

Data Mining



Introducción: Nociones básicas sobre Data Mining

Necessity is the mother of invention. —Plato

Organización del curso

Grupo docente:

Banchero, Santiago

Fernandez, Juan Manuel

Leonardo, Lucianna

Clases:

Teóricas y Prácticas de laboratorio. (3 hs semanales)

Cronograma y contenidos: (dmuba.github.io)

Evaluación:

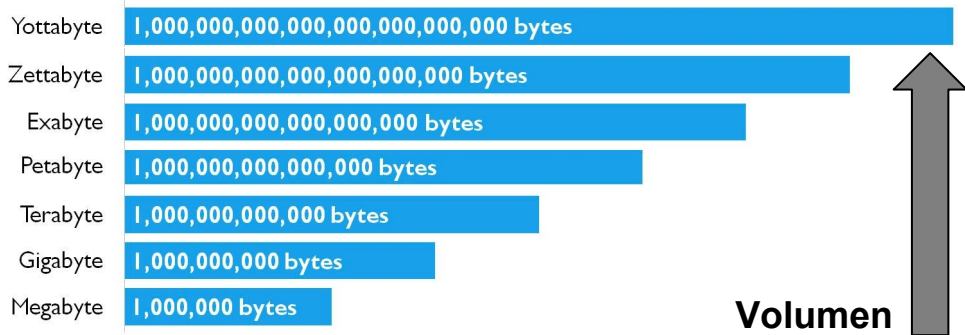
Trabajos prácticos obligatorios (x2). [Grupos de hasta 2]

Exámen escrito: Semana del 16/07

Recuperatorio: Semana del 23/07

¿Qué pasa con los datos?

We are drowning in data, but starving for knowledge!



Alta disponibilidad de datos y de colecciones de datos

Herramientas de recolección de datos automatizadas, sistemas de bases de datos, la Web, la sociedad informatizada.

Multiplicación de fuentes de datos

- ❑ **Negocios:** Web, e-Commerce, transacciones
- ❑ **Ciencia:** Remote sensing, bioinformática, simulaciones de escenarios globales, pronósticos climáticos, etc...
- ❑ **Sociedad:** noticias, dispositivos móviles, redes sociales, etc.

Ej: supermercados (el beep! beep!)

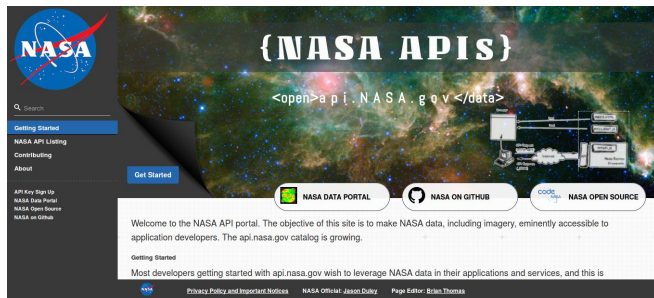
Cualquier empresa de EEUU con más de 1000 empleados tiene almacenado 200 TB de datos en promedio.

¿Cuál fue tu dataset más ?

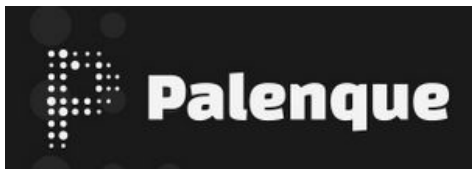


Disponibilidad de accesos

- Gran disponibilidad de accesos a través de Interfaces Programación de Aplicaciones (API).
- Cualquier usuarios puede conectarse a repositorios o streamings de datos.



facebook



Datos abiertos



Ministerio de Modernización
Presidencia de la Nación

Datasets Organizaciones APIs Acerca +

Datos Argentina

Ponemos a tu alcance **datos públicos en formatos abiertos** para que puedas usarlos, modificarlos y compartirlos. **Estos datos son tuyos.** Podés crear visualizaciones, aplicaciones y grandes herramientas con ellos.

625
CONJUNTOS DE DATOS

022
ORGANIZACIONES
CON DATOS

¿Qué datos buscás?

Agropecuaria, pesca y forestación

Asuntos internacionales

Ciencia y tecnología

Economía y finanzas

Educación, cultura y deportes

Energía

Gobierno y sector público

Justicia, seguridad y legales

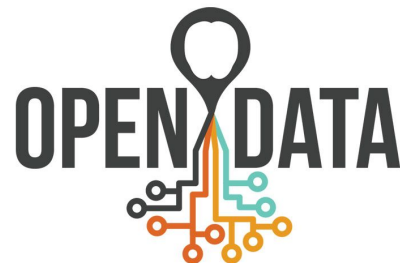
Medio ambiente

Población y sociedad

Regiones y ciudades

Salud

`datos.gob.ar/dataset?groups=agri`



Buenos Aires Data



Iniciativa de Datos Públicos y Transparencia de la Ciudad Autónoma de Buenos Aires.

Buscá entre los 211 datasets



Evolución de las tecnologías de Bases de Datos

Database Management Systems

- ❑ Relational database systems
- ❑ Data modeling: entity-relationship models, etc.
- ❑ Indexing and accessing methods
- ❑ Query processing and optimization
- ❑ Transactions, concurrency control and recovery
- ❑ Online transaction processing (OLTP)

Advanced Database Systems

- ❑ Advanced data models: extended-relational, object relational, etc.
- ❑ Managing complex data: spatial, temporal, multimedia...
- ❑ Data streams
- ❑ Web-based databases (XML, semantic web)
- ❑ Text database systems and integration with Information Retrieval.
- ❑ Cloud computing and parallel data processing
- ❑ Big Data: Extremely large (and complex) data manager.



Advanced Data Analysis

- ❑ Data warehouse and OLAP
- ❑ Data mining and knowledge discovery...
- ❑ Mining complex data: streams, sequence, text, spatial, temporal, Web, networks...
- ❑ Data mining applications: business, retail, banking, social networks, telecommunications, science, recommender systems...

Data mining puede ser visto como una evolución de las tecnologías de información.

¿Qué es Data Mining?

- ❑ Data mining (*knowledge discovery from data*)
- ❑ Es la extracción de **patrones interesantes** o conocimiento a partir de grandes cantidades de datos.
- ❑ Patrones interesantes:
 - ❑ No triviales
 - ❑ Implícitos
 - ❑ Previamente desconocidos
 - ❑ Potencialmente útiles

¿Qué es un patrón?

En Data Mining buscamos patrones.

- ❑ ¿Qué caracteriza a un pez del Río Paraná?
- ❑ ¿Cómo los distinguimos?



Características de un pez

Tiene bigote

Genero

Longitud (mts.)



SI

Pseudoplatystoma

1.8



NO

Salminus

1.3



SI

Patucu

1



Cada pez se convertirá en un vector de características

0.4

Data Mining y sus nombres alternativos

- ❑ Knowledge discovery (mining) in databases (KDD),
- ❑ knowledge extraction,
- ❑ data/pattern analysis,
- ❑ data archeology,
- ❑ data dredging,
- ❑ information harvesting,
- ❑ business intelligence

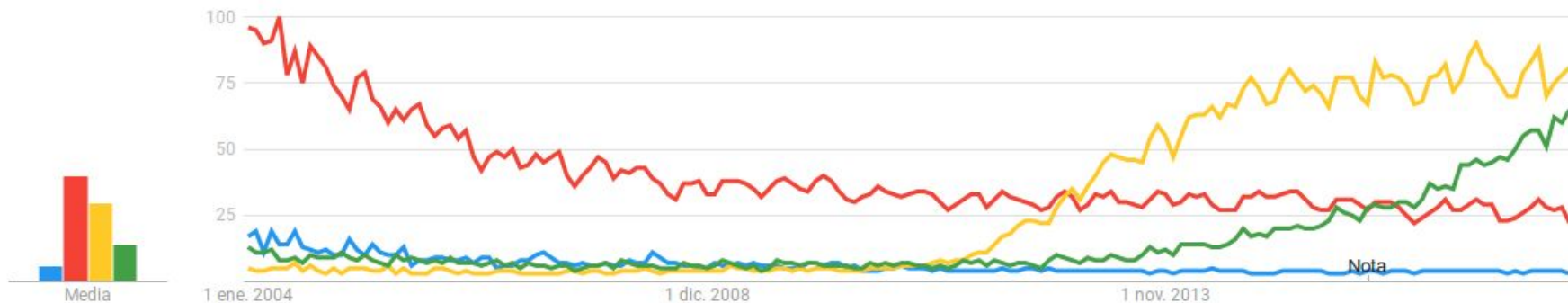


En la actualidad está representada por ***Data Science***

Data Mining: Evolución

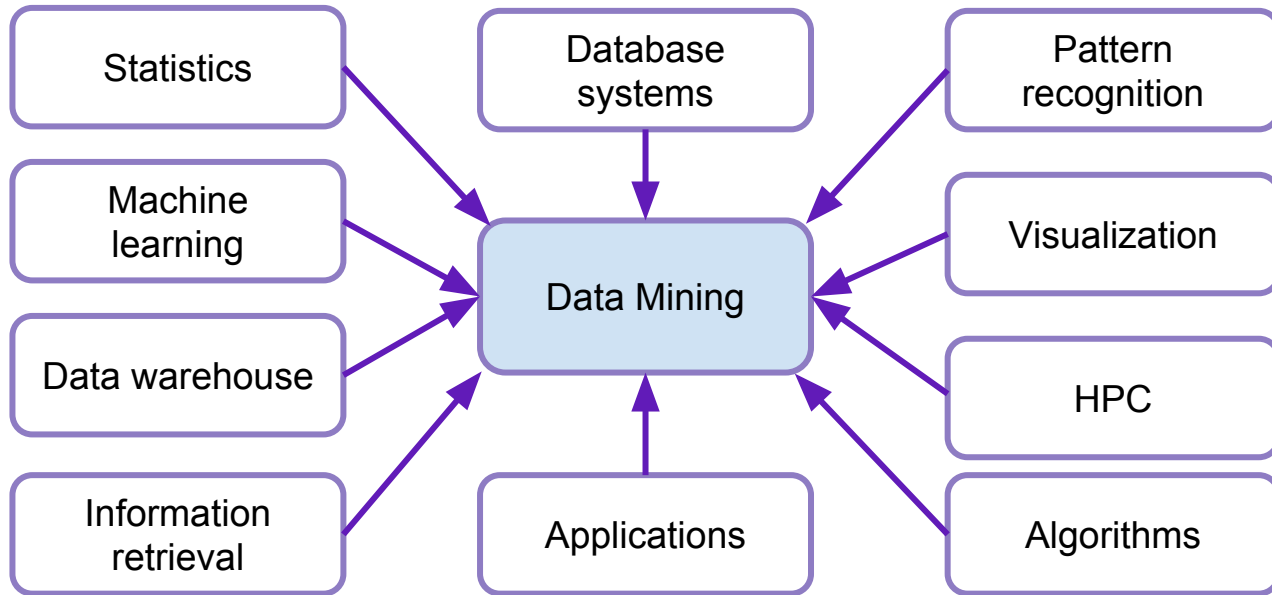
● KDD ● Data Mining ● Big Data ● Data Science

Todo el mundo, 2004 - hoy



¿Qué tecnologías se utilizan?

- ❑ La minería de datos ha incorporado muchas técnicas de otros dominios.
- ❑ Naturaleza interdisciplinaria.



¿Qué datos podemos minar?



Datos
estructurados

- ❑ Conjuntos de datos basados en aplicaciones de bases de datos.
 - ❑ Bases de datos relacionales, data warehouse, transactional database
- ❑ Conjuntos de datos avanzados y aplicaciones avanzadas
 - ❑ Streamings de datos y provenientes de sensores.
 - ❑ Time-series data, temporal data, sequence data.
 - ❑ Structure data, graphs, social networks and multi-linked data
 - ❑ Object-relational databases
 - ❑ Heterogeneous databases and legacy databases
 - ❑ Spatial data and spatiotemporal data
 - ❑ Multimedia database
 - ❑ Text databases
 - ❑ The World-Wide Web

Knowledge Discovery in Databases (KDD)

- ❑ Muchos tratan al Data Mining como un **sinónimo** de Knowledge Discovery in Databases (KDD).
- ❑ KDD fue acuñado en 1989 para enfatizar que el **conocimiento puede ser obtenido de un enfoque *data-driven*.**
- ❑ Otros ven al Data Mining simplemente como **un paso esencial** en el proceso de descubrimiento de conocimiento.

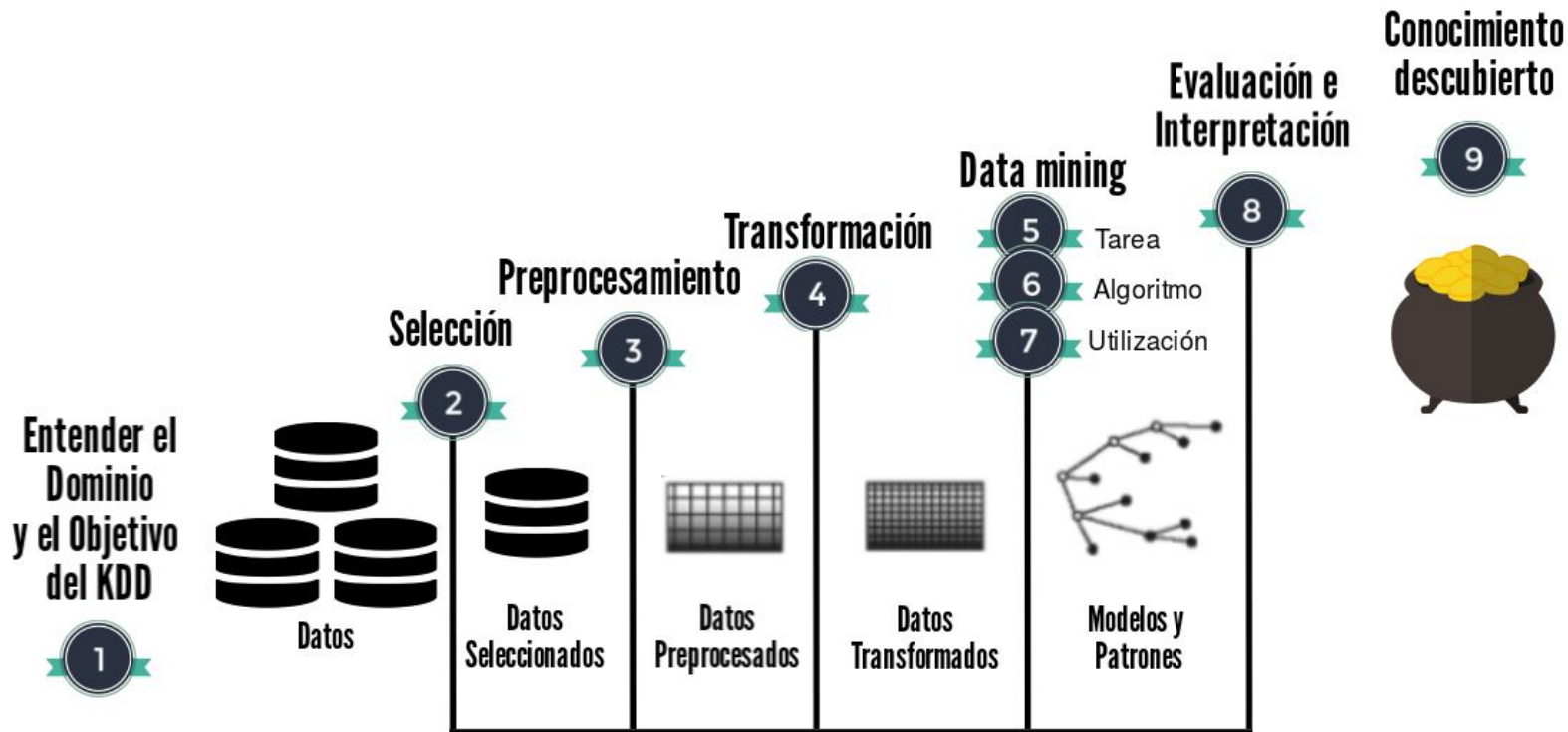
Proceso de KDD

El proceso KDD es **interactivo** e **iterativo**, con la participación de numerosos pasos con muchas decisiones tomadas por el usuario.

- ❑ Si el resultado final o intermedio no pasa los procesos de evaluación se puede repetir desde el principio o a partir de cualquiera de los pasos anteriores.
- ❑ Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.



Proceso de KDD



Entender el dominio

Entender el dominio de aplicación

- ❑ ¿En donde quiero aplicar el proceso de KDD?
- ❑ ¿Cuál es el problema a resolver?
- ❑ ¿Cuáles son los objetivos?



Seleccionar el conjunto de datos

¿Cuáles son los datos que voy a utilizar?

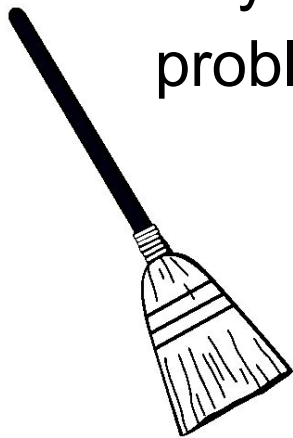
Ahora hay que crear el conjunto de datos (**dataset**) donde realizaremos el proceso de descubrimiento de conocimiento.

Esto incluye:

- ❑ Averiguar con qué datos disponemos.
- ❑ Obtener datos adicionales en el caso que sea necesario.
- ❑ Integrar todos los datos conseguidos de diferentes fuentes en un único dataset.

Preprocesamiento y limpieza de datos

- ❑ Se trata de mejorar la fiabilidad de los datos.
- ❑ Aquí se incluye limpieza de datos, tales como el manejo de los **datos faltantes** y la **eliminación de ruido o valores atípicos**.
- ❑ Hay que ser cauto y entender sobre el dominio del problema (no todo es anomalía, no todo es ruido)



Transformación

En esta etapa se trabaja en mejorar los datos para la etapa de Minería de Datos.

Reducción de dimensionalidad: PCA, Correlación, χ^2 , etc.

Suavizados: Discretización, Binning Methods, medias móviles, etc.

Agregación: *Group By*. Ejemplo ventas diarias agrupadas en ventas mensuales o anuales. (Granularidad)

Generalización: Datos de Bajo Nivel (raw data) son reemplazados por conceptos de Alto Nivel a través del concepto de jerarquía.

Normalización: Escalado de los atributos para unificar dominios. Puede ser llevar a un rango: -1 a 1 ó 0 a 1, también restar la media y dividir por el desvío estándar (z-score).

Construcción de Atributos: Se construyen nuevas variables que aportan mayor variabilidad favoreciendo al proceso de mining.

Selección del Método (o Tarea) de Mining

Es un proceso esencial donde se aplican métodos **“inteligentes”** para extracción de patrones de los datos.

¿Cuáles son las tareas que puede lograr la minería de los datos?

Descripción: Es la explicación de los patrones y tendencias existentes en los datos. Se aborda a través del Análisis exploratorio de datos y técnicas de visualización.

Selección del Método (o Tarea) de Mining

Estimación: Es la aproximación de una variable numérica (target) a partir de predictores numéricos o categóricos. Estimación puntual, regresiones, intervalos de confianza, etc.

Predicción: Permite pronosticar valores futuros de una variable, donde el target es numérico.

Clasificación: Utiliza una variable categórica como target.

Clustering: Es el agrupamiento de registros en grupos con similares características. Similares dentro y diferentes entre grupos.

Asociación: Es la acción de buscar qué atributos "van juntos". El análisis del changuito del super.

Selección del Método (o Tarea) de Mining

Estimación: Es la aproximación de una variable numérica (target) a partir de predictores numéricos o categóricos. Estimación puntual, regresiones, intervalos de confianza, etc.

Predicción: Permite pronosticar valores futuros de una variable, donde el target es numérico.

Clasificación: Utiliza una variable categórica como target.

Clustering: Es el agrupamiento de registros en grupos con similares características. Similares dentro y diferentes entre grupos.

Asociación: Es la acción de buscar qué atributos "van juntos". El análisis del changuito del super.

Selección del Algoritmo de Mining a utilizar

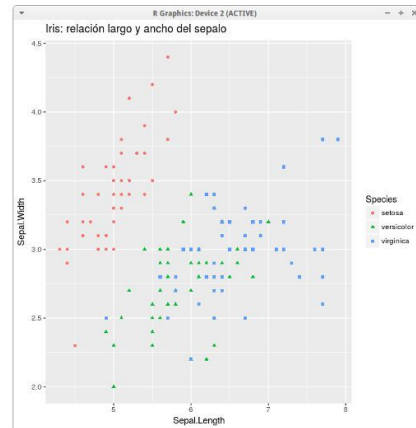
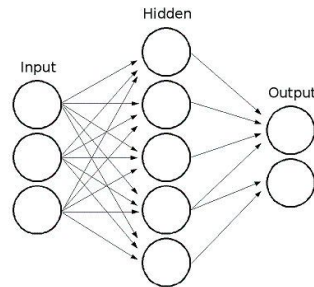
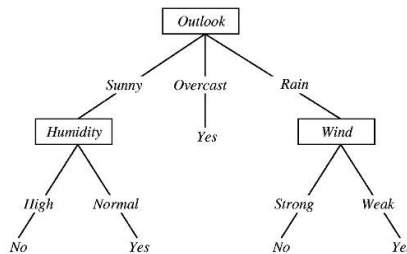
¿Qué Algoritmo de mining vamos a utilizar?

- ❑ Meta-Learning. ¿Cómo trabaja un algoritmo de aprendizaje? ¿Con qué datos estoy trabajando?

Cada algoritmo tiene:

- ❑ Un conjunto de parámetros
- ❑ sus táctica de aprendizaje.
 - ❑ Ejemplo: Clasificación y Predicción

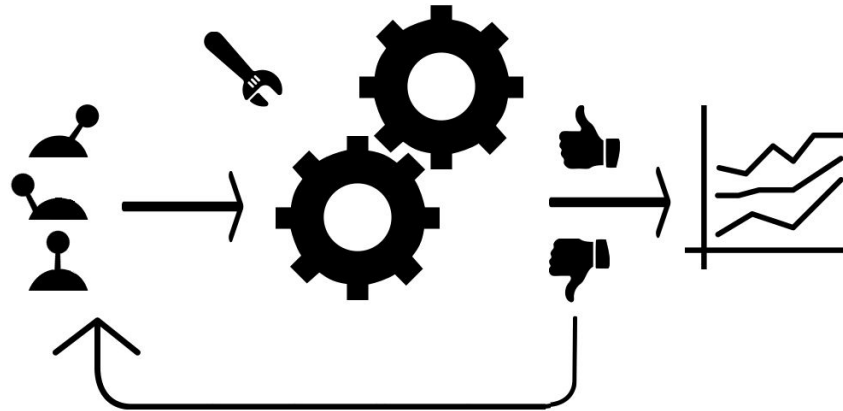
Hay que decidir entre **Precisión** vs **Capacidad de explicar**.



Utilización del algoritmo de data mining

En este paso será necesario emplear el algoritmo varias veces hasta obtener resultados satisfactorios.

Aquí se van a ajustar los parámetros de los algoritmos.



Evaluación

¿Qué tan buenos son los resultados conseguidos?

- ❑ Se evalúan e interpretan los patrones para determinar si se llegó a nuevos conocimientos.
- ❑ Este paso se centra en la comprensibilidad y la utilidad del modelo inducido.



Evaluación

- ❑ Los patrones descubiertos tienen que poder ser **validados** con nuevos datos con algún grado de certidumbre.
- ❑ Se necesitan **medidas cuantitativas** para evaluar los patrones.
- ❑ Además de precisión es posible evaluar con alguna **función de utilidad**. Ej: Maximizando/Minimizando una función de Ganancia.

$$\text{❑ } G = 3000 * P - 60 * N$$

Utilizar el conocimiento

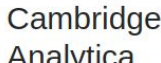

- ❑ El desafío aquí es superar las **condiciones de laboratorio**. Pasar de los datos seleccionados para ajustar el modelo al mundo real.
- ❑ Se deben verificar potenciales conflictos con conocimiento previos.
- ❑ El conocimiento puede ser incorporado en un sistema para dar soporte al proceso de toma de decisiones.

Ejemplo: Minería de la Web

Hacer DM a partir de datos de la Web involucra varias etapas:

- ☐ Data cleaning
- ☐ Data integration from multiple sources
- ☐ Warehousing the data
- ☐ Data cube construction
- ☐ Data selection for data mining
- ☐ Data mining
- ☐ Presentation of the mining results
- ☐ Patterns and knowledge to be used or stored into knowledge-base

Principales desafíos del Data Mining II

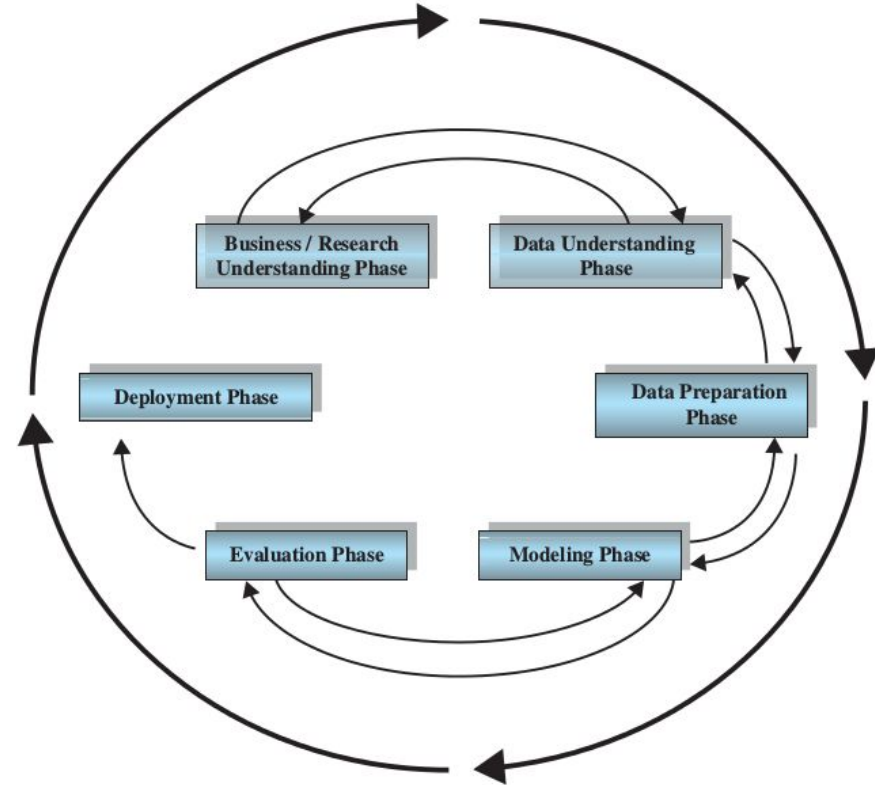
- ❑ Efficiency and Scalability
 - ❑ Efficiency and scalability of data mining algorithms
 - ❑ Parallel, distributed, stream, and incremental mining methods
 - ❑ Diversity of data types
 - ❑ Handling complex types of data
 - ❑ Mining dynamic, networked, and global data repositories
 - ❑ Data mining and society
 - ❑ Social impacts of data mining
 - ❑ Privacy-preserving data mining
 - ❑ Invisible data mining
- 



CRISP for DM

Cross-Industry Standard Process for Data Mining.

CRISP proporciona un proceso estándar no patentado y de libre acceso para adaptar la minería de datos a la estrategia general de resolución de problemas de una unidad comercial o de investigación.



Conferences and Journals on Data Mining

❑ KDD Conferences

- ❑ ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- ❑ SIAM Data Mining Conf. (SDM)
- ❑ (IEEE) Int. Conf. on Data Mining (ICDM)
- ❑ European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
- ❑ Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- ❑ Int. Conf. on Web Search and Data Mining (WSDM)

❑ Other related conferences

- ❑ DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
- ❑ Web and IR conferences: WWW, SIGIR, WSDM
- ❑ ML conferences: ICML, NIPS
- ❑ PR conferences: CVPR,

❑ Journals

- ❑ Data Mining and Knowledge Discovery (DAMI or DMKD)
- ❑ IEEE Trans. On Knowledge and Data Eng. (TKDE)
- ❑ KDD Explorations
- ❑ ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- ❑ Data mining and KDD (SIGKDD: CDROM)
 - ❑ Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - ❑ Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- ❑ Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
 - ❑ Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - ❑ Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- ❑ AI & Machine Learning
 - ❑ Conferences: Machine learning (ML), AAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
 - ❑ Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- ❑ Web and IR
 - ❑ Conferences: SIGIR, WWW, CIKM, etc.
 - ❑ Journals: WWW: Internet and Web Information Systems,
- ❑ Statistics
 - ❑ Conferences: Joint Stat. Meeting, etc.
 - ❑ Journals: Annals of statistics, etc.
- ❑ Visualization
 - ❑ Conference proceedings: CHI, ACM-SIGGraph, etc.
 - ❑ Journals: IEEE Trans. visualization and computer graphics, etc.



Referencias

- ❑ Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37. [[pdf](#)]
- ❑ Jiawei Han, Micheline Kamber, Jian Pei. 2011. Tercera edición. Data Mining: Concepts and Techniques.
- ❑ Daniel T. Larose. 2014. Segunda edición. Discovering Knowledge in Data: An Introduction to Data Mining.

Presentación inspirada en las clases de ([Han, 2011](#))