UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

Técnicas de análisis de datos para automatización

Jose Aguilar

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

Indice

1. Introducción
2. Minería de Datos
3. Minería Semántica
4. Minería de Procesos
5. BigData

Everything Mining

- Datos espaciales
- Espacio-temporal
- objetivos en movimiento
- datos multimedia
- flujos de datos

Minería de datos

Minería de procesos

Minería de dominios: salud, control del tráfico aéreo, alimentos, energía,

- Minería de texto
- Minería web
- Minería de la Web Semántica
- Ontología Minería
- Minería de grafos
- Datos Vinculados

Minería Semántica

El Mundo de la Información y sus Problemas.

- **Cada vez se genera más información** y se hace más fácil el acceso masivo a la misma (existen gran cantidad de bases de datos on-line)
  - ✓ Transacciones bancarias, Internet y la Web, observaciones científicas (biología, altas energías, etc.) "tranNASA's EOS (Earth Observation System)".
- La **tecnología es barata** y los **sistemas de gestión de bases de datos** son capaces de trabjar con cantidades masivas de datos (Terabytes).

Los datos contienen información útil **"CONOCIMIENTO"** !!!

- Necesitamos **extraer** información (**conocimiento**) de estos datos:
  - ✓**Rapidez y confiabilidad.**
  - ✓**Capacidad de modelización y escalabilidad.**
  - ✓**Explicación e Interpretación de los resultados (visualización, …).**

WalMart captura transacciones de 2900 tiendas en 6 países. Esta información e acumula en una base de datos masiva de 7.5 terabyte. WalMart permite que más de 3500 proveedores accedan a los datos relativos a sus productos para realizar distintos análisis. Así pueden identificar clientes, patrones de compras, etc. En 1995, WalMart computers procesó más de un millón de consultas complejas.

Introducción

- La revolución digital ha permitido que la captura de datos sea fácil, y su almacenamiento tenga un costo casi nulo.

- Enormes cantidades de datos son recogidas y almacenadas en BD en la vida diaria.

Resultado: Para analizar estas enormes cantidades de datos, las herramientas tradicionales de gestión de datos y las herramientas estadísticas no son adecuadas.

Introducción

- Los datos por sí solos no producen beneficio directo. Su verdadero valor consiste en poder extraer información útil para la toma de decisiones.

- Tradicionalmente se analizaban datos con la ayuda de técnicas estadísticas (resumiendo y generando informes) o validando modelos sugeridos manualmente por los expertos.

Introducción

- Estos procesos son irrealizables a medida que aumenta el tamaño de los datos.

- Bases de datos con un nº de registros del orden de 10^9 y 10^3 de dimensión, son fenómenos relativamente comunes.

La tecnología informática puede automatizar este proceso. ➔ Minería de datos

Estadística vs Minería de datos

	Estadistica	Mineria de datos
Construcción de modelos	Ceñido a premisas y teoremas	Mayor libertad en la construcción, interpretable
Score	Verosimilitud de los datos dado el modelo	Más directo, PBC por ejemplo
Búsqueda	Test de la razón de la verosimilitud	Metaheurísticos
Transparencia	Más complicados de interpretar	Más claros y sencillos
Validación	No	Sí
Selección de variables	Filter	Wrapper

Problemas

- Recolección masiva de datos:

  aumento dimensionalidad y nº observaciones
  históricos
  imperfectos

- Análisis de datos es crucial para el negocio
- Toma decisiones rápidas
- Dificultad para aplicar técnicas tradicionales
- Solamente un 5 % de la información es analizada
- Potentes computadoras con multiprocesadores

¿Qué es la Minería de Datos?

- Es un mecanismo de explotación que consiste en la búsqueda de información valiosa en grandes volúmenes de datos.

- Ligada a las bodegas de datos (información histórica) con la cual los algoritmos de minería de datos obtienen información necesaria para la toma de decisiones.

Definición de minería de datos

- Minería de datos es la exploración y análisis de grandes cantidades de datos con el objeto de encontrar patrones y reglas significativas (conocimiento)

¿Qué es la Minería de Datos?

Una definición de Minería de datos es:

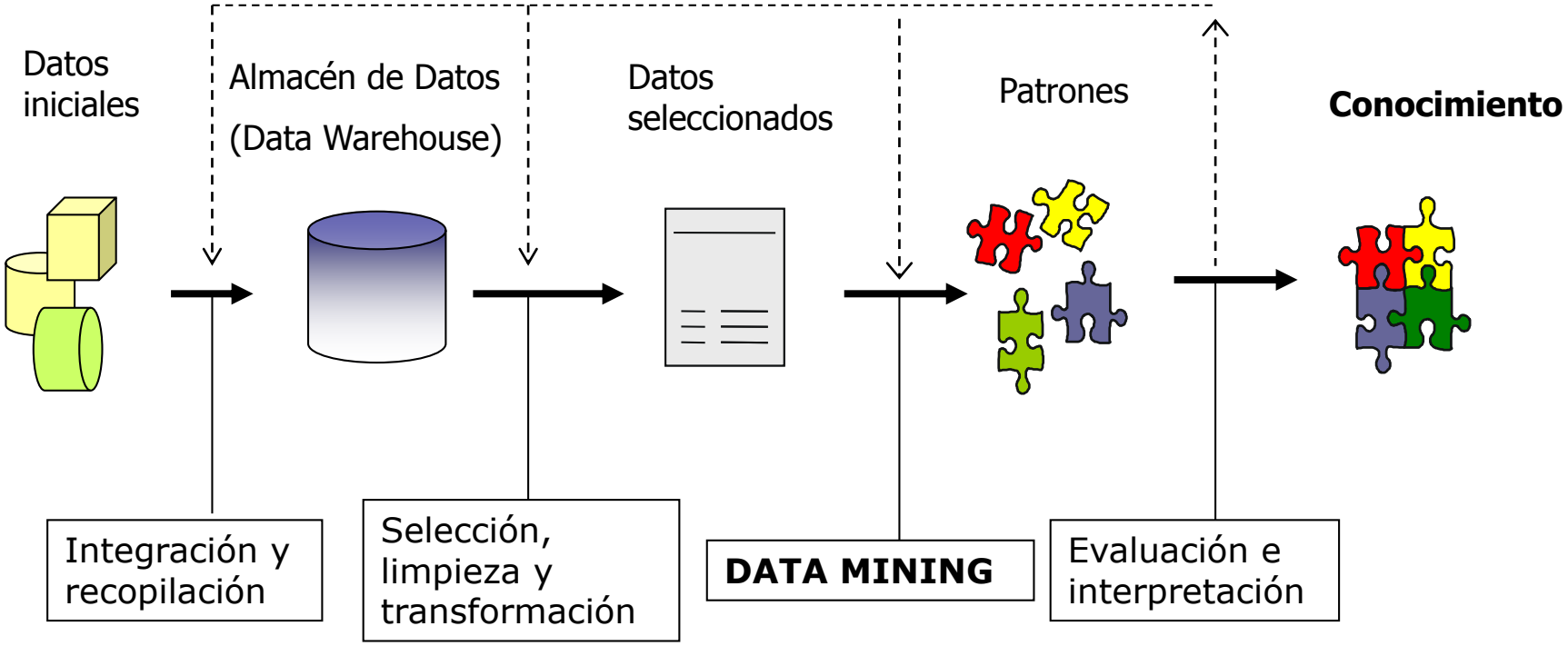
- "Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de obtención de patrones de los datos"

Un proceso más general es KDD (Knowledge Discovery on Databases/ Descubrimiento de conocimiento en Bases de Datos).

- *KDD* es empleado para describir el proceso de extracción de conocimiento de los datos.
- *Definición: "La extracción no-trivial de conocimiento implícito en los datos que resulte ser previamente desconocido y* potencialmente útil".
- El conocimiento debe ser nuevo, no obvio y debe estar disponible para el uso.

Knowledge Discovery from Databases

Proceso de KDD



¿Qué es la Minería de Datos?

- Definiciones necesarias:

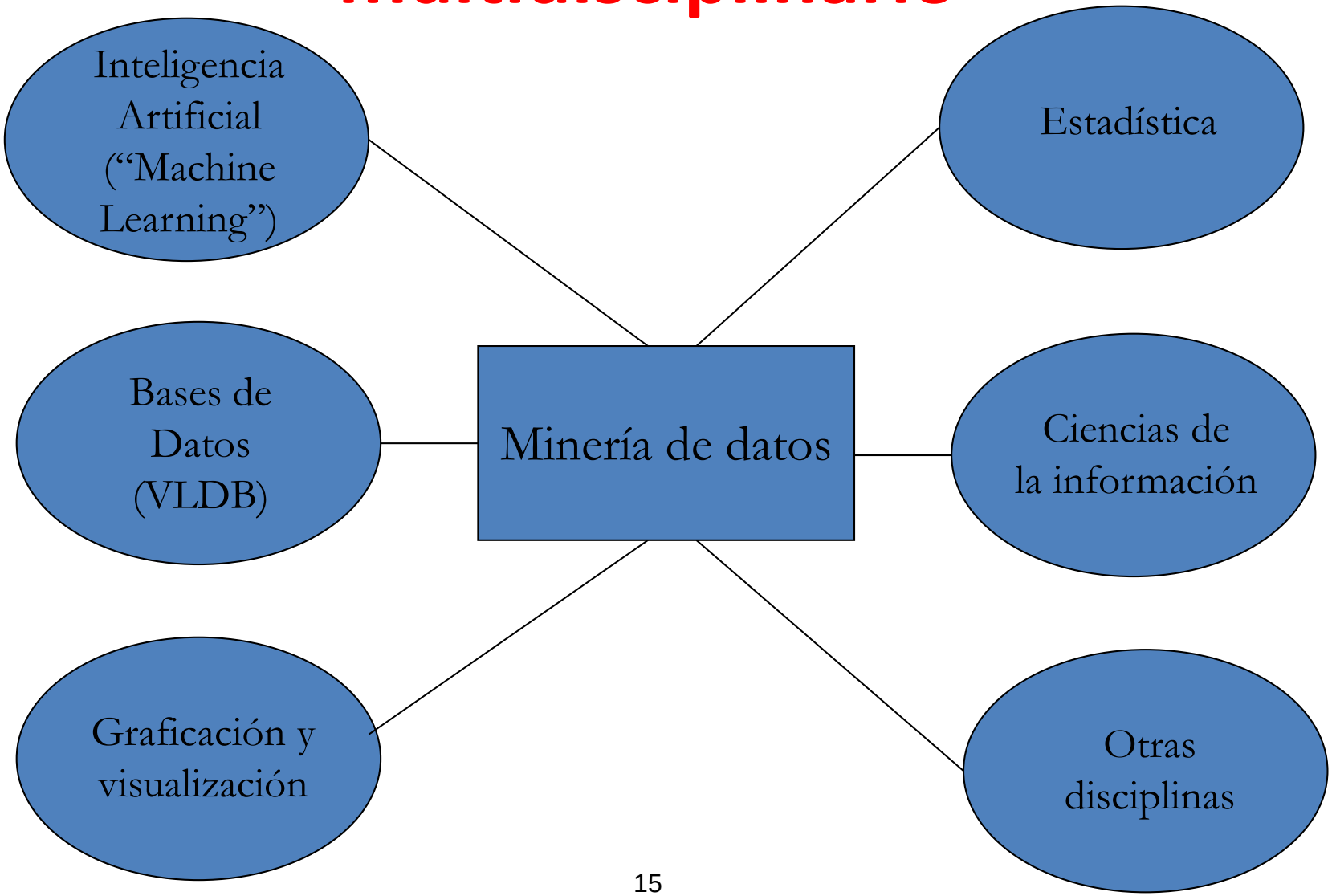
- **Datos:** hechos o medidas que describen características de objetos, eventos o personas, es la materia prima de la que se obtendrá la información.

- **Información:** Datos analizados y presentados en forma adecuada, de interés para un observador en un momento determinado.

- **Conocimiento:** información procesada para emitir juicios que llevan a conclusiones.

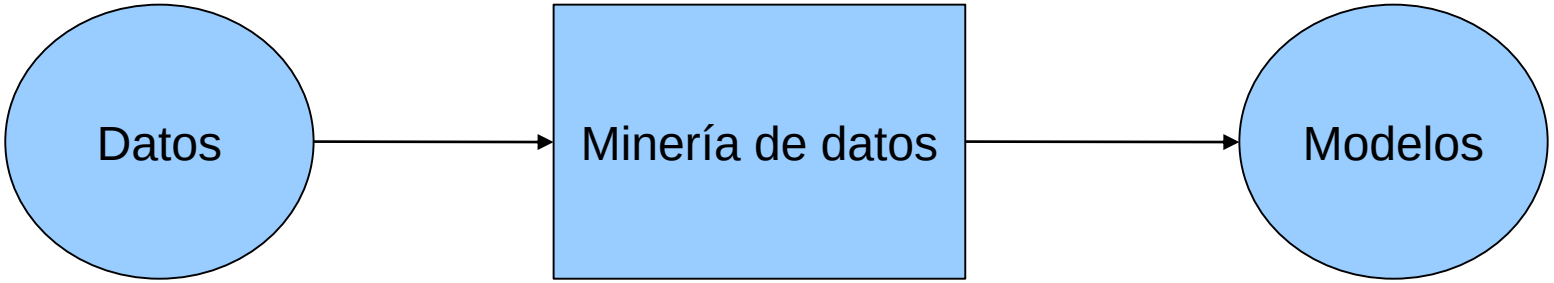
- **Meta Conocimiento:** Reglas que permiten obtener conocimiento.

La minería de datos es un campo multidisciplinario



Aproximación

- Una visión simplificada de la minería de datos



- Los “modelos” son el producto de la minería de datos...
- ...y dan soporte a las estrategias de decisión que se tomen

Minería de datos

- Proceso de utilizar datos “crudos” para inferir importantes relaciones entre ellos
- Colección de técnicas poderosas para analizar grandes volúmenes de datos
- No existe un solo enfoque para minería de datos sino un conjunto de técnicas que se pueden utilizar de manera independiente o en combinación
- Existe una relación con la estadística, aunque frecuentemente se separan las técnicas que no están basadas en métodos estadísticos

Tipos de tareas de la minería de datos

- Pueden clasificarse en las siguientes categorías

  - **Clasificación**
  - **Estimación**
  - **Pronóstico**
  - **Asociación**
  - **Agrupación o segmentación**

Datos y Modelos => Conocimiento

- Los datos se obtienen de:
  - Bases de datos (relacionales, espaciales, temporales, documentales, multimedia, etc)
  - World Wide Web

- **Modelos descriptivos**: identifican patrones que explican o resumen los datos
  - Reglas de asociación: expresan patrones de comportamiento en los datos
  - Clustering: agrupación de casos homogéneos
- **Modelos predictivos**: estiman valores de variables de interés (a predecir) a partir de valores de otras variables (predictoras)
  - Regresión: Variable a predecir continua
  - Clasificación supervisada: Variable a predecir discreta

¿Qué es la Minería de Datos?

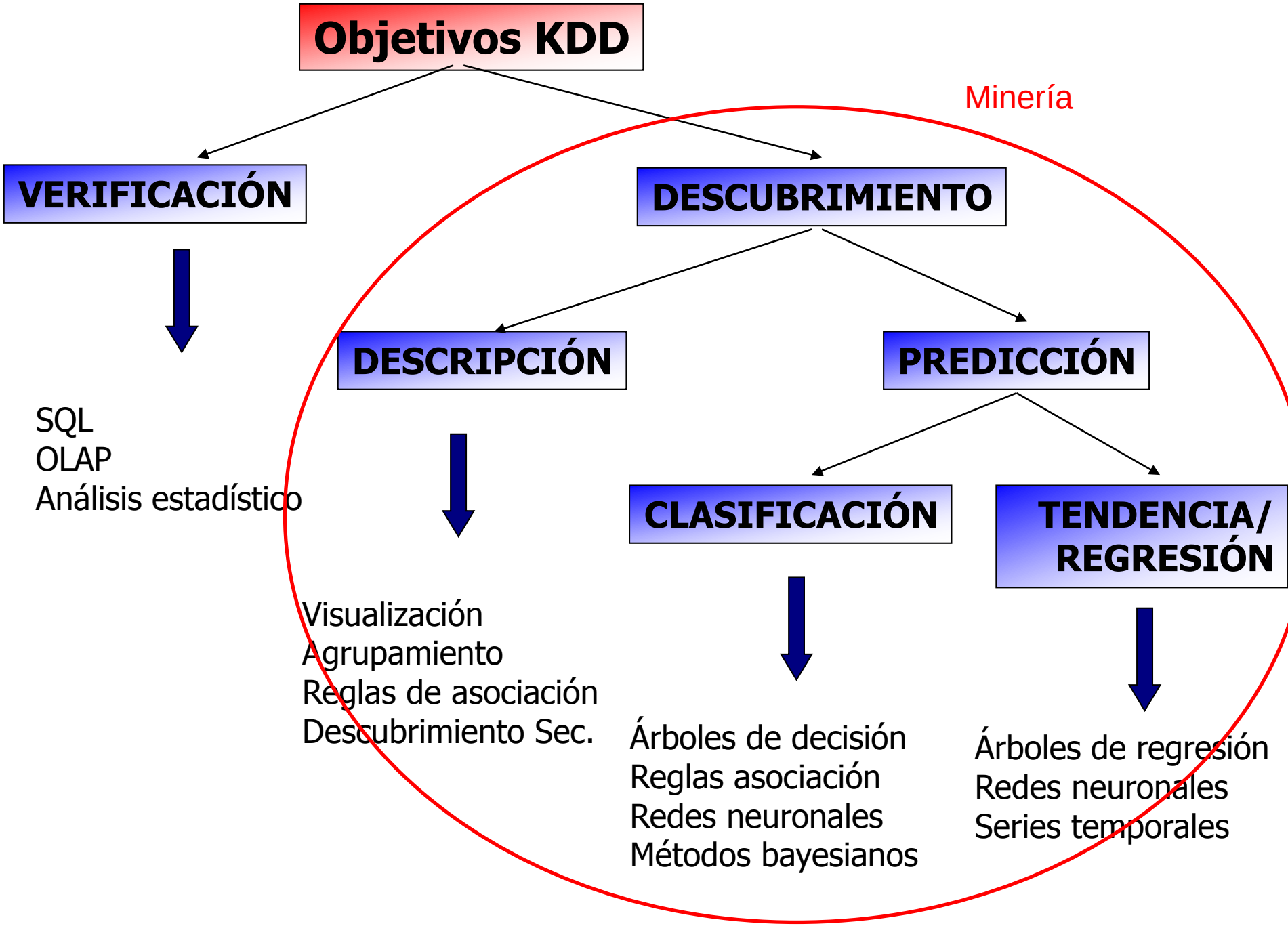
La minería de datos se puede dividir en:

- **Minería de datos predictiva (mdp):** usa primordialmente técnicas estadísticas.

- **Minería de datos para descubrimiento de conocimiento (mddc):** usa principalmente técnicas de inteligencia artificial.

Principales técnicas de minería de datos

- Análisis de canasta de supermercado
- K vecinos más cercanos
- Detección de grupos
- Análisis de encadenamiento
- Árboles de decisión
- Redes neuronales artificiales
- Algoritmos genéticos


Objetivos KDD
Minería
VERIFICACIÓN
DESCUBRIMIENTO
DESCRIPCIÓN
PREDICCIÓN
CLASIFICACIÓN
TENDENCIA/
REGRESIÓN
SQL
OLAP
Análisis estadístico
Visualización
Agrupamiento
Reglas de asociación
Descubrimiento Sec.
Árboles de decisión
Reglas asociación
Redes neuronales
Métodos bayesianos
Árboles de regresión
Redes neuronales
Series temporales

Asociaciones secuenciales

- Encuentra eventos que son inusualmente probables
- Requiere lista de eventos de "entrenamiento", eventos "interesantes" conocidos
- Debe ser robusto frente a eventos adicionales de "ruido“
- usos:
  - Análisis y predicción de fallas
  - Programación dinámica (deformación temporal dinámica)

“Encuentre secuencias comunes de fallas dentro de períodos de 10 minutos “

Alarma 2 en el interruptor C precedido por el error 21 en el interruptor B

Error 17 en cualquier interruptor precedido por Alarma 2 en cualquier interruptor

Time	Switch	Event
21:10	B	Fault 21
21:11	A	Warn 2
21:13	C	Warn 2
21:20	A	Fault 17

Detección de desviación

Encontrar comportamientos anormales

Usos
- Análisis de fallas
- Encontrar valores inesperados, valores atípicos

Usos:
- Análisis de fallas
- Descubrimiento de anomalías para el análisis

Técnicas:
- métodos de agrupamiento / clasificación
- Técnicas estadísticas
- visualización

"Encuentra ocurrencias inusuales en los precios de las acciones de IBM"

Sample date	*Event*	*Occurrences*
58/07/04	Market closed	317 times
59/01/06	2.5% dividend	2 times
59/04/04	50% stock split	7 times
73/10/09	not traded	1 time

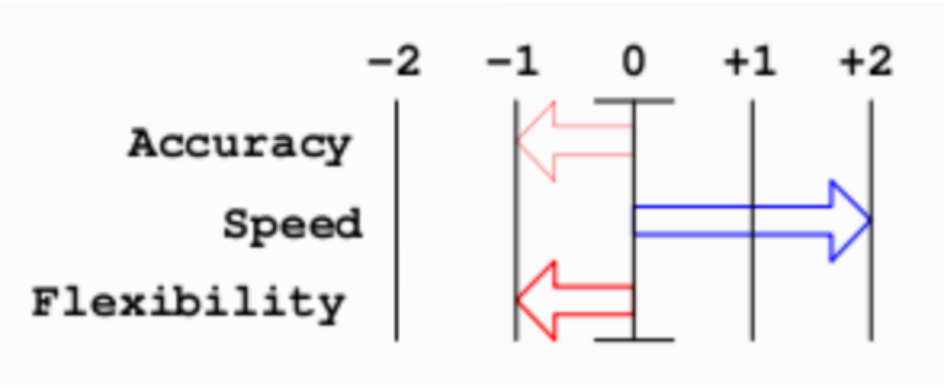
Date	**Close**	**Volume**	**Spread**
58/07/02	369.50	314.08	.022561
58/07/03	369.25	313.87	.022561
58/07/04	Market Closed		
58/07/07	370.00	314.50	.022561

Tipos de validación

- **Validación interna**
  - Es en la que se aprende, clasifica y valida con los datos de un mismo conjunto
- **Validación externa**
  - Se aprende un modelo con un conjunto de datos, y se valida con unos datos que no han sido empleados en el aprendizaje

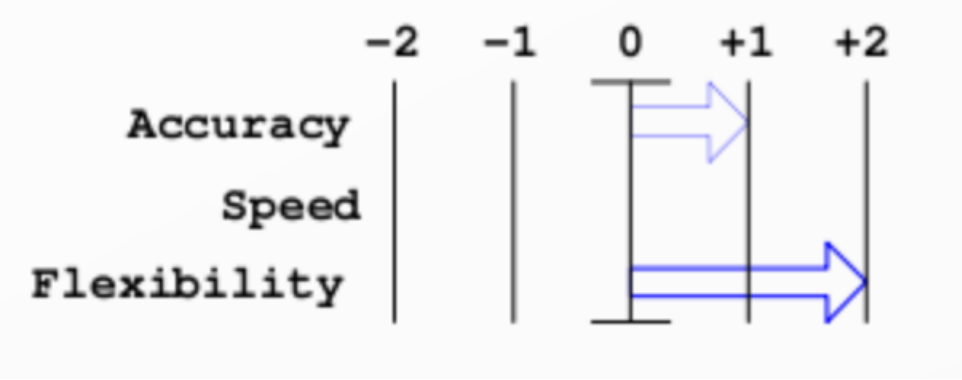
Hold-out

- Separar los datos disponibles en dos subconjuntos de datos: *training set* (para aprender un modelo) y *test set* (el resto de los datos)
- Se calcula la *accuracy* sobre el *test set* para estimar el error del modelo obtenido con el *training set*



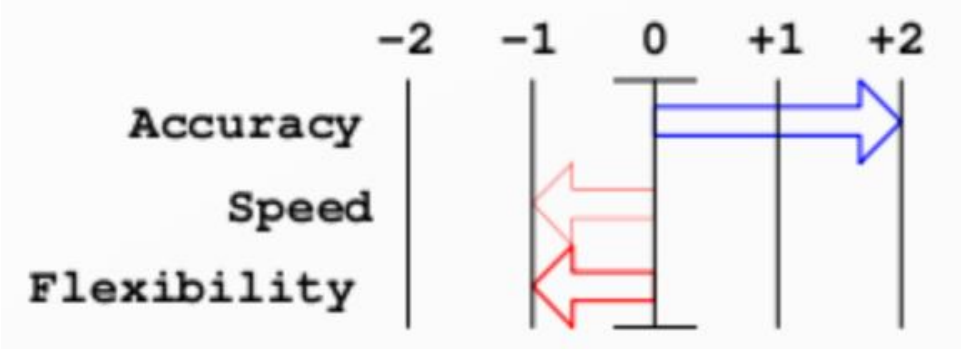
K-fold-Cross-Validation

- Se particiona aleatoriamente en k subconjuntos el conjunto de datos disponible.
- Para cada uno de los subconjuntos obtenidos, se utilizará de *test set* para evaluar el modelo obtenido con el resto de subconjuntos
- Se realiza la media de las evaluaciones realizadas para obtener el resultado final



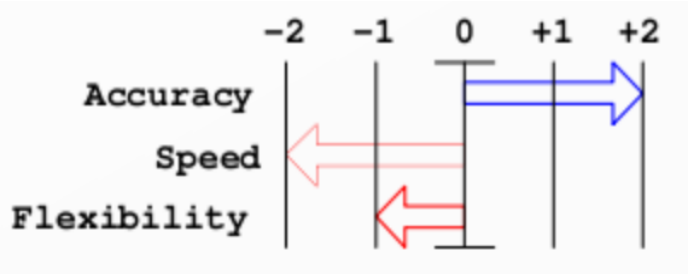
Leave one out

- Se deja una instancia de los datos como *test set* y se aprende con el resto del conjunto
- Este proceso se repite para cada instancia
- Se obtiene el resultado final realizando la media de todas las ejecuciones



0.632 Bootstrap

- Se divide en dos partes
  - Se aprende y se valida con el mismo conjunto de datos
  - N iteraciones de:
    - Se seleccionan con reemplazo el mismo número de instancias que se tengan del conjunto de datos inicial
    - Se utiliza el conjunto de datos creado como *training set* y se evalua con el conjunto formado por las instancias que no han sido seleccionadas en el paso anterior
    - Se obtiene la media de las N iteraciones
- Resultado final: $e = 0.632 \times E_{resubstitution} + 0.368 \times E_{iteraciones}$



Aplicaciones de la Minería de Datos.

Gran cantidad de información (financiera, servicios, empresas, universidades, libros y hobbies), con complejas interrelaciones.
El 99% de la información no le interesa al 99% de la gente.

Ambiente dinámico

En Internet

- ✓***E-bussines.*** ***Perfiles de clientes, publicidad dirigida, fraude.***
- ✓***Buscadores "inteligentes".*** ***Generación de jerarquías, bases de conocimiento web.***
- ✓***Gestión del tráfico de la red.*** ***Control de eficiencia y errores.***

- ***Reglas de asociación:***
  El 60% de las personas que esquían viajan frecuentemente a Europa.
- ***Clasificación:***
  Personas menores de 40 años y salario superior a 2000$ compran on-line frecuentemente.
- ***Clustering:***
  Los usuarios A y B tienen gustos parecidos (acceden URLs similares).
- ***Detección de "outliers"***
  El usuario A navega en Internet más del doble del tiempo promedio.

La publicidad en Internet es uno de los tópicos más actuales de Data Mining.

Los data warehouse de las empresas contienen enormes cantidades de información sobre sus clientes y gestiones.

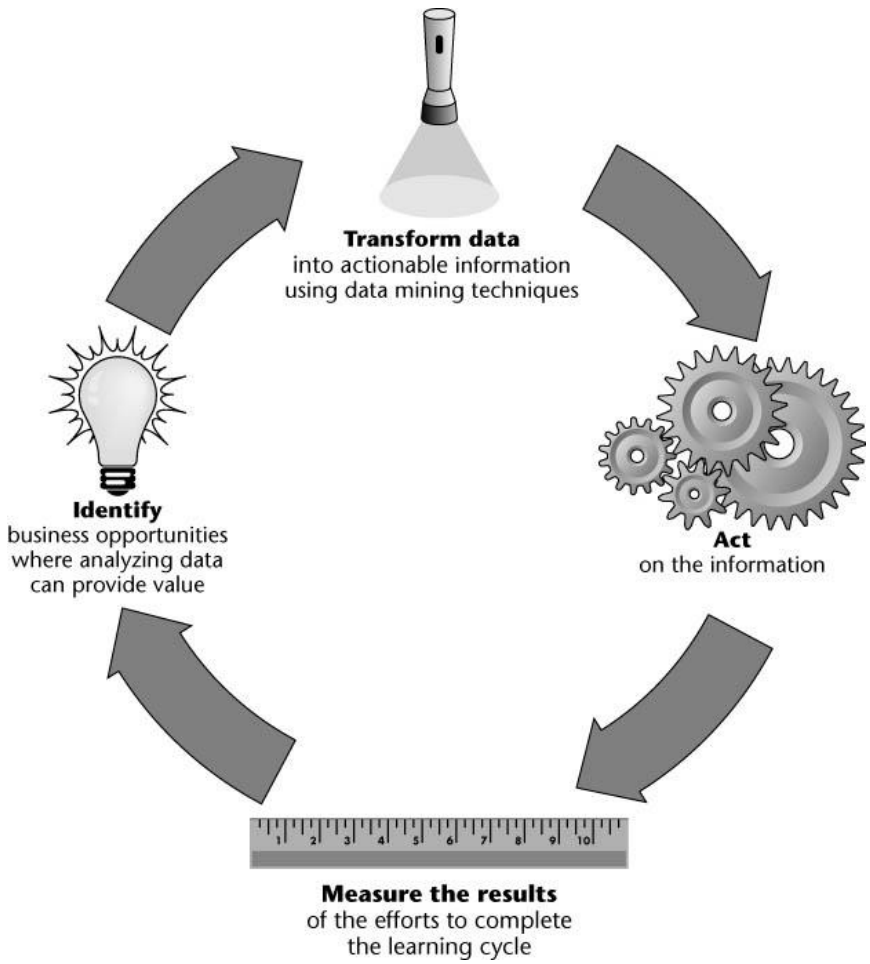
El Mundo de los Negocios

- ✓***Banca.*** ***Grupos de clientes, préstamos, oferta de productos.***
- ✓***Compañías de seguros.*** ***Detección de fraude, administración de recursos.***
- ✓***Marketing.*** ***Publicidad dirigida, estudios de competencia.***

Principales etapas en el proceso de Minería de Datos

- Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.
- El proceso parece secuencial con desarrollo lineal, pero en la práctica, en cualquier etapa se detiene y vuelve atrás.

Ciclo virtuoso de la minería de datos



Filtro de Datos

- El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto".

- Mediante el preprocesado, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, clustering,...).

Selección de variables

- Aún después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad bastante grande de datos.

- La selección de variables se realiza generalmente de una base de datos operacional. Para facilitar el proceso, los datos son copiados en otra base de datos denominada analítica.

- Las principales características de una Base de Datos Analítica, es que contienen gran cantidad de registros (información corporativa), son diseñadas para fines específicos y siempre son de consulta.

- El principal objetivo de la selección de variables es escoger datos que contengan la información o el conocimiento que se desea obtener

Extracción de Conocimiento

- Mediante una técnica de minería de datos (visualización, verificación y descubrimiento), se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.
- También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.
- El problema de la extracción de conocimiento en general se puede reducir a la forma como se manipulan los diferentes tipos de datos.

Interpretación y Evaluación

- Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.

- En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema.

- Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Resumen

- Minería de datos: descubriendo patrones interesantes a partir de grandes cantidades de datos
- Una evolución natural de la tecnología de bases de datos,
- Un proceso KDD incluye limpieza de datos, integración de datos, selección de datos, transformación, extracción de datos, evaluación de patrones y presentación de conocimientos.
- La minería se puede realizar en una variedad de repositorios de información
- Funcionalidades de minería de datos: caracterización, discriminación, asociación, clasificación, agrupamiento, análisis de tendencias y atípicas, etc.