

Análisis Inteligente de Datos

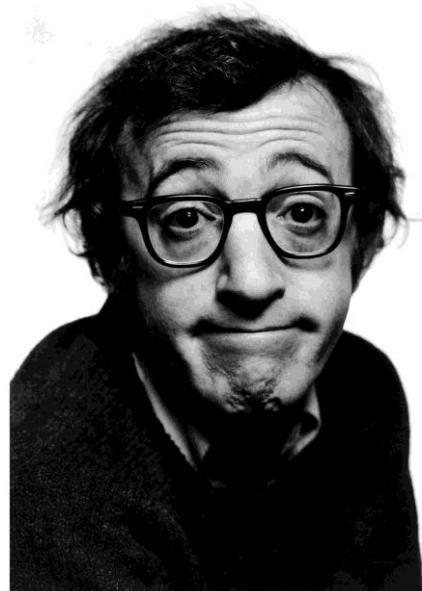
José A. Olivas



Universidad
Internacional
de Valencia

Desmontando a Harry, a Google y otros mitos de la era digital...

José A. Olivas



El aprendiz de Data Scientist...

José A. Olivas



Idea principal / objetivo

- Presentar aquellos retos que tienen que ver con la **gestión** y el **aprovechamiento inteligente** de la ingente cantidad de datos que se generan a partir de diversas fuentes y en diversos ámbitos de la **sociedad digital actual**, desde una perspectiva del **razonamiento aproximado**.
- Mostrar una visión panorámica de las **tecnologías** y **técnicas** emergentes para el **manejo masivo “inteligente”** de datos e información.
 - Las que provienen de la **estadística**.
 - Las que provienen de la **IA (Machine Learning)**.
- Introducir el concepto de **Soft-computing**.
- Mostrar la adecuación de las diferentes técnicas para los diferentes tipos de problemas a afrontar (Predicción, pronóstico...), siempre con una **posición crítica**.
- Describir algunos **ejemplos** de interés en el tema.

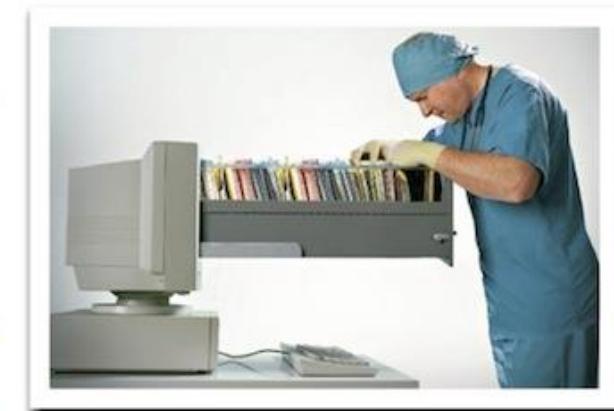
ii Aproximación Crítica !!

Algunas lecturas recomendables...

- Mayer-Schönberger, V.; Cukier, K.: Big data. La revolución de los datos masivos. Turner 2013.
- Piatetsky-Shapiro, G.; Frawley, W.: Knowledge Discovery in Databases. AAAI/MIT Press, Cambridge MA, 1991.
- Siegel E.: Analítica predictiva. Predecir el futuro utilizando Big Data. Anaya Multimedia-Anaya Interactiva, 2013.
- D. Agrawal, S. Das and A. E. Abbadi, “Big Data and Cloud Computing: Current State and Future Opportunities” ETDB 2011, Uppsala, Sweden.
- D. Agrawal, S. Das and A. E. Abbadi, “Big Data and Cloud Computing: New Wine or Just New Bottles?” VLDB 2010, Vol. 3, No. 2.

1. Origen: los datos

¿Qué son los datos?: DATOS / INFORMACIÓN /CONOCIMIENTO



Tipos de datos (I)



Tipos de datos (II)

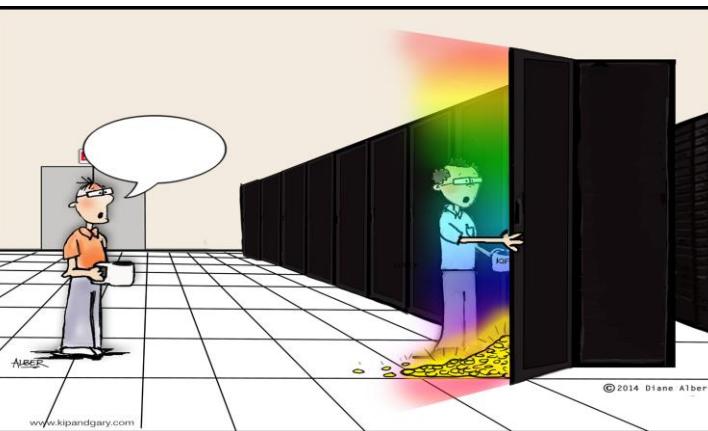
Document Database	Graph Databases
 	 
Wide Column Stores	Key-Value Databases
 	 
	      

¿ Dónde residen los datos ?

What is a data lake?

A repository for large quantities and varieties of data, both structured and unstructured.

Data generalists/
programmers can tap
the stream data for
real-time analytics.



The lake can serve as a staging area for the data warehouse, the location of more carefully “treated” data for reporting and analysis in batch mode.



2. Contexto: analítica, retos, límites

Business Intelligence (BI)

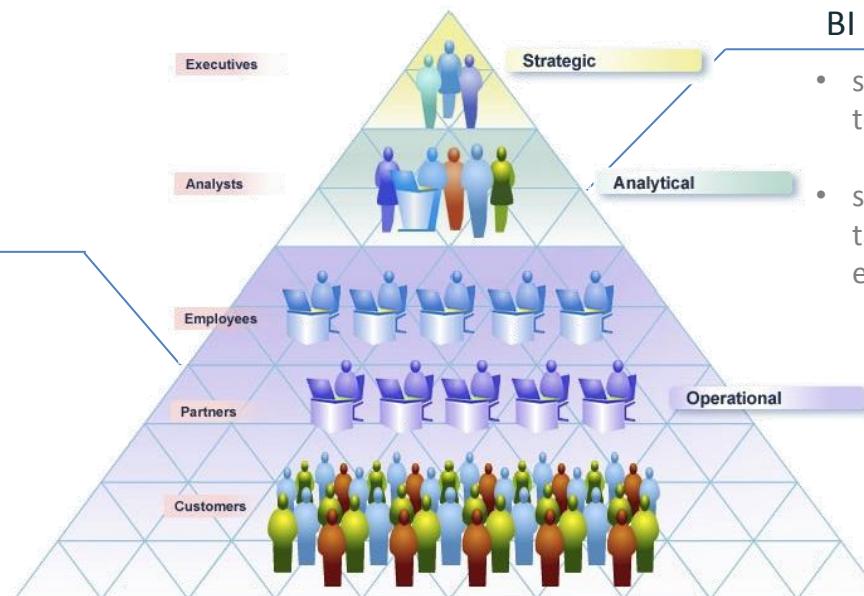
1 2 3

definición de business intelligence (BI)

La capacidad de transformar datos en información para ayudar a gestionar una empresa es el dominio de la **inteligencia empresarial de negocios (BI)**, que consiste en los procesos, aplicaciones y prácticas que apoyen la toma de decisiones ejecutivas

BI operacional

- soporta funciones al nivel operacional
- capacidad en tiempo real o cerca de real-time
- comprende y cubre los procesos.



BI analítico y estratégico

- soporta a los ejecutivos y en temas estratégicos
- soporta a los gestores en niveles tácticos que contribuyen a la estrategia

Crítica...

Demasiado restringido:

- ...transformar datos en información...
- ...apoyen la toma de decisiones...

¡¡ Hay muchas otras cosas que se pueden hacer !!

- Veamos las posibles salidas...

Outputs...

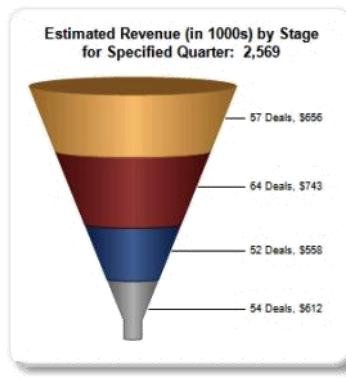


Top 10 Key Deals

Account	Est. Close Date	Est. Revenue (in 1000s)	Recent Activity
Wide World Importers	9/2/2008	\$725.00	0
Trey Research	8/4/2010	\$435.70	0
Fabrikam Inc.	2/21/2010	\$434.30	0
Tailspin Toys	8/25/2010	\$431.50	0
Consolidated Messenger	7/16/2010	\$395.00	0
City Power & Light	5/23/2009	\$339.69	0
Contoso Parts	3/23/2009	\$338.51	0
Fabrikam Inc.	4/10/2009	\$338.12	0
City Power & Light	7/9/2009	\$337.73	0
Madrona Solutions Group	8/9/2009	\$336.94	0

Top 10 Sales Leaders in 2009

Sales Representative	Actual Revenue (in 1000s)	Win Rate
Anton Kirillov	\$4,518.5821	47 %
System Administrator	\$3,640.8966	46 %
Simon Pearson	\$2,760.7613	43 %
Mark Hassall	\$1,867.9274	45 %
Brian Cox	\$1,798.2422	46 %
William Ngo	\$1,515.8373	47 %
Robert Lyon	\$1,454.1014	46 %
Lori Penor	\$1,372.6267	45 %
Steve Masters	\$1,359.9723	44 %
John Chen	\$799.0690	41 %



Crítica...

De nuevo demasiado restringido:

- Esto es sólo visualización
- Conocimiento...
 - Sistemas de Ayuda a la Decisión (DSS).
 - Sistemas Recomendadores (Recommender Systems).
 - Análisis de series temporales (Predicción vs Pronóstico).
- Segmentación.

¡¡ Patrones !!

- Las salidas condicionan todo el proceso.
- No se debe ir “a ciegas” hacia delante

Analítica de datos...



Tipos básicos de Analítica de datos...

DESCRIPTIVA

¿Qué está ocurriendo / Qué ha ocurrido?

DIAGNOSTICA

¿Porqué pasó / está ocurriendo lo que
está ocurriendo?



DESCUBRIMIENTO

PREDICTIVA

¿Qué es lo más probable que ocurre?

PRESCRIPTIVA

¿Qué es lo mejor que podemos hacer?

¡Predicción (estimación) vs Pronóstico (Hecho puntual)!

3. Analítica en entornos Big Data

Big Data

- Aproximación ingenua y crítica.
- Definición abierta de Big Data.

“**Big Data**” es en el sector de [tecnologías de la información y la comunicación](#) una referencia a los sistemas que manipulan grandes [conjuntos de datos](#). Las dificultades más habituales en estos casos se centran en la captura, el almacenamiento, búsqueda, compartición, análisis y visualización.

(Wikipedia)

“**Big data**” es un término aplicado a [conjuntos de datos](#) que superan la capacidad del [software habitual](#) para ser [capturados, gestionados y procesados](#) en un [tiempo razonable](#). Los tamaños del “big data” se hallan constantemente en [aumento](#).

(Wikipedia)

Big Data

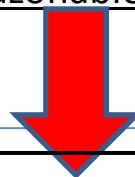
- Aproximación ingenua y crítica.
- Definición abierta de Big Data.

“**Big Data**” es en el sector de [tecnologías de la información y la comunicación](#) una referencia a los sistemas que manipulan grandes [conjuntos de datos](#). Las dificultades más habituales en estos casos se centran en la captura, el almacenamiento, búsqueda, compartición, análisis y visualización.

(Wikipedia)

"**Big data**" es un término aplicado a [conjuntos de datos](#) que superan la capacidad del [software habitual](#) para ser [capturados, gestionados y procesados](#) en un [tiempo razonable](#). Los tamaños del "big data" se hallan constantemente en [aumento](#).

(Wikipedia)



¡El nudo Gordiano de Google!

Definición

- Datos...
- Información...
- ¿Conocimiento?
 - Abstracción-Patrones...
 - Dimensión humana...
 - La Web...
 - Google...

¡¡Explosión en la cantidad de datos!!

- A380:
 - Más de 1 billón de líneas de código.
 - Cada motor genera 10 Tb cada media hora.
 - Mas de 640 Tb de información por vuelo.
- Twitter genera más de 15 Tb de datos al día.
- Las principales bolsas generan más de 1 Tb al día.
- La capacidad de almacenamiento de ha doblado cada 3 años desde los 80s.

¡¡Explosión en la cantidad de datos!!

- Historias Clínicas Electrónicas:
9.000.000.000 documentos sólo en España...

¡¡Problemas graves al gestionarlos!!

- A380 de Quantas (32-2009) ¡SATURACIÓN!
- A330 de Air France (447-2010) ¡INCONSISTENCIA!
- B777 Malayo (370-2014) ¡INCERTIDUMBRE!
- Twitter, ¡ANÁLISIS DE SENTIMIENTOS! PLN.
- No se usan las Historias Clínicas Electrónicas.

¡¡Explosión en la cantidad de datos!!

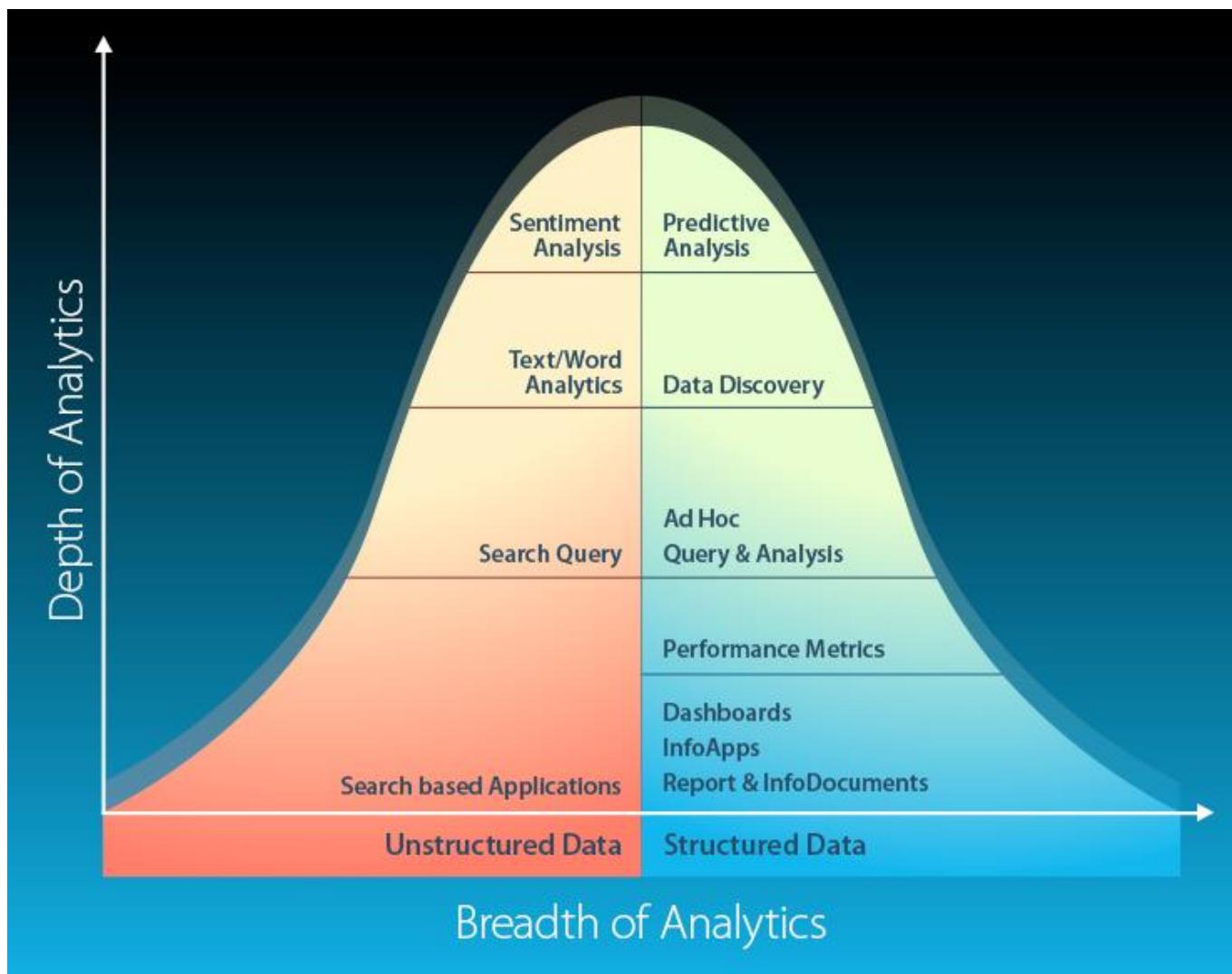
¿Habitualmente qué hacemos con todos estos datos?

¡¡Explosión en la cantidad de datos!!

¿Habitualmente qué hacemos con todos estos datos?

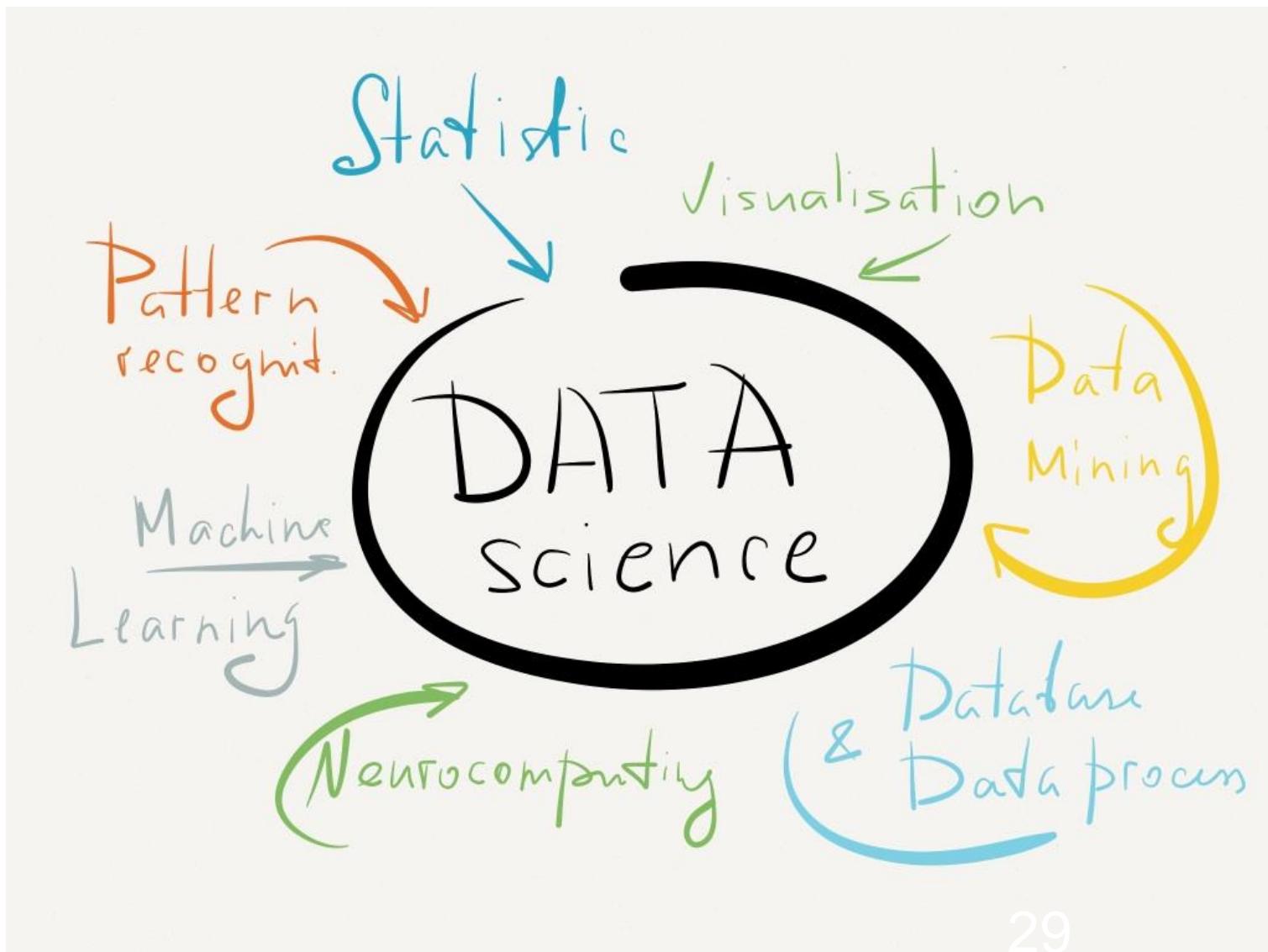
¡IGNORARLOS!

Evolución del análisis de datos:

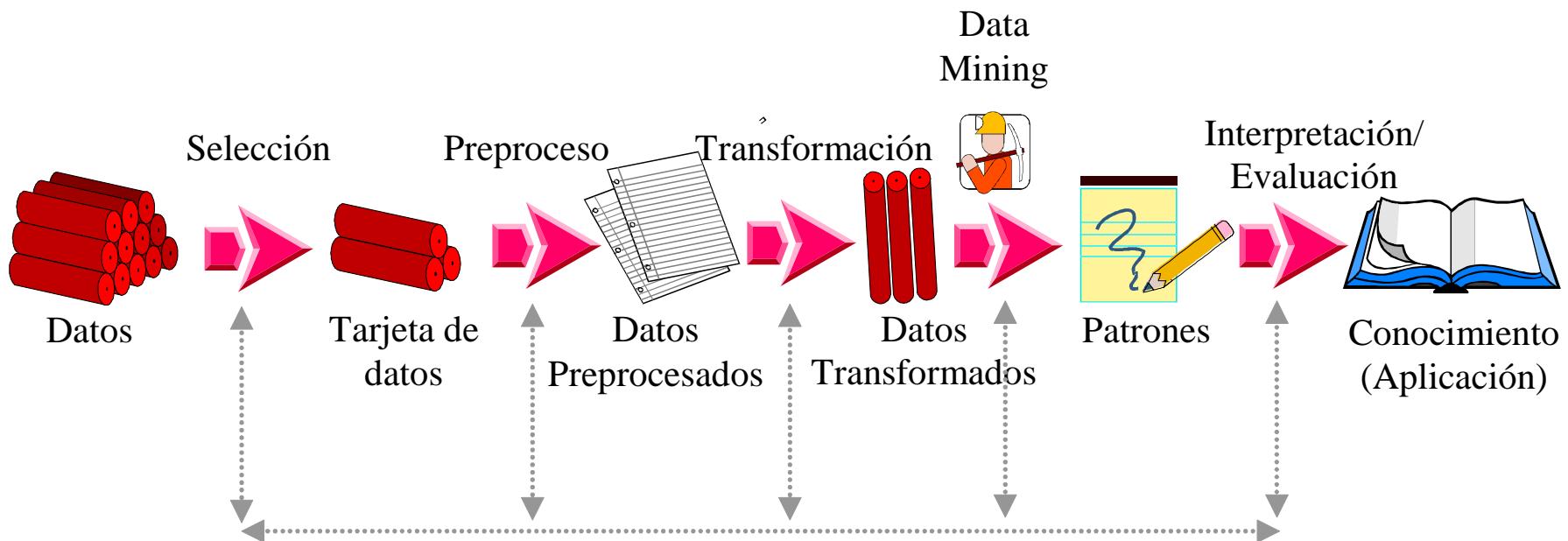


4. Ciencia de datos

Data Science

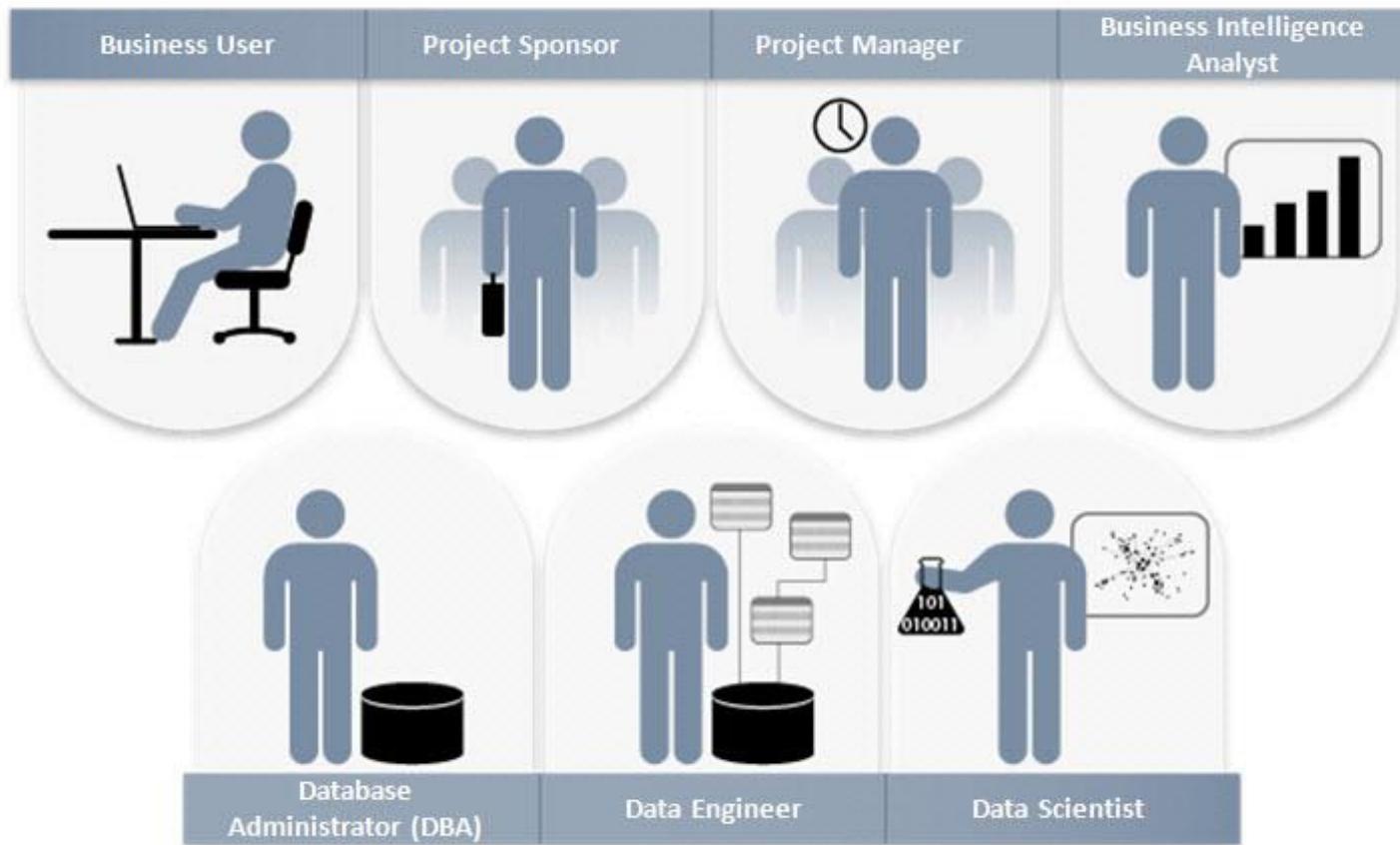


La importancia del KDD...



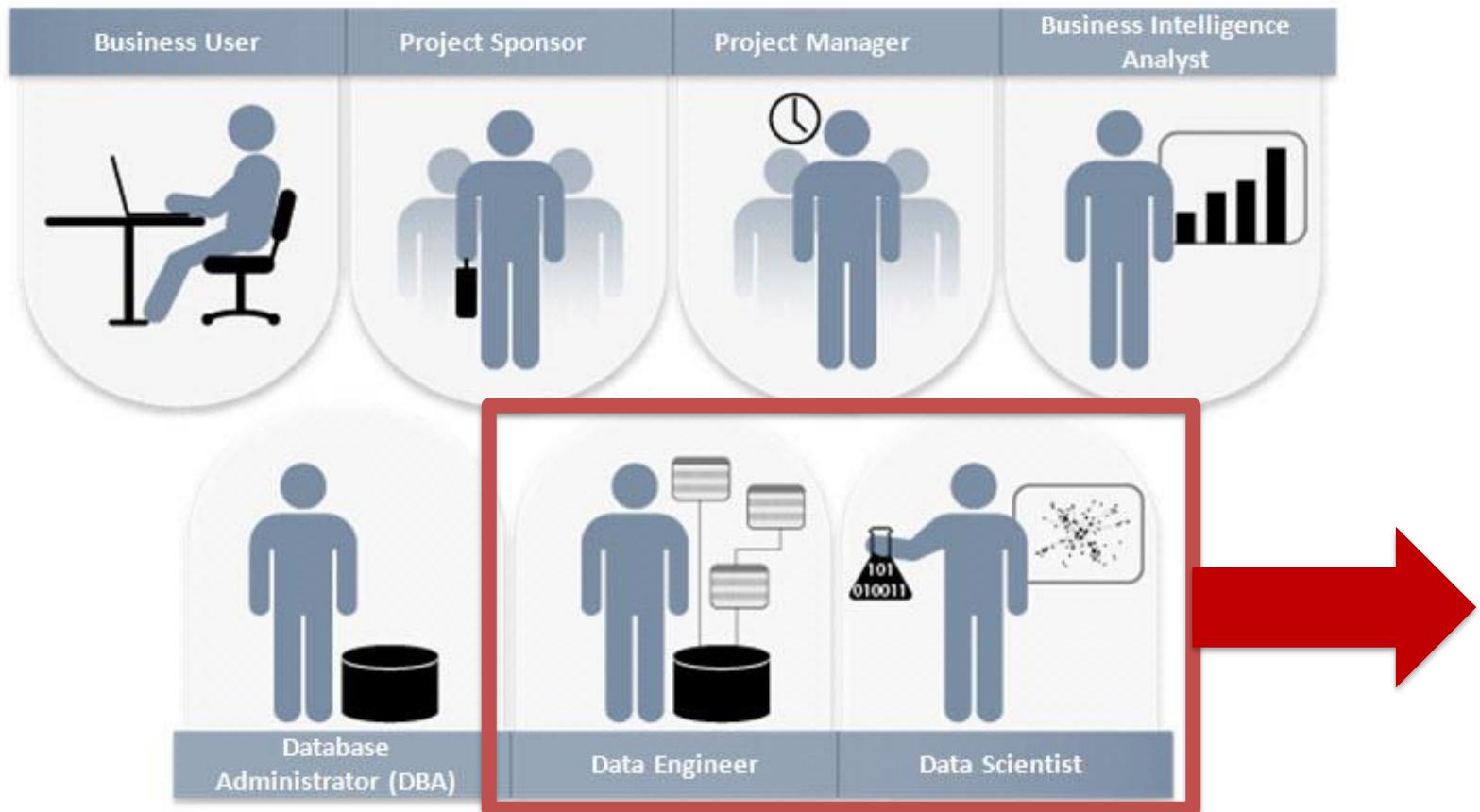
Equipo de Trabajo

Key Roles for a Successful Analytic Project

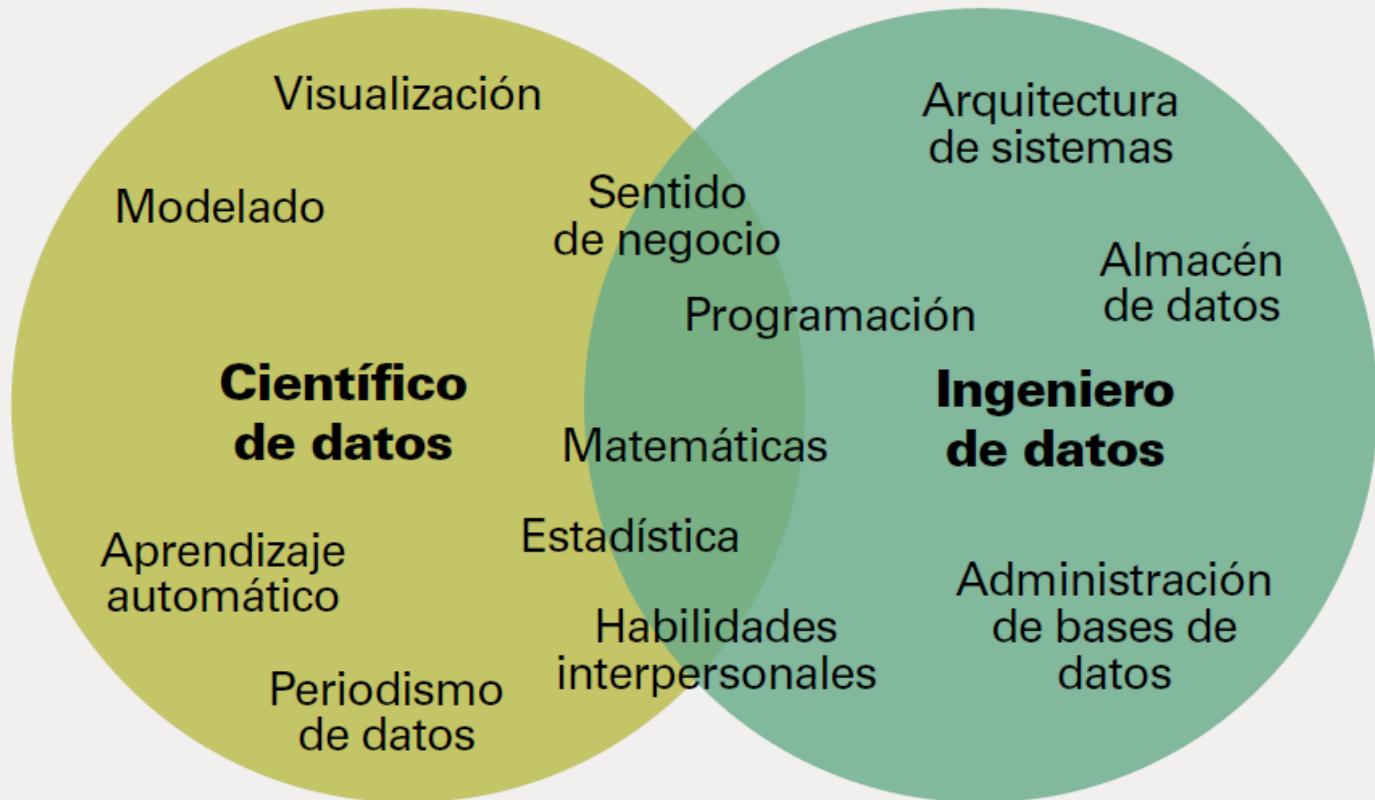


Equipo de Trabajo

Key Roles for a Successful Analytic Project



Equipo de Trabajo



Fuente: Universitat Oberta de Catalunya. Máster en *Business Intelligence* y Big Data (2016)

Equipo de Trabajo



4. Algunas otras aplicaciones/ejemplos

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:
 - “asíncrono” vs. “síncrono”
 - Reflexión/preparación vs. Inmediatez/visceralidad
- ¡Dimensión Humana!
- PLN
- Análisis de Sentimientos.

“Sentiment analysis: A review and comparative analysis of web services” (Information Sciences, 2015)

Acceso y la búsqueda de información digital.

- Los buscadores son eficientes, pero no eficaces.
- Ejemplos: sinonimia, veracidad (reputación), variedades diatópicas, operadores, tendencias...

Acceso y la búsqueda de información digital.

- Los buscadores son eficientes, pero no eficaces.
- Ejemplos: **sinonimia**, veracidad (reputación), variedades diatópicas, operadores, tendencias...

Acceso y la búsqueda de información digital.

- Los buscadores son eficientes, pero no eficaces.
- Ejemplos: **sinonimia, veracidad (reputación), variedades diatópicas, operadores, tendencias...**

Acceso y la búsqueda de información digital.

- Los buscadores son eficientes, pero no eficaces.
- Ejemplos: **sinonimia, veracidad (reputación), variedades diatópicas**, operadores, tendencias...

Acceso y la búsqueda de información digital.

- Los buscadores son eficientes, pero no eficaces.
- Ejemplos: **sinonimia, veracidad (reputación), variedades diatópicas, operadores, tendencias...**

Acceso y la búsqueda de información digital.

- Los buscadores son eficientes, pero no eficaces.
- Ejemplos: **sinonimia, veracidad (reputación), variedades diatópicas, operadores, tendencias...**

“Búsqueda eficaz de información en la Web” (EDULP, 2011)

5. Métodos basados en la estadística

Técnicas de Regresión y correlación

- Formalización de una relación significativa entre dos o más variables para calcular pronósticos a partir del conocimiento de los valores en un individuo concreto.
 - **Lineal** (aproximación de la dependencia entre una variable dependiente y variables independientes)
 - **Múltiple** (para predecir el valor de una variable dependiente a partir de variables independientes)
 - **Logística** (para predecir variables categóricas)
 - **CART** (Classification And Regression Trees, Leo Breiman)
 - Etc.

Otras Técnicas

- Técnicas de **extrapolación** de funciones.
- Técnicas de **aproximación y ajuste** de funciones.
- Técnicas de **agrupamiento** basadas en medidas estadísticas (**clustering**).
- Etc.

Muchas se pueden englobar tanto en Técnicas estadísticas como de *Machine Learning*...

6. Métodos basados en Inteligencia Artificial (*Machine Learning*)

Machine Learning (Aprendizaje automático)

- Rama de la **Inteligencia Artificial** en la que se diseñan mecanismos para dotar a los sistemas computacionales de capacidad de aprendizaje.
- Aprendizaje en el sentido de la capacidad de **descubrir regularidades (patrones)** en datos o situaciones anteriores y aplicarlos a nuevos problemas o situaciones análogas.

Principales paradigmas en Machine Learning

- **Paradigma Analógico** (Aprendizaje por analogía).
 - Pretende encontrar una solución a un problema que se presenta ahora **usando el mismo procedimiento** usado en la resolución de uno similar que se presentó en otra ocasión anterior.
 - Si dos problemas son similares en algún aspecto de su formulación entonces pueden serlo también en sus soluciones. Nuevos problemas pueden ser abordados reduciéndolos a problemas análogos resueltos.

Principales paradigmas en Machine Learning

- **Paradigma Analógico (Ejemplos).**
 - Analogía por transformación.
 - Analogía por derivación.
 - Razonamiento basado en casos.
 - Etc.

Principales paradigmas en Machine Learning

- **Paradigma Inductivo.**
 - Árboles de decisión, algoritmos de inducción pura...
- **Paradigma Conexionista.**
 - Redes Neuronales Artificiales...
- **Paradigma Evolutivo.**
 - Algoritmos Genéticos, otros métodos de optimización, colonias de insectos, descenso estocástico del gradiente...
- **Modelos gráficos probabilistas.**
 - Bayesianos, cadenas de Markov, Filtros de Kalman, redes de creencia, Máquinas de Soporte Vectorial (SVM), Metaheurísticas...

Técnicas de *Clustering* (Aprendizaje no supervisado)

- **Agrupar los elementos** de una colección en subconjuntos (clases, categorías, *clusters*), nítidos o borrosos, **en base a su similitud**. Es **no supervisado** porque las clases o categorías no se conocen a priori, las determinaran las propias similitudes entre los elementos.
- Por lo tanto, se centran en la “**medida de similitud**” entre elementos, de la que puede haber infinidad de variantes: **estadísticas**, distancias **euclídeas**, distancias **vectoriales** (coseno), distancias **borradas**, etc...

Técnicas de *Clustering*: EJEMPLOS

- **Paradigma Conexionista.**
 - Redes Neuronales Artificiales: **SOM** (**S**elf **O**rganized **M**aps, Mapas de Kohonen). Toolbox de Matlab SOM.
 - Etc.

Técnicas de *Clustering*: EJEMPLOS

- **Modelos estadísticos y probabilistas.**
 - K-means, c-means,
 - K-nearest neighbours (KNN),
 - Mean shift (ventanas circulares con un centroide),
 - Dirichlet process (estocásticos basados en distribuciones de probabilidad). LDA (Latent Dirichlet Allocation),
 - Modelos Gaussianos,
 - Etc.

Técnicas de *Clustering*: EJEMPLOS

- **Extensiones basadas en Lógica Borrosa.**
 - Fuzzy K-means,
 - Fuzzy c-means,
 - Isodata,
 - Etc.

Técnicas de Clasificación (Aprendizaje supervisado)

- Asignar una clase a un nuevo elemento en base a un conjunto de categorías previamente establecidas (**supervisado**), por ejemplo, evaluar los síntomas de un nuevo paciente y decir que tiene gripe (clase previamente establecida).
- Se basan en un **entrenamiento** en base a ejemplos con la **solución conocida** (supervisado) para crear **modelos** que permitan **clasificar nuevos casos** análogos.

Técnicas de Clasificación: EJEMPLOS

- **Paradigma Inductivo.** Árboles de decisión:
 - ID3,
 - CART,
 - C4.5,
 - See5,
 - **Random Forest** (de moda en Big Data), Leo Breiman 2001,
 - Etc.

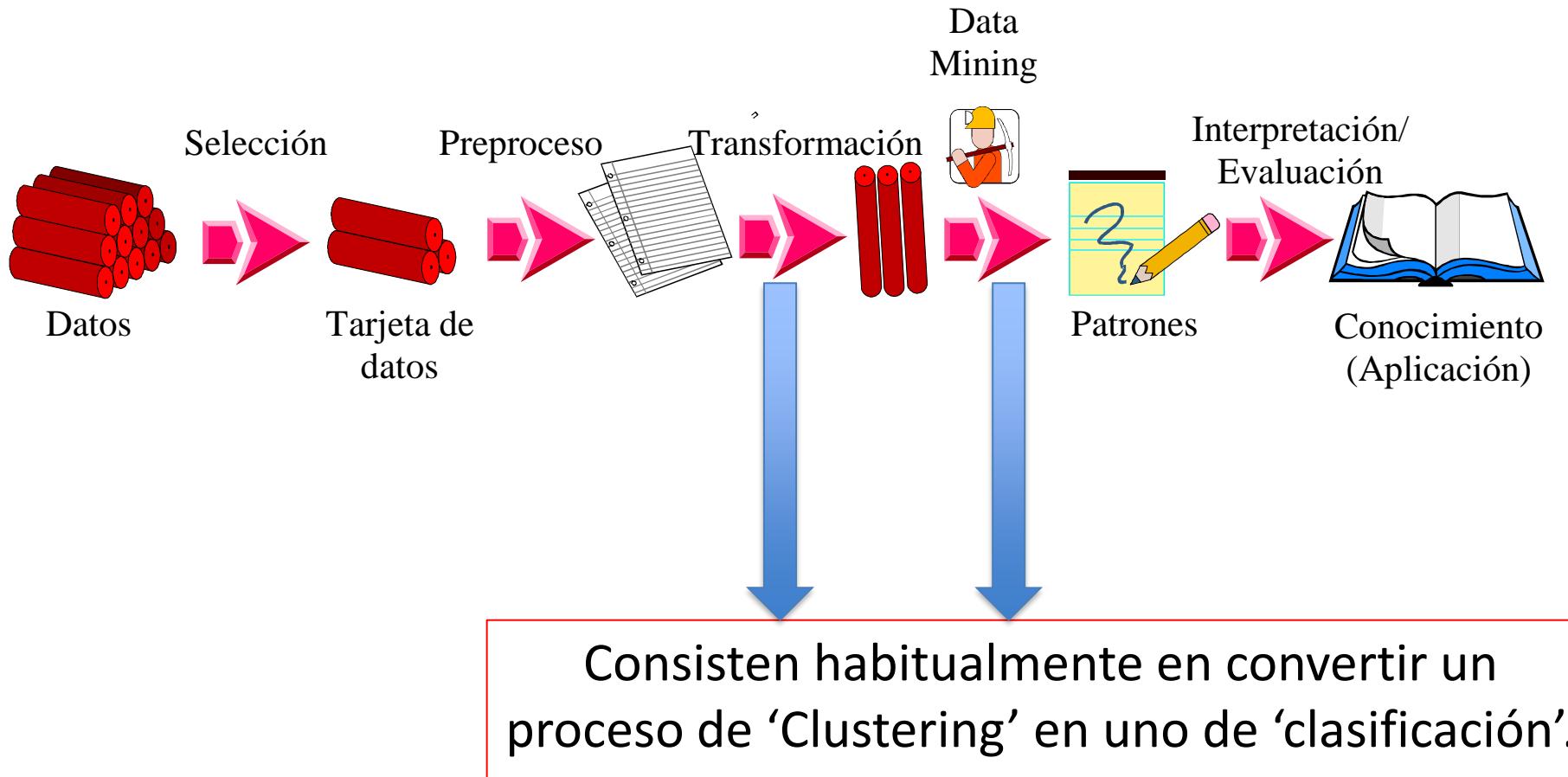
Técnicas de Clasificación: EJEMPLOS

- **Paradigma Conexionista.** Redes Neuronales Artificiales:
 - Perceptrón Multicapa (con backpropagation),
 - Convolucionales,
 - Neocognitrones,
 - Redes de Hopfield,
 - Redes recurrentes,
 - Adaline,
 - **Deep Learning** (de moda en Big Data),
 - Etc.

Técnicas de Clasificación: EJEMPLOS

- **Modelos estadísticos y probabilistas.**
 - Redes Bayesianas,
 - Naive-Bayes,
 - Máquinas de Soporte Vectorial (SVM),
 - Metaheurísticas,
 - Etc.

La importancia del KDD



7. Adecuación de los métodos a los problemas.

Análisis Descriptivo

- Resumen claro y fácil de entender de una colección de datos.
- Fundamento y concepto más básico de todas las estadísticas.
- Visualización de datos para entender el pasado y el presente.

Análisis Descriptivo

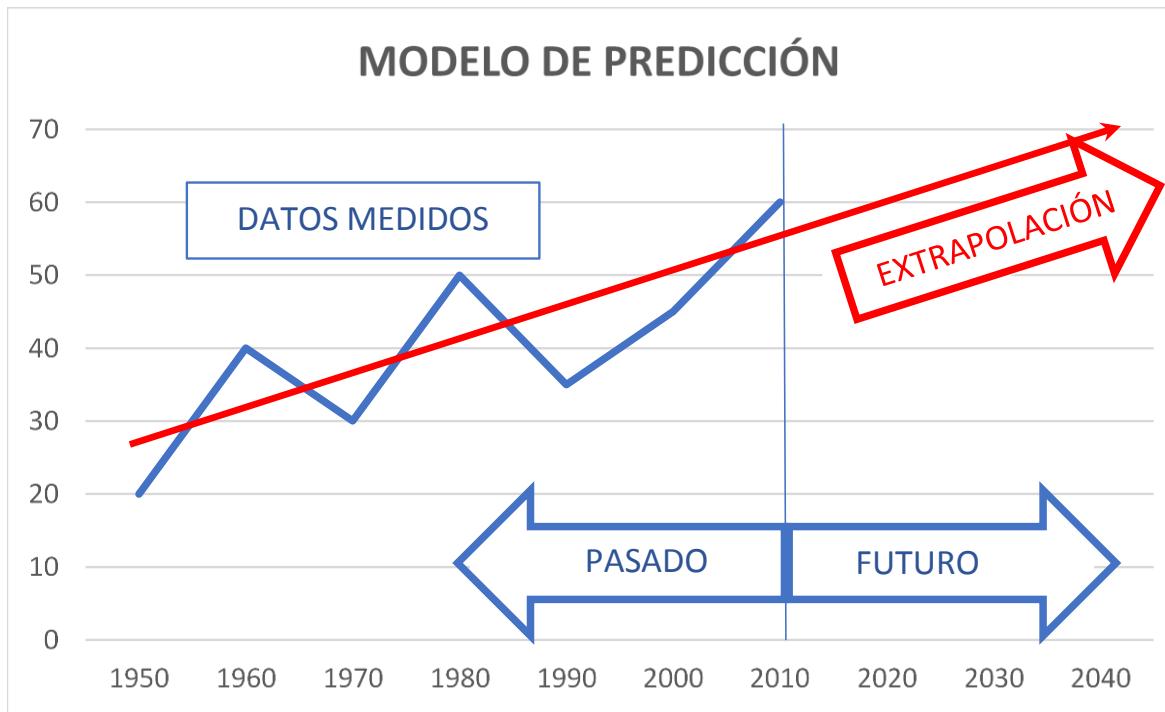
- Se describen los datos con tablas o gráficos.
- Descripciones numéricas de la variabilidad y la posición.
- Modelización descriptiva.

Análisis Predictivo

- **Extrapolación de funciones** (Tendencia para el futuro, pero no hay capacidad de pronóstico – hechos/cambios puntuales-).
- **Correlaciones entre variables** (Demasiado evidente, no suele funcionar de forma muy fina).
- Encontrar '**patrones**' en los datos que puedan ser aplicados a situaciones futuras (KDD y Data Mining).
- Métodos de **CLUSTERING Y CLASIFICACIÓN**.

Análisis Predictivo

- **Extrapolación de funciones (Por ejemplo Estimaciones o Líneas de Tendencia).**



Análisis Predictivo: Series Temporales

- **Estacionarias** (Medias y/o variabilidad se mantienen constantes).
- **No Estacionarias** (Medias y/o variabilidad NO se mantienen constantes, cambios de varianza/tendencias).
- **Tendencias:**
 - Método de Mínimos cuadrados.
 - Tendencias evolutivas.
 - Diferenciación estacional.

Análisis Predictivo: Series Temporales

- **Predicción:** Alisadores exponenciales.
 - Alisado exponencial simple.
 - Alisado exponencial lineal de Holt.
 - Alisado exponencial estacional de Holt-Winters.
- **Interpolación:** Predecir datos faltantes.

<http://www.statsoft.com/Textbook/Time-Series-Analysis>

Análisis Prescriptivo

- El análisis predictivo se centra en **un escenario futuro**.
- El prescriptivo se centra en **múltiples alternativas**.
- Por lo tanto, un modelo prescriptivo puede ser considerado como **una combinación de modelos predictivos** (uno por cada posible escenario), que **se ejecutan en paralelo**.
- El objetivo es encontrar la mejor opción posible: **OPTIMIZACIÓN**.

Análisis Prescriptivo

- Técnicas:
 - Técnicas de Investigación Operativa,
 - Algoritmos Genéticos,
 - Técnicas estocásticas,
 - Metaheurísticas,
 - Etc.

Análisis Prescriptivo



Fuente: Acotrend.com

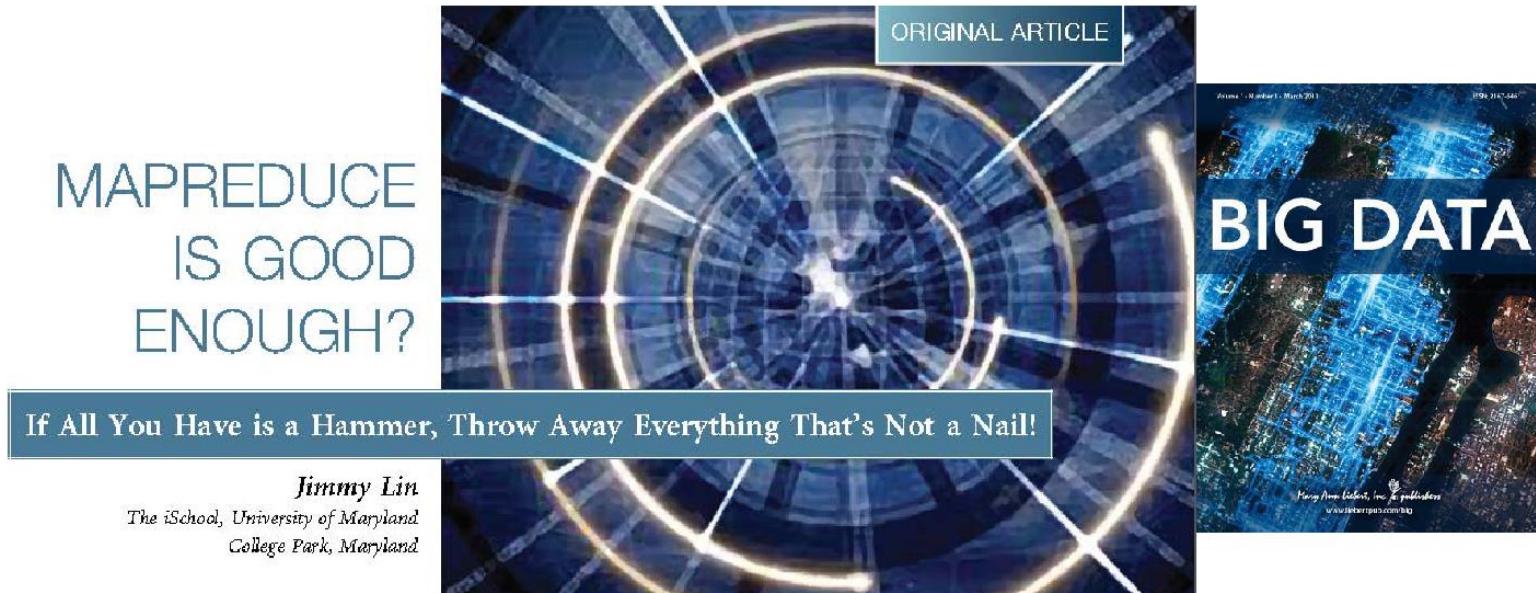
Análisis Prescriptivo



Source: Competing on Analytics: The New Science of Winning (Davenport / Harris): Gartner

8. Herramientas actuales (en entornos Big Data) para implementar estas soluciones.

MapReduce



Los siguientes tipos de algoritmos son ejemplos en los que MapReduce no funciona bien:

Iterative Graph Algorithms
Gradient Descent
Expectation Maximization



MapReduce

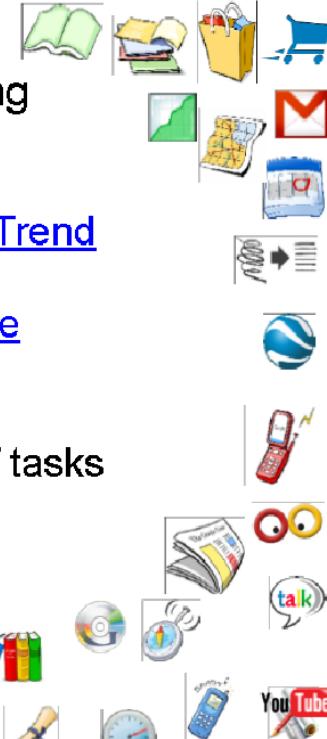
- Hay muchas **limitaciones** para los **algoritmos de grafos iterativos**.
- Por ejemplo, en PageRank, cada iteración es una ejecución completa de MapReduce.
- Por ello, se han propuesto una serie de **extensiones** para mejorar el cálculo iterativo.
- Por ejemplo **Pregel** (Google)
<http://www.michaelnielsen.org/ddi/pregel>

G. Malewicz, M. Austern, A. Bik, J. Dehnert, I. Horn, N. Leiser, G. Czajkowski:
“Pregel: A system for large scale graph processing”, ACM SIGMOD, 2010.

MapReduce

MapReduce inside Google

Google



Googlers' hammer for 80% of our data crunching

- [Large-scale web search indexing](#)
- Clustering problems for [Google News](#)
- Produce reports for popular queries, e.g. [Google Trend](#)
- Processing of [satellite imagery data](#)
- Language model processing for [statistical machine translation](#)
- Large-scale [machine learning problems](#)
- Just a plain tool to reliably spawn large number of tasks
 - e.g. parallel data backup and restore

The other 20%? e.g. [Pregel](#)



Reflexión

Una reflexión tecnológica (si resulta adecuada...):

- Algoritmos basados en el descenso del gradiente.
- Algoritmos dependientes y no dependientes de la iteración anterior, modelos aleatorios...
- Procesos con flujos acíclicos de procesamiento de datos.

<http://tez.apache.org/>

Reflexión

Una reflexión tecnológica (si resulta adecuada...):

- Leo Breiman...
 - Random Forest, Bagging (Bootstrap aggregation) y
 - Boosted Regresion Trees (secuencial, iterativo).
- Deep Learning...
 - Cybenko, backpropagation...

Librerías para Big Data

Mahout



Scalable machine learning
and data mining



Apache Mahout has implementations of a wide range of machine learning and data mining algorithms:
clustering, classification, collaborative filtering and frequent pattern mining



Mahout currently has

- Collaborative Filtering
- User and Item based recommenders
- K-Means, Fuzzy K-Means clustering
- Mean Shift clustering
- Dirichlet process clustering
- Latent Dirichlet Allocation
- Singular value decomposition

- Parallel Frequent Pattern mining
- Complementary Naive Bayes classifier
- Random forest decision tree based classifier
- High performance [java](#) collections (previously colt collections)
- A vibrant community
- and many more cool stuff to come by this summer thanks to Google summer of code



Biblioteca de código abierto en APACHE

<http://mahout.apache.org/>



<https://spark.apache.org/docs/latest/mllib-guide.html>

Machine Learning Library (MLlib) Guide

MLlib is Spark's scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives, as outlined below:

- [Data types](#)
- [Basic statistics](#)
 - summary statistics
 - correlations
 - stratified sampling
 - hypothesis testing
 - random data generation
- [Classification and regression](#)
 - [linear models \(SVMs, logistic regression, linear regression\)](#)
 - [naive Bayes](#)
 - [decision trees](#)
 - [ensembles of trees](#) (Random Forests and Gradient-Boosted Trees)
 - [isotonic regression](#)
- [Collaborative filtering](#)
 - alternating least squares (ALS)
- [Clustering](#)
 - [k-means](#)
 - [Gaussian mixture](#)
 - [power iteration clustering \(PIC\)](#)
 - [latent Dirichlet allocation \(LDA\)](#)
 - [streaming k-means](#)
- [Dimensionality reduction](#)
 - [singular value decomposition \(SVD\)](#)
 - [principal component analysis \(PCA\)](#)
- [Feature extraction and transformation](#)
- [Frequent pattern mining](#)
 - [FP-growth](#)
- [Optimization \(developer\)](#)
 - [stochastic gradient descent](#)
 - [limited-memory BFGS \(L-BFGS\)](#)



<http://0xdata.com/>

Data Science in H₂O

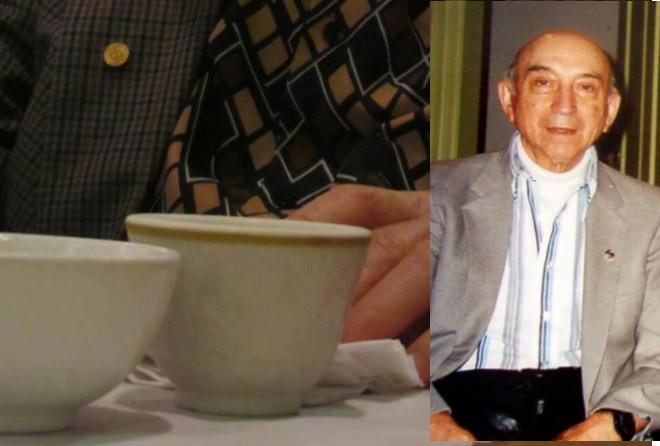
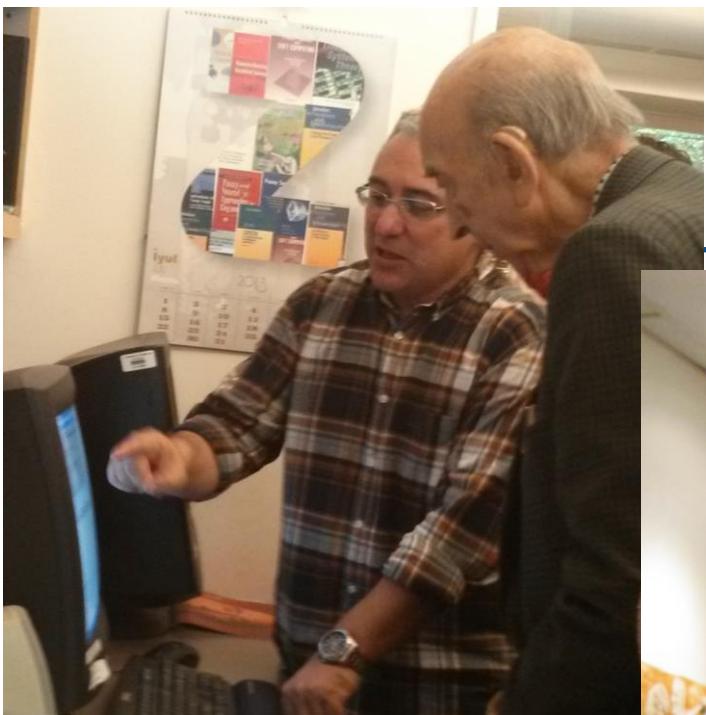
- Cox Proportional Hazards Model
- Deep Learning
- Generalized Linear Model
- Gradient Boosted Regression and Classification
- K-Means
- Naive Bayes
- Principal Components Analysis
- Random Forest
- Summary
- Data Science and Machine Learning
- Stochastic Gradient Descent
- References

Soporte para R, Hadoop y Spark

Funcionamiento: Crea una máquina virtual con Java en la que optimiza el paralelismo de los algoritmos

- Librería que contiene algoritmos de Deep Learning
 - Récord del mundo en el problema MNIST sin preprocesamiento

<http://0xdata.com/blog/2015/02/deep-learning-performance/>



Lógica Borrosa
Lotfi A. Zadeh
1921-2017

GRACIAS

