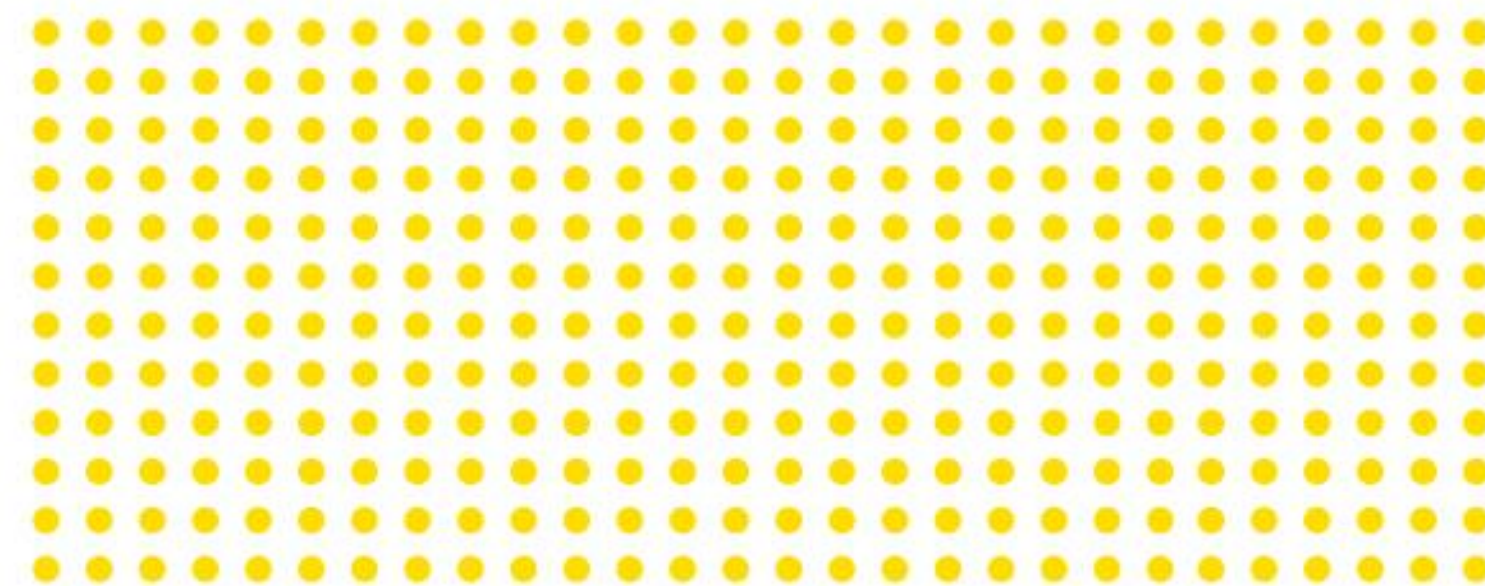


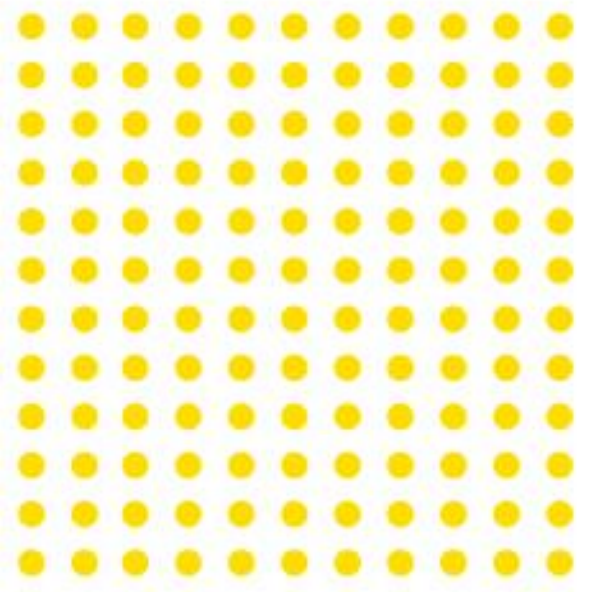


Universidad de
los Andes

Educación
Continua
Vicerrectoría Académica

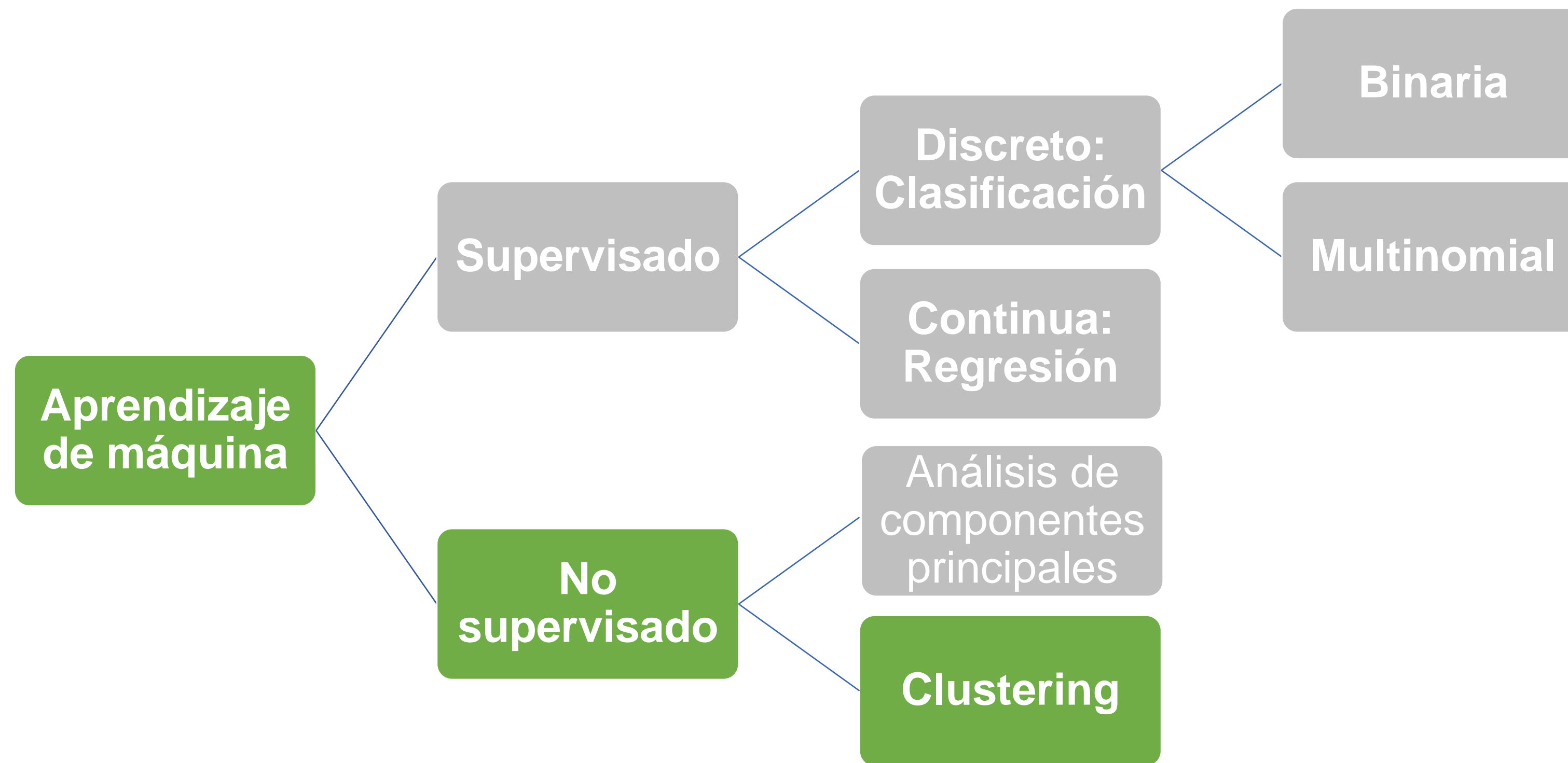


Agrupación

- 
- Concepto de distancia
 - K-medias
 - Agrupación jerárquica
 - Densidad



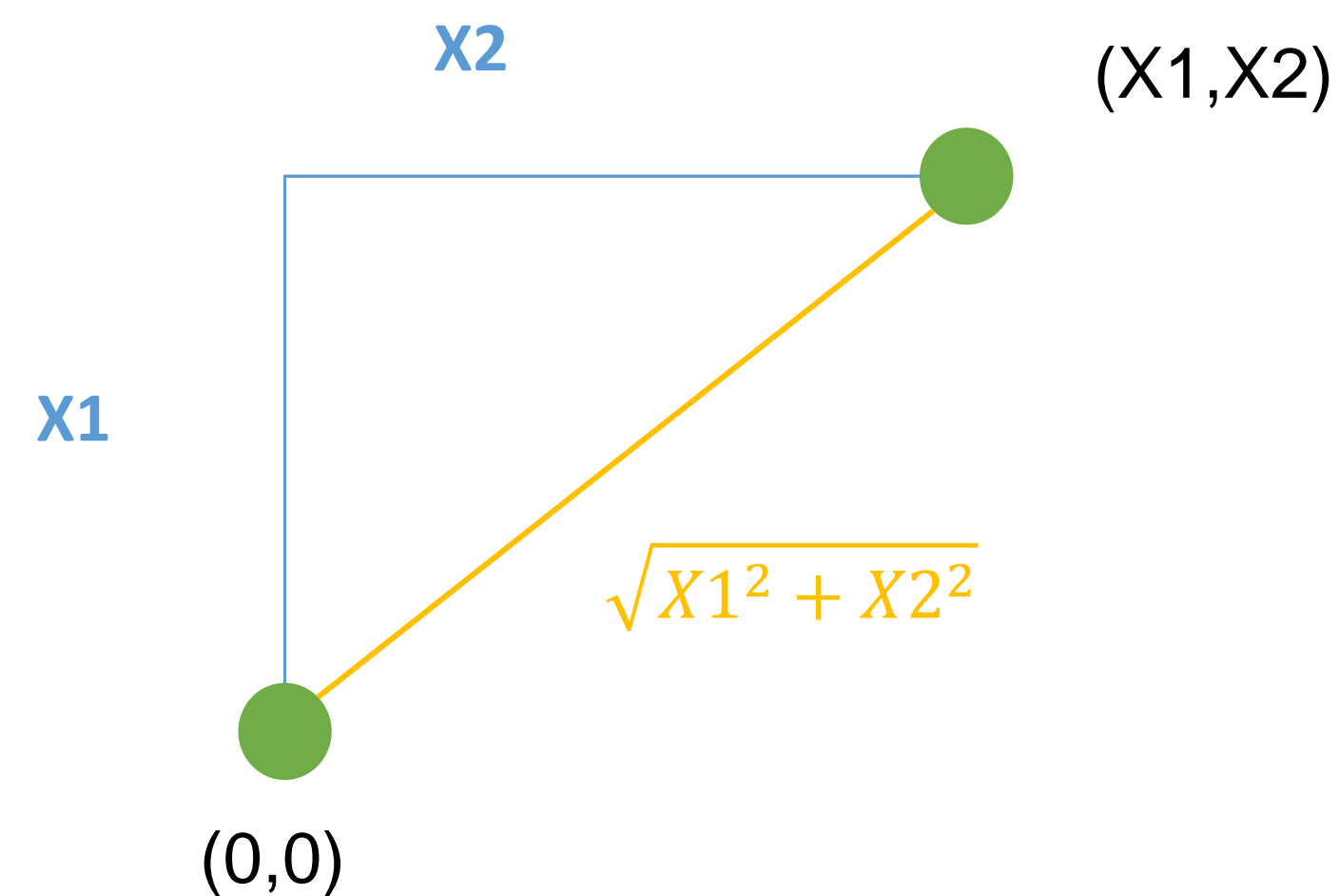
Estamos en:



Primero hablemos de distancia

Euclideana:

$$d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2}.$$



Manhattan:

$$d_{ij} = \sum_{k=1}^p |X_{ik} - X_{jk}|.$$

Ejemplo

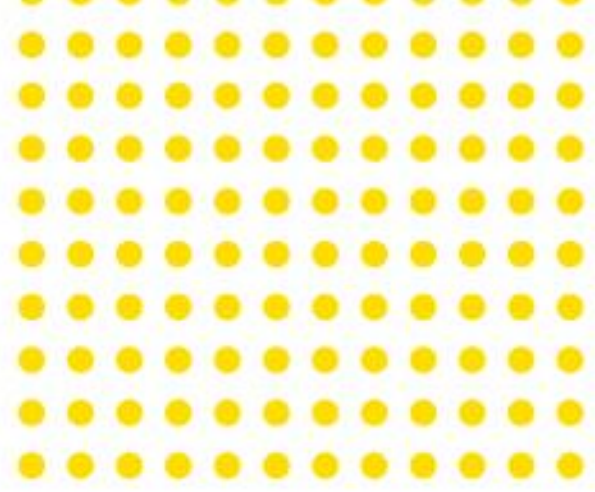
Debemos encontrar la distancia de 3 casas, a un centro comercial

Calcular la distancia de cada casa al centro comercial, con la distancia euclídeana y con la distancia Manhattan

	Norte	Oriente
Punto 1	200	100
Punto 2	150	150
Punto 3	-50	250
Centro comercial	0	0

Ejemplo: Distancia al centro comercial

	Norte	Oriente	Manhattan	Euclidiana
Punto 1	200	100	300	223,61
Punto 2	150	150	300	212,13
Punto 3	-50	250	300	254,95
Centro comercial	0	0	0	0



K medias



El paso a paso de lo que necesitamos

1. Definir el valor de K (Clusters).
2. Seleccionar unos puntos aleatorios para los centroides
3. Medir la distancia de cada punto al centroide
4. El centroide más cercano a cada punto toma “posesión” de ese punto (dato)
5. Se calcula la media de los datos
6. Se mueve el centroide a esa posición y se repiten los pasos anteriores
7. Si la media no cambia en los centroides, acabamos el proceso.

Ejemplo simple

Definir el valor de K
(Clusters)

Intentemos con 3



Ejemplo simple

Esta misma cantidad serán
los centroides

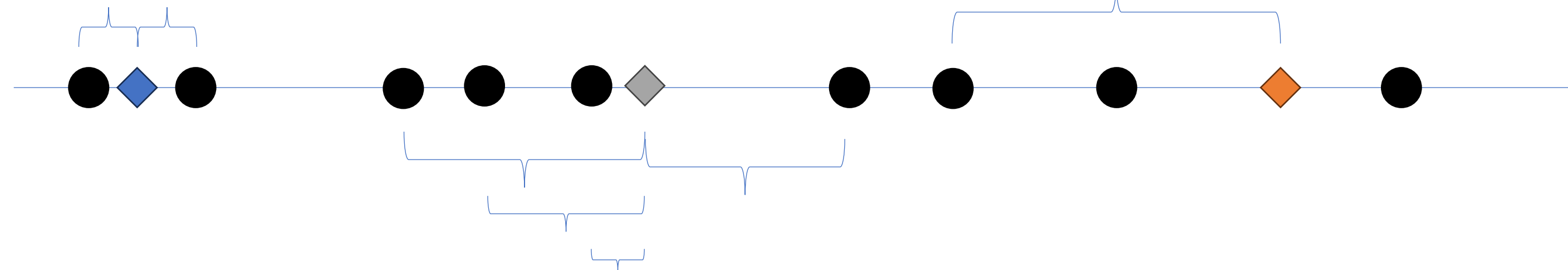
Ahora ponemos
aleatoriamente donde
inician los 3 centroides



Bueno para este ejemplo
no serán aleatorios, pero
supongamos

Ejemplo simple

Ahora calculamos la
distancia de cada punto a
nuestro centroide

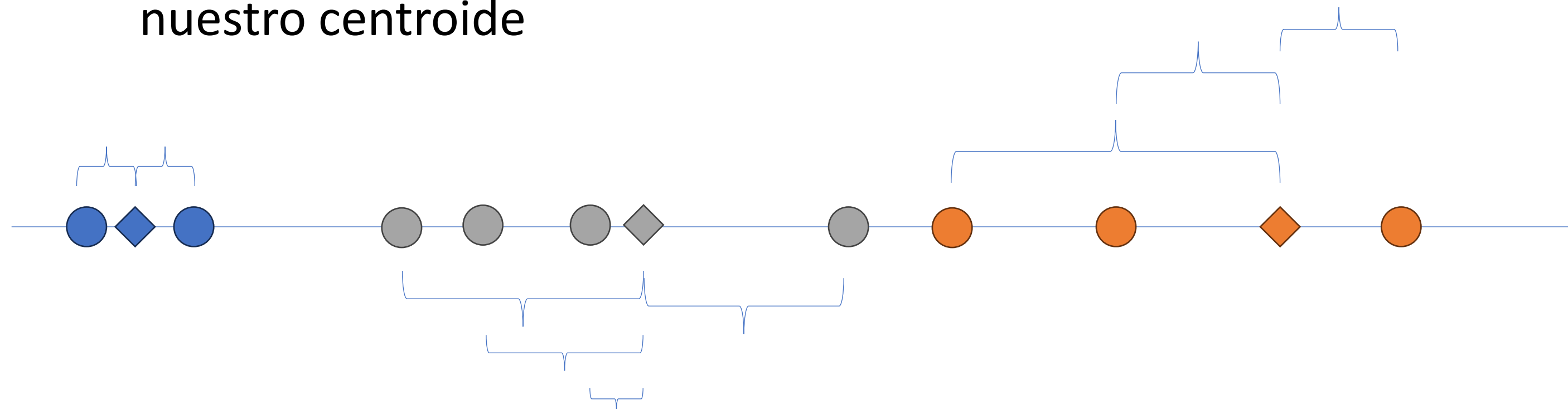


El más cercano toma
“posesión” del punto

Ejemplo simple

Ahora calculamos la
distancia de cada punto a
nuestro centroide

El más cercano toma
“posesión” del punto



Ejemplo simple

Luego se actualiza el
“centro” teniendo en
cuenta la media

Y recalculamos el punto del
centroide



Ejemplo simple

Es posible que al recalcular
la posición de los
centroides se cambie la
“posesión” de algunos
datos

Esto podría hacer que
tengamos que recalcular la
media nuevamente



Ejemplo simple

Si al repetir los pasos no
hay cambios podemos
dejar así

Y estos serían nuestros
Clusters



Ejemplo simple

¿Si era lo que nos
imaginábamos?

¿Cómo cambiaría con 4
Clusters?



Ejemplo simple

¿Si era lo que nos
imaginábamos?

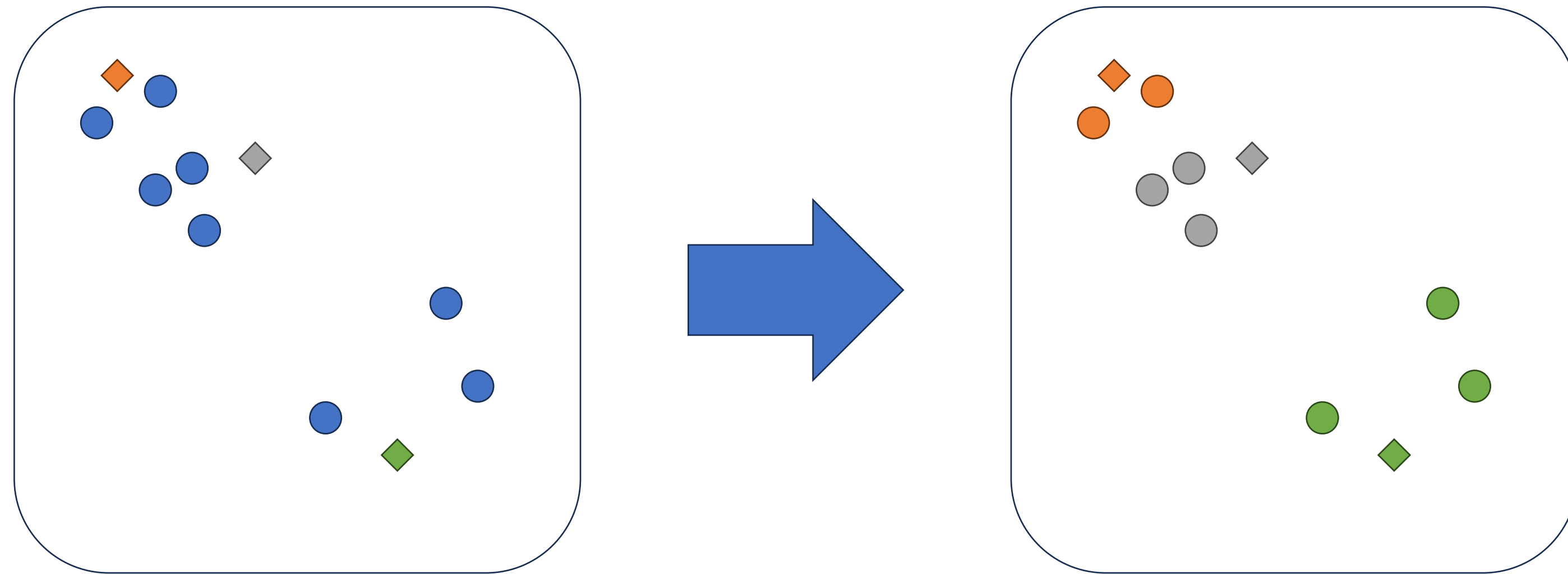
¿Cómo cambiaría con 4
Clusters?



¿Cambiaría si el centroide
empieza en otra parte?

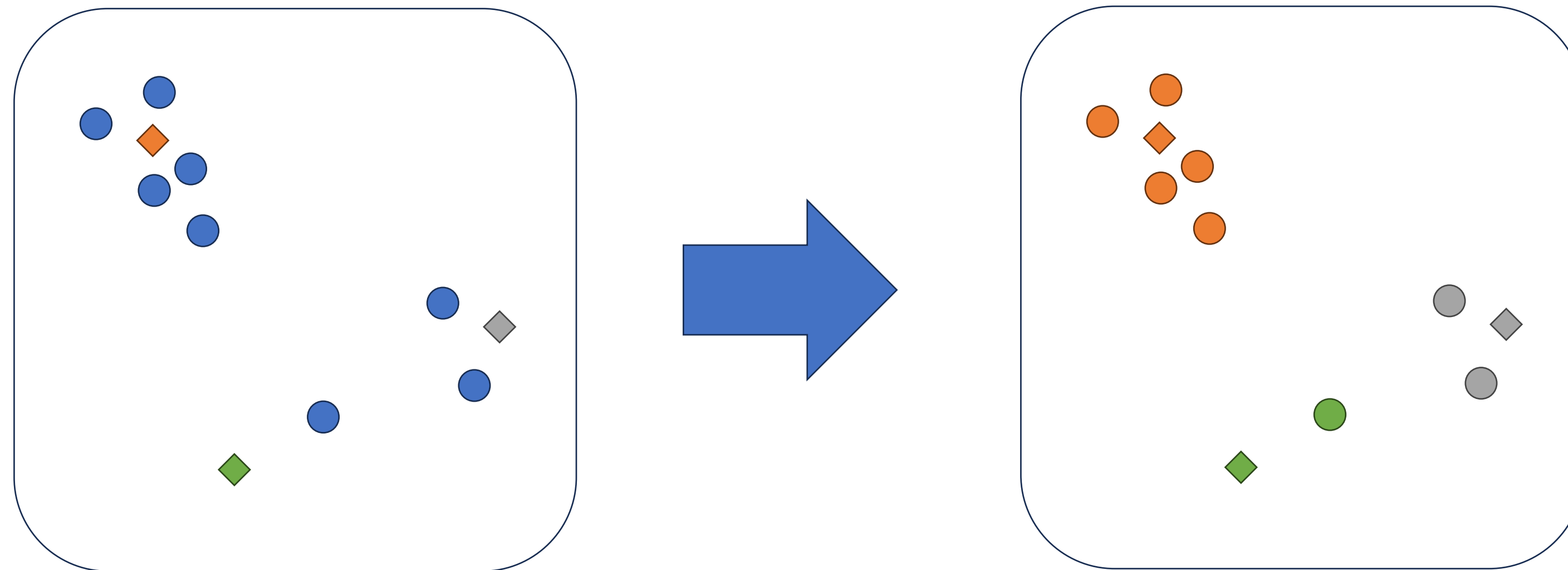
Evaluando la calidad de los clusters

Uno de los problemas iniciales que tiene este modelo es que la posición de los centroides puede cambiar todo.



Evaluando la calidad de los clusters

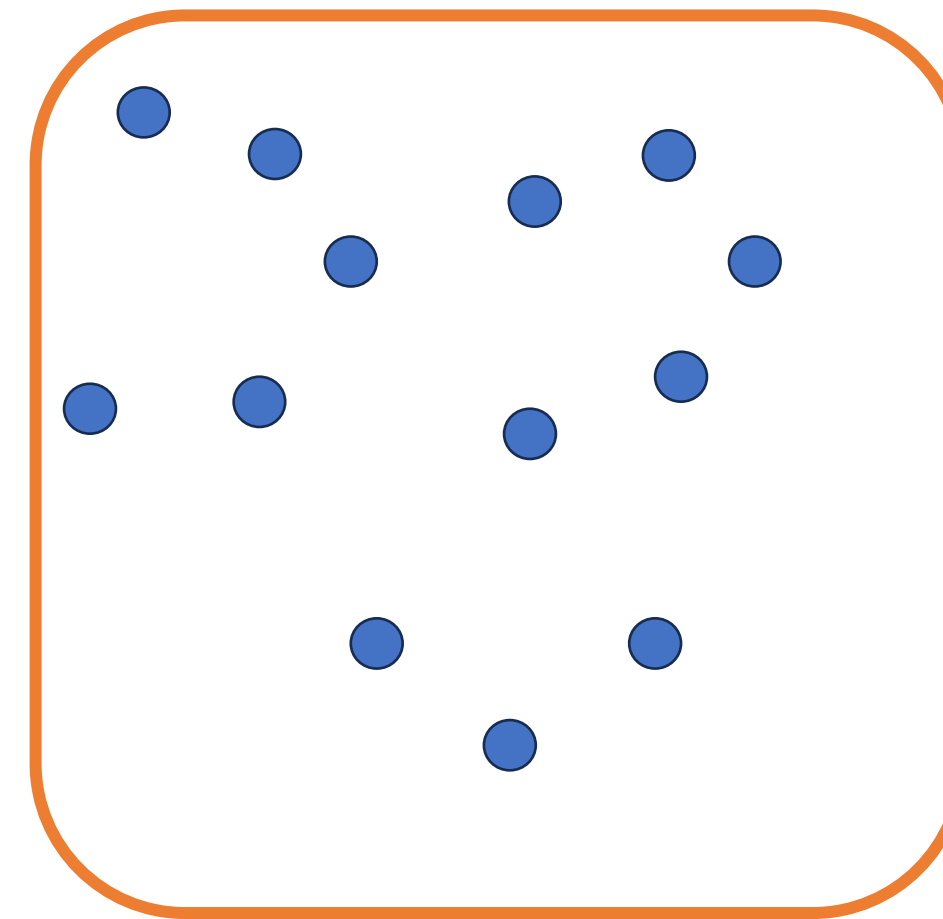
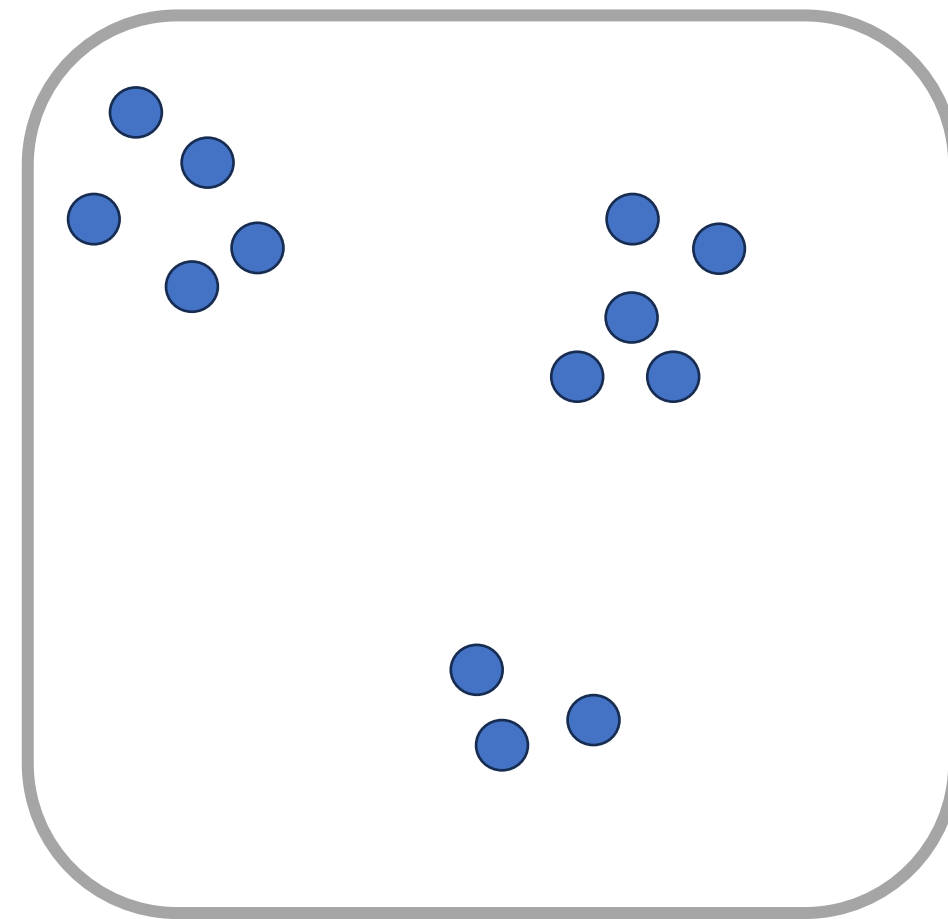
Uno de los problemas iniciales que tiene este modelo es que la posición de los centroides puede cambiar todo.



Para escoger la mejor agrupación

El objetivo del algoritmo es minimizar la diferencia intra-clusters y maximizar la diferencia inter-clusters.

Esto significa, definir y tener grupos con rasgos bien definidos que se diferencien fácilmente de los otros grupos.

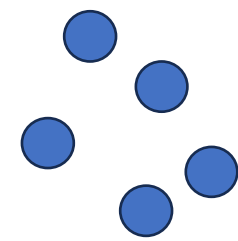


Inercia nos dice qué tan diferentes son

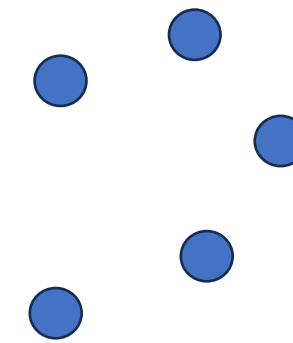
Igualdad de Fisher= Inercia total = Inercia inter + Inercia intra

¿Qué es la inercia?

Es la suma de cuadrados de cada punto con el centroide.



Nivel de inercia bajo



Nivel de inercia alto

¿En que afecta el número de clusters?

El número de Clusters es uno de los componentes principales del algoritmo.

Si creamos demasiados vamos a tener grupos muy específicos que pueden partir los datos en Clusters innecesarios.

Si creamos muy pocos no vamos a poder seccionar los datos de una forma correcta.

Lo ideal es crear una cantidad óptima de Clusters que permita extraer la mayor cantidad de información relevante.

Conocimiento del negocio

El número de Clusters se puede definir por el conocimiento del negocio.

Imaginemos que estamos usando este algoritmo para encontrar la ubicación donde podemos poner unas franquicias, buscando el punto óptimo del centroide como la sede.

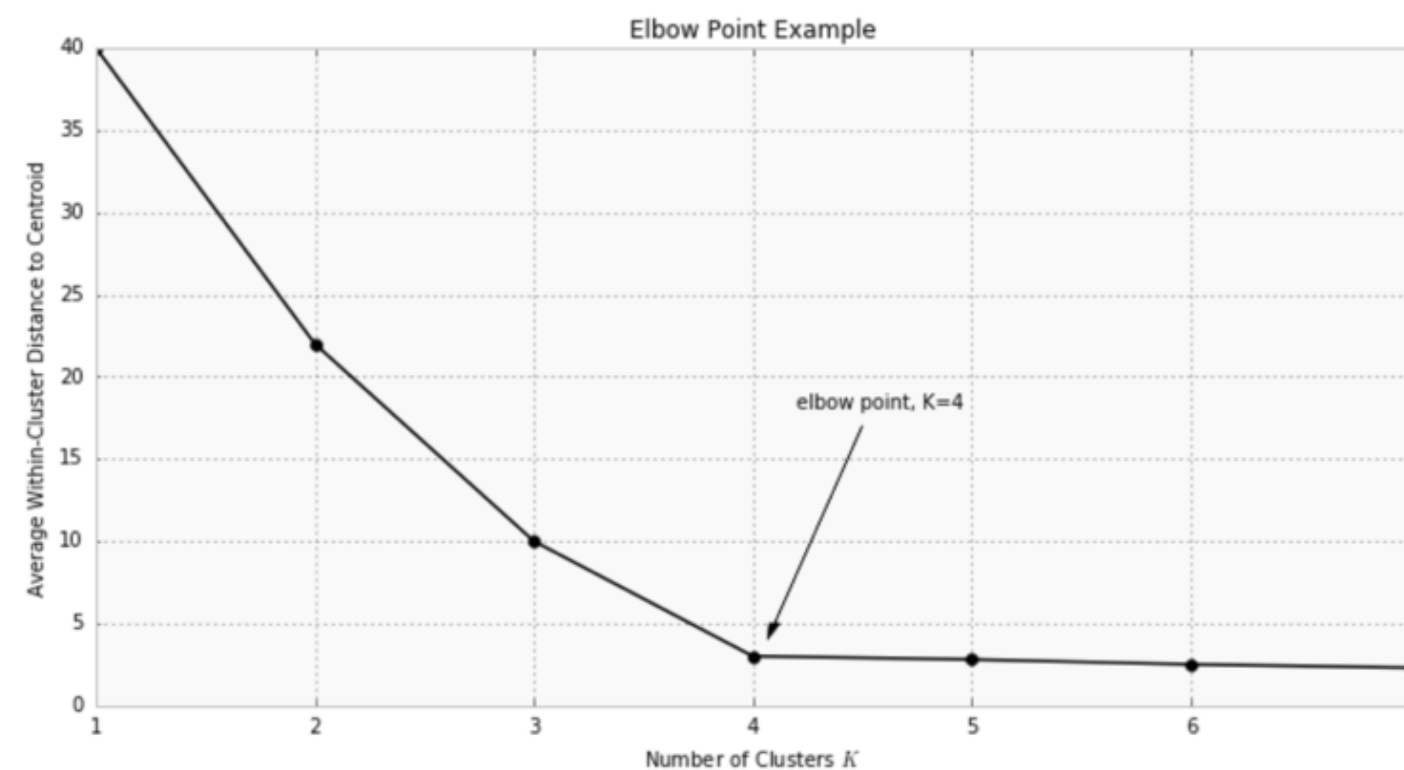
Cada dato puede representar puntos de interés a los que queremos llegar.

Pero si solo podemos abrir 3 sedes tenemos que limitar el número de Clusters.

Método del Codo

Otro método y tal vez el más conocido es el método del codo.
¿En qué consiste?

Vamos a crear el modelo de forma iterativa aumentando la cantidad de Clusters y midiendo la inercia en cada modelo.



Jerárquica

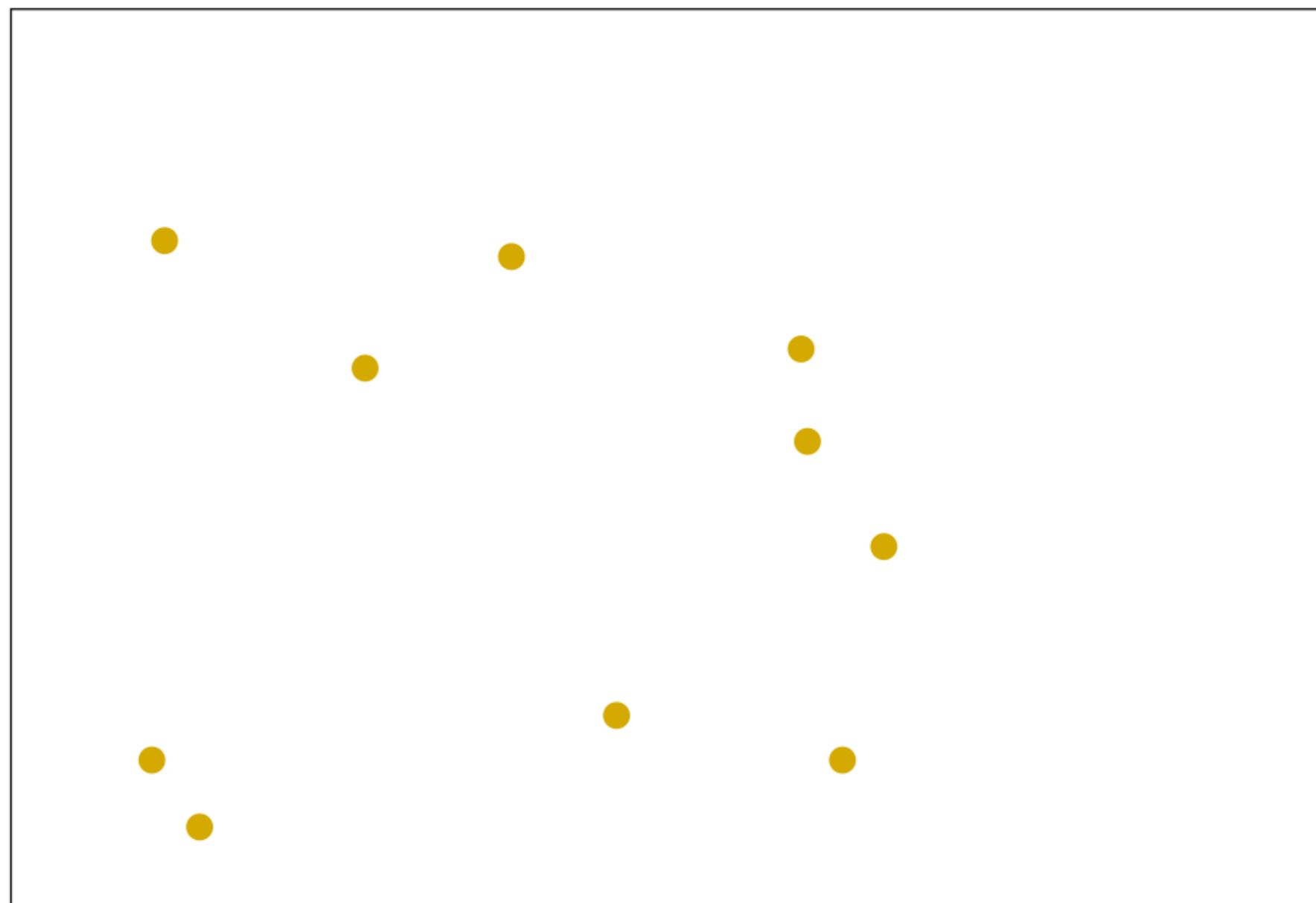


Intuición

- Armamos grupos con los datos más cercanos entre sí, y vamos **agrandando los grupos** conectando cada vez datos (y sus grupos) más lejanos: "integrándolos a la conversación".
- ¿Cuándo paramos? Cuando estemos conectando grupos que ya están "muuuy lejos".

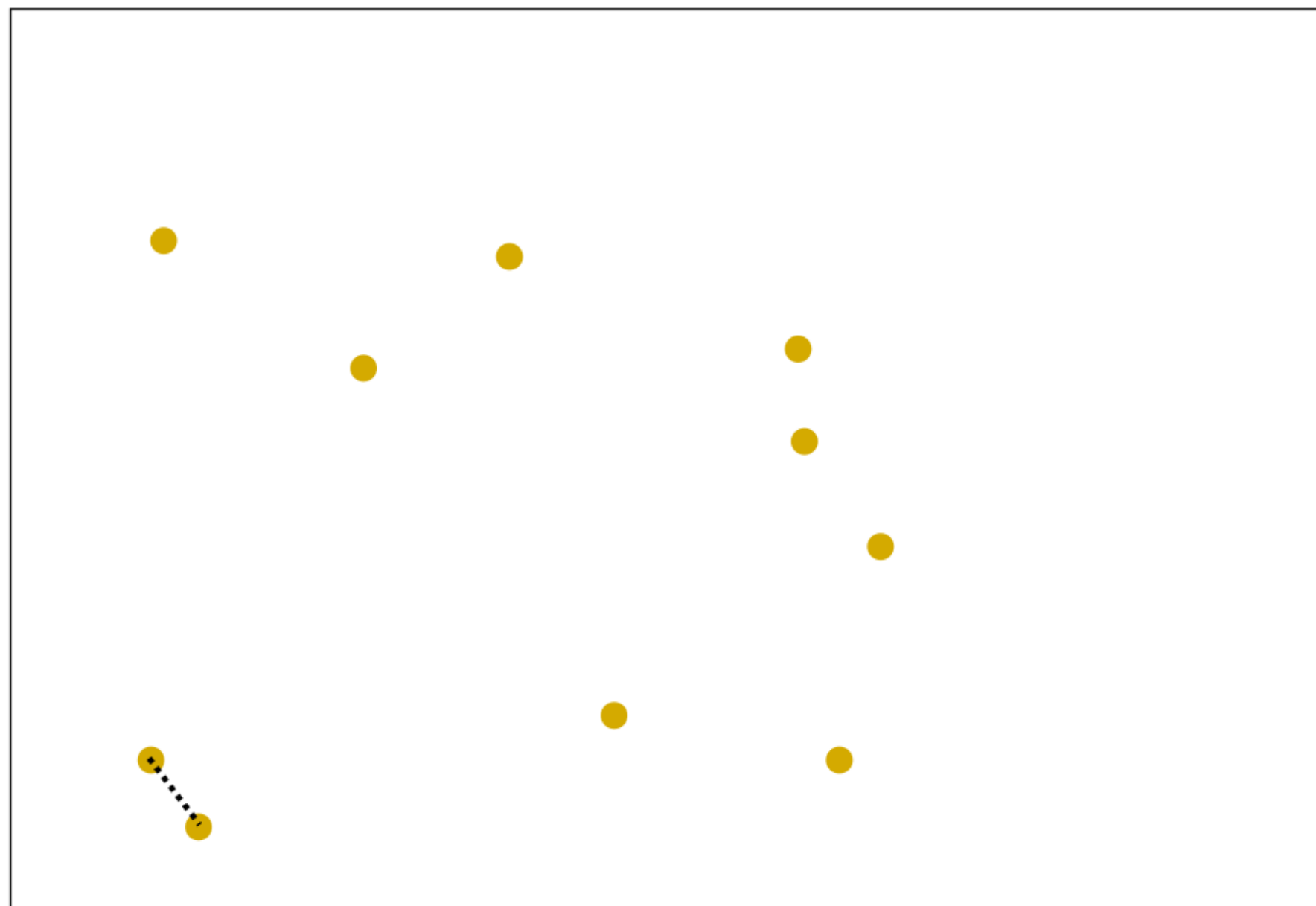


Agrupando



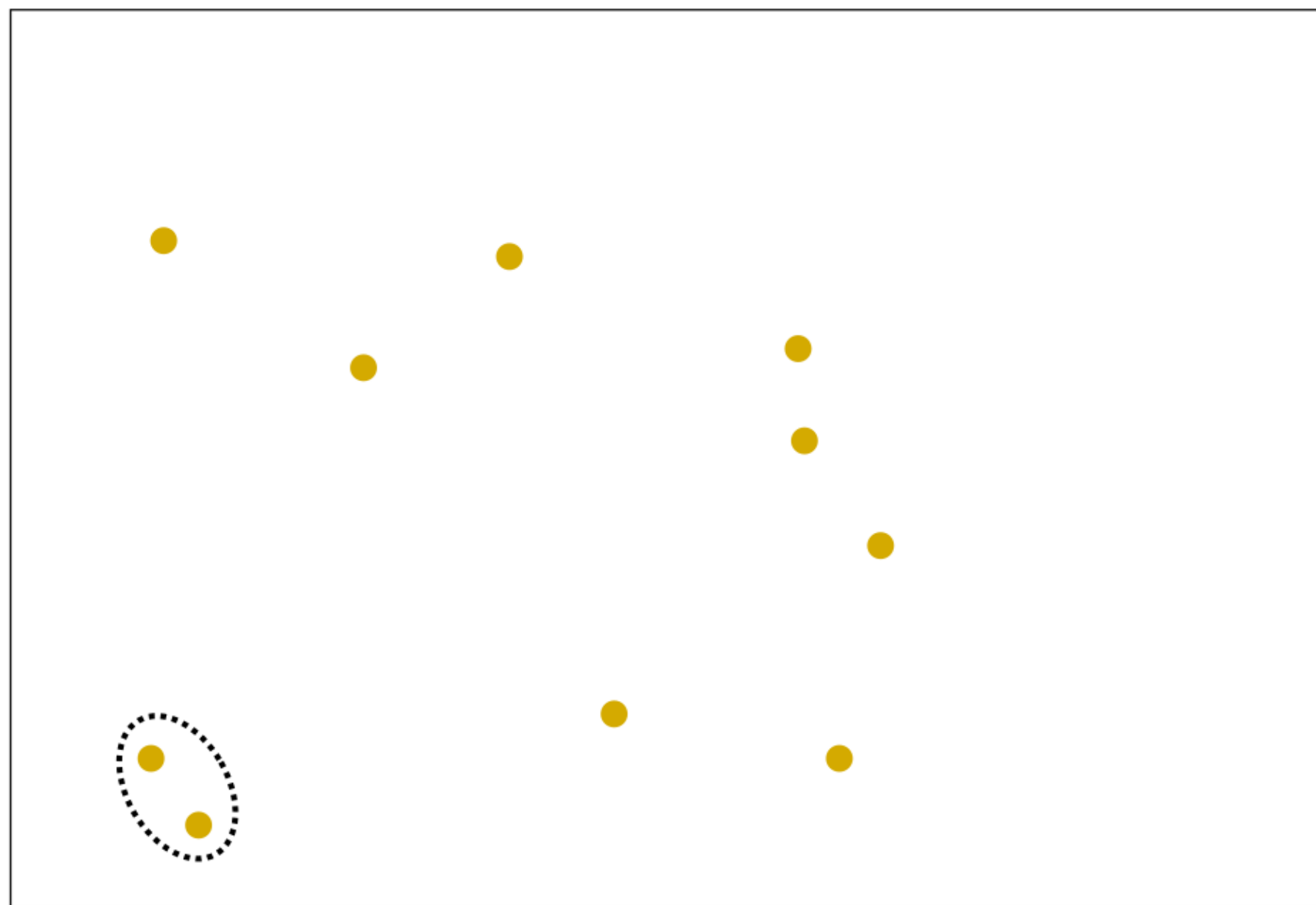
Partimos de los datos por agrupar.

Agrupando



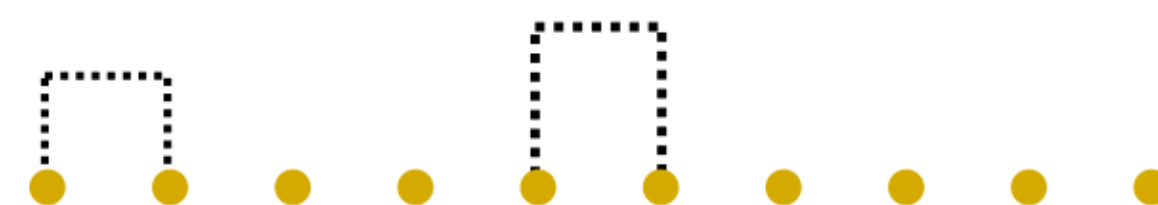
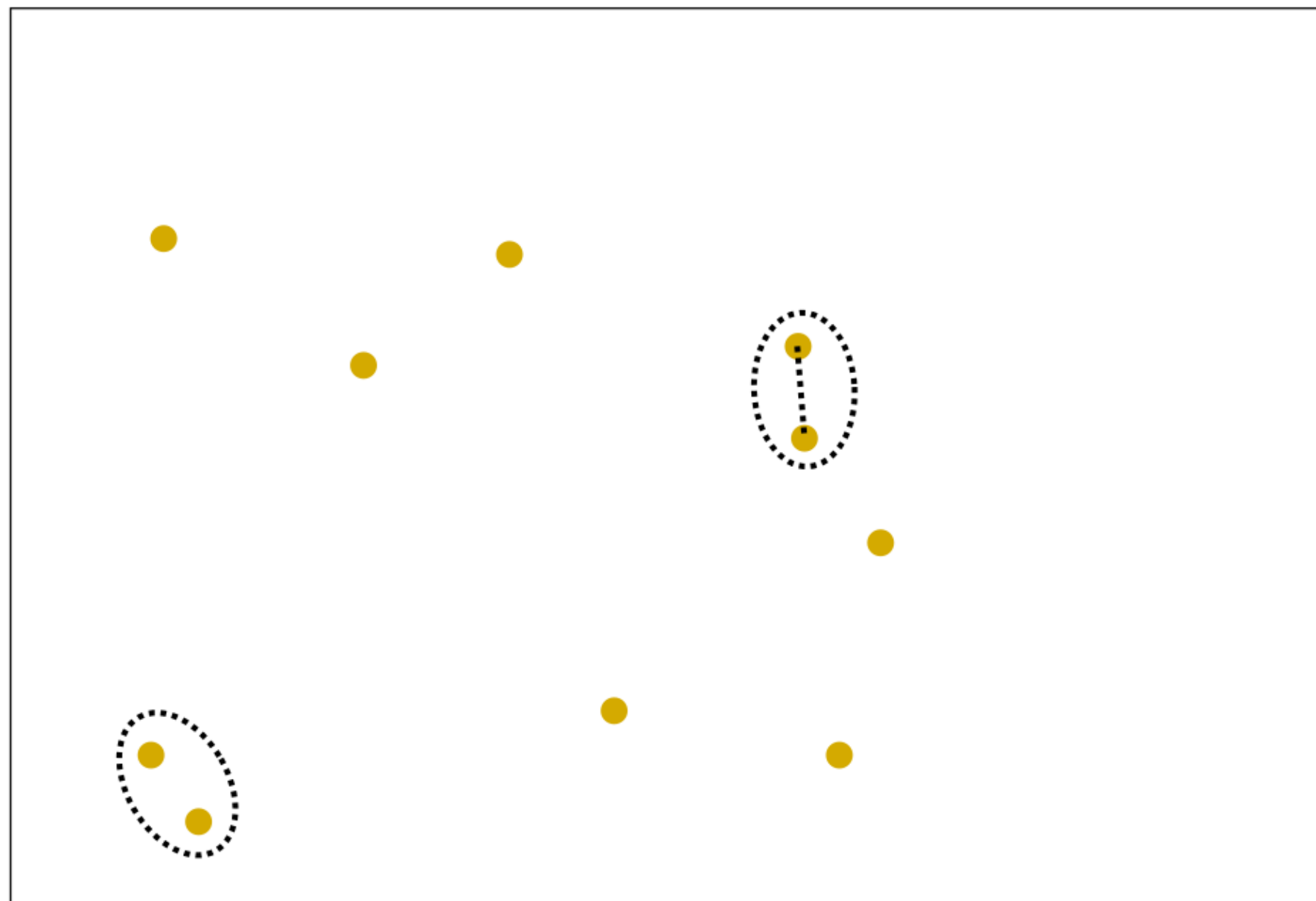
Identificamos el par de datos más cercano.

Agrupando



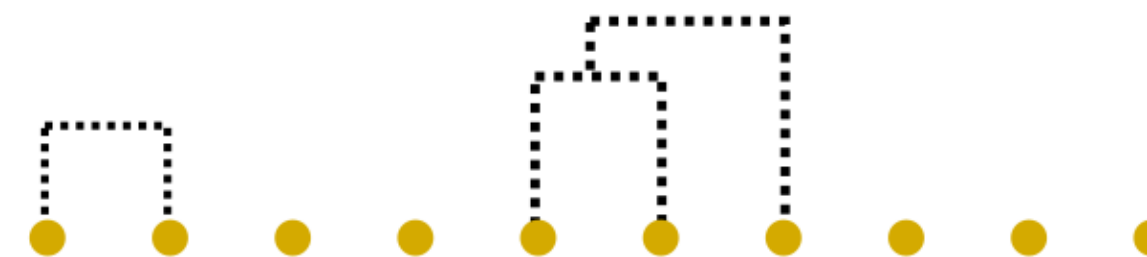
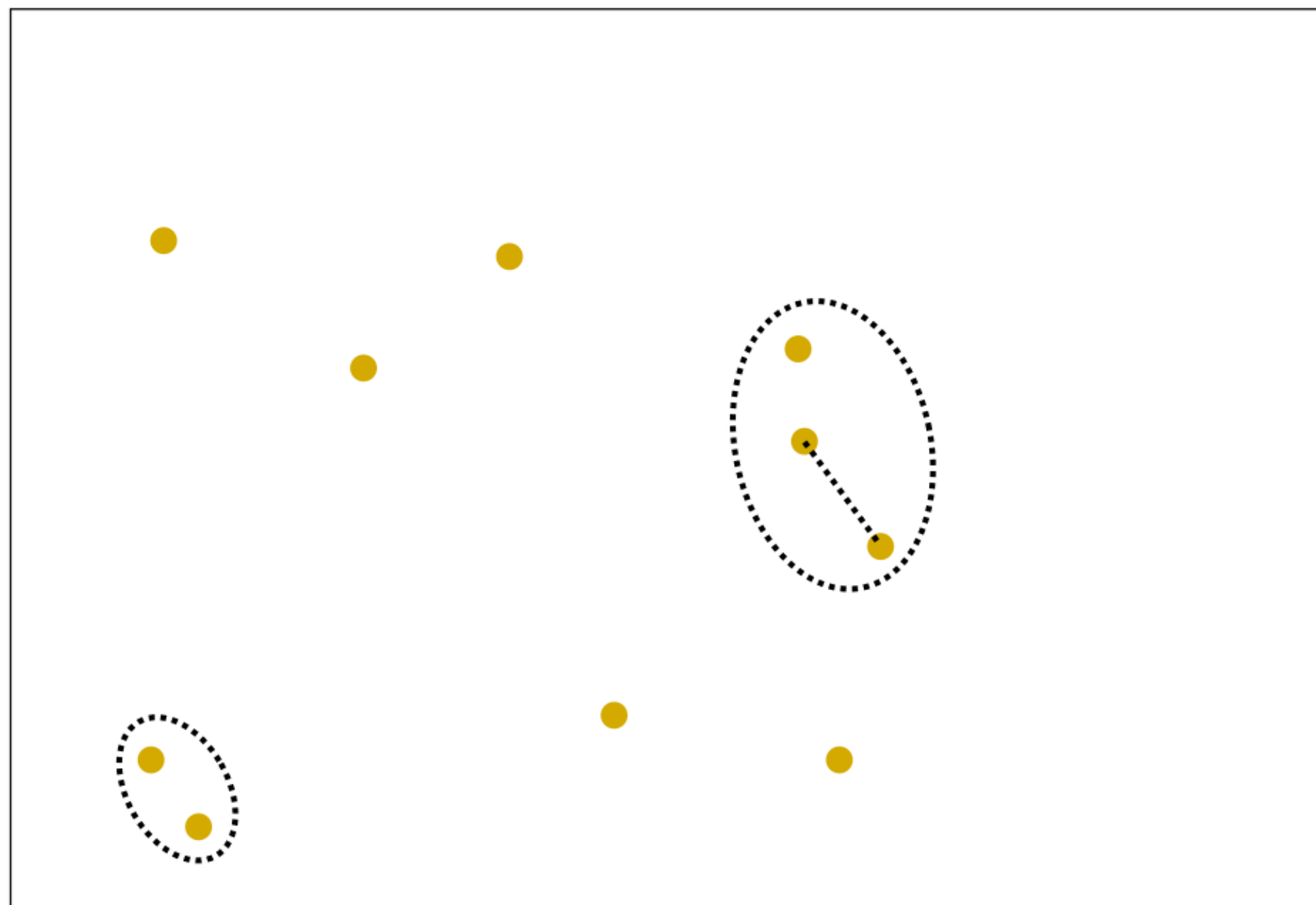
Representamos ese grupo graficando en el eje vertical la distancia entre los puntos.

Agrupando



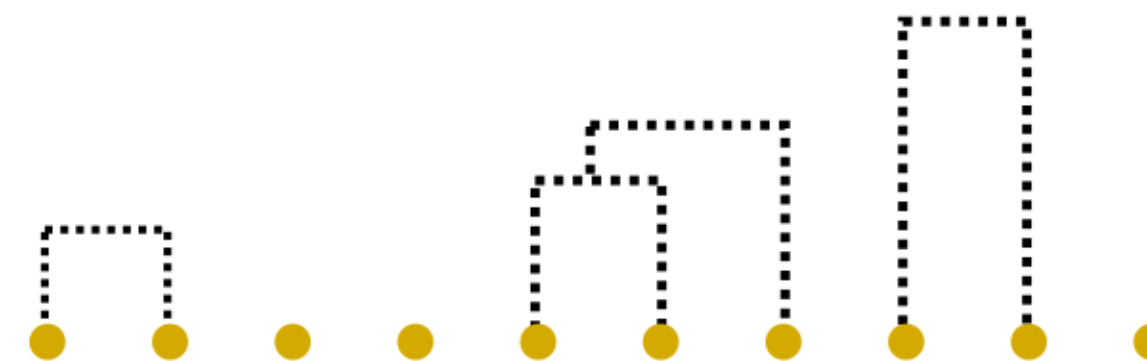
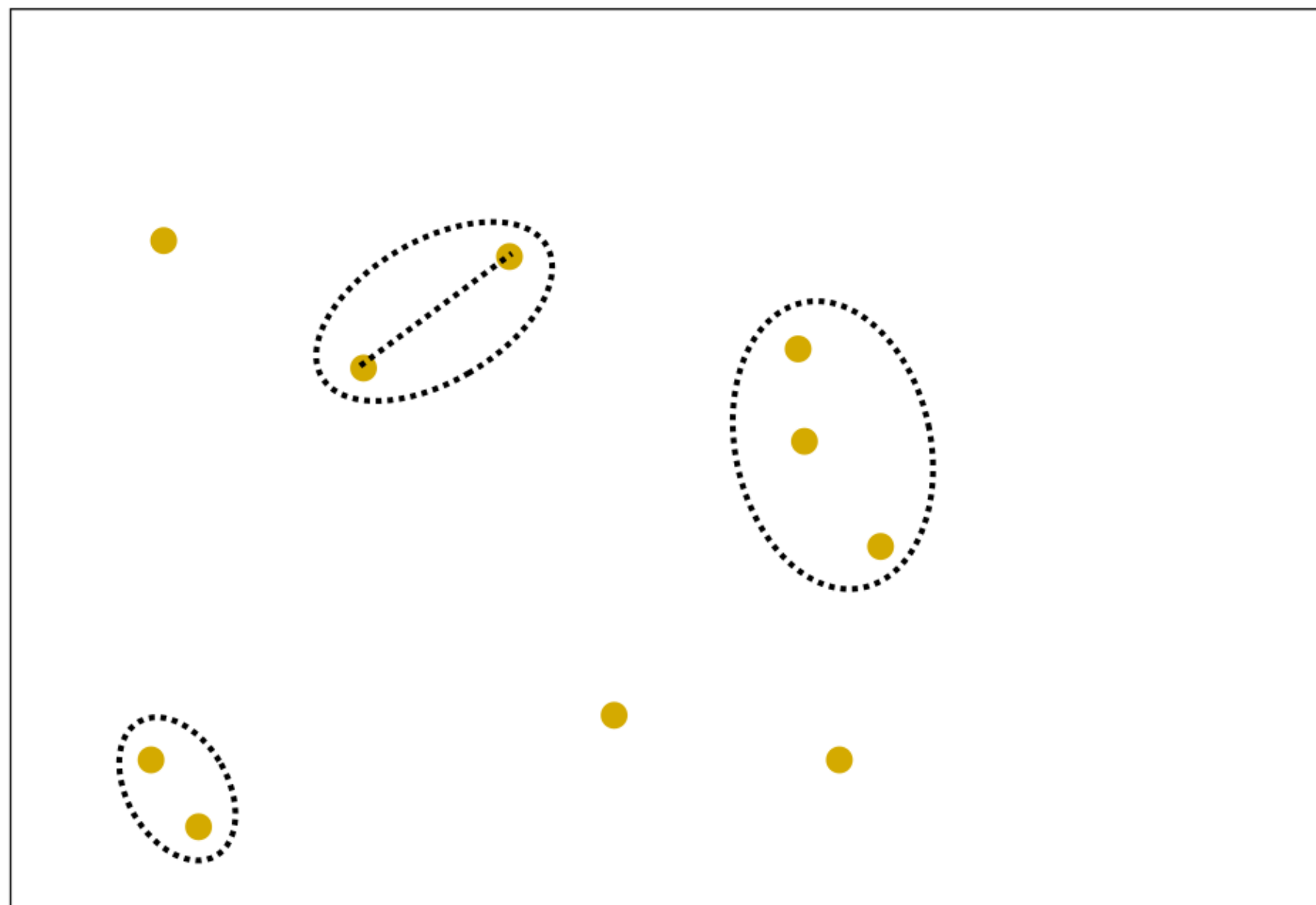
Hacemos lo mismo para el siguiente par de puntos más cercanos.

Agrupando



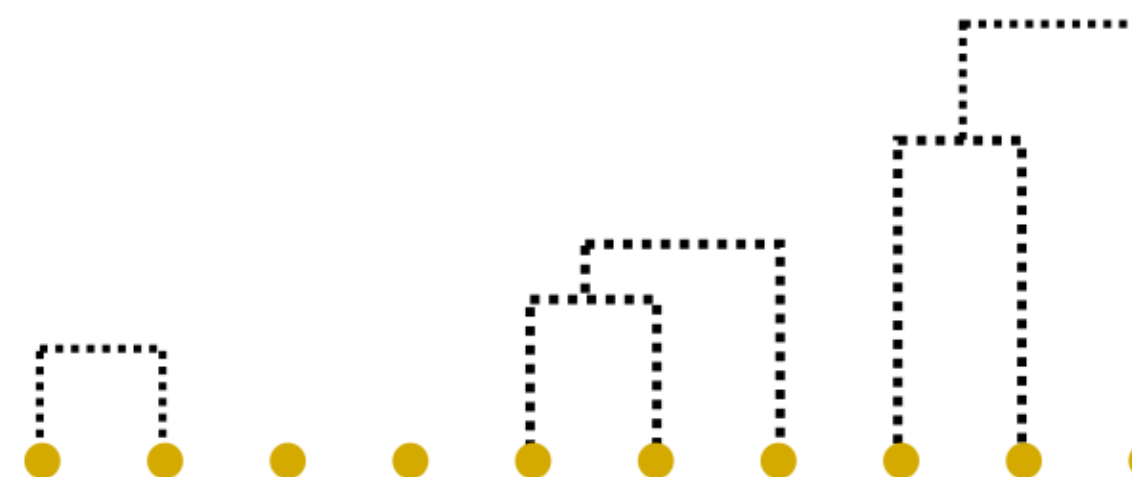
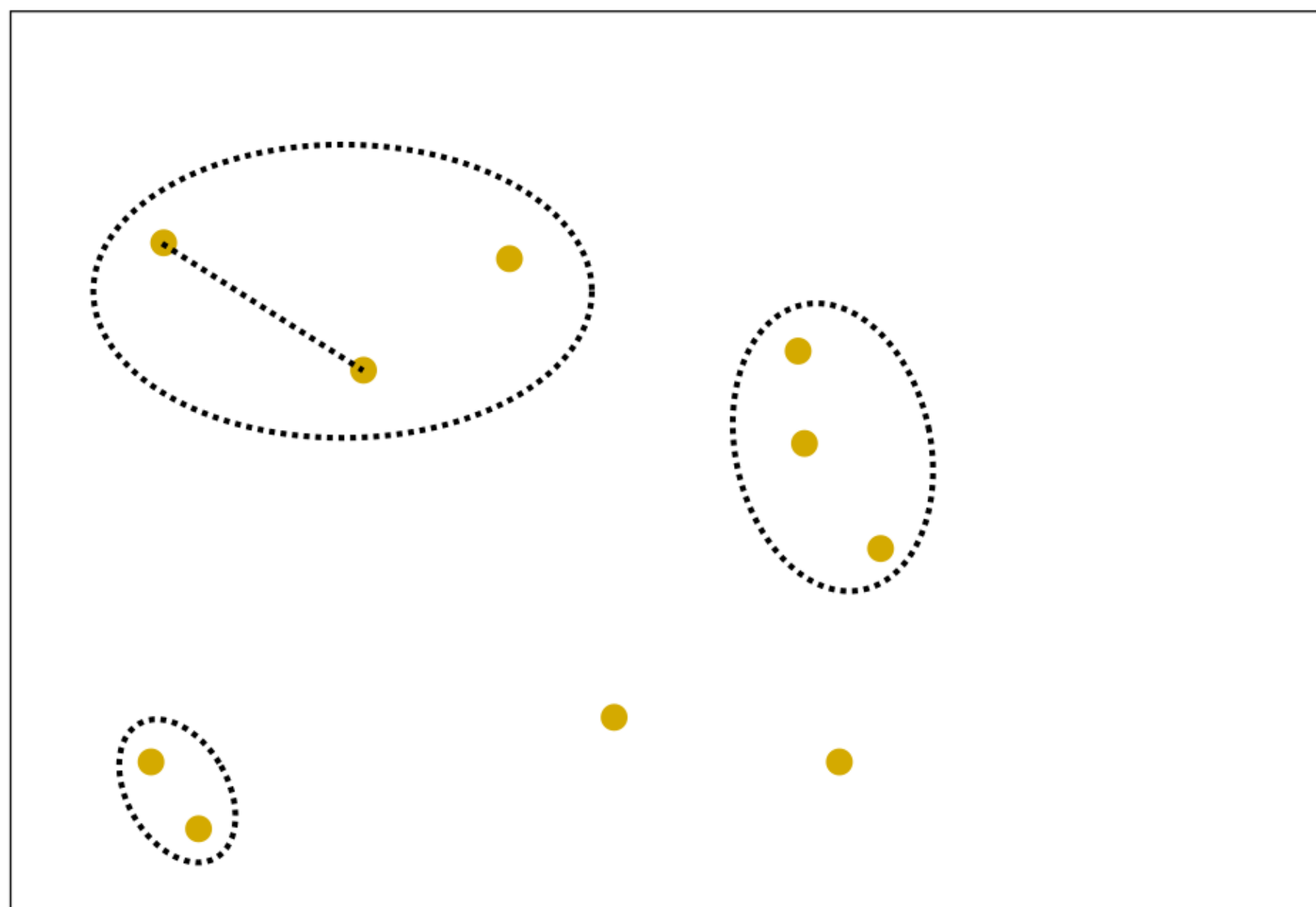
Anexamos el siguiente punto más cercano al clúster anterior. Y lo representamos con su altura.

Agrupando



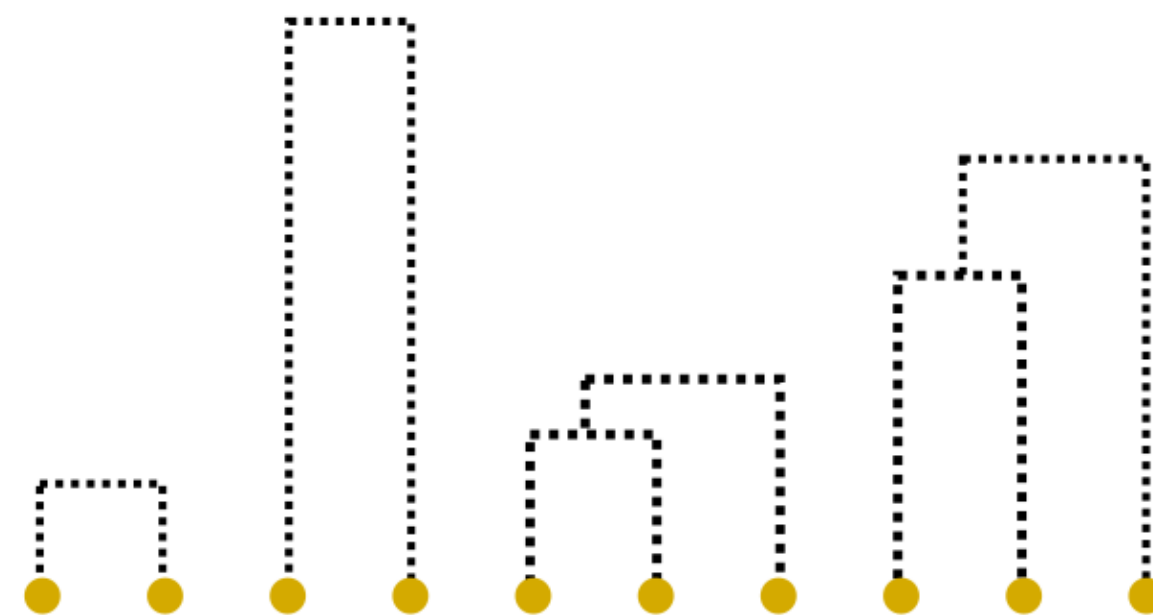
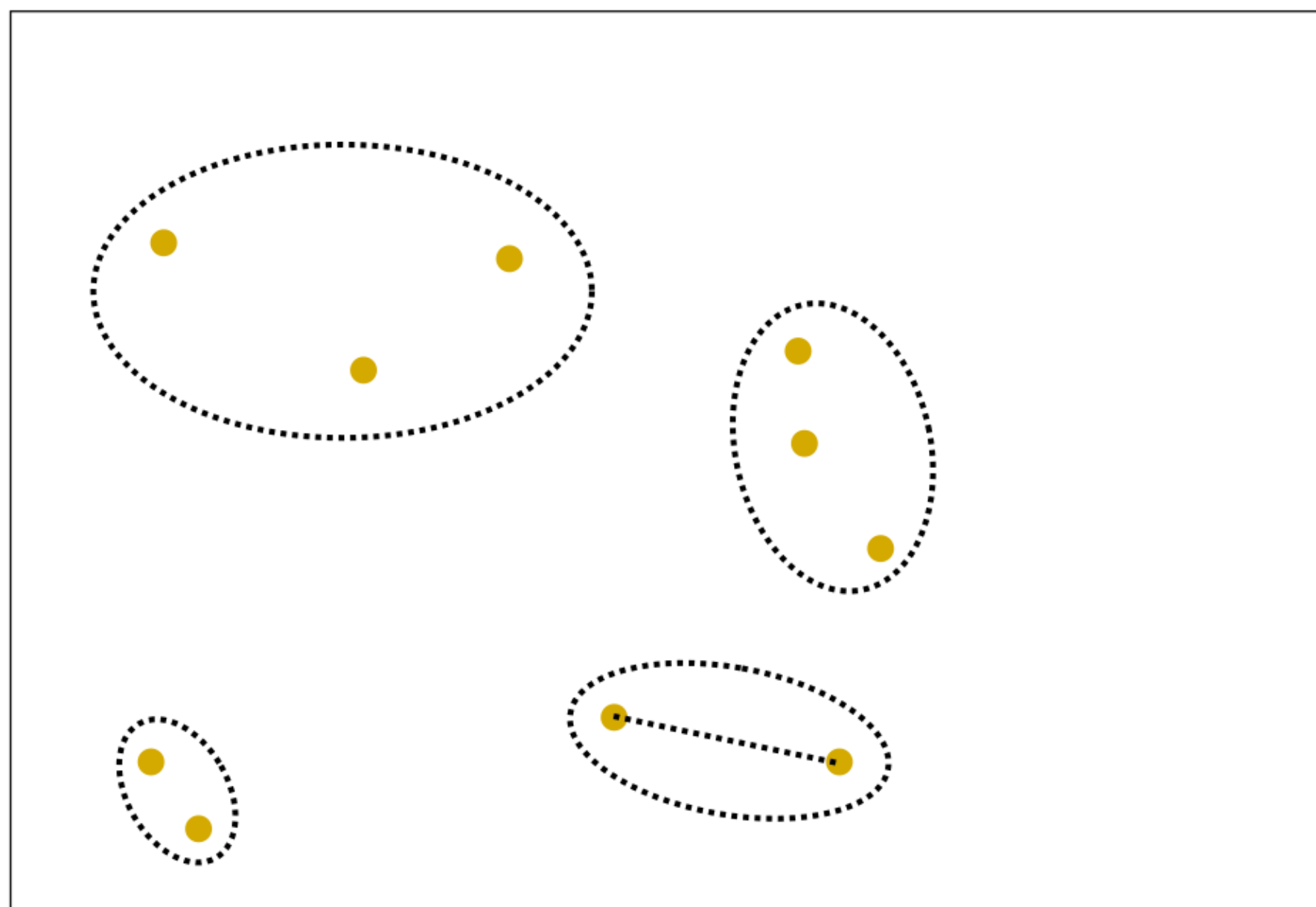
Seguimos adelante...

Agrupando



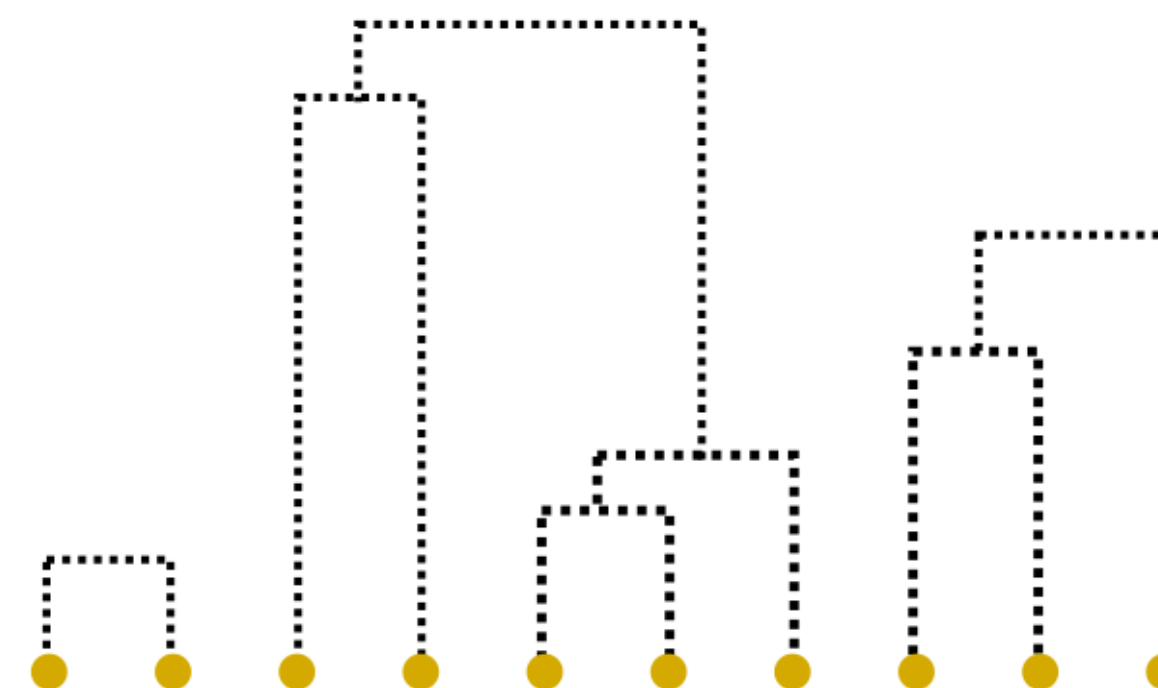
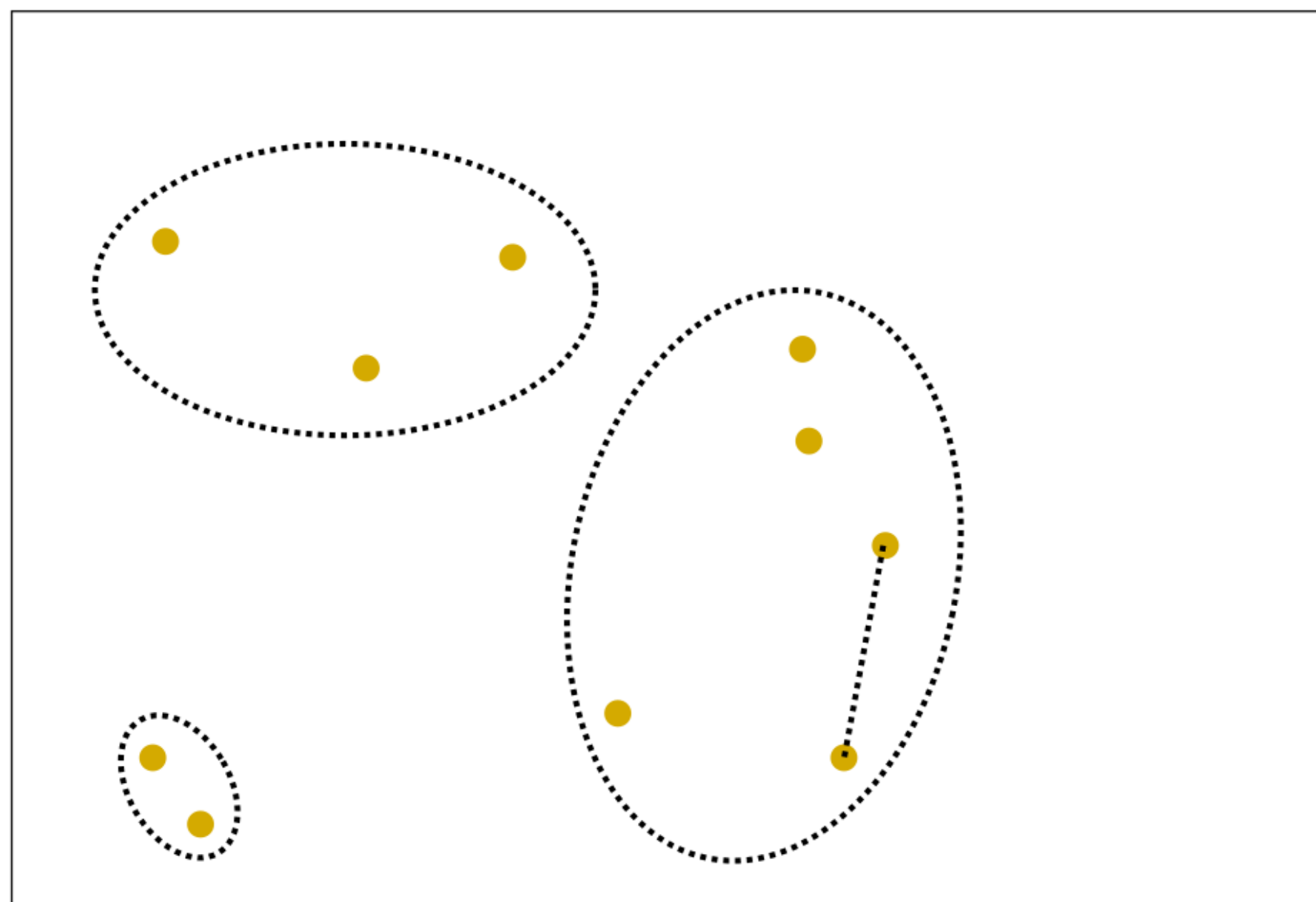
Seguimos adelante...

Agrupando



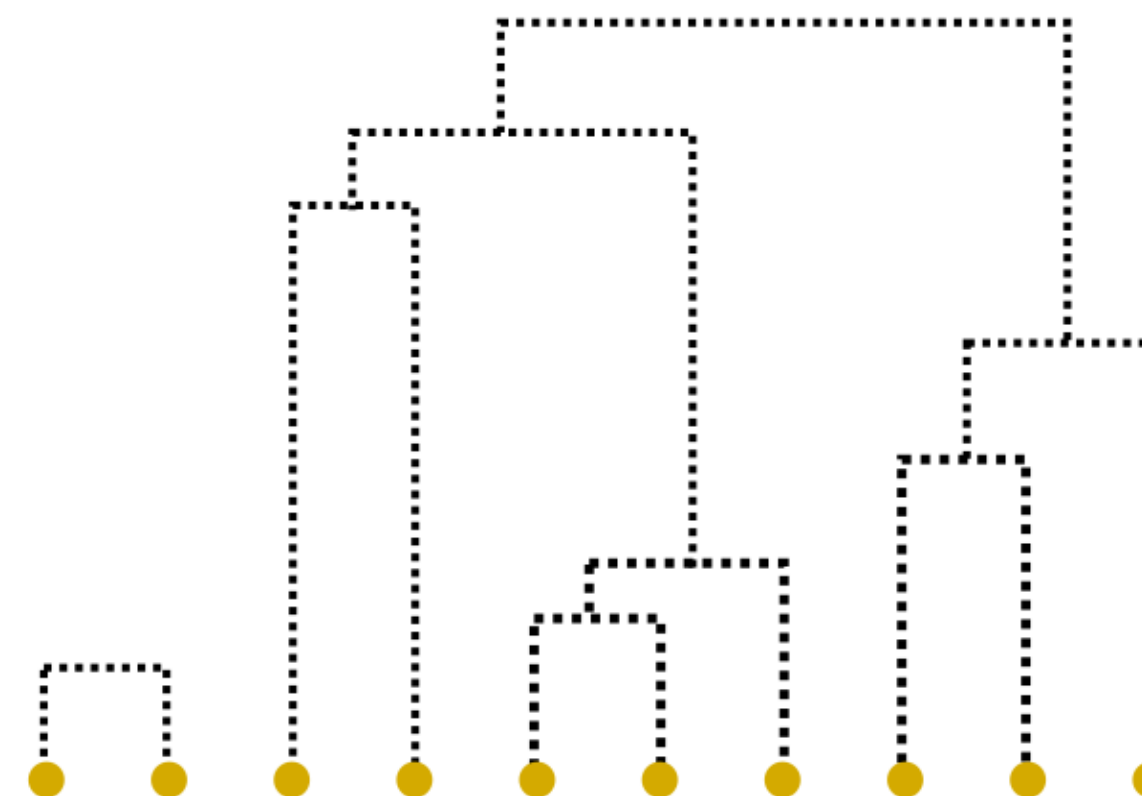
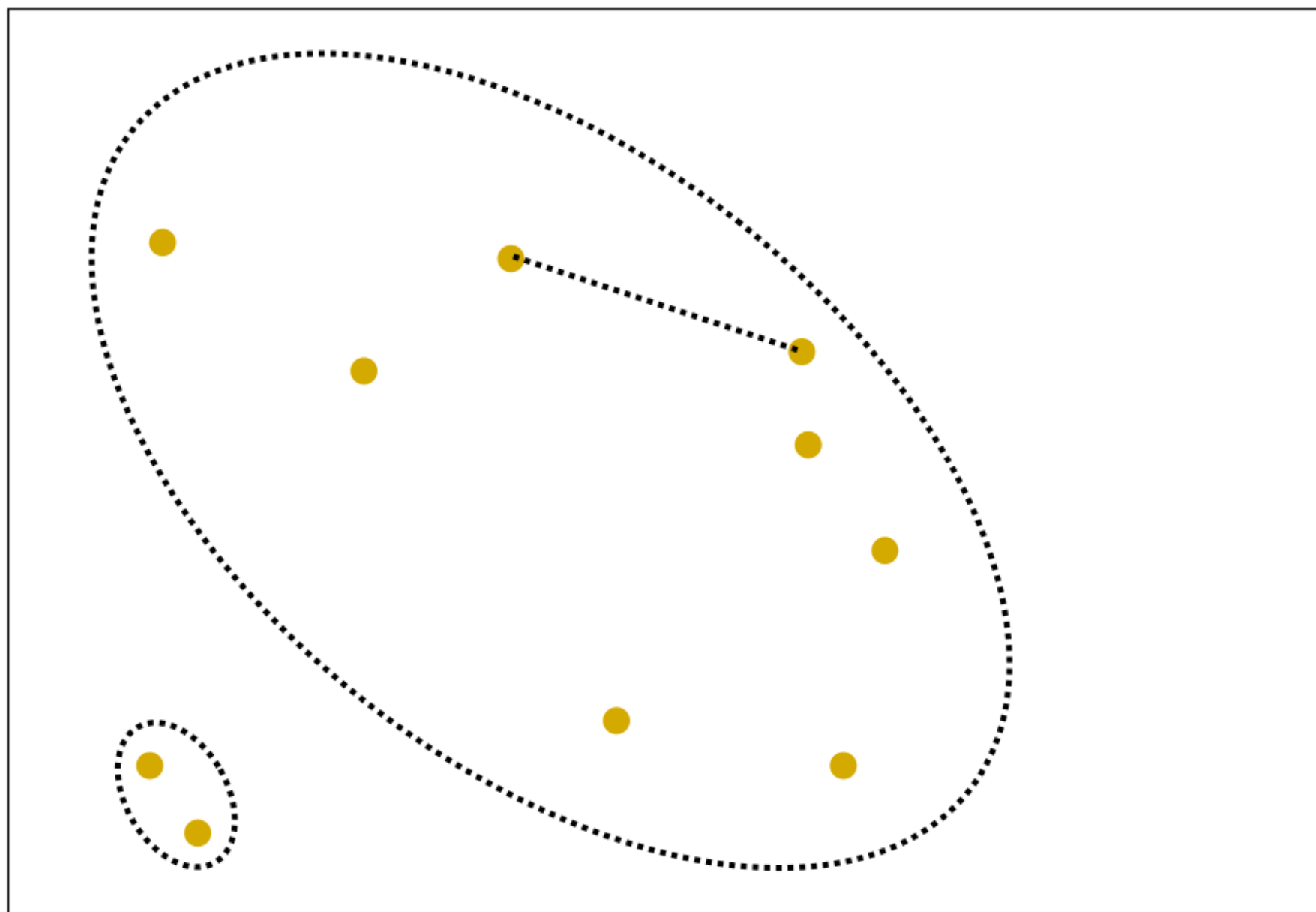
Seguimos adelante...

Agrupando



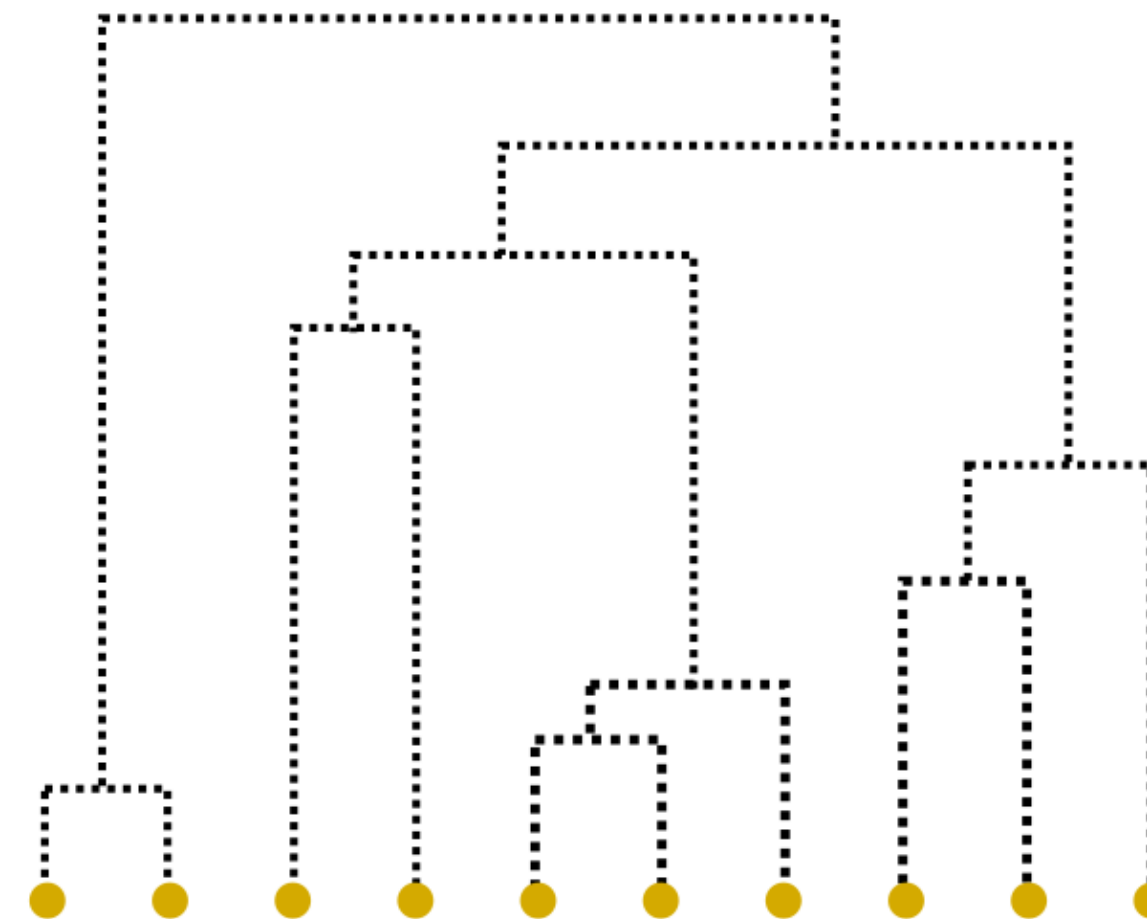
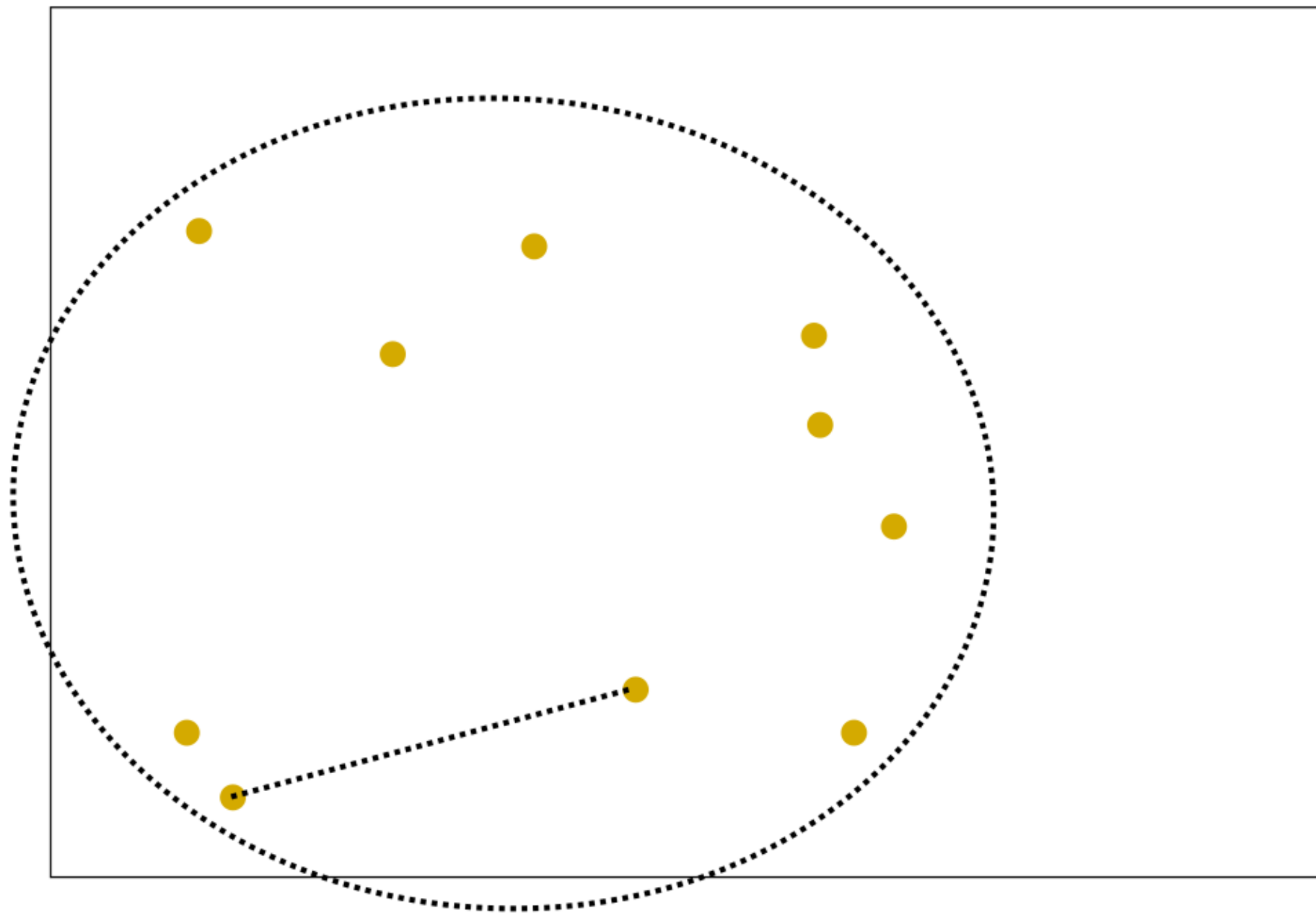
Seguimos adelante...

Agrupando



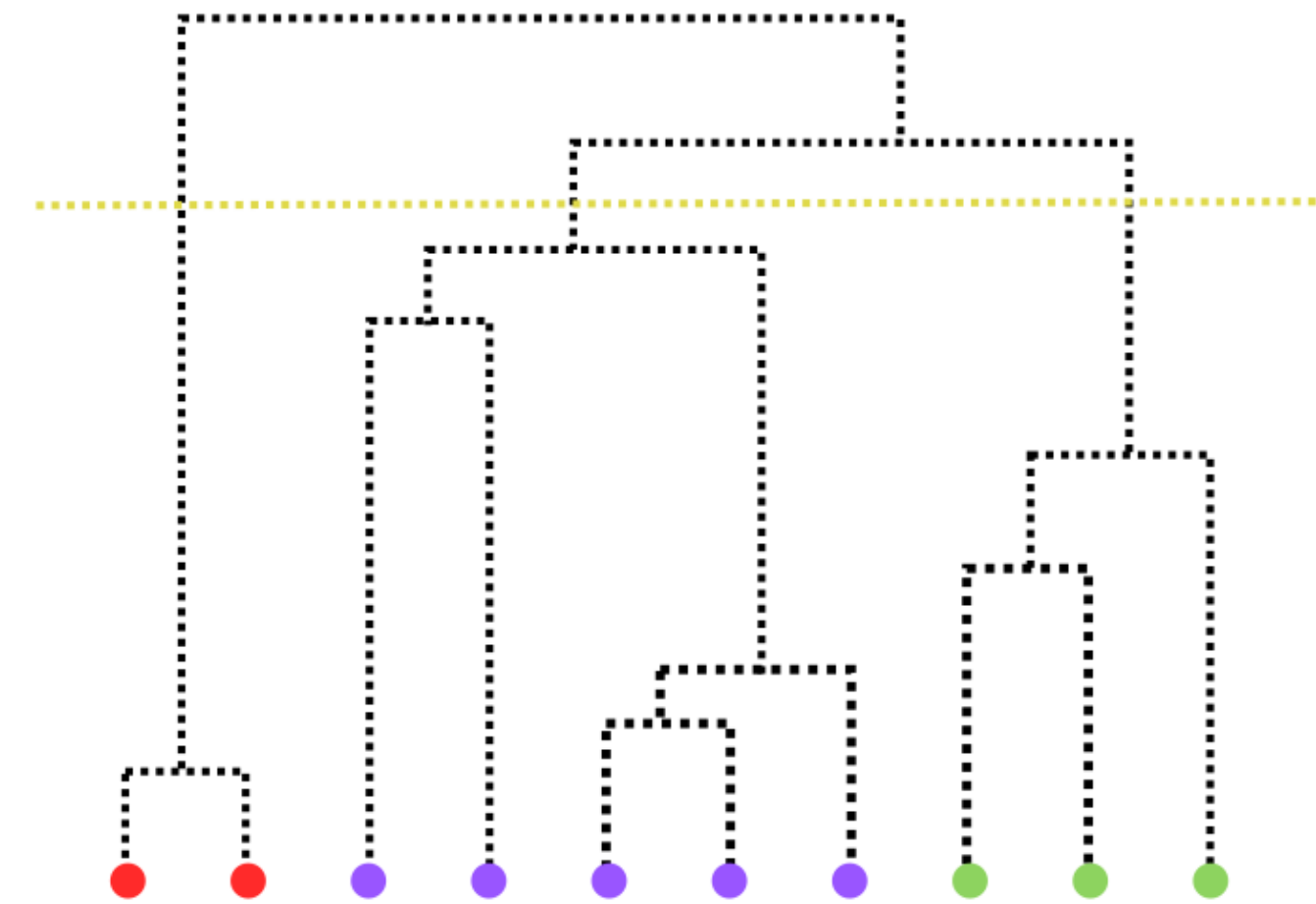
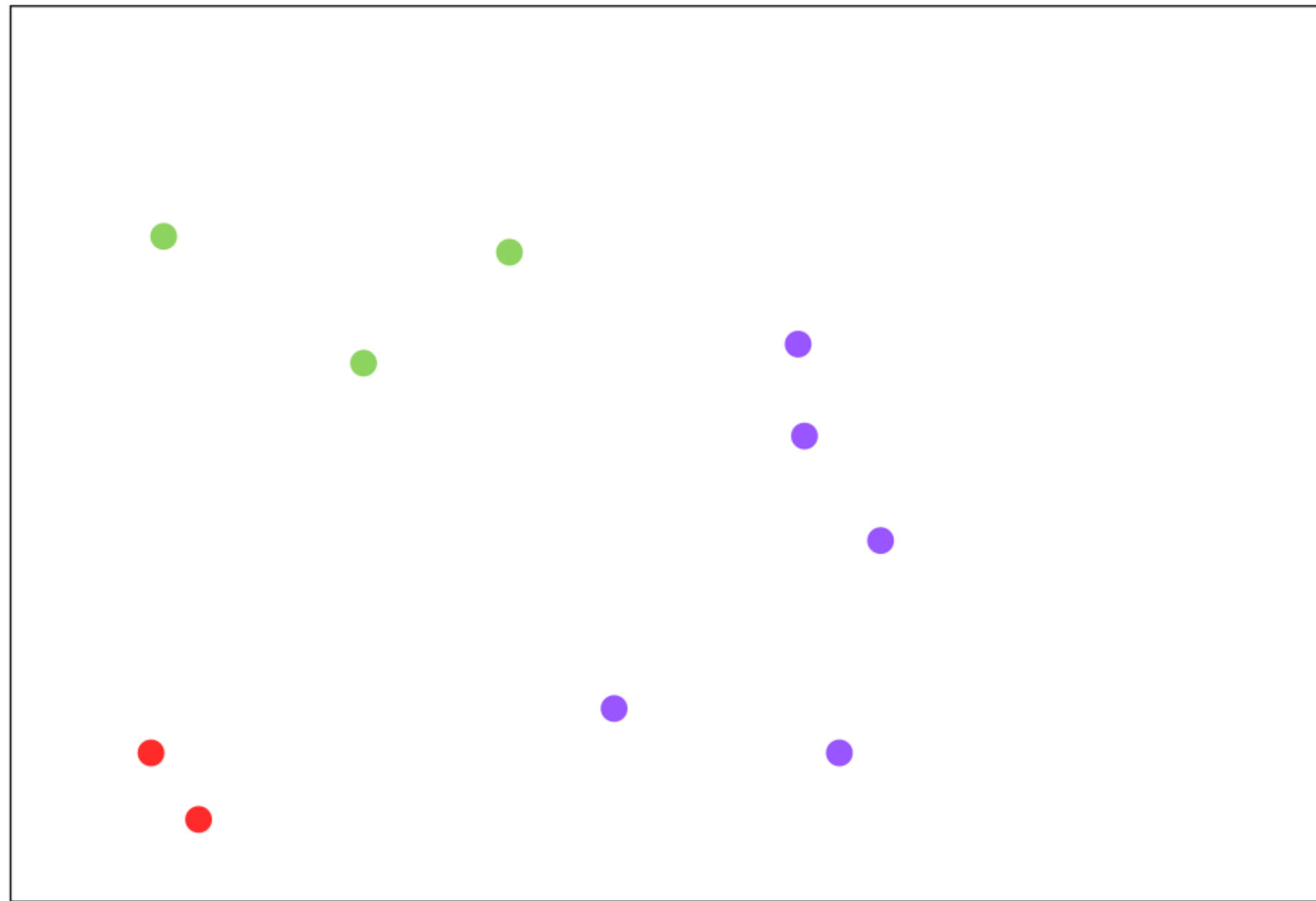
Seguimos adelante...

Agrupando



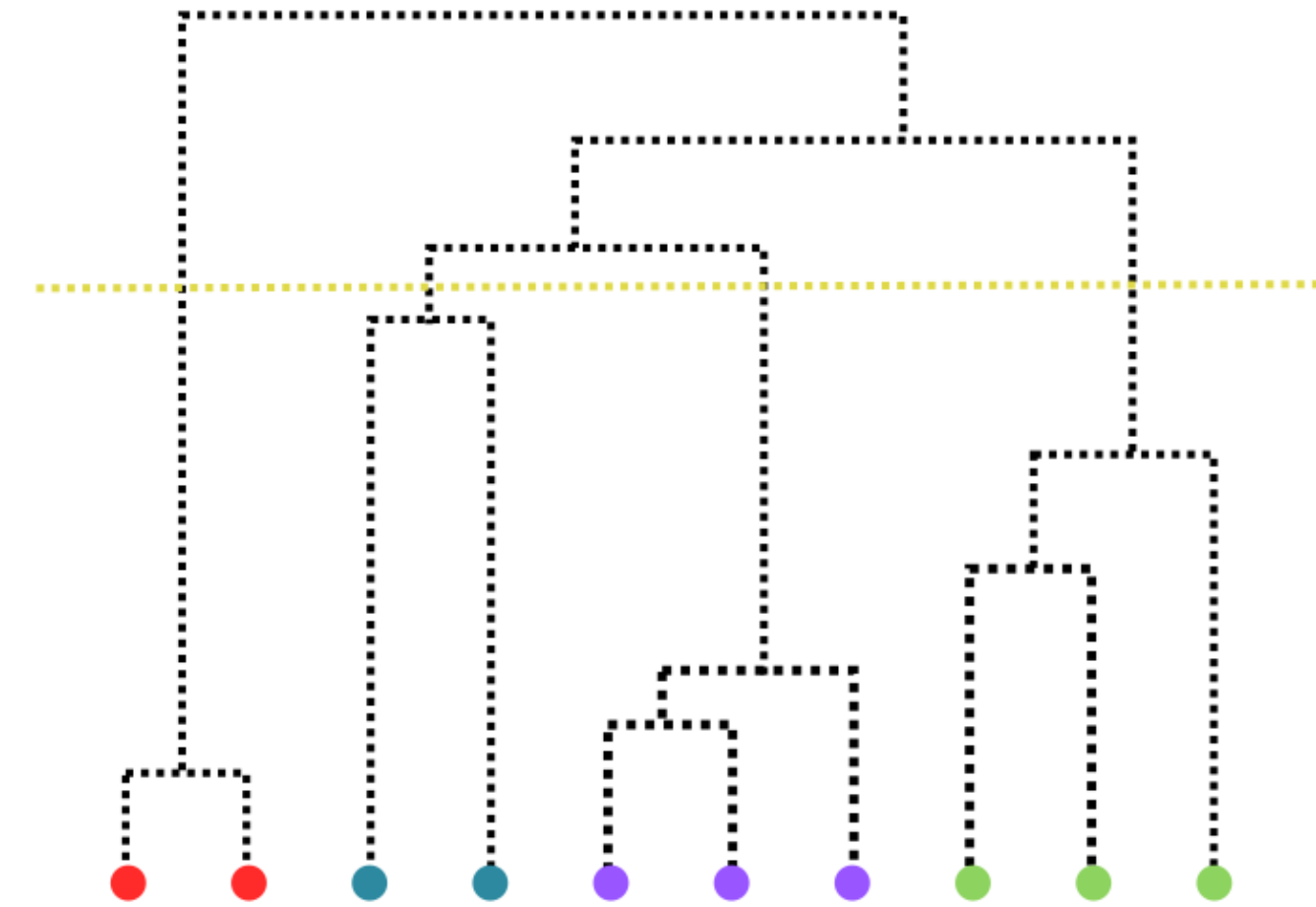
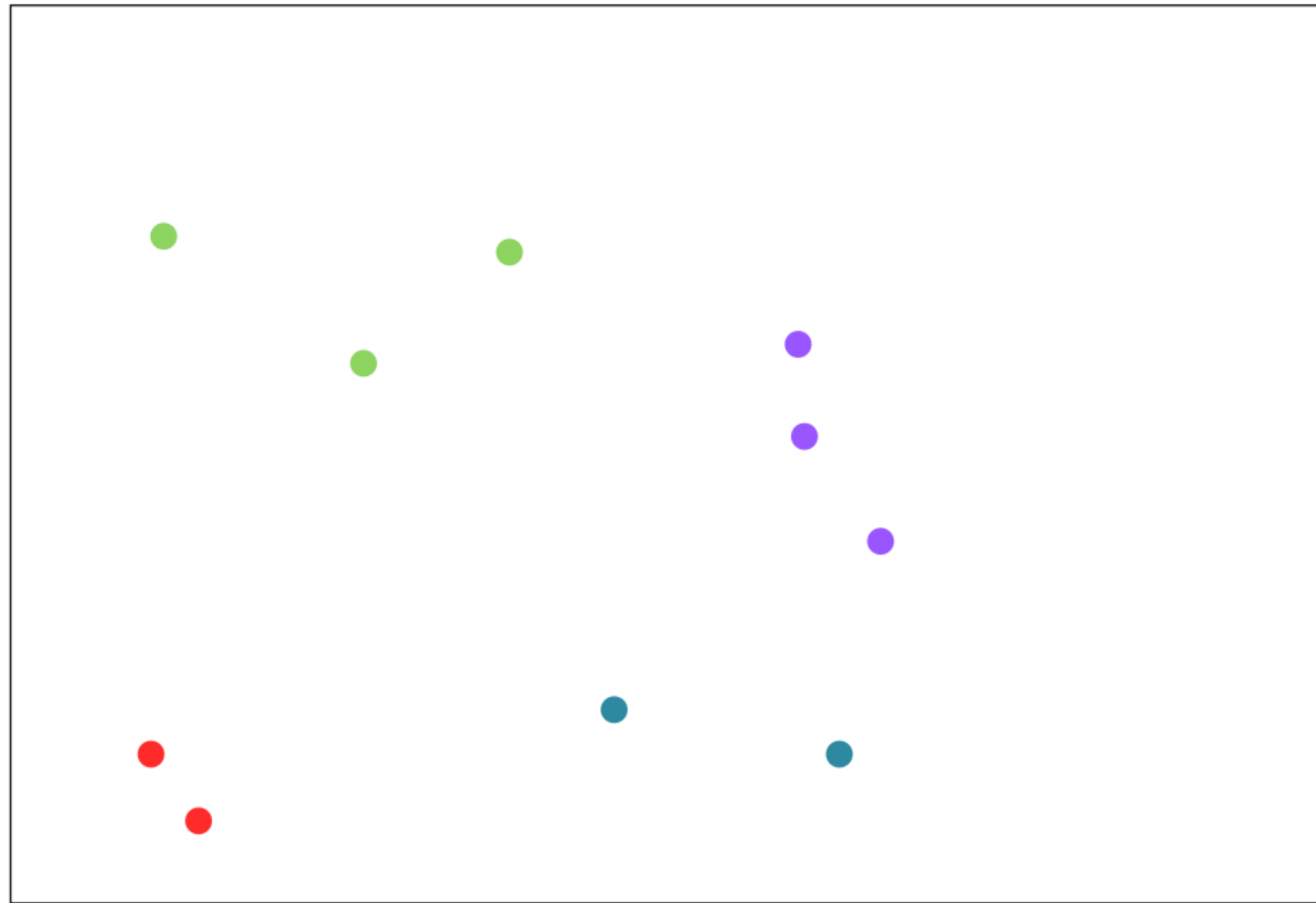
Hasta llegar a tener un único clúster.

Cuántos clústers



Donde definamos una altura máxima que es "muy lejos", va a quedar determinado un número de clústers.

Cuántos clústers



Algo muy útil de este algoritmo es que no importa si definimos distancia en 5 dimensiones, el dendrograma seguirá siendo una representación 2D.



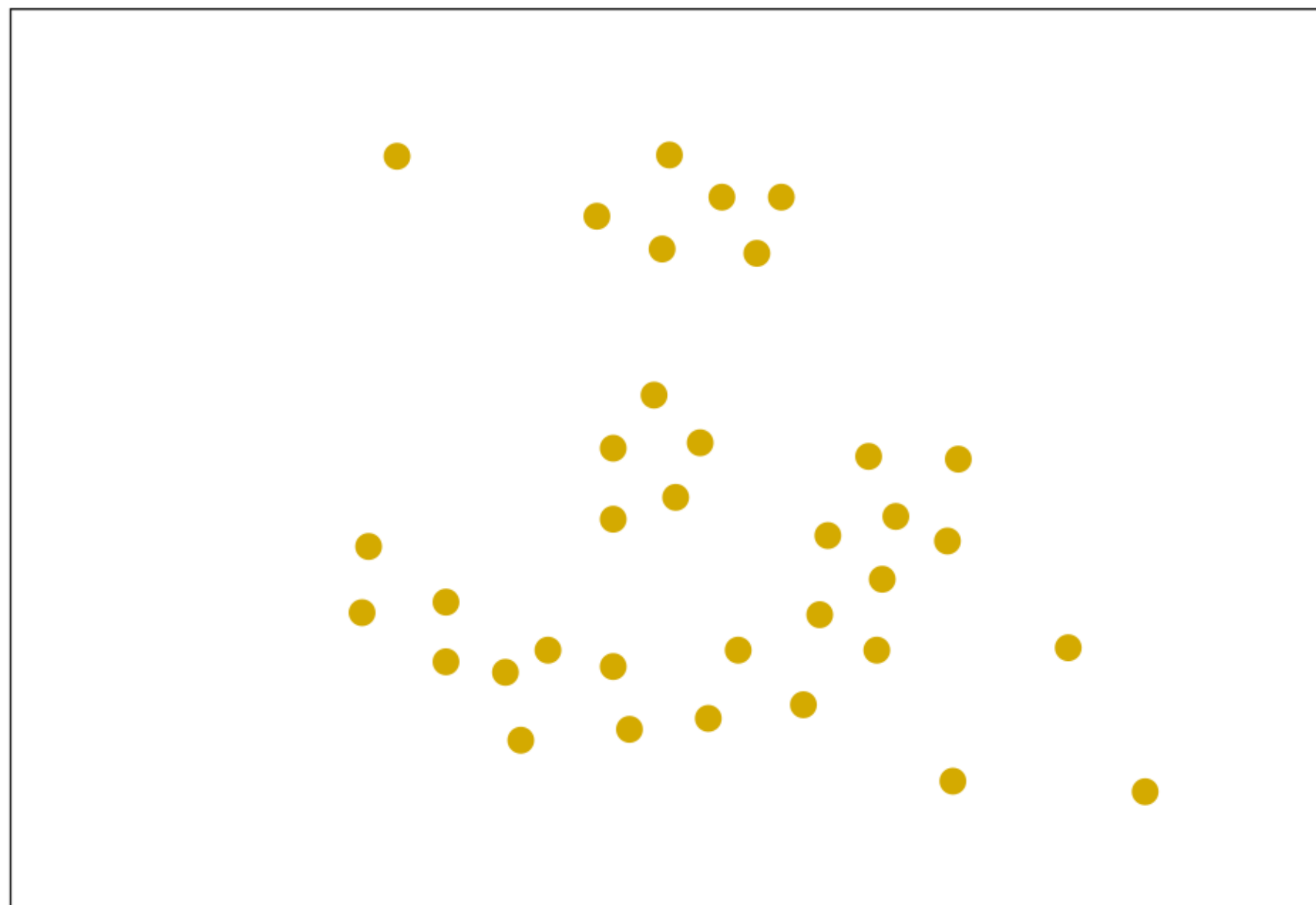
Agrupación por densidad

Intuición

- Identificamos los datos más "populares", los "extrovertidos", ellos invitan nuevos amigos al parche.
- Los más tímidos se unen porque los populares los invitan.
- Algunos se quedan solitos.

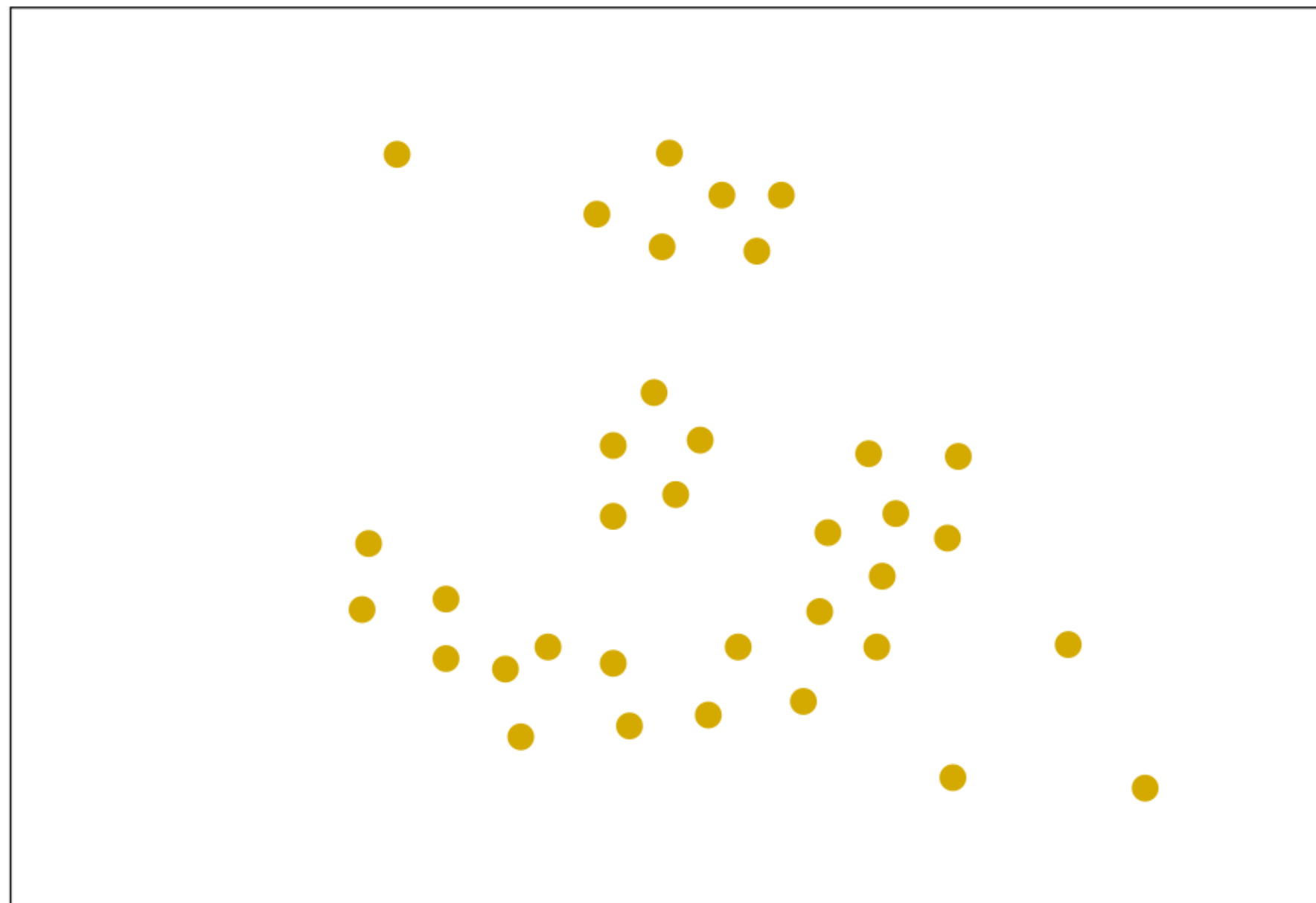


Agrupando



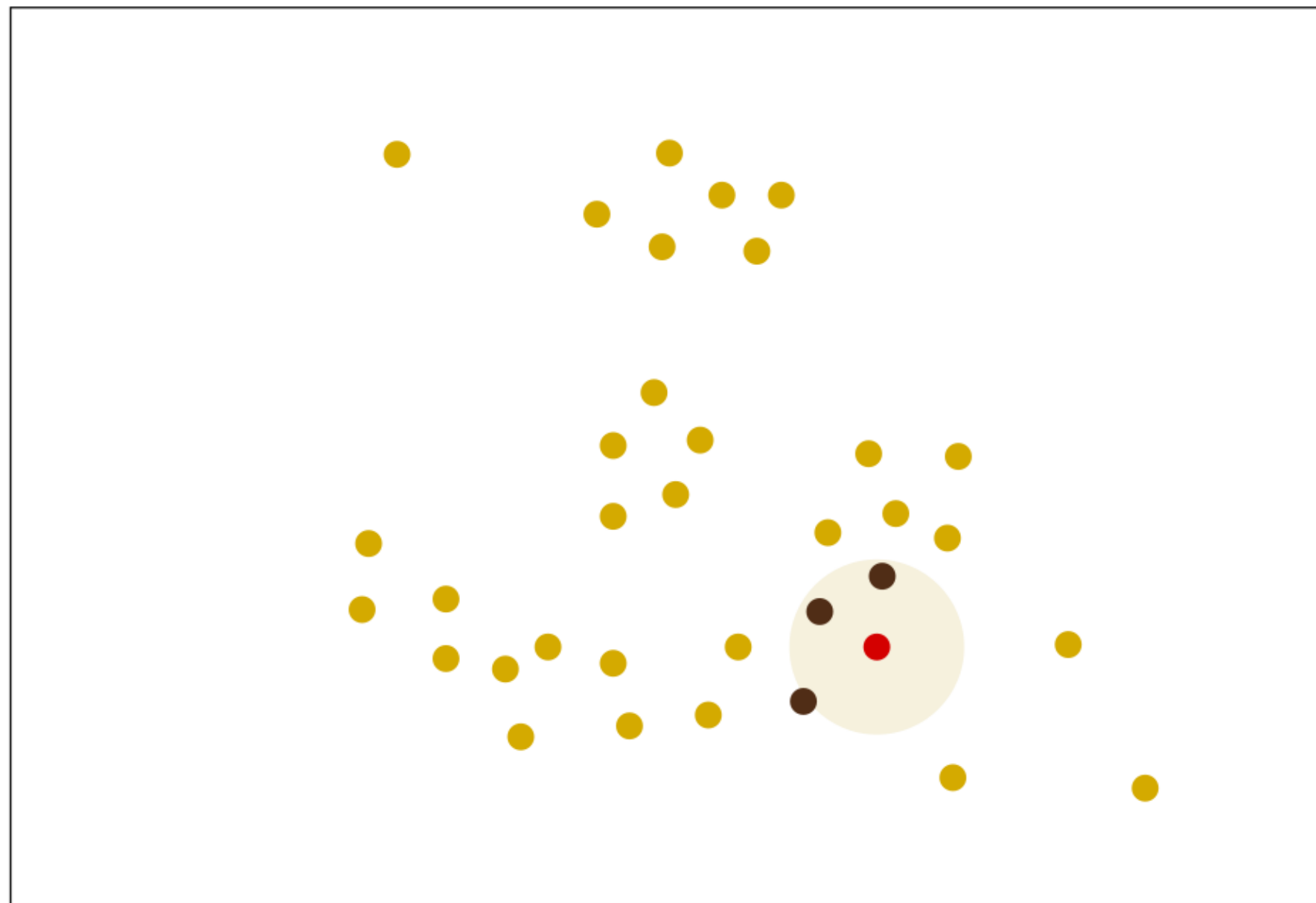
Partimos de los datos que vamos a agrupar.

Agrupando



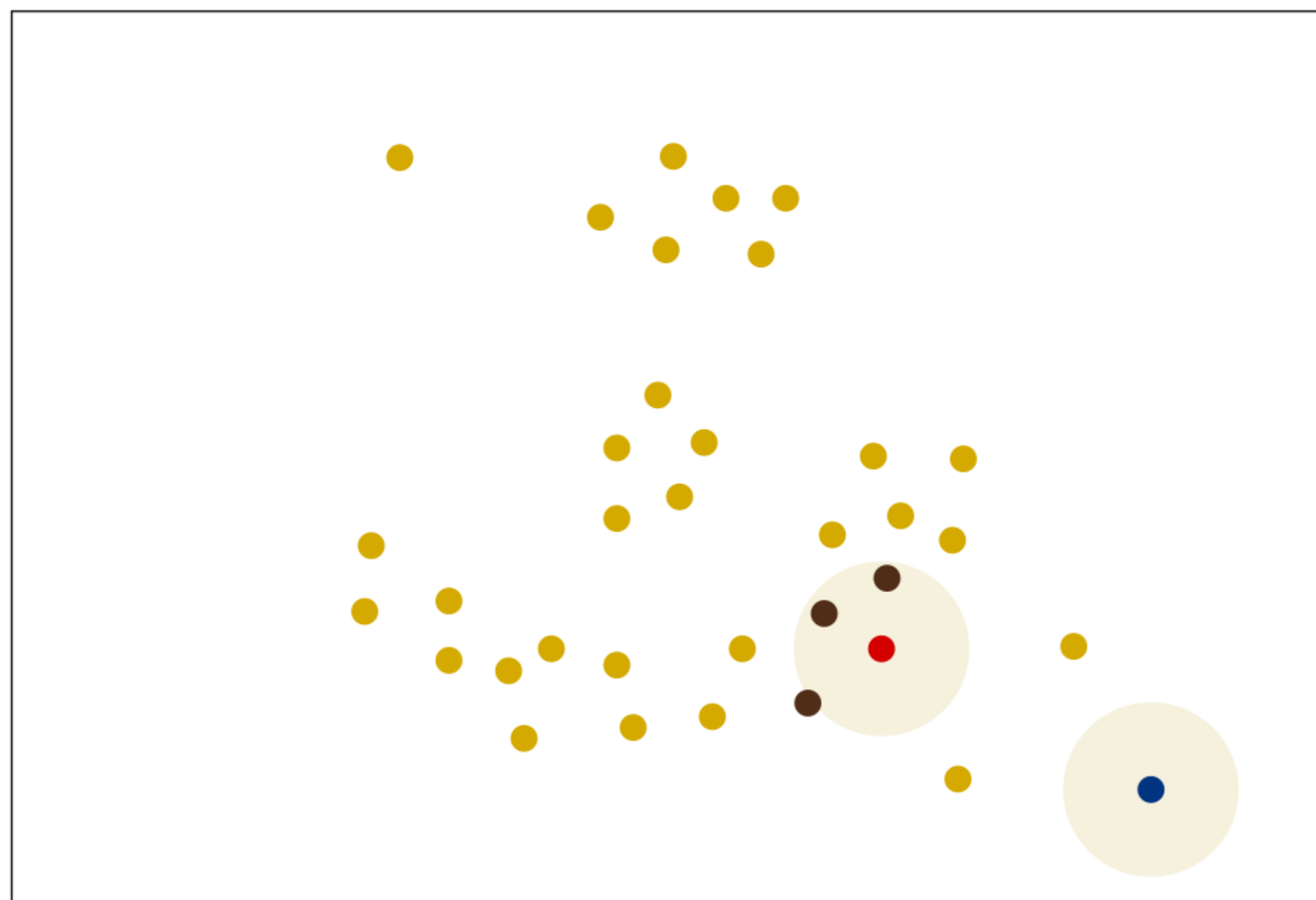
Definimos dos cosas: el **radio de contacto** entre los puntos, y qué significa ser "**popular**", por ejemplo tener "3 amigos".

Agrupando



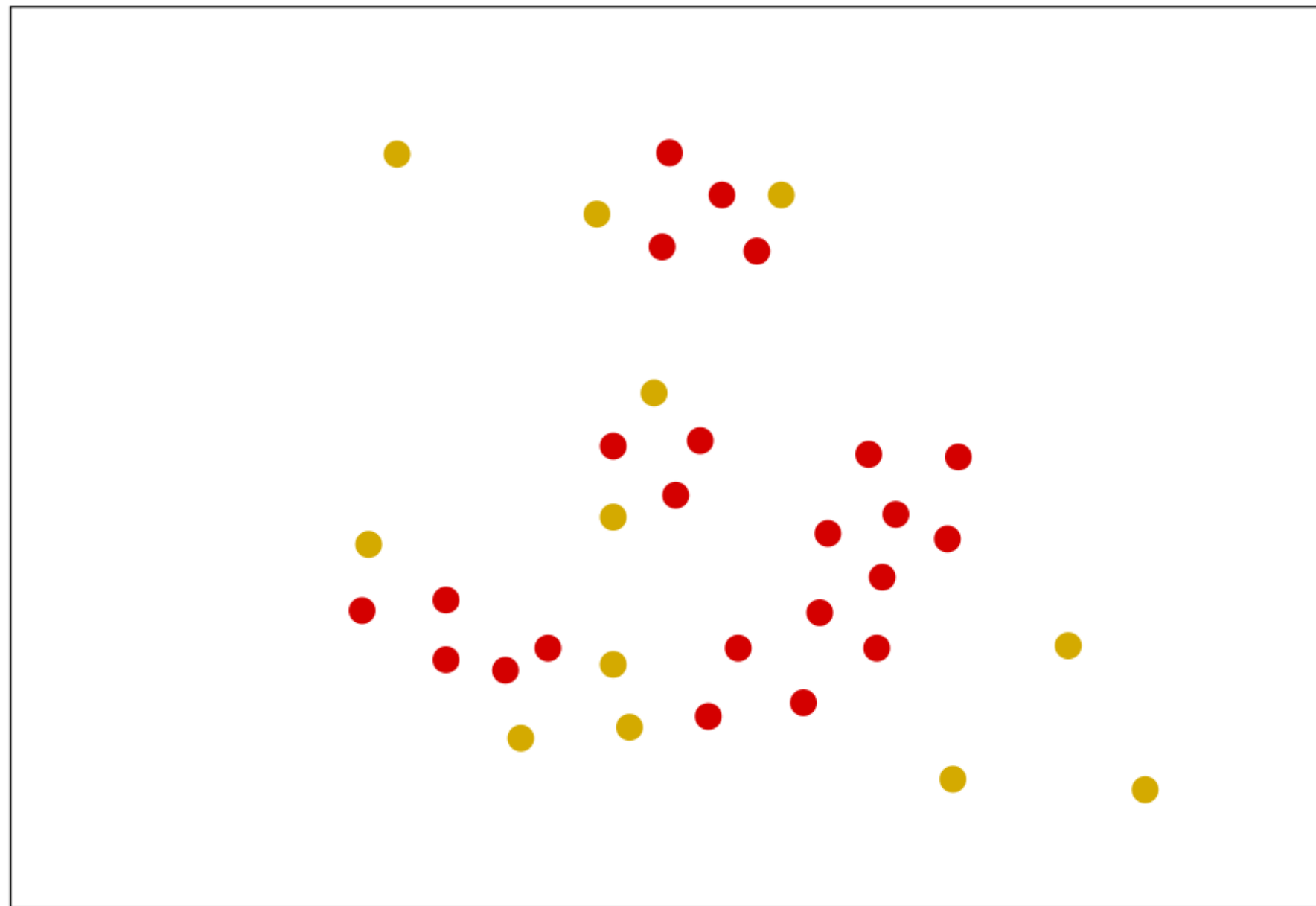
Este punto es popular, porque tiene contacto (así sea parcial) con tres amigos.

Agrupando



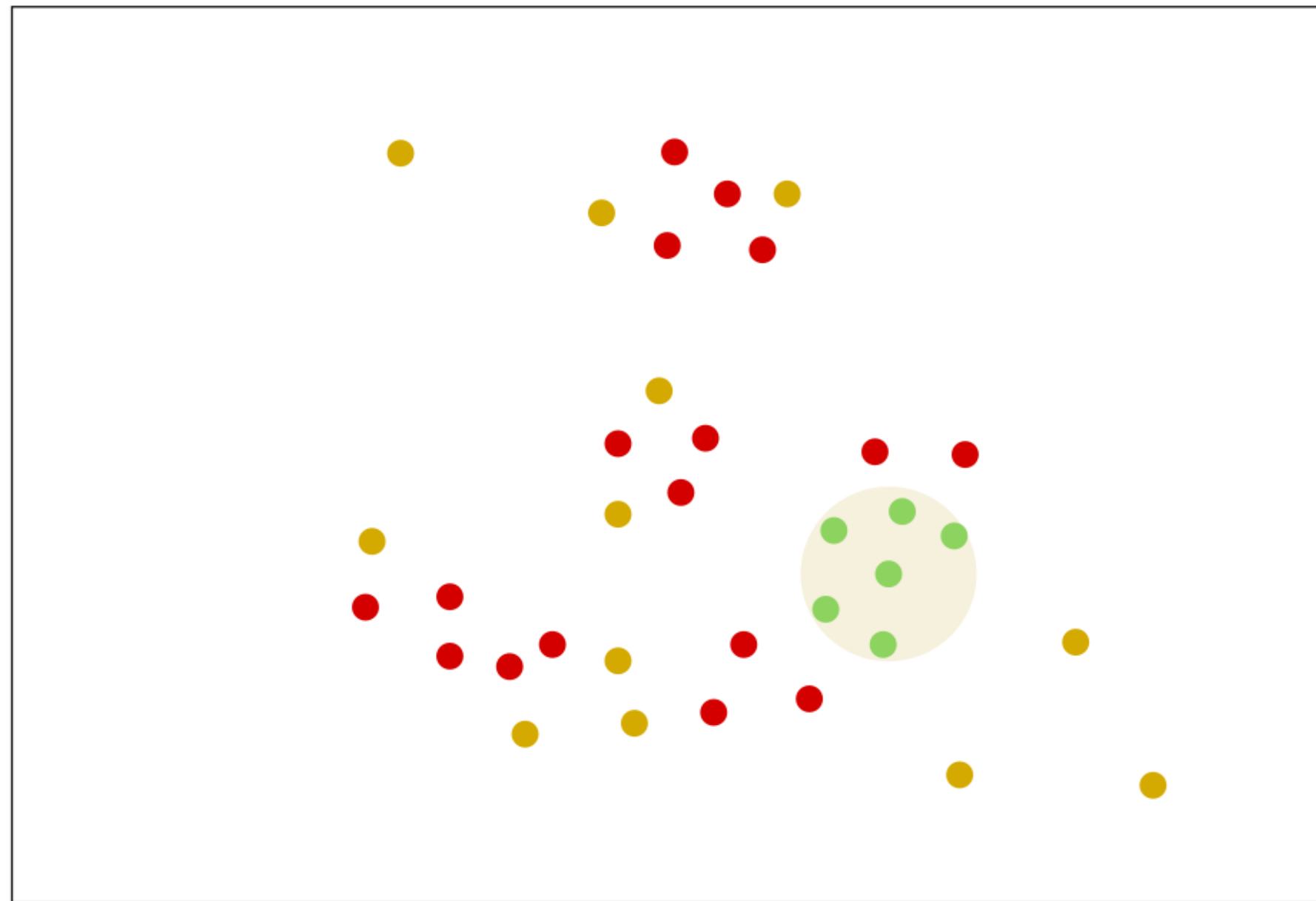
Este punto no es popular.

Agrupando



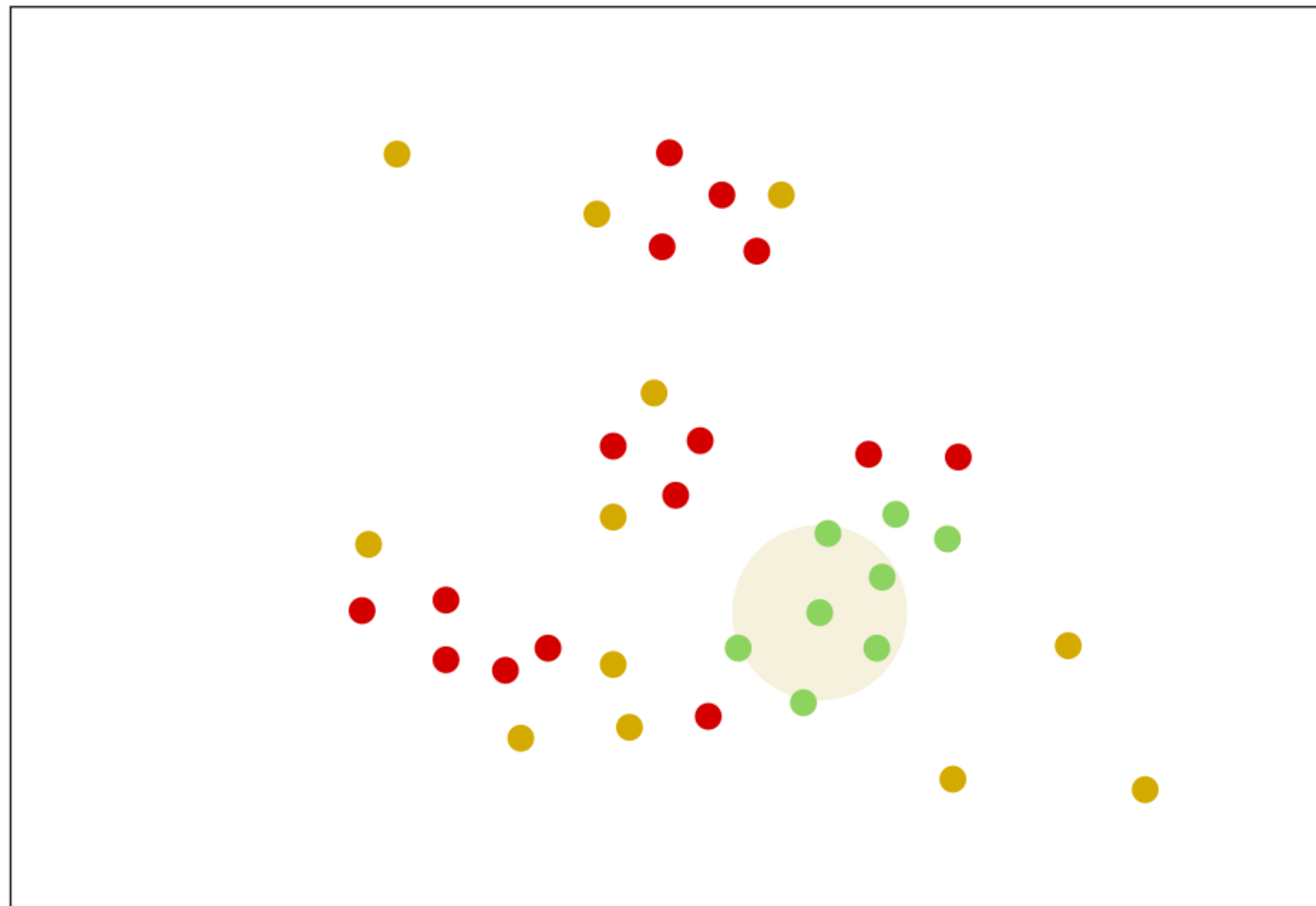
Todos los puntos "populares" se llaman **puntos core**.

Agrupando



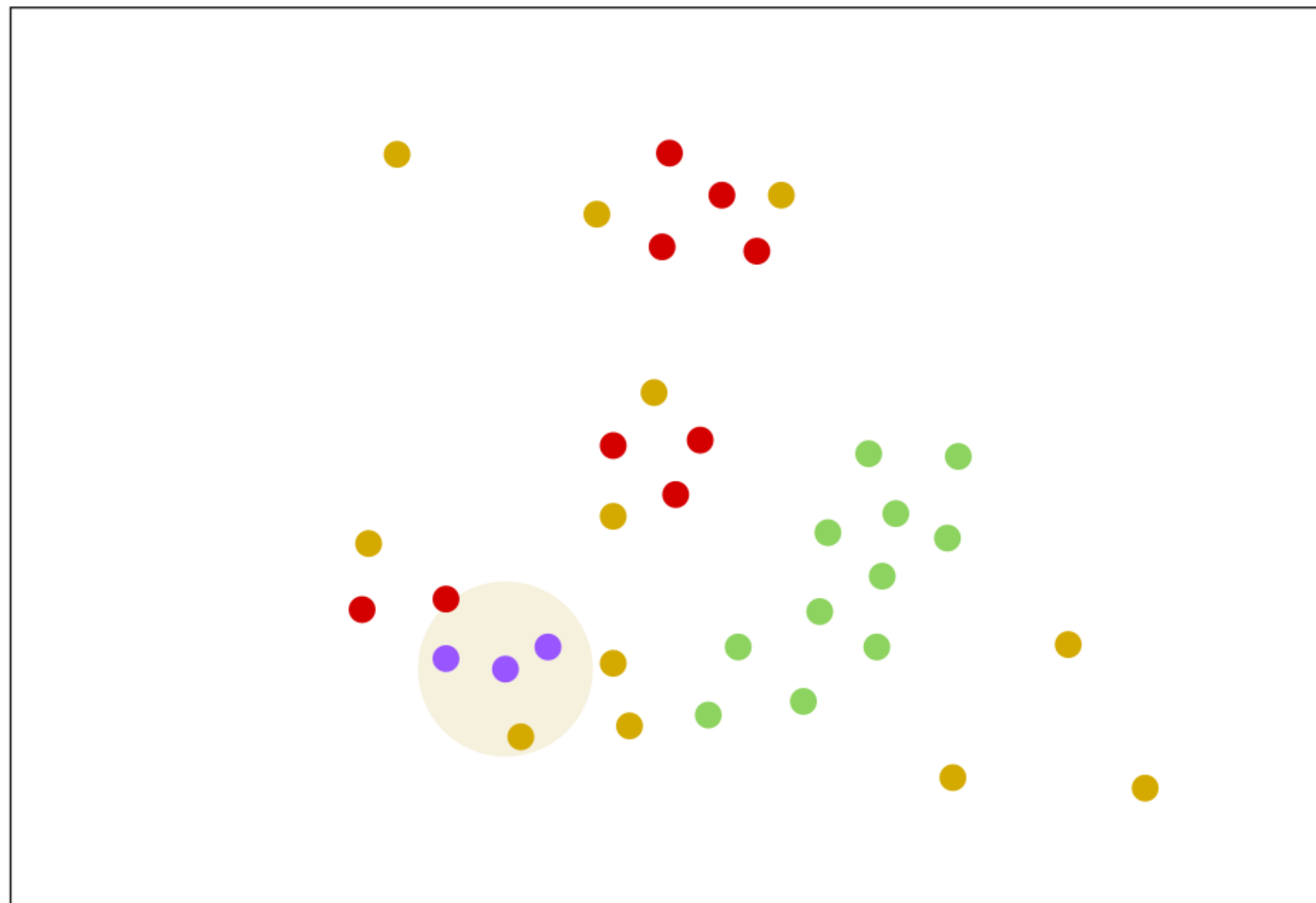
1. Escogemos un punto core aleatoriamente, y anexamos al cluster todos **sus puntos core** cercanos.

Agrupando



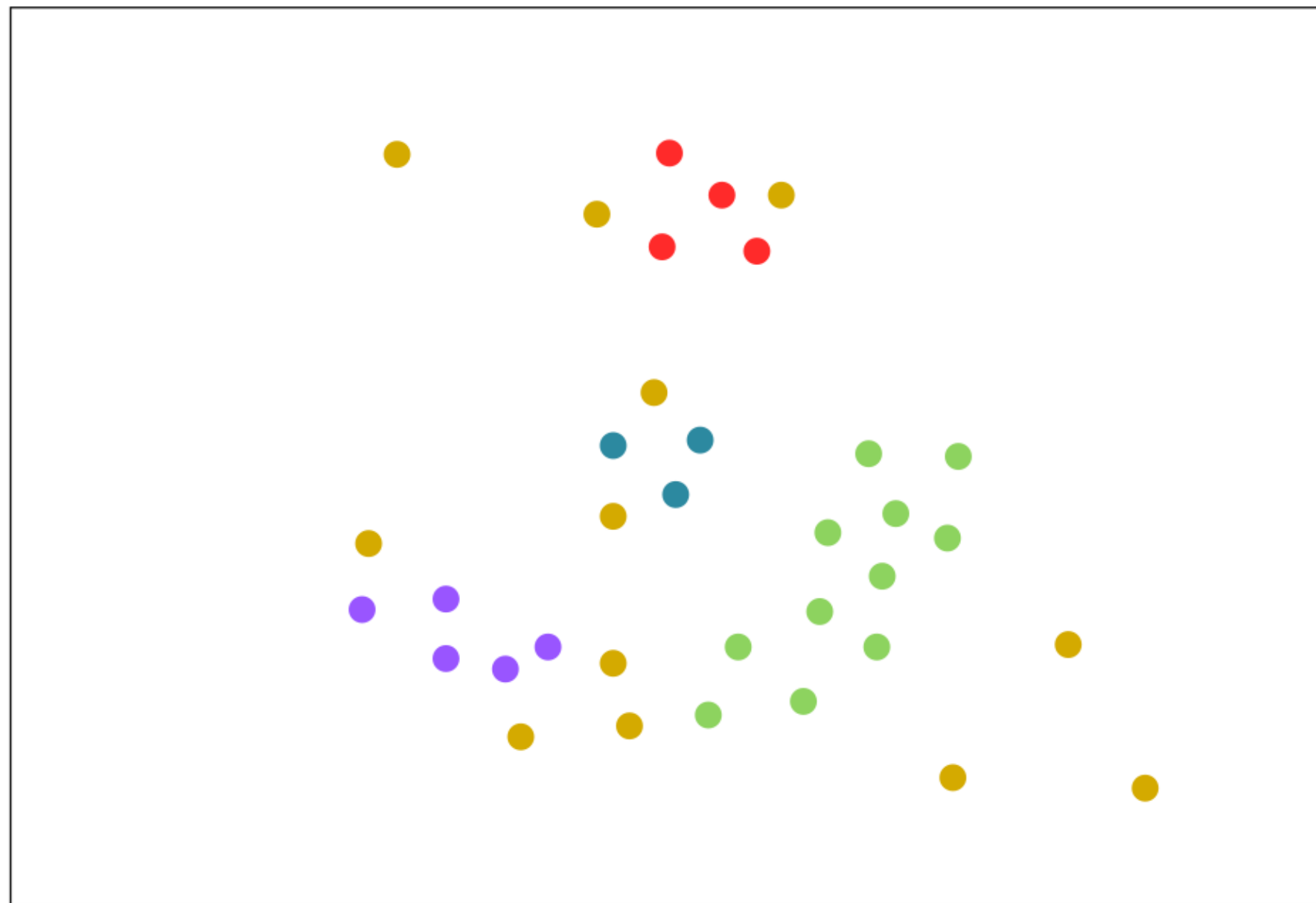
1. Los populares se juntan primero con los populares.

Agrupando



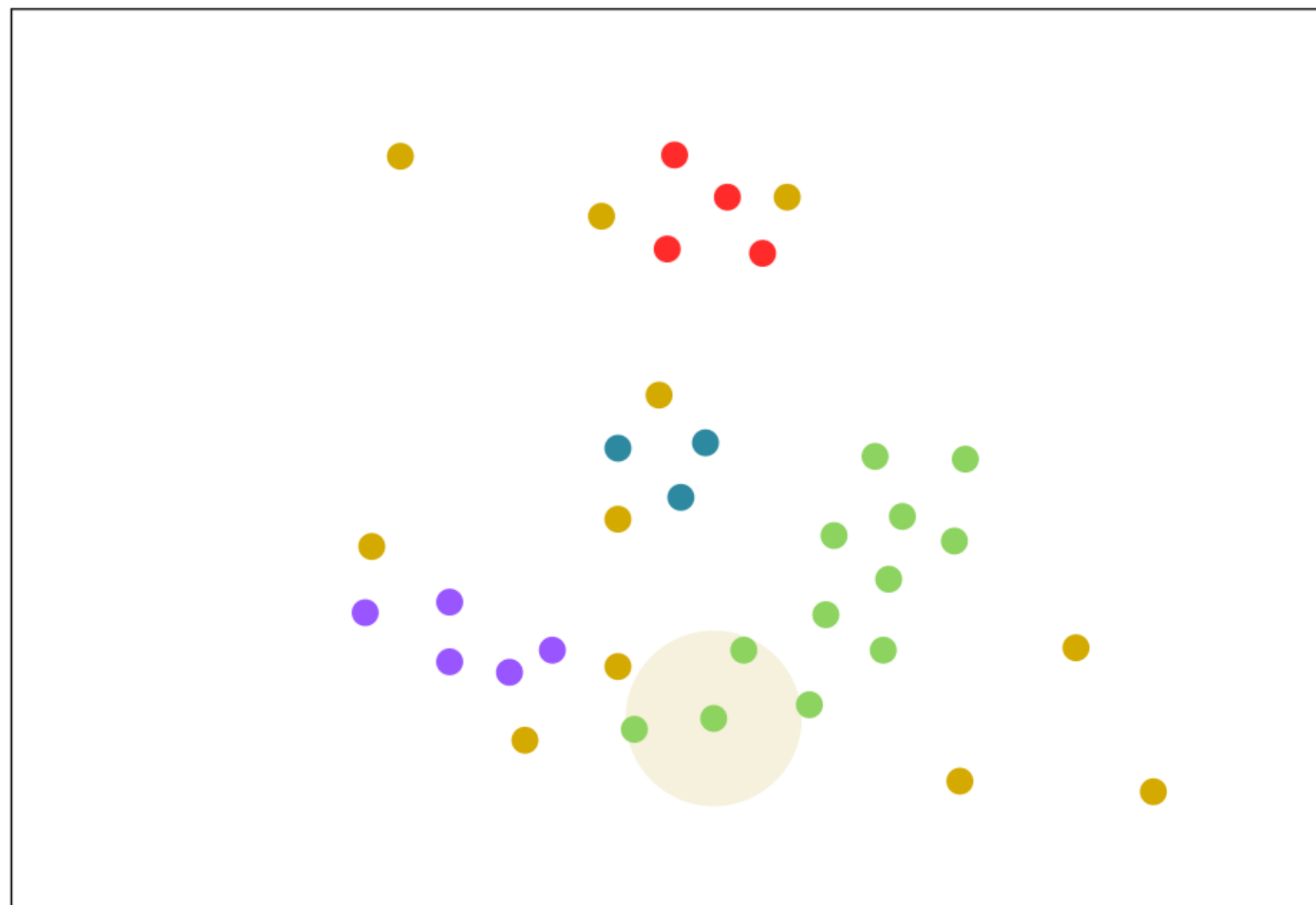
1. Cuando no quedan más populares en un grupo, escogemos **otro punto core aleatorio** (de los que no están ya en parche).

Agrupando



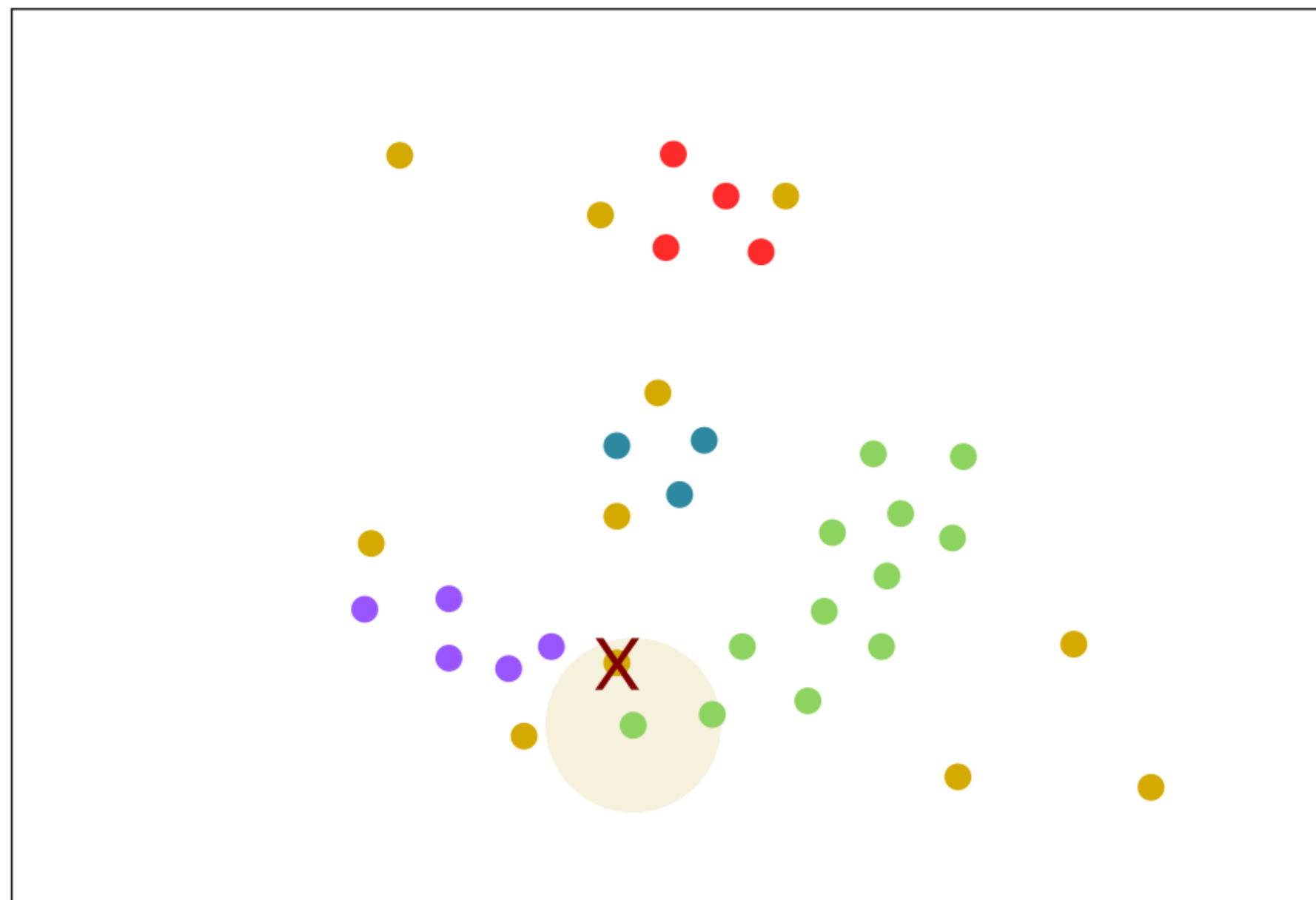
1. Se arman parches de todos los "puntos populares". Quedan los "puntos tímidos", o los que llamamos **no-core**.

Agrupando



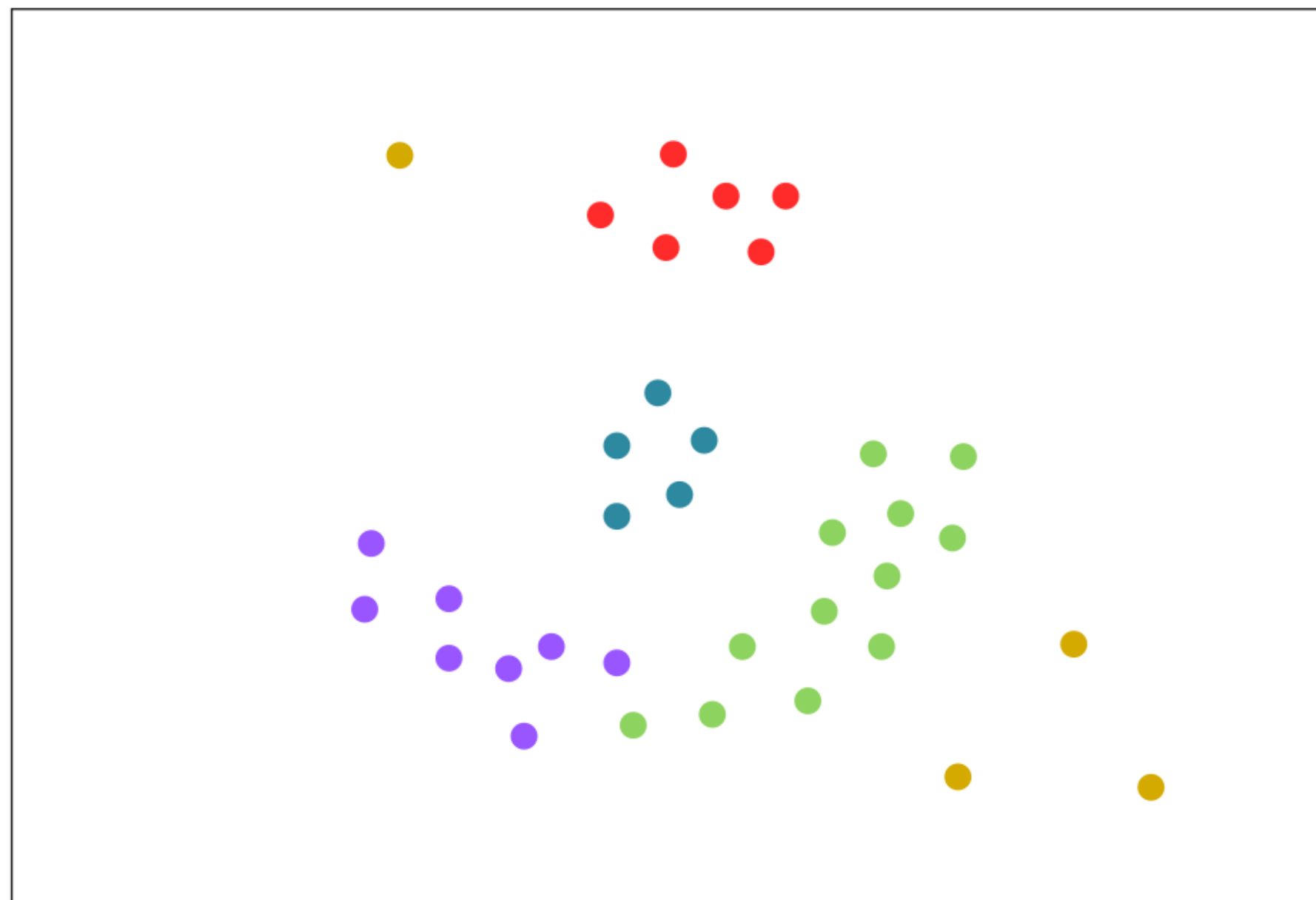
2. Luego, los puntos tímidos se anexan a los puntos populares que tienen cerca.

Agrupando



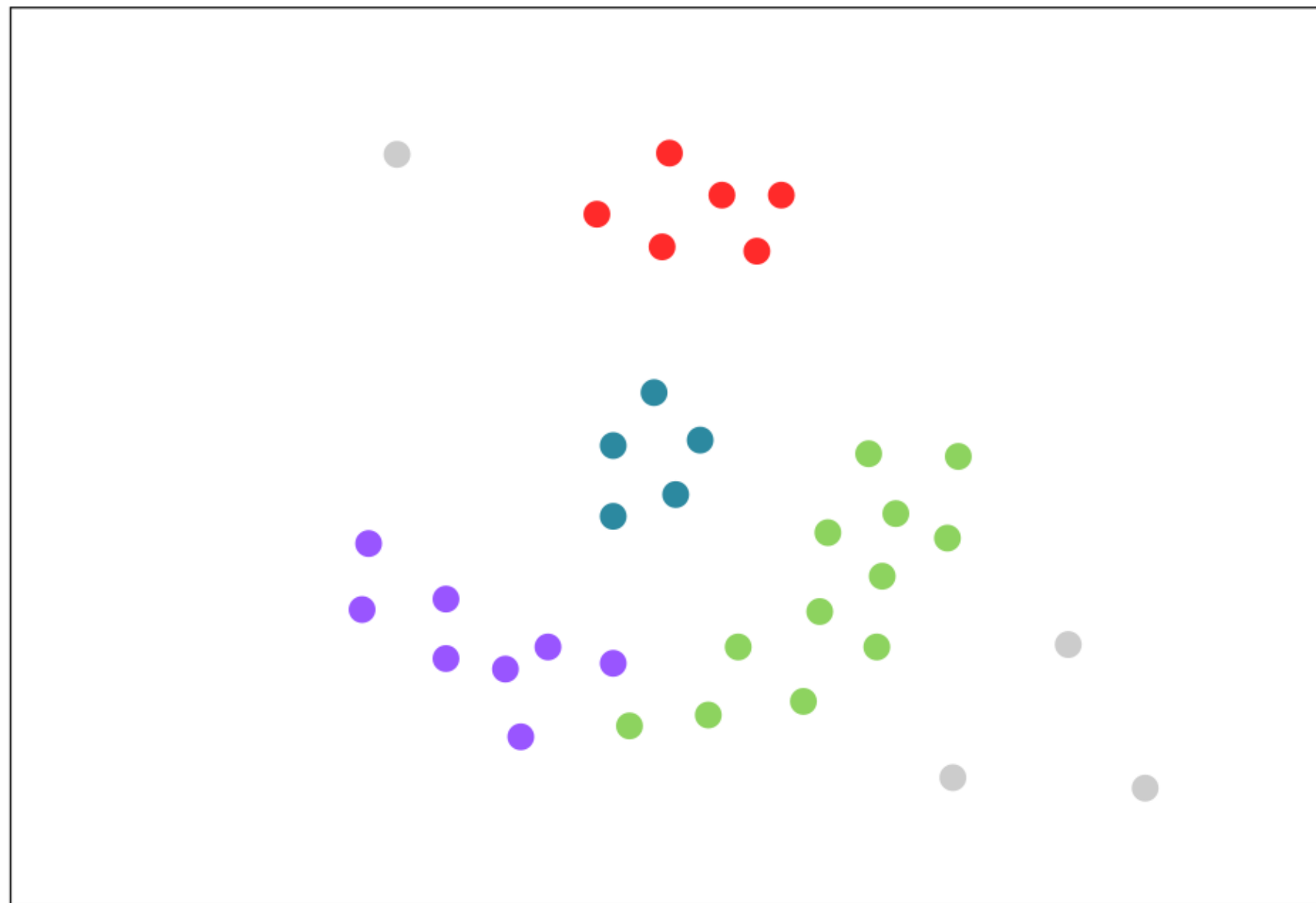
2. Noten: como el punto anexado es tímido, no es core, ese **sólo se une al clúster** de los populares, pero no invita a nadie más.

Agrupando



2. Todos los puntos tímidos se unen a los clústers de los puntos populares que tienen cerca.

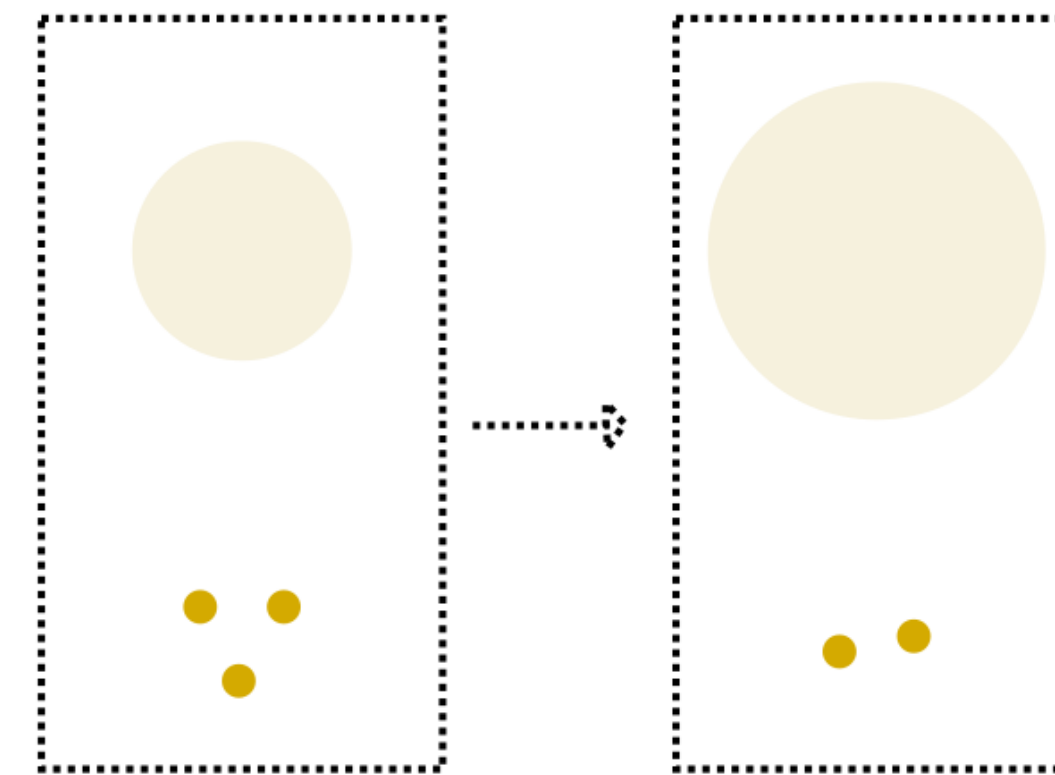
Agrupando



3. Los puntos tímidos que no tienen ningún contacto con algún punto popular se declaran ***outliers*** (datos atípicos). No pertenecen a ningún clúster.

Cuántos clústers

- En este modelo podemos jugar con el radio y la cantidad de contactos.
- Cuanto mayor sea el radio, o menor sean los contactos requeridos para ser popular, menor cantidad de clústers habrá.



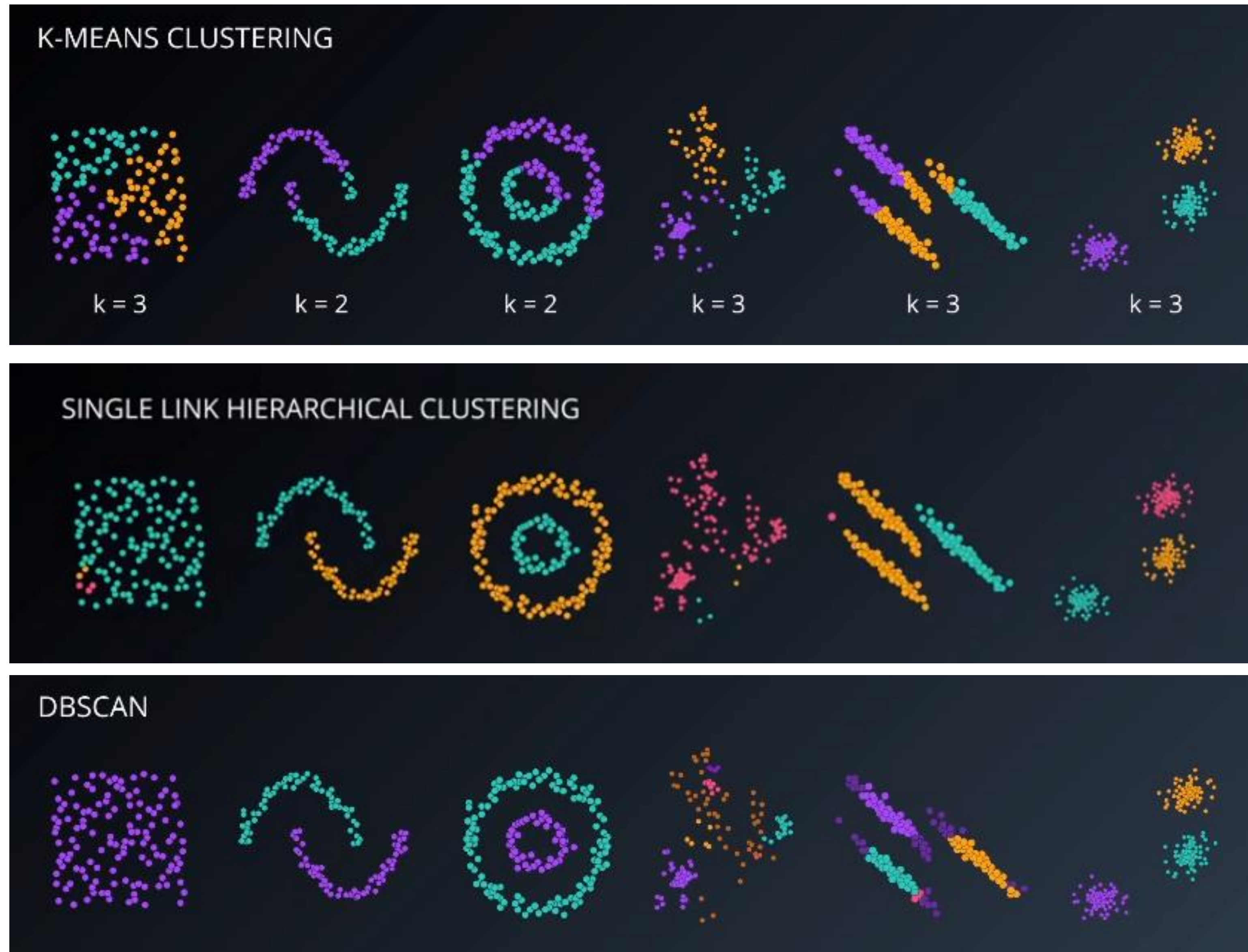
Conclusión



Al aplicar agrupación

- No sólo hace falta identificar cuántos grupos queremos conformar (jugar con los parámetros).
- Hay que probar varios métodos de *clustering* para encontrar el modelo que tenga más sentido para cada problema.
- A veces los grupos dan resultados intuitivos, interpretables. Tienen un sentido. A veces no. Aquí entra el **criterio del/la analista**.

Depende del problema, hay mejor modelo



Bibliografía

- [Estadística multivariada: inferencia y métodos \(unal.edu.co\)](http://unal.edu.co)

¡Gracias!

Aprendiendo juntos a lo largo de la vida

educacioncontinua.uniandes.edu.co

Síguenos: **EdcoUniandes**     



**Educación
Continua**
Vicerrectoría Académica

Universidad de los Andes | Vigilada Mineducación. Reconocimiento como Universidad: Decreto 1297 del 30 de mayo de 1964. Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.

