DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Tratamiento Inteligente de Datos

© Fernando Berzal, berzal@acm.org

IDA, KDD & Data Mining

- Datos en la empresa
- ¿Qué es la minería de datos?
- Aplicaciones
- KDD (Knowledge Discovery in Databases)
  - El proceso de extracción de conocimiento
  - Carácter multidisciplinar
- Técnicas de minería de datos
  - Modelos descriptivos y modelos predictivos
  - Clasificación de las técnicas de minería de datos
- Fuentes de datos
- Evaluación de resultados
- Ciencia de datos & Big Data
- Sistemas de minería de datos

Datos en la empresa

Datos

Representación formal de hechos, conceptos o instrucciones adecuada para su comunicación, interpretación y procesamiento por seres humanos o medios automáticos.

Información

El significado que un ser humano le asigna a los datos.

Datos en la empresa

Sistema de información

- Sistema, automatizado o manual, que engloba a personas, máquinas y/o métodos organizados para recopilar, procesar, transmitir datos que representan información.
- Infraestructura, organización, personal y componentes para la recopilación, procesamiento, almacenamiento, transmisión, visualización, diseminación y organización de información.

Datos en la empresa

Utilidad de los sistemas de información:

Gestión de los recursos de una empresa

Los sistemas de información sirven de apoyo en la realización de las actividades propias de una empresa:

- Comunicación:
  intranets/extranets, VANs [value-added networks], teletrabajo…
- Resolución de problemas:
  - MIS [Management Information Systems]
  - OLAP [Online Analytical Processing]
  - DSSs [Decision Support Systems]
  - KBSs [Knowledge-Based Systems]

Datos en la empresa

Sistemas de información gerencial

[MIS: Management Information Systems]

OBJETIVO

Proporcionar la información adecuada
a la persona adecuada
donde y cuando la necesite.

Datos en la empresa

La información como recurso empresarial

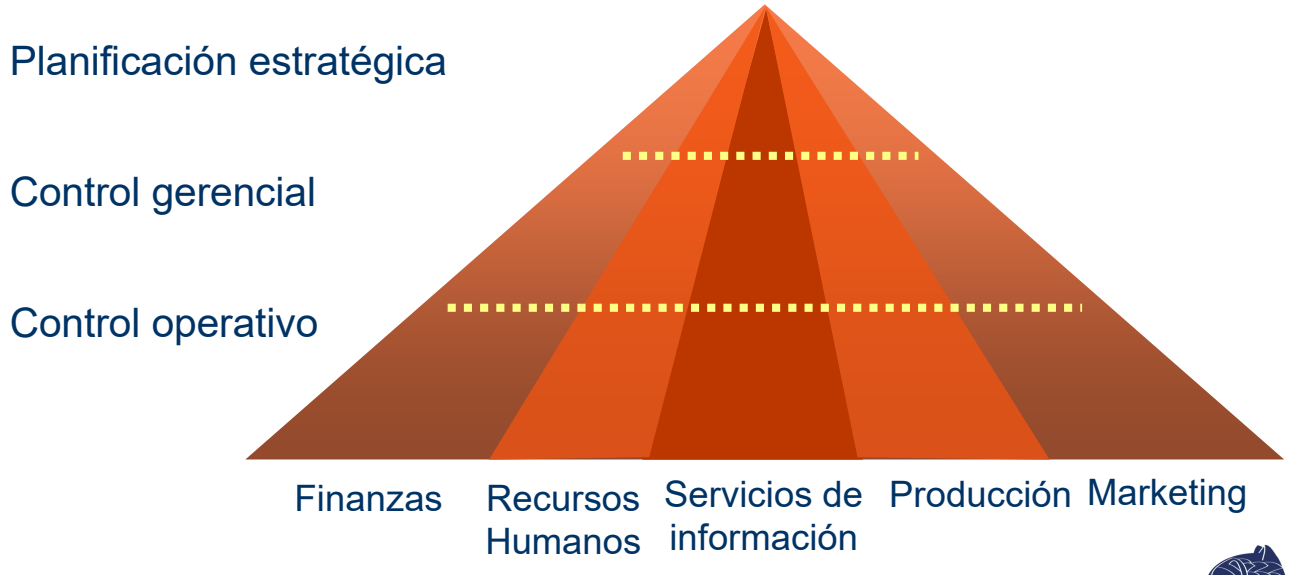
La información…

- … se obtiene a partir de datos (GIGO).
- … su producción es costosa.
- … no se consume ni agota al compartirla.
- … puede usarse simultáneamente.
- … es sinergética (cuanto más se usa, más se enriquece).
- … a corto plazo, no es posible medir el valor de su uso.

Una empresa no puede vivir sin usarla a todos sus niveles.

Datos en la empresa

Áreas funcionales de una empresa

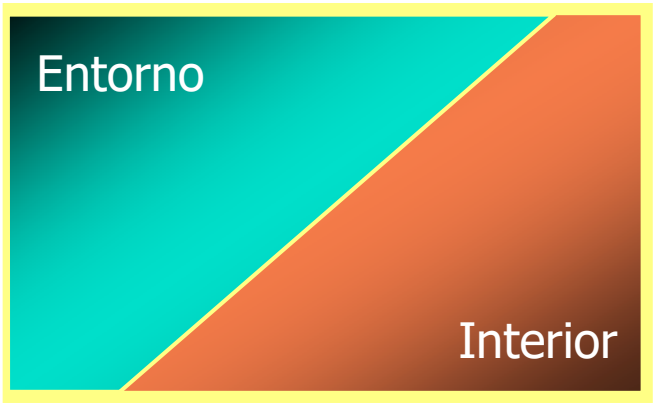


Datos en la empresa

Gestión de los recursos de una empresa:
Niveles gerenciales

Origen de la información

- Planificación estratégica
- Control gerencial
- Control operativo



Datos en la empresa

Gestión de los recursos de una empresa:
Niveles gerenciales

Presentación de la información

- Planificación estratégica
- Control gerencial
- Control operativo

¿Qué es la minería de datos?

Extracción de patrones ("conocimiento")
en **grandes** bases de datos.

¿Qué es la minería de datos?

Extracción de **conocimiento**
en grandes bases de datos.

Requisitos

- No trivial
- Implícito
- Previamente desconocido
- Potencialmente útil

¿Qué es la minería de datos?

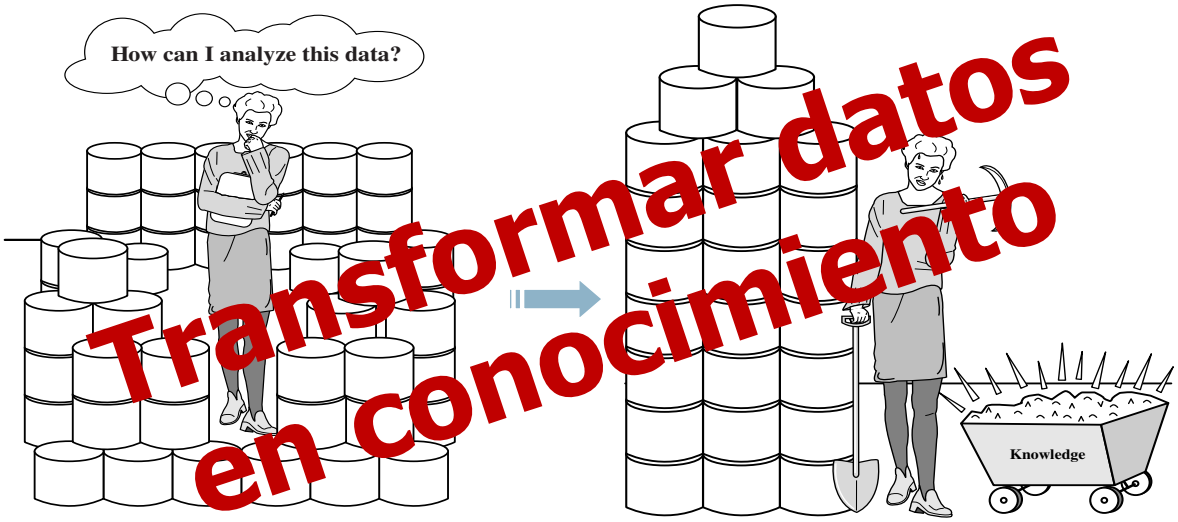
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

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

- Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

  Berry & Linoff:
  Data Mining Techniques.
  Wiley, 1997

¿Qué es la minería de datos?

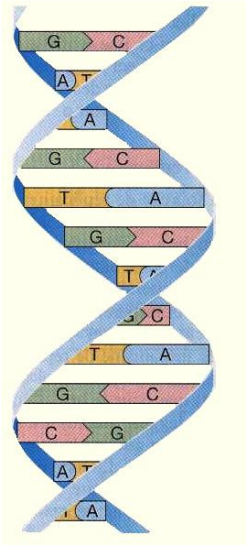


Transformar datos en conocimiento

"Data rich, Information poor" → Conocimiento (patrones interesantes)

Aplicaciones

- Market basket analysis (compras)
- Perfiles de usuario en la Web
- Segmentación de clientes
- Detección de fraudes / intrusos
- …



Google

amazon.com

KDD (Knowledge Discovery in Databases)

IDA [Intelligent Data Analysis]

Tratamiento inteligente de datos

- El uso de ordenadores permitió el desarrollo de nuevas técnicas de análisis de datos más allá de las técnicas estadísticas tradicionales.

- Como estas técnicas se desarrollaron como métodos de aprendizaje automático dentro de la I.A., recibieron el nombre de "análisis inteligente de datos".

KDD (Knowledge Discovery in Databases)

KDD [Knowledge Discovery in Databases]

Extracción de conocimiento en bases de datos

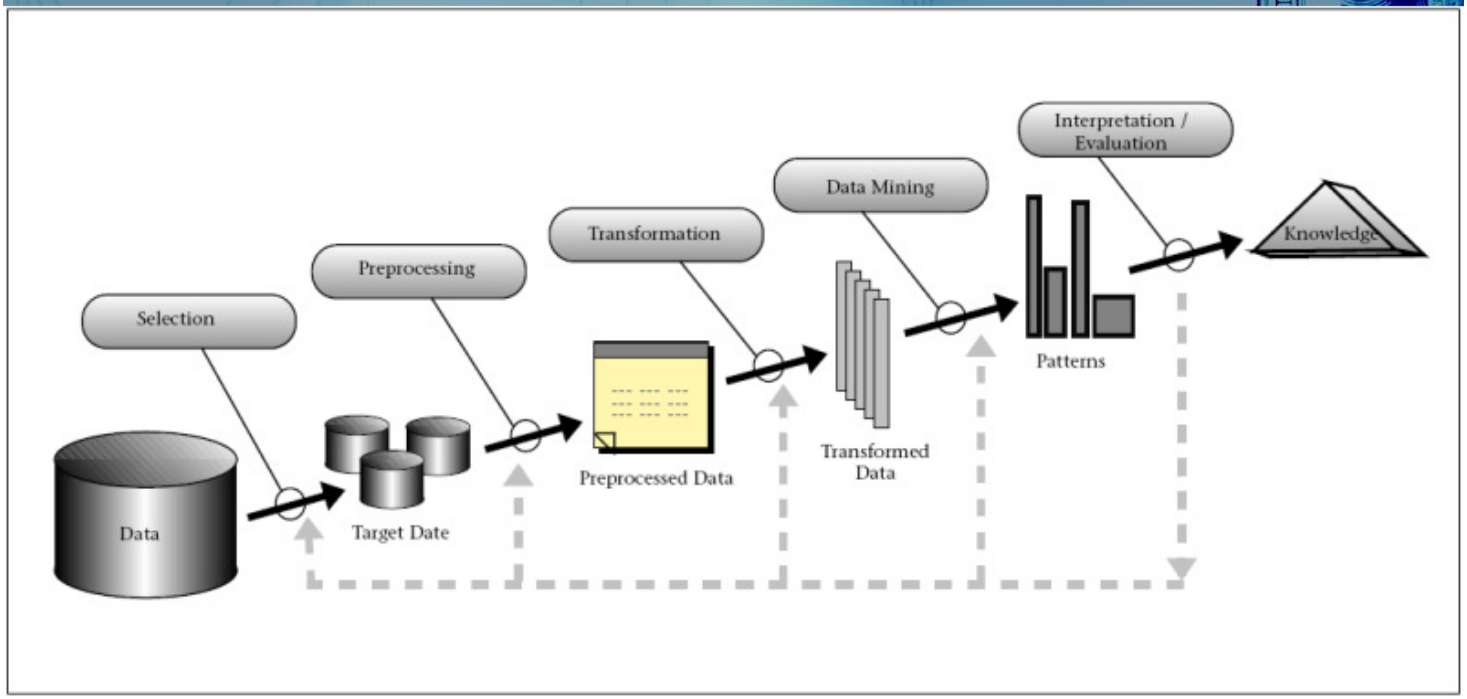
- El crecimiento vertiginoso de las bases de datos disponibles dio lugar, desde finales de los 80, al desarrollo de sistemas que ofrecen funcionalidades más allá de las típicas consultas de un DBMS/SGBD.
- La "minería de datos", englobada en un proceso más amplio de "extracción de conocimiento", permite analizar relaciones entre datos, proporcionar información resumida/agregada/clasificada y presentarla de forma inteligible.

KDD (Knowledge Discovery in Databases)

IDA vs. KDD

- IDA y KDD se consideran sinónimos.
- KDD tiene una connotación más empresarial.
- IDA tiene una connotación más científica.
- La minería de datos en una etapa en ambos procesos.

KDD (Knowledge Discovery in Databases)



Extracción de conocimiento en bases de datos

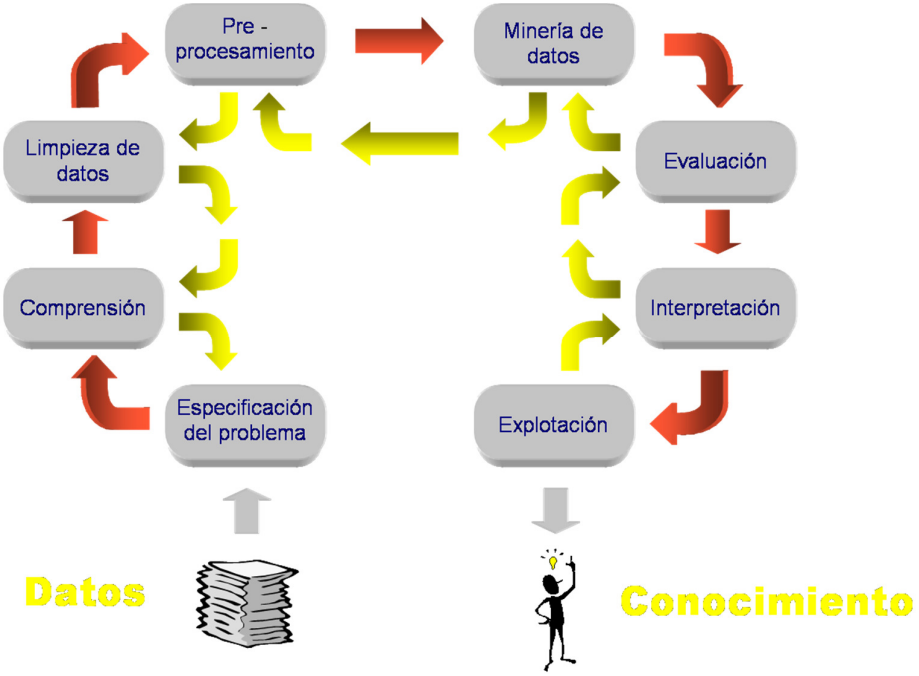
KDD (Knowledge Discovery in Databases)

El proceso de extracción de conocimiento

- Limpieza de datos
  (eliminación de ruido e inconsistencias)
- Integración de datos
  (combinación de múltiples fuentes de datos)
- Reducción/Selección de datos
  (identificación de datos relevantes para el problema)
- Transformación de datos
  (preparación de los datos para su análisis)
- **Minería de datos**
  (técnicas de extracción de patrones y medidas de interés)
- Presentación de resultados
  (técnicas de visualización y de representación del conocimiento)

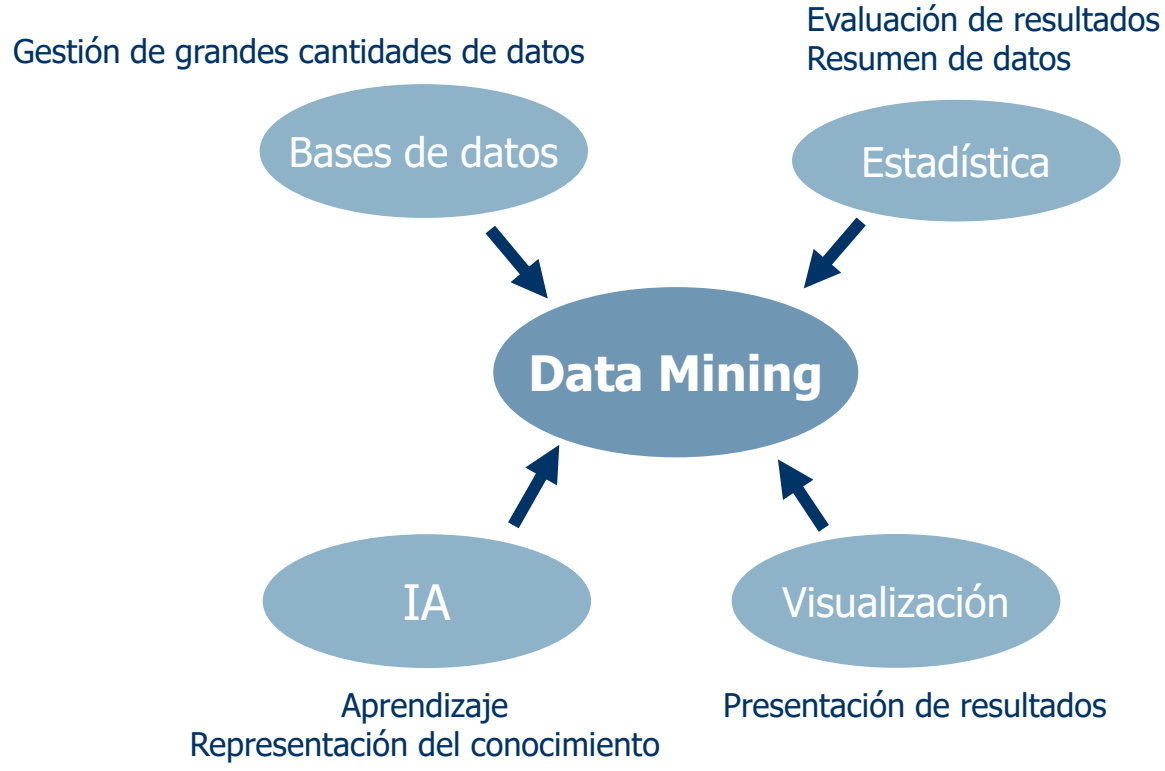
KDD (Knowledge Discovery in Databases)

Extracción de conocimiento en bases de datos:



KDD (Knowledge Discovery in Databases)

Carácter multidisciplinar



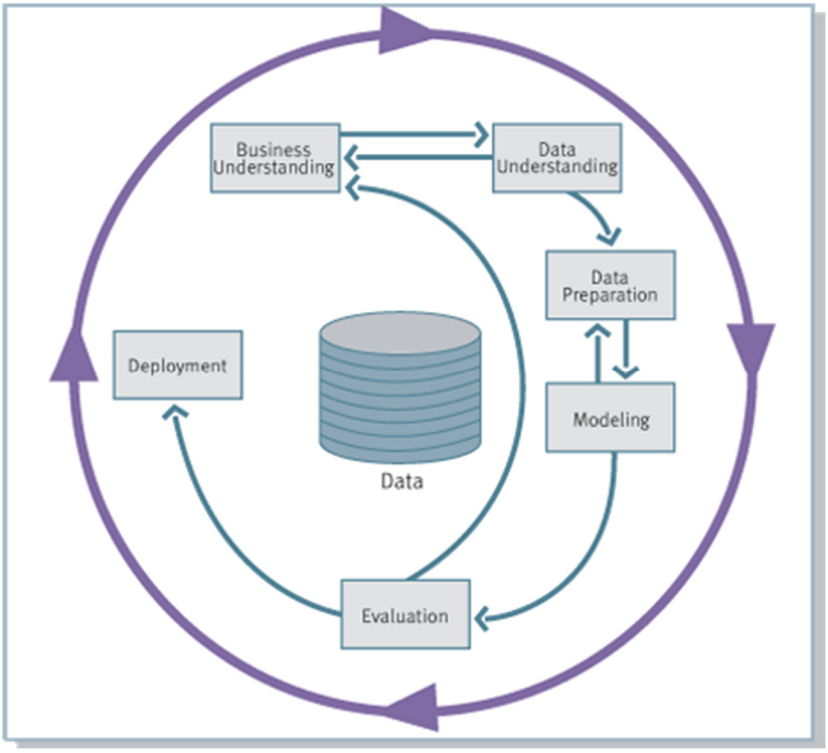
KDD (Knowledge Discovery in Databases)

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades…
Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

Hal R. Varian
Google's Chief Economist
Professor of Information Sciences, Business, and Economics
at the University of California at Berkeley

KDD (Knowledge Discovery in Databases)

CRISP-DM [Cross-Industry Standard Process for Data Mining]



KDD (Knowledge Discovery in Databases)

CRISP-DM

- **Comprensión del proyecto:**
  ¿Cuál es exactamente el problema?, ¿qué beneficios se esperan obtener con su resolución?, ¿qué tipo de solución estamos buscando? ¿qué respuestas nos piden?, ¿qué sabemos acerca del dominio del problema?, ¿cuál es el riego/coste de no resolverlo?

- **Comprensión de los datos:**
  ¿De qué datos disponemos?, ¿son relevantes para el problema?, ¿son fiables?, ¿válidos?, ¿suficientes en términos de calidad, cantidad y actualidad?

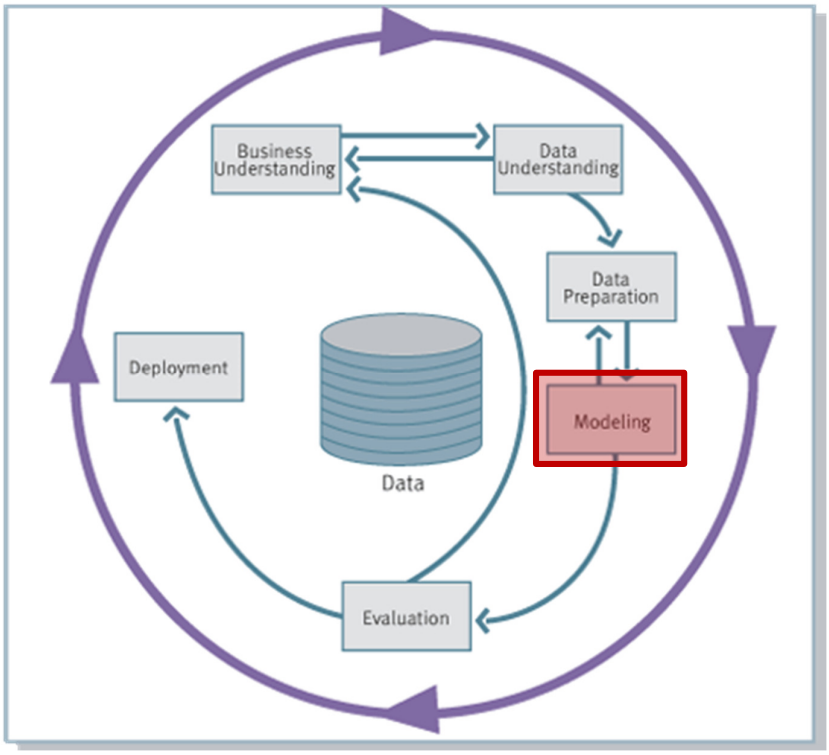
KDD (Knowledge Discovery in Databases)

CRISP-DM

- **Evaluación**
  ¿Satisface el modelo los requerimientos de nuestro proyecto?, ¿qué hemos aprendido acerca de nuestro problema a través del modelo?

- **Implantación / despliegue**
  ¿Cómo puede ser útil el conocimiento adquirido para la toma de decisiones?, ¿cómo puedo saber si el modelo sigue siendo válido?

Modelos de minería de datos



Técnicas de minería de datos

Clasificación de los modelos de minería de datos

En función de su propósito general:

- **Modelos descriptivos**
  (describen el comportamiento de los datos de forma que sea interpretable por un usuario experto).

- **Modelos predictivos**
  (además de describir los datos, se utilizan para predecir el valor de algún atributo desconocido).

Técnicas de minería de datos

Ejemplos

- Reglas de asociación (modelo descriptivo)
  Los compradores de pañales también suelen comprar cerveza.
- Clustering (modelo descriptivo)
  Segmentación de los clientes de un hipermercado:
  - Clientes ocasionales que gastan mucho.
  - Clientes habituales con presupuesto limitado.
  - Clientes ocasionales con presupuesto limitado.
- Clasificación (modelo predictivo):
  - Datagramas que corresponden a intentos de intrusión.
  - Perfil de un cliente de alto riesgo para préstamos bancarios.

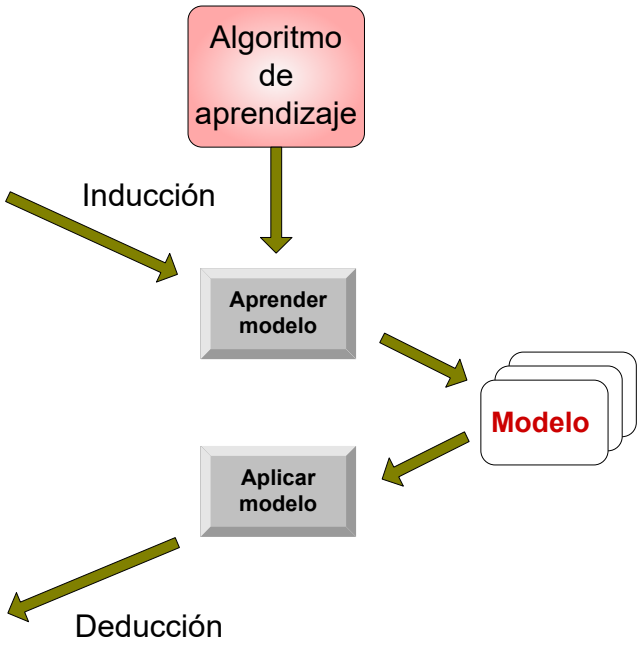
Técnicas de minería de datos

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de entrenamiento

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto de prueba



Aprendizaje automático

Técnicas de minería de datos

Clasificación de las técnicas de aprendizaje

- **Aprendizaje supervisado** (**clasificación & regresión**):
  Los casos del conjunto de entrenamiento aparecen etiquetados con la clase a la que corresponden.

- **Aprendizaje no supervisado** (**asociación & clustering**) :
  No se conocen las clases de los casos del conjunto de entrenamiento (ni siquiera su existencia).

Técnicas de minería de datos

Algunas técnicas de minería de datos

- Caracterización o resumen
- Discriminación o contraste
- Patrones frecuentes, asociaciones y correlaciones
- Clasificación y predicción
- Detección de agrupamientos (clustering)
- Detección de anomalías (outliers)
- Análisis de tendencias (series temporales)
- Análisis de secuencias, p.ej. Bioinformática

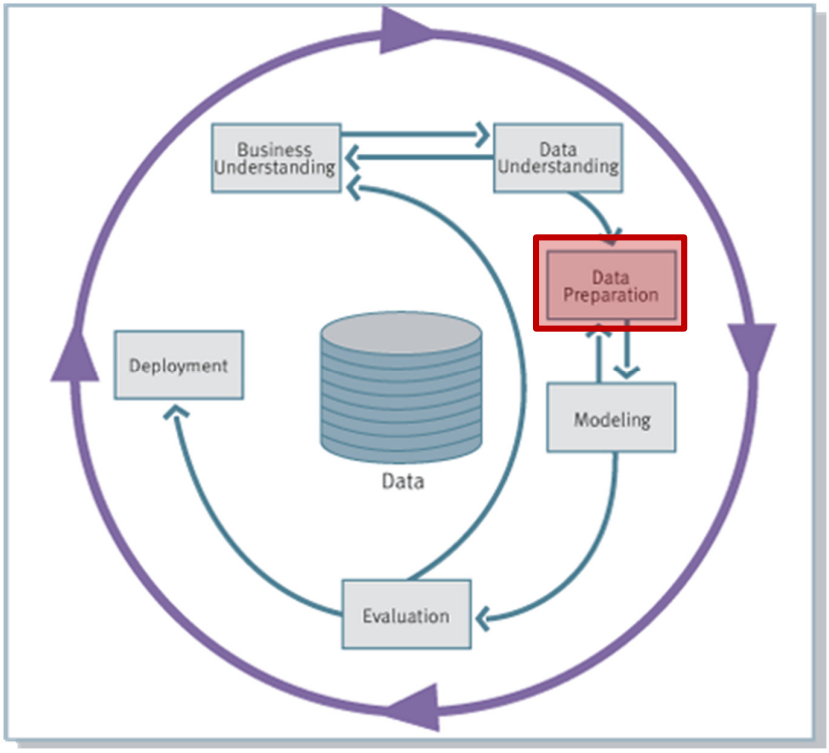
Técnicas de minería de datos

Las técnicas de minería de datos
también se pueden clasificar atendiendo a…

- el tipo de datos que hay que analizar
- el tipo de "conocimiento" que se obtiene
- el tipo de herramienta que se utiliza
- el dominio de aplicación

Fuentes de datos

Fuentes de datos



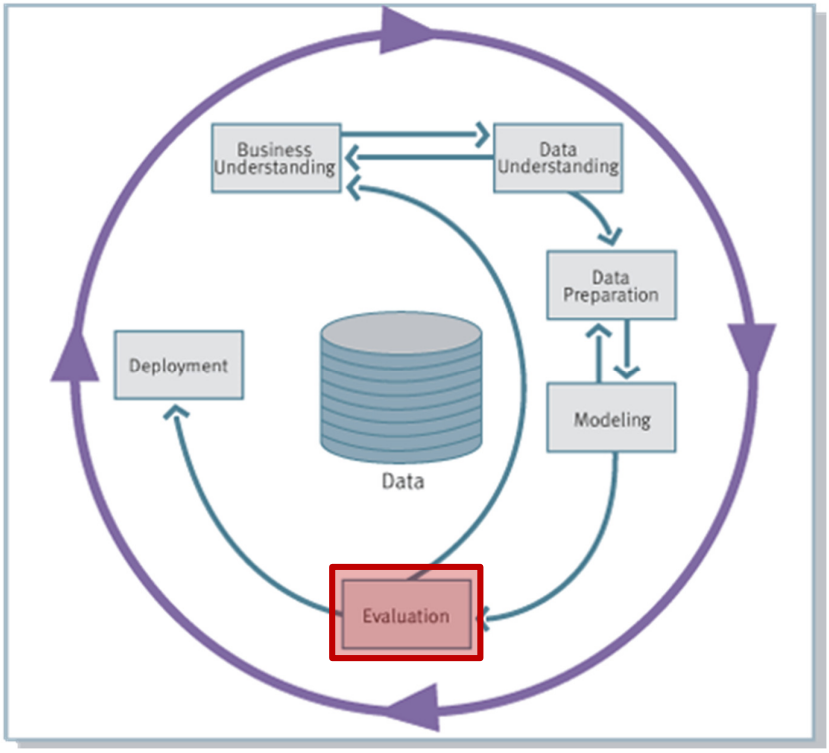
Fuentes de datos

Fuentes de datos

- Bases de datos relacionales
- Bases de datos multidimensionales (DW)
- Bases de datos transaccionales
- Series temporales, secuencias y data streams
- Datos estructurados (grafos, redes sociales)
- Datos espaciales y espaciotemporales
- Textos e hipertextos (p.ej. Web)
- Bases de datos multimedia (p.ej. Imágenes)

Evaluación de resultados

Evaluación de resultados



Evaluación de resultados

Un resultado es interesante si…

- es comprensible (por seres humanos)
- es válido con cierto grado de certeza
- es potencialmente útil
- es novedoso o sirve para validar una hipótesis

El interés de los resultados se puede evaluar

- objetivamente (criterios estadísticos)
- subjetivamente (perspectiva del usuario)

KDD (Knowledge Discovery in Databases)

CRISP-DM
[Cross-Industry Standard Process for Data Mining]

- Concebido en 1996, financiado por la UE en 1997 como un proyecto ESPRIT [European Strategic Programme on Research in Information Technology], publicado en 1999.
- Consorcio con ISL, Teradata, Daimler, NCR y OHRA.
  - ISL [Integral Solutions Ltd.] fue adquirida por SPSS en 1999.
  - SPSS fue adquirida por IBM en 2009.
- Adoptado por IBM en SPSS Modeler.

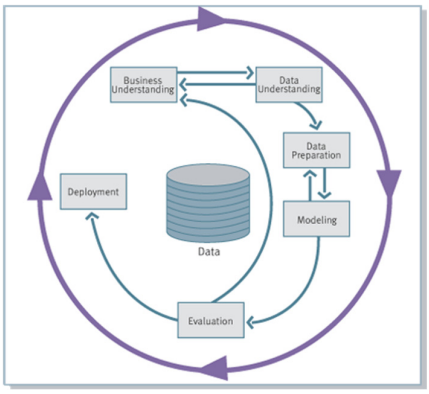
ASUM-DM

[Analytics Solutions Unified Method for Data Mining/Predictive Analytics]

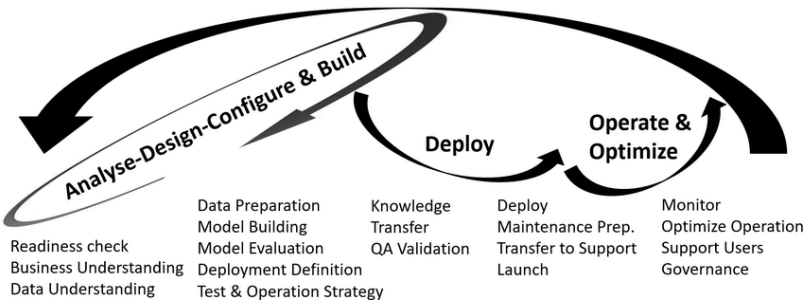
- Extensión de CRISP-DM propuesta por IBM en 2015.

KDD (Knowledge Discovery in Databases)

CRISP-DM

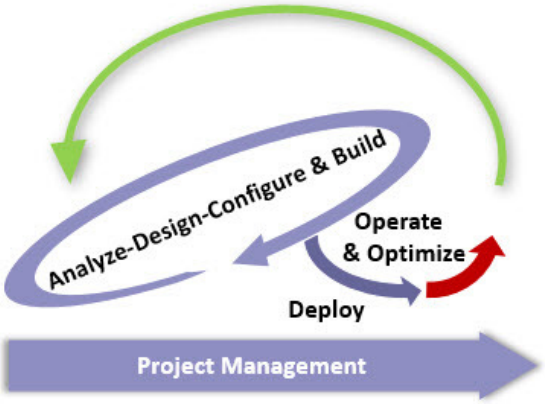


ASUM-DM

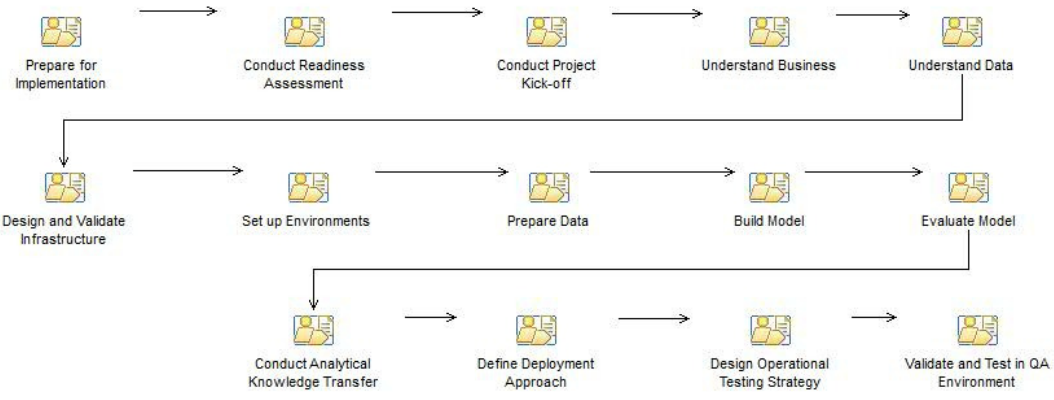


KDD (Knowledge Discovery in Databases)

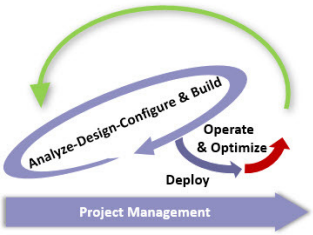
ASUM-DM Workflow



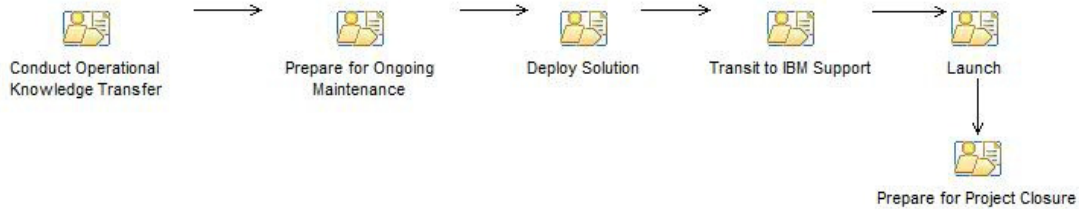
Analyze-Design-Configure & Build



ASUM-DM Workflow



Deploy

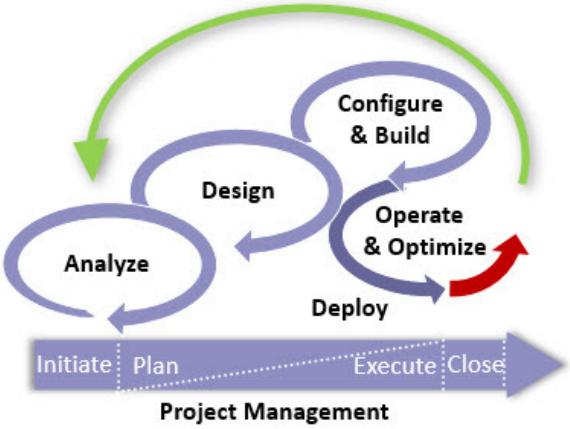


Operate

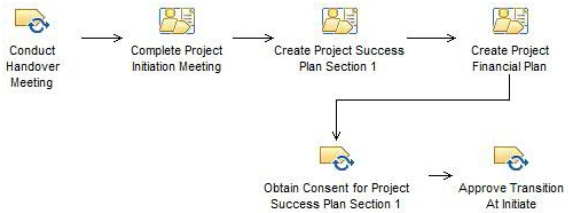
KDD (Knowledge Discovery in Databases)

ASUM-DM Workflow

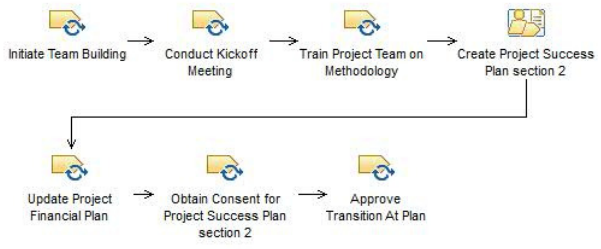
Project Management



- Initiate



- Plan

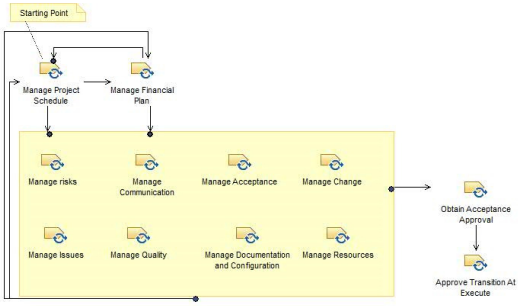


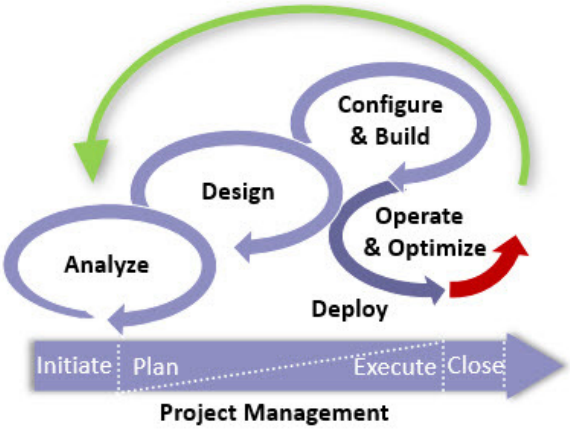
KDD (Knowledge Discovery in Databases)

ASUM-DM Workflow

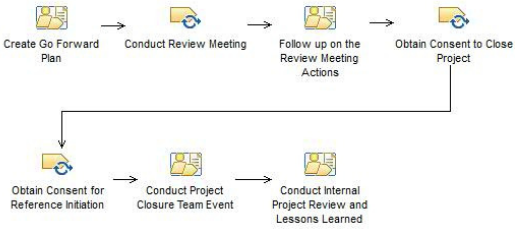
Project Management

- Execute

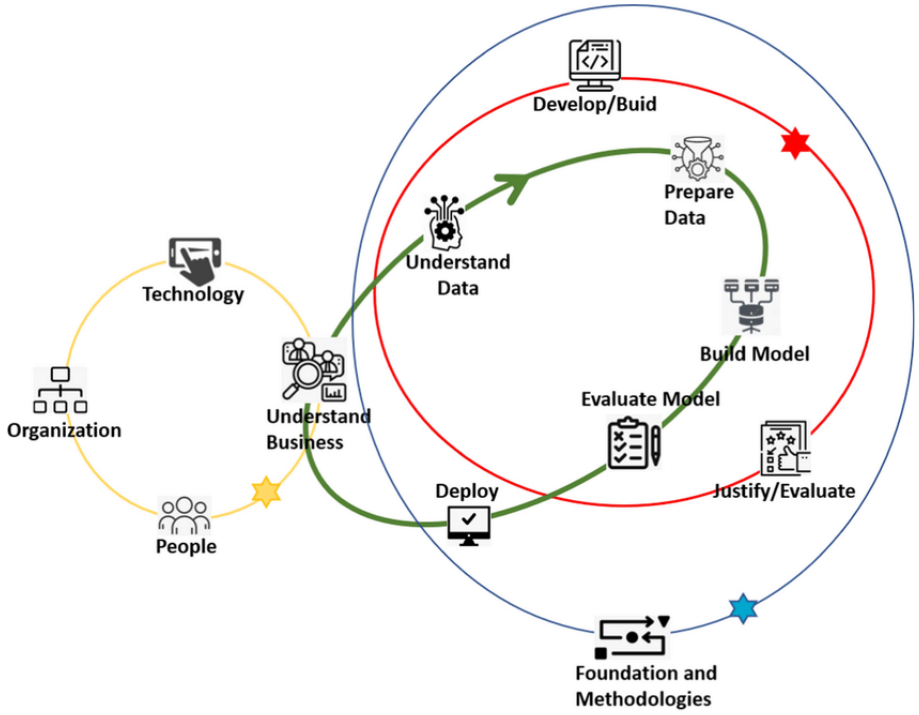




- Close



KDD (Knowledge Discovery in Databases)



Ciencia de datos

- **Análisis de datos [data analysis]**
  1962 John Tukey (Princeton & Bell Labs)
- **Ciencia de datos [data science]**
  1985 C.F. Jeff Wu (Georgia Institute of Technology)
- **KDD [Knowledge Discovery in Databases]**
  1989 Gregory Piatetsky-Shapiro (KDnuggets)
- **Minería de datos [data mining]**
  1990's @ Bases de datos

Ciencia de datos

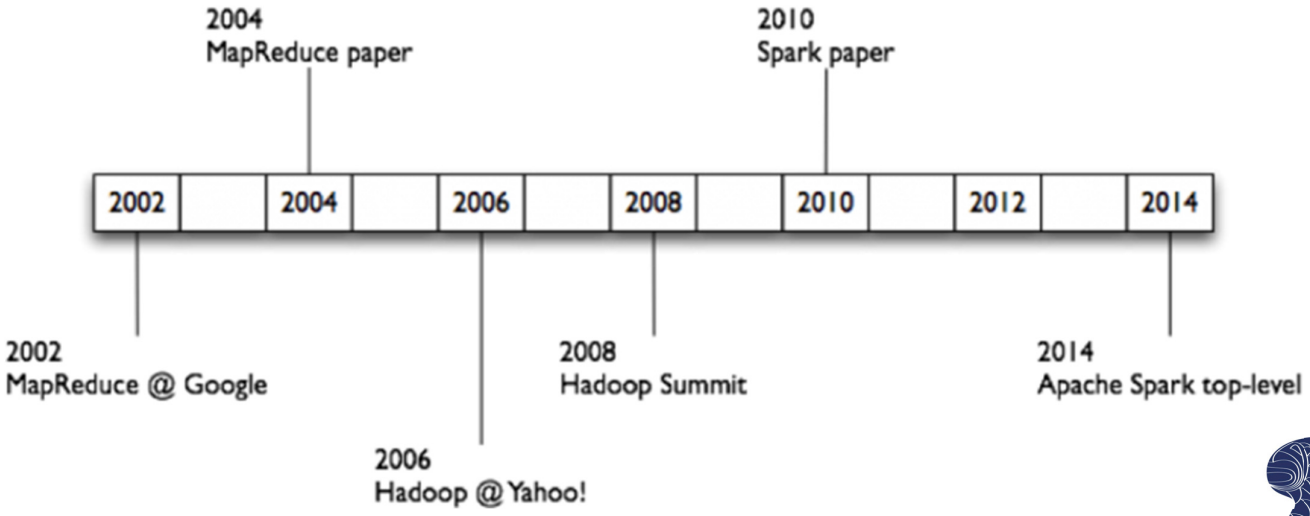
Más denominaciones afines…

- **Inteligencia de negocio [Business Intelligence]**
  1958 Hans Peter Luhn (IBM)
  → 1989 Howard Dresner (Gartner)
- **Aprendizaje automático [ML: Machine Learning]**
  1959 Arthur Samuel (IBM)
  @ IA
- **Reconocimiento de formas [pattern recognition]**
  1967 k-NN (término prestado de psicología)
  @ Visión artificial [CV: Computer Vision]

Big Data

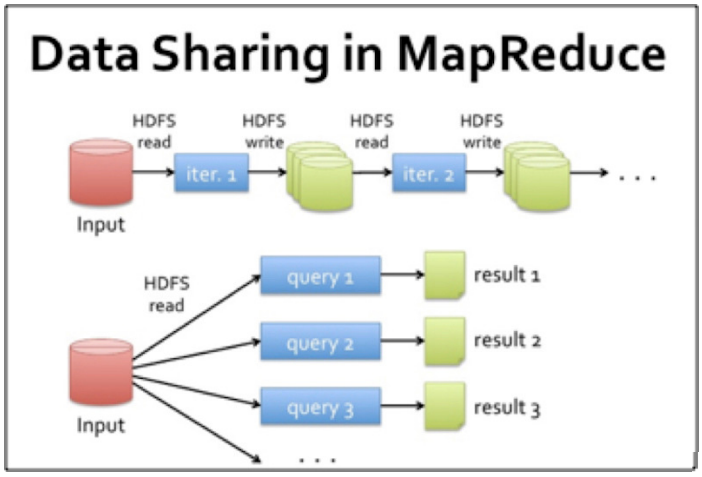
¿Cuándo pasa a ser "big data"?

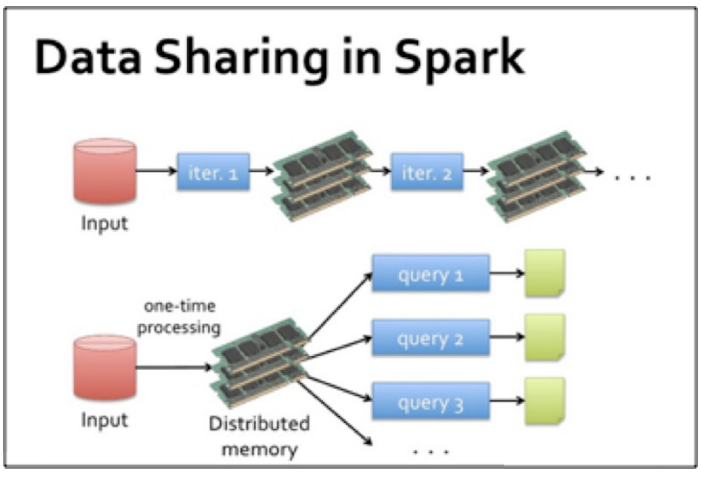
Cuando hacen falta técnicas paralelas para procesar los datos…



Big Data

Hadoop vs. Spark





hadoop MapReduce

Spark
Lightning-Fast Cluster Computing

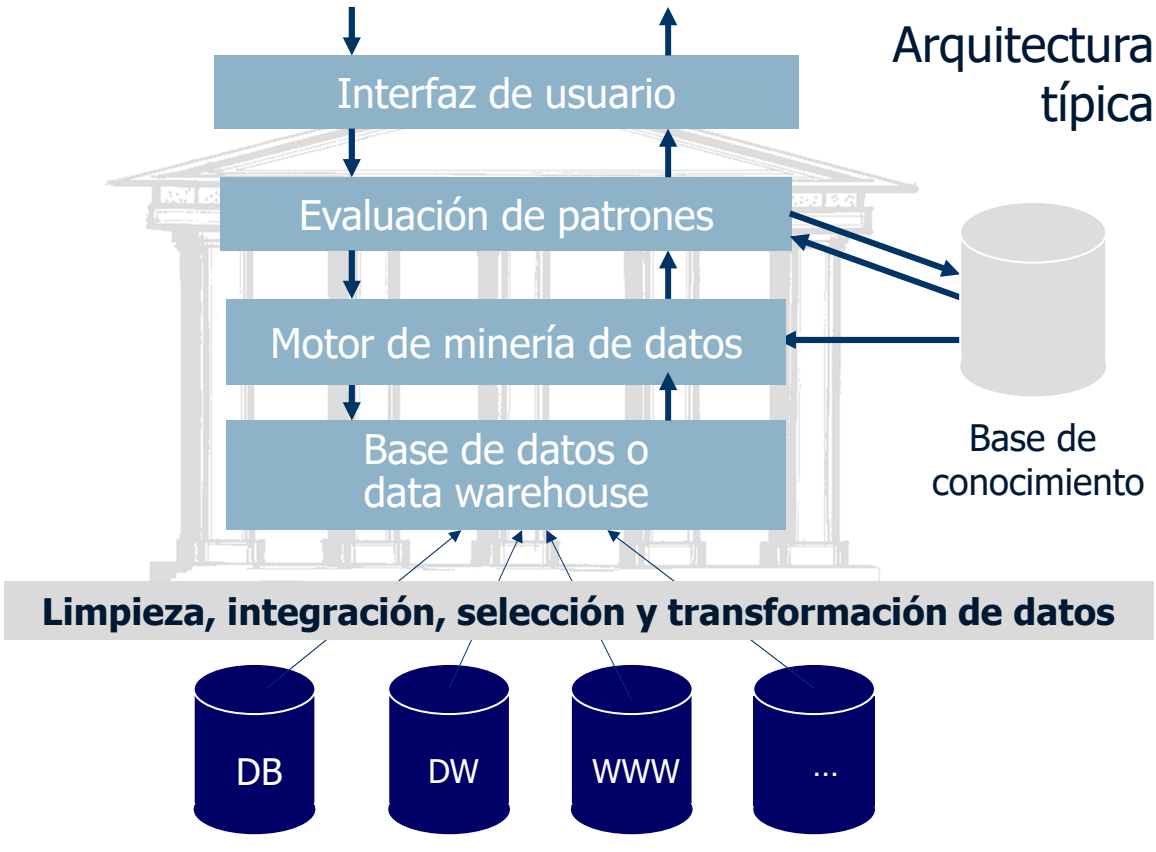
Big Data

Escalabilidad de las técnicas de minería de datos

La importancia de la eficiencia de un algoritmo...

n	10	100	1000	10000	100000
O(n)	10ms	0.1s	1s	10s	100s
$O(n \cdot \log_2 n)$	33ms	0.7s	10s	2 min	28 min
$O(n^2)$	100ms	10s	17 min	28 horas	115 días
$O(n^3)$	1s	17min	12 días	31 años	32 milenios

Sistemas de minería de datos



Sistemas de minería de datos

Descripción de una tarea de minería de datos:

- **Datos relevantes**
  (lo que hay que analizar)
- **Tipo de conocimiento**
  (lo que se desea obtener)
- **Conocimiento previo**
  (*background knowledge*, para guiar el proceso)
- **Medidas de interés**
  (para evaluar los resultados obtenidos)
- **Técnicas de representación**
  (para representar los resultados obtenidos)

Sistemas de minería de datos

Software de minería de datos

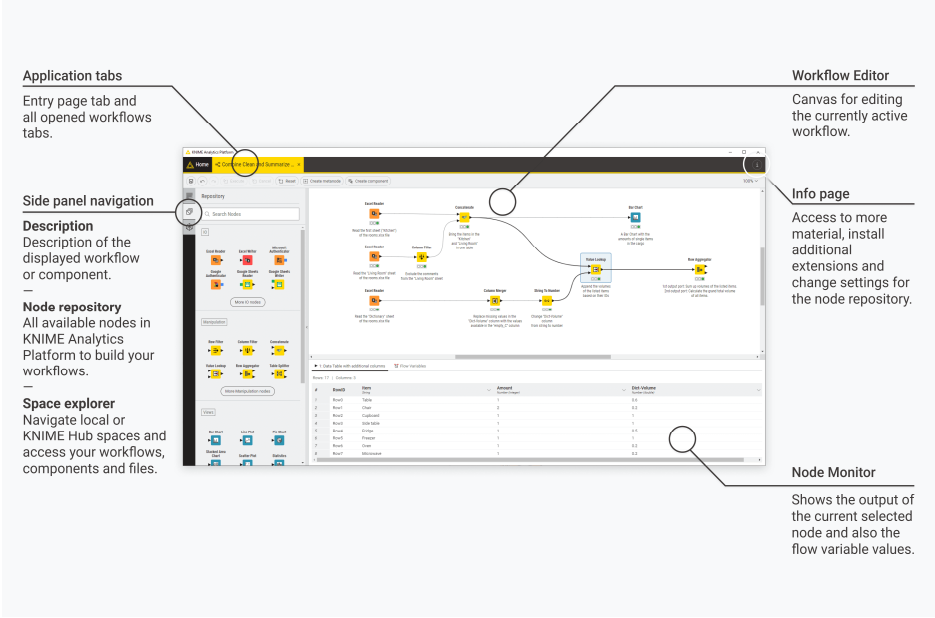
- **KNIME**
  http://www.knime.org/
- **RapidMiner**
  http://rapidminer.com/
- **SPSS Modeler**
  http://www.spss.com/software/modeler/
- **SAS Enterprise Miner**
  http://www.sas.com/
- **Weka**
  http://www.cs.waikato.ac.nz/ml/weka/
- **R & Python**
  http://www.r-project.org/ & https://www.python.org/

Sistemas de minería de datos

KNIME

Konstanz Information Miner @ 2004-

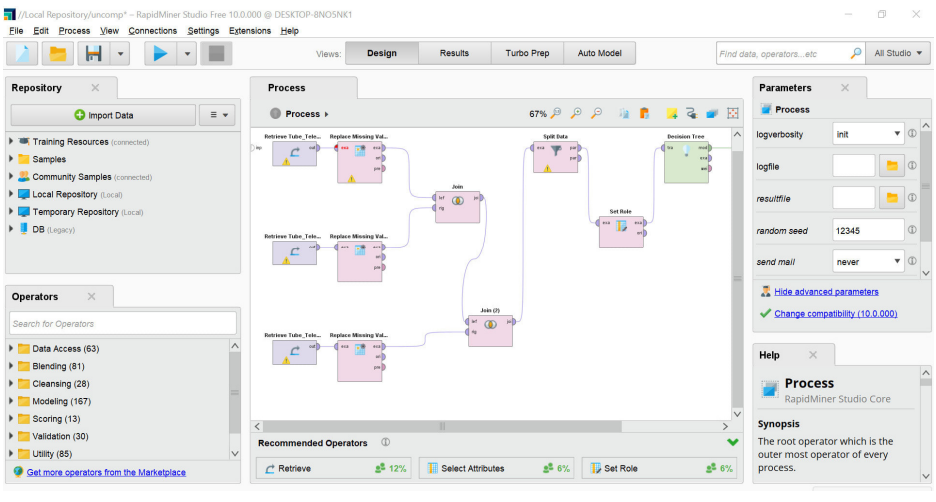
Michael Berthold (University of Konstanz)



Sistemas de minería de datos

RapidMiner

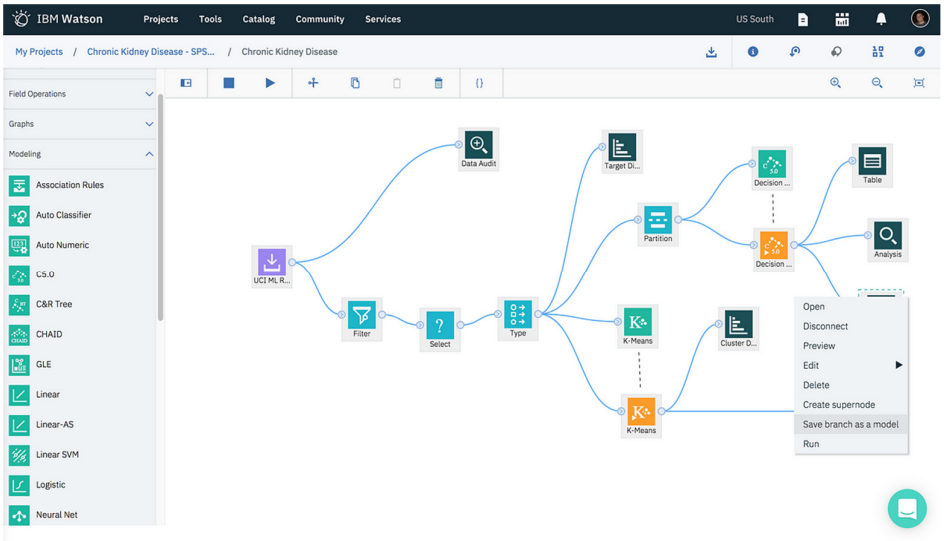
- Adquirida por Altair Engineering @ 2022
- Desarrollada por Rapid-I desde 2006
- Creada como YALE [Yet Another Learning Environment] @ 2001, Technical University of Dortmund



Sistemas de minería de datos

SPSS Modeler

- Clementine (ISL, UK) @ 1994-1999
- Clementine & PASW Modeler (SPSS) @ 1999-2009
- IBM SPSS Modeler (IBM) @ 2010-

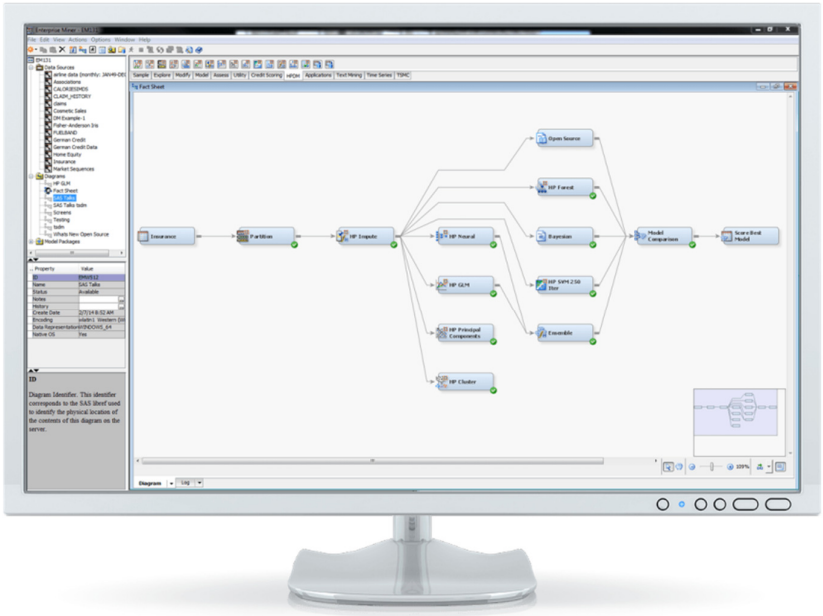


Sistemas de minería de datos

SAS Enterprise Miner

SAS Institute (Cary, NC, USA) @ 1998-

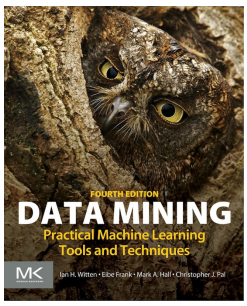
SAS [Statistical Analysis System] @ NCSU desde 1966!!!

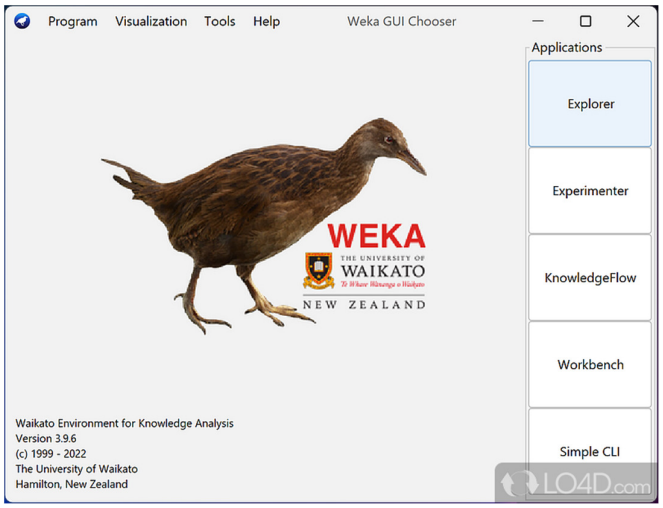


Sistemas de minería de datos

Weka

Waikato Environment for Knowledge Analysis
@ University of Waikato, New Zealand, 1993-





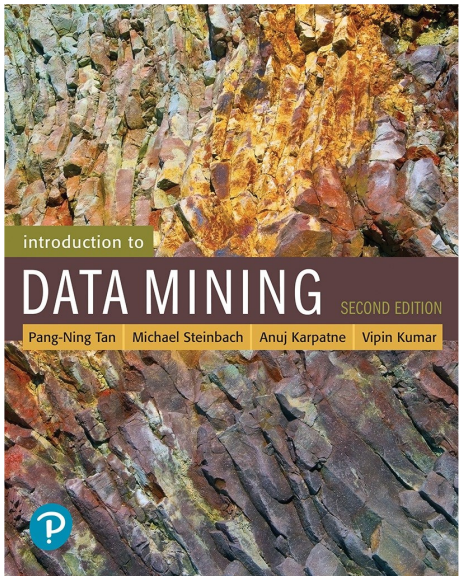
"Data Mining: Practical Machine Learning Tools and Techniques"

Temas de investigación

- Técnicas eficientes de minería de datos
  - Escalabilidad
  - Técnicas incrementales
  - Algoritmos paralelos
- Incorporación de conocimiento previo
- Evaluación de resultados (interés)
- Interacción con el usuario
  - Técnicas interactivas (a distintos niveles de abstracción)
  - Técnicas de presentación y visualización de resultados
- Análisis de "nuevos" tipos de datos
  - Estructuras complejas (grafos, redes sociales)
  - Bases de datos heterogéneas…

Bibliografía

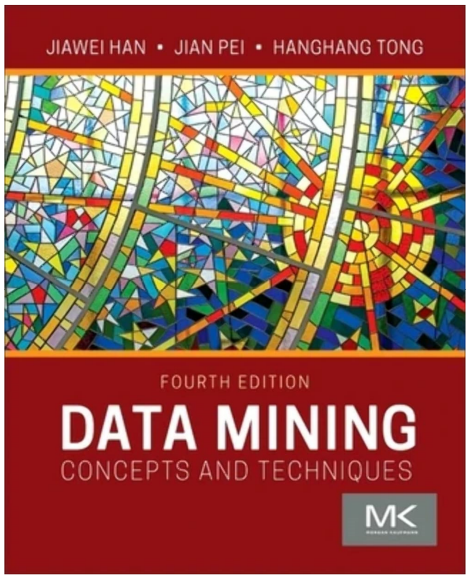
Pang-Ning Tan,
Michael Steinbach,
Vipin Kumar &
Anuj Karpatne:
Introduction to Data Mining,
2nd edition, Addison Wesley, 2018.
ISBN 0133128903



1.1 What Is Data Mining?
1.2 Motivating Challenges
1.3 The Origins of Data Mining
1.4 Data Mining Tasks

Bibliografía

Jiawei Han,
Jian Pei &
Hanghang Tong:
Data Mining:
Concepts and Techniques,
4th edition, Morgan Kaufmann, 2022.
ISBN 0128117605



1.1 What is data mining?
1.2 Data mining: An essential step in knowledge discovery
1.5 Data mining: Confluence of multiple disciplines
1.6 Data mining and applications

Bibliografía (investigación)

Revistas

- ACM Transactions on Knowledge Discovery from Data (TKDD)
- IEEE Transactions on Knowledge and Data Engineering (TKDE)
- Data Mining and Knowledge Discovery (DMKD)
- ACM SIGKDD Explorations
- Data & Knowledge Engineering (DKE)
- Knowledge and Information Systems (KAIS)

Congresos

- KDD (ACM SIGKDD International Conference on KDD)
- ICDM (IEEE International Conference on Data Mining)
- SDM (SIAM Data Mining Conference)
- PKDD (Principles and Practices of KDD)
- SIGMOD (Management of Data)
- CIKM (Information and Knowledge Management)