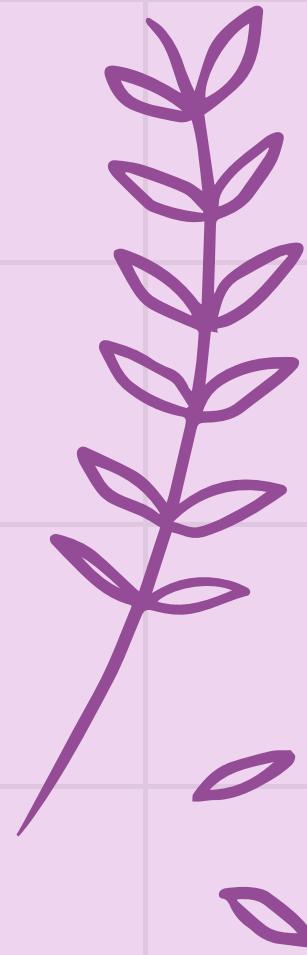


Introducción al Aprendizaje No Supervisado

R Ladies Medellín
Taller 4-2022



Agenda

1. R-Ladies
2. Equipo de Trabajo Capítulo Medellín
3. Diferencia entre Aprendizaje Supervisado y Aprendizaje no Supervisado
4. Clustering: K-Means
5. Clustering: jerárquico



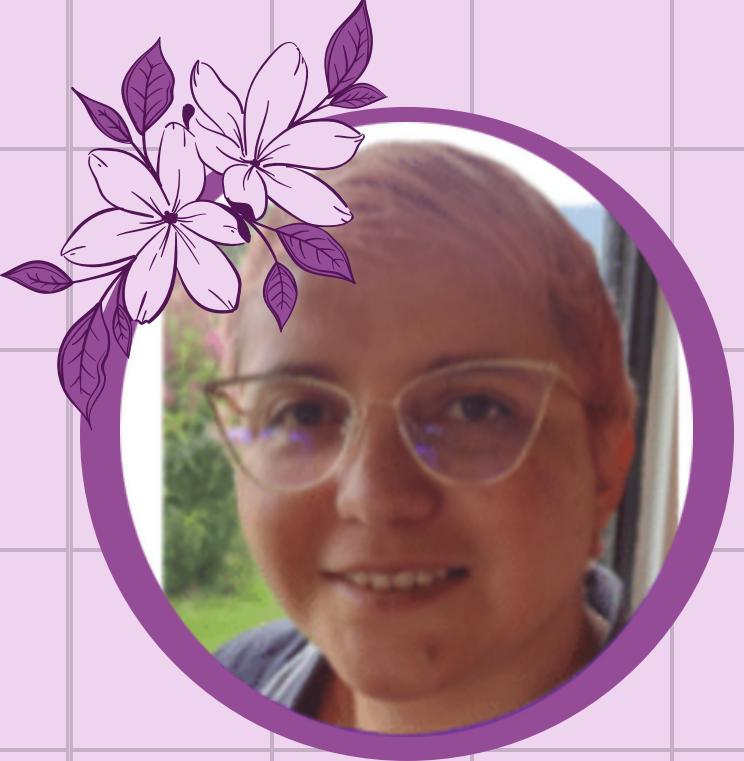
...

...



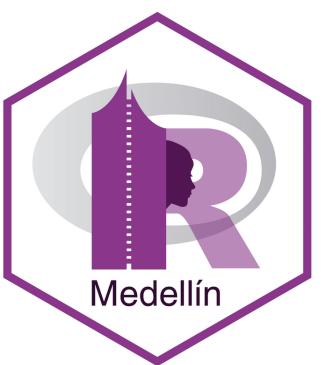
R-Ladies es una organización mundial cuya misión es promover la diversidad de género en la comunidad R.





Equipo de Trabajo Capítulo Medellín

Recuerda descargar el material del taller en Github



The screenshot shows a GitHub repository page for [RLadiesMedellin/Meetup-13-Clustering](https://github.com/RLadiesMedellin/Meetup-13-Clustering). The repository is public and contains one branch and no tags. The main file listed is `Clustering.Rmd`. A context menu is open over the `Code` button, showing options for cloning the repository via HTTPS, SSH, or GitHub CLI, and for opening it with GitHub Desktop or downloading a ZIP file. The repository has 0 stars, 1 watcher, and 0 forks. The `About` section provides a brief description of the repository's purpose.

RLadiesMedellin / Meetup-13-Clustering Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags

Go to file Add file Code Local Codespaces New

Clone

HTTPS SSH GitHub CLI

<https://github.com/RLadiesMedellin/Meetup-13-Clustering>

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

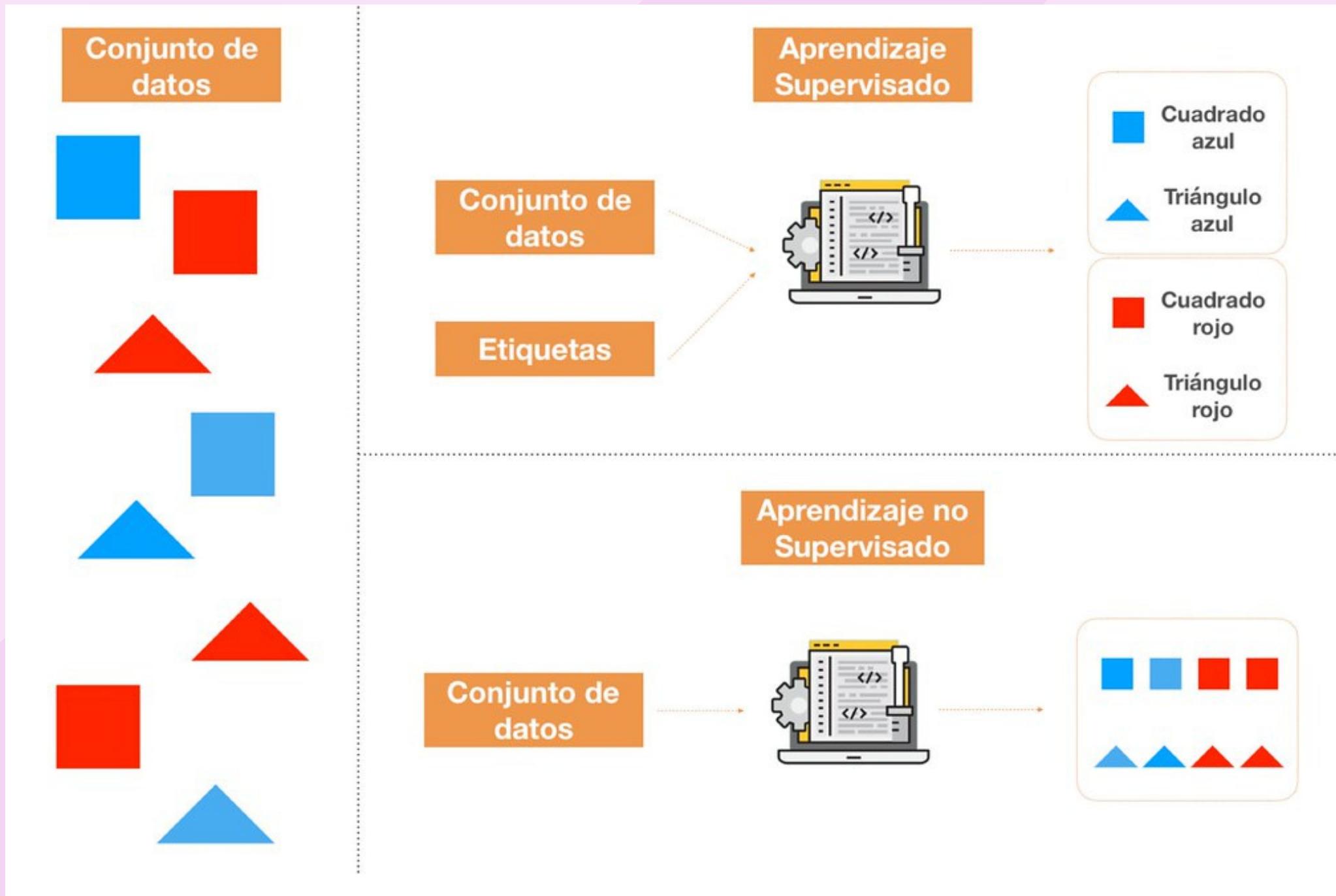
About

En este repositorio se encuentran todos los archivos utilizados durante el encuentro número 13 de R-Ladies capítulo Medellín: Introducción al Aprendizaje No Supervisado. Podrás explorar los algoritmos de clustering K-means y Jerarquización

Readme 0 stars 1 watching 0 forks

<https://github.com/RLadiesMedellin/Meetup-13-Clustering>

Diferencias entre aprendizaje supervisado y aprendizaje no supervisado



Necesidad de *datos etiquetados* en el aprendizaje supervisado

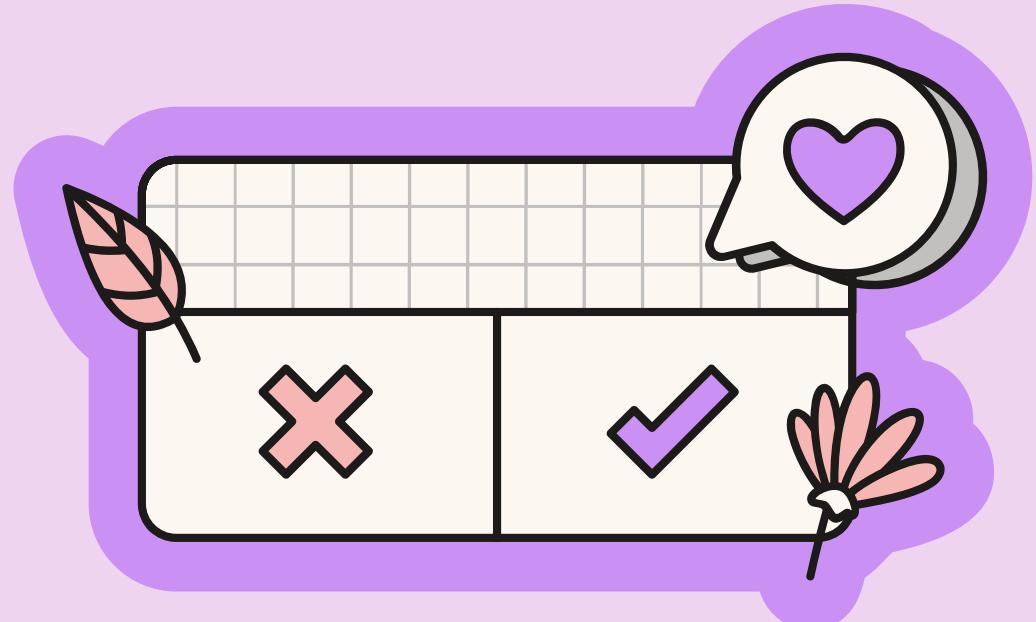
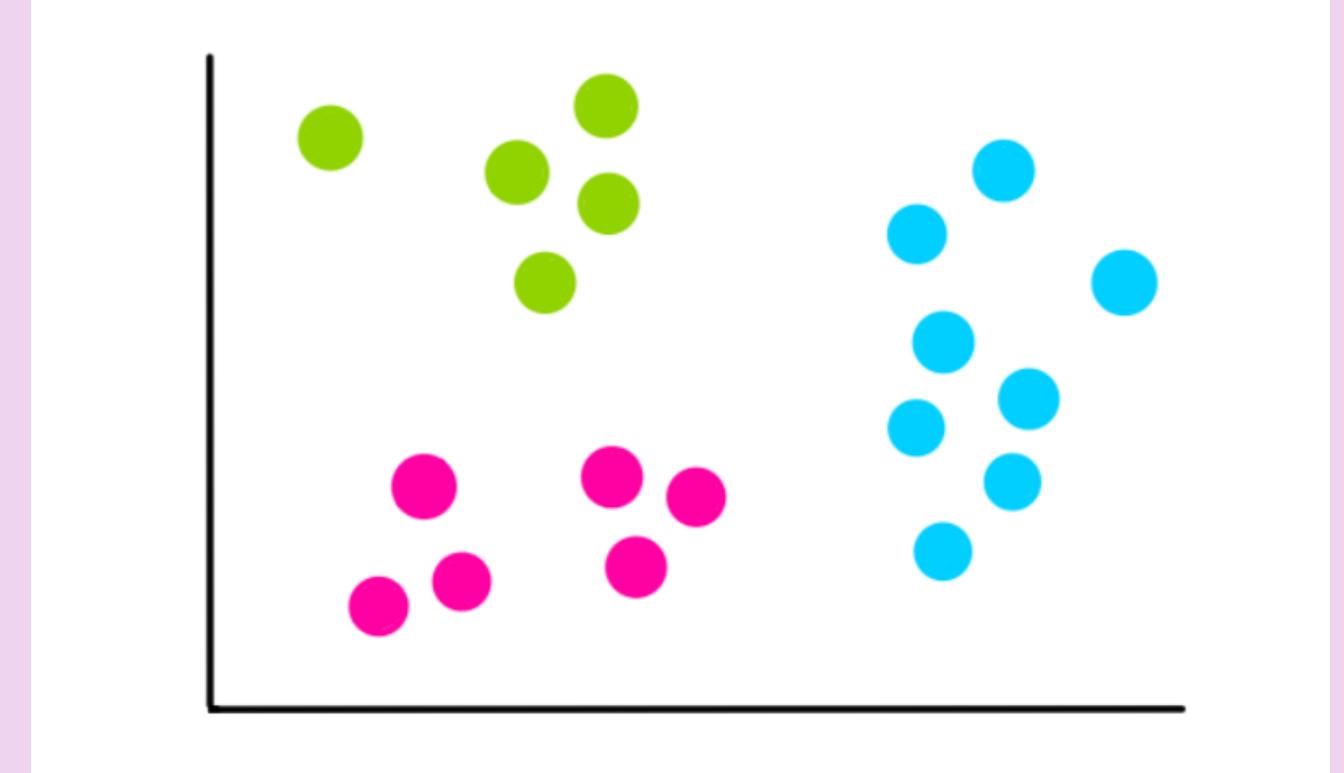
El *modelo* que se utiliza para solucionar un problema

El aprendizaje supervisado requiere muchos más recursos



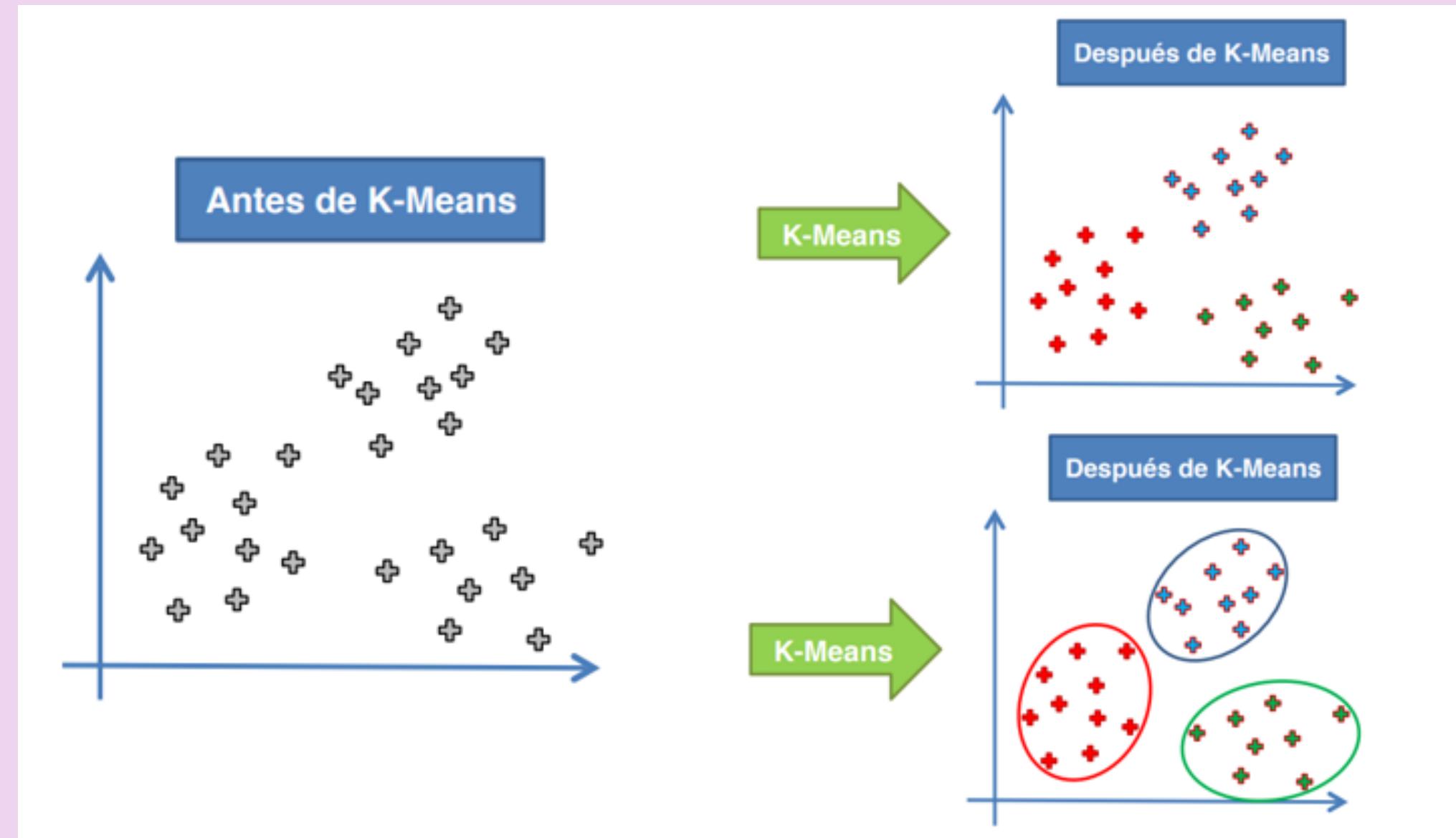
Clustering

Clustering es un proceso similar al de clasificación, pero con un fundamento diferente. En el Clustering no sabes cuales categorías estas buscando, si no que intentas crear una segmentación de tus propios datos en grupos más o menos homogéneos.





Clustering: K-Means



K-means agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster.

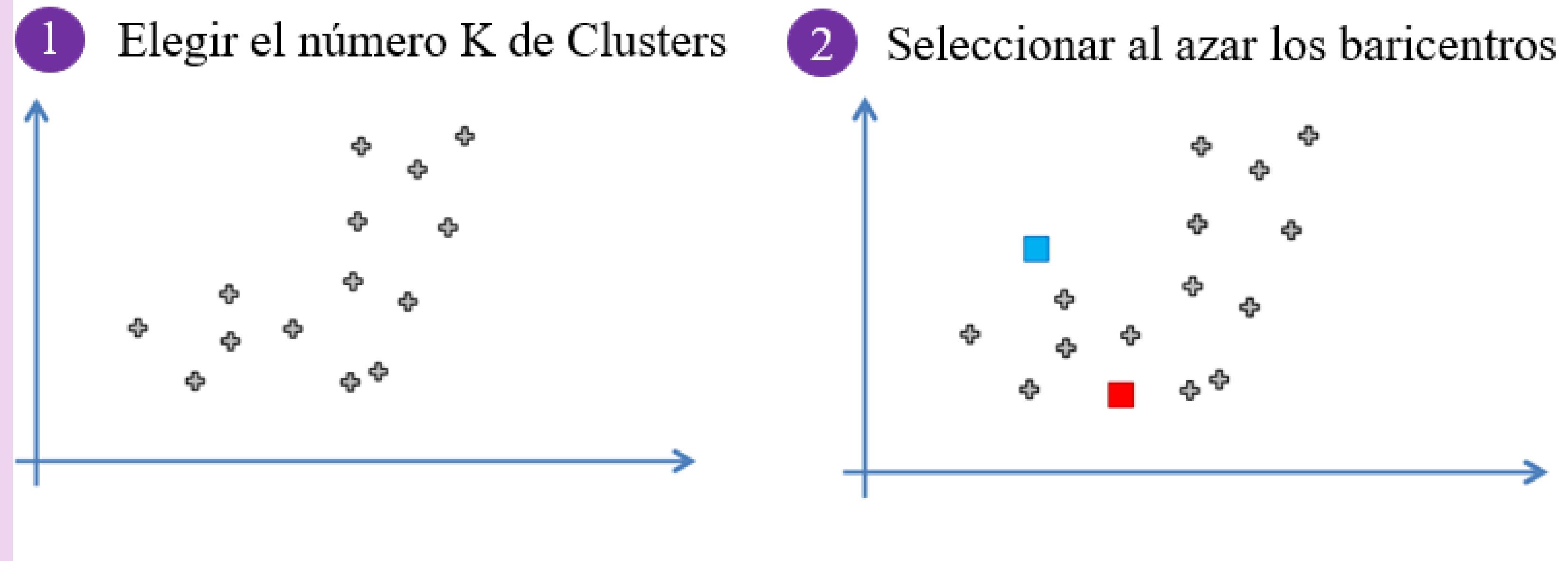


Clustering: K-Means

- Paso a paso**
- 1 Elegir el número K de Clusters
 - 2 Seleccionar al azar K puntos, los baricentros (no necesariamente de nuestro dataset)
 - 3 Asignar cada punto al baricentro más cercano
 - 4 Calcular y asignar el nuevo baricentro de cada cluster
 - 5 Reasignar cada punto de los datos a su baricentro más cercano, si ha habido nuevas asignaciones, ir al paso 4, sino ir FIN



Clustering: K-Means

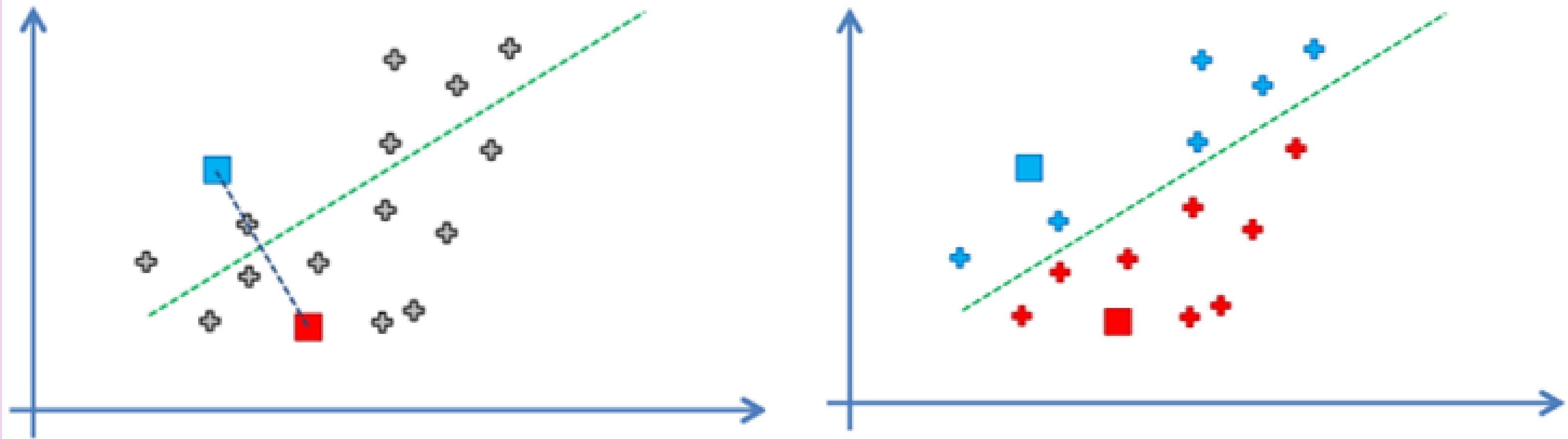




Clustering: K-Means

3

Asignar cada punto al baricentro más cercano

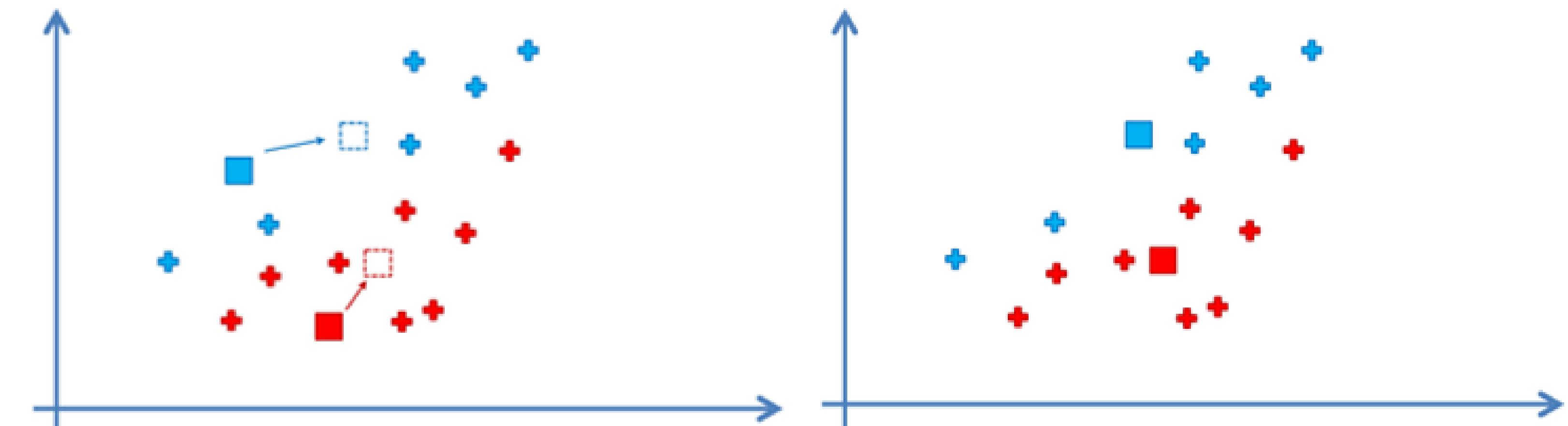




Clustering: K-Means

4

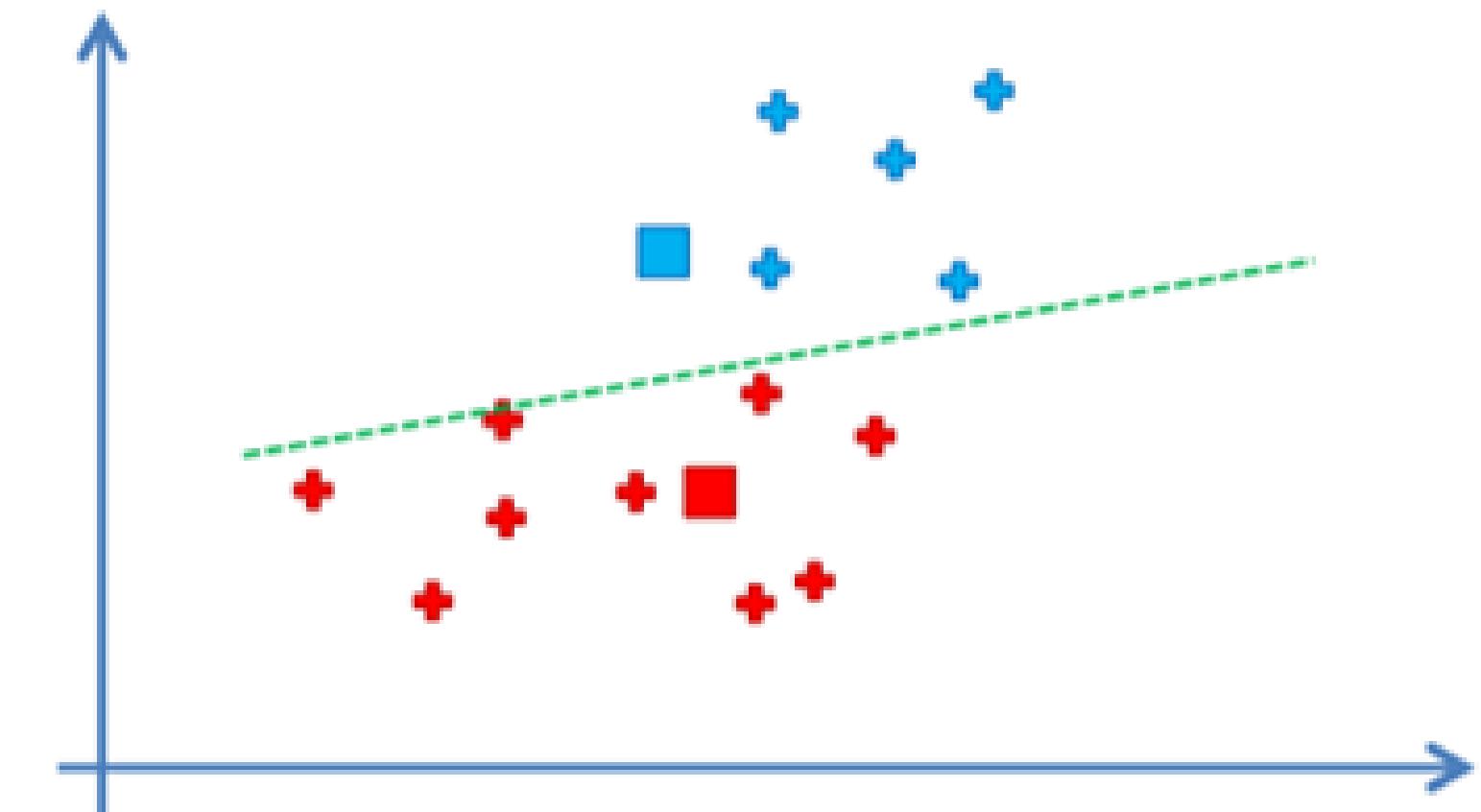
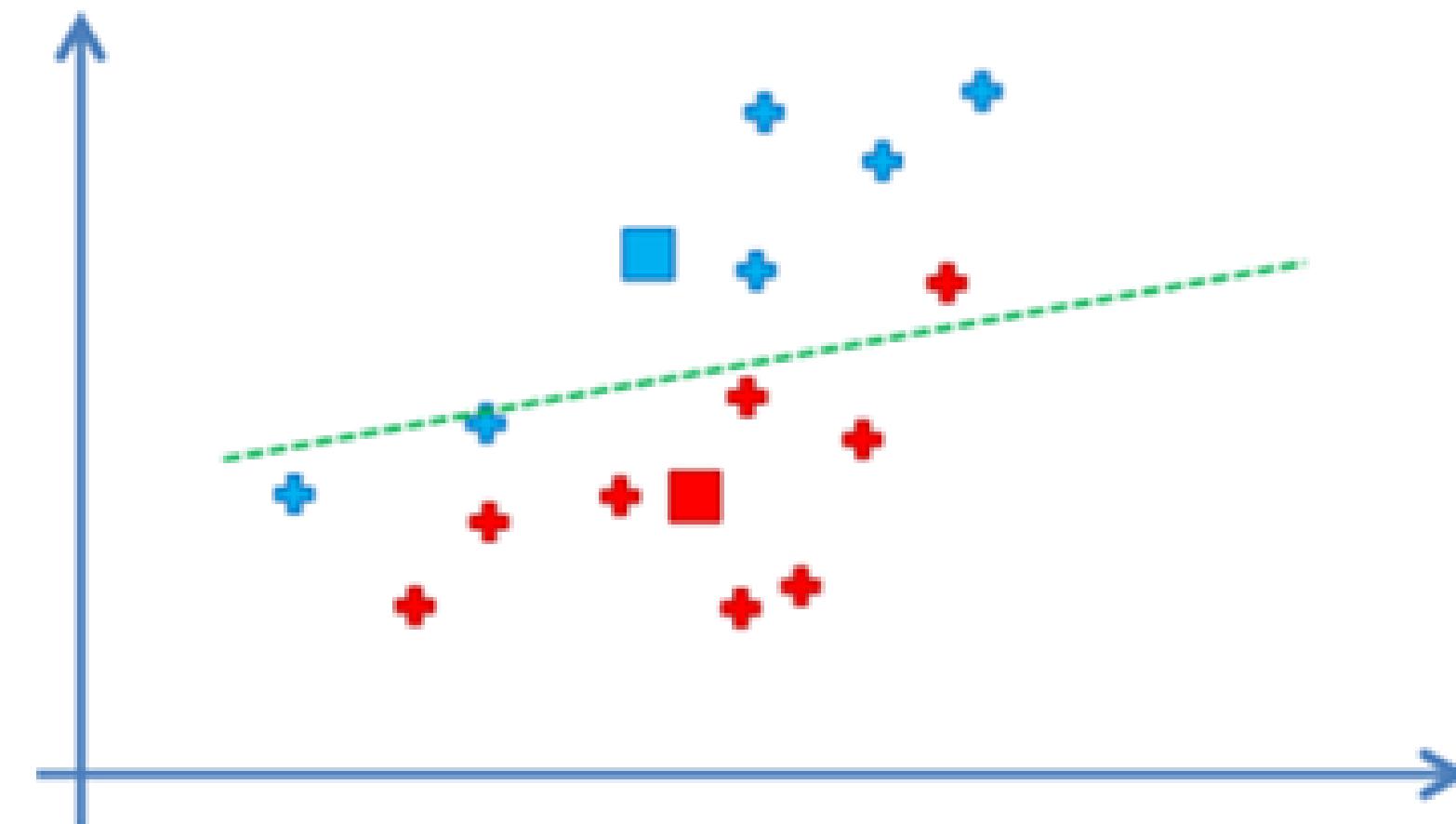
Calcular y asignar el nuevo baricentro de cada cluster





Clustering: K-Means

- 5 Reasignar cada punto de los datos a su baricentro más cercano

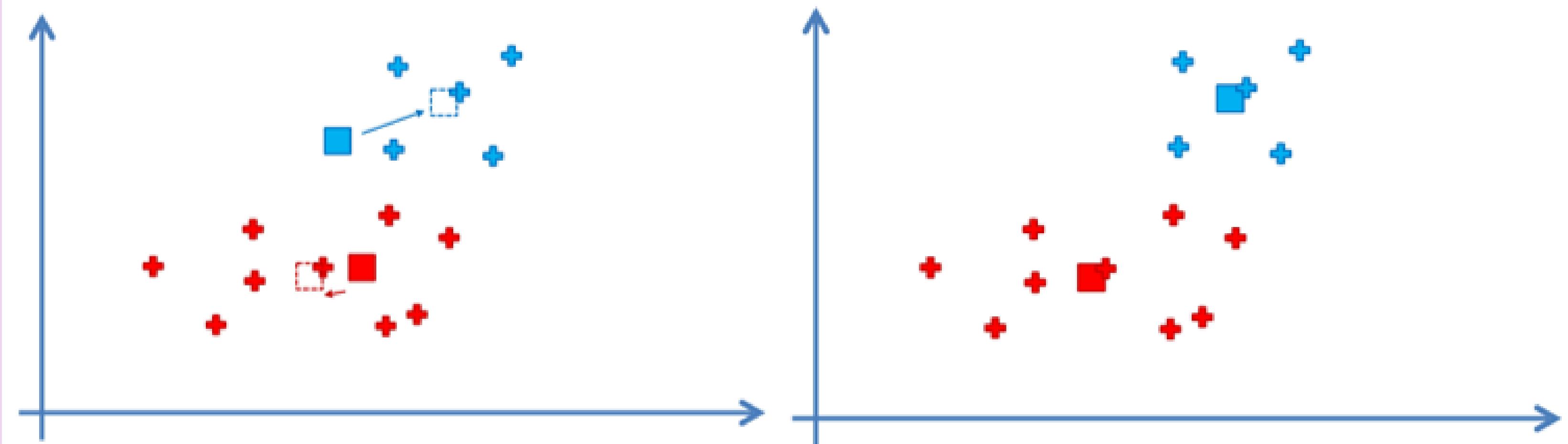




Clustering: K-Means

4

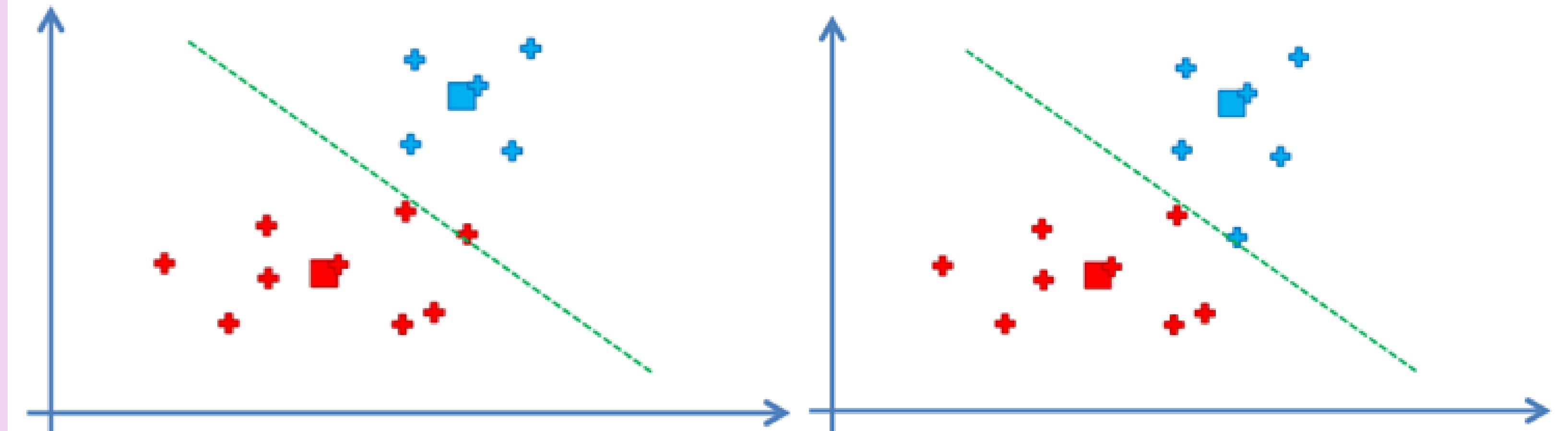
Calcular y asignar el nuevo baricentro de cada cluster





Clustering: K-Means

- 5 Reasignar cada punto de los datos a su baricentro más cercano

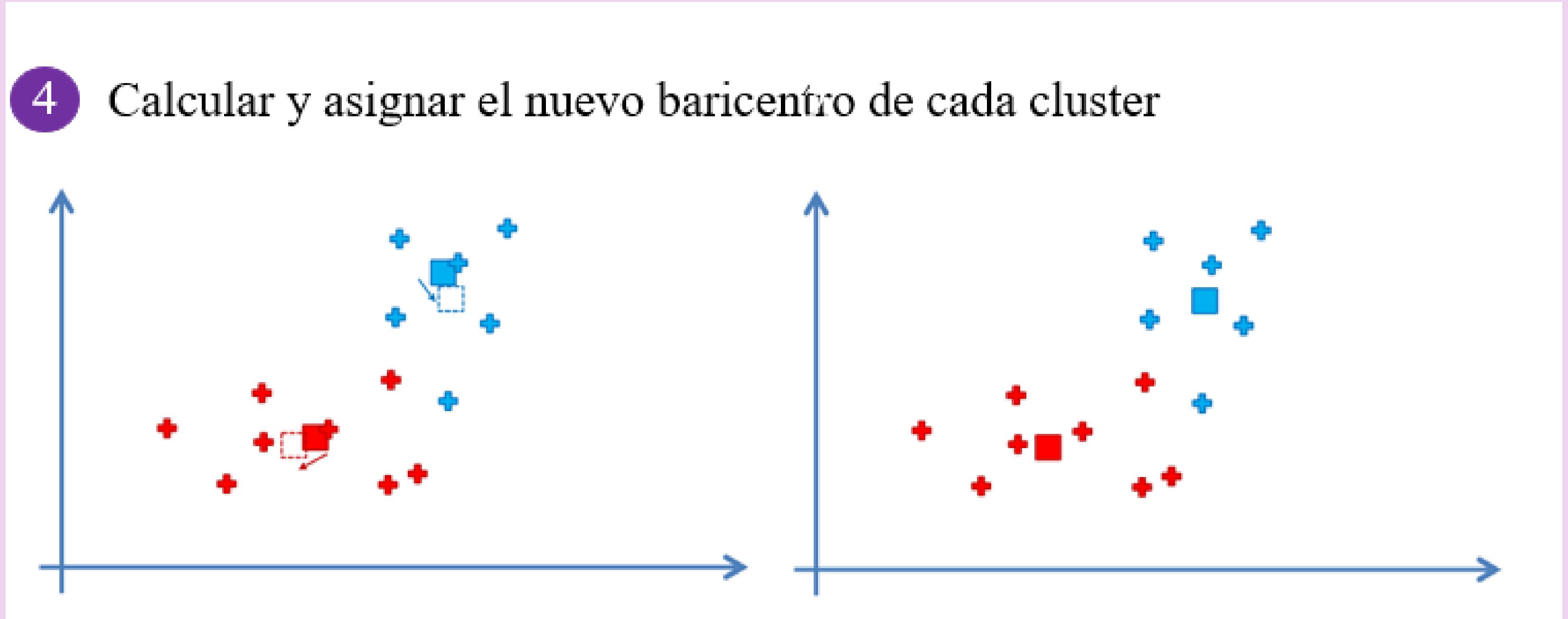




Clustering: K-Means

4

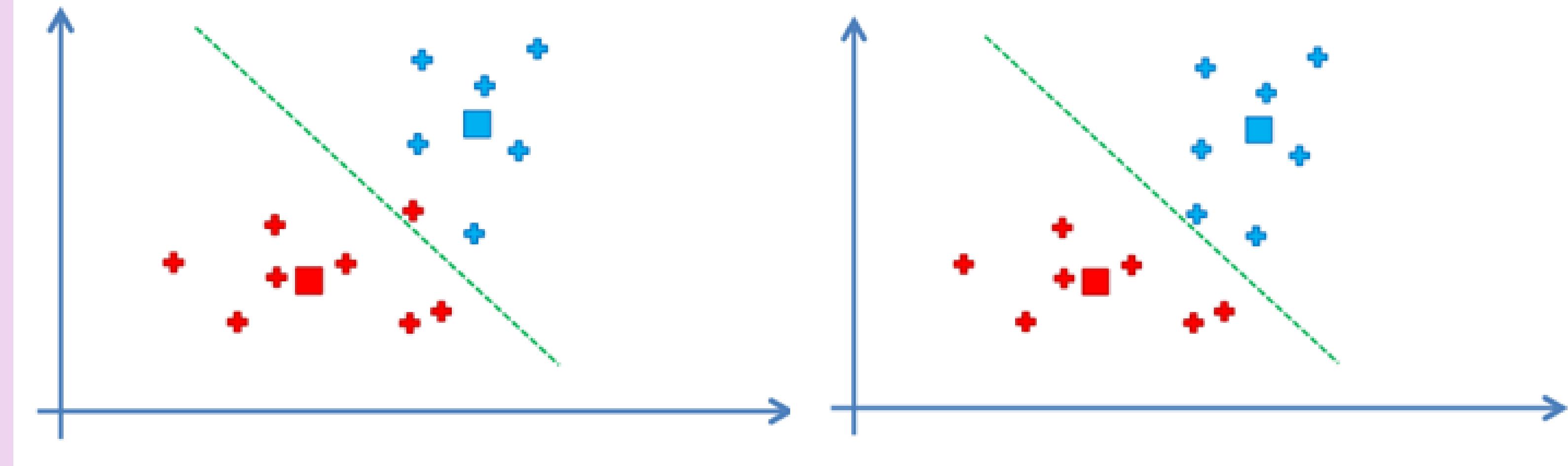
Calcular y asignar el nuevo baricentro de cada cluster





Clustering: K-Means

- 5 Reasignar cada punto de los datos a su baricentro más cercano

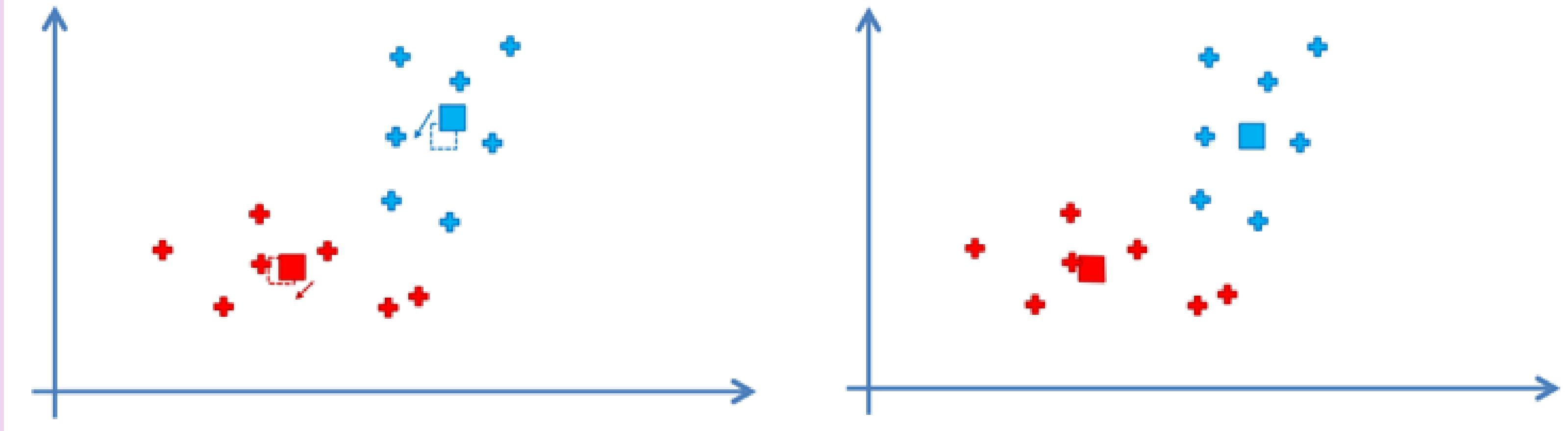




Clustering: K-Means

4

Calcular y asignar el nuevo baricentro de cada cluster

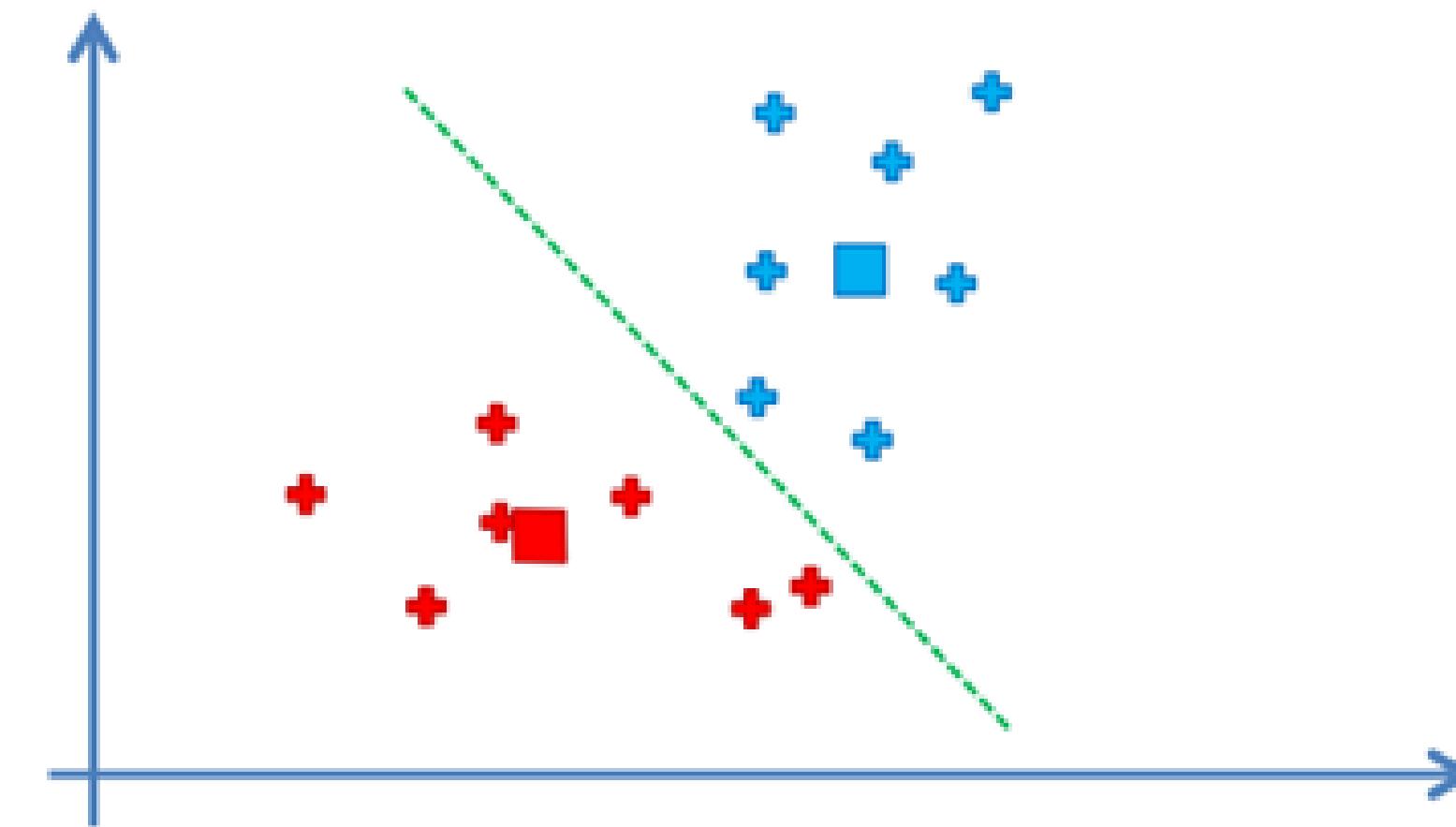




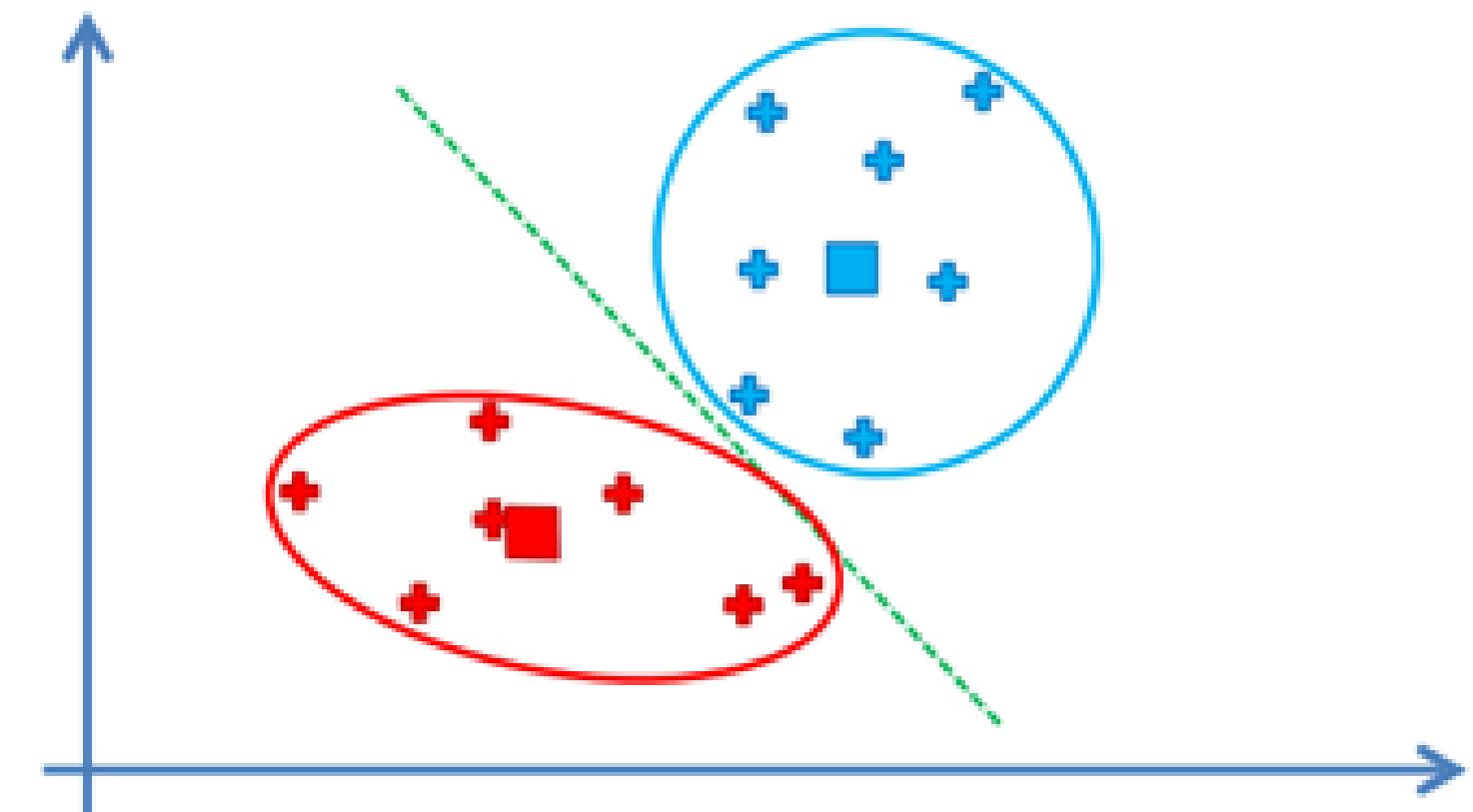
Clustering: K-Means

5

Reasignar cada punto de los datos a su baricentro más cercano



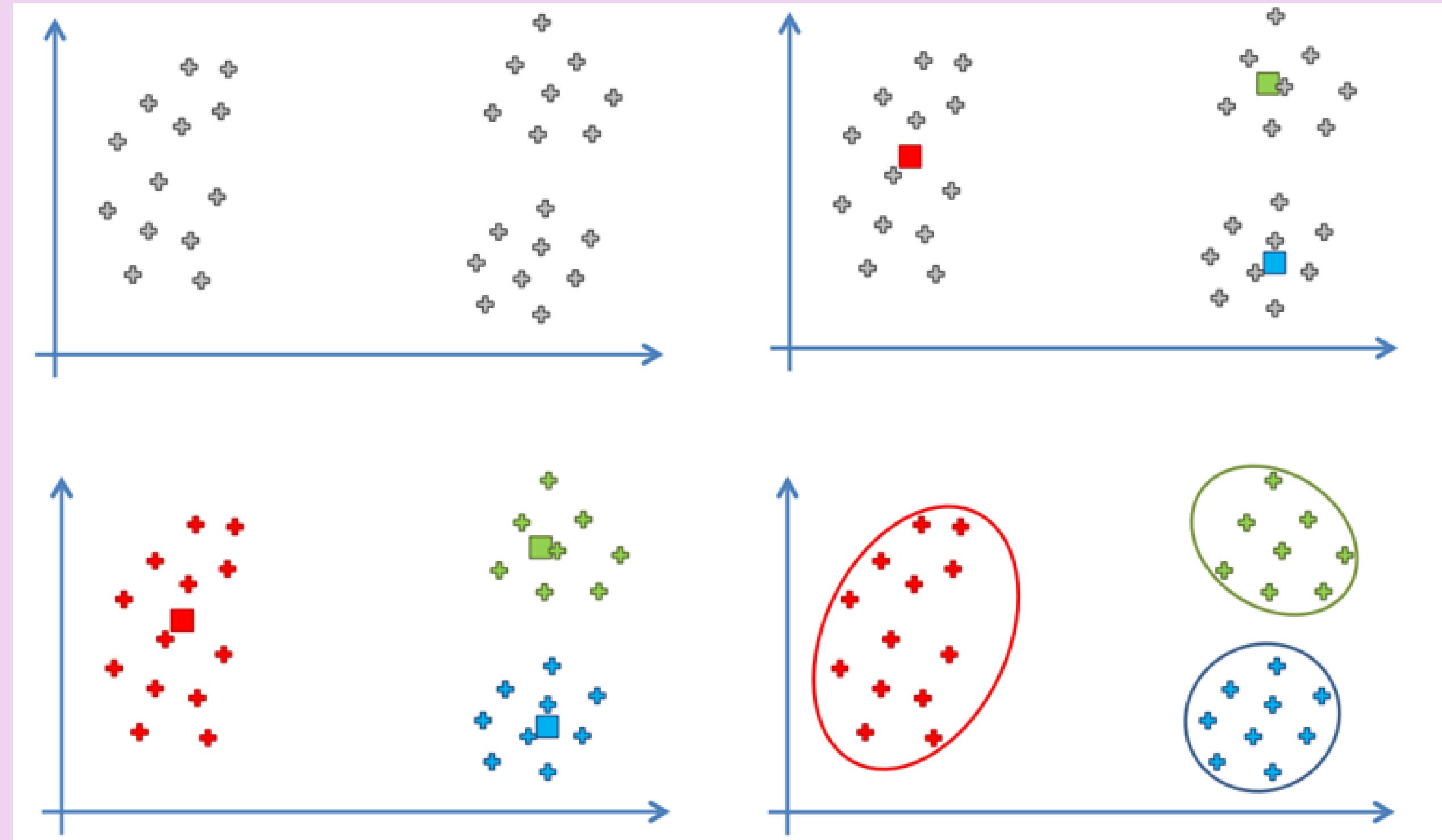
FIN: El modelo está listo





K-Means: la trampa de la inicialización aleatoria

¿Y qué pasaría si elegimos una mala inicialización aleatoria del baricentro?





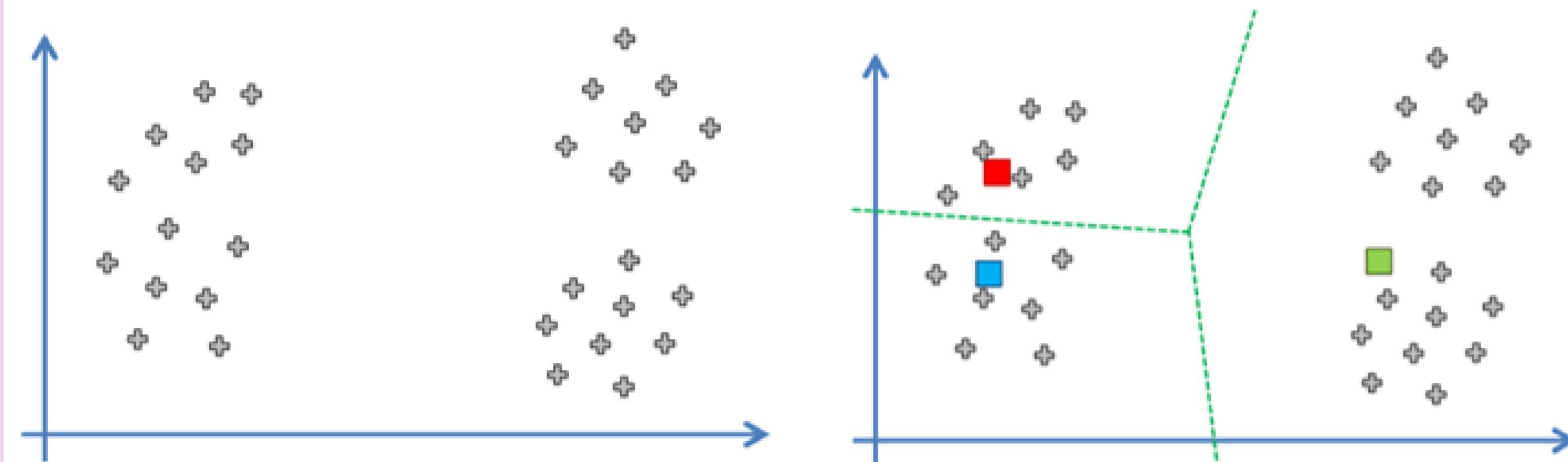
K-Means: la trampa de la inicialización aleatoria

- Paso a paso**
- 1 Elegir el número K de Clusters
 - 2 Seleccionar al azar K puntos, los baricentros (no necesariamente de nuestro dataset)
 - 3 Asignar cada punto al baricentro más cercano
 - 4 Calcular y asignar el nuevo baricentro de cada cluster
 - 5 Reasignar cada punto de los datos a su baricentro más cercano, si ha habido nuevas asignaciones, ir al paso 4, sino ir FIN



K-Means: la trampa de la inicialización aleatoria

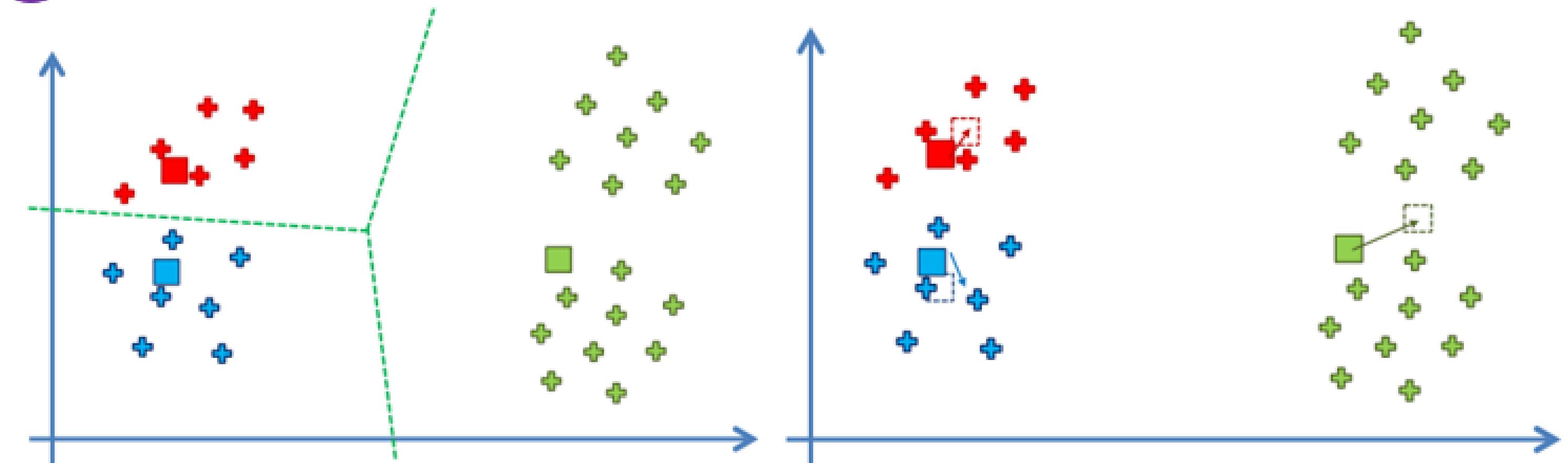
- 1 Elegir el número K de Clusters
- 2 Seleccionar al azar los baricentros





K-Means: la trampa de la inicialización aleatoria

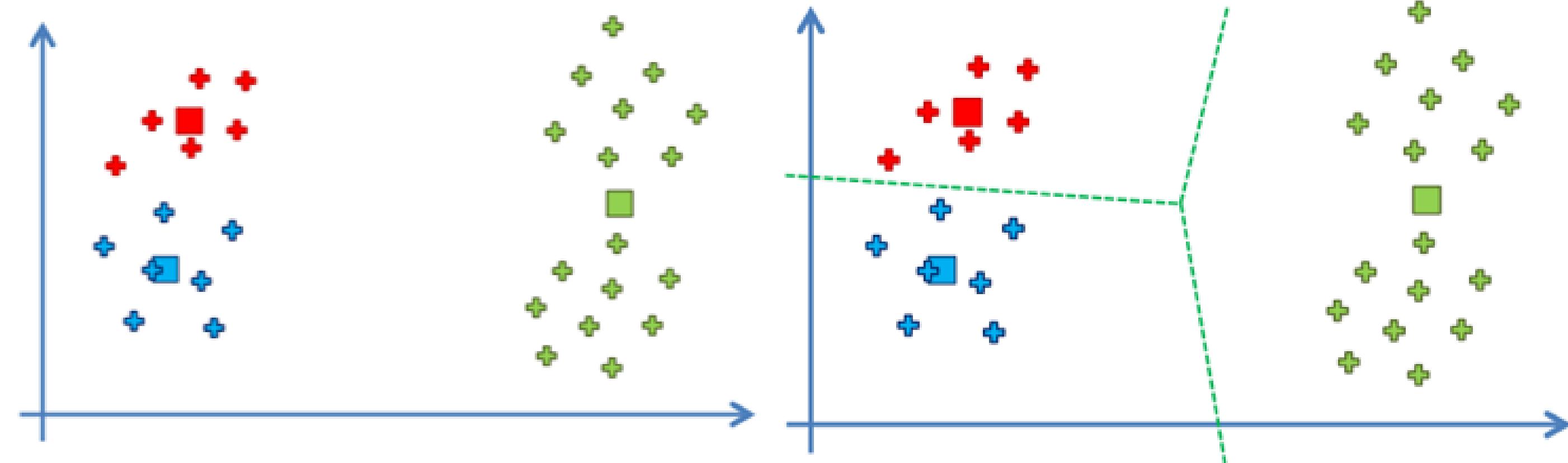
- 3 Asignar cada punto al baricentro más cercano





K-Means: la trampa de la inicialización aleatoria

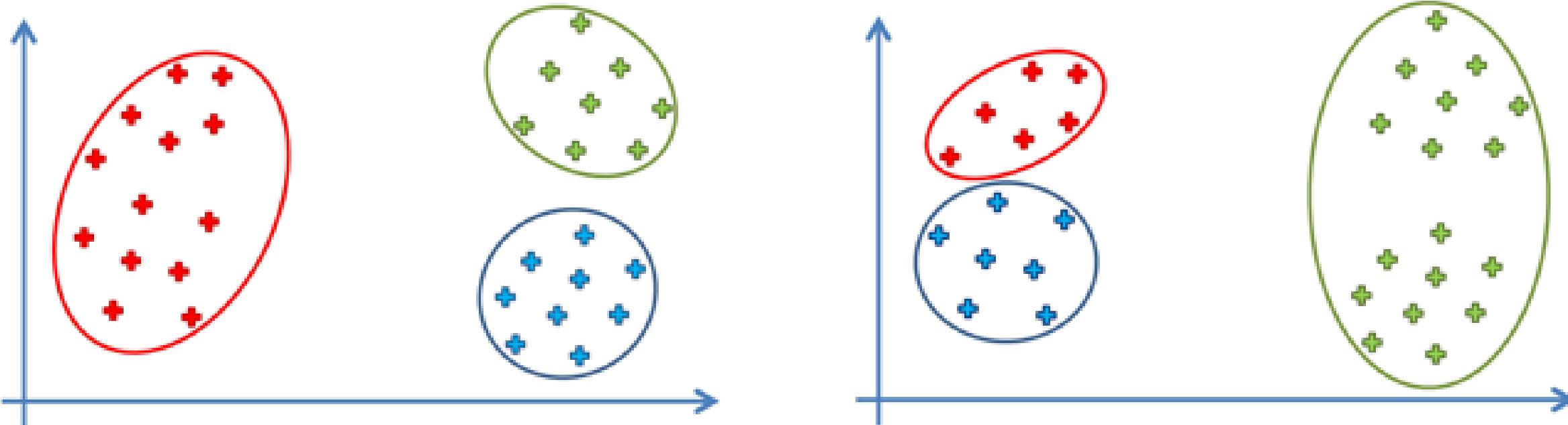
- 4 Calcular y asignar el nuevo baricentro de cada cluster
- 5 Reasignar cada punto de los datos a su baricentro más cercano





K-Means: la trampa de la inicialización aleatoria

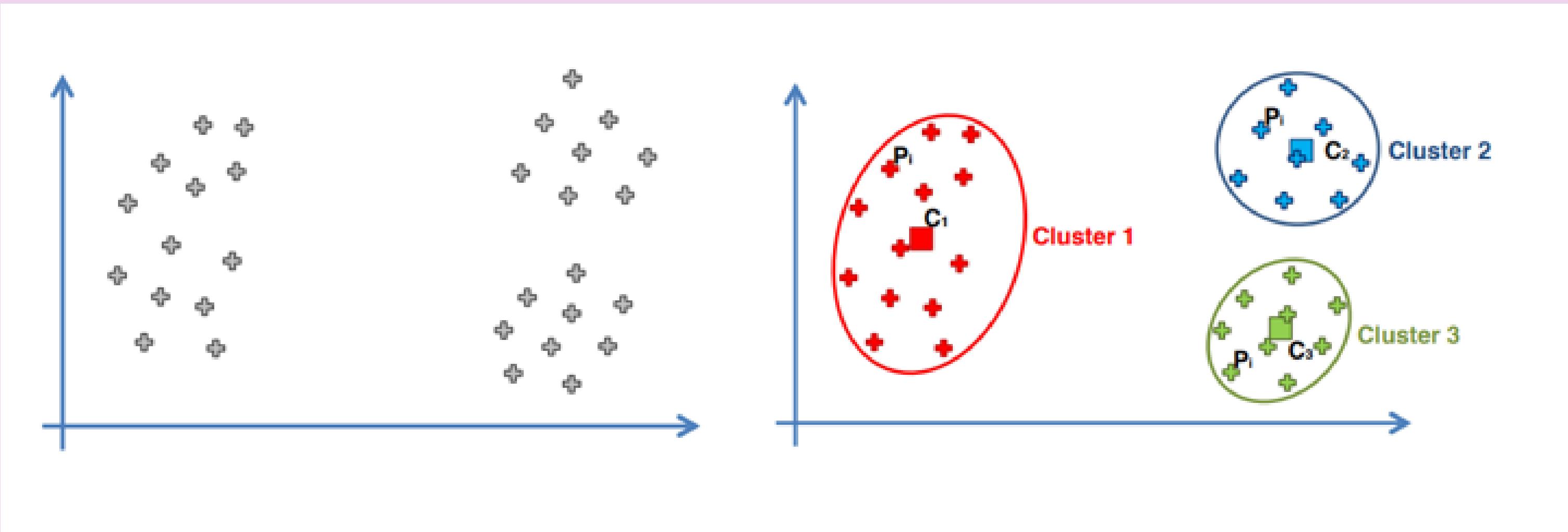
Una selección COMPLETAMENTE aleatoria del baricentro nos puede llevar a una clasificación incorrecta.



K-Means++



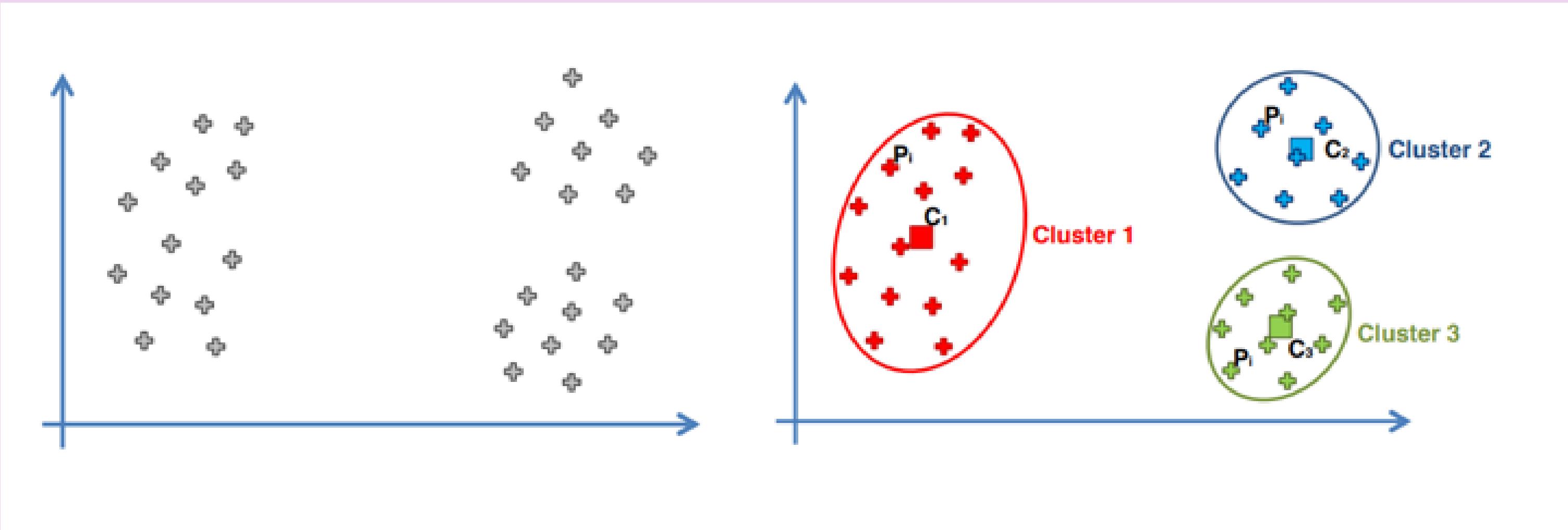
K-Means: ¿Cómo elegir el número correcto de clusters?



La Suma de los Cuadrados de los Centros de los Clusters WCSS (Within Cluster Sum of Squares) representa una solución para la selección adecuada del número de clusters



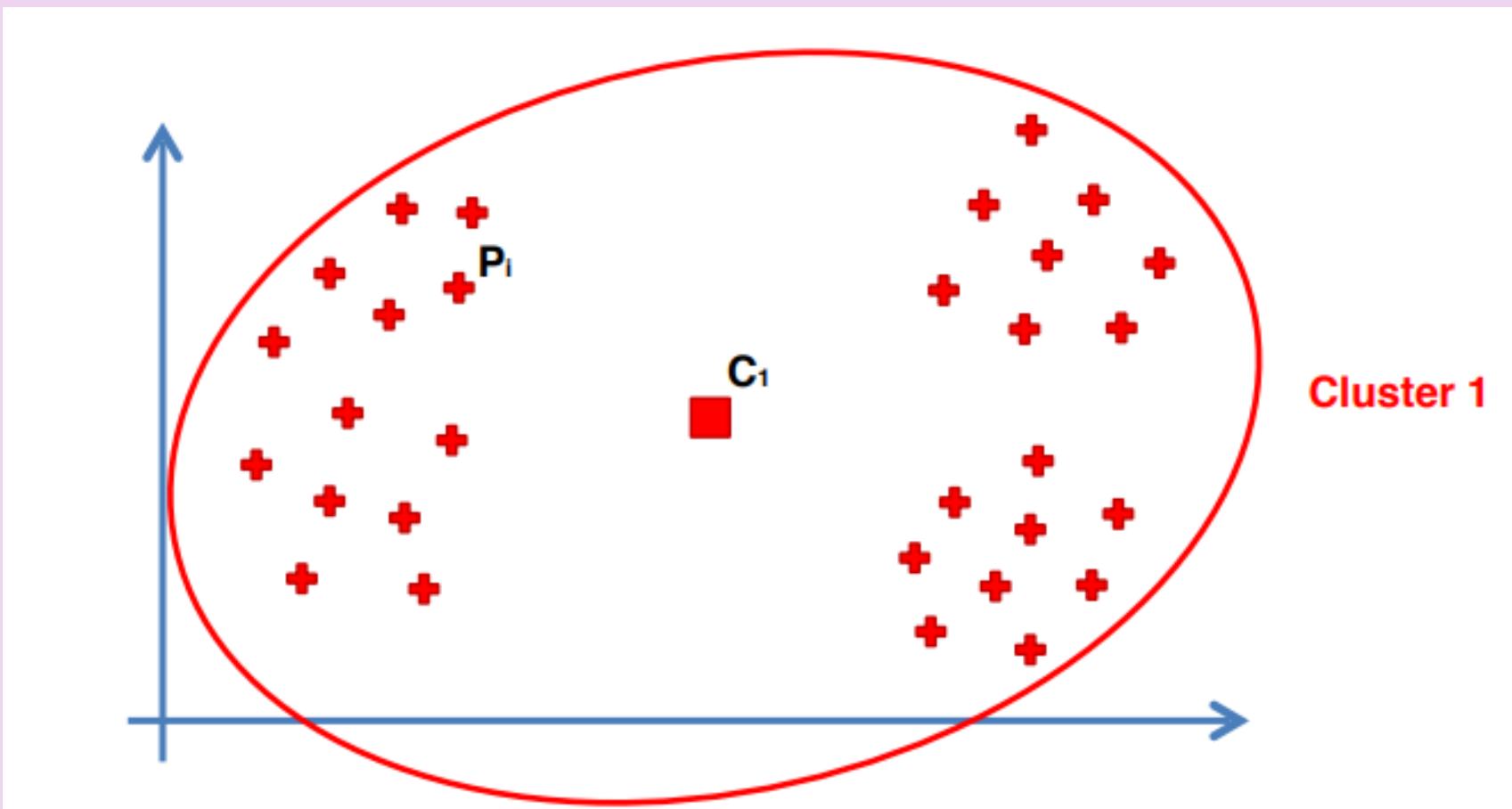
K-Means: ¿Cómo elegir el número correcto de clusters?



$$WCSS = \sum_{P_i \in Cluster1} d(P_i, C_1)^2 + \sum_{P_i \in Cluster2} d(P_i, C_2)^2 + \sum_{P_i \in Cluster3} d(P_i, C_3)^2$$



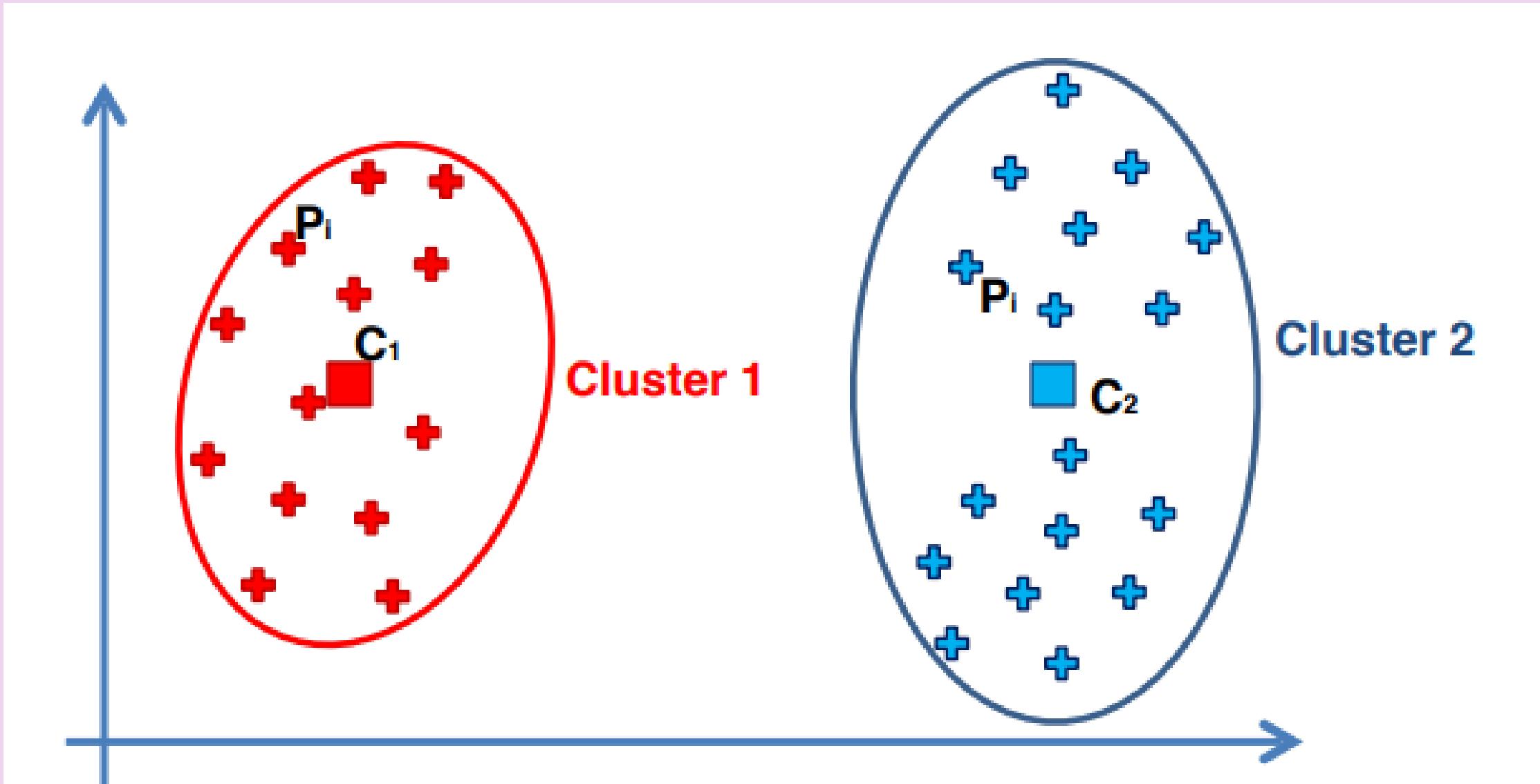
K-Means: ¿Cómo elegir el número correcto de clusters?



$$WCSS = \sum_{P_i \in Cluster1} d(P_i, C_1)^2$$



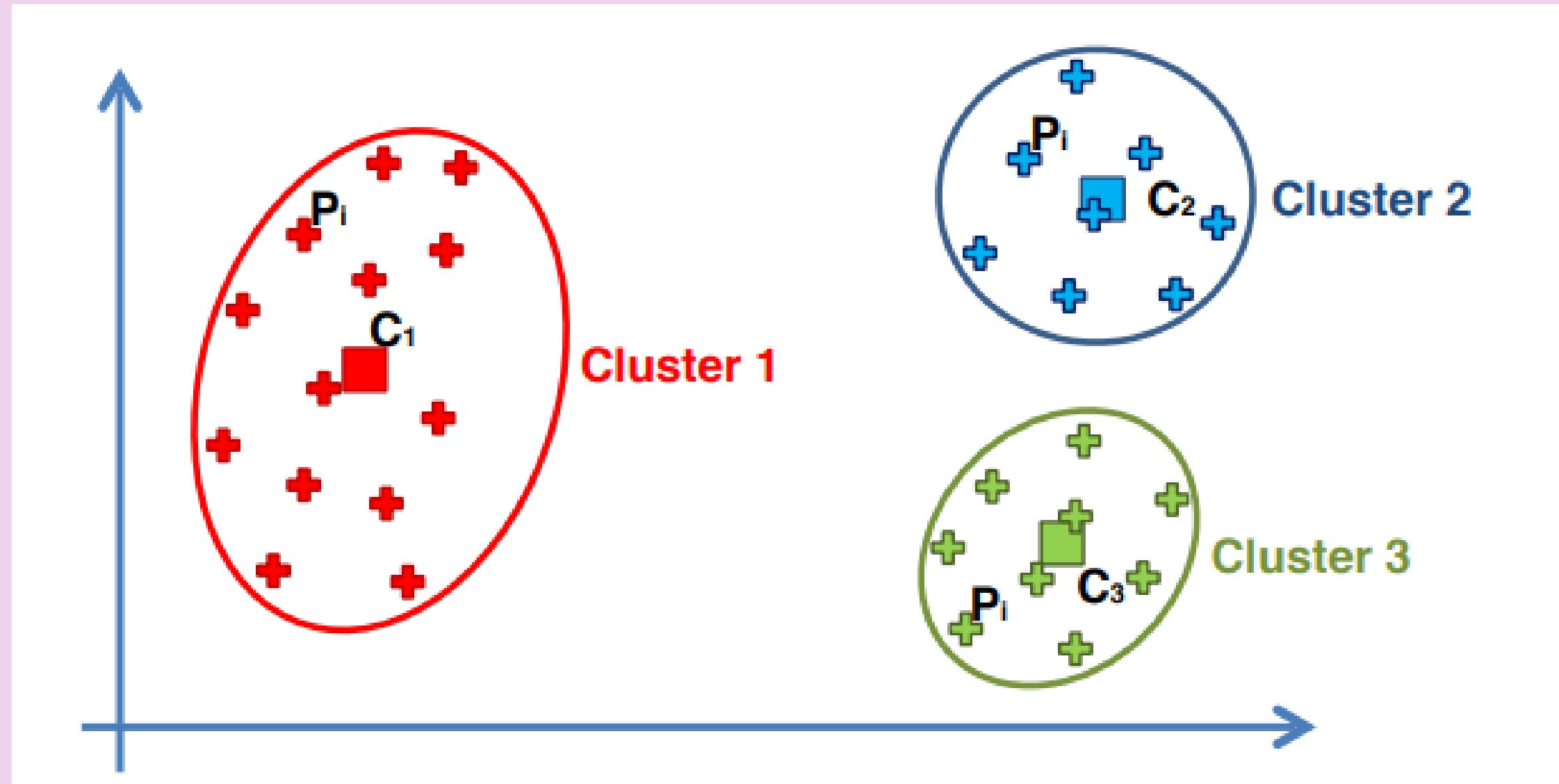
K-Means: ¿Cómo elegir el número correcto de clusters?



$$WCSS = \sum_{P_i \in Cluster1} d(P_i, C_1)^2 + \sum_{P_i \in Cluster2} d(P_i, C_2)^2$$



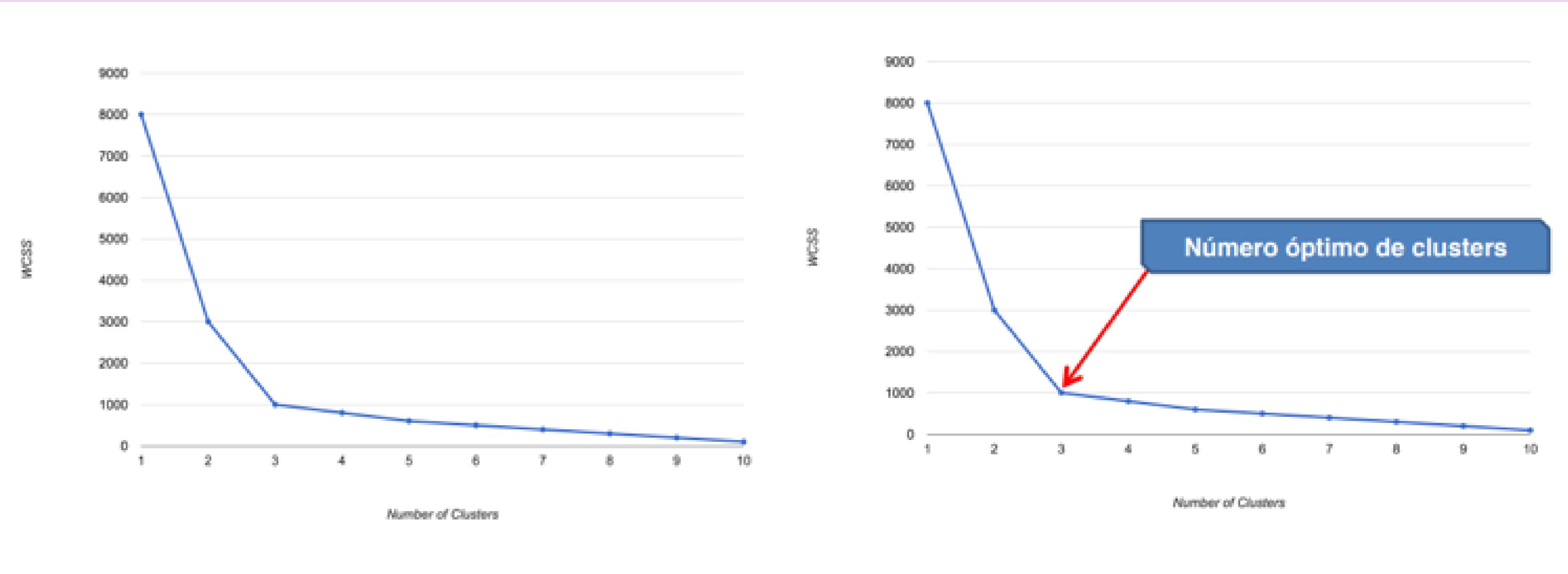
K-Means: ¿Cómo elegir el número correcto de clusters?



$$WCSS = \sum_{P_i \in Cluster1} d(P_i, C_1)^2 + \sum_{P_i \in Cluster2} d(P_i, C_2)^2 + \sum_{P_i \in Cluster3} d(P_i, C_3)^2$$



K-Means: ¿Cómo elegir el número correcto de clusters?





K-Means: Aplicación





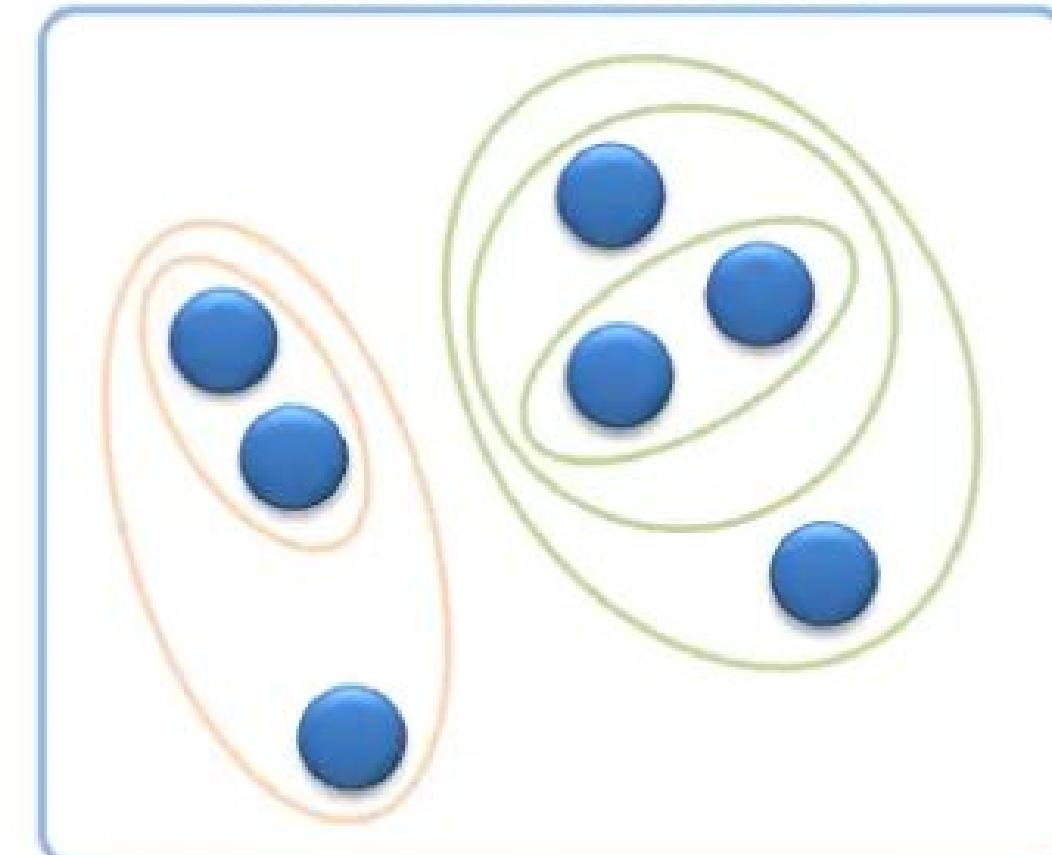
Cluster Jerárquico

Es un método de análisis de grupos puntuales, que busca construir una jerarquía de grupos en entre los elementos analizados.

Crear grupos de elementos homogéneos entre sí y heterogéneos entre grupos, para conseguirlo principalmente se puede hacer mediante estrategia aglomerativa o divisiva

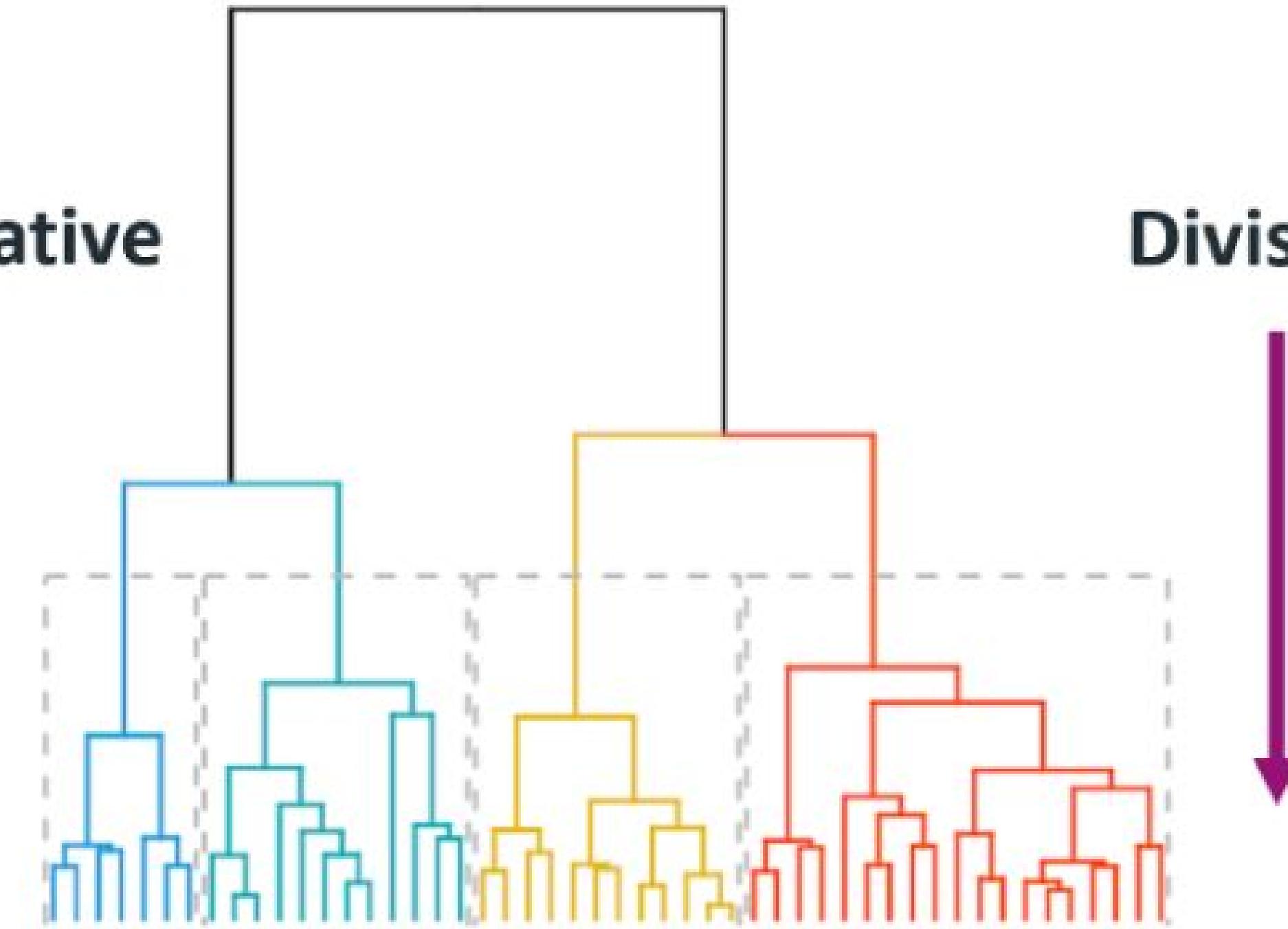
CLUSTER JERÁRQUICOS

HIERARCHICAL CLUSTER



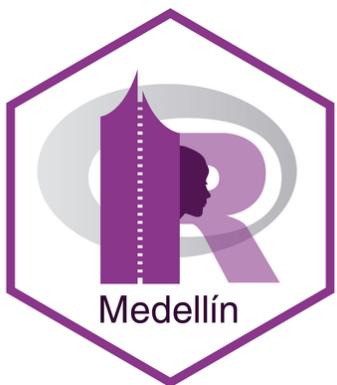
Métodos Jerárquicos

Agglomerative



Divisive

Agglomerative





Agglomerative

Paso a paso

- 1 Considerar cada una de las observaciones como un cluster individual, formando así la base del dendrograma (hojas).
- 2 Iterar hasta que todas las observaciones pertenecen a un único cluster:
 - a Se calcula la distancia entre cada posible par de los n clusters.
 - b Los dos clusters más similares se fusionan, de forma que quedan n-1 clusters.
- 3 Determinar dónde cortar la estructura de árbol generada (dendrograma).



Linkage

Completo o Máximo:

La mayor distancia entre todos los posibles pares formados por una observación del Cluster A y una del cluster B. Más conservadora

Promedio:

El promedio de las distancias entre todos los posibles pares formados por una observación del Cluster A y una del cluster B.



Linkage

Centroide:

Distancia entre los centroides de los cluster

Ward:

Minimiza la suma total de la varianza dentro del cluster



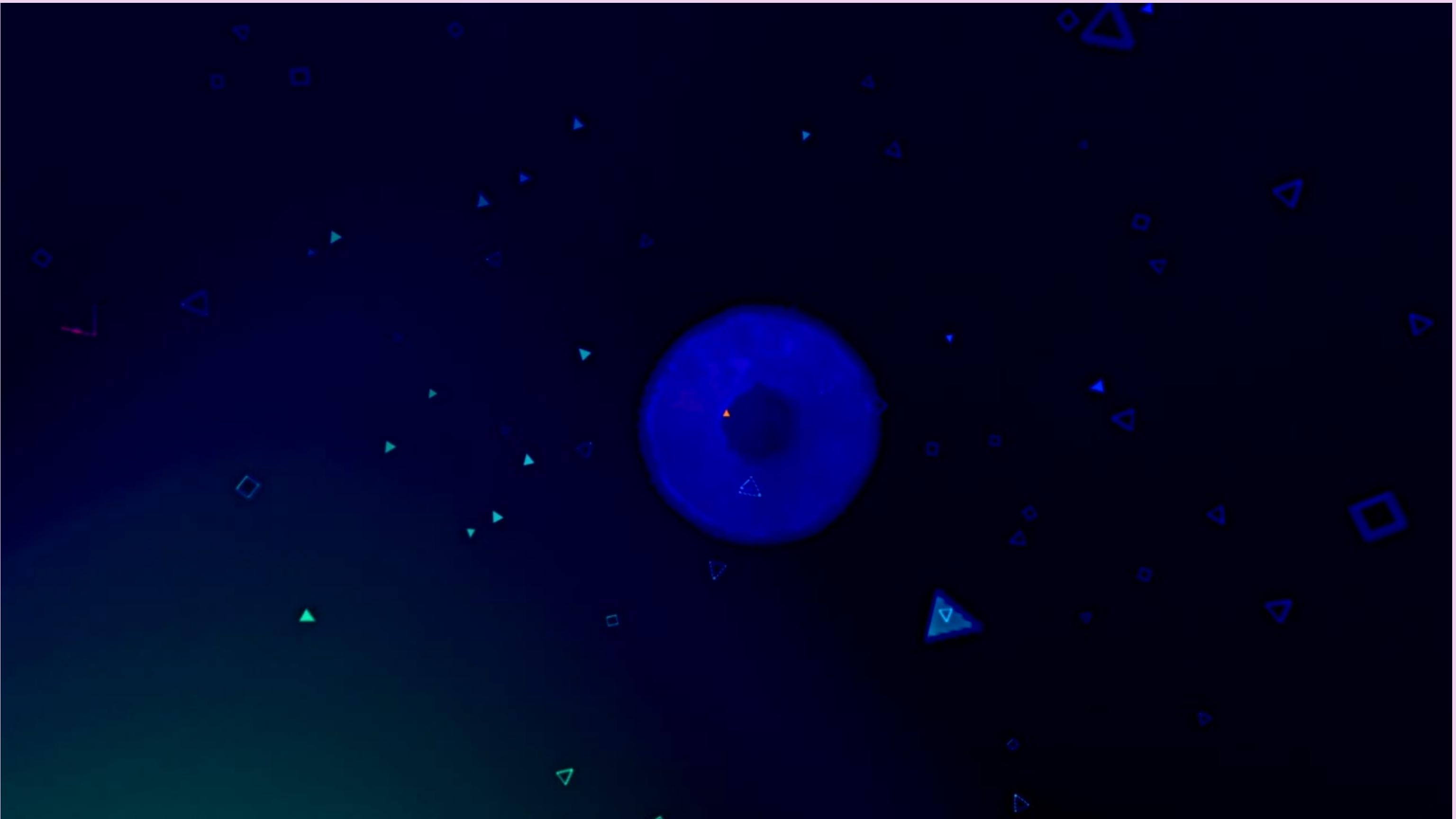
Coeficiente de Correlación

Útil para comparar métodos para saber si un dendograma resultante presenta relaciones aceptables al compararlo con la matriz de distancias originales

$$c = \frac{\sum_{i < j} [x(i, j) - \bar{x}][t(i, j) - \bar{t}]}{\sqrt{\sum_{i < j} [x(i, j) - \bar{x}]^2 \sum_{i < j} [t(i, j) - \bar{t}]^2}}$$



Divisive





Divisive

Paso a paso

- 1 Todas las observaciones forman UN cluster
- 2 Calcular para cada cluster la mayor de las distancias entre pares de observaciones
- 3 Seleccionar el cluster con mayor distancia
- 4 Calcular la distancia media de cada observación respecto a las demás
- 5 La observación más distante inicia un nuevo cluster
- 6 Se reasignan las observaciones entre el nuevo y el viejo cluster
- 7 Se repiten pasos del 2 al 6 hasta tener una observación por cluster



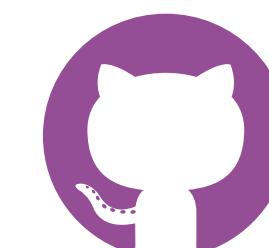
R'Ladies Medellín



R-Ladies Medellín



medellin@rladies.org



RLadiesMedellin



RLadiesMedellin



RLadiesMedellin



rladiesmedellin2

¡Muchas Gracias!

