Introduction to Clustering

Carla Brodley

Tufts University

Clustering (Unsupervised Learning)

Given: Examples: $< x_1 , x_2 , \ldots x_n >$

Find: A natural clustering (grouping) of the data

Example Applications:

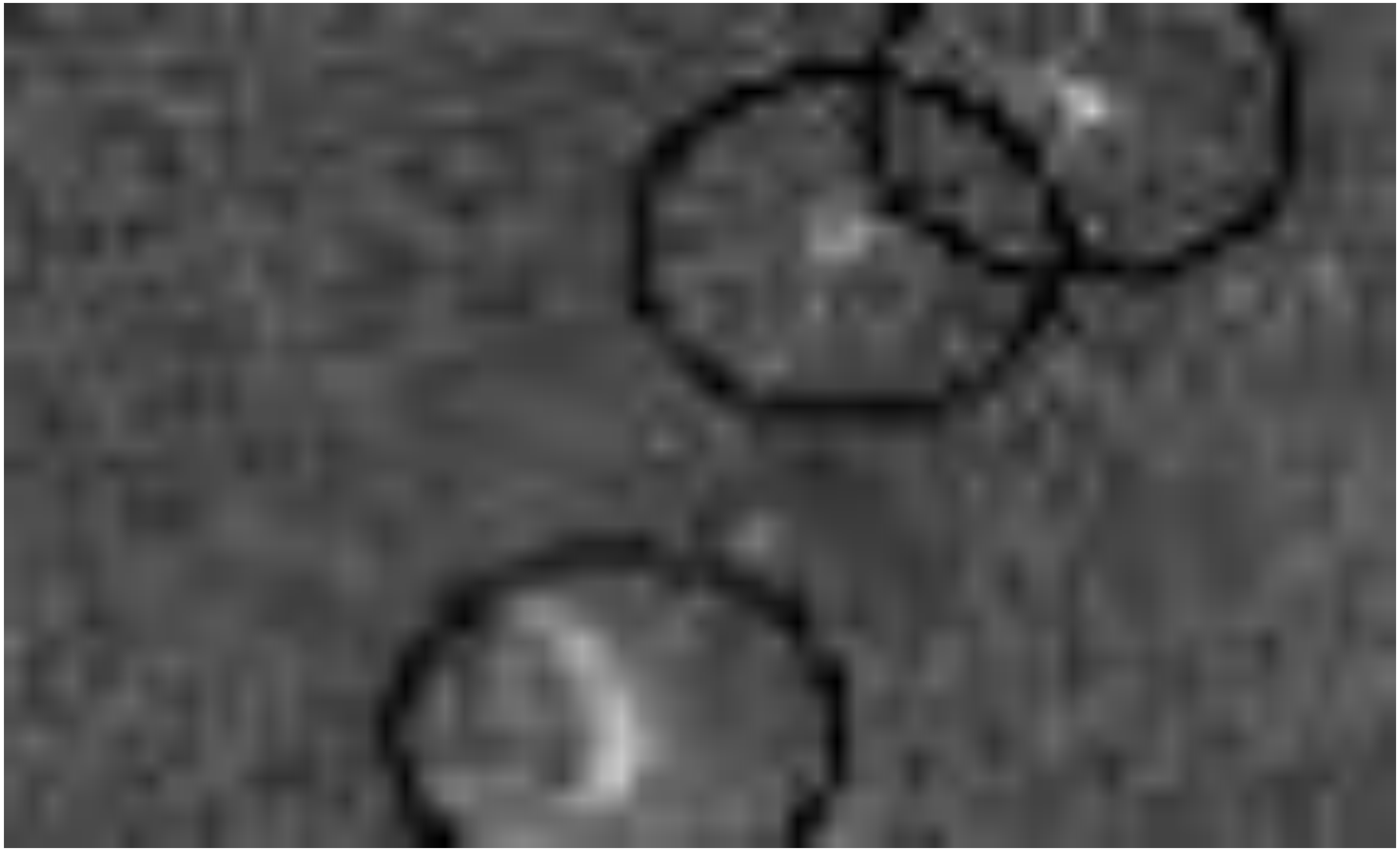
Identify similar energy use customer profiles

<**x**> = time series of energy usage

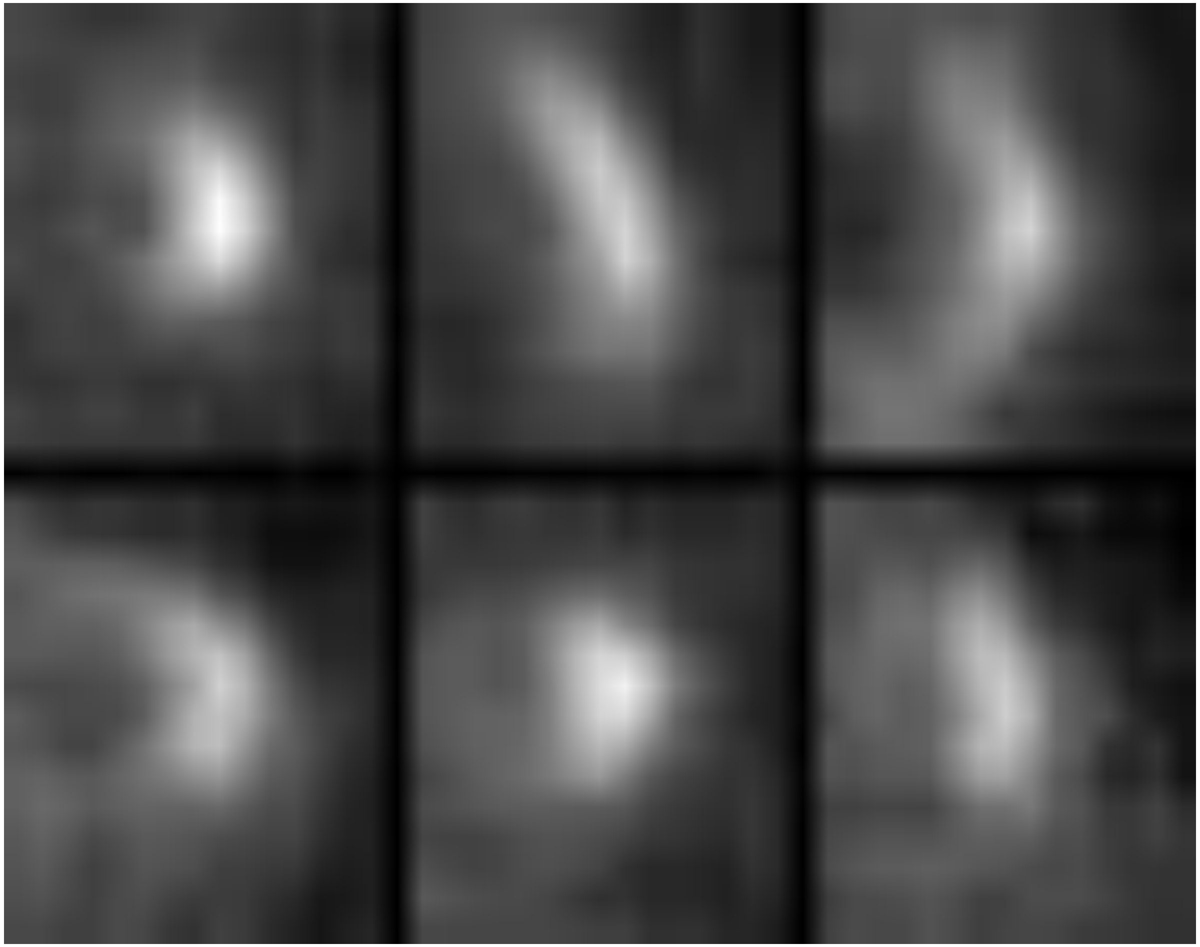
Identify anomalies in user behavior for computer security

<**x**> = sequences of user commands

Magellan Image of Venus



Prototype Volcanoes

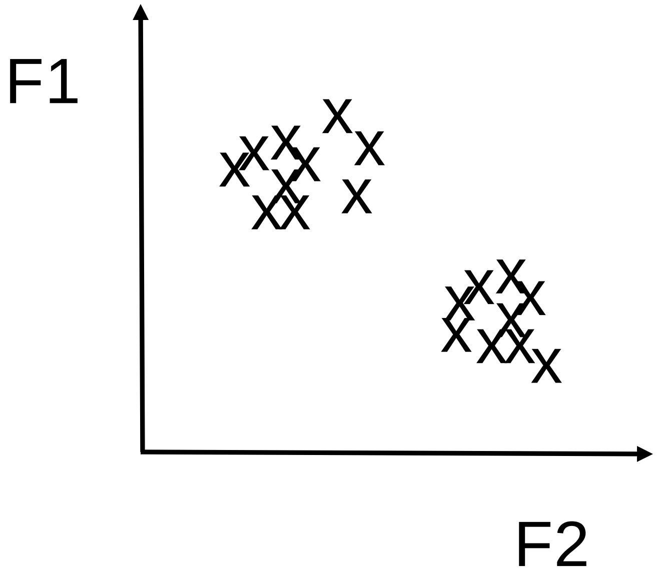


Why cluster?

- Labeling is expensive
- Gain insight into the structure of the data
- Find prototypes in the data

Goal of Clustering

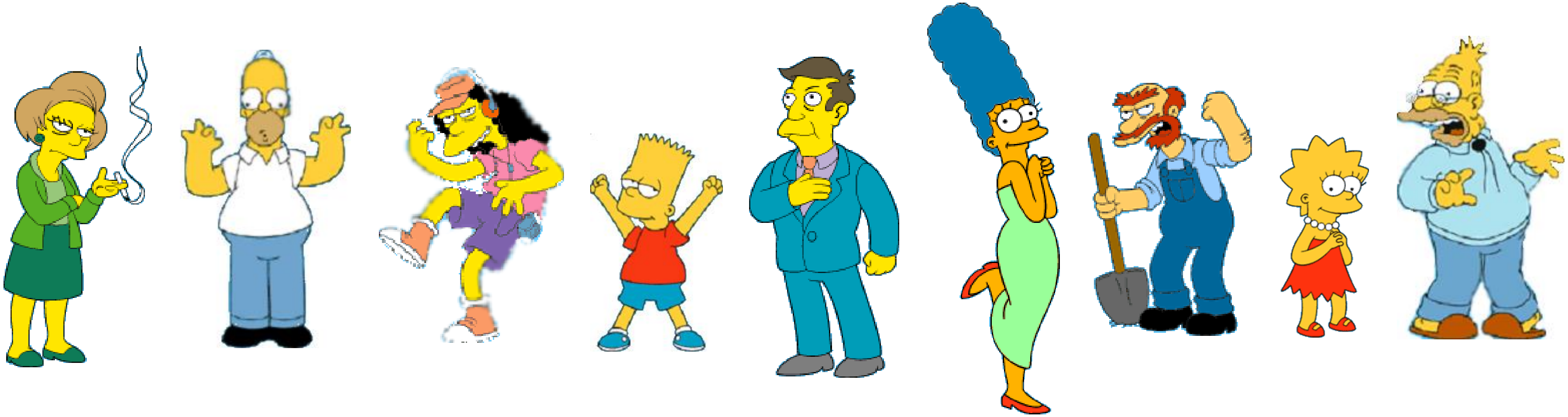
- Given a set of data points, each described by a set of attributes, find clusters such that:
  - Inter-cluster similarity is maximized
  - Intra-cluster similarity is minimized



- Requires the definition of a similarity measure

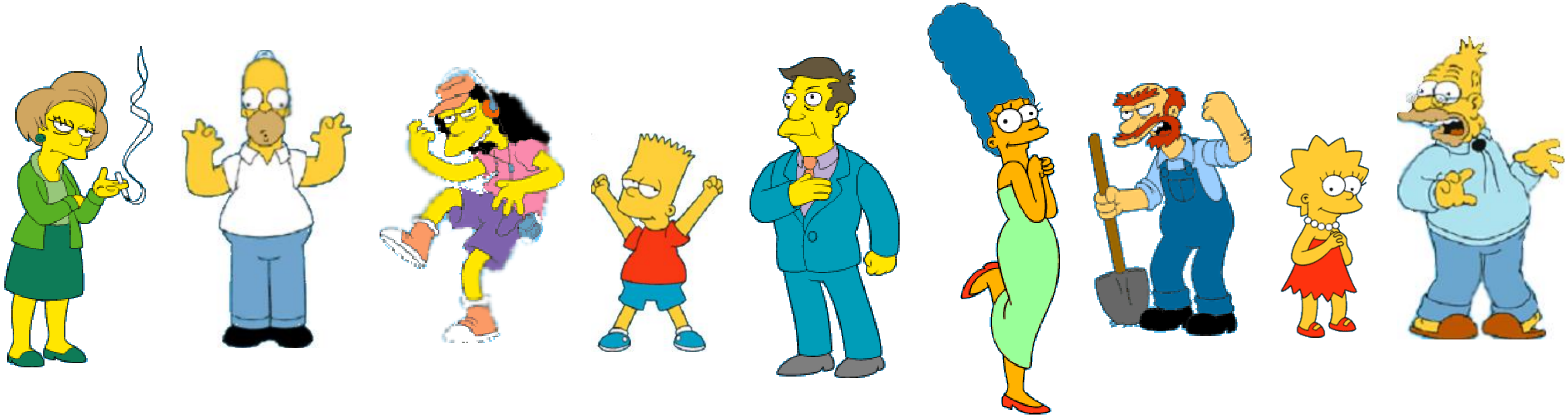
What is a natural grouping of these objects?

Slide from Eamonn Keogh

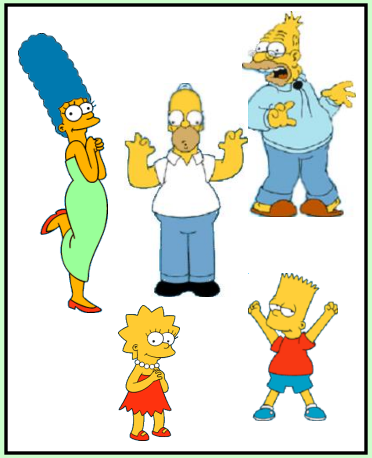


What is a natural grouping of these objects?

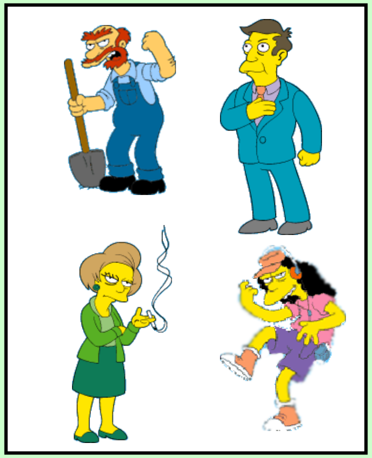
Slide from Eamonn Keogh



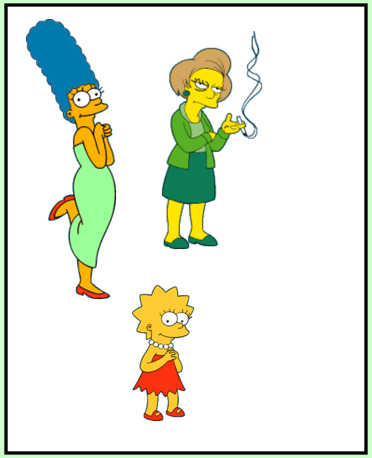
Clustering is subjective



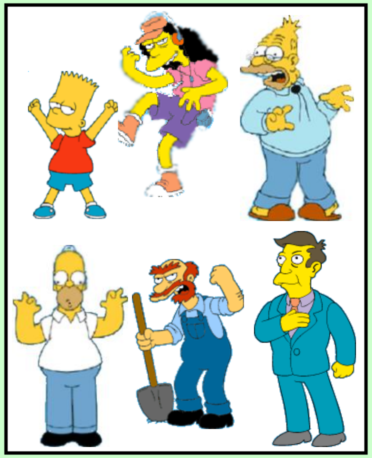
Simpson's Family



School Employees



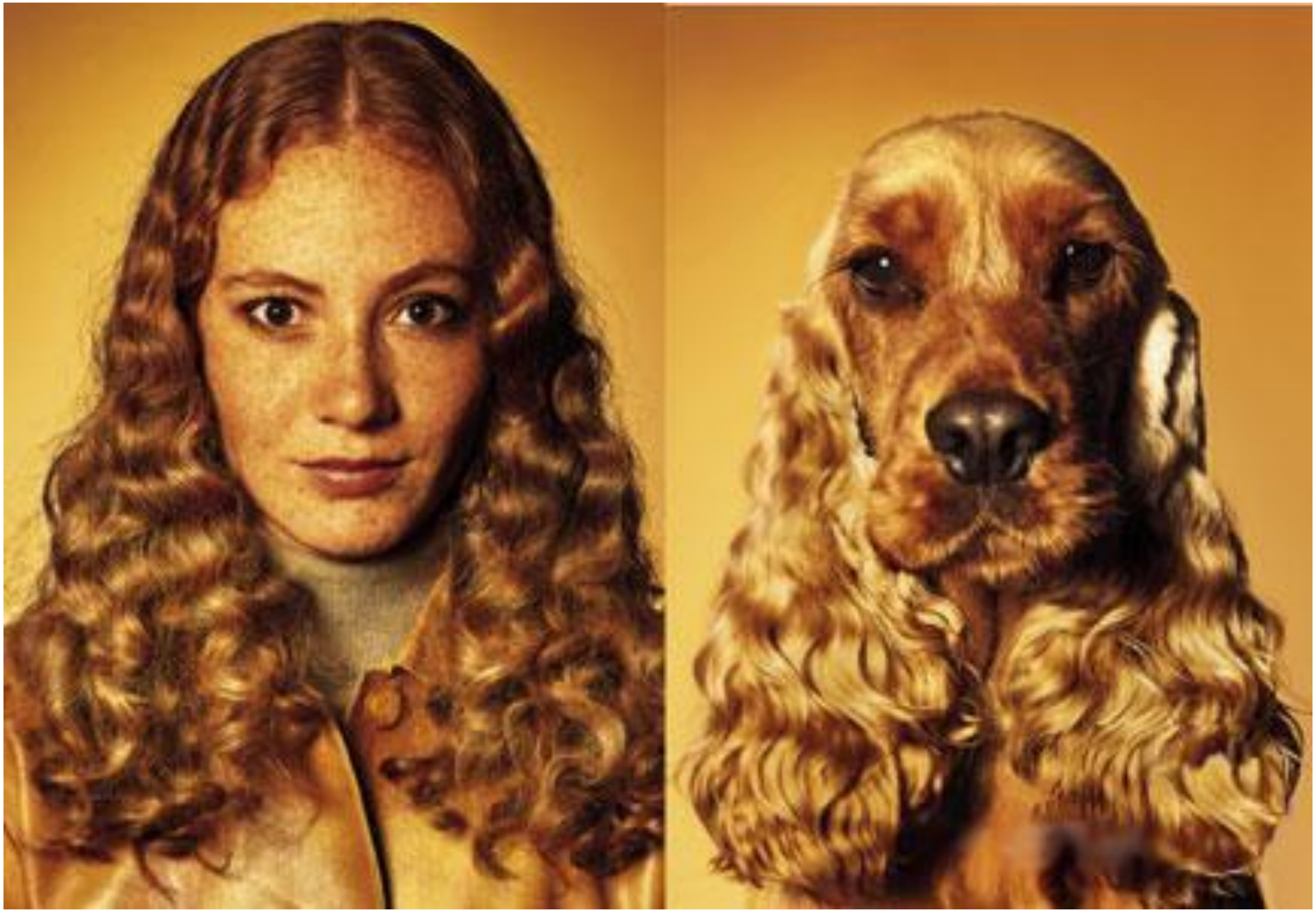
Females



Males

What is Similarity?

Slide based on one by Eamonn Keogh

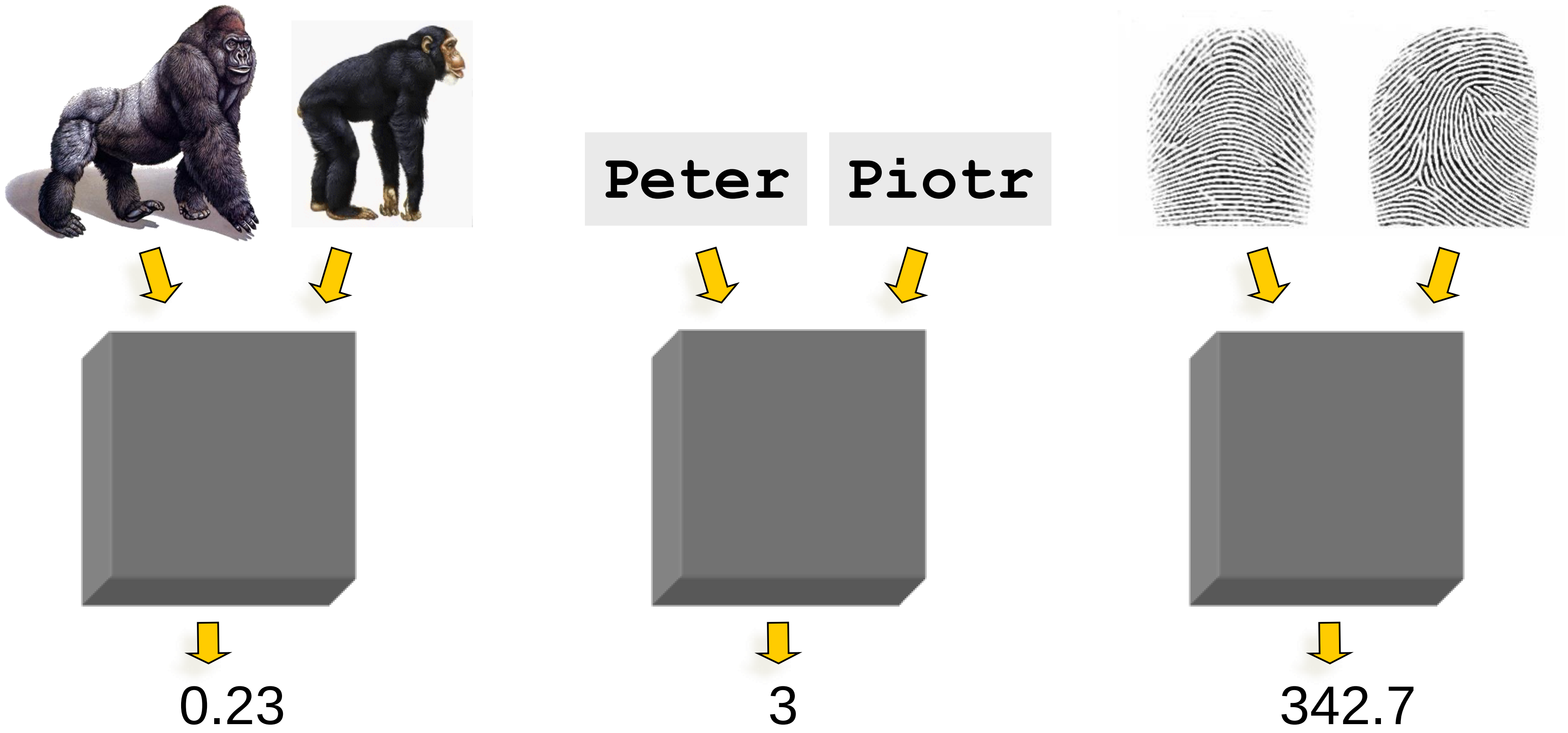


Similarity is hard to define, but…
“*We know it when we see it*”

Defining Distance Measures

Slide from Eamonn Keogh

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

Slide based on one by Eamonn Keogh

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*

Slide based on one by Eamonn Keogh

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."

Slide based on one by Eamonn Keogh

Intuitions behind desirable distance measure properties (continued)

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

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

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

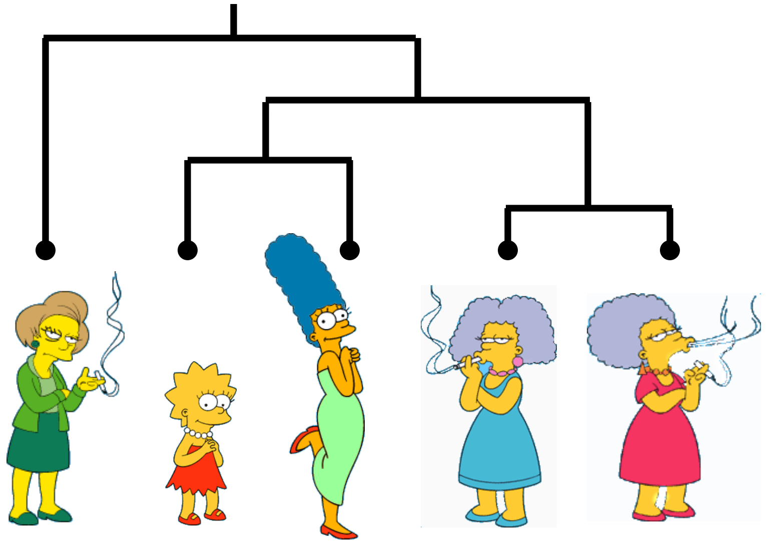
Otherwise you could claim “Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl.”

Slide based on one by Eamonn Keogh

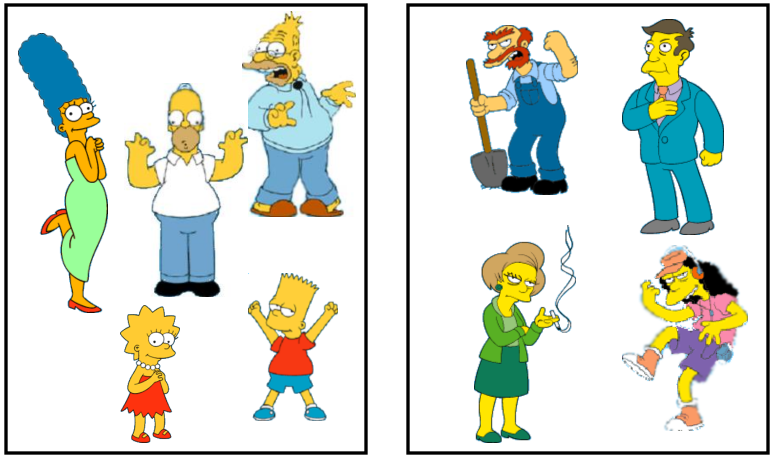
Two Types of Clustering

- **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

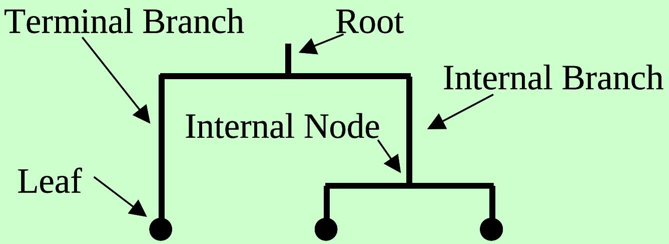


Partitional

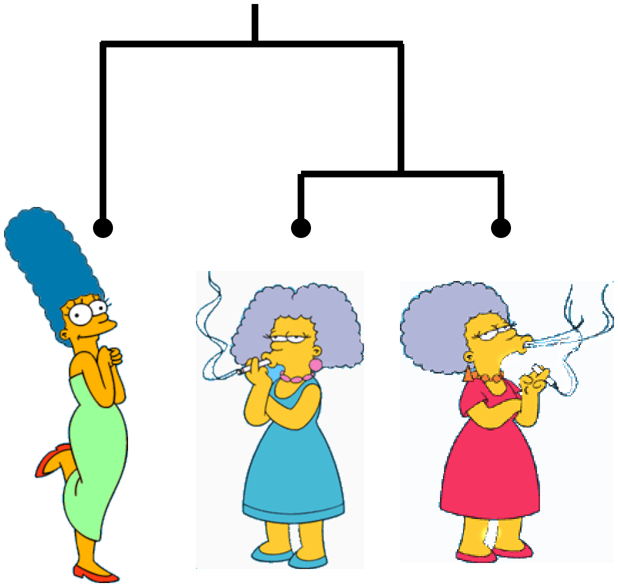


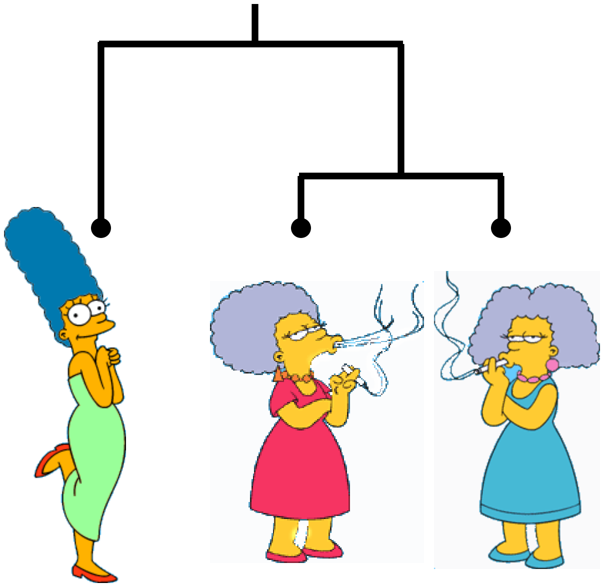
Slide based on one by Eamonn Keogh

Dendogram: A Useful Tool for Summarizing Similarity Measurements



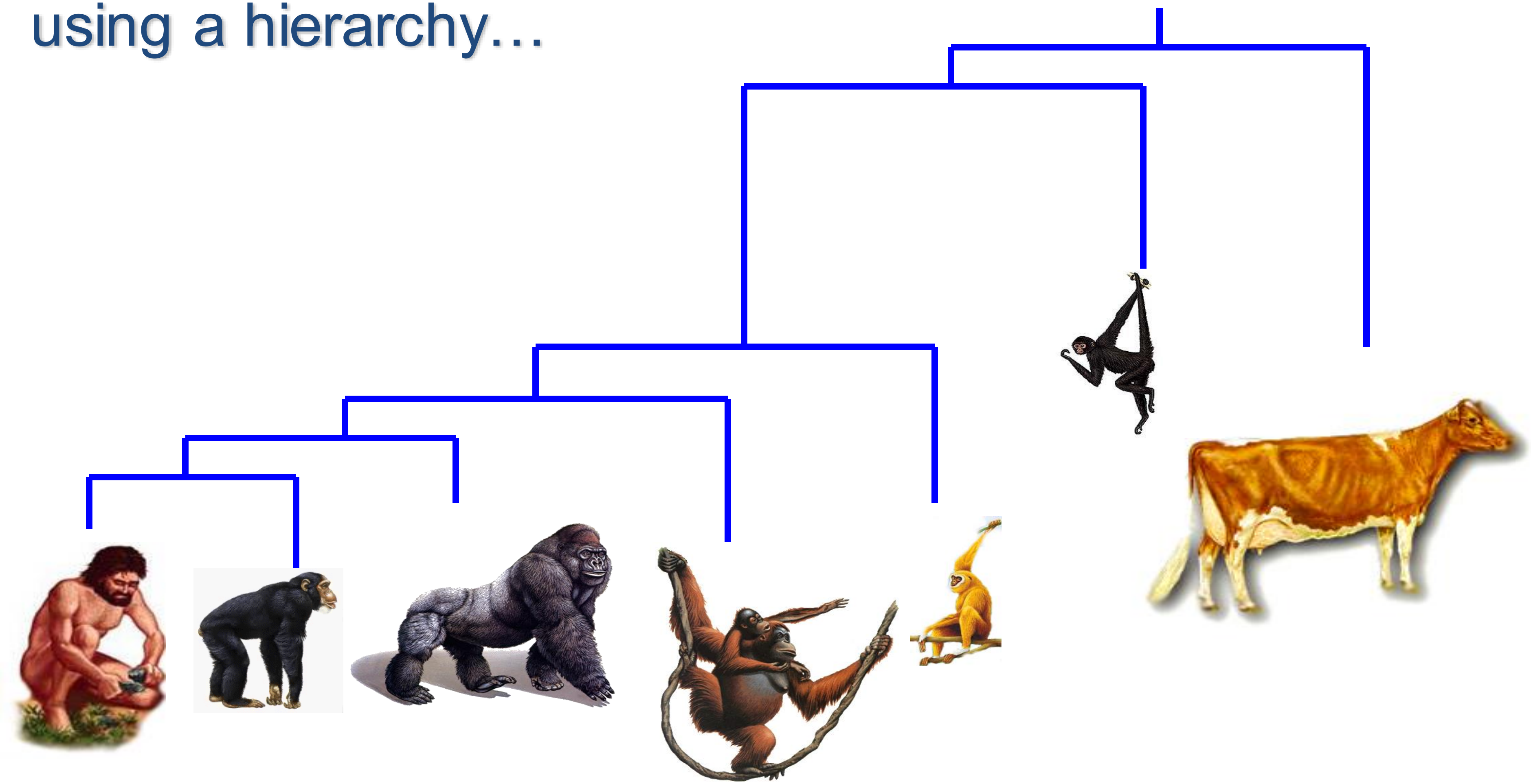
The similarity between two objects in a dendrogram is represented as the height of the lowest internal node they share.





Slide based on one by Eamonn Keogh

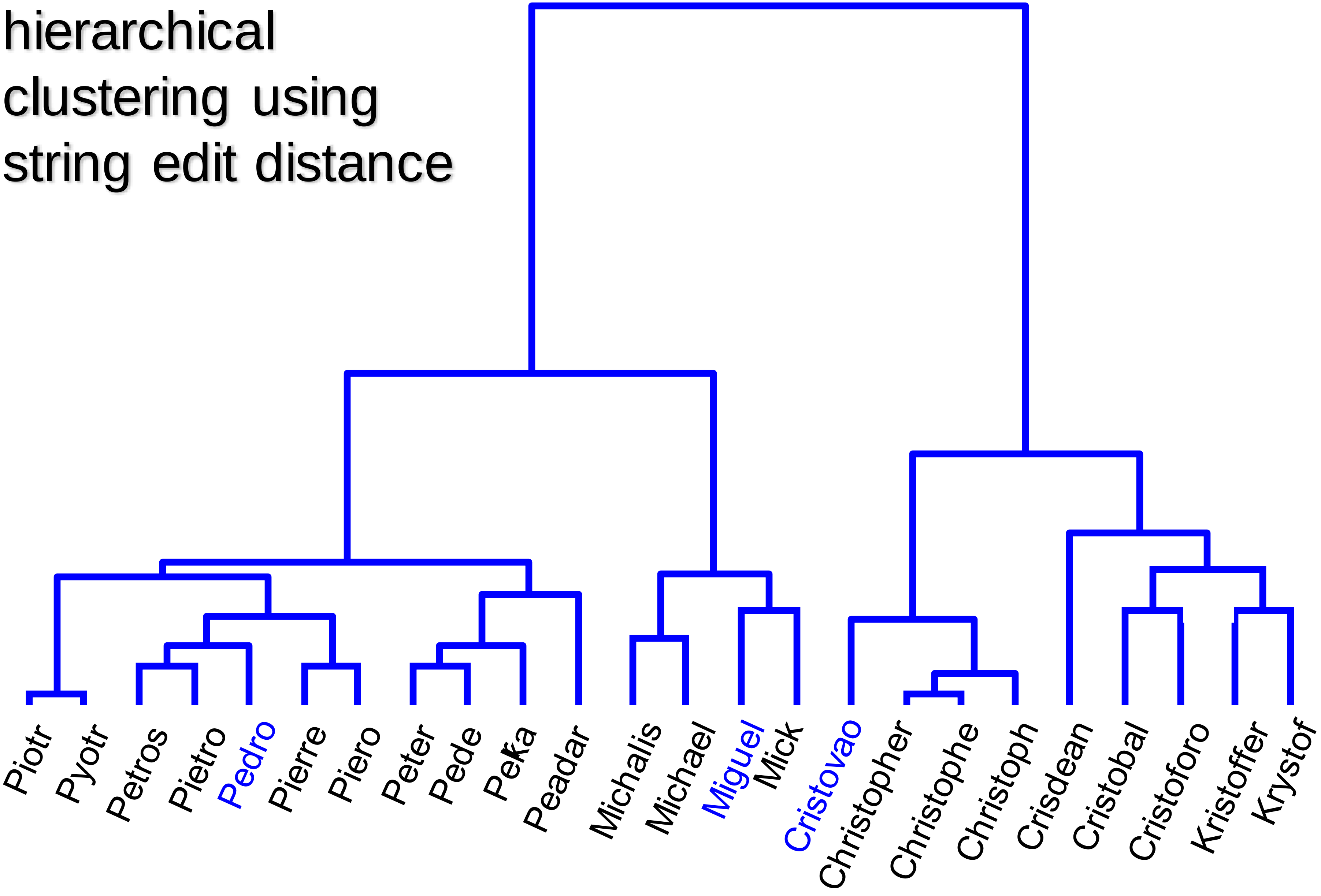
There is only one dataset that can be perfectly clustered using a hierarchy…



(Bovine:0.69395, (Spider Monkey 0.390, (Gibbon:0.36079,(Orang:0.33636,(Gorilla:0.17147,(Chimp:0.19268, Human:0.11927):0.08386):0.06124):0.15057):0.54939);

Slide based on one by Eamonn Keogh

A demonstration of hierarchical clustering using string edit distance

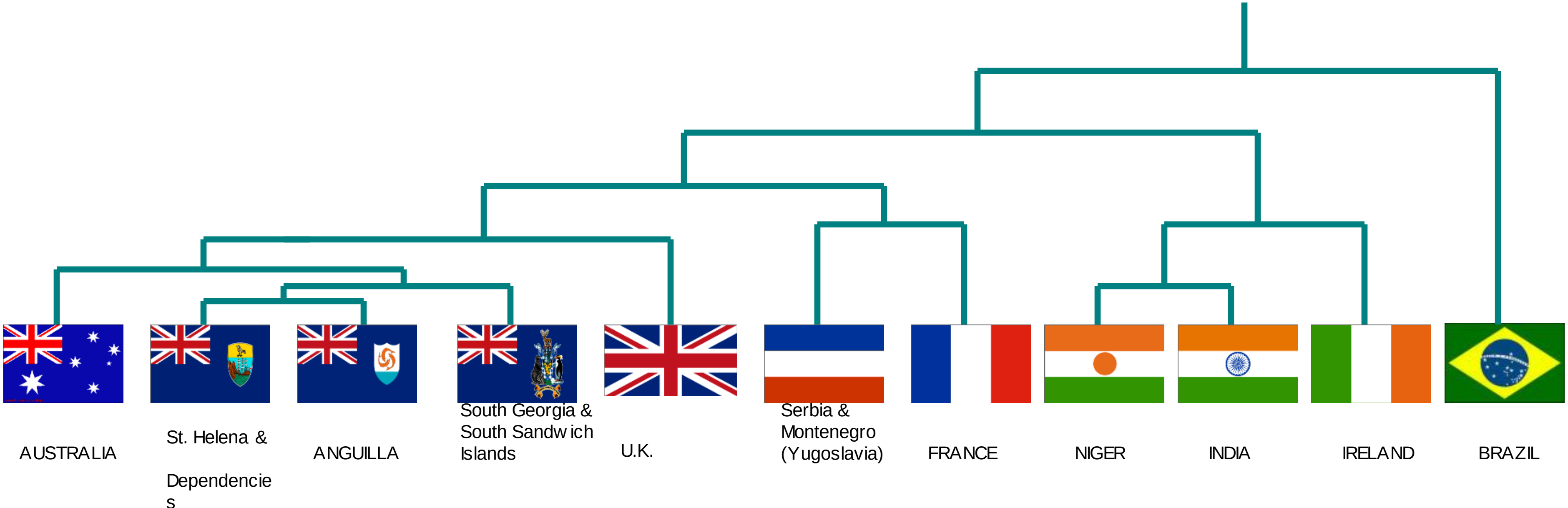


Slide based on one by Eamonn Keogh

Hierarchal clustering can sometimes show patterns that are meaningless or spurious

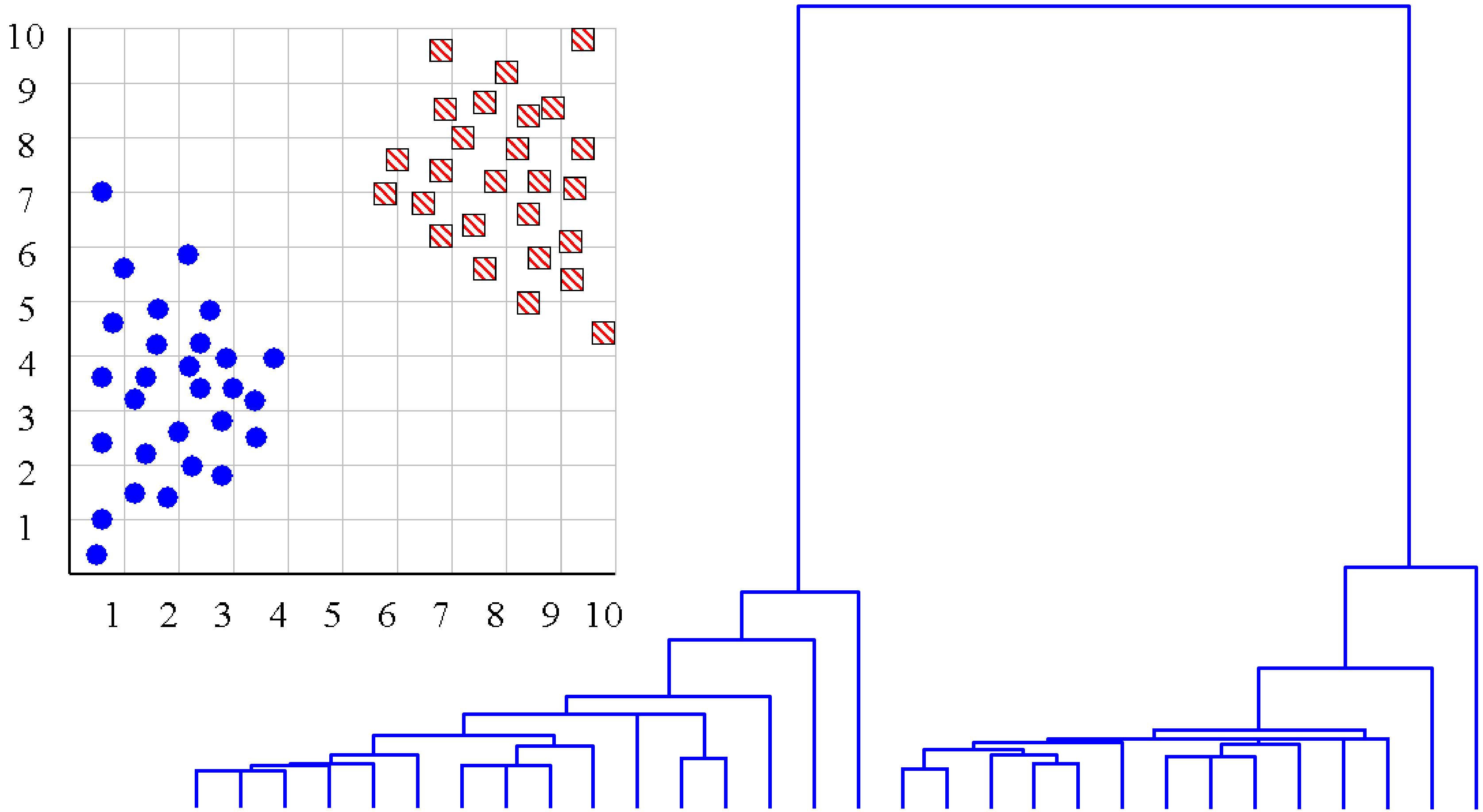
The tight grouping of Australia, Anguilla, St. Helena etc is meaningful; all these countries are former UK colonies

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



Slide based on one by Eamonn Keogh

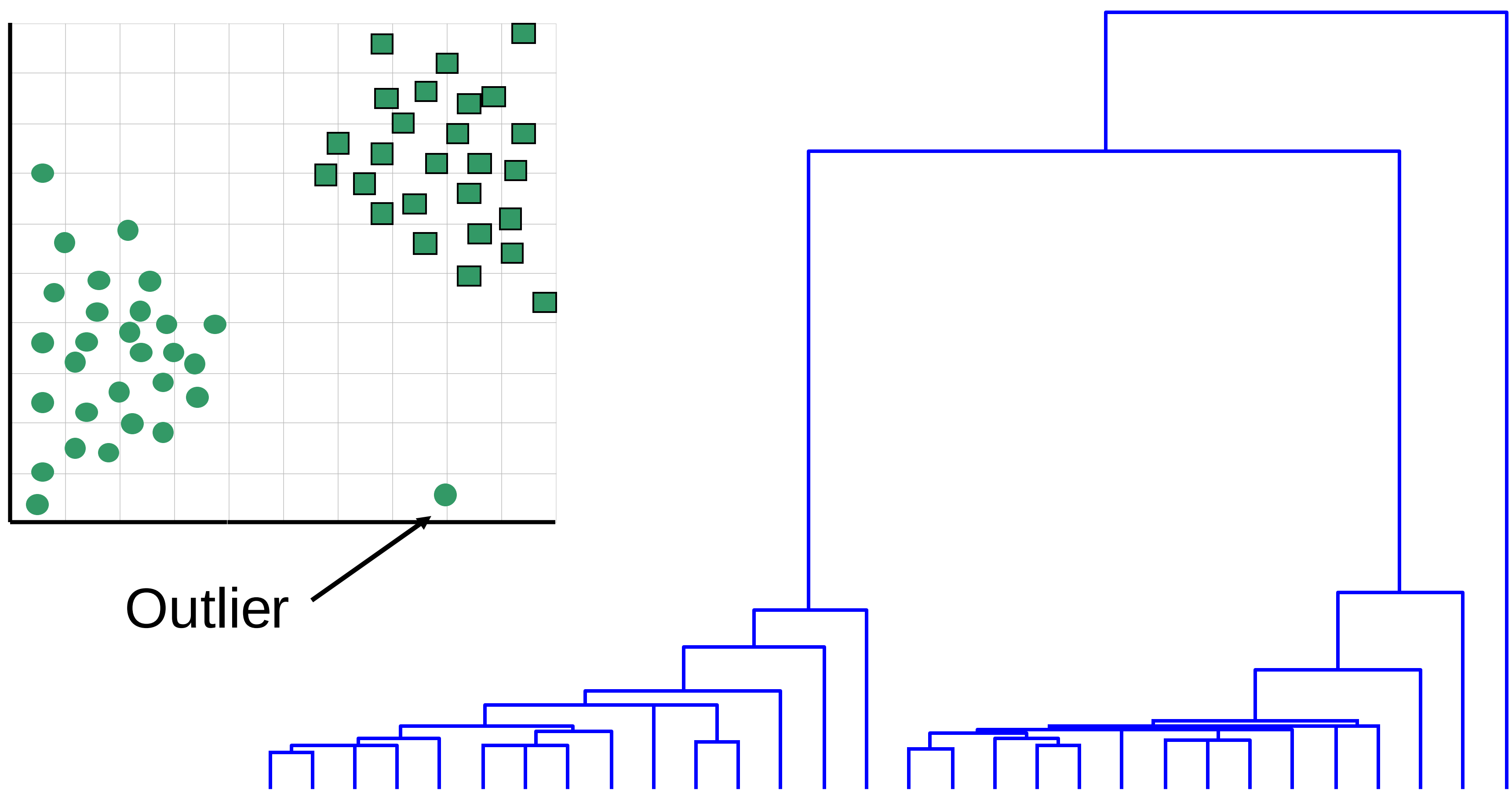
We can look at the dendrogram to determine the “correct” number of clusters.



Slide based on one by Eamonn Keogh

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



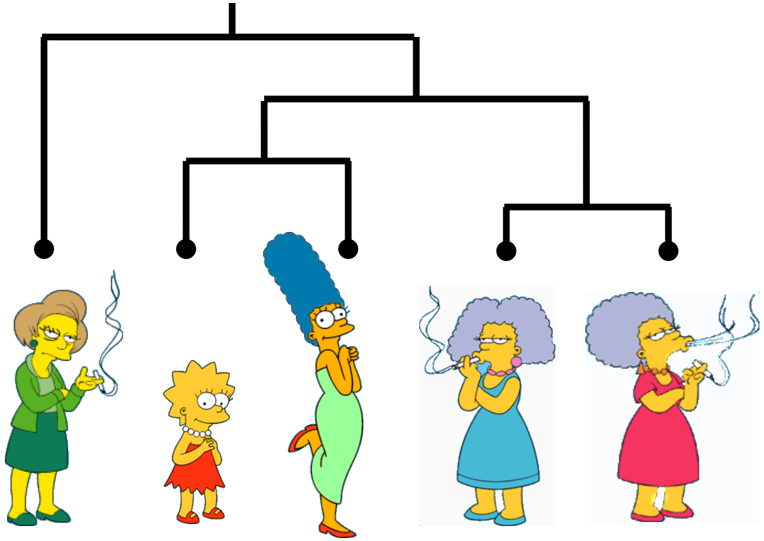
Outlier

Hierarchical Clustering

Slide based on one by Eamonn Keogh

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



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.

Slide based on one by Eamonn Keogh

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

D(,) = 8

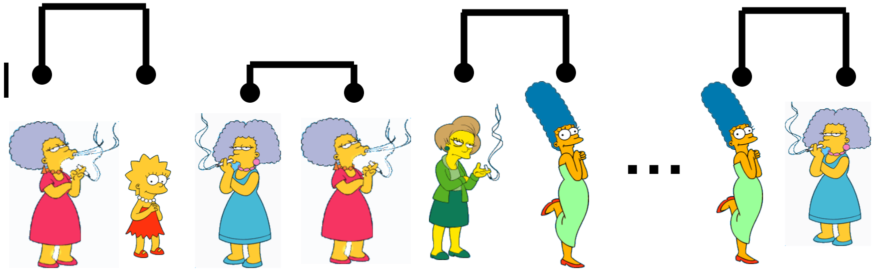
D(,) = 1

	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

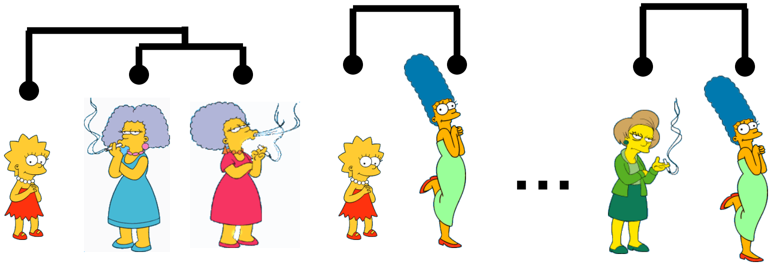
Consider all possible merges…



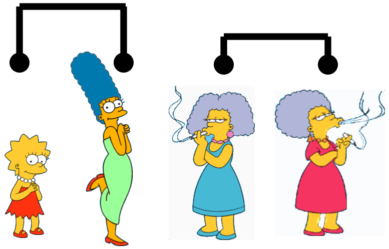
Choose the best

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.

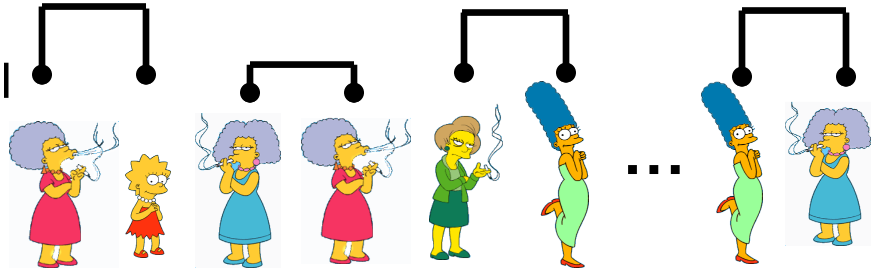
Consider all possible merges…



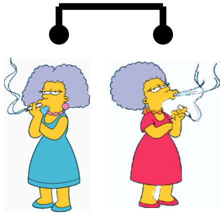
Choose the best



Consider all possible merges…

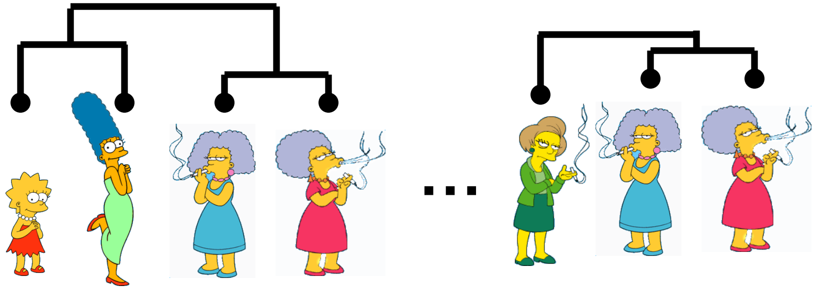


Choose the best

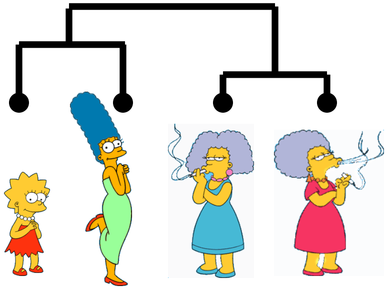


Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

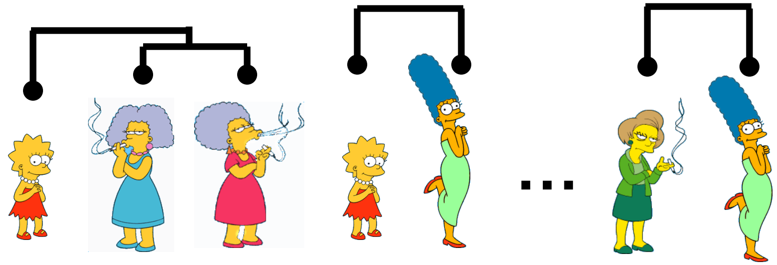
Consider all possible merges…



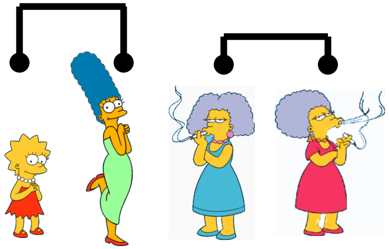
Choose the best



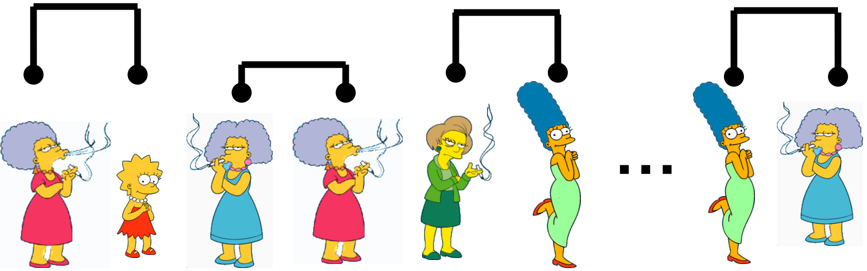
Consider all possible merges…



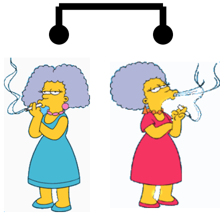
Choose the best



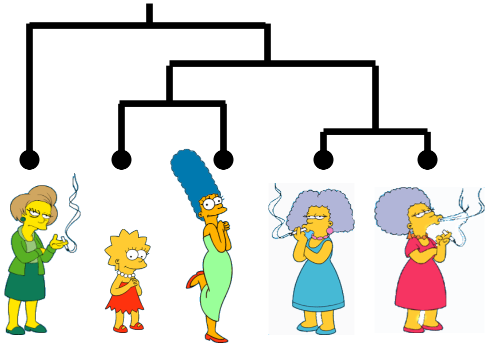
Consider all possible merges…



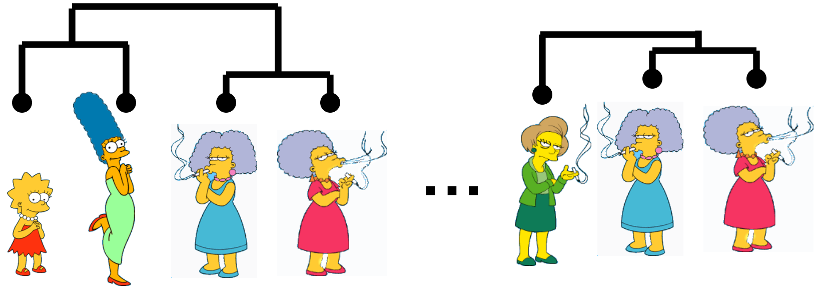
Choose the best



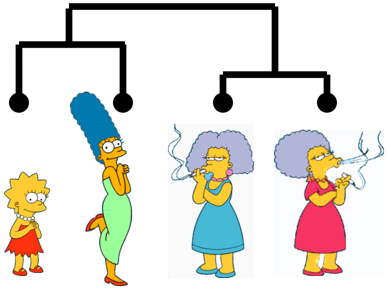
Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



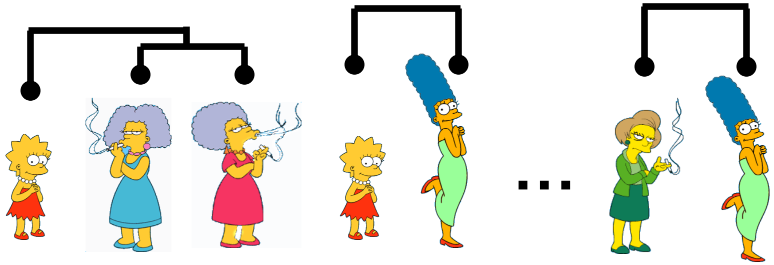
Consider all possible merges…



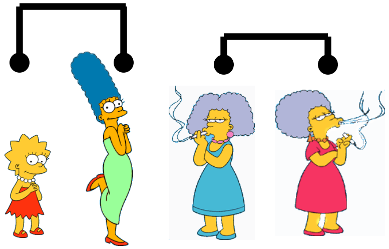
Choose the best



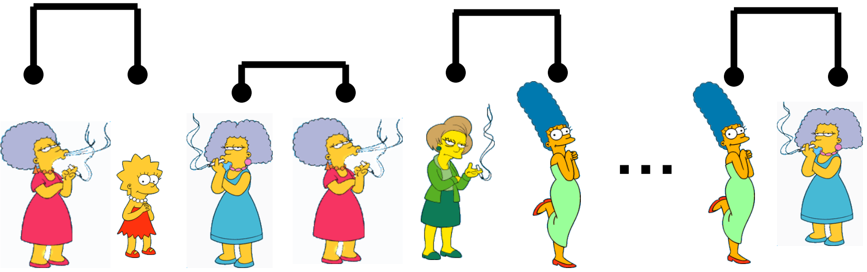
Consider all possible merges…



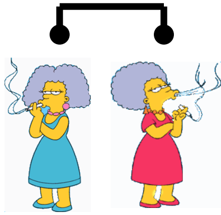
Choose the best



Consider all possible merges…



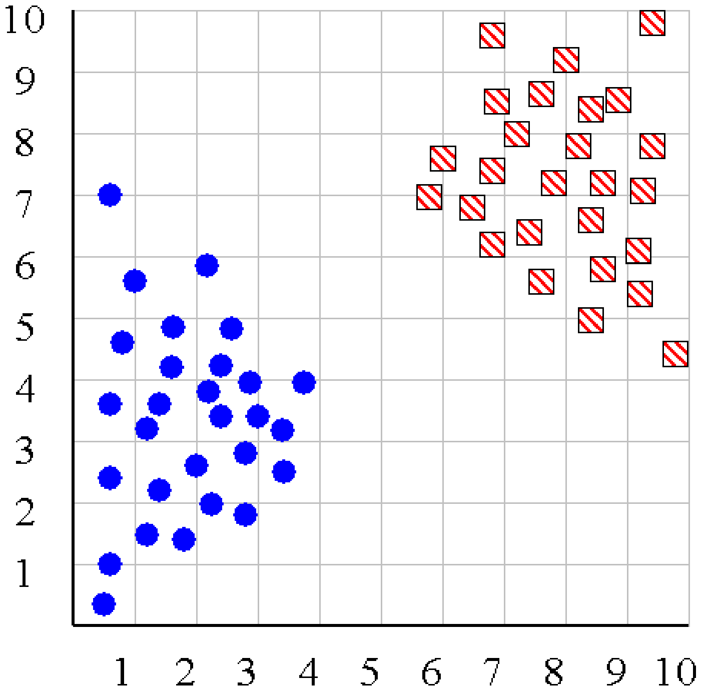
Choose the best



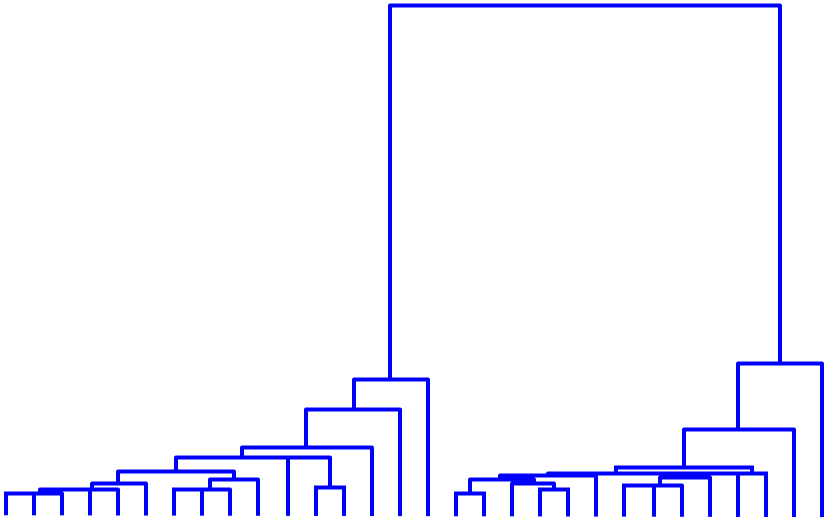
Slide based on one by Eamonn Keogh

We know how to measure the distance between two objects, but defining the distance between an object and a cluster, or defining the distance between two clusters is non obvious.

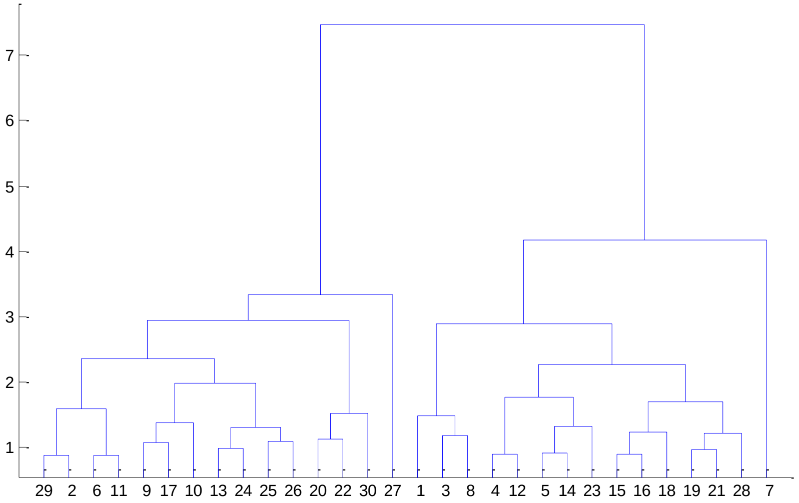
- **Single linkage (nearest neighbor):** In this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.
- **Complete linkage (furthest neighbor):** In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors").
- **Group average linkage:** In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters.



Slide based on one by Eamonn Keogh



Single linkage



Average linkage

Slide based on one by Eamonn Keogh

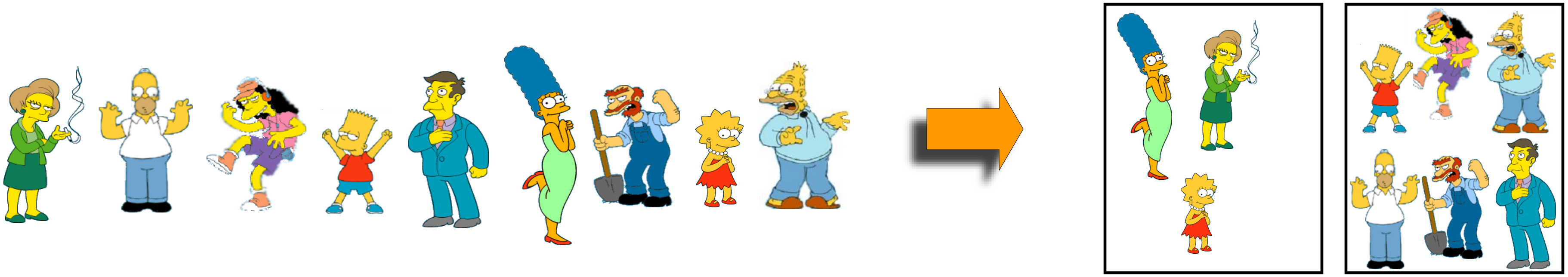
Hierarchal Clustering Methods Summary

- No need to specify the number of clusters in advance
- Hierarchal 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

Slide based on one by Eamonn Keogh

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.



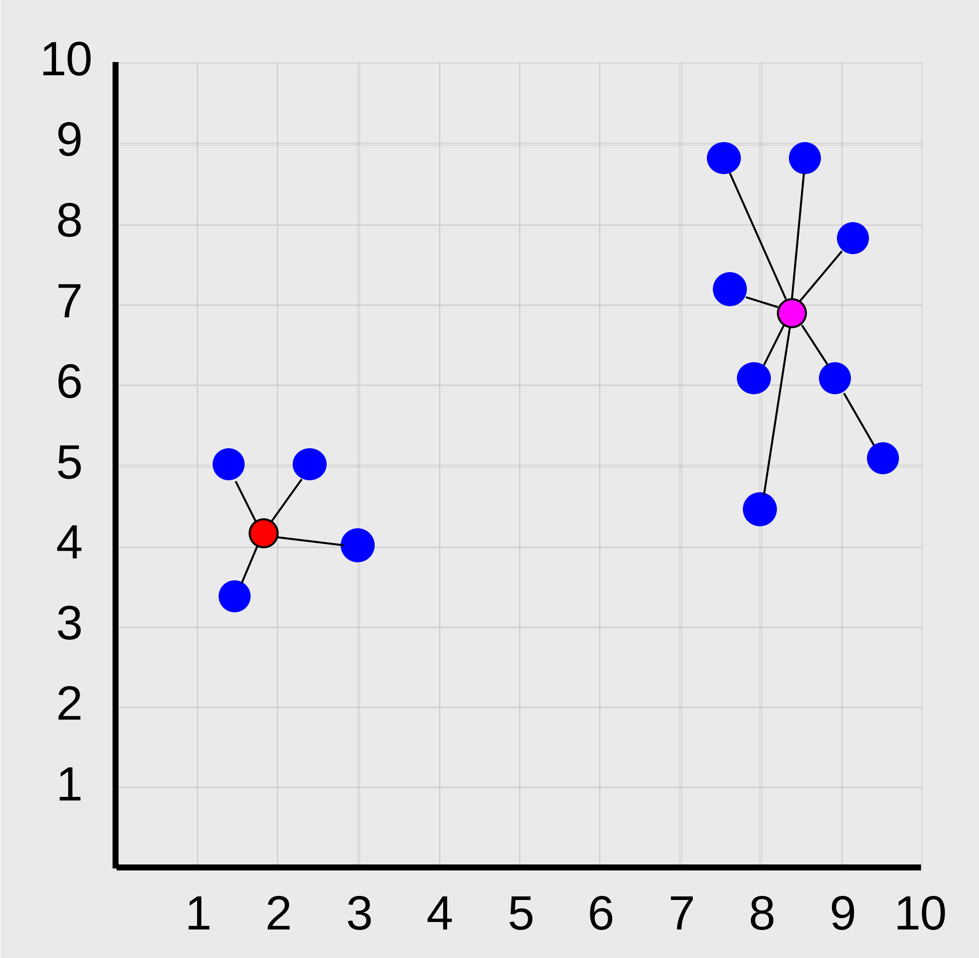
Slide based on one by Eamonn Keogh

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



Slide based on one by Eamonn Keogh

Partition Algorithm 1: k-means

1. Decide on a value for *k*.

2. Initialize the *k* cluster centers (randomly, if necessary).

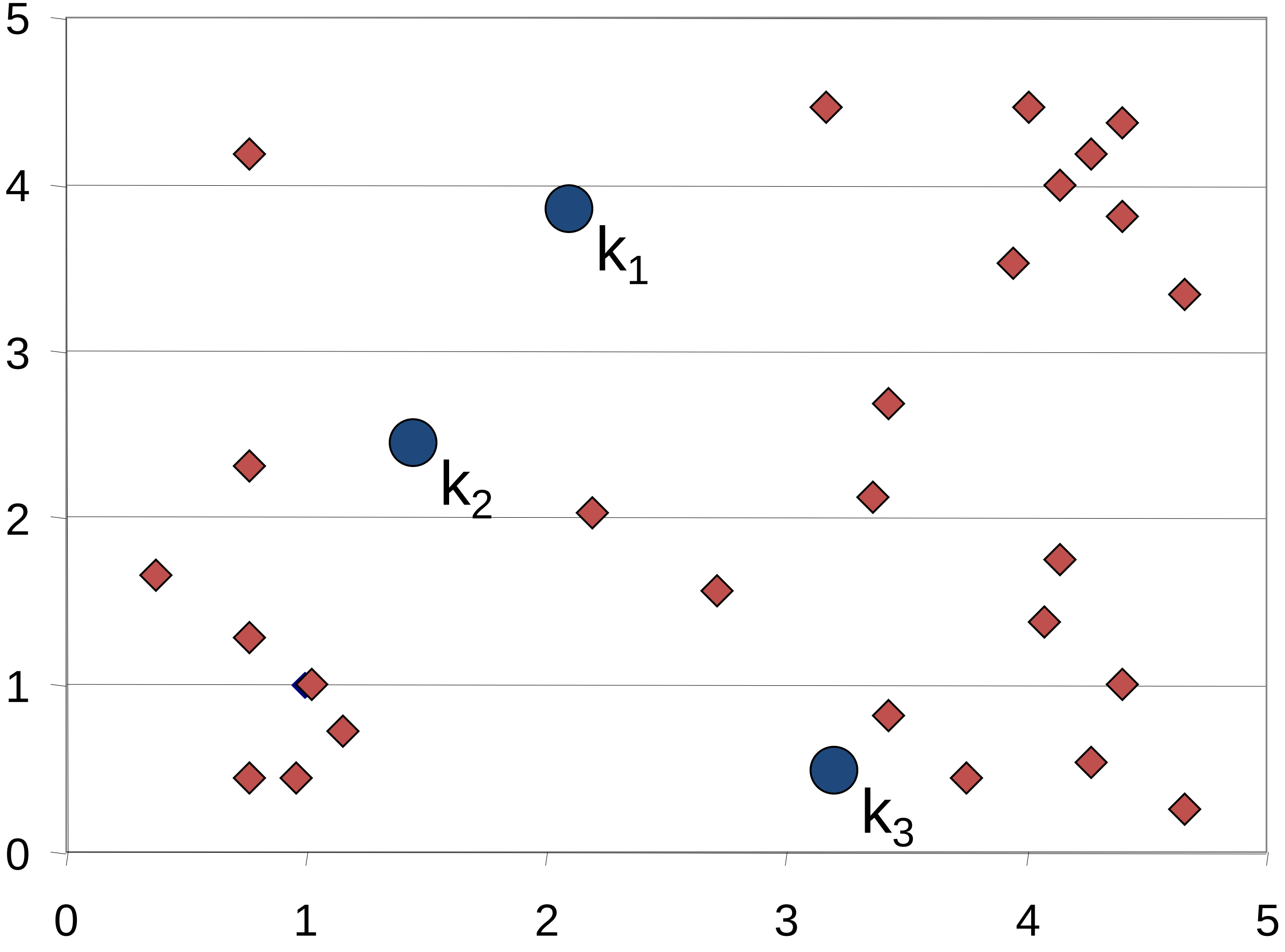
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.

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

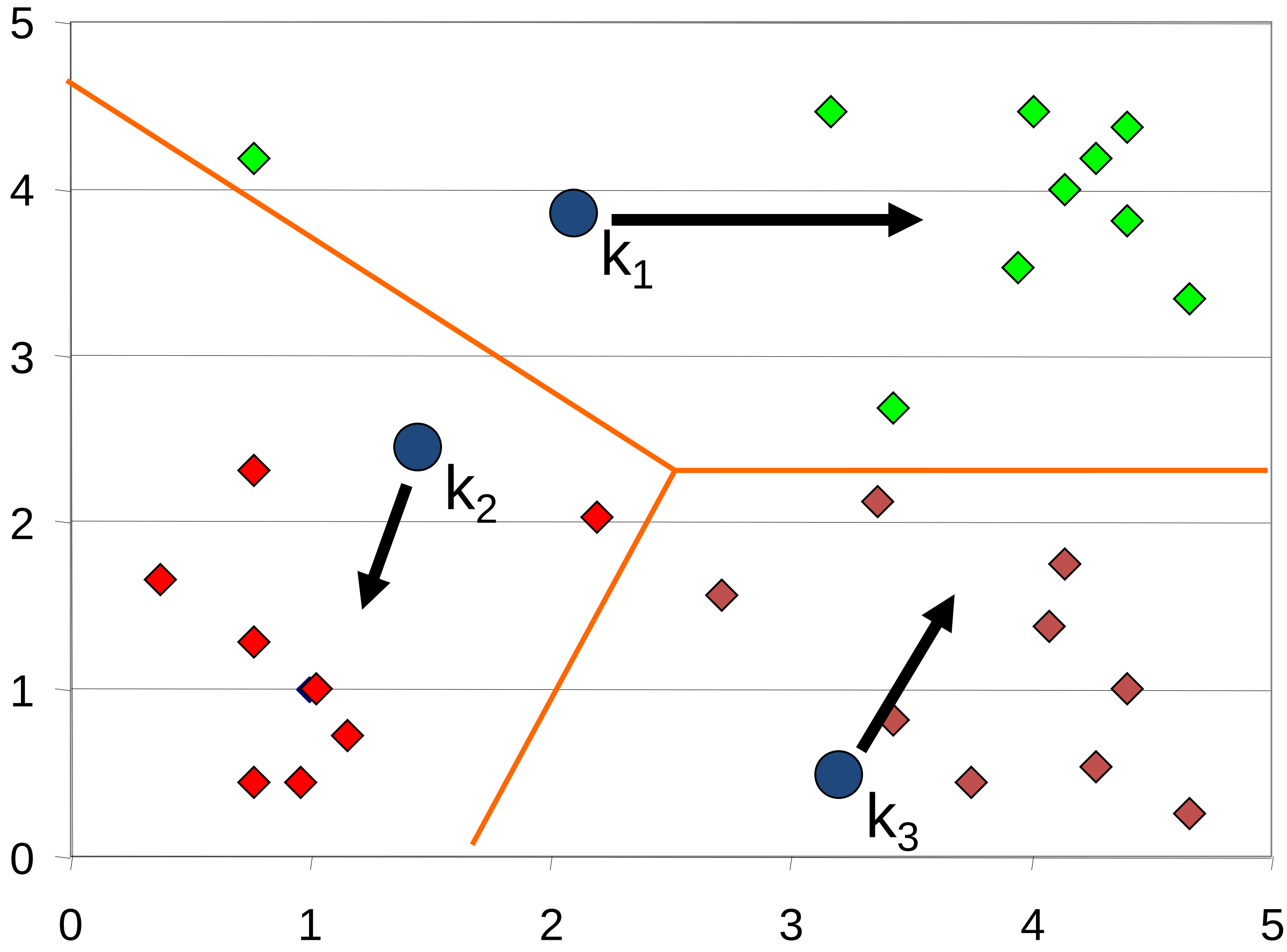
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 2

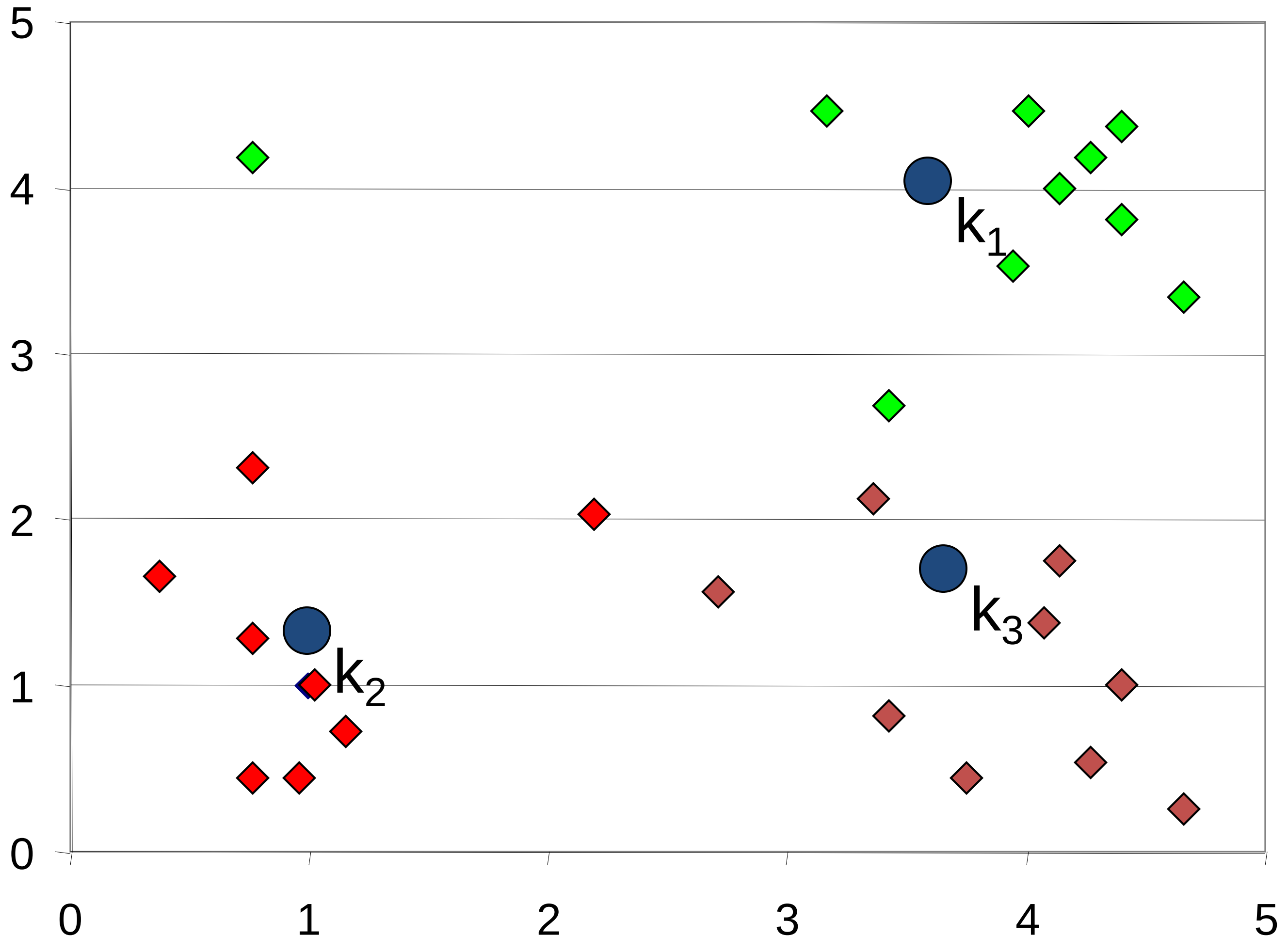
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 3

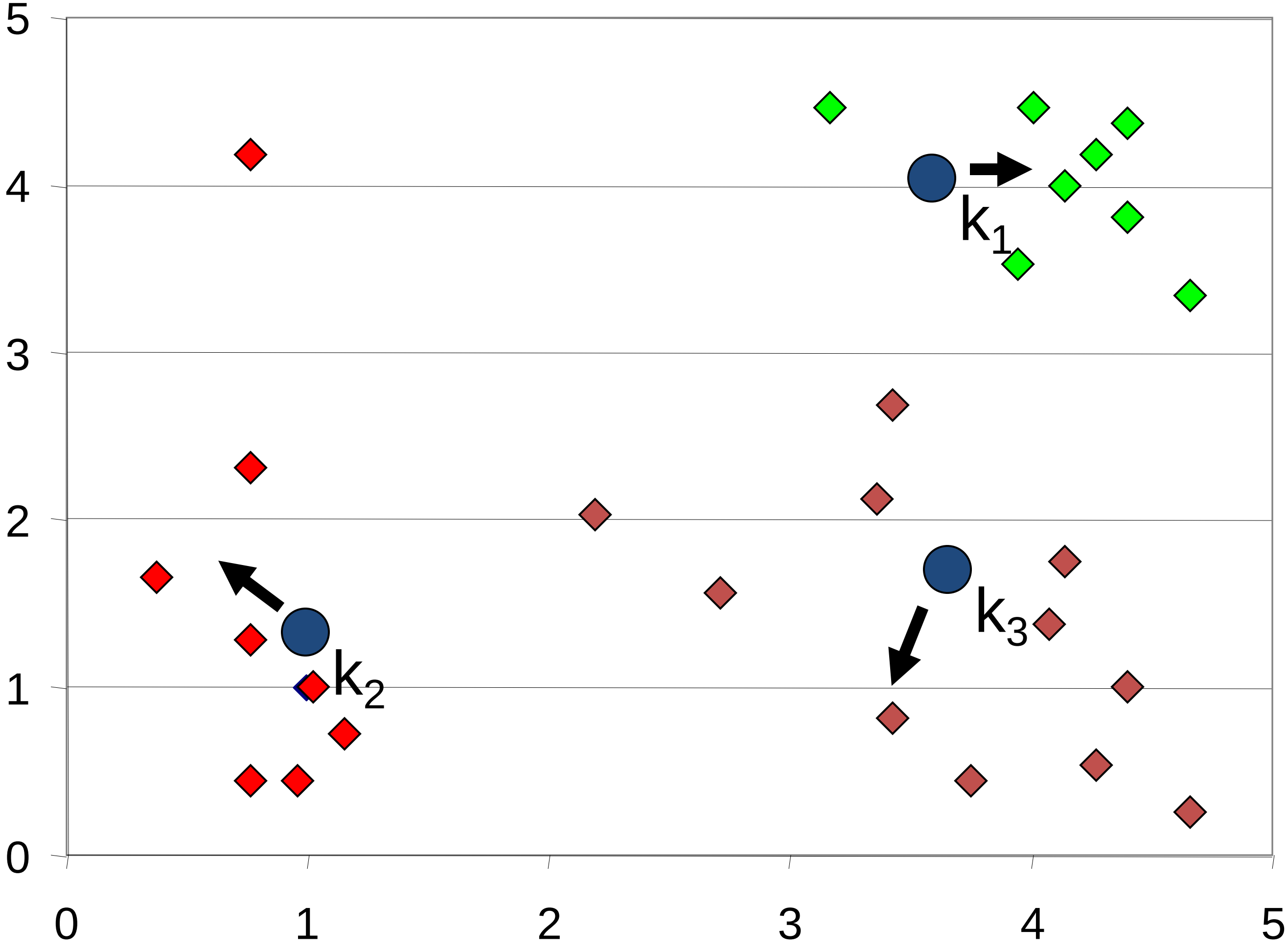
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 4

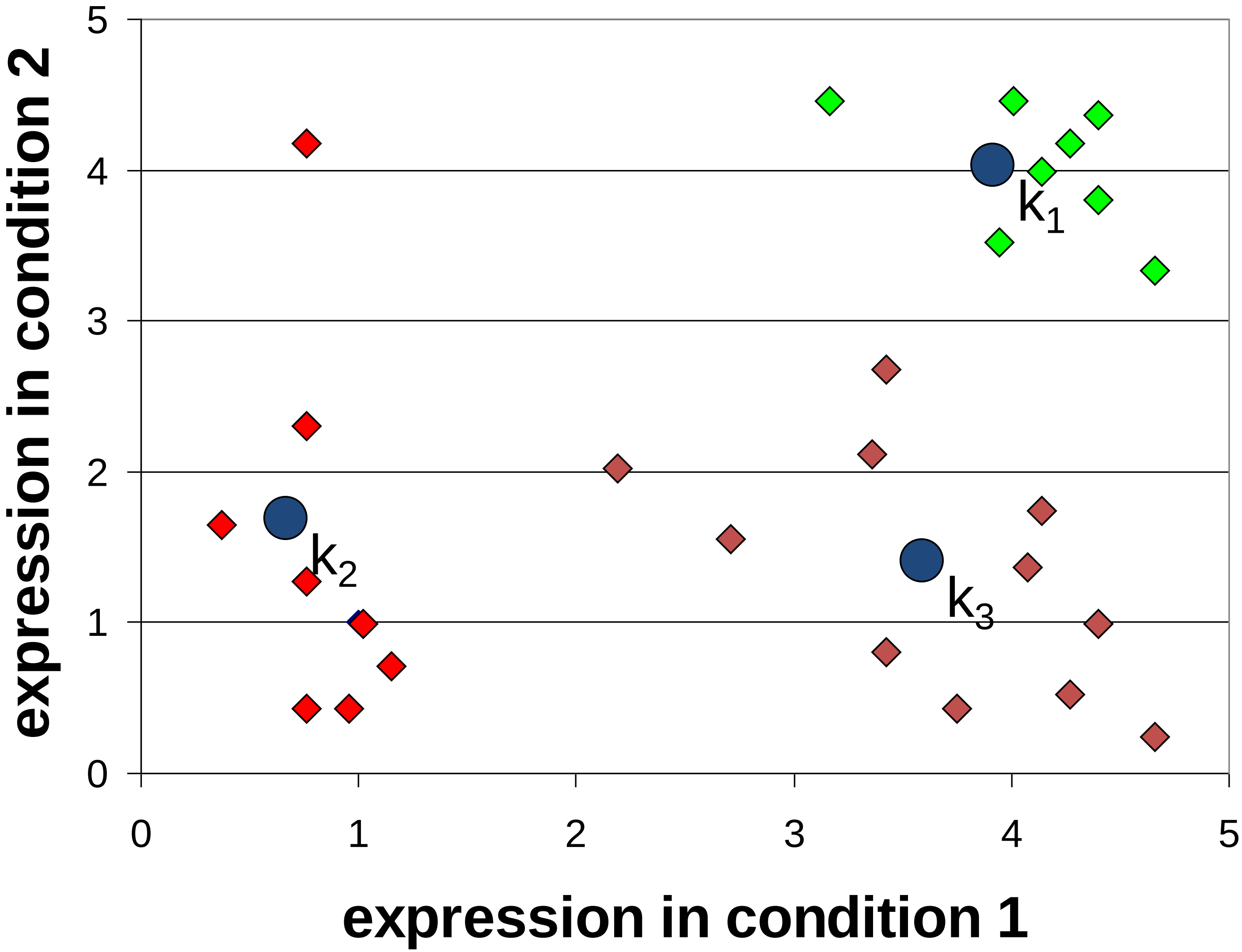
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



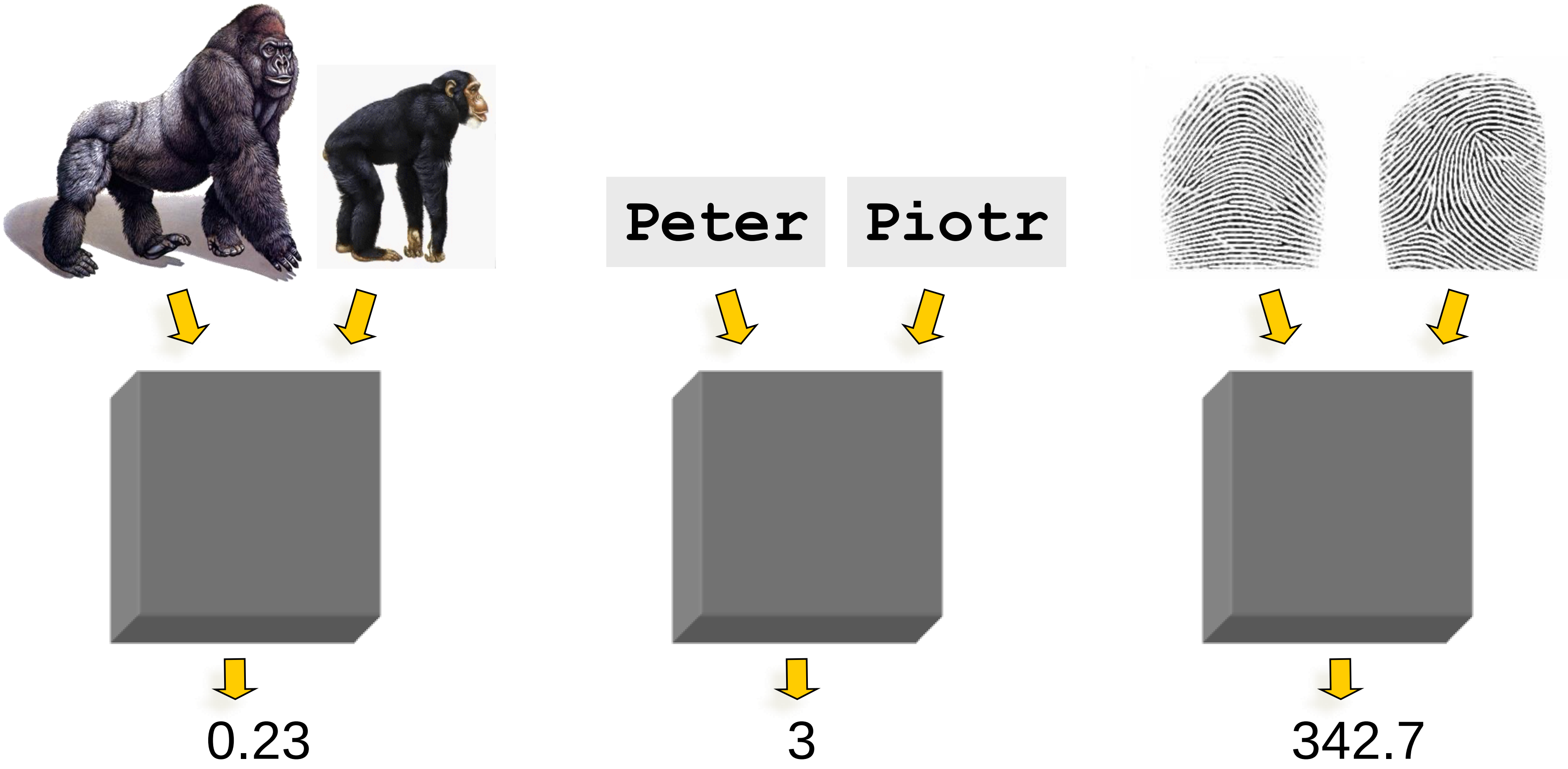
Slide based on one by Eamonn Keogh

Comments on k-Means

- Strengths
  - *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, $t << n$.
  - Often terminates at a local optimum.
- Weakness
  - Applicable only when mean is defined, then what about categorical data?
  - 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

Slide based on one by Eamonn Keogh

How do we measure similarity?



Peter Piotr

0.23

3

342.7

Slide based on one by Eamonn Keogh

A generic technique for measuring similarity

To measure the similarity between two objects, transform one into the other, and measure how much effort it took. The measure of effort becomes the distance measure.

The distance between Patty and Selma:

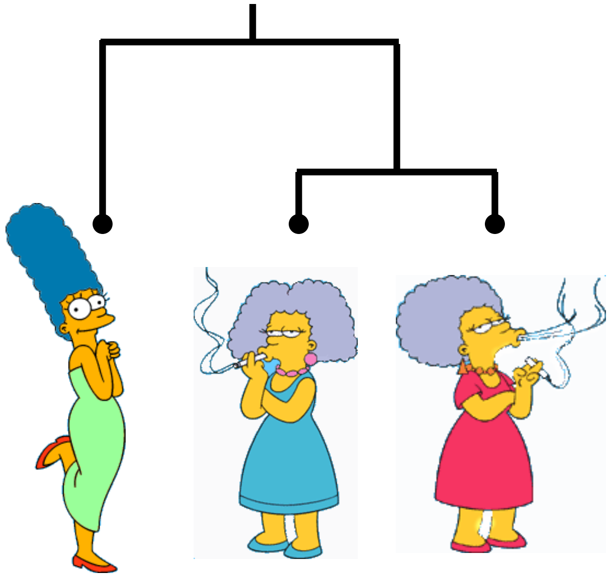
Change dress color,	1 point
Change earring shape,	1 point
Change hair part,	1 point

D(Patty,Selma) = **3**

The distance between Marge and Selma:

Change dress color,	1 point
Add earrings,	1 point
Decrease height,	1 point
Take up smoking,	1 point
Lose weight,	1 point

D(Marge,Selma) = **5**



This is called the “edit distance” or the “transformation distance”

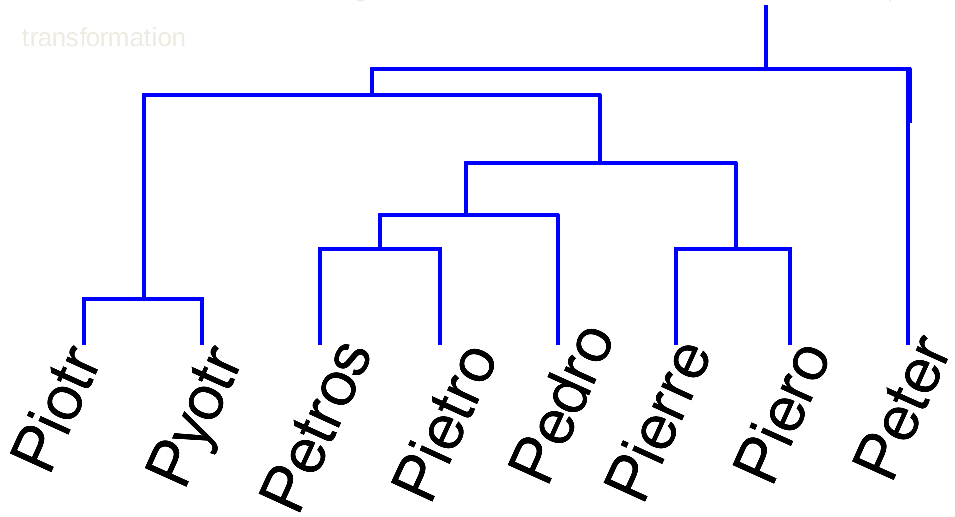
Slide based on one by Eamonn Keogh

Edit Distance Example

It is possible to transform any string *Q* into string *C*, using only *Substitution*, *Insertion* and *Deletion*.
Assume that each of these operators has a cost associated with it.

The similarity between two strings can be defined as the cost of the cheapest transformation from *Q* to *C*.

Note that for now we have ignored the issue of how we can find this cheapest transformation



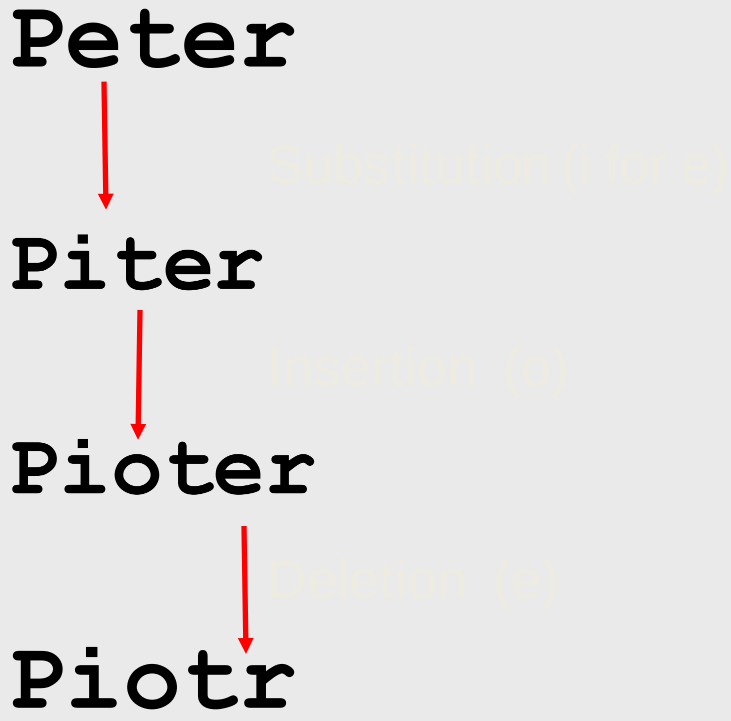
Slide based on one by Eamonn Keogh

How similar are the names "Peter" and "Piotr"?

Assume the following cost function

Substitution	1 Unit
Insertion	1 Unit
Deletion	1 Unit

D(**Peter**,**Piotr**) is 3


Peter

Substitution (i for e)

Piter

Insertion (o)

Pioter

Deletion (e)

Piotr

What distance metric did k-means use?

What assumptions is it making about the data?

Partition Algorithm 2: Using a Euclidean Distance Threshold to Define Clusters

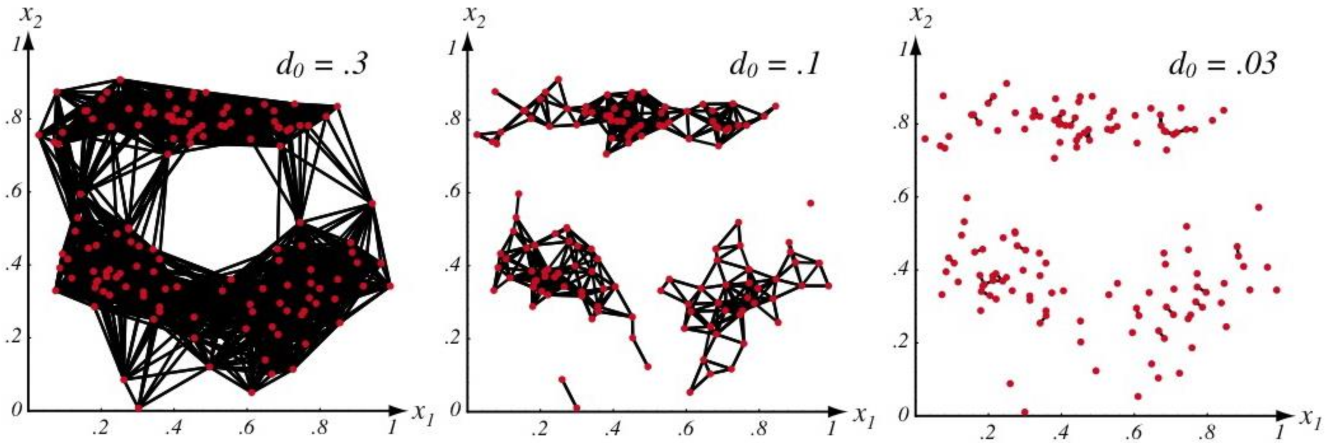


FIGURE 10.7. The distance threshold affects the number and size of clusters in similarity based clustering methods. For three different values of distance d_0, lines are drawn between points closer than d_0—the smaller the value of d_0, the smaller and more numerous the clusters. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

But should we use Euclidean Distance?

Good if data is isotropic and spread evenly along all directions

Not invariant to linear transformations, or any transformation that distorts distance relationships

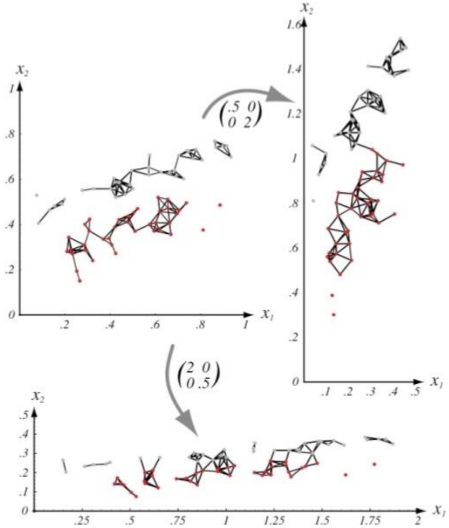
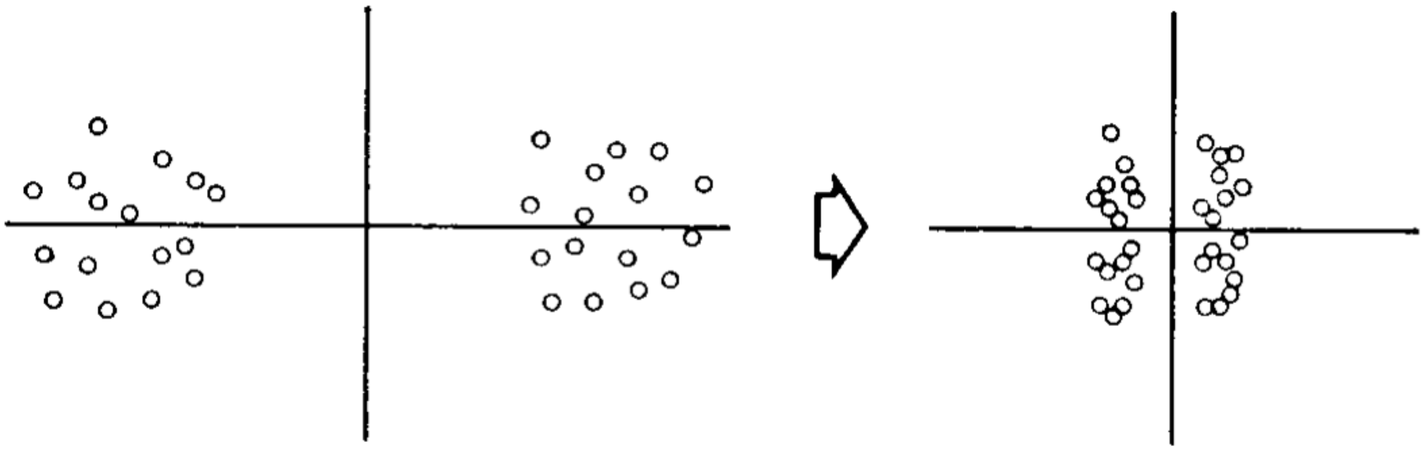


FIGURE 10.8. Scaling axes affects the clusters in a minimum distance cluster method. The original data and minimum-distance clusters are shown in the upper left; points in one cluster are shown in red, while the others are shown in gray. When the vertical axis is expanded by a factor of 2.0 and the horizontal axis shrunk by a factor of 0.5, the clustering is altered (as shown at the right). Alternatively, if the vertical axis is shrunk by a factor of 0.5 and the horizontal axis is expanded by a factor of 2.0, smaller more numerous clusters result (shown at the bottom). In both these scaled cases, the assignment of points to clusters differ from that in the original space. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Is normalization desirable?



(a) UNNORMALIZED

(b) NORMALIZED

Other distance/similarity measures

Distance between two instances x and x’, where q >= 1 is a selectable parameter and d is the number of attributes (called the Minkowski Metric)

$$d(x,x') = (\sum_{j=1}^{d} \left| x_j - x'_j \right|^q)^{1/q}$$

Cosine of the angle between two vectors (instances) gives a similarity function:

$$s(x,x') = \frac{x^t x'}{\|x\| \|x'\|}$$

Other distance/similarity measures

Distance between two instances x and x’, where q >= 1 is a selectable parameter and d is the number of attributes (called the Minkowski Metric)

$$d(x,x') = \left(\sum_{j=1}^{d} \left|x_j - x'_j\right|^q\right)^{1/q}$$

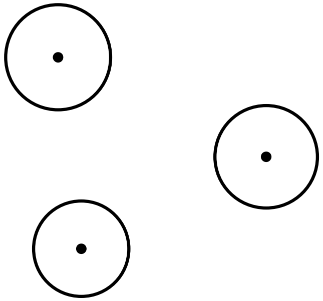
Cosine of the angle between two vectors
a similarity function:

$$s(x,x') = \frac{x^t x'}{\|x\|\|x'\|}$$

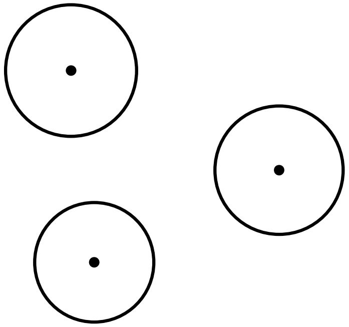
When features are binary this becomes the number of attributes shared by x and x’ divided by the geometric mean of the number of attributes in x and the number in x’. A simplification of this is:

$$s(x,x') = \frac{x^t x'}{d}$$

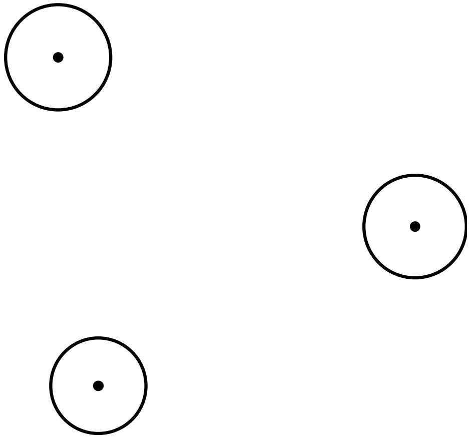
What is a good clustering?



(a)

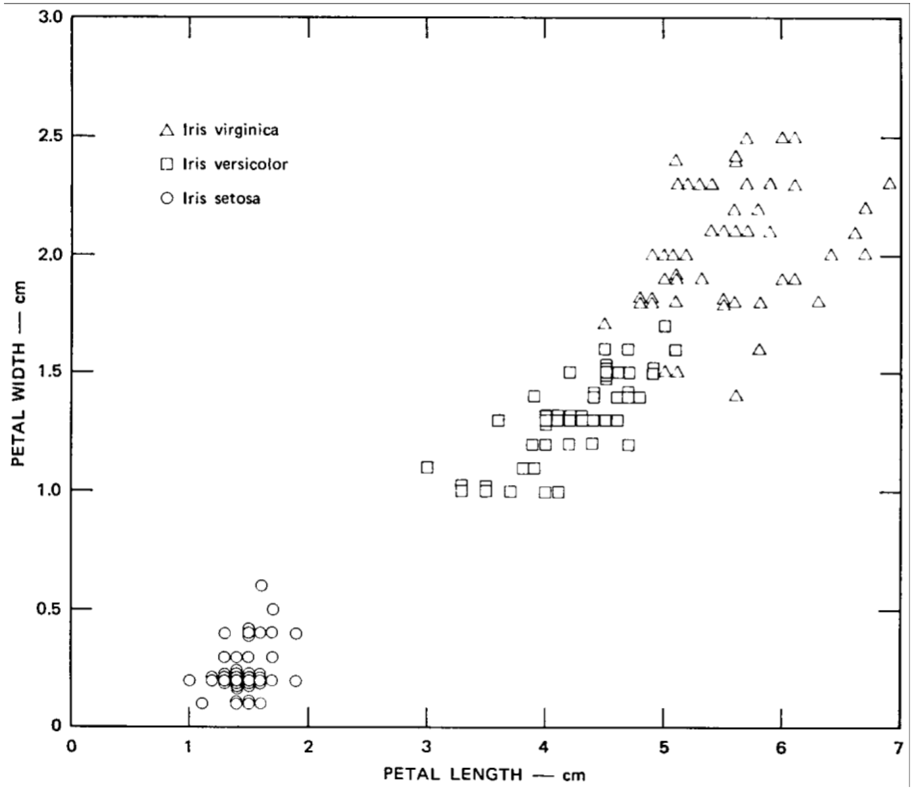


(b)

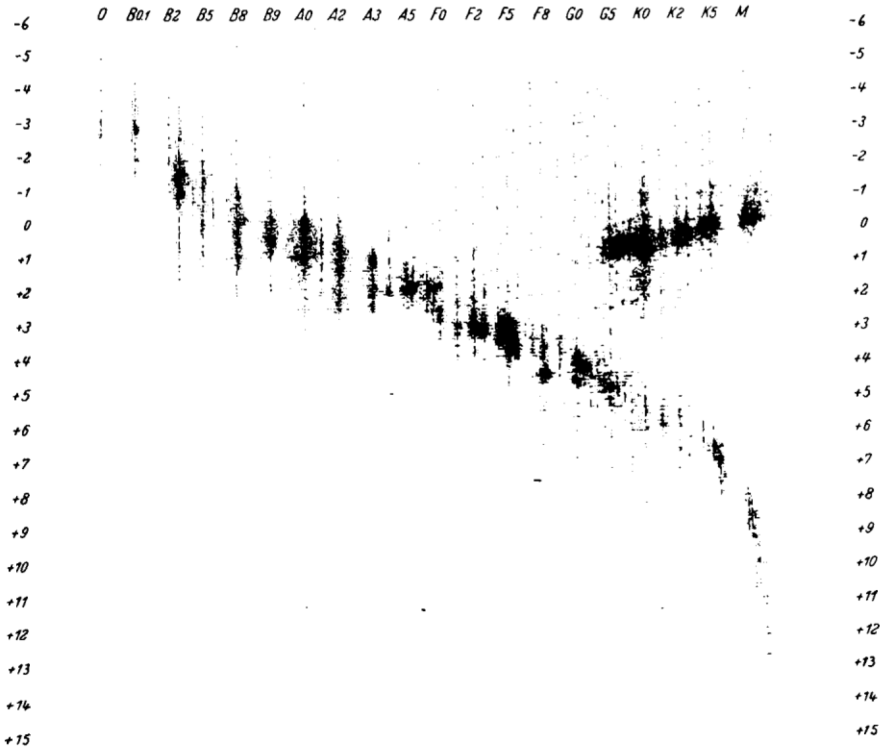


(c)

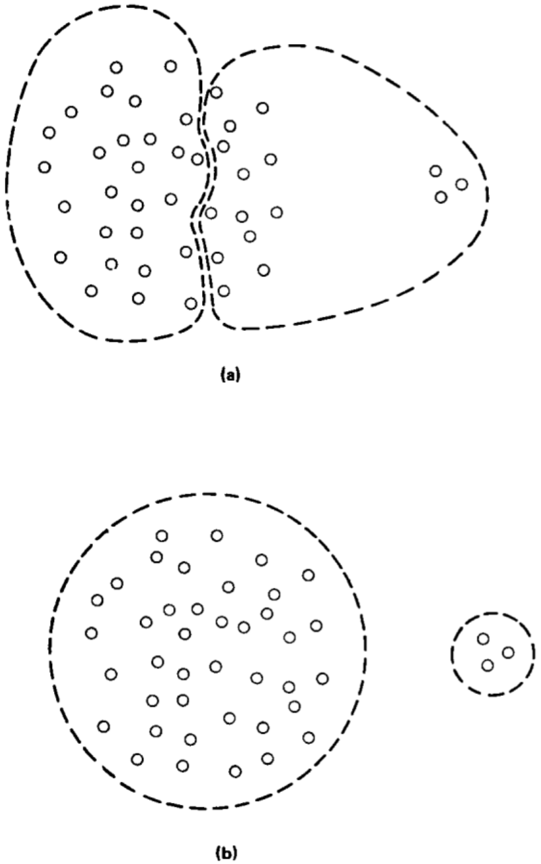
Clustering Criteria: Sum of Squared Error



A Dataset for which SSE is not a good criterion.



How does cluster size impact performance?



Scattering Criteria – on board

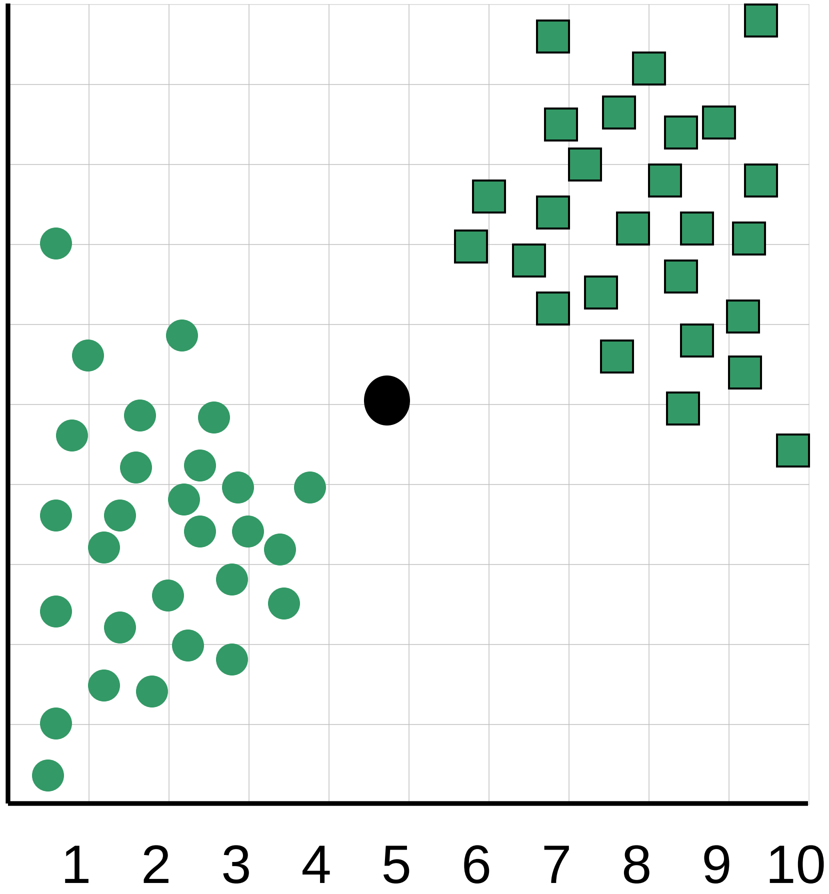
To apply partitional clustering we need to:

- Select features to characterize the data
- Collect representative data
- Choose a clustering algorithm
- Specify the number of clusters

Um, what about k?

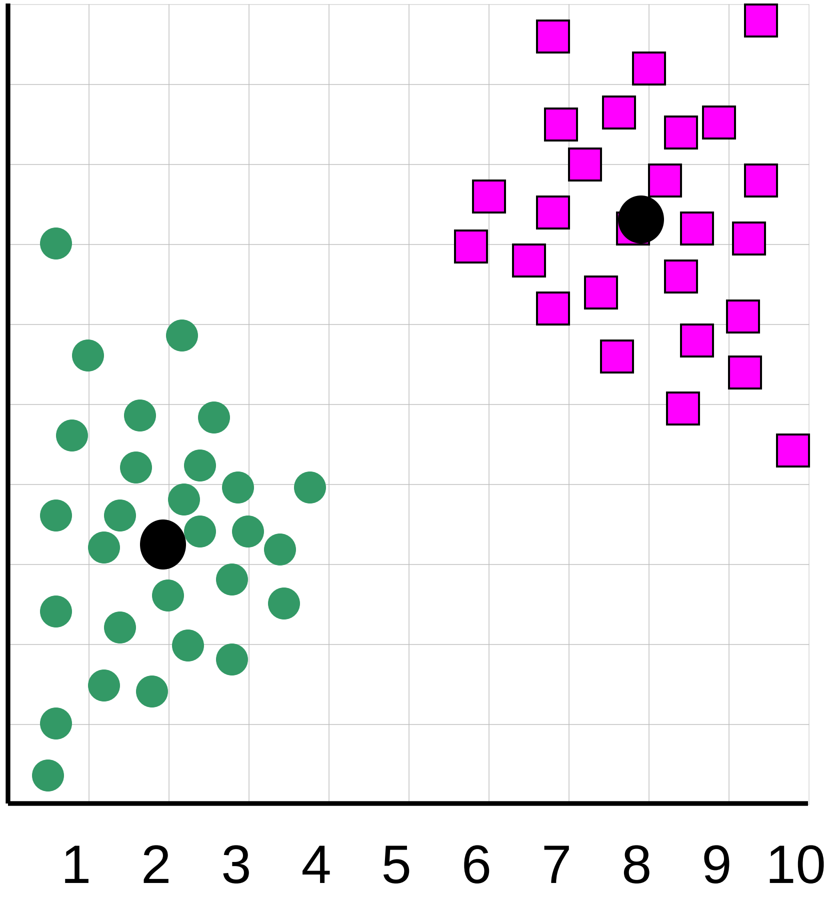
- Idea 1: Use our new trick of cross validation to select k
  - What should we optimize? SSE? Trace?
  - Problem?
- Idea 2: Let our domain expert look at the clustering and decide if they like it
  - How should we show this to them?
  - Problem?
- Idea 3: The “knee” solution

When k = 1, the objective function is 873.0



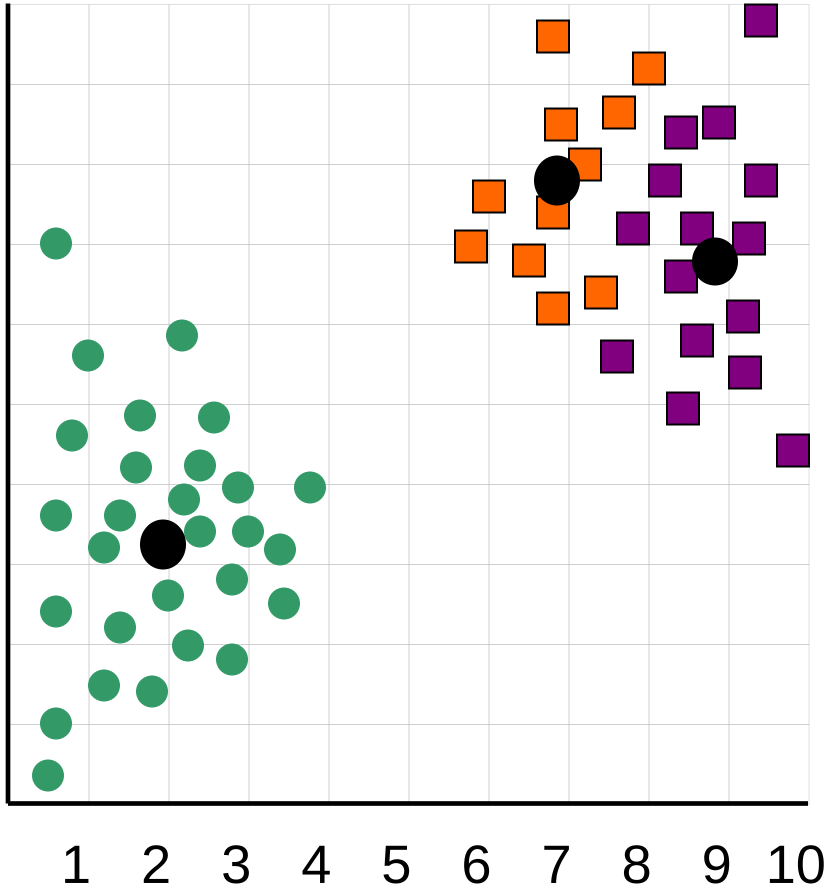
Slide based on one by Eamonn Keogh

When k = 2, the objective function is 173.1



Slide based on one by Eamonn Keogh

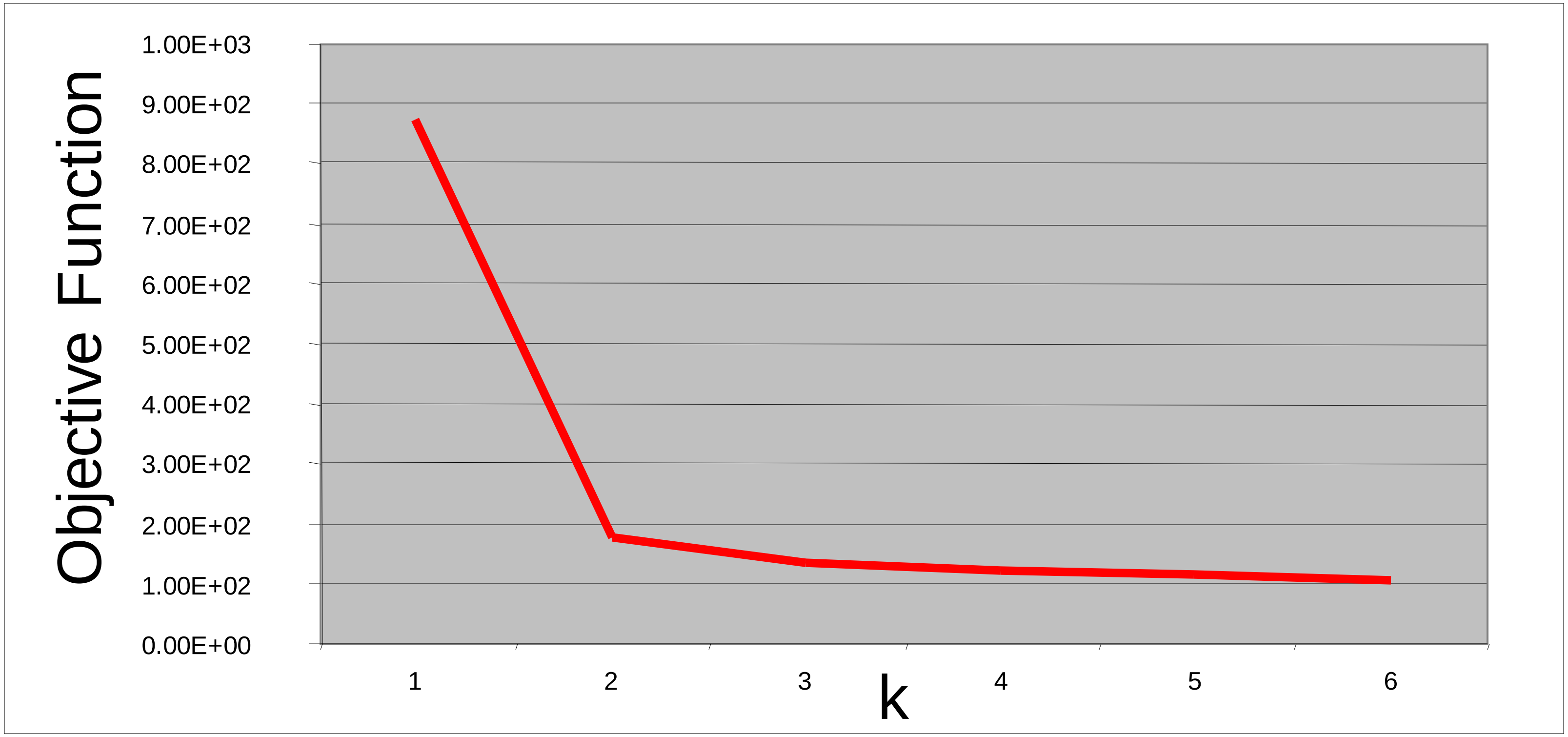
When k = 3, the objective function is 133.6



Slide based on one by Eamonn Keogh

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”.



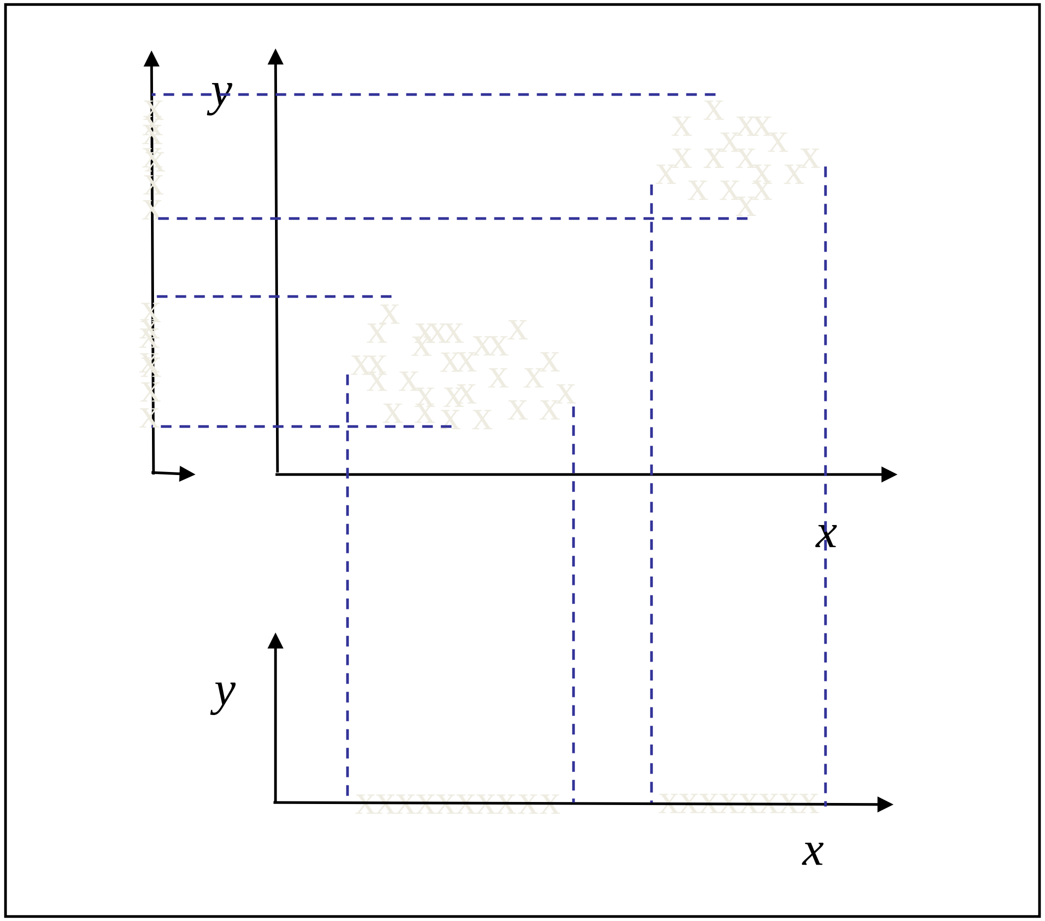
Slide based on one by Eamonn Keogh

High-Dimensional Data poses Problems for Clustering

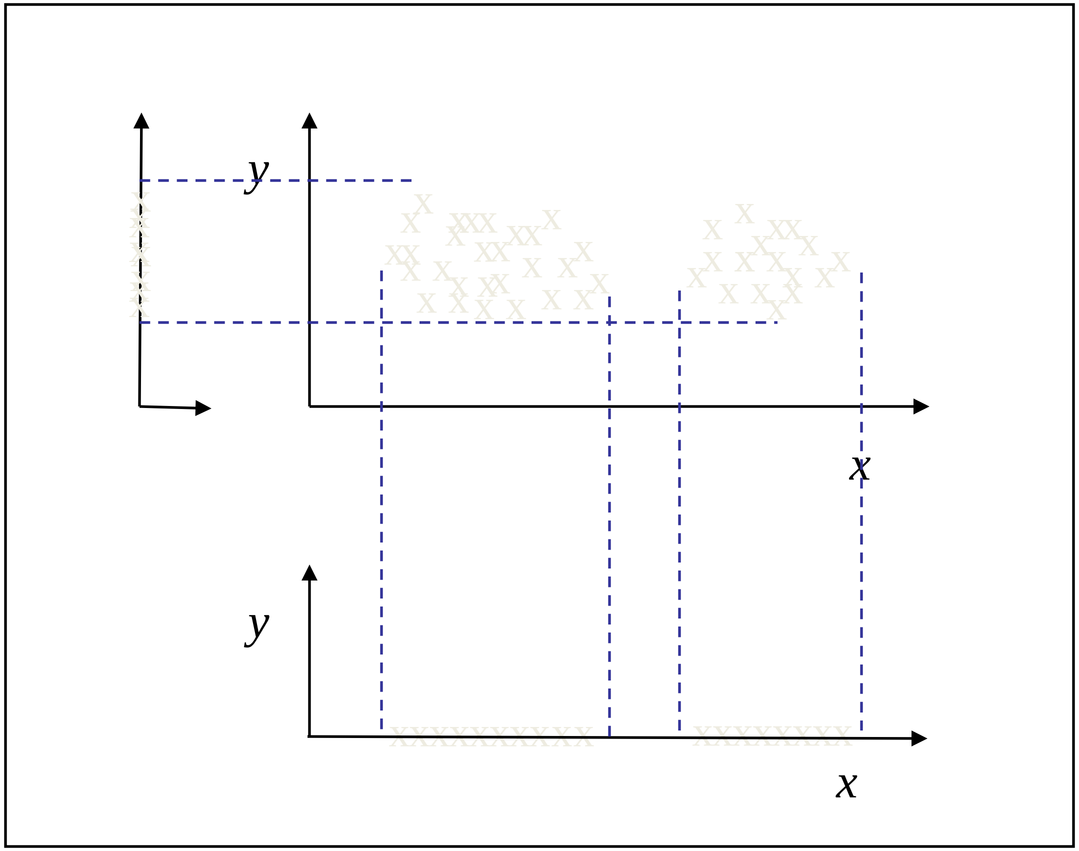
- Difficult to find true clusters
  - Irrelevant and redundant features
  - All points are equally close

- Solutions: Dimension Reduction
  - Feature subset selection
  - Cluster ensembles using random projection (in a later lecture….)

Redundant



Irrelevant



Curse of Dimensionality

100 observations cover the 1-D unit interval [0,1] well

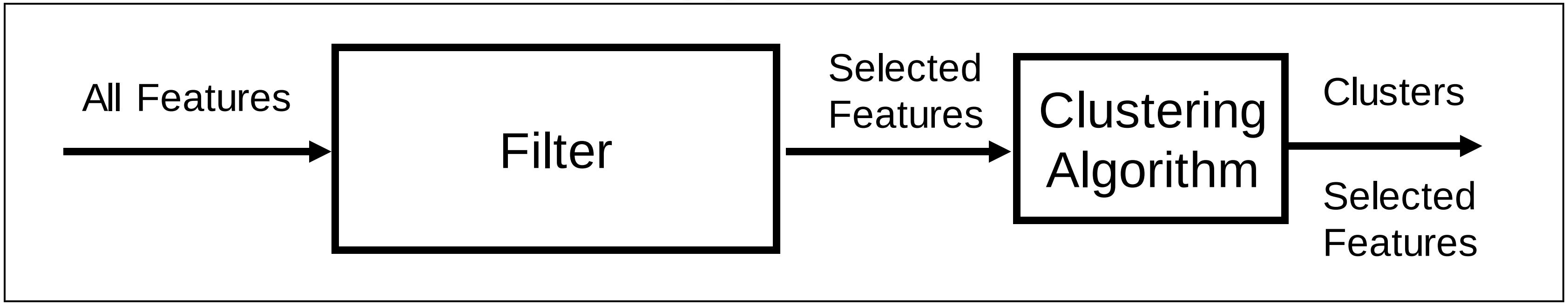
Consider the 10-D unit hypersquare, 100 observations are now isolated points in a vast empty space.

Consequence of the Curse

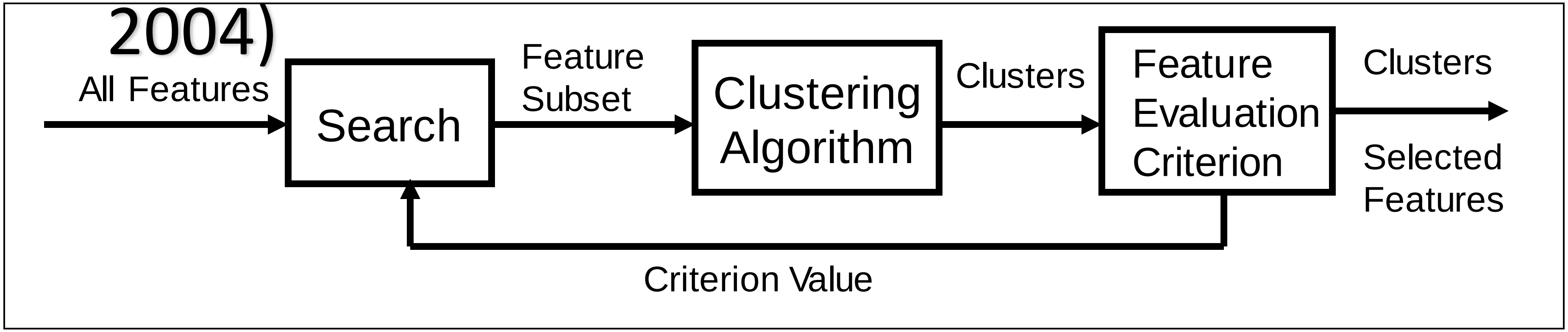
- Suppose the number of samples given to us in the total sample space is fixed
- Let the dimension increase
- Then the distance of the k nearest neighbors of any point increases

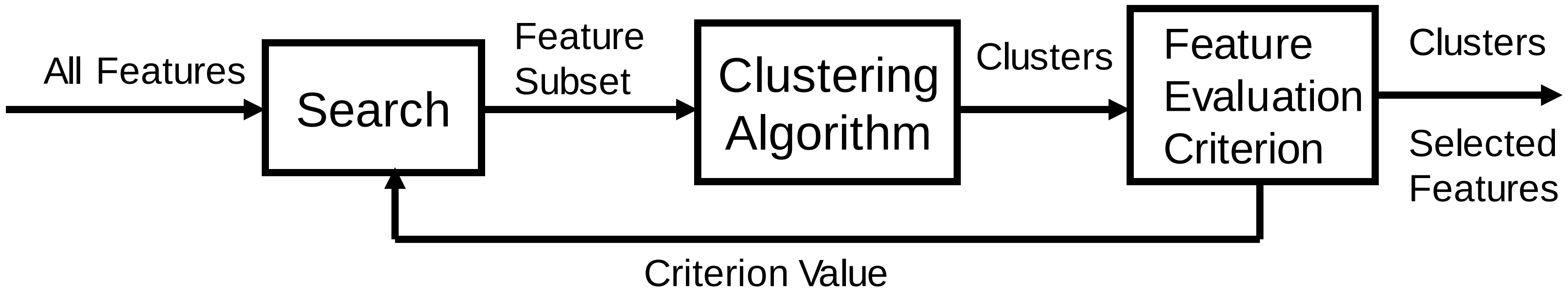
Feature Selection Methods

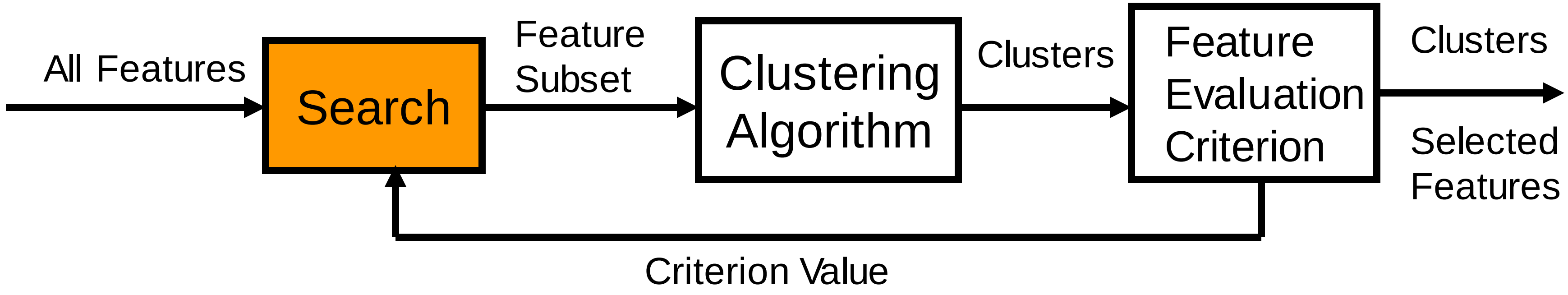
- Filter (Traditional approach)

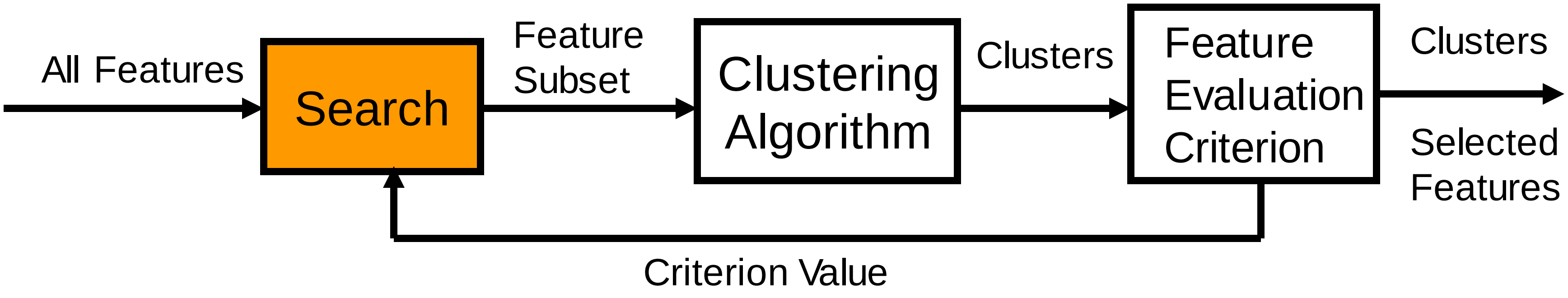


- Apply wrapper approach (Dy and Brodley, 2004)

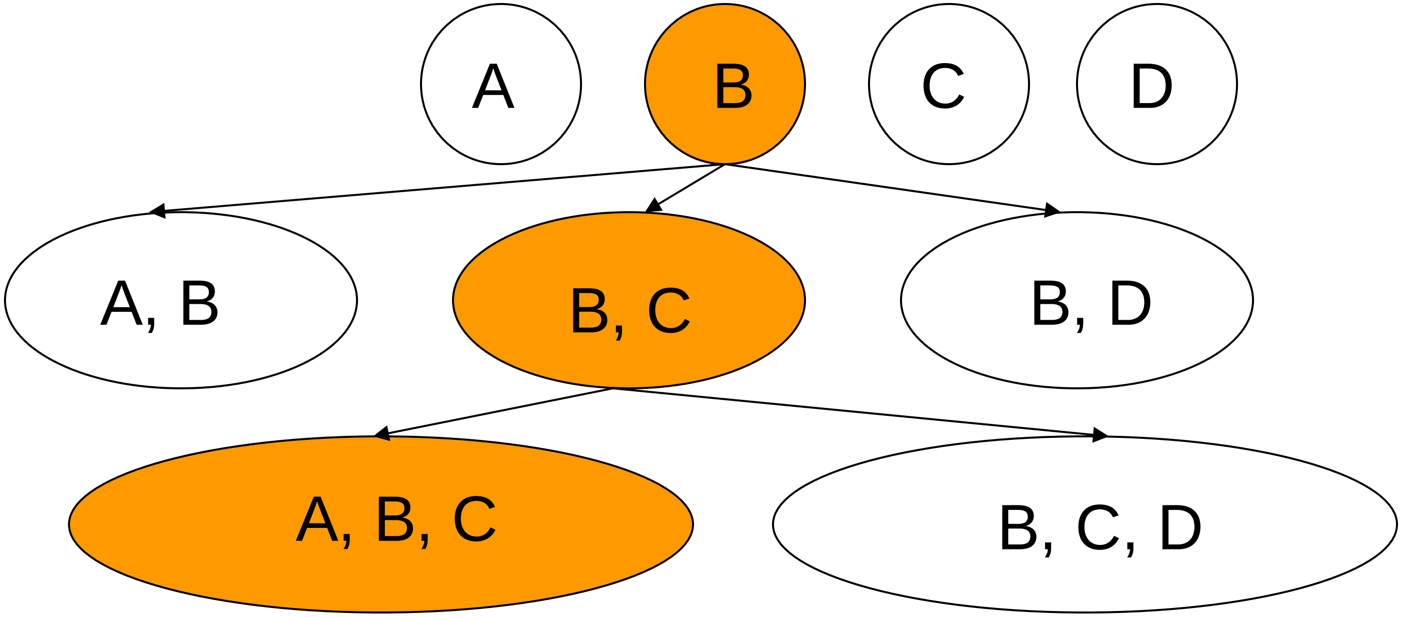


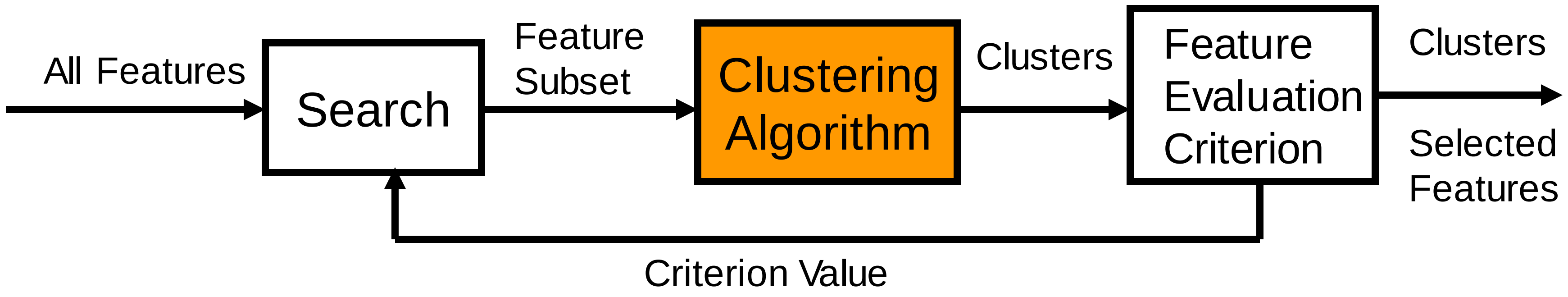

All Features
Search
Feature Subset
Clustering Algorithm
Clusters
Feature Evaluation Criterion
Clusters
Selected Features
Criterion Value

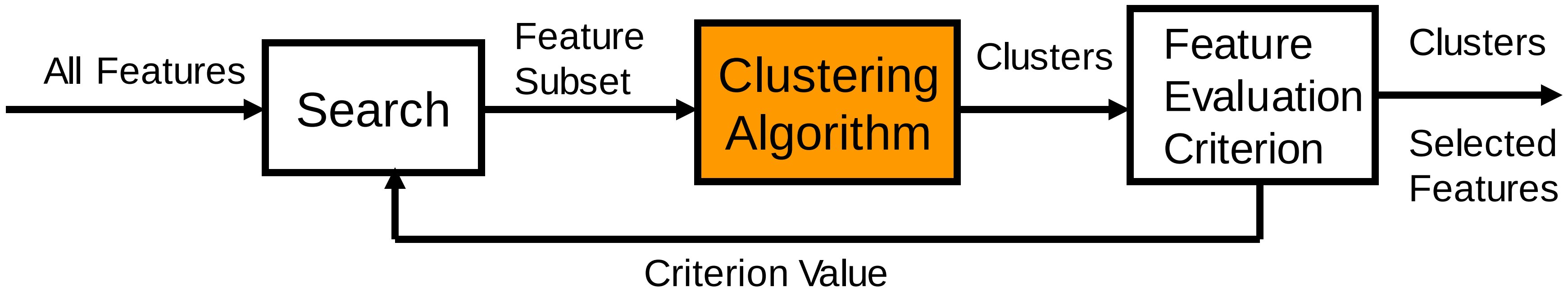

All Features
Search
Feature Subset
Clustering Algorithm
Clusters
Feature Evaluation Criterion
Clusters
Selected Features
Criterion Value



FSSEM Search Method: sequential forward search



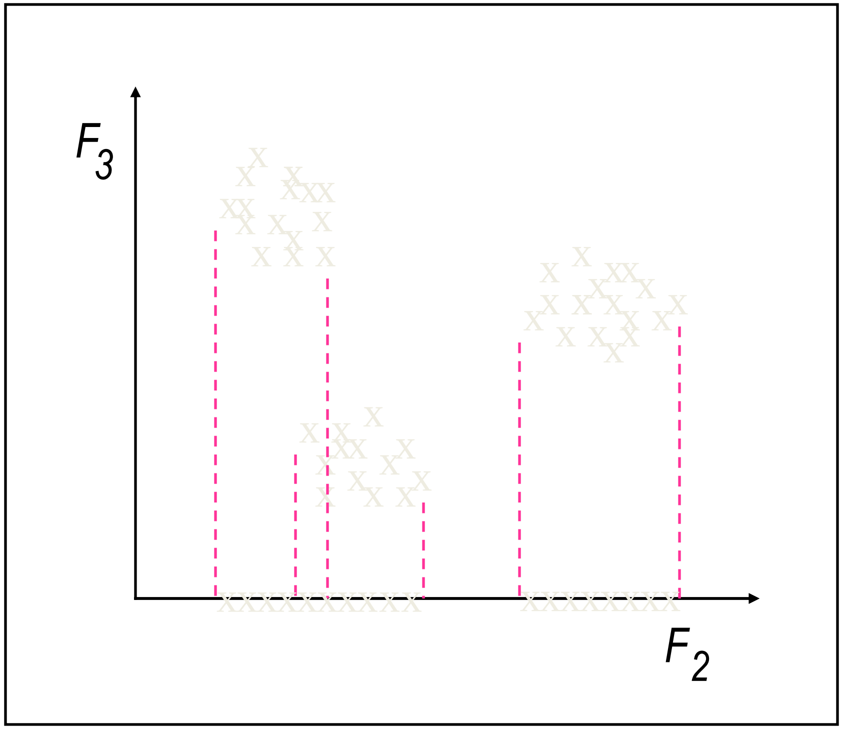

All Features
Search
Feature Subset
Clustering Algorithm
Clusters
Feature Evaluation Criterion
Clusters
Selected Features
Criterion Value



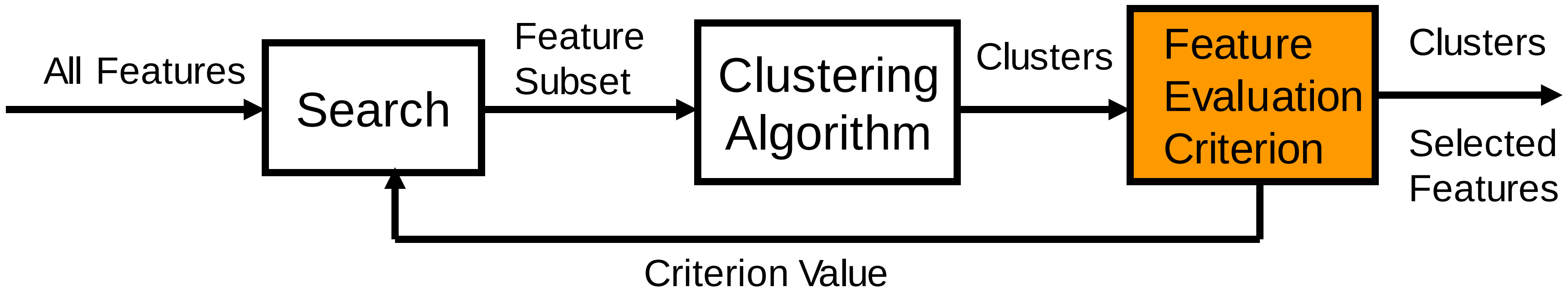
Clustering Algorithm:

Expectation Maximization (EM or EM-k) – coming soon

Searching for the Number of Clusters



Using a fixed number of clusters for all feature sets does not model the data in the respective subspace correctly.


All Features
Search
Feature Subset
Clustering Algorithm
Clusters
Feature Evaluation Criterion
Clusters
Selected Features
Criterion Value