

# Qué es DataMining?

Mg. Cecilia Ruz  
Luis Azaña Bocanegra

# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación

# Agenda

- ▶ **Qué es Data Mining?**
- ▶ **Cómo se integra en el proceso de Descubrimiento del conocimiento?**
- ▶ **Funcionalidades del Data Mining**
- ▶ **Técnicas**
  - ▶ **Supervisadas**
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ **No supervisadas**
    - ▶ Clustering
    - ▶ Reglas de Asociación

# Qué es Data Mining?

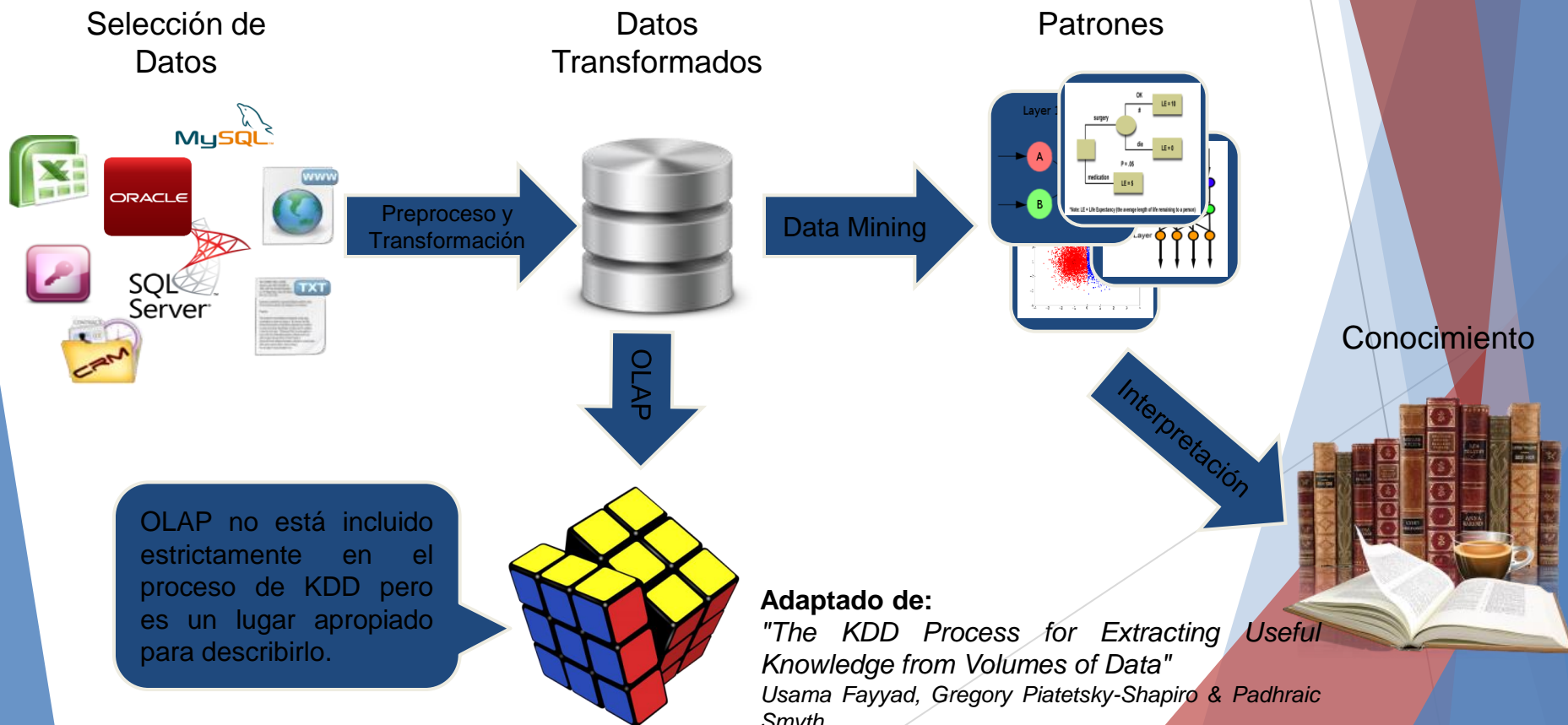
- ▶ “Es la extracción de patrones o información interesante ( no trivial, implícita, previamente desconocida y potencialmente útil) de grandes bases de datos”
- ▶ Esta definición tiene numerosas cosas a definir, que quiere decir no-trivial, útiles para quién?

# Agenda

- ▶ Qué es Data Mining?
- ▶ **Cómo se integra en el proceso de Descubrimiento del conocimiento?**
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación

# El Proceso de KDD

El proceso de descubrimiento de conocimiento en los datos, KDD (Knowledge Discovery in Data), se representa de la siguiente forma:



# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ **Funcionalidades del Data Mining**
- ▶ **Técnicas**
  - ▶ **Supervisadas**
    - ▶ Árboles
    - ▶ Redes neuronales
    - ▶ Regresión
  - ▶ **No supervisadas**
    - ▶ Clustering
    - ▶ Reglas de Asociación

# Funcionalidades del DM (1)

- ▶ Descripción de conceptos: Caracterización y discriminación
  - ▶ Generalizar, Resumir y contrastar las características de la información ( por ejemplo las regiones secas vs. Las regiones húmedas)
- ▶ Asociación (correlación y causalidad)
  - ▶ Multi-dimensionales vs. unica dimensión
  - ▶  $\text{age}(X, "20..29") \wedge \text{income}(X, "20..29K") \rightarrow \text{buys}(X, "PC")$   
[support = 2%, confidence = 60%]
  - ▶  $\text{contains}(T, "computer") \rightarrow \text{contains}(T, "mouse")$  [1%, 75%]



# Funcionalidades del DM (2)

## ▶ Clasificación y Predicción

- ▶ Encontrar modelos o funciones que describan y distingan clases para futuras predicciones
- ▶ Ej, Clasificar países de acuerdo a su clima, clientes de acuerdo a su comportamiento, contenedores de acuerdo a su riesgo.
- ▶ Presentación: árboles de decisión, reglas de clasificación, redes neuronales
- ▶ Predicción: Predecir valores numéricos desconocidos o faltantes.

## ▶ Cluster analysis

- ▶ No se sabe a que clase pertenecen los datos : se agrupan datos para formar clases,
- ▶ El Clustering se basa en el principio de maximizar la similitud dentro de la clase y minimizar la misma entre clases

# Funcionalidades del DM(3)

## ▶ Análisis de Outliers

- ▶ Outlier: un dato ( o un objeto) que no respeta el comportamiento general.
- ▶ Puede ser ruido o excepciones, pero son muy útiles en la detección de fraudes o eventos raros.

## ▶ Análisis de tendencias y evolución

- ▶ Tendencia y Desvíos: análisis de regresión
- ▶ Análisis de patrones secuenciales
- ▶ Análisis de similitudes
- ▶ Series de tiempo

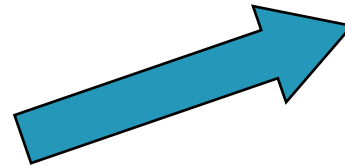
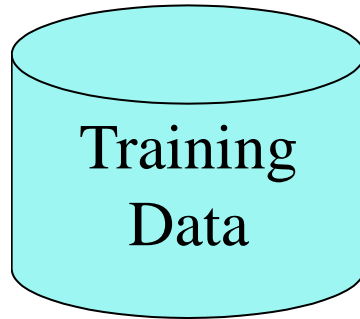
# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación

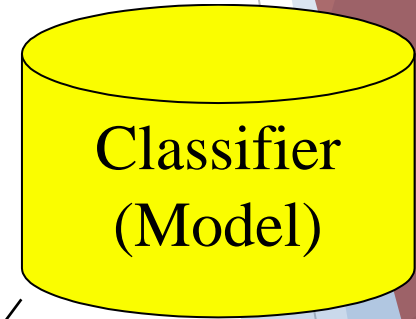
# Clasificación—Un proceso de dos pasos

- ▶ Construcción del modelo: descripción de las clases existentes
  - ▶ Cada ejemplo pertenece a una clase determinada
  - ▶ El training set es el conjunto de ejemplos que se usa para entrenar el modelo
  - ▶ El modelo se representa por medio de reglas de clasificación, árboles o fórmulas matemáticas
- ▶ Uso del modelo: para clasificar ejemplos futuros o desconocidos
  - ▶ Estimar la precisión del modelo
    - ▶ Para esto se aplica el modelo sobre un conjunto de test y se compara el resultado del algoritmo con el real.
    - ▶ Precisión es el porcentaje de casos de prueba que son correctamente clasificados por el modelo
    - ▶ El conjunto de entrenamiento debe ser independiente del de test para evitar “overfitting”

# Proceso de Clasificación (1): Construcción del Modelo



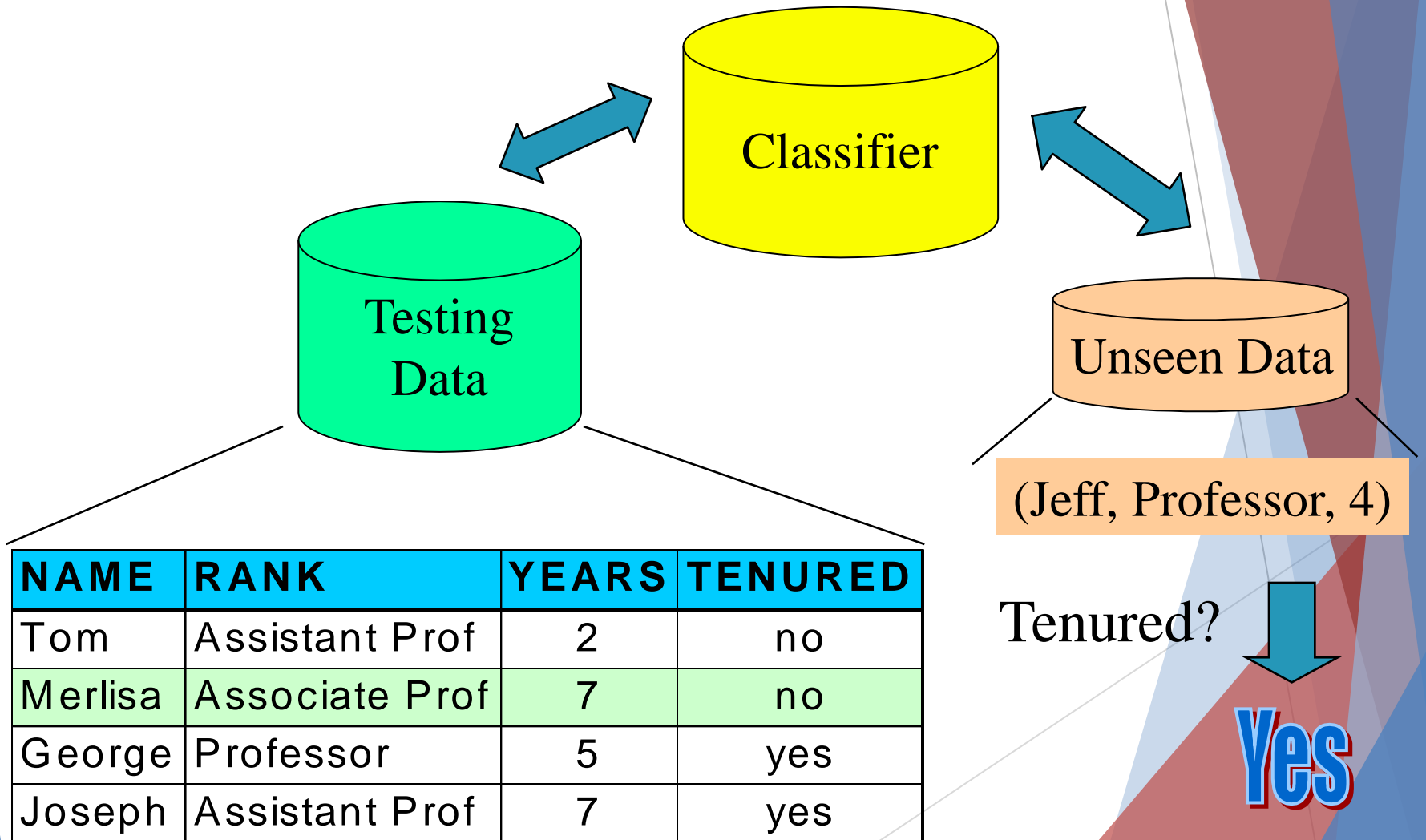
Classification Algorithms



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = 'professor'  
OR years > 6  
THEN tenured = 'yes'

# Proceso de Clasificación (2): Uso del modelo para predecir



# Redes Neuronales (1)

Son sistemas :

- ▶ Capaces de aprender
- ▶ Adaptarse a a condiciones variantes
- ▶ Adaptarse al ruido
- ▶ Predecir el estado futuro
- ▶ Enfrentar problemas que eran resueltos sólo por el cerebro humano

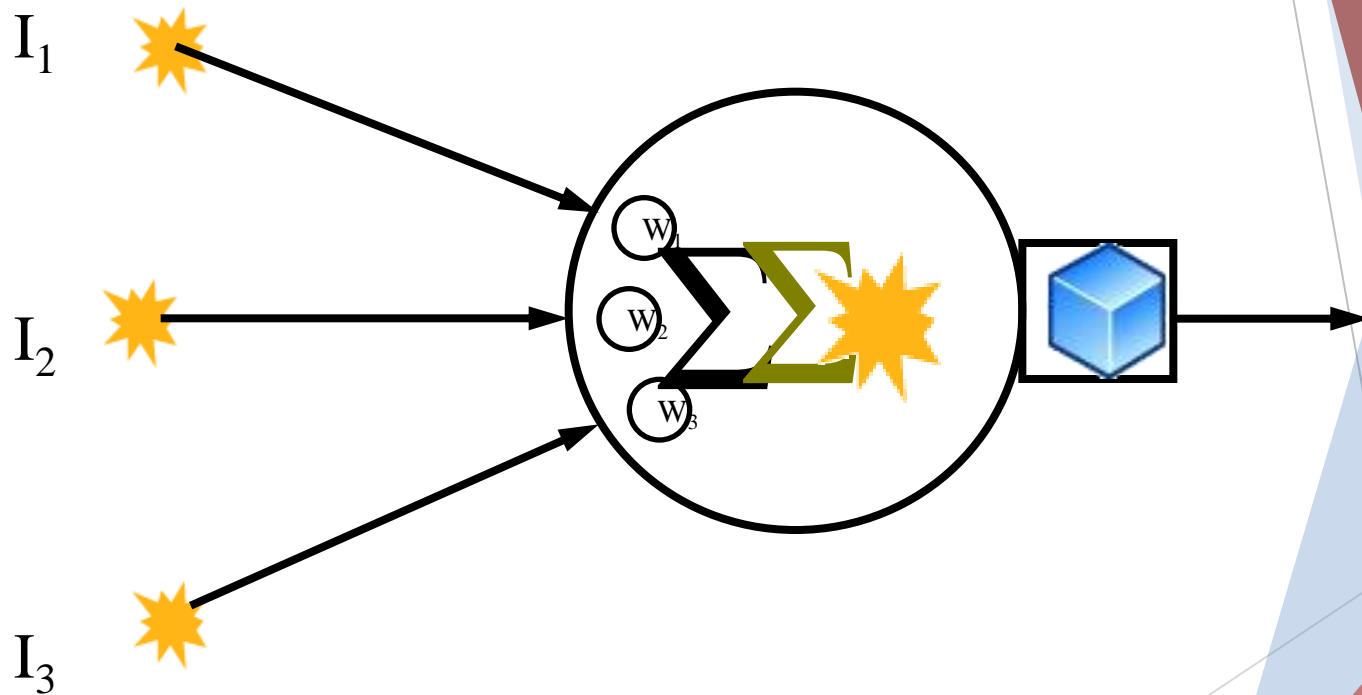
# Redes Neuronales (2)

No son algorítmicas

- ▶ No se programan haciéndoles seguir una secuencia predefinida de instrucciones.
- ▶ Las RNA generan ellas mismas sus propias "reglas", para asociar la respuesta a su entrada;
- ▶ Aprenden por ejemplos y de sus propios errores.
- ▶ Utilizan un procesamiento paralelo mediante un gran número de elementos altamente interconectados.



# Redes Neuronales - Neurona Modelo

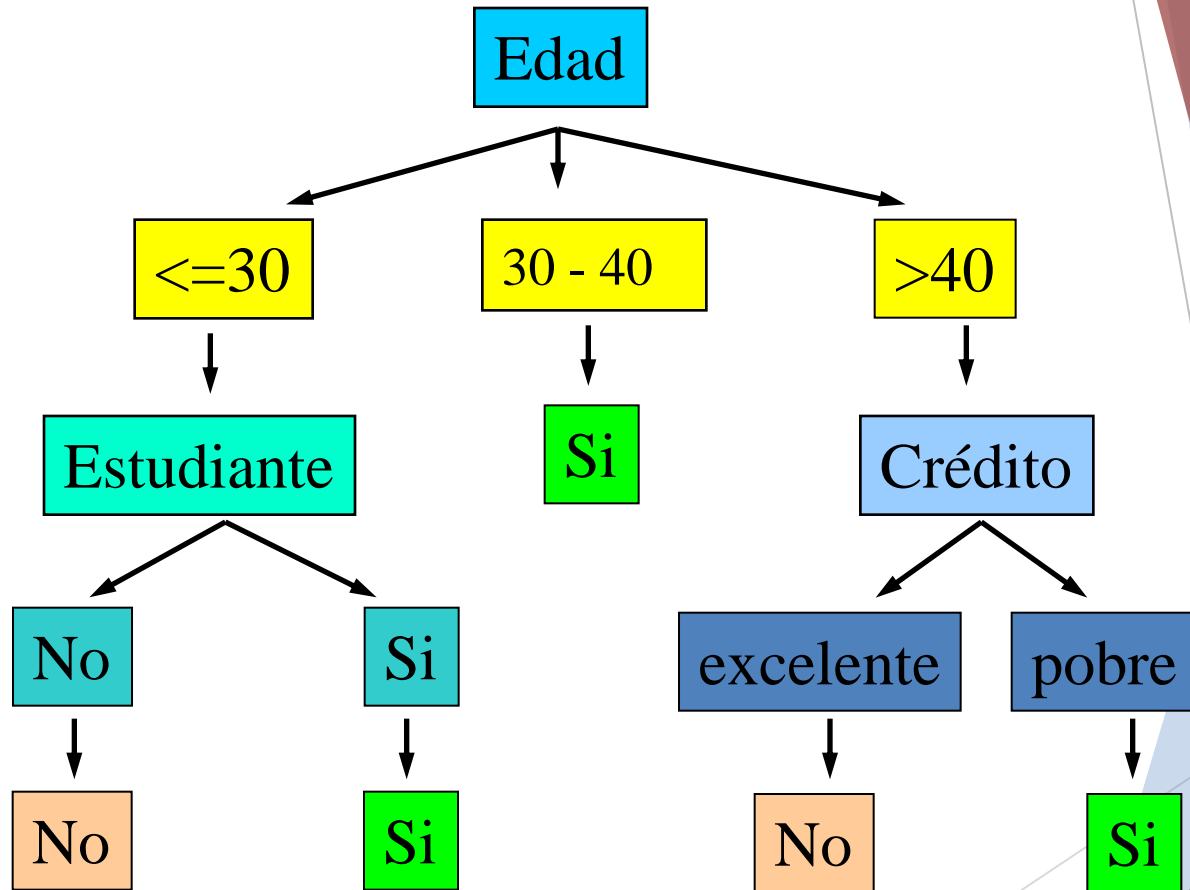


$$I_1 \cdot W_1 + I_2 \cdot W_2 + I_3 \cdot W_3$$

# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación

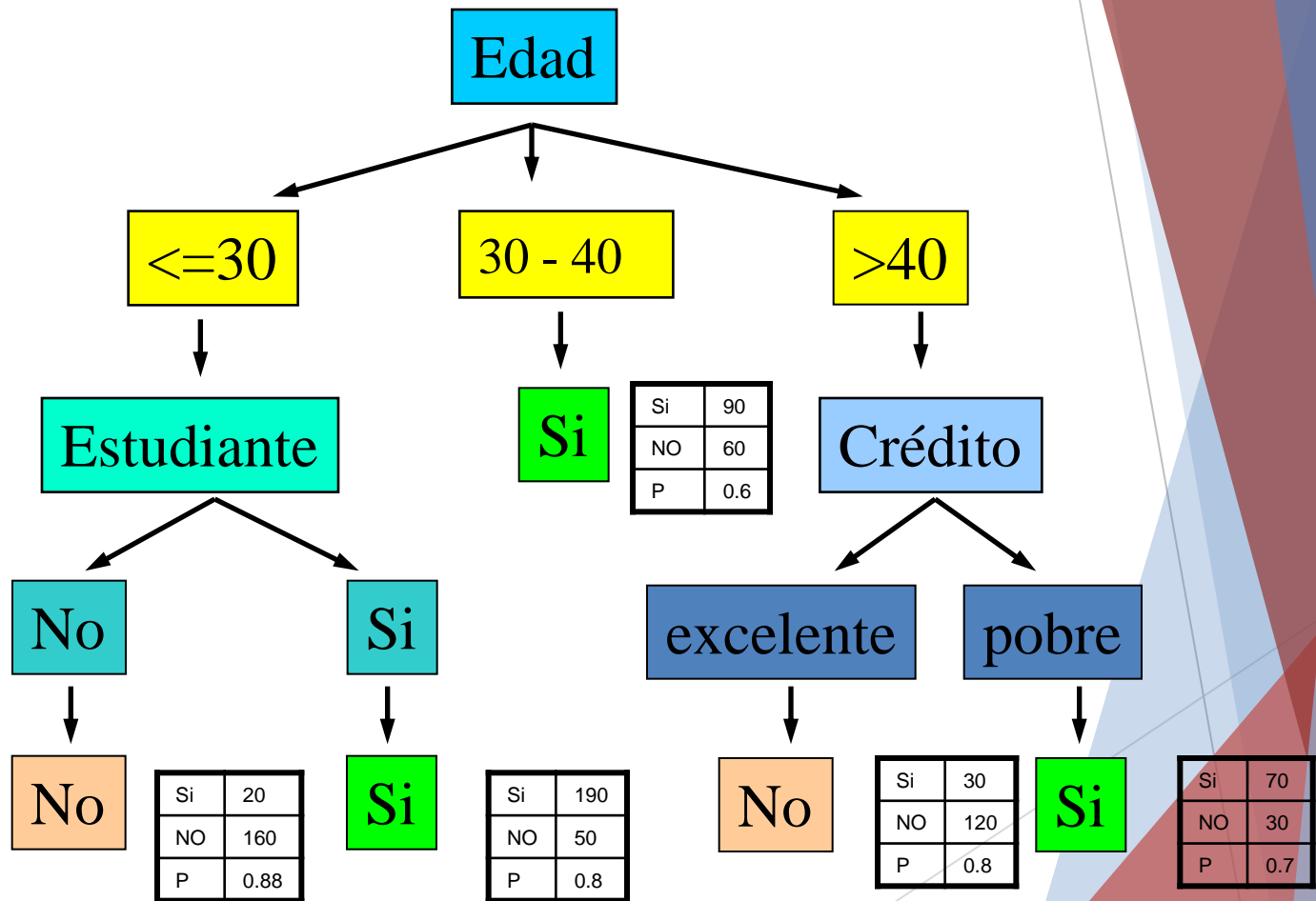
# Árbol de Decisión para ver quien compra una computadora



# Clasificación por medio de Árboles de Decisión

- ▶ Árboles de Decisión
  - ▶ Los nodos internos son preguntas sobre los atributos
  - ▶ Las hojas representan las etiquetas o clases resultantes
- ▶ La generación del árbol tiene fundamentalmente dos pasos
  - ▶ Construcción
    - ▶ Al comienzo todos los ejemplos están en la raíz del árbol
    - ▶ Se dividen los ejemplos en forma recursiva basado en atributos elegidos
  - ▶ Pruning
    - ▶ Identificar y remover ramas que representan outliers o ruido
- ▶ Uso de los árboles de decisión: clasificación de un ejemplo desconocido
  - ▶ Se controlan los valores de los atributos del ejemplo para asignarle la clase

# Árbol de Decisión con Probabilidad



# Extracción de reglas de clasificación a partir de los árboles

- ▶ Representa el conocimiento en la forma de reglas de **IF-THEN**
- ▶ Se genera una regla para cada camino desde la raíz hasta las hojas.
- ▶ Cada par atributo - valor forma una conjunción
- ▶ La hoja tiene la clase a predecir
- ▶ Las reglas son fácilmente entendibles por los seres humanos
- ▶ Ejemplos

```
IF edad = "<=30" AND estudiante = "no" THEN compra_PC = "no"
```

```
IF edad = "<=30" AND estudiante = "yes" THEN compra_PC = "si"
```

```
IF edad = "31 - 40" THEN compra_PC = "si"
```

```
IF edad = ">40" AND credito = "excelente" THEN compra_PC = "si"
```

```
IF edad = ">40" AND credito = "pobre" THEN compra_PC = "no"
```

## Evitar el Overfitting en la clasificación

- ▶ El árbol obtenido puede hacer overfitting sobre el conjunto de entrenamiento
  - ▶ Si hay demasiadas ramas algunas pueden reflejar anomalías
  - ▶ Como consecuencia de esto se tiene una performance muy mala sobre ejemplos nuevos
- ▶ Dos aproximaciones para evitar el overfitting
  - ▶ Prepruning: Interrumpir la construcción del árbol en forma anticipada. No partir un nodo si la mejora que esto produce está por debajo de un cierto umbral.
    - ▶ Es difícil encontrar el umbral adecuado
  - ▶ Postpruning: quitar ramas de un árbol ya contruido
    - ▶ Se puede usar un conjunto diferente del de entrenamiento para hacer esto.

# Matriz de confusión

		Clase Predicha	
		Bueno	Malo
Clase Real	Bueno	15	5
	Malo	10	115

**Error de la predicción =  $(10 + 5) / (15 + 5 + 10 + 115)$**

**Falsos Positivos (FP) =  $10/145$**

**Falsos Negativos (FN) =  $5/145$**

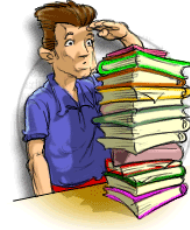


# Detección de Valores Extremos, Outliers

Los conjuntos de datos que analizamos generalmente proporcionan un subconjunto de datos en el que existe una variabilidad y/o una serie de errores. Estos datos siguen un comportamiento diferente al resto del conjunto ya sea en una o varias variables. Muchas veces es útil estudiarlos para detectar anomalías, mientras que otras veces es mejor descartarlos de los análisis porque ensucian o influyen en los resultados (por ejemplo en los promedios).



# Definiciones

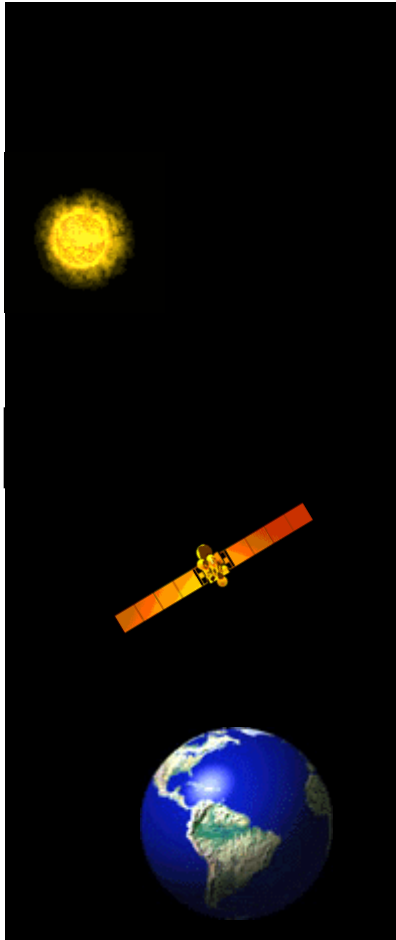


A la vista de lo anterior, podemos clasificar las observaciones atípicas o anómalas como:

- Observación **atípica**: Es aquel valor que presenta una gran variabilidad de tipo inherente.
- Observación **errónea**: Es aquel valor que se encuentra afectado de algún tipo de error, sea del medio, del experimentador, o de ambos.

Se llamará “outlier” a aquella observación que siendo atípica y/o errónea, tiene un comportamiento muy diferente respecto al resto de los datos, en relación al análisis que se desea realizar sobre las observaciones. Análogamente, se llamará “inlier” a toda observación no considerada como outlier.

# Outliers Peligrosos

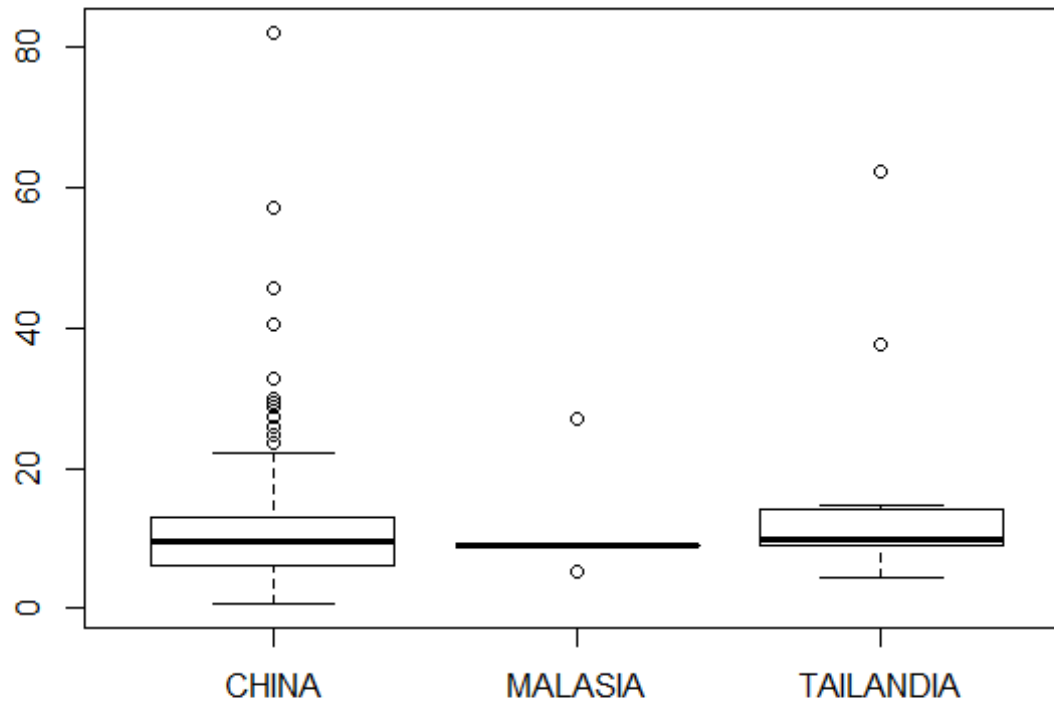


El “agujero de ozono” sobre la antártida es un ejemplo de uno de los outliers más infames de la historia reciente. Es también un buen ejemplo para decir a los que eliminan sistemáticamente los outliers de un dataset simplemente porque son outliers. En 1985 tres investigadores (Farman, Gardinar y Shanklin) fueron desconcertados por un ciertos datos recopilados por el “examen antártico británico” que demostraba que los niveles del ozono para la antártida habían caído el 10% debajo de los niveles normales de enero. El problema era, porqué el satélite Nimbo 7, que tenía instrumentos a bordo para medir con precisión los niveles del ozono, no había registrado concentraciones de ozono semejantemente bajas. Cuando examinaron los datos del satélite no les tomó mucho darse cuenta de que el satélite de hecho registraba estos niveles de concentraciones bajos y lo había estado haciendo por años. ¡Pero como las concentraciones de ozono registradas por el satélite fueron tan bajas eran tratadas como outliers por un programa de computadora y desechadas! El satélite Nimbo 7 de hecho había estado recolectando la evidencia de los niveles bajos de ozono desde 1976. El daño a la atmósfera causada por los clorofluocarburos pasó desapercibido y no fue tratado por nueve años porque los outliers fueron desechados sin ser examinados.

**Moraleja: No tirar los outliers sin examinarlos, porque pueden ser los datos más valiosos de un dataset.**

# Ejemplos

**Precio Unitario por Unidad Estadística posición 9503.00**



# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación

# Regresión Lineal

Para poder crear un modelo de regresión lineal, es necesario que se cumpla con los siguientes supuestos:

- ▶ La relación entre las variables es lineal.
- ▶ Los errores son independientes.
- ▶ Los errores tienen varianza constante.
- ▶ Los errores tienen una esperanza matemática igual a cero.
- ▶ El error total es la suma de todos los errores.

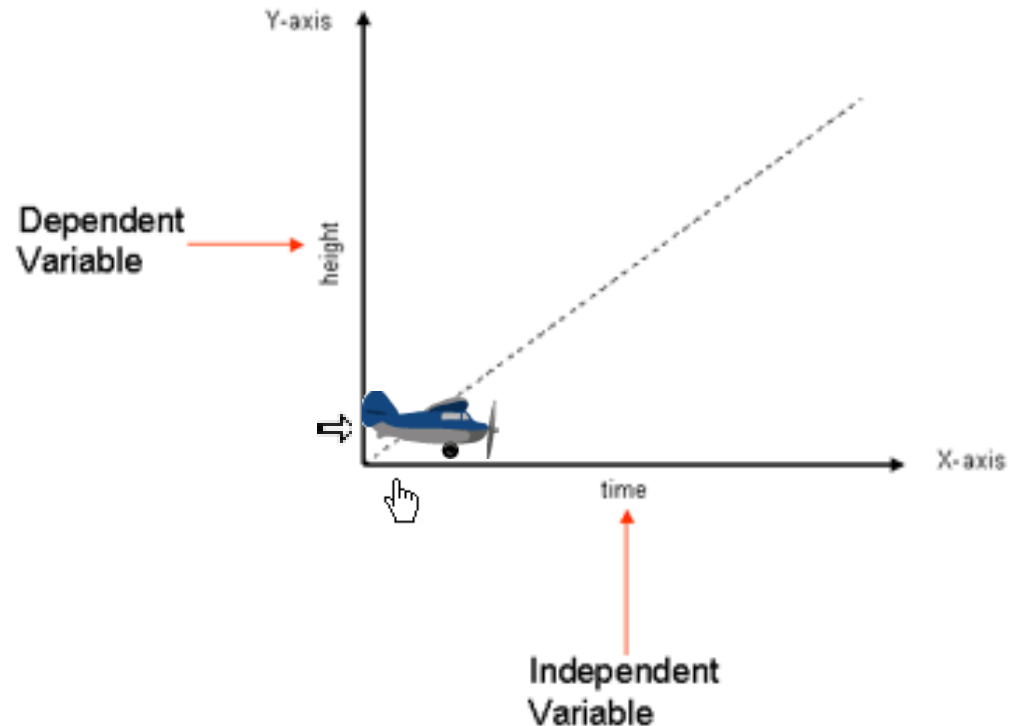
# Tipos de Regresión Lineal

- ▶ Regresión lineal simple. Sólo se maneja una variable independiente
- ▶ Regresión lineal múltiple. Maneja varias variables independientes.

# Represión Lineal Ejemplo

## **Variables dependientes:**

Son las variables de respuesta que se observan en el estudio y que podrían estar influidas por los valores de las variables independientes.



**Variables independientes:** Son las que se toman para establecer agrupaciones en el estudio, clasificando intrínsecamente a los casos del mismo



# Regresión Logística

- ▶ La regresión logística Se aplica cuando la variable dependiente es dicotómica o politómica y no numérica
- ▶ Para poder aplicar una regresión se asocia la variable dependiente a su probabilidad de ocurrencia.
- ▶ Por lo tanto el resultado de un regresión logística es la probabilidad de ocurrencia del suceso

# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación

# Qué es el análisis de clusters?

- ▶ Cluster: una colección de objetos
  - ▶ Similares dentro del cluster
  - ▶ Diferentes de los objetos en los otros clusters
- ▶ Cluster analysis
  - ▶ Agrupar un conjunto de datos en un cluster
- ▶ Clustering es **clasificación no supervisada** : no hay clases predefinidas
- ▶ Aplicaciones típicas
  - ▶ Como una herramienta independiente para tener una idea sobre la distribución de los datos
  - ▶ Como un proceso previo a usar otros algoritmos
  - ▶ Como una manera de detectar outliers

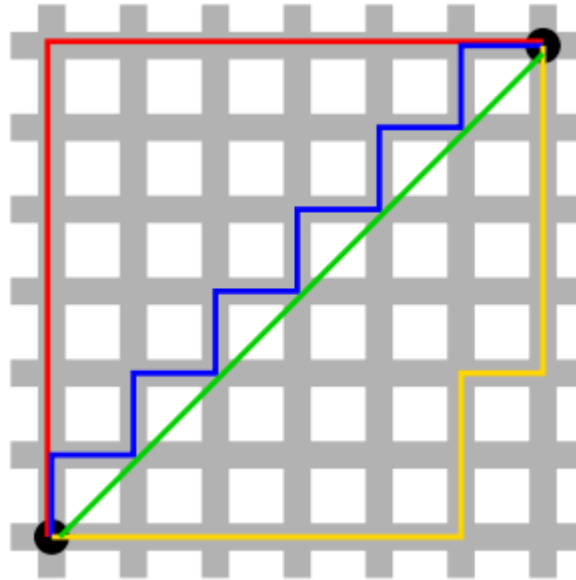
# Qué es un buen Clustering?

- ▶ Un buen método de clustering produce clusters de alta calidad con
  - ▶ Alta similitud en la clase
  - ▶ Baja similitud entre clases
- ▶ La calidad de un clustering depende de la medida de “similitud” usada por el método y de la forma en que está implementado.

# Medición de la calidad de un cluster

- ▶ Medida de similitud: La similitud está expresada en base a una función de distancia
- ▶ Hay una función separada que mide la bondad del clustering
- ▶ Las funciones de distancia a utilizar son muy diferentes de acuerdo al tipo de dato.
- ▶ Algunas veces es necesario asignarle “peso” a las variables dependiendo del significado que tienen para el problema

# Manhattan versus Euclidean



El rojo, azul, y amarillo representan la distancia Manhattan, todas tienen el mismo largo(12),mientras que la verde representa la distancia Euclidia con largo de  $6 \times \sqrt{2} \approx 8.48$ .

# Variables numéricas

- ▶ Estandarizar los datos

- ▶ Calcular la desviación absoluta de la media

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

donde  $m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf})$ .

- ▶ Normalizar (z-score)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

# Variables binarias

- ▶ Una tabla de contingencia

		Object $j$		
		1	0	<i>sum</i>
Object $i$	1	$a$	$b$	$a+b$
	0	$c$	$d$	$c+d$
	<i>sum</i>	$a+c$	$b+d$	$p$

- ▶ Coeficiente simple

$$d(i, j) = \frac{b+c}{a+b+c+d}$$

- ▶ Coeficiente de Jaccard :  $d(i, j) = \frac{b+c}{a+b+c}$

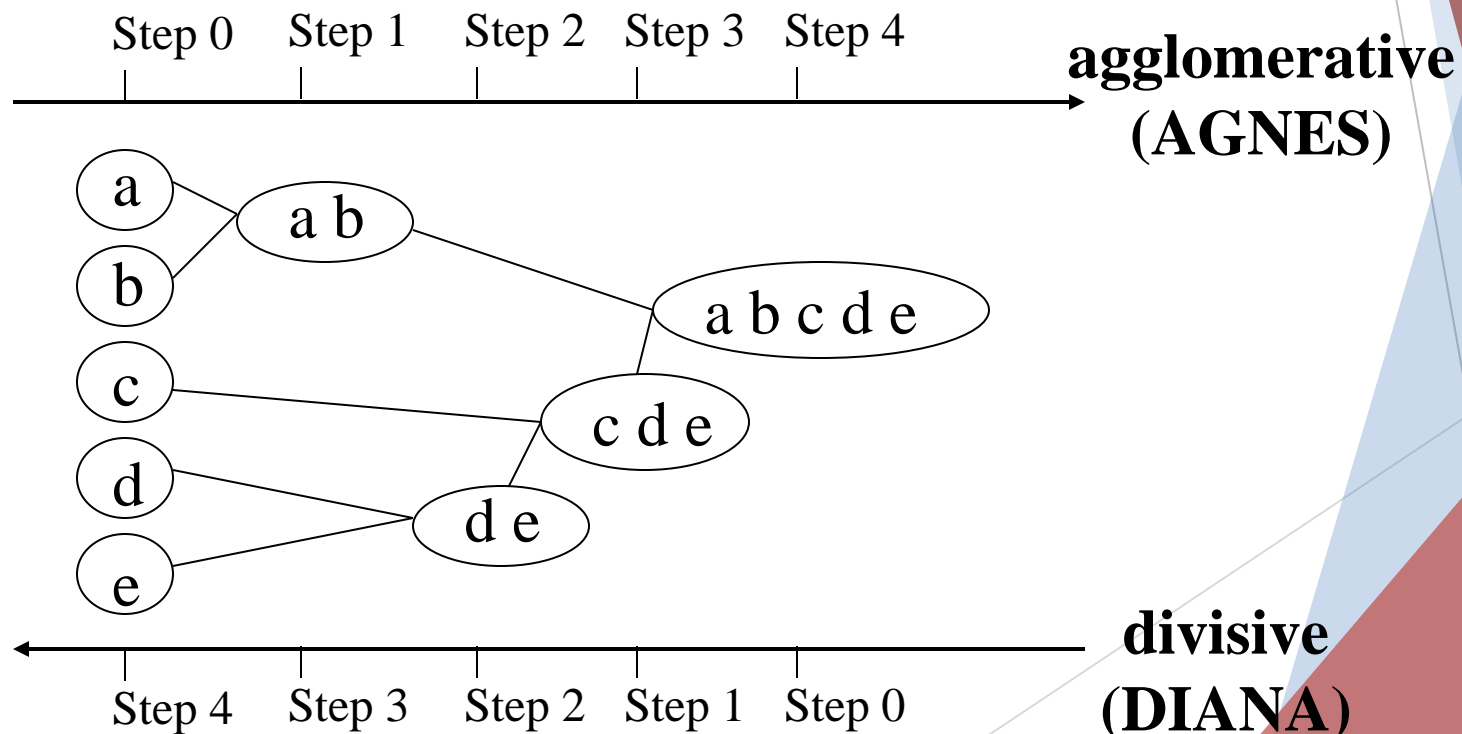


# Formas de obtener un cluster

- ▶ Jerárquicas
- ▶ No jerárquicas

# Clustering Jerárquico

- ▶ Usa la matriz de distancia como criterio. No requiere que el número de cluster sea uno de los parámetros de input



# Agrupamiento aglomerativo

- ▶ **Métodos de enlace**

- ▶ Enlace simple (distancia mínima)
- ▶ Enlace Completo (distancia máxima)
- ▶ Enlace promedio

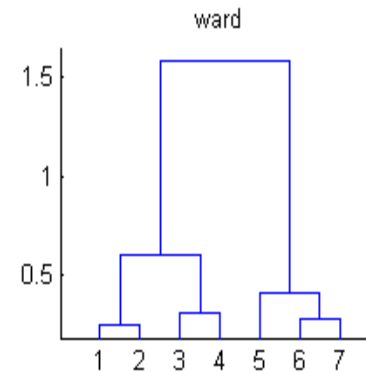
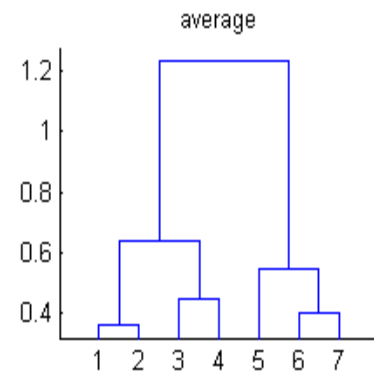
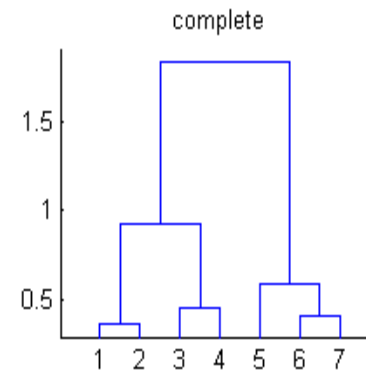
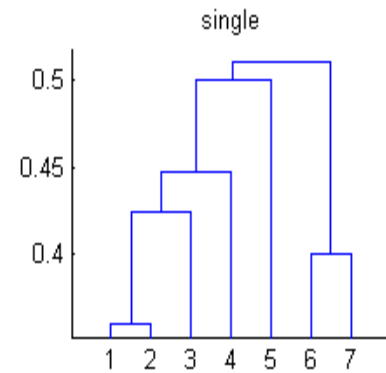
- ▶ **Método de Ward**

1. Calcular la suma de las distancias al cuadrado dentro de los clusters
2. Agregar clusters con incremento mínimo en la suma de cuadrados total

- ▶ **Método del centroide**

- ▶ La distancia entre dos clusters se define como la distancia entre los centroides (medias de los cluster)

# Dendrogramas: Otros Métodos



# No Jerárquicas: algoritmo básico

- ▶ Método de particionamiento: Construir una partición de la base de datos  $D$  de  $n$  objetos en  $k$  clusters
- ▶ Dado  $k$  encontrar una partición de  $k$  clusters que optimice el criterio de partición usado
  - ▶ Optimo Global: enumerar todas las particiones posibles
  - ▶ Métodos heurísticos:
    - ▶  $k$ -means (MacQueen'67): cada cluster esta representado por el centro del cluster
    - ▶  $k$ -medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): cada cluster está representado por uno de los objetos del cluster

# Métodos jerarquicos vs no jerarquicos

## Agrupamiento jerarquico

- ▶ No hay decisión acerca del número de clusters
- ▶ Existen problemas cuando los datos contienen un alto nivel de error
- ▶ Puede ser muy lento

## Agrupamiento no jerarquico

- Más rápido, más fiable
- Es necesario especificar el número de clusters (arbitrario)
- Es necesario establecer la semilla inicial (arbitrario)

# Ejemplos Clustering

# Agenda

- ▶ Qué es Data Mining?
- ▶ Cómo se integra en el proceso de Descubrimiento del conocimiento?
- ▶ Funcionalidades del Data Mining
- ▶ Técnicas
  - ▶ Supervisadas
    - ▶ Redes neuronales
    - ▶ Árboles
    - ▶ Regresión
  - ▶ No supervisadas
    - ▶ Clustering
    - ▶ Reglas de Asociación



# Propósito de MBA

- ▶ Generar reglas del tipo:
  - ▶ IF (SI) **condición** ENTONCES (THEN) **resultado**
- ▶ Ejemplo:
  - ▶ Si **producto B** ENTONCES **producto C**
- ▶ Association rule mining:
  - ▶ “Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.”

# Tipos de reglas según su utilidad

- ▶ **Útiles / aplicables** : reglas que contienen buena calidad de información que pueden traducirse en acciones de negocio.
- ▶ **Triviales** : reglas ya conocidas en el negocio por su frecuente ocurrencia
- ▶ **Inexplicables** : curiosidades arbitrarias sin aplicación práctica

# ¿Cuán buena es una regla?

- ▶ Medidas que califican a una regla:
  - ▶ Soporte
  - ▶ Confianza
  - ▶ Lift (Improvement)

# Ejemplo

$T1 = \{A, B, C, D\}$

$T2 = \{B, C\}$

$T3 = \{A, B, C\}$

$T4 = \{B, C, D\}$

$T5 = \{A, D\}$

$T6 = \{A, B\}$

# Soporte

- ▶ Es la cantidad (%) de transacciones en donde se encuentra la regla.
  - ▶ Ej : “Si B entonces C” está presente en 4 de 6 transacciones.
  - ▶ Soporte (B/C) : 66.6%

# Confianza

- ▶ Cantidad (%) de transacciones que contienen la regla referida a la cantidad de transacciones que contienen la cláusula condicional
  - ▶ Ej : Para el caso anterior, si B está presente en 5 transacciones (83.33%)
  - ▶ Confianza (B/C) =  $66.6/83.3 = 80\%$

# Mejora (Improvement)

- ▶ Capacidad predictiva de la regla:

- ▶ Mejora =  $p(B/C) / p(B) * p(C)$

- ▶ Ej:

- $p(B/C) = 0,67$  ;  $p(B) = 0,833$ ;  $p(C) = 0,67$

- $\text{Improv (B/C)} = 0,67(0,833*0,67) = 1.2$

- Mayor a 1 : la regla tiene valor predictivo

# Tipos de Reglas

- ▶ Booleanas o cuantitativas (de acuerdo a los valores que manejan)
  - ▶  $\text{buys}(x, \text{"SQLServer"}) \wedge \text{buys}(x, \text{"DMBook"}) \rightarrow \text{buys}(x, \text{"DBMiner"})$  [0.2%, 60%]
  - ▶  $\text{age}(x, \text{"30..39"}) \wedge \text{income}(x, \text{"42..48K"}) \rightarrow \text{buys}(x, \text{"PC"})$  [1%, 75%]
- ▶ Una dimensión o varias dimensiones
- ▶ Con manejo de jerarquías entre los elementos (taxonomías) o con elementos simples



# Ejemplos Reglas de Asociación

# Referencias

- ▶ <http://www.kdnuggets.com/>
- ▶ <http://www.acm.org/sigkdd/>
- ▶ [http://www.computer.org/portal/site/transactions/tkde/content/index.jsp?pageID=tkde\\_home](http://www.computer.org/portal/site/transactions/tkde/content/index.jsp?pageID=tkde_home)
- ▶ <http://domino.research.ibm.com/com/research.nsf/pages/r.kdd.html>
- ▶ <http://www.cs.waikato.ac.nz/~ml/weka/>
- ▶ [http://www.cs.umd.edu/users/nfa/dm\\_people\\_papers.html](http://www.cs.umd.edu/users/nfa/dm_people_papers.html)

# Preguntas



**Muchas Gracias!!!**

The background features abstract geometric shapes, primarily triangles, in shades of blue and red. These shapes overlap and intersect, creating a dynamic and modern aesthetic. The colors range from light, airy blues to deep, rich reds and dark blues.