

# **Artificial Intelligence**

## **Clustering**

**Dr. Uzma Jamil**

**Department of Computer Science**

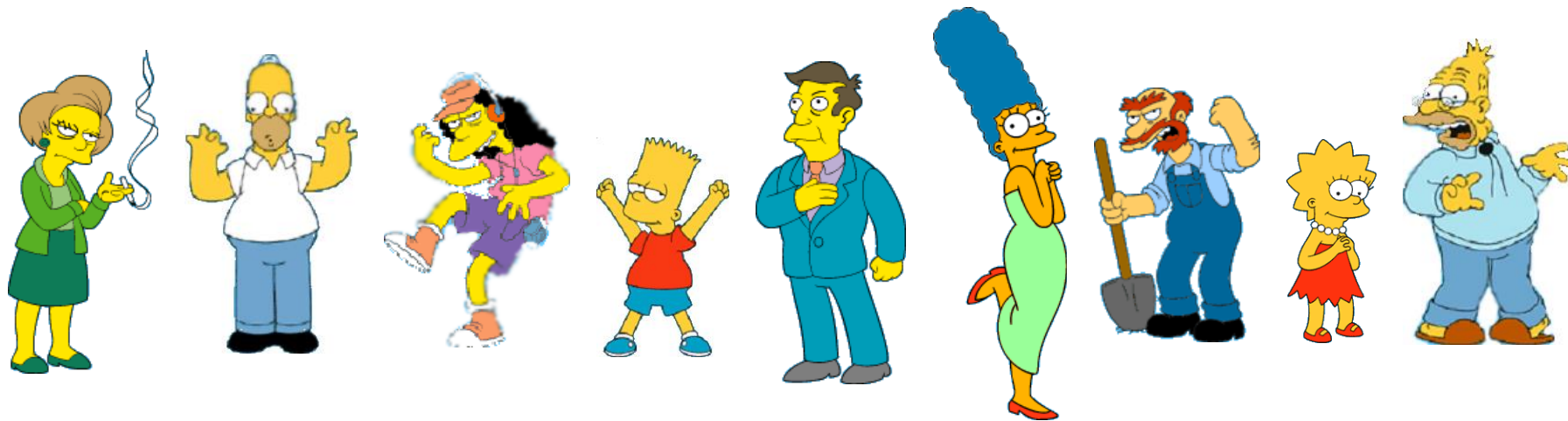
**Government College University, Faisalabad.**

# What is Clustering?

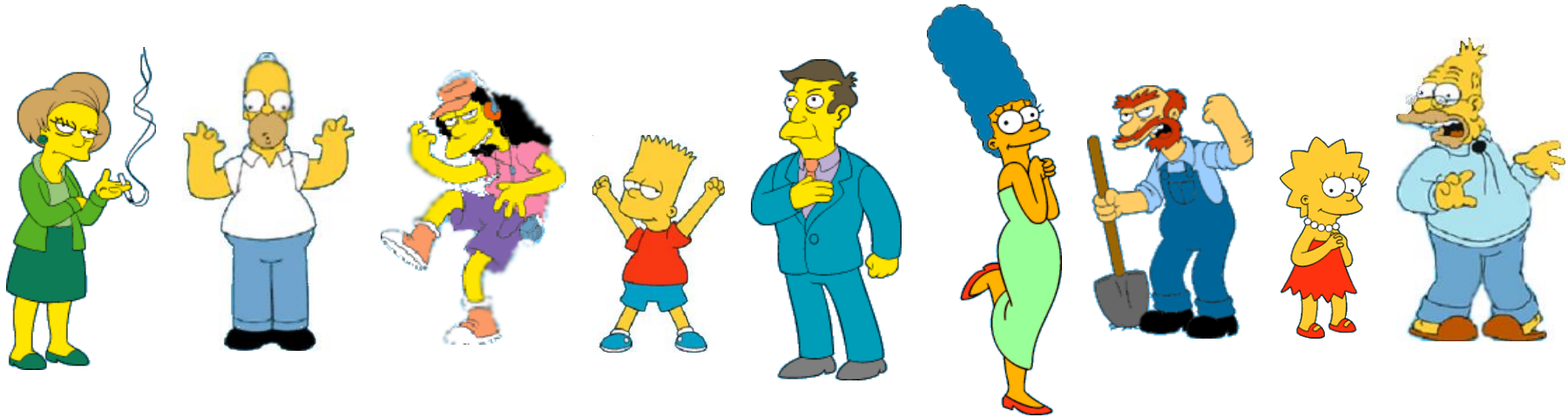
Also called *unsupervised learning*, sometimes called *classification* by statisticians and *sorting* by psychologists and *segmentation* by people in marketing

- 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.

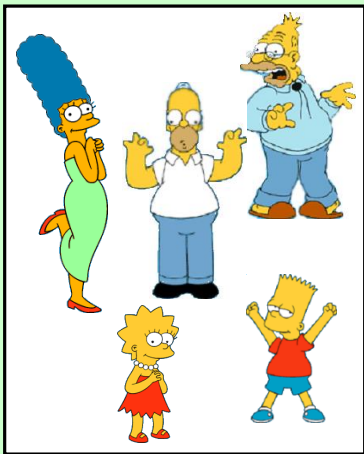
# What is a natural grouping among these objects?



# What is a natural grouping among these objects?



## Clustering is subjective



Simpson's Family



School Employees



Females



Males

# What is Similarity?

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

**Webster's Dictionary**



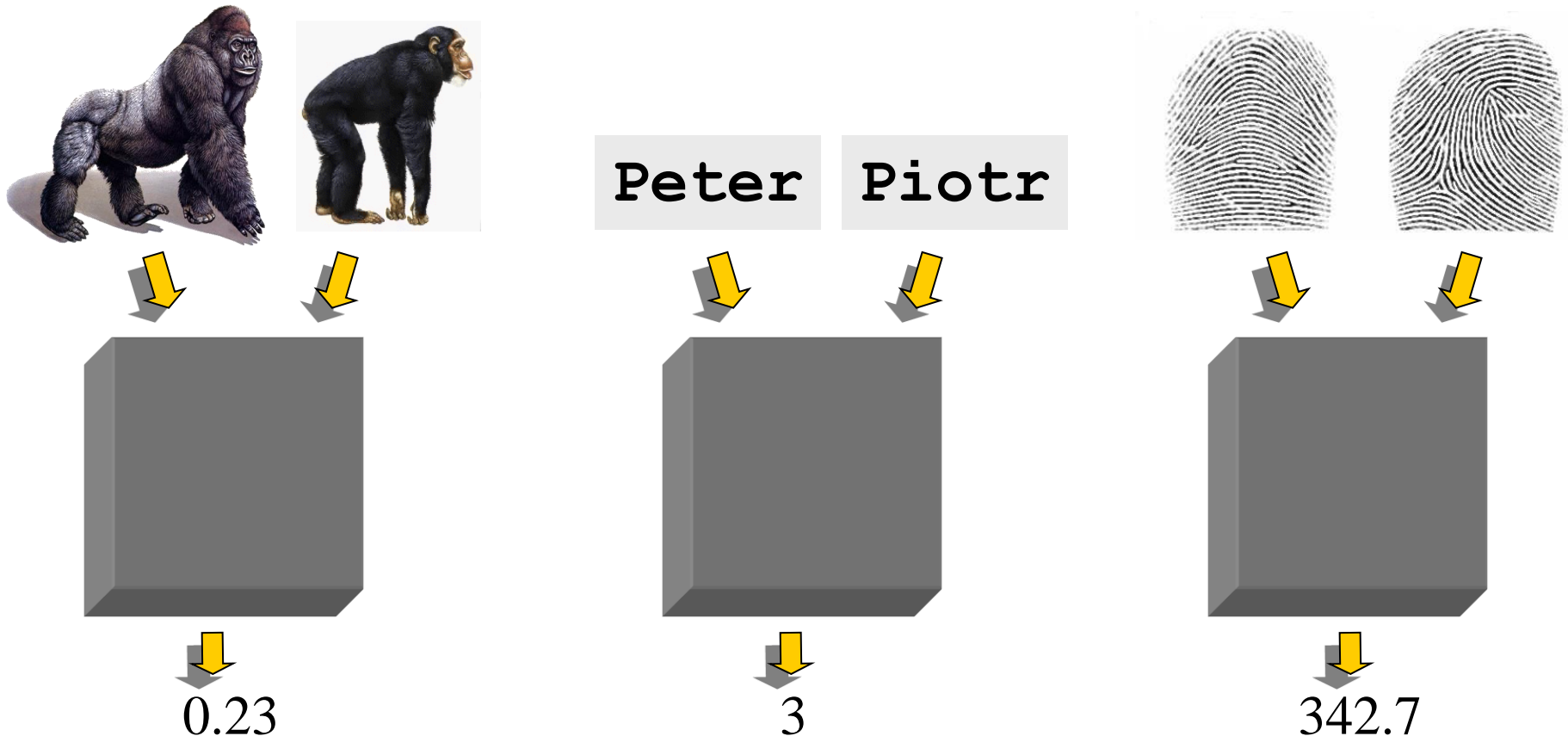
Similarity is hard to define, but...

*“We know it when we see it”*

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

# Defining Distance Measures

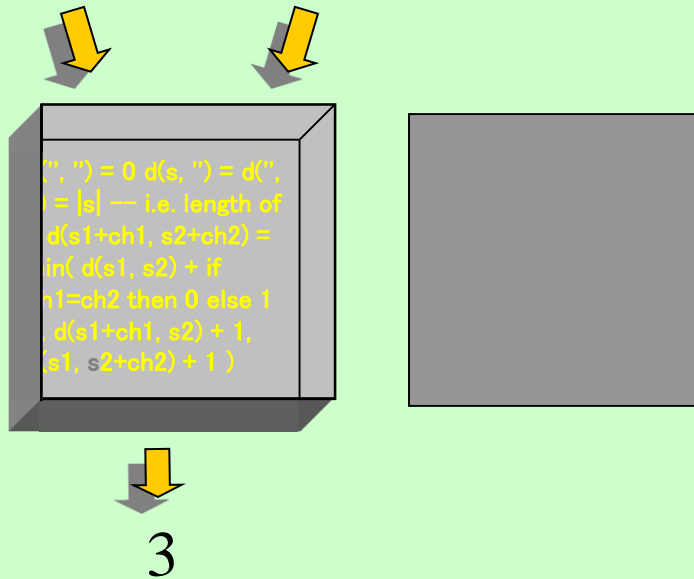
**Definition:** Let  $O_1$  and  $O_2$  be two objects from the universe of possible objects. The distance (dissimilarity) between  $O_1$  and  $O_2$  is a real number denoted by  $D(O_1, O_2)$





Peter

Piotr



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?

- $D(A,B) = D(B,A)$
- $D(A,A) = 0$
- $D(A,B) = 0$  IIf  $A = B$
- $D(A,B) \leq D(A,C) + D(B,C)$

*Symmetry*

*Constancy of Self-Similarity*

*Positivity (Separation)*

*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.”*

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

*Positivity (Separation)*

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

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

*Triangular Inequality*

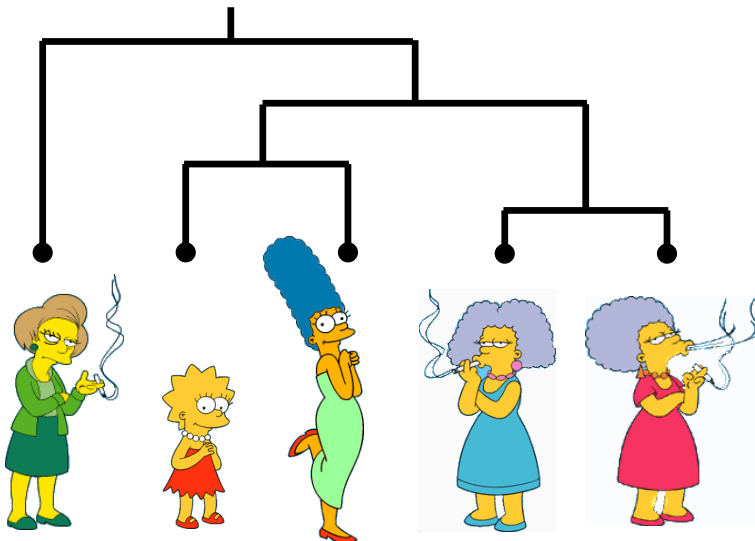
*Otherwise you could claim “Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl.”*



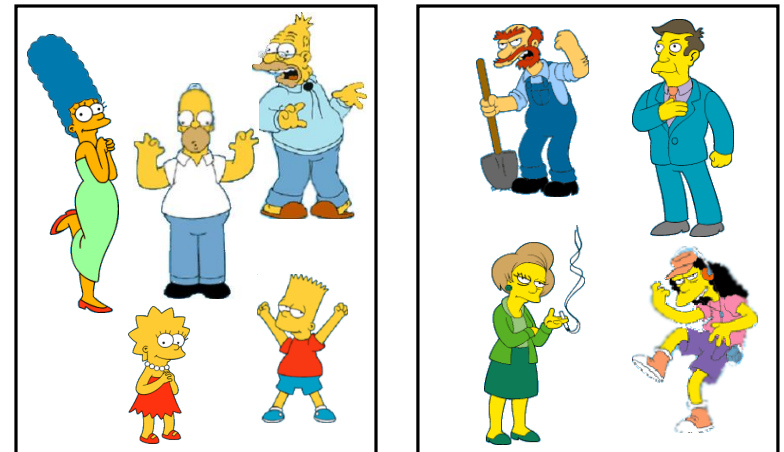
# Two Types of Clustering

- **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

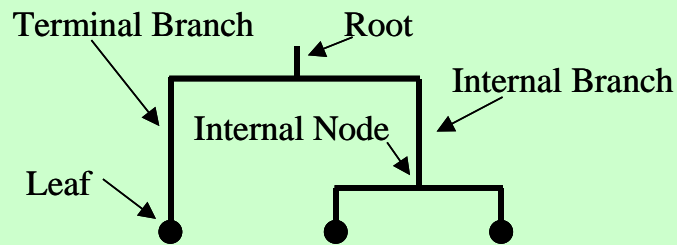


# Desirable Properties of a Clustering Algorithm

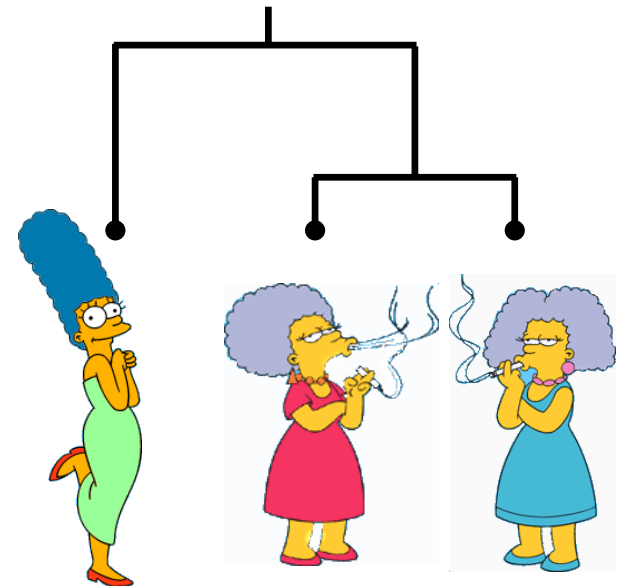
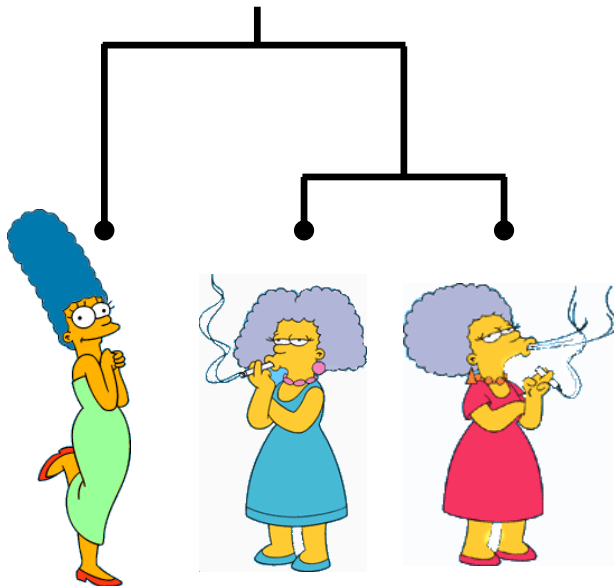
- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- Incorporation of user-specified constraints
- Interpretability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

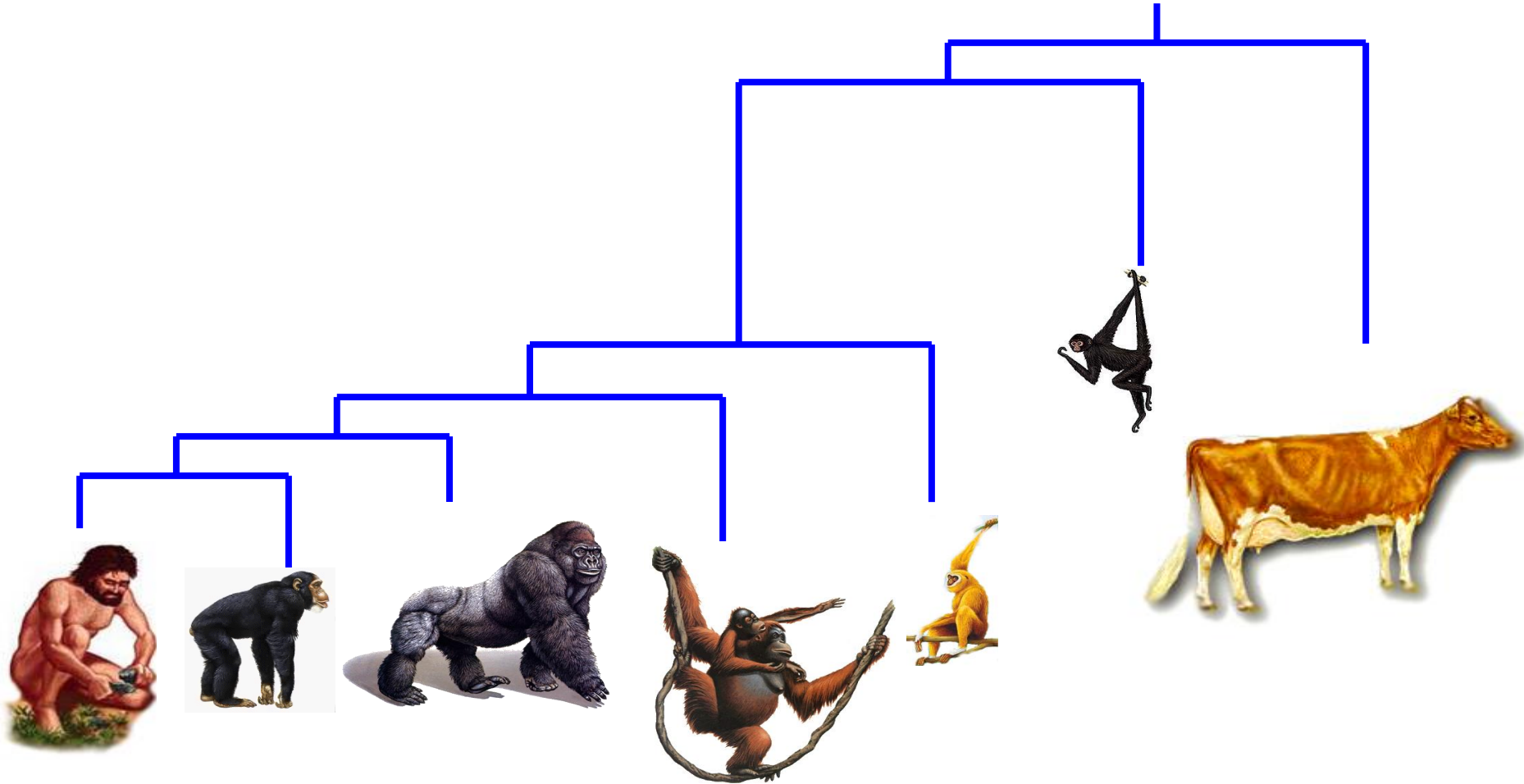
# A Useful Tool for Summarizing Similarity Measurements

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.





(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);

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.

Web Site Directory - Sites organized by subject

[Suggest your site](#)

**Business & Economy**

[B2B](#), [Finance](#), [Shopping](#), [Jobs](#)...

**Regional**

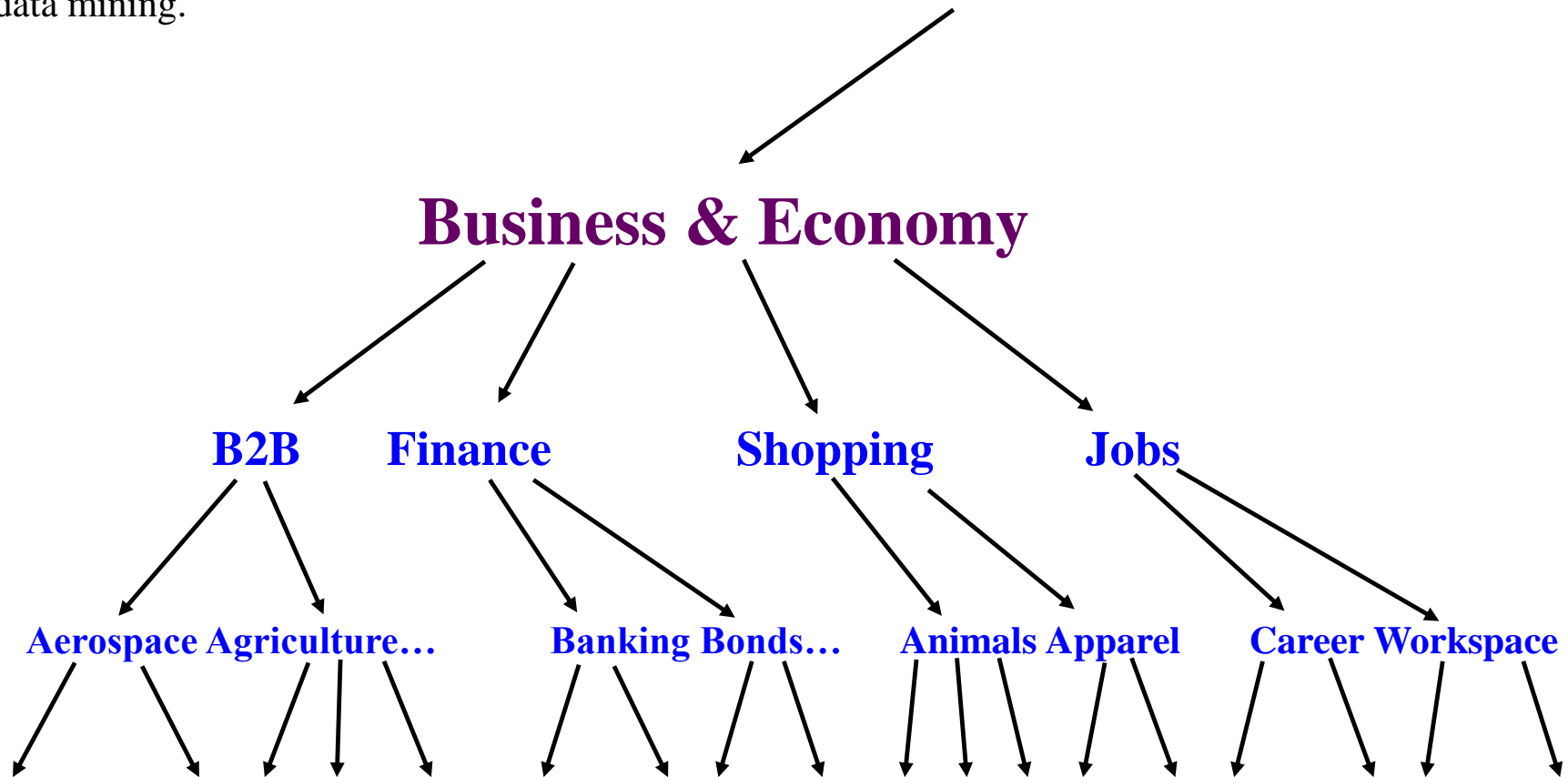
[Countries](#), [Regions](#), [US States](#)...

**Computers & Internet**

[Internet](#), [WWW](#), [Software](#), [Games](#)...

**Society & Culture**

[People](#), [Environment](#), [Religion](#)...



# A Demonstration of Hierarchical Clustering using String Edit Distance

## Pedro (Portuguese)

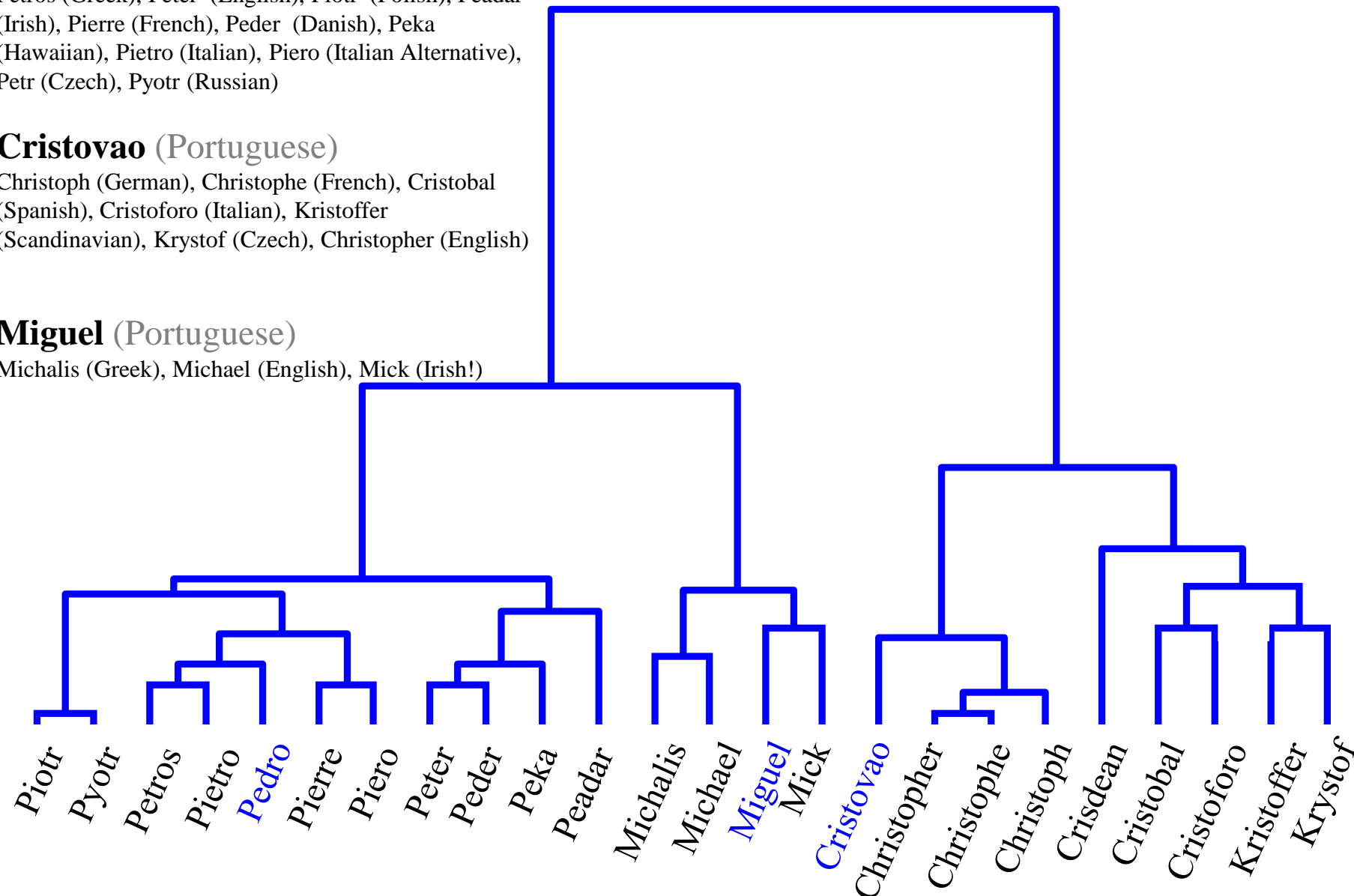
Petros (Greek), Peter (English), Piotr (Polish), Peadar (Irish), Pierre (French), Peder (Danish), Peka (Hawaiian), Pietro (Italian), Piero (Italian Alternative), Petr (Czech), Pyotr (Russian)

## Cristovao (Portuguese)

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

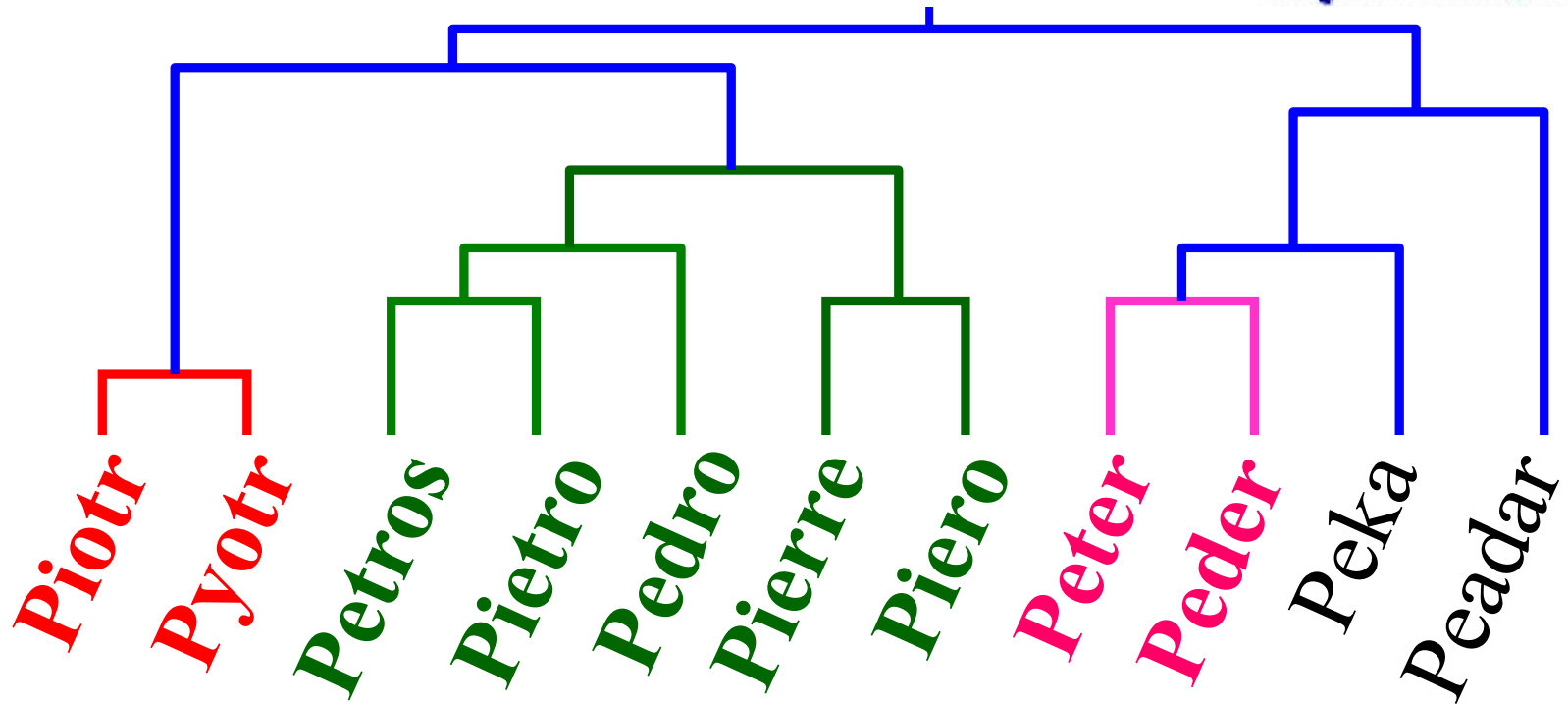
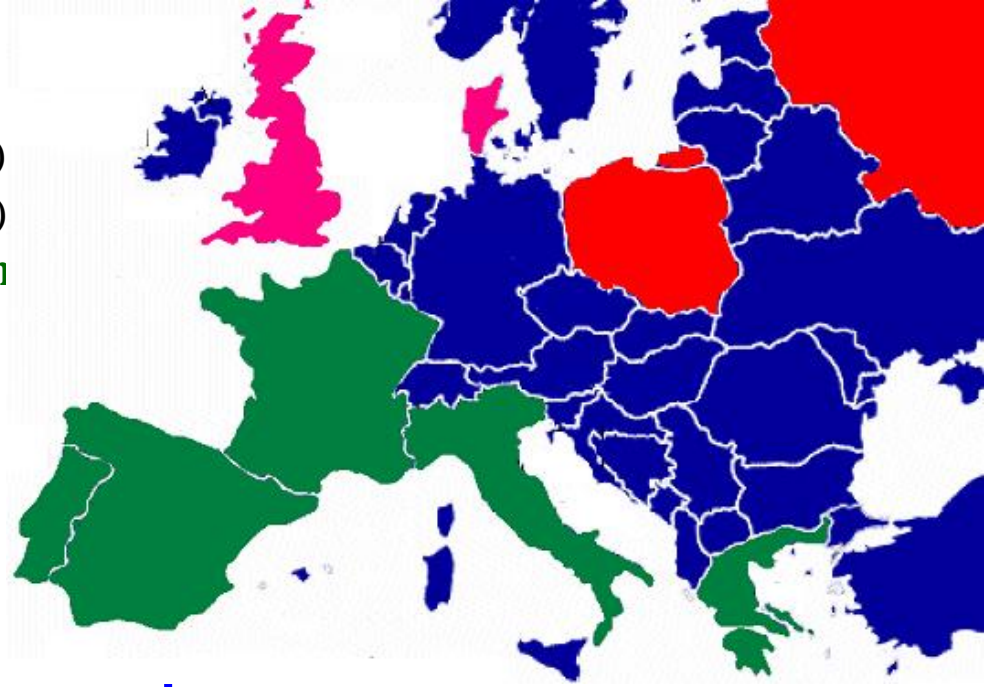
## Miguel (Portuguese)

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



# Pedro (Portuguese/Spanish)

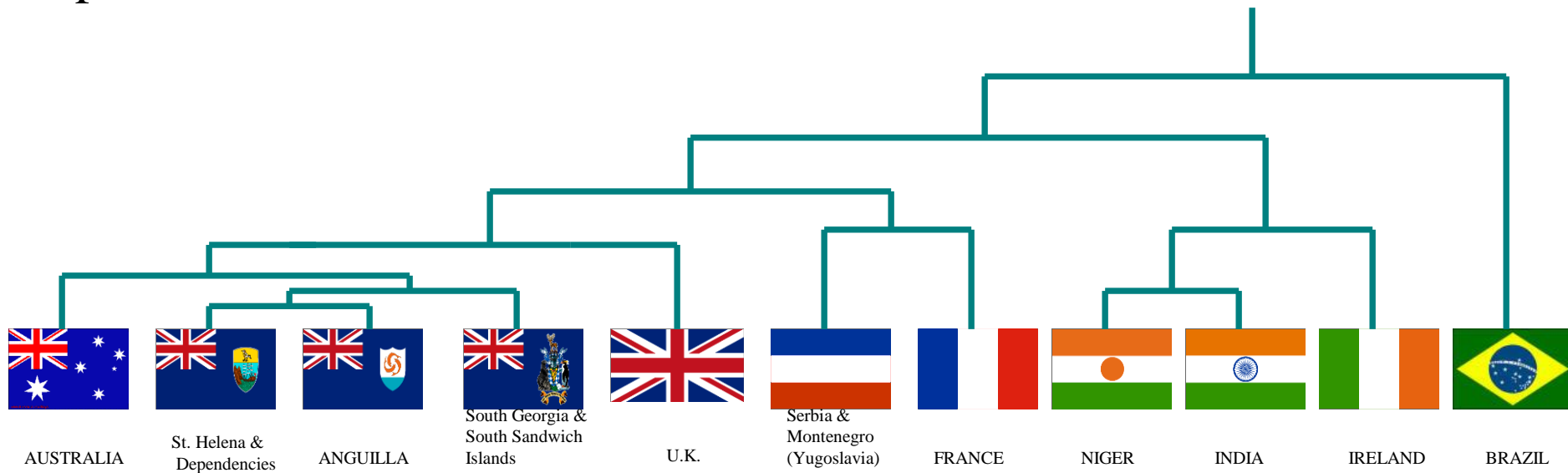
Petros (**Greek**), Peter (**English**), Piotr (**Polish**)  
Peadar (Irish), Pierre (**French**), Peder (**Danish**)  
Peka (Hawaiian), Pietro (**Italian**), Piero (**Italian Alternative**), Petr (Czech), Pyotr (**Russian**)



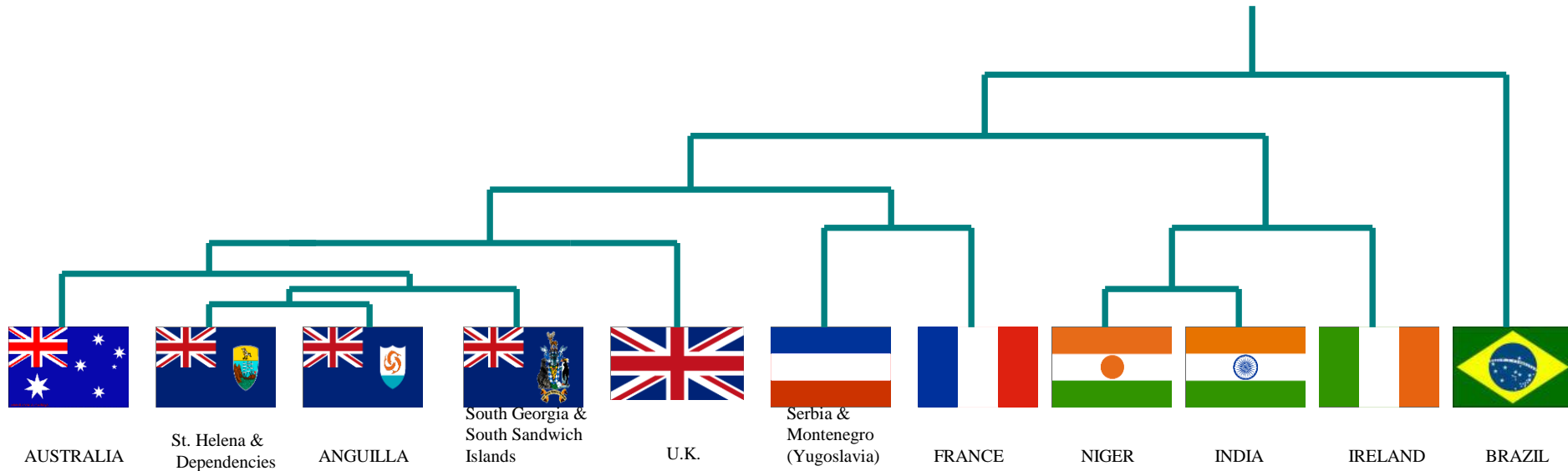


# Hierarchical clustering can sometimes show patterns that are meaningless or spurious

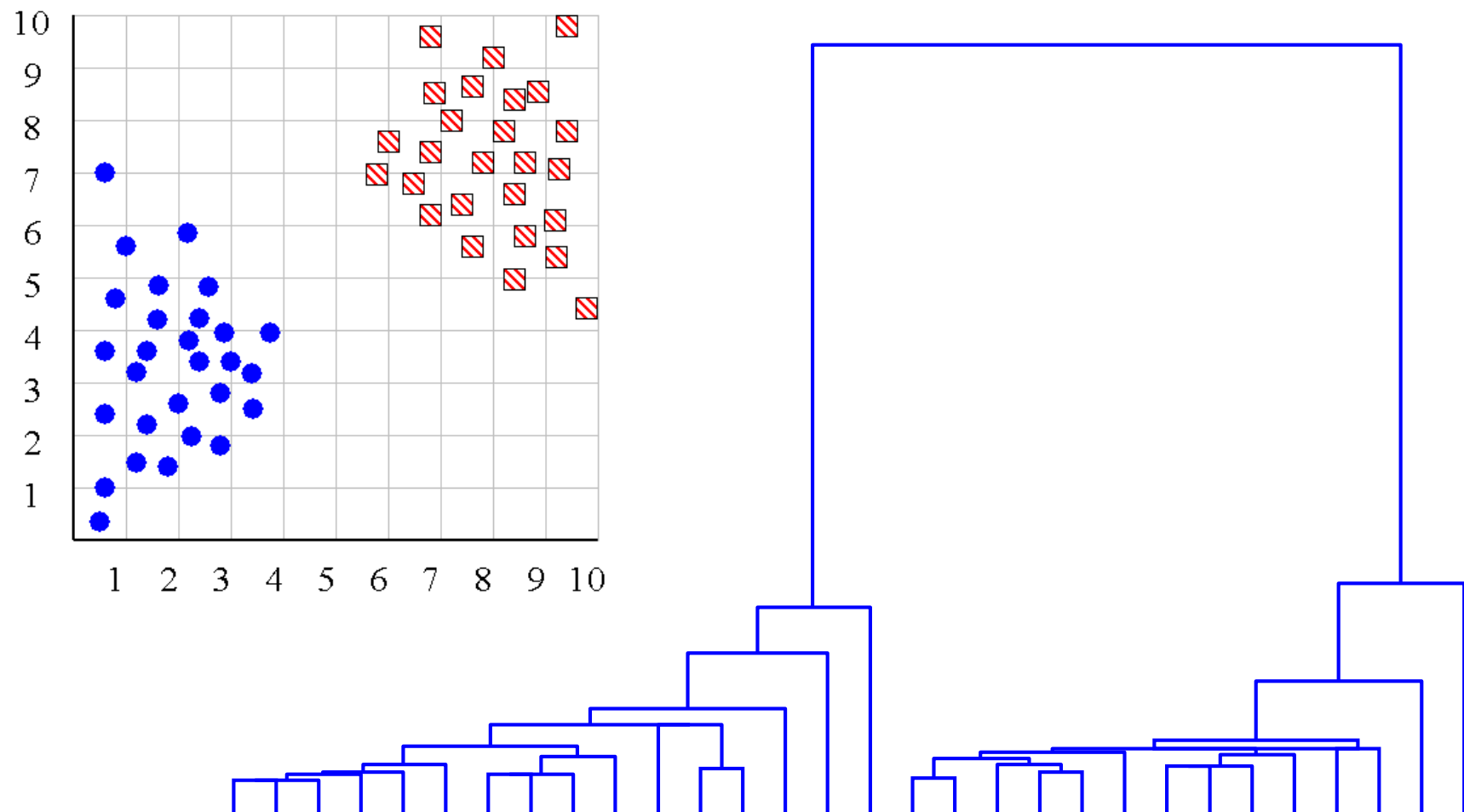
- For example, in this clustering, the tight grouping of Australia, Anguilla, St. Helena etc is meaningful, since all these countries are former UK colonies.
- However the tight grouping of Niger and India is completely spurious, there is no connection between the two.



- 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.

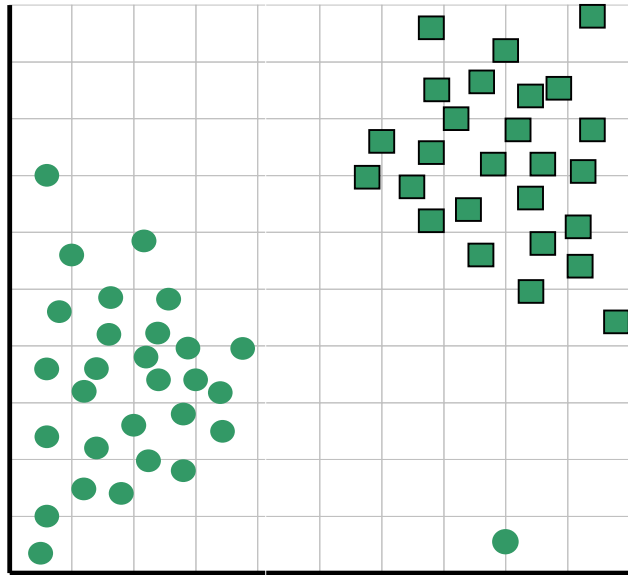


We can look at the dendrogram to determine the “correct” number of clusters. In this case, the two highly separated subtrees are highly suggestive of two clusters. (Things are rarely this clear cut, unfortunately)

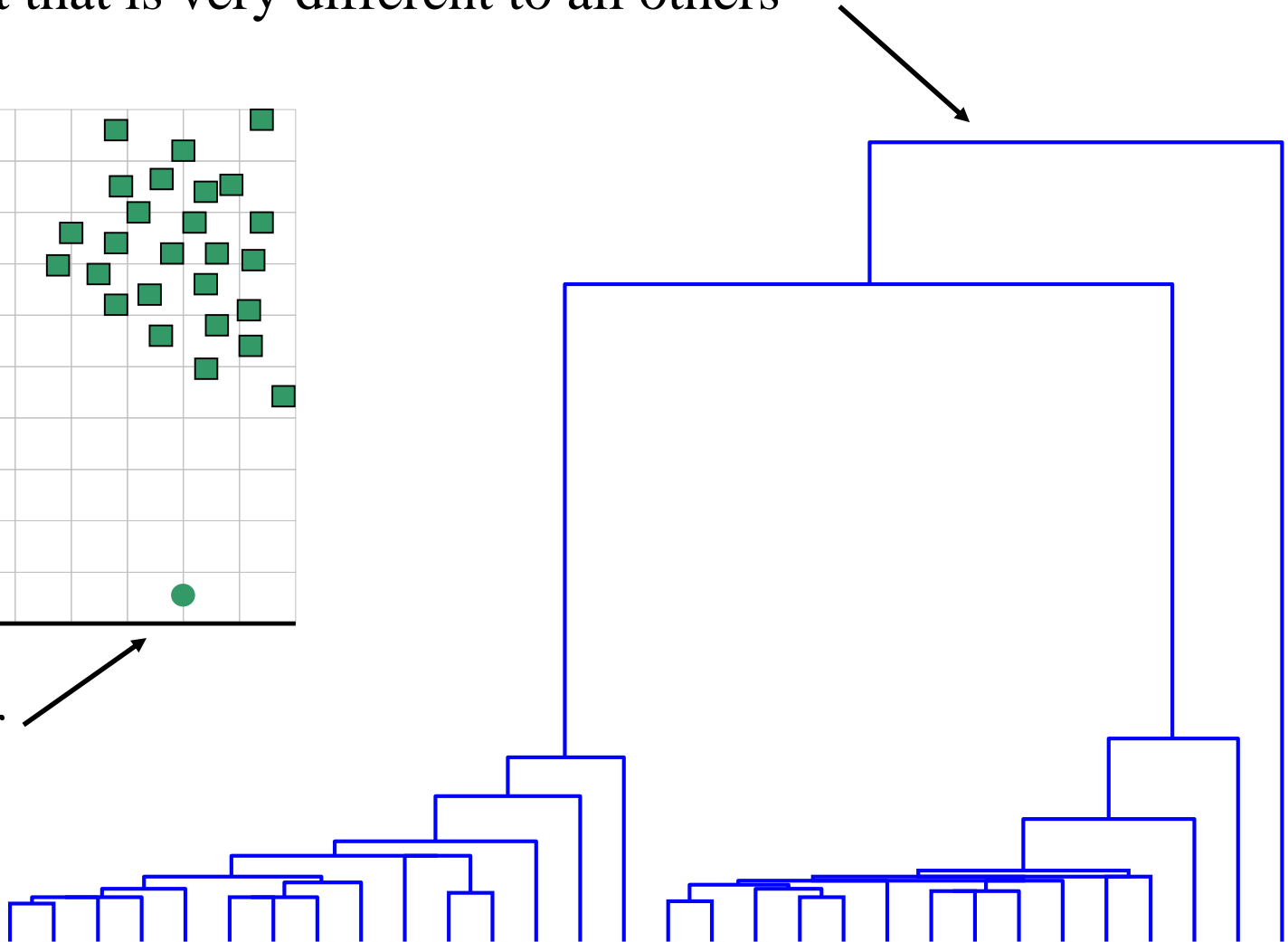


# One potential use of a dendrogram is to detect outliers

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



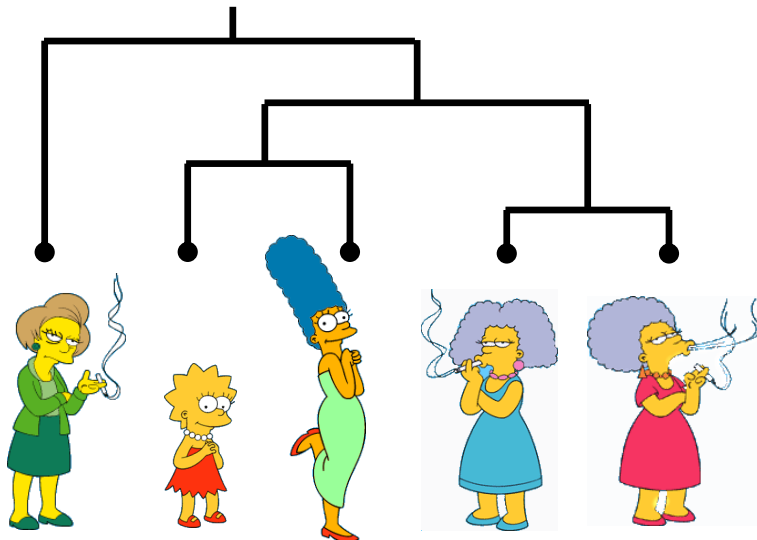
Outlier



# (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..

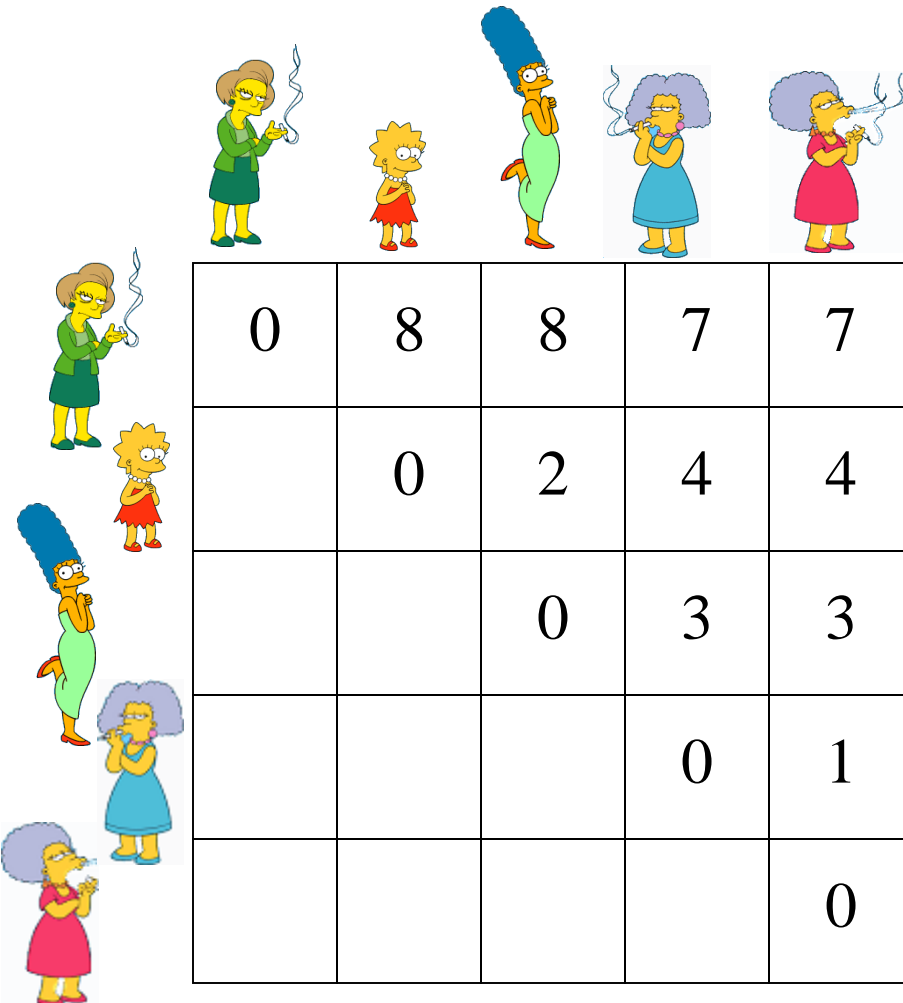
**Bottom-Up (agglomerative):** Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.











**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.


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


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

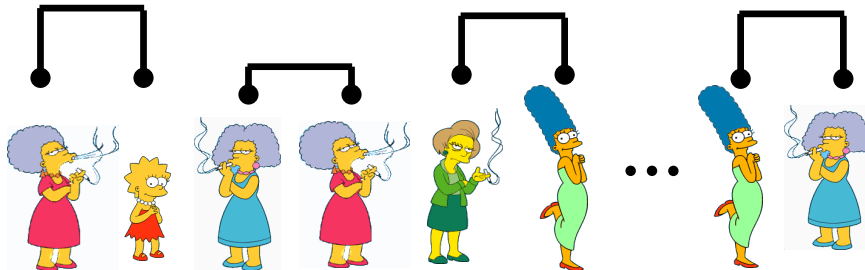


					
	0	8	8	7	7
		0	2	4	4
			0	3	3
				0	1
					0

# Bottom-Up (agglomerative):

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

Consider all possible merges...



Choose the best

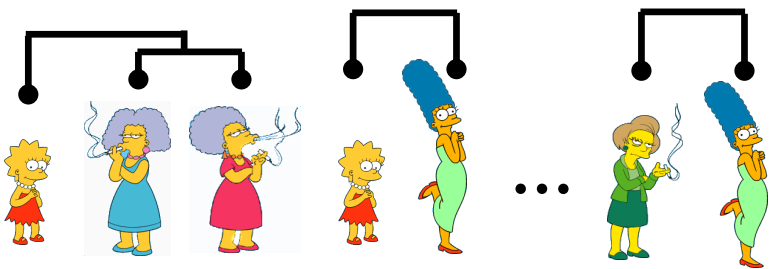




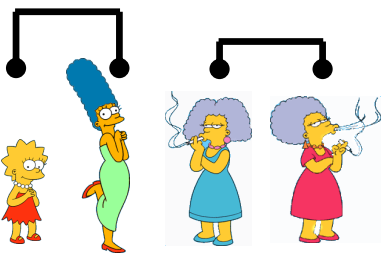
# Bottom-Up (agglomerative):

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

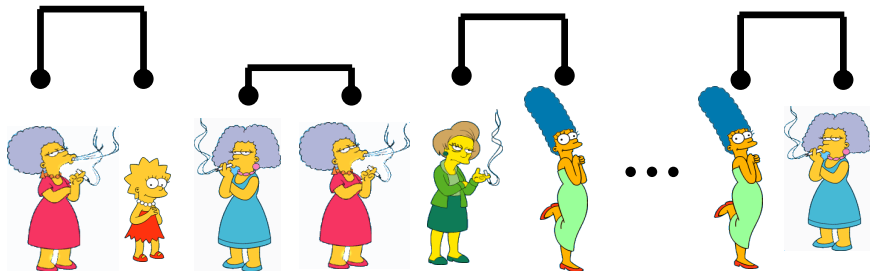
Consider all possible merges...



Choose the best



Consider all possible merges...



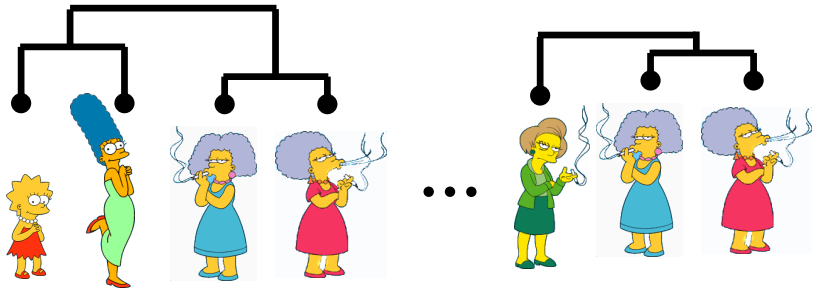
Choose the best



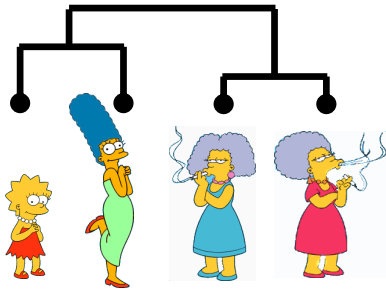
# Bottom-Up (agglomerative):

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.

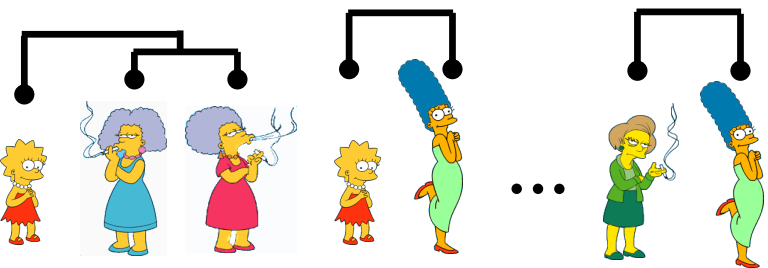
Consider all possible merges...



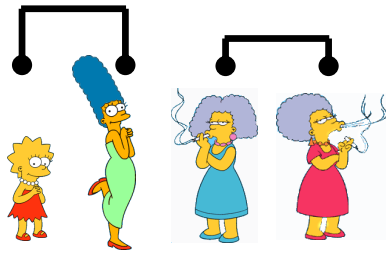
Choose the best



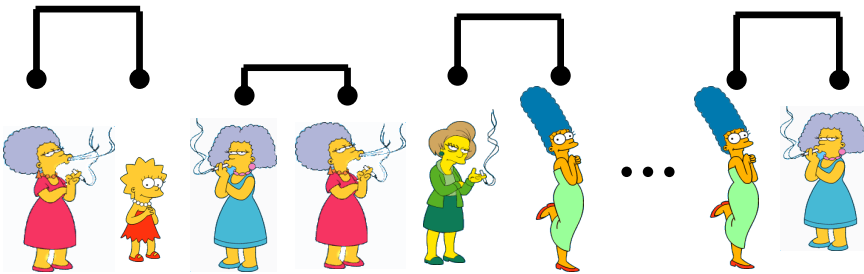
Consider all possible merges...



Choose the best



Consider all possible merges...

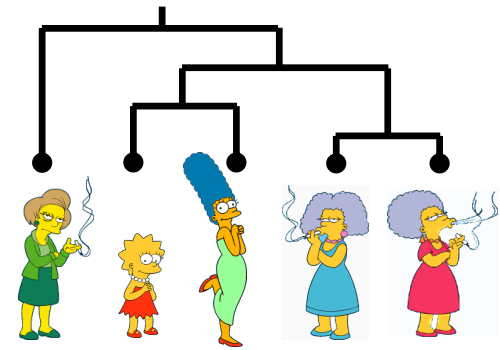


Choose the best

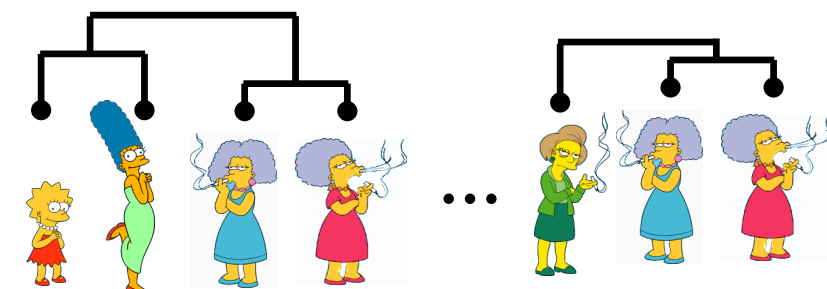


# Bottom-Up (agglomerative):

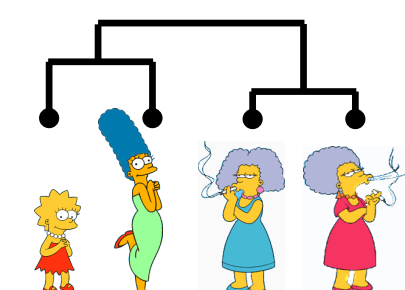
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



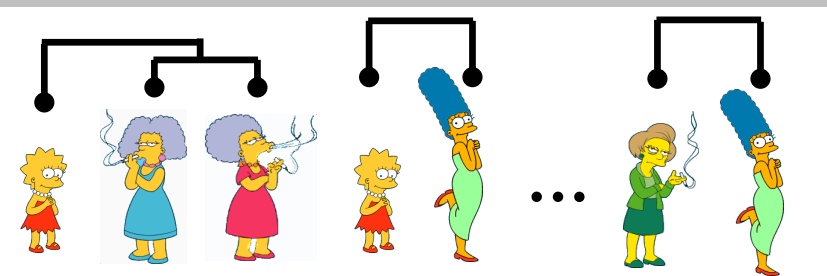
Consider all possible merges...



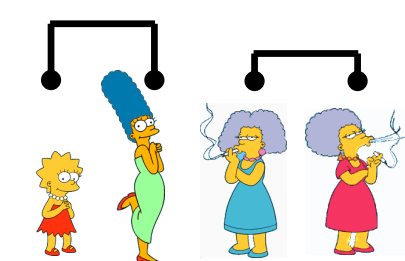
Choose the best



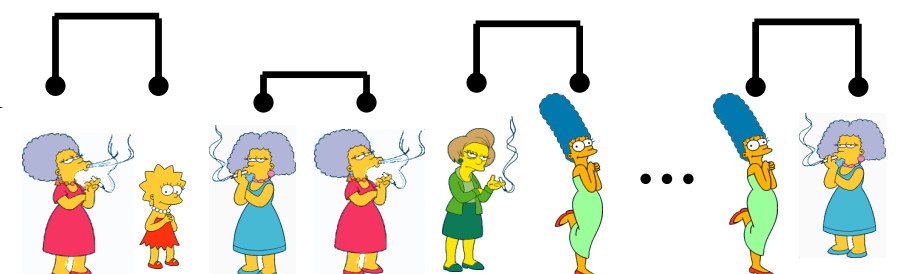
Consider all possible merges...



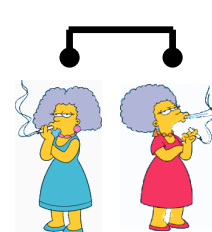
Choose the best



Consider all possible merges...

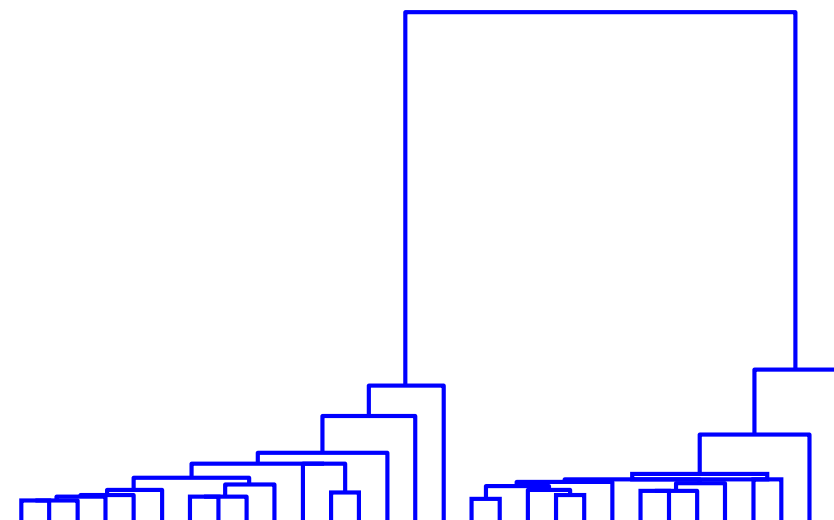
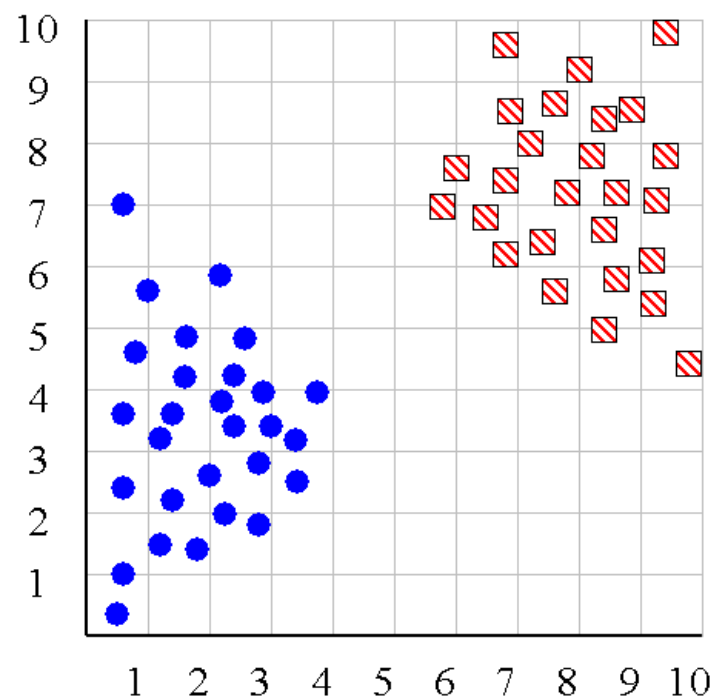


Choose the best

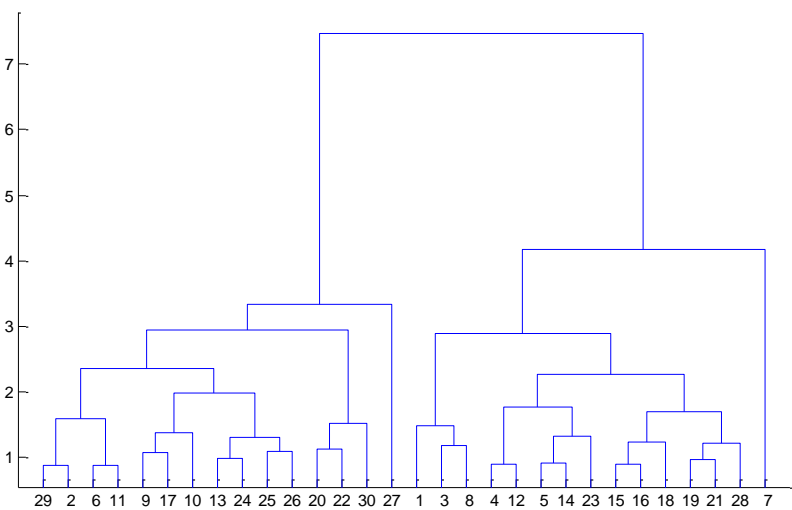


We know how to measure the distance between two objects, but defining the distance between an object and a cluster, or defining the distance between two clusters is non obvious.

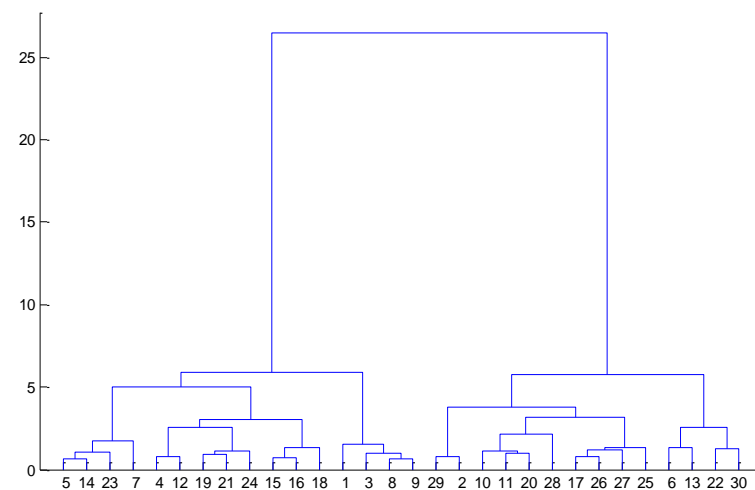
- **Single linkage (nearest neighbor):** In this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.
- **Complete linkage (furthest neighbor):** In this method, the distances between clusters are determined by 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



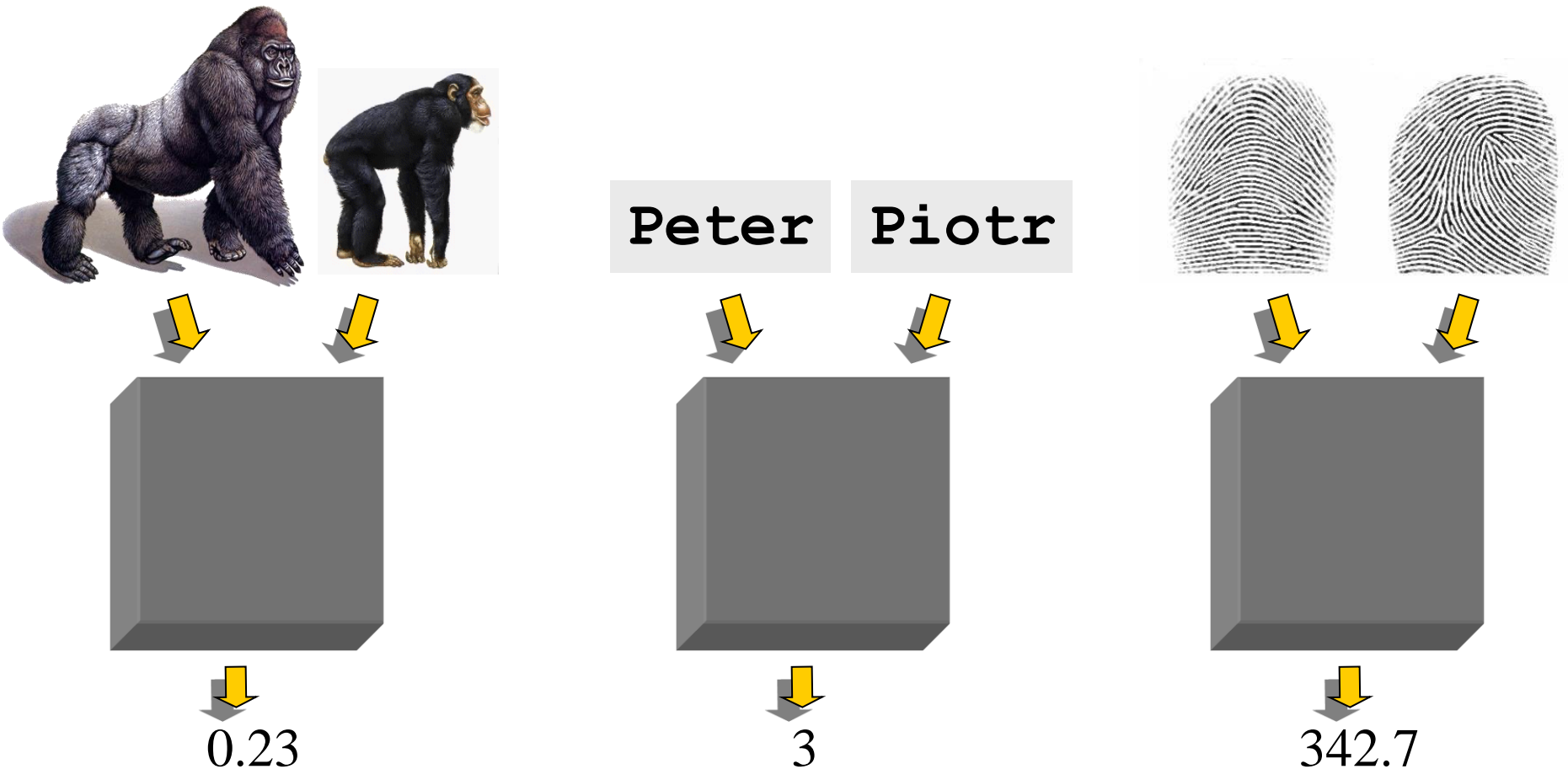
Wards linkage

# Summary of Hierarchical Clustering Methods

- 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.

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.

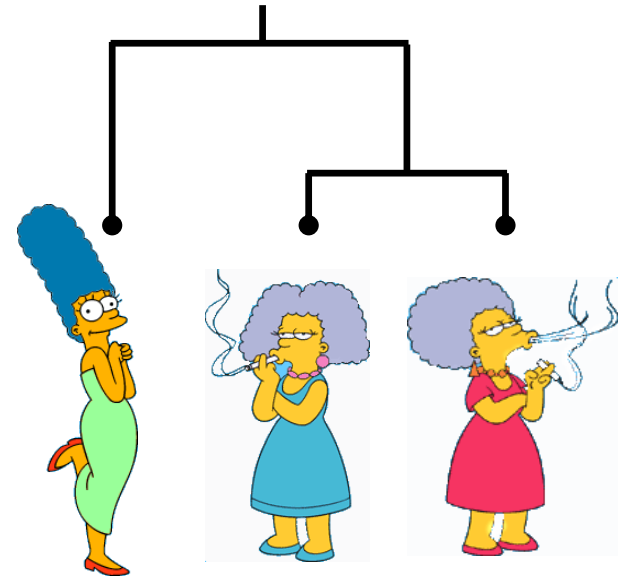
Change dress color,	1 point
Change earring shape,	1 point
Change hair part,	1 point

$D(\text{Patty}, \text{Selma}) = 3$

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

$D(\text{Marge}, \text{Selma}) = 5$



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

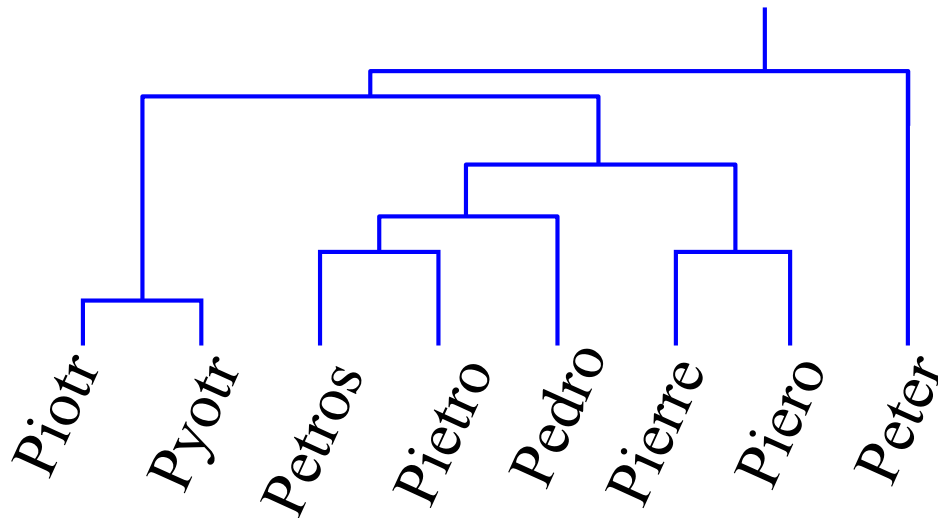
# Edit Distance Example

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

<i>Substitution</i>	1 Unit
<i>Insertion</i>	1 Unit
<i>Deletion</i>	1 Unit

$D(\text{Peter}, \text{Piotr})$  is 3

**Peter**



Substitution (i for e)

**Piter**



Insertion (o)

**Pioter**



Deletion (e)

**Piotr**