Jesús García Herrero

INTRODUCCIÓN

En esta clase introducimos la materia del Análisis de Datos, con una motivación general a estas técnicas, dominios más relevantes de aplicación, áreas de conocimiento y panorámica de los tipos de técnicas empleadas.

Se revisan los aspectos de infraestructura, metodología y elementos necesarios para abordar cualquier estudio de análisis de datos, para a continuación explicar los conceptos clave que se desarrollarán durante el resto del curso: tipos de modelos de aprendizaje (supervisado y no supervisado), preparación de datos, instancias y atributos y transformaciones más relevantes

Se detallan los principales tipos de modelos de aprendizaje (numérico, reglas, agrupamiento, asociación), y operaciones más frecuentes de limpieza y preparación de datos: eliminación de datos ruidosos, tratamiento de datos incompletos, normalización y proyecciones.

Finalmente se presentan los principales problemas abiertos en el análisis de datos, problemas técnicos dados por las limitaciones de los modelos utilizados, y problemas no técnicos según la información que se precise acceder y explotación de los modelos resultantes.

Análisis de Datos

Introducción y visión general

Jesús García Herrero
Universidad Carlos III de Madrid

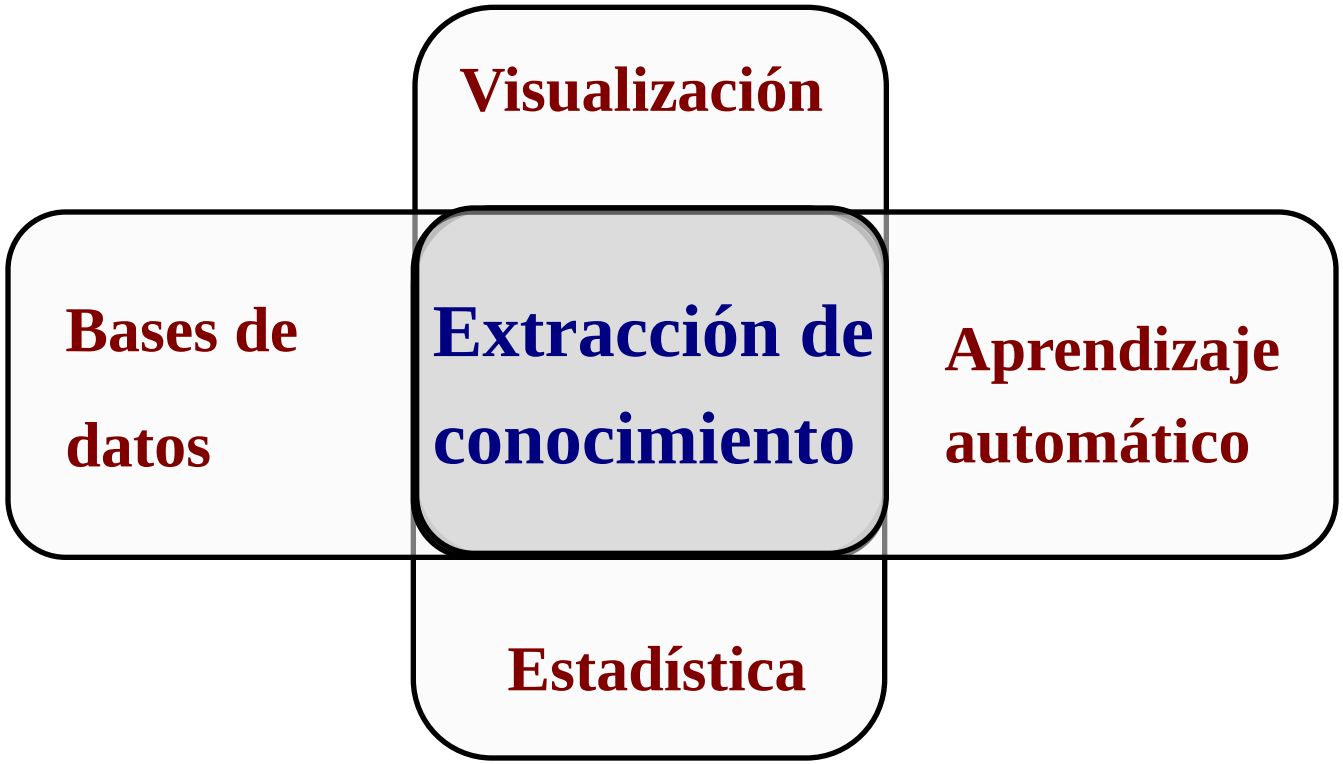
Universidad
Carlos III de Madrid

Introducción

- La sociedad de la información e Internet han generado una explosión de datos
  - No hay suficiente gente que pueda analizar tal cantidad de datos
  - Potencia de computación disponible
  - El desarrollo de software es un cuello de botella
- Extraer conocimiento a través de ejemplos es atractivo
  - Aprendizaje de la experiencia para tomar decisiones
  - Servicios comerciales, financieros, etc. tienden hacia la personalización: adaptación al individuo
- Es muy importante la comprensibilidad de la salida

Extracción de conocimiento

- Descubrimiento de patrones, relaciones y tendencias mediante análisis de gran cantidad de datos



Ejemplos de aplicaciones

- Toma de decisiones
  - Cuándo concedo un crédito hipotecario? por cuánto? Qué tipo de solicitante no devolverá el crédito?
  - Un cliente de tarjeta de crédito está realizando una compra, pagará? se la han robado?
- Diagnósticos
  - Determinación de enfermedades
  - Fallos en procesos industriales
- Marketing y ventas
  - Hábitos y fidelidad de clientes. Cuál es el perfil de los clientes que se gastan al mes más de 100.000 pts?
  - Análisis de compras. Qué productos de nuestra empresa es el que compran los clientes junto al detergente?
  - Análisis de perfil más adecuado para publicidad directa.

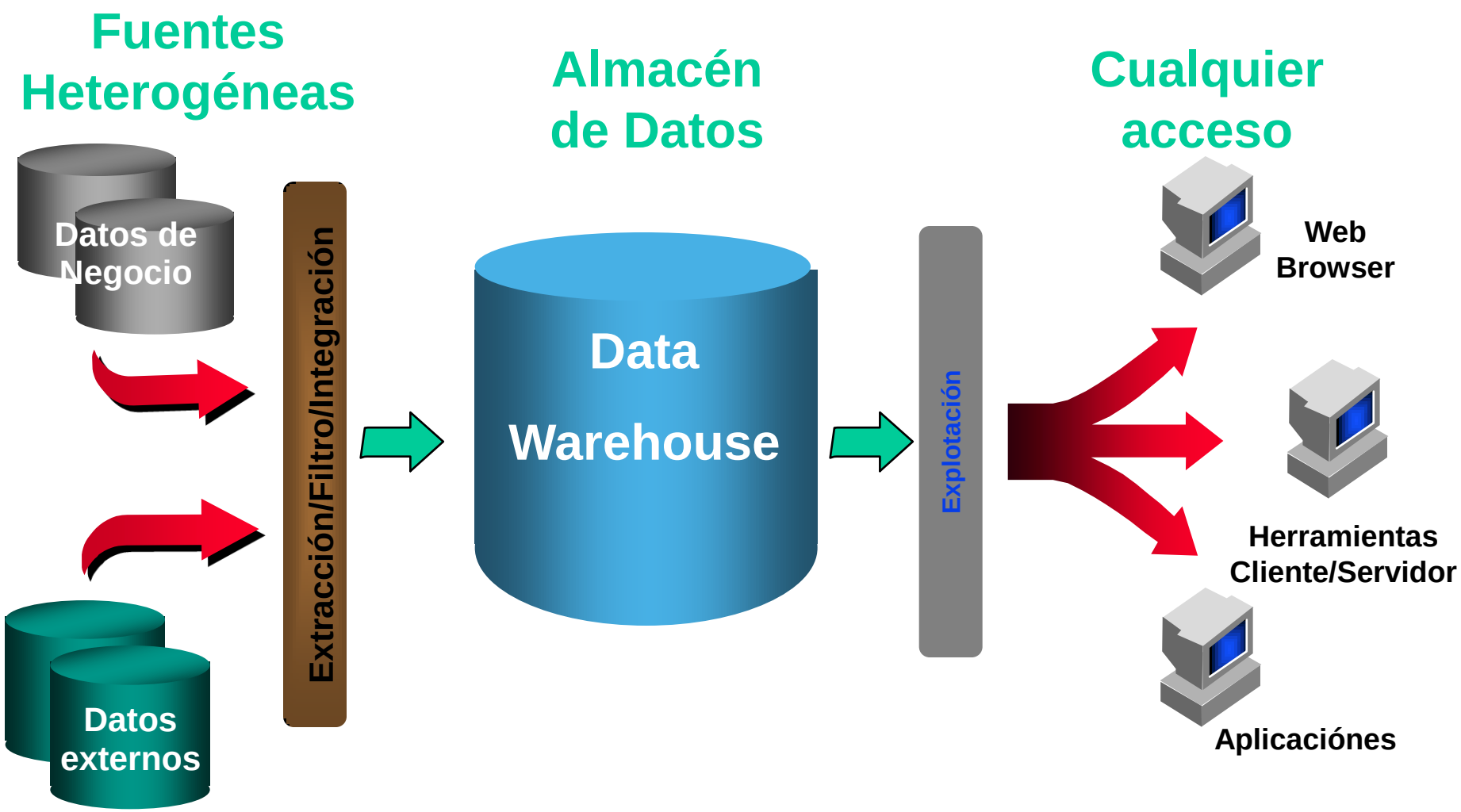
Ejemplos de aplicaciones

- Predicción
  - Cuánta energía se va a consumir en los próximos días?
- Agencia tributaria
  - Cuál es el perfil de los “defraudadores"?
  - Se puede subdividir en grupos homogéneos y caracterizar los diferentes tipos de contribuyentes?
  - Cuáles están más alejados de cada grupo?
- Herramienta de investigación. Ej.: imágenes:
  - Dada una imagen tomada por un telescopio, soy capaz de detectar y clasificar objetos interesantes?
  - Alerta de fuegos, fugas de combustible, militares, etc.?
- Mejora de procesos industriales
- ...

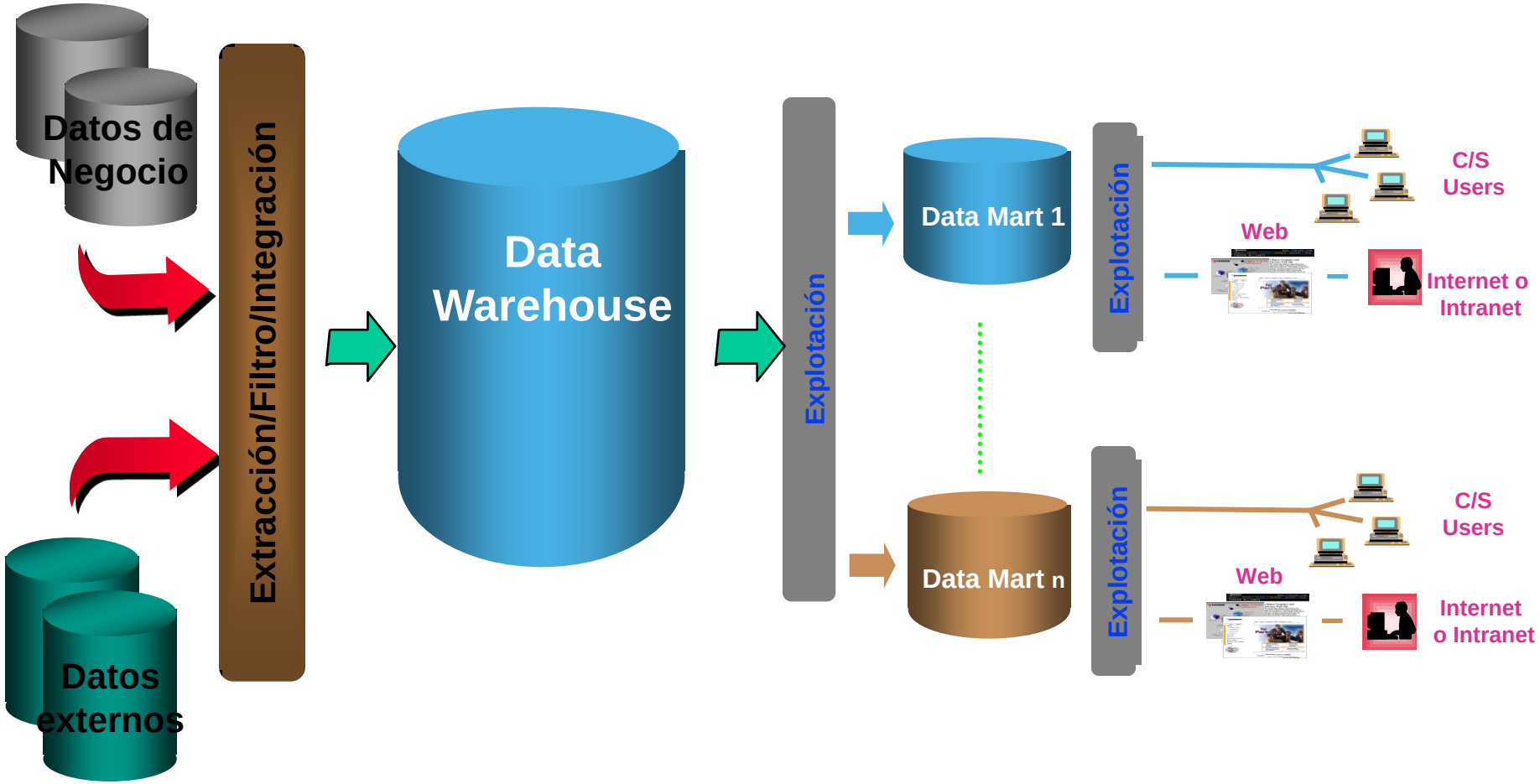
Análisis de datos en Internet

- Web Mining: análisis de páginas para extraer automáticamente información
- e-Mining: análisis de las interacciones de los clientes con mis páginas
- Web para extraer información
- Tipo de información que busco:
  - Qué tipo de clientes tengo
  - Cómo interacciona cada tipo de cliente con las páginas Web
  - Qué banners son los que siguen mis clientes (publicidad)
  - Descubrimiento de patrones de compra/navegación
- Herramientas de gestión automática del correo

Data Warehouse



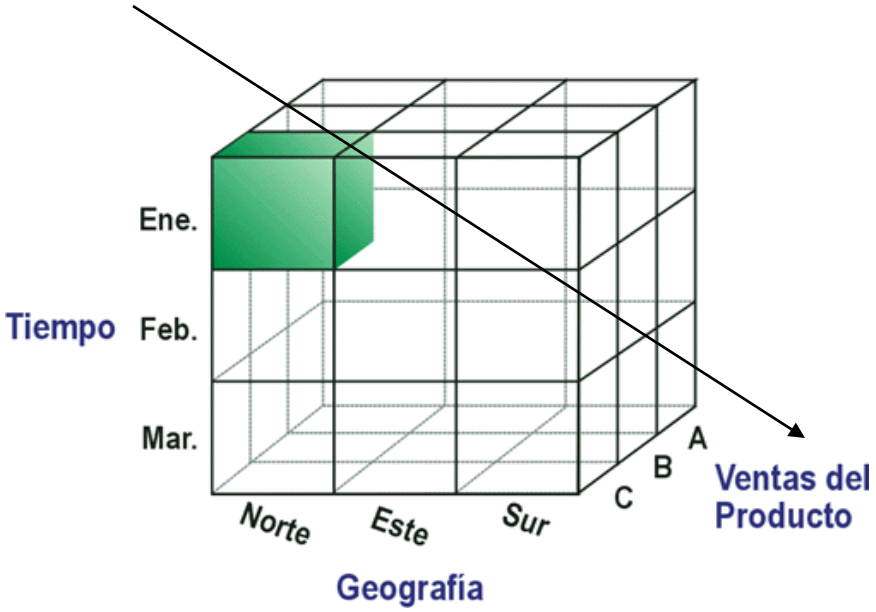
Data Warehouse y Data Marts



¿Qué es un CUBO de información?

Un “cubo” es una estructura para almacenar información que permite realizar análisis multidimensional y se basa en métricas y dimensiones.

Métrica: Medición matemática de una variable del negocio.
Qué quiero medir.

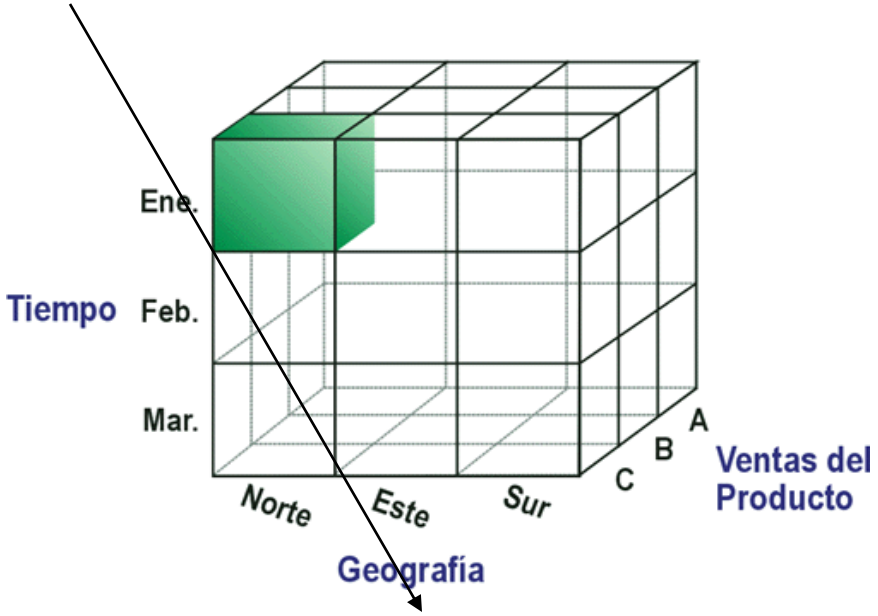


- cantidad de ventas
- unidades vendidas
- % desecho
- # productos en almacén
- etc.

¿Qué es un CUBO de información?

Un “cubo” es una estructura para almacenar información que permite realizar análisis multidimensional y se basa en métricas y dimensiones.

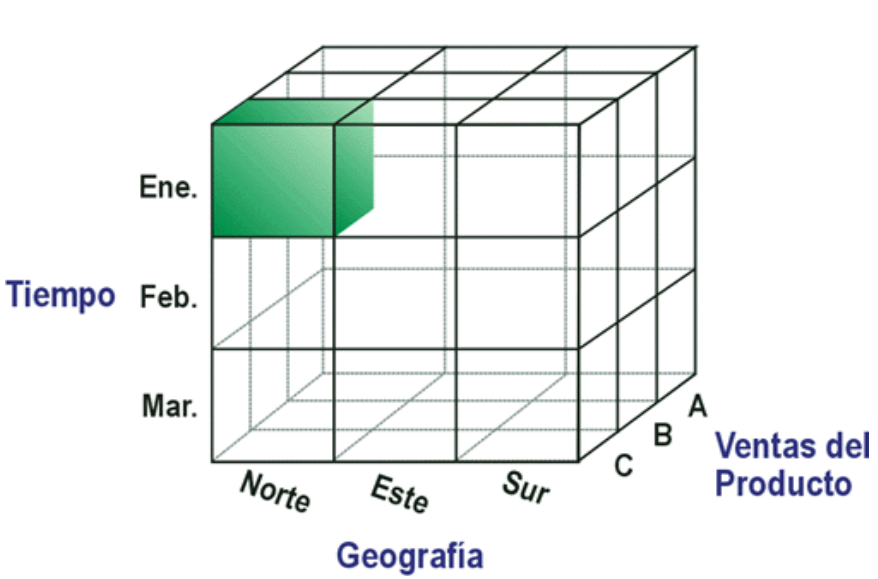
Dimensión. Contra qué quiero medir.

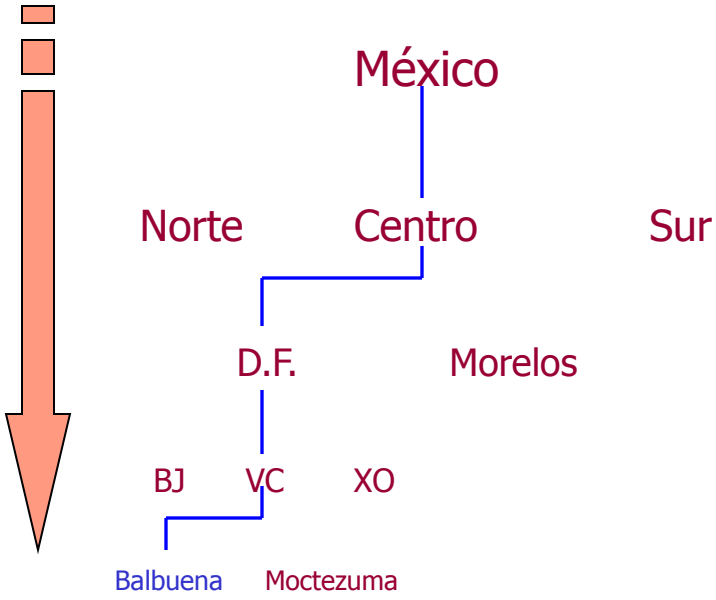


- sucursales
- zona
- clientes
- vendedores
- etc.

Operaciones con cubos OLAP

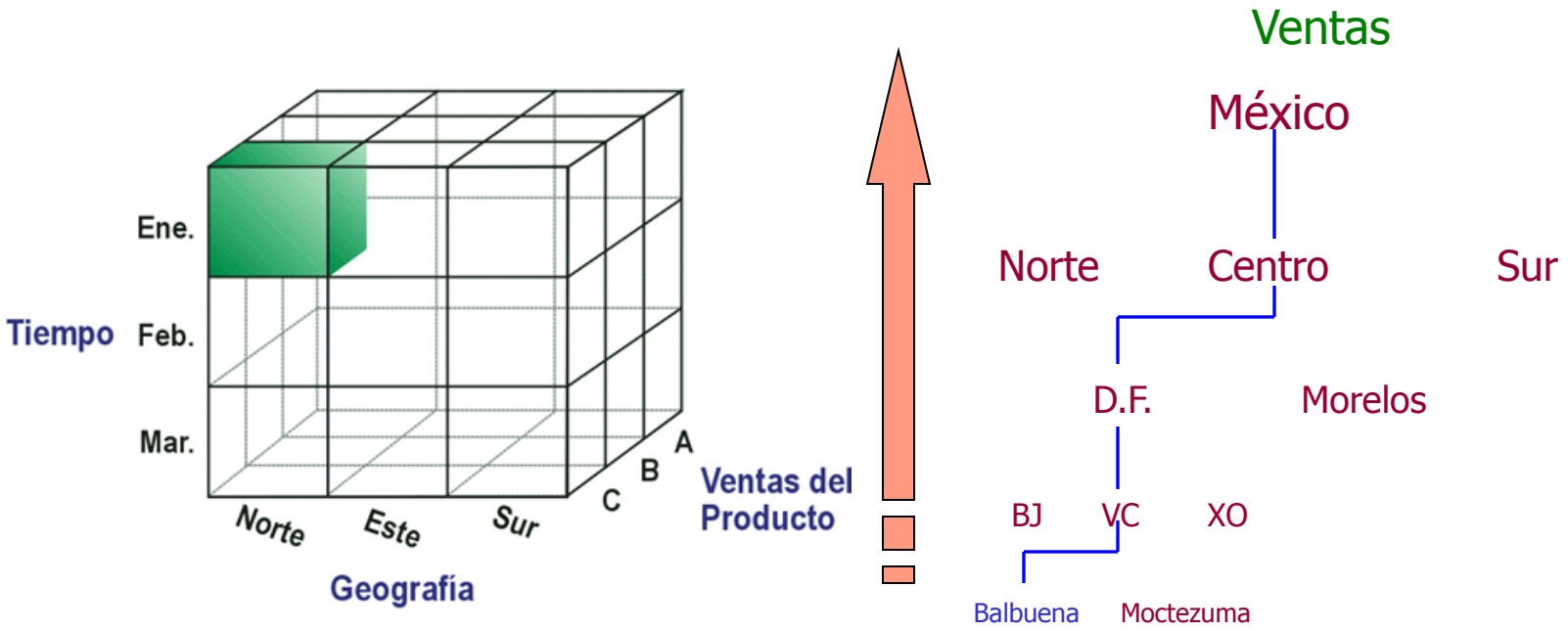
Drill Down. Desglosar una métrica de lo general a lo particular por la jerarquía de sus dimensiones





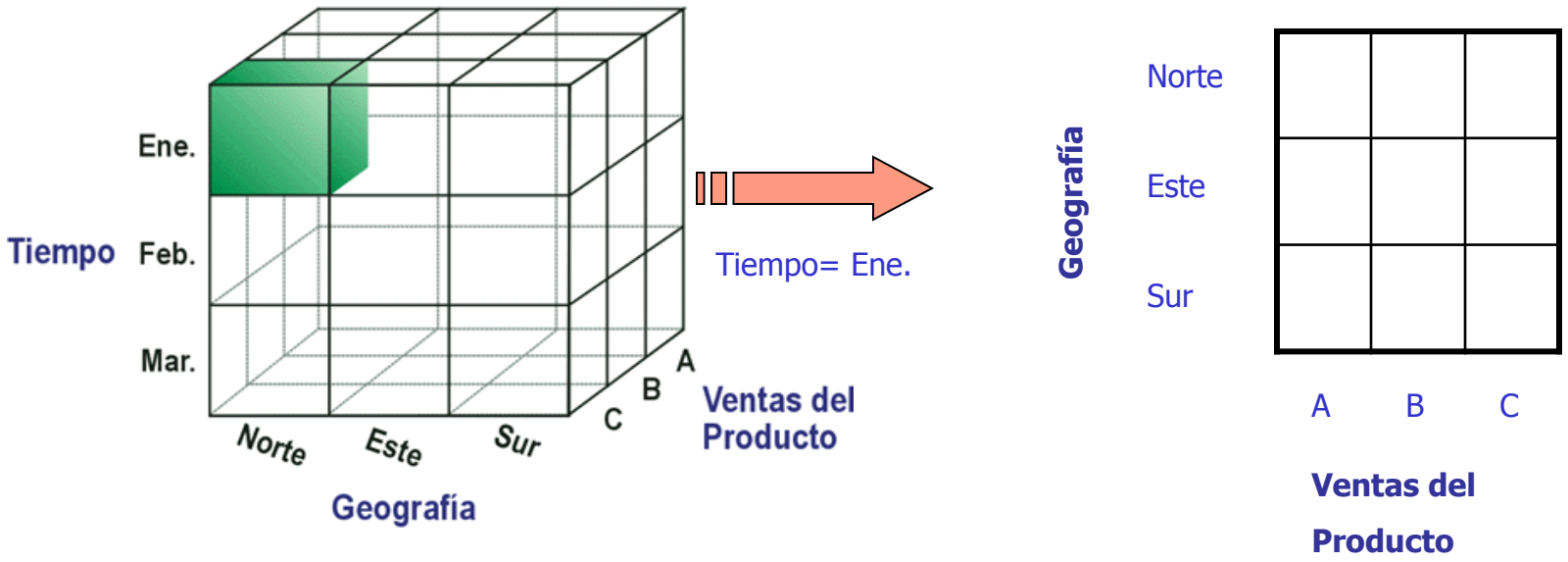
Operaciones con cubos OLAP

Drill Up. Agregar una métrica de lo particular a lo general por la jerarquía ascendente de sus dimensiones



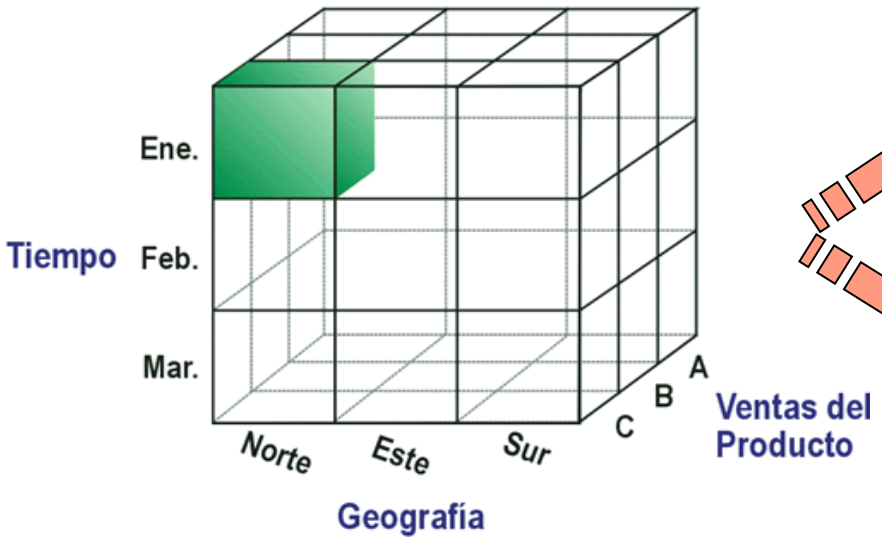
Operaciones con cubos OLAP

Slice. Obtener un sub-cubo fijando una de sus dimensiones



Operaciones con cubos OLAP

Dice. Obtener un sub-cubo fijando dos o mas de sus dimensiones



Tiempo = Ene. **or** Feb

Geografía =Norte **or** Este

Ventas del Producto = C **or** B

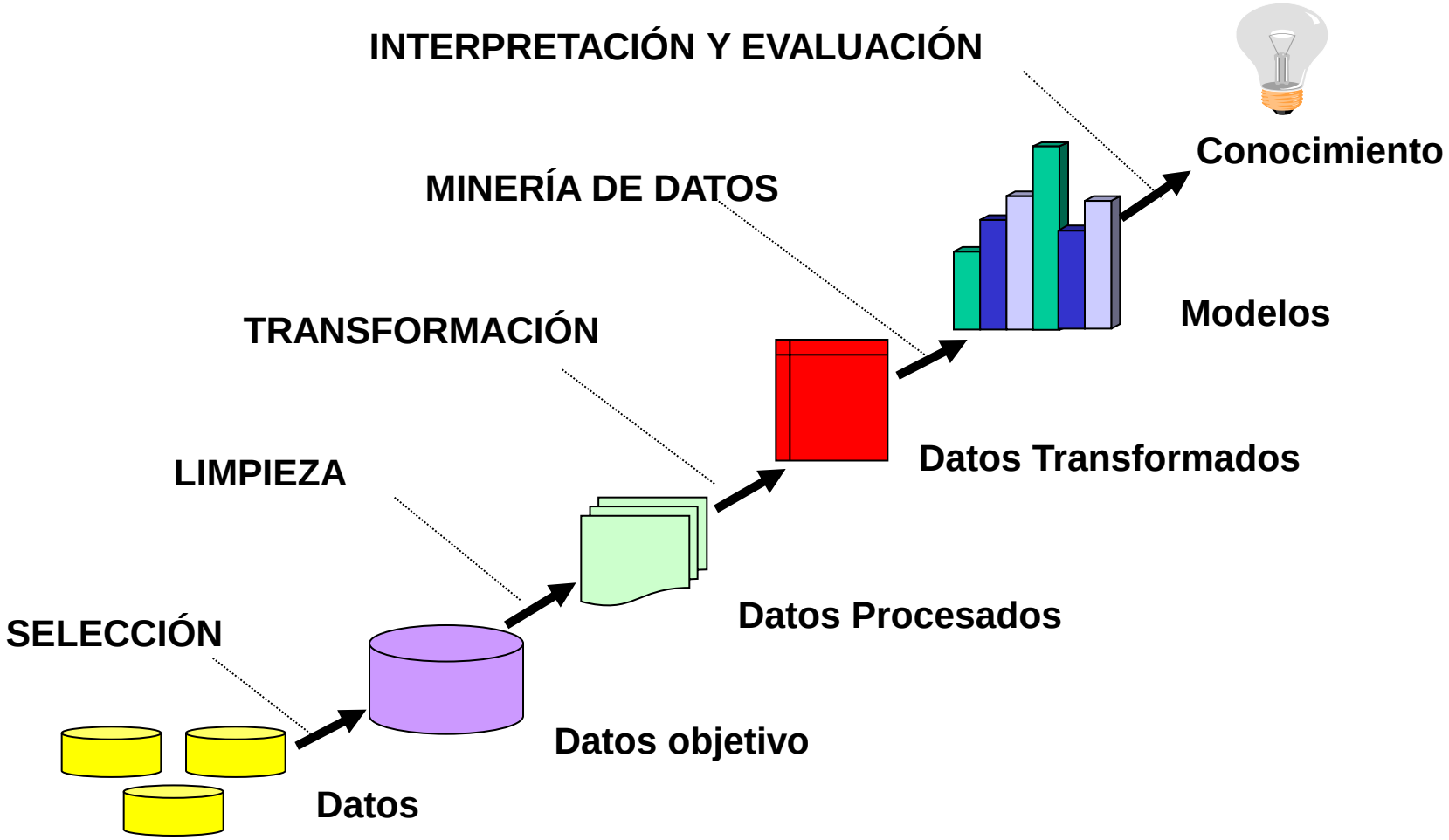
Obtenemos un cubo 2 x 2

Tiempo = Ene. **or** Feb

Geografía =Norte **or** Este

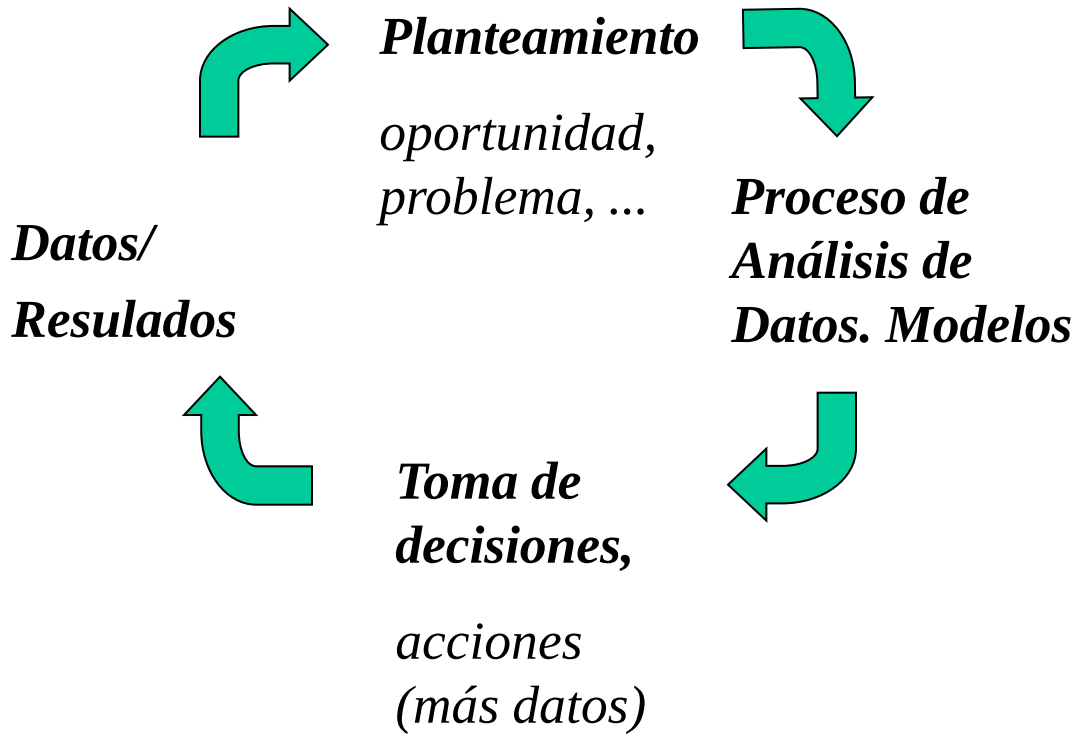
Obtenemos un cubo2 x 3

El Proceso de KDD



El Proceso de KDD

- Contexto de un aplicación con Análisis de Datos.
  - Proceso interactivo e iterativo. Ensayo y error



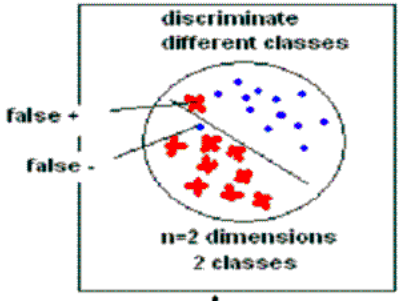
METODOLOGÍA

1. Formular el problema
2. Determinar la representación (atributos y clases)
   - directamente
   - hablando con expertos
   - a partir de otras técnicas (filtros)
3. Identificar y recolectar datos de entrenamiento (bases de datos, ficheros, …)
4. Preparar datos para análisis
5. Selección de modelo, construcción y entrenamiento
6. Evaluar lo aprendido
   - validación cruzada, expertos
7. Integrar la base de conocimiento a la espera de nuevos datos tras acciones

TÉCNICAS MD

Técnica

- Supervisada
  - Clasificación
  - Predicción



- No Supervisada
  - Agrupamiento
  - Asociación

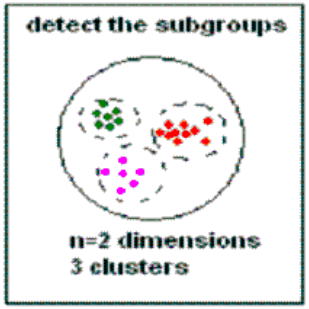


Tabla de Decisión
Árboles de Decisión
Reglas
Bayesiana
Basado en Ejemplares
Redes de Neuronas
Regresión
Árboles de Predicción
Estimador de Núcleos
Numérico K-MEDIAS
Conceptual
Probabilístico
A Priori

Tipos de técnicas

- Paramétricas, no paramétricas
- Grado de supervisión
  - Supervisadas, no supervisadas, por refuerzo
- Tipo de información resultante
  - Simbólica, subsimbólica/numérica, mixta
- Número de técnicas empleadas
  - Sencillos, meta-clasificadores
- Tipo de clases
  - Discretas, continuas, desconocidas

Estadística vs. aprendizaje automático

- Técnicas estadísticas
  - asumen una determinada estructura/ distribución
  - tienen que determinar los parámetros (técnicas paramétricas)
  - pueden funcionar con pocos datos. Verificación de hipótesis
  - basadas fundamentalmente en la estadística
- Técnicas no-paramétricas (aprendizaje automático)
  - están dirigidas por los datos. Realizan un mínimo de suposiciones
  - pueden capturar cambios estructurales en los datos
  - determinan los parámetros y además la estructura del modelo
  - admiten formulación genérica como búsqueda.
  - basadas fundamentalmente en la computación (algoritmos).

Elementos básicos de entrada

- **Concepto**: qué se quiere aprender (estructura inteligible y útil para cada tipo de problema). Salida: descripción del concepto
  - Clasificación
  - Predicción/Estimación
  - Asociación
  - Agrupamiento
- **Atributo**: qué características (variables) se van a utilizar para describir el concepto
  - Ej.: salario, crédito solicitado, categoría a la que pertenece, ...
  - Tipos: continuos, nominales/categóricos
- **Clase**: diferentes valores (etiquetas) del concepto aprendido
  - Ej.:sí, no, necesita-aval, etc.
- **Instancia** o ejemplo: cada muestra a partir de la cual se extrae el concepto

TABLA DE DATOS

Entrada: Instancias (o ejemplos):

SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
Poco	Sí	Joven	…	Uno	NO
Mucho	No	Adulto	…	Dos	SI
Mucho	No	Adulto	…	Dos	SI
Medio	Sí	Mayor	…	Tres	NO
⋮	⋮	⋮	⋱	⋮	⋮

Salida: Árboles, tablas de decisión, reglas, clusters, modelos regresión, etc.

Salida. Estructuras de conocimiento

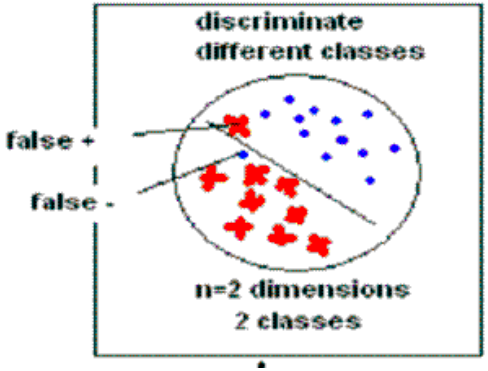
- Árboles, tablas de decisión, reglas, clusters, modelos regresión, etc.
- Instancias:

SALARIO	**CLIENTE**	**EDAD**	**…**	**HIJOS**	CRÉDITO
Poco	Sí	Joven	…	Uno	NO
Mucho	No	Adulto	…	Dos	SI
Mucho	No	Adulto	…	Dos	SI
Medio	Sí	Mayor	…	Tres	NO
⋮	⋮	⋮	⋱	⋮	⋮

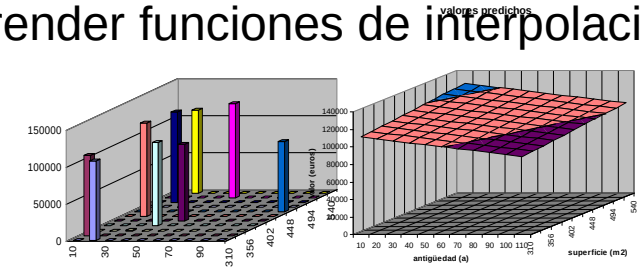
TÉCNICAS MD

Técnica

- Supervisado
  - Clasificación
    Separar instancias de cada categoría
    (aprender fronteras de clases)



  - Predicción
    Predecir valores numéricos
    (aprender funciones de interpolación)



- No Supervisado

Clasificadores bayesianos

- Tablas de prob. condicionales

SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
Poco	Sí	Joven	…	Uno	**NO**
Mucho	No	Adulto	…	Dos	**SI**
Mucho	No	Adulto	…	Dos	**SI**
Medio	Sí	Mayor	…	Tres	**NO**
⋮	⋮	⋮	⋱	⋮	⋮

Salario \ Crédito	No	Sí
Poco	0.4	0.2
Mucho	0.3	0.6
Medio	0.3	0.2

Cliente \ Crédito	No	Sí
Sí	0.4	0.3
No	0.6	0.7

Hijos \ Crédito	No	Sí
Uno	0.2	0.1
Dos	0.3	0.0
Tres	0.5	0.9

Edad \ Crédito	No	Sí
Joven	0.4	0.2
Adulto	0.4	0.6
Mayor	0.2	0.2

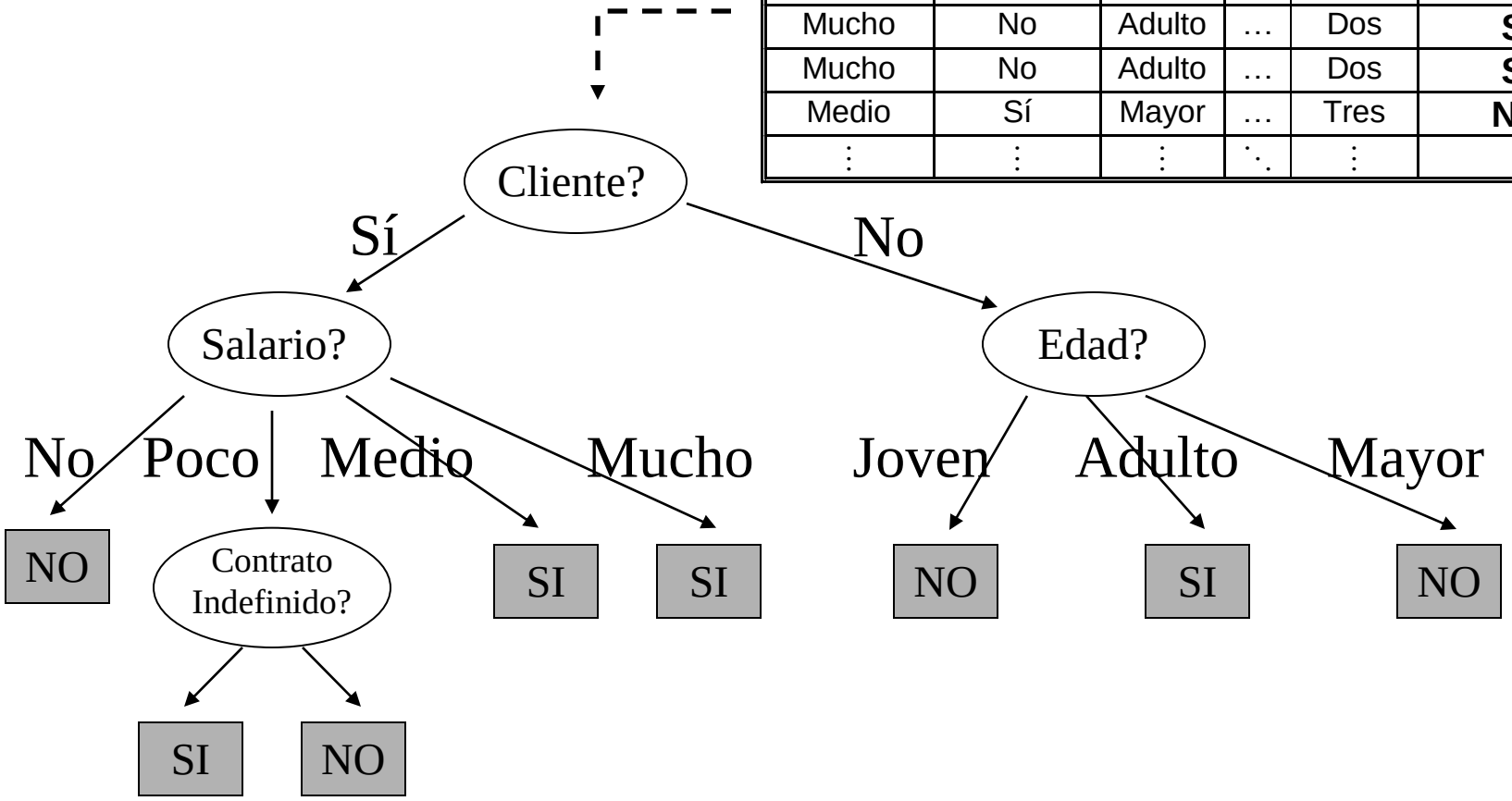
. . .

p(SI) = 0.63

p(NO) = 0.37

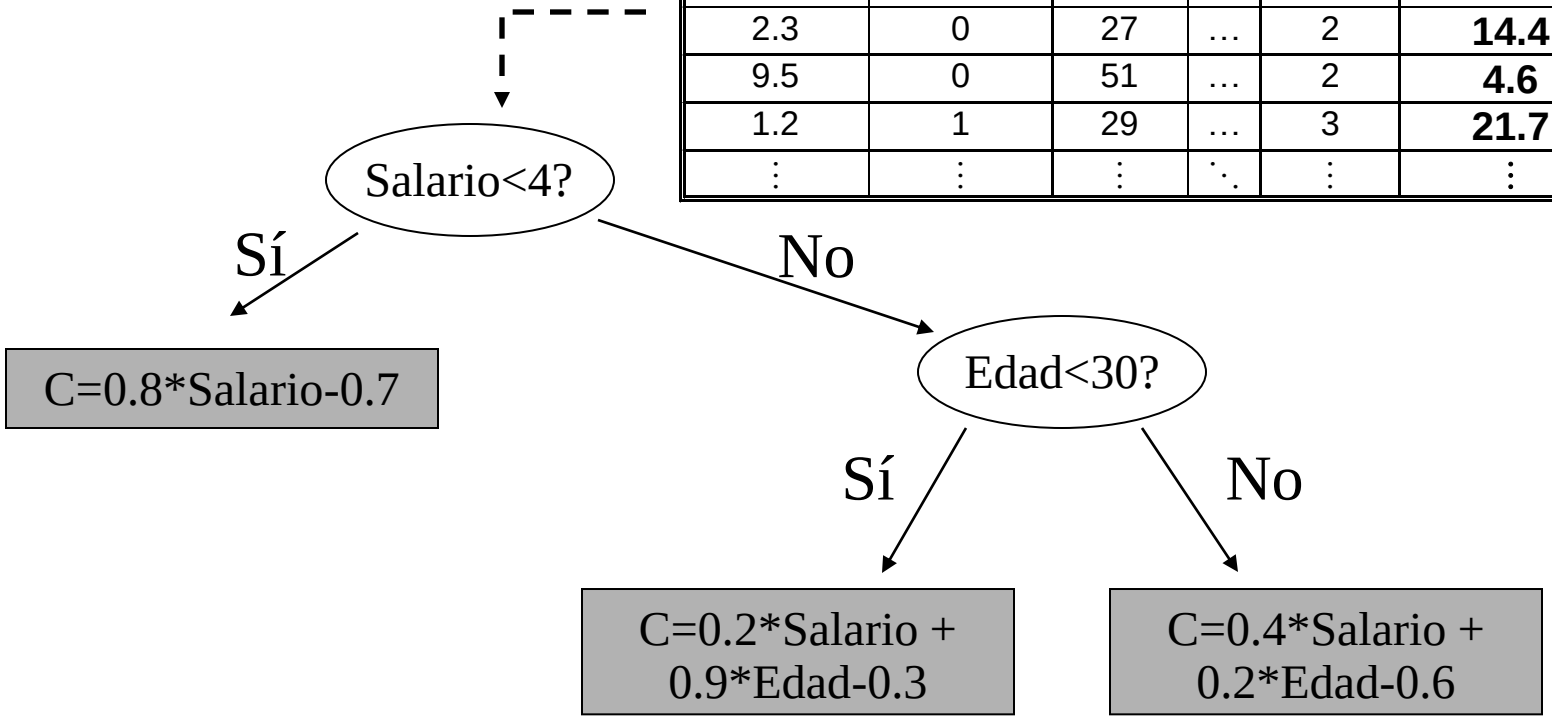
Árboles de decisión

SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
Poco	Sí	Joven	…	Uno	**NO**
Mucho	No	Adulto	…	Dos	**SI**
Mucho	No	Adulto	…	Dos	**SI**
Medio	Sí	Mayor	…	Tres	**NO**
⋮	⋮	⋮	⋱	⋮	⋮



Árboles de regresión

SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
4.5	1	34	…	1	**12.3**
2.3	0	27	…	2	**14.4**
9.5	0	51	…	2	**4.6**
1.2	1	29	…	3	**21.7**
⋮	⋮	⋮	⋱	⋮	⋮



Reglas de decisión

SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
Poco	Sí	Joven	…	Uno	NO
Mucho	No	Adulto	…	Dos	SI
Mucho	No	Adulto	…	Dos	SI
Medio	Sí	Mayor	…	Tres	NO
⋮	⋮	⋮	⋱	⋮	⋮

Si (Cliente=No) Y (Edad=Joven) Entonces NO
Si (Cliente=No) Y (Edad=Adulto) Entonces SI
Si (Cliente=Sí) Y (Salario=Ninguno) Entonces NO
Si (Cliente=Si) Y (Salario=Medio) Entonces SI
…

- Reglas proposicionales y relacionales
- Convertibiliad de árboles en reglas
- Tablas secuenciales y no secuenciales

Aprendizaje vago

- Se almacena toda la tabla de instancias:

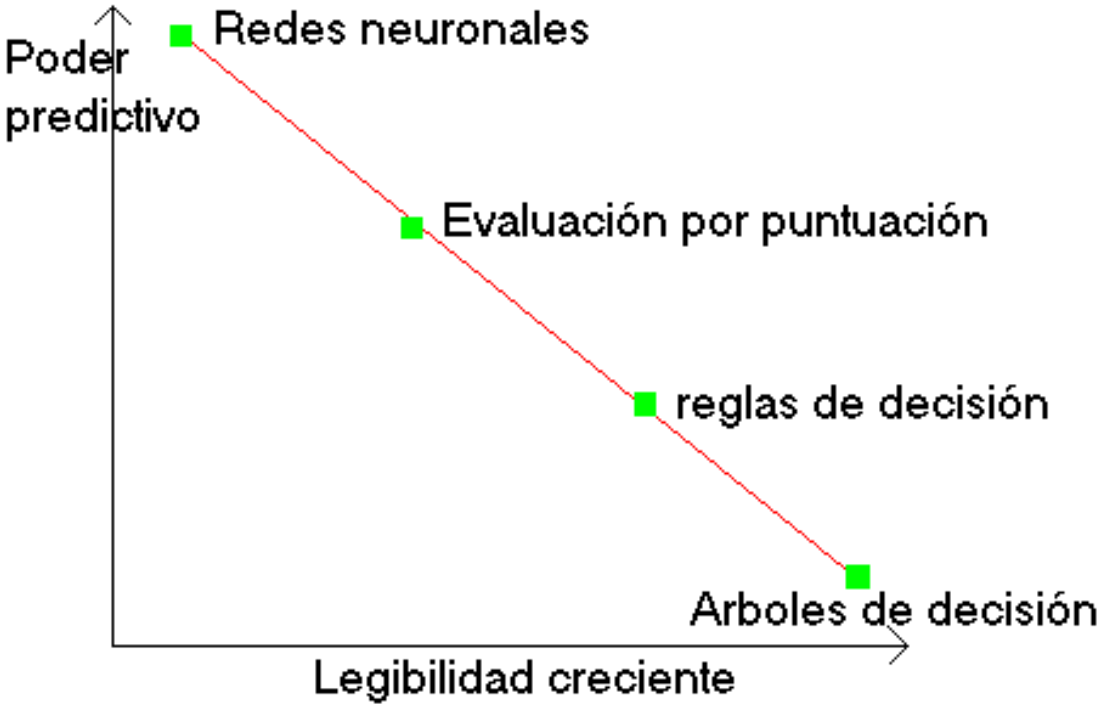
SALARIO	CLIENTE	EDAD	…	HIJOS
Medio	Sí	Adulto	…	Cuatro

SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
Poco	Sí	Joven	…	Uno	NO
Mucho	No	Adulto	…	Dos	SI
Mucho	No	Adulto	…	Dos	SI
Medio	Sí	Mayor	…	Tres	NO
⋮	⋮	⋮	⋱	⋮	⋮

NO

  - Caso particular: *tabla de decisión*
  - Poca eficiencia de cómputo
  - Ausencia de análisis de información

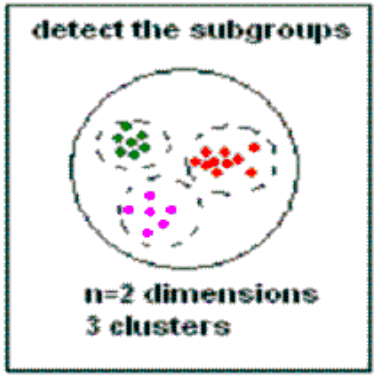
Propiedades (discutible)



TÉCNICAS MD

Técnica

- Supervisado
- No Supervisado
  - **Asociación**
    Relaciones entre atributos (columnas)
  - **Agrupamiento**
    relaciones entre objetos (filas)

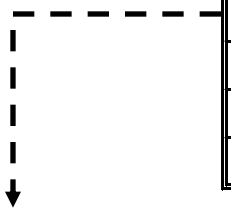


REGLAS DE ASOCIACIÓN

Búsqueda general de relaciones significativas entre atributos

No supervisado

Mayor número de grados de libertad para búsqueda (subconjuntos)



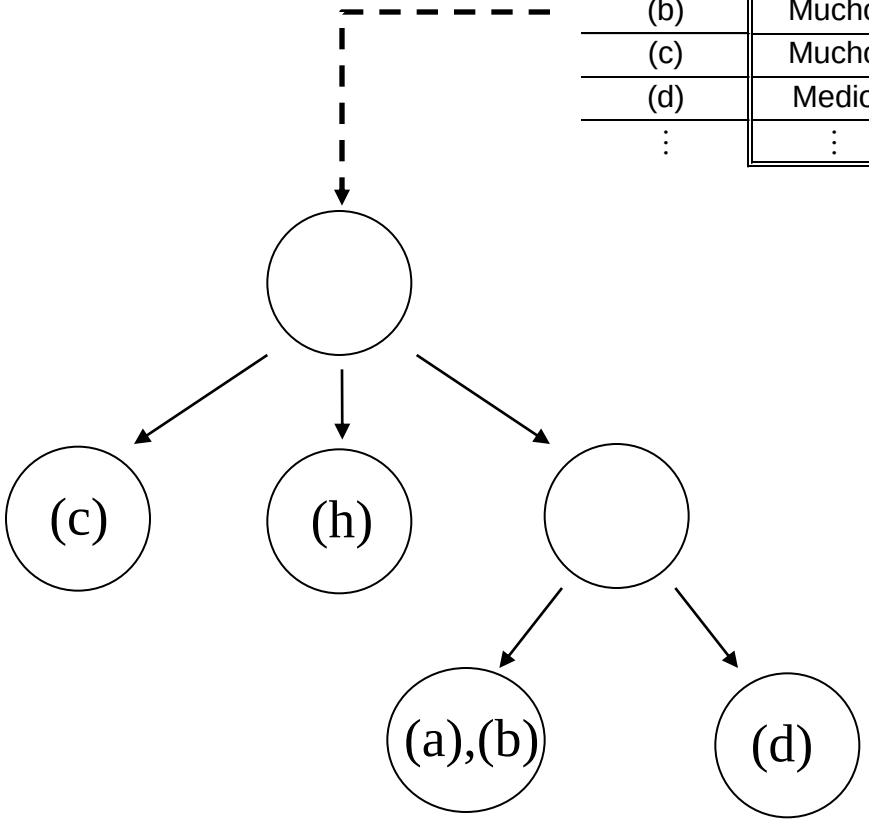
SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
Poco	Sí	Joven	…	Uno	NO
Mucho	No	Adulto	…	Dos	SI
Mucho	No	Adulto	…	Dos	SI
Medio	Sí	Mayor	…	Tres	NO
⋮	⋮	⋮	⋱	⋮	⋮

Si (Edad=Joven) Entonces Hijos < Tres
Si (Residencia=X) Entonces (Salario=Medio) Y (Hijos>2)
Si (Salario=NO) Entonces (Salario=Bajo) O (Salario=Mucho)
Si (Crédito=Si) Entonces (Cliente=Si)
…

Algoritmo
“A priori”

Agrupación (clustering)

	SALARIO	CLIENTE	EDAD	…	HIJOS	CRÉDITO
(a)	Poco	Sí	Joven	…	Uno	NO
(b)	Mucho	No	Adulto	…	Dos	SI
(c)	Mucho	No	Adulto	…	Dos	SI
(d)	Medio	Sí	Mayor	…	Tres	NO
⋮	⋮	⋮	⋮	⋱	⋮	⋮



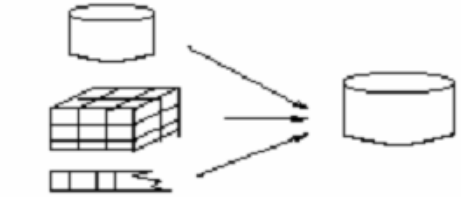
PREPARACIÓN DE DATOS

- La preparación de datos puede suponer 60-90% del tiempo
  - objetivo: única tabla de datos (instancias, atributos)
  - ensamblar, integrar formatos, agregar, …

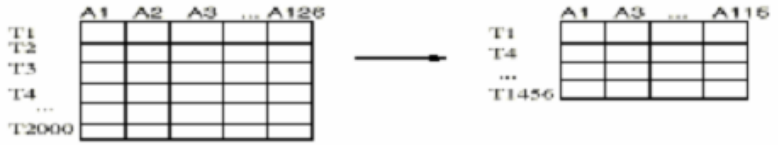
- **Filas**: Agregación: selección de dato unitario
  - asociar datos, calcular resúmenes, …
- **Columnas**: Selección de atributos
  - eliminar campos redundantes o inapropiados (ID)
  - crear atributos de interés de campos textuales
    - fecha/hora->edad, estación, vacaciones, mañana/tarde/noche,
    - dirección/código postal-> lugar geográfico, área, ciudad
  - transformar atributos

PREPARACIÓN DE DATOS

- Integración
  - Múltiples fuentes (bases de datos, ficheros, …)

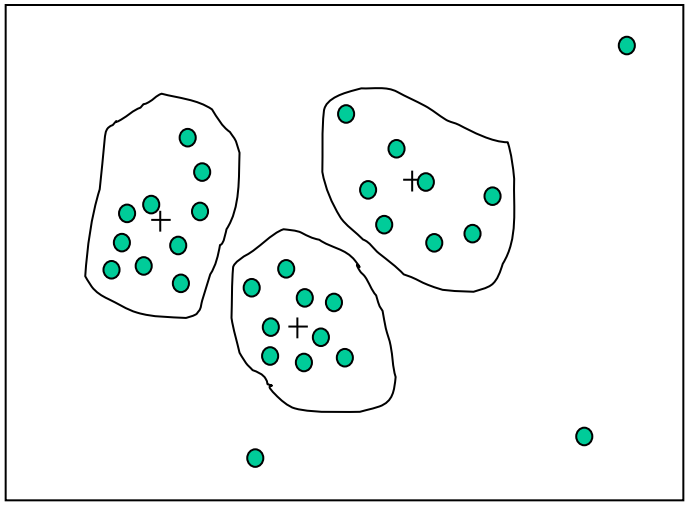


- Limpieza
  - Valores faltantes, outliers, inconsistencias
- Transformación
  - Normalizar, proyectar, discretizar
- Reducción
  - Representación reducida

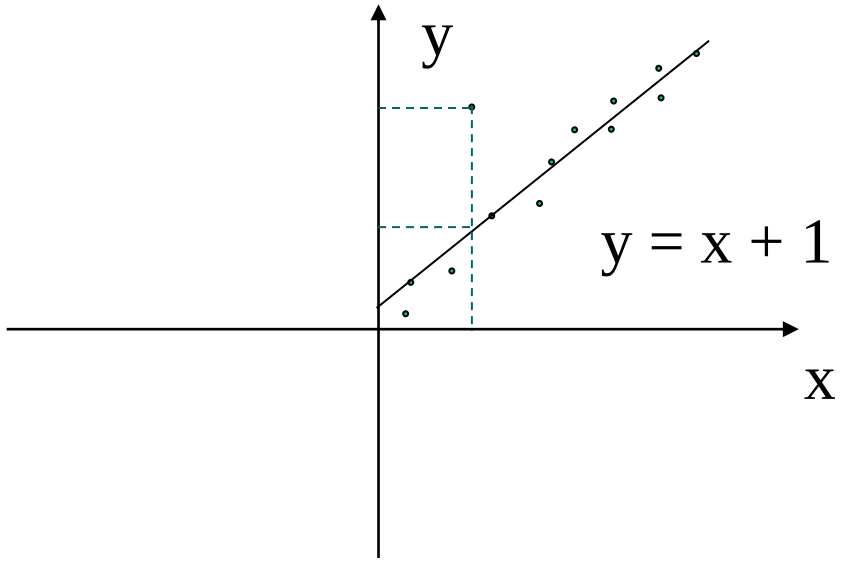


LIMPIEZA: DATOS RUIDOSOS

- Outliers : detección y eliminación (o tratamiento individual)
  - Analisis de clusters



Regresión



LIMPIEZA: DATOS INCOMPLETOS, REDUNDANTES

- Datos Redundantes: detectar relaciones causales o funcionales en datos
  - Frecuente al integrar múltiples bases de datos
    - El mismo atributo con diferente nombre
    - Relaciones directas: atributo “calculado”
  - Se detecta con análisis de correlación

- Datos incompletos: faltan atributos en algunos ejemplos
  - Relleno manual: tedioso o no posible
  - Ignorar ejemplo: cuando son pocos casos
  - Nuevo valor “desconocido”
  - Valor medio, valor más probable según resto, …

TRANSFORMACION DE DATOS

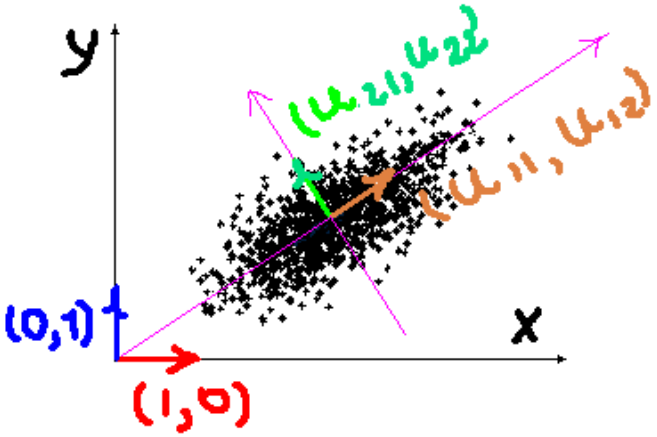
- Agregación: resumen, cubos de datos
- Normalización: re-escalar las variables (para distancias)
  - normalización min-max $v' = \frac{v - min_A}{max_A - min_A}$
  - normalización estadística (tipificar) $v' = \frac{v - media_A}{stand\_dev_A}$
- Selección/transformación de atributos
  - Discretizar
  - Quitar atributos redundantes
  - Proyectar a espacios reducidos: PCA

TRANSFORMACIÓN: DISCRETIZACION

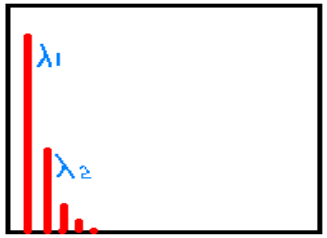
- Discretización
  - Reducir número de valores, o poner intervalos a variables continuas.
    Ej.: edad->(joven, adulto, mayor)
  - Reduce el tamaño de datos y mejora precisión
  - Misma amplitud:
    - Cajas: *W* = (Max-min)/*N.*
    - El más directo. Problemas de escala y con outliers
  - Misma frecuencia:
    - Cada caja el mismo numero de muestras
  - Métodos supervisados

TRANSFORMACIÓN: PROYECCIÓN PCA

- Dados vectores *k*-dimensionales, buscar vectores ortogonales de dimensión *c* <= *k*
- Aplicable a datos numéricos de muchas dimensiones

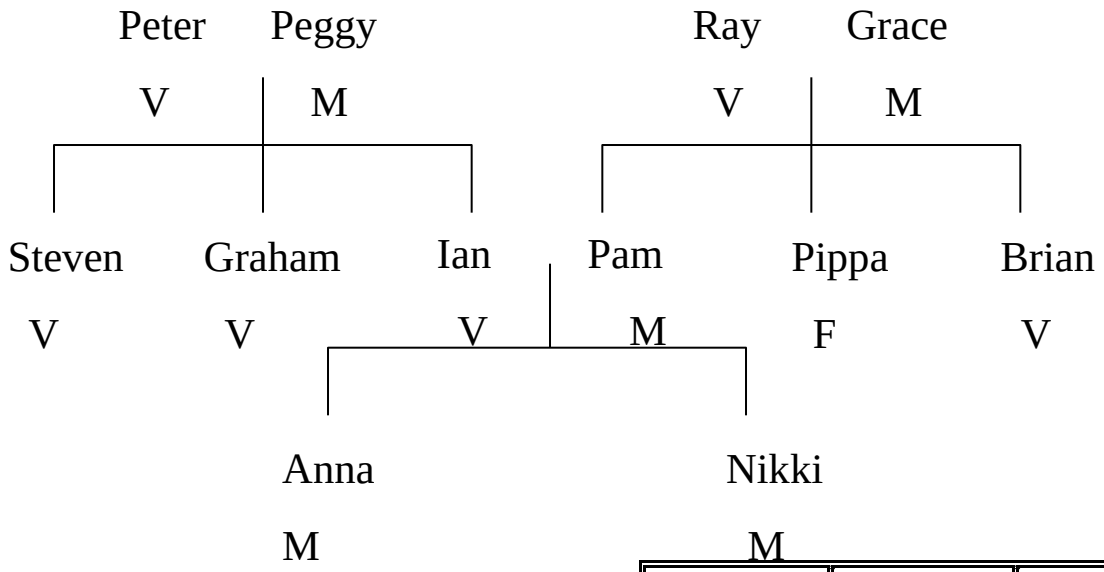


$$C = \frac{1}{n} X^t X = U \Lambda U^t = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix}$$



Otros aspectos de preparación

- Formulación problema, atributos, representación…



persona1	persona2	hermana?
Peter	Peggy	**No**
Peter	Steven	**No**
⋮	⋮	⋮
Brian	Pam	**Sí**
⋮	⋮	⋮

persona1	persona2	abuelo?
Peter	Peggy	**No**
Peter	Steven	**No**
⋮	⋮	⋮
Anna	Peter	**Sí**
⋮	⋮	⋮

Desnormalización de relaciones

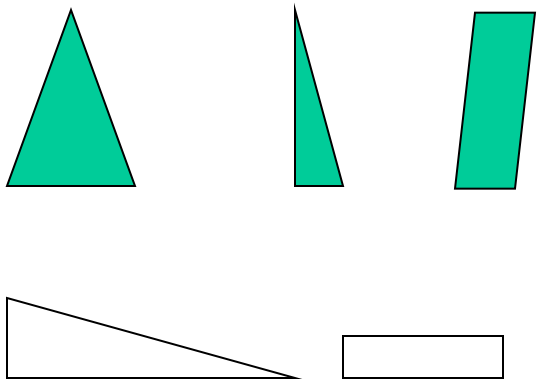
persona1	género	padre	madre	persona2	género	padre	madre	hermana?
Brian	V	Ray	Grace	Pam	**M**	**Ray**	**Grace**	**Sí**
Pippa	M	Ray	Grace	Pam	**M**	**Ray**	**Grace**	**Sí**
⋮				⋮				⋮
Nikki	M	Ian	Pam	Anna	**M**	**Ian**	**Pam**	**Sí**
Resto								**No**

Si (género persona2=M) Y (padre persona 1=padre persona 2) Entonces Si
Si (género persona2=M) Y (madre persona 1=madre persona 2) Entonces Si
Resto: NO

- Abuelo? Más atributos
- Descendiente? Programación lógica inductiva

Reglas proposiocionales y relacionales

- A veces se precisan reglas con relaciones entre atributos para aprender un tipo de problema



Proposicional

Si (ancho>=4) y (alto<3) =>tumbado

Si (alto>6) =>levantado

…

Relacional

Si (ancho>alto) =>tumbado

Si (ancho<alto) =>levantado

Problemas técnicos

- Relaciones espúreas, casuales o coincidencias
- Datos incorrectos o incompletos
- Datos distribuidos
- Conceptos que cambian con el tiempo
- Datos que llegan en el tiempo
- Datos o formulación del problema sesgados
- Conceptos en varios niveles de generalización
- Integración de información cualitativa y cuantitativa

Problemas no técnicos

- Las máquinas no son responsables de las predicciones o clasificaciones
- Los humanos no se fían de los resultados
- La salida no es entendible por el humano
- Hay que proporcionar al sistema los valores de los atributos
- Miedo de los humanos a la pérdida de control
- Cuestiones legales y éticas: privacidad, discriminación, etc.