

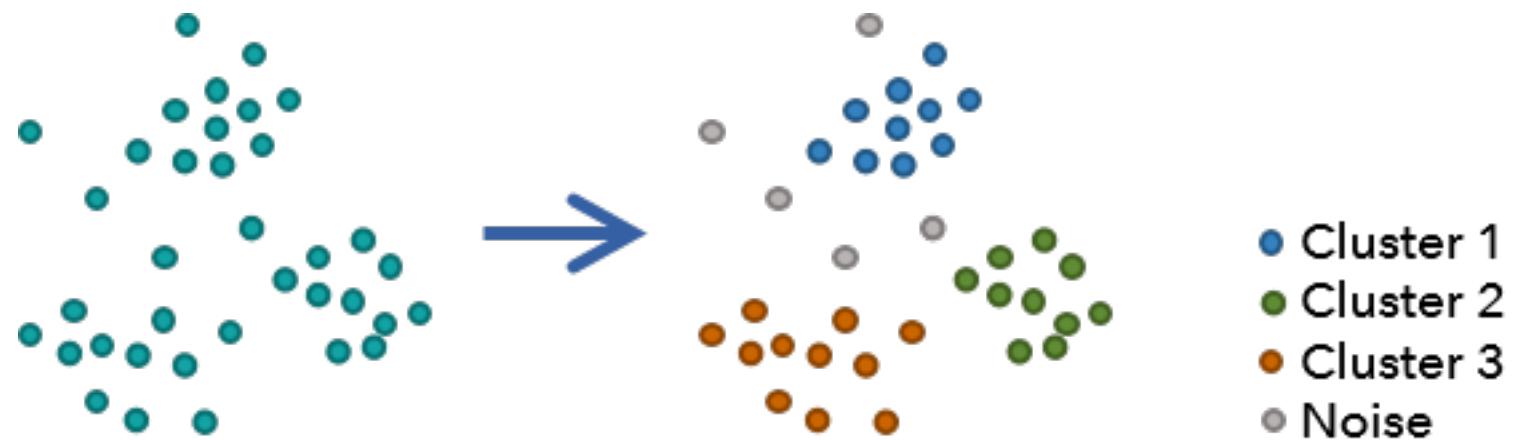


ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA

# IIC2433 Minería de Datos Clustering

Profesor: Mauricio Arriagada  
Minería de Datos

# Algoritmo de Clustering

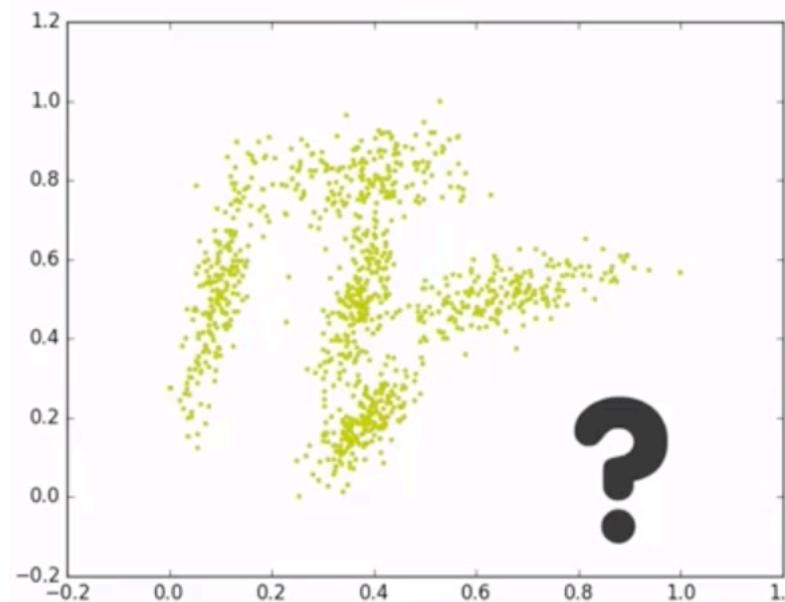


# OBJETIVO

- ▶ Algoritmo K-means consiste en encontrar grupos de datos similares o clusters

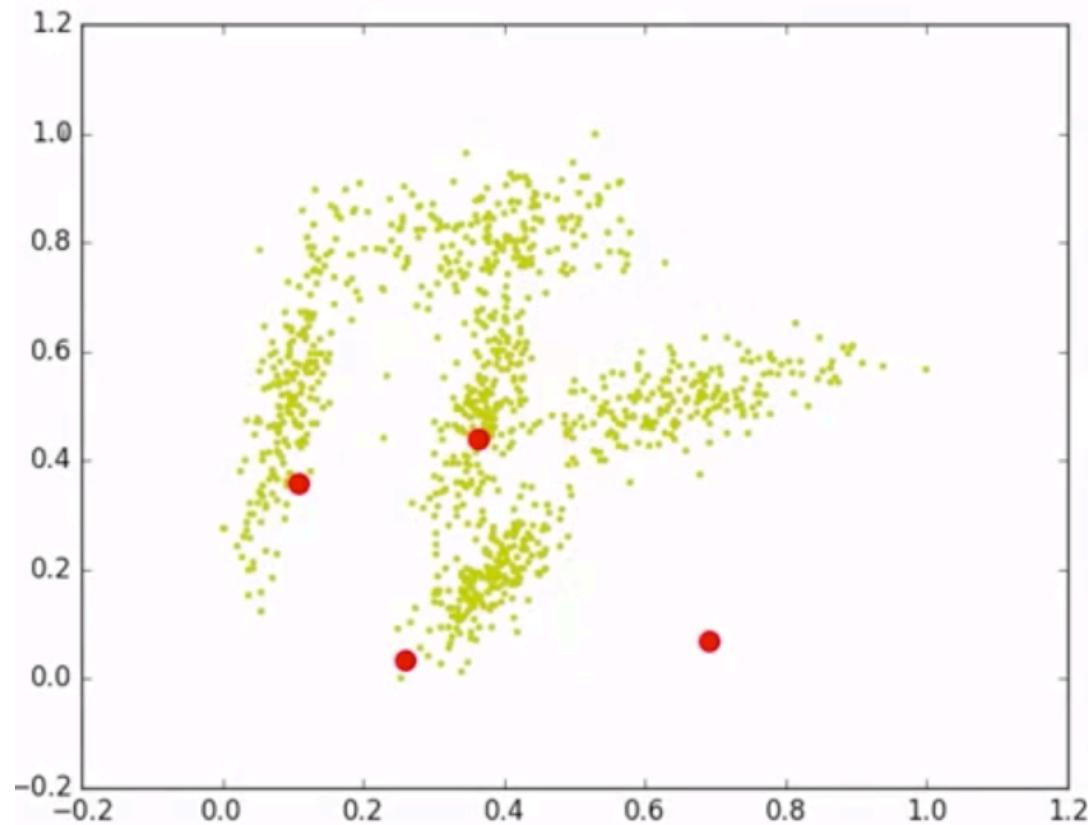
# Definir el valor de k

- ▶ Suponiendo que se tiene un set de datos de dos variables como se muestra en la figura
- ▶ El primer paso es definir el valor de k (número de clusters). Este valor es un parámetro del algoritmo que se debe definir antes de ejecutar el algoritmo. No existe un valor definido para k



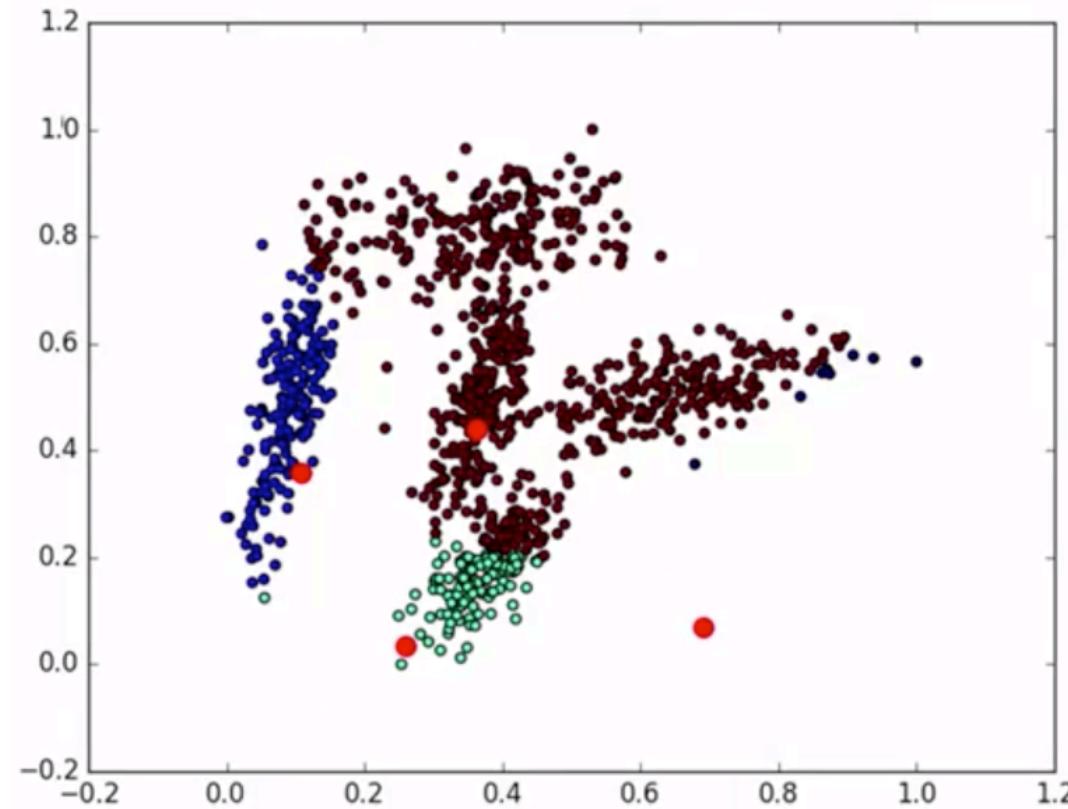
# Una vez definido el número k

- ▶ Se generan k centros alatorios



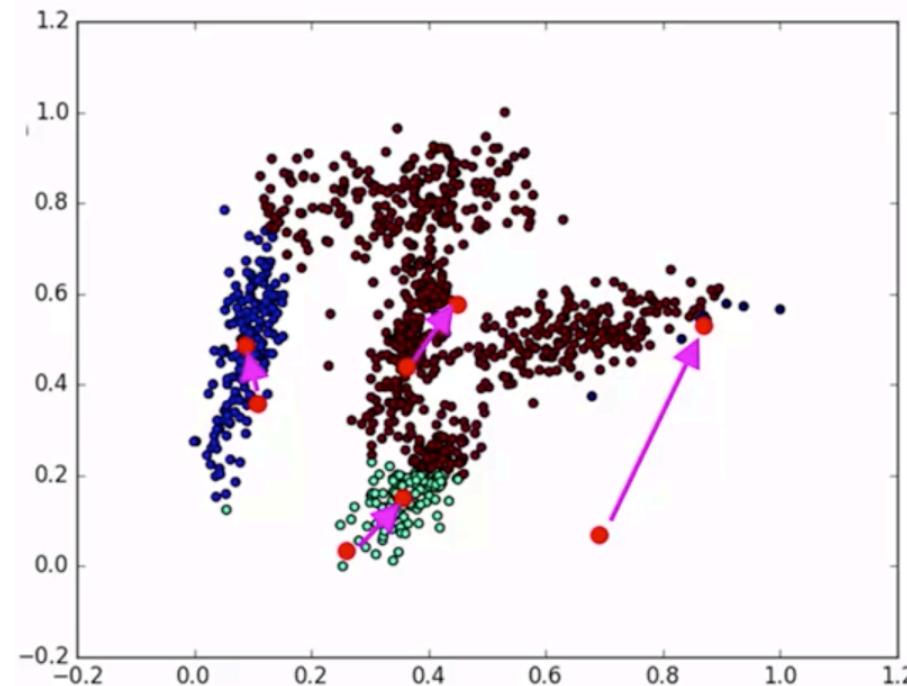
# Una vez definido el número k

- ▶ Se generan k centros alatorios
- ▶ Luego, se asigna cada punto del set de datos al punto más cercano



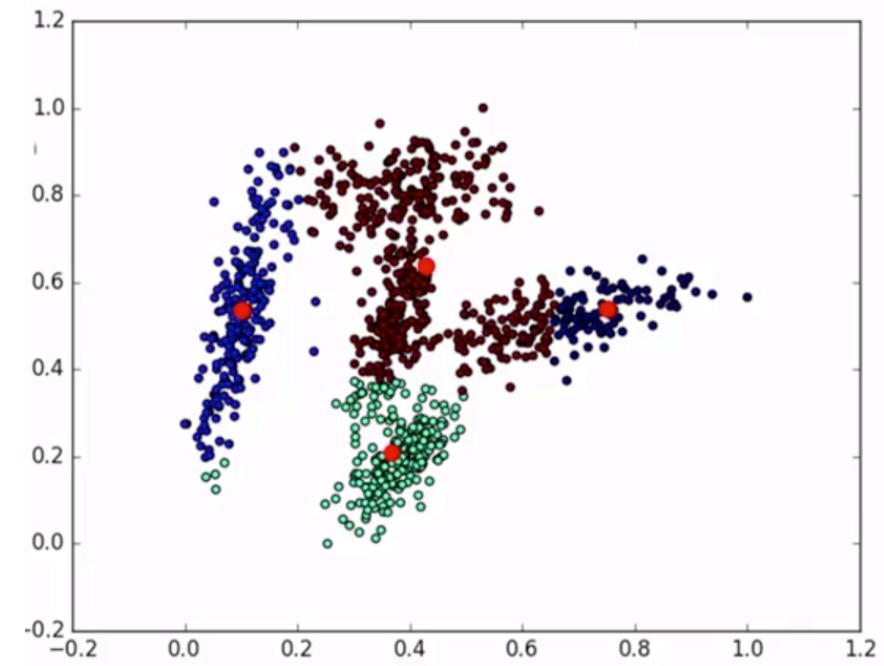
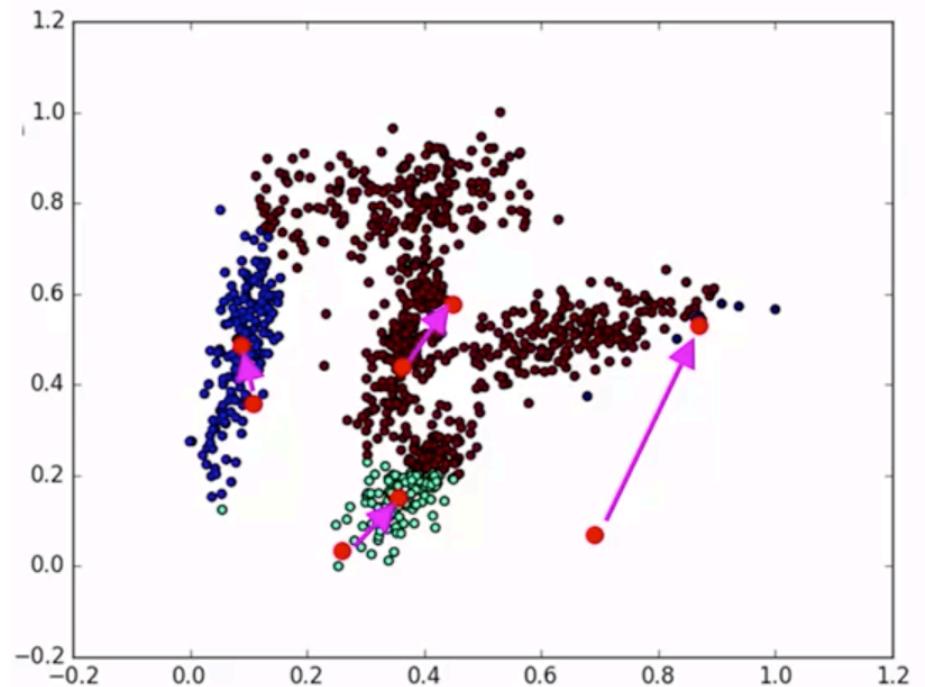
# Una vez definido el número k

- ▶ Ahora, los grupos se pueden reconfigurar. Es posible que cada grupo encuentre un nuevo centro en su cluster
- ▶ Es común usar distancia Euclíadiana para medir longitud de puntos



# Una vez definido el número k

- ▶ Cada vez que se actualizan los centros, cada cluster tienen un nuevo centro.
- ▶ Esto es un proceso iterativo hasta converger a mejores centros

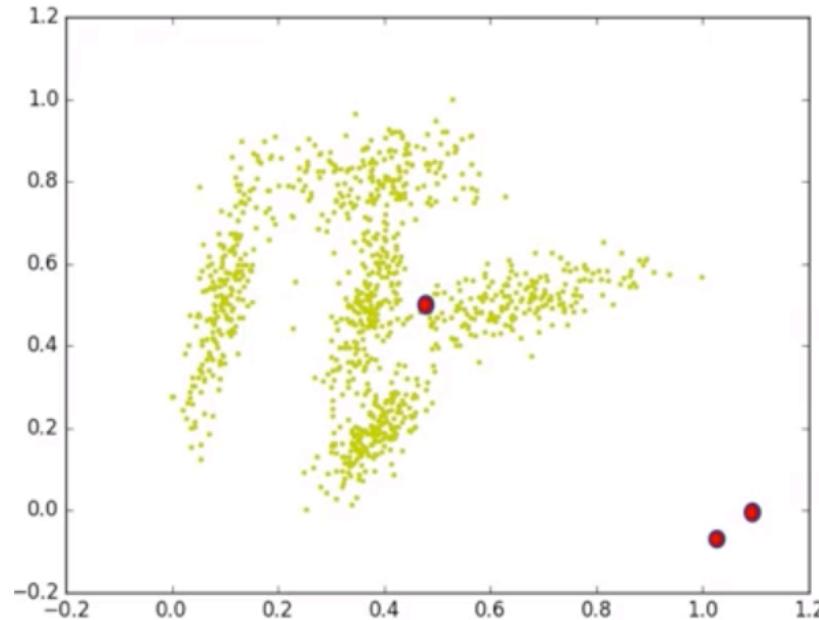


# Una vez definido el número k

- ▶ Este proceso iterativo termina cuando la posición de los cluster no cambia significativamente
- ▶ Es posible definir criterios de detención del algoritmo
  - ▶ Por ejemplo, definir un delta (distancia muy pequeña) que se usará para determinar si la posición de los centros es mayor o menor a este delta.

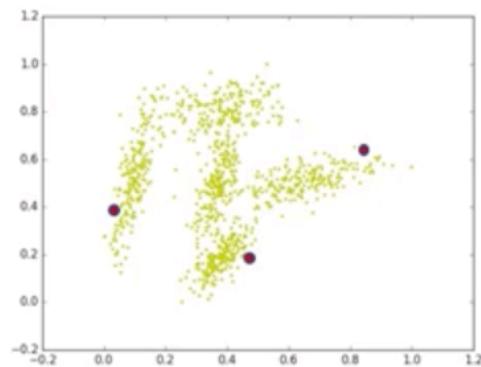
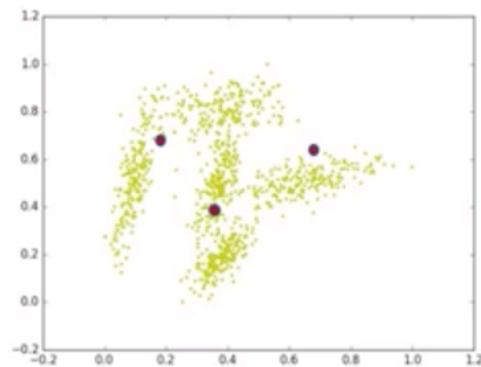
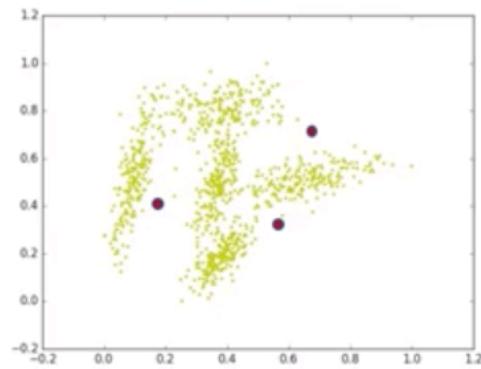
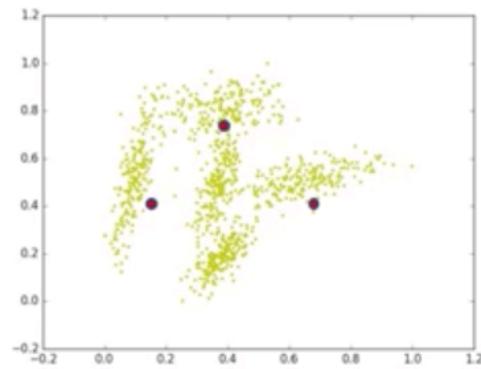
# Consideraciones antes de ejecutar k-means

- ▶ ¿Cómo se puede asegurar un resultado más cercano al correcto luego de haber ejecutado K-means?
- ▶ El algoritmo no siempre converge al mismo resultado. Esto es porque la posición inicial de los centros es aleatoria
- ▶ Por ejemplo en algunos casos, algunos centros pueden quedar muy lejos de las masas de datos impidiendo que estos puedan actualizarse y moverse a las zonas de mayor densidad



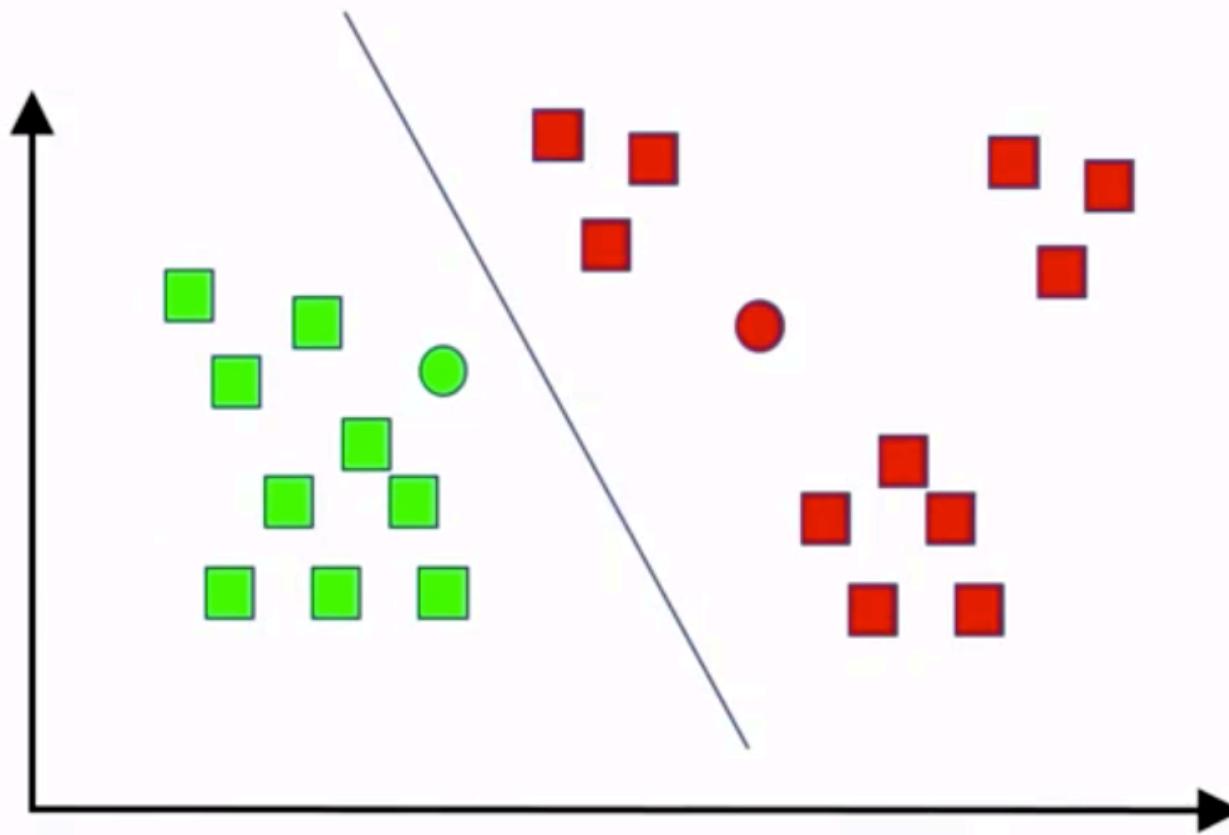
# Possible solución

Ejecutar el algoritmo K-Means varias veces,  
de tal forma de reducir la probabilidad de que nos estemos  
quedando con un resultado de clustering muy extremo



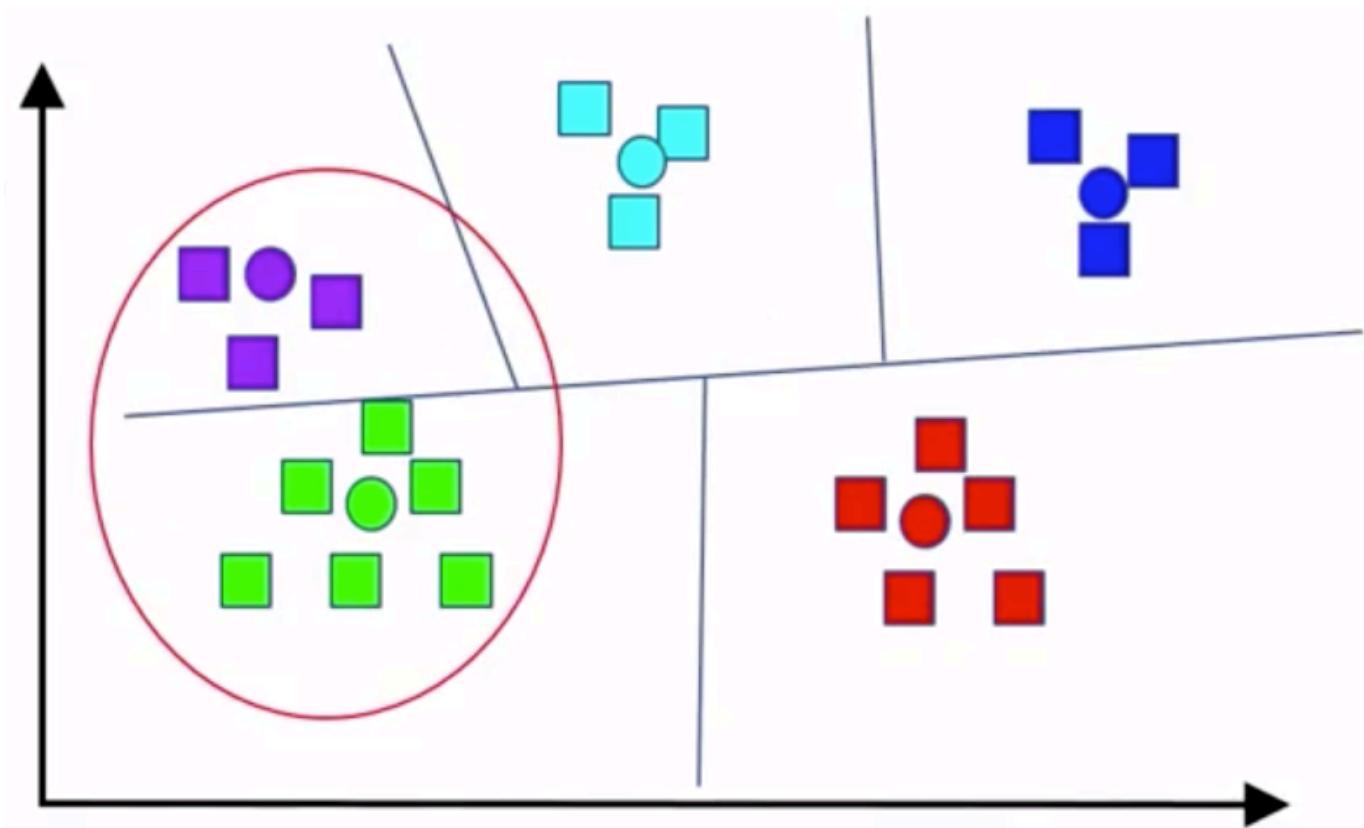
# Determinado el valor de k

- ▶ Ejemplo si  $k = 2$



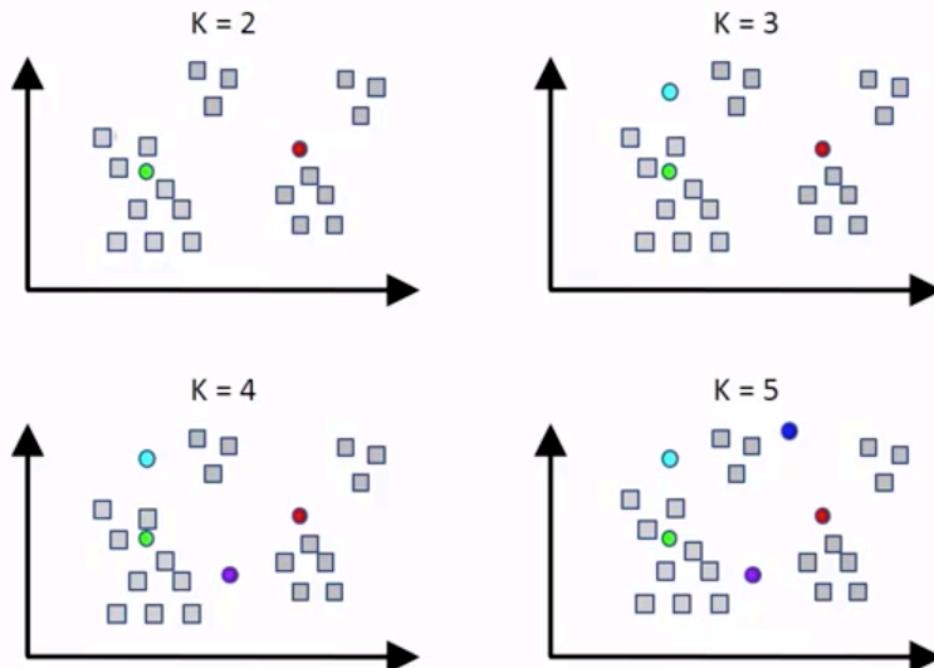
# Determinado el valor de k

- ▶ Ejemplo si  $k = 5$
- ▶ Comparando  $k=2$  y  $k=5$  los resultados son totalmente diferentes



# Determinado el valor de k

- ▶ Una alternativa es ejecutar el algoritmo varias veces cambiando el valor de K y analizando los resultados
- ▶ Si al variar K vemos que ciertos puntos pasan a otros clusters (cuando sabemos que no deberían) es probable que el valor de K sea muy alto
- ▶ Asimismo, si los clusters tienen pocos puntos, es un indicador que se debe reducir el valor de K



# Determinado el valor de k

- ▶ K-means trabaja con distancia Euclíadiana es importante usar variables que solo son importantes, de lo contrario, mantener variables sin importancia podría causar ruido en el cálculo de los centros.

# Ejemplo

- ▶ Supongamos que contamos con la siguiente información financiera perteneciente a un/a jefe/a de familia

Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Juan	9.8	\$290	\$92
Pedro	3.5	\$300	\$200
María	4.1	\$130	\$158
Angela	10	\$440	\$86
Nicolás	1.6	\$1000	\$164
Carlos	4.4	\$270	\$84

# Ejemplo

- ▶ El primer paso es normalizar

Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Juan	0.98	0.29	0.46
Pedro	0.35	0.30	1
María	0.41	0.13	0.79
Angela	1	0.44	0.43
Nicolás	0.16	1	0.82
Carlos	0.44	0.27	0.42

# Ejemplo

- ▶ Se definirán aleatoriamente los centros para 2 clusters

Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Cluster 1	0.31	0.3	0.65
Cluster 2	0.28	0.4	0.3

# Ejemplo

- Se debe calcular la distancia euclíadiana de todos los registros hacia los centros seleccionados

Centro 1

$$d(C_1, \text{Juan}) = 0.7$$

$$d(C_1, \text{Pedro}) = 0.35$$

$$d(C_1, \text{María}) = 0.24$$

$$d(C_1, \text{Angela}) = 0.74$$

$$d(C_1, \text{Nicolás}) = 0.74$$

$$d(C_1, \text{Carlos}) = 0.27$$

Centro 2

$$d(C_2, \text{Juan}) = 0.73$$

$$d(C_2, \text{Pedro}) = 0.71$$

$$d(C_2, \text{María}) = 0.57$$

$$d(C_2, \text{Angela}) = 0.73$$

$$d(C_2, \text{Nicolás}) = 0.80$$

$$d(C_2, \text{Carlos}) = 0.24$$

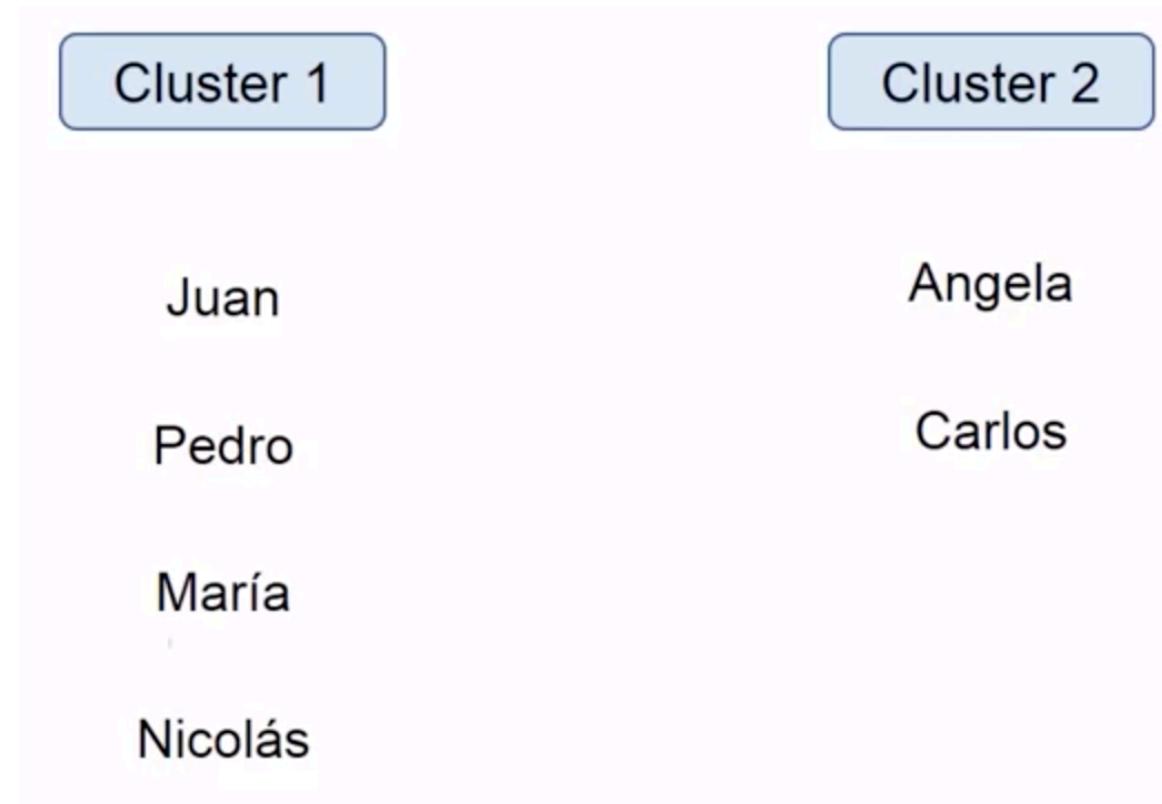
# Ejemplo

- Por registro, se selecciona el menor valor existente entre clusters

Centro 1	Centro 2
$d(C1, Juan) = 0.7$	$d(C2, Juan) = 0.73$
$d(C1, Pedro) = 0.35$	$d(C2, Pedro) = 0.71$
$d(C1, María) = 0.24$	$d(C2, María) = 0.57$
$d(C1, Angela) = 0.74$	$d(C2, Angela) = 0.73$
$d(C1, Nicolás) = 0.74$	$d(C2, Nicolás) = 0.80$
$d(C1, Carlos) = 0.27$	$d(C2, Carlos) = 0.24$

# Ejemplo

- ▶ Esto genera que los clusters tienen registros asociados a ellos



# Ejemplo

- Se debe actualizar el centro del nuevo cluster. Esto se puede obtener a través del promedio de cada registro asociado al cluster en cuestión. En este caso, cluster 1

Cluster 1

Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Juan	0.98	0.29	0.46
Pedro	0.35	0.30	1
María	0.41	0.13	0.79
Nicolás	0.16	1	0.82

$$\left( \frac{0.98 + 0.35 + 0.41 + 0.16}{4}, \frac{0.29 + 0.3 + 0.13 + 1}{4}, \frac{0.46 + 1 + 0.79 + 0.82}{4} \right)$$

$$(0.475, 0.43, 0.767)$$

# Ejemplo

- ▶ Se debe actualizar el centro del nuevo cluster. Esto se puede obtener a través del promedio de cada registro asociado al cluster en cuestión. En este caso, cluster 2

Cluster 2			
Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Angela	1	0.439	0.43
Carlos	0.435	0.268	0.422

$(0.72, 0.355, 0.767)$

# Ejemplo

- Este proceso se debe repetir usando los nuevos centros

Centro 1	Centro 2
$d(C1, Juan) = 0.61$	$d(C2, Juan) = 0.27$
$d(C1, Pedro) = 0.29$	$d(C2, Pedro) = 0.69$
$d(C1, María) = 0.31$	$d(C2, María) = 0.53$
$d(C1, Angela) = 0.62$	$d(C2, Angela) = 0.29$
$d(C1, Nicolás) = 0.65$	$d(C2, Nicolás) = 0.94$
$d(C1, Carlos) = 0.38$	$d(C2, Carlos) = 0.29$

# Ejemplo

- Este proceso se debe repetir usando los nuevos centros

Centro 1

$$d(C1, Juan) = 0.61$$

$$d(C1, Pedro) = 0.29$$

$$d(C1, María) = 0.31$$

$$d(C1, Angela) = 0.62$$

$$d(C1, Nicolás) = 0.65$$

$$d(C1, Carlos) = 0.38$$

Centro 2

$$d(C2, Juan) = 0.27$$

$$d(C2, Pedro) = 0.69$$

$$d(C2, María) = 0.53$$

$$d(C2, Angela) = 0.29$$

$$d(C2, Nicolás) = 0.94$$

$$d(C2, Carlos) = 0.29$$

# Ejemplo

- ▶ Los registros asociados a los nuevos centros son:

Cluster 1	Cluster 2
Pedro	Juan
María	Angela
Nicolás	Carlos

# Ejemplo

- ▶ Se vuelve a actualizar los centros, en este caso de cluster 1:

Cluster 1			
Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Pedro	0.35	0.30	1
María	0.41	0.13	0.79
Nicolás	0.16	1	0.82

$(0.31, 0.48, 0.87)$

# Ejemplo

- ▶ Se vuelve a actualizar los centros, en este caso de cluster 1:

Cluster 2

Nombre	Antigüedad (Años)	Sueldo	Gasto Mensual
Juan	0.98	0.29	0.46
Angela	1	0.439	0.43
Carlos	0.435	0.268	0.422

$(0.81, 0.33, 0.44)$

# Ejemplo

- Se vuelve a iterar con los nuevos centros.:

Centro 1

$$d(C_1, \text{Juan}) = 0.81$$

$$d(C_1, \text{Pedro}) = 0.22$$

$$d(C_1, \text{María}) = 0.37$$

$$d(C_1, \text{Angela}) = 0.82$$

$$d(C_1, \text{Nicolás}) = 0.55$$

$$d(C_1, \text{Carlos}) = 0.51$$

Centro 2

$$d(C_2, \text{Juan}) = 0.18$$

$$d(C_2, \text{Pedro}) = 0.73$$

$$d(C_2, \text{María}) = 0.57$$

$$d(C_2, \text{Angela}) = 0.22$$

$$d(C_2, \text{Nicolás}) = 1$$

$$d(C_2, \text{Carlos}) = 0.37$$

# Ejemplo

- Se seleccionan los registros asociados al cluster:

Centro 1

$$d(C_1, \text{Juan}) = 0.81$$

$$d(C_1, \text{Pedro}) = 0.22$$

$$d(C_1, \text{María}) = 0.37$$

$$d(C_1, \text{Angela}) = 0.82$$

$$d(C_1, \text{Nicolás}) = 0.55$$

$$d(C_1, \text{Carlos}) = 0.51$$

Centro 2

$$d(C_2, \text{Juan}) = 0.18$$

$$d(C_2, \text{Pedro}) = 0.73$$

$$d(C_2, \text{María}) = 0.57$$

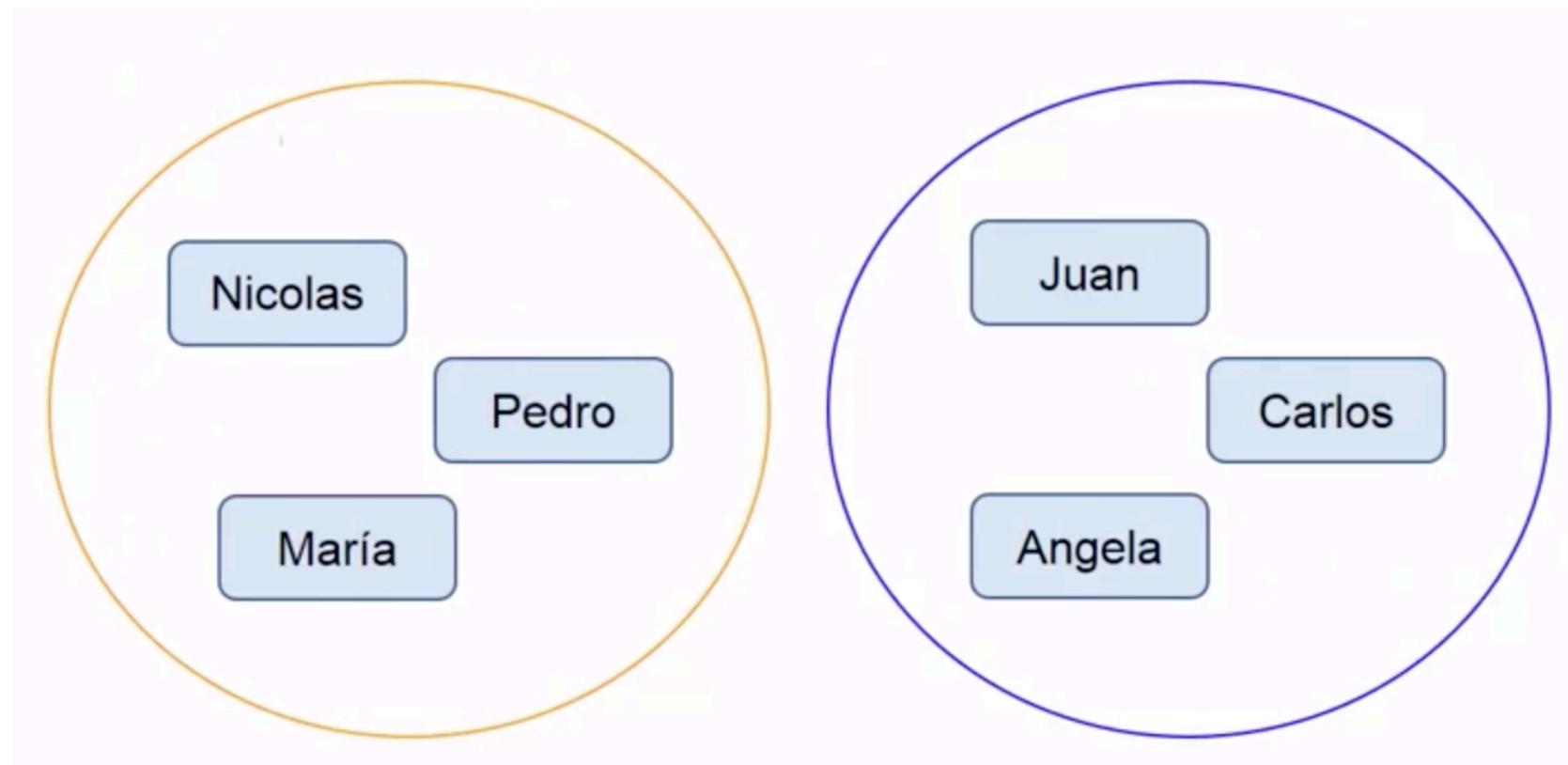
$$d(C_2, \text{Angela}) = 0.22$$

$$d(C_2, \text{Nicolás}) = 1$$

$$d(C_2, \text{Carlos}) = 0.37$$

# Ejemplo

- ▶ Se seleccionan los registros asociados al cluster:

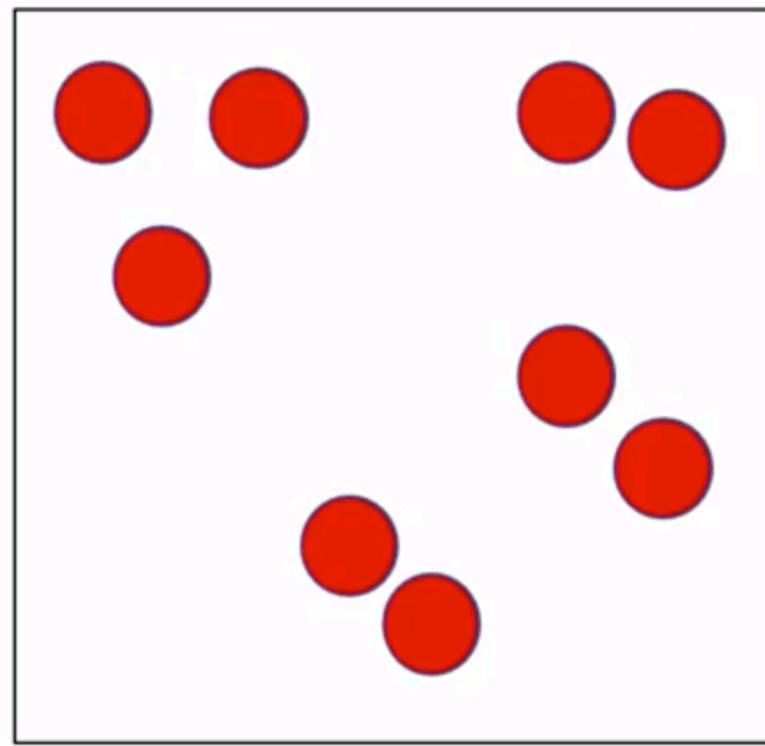


# Clustering jerárquico aglomerativo

- ▶ Es un método bastante sencillo y útil en la práctica
- ▶ La idea principal del clustering jerárquico es que a partir de una medida de similaridad, se van juntando paso a paso los puntos más cercanos dentro de los datos, generando una jerarquía de resultados de clustering.

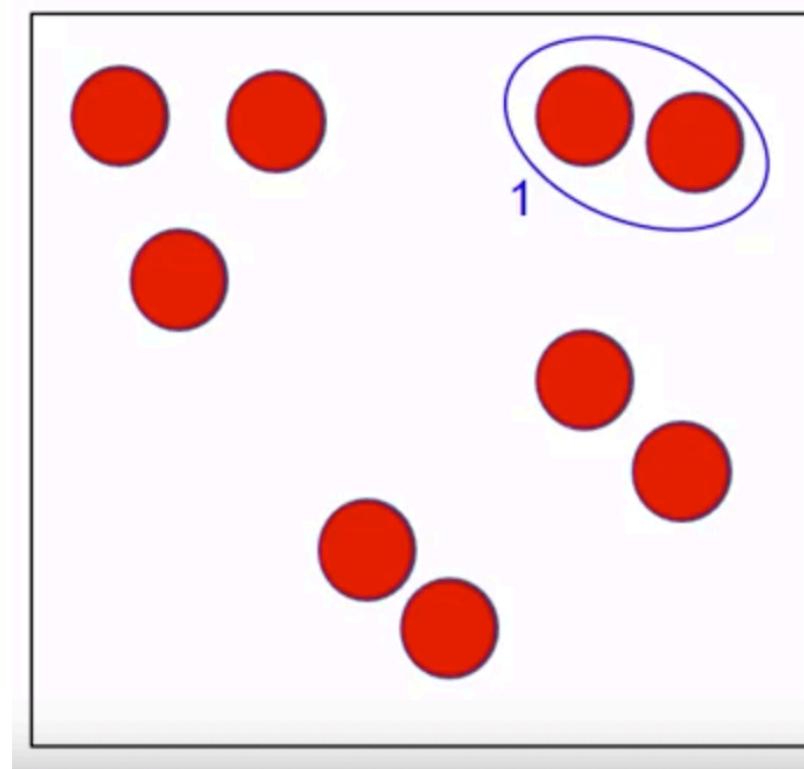
# Clustering jerárquico aglomerativo

- ▶ En cada iteración se van juntando el par de cluster más cercano
- ▶ Cada punto es un cluster por si solo



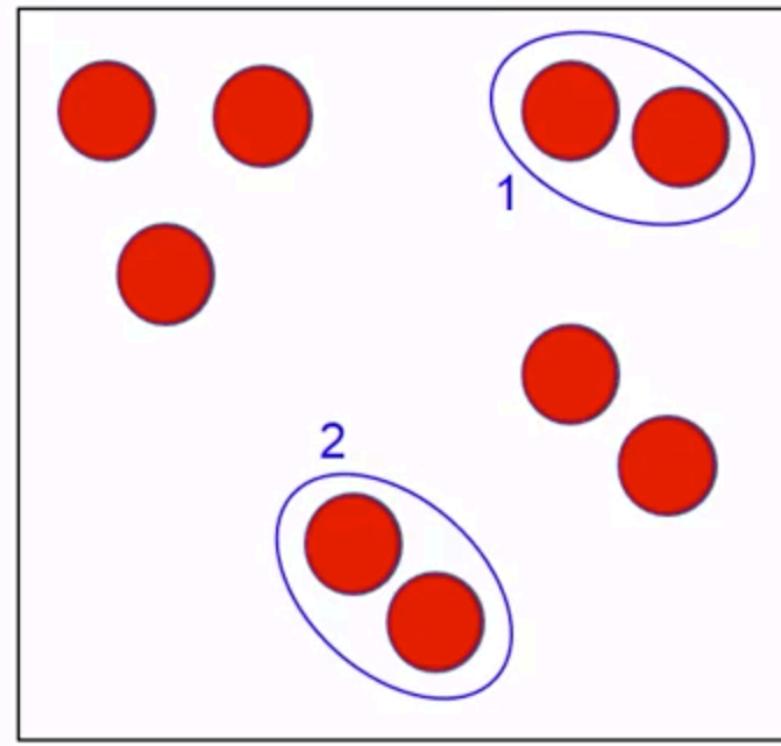
# Clustering jerárquico aglomerativo

- ▶ Los 2 puntos pasan a formar un cluster llamado 1



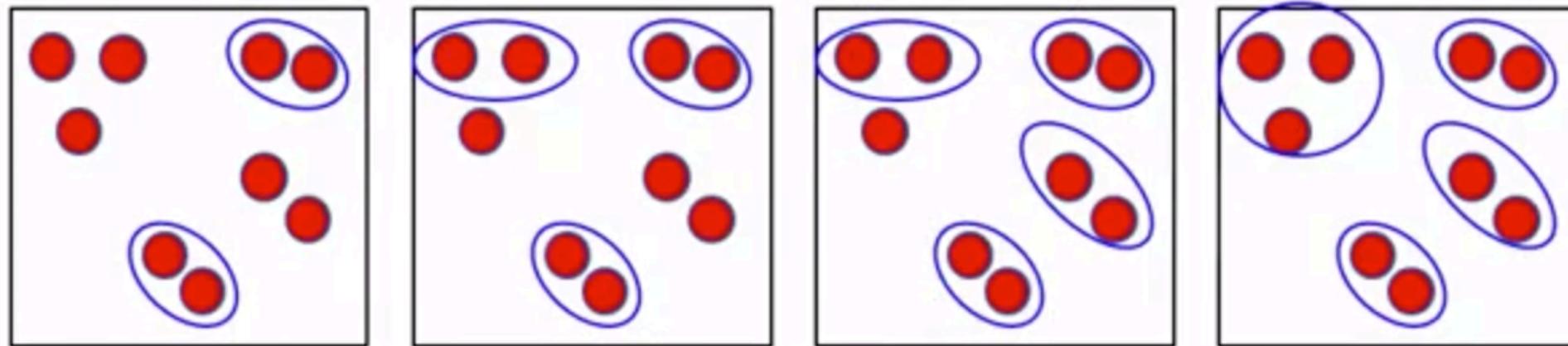
# Clustering jerárquico aglomerativo

- Se debe continuar con juntar los clusters más cercanos, En este ejemplo es el cluster 2



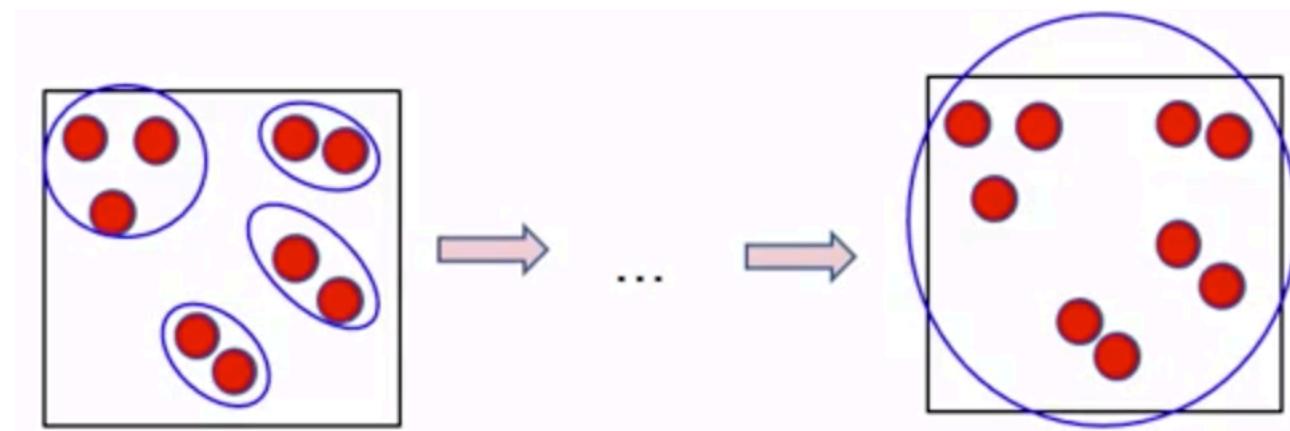
# Clustering jerárquico aglomerativo

- Se continúa juntando el par de clusters más cercano



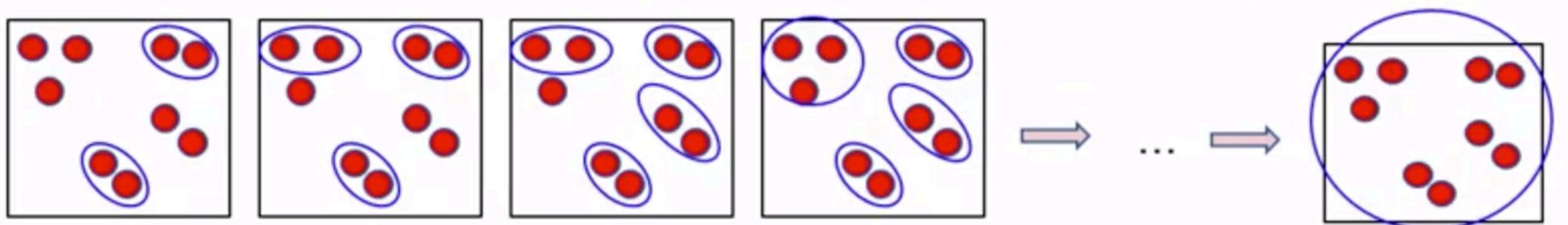
# Clustering jerárquico aglomerativo

- ▶ Finalmente, se llegará a un solo cluster que contendrá todos los registros
- ▶ Siempre es posible detener el algoritmo antes de llegar a crear un gran cluster. En este caso se debe contar con algún criterio



# Clustering jerárquico aglomerativo

- ▶ Entre los criterios de detención del algoritmo se encuentran:
  - ▶ Número mínimo de clusters
  - ▶ Umbral de distancia máxima
  - ▶ Número máximo de iteraciones.
- ▶ Estos criterios aplican de manera diferente en diversas bases de datos por lo que requiere un entendimiento del set de datos antes de ser aplicado



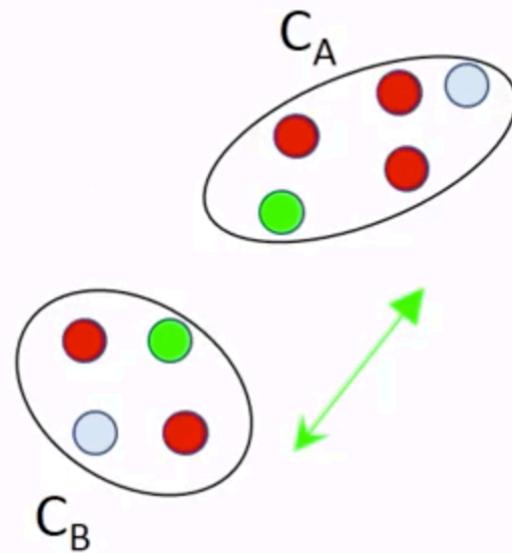
# Clustering jerárquico aglomerativo

- ▶ Además se necesita medir la distancia entre clusters. Algunas son:
  - ▶ Conexión simple
  - ▶ Conexión completa
  - ▶ Distancia entre medias
  - ▶ Distancia promedio entre pares

# Conexión Simple

- ▶ Se calcula la mínima distancia entre todos los pares de puntos de ambos clusters y se selecciona la menor

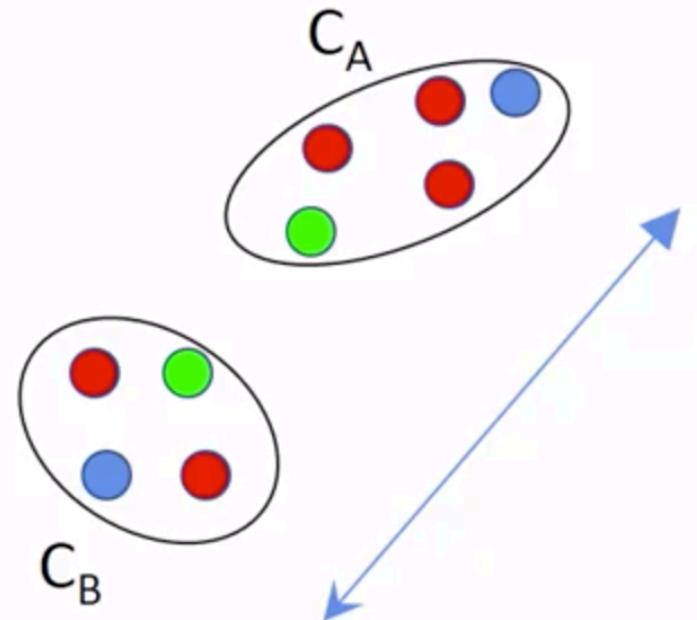
$$D(C_a, C_b) = \text{Min}\{d(i, j)\}, \forall i \in C_a, \forall j \in C_b$$



# Conexión completa

- Es lo opuesto a conexión simple

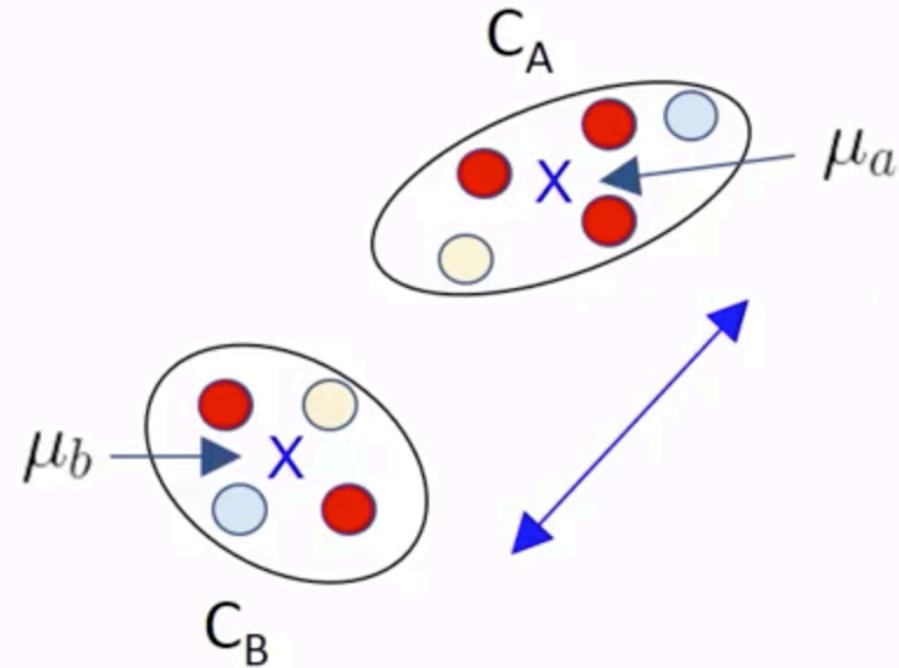
$$D(C_a, C_b) = \text{Max}\{d(i, j)\}, \forall i \in C_a, \forall j \in C_b$$



# Conexión distancia entre medias

- Es la distancia que existe entre los centroides de cada grupo

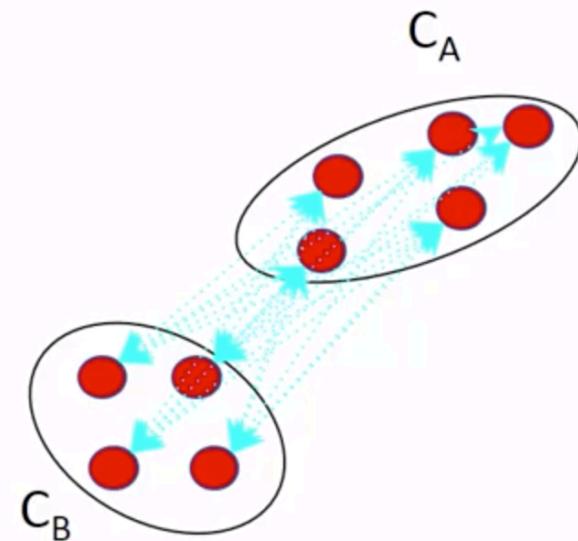
$$D(C_a, C_b) = d(\mu_a, \mu_b)$$



# Conexión distancia entre medias

- ▶ Corresponde al promedio de distancias

$$D(C_a, C_b) = \text{promedio}\{d(i, j)\}, \forall i \in C_a, \forall j \in C_b$$



# REFERENCIAS

- ▶ Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- ▶ Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- ▶ Hand, D. J. (2006). Data Mining. *Encyclopedia of Environmetrics*, 2.
- ▶ Introducción a la Minería de Datos, Karim Pichara, coursera.