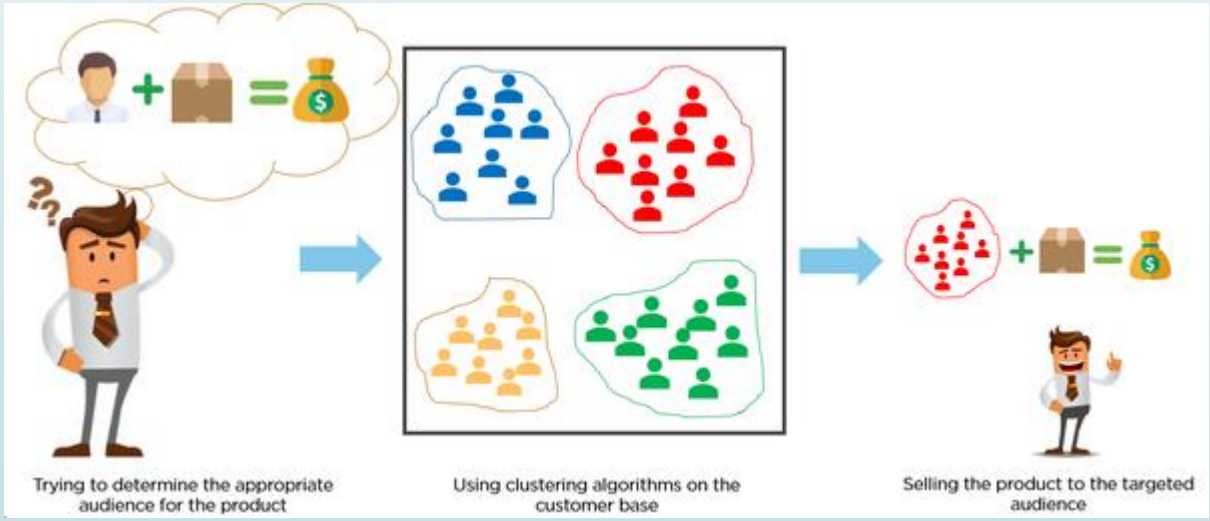
BASE DE DATOS II – LSI- 2019

AGRUPAMIENTO

CLUSTERING



AGRUPAMIENTO O CLUSTERING

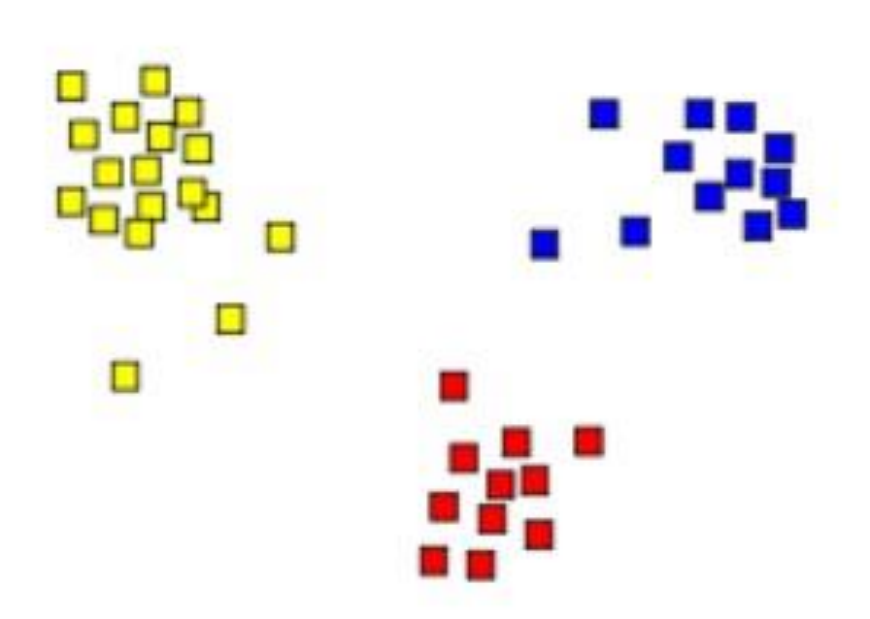
- El *clustering* es uno de los métodos de aprendizaje **no supervisado** más importantes y busca caracterizar conceptos desconocidos a partir de los ejemplos disponibles.
- Generalmente, en un problema real **se desconoce la clase** y es allí donde el agrupamiento puede ayudar a identificar las características comunes entre instancias.
- Al no disponer de la clase utiliza una **medida de similitud (distancia)** para determinar el parecido entre instancias.

AGRUPAMIENTO O CLUSTERING

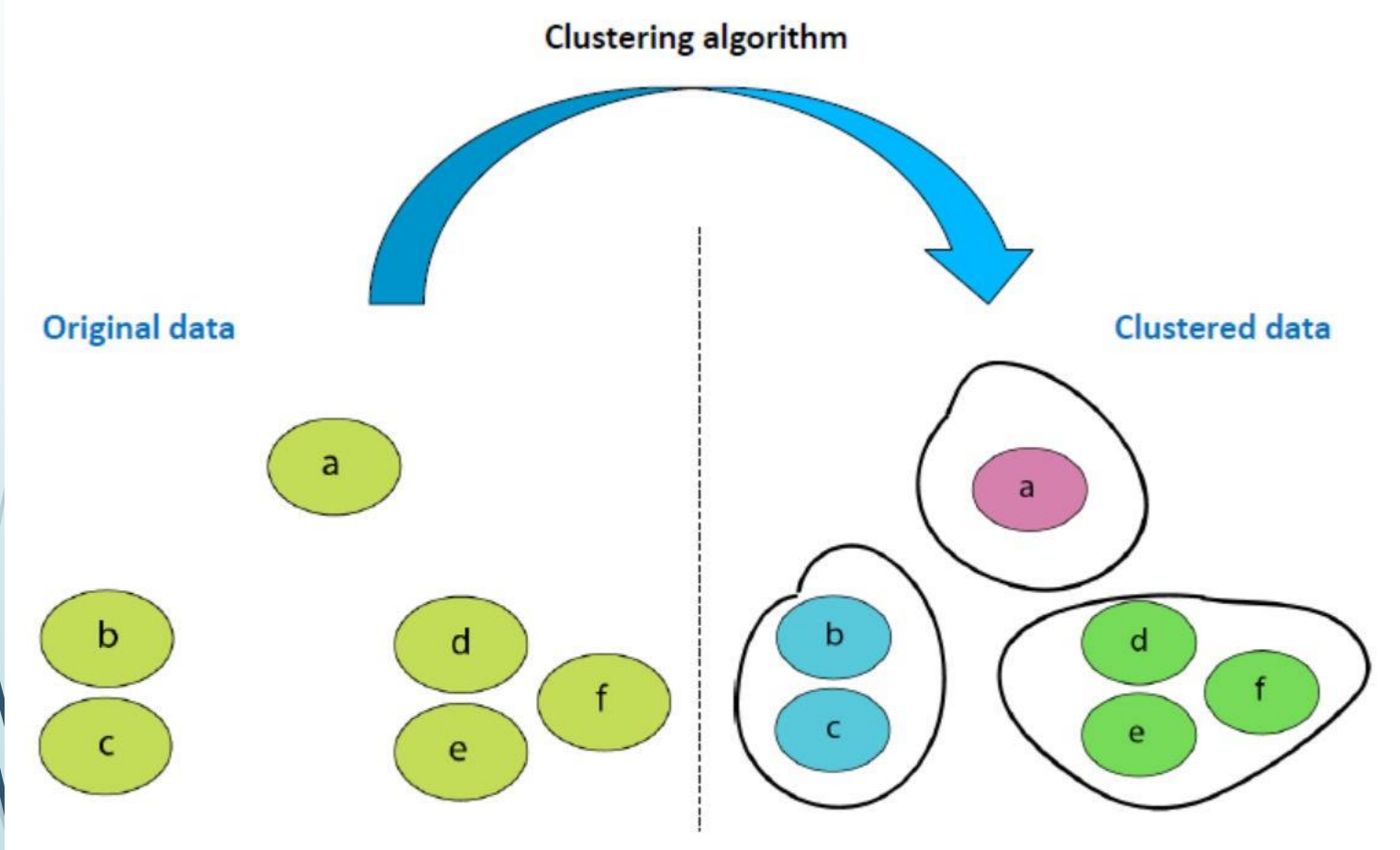
- Permite encontrar grupos de instancias con características similares.
- Aplicaciones
  - Identificar grupos y describirlos
    - Detectar clientes con características similares para ofrecer servicios adecuados.
    - Identificar alumnos con rendimientos académicos similares con el objetivo de reducir la deserción escolar.
  - Detección de casos anómalos
    - Detección de fraudes.

AGRUPAMIENTO O CLUSTERING

- El resultado de aplicar una técnica de *clustering* es una serie de agrupamientos o *clusters* formados al particionar las instancias.



AGRUPAMIENTO - OBJETIVO

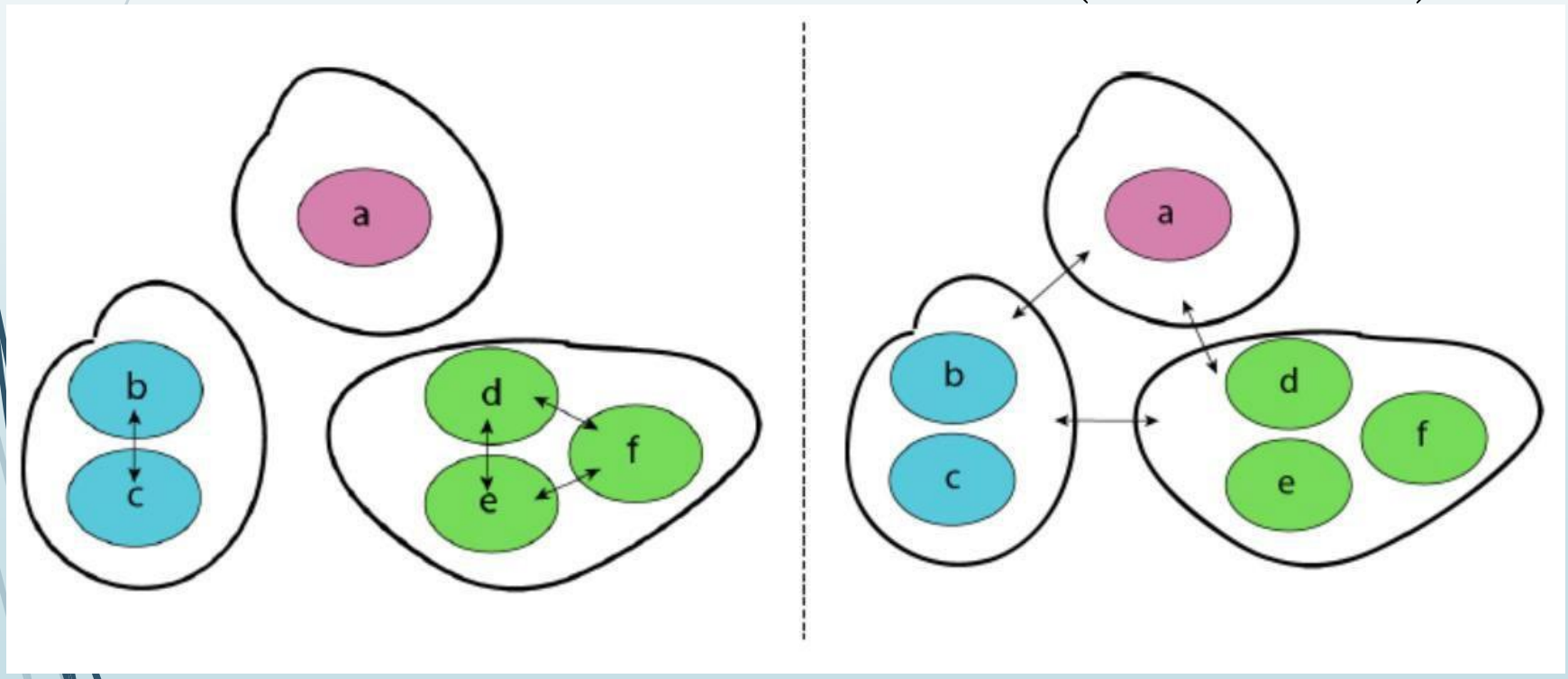


Métricas del agrupamiento obtenido

- Un buen método de agrupamiento producirá grupos de alta calidad en los cuales
  - El parecido entre los elementos que componen un mismo grupo es alto (intra-cluster).
  - El parecido entre los elementos de grupos distintos es bajo (inter-cluster).

AGRUPAMIENTO - OBJETIVO

- Minimizar la distancia entre los elementos de un mismo cluster (intra-cluster)
- Maximizar la distancia entre clusters (inter-cluster)

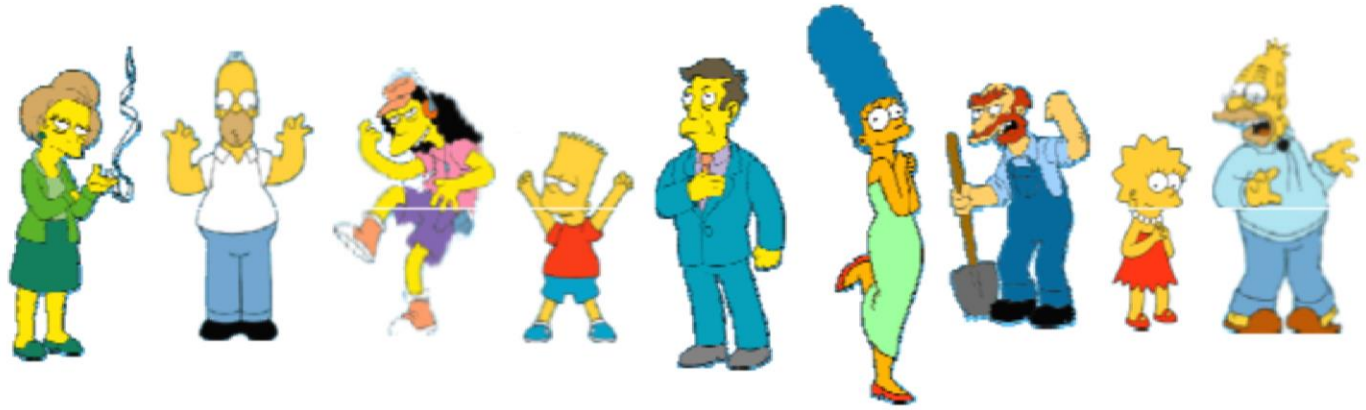


Tipos de algoritmos de agrupamiento

- Algoritmo Partitivo
  - Particionan los datos creando un número K de clusters.
  - Una instancia pertenece a un único grupo.
- Algoritmo Jerárquico
  - Generan una estructura jerárquica de clusters que permiten ver las particiones de las instancias con distinta granularidad.
  - Una instancia pertenece a un único grupo.
- Algoritmo probabilista
  - Los clusters se generan con un método probabilístico

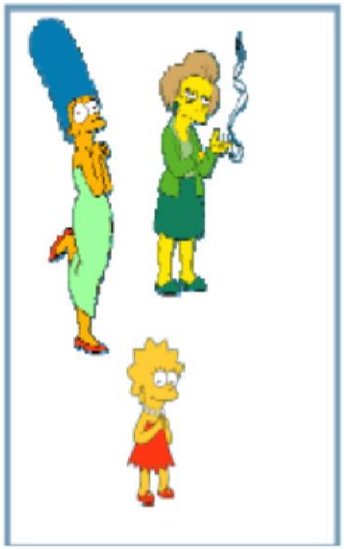
MEDIDAS DE DISTANCIA

¿Cuál es la forma natural de agrupar los personajes?

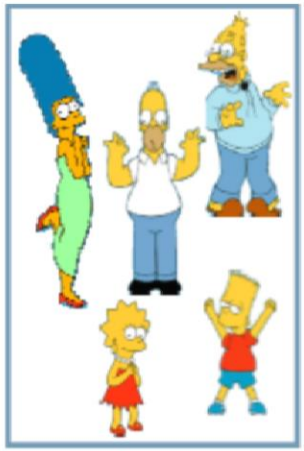


MEDIDAS DE DISTANCIA

Mujeres
vs.
Hombres



Simpsons
vs.
Empleados de
la escuela de
Springfield



MEDIDAS DE DISTANCIA

- A la hora de calcular la distancia entre dos objetos
  - No tienen porque utilizarse todos los atributos disponibles del conjunto de datos.
  - Hay que tener cuidado con las magnitudes de cada variable.
- Usualmente se expresan en términos de distancia:

$$d(i,j) > d(i,k)$$

Nos indica que el objeto i es mas parecido al objeto k que a j.

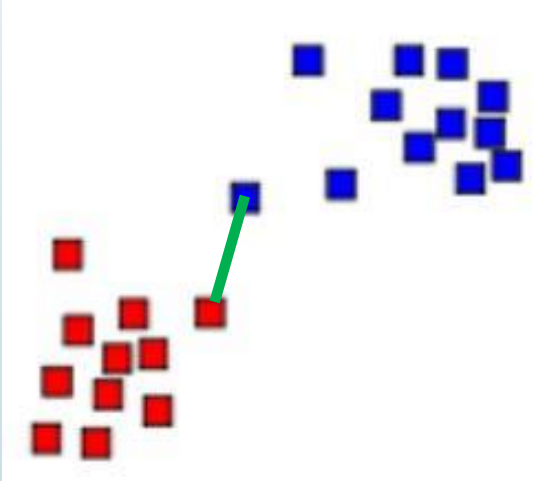
MEDIDAS DE DISTANCIA

- Se utilizan para estimar la similitud entre instancias al momento de decidir si deben ser incluidas en el mismo grupo o en grupos diferentes.
- La más utilizada suele ser la Distancia Euclídea

$$distancia\ (X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

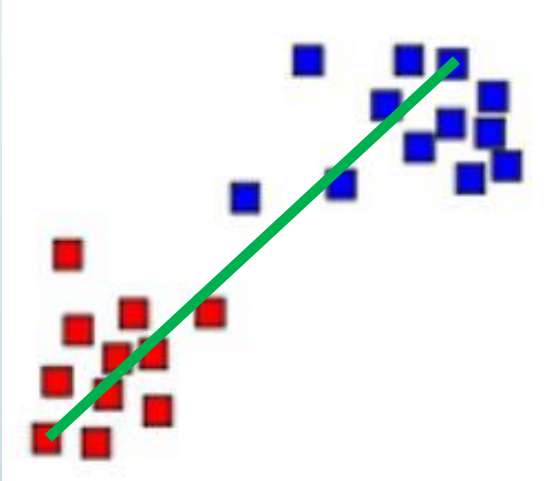
MEDIDAS DE CONECTIVIDAD (*LINKAGE MEASURES*)

- **Enlace sencillo (*single-linkage*)**
  - La similitud entre dos clusters se calcula como la similitud de los **dos puntos más cercanos** pertenecientes a los diferentes clusters.



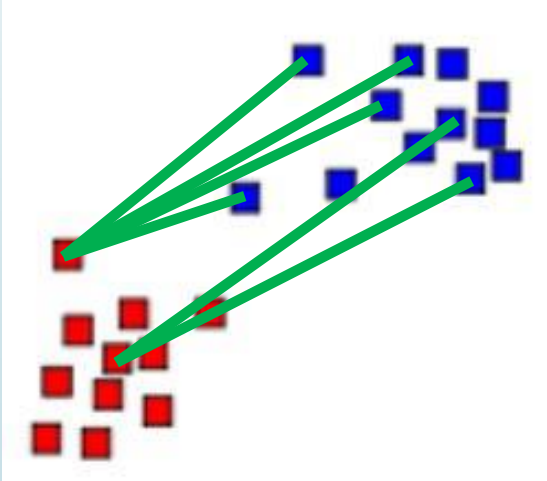
MEDIDAS DE CONECTIVIDAD (*LINKAGE MEASURES*)

- **Enlace completo (*complete-linkage*)**
  - La similitud entre dos clusters se calcula como la similitud de los **dos puntos más lejanos** pertenecientes a los diferentes clusters.



MEDIDAS DE CONECTIVIDAD (*LINKAGE MEASURES*)

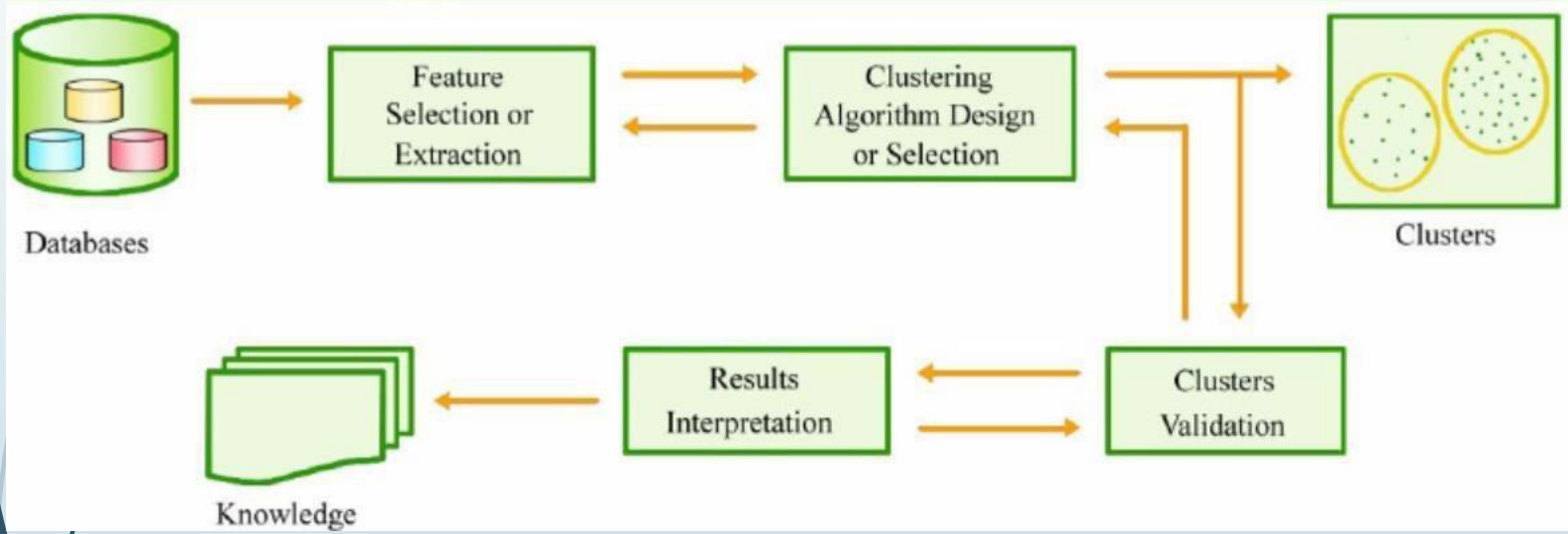
- **Enlace promedio (*average-linkage*)**
  - La distancia entre dos grupos se calcula promediando las distancias entre todos los pares que se puedan formar tomando una instancia de cada cluster.



Proceso de agrupamiento

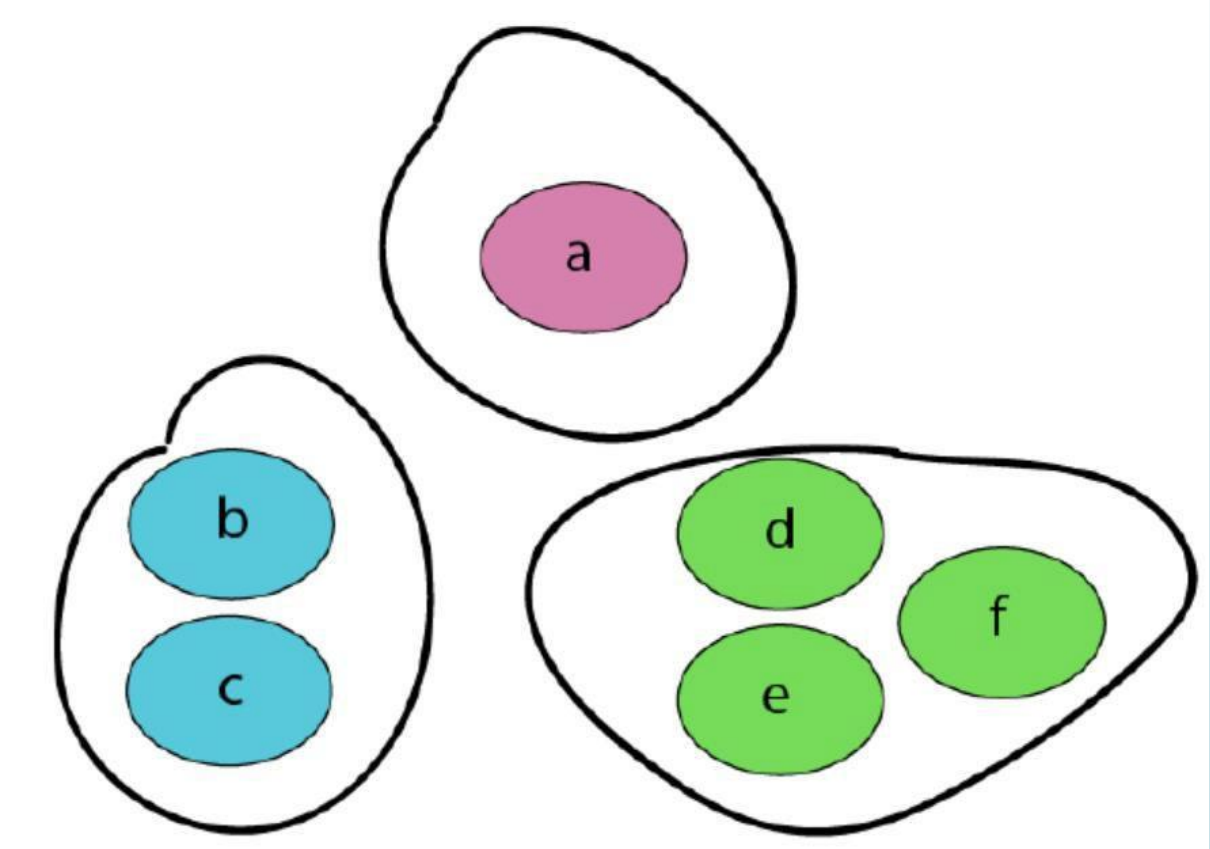
- Seleccionar las características relevantes
- Definir una representación adecuada.
- Definir la medida de similitud a utilidad (medida de distancia). Depende del problema.
- Aplicar un algoritmo de agrupamiento
- Validar los grupos obtenidos y de ser necesario volver a repetir el proceso.

PROCESO DE AGRUPAMIENTO



ALGORITMOS DE CLUSTERING PARTITIVOS

- Obtiene una única partición de los datos



K-MEDIAS

- El algoritmo K-Medias fue propuesto por MacQueen, en 1967.
- Requiere conocer a priori el número K de grupos a formar.
- El algoritmo está basado en la minimización de la distancia interna (la suma de las distancias de los patrones asignados a un agrupamiento al centroide de dicho agrupamiento).
- De hecho, este algoritmo minimiza la suma de las distancias al cuadrado de cada patrón al centroide de su agrupamiento.

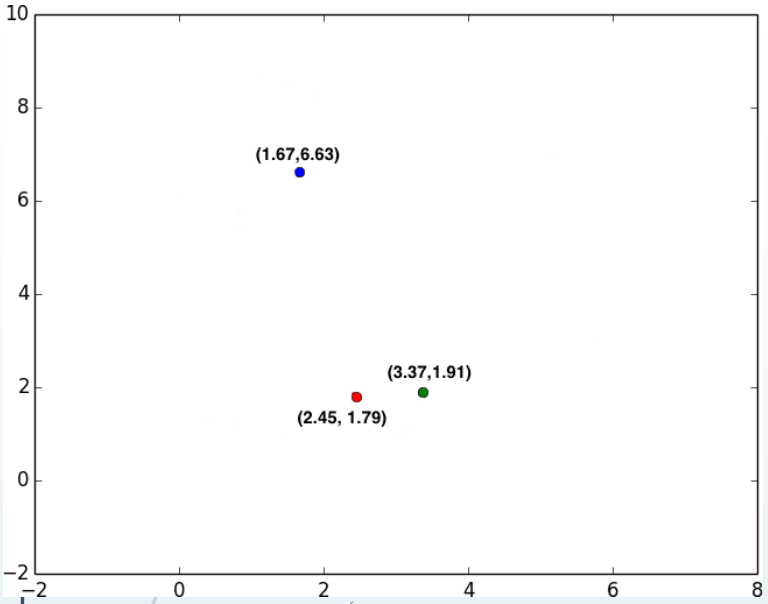
K-MEDIAS

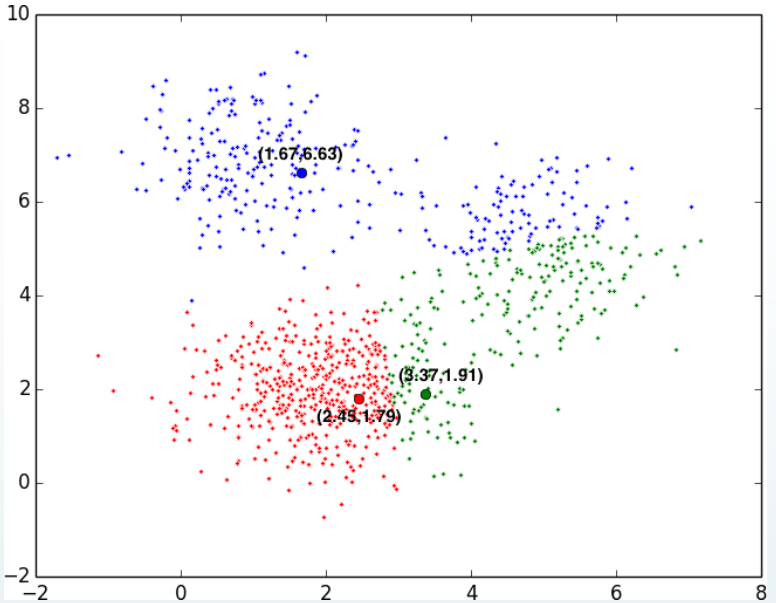
- Características
  - El algoritmo es sencillo y eficiente.
  - Procesa los patrones secuencialmente (por lo que requiere un almacenamiento mínimo).
  - Está sesgado por el orden de presentación de los patrones (los primeros patrones determinan la configuración inicial de los agrupamientos)
  - Su comportamiento depende enormemente del parámetro K.

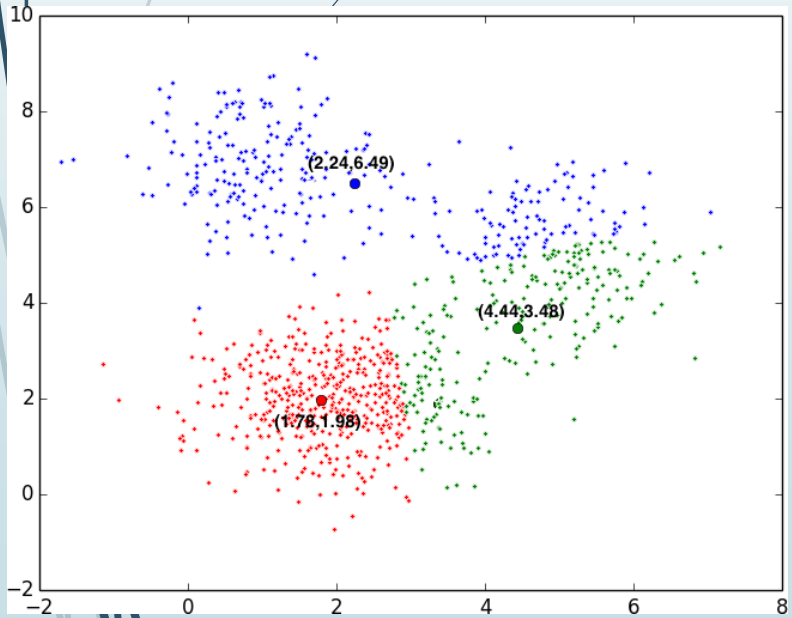
Algoritmo K-Medias

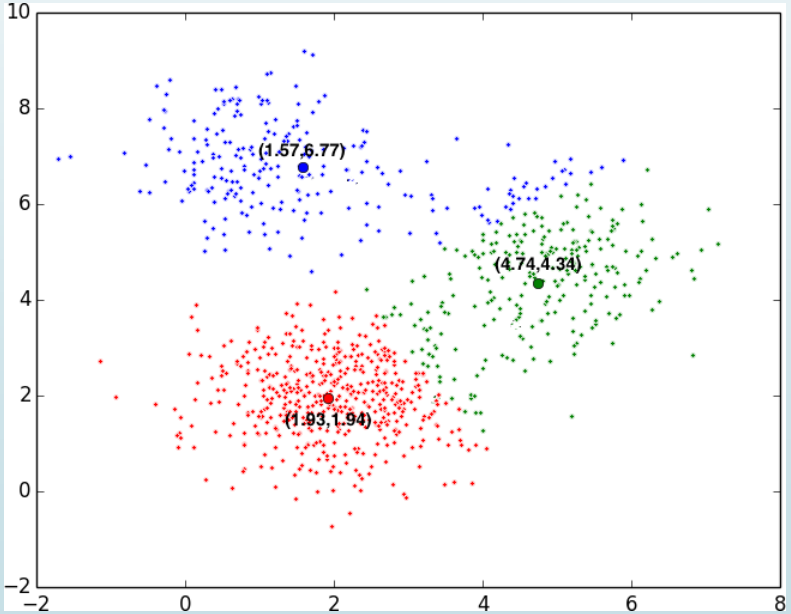
- Elegir aleatoriamente K vectores de entrada como centros iniciales.
- Repetir
  - Calcular los centros de los K clusters.
  - Redistribuir los patrones entre los clusters utilizando la mínima distancia euclídea al cuadrado como clasificador.

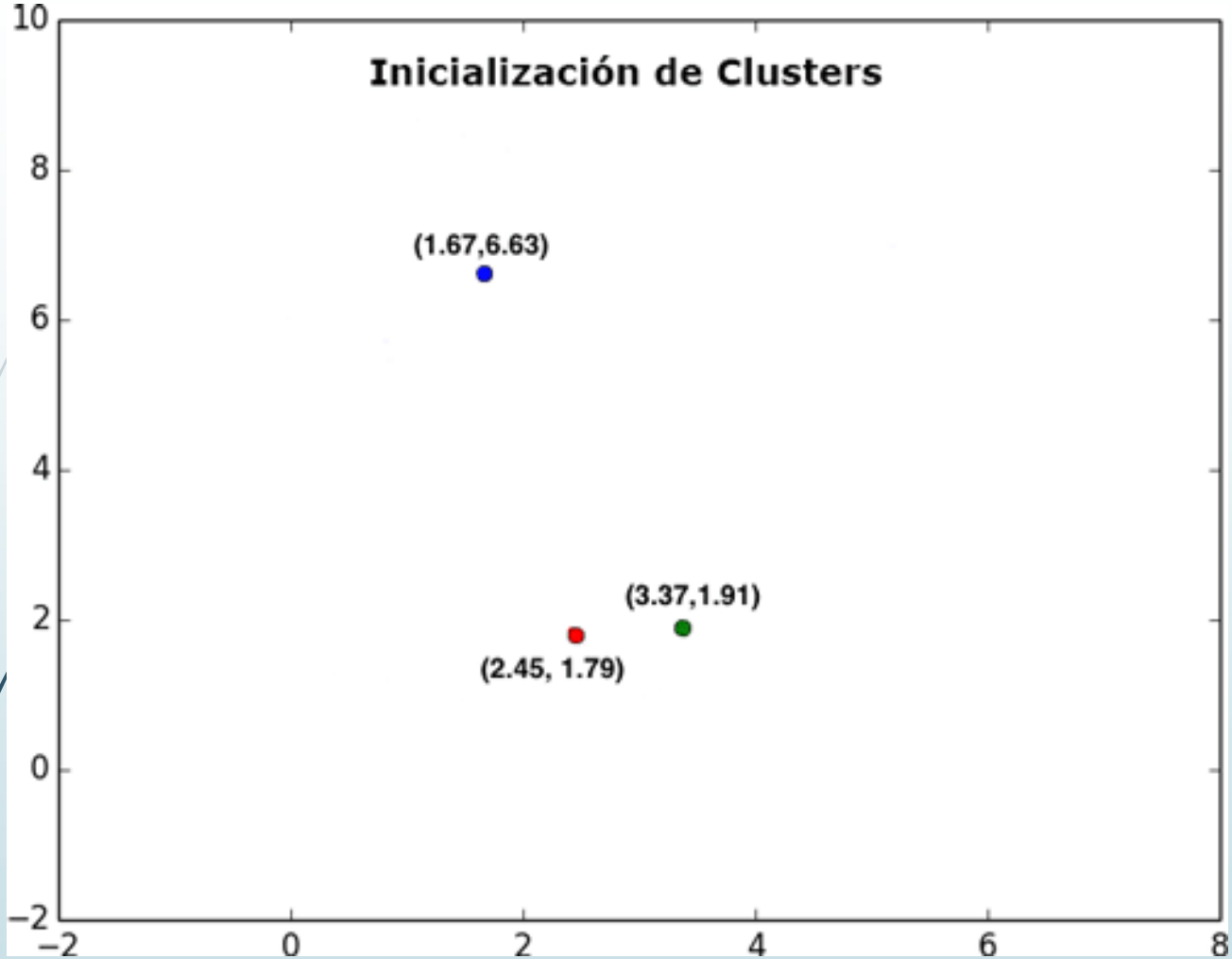
hasta que no cambien los centros de los clusters


(1.67,6.63)
(3.37,1.91)
(2.45, 1.79)

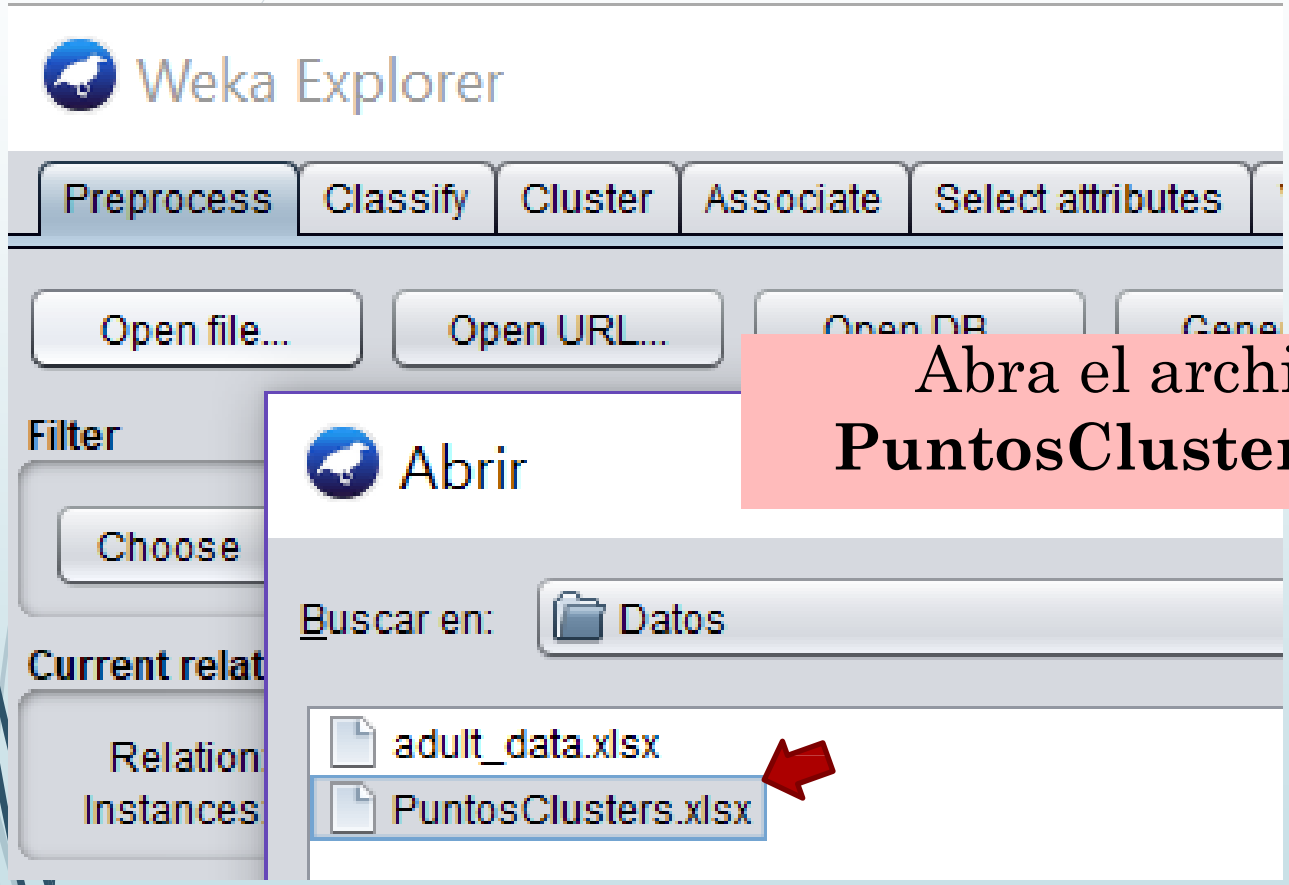

(1.67,6.63)
(3.37,1.91)
(2.45,1.79)


(2,24,6.49)
(4.44,3.48)
(1.78,1.98)


(1.57,6.77)
(4.74,4.34)
(1.93,1.94)

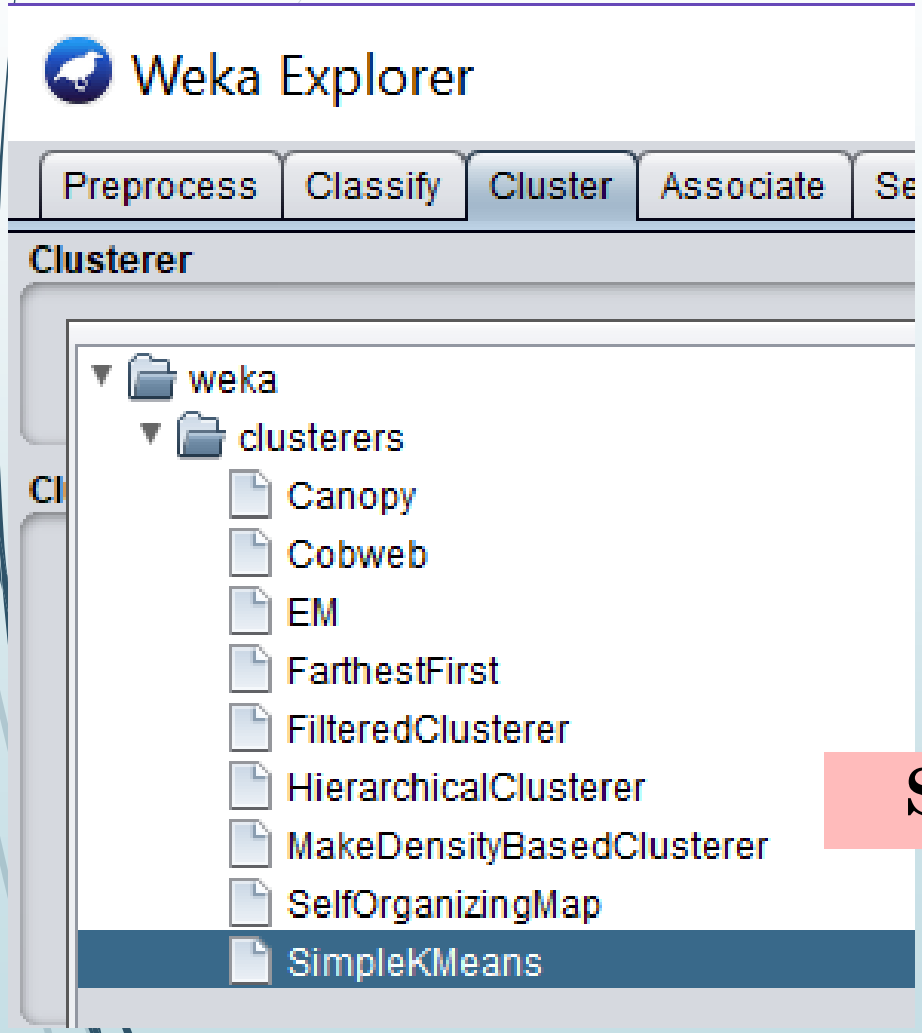

Inicialización de Clusters
(1.67,6.63)
(3.37,1.91)
(2.45, 1.79)
10
8
6
4
2
0
−2
−2
0
2
4
6
8

K-MEDIAS DESDE WEKA

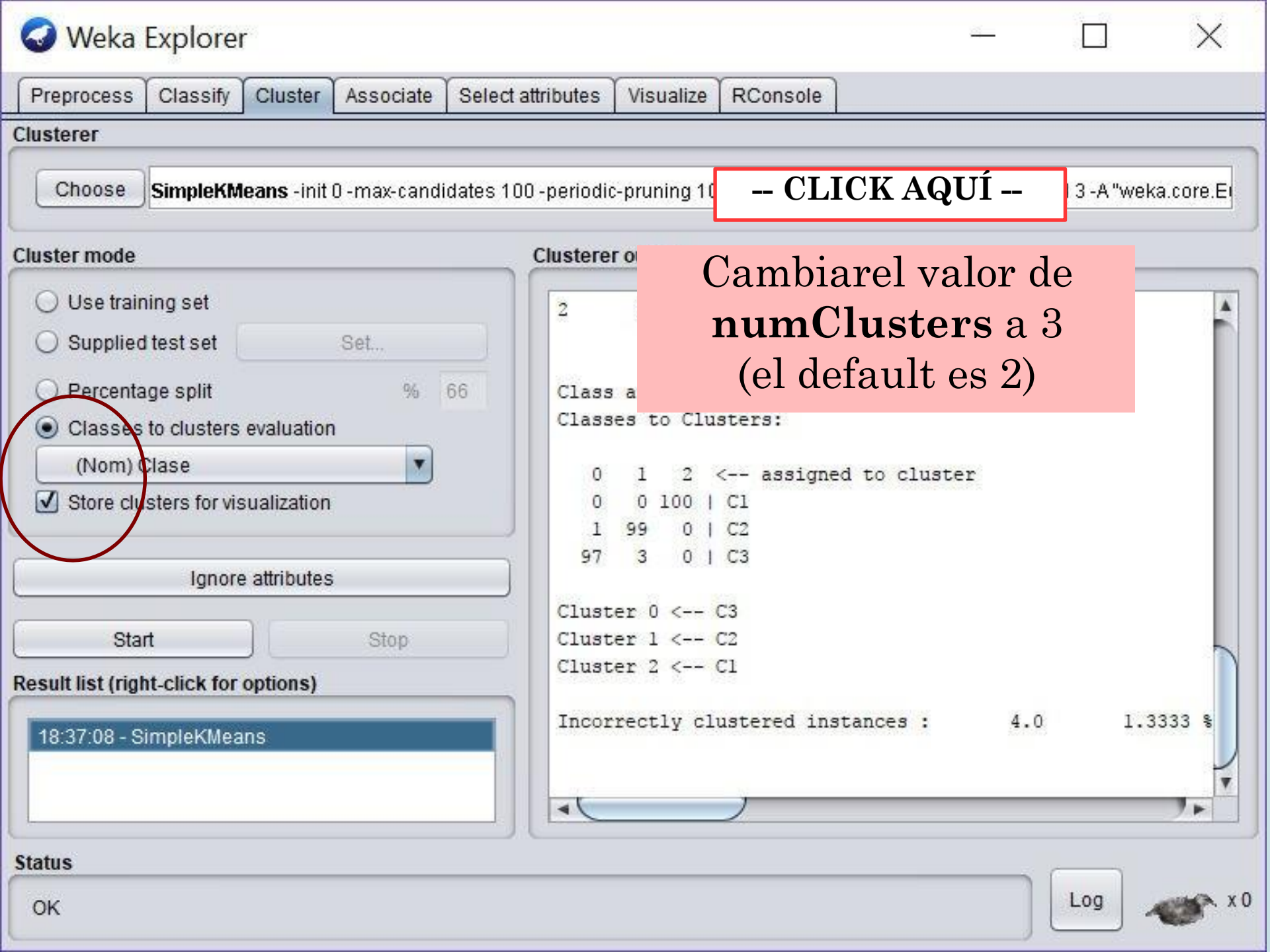


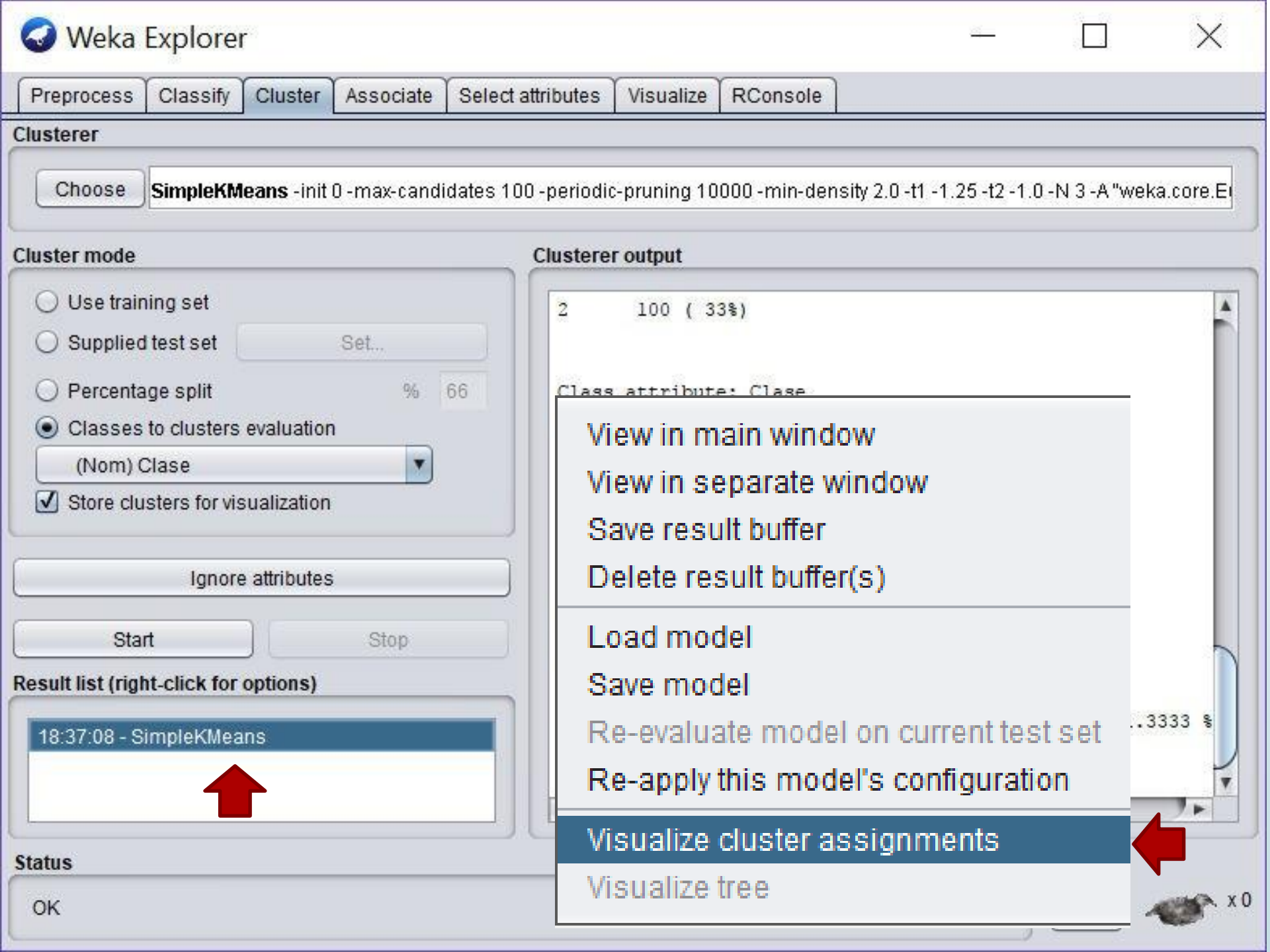
Abra el archivo
PuntosClusters.xls

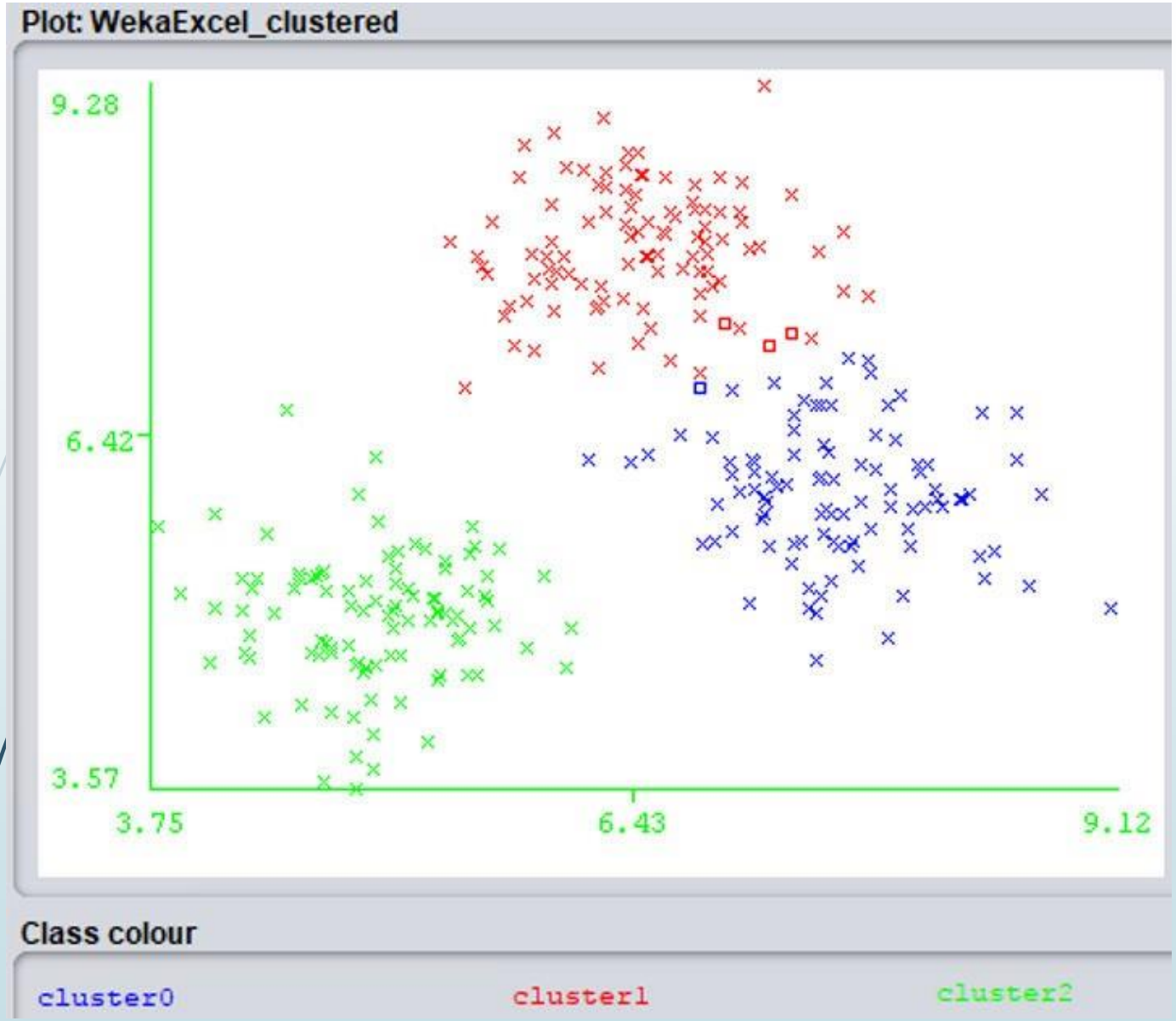
K-MEDIAS DESDE WEKA



SimpleKMeans


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
RConsole
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10
3 -A "weka.core.E
-- CLICK AQUÍ --
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Clase
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
18:37:08 - SimpleKMeans
Clusterer o
2
Class a
Classes to Clusters:
0 1 2 <-- assigned to cluster
0 0 100 | C1
1 99 0 | C2
97 3 0 | C3
Cluster 0 <-- C3
Cluster 1 <-- C2
Cluster 2 <-- C1
Incorrectly clustered instances : 4.0 1.3333 %
Cambiarel valor de
numClusters a 3
(el default es 2)
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
RConsole
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.E
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Clase
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
18:37:08 - SimpleKMeans
Clusterer output
2 100 (33%)
Class attribute: Clase
View in main window
View in separate window
Save result buffer
Delete result buffer(s)
Load model
Save model
Re-evaluate model on current test set
Re-apply this model's configuration
Visualize cluster assignments
Visualize tree
.3333 %
Status
OK
x 0


Plot: WekaExcel_clustered
9.28
6.42
3.57
3.75
6.43
9.12
Class colour
cluster0
cluster1
cluster2

K-MEDIAS DESDE WEKA

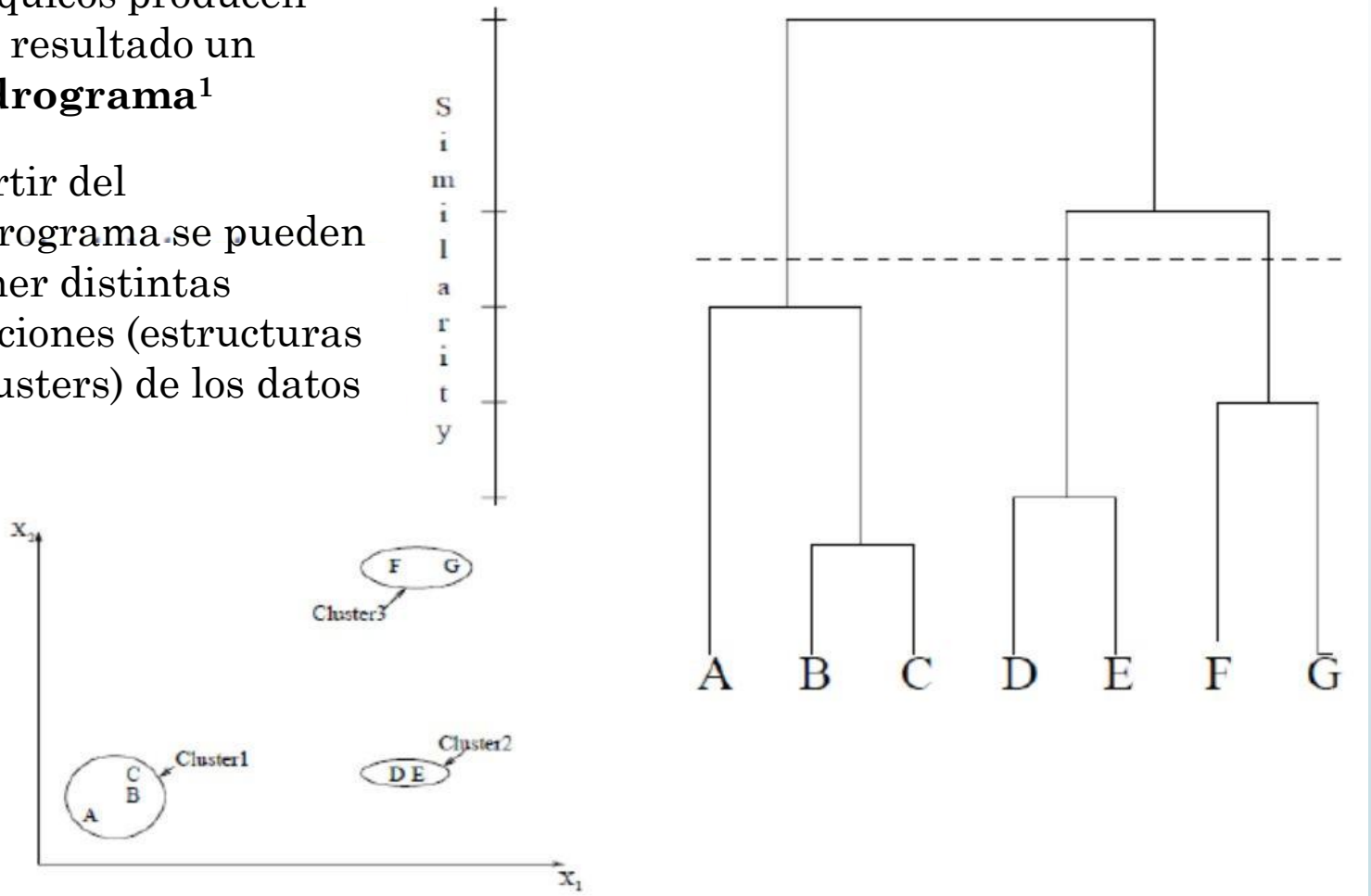
- Centroides obtenidos como resultado del agrupamiento con k=3

```
Final cluster centroids:
                          Cluster#
Attribute     Full Data           0           1           2
                (300.0)      (98.0)     (102.0)     (100.0)
=============================================================
X1                6.3403      7.5672      6.4782      4.9972
X2                6.2868      5.9504      7.9324       4.938
```

ALGORITMO DE CLUSTERING JERÁRQUICOS

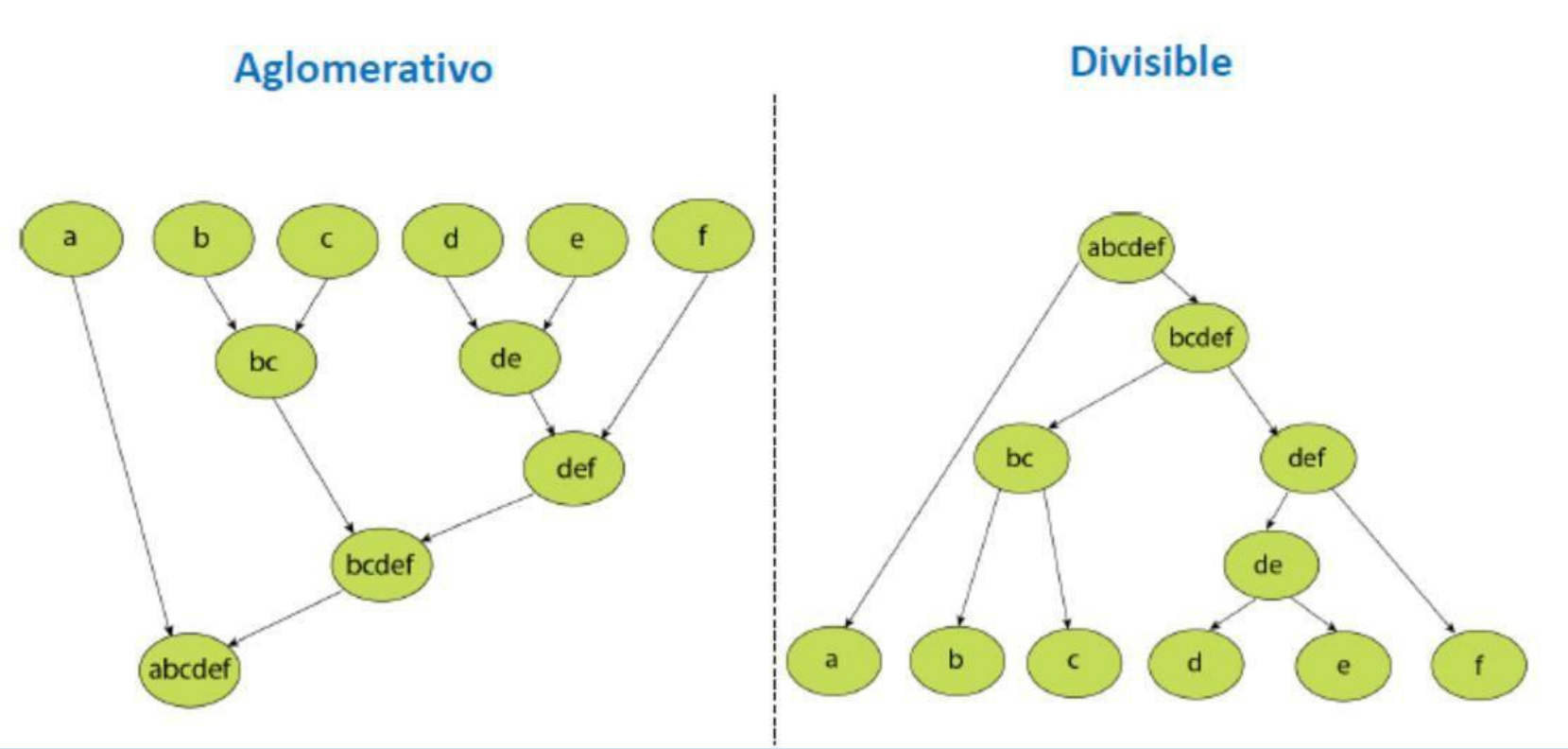
Todos los algoritmos jerárquicos producen como resultado un **dendrograma**[1]

A partir del dendrograma se pueden obtener distintas particiones (estructuras de clusters) de los datos



Dendrograma es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo en otras sucesivamente).

ALGORITMO DE CLUSTERING JERÁRQUICOS



Algoritmo Jerárquico aglomerativo

- **Paso 1**: A cada instancia se le asigna un cluster, de modo que inicialmente si hay N instancias se tienen N clusters de 1 elemento cada uno.
- **Paso 2**: Calcular la distancia entre clusters y unir en uno solo a los dos más cercanos.
- **Paso 3**: Calcular la distancia del nuevo cluster a los restantes.
- **Paso 4**: Repetir los pasos 2 y 3 hasta que todas las instancias pertenezcan al mismo cluster

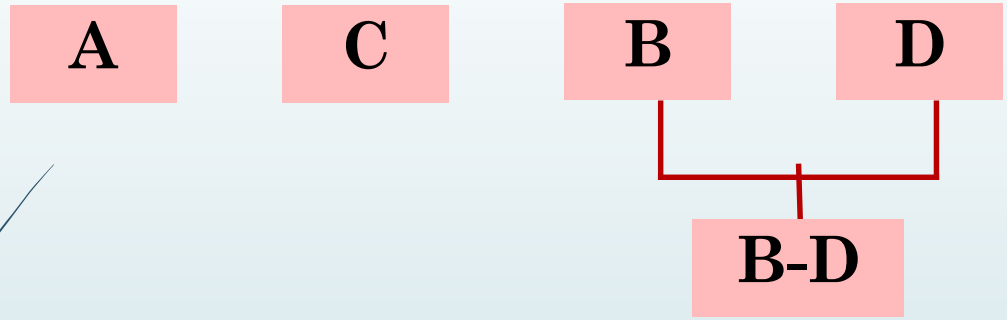
Ejemplo

- Aplique un algoritmo jerárquico aglomerativo para agrupar las instancias A, B, C y D cuya matriz de distancias se indica a continuación

	A	B	C	D
A	0	14	10	6
B	14	0	8	4
C	10	8	0	5
D	6	4	5	0

Ejemplo

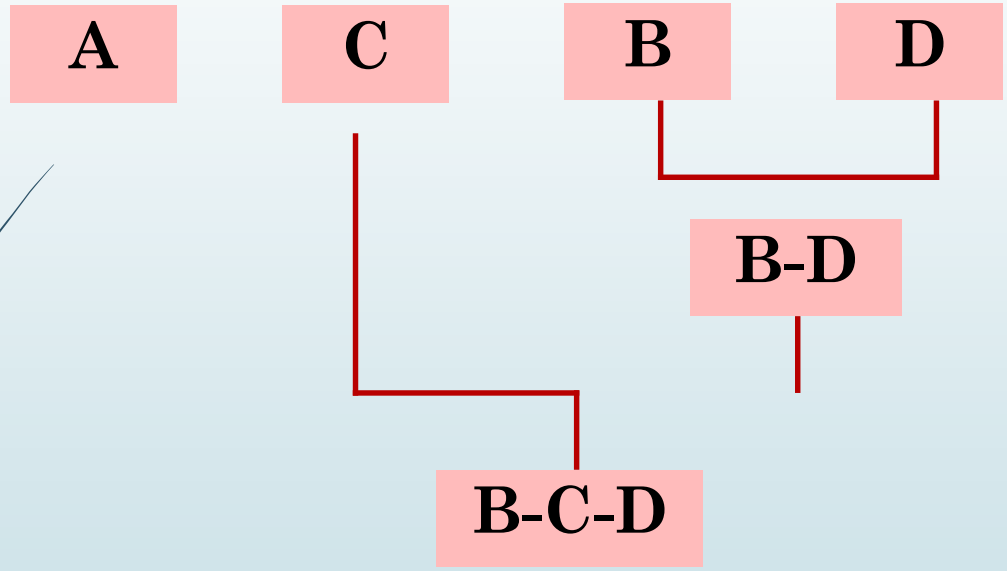
	A	B	C	D
A	0	14	10	6
B	14	0	8	4
C	10	8	0	5
D	6	4	5	0



	A	C	B-D
A	0	10	6
C	10	0	5
B-D	6	5	0

EJEMPLO

	A	C	B-D
A	0	10	6
C	10	0	5
B-D	6	5	0



	A	B-C-D
A	0	6
B-C-D	6	0

EJEMPLO

	A	B-C-D
A	0	6
B-C-D	6	0