TEMA 1: INTRODUCCIÓN A LA MINERÍA DE DATOS

¿Qué son los datos?

- Colecciones de objetos y sus atributos
- Un atributo es una propiedad o característica de un objeto
  - Ejemplo: edad, sexo, altura, estado civil, peso, etc…
  - Los atributos también son conocidos como variables, campos, características o aspectos
- Una colección de atributos describe un objeto
  - El objeto es también conocido como registro, punto, caso, ejemplo o instancia

¿Qué son los datos?

- Ejemplo de conjunto de datos

Atributos
Variables

Instancias
Objetos

CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

Tipos de atributos

- Existen diferentes tipos de atributos en función de los valores que pueden tomar
- Tipos
  - Atributos discretos: Tienen un número finito de valores o un conjunto numerable
    - Nominal
      - Ejemplos: sexo, país de origen, marca de coche
    - Ordinal
      - Ejemplos: nivel de consumo, nivel educativo
  - Numéricos (continuos): Su dominio son los números reales, normalmente de tipo punto flotante
    - Ejemplos: precio, peso, altura

Motivación de la ciencia de datos

El problema de la explosión de información:

- existencia de herramientas para la recolección de información
- madurez de la tecnología de bases de datos
- bajo precio del hardware

→ cantidades gigantescas de datos almacenados en bases de datos, *data warehouses* y otros tipos de almacenes de información

Somos ricos en datos pero pobres en conocimiento

El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

Ciencia de datos

- Data Science o la Ciencia de Datos incorpora diferentes elementos y se basa en las técnicas y teorías de muchos campos, incluyendo las matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelado de la incertidumbre, almacenamiento de datos y la informática de alto rendimiento con el objetivo de extraer el significado de datos y la creación de productos de datos.
- Es un término relativamente nuevo que se utiliza a menudo de manera intercambiable con analítica de negocio. La ciencia de datos busca utilizar todos los datos disponibles y relevantes para "extraer conocimiento" que pueda ser fácilmente comprendido por los expertos en el área de aplicación. Un experto de la ciencia de datos se denomina un científico de datos.

Ciencia de datos

José Antonio Guerrero: uno de los mejores científicos de datos del mundo (Plataforma Kaggle)

¿Qué es un científico de datos?

"Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos. De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco."

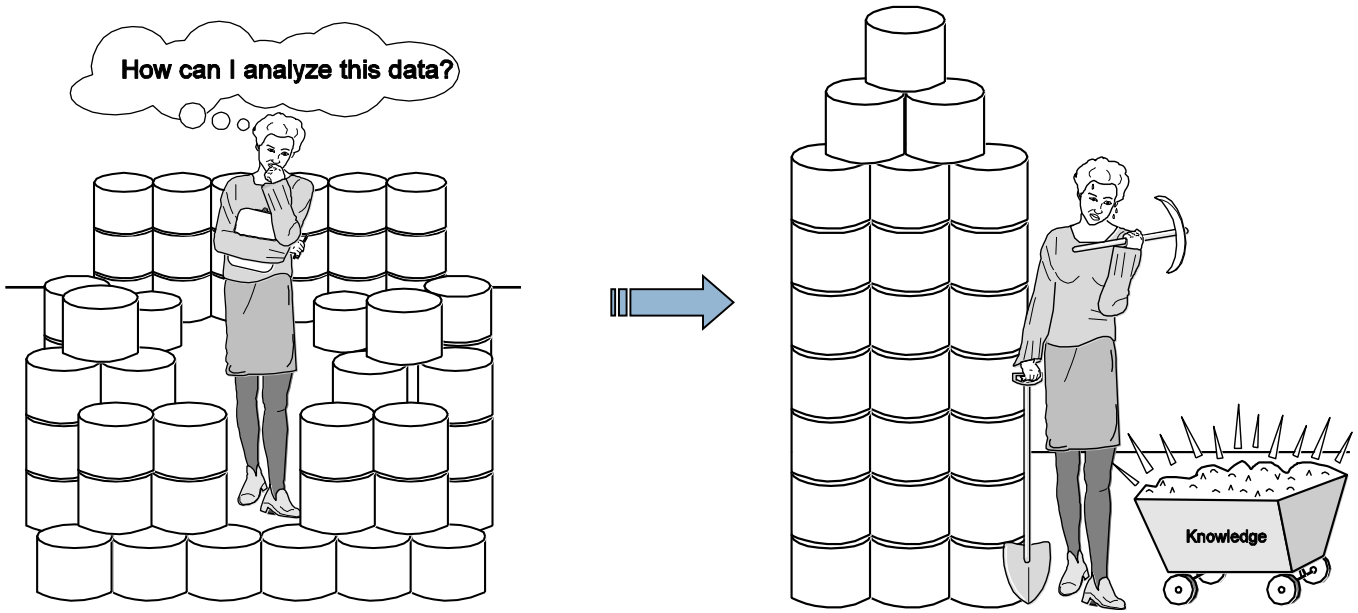
Leer más: http://www.elconfidencial.com/tecnologia/2013-12-19/un-matematico-andaluz-desconocido-es-el-mejor-cientifico-de-datos-del-mundo_67675/

Ciencia de datos

- Carácter interdisciplinar

Ciencia de datos: Minería de datos

- Metáfora



"Muchos datos,
Poca información"

Minería (búsqueda) de datos para extraer conocimiento de ellos (patrones interesantes)

¿Qué es la Minería de Datos?

- Proceso de extracción de conocimiento a partir de grandes cantidades de datos
  - No trivial
  - Implícito
  - Potencialmente útil
  - Previamente desconocido

Frawley, Piatetsky-Shapiro & Matheus:
Knowledge Discovery in Databases: An Overview.
MIT Press, 1991.

- Exploración y análisis de grandes cantidades de datos para descubrir patrones significativas utilizando medios automáticos o semi-automáticos

Berry & Linoff:
Data Mining Techniques.
Wiley, 1997

¿Qué es la Minería de Datos?

- Muchas de las técnicas utilizadas en MD ya se conocían previamente, ¿a qué se debe?
- En los 90's convergen los siguientes factores:
  1. Los datos se están produciendo
  2. Los datos se están almacenando
  3. La potencia computacional necesaria es abordable
  4. Existe una gran presión competitiva a nivel empresarial
  5. Las herramientas software de MD están disponibles

¿Qué es la Minería de Datos?

¿Para qué se utiliza el 'conocimiento' obtenido?

- Hacer predicciones sobre nuevos datos
- Explicar los datos existentes
  - Resumir una base de datos masiva para facilitar la toma de decisiones
  - Visualizar datos altamente dimensionales
  - …

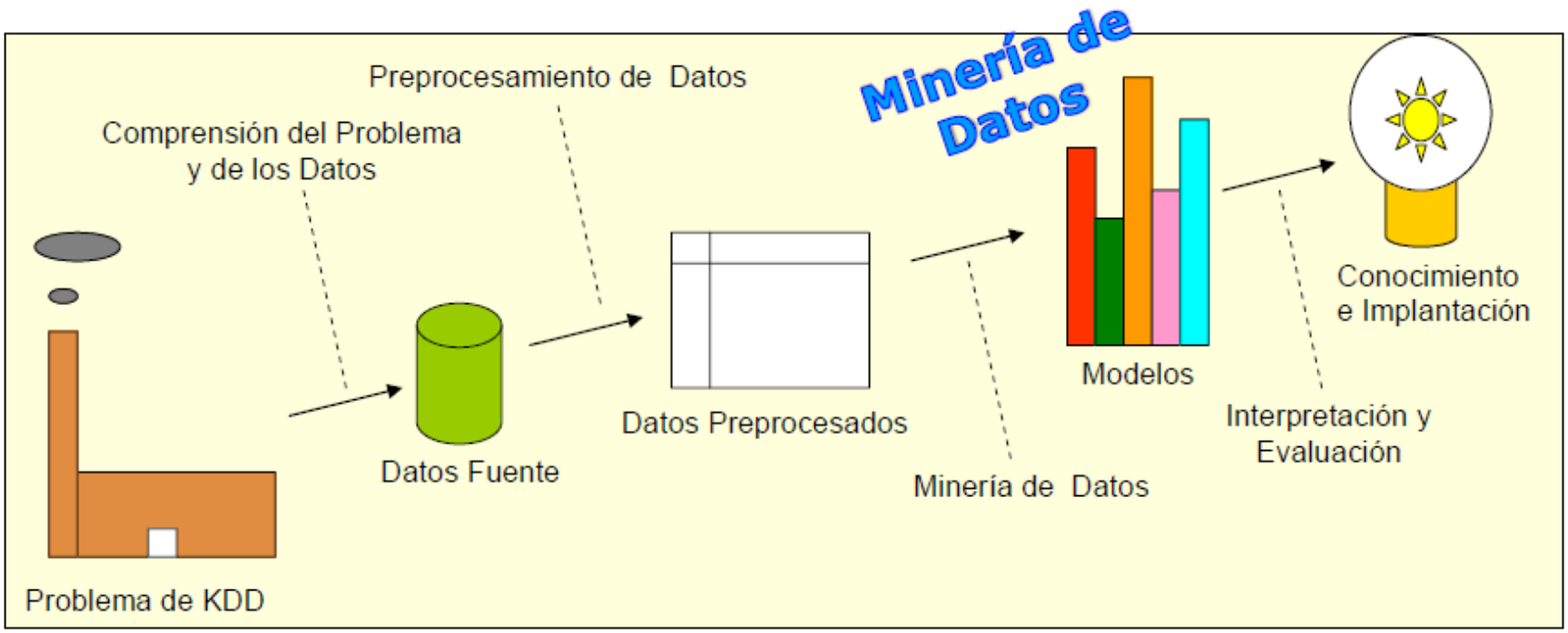
Nuevas necesidades de análisis datos

¿Qué es la Minería de Datos?

- KDD = *Knowledge Discovery from Databases*
- El KDD es el proceso completo de extracción de conocimiento a partir de bases de datos
- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La Minería de Datos es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

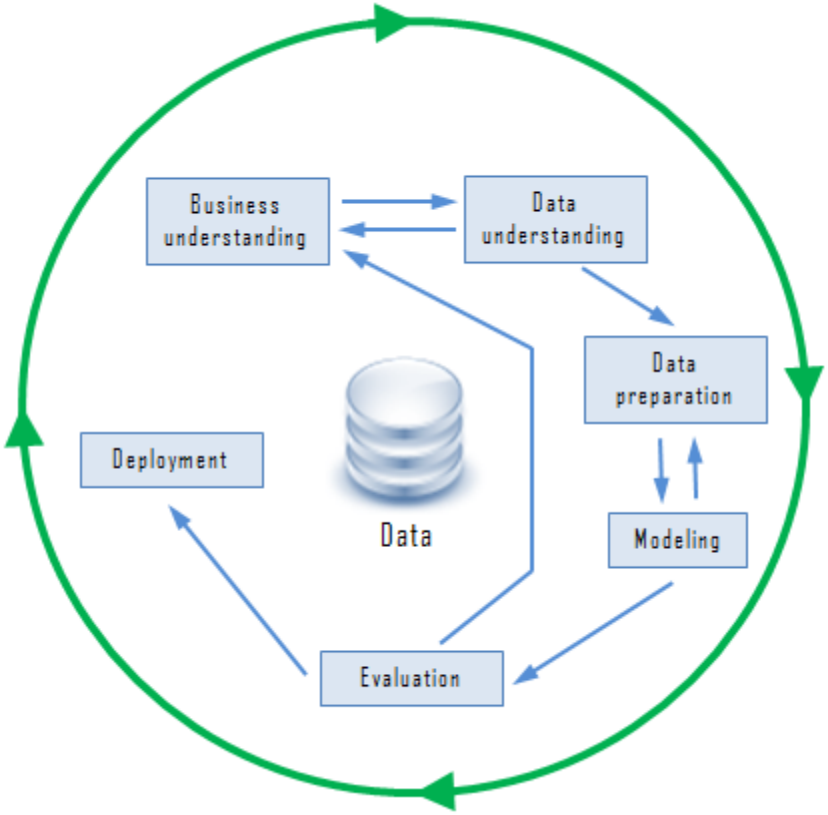
¿Qué es la Minería de Datos?

Etapas en un proceso de KDD



¿Qué es la Minería de Datos?

- KDD es un proceso iterativo



Minería de Datos. Tipos de datos

¿A qué tipos de datos puede aplicarse las técnicas de Minería de Datos?

En principio, a cualquier tipo

- Bases de datos relacionales
- Bases de datos espaciales
- Bases de datos temporales
- Bases de datos documentales (Text mining)
- Bases de datos multimedia
- World Wide Web (Web mining)
  - El almacén de información más grande y diverso de los existentes
  - Existe gran cantidad de datos de los que extraer información útil

…. Grandes volúmenes de datos: Big Data

¿Qué es la Minería de Datos?

- **La Minería de Datos es una forma de aprender del pasado para tomar mejores decisiones en el futuro**

- http://www.youtube.com/watch?v=9maeZ9slKwE

Minería de Datos. Aplicaciones

Análisis y gestión de mercados

- **Análisis de cestas de mercado**: asociaciones / co-relaciones entre ventas de producto, predicción basada en asociación de informaciones,…
- **Perfiles de cliente**: Identificar **qué tipo de clientes compra qué productos** (clustering y/o clasificación), usar predicción para encontrar **factores que atraigan nuevos clientes, retención de clientes,…**

Minería de Datos. Aplicaciones

Web mining / minería de datos web

- La mayoría de las herramientas actuales analizan los ficheros .log y generan estadísticas, pero ningún conocimiento acerca de las características del cliente ni de su comportamiento
- Minería de datos web en un sitio de e-comercio, generaría **análisis del comportamiento y perfiles del visitante**
- Lo que interesa es responder preguntas del tipo: **¿quién compra qué producto y en qué porcentaje?**
- Hay que capturar información en el servidor desde los **.log, cookies, formularios,** y completar con **información geográfica,** etc.,…
- En función de esto y de su actividad, generar perfiles de cliente y estudiar posibilidades de venta cruzada (*cross-selling*)

Minería de Datos. Aplicaciones

Análisis de riesgo en banca y seguros

- **Banca**
  - Detectar patrones de **uso fraudulento en tarjetas**
  - Estudio de **concesión de créditos** y/o tarjetas
  - Determinación del gasto en tarjeta por grupos
  - Identificar reglas de **comportamiento del mercado de valores** a partir de históricos
- **Seguros**
  - Predicción de **clientes propensos a suscribir nuevas pólizas**
  - Identificar grupos/patrones de riesgo
  - Identificar **tendencias de comportamiento fraudulento**
- **Ambos**: Identificación de clientes leales, identificación de fuga de clientes

Minería de Datos. Aplicaciones

Medicina / diagnóstico

- Identificación de terapias para diferentes enfermedades
- Estudio de **factores de riesgo en distintas patologías**
- **Segmentación de pacientes** en grupos afines
- Gestión hospitalaria y planificación temporal de salas, urgencias,…
- Recomendación priorizada de fármacos para una misma patología
- **Estudios en genética** (ADN,…)
- Selección de embriones en reproducción artificial

Minería de Datos. Aplicaciones

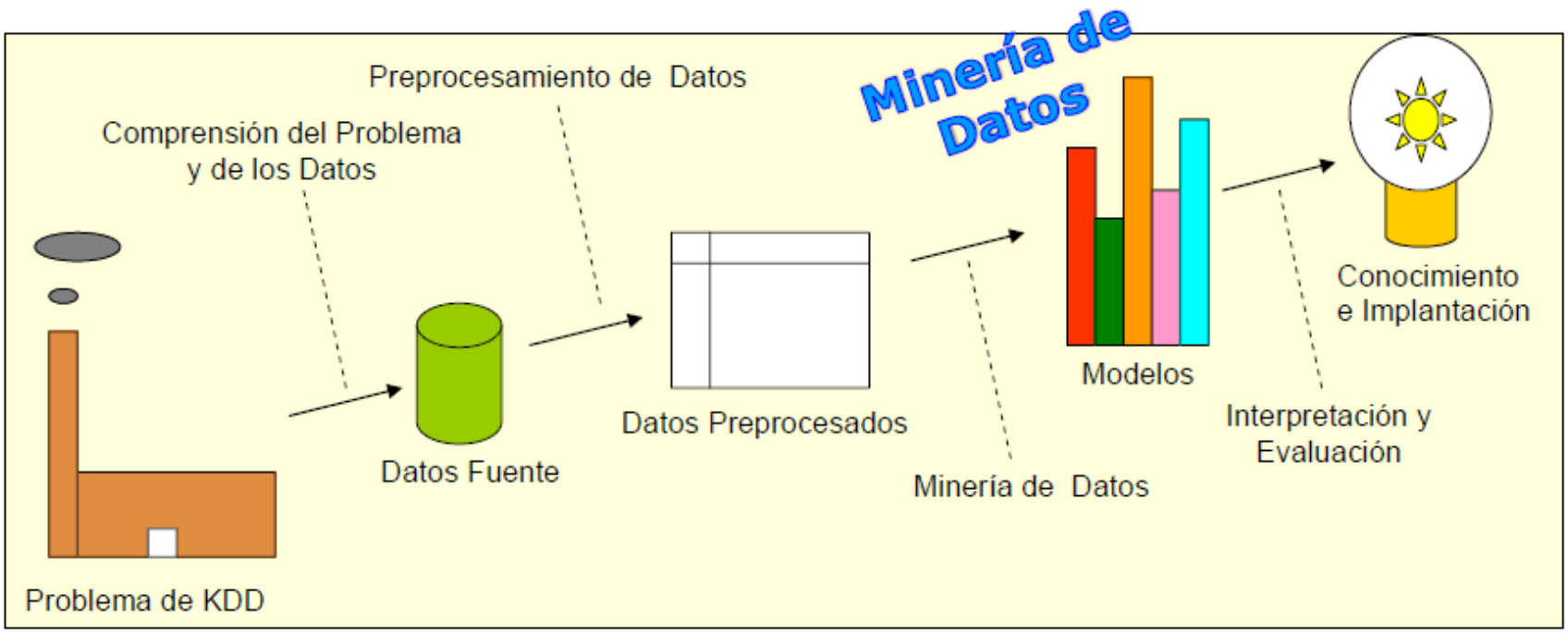
Minería de datos en industria

- **Control de calidad**
  - Detección precisa de productos defectuosos
  - Localización precoz de defectos
  - Identificación de causas de fallos

- **Procesos industriales**
  - Automatizar el control del proceso
  - Optimización del rendimiento de forma adaptativa
  - Implementar programas de mantenimiento predictivo

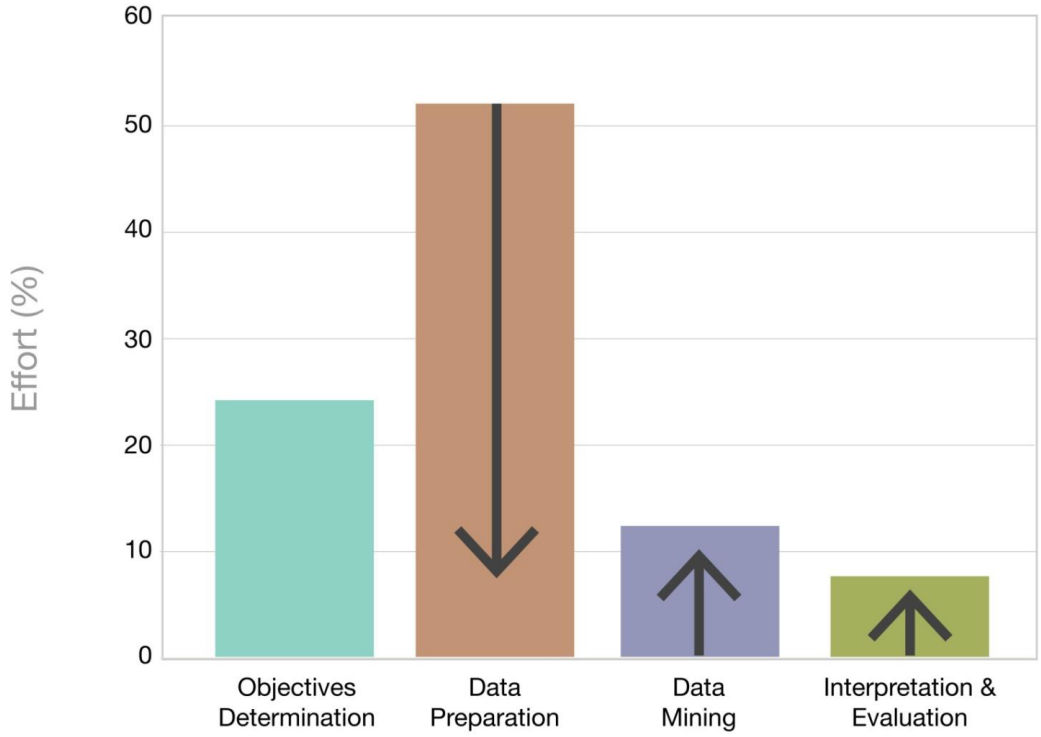
¿Qué es la Minería de Datos?

Etapas en un proceso de KDD



Etapas en el proceso de KDD

- Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos



Etapas en el proceso de KDD

- **Integración y recopilación**: Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori y creación del Datawarehouse
- **Preprocesamiento**: Selección de datos, limpieza, reducción y transformación
- **Selección de la técnica de MD** y aplicación de algoritmos concretos de MD
- **Evaluación**, interpretación y presentación de los resultados obtenidos
- **Difusión** y utilización del nuevo conocimiento

Integración y recopilación

- La familiarización con el dominio del problema y la obtención de conocimiento a priori disminuye el espacio de soluciones posibles

  → más eficiencia en el resto del proceso

- En problemas de KDD se suele trabajar con datos de diferentes departamentos de una entidad

  → es conveniente agrupar y unificar la información

- Unificación de la información en un Datawarehouse (DW) a partir de:
  - Información interna: distintas BBDD diseñadas para trabajo transaccional y de otro tipo (hojas de cálculo, informes,…)
  - Estudios publicados (demografía, catálogos, páginas, …)
  - Otras bases de datos (compradas, industrias/empresas afines,…)
- El resto del proceso de KDD será más cómodo si la fuente de datos está unificada, es accesible y dedicada (desconectada del trabajo transaccional)
- El DW es conveniente para KDD aunque no imprescindible. A veces se trabaja directamente con la BD o con las BBDD en formatos heterogéneos

Importancia del preprocesamiento

- 40% de los datos impuros
  - Sin limpiar (pre-procesar) los datos la calidad del conocimiento (patrones/reglas) obtenido por las técnicas de data mining se reduce en gran medida (poco útil)
- Limpiarlos por humanos
  - Muchas personas
  - Tarea laboriosa y complicada
  - Mucho tiempo
  - Suele llevar a errores
- Objetivo general del preprocesamiento: seleccionar/tratar el conjunto de datos adecuado para el resto del proceso de KDD

Importancia de la Preparación de Datos

- Tipos de impurezas (errores) de los datos
  - Valores inconsistentes (inexactos): valores para los que no se ha comprobado su validez
  - Ejemplos redundantes: ejemplos duplicados
  - Valores incompletos (valores perdidos):
    - Desconocidos
    - No almacenados
    - Irrelevantes
  - Valores "Outliers": valores fuera del rango habitual de la variable en estudio
  - Valores con ruido

Minería de datos

- **Objetivo**: Producir nuevo conocimiento que pueda utilizar el usuario
- **¿Cómo?** Construyendo un modelo, basado en los datos recopilados, que sea una descripción de los patrones y relaciones entre los datos con los que se puedan hacer predicciones, entender mejor los datos o explicar situaciones pasadas
- **Decisiones a tomar:**
  - ¿Qué tipo de conocimiento buscamos?
    - Predictivo, Descriptivo
  - ¿Qué técnica es la más adecuada?
    - Clasificación, Regresión, Clustering, Asociaciones, …
  - ¿Qué tipo de modelo?
    - P.e. Clasificación: reglas, árboles de decisión, SVM, etc.
  - ¿Qué algoritmo es el más adecuado?

Evaluación, interpretación y presentación de resultados

- La fase de MD puede producir varias hipótesis de modelos
- Es necesario establecer qué modelos son los más válidos
- **Criterios**: los patrones descubiertos deben ser
  - precisos,
  - comprensibles, e
  - interesantes (útiles, novedosos)
- **Técnicas de evaluación**: Al menos se divide el conjunto de datos en dos (entrenamiento y test)
  - Entrenamiento: Para extraer el conocimiento
  - Test: Para probar la validez del conocimiento extraído
  - Alternativas:
    - Validación simple
    - n-validación cruzada
    - *Bootstrapping*,…

Evaluación, interpretación y presentación de resultados

- **Medidas de evaluación de modelos**: Dependen de la tarea:
  - Clasificación: precisión predictiva (%acierto)
  - Regresión: Error cuadrático medio
  - Agrupamiento: Medidas de cohesión y separación entre grupos
  - Reglas de asociación: cobertura, confianza…

- La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integración, ventajas colaterales,…) ayuda a la selección del modelo(s) final(es)

Difusión y utilización del nuevo conocimiento

Una vez construido y validado el modelo puede utilizarse:

- para recomendar acciones
- para aplicar el modelo a diferentes conjuntos de datos

En cualquier caso, es necesario:

- **Difusión**: Elaboración de informes para su distribución
- **Utilización** del nuevo conocimiento de forma independiente
- **Incorporación** a sistemas ya existentes

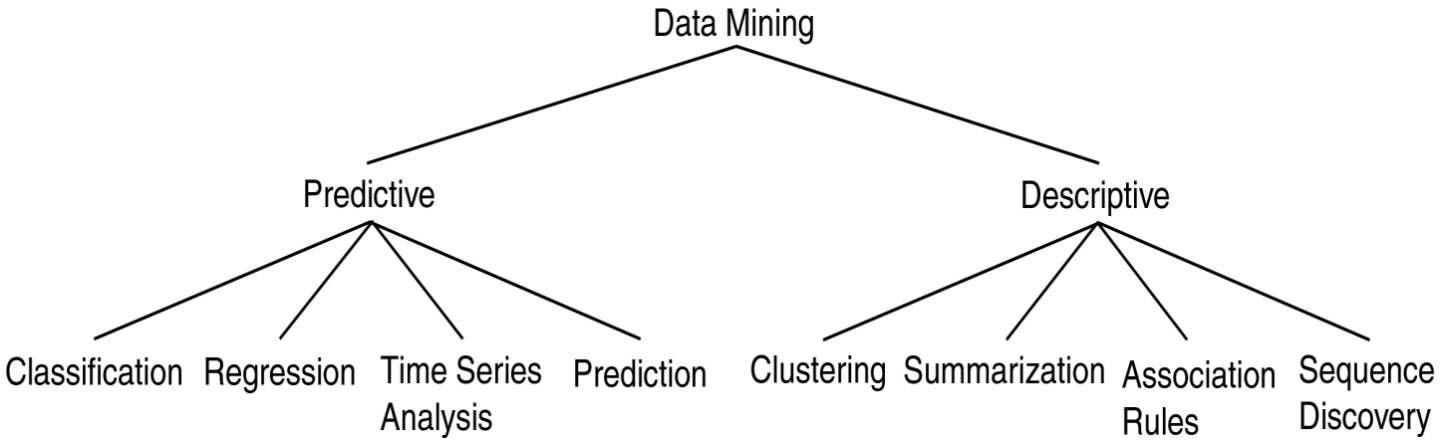
comprobar con el conocimiento ya utilizado para evitar inconsistencias y posibles conflictos

La monitorización del sistema en acción dará lugar a nuevos casos que realimentarán el ciclo del KDD

Las conclusiones iniciales pueden variar, invalidando el modelo adquirido

Objetivos de los métodos de Minería de Datos

- **Métodos predictivos**
  - Se utilizan algunas variables para predecir valores desconocidos de otras variables
- **Métodos descriptivos**
  - Encuentran patrones interpretables que describen los datos



Aprendizaje Supervisado vs No Supervisado

- **Aprendizaje supervisado**: Aprende, a partir de un conjunto de instancias pre-etiquetadas un método para predecir (Ejemplo, clasificación: la clase a que pertenece una nueva instancia)

	A	B	C	D	E
1	**cielo**	**Temperatura**	**Humedad**	**Viento**	**JuegaTenis**
2	soleado	calor	alta	débil	no
3	soleado	calor	alta	fuerte	no
4	nublado	calor	alta	débil	si
5	lluvioso	cálido	alta	débil	si
6	lluvioso	fresco	normal	débil	si
7	lluvioso	fresco	normal	fuerte	no
8	nublado	fresco	normal	fuerte	si
9	soleado	cálido	alta	débil	no
10	soleado	fresco	normal	débil	si
11	lluvioso	cálido	normal	débil	si
12	soleado	cálido	normal	fuerte	si
13	nublado	cálido	alta	fuerte	si
14	nublado	calor	normal	débil	si
15	lluvioso	calor	alta	fuerte	no

Aprendizaje Supervisado vs No Supervisado

- **Aprendizaje no supervisado**: **No hay conocimiento a priori sobre el problema, no hay instancias etiquetadas, no hay supervisión sobre el procedimiento**. **(Ejemplo, clustering: Encuentra un agrupamiento de instancias "natural" dado un conjunto de instancias no etiquetadas)**

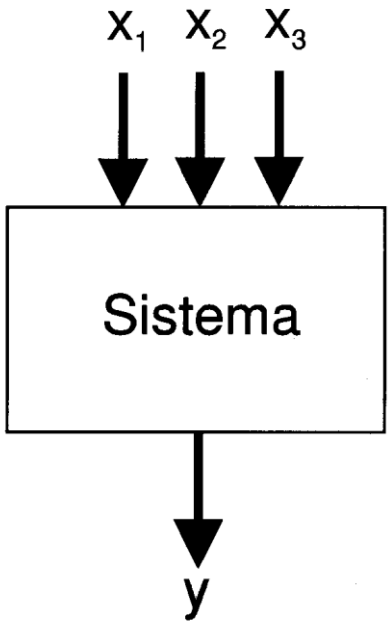
	A	B	C	D
1	**cielo**	**Temperatura**	**Humedad**	**Viento**
2	soleado	calor	alta	débil
3	soleado	calor	alta	fuerte
4	nublado	calor	alta	débil
5	lluvioso	cálido	alta	débil
6	lluvioso	fresco	normal	débil
7	lluvioso	fresco	normal	fuerte
8	nublado	fresco	normal	fuerte
9	soleado	cálido	alta	débil
10	soleado	fresco	normal	débil
11	lluvioso	cálido	normal	débil
12	soleado	cálido	normal	fuerte
13	nublado	cálido	alta	fuerte
14	nublado	calor	normal	débil
15	lluvioso	calor	alta	fuerte

Problemas de Minería de Datos

- Tipos de problemas:
  - Regresión (Predictiva)
  - Clasificación (Predictiva)
  - Series temporales (Predictiva)
  - Caracterización o resumen (Descriptiva)
  - Reglas de asociación (Descriptiva)
  - Clustering (Descriptiva)
  - Detección de anomalías / desviaciones (Descriptiva)
  - Descubrimiento de patrones secuenciales (Descriptiva)

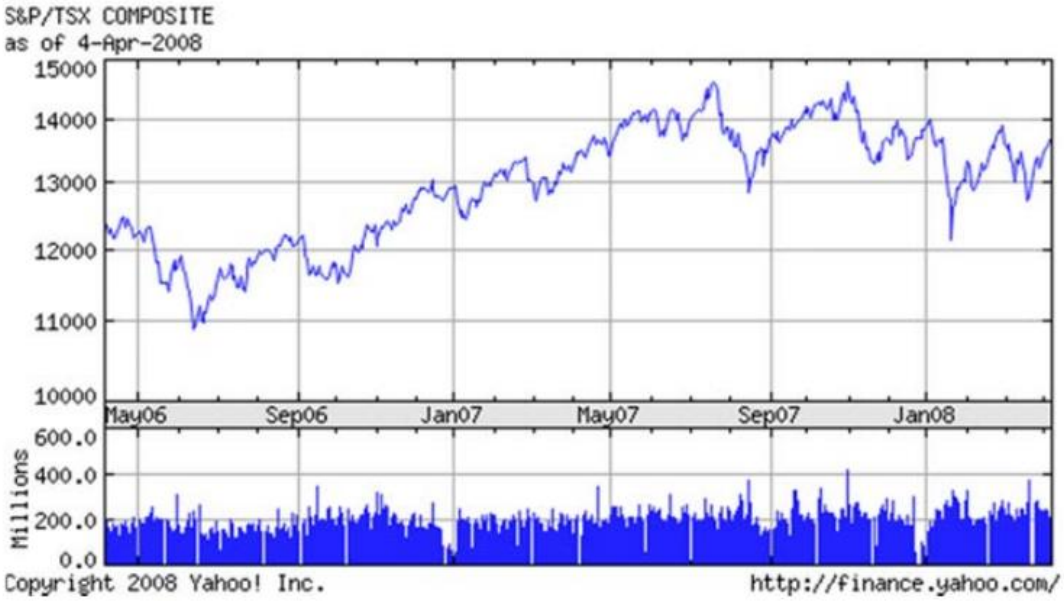
Regresión

- **Regresión**: El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable de control (continuo)



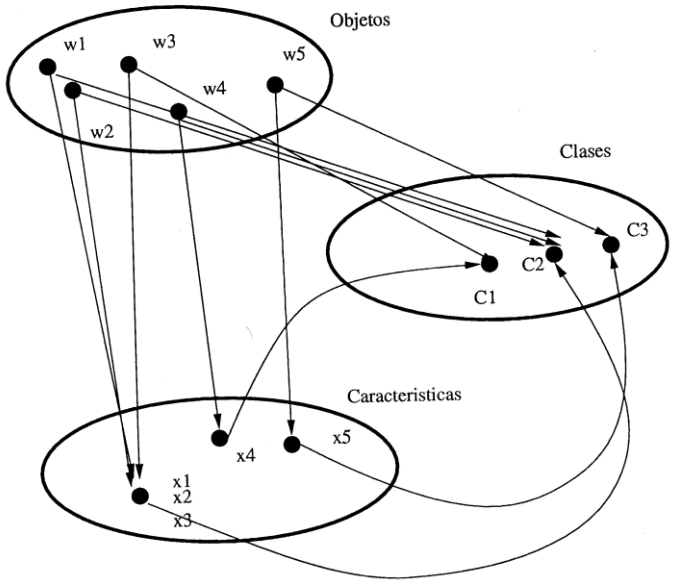
Ejemplo: predicción del valor de las acciones

- Problema de regresión (supervisado)
- Dada una base de datos con transacciones pasadas
  - Obtener las variables más influyentes en el precio de venta de las acciones
  - Aprender un modelo que permita predecir el valor de las acciones
- Objetivo: obtener el menor error posible en la predicción del valor



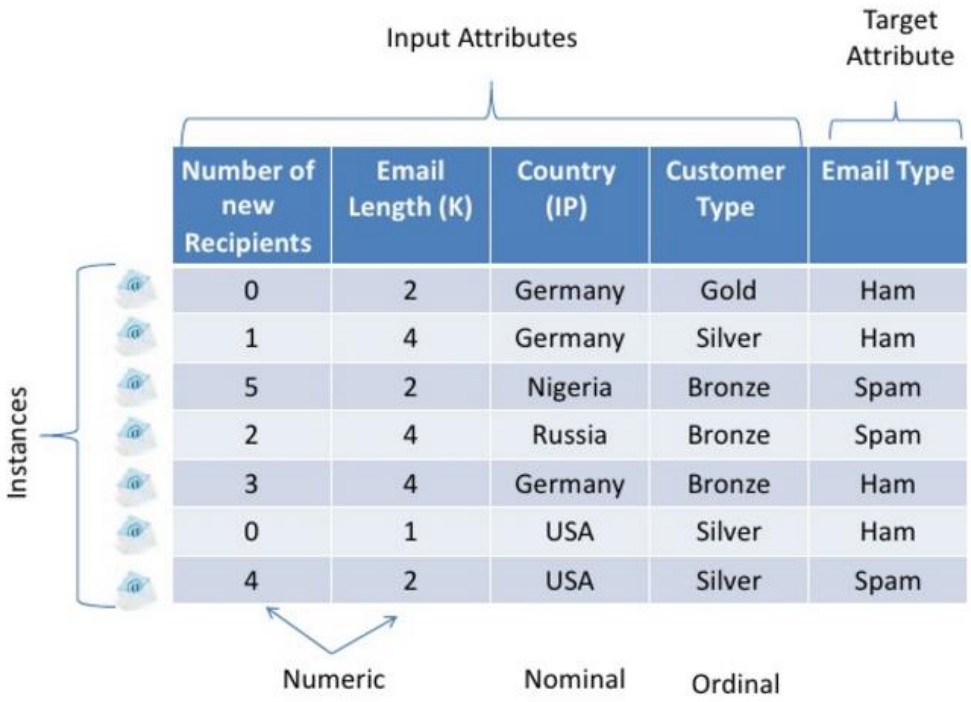
Clasificación

- **Clasificación:** El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable de control (discreto).
  - El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.



Ejemplo: filtro de spam

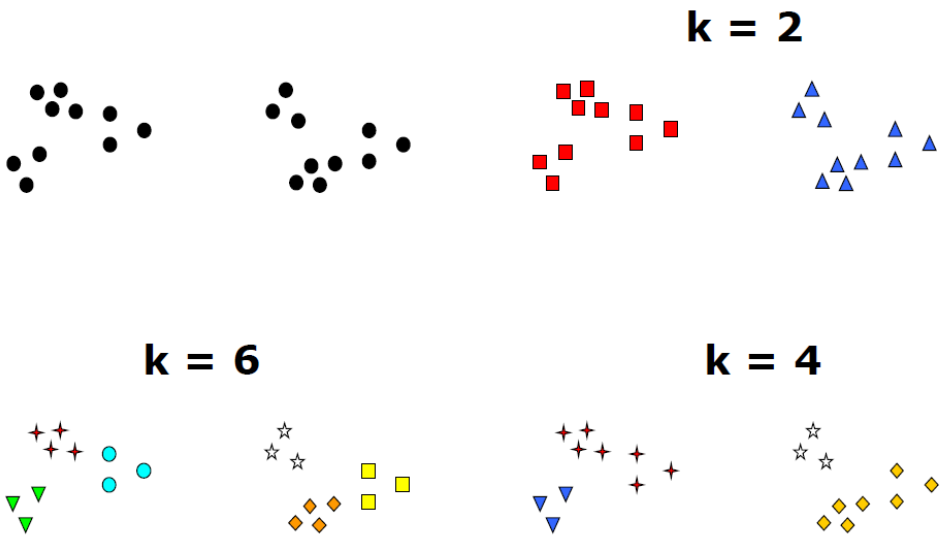
- Problema de clasificación supervisada
- Dada una base de datos con información de correos electrónicos
  - Procesar la información del servidor de correo para obtener el conjunto de datos
- Objetivo: obtener la mejor tasa de acierto posible



Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Clustering

- **Clustering**: Hay problemas en los que deseamos agrupar las instancias creando clústeres de similares características
  - **Objetivo:** Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos [*clústeres*].
  - La decisión del número de clústeres es uno de los retos en agrupamiento



Reglas de asociación

Descubrimiento de reglas de asociación

- Búsqueda de patrones frecuentes, asociaciones, correlaciones, o estructuras causales entre conjuntos de artículos u objetos (datos) a partir de bases de datos transaccionales, relacionales y otros conjuntos de datos
- **Objetivo**: determinar grupos de ítems que tienden a ocurrir juntos en transacciones (=tickets de compra pagados con o sin tarjeta)

Id	*Productos*
1	Pan, Coca-cola, Leche
2	Cerveza, Pan
3	Cerveza, Coca-cola, Pañal, Leche
4	Cerveza, Pan, Pañal, Leche
5	Coca-cola, Pañal, Leche

Reglas encontradas:

{Leche} --> {Coca-cola}
{Pañal, Leche} --> {Cerveza}

Minería de Datos. Aplicaciones

- Aplicaciones: *Market Basket analysis*
- Acciones a realizar:
  - Planificar disposiciones alternativas en el almacén
  - Limitar descuentos especiales a sólo uno de los dos productos que tienden a comprarse juntos
  - Poner los aperitivos que más margen dejan entre los pañales y las cervezas
  - Poner productos de bebé en oferta cerca de las cervezas
  - Ofrecer cupones descuento para el producto "complementario", cuando uno de los productos se venda por separado…

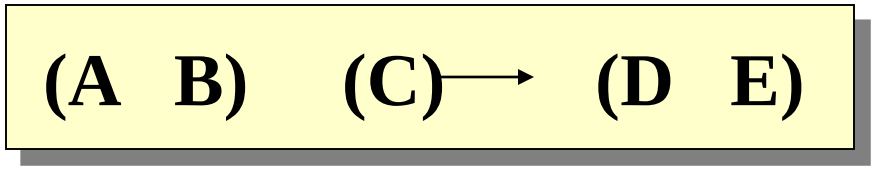
La proliferación de "tarjetas de lealtad" se debe al interés por identificar el historial de ventas individual del cliente…

Detección de Desviaciones/Anomalías

- **Detección de anomalías**
  - Método: detectar comportamientos que se salgan fuera de los parámetros normales.
  - Aplicaciones:
    - Detección de fraudes en tarjetas de crédito
    - Detección de intrusos en redes de ordenadores

Aplicaciones de la Minería de Datos

- **Patrones secuenciales**
  - Método: encontrar reglas que predigan dependencias secuenciales



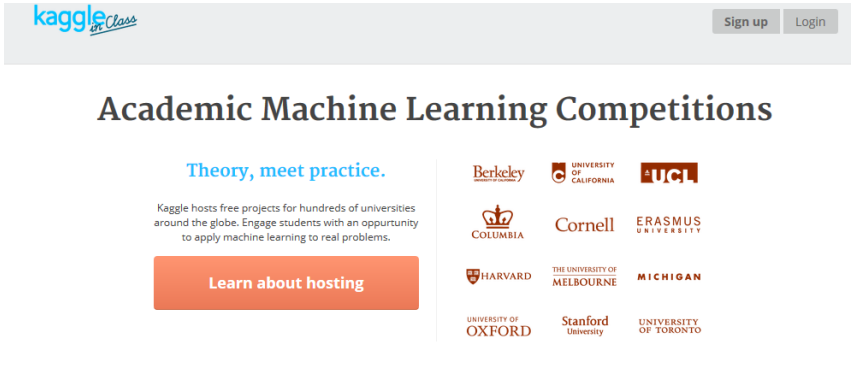
  - Aplicación: predicción de eventos

Herramientas, Lenguajes, Kaggle

- Herramientas
  - KNIME (o Konstanz Information Miner)
    - Plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual (Java)
    - Universidad de Constanza, Alemania
    - *https://www.knime.org/*
  - Weka
    - Machine learning software in Java implementation
    - The University of Waikato, New Zealand
    - http://www.cs.waikato.ac.nz/ml/weka/
  - KEEL
    - Machine learning software in Java implementation
    - University of Granada, España
    - http://www.keel.es/
  - Power BI: https://powerbi.microsoft.com/es-es/
  - RapidMiner: https://rapidminer.com/
  - Orange: http://orange.biolab.si/
- Lenguajes de programación a aprender por científicos de datos en 2018

Herramientas, Lenguajes, Kaggle

- *Kaggle*: The Home of Data Science
  - Es un portal web que ofrece competiciones, tutoriales, actividades académicas …
  - http://www.kaggle.com/



Comentarios

- Tendencias de empleo para 2019: uno de los perfiles con más demanda será el especialista en Machine Learning
- Los puestos de trabajo que se pondrán de moda en 2019: Analista y científico de datos
- 3 de los 10 trabajos con más futuro están relacionados con esta temática. El estudio está realizado por InfoJobs y ESADE y usa como fuentes datos de EPA, SEPE, Eurostat e INE
- El puesto mejor pagado del sector IT es el big data architect