INTRODUCCIÓN A LA MINERÍA DE DATOS

UNA PERSPECTIVA ANALÍTICA

JAIME MIRANDA

(jmiranda@dii.uchile.cl)

Departamento de Ingeniería Industrial

Universidad de Chile

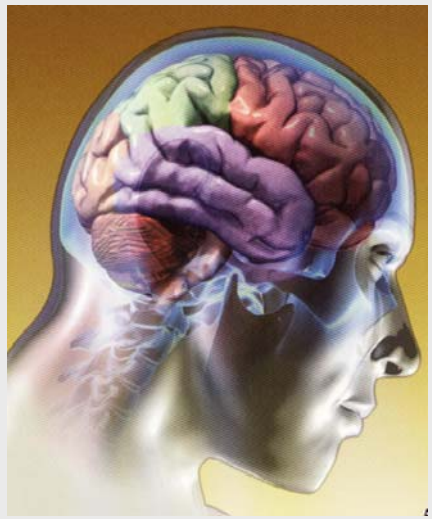
ALGUNAS DEFINICIONES …

APRENDIZAJE

→ **“El aprendizaje es una habilidad de la que disponen gran parte de los sistemas naturales para adaptarse al entorno en el que vive”.**

→ **“Adquisición de conocimiento de un proceso por medio del análisis, ejercicio o experiencia“.**

→ **“Un proceso por el cual los parámetros libres del sistema se adaptan a través de un proceso continuo de estimulación a partir del entorno en el que el sistema está inmerso”.**



¿QUÉ ES DATA-MINING?

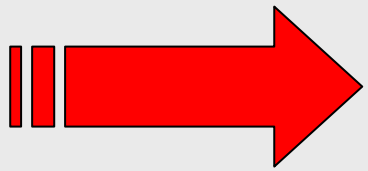
ALGUNAS DEFINICIONES:

→ *"Proceso de extracción de información y patrones de comportamientos que permanecen ocultos entre grandes cantidades de información."*

→ *"Proceso que a través del descubrimiento y cuantificación de relaciones predictivas en los datos, permite transformar la información disponible en conocimiento útil."*

Información

Relaciones



Conocimiento útil

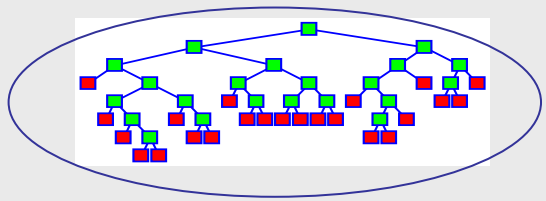
Patrones ocultos

¿POR QUÉ ES NECESARIO?

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.



ALMACENAMIENTO Y CAPACIDAD DE PROCESAMIENTO

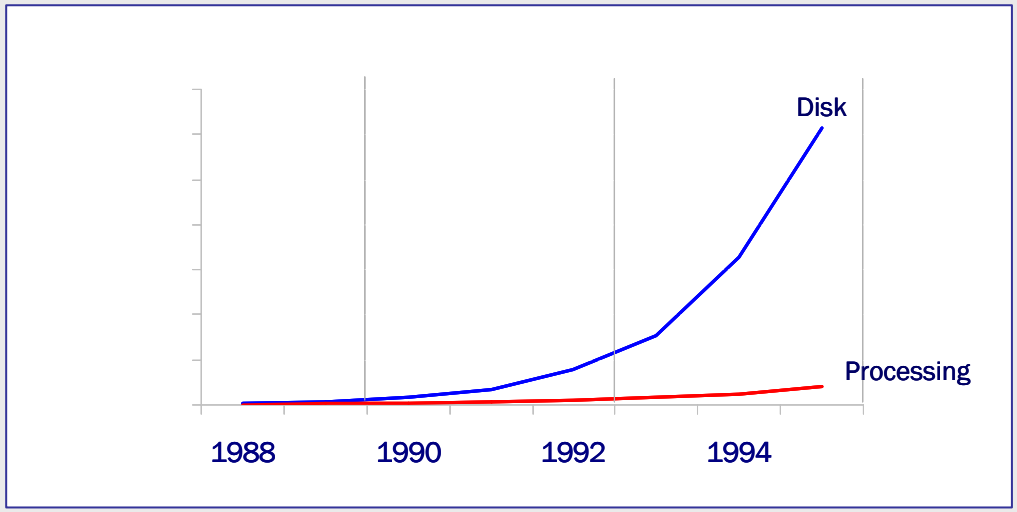
UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

LEY DE MOORE:

→ **"La capacidad de procesamiento se duplica cada 18 meses"**

RESPECTO AL ALMACENAMIENTO:

→ **"La capacidad de almacenamiento se duplica cada 9 meses"**



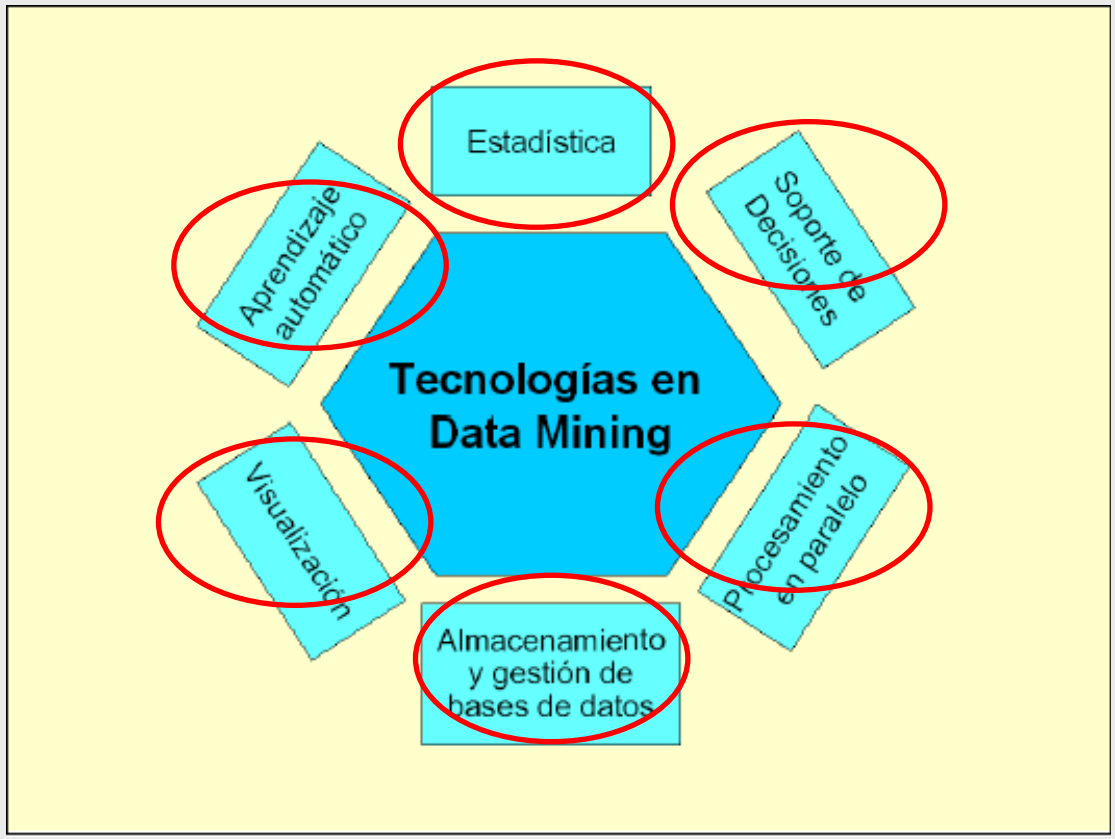
La brecha entre capacidad de procesar lo que almacenamos,

aumenta con el tiempo

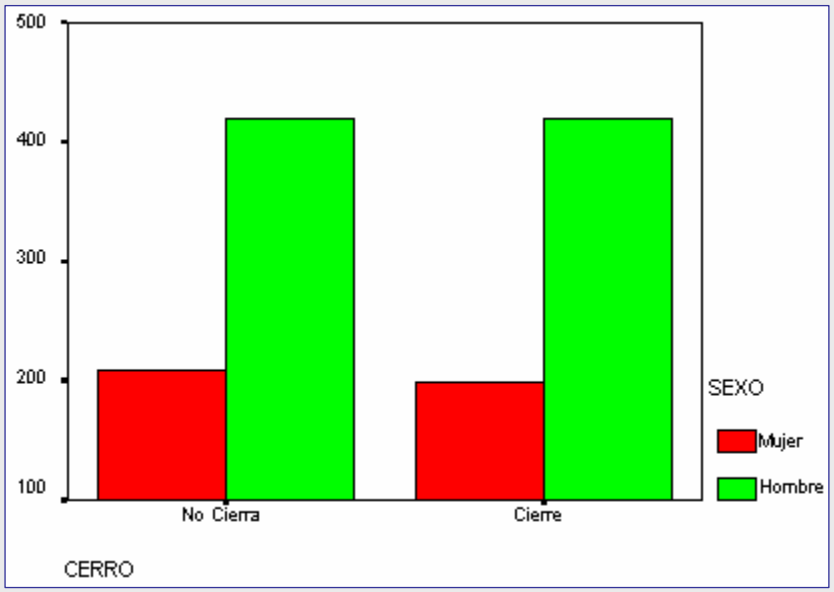
¿DE DONDE SURGE EL DATA-MINING ?

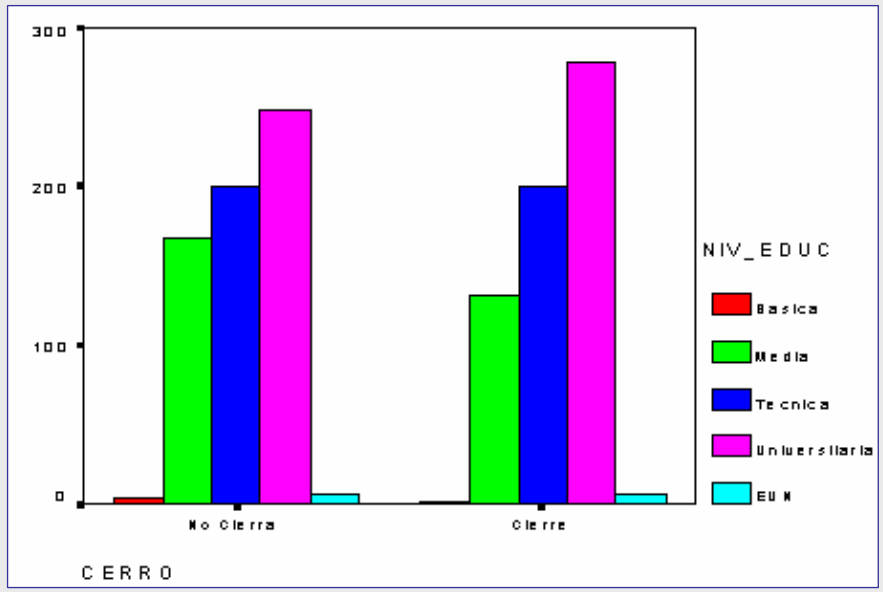
UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

De la integración múltiple...



UN PEQUEÑO EJEMPLO …





EL VÉRTIGO DE LA INTELIGENCIA DE NEGOCIOS

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

BUSINESS INTELLIGENCE – DEFINICIÓN

BUSINESS INTELLIGENCE

→La Inteligencia del Negocio (BI) representa las herramientas y sistemas que juegan un papel clave en el proceso estratégico de la planificación de una compañía. Estos sistemas permiten reunir, almacenar, y analizar los datos corporativos siendo una importante ayuda en la toma de decisiones.

Generalmente estos sistemas ilustrarán los perfiles de los clientes, estudios de mercado, segmentación de clientes, predicción de comportamientos.

POTENCIAL DE LA INTELIGENCIA DE NEGOCIOS
PARA EL MERCADO CHILENO

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

SECTOR FINANCIERO

→Retención de clientes.
→Detección de fraudes.
→Credit Scoring

SECTOR ISAPRES

→Detección de fraudes.
→Retención de clientes.

SECTOR RETAIL

→Venta cruzada.
→Predicción de demanda.
→Segmentación de clientes.

APLICACIONES DE NEGOCIO

DETECCIÓN DE FRAUDES:

→ **Identificar transacciones fraudulentas**

MARKETING Y VENTAS:

→ **Identificar potenciales clientes; establecer la efectividad de las campañas de marketing**

ANÁLISIS DE PROCESOS DE MANUFACTURA:

→ **Identificar las causas de fallas en máquinas**

ENTENDIENDO COMPORTAMIENTO DE CONSUMIDORES:

→ **modelos de retención de clientes, afinidades, clustering**

APLICACIONES DE NEGOCIO (2)

APROBAR CRÉDITOS:

→ **Establecer Credit Scoring para un cliente a la hora de pedir un préstamo**

GESTIÓN DE PORTAFOLIO:

→ **optimizar un portafolio de instrumentos financieros maximizando el retorno o minimizando el riesgo**

ANÁLISIS DE WEBSITES:

→ **modelar preferencias de usuarios desde logs, filtros colaborativos, caminos preferidos, etc.**

¿QUÉ SIGNIFICA DATABASE MARKETING?

→ **Es una colección de datos que proporciona Información para los expertos del negocio ayudándolos a tomar las mejores decisiones de trabajo cumpliendo con los objetivos del negocio**

→ **Más específico, database marketing puede definirse como reunir, guardar y utilizar la máxima cantidad de conocimiento de tus clientes y prospectos, para su beneficio y tú ganancia**

ALGUNAS ÁREAS TRATADAS EN DM

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

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.**

PROBLEMAS DE 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.

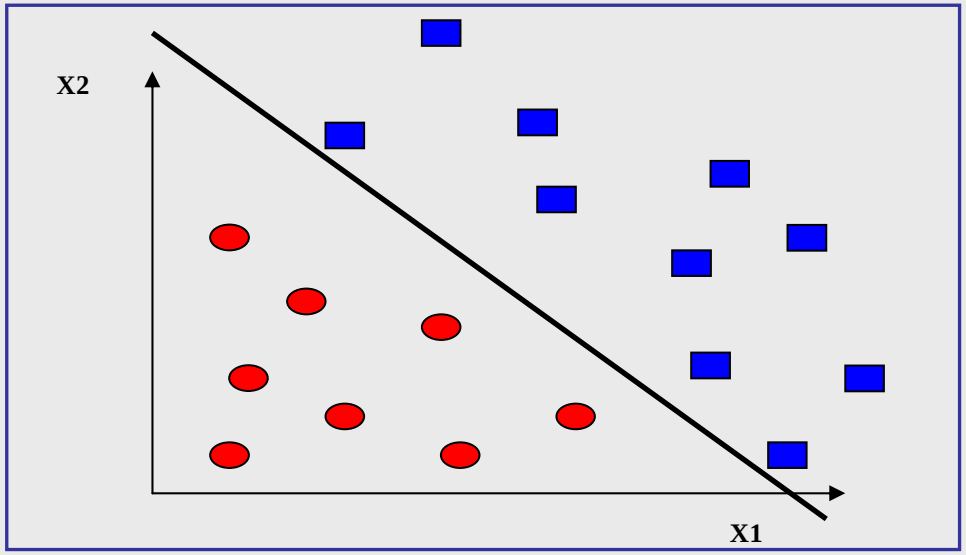
Se tiene siempre un número limitado de clases y se esperar poder asignar

PROBLEMAS DE 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.



PROBLEMAS DE 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)**

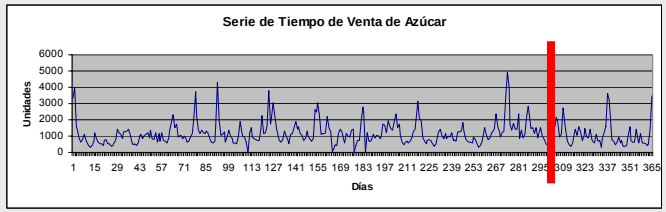
Ejemplos

→ **Estimar el número de hijos de una familia.**

→ **Estimar la probabilidad de que alguien conteste a un mailing.**

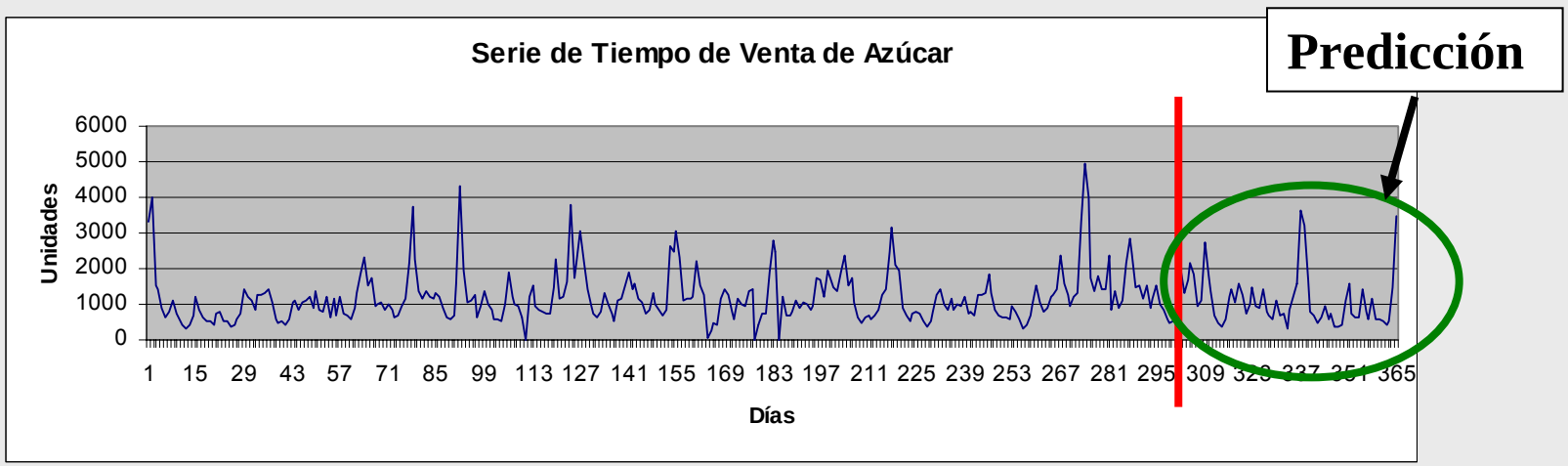
→ **Estimar el tiempo de vida de un cliente.**

→ **Estimar los ingresos totales de una familia.**



PROBLEMAS DE 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.**



ASOCIACIÓN O CANASTA DE LA COMPRA

IDEA CENTRAL: **Determinar que cosas van juntas.**

→ **Pañales y cerveza se compran juntos los fines de semana**

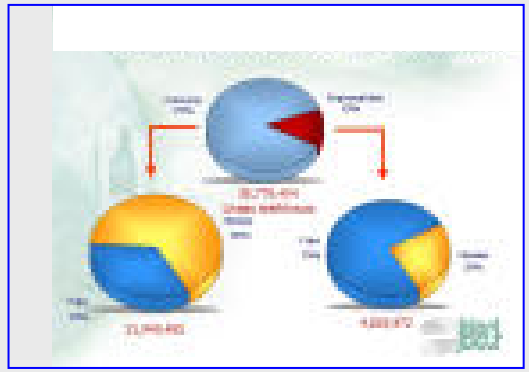
El ejemplo típico es observar qué productos suelen ir juntos en la cesta de

la compra

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

PROBLEMAS DE SEGMENTACIÓN

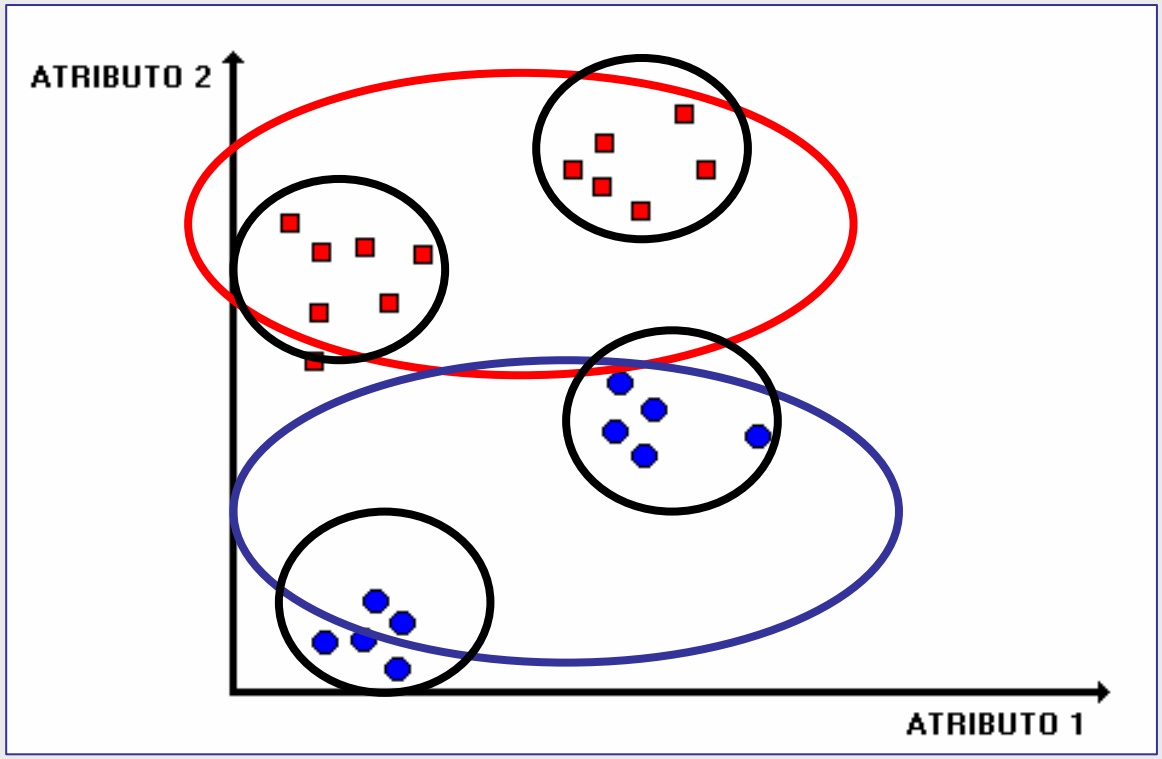
- **Segmentar una población heterogénea en un número de subgrupos homogéneos o clusters.**
- **No hay clases predefinidas**
- **Registros agrupados en base a su similitud.**
- **Se realiza a menudo antes de otras tareas de descubrimiento.**
  → **Encontrar clientes con hábitos de compra similares**



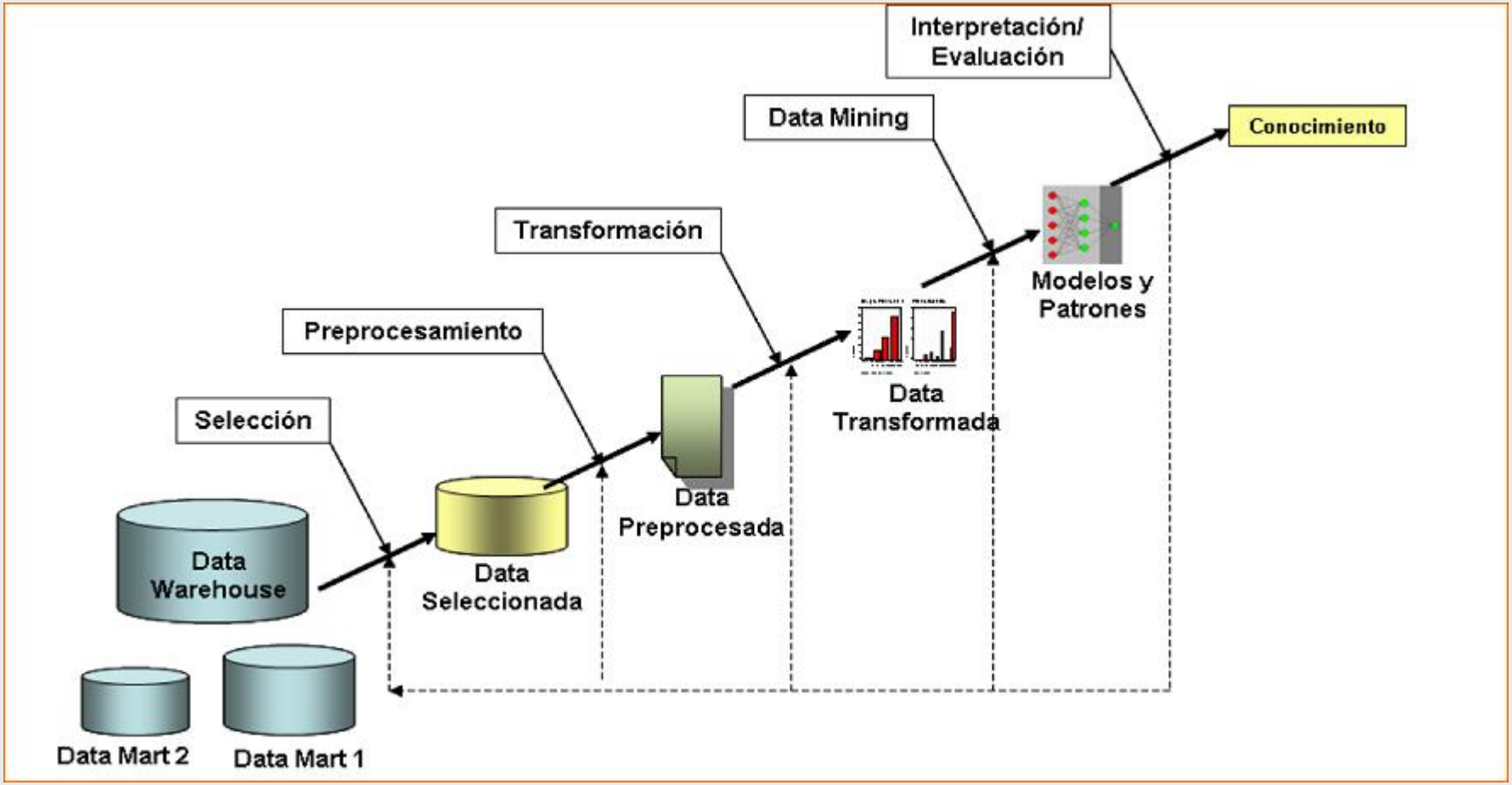
PROBLEMAS DE SEGMENTACIÓN (2)

Encontrar patrones característicos no visibles a simple vista.

Encontrar soluciones entre subconjuntos o subpoblaciones.



PROCESO DE KDD
KNOWLEDGE DISCOVERY IN DATABASES



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

Teoría de Bases de Datos

NOCIONES BÁSICAS

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile

BASE DE DATOS RELACIONALES (RDBMS)

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

- **Nacieron en los 70´S creadas por E.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.**

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FISICAS Y MATEMATICAS
DEPARTAMENTO DE
INGENIERIA INDUSTRIAL

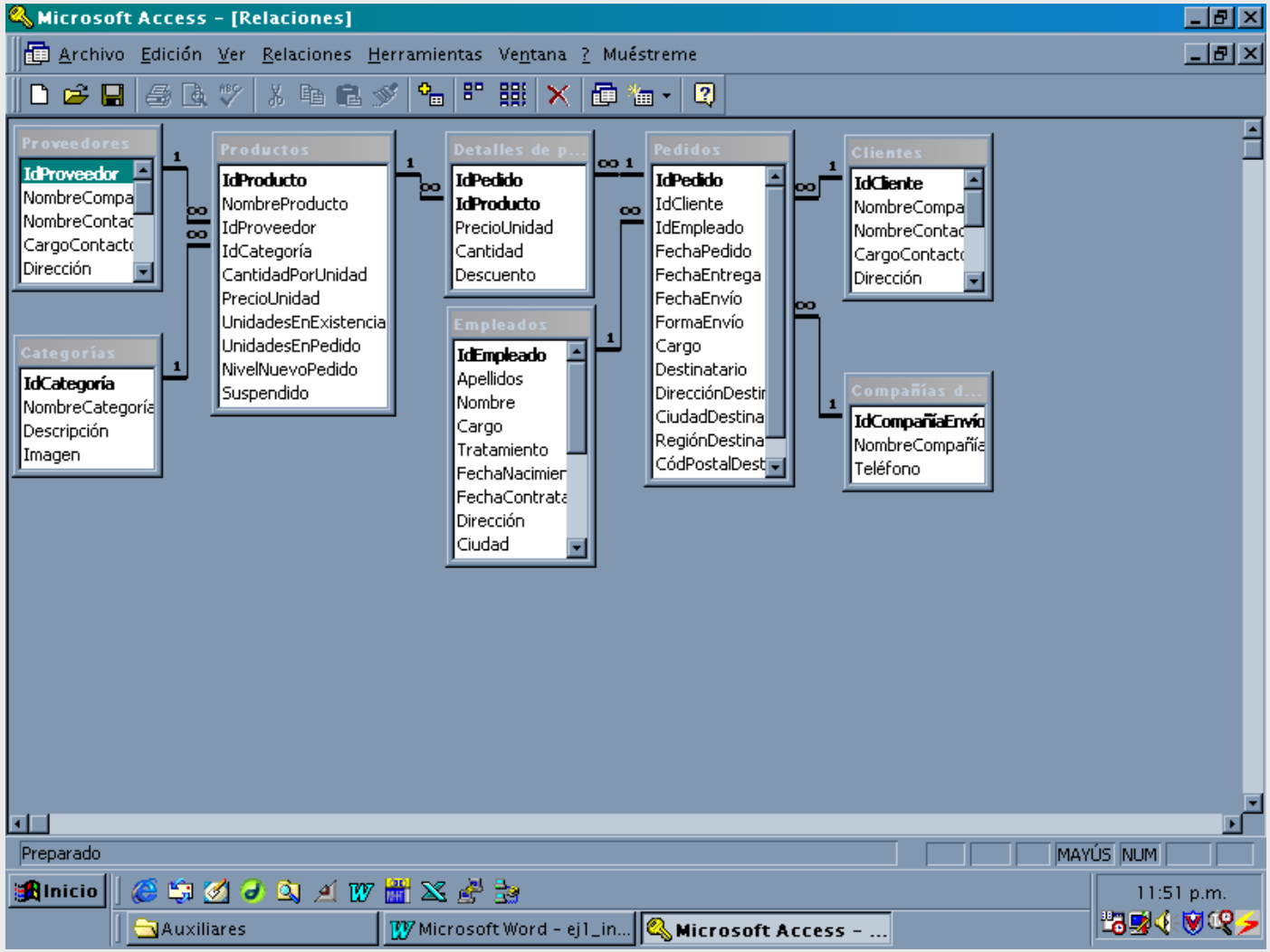
Relación Empleado

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

ALGUNOS EJEMPLOS …



SINTAXIS SQL : “Structured Query Language”

→ **SELECT < Lista de atributos >**

→ **FROM < Lista de tablas >**

→ **WHERE < Condición >**

EJEMPLOS SQL

Recuperar todos los números de RUT de los empleados.

→ **SELECT RUT**

→ **FROM EMPLEADO**

Recuperar los valores de todos los atributos de EMPLEADO que trabajan en el departamento número "5".

→ **SELECT * **

→ **FROM EMPLEADO**

→ **WHERE NDEPTO = 5**

Ejemplos SQL (2)

Comando AND: Recuperar la fecha de nacimiento y dirección del empleado Juan Pérez

→ **SELECT FNAC, DIRECCION**

→ **FROM EMPLEADO**

→ **WHERE NPILA = "Juan" AND APPAT = "Pérez"**

Data Warehouse y Data Mart

NOCIONES BÁSICAS

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile

ALGUNAS FUNCIONES...

- **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**

ALGUNAS DEFINICIONES…

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

DATAWAREHOUSE: Colección de objetos

→ **Orientada al sujeto:**

→ **Organizada en torno a los datos más importantes de la empresa.**

→ **Es bueno para realizar filtros y eliminar información poco importante.**

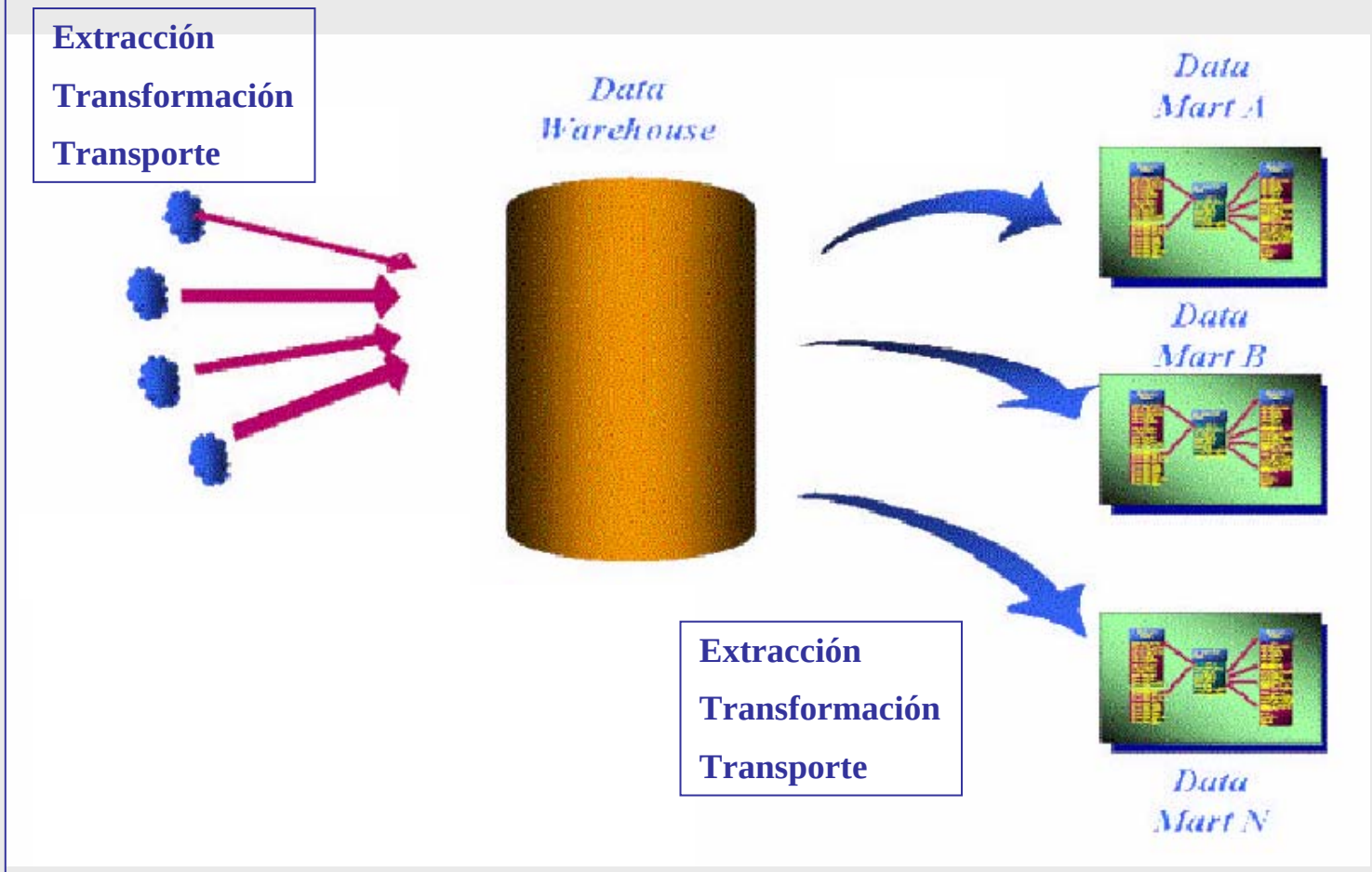
→ **El modelamiento se enfoca en el análisis y toma de decisiones basadas en estos datos particulares y no en el procesamiento diario de las transacciones.**

→ **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.**

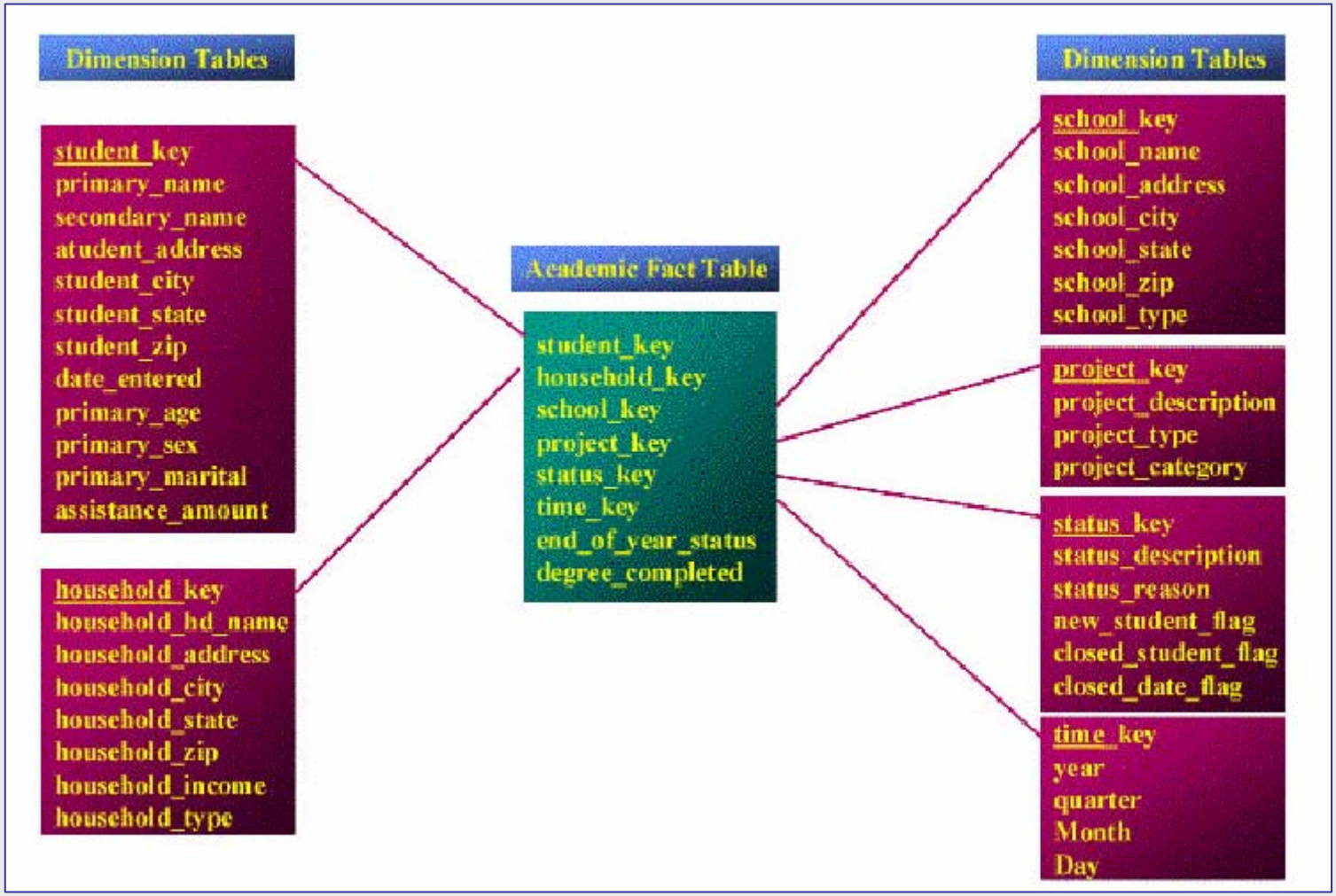
DATAWAREHOUSE: Colección de objetos

- → **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.**

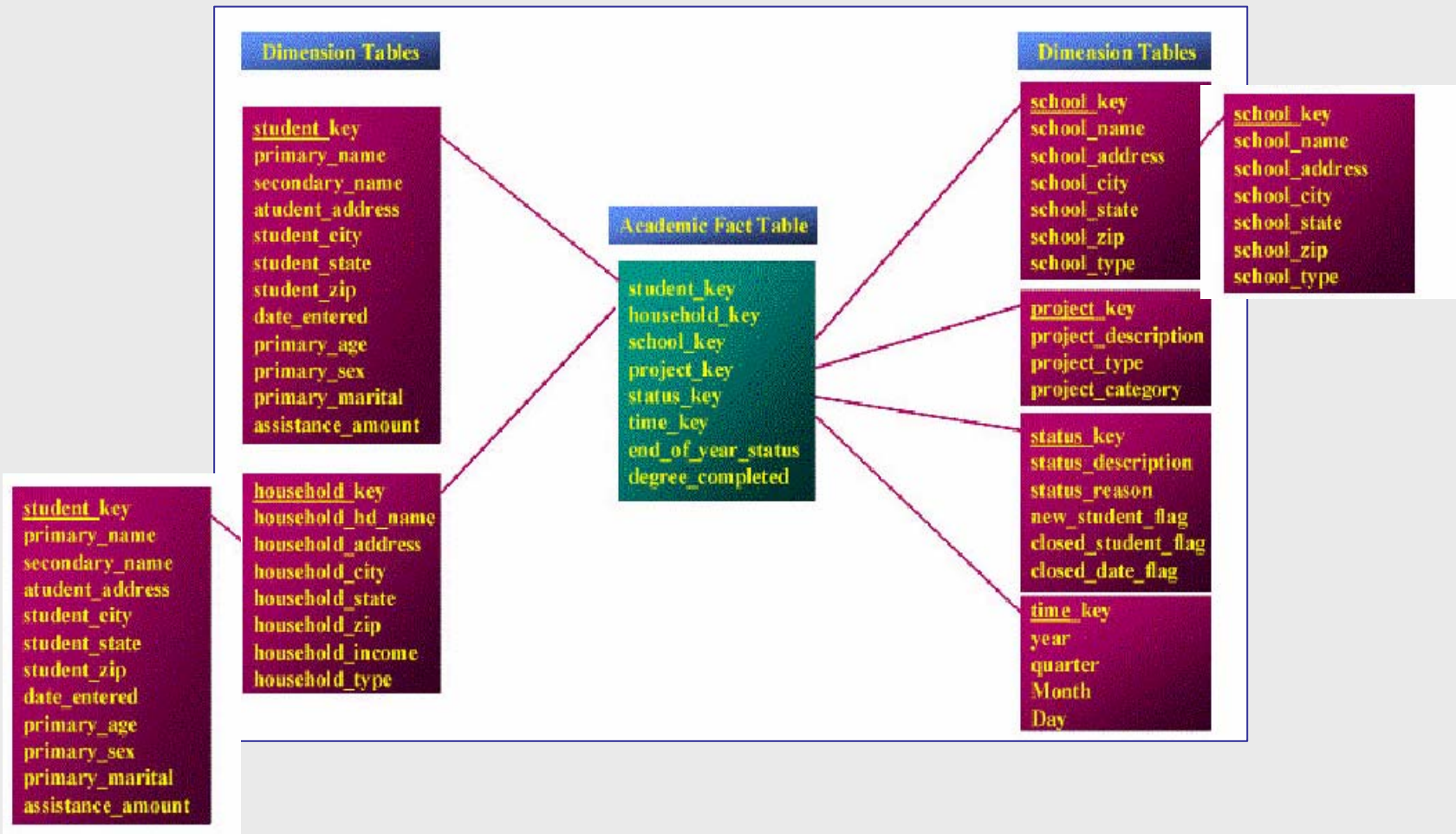
GRÁFICAMENTE …



ESQUEMA “ESTRELLA”



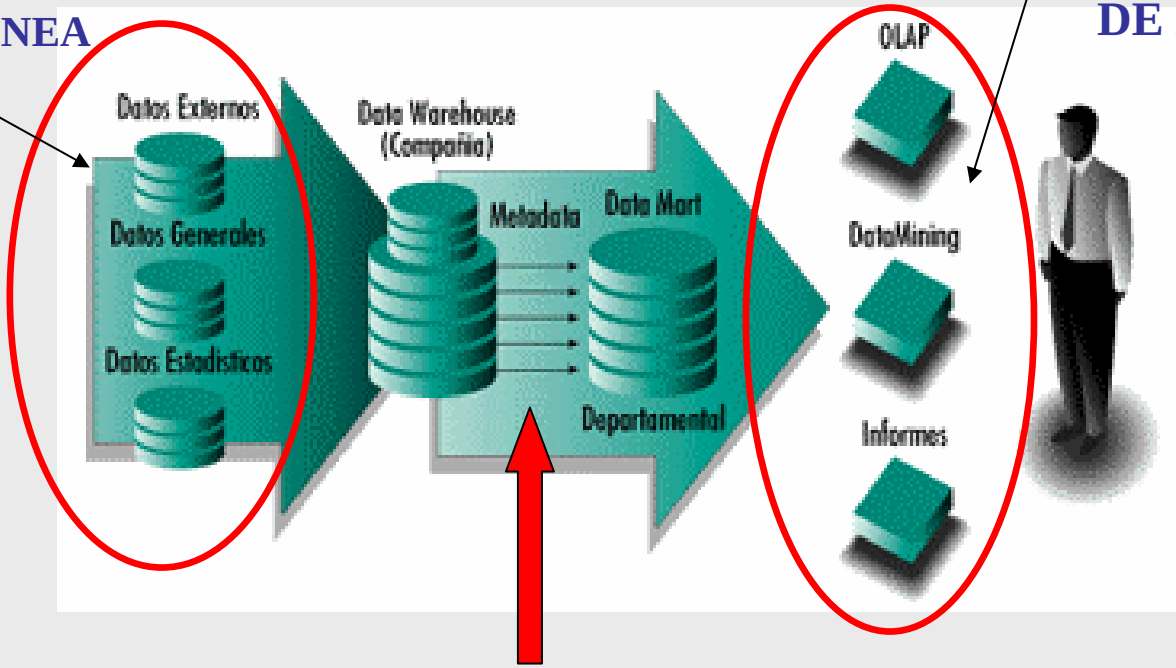
ESQUEMA “COPO DE NIEVE”



ARQUITECTURA MULTICAPAS

INFORMACION HETEROGENEA

HERRAMIENTAS DE ANALISIS



METADATOS

Cubos Multidimensionales

NOCIONES BÁSICAS

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile

ALGUNAS NOCIONES...

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

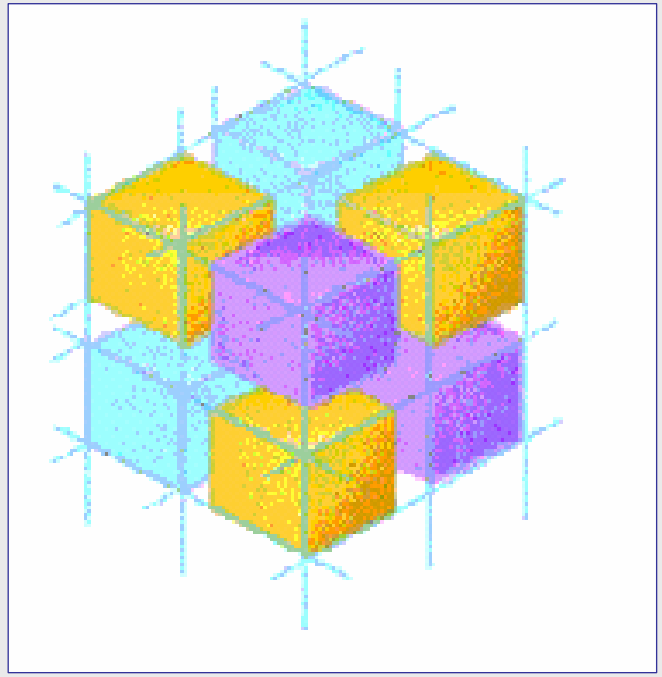
Consiste en una representación multidimensional de datos de detalle y resumen.

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

ALGUNAS NOCIONES...

Son un subconjunto de datos de la base de datos original.

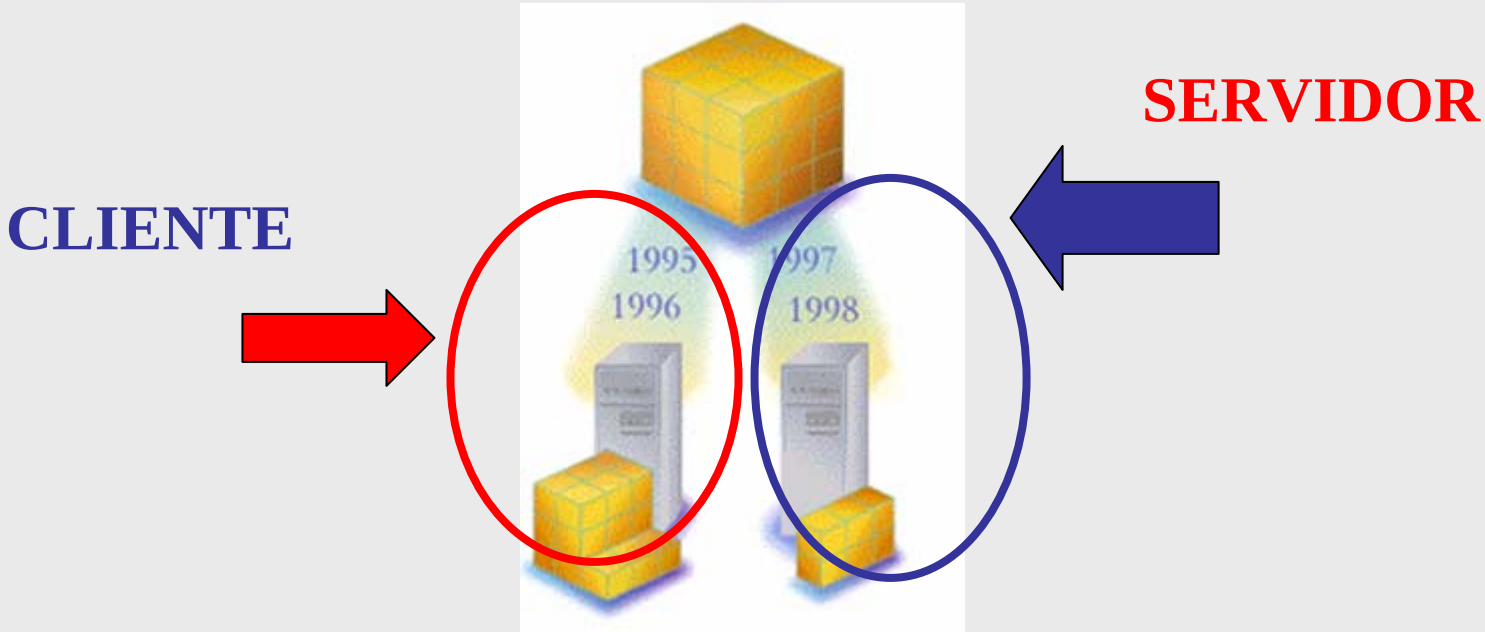
Son capaces de administrar de forma rápida y eficiente grandes cantidades de información.



COMPONENTES DE UN CUBO

ORIGEN DE LOS DATOS

→ **Identifica y conecta donde se encuentra el almacén de datos la información relevante para resolver un problema.**



COMPONENTES DE UN CUBO (2)

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

MEDIDAS

→ **Datos numéricos de interés para los usuarios.**

→ **Que queremos medir o seleccionar.**

→ **Algunos ejemplos:**

- **Ventas.**
- **Costos.**
- **Unidades vendidas.**

→ **Se pueden crear algunas medias:**

- **Beneficios= Ventas-Costos**

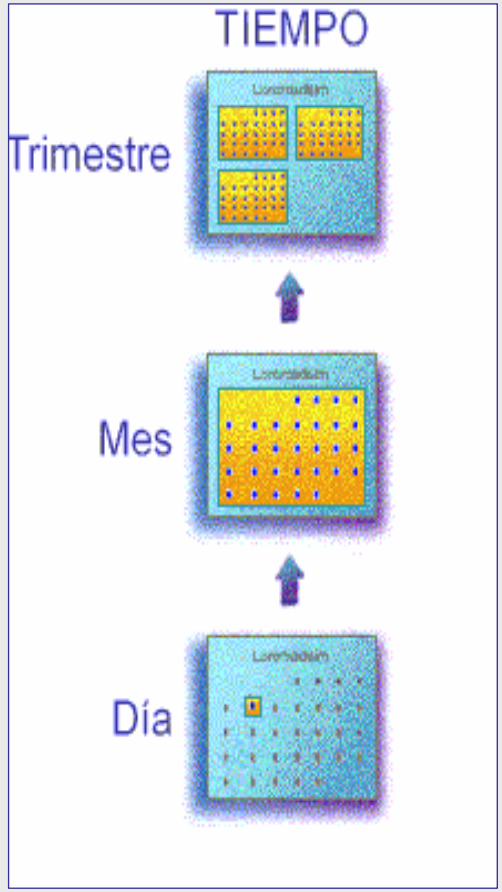
Componentes de un cubo (3)

DIMENSIONES

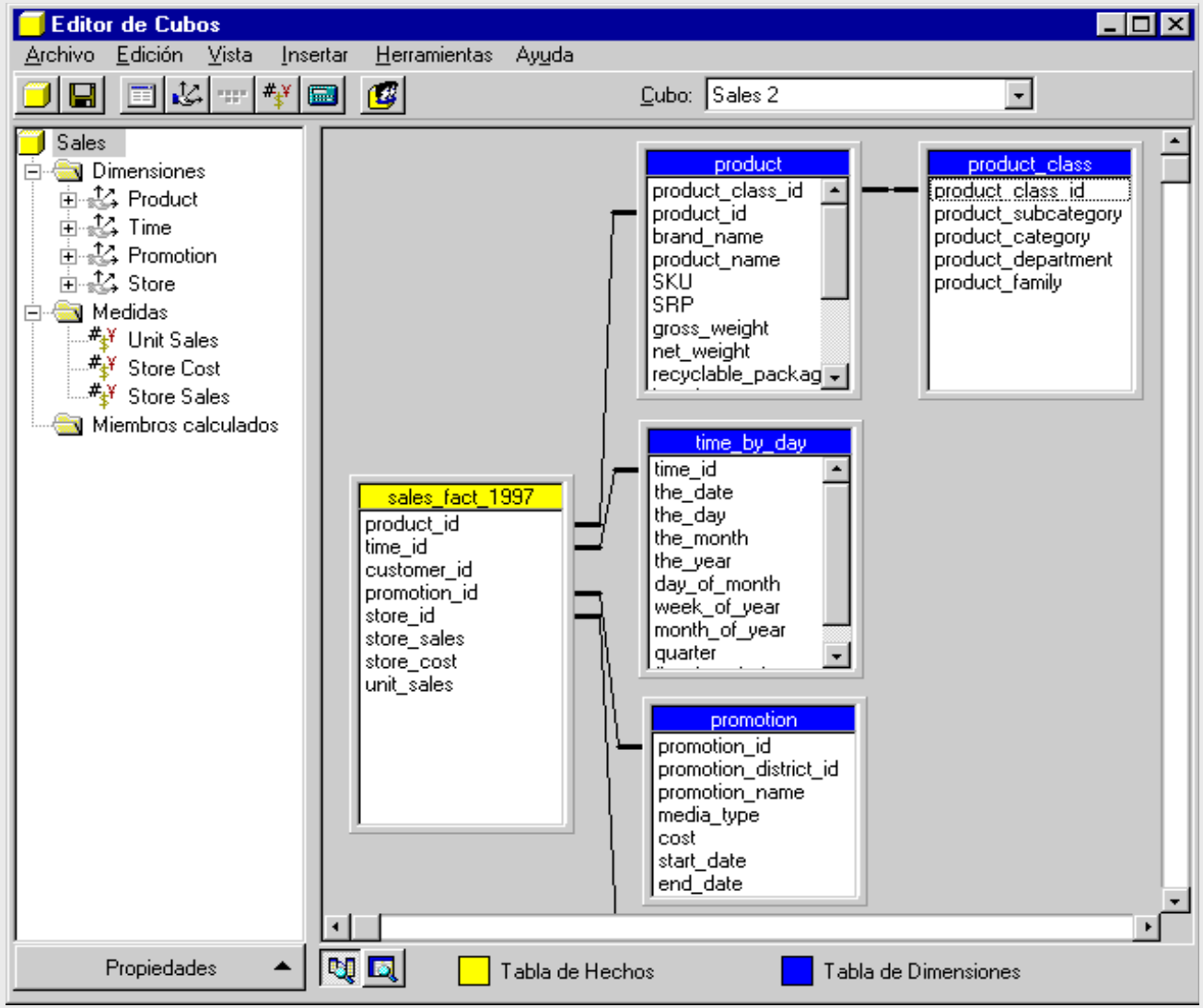
→ **Representan columnas que describen las categorías a través de las cuales se separan las medidas.**

→ **Similitud con los ejes de un sistema cartesiano.**

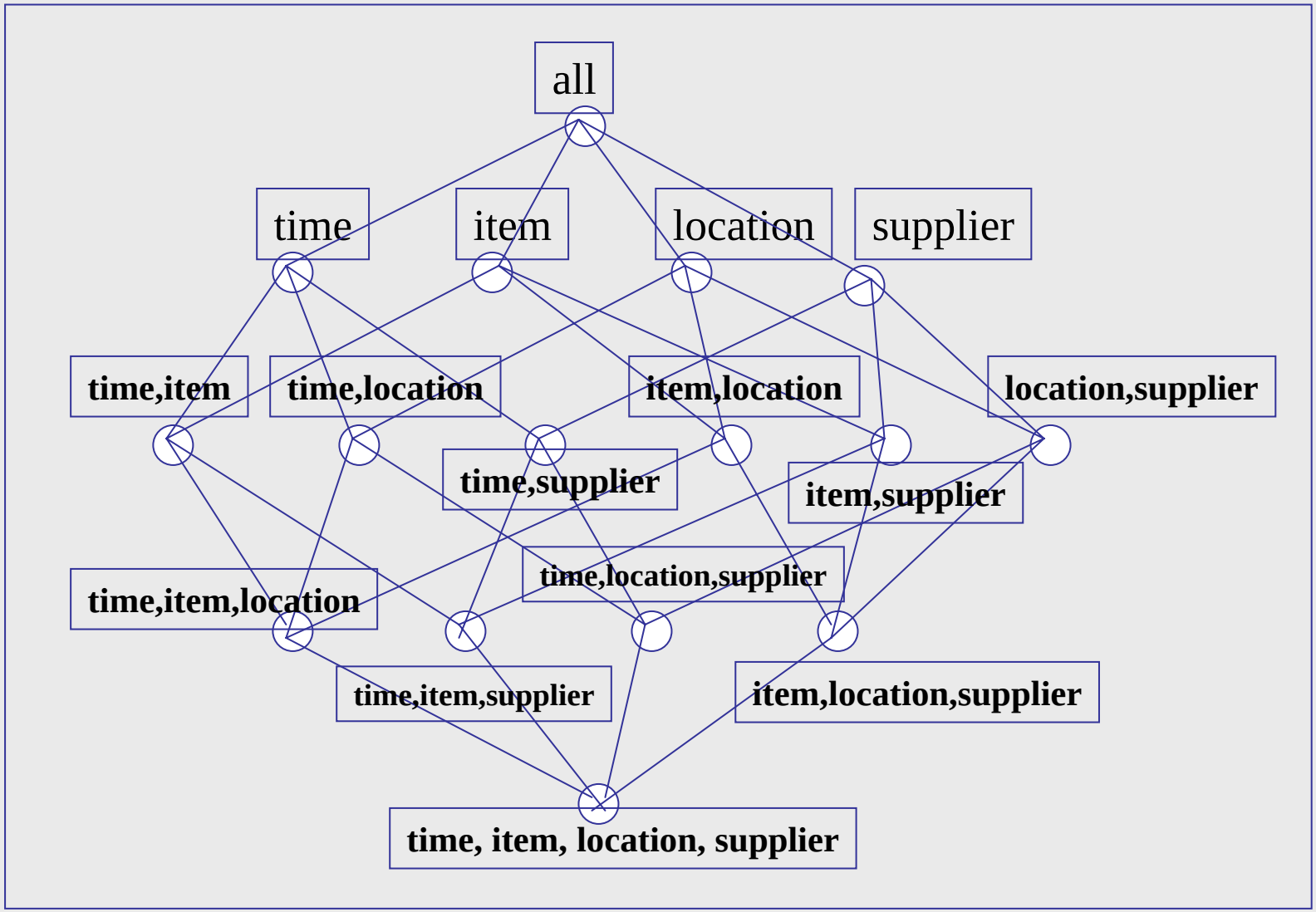
→ **Tienen un limite máximo de 64 dimensiones.**



EJEMPLO DE UN CUBO



OTRO EJEMPLO DE UN CUBO …



MODOS DE ALMACENAMIENTO

MOLAP (OLAP multidimensional)

→ **Formato de almacenamiento de alto rendimiento.**

→ **Esta altamente especializado a datos multidimensionales.**

→ **Se aconseja para conjuntos de datos pequeños o medios.**

→ **Es recomendable para cubos de uso frecuente , pues presenta tiempos de respuesta rápidos y eficientes.**

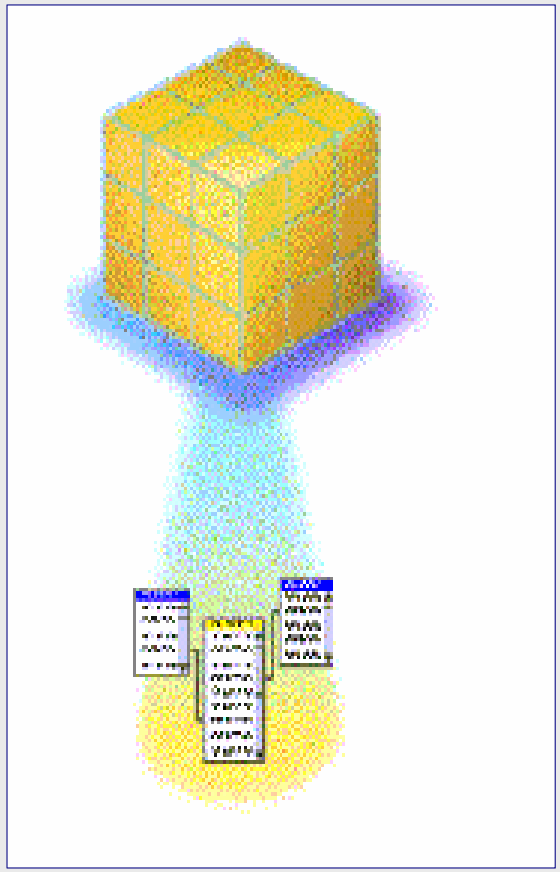
MODOS DE ALMACENAMIENTO (2)

ROLAP

→ **Los datos permanecen en las tablas originales.**

→ **Se utiliza un conjunto separado de tablas relacionales para hacer referencia a los datos agregados.**

→ **Ideal para bases grandes o datos antiguos que se consultan con poca frecuencia.**



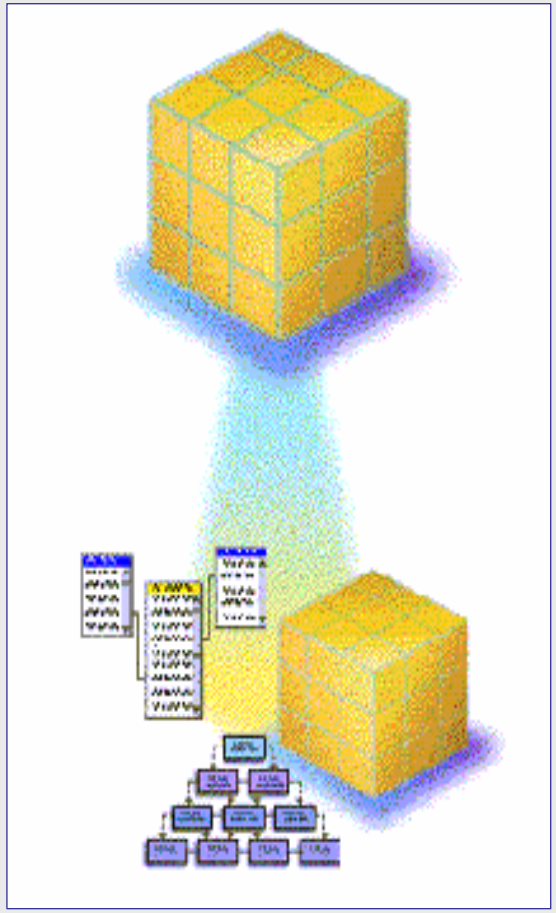
Modos de almacenamiento (3)

HOLAP

→ **Combinación de ambos modos (OLAP y ROLAP).**

→ **Mantiene los datos originales en tablas relacionales (ROLAP).**

→ **Mantiene los datos agregados en formato multidimensional (MOLAP)**



PROCESAMIENTO DE CUBOS

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

Esto implica:

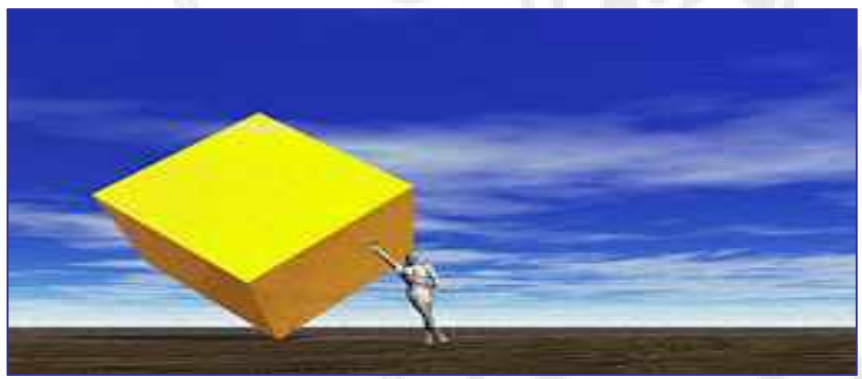
- → **Leer las tablas de dimensiones para llenar los miembros con los datos actuales.**
- → **Leer la tabla de hechos.**
- → **Almacenar los datos en el cubo.**
- → **Se debe procesar un cubo cada ves que se ingresen nuevos valores o cuando se modifiquen alguna dimensión o medida.**

Ventajas:

- → **Mejora la eficiencia de las consultas**
- → **Reducen los tiempos de respuesta.**

PARTE Nº1

ANÁLISIS MULTIDIMENSIONAL DE DATOS



ANÁLISIS MULTIDIMENCIONAL

→ **Trabaja con un conjunto reducido de datos.**

→ **Genera reportes mas claros y más específicos.**

→ **Responden con mayor rapidez a las consultas**

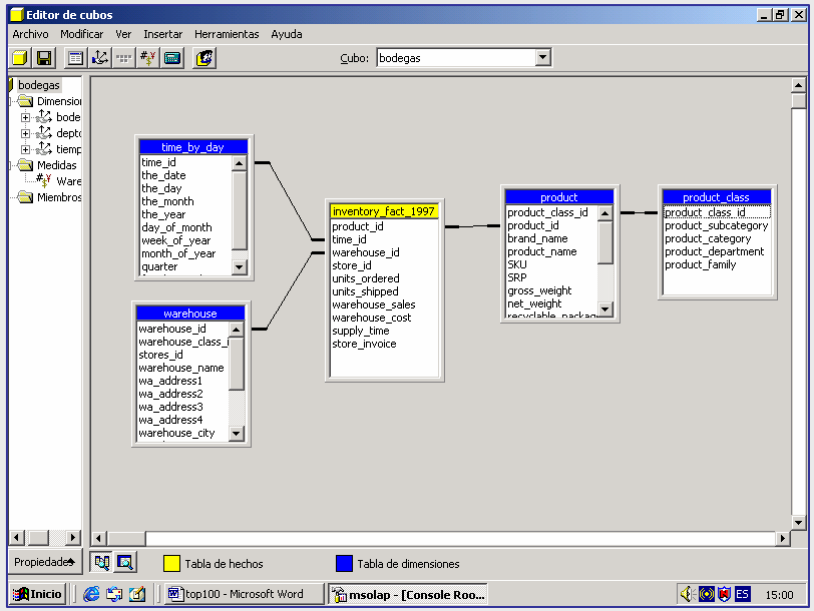
Mejorar el rendimiento en línea y
el rendimiento de las consultas a las bases

CUBOS MULTIDIMENSIONALES

COMPONENTES PRINCIPALES

→ Dimensiones

→ Medidas



PRINCIPALES ARQUITECTURAS

→ Estrella

→ Copo de nieve

PRINCIPALES TIPOS

→ OLAP

→ MOLAP

→ ROLAP

SQL Server 7.0

TAREA Nº1: ANÁLISIS MULTIDIMENCIONAL

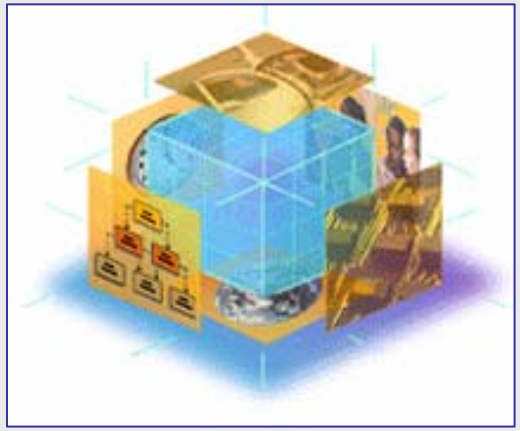
OBJETIVO GENERAL

→ **“Entender y usar las herramientas de análisis multidimensional con el fin de comprender y describir el comportamiento de una variable dada”**

SE TENDRÁ:

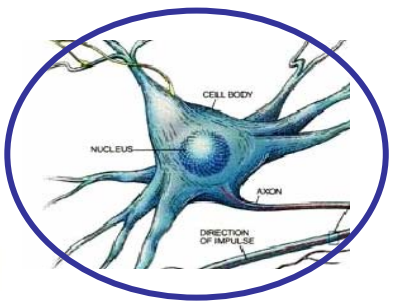
→ **Una base de datos relacional.**

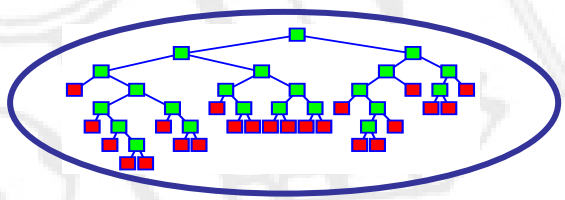
→ **Una serie de consultas y preguntas sobre el comportamiento de una o más variables.**



PARTE Nº2

TÉCNICAS DE MINERÍA DE DATOS



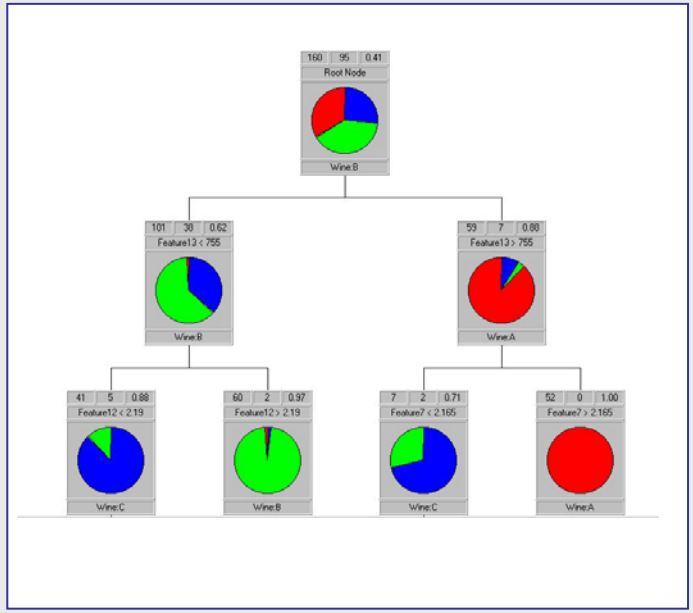


TÉCNICAS EMPLEADAS

MÉTODOS SUPERVISADOS

→ **Redes neuronales**

→ **Árboles de decisión**



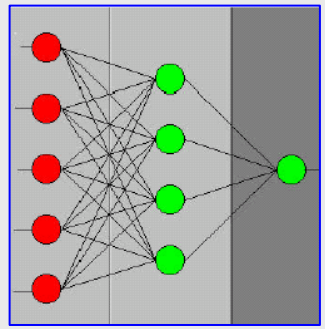
MÉTODOS NO SUPERVISADOS

→ **Fuzzy C-means (Cluster)**

→ **Mapas Kohonen**

REDES NEURONALES MLP

APLICACIONES



→ Retención o fuga de clientes
→ Detección de fraudes
→ Scoring (varios)

FORTALEZAS

→ Fuerte en lo referente a la modelación no lineal
→ Trabaja tanto con variables categóricas como continuas
→ Alta aplicabilidad (variadas áreas de estudio)

DEBILIDADES

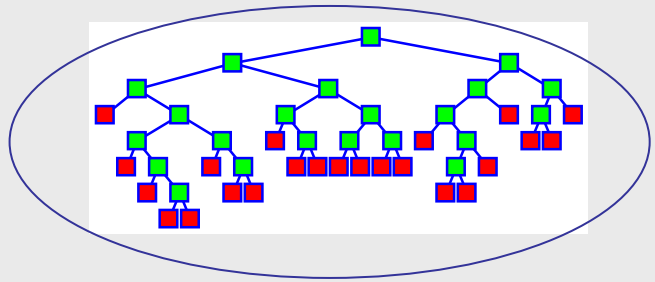
→ Difícil interpretación de las relaciones entre las variables (Heurísticas)
→ Sobreajuste

ÁRBOLES DE DECISIÓN

APLICACIONES

→ **Segmentación de clientes**

→ **Generación de reglas de clasificación en general**



FORTALEZAS

→ **Fácil interpretación y entendimiento**

→ **Genera un ranking automático de variables**

→ **Rápida convergencia del algoritmo**

DEBILIDADES

→ **Si poseen mucha “profundidad” son difíciles de interpretar**

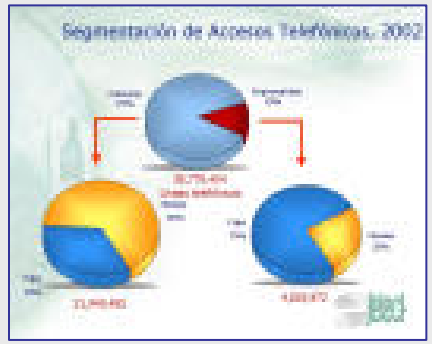
→ **Posibilidades discretas: relacionado a variables con muchas categorías**

ALGORITMO FUZZY C-MEANS

APLICACIONES

→ **Marketing**

- **Segmentación de clientes**
- **Ofertas focalizadas**



FORTALEZAS

→ **No asume ninguna distribución estadística entre los datos.**

→ **Los resultados son mas intuitivos y fáciles de entender e interpretar**

DEBILIDADES

→ **Son muy sensitivos a los valores fuera de rango (outliers)**

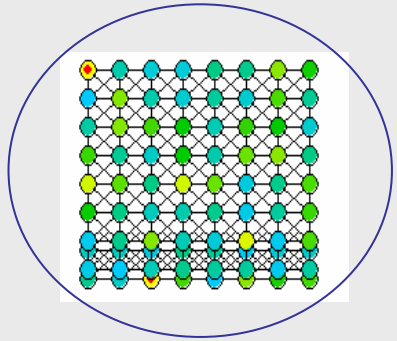
→ **No trabajan bien con variables categóricas**

MAPAS DE KOHONEN

APLICACIONES

→ **Marketing**

- **Segmentación de clientes**
- **Ofertas focalizadas**



FORTALEZAS

→ **Reduce la dimensionalidad del espacio.**

→ **Los resultados son mas intuitivos y fáciles de entender e interpretar**

DEBILIDADES

→ **Sólo nos da una visón espacial de los resultados**

→ **No trabajan bien con variables categóricas**

TAREA Nº2: CONSTRUCCIÓN DE MODELOS DE DM

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

OBJETIVO GENERAL

→ Comprender y estudiar los distintos métodos de minería de datos

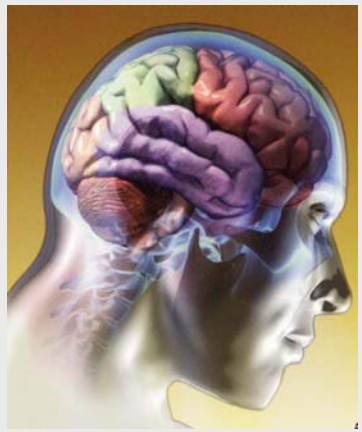
→ Hacer sensibilidad de parámetros y obtener configuraciones optimas

→ Seleccionar de atributos relevantes para el análisis

SE TENDRÁ:

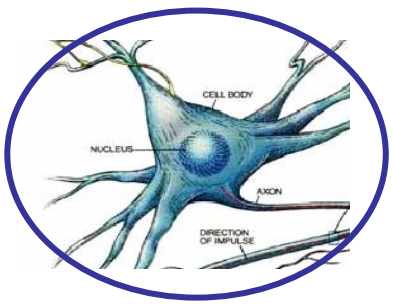
→ Una base de datos de un problema de clasificación con la clase especifica

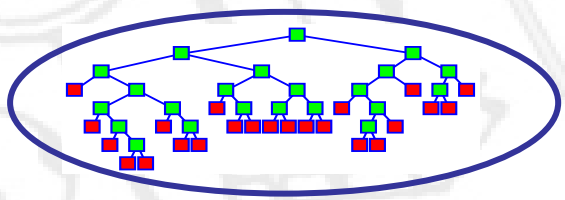
→ Una base sin la clase y el problema será predecir dicha clase



PARTE Nº3

CASOS APLICADOS A LA INDUSTRIA





OBJETIVOS DE LOS CASOS APLICADOS

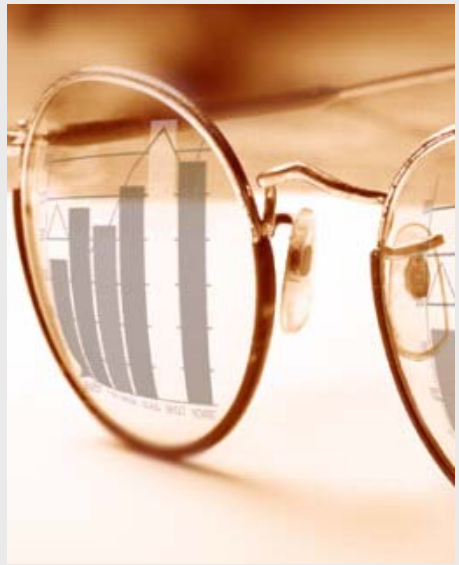
OBJETIVOS

→ **Lograr un a primera aproximación a un problema real en Data-Mining.**

→ **Desarrollar de habilidades en el uso practico de algunas de las técnicas usadas en la minería de datos.**

→ **Encontrar el mejor modelo para la resolución del problema.**

→ **Generar políticas comerciales asociadas a cada problema en particular.**



CASOS PRELIMINARES DE ESTUDIO

CASO Nº1: Modelo Predictivo de Ofertas Focalizadas

CASO Nº2 : Modelo Predictivo para Créditos de Consumo

CASO Nº3 : Modelo Predictivo de Fugas de Cliente

CRITERIOS DE EVALUACIÓN

TAREAS

Nota Tarea = 0.85* Nota Informe + 0.15* Nota Interrogación.
Nota Tarea Final = Promedio de tareas.

CTP´S

6 * (número de CTPs aprobados/número total de CTPs) +1

NOTA FINAL

Nota examen 40%
Nota promedio de tareas 40%
Nota CTP 20%

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?

13

7244Q4 1