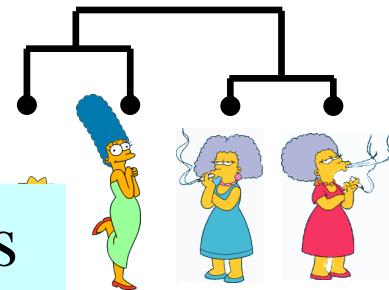
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



Consider all possible merges...

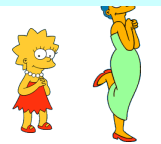
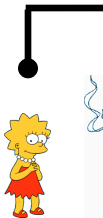


Choose



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

Consider all possible merges...



...



the best



Consider all possible merges...



...

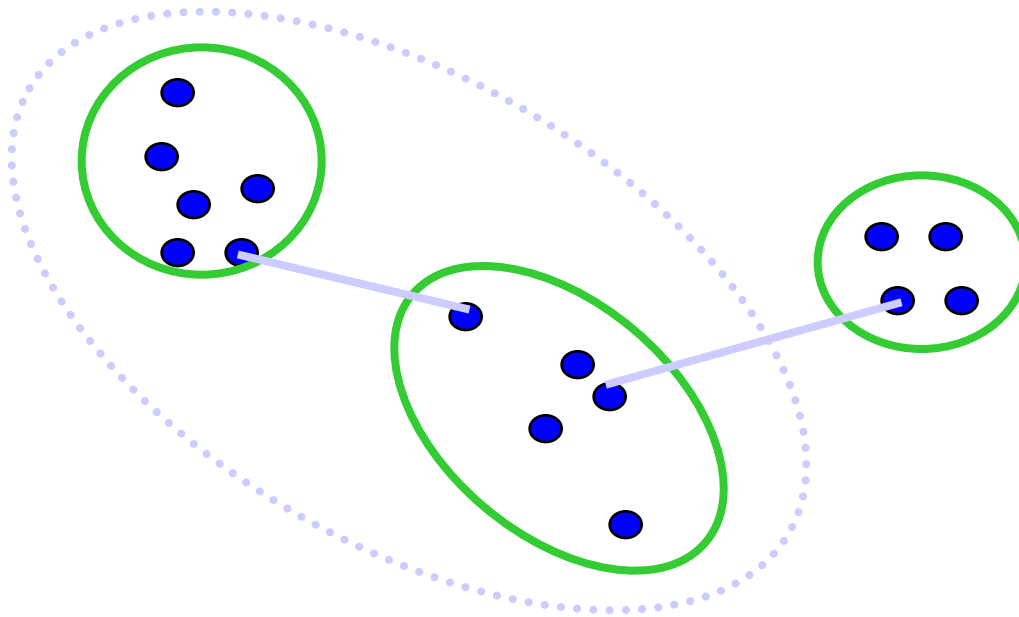


Choose the best



# Computing distance between clusters: Single Link

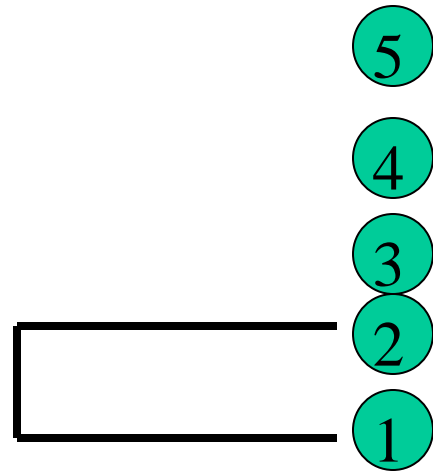
- cluster distance = distance of two **closest** members in each class



- Potentially long and skinny clusters

# Example: single link

$$\begin{array}{c} 1 \quad 2 \quad 3 \quad 4 \quad 5 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array}$$



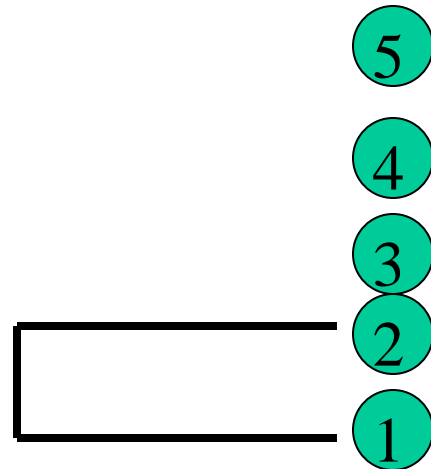
# Example: single link

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

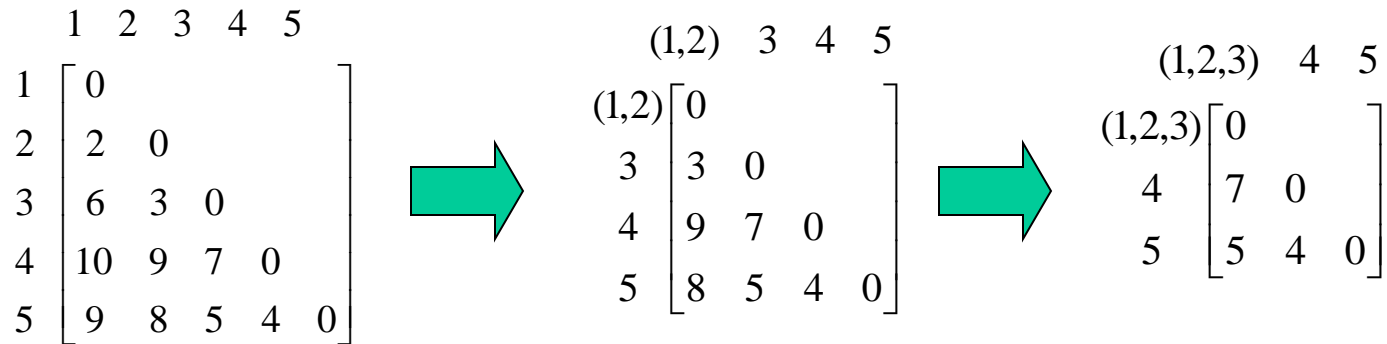
$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6, 3\} = 3$$

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

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

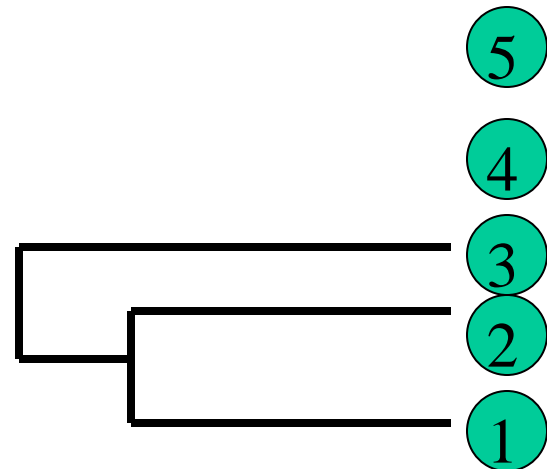


# Example: single link

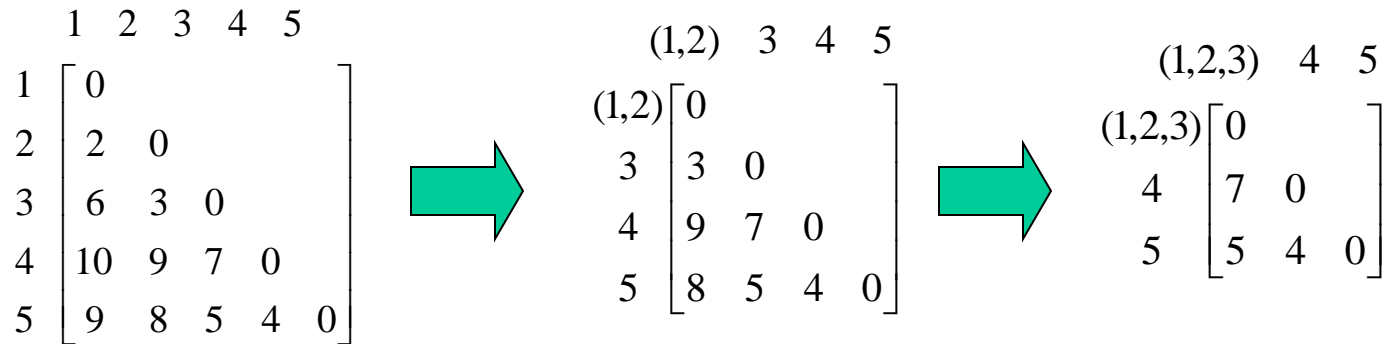


$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9, 7\} = 7$$

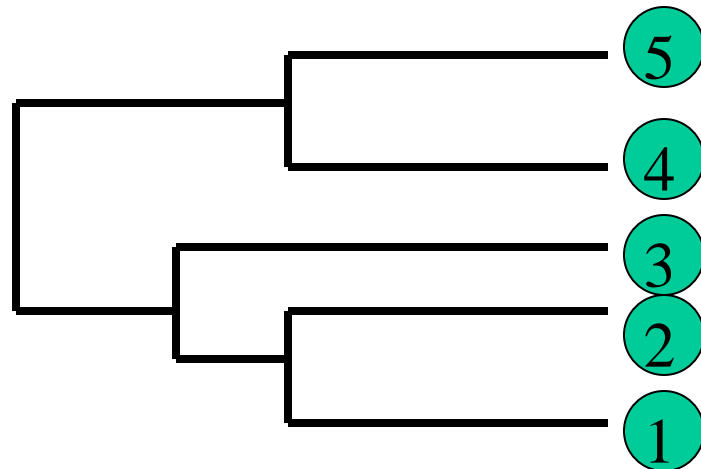
$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8, 5\} = 5$$



# Example: single link



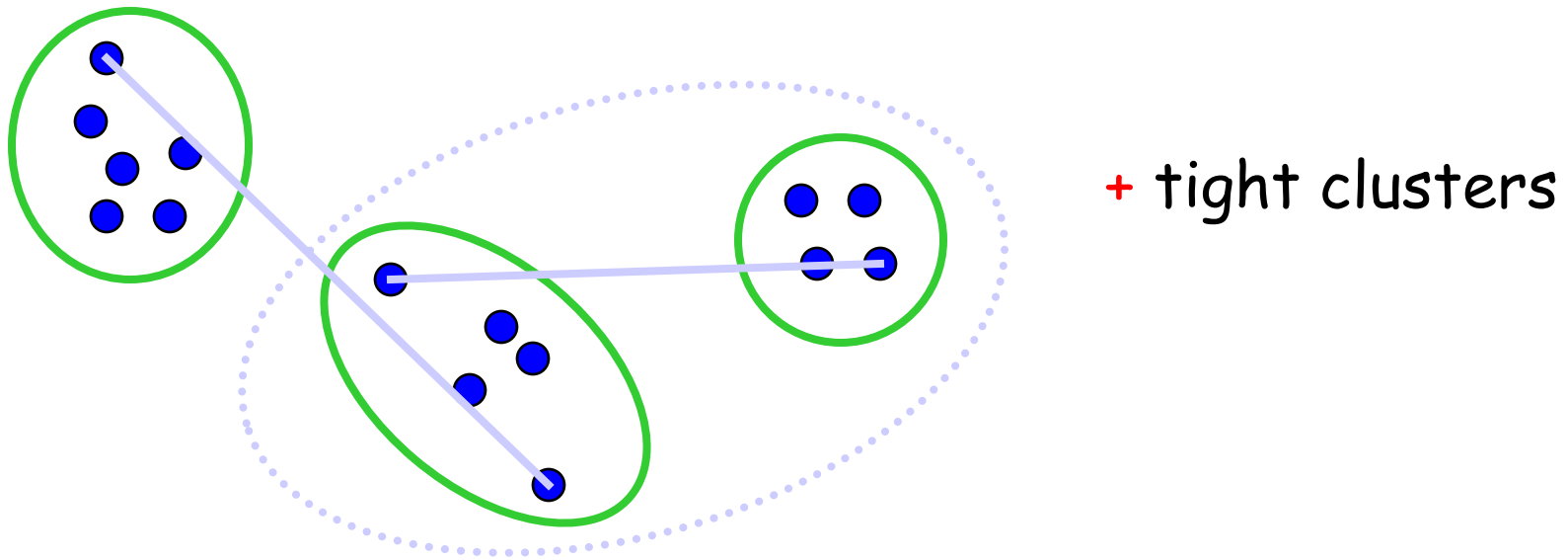
$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$





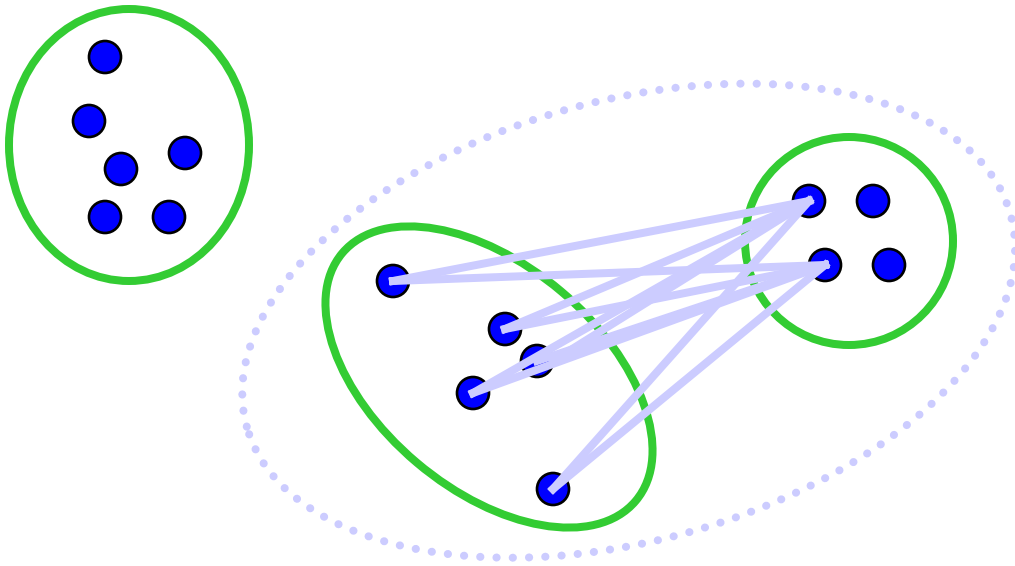
# Computing distance between clusters: : Complete Link

- cluster distance = distance of two farthest members



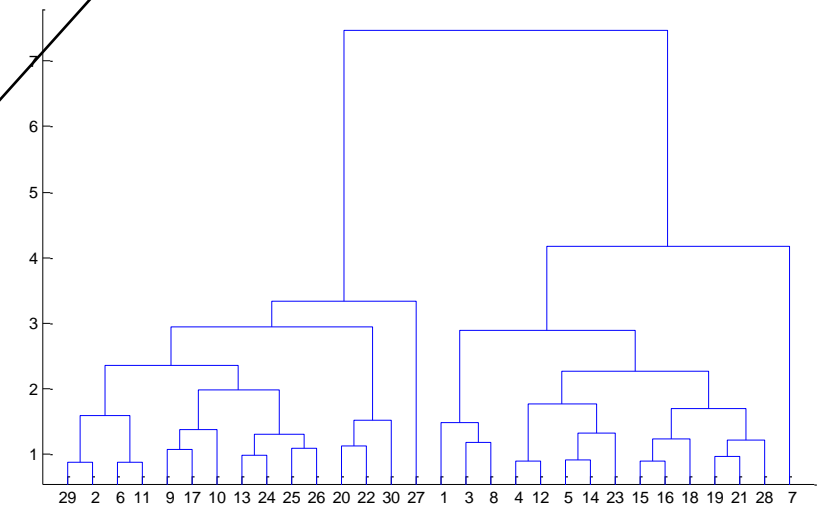
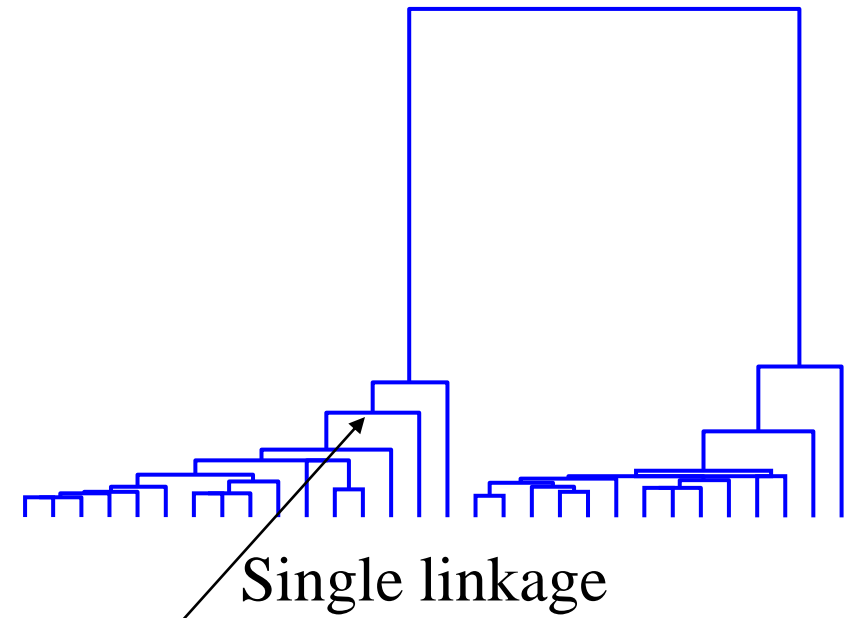
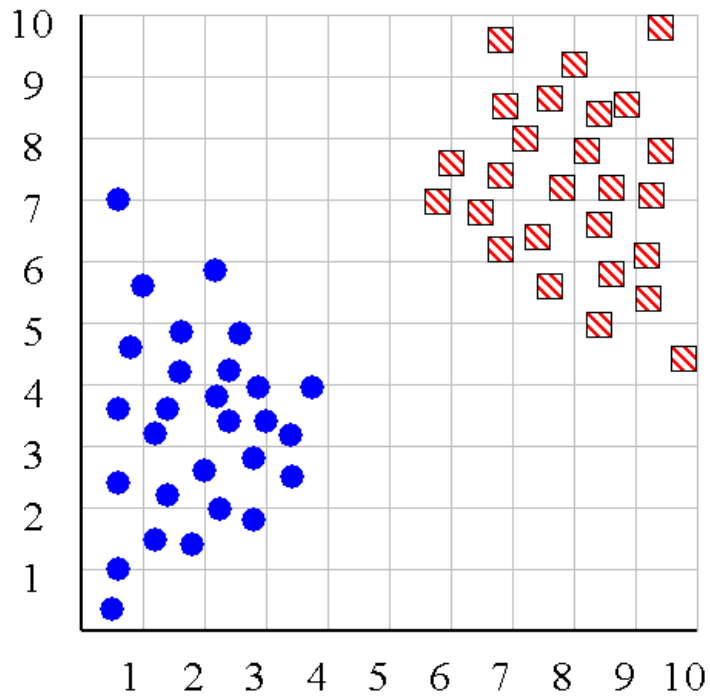
# Computing distance between clusters: Average Link

- cluster distance = average distance of all pairs



**the most widely  
used measure**

**Robust against  
noise**



Height represents  
distance between objects  
/ clusters

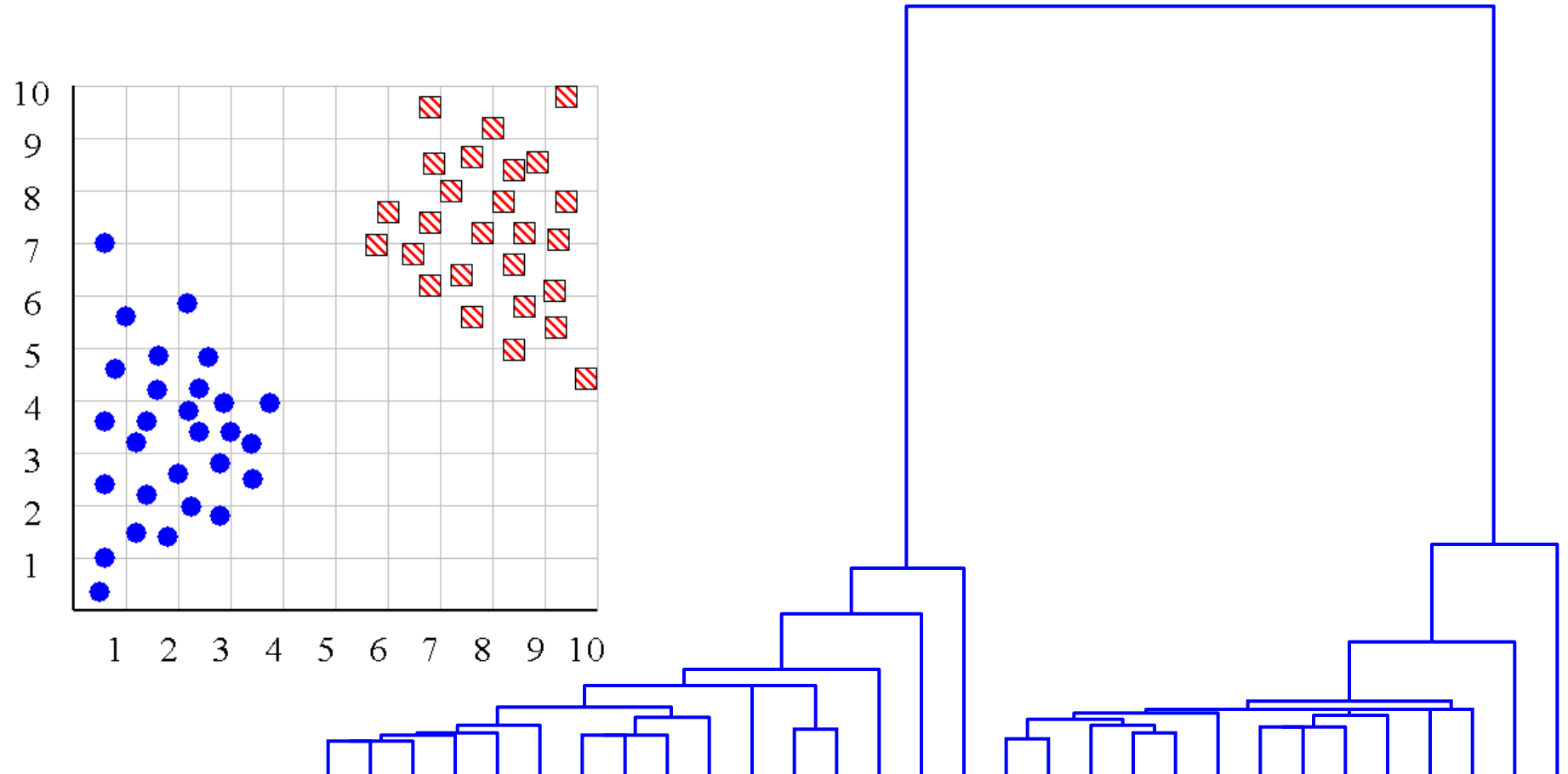
Average linkage

# 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.

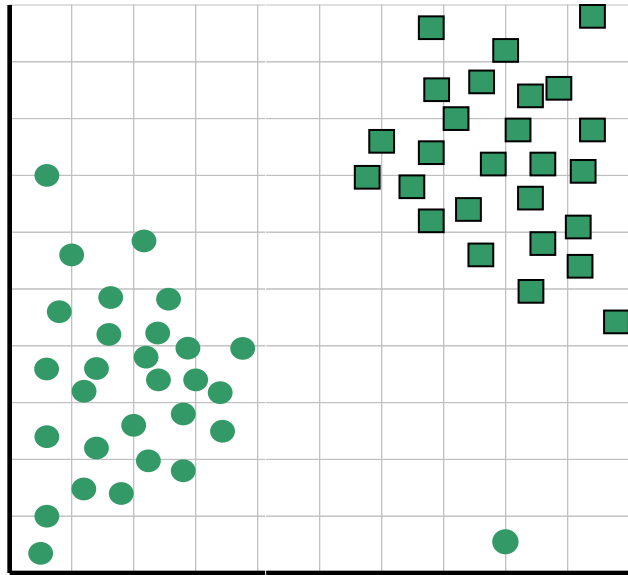
# But what are the clusters?

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

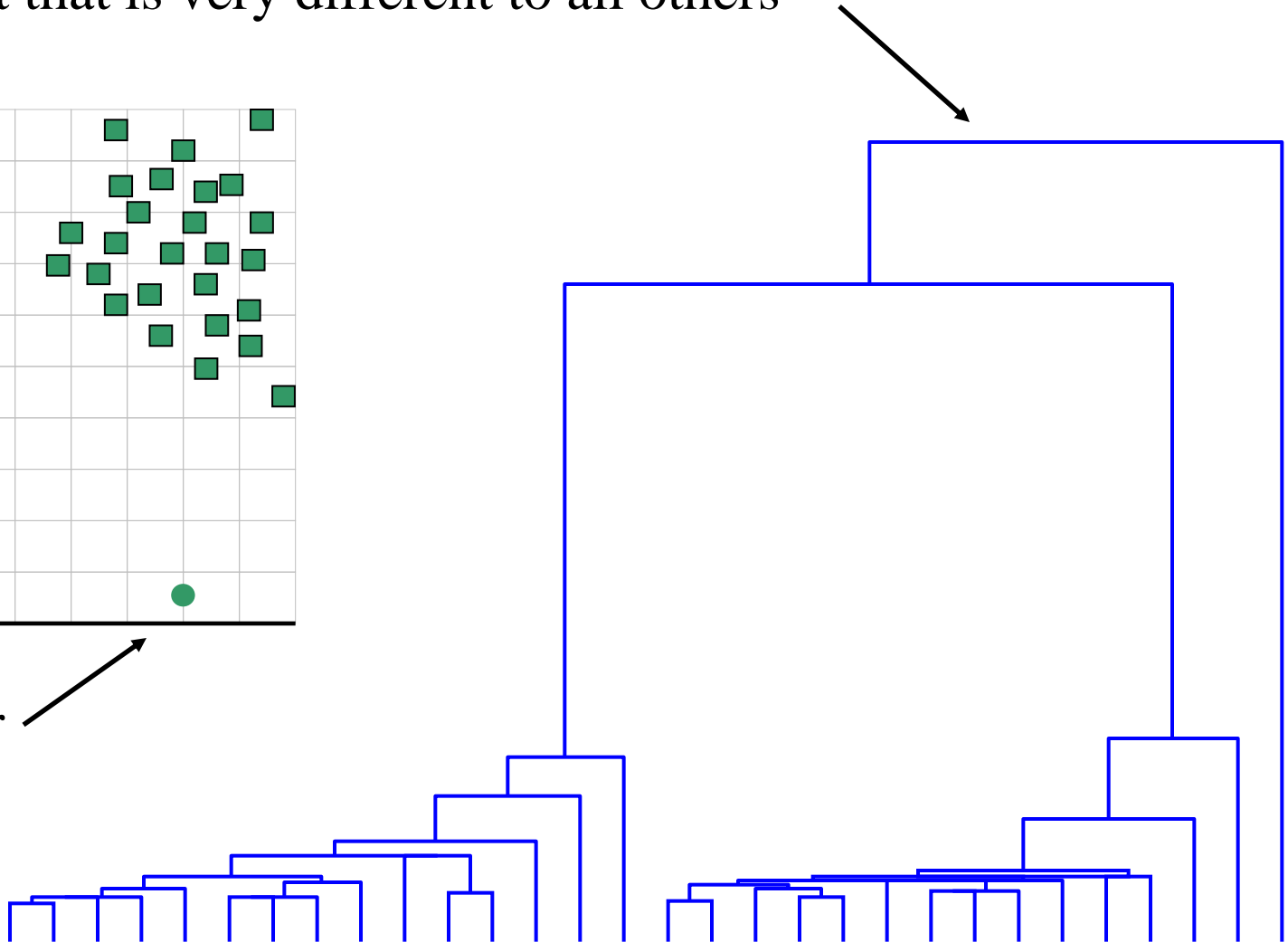


# One potential use of a dendrogram is to detect outliers

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

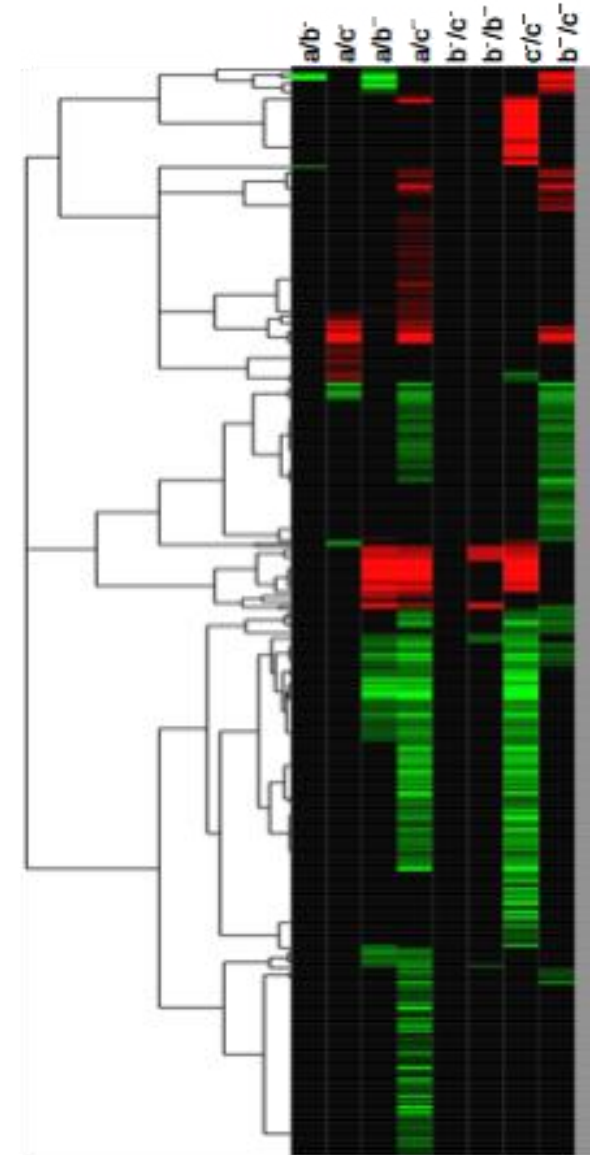


Outlier



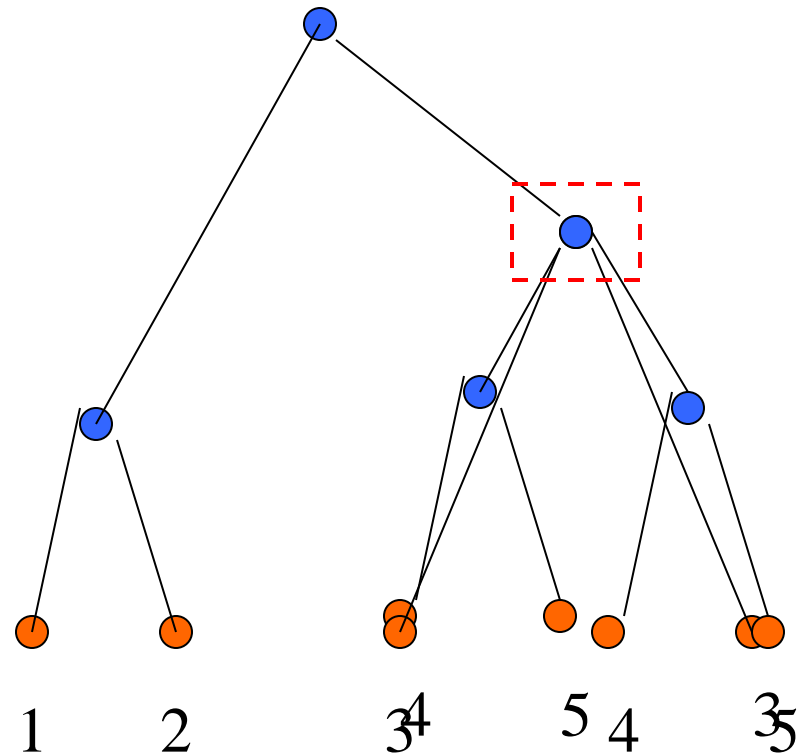
# Example: clustering genes

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



# Clustering tree

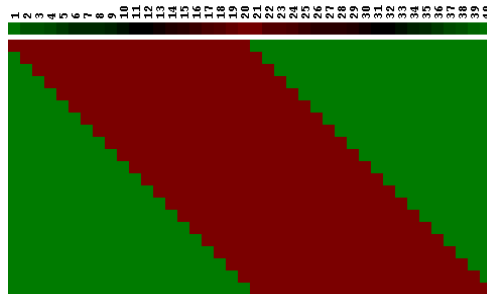
- For  $n$  leaves there are  $n-1$  internal nodes
- Each flip in an internal node creates a new linear ordering
- There are  $2^{n-1}$  possible linear ordering of the leafs of the tree



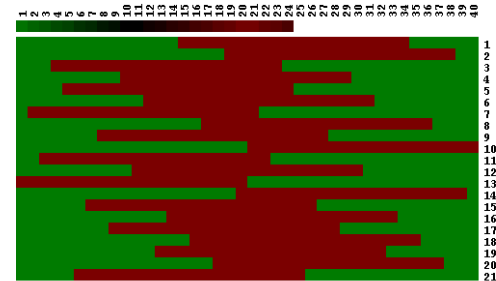


# Importance of the Ordering

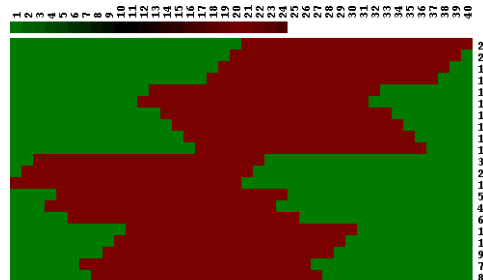
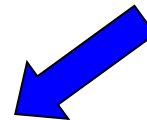
- Samples that are adjacent in the linear ordering are often hypothesized to be related.
- Ordering can help determine relationships between samples and clusters in time series data analysis.



Initial structure

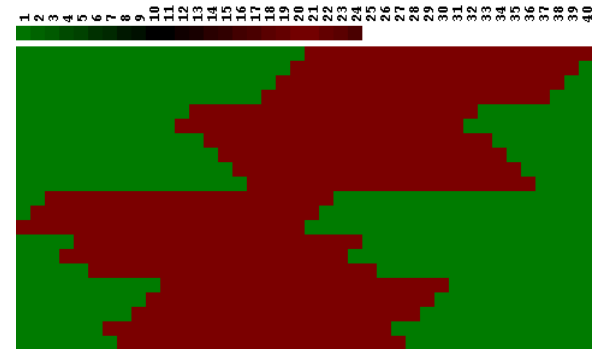


Permuted

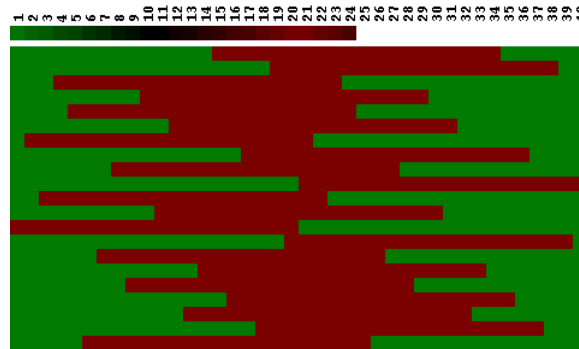


Hierarchical clustering

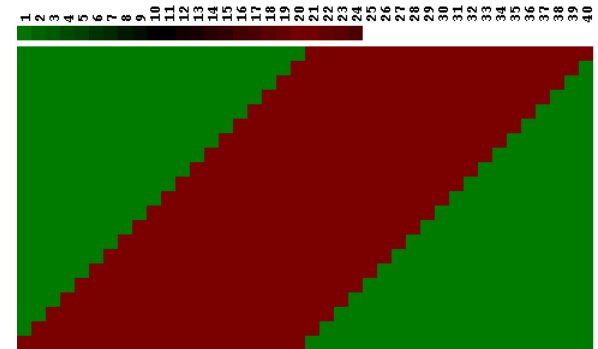
# Results – Synthetic Data



Hierarchical clustering



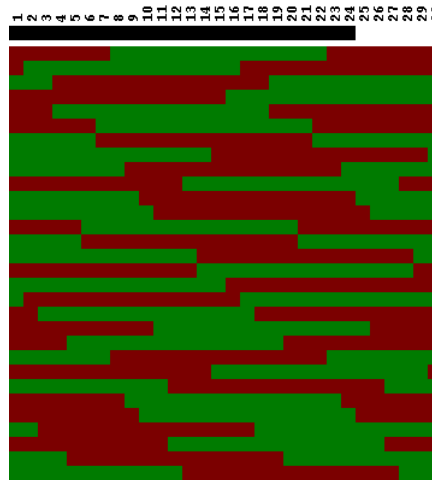
Input



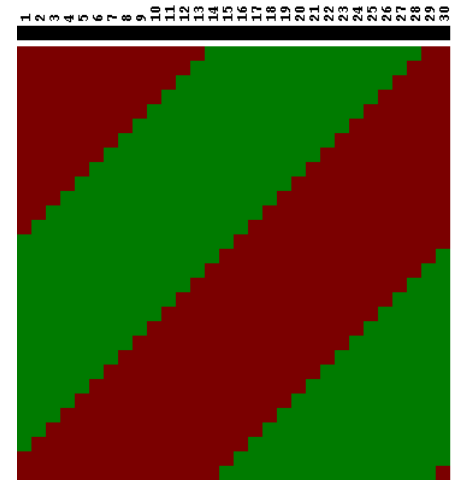
Optimal ordering



Hierarchical clustering



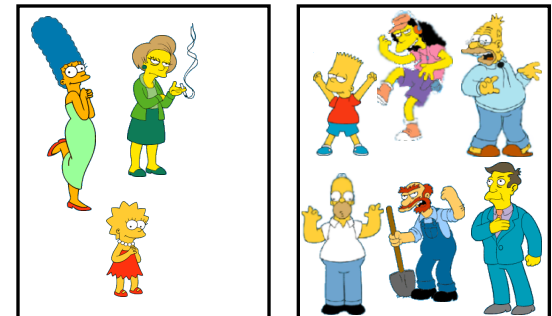
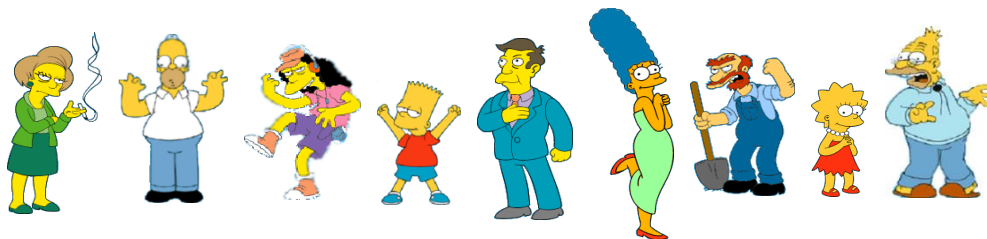
Input



Optimal ordering

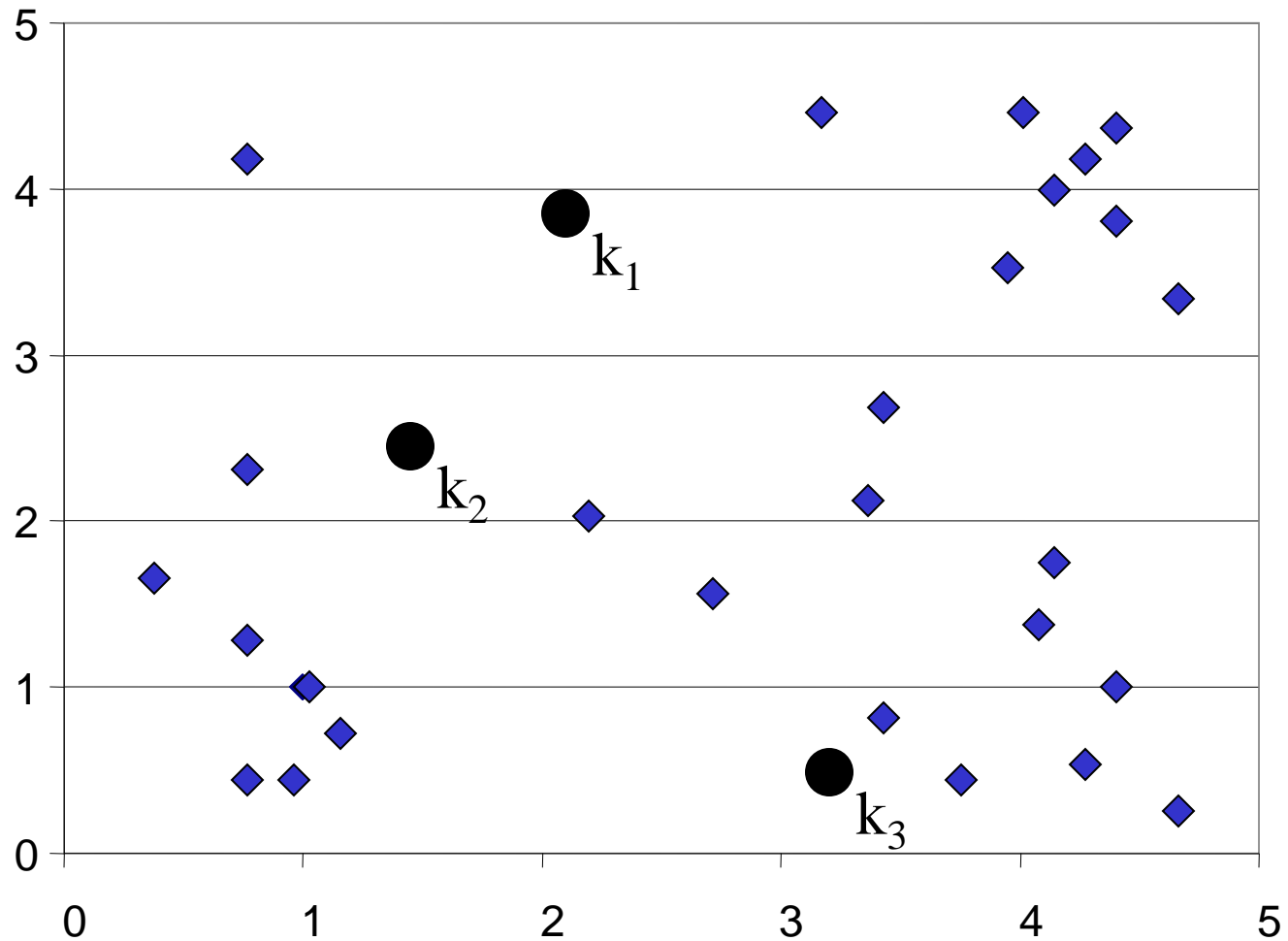
# Partitional Clustering

- 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: Initialization

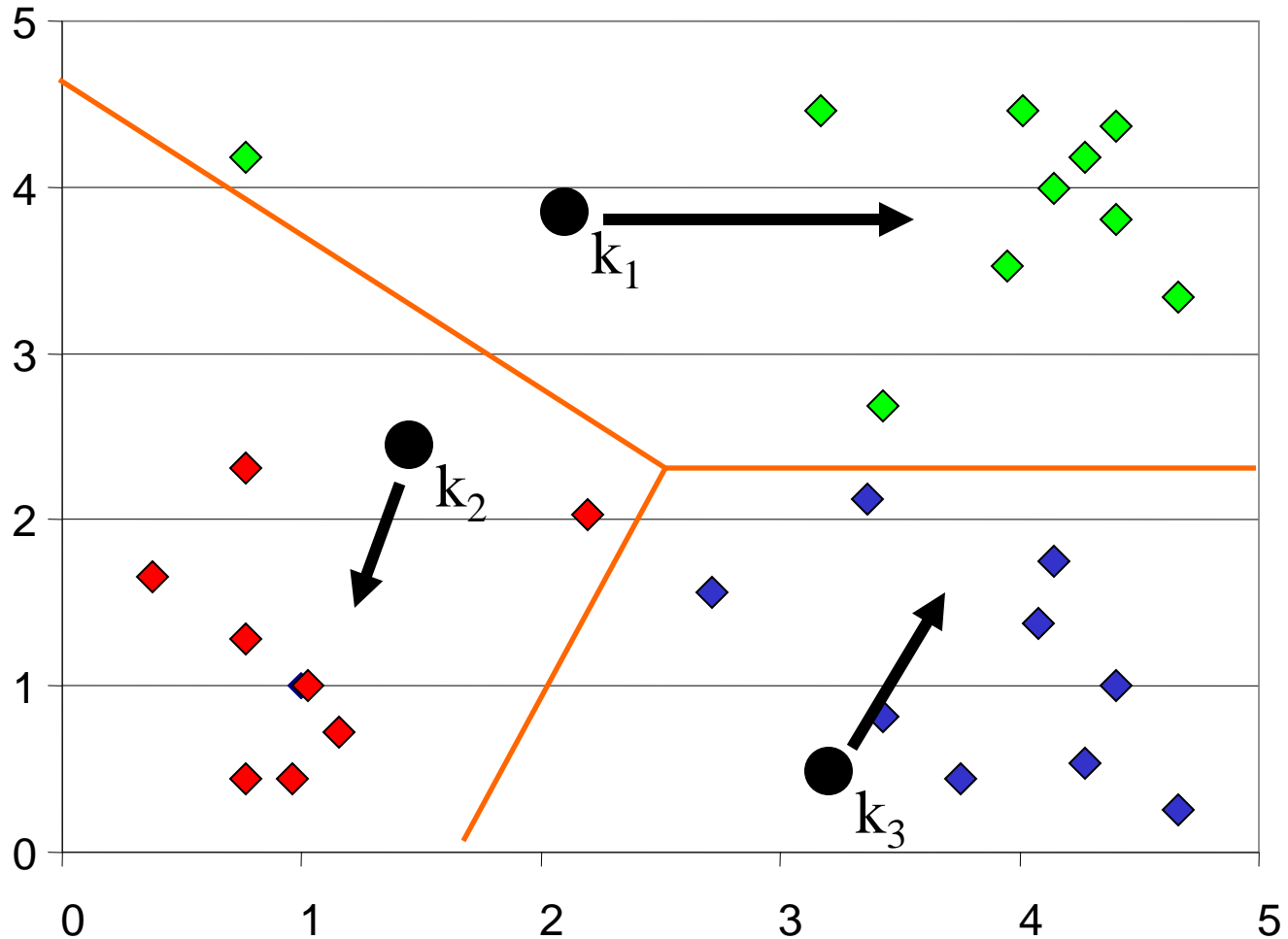
Decide  $K$ , and initialize  $K$  centers (randomly)



# K-means Clustering: Iteration 1

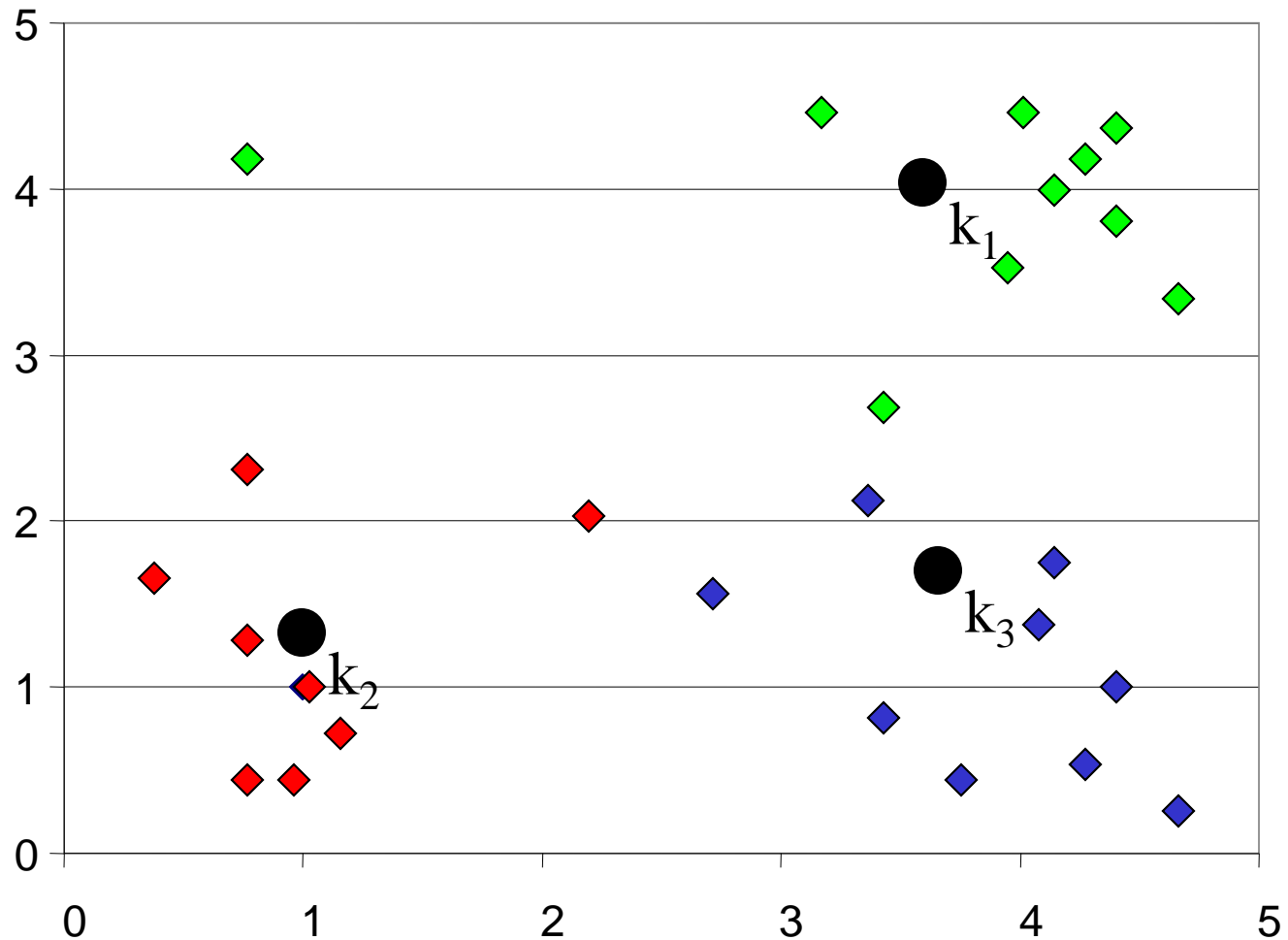
Assign all objects to the nearest center.

Move a center to the mean of its members.



# K-means Clustering: Iteration 2

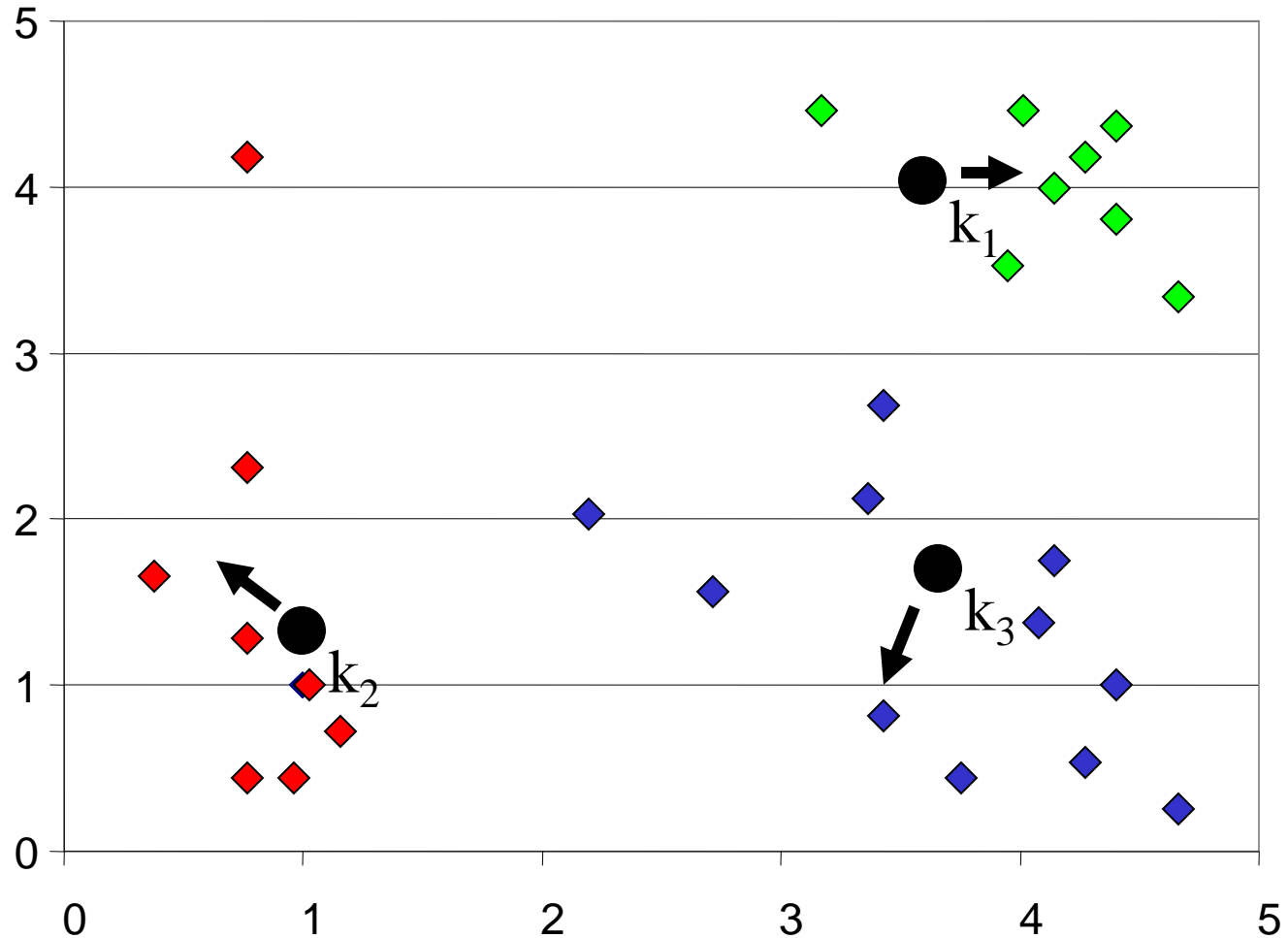
After moving centers, re-assign the objects...



# K-means Clustering: Iteration 2

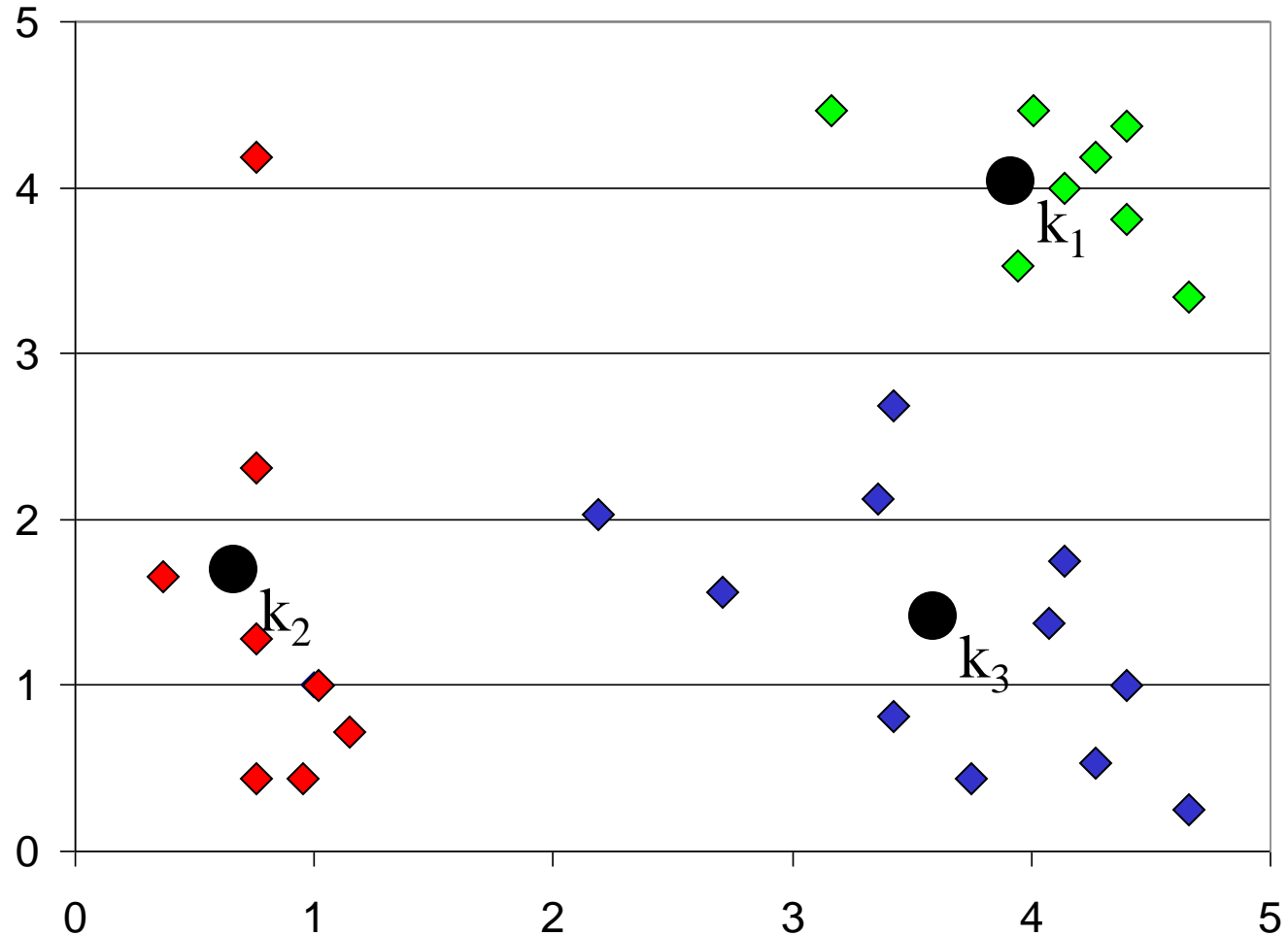
After moving centers, re-assign the objects to nearest centers.

Move a center to the mean of its new members.



# K-means Clustering: Finished!

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





# Algorithm *k-means*

1. Decide on a value for  $K$ , the number of clusters.
2. Initialize the  $K$  cluster centers (randomly, if necessary).
3. Decide the class memberships of the  $N$  objects by assigning them to the nearest cluster center.
4. Re-estimate the  $K$  cluster centers, by assuming the memberships found above are correct.
5. Repeat 3 and 4 until none of the  $N$  objects changed membership in the last iteration.

# Algorithm *k-means*

1. Decide on a value for  $K$ , the number of clusters (if necessary).
2. Initialize the  $K$  cluster centers (e.g., randomly or by hand).  
Use one of the distance / similarity functions we discussed earlier
3. Decide the class memberships of the  $N$  objects by assigning them to the nearest cluster center.
4. Re-estimate the  $K$  cluster centers, by assuming the memberships found above are correct.  
Average / median of class members
5. Repeat 3 and 4 until none of the  $N$  objects changed membership in the last iteration

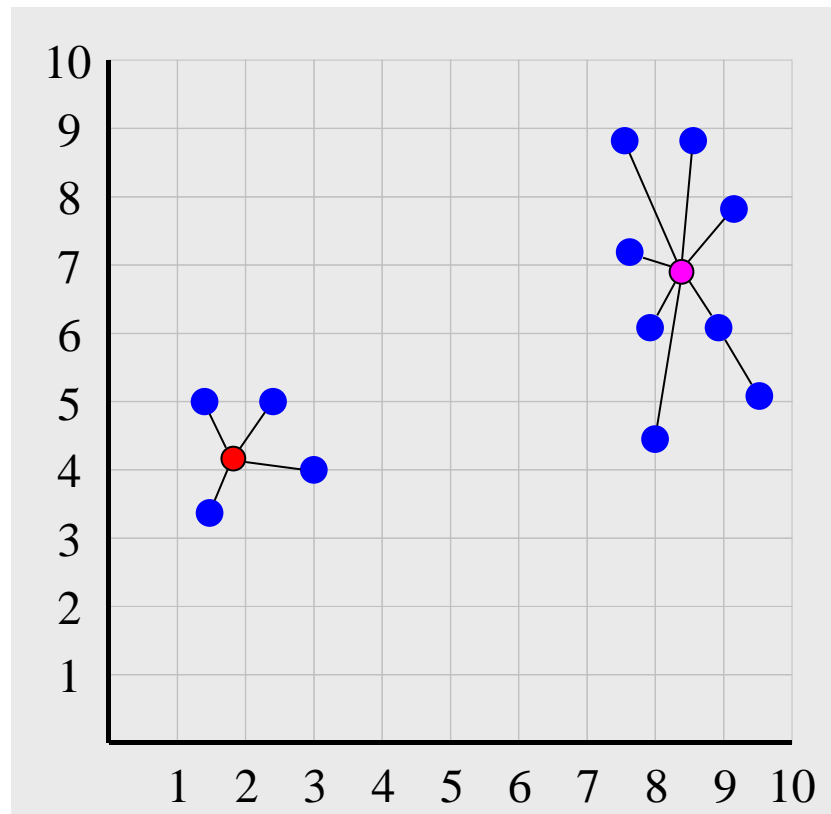
# Why K-means Works

- What is a good partition?
- High intra-cluster similarity
- K-means optimizes
  - the average distance to members of the same cluster

$$\sum_{k=1}^K \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|x_{ki} - x_{kj}\|^2$$

- which is twice the total distance to centers, also called squared error

$$se = \sum_{k=1}^K \sum_{i=1}^{n_k} \|x_{ki} - \mu_k\|^2$$



# Summary: *K-Means*

- Strength
  - Simple, easy to implement and debug
  - Intuitive objective function: optimizes intra-cluster similarity
  - *Relatively efficient*:  $O(tkn)$ , where  $n$  is # objects,  $k$  is # clusters, and  $t$  is # iterations. Normally,  $k, t \ll n$ .
- Weakness
  - Applicable only when *mean* is defined, what about categorical data?
  - Often terminates at a *local optimum*. Initialization is important.
  - Need to specify  $K$ , the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*
- Summary
  - Assign members based on current centers
  - Re-estimate centers based on current assignment

# 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 (to be discussed next lecture): number of clusters, initialization