Universidad de Los Andes
Facultad de Ciencias Económicas y Sociales
Escuela de Estadística

Minería de Datos

Prof. Angel A. Zambrano

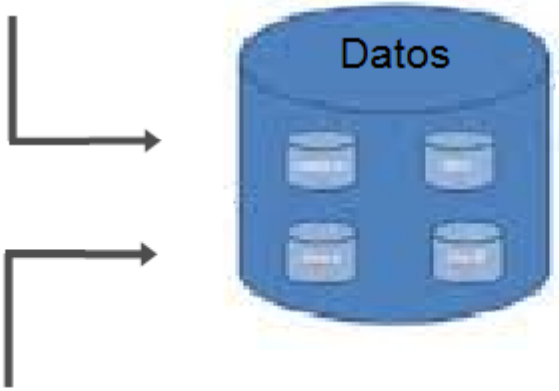
Agenda

- Qué es la Minería de Datos
- Ejemplos
- Herramientas
- Contenido
- Bibliografía
- Evaluación

Qué es la Minería de Datos

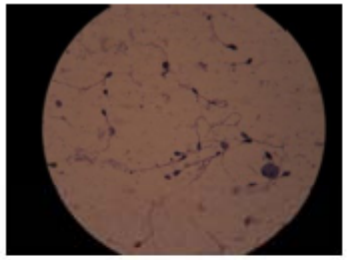
Nuevas Tecnoloías en la Recolección y Almacenamiento de Datos producen grandes conjuntos de Datos

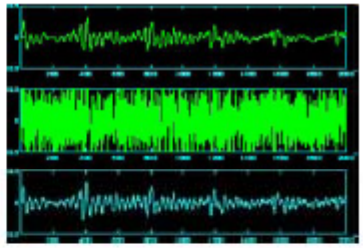
Diversas fuentes de adquisición de datos



Ciencia, Ingeniería, Medicina
Negocios, Bancos, La Web
Redes de telecomunicaciones
Instituciones públicas y privadas
Comunidades virtuales y redes sociales, entre otras

Diferentes tipos de Datos: Numericos, textos, imagénes, videos, sonidos, ondas, espaciales, etc

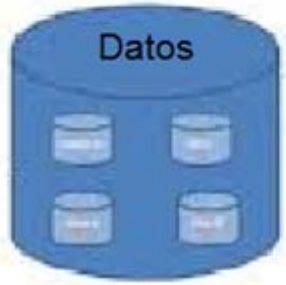




Qué es la Minería de Datos

- Registros médicos
- Registros del uso de tarjetas de crédito, historiales crediticios
- Registros estudiantiles, registros de actividades en sistemas e-learning
- Transacciones en supermercados, tiendas, sitios de comercio electrónico, entre otras
- Accesos a la Web
- Detalles del uso de telefonía móvil
- Imágenes espaciales, médicas, entre otras
- Genoma humano
- Librerías digitales, textos, páginas Web
- Datos climáticos, sísmicos, geográficos
- Interacciones en redes sociales … ¡y muchos más!

Qué es la Minería de Datos



En los datos hay conocimiento oculto

Registros médicos asociados a pacientes	→	Identificar los pacientes con riesgo de sufrir una patología.
Archivos de registro de servidores Web	→	Detectar patrones de comportamiento de los usuarios
Variables financieras de compañías	→	Determinar la capacidad de pagos de crédito
Tweets de los usuarios	→	Determinar el impacto en la población de eventos naturales o epidemias
Valores de métricas que caracterizan un proyecto de software	→	Estimar el costo y tiempo de desarrollo de un proyecto
Transacciones en tiendas de comercio electrónico	→	Detectar patrones de compra de los usuarios

Qué es la Minería de Datos

Cómo extraer el conocimiento oculto en los datos y que sea útil para la toma de desiciones?

KDD (Knowledge Discovery in Databases)

Proceso de Descubrimiento de Conocimiento a partir de Datos

Definición: **según Fayyad *et al*., (1996)**

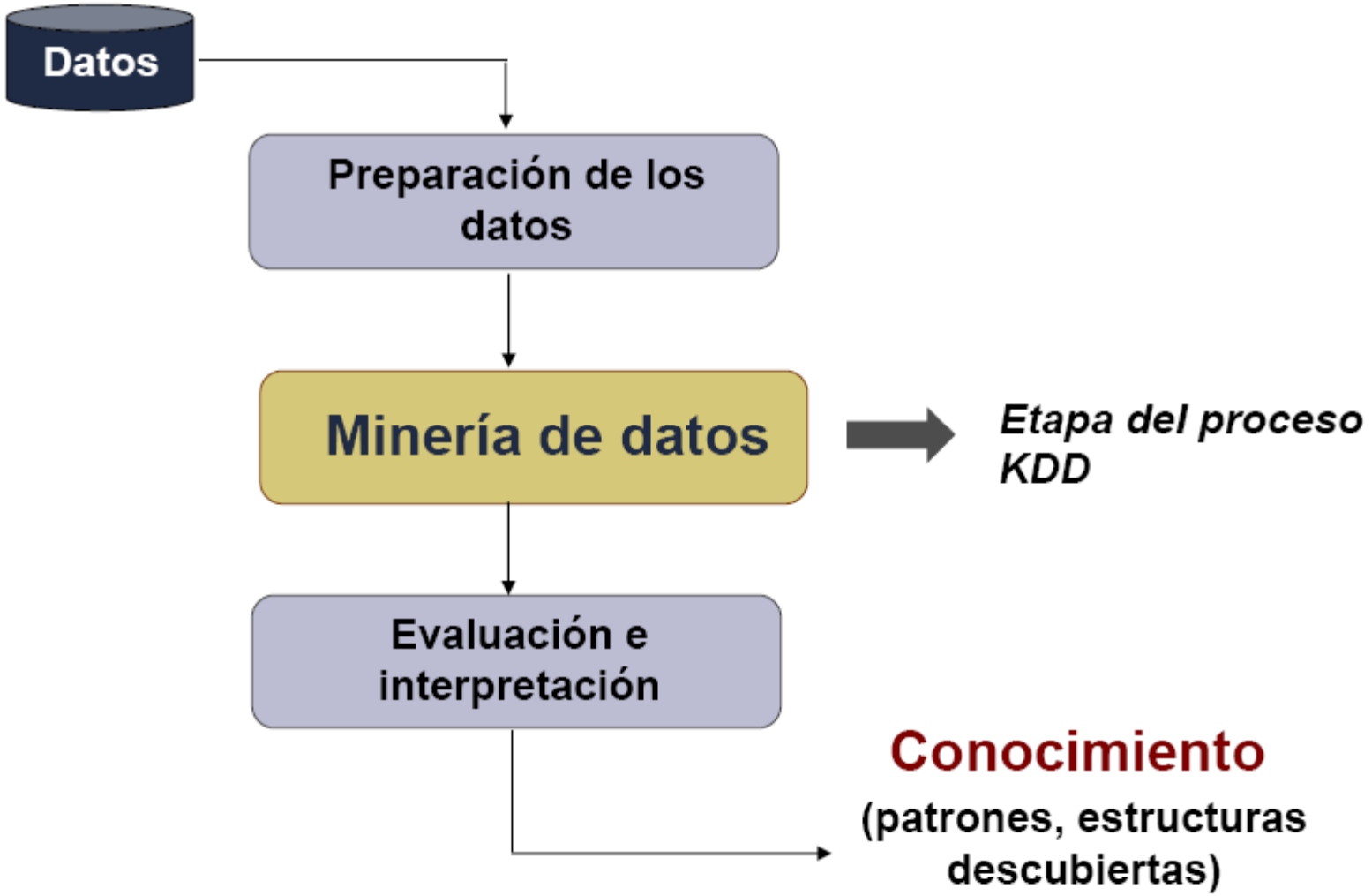
"Proceso NO TRIVIAL de identificar, a partir de datos, patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles"

Entonces, el objetivo del KDD es:

Procesar automáticamente grandes cantidades de datos, identificar los patrones más significativos y relevantes, y presentarlos como conocimiento apropiado para satisfacer las necesidades de los usuarios

Qué es la Minería de Datos

El proceso KDD envuelve:

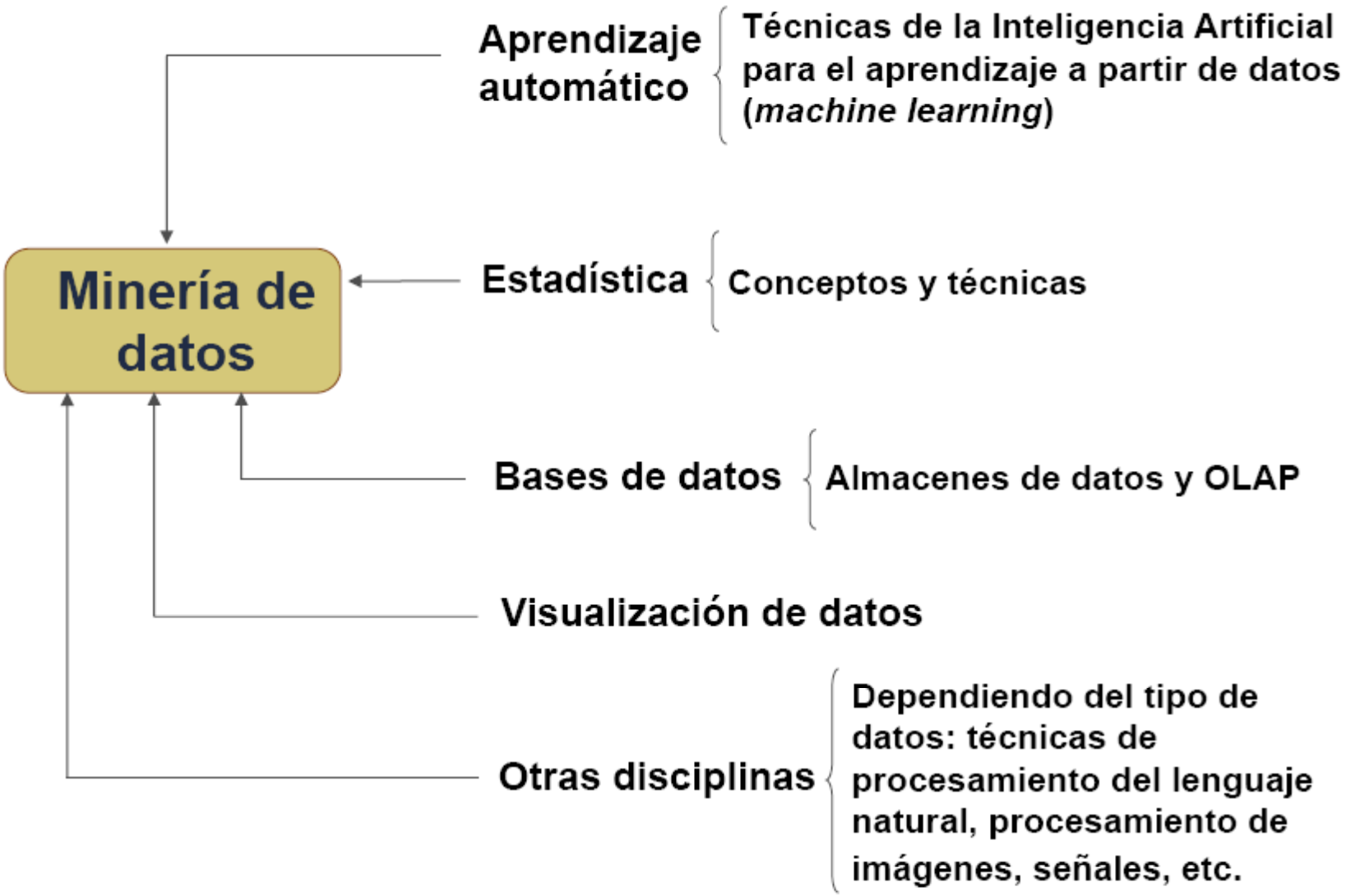


Qué es la Minería de Datos

Definiciones de minería de datos:

- **Aplicación de algoritmos específicos para extraer patrones a partir de datos (*Fayyad et al., 1996*).**
- **Proceso de identificar patrones interesantes a partir de grandes cantidades de datos (*Han et al., 2011*)**
- **Extracción de información implícita, previamente desconocida y potencialmente útil a partir de datos (*Witten y Frank, 2011*)**
- **Conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de conjuntos de datos (*Orallo et al., 2004*).**
- **Proceso de descubrir automáticamente información útil a partir de grandes repositorios de datos (*Tan, P., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining". Pearson*).**

Qué es la Minería de Datos



OLAP =On-Line Analytical Processing - Procesamiento Análitico en Línea

Ejemplos y Aplicaciones

- **Medicina:** diagnóstico de enfermedades, gestión hospitalaria, recomendación de medicinas, …
- **Educación:** Construcción de perfiles de estudiantes, predicción del rendimiento estudiantil, …
- **Industria:** Detección de piezas defectuosas, predicción de fallas, estimación de modelos de calidad, …
- **Mercado:** Segmentación de clientes, evaluación de campañas publicitarias, análisis de mercados, …
- **Inteligencia de negocios**: determinación del comportamiento de clientes, predicción de ventas, …
- **Asistencia personalizada**: buscadores Web adaptados a los usuarios, periódicos electrónicos con noticias personalizadas, …

Ejemplos y Aplicaciones

- **Banca y Finanzas:** Análisis del riesgo de asignación de créditos, detección de fraude, …
- **Biología y Ambiente:** Modelos de calidad del agua, construcción de indicadores ecológicos, clasificación de especies, …
- **Turismo:** Identificación de patrones de reserva, segmentación de clientes, sistemas de recomendación turísticos, …
- **Telecomunicaciones:** Determinación de patrones de llamadas y uso de telefonía móvil, detección de intrusos, …
- **Web:** Análisis del comportamiento de usuarios, clasificación de sitios Web, …
- **Ingeniería de Software**: Predicción y estimación de índices de calidad del software, costo, duración de proyectos, ...
- **Entretenimiento**: juegos adaptados a los usuarios, …

Ejemplos y Aplicaciones

- *Predicción de fallas* (por ejemplo, a partir de los registros de los usuarios de una compañía de telefonía celular, identificar las posibles fallas que se puedan presentar en el servicio. Este conocimiento permitiría mejorar los tiempos de respuesta de los casos hacia los usuarios y mejorar la calidad del servicio).

- *Sistemas de recomendación* (por ejemplo, partir de las búsquedas que realiza un usuario en la Web es posible determinar sus preferencias o perfil, para construir un buscador personalizado que realice recomendaciones de diversos items de acuerdo al perfil detectado)

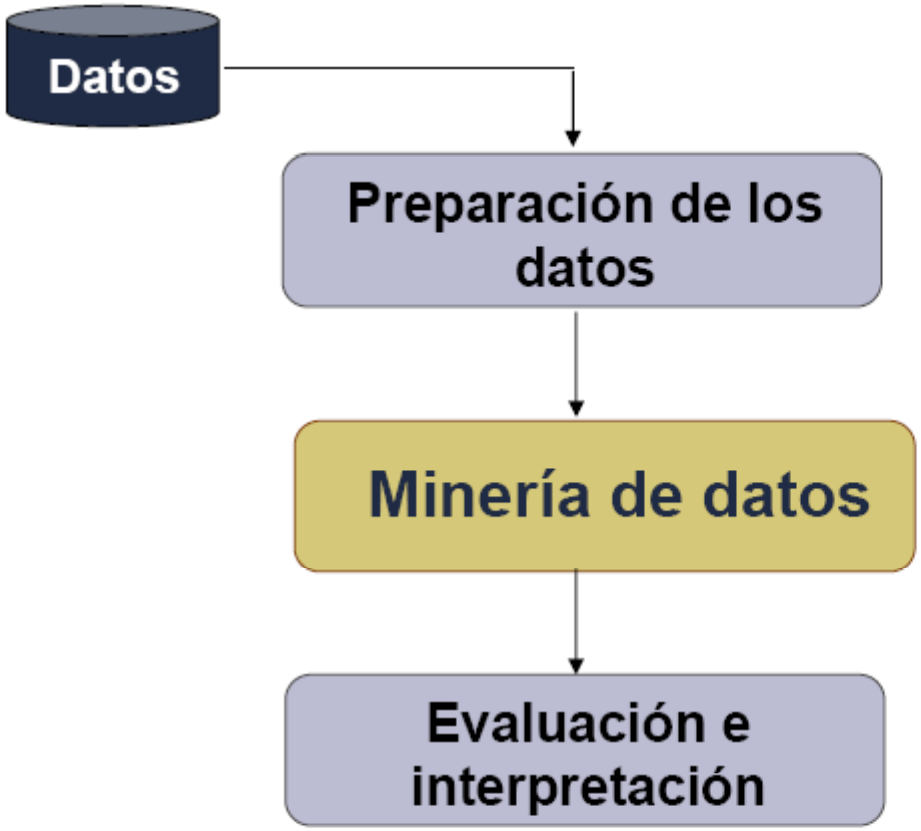
- *Identificación de grupos con características similares* (por ejemplo, a partir de los viajes que han realizado clientes de una agencia de viajes por Internet, identificar grupos de personas con preferencias similares, con el objetivo de sugerir paquetes/ofertas más acertadas)

- *Detección de fraude* (por ejemplo, a partir de los registros de uso de las tarjetas de crédito por parte de los usuarios de un entidad bancaria, detectar eventos anómalos que puedan indicar posibilidad de fraude).

- *Videojuegos adaptativos*: (por ejemplo, a partir de las sesiones de juego de los usuarios se podría detectar su nivel de experticia para realizar una recomendación de niveles de acuerdo al perfil).

Herramientas de la Minería de Datos

Sobre la base del proceso de extracción de conocimiento a partir de datos, en general estas herramientas proporcionan:



Funcionalidades para limpieza y preparación de los datos, así como para la visualización y análisis exploratorio.

Implantación de diversos algoritmos aprendizaje para resolver diferentes tareas de la minería de datos

Funcionalidades para la evaluación de los modelos obtenidos y visualización de resultados.

Herramientas de la Minería de Datos

- **WEKA** (*Universidad de Waikato, Nueva Zelandia*)
- **Suite de Minería de Datos de PENTAHO** (*basado en Weka*)
- **Orange** (*Facultad de informática de la Universidad de Ljubljana, Eslovenia*)
- **RapidMiner** (*Universidad de Dortmund, Alemania, puede utilizar los algoritmos de Weka*)
- **KNIME** (*University of Konstanz, Alemania*)
- **R** (*R Development Core Team*)
- **DBMiner** (*Universidad Simon Fraser, Canadá*)

Otros:

- **SPSS Clementine**
- **ORACLE Data Miner**
- **SAS Enterprise Miner**
- **STATISTICA Data Miner**

Contenido

1. Introducción a la minería de datos.

Concepto de Minería de Datos (MD). Datos y tipos de patrones. Tareas de la MD. Proceso de descubrimiento de conocimiento a partir de datos. Técnicas de MD. Aplicaciones de la minería de datos.

2. Preparación de los datos.

Limpieza y depuración de los datos. Transformaciones. Reducción de la dimensionalidad y selección de variables.

3. Análisis exploratorio de datos.

Conceptos básicos. Resúmenes estadísticos. Técnicas de visualización para análisis de datos.

4. Técnicas de minería de datos para tareas predictivas.

Técnicas para clasificación: Introducción. Aprendizaje de árboles de clasificación. Aprendizaje de reglas de clasificación. Métodos basados en vecindad. Técnicas para regresión: Introducción. Aprendizaje de modelos de regresión.

5. Técnicas de minería de datos para tareas descriptivas.

Técnicas para agrupación: Introducción. Métodos basados en particiones. Métodos jerárquicos. Técnicas para análisis de asociación: Introducción. Soporte y confianza de una regla. Aprendizaje de reglas de asociación.

6. Evaluación de modelos.

Medidas de evaluación. Técnicas de evaluación.

Evaluación

- Examen Teórico.
- Ejercicios en Clases.
- Tareas y Ejercicios para casa.
- Trabajo Final.

Bibliografía

- César Pérez López; Daniel Santin Gónzalez. **Minería de datos. Técnicas y herramientas**. Thomson Editorial Paraninfo. 2007.

- José Hernández Orallo, Ma José Ramírez Quintana, César Ferri Ramírez. **Introducción a la Minería de Datos**. Prentice-Hall. 2004.

- Jiawei Han, Micheline Kamber, Jian Pei. **Data Mining. Concepts and Techniques**. Morgan Kaufmann. Elsevier. Tercera Edición. 2011.