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

Introducción al Data Mining

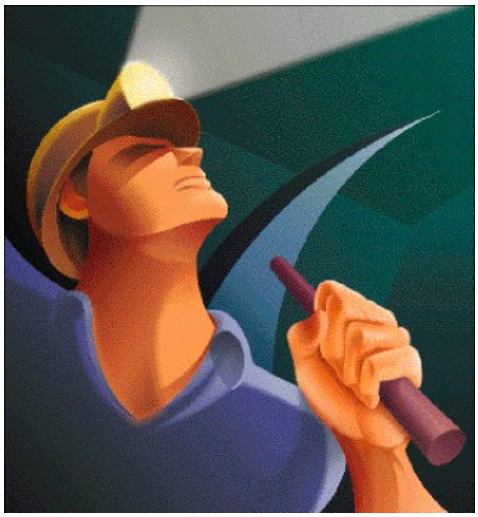
© Fernando Berzal, berzal@acm.org

Introducción al Data Mining

- ¿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
- Sistemas de minería de datos

¿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?

Definiciones

- "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

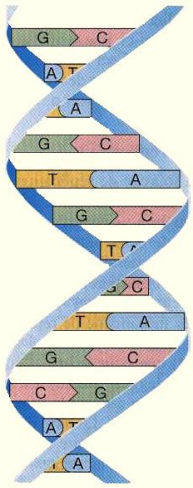
¿Qué es la minería de datos?

Transformar datos en conocimiento

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

Aplicaciones

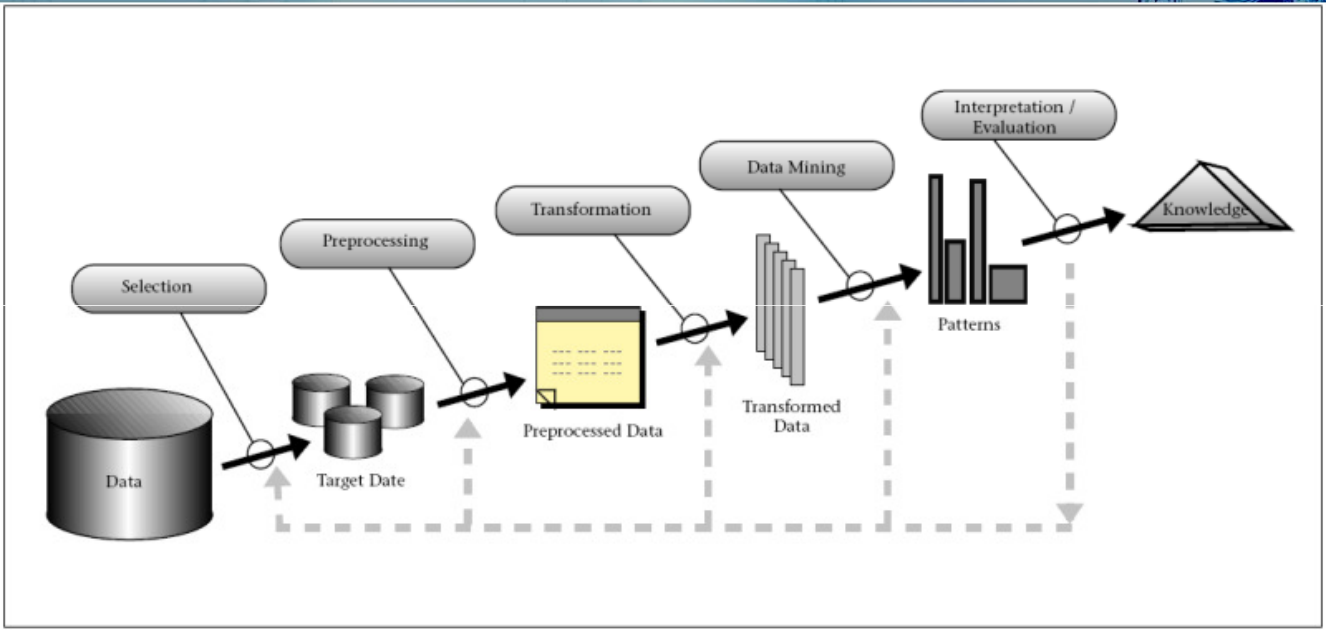
- Market basket analysis (compras)
- Perfiles de usuario en la Web
- Segmentación de clientes
- Detección de fraudes / intrusos
- …



Google

amazon.com

KDD (Knowledge Discovery in Databases)



Extracción de conocimiento en bases de datos

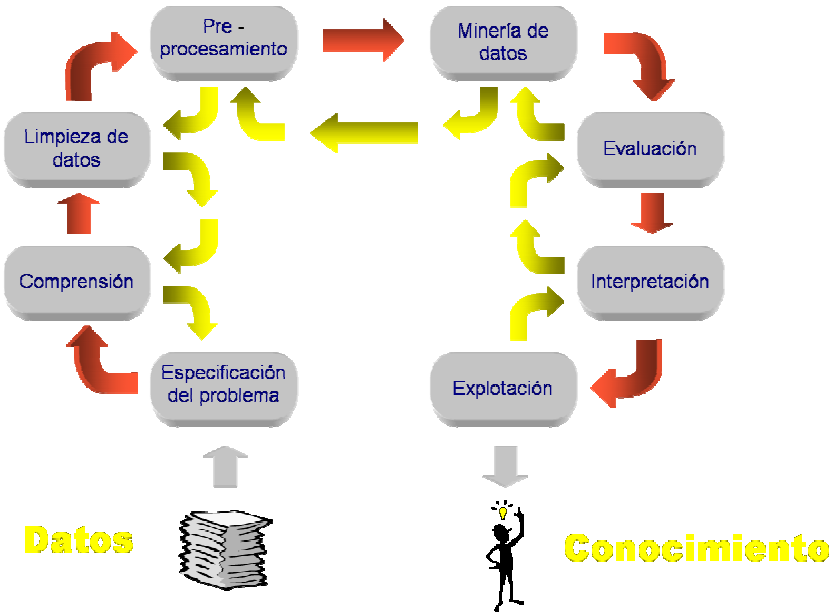
KDD (Knowledge Discovery in Databases)

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)

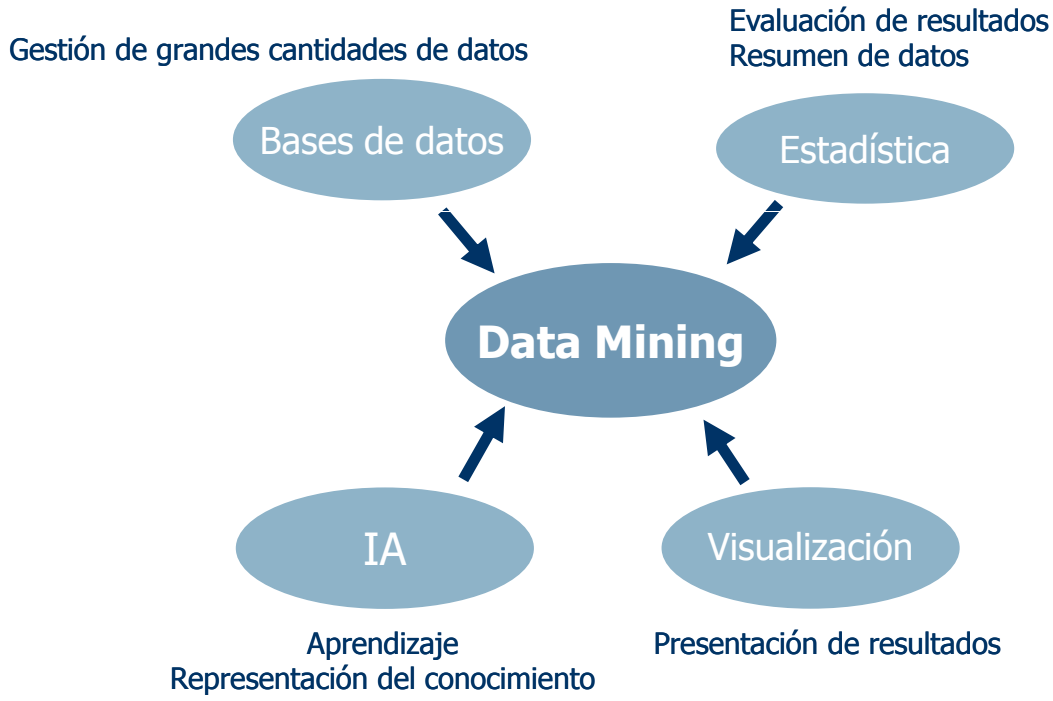
KDD (Knowledge Discovery in Databases)

Extracción de conocimiento en bases de datos:



KDD (Knowledge Discovery in Databases)

Carácter multidisciplinar



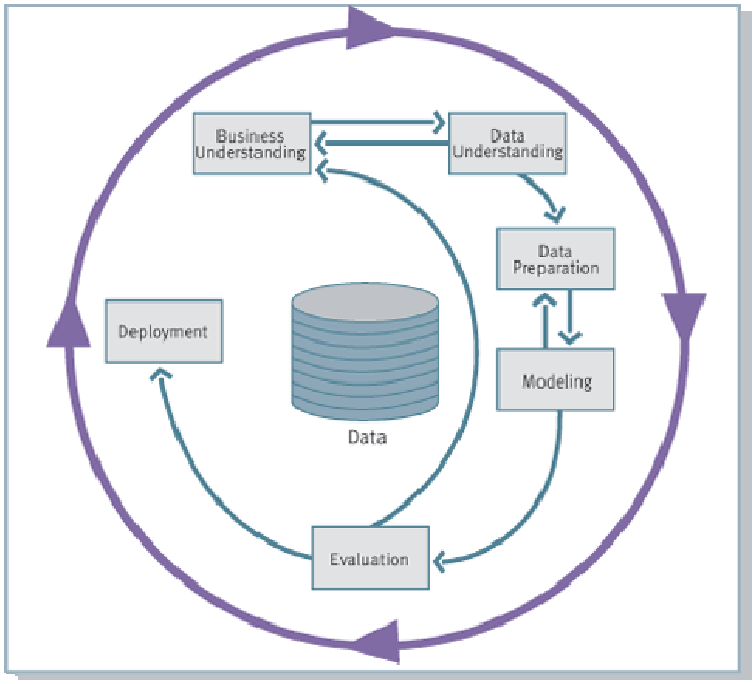
KDD (Knowledge Discovery in Databases)

"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

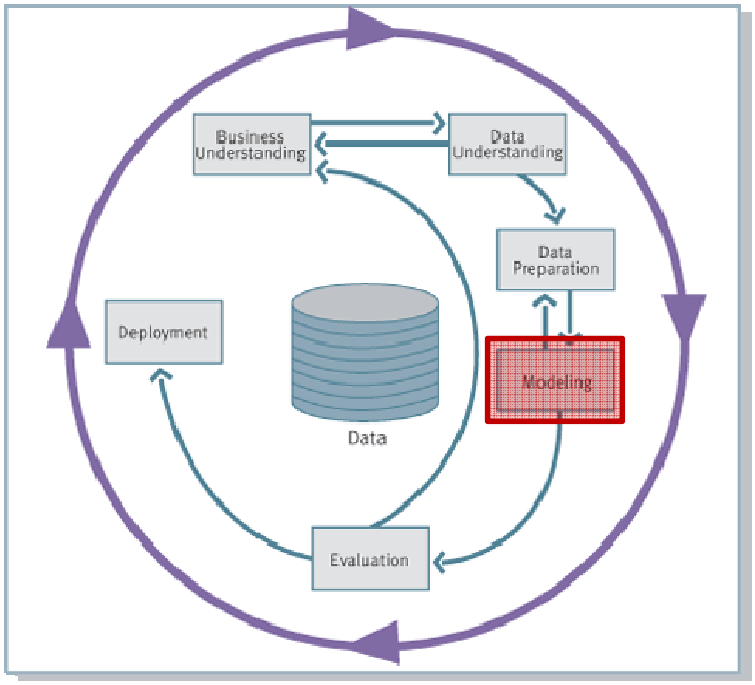
KDD (Knowledge Discovery in Databases)

Extracción de conocimiento en bases de datos:



Técnicas de minería de datos

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

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)

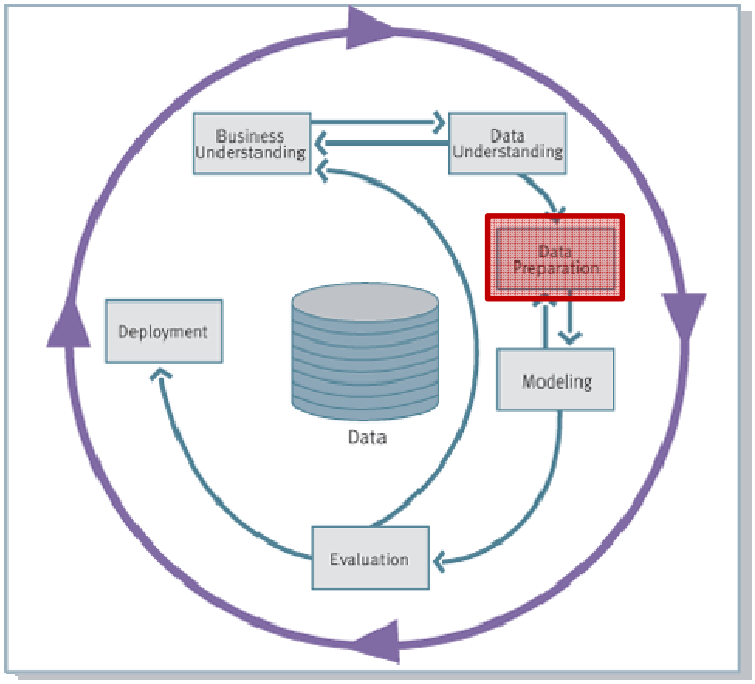
Técnicas de minería de datos

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



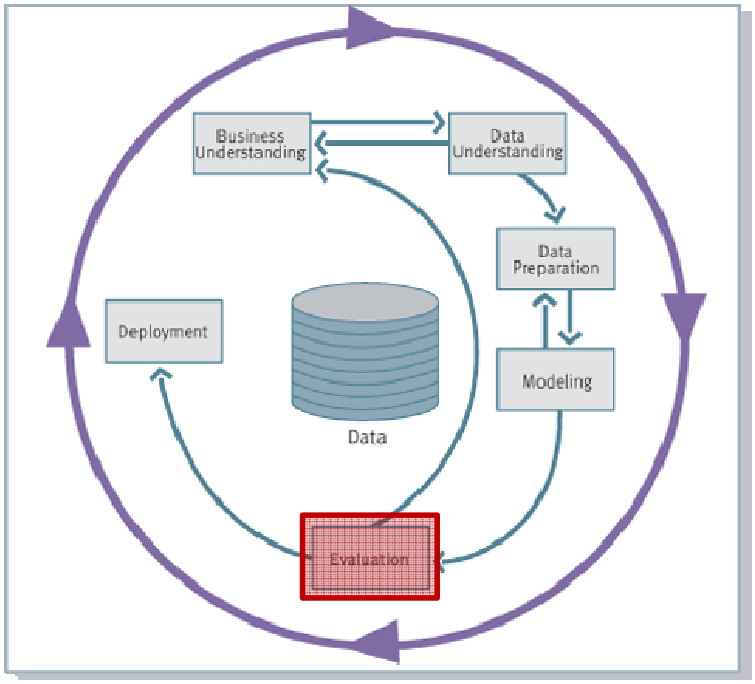
Fuentes de datos

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

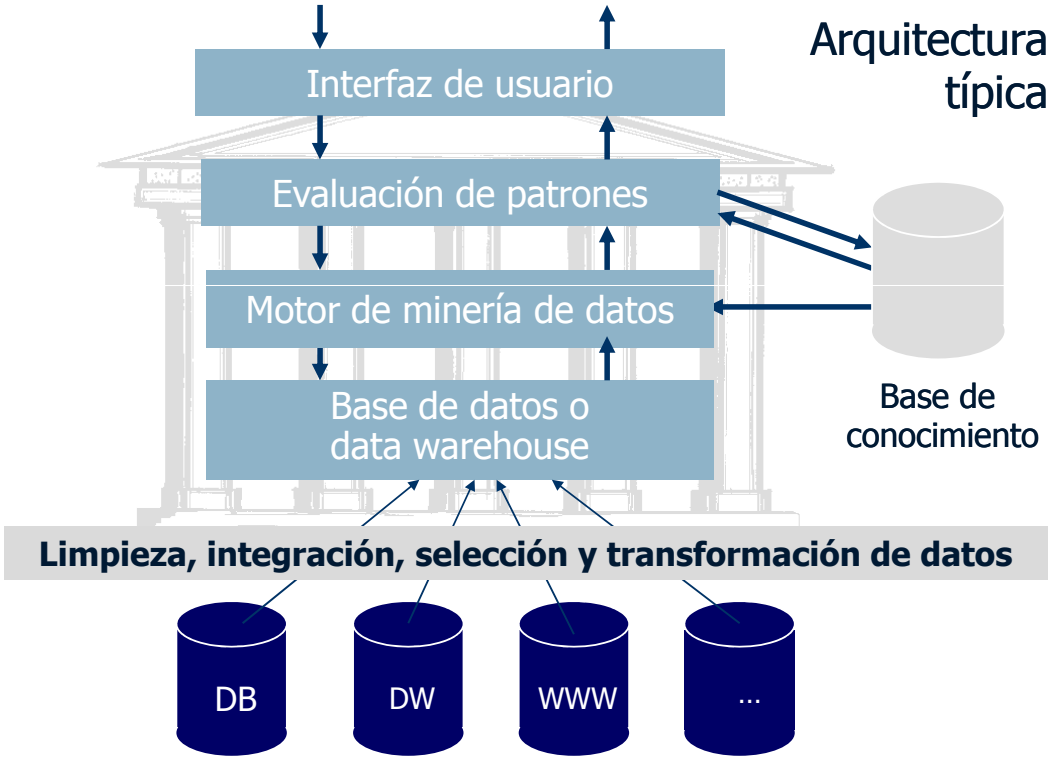
Un resultado es interesante si…

- 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)

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

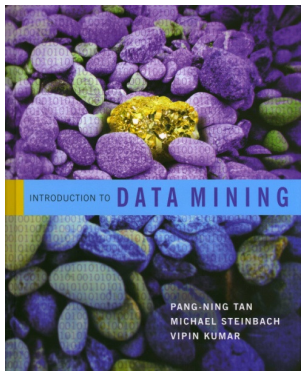
- KNIME
  http://www.knime.org/
- RapidMiner
  http://rapidminer.com/
- Weka
  http://www.cs.waikato.ac.nz/ml/weka/
- R
  http://www.r-project.org/
- SPSS Modeler
  http://www.spss.com/software/modeler/
- SAS Enterprise Miner
  http://www.sas.com/

Temas de investigación

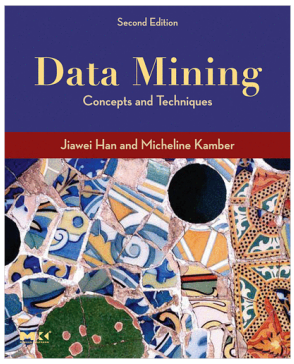
- Técnicas eficientes de minería de datos
  - Escalabilidad
  - Técnicas incrementales
  - Algoritmos paralelos
- Incorporación de conocimiento previo
- Evaluación de resultados (interés)
- Interacción con el usuario
  - Técnicas interactivas (a distintos niveles de abstracción)
  - Técnicas de presentación 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: **Introduction to Data Mining** Addison-Wesley, 2006. ISBN 0321321367



- Jiawei Han & Micheline Kamber: **Data Mining: Concepts and Techniques** Morgan Kaufmann, 2006. ISBN 1558609016



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)