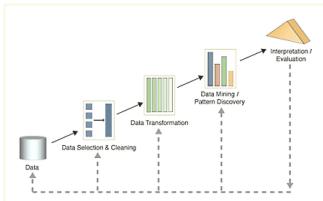
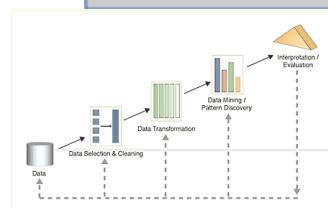
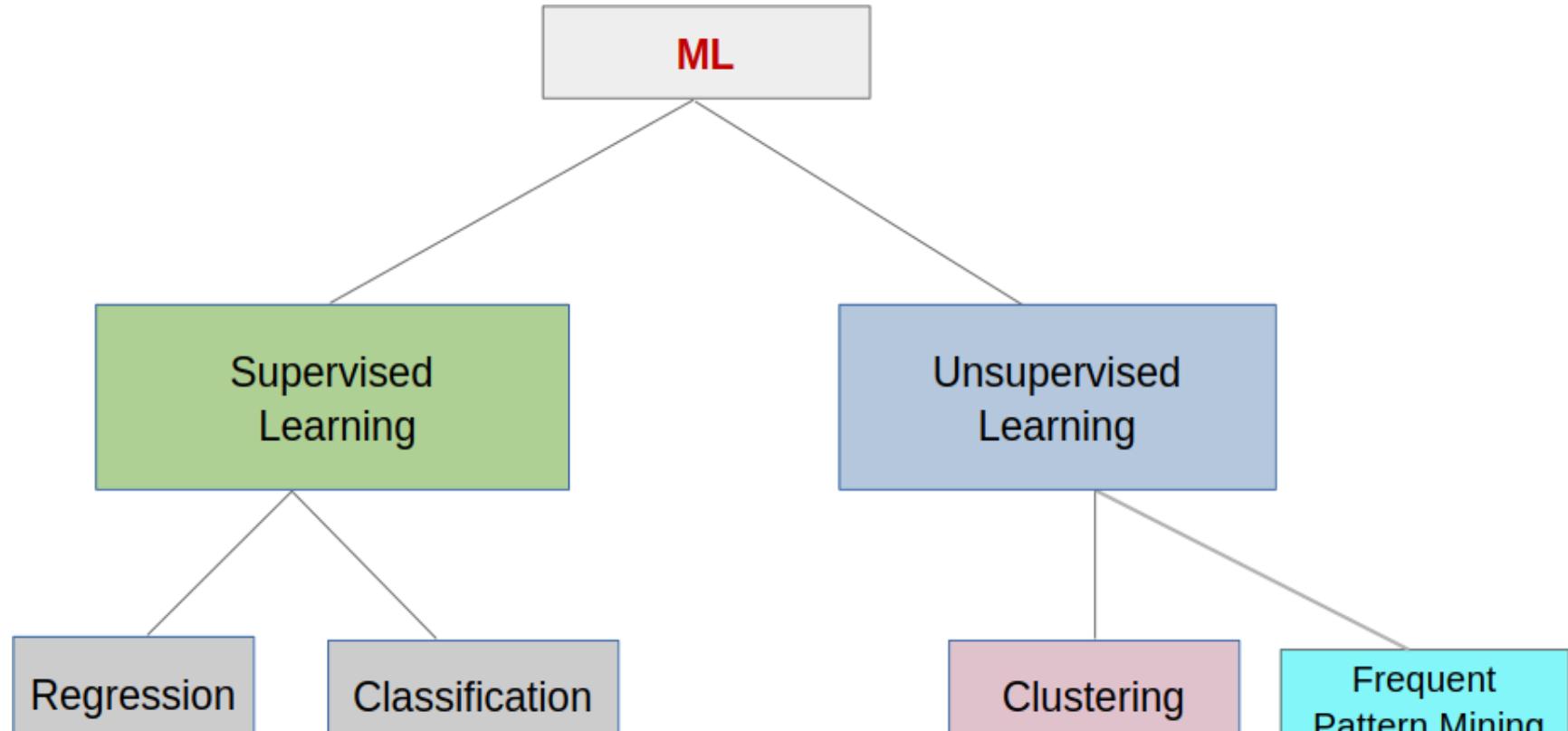


# Segmentación o Clustering



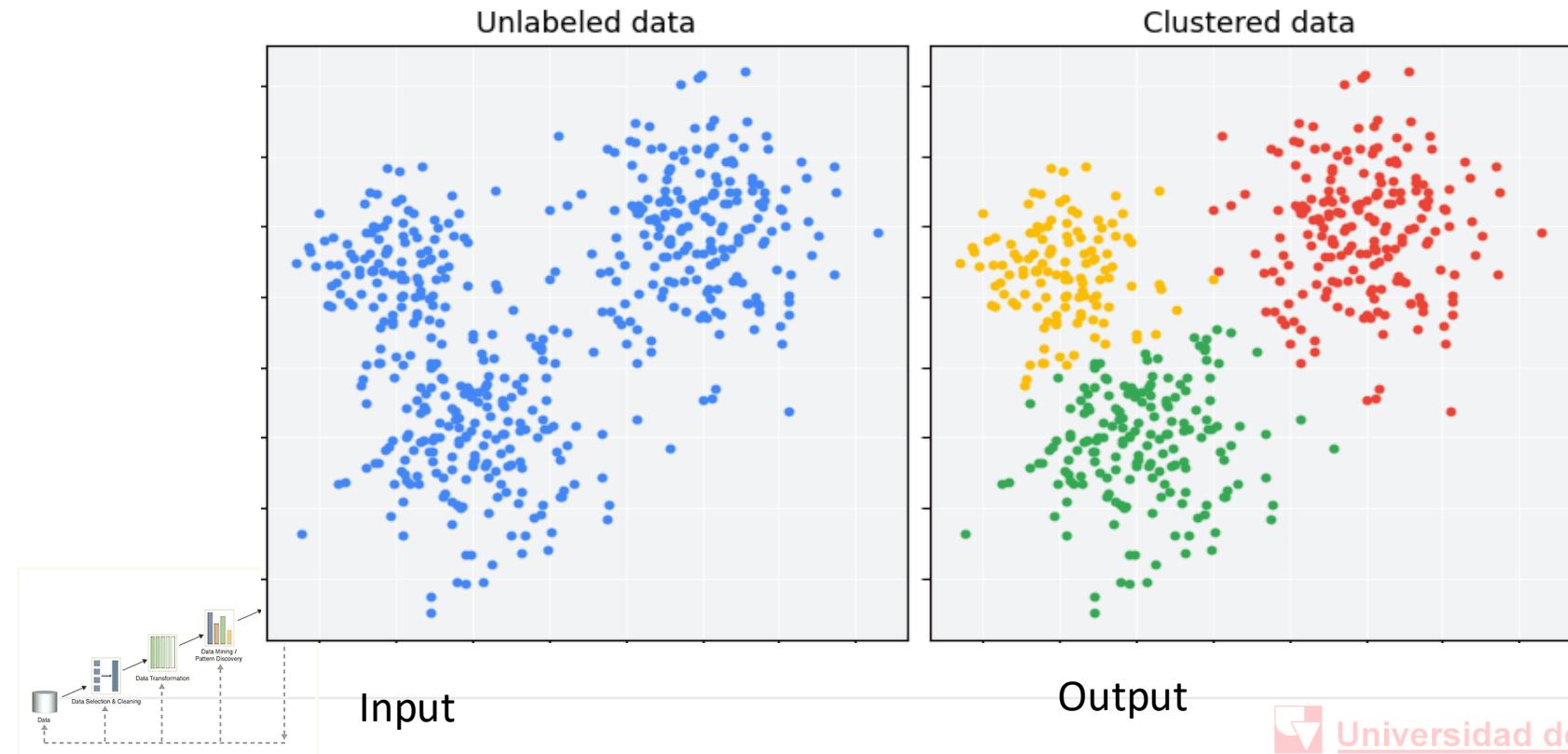
# Segmentación o Clustering



# Segmentación o Clustering

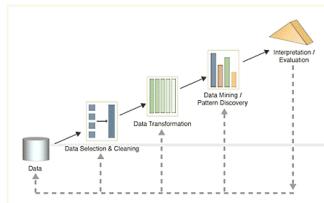
Es el proceso de dividir un conjunto de objetos de datos en subconjuntos.

Cada subconjunto es un grupo, de modo que los objetos de un grupo son **similares entre sí, pero diferentes a los objetos de otros grupos.**



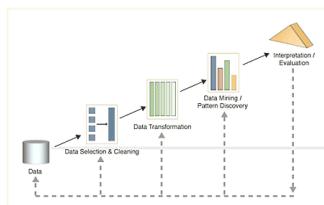
# Segmentación o Clustering

- No se requiere datos etiquetados
- Clustering: proceso de **agrupar** objetos similares: clusters
- Medidas de Similitud y distancia entre objetos
- Aprendizaje no supervisado



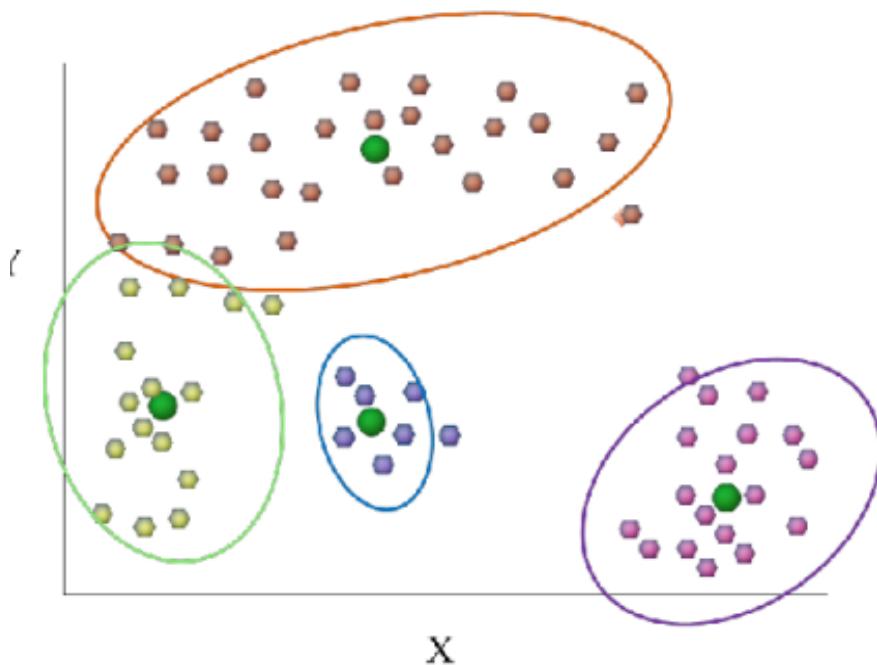
# Segmentación o Clustering

- MÉTODOS
  - Particionales
  - Jerárquicos
- Basados en
  - Densidad
  - Grid
  - Modelos



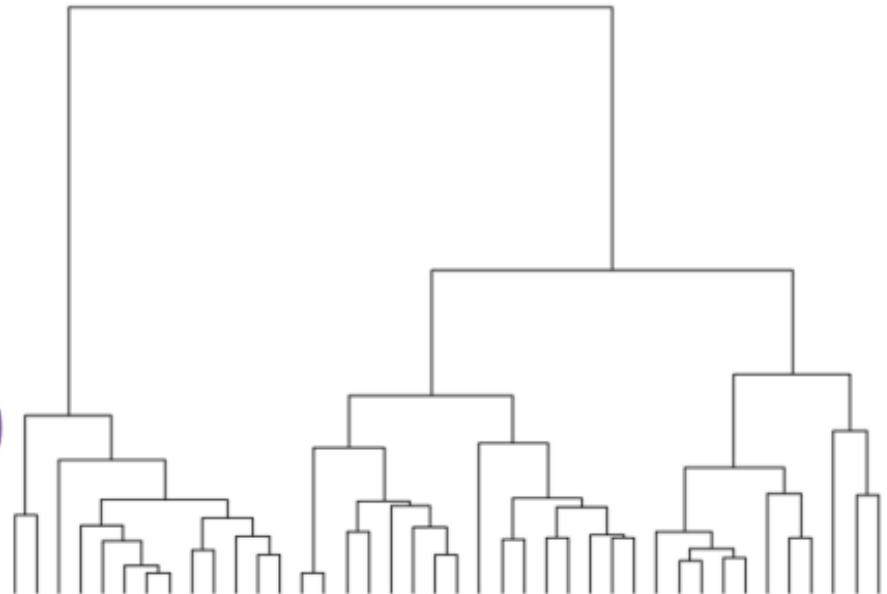
# Tipos de Clustering

**Centroid-Based  
Clustering**



**Particionales**

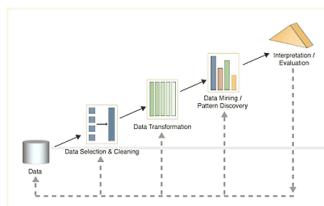
**Hierarchical  
Clustering**



**Jerárquico (Dendograma)**

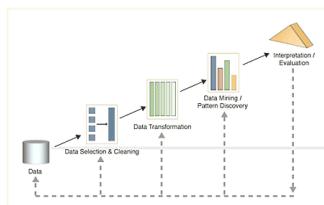
# Clustering: requisitos

- No necesita etiqueta de clase
- Número máximo de clusters
- Número de iteraciones



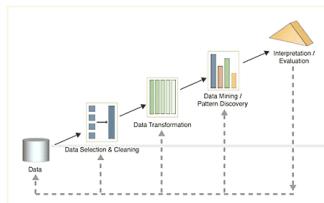
# Aplicaciones

- Segmentación de mercado
- Análisis de redes sociales
- Agrupamiento de resultados de búsqueda (Eg. Google)
- Imágenes médicas
- Segmentación de imágenes
- Detección de anomalías



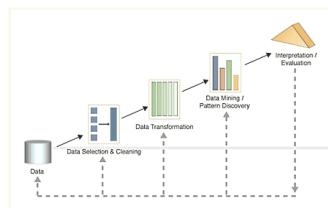
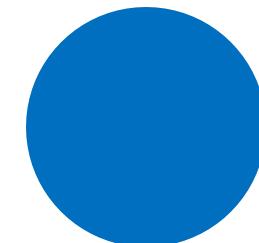
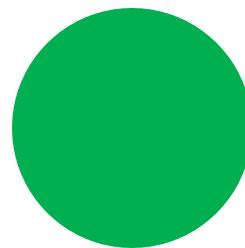
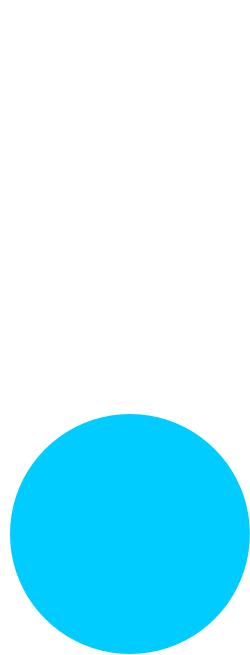
# Bondad del cluster

- Un buen método de clustering : clusters de calidad
  - Alta similitud dentro de cada clase
  - Baja similitud de elementos de distintas clases
- La calidad depende de la medida de similitud que se utilice
- La calidad medida por su capacidad para descubrir patrones ocultos



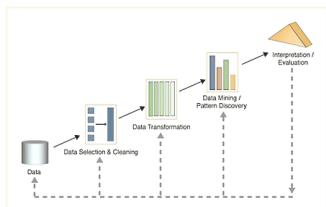
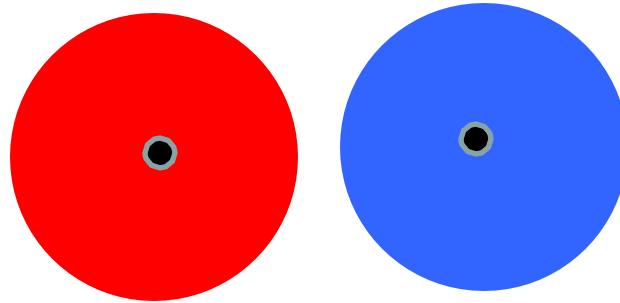
# Tipos de cluster

- Cluster bien separados



# Tipos de cluster

- Basados en un centroide

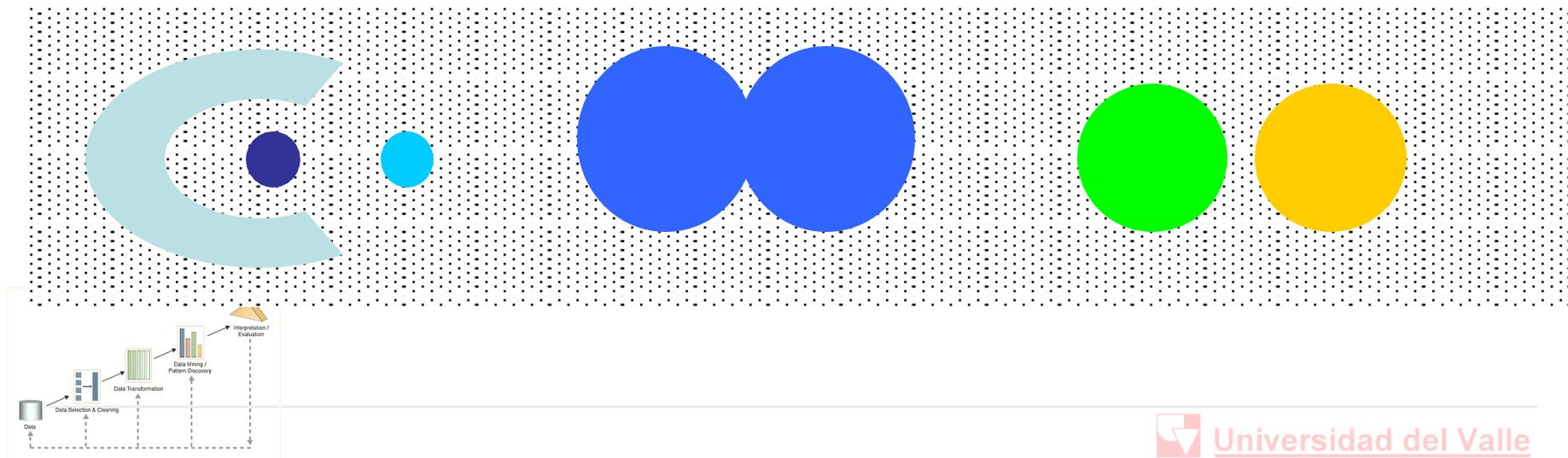


# Tipos de cluster

- Basados en densidad

Un clúster es una región densa de puntos, que está separada por las regiones de baja densidad, de otras regiones de alta densidad

Se utiliza cuando los grupos son irregulares o entrelazados, y cuando el ruido y valores atípicos están presentes



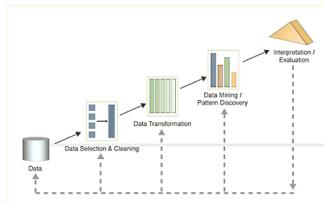
# Estructuras de datos

## Matriz de similitud

	G1	G2	G3	G4
G1	1	0.83	0	0
G2	0.83	1	0	0
G3	0	0	1	0.32
G4	0	0	0.32	1

## Matriz de distancias

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



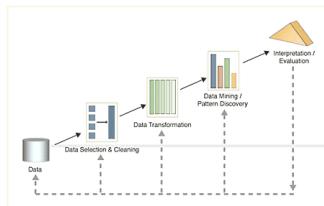
# Estructuras de datos

## Matriz de similitud

	G1	G2	G3	G4
G1	1	0.83	0	0
G2	0.83	1	0	0
G3	0	0	1	0.32
G4	0	0	0.32	1

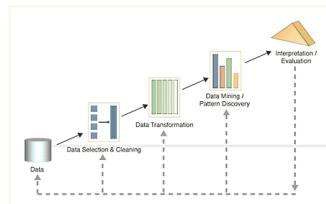
## Matriz de distancias

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



# Tipos de datos posibles

- Variables de intervalo: medidas continuas
- Variables binarias
- Nominales, ordinales
- Variables mixtas

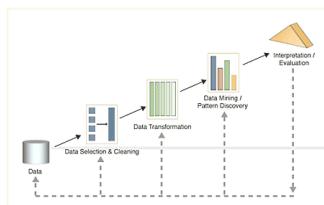


# Similitud entre objetos

Distancia Euclídea

$$d_{(p1,p2)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

También se pueden utilizar pesos para dar más importancia a ciertas variables



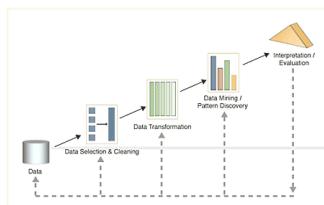
# Distancia entre objetos

En general, la distancia euclíadiana entre los puntos:

$$P = (p_1, p_2, \dots, p_n) \quad \text{y} \quad Q = (q_1, q_2, \dots, q_n)$$

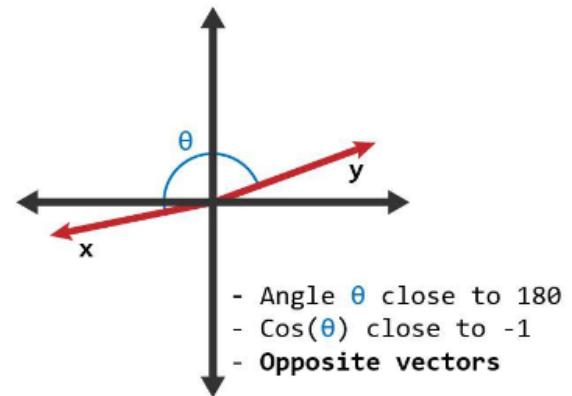
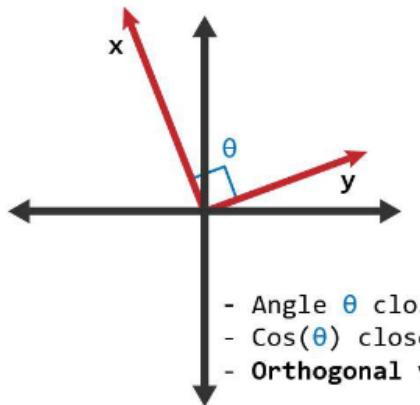
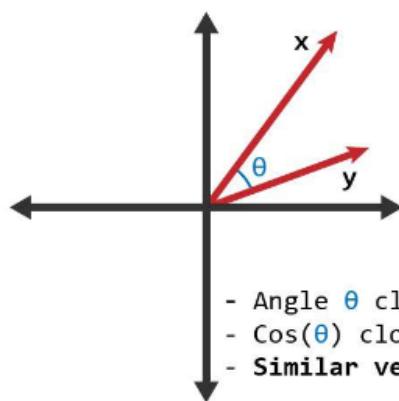
Esta dada por:

$$d_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}.$$



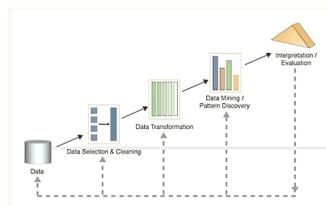
# Similitud entre objetos

## Similitud Coseno



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

1,352 x 482

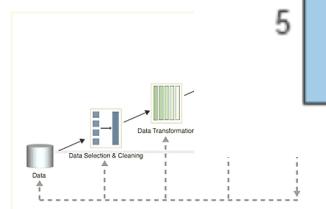
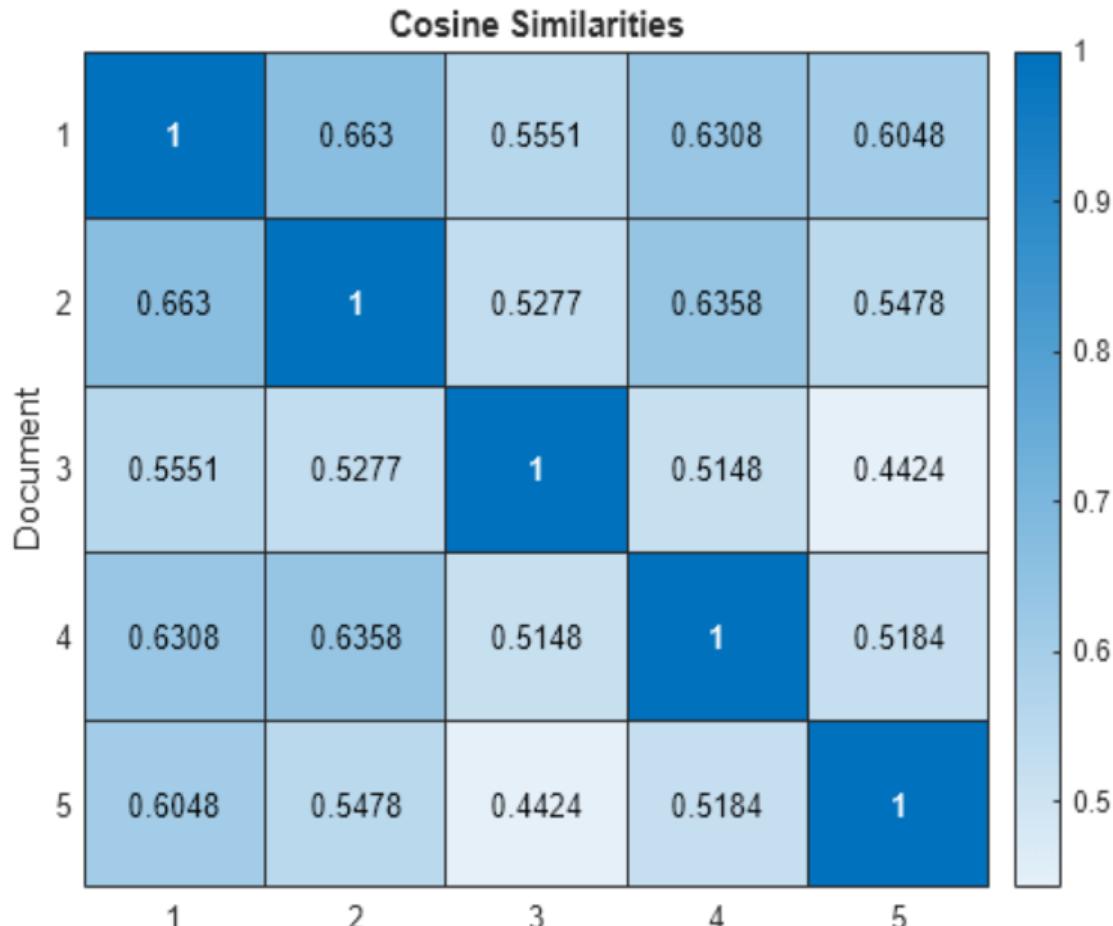


# Similitud entre objetos

## Similitud Coseno

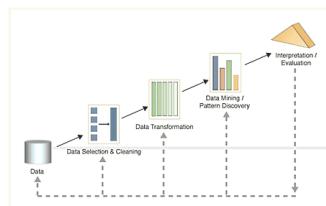
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

1,352 × 482

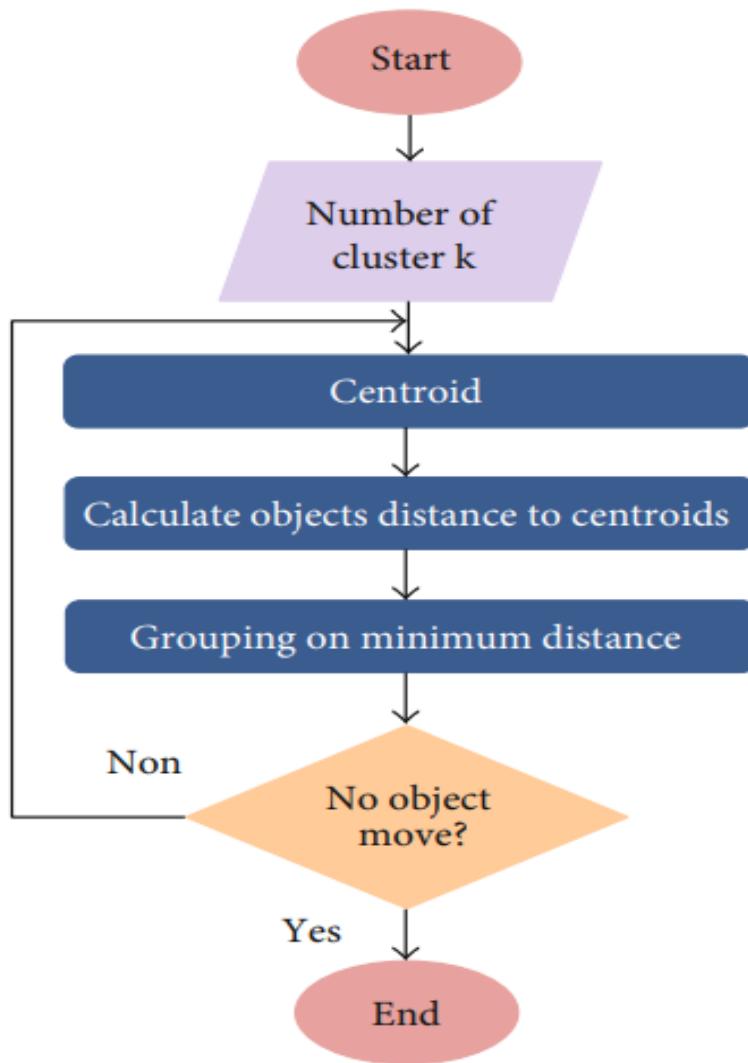


# Enfoques más importantes

- **Algoritmos divisivos:** Construyen varias particiones y luego las evalúan
- **Jerárquicos:** Crean una descomposición jerárquica
- **Basados en Densidad:** utilizan funciones de densidad



# El método K-Means



Iterate until stable (= no object move group):

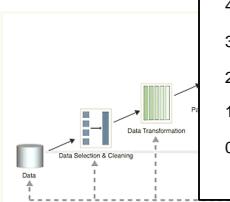
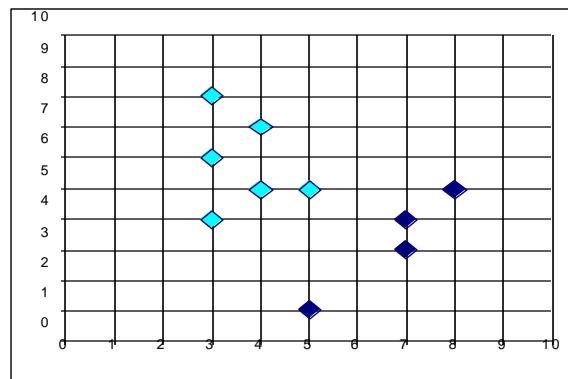
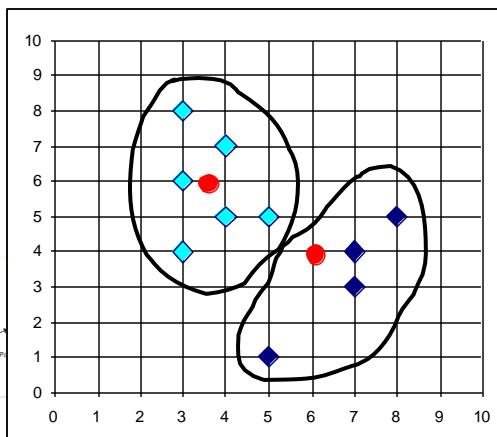
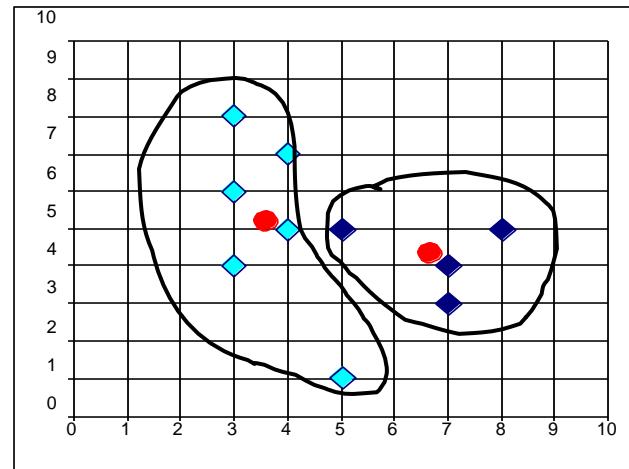
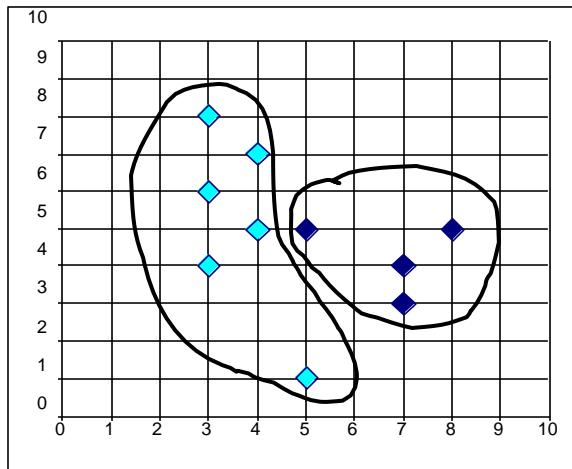
1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance

$$d_{(p_1,p_2)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

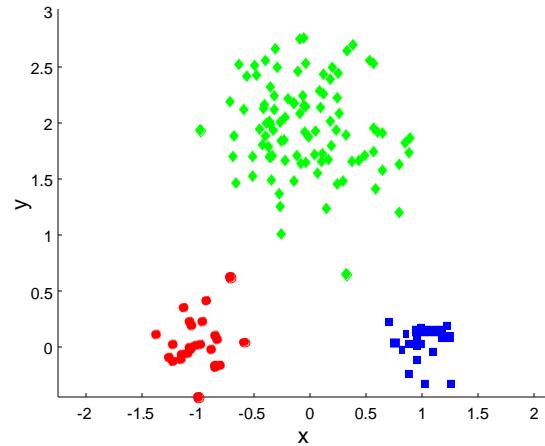


# El método de las K-medias

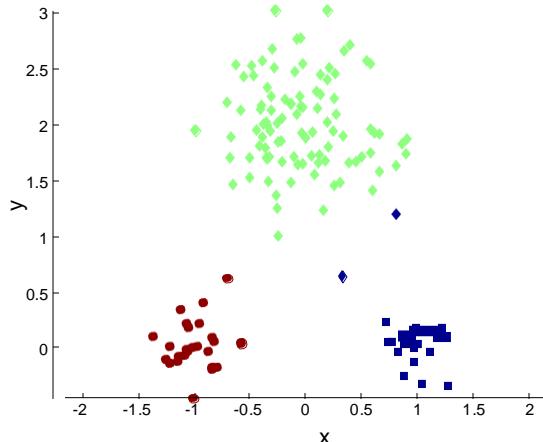
## Ejemplo 1



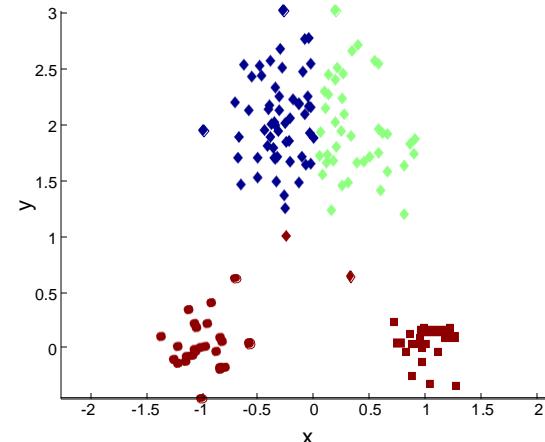
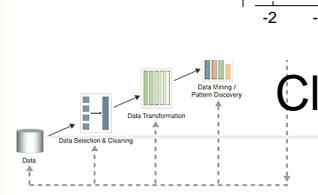
# El método de las K-medias



Puntos originales



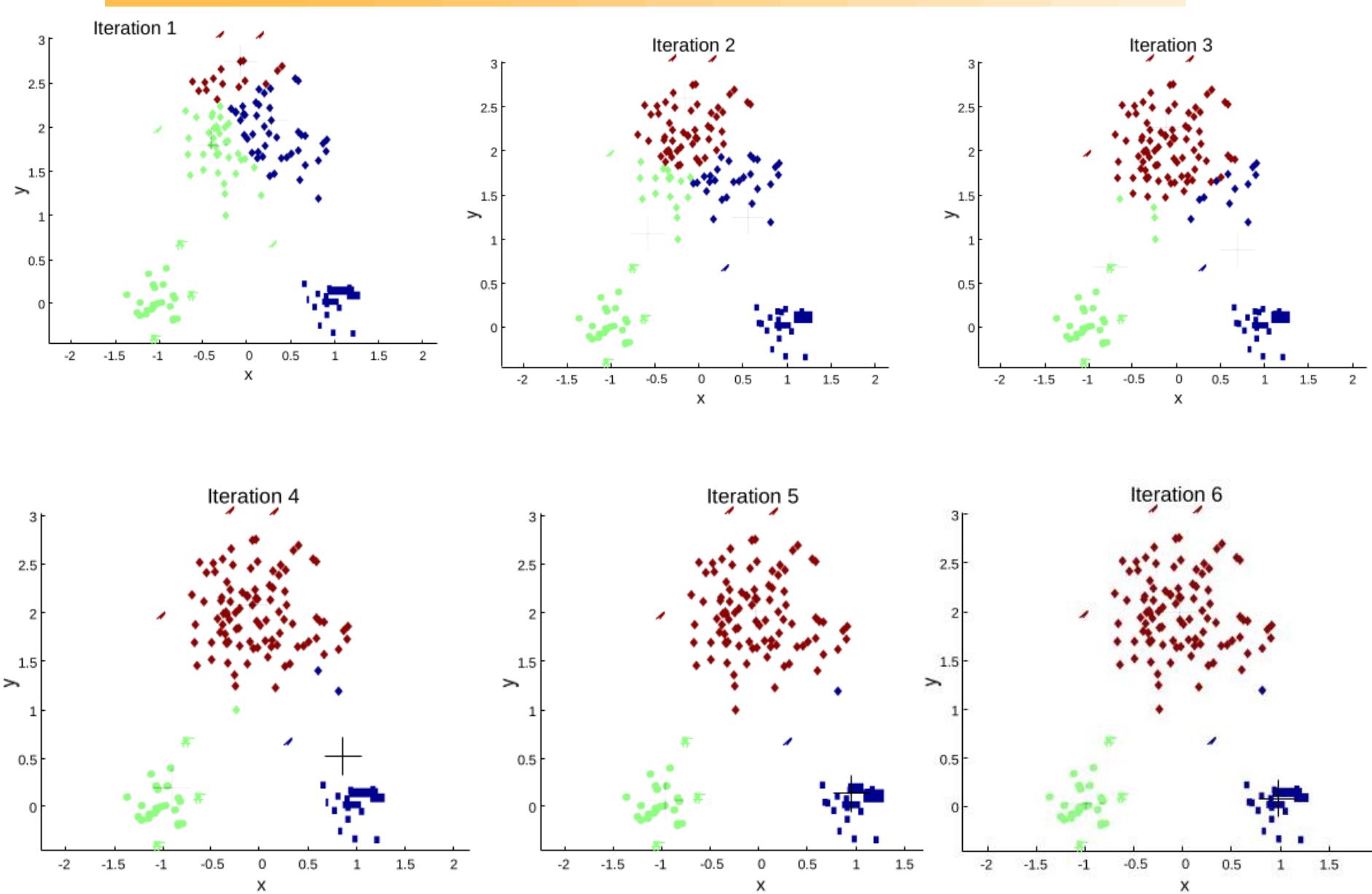
Clustering óptimo



Clustering sub óptimo

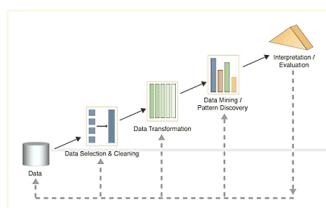
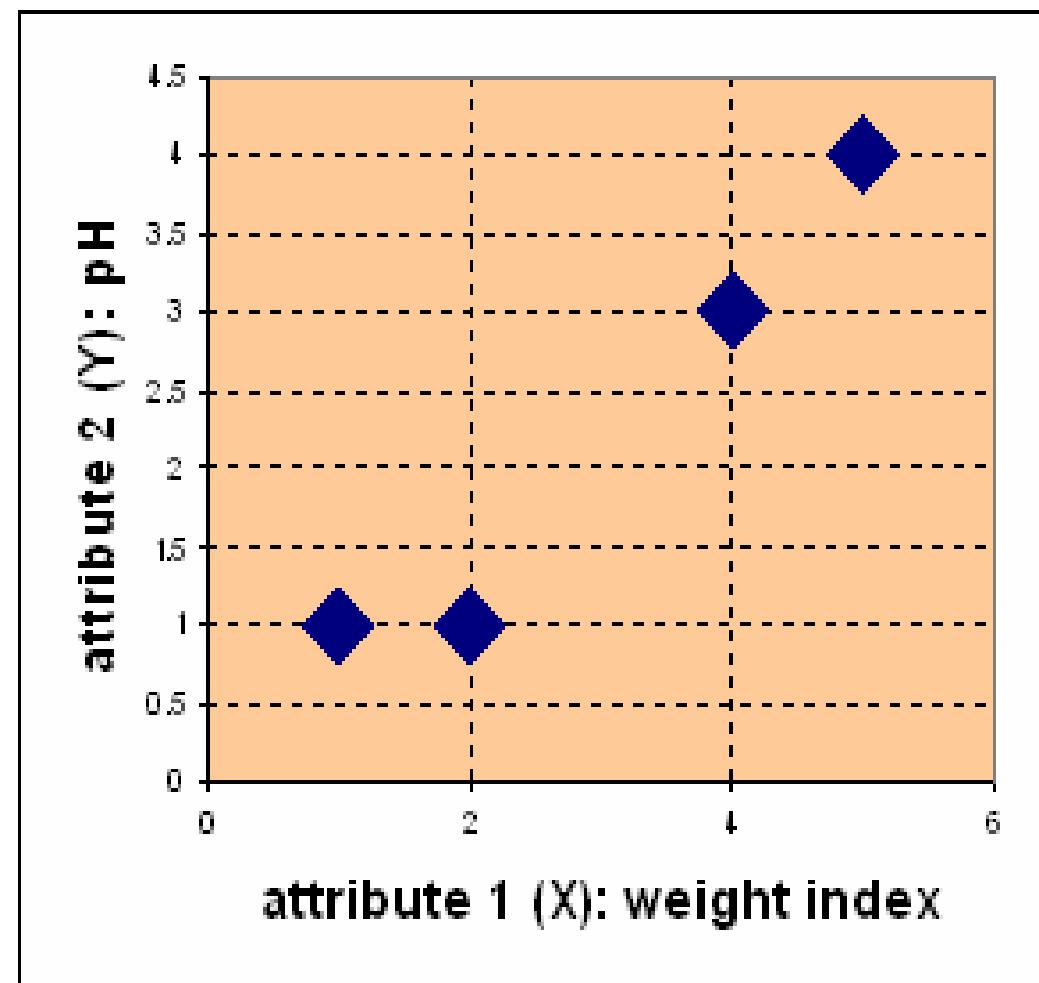


# K-means:Ejemplo



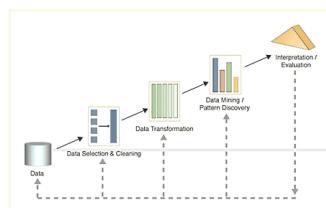
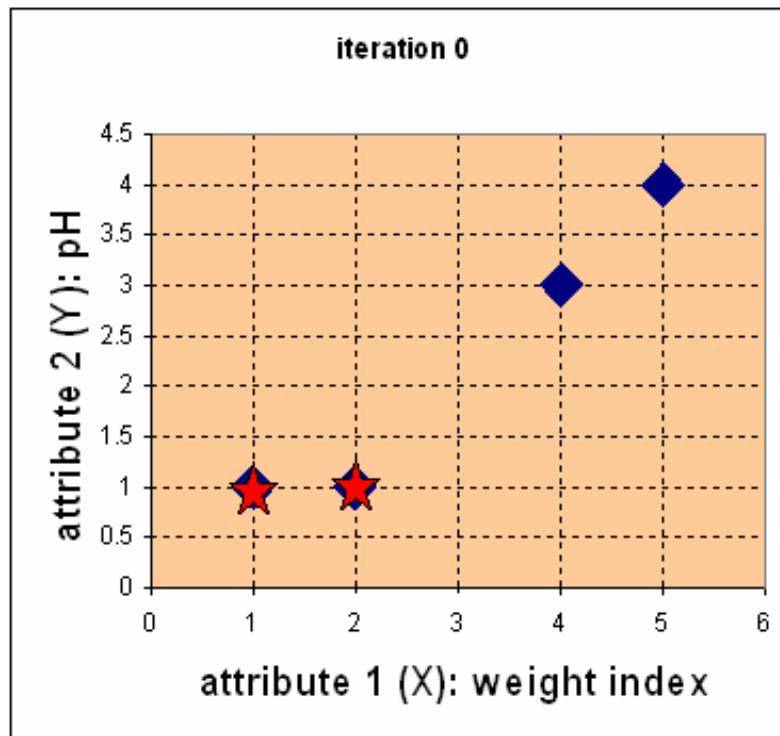
# K-means:Ejemplo

Medicine	Weight (X)	PH (Y)
A	1	1
B	2	1
C	4	3
D	5	4



# K-means:Ejemplo

Centroides Iniciales  $c_1 = (1, 1)$        $c_2 = (2, 1)$



# K-means:Ejemplo

2. Calcular la distancia entre los centroides y los demás objetos

$$c_1 = (1,1) \quad \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

Distancia de la medicina C (4,3) con el centroide (1,1) y (2,1)

$$c_2 = (2,1) \quad \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

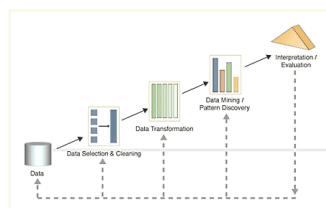
A    B    C    D

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad X \quad Y$$

$$c_1 = (1,1) \quad group - 1$$

$$c_2 = (2,1) \quad group - 2$$

Como está más cerca de **c2**, la medicina C se asigna al cluster 2



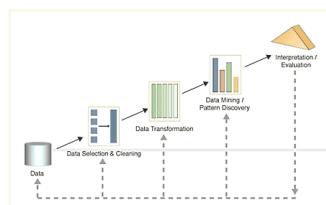
# K-means:Ejemplo

## 3. Asignar los objetos a los cluster de acuerdo a la distancia mínima

La medicina A se asigna al cluster 1, la medicina B, C y D al cluster 2

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{array}{l} group - 1 \\ group - 2 \end{array}$$

*A    B    C    D*



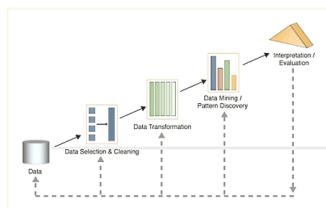
# K-means:Ejemplo

4. (Iteracion-1) Ajustar los centroides con la información del cluster

Cluster 1:  $c_1 = (1, 1)$

Cluster 2:  $c_2 = \left( \frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left( \frac{11}{3}, \frac{8}{3} \right)$

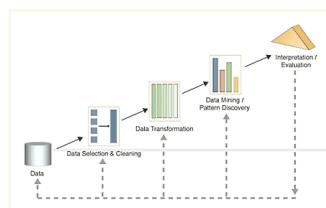
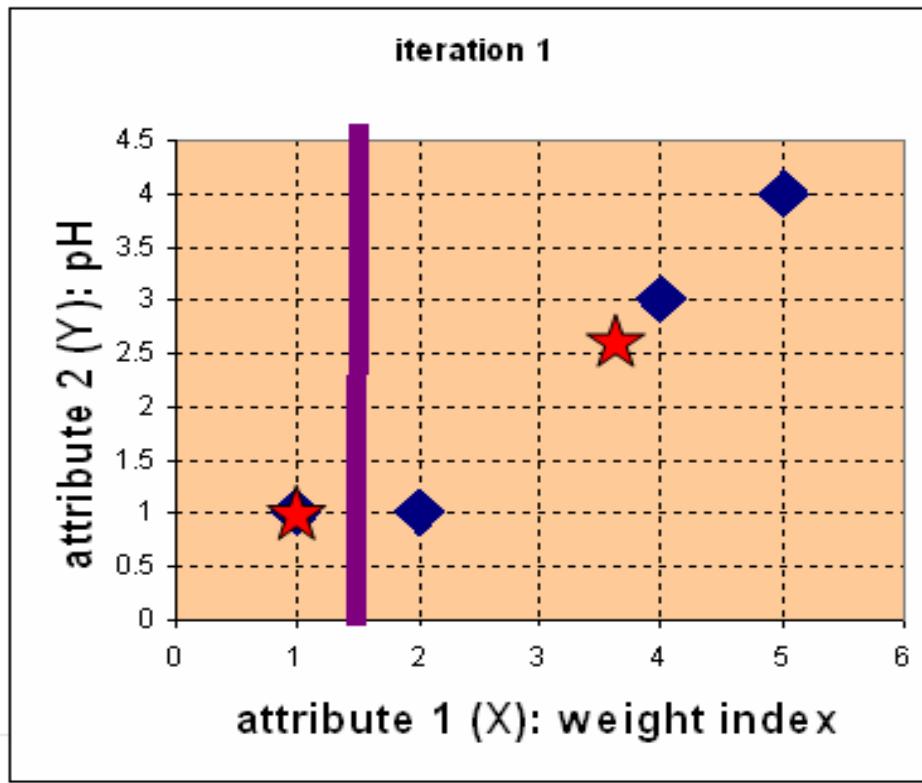
Medicine	Weight (X)	PH (Y)
A	1	1
B	2	1
C	4	3
D	5	4



# K-means:Ejemplo

4. . (Iteracion-1) Ajustar los centroides con la información del cluster

$$\star c_1 = (1, 1) \quad \star c_2 = \left( \frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left( \frac{11}{3}, \frac{8}{3} \right)$$

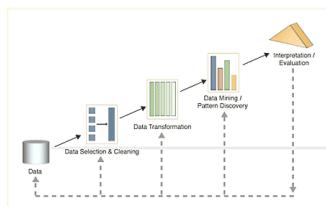


# K-means:Ejemplo

5. . (Iteracion-1) Calcular la distancia entre todos los objetos y los nuevos centroides, como en el paso 2

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \mathbf{c}_1 = (1,1) \quad group - 1$$
$$\mathbf{c}_2 = \left(\frac{11}{3}, \frac{8}{3}\right) \quad group - 2$$

$$\begin{array}{cccc} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \end{bmatrix} & X \\ \begin{bmatrix} 1 & 1 & 3 & 4 \end{bmatrix} & Y \end{array}$$



# K-means:Ejemplo

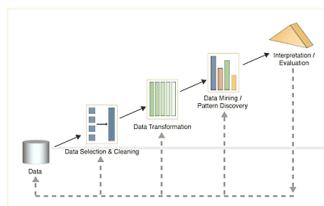
6. . (Iteracion-1) Asignar los objetos a cada cluster basándose en la matriz de distancia, como en el paso 3.

$$\mathbf{G}^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group - 1 \\ group - 2 \end{matrix}$$

*A    B    C    D*

***La medicina B, se mueve al cluster 1***

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

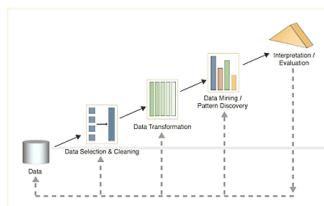
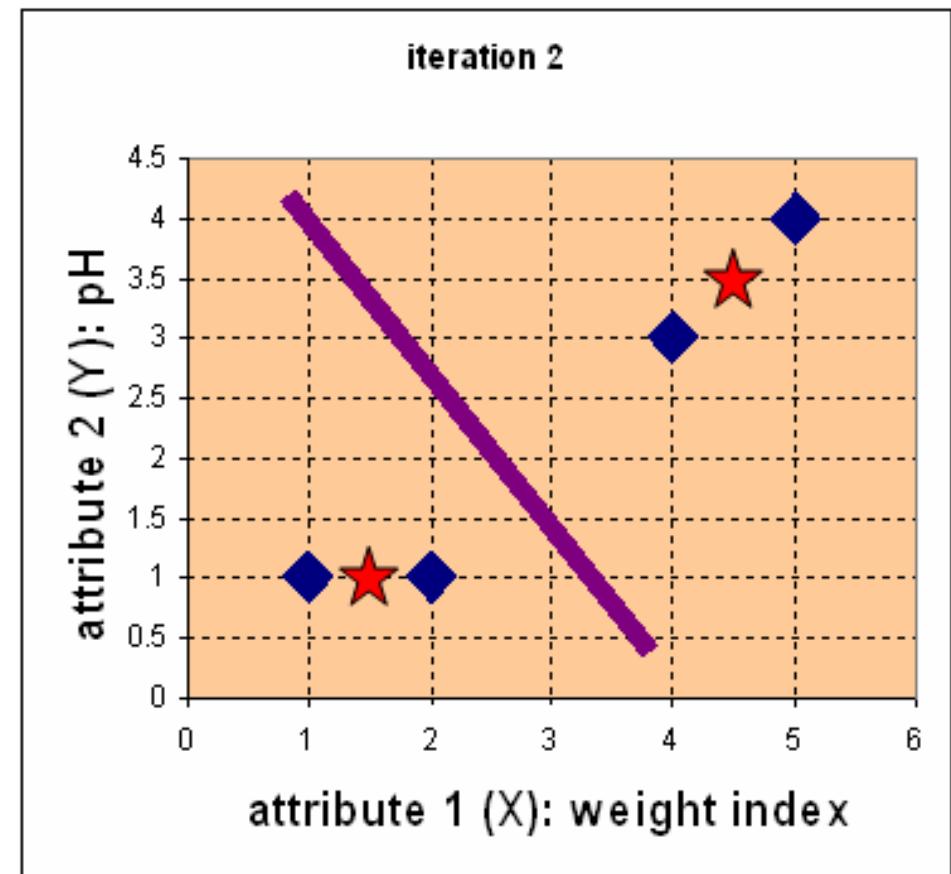


# K-means:Ejemplo

7. . (Iteracion-2) Determinar los centroides nuevamente

$$c_1 = \left( \frac{1+2}{2}, \frac{1+1}{2} \right) = (1\frac{1}{2}, 1)$$

$$c_2 = \left( \frac{4+5}{2}, \frac{3+4}{2} \right) = (4\frac{1}{2}, 3\frac{1}{2})$$



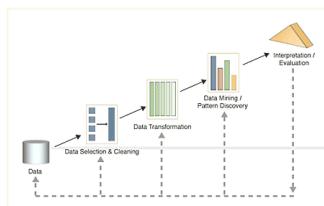
# K-means:Ejemplo

8. . (Iteracion-2) Repetir el paso 2, calcular la distancia.....

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \mathbf{c}_1 = (1\frac{1}{2}, 1) \quad group - 1$$
$$\mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) \quad group - 2$$

*A      B      C      D*

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad X$$
$$Y$$



# K-means:Ejemplo

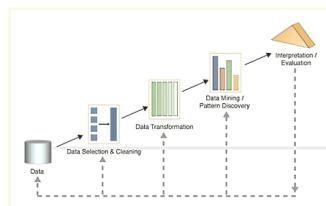
8. . (Iteracion-2) Asignar los objetos a cada cluster

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{array}{l} group - 1 \\ group - 2 \end{array}$$

*A    B    C    D*

**Los objetos no cambian respecto a la iteración anterior, entonces:**

**El algoritmo termina.**



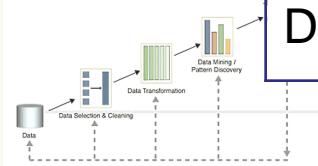
# K-means:Ejemplo

9. (Iteracion-2) Asignar los objetos a cada cluster

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{array}{l} group - 1 \\ group - 2 \end{array}$$

*A    B    C    D*

Medicine	Weight (X)	PH (Y)	Result
A	1	1	1
B	2	1	1
C	4	3	2
D	5	4	2



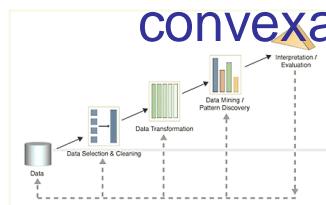
# Comentarios sobre las K-means

## Ventajas

- Relativamente eficiente
- Generalmente termina con un óptimo local.

## Debilidades

- Solo es aplicable cuando la media está definida. ¿datos categóricos?
- Se necesita especificar K de antemano.
- No es capaz de tratar con ruido
- No es apropiado para descubrir cluster que no tengan formas no convexas



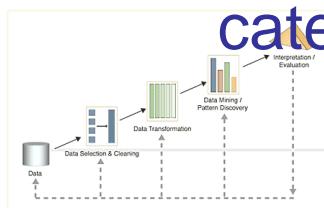
# Variaciones del método de las K-Means

Las variantes se diferencian por:

- Selección de las k medias iniciales
- Cálculo de similitudes
- Estrategias para calcular las medias

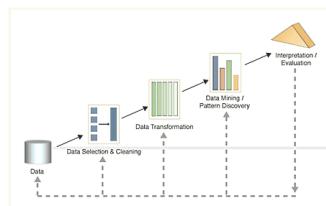
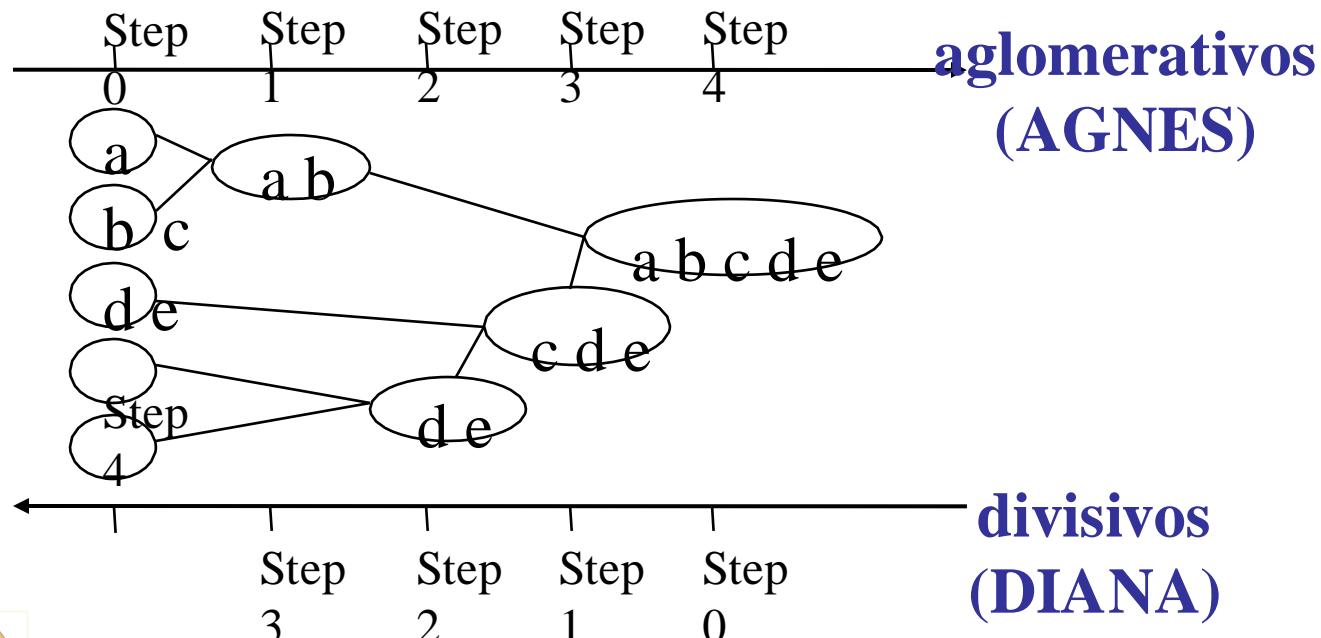
## Tratamiento de datos categóricos

- Usar las medidas de similitud para los objetos categóricos



# Métodos jerárquicos

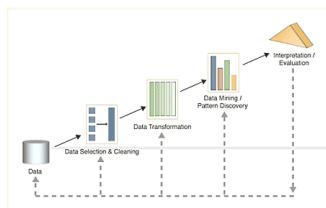
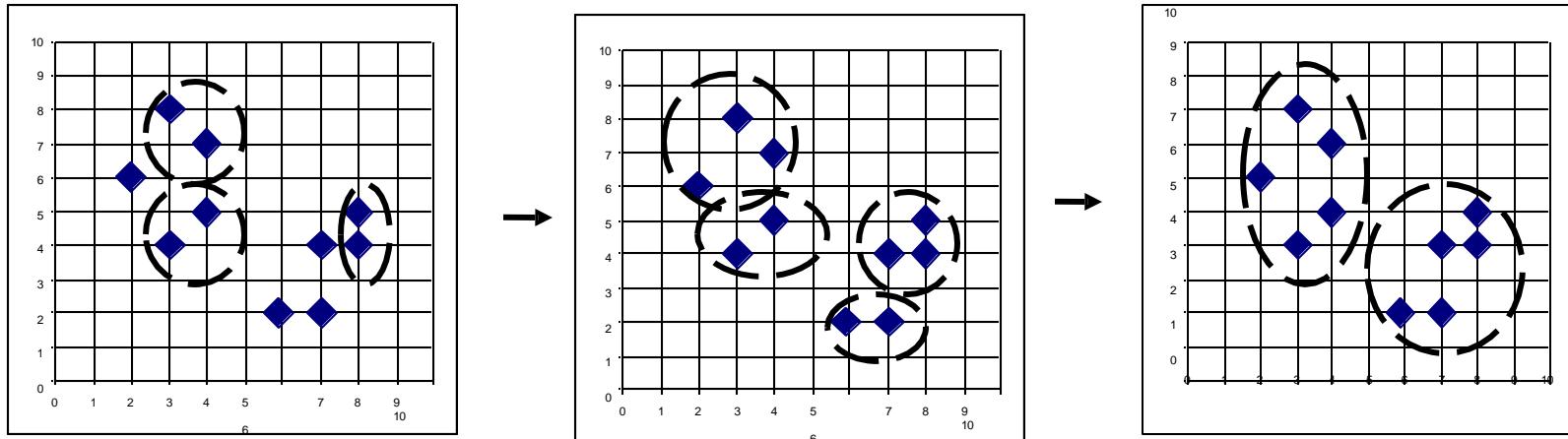
Utilizan la matriz de distancias. NO requiere el número de cluster como entrada pero necesita una condición de terminación



# AGNES (Agglomerative Nesting)

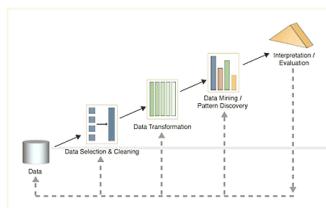
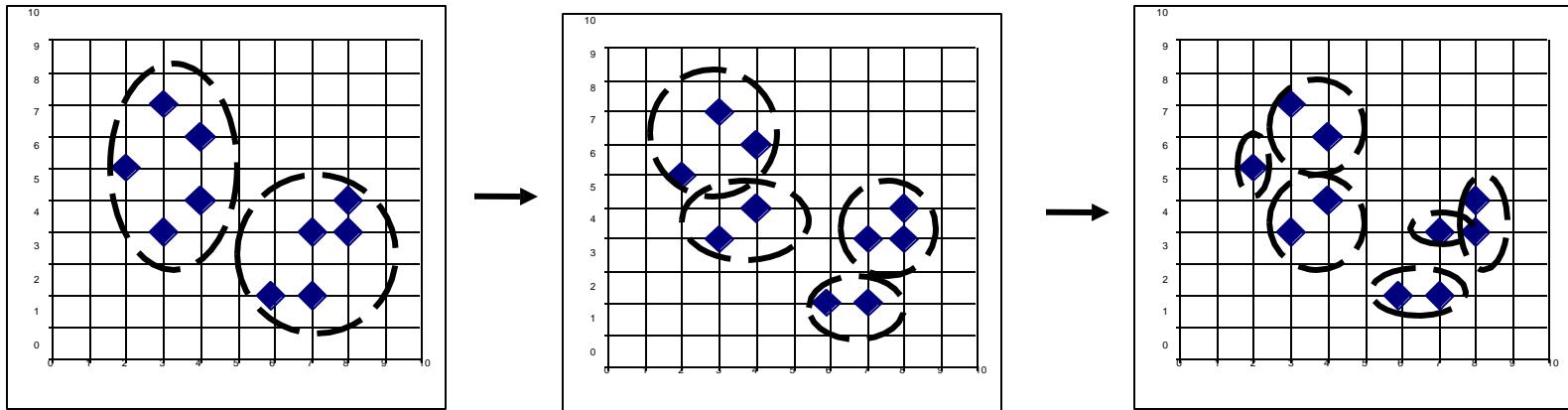
Está en los paquetes estadísticos

Une los nodos que tengan la menor disimilaridad.

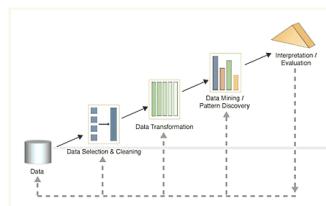
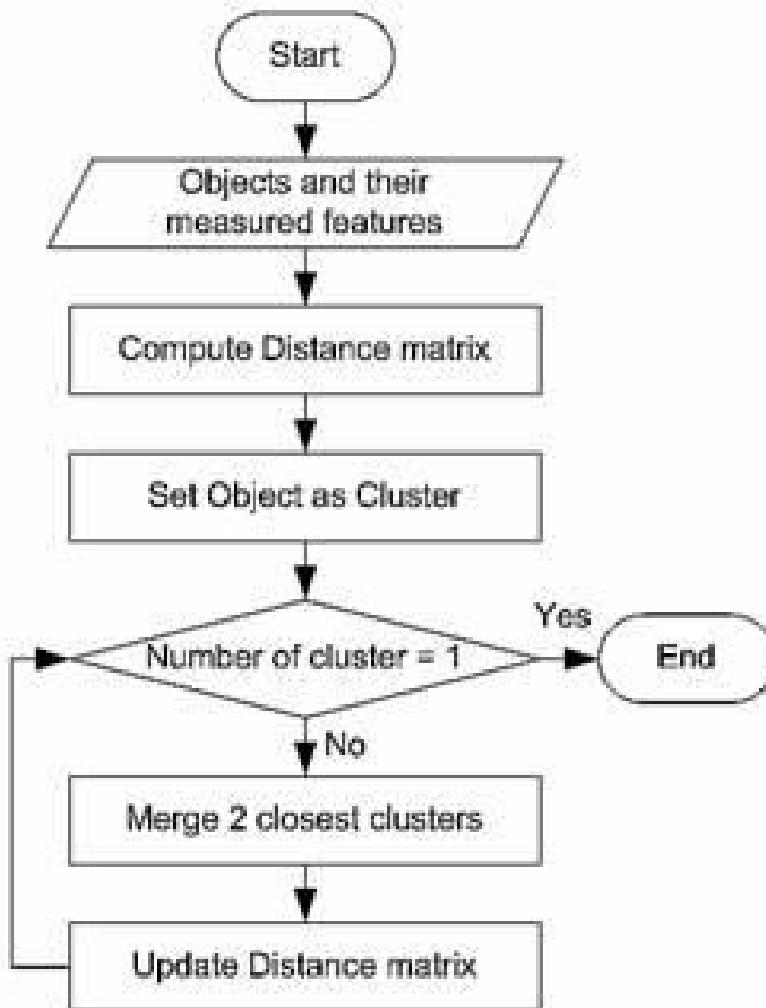


# DIANA (Divisive Analysis)

## Orden inverso de AGNES

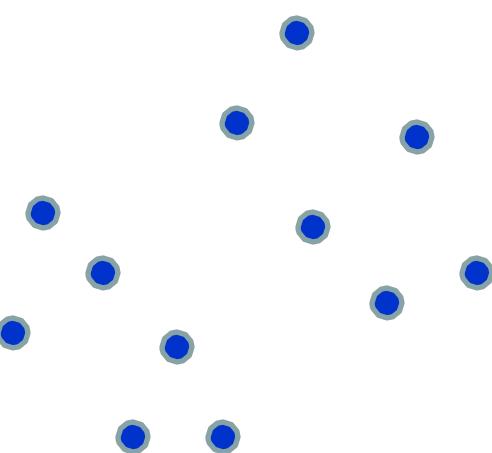


# Métodos jerárquicos aglomerativos

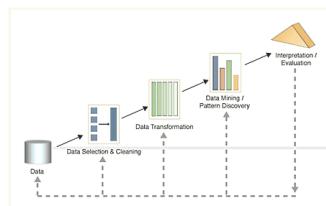


# Métodos jerárquicos

- Representa los puntos en una matriz de distancia o de similitud

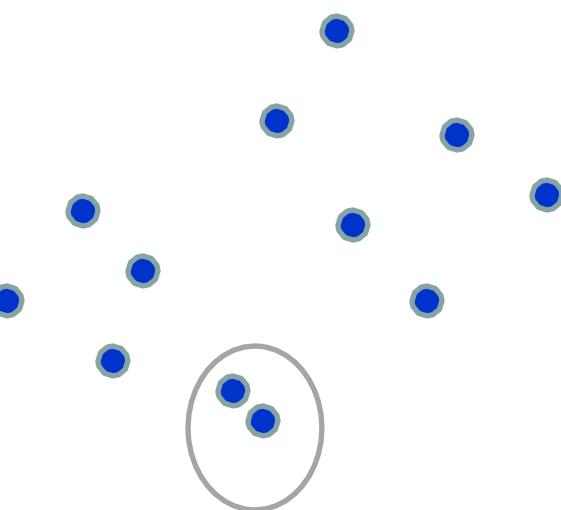


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
:						

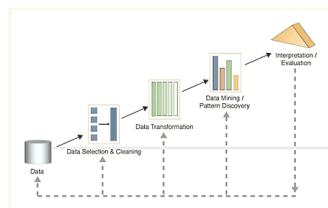


# Métodos jerárquicos

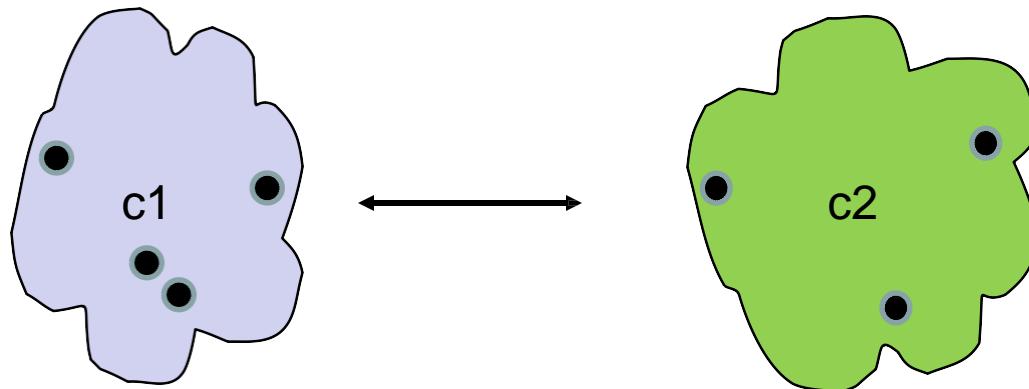
- Une los dos cluster más cercanos y actualiza la matriz



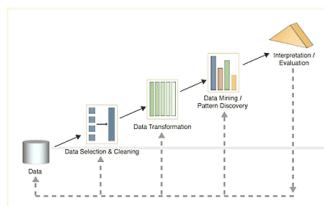
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
:						
.						



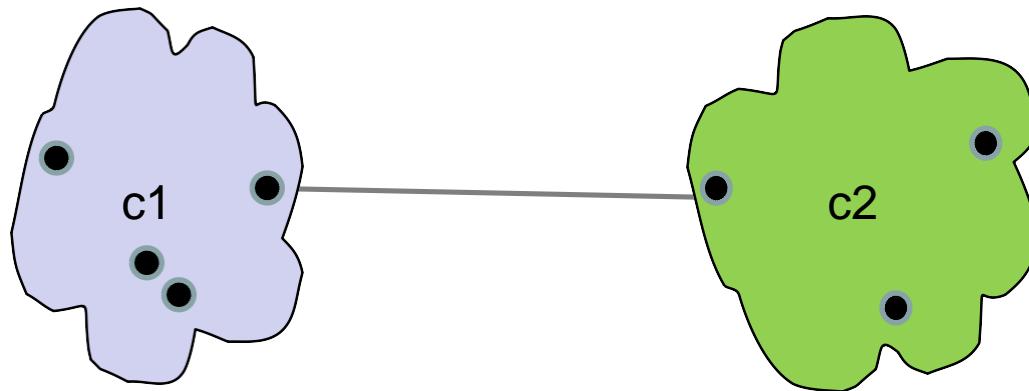
# Cómo definir la similitud entre clusters?



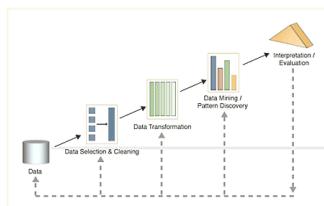
- | MIN
- | MAX
- | Promedio del grupo
- | Distancia entre centroides



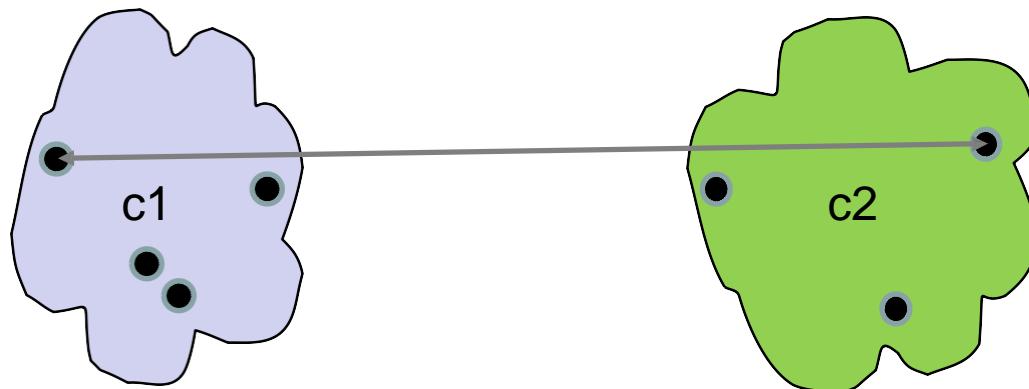
# Cómo definir la similitud entre clusters?



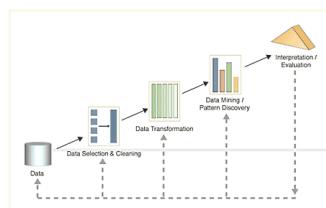
- | MIN
- | MAX
- | Promedio del grupo
- | Distancia entre centroides



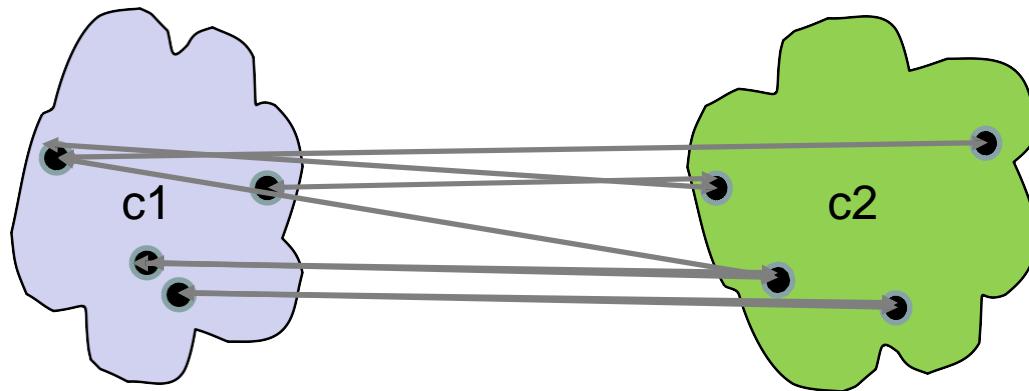
# Cómo definir la similitud entre clusters?



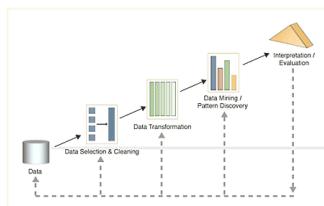
- | MIN
- | MAX
- | Promedio del grupo
- | Distancia entre centroides



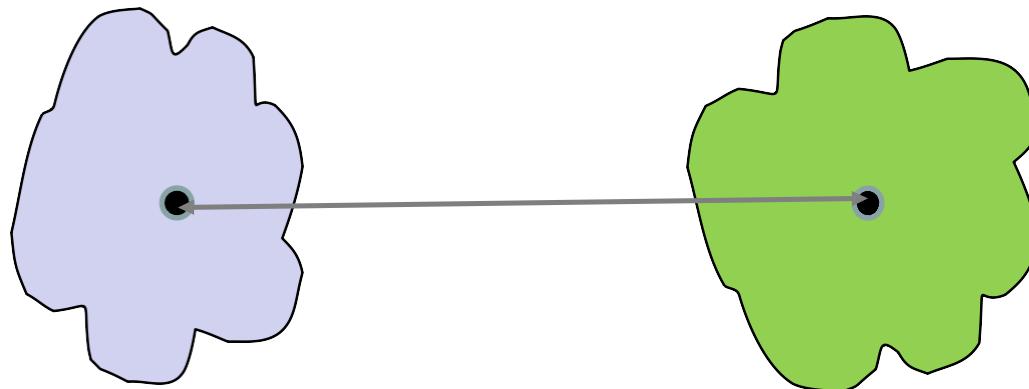
# Cómo definir la similitud entre clusters?



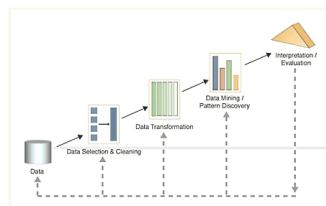
- | MIN
- | MAX
- | Promedio del grupo
- | Distancia entre centroides



# Cómo definir la similitud entre clusters?

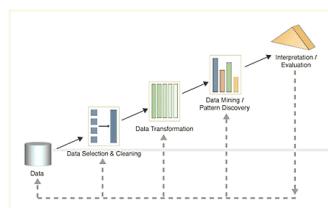
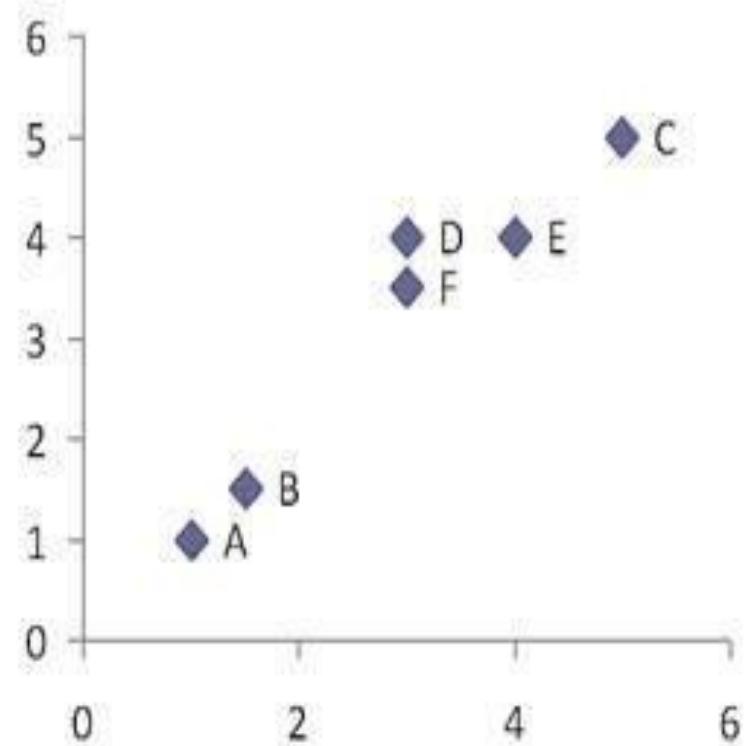


- | MIN
- | MAX
- | Promedio del grupo
- | Distancia entre centroides



# Métodos jerárquicos: Ejemplo

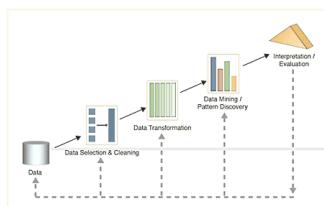
	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



# Métodos jerárquicos: Ejemplo

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

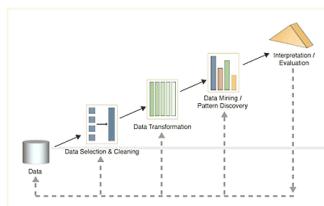
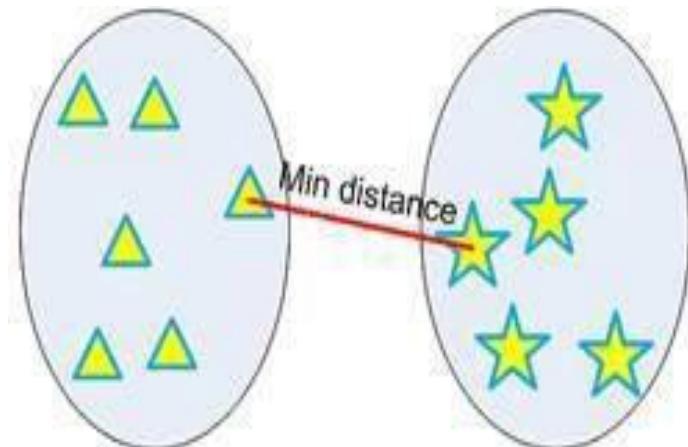
Matriz de distancias



# Métodos jerárquicos: Ejemplo

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

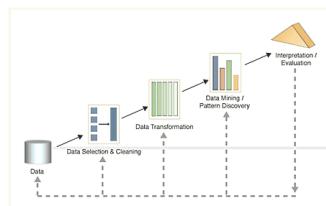
*Linkage hierarchical clustering (MIN)*



# Métodos jerárquicos: Ejemplo

1. Cada objeto se considera un cluster. En este caso, 6 cluster

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00



# Métodos jerárquicos: Ejemplo

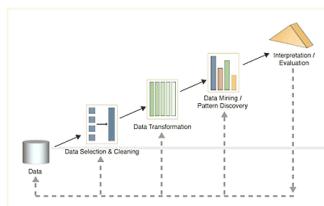
2. En cada paso de la iteración se buscan los dos cluster más cercanos usando la matriz de distancias

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Los cluster *D* y *F* son los más cercanos

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00



# Métodos jerárquicos: Ejemplo

3. Calcular las distancias entre el nuevo cluster (D F) y los demás cluster.

Min Distance (Single Linkage)

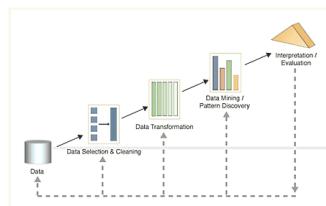
Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

$$d_{(D,F) \rightarrow A} = \min(d_{DA}, d_{FA}) = \min(3.61, 3.20) = 3.20$$

$$d_{(D,F) \rightarrow B} = \min(d_{DB}, d_{FB}) = \min(2.92, 2.50) = 2.50$$

$$d_{(D,F) \rightarrow C} = \min(d_{DC}, d_{FC}) = \min(2.24, 2.50) = 2.24$$

$$d_{E \rightarrow (D,F)} = \min(d_{ED}, d_{EF}) = \min(1.00, 1.12) = 1.00$$



# Métodos jerárquicos: Ejemplo

- Calcular las distancias entre el nuevo cluster (D F) y los demás cluster.

Min Distance (Single Linkage)

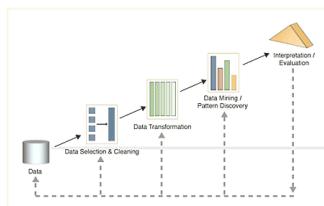
Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00

$$d_{(D,F) \rightarrow A} = \min(d_{DA}, d_{FA}) = \min(3.61, 3.20) = 3.20$$

$$d_{(D,F) \rightarrow B} = \min(d_{DB}, d_{FB}) = \min(2.92, 2.50) = 2.50$$

$$d_{(D,F) \rightarrow C} = \min(d_{DC}, d_{FC}) = \min(2.24, 2.50) = 2.24$$

$$d_{E \rightarrow (D,F)} = \min(d_{ED}, d_{EF}) = \min(1.00, 1.12) = 1.00$$



# Métodos jerárquicos: Ejemplo

- Similar al paso 2, se buscan los dos cluster más cercanos.

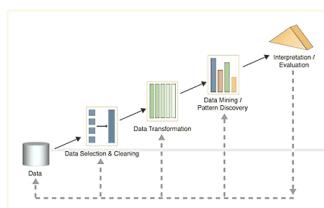
Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00



Ahora los cluster más cercanos son A y B

Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0



# Métodos jerárquicos: Ejemplo

- Calcular las distancias del nuevo cluster con los demás

Min Distance (Single Linkage)

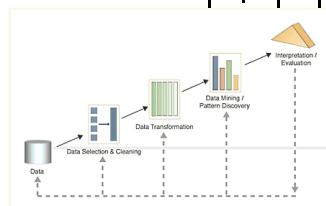
Dist	A,B	C	(D,F)	E
A,B	0	4.95	2.50	3.54
C	4.95	0	2.24	1.41
(D,F)	2.50	2.24	0	1.00
E	3.54	1.41	1.00	0

$$d_{C \rightarrow (A,B)} = \min(d_{CA}, d_{CB}) = \min(5.66, 4.95) = 4.95$$

$$d_{E \rightarrow (A,B)} = \min(d_{EA}, d_{EB}) = \min(4.24, 3.54) = 3.54$$

$$d_{E \rightarrow (A,B)} = \min(d_{EA}, d_{EB}) = \min(4.24, 3.54) = 3.54$$

$$d_{(D,F) \rightarrow (A,B)} = \min(d_{DA}, d_{DB}, d_{FA}, d_{FB}) = \min(3.61, 2.92, 3.20, 2.50) = 2.50$$



# Métodos jerárquicos: Ejemplo

5. Continua iterativamente

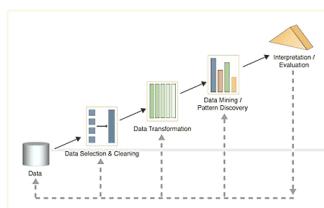
Min Distance (Single Linkage)

Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	4.95
C	4.95	0.00	1.41
(D, F), E	2.50	1.41	0.00



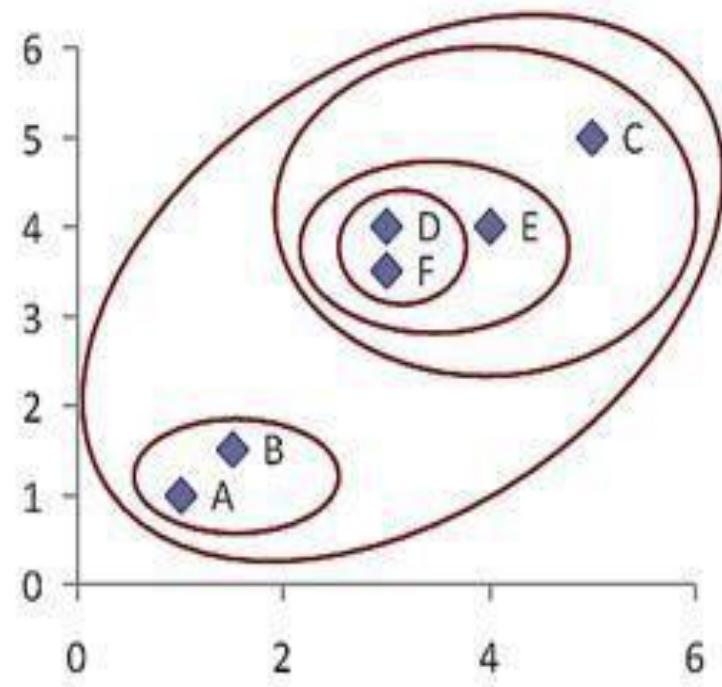
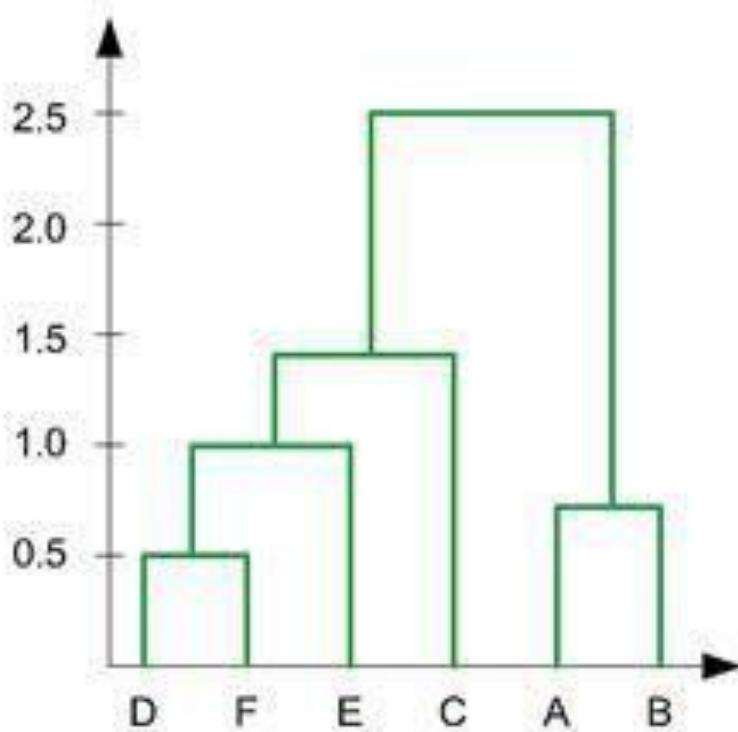
Min Distance (Single Linkage)

Dist	(A,B)	(D, F), E, C
(A,B)	0.00	2.50
((D, F), E), C	2.50	0.00



# Métodos jerárquicos: Ejemplo

## 5. Resultado



# Métodos basados en la densidad

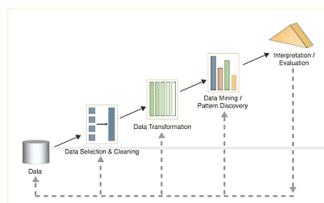
Descubre clusters de formas arbitrarias  
Trata el ruido

Una sola pasada

Necesita parámetros como condición de terminación

DBSCAN: Ester, et al. (KDD'96)

OPTICS: Ankerst, et al (SIGMOD'99). DENCLUE: Hinneburg & D. Keim (KDD'98) CLIQUE: Agrawal, et al. (SIGMOD'98)



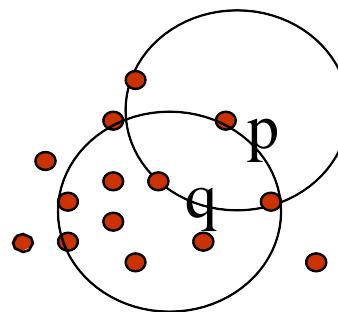
# Parámetros de los métodos basados en densidad

Eps: radio máximo del vecindario

MinPts: número mínimo de puntos en el vecindario de ese punto

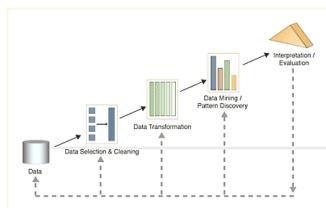
NEps(p):  $\{q \text{ pertenecientes a } D \mid \text{dist}(p,q) \leq Eps\}$  Un punto p es alcanzable directamente desde q con Eps, MinPts si

- 1)p pertenece NEps(q)
- 2) $|NEps (q)| \geq MinPts$



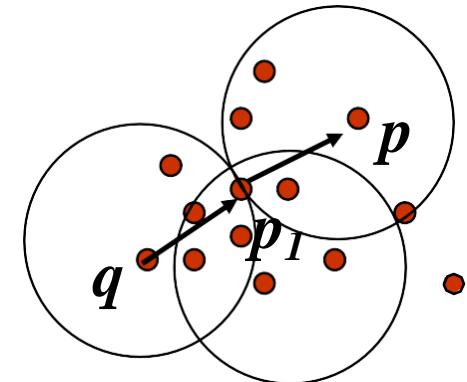
MinPts = 5

Eps = 1 cm

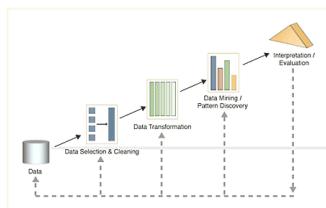
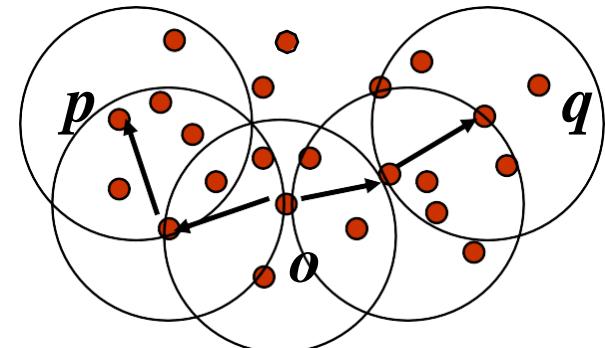


# Parámetros de los métodos basados en densidad

Un pto  $p$  es Densidad-alcanzable desde un pto  $q$  si hay una cadena de puntos  $p_1, \dots, p_n$ ,  $p_1 = q$ ,  $p_n = p$  tales que  $p_{i+1}$  es directamente alcanzable desde  $p_i$



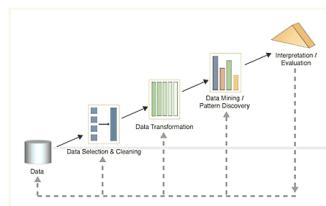
Un pto  $p$  es Densidad-conectado con  $q$  si hay un pto  $o$  tal que,  $p$  y  $q$  son densidad alcanzables desde  $o$



# Resumen

El análisis de cluster permite agrupar objetos basándose en su similitud y tiene muchas aplicaciones

La medida de similitud se calcula dependiendo del tipo de datos



# Bibliografía

Data mining concepts and techniques

*Jiawei Han and Micheline Kamber*

***The Morgan Kaufmann Series in Data Management Systems***

*Introduction to Data Mining.*

**Pang-Ning Tan, Michigan State University,  
Michael Steinbach, University of Minnesota  
Vipin Kumar, University of Minnesota**

