



Datamining Introducción

Yerko Halat
2 de Octubre del 2001

1

¿Cuál es la diferencia entre datos,
información y conocimiento?



- 3 ...es un dato
- 3 perros ...es un dato
- 3 perros guardianes ...es un dato
- 3 perros guardianes cuidando una casa en verano ...es información

2

¿Cuál es la diferencia entre datos,
información y conocimiento?



- 3 perros guardianes cuidando una casa en verano, **implica** que no hay moradores
...esto es conocimiento!!!!

Pues existen reglas de asociación no trivial

3

...los pañales y la cerveza...



- Una tienda estadounidense descubrió un patrón extraño: existía una alta correlación entre las ventas de pañales y cervezas los días Jueves de 18:00 a 22:00
- Acomodaron ambos juntos y la venta se acrecentó significativamente.

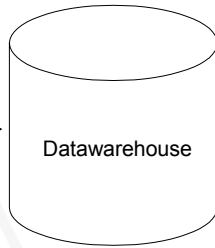
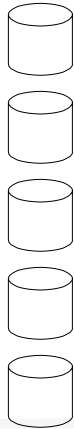
Esto es Datamining!!!!!!

4

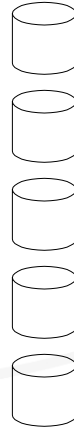
Modelo para la Gestión de la Información



Base de datos Operacionales



Datamarts



Reporting
Análisis
Olap
Cubos
InfoPortales
Consultas
Proyecciones
Tendencias

“Gestión de la Información”

5

Motivaciones para almacenar datos



Razones iniciales:	Potenciales:
En telecomunicación: Facturación de llamadas	En telecomunicación: Detección de fraude
En supermercados: Gestión del inventario	En supermercados: Asociación de ventas
En bancos: Manejo de cuentas	En bancos: Segmentación de clientes
En empresas de producción: Control de procesos	En empresas de producción: Mantención preventivo

6

Evolución del Marketing/Potencial del Conocimiento



(-) “Conocimiento” del cliente (+)



- 1º Etapa: Producto único sin diferenciación de clientes
- 2º Etapa: Diferenciación de productos basados en sus atributos
- 3º Etapa: Segmentación, Ofertas de productos en base a una diferenciación por clases de clientes
- 4º Etapa: Oferta de productos especializada en base a un **conocimiento específico** del cliente

7

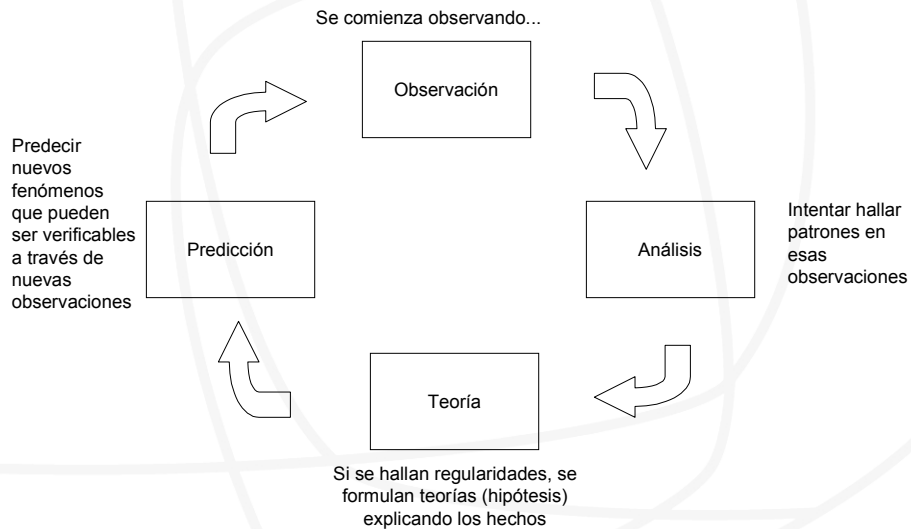
Idea básica y potenciales del Datamining



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

8

Proceso de Aprendizaje

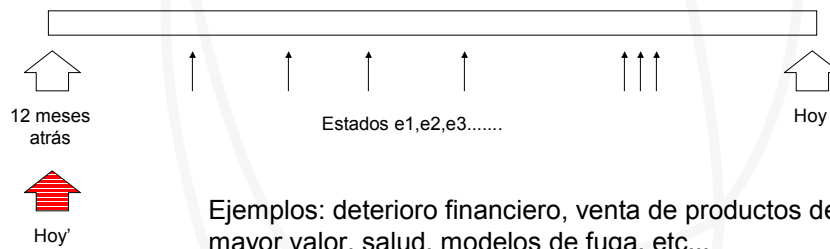


9

Proceso de Aprendizaje con "Consecuencia Conocida"



Con suficiente información es posible reconstruir el pasado para identificar los factores y elementos que implicaron un estado futuro conocido



10

Proceso 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“

11

KDD y el Negocio



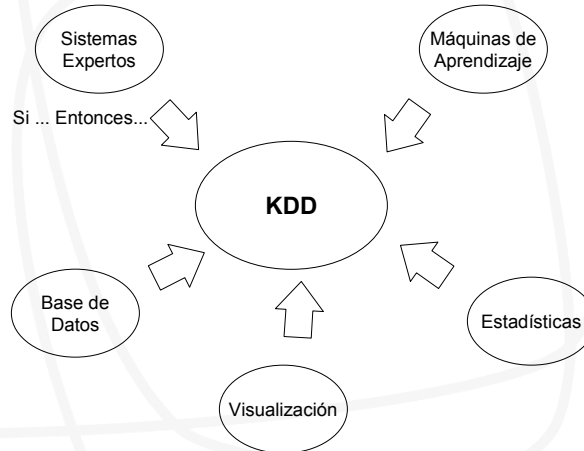
Debe contribuir al crecimiento y mejoramiento del negocio a través de **planes/acciones concretos.**

12

KDD y Datamining



Datamining es una técnica multidisciplinaria:



13

Aplicaciones del Datamining



- Customer Relationship Management (CRM)
 - Segmentación de clientes
 - Database Marketing
 - Predicción de compra
 - Retención de clientes
 - Predicción de fuga
- Detección de Fraude
 - Tarjetas de crédito
 - Uso de teléfonos (celulares)
- Predicción de series de tiempo

14

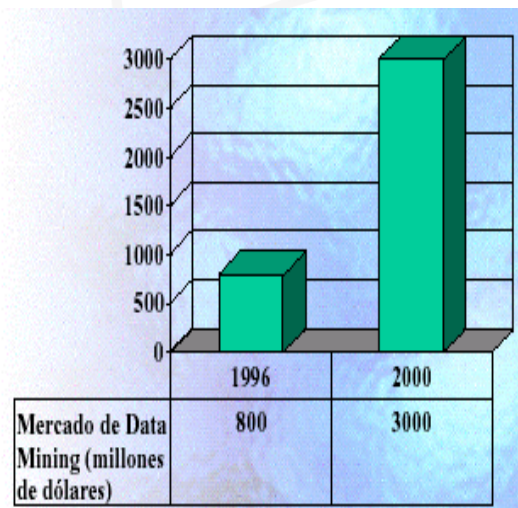
Métodos de Datamining



- Estadística
 - Agrupamiento (Clustering)
 - Análisis Discriminante
- Redes Neuronales
- Árboles de Decisión
- Reglas de Asociación
- Bayesian (Belief) Networks

15

Relevancia del Datamining



- Descubrimiento de herramientas y técnicas del datamining como una fuente de ventaja competitiva.
- Sin embargo, en un futuro cercano se tornará un factor crítico en vez de un factor clave

16

Datamining y Datawarehouse



- Para aplicar técnicas de datamining no es estrictamente necesario contar con datawarehouse, pero en la práctica es absolutamente habilitador.
- Las ventajas y beneficios de contar con un DWH como soporte son:
 - El tiempo como dimensión
 - No volátiles: no son actualizables, modificables ni borradas
 - Orientadas al asunto: modelamiento integral

17

Datamining y Datawarehouse



- Fuentes integradas, de las distintas bases operacionales
- Metadata, existe una máscara que cubre los tecnicismos de los programadores
- Riesgo de alterar alguna fuente operacional
- Costos (velocidad, complejidad, repetición, perfil de los analistas)

18

SQL (structured query language)



- Lenguaje de consulta de base de datos relacionales

The screenshot shows a database management system interface. The top part displays a query window with a SQL query. The bottom part shows a results table with columns for País, Apellidos, Nombre, IdPedido, and Subtotal. The table contains several rows of data.

```
SELECT DISTINCTROW Empleados.País,  
Empleados.Apellidos, Empleados.Nombre,  
Pedidos.IdPedido, Sum([Subtotales por  
pedido].Subtotal) AS ImporteVenta
```

```
FROM Empleados INNER JOIN (Pedidos  
INNER JOIN [Subtotales por pedido] ON  
Pedidos.IdPedido = [Subtotales por  
pedido].IdPedido) ON  
Empleados.IdEmpleado =  
Pedidos.IdEmpleado
```

```
GROUP BY Empleados.País,  
Empleados.Apellidos, Empleados.Nombre,  
Pedidos.IdPedido;
```

19

Datamining vs query tools (SQL)



- Son complementarias y no reemplazables
- SQL responde a preguntas preformuladas, datamining permite hallar patrones.
- En queries el conocimiento es aplicado/ratificado en base a la experiencia del analista. Datamining deduce conocimiento
- Datamining es optimizante

20

Datamining vs query tools (SQL)



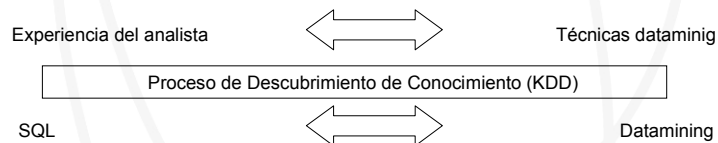
- Datamining permite relacionar un conjunto indefinido de variables. SQL maneja un número limitado de variables, pues la mente humana soporta 8 (+- 2) variables relacionadas simultáneamente.
- SQL es útil para pruebas aproximativas de análisis.
- Problemática del datamining: "garbage in, garbage out", que enturbian el KDD.

21

Datamining vs query tools (SQL)



- Conclusión
 - Desde el punto de vista del tipo y complejidad del análisis, la complementariedad se observa como un continuo:



22