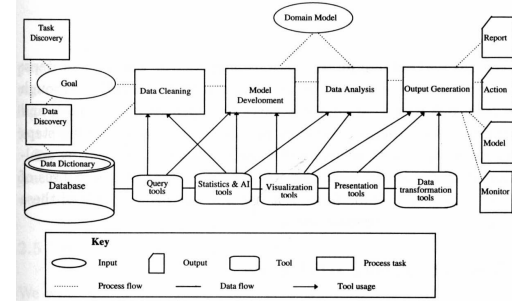


## 3.2. El Proceso de KDD

### Fases y Técnicas del KDD

Las distintas técnicas de distintas disciplinas se utilizan en distintas fases:



### Fases del KDD: Recogida de Datos

Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra:

- en bases de datos y otras **fuentes muy diversas**,
- tanto internas como **externas**.
- muchas de estas fuentes son las que se utilizan para el trabajo **transaccional**.

El análisis posterior será mucho más sencillo si la fuente es **unificada, accesible** (interna) y desconectada del trabajo **transaccional**.

3

### Fases del KDD: Recogida de Datos

El proceso subsecuente de minería de datos:

- Depende mucho de la fuente:
  - OLAP u OLTP.
  - Datawarehouse o copia con el esquema original.
  - ROLAP o MOLAP.
- Depende también del tipo de usuario:
  - ‘picapedreros’ (o ‘granjeros’): se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc.
  - ‘exploradores’: encargados de encontrar nuevos patrones significativos utilizando técnicas de minería de datos.

4

### Fases del KDD: Recogida de Datos

Recogida de Información Externa:

- Aparte de información interna de la organización, los almacenes de datos pueden recoger información externa:
  - Demografías (censo), páginas amarillas, psicografías (perfiles por zonas), uso de Internet, información de otras organizaciones.
  - Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
  - Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas-deportivas, catástrofes,...
  - Bases de datos externas compradas a otras compañías.

5

### Fases del KDD: Selección, Limpieza y Transformación de Datos

**Limpieza (data cleansing) y criba (selección) de datos:**

Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba).

Métodos estadísticos casi exclusivamente.

- histogramas (detección de datos anómalos).
- selección de datos (muestreo, ya sea verticalmente, eliminando atributos, u horizontalmente, eliminando tuplas).
- redefinición de atributos (agrupación o separación).

6

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Acciones ante datos anómalos (outliers):

- ignorar: algunos algoritmos son robustos a datos anómalos (p.ej. árboles)
- filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna discreta diciendo si el valor era normal u outlier (por encima o por debajo).
- filtrar la fila: puede sesgar los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
- reemplazar el valor: por el valor 'nulo' si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ML.
- discretizar: transformar un valor continuo en uno discreto (p.ej. muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

7

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Acciones ante datos faltantes (missing values):

- ignorar: algunos algoritmos son robustos a datos faltantes (p.ej. árboles).
- filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna booleana diciendo si el valor existía o no.
- filtrar la fila: claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.
- reemplazar el valor: por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ML.
- segmentar: se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
- modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

8

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Razones sobre datos faltantes (missing values):

A veces es importante examinar las razones tras datos faltantes y actuar en consecuencia:

- algunos valores faltantes expresan características relevantes: p.ej. la falta de teléfono puede representar en muchos casos un deseo de que no se moleste a la persona en cuestión, o un cambio de domicilio reciente.
- valores no existentes: muchos valores faltantes existen en la realidad, pero otros no. P.ej. el cliente que se acaba de dar de alta no tiene consumo medio de los últimos 12 meses.
- datos incompletos: si los datos vienen de fuentes diferentes, al combinarlos se suele hacer la unión y no la intersección de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una/s fuente/s diferente/s al resto.

9

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Transformación del Esquema:

- Esquema Original:
  - Ventajas: Las R.I. se mantienen (no hay que reaprenderlas, no despidan)
  - Inconvenientes: Muchas técnicas no se pueden utilizar.
- Tabla Universal: *Cualquier Esquema Relacional se puede convertir (en una correspondencia 1 a 1) a una tabla universal.*
  - Ventajas: Modelos de aprendizaje más simples (proposicionales).
  - Desventajas: Muchísima Redundancia (tamaños ingentes). La información del esquema se pierde. Muchas dependencias funcionales se vuelven a re-descubrir!! Se debe añadir metainformación.
- Desnormalizado Tipo Estrella o Copo de Nieve (*datamarts*):
  - Ventajas: Se pueden buscar reglas sobre información sumariada y si resultan factibles se pueden comprobar con la información detallada. Con operadores propios: *Roll-up, Drill-down, Slicing and Dicing*.
  - Desventajas: Orientadas a extraer un tipo de información (granjeros)<sub>10</sub>

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Intercambio de Dimensiones: (filas por columnas)

#### EJEMPLO:

Una tabla de cestas de la compra, donde cada atributo indica si el producto se ha comprado o no.

- Objetivo: Ver si dos productos se compran conjuntamente (regla de asociación).

Es muy costoso: hay que mirar al menos la raíz cuadrada de todas las relaciones (cestas).

Y puede haber millones en una semana...

Sin embargo...

Productos sólo hay unos 10.000.

11

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Intercambio de Dimensiones: EJEMPLO

Si se intercambian filas por columnas tenemos:

	B1	B2	B3	B4	B5	B6	...
Jabón	X		X				
Huevos		X			X		
Patatas Fritas		X			X		
Champú							
Jabón + Champú	X		X				
Huevos + Patatas							

Sólo es necesario hacer XOR entre dos filas para saber si hay asociación.

12

## Fases del KDD: Selección, Limpieza y Transformación de Datos

### Transformación de los Campos:

- Numerización / Etiquetado
  - Ventajas: Se reduce espacio. Ej: apellido  $\Rightarrow$  entero. Se pueden utilizar técnicas más simples.
  - Desventajas: Se necesita meta-información para distinguir los datos inicialmente no numéricos (la cantidad no es relevante) de los inicialmente numéricos (la cantidad es relevante: precios, unidades, etc.)  
A veces se puede "sesgar" el modelo (*biasing*).
- Discretización:
  - Ventajas: Se reduce espacio. Ej. 0..10  $\Rightarrow$  (pequeño, mediano, grande). Se pueden utilizar árboles de decisión y construir reglas discretas.
  - Desventajas: Una mala discretización puede invalidar los resultados.

13

## Fases del KDD: La Minería de Datos

### Características Especiales de los Datos:

Aparte del gran volumen, ¿por qué las técnicas de aprendizaje automático y estadística no son *directamente* aplicables?

- Los datos residen en el disco. No se pueden escanear múltiples veces.
- Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales.
- Muy alta dimensionalidad (muchos campos).
- Evidencia POSITIVA.
- DATOS IMPERFECTOS...

*Aunque algunos se aplican casi directamente, el interés en la investigación en minería de datos está en su adaptación.*

14

## Fases del KDD: La Minería de Datos

### Patrones a descubrir:

- Una vez recogidos los datos de interés, un explorador puede decidir qué tipo de patrón quiere descubrir.
- El tipo de conocimiento que se desea extraer va a marcar claramente la *técnica* de minería de datos a utilizar.
- Según como sea la búsqueda del conocimiento se puede distinguir entre:
  - *Directed data mining*: se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases.
  - *Undirected data mining*: no se sabe lo que se busca, se trabaja con los datos (*hasta que confiesen!*).
- En el primer caso, algunos sistemas de minería de datos se encargan generalmente de elegir el *algoritmo* más idóneo entre los disponibles para un determinado tipo de patrón a buscar.

15

## Fases del KDD: Evaluación y Validación

La fase anterior produce una o más hipótesis de modelos.

Para seleccionar y validar estos modelos es necesario el uso de **criterios de evaluación de hipótesis**.

Por ejemplo:

1ª Fase: Comprobación de la precisión del modelo en un **banco de ejemplos independiente** del que se ha utilizado para aprender el modelo. Se puede elegir el mejor modelo.

2ª Fase: Se puede realizar una **experiencia piloto** con ese modelo. Por ejemplo, si el modelo encontrado se quería utilizar para predecir la respuesta de los clientes a un nuevo producto, se puede enviar un mailing a un subconjunto de clientes y evaluar la *fiabilidad del modelo*.

16

## Fases del KDD: Interpretación y Difusión

El despliegue del modelo a veces a veces es trivial pero otras veces requiere un proceso de implementación o interpretación:

- El modelo puede requerir **implementación** (p.ej. tiempo real detección de tarjetas fraudulentas).
- El modelo es descriptivo y requiere **interpretación** (p.ej. una caracterización de zonas geográficas según la distribución de los productos vendidos).
- El modelo puede tener muchos usuarios y necesita **difusión**: el modelo puede requerir ser expresado de una manera comprensible para ser distribuido en la organización (p.ej. las cervezas y los productos congelados se compran frecuentemente en conjunto  $\Rightarrow$  ponerlos en estantes distantes),

## Fases del KDD: Actualización y Monitorización

Los procesos derivan en un mantenimiento:

- Actualización: Un modelo válido puede dejar de serlo: cambio de contexto (económicos, competencia, fuentes de datos, etc.).
- Monitorización: Consiste en ir revalidando el modelo con cierta frecuencia sobre nuevos datos, con el objetivo de detectar si el modelo requiere una actualización.

Producen realimentaciones en el proceso KDD.

18

## Tipología de Técnicas de Minería de Datos

Las técnicas de minería de datos crean modelos que son **predictivos** y/o **descriptivos**.

Un modelo predictivo responde preguntas sobre datos futuros.

- ¿Cuáles serán las ventas el año próximo?
- ¿Es esta transacción fraudulenta?
- ¿Qué tipo de seguro es más probable que contrate el cliente X?

Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características. Genera información del tipo:

- Los clientes que compran pañales suelen comprar cerveza.
- El tabaco y el alcohol son los factores más importantes en la enfermedad Y.
- Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto.

19

## Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Predictivo:

- Queremos saber si jugar o no jugar esta tarde al tenis.
- Hemos recogido datos de experiencias anteriores:

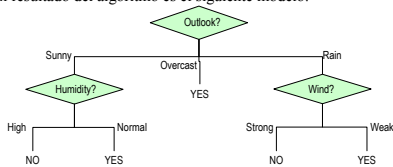
Example	Sky	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

20

## Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Predictivo:

- Pasamos estos ejemplos a un algoritmo de aprendizaje de árboles de decisión, señalando el atributo "PlayTennis" como la clase (output).
- El resultado del algoritmo es el siguiente modelo:



- Ahora podemos utilizar este modelo para predecir si esta tarde jugamos o no al tenis. P.ej., la instancia:

(Outlook = sunny, Temperature = hot, Humidity = high, Wind = strong) es NO.

21

## Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Descriptivo:

- Queremos categorizar nuestros empleados.
- Tenemos estos datos de los empleados:

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo
1	10000	Si	No	0	Alquiler	No	7	15	H
2	20000	No	Si	1	Alquiler	Si	3	3	M
3	15000	Si	Si	2	Prop	Si	5	10	H
4	30000	Si	Si	1	Alquiler	No	15	7	M
5	10000	Si	Si	0	Prop	Si	1	6	H
6	40000	No	Si	0	Alquiler	Si	3	16	M
7	25000	No	No	0	Alquiler	Si	0	8	M
8	20000	No	Si	0	Prop	Si	2	6	H
9	20000	Si	Si	3	Prop	No	7	5	H
10	30000	Si	Si	2	Prop	No	1	20	H
11	50000	No	No	0	Alquiler	No	2	12	M
12	8000	Si	Si	2	Prop	No	3	1	H
13	20000	No	No	0	Alquiler	No	27	5	M
14	10000	No	Si	0	Alquiler	Si	0	7	H
15	8000	No	Si	0	Alquiler	No	3	2	H

22

## Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Descriptivo:

- Pasamos estos ejemplos a un algoritmo de clustering K-means.
- Se crean tres clusters, con la siguiente descripción:

<b>cluster 1: 5 examples</b> Sueldo : 22600 Casado : No -> 0.8 Si -> 0.2 Coche : No -> 0.8 Si -> 0.2 Hijos : 0 Alq/Prop : Alquiler -> 1.0 No -> 0.8 Si -> 0.2 Bajas/Año : 8 Antigüedad : 8 Sexo : H -> 0.6 M -> 0.4	<b>cluster 2: 4 examples</b> Sueldo : 22500 Casado : No -> 1.0 Si -> 1.0 Coche : Si -> 1.0 Hijos : 0 Alq/Prop : Alquiler -> 0.75 Prop -> 0.25 Sindic. : Si -> 1.0 Bajas/Año : 2 Antigüedad : 8 Sexo : H -> 0.25 M -> 0.75	<b>cluster 3: 6 examples</b> Sueldo : 18833 Casado : Si -> 1.0 Si -> 1.0 Coche : Si -> 1.0 Hijos : 2 Alq/Prop : Alquiler -> 0.17 Prop -> 0.83 Sindic. : No -> 0.67 Si -> 0.33 Bajas/Año : 5 Antigüedad : 8 Sexo : H -> 0.83 M -> 0.17
--	---	--

- GRUPO 1: Sin hijos y de alquiler. Poco sindicados. Muchas bajas.
- GRUPO 2: Sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente de alquiler y mujeres.
- GRUPO 3: Con hijos, casados y con coche. Propietarios. Poco sindicados. Hombres.

## Tipología de Técnicas de Minería de Datos

Tipos de conocimiento:

- **Asociaciones:** Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.

- Ejemplo, en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.

- **Dependencias:** Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Ojo! Existen muchas dependencias nada interesantes (causalidades inversas).

- Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

La búsqueda de asociaciones y dependencias se conoce a veces como análisis exploratorio.

24

## Tipología de Técnicas de Minería de Datos

Tipos de conocimiento (cont.):

- **Clasificación:** Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.
  - Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, número de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria.
  - Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.
- **Agrupamiento / Segmentación:** El agrupamiento (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

25

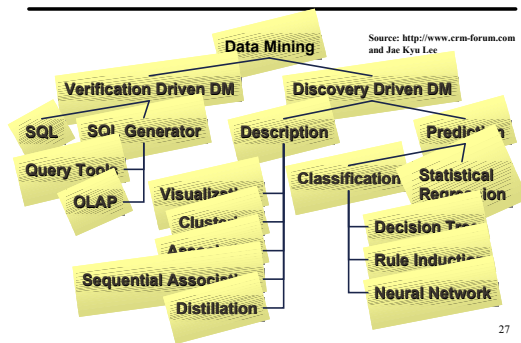
## Tipología de Técnicas de Minería de Datos

Tipos de conocimiento (cont.):

- **Tendencias/Regresión:** El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo.
  - Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.
- **Información del Esquema:** (descubrir claves primarias alternativas, R.I.).
- **Reglas Generales:** patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

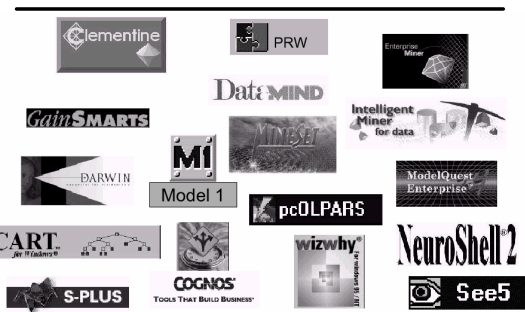
26

## Taxonomía Técnicas de Minería de Datos.



27

## Sistemas



Elder Research,  
[www.datamininglab.com](http://www.datamininglab.com)

28

## Sistemas

Tipos de Sistemas:

- **Standalone:** Los datos se deben exportar/convertir al formato interno del sistema de data mining: Knowledge Seeker IV (Angoss International Limited, Groupe Bull).
- **On-top:** pueden funcionar sobre un sistema propietario (Clementine sobre ODBC, microstrategy sobre Oracle).
- **Embedded** (propietarios): Oracle Discoverer, Oracle Darwin, IBM...
- **Extensible** (Tecnología *Plug-ins*): proporcionan unas herramientas mínimas de interfaz con los datos, estadísticas y visualización, y los algoritmos de aprendizaje se pueden ir añadiendo con plug-ins. (ej. KEPLER).

29

## Sistemas

Producto	Compañía	Técnicas	Plataformas	Interfaz
Knowledge Seeker	Angoss	Decision Trees, Statistics	Win NT	ODBC
CART	Salford Systems	Decision Trees	UNIX/NT	
Clementine	SPSS-Integral Solutions Limited (ISI)	Decision Trees, ANN, Statistics, Rule Induction, Association Rules, K Means, Linear Regression	UNIX/NT	ODBC
Data Surveyor	Data Explorer	Amplio Abanico	UNIX	ODBC
GainSmarts	Urban Science	Especializado en gráficos de ganancias en campañas de clientes (solo Decision Trees, Linear Statistics y Logistic Regression)	UNIX/NT	
Intelligent Miner	IBM	Decision Trees, Association Rules, ANN, RBF, Time Series, K Means, Linear Regression	UNIX (AIX)	IBM, DB2
Microstrategy	Microstrategy	Datawarehouse solo	Win NT	Oracle
Polyanalyst	Megaputer	Symbolic, Evolutionary	Win NT	Oracle, ODBC
Darwin	Oracle	Amplio Abanico (Decision Trees, ANN, Nearest Neighbor)	UNIX/NT	Oracle
Enterprise Miner	SAS	Decision Trees, Association rules, ANN, regression, clustering	UNIX (Sun), NT, Mac	Oracle, ODBC
SGI MMSet	Silicon Graphics	association rules and classification models, used for prediction, scoring, segmentation and profiling	UNIX (Irix)	Oracle, Sybase, Informatica
Wizsoft/Wizwhy				

30

## Sistemas

- Más software comercial DM:
  - [http://www.kdcentral.com/Software/Data\\_Mining/](http://www.kdcentral.com/Software/Data_Mining/)
  - <http://www.the-data-mine.com/bin/veiw/Software/WebIndex>
- Algunos Prototipos No Comerciales o Gratuitos:
  - Kepler: sistema de plug-ins del GMD (<http://ais.gmd.de/KD/kepler.html>).
  - Rproject: herramienta gratuita de análisis estadístico (<http://www.R-project.org/>)
  - Librerías WEKA (<http://www.cs.waikato.ac.nz/~ml/weka/>) (Witten & Frank 1999)

31

## Sistemas

EJEMPLO: **Clementine** (Integral Solutions Limited (ISL))  
[www.spss.com](http://www.spss.com)

- Herramienta que incluye:
  - fuentes de datos (ASCII, Oracle, Informix, Sybase e Ingres).
  - interfaz visual.
  - distintas herramientas de minería de datos: redes neuronales y reglas.
  - manipulación de datos (pick & mix, combinación y separación).

32

## Sistemas

EJEMPLO: **Clementine**

Ejemplo Práctico: Ensayo de Medicamentos

[http://www.pcc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final\\_3.html](http://www.pcc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final_3.html)

- Un número de pacientes hospitalarios que sufren todos la misma enfermedad se tratan con un abanico de medicamentos.
- 5 medicamentos diferentes están disponibles y los pacientes han respondido de manera diferente a los diferentes medicamentos.
- Problema:

¿qué medicamento es apropiado para un nuevo paciente?

33

## Sistemas

EJEMPLO: **Clementine**. Ejemplo Práctico: Ensayo de Medicamentos

Primer Paso: **ACCEDIENDO A LOS DATOS:**

- Se leen los datos. Por ejemplo de un fichero de texto con delimitadores.
- Se nombran los campos:

age	edad
sex	sexo
BP	presión sanguínea (High, Normal, Low)
Cholesterol	colesterol (Normal, High)
Na	concentración de sodio en la sangre.
K	concentración de potasio en la sangre.
drug	medicamento al cual el paciente respondió satisfactoriamente.

SE PUEDEN COMBINAR LOS DATOS:

P.ej. se puede añadir un nuevo atributo: Na/K

34

## Sistemas

EJEMPLO: **Clementine**

Segundo Paso: **Familiarización con los Datos.** Visualizamos los registros:

Age	Sex	BP	Cholesterol	Na	K	Drug	Na_to_K
23	F	HIGH	HIGH	0.79	0.05	drugY	25.35
47	M	LOW	HIGH	0.74	0.06	drugC	13.09
47	M	LOW	HIGH	0.7	0.07	drugC	10.11
28	F	NORMAL	HIGH	0.56	0.07	drugX	7.9
51	F	LOW	HIGH	0.56	0.03	drugY	18.04
22	F	NORMAL	HIGH	0.68	0.08	drugX	8.61
49	F	NORMAL	HIGH	0.79	0.05	drugY	16.28
41	M	LOW	HIGH	0.77	0.07	drugC	11.04
60	M	NORMAL	HIGH	0.78	0.05	drugY	15.17
43	M	LOW	NORMAL	0.53	0.03	drugY	19.37

35

## Sistemas

EJEMPLO: **Clementine**

- Permite seleccionar campos o filtrar los datos
- Permite mostrar propiedades de los datos. Por ejemplo:  
 ¿Qué proporción de casos respondió a cada medicamento?

Value	Proportion	%	Occurrences
drugB		11.5	23
drugB		8.0	16
drugC		8.0	16
drugX		27.0	54
drugY		45.5	91

36





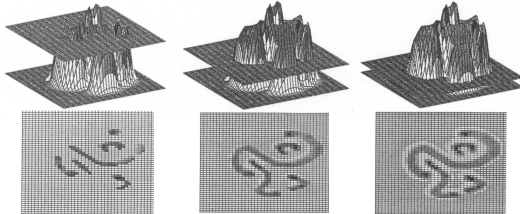


## Visualización

### Visualización Previa:

Ejemplo: segmentación mediante funciones de densidad, generalmente representadas tridimensionalmente.

Los seres humanos ven claramente los segmentos (clusters) que aparecen con distintos parámetros



## Visualización

### Visualización Previa:

Mayor problema: dimensionalidad > 3.

Objetivo: conseguir proyectar las dimensiones en una representación en 2 (ó 3 simuladas) dimensiones.

Solución:

Uso de proyecciones geométricas:

50

## Visualización

### Visualización Previa: Proyecciones geométricas:

- técnica de visualización de coordenadas paralelas [Inselberg & Dimsdale 1990]. Se mapea el espacio  $k$ -dimensional en dos dimensiones mediante el uso de  $k$  ejes de ordenadas (escalados linealmente) por uno de abscisas. Cada punto en el espacio  $k$ -dimensional se hace corresponder con una línea poligonal (polígono abierto), donde cada vértice de la línea poligonal intersecta los  $k$  ejes en el valor para la dimensión.
  - Cuando hay pocos datos cada línea se dibuja de un color.
  - Cuando hay muchos datos se utiliza una tercera dimensión para los casos.

- técnica radial (igual que la anterior pero los ejes se ponen circularmente) →



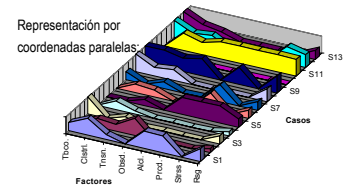
51

## Visualización

### Visualización Previa: Ejemplo: dimensionalidad...

Dados ciertos atributos de pacientes (tabaquismo, colesterol, tensión, obesidad, alcoholismo, precedentes, estrés) y su riesgo (muy bajo, bajo, medio, alto, muy alto) de enfermedades coronarias:

Tabac	Cholest	Ten	Obes	Alco	Preced	Estrés	Ries
Med	Alto	8	No	Si	Si	No	Alto
Bajo	Med	9	Si	No	No	No	Bajo
Alto	Bajo	0.5	No	No	No	No	Med
Bajo	Med	7	No	No	No	No	Bajo
Bajo	Bajo	0.5	No	Si	Si	Si	Med
Bajo	Med	9	No	No	Si	No	Med
Med	Bajo	9	No	No	Si	No	Med
Alto	Med	11	No	No	No	No	Alto
Alto	Alto	13	Si	No	Si	No	M.A
Bajo	Alto	7	No	No	No	No	M.B
Bajo	Med	12	Si	Si	Si	Si	M.A
Alto	Med	11	No	No	No	No	Alto
Alto	Med	8	No	No	No	No	Med



El mayor problema de estas representaciones (y de otras muchas) es que no acomodan bien las variables discretas.

## Visualización

### Visualización Previa:

- Iconicas: Existen otro tipo de técnicas que sí permiten combinar atributos continuos y discretos, mediante el uso de transformaciones menos estándar y el uso de iconos.
  - Se utilizan rasgos compatibles y diferenciados para distintas dimensiones, como son círculos, estrellas, puntos, etc., con la ventaja de que se pueden combinar más convenientemente valores discretos y continuos.
- Otras aproximaciones más sofisticadas se basan en estructuras jerárquicas, como por ejemplo, los Cone Trees [Robertson et al. 1991].

53

## Visualización

### Visualización Posterior:

Se utiliza para mostrar los patrones y entenderlos mejor.

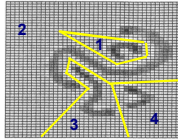
- Un árbol de decisión es un ejemplo de visualización posterior.
- Otros gráficos de visualización posterior de patrones:
  - muestran una determinada segmentación de los datos, una asociación, una determinada clasificación.
  - utilizan para ello gráficos de visualización *previa* en los que además se señala el patrón.
  - permiten evaluar gráficamente la calidad del modelo.

54

## Visualización

### Visualización Posterior:

EJEMPLO: se muestra una segmentación lineal para el corte del ejemplo anterior:



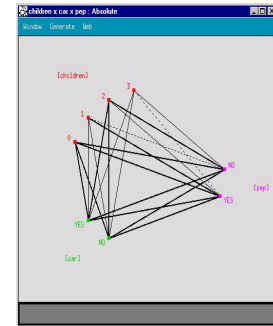
55

## Visualización

### Visualización Posterior:

EJEMPLO:

se muestra el grado de asociación según la línea que conecta los valores (continua gruesa, continua, discontinua o inexistente):



## Visualización

### Visualización Posterior:

EJEMPLO:

representación de ganancias acumulativas de un árbol de decisión:

$$lift^a = \arcsen No/Total$$

El árbol óptimo sería así:

