

# Introducción a la Ciencia de Datos y Minería de Datos



# Introducción a la Ciencia de Datos

## CONTENIDO:

Introducción a la Ciencia de Datos

Software y Lenguajes de Programación:

Lenguajes Python y R

Paquetes de R para el Análisis de Datos

Aprendizaje Supervisado: Clasificación y Regresión:

Modelos estadísticos (Regresión lineal, Regresión no lineal, KNN,...)

Modelos de la Inteligencia Artificial (Arboles de Decisión, Reglas, Random Forest, SVM,...)

Algoritmos avanzados (Deep Learning,...)

Problemas (Clasificación no Balanceada,...)

Uso de Datos Masivos

Aprendizaje no Supervisado:

Reglas de Asociación

Clustering

Modelos Básicos

Uso de Datos Masivos

Análisis Estadístico de Experimentos

Preprocesamiento de Datos

Detección de Anomalías

Big Data

Técnicas (Cloud computing, Hadoop, MongoDB, Hive)

Herramientas (Pig, Spark,...)

www.kdnuggets.com/2016/10/top-10-data-science-videos-youtube.html

Data Mining, Analytics, Big Data, and Data Science

KDnuggets™ Subscribe to [KDnuggets News](#) | Follow [Twitter](#) [Facebook](#) [LinkedIn](#) | [Contact](#)

SOFTWARE | NEWS | Top stories | Opinions | Tutorials | JOBS | Academic | Companies | Courses | Datasets | EDUCATION  
KDnuggets Home » News » 2016 » Oct » Tutorials, Overviews » **Top 10 Data Science Videos on Youtube ( 16:n37 )**

**Latest News, Stories**

- Intellectual Ventures: Sr. Machine Learning Algorithm ...
- European Machine Intelligence Landscape
- Clustering Key Terms, Explained
- PAW Business, NYC Oct 23-27: Last Chance to Save
- LinkedIn Knowledge Graph – KDnuggets Interview

**Top 10 Data Science Videos on Youtube**

◀ Previous post Next post ▶

[f](#) [in](#) [G+1](#) 10 Share 26 [Tweet](#)

Tags: [Data Science](#), [Data Scientist](#), [DJ Patil](#), [Online Education](#), [R](#), [Videolectures](#), [Youtube](#)

Learning and the future are the key topics in the recent Youtube videos on Data Science. The main questions revolve around: "how to become a Data Scientist", "what is a data scientist", and "where data science is going". But why there is so little explanation of data science to the masses?

<http://www.kdnuggets.com/2016/10/top-10-data-science-videos-youtube.html>



## The Top Skills of 2016 on LinkedIn Global

1	Cloud and Distributed Computing	0	6	Network and Information Security	+1
2	Statistical Analysis and Data Mining	0	7	Mobile Development	-1
3	Web Architecture and Development Framework	+6	8	Data Presentation	NR
4	Middleware and Integration Software	+1	9	SEO/SEM Marketing	-5
5	User Interface Design	+5	10	Storage Systems and Management	-2

\* NR (Not recorded in 2015)



## Analytics Trends 2016 | The Next Evolution

Forty percent of respondents to a 2015 MIT Sloan Management Review survey say they have difficulty hiring analytical talent. Only 17 percent of “analytically challenged” firms say they have the talent they need. Among companies reported to be “analytics innovators,” 74 percent said they had the analytics talent needed.

[https://www2.deloitte.com/content/dam/Deloitte/za/Documents/risk/ZA\\_Analytics\\_Trends\\_Risk\\_100316.pdf](https://www2.deloitte.com/content/dam/Deloitte/za/Documents/risk/ZA_Analytics_Trends_Risk_100316.pdf)

# InformationWeek

Join us live at  
**Interop** ITX

IT Leadership

DevOps

Security

Cloud

Data Ma

DATA MANAGEMENT // BIG DATA ANALYTICS

NEWS

5/24/2016  
09:06 AM

## Big Data, Analytics Sales Will Reach \$187 Billion By 2019



Jessica Davis  
News

Market research firm IDC forecasts a 50% increase in revenues from the sale of big data and business analytics software, hardware, and services between 2015 and 2019. Services will account for the biggest chunk of revenue, with banking and manufacturing-led industries poised to spend the most.

[http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-sales-will-reach-\\$187-billion-by-2019/d/d-id/1325631](http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-sales-will-reach-$187-billion-by-2019/d/d-id/1325631)



# 2016 - 2026 Worldwide Big Data Market Forecast

by Ralph Finos | 30 March 2016 | Big Data, Featured, Forecasts, Premium

The big data market grew 23.5% in 2015, led by Hadoop platform revenues. We believe the market will grow from \$18.3B in 2014 to \$92.2B in 2026 – a strong 14.4% CAGR. Growth throughout the next decade will take place in three successive and overlapping waves of application patterns – Data Lakes, Intelligent Systems of Engagement, and Self-Tuning Systems of Intelligence. Increasing amounts of data generated by sensors from the Internet of

<http://wikibon.com/2016-2026-worldwide-big-data-market-forecast/>

Portada | EcoDiario | Ecoteuve | Informalia | Evasión | Ecomotor | Ecoley | Ecotrader | elMonitor | Economíahc



Jueves, 20 de Octubre de 2016 Actualizado a las 20:18

Innovacion

Portada Mercados y Cotizaciones ▾ Empresas ▾ Economía ▾ Tecnología ▾ Vivienda Opinión/

IBEX 35 ▾ -0,08% | EURUSD ▾ -0,82% | I. GENERAL DE MADRID ▾ +0,12% | DOW JONES ▾ -0,22% | ECO10 ▲ +0,08%

DESTACAMOS

¿Subida de tipos de interés del BCE? Antes llegarán los coche

EN ECODIARIO.ES

La estatua de Franco en el Born, al 'desguace': la tumban tras

# Luca: la nueva filial de Telefónica para ofrecer servicios de Big Data a las empresas

<http://www.eleconomista.es/negocio-digital/innovacion/noticias/7904727/10/16/Telefonica-crea-una-filial-para-ofrecer-servicios-de-Big-Data.html>

# Objetivos:

- Introducir los conceptos de Ciencia de Datos, Minería de Datos, Big Data
  - Conocer las etapas del proceso de minería de datos
  - Conocer los problemas clásicos de minería de datos



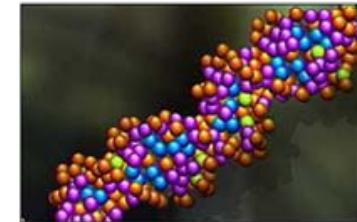
Ben Chams - Fotolia

# Índice

- 
- ¿Qué es la Ciencia de Datos?
  - Minería de Datos
  - Técnicas de Minería de Datos
  - Herramientas y Lenguajes en Ciencia de Datos.

# Nuestro mundo gira en torno a los datos

- Ciencia
  - Bases de datos de astronomía, genómica, datos medio-ambientales, datos de transporte, ...
- Ciencias Sociales y Humanidades
  - Libros escaneados, documentos históricos, datos sociales, ...
- Negocio y Comercio
  - Ventas de corporaciones, transacciones de mercados, censos, tráfico de aerolíneas, ...
- Entretenimiento y Ocio
  - Imágenes en internet, películas, ficheros MP3, ...
- Medicina
  - Datos de pacientes, datos de escaner, radiografías ...
- Industria, Energía, ...
  - Sensores, ...



# Motivación

El problema de la explosión de información:

- existencia de herramientas para la recolección de información
  - madurez de la tecnología de bases de datos
  - bajo precio del hardware
- ➔ gigantescas cantidades de datos almacenados en bases de datos, *data warehouses* y otros tipos de almacenes de información

**Somos ricos en datos pero pobres en conocimiento**

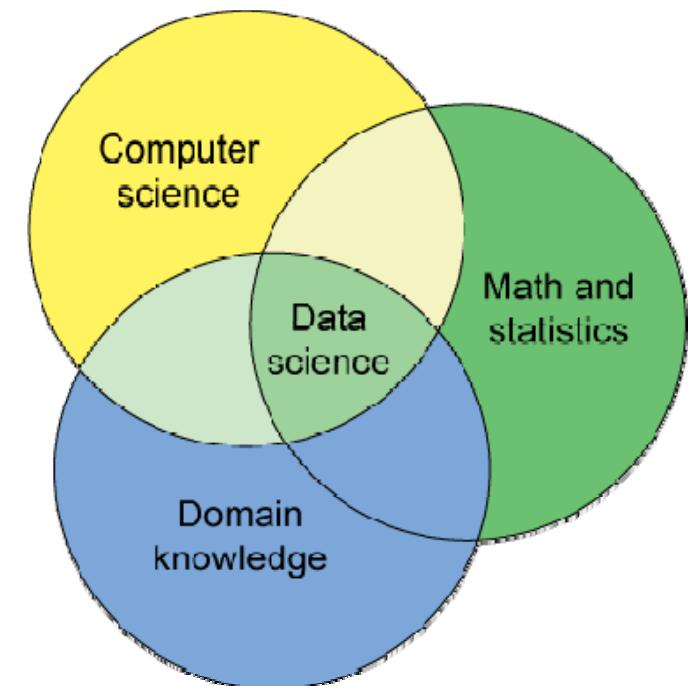
El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

## ¿Qué es la Ciencia de Datos -Data Science-?

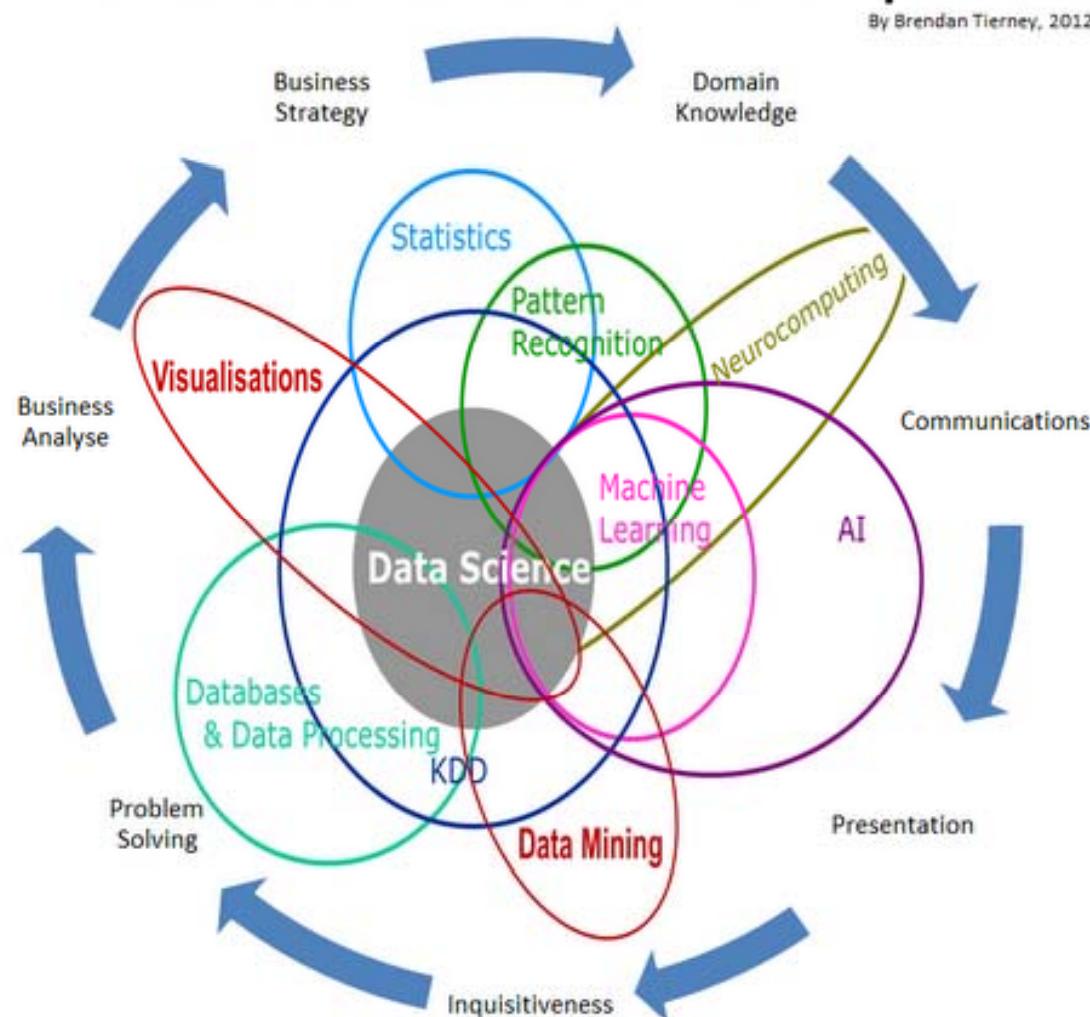
Es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos

## ¿Qué es un Científico de Datos?

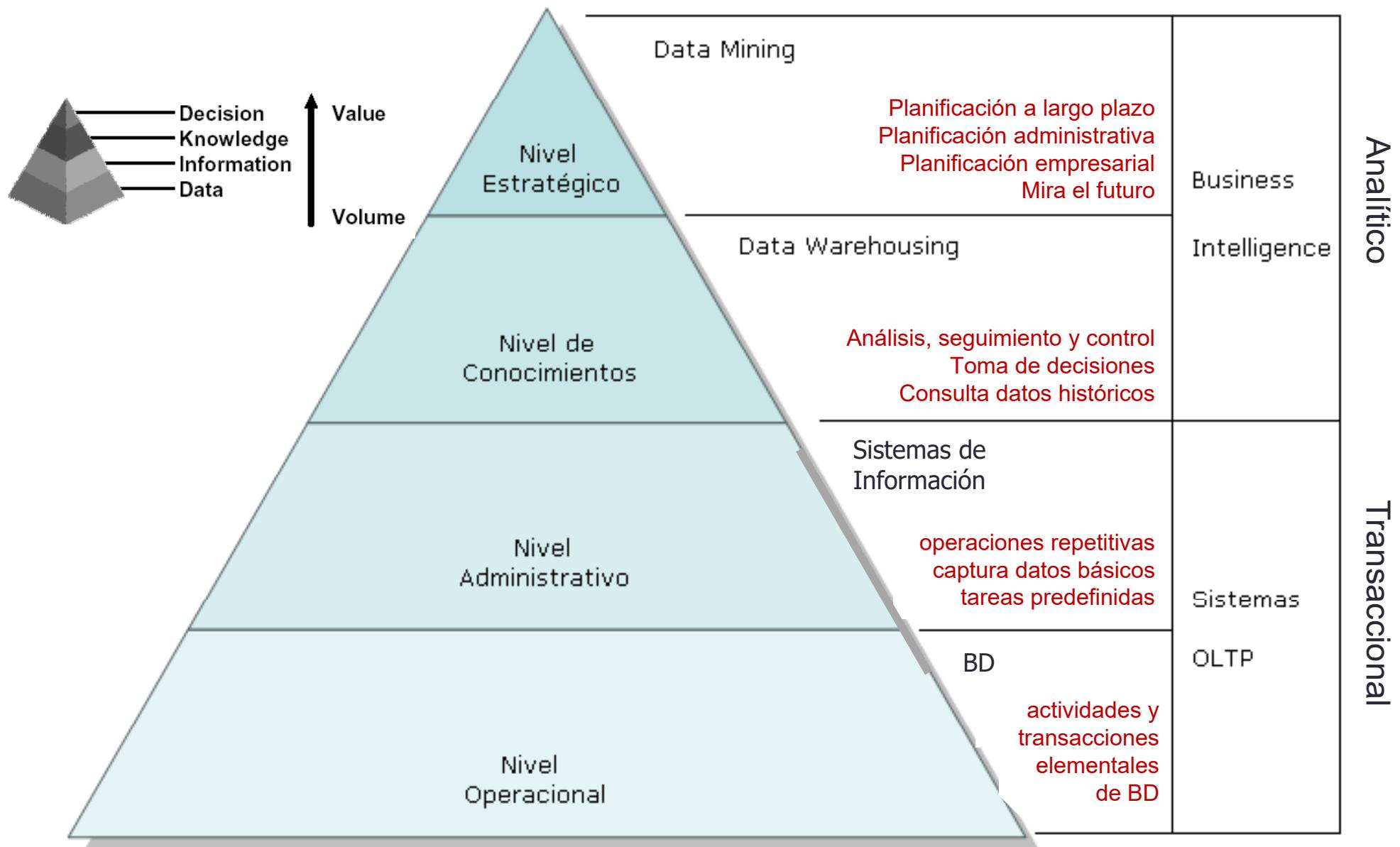
Un profesional que debe dominar las Ciencias Matemáticas y la Estadística, Ciencias de la Computación y analítica del dominio específico del problema abordado.



# Data Science Is Multidisciplinary



# Contexto



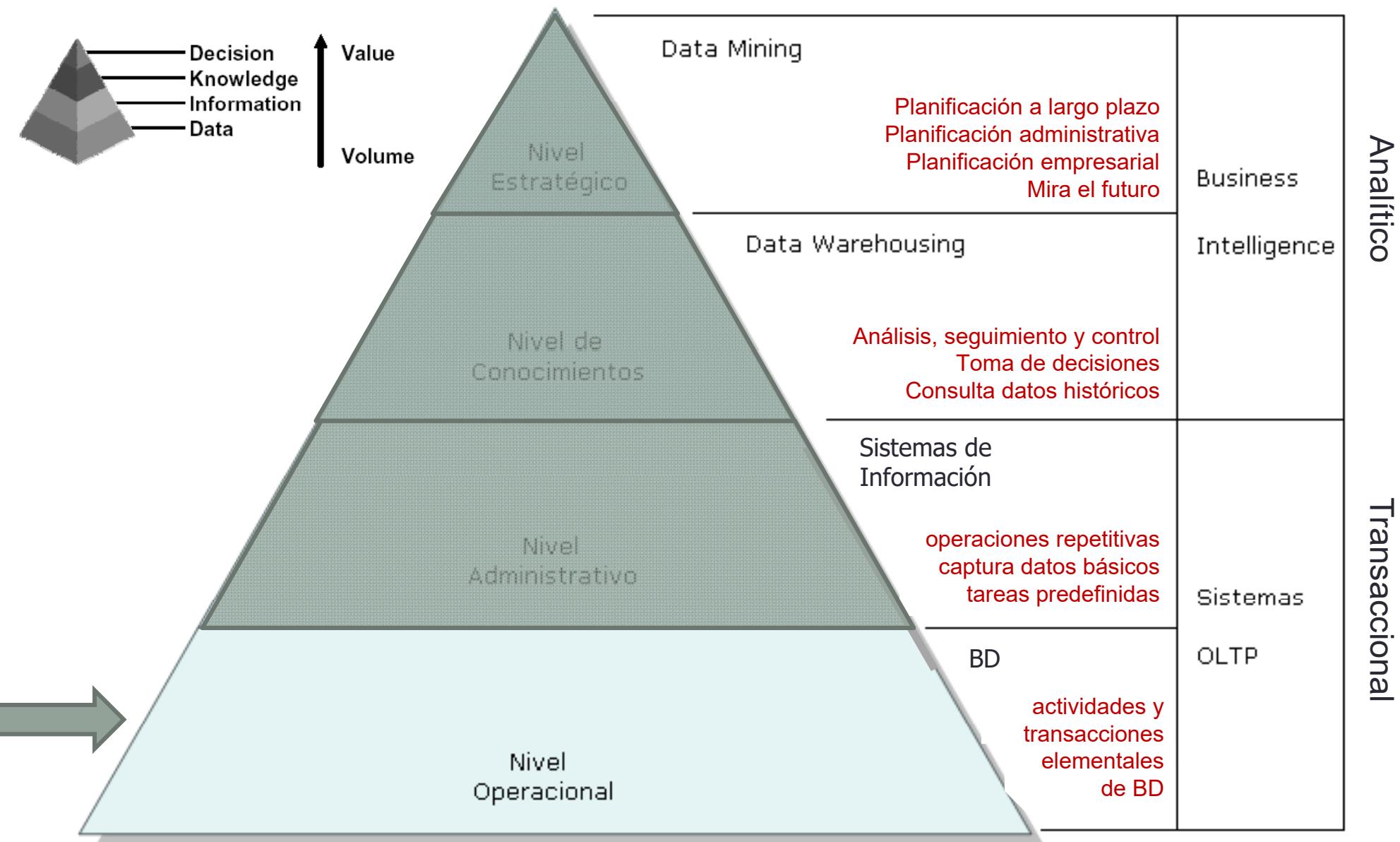


Tabla *Store\_Information*

Store_Name	Sales	Txn_Date
Los Angeles	1500	05-Jan-1999
San Diego	250	07-Jan-1999
Los Angeles	300	08-Jan-1999
Boston	700	08-Jan-1999

```
SELECT Store_Name FROM Store_Information;
```

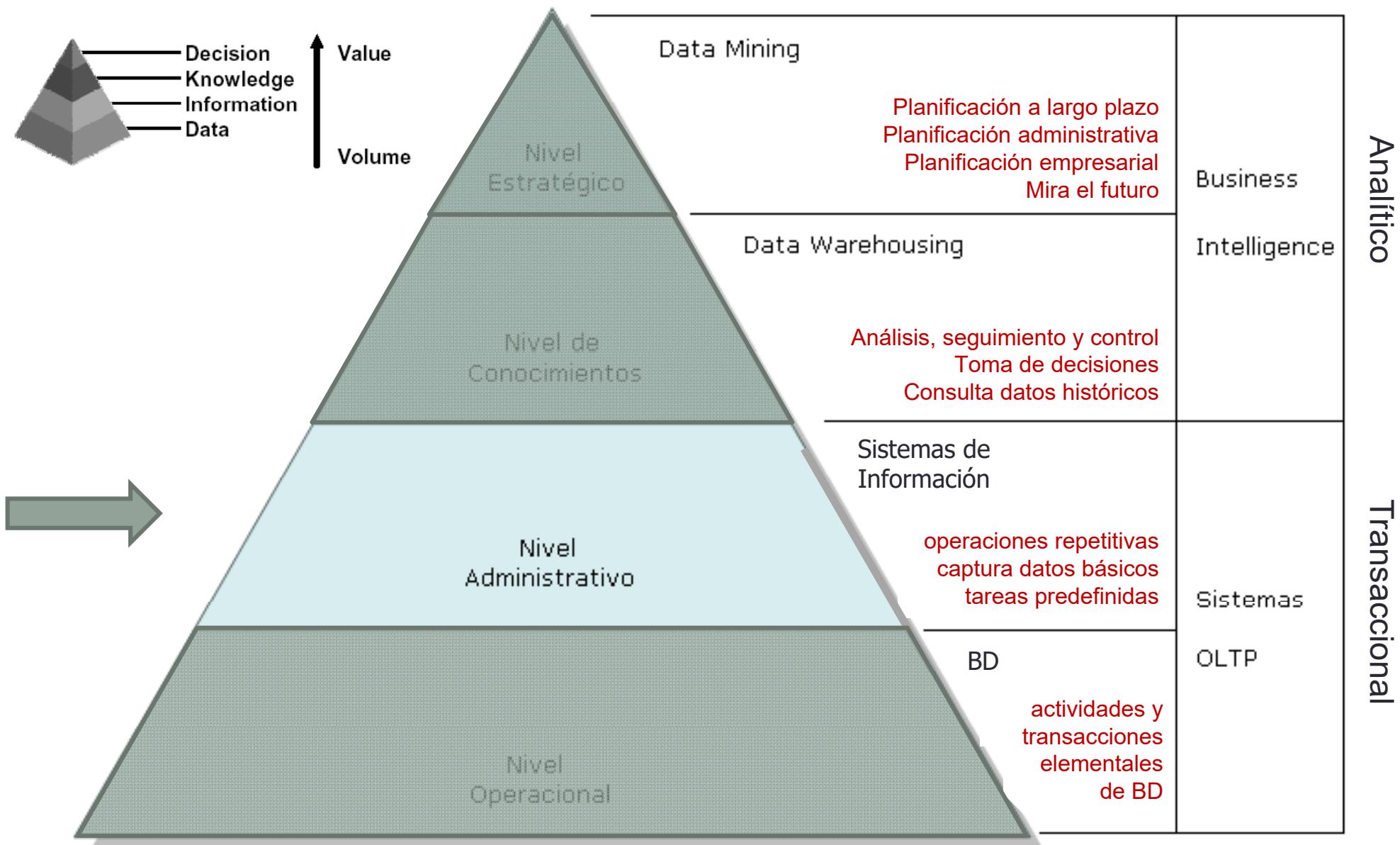
Store\_Name  
Los Angeles  
San Diego  
Los Angeles  
Boston

```
SELECT Store_Name  
FROM Store_Information  
WHERE Sales > 1000;
```

Store\_Name  
Los Angeles

```
SELECT Store_Name, SUM(Sales)  
FROM Store_Information  
GROUP BY Store_Name;
```

<u>Store_Name</u>	<u>SUM(Sales)</u>
Los Angeles	1800
San Diego	250
Boston	700





ERP: Integra en un mismo sistema todas las áreas involucradas en un negocio, pero no hay *inteligencia*

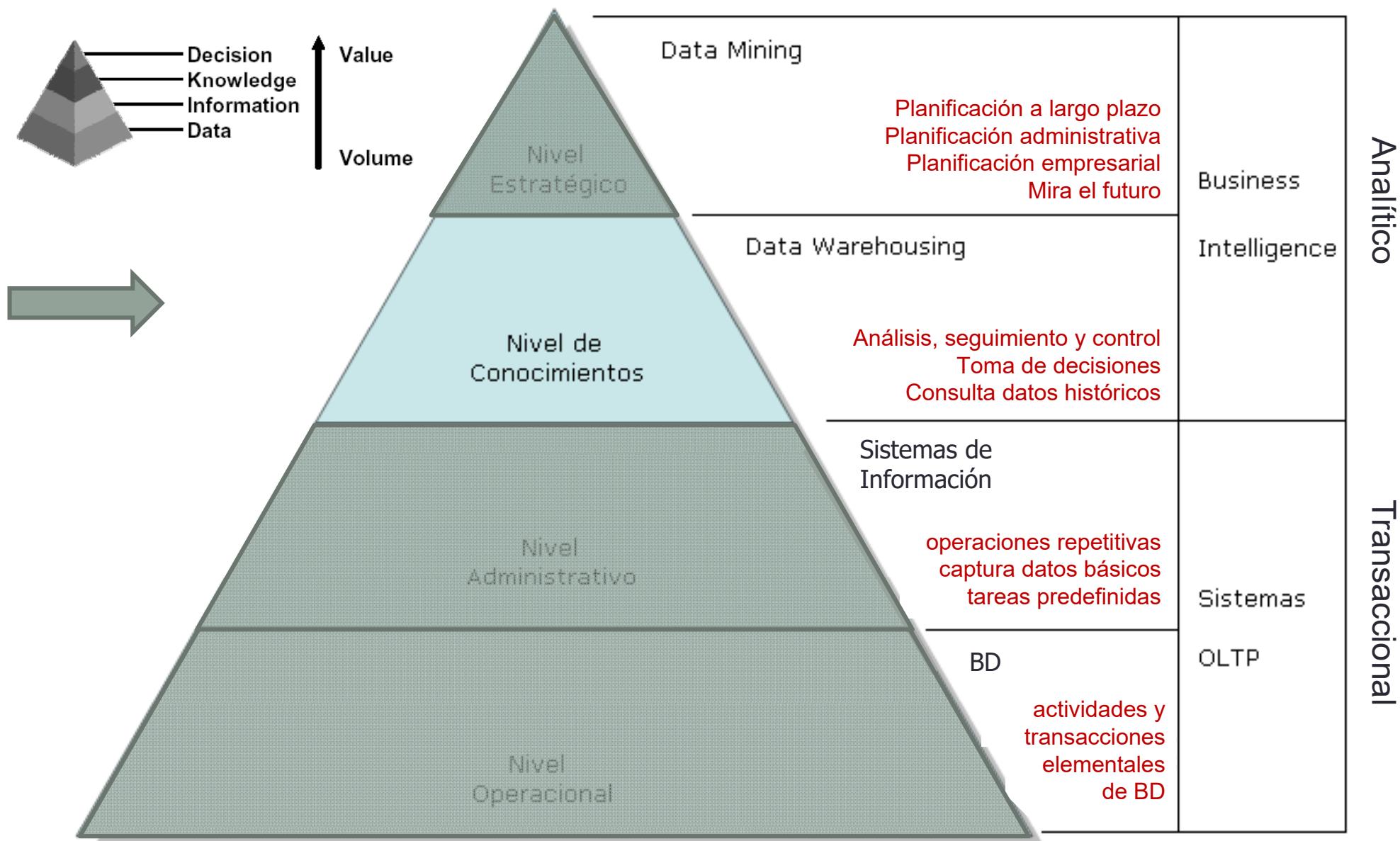


Tabla *Store\_Information*

Store_Name	Sales	Txn_Date
Los Angeles	1500	05-Jan-1999
San Diego	250	07-Jan-1999
Los Angeles	300	08-Jan-1999
Boston	700	08-Jan-1999

Store\_Name  
Los Angeles  
San Diego  
Los Angeles  
Boston

```
SELECT Store_Name FROM Store_Information;
```

```
SELECT Store_Name  
FROM Store_Information  
WHERE Sales > 1000;
```

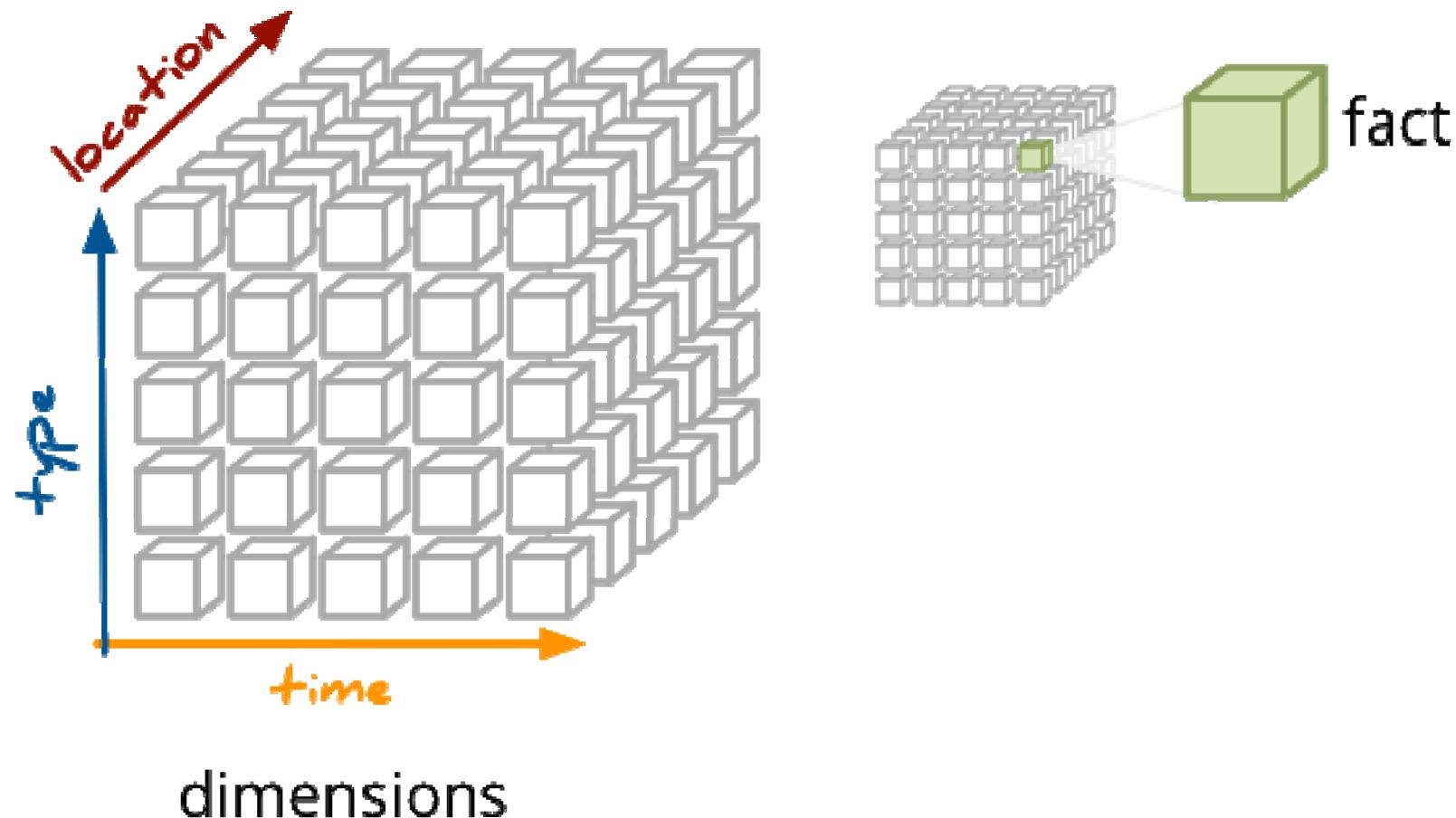
Store\_Name  
Los Angeles

```
SELECT Store_Name, SUM(Sales)  
FROM Store_Information  
GROUP BY Store_Name;
```

Store\_Name SUM(Sales)  
Los Angeles 1800  
San Diego 250  
Boston 700

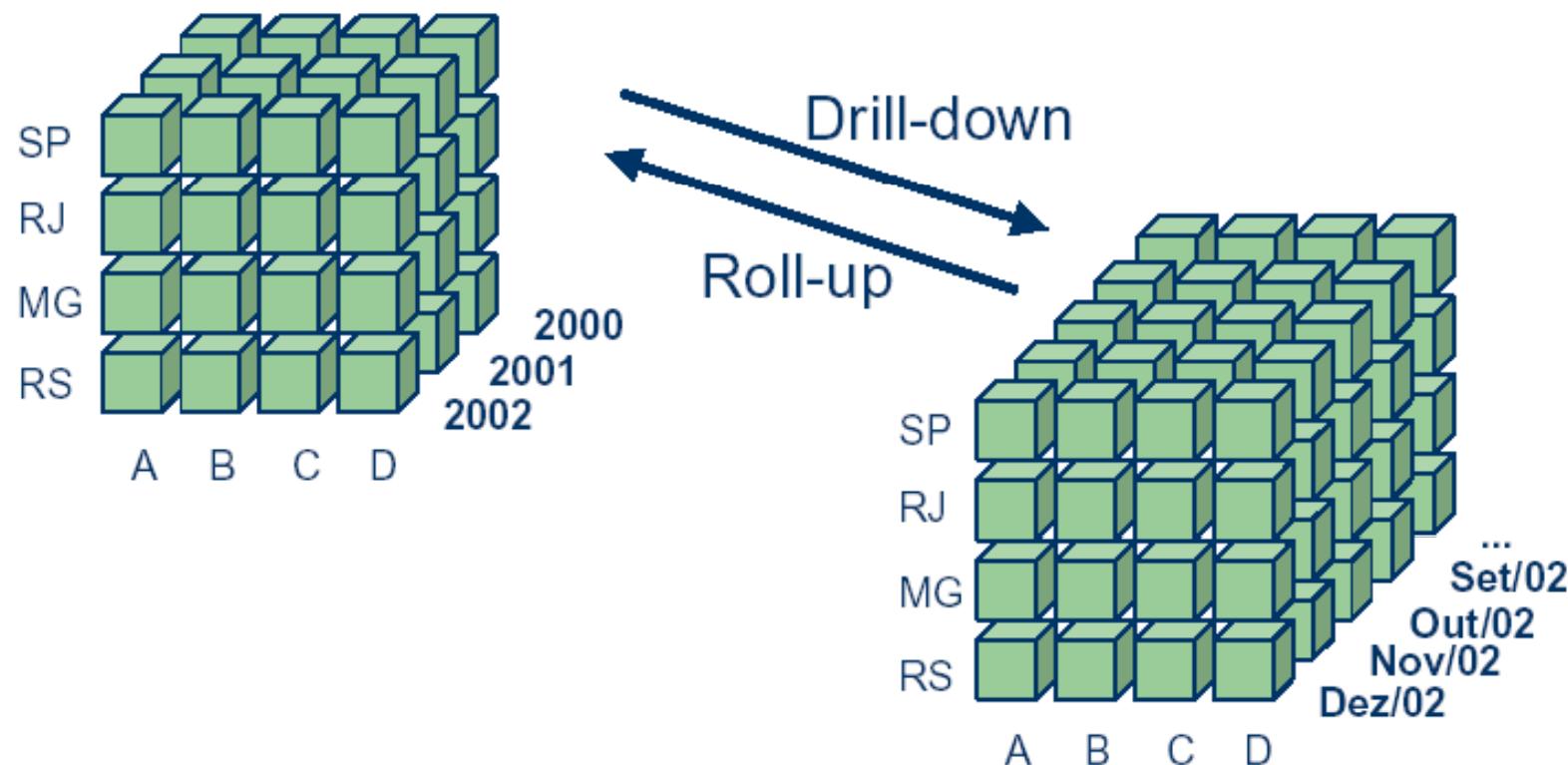
# OLAP/Data Warehousing

## Hechos y dimensiones

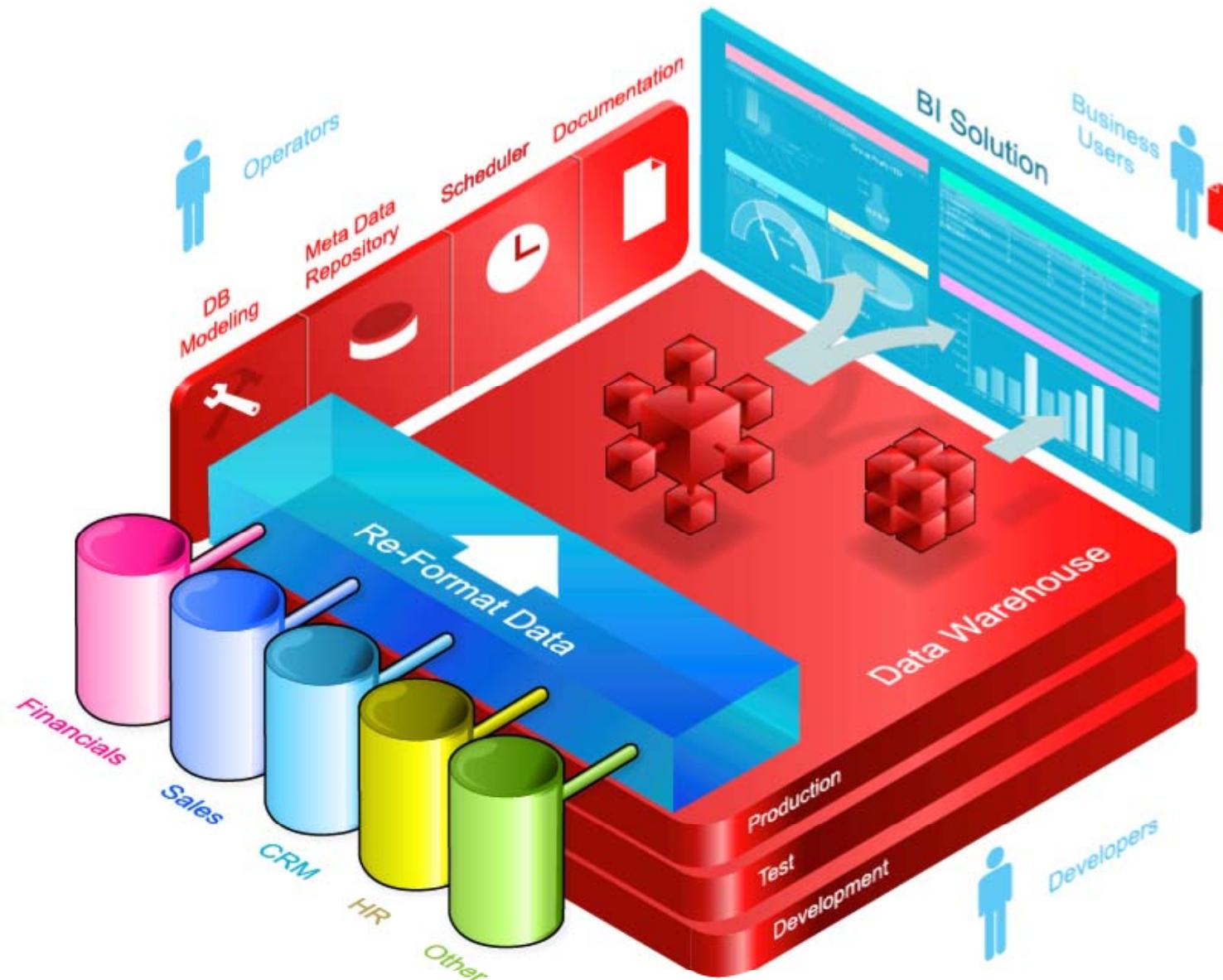


# OLAP/Data Warehousing

## Agregación



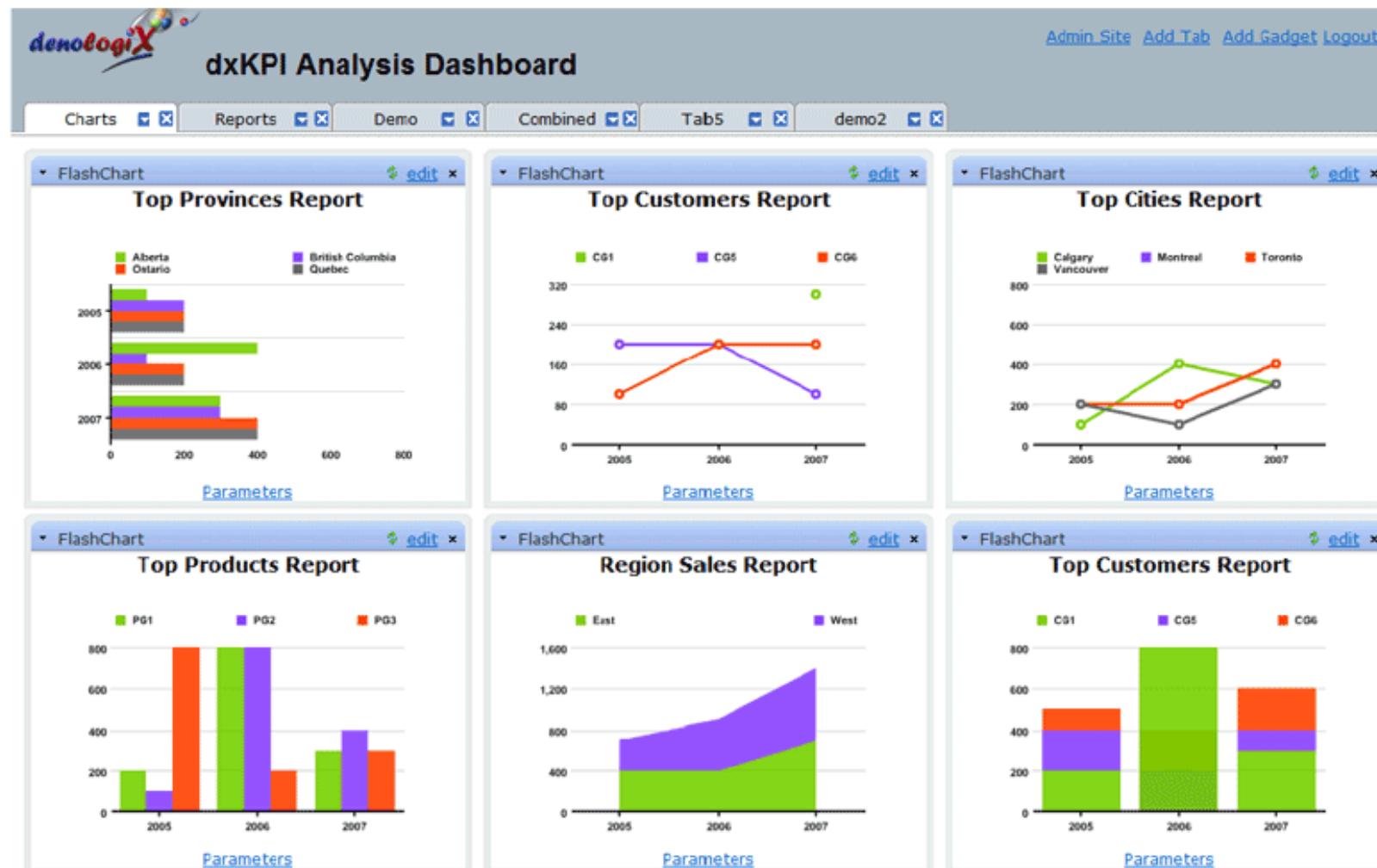
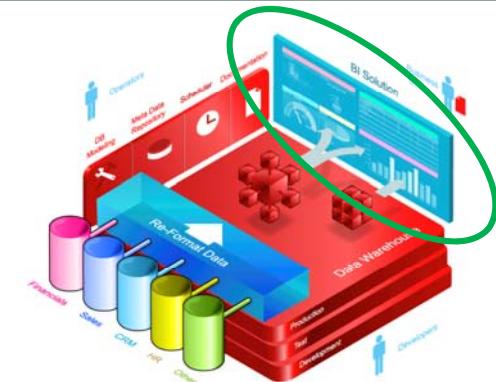
# OLAP/Data Warehousing

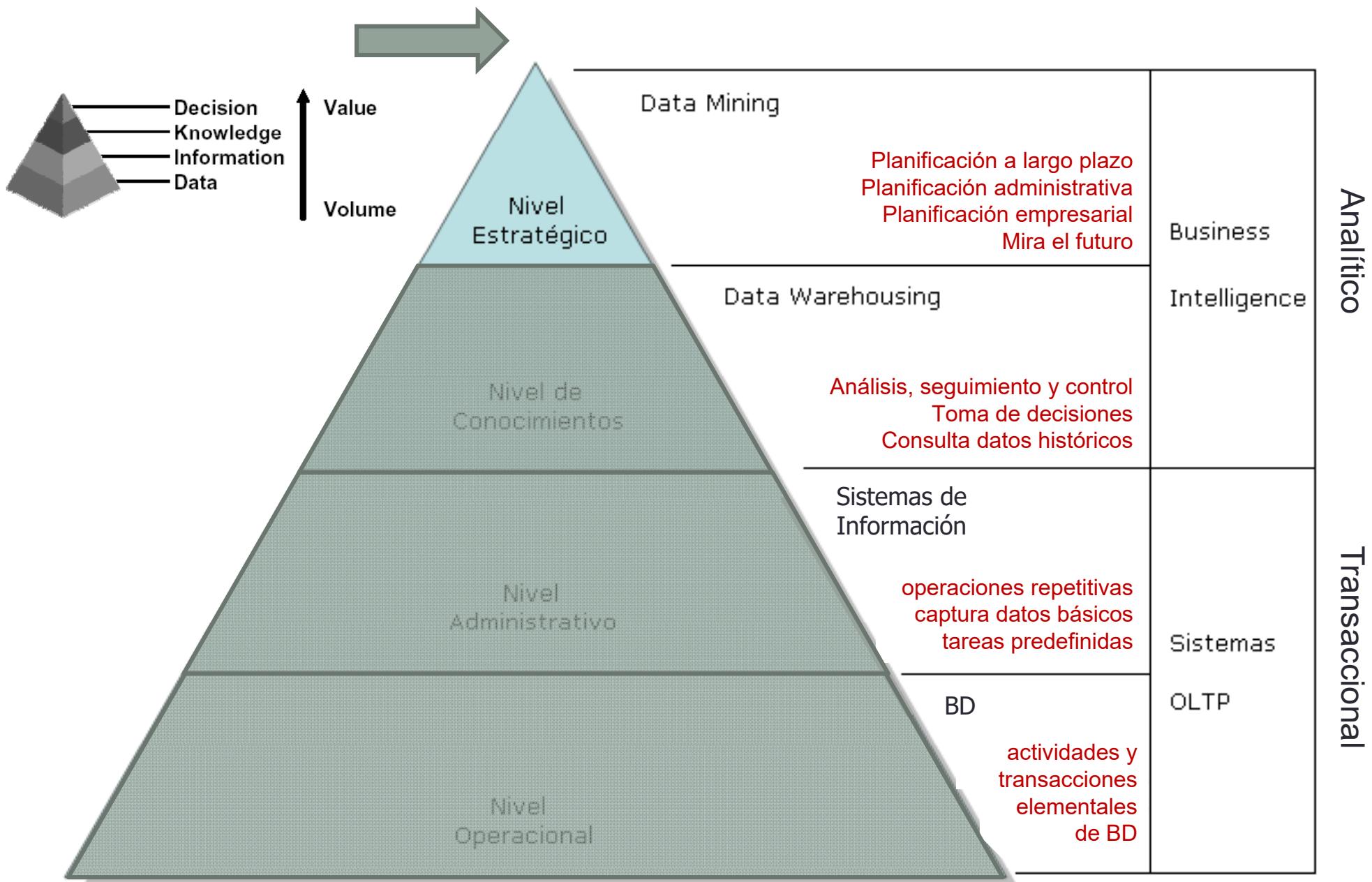


# OLAP/Data Warehousing

ScoreBoards/KPI (Key Performance Indicator)

**Real-time access to your KPIs**





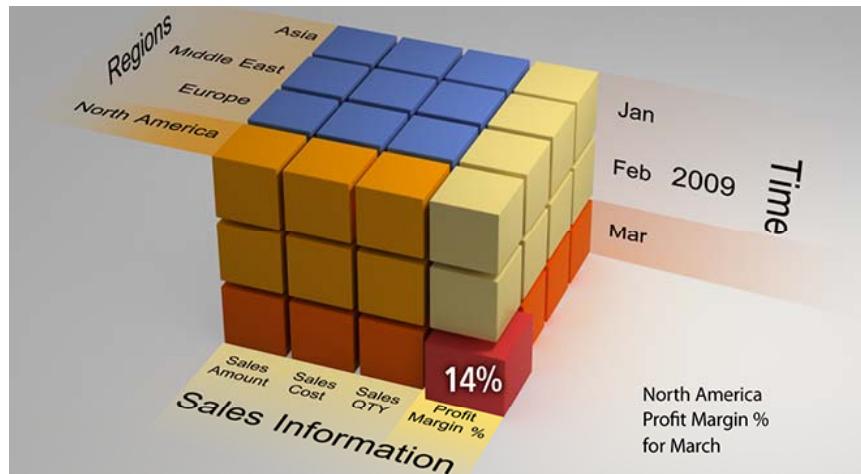
## Sistemas OLTP

Outcome	Type	Velocity	pfx_pfxX	pfx_pfxZ	Loc_X	Loc_Z	Count
LD	CH	87.3	-7.7	3.9	-0.3	2.4	11
FB	CH	87.5	-8.5	4.3	-0.2	2.5	10
GB	CH	87.2	-7.9	3.6	-0.4	2.4	37
ball	CH	87.7	-8.2	4.0	0.1	1.7	104
StrF	CH	87.7	-7.6	3.8	-0.3	2.1	56
StrL	CH	87.5	-8.4	3.7	-0.2	2.6	39
StrS	CH	88.0	-7.9	3.6	-0.3	1.8	58
LD	CU	82.1	5.7	-7.0	-0.2	2.1	6
FB	CU	82.2	5.8	-6.1	-0.2	2.2	7
GB	CU	82.4	5.6	-7.4	-0.1	2.2	12
ball	CU	82.1	5.8	-6.6	-0.6	2.3	86
StrF	CU	82.4	6.0	-6.5	-0.1	2.3	21
StrL	CU	81.8	6.1	-6.1	-0.3	2.6	59
StrS	CU	83.8	4.7	-7.4	0.4	1.2	14
LD	FA	95.2	-7.8	7.6	-0.3	2.5	59
FB	FA	94.9	-6.9	8.4	-0.1	2.7	92
GB	FA	94.9	-7.8	7.5	-0.3	2.4	140
ball	FA	94.9	-7.4	8.1	-0.2	2.4	527
StrF	FA	95.2	-7.1	8.0	-0.4	2.6	321
StrL	FA	94.5	-7.9	7.9	-0.2	2.5	264
StrS	FA	95.4	-6.5	8.4	-0.3	2.6	89
LD	SL	86.2	1.1	-1.1	0.4	2.2	8
FB	SL	87.9	1.5	-0.9	0.2	2.4	9
GB	SL	87.1	1.0	-1.1	0.1	2.4	17
ball	SL	87.7	0.5	-0.7	0.8	2.0	100
StrF	SL	87.1	0.8	-1.1	0.2	2.3	33
StrL	SL	87.0	1.5	-0.9	0.1	2.7	36
StrS	SL	87.7	1.0	-1.1	0.8	1.6	47

## Data Analysis

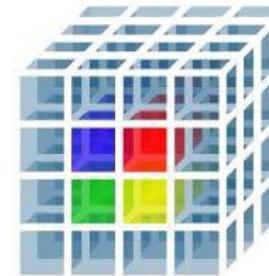
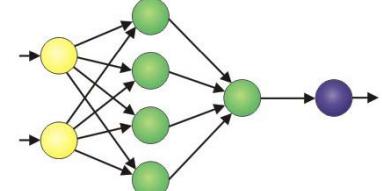


## Data Warehousing/OLAP



## Data Mining



Plazo	Uso	Técnica	Tecnología	Tecnología	Conocimiento
Corto Plazo	Gestión de datos Obtención y control ¿Total de ventas en Granada?	Legacy Systems	OLTP On-Line Transaction Processing		Datos Operativo
Mediano Plazo	Decisiones tácticas ¿Total de ventas en Granada por cuatrimestres y por categoría de producto?	Data Warehouse	OLAP On-Line Analytical Processing		Información Toma de Decisiones
Largo Plazo	Estratégico, Pronóstico ¿Cómo evolucionarán las ventas el próximo año en Granada?	Minería de Datos	Agrupamiento Clasificación Secuenciación Reglas de asociación		Patrones Nuevos Conocimientos

# Índice

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
  - ❑ Técnicas de Minería de Datos
  - ❑ Herramientas y Lenguajes en Ciencia de Datos.

**“The key in business is to know something that nobody else knows.”**

— Aristotle Onassis



PHOTO: [HULTON-DEUTSCH COLL](#)

**“To understand is to perceive pattern**



PHOTO: [LUCINDA DOUGLAS-MENZIES](#)

— Sir Isaiah Berlin

The screenshot shows the homepage of CincoDías, a Spanish business newspaper. At the top, there's a teal header bar with the website's name and a navigation menu. Below the header, the main title of the article is displayed, followed by a list of bullet points summarizing its content. At the bottom, there's a footer section with author information and a timestamp.

Recorte rectangular

# CincoDías

MIÉRCOLES, 19 DE OCTUBRE DE 2016

Inicio    Mercados    Empresas    Economía    Tecnología

ESTÁ PASANDO: IBEX 35    Calendario laboral 2016-2017    Declaración IVA    Elecciones EE

*Tribuna*

## *El 'big data' se dedica a simplificar*

- Hay que tener clara la información que necesitamos y cuándo se convierte en conocimiento que nos ayuda a tomar las decisiones correctas

DAVID CASCANT | 06-06-2016 21:32

<http://flip.it/yUAnZ>

# Minería de Datos. ¿Qué es?

La Minería de datos (MD) es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos



También se conoce como:

- Descubrimiento de conocimiento en bases de datos (KDD),
  - extracción del conocimiento,
  - análisis inteligente de datos /patrones,
  - ...

# Minería de Datos. ¿Qué es?

- Muchas de las técnicas utilizadas en MD ya se conocían previamente, ¿a qué se debe?
- En los 90's convergen los siguientes factores:
  1. Los datos se están produciendo
  2. Los datos se están almacenando
  3. La potencia computacional necesaria es abordable
  4. Existe una gran presión competitiva a nivel empresarial
  5. Las herramientas software de MD están disponibles



# Minería de Datos. ¿Qué es?

*¿Para qué se utiliza el ‘conocimiento’ obtenido?*

- hacer predicciones sobre nuevos datos
- explicar los datos existentes
- resumir una base de datos masiva para facilitar la toma de decisiones
- visualizar datos altamente dimensionales, extrayendo estructura local simplificada, ...

**Nuevas necesidades de análisis datos**

# Minería de Datos. ¿Qué es?

Minería de datos NO es:

- Procesamiento deductivo de consultas en bases de datos
- Un sistema experto
- Análisis estadístico
- Visualización de datos
- Pequeños programas de aprendizaje

# Minería de Datos. ¿Qué es?

*¿A qué tipos de datos puede aplicarse DM?*

En principio, a cualquier tipo

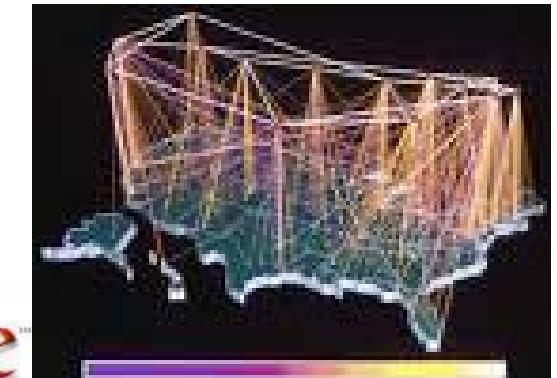
- Bases de datos relacionales
- Bases de datos espaciales
- Bases de datos temporales
- Bases de datos documentales (**Text mining**)
- Bases de datos multimedia
- World Wide Web (**Web mining**)
  - El almacén de información más grande y diverso de los existentes
  - Existe gran cantidad de datos de los que extraer información útil
- .... **Grandes volúmenes de datos: Big Data**

# Minería de Datos. ¿Qué es?

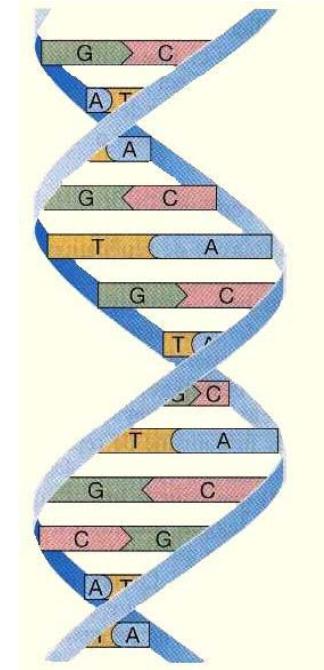
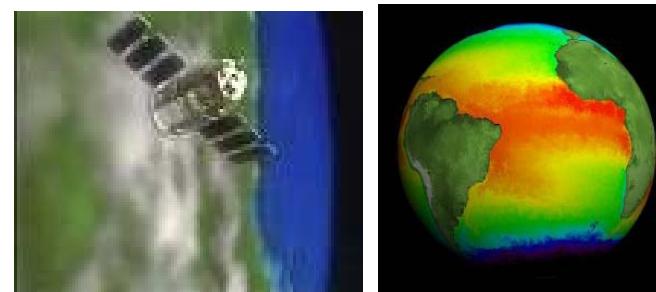
amazon.com®

Tera/Peta bytes de datos:

- Compras relacionadas.
- Perfiles de usuario en la Web
- Segmentación de clientes
- Detección de intrusos / Fraudes
- Pronóstico Tiempo
- Predicción estructura ADN
- etc



Google™



# Minería de Datos. Áreas de aplicación

## Análisis y gestión de mercados (I)

- *Fuentes:* transacciones con tarjetas de crédito, tarjetas de descuento, quejas de cliente, estilos de vida publicados, comentarios en redes sociales...
- *Identificación de objetivos para marketing:* encontrar grupos (*clusters*) que identifiquen un modelo de cliente con características comunes (intereses, nivel de ingresos, hábitos de gasto, ...)

# Minería de Datos. Áreas de aplicación

## Análisis y gestión de mercados (II)

- *Análisis de cestas de mercado:* asociaciones / co-relaciones entre ventas de producto, predicción basada en asociación de informaciones,...
- *Perfiles de cliente:* Identificar qué tipo de clientes compra qué productos (*clustering* y/o clasificación), usar predicción para encontrar factores que atraigan nuevos clientes, retención de clientes,...
- *Generar información resumida:* informes multidimensionales, información estadística (tendencia central y variación), ...

# Minería de Datos. Áreas de aplicación

## Análisis de riesgo en banca y seguros

- Banca
  - Detectar patrones de uso fraudulento en tarjetas
  - Estudio de concesión de créditos y/o tarjetas
  - Determinación del gasto en tarjeta por grupos
  - Identificar reglas de comportamiento del mercado de valores a partir de históricos
- Seguros
  - Predicción de clientes propensos a suscribir nuevas pólizas
  - Identificar grupos/patrones de riesgo
  - Identificar tendencias de comportamiento fraudulento
- Ambos: Identificación de clientes leales, identificación de fuga de clientes

# Minería de Datos. Áreas de aplicación

## Minería de datos en industria

- Control de calidad
  - Detección precisa de productos defectuosos
  - Localización precoz de defectos
  - Identificación de causas de fallos
- Procesos industriales
  - Automatizar el control del proceso
  - Optimización del rendimiento de forma adaptativa
  - Implementar programas de mantenimiento predictivo

# Minería de Datos. Áreas de aplicación

## Web mining / minería de datos web

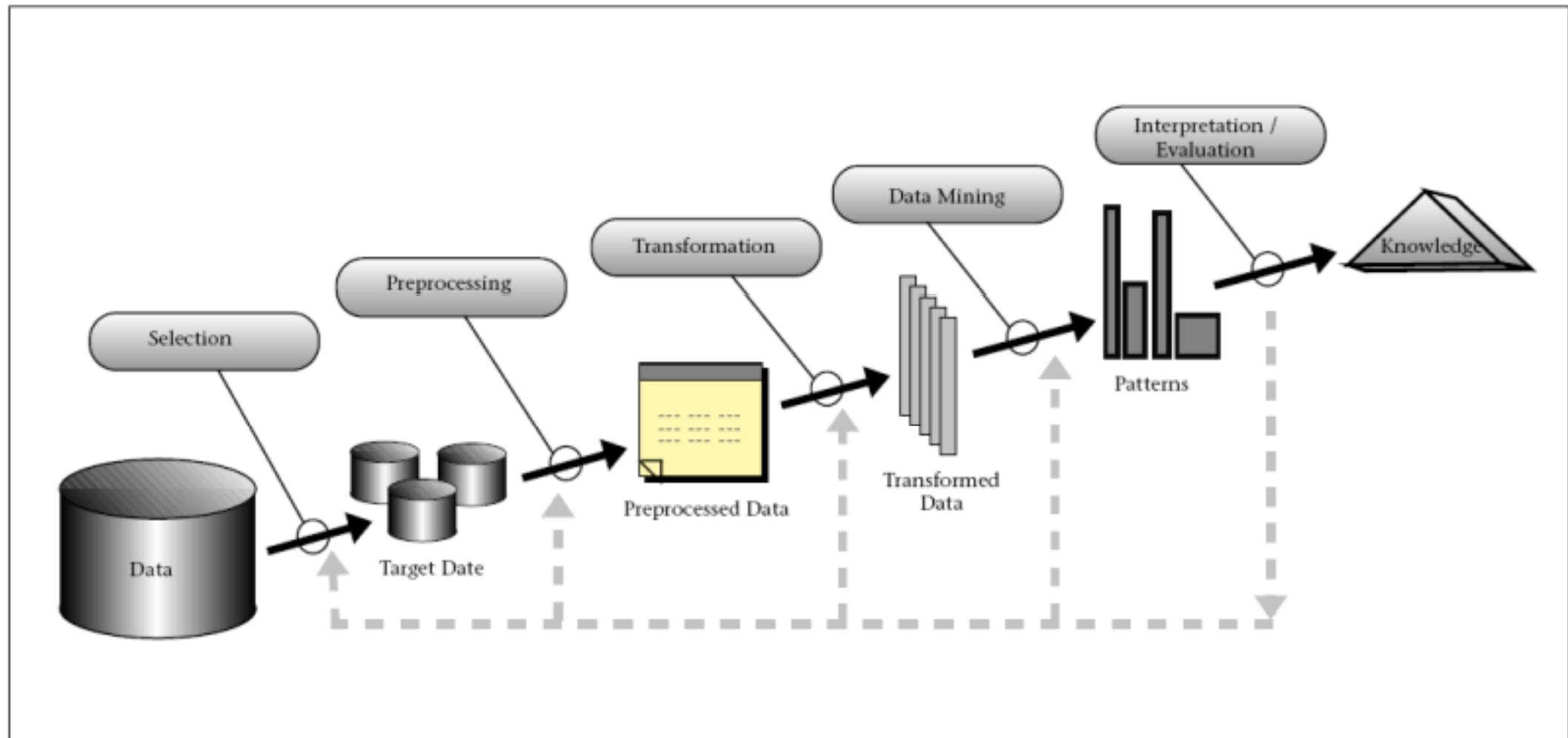
- Análisis del comportamiento y perfiles del visitante
- Potenciar la venta cruzada (cross-selling)
- Generación de respuestas agrupadas según el tipo de contenido
- Recuperación de información (information retrieval) Búsqueda de metadatos que describan los documentos.
- Recuperación inteligente de datos complejos (texto, imágenes, etc)
- Análisis de grupos en redes sociales

# Minería de Datos. Áreas de aplicación

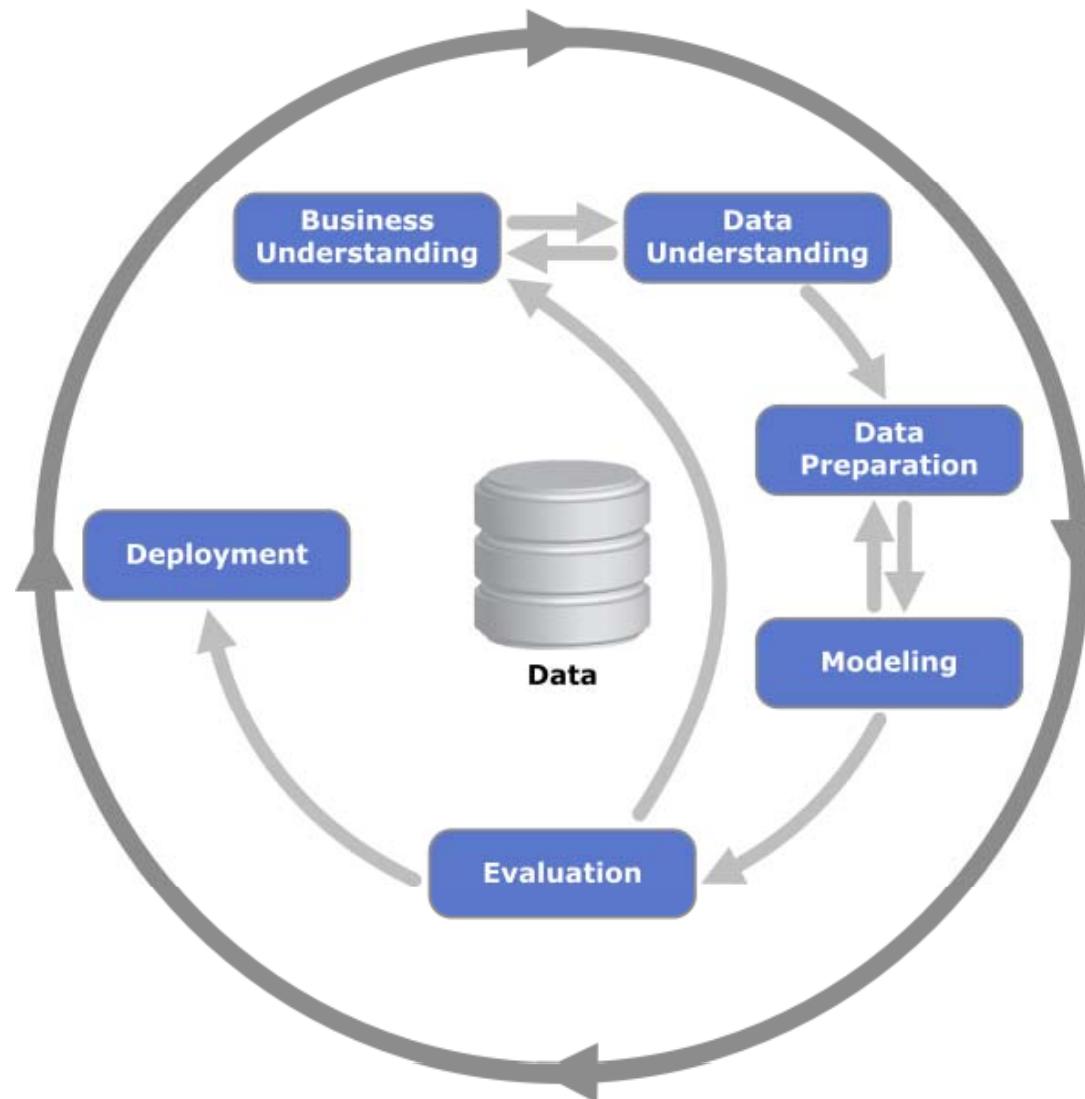
## Medicina / diagnóstico

- Identificación de terapias para diferentes enfermedades
- Estudio de factores de riesgo en distintas patologías
- Segmentación de pacientes en grupos afines
- Gestión hospitalaria y planificación temporal de salas, urgencias,...
- Recomendación priorizada de fármacos para una misma patología
- Estudios en genética (ADN,...)

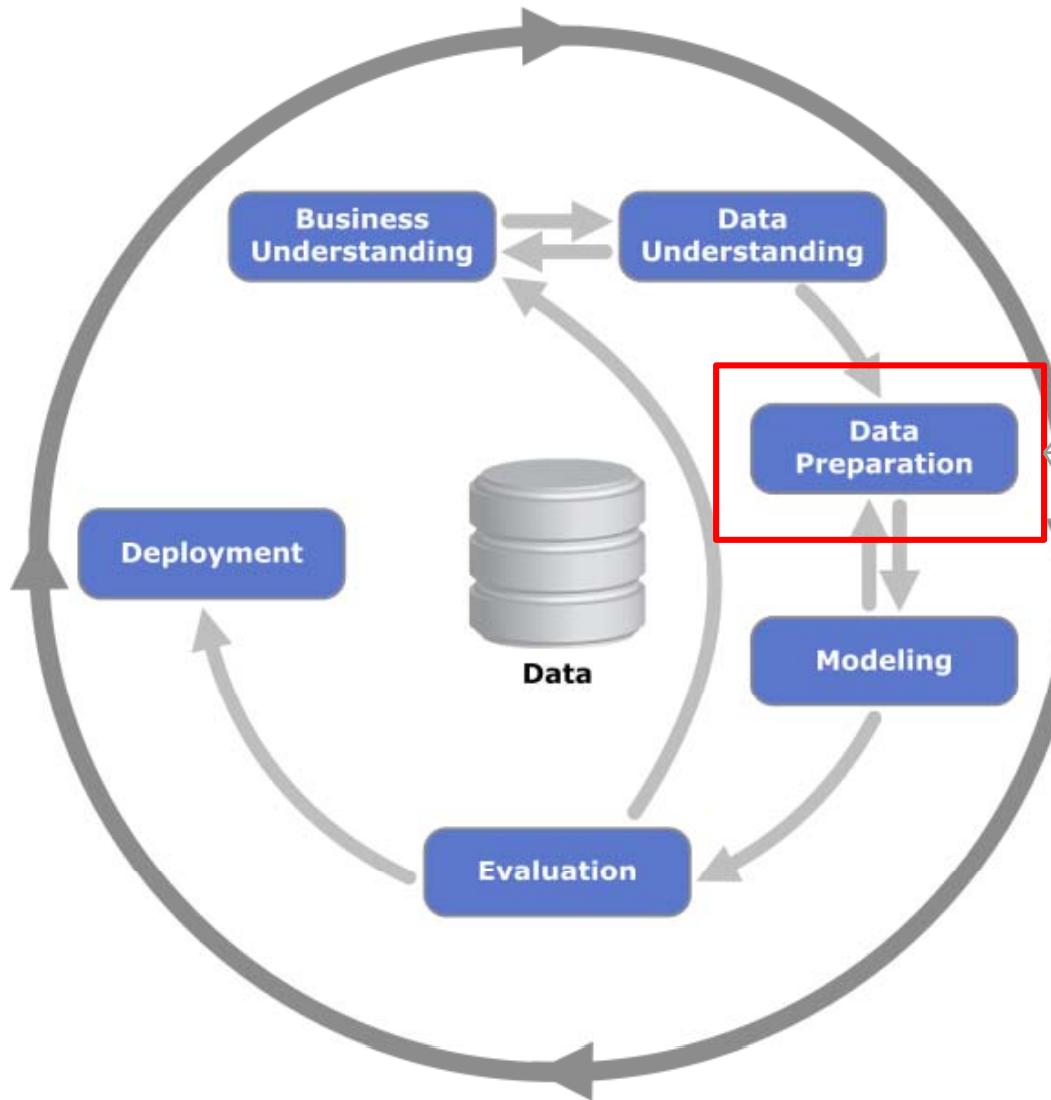
# Minería de Datos. Fases



# Minería de Datos. Fases



# Minería de Datos. Fases



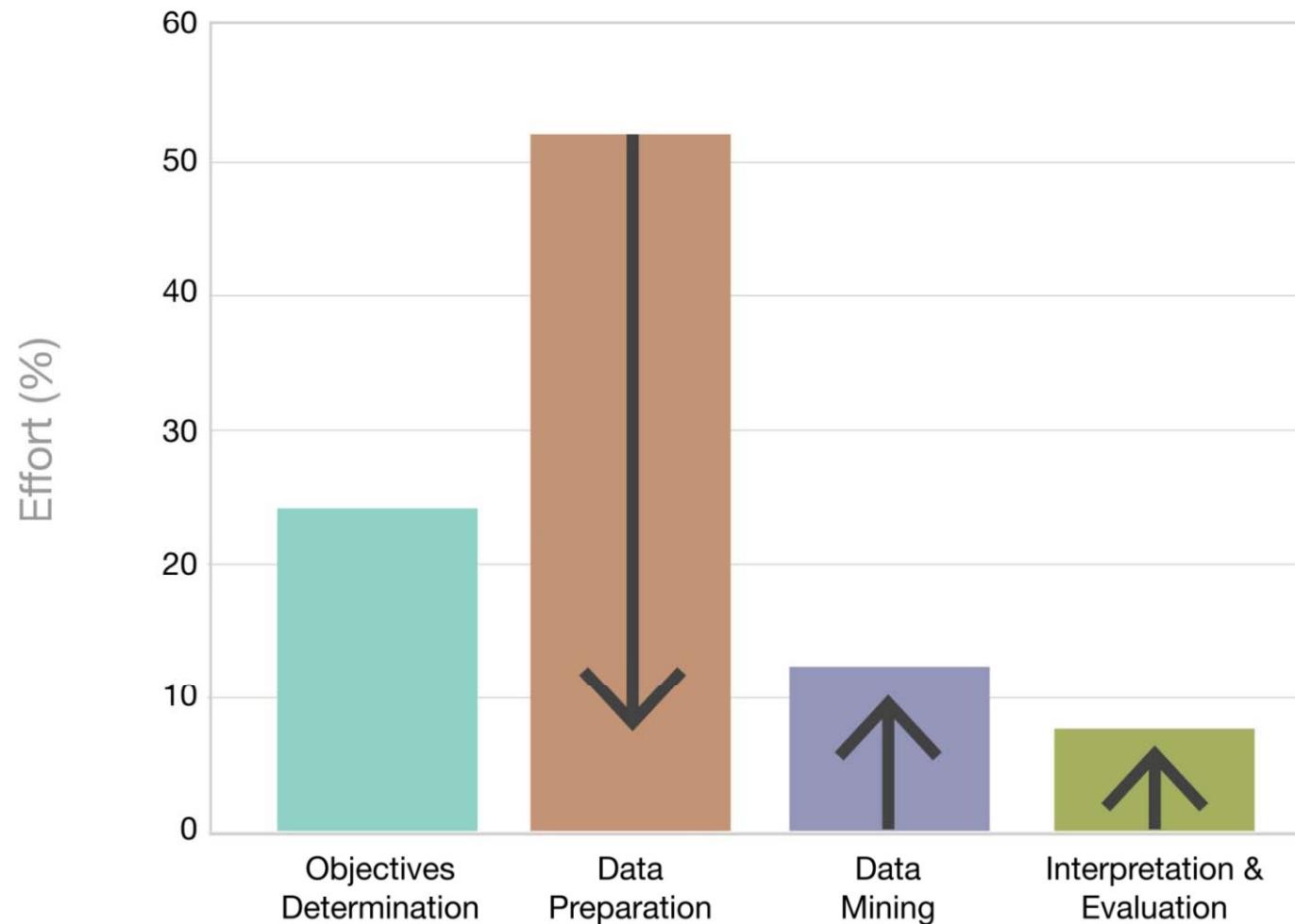
Cursos del Máster:

Introducción a la ciencia de datos

Minería de datos:  
preprocesamiento y  
clasificación

Minería de datos: aprendizaje  
no supervisado y detección de  
anomalías

# Minería de Datos. Fases

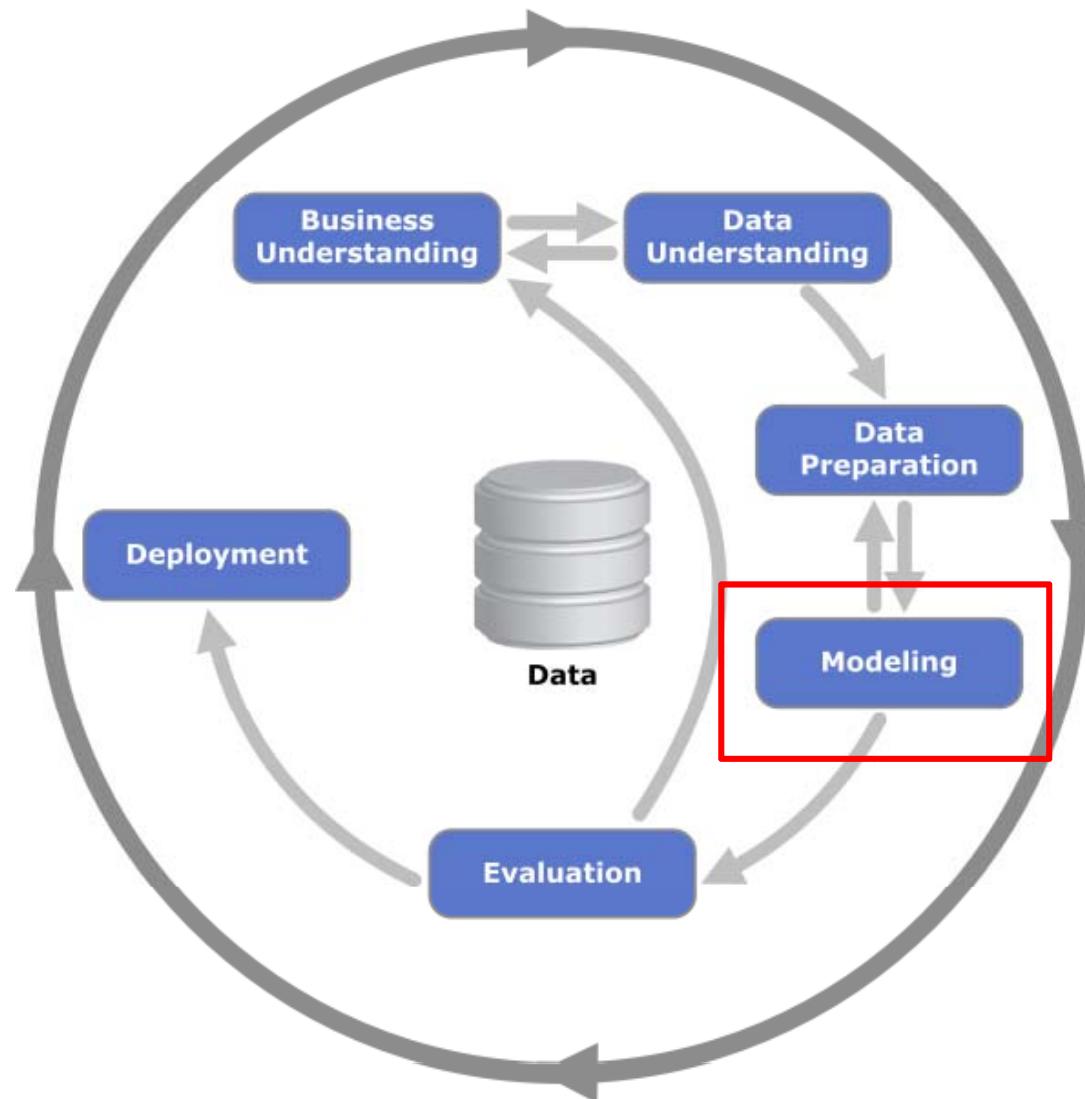


Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos

# Índice

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Técnicas de Minería de Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos.

# Minería de Datos. Modelos



# Minería de Datos. Modelos

Clasificación En función de su propósito general:

- ▶ Modelos descriptivos
- ▶ Modelos predictivos

## Modelos descriptivos

- Describen el comportamiento de los datos de una forma fácilmente interpretable.

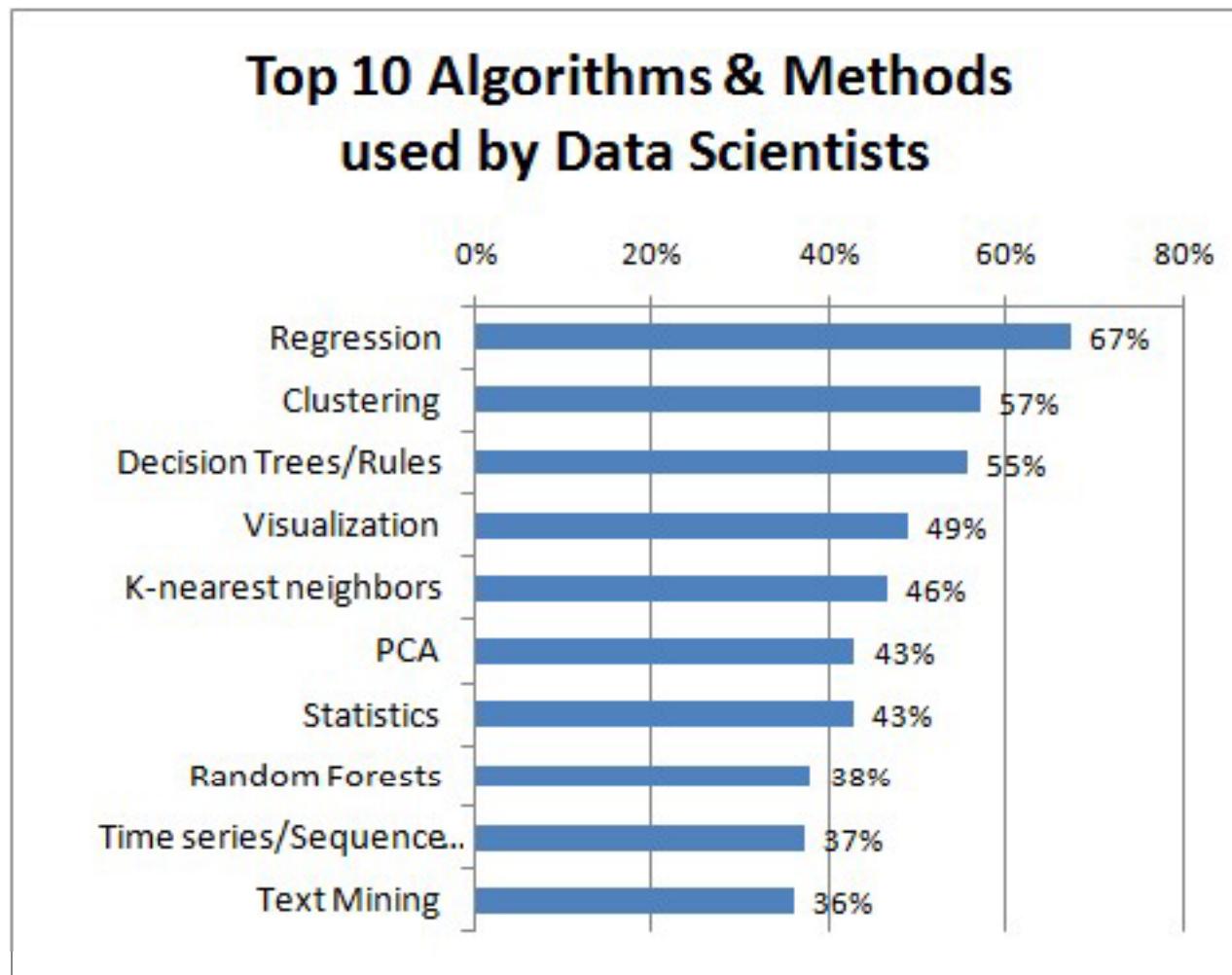
## Modelos predictivos

- Además de describir los datos, el modelo construido se usa para predecir el valor de algún atributo de una nueva entrada

# Minería de Datos. Modelos

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation/Anomaly Detection [Predictive]
- Time Series [Predictive]
- Summarization [Descriptive]

# Minería de Datos. Modelos

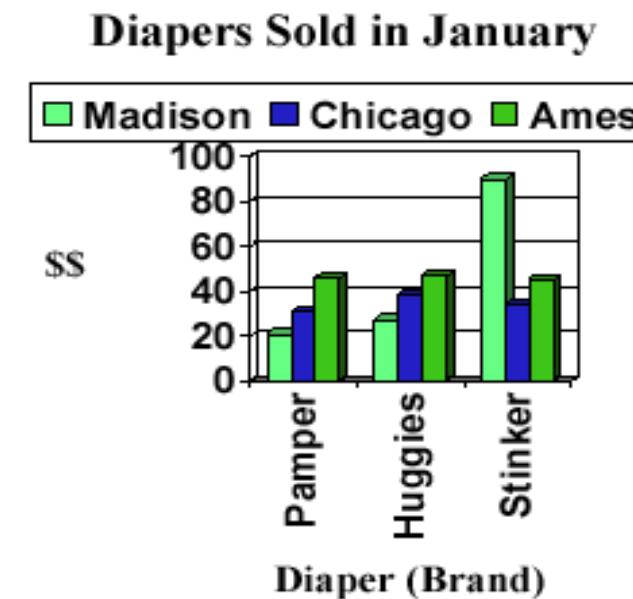


<http://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html>

# Minería de Datos. Modelos → Reglas de Asociación

Supongamos ventas de una tienda 24h.

Podemos plantear un cubo OLAP y ver los informes de ventas sobre cervezas y pañales por separado  
→ de poca utilidad



Más interesante: ¿Influye la venta de un producto en otro?

# Minería de Datos. Modelos → Reglas de Asociación

Modelo descriptivo

Asociación (Análisis de tendencias)

→ Market basket Analysis



Longitud variable

Transaction Id	Products Id
Madrid_3_2013_03_13_T0000134278	PK10056, TKN100UG, JG20045
Barcelona_4_2013_05_23_T259034439	PK10056, TKN100UG, UTR567, PLG345, UTG6003, JKOP345
Madrid_1_2013_04_15_T1779234445	TKN100UG, JG20045

# Minería de Datos. Modelos → Reglas de Asociación

- Association Rule
  - An expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets
  - Example:  
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$
- Rule Evaluation Metrics
  - Rule Support ( $s$ )
    - ◆ Fraction of transactions that contain both  $X$  and  $Y$
$$s(X \rightarrow Y) = s(XY) = \frac{\#(XY)}{|T|}$$
  - Confidence ( $c$ )
    - ◆ Measures how often items in  $Y$  appear in transactions that contain  $X$
$$c(X \rightarrow Y) = P(Y | X) = \frac{\#(XY)}{\#(X)}$$

<i>TID</i>	<i>Items</i>
1	<b>Bread, Milk</b>
2	<b>Bread, Diaper, Beer, Eggs</b>
3	<b>Milk, Diaper, Beer, Coke</b>
4	<b>Bread, Milk, Diaper, Beer</b>
5	<b>Bread, Milk, Diaper, Coke</b>

Example:

$$\{\text{Milk, Diaper}\} \rightarrow \text{Beer}$$

$$s = \frac{\#(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\#(\text{Milk, Diaper, Beer})}{\#(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

# Minería de Datos. Modelos → Reglas de Asociación

Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.

- ✓ ¿Qué significa?
- ✓ ¿A qué se debe?
- ✓ Acciones a realizar



# Minería de Datos. Modelos → Reglas de Asociación

Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.

- Se acerca el fin de semana
- Hay un bebé en casa
- No quedan pañales
- El padre/madre compra pañales al salir del trabajo
- ¡No pueden salir!
- Comprar cervezas para el fin de semana (y un partido/película PPV)

- Se acerca el fin de semana
- Hay un bebé en casa luego nada de ir fuera
- Hay que comprar pañales
- Quedarse en casa → ver partido/película
- Comprar cervezas para el partido/película

Pañales → Cerveza



# Minería de Datos. Modelos → Reglas de Asociación

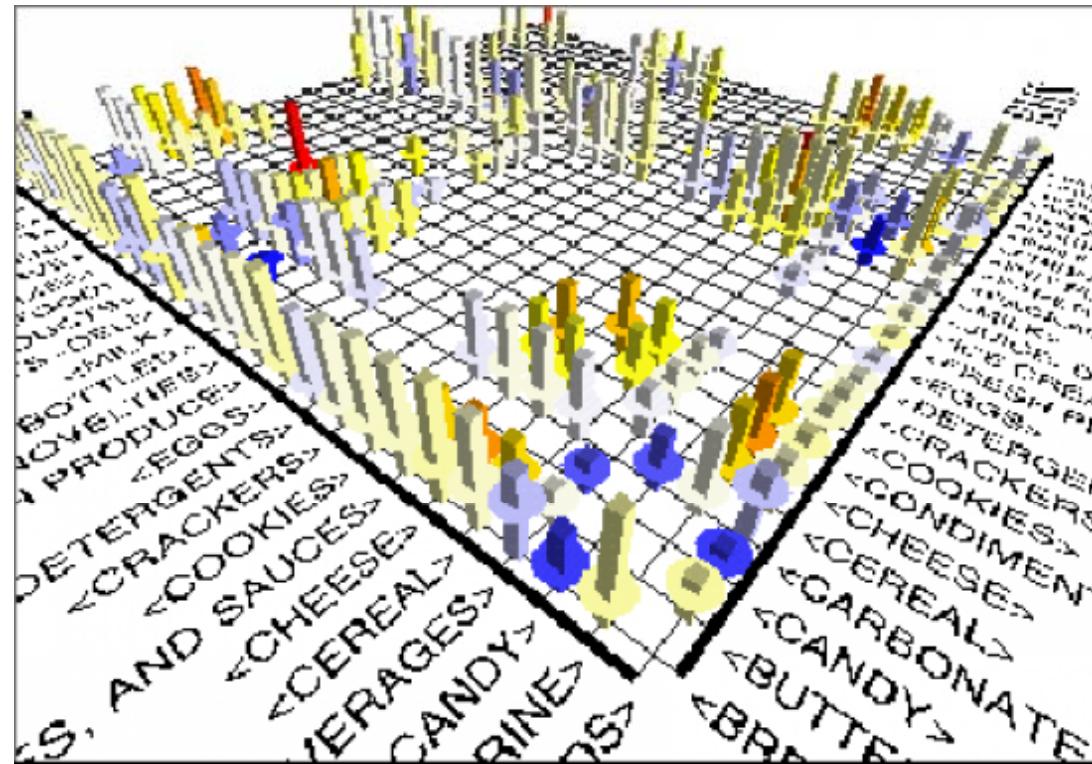


Acciones a realizar:

- Planificar disposiciones alternativas en el almacén
- Limitar descuentos especiales a sólo uno de los dos productos que tienden a comprarse juntos
- Poner los aperitivos que más margen dejan entre los pañales y las cervezas
- Poner productos de bebé en oferta cerca de las cervezas
- Ofrecer cupones descuento para el producto “complementario”, cuando uno de los productos se venda por separado...

# Minería de Datos. Modelos → Reglas de Asociación

Visualización de resultados:

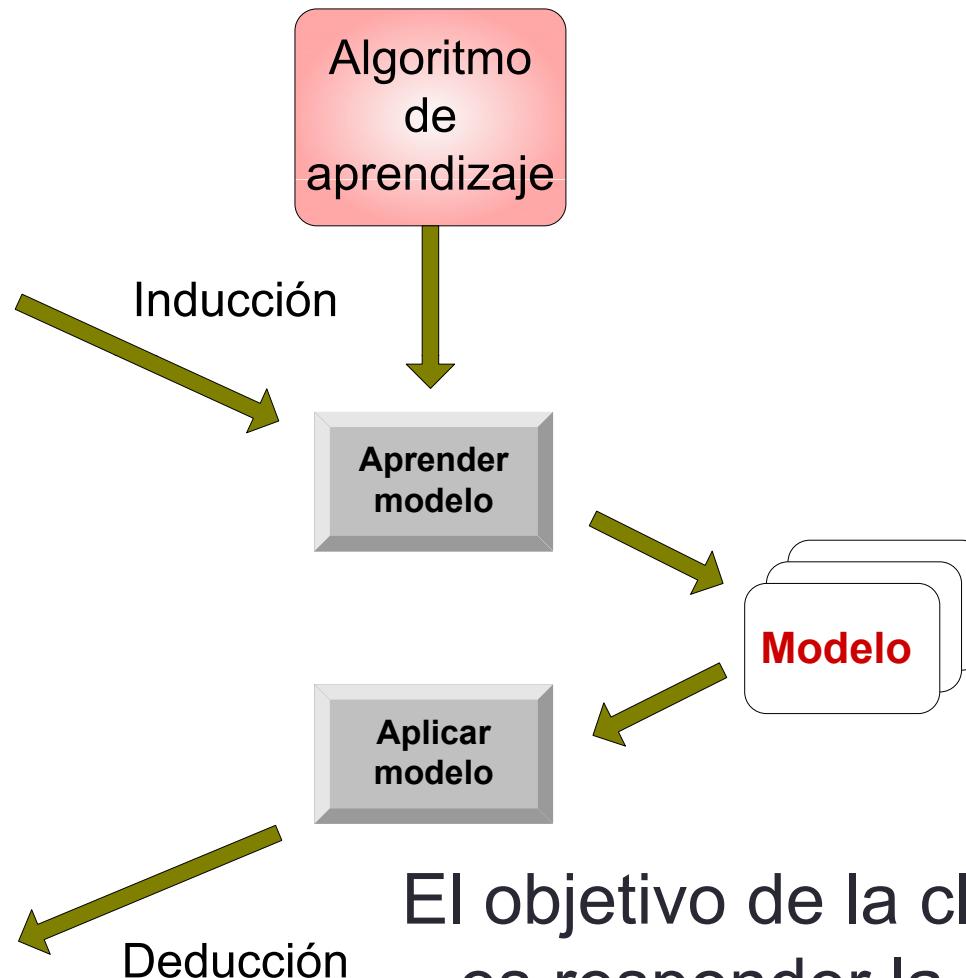


# Minería de Datos. Modelos → Clasificación

## Clasificación: Modelo predictivo

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?



El objetivo de la clasificación es responder la pregunta  
¿Cuál es el valor de la clase para los registros nuevos?

# Minería de Datos. Modelos → Clasificación

Bank  
customers:  
Predict  
fraudulent  
customers



Attributes

Target attribute

Name	Balance	Age	Employed	Write-off
Mike	\$200,000	42	no	yes
Mary	\$35,000	33	yes	no
Claudio	\$115,000	40	no	no
Robert	\$29,000	23	yes	yes
Dora	\$72,000	31	no	no

This is one row (example).  
Feature vector is: <Claudio,115000,40,no>  
Class label (value of Target attribute) is no

# Minería de Datos. Modelos → Clasificación

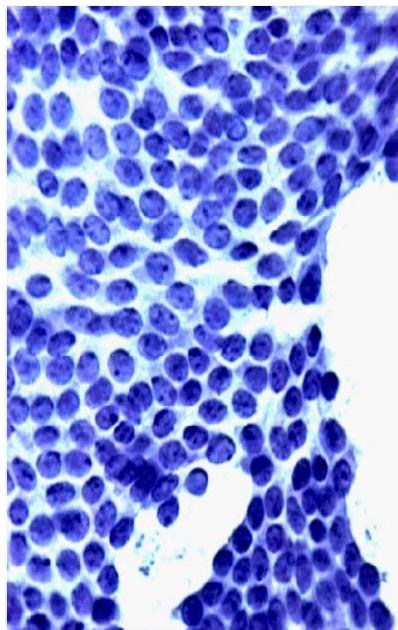
Cell phone  
company:  
Predict  
Customer  
Churn



Variable	Explanation
COLLEGE	Is the customer college educated?
INCOME	Annual income
OVERAGE	Average overcharges per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
HANDSET_PRICE	Cost of phone
LONG_CALLS_PER_MONTH	Average number of long calls (15 mins or over) per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self-reported usage level
LEAVE ( <i>Target variable</i> )	Did the customer stay or leave (churn)?

# Minería de Datos. Modelos → Clasificación

Wisconsin Breast  
Cancer: Predict  
malignant/benign



Attribute name	Description
RADIUS	<i>Mean of distances from center to points on the perimeter</i>
TEXTURE	<i>Standard deviation of grayscale values</i>
PERIMETER	<i>Perimeter of the mass</i>
AREA	<i>Area of the mass</i>
SMOOTHNESS	<i>Local variation in radius lengths</i>
COMPACTNESS	<i>Computed as: <math>\text{perimeter}^2/\text{area} - 1.0</math></i>
CONCAVITY	<i>Severity of concave portions of the contour</i>
CONCAVE POINTS	<i>Number of concave portions of the contour</i>
SYMMETRY	<i>A measure of the nuclei's symmetry</i>
FRACTAL DIMENSION	<i>'Coastline approximation' – 1.0</i>
DIAGNOSIS (Target)	<i>Diagnosis of cell sample: malignant or benign</i>

# Minería de Datos. Modelos → Clasificación

Handwriting  
recognition.

Assign a digit  
from 0 to 9.



0 0 0 0 0 0 0 0 0 0 0 0 0  
1 1 1 1 1 1 1 1 1 1 1 1 1  
2 2 2 2 2 2 2 2 2 2 2 2 2 0  
3 3 3 3 3 3 3 3 3 3 3 3 3 3  
4 4 4 4 4 4 4 4 4 4 4 4 4 4  
5 5 5 5 5 5 5 5 5 5 5 5 5 5  
6 6 6 6 6 6 6 6 6 6 6 6 6 6  
7 7 7 7 7 7 7 7 7 7 7 7 7 7  
8 8 8 8 8 8 8 8 8 8 8 8 8 8  
9 9 9 9 9 9 9 9 9 9 9 9 9 9

# Minería de Datos. Modelos → Clasificación

Se pueden construir distintos tipos de clasificadores:

## Modelos Interpretables:

- Árboles de decisión (decision trees)
- Reglas (p.ej. listas de decisión)

## Modelos no interpretables:

- Clasificadores basados en casos (k-NN)
- Redes neuronales
- Redes bayesianas
- SVMs (Support Vector Machines)
- ...

# Minería de Datos. Modelos → Clasificación

## → Árboles de decisión

### 1. Preprocessing: Remove keys

Attributes

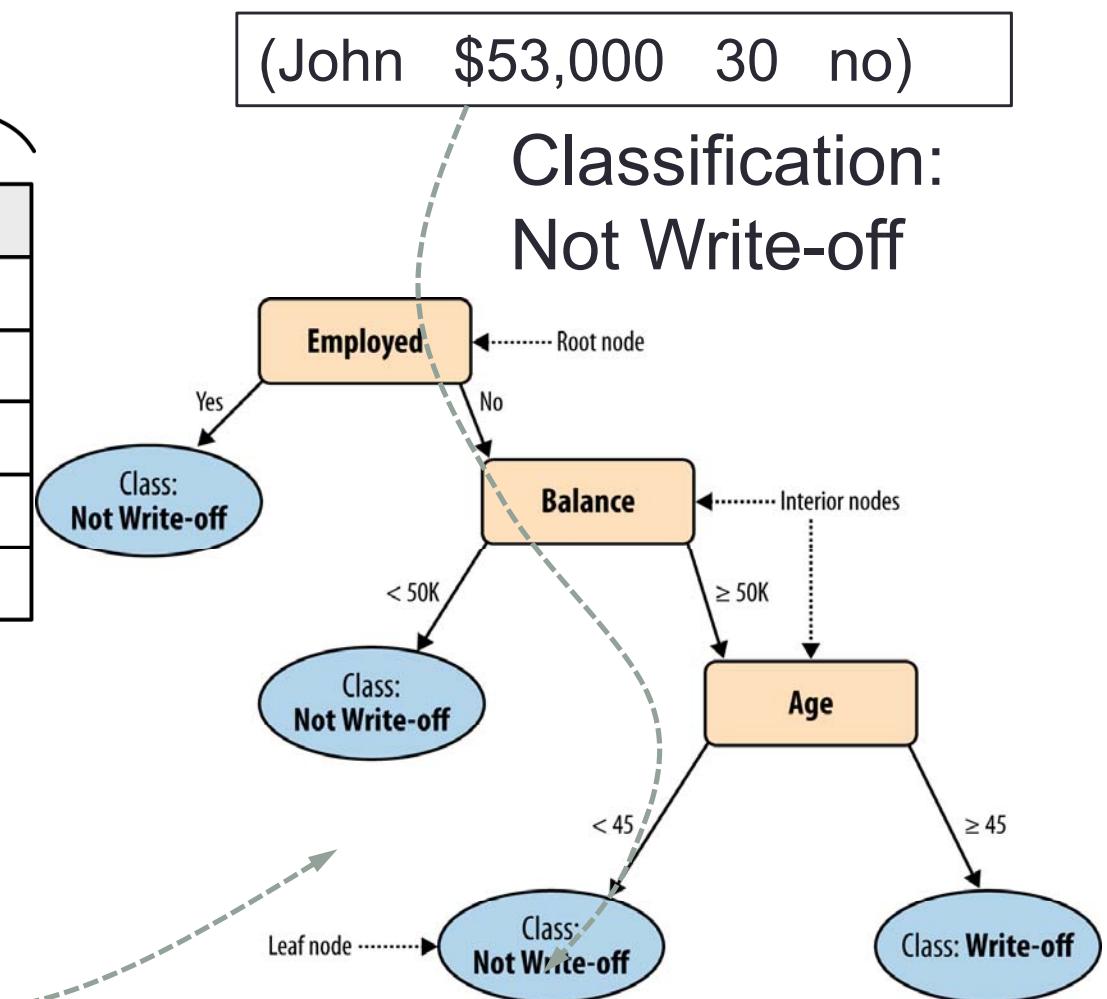
Target attribute

Name	Balance	Age	Employed	Write-off
Mike	\$200,000	42	no	yes
Mary	\$35,000	33	yes	no
Claudio	\$115,000	40	no	no
Robert	\$29,000	23	yes	yes
Dora	\$72,000	31	no	no

This is one row (example). Feature vector is: <Claudio,115000,40,no> Class label (value of Target attribute) is no

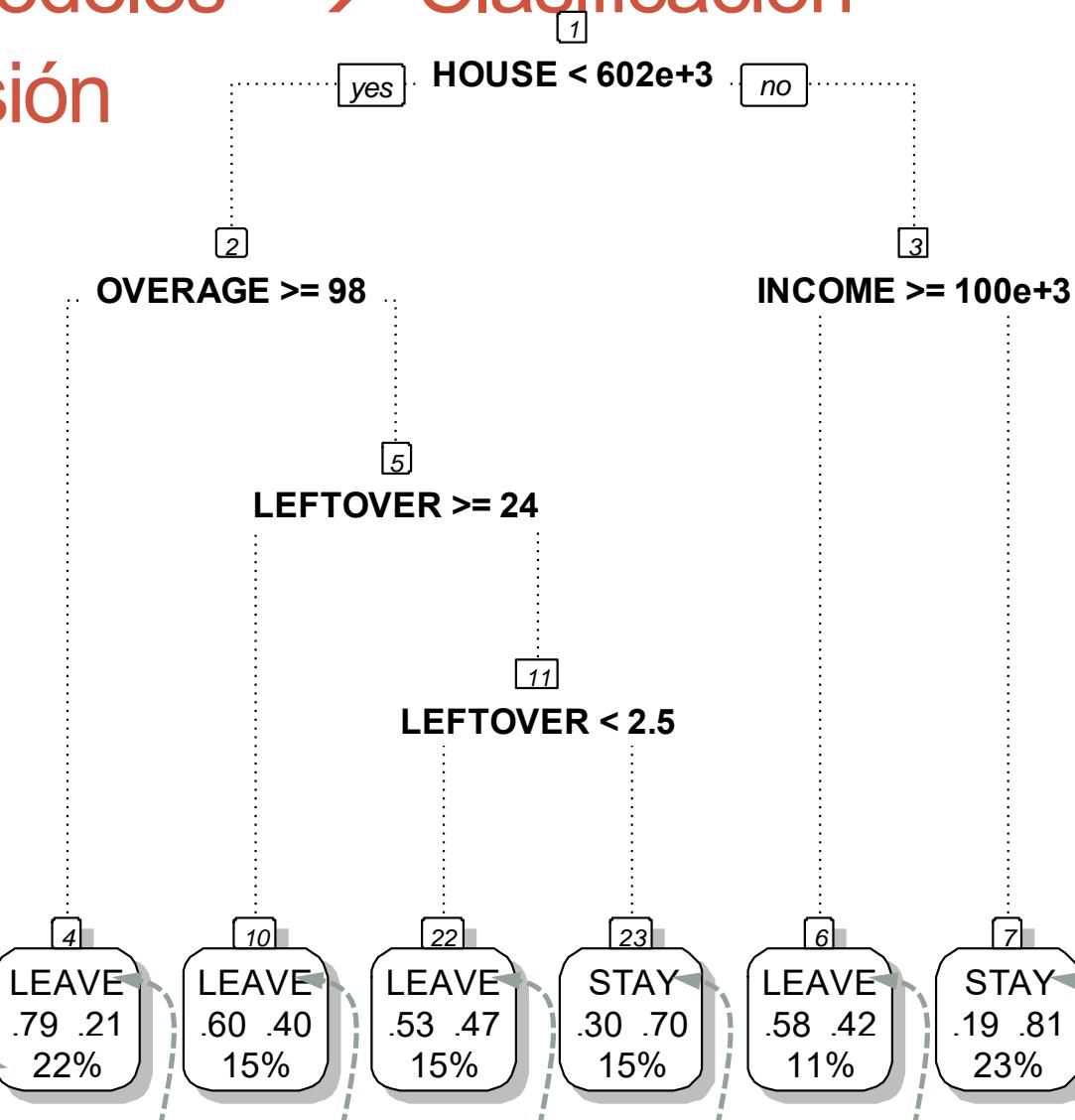
### 2. The model (decision tree) is constructed

3. A new record is “parsed” by the tree. The leaf node gives the assigned class label



# Minería de Datos. Modelos → Clasificación → Árboles de decisión

Customer Churn example



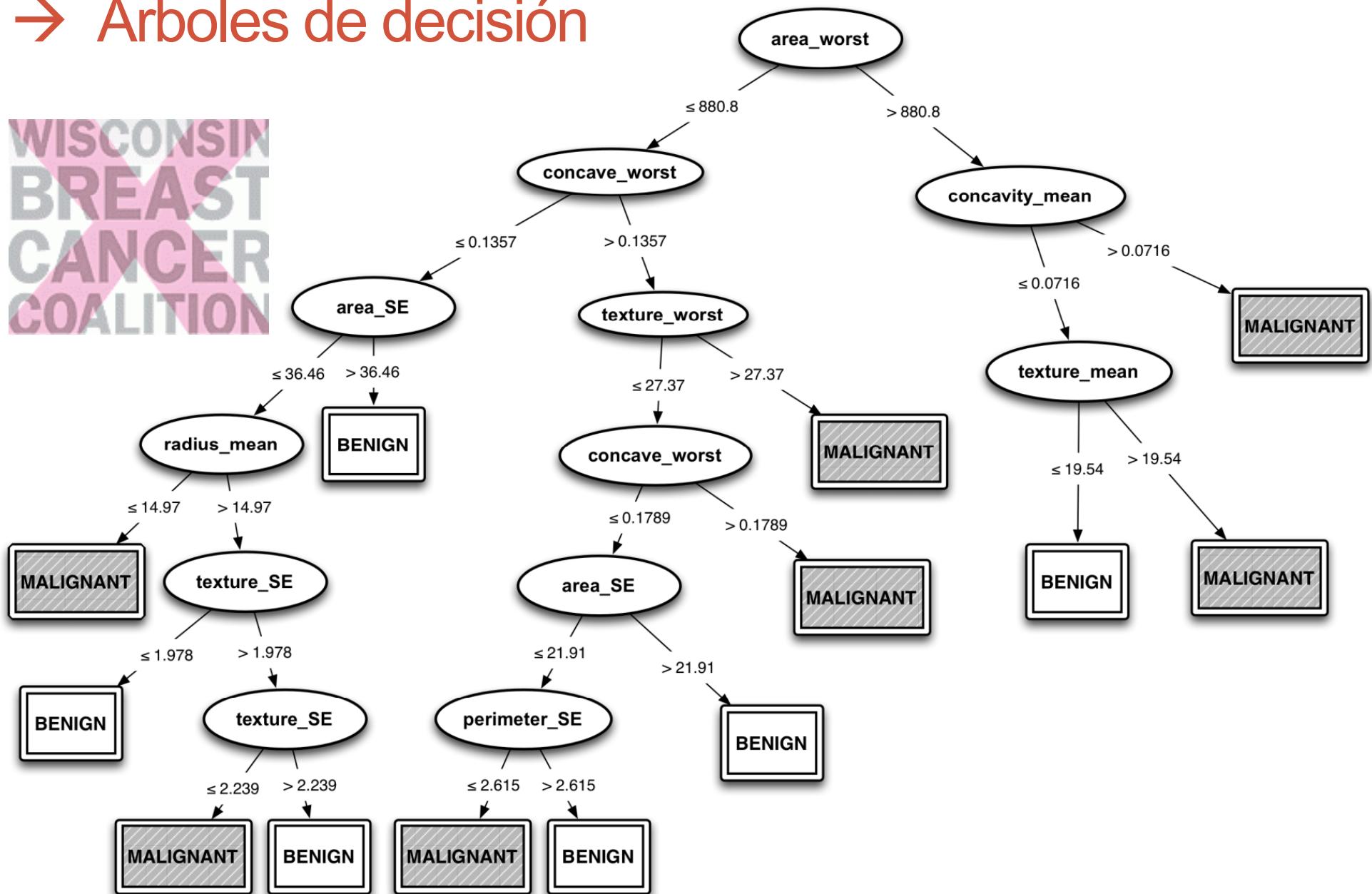
Total percentage of covered examples

Leaf nodes are not “pure”. They contain examples of several classes.

The majority one is chosen to label the node

# Minería de Datos. Modelos → Clasificación → Árboles de decisión

**WISCONSIN  
BREAST  
CANCER  
COALITION**



# Minería de Datos. Modelos → Clasificación

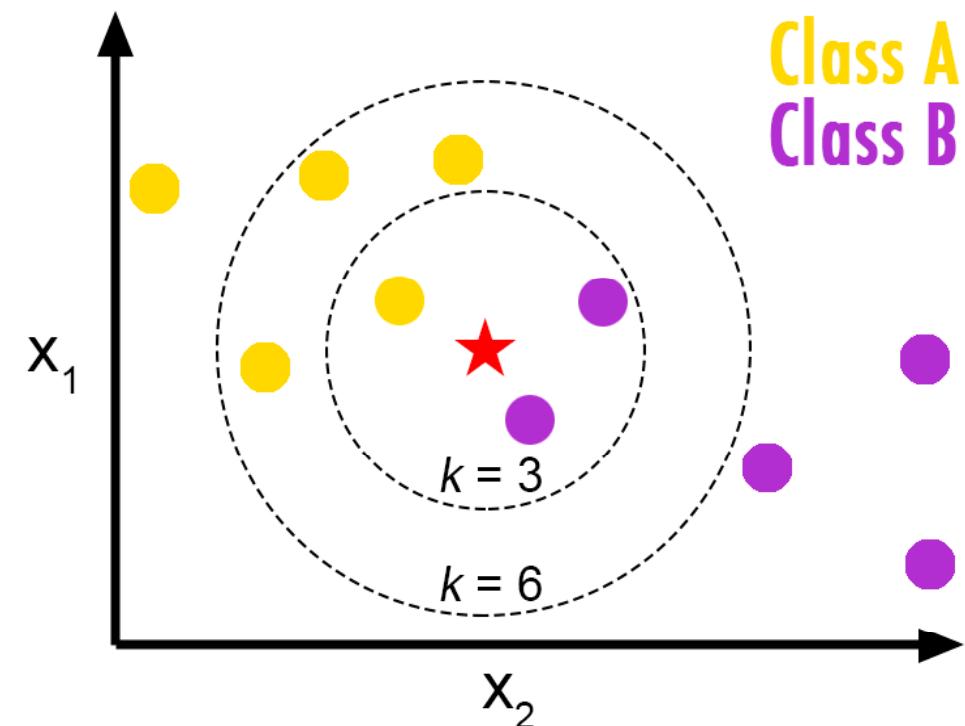
## → KNN

KNN = K-nearest neighbors, k vecinos más cercanos.

Es un modelo no interpretable

La clasificación consiste en encontrar los k vecinos más cercanos y se le asigna al nuevo dato la clase más común entre los k vecinos.

Cercanía → Medida de distancia.



Ejemplo con 2 Atributos.  
Distancia Euclídea.

Datos de referencia: ○

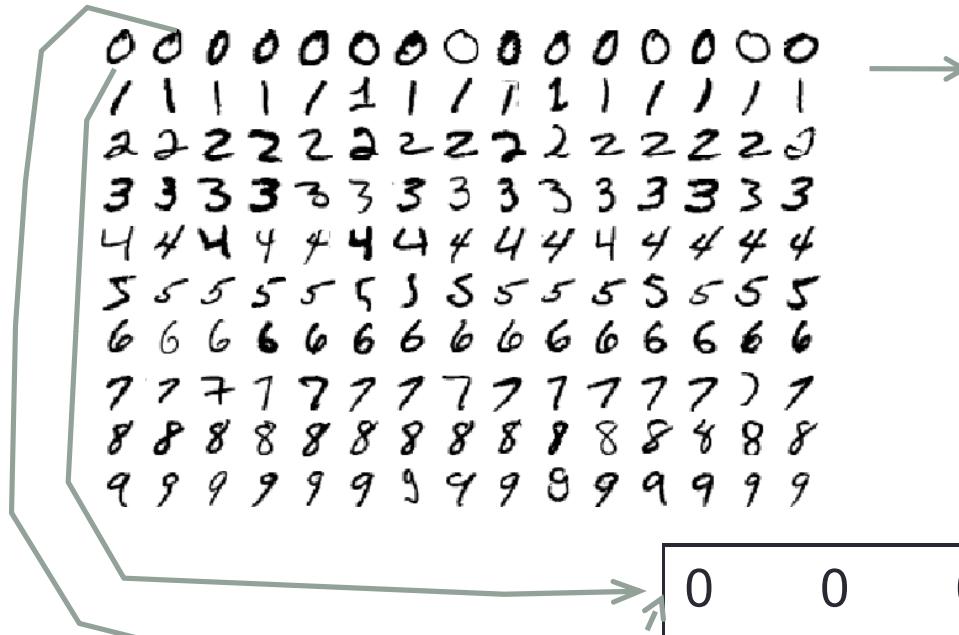
Nuevo punto: ☆

$k=3 \rightarrow$  Clase asignada: B

$k=6 \rightarrow$  Clase asignada: A

# Minería de Datos. Modelos → Clasificación

## → KNN



Representación por pixels. Cada dígito es una matriz de  $n \times m$  pixeles → Se representa como un vector de longitud  $n \times m$



0	0	0	0	1	1	...	0	0	Digit 0
0	0	0	0	0	1	...	0	0	Digit 0
...	...	...	...	...	...	...	...	...	...
0	0	1	1	0	1	...	0	0	Digit 9

Nuevo dato.  
Se calcula la  
distancia a todos  
ellos → Clase  
mayoritaria de  
los  $k$  más cercanos

0	0	0	...	1	1	...	0	0	0
---	---	---	-----	---	---	-----	---	---	---

Clasificación: Digit 7

# Minería de Datos. Modelos → Agrupamiento

Agrupamiento o Clustering: Modelo descriptivo

**Clasificación (Aprendizaje supervisado, Classification):**

El experto ha de identificar la clase, es decir, un atributo objetivo (target)

**Agrupamiento (Aprendizaje no supervisado, Clustering):**

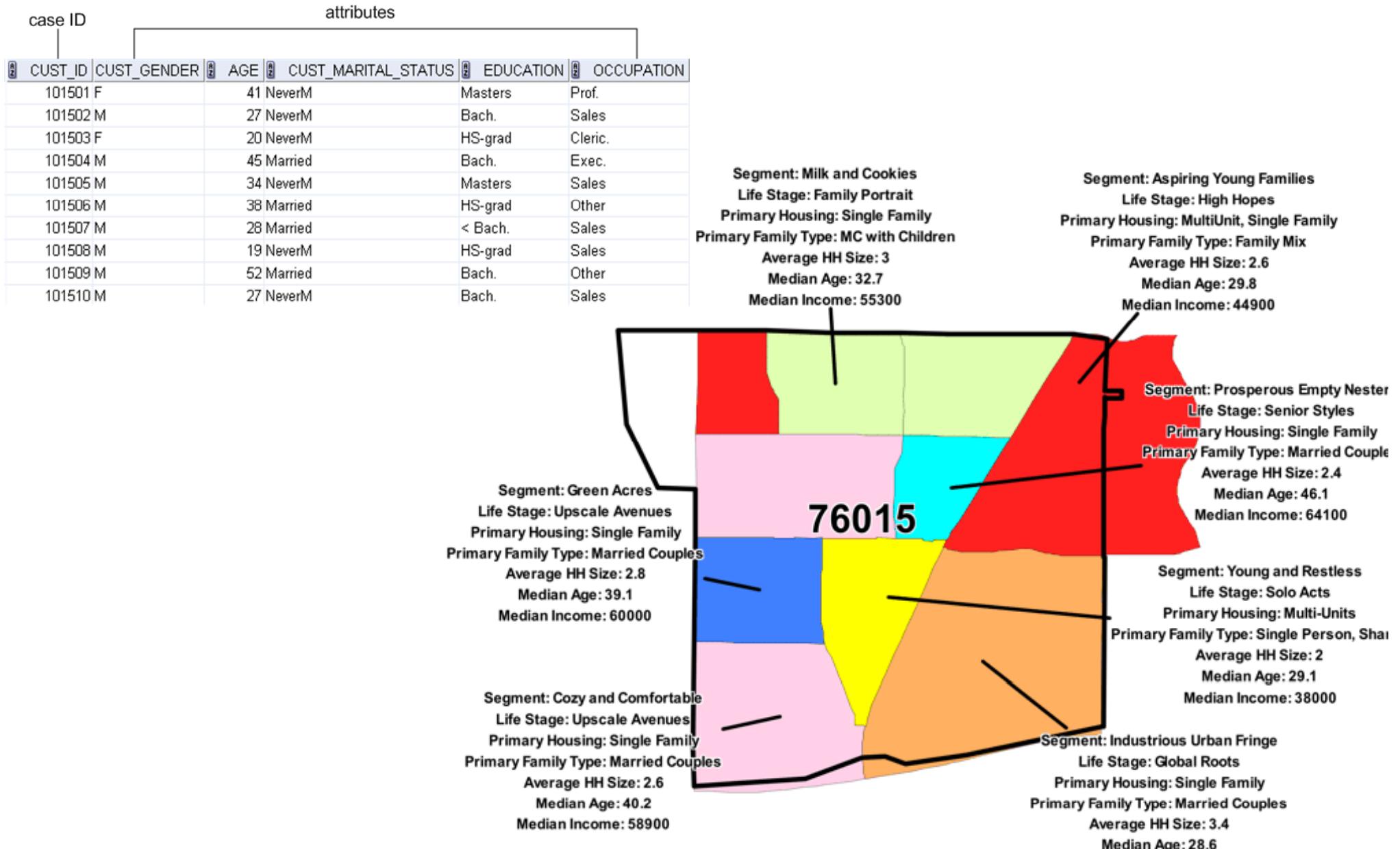
No existe tal atributo. El objetivo es agrupar registros (filas, tuplas) en base a sus semejanzas.

# Minería de Datos. Modelos → Agrupamiento

Segmentación de Clientes. Objetivo: Encontrar grupos (clusters) que identifiquen un modelo de cliente con características comunes (intereses, nivel de ingresos, hábitos de gasto, ...)



# Minería de Datos. Modelos → Agrupamiento



# Minería de Datos. Modelos → Agrupamiento

Clustering de resultados en un motor de búsqueda

Clusty  
→  
Yippy

The screenshot shows the Yippy search interface. At the top, there's a navigation bar with links for 'web', 'news', 'blogs', 'wikipedia', 'jobs', and 'more'. Below that is a search bar containing the query 'data mining', a 'Search' button, and 'advanced preferences' link. On the left, a sidebar titled 'clouds' lists various categories with their counts: All Results (455), Australia (22), Consulting (36), Business Intelligence (31), Jobs (20), Projects (16), Visualization (17), Gold (13), Freedoms (15), Sales (12), Data gathering or extraction (15), Knowledge discovery (8), Mobile (9), Computer Science (12), Protection, Privacy (13), Reviews (11), Conference (10), Programmers (8), Automotive, CRM (7), Predictive Modeling (9), and Market Research (7). The main content area displays the top 455 results out of 57,638. The first result is 'Kiribati', followed by several news articles and research links related to data mining, such as 'Japanese stocks, dollar shine on upbeat U.S. jobs', 'Data Mining Research - www.dataminingblog.com', 'Big Data Analytics, Enterprise Analytics, Data Mining Software, Statistical Analysis, Predictive Analytics', and 'Australia stocks sink ever lower after last week's retreat'.

Top 455 results of at least 57,638 retrieved for the query **data mining** ([details](#))

**Kiribati**

KiribatiArea: 277 sq mi (717 sq km) / World Rank: 179Location: Group of islands in the Pacific Ocean, between the equator and international date line intersect.Coordinates: 1°25'N, 173°00'EBorders: No international bordersCoastline:

[Japanese stocks, dollar shine on upbeat U.S. jobs](#)

... denominated index of Asia-Pacific shares outside Japan was down 0.1 percent, as the boost from positive U.S. data was offset by weaker Asian currencies and falls in Australian mining shares. The U.S. Labor Department reported nonfarm payrolls rose 248,000 in September, 33,000 more than median forecast ...

[www.reuters.com/...obal-idUSKCN0HV00Z20141006?feedType=RSS&feedName=businessNews](http://www.reuters.com/...obal-idUSKCN0HV00Z20141006?feedType=RSS&feedName=businessNews) - [cache]  
- Reuters, Reuters

[Data Mining Research - www.dataminingblog.com](#)

... purchase be used to Continue reading... Next Page » **Data Mining** Research **Data Mining** Links RapidMiner Open Source **Data Mining** Links PROS ... **Data Mining** Reading Recommandations T-shirts, Mugs & Mousepads **Data Mining** Search Engine Supported by AnalyticBridge Archives Select Month ...

[www.dataminingblog.com](http://www.dataminingblog.com) - [cache] - Yippy Index

[Big Data Analytics, Enterprise Analytics, Data Mining Software, Statistical Analysis, Predictive Analytics](#)

... STATISTICA Product Overview Connectivity and **Data Integration** Solutions **Data Mining** Solutions Decisioning Platform Desktop Solutions Enterprise Solutions Power ...

[www.statsoft.com](http://www.statsoft.com) - [cache] - Yippy Index

[Australia stocks sink ever lower after last week's retreat](#)

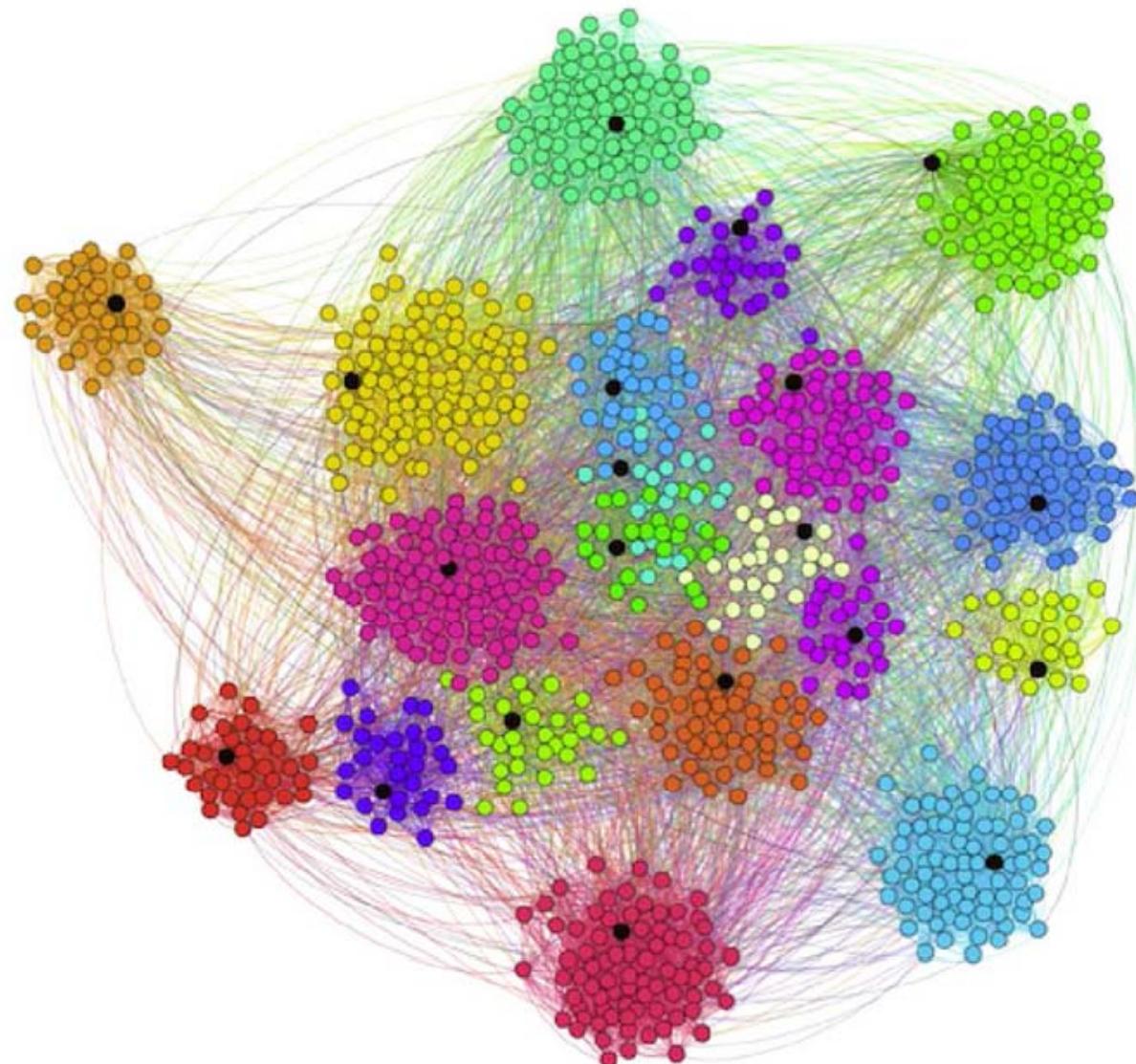
[www.marketwatch.com/...h/marketpulse](http://www.marketwatch.com/...h/marketpulse) (MarketWatch.com - MarketPulse) - [cache] - Yippy News

[China, Japan data deluxe on tap this week](#)

# Minería de Datos. Modelos → Agrupamiento

Detección de *comunidades* en redes sociales

Gephi  
(herramienta de visualización)



# Minería de Datos. Modelos → Agrupamiento

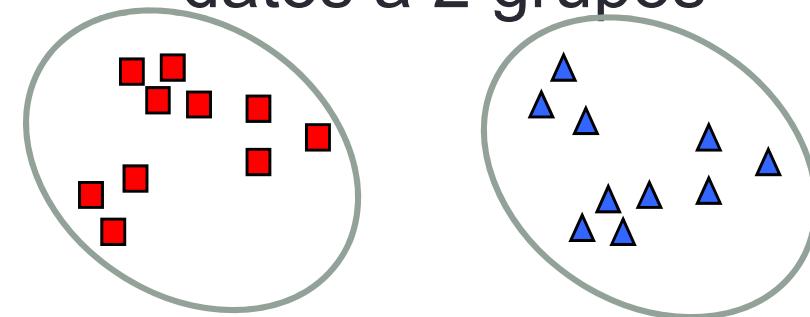
¿Cómo realizar el agrupamiento?

Supongamos sólo dos atributos:

Datos originales



Asignación de los  
datos a 2 grupos



# Minería de Datos. Modelos → Agrupamiento

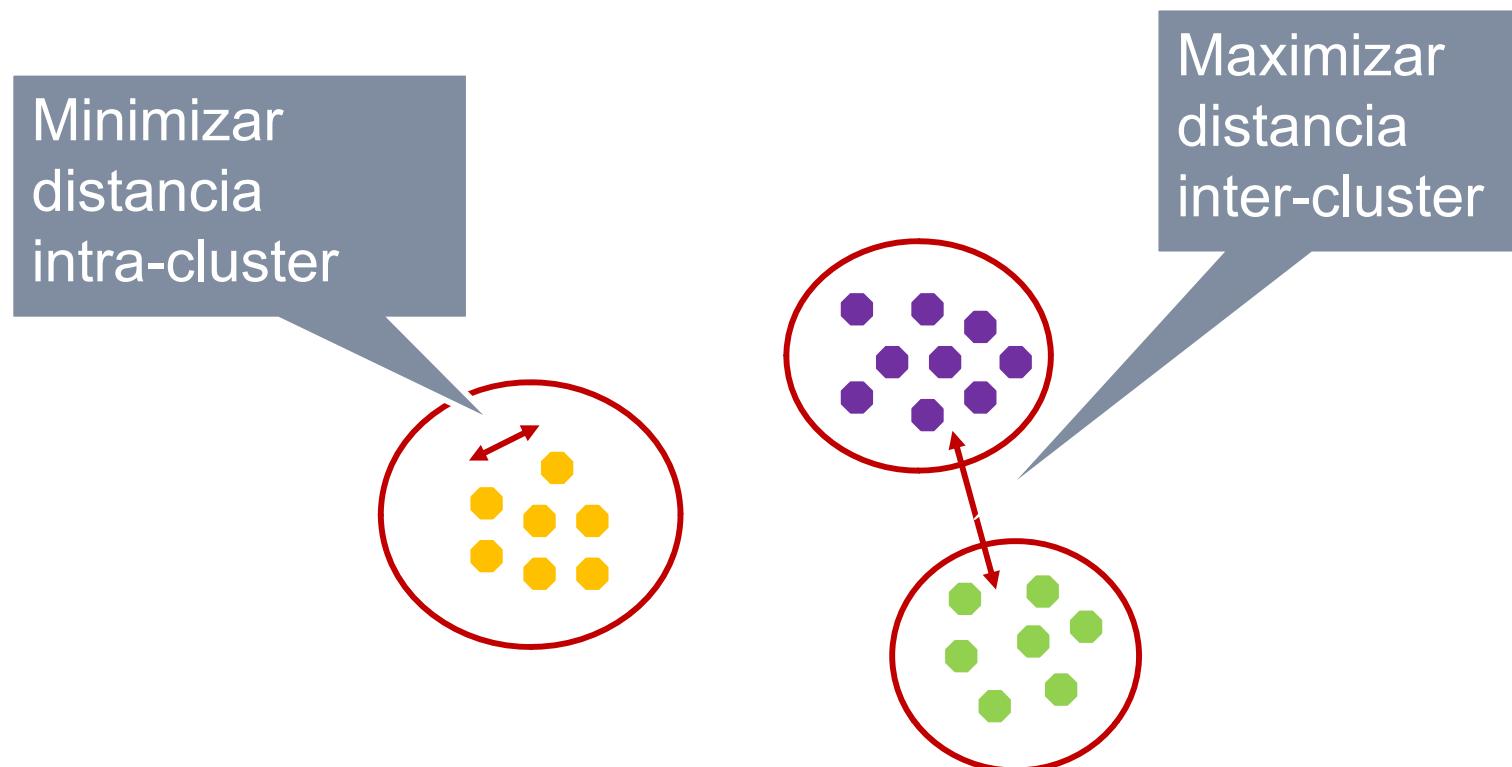
Más de dos atributos:

	id	sexo	fechnac	educ	catlab	salario	salini	tiempem p	expprev	minoría
Grupo 1	121	Mujer	6-agosto-1936	15	Administrativo	\$18.750	\$10.500	90	54	No
	122	Mujer	26-septiembre-1965	15	Administrativo	\$32.550	\$13.500	90	22	No
	123	Mujer	24-abril-1949	12	Administrativo	\$33.300	\$15.000	90	3	No
	124	Mujer	29-mayo-1963	16	Administrativo	\$38.550	\$16.500	90	Ausente	No
	125	Hombre	6-agosto-1956	12	Administrativo	\$27.450	\$15.000	90	173	Sí
Grupo 2	126	Hombre	21-ene-1951	15	Seguridad	\$24.300	\$15.000	90	191	Sí
	127	Hombre	1-septiembre-1950	12	Seguridad	\$30.750	\$15.000	90	209	Sí
Grupo 3	128	Mujer	25-julio-1946	12	Administrativo	\$19.650	\$9.750	90	229	Sí
	129	Hombre	18-julio-1959	17	Directivo	\$68.750	\$27.510	89	38	No
	130	Hombre	6-septiembre-1958	20	Directivo	\$59.375	\$30.000	89	6	No
	131	Hombre	8-febrero-1962	15	Administrativo	\$31.500	\$15.750	89	22	No
	132	Hombre	17-mayo-1953	12	Administrativo	\$27.300	\$17.250	89	175	No
	133	Hombre	12-septiembre-1959	15	Administrativo	\$27.000	\$15.750	89	87	No

# Minería de Datos. Modelos → Agrupamiento

## Objetivo

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos [*clusters*].



## Minería de Datos. Modelos → Agrupamiento

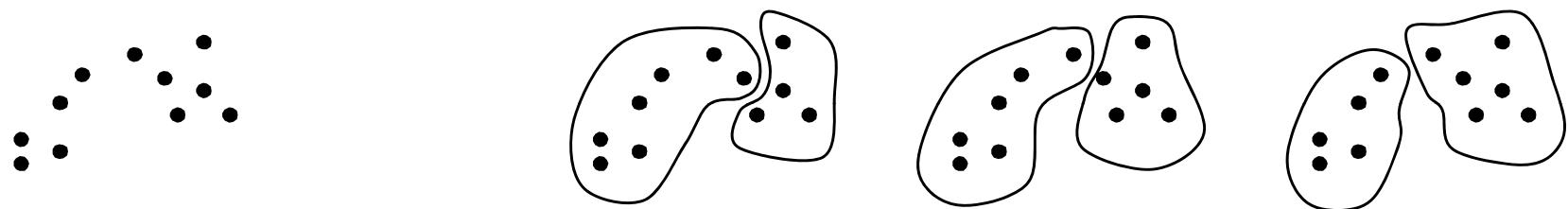
¿Cuántos grupos -k- se han de considerar? Opciones:

- Se fija a priori k → Algoritmos eficientes
- No se fija ningún valor de k sino que se fija un criterio de fusión de grupos y se ofrece la posibilidad de elegir cualquier k → Métodos jerárquicos (poco eficientes)
- Otros métodos (basados en densidad, por ejemplo) fijan un umbral de distancia entre registros y se construyen los grupos en base a dicha información.

# Minería de Datos. Modelos → Agrupamiento

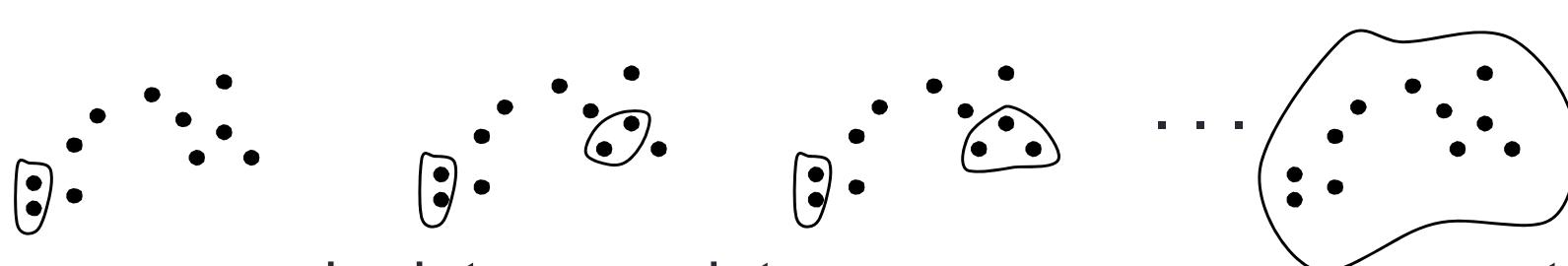
## Tipos de Clustering.

**Clustering por particiones** (suele fijarse k)



Se parte de una división inicial y se van recolocando los puntos

**Clustering jerárquico** (no se fija k)

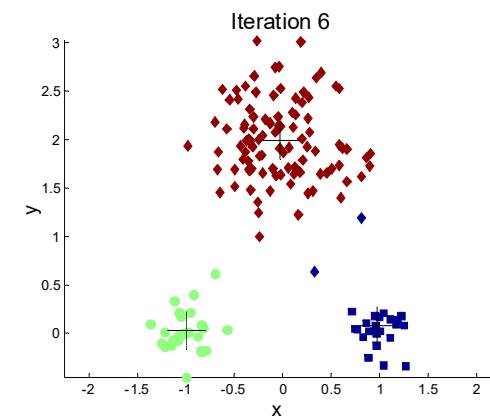
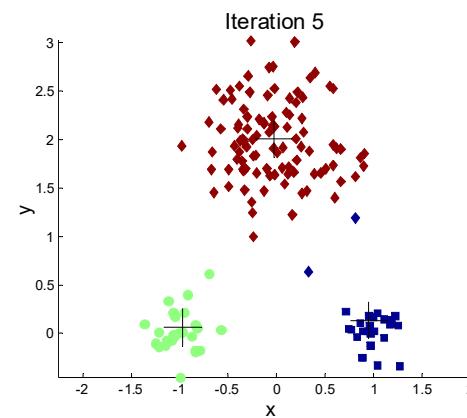
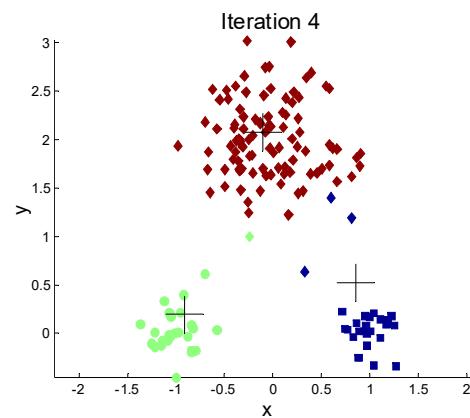
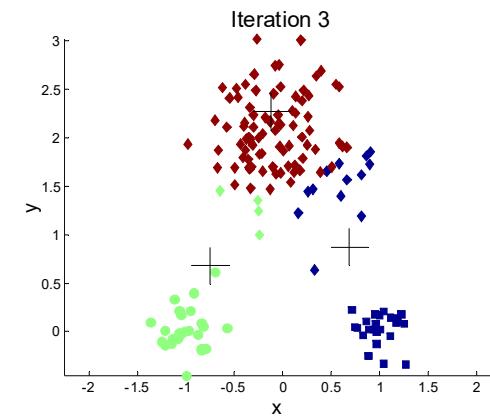
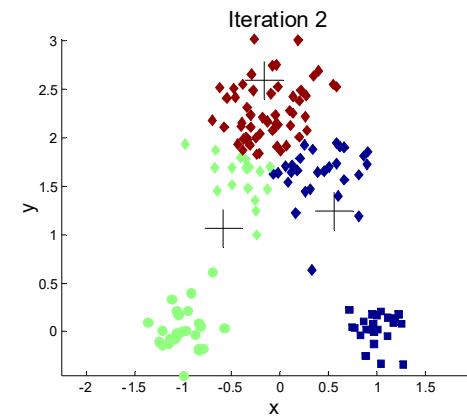
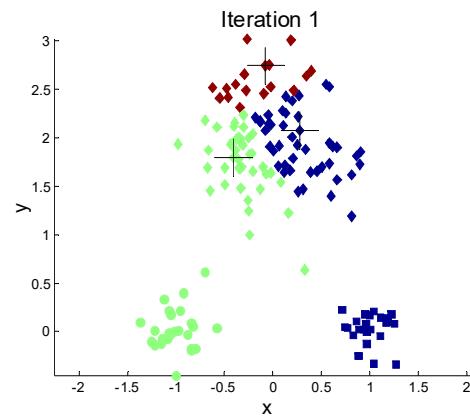


Se van agrupando datos con datos, con grupos y grupos entre sí.

# Minería de Datos. Modelos → Agrupamiento

## Clustering por particiones. K-Means

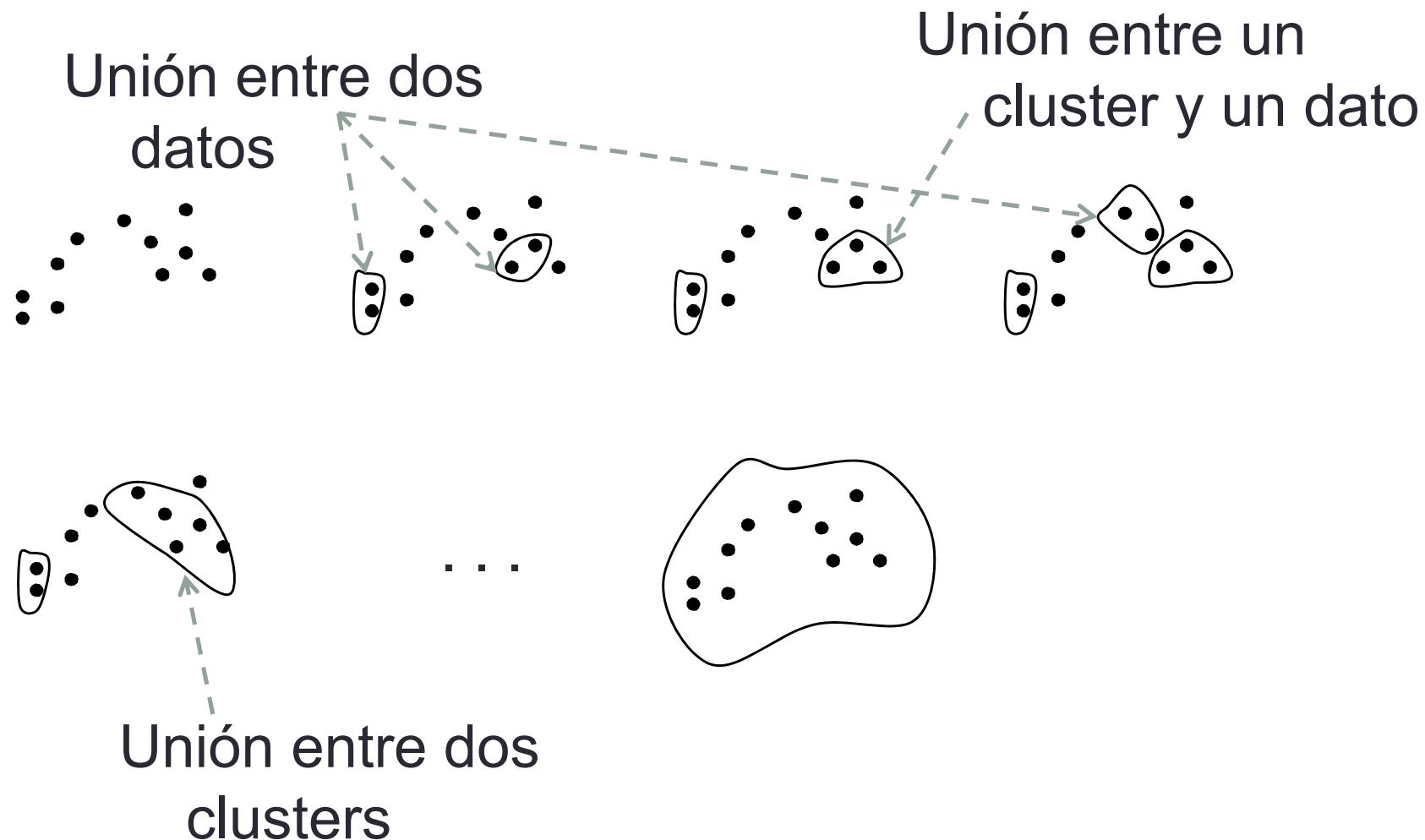
Se eligen k centroides y se van recolocando en un proceso iterativo



# Minería de Datos. Modelos → Agrupamiento

## Clustering Jerárquico

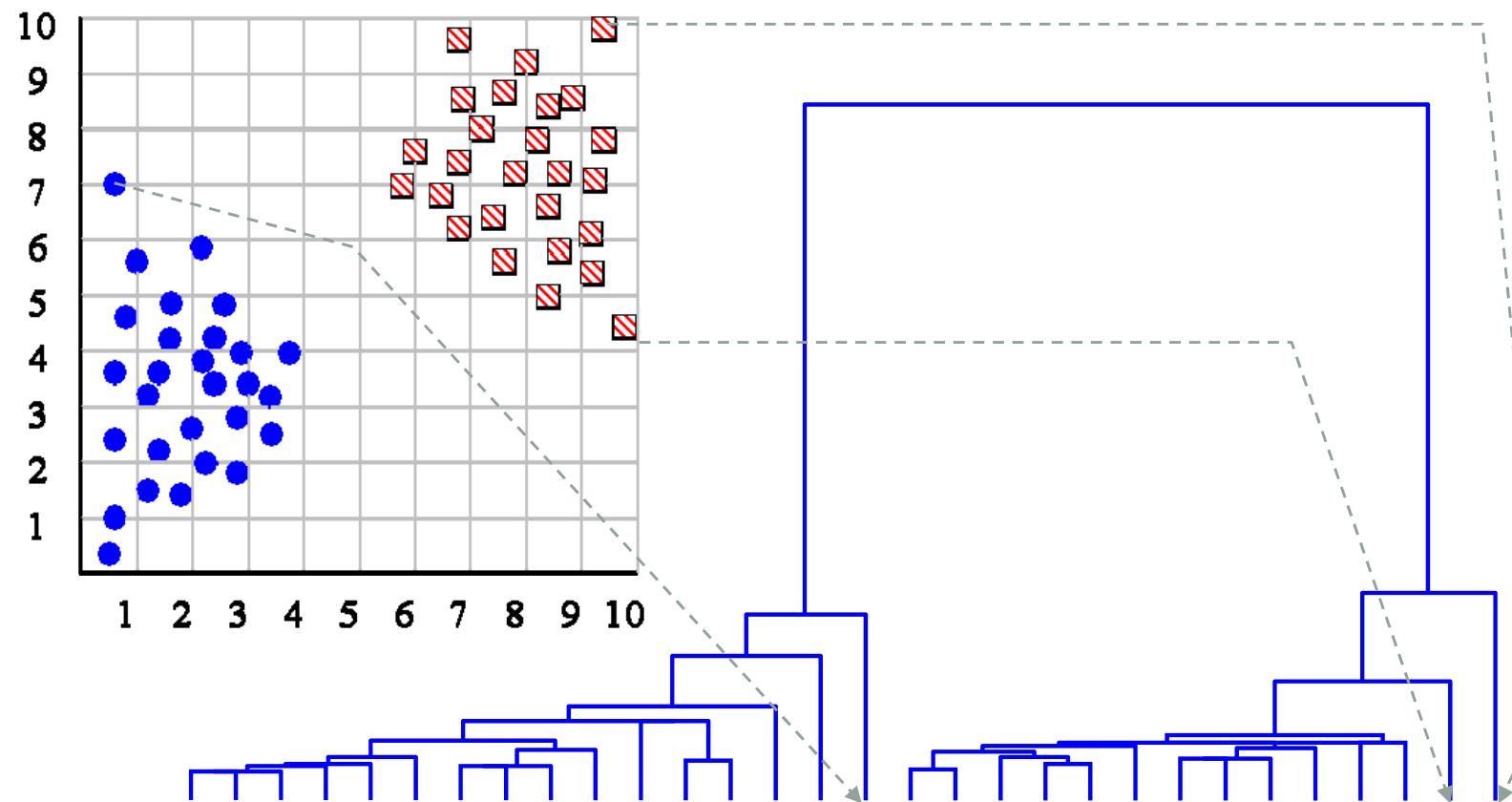
En los métodos jerárquicos, se van agrupando datos con datos, datos con grupos y grupos entre sí.



# Minería de Datos. Modelos → Agrupamiento

Representación gráfica: Dendrograma

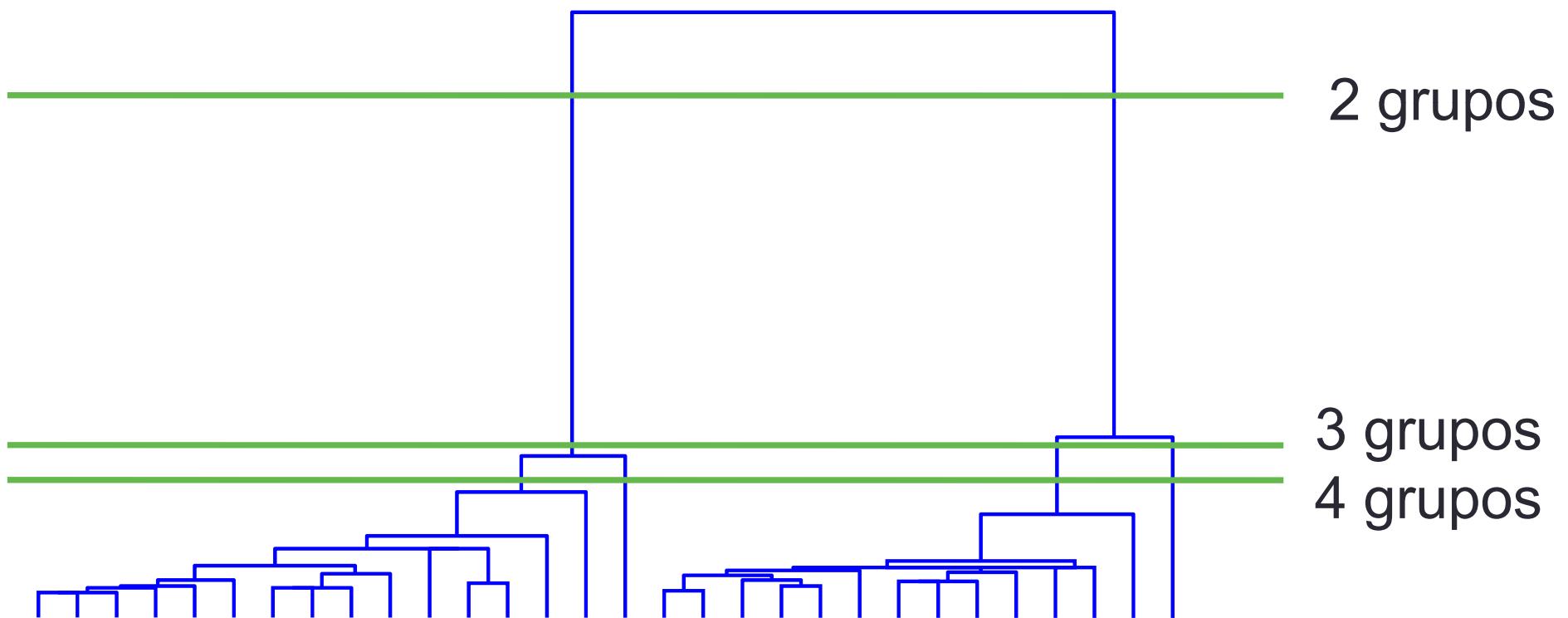
La altura representa la distancia entre los datos y/o grupos



# Minería de Datos. Modelos → Agrupamiento

¿Número de grupos?

Inspección visual, bajando progresivamente una línea horizontal



# Minería de Datos. Modelos → Detección de Anomalías

*Finding a needle in a haystack* is not a correct phrase to refer to the problem of finding anomalies because I know what a needle looks like



[www.jolyon.co.uk](http://www.jolyon.co.uk)

# Minería de Datos. Modelos → Detección de Anomalías

I know what I have to find

I have a complete and accurate  
description of the anomalous  
entity to be found



I don't know what I have to find

An anomaly is an abnormal  
entity



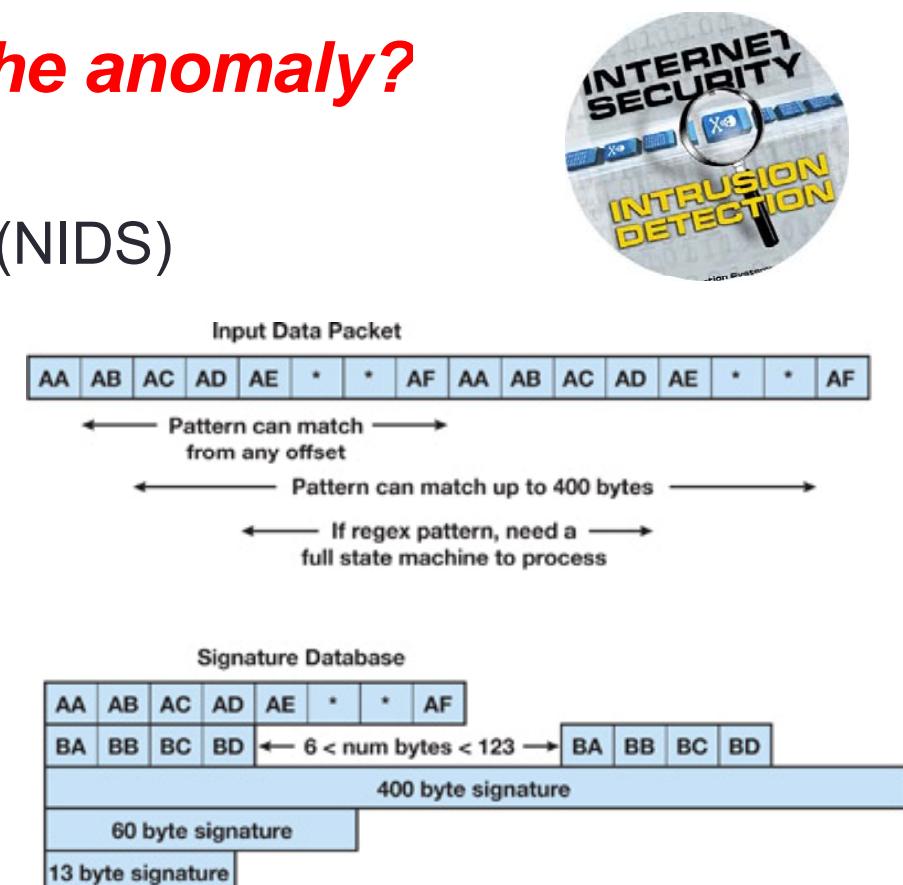
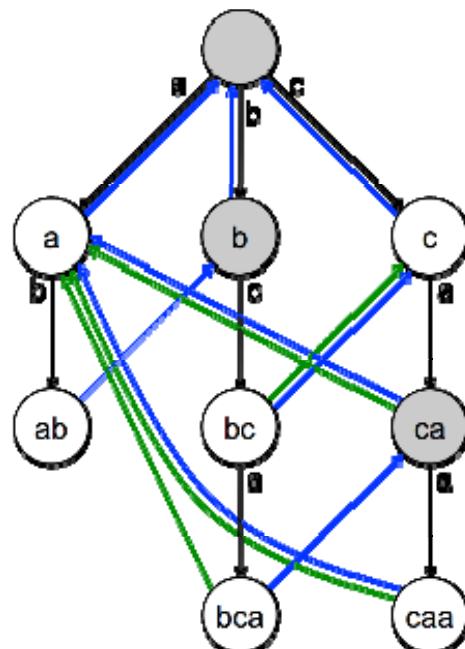
# Minería de Datos. Modelos → Detección de Anomalías

***Do I have the description of the anomaly?***

Example:

Network Intrusion Detection Systems (NIDS)

- NIDS Signature based:  
I know what I have to find.  
Main problem: String Matching



# Minería de Datos. Modelos → Detección de Anomalías

***Do I have the description of the anomaly?***

Example:

Network Intrusion Detection Systems (NIDS)

- NIDS Anomaly detection based:  
I don't know what I have to find.

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

# Minería de Datos. Modelos → Detección de Anomalías

## *Supervised Methods* →

I have anomalies in my training set and they are labelled

A classification model  
(including the anomaly class)  
is built.

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

# Minería de Datos. Modelos → Detección de Anomalías

*SemiSupervised Methods* →  
I do not have anomalies  
in my training set

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No

# Minería de Datos. Modelos → Detección de Anomalías

*UnSupervised Methods* →

I have anomalies in my training set but they are not labelled  
I don't know if a record is an anomaly or not)

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	
1	206.135.38.95	11:07:20	160.94.179.223	139	192	
2	206.163.37.95	11:13:56	160.94.179.219	139	195	
3	206.163.37.95	11:14:29	160.94.179.217	139	180	
4	206.163.37.95	11:14:30	160.94.179.255	139	199	
5	206.163.37.95	11:14:32	160.94.179.254	139	19	
6	206.163.37.95	11:14:35	160.94.179.253	139	177	
7	206.163.37.95	11:14:36	160.94.179.252	139	172	
8	206.163.37.95	11:14:38	160.94.179.251	139	285	
9	206.163.37.95	11:14:41	160.94.179.250	139	195	
10	206.163.37.95	11:14:44	160.94.179.249	139	163	

# Minería de Datos. Modelos → Detección de Anomalías

Detección de Anomalías

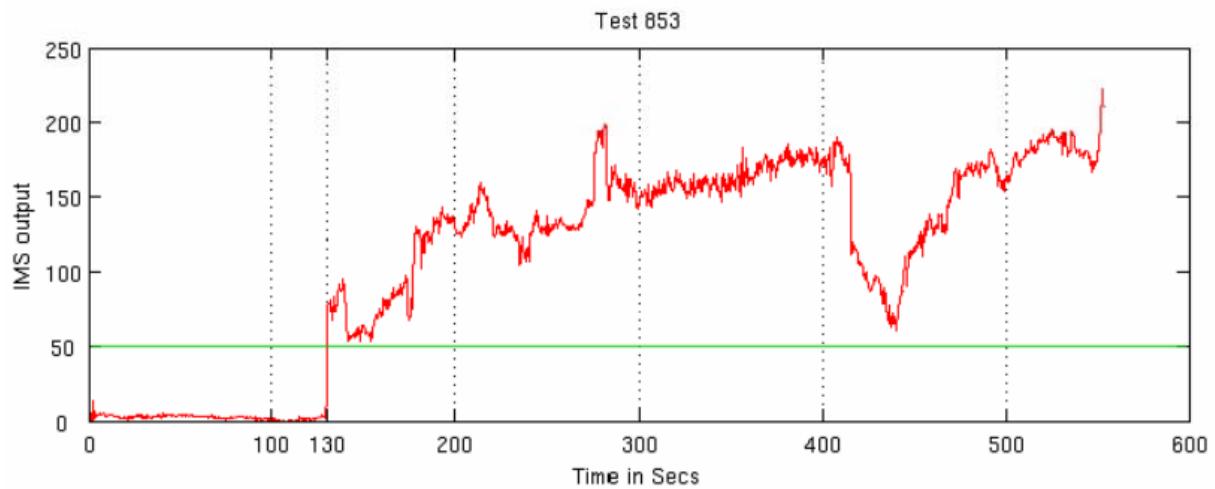
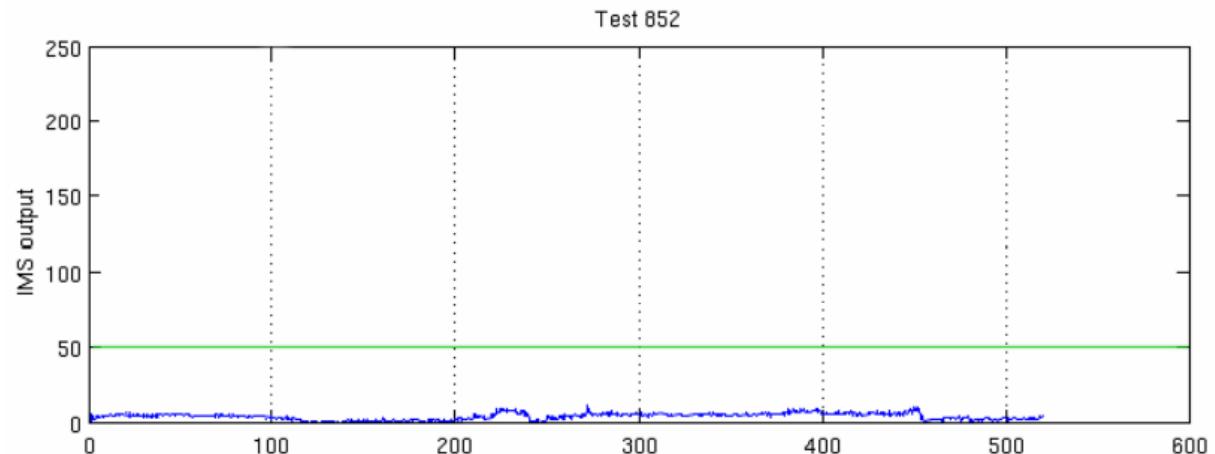
→ Fallos en sistemas

SSME: Space Shuttle Main Engines (reusable)

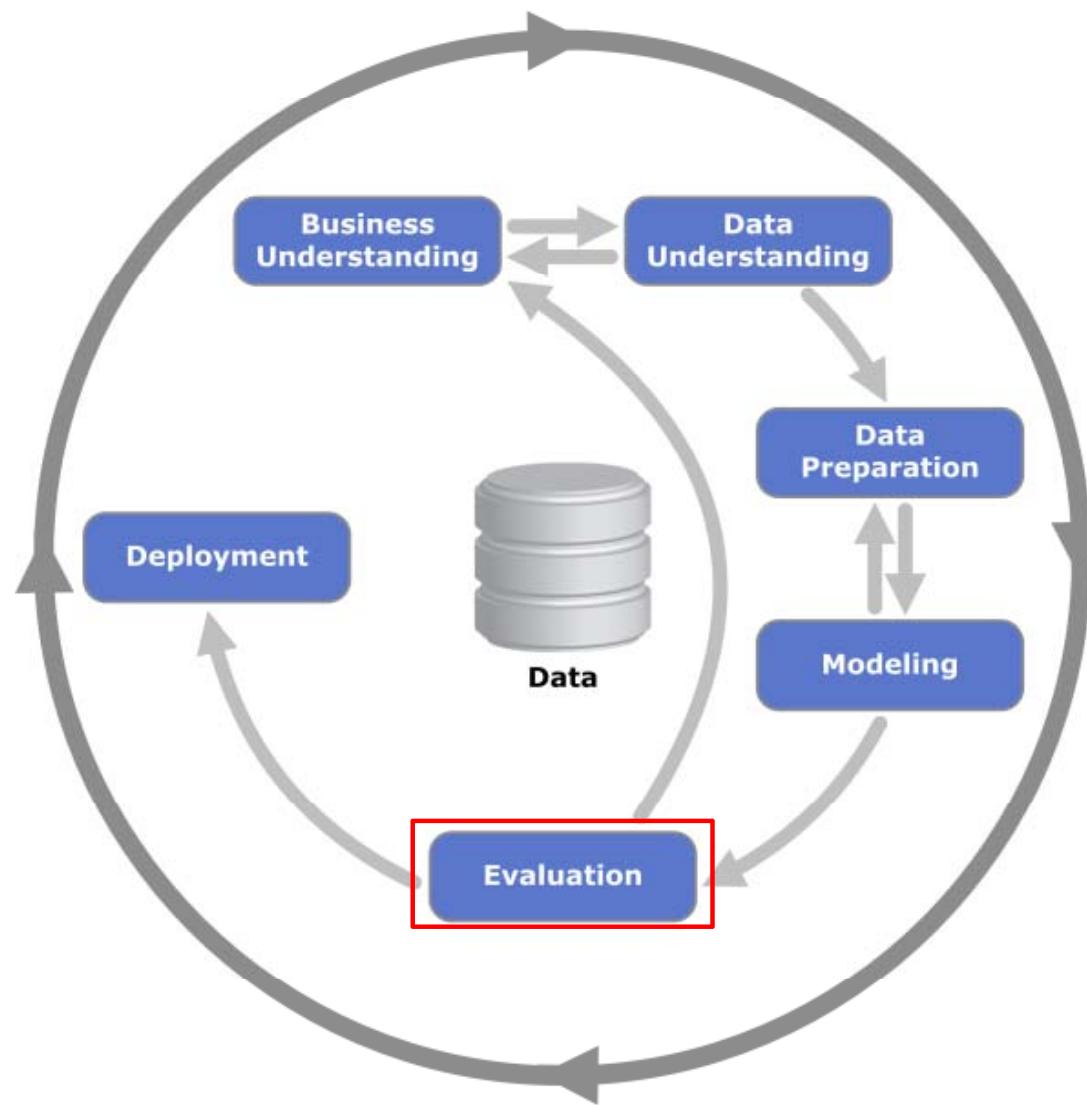


Inductive Monitoring System(IMS)

Pressure, temperature, vibration, etc.  
Sensors. → 147 variables. 13000 time steps



# Minería de Datos. Evaluación



## Minería de Datos. Evaluación

- Reglas de asociación: Medidas del interés de las reglas obtenidas, complejidad de las mismas, etc.
- Clustering: Número de clusters obtenido, cohesión de éstos, etc.
- Clasificación: Estimación del error de clasificación, interpretabilidad del modelo, complejidad del mismo, etc.

En los distintos cursos del Máster se verán técnicas de evaluación de los distintos modelos.

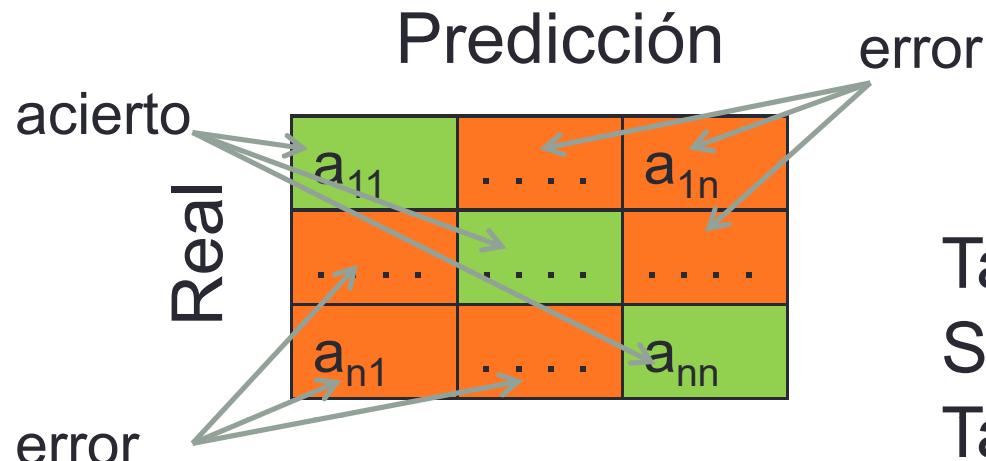
Nos centramos ahora en Clasificación.

## Minería de Datos. Evaluación → Clasificación

El conjunto de datos disponible se divide en:

- un **conjunto de entrenamiento -training-** (con el que se construye el modelo de clasificación) 70% p.ej.
- un **conjunto de prueba -testing-** (para evaluar el modelo) 30% p.ej.

Se aplica el modelo sobre el conjunto de prueba y se comprueba el número de aciertos.



$a_{ij}$  = Número de veces en los que la predicción es j cuando la clase correcta (real) es i

Tasa de Aciertos (**Accuracy**):  
Suma Diagonal Principal / Tamaño del conjunto de prueba

## Minería de Datos. Evaluación → Clasificación

- Obtención del conjunto de entrenamiento y de prueba:  
Debe contener un número de ejemplos de cada valor de la clase, proporcional a la presencia de éste → ***partición estratificada***
- Se consigue una mejor estimación de la tasa de aciertos si se repite la evaluación con distintas particiones → ***validación cruzada***

## Minería de Datos. Evaluación → Clasificación

- Frecuentemente, se está interesado en un valor específico de la clase (*P -Positive-*). En este caso (o cuando sólo hay dos valores de la clase), se usan otras medidas de evaluación sobre una tabla 2x2:

		Predicción	
		TP	FN
Real	acíerto	TP	FN
	error	FP	TN

P: Positive (un valor de la clase)  
N: Negative = Not P (cualquier otro)  
T: True = Acierto  
F: False = Fallo

**Precision for P =  $TP/(TP + FP)$**

**Recall for P =  $TP/(TP + FN)$**

# Índice

---

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Técnicas de Minería de Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos.

# Compendio de referencias a plataformas y lenguajes



---

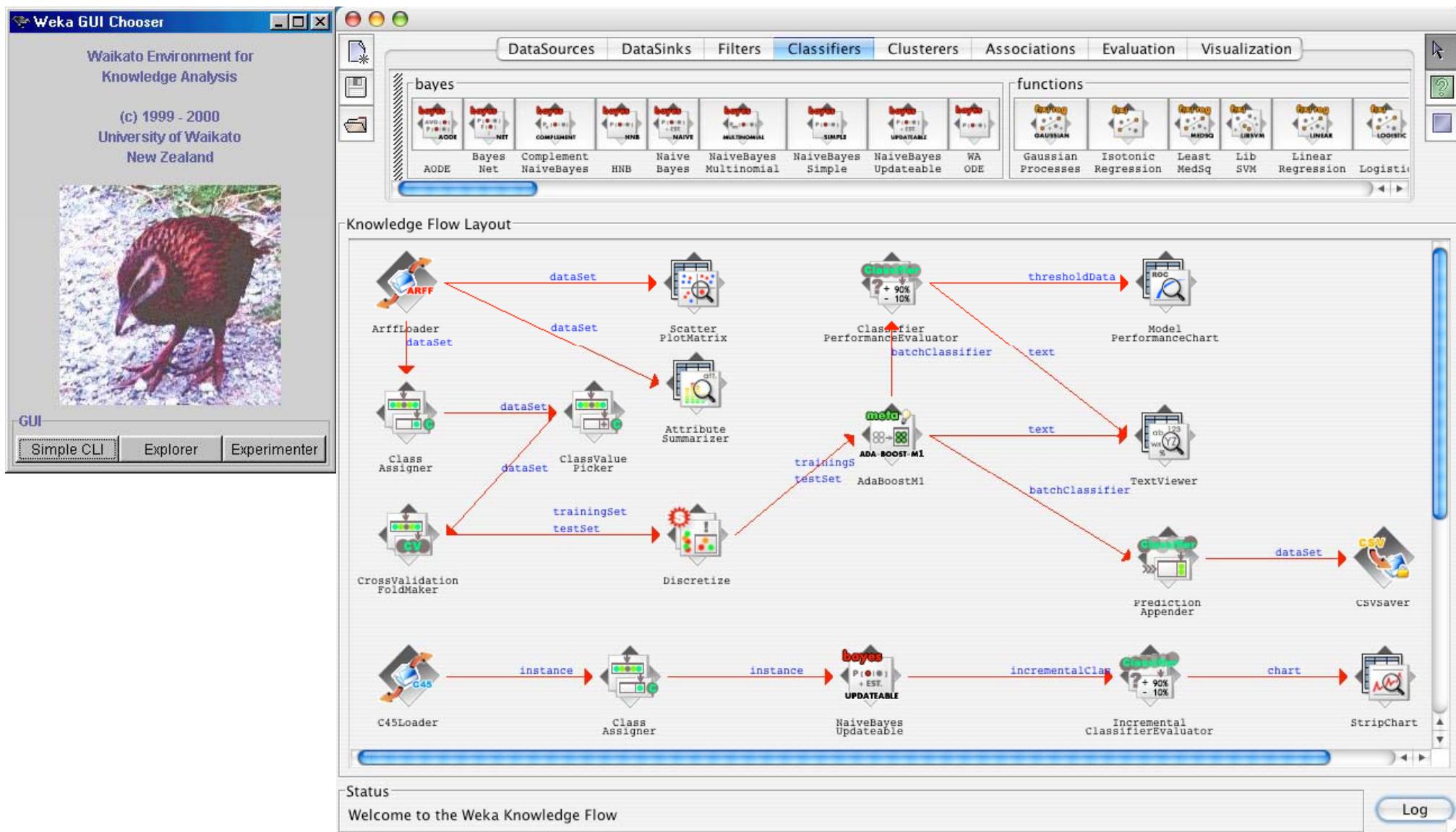
BLOG   BIG DATA COURSE   ADVICE   STARTUPS   USE CASES   SPEAKER   OPEN SOURCE   PUBLIC DATA   EVENTS   FORUM   ABOUT

---

<http://www.bigdata-startups.com/open-source-tools/>

# Entornos de Desarrollo Data Mining

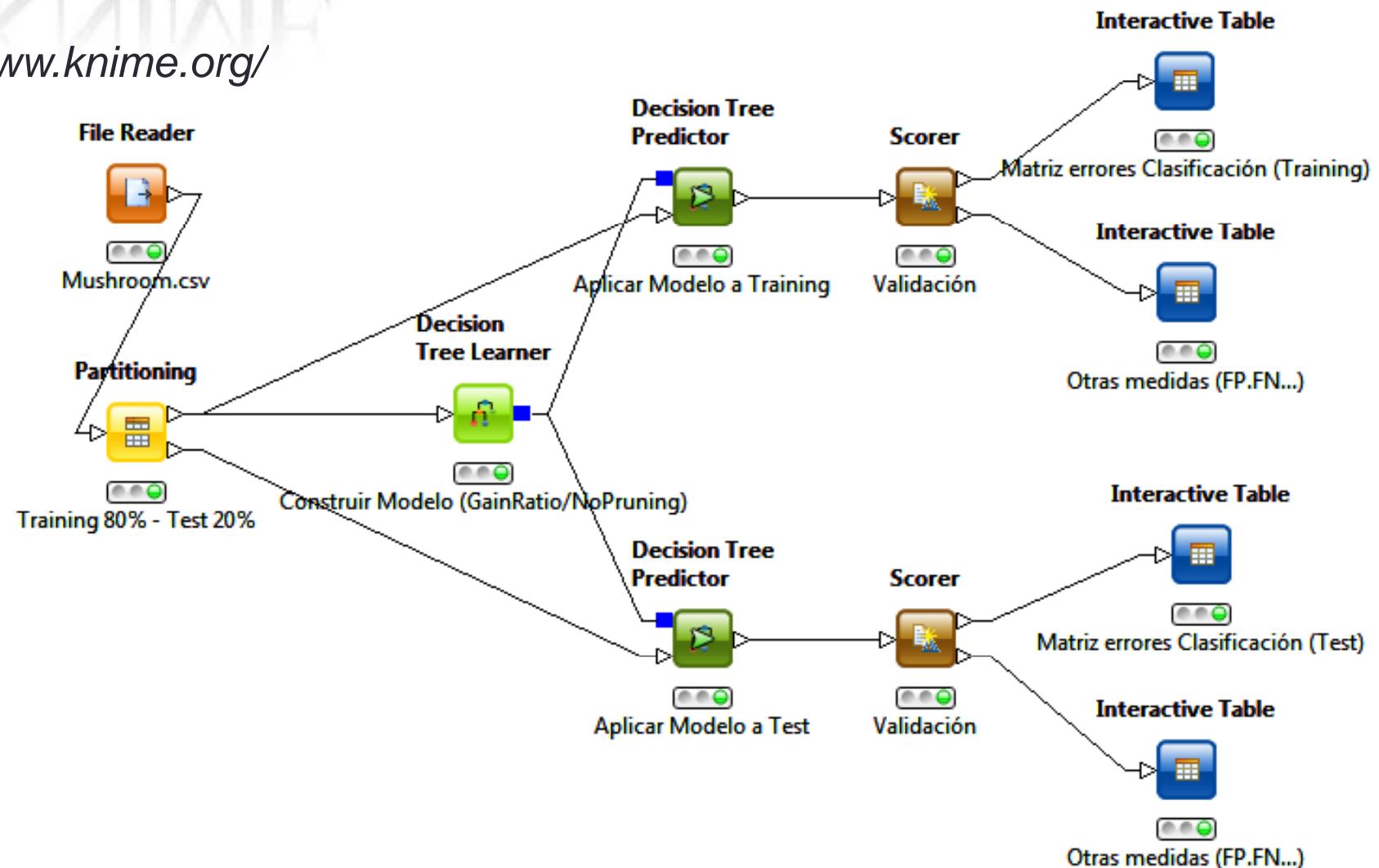
Weka <http://www.cs.waikato.ac.nz/ml/weka/>



# Entornos de Desarrollo Data Mining



<https://www.knime.org/>



# Entornos de Desarrollo Data Mining

Entornos similares propietarios



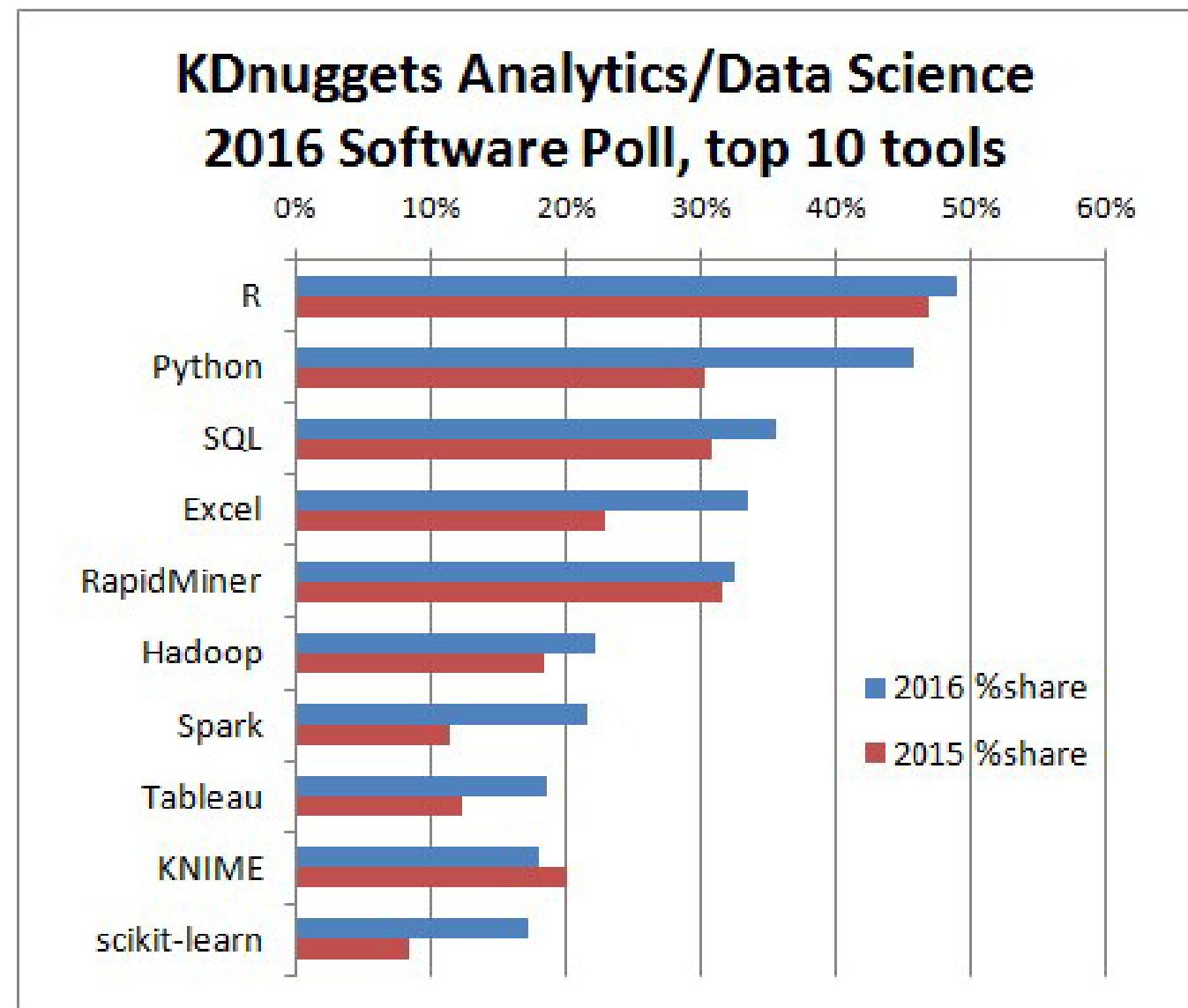
**Oracle Data Mining**  
Powering Next-Generation Predictive Applications



SAS® Enterprise  
Miner™

# Lenguajes Data Mining

The current data suggests that while Python is more popular than R as a general-purpose programming language, R is more popular than Python for data analysis.



# Lenguajes Data Mining



The Comprehensive R Archive Network

*cran.r-project.org/*

```
mydata      = read.arff(paste0(DIRECTORIO_DATOS,"\\Otros\\churn.arff"))
class.name  = "LEAVE"
class.index = grep(class.name, colnames(mydata))
myformula   = formula(paste(class.name,"~ ."))
myclass     = mydata[,class.index]
```

```
set.seed(123)
trainIndex = createDataPartition(myclass, p = .7, list = FALSE)
training   = mydata[trainIndex, ]
testing    = mydata[-trainIndex, ]
```

```
decision.tree.model = rpart(myformula, training, parms=list(split="information"))
decision.tree.predictions = predict(decision.tree.model, newdata = testing,
                                     type = "class")
prp (decision.tree.model, type = 2, , extra = 104 ,nn=TRUE,
      fallen.leaves=TRUE,faclen=0,varlen=0,shadow.col="grey",branch.lty=3)
```

# Lenguajes Data Mining



The Comprehensive R Archive Network

*cran.r-project.org/*

The screenshot shows the RGui interface. On the left, the R console window displays the following R code:

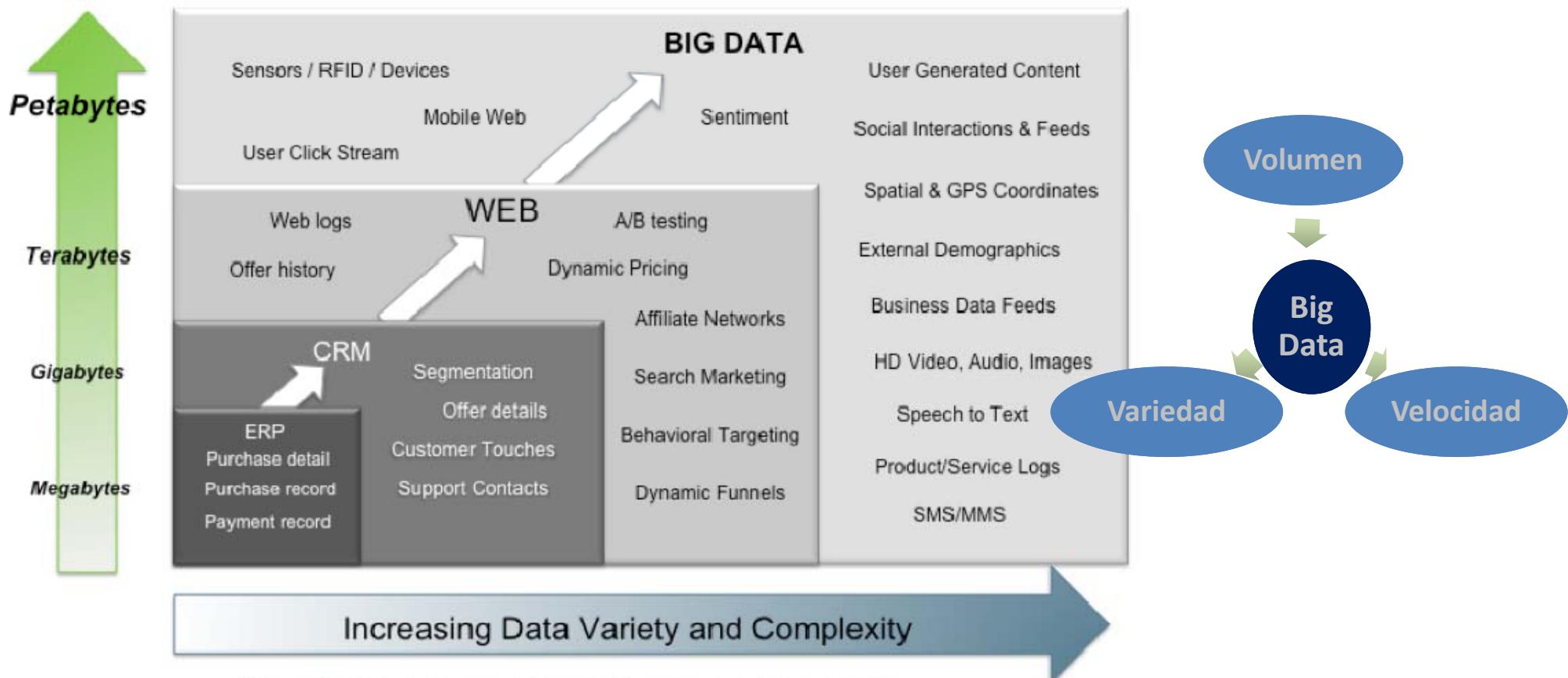
```
> mydata = read.csv(paste0(DIRECTORIO_DATOS, "\\VOTOS\\Chus"))
> class.name = "LEAVE"
> class.index = grep(class.name, colnames(mydata))
> myformula = formula(paste(class.name,"~ ."))
> myclass = mydata[,class.index]
>
> set.seed(123)
> trainIndex = createDataPartition(myclass, p = .7, list = FALSE)
> training = mydata[trainIndex, ]
> testing = mydata[-trainIndex, ]
>
> #####
> # rpart
>
> decision.tree.model = rpart(myformula,
+                             training,
+                             parms=list(split="information"))
>
> # label núm.tuplas en el nodo | núm.errores abs. | classific$
> # 1) root 14001 6897 STAY (0.4926077 0.5073923)
> # 2) HOUSE< 601793 9289 3893 LEAVE (0.5809021 0.4190979)
>
> decision.tree.predictions = predict(decision.tree.model,
+                                       newdata = testing,
+                                       type = "class") # "probs"
>
> prp (decision.tree.model,
+       type = 2, , extra = 104,nn =TRUE,fallen.leaves=TRUE,fact$
```

On the right, the R Graphics window titled "R Graphics: Device 2 (ACTIVE)" displays a decision tree diagram. The tree has a root node labeled "STAY 6897 0.5073923". It branches into two nodes: "HOUSE< 601793" leading to "STAY 9289 0.4190979" and "LEAVE 3893". The "STAY" node further branches into two leaf nodes labeled "STAY 3893 0.4190979". The "LEAVE" node also branches into two leaf nodes labeled "LEAVE 3893 0.4190979". The nodes are represented by rounded rectangles with a small tree icon above them.

# Big Data

“*Big Data*” son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos ...

Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

# Big Data → NoSql Databases

## BD SQL:

- Es crucial mantener la integridad referencial

## BD NoSQL:

- El rendimiento y poder responder en tiempo real prevalece sobre el mantenimiento de la integridad.
- Optimizadas para recuperar y agregar datos



# Big Data → Frameworks

Objetivo: Distribuir automáticamente la ejecución de procesos de análisis de datos y encargarse de las tareas de seguridad en red, de comunicación con el sistema de ficheros (HDFS), acceso a BD NoSQL (Hbase, Cassandra, etc), copias de seguridad, etc.



**HPCC Systems**



OpenDremel: Open source java implementation of Google BigQuery

[Project Home](#)

[Downloads](#)

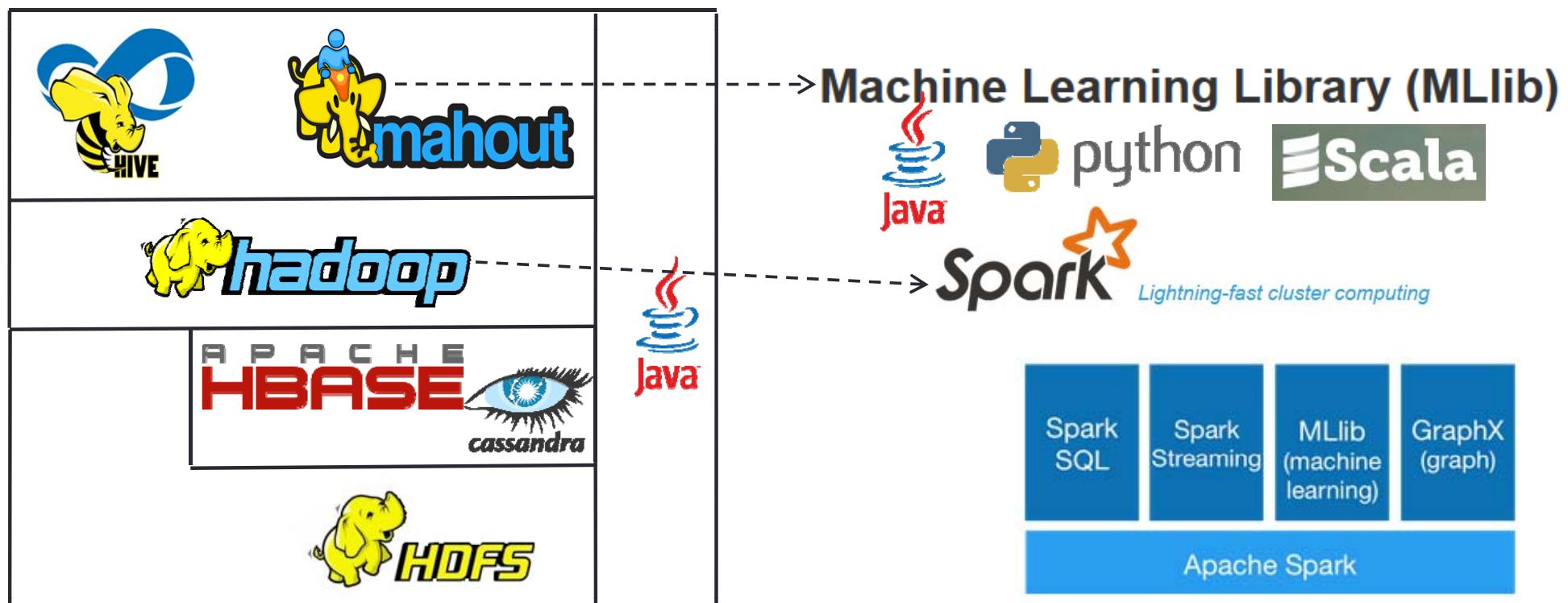
[Wiki](#)

[Issues](#)

[Source](#)

# Big Data → Frameworks

Sobre los Frameworks se han desarrollado otro nivel de librerías y plataformas para gestión de bases de datos multidimensionales, desarrollo de algoritmos de DataMining, etc



# Testing → Data Mining en general

<https://archive.ics.uci.edu/ml/datasets.html>

The screenshot shows the homepage of the UCI Machine Learning Repository. At the top, there is a logo for UCI (University of California, Irvine) featuring a stylized anteater. Below the logo, the text "Machine Learning Repository" is displayed in large yellow letters, with "Center for Machine Learning and Intelligent Systems" in smaller text underneath. A sidebar on the left lists categories for "Default Task" (Classification 214, Regression 42, Clustering 36, Other 50) and "Attribute Type" (Categorical 36, Numerical 162, Mixed 56). On the right, there are two examples: "Abalone" with a small image of an abalone shell and "Adult" with a small image of a person's face.

<http://people.sc.fsu.edu/~jb Burkardt/datasets/datasets.html>

<http://www.inf.ed.ac.uk/teaching/courses/dme/2014/datasets.html>

<http://vincentarelbundock.github.io/Rdatasets/>

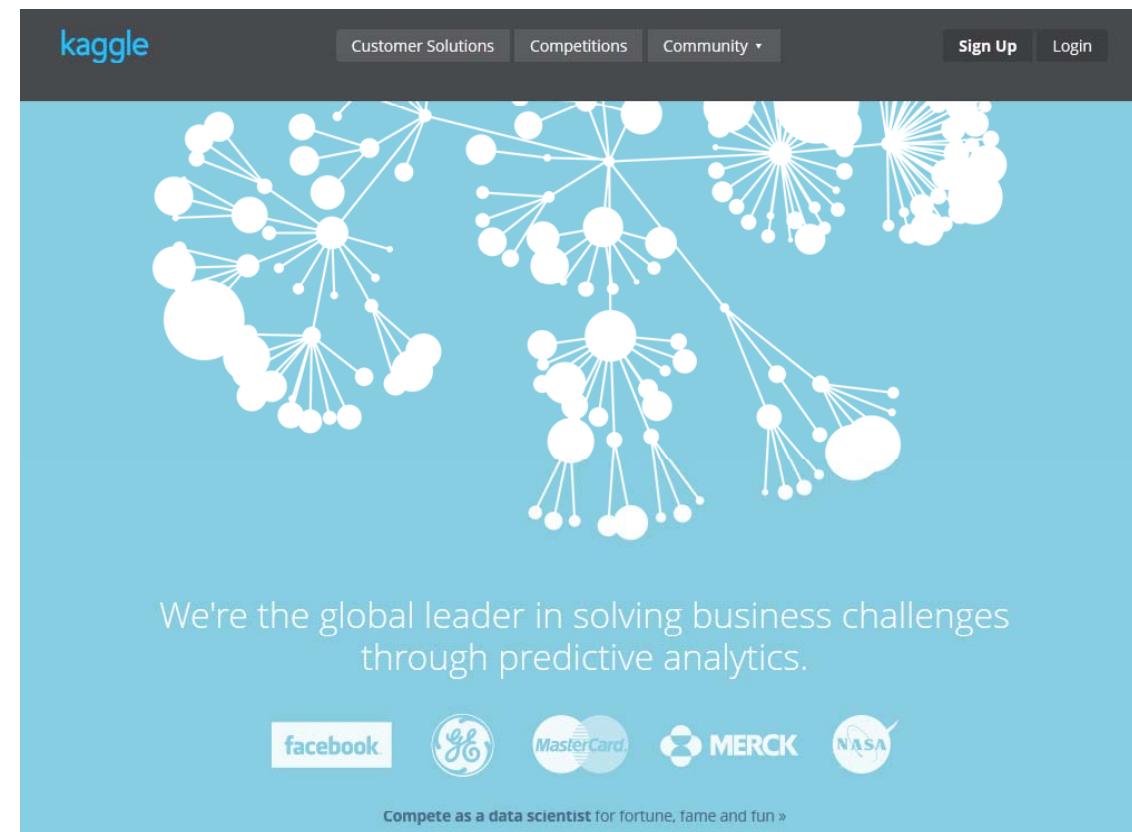
# Testing → Clasificación

## Kaggle: Go from Big Data to Big Analytics

<http://www.kaggle.com/>

Es una empresa con  
un website que ofrece  
competiciones,  
ofertas de empleo,

...



# Testing → Clasificación

## Kaggle: Go from Big Data to Big Analytics

Active Competitions	Competition Name	Reward	Teams	Deadline
All Competitions	 American Epilepsy Society Seizure Prediction Challenge Predict seizures in intracranial EEG recordings	\$25,000	306	28 days
14 found, 14 active	 Africa Soil Property Prediction Challenge Predict physical and chemical properties of soil using spectral measurements	\$8,000	1241	37 hours
<input type="radio"/> All competitions <input checked="" type="radio"/> Enterable	 Tradeshift Text Classification Classify text blocks in documents	\$5,000	185	21 days
Status	 Learning Social Circles in Networks Model friend memberships to multiple circles	Knowledge	191	8.6 days
Sponsor	 Digit Recognizer Classify handwritten digits using the famous MNIST data	Knowledge	413	2 months
InClass (student competition)				

# Testing → Clasificación

Netflix Prize (2009):  
1 Millón dólares

Objetivo: Predecir la calificación de usuarios (user's ratings) sobre películas, basándose únicamente en calificaciones previas sobre otras películas.

Business Understanding:  
Fue crucial darse cuenta de cómo influía el factor tiempo en las calificaciones

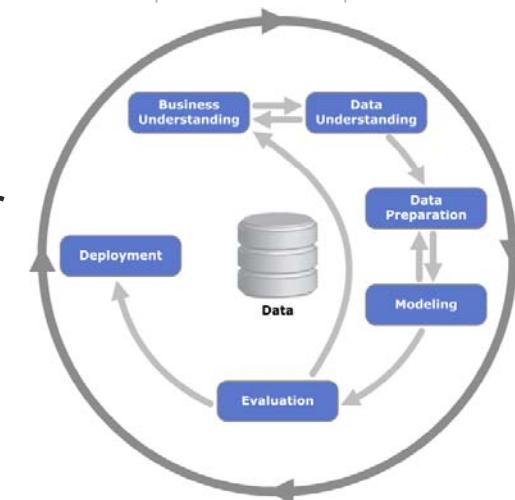


## Leaderboard

Showing Test Score. [Click here to show quiz score](#)

Display top  leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
<b>Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos</b>				
1	<a href="#">BellKor's Pragmatic Chaos</a>	0.8567	10.06	2009-07-26 18:18:28
2	<a href="#">The Ensemble</a>	0.8567	10.06	2009-07-26 18:38:22
3	<a href="#">Grand Prize Team</a>	0.8582	9.90	2009-07-10 21:24:40
4	<a href="#">Opera Solutions and Vandelay United</a>	0.8588	9.84	2009-07-10 01:12:31
5	<a href="#">Vandelay Industries !</a>	0.8591	9.81	2009-07-10 00:32:20
6	<a href="#">PragmaticTheory</a>	0.8594	9.77	2009-06-24 12:06:56
7	<a href="#">BellKor in BigChaos</a>	0.8601	9.70	2009-05-13 08:14:09
8	<a href="#">Dacc</a>	0.8612	9.60	2009-07-24 17:10:43



## Business Exploitation:

La explotación del conocimiento servirá a un experto en la toma de decisiones, pero no siempre será adecuada!

The screenshot shows a news article from [www.elmundo.es/papel/pantallas/2016/09/05/57cd40d5268e3e3f248b4633.html](http://www.elmundo.es/papel/pantallas/2016/09/05/57cd40d5268e3e3f248b4633.html). The title is "Big Data: el 'asesino' de los guionistas". The article features a photo of four men in suits walking down a hallway, each looking at their phone. Below the photo is a caption: "Después de un análisis de datos, Amazon lanzó Alpha House, pero no duró más de una temporada." To the left of the main content is a sidebar for the Lexus RX 450h Hybrid advertisement.

PANTALLAS

# Big Data: el 'asesino' de los guionistas

COMPARTIDO 425

4 COMENTARIOS

LIVE THE LIFE RX  
Lexus RX 450h Híbrido

DESCUBRA MÁS

Después de un análisis de datos, Amazon lanzó Alpha House, pero no duró más de una temporada.

→ Netflix, Marvel y los estudios de Hollywood utilizan algoritmos para ajustar sus tramas a los gustos del público. Internet y las redes sociales son su oráculo.

House of Cards: Ajuste del guión según el análisis de redes sociales 😊

Serie nueva: Alpha house.  
Sólo duró 1 temporada 😞

<http://www.elmundo.es/papel/pantallas/2016/09/05/57cd40d5268e3e3f248b4633.html>