



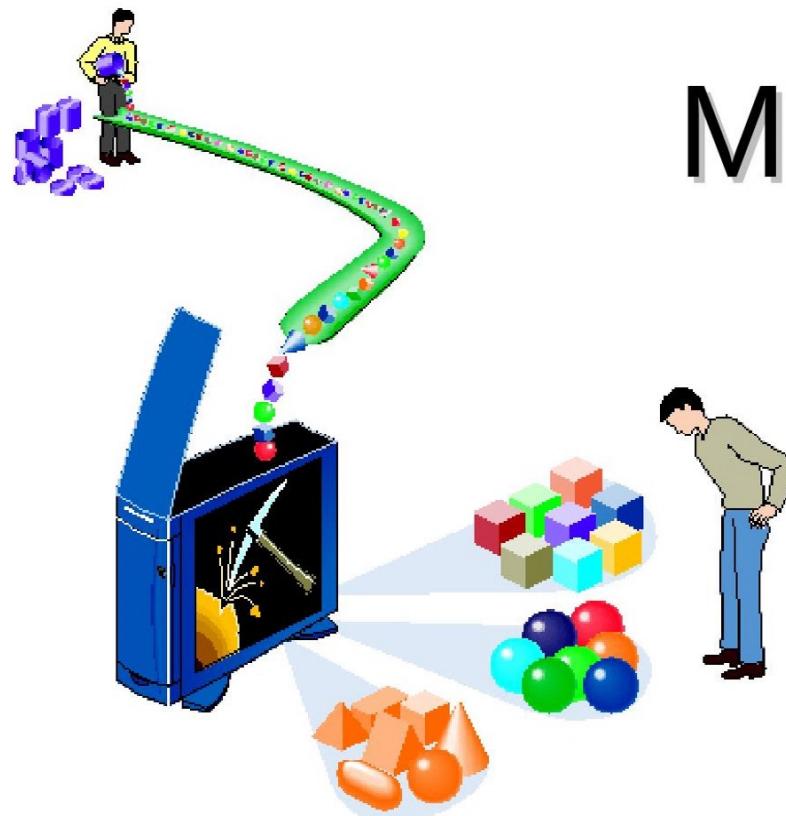
Módulo Minería de Datos Diplomado

Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS

Este documento se desarrolló a partir de otras fuentes que se encuentran citadas tanto dentro del contenido como en los espacios reservados para la bibliografía.

Si usted es autor de los documentos que se tomaron como bibliografía y considera que las referencias a su trabajo no están adecuadamente descritas, por favor comuníquese con la profesora Elizabeth León Perdomo a través del correo electrónico: eleonguz@unal.edu.co.

Introducción Minería de Datos



Agenda

- 1.** ¿Qué es la minería de datos?
Datos, Información, Conocimiento
- 2.** KDD: Knowledge Discovery Databases
Proceso de descubrir conocimiento en bases de datos
- 2.** Técnicas de Minería de Datos
- 3.** Tareas de Minería de Datos
- 4.** Aplicaciones de Minería de Datos

¿Qué es un dato?

Hecho individual acerca de algo de interés para alguien

¿Qué es información?

Datos relacionados

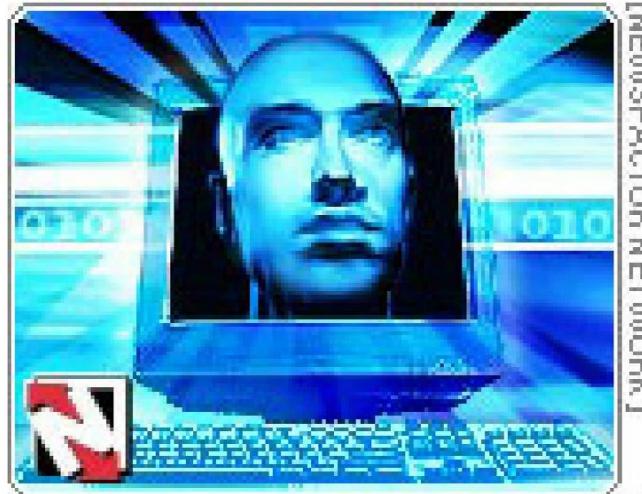
¿Qué es conocimiento?

Información co-relacionada
Patrones!

Generación de Datos

Comercial

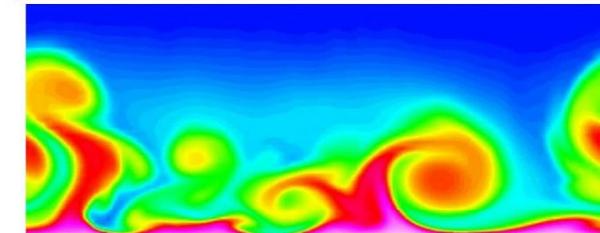
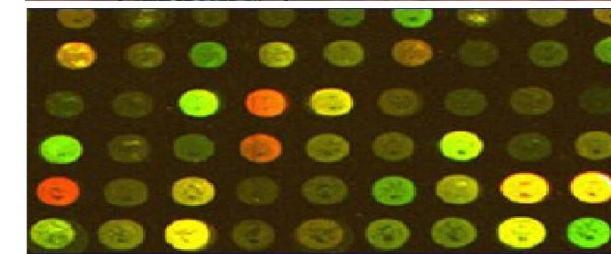
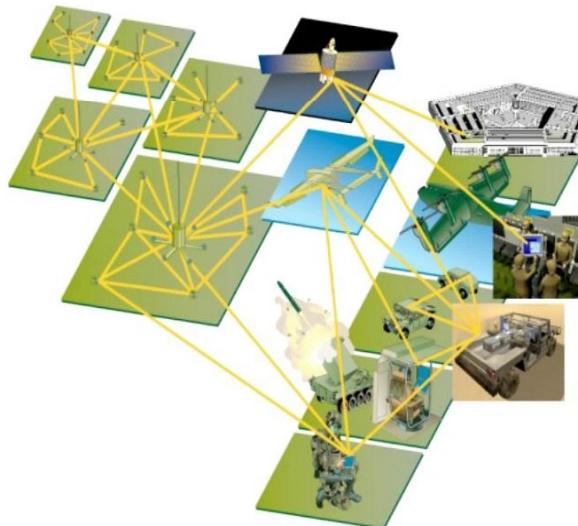
- Web (e-commerce)
- Supermercados(compras)
- Bancos (transacciones con tarjetas)



Generación de Datos

Científico

- Satélites (sensores)
- Telescopios
- Microarrays (información genética)
- Simulaciones



Datos



Datos almacenados

- Bases de datos
- Web
- Archivos (excel, pdf, txt, etc)

Información

Algo peor que no tener información disponible es tener mucha información

Y no saber qué hacer con ella.

Minería de Datos



- Grandes bases de datos contienen información no plenamente explotada

Información oculta

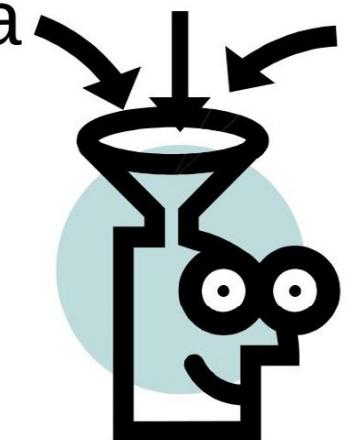
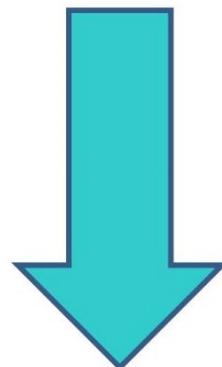
Información valiosa: “Conocimiento” oro!

- Esta información puede ser encontrada entre los datos haciendo uso de **minería en los datos**.
 - Pto comercial: competencia, servicios con el cliente
 - Pto científico: clasificar y segmentar los datos

KDD

Descubrimiento de Conocimiento en Bases de Datos

Los datos son la materia prima bruta



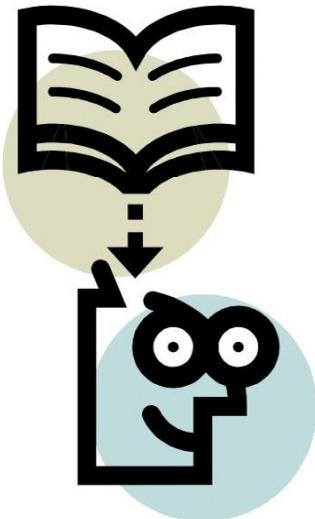
INFORMACIÓN

¿En que momento?

KDD

Descubrimiento de Conocimiento en Bases de Datos

Nos referimos al

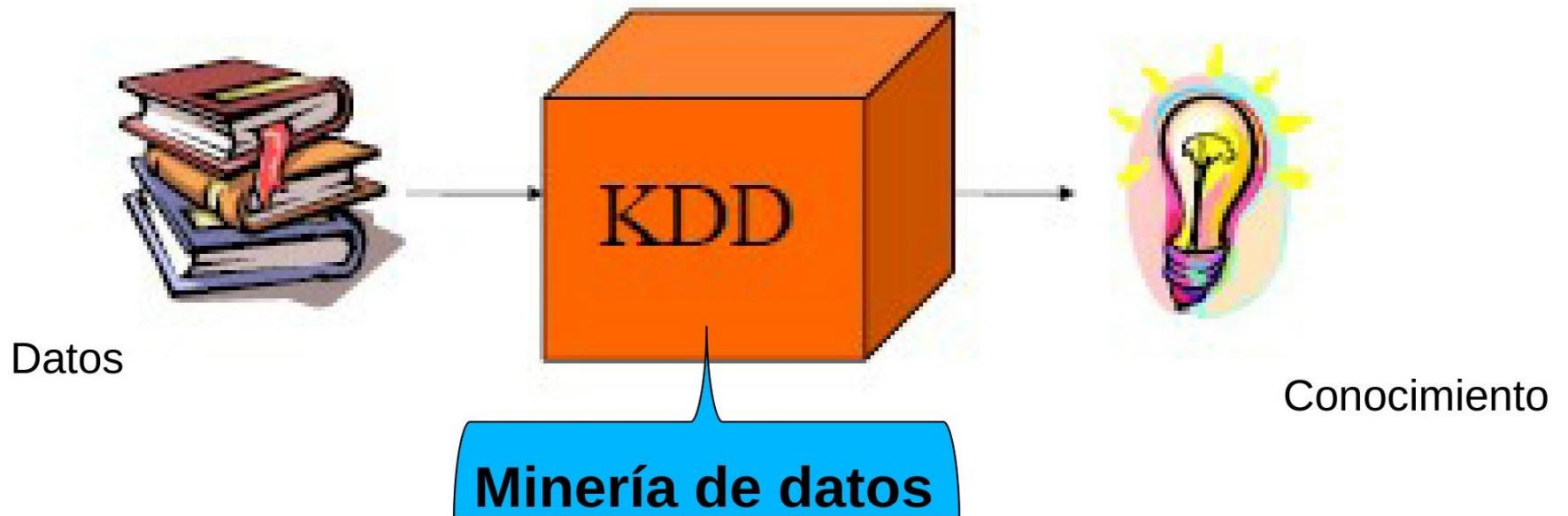


Conocimiento

KDD

Descubrimiento de Conocimiento en Bases de Datos

Proceso para descubrir información útil o conocimiento (patrones, asociaciones) desde grandes repositorios de datos.



KDD

Descubrimiento de Conocimiento en Bases de Datos



KDD

Descubrimiento de Conocimiento en Bases de Datos

El valor real de los datos reside en la información que **se puede extraer de ellos**, información que ayude a **tomar decisiones** o mejorar nuestra comprensión de los fenómenos que nos rodean



KDD

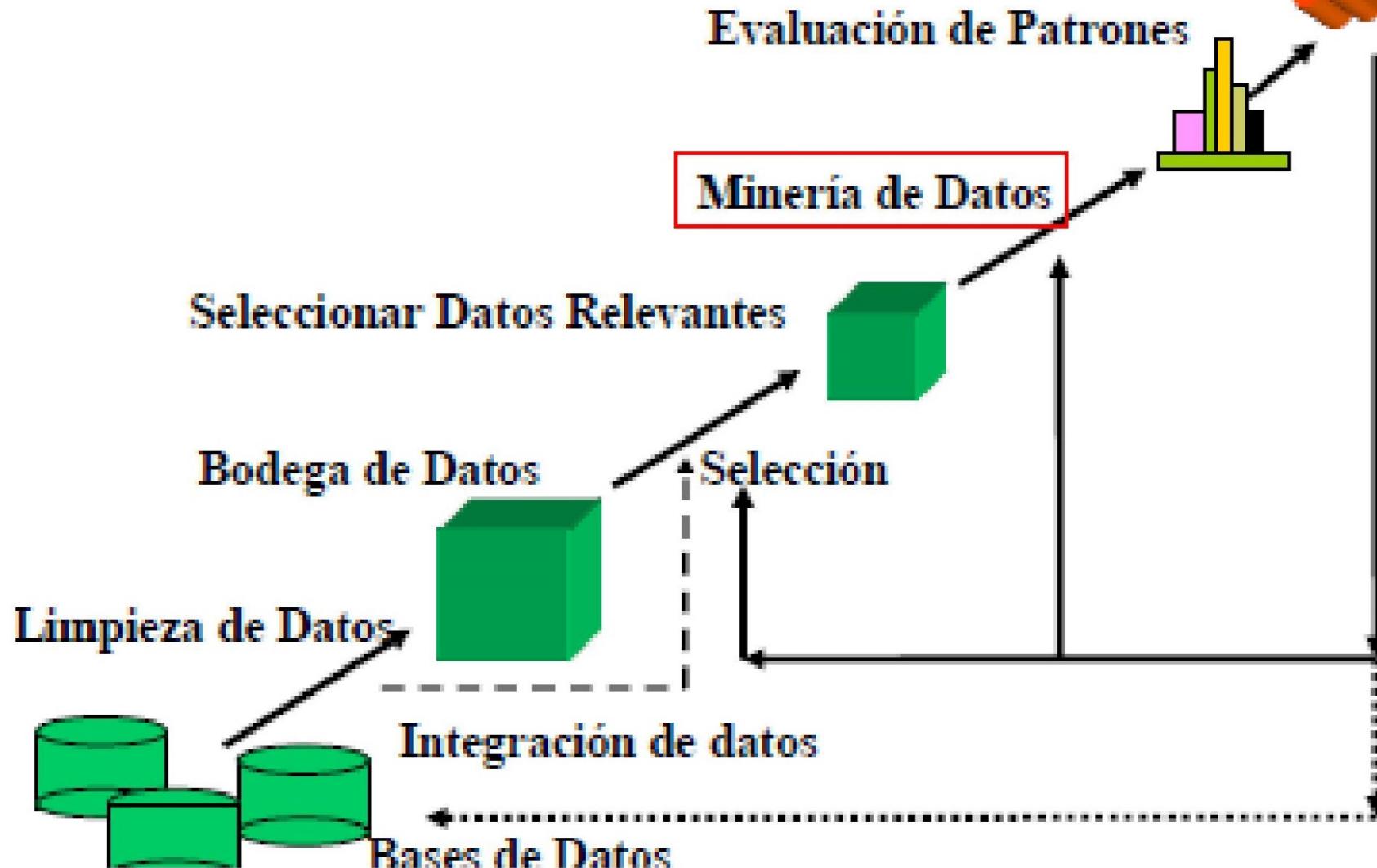
Descubrimiento de Conocimiento en Bases de Datos



- **El KDD es el Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos, teniendo como objetivo encontrar conocimiento útil relevante y nuevo sobre un fenómeno o actividad, presentando los resultados de manera visual.**

Proceso KDD

Conocimiento

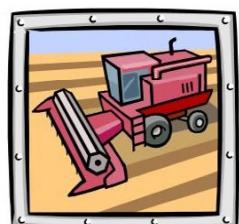


KDD

Descubrimiento de Conocimiento en Bases de Datos



1. Pre-procesamiento de Datos: Limpieza, integración y transformación.



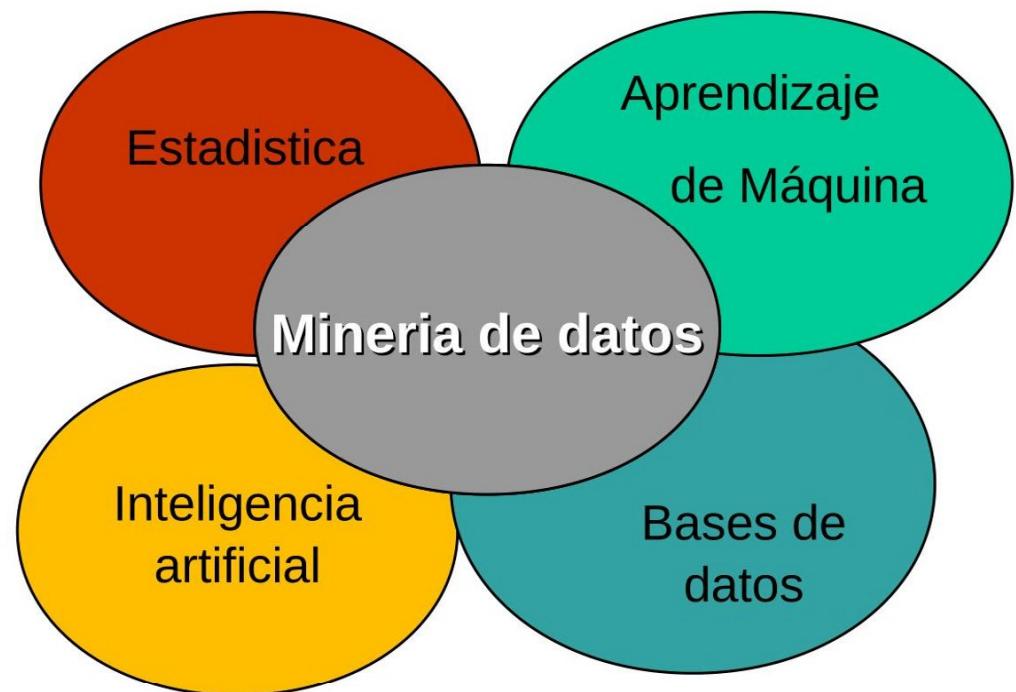
2. Minería de Datos: Uso de métodos inteligentes para extraer conocimiento (búsqueda de oro) .



3. Evaluación de patrones encontrados y presentación

Minería de datos

Paso del KDD,
Que **descubre “conocimiento”** en grandes conjunto de datos
Usa métodos como:





Minería de datos

No es...



- Buscar un número telefónico en un directorio
 - Buscar en Google
- Generar histogramas de salarios por grupos de edades diferentes



Minería de datos



es...

- Encontrar grupos de personas con similares hobbies.
- ¿Hay mas probabilidad de desarrollar cáncer si se vive cerca de una línea de poder?

Ejemplos

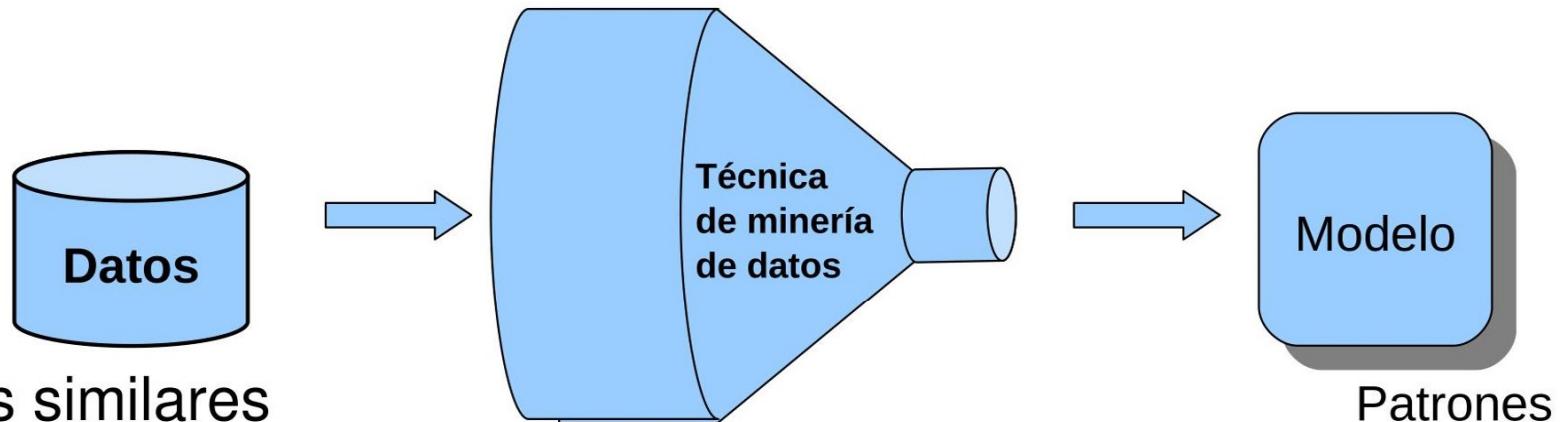
Base de datos

- Encontrar las personas que aplicaron a crédito con apellido Smith.
 - Identificar los clientes que compraron mas de \$1,000,000 en el último mes.
 - Encontrar todos los clientes que han comprado leche
- ## Minería de datos

- Encontrar las personas que aplicaron a crédito con poco riesgo de pago del crédito.
- Identificar clientes con tendencias similares de compra.
- Encontrar todos los artículos que son comprados frecuentemente con leche.

Minería de Datos

- Generar un modelo con datos



- Términos similares

“Exploratory data analysis”

“Data driven discovery”

“Deductive learning”

Algoritmo de minería de datos

Objetivo: Construir un modelo de los datos

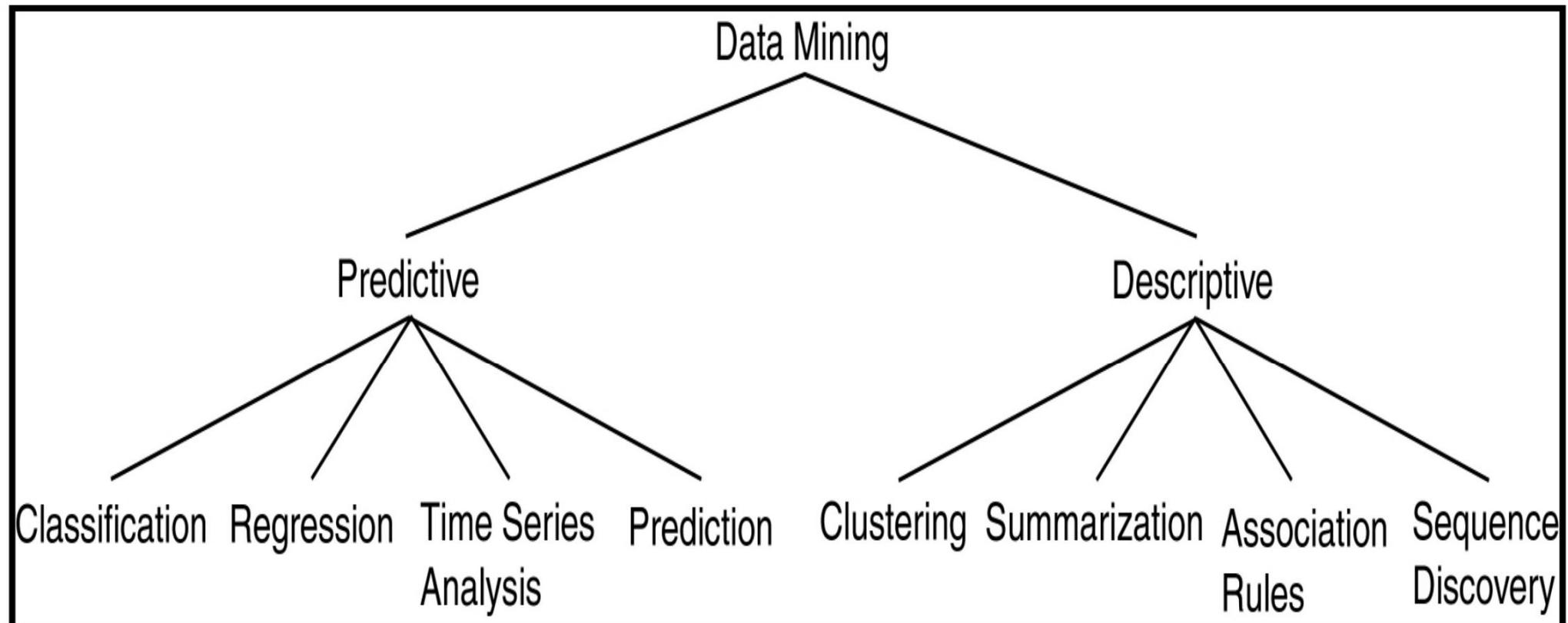
- Descriptivo
- Predictivo

Preferencia – Técnica para escoger el **mejor modelo**

Tareas de la Minería de Datos & Aplicaciones

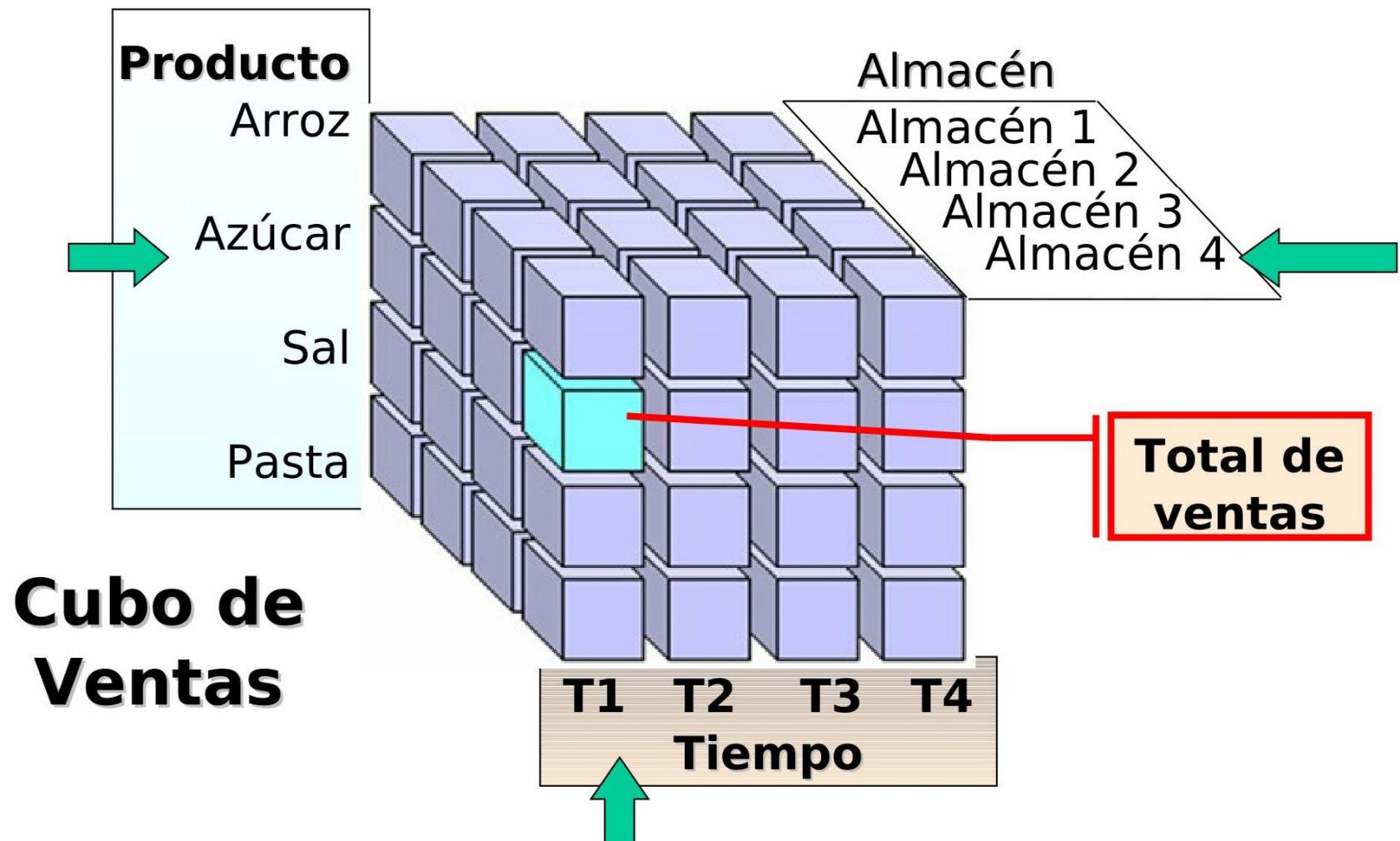


Tareas de la Minería de Datos



Sumarización

- Caracterización de la colección de datos.
- OLAP (On Line Analytical Process).
- Tendencias.
- Reportes.



¿Cuál fue el total de ventas de azúcar en el almacén 4 durante el tiempo T1?

Análisis OLAP

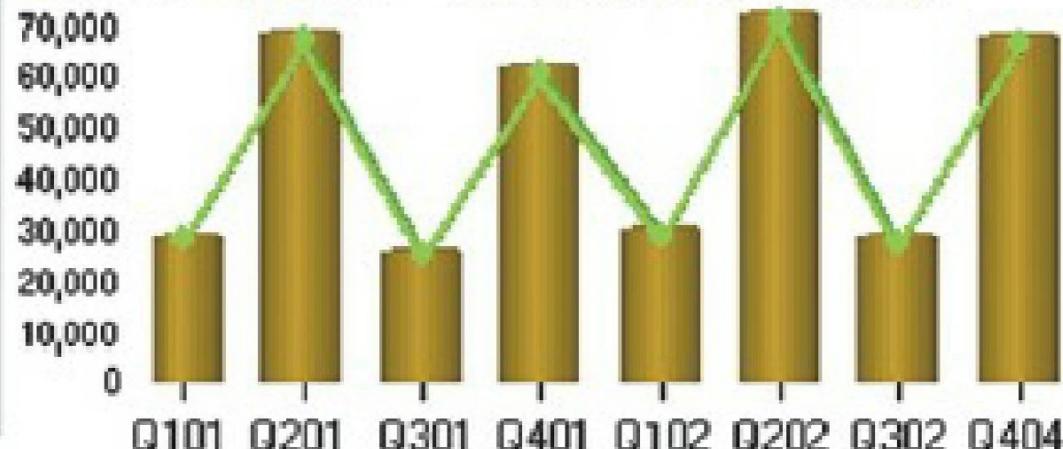
Agregaciones

Ventas**Recursos Humanos****Cuadro de mando****Mapa de ventas****Indicadores Estratégicos****Notificaciones inteligentes**

- ✗ Ventas de 2 categorías debajo de la meta
- ✗ Existen 4 marcas que no alcanzaron objetivo
- ✓ El ingreso bruto se incrementó en un 15%
- ✗ La región Norte obtuvo el menor ingreso en el periodo

Recursos Humanos

4-2002-Compañía	Empleados	Nómina	Salario Base
Acme Chemical	17	494,880	399,000
Acme Distributor	11	294,136	257,000
Acme Mining	11	323,697	279,000
Acme Paints	18	696,894	546,250
Acme Petroleum	11	264,733	232,000

Tendencia de Ventas

Asociacion

Descubrir asociaciones, relaciones / correlaciones entre un conjunto de “items”

Id	Items
1	{pan, leche}
2	{pan,pañales,cerveza,huevos}
3	{leche, pañales,cerveza, gaseosa}
4	{pan,leche,pañales,cerveza}
5	{pan,leche,pañales,gaseosa}

Ítems comprados por cliente

Interés en analizar los datos para aprender el comportamiento de las compras de sus clientes



- Promociones de mercadeo
- Manejo de inventario
- Relacion con el cliente

Encontrar Asociaciones

Asociaciones, relaciones / correlaciones en **forma de reglas**: $X \Rightarrow Y$

(registros en BD que satisfacen X, también satisfacen Y)

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule $A \Rightarrow C$: support = support($\{A \cup C\}$) = 50%

confidence = support($\{A \cup C\}$)/support($\{A\}$) = 66.6%

Support = 50% confidence = 100%

For rule $C \Rightarrow A$: ?

Asociacion

- *Supermercados (Canasta de mercado)*

Contratos de Mantenimiento (Que debe hacer el almacén para potenciar las ventas de contratos de mantenimiento)

98% de la gente que compra llantas y accesorios de autos también obtiene servicios de mantenimiento

- *Recomendaciones de paginas Web:*

URL1 & URL3 -> URL5

60% de usuarios de la Web quien visita la Pagina A y B compra el ítem T1

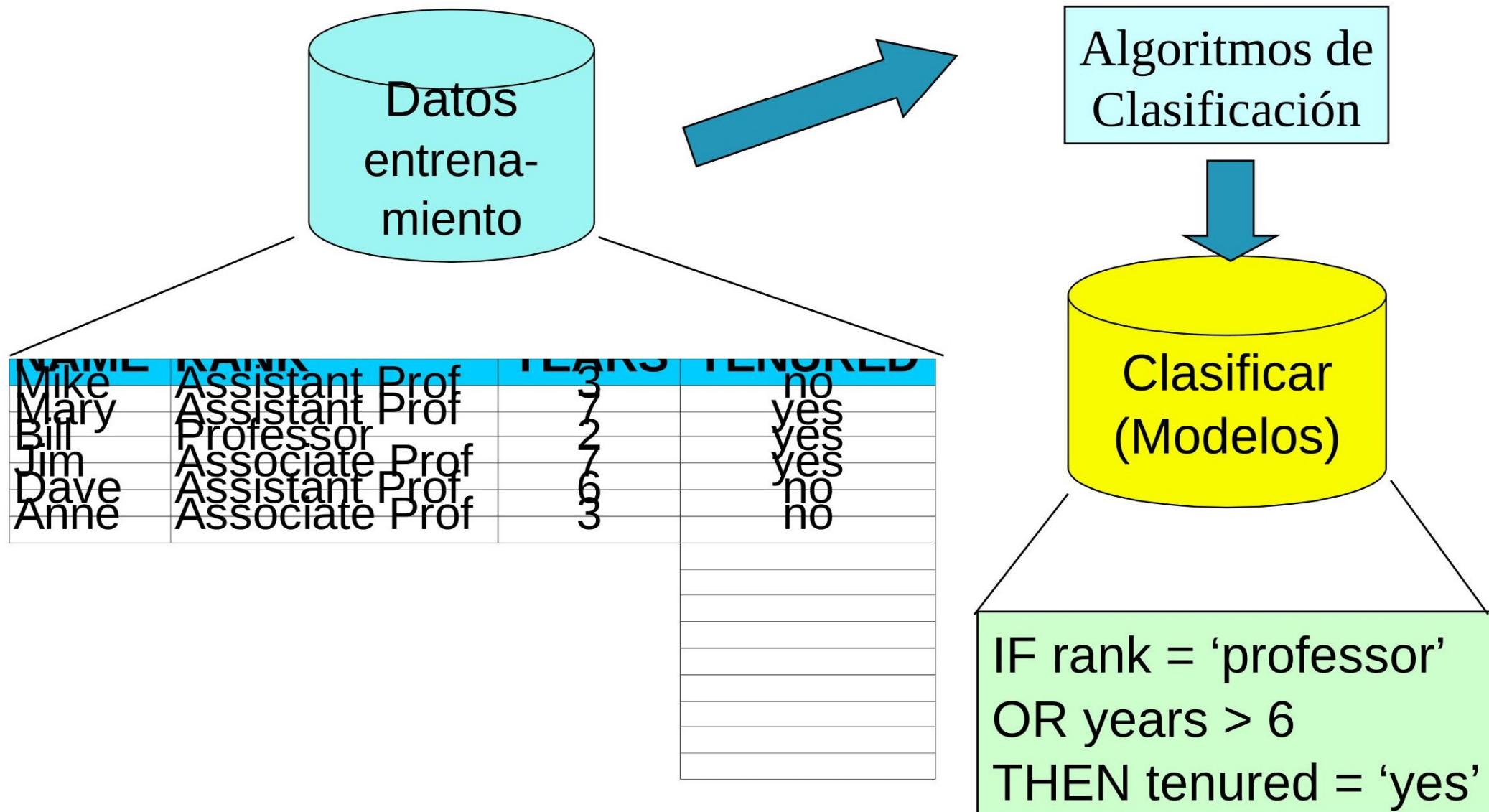
•Clasificación y Predicción



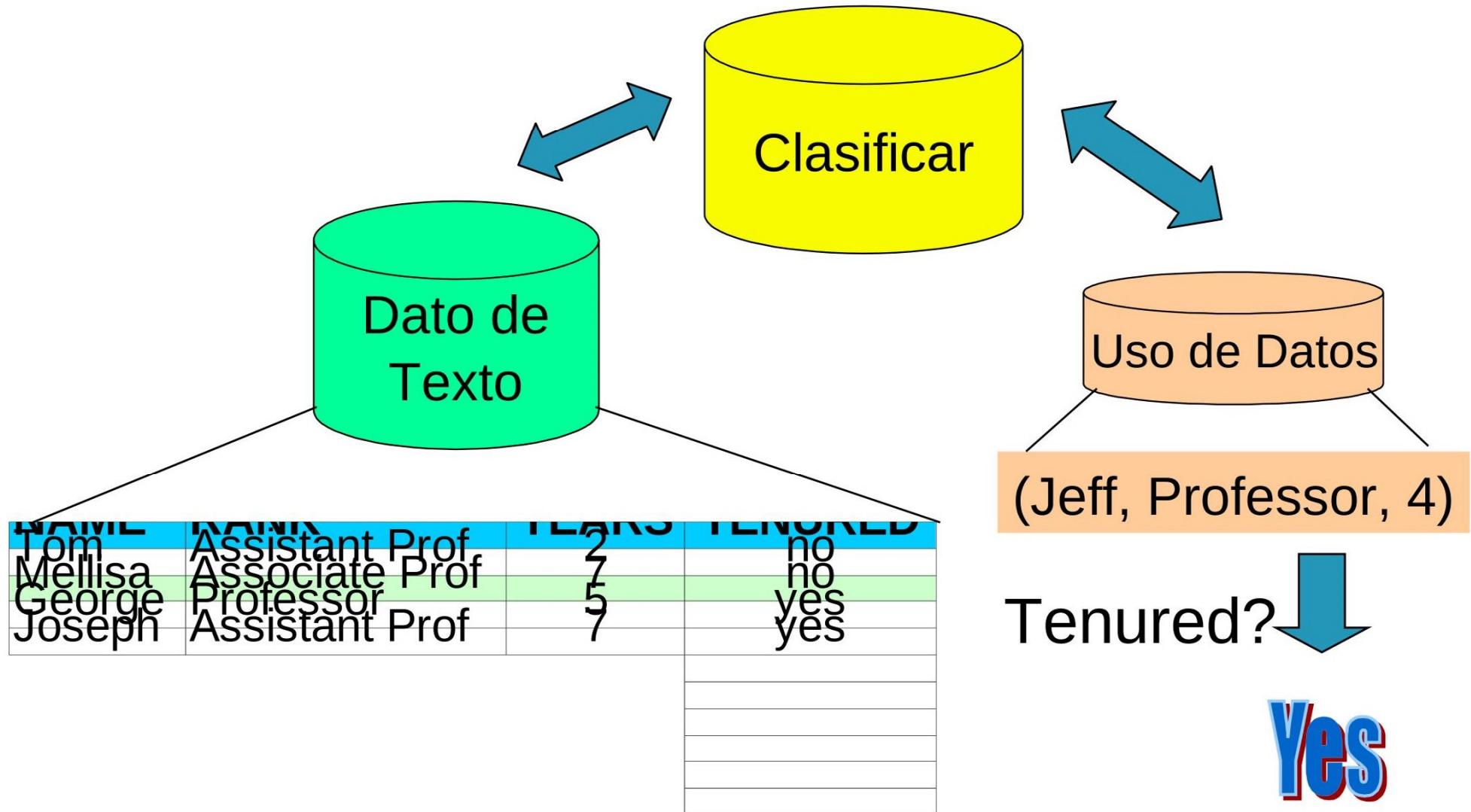
Clasificación: Construir un **modelo** por cada clase de dato etiquetado usado en el entrenamiento del modelo. Basado en sus características y usado para clasificar futuros datos

Predicción: Predecir valores posibles de datos/atributos basados en similar objetos.

Proceso de Clasificación (Paso 1): Construcción de modelo (2-clases)



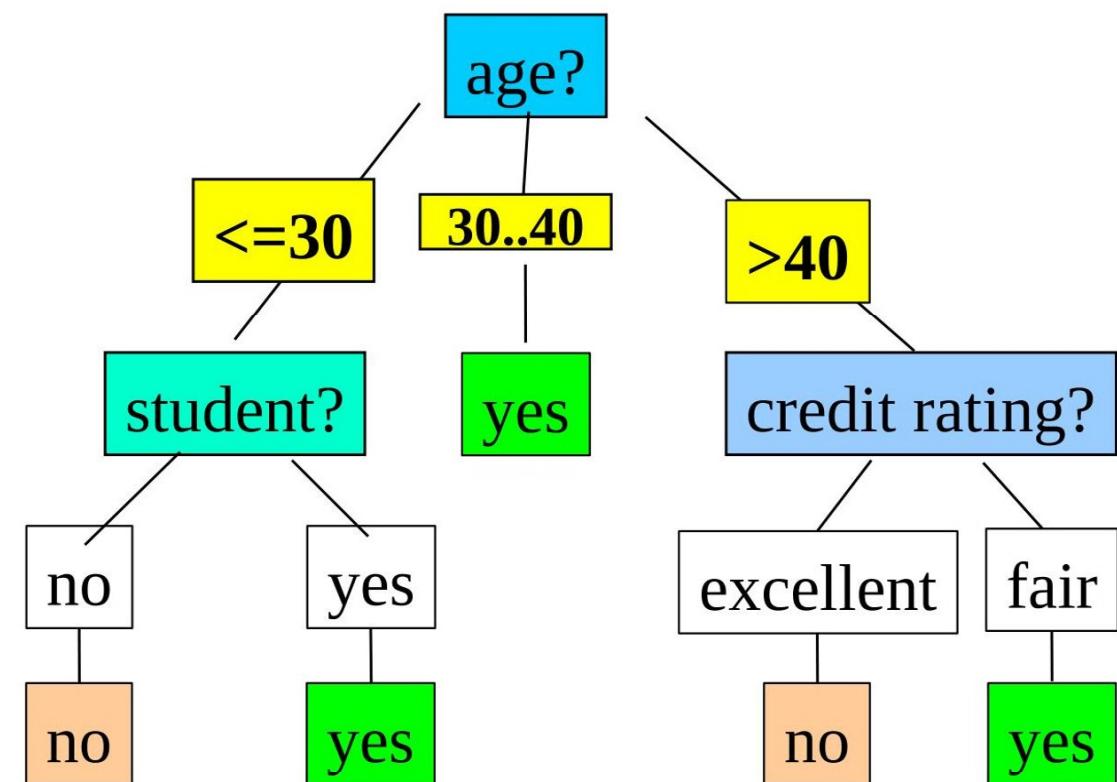
Proceso de Clasificación (Paso 2): Uso de Modelos en Predicción



Clasificación

Determinar si clientes compraran determinado producto

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
0..40	high	no	fair	yes
10..40	medium	no	fair	yes
10..40	low	yes	fair	yes
10..40	low	yes	excellent	no
10..40	medium	no	fair	yes
10..40	low	yes	fair	yes
10..40	medium	yes	fair	yes
10..40	medium	yes	excellent	yes
10..40	medium	no	excellent	yes
10..40	high	yes	fair	yes
10..40	medium	no	excellent	no

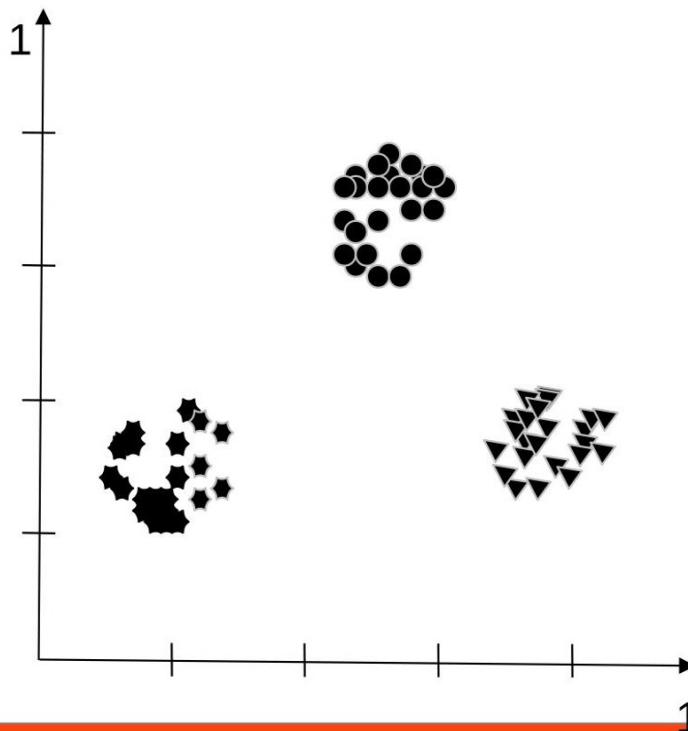


Clasificación

- Aprobación de créditos
- Diagnóstico médico
- Clasificación de documentos de texto (text mining)
- Recomendación de páginas Web automáticamente
- Seguridad

Agrupamiento

Dividir datos sin etiqueta en **grupos** (clusters) de tal forma que datos que pertenecen al mismo grupo son similares, y datos que pertenecen a diferentes grupos son diferentes

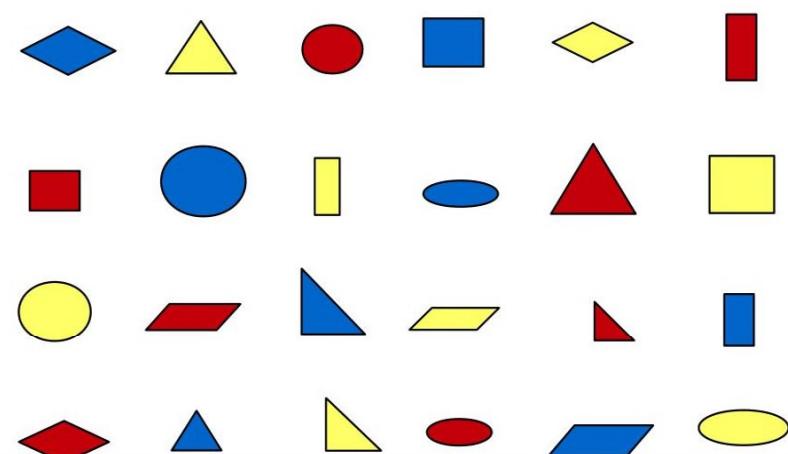


Agrupamiento

Las clases (grupos con significado) indican como las personas **analizan** y **describen** el mundo

Los humanos tienen la habilidad de dividir los objetos en grupos (**agrupamiento**) y asignar objetos particulares a esos grupos (**clasificación**)

Ej: los niños dividen objetos en fotografías: edificios, vehículos, gente, animales, plantas



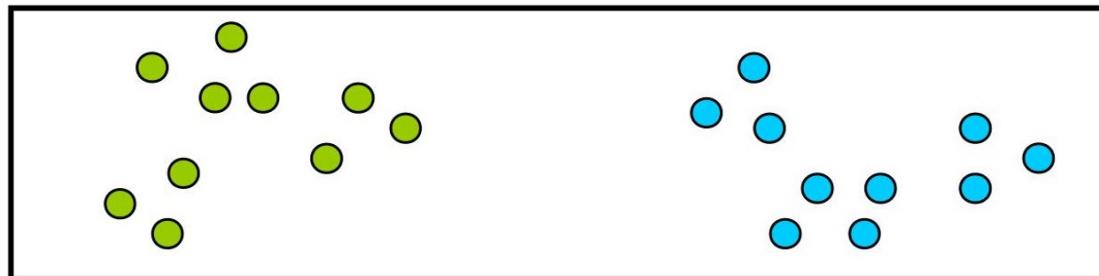
Cluster Analysis: Clustering

Cluster Análisis (clustering) es el estudio de técnicas para encontrar las clases automáticamente.

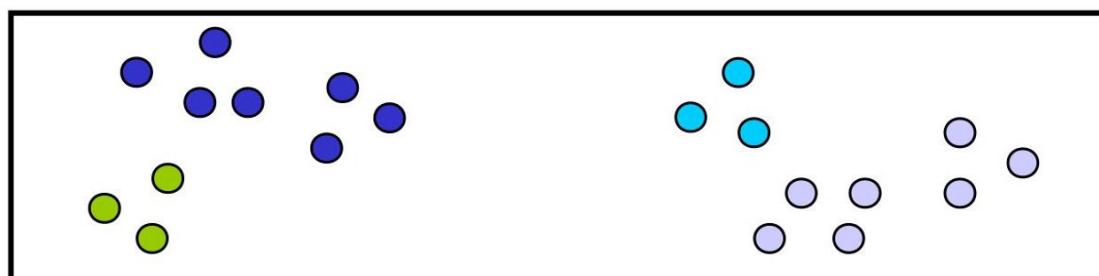
Diferentes formas de agrupar el mismo conjunto de datos



Puntos originales



Dos clusters



Cuatro clusters



Agrupamiento

Sistema visual del humano (espacio **Euclídeano**)

La arbitrariedad en el número de clusters es el mayor problema en clustering.

Grupos tienen diferentes formas, tamaños en un espacio n-dimensional

Definición de cluster es impreciso y la mejor definición depende de la naturaleza de los datos y de los resultados deseados

Clasificación NO supervisada (contraste con clasificación)

Medidas de similaridad/distancia

- La medida de **similaridad** es fundamental en la definición del cluster
- Debe ser escogida muy cuidadosamente, ya que la calidad de los resultados dependen de ella
- Se puede usar la **disimilaridad** (distancia)
- Dependen de los tipos de datos

Numéricos:

- Distancia Euclideana, Minskonski, Mahalanobis, Cosine

Binarios

- Jaccard, Cosine, Hamming

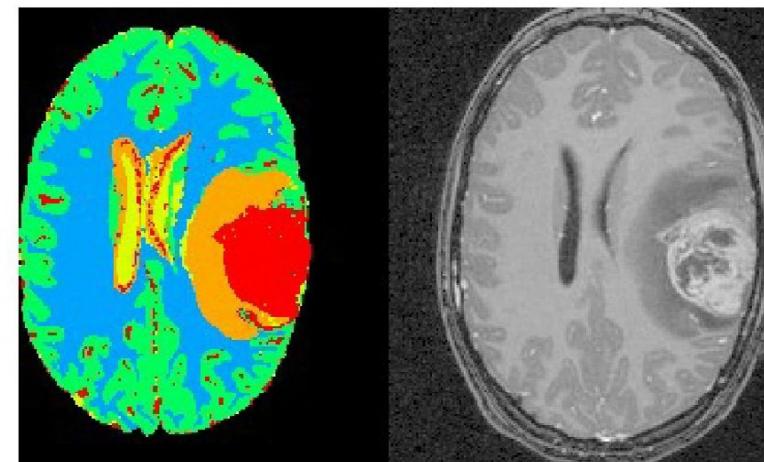
Agrupación (Aplicaciones)

Procesamiento de Imágenes (segmentar imágenes a color en regiones)

Indexamiento de texto

e imágenes

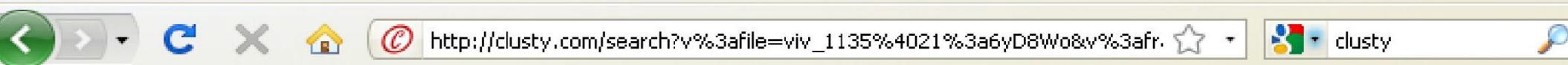
www



- Clasificación de páginas Web (usados por motores de búsqueda -Google)
- Agrupar web log para descubrir grupos de patrones de acceso similares (web usage profiles)

Agrupación (Aplicaciones)

- **Biología:** **taxonomía (especies)**, análisis de información genética(grupos de genes que tienen funciones similares)
- **Recuperación de Información (Information retrieval):**
Agrupar resultados de búsquedas en la web (cada grupo contiene aspectos particulares de la consulta) Ej: cine (comentarios, estrellas, teatros)
- **Psicología y Medicina:** Agrupar diferentes tipos de depresión, detectar patrones en la distribución temporal de una enfermedad.

[web](#) news images wikipedia blogs jobs more »

data mining

[advanced preferences](#)[clusters](#) [sources](#) [sites](#)[All Results \(148\)](#)[remix](#)[+ Software, Analysis \(34\)](#)[+ Discovery, Knowledge \(15\)](#)[- Research \(13\)](#)[• Discovery And Data Mining \(4\)](#)[• Industry, Statistical \(2\)](#)[• Research Laboratory \(2\)](#)[• Other Topics \(5\)](#)[+ Management, Data warehousing \(12\)](#)[+ Techniques \(12\)](#)

Search Results

Cluster Research contains 13 documents.

1. [PS-Explore: Data Mining, Data Analysis, Statistics](#)
PS-Explore is an advanced and easy to use statistical and database software determined for use in industry and scientific research both.
www.ps-explore.de - [cache] - Open Directory
2. [RuleQuest Research Data Mining Tools](#)
Knowledge discovery and data mining tools for Unix and Windows, including C5.0/See5 (decision trees) and Magnum Opus (association rules).
www.rulequest.com - [cache] - Open Directory
3. [Data Mining Research Laboratory @ LaTech](#)
Data Mining Research Laboratory at Louisiana Tech University has research specialization in the arena of Bioinformatics, Clinical Imaging and knowledge discovery applications in distributed and heterogeneous data domains.
dmrl.latech.edu - [cache] - Open Directory

terminado

McAfee SiteAdvisor®



EN



06:02 a.m.

Agrupación (Aplicaciones)

Seguridad: Descubriendo patrones de acceso a redes
(Detección de intrusos)

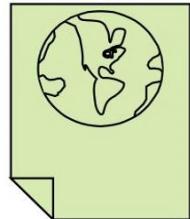
Algorithm	FA %	DR %
FAD+ with PCA	2.20	99.20
FAD+ without PCA	7.84	94.09
RIPPER-AA [11]	2.02	94.26
SMARTSIFTER [27]	-	82.0

(a)

	DOS	PRB	R2L	U2R
DR %	95.9	93.9	98.6	90.9
FA %	1.0	12.2	28.6	20.6

(b)

Aplicaciones de “Web Mining”



- Personalización automática: Sitios adaptativos facilitan navegación, búsqueda, etc.
- E-commerce: Sitios Web pueden ser construidos mas amigables
- Mercadeo Optimizado (publicidad) para negociar productos, servicios e información
- Mejores Motores de Búsqueda



Bibliografia

Introduction to Data Mining. Tan, Steinbach, Kumar. 2006

Gracias!