Introducción a la Minería de Datos y al Aprendizaje Automático

Carlos Alonso González
Grupo de Sistemas Inteligentes
Departamento de Informática
Universidad de Valladolid

Juan José Rodriguez Diez
Grupo de Sistemas Inteligentes
Departamento de Ingeniería Civil
Universidad de Burgos

Contenidos

1. **Interés**
2. **Definición de aprendizaje**
3. **Tareas Básicas de aprendizaje**
4. **Dimensiones de Análisis**
5. **Paradigmas de aprendizaje**
6. **Minería de datos**
    1. Motivación
    2. ¿Qué es la minería de datos?
    3. Etapas
    4. Ejemplos
    5. Ética y Minería de da datos

1 Interés

- No hay inteligencia sin aprendizaje (adaptación, mejora, descubrimiento…)
- En la práctica
  - Exceso de información
  - Escasez de conocimiento
  - Necesidad de automatizar la obtención de conocimiento a partir de información

Nichos de aplicación

- Minería de datos: uso de datos históricos para mejorar la toma de decisiones
  - Registros médicos → Conocimiento médico
  - Imágenes del firmamento -> catálogo de objetos estelares
- Aplicaciones software que no se pueden programar con técnicas convencionales
  - Reconocimiento del habla
  - Vehículos autónomos
- Software personalizado
  - Filtro de noticias de interés
  - Gestión de Agenda

2 Una definición de aprendizaje

- Un programa de ordenador APRENDE de la experiencia E con respecto a una clase de tareas T y medida de desempeño P si su rendimiento en tareas de T, según la medida P, mejora con la experiencia E (Mitchell, 97)

Ejemplos

- Aprender a Jugar a las Damas
  - T: jugar a las damas
  - P: porcentaje de juegos ganados al adversario
  - E: juegos de entrenamiento consigo mismo
- Aprender a reconocer la escritura manual
  - T: reconocer y clasificar palabras manuscritas en una imagen
  - P: porcentaje de palabras reconocidas correctamente
  - E: base de datos de imágenes de palabras manuscritas, clasificadas
- Aprender a conducir
  - T: conducir en una autopista pública de 4 carriles utilizando sensores de visión
  - P: distancia media viajada antes de un error (según instructor humano)
  - E: secuencia de imágenes y comandos de guiado registrados a partir de la observación de un conductor humano

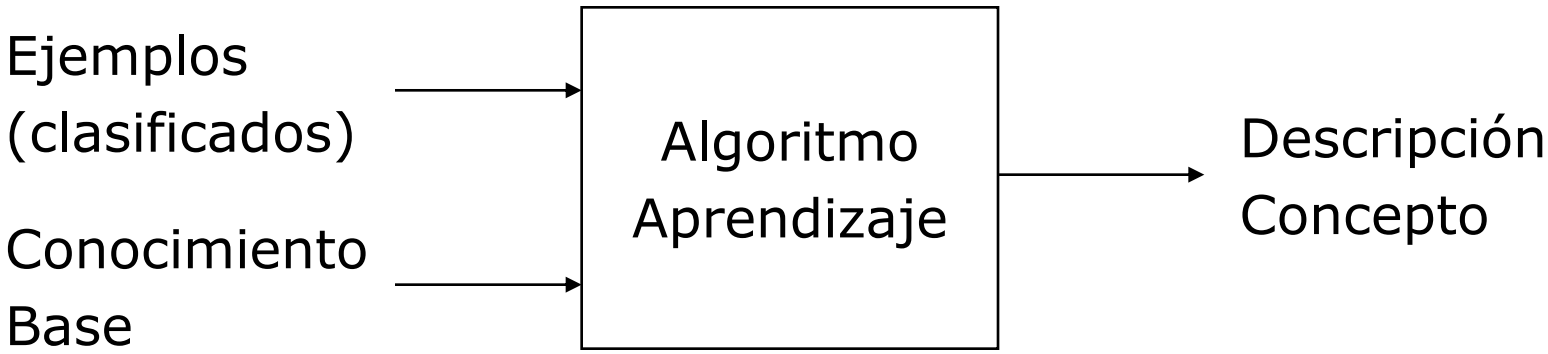
3 Tareas básicas en el aprendizaje automático

- Descripción de conceptos
- Formación de conceptos
- Mejora de la eficiencia

- Análisis de regularidades en datos

Descripción de conceptos

- Planteamiento general

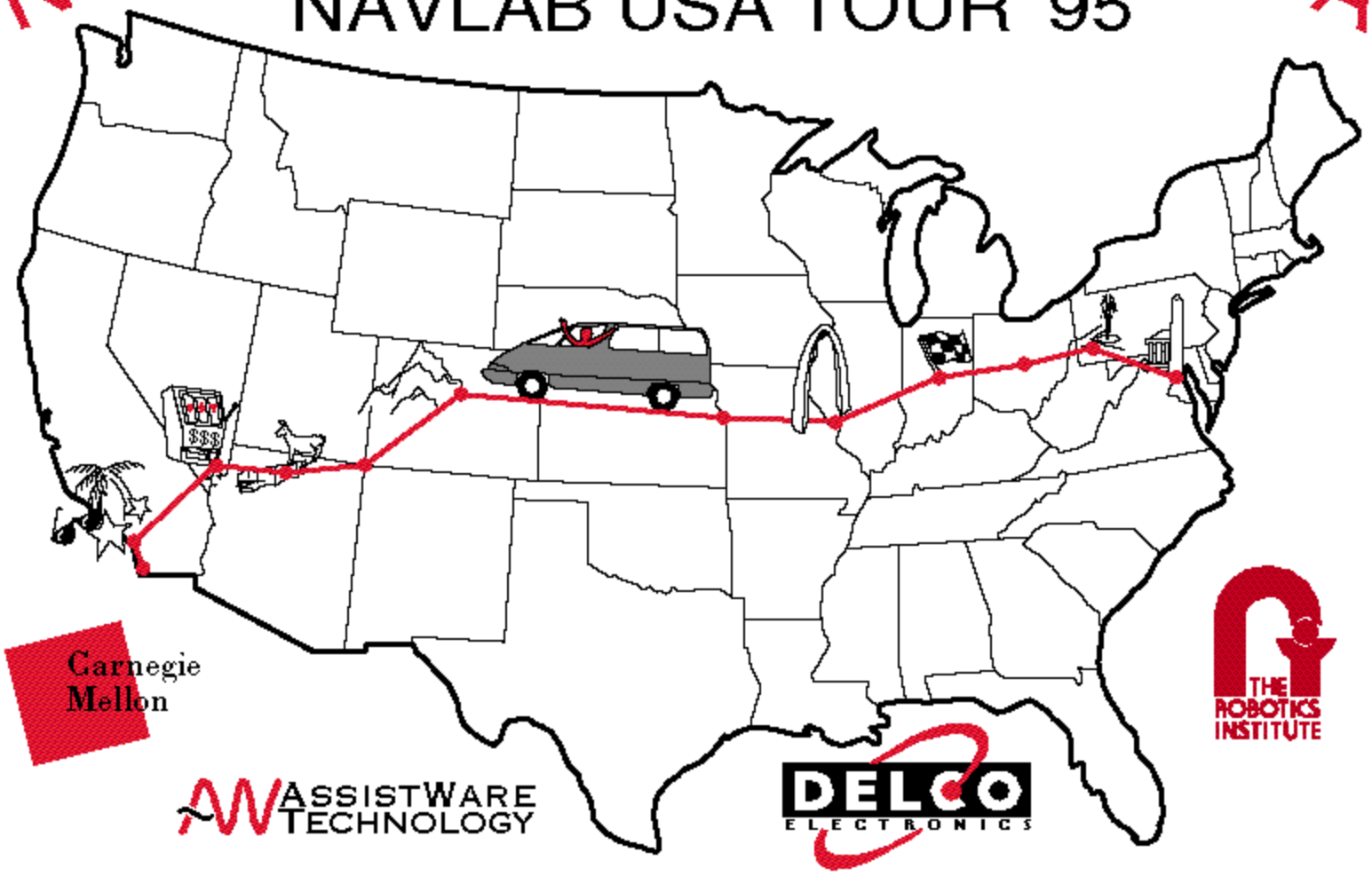


Descripción de conceptos

- Dado
  - Concepto objetivo
  - Instancias del mismo
  - Conocimiento base
- Obtener
  - Caracterización del concepto
    - Típicamente clasificador a partir de atributos (identificar/predecir el valor de la clase)
    - También regresión (predecir valor atributo numérico)
- Ejemplos
  - Análisis de riesgos en asignación de créditos
  - Diagnosis
  - Vehículos autónomos

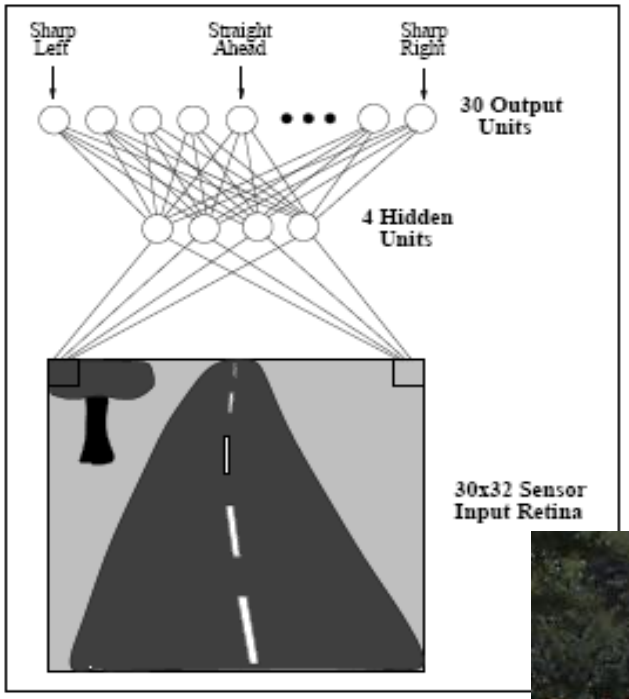
NO HANDS ACROSS AMERICA

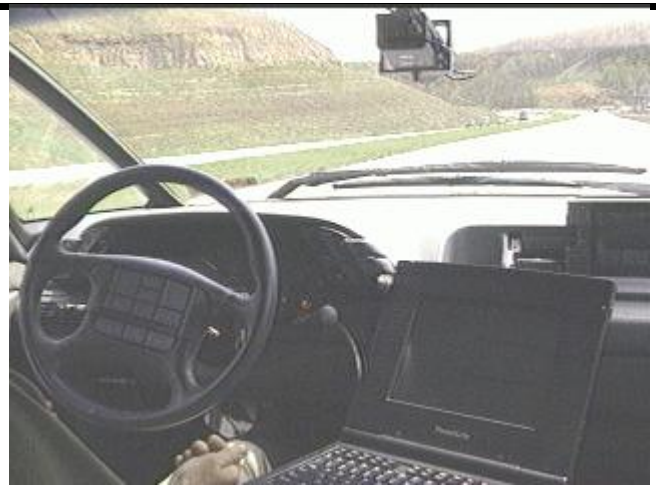
NAVLAB USA TOUR '95



- Washington DC
- Pittsburgh PA
- Columbus OH
- Indianapolis IN
- Kokomo IN
- Saint Louis MO
- Kansas City KA
- Denver CO
- Four Corners
- Grand Canyon
- Las Vegas NV
- Los Angeles CA

ALVINN, RALPH

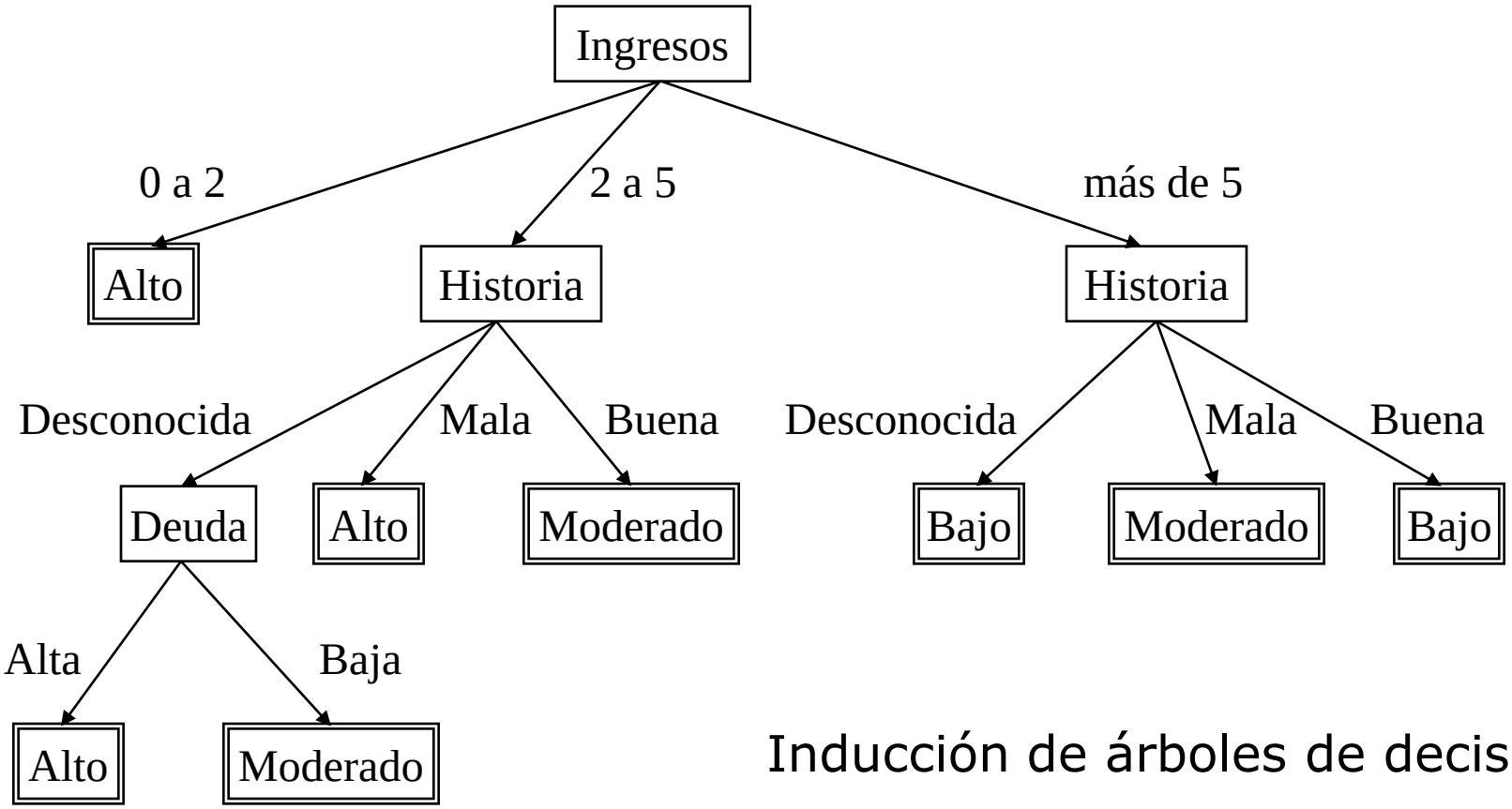




Análisis riesgos concesión de créditos

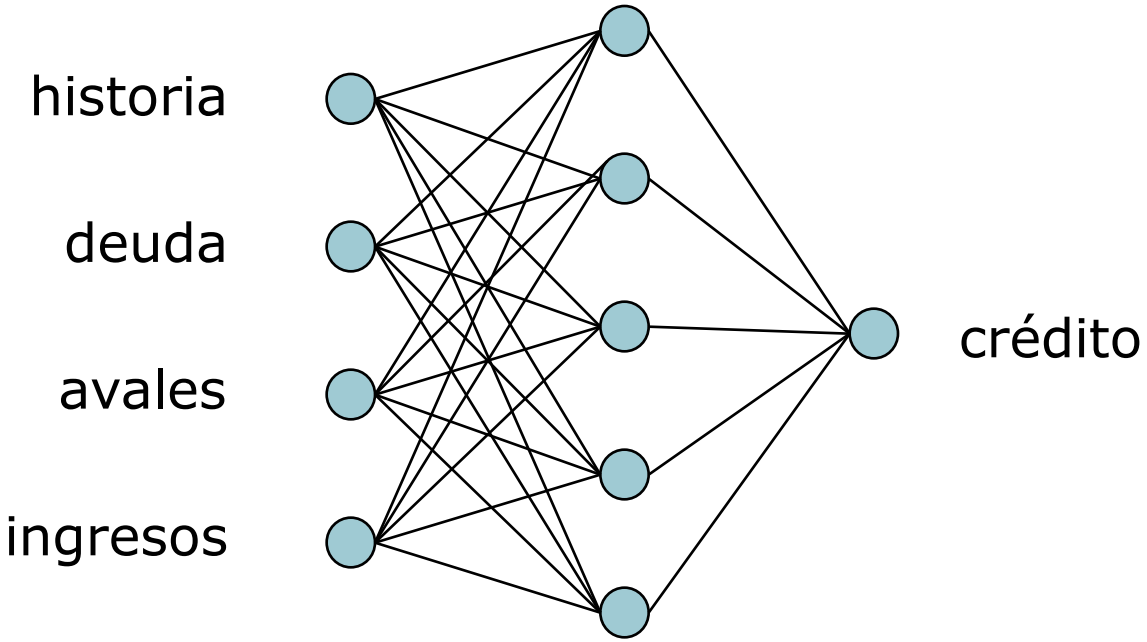
Nº	Riesgo	Historia	Deuda	Avales	Ingresos
1	alto	mala	alta	no	0 a 2M
2	alto	desconocida	alta	no	2 a 5M
3	moderado	desconocida	baja	no	2 a 5M
4	alto	desconocida	baja	no	0 a 2M
5	bajo	desconocida	baja	no	más de 5M
6	bajo	desconocida	baja	adecuados	más de 5M
7	alto	mala	baja	no	0 a 2M
8	moderado	mala	baja	adecuados	más de 5M
9	bajo	buena	baja	no	más de 5M
10	bajo	buena	alta	adecuados	más de 5M
11	alto	buena	alta	no	0 a 2M
12	moderado	buena	alta	no	2 a 5M
13	bajo	buena	alta	no	más de 5M
14	alto	mala	alta	no	2 a 5M

Análisis riesgos concesión de créditos



Inducción de árboles de decisión

Análisis riesgos concesión de créditos



Redes de neuronas

Concepto: Política accesos (Ejemplos, conocimiento base)

- Ejemplos
  - puede_operar(smith, pabxb_17),
  - puede_operar(miller, lod_2)...
- Conocimiento base
  - manager(smith),
  - trabaja_para(smith, betecom),
  - alquila(betecom, pabxb_17)...

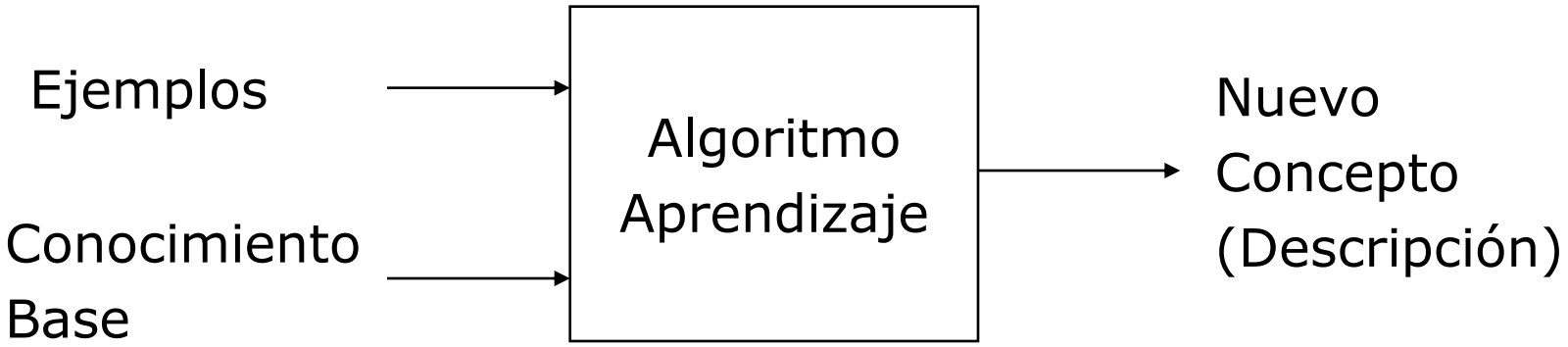
Concepto: Política accesos
(Concepto)

trabaja_para(Persona, Compañía),
alquila(Compañía, Sistema,)
manager(Persona)
→
puede_operar(Persona, Sistema)

Programación lógica inductiva

Formación de conceptos

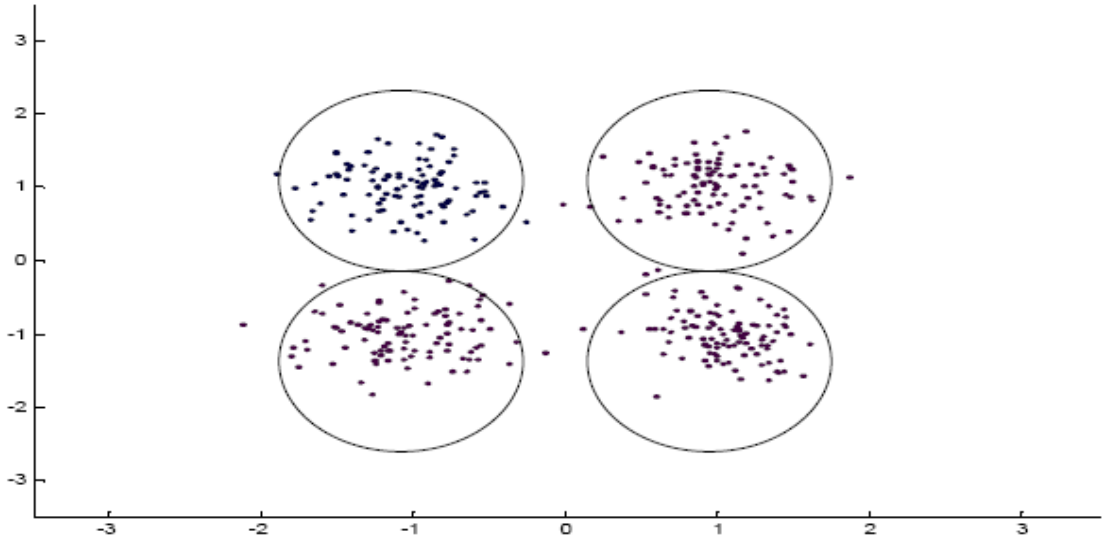
- Planteamiento general



Formación de conceptos

- Dado
  - Instancias de un concepto, posiblemente desconocido
- Obtener
  - Caracterización del concepto
- Tres aproximaciones principales
  - Agrupamiento (clustering)
  - Asociación
  - Descubrimiento

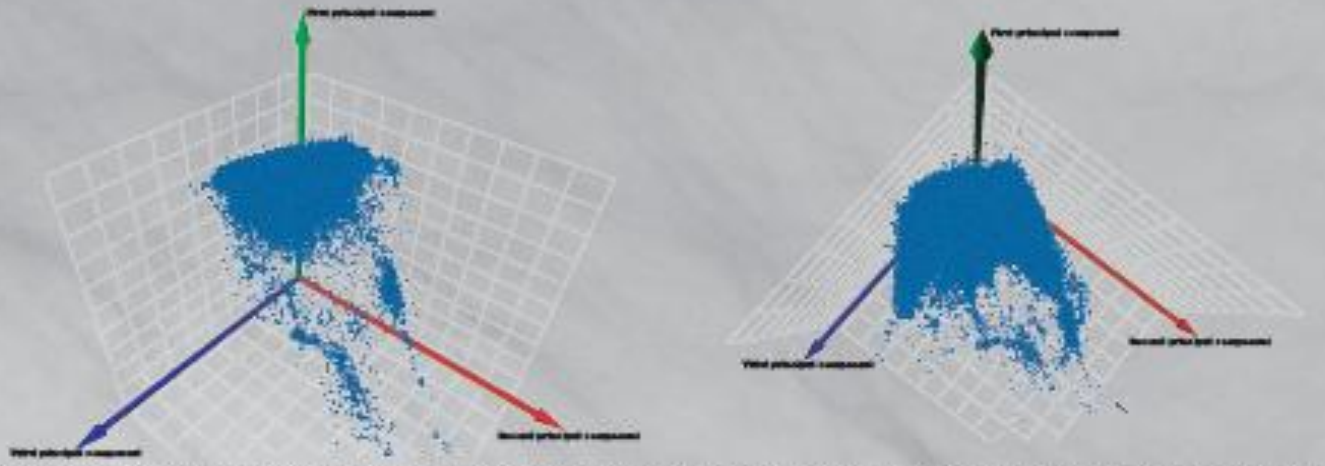
Agrupamiento



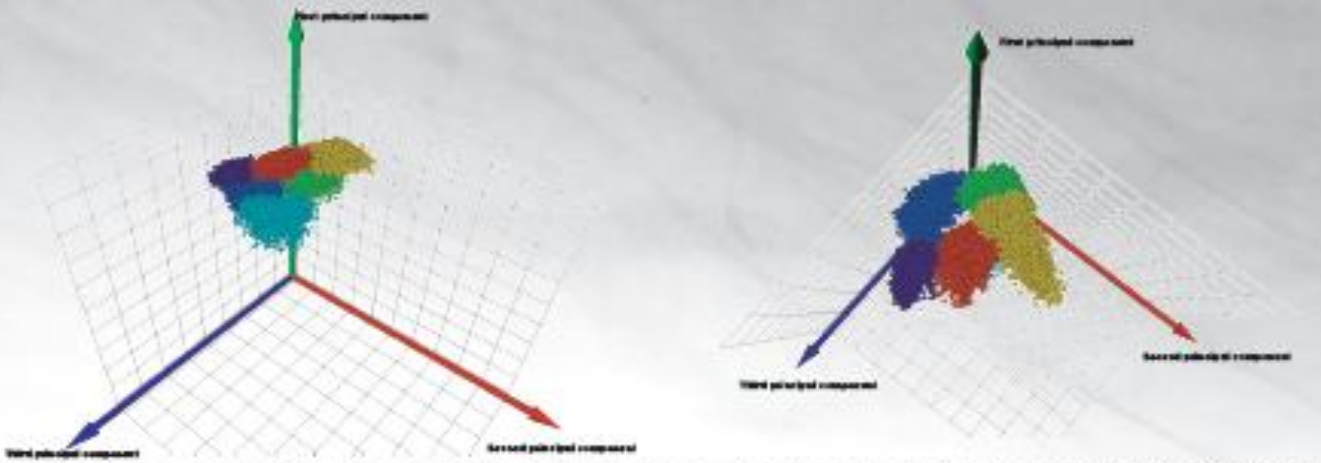
Results of Clustering

The selection of features was carried out using about 93,000 suspicious regions extracted from 200 mammograms. For each of the regions a candidate set of 46 statistical features was computed. Using k-means algorithm we discarded 19 least descriptive features. Next, we used our clustering algorithm to partition the resultant vectors, and compute the score values of the remaining features. Using only the 18 features with the highest score values we achieved classification accuracy close to the one given by all 46 features.

The analysis of the original, 46 dimensional data set, as well as clustering results was made significantly easier by using interactive visualization tools. Furthermore, by using dimensionality reduction methods, visualization of multidimensional datasets was possible.



Original set of feature vectors from mammogram regions. Two different views are presented. The projection into 3D space was done using the PCA technique.



Result of clustering mammogram regions using our method. Two different views are presented. The projection into 3D space was done using the PCA technique.

Reglas de asociación

- Ejemplo: análisis de la cesta de la compra

Si edad < 40 **Y**

contiene pañales

Entonces contiene cerveza

4 Dimensiones de Análisis

- Ejemplos
  - Nº ejemplos: múltiples / único (pocos)
  - Clasificación: supervisado / no supervisado
  - Procesamiento: no incremental (lotes) / incremental
- Conocimiento básico
  - No utilizan /utilizan
- Representación conocimiento
  - Simbólico / subsimbólico
- Sesgos (bias) inductivos: factores adicionales que determinan que conceptos se pueden aprender
  - Representacionales
  - Restrictivos
  - De preferencia

Papel del Bias

- Propiedad de la inferencia inductiva

  "Un sistema de aprendizaje que no haga suposiciones *a priori* sobre la identidad del concepto objetivo no tiene ninguna base racional para clasificar instancias no vistas" (Mitchell, 97)

5 Paradigmas principales

- Aprendizaje memorístico
- Aprendizaje inductivo
- Aprendizaje deductivo
- Aprendizaje multiestrategia
- Aprendizaje por analogía
- Aprendizaje por refuerzo

- En este curso: aprendizaje inductivo

6 Minería de datos
6.1 Motivación I

- Crecimiento explosivo de los datos.
  - De terabytes a petabyes.
- Cada día se crean 52.000.000.000 MB de datos (1997).
- La cantidad de datos almacenados se duplica cada 10 meses.
- Cada persona está en 800–1000 bases de datos.
- Sólo el 4% de los datos se usa para algo (IBM).

Aprendizaje memorístico

- Discutible: no tiene capacidad de generalización
- Primer paradigma utilizado con éxito: Samuel (Damas, 50)

Aprendizaje inductivo

- También denominado basado en ejemplos
  - Se caracteriza por utilizar (numerosos) ejemplos de un concepto
- Objetivo: caracterizar un (nuevo) concepto
- Numerosas aproximaciones de interés

- GENERALIZACION a PARTIR DE EJEMPLOS

Aprendizaje deductivo

- Objetivo: aumentar eficiencia, mediante caracterizaciones alternativas de un concepto conocido
- Requiere:
  - Definición inicial del concepto
  - 1+ ejemplos del concepto
  - Teoría del dominio
  - Criterios operacionales
- Paradigmas: aprendizaje basado en explicaciones

- ESPECIALIZACION de una teoría general a EJEMPLOS

Aprendizaje por analogía

- Objetivo: encontrar la solución a partir de soluciones previas a problemas similares
- Requiere: Ejemplo de problemas y sus soluciones
- Paradigmas: Razonamiento basado en casos

- PROBLEMAS Y SOLUCIONES PASADOS adaptados a PROBLEMA y SOLUCION ACTUAL

Aprendizaje por refuerzo

- No hay ejemplos
- El sistema aprende mediante prueba y error
- Especialmente orientado a agentes que interaccionan con el entorno
  - El entorno ha de cuantificar el éxito o fracaso de las acciones
- EXPLORACIÓN del ENTORNO para obtener MODELO de COMPORTAMIENTO

Motivación II

- Recogida de datos y disponibilidad de los mismos.
  - Recogida de datos automática, sistemas de bases de datos, web, sociedad informatizada.
- Principales fuentes de datos.
  - Negocios: web, comercio electrónico, transacciones, stocks. . .
  - Ciencia: teledetección, bioinformática, simulaciones. . .
  - Sociedad, todos: noticias, cámaras digitales.

Motivación III

- *We are drowing in data, but starving for knowledge*
- Si se pudiera hacer algo útil con tanto dato. . .
- La necesidad es la madre de la invención.
  - Minería de Datos: análisis automático de conjuntos de datos masivos.

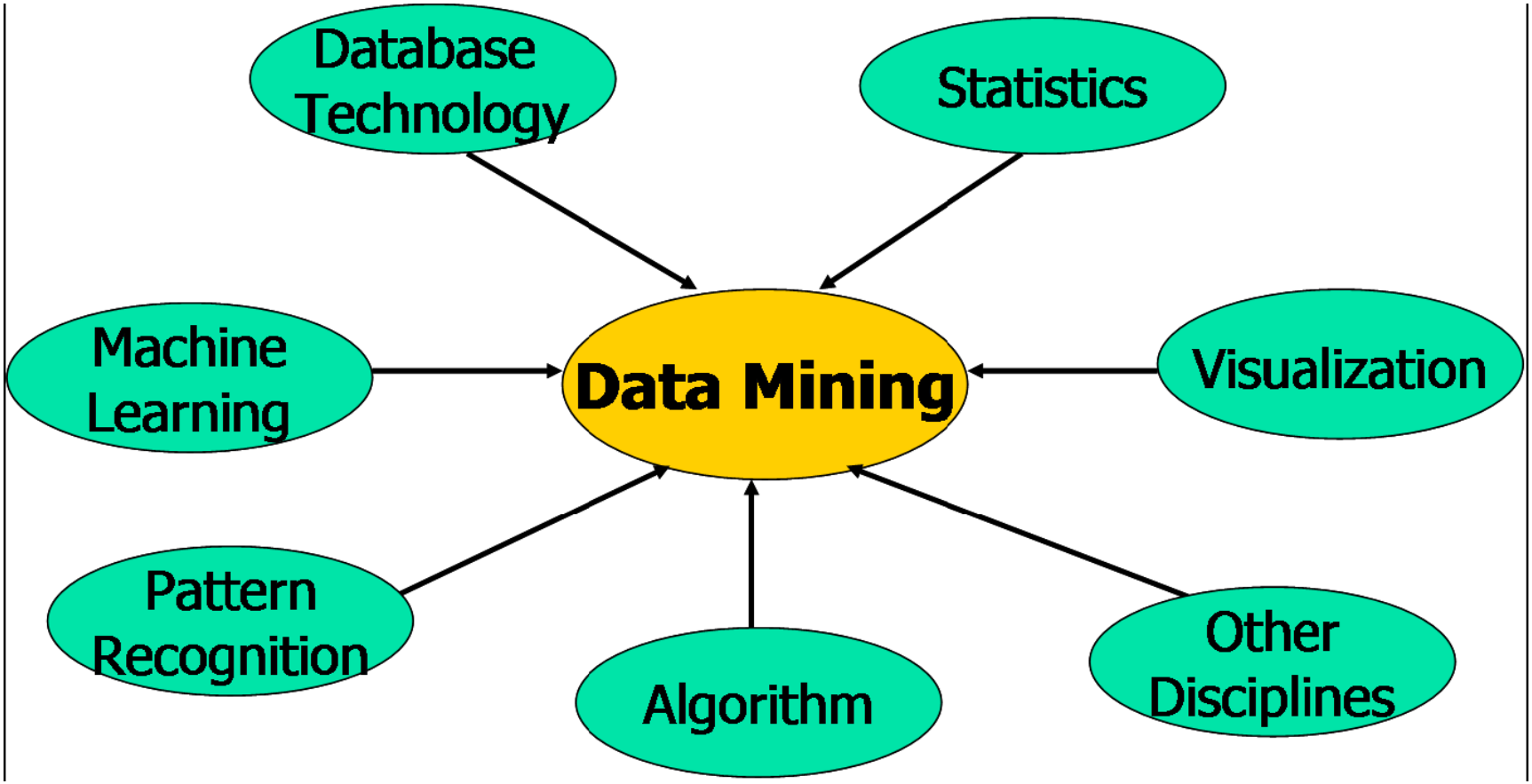
6.2 ¿Qué es la minería de datos?

- La **aplicación de técnicas de la inteligencia artificial** sobre grandes cantidades de datos, con el objetivo de descubrir tendencias, patrones, o relaciones ocultas.
- Un paso en el proceso de **descubrimiento de conocimiento en bases de datos (KDD)** que consiste en la aplicación de algoritmos de análisis de datos y descubrimiento que, sometidos a restricciones de eficiencia, producen una enumeración particular de patrones sobre los datos.
- Un área en la **intersección del aprendizaje computacional, la estadística y las bases de datos**.
- El proceso de **seleccionar, explorar y modelar** grandes cantidades de datos para descubrir patrones, previamente desconocidos, que proporcionen una ventaja competitiva.

Definición

- El análisis de conjuntos de datos (a menudo grandes) observados con el objetivo de
  - encontrar relaciones insospechadas
  - resumir los datos de maneras novedosas que sean
    - comprensibles
    - útiles
- Típicamente los datos se han recopilado para algún otro propósito

Relación con otras disciplinas I



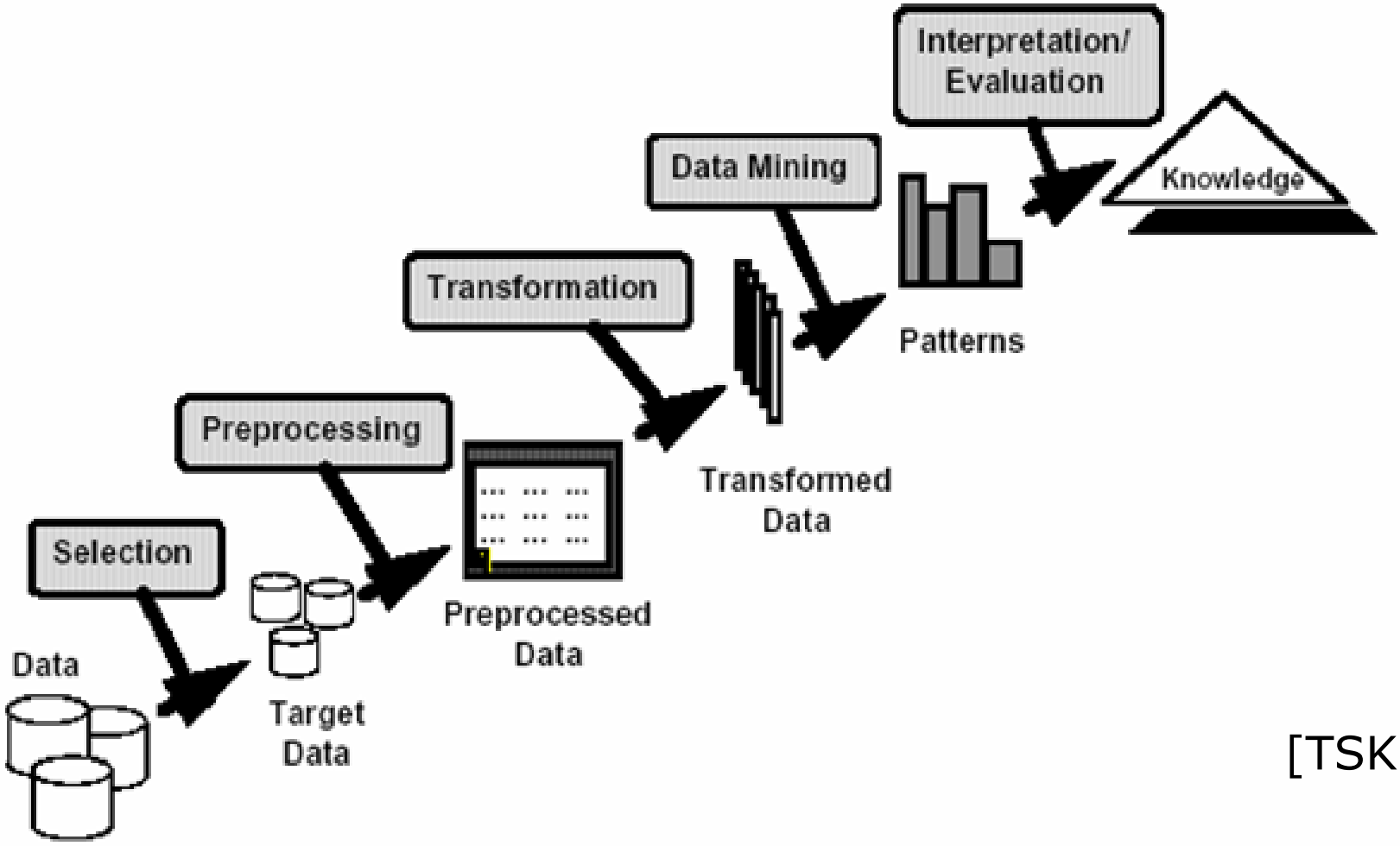
Relación con otras disciplinas II

- **Bases de datos.** De donde provienen los datos. Técnicas de indexación y acceso a datos.
  - Diferencia: extraer conocimiento novedoso y comprensible.
- **Recuperación de la información.** Obtener información a partir de datos textuales.
  - E.g., clasificación de documentos en función de palabras clave.
- **Estadística.** Fuente de conceptos, algoritmos, técnicas.
  - Comprobar hipótesis frente a encontrar hipótesis.
- **Aprendizaje automático.** Área de la IA, algoritmos capaces de aprender.

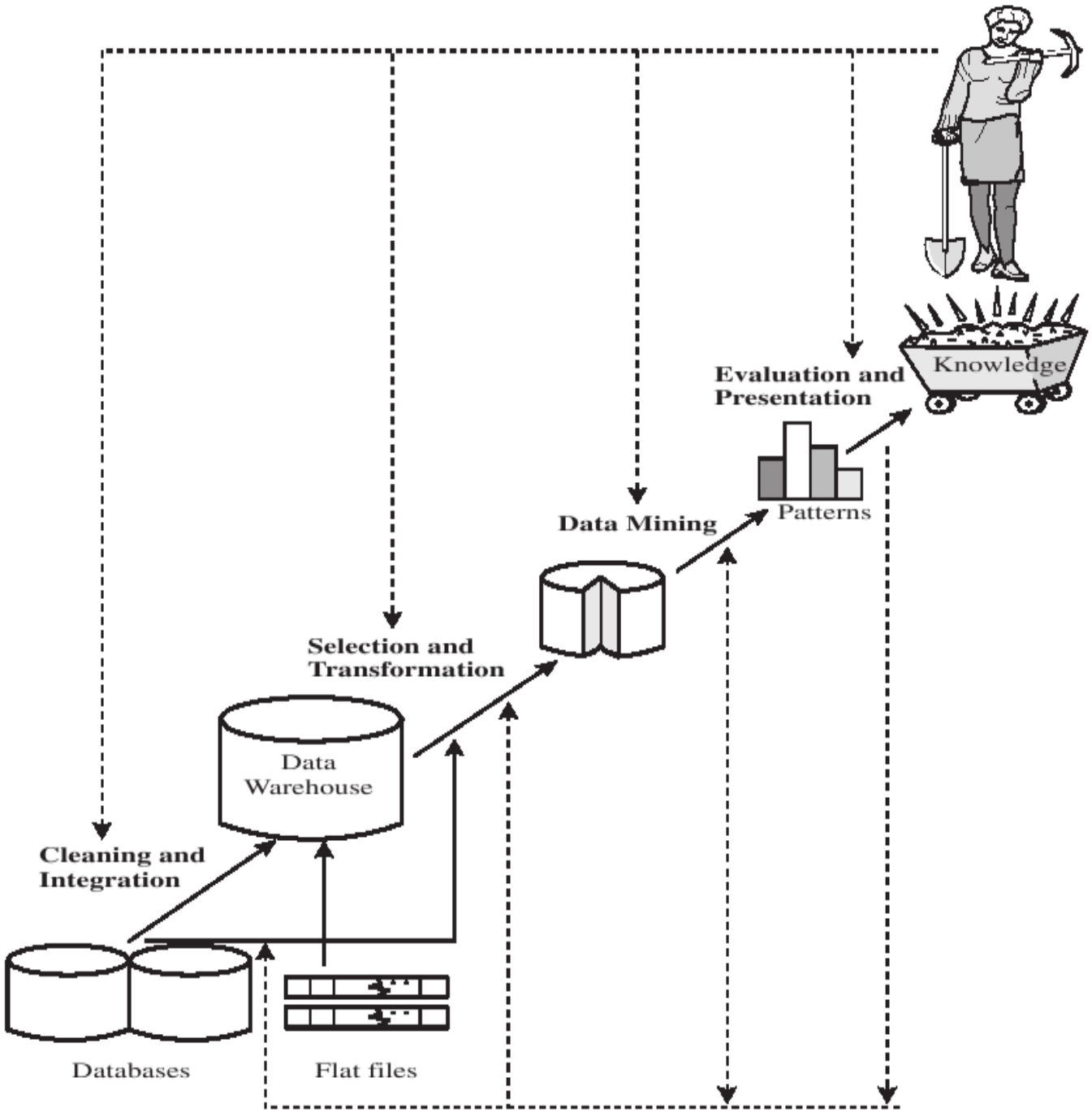
Relación con otras disciplinas III

- **Sistemas para la toma de la decisión.** Asistencia a directivos, diagnóstico. . .
- **Visualización de datos.** Describir, intuir o entender patrones. Difíciles de comprender a partir de fórmulas matemáticas o descripciones textuales.
- **Computación paralela y distribuida.** Elevado coste computacional de las tareas más complejas en MD, BD distribuídas.
- **Otras.** Dependientes del tipo de datos. Procesamiento del lenguaje natural, análisis de imágenes, procesamiento de señales. . .

6.3 Etapas del KDD



[TSK06]


Knowledge
Evaluation and
Presentation
Patterns
Data Mining
Selection and
Transformation
Data
Warehouse
Cleaning and
Integration
Databases
Flat files

[HK06]

Etapas del KDD

- **Limpieza de datos**. Eliminar ruido y datos inconsistentes
- **Integración de datos**. De distintas fuentes
- **Selección de datos**. Recuperar de la BD los datos relevantes para la tarea de análisis
- **Transformación de datos**. Los datos se transforman o consolidan en formas apropiadas para su minería (e.g., sumarios, agregación)
- **Minería de datos**. Aplicación de métodos inteligentes con el objetivo de extraer patrones
- **Evaluación de patrones**. Identificar los patrones verdaderamente interesantes
- **Presentación del conocimiento**. Visualización y representación del conocimiento para presentar el conocimiento extraído del usuario

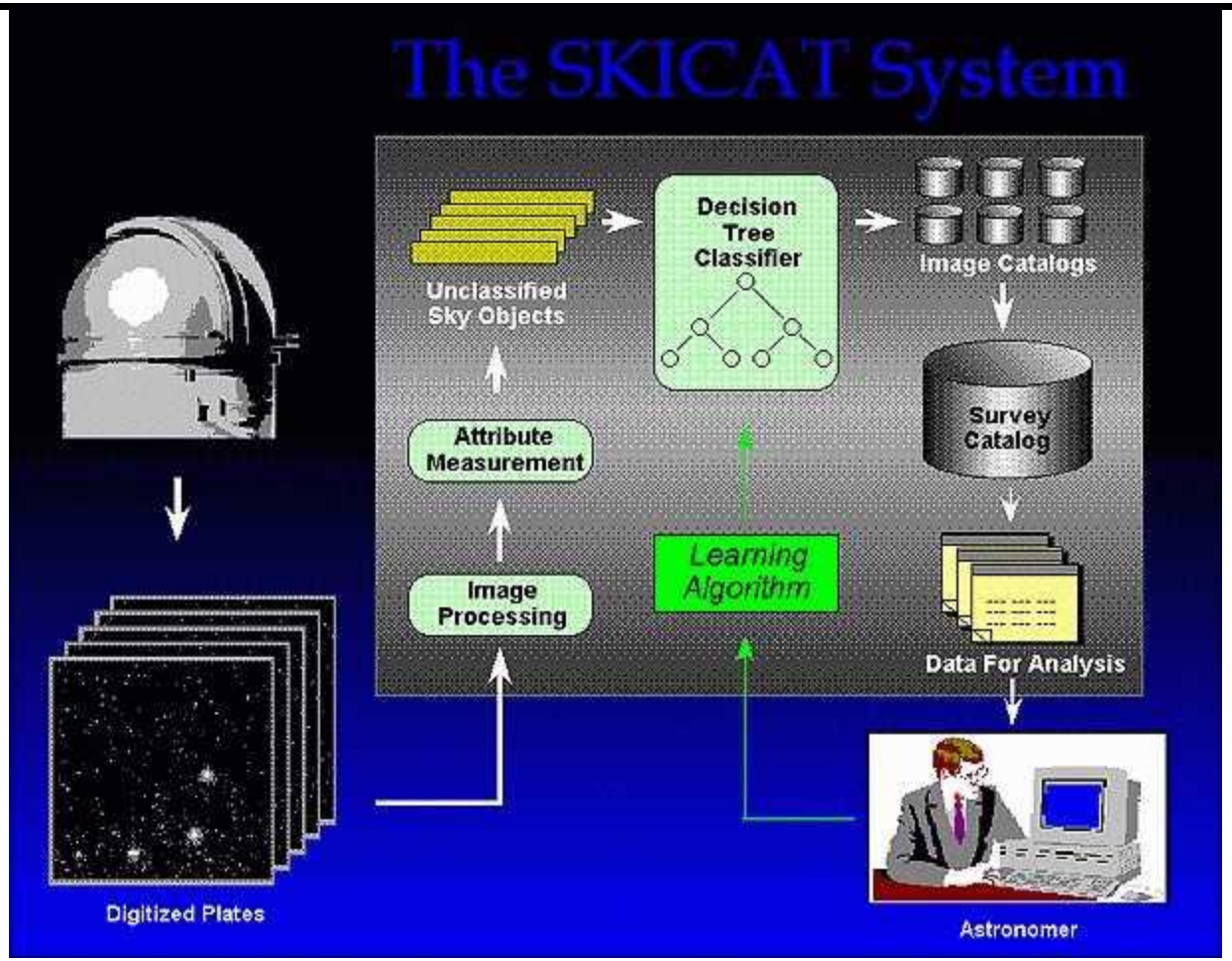
6.4 Posibles aplicaciones

- Análisis de dato y soporte a la decisión.
  - Análisis y gestión del mercado.
    - Marketing personalizado, CRM (Customer Relationship Management), market basket analysis, cross selling, segmentación del mercado
  - Análisis y gestión de riesgos.
    - Predicción, retención de clientes, aseguración mejorada, control de calidad, análisis competitivo.
  - Detección de fraudes y patrones inusuales (outliers).
- Otros.
  - Text mining, minería sobre flujos de datos, bioinformática.

Ejemplos de aplicación

- Diagnosis: lentes duras o blandas
- Bancarias: conceder o no un crédito
- Detección de fraudes: ¿es una transacción sospechosa?
- Mailings: ¿a quién?
- Rendimiento de ordenadores: como configurar
- Teledetección: polución del agua
- Predicción de carga: demanda de electricidad
- Cajeros inteligentes: cuanto dinero necesito
- Identificar grupos de usuarios similares de tarjetas
- Organizar e-mails
- Caracterizar intereses de un usuario de internet

Skicat
Sky Image Cataloging and Analysis Tool



6.5 Ética y minería de datos I

- La minería de datos se suele usar para discriminar
- La discriminación por ciertos criterios no es ética, e incluso puede ser ilegal
- Todo depende de la aplicación
  - Sí que se puede utilizar el sexo o la raza para diagnosis médica
- Aunque se eliminen ciertas variable, otras pueden indicar dicha información indirectamente
  - E.g.: código postal
- Al suministrar información, debe conocerse para que va a usarse
  - Muchas veces, en minería de datos, se pretende extraer información de datos que fueron recopilados para otro propósito

Ética y minería de datos II

- Resultados sorprendentes
  - Las personas que compran coches rojos son más propensas fallar en el pago del crédito
- Al trabajar con un conjunto de datos
  - Quién tiene permitido el acceso
  - Para que propósito se recopiló
  - Qué tipo de conclusiones es legítimo obtener
- Normas de los usuarios habituales de los datos
  - E.g.: privacidad de los usuarios de bibliotecas

Referencias

- Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan. Data Mining: A Knowledge discovery Approach. Spriner, 2007.
- Margaret H. Dunham. Data Mining: Introductory and Advanced Topics. Prentice Hall, 2003.
- Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd edition, 2006.
- David Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining. The MIT Press, 2001.
- José Hernández Orallo, M. José Ramírez Quintana, and César Ferri Ramírez, editors. Introducción a la Minería de Datos. Pearson Educación, 2004.
- Tom M. Mitchell. Machine Learning. McGraw-Hill, 1997.
- Sankar K. Pal, Pabitra Mitra. Pattern Recognition Algorithms for Data Mining. Chapman & Hall/CRC, 2004.
- Basilio Sierra. Aprendizaje Automático: conceptos básicos y avanzados. Pearson Educación, 2006.
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. Addison Wesley, 2006.
- I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd edition, 2005.