Reconocimiento de Patrones, Aprendizaje Automático y Minería de Datos: Introducción

Ivan Olmos Pineda

Contenido

- Antecedentes Históricos
- Conceptos Básicos
- Ejemplos
- Relación con otras Disciplinas

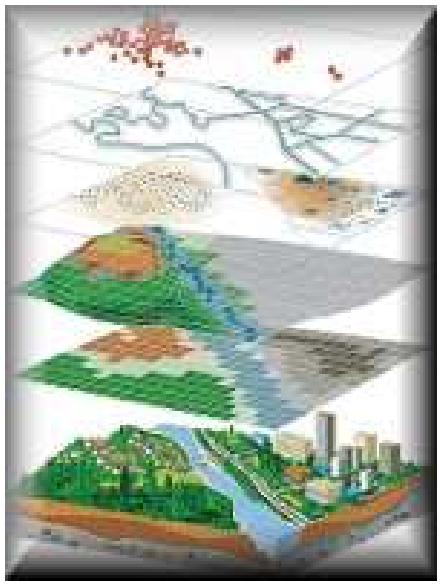
Antecedentes Históricos

Antecedentes Históricos

- Bases de Datos
  - 60’s: bases de datos primitivas
  - 70’s hasta principios de los 80’s:
    - Bases de datos relacionales
    - Bases de datos jerárquicas y de red
    - Herramientas de modelado de datos
    - Indexación de los datos: tablas hash
    - Técnicas de Organización: B-trees

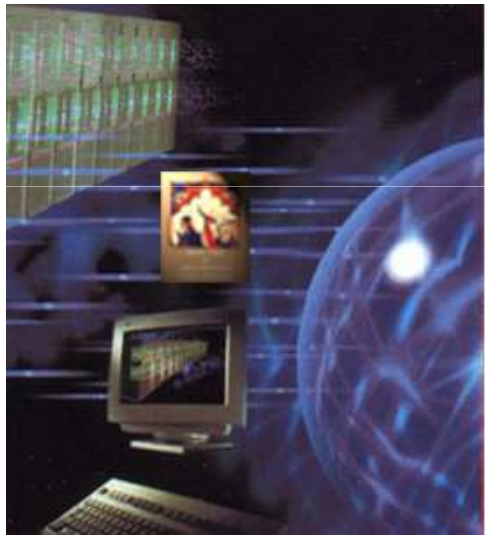
Antecedentes Históricos

- 70’s hasta principios de los 80’s:
  - Lenguajes de manipulación y consulta: SQL
- 80’s a finales de los 90’s:
  - Bases de Datos avanzadas: modelos de datos avanzados (extended-relational, OO)
  - Orientada a aplicaciones: espaciales, temporales, multimedia, científicas



Antecedentes Históricos

- 80’s a finales de los 90’s:
  - Datawhare house y OLAP (On Line Analytical Processing)
  - Minería de Datos y descubrimiento de conocimiento
- 2000 al presente:
  - Sistemas en XML
  - Sistemas de Información Integrados



Conceptos Básicos

Minería de Datos

La Minería de Datos se define como el proceso de descubrir patrones en los datos (Witten, et.al., 2000)

- El proceso debe ser automático o semi-automático
- Los patrones descubiertos deben ser de utilidad, que generen alguna ventaja en su uso

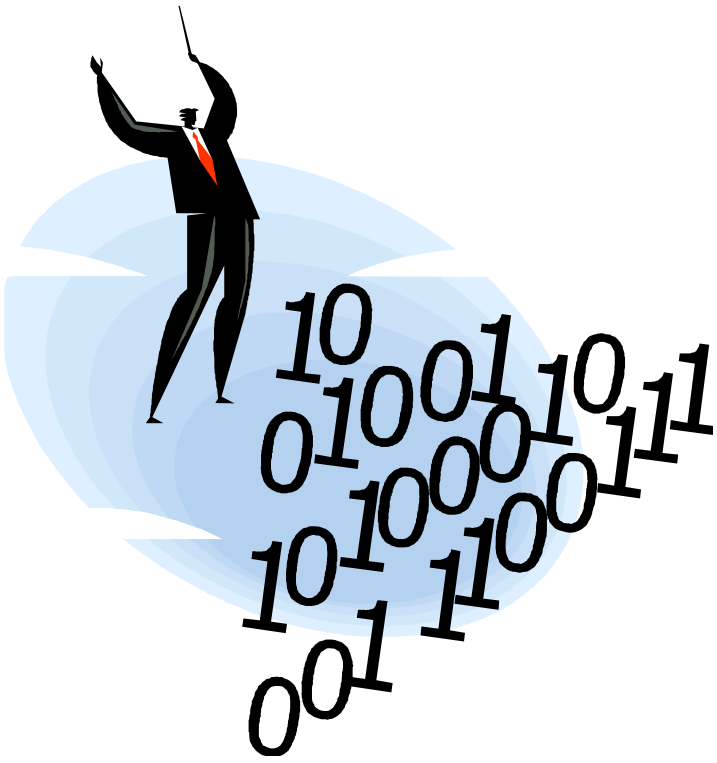
Aprendizaje Automático

Estudiar y modelar computacionalmente los procesos de aprendizaje en sus diversas manifestaciones

- Que aprendan conceptos con el paso del tiempo (cambios adaptivos)

Tareas de la MD – Aprendizaje Automático

- Predecir
  - Con base en valore existentes en la BD, se estiman posibles valores futuros
- Describir
  - Encontrar patrones que describan a los datos de la base



Algoritmos en el Aprendizaje Automático – Reconocimiento de Patrones

- Árboles de decisión y reglas de clasificación
- Métodos de Clasificación de funciones lineales y no lineales
- Agrupamiento o clustering
- Sumarización (describir clases o conceptos)
- Modelos de Dependencias probabilísticas
- Detección de Cambios o Desviaciones
- Asociación
- Análisis de Evolución

Conceptos Relacionados

- Análisis (inteligente) de datos [Berthold & Hand 2003] (fundamentalmente análisis estadísticos)
- Extracción o descubrimiento de conocimiento en BD's (KDD)

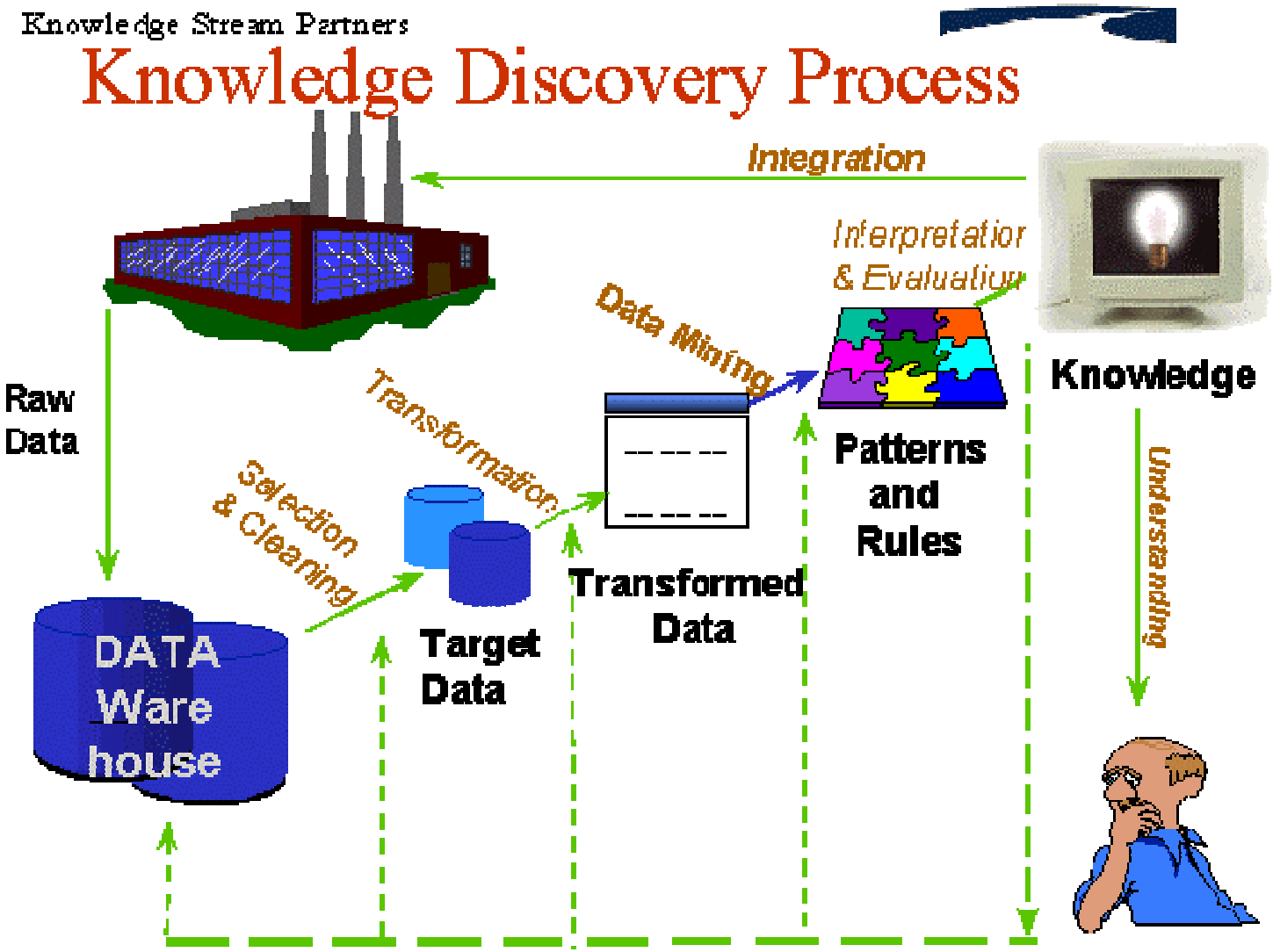
Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles, comprensibles a partir de los datos

KDD

- Válido: patrones precisos para nuevas instancias de la base
- Novedoso: aporte algo desconocido tanto para el sistema como para el usuario
- Potencialmente útil: la información debe conducir a acciones que reporten algún beneficio
- Comprensible: de fácil interpretación para el usuario

Knowledge Stream Partners

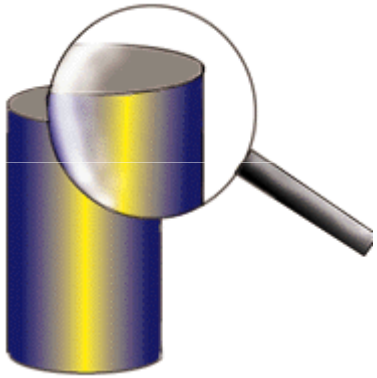
Knowledge Discovery Process



Minería de Datos: Motivación

Almacenaje Masivo de Información

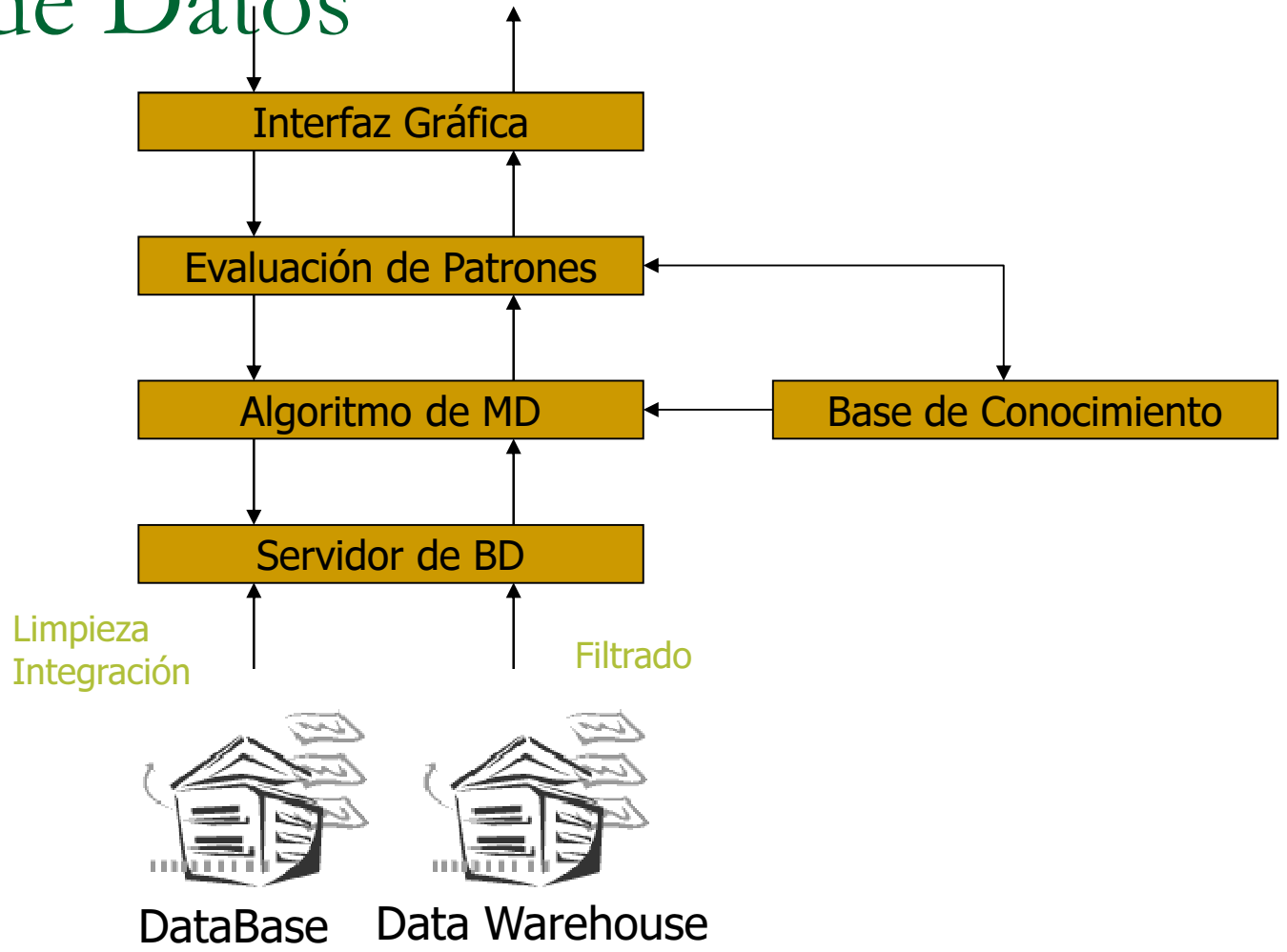
•Automatización de Procesos
•Lector código de barras
•Nuevos instrumentos científicos



Necesidad de nuevas Herramientas para analizar la información

Las herramientas estándares como la estadística no son suficientes

Arquitectura de un Sistema Típico de Minería de Datos



Herramientas de la Minería de Datos

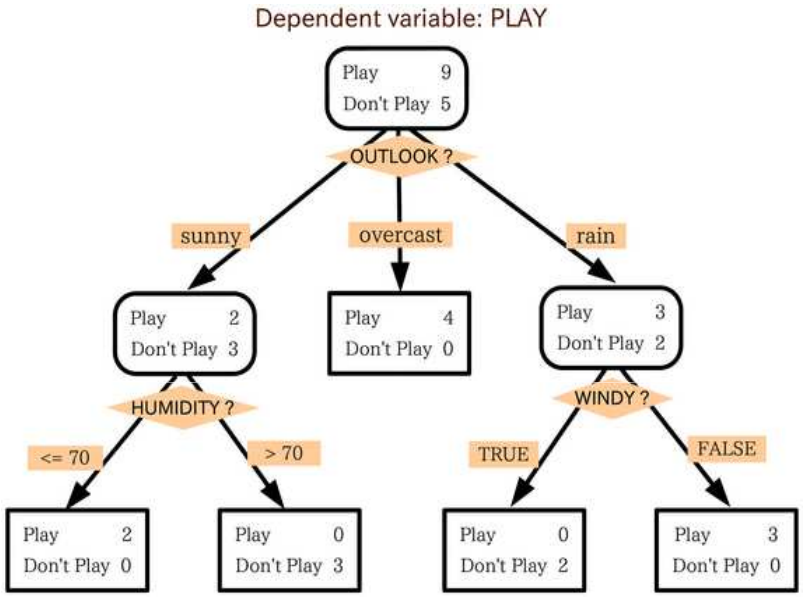
- Tecnología de Base de Datos
- Estadística
- Aprendizaje Automático
- Cómputo de Alto Rendimiento
- Reconocimiento de Patrones
- Redes Neuronales
- Análisis de Datos Espaciales

Componentes de un Algoritmo de Aprendizaje Automático

- 3 elementos:
  - Modelo de Representación
  - Modelo de Evaluación
  - Modelo de Búsqueda

Modelo de Representación

- Lenguaje para describir los patrones
  - Árbol de decisión
  - Lógica de primer grado
  - Gráfico



Si cuentas-morosas > 0 entonces Devuelve-crédito = no
Si cuentas-morosas = 0 y (salario > 2500) ó (D-crédito > 10) entonces
Devuelve-Crédito = si

Modelo de Evaluación

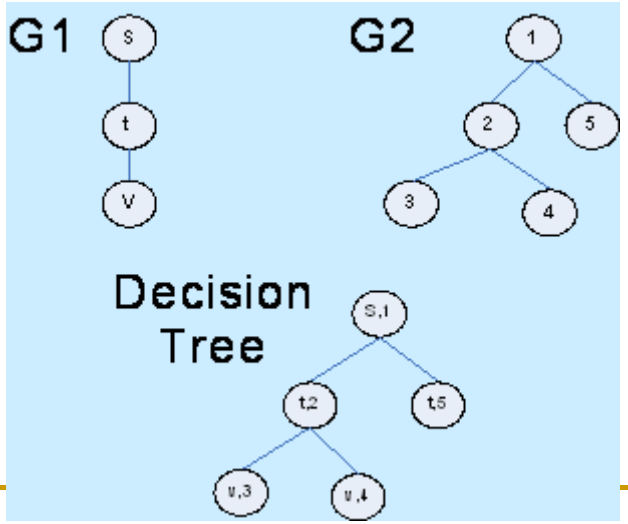
- Características del patrón encontrado:
  - Útil
  - Novedoso
  - Entendible
  - Efectivo para tareas de predicción
  - Medidas
    - Soporte
    - Confianza
  - Dominios médicos
    - Sensibilidad
    - Especificidad

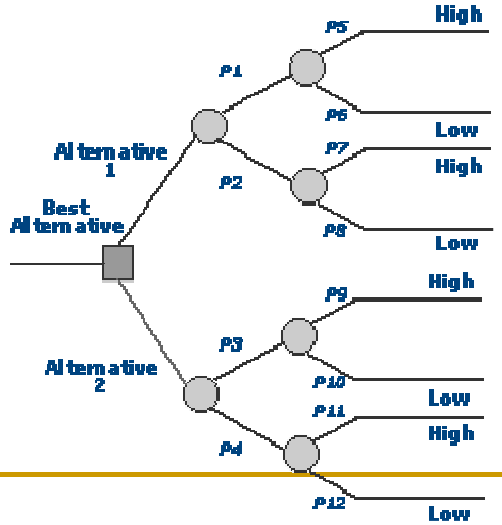
Métodos de Búsqueda (1)

- Búsqueda de parámetros (el num. De parámetros depende directamente del algoritmo seleccionado)
  - Parámetros en algoritmos de árboles de clasificación
  - Parámetros de espacio en “beam search”

Métodos de Búsqueda (2)

- Búsqueda del modelo
  - Itera sobre la búsqueda de parámetros y elige el mejor resultado





Algoritmos de Minería de Datos / Aprendizaje Automático (1)

- Reglas de Asociación
  - Fast Association Rules
- Métodos Bayesianos
  - Redes Bayesianas
- Árboles de Decisión
  - Induction of Decision Trees
- Métodos Relacionales y Estructurales
  - Programación Lógica
- Redes Neuronales Artificiales
  - Backpropagation

Algoritmos de Minería de Datos / Aprendizaje Automático (2)

- Máquinas de Vector de Soporte
  - SVM
- Algoritmos Evolutivos
  - Teoría Evolutiva, Lógica Difusa
- Métodos Basados en Casos y Vecindad
  - Vecinos más cercanos

Usuarios (1)

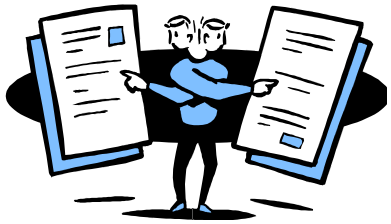
- ¿A que tipo de usuarios esta dirigida la Minería de Datos / Aprendizaje Automático?
  - Negocios: Para construir modelos a partir de grandes bases de datos

Usuarios (2)

- ¿A que tipo de usuarios esta dirigida la Minería de Datos / Aprendizaje Automático?
  - Consumidores: para filtrar información de grades bases de datos

Usuarios (3)

- ¿A que tipo de usuarios esta dirigida la Minería de Datos / Aprendizaje Automático?
  - Investigadores: análisis de grandes bases de datos

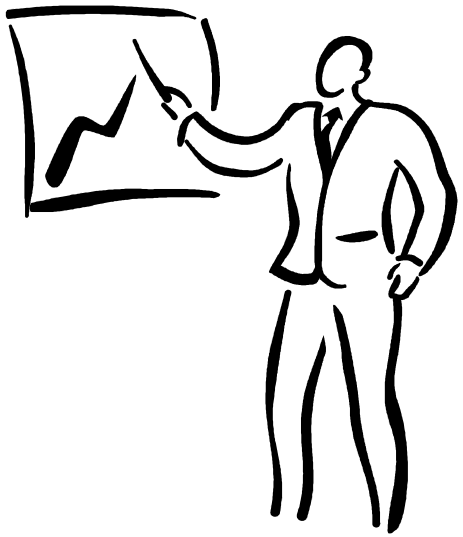


Aplicaciones (1)

- Astronomía
  - Clasificación de estrellas y galaxias

Aplicaciones (2)

- Análisis de Mercado y Administración
  - Perfil del cliente
    - ¿qué tipos de clientes compras que tipo de productos? (clustering)
    - ¿Qué productos se compran normalmente juntos? (reglas de asociación)
    - Descubrir las relaciones entre características personales y el tipo de productos que se compran



Aplicaciones (3)

- Finanzas
  - Inversiones a partir del análisis de minería de datos
  - Análisis de clientes para otorgar crédito

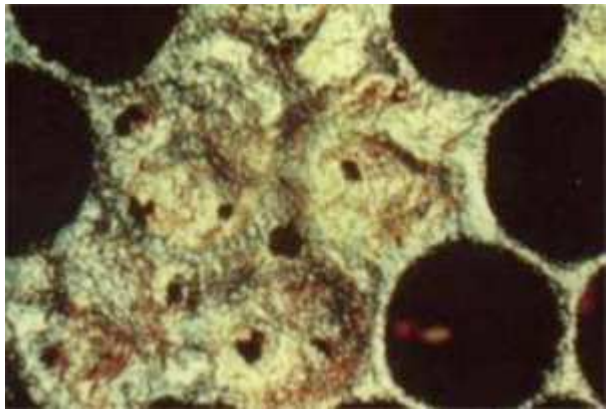
- Fraudes
  - A partir de datos históricos detectar comportamientos anómalos (construcción de modelos), para la detección de nuevos fraudes
    - Seguros de autos
    - Seguros médicos
    - Sector bancario

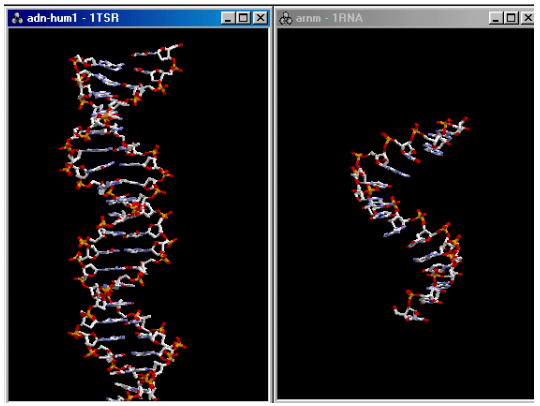
Aplicaciones (4)

- Deportes (interpretación de estadísticas)
- Web
  - Análisis de registros log
  - Análisis de comportamiento en un sitio
- e-mail
  - Clasificación (spam)
- Recursos Humanos
  - Selección de personal

Aplicaciones (5)

- Medicina y Farmacéutica
  - Detección de Enfermedades
  - Desarrollo de nuevo medicamento
  - Análisis de ADN
  - Secuenciación





Ejemplos de Aplicaciones del Aprendizaje Automático

Ejemplo (1)
El Juego del Tenis

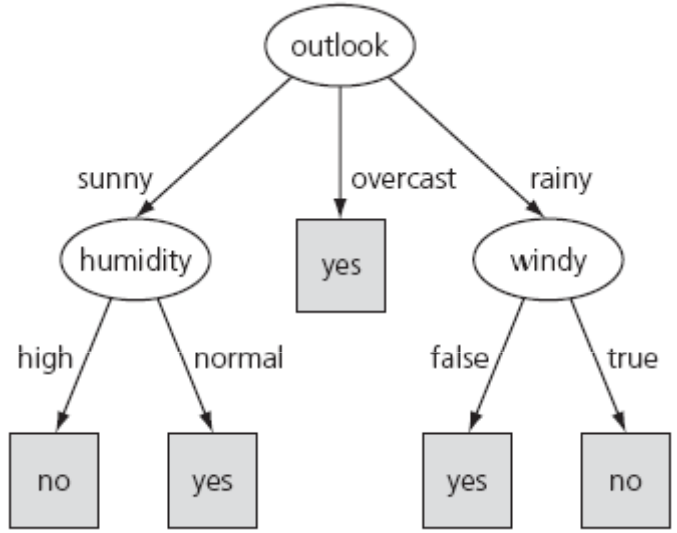
- ¿será un buen día para jugar tenis?

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Ejemplo (1)
El Juego del Tenis

- Datos
  - 4 atributos:
    - Outlook: sunny, overcast, rainy
    - Temperature: hot, mild, cool
    - Humidity: high, normal
    - Windy: true, false
  - ¿Tamaño del espacio de búsqueda?

Ejemplo (1)
El Juego del Tenis




```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

Ejemplo (2)
Lentes de Contacto

- ¿Puedo usar lentes de contacto?

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Ejemplo (2)
Lentes de Contacto

- ¿condiciones bajo las cuales un oftalmólogo puede prescribir lentes de contacto suaves, duros?

```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no and
   tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no and
   tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope and
   astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no and
   tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes and
   tear production rate = normal then recommendation = hard
If age = young and astigmatic = yes and
   tear production rate = normal then recommendation = hard
If age = pre-presbyopic and
   spectacle prescription = hypermetrope and astigmatic = yes
   then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
```

Ejemplo (3)
Identificación de Leucemia

- A partir de imágenes digitales de leucemia, ¿cómo clasificar nuevas muestras que determinen la presencia de la enfermedad?

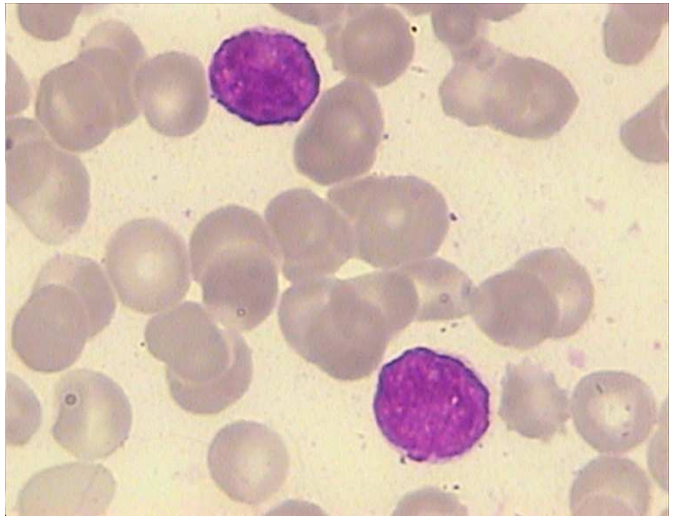


Fig.1 Ejemplo de una imagen del tipo LLA

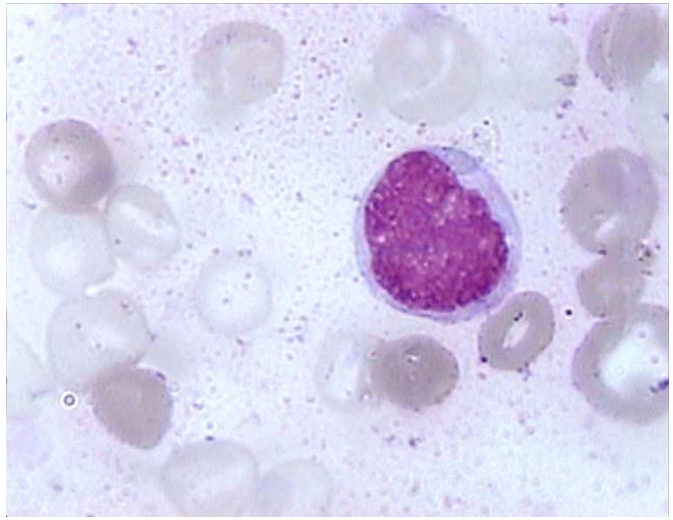
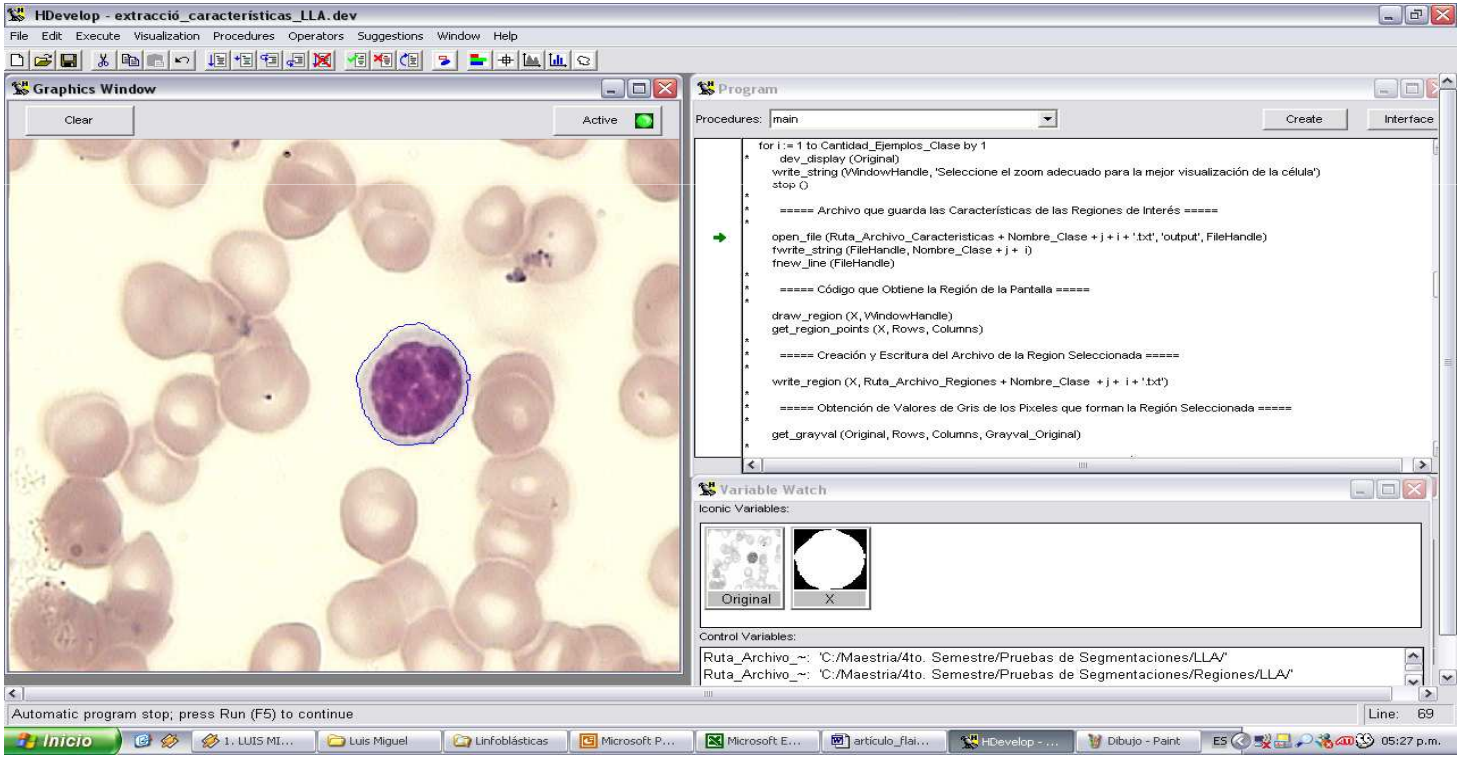


Fig.2 Ejemplo de una imagen del tipo LMA

Ejemplo (3)
Identificación de Leucemia

- ¿Cómo trabajar las imágenes?
  - → Transformación de las imágenes



Ejemplo (3)
Identificación de Leucemia

Reglas Obtenidas con la BD Desbalanceada

ADDTree
If grayvalue < 16419.5 **then** LLA
If grayvalue >= 16419.5 **and** contraste < 5.244 **and** Mn_VG < 76.5 **and** Homogeneidad < 0.551 **then** LLA
If grayvalue >= 16419.5 **and** contraste < 5.244 **and** Mn_VG < 76.5 **and** Homogeneidad >= 0.551 **then** LMA
If grayvalue >= 16419.5 **and** contraste < 5.244 **and** Mn_VG >= 76.5 **and** orientacion >= 0.436 **and** area < 17728.5 **then** LLA
If grayvalue >= 16419.5 **and** contraste < 5.244 **and** Mn_VG >= 76.5 **and** orientacion >= 0.436 **and** area >= 17728.5 **then** LMA
If grayvalue >= 16419.5 **and** contraste >= 5.244 **and** asintropia < -0.496 **and** Homogeneidad < 0.568 **and** Factor_Forma_Exc < 0.121 **then** LMA
If grayvalue >= 16419.5 **and** contraste >= 5.244 **and** asintropia < -0.496 **and** Homogeneidad >= 0.568 **and** Factor_Forma_Exc >= 0.121 **then** LLA
If grayvalue >= 16419.5 **and** contraste >= 5.244 **and** asintropia >= -0.496 **and** convexidad < 0.948 **then** LMA
If grayvalue >= 16419.5 **and** contraste >= 5.244 **and** asintropia >= -0.496 **and** convexidad >= 0.948 **then** LLA

Conocimiento es poder!!!

Interfaz Gráfica

- Interacción con el usuario
- Selección de algoritmo de MD
- Ayuda a establecer parámetros
- Permite la exploración de los patrones encontrados
- Visualización de los resultados en diferentes formatos

Módulo de Evaluación de Patrones

- Medidas de que tan Interesante es un patrón
  - Tendencias
  - Patrones
  - Desviaciones

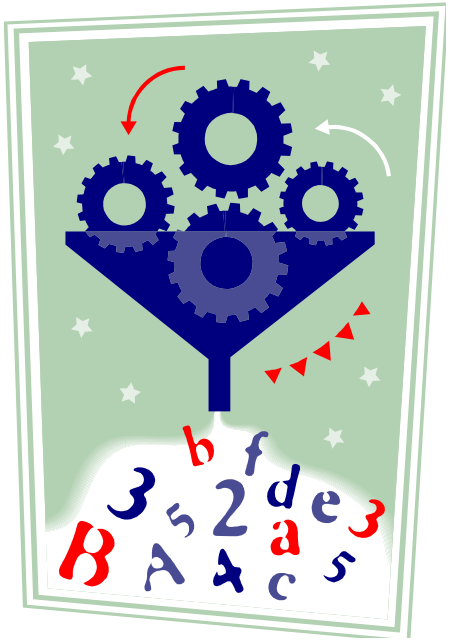
- Interactúa con el algoritmo de MD - guía el proceso de búsqueda

Arquitectura de un Sistema de Minería de Datos

- Algoritmo de Minería de Datos (esquema modular)
  - Caracterización
  - Asociación
  - Clasificación
  - Análisis de Grupos
  - Análisis de desviaciones

Servidor de Base de Datos

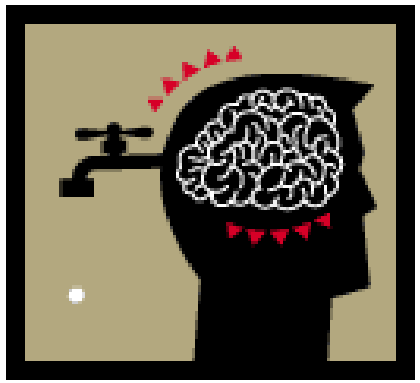
- Usado para extraer información relevante



Lenguajes de Consulta
- SQL

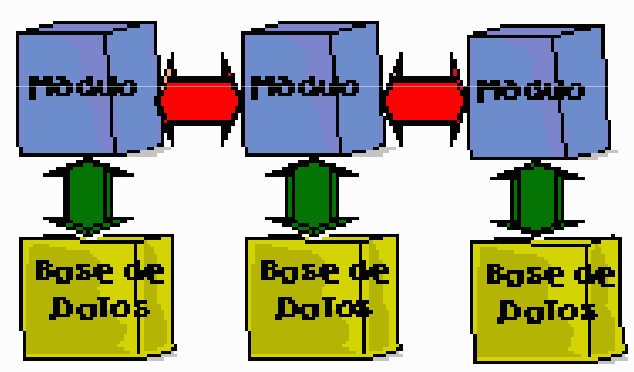
Base de Conocimiento

- Conocimiento del dominio para guiar la búsqueda (expertos)
- Creencias sobre los datos
- Umbrales de evaluación



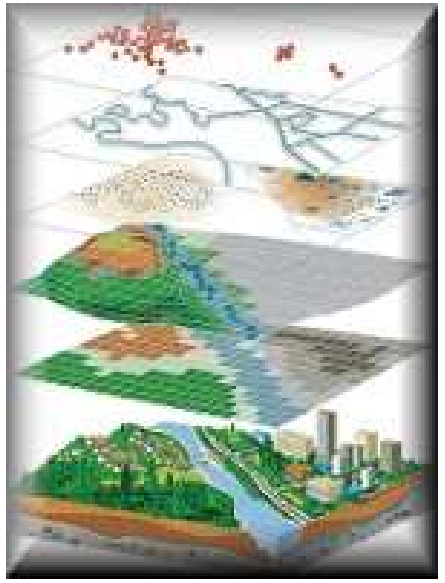
Base de Datos

- Base de Datos Relacional
  - Modelo Entidad - Relación
  - DBMS
  - Lenguajes (DDL, DML, …)
  - Tablas (tuplas, campos)
  - Normalización



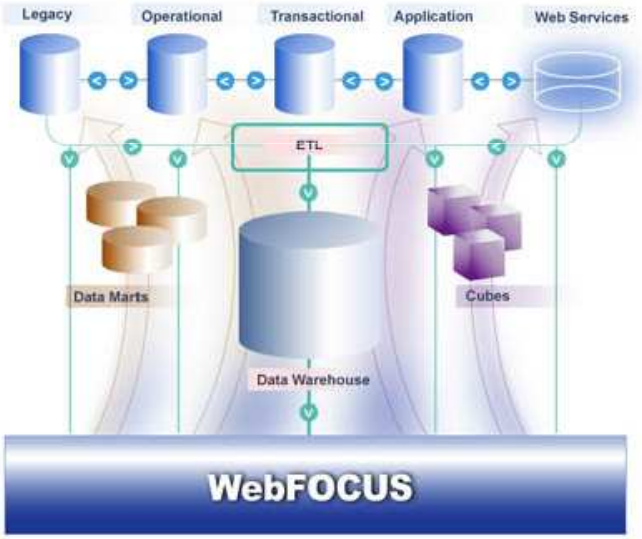
Base de Datos

- Bases de Datos Transaccionales
- Base de Datos Orientada a Objetos
- Bases de Datos Espaciales
- Bases de Datos Temporales
- Bases de Datos Multimedia
- Base de Datos de texto
- Base de Datos Heterogéneas
- WWW



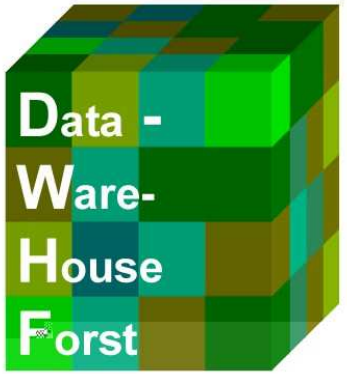
Data Warehouse

- Repositorio de información recopilado de varias fuentes bajo un esquema unificado
  - Usualmente residen en un solo sitio
- Construcción
  - Limpieza de datos
  - Transformación de datos
  - Integración de datos
  - Carga de los datos
  - Actualización periódica



Data Warehouse

- Datos organizados por temas de alto nivel (cliente, proveedor, transacción…)
- Datos desde una perspectiva histórica (resumen de varios años)
- Modelos sobre una estructura multidimensional
  - Cubos de datos



Data Warehouse

- Análisis de Datos
  - OLAP: utiliza información previa del dominio para presentar los datos a diferentes niveles de abstracción
  - Se requiere un análisis de datos profundo