

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,...)

The screenshot shows a web browser window with the URL www.kdnuggets.com/2016/10/top-10-data-science-videos-youtube.html. The page is titled "Data Mining, Analytics, Big Data, and Data Science" and features the KDnuggets logo. It includes navigation links for Software, News, Top stories, Opinions, Tutorials, Jobs, Academic, Companies, Courses, Datasets, and Education. The main content is titled "Top 10 Data Science Videos on Youtube". On the left, there's a sidebar for "Latest News, Stories" with several links. Below the main content are social sharing buttons and a summary of the video topics.

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

More News & Stories | Top Stories

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](#)

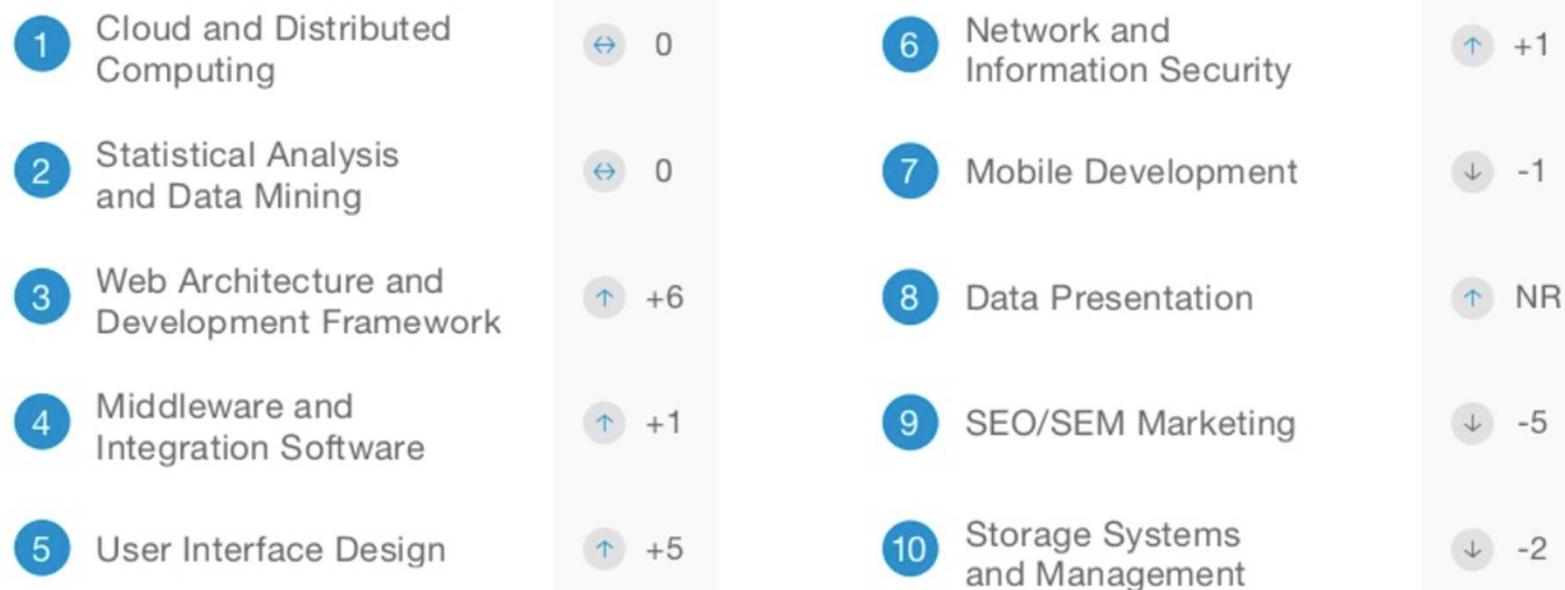
Learning and the future are the key topics in the recent Youtube videos on Data Science. The main questions revolve around: "how to become a Data Scientist", "what is a data scientist", and "where data science is going". But why there is so little explanation of data science to the masses?

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

LinkedIn



The Top Skills of 2016 on LinkedIn Global



* NR (Not recorded in 2015)



https://cincodias.elpais.com/cincodias/2018/04/09/midinero/1



t Visited



Getting Started

Más visitados

Comenzar a usar Firefox

Últimas noticias

Comenzar a usar Firefox

Mi dinero

[Empleo >](#)

Analista de datos, el puesto más demandado (y mejor pagado) por las empresas

- Las carreras con más salidas son ADE, Matemáticas y Estadística
- También crece la demanda de empleos en los sectores relacionados con el comercio 'online' y el turismo

☰ Harvard Business Review SEARCH

DATA

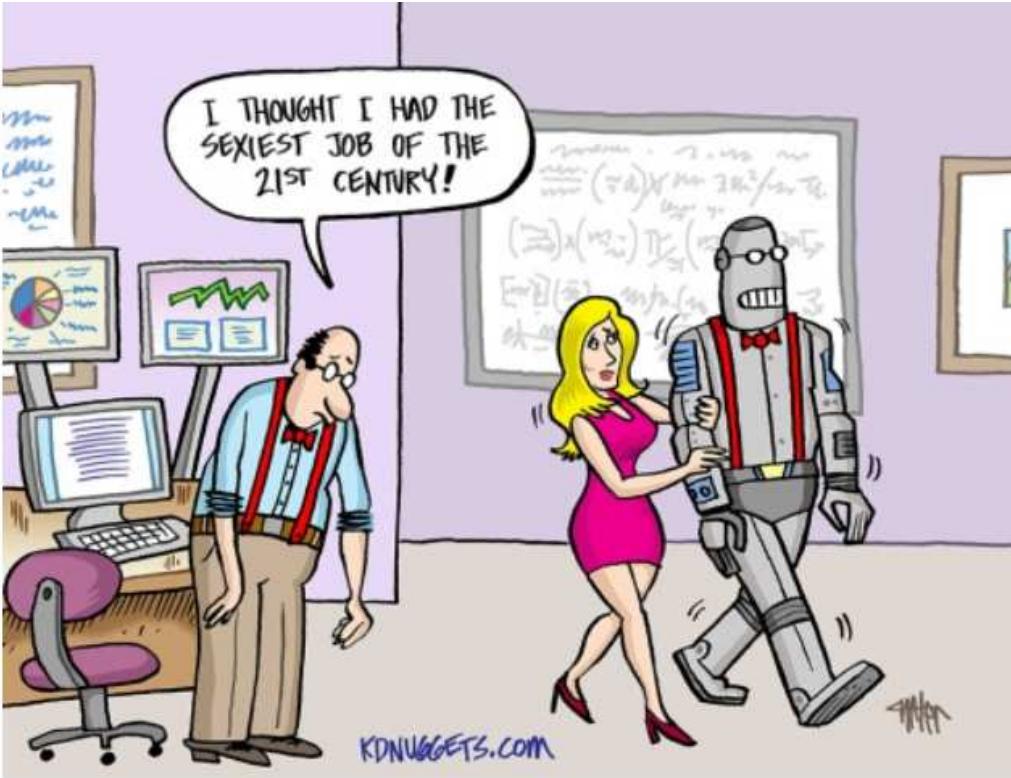
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

[SUMMARY](#) [SAVE](#) [SHARE](#) [COMMENT 16](#) [TEXT SIZE](#) [PRINT](#)

When Jonathan Goldman arrived for work in 2003 at LinkedIn, the business networking site, the company still felt like a start-up. The company had just 1 million accounts, and the number was growing quickly as early members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were



A cartoon illustration in a office setting. A woman with blonde hair in a pink dress is walking towards a man in a white shirt and red bow tie who is walking away. The man is carrying a briefcase and has a speech bubble above him that says "I THOUGHT I HAD THE SEXIEST JOB OF THE 21ST CENTURY!". In the background, there's a computer monitor displaying a pie chart, a whiteboard with mathematical equations, and a framed picture on the wall.

Juan Carlos Cubero. Universidad de Granada. Más

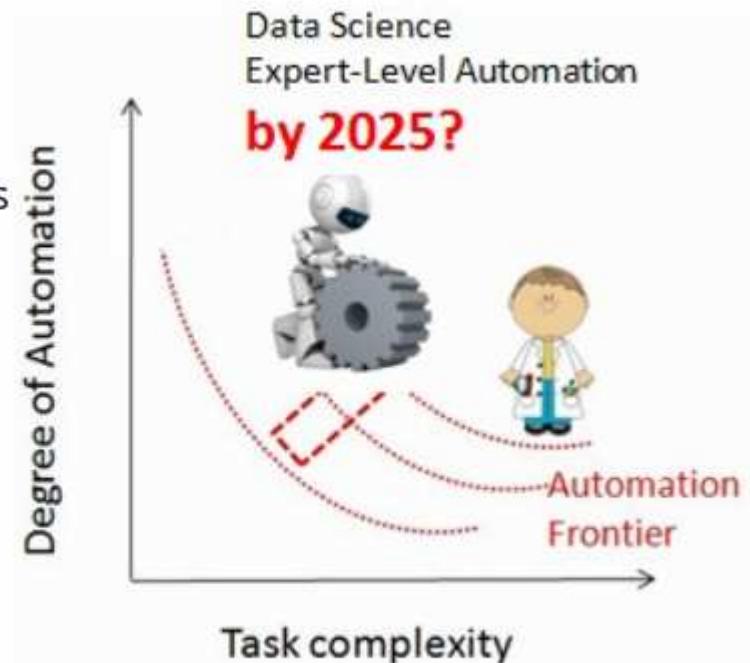
Data Scientist has been called the sexiest job of the 21st century. But perhaps the century will last only 25 years.

With even knowledge-based jobs like lawyers and accountants being **automated**, will Data Scientists prove to be an exception?

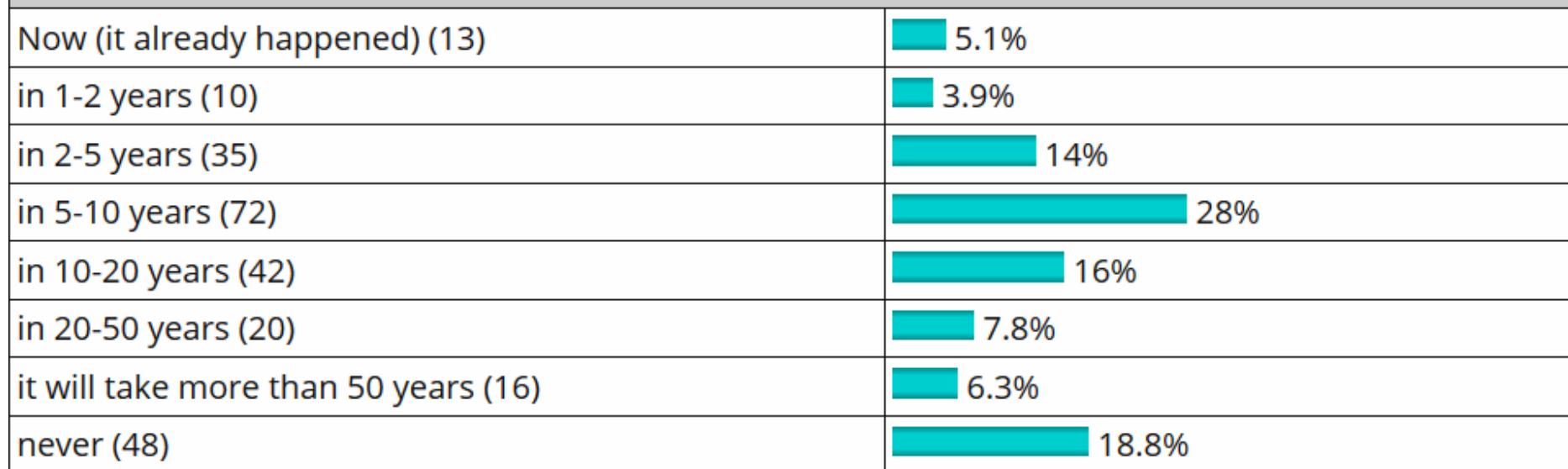
What predictive analytics professionals predict about the future of their profession?

Latest KDnuggets Poll asked:

When will most expert-level Predictive Analytics/Data Science tasks - currently done by human Data Scientists - be automated?



When will most expert-level Predictive Analytics/Data Science tasks - currently done by human Data Scientists - be automated: [255 voters]



InformationWeek

Join us live at
Interop 

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)

WIKIBON ANALYSTS AGENDA BLOG EVENTS RESEARCH ▾

The worldwide BDA market grew at 24.5% in 2017 vs 2016, faster than we forecasted in last year's report as a result of better than expected public cloud deployment and utilization as well as progress in convergence of tools. Enterprises are moving more rapidly out of the experimentation and proof-of-concept phases to achieving higher levels of business value from their big data deployments.

Looking forward, the overall the BDA market will grow at an 11% compounded annual growth rate (CAGR) to \$103B by 2027 (Figure 1). Edge computing – including streaming and ML app deployments on smart devices will boost the market in the out years.

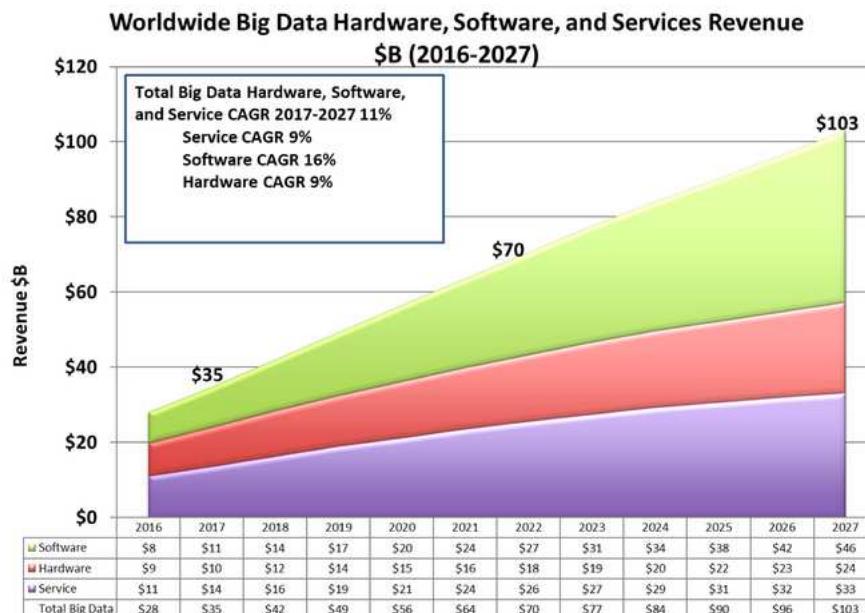
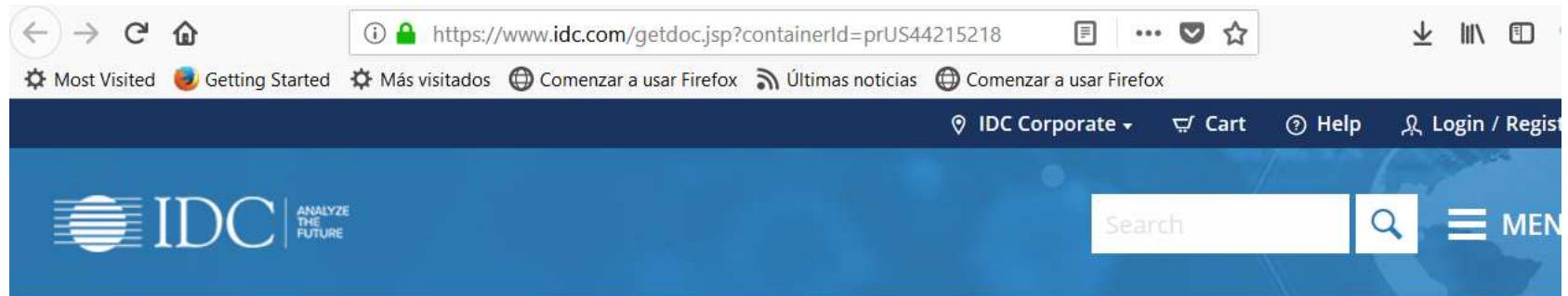


Figure 1: Worldwide Big Data Hardware, Software, and Services Revenue \$B
2016-2027



15 Aug 2018

Revenues for Big Data and Business Analytics Solutions Forecast to Reach \$260 Billion in 2022, Led by the Banking and Manufacturing Industries, According to IDC

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



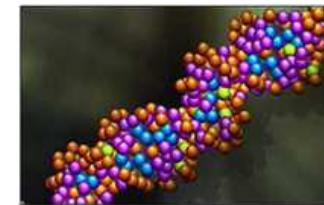
Ben Chams - Fotolia

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

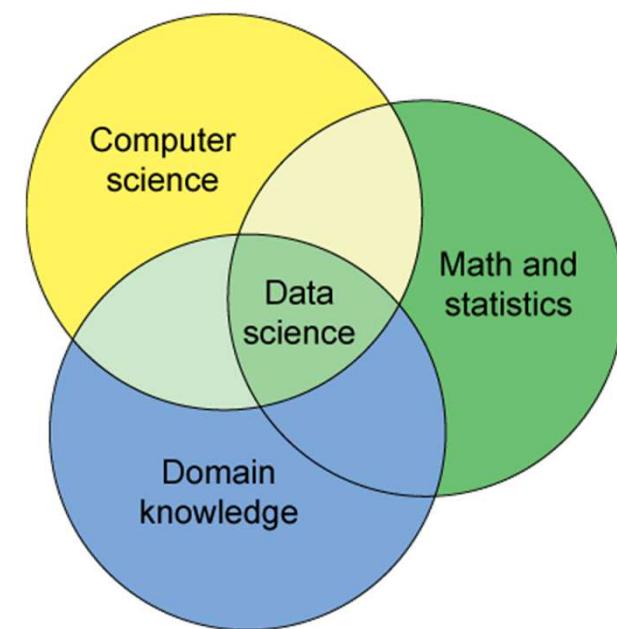
El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

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

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

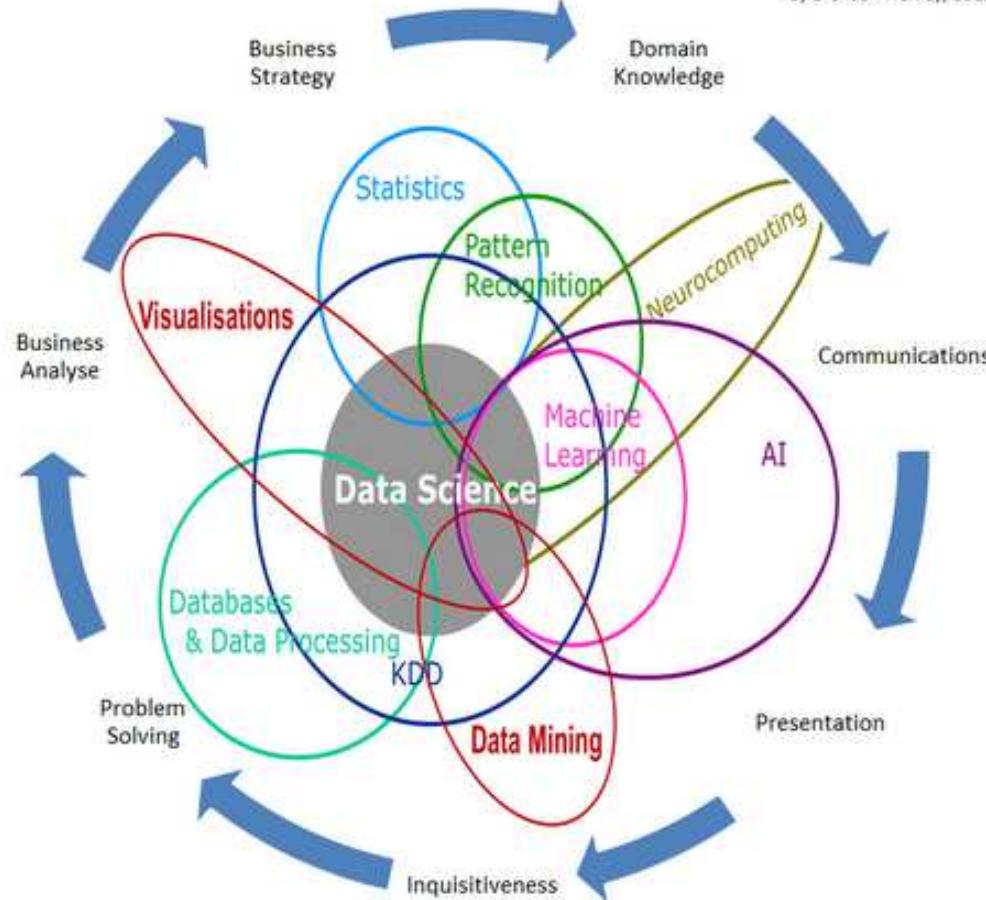
¿Qué es un Científico de Datos?

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

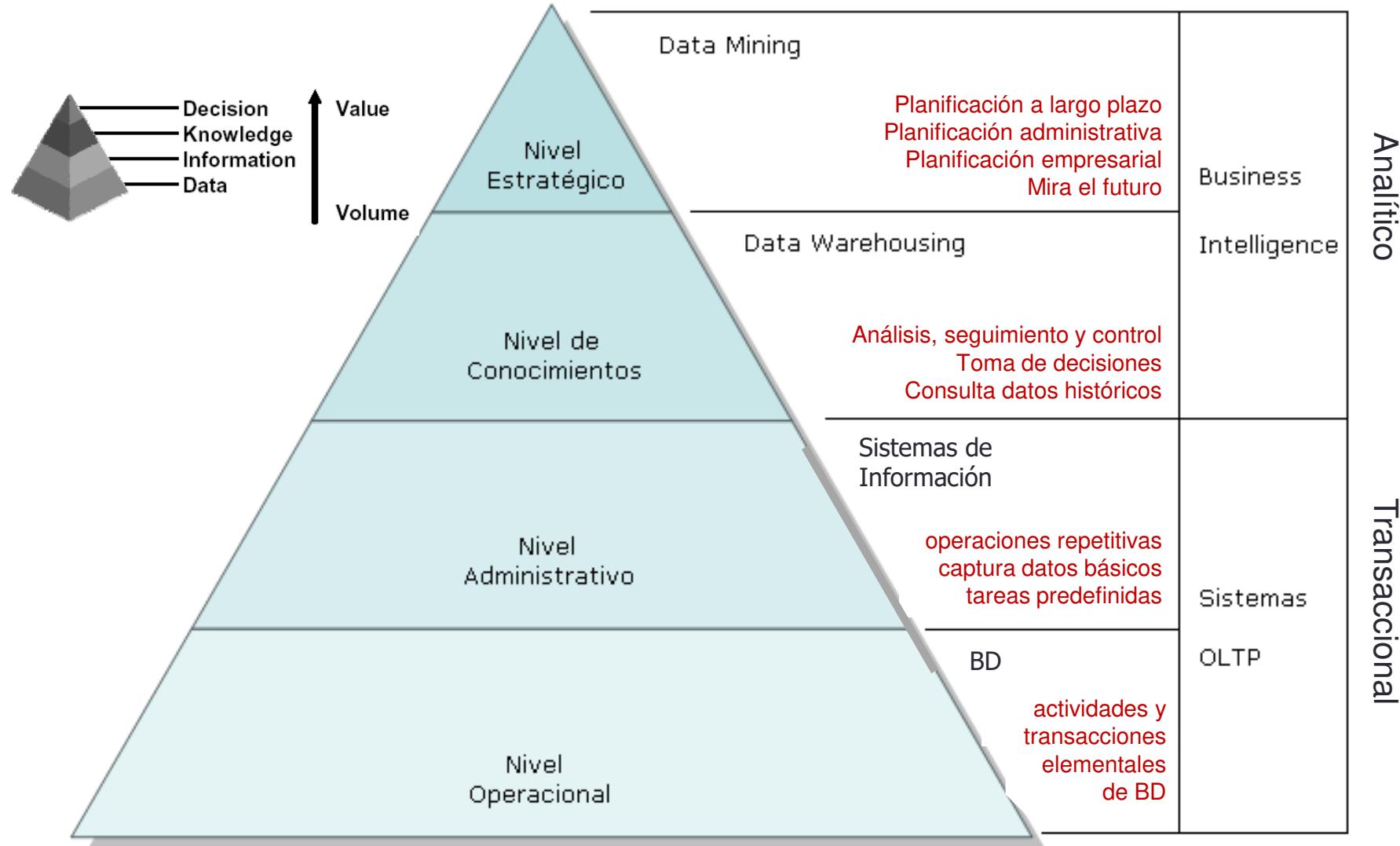


Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Contexto



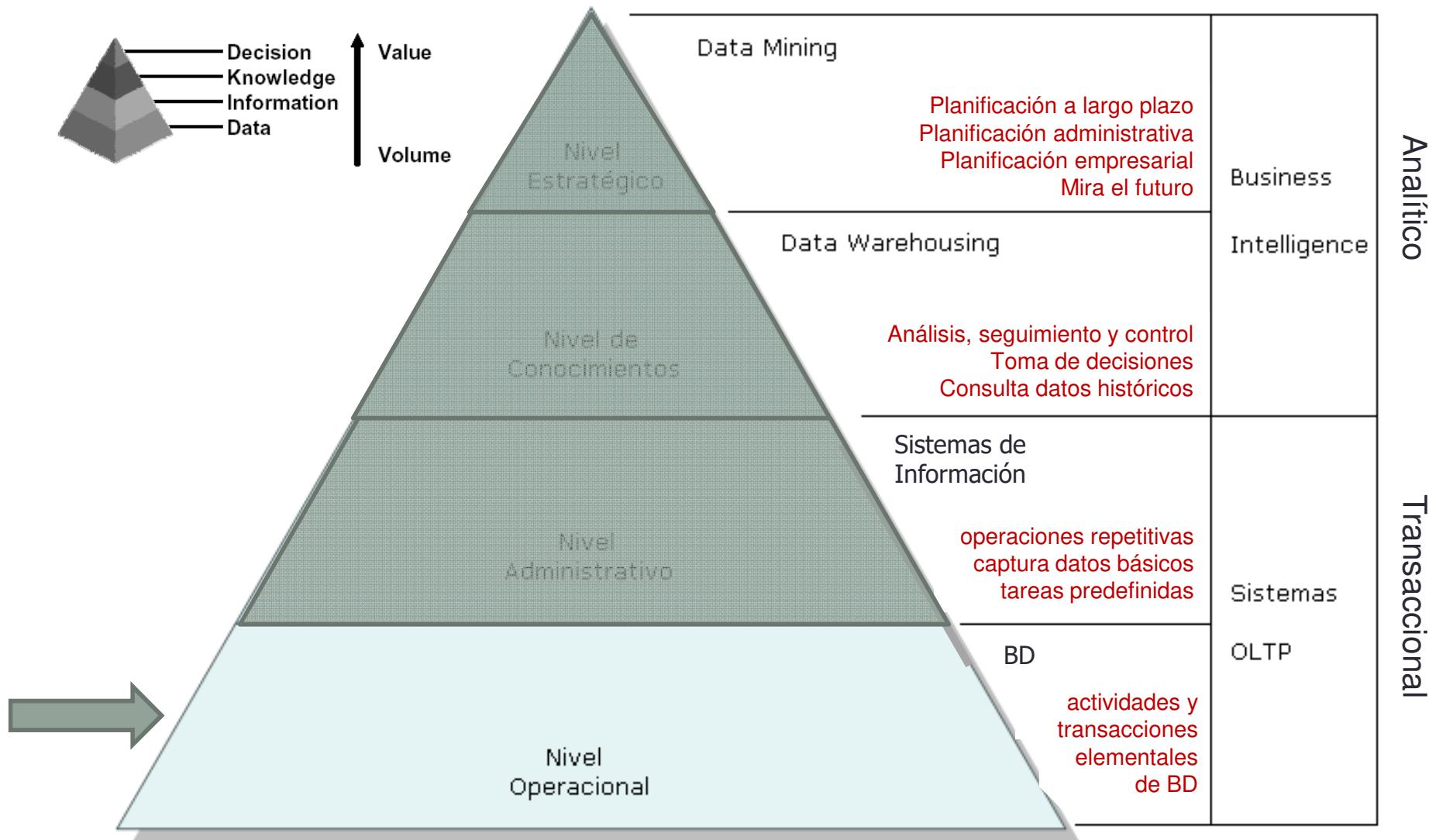


Tabla *Store_Information*

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

Store_Name
Los Angeles
San Diego
Los Angeles
Boston

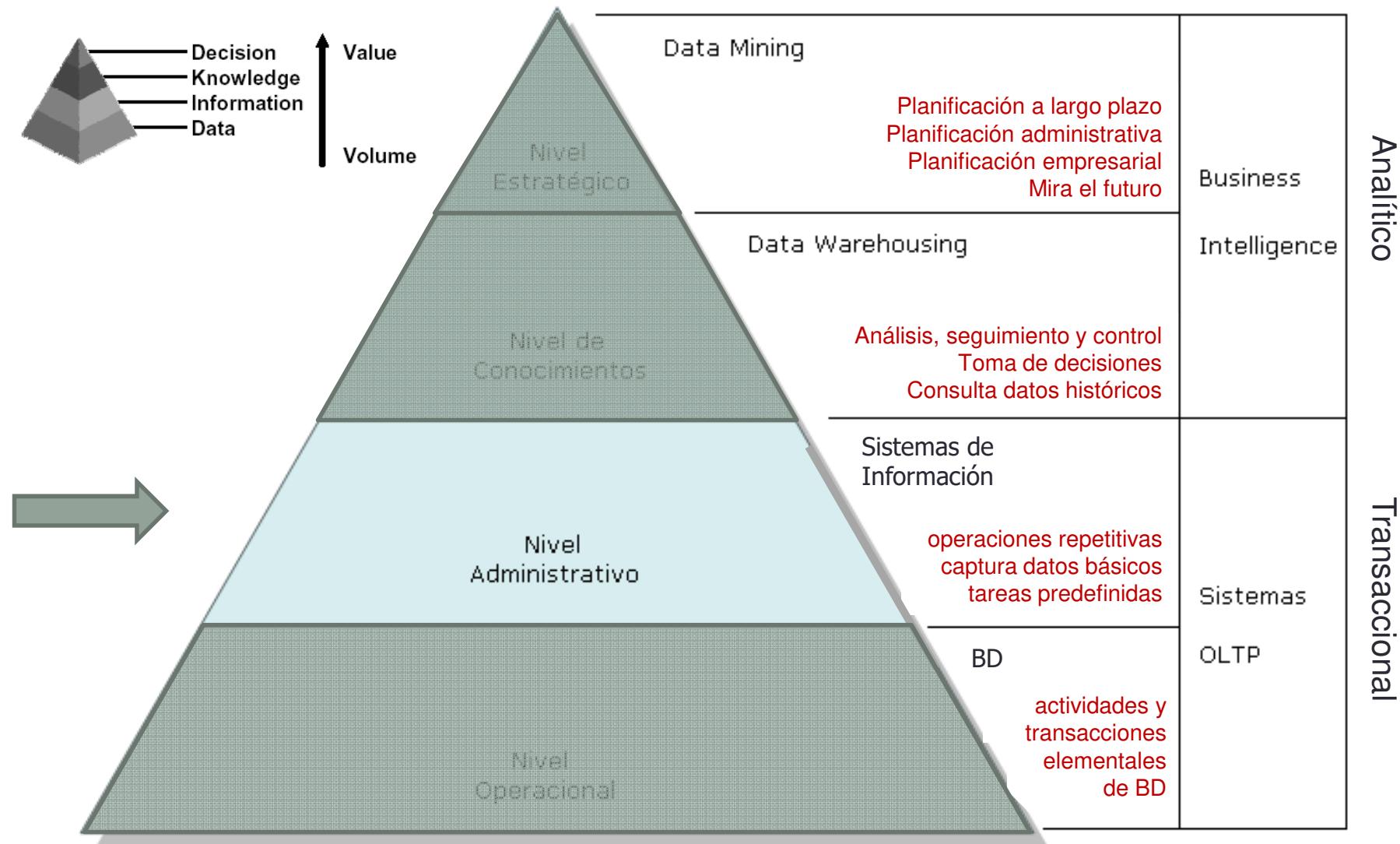
SELECT Store_Name FROM Store_Information;

Store_Name
Los Angeles

SELECT Store_Name
FROM Store_Information
WHERE Sales > 1000;

Store_Name SUM(Sales)
Los Angeles 1800
San Diego 250
Boston 700

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





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

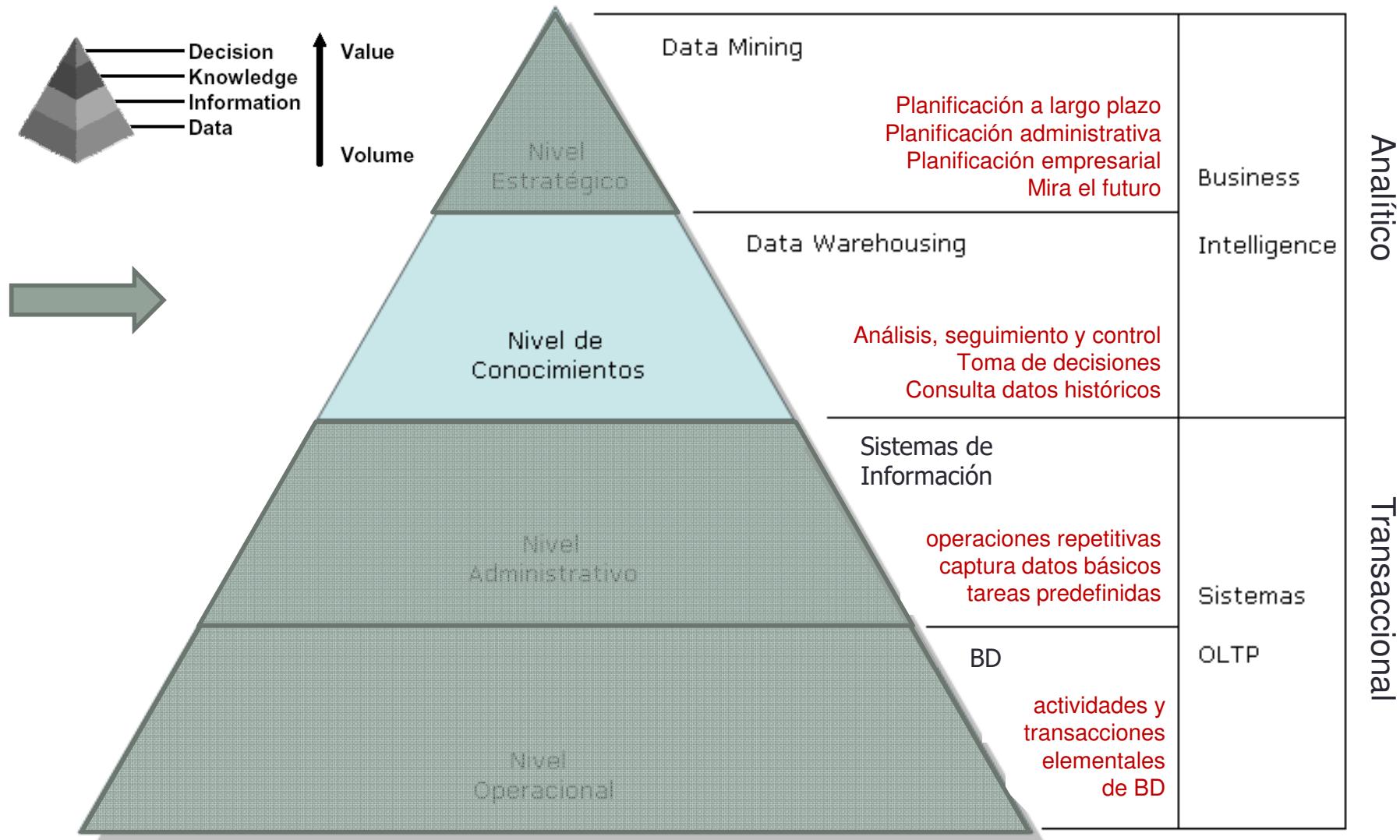


Tabla *Store_Information*

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;
```

Store_Name
Los Angeles

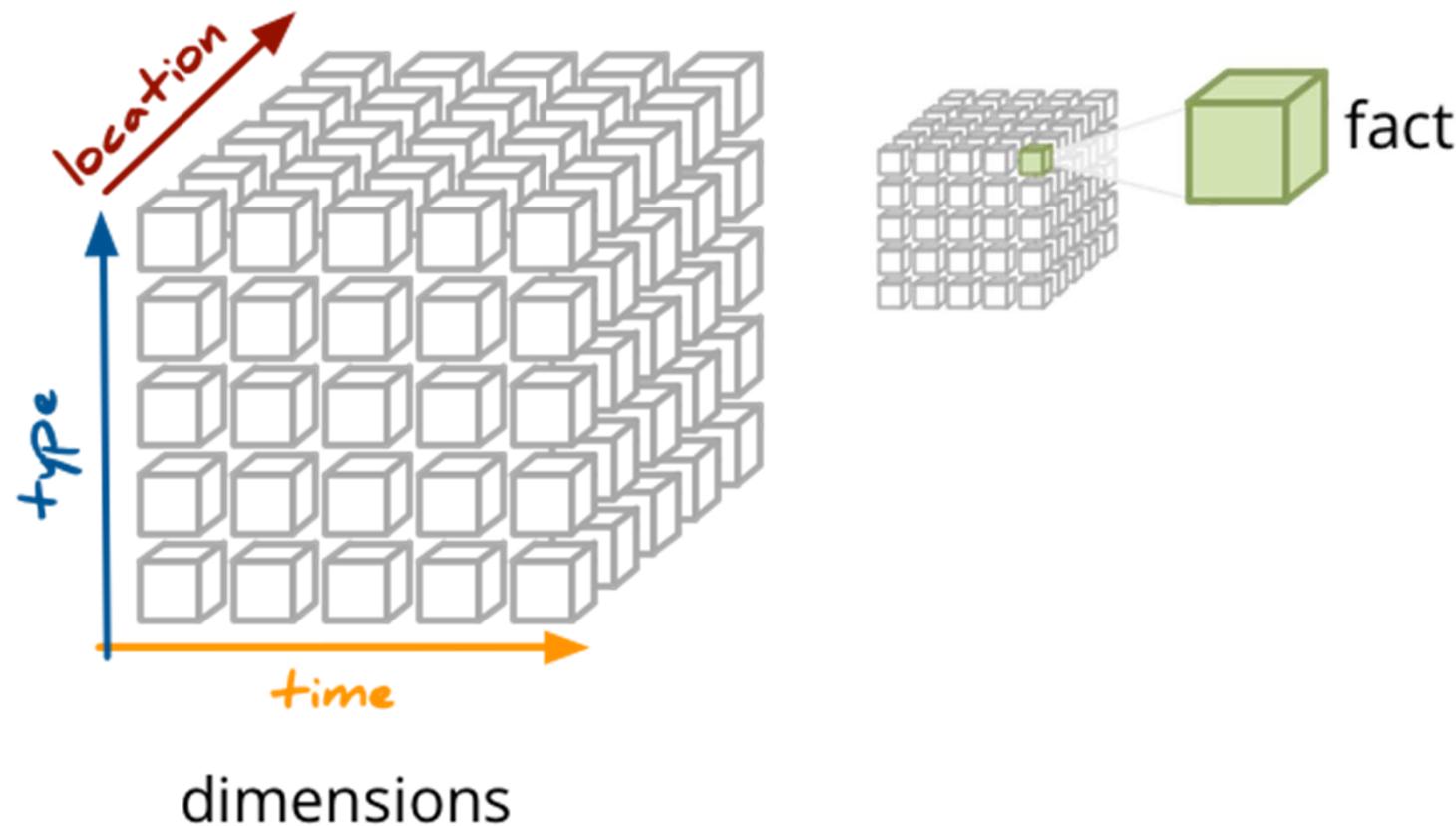
```
SELECT Store_Name  
FROM Store_Information  
WHERE Sales > 1000;
```

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

Store_Name SUM(Sales)
Los Angeles 1800
San Diego 250
Boston 700

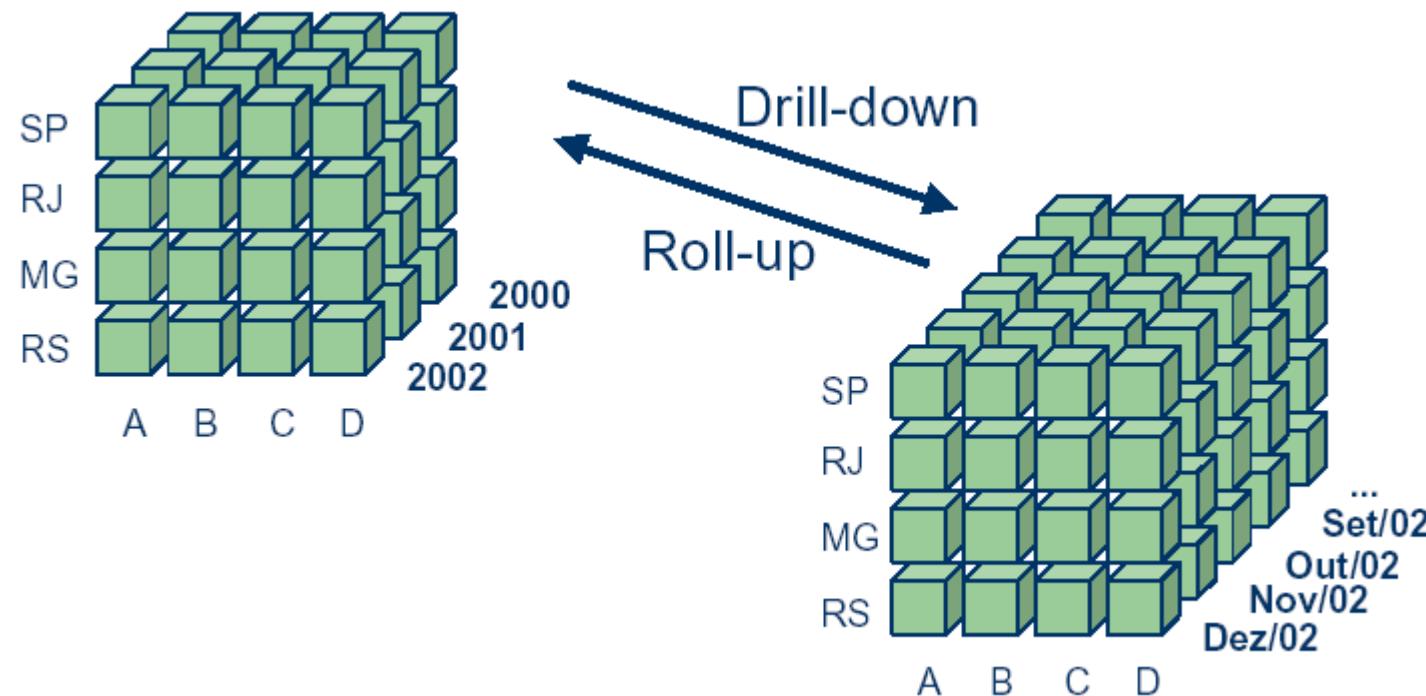
OLAP/Data Warehousing

Hechos y dimensiones

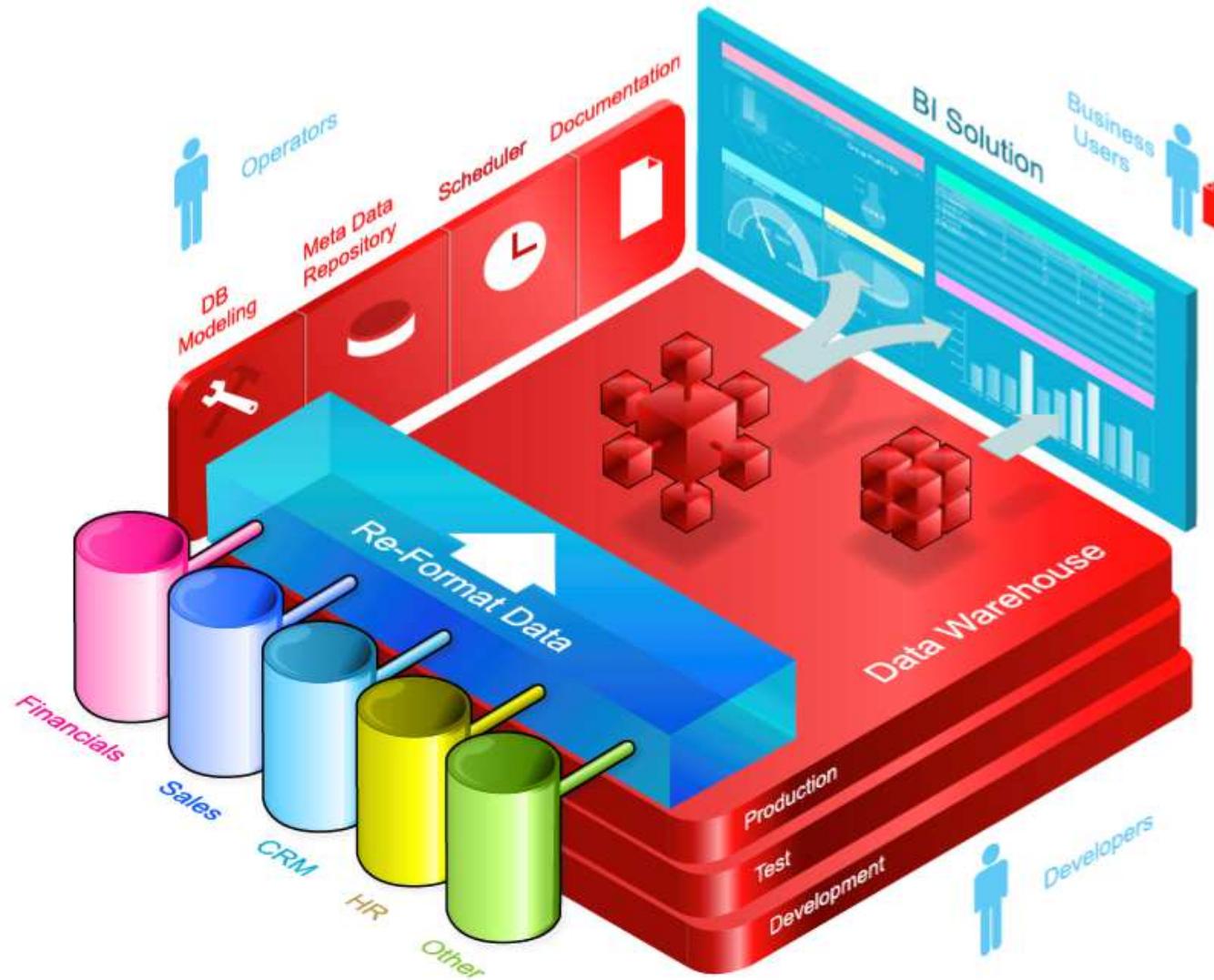


OLAP/Data Warehousing

Agregación



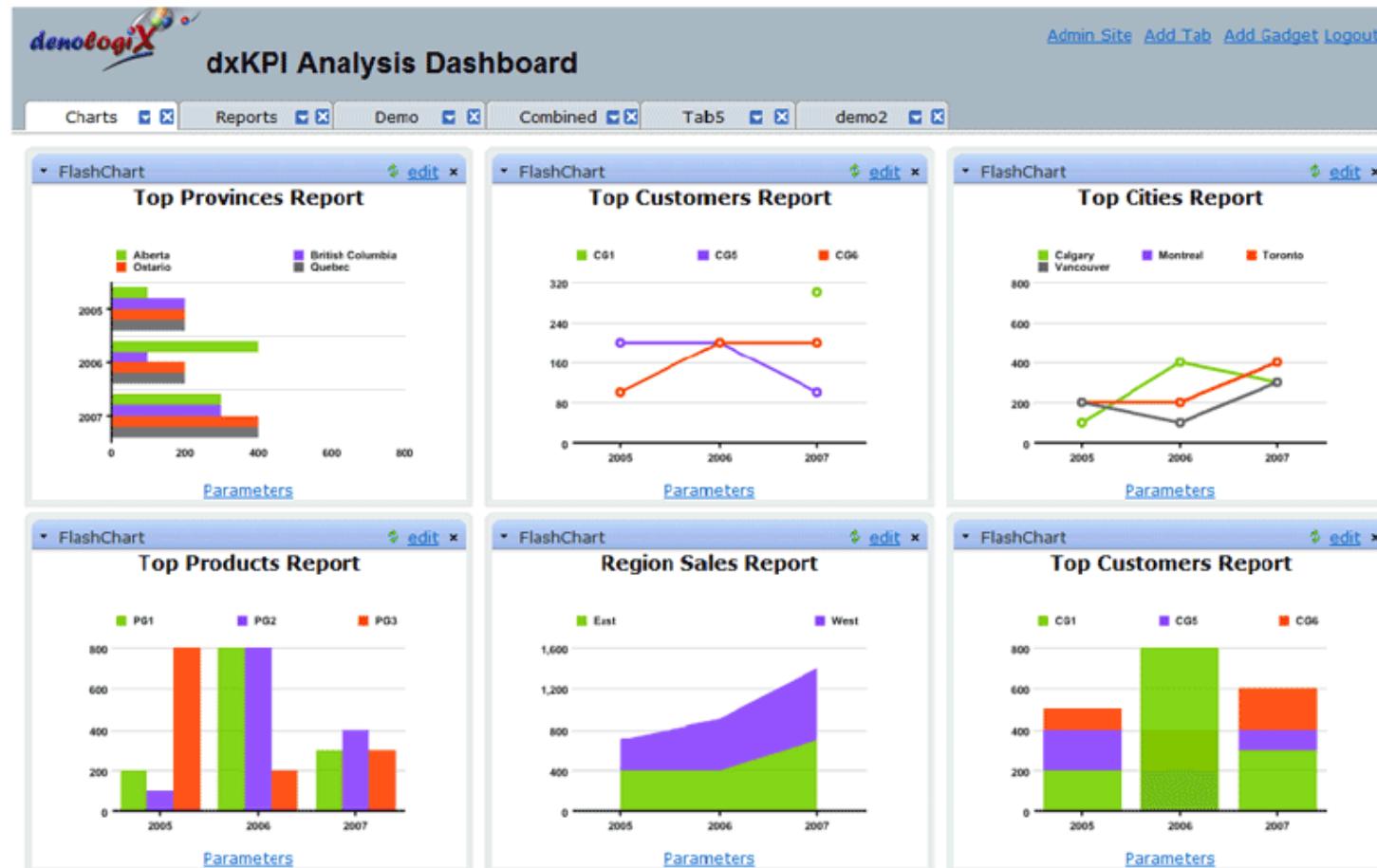
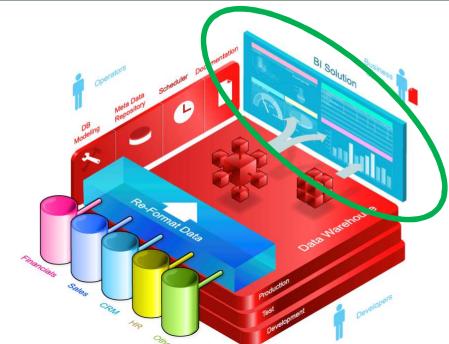
OLAP/Data Warehousing

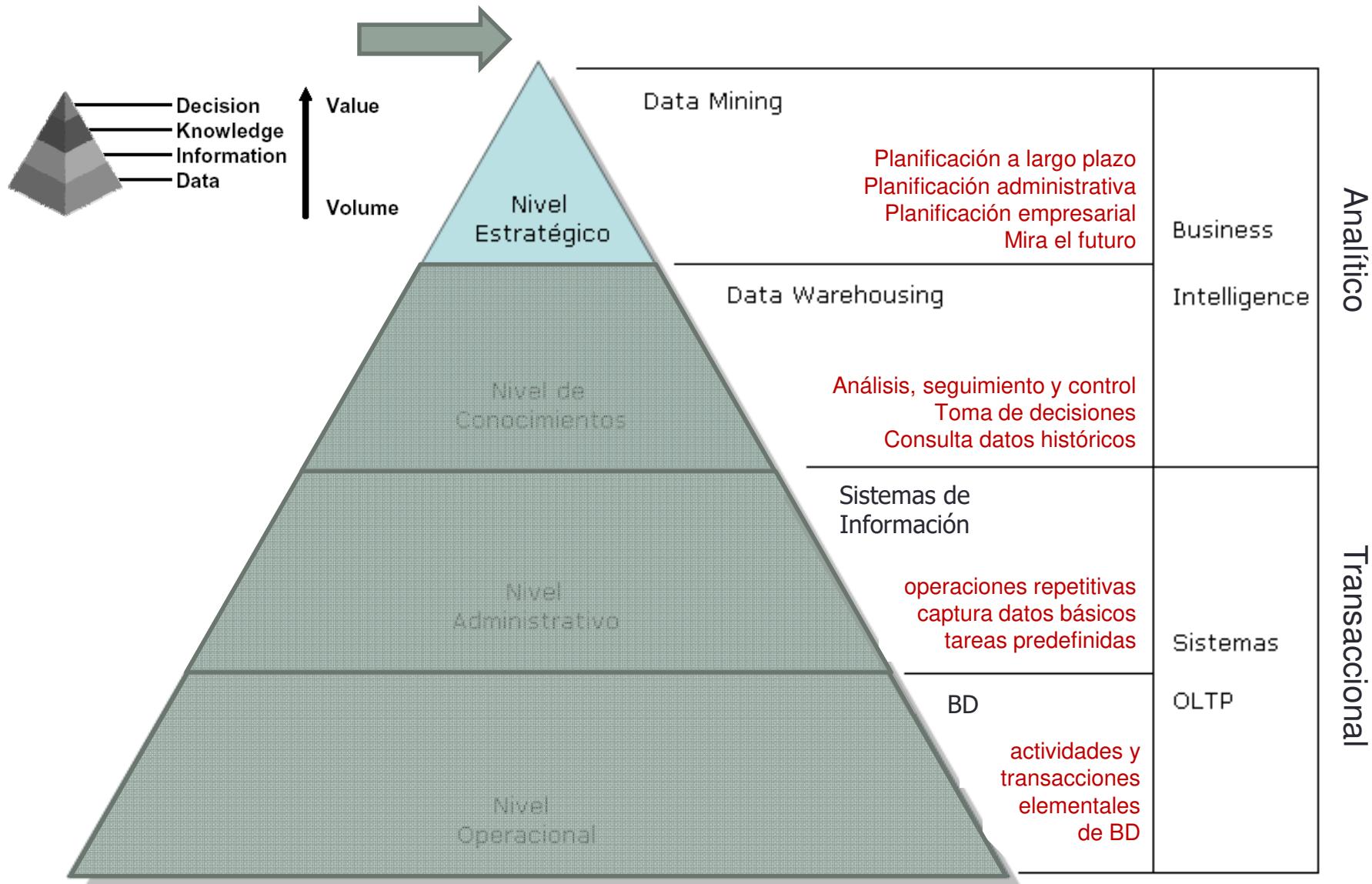


OLAP/Data Warehousing

ScoreBoards/KPI (Key Performance Indicator)

Real-time access to your KPIs





Sistemas OLTP

Outcome	Type	Velocity	pfx	pfxX	pfx_xfxZ	Loc_X	Loc_Z	Count
LD	CH	87.3	-7.7	5.9	-0.3	2.4	11	
FB	CH	87.5	-8.5	4.3	-0.2	2.5	10	
GB	CH	87.2	-7.9	5.6	-0.4	2.4	37	
ball	CH	87.7	-8.2	4.0	-0.1	2.1	104	
StrF	CH	87.7	-7.6	3.8	-0.3	2.1	56	
StL	CH	87.5	-8.4	3.7	-0.2	2.6	39	
StrS	CH	88.0	-7.9	5.6	-0.3	1.8	58	
LD	CU	82.1	5.7	-7.0	-0.3	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	
StL	CU	81.8	6.1	-6.0	-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	2.1	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	
StL	FA	94.5	-7.9	7.9	-0.2	2.5	264	
StrS	FA	95.4	-6.5	6.4	-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	
StL	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 Warehousing/OLAP

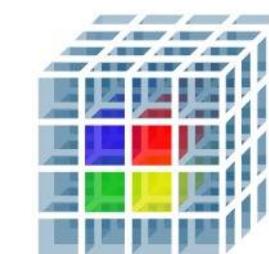
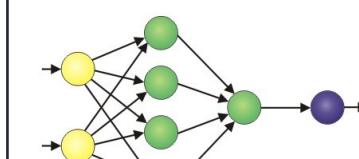


Data Analysis



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.

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
- visualizar datos altamente dimensionales, extrayendo estructura local simplificada, ...

Nuevas necesidades de análisis datos

Minería de Datos. ¿Qué es?

¿A qué tipos de datos puede aplicarse DM?

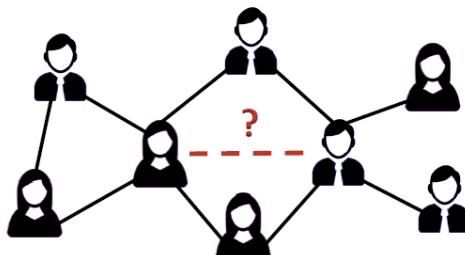
En principio, a cualquier tipo

- 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?

Tera/Peta bytes de datos:

- Compras relacionadas.
- Perfiles de usuario
- Segmentación de clientes
- Detección de intrusos / Fraudes
- Pronóstico Tiempo
- Predicción estructura AND
- Predicción de enlaces
- etc



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

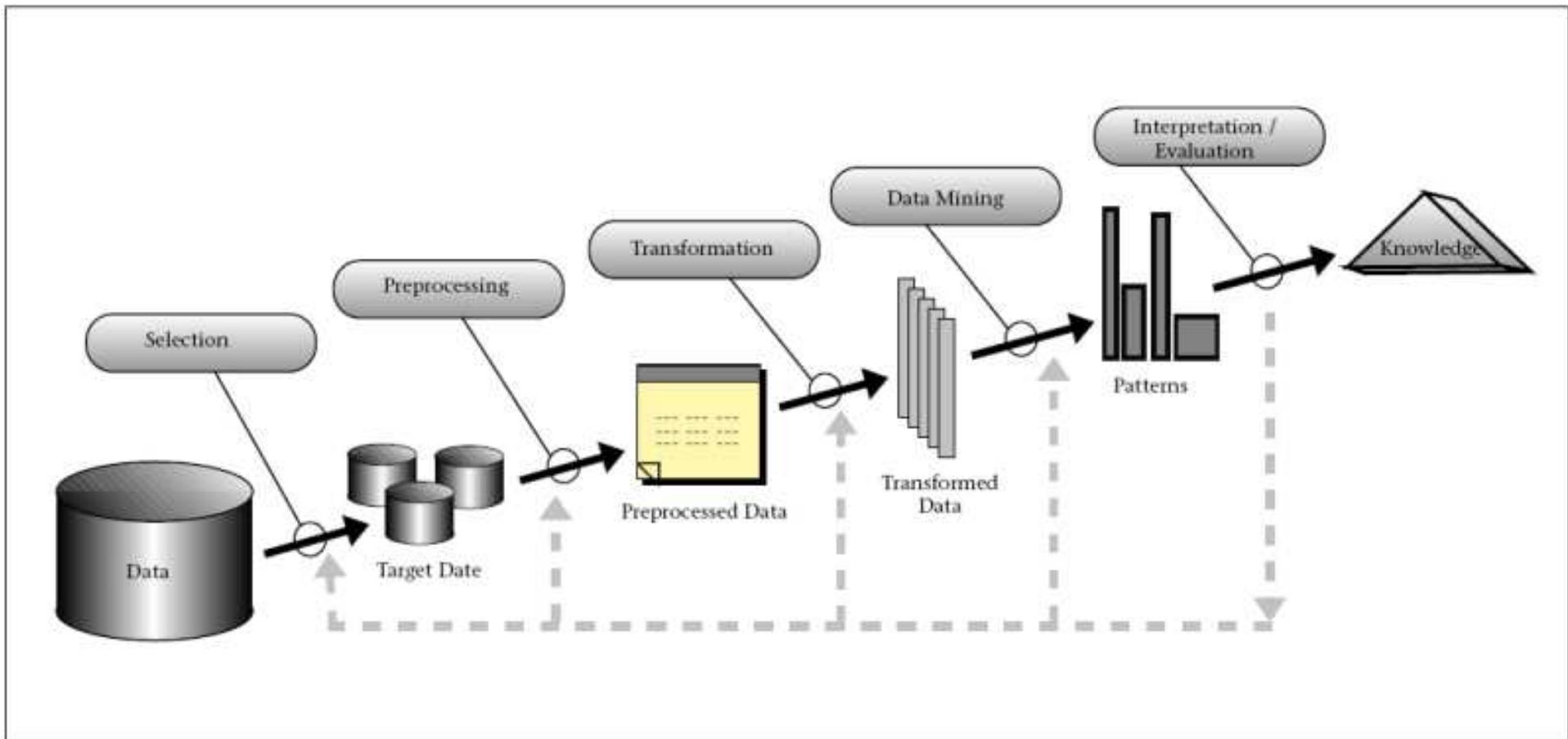
- Análisis del comportamiento y perfiles del visitante
- Potenciar la venta cruzada (cross-selling)
- Generación de respuestas agrupadas según el tipo de contenido
- Recuperación de información (information retrieval) Búsqueda de metadatos que describan los documentos.
- Recuperación inteligente de datos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

Minería de Datos. Áreas de aplicación

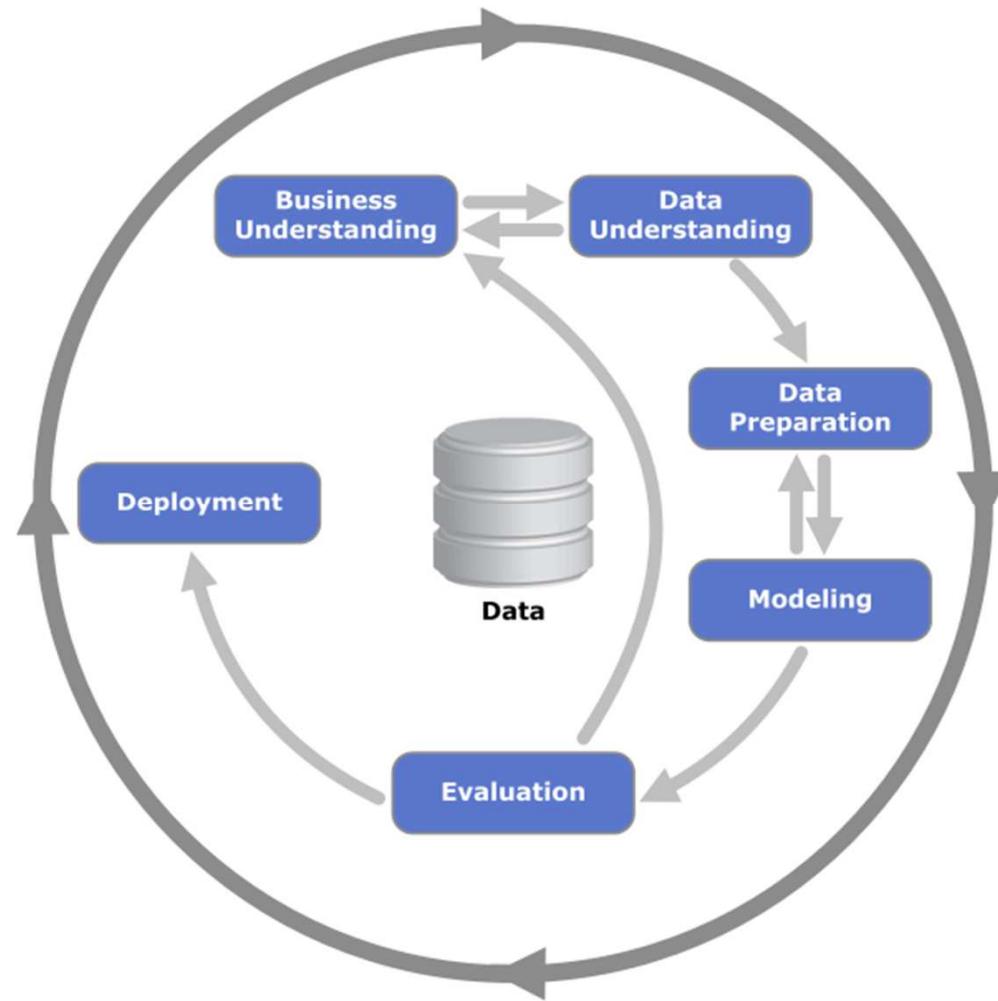
Medicina / diagnóstico

- 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,...)

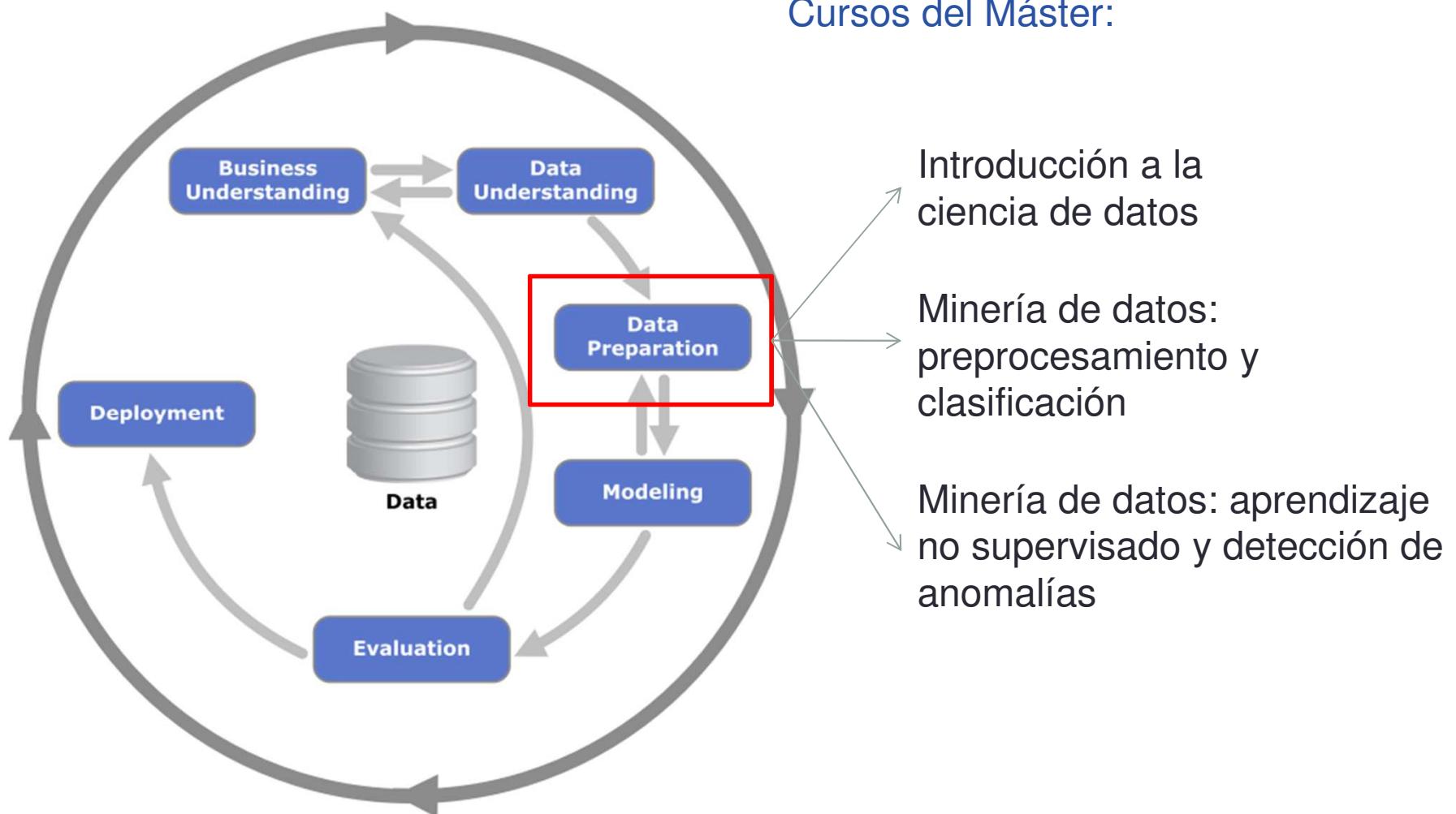
Minería de Datos. Fases



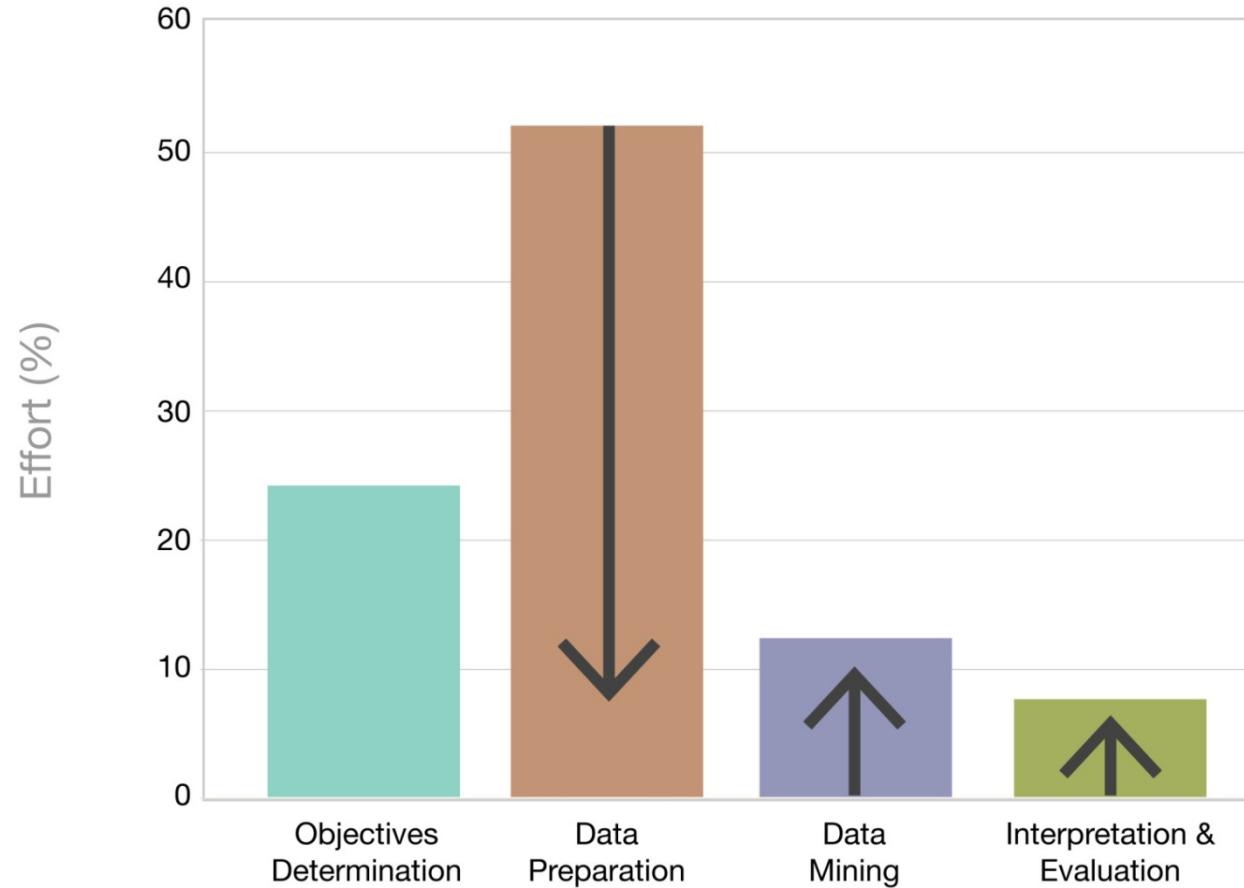
Minería de Datos. Fases



Minería de Datos. Fases



Minería de Datos. Fases

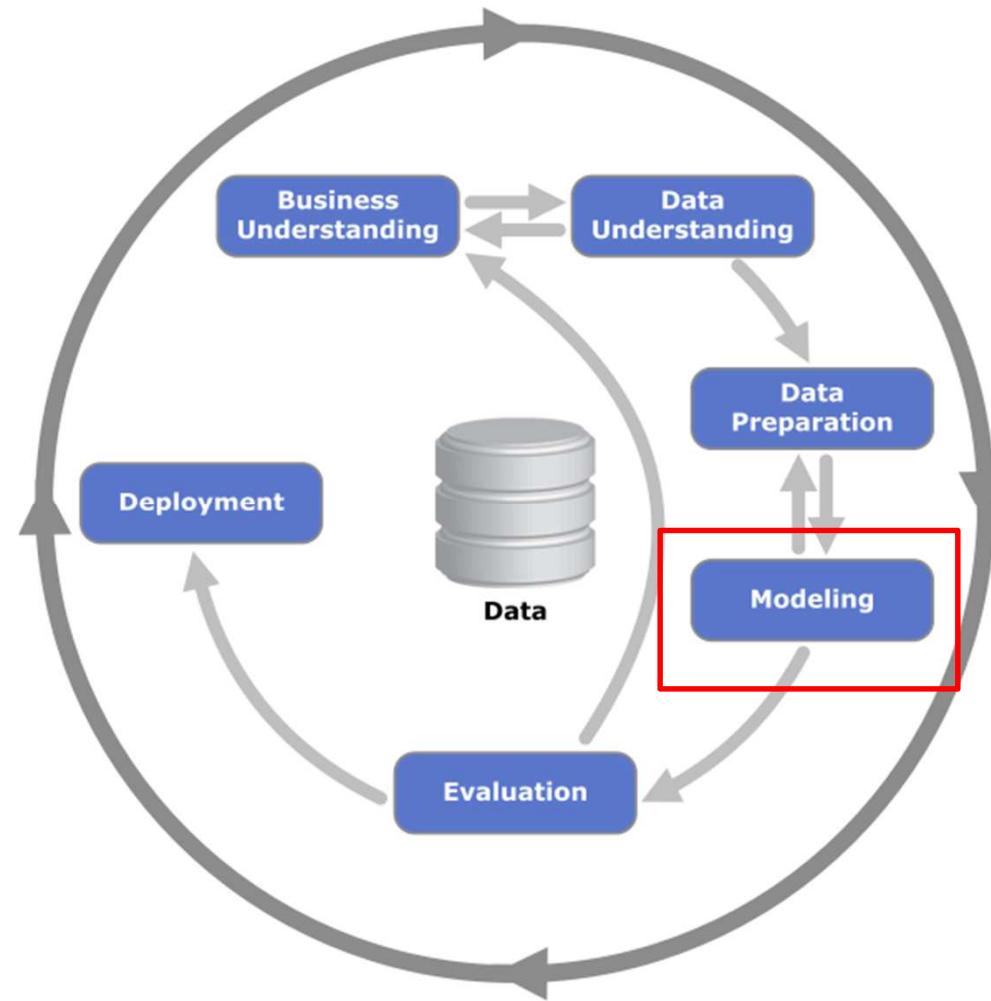


Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos

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.

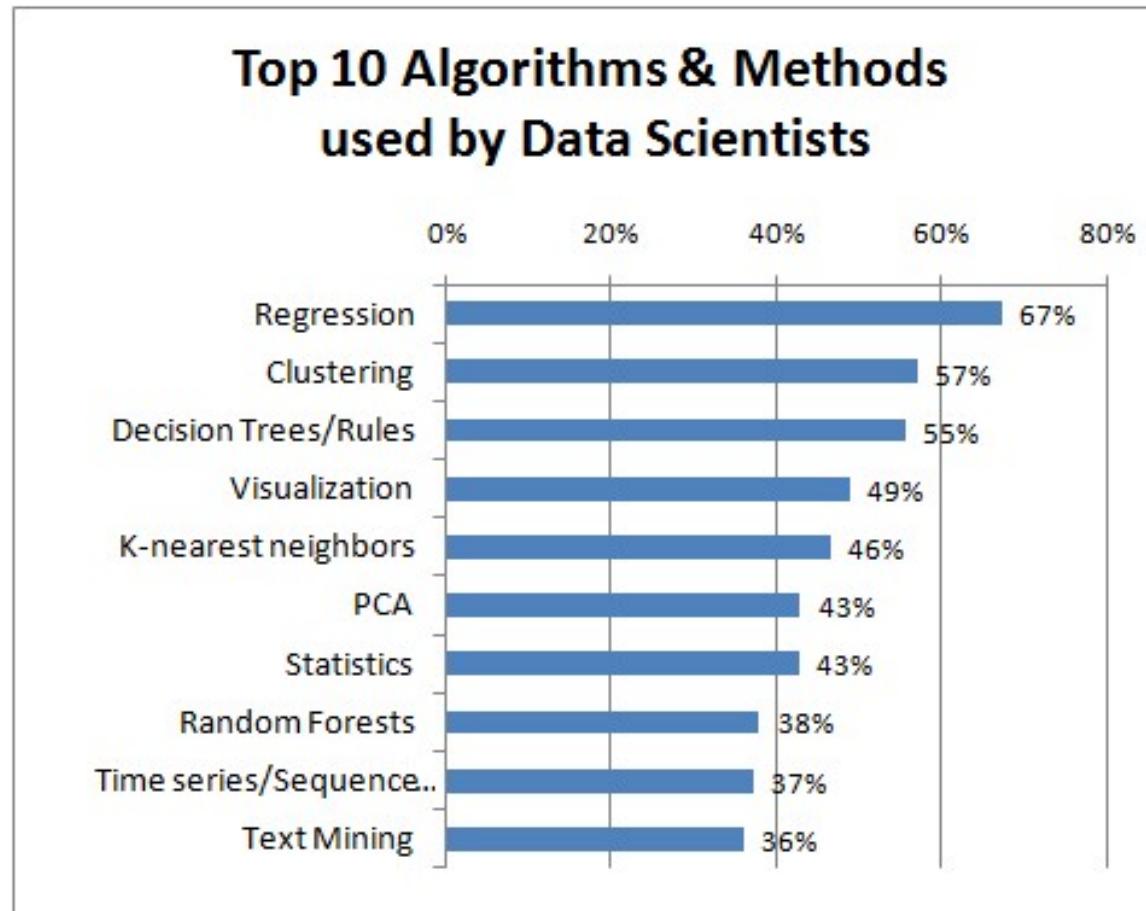
Modelos predictivos

- 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

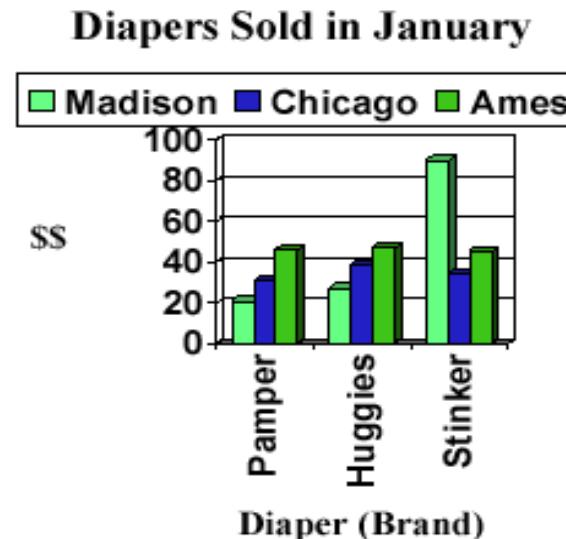


<http://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html>

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	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

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:

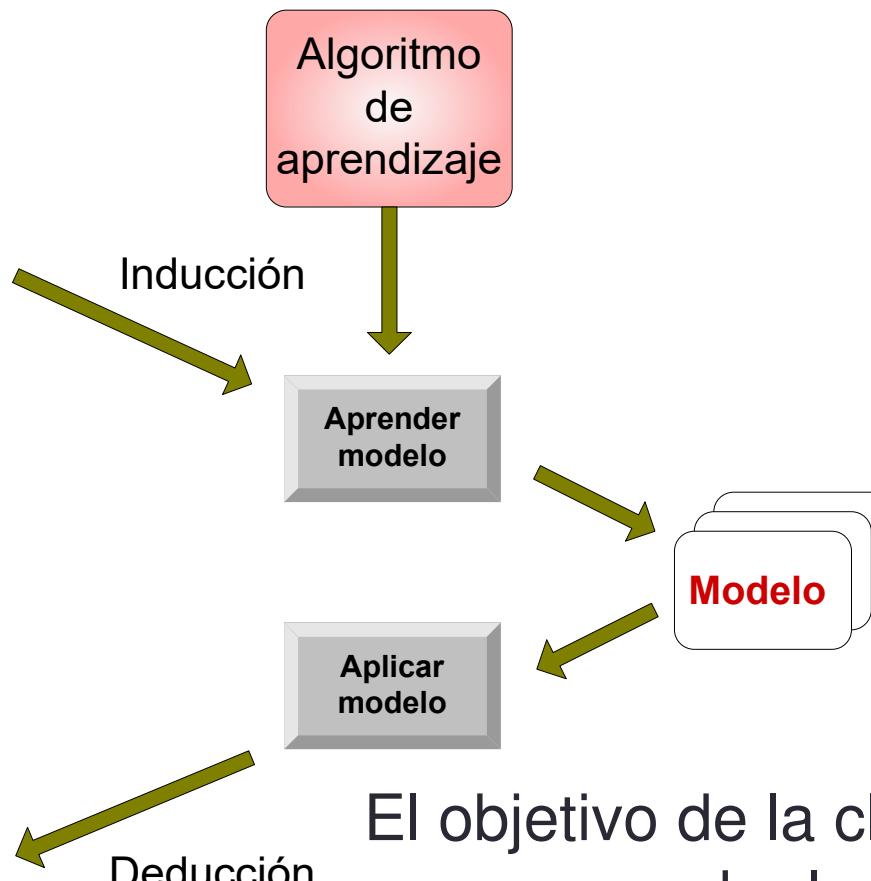
- Planificar disposiciones alternativas en el almacén
- Limitar descuentos especiales a sólo uno de los dos productos que tienden a comprarse juntos
- Poner los aperitivos que más margen dejan entre los pañales y las cervezas
- Poner productos de bebé en oferta cerca de las cervezas
- Ofrecer cupones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

Minería de Datos. Modelos → 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

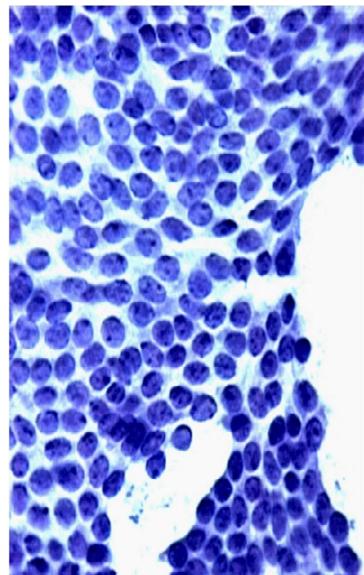
Cell phone
company:
Predict
Customer
Churn



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: $\text{perimeter}^2/\text{area} - 1.0$</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

Handwritting
recognition.
Assign a digit
from 0 to 9.



0 0 0 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 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 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 4 4
5 5 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 6 6
7 7 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 8 8
9 9 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 / Deep Learning
- 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

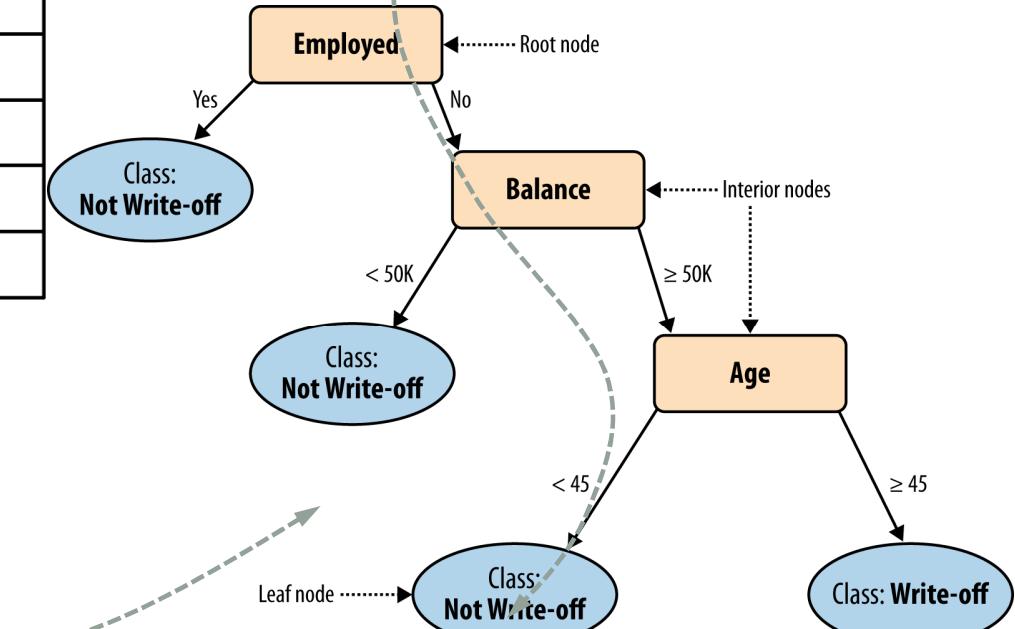
2. The model (decision tree) is constructed

→ Clasificación

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

(John \$53,000 30 no)

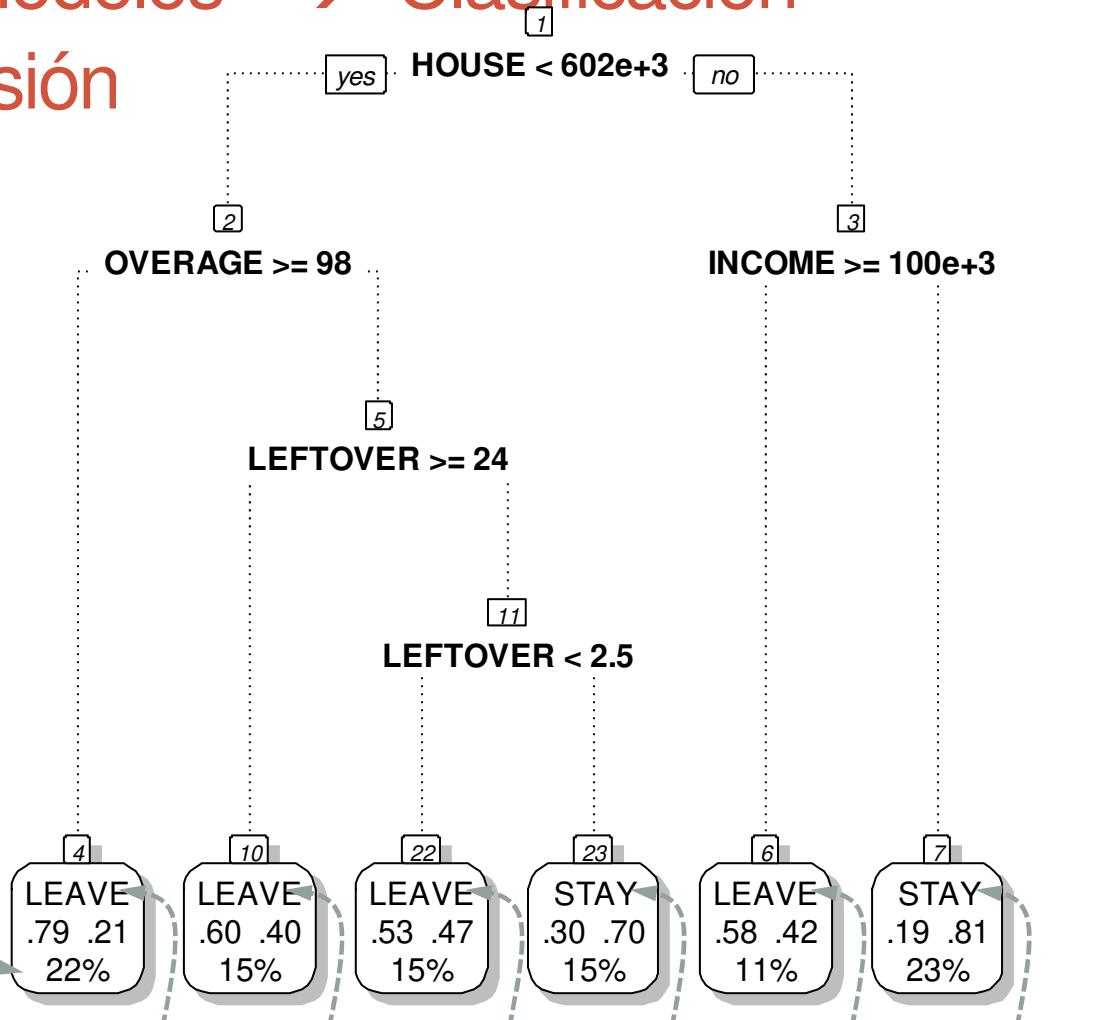
Classification:
Not Write-off



Minería de Datos. Modelos → Clasificación → Árboles de decisión

Customer Churn
example

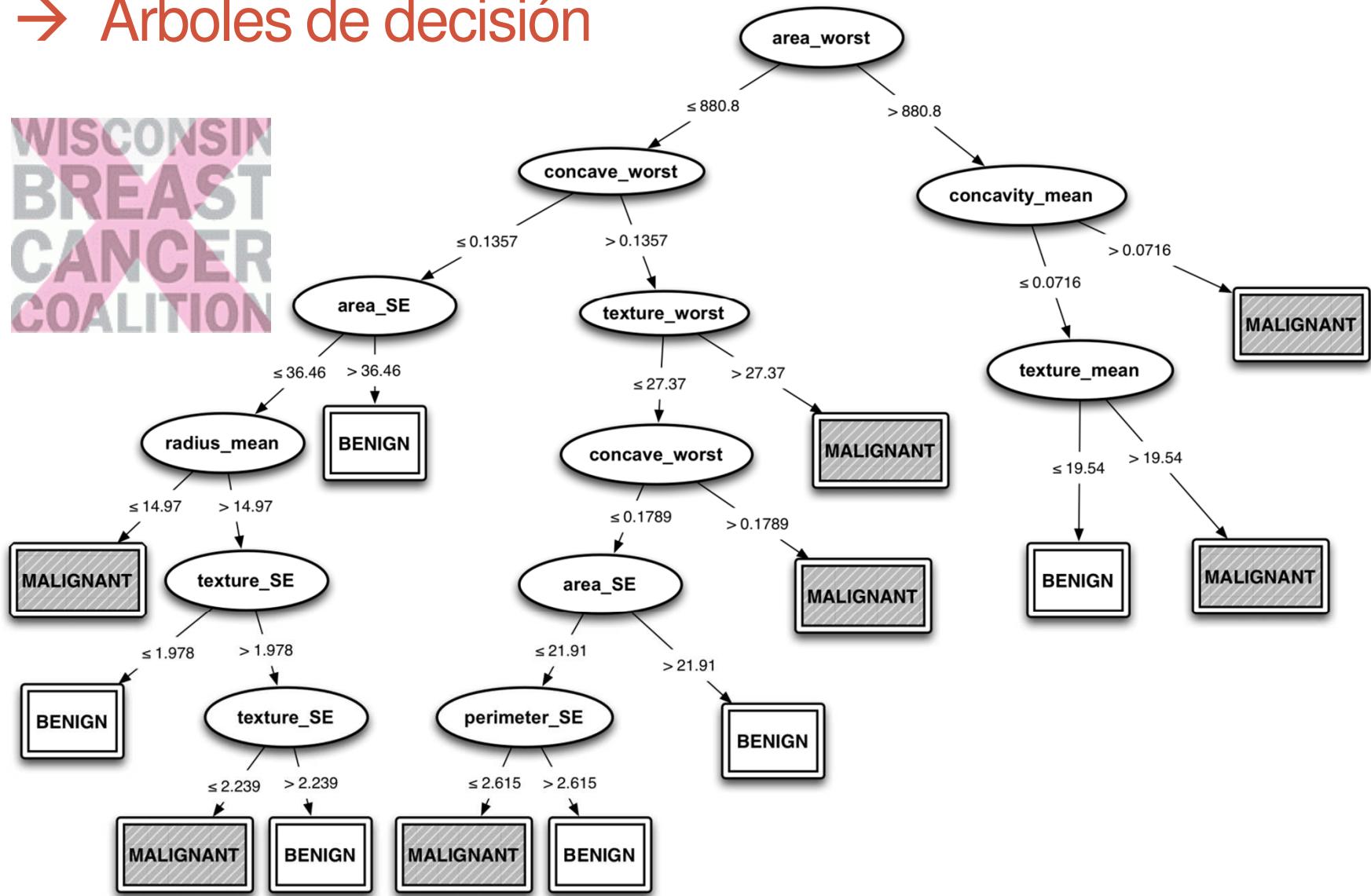
Total percentage of
covered examples



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



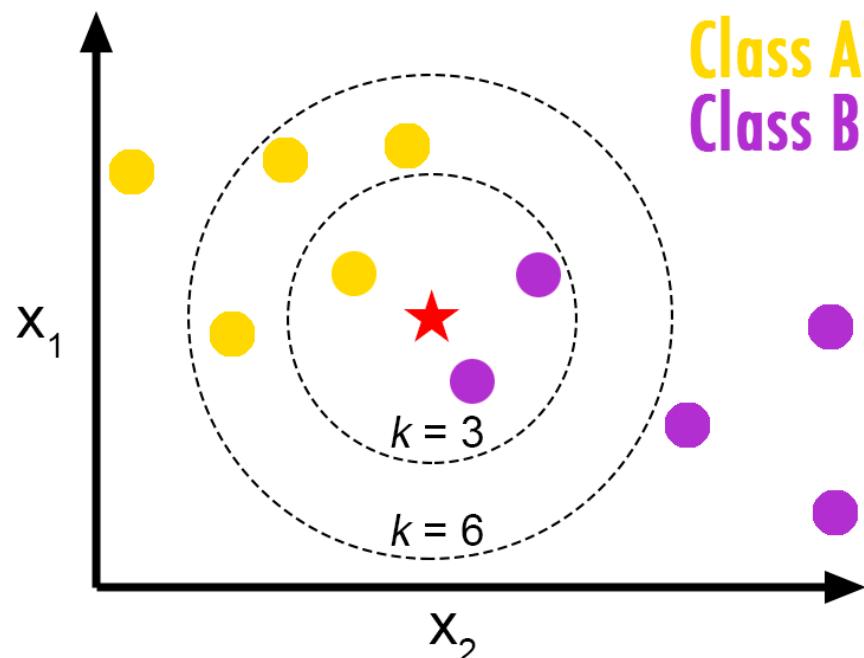
Minería de Datos. Modelos → Clasificación → KNN

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.

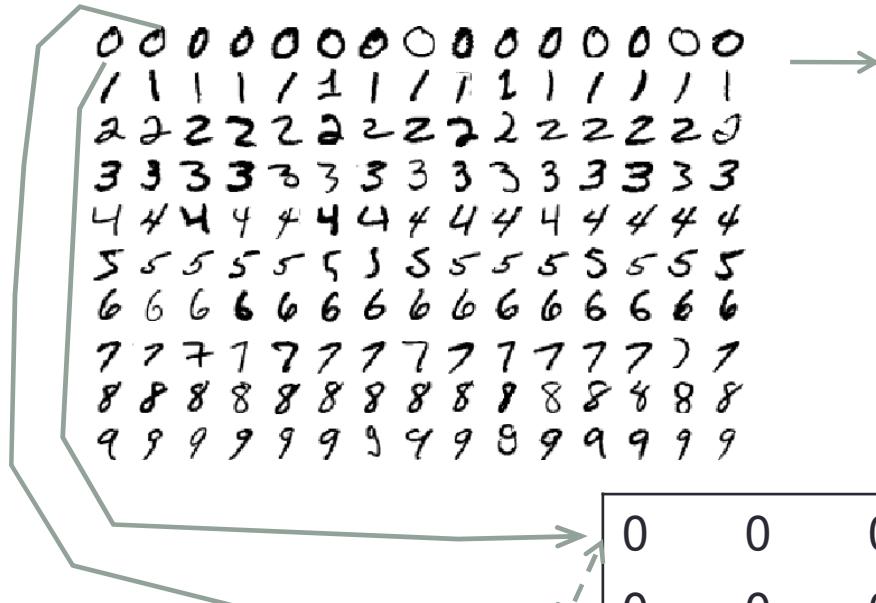
Cercanía → Medida de distancia.



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 matrix 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

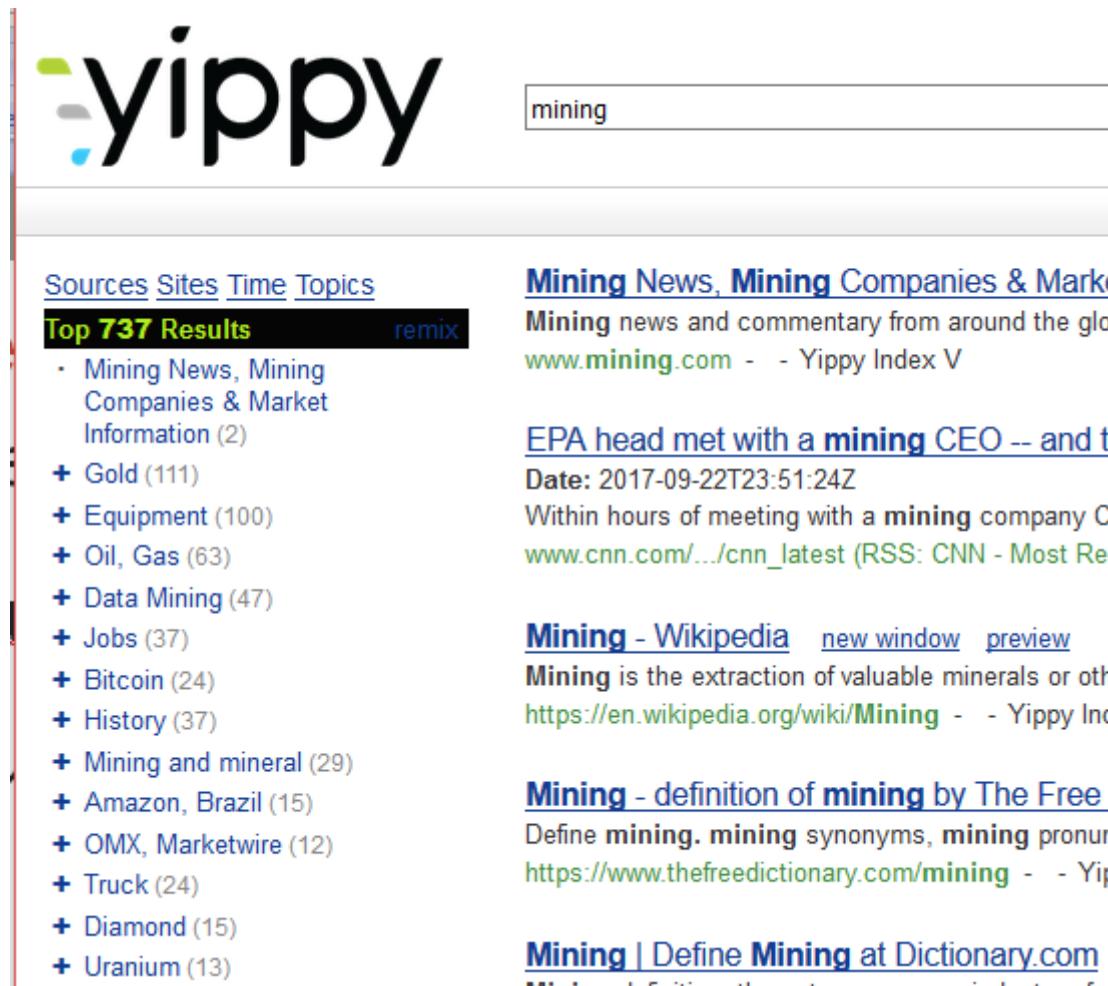
Segmentación de Clientes. Objetivo: Encontrar grupos (clusters) que identifiquen un modelo de cliente con características comunes (intereses, nivel de ingresos, hábitos de gasto, ...)



Minería de Datos. Modelos → Agrupamiento

Clustering de resultados en un motor de búsqueda

Clusty
→
Yippy



The screenshot shows the Yippy search interface. At the top, there's a search bar with the word "mining". Below it, a navigation bar offers options to switch between "Sources", "Sites", "Time", and "Topics". A prominent button labeled "Top 737 Results" is highlighted in green. To its right is a "remix" link. The main content area displays a list of 737 results, each with a title, a brief description, and a link. The results are categorized by topic, such as "Mining News, Mining Companies & Market Information", "Gold", "Equipment", "Oil, Gas", "Data Mining", "Jobs", "Bitcoin", "History", "Mining and mineral", "Amazon, Brazil", "OMX, Marketwire", "Truck", "Diamond", and "Uranium". Each entry includes a small number in parentheses indicating the count of results. The results are presented in a clean, modern style with blue links and a light gray background.

mining

Sources Sites Time Topics

Top 737 Results remix

- Mining News, Mining Companies & Market Information (2)
- + Gold (111)
- + Equipment (100)
- + Oil, Gas (63)
- + Data Mining (47)
- + Jobs (37)
- + Bitcoin (24)
- + History (37)
- + Mining and mineral (29)
- + Amazon, Brazil (15)
- + OMX, Marketwire (12)
- + Truck (24)
- + Diamond (15)
- + Uranium (13)

[Mining News, Mining Companies & Market Information](#) Mining news and commentary from around the globe. www.mining.com - - Yippy Index V

[EPA head met with a mining CEO -- and they talked about water](#) Date: 2017-09-22T23:51:24Z Within hours of meeting with a mining company CEO, EPA head Scott Pruitt was in a meeting with another mining company CEO. www.cnn.com/.../cnn_latest (RSS: CNN - Most Recent)

[Mining - Wikipedia](#) new window preview Mining is the extraction of valuable minerals or other geological materials from the earth, including the mining of coal, oil, natural gas, gold, silver, diamonds, and many others. <https://en.wikipedia.org/wiki/Mining> - - Yippy Index V

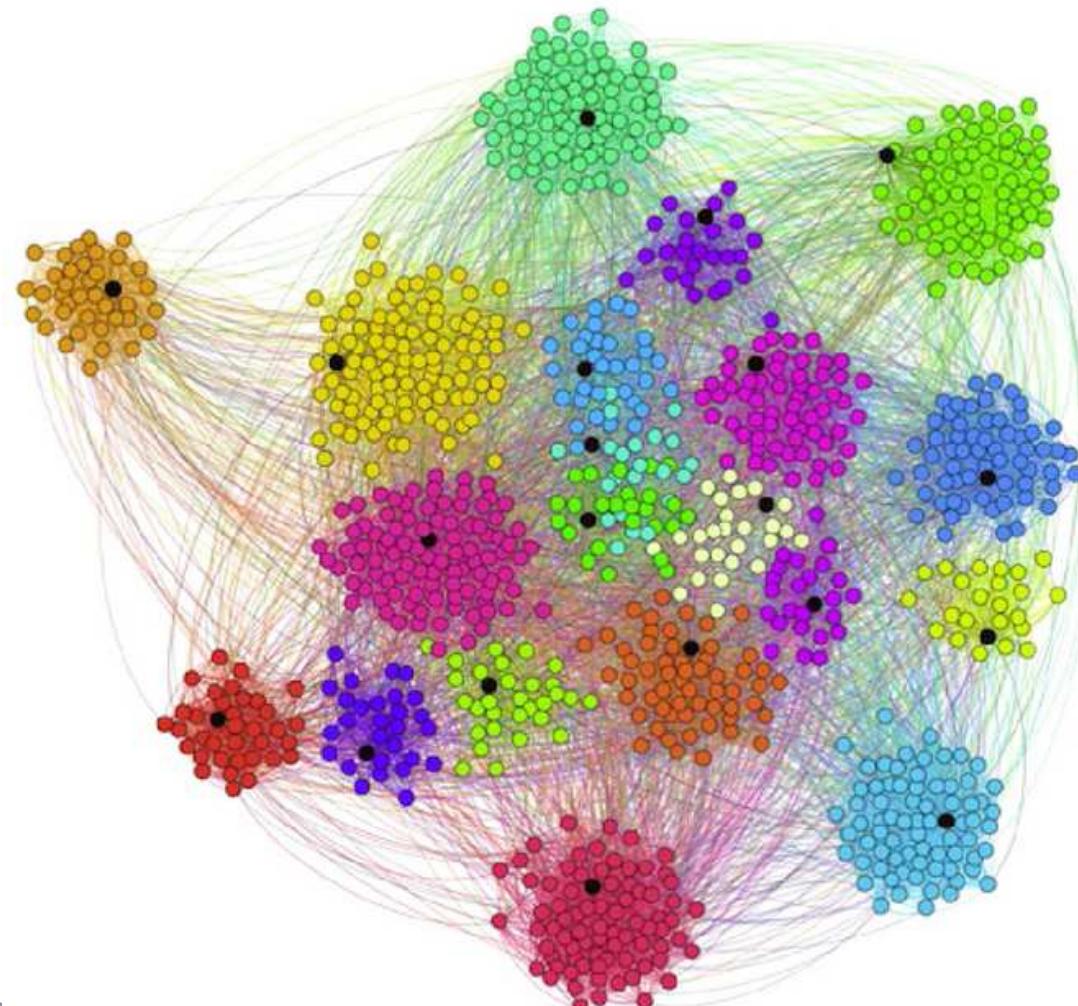
[Mining - definition of mining by The Free Dictionary](#) Define mining. mining synonyms, mining pronunciation, mining translation, English dictionary. <https://www.thefreedictionary.com/mining> - - Yippy Index V

[Mining | Define Mining at Dictionary.com](#) Mining definition, Mining meaning, what is Mining? <http://dictionary.reference.com/browse/mining> - - Yippy Index V

Minería de Datos. Modelos → Agrupamiento

Detección de *comunidades* en redes sociales

Gephi
(herramienta de
visualización)

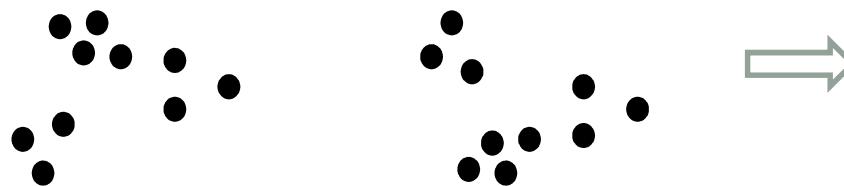


Minería de Datos. Modelos → Agrupamiento

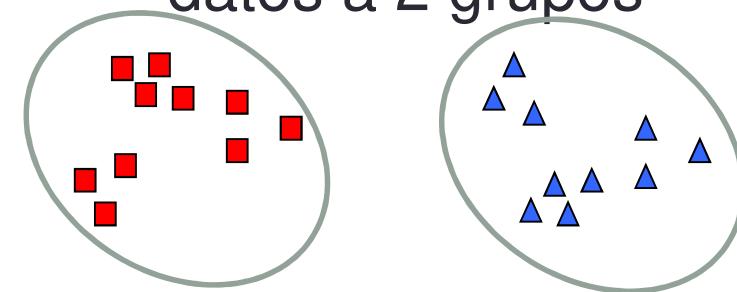
¿Cómo realizar el agrupamiento?

Supongamos sólo dos atributos:

Datos originales



Asignación de los
datos a 2 grupos



Minería de Datos. Modelos → Agrupamiento

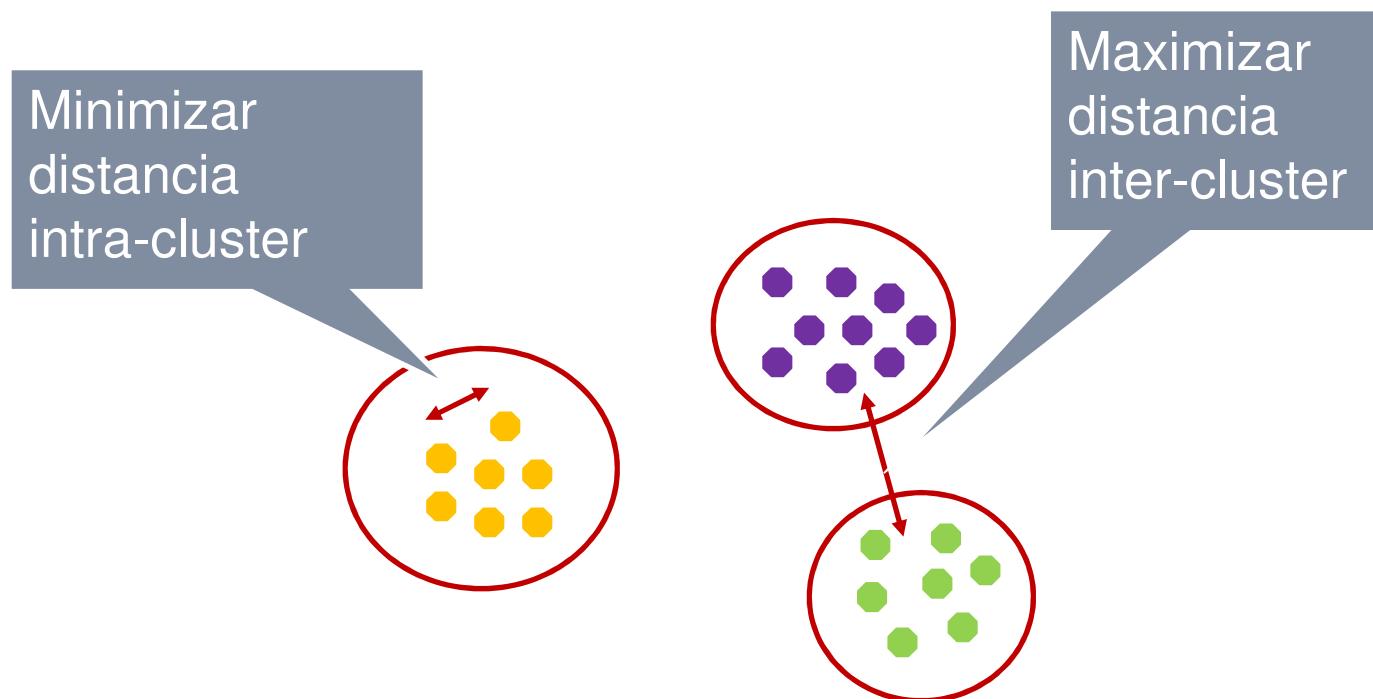
Más de dos atributos:

id	sexo	fechnac	educ	catlab	salario	salini	tiempem p	expprev	minoría
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
126	Hombre	21-enero-1951	15	Seguridad	\$24.300	\$15.000	90	191	Sí
127	Hombre	1-septiembre-1950	12	Seguridad	\$30.750	\$15.000	90	209	Sí
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 → 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



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

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

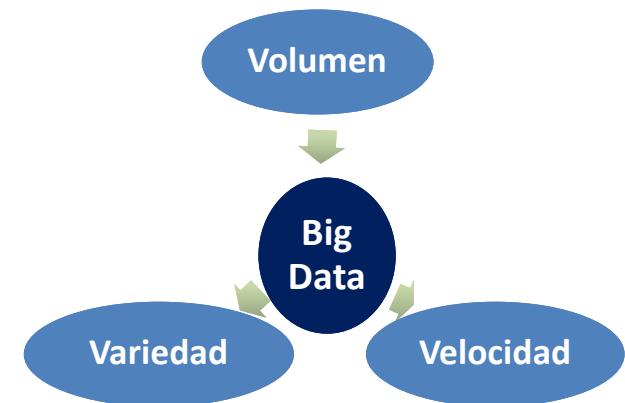
*UnSupervised Methods →
I have anomalies in my training
set but they are not labelled
I don't know if a record is an
anomaly or not)*

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	

Data Mining vs Big Data

Términos similares, aunque Big Data hace especial énfasis en buscar soluciones a los problemas planteados por:

- **Volumen** enorme de datos
- **Velocidad** con la que cambian
- **Variedad** de los tipos de datos
- Otras “uves”: **Veracidad**, **Volatilidad**



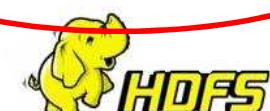
Big Data → NoSql Databases

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



Lightning-fast cluster computing



OpenDremel: Open source java implementation of Google BigQuery

Jug

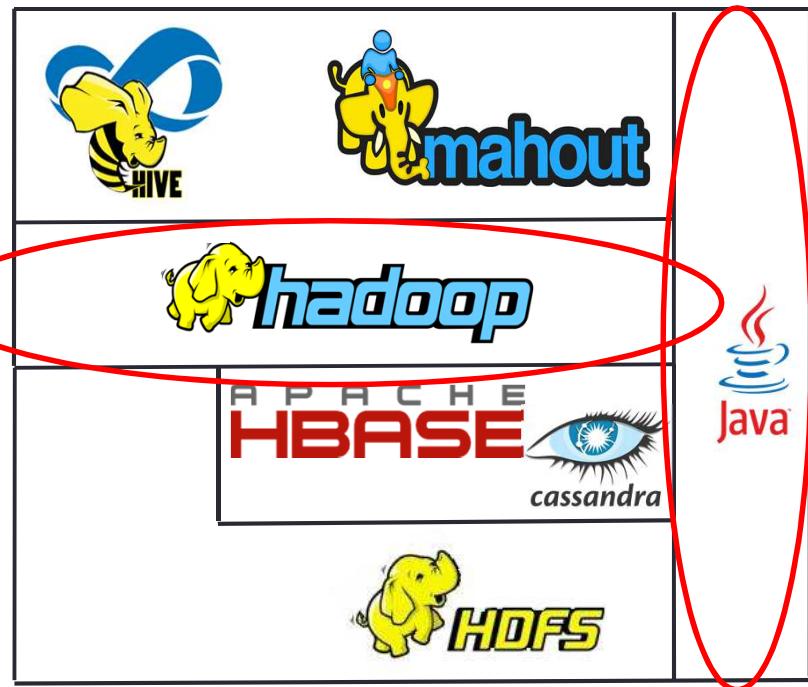
[Project Home](#)

[Downloads](#)

[Wiki](#)

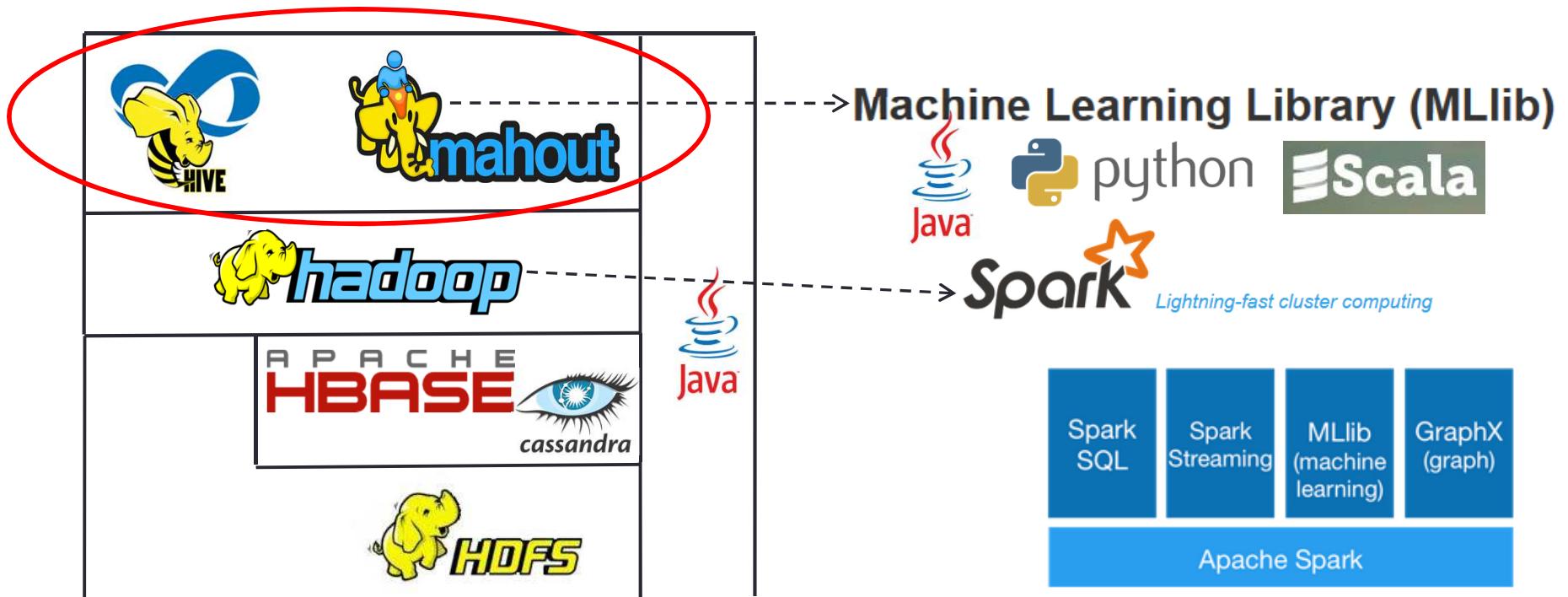
[Issues](#)

[Source](#)

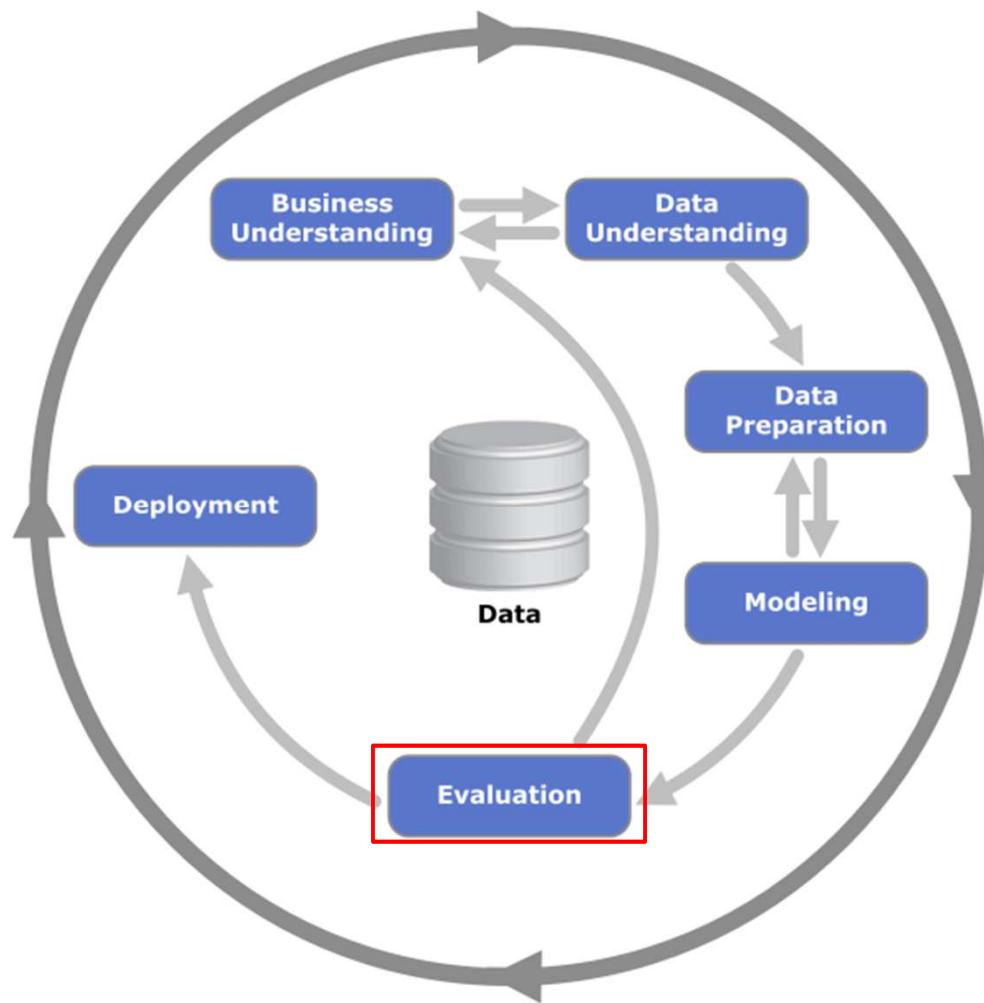


Big Data → Frameworks

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



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.

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 featuring the letters 'UCI' in yellow and a blue illustration of an antechinus (a small marsupial) pointing to the right. Below the logo, the text 'Machine Learning Repository' is displayed in yellow, with 'Center for Machine Learning and Intelligent Systems' in smaller text underneath. A navigation bar at the top includes a back button, a warning icon, and the URL 'https://archive.ics.uci.edu/ml/datasets.html'. The main content area has a dark blue header with the text 'Browse Through: 299 Data Sets' in white. Below this, there are two columns of filters. The left column, titled 'Default Task', lists 'Classification (214)', 'Regression (42)', 'Clustering (36)', and 'Other (50)'. The right column, titled 'Name', shows two examples: 'Abalone' with a thumbnail image of a shell and 'Adult' with a thumbnail image of a person's face.

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

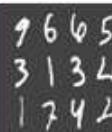
<http://www.inf.ed.ac.uk/teaching/courses/dme/2014/datasets.html>

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

Testing → Clasificación

Kaggle: Go from Big Data to Big Analytics

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

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

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

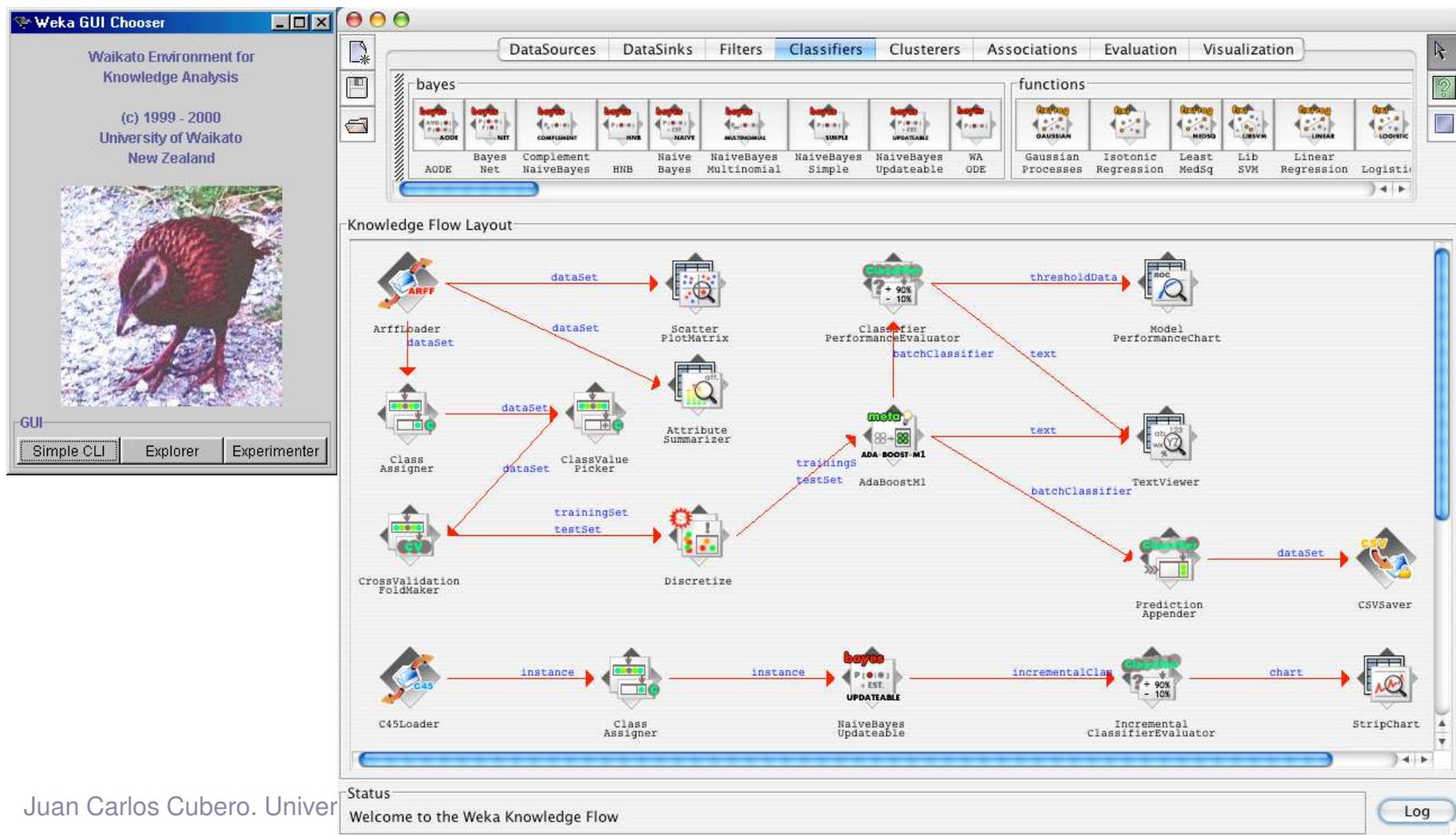


BLOG BIG DATA COURSE ADVICE STARTUPS USE CASES SPEAKER OPEN SOURCE PUBLIC DATA EVENTS FORUM ABOUT

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

Entornos de Desarrollo Data Mining

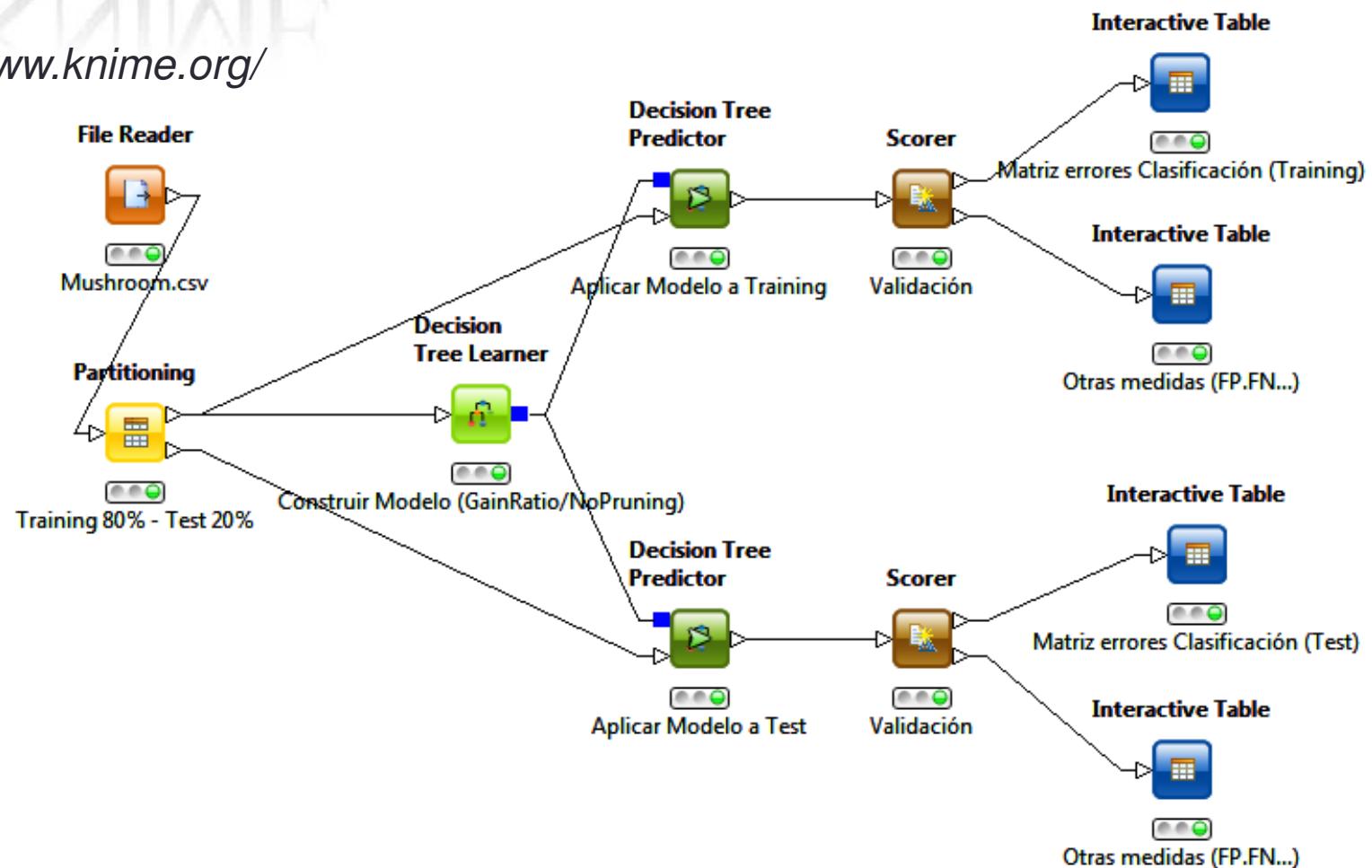
Weka <http://www.cs.waikato.ac.nz/ml/weka/>



Entornos de Desarrollo Data Mining

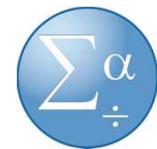
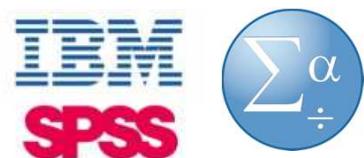


<https://www.knime.org/>



Entornos de Desarrollo Data Mining

Entornos similares propietarios



SAS® Enterprise
Miner™

Lenguajes Data Mining



[The Comprehensive R Archive Network](http://cran.r-project.org/)

cran.r-project.org/

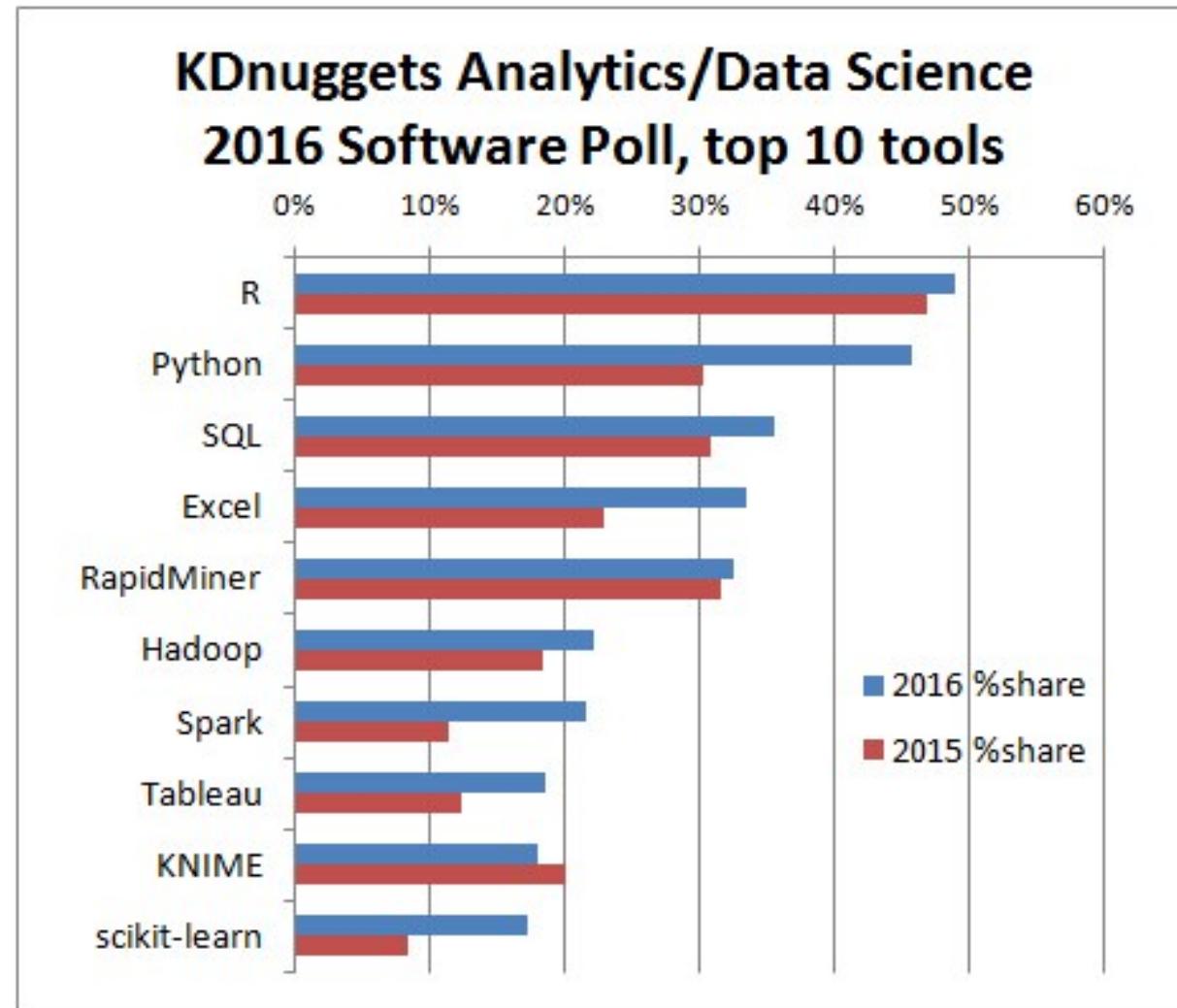
```
mydata      = read.arff(paste0(DIRECTORIO_DATOS,"\\Otros\\churn.arff"))
class.name  = "LEAVE"
class.index = grep(class.name, colnames(mydata))
myformula   = formula(paste(class.name,"~ ."))
myclass     = mydata[,class.index]
```

```
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 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.



Epílogo

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.

El equipo ganador presentó la solución 20 minutos antes que otro igual de bueno

Business Understanding:

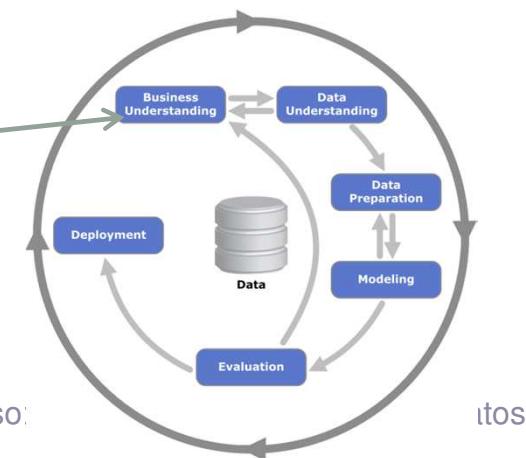
Fue crucial darse cuenta de cómo influía el factor tiempo en las calificaciones

Juan Carlos Cubero. Universidad de Granada.

The screenshot shows the Netflix Prize Leaderboard page. At the top, a large red stamp says "COMPLETED". Below it, the title "Leaderboard" is displayed in blue. A sub-instruction "Showing Test Score. [Click here to show quiz score](#)" is present. A dropdown menu indicates "Display top 20 leaders". The main table lists the top 8 teams with their names, best test scores (RMSE), percentage improvement, and best submit time. The winning team, "BellKor's Pragmatic Chaos", is highlighted in the first row.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries!	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries!	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43



Máster en Ciencia de Datos. Curso:

itos

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 [elmundo.es](http://www.elmundo.es/papel/pantallas/2016/09/05/57cd40d5268e3e3f248b4633.html). The title is "Big Data: el 'asesino' de los guionistas". Below the title are social sharing icons for Twitter, Facebook, and Email, and a "COMPARTIDO 425" counter. There are also "4 COMENTARIOS" and a "PANTALLAS" link. A small advertisement for a Lexus RX 450h Híbrido is visible on the left. The main image shows four men in suits walking down a hallway, each looking at their smartphone. Below the image is a caption: "Después de un análisis de datos, Amazon lanzó Alpha House, pero no duró más de una temporada." At the bottom, there is a note: "→ 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>