Business Intelligence

Clase 1:
Introducción a la Minería de Datos

Dr. Sebastián Maldonado A.

13 de Diciembre de 2011

Introducción: Idea Básica y Potencial

"We are drowning in information and starving for knowledge"

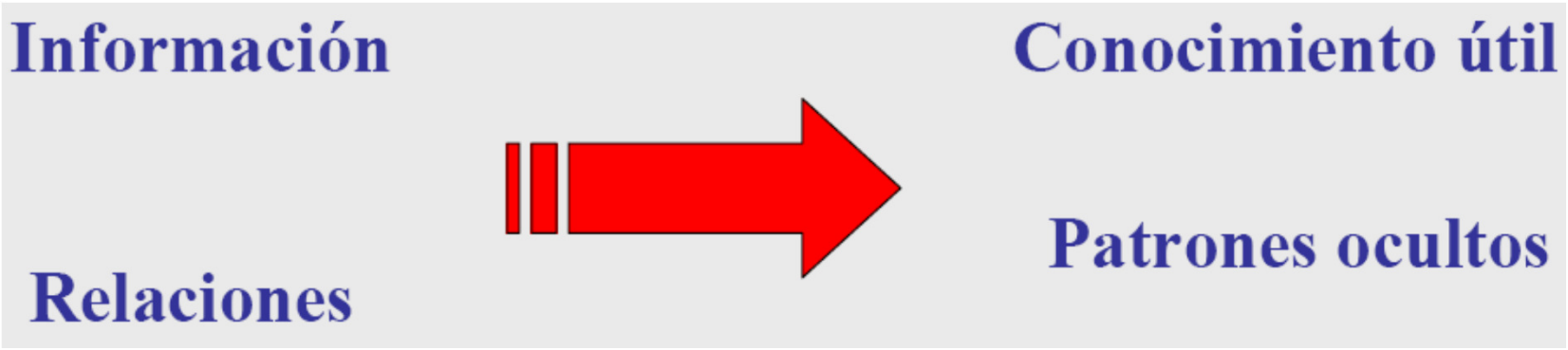
- Rutherford D. Roger

Introducción: Idea Básica y Potencial

- Empresas y Organizaciones tienen gran cantidad de datos almacenados.
- Los datos disponibles contienen información importante.
- La información está escondida en los datos.
- ***Data Mining*** puede encontrar información nueva y potencialmente útil en los datos.

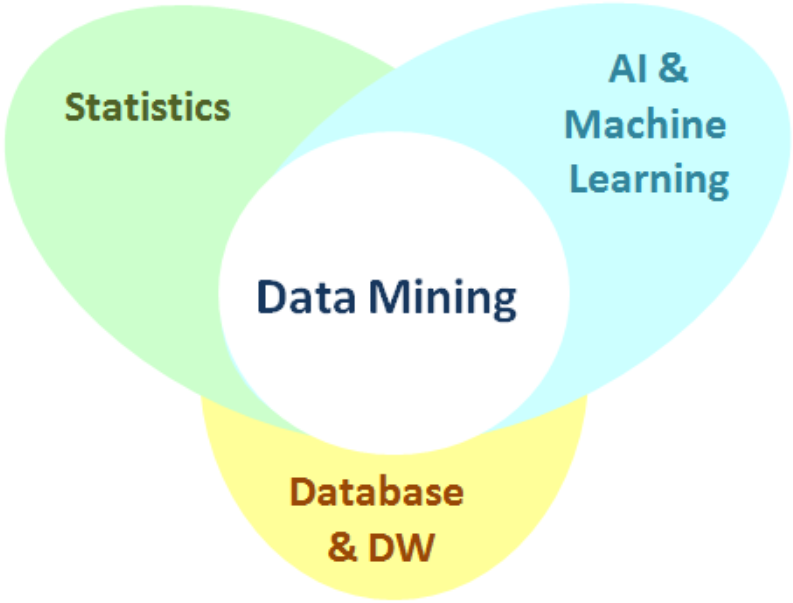
Introducción: Definición Data Mining

- "Proceso de extracción de información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información."



Introducción: Definición Data Mining

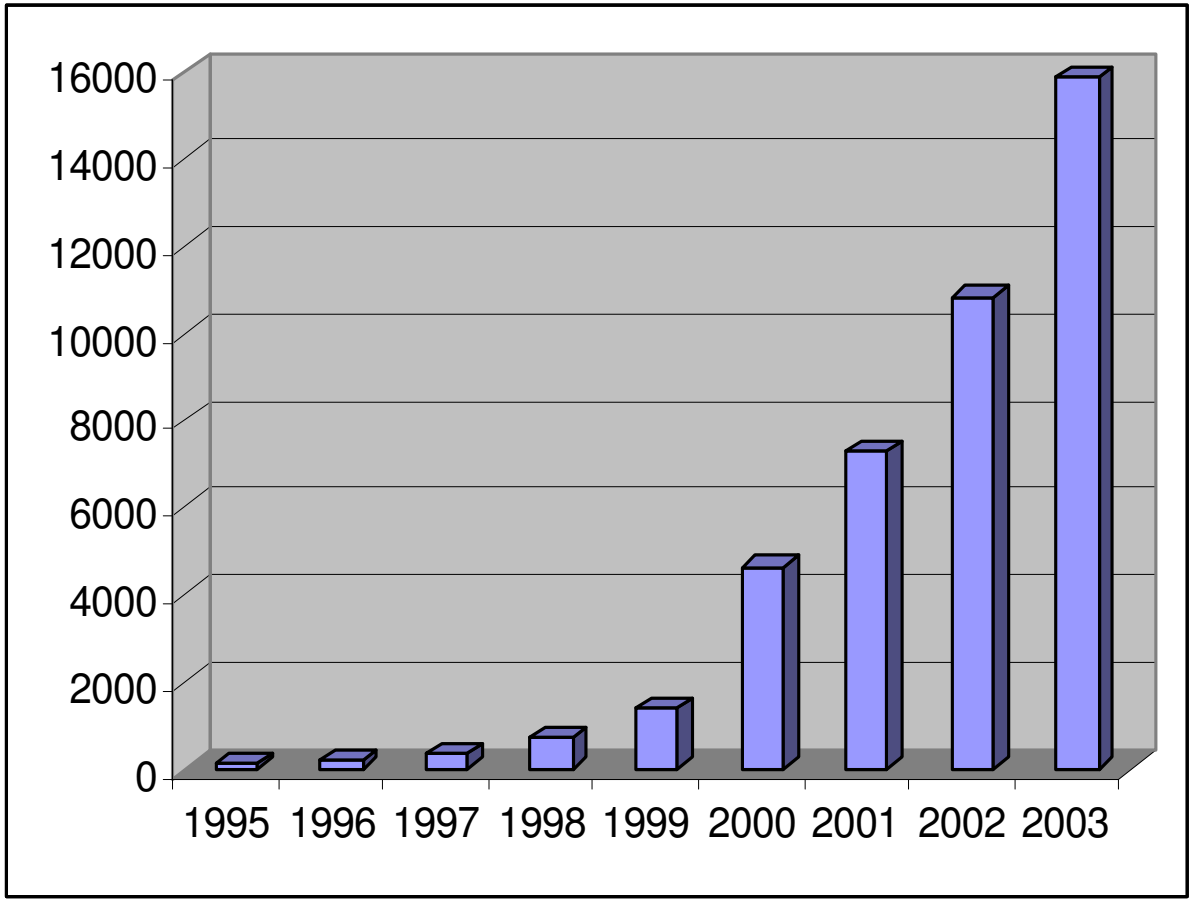
- Data Mining busca explorar el pasado y predecir el futuro mediante el análisis de datos.
- Es un campo multidisciplinario que combina estadísticas, aprendizaje computacional y tecnología de base de datos.



Introducción: Importancia Data Mining

- Las empresas de todos los tamaños necesitan aprender de sus datos para crear una relación "one-to-one" con sus clientes.
- Las empresas recogen datos de todos sus procesos.
- Los datos recogidos se tienen que analizar, comprender y convertir en información con la que se pueda actuar y aquí es donde Data Mining juega su papel.

Introducción: Relevancia Data Mining

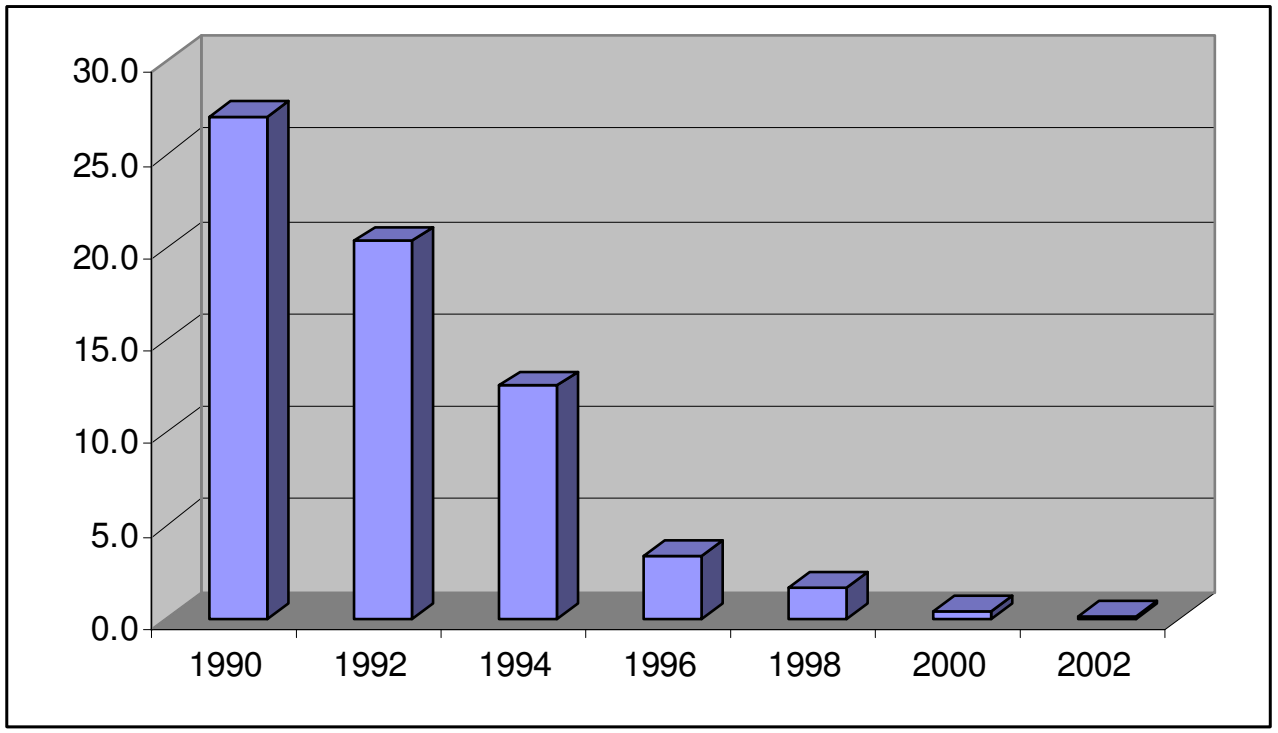


Capacidad de nuevos discos duros (PB)

Fuente:

http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Introducción: Relevancia Data Mining



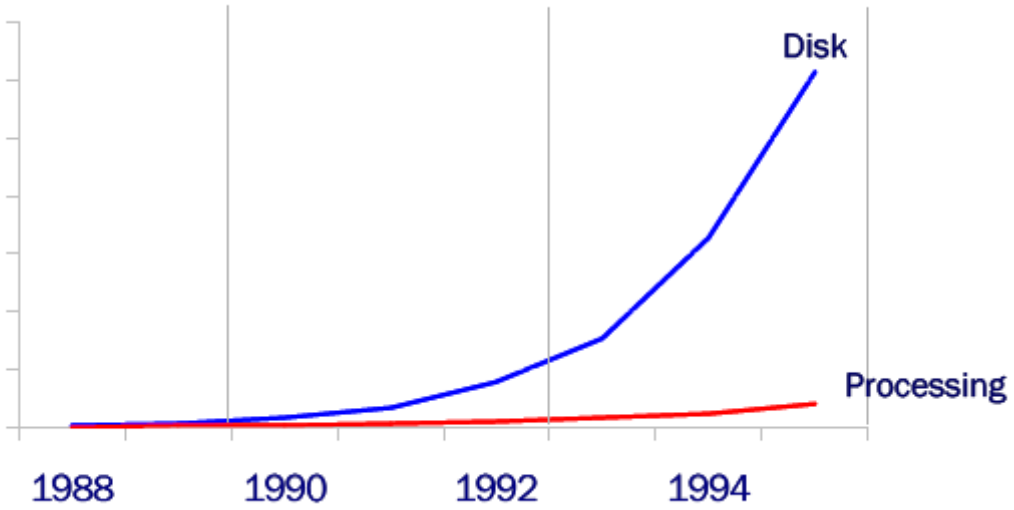
Costos de un disco duro (US-$) / Capacidad (MB)

Fuente:
http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Introducción: Almacenamiento y Procesamiento

LEY DE MOORE: "La capacidad de procesamiento se duplica cada 18 meses"

ALMACENAMIENTO: "La capacidad de almacenamiento se duplica cada 9 meses"



- La brecha entre capacidad de procesar lo que almacenamos, aumenta con el tiempo.

Introducción: Potenciales (1)

Introducción: Potenciales (2)

Evolución Data Mining

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	SPSS, Comshare, Arbor, Cognos, Microstrategy,NCR	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	SPSS/Clementine, Lockheed, IBM, SGI, SAS, NCR, Oracle, numerous startups	Prospective, proactive information delivery

Introducción: Motivación Almacenar Datos

Razones iniciales:

En telecomunicación:

Facturación de llamadas

En supermercados:

Gestión del inventario

En bancos:

Manejo de cuentas

En empresas de producción:

Control de procesos

Potenciales:

En telecomunicación:

Detección de fraude

En supermercados:

Asociación de ventas

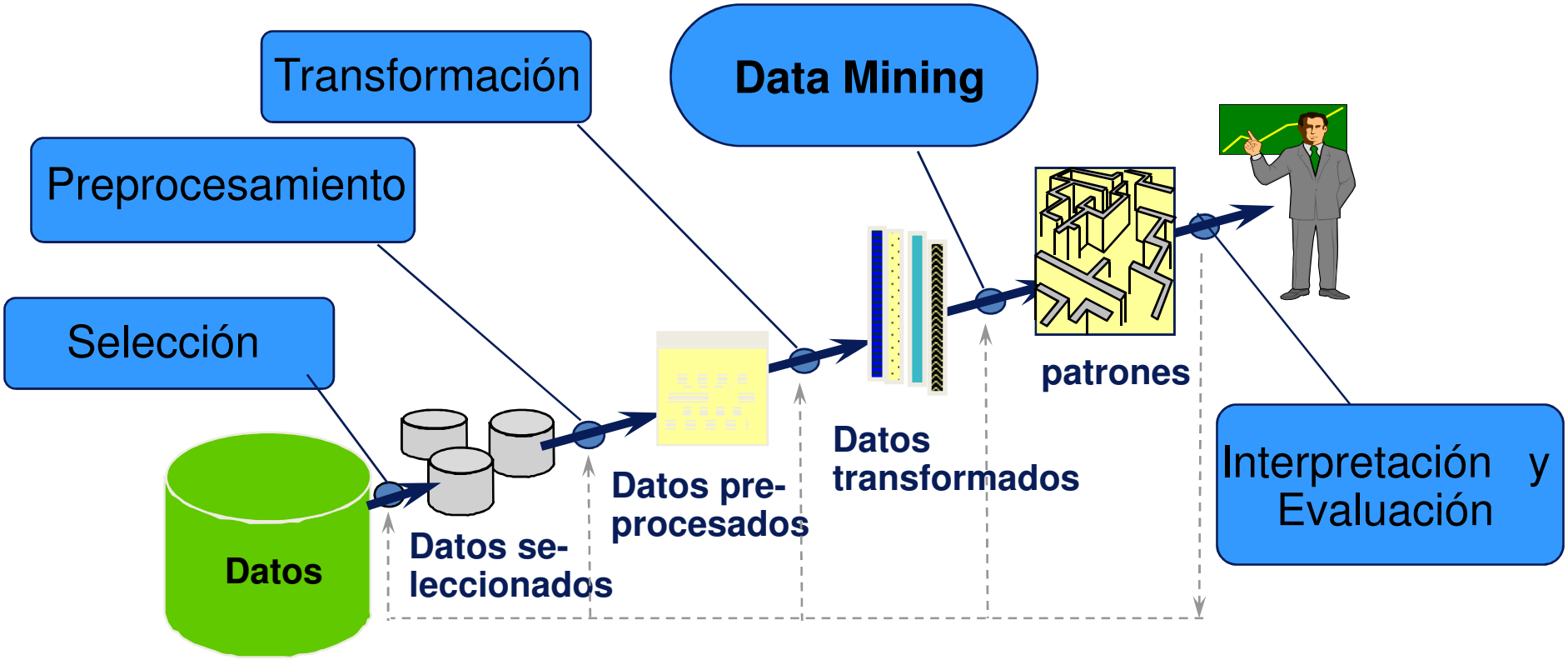
En bancos:

Segmentación de clientes

En empresas de producción:

Mantención preventiva

Introducción: Proceso KDD



"KDD es el proceso no-trivial de identificar patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos"

Proceso KDD: Paso 0

Entendimiento del negocio

- Establecer objetivos del negocio
- Establecer objetivos del minado de datos
- Establecer criterios de éxito

Entendimiento de los datos

- Explorar los datos y verificar calidad
- Establecer rangos

Preparación de los datos

Normalmente requiere el 90% de nuestro tiempo:

- Recolección
- Consolidación (niveles de agregación, unión de tablas, datos de panel)
- Limpieza
- Selección
- Muestreo
- Visualización
- Transformación
- Selección de Atributos (primer filtro)

Introducción: Áreas Data Mining

CLASIFICACIÓN

Consiste en etiquetar los objetos y crear un modelo que los clasifique bajo algún criterio.

ESTIMACIÓN O REGRESIÓN

Es la asignación de un valor ausente en un campo, en función de los demás campos presentes en el registro o de los mismos registros existentes.

SEGMENTACIÓN:

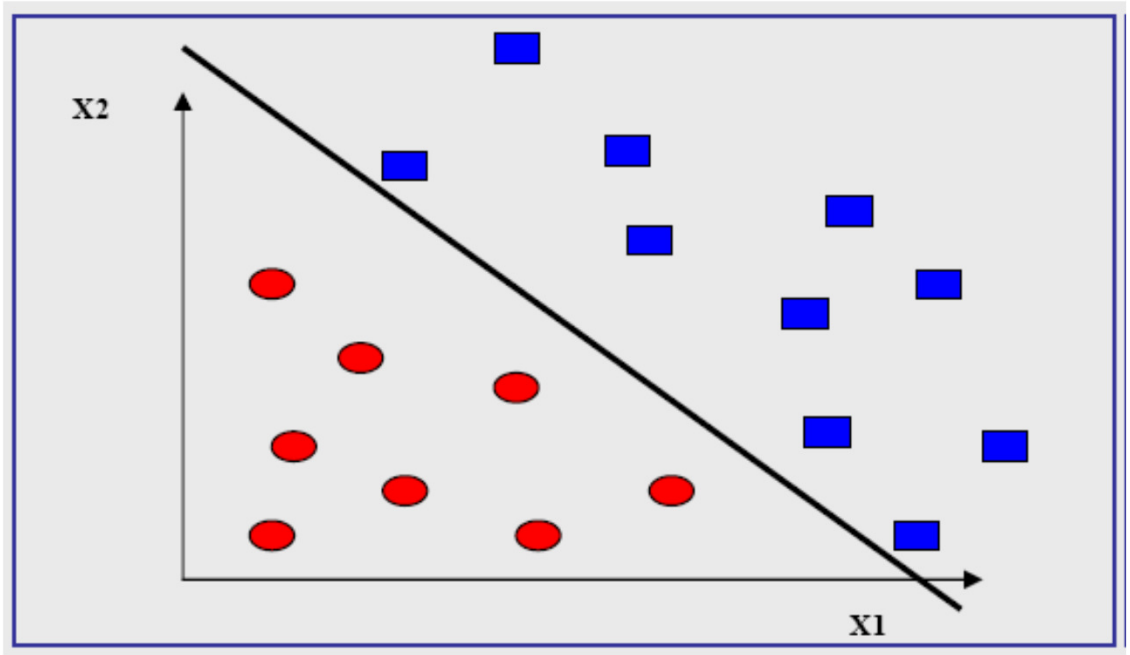
Consiste en fraccionar el conjunto de los registros (población) en subpoblaciones de comportamiento similar.

Introducción: Clasificación

- Examinar las características de un nuevo objeto y asignarlo a una clase dentro de un conjunto de clases predefinido.
  - Clasificar personas que piden créditos como alto medio o bajo riesgo
  - Determinar el patrón de las quejas de seguros fraudulentas
  - Patrón de los clientes que nos dejarán en los próximos 6 meses
- Se ha de disponer de un conjunto de entrenamiento en el que todos los registros estén clasificados.
- El problema consiste en construir un modelo que, aplicado a un nuevo ejemplo sin clasificar, lo clasifique.

Introducción: Clasificación(2)

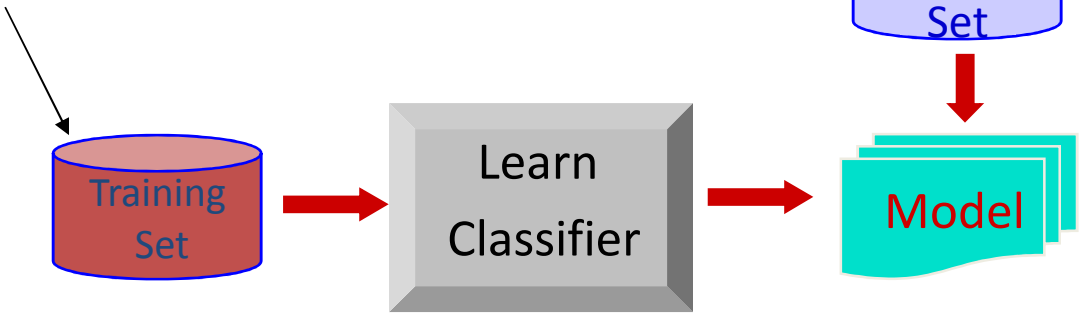
- Determinación de la pertenencia de un objeto a una cierta clase específica.
- Encontrar la mejor función que discrimine este fenómeno.
- Aplicar la función encontrada a nuevos objetos.



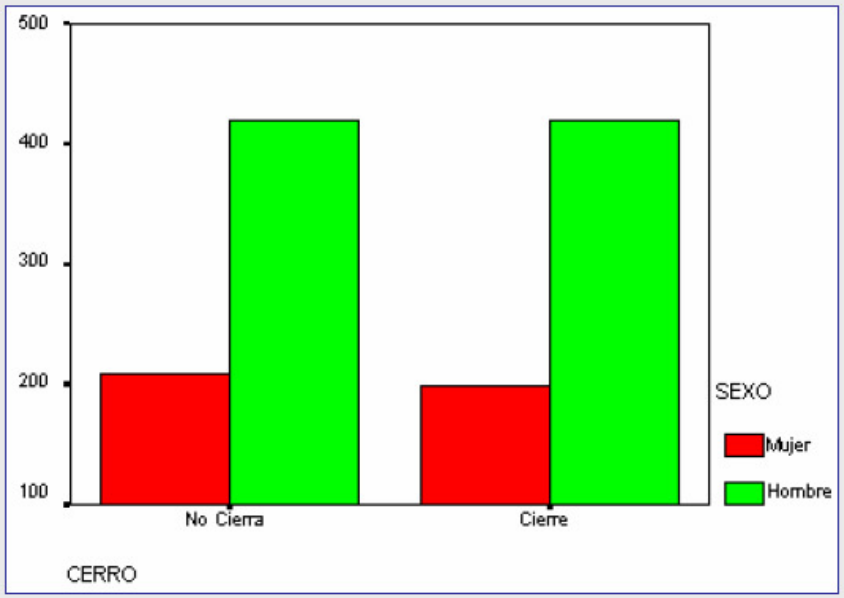
Introducción: Clasificación (Ejemplo)

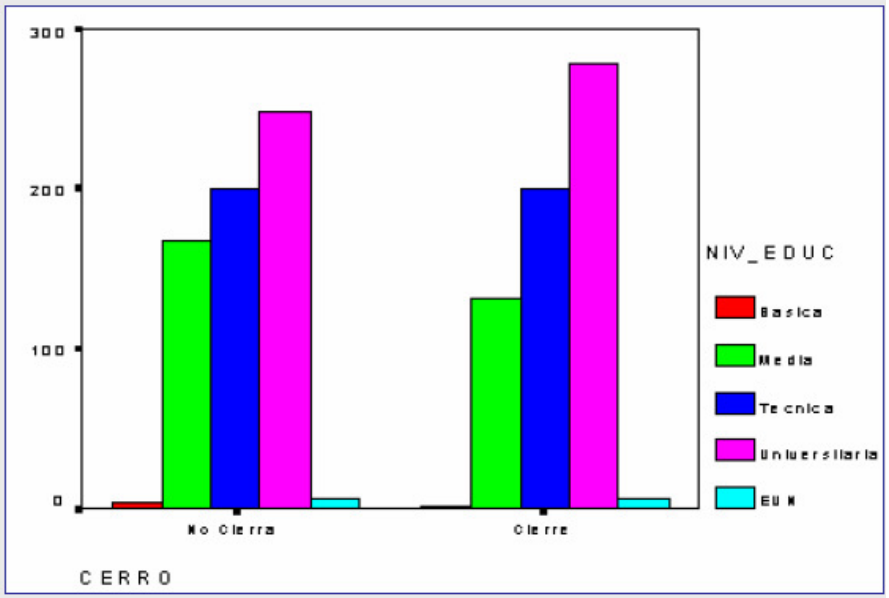
Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Introducción: Clasificación (Ejemplo)





Clasificación: Aplicaciones

Marketing Directo

- Objetivo: Reducir costos por *mailing,* enfocándose en sólo los consumidores que con mayor probabilidad comprarán un nuevo celular.
- Enfoque:
  - Usar los datos de productos similares introducidos con anterioridad.
  - Sabemos qué clientes deciden comprar y quiénes no. Esta decisión forma la variable objetivo.
  - Se recolecta información demográfica, estilos de vida e interacción con la compañía para todos los clientes.
  - Se utiliza esta información como variables de entrada para entrenar un modelo.

Clasificación: Aplicaciones

Detección de Fraude

- Objetivo: predecir casos fraudulentos en transacciones de tarjetas de crédito.
- Enfoque:
  - Usar transacciones e información de propietarios como atributos (cuándo compra, qué compra, con qué frecuencia paga a tiempo, etc.)
  - Se etiquetan las transacciones pasadas como fraudulentas o no fraudulentas.
  - Se construye un modelo a partir de esta información.

Introducción: Regresión

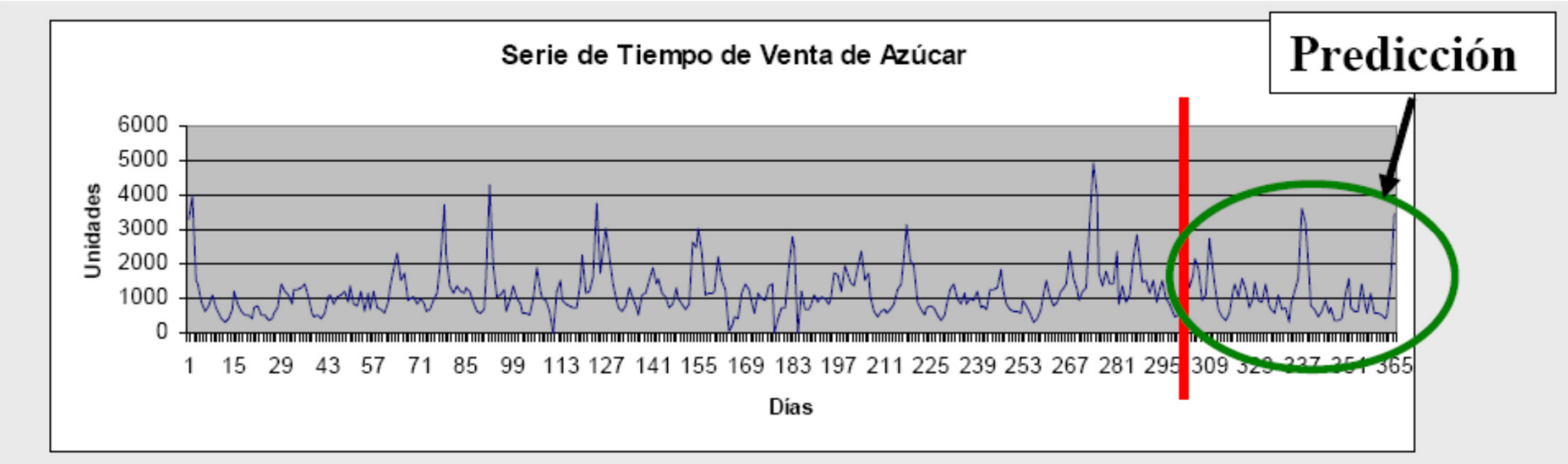
- La clasificación trata con problemas de salidas discretas (si o no, alto, medio o bajo riesgo, responderá o no responderá...ETC)

- La estimación trata con problemas donde el valor a clasificar puede tomar valores en un rango continuo (ingresos, balance de la tarjeta de crédito, probabilidad de que sea jugador)

- Ejemplos
  - Estimar el número de hijos de una familia.
  - Estimar la probabilidad de que alguien conteste a un *mailing*.
  - Estimar el tiempo de vida de un cliente.
  - Estimar los ingresos totales de una familia.

Introducción: Regresión(2)

- Estudiar el comportamiento temporal y dinámico de alguna variable.
- Encontrar la mejor función que describa este fenómeno.
- Aplicar la función encontrada a la predicción de nuevos valores de la serie.



Introducción: Reglas de Asociación

- IDEA CENTRAL: Determinar que cosas van juntas.
  - Pañales y cerveza se compran juntos los fines de semana

- El ejemplo típico es observar qué productos suelen ir juntos en la cesta de la compra.

- Se puede utilizar para establecer los almacenes, góndolas y estrategias de Cross-selling.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

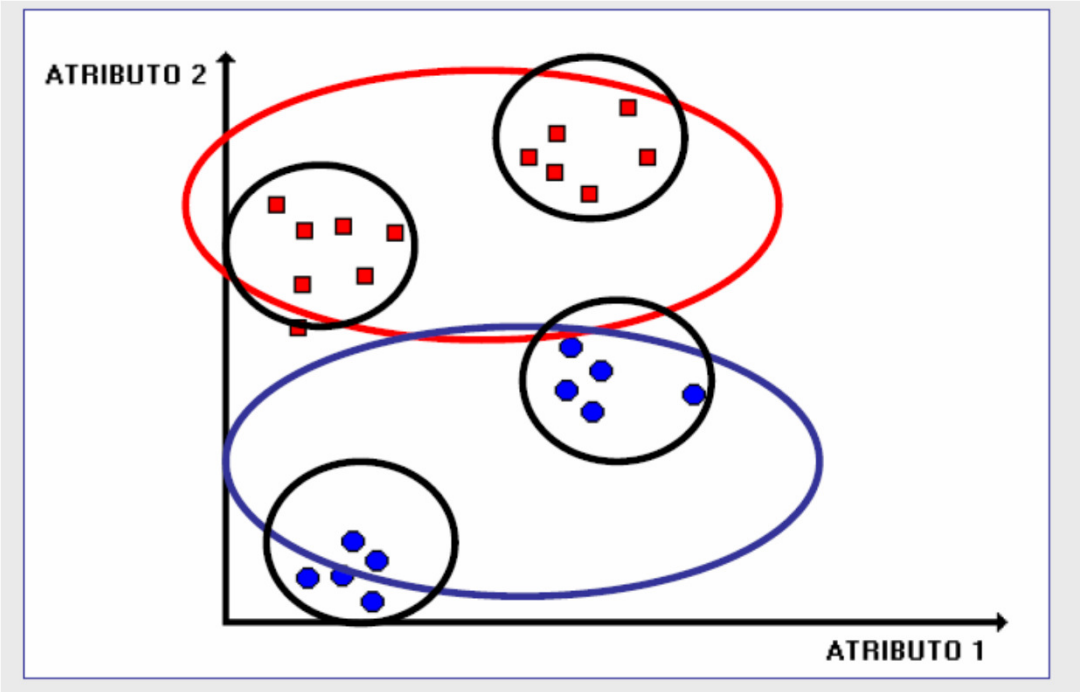
Introducción: Segmentación

- Segmentar una población heterogénea en un número de subgrupos homogéneos o clusters.
- No hay clases predefinidas
- Registros agrupados en base a su similitud.
- Se realiza a menudo antes de otras tareas de descubrimiento.
  - Encontrar clientes con hábitos de compra similares

Introducción: Segmentación(2)

- Encontrar patrones característicos no visibles a simple vista.

- Encontrar soluciones entre subconjuntos o subpoblaciones.



Evaluación de los Modelos

- Evaluando un modelo: que tan bien se comporta éste en datos de testeo.

- Medidas y criterios dependen del tipo del tipo de modelo.
  - Matriz de coincidencias para clasificación
  - Tasa de error promedio para regresión
  - ¿Segmentación?

- Interpretación del modelo: importancia depende de la aplicación, dificultad depende del algoritmo.

Implementación de los Modelos

- Determinar como los modelos deben ser utilizados. Preguntas clave:
  - Quién debe utilizar los resultados?
  - Con qué frecuencia?
  - Seguimiento

- Estrategias
  - Reglas de Negocio
  - *Scorecards*
  - *Scoring* interactivo online

Aplicaciones Exitosas: Detección de Fraude

Credicard Brazil, S.A.:

- Sistema para detección de fraude usando redes neuronales
- Reducción de fraude por 40% en un año
- Observación de transacciones de 4.5 millones de tarjetas de Credicard en tiempo real.

Aplicaciones Exitosas: Cross-Selling

Banco HSBC (USA):

- Dada la fuerte competencia en la industria financiera, la importancia que tiene la retención de clientes es mayor que nunca.
- HSBC buscaba incentivar clientes a mejorar sus productos o a adquirir nuevos.
- Se reducen los costos por mailing en un 30%, capturando el 95% de las utilidades de la campaña.
- Menos correo basura se traduce en clientes más leales.

Programa del Curso

- Clase 1 (13/12): Introducción, Teoría Bases de Datos.
- Clase 2 (20/12): Consolidación de Bases de Datos.
  - ¿Cómo partir con un proyecto de BI?
- Clase 3 (21/12): Limpieza/Transformación/ Selección de Atributos.
- Clase 4 (27/12): Métodos de Clasificación (1) .
- Clase 5 (03/01): Métodos de Clasificación (2), Validación. Tarea
- Clase 6 (10/01): Métodos de Segmentación.
- Clase 7 (17/01): Métodos de Regresión, Examen.

*Entrega Tarea: Fines de enero.

Evaluaciones

- Una evaluación personal (examen) y una tarea computacional grupal.
- Grupos de tres integrantes. La tarea será entregada al final de la clase 5, con plazo para fines de enero.
- Software: Rapid Miner 5 (Open Source)
- Notas:
  - Nota Examen: 60% Nota Final.
  - Nota Tarea: 40% Nota Final.

Teoría de Bases de Datos

Bases de Datos Relacionales

- Nacieron en los 70´S creadas por Edgar F. Codd.
- Están basadas en relaciones (tablas) como estructura de almacenamiento, con atributos o campos (columnas) y una serie de tuplas o registros (filas).
- Estandarizaron el lenguaje de manipulación, usando SQL, creado por IBM en los 80´S.

Bases de Datos Relacionales: Ejemplo

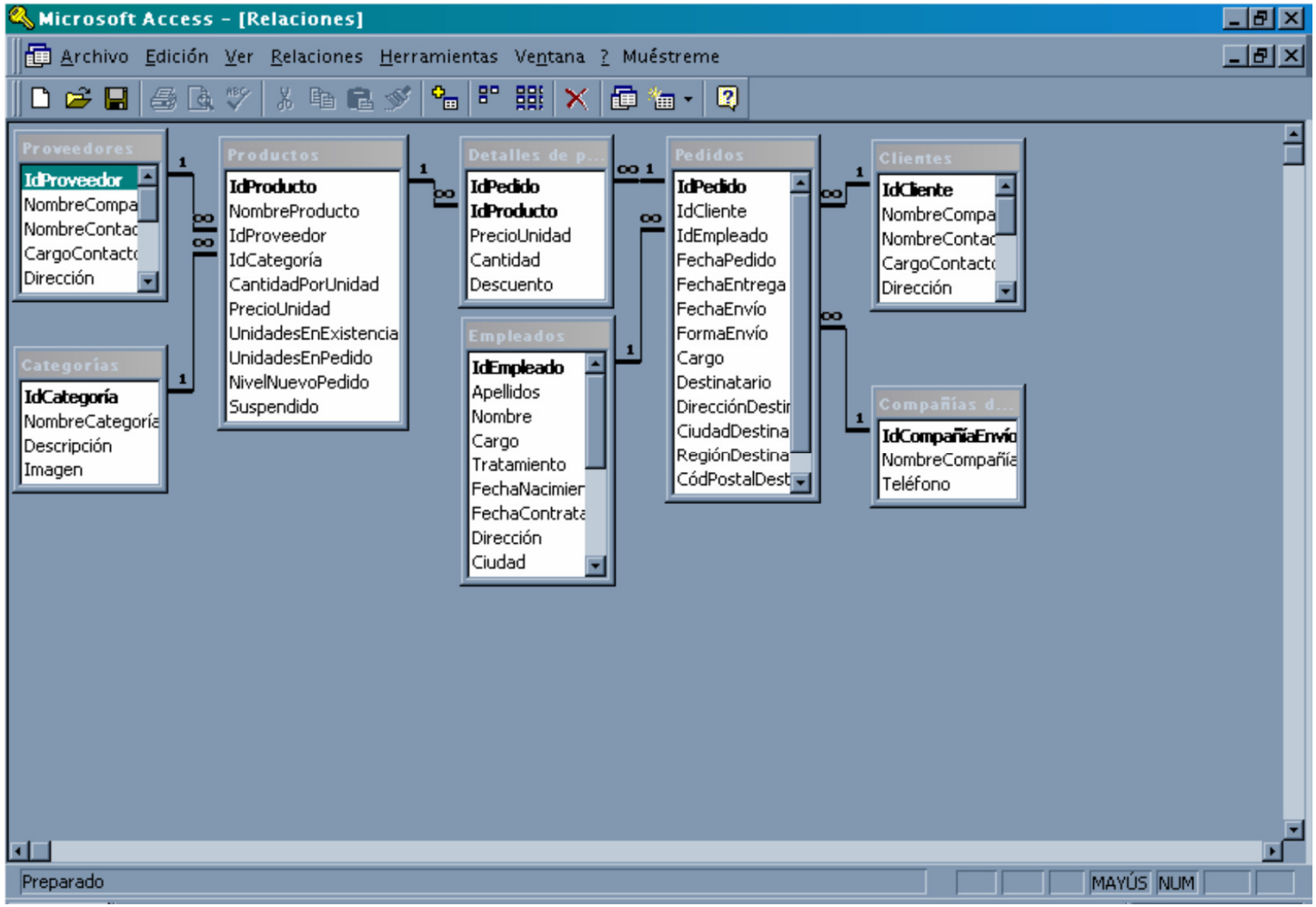
Relación Empleado

EMPLEADO	NPILA	APPAT	APMAT	RUT	FNAC	DIRECCION	SEXO	SUELDO	RUTSUPERV	NDEPTO
	Juan	Perez	Martinez	13.463.530-4	12-01-78	Av.Matta 223	M	120.000	123654	5
	Alicia	Rubio	Jara	15.356.345-8	25-06-65	Alameda 123	F	190.000	852647	4
	Sebastian	Carrasco	Claro	10.254.269-7	18-12-50	San Diego 654	M	250.000	843601	1

Relación Departamento

DEPARTAMENTO	DNOMBRE	DNUMERO	RUTGERENTE	GERFECHAINIC
	Of. Central	1	88866555	19-06-71
	Administración	4	98765432	01-01-85
	Investigación	5	33344555	22-05-78

Bases de Datos Relacionales: Ejemplo



Data Warehouse

- Reúne datos esenciales provenientes de bases de datos heterogéneas desde todas las áreas de negocio (Ventas, finanzas, RRHH, etc.)
  - Una base de datos para apoyar decisiones (DS-DB) que es mantenida separadamente de la BD transaccional de la empresa.
  - Procesamiento de información de soporte mediante una plataforma sólida, de datos históricos y consolidados listos para ser analizados.
- Organiza los datos para apoyar decisiones de gestión.
- Maneja elevados volúmenes de información.
- Permite el mejor funcionamiento de los métodos de Data Mining.
- Data Warehousing: el proceso para construir DW

Data Warehouse

- DATAWAREHOUSE: Colección de objetos
  - Orientada al objetivo:
    - Organizada en torno a los datos más importantes de la empresa.
    - Es bueno para realizar filtros y eliminar información poco importante.
    - El modelamiento se enfoca en el análisis y toma de decisiones basadas en estos datos particulares y no en el procesamiento diario de las transacciones.
    - Provee una vista simple y concisa a cerca de los datos de interés, siendo capaz de verlos desde distintos puntos de vista o dimensiones. A la vez se filtra todo dato que no aporta a la toma de decisiones.

Data Warehouse

- DATAWAREHOUSE: Colección de objetos
  - Unificada:
    - Basada en unión de información de varias fuentes.
    - Asegura la consistencia de la información.
  - Variante en el tiempo
    - Guarda información a través del tiempo.
    - Posee actualizaciones temporales agregadas: no hay actualizaciones diarias.

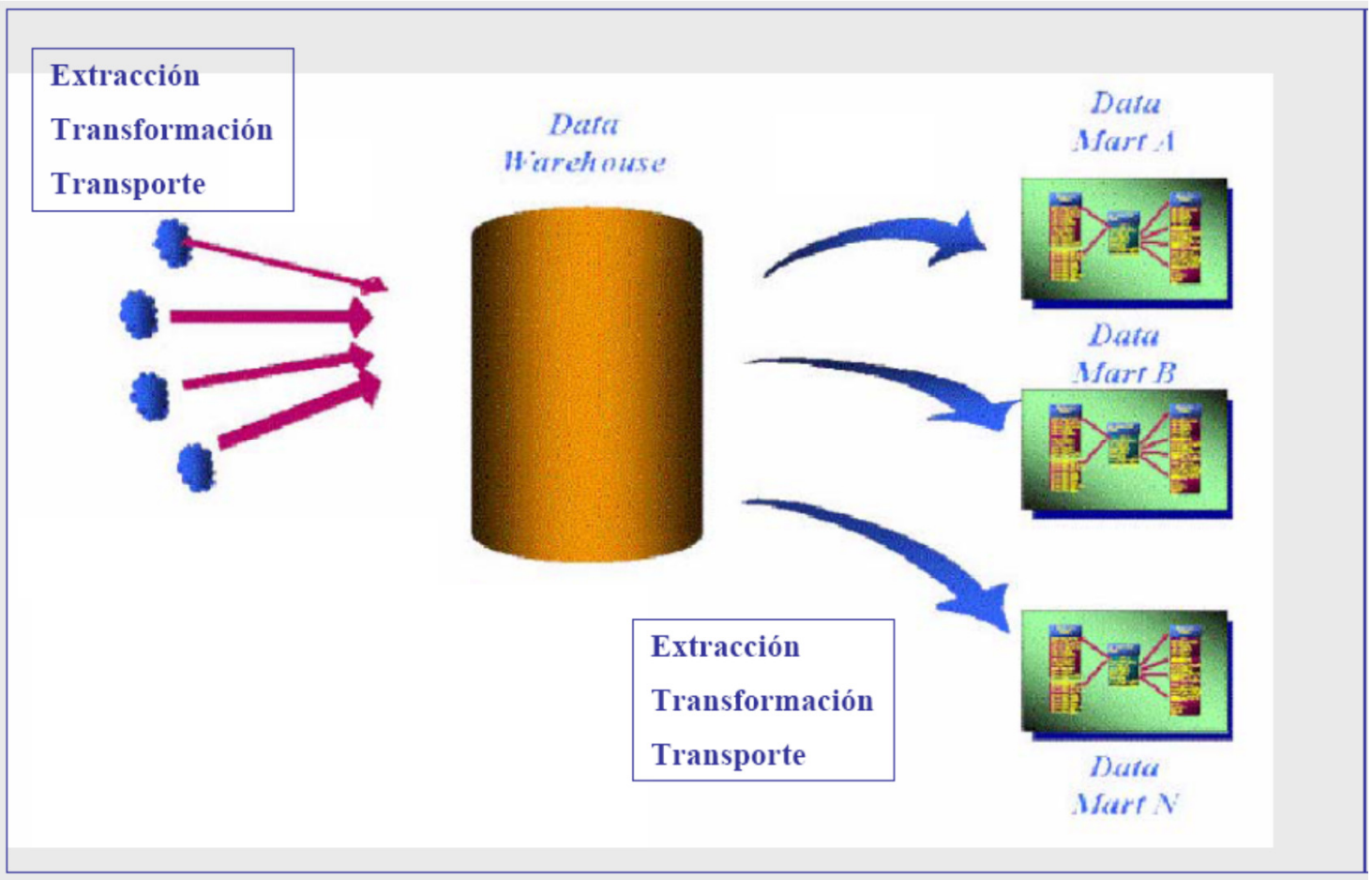
Diferencias Data Warehouse y Base de Datos

Características	Bases de Datos	Data Warehouse
Volumen	alto	bajo o medio
Tiempo de respuesta	muy rápido	normal
Frecuencia de actualizaciones	alta, permanentemente	baja
Nivel de los datos	en detalle	agregado

Data Mart: Forma simple de un Data Warehouse

Data Mart	Data Warehouse
Nivel de Departamentos (Finanzas, Ventas, ...)	Nivel corporativo
pocas fuentes	muchas fuentes
<100 GB	100GB - TB+

Data Mart y Data Warehouse



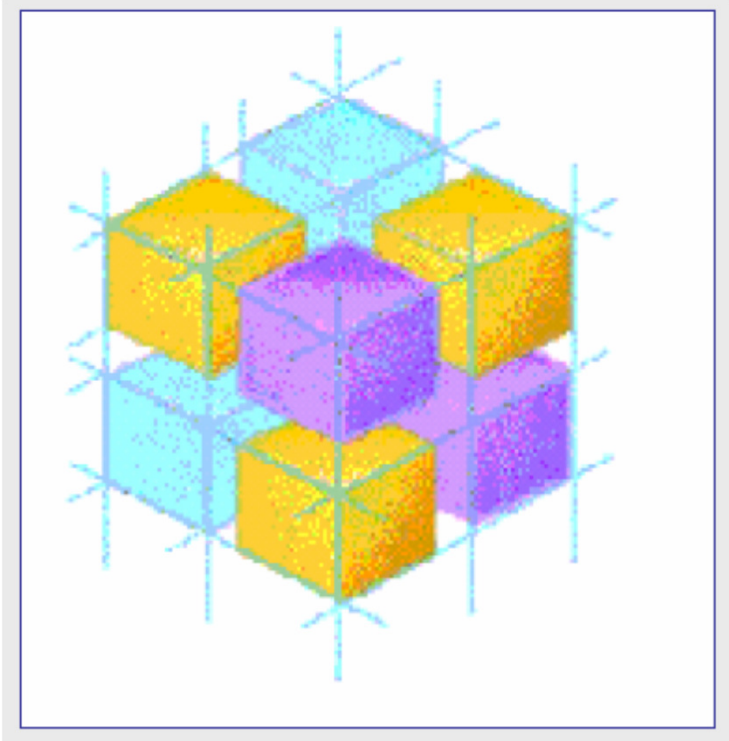
Cubos Multidimensionales

- Consiste en una representación multidimensional de datos de detalle y resumen.

- Tiene como objetivo mejorar el rendimiento empresarial en línea y mejorar el rendimiento de las consultas.

Cubos Multidimensionales

- Son un subconjunto de datos de la base de datos original.
- Son capaces de administrar de forma rápida y eficiente grandes cantidades de información.

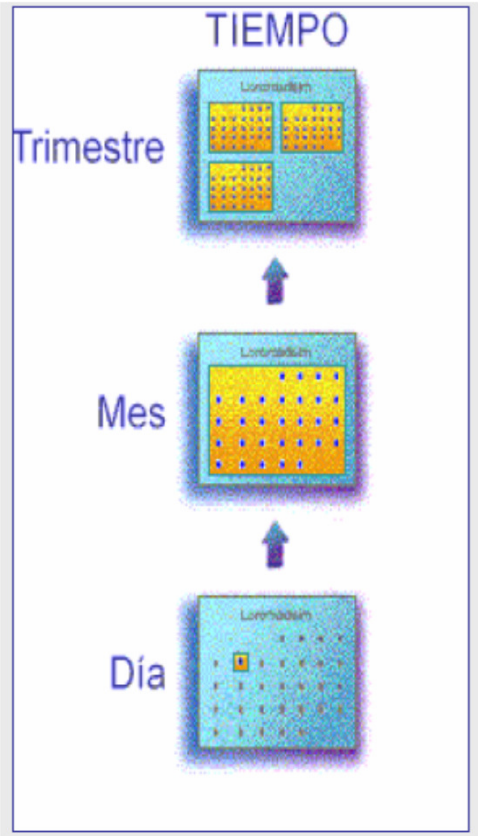


Cubos Multidimensionales : Componentes

- ORIGEN DE LOS DATOS
  - Identifica y conecta donde se encuentra el almacén de datos la información relevante para resolver un problema.

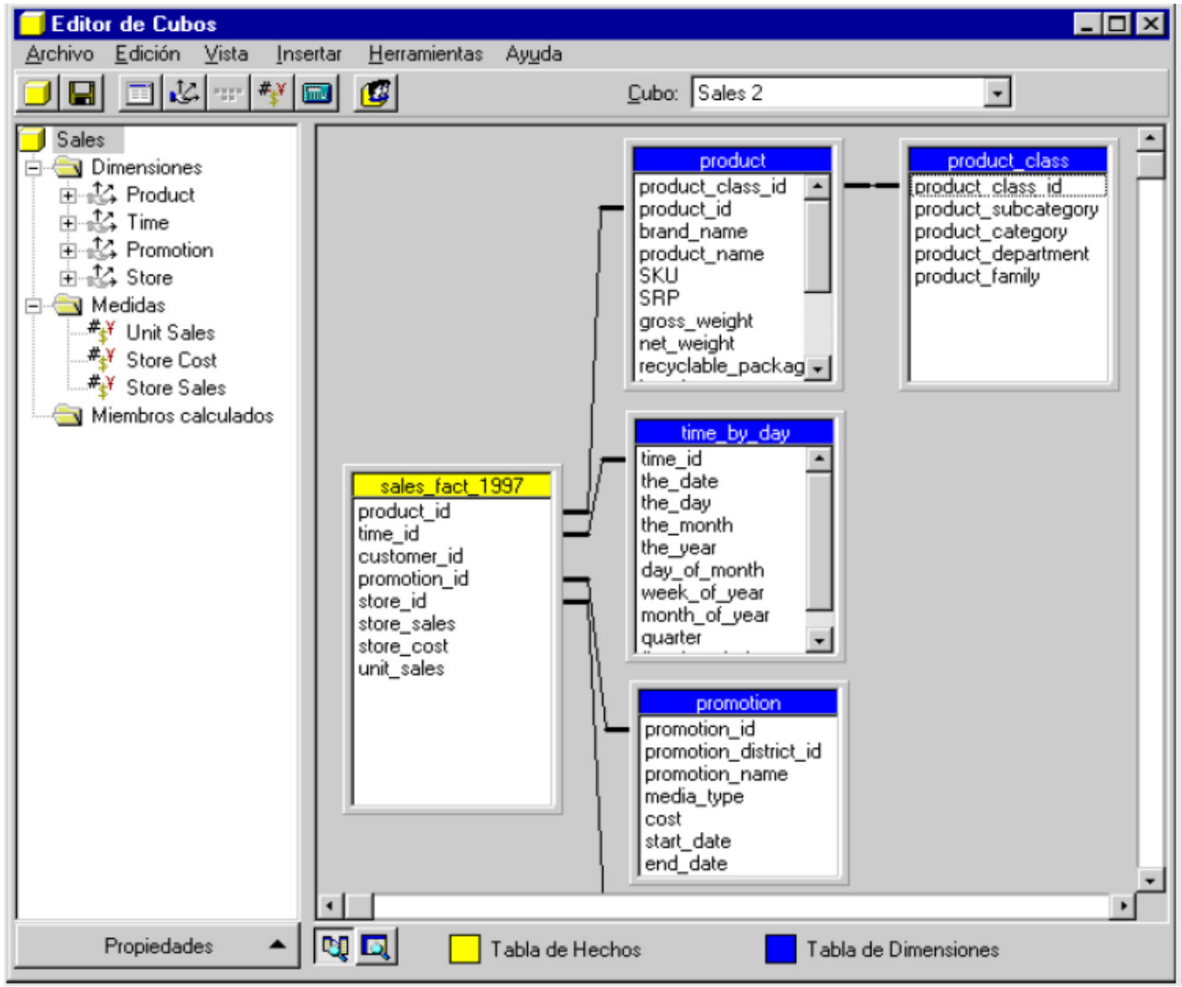
- MEDIDAS
  - Datos numéricos de interés para los usuarios.
  - Lo que queremos medir o seleccionar.
  - Algunos ejemplos:
    - Ventas.
    - Costos.
    - Unidades vendidas.
  - Se pueden crear algunas medias:
    - Beneficios= Ventas-Costos

Cubos Multidimensionales : Componentes



- DIMENSIONES
  - Representan columnas que describen las categorías a través de las cuales se separan las medidas.
  - Similitud con los ejes de un sistema cartesiano.
  - Tienen un limite máximo de 64 dimensiones.

Cubos Multidimensionales : Ejemplo




MARCH 25, 1996 $2.95
TIME
CHINA: Our New Enemy?
THE HENSEL TWINS: Sharing a Body
Can Machines Think?
They already do, say scientists. So what (if anything) is special about the human mind?