

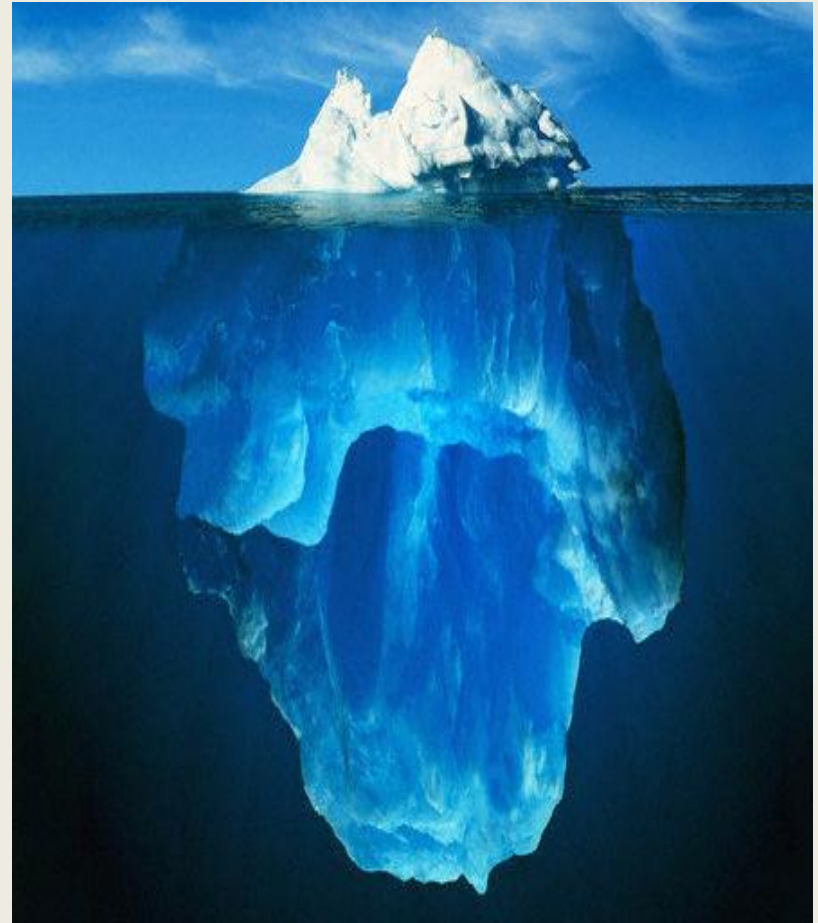
MINERIA DE DATOS Y Descubrimiento del Conocimiento

UNA APLICACIÓN EN DATOS AGROPECUARIOS
INTA EEA Corrientes

Maximiliano Silva

La información

- Herramienta estratégica para el desarrollo de:
 - Sociedad de la información.
 - Economía cuya base es el conocimiento.



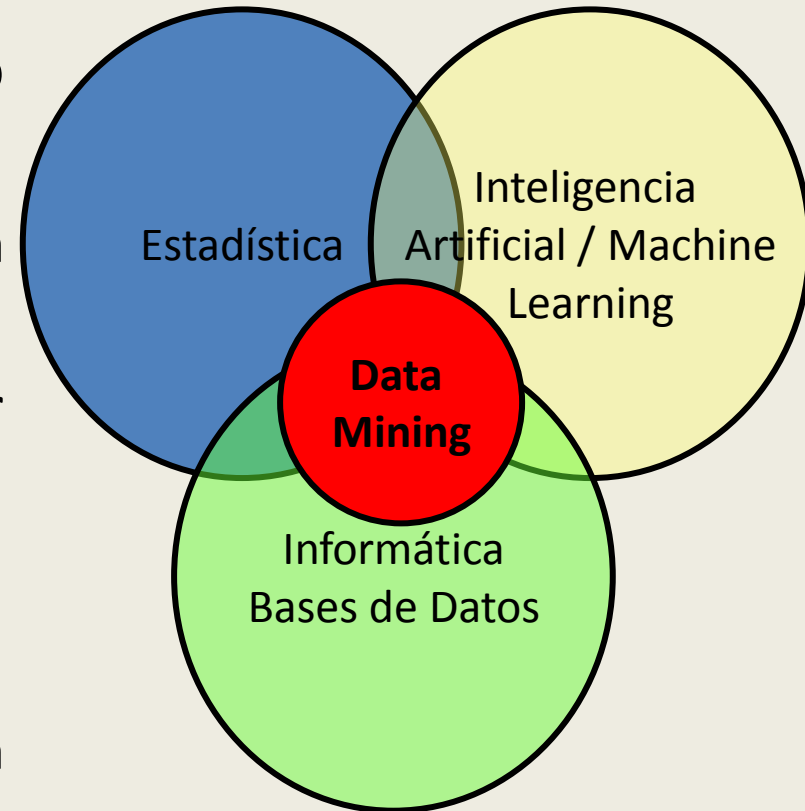
Problemática

- Incremento en dos sentidos en Bases de Datos:
 - Número N de registros u objetos.
 - Número D de campos u atributos por objeto.
- Crecimiento BD (tamaño y número)
 - Supera a las habilidades humanas para analizar.
 - Necesidad y oportunidad de extraer conocimiento.



Descubrimiento de Conocimiento en Bases de Datos (KDD)

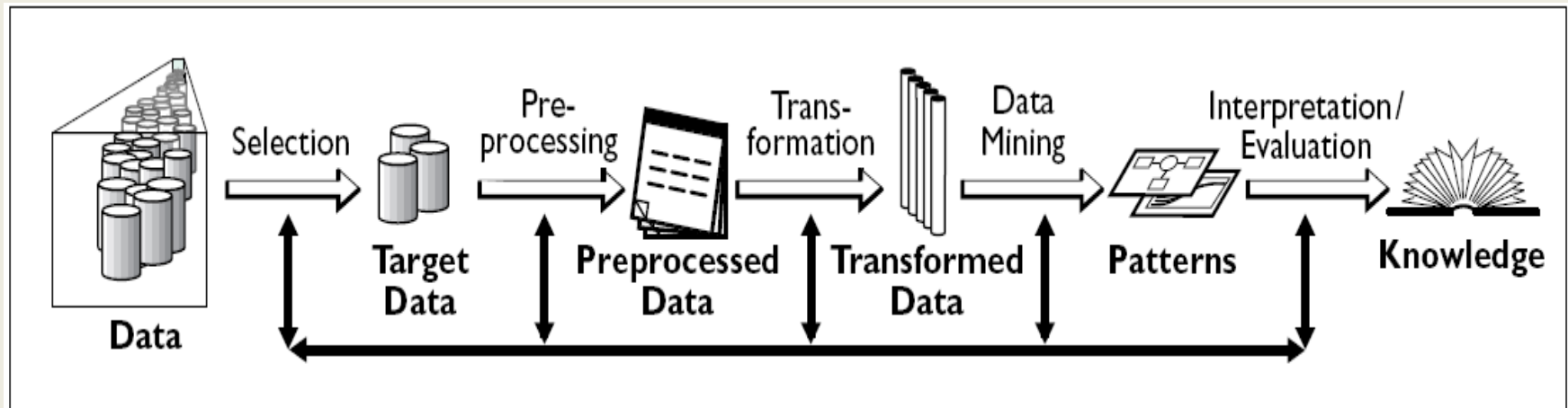
- Descubrimiento del conocimiento en Base de Datos :
 - KDD: Knowledge Discovery in Database. 1989.
- El método tradicional de convertir datos en conocimiento:
 - análisis e interpretación manual.
 - lento, costoso y altamente subjetivo .
 - volúmenes de datos crecen exponencialmente.



KDD

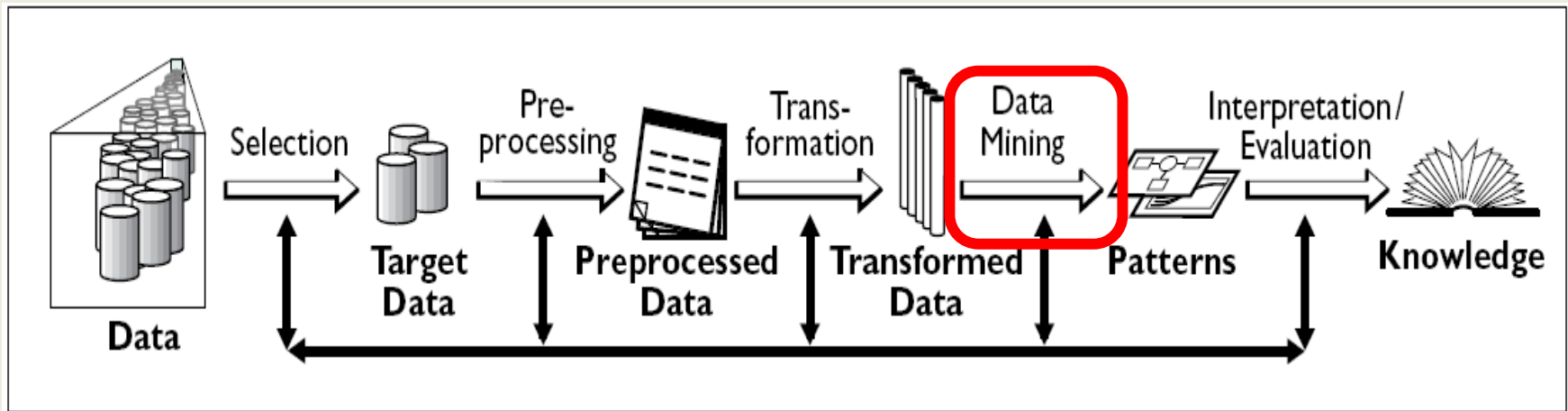
- *“El proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en ultima instancia comprensible en los datos”*

Usama Fayyad 1996



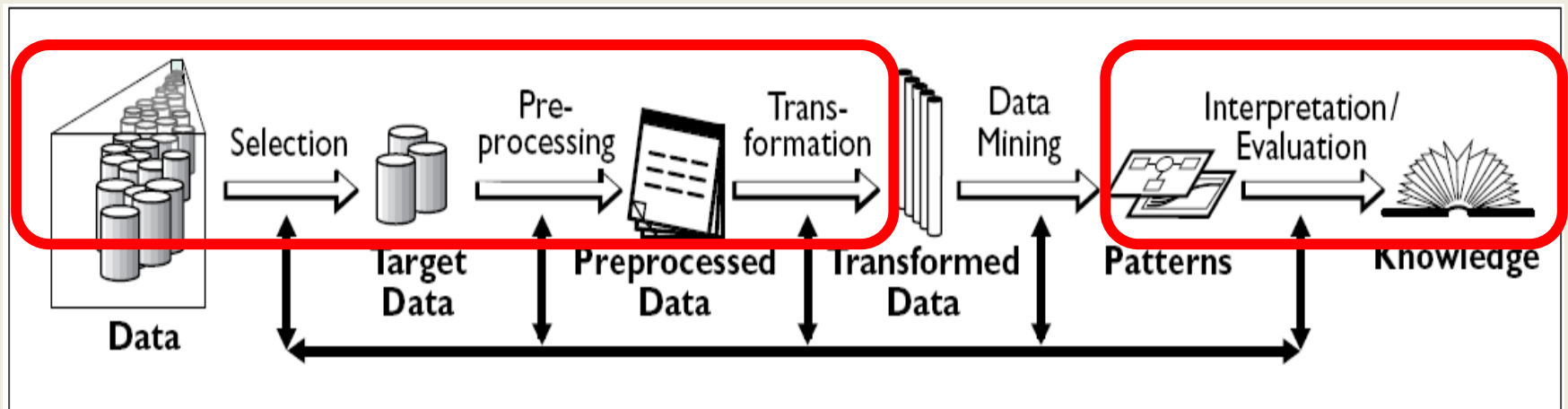
Minería de Datos (DM)

- Minería de Datos: Data Mining – DM.
- Es la aplicación de algoritmos específicos para extraer patrones desde los datos

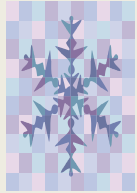


KDD: Pasos adicionales

- Selección
- Limpieza .
- Reducción.
- Interpretación.
- Uso del conocimiento.

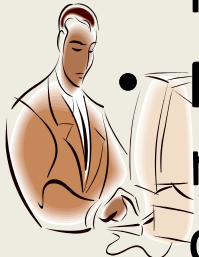


Aplicaciones del Data Mining



Aspectos climatológicos: predicción de tormentas, etc.

- **Medicina:** encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico.



- **Mercadotécnia:** identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de productos, etc.

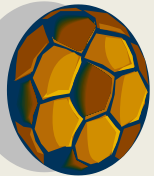
- **Inversión en casas de bolsa y banca:** análisis de clientes, aprobación de préstamos, determinación de montos de crédito.





Detección de fraudes y comportamientos inusuales: telefónicos, seguros, en tarjetas de crédito, evasión fiscal, electricidad, etc.

- **Análisis de canastas de mercado para mejorar la organización de tiendas, segmentación de mercado (clustering).**



Deporte profesional: determinar puntos, expulsiones/tarjetas que tiene cada jugador, tomar mejores decisiones para siguientes temporadas.

Algoritmos de Minería de Datos

- Supervisados o predictivos:
 - Dado un conjunto de variables predictoras, se desea conocer el comportamiento de la variable a predecir. Predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos.
 - Una vez entrenado el modelo, sirve para realizar la predicción de datos cuyo valor es desconocida.

Variables Predictoras

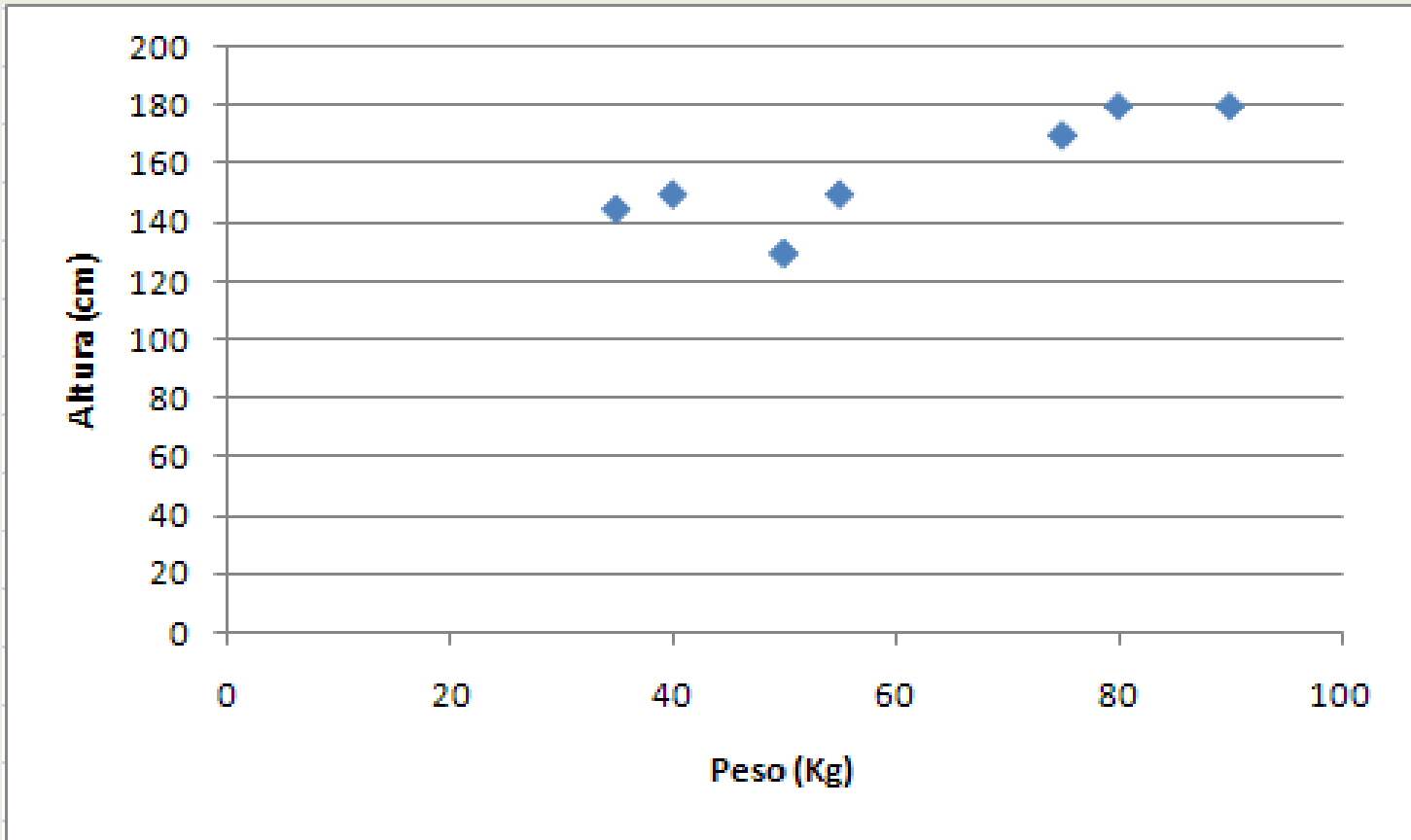
cielo	Temperatura	Humedad	Viento	JuegaTennis
soleado	calor	alta	débil	no
soleado	calor	alta	fuerte	no
nublado	calor	alta	débil	si
lluvioso	cálido	alta	débil	si
lluvioso	fresco	normal	débil	si
lluvioso	fresco	normal	fuerte	no
nublado	fresco	normal	fuerte	si
soleado	cálido	alta	débil	no
soleado	fresco	normal	débil	si
lluvioso	cálido	normal	débil	si
soleado	cálido	normal	fuerte	si
nublado	cálido	alta	fuerte	si
nublado	calor	normal	débil	si
lluvioso	calor	alta	fuerte	no
soleado	calor	normal	fuerte	✘
lluvioso	fresco	normal	fuerte	✘

Variable a predecir

Algoritmos de Minería de Datos

- No supervisados:
 - Descubren patrones y tendencias en los datos, que no poseen variable a predecir.
 - Agrupar registros por similitud.
 - Descubrimiento de conocimiento: tomar acciones y obtener un beneficio (científico o de negocio) de ellas.

Peso	Altura
80	180
50	130
55	150
90	180
75	170
35	145
40	150



Árbol de Decisión

- Herramienta potentísima de clasificación. Construyen un árbol del que se pueden extraer reglas.
- Validaciones. Detectar elementos anómalos en función de si encajan o no con las reglas surgidas del árbol.
- Predecir el valor de un atributo con precisión, encontrando correlaciones entre las variables predictoras y la variable a predecir.

Ejemplo:

	A	B	C	D	E
1	cielo	Temperatura	Humedad	Viento	JuegaTennis
2	soleado	calor	alta	débil	no
3	soleado	calor	alta	fuerte	no
4	nublado	calor	alta	débil	si
5	lluvioso	cálido	alta	débil	si
6	lluvioso	fresco	normal	débil	si
7	lluvioso	fresco	normal	fuerte	no
8	nublado	fresco	normal	fuerte	si
9	soleado	cálido	alta	débil	no
10	soleado	fresco	normal	débil	si
11	lluvioso	cálido	normal	débil	si
12	soleado	cálido	normal	fuerte	si
13	nublado	cálido	alta	fuerte	si
14	nublado	calor	normal	débil	si
15	lluvioso	calor	alta	fuerte	no
16	soleado	calor	normal	fuerte	
17	lluvioso	fresco	normal	fuerte	

Stream2* - Clementine 12.0

File Edit Insert View Tools SuperNode Window Help

```
graph LR; Libro1[Libro1.xls] --> Type[Type]; Type --> JuegaTennis1[JuegaTennis]; JuegaTennis1 --> JuegaTennis2[JuegaTennis]; JuegaTennis2 --> Table[Table];
```

Streams Outputs Models

JuegaTennis

CRISP-DM Classes

- (unsaved project)
 - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modeling
 - Evaluation
 - Deployment

Favorites Sources Record Ops Field Ops Graphs Modeling Output Export

Table Custom Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals SPSS Output

Server: Local Server 104MB / 165MB

Table (7 fields, 16 records) #16

	cielo	Temperatura	Humedad	Viento	JuegaTennis	\$C-JuegaTennis	\$CC-JuegaTennis
1	soleado	calor	alta	débil	no	no	0.800
2	soleado	calor	alta	fuerte	no	no	0.800
3	nublado	calor	alta	débil	si	si	0.833
4	lluvioso	cálido	alta	débil	si	si	0.800
5	lluvioso	fresco	normal	débil	si	si	0.800
6	lluvioso	fresco	normal	fuerte	no	no	0.750
7	nublado	fresco	normal	fuerte	si	si	0.833
8	soleado	cálido	alta	débil	no	no	0.800
9	soleado	fresco	normal	débil	si	si	0.750
10	lluvioso	cálido	normal	débil	si	si	0.800
11	soleado	cálido	normal	fuerte	si	si	0.750
12	nublado	cálido	alta	fuerte	si	si	0.833
13	nublado	calor	normal	débil	si	si	0.833
14	lluvioso	calor	alta	fuerte	no	no	0.750
15	soleado	calor	normal	fuerte	\$null\$	si	0.750
16	lluvioso	fresco	normal	fuerte	\$null\$	no	0.750

Table

Annotations

OK



JuegaTennis

Nodo 0		
Categoría	%	n
no	35,714	5
si	64,286	9
Total	100,000	14

cielo

lluvioso

nublado

soleado

Nodo 1		
Categoría	%	n
no	40,000	2
si	60,000	3
Total	35,714	5

Nodo 4		
Categoría	%	n
no	0,000	0
si	100,000	4
Total	28,571	4

Nodo 5		
Categoría	%	n
no	60,000	3
si	40,000	2
Total	35,714	5

Viento

Humedad

fuerte

débil

normal

alta

Nodo 2		
Categoría	%	n
no	100,000	2
si	0,000	0
Total	14,286	2

Nodo 3		
Categoría	%	n
no	0,000	0
si	100,000	3
Total	21,429	3

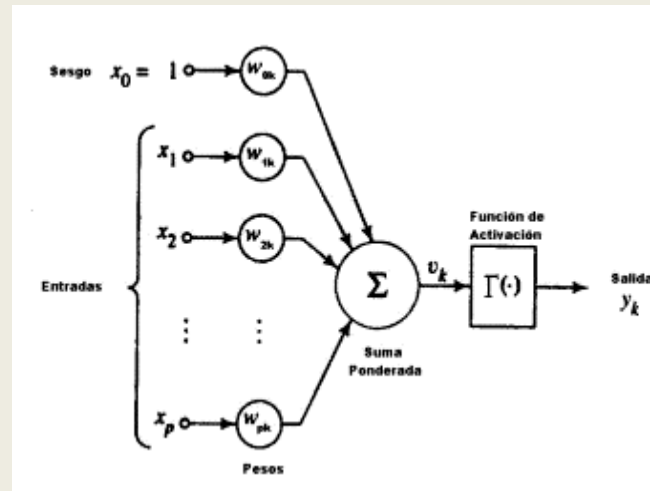
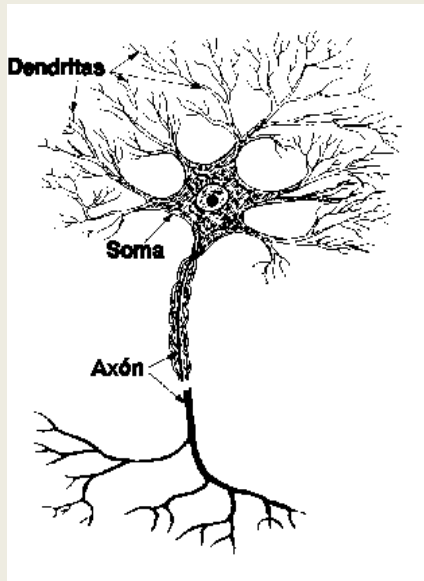
Nodo 6		
Categoría	%	n
no	0,000	0
si	100,000	2
Total	14,286	2

Nodo 7		
Categoría	%	n
no	100,000	3
si	0,000	0
Total	21,429	3

Redes neuronales

- Se basan en la analogía que existe en el comportamiento y función del cerebro humano, en particular del sistema nervioso.
 - Aprende variando el peso sináptico.

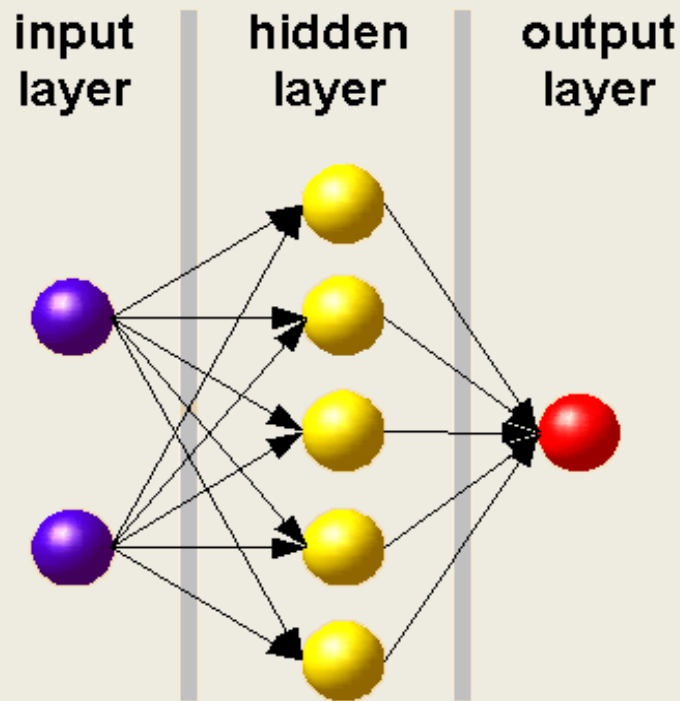
Neurona
Biológica



Modelo
Matemático
de la Neurona
Biológica

Redes neuronales

- Aprenden a través del entrenamiento.
- Objetivo: balance entre
 - Habilidad para responder correctamente en relación a la entrada de patrones es decir usado para el entrenamiento .
 - Habilidad de dar una respuesta (buena) razonable para la entrada que es similar.



- El entrenamiento de la red por *backpropagation* implica tres etapas:
- Feedforward (red de alimentación hacia adelante) del entrenamiento de patrones de entrada.
- *Backpropagation* del error asociado y
- El ajuste de los pesos.

Agrupamiento (Clustering)

- Es una técnica cuya idea básica es agrupar un conjunto de observaciones en un número dado de clusters o grupos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones.
- La idea es que los elementos en un grupo sean similares y en grupos diferentes tengan la menor similitud posible.



Clustering



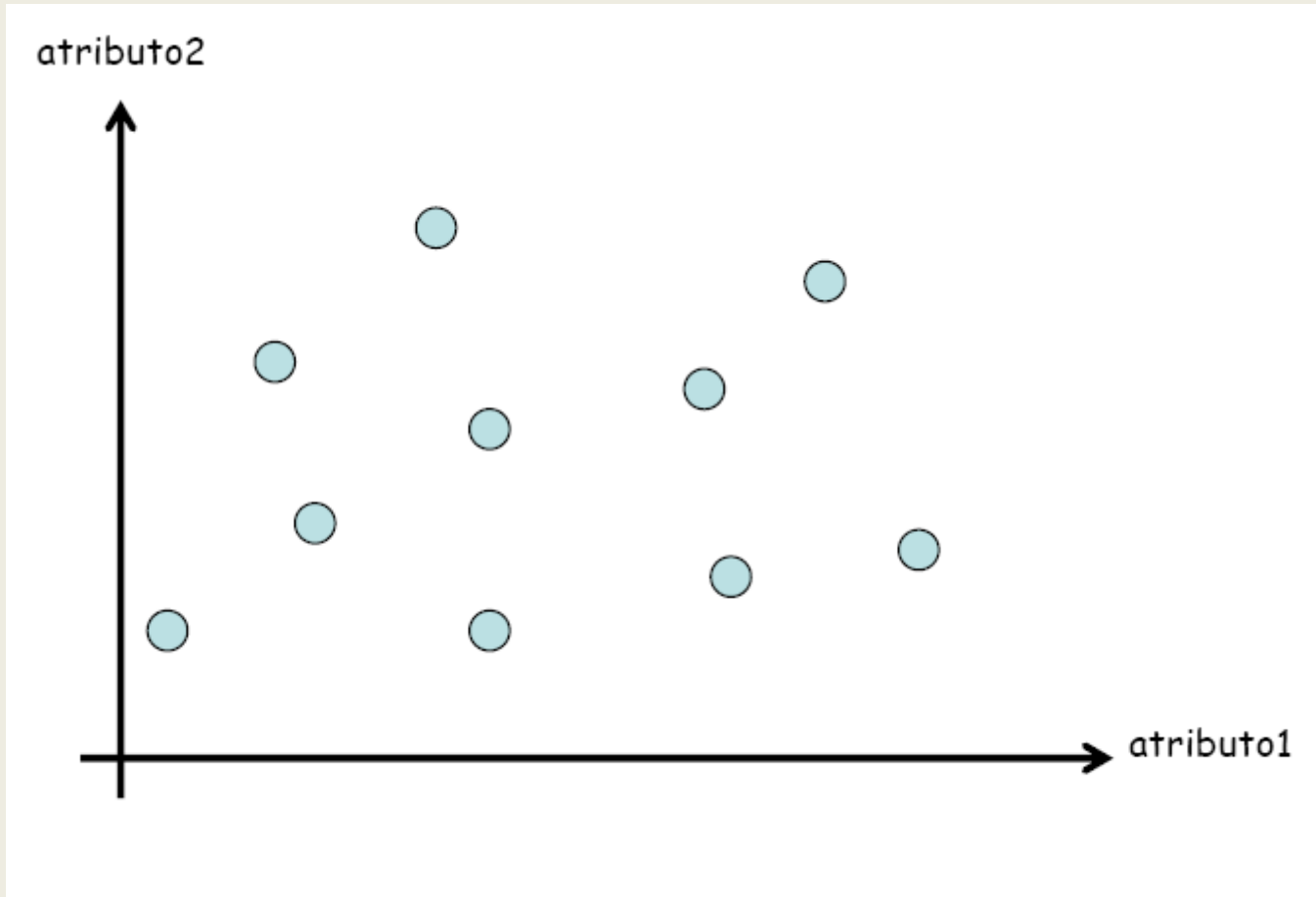
**1000 clientes
en una BD**



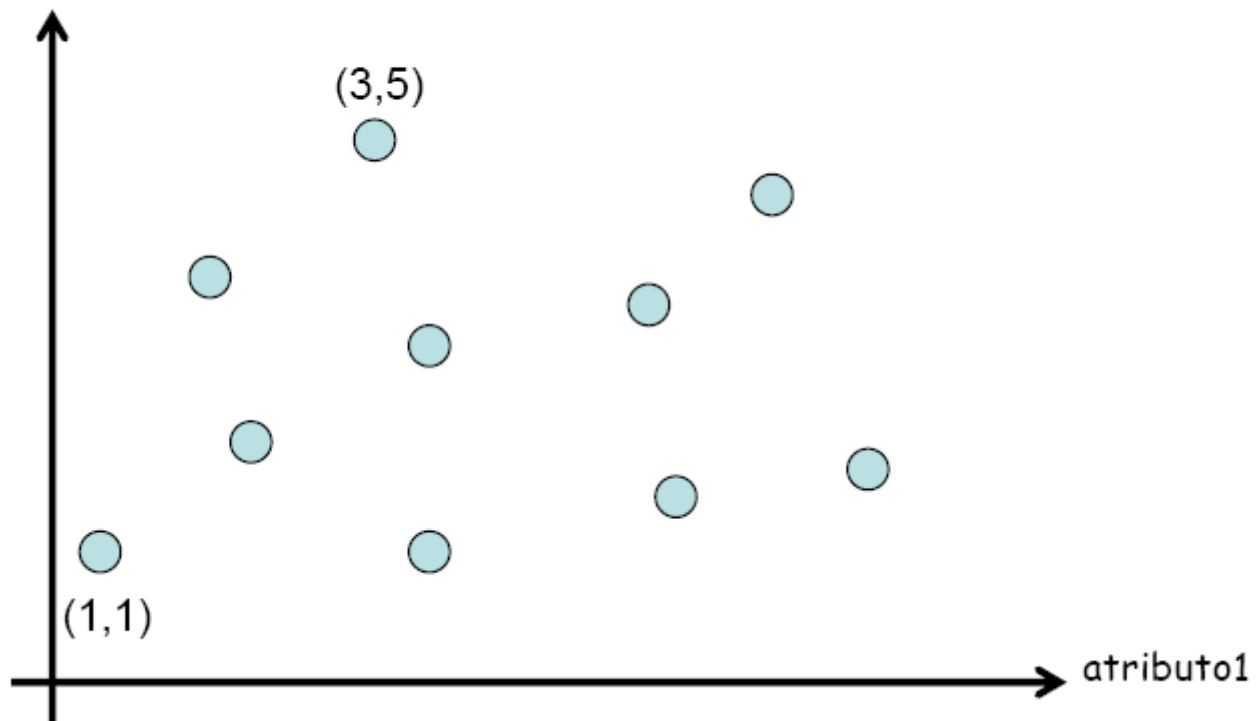
K-medias

- El algoritmo de las K-medias es un algoritmo de partición. Básicamente este algoritmo busca formar clusters (grupos) los cuales serán representados por K objetos (centroides)
- La cantidad de K es un valor ingresado por el usuario.
- Utiliza la noción de centroide.
- Cada uno de estos centroides es el valor medio de los objetos que pertenecen a dicho grupo.
- Es un algoritmo iterativo por naturaleza.

Ejemplo

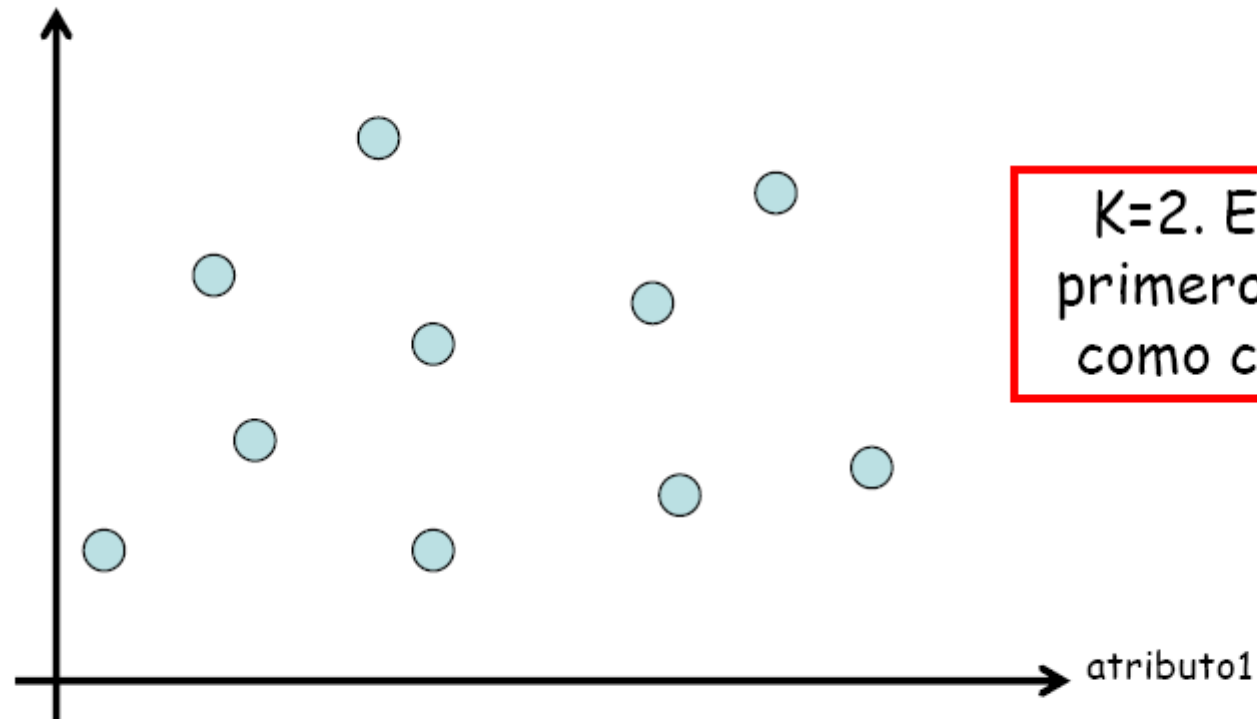


atributo2



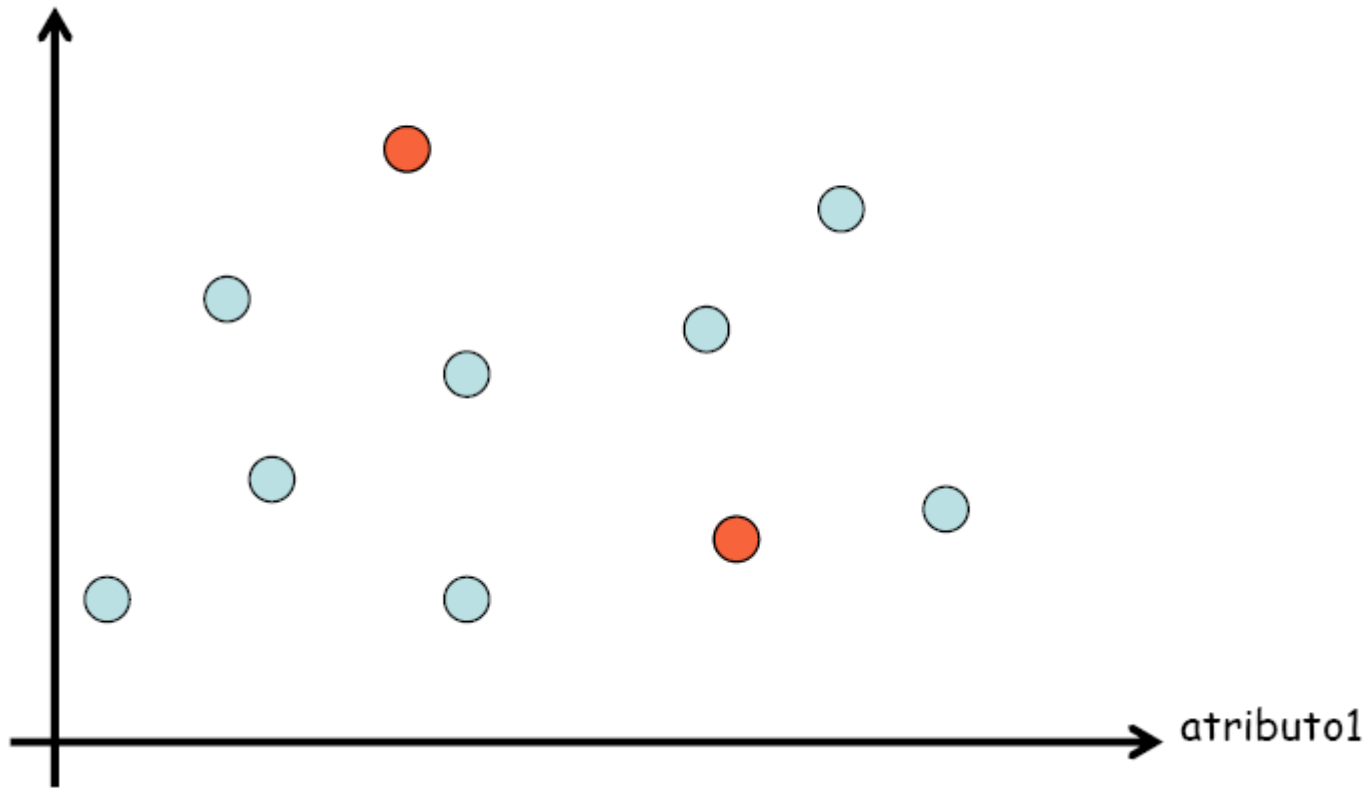
atributo1

atributo2



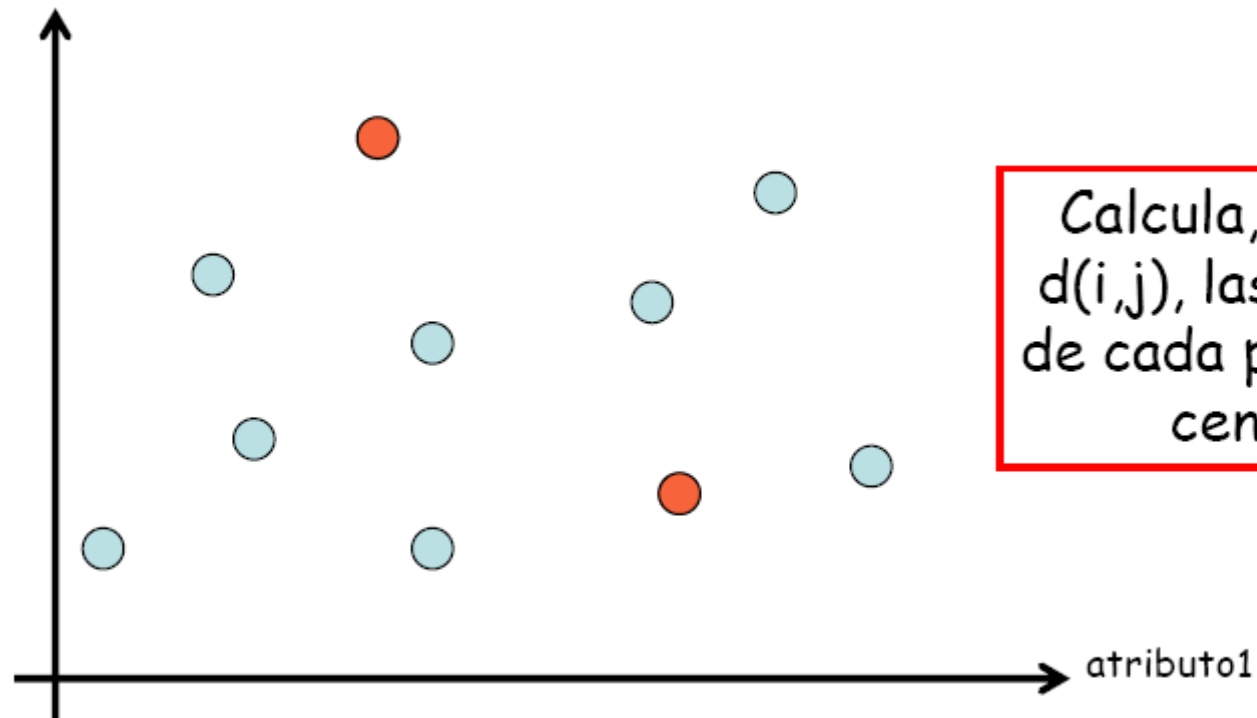
$K=2$. Escoge los primeros k puntos como centroides

atributo2



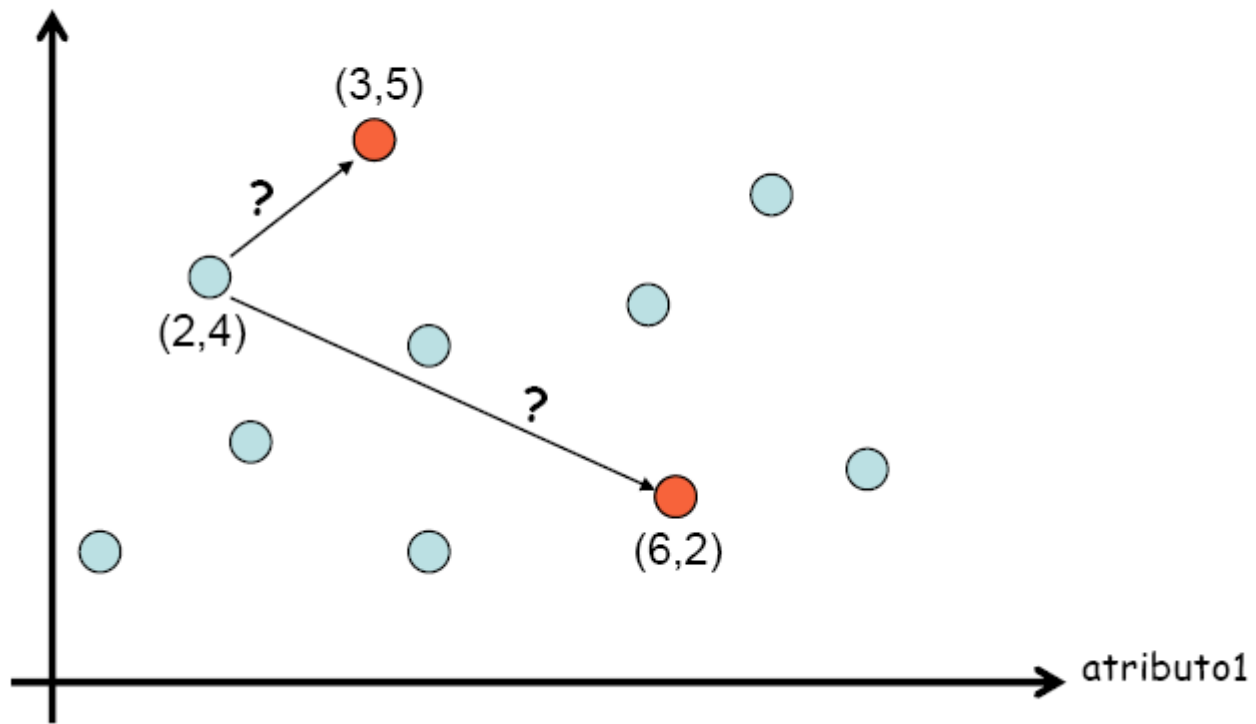
atributo1

atributo2



Calcula, utilizando $d(i,j)$, las distancias de cada punto a cada centroide

atributo2



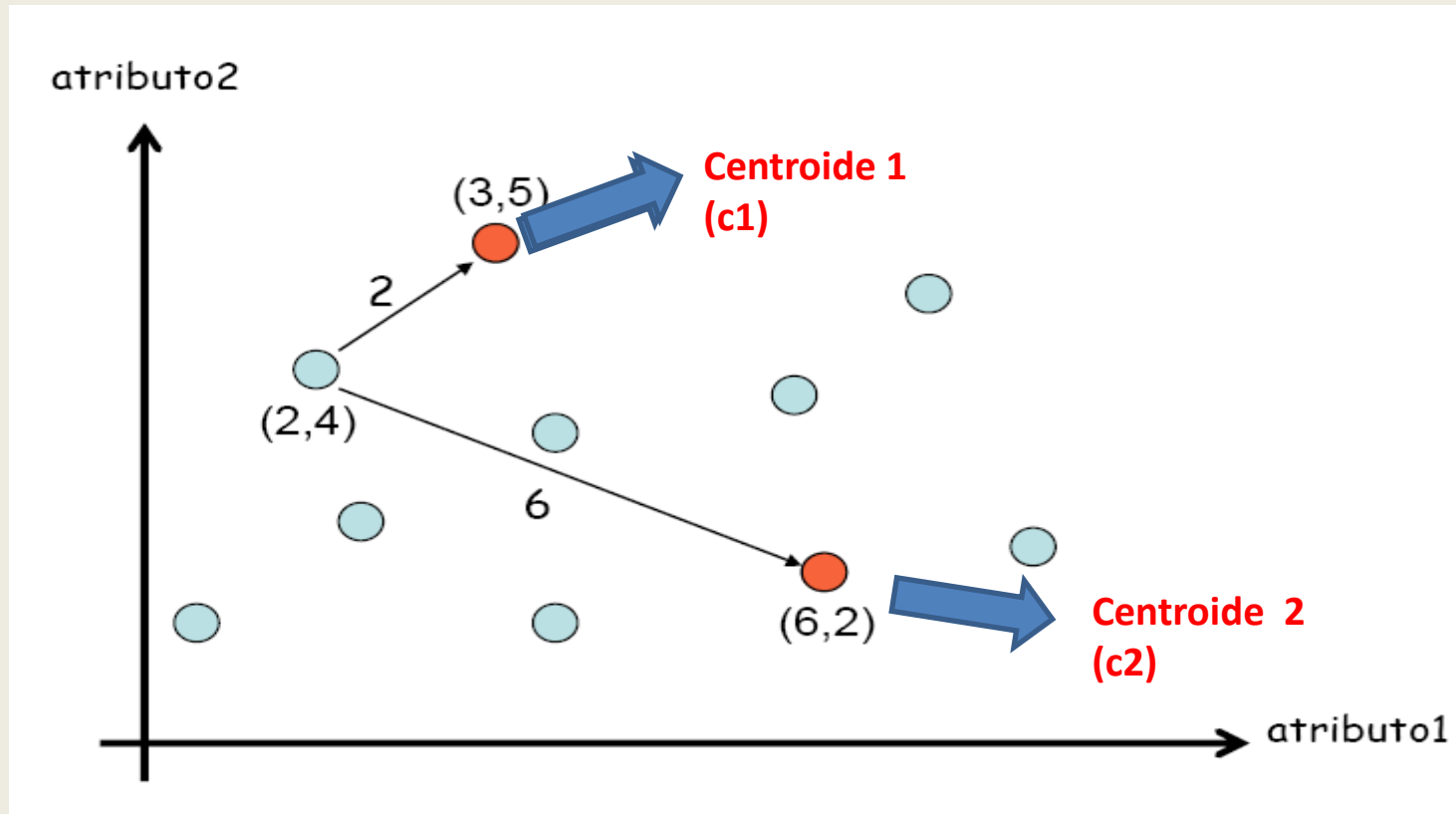
atributo1

Distancia de Manhattan

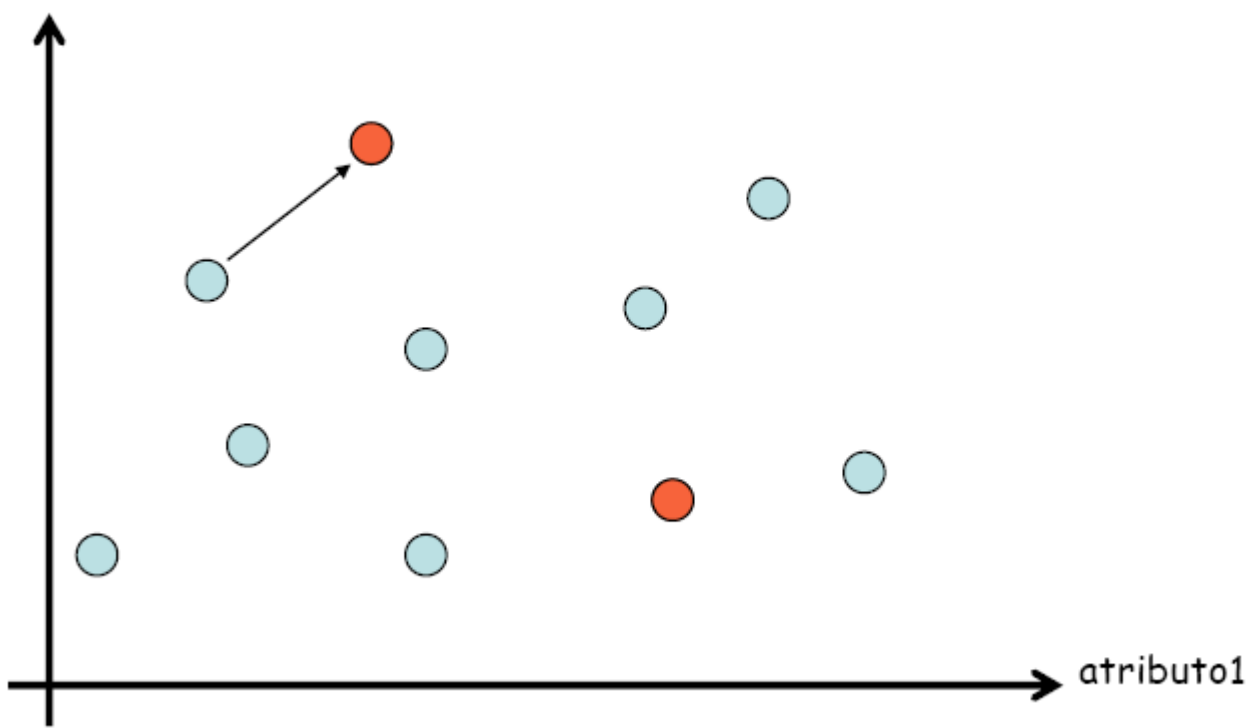
$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

$$D(p1, c1) = |2-3| + |4-5| = 2$$

$$D(p1, c2) = |2-6| + |4-2| = 6$$

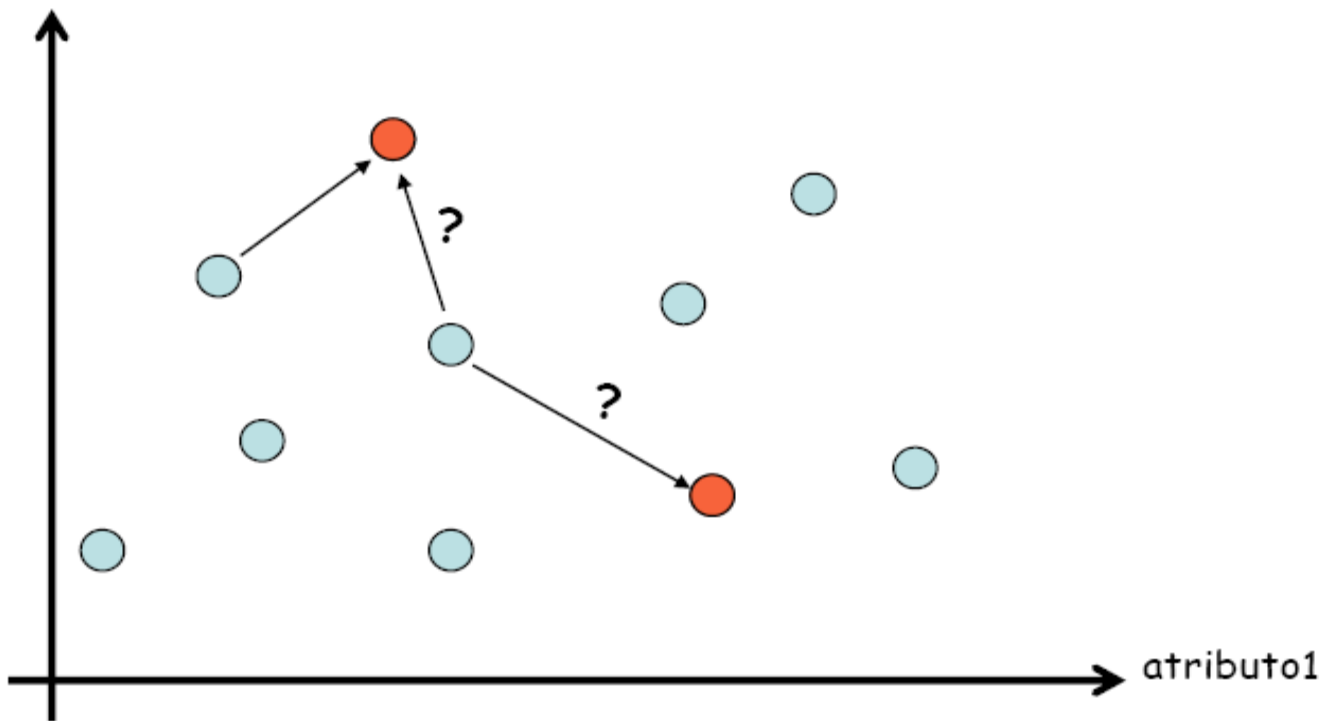


atributo2



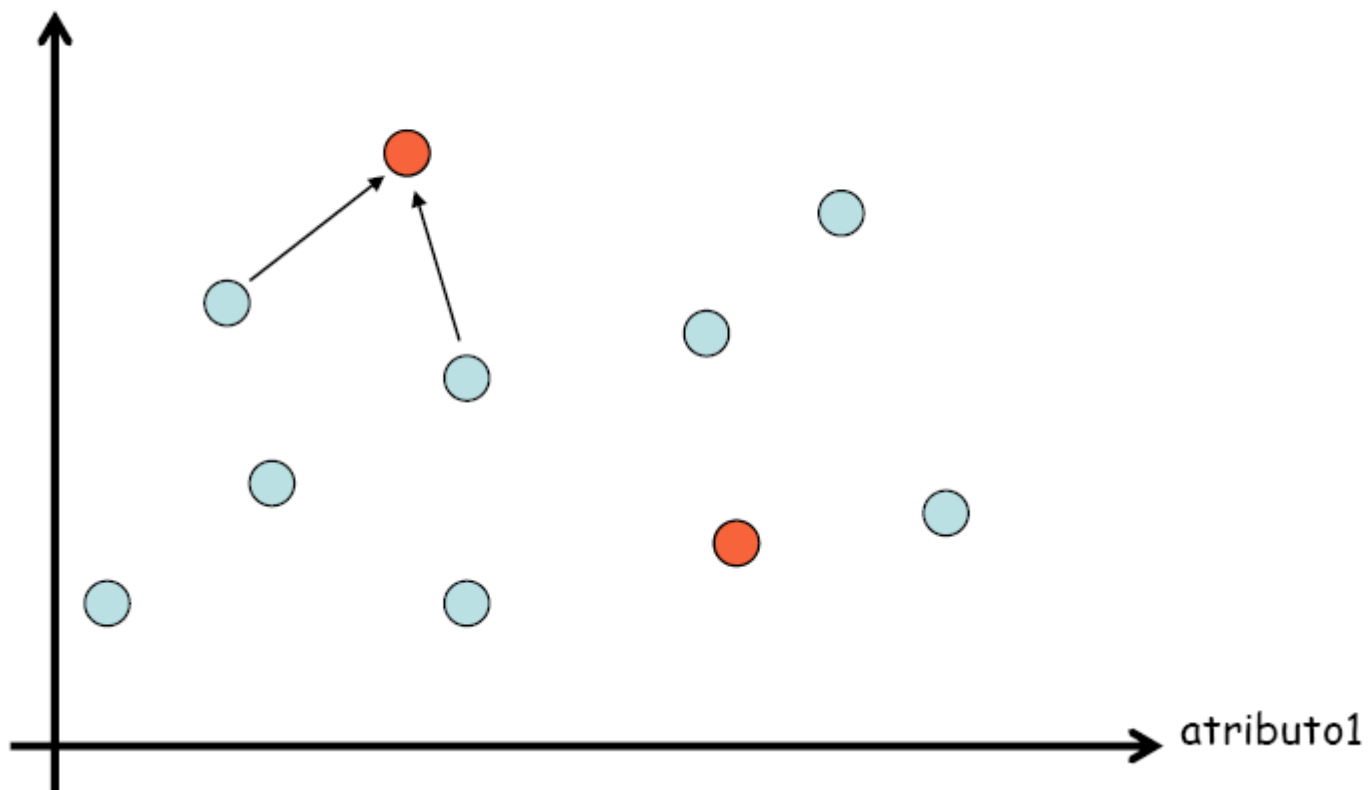
atributo1

atributo2



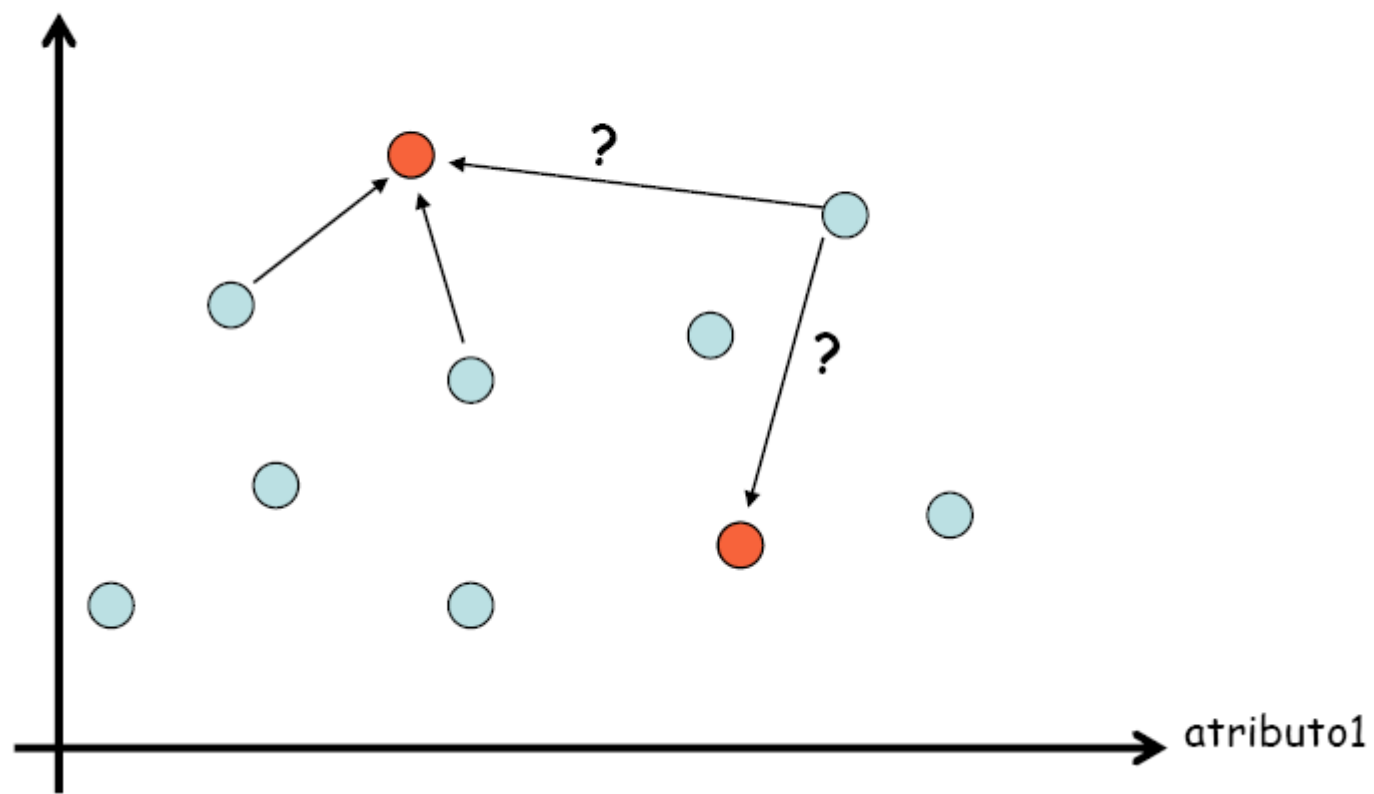
atributo1

atributo2



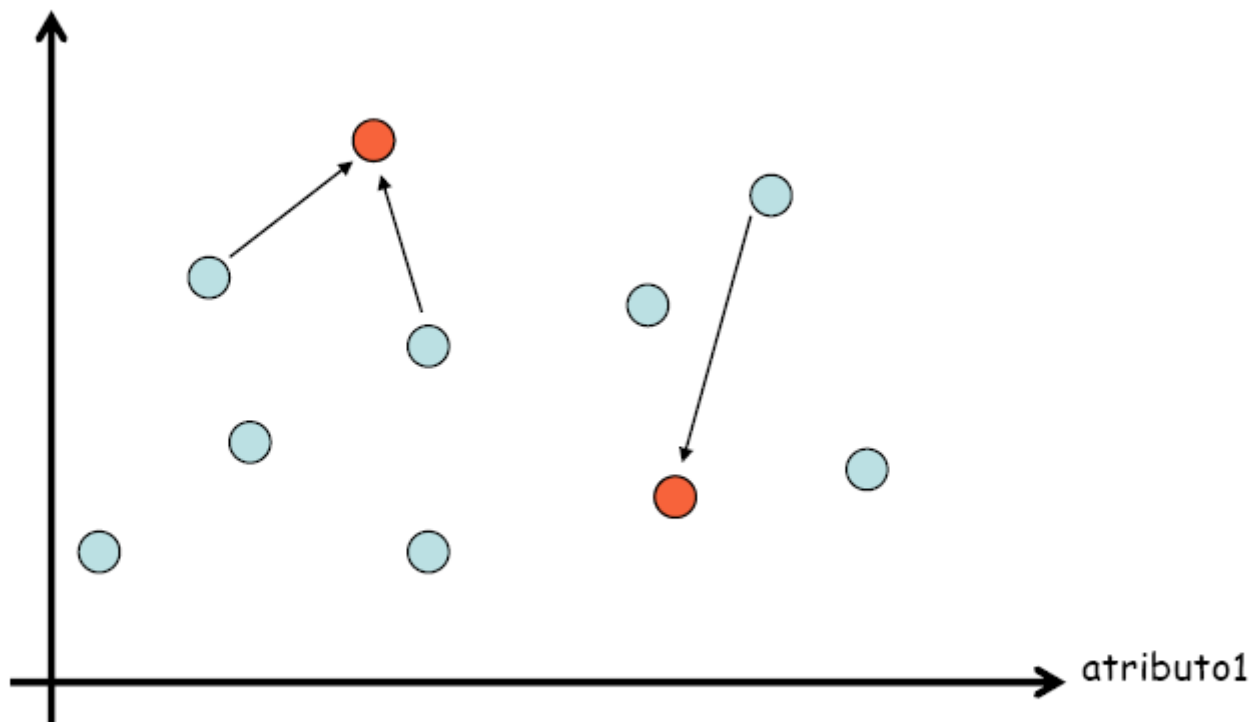
atributo1

atributo2

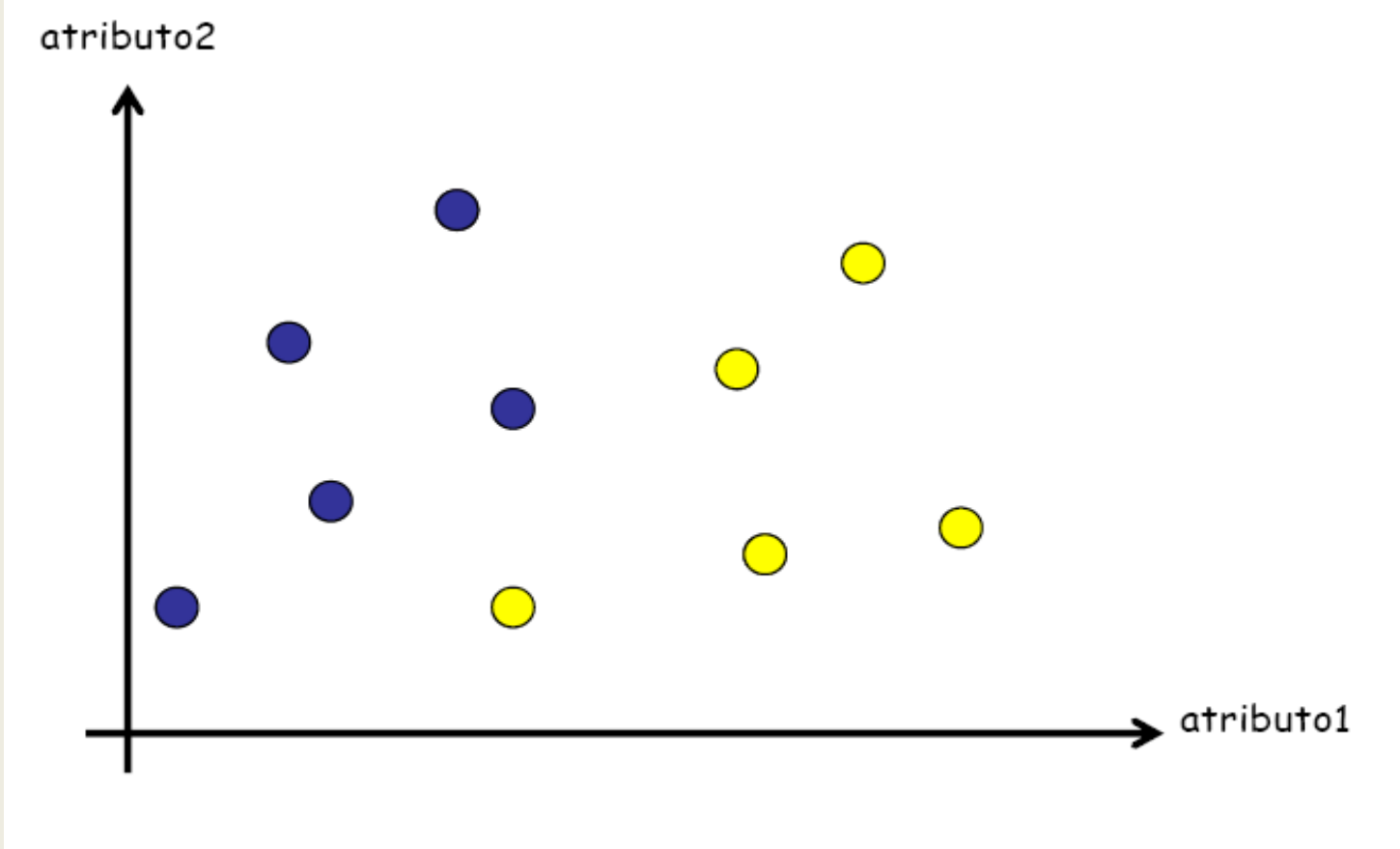


atributo1

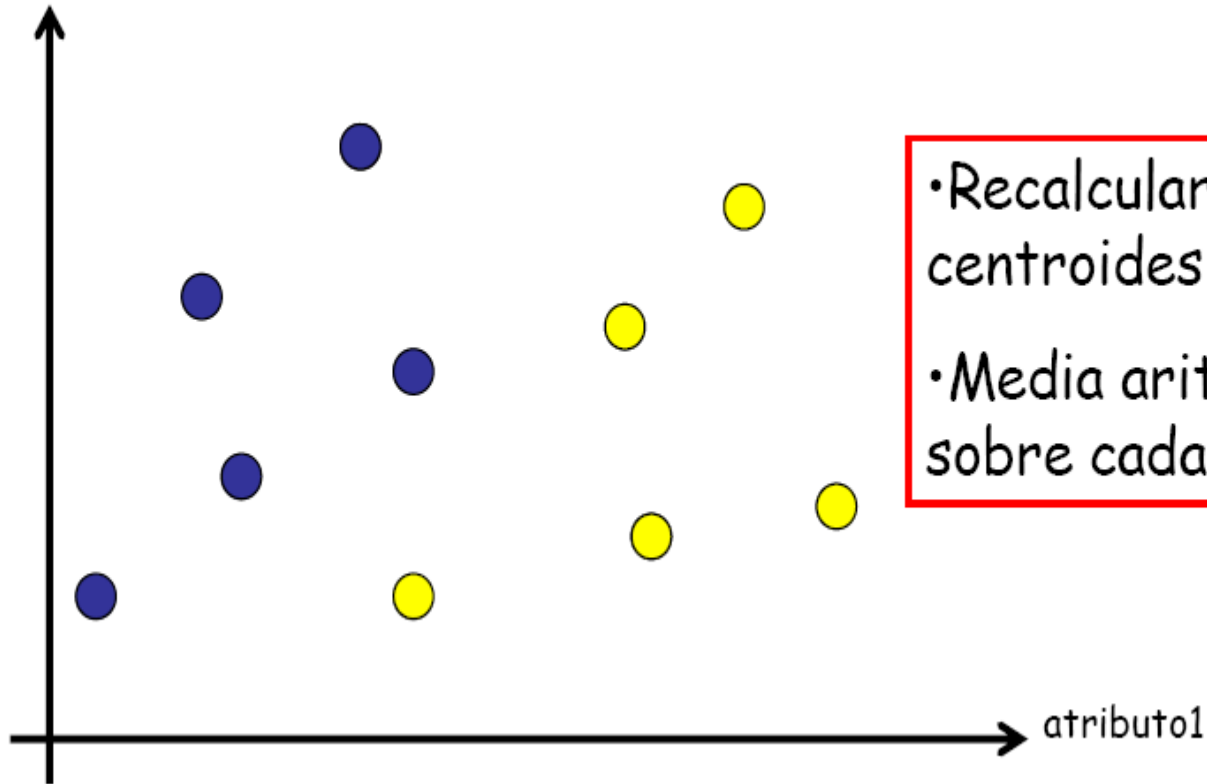
atributo2



atributo1

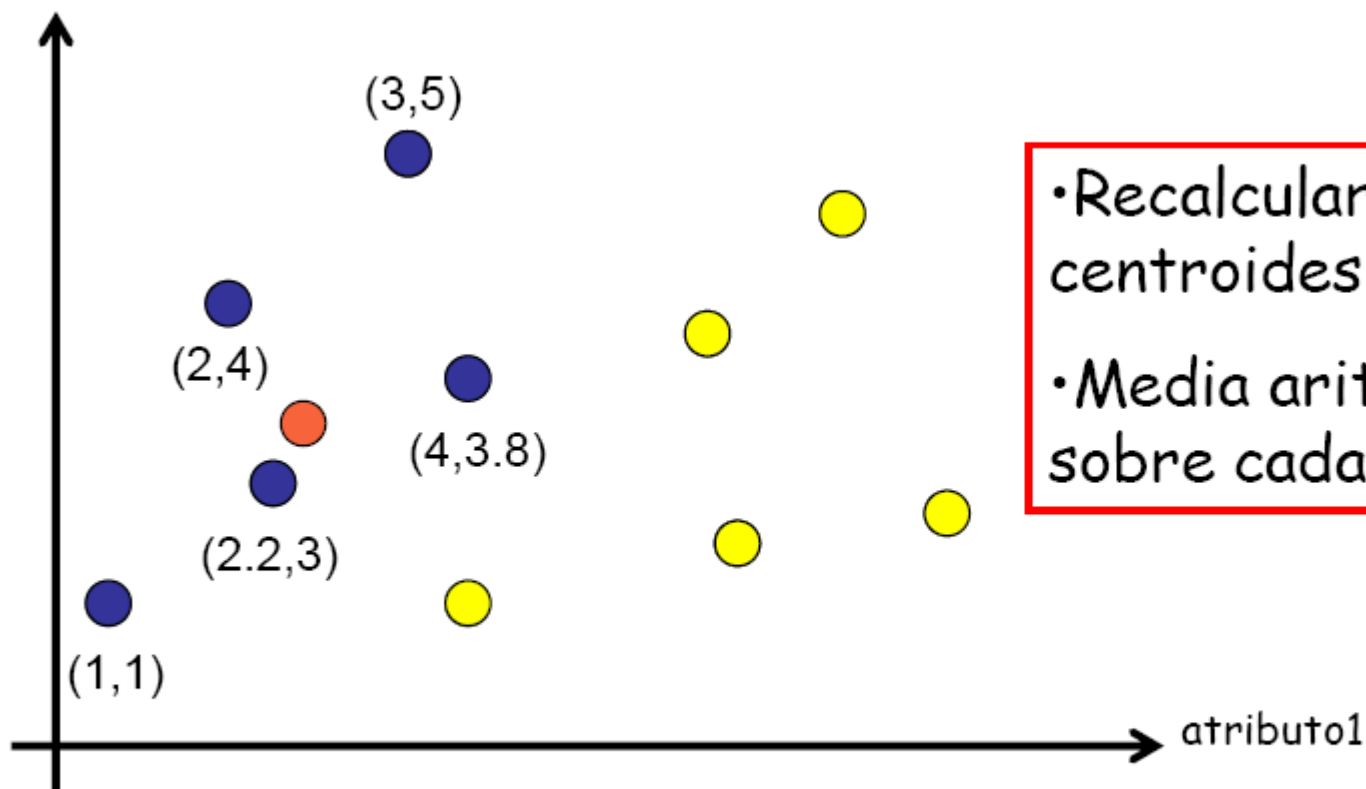


atributo2



- Recalcular los centroides
- Media aritmética sobre cada dimensión

atributo2

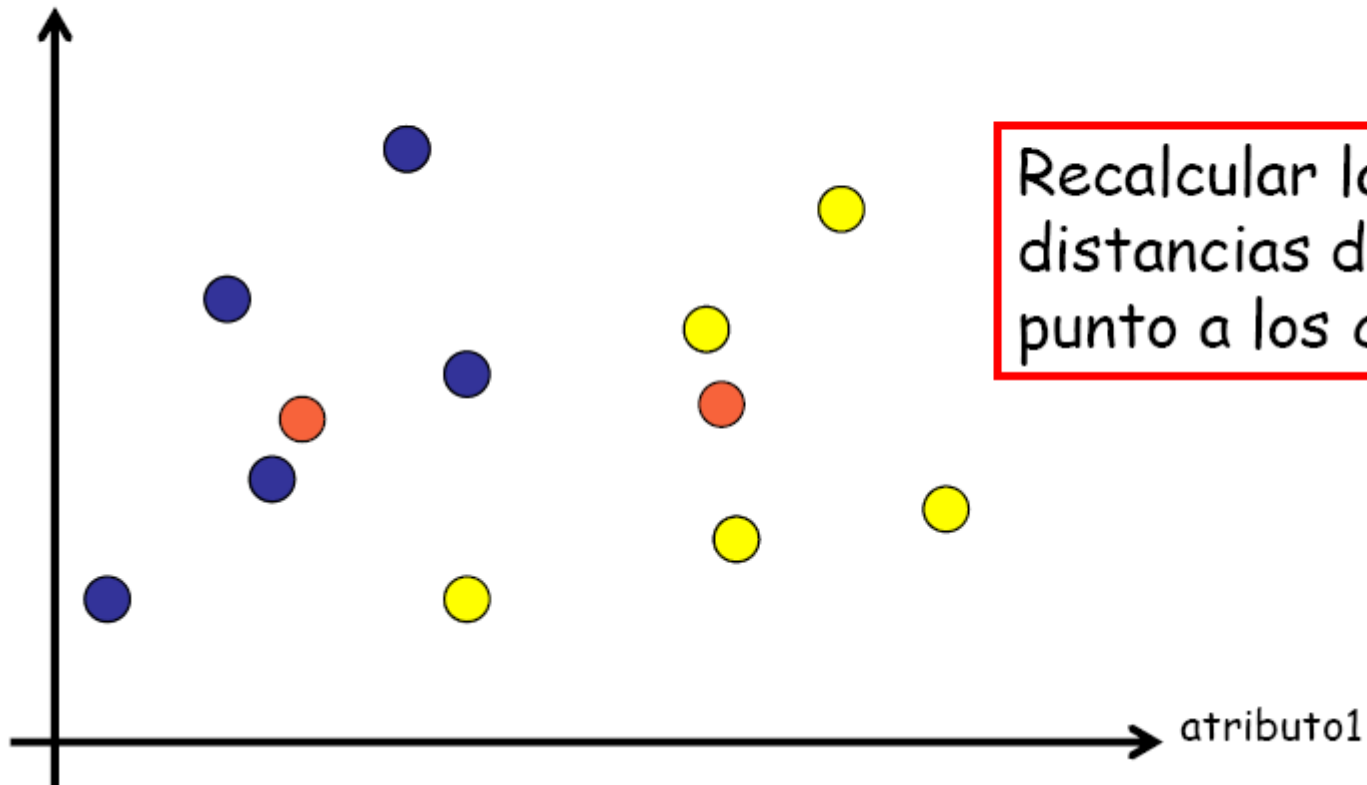


- Recalcular los centroides
- Media aritmética sobre cada dimensión

$$\text{Centroide}_{1_x} = (1+2+2.2+4+3)/5 = 2.4$$

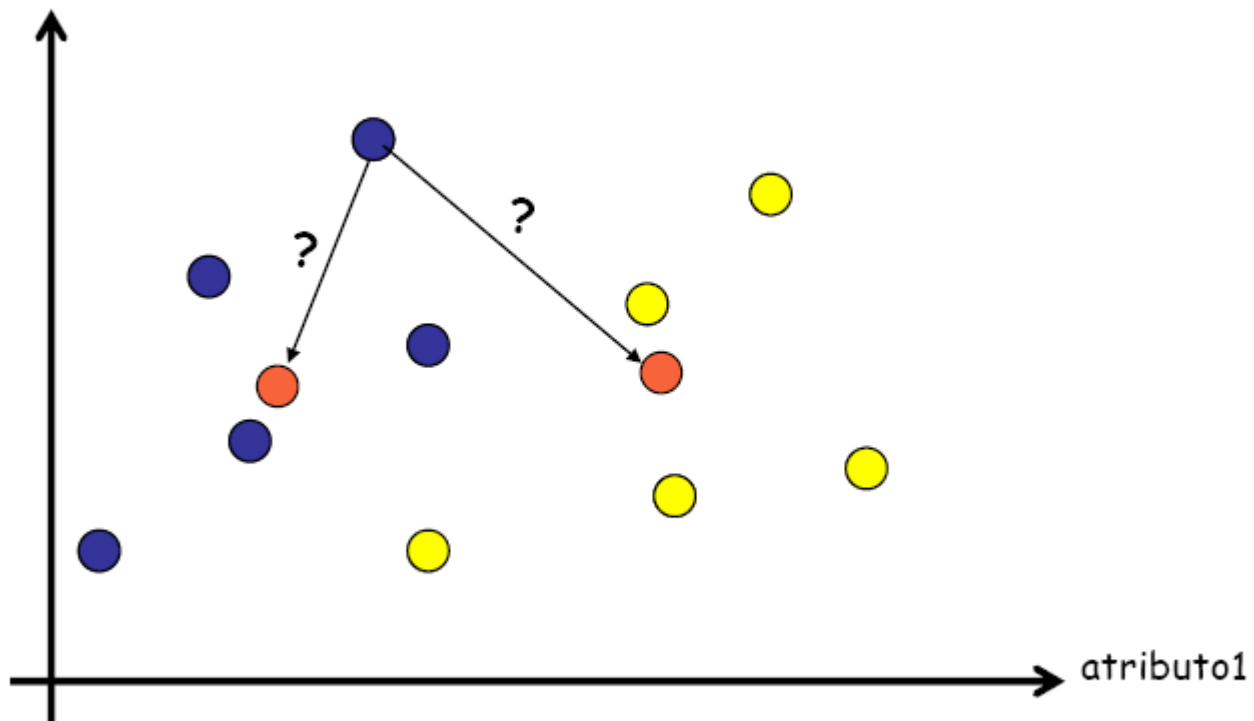
$$\text{Centroide}_{1_y} = (1+3+3.8+4+5)/5 = 3.3$$

atributo2



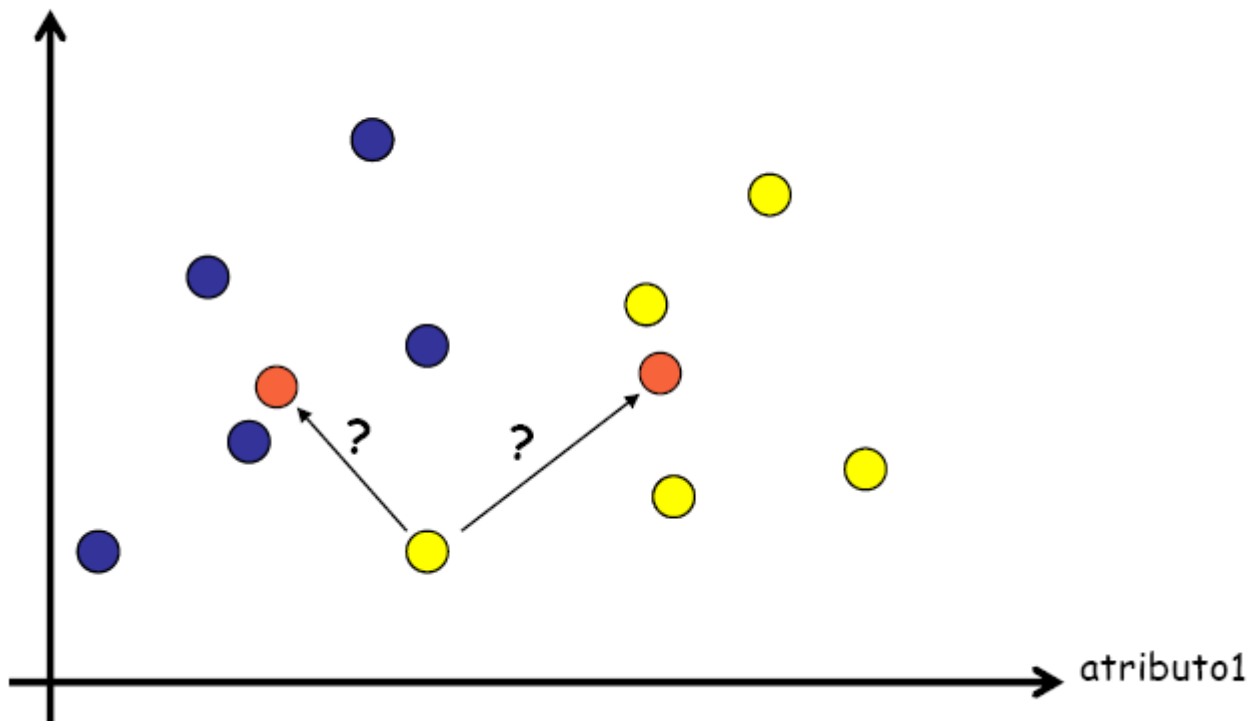
Recalcular las distancias de cada punto a los centroides

atributo2



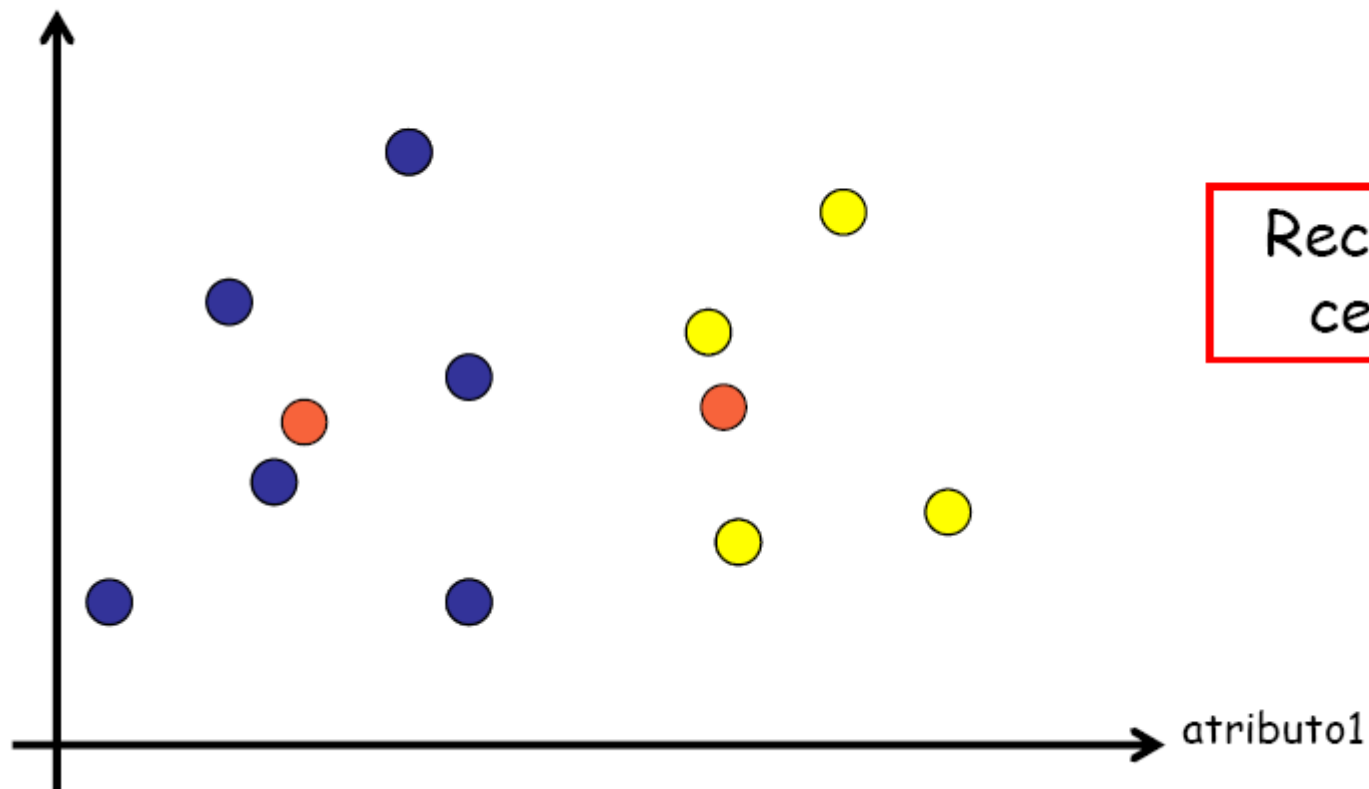
atributo1

atributo2



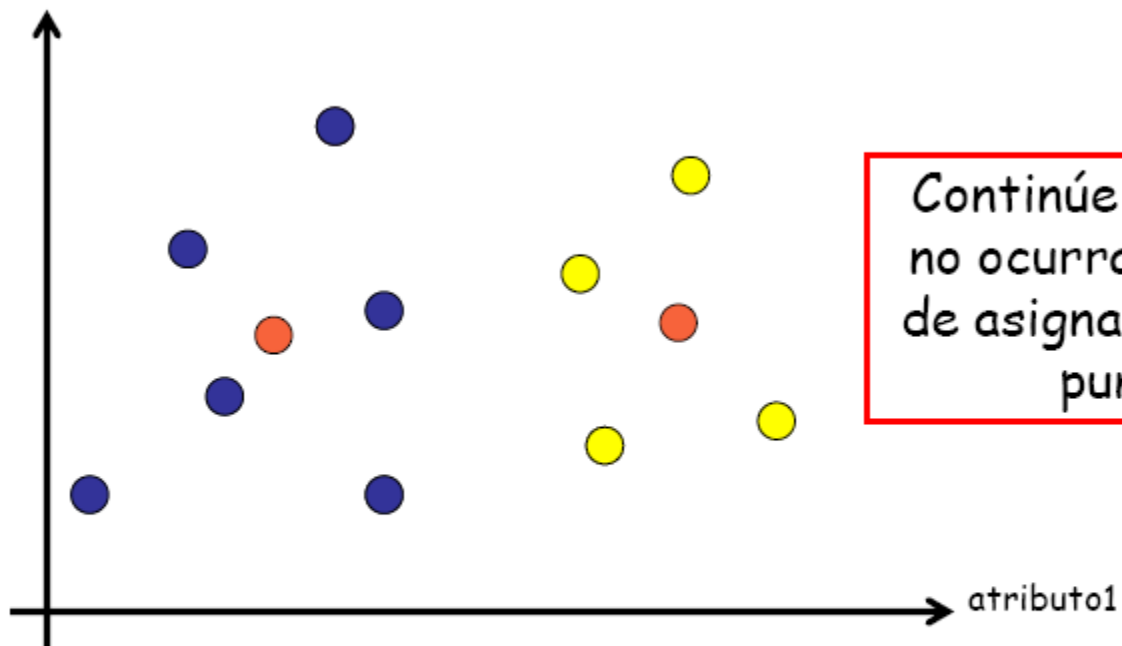
atributo1

atributo2



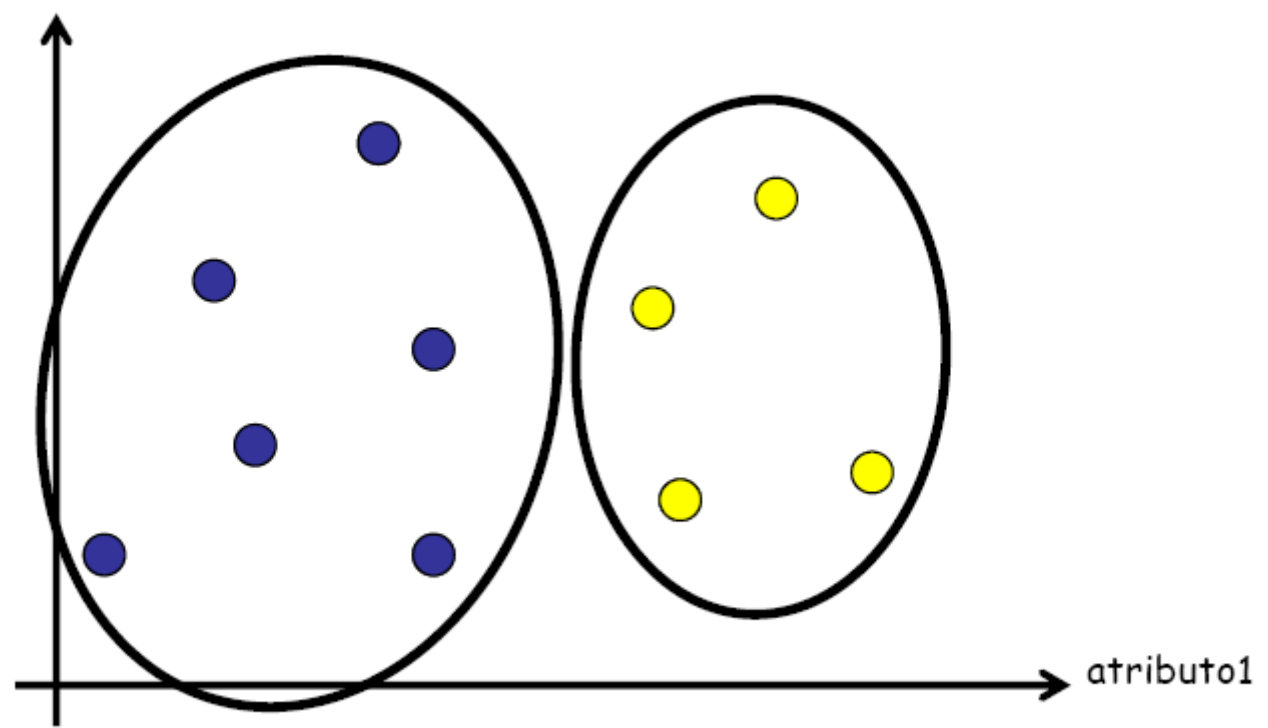
Recalcular los
centroides

atributo2



Continúe hasta que
no ocurran cambios
de asignación de los
puntos

atributo2



atributo1

Algoritmo Jerárquico

- El primer paso es calcular las distancias entre todos los pares de objetos. Esto es lo mismo que asumir que cada objeto constituye un cluster: $\{C_1, \dots, C_N\}$.
- Se buscan los dos clusters más cercanos (C_i, C_j) , éstos se juntan y constituyen uno solo C_{ij} .
- Se repite el paso 2 hasta que no quedan pares de comparación.

- Para clasificar los elementos en clusters, este algoritmo tiene dos variantes que pueden ser:
 - Acumulativos: se forman grupos haciendo *clusters cada vez más grandes*.
 - Disminutivos: partiendo de un solo grupo se separan los elementos en *clusters cada vez más pequeños*.

- Entre los algoritmos jerárquicos *acumulativos* destacan los siguientes métodos:
 - Método de las distancias mínimas: se busca la mayor semejanza entre los elementos o grupos más cercanos.
 - Método de las distancias máximas: se calcula la mínima distancia entre los elementos más alejados.
 - Método de las distancias medias: se calcula la media de las distancias entre elementos.

Ejemplo con mínima distancia

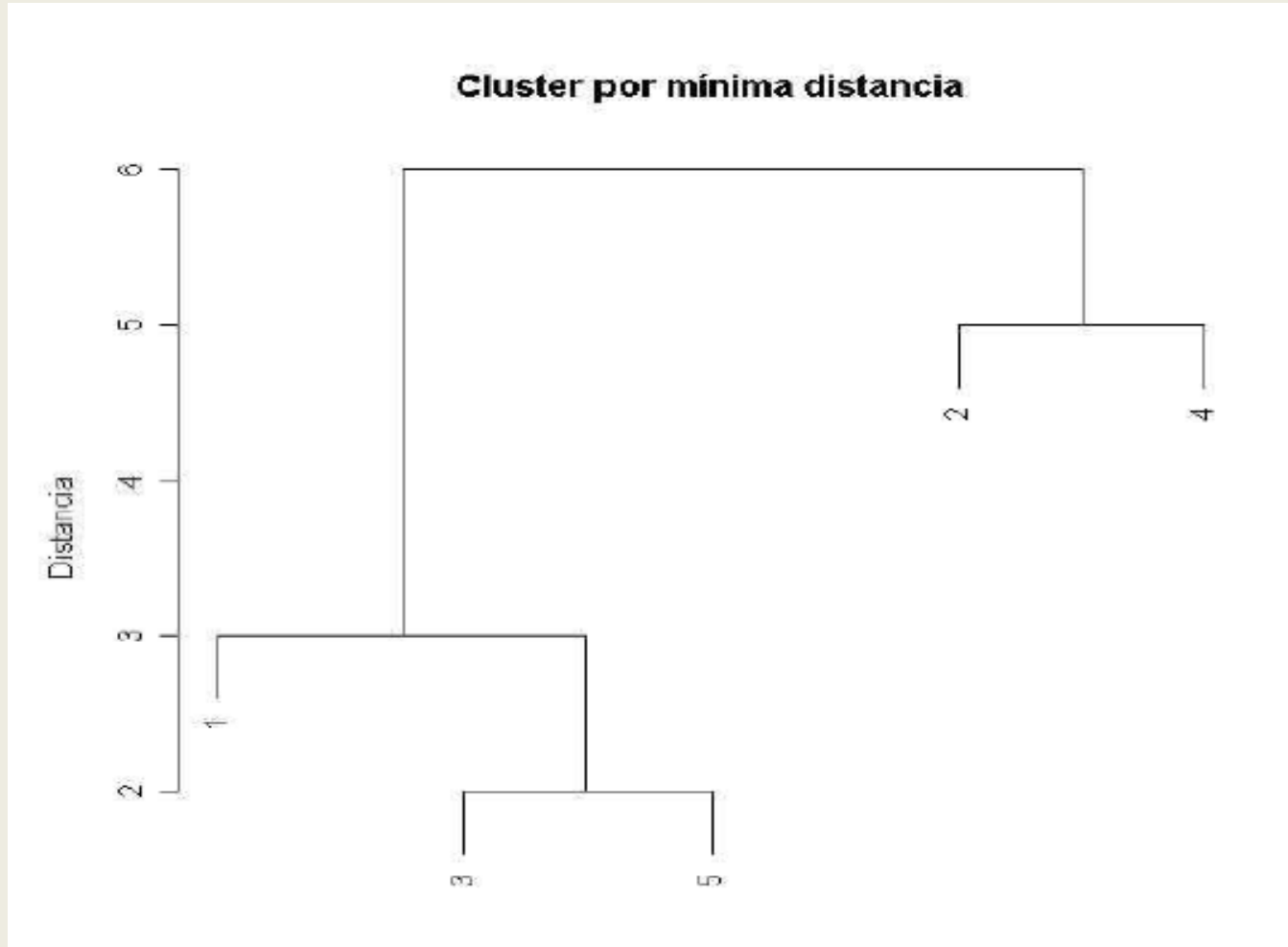
$$D = [d_{ik}]_{ik} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \mathbf{2} & 8 & 0 \end{bmatrix}$$

$$D = [d_{ik}]_{ik} = \begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{bmatrix} 0 & & & \\ \mathbf{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix}$$

$$D = [d_{ik}]_{ik} = \begin{matrix} & & (351) & 2 & 4 \\ (351) & & & & \\ 2 & & & & \\ 4 & & & & \end{matrix} \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix}$$

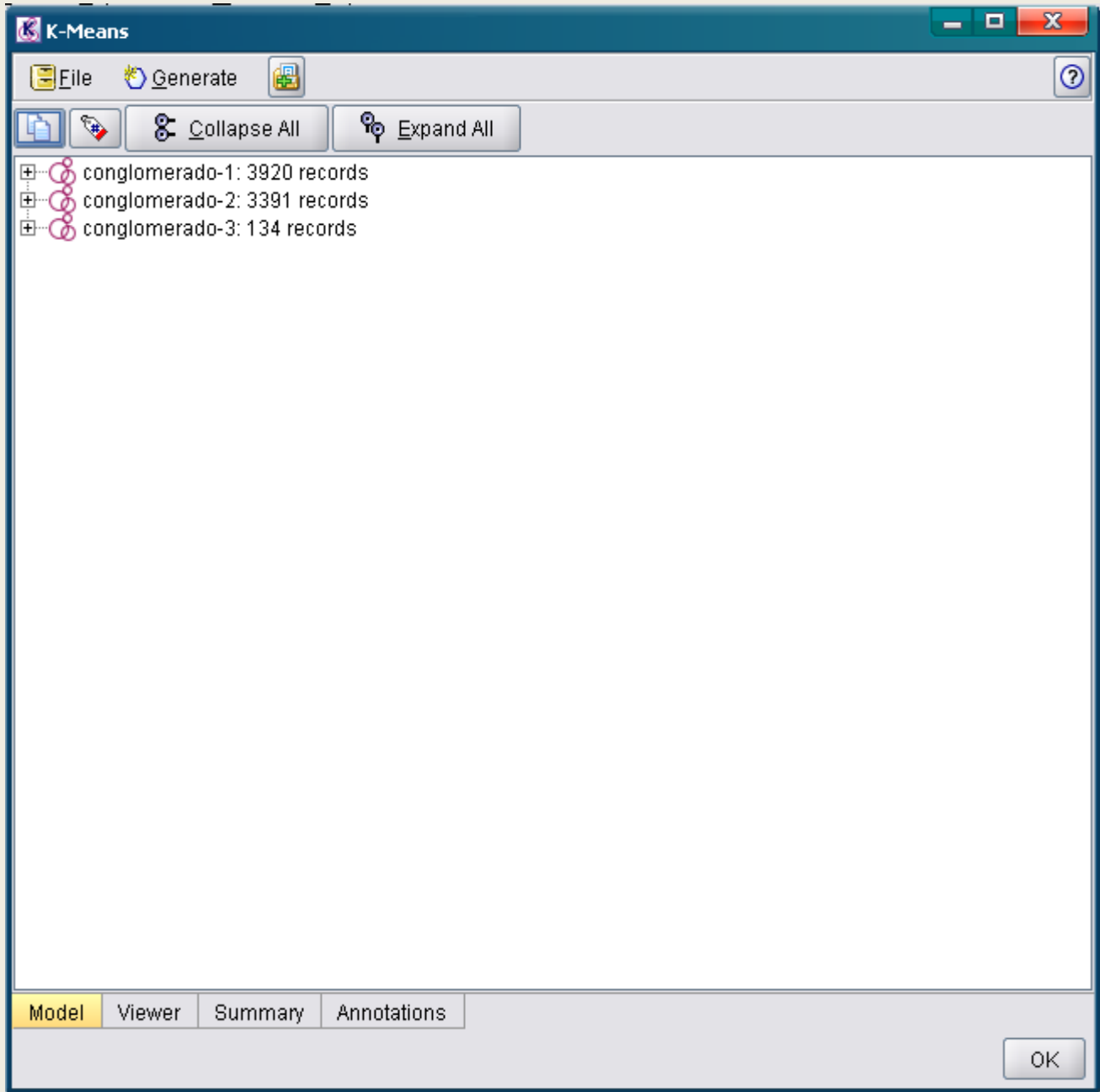
$$D = [d_{ik}]_{ik} = \begin{matrix} & & (351) & (24) \\ (351) & & & \\ (24) & & & \end{matrix} \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix}$$

Dendrograma




























Ejemplo de aplicación

- Utilización del algoritmo k- medias para clasificación de sistemas productivos en una provincia Argentina.
- Software utilizado: SPSS Clementine.
- Dimensiones del conjunto de datos:
 - Más de 9000 registros.
 - Más de 400 variables (atributos).
- Preprocesamiento y transformación de datos:
 - Reducción de cantidad de variables a 40.



Aclaración: los tipos de cultivos y bovinos son nombrados en forma genérica debido a que el resultado del proceso de Minería de Datos se encuentra en etapa de interpretación y el mismo no ha sido publicado.

K-Means
KMeans

	conglomerado-1	conglomerado-2	conglomerado-3	Importanc
				<ul style="list-style-type: none">  $\geq 0,05$  $\geq 0,90$  $< 0,90$  Desconocido
Cultivo a	—	—		Importanc  1,00
Cultivo b	—	—		Importanc  1,00
Cultivo c	—	—		Importanc  1,00
Cultivo d	—	—		Importanc  1,00
Bovino V	—	—		Importanc  1,00
Bovino W	—	—		Importanc  1,00
Bovino X	—	—		Importanc  1,00
Bovino Y	—	—		Importanc  1,00
Bovino Z	—	—		Importanc  1,00

Minería de Datos y Grid Computing

- Como los datos día a día crecen en dimensiones descomunales, las computadoras convencionales son muy limitadas para ofrecer un buen rendimiento a los procesos de minería.
- Una posible solución es la Computación Grid, la cual busca solucionar problemas que no pueden ser resueltos en un tiempo razonable con computadoras convencionales, mediante el uso de diferentes procesadores y/o máquinas conectados a una red que se distribuyen las tareas y finalmente se obtengan resultados más rápida y eficientemente.

Conclusión

- La Minería de datos es una herramienta que permite convertir los datos almacenados en información valiosa.
- Los campos en los que se pueden aplicar estas técnicas son extremadamente variados, siempre que se disponga de un conjunto de datos.
- En el INTA, permitiría crear modelos para predecir lluvias, rendimiento de cultivos, etc.

Preguntas

¿?

**Muchas
Gracias!**

Bibliografía

- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Cambridge, Mass., 1996.
- Apuntes proporcionados por la Cátedra.