



Introducción a la Minería de Datos

DE LOS DATOS AL CONOCIMIENTO...

JAIME MIRANDA

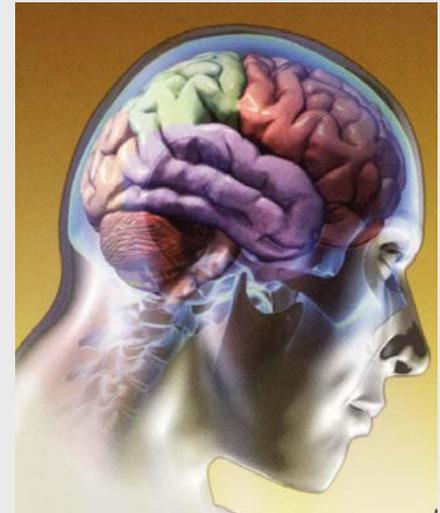
(jmiranda@dii.uchile.cl)

Departamento de Ingeniería Industrial

Universidad de Chile

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

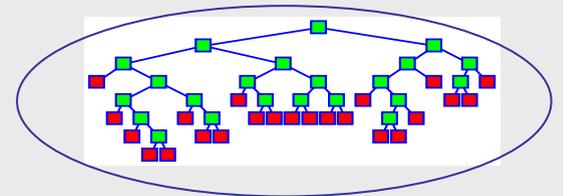
Patrones ocultos

¿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.



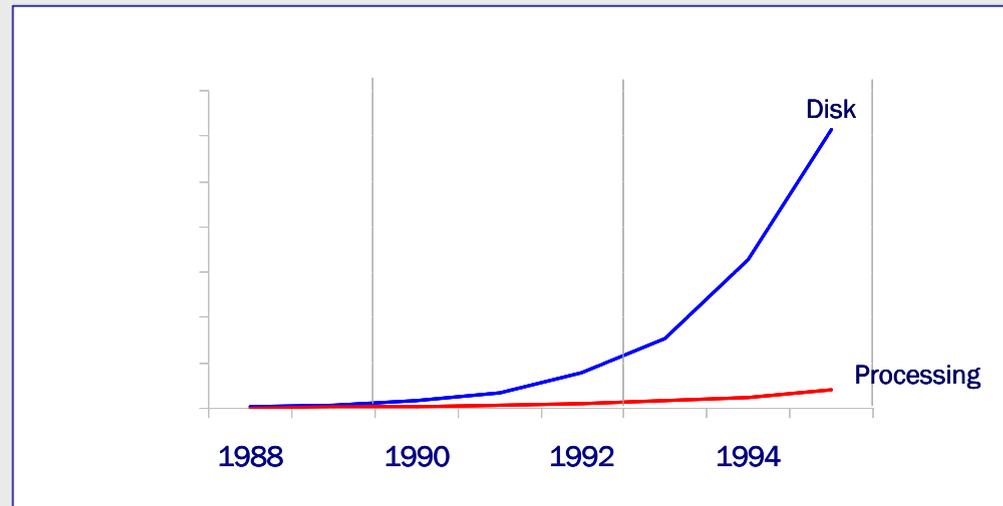
ALMACENAMIENTO Y CAPACIDAD DE PROCESAMIENTO

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

¿CÓMO SE USA DATA MINING HOY?

DETECTAR SEGMENTOS

CALCULAR PERFILES

CROSS-SELLING

DETECTAR BUENOS CLIENTES

EVITAR EL “CHURNING”, “ATTRITION”

DETECCIÓN DE MOROSIDAD

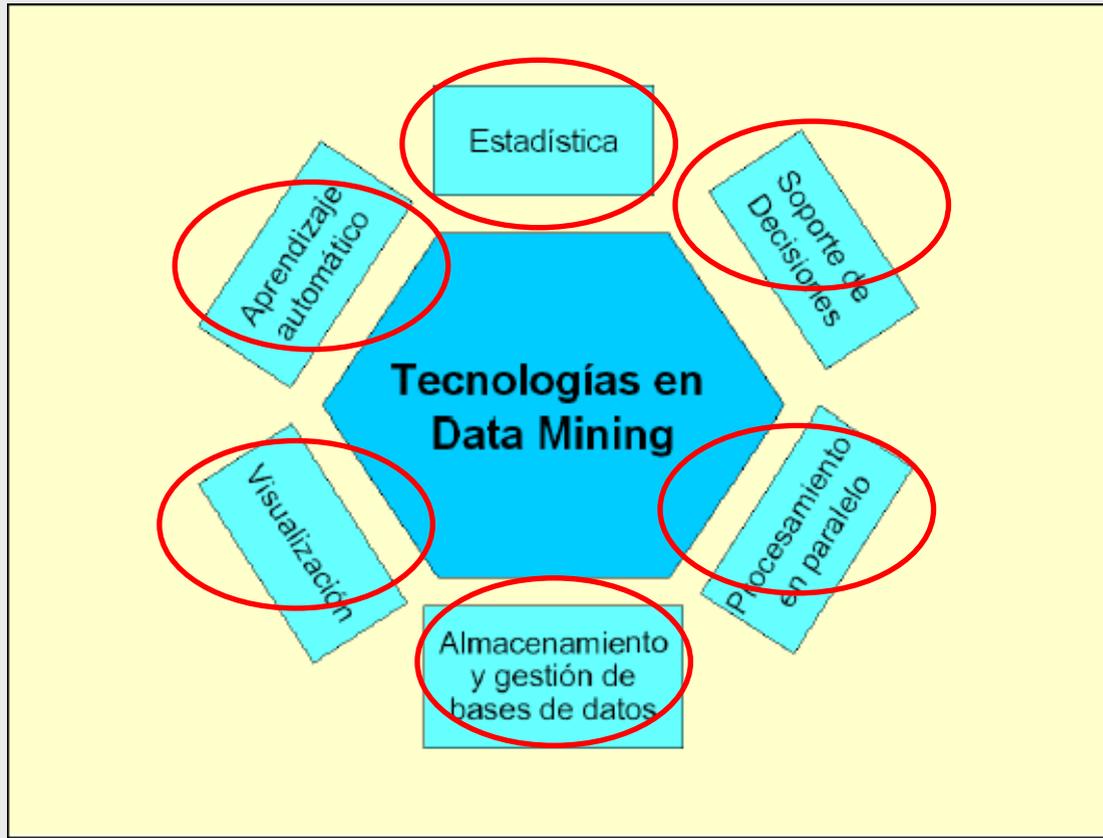
MEJORA DE RESPUESTA DE MAILINGS

CAMPAÑAS DE ADQUISICIÓN DE CLIENTES



¿DE DONDE SURGE EL DATA-MINING ?

De la integración múltiple...



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

APROBAR CRÉDITOS:

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

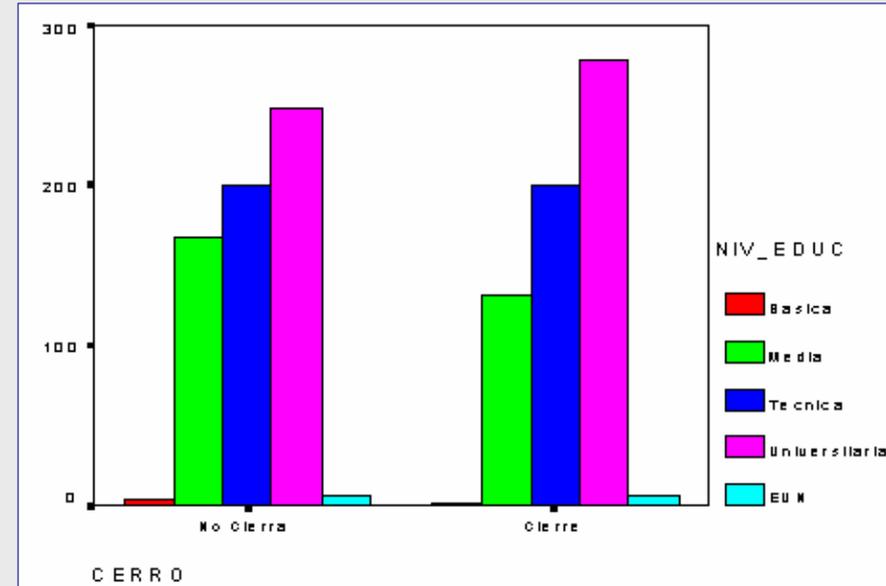
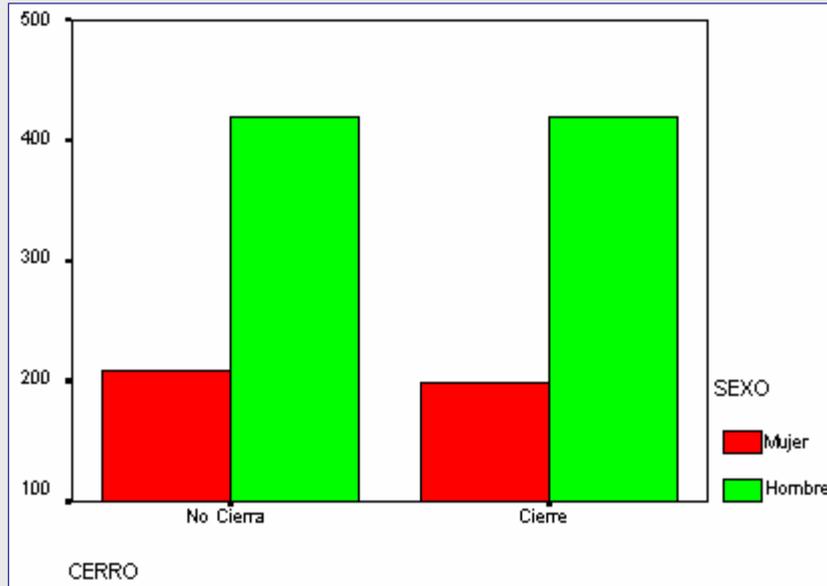
GESTIÓN DE PORTAFOLIO:

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

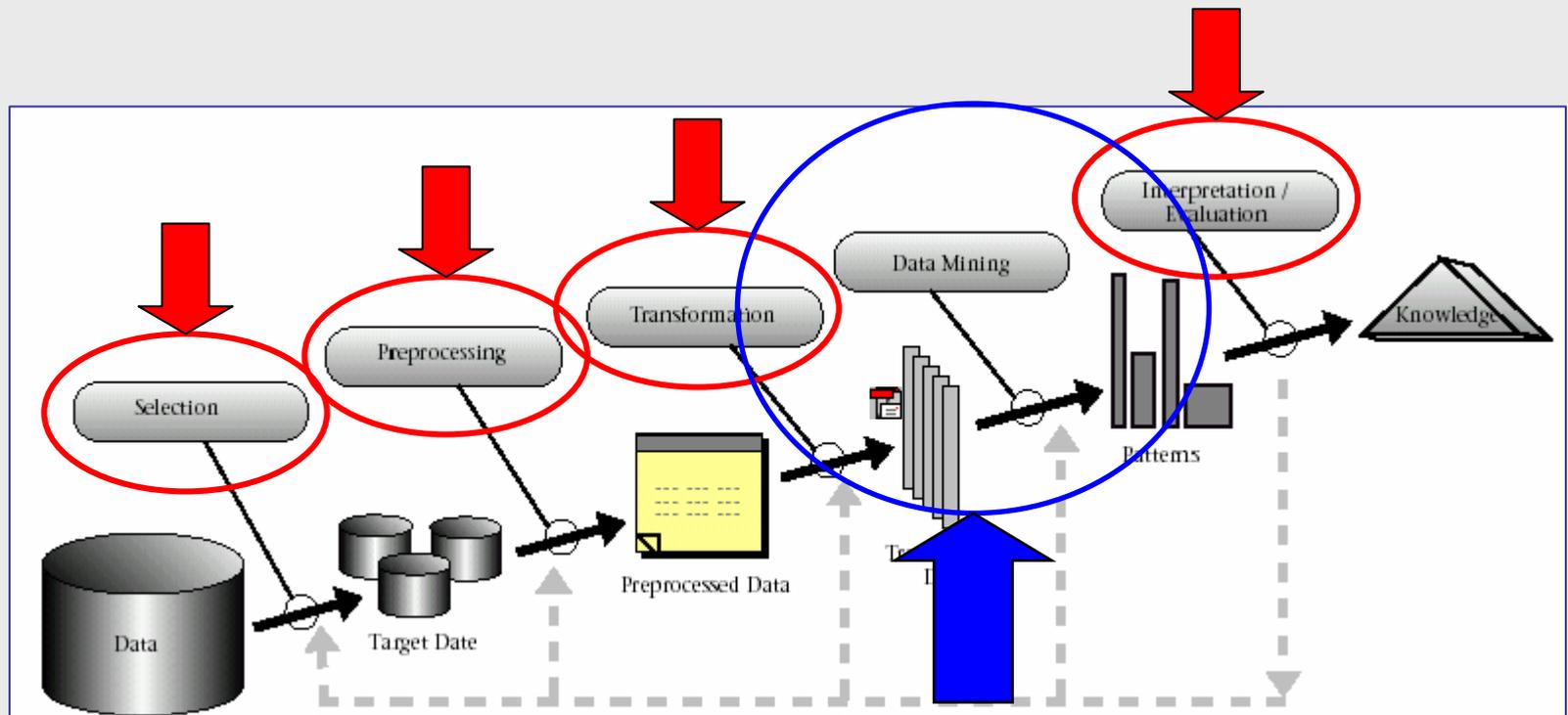
ANÁLISIS DE WEBSITES:

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

UN PEQUEÑO EJEMPLO ...

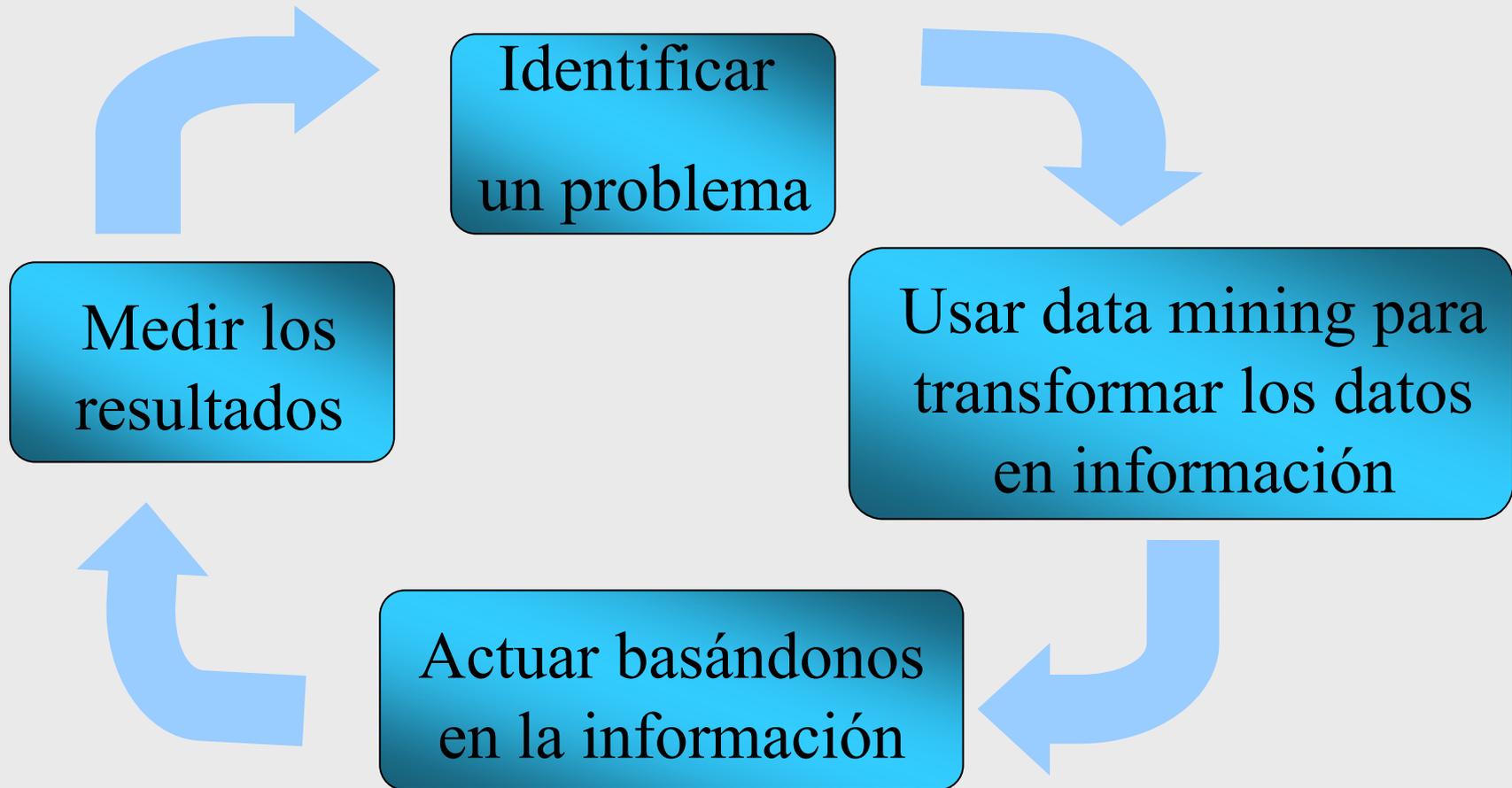


¿DÓNDE ESTAMOS PARADOS?

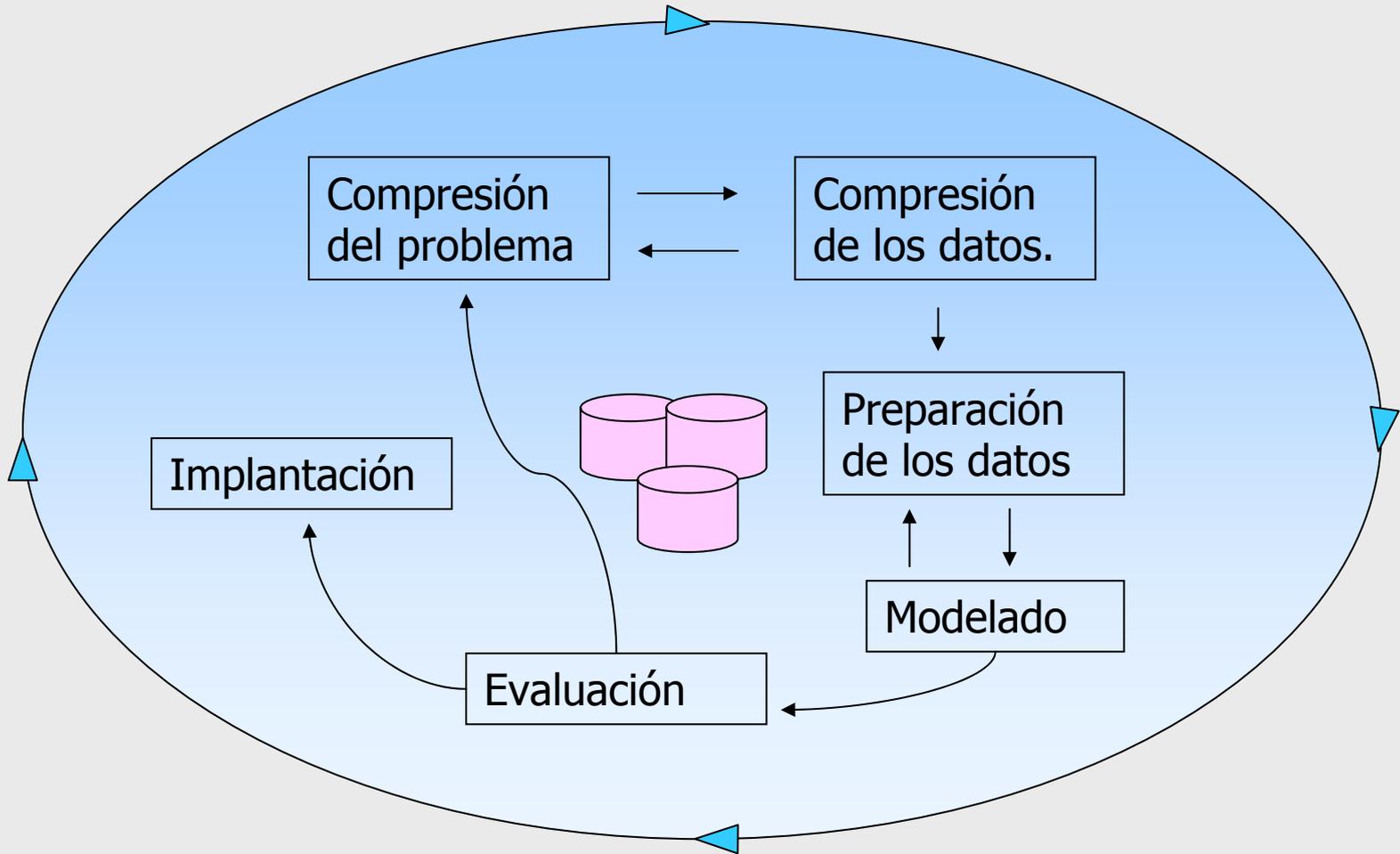


80% tiempo

EL CICLO DE DATA MINING



MÁS EN DETALLE ...



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.

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.

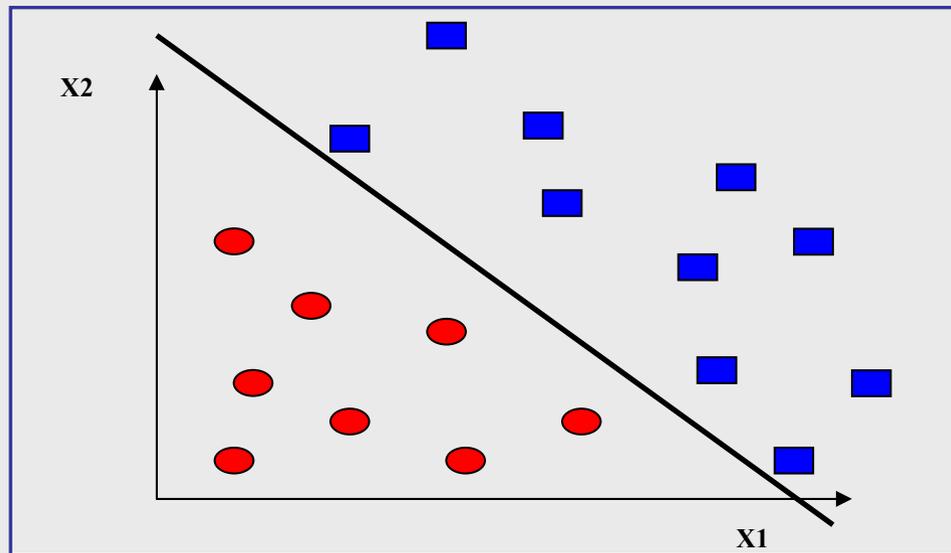
Se tiene siempre un número limitado de clases y se espera poder asignar cualquier nuevo objeto en una de esas clases.

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.



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

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

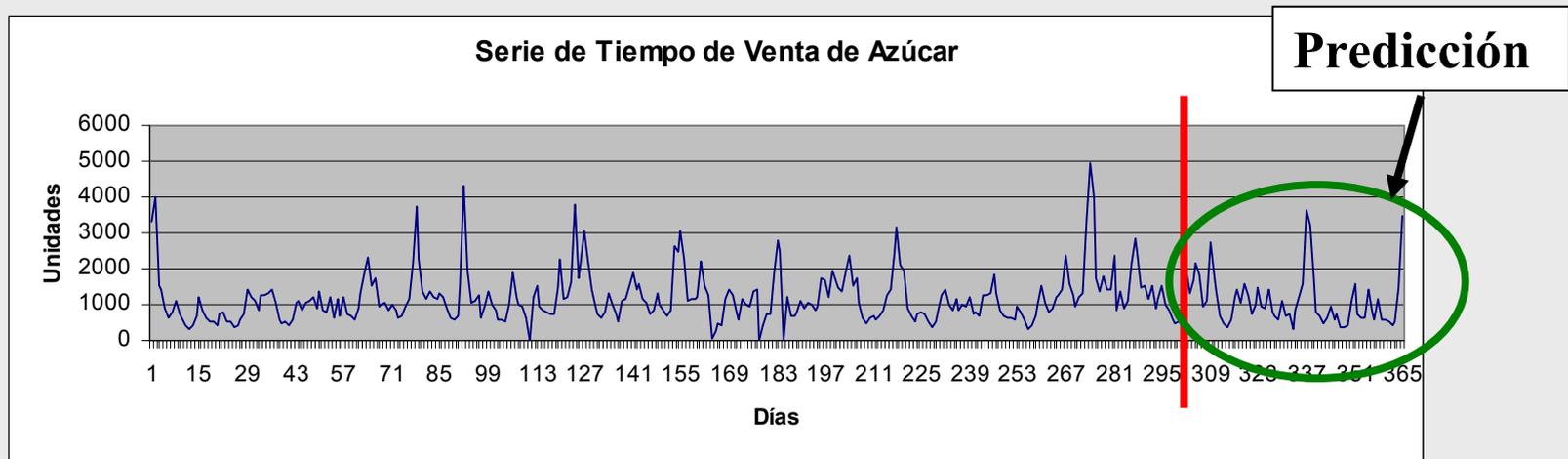
- Estimar el número de hijos de una familia
- Estimar la probabilidad de que alguien conteste a un mailing
- Estimar el tiempo de vida de un cliente
- Estimar los ingresos totales de una familia

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.



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.



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.

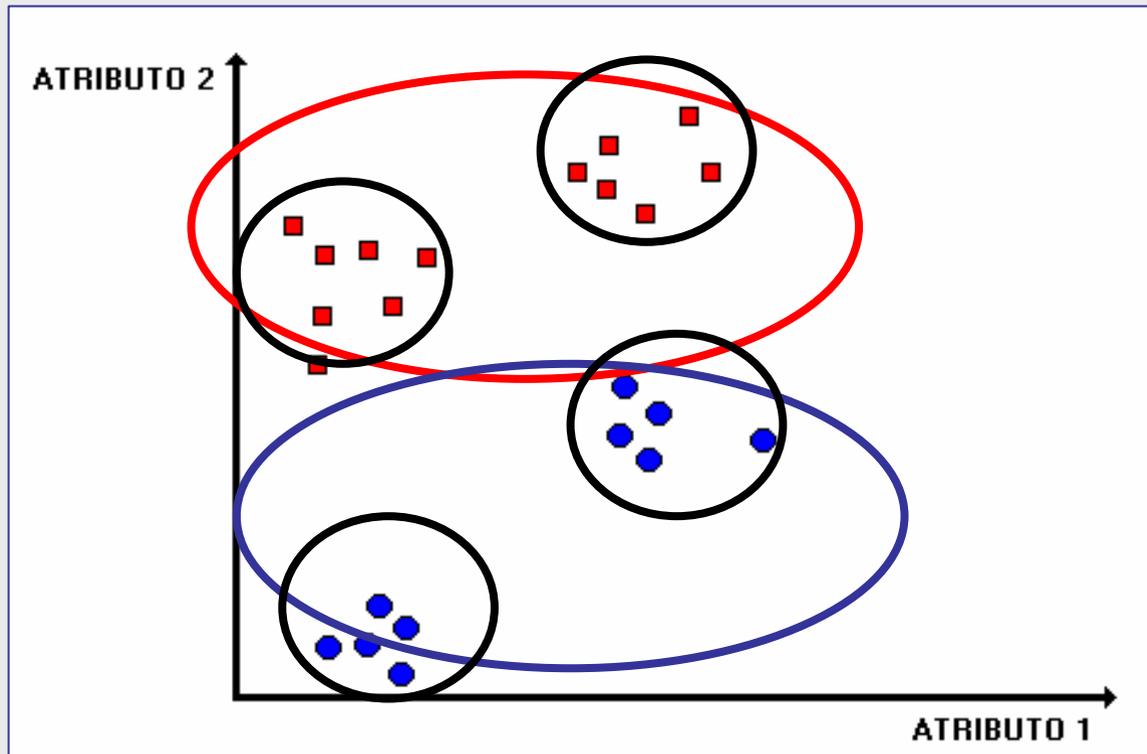
→ Encontrar clientes con hábitos de compra similares



PROBLEMAS DE SEGMENTACIÓN (2)

Encontrar patrones característicos no visibles a simple vista.

Encontrar soluciones entre subconjuntos o subpoblaciones.





Introducción a la Minería de Datos

DE LOS DATOS AL CONOCIMIENTO...

JAIME MIRANDA

(jmiranda@dii.uchile.cl)

Departamento de Ingeniería Industrial

Universidad de Chile