

Big Data y Minería de Datos

Unidad 1. Introducción a la minería de datos

Autores: José Antonio Ruipérez Valiente (jruiperez@um.es), Mariano Albaladejo González (mariano.albaladejog@um.es) y Manuel Jesús Gómez Moratilla (manueljesus.gomezm@um.es)

UNIVERSIDAD DE
MURCIA

Unidad 1. Introducción a la minería de datos

Estructura

- El nacimiento de la minería de datos
- ¿Qué es la minería de datos?
- El ciclo de la minería de datos
- Principales desafíos de la minería



Introducción a la minería de datos

Técnicas de inteligencia artificial

Reglas

Fundamentos lógicos
Sistemas basados en reglas
Reglas de asociación

Aprendizaje computacional

Fundamentos y evaluación
Regresión y clasificación
No supervisado

Procesado del
lenguaje natural
Fundamentos
Representación y
tópicos
Sentimientos y NER

Introducción al big data

Unidad 1. Introducción a la minería de datos

EL NACIMIENTO DE LA MINERÍA DE DATOS



Unidad 1. Introducción a la minería de datos

Nuevas necesidades

- “We are living in the ~~information~~ data age”
- Crecimiento explosivo del volumen de datos por la digitalización de servicios y productos
 - Walmart y cientos de millones de transacciones por semana
 - Pero también otras fuentes como observaciones espaciales con sensores, flujos de cámara y audio, comunidades online...
- Necesidad de explotar esos datos, tanto a nivel de negocio, como por razones públicas o científicas, por ejemplo
- El objetivo es transformar colecciones de datos en conocimiento
 - Por ejemplo, el buscador de Google y la gripe

Unidad 1. Introducción a la minería de datos

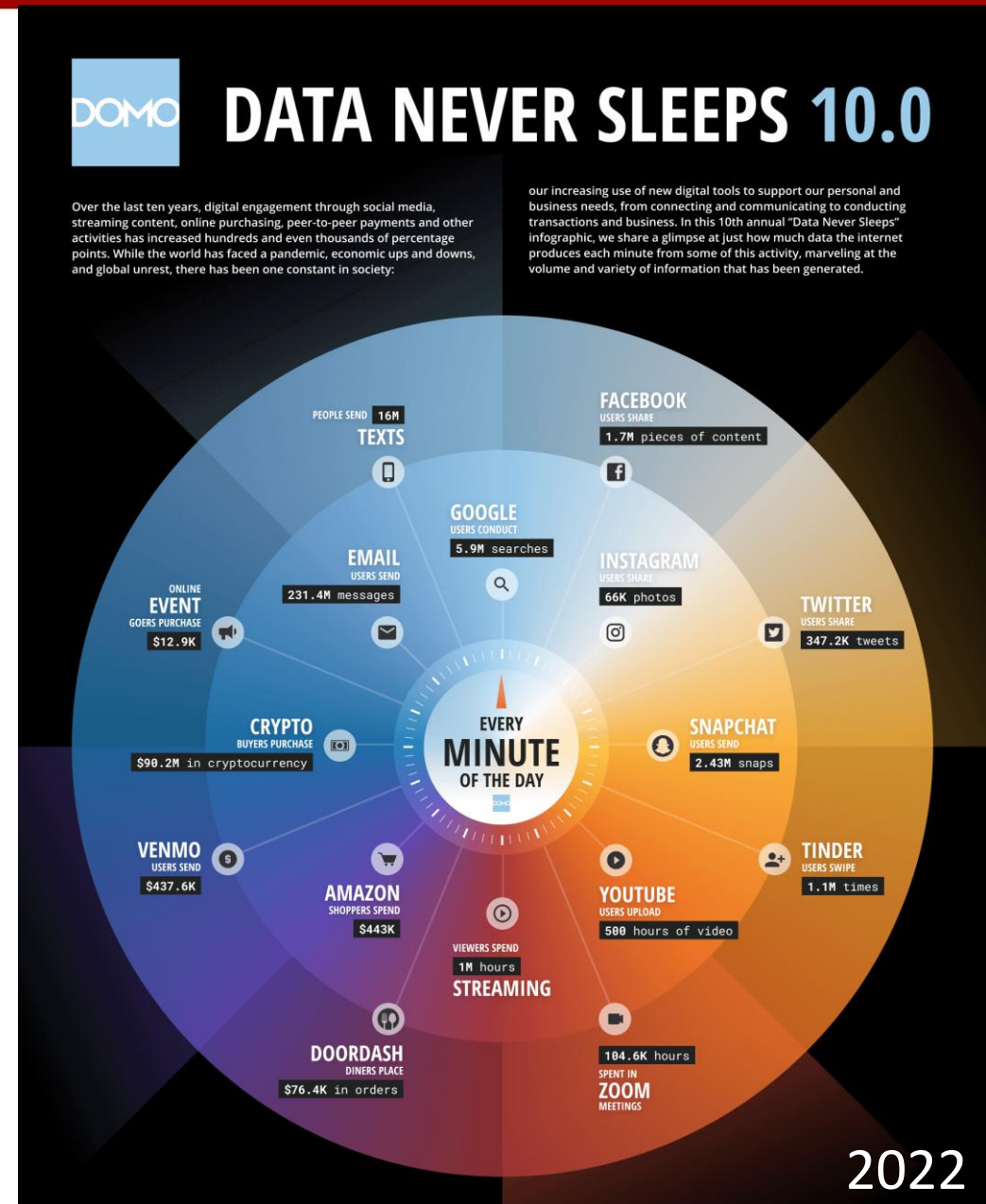
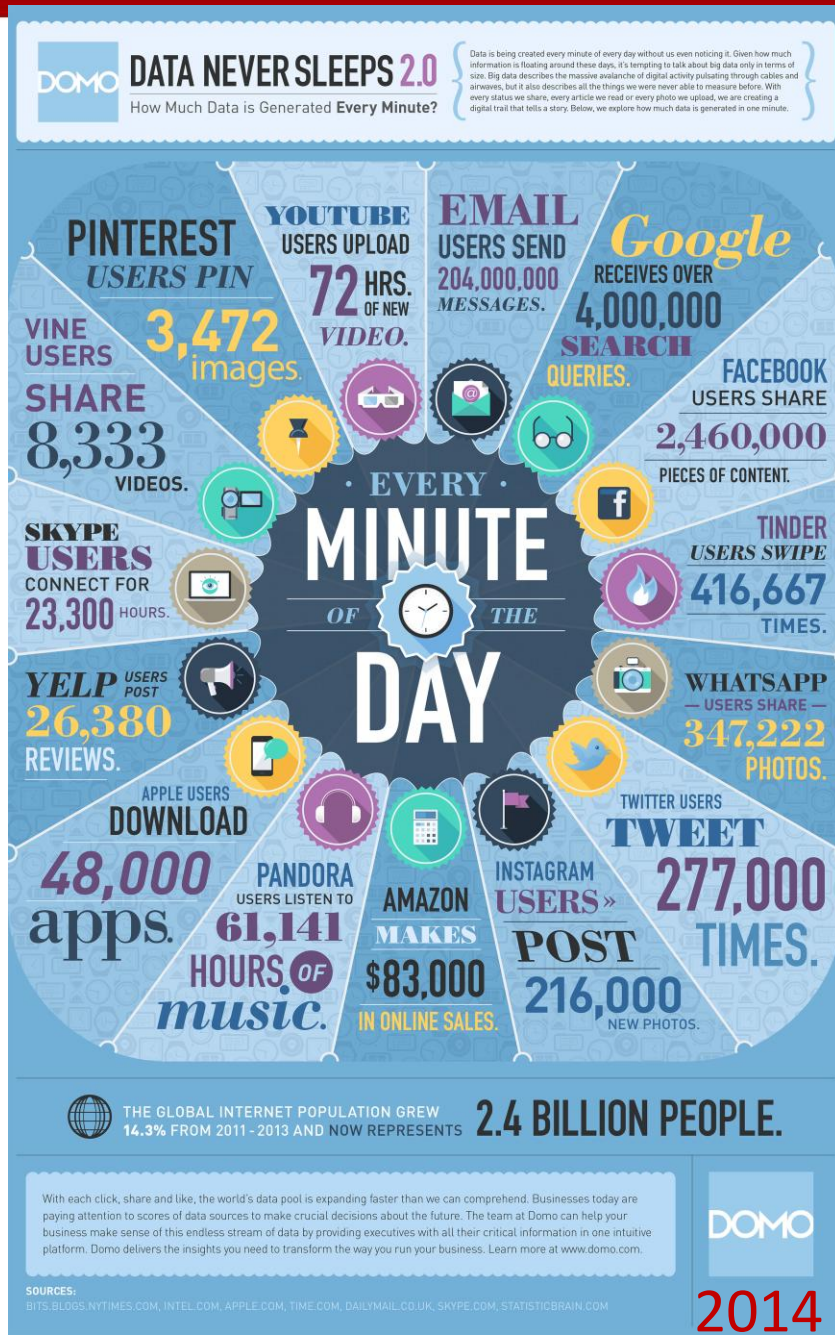
Cantidades de datos

UNIDAD	ÓRDENES DE MAGNITUD
bit (b)	0 o 1, sí o no
Byte (B)	Una letra: a, b, c, d, ...
Kilobyte (kB) = 1024 bytes	
Megabyte (MB) = 1024^2 bytes $\cong$ 1 millón bytes	5 MB = una canción (formato mp3)
Gigabyte (GB) = 1024^3 bytes $\cong$ 1 mil millones bytes	6.4 GB = genoma humano (2 bits par)
Terabyte (TB) = 1024^4 bytes $\cong$ 1 billón bytes	1.9 TB = Wikipedia Inglés (2012)
Petabyte (PB) = 1024^5 bytes $\cong$ 1 mil billones bytes	300 PB = almacén Facebook (2013)
Exabyte (EB) = 1024^6 bytes $\cong$ 1 trillón bytes	1.2 EB = almacén Google (2013)
Zettabyte (ZB) = 1024^7 bytes	4.8 ZB = todos HD vendidos (2016)
Yottabyte (YB) = 1024^8 bytes	
Ronnabyte, Quettabyte...	

Unidad 1. Introducción a la minería de datos

La velocidad de los datos

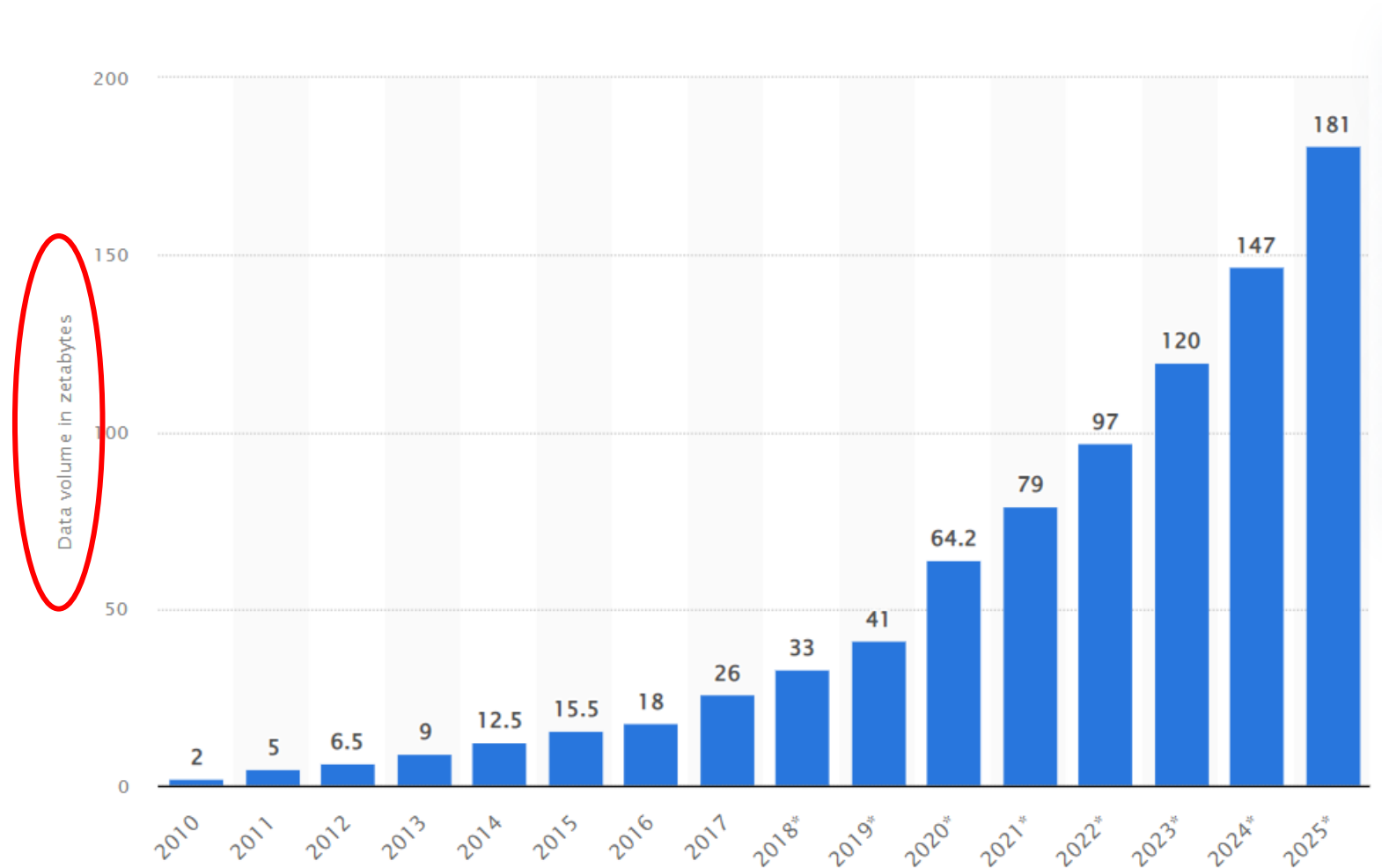
UNIVERSIDAD DE
MURCIA



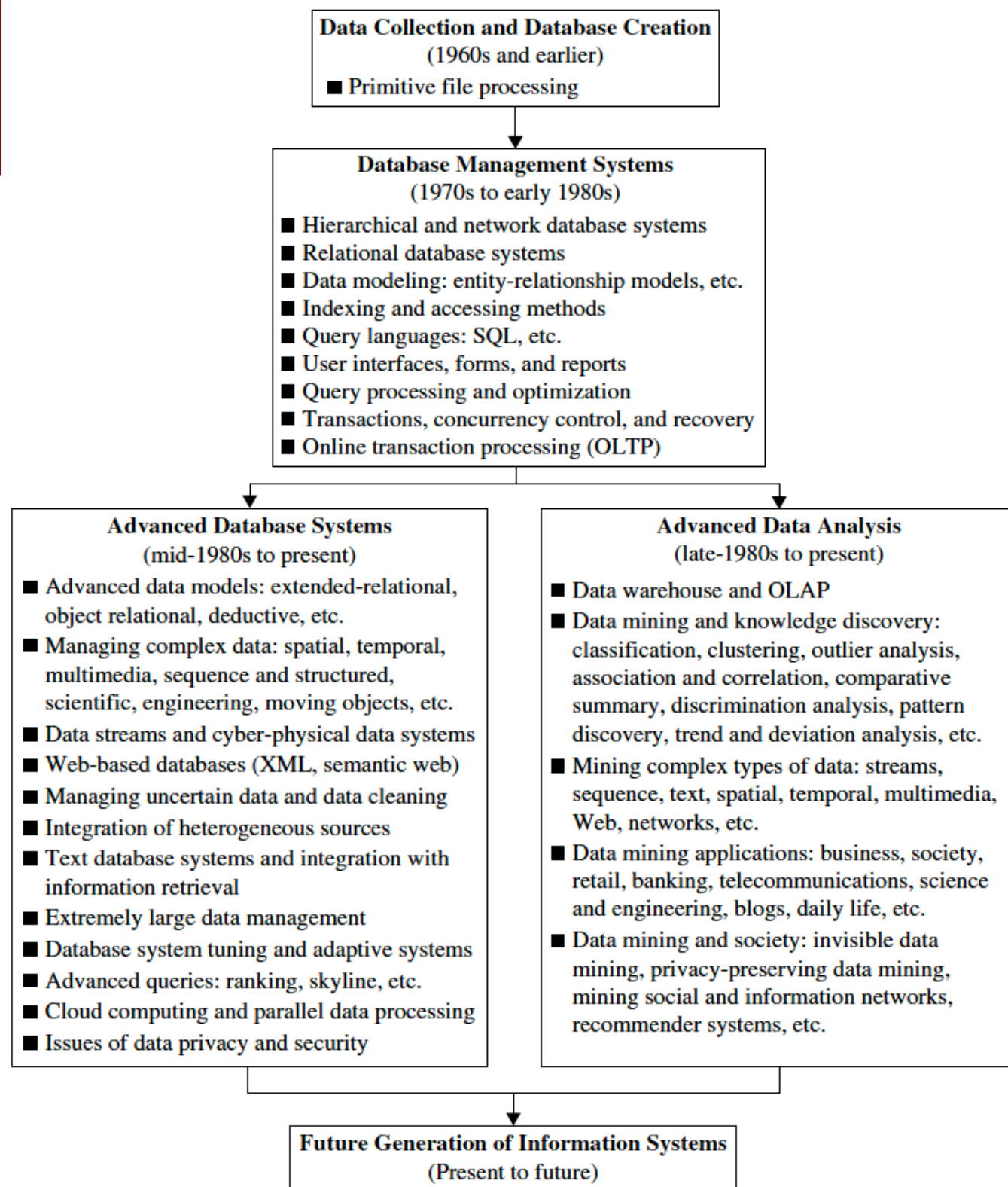
Unidad 1. Introducción a la minería de datos

Nuevas necesidades

- Estimación de volumen de datos creados, guardados, copiados y consumidos de 2010 a 2020, predicciones para 2021 al 2025.



Unidad 1. Introducción Evolución



Unidad 1. Introducción a la minería de datos

¿QUÉ ES LA MINERÍA DE DATOS?



Unidad 1. Introducción a la minería de datos

El concepto de la minería de datos

- [Witten & Frank 2000] se define la minería de datos como “*el proceso de **extraer conocimiento útil y comprensible**, previamente desconocido, desde grandes cantidades de **datos** almacenados en **distintos formatos**.*”
- La tarea principal es encontrar modelos inteligibles:
 - Debería ser un proceso automático o semi-automático (asistido)
 - Los patrones descubiertos tendrían que tener utilidad
- Por lo tanto, dos de los desafíos principales será:
 - Trabajar con grandes volúmenes de datos (ruido, datos ausentes, intratabilidad, volatilidad...)
 - Usar las técnicas adecuadas para analizarlos y extraer info
- Utilidad del conocimiento y cercanía al dominio
- El usuario final no tiene por qué ser un experto en minería de datos (interpretabilidad y comunicación información)

Unidad 1. Introducción a la minería de datos

Ejemplo 1: análisis de créditos bancarios

- Un banco por Internet desea obtener reglas para predecir qué personas que solicitan un crédito no lo devuelven
- ¿Qué tipo de datos necesitaríamos?

IDC	D-crédito (años)	C-crédito (euros)	Salario (euros)	Casa propia	Cuentas morosas	...	Devuelve- crédito
101	15	60.000	2.200	sí	2	...	no
102	2	30.000	3.500	sí	0	...	sí
103	9	9.000	1.700	sí	1	...	no
104	15	18.000	1.900	no	0	...	sí
105	10	24.000	2.100	no	0	...	no
...	...	...	...	...	...	...	...

- Aplicando minería de datos para obtener reglas...

SI Cuentas-Morosas > 0 **ENTONCES** Devuelve-crédito = no

SI Cuentas-Morosas = 0 **Y** [(Salario > 2.500) **O** (D-crédito > 10)] **ENTONCES**
Devuelve-crédito = sí

- El banco debe determinar acciones en caso de no

Unidad 1. Introducción a la minería de datos

Ejemplo 2: análisis de la cesta de la compra

- Un supermercado quiere obtener información del comportamiento de compra
 - Mejorar servicios, reubicación de productos, localizar buen emplazamiento
- ¿Qué tipo de información necesitamos?

Idcesta	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Lechugas	...
1	sí	no	no	sí	no	sí	sí	sí	...
2	no	sí	no	no	sí	no	no	sí	...
3	no	no	sí	no	sí	no	no	no	...
4	no	sí	sí	no	sí	no	no	no	...
5	sí	sí	no	no	no	sí	no	sí	...
6	sí	no	no	sí	sí	sí	sí	no	...
7	no	no	no	no	no	no	no	no	...
8	sí	sí	sí	sí	sí	sí	sí	no	...
...	...	...	...	...	...	...	...	...	...

- ¿Podrías encontrar patrones?
- El 100% de las veces que se compran pañales, también leche
- También la frecuencia de las asociaciones, ¿es frecuente?

Unidad 1. Introducción a la minería de datos

Ejemplo 3: determinar las ventas de un producto

- Cadena de electrodomésticos quiere optimizar el mantenimiento del stock, para servirlo rápidamente
- ¿Qué tipo de información necesitamos?

Producto	mes-12	...	mes-4	mes-3	mes-2	mes-1
televisor plano 30' Phlipis	20	...	52	14	139	74
vídeo-dvd-recorder Miesens	11	...	43	32	26	59
discman mp3 LJ	50	...	61	14	5	28
frigorífico no frost Jazzussi	3	...	21	27	1	49
microondas con grill Sanson	14	...	27	2	25	12
...	...	...	...	...	...	...

- Esta información permite generar un modelo para predecir las ventas en base a los meses anteriores
- En función de las ventas y el stock actual, saben si deben pedir más o no

Unidad 1. Introducción a la minería de datos

Ejemplo 4: grupos diferenciados de empleados

- El departamento de RRHH desea categorizar empleados en distintos grupos para adaptar sus políticas
- ¿Qué información necesitamos?

Id	Sueldo	Casado	Coche	Hijos	Alq/prop	Sindicado	Bajas/año	Antigüedad	Sexo
1	1.000	Sí	No	0	Alquiler	No	7	15	H
2	2.000	No	Sí	1	Alquiler	Sí	3	3	M
3	1.500	Sí	Sí	2	Prop	Sí	5	10	H
4	3.000	Sí	Sí	1	Alquiler	No	15	7	M
5	1.000	Sí	Sí	0	Prop	Sí	1	6	H
6	4.000	No	Sí	0	Alquiler	Sí	3	16	M
7	2.500	No	No	0	Alquiler	Sí	0	8	H
8	2.000	No	Sí	0	Prop	Sí	2	6	M
9	2.000	Sí	Sí	3	Prop	No	7	5	H
10	3.000	Sí	Sí	2	Prop	No	1	20	H
11	5.000	No	No	0	Alquiler	No	2	12	M
12	800	Sí	Sí	2	Prop	No	3	1	H
13	2.000	No	No	0	Alquiler	No	27	5	M
14	1.000	No	Sí	0	Alquiler	Sí	0	7	H
15	8 00	No	Sí	0	Alquiler	No	3	2	H
...	...	...	...	...	...	...	...	...	...

- Se puede construir un sistema de minería para agruparlos

Unidad 1. Introducción a la minería de datos

Ejemplo 4: grupos diferenciados de empleados

Grupo 1:

Sueldo: 1.535,2€
Casado: No -> 0,777
 Sí -> 0,223
Coche: No -> 0,82
 Sí -> 0,18
Hijos: 0,05
Alq/Prop: Alquiler -> 0,99
 Propia -> 0,01
Sindic.: No -> 0,8
 Sí -> 0,2
Bajas/Año: 8,3
Antigüedad: 8,7
Sexo: H -> 0,61
 M -> 0,39

Grupo 2:

Sueldo: 1.428,7€
Casado: No -> 0,98
 Sí -> 0,02
Coche: No -> 0,01
 Sí -> 0,99
Hijos: 0,3
Alq/Prop: Alquiler -> 0,75
 Propia -> 0,25
Sindic.: Sí -> 1,0
Bajas/Año: 2,3
Antigüedad: 8
Sexo: H -> 0,25
 M -> 0,75

Grupo 3:

Sueldo: 1.233,8€
Casado: Sí -> 1,0
Coche: No -> 0,05
 Sí -> 0,95
Hijos: 2,3
Alq/Prop: Alquiler -> 0,17
 Propia -> 0,83
Sindic.: No -> 0,67
 Sí -> 0,33
Bajas/Año: 5,1
Antigüedad: 8,1
Sexo: H -> 0,83
 M -> 0,17

- Estos grupos podrían ser interpretados como:
 - Grupo 1: sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas
 - Grupo 2: sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en casas de alquiler
 - Grupo 3: con hijos, casados y con coche. Mayoritariamente hombres propietarios de su vivienda. Poco sindicados

Unidad 1. Introducción a la minería de datos

Tipos de datos minables. BBDD relacionales

- Bases de datos relacionales. Una empresa de electrónica tiene sus sistemas OLTP

customer (*cust_ID, name, address, age, occupation, annual_income, credit_information, category, ...*)

item (*item_ID, brand, category, type, price, place_made, supplier, cost, ...*)

employee (*empl_ID, name, category, group, salary, commission, ...*)

branch (*branch_ID, name, address, ...*)

purchases (*trans_ID, cust_ID, empl_ID, date, time, method_paid, amount*)

items_sold (*trans_ID, item_ID, qty*)

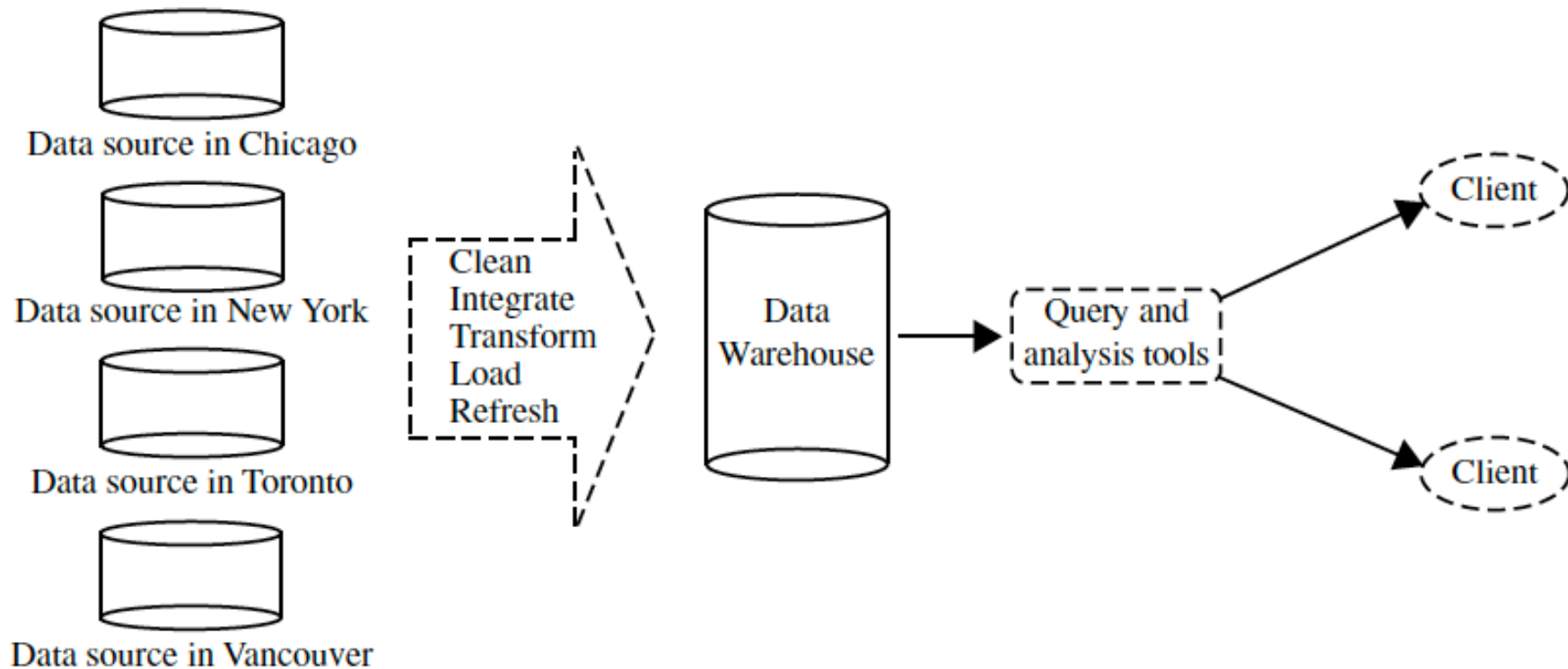
works_at (*empl_ID, branch_ID*)

- Se pueden realizar peticiones a través de una GUI con SQL y posteriormente aplicar algoritmos de minería

Unidad 1. Introducción a la minería de datos

Tipos de datos minables. Almacenes de datos

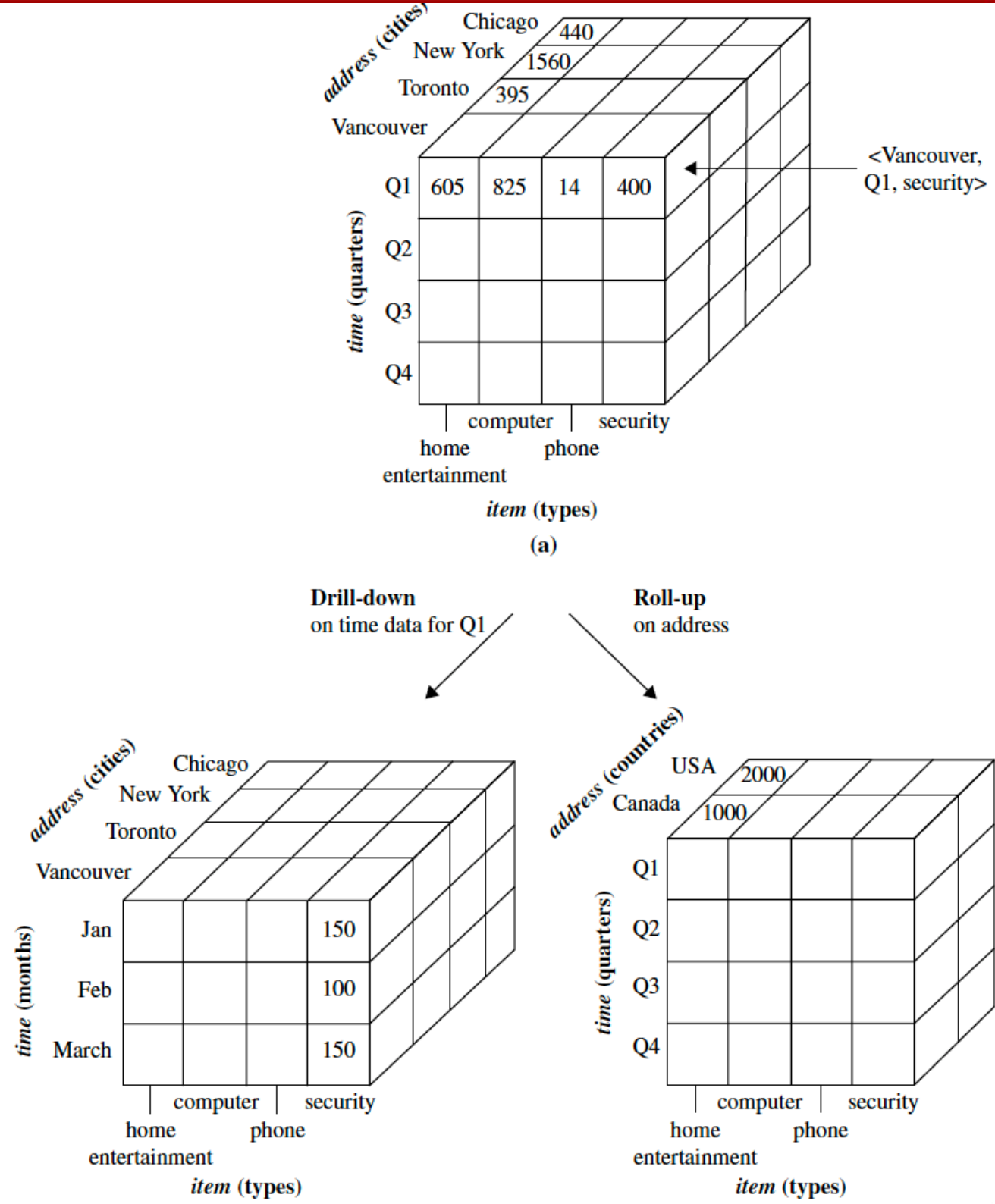
- Preparar los datos para un análisis en un almacén de datos OLAP



Unidad 1. Introducción

Tipos de datos minados

- Y posteriormente minar los cubos



Unidad 1. Introducción a la minería de datos

Tipos de datos minables. Datos transaccionales

- Datos de las transacciones en relación a las compras conjuntas que se han realizado

<i>trans_ID</i>	<i>list_of_item_IDs</i>
T100	I1, I3, I8, I16
T200	I2, I8
...	...



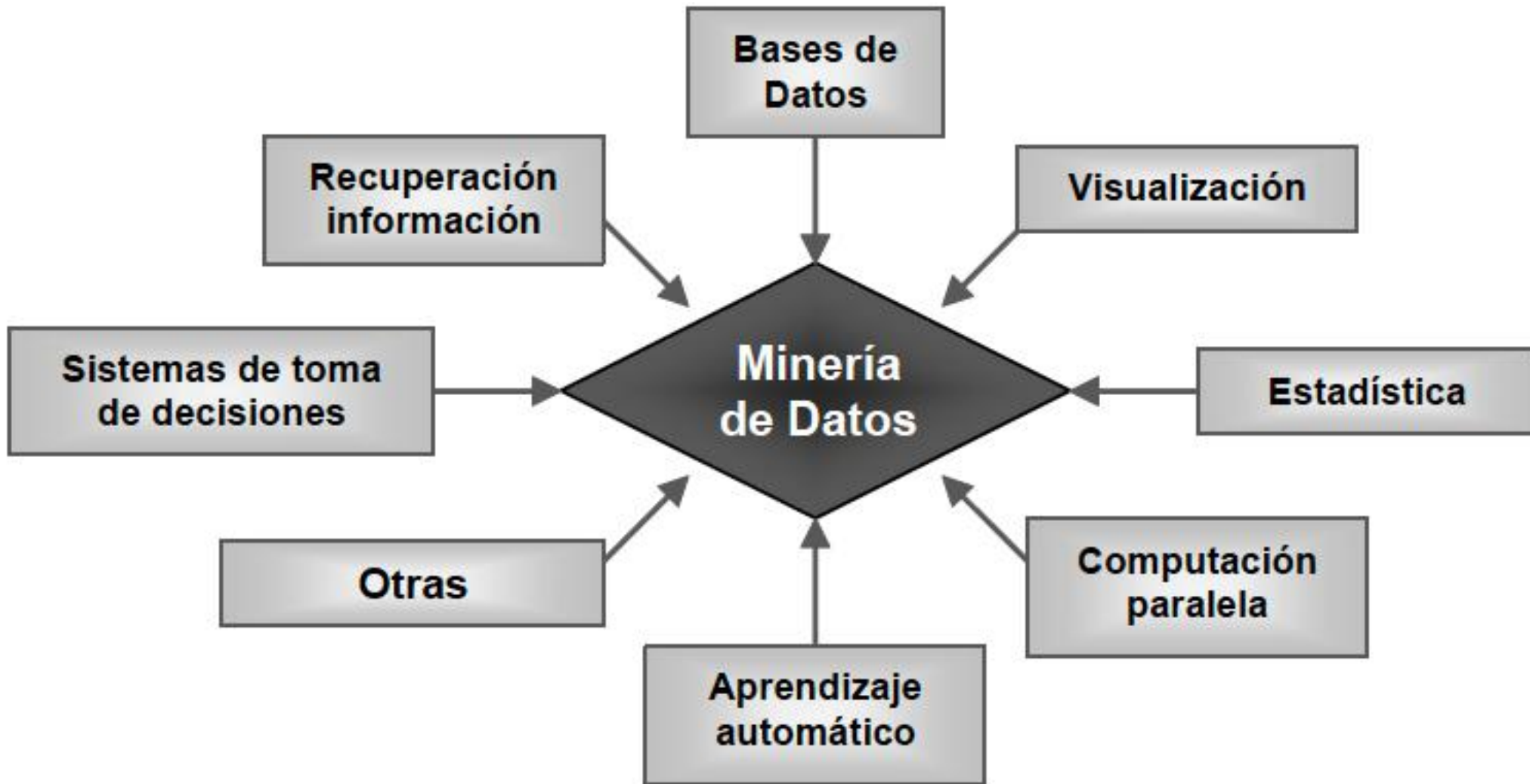
Unidad 1. Introducción a la minería de datos

Tipos de datos minables. Otros datos

- Pero todo esto es sólo la punta del iceberg
- Datos temporales y secuenciales
 - El histórico de bolsa, datos biológicos secuenciales
- Flujos de datos
 - Cámaras de seguridad o sensores como audio/temperatura
- Datos espaciales
 - Mapas o coordenadas GPS
- Hipertexto y multimedia
 - Todo tipo de texto, imágenes, vídeo, y audio
- La Web
 - Minar ese gran repositorio completo de todo tipo de información
- Cualquier aplicación que genere datos, es minable

Unidad 1. Introducción a la minería de datos

Campo multidisciplinar



Unidad 1. Introducción a la minería de datos

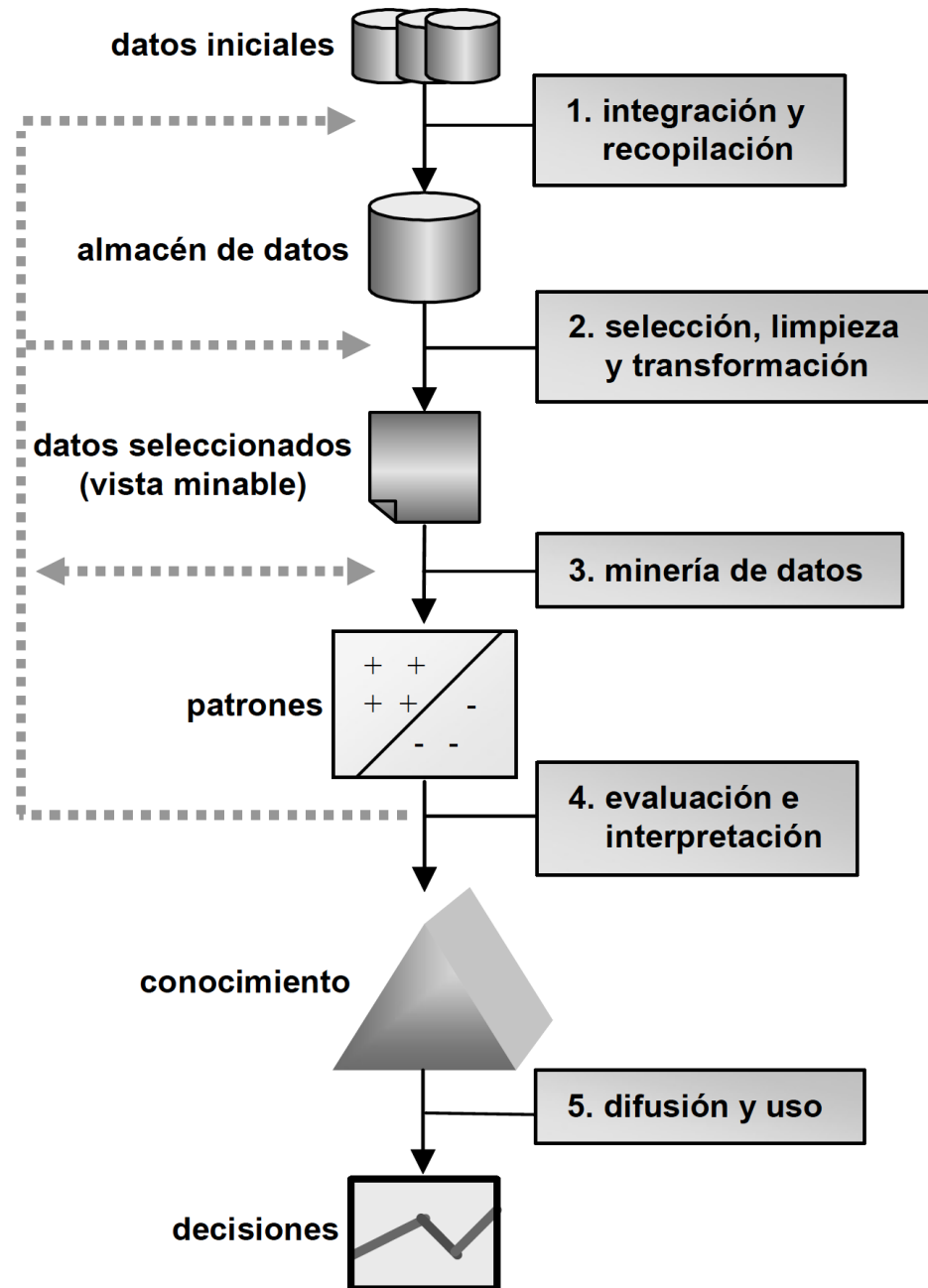
EL CICLO DE LA MINERÍA DE DATOS



Unidad 1. Introducción a la minería de datos

El descubrimiento de conocimiento en bases de datos

- Esta parte de minería es el paso esencial del descubrimiento de conocimiento en bases de datos (KDD)
- Secuencia iterativa de fases
- Las dos primeras fases se engloban como preparación de datos
- Fase previa de necesidades y requisitos

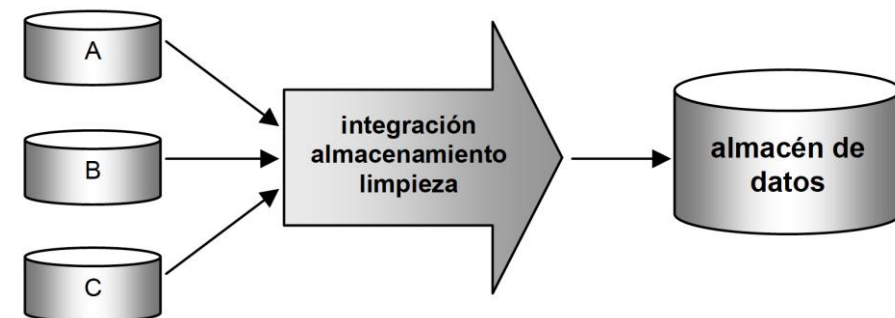


UNIVERSIDAD DE
MURCIA

Unidad 1. Introducción a la minería de datos

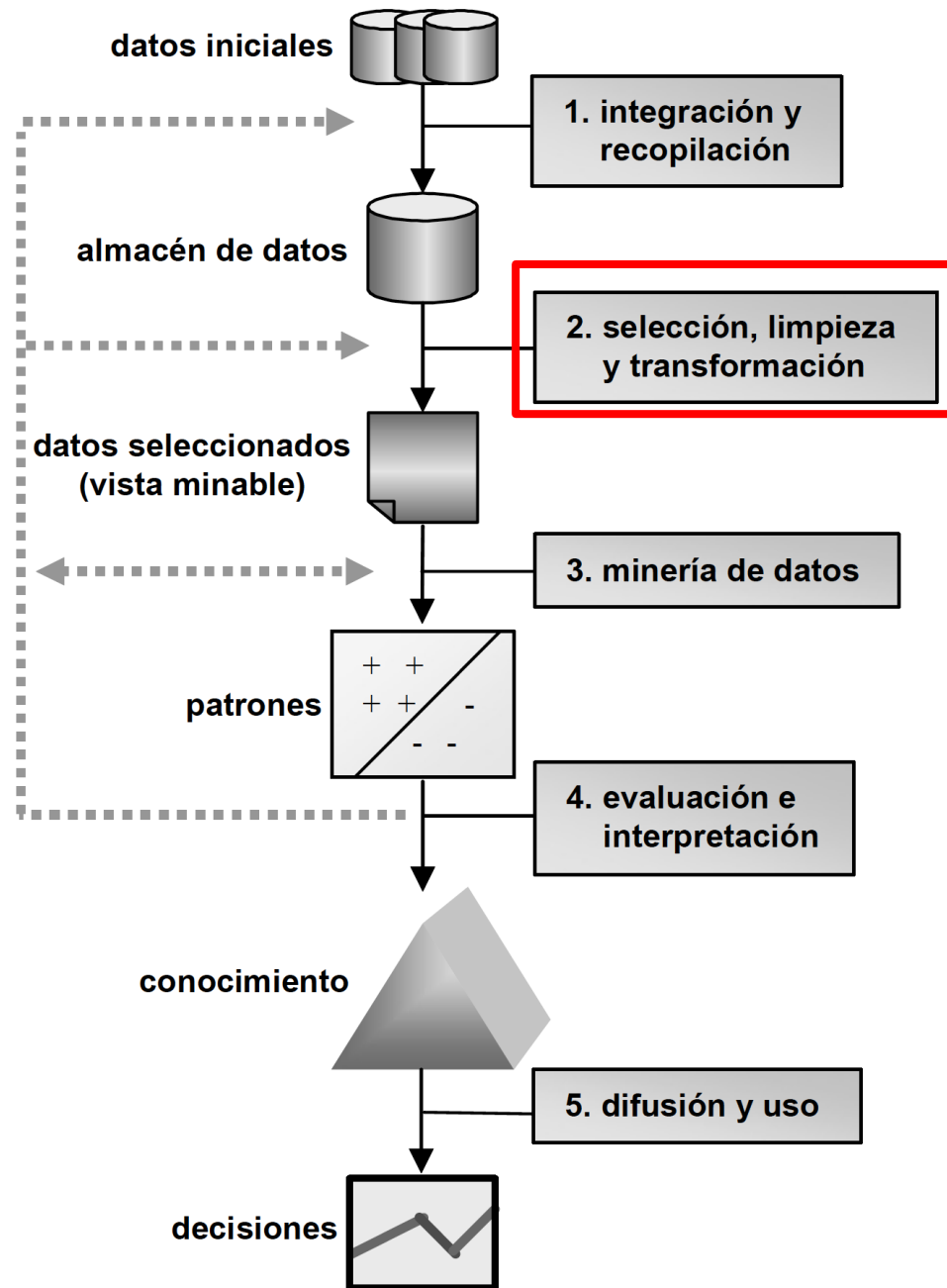
Fase de integración y recopilación

- El software empresarial realiza las tareas de procesamiento transaccional en línea (OLTP)
 - Orientado a realizar operaciones diarias, no minería
 - Necesidad de integrar diversas fuentes como datos públicos, bases de datos privadas, o de otros departamentos
- Integrar fuentes en un almacén de datos para permitir OLAP
 - Orientado a hechos y al procesado multidimensional de los datos
 - Muy superior para minar datos
- Proceso deductivo (OLAP) vs. Proceso inductivo (minería)
- En algunos casos (e.g. datos reducidos), se puede minar la BBDD



Unidad 1. Introducción a la minería de datos

Fase de selección, limpieza y transformación



Unidad 1. Introducción a la minería de datos

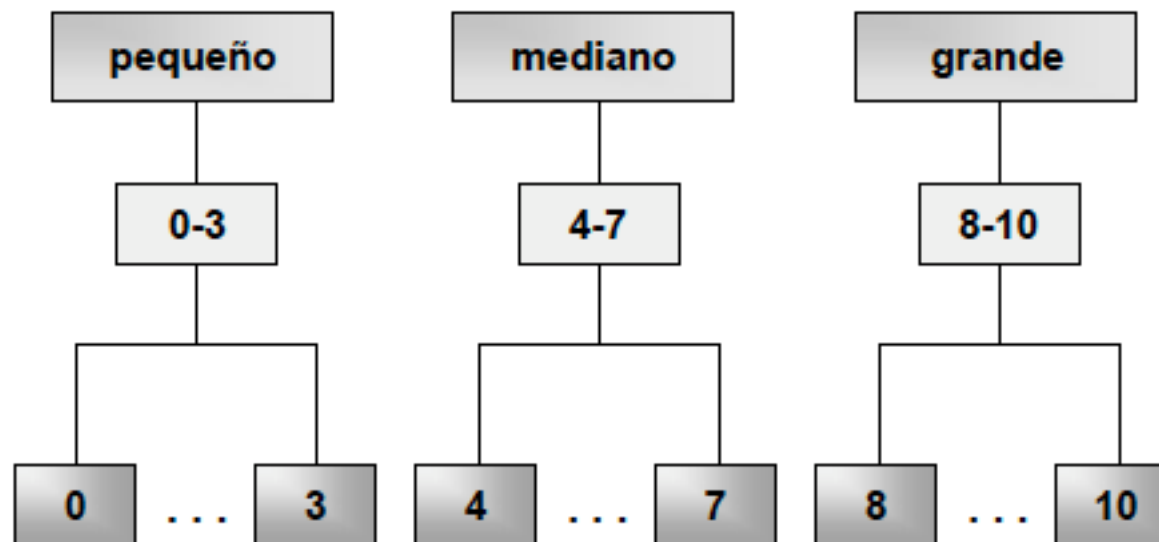
Fase de selección, limpieza y transformación

- La calidad del conocimiento no es sólo el algoritmo, sino también la calidad de los datos
 - Hay que preparar los datos para generar la vista minable
 - Relevancia de datos, gestión de atípicos (eliminarlos o no), datos faltantes
- Selección de variables en el preprocesamiento es clave
 - Imaginemos en un torneo de Wimbledon, ¿Cuándo se juega en base a climatología? ¿Qué variables necesitamos?
 - Se podrían usar todas las variables y dejar que el algoritmo optimice. ¿Qué problemas hay?
 - El conocimiento de negocio es clave para este filtrado inicial

Unidad 1. Introducción a la minería de datos

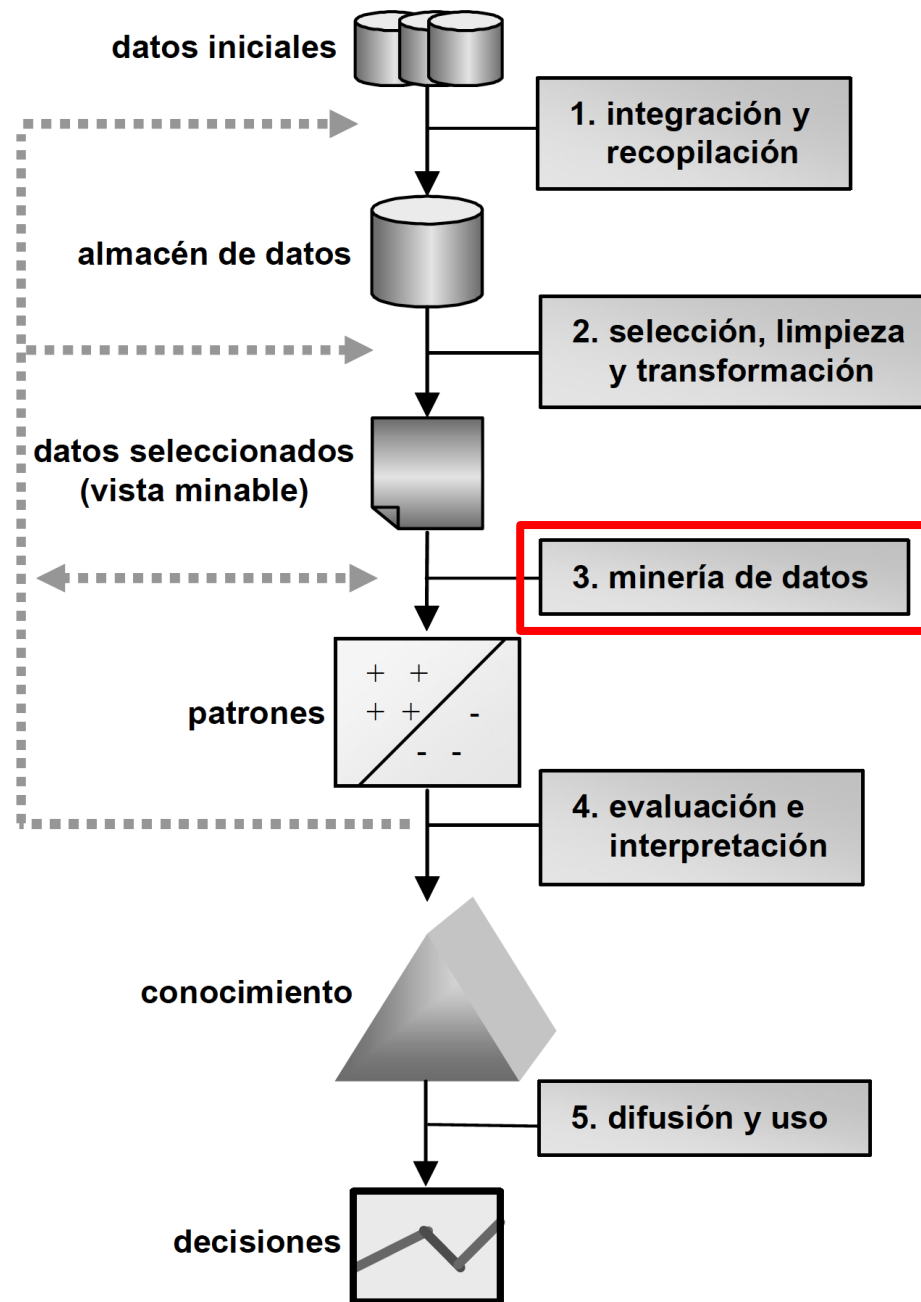
Fase de selección, limpieza y transformación

- Usar todos los datos vs. obtener muestra (representativa)
- Construcción de nuevas variables con procesado/funciones
 - Pueden tener mejor poder predictivo que originales
 - Por ejemplo, precio de viviendas de una zona si tenemos la población absoluta y la superficie
- Modificar el tipo de datos para uso de técnicas
 - Numerizar atributos para usar técnicas numéricas
 - O el inverso, discretizar atributos continuos a discretos/nominales



Unidad 1. Introducción a la minería de datos

Fase de minería de datos



Unidad 1. Introducción a la minería de datos

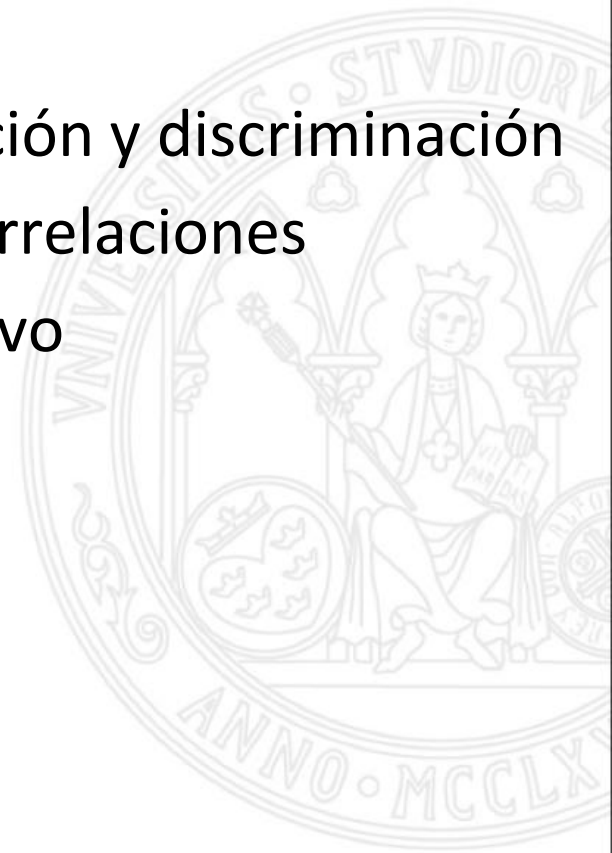
Fase de minería de datos

- La fase característica del KDD, usando este término a veces para englobar todo el proceso
- Objetivo de producir el nuevo conocimiento a partir de un modelo que se adaptará a los patrones de los datos
- Hay que tomar una serie de decisiones:
 1. ¿Qué tipo de tarea es el más apropiado? Por ejemplo, ¿para predecir que clientes dejaran de serlo en un banco?
 2. ¿Qué tipo de modelo aplicamos? Por ejemplo, ¿en clasificación si queremos que sea interpretable?
 3. ¿Qué algoritmo que resuelva la tarea aplicamos?
- Ahora vamos a desglosar todo esto

Unidad 1. Introducción a la minería de datos

Fase de minería. Tareas

- Hay que distinguir entre tipos de tareas, analogía con tipo de problema resuelto por un algoritmo
- Las tareas pueden ser **predictivas** o **descriptivas**
- Tipos de tareas:
 - Descripción de clase o concepto: Caracterización y discriminación
 - Minar patrones frecuentes, asociaciones y correlaciones
 - Clasificación y regresión para análisis predictivo
 - Agrupamiento
 - Análisis de atípicos



- Los casos pueden estar asociados con clases o conceptos
 - Por ej, en una tienda de electrónicos las clases de los objetos pueden ser el tipo de producto o el concepto de los clientes podría ser categoría de gasto
- La caracterización realiza un resumen de las características generales de dicha clase (por ejemplo, los tipos de productos que han incrementado ventas)
 - Las operaciones con cubos OLAP roll-up pueden ser usadas
 - Se puede representar de varias formas, con tablas, cubos, visualizaciones...
 - Ejemplo: “Resumen las características de clientes que gastan más de 5000€ anualmente en toda la electrónica”
 - Se obtiene un perfil de “Clientes entre 40 y 50 años, con trabajo, y con excelente crédito”

- La discriminación de datos es una comparación general con la clase objetivo, con respecto de otras clases
- Por ej, comparación entre productos de software que incrementaron un 10% vs. aquellos que disminuyeron 30% en el mismo periodo
- Por ej, en la tienda de electrónica, se quieren comparar clientes que compren regularmente vs. aquellos que compren con baja regularidad
 - Se obtiene que el 80% de los que compren regularmente tienen entre 20 y 40 años con titulación universitaria, vs. 60% de los que compran con baja frecuencia, son mayores sin título universitario

Unidad 1. Introducción a la minería de datos

Fase de minería. Tareas. Patrones frecuentes

- Son aquellos que ocurren de forma frecuente en los datos
- Itemset frecuentes, aquellos conjuntos de ítems que aparecen frecuentemente juntos en datos transaccionales
 - Por ejemplo, hamburguesas y pan de hamburguesas comprados
- Patrones secuenciales frecuentes en acciones teniendo en cuenta el orden
 - Por ejemplo, análisis secuencial obtiene que 30% de clientes que compraron un televisor, compraron un DVD antes de 2 meses
- Correlaciones, grado de similitud entre dos variables numéricas, mediante un índice de correlación (entre -1 y 1)
 - Por ejemplo, análisis de correlación entre grosor de protector de material electrónico y frecuencia de incendios

- La clasificación es el proceso de encontrar un modelo (o función) que describe y distingue entre clases o conceptos. Se quiere maximizar la precisión y especificidad
 - Oftalmólogo desea un sistema para determinar si alguien debe operarse o no en base al paciente
- Por otra parte, la regresión se centra en un valor real, es decir, predecir un número. Se quiere minimizar el error medio entre el valor real y el valor predicho
 - Un empresario quiere conocer el costo de un nuevo contrato en función de anteriores
- Los modelos se entrenan sobre juegos de datos de entrenamiento y se evalúan sobre nuevos datos no vistos

- Típica tarea descriptiva que busca grupos “naturales” a partir de **datos no etiquetados**
 - Los datos se agrupan en función de un criterio de similitud:
 - Maximizar la similaridad intragrupo
 - Minimizar la similaridad entre grupos
 - No confundir con la clase
 - También se le conoce como segmentación (segmentación en grupos)
 - Puede ayudar a la generación de taxonomías
- Por ej, una librería online utiliza agrupamiento para encontrar perfiles de clientes y ofrecer un servicio personalizado. Por ejemplo, recomendar libros de otros clientes del grupo

- Objetos del dataset que no concuerdan en el comportamiento o modelo general
 - Se los conoce como atípicos o anomalías (*outliers* o *anomalies*)
- Muchos métodos los descartan como ruido o excepciones
- En otros casos, estos eventos atípicos son los que realmente queremos identificar
 - Por ejemplo, ataques a la red o terremotos
- El funcionamiento suele ser diferente, se configuran modelos con la distribución normal, y se analiza cuando un nuevo caso es con alta probabilidad diferente de dicha distribución
- Ejemplo de aplicación, usar grandes cantidades de transacciones para encontrar atípicos, por ej, en función de tipo de compra, localización, frecuencia o cantidad
 - Aplicaciones actuales avisan con este tipo de atípicos

Unidad 1. Introducción a la minería de datos

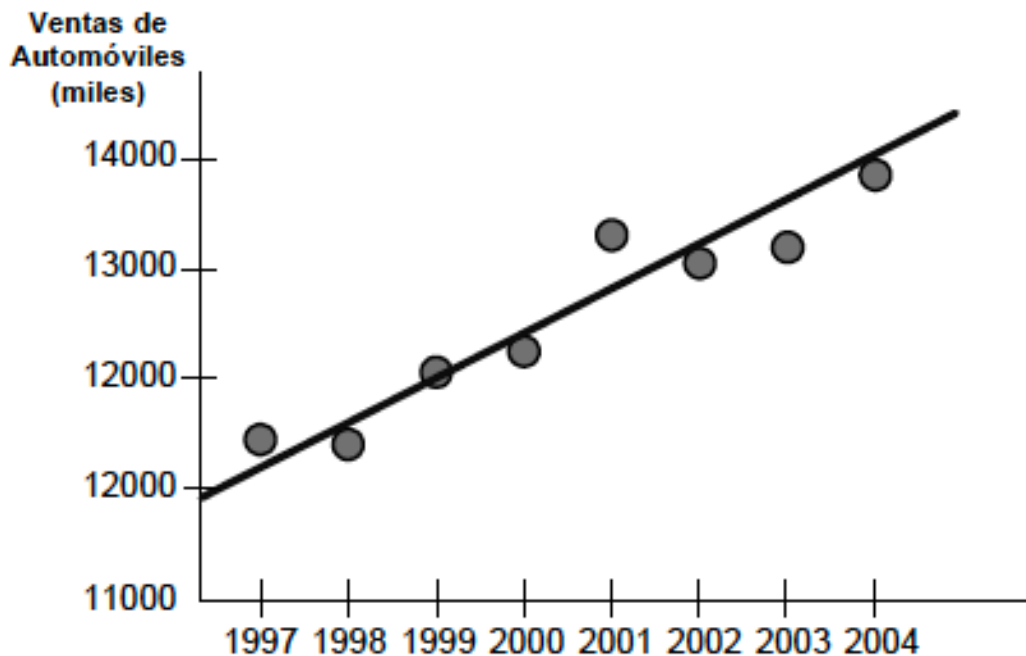
Fase de minería. Técnicas de minería de datos

- Existen numerosos paradigmas detrás de las técnicas utilizadas para cumplir dichas tareas, por ejemplo:
 - Técnicas de inferencia estadística
 - Árboles de decisión
 - Redes neuronales
 - Inducción de reglas
 - Aprendizaje computacional
 - Aprendizaje bayesiano
 - Programación lógica inductiva
- Algunas técnicas pueden tener más de una aplicación, otras van directamente orientadas a una aplicación
- Tienen fundamentos diferentes en estadística o inteligencia artificial
- Ponemos algunos ejemplos

Unidad 1. Introducción a la minería de datos

Fase de minería. Técnicas. Regresión lineal

- La técnica tradicional por excelencia, donde se ajusta una recta en función de variables que minimicen el error



$$Y = c_0 + c_1X_1 + \dots + c_nX_n$$

- Se pueden usar funciones cuadrado, inversa, logaritmos...) para que sea no lineal. Flexibiliza la salida

$$Y = c_0 + f_1(x_1) + \dots + f_n(x_n)$$

Unidad 1. Introducción a la minería de datos

Fase de minería. Técnicas. Reglas de asociación

- Ejemplo de reglas de asociación. Eres un manager en la tienda de electrónica y quieres saber que ítems se compran conjuntamente (misma transacción). Encuentras que:

*compra(X, "ordenador") => compra(X, "software")
[soporte=1%, confianza=50%]*

- Regla con un solo predicado (compra) que se repite (itemset)
- Si tuviéramos toda la base de datos, podríamos encontrar que:

*edad(X, "20..29") ^ ingresos(X, "40k..49K") => compra(X, "portátil")
[soporte=2%, confianza=60%]*

- Regla de asociación multidimensional con tres predicados
- Las reglas de asociación se descartan como no interesantes si no tienen un soporte o confianza mínimos

Unidad 1. Introducción a la minería de datos

Fase de minería. Técnicas. Naive bayes

- Método probabilístico para ver que una instancia sin clasificar pertenezca a una clase
- Se basa en la regla de Bayes, si tenemos una hipótesis H sustentada por una evidencia E, entonces:

$$p(H|E) = p(E | H) * p(H) / p(E)$$

#Instancia	edad	hijos	practica_deporte	salario	buen_cliente
1	joven	sí	no	alto	sí
2	joven	no	no	medio	no
3	joven	sí	sí	medio	no
4	joven	sí	no	bajo	sí
5	mayor	sí	no	bajo	sí
6	mayor	no	sí	medio	sí
7	joven	no	sí	medio	sí
8	joven	sí	sí	alto	sí
9	mayor	sí	no	medio	sí
10	mayor	no	no	bajo	no

$$p(sí | E) = \frac{[p(edad_E | sí) \cdot p(hijos_E | sí) \cdot p(practica_deporte_E | sí) \cdot p(salario_E | sí)] \cdot p(sí)}{p(E)}$$

Fase de minería. Técnicas. Árboles de decisión

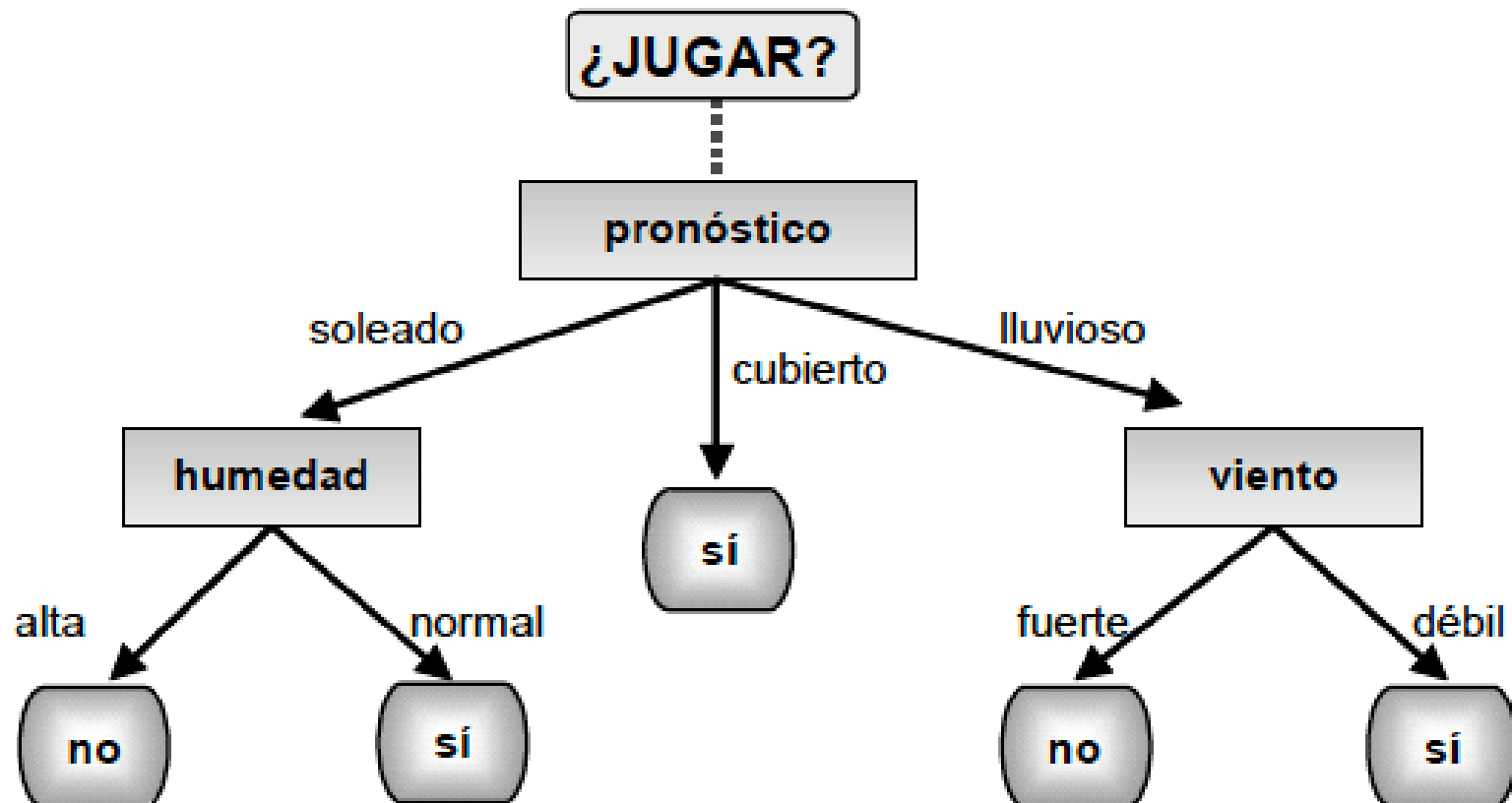
- Serie de decisiones o condiciones organizadas de forma jerárquica a modo de árbol
 - Útiles que alta dimensionalidad, que mezclen datos categóricos y numéricos
 - Pueden usarse tanto para predecir variables categóricas, como para predecir variables continuas (árboles de regresión)

#instancia	pronóstico	humedad	viento	jugar
1	soleado	alta	débil	No
2	cubierto	alta	débil	Sí
3	lluvioso	alta	débil	Sí
4	lluvioso	normal	fuerte	No
5	soleado	normal	débil	Sí
...	...	...	...	...

Unidad 1. Introducción a la minería de datos

Fase de minería. Técnicas. Árboles de decisión

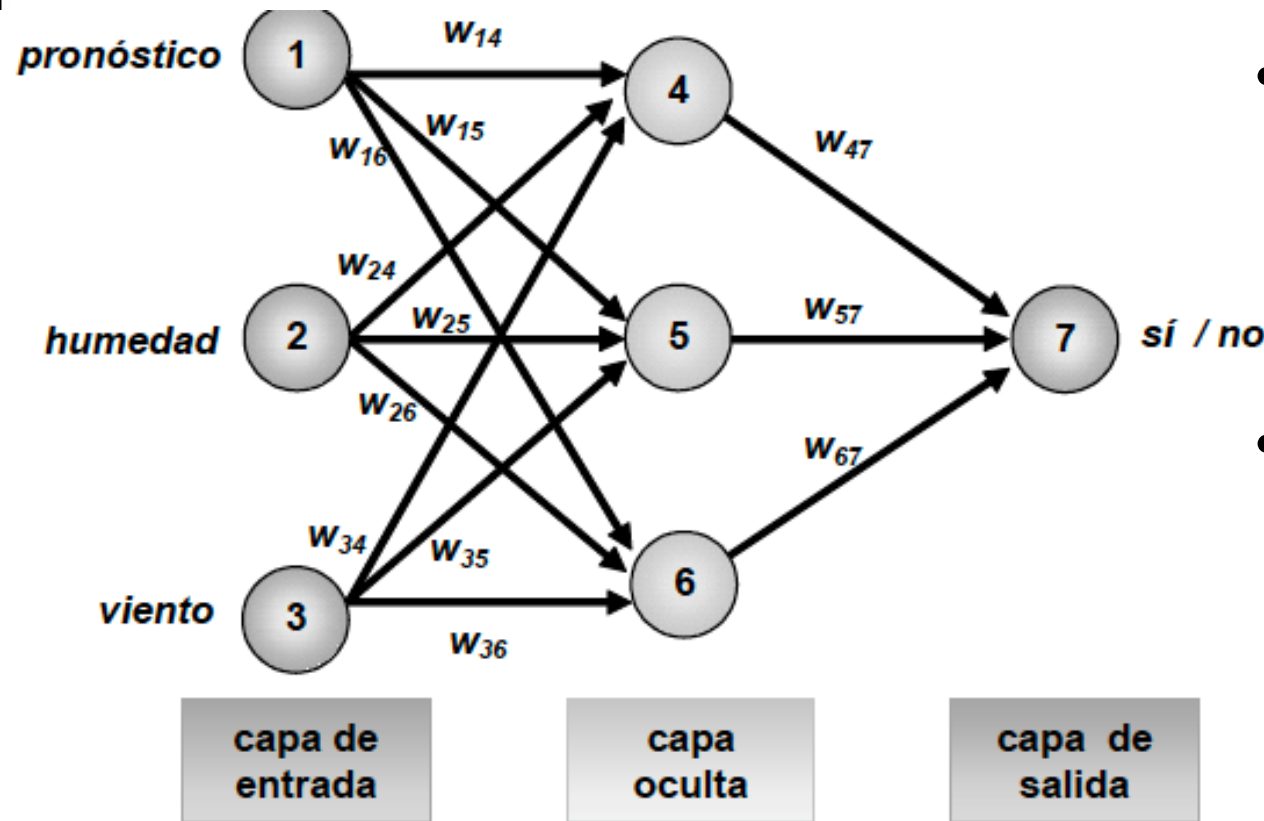
- Aproximación de “divide y vencerás” partiendo el espacio del problema en subconjuntos
- Arriba el nodo raíz con el problema resolver
- Los nodos internos corresponden a particiones de atributos particulares, conduciendo a nodos de decisión



Unidad 1. Introducción a la minería de datos

Fase de minería. Técnicas. Redes neuronales

- Un paradigma de computación que modeliza problemas complejos cuando hay interacciones no lineales
 - Para usar datos nominales tienen que numerizarse
 - Pueden verse como un grafo dirigido con muchos nodos y arcos entre ellos

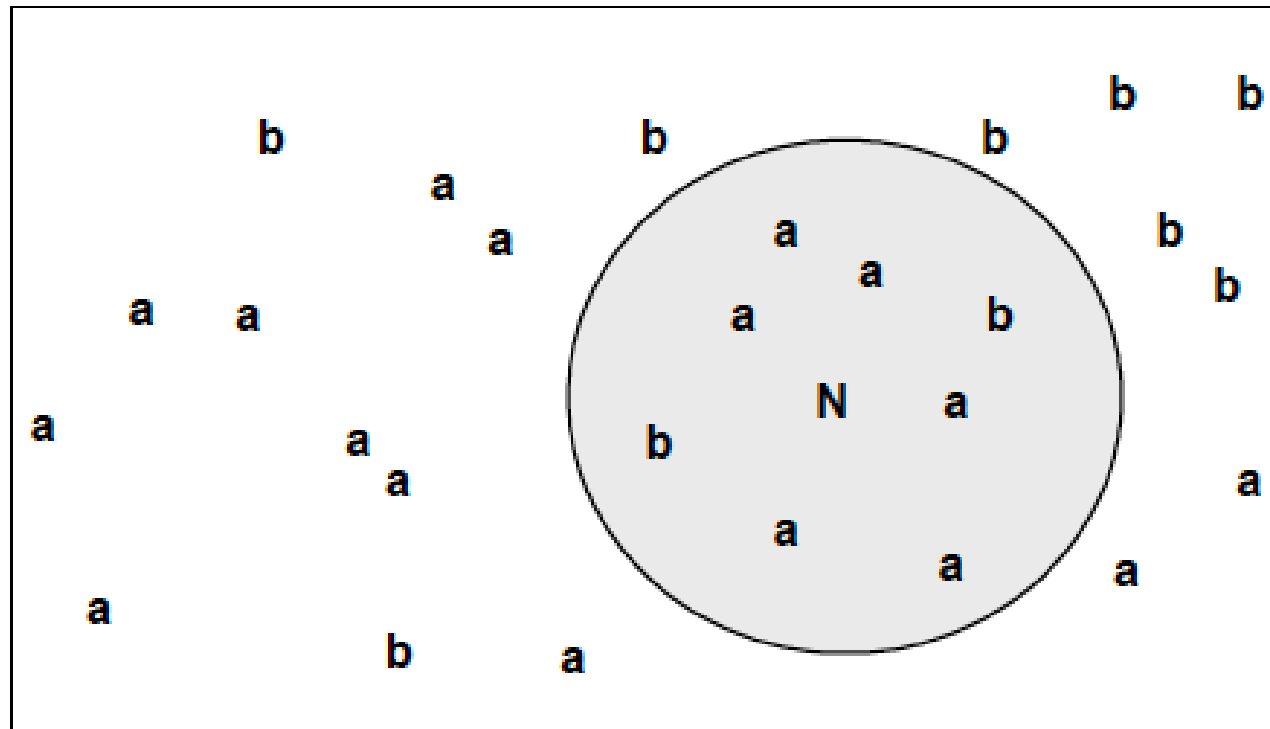


- Estimar pesos mediante backpropagation (reducir error en la salida)
- Requieren muchos datos y son poco interpretables

Unidad 1. Introducción a la minería de datos

Fase de minería. Técnicas. k vecinos más próximos

- K vecinos más próximos, usa una votación mayoritaria para etiquetar un nuevo caso
- Se tiene que hacer por medio de la definición de distancias, que pueden ser sencillas o más sofisticadas (problemática con categóricas)



Unidad 1. Introducción a la minería de datos

Fase de minería. Construcción del modelo

- La construcción tiene un carácter iterativo, para explorar modelos alternativos y encontrar el que resuelve
 - Se usan distintos parámetros o técnicas de entrenamiento
- Algunos algoritmos requieren de pre-procesados específicos
- El proceso de construcción de modelos predictivos tiene unas etapas bien de finidas de entrenamiento y validación
 - Imprescindible para tener modelos robustos y que generalicen, evitando sobreajuste
 - La idea básica es usar juegos de datos de entrenamiento y de evaluación.
 - También usar técnicas de cros-validación para ajustar los parámetros de los modelos

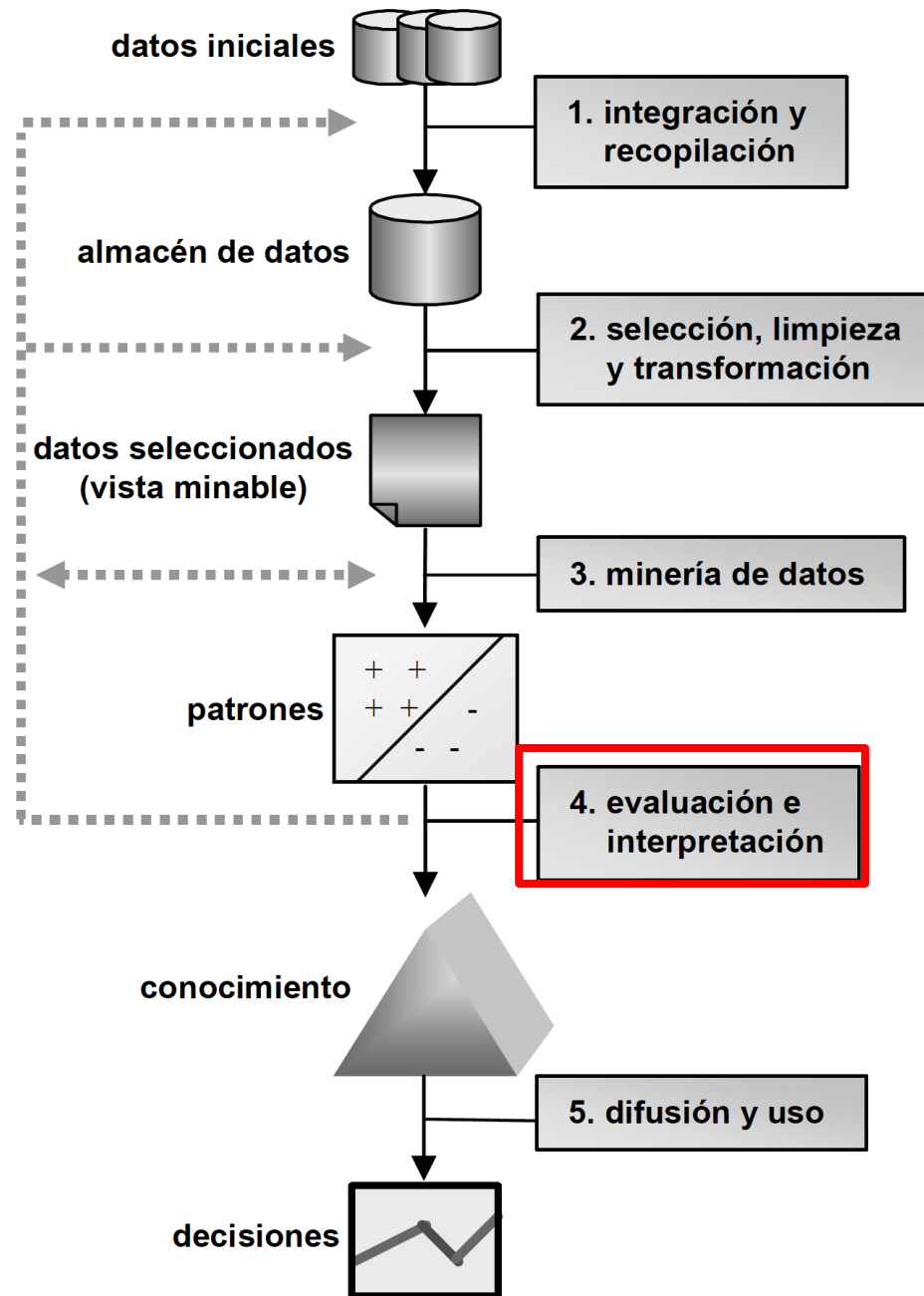
Fase de minería. ¿Son todos los patrones interesantes?

- ¿Qué hace que un patrón encontrado sea interesante?
 - Fácilmente entendible por humanos
 - Fiable al ser aplicado a nuevos datos
 - Potencialmente útil y novel
 - Valida una hipótesis que se querían confirmar
- Medidas objetivas como el soporte y confianza, o la exactitud
- Medidas subjetivas como si son sorprendes o accionables
- ¿Se pueden encontrar todas las interesantes?
 - Es ineficiente, nos centramos en dar medidas y criterios de lo que queremos encontrar
- ¿Pueden los sistemas generar solo las interesantes?
 - Sigue siendo un desafío de optimización en el que se está trabajando

$$\text{support}(X \Rightarrow Y) = P(X \cup Y),$$
$$\text{confidence}(X \Rightarrow Y) = P(Y|X).$$

Unidad 1. Introducción a la minería de datos

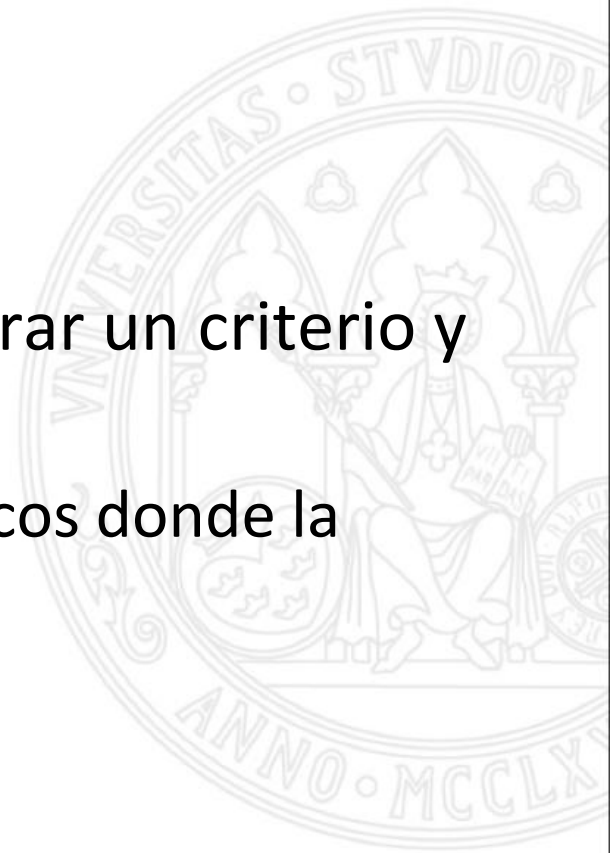
Fase de evaluación e interpretación



Unidad 1. Introducción a la minería de datos

Fase de evaluación e interpretación

- Medir la calidad de los patrones descubiertos no es un problema trivial
- Idealmente, deben tener tres cualidades
 - Precisos
 - Comprensibles (es decir, inteligibles)
 - Interesantes (útiles y novedosos)
- Según las aplicaciones, puede interesar mejorar un criterio y sacrificar otro
 - Por ejemplo, en el caso de diagnósticos médicos donde la comprensibilidad es clave



Unidad 1. Introducción a la minería de datos

Fase de evaluación/interpr. Técnicas evaluación

- Para entrenar y probar un modelo, se parte de los conjuntos de entrenamiento (*training set*) y test (*test set*), evita sobreajuste
 - Se entrena con uno, y se evalúa con el otro para tener una estimación realista
- El método de evaluación más básico es la validación simple, que reserva un conjunto de datos para cada parte
- Si la muestra es pequeña, se suele usar validación cruzada
 - Se denomina de *n*-pliegues (*n-fold cross validation*)
 - Se entrena con $n-1$ y se usa el resto para evaluar iterativamente
- Otra técnica cuando hay pocos datos es el *bootstrapping*
 - Se construye un modelo con todos los datos
 - Se crean sub-conjuntos haciendo muestreo con reemplazo llamados *bootstrap samples* (pudiendo repetir casos)
 - Se construyen modelo con cada conjunto y se evalúa en test

Unidad 1. Introducción a la minería de datos

Fase de evaluación/interpr. Medidas evaluación

- Dependiendo de la tarea existen diferentes evaluaciones
- En clasificación, se puede usar la exactitud (muestras clasificadas correctamente con respecto al total)
- En reglas de asociación, también se usa la cobertura y confianza
- En la regresión con una salida numérica, se suele usar el error cuadrático medio
- En el agrupamiento, se suelen usar medidas de cohesión de cada grupo y de separación entre grupos
 - Distancia media al centro del grupo
 - Distancia entre los grupos
- También hay medidas subjetivas, como el interés, novedad, simplicidad o comprensibilidad

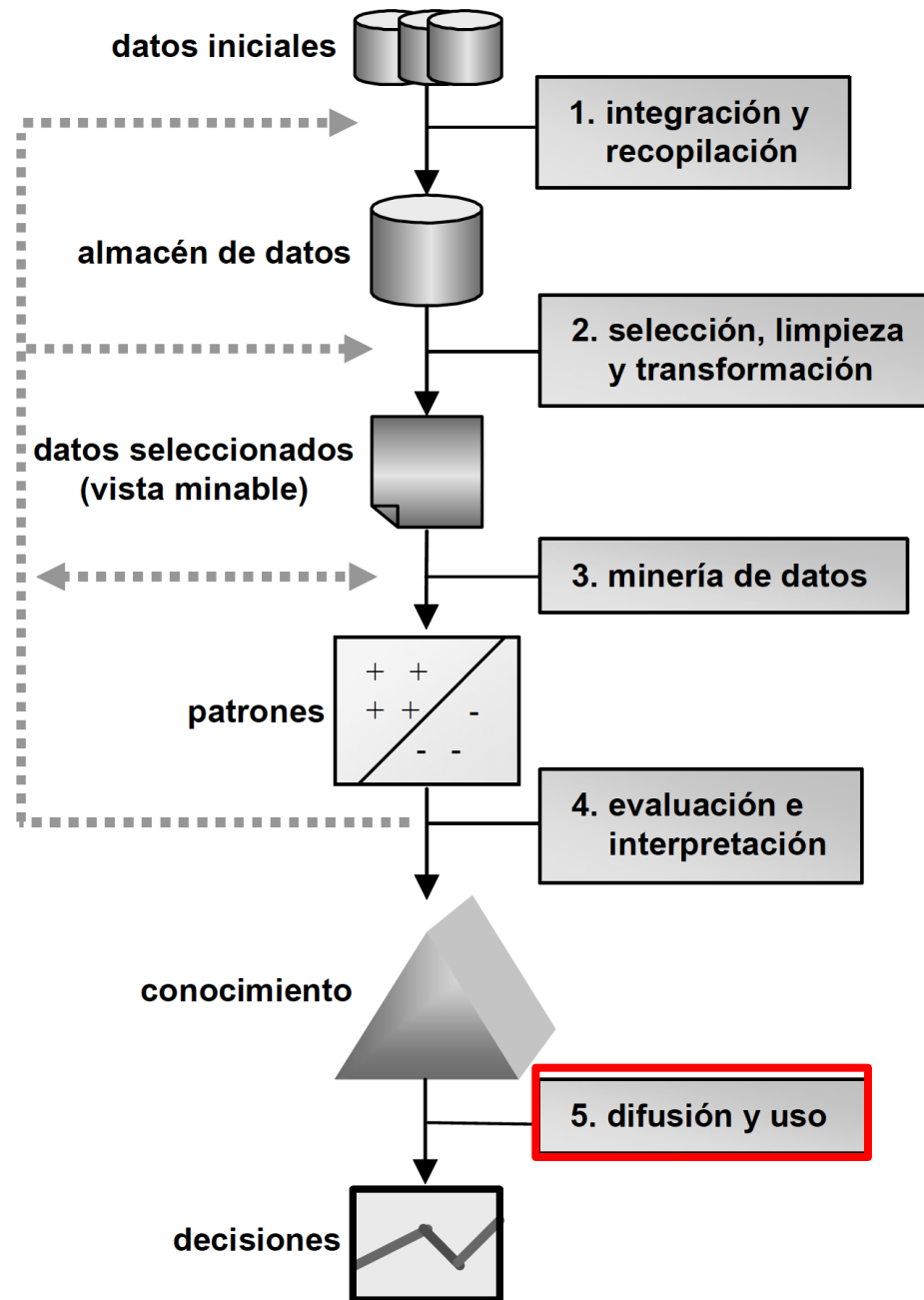
Unidad 1. Introducción a la minería de datos

Fase de eval/interpretación. Interpretación y contexto

- También hay que evaluar el contexto de aplicación
- Por ejemplo, en clasificación, usar la exactitud, no tiene en cuenta las distribuciones de clases no balanceadas
 - Por ejemplo, detectar transacciones que sea ilícitas
 - Por ejemplo, diagnóstico médico para tomar decisión de ser dar de alta, hospitalizado o UCI (86.5%, 13% y 0,5% respectivamente)
- Pone de relevación el tipo de error y coste asociado
- Se suele usar la matriz de confusión y se pueden adaptar los costes de los distintos errores
- La predicción de un modelo, no garantiza que funcione así en el mundo real
 - Imaginemos modelos de consumo energético implementados con datos antes de la guerra de Ucrania y su comportamiento después

Unidad 1. Introducción a la minería de datos

Fase de difusión y uso



Unidad 1. Introducción a la minería de datos

Fase de difusión y uso

- Una vez construido y validado el modelo, puede usarse para:
 - Un analista recomienda acciones en base al modelo/resultados
 - Aplicar el modelo a diferentes conjuntos de datos
- También puede utilizarse para ser incorporado a un sistema
- Se debe explotar dicho nuevo conocimiento para obtener ventajas
- Tanto en situaciones manuales o automáticas, tiene que distribuirse y ser comunicado a los usuarios
- Hay que medir cómo evoluciona el modelo, dando mantenimiento (por ej. entrenamientos automatizados)
 - Incluso hay algoritmos que tienen en cuenta la temporalidad de los casos

Unidad 1. Introducción a la minería de datos

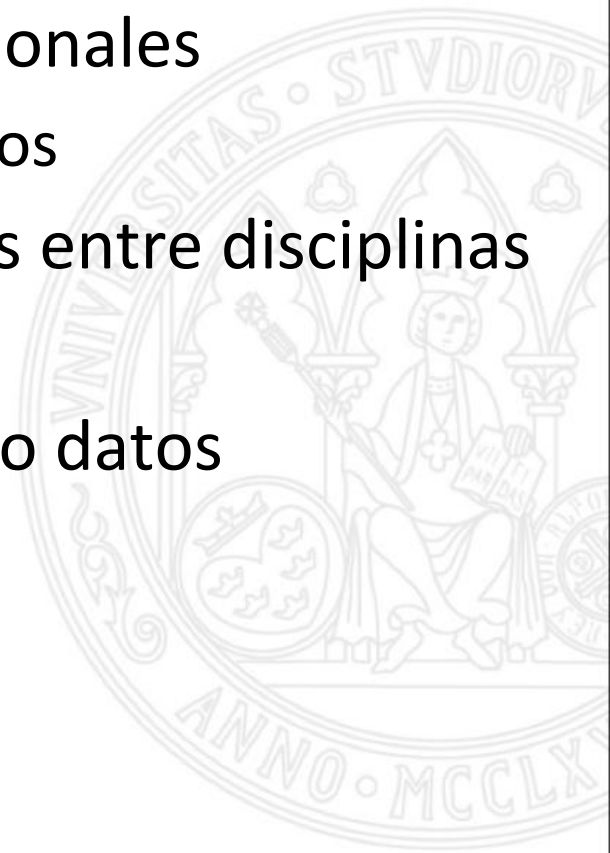
PRINCIPALES DESAFÍOS DE LA MINERÍA DE DATOS



Unidad 1. Introducción a la minería de datos

Principales desafíos. Metodología

- Investigación básica en el desarrollo de nuevos métodos
- Minar nuevos tipos de conocimiento en el mundo cambiante
 - Por ejemplo, en redes de información y difusión
- Minar conocimiento en espacios multidimensionales
 - Diferentes dimensiones y agregaciones en cubos
- Un esfuerzo interdisciplinar, llevando métodos entre disciplinas
 - Por ejemplo, *bug mining* o *bioinformatics*
- Mejorar el manejo de la incertidumbre, ruido o datos incompletos



Unidad 1. Introducción a la minería de datos

Principales desafíos. Interacción de usuarios

- Minería interactiva. Implementar interfaces flexibles y entornos de minado exploratorios
 - Cambiar el foco de la búsqueda, refinar las peticiones, pivotar una tabla...
- Incorporar antecedentes conocidos. Incorporar el conocimiento experto de los usuarios para ser usado
- Minería bajo demanda y lenguajes de minería. Peticiones sencillas que respondan a preguntas de minería
- Presentación y visualización de resultados de minería
 - ¿Cómo presentar los resultados de forma viva, flexible y que sea fácilmente interpretable por humanos?

Unidad 1. Introducción a la minería de datos

Principales desafíos. Minería de datos y sociedad

- El impacto social de la minería en la sociedad
 - Usarla para mejorar a la sociedad y proteger el uso incorrecto
 - Ataques adversariales
 - Sesgos sociales
- Preservar la privacidad en la minería
 - Evitar revelar la información de individuos
 - Dar la oportunidad de optar a dar datos o no
- Minería de datos invisible
 - Democratizar el uso de herramientas de minería de datos
 - Implementar aplicaciones usables por usuarios no técnicos (e.g. búsqueda en internet)
 - Desconocimiento de los usuarios del funcionamiento de algoritmos que les impactan