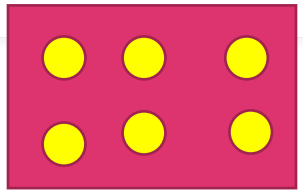
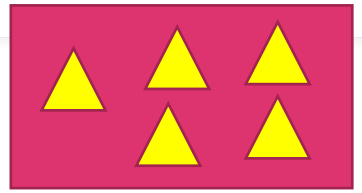
Análisis de Conglomerados

Dr. Misael Erikson
Maguiña Palma

Propósito de la segmentación





- **Encontrar grupos en que sus elementos sean:**
  - Lo más heterogéneos posible respecto a los elementos pertenecientes a los otros grupos.
  - Lo más homogéneos posible respecto a los elementos que pertenecen al mismo grupo.

- **Encontrar grupos:**
  - Significativos (tamaño justificable)
  - Alcanzables (accesibles para la compañía de acuerdo a sus recursos y experiencia)
  - Identificables (interpretables)

¿Cuándo y para qué usar Clustering?

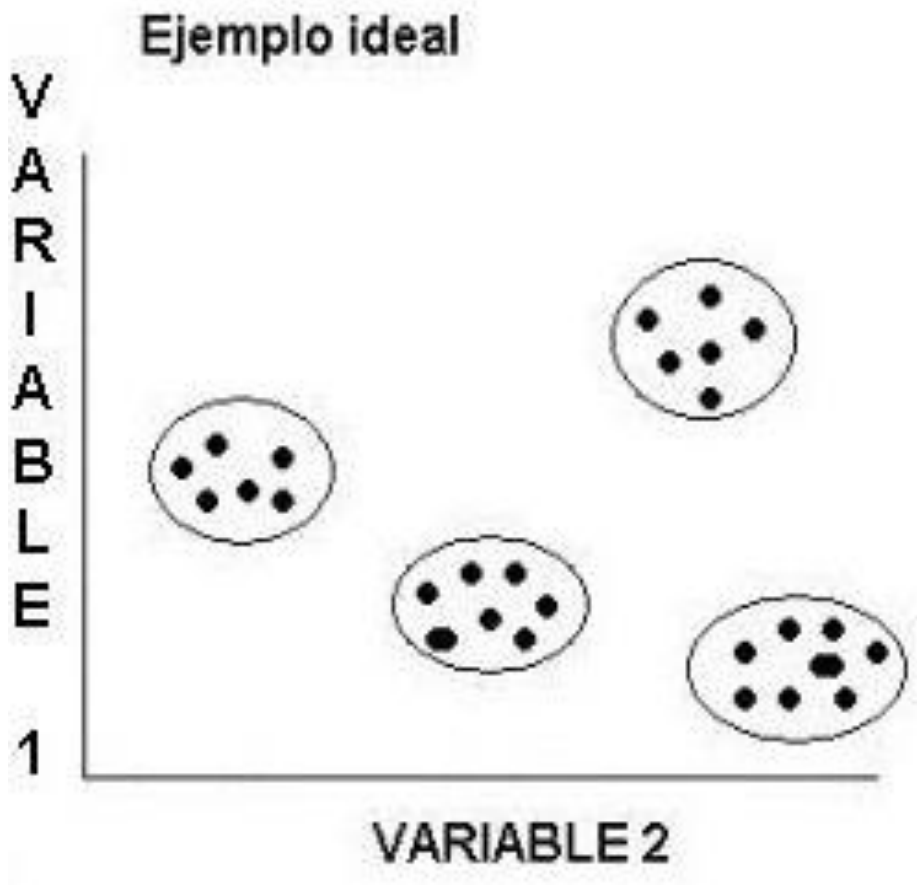
Cuando necesitemos dividir nuestros datos en grupos que sean:

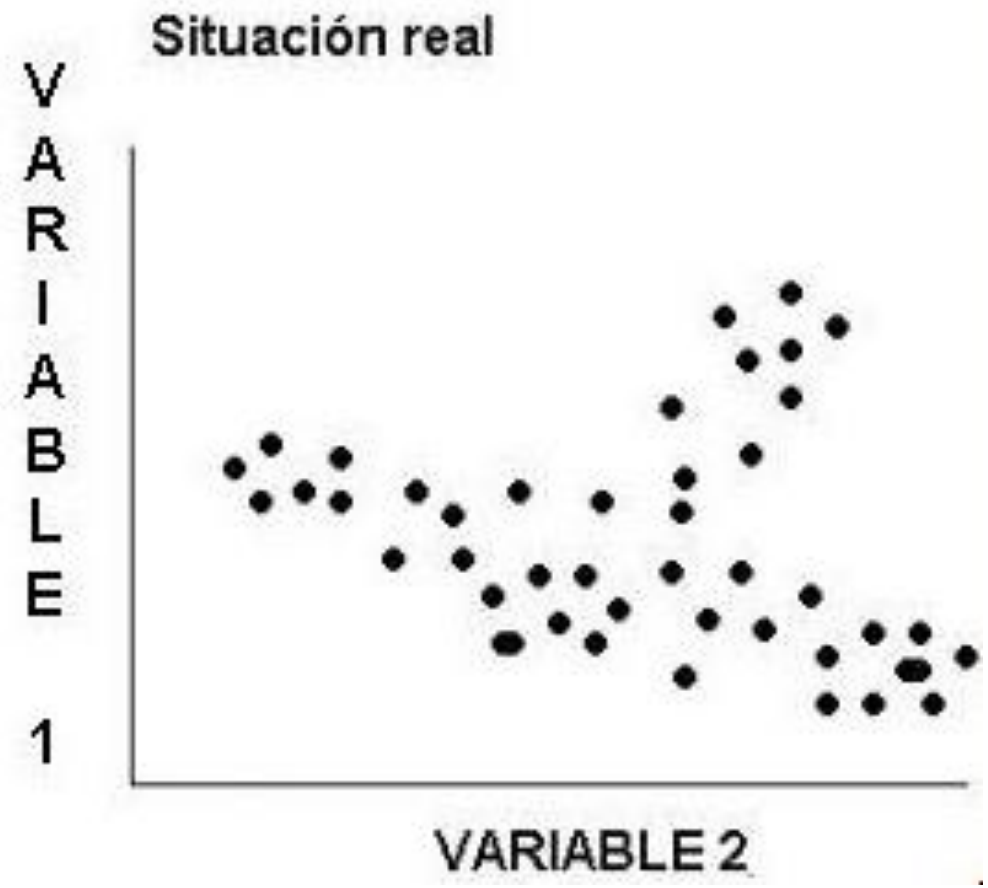
- Significativos y/o útiles
- Debemos preocuparnos de capturar la estructura natural de los datos

Clasificación vs Clustering?

- Clasificación: aprendizaje supervisado,
- Clustering: aprendizaje no-supervisado

Buscamos Capturar agrupaciones naturales en los datos





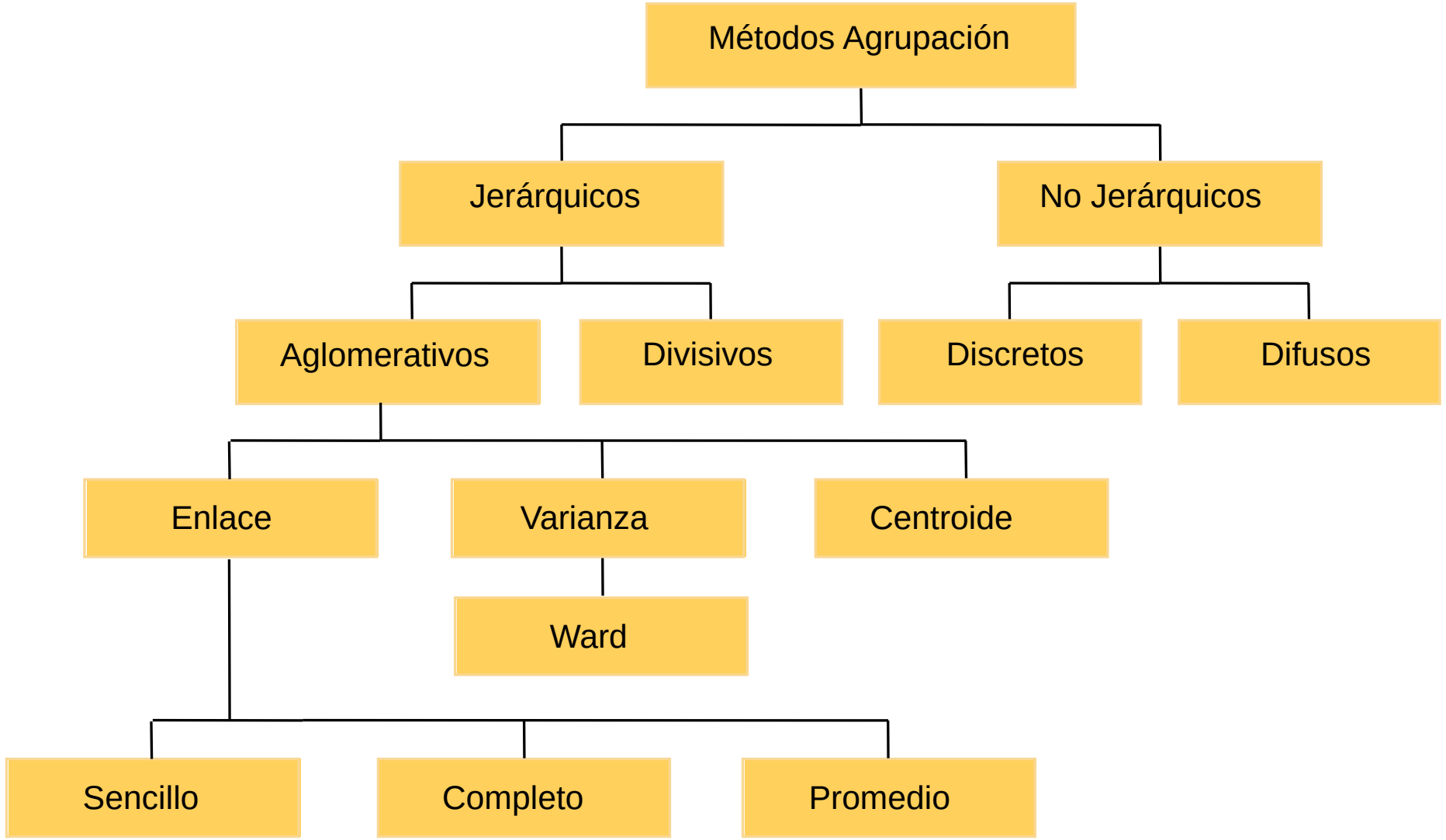
Análisis de clusters en una tarea esencial para muchas aplicaciones

Por ejemplo:

- Encontrar clusters naturales y describir sus propiedades (data understanding)
- Encontrar agrupamientos útiles (data class identificaction)
- Encontrar representantes para grupos homogéneos (data reduction)
- Encontrar objetos iniciales (outliers detection)
- Encontrar pertubaciones aleatorias de los datos (noise detection)

Métodos de asignación

Corresponden a la lógica en que los objetos se asignan a cada grupo.



Métodos de asignación

Existen dos grandes tipos de métodos de agrupación:

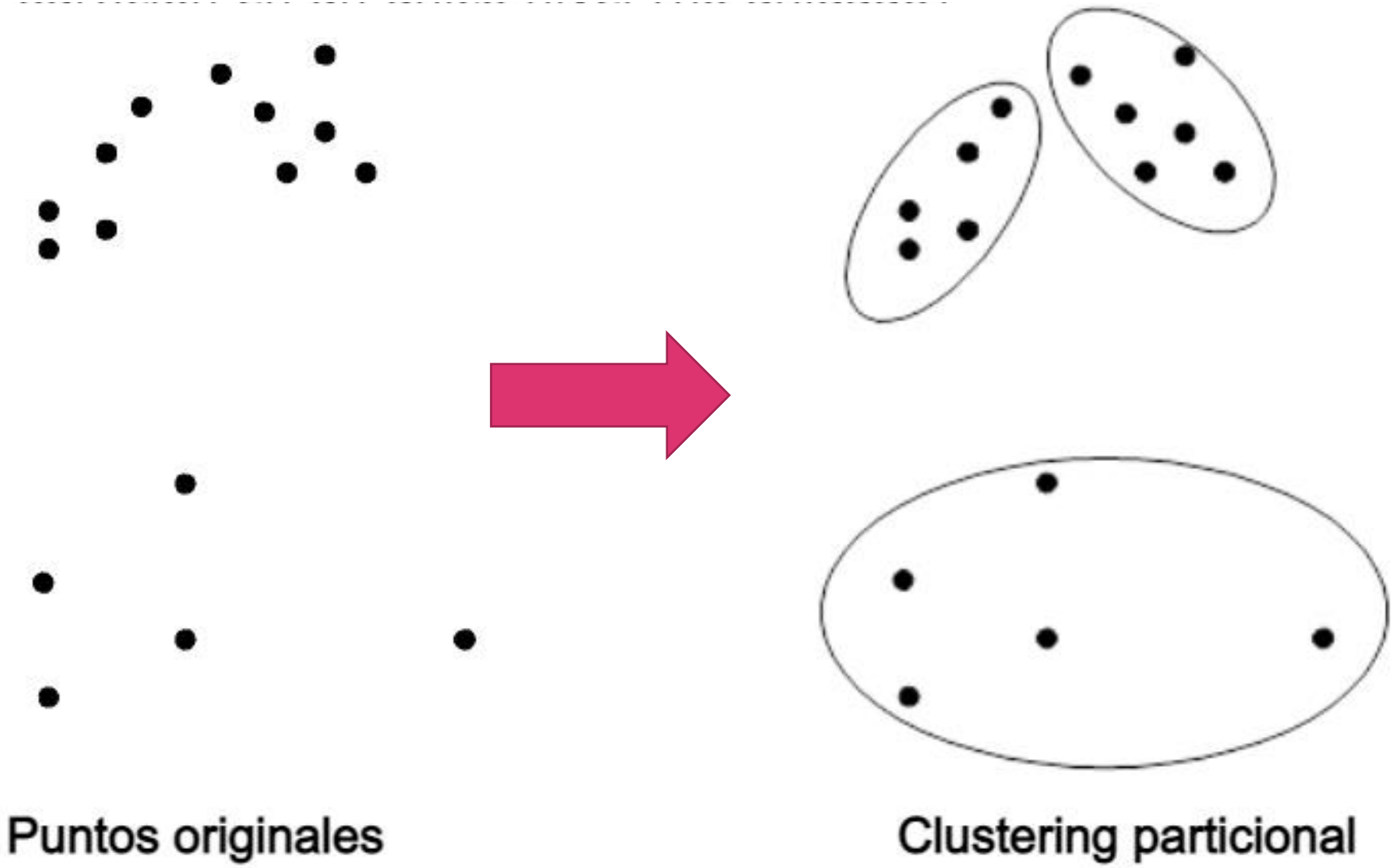
- **Métodos jerárquicos:**

  Los objetos se agrupan (dividen) por partes hasta clasificar todos los objetos.
  - De una iteración a otra, se modifica el valor de pertenencia a grupos de un único objeto.
  - No requiere a priori fijar un número de clusters.

- **Métodos no jerárquicos (de partición):**

  Se tiene un número de grupos predefinidos y cada objeto se ubica en un grupo hasta alcanzar estabilidad.
  - De una iteración a otra, se puede modificar el valor de pertenencia a grupos de todos los objetos.
  - Requiere a priori fijar un número de clusters.

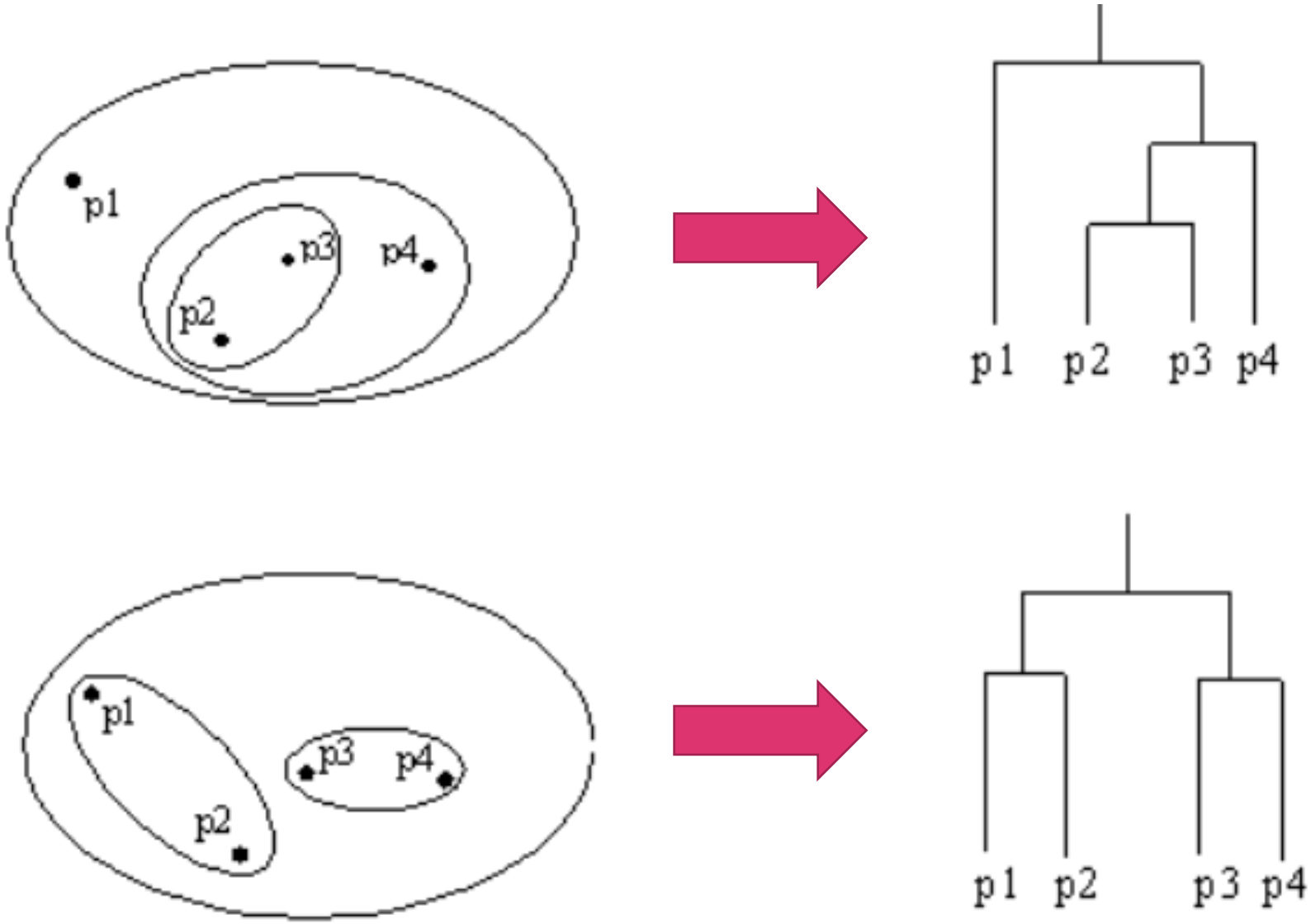
Clustering Particional



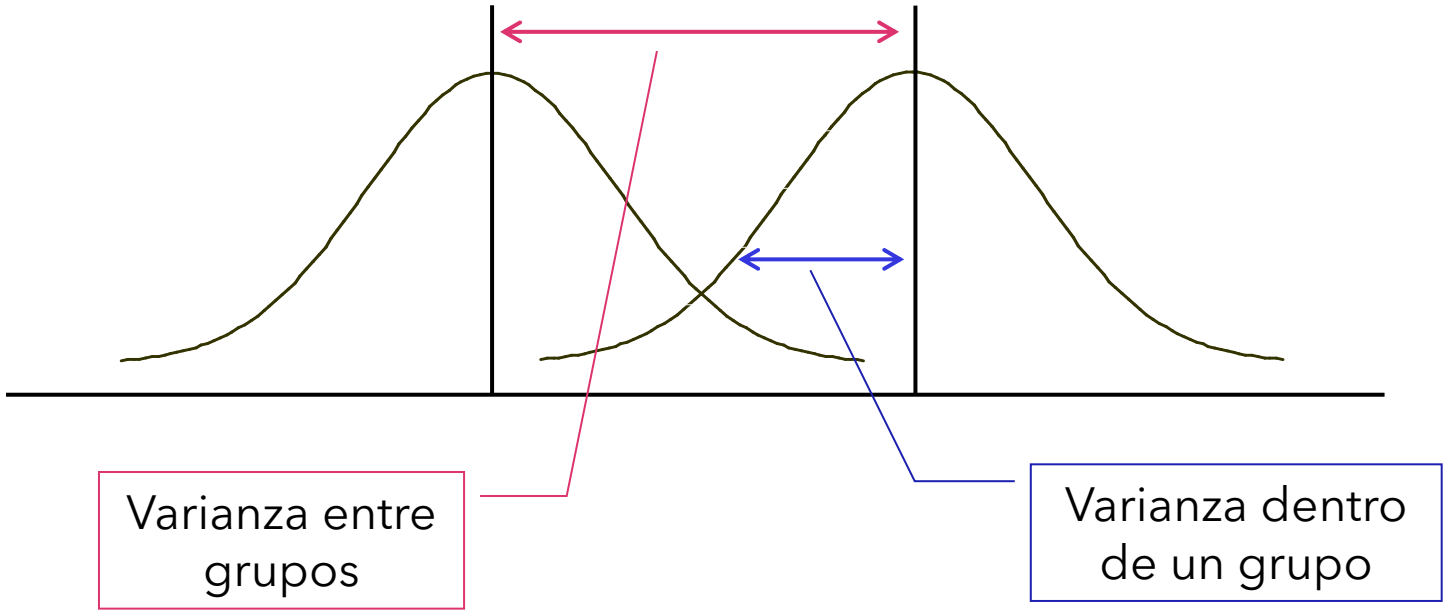
Puntos originales

Clustering particional

Clustering

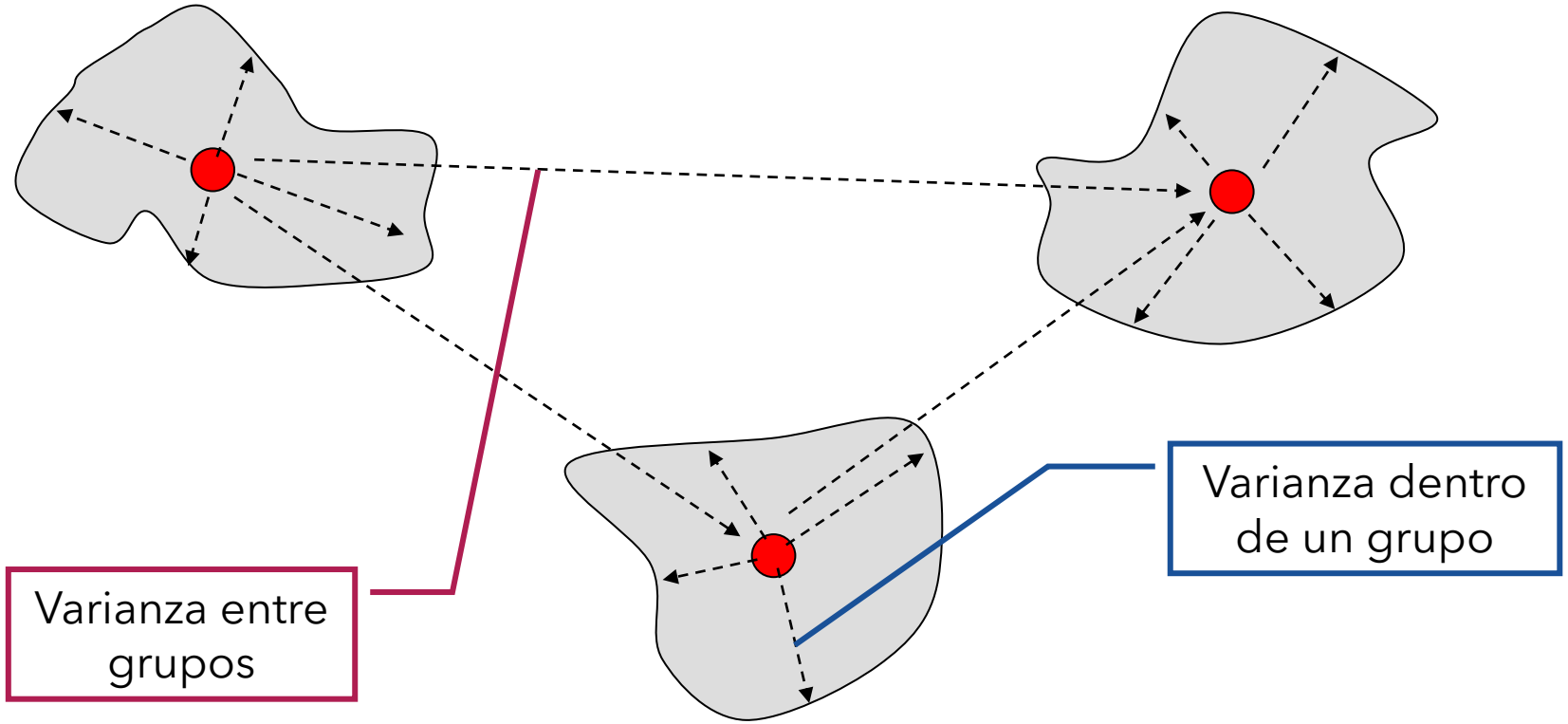


Gráficamente



$$\max_{c \in C} \left\{ \frac{\text{varianza entre grupos}}{\text{varianza en los grupos}} \right\} \quad C = \text{Conjunto de clusters posibles}$$

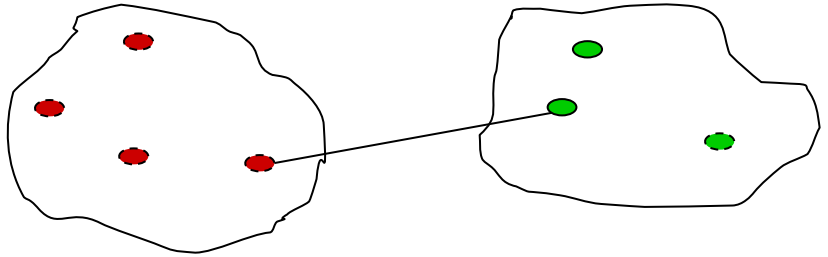
Gráficamente



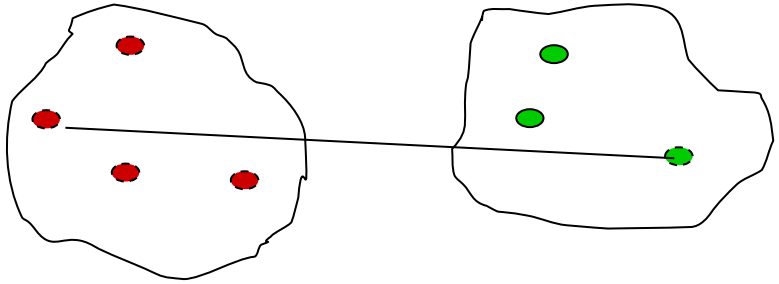
$$\max_{c \in C} \left\{ \frac{\text{varianza entre grupos}}{\text{varianza en los grupos}} \right\} \quad C = \text{Conjunto de clusters posibles}$$

Criterios de cercanía de enlace

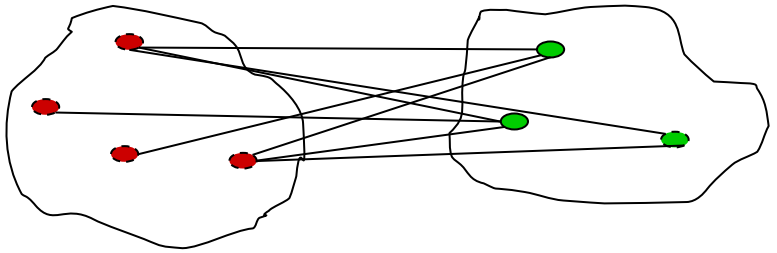
- **Enlace sencillo (vecino más cercano):** se juntan los grupos que presentan la mínima distancia entre objetos.



- **Enlace completo (vecino más lejano):** se juntan los grupos que presentan la mínima distancia entre los objetos más distantes del grupo.

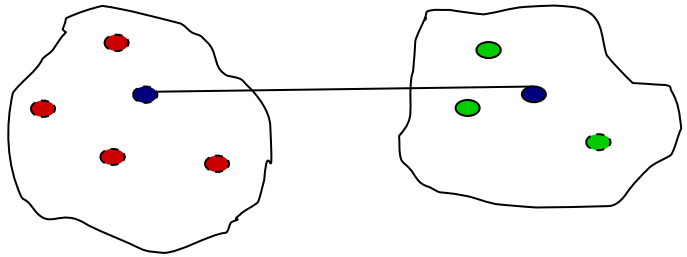


- **Enlace promedio (vecino promedio):** se juntan los grupos que presentan la mínima distancia promedio entre grupos.

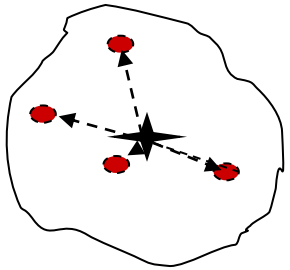


Criterios de cercanía del centroide y de la varianza (Ward)

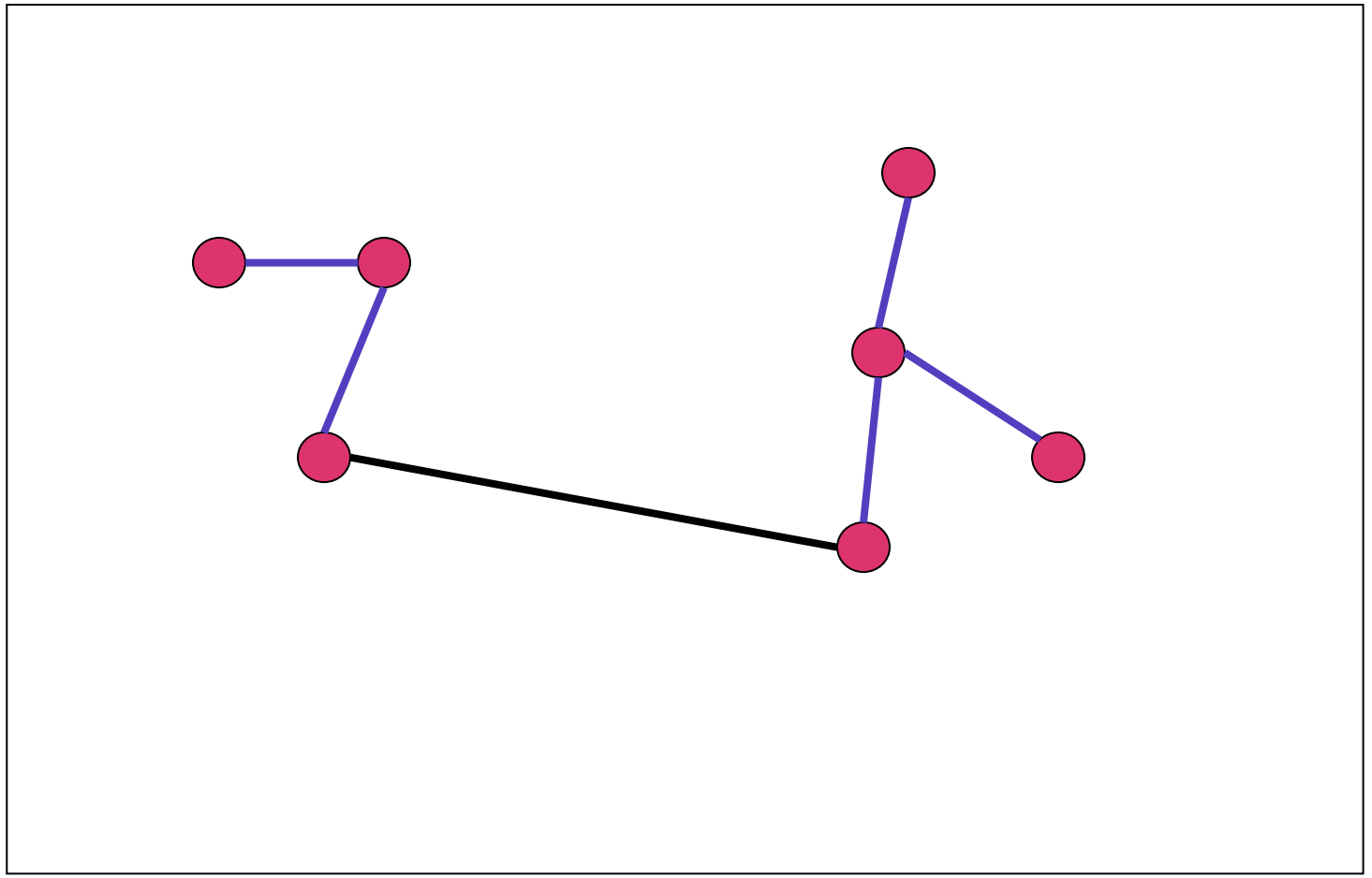
- **Método del centroide:** se juntan los grupos que presentan la mínima distancia entre sus centroides (medias para todas las variables).



- **Método de Ward:** se juntan los grupos que presentan la mínima varianza dentro de los grupos.
  - Para cada grupo, se calculan las medias para todas las variables.
  - Para cada objeto, se calcula la distancia euclideana cuadrada a las medias de los grupos.
  - Estas distancias se suman para todos los objetos del grupo.
  - Se combinan los grupos con el menor incremento en la suma total de los cuadrados de las distancias dentro de los conglomerados.



Ejemplo gráfico



7 clusters

6 clusters

5 clusters

4 clusters

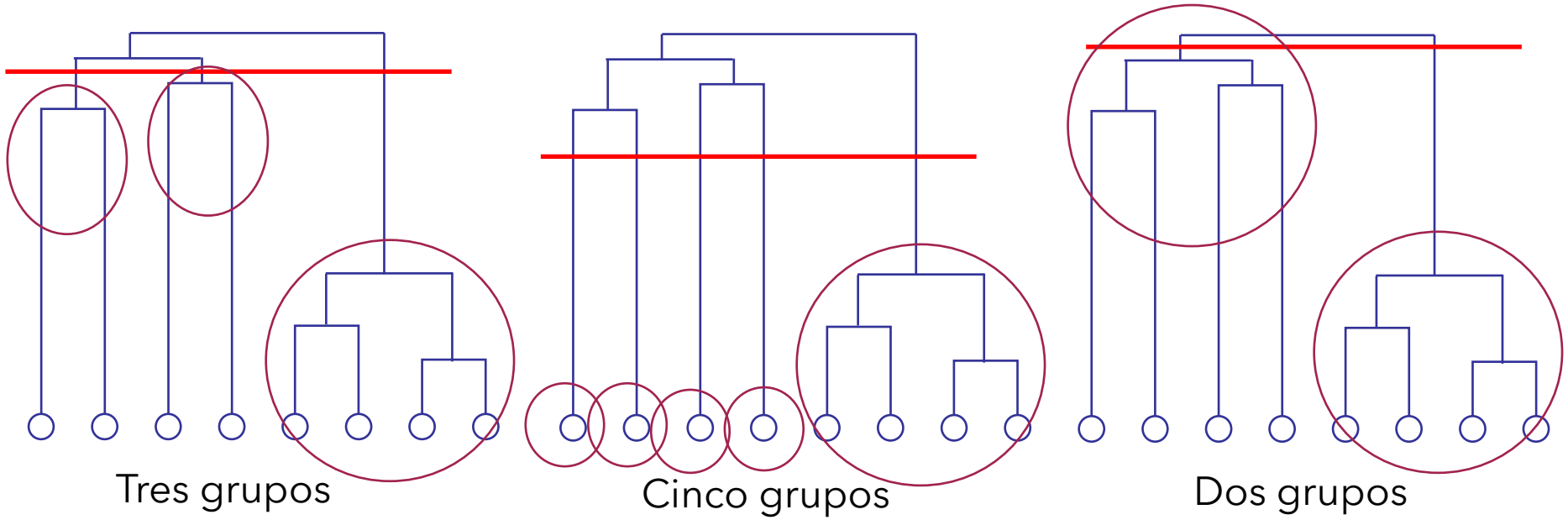
3 clusters

2 clusters

1 cluster

Dendogramas

Un dendograma es un árbol en el que el largo de las ramas está asociado inversamente a la fortaleza de la relación.

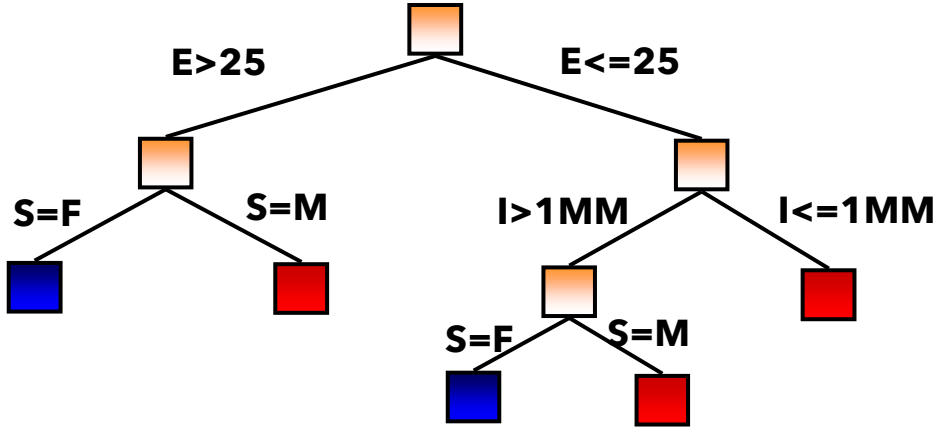


Métodos de asignación jerárquicos de tipo divisivo

- Esquema general algoritmo:
  - Todos los objetos pertenecen a un mismo grupo.
  - Cada grupo se separa bajo algún criterio de maximización de varianza entre grupos.
  - Se divide cada uno de los grupos hasta que:
    - Los objetos que forman parte de cada grupo son tan homogéneos que no vale la pena seguir dividiendo.
    - Los grupos son tan pequeños que no vale la pena seguir dividiendo (en el extremo se puede iterar hasta que queden tantos grupos como elementos).
- El proceso genera junto con la estructura de árbol, **reglas explícitas** de clasificación para los objetos.
- El método de árbol es un procedimiento que trata de clasificar casos, en base a un conjunto de variables independientes, de modo de discriminar mejor una variable dependiente.

Árboles de clasificación

- **Idea:** dada una clasificación de objetos se busca determinar una estructura jerárquica de reglas que permitan discriminar de la mejor manera posible entre las categorías.
- **Ejemplo:** en las bases de datos de una empresa se cuenta con una gran información de clientes (edad, ingreso, sexo) y se desea saber el perfil de las personas que serán receptivas a la realización de una campaña de promoción.



Receptivo

No Receptivo

Los hombres mayores de 25 no serán receptivos

Métodos de asignación no jerárquicos

- En cada iteración ubican a los objetos en el grupo más cercano a él.
- Se determina un número predefinido de grupos, en donde pueden ser asignados todos los casos.
- Es un tipo de asignación menos explicativa, pero poco sumamente eficiente cuando se trabaja con muchos casos.
- Distinguimos dos tipos de métodos de asignación no jerárquicos:
  - **Discretos:** cada objeto sólo puede pertenecer a un único grupo.
    - K-Means, Two Step Cluster.
  - **Difusos:** cada objeto tiene un grado de pertenencia a cada uno de los grupos.
    - Fuzzy C-Means, Clase Latente.

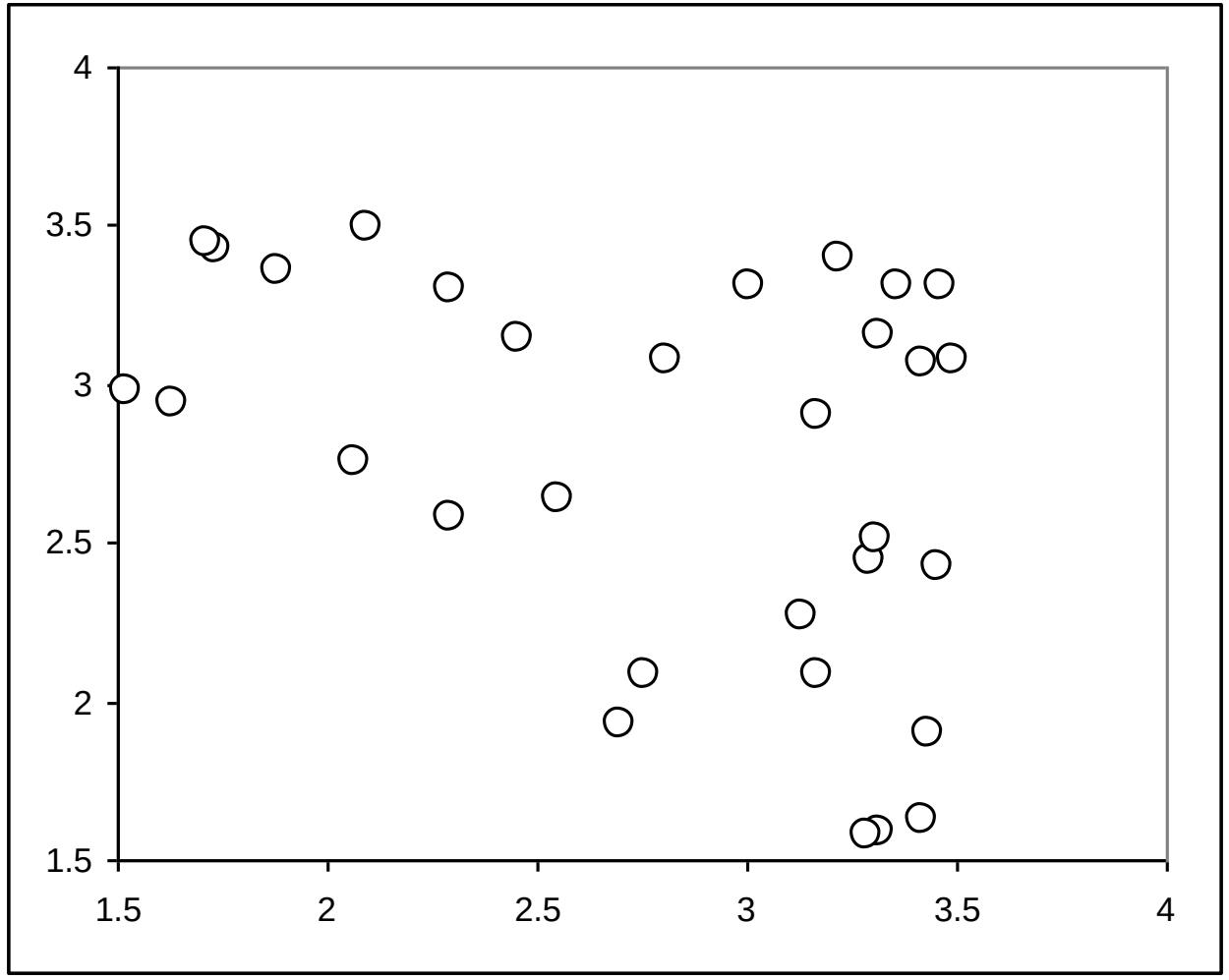
Métodos de asignación no jerárquicos de tipo discreto (K-Means)

- Es por lejos el método de asignación más utilizado en la **segmentación de mercados**.
- Es en este método donde centraremos nuestro estudio de la **agrupación no jerárquica de elementos**.
- Esquema general algoritmo:
  - Se tiene un conjunto de N objetos y K grupos.
  - Antes de la primera iteración se eligen arbitrariamente los centros de cada grupo.
  - En cada iteración se asigna cada objeto a su grupo más cercano y luego se recalculan los centros de cada grupo con los nuevos elementos asignados.
  - Iterar hasta que los cambios en los centros de cada grupo no sean significativos.
- El método entrega los elementos que pertenecen a cada grupo y los centros de cada grupo.

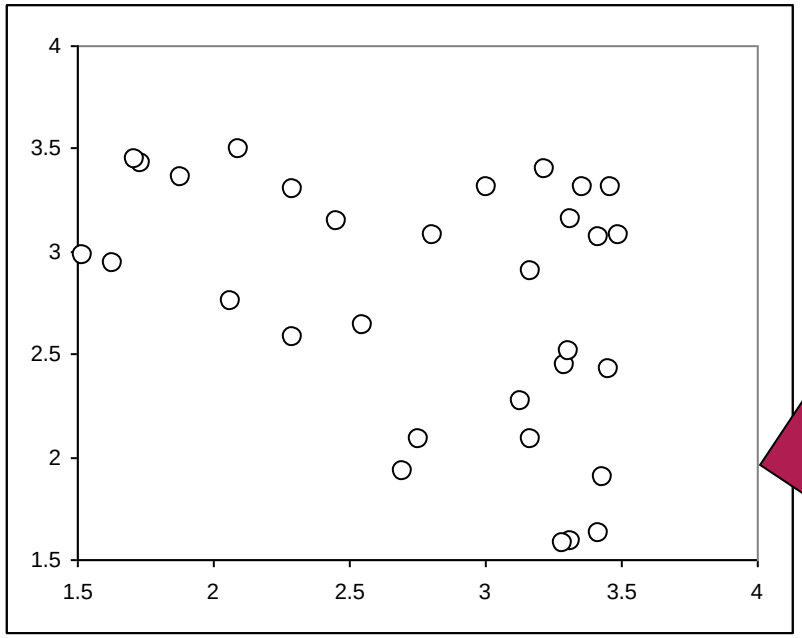
Métodos de asignación no jerárquicos de tipo discreto (K-Means)

- Veamos su formulación matemática:
  - Entradas
    - X conjunto de N objetos
    - K número de grupos
  - Salidas
    - S1,...,Sk K conjuntos
    - Z1,...,Zk Los centros de cada grupo
  - Inicialización.
    - t=0
    - Elegir arbitrariamente Zj(t).
  - Asignación y actualización de centros.
    - Asignar Xi al grupo mas cercano para todo i=1...N.
    - Recalcular Zj j=1...K
    - t=t+1
  - Criterio de parada.
    - Si Zi(t)-Zi(t+1) < e para todo i, parar.

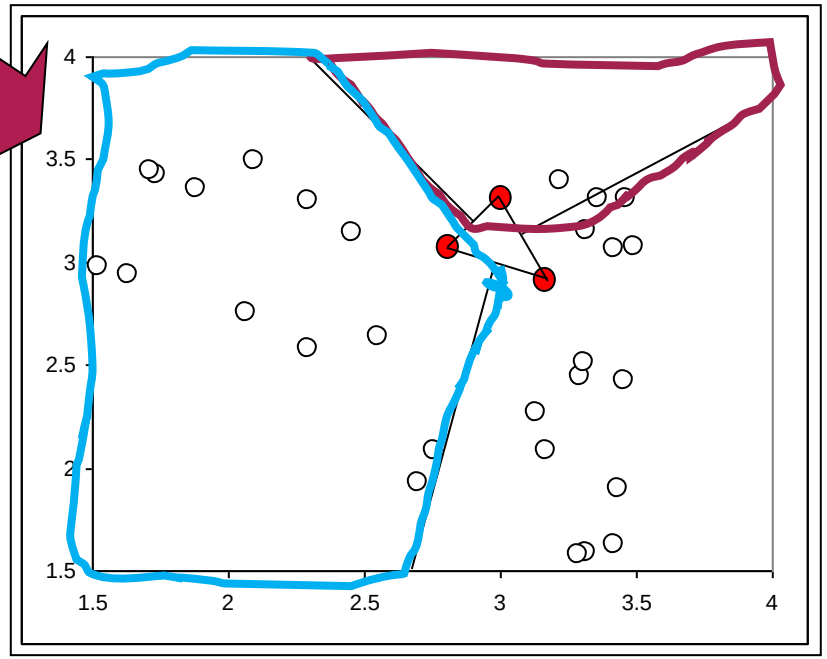
Ejemplo Gráfico



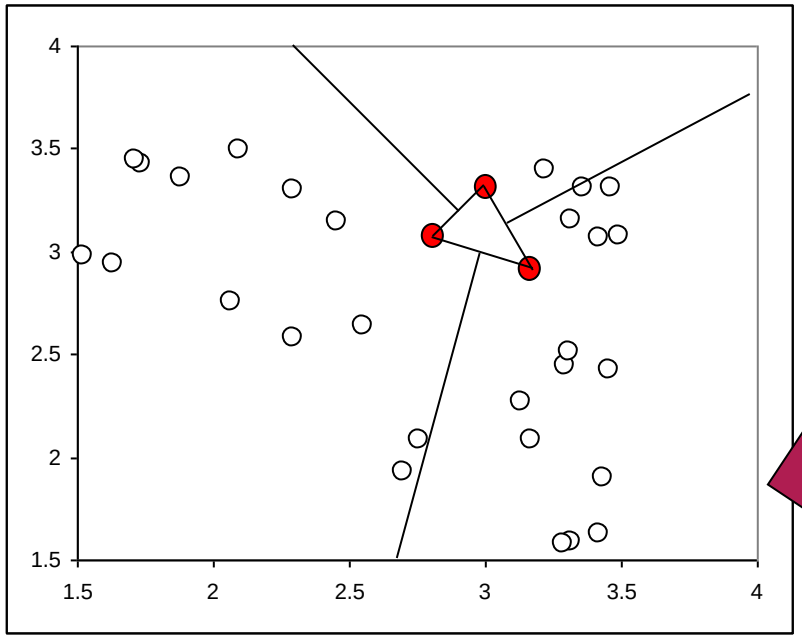
Iteración 1



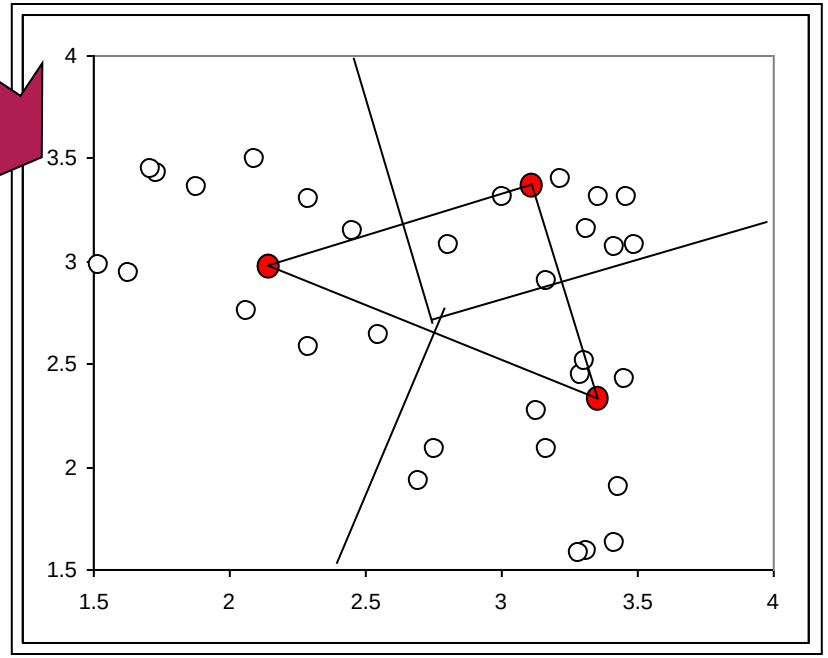
- Elegir los centros de los tres clusters aleatoriamente
- Localizar cada punto en su centro de cluster más cercano



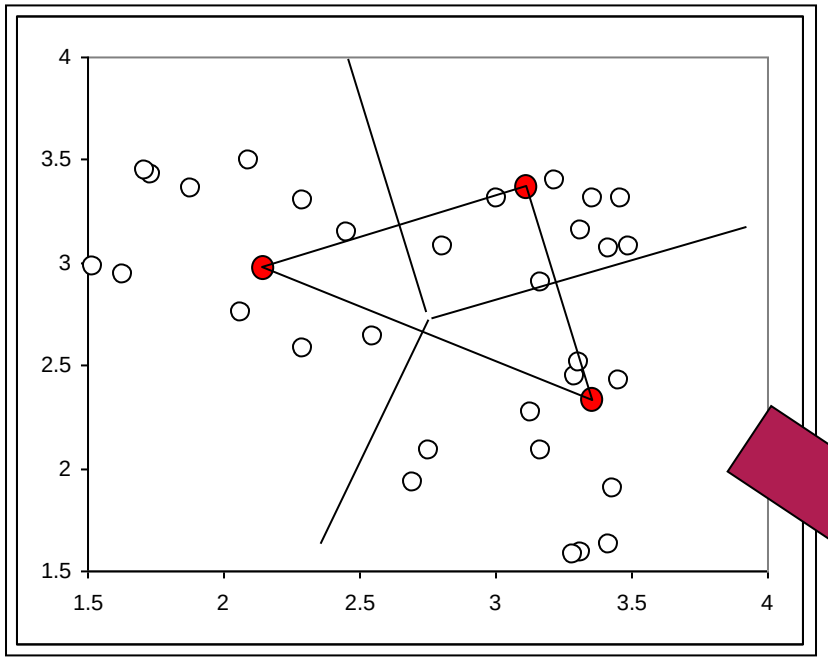
Iteración 2



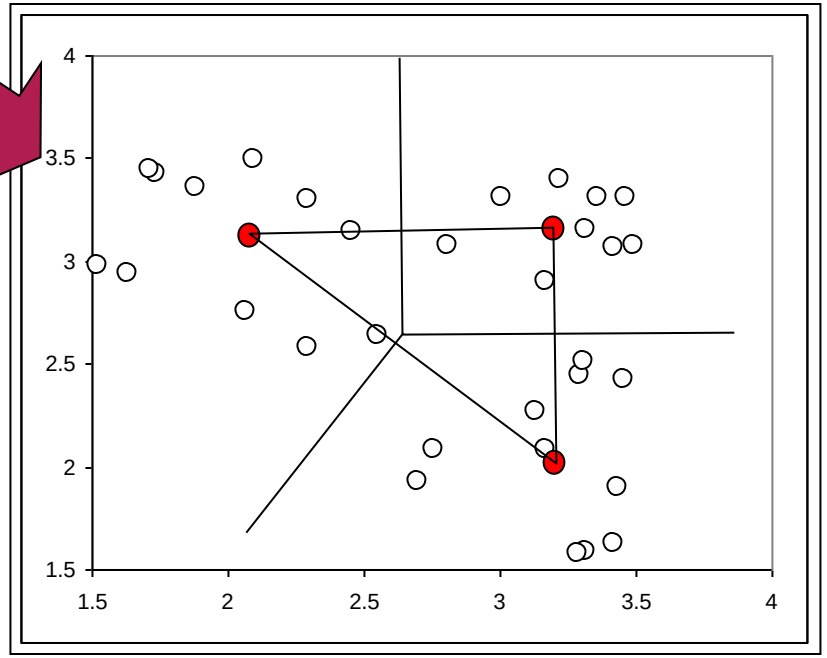
- Calcular **nuevamente los centros de los clusters desde los centroides escogidos en la iteración 1**.
- Localizar cada punto en el centroide que está más cerca a él.



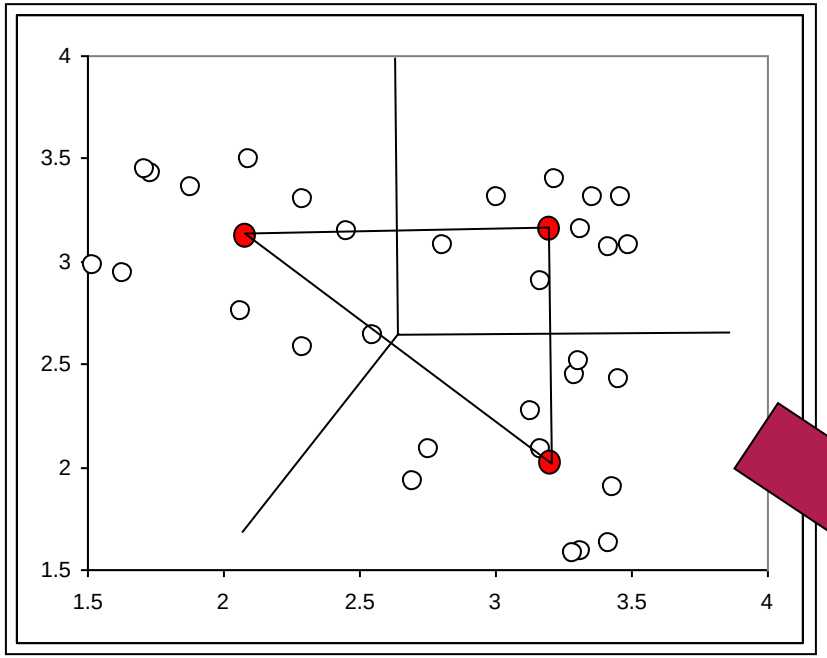
Iteración 3



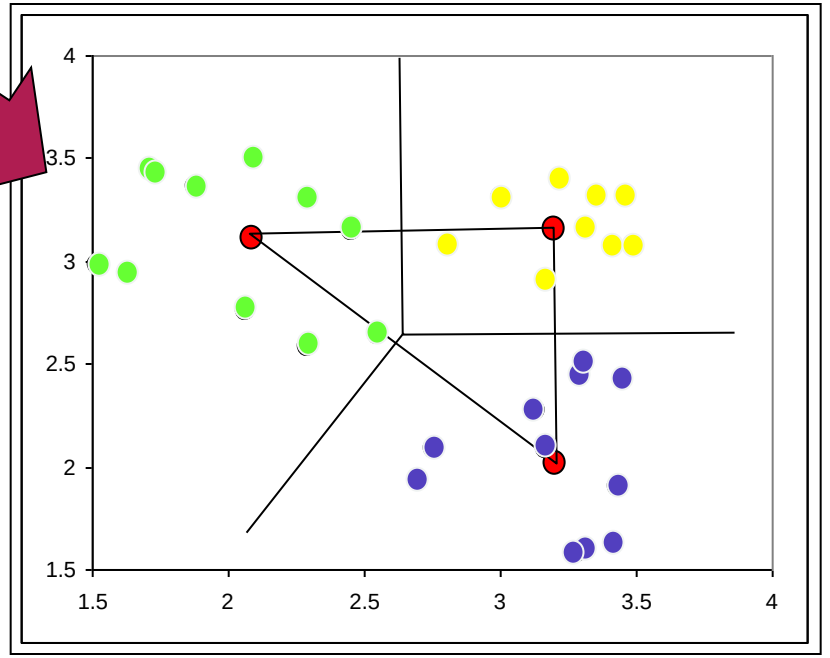
- Recalcular los **centros como los centroides encontrados en la iteración 2**.
- Localizar cada punto en el centroide que está más cerca a él.



Iteración 4 y etapa final



- Recalcular los centros como los centroides de los clusters desde la iteraciónn 3.
- Nada cambió!!
- Ok, está listo.



Practicando en R Studio...!