TEMA 6. TÉCNICAS DE CLUSTERING

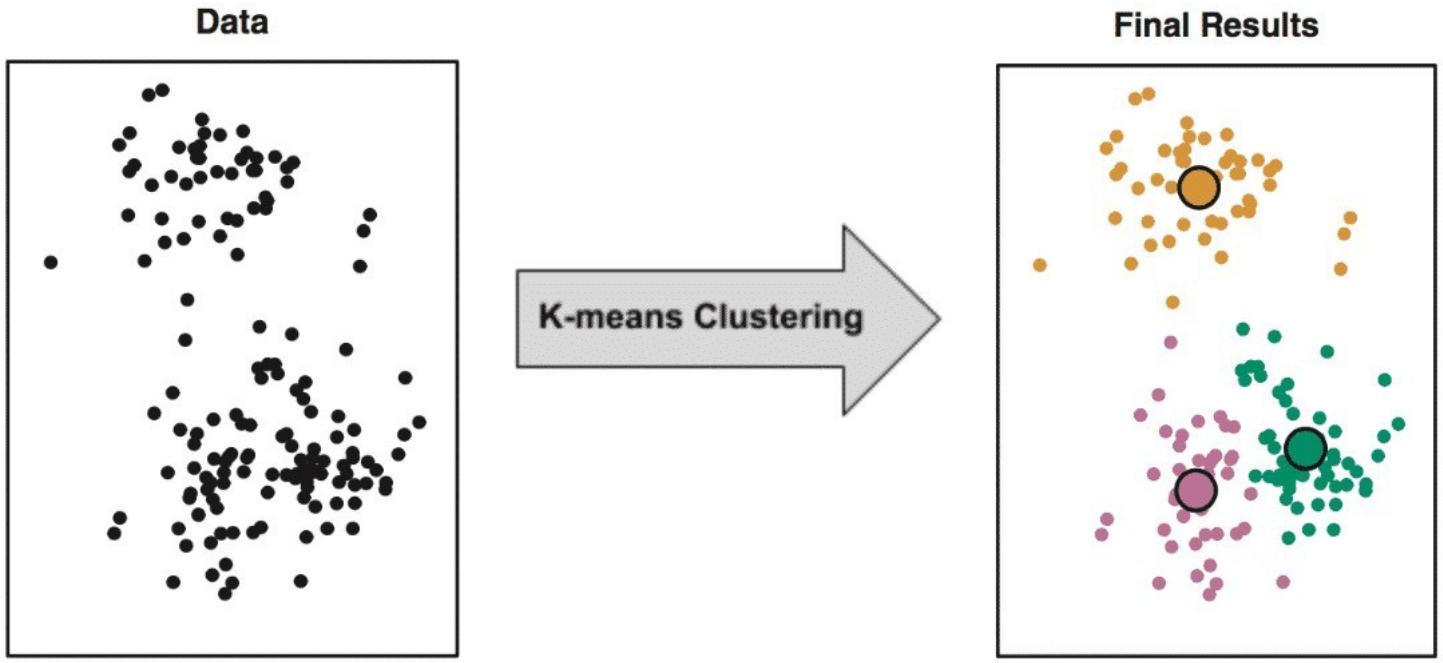
Contenidos

1. Introducción
2. Definiciones y representación de patrones
3. Validación en clustering
4. Algoritmos de clustering
   a) Clustering particional
   b) Clustering jerárquico
   c) Jerárquico frente a particional
5. Ejercicio de clustering

1. Introducción

- **¿Qué es clustering?**

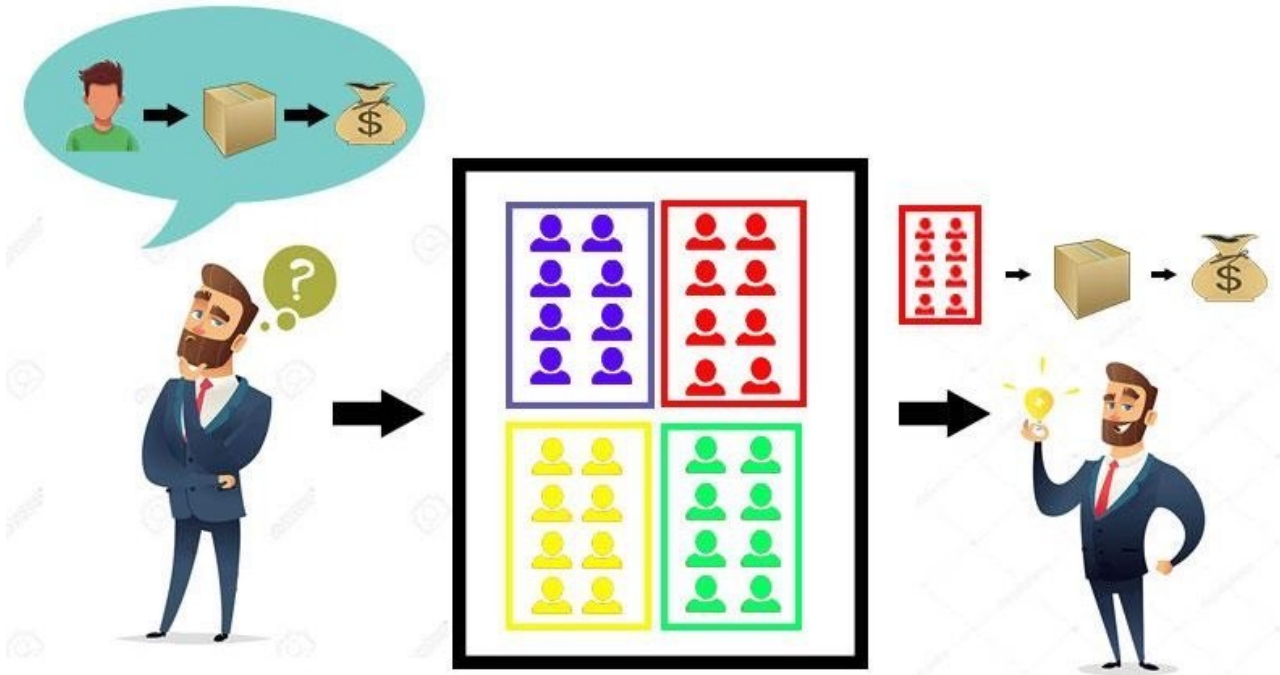
  Consiste en organizar datos no etiquetados en grupos similares llamados clusters



1. Introducción

- **Clustering**
  - Proceso de **agrupamiento** o clasificación no supervisada de objetos
- **Planteamiento del problema**
  - Se pretende clasificar los individuos de una BD en grupos según la similitud encontrada entre ellos.
  - En principio no se conoce la distribución de los datos.
  - No se conoce el número de grupos.
  - Se dispone sólo de la información de los atributos.
- **Aplicación**
  - Áreas de trabajo con poca o ninguna información sobre las etiquetas de los individuos.
  - Probar la calidad del clustering y su relación con la clase en aprendizaje supervisado (análisis cluster-clase).

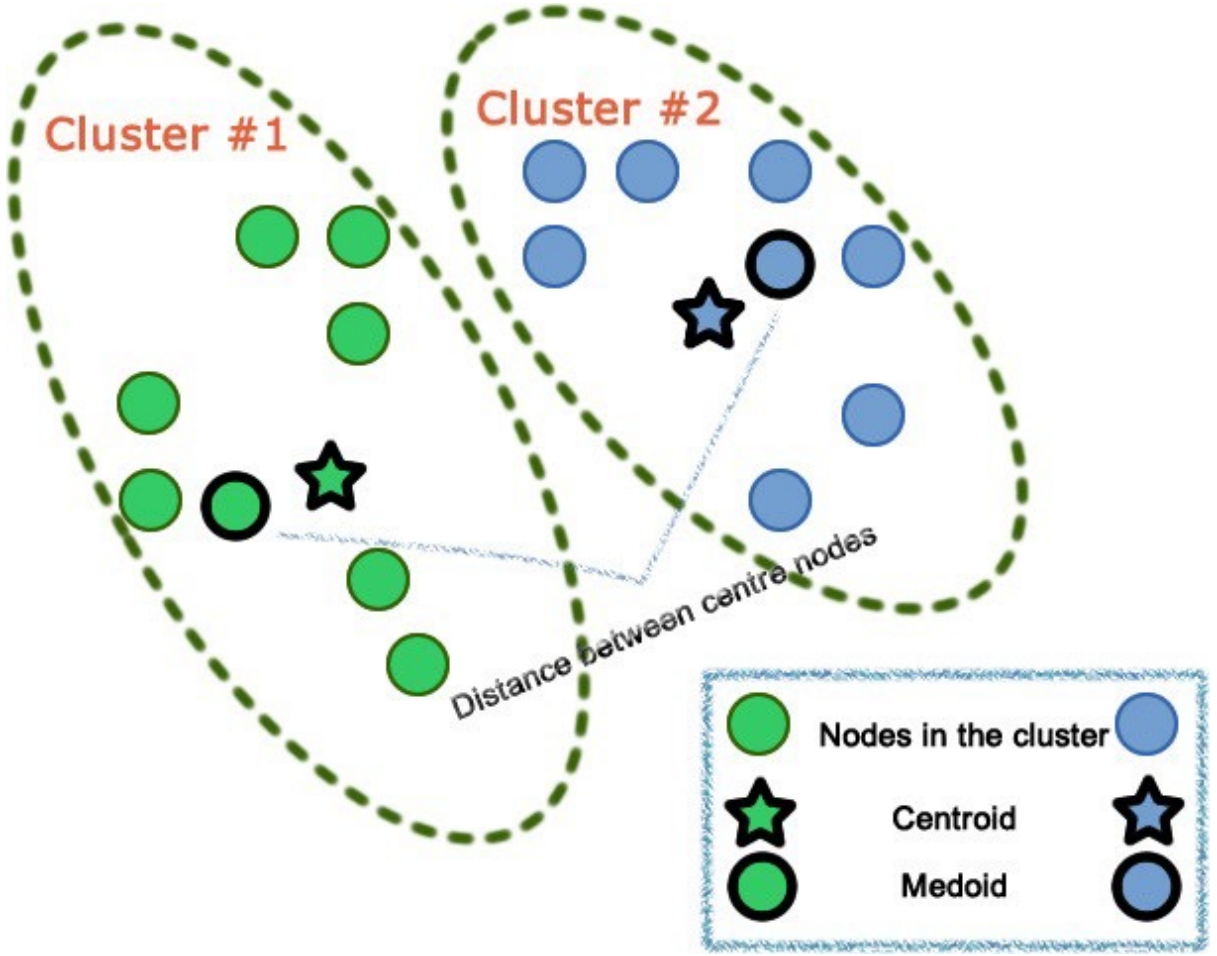
1. Introducción



2. Definiciones y representación de patrones

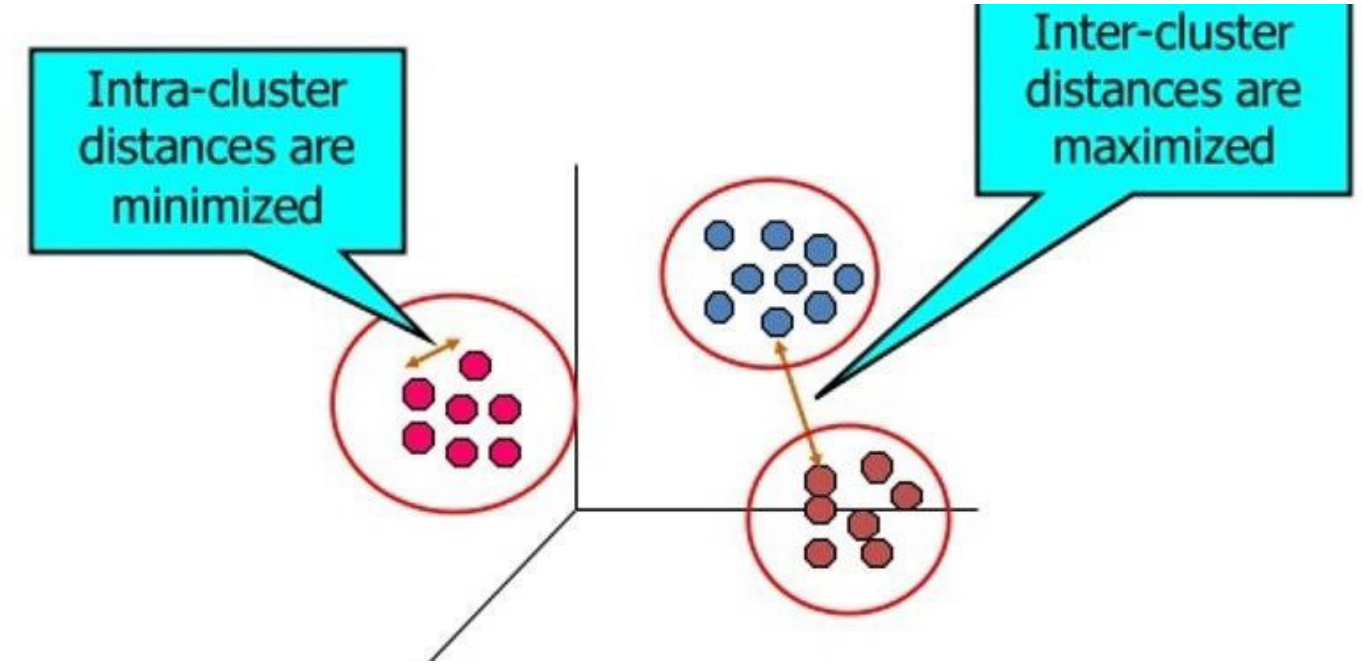
- **Clustering**: Proceso de agrupación de las instancias en un dataset.
- **Clúster**: Conjunto de instancias agrupadas tras el proceso de clustering.
- **Centroide**: Punto de dimensión *d* que identifica el centro de gravedad de un clúster: $C = \{ c_1 , ..., c_d \}$
- **Función de distancia**: Métrica que cuantifica la similitud (atributos discretos) o proximidad (atributos continuos) entre ejemplos (o entre clusters).

2. Definiciones y representación de patrones



2. Definiciones y representación de patrones

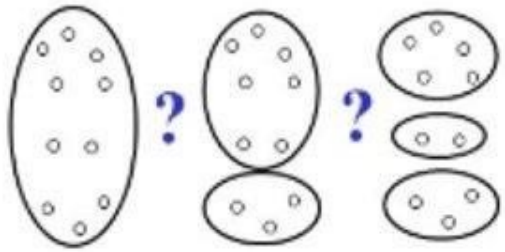
¿Qué persigue un buen clustering?



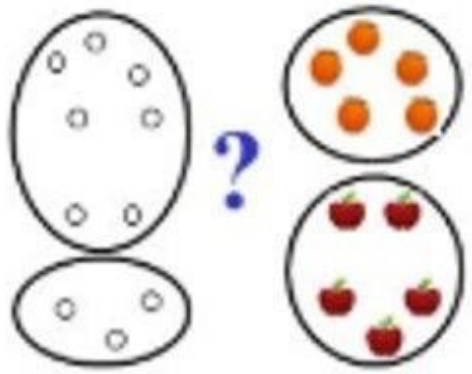
3. Validación en clustering

MEDIDAS DE BONDAD

1. **Índices internos**
    1. Validan sin información externa, es decir, sólo usando los datos
    2. Se usan para elegir el mejor número de clusters



2. **Índices externos**
    1. Se valida contra el “ground truth”
    2. Se usan para obtener el mejor algoritmo de clustering



3. Validación en clustering

MEDIDAS DE BONDAD **INTERNAS**

1. Cohesión (intra-cluster) (minimizar)

- Mide cómo de cerca están los puntos de los clusters con respecto a su centroide:

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

m_i: centroide
m: centroide más cercano

2. Separación (inter-cluster) (maximizar)

- Mide cómo de bien separados están los clústeres entre sí (distancias desde su centroide al centroide más cercano):

$$BSS = \sum_i |C_i| (m - m_i)^2$$

3. Validación en clustering

MEDIDAS DE BONDAD **INTERNAS**

3. Índice Silhouette

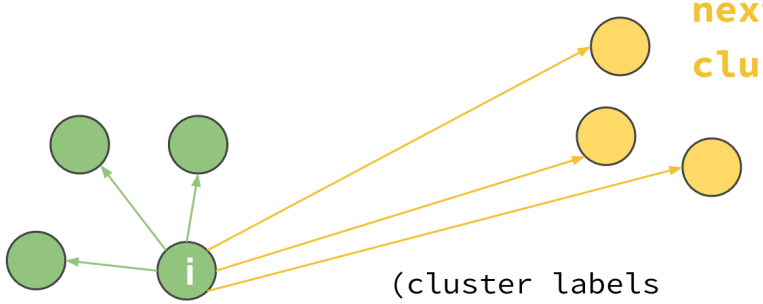
1. Combina tanto las ideas de cohesión *a(x)* como de separación *b(x)*
2. De forma matemática, para un punto *x*, se define como:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

3. Validación en clustering

SILHOUETTE INDEX

b(i): distance between i and its **next nearest cluster centroid**



(cluster labels assigned from ground-truth classes)

a(i): average distance between i and all of the other points in **its own cluster**

For a single point, i

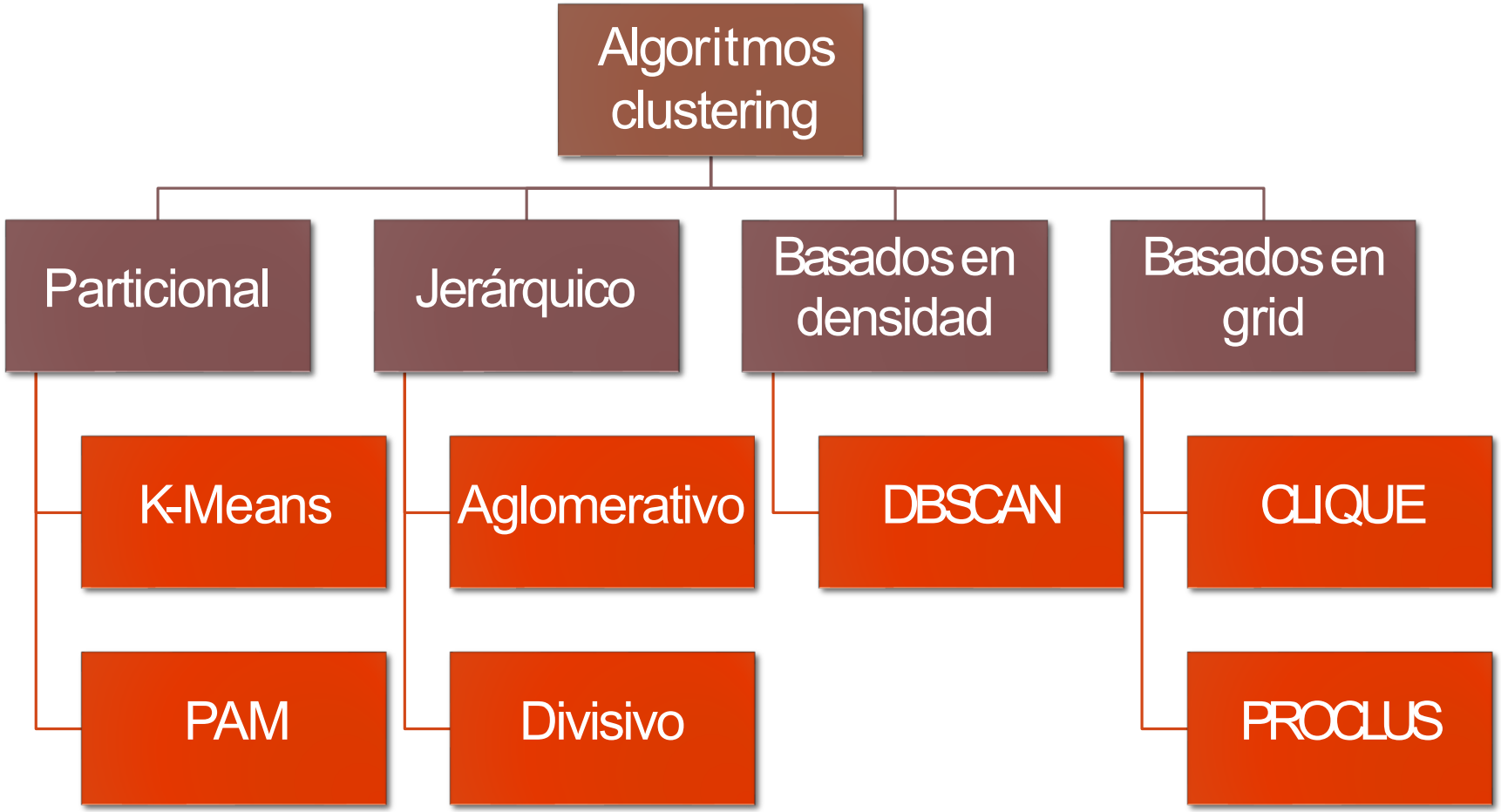
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

Sprawling, overlapped clusters

Tight, well-separated clusters

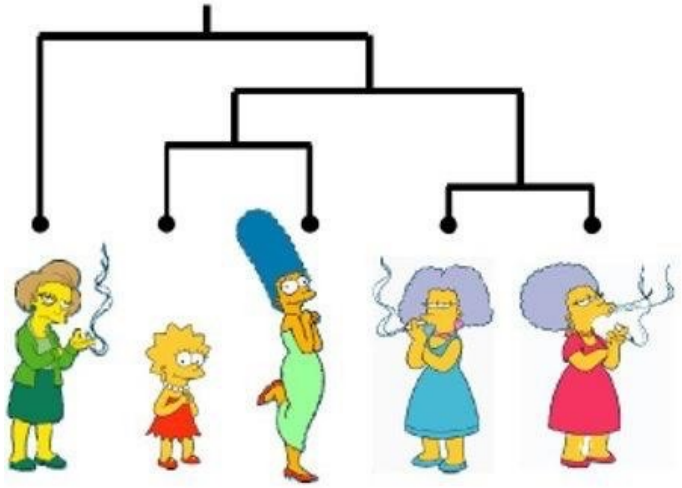
4. Algoritmos de clustering



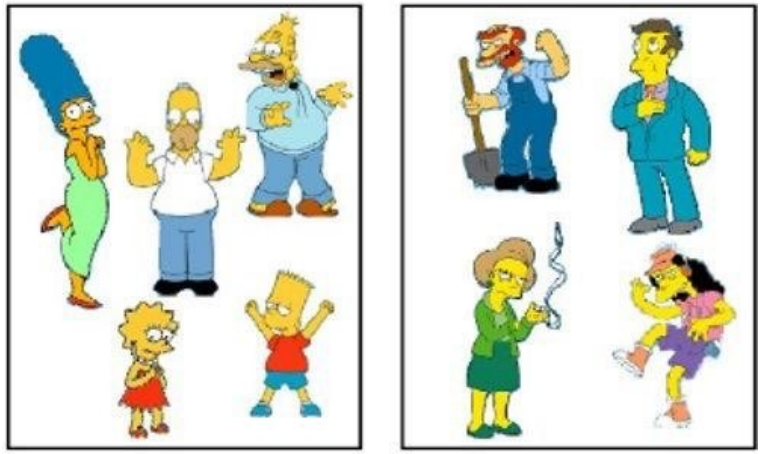
4. Algoritmos de clustering

JERÁRQUICO VS PARTICIONAL

Clustering Jerárquico



Clustering de Partición



4. Algoritmos de clustering: particional (1/11)

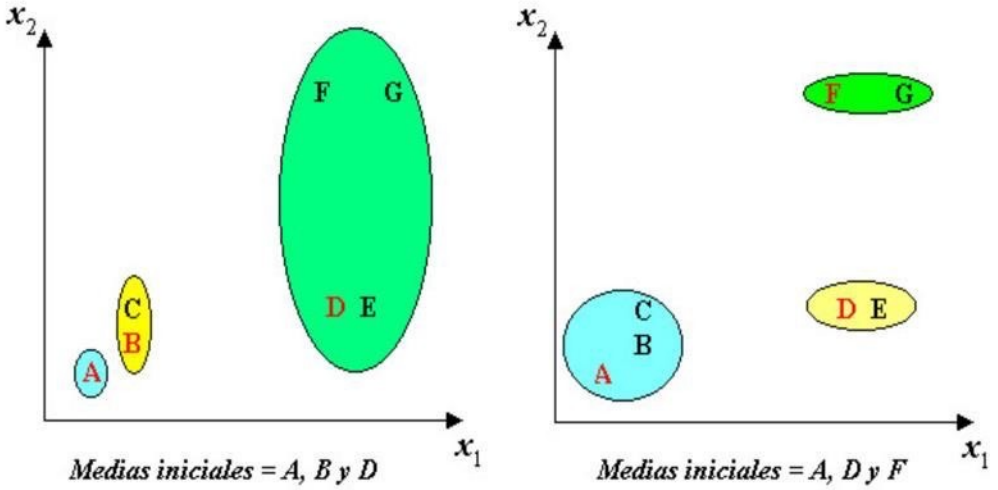
- Producen una única partición de los datos en lugar de una estructura
- Hay que definir un criterio de partición
- Es necesario establecer el número de clusters final
- Los algoritmos más conocidos son:
  - Error cuadrático (square error) → K-medias (k-means)
  - Grafo teórico (Graph-Theoretic)
  - Mixture resolving →Expectation Maximization
  - Mode seeking

4. Algoritmos de clustering: particional (2/11)

- K-MEDIAS
  - Comienza con una partición inicial aleatoria, reasignando los patrones a clusters basándose en la similitud de los patrones y el centro de los clusters, hasta que se alcanza el criterio de convergencia

- **Ventajas**: Sencillo de implementar y baja complejidad en tiempo O(n) (n = nº patrones)

- **Inconveniente**: Es sensible a la partición inicial (puede converger a un mínimo local del criterio)



Partición óptima = {{A,B,C} {D,E} {F,G}}

4. Algoritmos de clustering: particional (3/11)

- K-MEDIAS

Seleccionar una K centros de clusters *

Repetir:

Asigna cada ejemplo al cluster cuyo centro sea el más cercano

Calcula los nuevos centros de gravedad

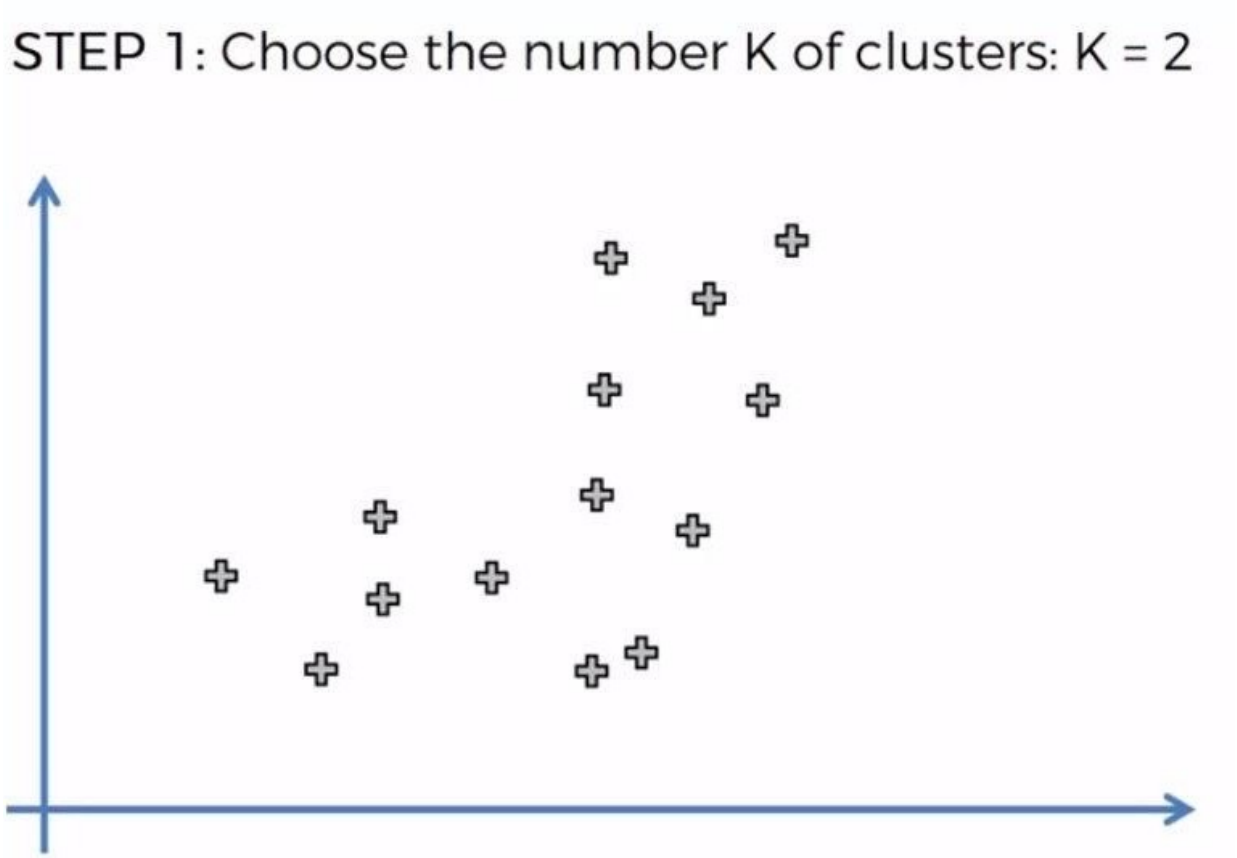
hasta que se cumpla el criterio de convergencia **

(*) Pueden coincidir con K patrones seleccionados al azar,o bien con K puntos aleatorios definidos dentro del hipercubo que contiene el sistema de patrones

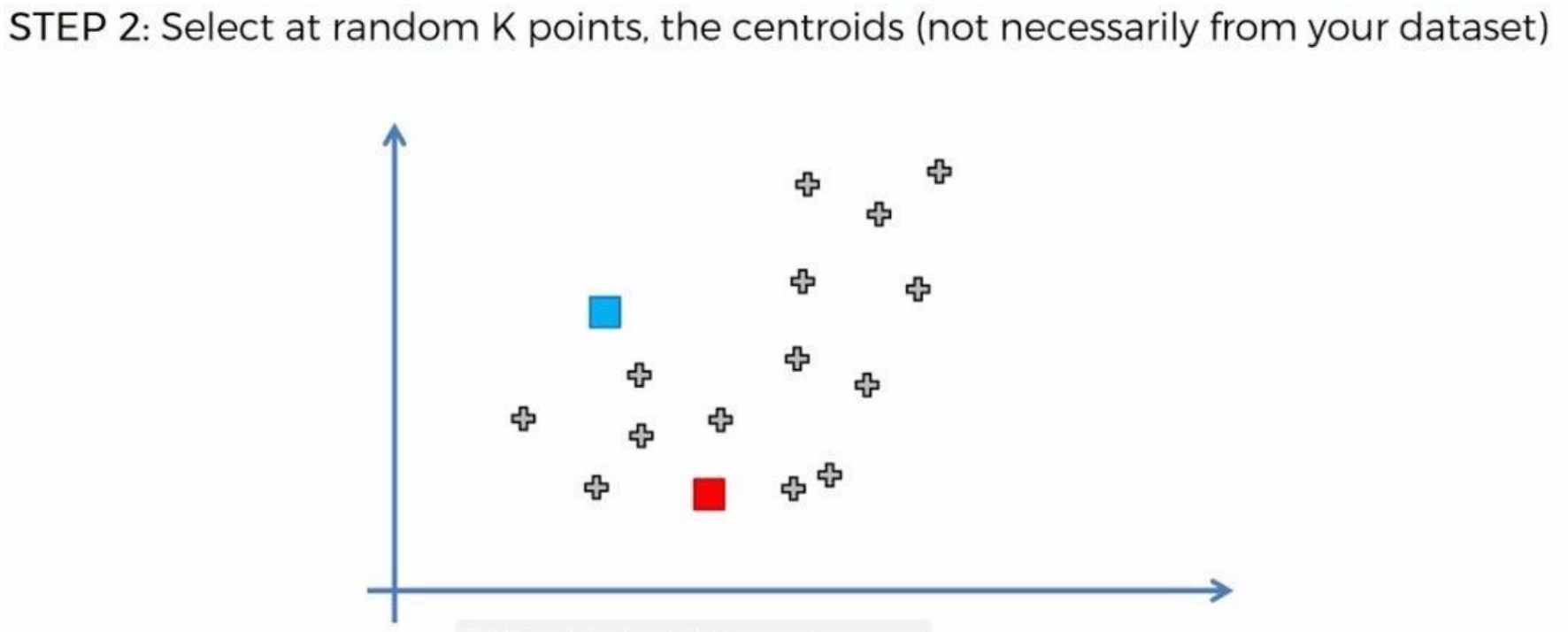
(**) Los criterios de convergencia típicos son:

- No reasignación o reasignación mínima de patrones a los nuevos clusters
- Mínima disminución del error cuadrático

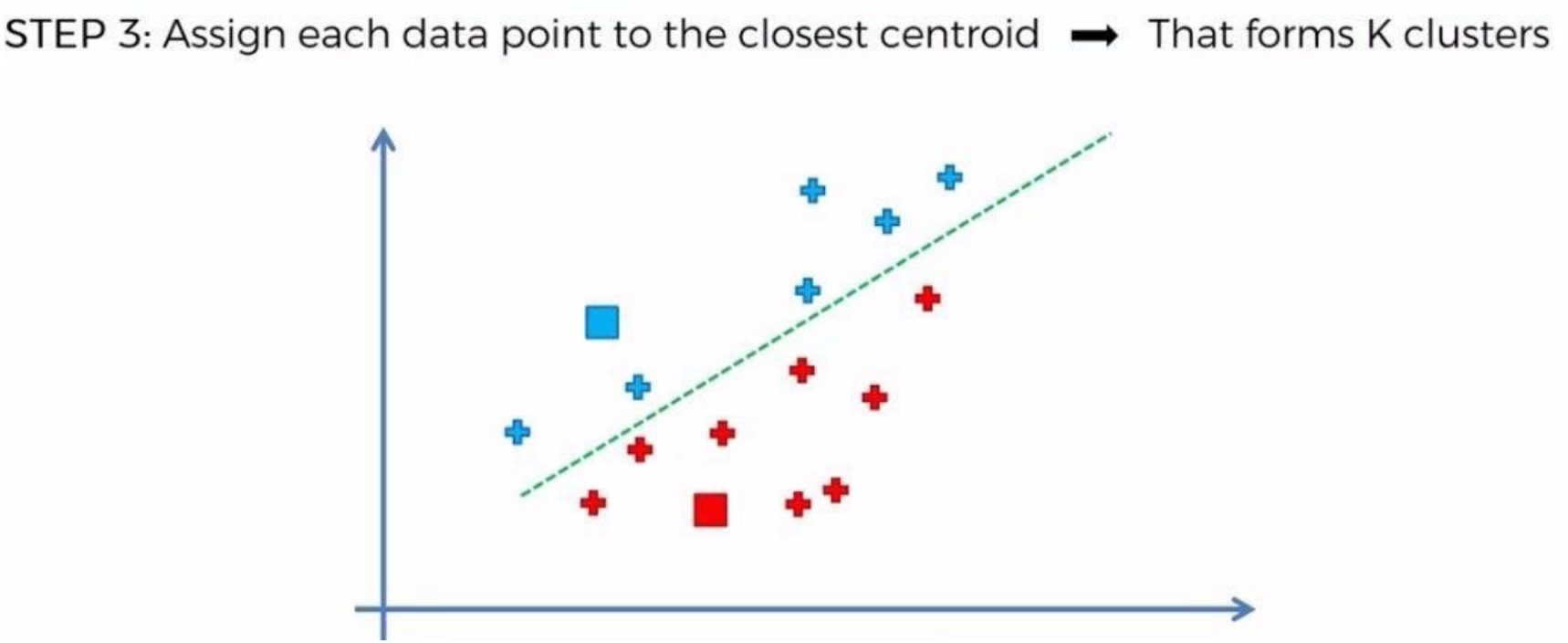
4. Algoritmos de clustering: particional (4/11)



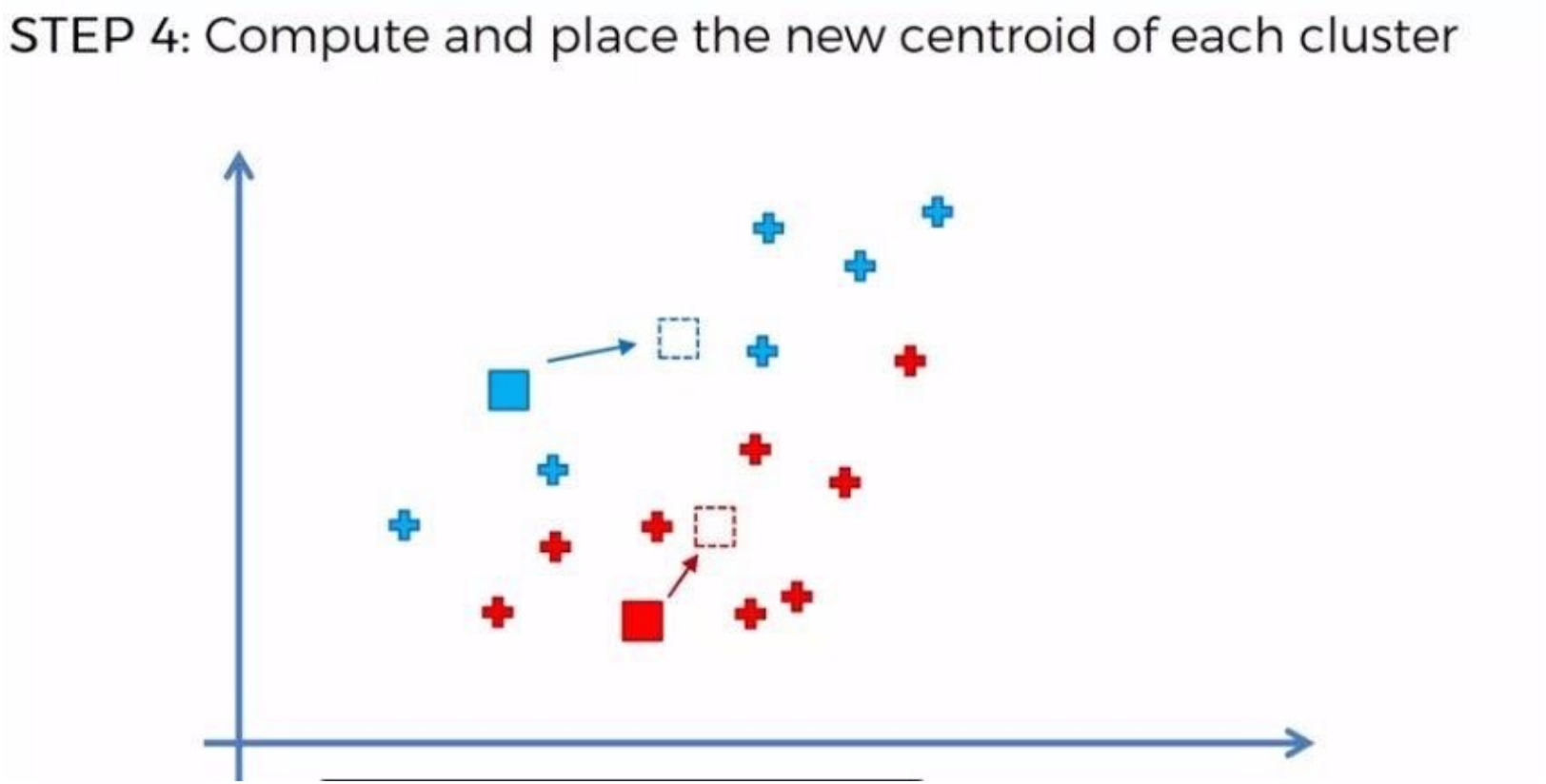
4. Algoritmos de clustering: particional (5/11)



4. Algoritmos de clustering: particional (6/11)

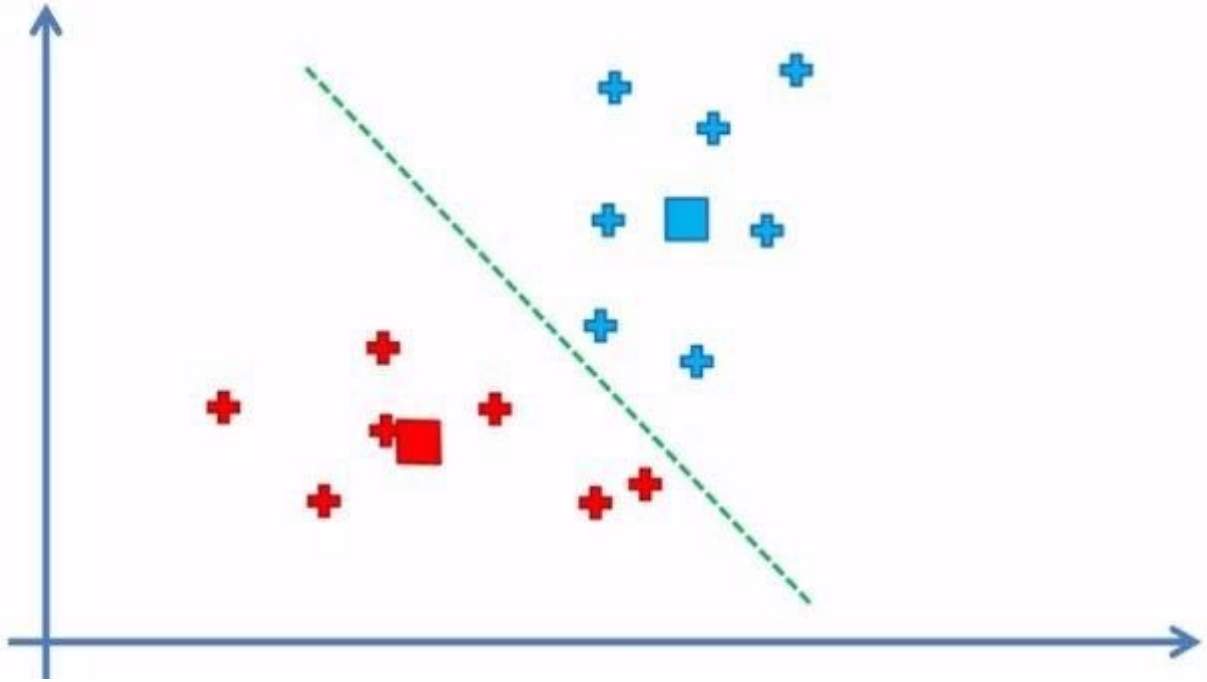


4. Algoritmos de clustering: particional (7/11)



4. Algoritmos de clustering: particional (8/11)

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



4. Algoritmos de clustering: particional (9/11)

La pregunta del millón…

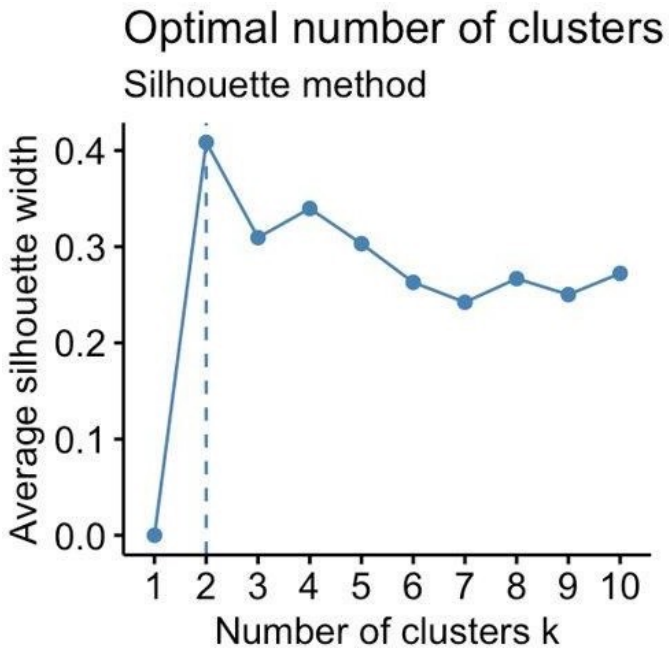
¿Qué valor de K?

4. Algoritmos de clustering: particional (10/11)

Método Silhouette

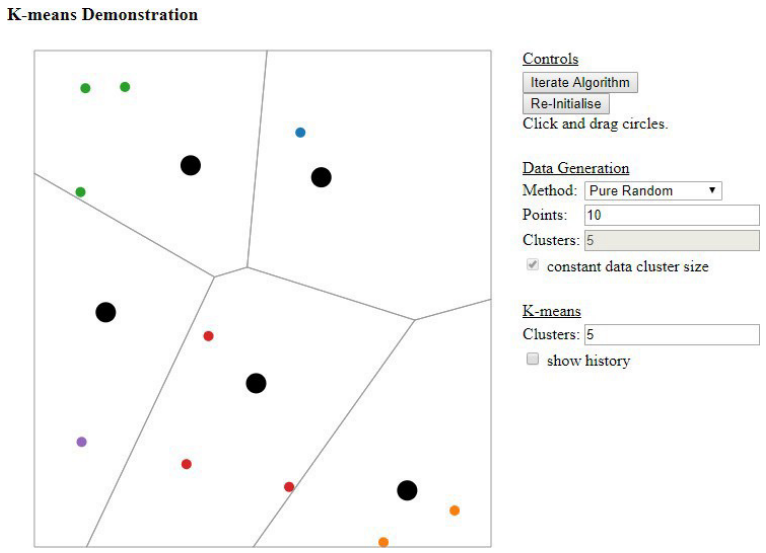
Pasos:

1. Ejecutar el algoritmo para varios valores de k
2. Para cada k calcular la media de los índices silhouette de las observaciones
3. Representar la curva obtenida
4. El mejor número de k es el que proporciona un máximo en la gráfica



4. Algoritmos de clustering: particional (11/11)

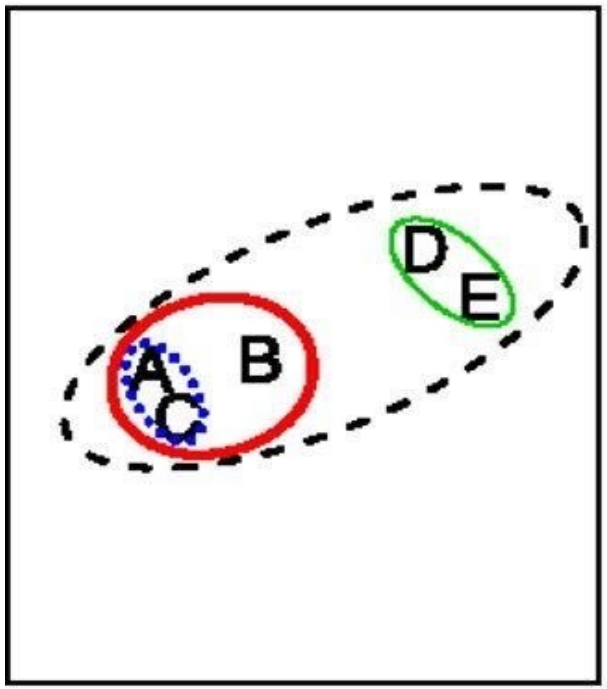
Ejemplo interactivo

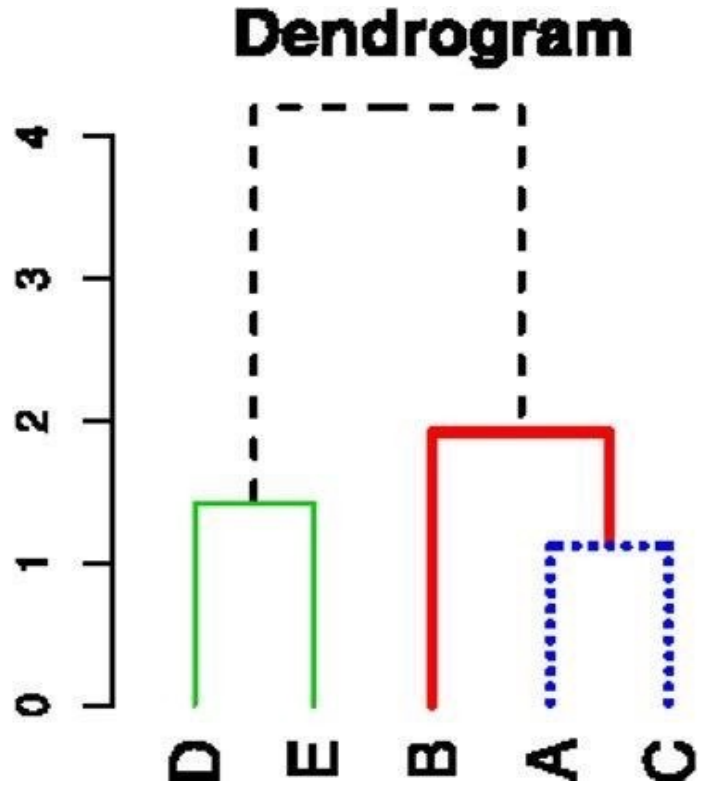


http://alekseynp.com/viz/k-means.html

4. Algoritmos de clustering: jerárquico (1/4)

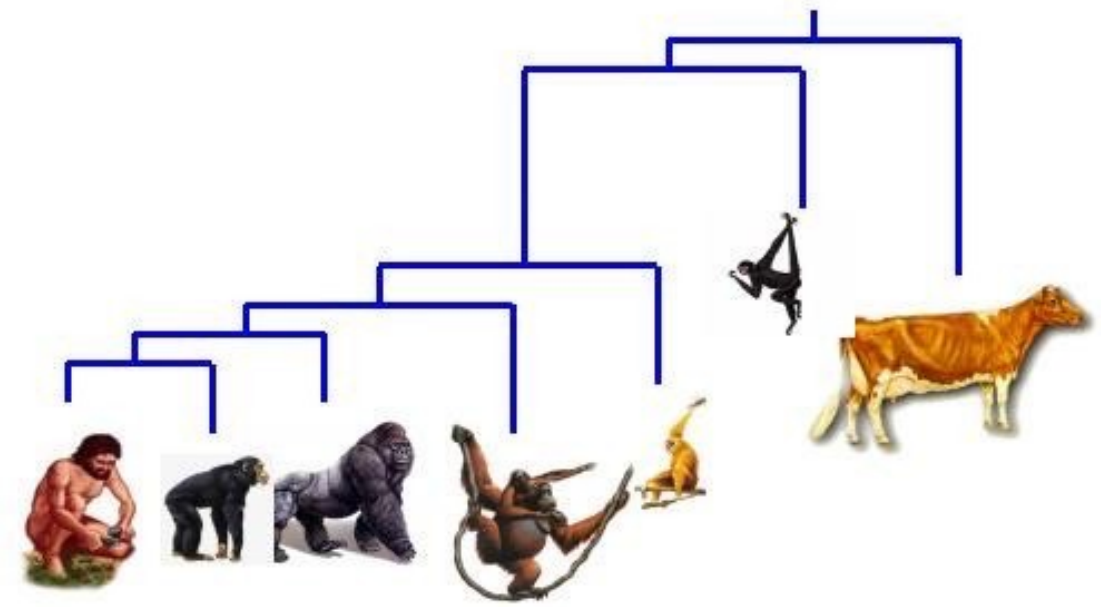
- Producen una serie jerarquizada de particiones que dependen de la similitud (distancia) entre los clusters/patrones
- Ejemplo:





4. Algoritmos de clustering: jerárquico (2/4)

Dendrograma

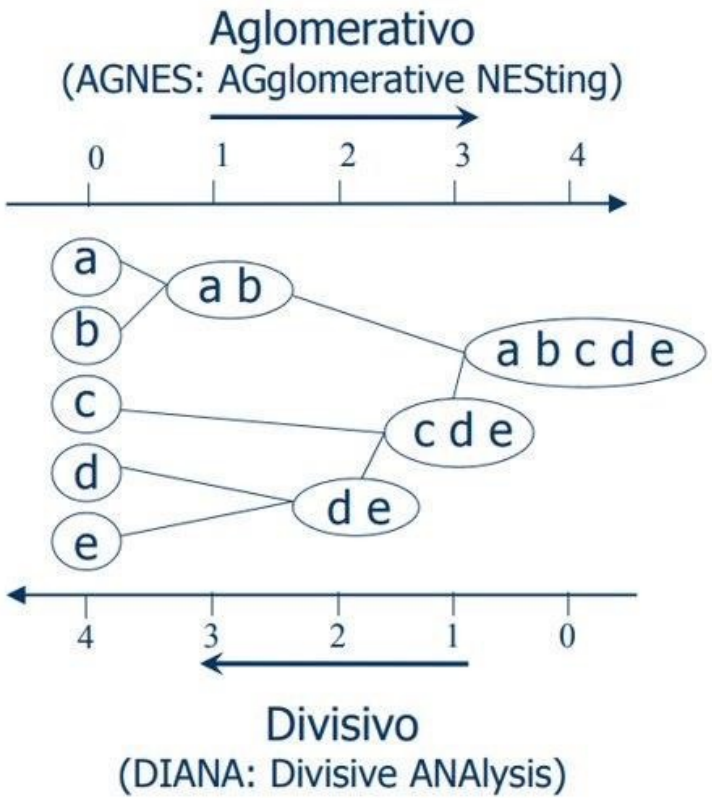


La similitud entre dos objetos viene dada por la "altura" del nodo común más cercano

4. Algoritmos de clustering: jerárquico (3/4)

- Según la forma en la que se vayan agrupando las instancias se pueden definir dos tipos.

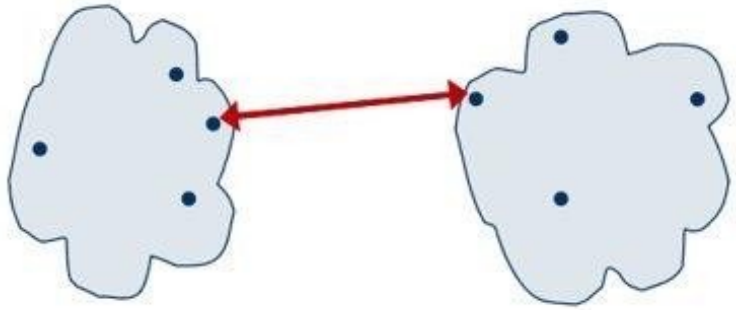
- En ambos métodos hay que tener en cuenta:
  - **Distancia** usada para medir la similitud
  - Criterio de **enlace**



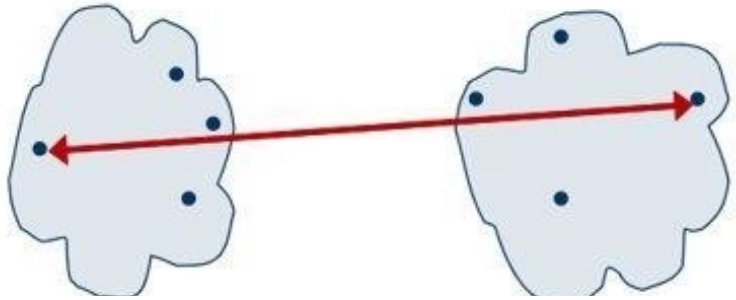
4. Algoritmos de clustering: jerárquico (4/4)

Criterio de enlace:

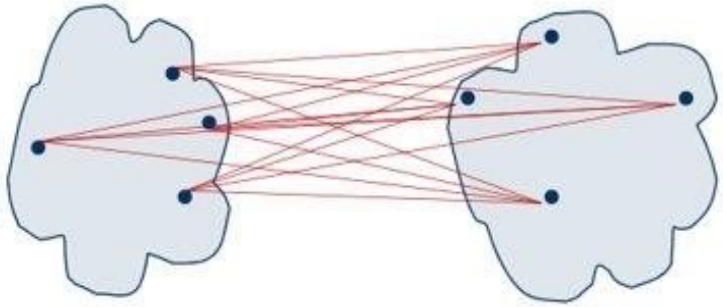
- MIN
  single-link



- MAX
  complete
  linkage
  (diameter)



- Promedio



4. Clustering jerárquico vs particional

Jerárquico	Particional
✓Mayor versatilidad	✓ Menor complejidad en tiempo y espacio
✗ Mayor complejidad en tiempo y espacio	✗ Necesita un número de clusters final
✓ Una vez obtenido el árbol, se corta por el nivel deseado	

5. Actividad 1 – Análisis clustering k-Means (1/3)

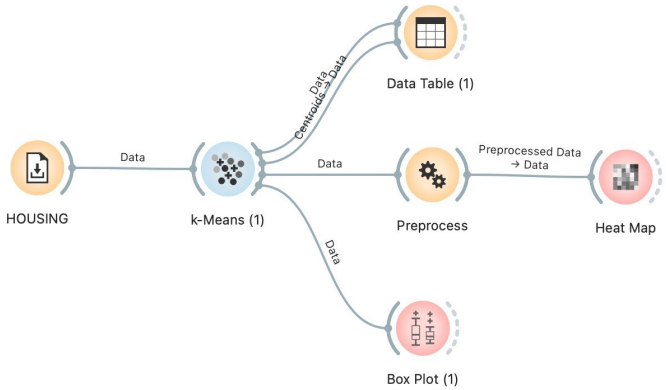
- **Objetivo**
  - Realizar agrupamientos de observaciones en dos conjuntos de datos:
    - Atributos/clase cualitativos: **tennis.csv** (nodo File)
    - Atributos/clase numéricos: **Housing** (nodo Datasets)
- **Procedimiento**
  1. Cargue el conjunto de datos **tennis.csv**.
  2. Aplique el nodo **k-Means** de clustering a los datos usando k desde 2 hasta 10.
  3. Varíe el número de clusters manualmente a partir de la tabla de índices Silhouette.
  4. Observe la distribución de los índices Silhouette de cada instancia con el nodo **Silhouette Plot**.
  5. Conecte un nodo **Data Table** para observar tanto el dataset ampliado con los nuevos atributos "Cluster" y "Silhouette" añadidos, como la tabla de centroides.
  6. Visualice la distribución conjunta entre los atributos (incluida la clase) y el número de cluster usando el nodo **Box Plot** (o el nodo **Distributions**).

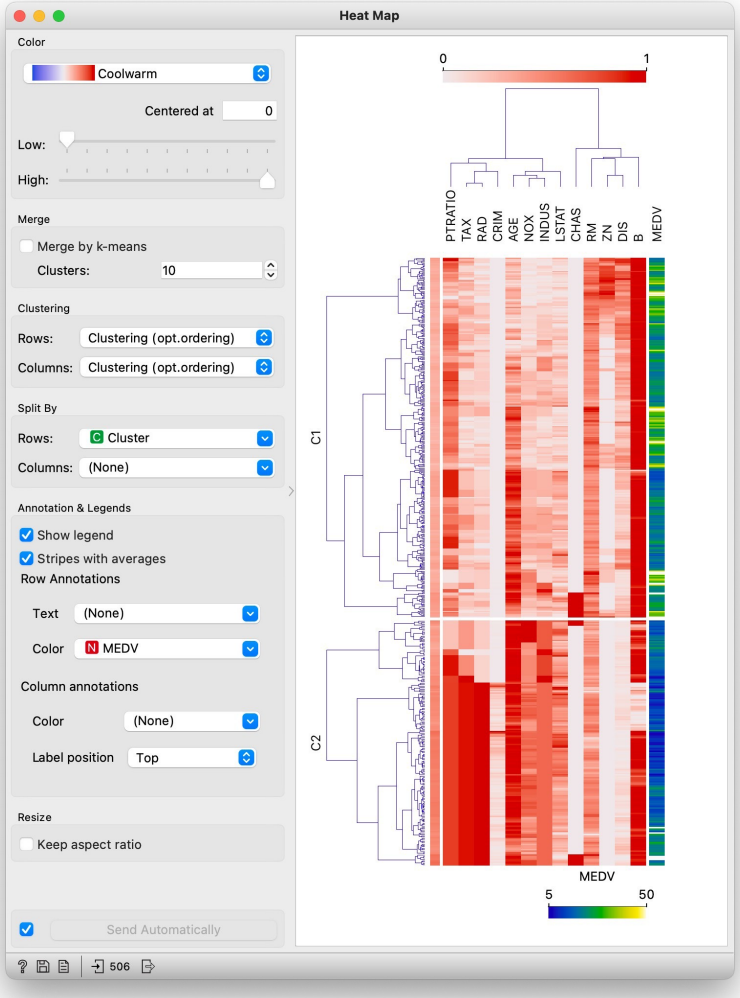
5. Actividad 1 – Análisis clustering k-Means (2/3)

- **Objetivo**
  - Realizar agrupamientos de observaciones en dos conjuntos de datos:
    - Atributos/clase cualitativos: **tennis.csv** (nodo File)
    - Atributos/clase numéricos: **Housing** (nodo Datasets)
- **Procedimiento**
  7. Cargue el conjunto de datos **Housing**.
  8. Aplique el algoritmo k-Means de clustering a los datos usando k desde 2 hasta 16.
  9. Fije un número de clusters manualmente a partir de la tabla de índices Silhouette.
  10. Conecte un nodo **Data Table** para observar tanto el dataset ampliado con los nuevos atributos "Cluster" y "Silhouette" añadidos, como la tabla de centroides.
  11. Visualice la distribución conjunta entre los atributos (incluida la clase) y el número de cluster usando el nodo **Box Plot** (o el nodo **Distributions**).

5. Actividad 1 – Análisis clustering k-Means (3/3)

- **Objetivo**
  - Agrupar observaciones en dos conjuntos de datos:
    - Atributos/clase cualitativos: **tennis.csv** (nodo File)
    - Atributos/clase numéricos: **Housing** (nodo Datasets)
- **Procedimiento**
  12. Normalice los datos entre 0 y 1.
  13. Visualice la distribución de los datos con Heat Map.





5. Actividad 2 – Silhouette y k-Means interactivo con datos sintéticos

- **Objetivos**
  - Analizar la distribución de los puntos según su valor de Silhouette
  - Observar la ejecución del algoritmo k-Means paso a paso
  - Generar datos sintéticos, tanto en grupos separados como mezclados
- **Procedimiento**
  1. Dibuje tres grupos de puntos conglomerados y **separados entre sí**.
  2. Aplique el nodo **Interactive k-Means** a los datos usando 3 centroides. Ejecute paso a paso el algoritmo hasta que converja.
  3. Conecte la salida de **Interactive k-Means** a un **Data Table** y a un nodo **Box Plot**, para analizar los resultados.
  4. Aplique el nodo **k-Means** a los datos usando k = 3 clusters.
  5. Conecte un nodo **Scatter Plot** a la salida de k-Means y coloree los puntos en función del índice Silhouette.
  6. Repita los pasos anteriores para tres grupos de puntos conglomerados **mezclados en la misma región** del plano.