What is Clustering?

Also called *unsupervised learning*, sometimes called *classification* by statisticians and *sorting* by psychologists and *segmentation* by people in marketing

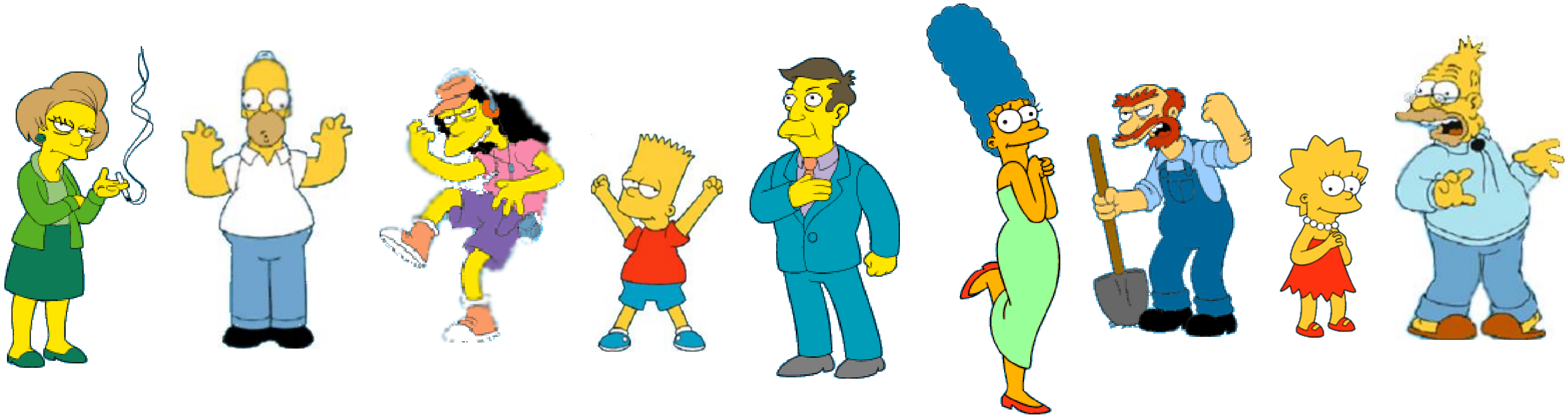
Clustering = Συσταδοποίηση, Ομαδοποίηση

- Organizing data into classes such that there is
  - high intra-class similarity
  - low inter-class similarity
- Finding the class labels and the number of classes directly from the data (in contrast to classification).
- More informally, finding natural groupings among objects.

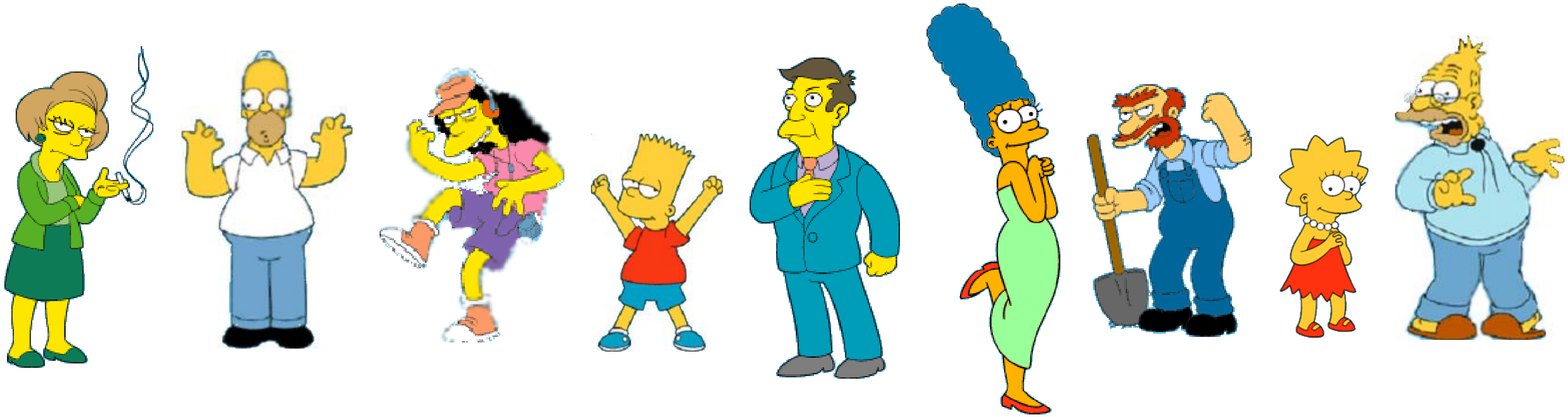
Useful for Clustering Understanding

- Market Segmentation
- [Social Network Analysis](https://immersion.media.mit.edu/viz)
- Organising Computing Clusters
- Astronomical Data Analysis
- Summarization
- Compression
- Efficiently Finding Nearest Neighbours

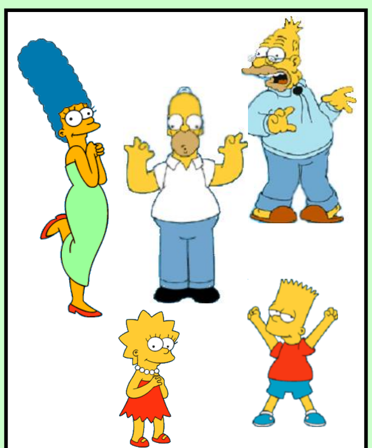
What is a natural grouping among these objects?



What is a natural grouping among these objects?

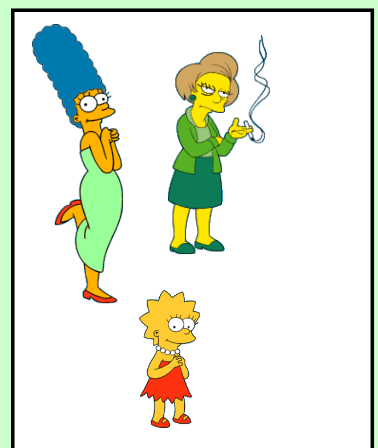


Clustering is subjective

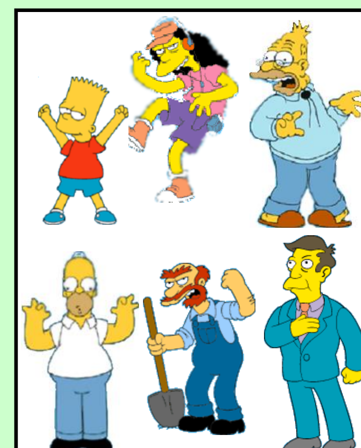


Simpson's Family

School Employees



Females



Males

What is Similarity?

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

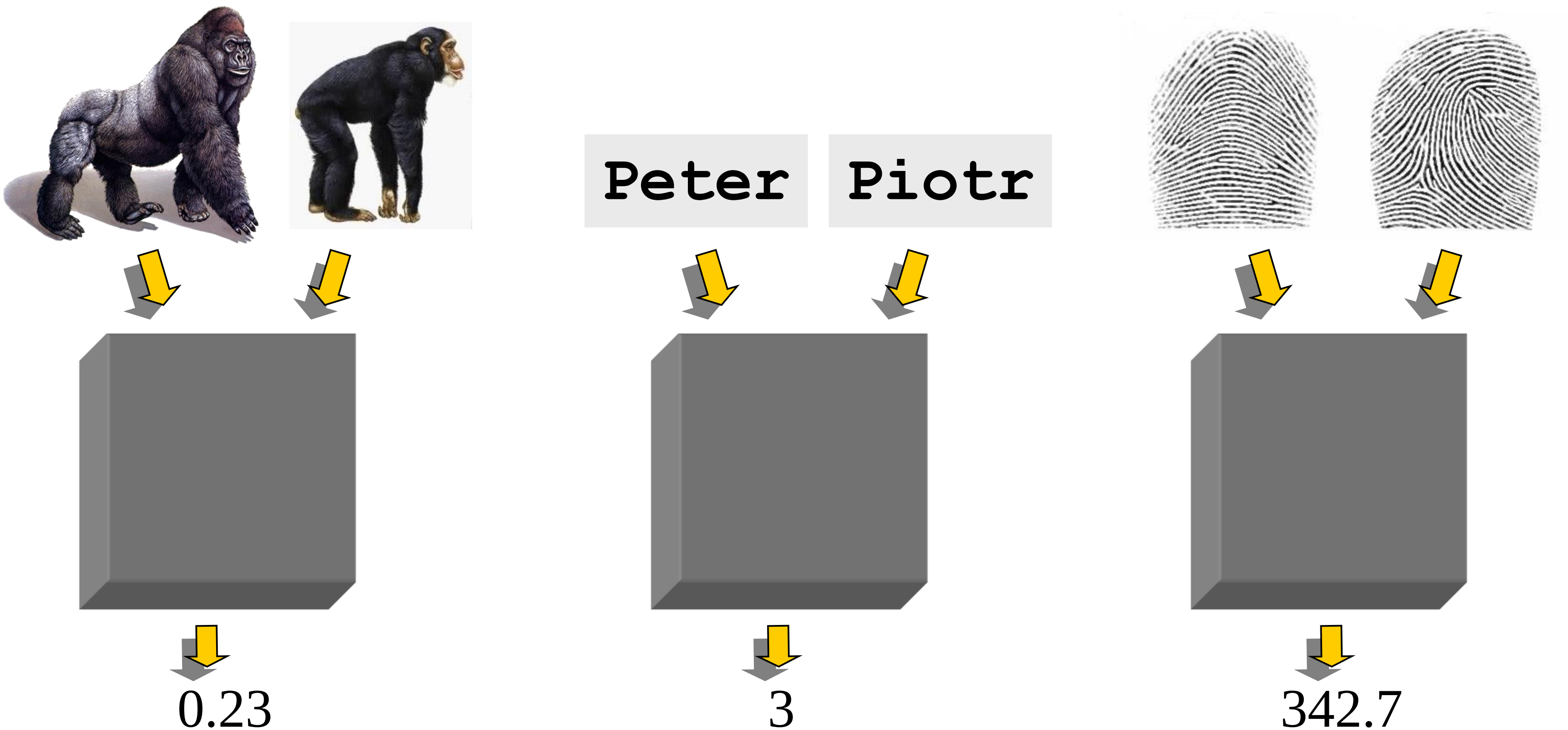
Webster's Dictionary

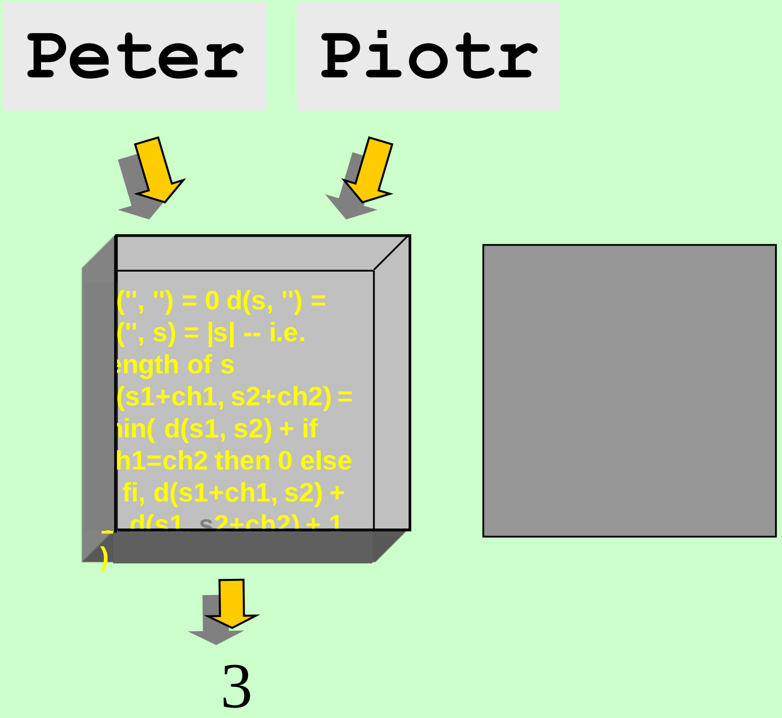
Similarity is hard to define, but… "*We know it when we see it*"

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-ανομοιότητα) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$





When we peek inside one of these black boxes, we see some function on two variables. These functions might very simple or very complex.

In either case it is natural to ask, what properties should these functions have?

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) *Symmetry*

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

D(A,A) = 0 *Constancy of Self-Similarity*

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

D(A,B) = 0 IIf A=B *Positivity (Separation)*

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

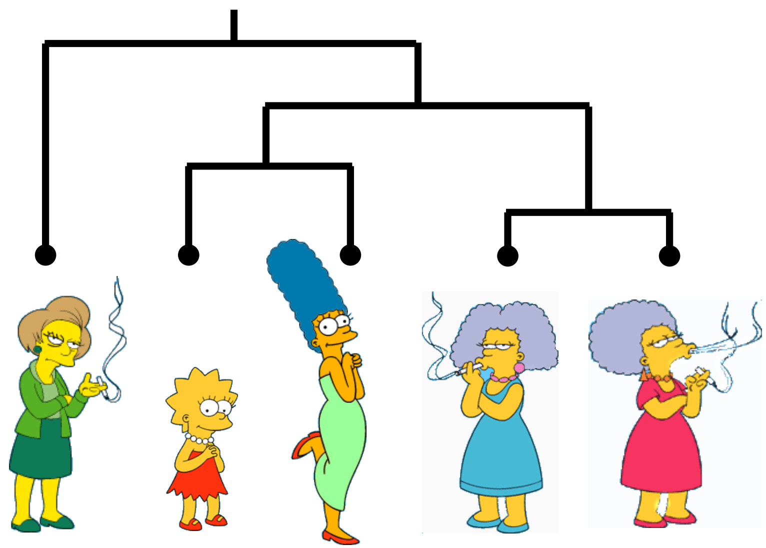
D(A,B) ≤ D(A,C) + D(B,C) *Triangular Inequality*

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

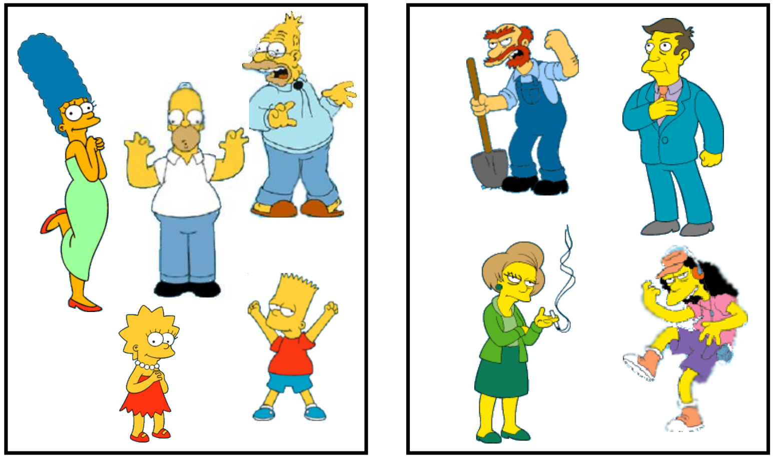
Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion (we will see an example called BIRCH)
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion

Hierarchical



Partitional

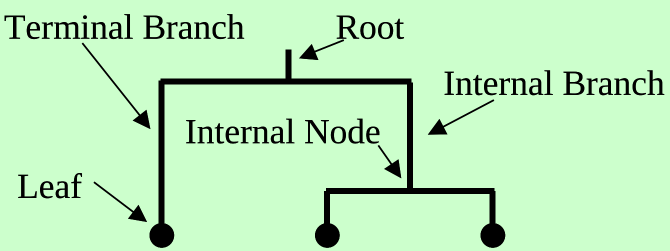


Desirable Properties of a Clustering Algorithm

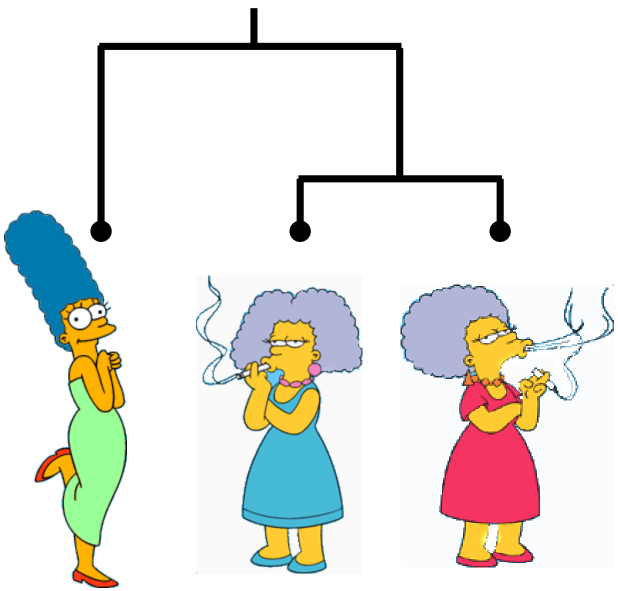
- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- Incorporation of user-specified constraints
- Interpretability and usability

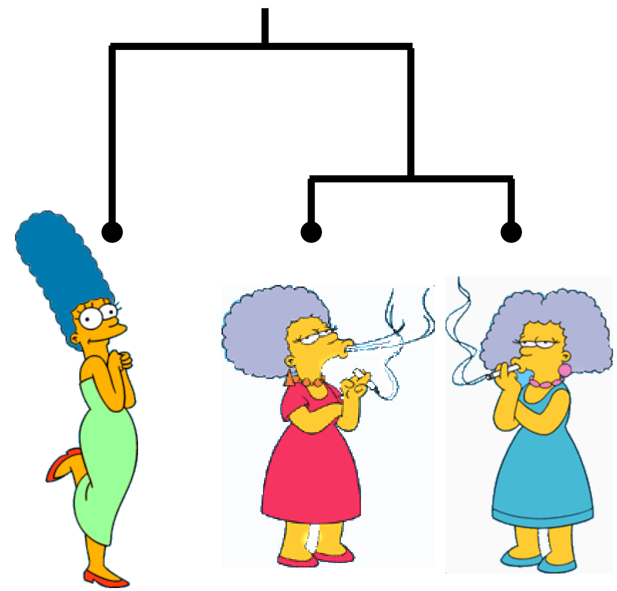
A Useful Tool for Summarizing Similarity Measurements

In order to better appreciate and evaluate the examples given in the early part of this talk, we will now introduce the *dendrogram*.



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

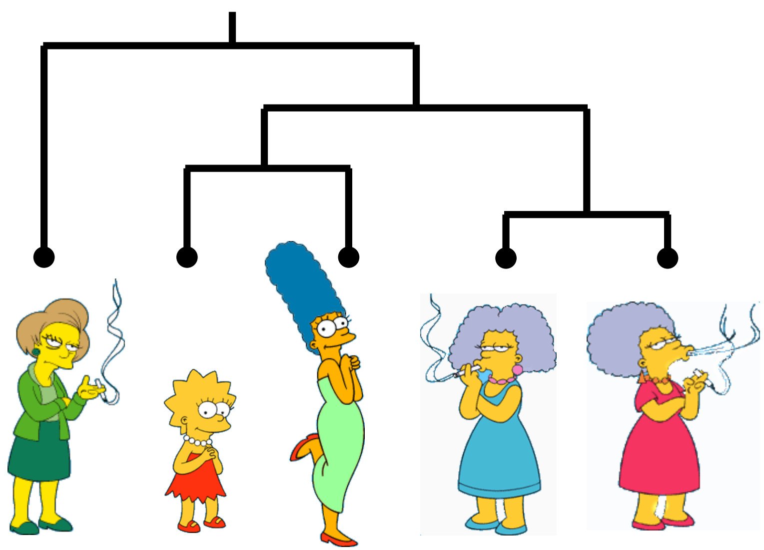




(How-to) Hierarchical Clustering

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

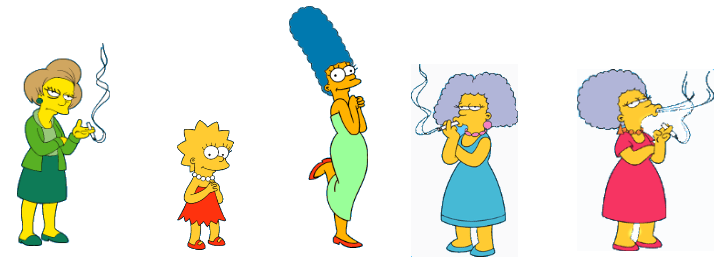


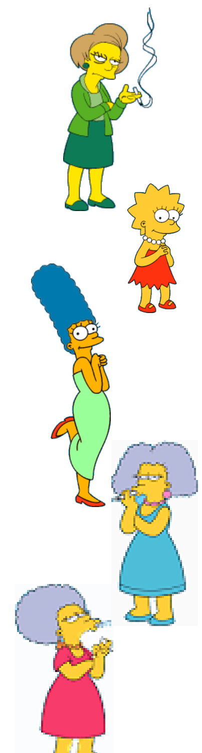
Since we cannot test all possible trees we will have to heuristic search of all possible trees. We could do this..

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.





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

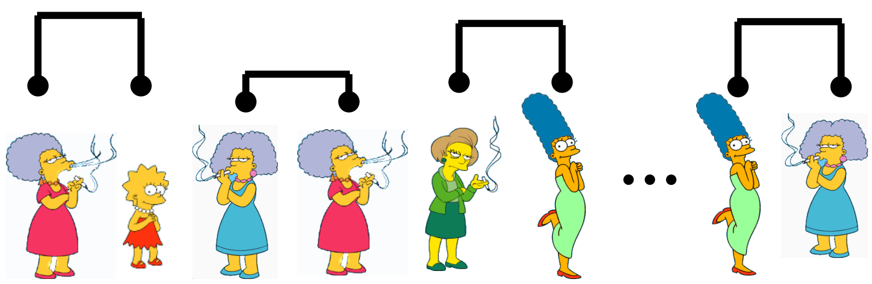
D(,) = 8

D(,) = 1

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…

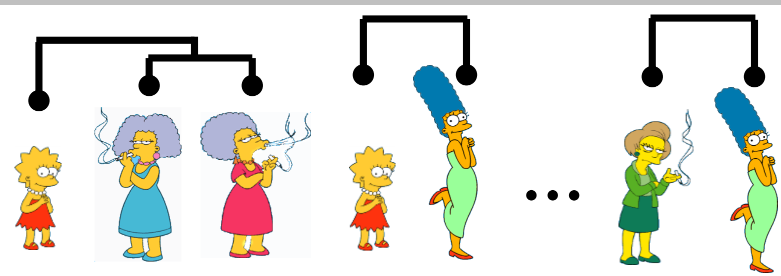


Choose the best

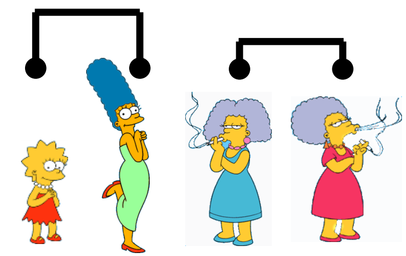
Bottom-Up (agglomerative):

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.

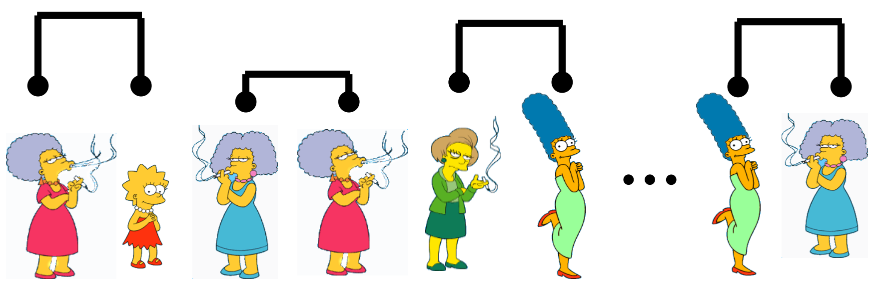
Consider all possible merges…



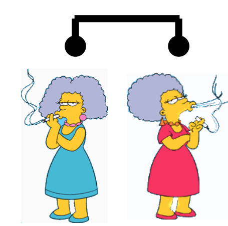
Choose the best



Consider all possible merges…



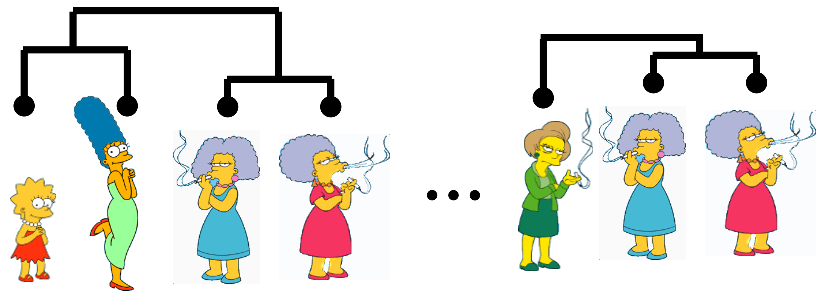
Choose the best



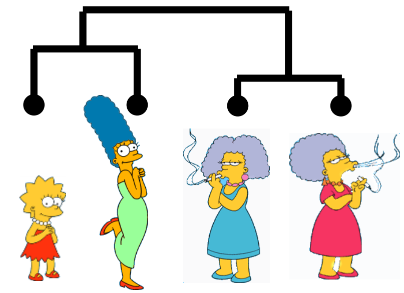
Bottom-Up (agglomerative):

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.

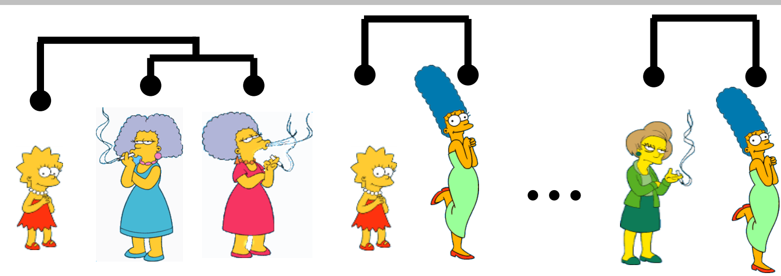
Consider all possible merges…



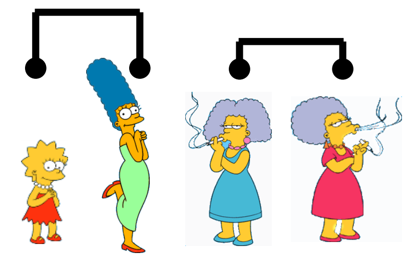
Choose the best



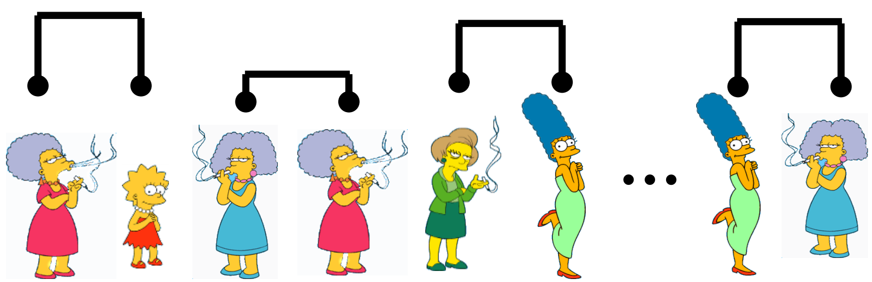
Consider all possible merges…



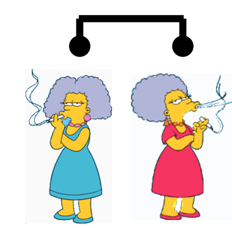
Choose the best



Consider all possible merges…

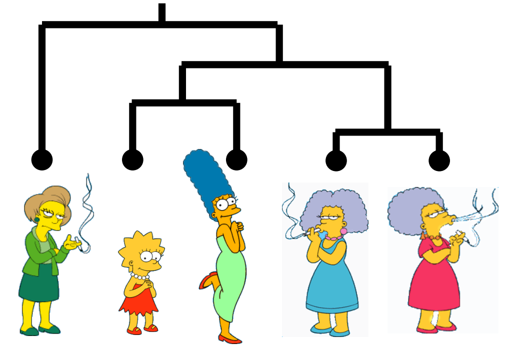


Choose the best

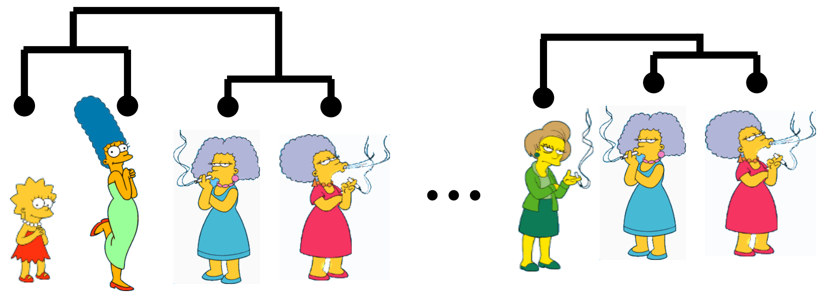


Bottom-Up (agglomerative):

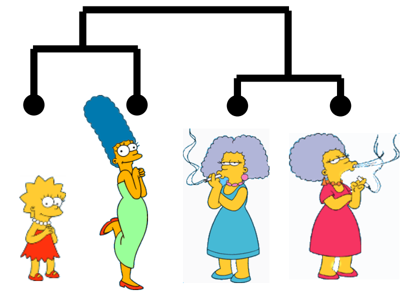
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



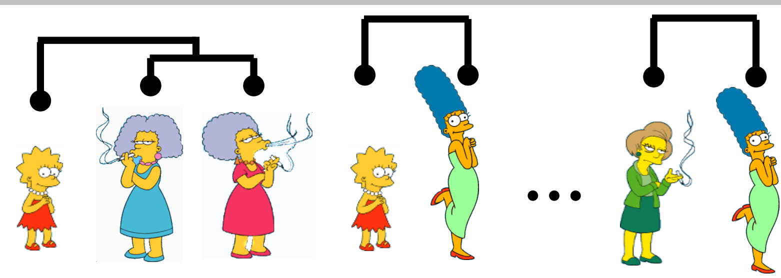
Consider all possible merges…



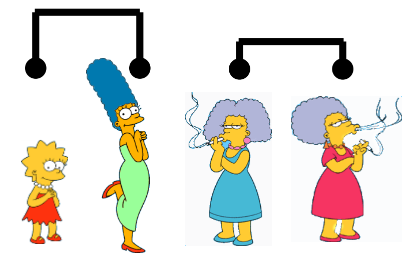
Choose the best



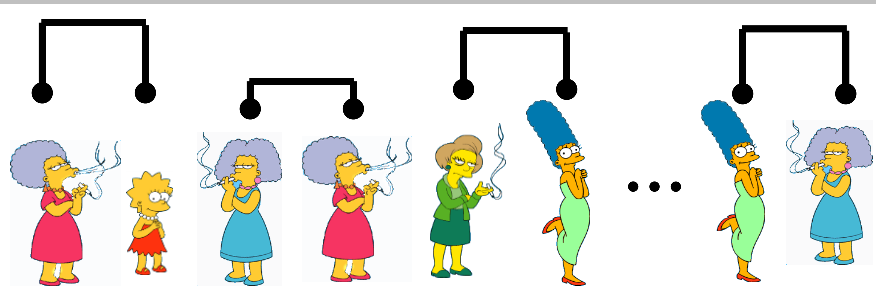
Consider all possible merges…



Choose the best



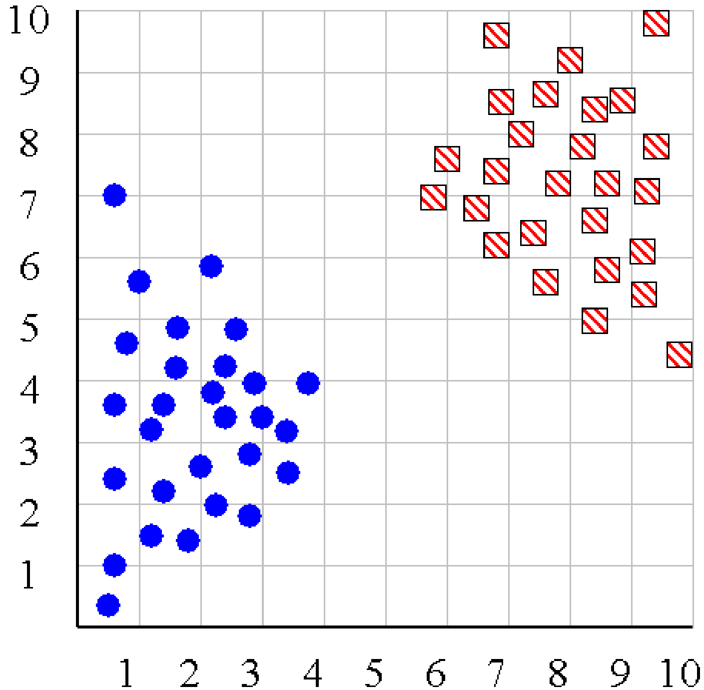
Consider all possible merges…

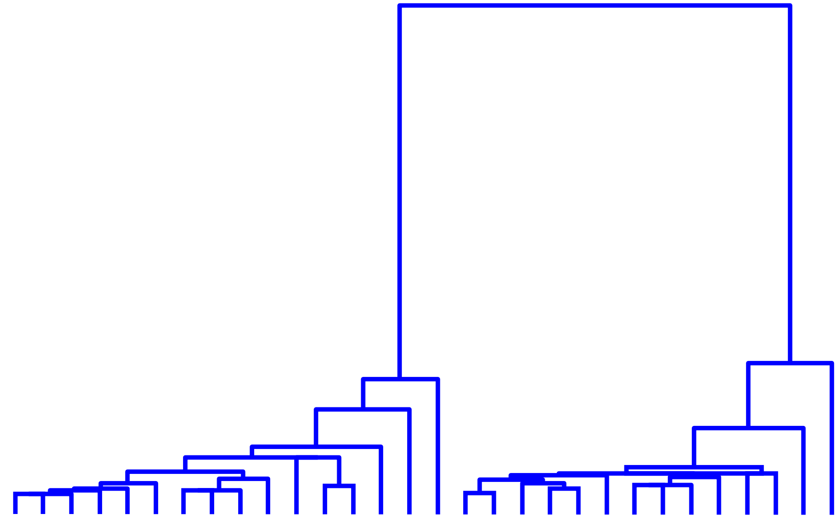


Choose the best

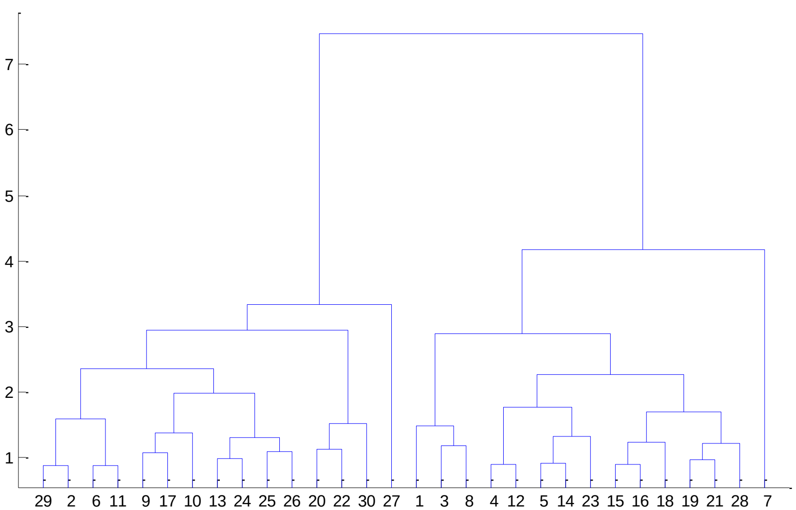
We know how to measure the distance between two objects, but defining the distance between an object and a cluster, or defining the distance between two clusters is non obvious.

- **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.
- **Wards Linkage**: In this method, we try to minimize the variance of the merged clusters

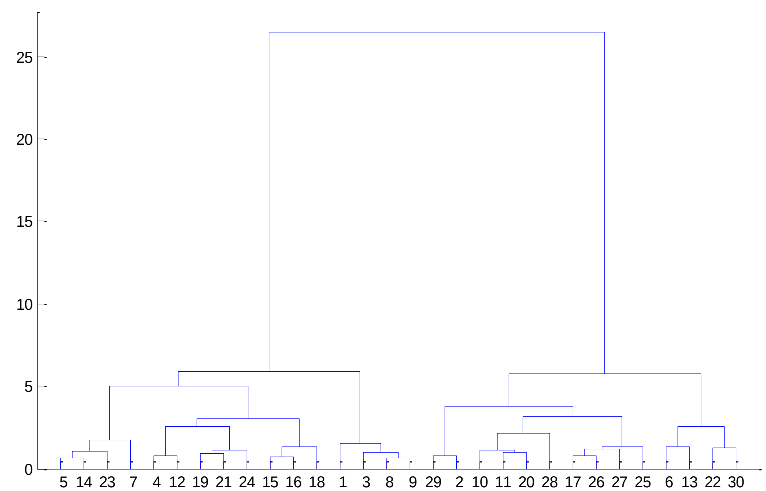




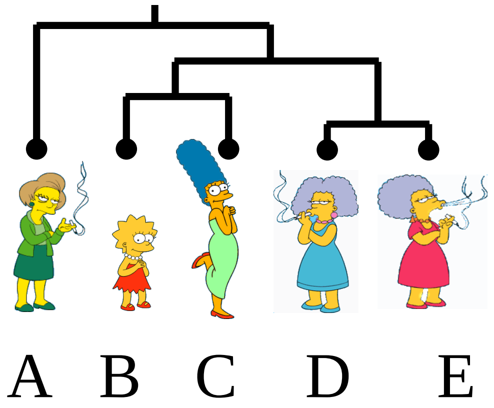
Single linkage



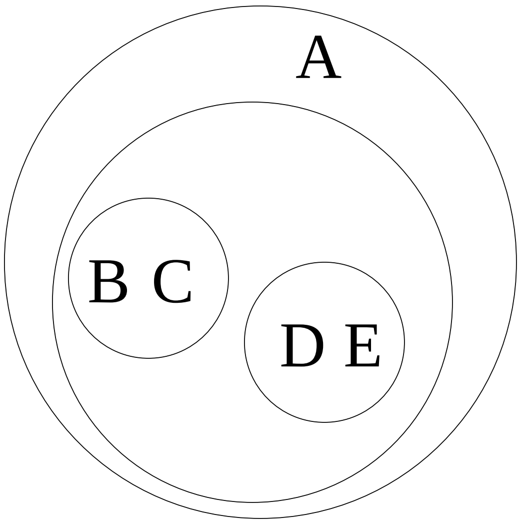
Average linkage

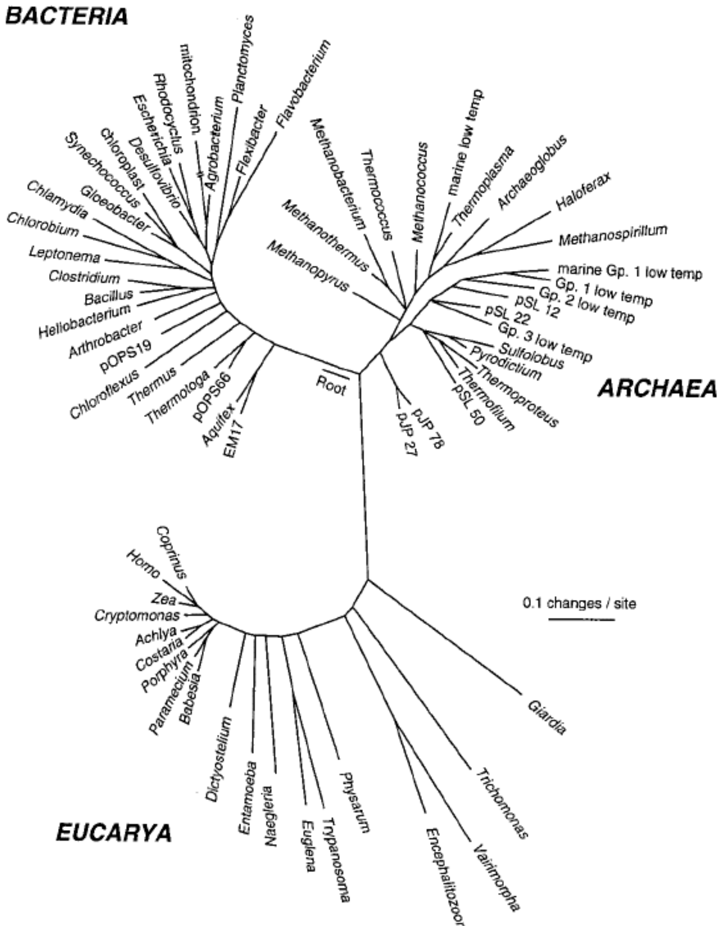


Wards linkage

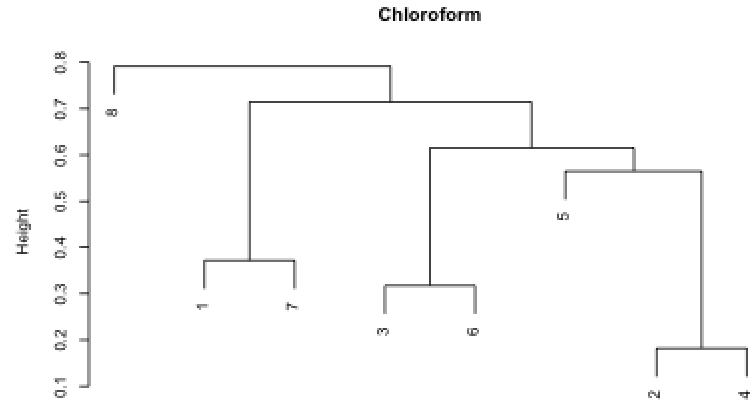


(A,((B,C),(D,E)))

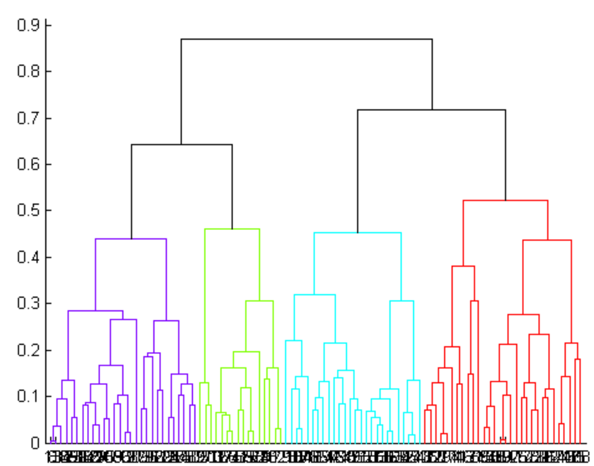




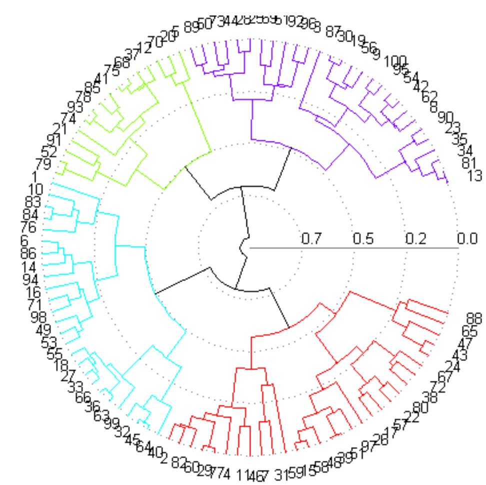
BACTERIA
ARCHAEA
EUCARYA
mitochondrion
Planctomyces
Flavobacterium
Rhodocyclus
Agrobacterium
Escherichia
Flexibacter
Desulfovibrio
chloroplast
Synechococcus
Gloeobacter
Chlamydia
Chlorobium
Leptonema
Clostridium
Bacillus
Heliobacterium
Arthrobacter
pOPS19
Chloroflexus
Thermus
Thermotoga
pOPS66
Aquifex
EM17
Root
Methanobacterium
Thermococcus
Methanococcus
marine low temp
Thermoplasma
Archaeoglobus
Haloferax
Methanospirillum
Methanothermus
Methanopyrus
marine Gp. 1 low temp
Gp. 1 low temp
Gp. 2 low temp
pSL 12
pSL 22
Gp. 3 low temp
Sulfolobus
Pyrodictium
Thermoproteus
Thermofilum
pSL 50
pJP 78
pJP 27
Coprinus
Homo
Zea
Cryptomonas
Achlya
Costaria
Porphyra
Paramecium
Babesia
Dictyostelium
Entamoeba
Naegleria
Euglena
Trypanosoma
Physarum
Encephalitozoon
Vairimorpha
Trichomonas
Giardia
0.1 changes / site



Chloroform
Height
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
8
1
7
3
6
5
2
4

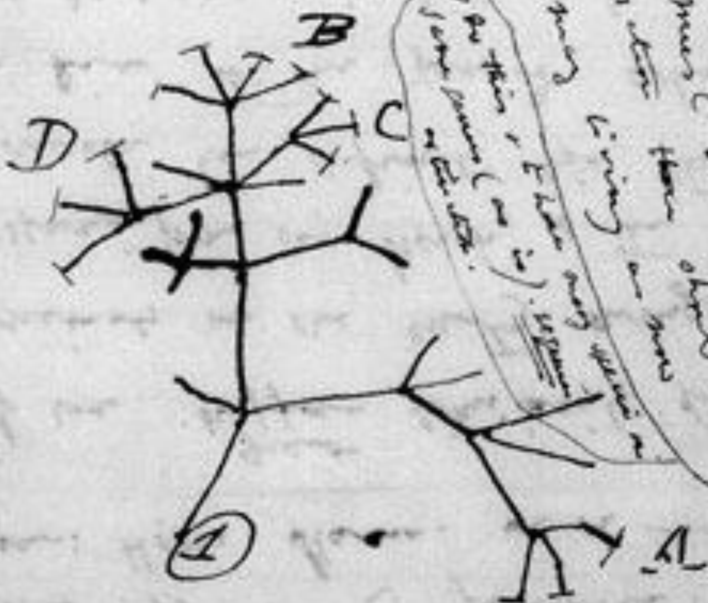


0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9



0.7
0.5
0.2
0.0

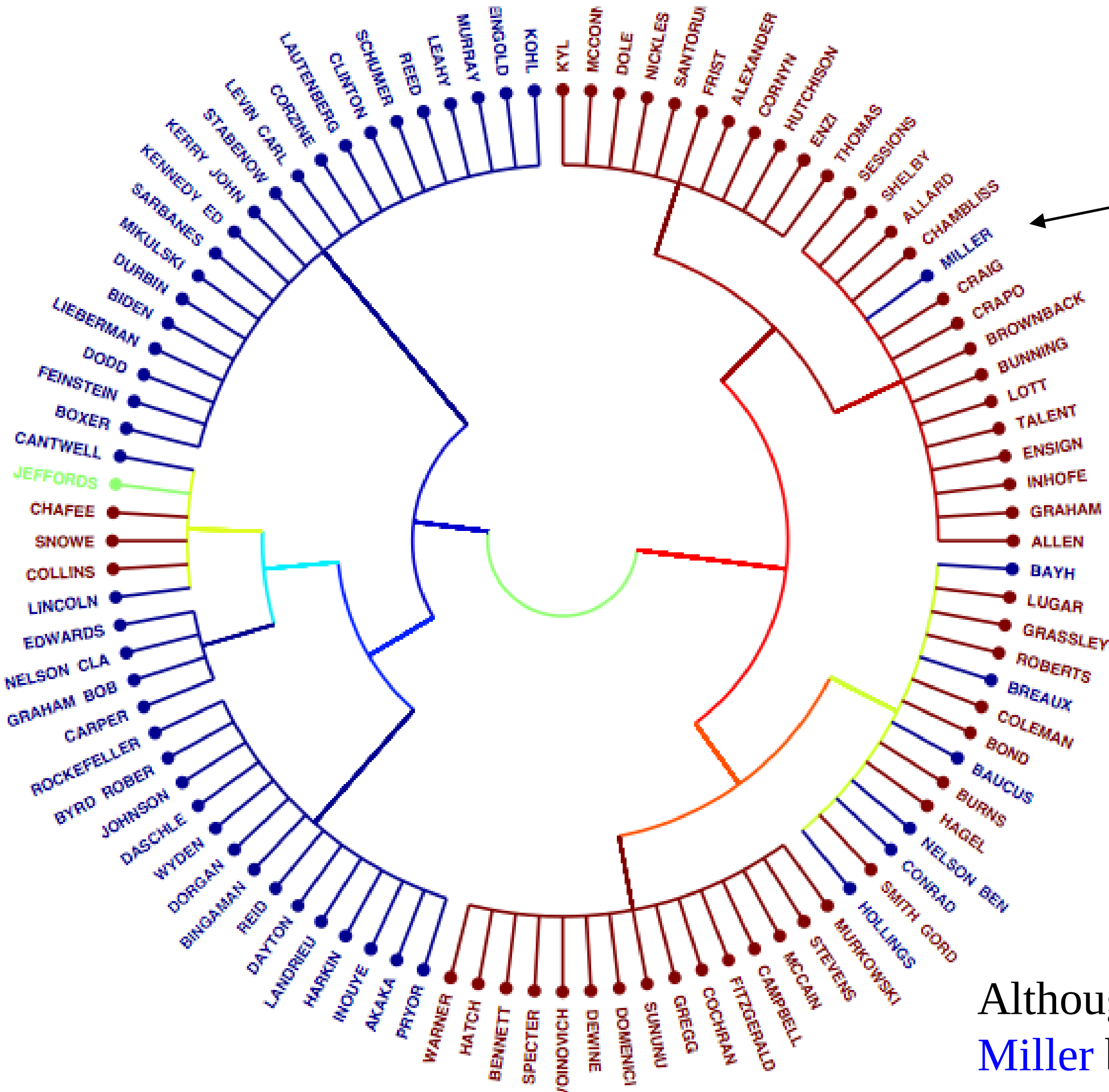
I think



Case must be that one generation then should be as many living as now

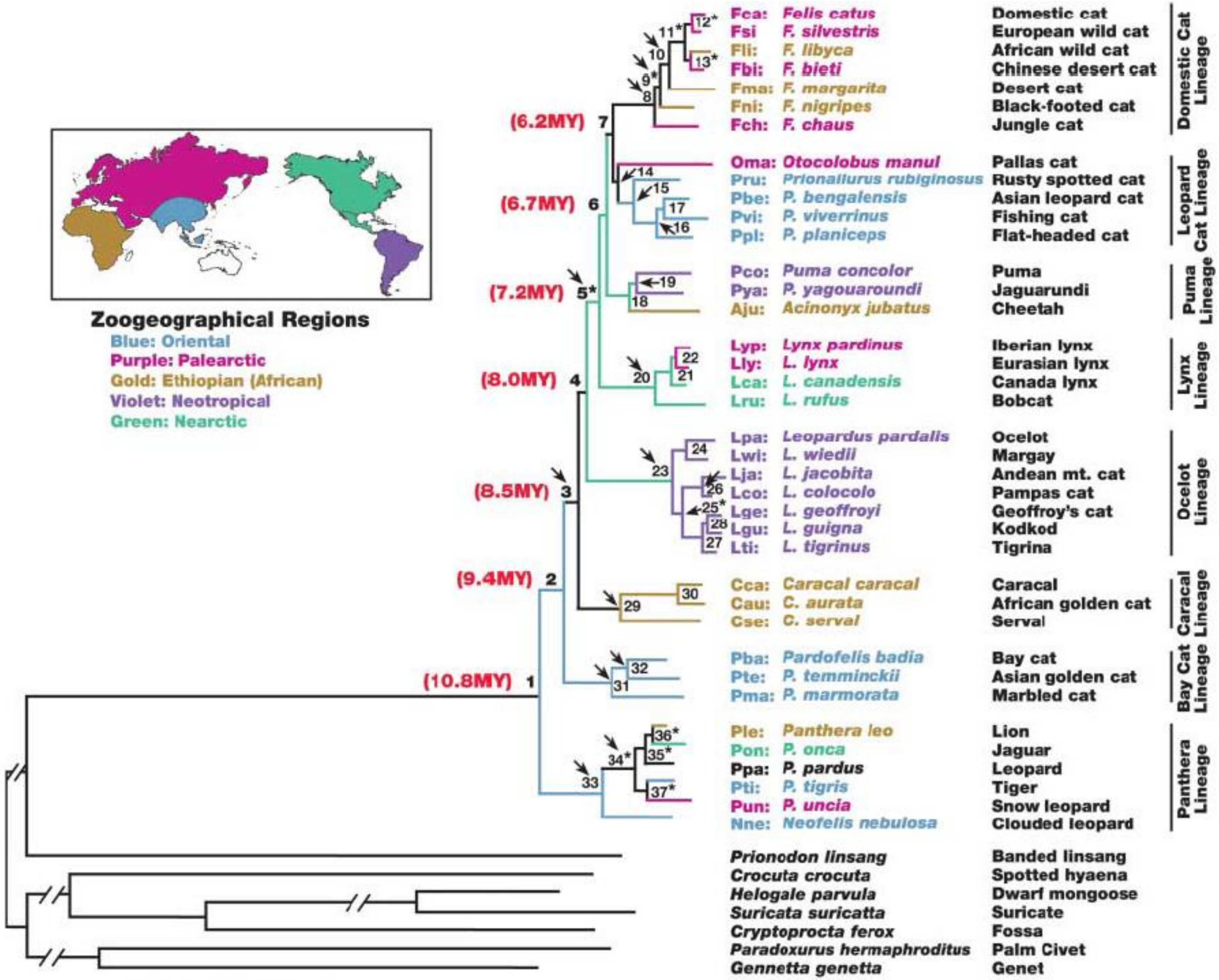
To do this & to have many species in same genus (as is). requires extinction.

Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation



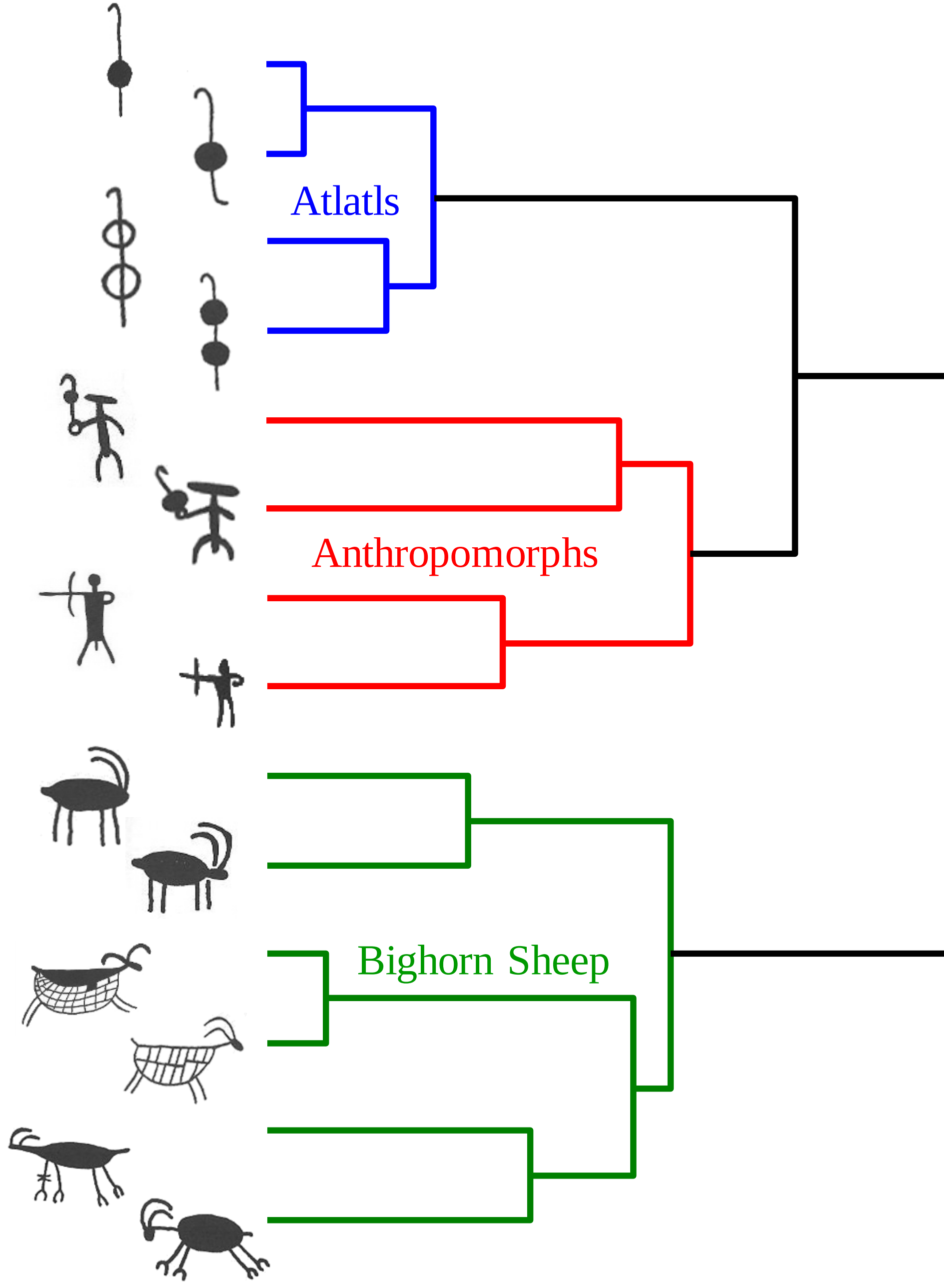
Co-Sponsorship Networks– Senators of the 108th Congress

Although a member of the Democratic Party, Miller backed Republican President George W. Bush over Democratic nominee John Kerry in the 2004 presidential election and since 2003 has frequently criticized the Democratic Party, and has publicly supported several Republican candidates.



The Late Miocene Radiation of Modern Felidae: A Genetic Assessment .

Warren E. Johnson, *Science* 6 January 2006: Vol. 311. no. 5757, pp. 73 - 77


Atlatls
Anthropomorphs
Bighorn Sheep

Hierarchical clustering

- Input: a pairwise matrix involved all instances in S
- Algorithm

1. Place each instance of S in its own cluster (singleton), creating the list of clusters L (initially, the leaves of T):

   $L= S_1, S_2, S_3, ..., S_{n-1}, S_n$.

2. Compute a **merging cost function** between every pair of elements in L to find the two closest clusters $\{S_i, S_j\}$ which will be the cheapest couple to merge.
3. Remove S_i and S_j from L.
4. Merge S_i and S_j to create a new internal node S_{ij} in T which will be the parent of S_i and S_j in the resulting tree.
5. Go to **Step 2** until there is only one set remaining.

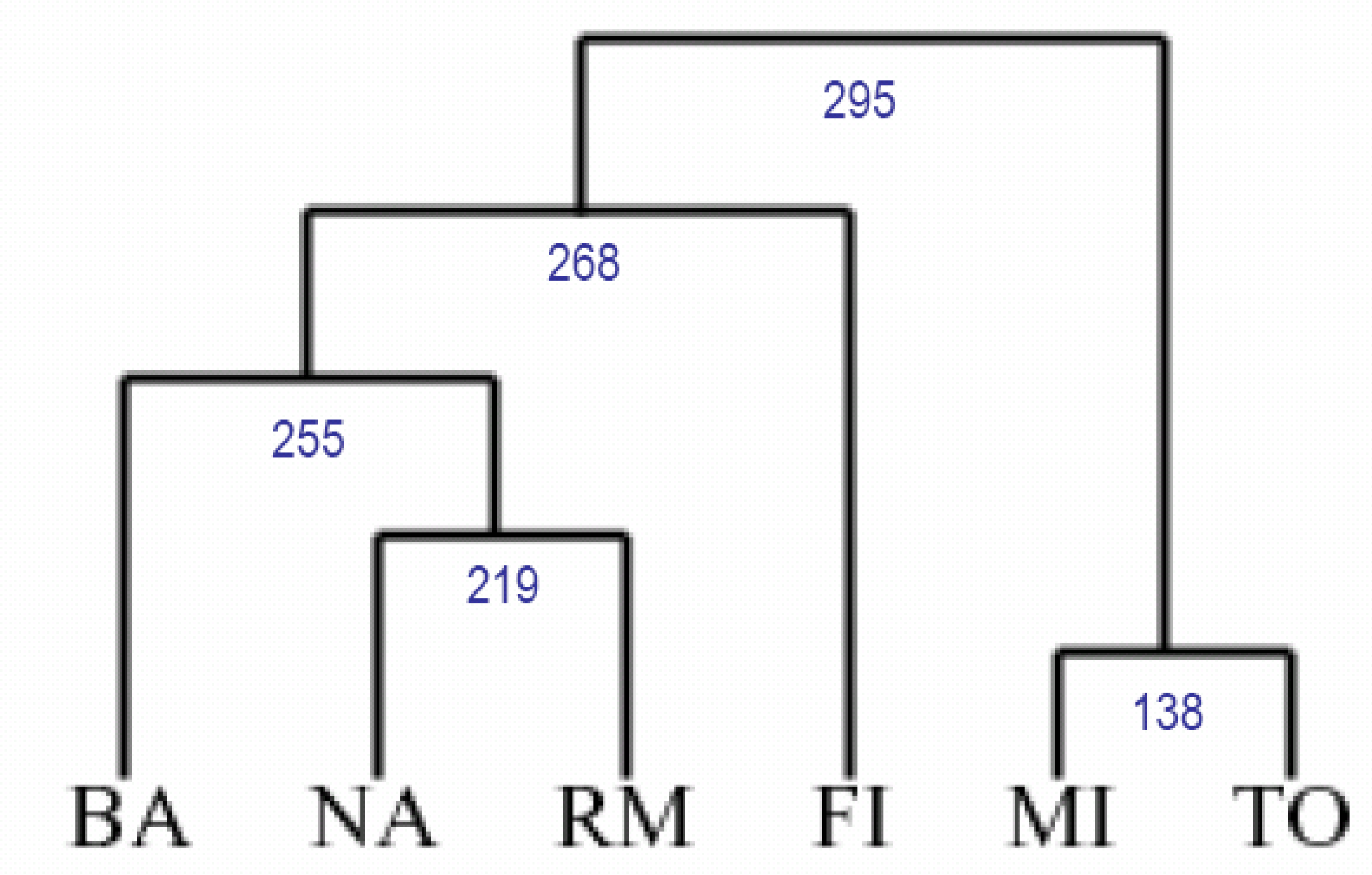
Hierarchical clustering

- **Step 2** can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering.
  - In single-linkage clustering (also called the connectedness or minimum method): we consider the distance between one cluster and another cluster to be equal to the **shortest** distance from any member of one cluster to any member of the other cluster.
  - In complete-linkage clustering (also called the diameter or maximum method), we consider the distance between one cluster and another cluster to be equal to the **greatest** distance from any member of one cluster to any member of the other cluster.
  - In average-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the **average** distance from any member of one cluster to any member of the other cluster.

Hierarchical clustering: example

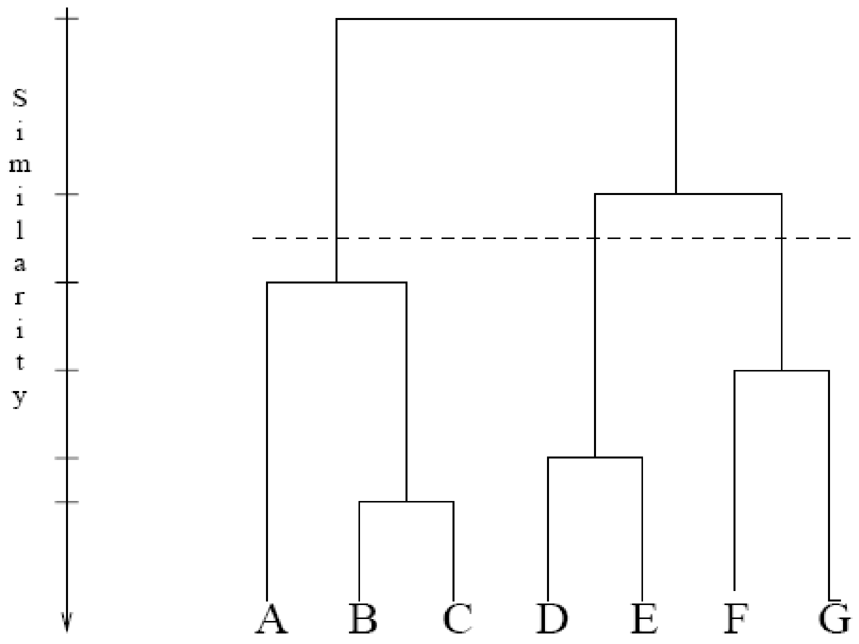
	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

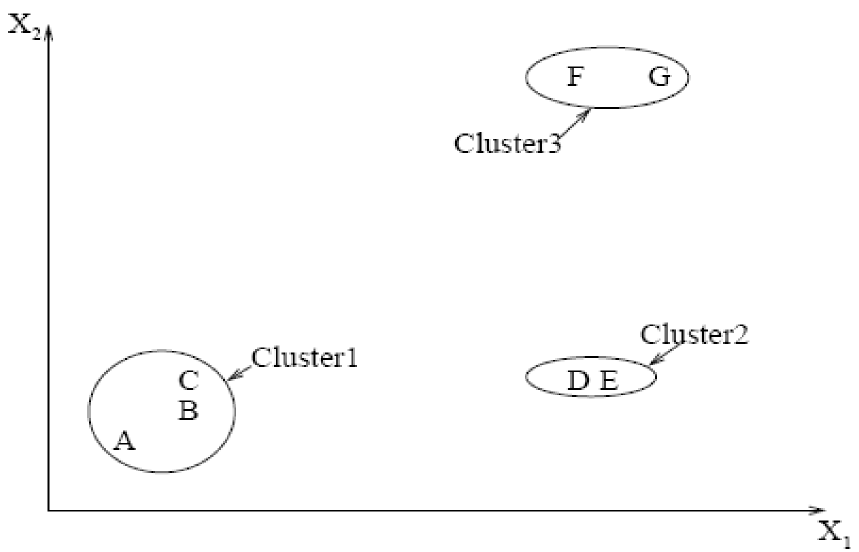
Hierarchical clustering: example using single linkage



Hierarchical clustering: forming clusters

- Forming clusters from dendograms





Hierarchical clustering

- Advantages
  - Dendograms are great for visualization
  - Provides hierarchical relations between clusters
  - Shown to be able to capture concentric clusters
- Disadvantages
  - Not easy to define levels for clusters
  - Experiments showed that other clustering techniques outperform hierarchical clustering

Summary of Hierarchal Clustering Methods

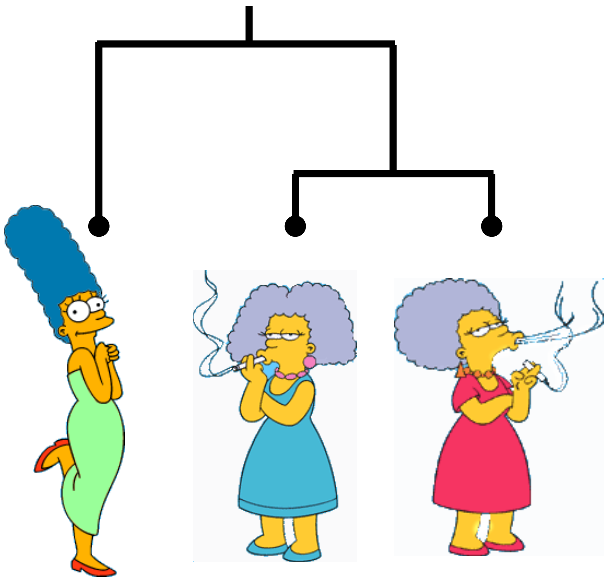
- No need to specify the number of clusters in advance.
- Hierarchal nature maps nicely onto human intuition for some domains
- They do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects.
- Like any heuristic search algorithms, local optima are a problem.
- Interpretation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

Up to this point we have simply assumed that we can measure similarity, but

How do we measure similarity?

A generic technique for measuring similarity

To measure the similarity between two objects, transform one of the objects 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**

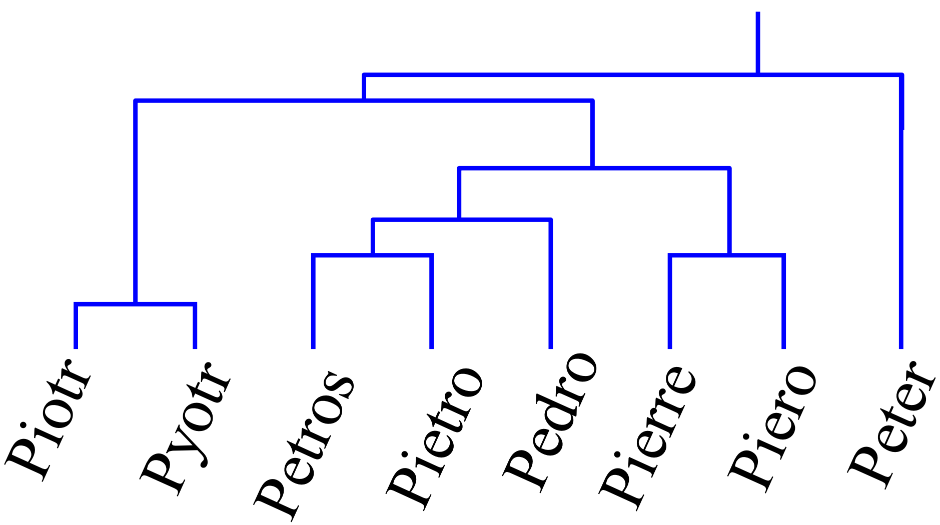
This is called the "edit distance" or the "transformation distance"

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



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)

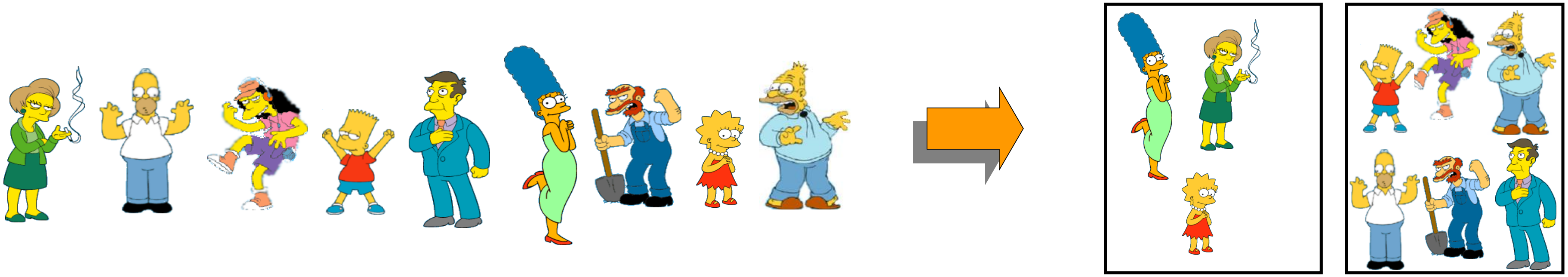
`Pioter`

Deletion (e)

`Piotr`

Partitional Clustering

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



K-means

- Input: n objects (or points) and a number k
- Algorithm
  1. Randomly place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
  2. Assign each object to the group that has the closest centroid.
  3. When all objects have been assigned, recalculate the positions of the K centroids.
  4. Repeat Steps 2 and 3 until the stopping criteria is met.

K-means

- Stopping criteria:
  - No change in the members of all clusters
  - when the squared error is less than some small threshold value α
    - Squared error *se*

$$se = \sum_{i=1}^{k} \sum_{p \in c_i} \| p - m_i \|^2$$

      - where m_i is the mean of all instances in cluster c_i
    - $se < \alpha$
- Properties of k-means
  - Guaranteed to converge
  - Guaranteed to achieve local optimal, not necessarily global optimal.
- Example: http://www.kdnuggets.com/dmcourse/data_mining_course/mod-13-clustering.ppt.

K-means

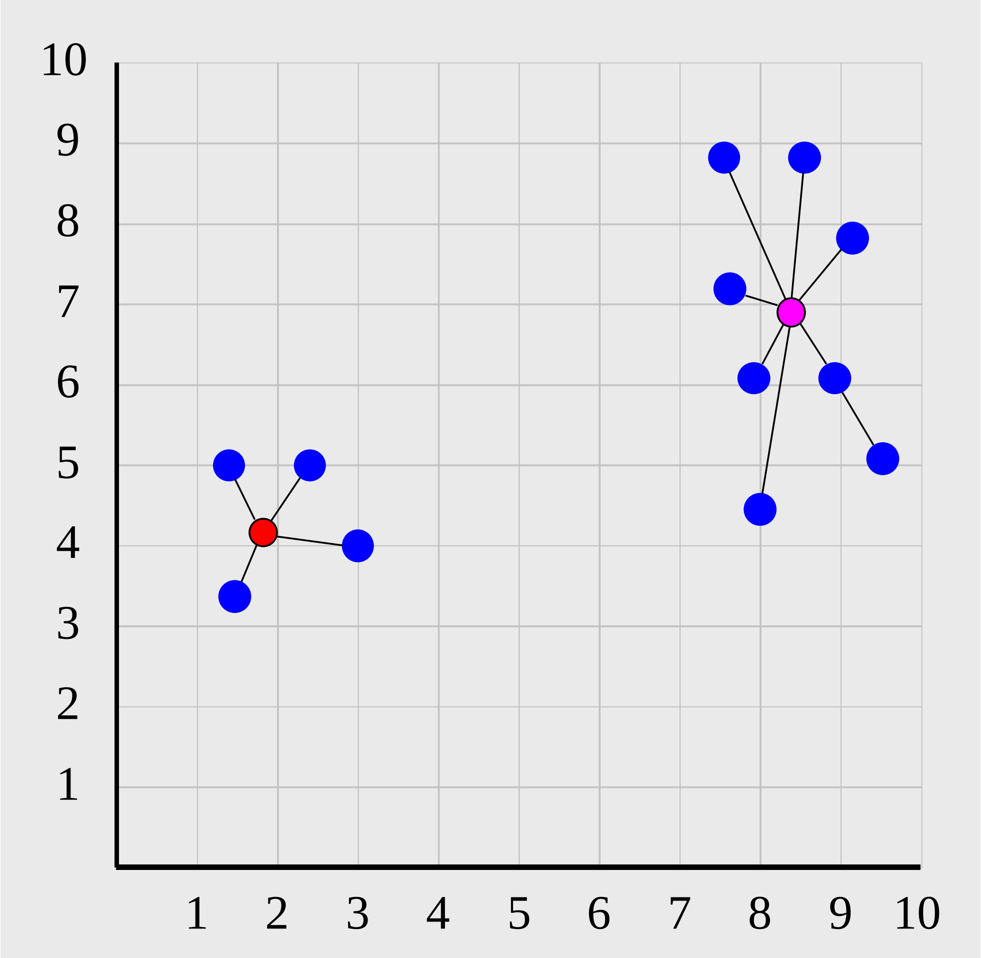
- Pros:
  - Low complexity
    - *complexity is O(nkt), where t = #iterations*
- Cons:
  - Necessity of specifying k
  - Sensitive to noise and outlier data points
    - Outliers: a small number of such data can substantially influence the mean value)
  - Clusters are sensitive to initial assignment of centroids
    - K-means is not a deterministic algorithm
    - Clusters can be inconsistent from one run to another

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



Algorithm *k-means*

1. Decide on a value for k.

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

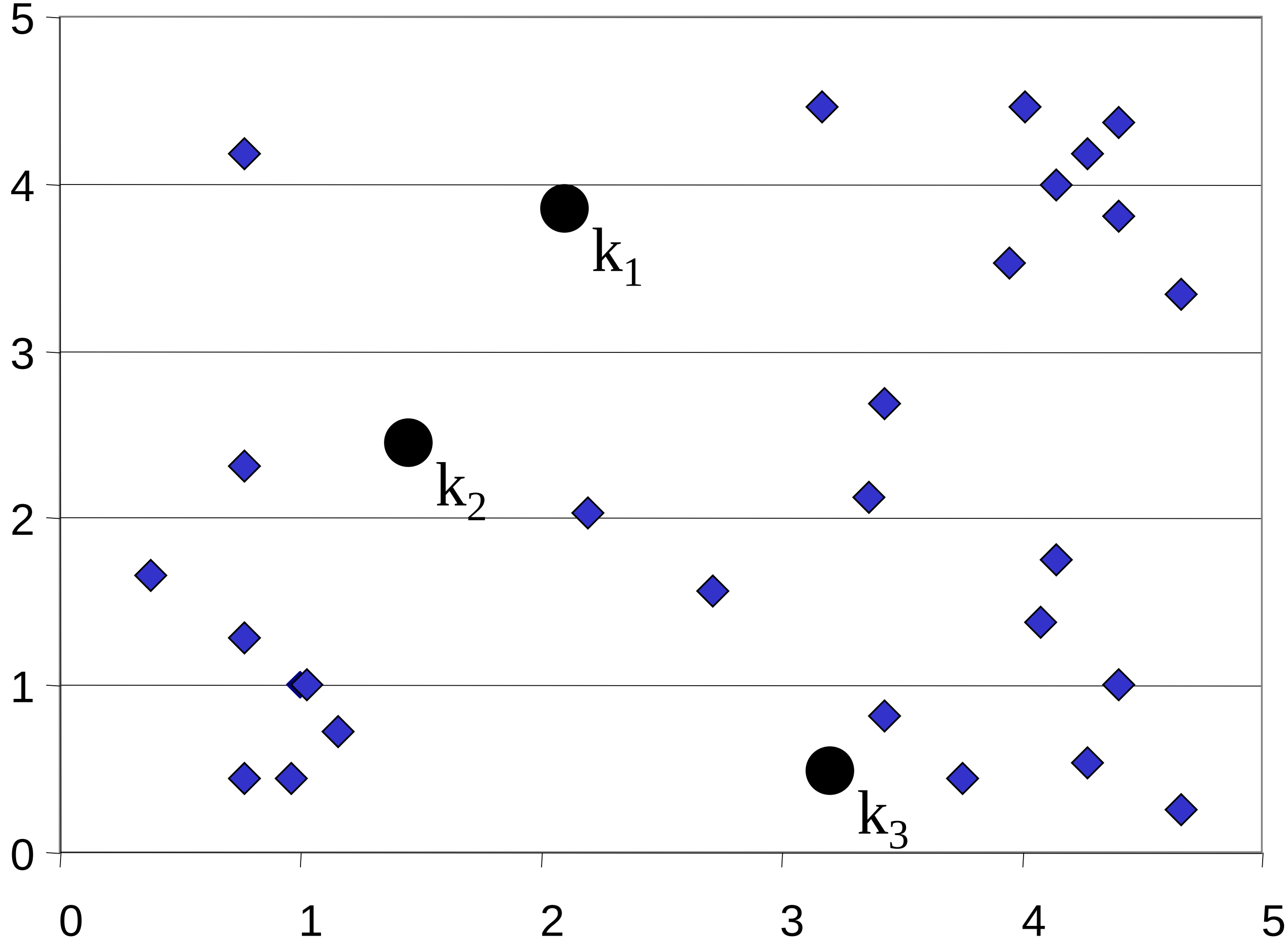
3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.

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.

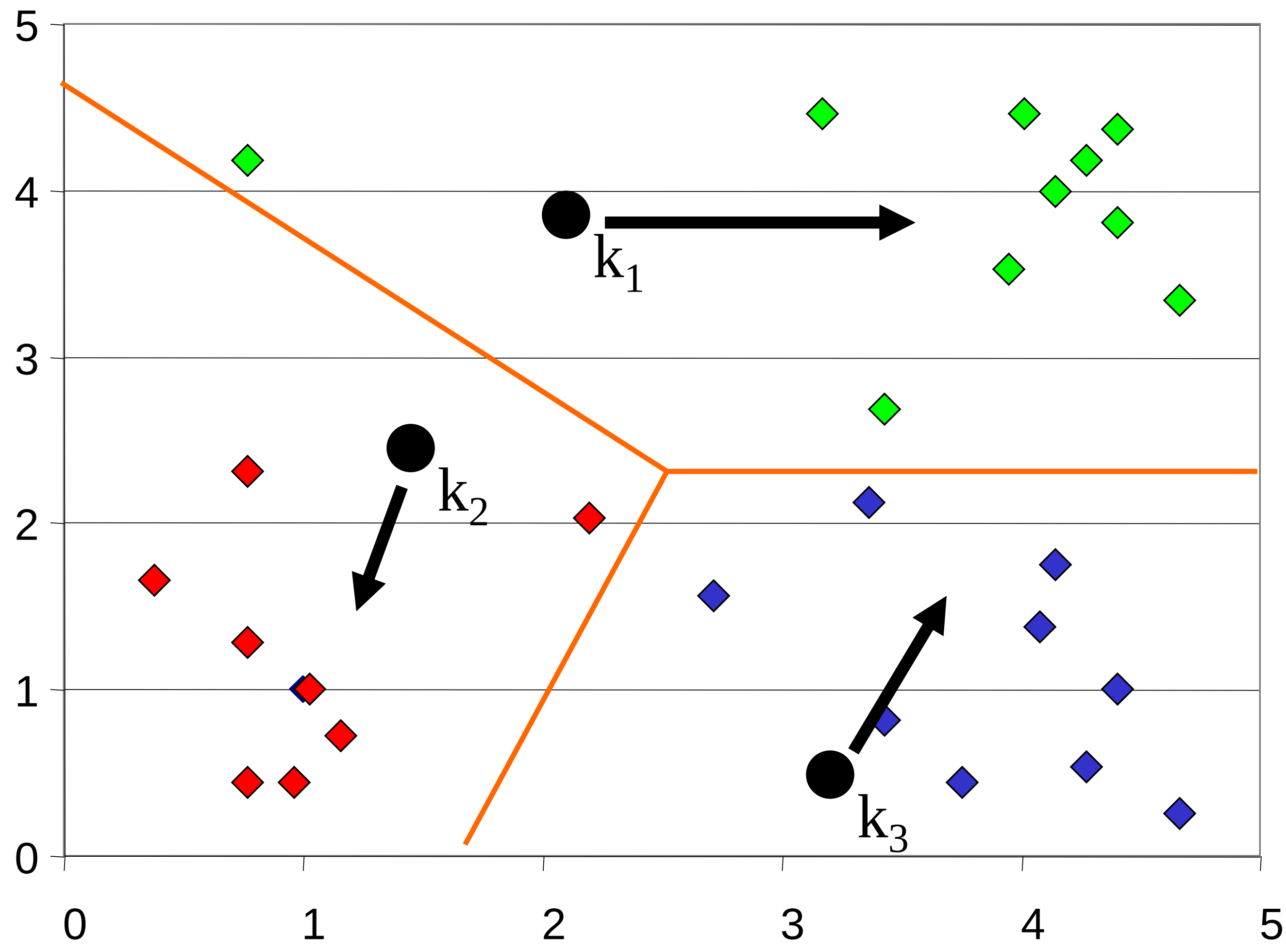
K-means Clustering: Step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



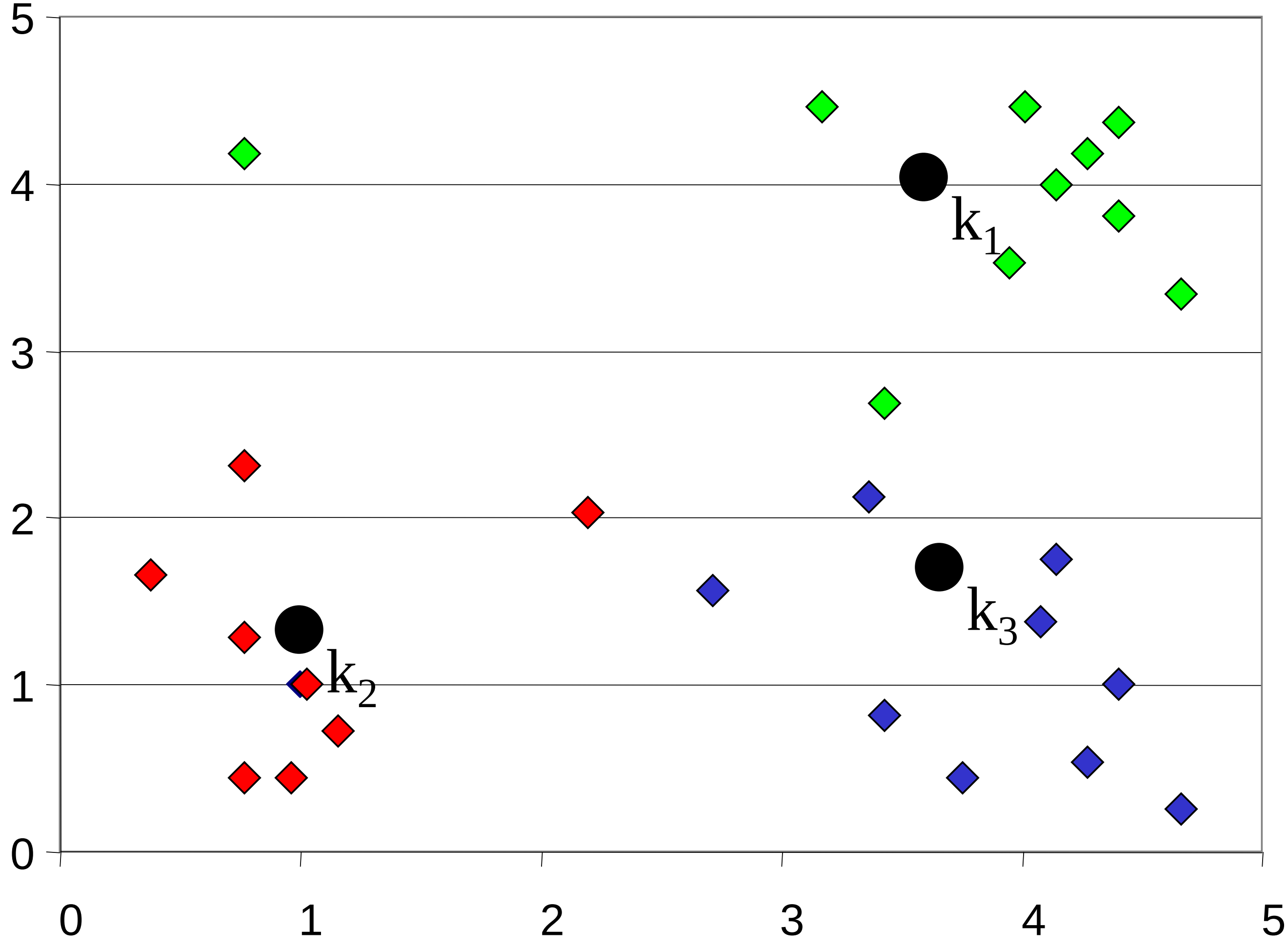
K-means Clustering: Step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



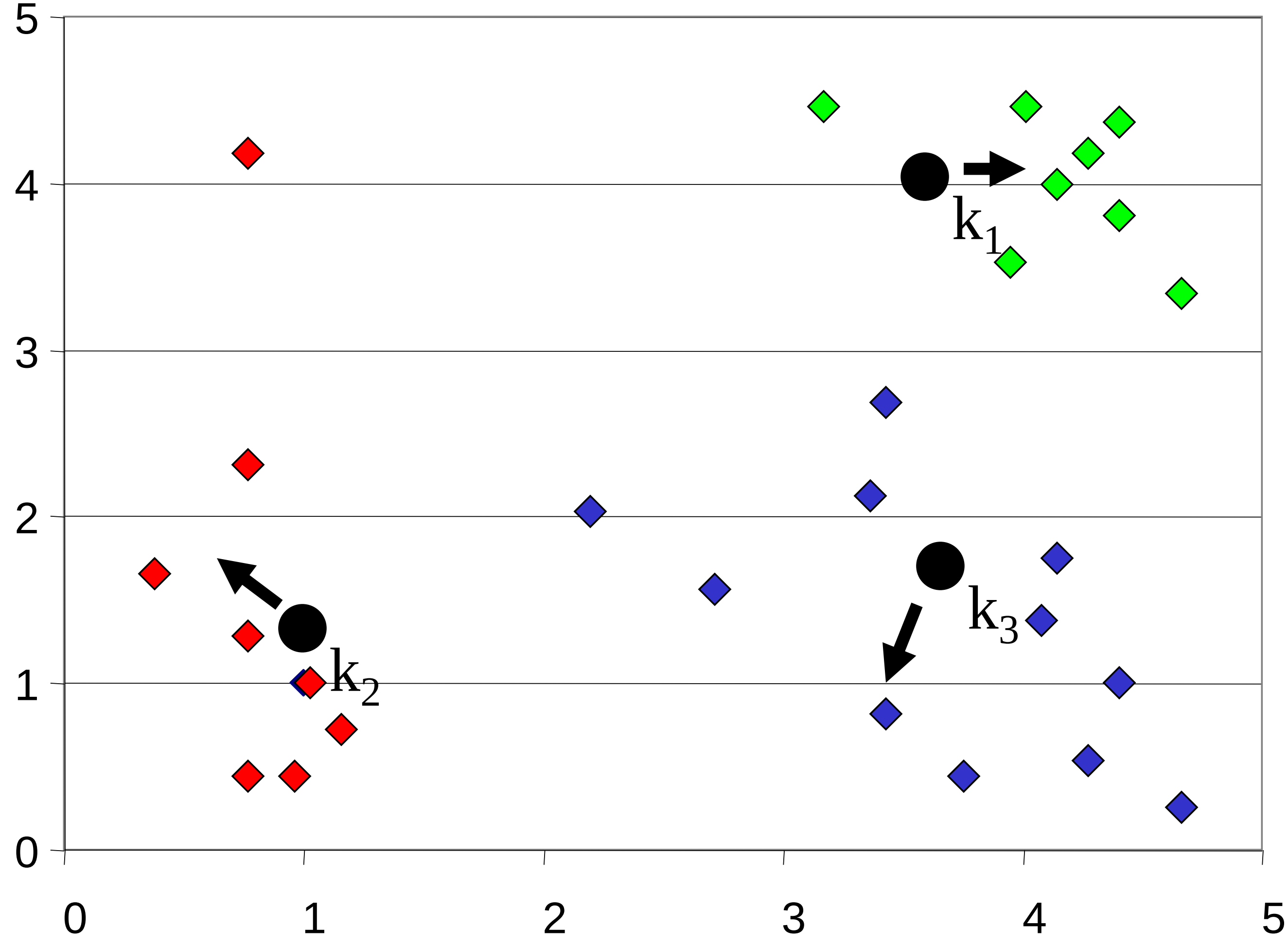
K-means Clustering: Step 3

Algorithm: k-means, Distance Metric: Euclidean Distance



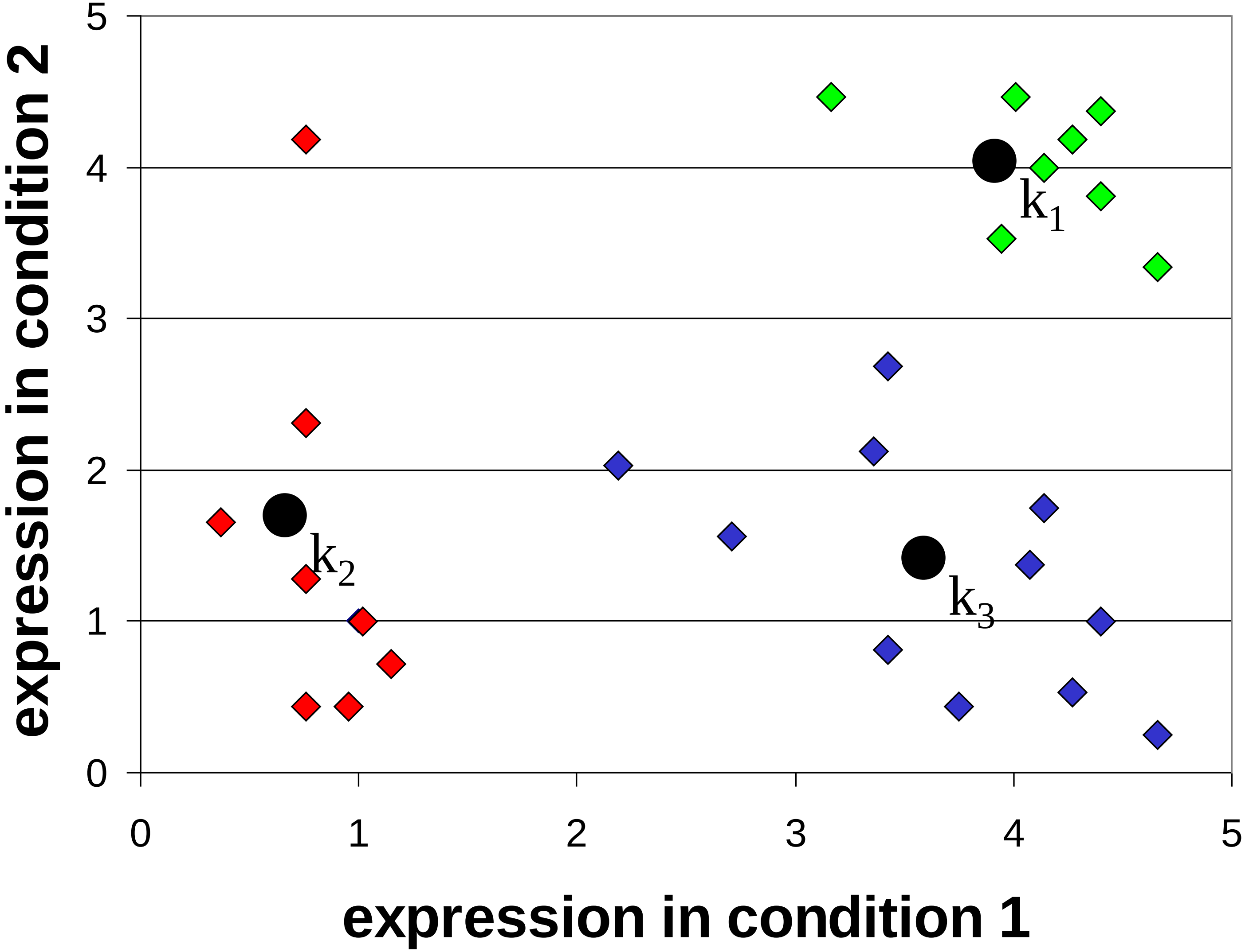
K-means Clustering: Step 4

Algorithm: k-means, Distance Metric: Euclidean Distance



K-means Clustering: Step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



http://mlehman.github.io/kmeans-javascript/

Comments on the *K-Means* Method

- Strength
  - *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, $t \ll n$.
  - Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- 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*

The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets
- **Medoids** are representative objects of a data set or a cluster with a data set whose average dissimilarity to all the objects in the cluster is minimal. **Medoids** are similar in concept to means or centroids, but **medoids** are always members of the data set.

EM Algorithm

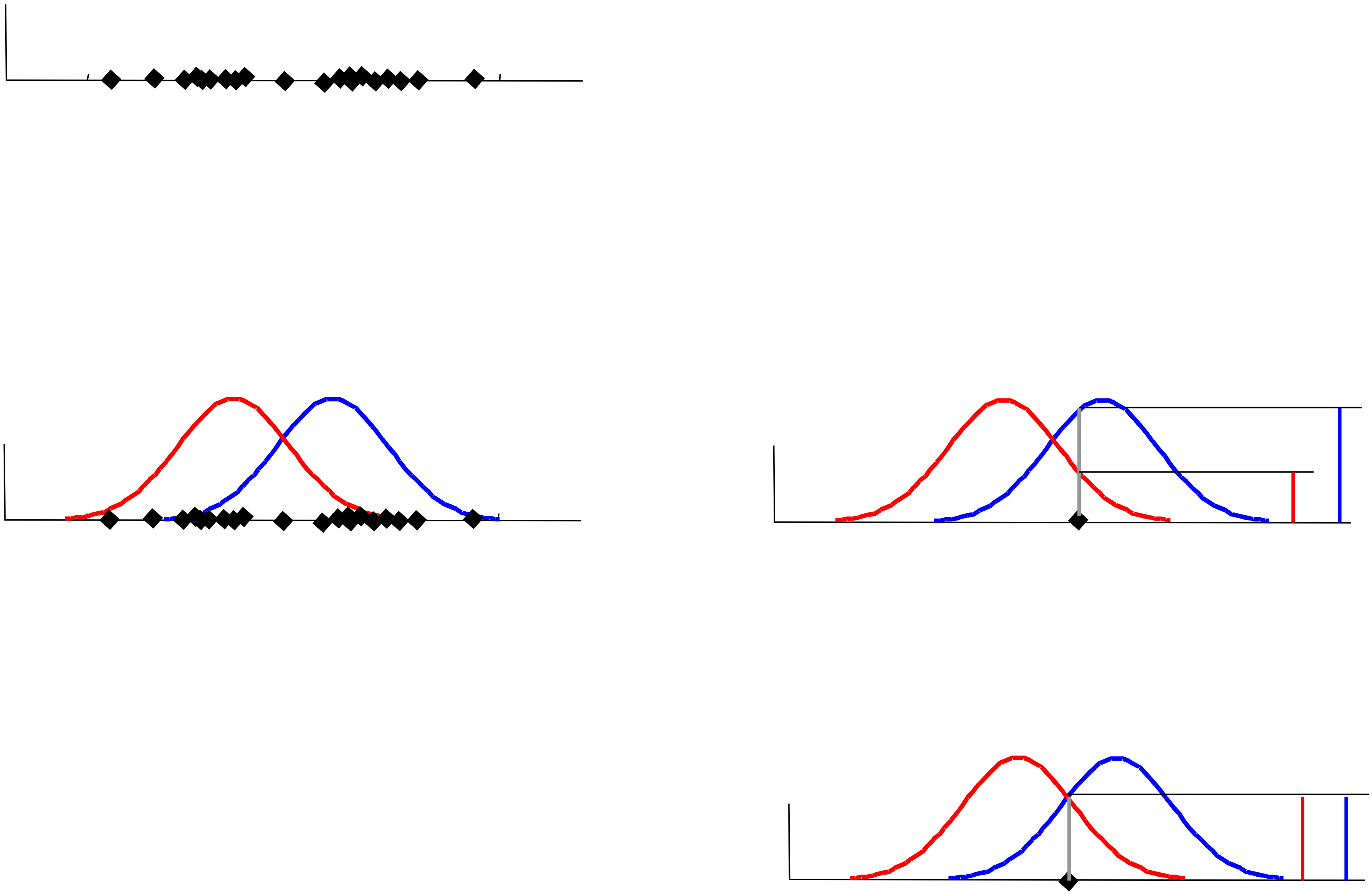
- Initialize K cluster centers
- Iterate between two steps
  - **E**xpectation step: assign points to clusters

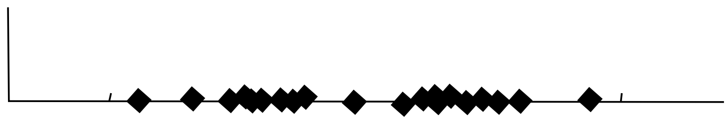
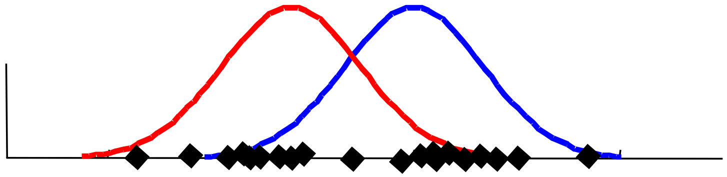
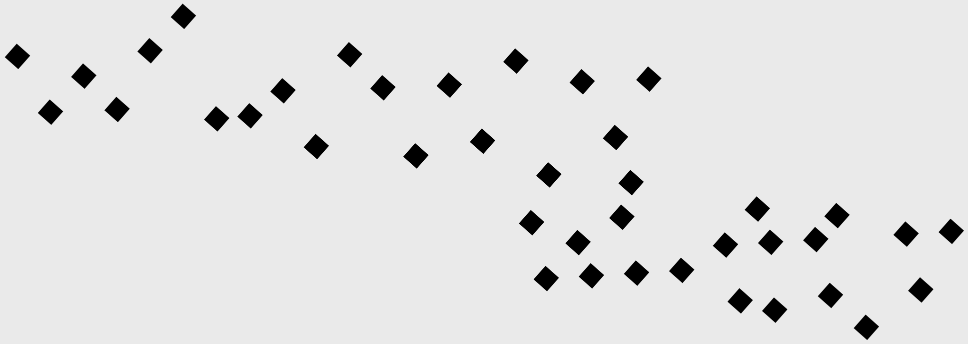
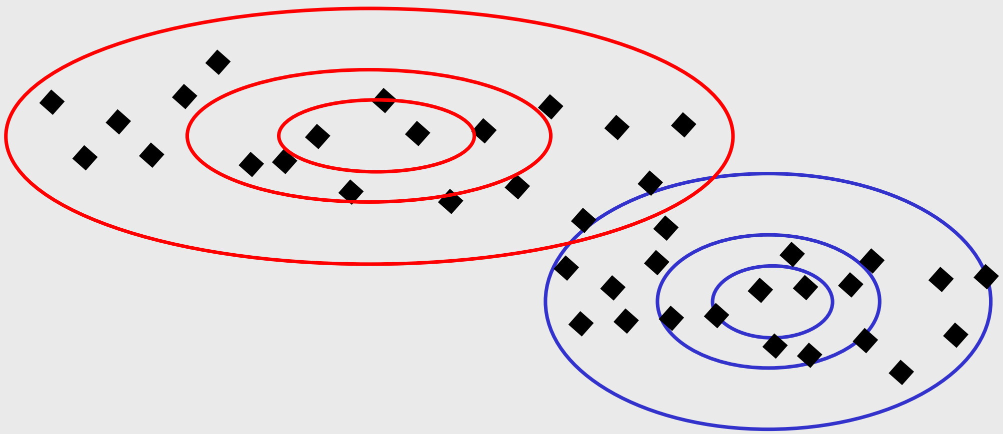
$$P(d_i \in c_k) = w_k \Pr(d_i \mid c_k) \Big/ \sum_j w_j \Pr(d_i \mid c_j)$$

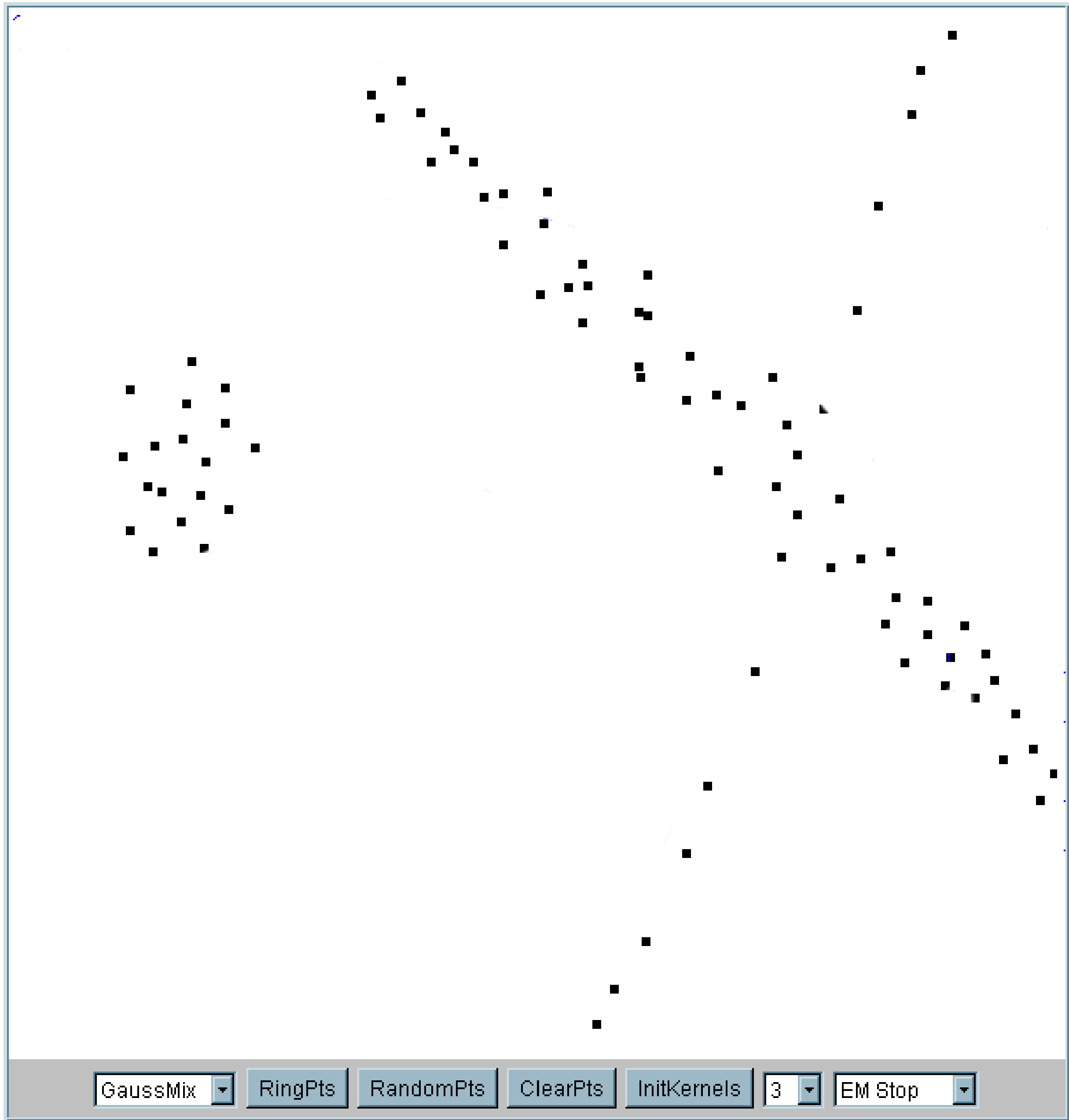
$$w_k = \frac{\sum_i \Pr(d_i \in c_k)}{N}$$

  - **M**aximation step: estimate model parameters

$$\mu_k = \frac{1}{m} \sum_{i=1}^{m} \frac{d_i P(d_i \in c_k)}{\sum_k P(d_i \in c_j)}$$

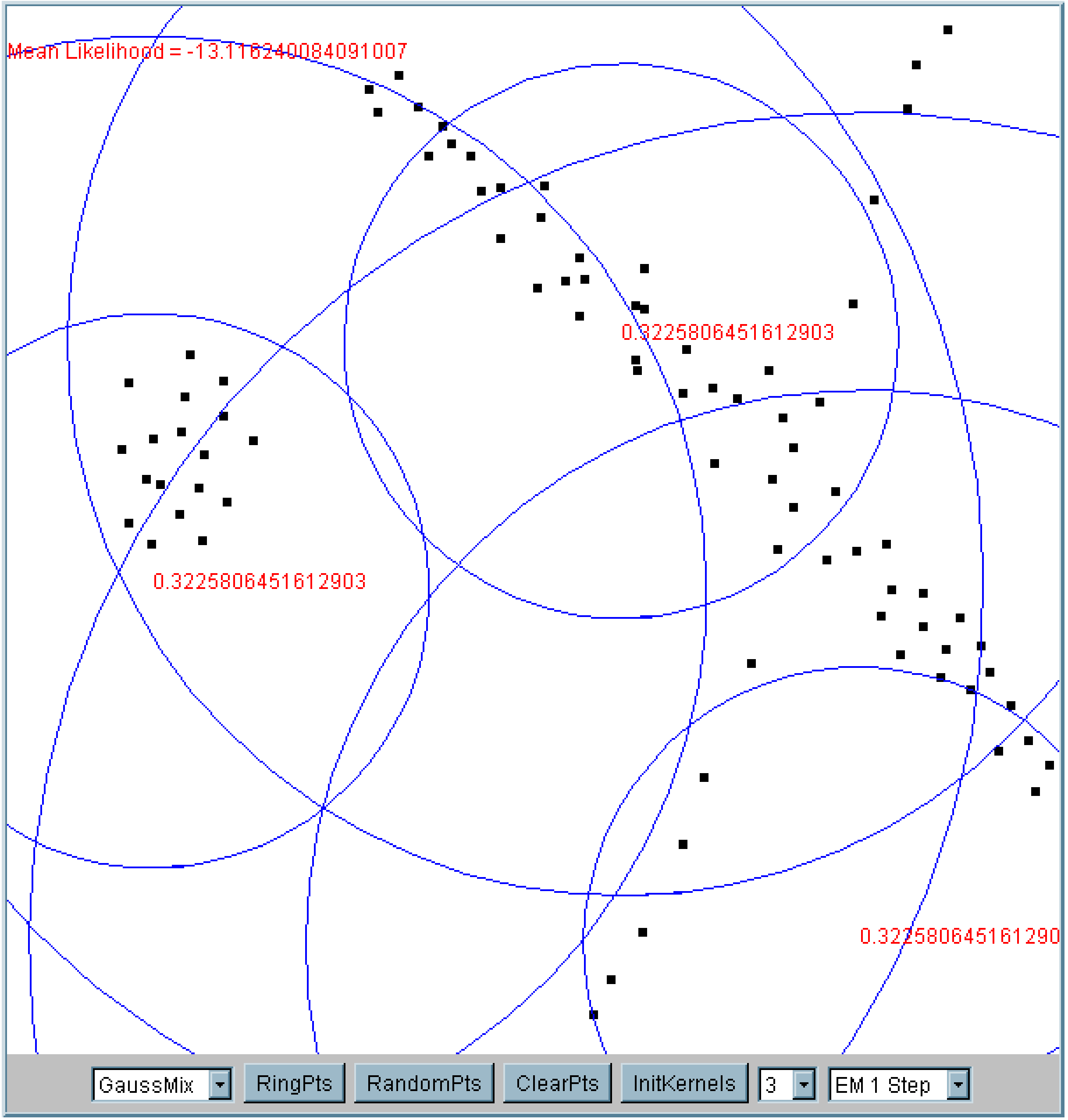


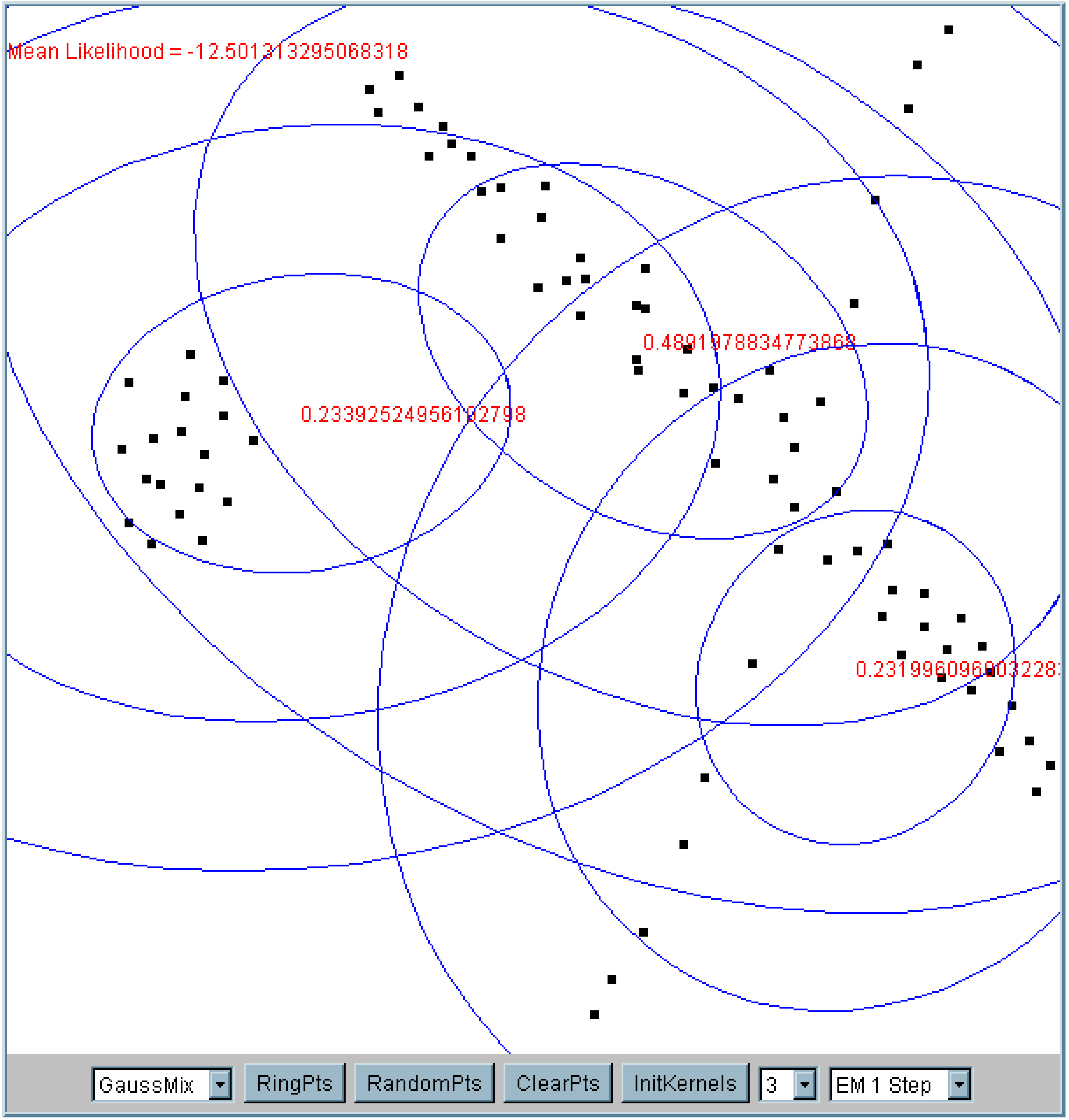

GaussMix
RingPts
RandomPts
ClearPts
InitKernels
3
EM Stop

Iteration 1

The cluster means are randomly assigned

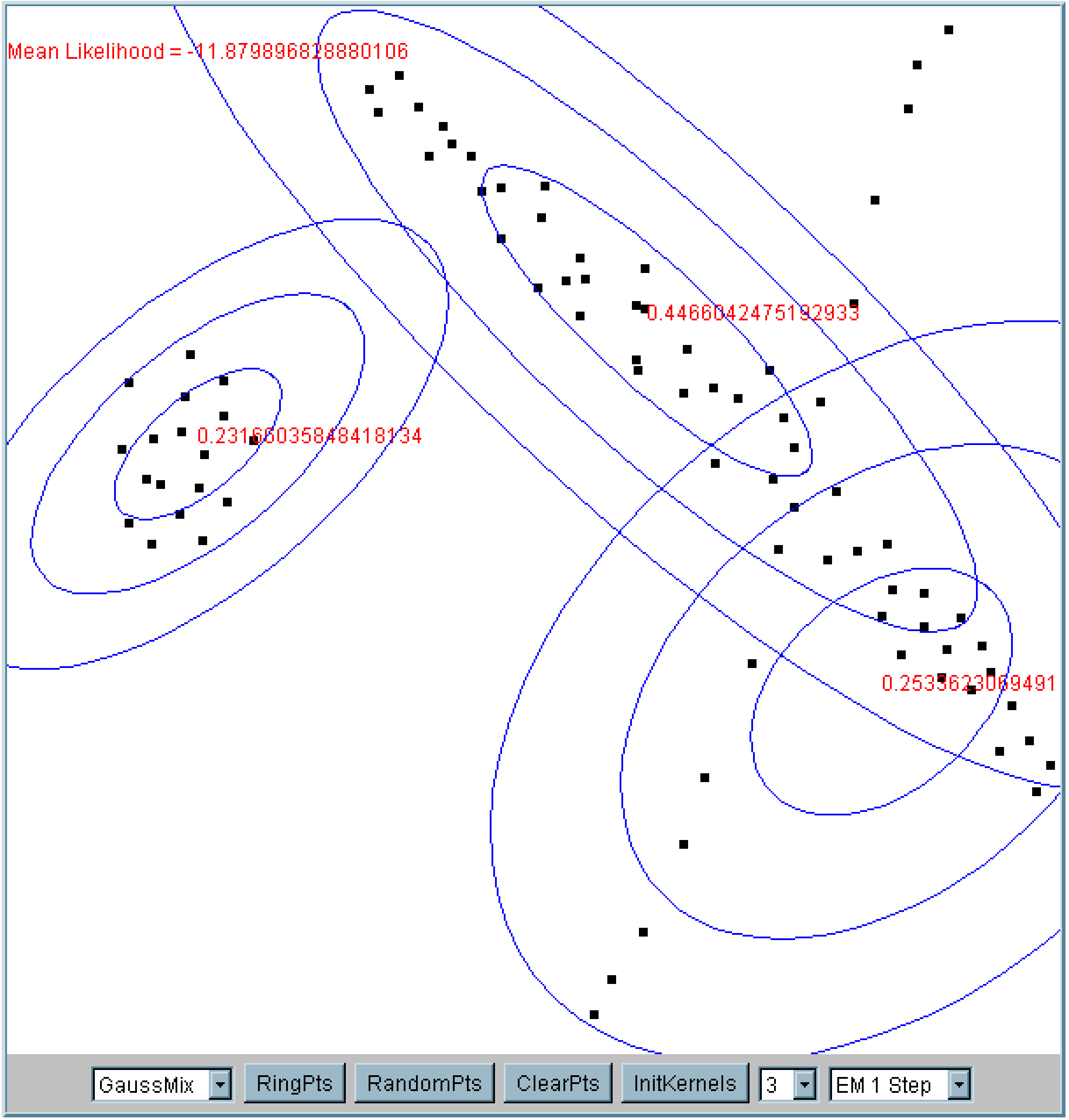


Iteration 2



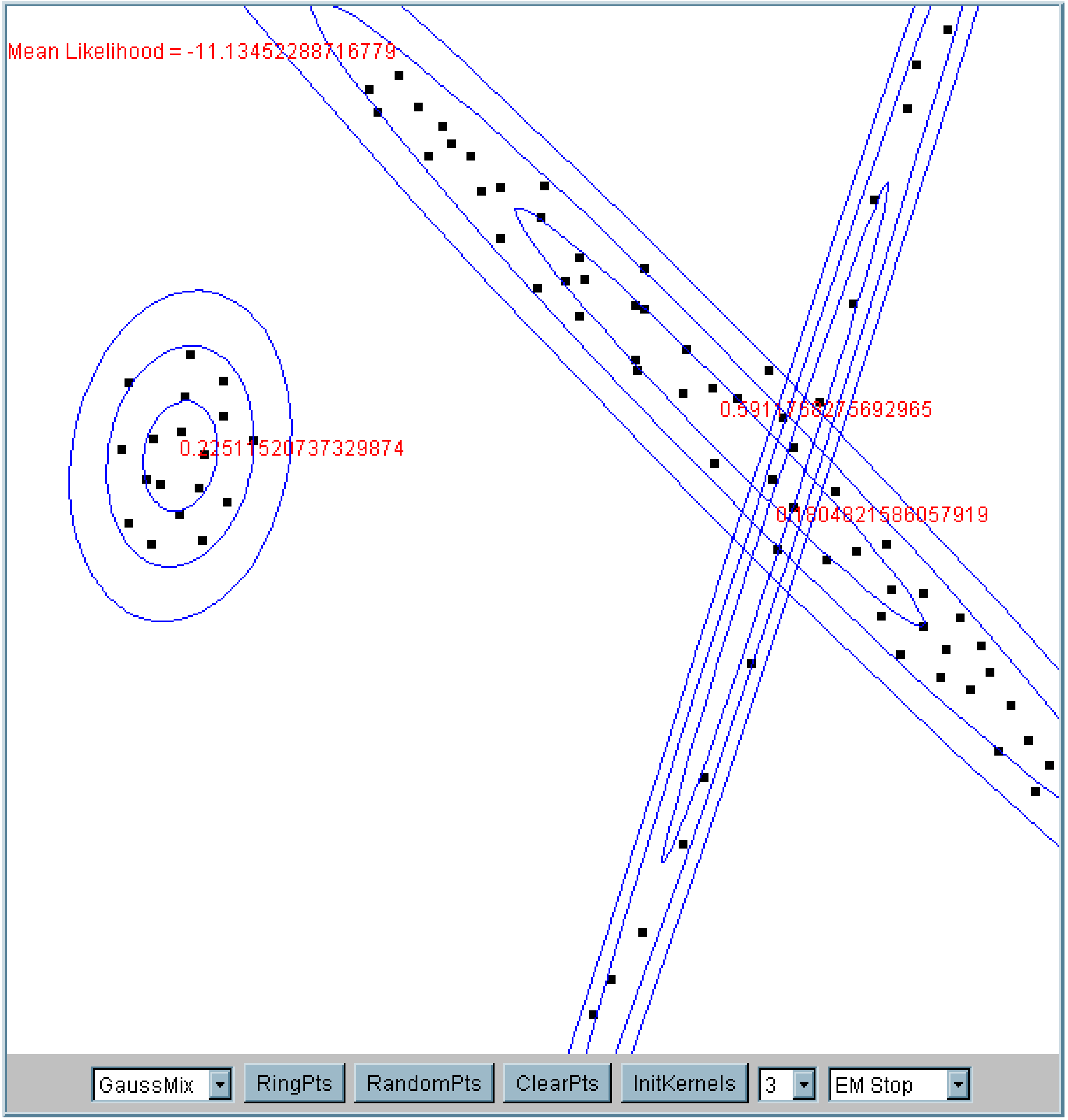
Mean Likelihood = -12.501313295068318
0.4891978834773868
0.2339252495610279
0.231996096003228
GaussMix
RingPts
RandomPts
ClearPts
InitKernels
3
EM 1 Step

Iteration 5



Mean Likelihood = -11.879896828880106
0.4466042475192933
0.2316603584841813 4
0.2536230 69491
GaussMix
RingPts
RandomPts
ClearPts
InitKernels
3
EM 1 Step

Iteration 25



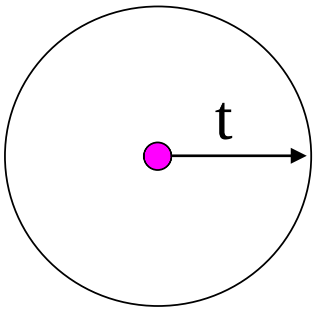
What happens if the data is streaming…

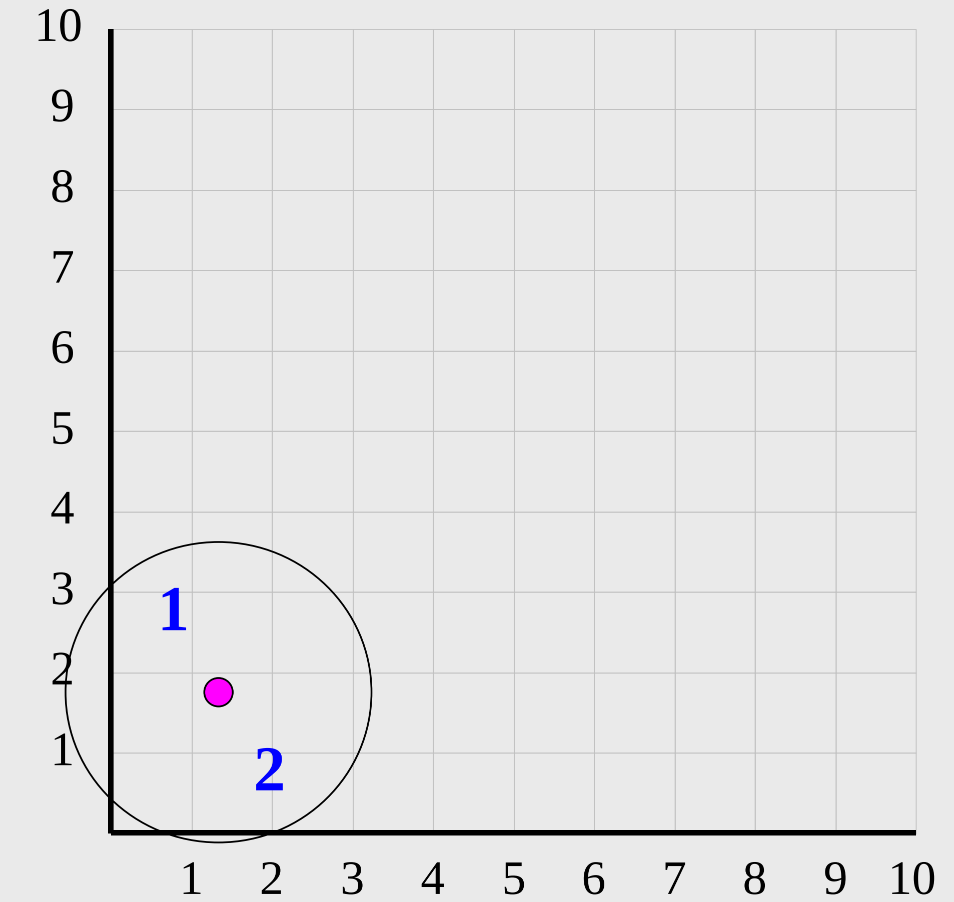
Nearest Neighbor Clustering

Not to be confused with Nearest Neighbor **Classification**

- Items are iteratively merged into the existing clusters that are closest.
- Incremental
- Threshold, t, used to determine if items are added to existing clusters or a new cluster is created.

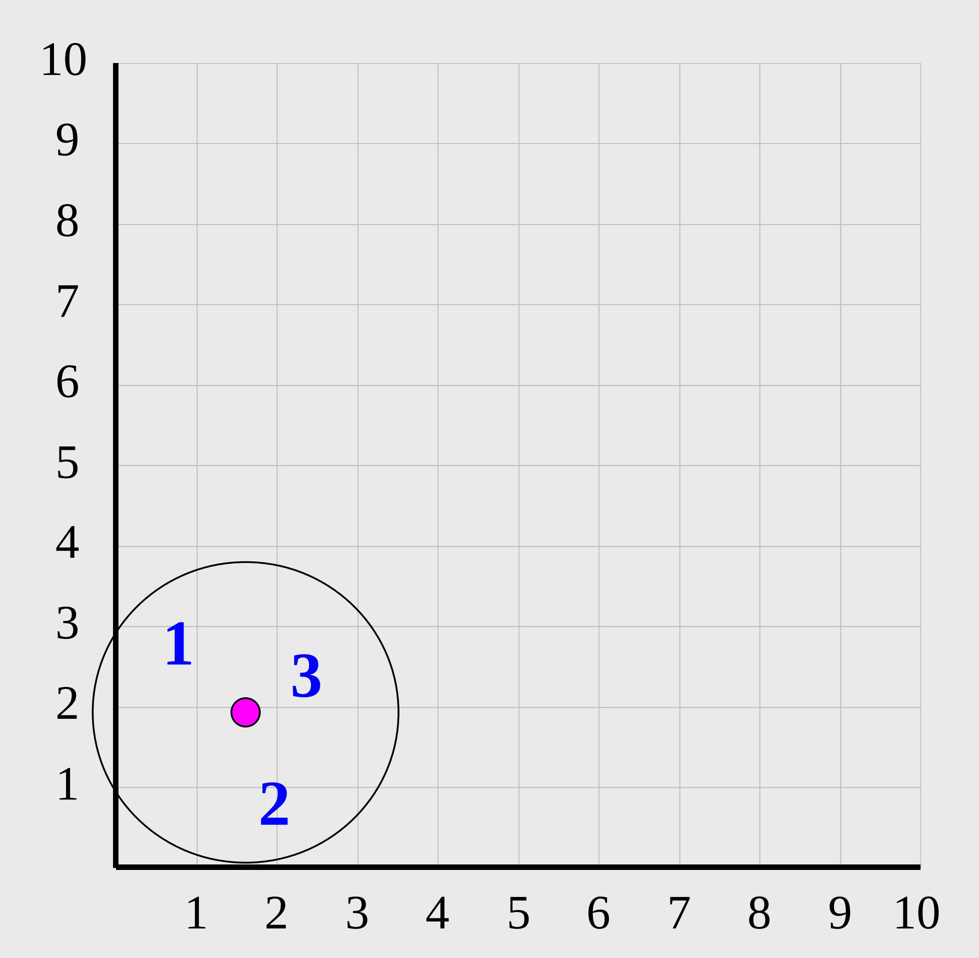
Threshold t





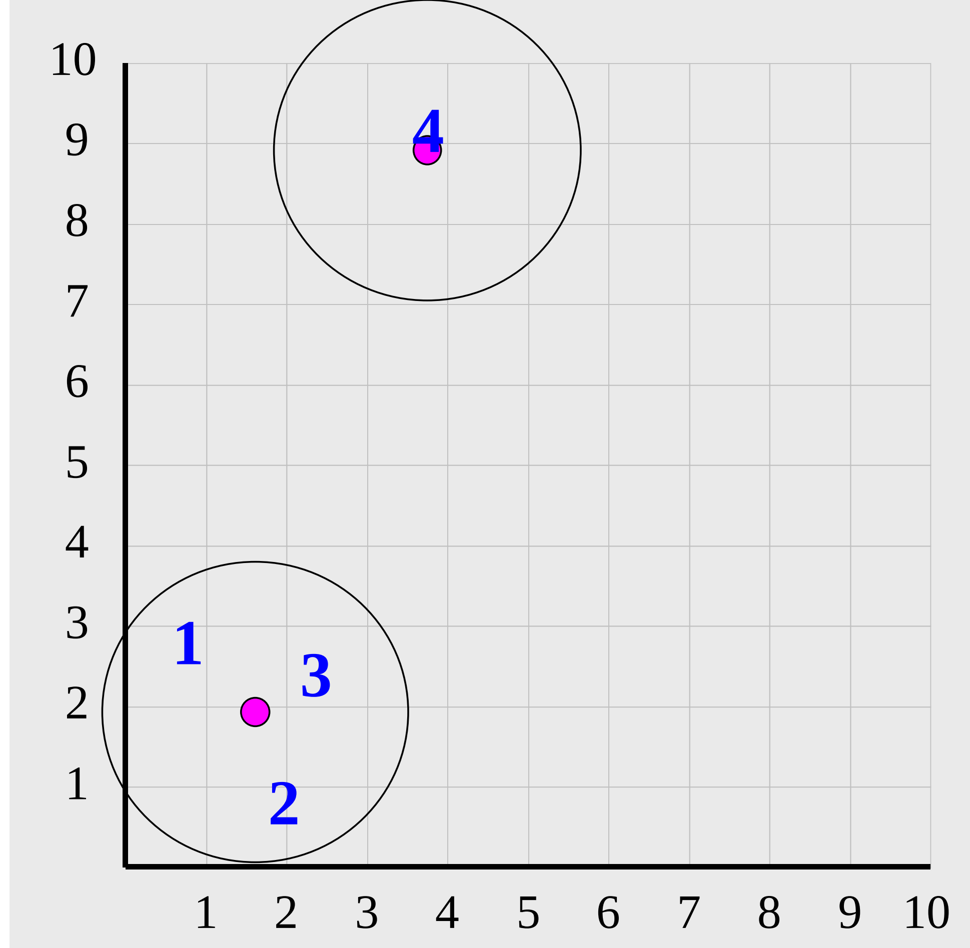
New data point arrives…

It is within the threshold for cluster 1, so add it to the cluster, and update cluster center.



New data point arrives…

It is **not** within the threshold for cluster 1, so create a new cluster, and so on..



Algorithm is highly order dependent…

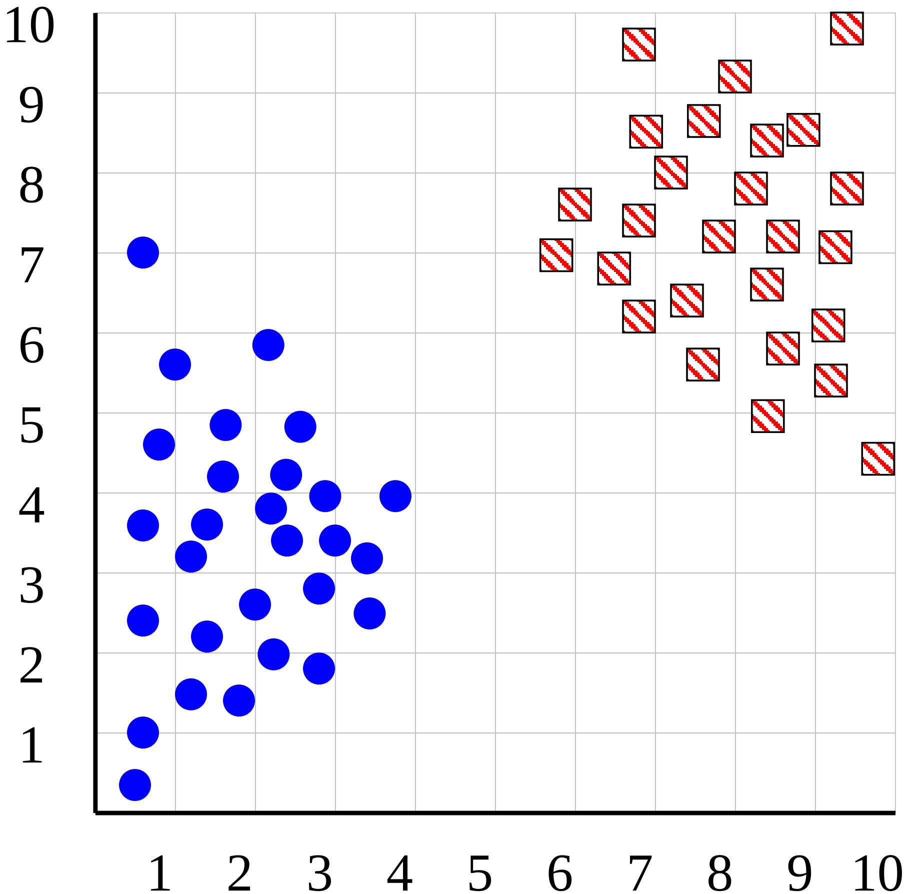
It is difficult to determine t in advance…

Partitional Clustering Algorithms

- Clustering algorithms have been designed to handle very large datasets
- E.g. the Birch algorithm
  - Main idea: use an in-memory R-tree to store points that are being clustered
  - Insert points one at a time into the R-tree, merging a new point with an existing cluster if is less than some δ distance away
  - If there are more leaf nodes than fit in memory, merge existing clusters that are close to each other
  - At the end of first pass we get a large number of clusters at the leaves of the R-tree
    - Merge clusters to reduce the number of clusters

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

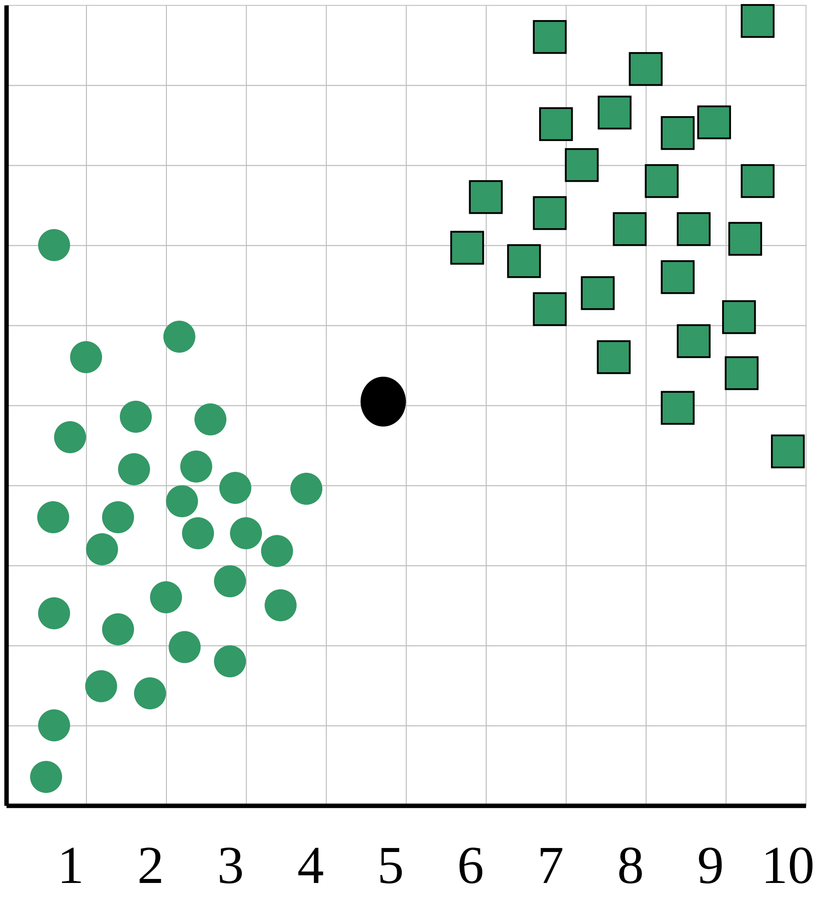
In general, this is a unsolved problem. However there are many approximate methods. In the next few slides we will see an example.



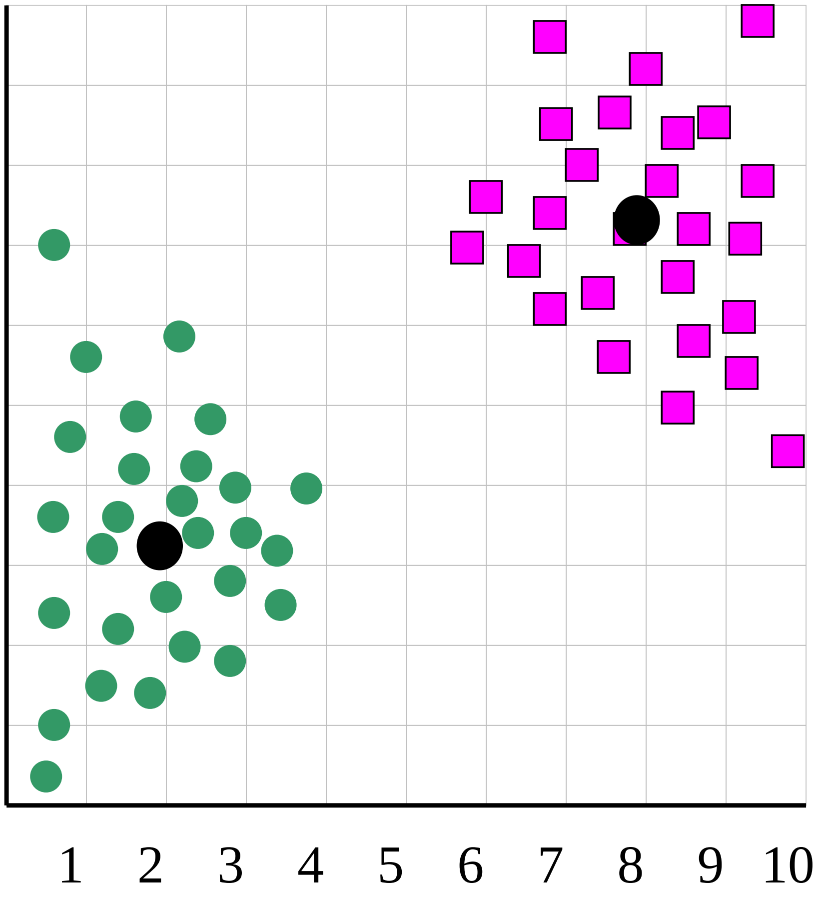
For our example, we will use the familiar katydid/grasshopper dataset.

However, in this case we are imagining that we do NOT know the class labels. We are only clustering on the X and Y axis values.

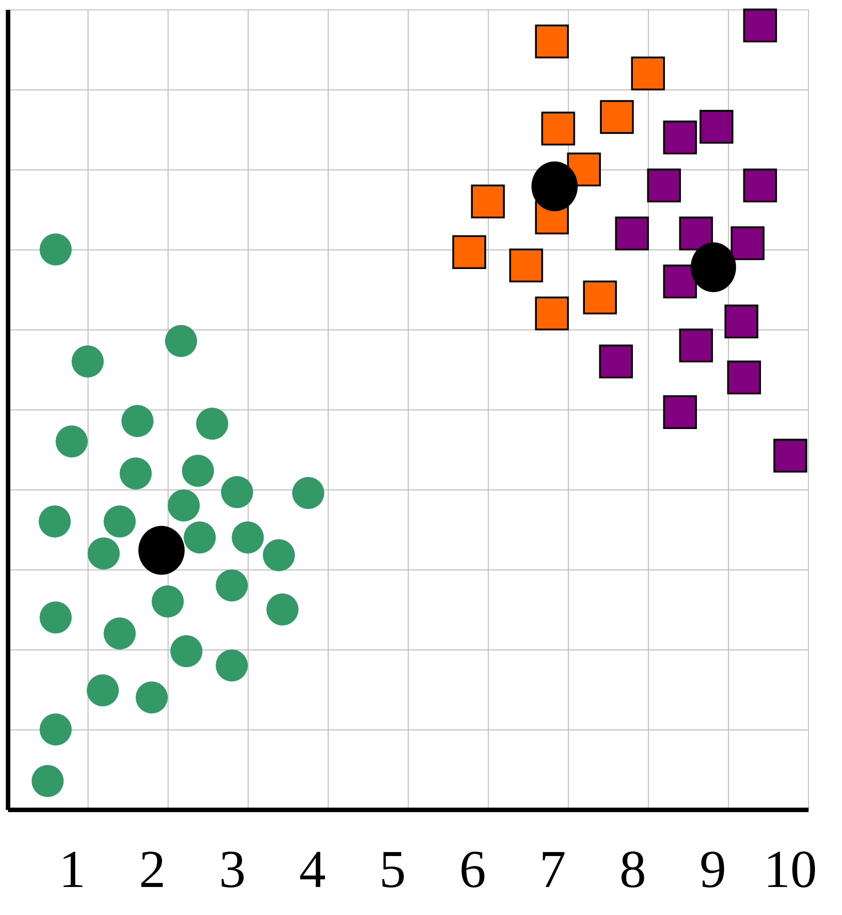
When k = 1, the objective function is 873.0



When k = 2, the objective function is 173.1

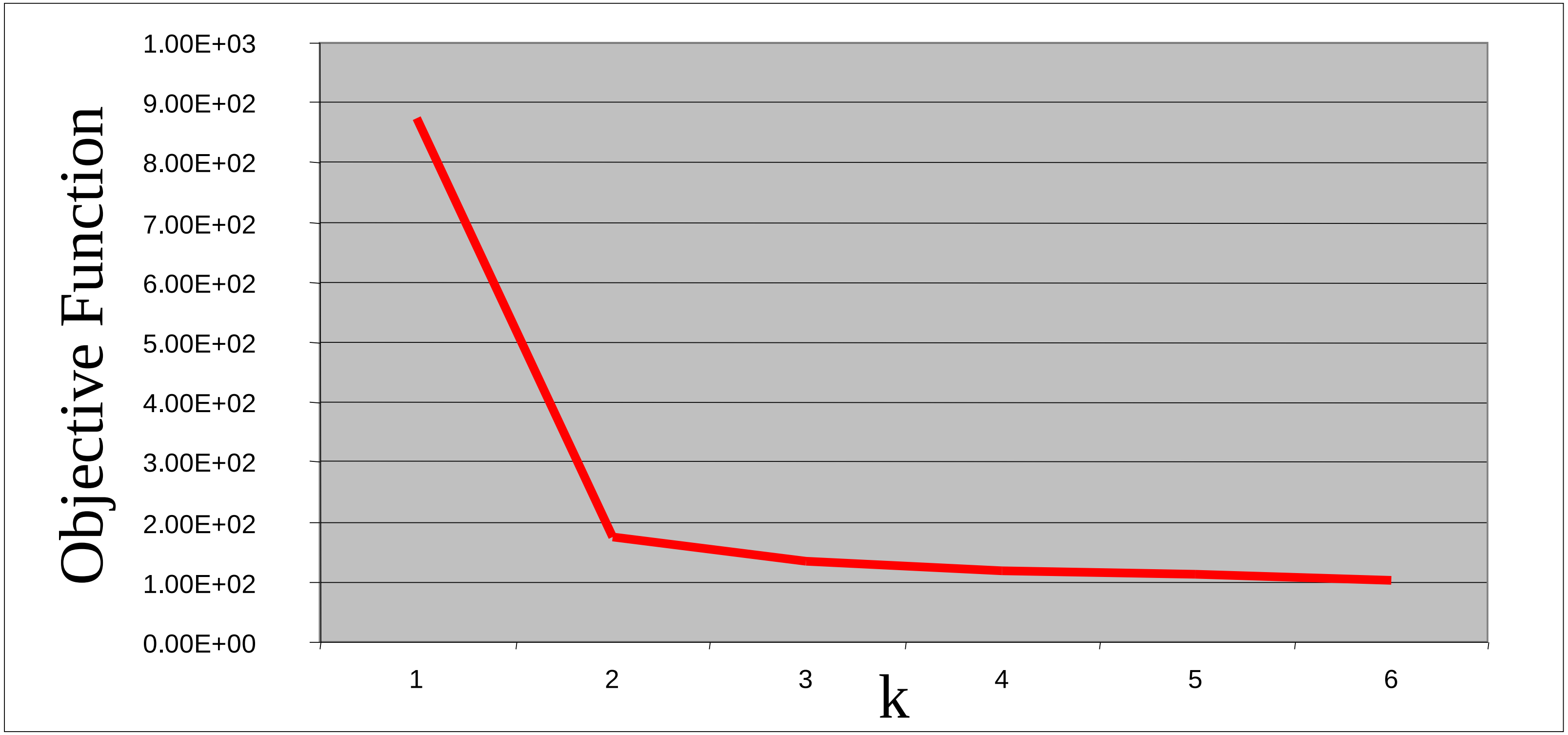


When k = 3, the objective function is 133.6



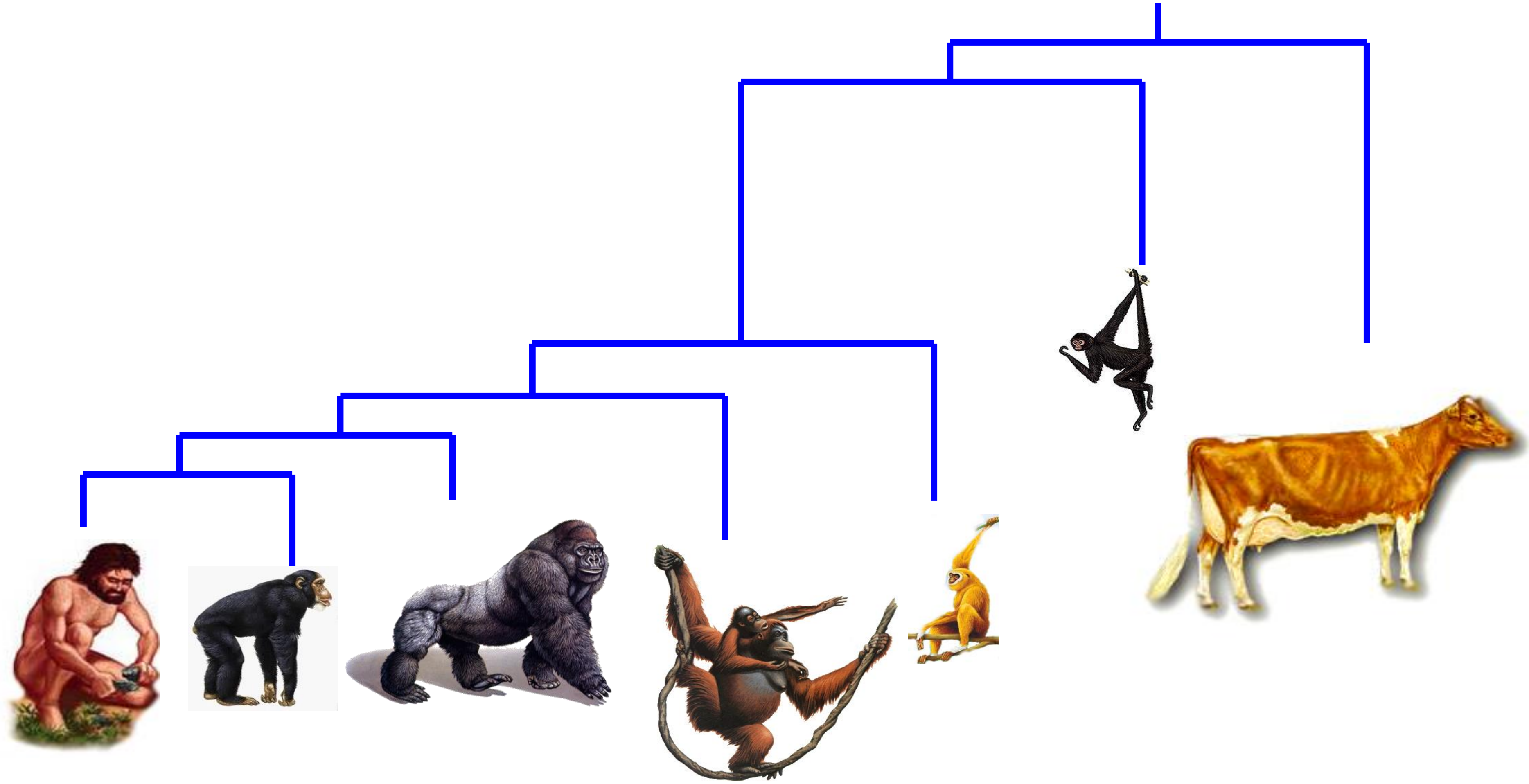
We can plot the objective function values for k equals 1 to 6…

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”.



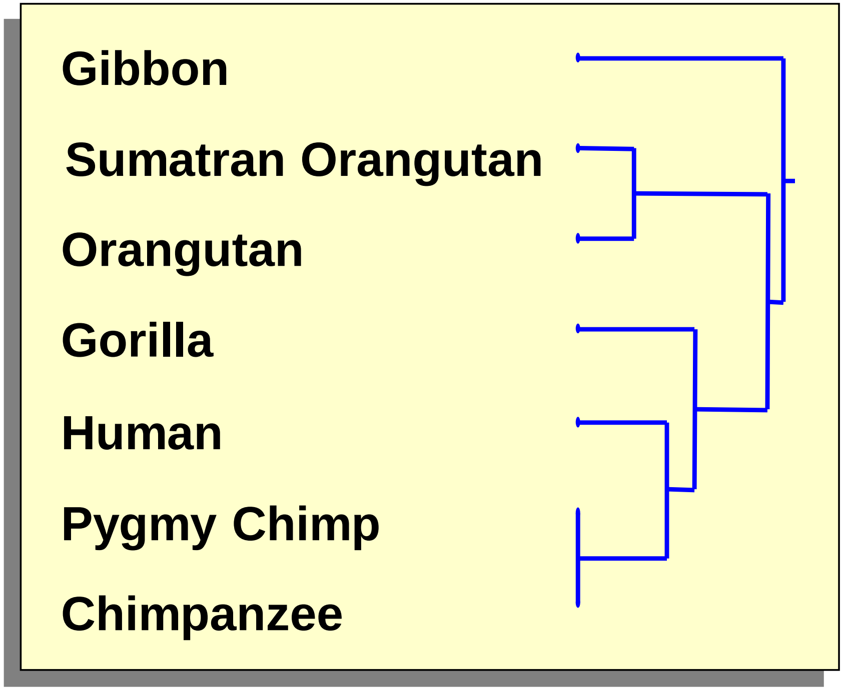
Note that the results are not always as clear cut as in this toy example

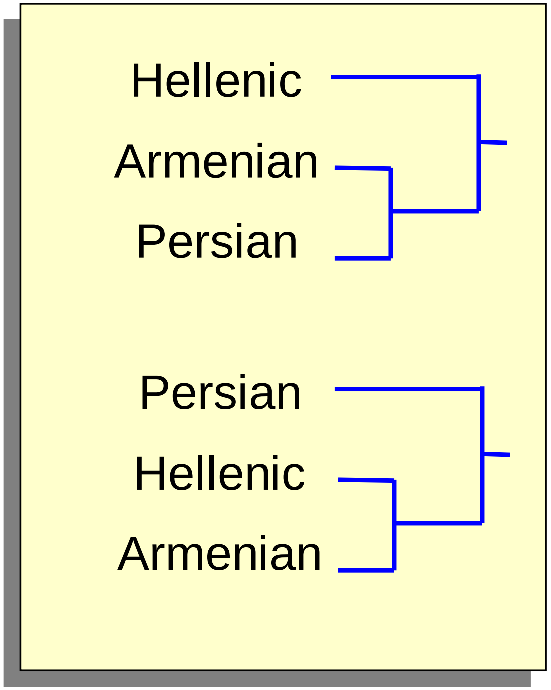
There is only one dataset that can be perfectly clustered using a hierarchy…



(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);

Do Trees Make Sense for non-Biological Objects?





"*Armenian borrowed so many words from Iranian languages that it was at first considered a branch of the Indo-Iranian languages, and was not recognized as an independent branch of the Indo-European languages for many decades*"

The answer is "Yes".
There are increasing theoretical and empirical results to suggest that phylogenetic methods work for cultural artifacts.

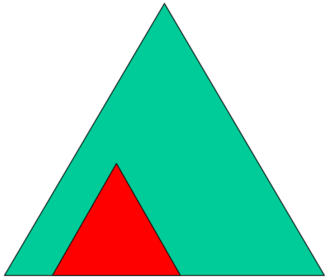
- *Does horizontal transmission invalidate cultural phylogenies?* Greenhill, Currie & Gray.
- *Branching, blending, and the evolution of cultural similarities and differences among human populations*. Collard, Shennan, & Tehrani.

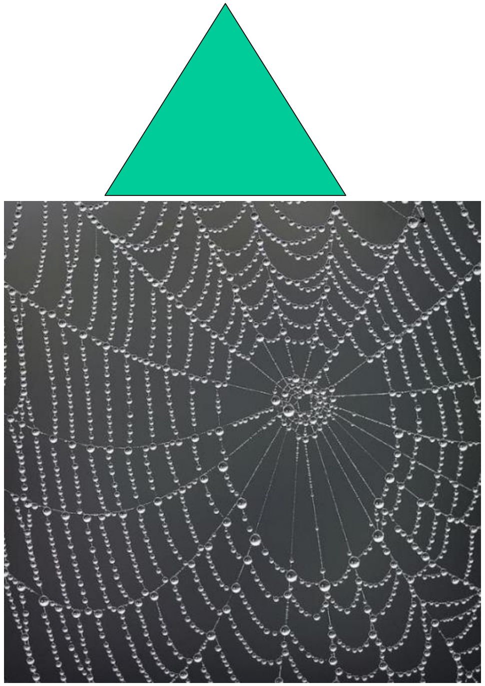
..*results show that trees constructed with Bayesian phylogenetic methods are robust to realistic levels of borrowing*

Why would we want to use trees for non biological things?

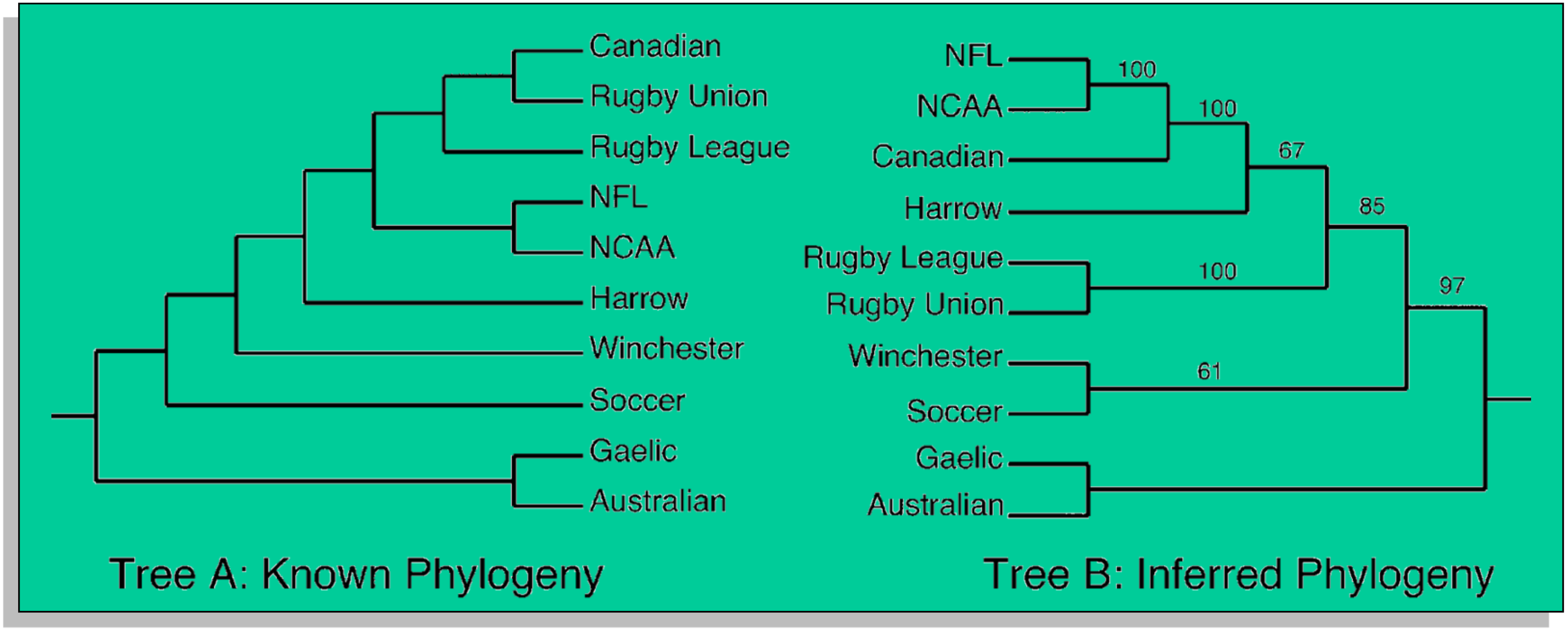
Because trees are powerful in biology

- They make predictions
  - Pacific Yew produces *taxol* which treats some cancers, but it is expensive. Its nearest relative, the European Yew was also found to produce *taxol*.
- They tell us the order of events
  - Which came first, classic geometric spider webs, or messy cobwebs?
- They tell us about..
  - “Homelands”, where did it come from.
  - “Dates” when did it happen.
  - Rates of change
  - Ancestral states





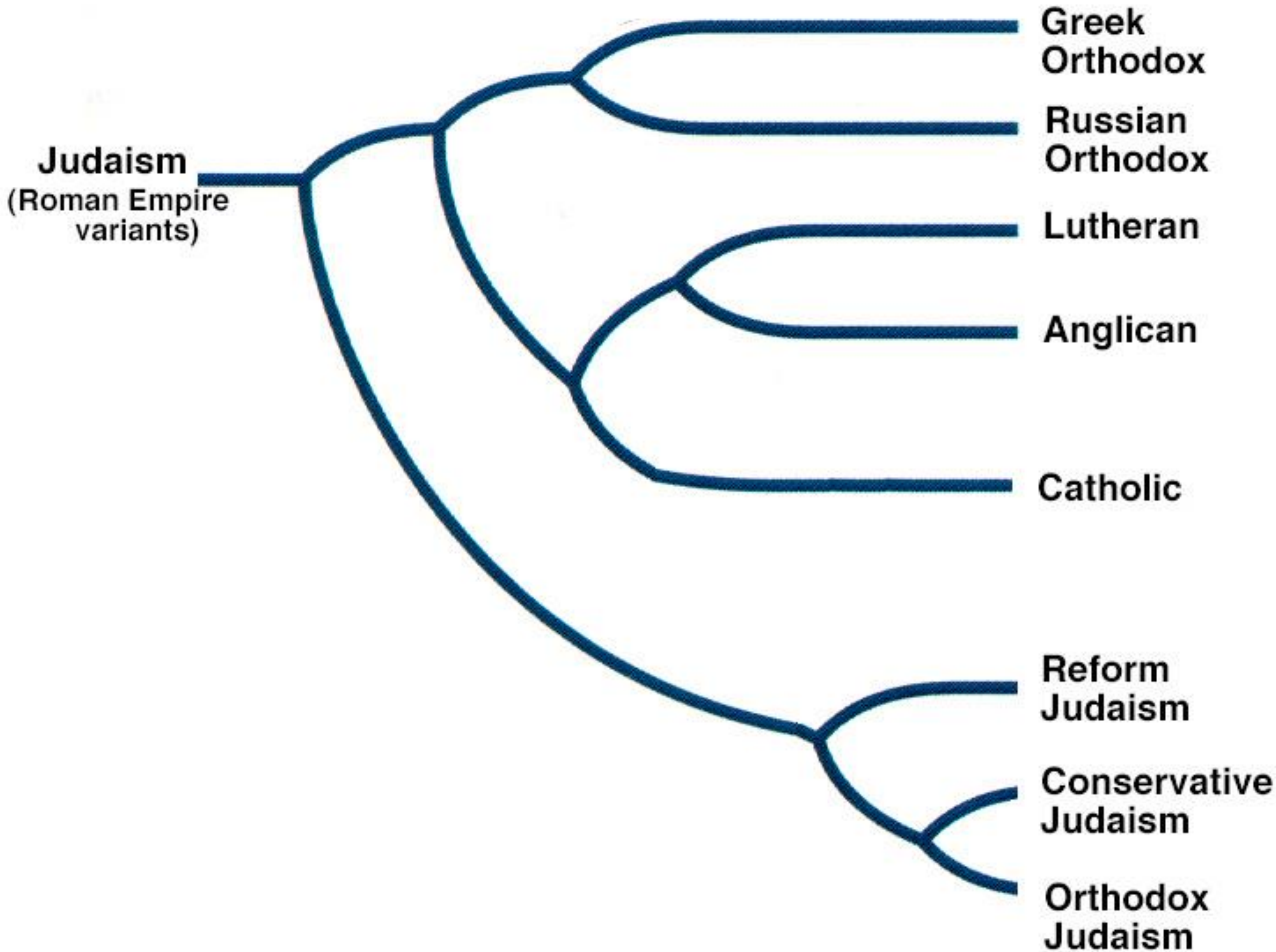
On trick to text the applicably of phylogenetic methods outside of biology is to test on datasets for which you know the right answer by other methods.



"*Canadian Football is historically derived from the ancestor of rugby, but today closely resembles the American versions of the game. In this branch of the tree geography has trumped deeper phylogenetic history*."

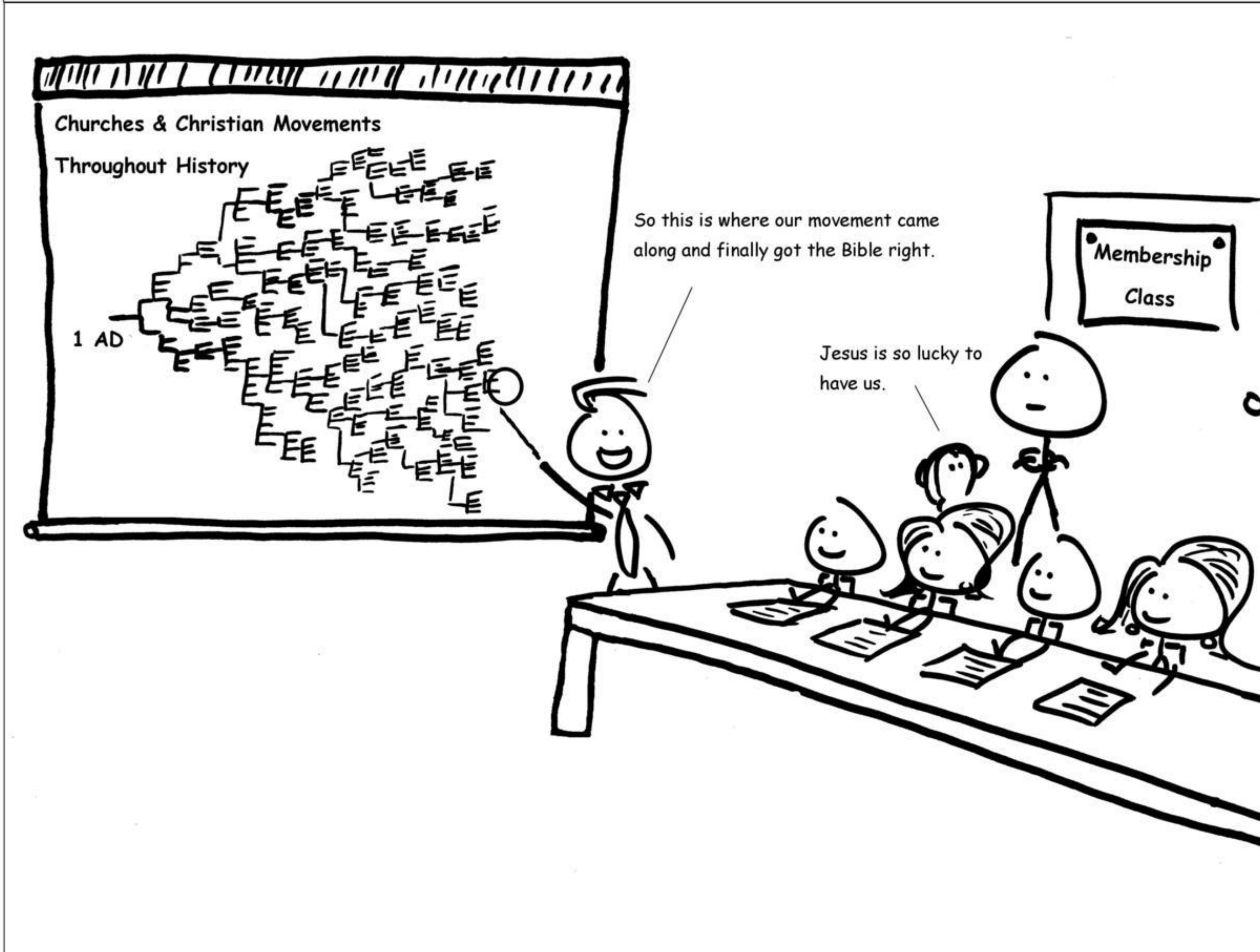
Here the results are very good, but not perfect.

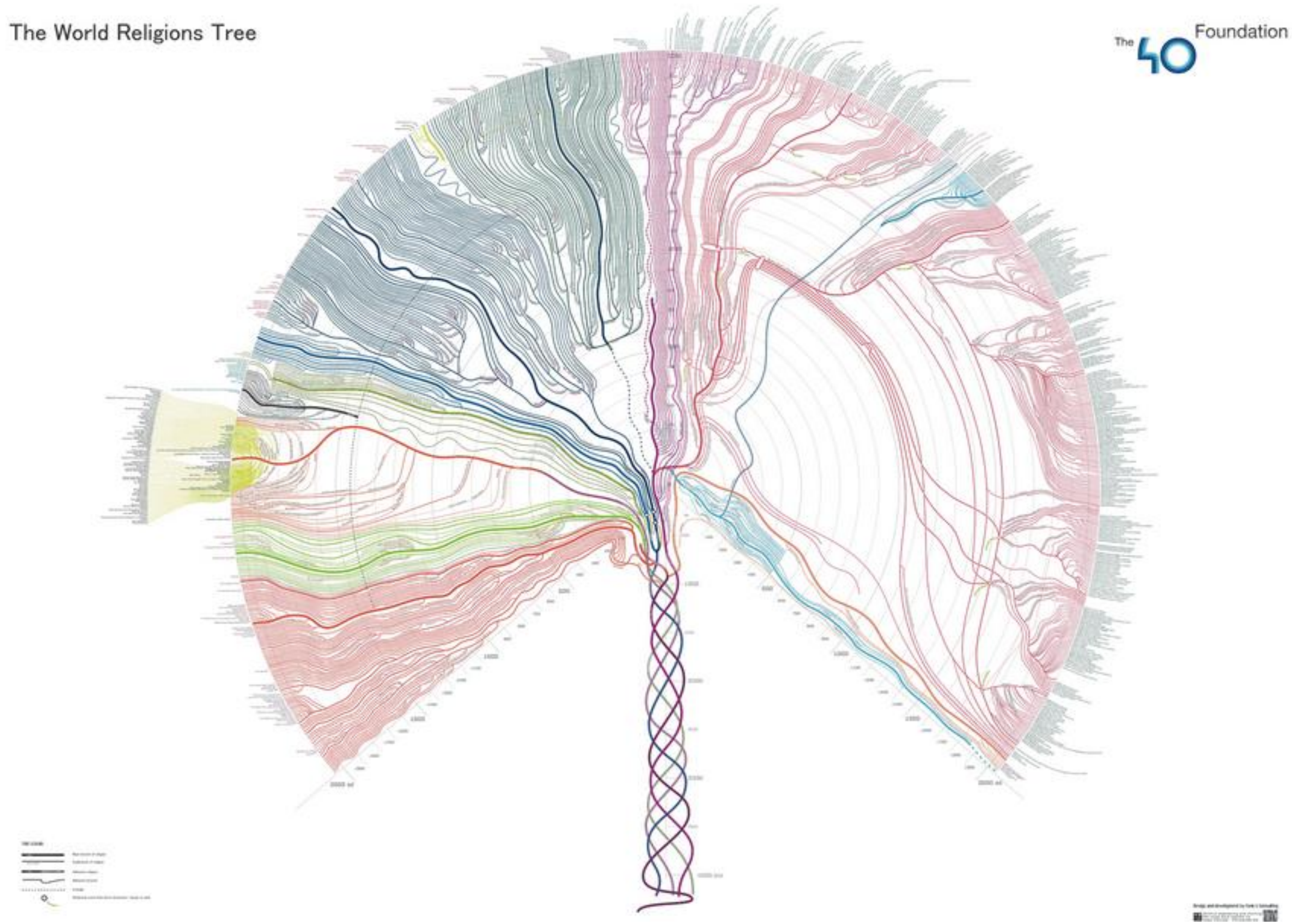
Gray, RD, Greenhill, SJ, & Ross, RM (2007). The Pleasures and Perils of Darwinizing Culture (with phylogenies). Biological Theory, 2(4)


Judaism
(Roman Empire variants)
Greek Orthodox
Russian Orthodox
Lutheran
Anglican
Catholic
Reform Judaism
Conservative Judaism
Orthodox Judaism

Tom's Doubts, #14

by Saji

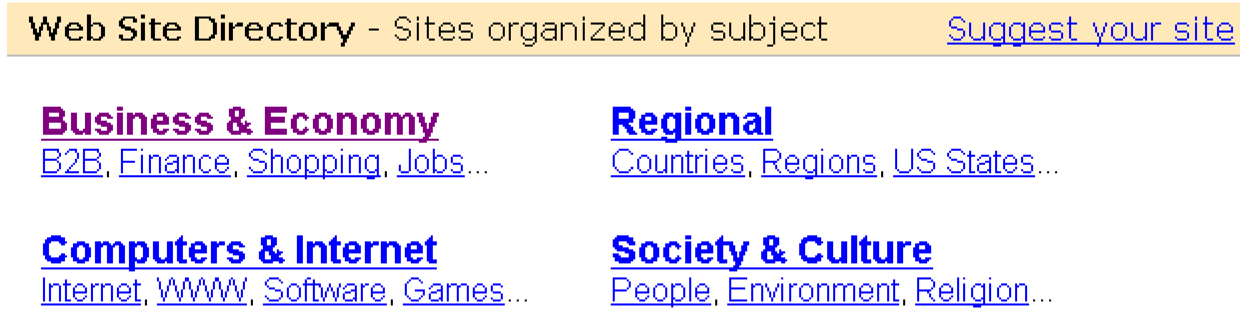


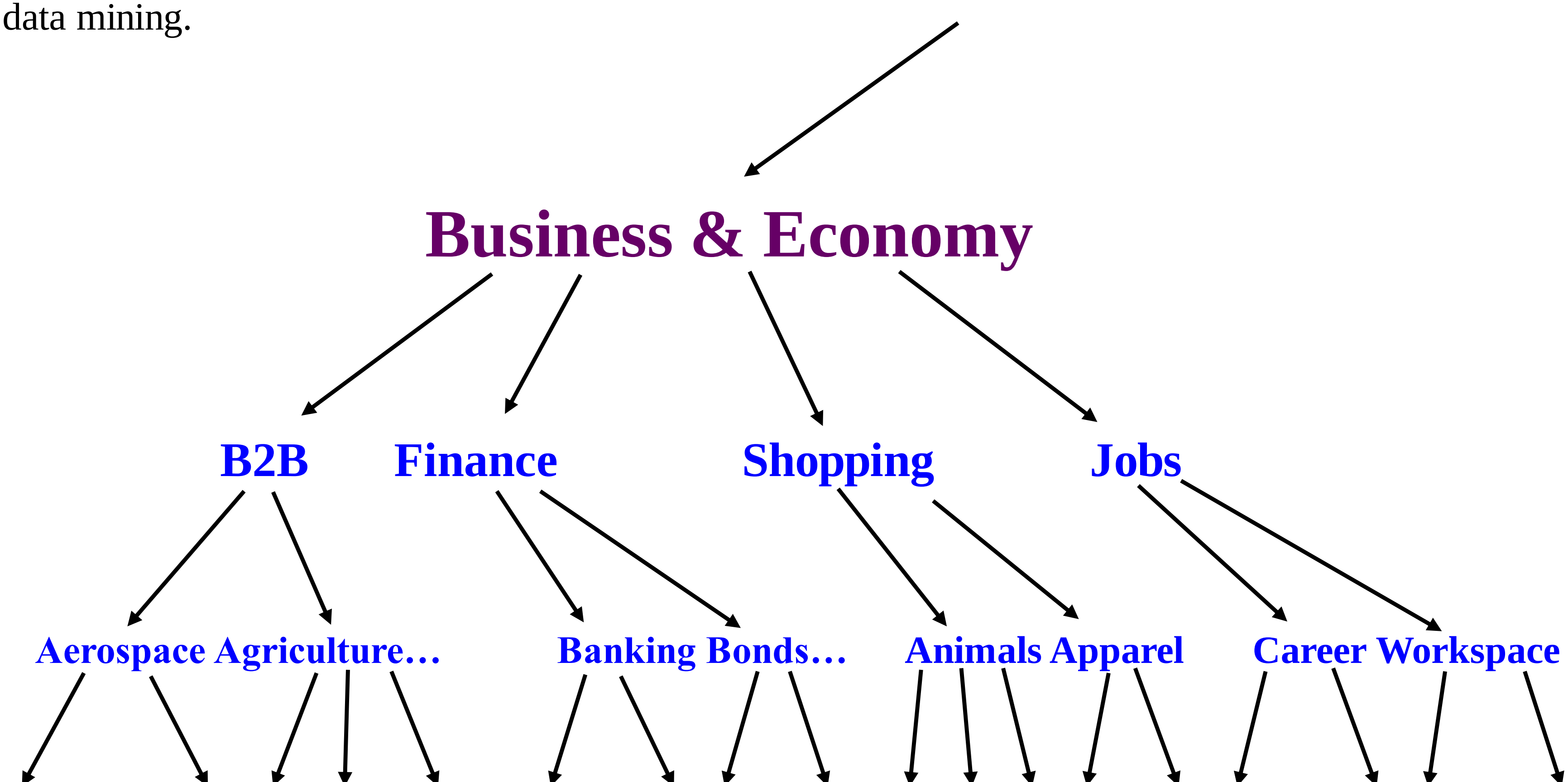


http://000024.org/religions_tree/

Note that hierarchies are commonly used to organize information, for example in a web portal.

Yahoo's hierarchy is manually created, we will focus on automatic creation of hierarchies in data mining.





A Demonstration of Hierarchical Clustering using String Edit Distance

Pedro (Portuguese)

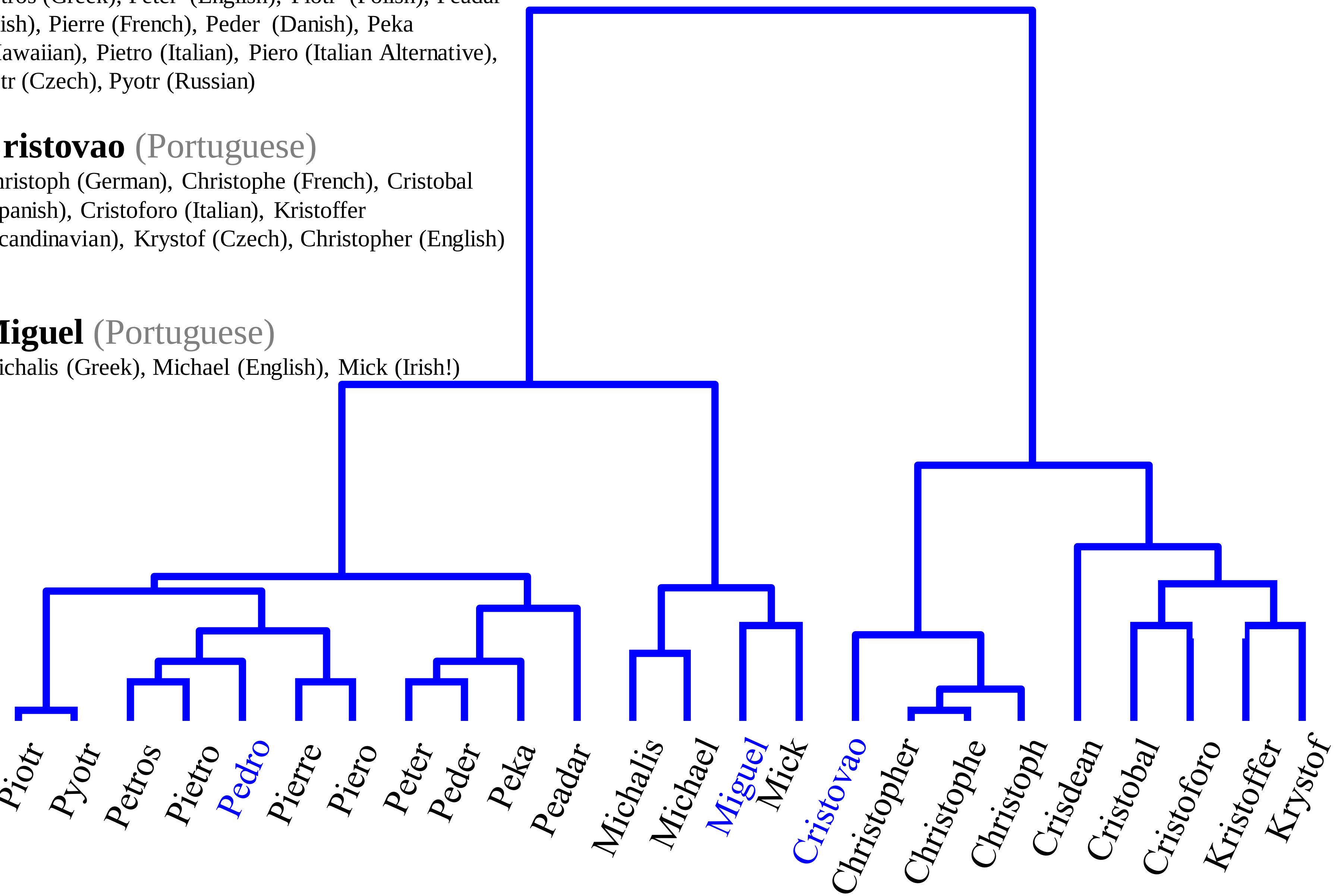
Petros (Greek), Peter (English), Piotr (Polish), Peadar (Irish), Pierre (French), Peder (Danish), Peka (Hawaiian), Pietro (Italian), Piero (Italian Alternative), Petr (Czech), Pyotr (Russian)

Cristovao (Portuguese)

Christoph (German), Christophe (French), Cristobal (Spanish), Cristoforo (Italian), Kristoffer (Scandinavian), Krystof (Czech), Christopher (English)

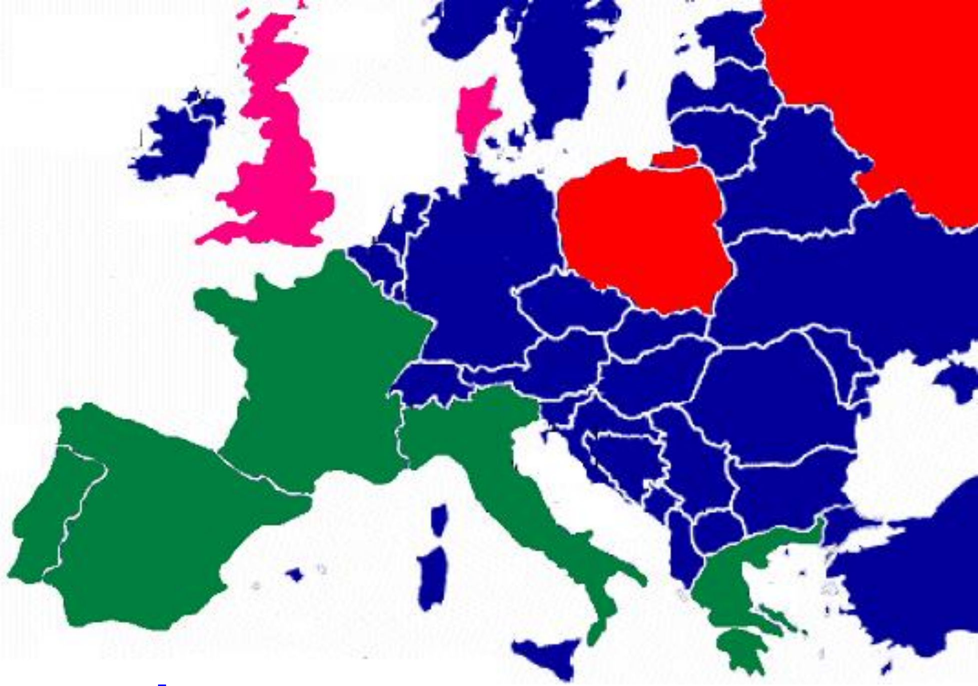
Miguel (Portuguese)

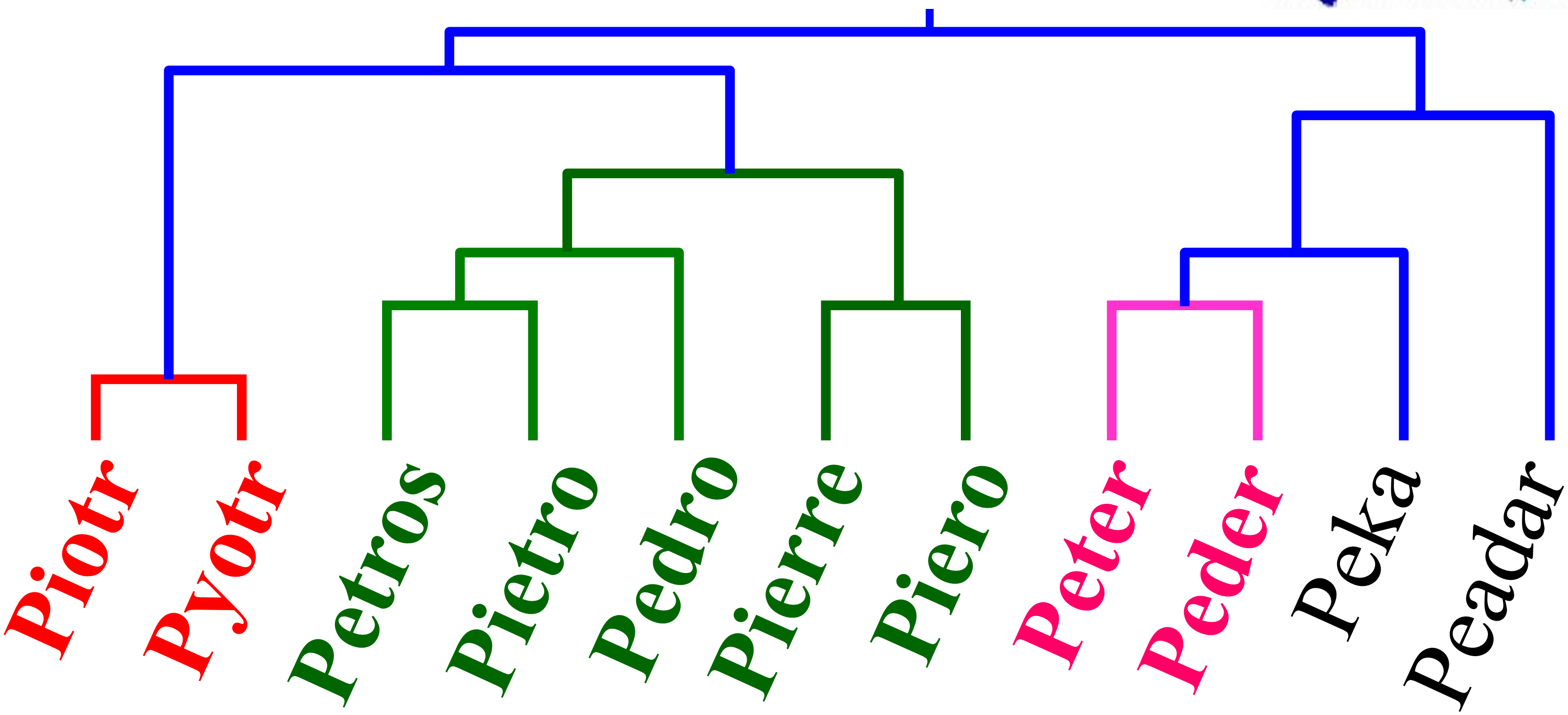
Michalis (Greek), Michael (English), Mick (Irish!)



Pedro (**Portuguese/Spanish**)

Petros (**Greek**), Peter (**English**), Piotr (**Polish**)
Peadar (Irish), Pierre (**French**), Peder (**Danish**)
Peka (Hawaiian), Pietro (**Italian**), Piero (**Italian Alternative**), Petr (Czech), Pyotr (**Russian**)

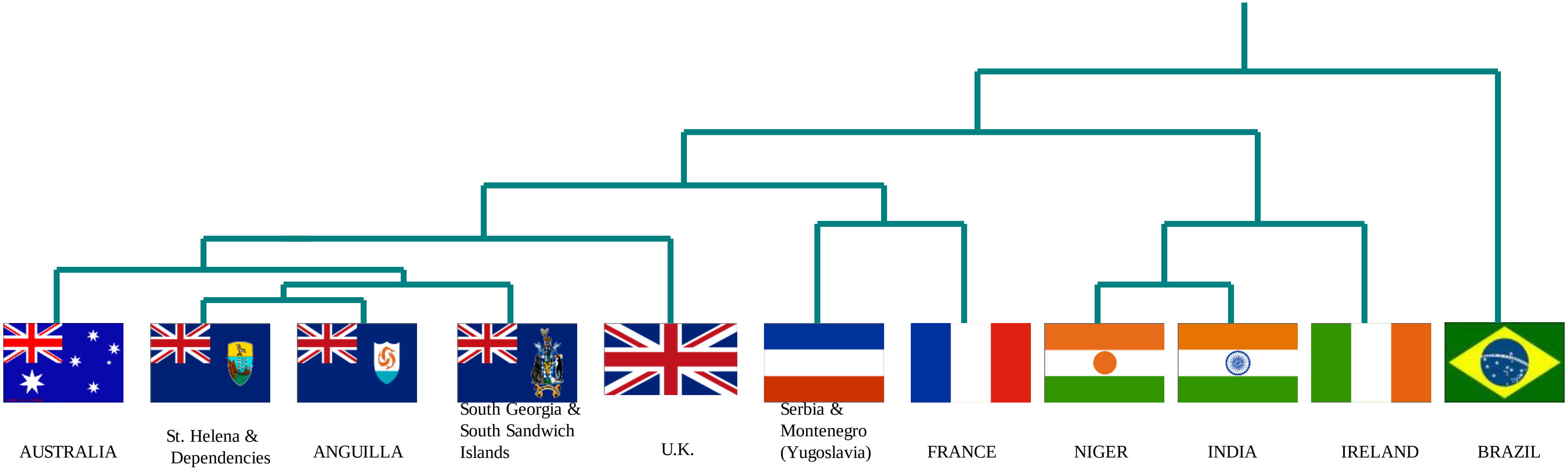




Hierarchal clustering can sometimes show patterns that are meaningless or spurious

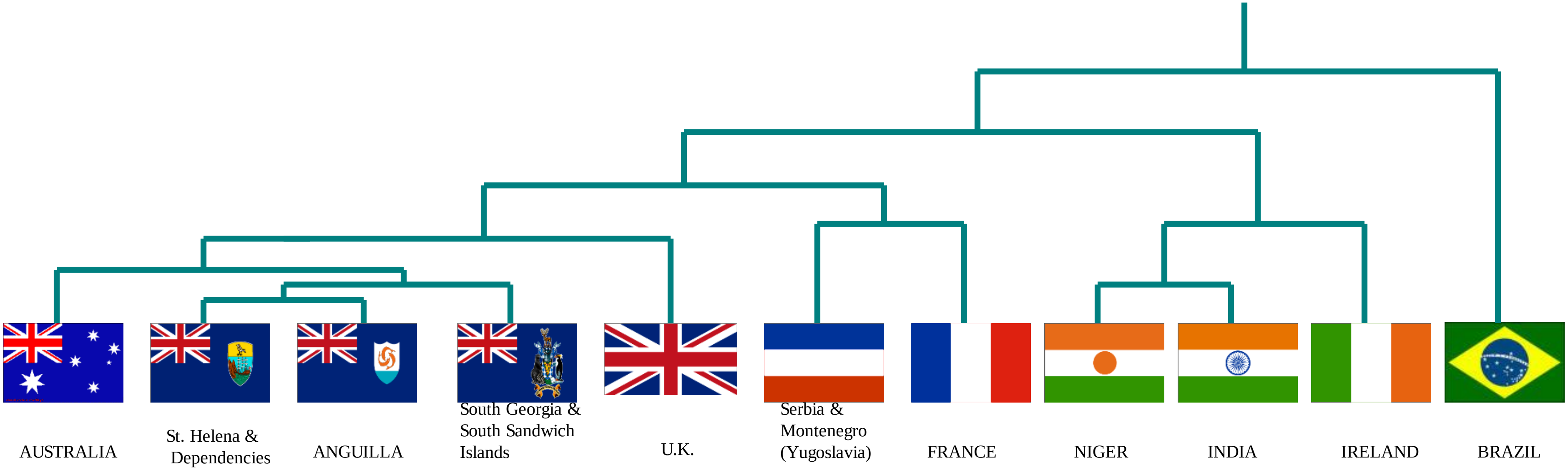
- For example, in this clustering, the tight grouping of Australia, Anguilla, St. Helena etc is meaningful, since all these countries are former UK colonies.

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

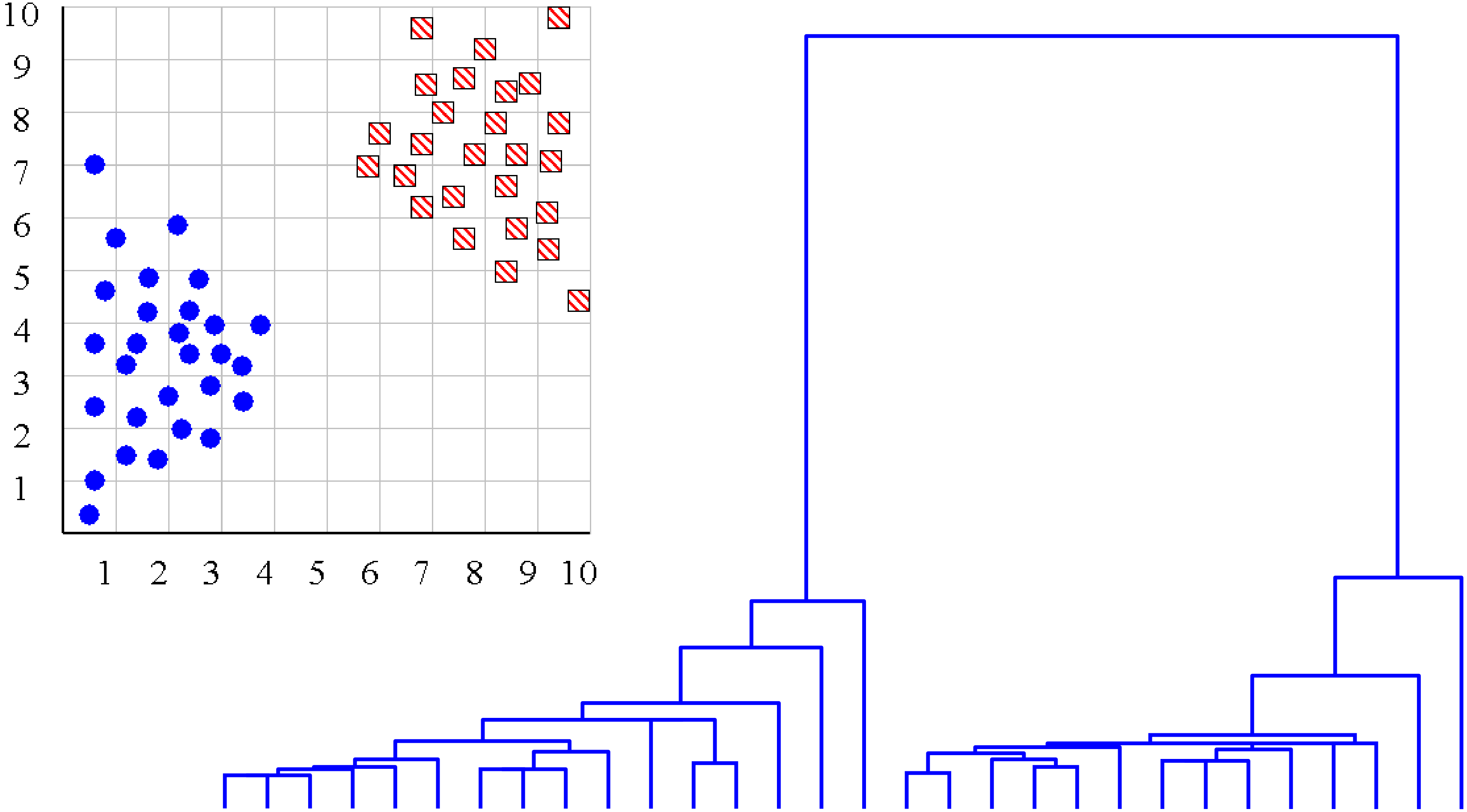


• The flag of Niger is orange over white over green, with an orange disc on the central white stripe, symbolizing the sun. The orange stands the Sahara desert, which borders Niger to the north. Green stands for the grassy plains of the south and west and for the River Niger which sustains them. It also stands for fraternity and hope. White generally symbolizes purity and hope.

• The Indian flag is a horizontal tricolor in equal proportion of deep saffron on the top, white in the middle and dark green at the bottom. In the center of the white band, there is a wheel in navy blue to indicate the Dharma Chakra, the wheel of law in the Sarnath Lion Capital. This center symbol or the 'CHAKRA' is a symbol dating back to 2nd century BC. The saffron stands for courage and sacrifice; the white, for purity and truth; the green for growth and auspiciousness.

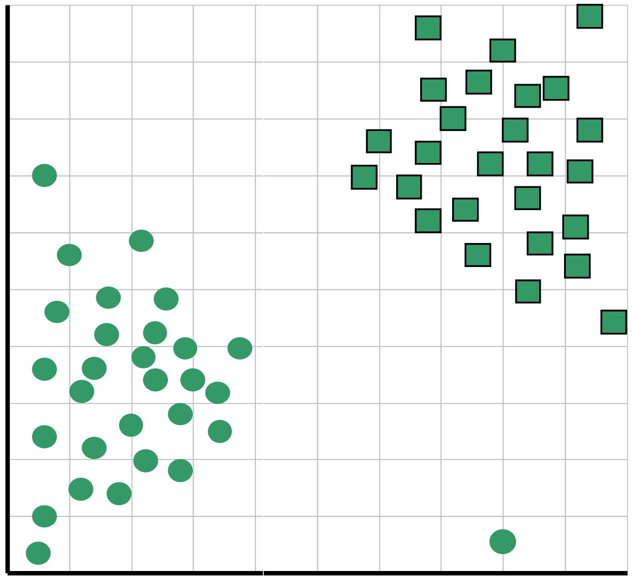


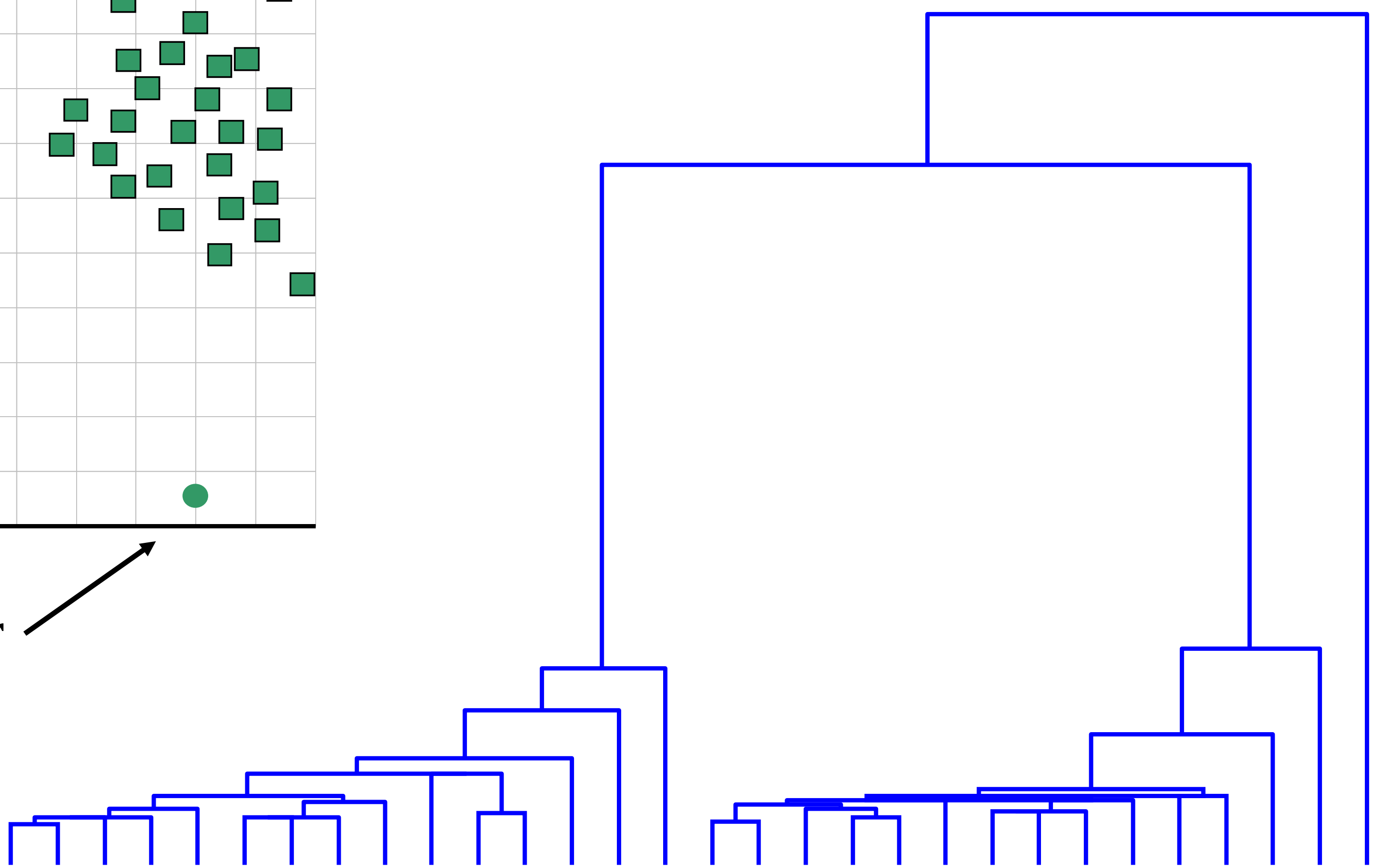
We can look at the dendrogram to determine the “correct” number of clusters. In this case, the two highly separated subtrees are highly suggestive of two clusters. (Things are rarely this clear cut, unfortunately)



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