MÁSTER EN DIRECCIÓN Y GESTIÓN DE MARKETING DIGITAL

NUEVAS TECNOLOGÍAS Y MARKETING

2015

Profesor: Dr. LEONARDO FRANCO

MÁSTER EN DIRECCIÓN Y GESTIÓN DE MARKETING DIGITAL

NUEVAS TECNOLOGÍAS Y MARKETING

INTRODUCCIÓN A LA MINERÍA DE DATOS

2015

Profesor: Dr. LEONARDO FRANCO

DATA MINING

- Data Mining y Big Data
- Introducción al problema de minería de datos
- Software de uso estándar
- Análisis inicial y pre-procesamiento de los datos
- Modelos para predicción
- Visualización y análisis de resultados
- **Caso práctico**

BIG DATA y DATA MINING

DATA MINING

TRATAMIENTO DE DATOS PARA SU APROVECHAMIENTO COMO **INFORMACIÓN ÚTIL**

Grandes Volúmenes de datos

Con el BIG DATA aparecen nuevos factores relacionados con:

Recogida, Almacenamiento y Transmisión de la información
Tiempo de procesamiento
Variabilidad de los datos

Y entonces el proceso de minería de datos pasa a ser una parte del fenómeno de BIG DATA: hacen faltan nuevas formas de manipular la información y métodos de data mining que se adapten a grandes volúmenes

¿Qué procesos involucra BIG Data?

Tareas		Tecnologías
• Almacenar y acceder a grandes volúmenes de datos no-estructurados	→	Manejo de base de datos
• Procesar grandes volúmenes de datos on-line (data streams)	→	Modelos de programación paralela y distribuida
• **Dar utilidad a los datos**	→	**Técnicas de minería de datos**
• **Tecnologías para hacer predicciones**	→	**Machine learning**

DATA MINING

TRATAMIENTO DE DATOS PARA SU APROVECHAMIENTO COMO **INFORMACIÓN ÚTIL**

GENERAR CONOCIMIENTO

- Entendible
- Válido
- Nuevo
- Útil

DESCRIBIR LOS DATOS DE UNA FORMA SENCILLA, POR EJEMPLO A TRAVÉS DE REGLAS

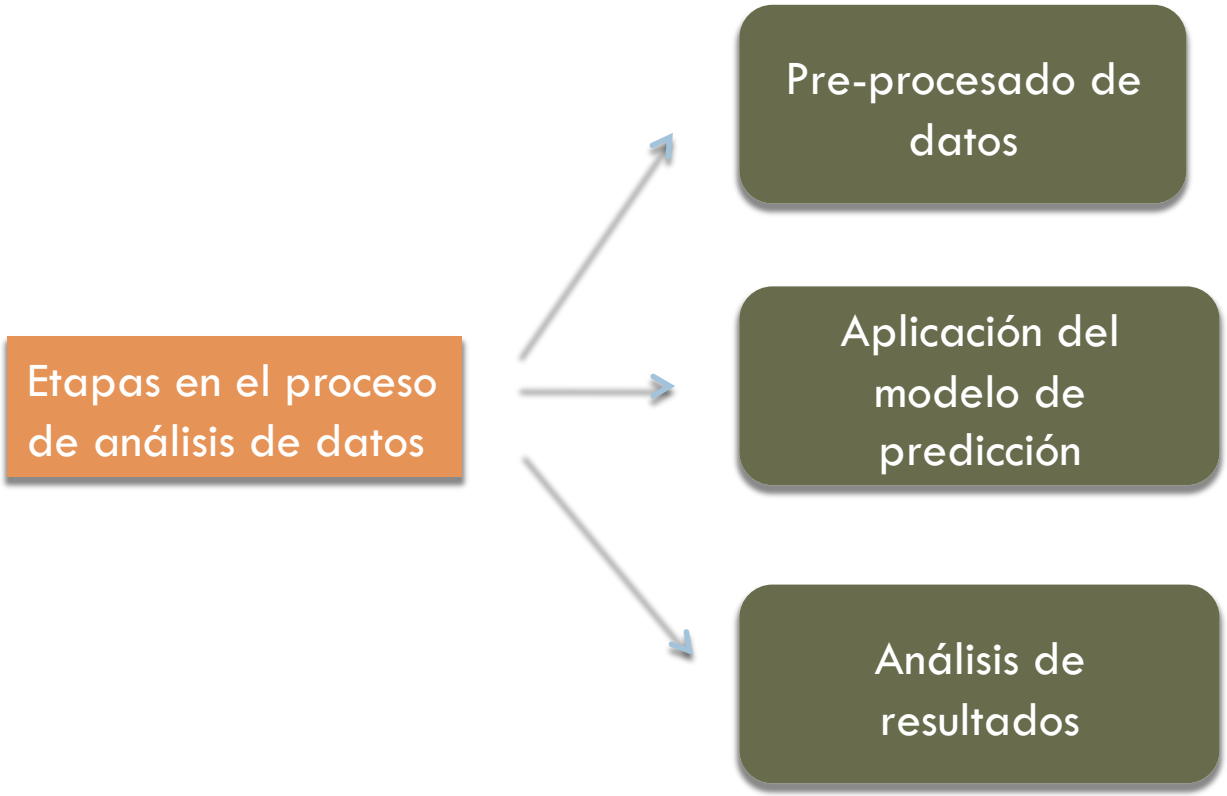
IF A >15 **&** B < 70

THEN SUCEDE C

DM y BD SON DISCIPLINAS GUIADA POR LOS DATOS Y NO POR LOS MODELOS

Etapas en el análisis de los datos

PODEMOS AGRUPAR LAS DIFERENTES TÉCNICAS DE LA MINERÍA DE DATOS EN LOS SIGUIENTES TRES SUB-PROCESOS



Marco general del problema de DATA MINING

Pre-procesado de datos → Aplicación del modelo de predicción → Análisis de resultados

Pre-procesado de datos

- Análisis estadístico de los datos
- Normalización ó Discretización
- Selección de variables
- Tratamiento de datos perdidos

Aplicación del modelo de predicción

- Elección del modelo
- Ajuste de parámetros
- Conjunto de entrenamiento
- Métodos de regularización

Análisis de resultados

- Medidas de performance
- Comparación con otros modelos
- Reglas generadas por el modelo e interpretación
- Visualización

Preprocesado de datos

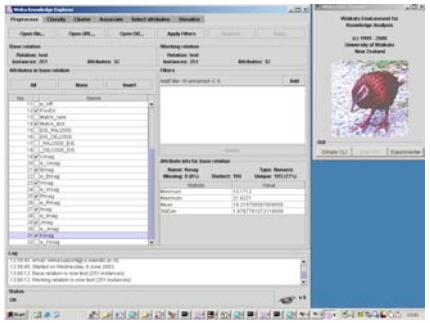
etapa fundamental si quieren obtenerse buenos resultados en los siguientes pasos

Análisis estadístico de los datos	Conocer los datos que vamos a tratar: media y rango de cada variable, distribución de las clases, outliers
Normalización ó Discretización	Aconsejable para que no predominen ciertas variables sobre otras
Selección de variables	Utilizar todas las variables disponibles puede ser perjudicial para nuestro modelo, además de complejo
Tratamiento de datos perdidos	Problema muy frecuente y complejo de tratar. Muy útil en casos donde no es aconsejable la eliminación de instancias

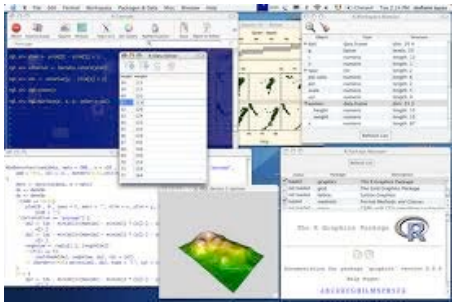
Big data requiere algoritmos rápidos y eficientes

Software para Data Mining

XLMiner Software utilizado en business intelligence, interfaz sencilla es un plug-in para Excel



WEKA: Desarrollado en JAVA, permite la comparación rápida y sencilla entre diferentes algoritmos



R: Entorno para análisis estadístico. Paquetes para efectuar análisis muy avanzados.

Fin presentación día 1

MÁSTER EN DIRECCIÓN Y GESTIÓN DE MARKETING DIGITAL

NUEVAS TECNOLOGÍAS Y MARKETING

INTRODUCCIÓN A LA MINERÍA DE DATOS

2015

Profesor: Dr. LEONARDO FRANCO

DATA MINING Día 2

- Software WEKA
- Pre-procesamiento de datos: Outliers, Normalización y datos perdidos

Tipos de datos

- Numéricos
  - Valores **Continuos** (Ej. : Altura de una persona en cm)
  - Valores **Enteros** (Ej.: Edad de una persona en años)
  - Valores **Categóricos**
    - **Nominales** (no ordenados)
    - **Ordinales**
- Texto

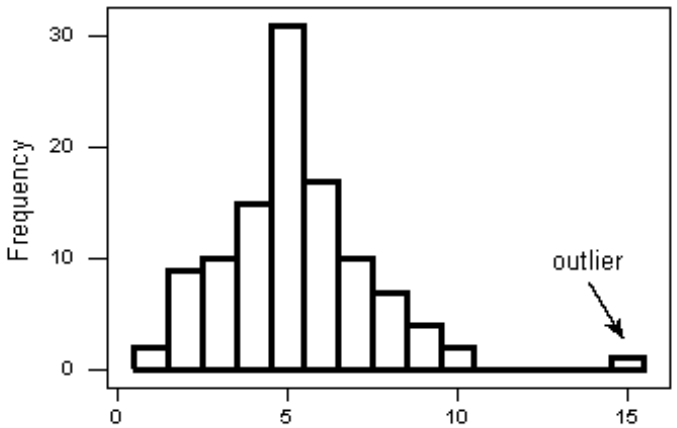
Los datos que suelen dar mayores problemas son los de tipo categórico (y mayor aún siendo nominales), en estos casos suelen usarse variables tipo dummy

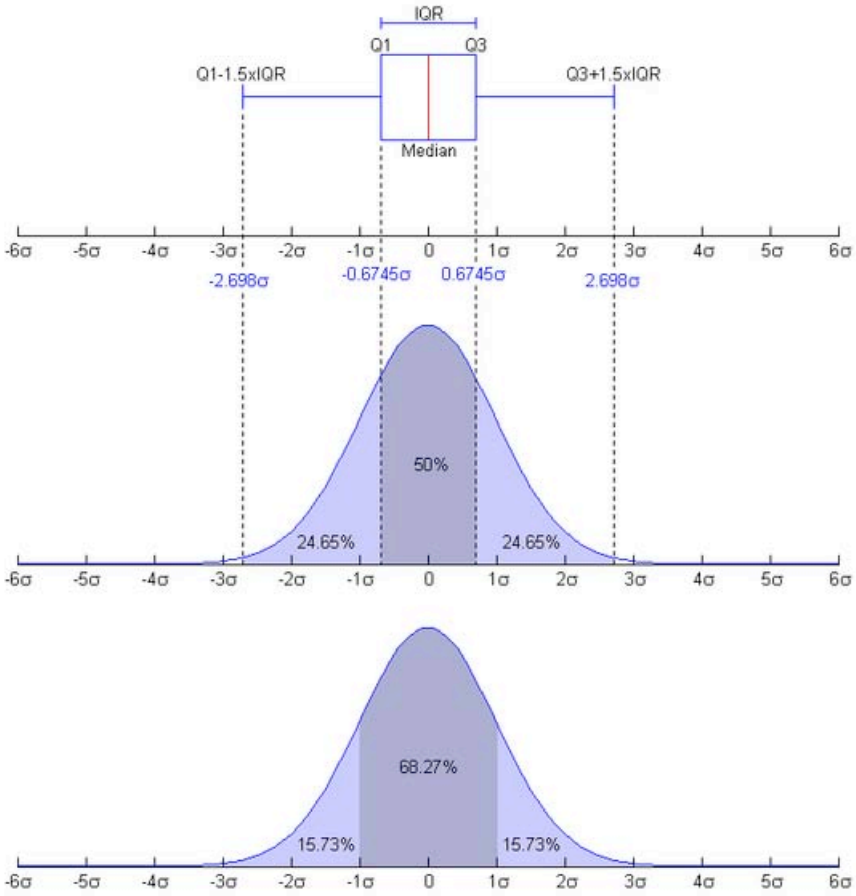
Ejemplo la variable Empleo (Estudiante, desempleado, Empleado, Jubilado) suele ser representada por4 variables :

Estudiante (1,0,0,0) Desempleado (0,1,0,0) Empleado (0,0,1,0) Jubilado (0,0,0,1)

Tratamiento de Outliers

- Valor atípico (outlier): una observación que se encuentra fuera de la distribución global de los datos.

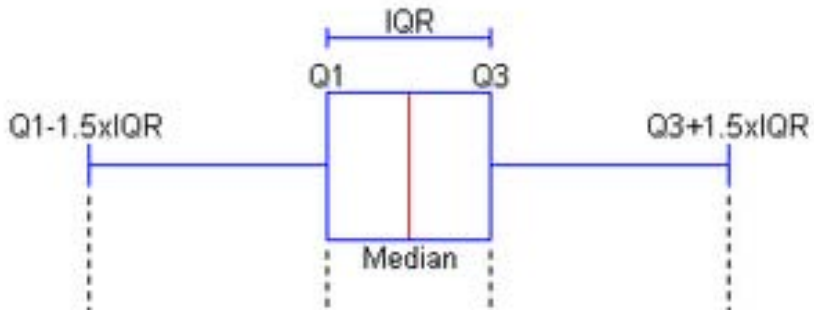




Tratamiento de Outliers

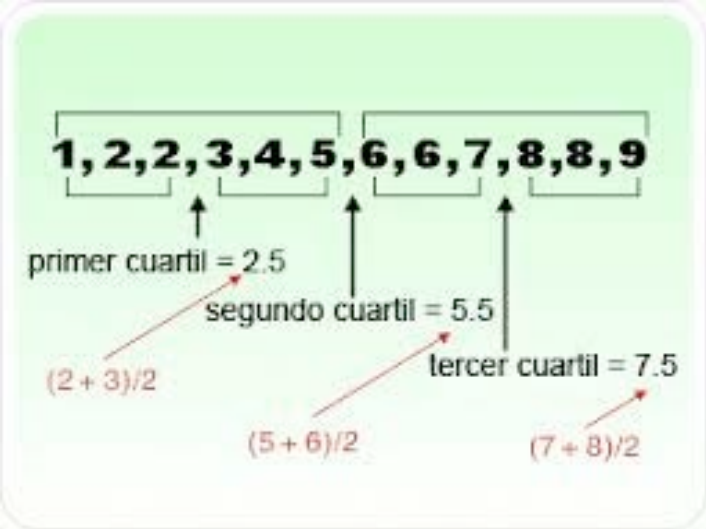
- Suelen eliminarse los patrones considerados outliers, aunque dependerá del tamaño y de la distribución del conjunto de datos. En el caso de Big Data, como tenemos muchos datos, no es un problema eliminarlos, aunque deberíamos tener cuidado de que no sea un grupo nuevo de consumidores.

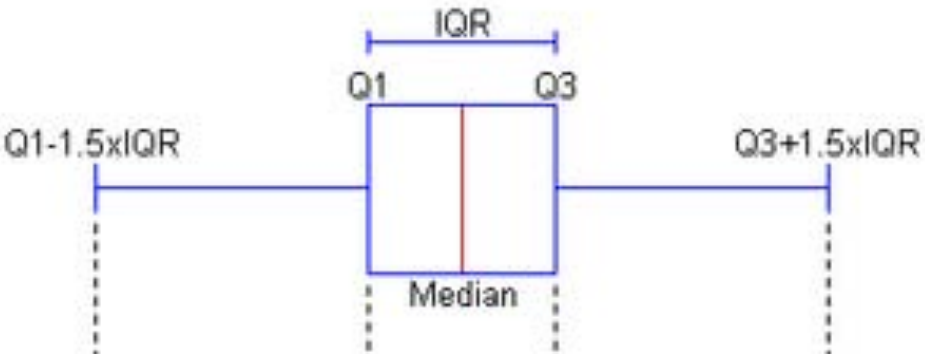
Una estrategia es utilizar los dos conjuntos y ver las diferencias que obtenemos cuando hacemos predicción.



Patrones con valores menores de Q1-1.5 * IQR o mayores que Q3 + 1.5 *IQR serán candidatos a ser eliminados

Rango Intercuartílico





Límite inferior: Q1-1.5*5 = 2.5 -7.5 = -5

Límite superior: 7.5 +1.5* 5 = 7.5 +7.5 = 15

Mediana: 5.5

Q1= 2.5 Q3= 7.5

IQR= Q3 – Q1 = 5

Cualquier dato menor a -5 o mayor a 15 será considerado un outlier y será candidato a ser eliminado

Ejercicio Outliers

a) Encontrar el Máximo, mínimo, Q1, Q2 (mediana), Q3 y Máximo.
b) Calcular el rango y el rango intercuartílico.
c) Obtener entonces los límites de los outliers y los outliers.

Datos: {79,81,83, 86,87,88,89,90,91,95,108,111}

Mínimo:
Q1:
Mediana:
Q3:
Máximo:
Rango:
IQR:

Límites de los outliers:

Outliers:

Ejercicio Outliers

a) Encontrar el Máximo, mínimo, Q1, Q2 (mediana), Q3 y Máximo.
b) Calcular el rango y el rango intercuartílico.
c) Obtener entonces los límites de los outliers y los outliers.

Datos: {79,81,83, 86,87,88,89,90,91,95,108,111}

Mínimo: 79
Q1: 84.5
Mediana: 88.5
Q3: 93
Máximo: 111
Rango: 32
IQR: 8.5

Límites de los outliers: (71.75, 105.75)

Outliers: {108,111}

Normalización y Estandarización

Un aspecto importante de los datos es también el rango de valores cada uno de las variables, ya que grandes diferencias puede hacer que los métodos a aplicar prioricen para su análisis las variables con valores mayores. También es útil aplicarlas para evitar errores de precisión y de desbordamientos aritmético ("overflow") en los procedimientos a aplicar.

Existen dós técnicas bastante sencillas que es recomendable aplicar luego de eliminar outliers: Normalización por rango y normalización por Z-score (estandarización).

$$V' = \frac{(V - min)}{(max - min)}$$

Normalización a [0, 1]

$$V' = \frac{(V - media)}{desv_std}$$

Z-score

Tutorial WEKA pre-processing
https://www.youtube.com/watch?v=Ys3x2E9WZXQ

Datos faltantes (Missing data)

- Pueden eliminarse instancias con valores perdidos
- Posibilidad de reemplazar valores faltantes por la media (valores numéricos) y moda (valores categóricos)
- Dependerá del porcentaje de valores faltantes y del tamaño de nuesra base de datos

Introduction to Weka

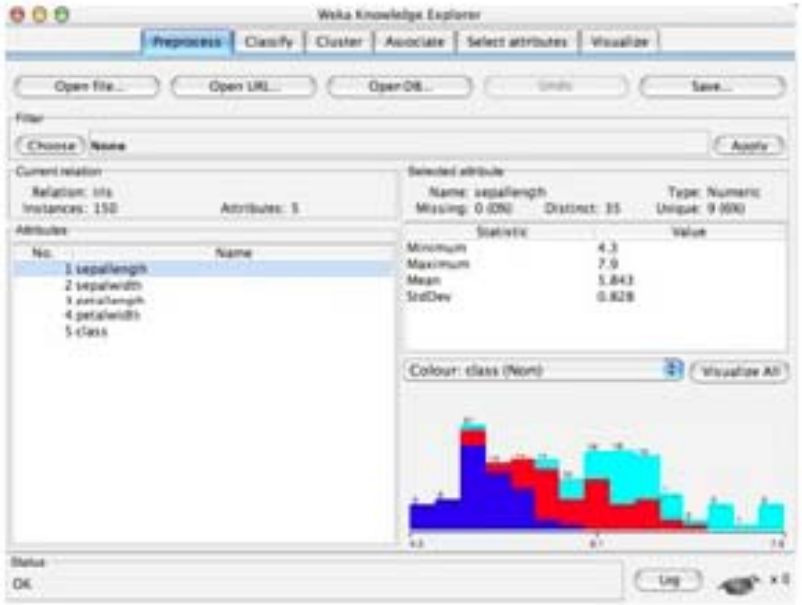
Software de código abierto y sencilla operación.

Colección de algoritmos de machine learning para tareas de minería de datos

What is Weka?

- Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

CLI Vs GUI



- Recommended for in-depth usage
- Offers some functionality not available via the GUI

- Explorer
- Experimenter
- Knowledge Flow

Datasets in Weka

- Each entry in a dataset is an instance of the java class:
  - weka.core.Instance
- Each instance consists of a number of attributes

Attributes

- *Nominal*: one of a predefined list of values
  - e.g. red, green, blue
- *Numeric*: A real or integer number
- *String:* Enclosed in “double quotes”

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose | **RemoveWithValues** -S 0.0 -C 15 -L last | Apply

Current relation

Relation: housing-weka.filters.unsupervised.attribut...
Instances: 383 Attributes: 16

Attributes

All | None | Invert | Pattern

No.	Name
6	RM
7	AGE
8	DIS
9	RAD
10	TAX
11	PTRATIO
12	B
13	LSTAT
14	class
15	Outlier
16	ExtremeValue

Remove

Selected attribute

Name: Outlier Type: Nominal
Missing: 0 (0%) Distinct: 1 Unique: 0 (0%)

No.	Label	Count
1	no	383
2	yes	0

Class: ExtremeValue (Nom) | Visualize All

Status

OK

Log | x 0

Tipos de Algoritmos de Aprendizaje

Aprendizaje Supervisado

- Establece una correspondencia entre la entrada y la salida deseada (Clasificación, Regresión) [x, f(x)]

Aprendizaje No-supervisado

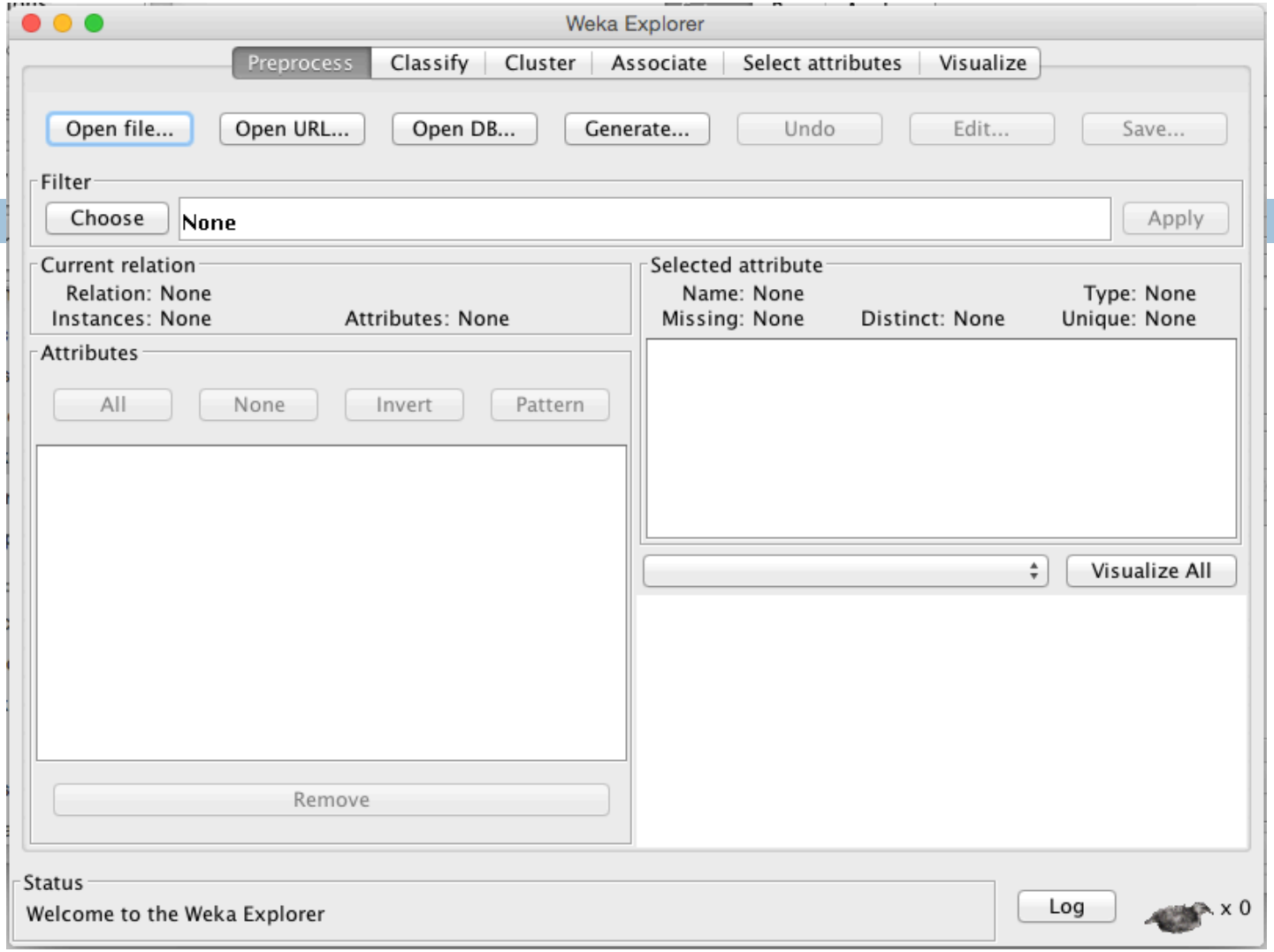
- Todo el proceso de modelado se realiza sobre un conjunto de ejemplos donde se tiene solo las entradas [x]

Aprendizaje por refuerzo

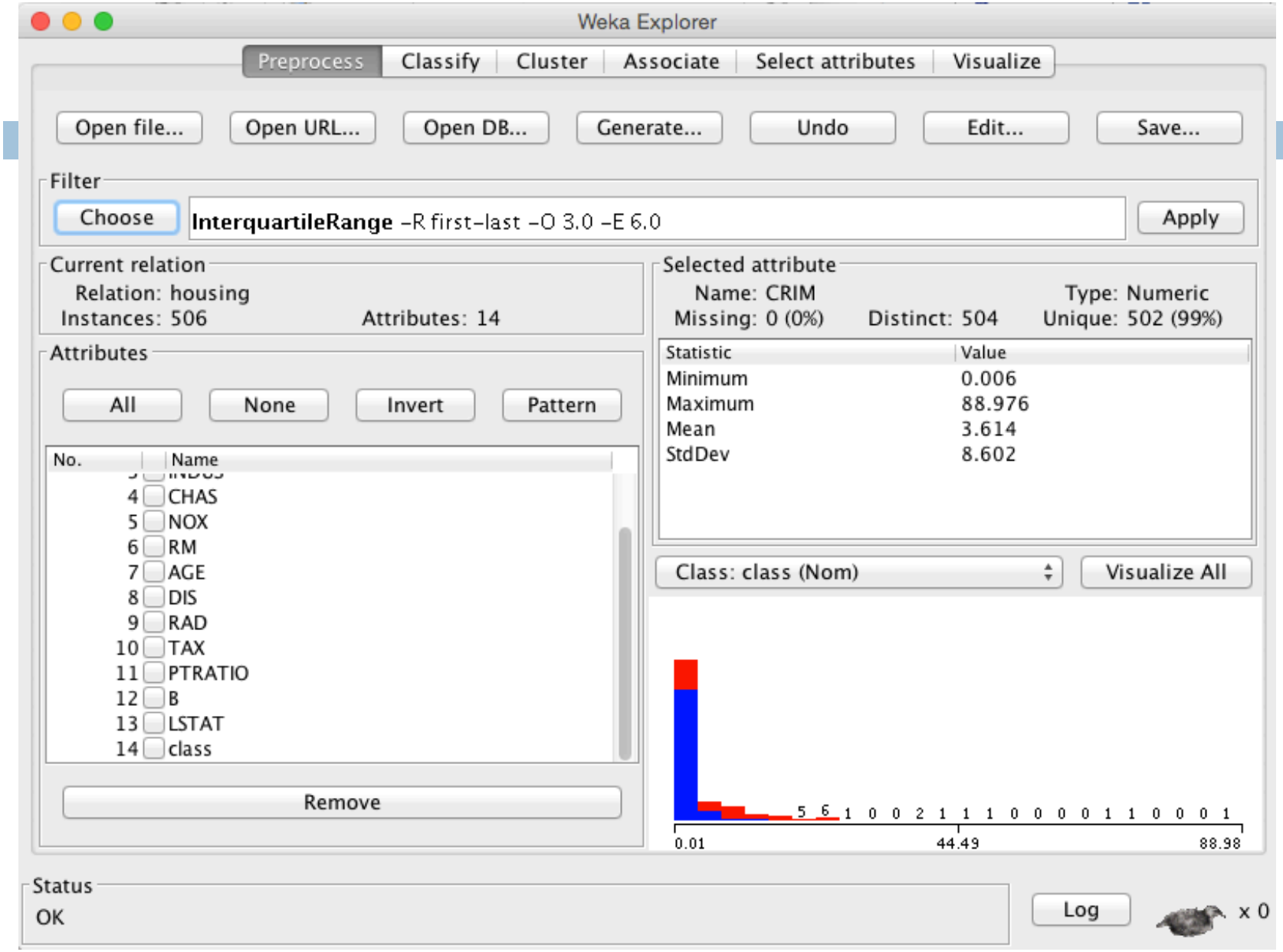
- El algoritmo re-aprende constantemente, en función a sus experiencias, refuerza el aprendizaje si tiene éxito (Feedback) [x, f(x)]

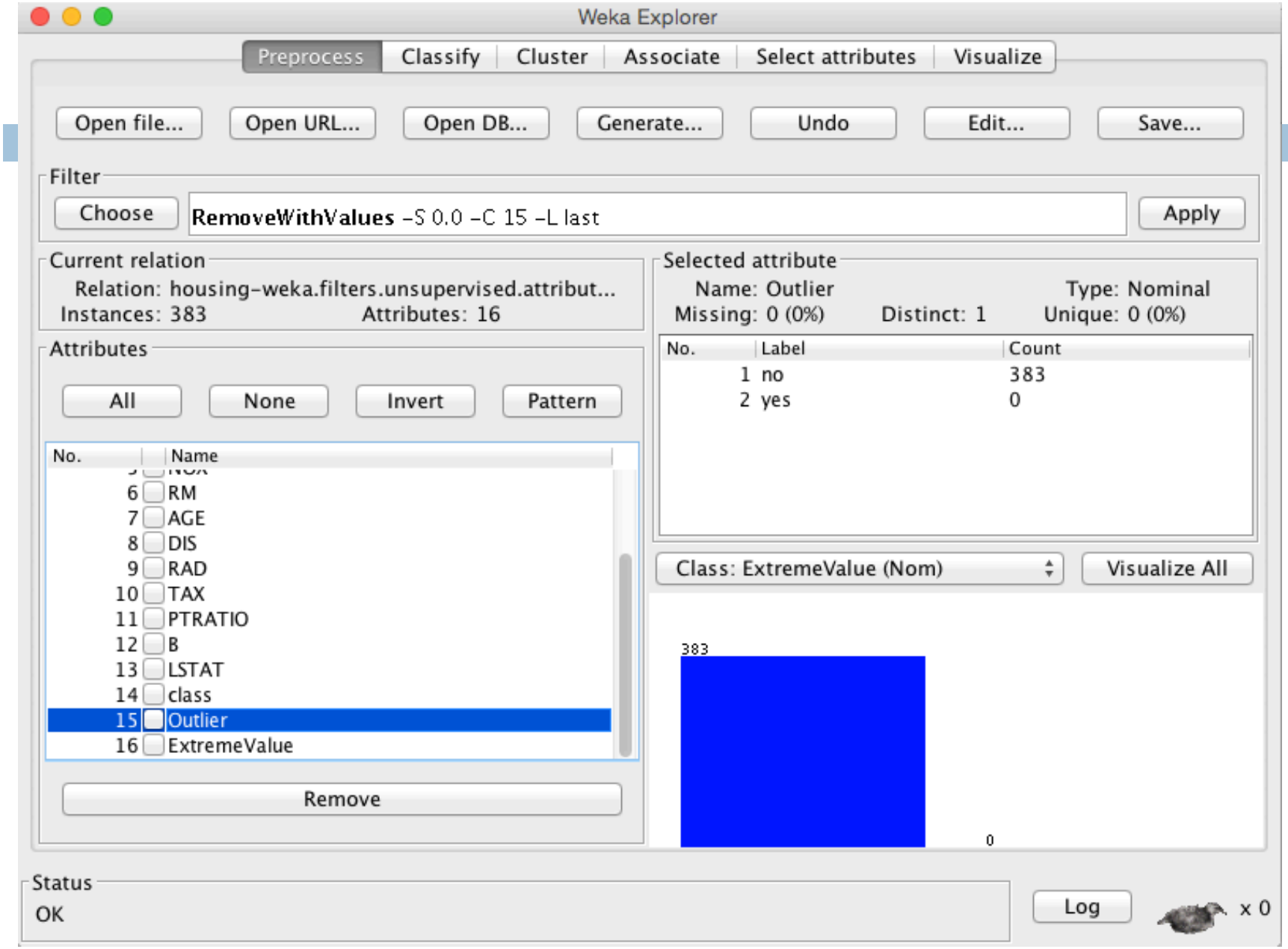
Comenzando con WEKA

Abrimos la aplicación "Explorer"

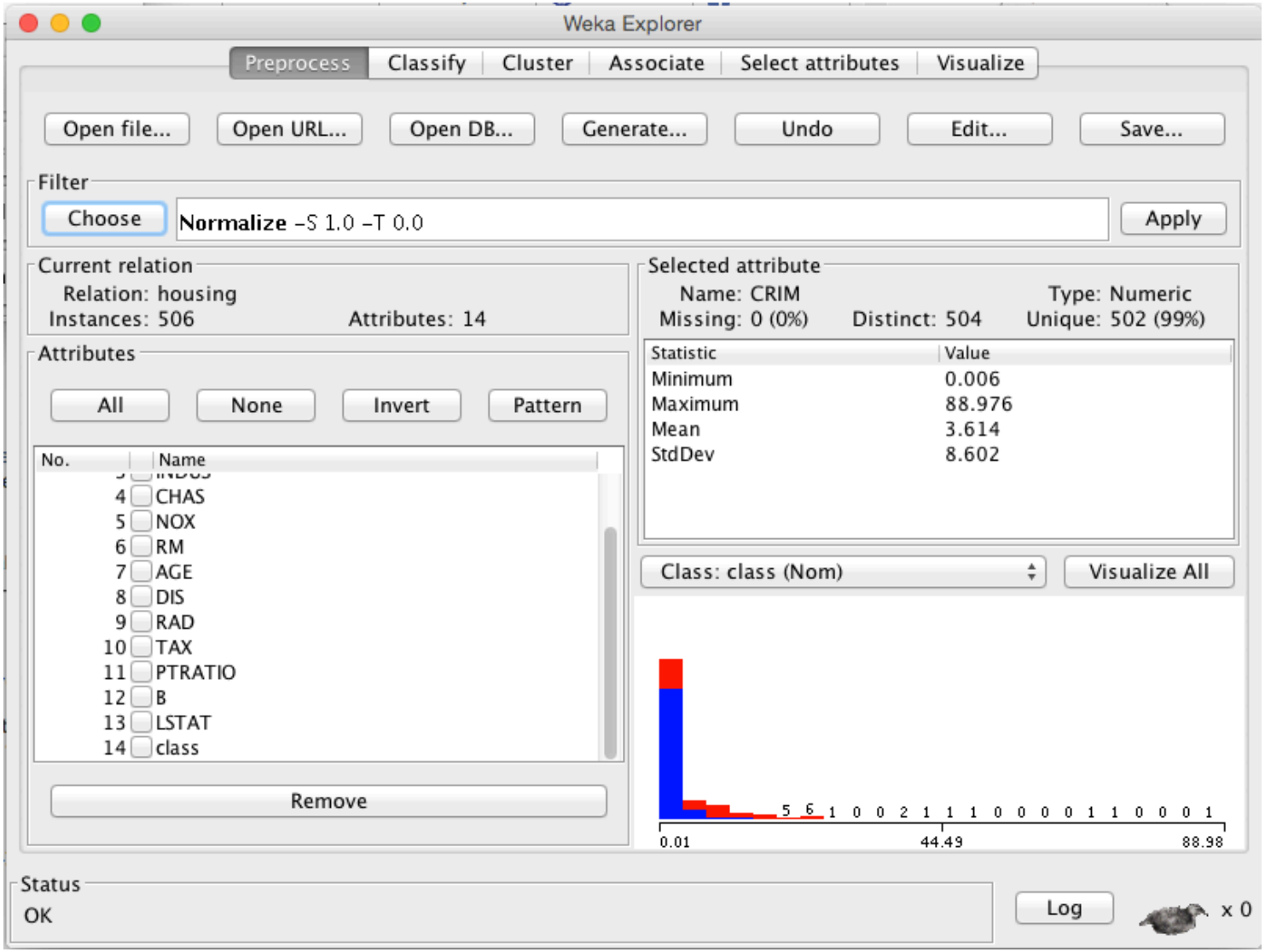


Usando la opción “Open File” cargaremos el archivo “Boston-housing.arff”


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
InterquartileRange -R first-last -O 3.0 -E 6.0
Apply
Current relation
Relation: housing
Instances: 506
Attributes: 14
Selected attribute
Name: CRIM
Type: Numeric
Missing: 0 (0%)
Distinct: 504
Unique: 502 (99%)
Statistic
Value
Minimum
0.006
Maximum
88.976
Mean
3.614
StdDev
8.602
Attributes
All
None
Invert
Pattern
No.
Name
4 CHAS
5 NOX
6 RM
7 AGE
8 DIS
9 RAD
10 TAX
11 PTRATIO
12 B
13 LSTAT
14 class
Remove
Class: class (Nom)
Visualize All
5 6 1 0 0 2 1 1 1 0 0 0 0 1 1 0 0 0 1
0.01
44.49
88.98
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemoveWithValues -S 0.0 -C 15 -L last
Apply
Current relation
Relation: housing-weka.filters.unsupervised.attribut...
Instances: 383
Attributes: 16
Selected attribute
Name: Outlier
Type: Nominal
Missing: 0 (0%)
Distinct: 1
Unique: 0 (0%)
No. Label Count
1 no 383
2 yes 0
Attributes
All
None
Invert
Pattern
No. Name
6 RM
7 AGE
8 DIS
9 RAD
10 TAX
11 PTRATIO
12 B
13 LSTAT
14 class
15 Outlier
16 ExtremeValue
Remove
Class: ExtremeValue (Nom)
Visualize All
383
0
Status
OK
Log
x 0

Normalización de datos

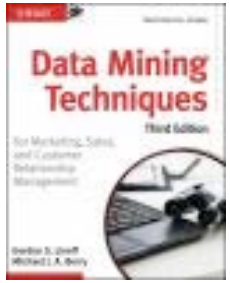


Valores perdidos

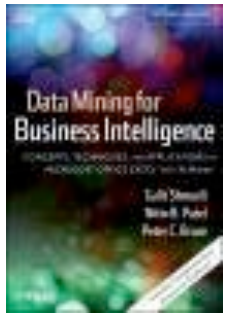
- Eliminar registros (No siempre es conveniente)
- Métodos sencillos de imputación (Media, valor aleatorio en el rango)
- Métodos sofisticados de imputación (Redes neuronales, imputación múltiple, etc.)

Bibliografía Data Mining

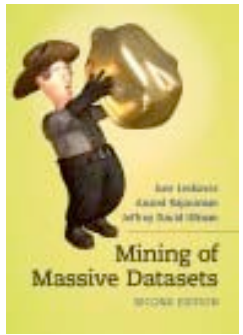
[DM-R1] Data mining techniques: for marketing, sales, and customer relationship management por Berry, Michael J. A; Linoff, Gordon, Wiley, (2011).



[DM-R2] Data mining for business intelligence: concepts, techniques. Shmueli, Galit; Patel, Nitin R; Bruce, Peter C., Wiley. (2010) Disponible en Jábega.

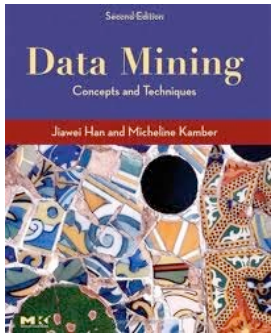


[DM-R3] Mining of Massive Datasets J. Leskovec, A. Rajaraman, J. Ullman Cambridge University Press 2011
http://www.cambridge.org/us/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets?format=HB

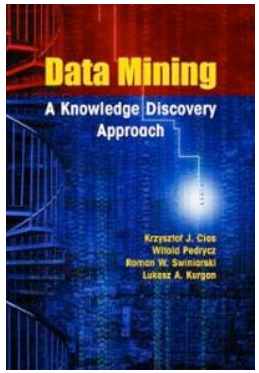


Bibliografía Adicional Data Mining

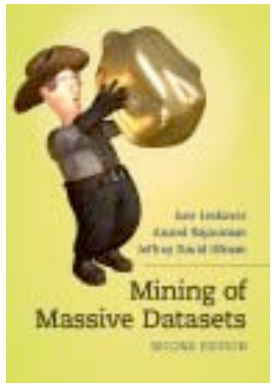
Data Mining: Concepts and Techniques, 2nd ed.
Jiawei Han and Micheline Kamber
Morgan Kaufmann 2006
http://www.cs.uiuc.edu/~hanj/bk2/



Data Mining: A Knowledge Discovery Approach
Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, and Lukasz A. Kurgan.
Springer-Verlag New York 2007
http://link.springer.com/book/10.1007/978-0-387-36795-8/page/1



Mining of Massive Datasets
J. Leskovec, A. Rajaraman, J. Ullman
Cambridge University Press 2011
http://www.cambridge.org/us/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets?format=HB



Predictive Analytics

- Introducción
- Aprendizaje supervisado vs no supervisado
- ¿Como hacer predicciones?
- Sobre-entrenamiento
- Data Splitting

Analicemos a modo de introducción el siguiente artículo

http://www.cmo.com/articles/2013/3/20/predictive_analytics.html

Con los datos tradicionales, podemos responder a :

- Which of my customers are most likely to terminate service? What steps could I take now to remediate this?
- Of the thousands of prospects my sales team are speaking to, which ones are most/least likely to buy? How can I align my sales team's time with those most likely to buy?
- Which of my current customers are most likely to buy some of my other products? Which products should I try to cross-market to which customers?

Tal vez SÍ, pero dependerá de cómo estemos analizándolos

Y aún podemos hacer más ...

- Healthcare: Which patients are at risk of developing certain conditions?
- Fraud detection: Which transactions are most likely to be fraudulent?
- Insurance: What is the risk of illness and death? Of a car accident?
- Economics: What is the likely unemployment rate next year?
- Online retailing: What book or movie recommendation is most likely to result in an incremental purchase?
- Cable television: What offer can I make to this subscriber to prevent her from terminating service?

In each example, past data is used to predict future outcomes.

Pasado, presente y futuro

- The key difference between successful and unsuccessful marketing organizations is how they are scaling up for the data challenge.
- Five years ago, it was probably OK to review the data each month (or quarter) for “insights” to incorporate into future marketing programs.
- Today marketing organizations are scaling these processes by developing predictive models that identify patterns in historical and transactional data.

Predictive analytics is a reality for three main reasons:

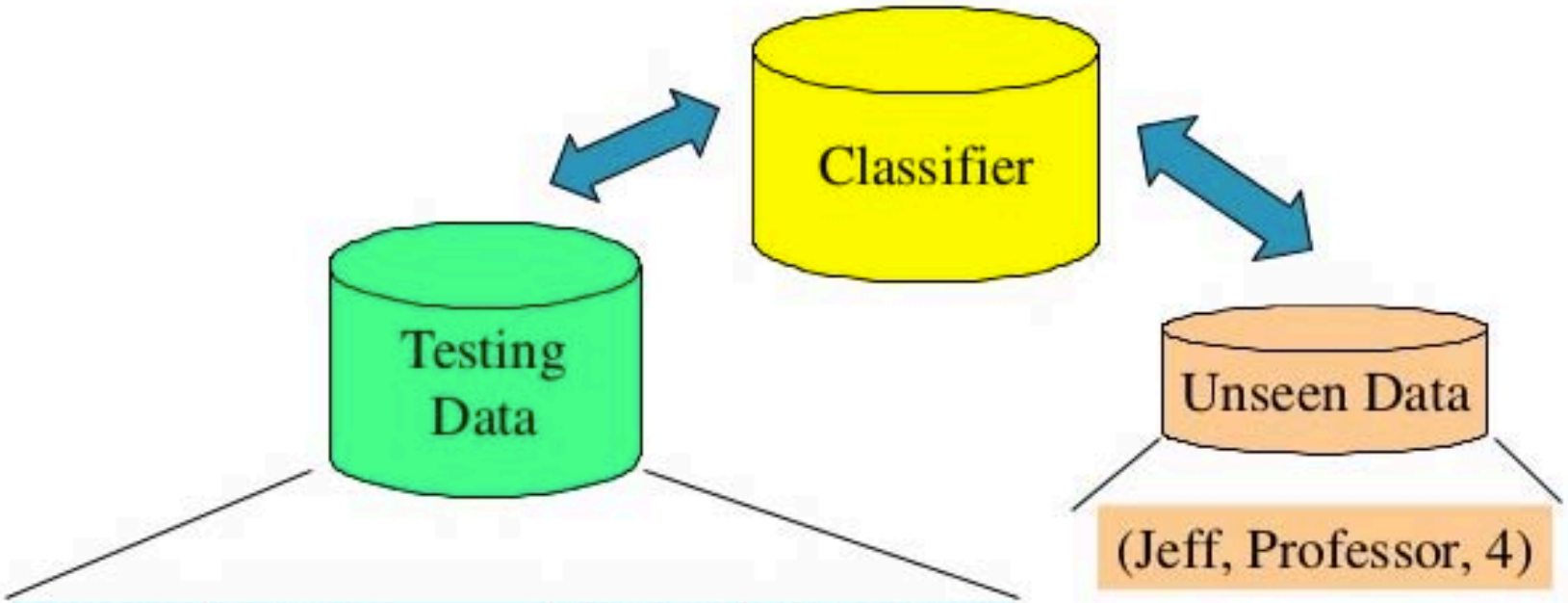
- **The era of big data is here:** until recently we didn't have enough marketing data to confidently predict the future.

  Today, companies and individuals are spewing out massive amounts of information in social networks, on the Web, etc.

- **Technology is effectively harnessing big data:** There have been giant technological advances in computer hardware–faster CPUs, cheaper memory, and massively parallel processing (MPP) architectures. Today, exploring big data and using predictive analytics is within reach of more organizations than ever before.

- **The democratization of the math:** Until recently big data and predictive analytics were almost exclusively the domain of highly skilled data scientists. Today, software makes even the most exotic of techniques within sight–from simple linear and multivariate regression to classification and regression trees (CART), conditional mutual information algorithms, random forests, and neural networks.

¿Cómo obtenemos predicciones?



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Difrentes tipos de clasificadores

- Regresión logística
- Árboles de decisión
- Redes Neuronales
- k-vecinos más cercanos
- etc., etc.

Debemos mirar los tipos de datos que tenemos, potencia de cómputo, capacidad de interpretación del resultado, etc.

Aprendizaje - Entrenamiento

Redes neuronales ←→ **capacidad de aprender**

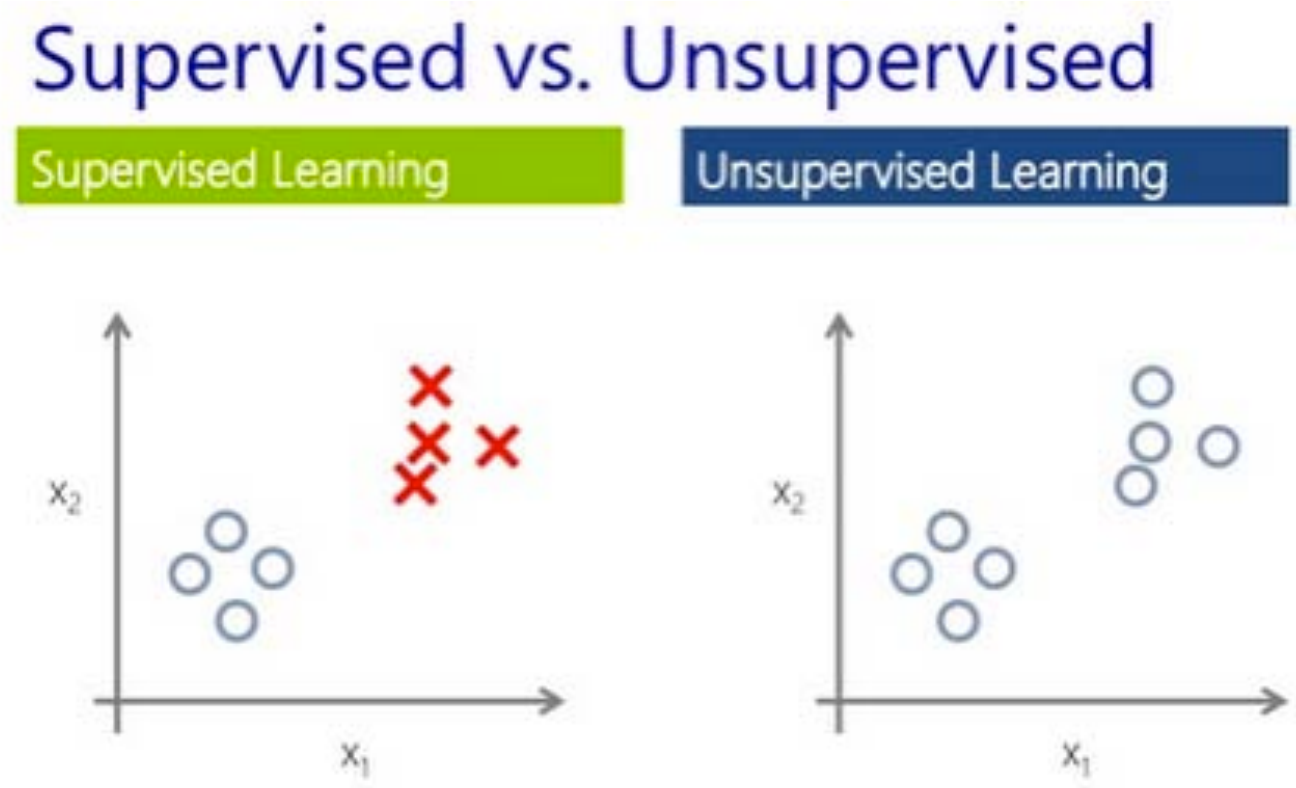
Mediante un **proceso adaptativo** en el cual modificamos los pesos sinápticos o conexiones.

Aprendizaje ⟷ Modificación de pesos sinápticos

Proceso de aprendizaje:

- **Aprendizaje no- supervisado**
- **Aprendizaje por refuerzo**
- **Aprendizaje supervisado**

Aprendizaje supervisado y no -supervisado



Conocemos la clase (etiqueta) a la que pertencen los patrones

No conocemos la clase a la que pertencen los patrones

Redes neuronales artificiales

Modelos simplificados que emulan el funcionamiento del cerebro

Las redes de neuronas artificiales son modelos computacionales **paralelos** que constan de **unidades de proceso** (neuronas) **adaptativas** y **masivamente interconectadas**.

Son, por tanto, procesadores **distribuidos masivamente en paralelo** para almacenar conocimiento experimental y hacerlo disponible para su uso.

¿Qué es una Red Neuronal?

Características importantes de las RNAs:

-- **Aprender a través de ejemplos**

-- **No linealidad** → función de activación → capacidad de cómputo

-- **Adaptabilidad**: Modifican sus pesos sinápticos para adaptarse a los cambios del entorno.

-- Tolerancia a fallos: Por una parte, si una neurona o un enlace de la red neuronal son dañados, la respuesta de la red probablemente no quedará afectada. Una persona nace con unos 100 billones de neuronas y a partir de los cuarenta años, o antes, se pierden más de 40.000 neuronas diarias, que no son reemplazadas y, sin embargo, continuamos aprendiendo.

Cerebro vs. Ordenadores

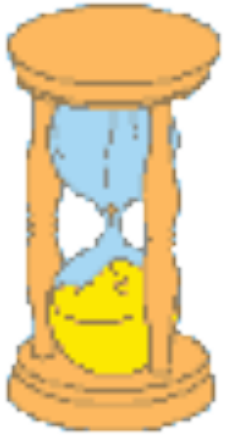
Memoria

Procesamiento de la información

Adaptarse a nuevas tareas

Necesitan energía

Pueden dañarse



Evolucionan

Inteligencia artificial tradicional	Redes Neuronales
Enfoque descendente	Enfoque ascendente
Basado en la psicología	Basado en la biología
Qué hace el cerebro	Cómo lo hace el cerebro
Reglas Si/Entonces	Generalización a partir de ejemplos
Sistemas programados	Sistemas entrenados
Lógica, conceptos, reglas	Reconocimiento de patrones
Arquitectura von Neumann	Arquitecturas paralelas, distribuidas, adaptativas, autoorganización

¿ Para qué problemas son útiles las redes neuronales ?

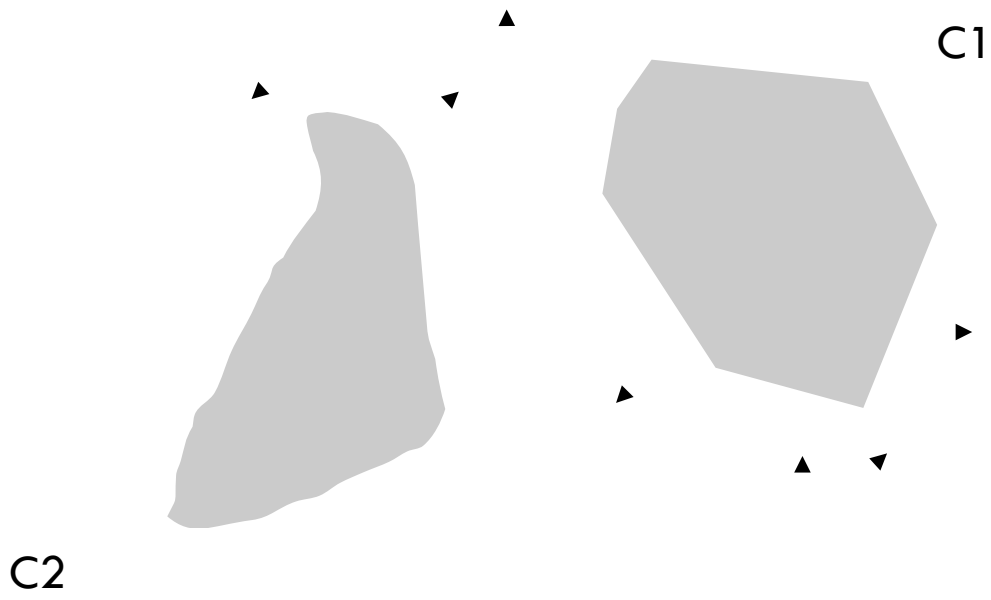
Problemas en los que desconocemos la dinámica interna pero tenemos datos para entrenar el modelo.

- Reconocimiento de patrones
- Diagnóstico médico
- Predicción de bolsa
- Fraude con tarjetas de crédito

Problemas de optimización

Los ordenadores tradicionales son muy efectivos en problemas en los cuales conocemos el algoritmo.

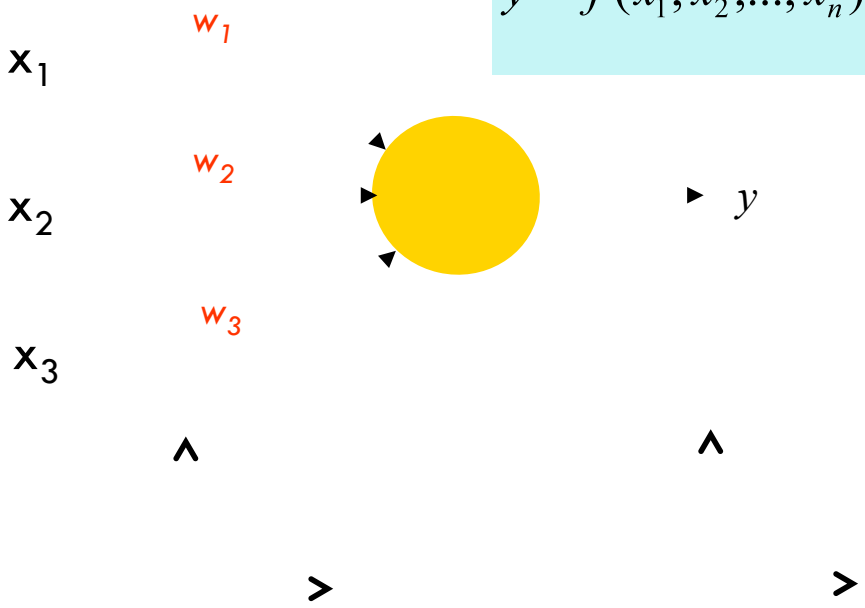
Problemas de clasificación o reconocimiento de patrones en categorías

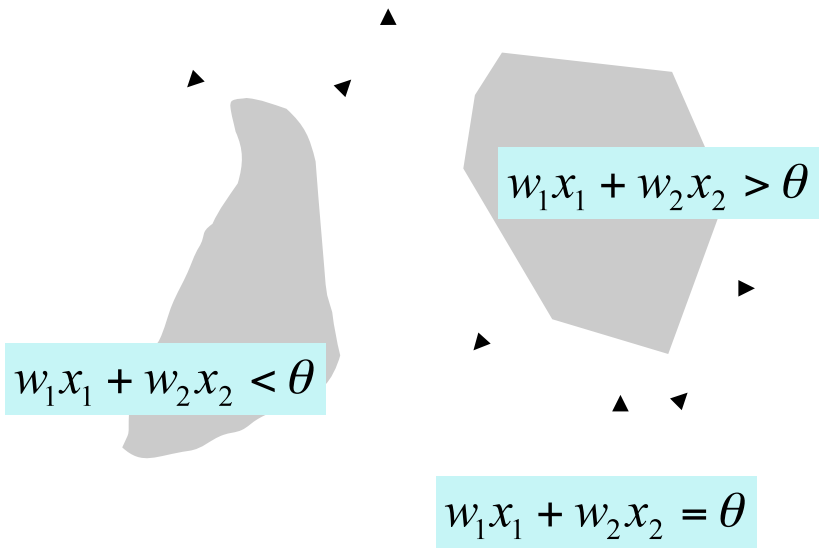


El Perceptrón

1958 - **El psicólogo Frank Ronsenblant desarrolló un modelo simple de neurona basado en el modelo de McCulloch y Pitts que utilizaba una regla de aprendizaje basada en la corrección del error:** **Perceptrón**

$$y = f(x_1, x_2, \ldots, x_n) = \begin{cases} 1 & \text{si} \quad w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \geq \theta \\ -1 & \text{si} \quad w_1 x_1 + w_2 x_2 + \ldots + w_n x_n < \theta \end{cases}$$

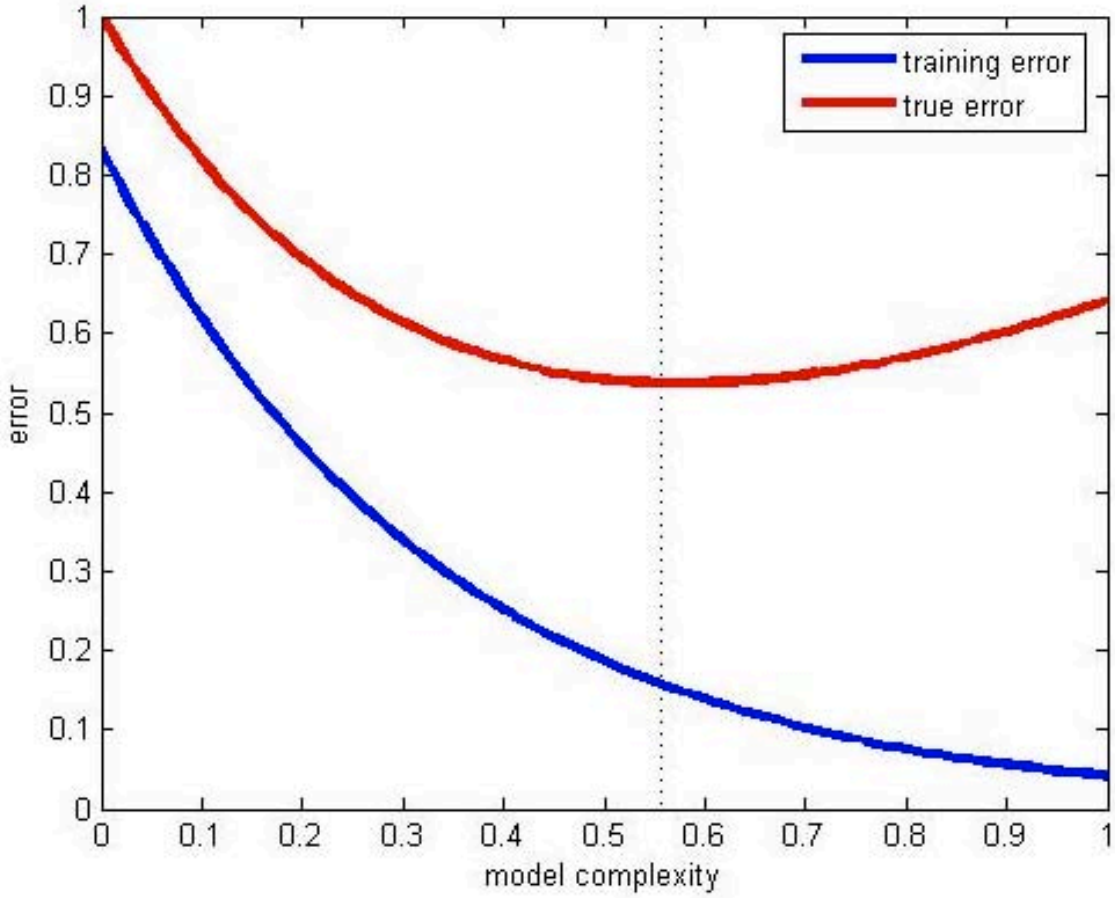




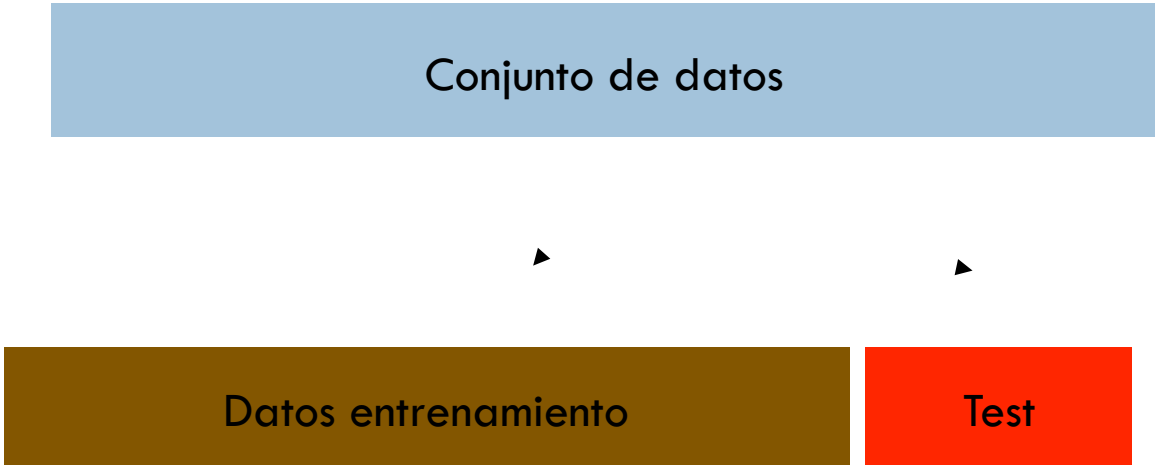
Función paso o *De Heaviside*

Función signo

Un grave peligro: Sobre-entrenamiento



Conjuntos de entrenamiento y Generalización



Dicho método consiste en dividir los datos muestrales en dos partes; una parte se utiliza como **conjunto de entrenamiento** para determinar los parámetros del clasificador neuronal y la otra parte, llamada **conjunto de prueba (ó test ó conjunto de generalización)** se utiliza para estimar el error de generalización o predicción.

Modelos muy evolucionados

- Deep Neural Networks (Google)
- Watson IBM

Modelos muy complejos pero que están demostrando una gran capacidad para tareas complejas y comienzan a usase en Data Analytics

Preguntas para revisar el tema

¿Qué son las redes neuronales ?

Diferencia entre un método de aprendizaje supervisado y no supervisado

¿Qué es el sobre-entrenamiento?

Ejemplos prácticos de utilidad de las predicciones en marketing

Presentaciones

- Nombre: Leonardo Franco
  http://www.lcc.uma.es/~lfranco/

- email: lfranco@lcc.uma.es
- Profesor del Departamento de Lenguajes y Ciencias de la computación en al área de Ciencias de la Computación e Inteligencia Artificial.
- Temas de investigación: Minería de datos, Redes Neuronales, Problemas de predicción, Computación de altas prestaciones (FPGA), Redes de sensores inteligentes, Neurociencia computacional.

Presentaciones de alumnos

- Nombre
- Titulación / Trabajo actual
- Interés en este máster
- Grado de satisfacción, inconvenientes hasta el momento
- Experiencia previa en temas de *Data Mining, Business Intelligence*
- Experiencia con Lenguajes de programación, softwares utilizados

Objetivos de la Asignatura

Al final de este módulo deberíais de ser capaces de:

- Aplicar técnicas de minería de datos a problemas reales en el ámbito de Marketing.
- Entender el fenómeno de Big Data y sus implicaciones en la problématica actual de las empresas.

Programación de la Asignatura

Las actividades diarias estarán divididas en dos temáticas:

19:!5

21:30

Tópicos

BIG DATA

- Las 4 V's
- SMEC
- Casos de estudio
- Aplicaciones

Métodos de Data Mining

- Pre-procesamiento
- Modelado
- Visualización

Tratando de hacer las clases lo más **participativas** posibles, con trabajos en grupo, discusiones, búsqueda de información, tests on-line, etc.

Evaluación de la asignatura

NOTA FINAL 100 %

Evaluación Continua 30 %

- Participación en clase
- Cuaderno "on-line".
- Presentaciones de temas.

Examen* 30 %

- Test de preguntas sobre lo visto en el curso (Data Mining y Big Data).

Informe a entregar 40 %

- Aplicación de métodos de "data mining" a un caso práctico.

* 2 posibilidades: Al final del módulo (1-2 de Junio) ó en el examen del día 30 de Junio

Cronograma Primera Semana

Martes19	Miércoles 20	Jueves 21
• Presentación de la asignatura Big Data • Introducción • Cuaderno "on-line"	Big Data • Breve historia • Casos de estudio • Presentación actividad "Debate"	Big Data • Caso de estudio: •Presentación Airlines sentiment analysis • Tarea aplicación a casos de estudio
Minería de datos • Introducción • Pre-procesamiento de datos	Minería de datos • Software Weka • Ejercicios	Minería de datos • Modelos de predicción

Cronograma Segunda Semana

Martes 26	Miércoles 27	Jueves 28
Big Data • Presentación breve sobre casos de estudio	Minería de datos • Análisis de resultados y visualización	Big Data • Presentaciones y debate
Minería de datos • Aplicaciones de modelos de predicción	Minería de datos • Ejercicios	Minería de datos • Propuestas de temas para “Trabajos a entregar”

Día adicional

- Test “on-line” sobre lo visto en el curso:

  -- Big Data

  -- Data Mining

Preguntas de opción múltiple sobre los temas analizados.

El mismo tipo de test se volverá a realizar en el examen del día 30 de Junio.

MÁSTER EN DIRECCIÓN Y GESTIÓN DE MARKETING DIGITAL

NUEVAS TECNOLOGÍAS Y MARKETING

BIG DATA

2015

Profesor: Dr. LEONARDO FRANCO

BIG DATA

- **Grandes volúmenes de datos complejos y dinámicos que no pueden almacenarse y tratarse con los métodos tradicionales**

El 90% de todos los datos existentes fue creado en los últimos dos años y este total se duplicará cada dos años.

En 2020 tendremos 50 veces más datos que los que hubo en 2011.

BIG DATA

- **Origen de los datos**

Internet : Transacciones, reservas de viaje, comentarios, periódicos, redes sociales.

Cámaras web: Videovigilancia, móviles.

Sensores: Wearables, estaciones de control meteorológico, etc.

BIG DATA - Desafíos

- Así como Henry Ford cambió la forma de fabricar coches transformando al mismo tiempo la forma de trabajo, Big Data emerge como un nuevo paradigma de conocimiento que puede cambiar la forma en la que entendemos las relaciones sociales.

'Change the instruments, and you will change the entire social theory that goes with them` (Latour 2009)

¿Es el BIG DATA un fenómeno real o una moda ?

BIG DATA - Desafíos

- **¿ Será una herramienta útil al alcance de todas las empresas?**

ó

- **¿Llevará a una mayor concentración, en la cual sólo salen beneficiadas las empresas más grandes dueñas de los datos?**

BIG DATA - Desafíos

- **¿Tener acceso a tantos datos ayudará a tener mejores herramientas, servicios y bienes públicos ?**

ó

- **¿Llevará a una nueva ola de invasión de la privacidad con ofertas invasivas de marketing ?**

BIG DATA - Desafíos

- **¿ Nos ayudará a entender las comunidades digitales y los nuevos movimientos sociales ?**

ó

- **¿Será una herramienta para monitorizar los grupos políticos y suprimir la libertad de expresión?**

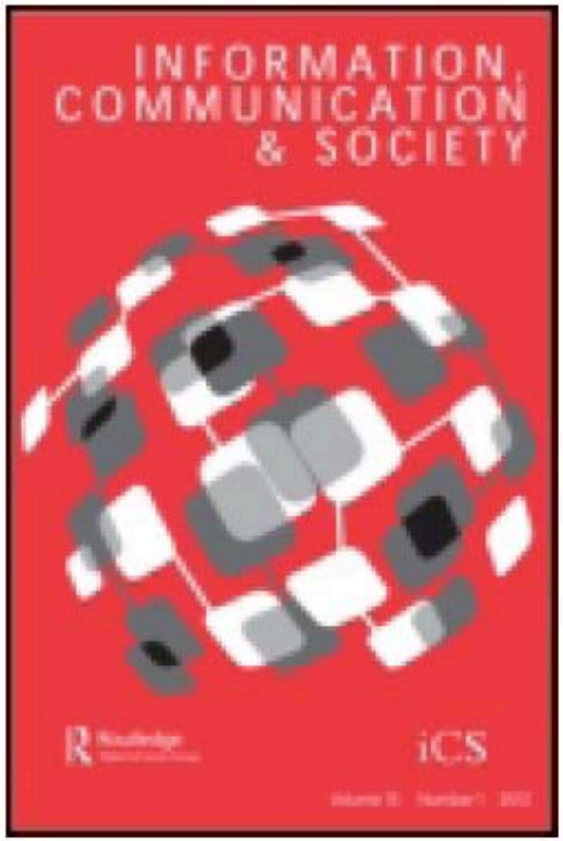
BIG DATA - Desafíos

- **¿ Modificará la forma de estudiar la comunicación y conductas humanas?**

ó

- **¿Estrechará la paleta de opciones de investigación, alterando lo que significa "investigar"?**

Critical questions for Big Data



Information, Communication & Society

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rics20

CRITICAL QUESTIONS FOR BIG DATA

danah boyd [a] & Kate Crawford [b]

[a] Microsoft Research , One Memorial Drive , Cambridge , MA , 02142 , USA

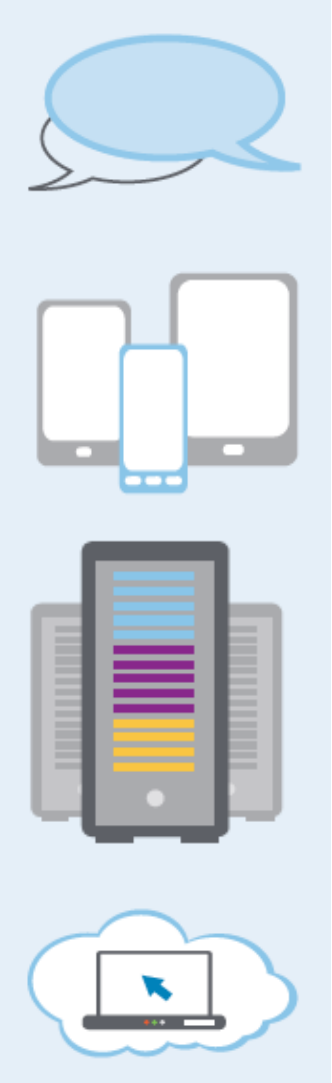
[b] Microsoft Research , One Memorial Drive , Cambridge , MA , 02142 , USA E-mail:

Published online: 10 May 2012.

Essential Reading: ***BD-R1-critical-questions.pdf***

Nuevas Tecnologías para los negocios del futuro presente

S



Social Networks

M

Mobility

A

Analytics

C

Cloud Computing

Pre-digital, digital and post-digital eras in Marketing

	Pre-digital era	Digital era	Post-digital era
Organization structure	Highly centralized	Semi-decentralized	Highly decentralized
Decision making process	Management driven	Management driven, with customer input	In collaboration with the customer
Customers	Lack information	More informed	Highly informed
Role of technology	Support function	Strategic function	Business enabler
Disruptive technologies	PC's	Internet, mobility, social media	Artificial intelligence, Robotics, Augmented reality, cloud computing, BYOD

Referencia Esencial
BD-R2-TheSmacCode.pdf

Era Pre-digital

- IT housed in small, internal teams within enterprises
- IT is seen mainly as a support function
- Computing power is scare, and extremely expensive

- In the absence of information, innovation is a top-down process, based more of what the management believes, rather than what the customer is looking for
- Innovation becomes a long drawn out process, with little or no feedback during the innovation lifecycle

- PC's emerged as the primary disruptive technology of this era. It reduced the cost of computing drastically, and made it affordable for enterprises
- The emergence of digital devices across the spectrum laid down the way for the digital era to follow

Era Digital

- IT becomes a strategic function in leading enterprises
- Outsourcing and off-shoring become the buzzwords; the era sees the rise of large IT-BPO players like IBM, Accenture and TCS

- The availability of timely information across the value chain leads to a collaborative innovation process
- The pace of innovation also increases, as enterprises are more rapidly able to respond to changes in the marketplace

- The emergence of the internet was the foundation upon which the digital era was based
- The spread of mobile communication networks, as well as affordable computing power added to the digital age

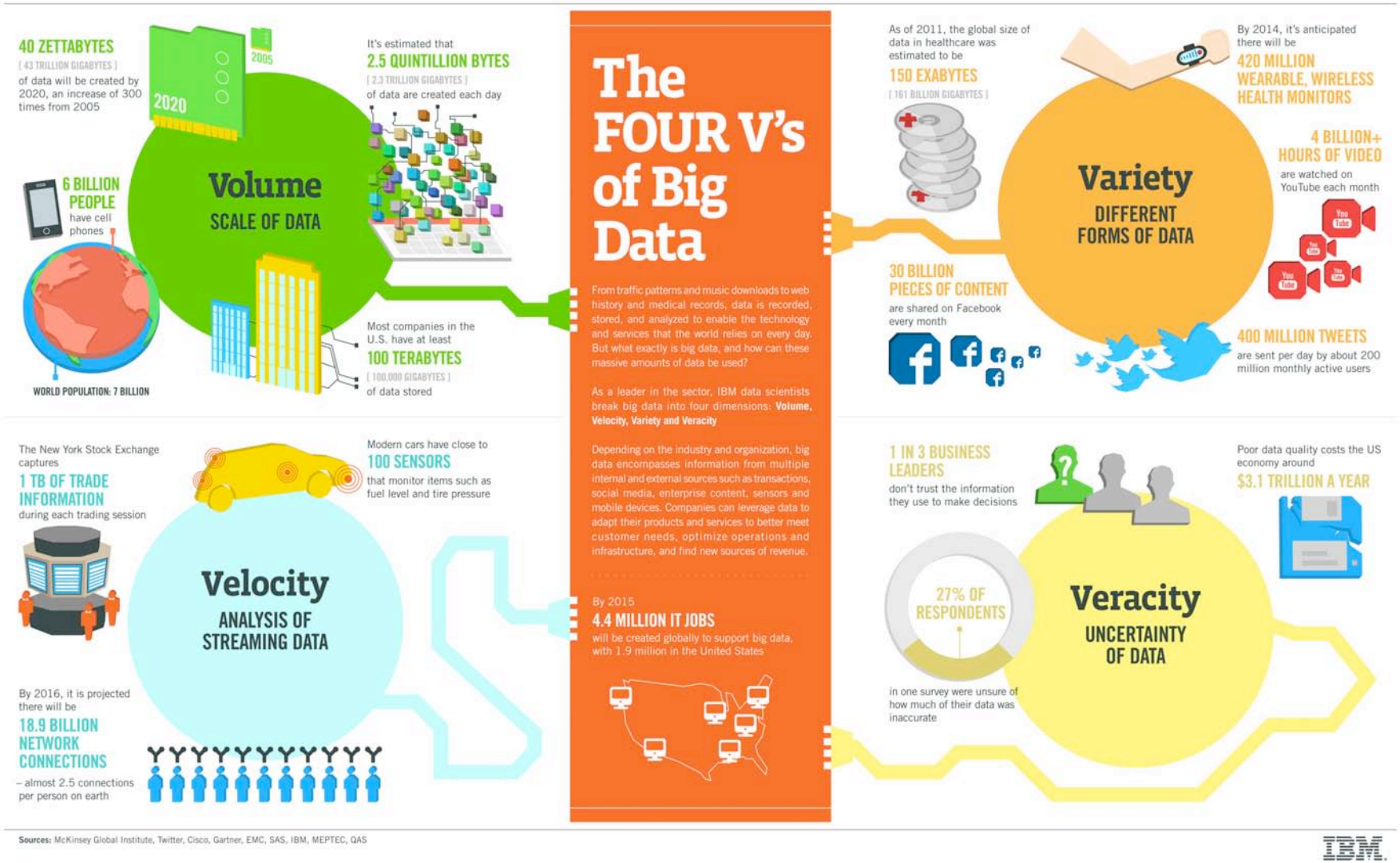
Era Post-digital

- IT becomes a business enabler - IT decision making shifts from the CIO to the CFO/CMO
- Enterprises can use IT to emerge as leaders in the marketplace, Business models, operational processes, and IT ecosystems will change to reflect the new ground realities

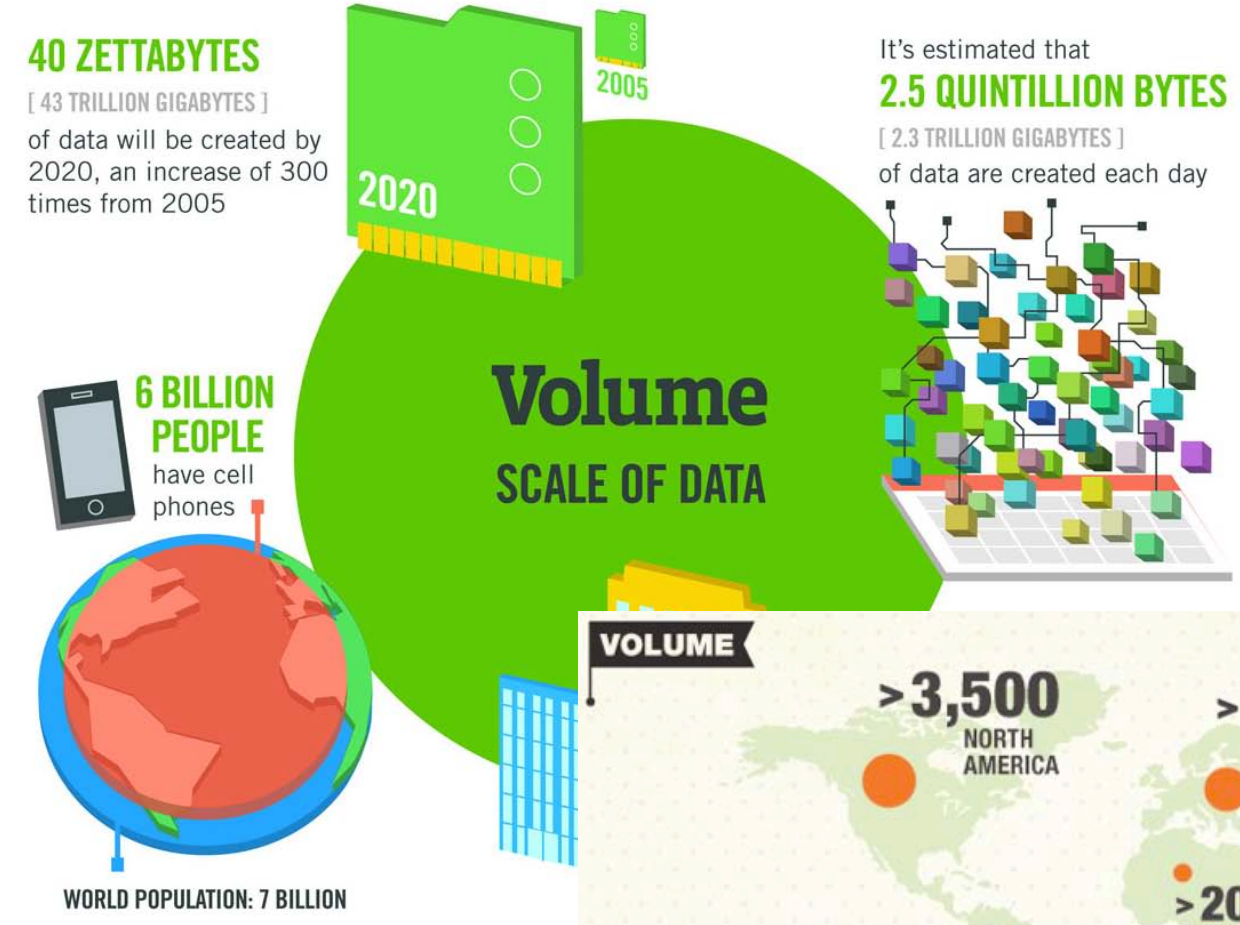
- Innovation will become a bottom-up process, with customer input driving innovation
- The convergence of social media and analytics can lead to significant reduction in the innovation process

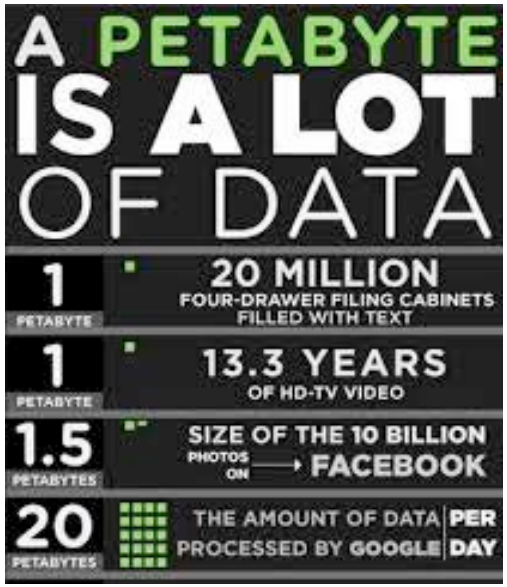
- The convergence of social media, mobility, analytics and cloud (SMAC) will drive the next generation of innovations
- Moreover, the increasing spread of these technologies to people across the world is bridging the 'digital divide'

Las 4 V's del Big Data



Volumen



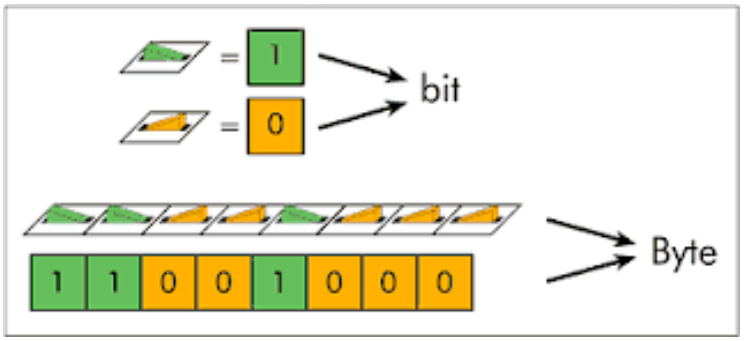


~ exabytes

100 PB

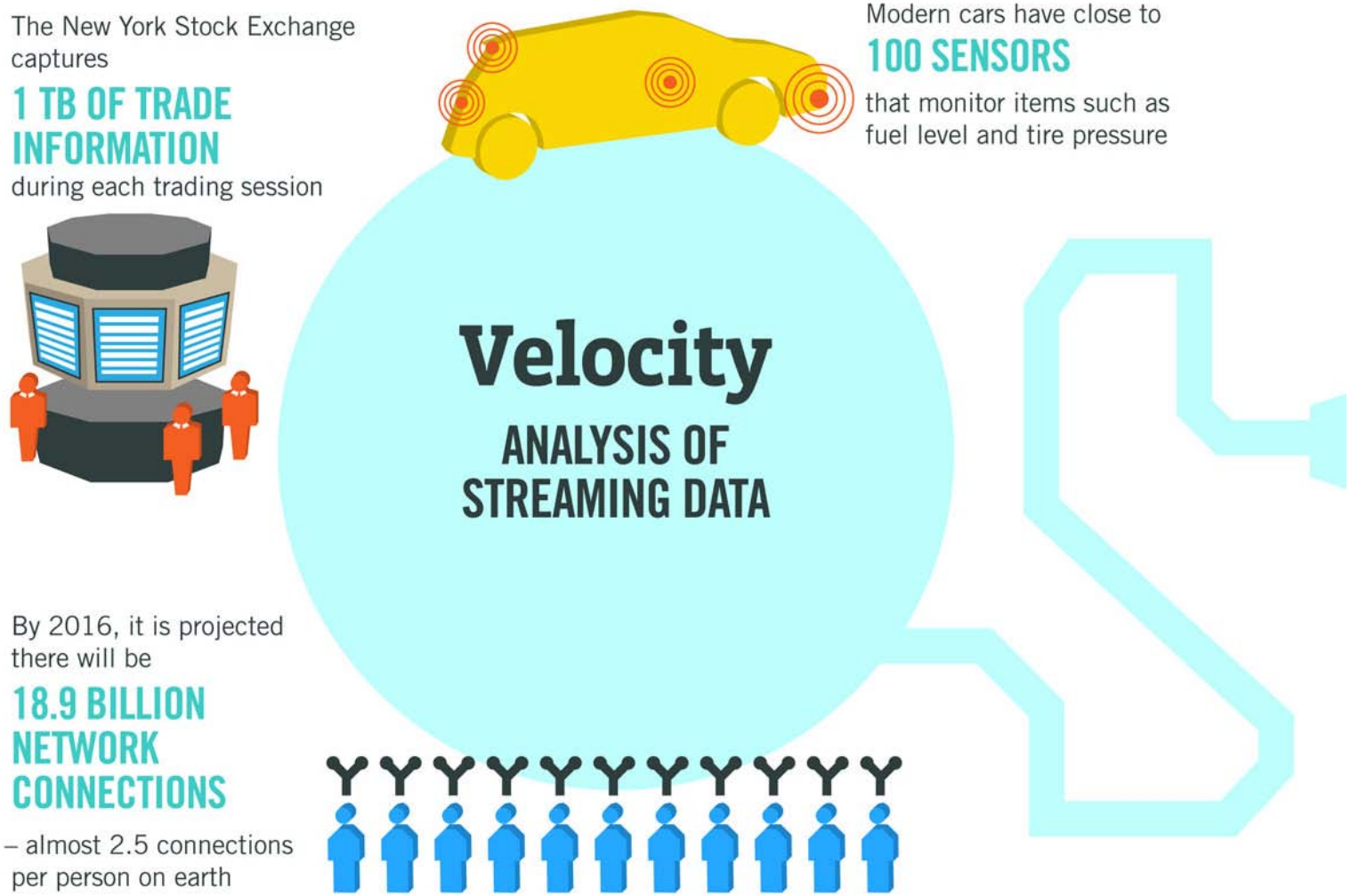
Unidades de almacenamiento de datos

1 Byte = 8 bits



	Unidad		Equivalencia
2^{10}	1 KB	1024 B	Imagen
2^{20}	1 MB	1024 KB	Música
2^{30}	1 GB	1024M B	Video
2^{40}	1 TB	1024 GB	Datos personales
2^{50}	1 PB	1024 TB	Amazon
2^{60}	1 EB	1024 PB	Google
2^{70}	1 ZB	1024 EB	The World

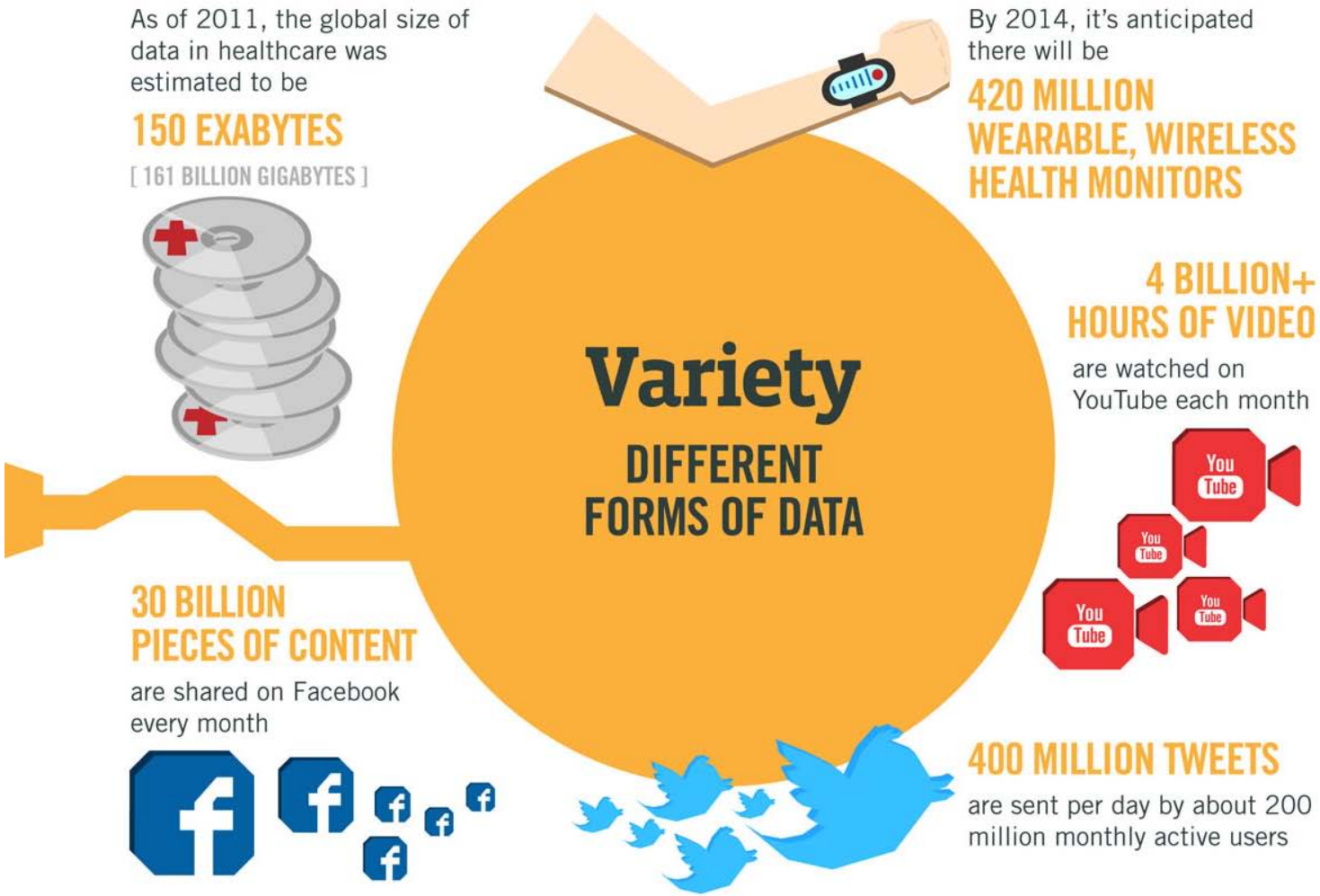
Velocidad



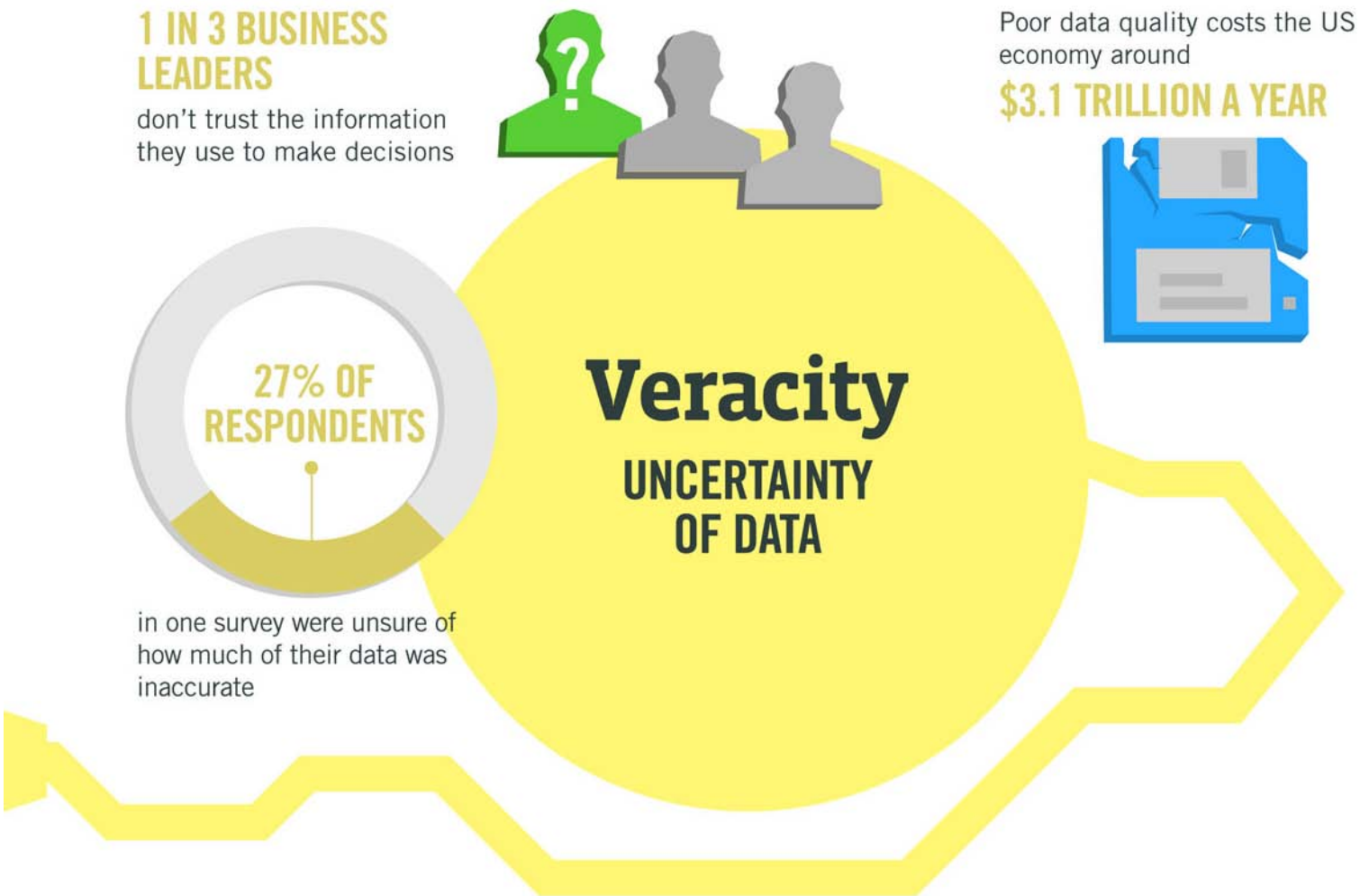
Alibaba vendió 2.000 millones de dólares en una hora (Nov 2014)

Variedad

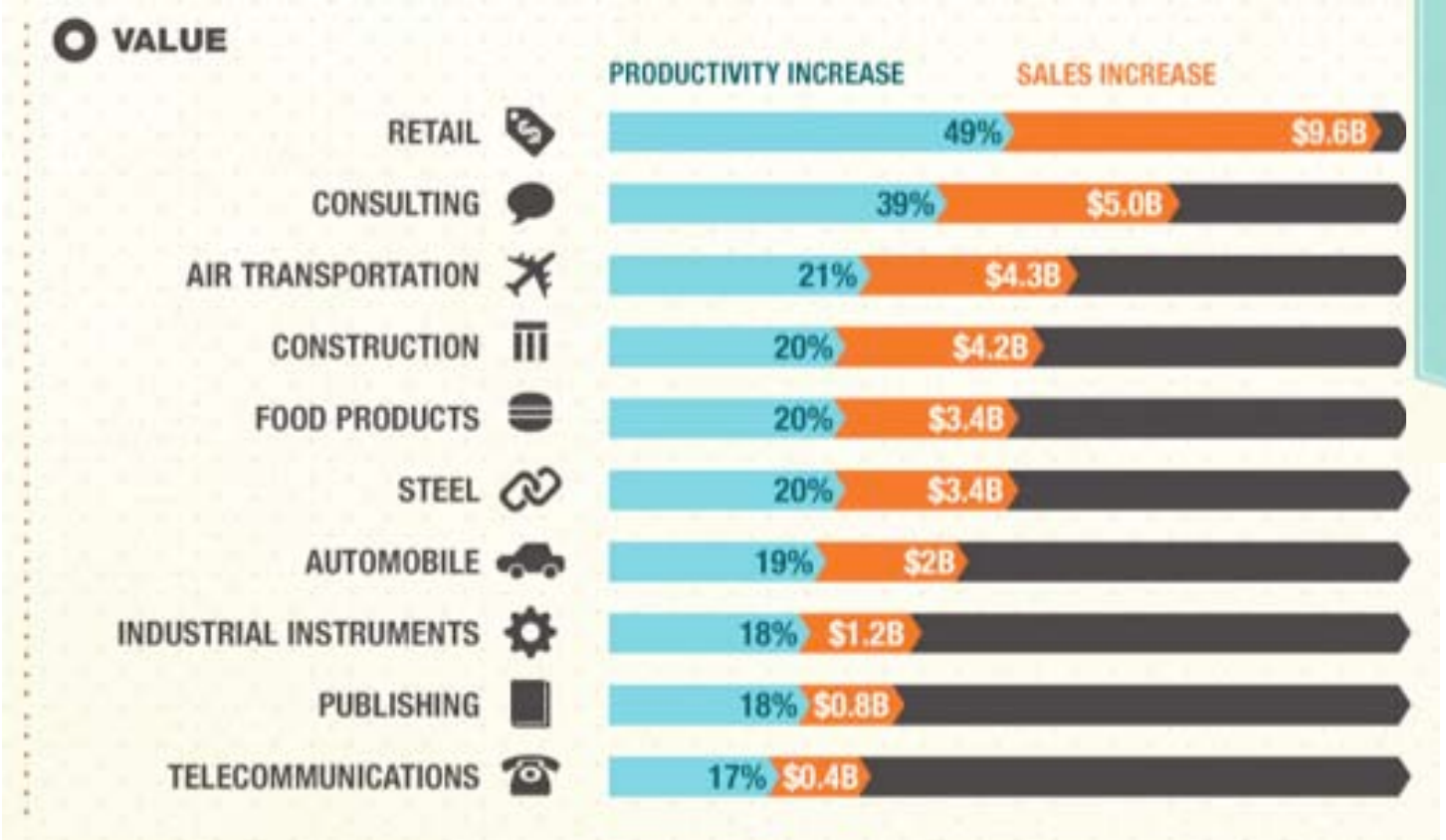
(Heterogeneidad de los datos)

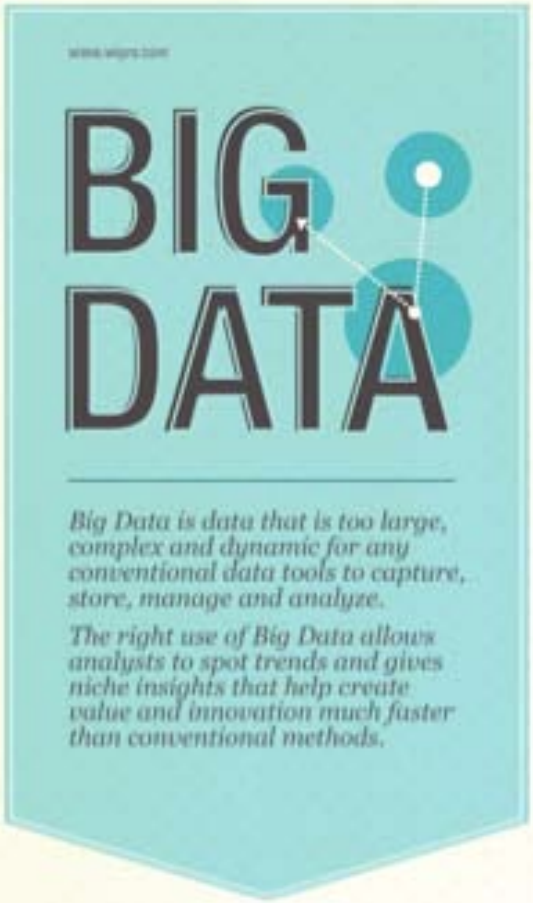


Veracidad


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Más cifras del Big Data





Aumento en productividad y ventas previstos por el efecto de Big Data

Big Data y Marketing Digital

El fenómeno de Big Data constituye un nuevo horizonte para el marketing digital.

Es un nuevo reto, una nueva herramienta, que ofrece un potencial enorme, y que habrá que saber explotar de la mejor forma posible.
Son dos ámbitos que estan condenados a entenderse y a trabajar juntos.

Pasos fundamentales para la explotación del Big Data en Marketing

- Escuchar y conversar en medios sociales (CRMs u otra fuente).
- Establecer mecanismos apropiados de recolección de datos, analizando qué mensaje es el más apropiado para cada tipo de cliente.
- Agrupar los diferentes mensajes por cada tipo de cliente
- Enviar el mensaje a los clientes en función del dispositivo en el que los recibirá

¿ Qué implica el BIG Data para el Marketing Digital ?

Otros aspectos relevantes el futuro del Big Data y el marketing digital son: **una mejor experiencia de usuario unida a una mayor personalización de los contenidos,** lo que conlleva inevitablemente a un **aumento del grado de relevancia y de valor de los mensajes.**

Entonces podemos ir extrayendo las siguientes conclusiones:

el Big Data es precisamente eso, datos: **Datos en bruto,** que habrá que convertir en información, para posteriormente, **adquirir conocimiento.**

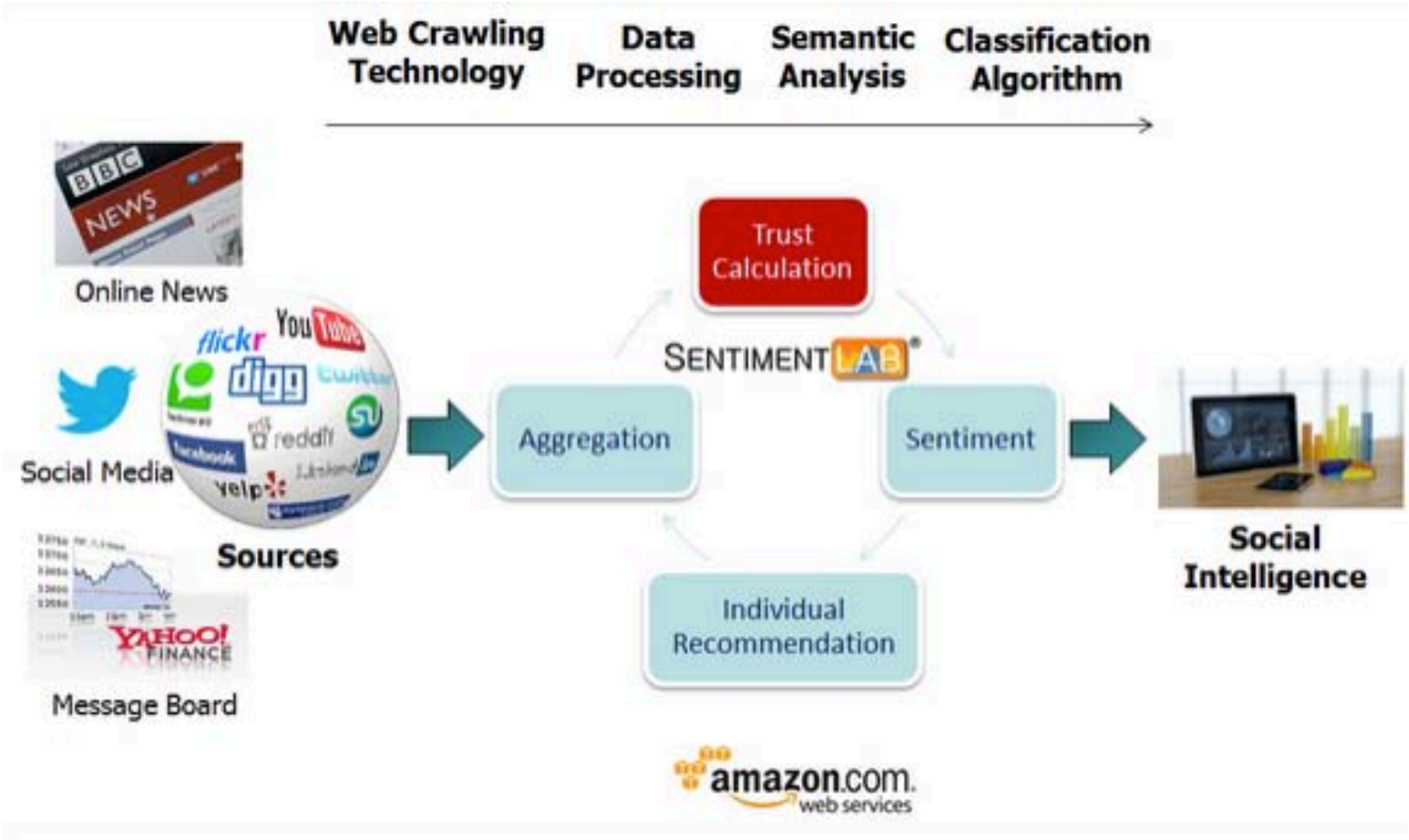
el Big Data proporciona un **marketing en tiempo real, y añade la inmediatez como elemento clave** y ampliamente demandado por los usuarios y consumidores finales.

Lectura adicional: [RA-BD4] García Martínez (2013)
"El Big Data es ahora el gran reto ..."

Áreas de Aplicación de Big Data

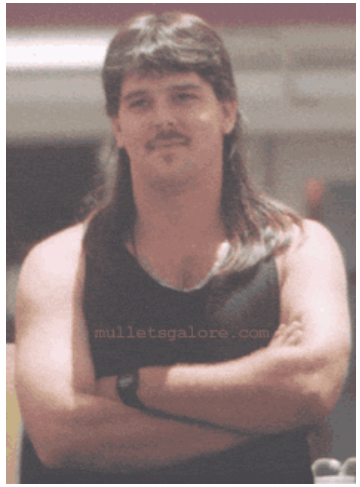
- Sentiment analysis (imagen de políticos, mercados financieros, empresas)
- Recomendaciones on-line: de compra (Amazon, ebay, Netflix, supermercados), de amistades (FB, Linkedin), publicidad dirigida.
- Optimización de costes (”General Electric”, “Smart cities”).
- Predicciones (Potenciales clientes, Bolsa, etc.)

Métodos involucrados en sentiment analysis



http://www.h2oconsulting.ch/sentiment-home/

Métodos involucrados en sistemas de recomendación



- **Cliente X**
  - Compra un CD de Metallica
  - Compra un Cd de Megadeth
- **ClienteY**
  - Busca sobre Metallica
  - El sistema le recomienda Megadeth fde acuerdo a los datos del usuario X

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

¿Qué procesos involucra BIG Data?

Tareas	Tecnologías
• Almacenar y acceder a grandes volúmenes de datos no-estructurados	Manejo de base de datos
• Procesar grandes volúmenes de datos on-line (data streams)	Modelos de programación paralela y distribuida
• **Dar utilidad a los datos**	**Técnicas de minería de datos**
• **Tecnologías para hacer predicciones**	**Machine learning**

Tarea grupal sobre tópicos del Big Data

- Tipo: Presentación oral
- Fecha: Jueves 28/5
- Detalles: preparar una presentación grupal sobre el tema asignado para una duración aproximada de 10 minutos (entre 7 y 12 diapositivas).
- Se tendrá en cuenta para la evaluación:

a) Presentación visual: diseño de las diapositivas, estructura de la presentación siguiendo un hilo argumental.

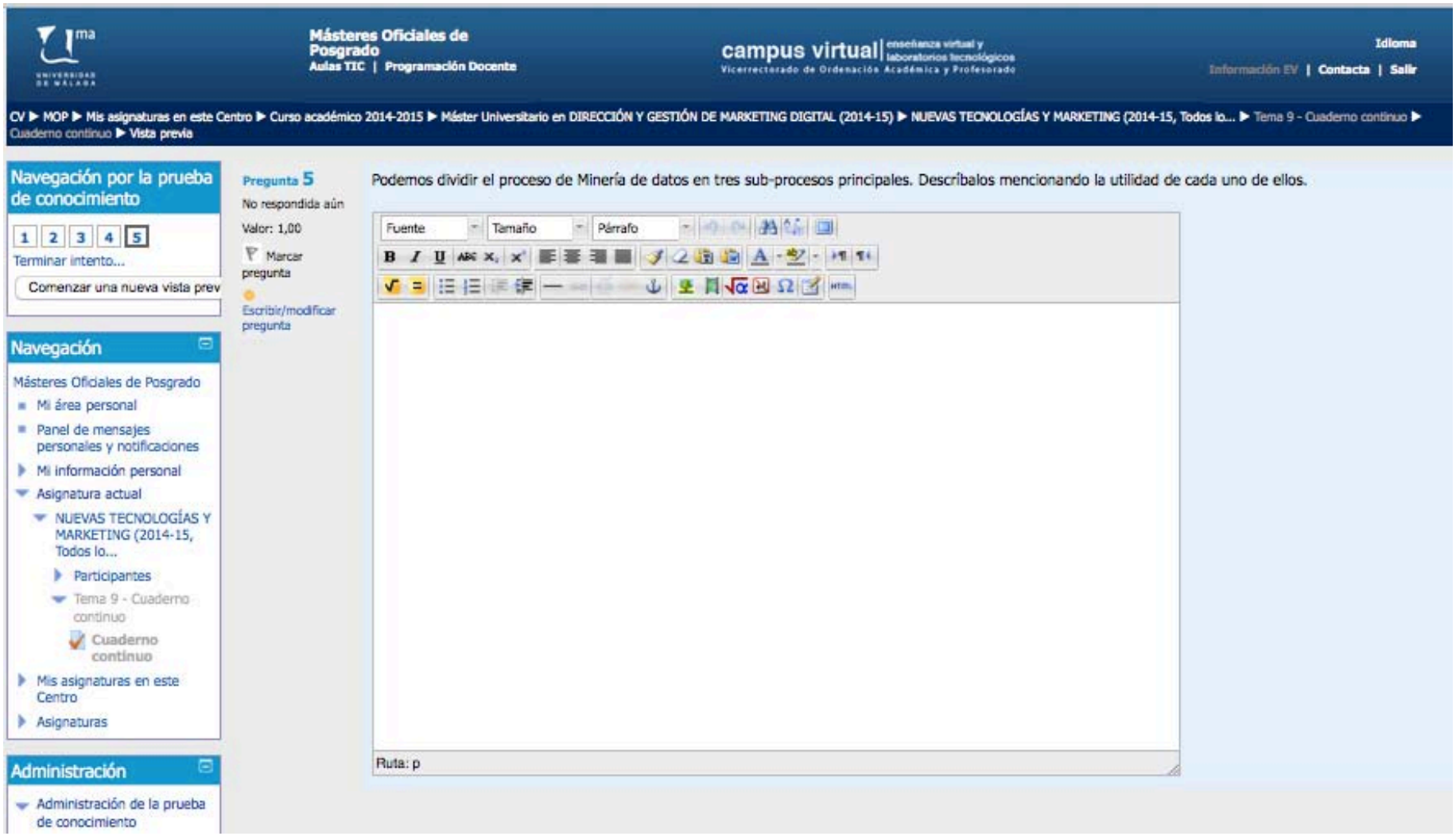
b) Presentación oral: contenido al margen de lo mostrado visualmente, y forma de expresión.

c) Contenido global de la presentación: Si el material presentado es interesante y relevante al tema, y si se utilizan y citan diversas fuentes.

Tarea individual sobre casos de Big Data y Data Mining

- Tipo: Presentación oral
- Fecha: Martes 26/5
- Detalles: muy breve presentación oral individual sobre el caso elegido. Duración aproximada de 3 minutos.
- Buscar información relevante e interesante sobre los métodos y tecnologías novedosas utilizadas en el caso en concreto y en los resultados obtenidos.

Cuaderno de la asignatura



Sitios web de interés

http://www.educacionline.com/instituto-de-marketing-online/

http://www.puromarketing.com/

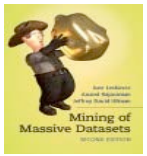
http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Diference between big data and data mining: http://stackoverflow.com/questions/22419958/what-is-the-difference-between-big-data-and-data-mining

http://blog.data-miners.com/ G. Linnoff blog

Scaling big data – Twitter
http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-02-Lin.pdf

Facebook de la Asignatura con links a noticias sobre Big Data
Nuevas Tecnologías https://www.facebook.com/profile.php?id=100009746673220



Libro: Mining Massive Data Sets
http://infolab.stanford.edu/~ullman/mmds/book.pdf

Más referencias

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

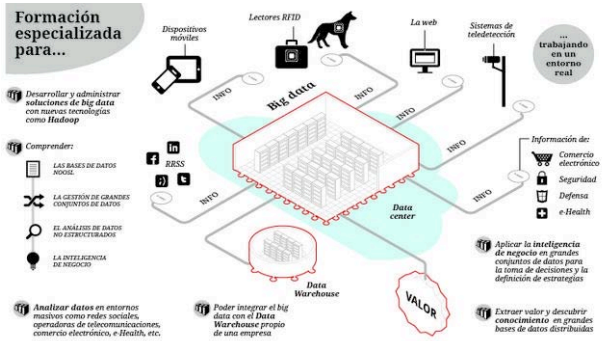
Las claves del sistema de recomendación de Netflix

http://www.bigdatanews.com/profiles/blogs/fast-clustering-algorithms-for-massive-datasets

http://www.ibermatica.com/sala-de-prensa/opinion/big-data-el-termino-de-moda-en-el-mundo-de-la-informatica

Big Data día 2

- Resumen visual del BIG DATA



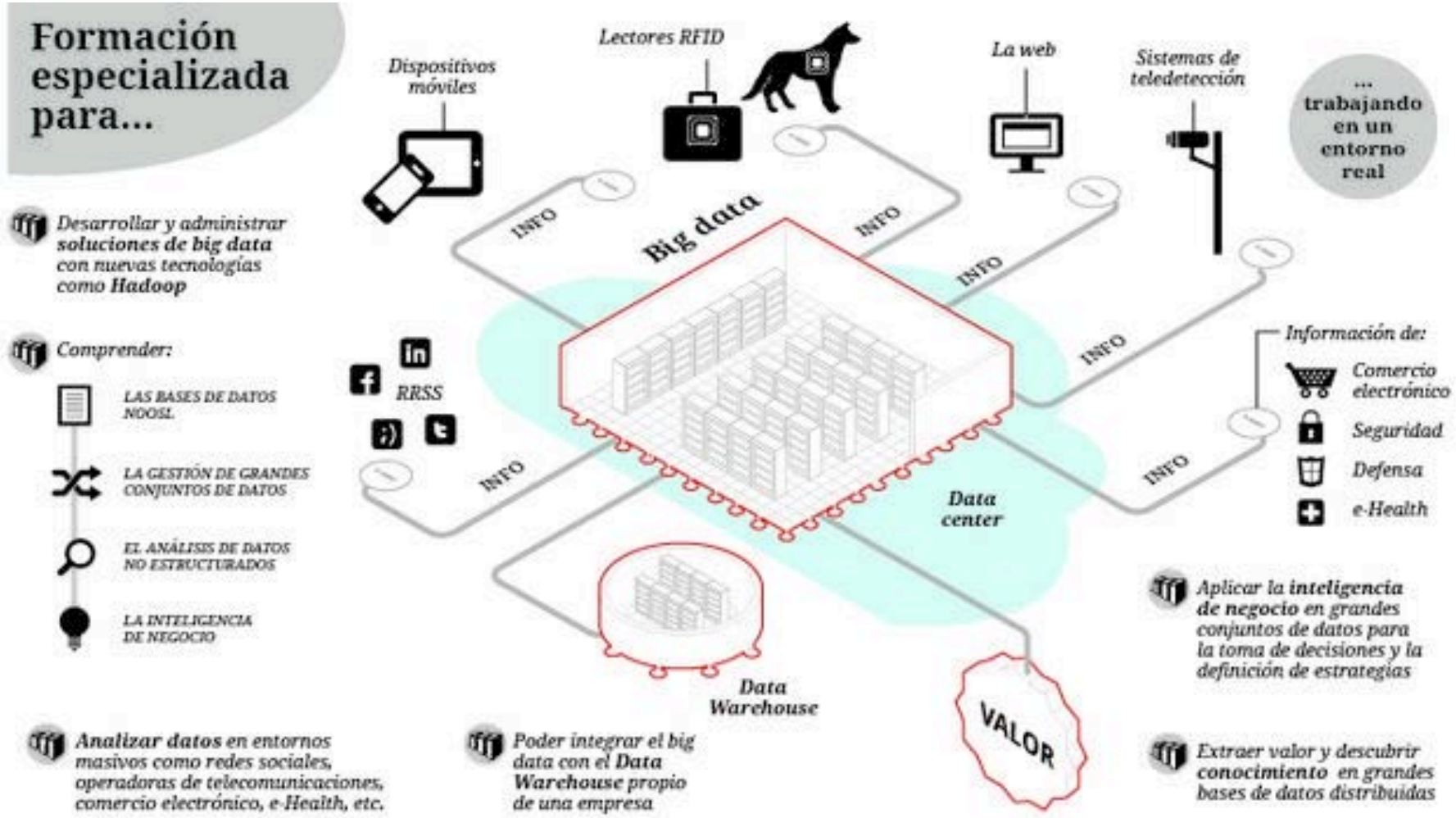
- 3 promesas y riesgos del Big Data

Advanced customer analytics can be a powerful business tool, but companies need to avoid common pitfalls before investing.

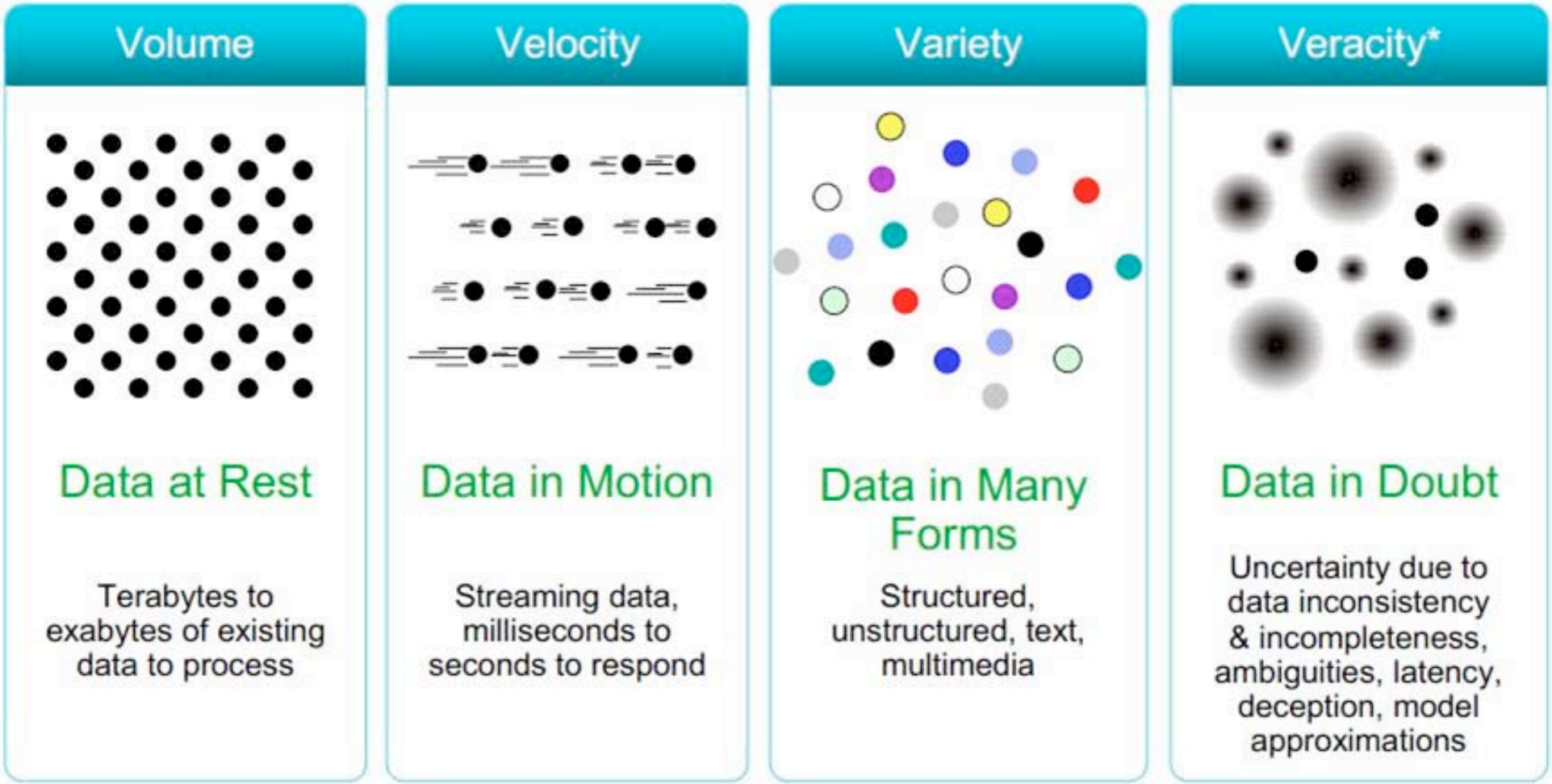
By Eric Almquist, John Senior and Tom Springer

- Temas Presentaciones individuales Martes 26/5
- Temas Presentaciones grupales Jueves 28/5

Infografía del BIG DATA



4 V's del Big Data



The Analytics Big Bang

Predictive analytics reaches critical mass as Big Data and new technologies collide

Key Innovations

New Users

+ Monte Carlo simulations
+ Computational models for neural networks
+ Linear programming

Government Agencies

+ Non-linear programming
+ Computer-based heuristic problem solving

Corporations & Research Institutions

+ Real-time analytics
+ Prescriptive analytics

Mid-size Businesses & Tech Startups

+ R version 1.0
+ Standardization of natural language processing
+ Apache Hadoop

Small Businesses & Analytic Experts

+ Analytic exchanges that enable global collaboration
+ Anticipatory analytics

Anyone

Analytic Innovation Accelerates
2000–2009: Production version of R language for analytic software grows from 0 to 1,000,000 users[1]

Buy! Buy! Buy!
2000–2012: Analytic software market grows from $11 billion to $35 billion[2]

21st Century's Sexiest Job[3]
2011–2012: Data scientist job posts jump 15,000%[4]

Hyper-connectivity
2012: 1.7 billion mobile devices sold and 2+ billion people on social networks add to data explosion

Collaboration leads to massive innovation

Personalization of every event, experience, offer

Solving the Unsolvable in medicine, energy, agriculture and more

Analytic Talent Gap grows as demand skyrockets

Affordable and Accessible analytics as tools become broadly adopted

Milestones

1930s–40s

Dawn of Computer Age

1940s: Turing and Good conduct groundbreaking work with "weights of evidence" to decode German messages in WWII

1940: Kerrison Predictor automates targeting of anti-aircraft weapons against enemy planes

1944: Manhattan Project team runs computer simulations to predict behavior of nuclear chain reactions

1950s–1960s

Commercialization of Analytics

1950: ENIAC computer generates first models to forecast weather

1951: First university degree program in Operations Research (Case Institute of Technology)

1956: Analytics solves "shortest path problem," improving air travel and logistics

1958: FICO applies predictive modeling to credit risk decisions

1966: Future SAS Institute starts as research project funded by US Department of Agriculture

1970s–1990s

Analytics Goes Mainstream

1973: Black-Scholes model created to predict optimal price for stock options over time

1980: First commercial tool for building model-driven Decision Support Systems is marketed

1992: FICO deploys real-time analytics to fight credit card fraud

1995: Amazon and eBay go live; race to personalize online experience is on

1998: Google applies algorithms to web searches to maximize results relevance

1998: Moneyball changes pro sports as Oakland A's use analytics for a competitive edge

2000–Present

Analytics' Deep Impact

- **Widespread analytics use:** dynamic ticket pricing, shopping and movie recommendations, traffic management and much more
- **Natural language processing:** unlocks analytic value of unstructured data (e.g., Facebook posts, web pages, PDFs, email, Word docs)
- **Big Data arrives:** 2.5 quintillion bytes of data created each day[5]
- **Server farms and low-cost, high-speed processing:** make distributed computing and Big Data analytics viable for most organizations
- **Growing demand for talent:** 190,000 more analytics experts and 1.5 million more data-literate managers needed in US alone by 2018[6]

Next

Ubiquitous Analytics

- **Cloud-based analytic exchanges:** lead to Collaboration Economy between developers, businesses, researchers, scientists and entrepreneurs
- **Individuals use analytics in everyday decisions:** about education, careers, finances, healthcare, peer-to-peer renting and lending ("share economy")
- **Curing rare diseases:** becomes a financial winner
- **Predictive policing:** preempts many crimes
- **Anticipatory analytics:** makes it nearly impossible to crash a car or burn dinner
- **Mass marketing campaigns are dead:** all customer interaction is personal
- **Data licensing trumps data purchasing:** in a world where data gets stale in minutes

FICO

www.fico.com/analytics

1 http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/
2 IDC Market Analysis: Worldwide Business Analytics Software
3 http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
4 http://www.indeed.com/jobtrends?q=%22Data+Scientist%22&l=1&relative=1
5 http://marciaconner.com/blog/data-on-big-data/
6 http://spotfire.tibco.com/blog/?p=6886

© 2013 Fair Isaac Corporation. All rights reserved.

Promesas y riesgos del BIG DATA

Advanced customer analytics can be a powerful business tool, but companies need to avoid common pitfalls before investing.

By Eric Almquist, John Senior and Tom Springer

Ser realista

Distinguir el potencial real del BIG Data de las promesas extravagantes

- Big Data por sí misma no identificará nuevos negocios

Feel like you've seen this movie before? If you were caught up in the initial euphoria of the customer relationship management (CRM) revolution, then you did. Starting in the early 1990s, many companies bought into the hype and the technology, only to wind up with unusable databases, rebellious sales teams and depleted capital budgets.

The CRM industry has since matured, and there is no doubt that CRM solutions can now deliver real value to many organizations. As evidence, CRM was the sixth most popular business tool in Bain's 2015 Management Tools & Trends survey. And global CRM spending totaled $20.4 billion in 2014, up from $18 billion the previous year, according to Gartner research.

Yet CRM failure rates remain high. A 2014 report from C5 Insight found that more than 30% of all CRM implementations fail—and second and third CRM implementations at the same companies had only slightly lower failure rates. And this is 20 years into the "revolution"!

CRM history

We see Big Data going down a similar path, making big promises about customer impact and value creation predicated on large investments in technology and expertise. In a recent report, Gartner predicted that "through 2017, 60% of Big Data projects will fail to go beyond piloting and experimentation and will be abandoned." Why is history repeating itself? It's not for lack of interest, effort or investment. Instead, it reflects the difficulty of generating value from existing customer, operational and service data—let alone the reams of unstructured, internal and external data generated from social media, mobile devices and online activity.

Leading users of Big Data set a high bar for success. They have assembled deep benches of analytical talent and created processes that allow their organizations to glean useful insights from advanced analytics. They have built technology platforms that deliver timely data and insights when and where they are needed in the organization.

More data → More Business ?

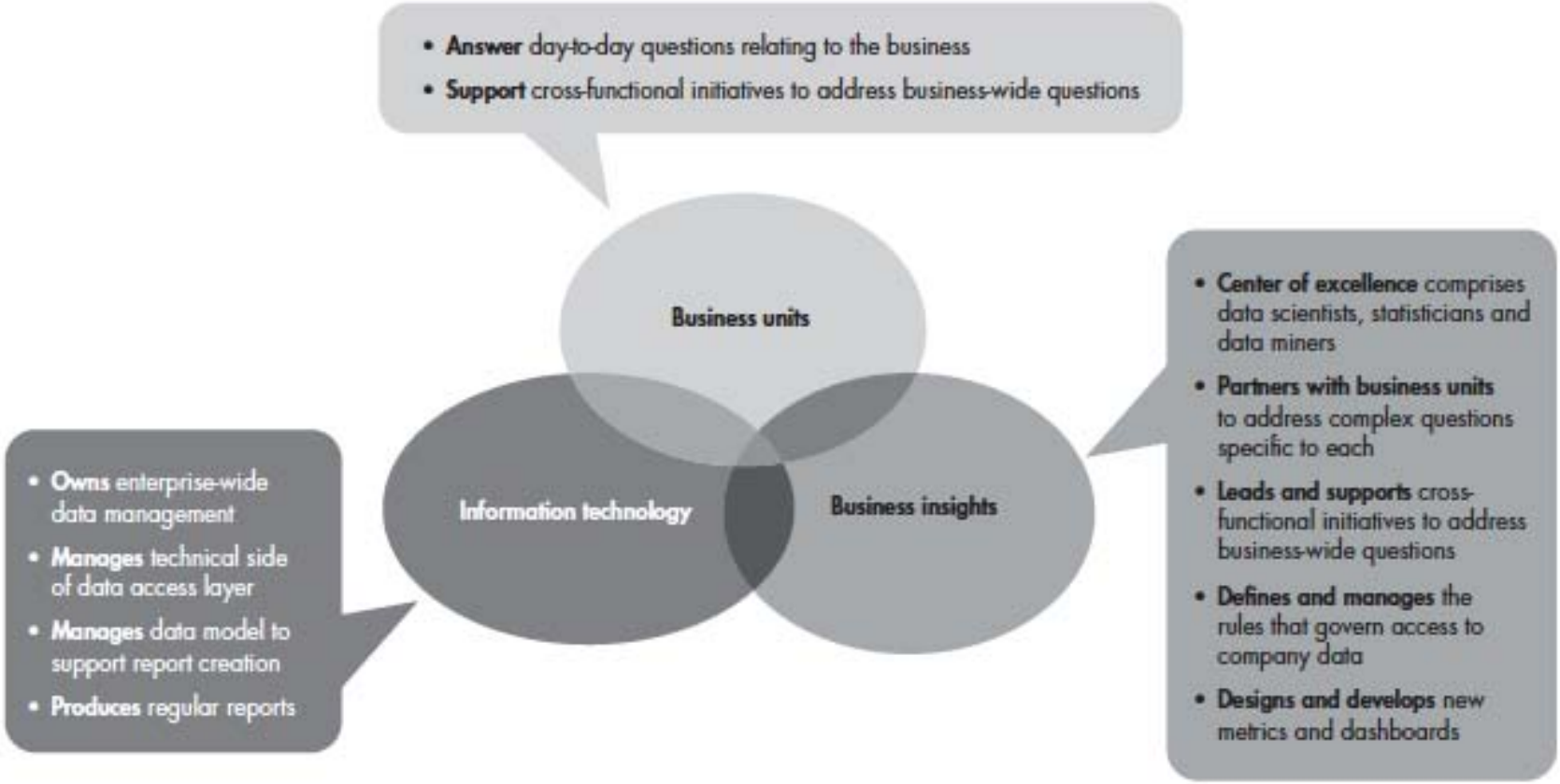
From an analytic perspective, it is generally easier to work with data that has some history than it is to attack brand-new data sets.

Comenzar analizando datos históricos, en lugar de comenzar directamente con nuevos datos

The temptation to acquire and mine new data sets has intensified with the explosion of social media and mobile devices. And yet many large organizations are already drowning in data, much of it held in silos where it cannot easily be accessed, organized, linked or interrogated. We've found that successful Big Data journeys tend to start by fully exploiting the organization's existing data.

¿Está tu empresa preparada para aprovechar un analista de datos ?

Figure 2: Big Data initiatives work best when data scientists partner closely with business units and IT



Source: Bain & Company

Conclusión

La revolución del Big Data ya está aquí. Hay empresas innovadoras que han logrado beneficiarse de este fenómeno pero otras más tradicionales han fracasado.

Compañias exitosas han seguido los siguientes tres reglas:

1. Prove your organization can apply advanced analytics to solve a few high-value business problems before investing in Big Data technology solutions.

2. Create value from your in-house data before expanding to new data sources. Then use test-and-learn approaches to inject forward-looking data sets into your historical data.

3. Align your operating model to enable your organization, particularly the front line, to act quickly and with confidence on the insights from your advanced analytics teams.

Temas para presentación día Martes

- Netflix: Big data en acción (Violeta)
- Shazam: Sistema de reconocimiento de canciones (Pedro)
- Las subastas y la maldición del ganador (Laura)
- Principio de la Navaja de Occam y la paradoja de Simpson (Oscar)
- El sistema IBM Watson Analytics (José Manuel)
- Reconocimiento de imágenes y predicciones usando Deep Learning (Marta)
- Algoritmos de Clustering en el ejército americano (Estefanía)
- Small world – 6 grados de separación y el porqué de que tus amigos tengan más amigos que tú (Ma. José)

Presentación grupal

Tema 3 - Presentaciones BIG DATA

Preparar una presentación grupal sobre el tema asignado para una duración aproximada de 10 minutos (entre 7 y 12 diapositivas).

Se evaluará:

a) Presentación visual: diseño de las diapositivas, estructura de la presentación siguiendo un hilo argumental.

b) Presentación oral: contenido al margen de lo mostrado visualmente, y forma de expresión.

c) Contenido global de la presentación: Si el material presentado es interesante y relevante al tema, y si se utilizan y citan diversas fuentes.

Temas

BIG DATA en pequeñas empresas	grupo de 3 estudiantes
En contra del Big Data	grupo de 3 estudiantes
Empleabilidad del Big Data	grupo de 2 estudiantes

Tema 1

Presentación 1 - BD en pequeñas empresas

Debéis analizar la utilidad de la aplicación de técnicas Big Data en pequeñas empresas, analizando entre otros aspectos las diferencias entre la aplicación de BD en grandes y pequeñas empresas, las ventajas/desventajas de su aplicación en pequeñas empresas, y otros factores que consideréis relevantes a la temática.

[Debate1-R1] https://www.google.com /search?q=how+to+implement+big+data+in+small+companies&ie=utf-8&oe=utf-8&aq=t& rls=org.mozilla:es-ES:official&client=firefox-a&channel=sb

Tema 2

Presentación 2 - En contra del BD

Debéis convencernos de que las empresas no deben apostar por el Big Data. Enfocando en las desventajas/inconvenientes que presenta está tecnología, debeís mostrar el lado negativo de la misma de forma seria y convincente. Algunos aspectos a considerar son:

--Privacidad, monopolio de los datos por parte de las grandes empresas.

-- Coste/beneficio de su aplicación. Dificultad de obtener mejoras.

--Marketing sin sentido, enfocando el BD como un fenómeno de datos pero sin conocimiento.

[Debate2-R1] http://big-data-gurus.com/2014/02/big-data-a-big-problem/

[Debate2-R2] http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=0

[Debate2-R3] http://venturebeat.com/2013/12/04/the-3-big-problems-in-big-data-hint-theyre-all-about-people/

Tema 3

Presentación 3 - Empleabilidad y Big Data

Presentar cifras e información sobre la creación de empleo en torno al fenómeno de Big Data. Analizar cuestiones como:

¿Qué se necesita para ser un experto en BD?

¿Qué perfil es el solicitado por las empresas?

¿Qué se enseña en los diferentes másteres de BD?

[Debate3-R1] Debate3-R1.pdf

[Debate3-R2] http://bigdata.lcc.uma.es/

[Debate3-R3] http://masteres.ugr.es/datcom/pages/ficha

Fin Big Data día 2

Sentiment Analysis

Conocer y/o Monitorizar la opinión de una audiencia respecto de algún tema:

- Valoración de un político
- Agrado de una película de cine
- Funcionamiento de la red de metro
- Experiencia de vuelo con una aerolínea

Otros nombres

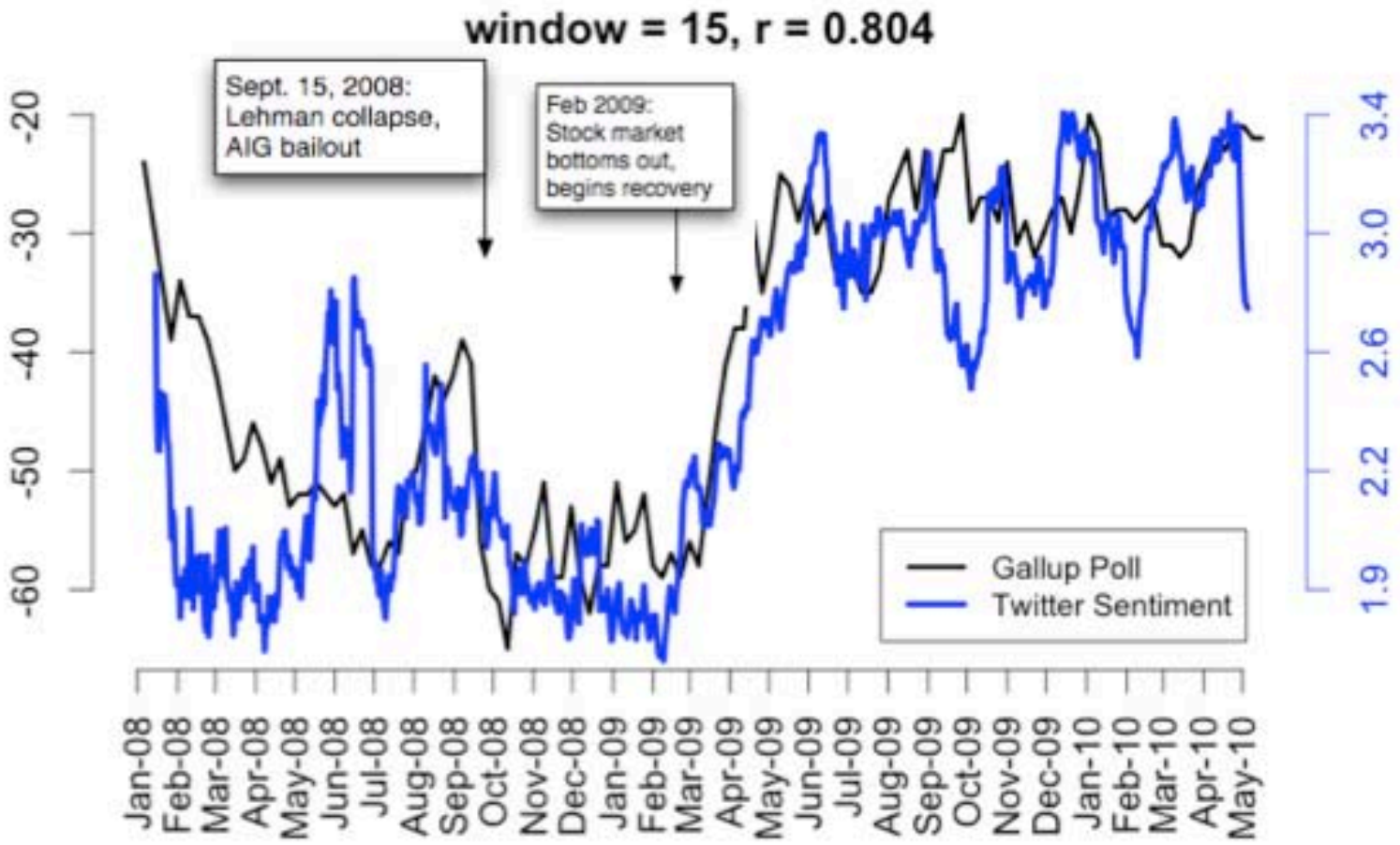
- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

Diferentes grados de análisis

- Simplest task:
  - Is the attitude of this text positive or negative?
- More complex:
  - Rank the attitude of this text from 1 to 5
- Advanced:
  - Detect the target, source, or complex attitude types

Twitter: una herramienta poderosa para sentiment analysis

Confianza del consumidor



Satisfacción de los consumidores

Airlines, Consumers, and Twitter

Anyone who travels regularly recognizes that airlines struggle to deliver a consistent, positive custom experience. Through extensive interview and survey work, the American Customer Satisfaction Index (http://theacsi.org/) quantifies this impression. As a group, airlines falls at the bottom of their industry rankings, below the Post Office and insurance companies:

Hotels	75	73	72	71	71	72	72	71	71	73	72	73	75	71	75	75	75		0.0	0.0
Limited Service Restaurants	69	70	66	68	69	69	70	71	71	74	NM**	76	77	77	78	78	75		-3.8	8.7
Supermarkets	76	75	74	73	73	74	73	75	75	74	73	74	75	76	76	76	75		-1.3	-1.3
Fixed Line Telephone Service	81	80	79	75	74	73	72	70	71	72	71	70	70	70	73	72	75	73	-2.7	-9.9
Internet News & Information									73	74	75	75	73	75	75	74	74		0.0	1.4
Network Cable TV News	77	76	70	62	65	62	64	62	65	68	66	68	69	67	69	71	74	77	4.1	0.0
Health Insurance								68	69	70	67	68	72	71	73	75	73		-2.7	7.4
Hospitals	74	74	71	67	72	70	69	68	70	73	76	71	74	77	75	77	73	77	5.5	4.1
Wireless Telephone Service											65	63	66	68	68	69	72	71	-1.4	9.2
U.S. Postal Service	61	69	74	69	71	71	72	70	73	72	74	73	71	73	74	74	71		-4.1	16.4
Internet Social Media																	70		N/A	N/A
Gasoline Stations	78	80	77	78	79	76	75	77	76	75	70	69	71	70	74	76	70		-7.9	-10.3
Subscription Television Service								64	61	61	61	61	63	62	64	63	66	66	0.0	3.1
Airlines	72	69	69	67	65	63	63	61	66	67	66	66	65	63	62	64	66		3.1	-8.3
Newspapers	72	68	69	69	66	69	68	68	63	64	68	63	63	66	64	63	65	65	0.0	-9.7
Municipal Utilities																		73	N/A	N/A
Investor-Owned Utilities																		74	N/A	N/A
Cooperative Utilities																		82	N/A	N/A

Consumidores utilizan twitter para expresar su malestar/agrado

@tkibbe
Terry Kibbe

4 1/2 hour delay IND - DCA. Told weather problems turned out to be mechanical. @Usairways #fail. This star gold flier hates the dishonesty

16 Jun via web

@johnfanelli
johnfanelli

@united #fail on wifi in red carpet clubs (too slow), delayed flight, customer service in red carpet club (too slow), hmmm do u see a trend?

17 Jun via web

Acceso y análisis a los datos

- Twitter permite de forma gratuita y cuasi-libre

(previo registro) acceder a un porcentaje de los tweets de forma instantánea.

el mismo sistema lleva incorporada una etiqueta

Hashtag (#) que indica los temas de los que trata el tweet, a fin de conectar con otros usuarios

(@united, @delta)

Análisis semántico

- Recolectamos los tweets y ahora deberíamos al menos clasificarlos en:
  - positivos
  - neutros
  - negativos

Es una tarea muy compleja, como veremos en un caso real, sobre todo si utilizamos algunos métodos rápidos y simples como el “bag of words”

Bag of words

- This model focuses completely on the words, or sometimes a string of words,
- It pays no attention to the "context" so-to-speak. The bag of words model usually has a large list, probably better thought of as a sort of "dictionary," which are considered to be words that carry sentiment. These words each have their own "value" when found in text.
- The values are typically all added up and the result is a sentiment valuation.

Natural Language Processing

- This model attempts understand the sentences structures, context, and is more focused on the succession of a string of words.
- Usually, this structure requires the machine to have understanding of grammar principles.
- Natural Language Processing (NLP) techniques are used to tag parts of speech, named entities, and more, in order to actually understand the "language" of the text, and not just look for target words.

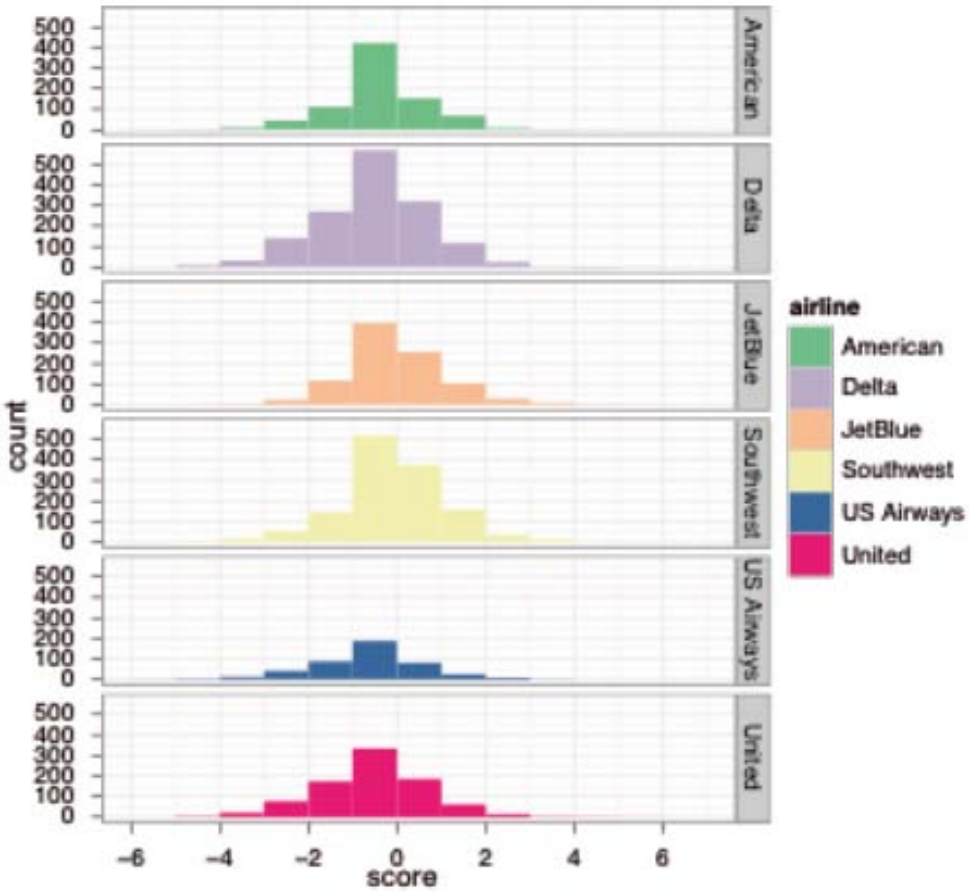
Evaluando algunos tweets

2 You're awesome and I love you

-5 I hate and hate and hate. So angry. Die!

4 Impressed and amazed: you are peerless in your achievement of unparalleled mediocrity.

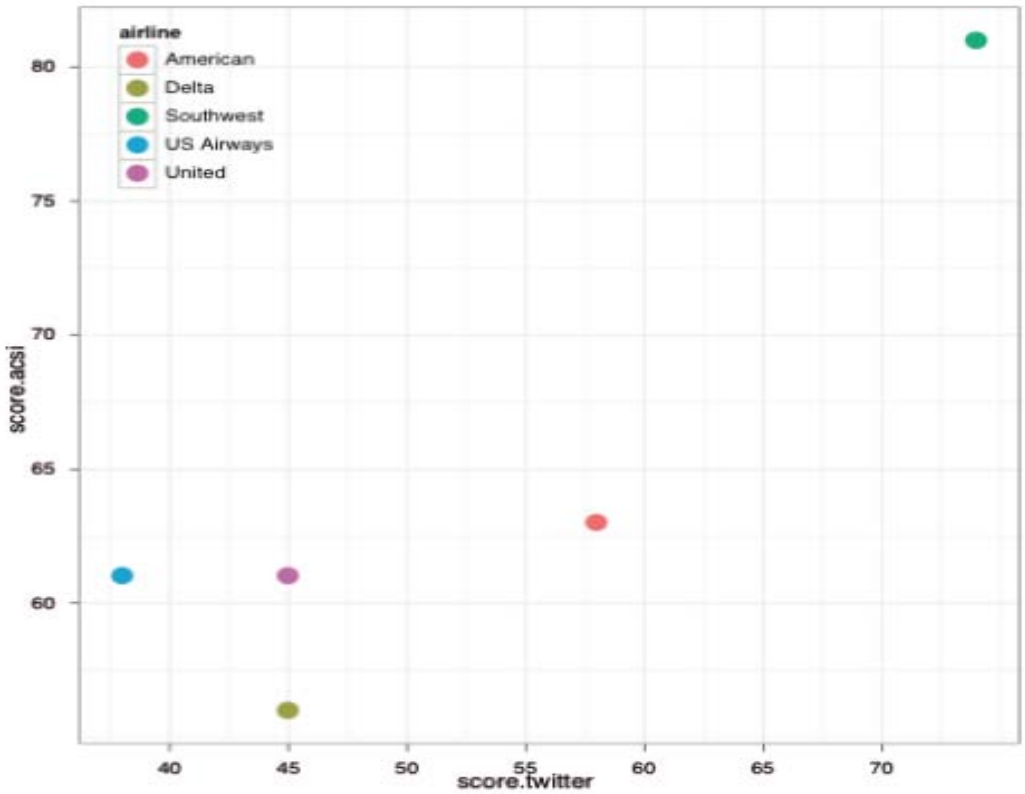
Sentiment analysis sobre empresas de aviación



Distribución de la valoración de tweets

Comparación con la encuesta de consumidores

code	airline	very.pos.count	very.neg.count	very.tot	score.twitter	score.acsi
AA	American	80	57	137	58	63
DL	Delta	152	185	337	45	56
UA	United	82	102	184	45	61
US	US Airways	38	62	100	38	61
WN	Southwest	207	72	279	74	81



Conclusiones

- Aún cuando utilicemos métodos aproximados (análisis semántico de los tweets), dado que utilizaremos una gran cantidad de datos, un Sentiment analysis relativamente simple permite obtener resultados parecidos a un método tradicional y mucho más costoso como una encuesta.