

Lecture 12: Clustering

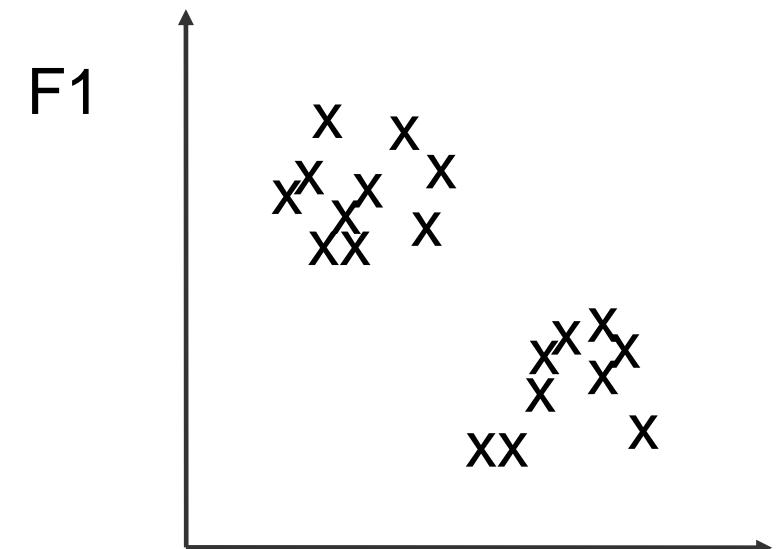
Instructor: Sergei V. Kalinin

This and that

1. Please let me know if you have submitted HW 1, but do not see the rate on CANVAS
2. Please add the non-programming part of HW2 to the submitted Colabs and reshare
3. As a reminder, please submit all assignments as Colabs to sergei2vk@gmail.com
4. In October, I will have some immutable travel. The plan is to have several lectures to be given by guest lecturers on specific use cases for ML in materials and imaging (there is uncertainty given reliability of late flights)

Clustering

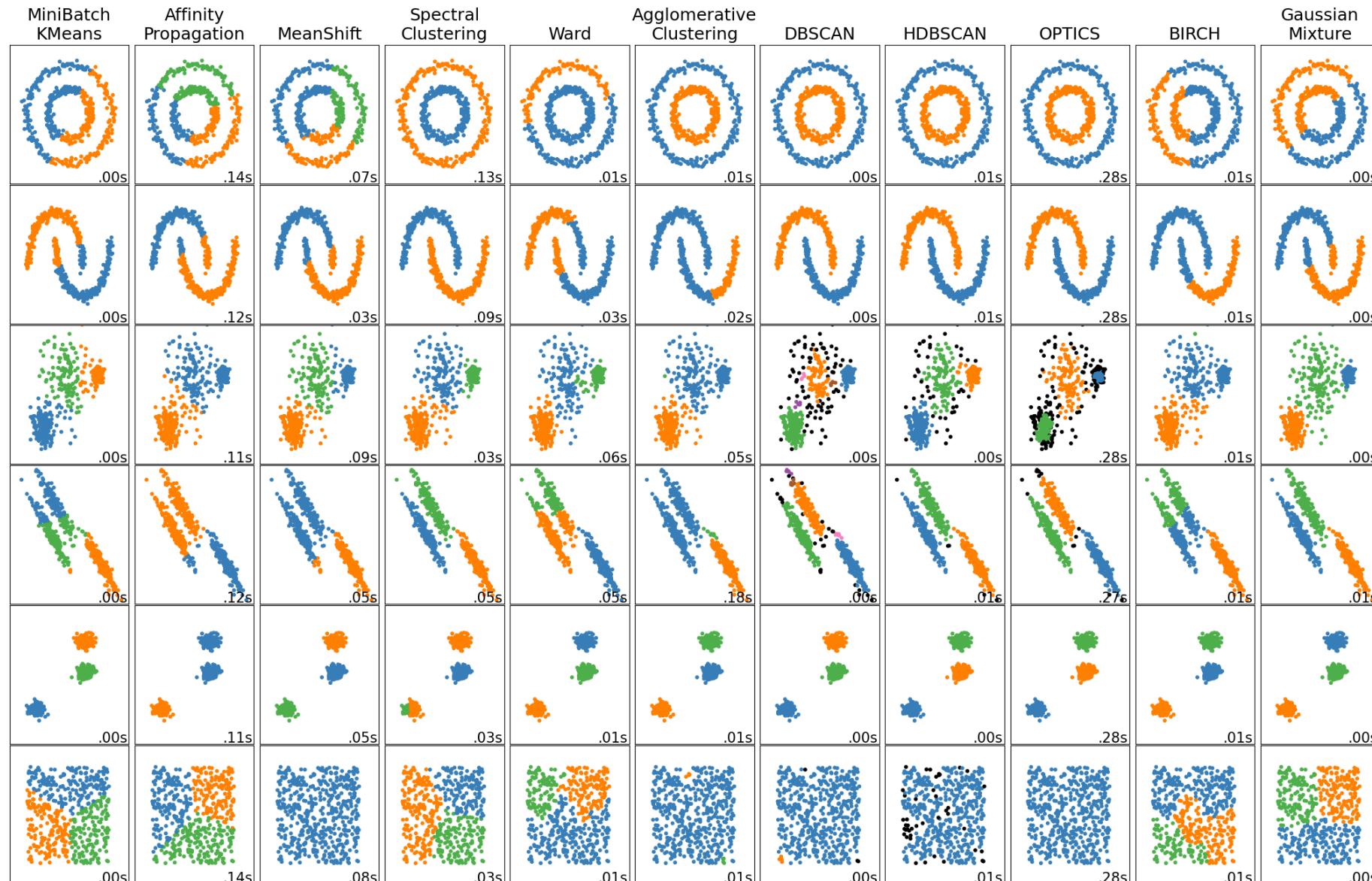
- The process of grouping a set of objects into classes of similar objects
 - Objects within a cluster should be similar.
 - Objects from different clusters should be dissimilar.
- The most common form of *unsupervised learning*
- Given a set of data points, each described by a set of attributes, find clusters such that:
 - Inter-cluster similarity is maximized
 - Intra-cluster similarity is minimized
- Requires the definition of a similarity measure



https://en.wikipedia.org/wiki/Cluster_analysis

CS276: Information Retrieval and Web Search Pandu Nayak and Prabhakar Raghavan

Clustering: good news and bad news



Clustering: Data Structures

- Hierarchical clustering
- K-means clustering
- Gaussian Mixture Models
- Density-based clustering
- Spectral clustering

Data matrix (two modes)

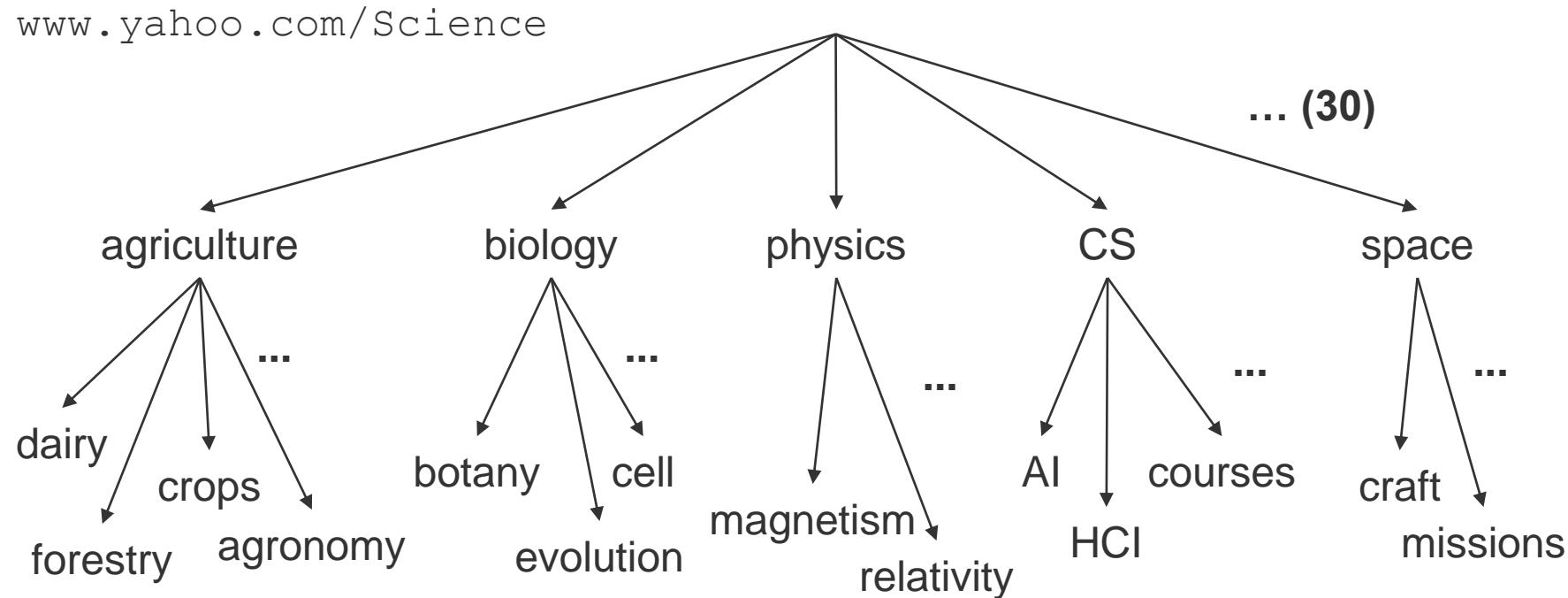
$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Dissimilarity matrix (one mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

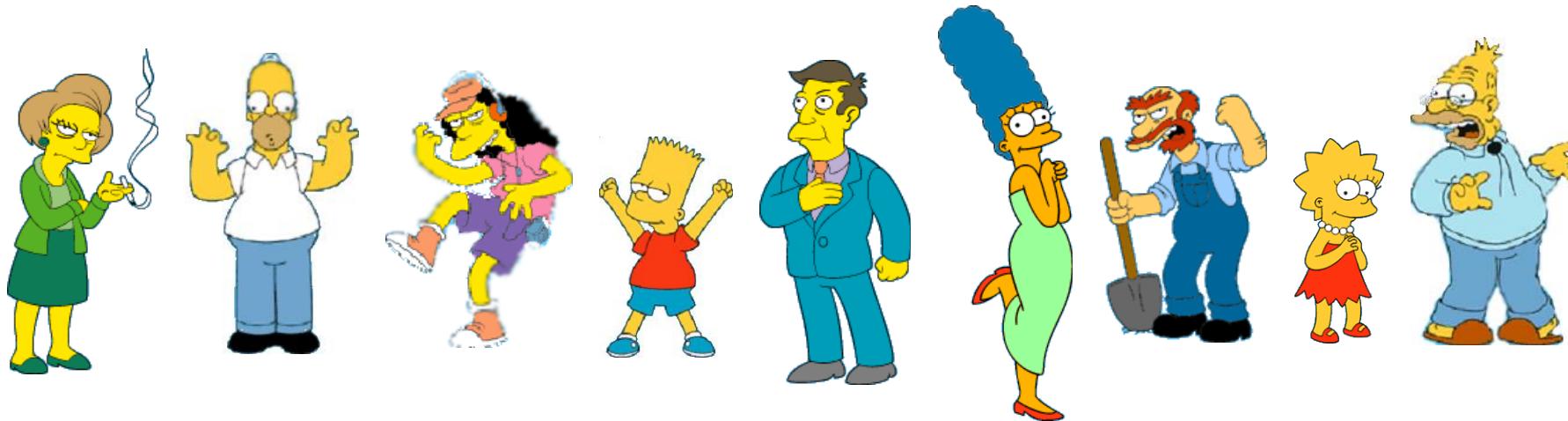
Clustering

Yahoo! Hierarchy *isn't* clustering but *is* the kind of output you want from clustering



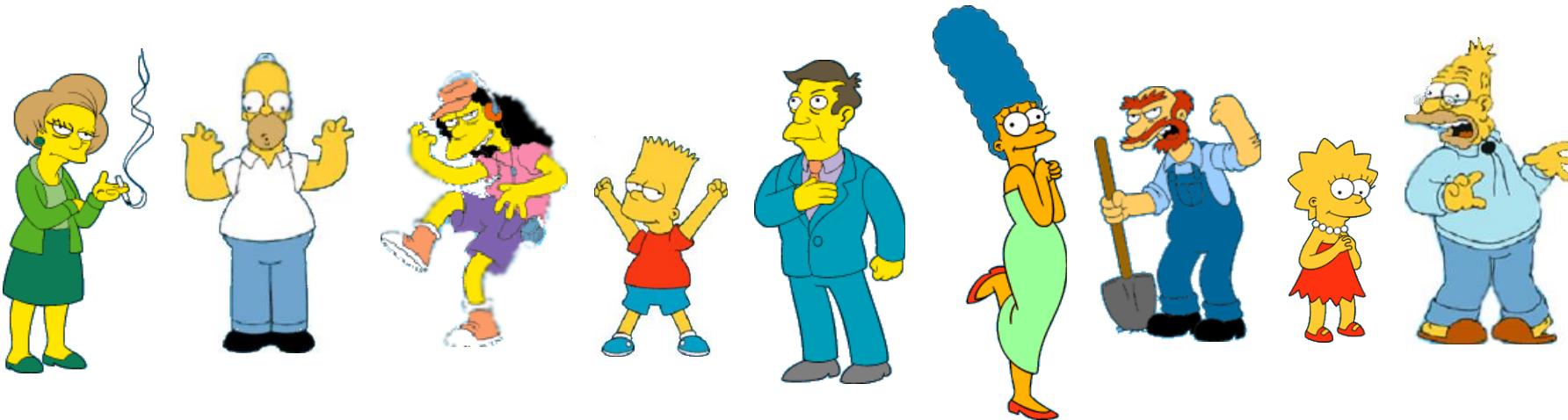
What is a natural grouping of these objects?

Slide from Eamonn Keogh, from lecture by Carla Brodley, Tufts University

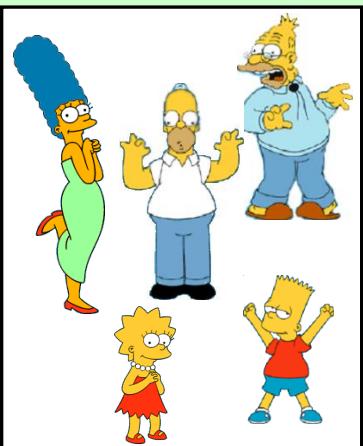


What is a natural grouping of these objects?

Slide from Eamonn Keogh, from lecture by Carla Brodley, Tufts University



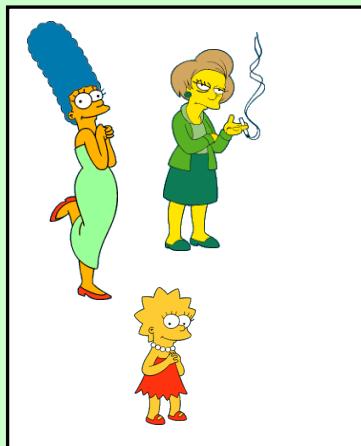
Clustering is subjective



Simpson's Family



School Employees



Females



Males

What is Similarity?

Slide from Eamonn Keogh, from lecture by Carla Brodley, Tufts University



Similarity is
hard to define,
but...
*"We know it
when we see it"*

Defining Distance Measures

Slide from Eamonn Keogh, from lecture by Carla Brodley, Tufts University

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 $D(O_1, O_2)$



Peter Piotr



0.23



3



342.7

What properties should a distance measure have?

- $D(A,B) = D(B,A)$ *Symmetry*
- $D(A,A) = 0$ *Constancy of Self-Similarity*
- $D(A,B) = 0$ iif $A = B$ *Positivity (Separation)*
- $D(A,B) \leq D(A,C) + D(B,C)$ *Triangular Inequality*

Intuitions behind desirable distance measure properties

$$D(A,B) = D(B,A)$$

Otherwise you could claim “Alex looks like Bob, but Bob looks nothing like Alex.”

$$D(A,A) = 0$$

Otherwise you could claim “Alex looks more like Bob, than Bob does.”

Intuitions behind desirable distance measure properties (continued)

$$D(A, B) = 0 \text{ IIf } A=B$$

Otherwise there are objects in your world that are different, but you cannot tell apart.

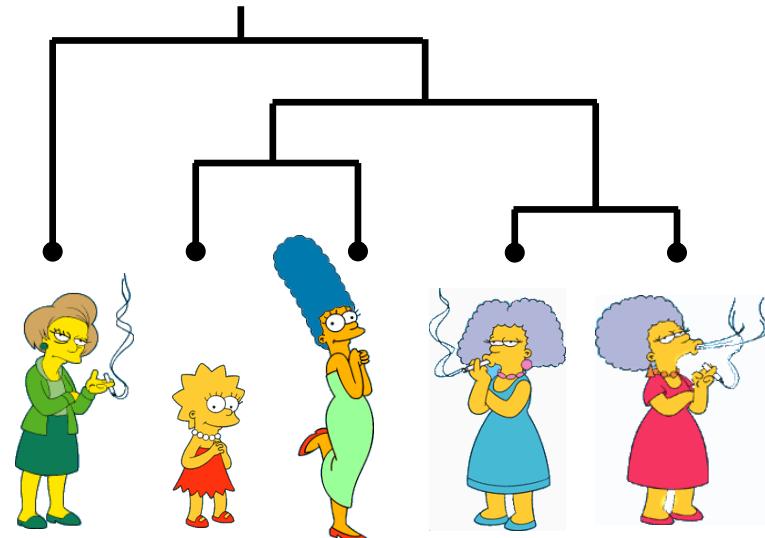
$$D(A, B) \leq D(A, C) + D(B, C)$$

Otherwise you could claim “Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl.”

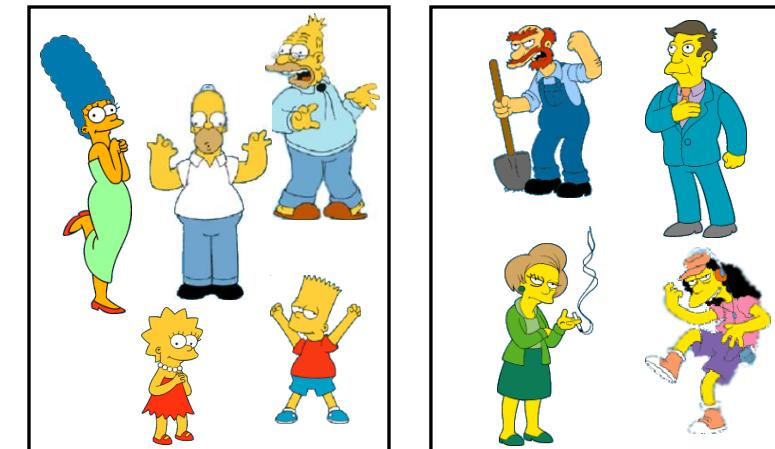
Two Types of Clustering

- **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

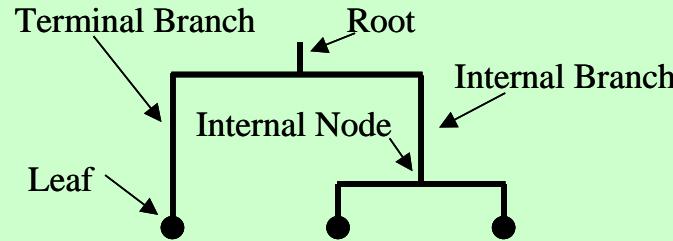
Hierarchical



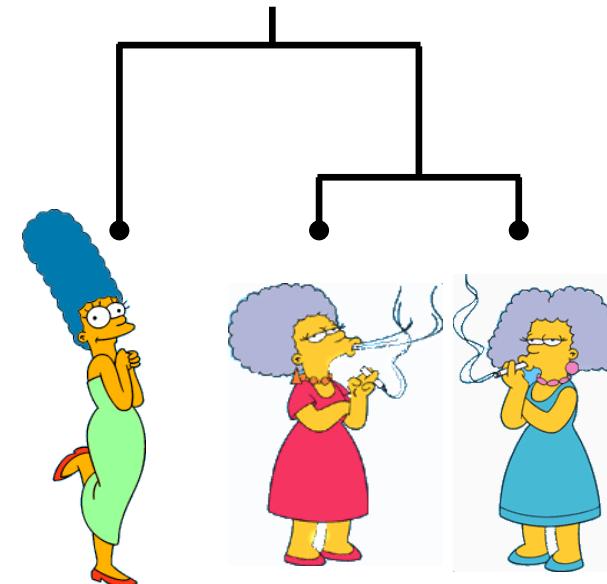
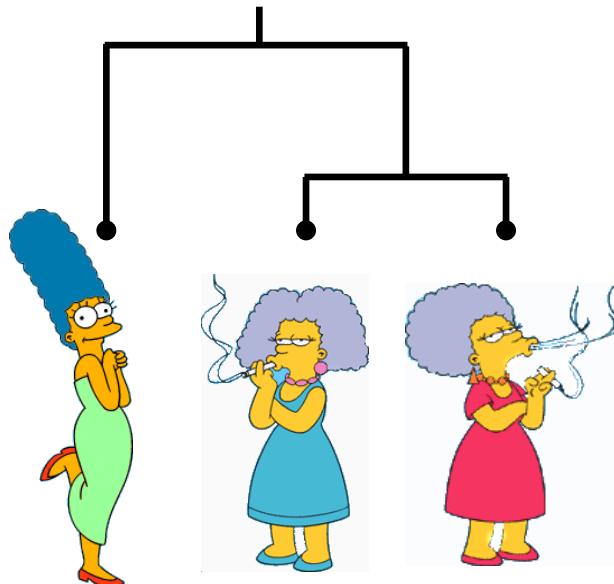
Partitional



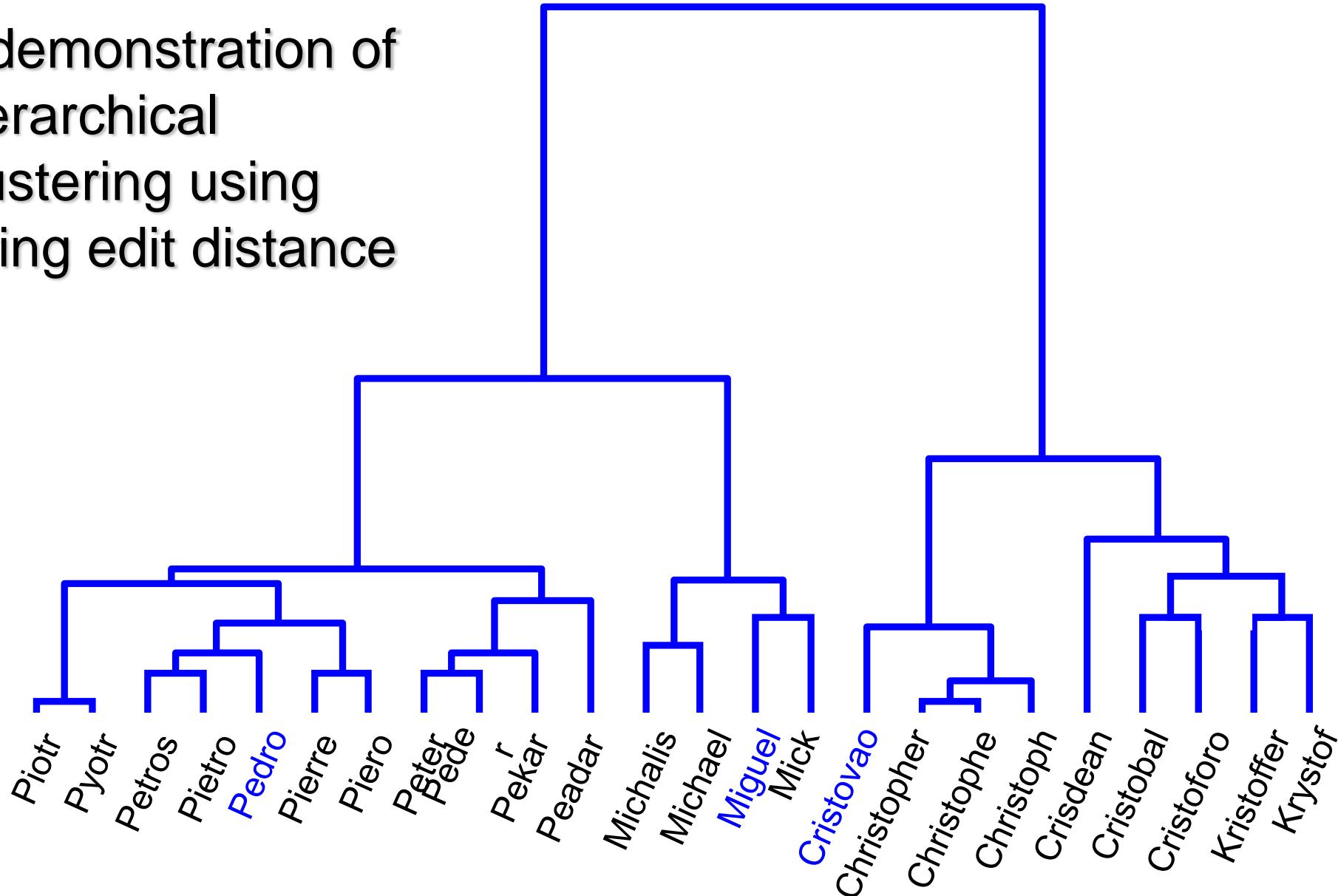
Dendrogram: A Useful Tool for Summarizing Similarity Measurements



The similarity between two objects in a dendrogram is represented as the height of the lowest internal node they share.



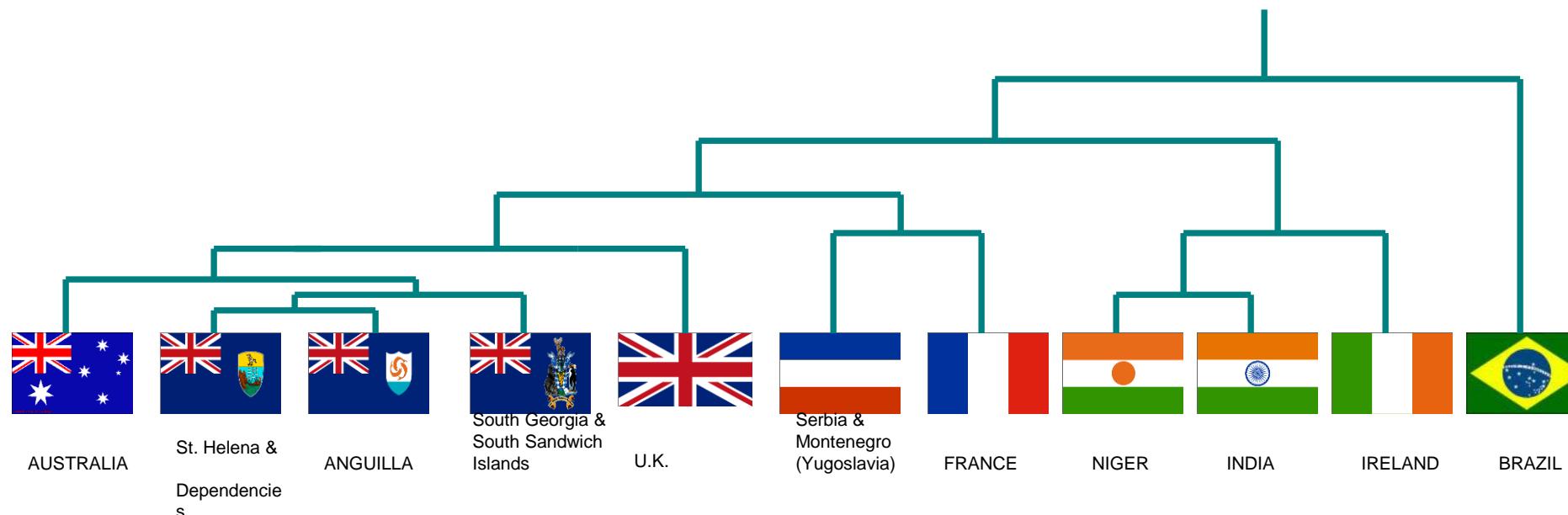
A demonstration of hierarchical clustering using string edit distance



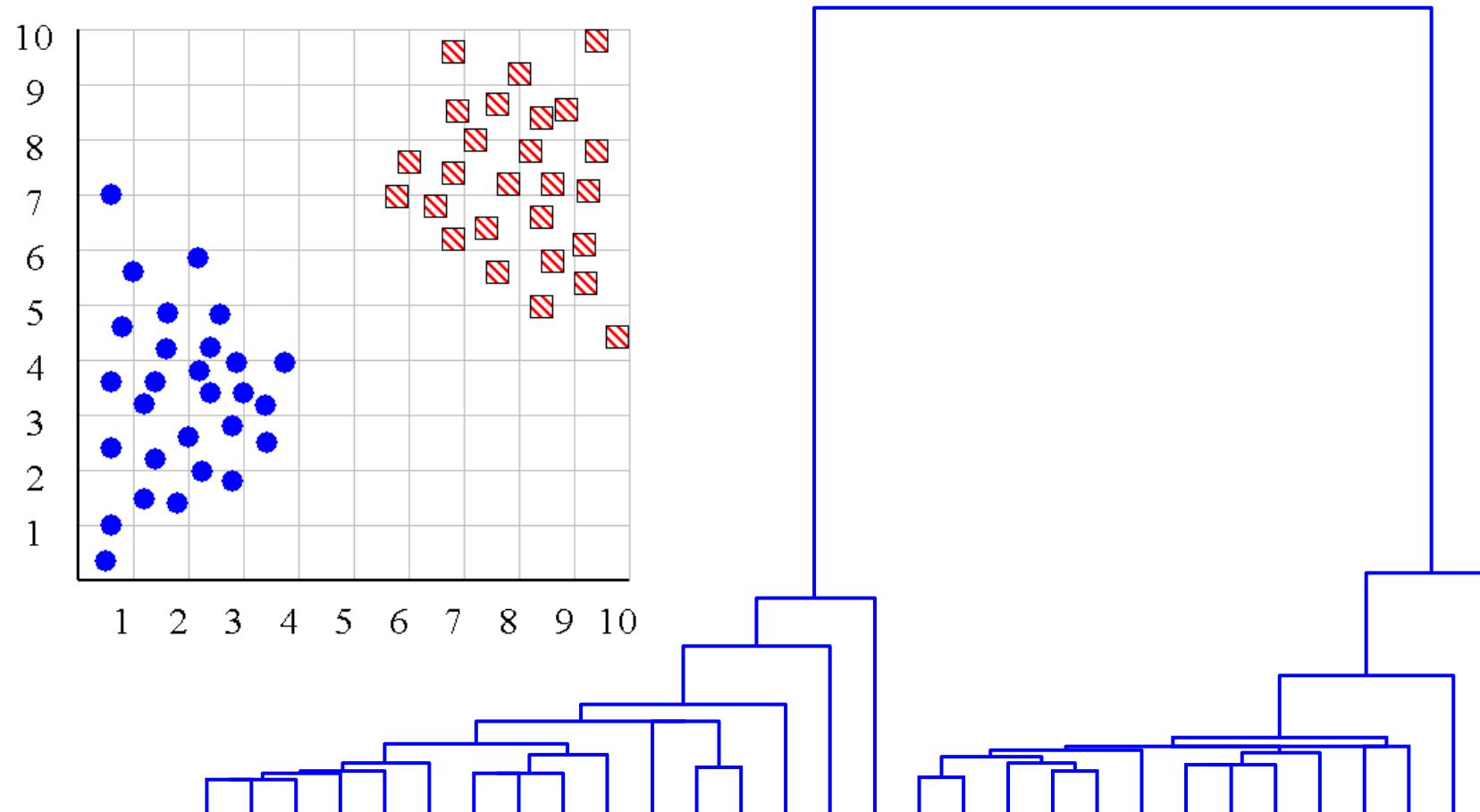
Hierarchal clustering can sometimes show patterns that are meaningless or spurious

The tight grouping of Australia, Anguilla, St. Helena etc is meaningful; all these countries are former UK colonies

However the tight grouping of Niger and India is completely spurious; there is no connection between the two.

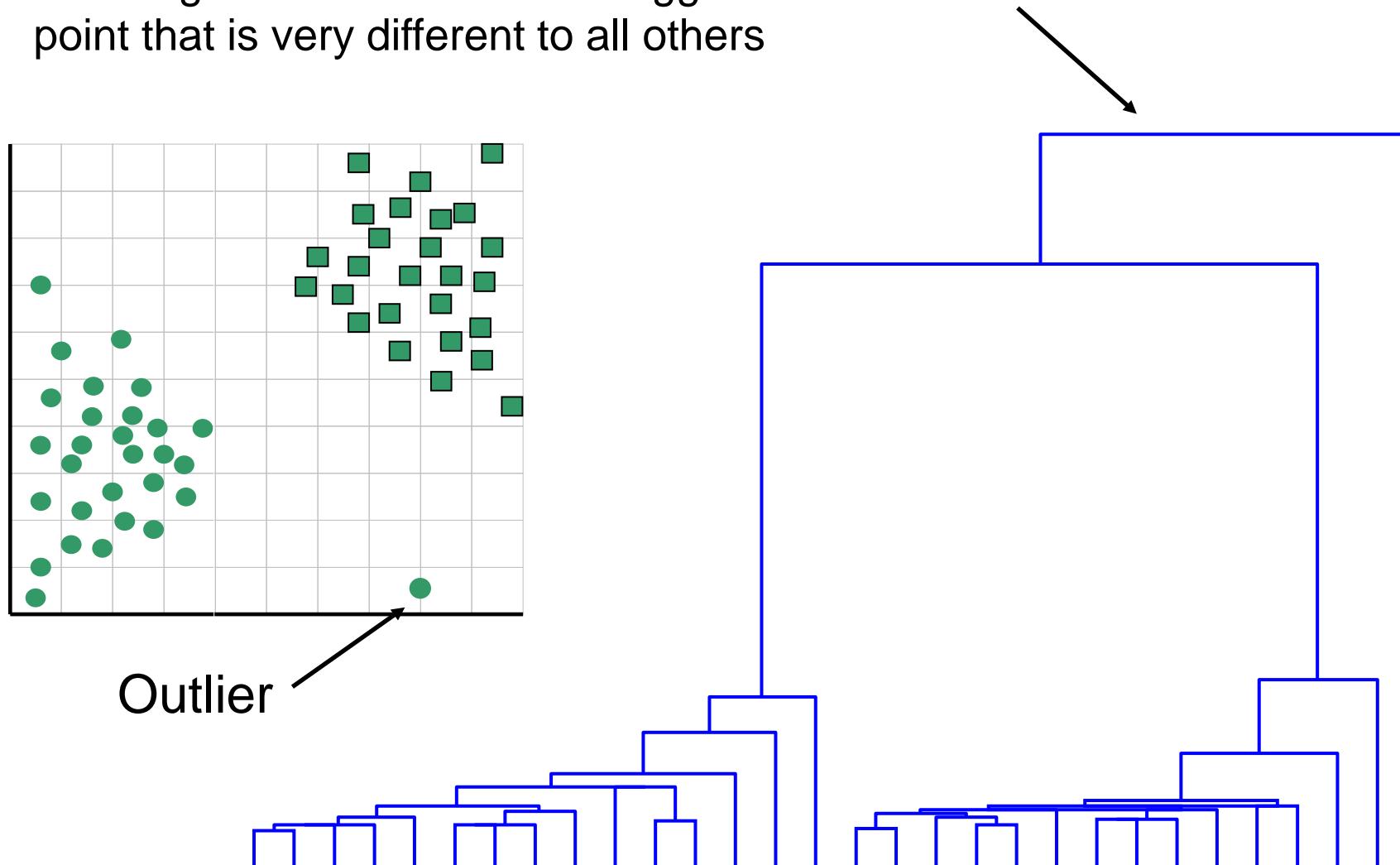


We can look at the dendrogram to determine the “correct” number of clusters.



One potential use of a dendrogram: detecting outliers

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



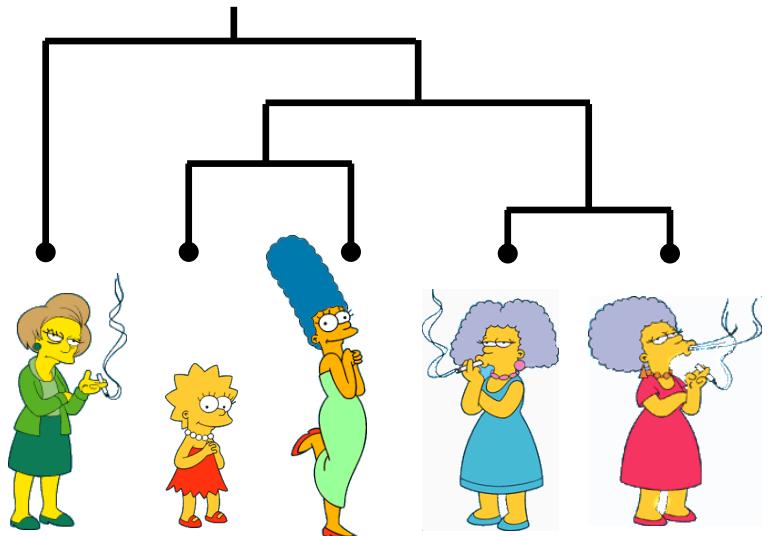
Hierarchical Clustering

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

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

Since we cannot test all possible trees we will have to heuristic search of all possible trees. We could do this..

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.



Top-Down (divisive): Starting with all the data in a single cluster, consider every possible way to divide the cluster into two. Choose the best division and recursively operate on both sides.

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

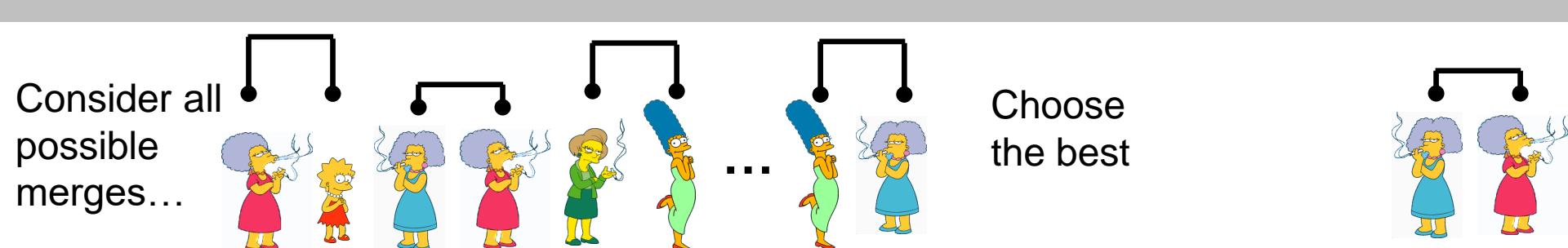
$$D(\text{Marge, Lisa}) = 8$$

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

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

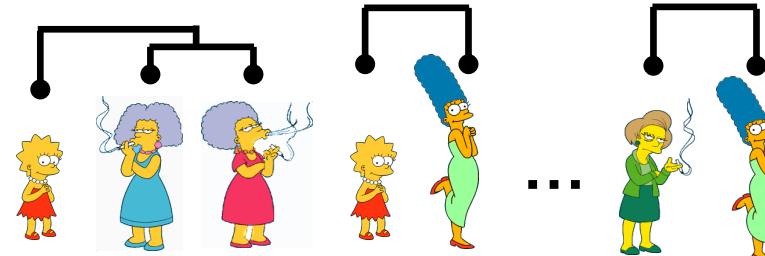
Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

This slide and next 4 based on slides by Eamonn Keogh

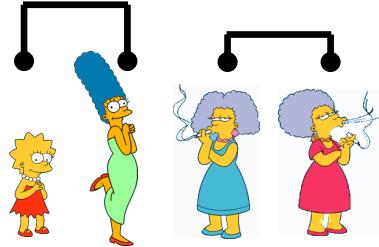


Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

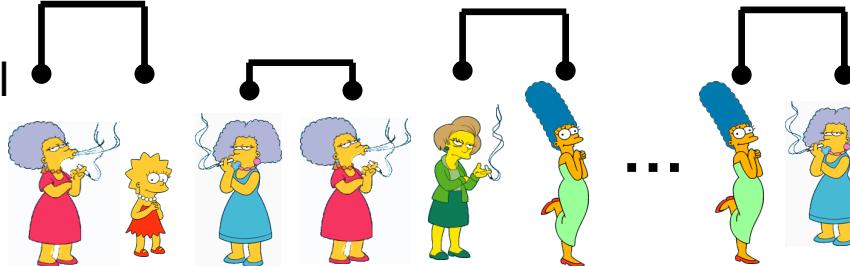
Consider all
possible
merges...



Choose
the best



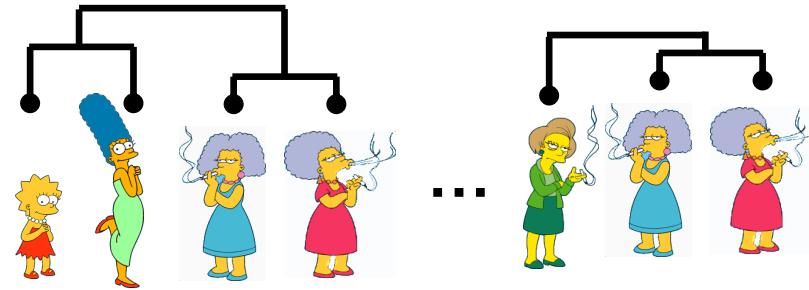
Consider all
possible
merges...



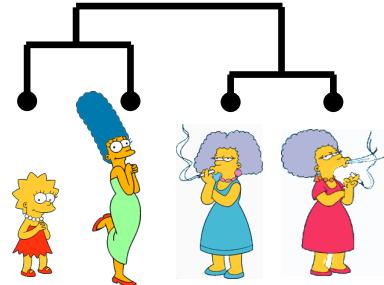
Choose the best

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.

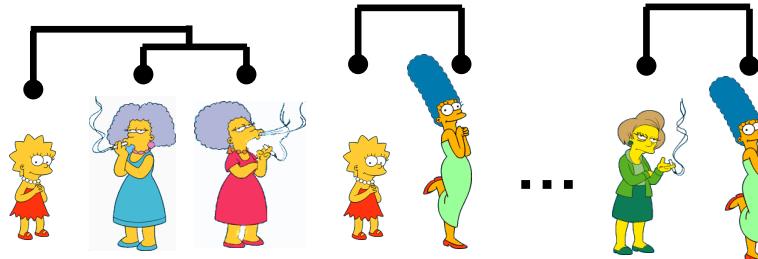
Consider all possible merges...



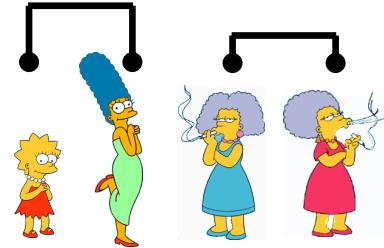
Choose the best



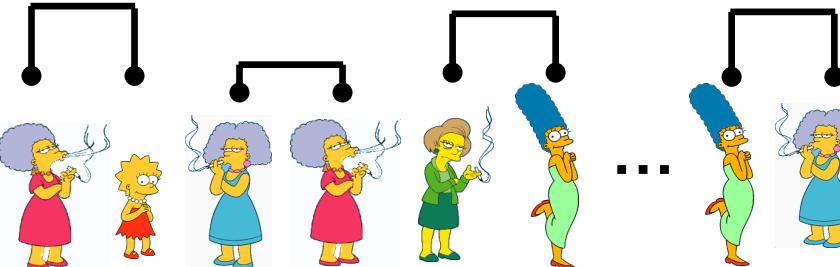
Consider all possible merges...



Choose the best



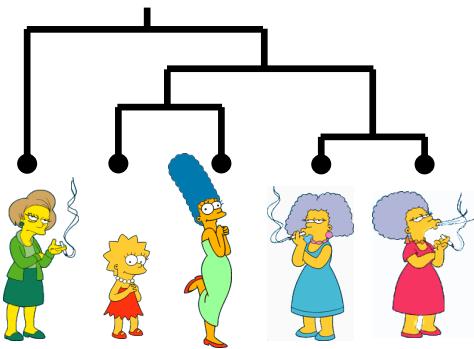
Consider all possible merges...



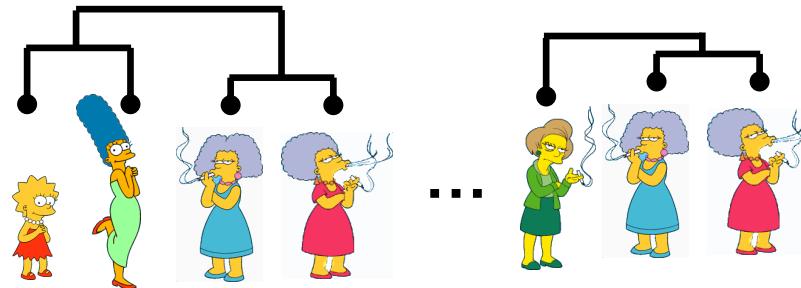
Choose the best



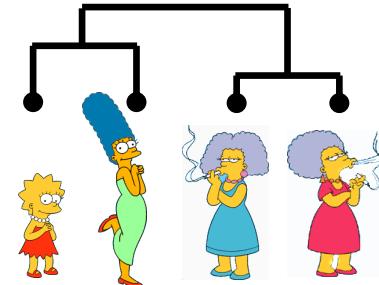
Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



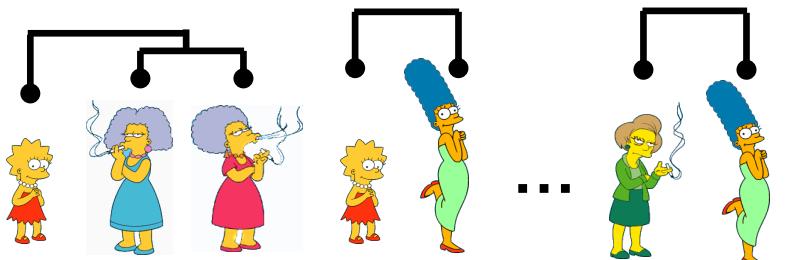
Consider all possible merges...



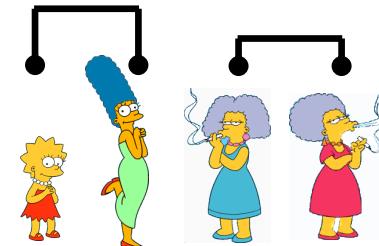
Choose the best



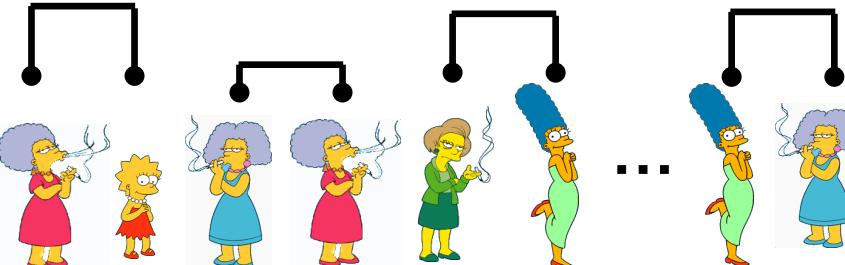
Consider all possible merges...



Choose the best



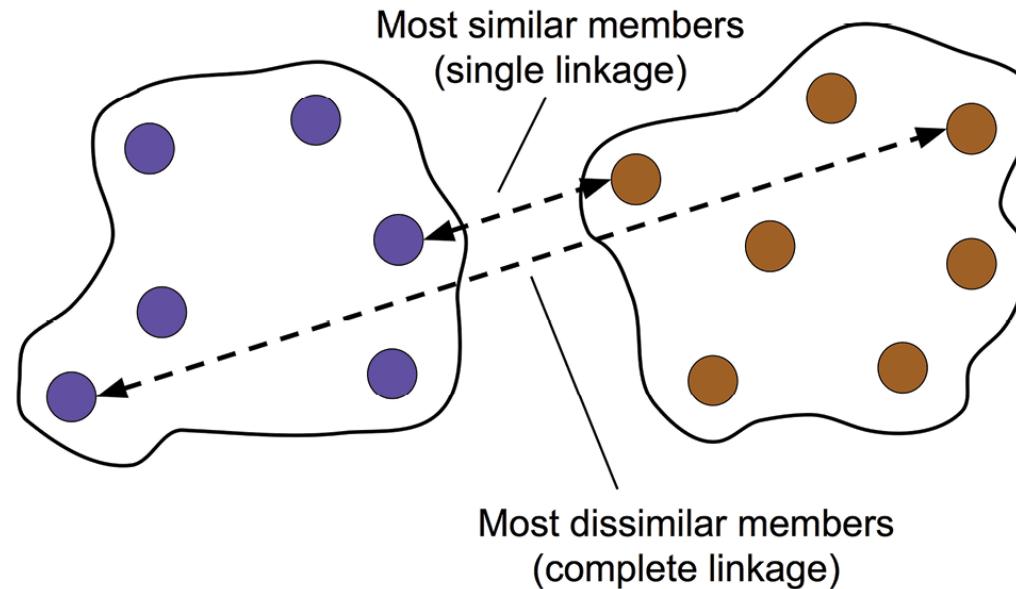
Consider all possible merges...



Choose the best



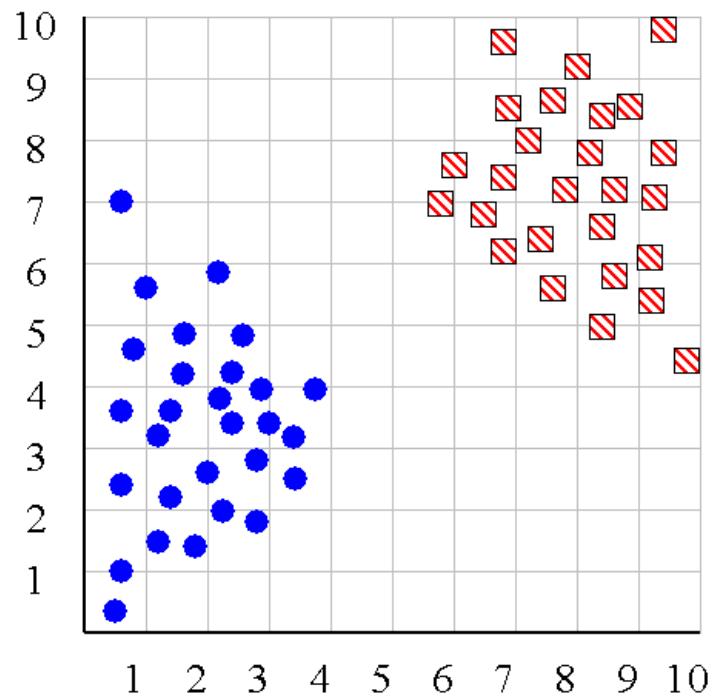
Distances



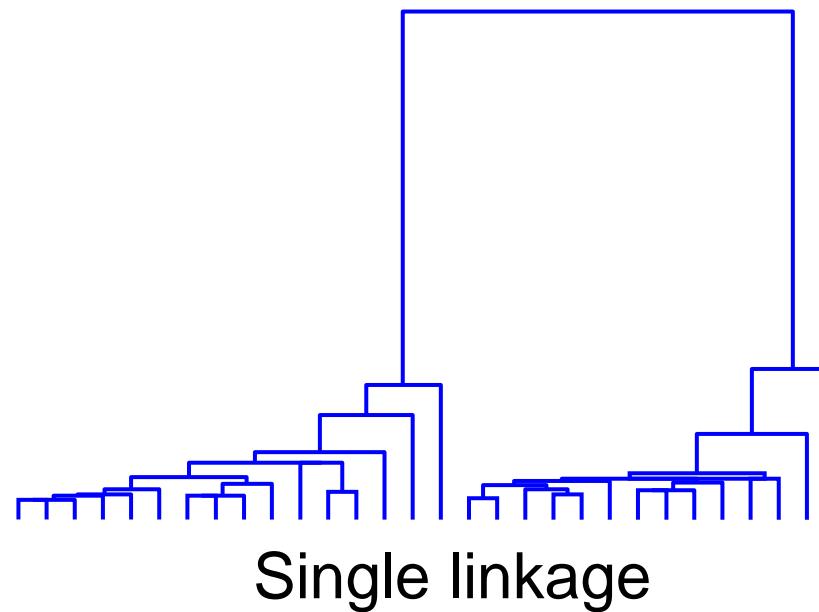
- Have a distance measure on pairs of objects, $d(x, y)$:
- Single linkage: $\text{dist}(A, B) = \min_{x \in A, x' \in B} d(x, x')$
- Complete linkage: $\text{dist}(A, B) = \max_{x \in A, x' \in B} d(x, x')$
- Average linkage: $\text{dist}(A, B) = \text{average } d(x, x')_{x \in A, x' \in B}$
- Ward's method $\text{dist}(A, B) = \frac{|A| |B|}{|A| + |B|} \|\text{mean}(A) - \text{mean}(B)\|^2$

What difference does it make?

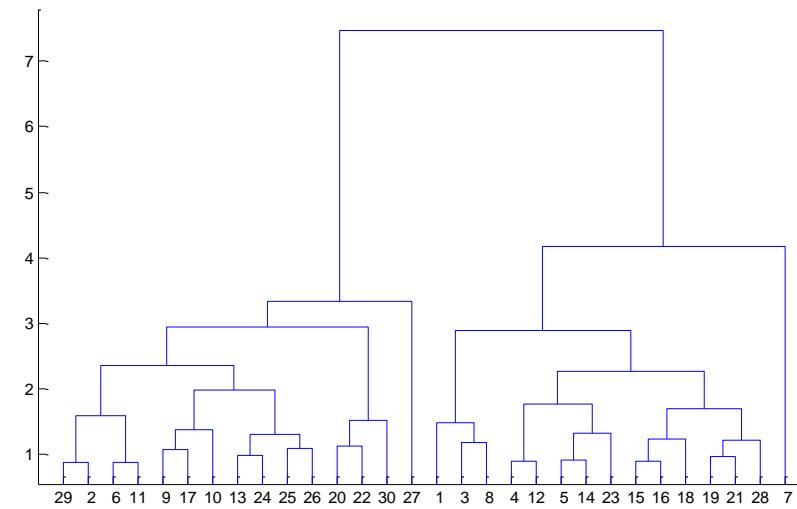
- **Single linkage:** $\text{dist}(A, B) = \min_{x \in A, x' \in B} d(x, x')$
 - At any time, distance between any two points in a connected components $< r$.
- **Complete linkage:** $\text{dist}(A, B) = \max_{x \in A, x' \in B} d(x, x')$
 - Keep max diameter as small as possible at any level
- **Ward's method** $\text{dist}(A, B) = \frac{|A| |B|}{|A| + |B|} \|\text{mean}(A) - \text{mean}(B)\|^2$
 - Merge the two clusters such that the increase in k-means cost is as small as possible.
 - Works well in practice



Slide from Eamonn Keogh, from lecture by Carla Brodley, Tufts University

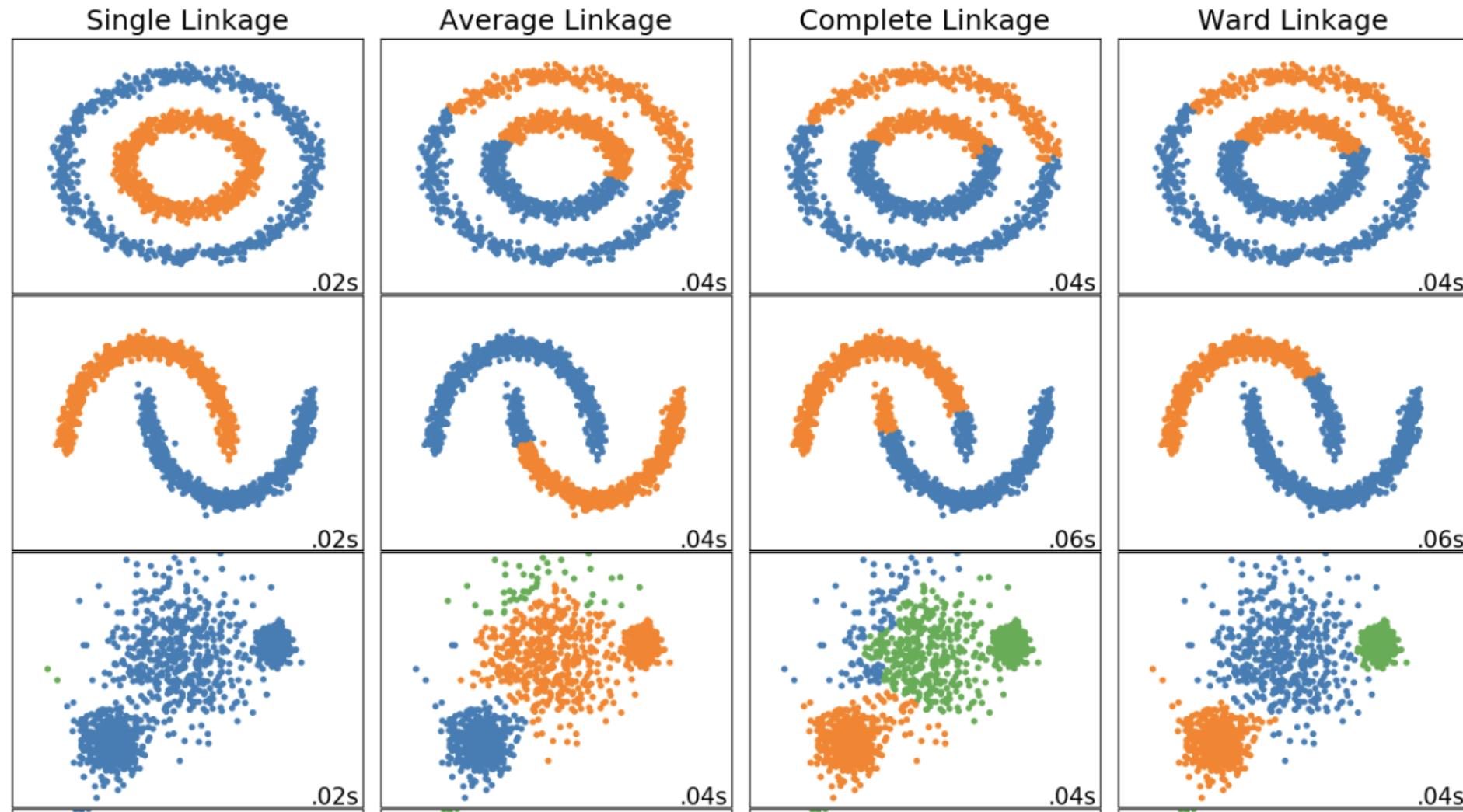


Single linkage



Average linkage

What difference does it make?



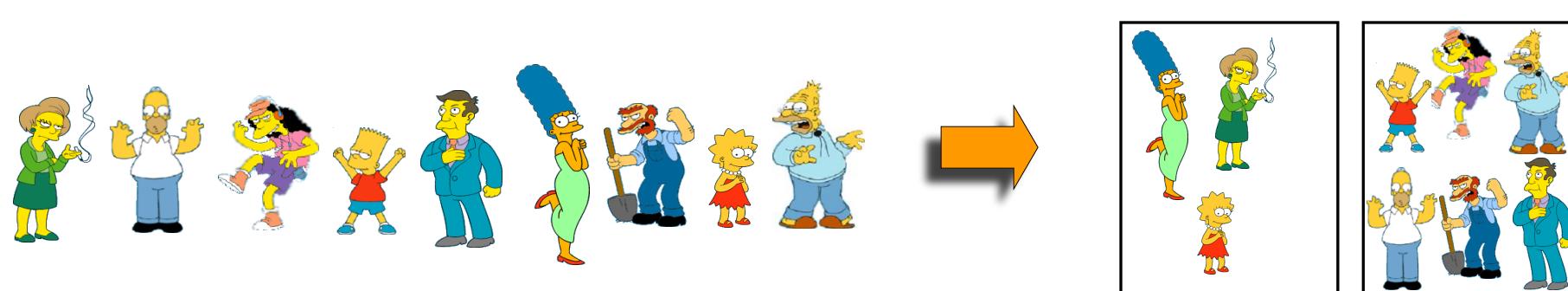
Ward linkage measures variance of clusters. The distance between two clusters, A and B, is how much the sum of squares would increase if we merged them.

Hierarchical Clustering Methods Summary

- No need to specify the number of clusters in advance
- Hierarchical nature 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

Partitional Clustering

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



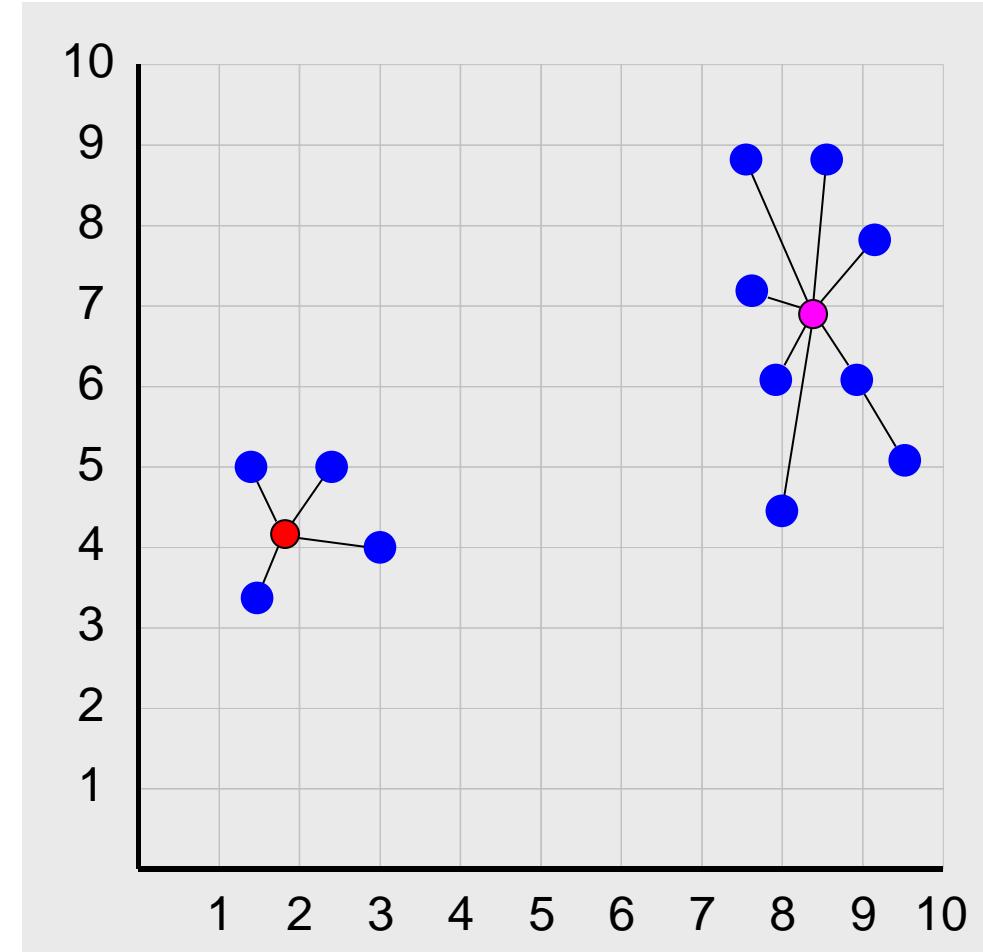
Squared Error

$$se_{K_i} = \sum_{j=1}^m \|t_{ij} - C_k\|^2$$

$$se_K = \sum_{j=1}^k se_{K_j}$$



Objective Function

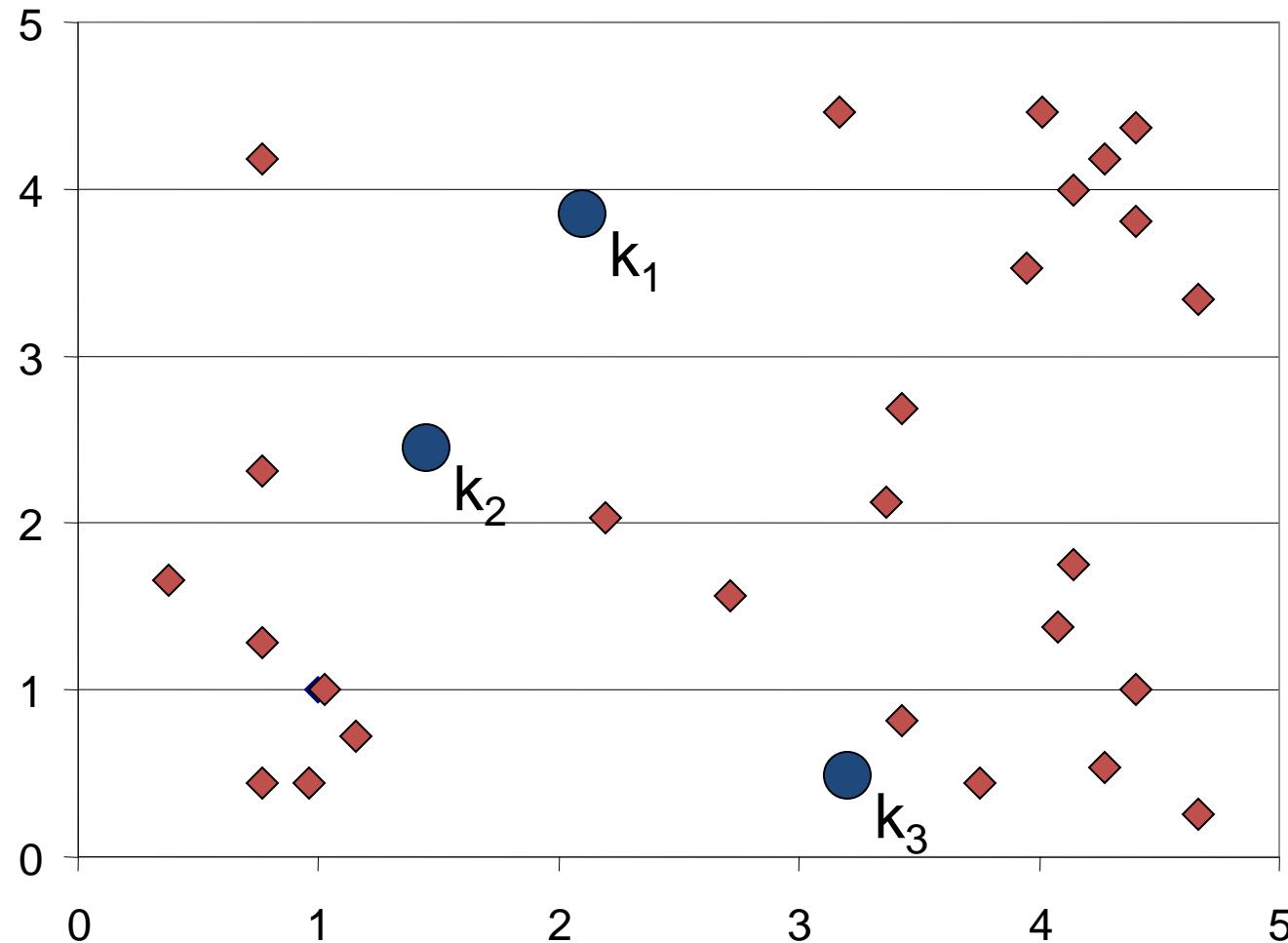


Partition Algorithm 1: k-means

1. Decide on a value for k .
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. If none of the N objects changed membership in the last iteration, exit. Otherwise goto 3.

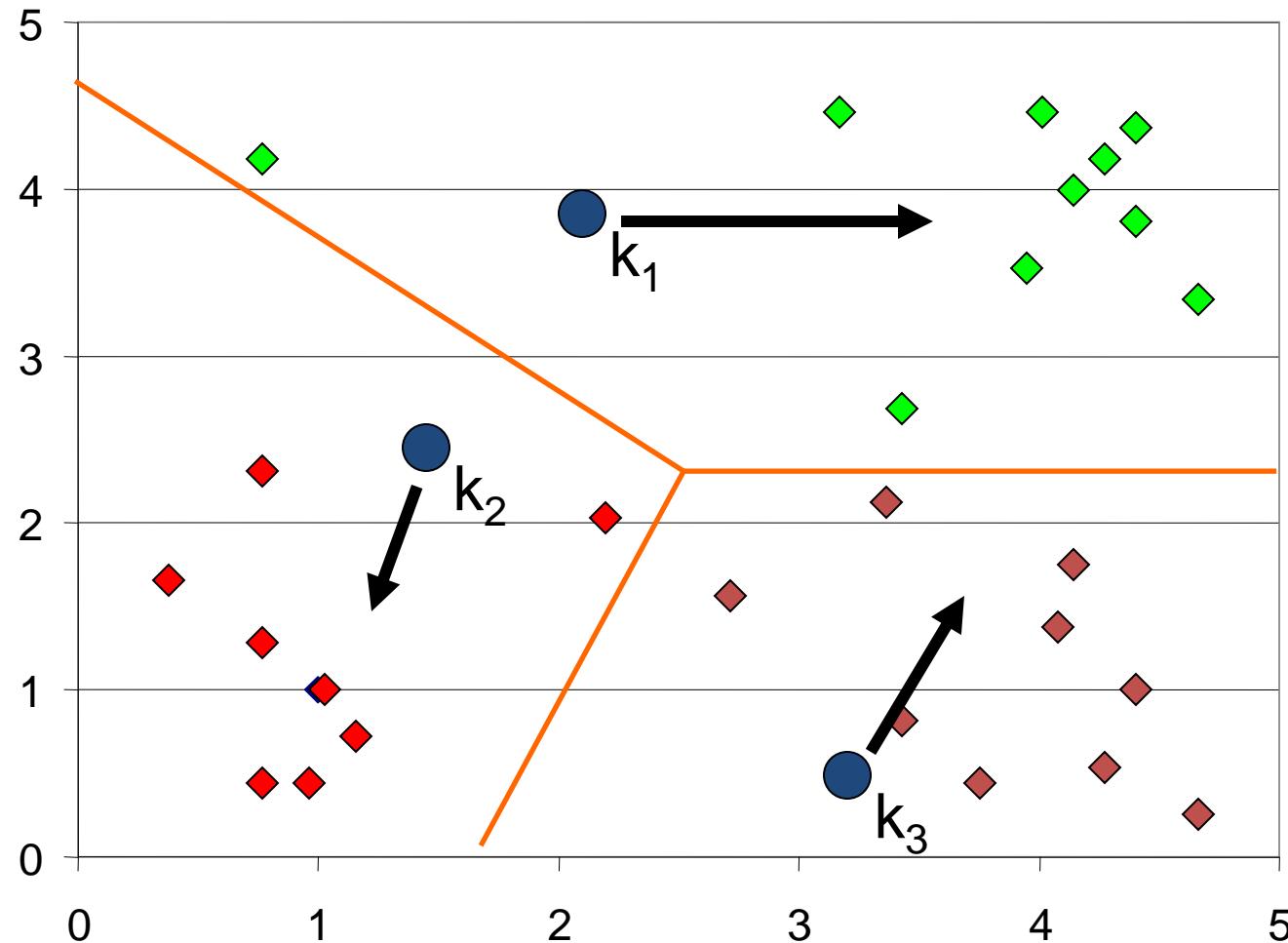
K-means Clustering: Step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



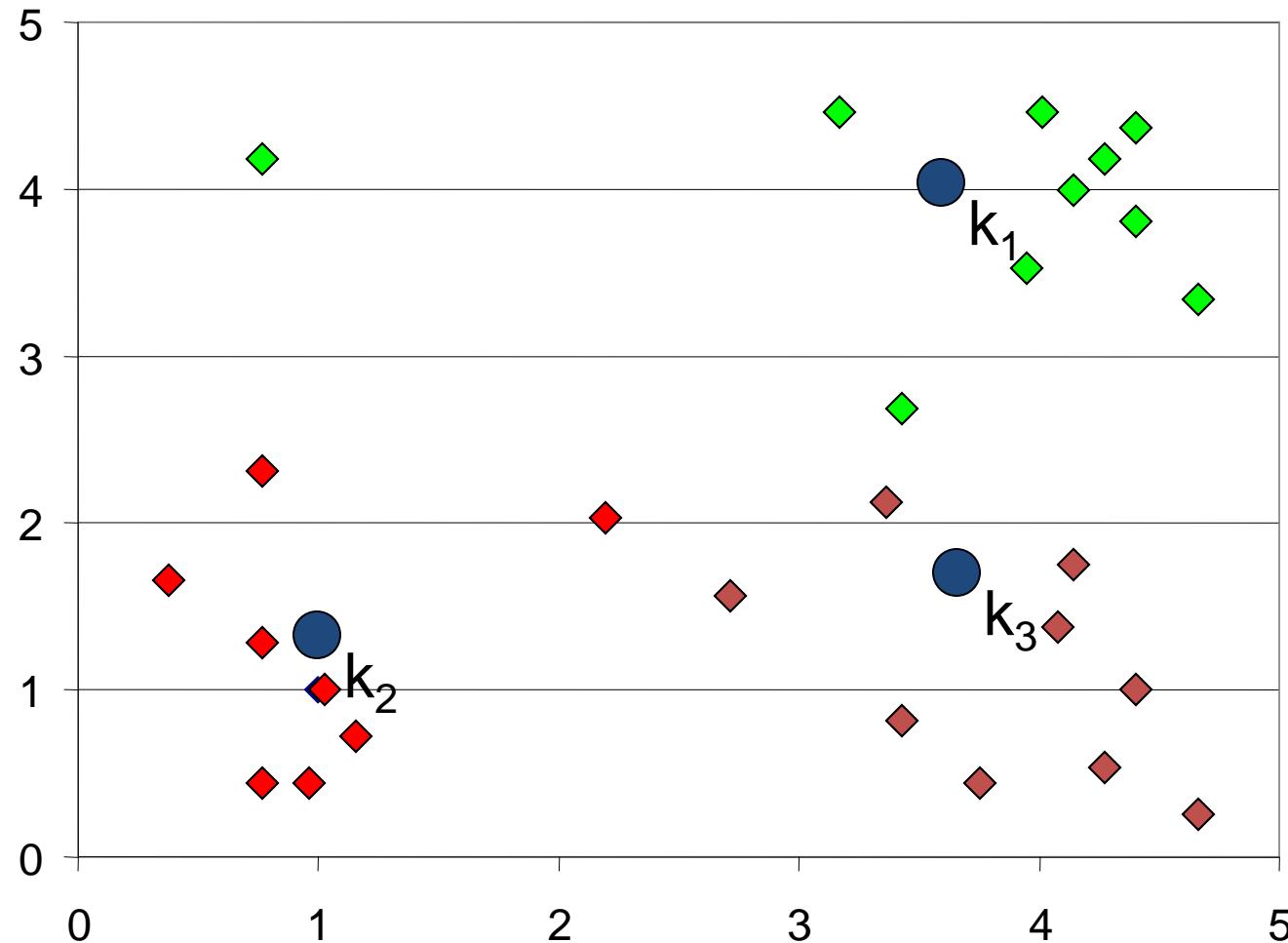
K-means Clustering: Step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



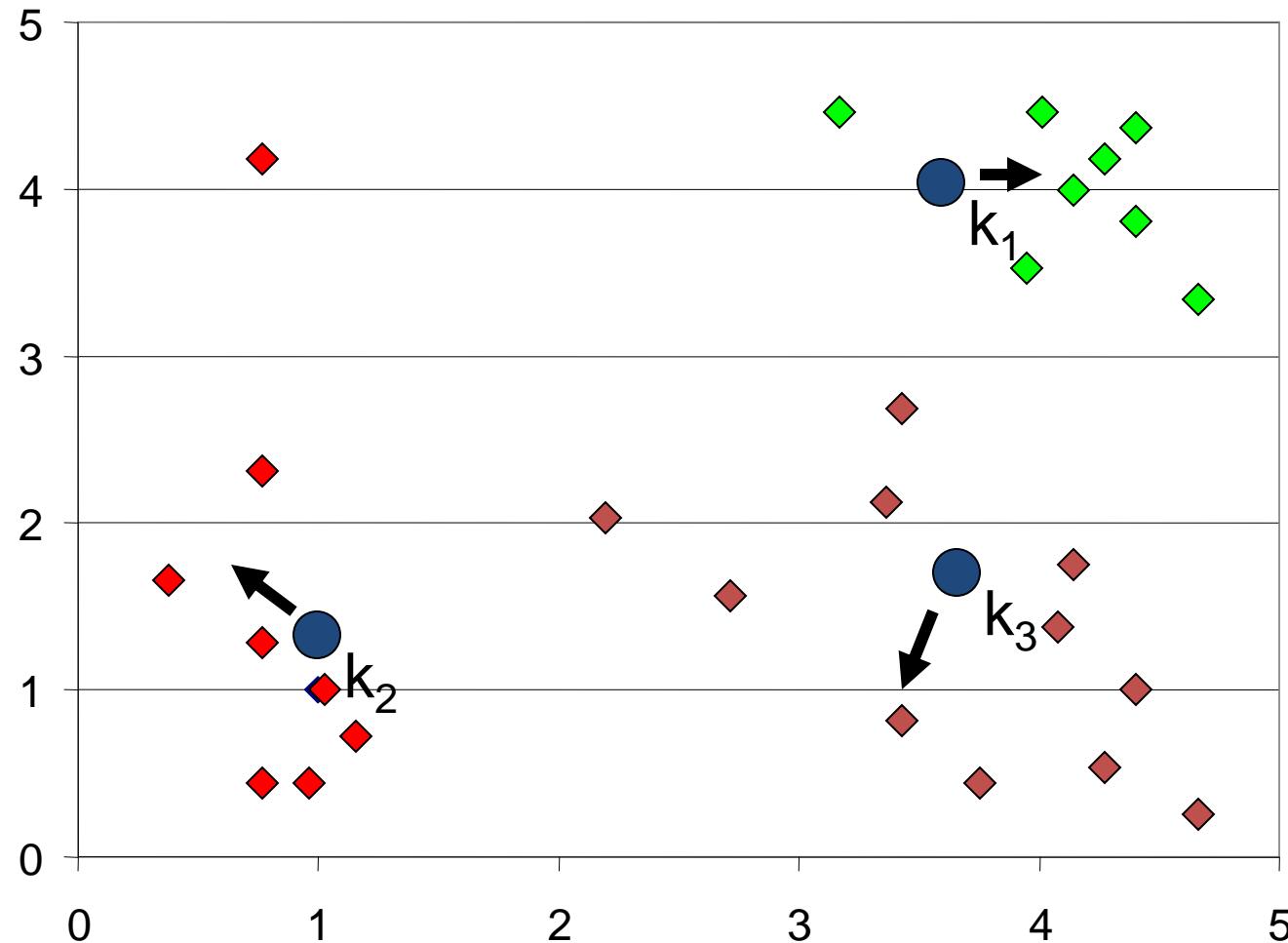
K-means Clustering: Step 3

Algorithm: k-means, Distance Metric: Euclidean Distance



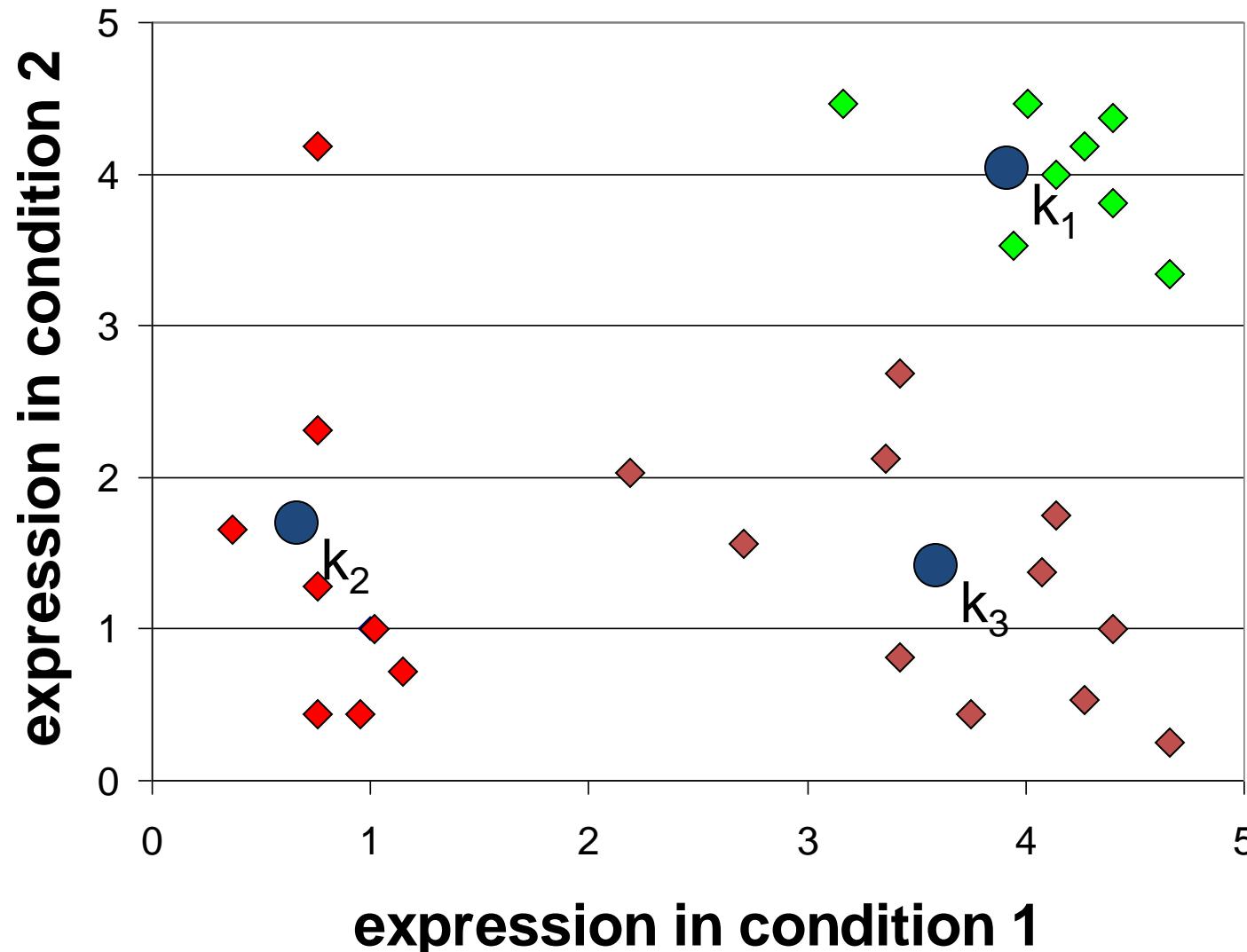
K-means Clustering: Step 4

Algorithm: k-means, Distance Metric: Euclidean Distance

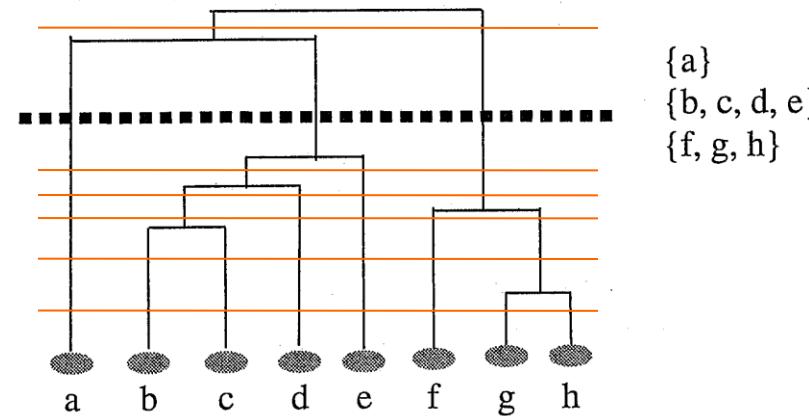


K-means Clustering: Step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



How many clusters to choose?



- Depends on goals
 - May know beforehand how many clusters you want - or at least a range (e.g. 2-10)
 - Could analyze the dendrogram and data after the full clustering to decide which subclustering level is most appropriate for the task at hand
 - Could use automated *cluster validity* metrics to help
- Could do stopping criteria during clustering

How many clusters to choose?

Compactness: members of a cluster are all similar and close together

- One measure of compactness of a cluster is the square distance of the cluster instances compared to the cluster centroid

$$Comp(C) = \sum_{i=1}^{|X_c|} (\mathbf{c} - \mathbf{x}_i)^2$$

- where \mathbf{c} is the centroid of a cluster C , made up of instances X_c . Lower is better.
- The overall compactness of a particular clustering is just the sum of the compactness of the individual clusters
- Gives a numeric way to compare different clusterings by seeking clusterings which minimize the compactness metric

How many clusters to choose?

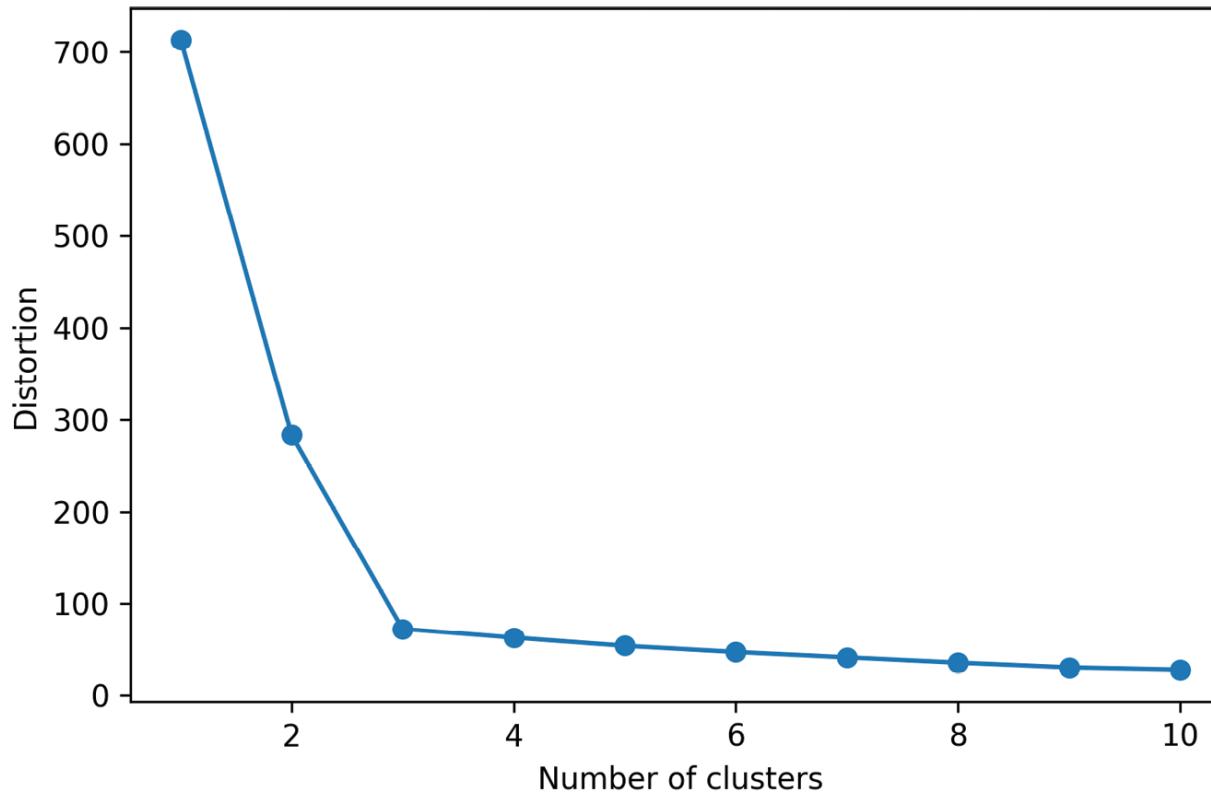
Separability: members of one cluster are sufficiently different from members of another cluster (cluster dissimilarity)

- One measure of the separability of two clusters is their squared distance. The bigger the distance the better.
- $dist_{ij} = (\mathbf{c}_i - \mathbf{c}_j)^2$ where \mathbf{c}_i and \mathbf{c}_j are two cluster centroids
- For a clustering which cluster distances should we compare?
- For each cluster we add in the distance to its closest neighbor cluster

$$Separability = \min_{i=1}^{|C|} \min_j dist_{ij}(\mathbf{c}_i, \mathbf{c}_j)$$

- We would like to find clusterings where separability is maximized
- separability is usually maximized when there are very few clusters

Elbow methods



Calculate cluster size as a function of number of clusters