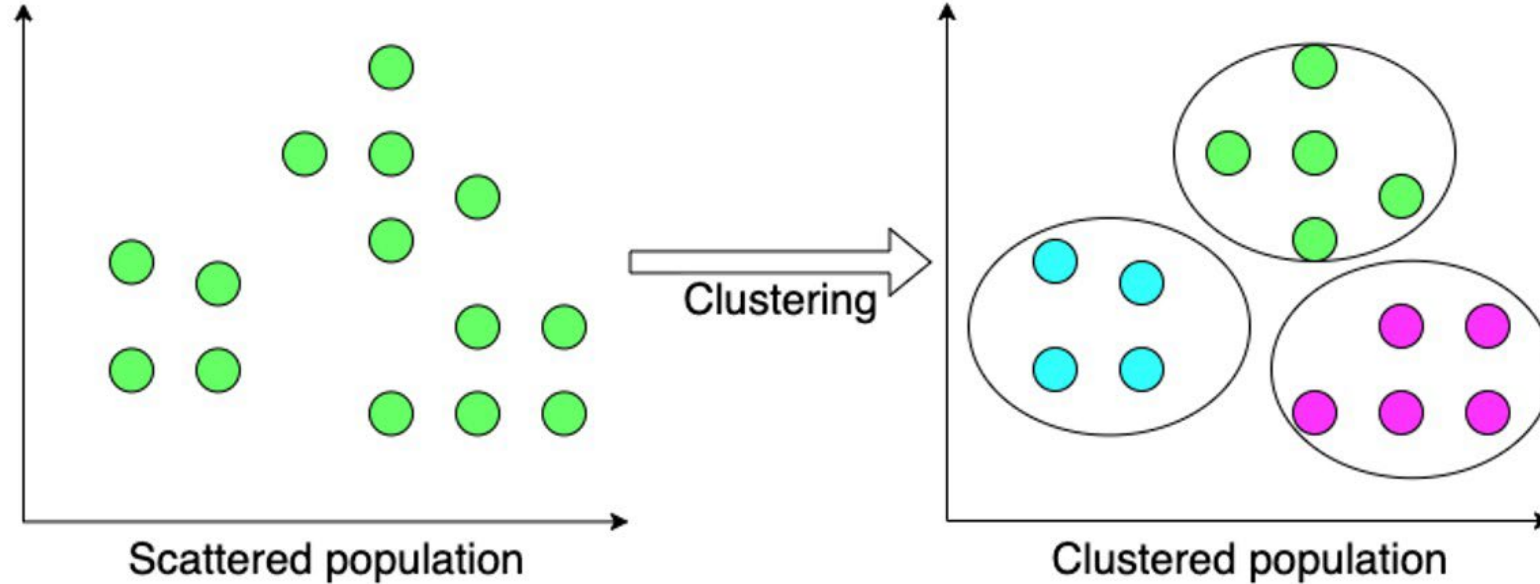


Unsupervised Learning - Clustering



Aprendizaje no Supervisado y Clustering Jerárquico

PhD(e). Jonnatan Arias Garcia – jonnatan.arias@utp.edu.co –
jariasg@uniquindio.edu.co

PhD. David Cardenas peña - dcardenasp@utp.edu.co

PhD. Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Clustering

(Unsupervised Learning)

Dado: unas muestras $\langle X_1, X_2, X_3, \dots, X_n \rangle$

Encontrar el agrupamiento natural de los datos.

No tenemos etiquetas (Y)

Ejemplo de aplicación:

Identificar perfiles de clientes de uso de energía similares

$\langle x \rangle$ = serie temporal de uso de energía

Identificar anomalías en el comportamiento de los usuarios para la seguridad informática

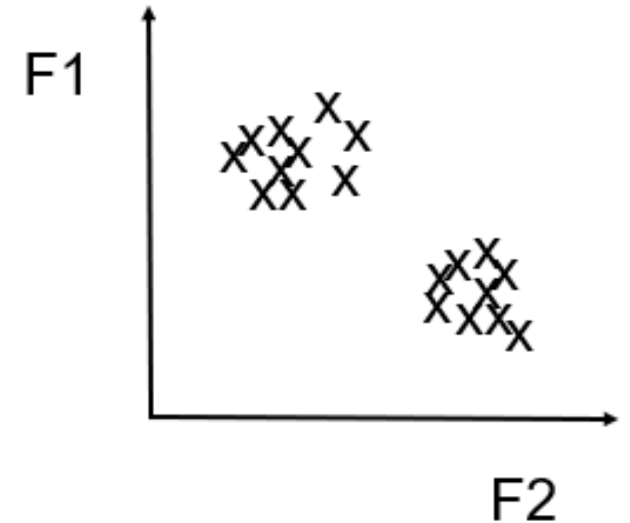
$\langle x \rangle$ = secuencias de comandos de usuario

Porque cluster?

- El etiquetado suele ser una tarea costosa
- Ganamos información estructural de los datos
- Encontramos posibles características poco comunes de los datos.

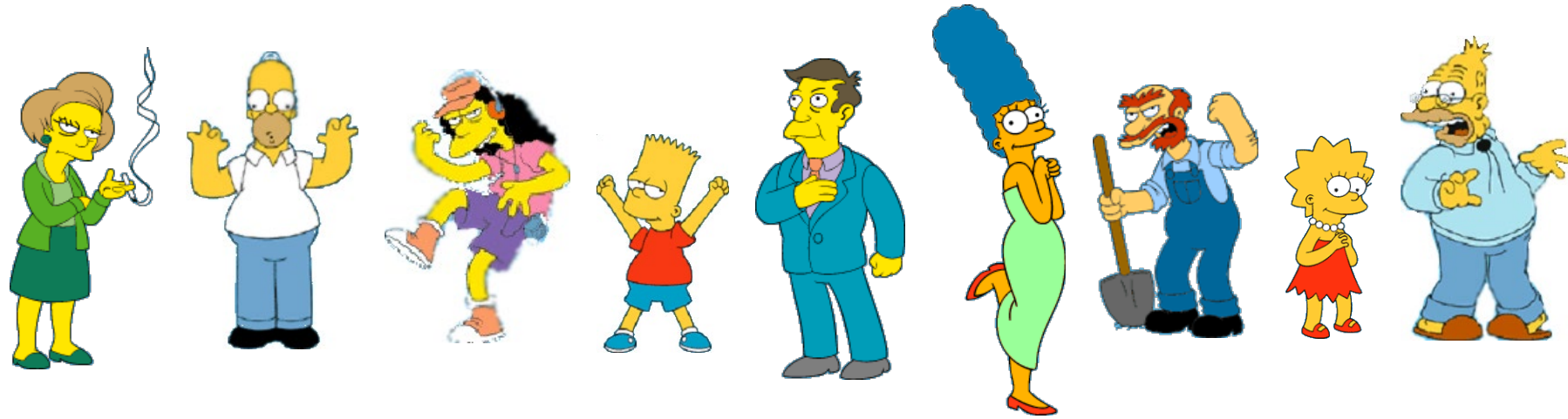
Objetivo del Cluster

- Dado un conjunto de puntos, cada uno descrito por unos atributos, debemos encontrar un cluster tal que:
 - Maximicemos la similitud entre inter-cluster (misma clase)
 - Minimicemos la similitud intra-clues (diferentes clases)
- Es necesario definir una medida de similitud



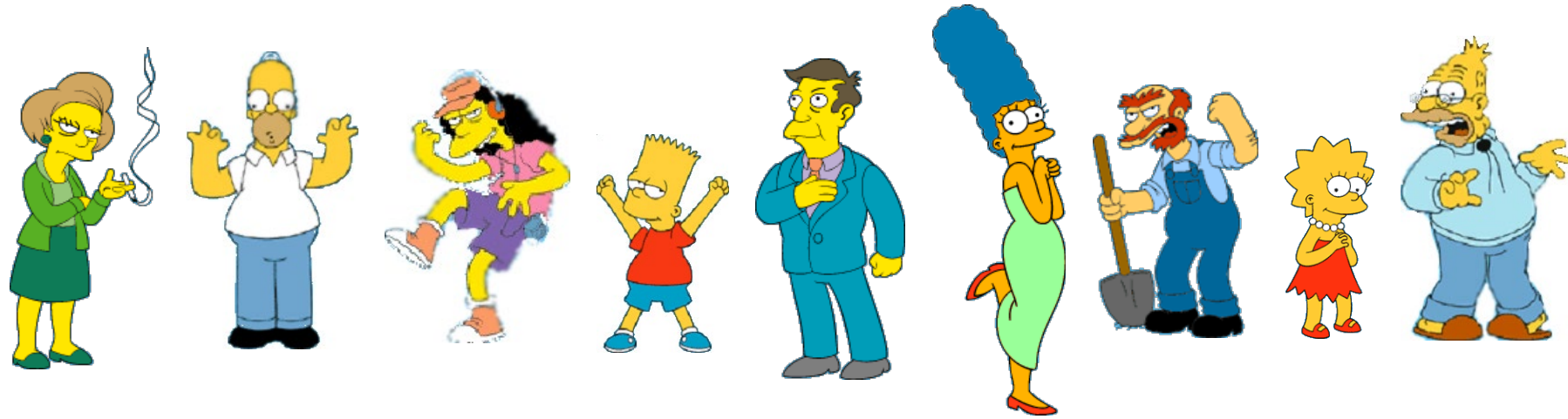
Cual es el agrupamiento natural de estos personajes?

Slide from Eamonn Keogh

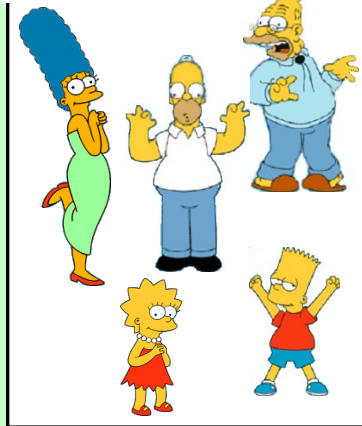


Cual es el agrupamiento natural de estos personajes?

Slide from Eamonn Keogh



Clustering es subjective



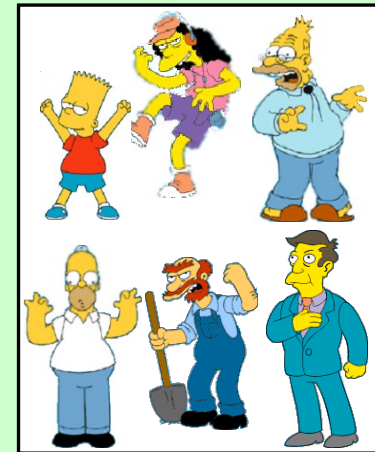
Simpson's Family



School Employees



Females



Males

Que es similitud?

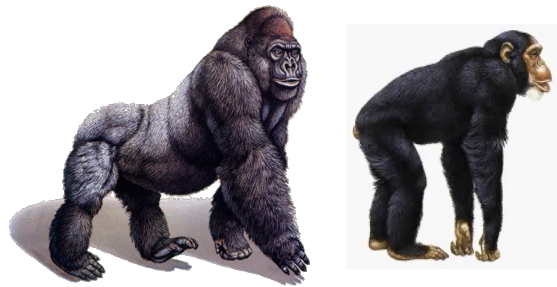
Slide based on one by Eamonn Keogh



Similitud es
difícil de definir,
pero,...
*"We know it
when we see it"*

Definiendo Medida de Distancia

Dados O_1 y O_2 dos objetos del universe de posibles objetos.
La distancia (similitud) entre O_1 y O_2 es un numero real
denotado $D(O_1, O_2)$



0.23

Peter

Piotr



3



342.7

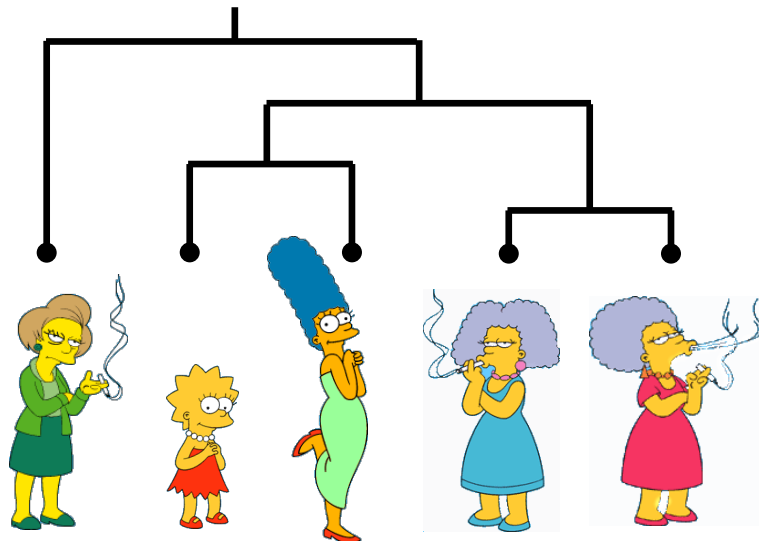
Que propiedades tiene la medida de distancia?

- $D(A,B) = D(B,A)$ *Simetría*
- $D(A,A) = 0$ *Constancia de autosimilitud*
- $D(A,B) = 0$ if $A = B$ *Positividad (Separación)*
- $D(A,B) \leq D(A,C) + D(B,C)$ *inequalidad Triangular*

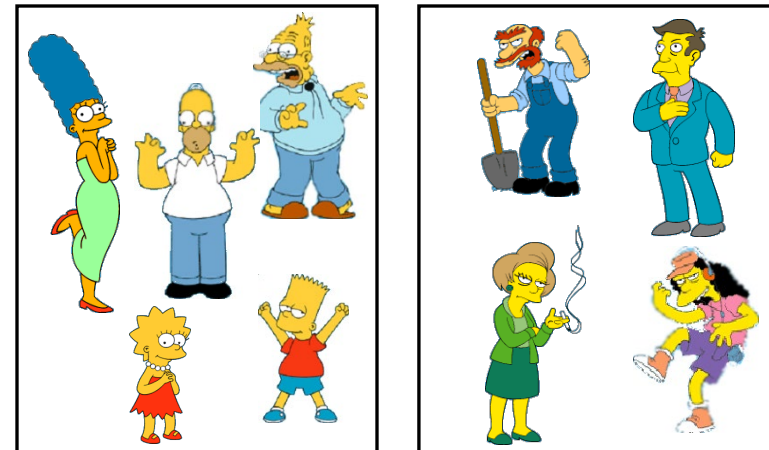
Dos Tipos de Clustering

- **Algoritmos Particionales:** Construye varias particiones y luego evalúa según algún criterio.
- **Algoritmos Jerárquicos:** Crea una descomposición jerárquica del conjunto de objetos usando algún criterio.

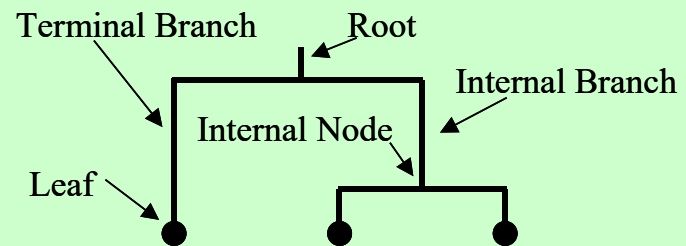
Hierarchical



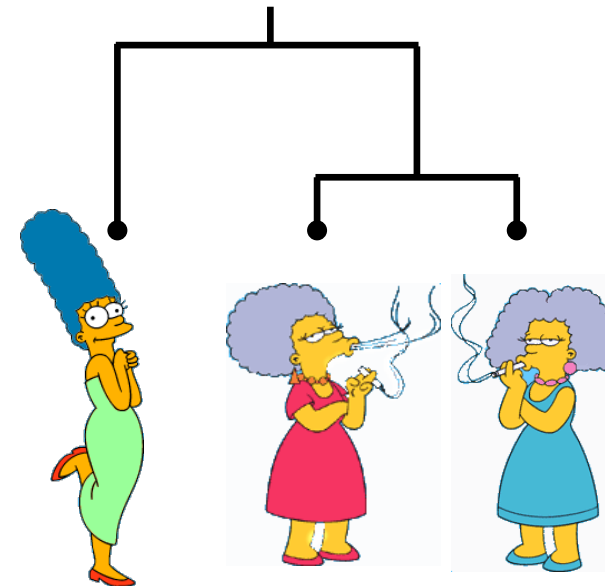
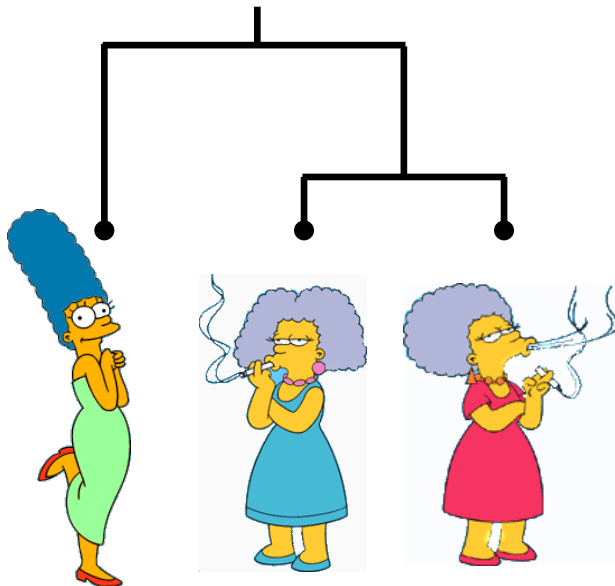
Partitional



1. Dendrogram: Una herramienta para sumarizar medidas de similitud



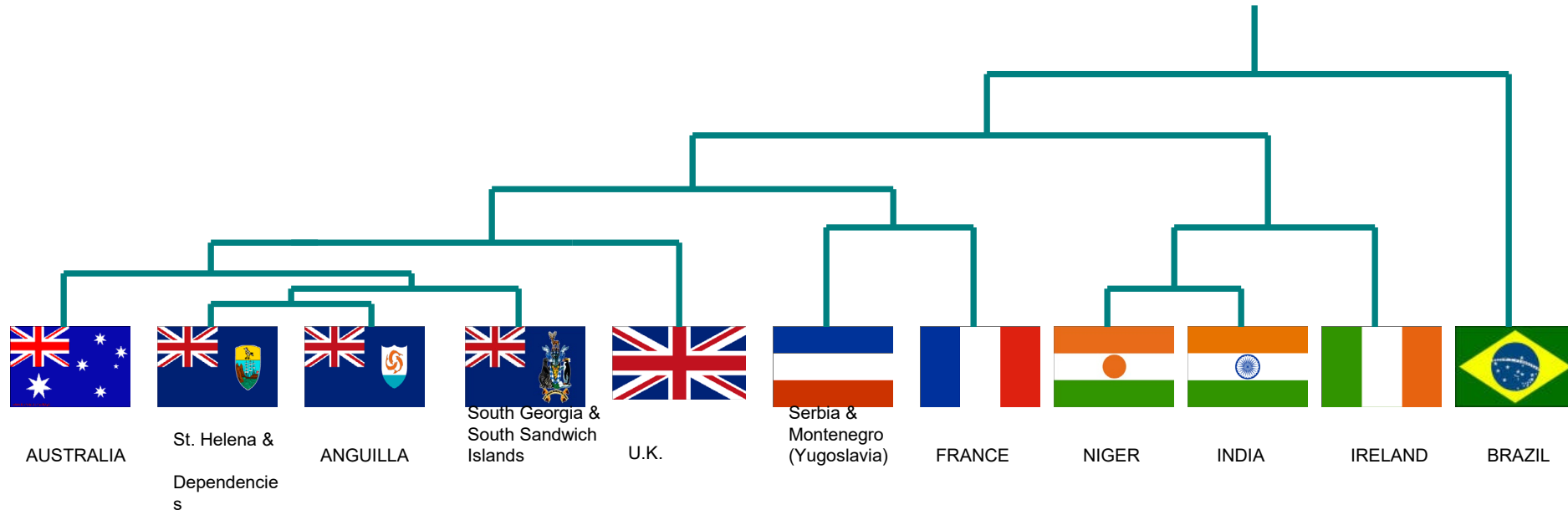
La similitud entre dos objetos en un Dendrogram es representado por el nodo interno mas alto que comparte



La agrupación jerárquica a veces puede mostrar patrones que no tienen sentido o son falsos.

La estrecha agrupación de Australia, Anguila, Santa Elena, etc., es significativa; Todos estos países son antiguas colonias del Reino Unido.

Sin embargo, la estrecha agrupación de Níger e India es completamente espuria; no hay conexión entre los dos.



Clúster Jerárquico

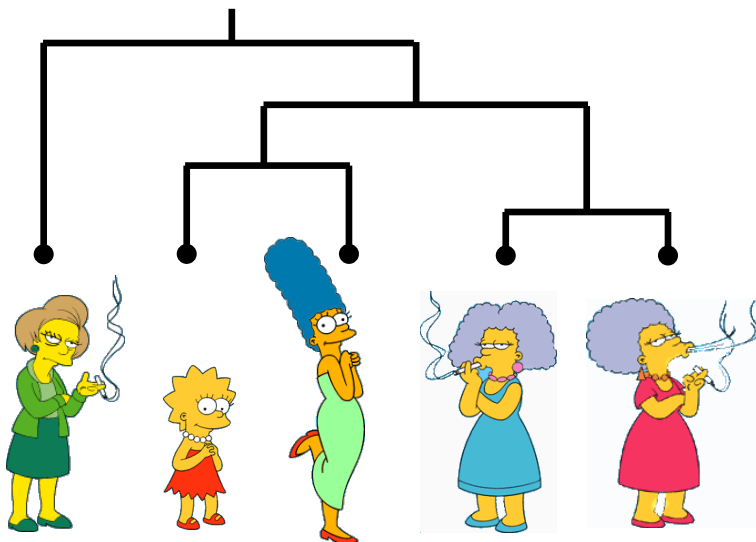
El numero de Dendogramas con n
Hojas = $(2n - 3)! / [(2^{(n-2)}) (n - 2)!]$

Numero De Hojas	Numero de posibles Dendogramas
2	1
3	3
4	15
5	105
...	...
10	34,459,425

Como no podemos probar todos los árboles posibles, tendremos que realizar una búsqueda heurística de todos los árboles posibles. Podríamos hacer esto...

Ascendente (aglomerativo): comenzando con cada elemento en su propio grupo, encuentre el mejor par para fusionarlo en un nuevo grupo. Repita hasta que todos los grupos estén fusionados.

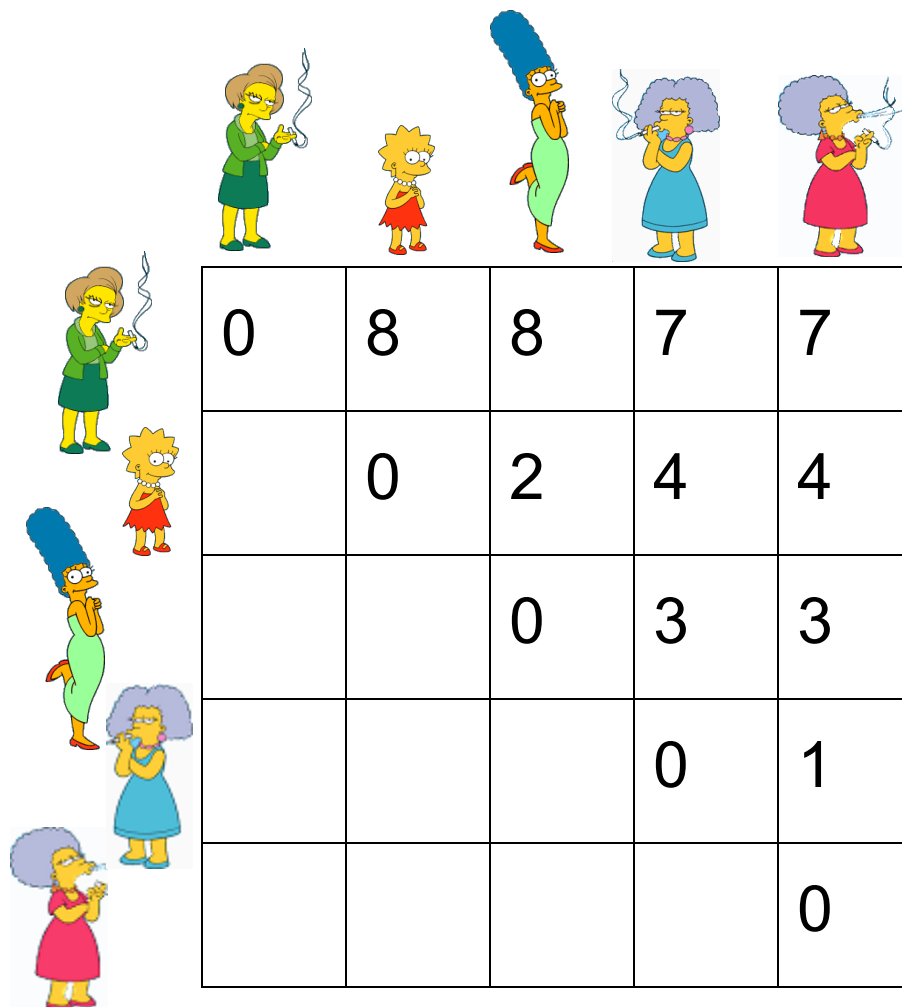
De arriba hacia abajo (divisivo): comenzando con todos los datos en un solo grupo, considere todas las formas posibles de dividir el grupo en dos. Elija la mejor división y opere recursivamente en ambos lados.












Comenzamos con una matriz de distancias entre todos los pares de objetos de nuestra base de datos

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

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

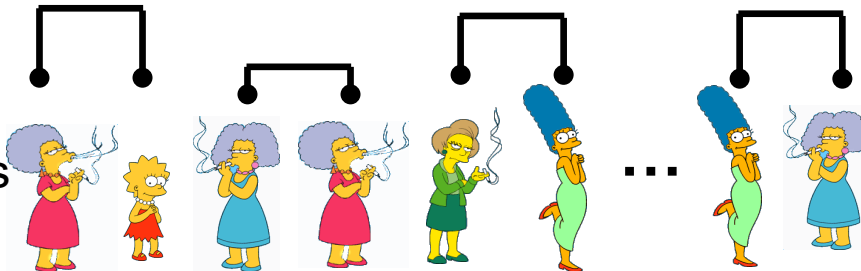


				
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

Ascendente (aglomerativo):

comenzando con cada elemento en su propio grupo, encuentre el mejor par para fusionarlo en un nuevo grupo. Repita hasta que todos los grupos estén fusionados.

Considere todas las combinaciones posibles...



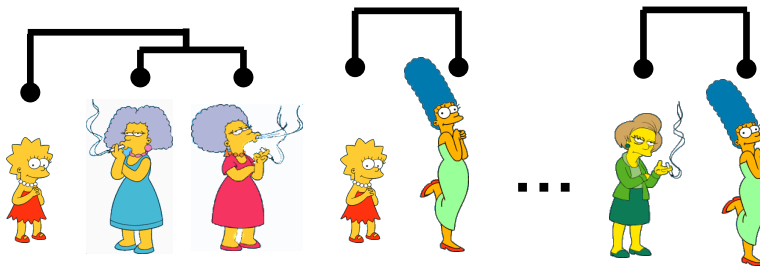
Elige la mejor



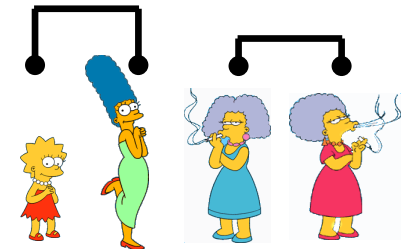
Ascendente (aglomerativo):

comenzando con cada elemento en su propio grupo, encuentre el mejor par para fusionarlo en un nuevo grupo. Repita hasta que todos los grupos estén fusionados.

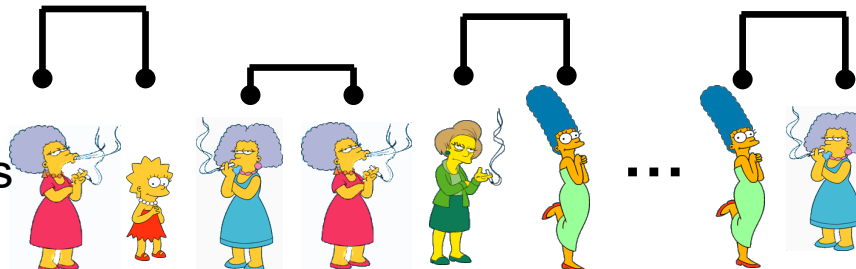
Considere todas las combinaciones posibles...



Elige la mejor



Considere todas las combinaciones posibles...



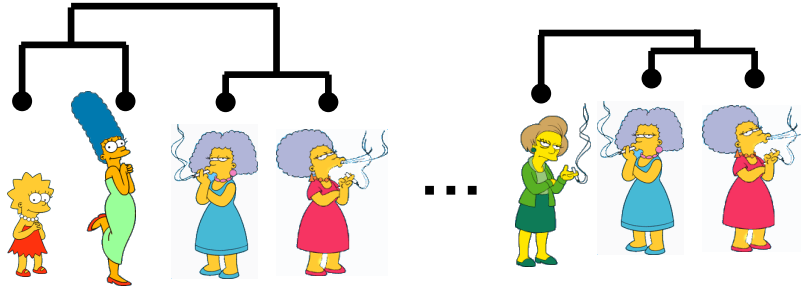
Elige la mejor



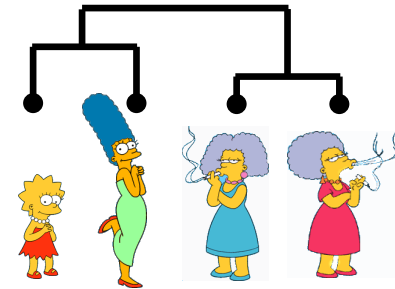
Ascendente (aglomerativo):

comenzando con cada elemento en su propio grupo, encuentre el mejor par para fusionarlo en un nuevo grupo. Repita hasta que todos los grupos estén fusionados.

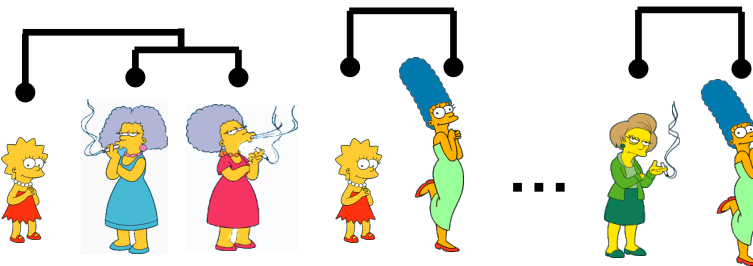
Considere todas las combinaciones posibles...



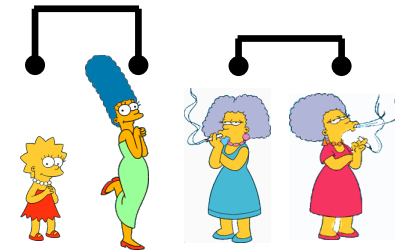
Elige la mejor



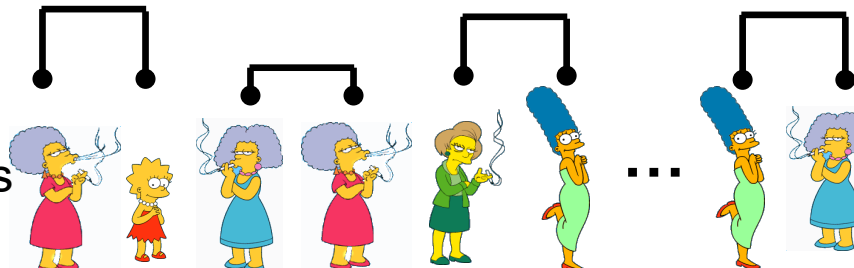
Considere todas las combinaciones posibles...



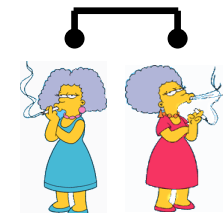
Elige la mejor



Considere todas las combinaciones posibles...

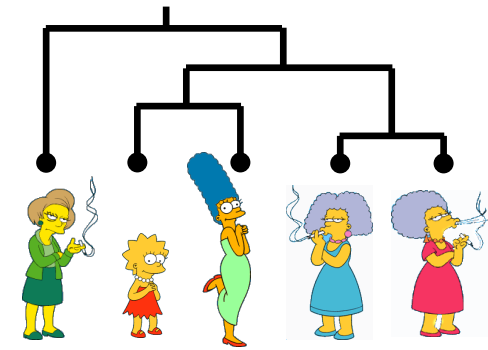


Elige la mejor

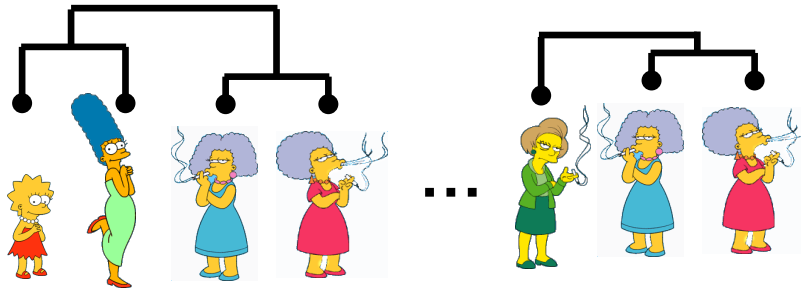


Ascendente (aglomerativo):

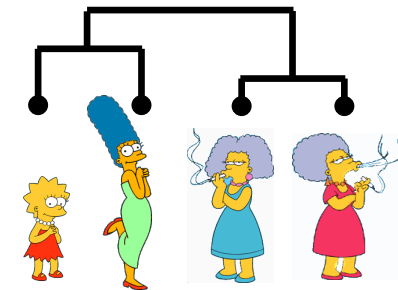
comenzando con cada elemento en su propio grupo, encuentre el mejor par para fusionarlo en un nuevo grupo. Repita hasta que todos los grupos estén fusionados.



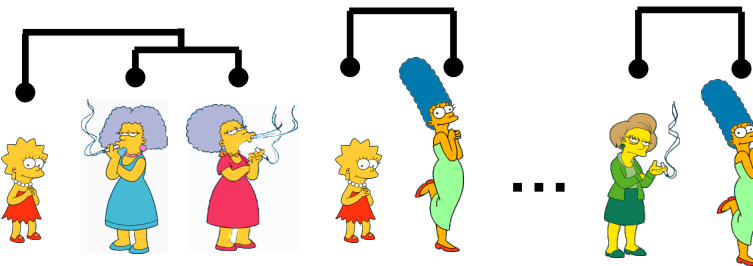
Considere todas las combinaciones posibles...



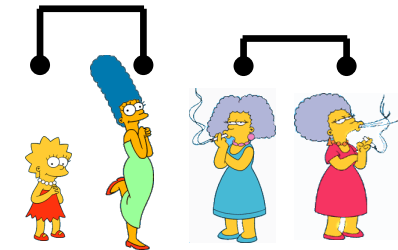
Elige la mejor



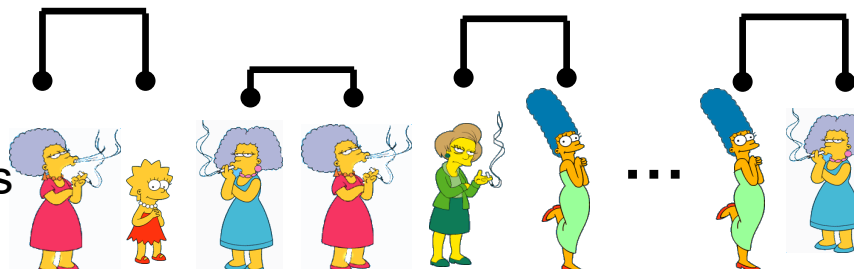
Considere todas las combinaciones posibles...



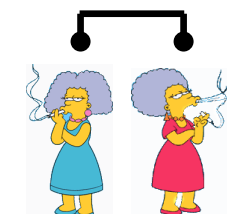
Elige la mejor



Considere todas las combinaciones posibles...

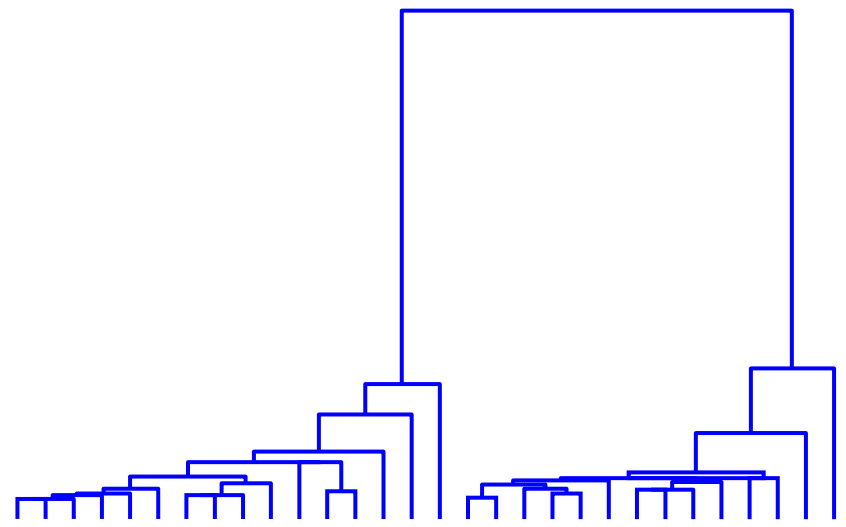
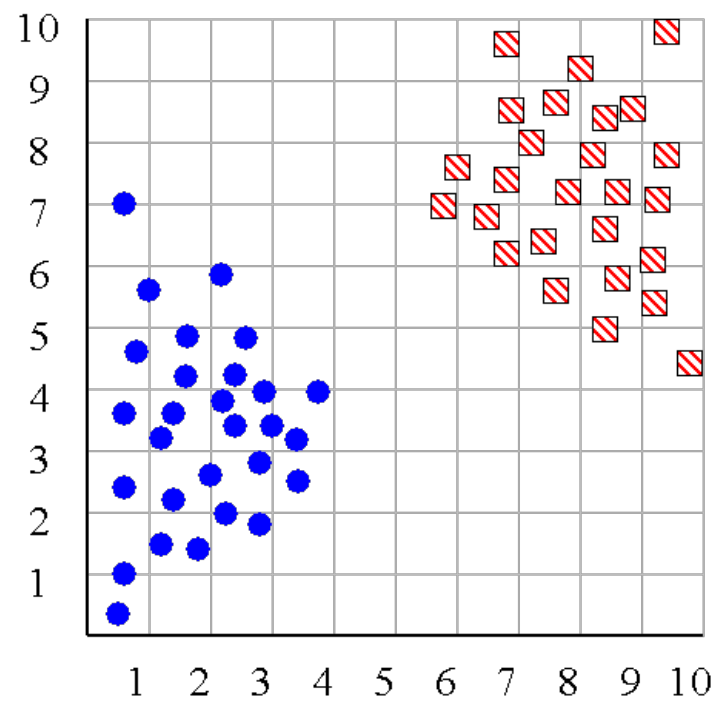


Elige la mejor

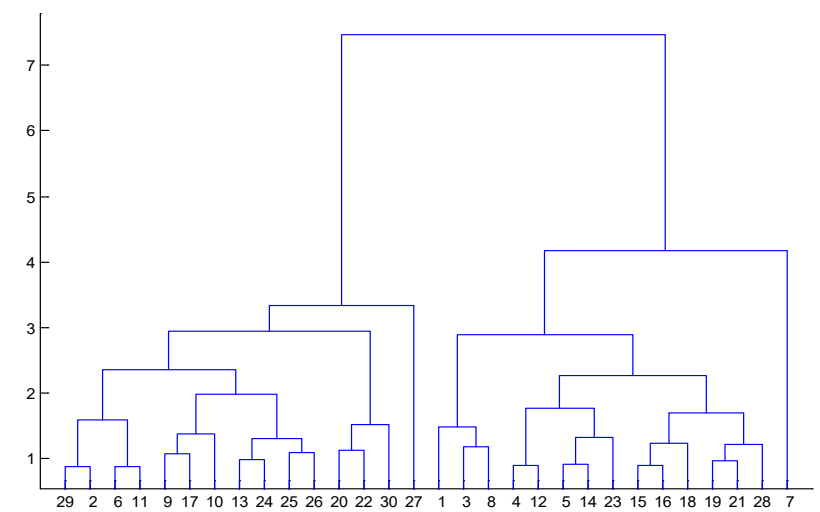


Sabemos cómo medir la distancia entre dos objetos, pero definir la distancia entre un objeto y un grupo, o definir la distancia entre dos grupos, no es obvio.

- **Single linkage (nearest neighbor):** en este método, la distancia entre dos grupos está determinada por la distancia de los dos objetos más cercanos (vecinos más cercanos) en los diferentes grupos.
- **Complete linkage (furthest neighbor):** en este método, las distancias entre grupos están determinadas por la mayor distancia entre dos objetos cualesquiera en los diferentes grupos (es decir, por los "vecinos más lejanos").
- **Group average linkage:** en este método, la distancia entre dos grupos se calcula como la distancia promedio entre todos los pares de objetos en los dos grupos diferentes.



Single linkage



Average linkage

Clustering Jerárquico - Summary

- No necesitamos especificar el numero de clusters
- La naturaleza jerárquica se adapta muy bien a la intuición humana en algunos dominios
- No escalan bien: complejidad temporal de al menos $O(n^2)$, donde n es el número total de objetos
- Como cualquier algoritmo de búsqueda heurística, los óptimos locales son un problema
- La interpretación de los resultados es (muy) subjetiva.