Inteligencia en Redes de Comunicaciones

Tema 7

Minería de Datos

Julio Villena Román, Raquel M. Crespo García, José Jesús García Rueda
{jvillena, rcrespo, rueda}@it.uc3m.es

Universidad Carlos III de Madrid

Índice

- Definición y conceptos
- Técnicas y modelos

Descubrimiento de conocimiento

- Para decidir cuál es la técnica más adecuada para una determinada situación es necesario distinguir el tipo de información que se desea extraer de los datos. Según su nivel de abstracción, el conocimiento contenido en los datos puede clasificarse en distintas categorías y requerirá una técnica más o menos avanzada para su recuperación:

Fuente: “Data Mining. DAEDALUS White Paper”, Daedalus – Data, Decisions and Language, S.A. (www.daedalus.es)

Tipos de conocimiento

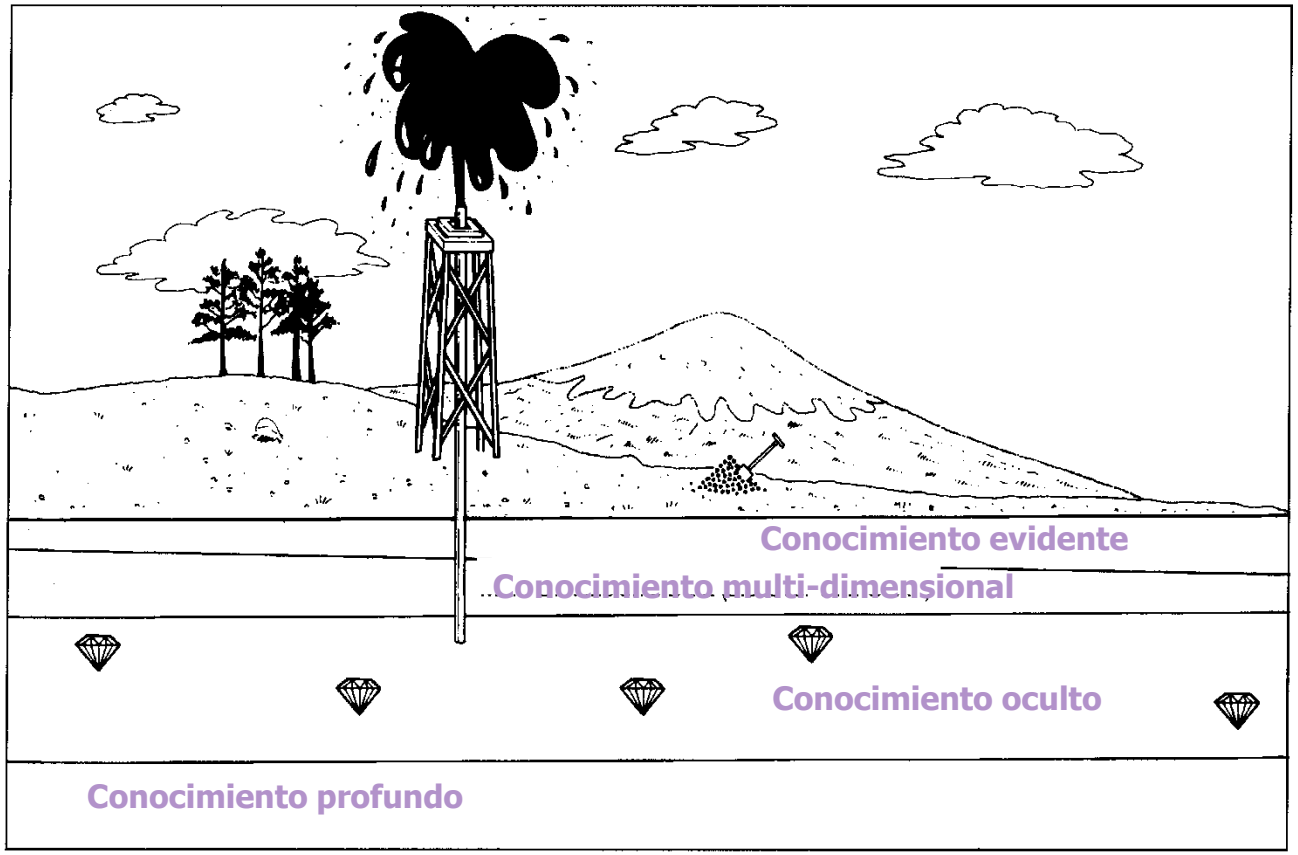


Imagen original: P. Adriaans, D. Zantinge. Addison-Wesley: “Data Mining”, 1996.

Tipos de conocimiento (2)

(según su nivel de abstracción)

- Conocimiento evidente
  - Información fácilmente recuperable mediante una simple consulta SQL
  - Un ejemplo de este tipo de conocimiento es una pregunta como "¿Cuántos alumnos obtuvieron su título en la UC3M en el año 2011?" o "¿Cuál es la nota media de mis alumnos?".
  - Técnica: consulta SQL

Tipos de conocimiento (3)

- (según su nivel de abstracción)

- **Conocimiento multi-dimensional**
  - El siguiente nivel de abstracción consiste en considerar los datos con una cierta estructura.
  - Por ejemplo, en vez de considerar cada transacción individualmente, las ventas de una compañía pueden organizarse en función del tiempo y de la zona geográfica, y analizarse con diferentes niveles de detalle (país, región, localidad...).
  - Técnicamente, se trata de reinterpretar una tabla con n atributos independientes como un espacio n-dimensional, lo que permite detectar algunas regularidades difíciles de observar con la representación monodimensional clásica.
  - Este tipo de información es la que analizan las herramientas OLAP, que resuelven de forma automática cuestiones como "¿Cuáles fueron las ventas en España el pasado marzo? Aumentar el nivel de detalle: mostrar las de Madrid."

  - Técnica: OLAP (análisis multidimensional)

Fuente: "Data Mining. DAEDALUS White Paper", Daedalus – Data, Decisions and Language, S.A. (www.daedalus.es)

Tipos de conocimiento (3)

(según su nivel de abstracción)

- Conocimiento oculto
  - Información no evidente, desconocida a priori y potencialmente útil
  - Que puede recuperarse mediante técnicas de minería de datos, como reconocimiento de regularidades o algoritmos de aprendizaje automático
  - Esta información es de gran valor, puesto que no se conocía y se trata de un descubrimiento real de nuevo conocimiento, del que antes no se tenía idea y que abre la posibilidad de descubrir una nueva visión del problema.
  - Un ejemplo de este tipo de información sería "¿Qué tipos de clientes tenemos? ¿Cuál es el perfil típico de cada clase de usuario?".
  - Técnica: minería de datos

Fuente: "Data Mining. DAEDALUS White Paper", Daedalus – Data, Decisions and Language, S.A. (www.daedalus.es)

Tipos de conocimiento (3)

(según su nivel de abstracción)

- Conocimiento profundo
  - Información que está almacenada en los datos, pero que resulta imposible de recuperar a menos que se disponga de alguna clave que oriente la búsqueda
  - Un ejemplo típico sería un mensaje cifrado. Es fácil recuperar la información codificada si se dispone de la clave, pero imposible o muy difícil si no se tiene.

Aprender…

Objetivo:

- Construir un sistema computacional que sea capaz de encontrar y modelar el conocimiento oculto que a los seres humanos nos resulta difícil ver

¿Cómo?

- Dotando a ese sistema de algoritmos o técnicas que imiten la cualidad humana del aprendizaje, esto es, ser capaz de extraer nuevos conocimientos a partir de las experiencias (ejemplos)

KD

Knowledge Discovery *is the nontrivial extraction of implicit, previously unknown and potentially useful information from data*

W. J. Frawley,
G.Piatetsky-Shapiro,
C. J. Matheus

KDD

Knowledge Discovery in Databases: nombre técnico con que se denomina al proceso global de extracción de conocimiento de bases de datos

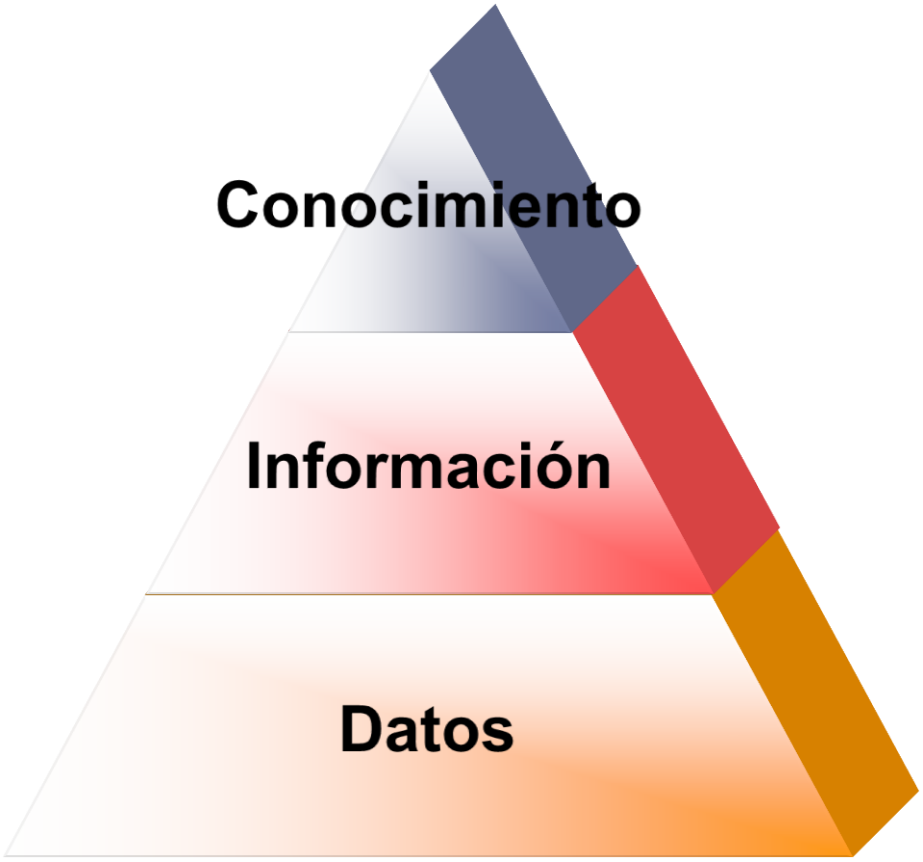
Data Mining

La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos con vistas al descubrimiento de información previamente desconocida y que pueda servir de ayuda en el proceso de toma de decisiones, formando parte del conjunto de tecnologías de la Inteligencia de Negocio

Business Intelligence

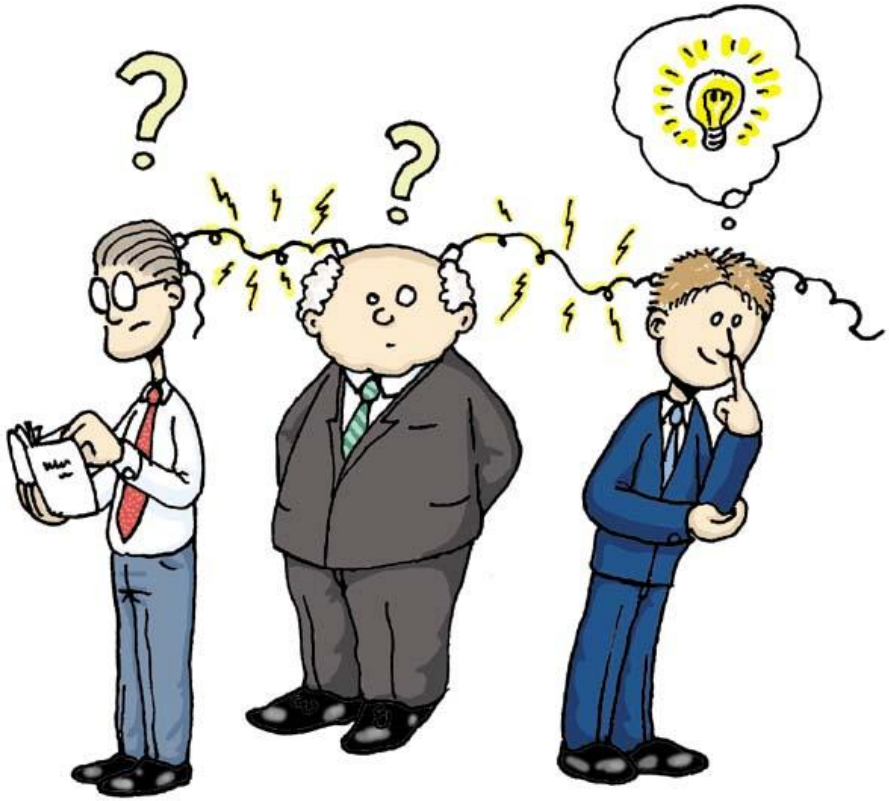
Realización eficiente de todas las actividades relacionadas con la generación, extracción, organización, análisis, compartición y distribución del conocimiento de una organización

Datos, Información y Conocimiento



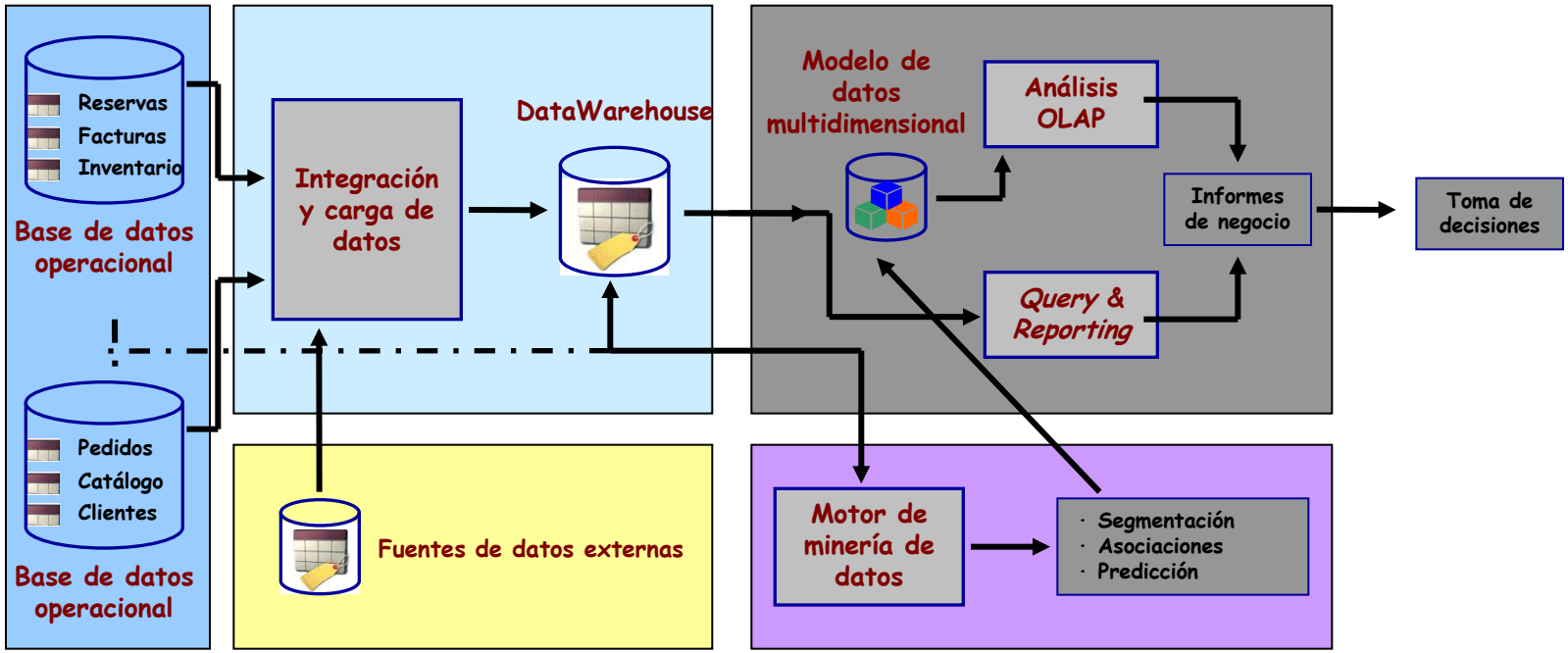
Conocimiento: capacidad de convertir datos e información en acciones efectivas

Objetivo



Poner al alcance
de cada individuo
lo que necesita
en el momento preciso
para que su actividad
sea efectiva

Arquitectura tecnológica



Verificación vs. descubrimiento

Verificación

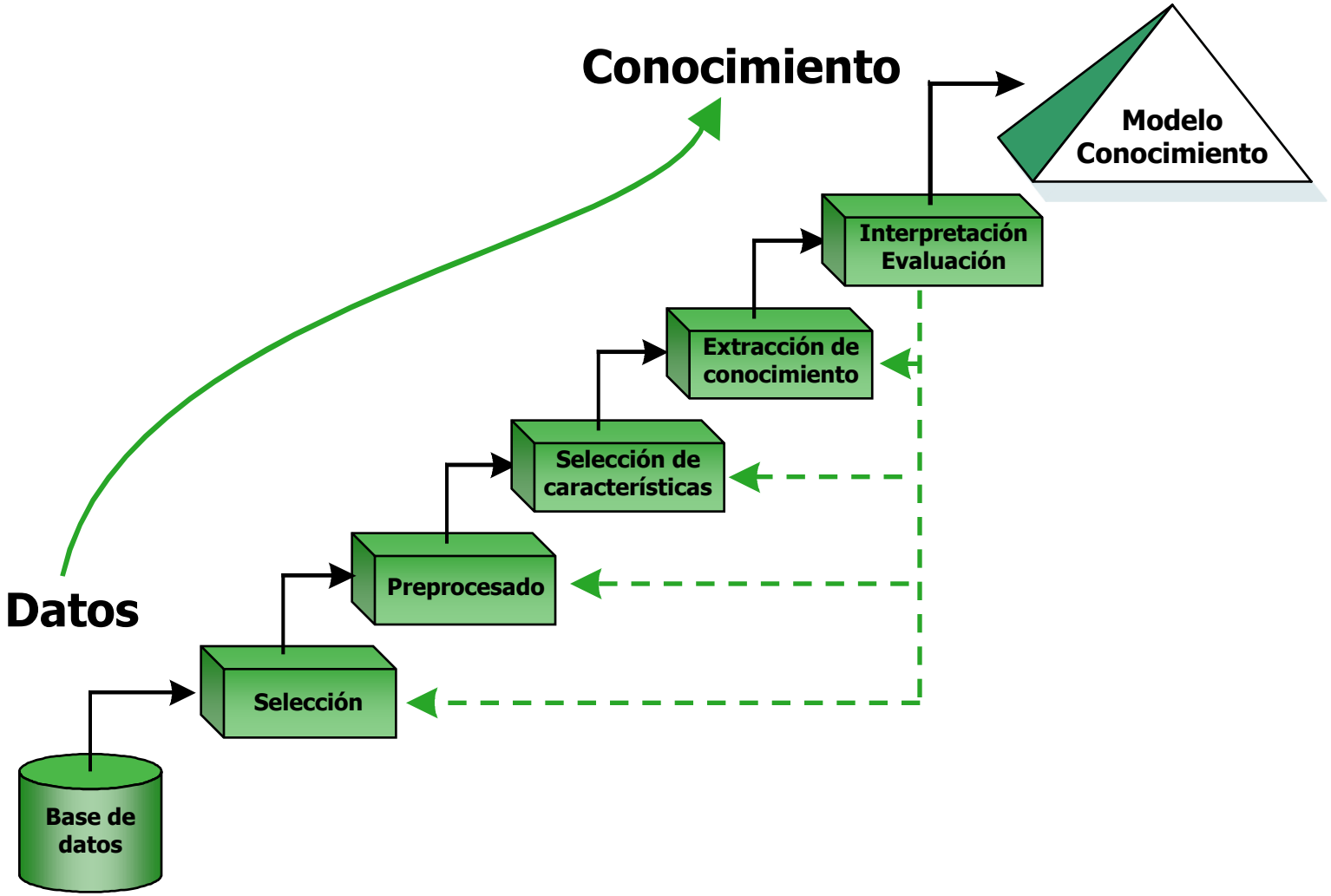
1. Elaborar una hipótesis sobre la existencia de una información de interés
2. Convertir la hipótesis en una consulta
3. Ejecutar la consulta contra un sistema de información
4. Interpretar los resultados
5. Refinar la hipótesis y repetir la ejecución

Descubrimiento

1. Identificar un objetivo o problema de negocio
2. Habilitar un acceso a los datos de interés y acondicionarlos
3. Seleccionar una técnica de explotación de los datos adecuada para el problema
4. Ejecutar la técnica contra los datos
5. Interpretar los resultados

Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de la información

Proceso de minería de datos



Metodología CRISP-DM
CRoss Industry Standard Process for Data Mining

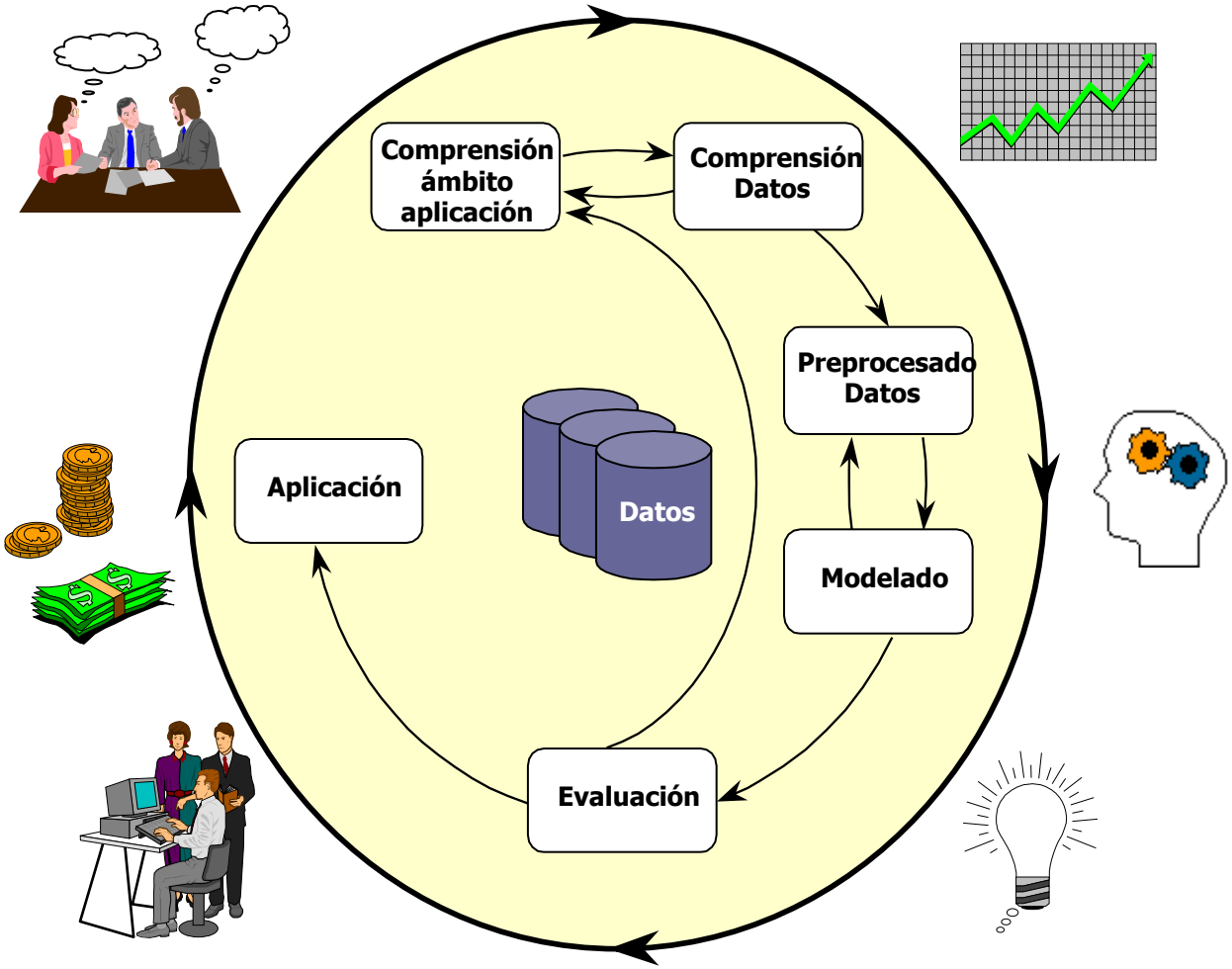


Imagen original: http://www.crisp-dm.org/

Dominios de aplicación

- Aplicaciones en campos muy diversos
  - Medicina
  - Economía
  - Comercio
  - Marketing
  - Telecomunicaciones
  - Seguridad
  - Etc.

Herramientas

- Comerciales
  - Intelligent Miner / DB2 Data Warehouse Edition (IBM)
  - Clementine (SPSS)
  - Enterprise Miner (SAS)
  - DataEngine
- De código libre
  - Weka

IBM Intelligent Miner

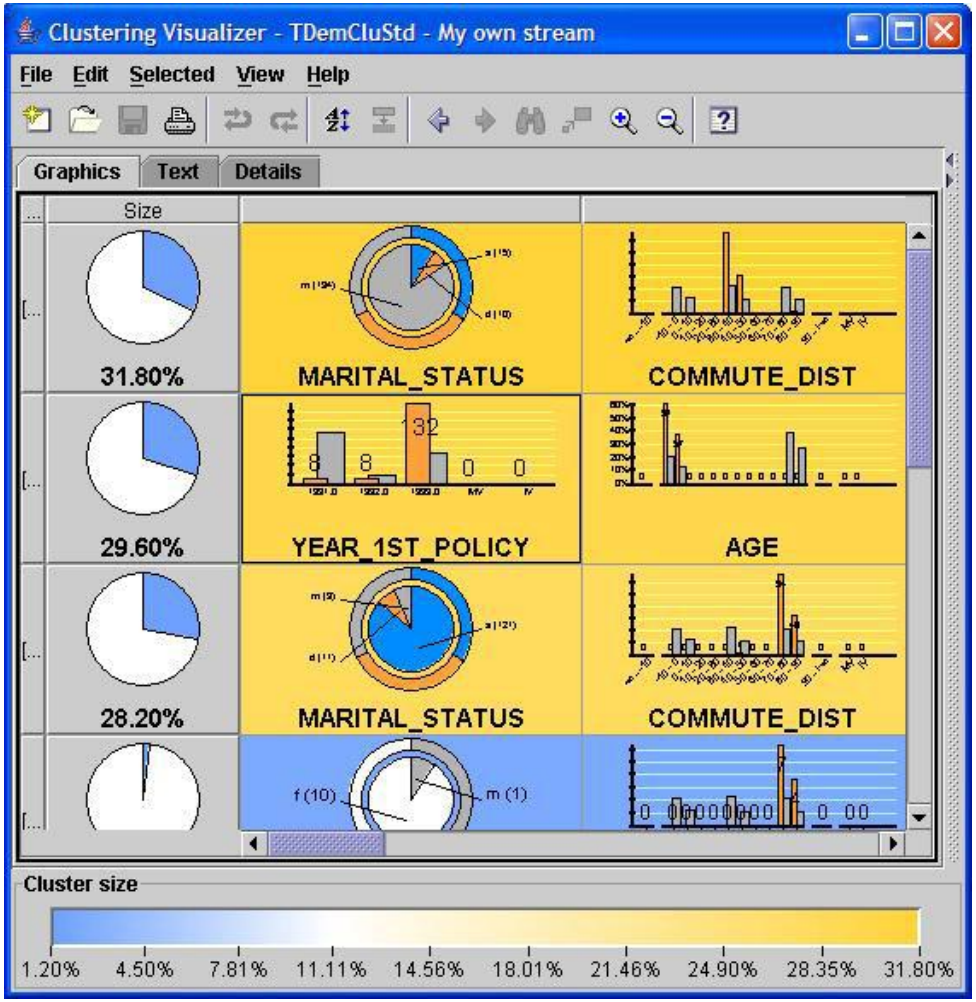


Imagen original: ibm.com

SPSS Modeler (antes Clementine)

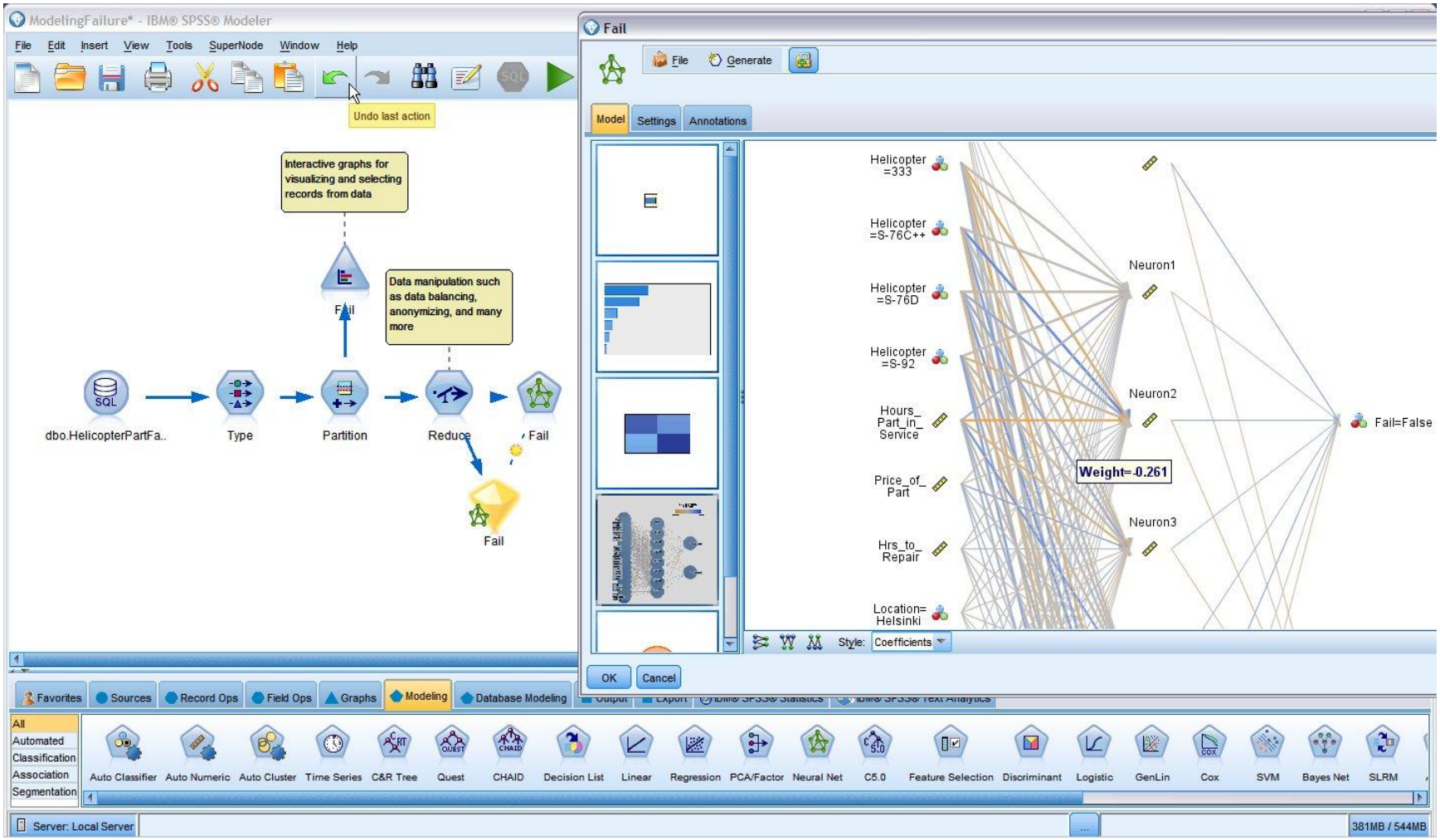
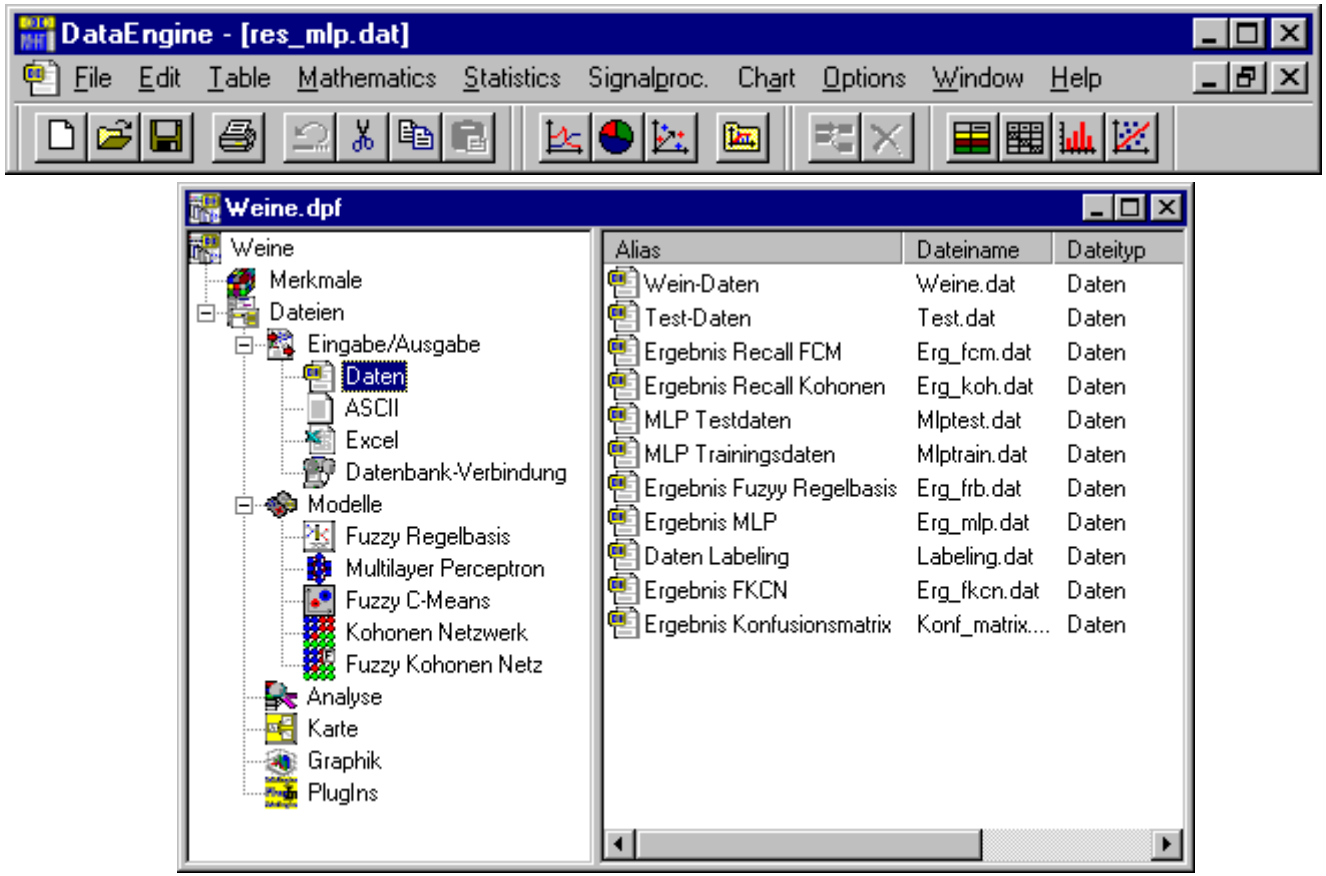


Imagen original: ibm.com

SAS Enterprise Miner

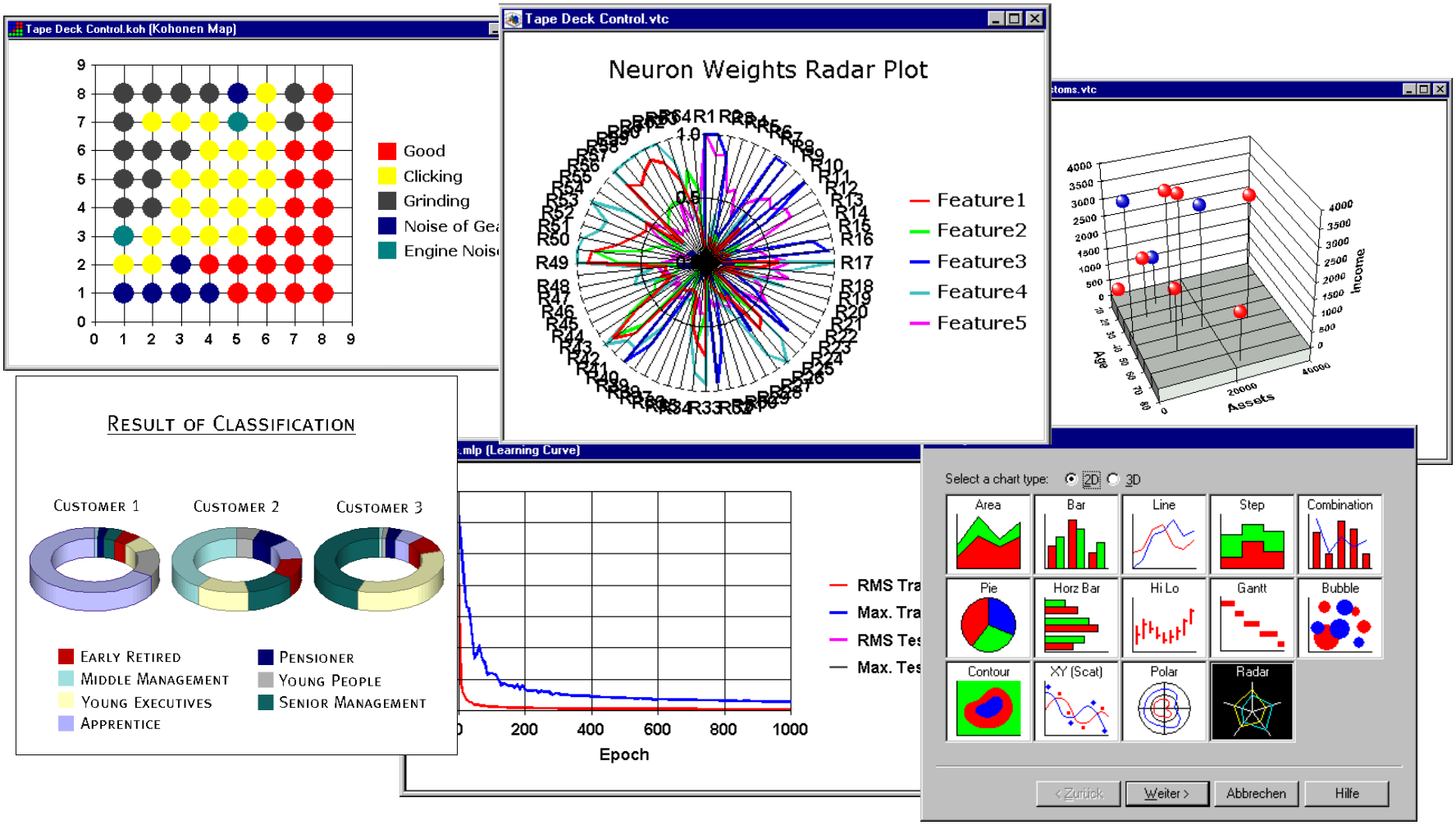
Imagen original: sas.com

MIT DataEngine



Fuente: DataEngine, MIT GmbH

MIT DataEngine (2)



Fuente: DataEngine, MIT GmbH

Weka

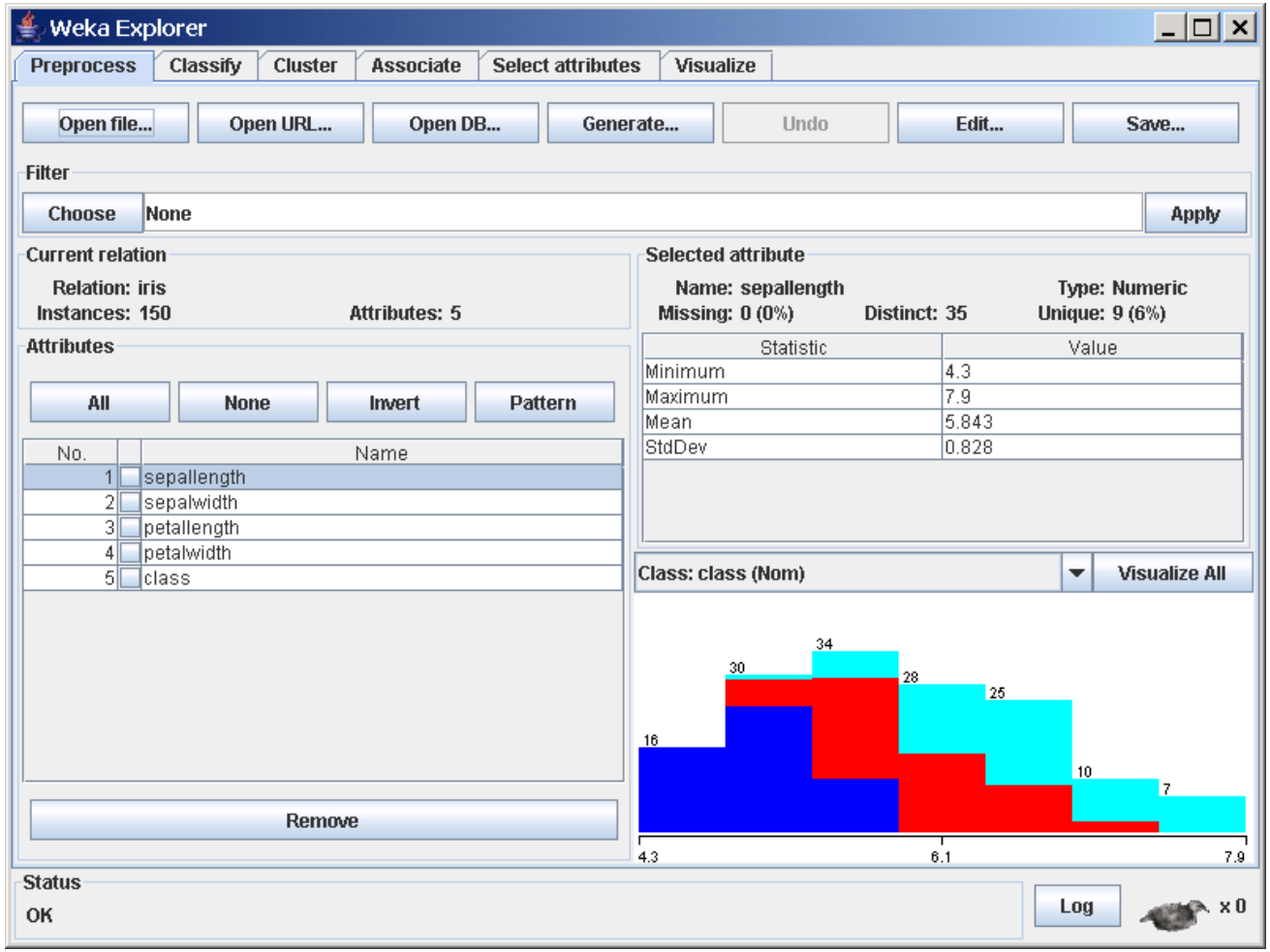
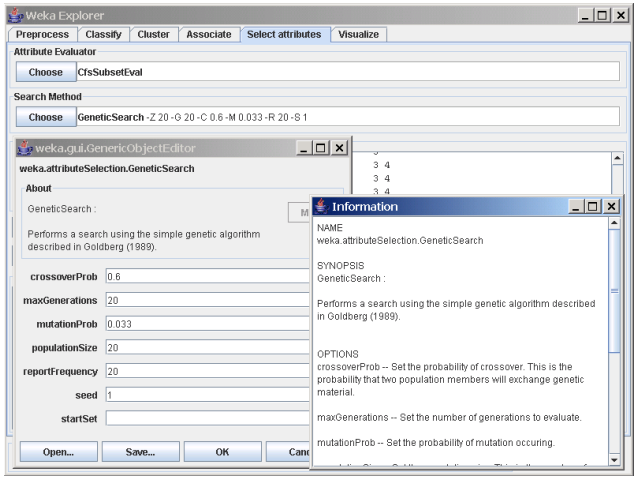
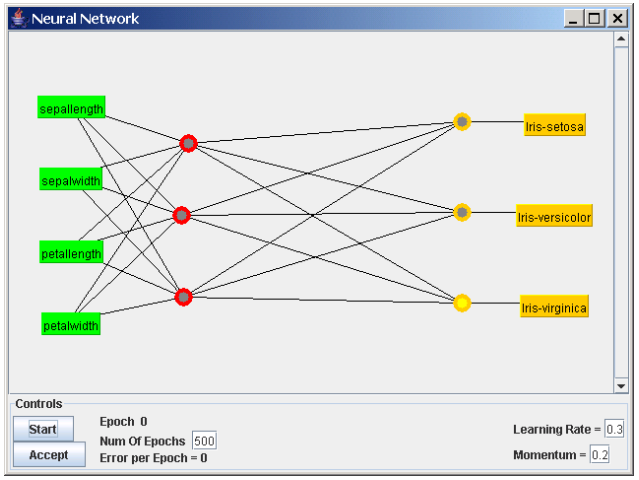


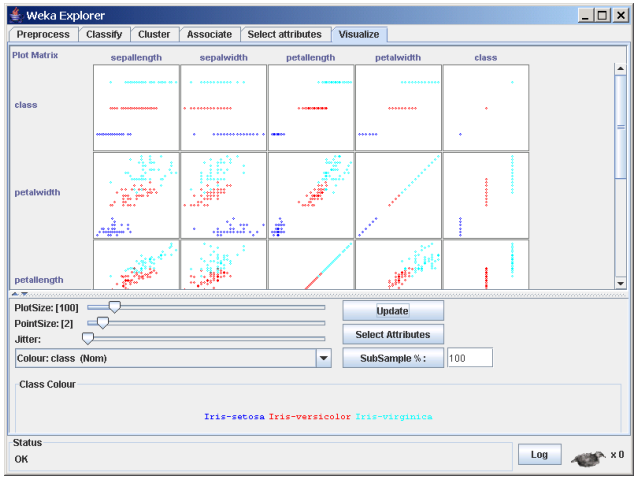
Imagen original: Weka Knowledge Explorer (http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html)

http://www.cs.waikato.ac.nz/~ml/sounds/weka-long.au

Weka (2)







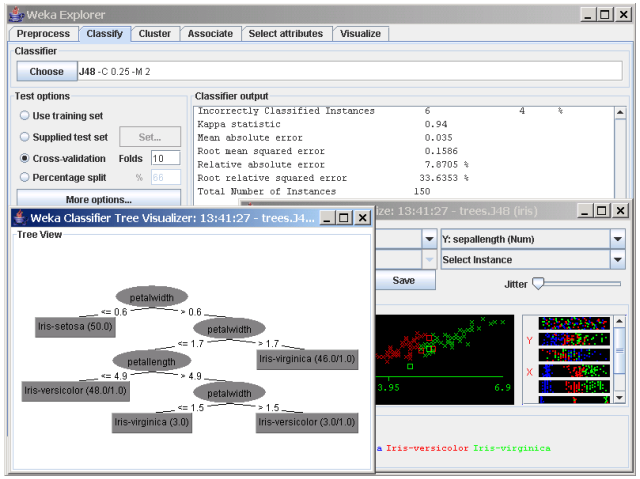
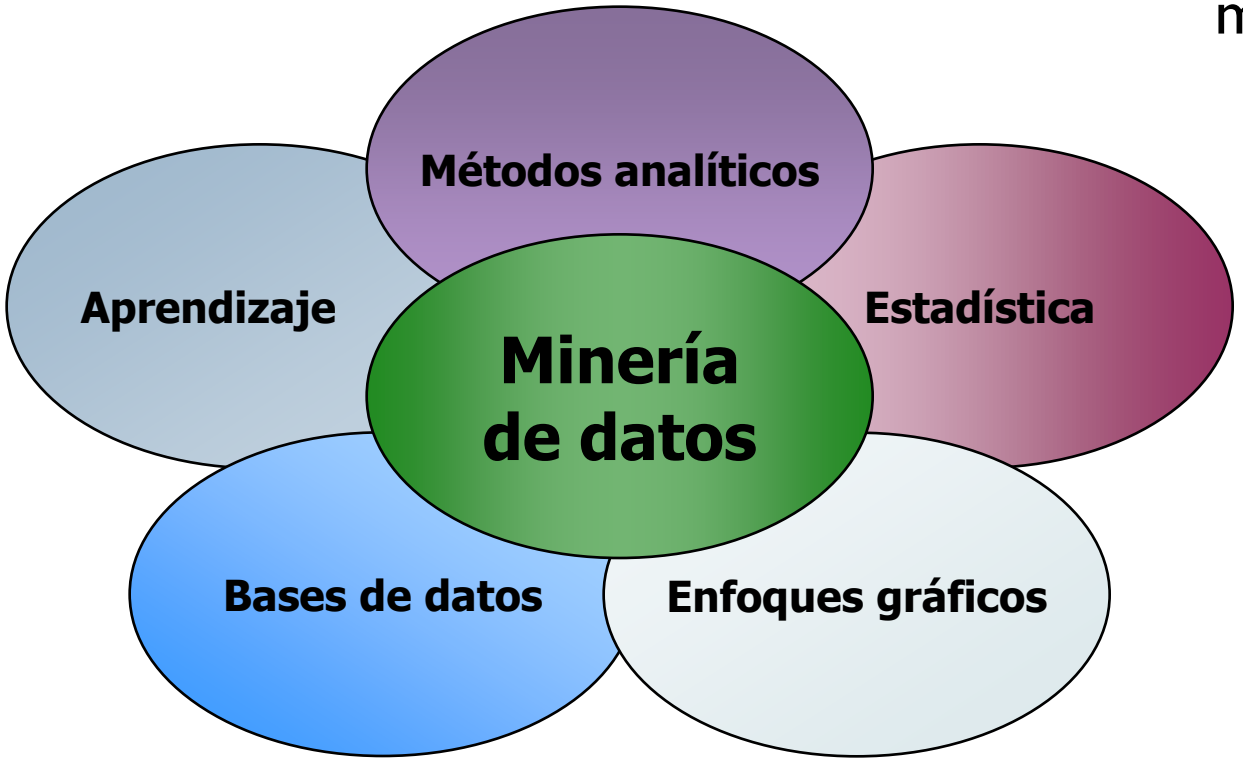


Imagen original: Weka Knowledge Explorer (http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html)

Enfoque multidisciplinar



- Enfoque integrador multidisciplinar
  - Diferentes técnicas según el tipo de información a extraer

Técnicas de minería de datos

- Técnicas descriptivas
  - Orientadas a describir un conjunto de datos
- Técnicas predictivas
  - Orientadas a estimar valores de salida

- Técnicas de modelado
  - Orientadas a la comprensión del sistema: obtener una representación del sistema que permita imitar su comportamiento
  - Emplea cualquier técnica que no funcione como "caja negra"
    - Agrupamiento
    - Árboles de decisión
    - Análisis de secuencias/asociaciones

Técnicas de extracción de conocimiento

Técnicas descriptivas

Segmentación de datos
- Agrupación no supervisada de clientes
- Categorización automática de sucursales

Clasificación
- Asignación de nuevos clientes a segmentos predefinidos
- Identificación de alarmas

Análisis de asociaciones
- Análisis de venta cruzada de productos
- Correlación de hábitos de consumo en base a su ocurrencia

Técnicas predictivas

Análisis de patrones secuenciales
- Detección de secuencias de compra en el tiempo

Análisis de similitud en series temporales
- Identificación de pautas de compra en el tiempo

Predicción
- Asignación de probabilidades de fraude con tarjetas
- Estimación de la demanda y el rendimiento por cliente

Segmentación

Para la agrupación automática de registros que comparten rasgos similares (no supervisados), existen diversas técnicas:

Segmentación o clustering

- El nº de segmentos se determina durante la ejecución del algoritmo
- Procesa tanto variables cuantitativas como cualitativas
- Maximiza la similitud entre los miembros de un mismo segmento y las diferencias entre los miembros de segmentos diferentes, en base a métricas de similitud, no de distancia
- Es eficiente para la detección de nichos de registros

Segmentación neuronal (mapas autoorganizativos de Kohonen)

- Es necesario predefinir el nº de segmentos que se desean obtener y su distribución bidimensional
- Procesa tanto variables cualitativas como cuantitativas, aunque funciona mejor cuando dominan estas últimas
- Es eficiente cuando se desea particionar una población imponiendo cierta relación entre los segmentos obtenidos

Clasificación

Como métodos de clasificación supervisada (predicción de variables cualitativas), algunas técnicas son:

Clasificación basada en árboles de decisión

- Modelo de clasificación en forma de árbol de decisión
- Procesando tanto variables cuantitativas como cualitativas
- Técnicas de podado, que proporciona árboles de menor tamaño
- Son escalables, pudiendo procesar conjuntos con independencia del número de clases, atributos y registros

Clasificación neuronal

- Basada en redes neuronales de propagación hacia atrás
- Detecta de forma automática la topología más adecuada para cada problema, aunque permite especificar una concreta
- Realiza un análisis de sensibilidad para detectar las variables más significativas para cada topología

Predicción

Para la estimación de variables cuantitativas, los métodos más empleados son:

Funciones de base radial

- Pueden procesar variables cuantitativas y cualitativas a la vez
- Detecta el número de centroides óptimo, predefiniendo el número máximo de éstos y el número mínimo de registros asignados a cada centro
- Funciona especialmente bien cuando la estructura de los datos tiende a agruparse en conjuntos, ya que implementa cierto tipo de segmentación

Predicción neuronal

- Basada en redes neuronales de propagación hacia atrás
- Detecta de forma automática la topología más adecuada para cada problema, aunque permite especificar una concreta
- Permite predecir datos en forma de series temporales
- Permite implementar regresión logística

Análisis de asociaciones

Los análisis de asociaciones y patrones secuenciales permiten extraer información desconocida de los hábitos de compra:

Análisis de asociaciones

- Detecta elementos en una transacción que implican la presencia de otros elementos en ésta misma
- Expresa las afinidades entre elementos en forma de reglas de asociación X→Y, facilitando una serie de métricas como el soporte y confianza

Patrones secuenciales

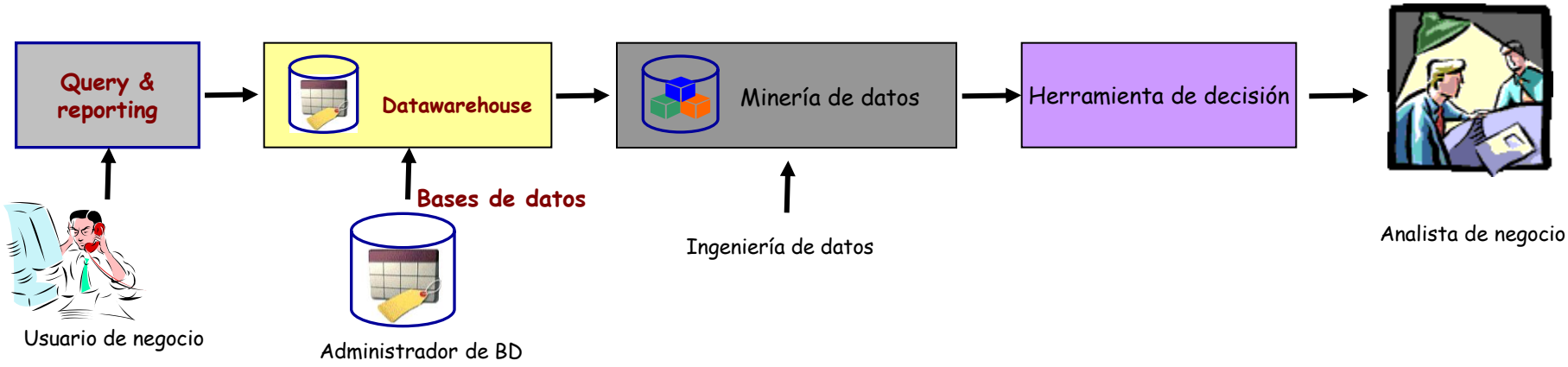
- Detectan patrones entre transacciones, lo que permite optimizar las ventas a lo largo del tiempo

Análisis de similitud en series temporales

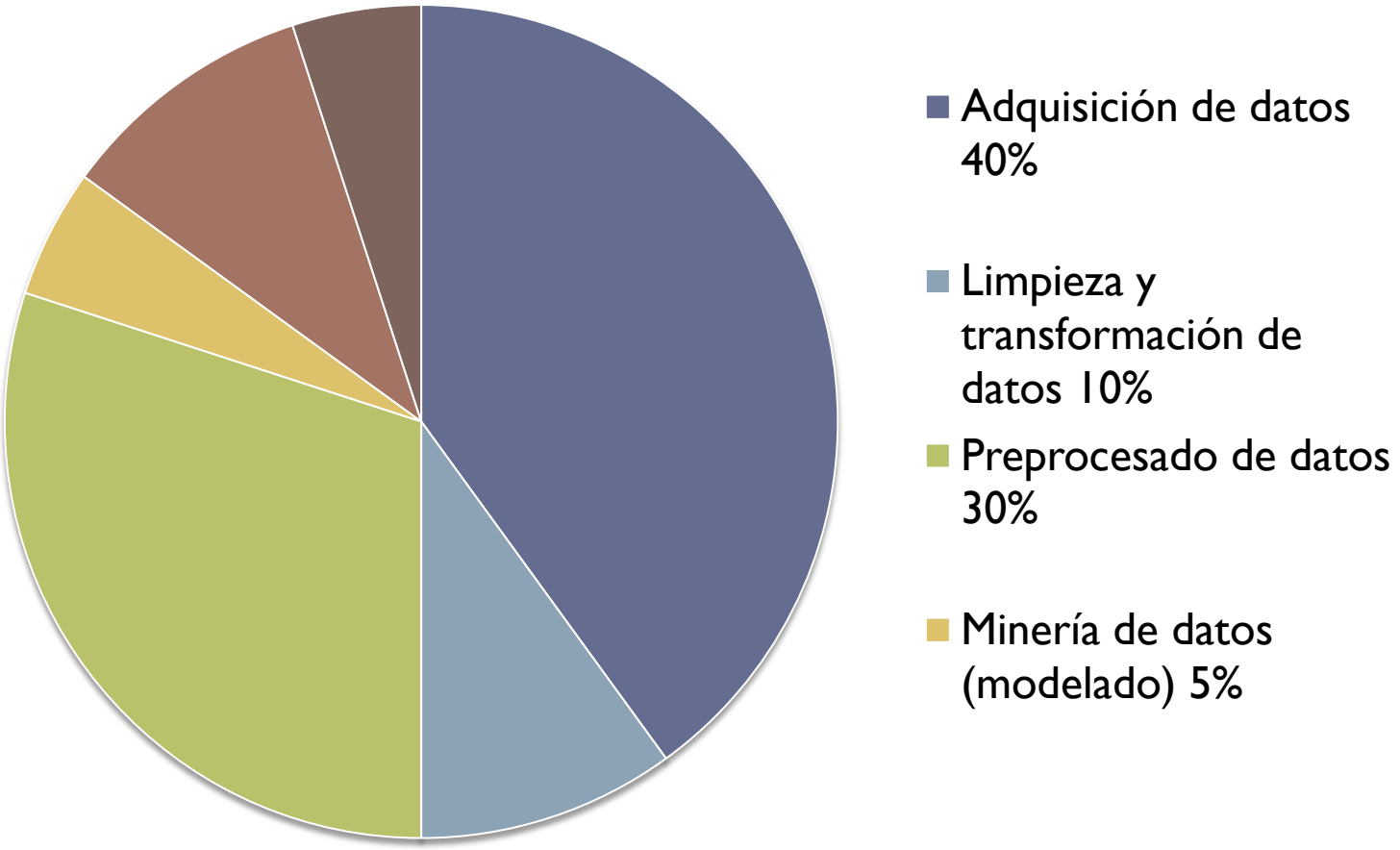
- Detecta todas las ocurrencias de secuencias similares en una colección de series temporales

Equipo de trabajo

- Para lograr un resultado óptimo en un proyecto de minería de datos, el equipo de trabajo debe incluir:
  - expertos en manipulación de datos
  - expertos en inteligencia artificial y en algoritmos de extracción de conocimiento
  - conocedores del dominio de aplicación o con habilidades para comunicarse con los expertos
  - analistas de negocio



Esfuerzo requerido



Submodelos

- En la mayoría de las ocasiones, un único modelo no sirve para representar el sistema completo de manera fiable
- Lo habitual es aplicar la técnica de "divide y vencerás" y construir submodelos que cubren aspectos parciales del sistema
  - Estos submodelos en conjunto resultan más precisos o, al menos, acotan de forma más precisa el error en los aspectos que cubren
  - Para realizar la división en submodelos, se suele aplicar segmentación (clustering) y luego se construye un modelo de predicción para cada uno de los grupos encontrados

Segmentación + Predicción

- El proceso habitual suele ser una primera segmentación de la población en grupos, y luego aplicar a cada uno de ellos un modelo adaptado, por ejemplo de predicción.
- Así el modelo de cada grupo será mejor que si hubiera un único modelo para toda la población.