

# **Sistemas Inteligentes para la Gestión de la Empresa**

**2016 - 2017**



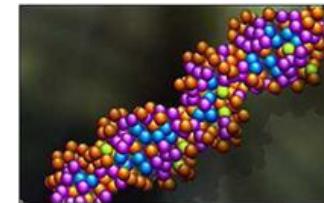
- **Tema 1. Introducción a la Ciencia de Datos**
- **Tema 2. Depuración y Calidad de Datos**
- **Tema 3. Análisis Predictivo para la Empresa**
- **Tema 4. Análisis de Transacciones y Mercados**
- **Tema 5. Modelos avanzados de Analítica de Empresa**
- **Tema 6. Big Data**
- **Tema 7. Aplicaciones de la Ciencia de Datos en la Empresa**

# Ciencia de Datos, Minería de Datos, Big Data

---

## 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, ...



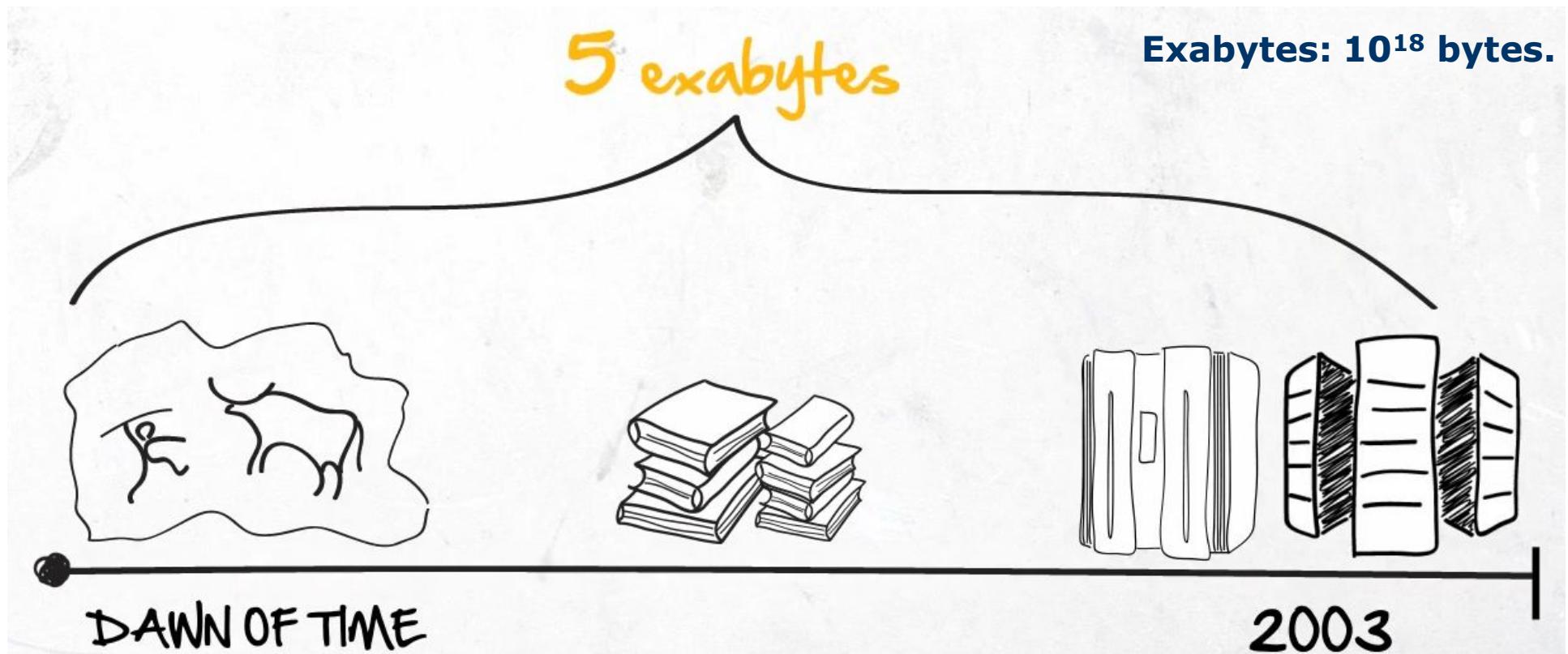
- Entretenimiento y Ocio
  - Imágenes en internet, películas, ficheros MP3, ...

- Medicina
  - Datos de pacientes, datos de escaner, radiografías ...



- Industria, Energía, ...
  - Sensores, ...

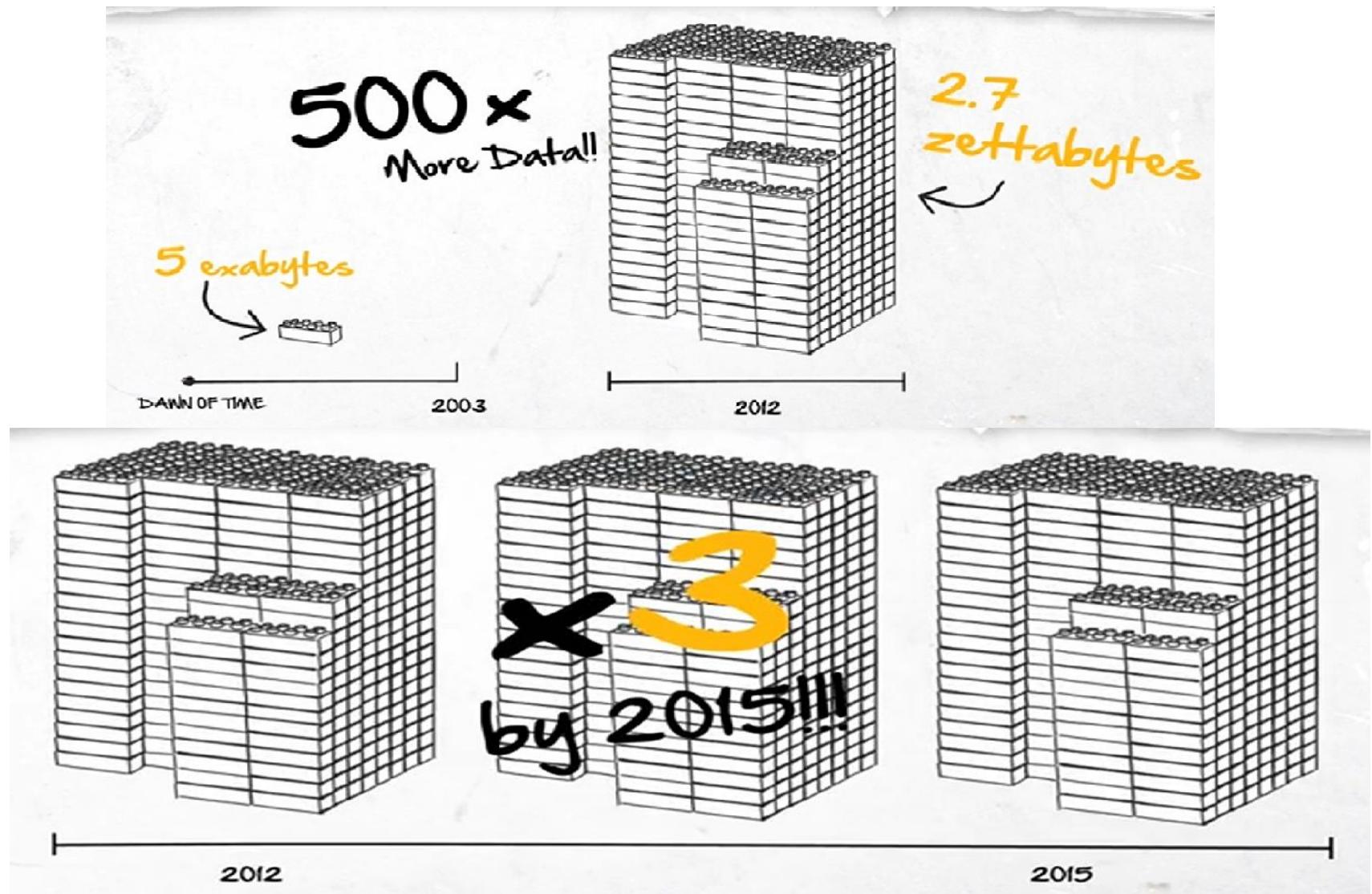
# La explosión de los datos

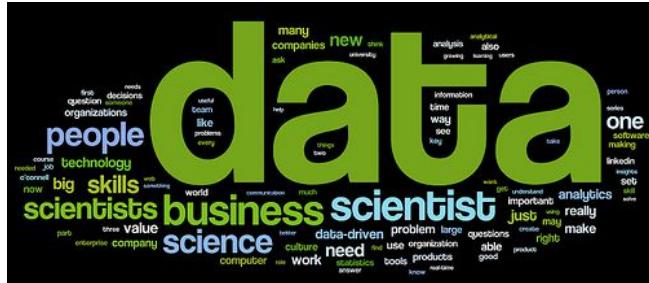


$1 \text{ EB} = 10^3 \text{ PB} = 10^6 \text{ TB} = 10^9 \text{ GB} = 10^{12} \text{ MB} = 10^{15} \text{ KB} = 10^{18} \text{ bytes.}$

# La explosión de los datos

Zettabytes:  $10^{21}$  bytes.





# 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

→ cantidades gigantescas 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**

# Ciencia de Datos, Minería de Datos, Big Data



INTERNET | Campus Party Europa 2013

'Es la década de los datos y de ahí vendrá la revolución'



<http://www.elmundo.es/elmundo/2013/09/03/navegante/1378>

Alex ' Sandy' Pentland, director del programa de emprendedores del 'Media Lab' del Massachusetts Institute of Technology (MIT)

Considerado por 'Forbes' como uno de los siete científicos de datos más poderosos del mundo



Ben Chams - Fotolia

# Índice



- ¿Qué es la Ciencia de Datos?
  - El poder de los datos y su impacto en nuestra sociedad
  - Herramientas y Lenguajes en Ciencia de Datos.  
Repositorio de Kaggle
  - Comentarios Finales

# Índice



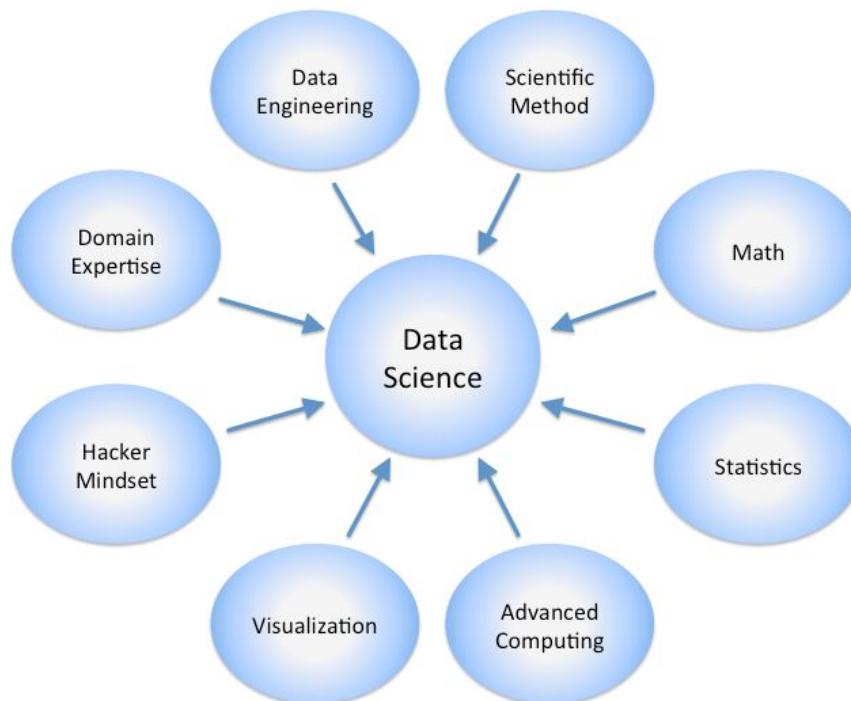
- ¿Qué es la Ciencia de Datos?
  - El poder de los datos y su impacto en nuestra sociedad
  - Herramientas y Lenguajes en Ciencia de Datos.  
Repositorio de Kaggle
  - Comentarios Finales

# Ciencia de Datos

---

## Data Science

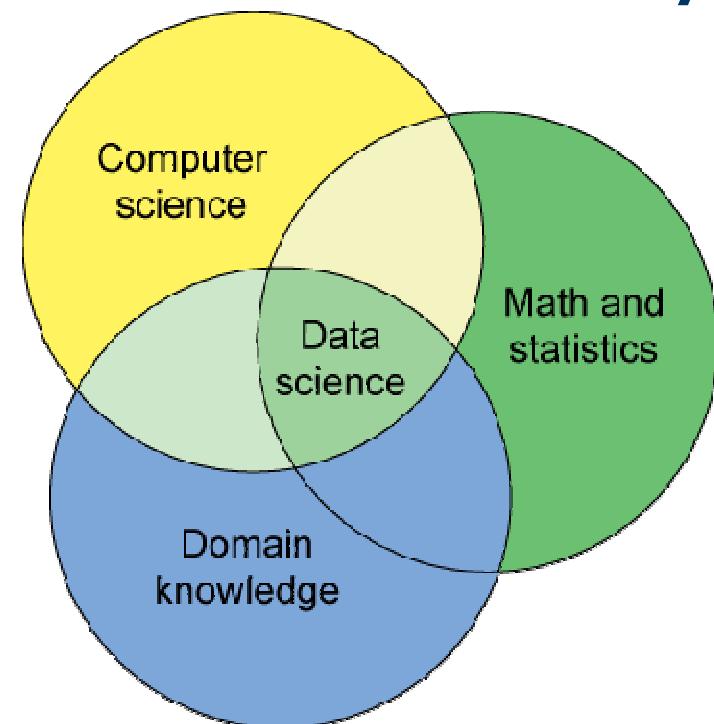
**Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas a la extracción de conocimiento de datos, incluyendo Big Data**



# Ciencia de Datos

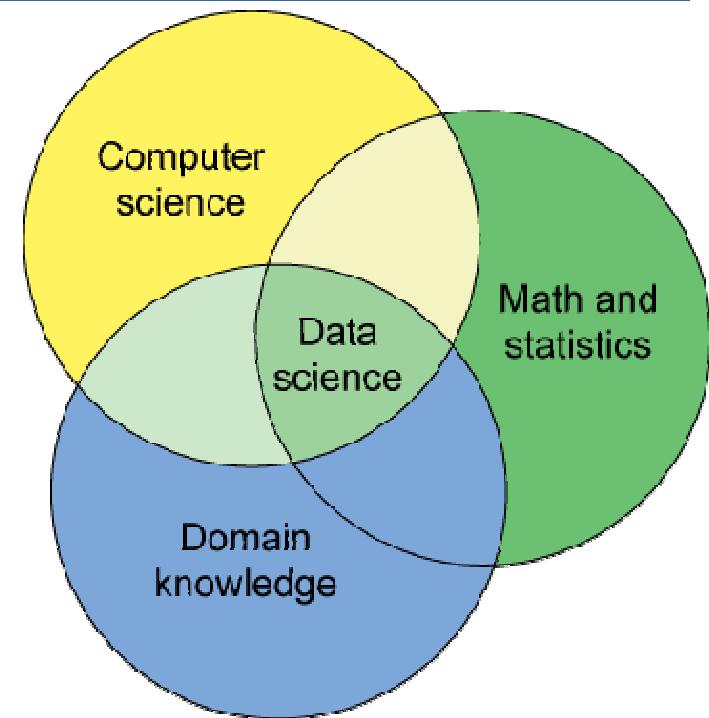
---

**Data Science o la Ciencia de Datos incorpora diferentes elementos y se basa en las técnicas y teorías de muchos campos, incluyendo las matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelado de la incertidumbre, almacenamiento de datos y la informática de alto rendimiento con el objetivo de extraer el significado de datos y la creación de productos de datos.**



# Ciencia de Datos

---



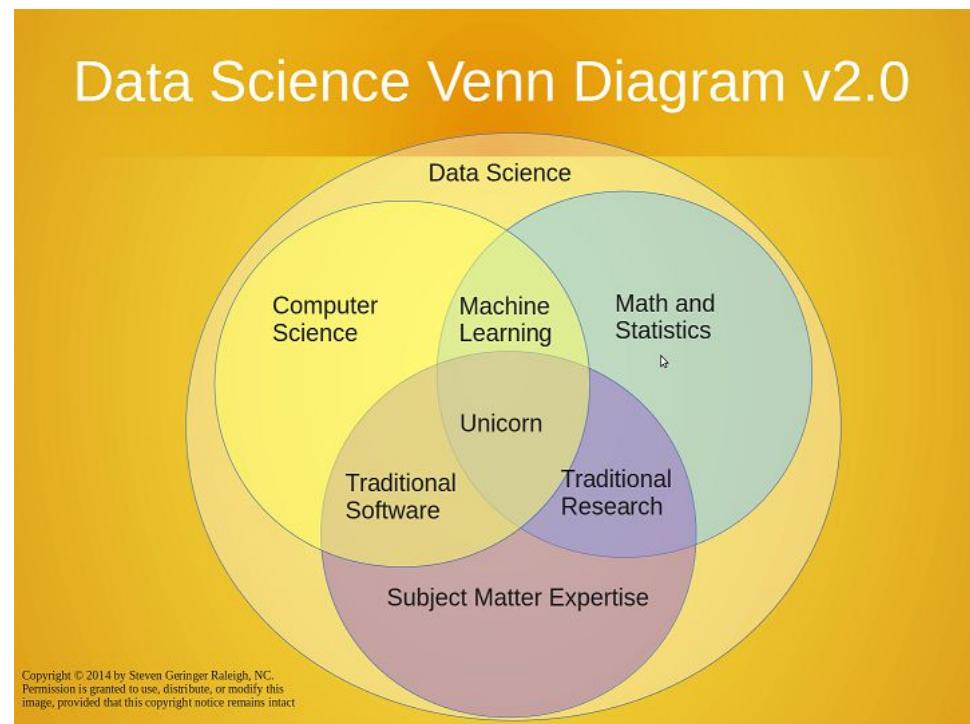
**Es un término relativamente nuevo que se utiliza a menudo de manera intercambiable con analítica de negocio. La ciencia de datos busca utilizar todos los datos disponibles y relevantes para “extraer conocimiento” que pueda ser fácilmente comprendido por los expertos en el área de aplicación. Un experto de la ciencia de datos se denomina un científico de datos.**

# Ciencia de Datos

---

## ¿Qué es un Científico de Datos?

**Un científico de datos es un profesional que debe dominar las ciencias matemáticas y la estadística, acabados conocimientos de programación (y sus múltiples lenguajes), ciencias de la computación y analítica.**



# Ciencia de Datos

---



**José Antonio Guerrero: uno de los mejores científicos de datos del mundo (Plataforma Kaggle)**

**¿Qué es un científico de datos?**

*"Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos.*

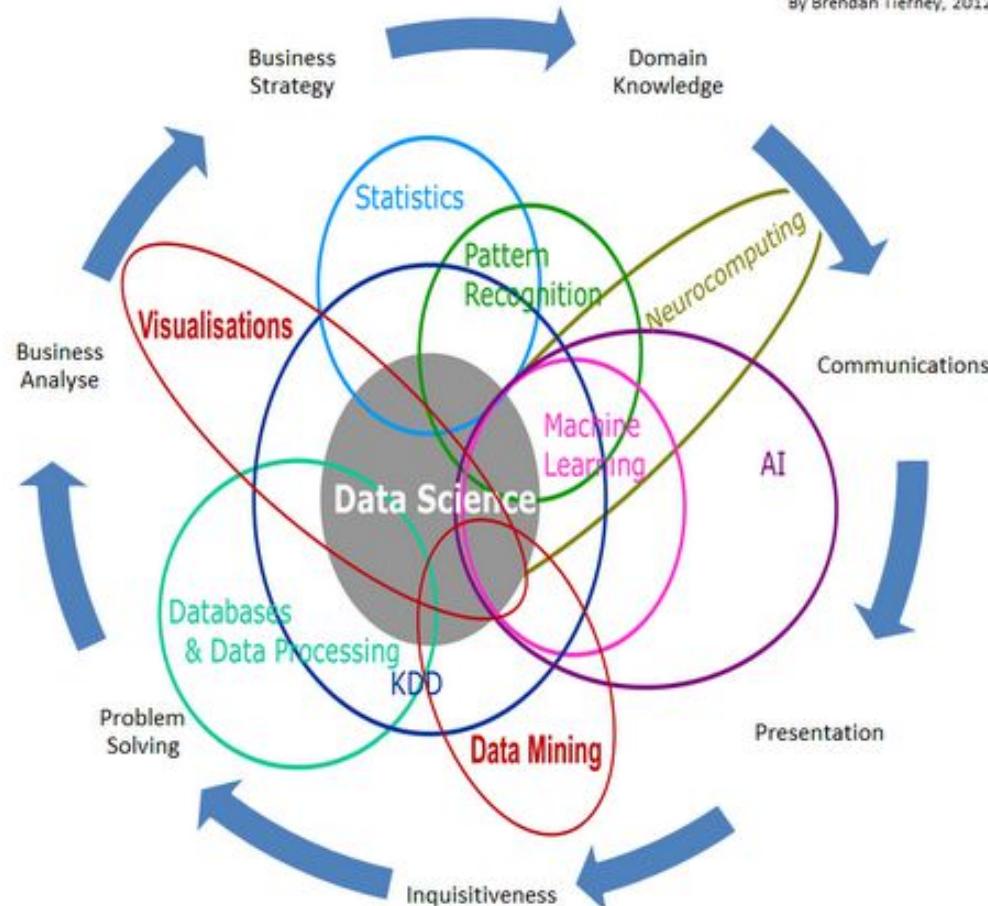
*De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco."*

Leer más: [http://www.elconfidencial.com/tecnologia/2013-12-19/un-matematico-andaluz-desconocido-es-el-mejor-cientifico-de-datos-del-mundo\\_67675/](http://www.elconfidencial.com/tecnologia/2013-12-19/un-matematico-andaluz-desconocido-es-el-mejor-cientifico-de-datos-del-mundo_67675/)

# Ciencia de Datos

## Data Science Is Multidisciplinary

By Brendan Tierney, 2012

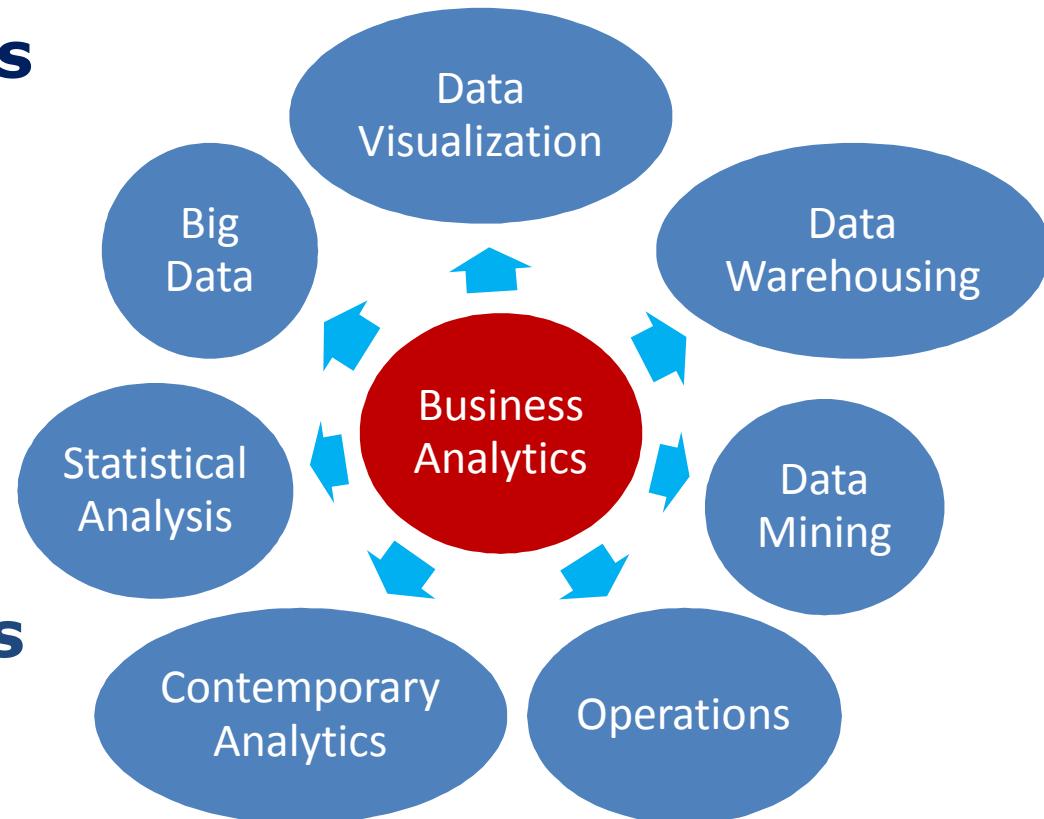


## Business Analytics

### Data Science



### Business Analytics



**Data mining:** Data Preprocessing, Supervised learning, unsupervised learning, forecasting

**Contemporary Analytics:** text mining, network analytics, social analytics, customer analytics, web analytics, risk analytics, information retrieval and recommendations

**Statistical Analysis:** Estimation and inference; and regression models

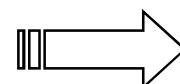
**Operations:** Simulation and optimization

# Ciencia de Datos

## Minería de Datos



We have rich data,  
but poor information

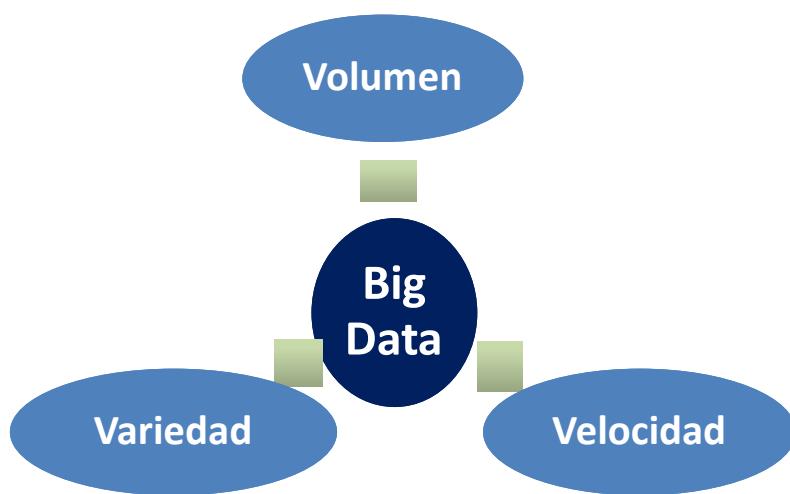


Data mining-searching for knowledge  
(interesting patterns) in your data.

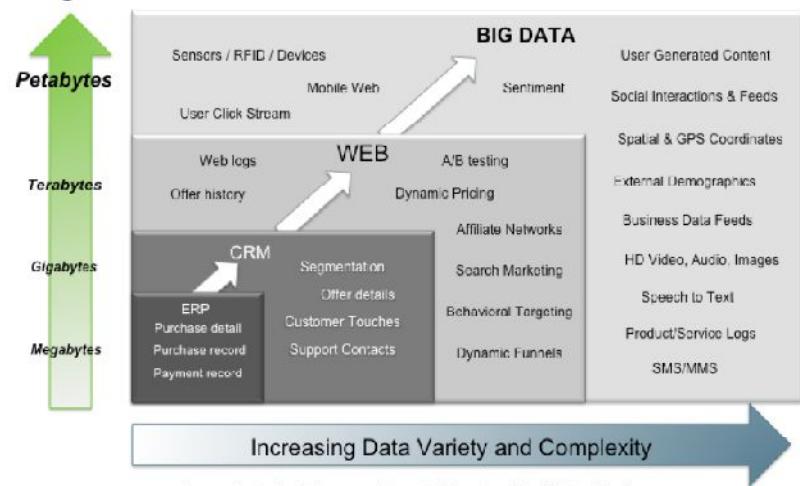
# Ciencia de Datos

## Big Data

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



Big Data = Transactions + Interactions + Observations



# Índice



- ¿Qué es la Ciencia de Datos?
  - **El poder de los datos y su impacto en nuestra sociedad**
  - Herramientas y Lenguajes en Ciencia de Datos.  
Repositorio de Kaggle
  - Comentarios Finales

# El poder de los datos

## Análisis de asociaciones

Pañales y cerveza. Ficción y leyenda para ilustrar el análisis de transacciones



Si compro cerveza, entonces compro pañales

60%

Si compro pañales, entonces compro cerveza

100% ✓

# El poder de los datos

## Análisis de asociaciones en transacciones de tarjetas de crédito



Fuente: Big Data. La revolución de los datos masivos. Pag. 160

# El poder de los datos

# Análisis de transacciones



Fuente: Big Data. La revolución de los datos masivos. Pag. 77

# El poder de los datos

## Análisis de transacciones: Un chivo expiatorio



**Target (cadena de grandes almacenes) que utiliza el análisis de transacciones y asociaciones.**



**Unos días después el director llamó al padre para disculparse.**

**Respuesta conciliadora del padre:**

**“He estado hablando con mi hija –dijo el padre– Resulta que en mi casa han tenido lugar ciertas actividades de las que yo no estaba del todo informado. Mi hija sale de cuentas en agosto. Soy yo el que les debe una disculpa”.**



# El poder de los datos

## Análisis de transacciones

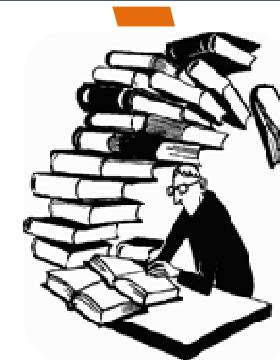
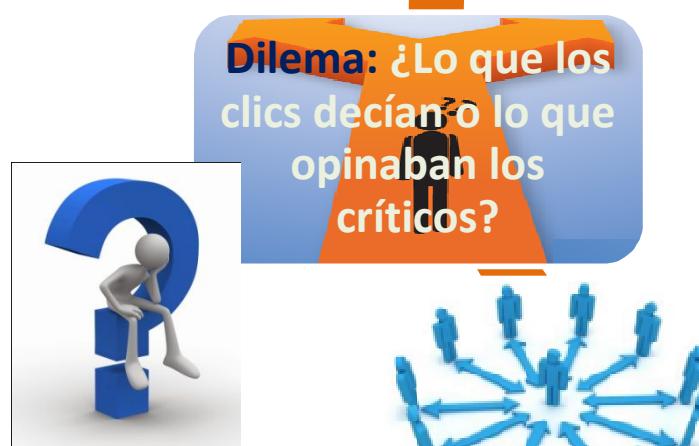
### Amazon: Sistema de recomendación



Los datos incrementaron tremadamente las ventas  
Ahora más de 1/3 de las ventas son gracias a las recomendaciones



Críticos y editores literarios  
La voz de Amazon (1995)



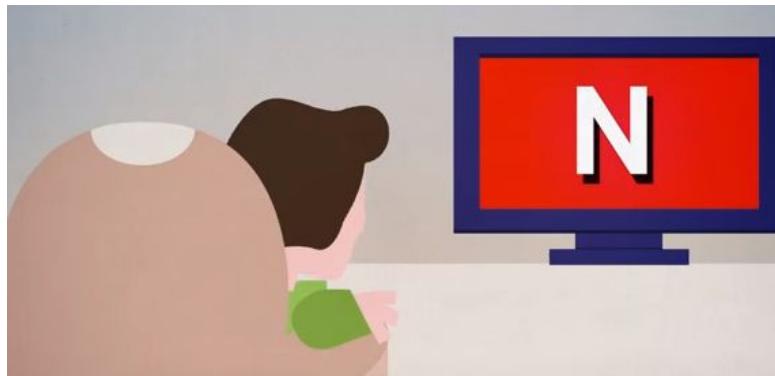
Greg Linde (1997) propuso un sistema de recomendaciones, filtrado colaborativo “artículo a artículo”



Fuente: Big Data. La revolución de los datos masivos. Pag. 69.

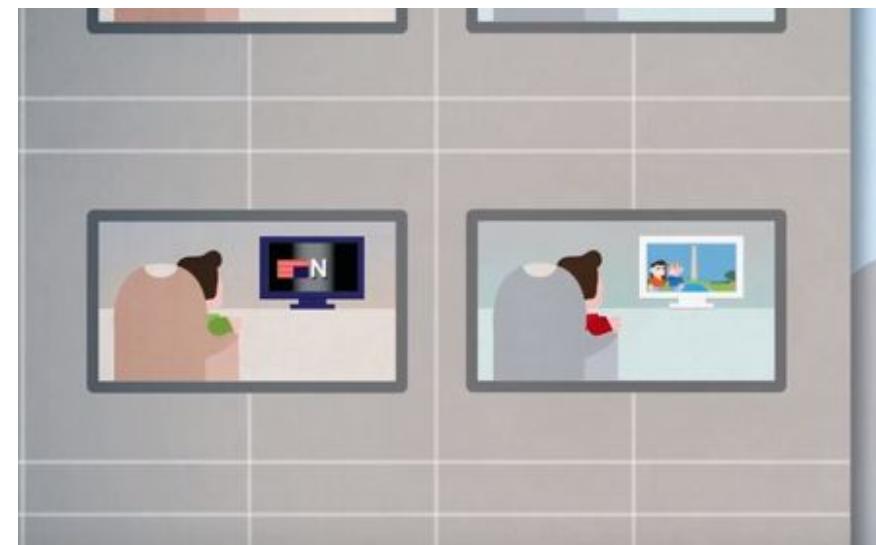
# El poder de los datos

## Netflix: Sistema de recomendación



Para Netflix, compañía de alquiler de películas online, las tres cuartas partes de los pedidos nuevos surgen de las recomendaciones.

Netflix y Amazon son dos empresas cuyo plan de negocio está basada en big data y sistemas de recomendación



# Analizando Twitter para medir la Salud Pública

**You Are What You Tweet**

Un sistema de filtrado de datos de Twitter puede inferir aspectos de salud analizando 144M de tuits (2011-2013)



Se obtienen 13 grupos coherentes de mensajes correlacionados

- Gripe estacional ( $r= 0.689$ ) y alergias ( $r = 0.810$ )
- Ejercicio y obesidad relacionados con datos geográficos, ..

**Discovering Health Topics in Social Media Using Topic Models**

Michael J. Paul, Mark Dredze, Johns Hopkins University, Plos One 9(8)

**e103408**, 2014

**doi:10.1371/journal.pone.0103408**

# Banca: Identificación de personas con las compras de tarjetas de crédito

[http://elpais.com/elpais/2015/01/29/ciencia/1422520042\\_066660.h](http://elpais.com/elpais/2015/01/29/ciencia/1422520042_066660.h)

PRIVACIDAD EN INTERNET »

## Cuatro compras con la tarjeta bastan para identificar a cualquier persona

- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT



## Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,<sup>1,\*</sup> Laura Radaelli,<sup>2</sup> Vivek Kumar Singh,<sup>1,2</sup> Alex “Sandy” Pentland<sup>1</sup>

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that women are more reidentifiable than men in credit card metadata.

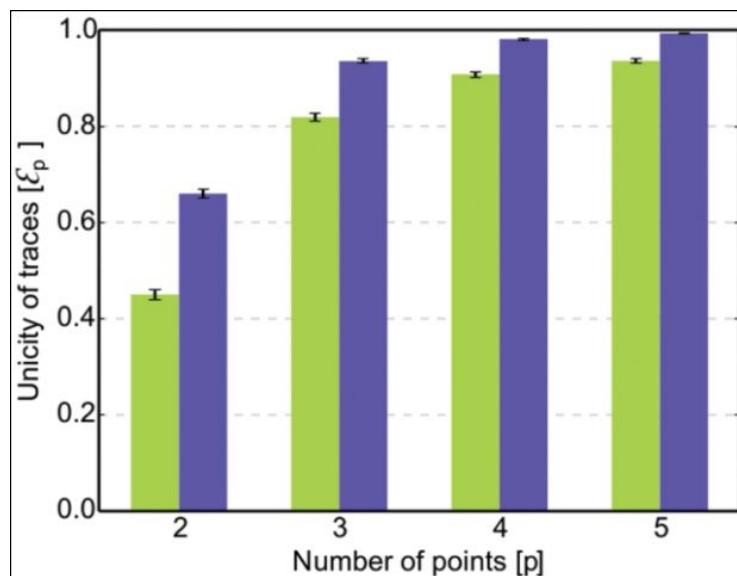
<http://www.sciencemag.org/content/347/6221/536>

# Banca: Identificación de personas con las compras de tarjetas de crédito

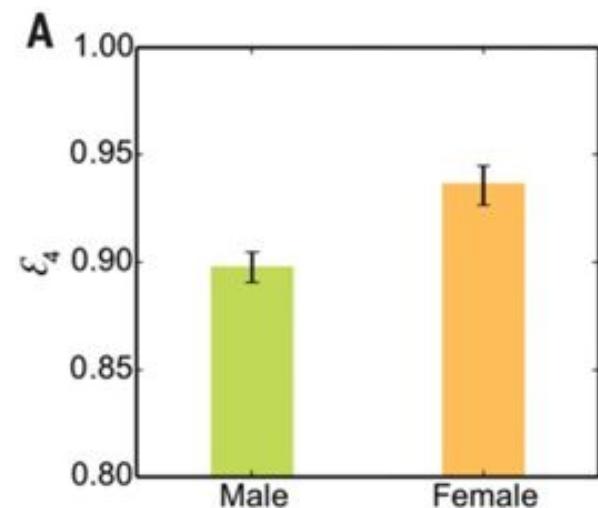
PRIVACIDAD EN INTERNET »

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

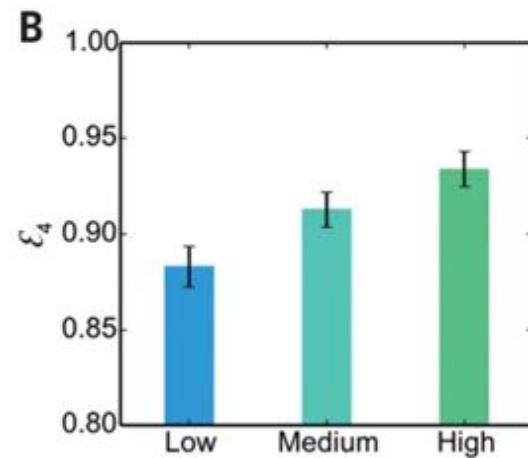
- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT



Identificación por el número de compras



Identificación por el género



Identificación por el poder adquisitivo

# Impacto en la sociedad

---

**Salud**

**Redes sociales como fuente de datos**

**Industria, comercio, banca, ...**

**Ocio y cultural (Ej. Recomendaciones)**

**Política, Bien social (Social good)**



 **UNITED NATIONS GLOBAL PULSE**  
Harnessing big data for development and humanitarian action

**GLOBAL PULSE PROJECT SERIES**  
The Global Pulse Project Series is a collection of case studies of 20 data innovation projects covering global issues ranging from public health to climate change, food security to employment.

[Read More /](#)

• • • •



**Global Pulse Project Series**

# Gran Impacto en la Sociedad y presencia en los medios de comunicación



INTERNET | Campus Party Europa 201

'Es la década d  
ahí vendrá la re



<http://www.elmundo.es/elm>

EL PAÍS

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOL

El maná de los datos

- La conversión de datos en información útil para miles de dólares en 2015. La herramienta 'I

SUSANA BLÁZQUEZ | Madrid | 29 SEP 2013 - 01:0

Archivado en: Citigroup Cap Gemini Sogeti S/ IBM Telefónica Aplicaciones informáticas Tecnol



EL PAÍS

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOLSA MIS AHORROS VIVIENDA TECNOLOGÍA OF

EMPRENDEDORES »

El Big Data echa una mano al campo

- Una empresa española recoge miles de datos para predecir las cosechas

MARÍA FERNÁNDEZ | 30 NOV 2014 - 00:00 CET

Archivado en: Bases datos Emprendedores Aplicaciones informáticas Empresas Programas informáticos Economía Informática Industria



# Impacto Económico

**La demanda de profesionales formados en Ciencia de Datos y *Big Data* es enorme.**

**Se estima que la conversión de datos en información útil generó un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos.**

**España necesitaba para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y *Big Data*.**

**España necesitará 60.000 profesionales de Big Data hasta 2015**

22 octubre, 2013 Eventos 18



España necesitará 60.000 profesionales de Big Data hasta 2015

Toledo.

## EL PAÍS ECONOMÍA

PORQUÉ | INTERNACIONAL | POLÍTICA | CULTURA | DEPORTES | SOCIEDAD | TECNOLOGÍA |

ESTÁ PASANDO | Multa a la banca | Revuelo en Hacienda | Eléctricas y renovables | Paro |

### El maná de los datos

- La conversión de datos en información útil para las empresas generará un mercado de 132.000 millones de dólares en 2015. La herramienta 'big data' sacará del mercado a quien no la use

SUSANA BLÁZQUEZ | Madrid | 29 SEP 2013 - 01:00 CET

Archivado en: CitiGroup Cap Gemini Sogeti SAP Oracle ING Bank BBVA Mapfre Bases de datos IBM Telefónica Aplicaciones informáticas Tecnología Empresas Programas informáticos Economía



[http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725\\_938376.html](http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725_938376.html)

[http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde\\_4377072\\_member\\_5811011886832984067#!](http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde_4377072_member_5811011886832984067#!)

# Índice



- ¿Qué es la Ciencia de Datos?
- El poder de los datos y su impacto en nuestra sociedad
- **Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle**
- Comentarios Finales

# Herramientas, Lenguajes, Kaggle

Una web sobre el software libre para Ciencia de Datos ...

---

## Software (open source tools)



---

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

---

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

# Herramientas, Lenguajes, Kaggle

Una web sobre el software libre para Ciencia de Datos ...

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

The image displays a grid of logos for various open-source data science tools, categorized into 18 boxes:

- Data Analysis & Platforms**: Includes Hadoop, PARACCEL, Storm, HPCC Systems, GridGain, Dremel, Hortonworks, Zettaset, calport, ORACLE, and HD.
- Databases / Data warehousing**: Includes INFOBRIGHT, Hibari, Bigdata, OrientDB, Neo4j, HIVE, redis, and Hypertable.
- Operational**: Includes Versant JPA, MarkLogic, mObject, and iBASE INTERNATIONAL.
- Multivalue database**: Includes Rocket, U2, REVELATION SOFTWARE, northgate, OpenQM, and Ladybridge Systems.
- Big Data search**: Includes Lucene, Apache Solr, and elasticsearch.
- Data aggregation**: Includes oqoop, Quarkus, and chukwa.
- Business Intelligence**: Includes talend\*, JASPERSOFT, pentaho, Jedox, Exchange, ACTUATE, and KNIME.
- Data Mining**: Includes RAPID MINER, RAPID ANALYTICS, mahout, WEKA, KEEL, togaware, and SPMF.
- Social**: Includes Apache Kafka, ThinkUp, and Corona.
- Multidimensional**: Includes GT.M, QSciDB, and rasdaman.
- KeyValue**: Includes AEROSPIKE, leveldb, GENIE DB, Chordless, Tokyo Cabinet, Scalaris, SCALIEN, Project Voldemort, hamsterdb, RAPTORDB, FairCom, STS DB, HyperDex, IQLECT, OpenLDAP, and ioremap.net.
- Document Store**: Includes mongoDB, CLUSTERPOINT, JasDB, SchemafreeDB, COUCHBASE, RaptorDB, ejdb, SISDb, denso db, CouchDB, Raven DB, and Riak.
- Graphs**: Includes Gephi, InfiniteGraph, AllegroGraph® 4.9, FlockDB, GraphBuilder, Gremlin, HYPERGRAPHDB, GraphBase, and BrightstarDB.
- Grid Solutions**: Includes GIGASACES, HAZELCAST, and Galaxy.
- Object databases**: Includes db4objects, Picolisp, STARCOUNTER, siaqodb, HSS Database™, ZOPE, MORANTEK INFORMATION SYSTEMS, Magma, NEOPPOD, Sterling, EyeDB, and NDatabase.
- Multimodel**: Includes ArangoDB, alchemydatabase, and sedna.
- XML Databases**: Includes existdb, BASE, Qizx, and xindice.

Created by: www.bigdata-startups.com

# Herramientas, Lenguajes, Kaggle

---

**KNIME (o Konstanz Information Miner) es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. KNIME está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en java.**

**Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com GmbH, radicada en Zúrich, Suiza, continúa su desarrollo además de prestar servicios de formación y consultoría.**

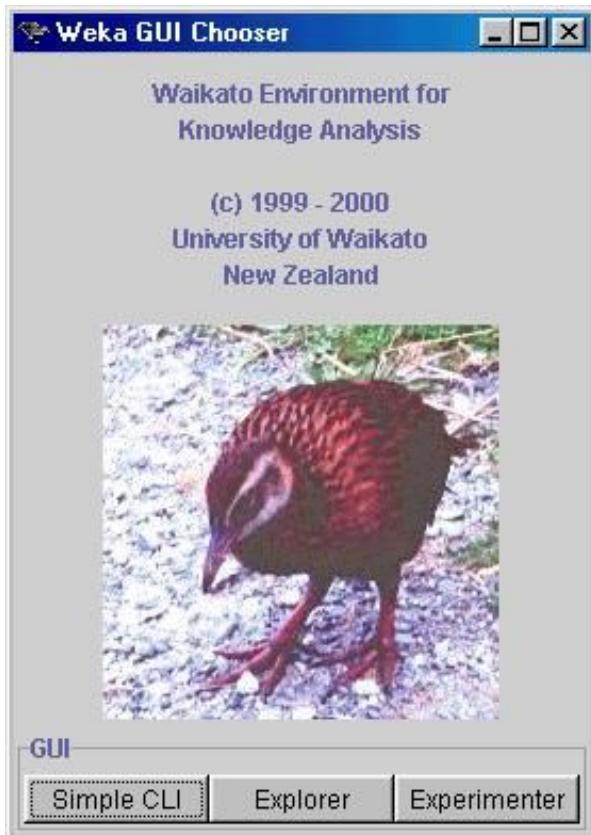


<https://www.knime.org/>

# Herramientas, Lenguajes, Kaggle

---

## Weka



- **The University of Waikato, New Zealand**
- **Machine learning software in Java implementation**

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

# Herramientas, Lenguajes, Kaggle

---

## KEEL

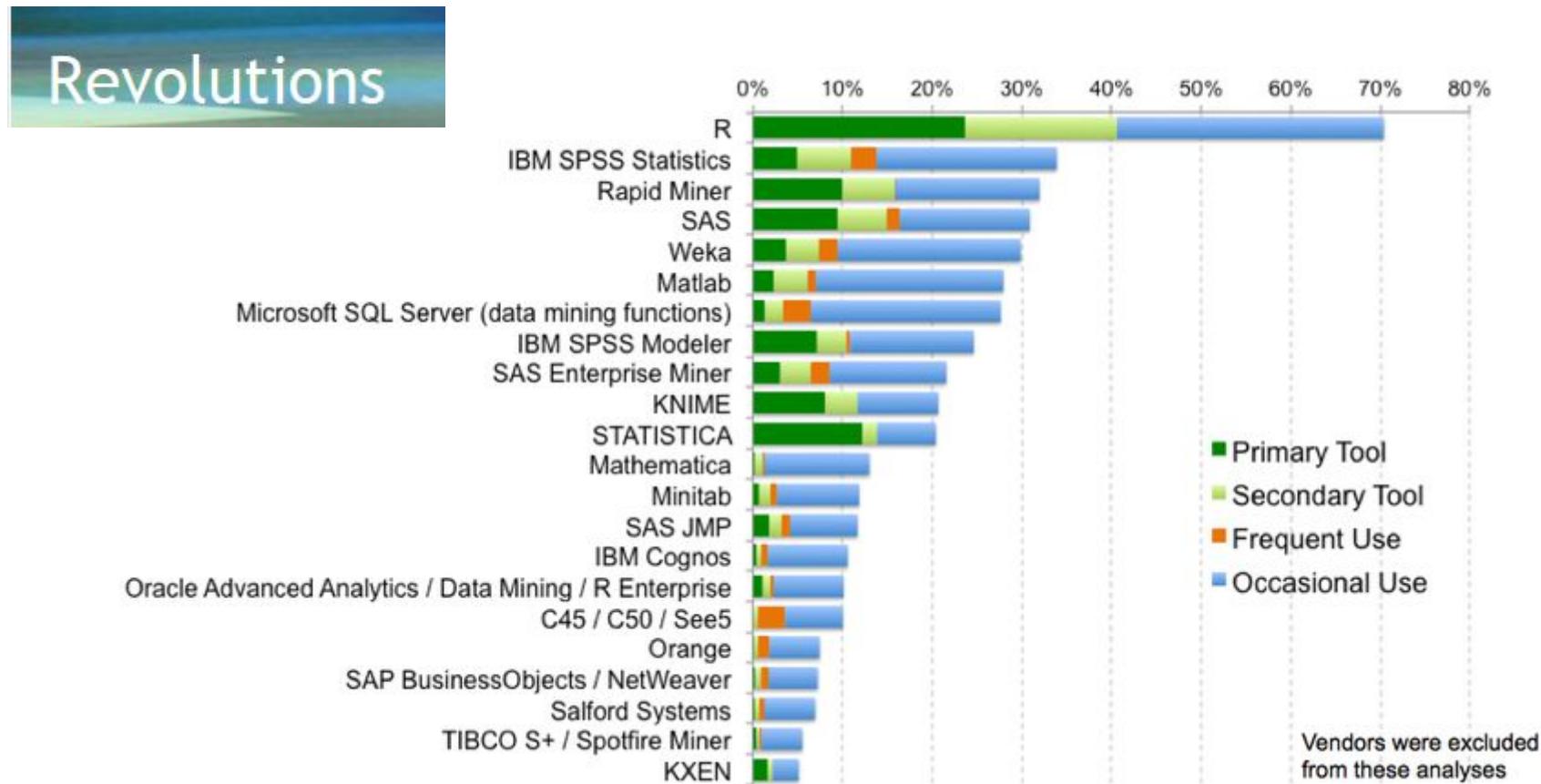


- . University of Granada
- . Machine learning  
software in Java  
implementation

<http://www.keel.es/>

# Herramientas, Lenguajes, Kaggle

## Sobre herramientas de minería de datos

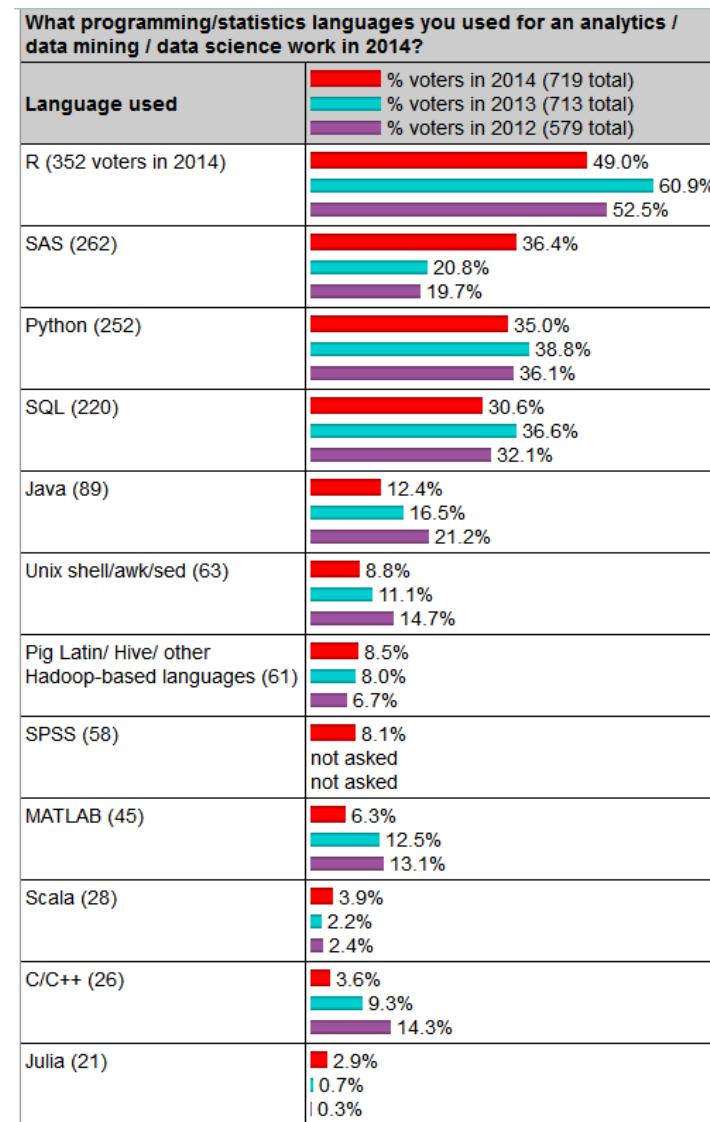
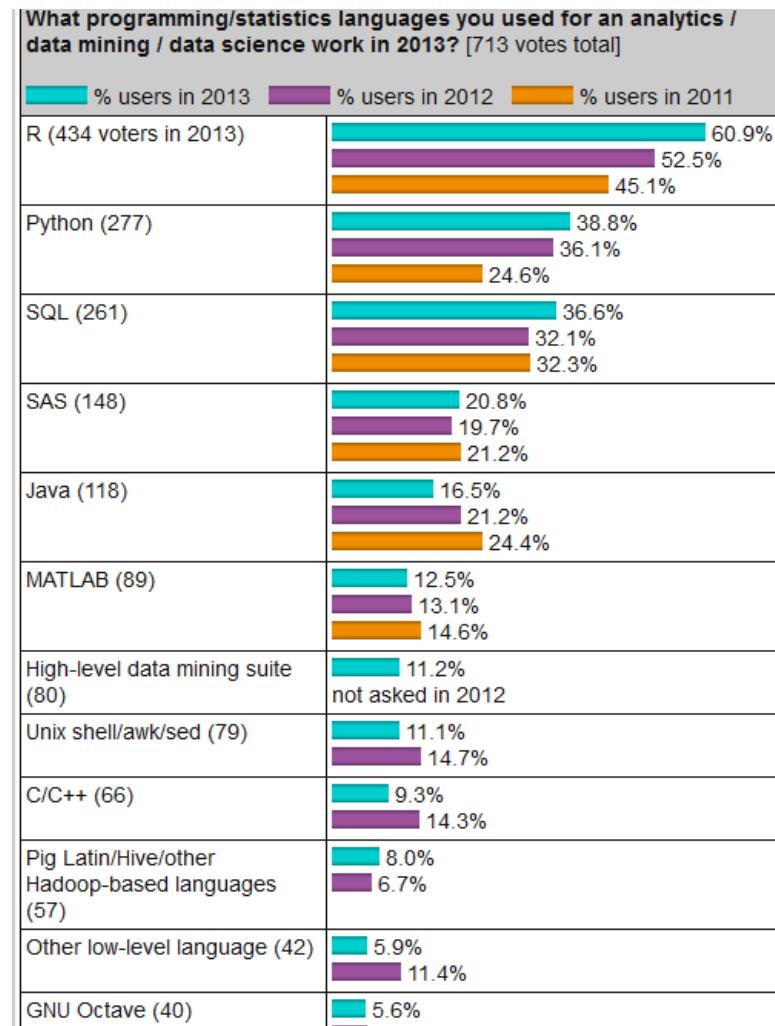


<http://blog.revolutionanalytics.com/2013/10/r-usage-skyrocketing-rexer-poll.html>

# Herramientas, Lenguajes, Kaggle

## Sobre los lenguajes de programación (R, Phyton, ...).

### Lenguajes a usar para Data Science



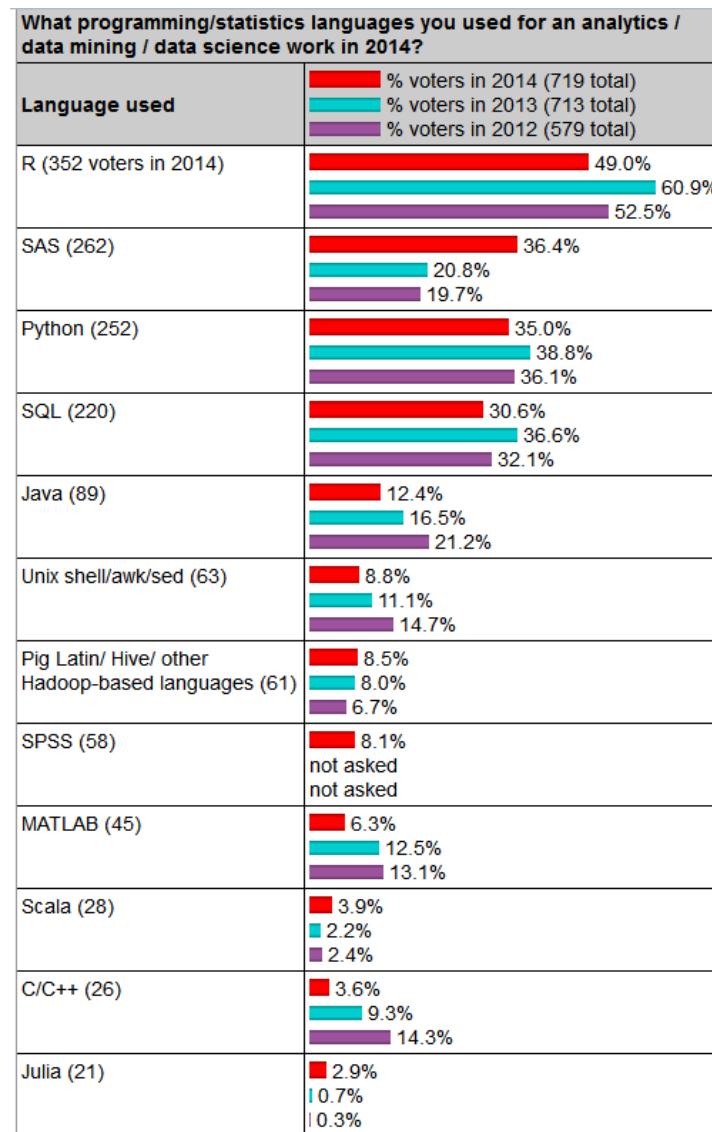
# Herramientas, Lenguajes, Kaggle

Sobre los lenguajes de programación (R, Phyton, ...).

**Consolidation among top 4 languages: R, SAS, Python, and SQL, and decline in usage of less popular languages for data mining: Java, Unix shell, MATLAB, C/C++, Perl, Octave, Ruby, Lisp, F.**

By Gregory Piatetsky, Aug 18, 2014.

<http://www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html>



# Herramientas, Lenguajes, Kaggle

Sobre los lenguajes de programación (R, Phyton, ...). El website CRAN

---

*[cran.r-project.org/](http://cran.r-project.org/)*

## The Comprehensive R Archive Network



[CRAN](#)  
[Mirrors](#)  
[What's new?](#)

### Contributed Packages

#### Available Packages

Currently, the CRAN package repository features 5799 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

**<http://cran.r-project.org/web/views/MachineLearning.html>**

# Herramientas, Lenguajes, Kaggle

## Sobre herramientas de minería de datos



Python Software  
Foundation



<http://scikit-learn.org/stable/>

### Classification

Identifying to which set of categories a new observation belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ...

— Examples

### Regression

Predicting a continuous value for a new example.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, ridge regression, Lasso, ...

— Examples

### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** K-Means, spectral clustering, mean-shift, ...

— Examples

### Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization. — Examples

### Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics.

— Examples

### Preprocessing

Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction.

— Examples

# Herramientas, Lenguajes, Kaggle

## Sobre herramientas de minería de datos

---



Anaconda



Anaconda

<http://docs.continuum.io/anaconda/>

Anaconda is a free collection of powerful packages for Python that enables large-scale data management, analysis, and visualization for Business Intelligence, Scientific Analysis, Engineering, Machine Learning, and more.

\*\*\*\*\*

## Large learning problems

theano

Phytон library

<https://pypi.python.org/pypi/Theano>



Deep Learning

Pylearn is a Python library for machine learning, built on top of Theano, our library for defining, optimizing and evaluating mathematical expressions involving multi-dimensional arrays.

# Herramientas, Lenguajes, Kaggle

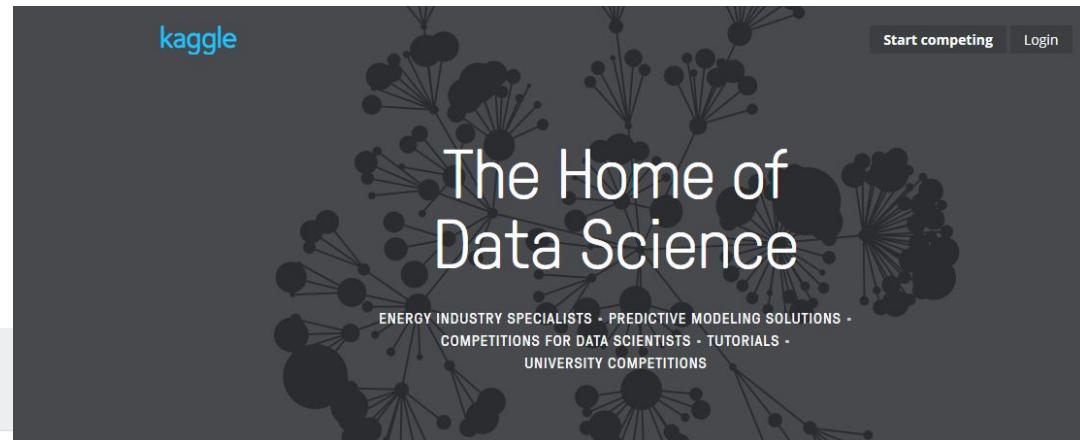
... y un buen enlace para comenzar a practicar, **KAGGLE**

## **Kaggle: The Home of Data Science**

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

**Es un portal web que ofrece competiciones, tutoriales, actividades académicas ...**

kaggle  
*in Class*



### Academic Machine Learning Competitions

Theory, meet practice.

Kaggle hosts free projects for hundreds of universities around the globe. Engage students with an opportunity to apply machine learning to real problems.

[Learn about hosting](#)

Berkeley  
UNIVERSITY OF CALIFORNIA

UNIVERSITY  
OF  
CALIFORNIA

UCL

COLUMBIA

Cornell

ERASMUS  
UNIVERSITY

HARVARD

THE UNIVERSITY OF  
MELBOURNE

MICHIGAN

UNIVERSITY OF  
OXFORD

Stanford  
University

UNIVERSITY  
OF TORONTO

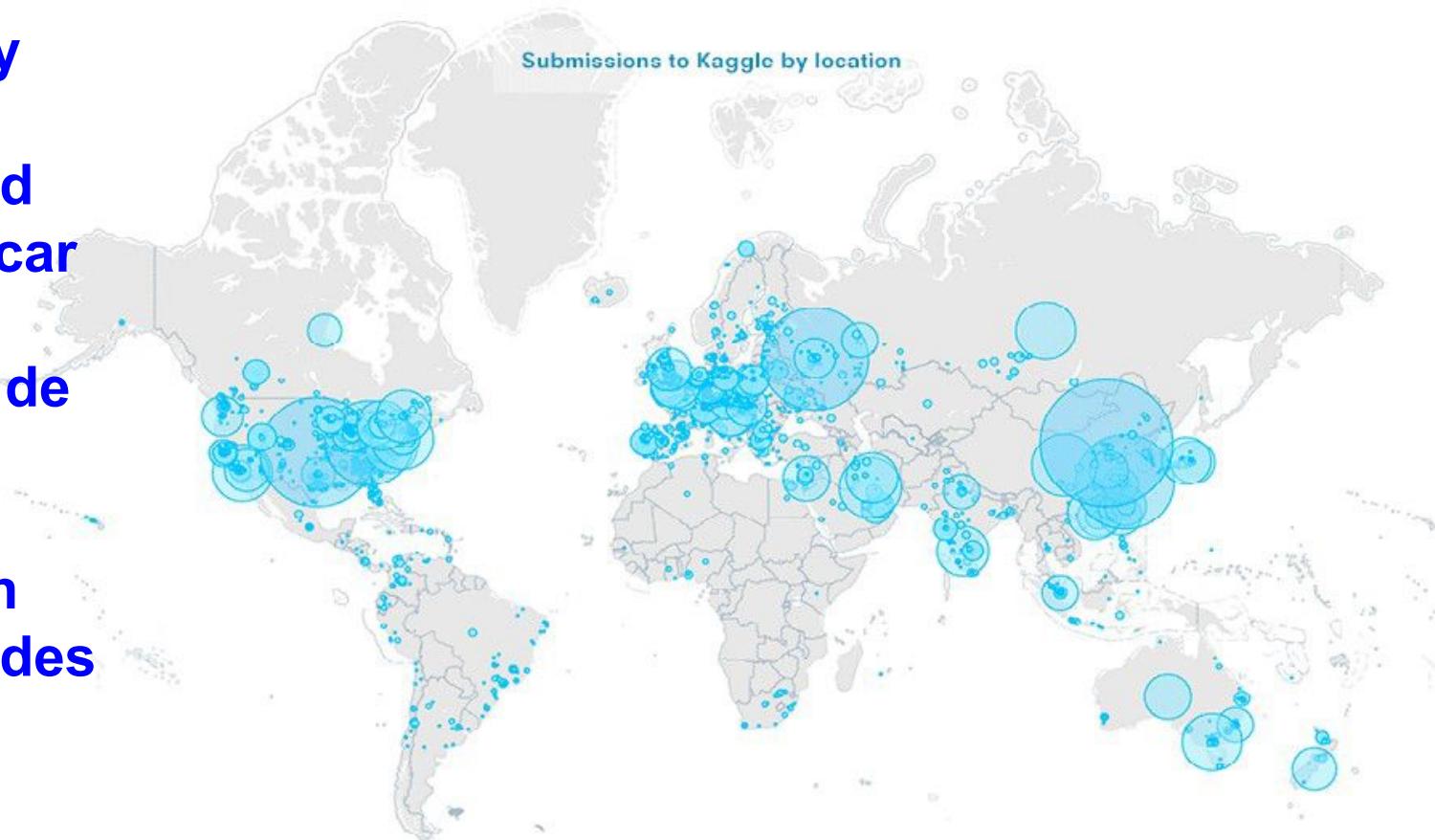
# Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, **KAGGLE**

---

## **Kaggle: The Home of Data Science**

Es una muy buena oportunidad para practicar en la resolución de problemas reales y la adquisición de habilidades en Data Science.



# Herramientas, Lenguajes, Kaggle

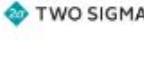
... y un buen enlace para comenzar a practicar, **KAGGLE**

## Kaggle: The Home of Data Science

12 active competitions      **12 Febrero 2017**      Active Competitions

Active All Entered      All Categories      Sort By Prize

All Competitions     

	<b>Data Science Bowl 2017</b> Can you improve lung cancer detection? Featured • 2 months to go • 442 kernels	\$1,000,000 1,079 teams
	<b>The Nature Conservancy Fisheries Monitoring</b> Can you detect and classify species of fish? Featured • 2 months to go • 258 kernels	\$150,000 1,475 teams
	<b>Dstl Satellite Imagery Feature Detection</b> Can you train an eye in the sky? Featured • 23 days to go • 133 kernels	\$100,000 243 teams
	<b>Two Sigma Financial Modeling Challenge</b> Can you uncover predictive value in an uncertain world? Featured • 17 days to go • 192 kernels	\$100,000 1,797 teams

# Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

## Kaggle: Go from Big Data to Big Analytics

12 Febrero 2017

	<b>Two Sigma Connect: Rental Listing Inquiries</b> How much interest will a new rental listing on RentHop receive? <a href="#">Recruitment</a> · 2 months to go · 143 kernels	<b>Jobs</b> 292 teams
	<b>Dogs vs. Cats Redux: Kernels Edition</b> Distinguish images of dogs from cats <a href="#">Playground</a> · 18 days to go · 183 kernels	1,009 teams
	<b>Transfer Learning on Stack Exchange Tags</b> Predict tags from models trained on unrelated topics <a href="#">Playground</a> · A month to go · 93 kernels	263 teams
	<b>March Machine Learning Mania 2017</b> Predict the 2017 NCAA Basketball Tournament <a href="#">Playground</a> · A month to go · 15 kernels	<b>Swag</b> 156 teams
	<b>House Prices: Advanced Regression Techniques</b> Sold! How do home features add up to its price tag? <a href="#">Playground</a> · 17 days to go · 860 kernels	4,264 teams
	<b>Leaf Classification</b> Can you see the random forest for the leaves? <a href="#">Playground</a> · 16 days to go · 376 kernels	1,465 teams

# Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, **KAGGLE**

## **Kaggle: Go from Big Data to Big Analytics**

**12 Febrero 2017**



### **Digit Recognizer**

Classify handwritten digits using the famous MNIST data

[Getting Started](#) · 3 years to go · 2,318 kernels

1,425 teams



### **Titanic: Machine Learning from Disaster**

Predict survival on the Titanic using Excel, Python, R & Random Forests

[Getting Started](#) · 3 years to go · 5,814 kernels

5,870 teams

# Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, **KAGGLE**

## Kaggle: The Home of Data Science

9665407401    Digit Recognizer  
3134727121  
1742351244

Classify handwritten digits using the famous MNIST data  
1,425 teams · 3 years to go

[Overview](#)   [Data](#)   [Kernels](#)   [Discussion](#)   [Leaderboard](#)   [More](#)

0000000000000000  
1111111111111111  
2222222222222222  
3333333333333333  
4444444444444444  
5555555555555555  
6666666666666666  
7777777777777777  
8888888888888888  
9999999999999999

**MNIST data**

# Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

## Kaggle: Go from Big Data to Big Analytics

<https://www.kaggle.com/c/titanic>

The screenshot shows the Kaggle competition page for 'Titanic: Machine Learning from Disaster'. The page features a large background image of the Titanic ship at night. At the top left, there's a competition icon and the text 'Getting Started Prediction Competition'. The main title 'Titanic: Machine Learning from Disaster' is prominently displayed in large, bold, white font. Below it, a subtitle reads 'Predict survival on the Titanic using Excel, Python, R & Random Forests'. On the left side, there's a 'Kaggle' logo and the text 'Kaggle · 5,870 teams · 3 years to go'. Below this, a navigation bar includes tabs for 'Overview' (which is underlined in blue), 'Data', 'Kernels', 'Discussion', 'Leaderboard', 'More', and a large blue 'Submit Predictions' button. At the bottom, there's a horizontal menu with links for 'Description' (underlined in blue), 'Evaluation', 'Frequently Asked Questions', 'Getting Started', and 'Submission Instructions'.

# Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

[kaggle](#) in Class

[Sign Up](#)



## Predicting unemployment in the Great Recession

Monday, September 23, 2013      Knowledge • 84 teams      Friday, December 6, 2013

[Learn about hosting](#)

**Academic Machine Learning Competitions**

**Theory, meet practice.**

Kaggle hosts free projects for hundreds of universities around the globe. Engage students with an opportunity to apply machine learning to real problems.

Berkeley      University of California      UCI  
Columbia      Cornell      Erasmus University  
Harvard      University of Melbourne      Michigan  
University of Oxford      Stanford      University of Toronto

**Dashboard**

- Home
- Data
- Make a submission

**Information**

- Description
- Evaluation
- Rules
- Prizes

**Forum**

**Leaderboard**

- Public
- Private

**Leaderboard**

1. GobbleGobble
2. Bryan
3. Jim Monteleone
4. DMFK
5. mudkipz
6. Steven Balough

**Competition Details** » [Get the Data](#) » [Make a submission](#)

 **This competition is private-entry.** You can view but not participate.

### Stats 202 Prediction Challenge.

**Who can participate?**

This challenge is restricted to students enrolled in [Stats 202](#) at Stanford University in the fall of 2013.

**The data**

We will use the National Longitudinal Study of Youth (NLSY79). This landmark study published by the National Bureau of Labor Statistics has followed a cohort of baby boomers since 1979, recording various data on employment, education, poverty, family, attitudes, etc.

**What is your goal?**

Based on interviews spanning from 1979 to 2006, you will be tasked with predicting the number of weeks that each person was unemployed during 2010, at the peak of the Great Recession.

# Índice



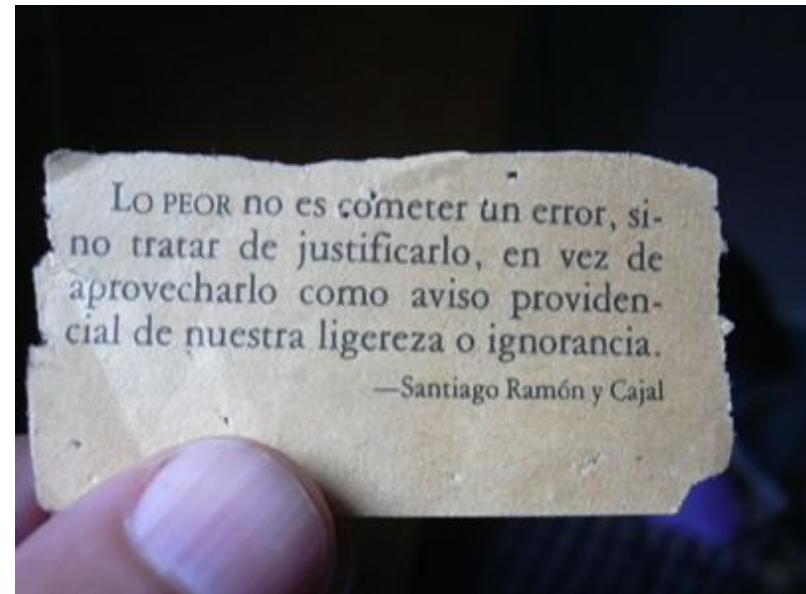
- ¿Qué es la Ciencia de Datos?
- El poder de los datos y su impacto en nuestra sociedad
- Herramientas y Lenguajes en Ciencia de Datos.  
Repositorio de Kaggle
- **Comentarios Finales**

# Comentarios Finales

---

## **Hay que evitar los errores comunes**

- Aprender de cosas que no son ciertas
  - Patrones que no representan ninguna regla subyacente
  - Datos que no reflejan lo relevante
  - Datos con un nivel de detalle erróneo
  
- Aprender cosas ciertas, pero inútiles
  - Aprender información ya conocida
  - Aprender cosas que no se pueden utilizar



**Hay que obtener  
conocimiento útil**

# Comentarios Finales

**Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”**

## Oportunidades

**La demanda de profesionales formados en Ciencia de Datos y Big Data es enorme.**

**Se estima que la conversión de datos en información útil generará un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos.**

**España necesitará para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y Big Data.**



The screenshot shows the header of the EL PAÍS Economía website. The main title 'EL PAÍS' is at the top left, followed by 'ECONOMÍA'. Below the title are several navigation links: 'PORTADA', 'INTERNACIONAL', 'POLÍTICA', 'ECONOMÍA', 'EMPRESAS', 'MERCADOS', 'BOLSA', 'FINANZAS PERSONALES', 'VIVIENDA', 'TECNOLOGÍA', 'ESTÁ PASANDO', 'Multa a la banca', 'Revuelto en Hacienda', 'Eléctricas y renovables', and 'Paro industrial'.

### El maná de los datos

■ La conversión de datos en información útil para las empresas generará un mercado de 132.000 millones de dólares en 2015. La herramienta 'big data' sacará del mercado a quien no la use.

SUSANA BLÁZQUEZ | Madrid | 29 SEP 2013 - 01:00 CET

Archivado en: Citigroup Cap Gemini Sogeti SAP Oracle ING Bank BBVA Mapfre Bases de datos IBM Telefónica Aplicaciones informáticas Tecnología Empresas Programas informáticos Economía



[http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725\\_938376.html](http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725_938376.html)

# Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”

## Oportunidades (en España)

[http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde\\_4377072\\_member\\_5811011886832984067#!](http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde_4377072_member_5811011886832984067#!)

**España necesitará 60.000 profesionales de Big Data hasta 2015**

22 octubre, 2013 Eventos 18



España necesitará 60.000 profesionales de Big Data hasta 2015

Toledo.

“España va a necesitar alrededor de sesenta mil profesionales del Big Data de aquí a 2015”, así lo ha asegurado Francisco Javier Antón, Subdirector General de Tecnologías del Ministerio de Educación, Cultura y Deportes en una mesa redonda sobre beneficio y aplicación de Big Data en pymes, moderada por Daniel Tapias de **Sigma Technologies**, celebrada durante el 4º Congreso Nacional de CENTAC de

**“Existe una demanda mundial para formar a 4,4 millones de profesionales de la gestión Big Data desde ingenieros, gestores y científicos de datos”, comenta Antón. Sin embargo, “las empresas todavía no ven en el Big Data un modelo de negocio”, lamenta. “Solo se extrae un 1% de los datos disponibles en la red”, añade. “Hace falta formación y concienciación.**

# **Sistemas Inteligentes para la Gestión de la Empresa**

**2016 - 2017**



- **Tema 1. Introducción a la Ciencia de Datos**
- **Tema 2. Depuración y Calidad de Datos.**  
**Preprocesamiento de datos**
- **Tema 3. Análisis Predictivo para la Empresa**
- **Tema 5. Análisis de Transacciones y Mercados**
- **Tema 4. Modelos avanzados de Analítica de Empresa**
- **Tema 6. Big Data**
- **Tema 7. Aplicaciones de la Ciencia de Datos en la Empresa**