

Statistics, Data Mining, Big Data, Data Science: More than a play of names

Edgar Acuna
Department of Mathematical Science
University of Puerto Rico-Mayaguez

E-mail: edgar.acuna@upr.edu , eacunaf@gmail.com

Website: academic.uprm.edu/eacuna

Outline

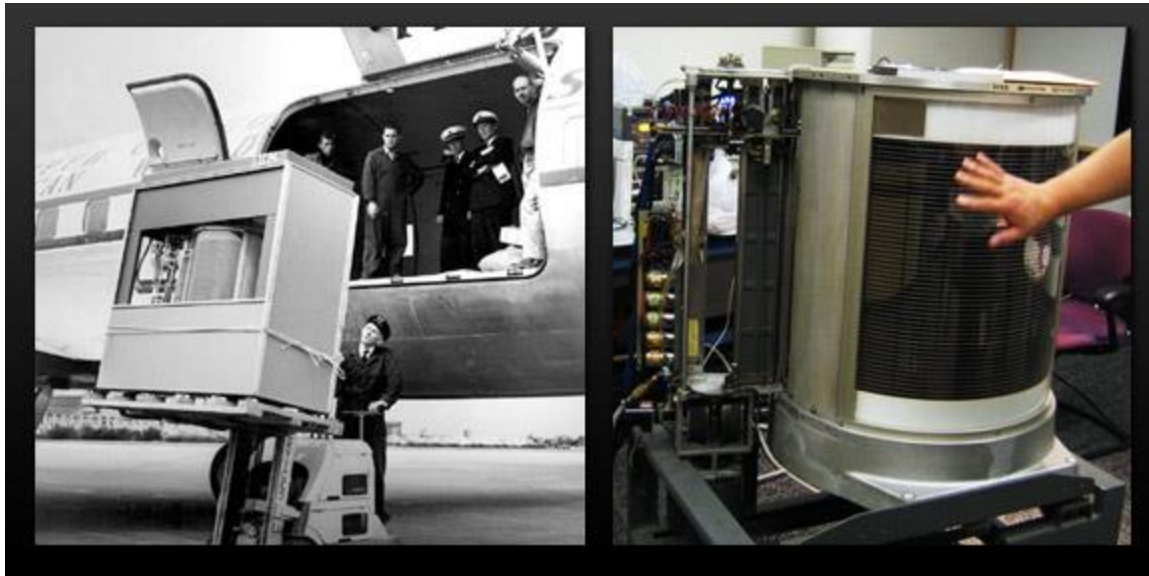
- I. Introduction
- II. Data Mining
- III. From Statistics to Data Mining
- IV. From STAT/DM to Big Data
- V. From Statistics to Data Science

Introduction

The mechanisms for automatic recollection of data and the development of databases technology has made possible that a large amount of data can be available in databases, data warehouses and other information repositories.

Nowdays, there is the need to convert this data in knowledge and information.

The first hard drive, 1956



IBM 350, had the size of two refrigerators and about 3.75MB of storage. It weighted over a ton. Approximately price of 50,000 dollars. Today, Seagate sells a 2TB hard drive it weights only .33 pounds. It costs around \$100.

Size (in Bytes) of datasets

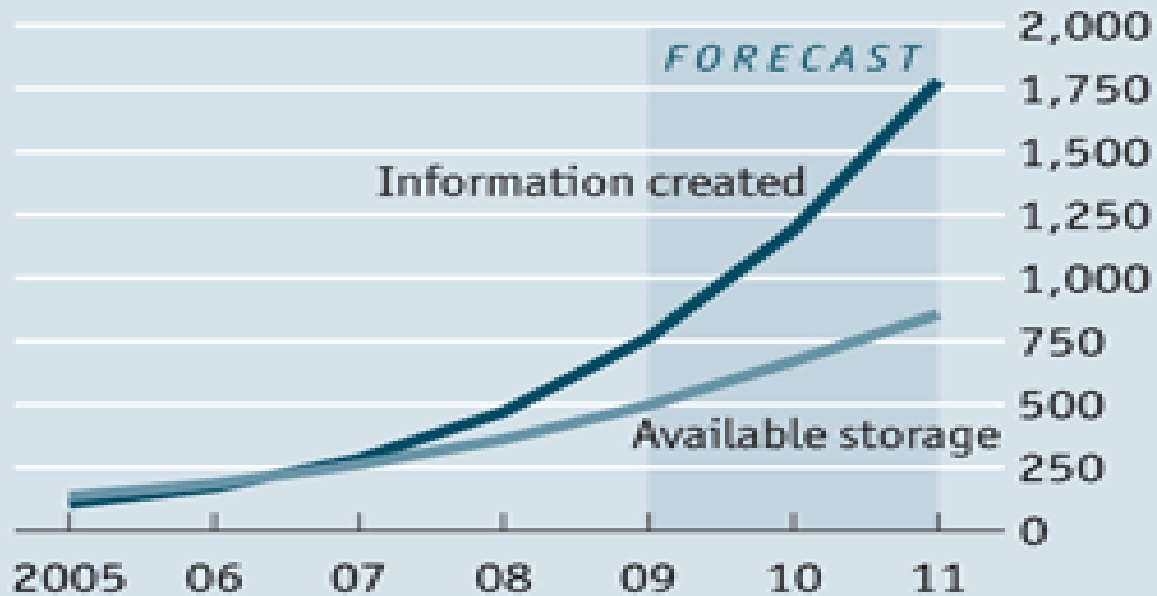
Description	Size	Storage Media
Very small	10^2	Piece of paper
Small	10^4	Several sheets of paper
Medium	10^6 (megabyte)	Floppy Disk
Large	10^9 (gigabite)	USB/Hard Disk
Massive	10^{12} (Terabyte)	Hard disk/USB
Super-massive	10^{15} (Petabyte)	File of distributed data
Exabyte(10^{18}), Zettabytes(10^{21}), Yottabytes(10^{24})		

The economist, February 2010

Overload

1

Global information created and available storage
Exabytes



Source: IDC

Data Revolution

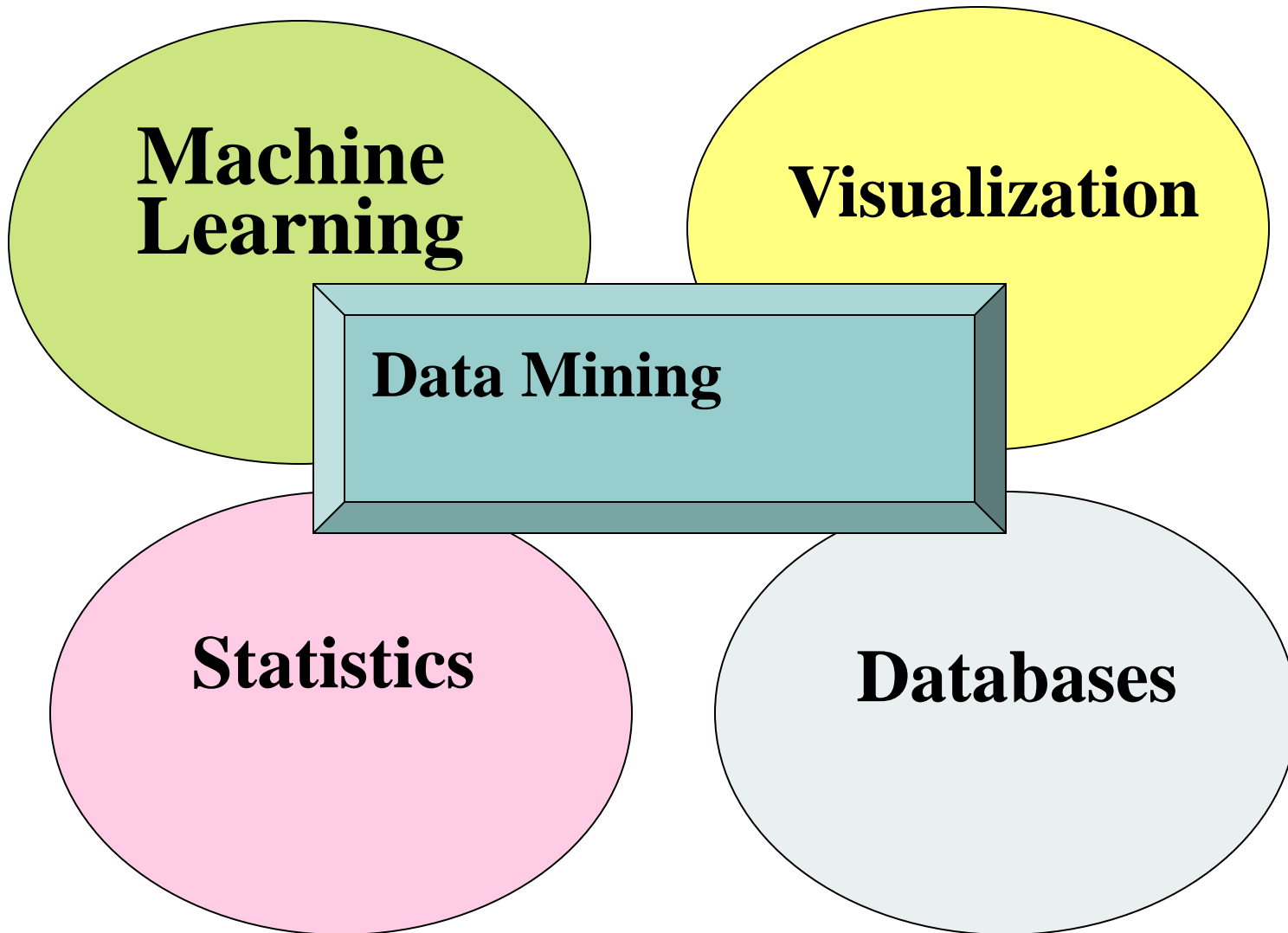
Year	Digital	Analog	Amount
2000	25%	75%	2 Exabytes
2007	93%	7%	300 Exaby
2013	98%	2%	1,200 Exaby

Source: Viktor Mayer-Schönberger and Kenneth Cukier: Big Data: A Revolution that will Transform how We Live, Work and Think (2013)

What is Data Mining?

- It is the process of extracting valid knowledge/information from a very large dataset. The knowledge is given as patterns and rules that are non-trivial, previously unknown, understandable and with a high potential to be useful.
- Other names: Knowledge discovery in databases (KDD), Intelligent Data Analysis, Business Intelligence.
- The first paper in Data Mining: Agrawal et al. Mining Association rules, ACM SIGMOD 1993.

Areas related to Data Mining



Contribution of of each area to Data Mining

- Statistics (~35%): Estimation of prediction models. Assume distribution for the features used in the model. Use of sampling
- Machine learning: (~30 %): Part of Artificial Intelligence. More heuristic than Statistics. Small data and complex models
- Databases: (~20%):large Scale data, simple queries. The data is maintained in tables that are accessed quickly.
- Visualization: (~ 5%).It can be used in either the pre-processing o post-processing step of the KDD process.
- Other Areas: (~10%): Pattern Recognition, Expert Systems, High Performance Computing.

Data Mining Applications

Science: Astronomy, Bioinformatics (Genomics, Proteonomics, Metabolomics), drug discovery.

Business: Marketing, credit risk, Security and Fraud detection,

Government: detection of tax cheaters, anti-terrorism.

Text Mining:

Discover distinct groups of potential buyers according to a user text based profile. Draw information from different written sources (e-mails).

Web mining: Identifying groups of competitors web pages. Recomemder systems(Netflix, Amazon, Ebay)

Data Mining: Type of tasks

- **Descriptive:** General properties of the database are determined. The most important features of the databases are discovered. It includes outlier detection, visualization, clustering and association rules.

Predictive: The collected data is used to train a model for making future predictions. Never is 100% accurate and the most important matter is the performance of the model when is applied to future data. It Includes Regression and Classification

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data.
- Quality of Data
- Privacy Data
- **Streaming Data**

Software for Data Mining

- **Open source**
- R (cran.r-project.org). Related to Statistics (38.5% users, Kdnuggets June 2014).
- RapidMiner (<http://rapidminer.com>). (44.2%) Related to the Database community. **
- Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) (17.0%): Related to Machine Learning. Written in Java.
- Python, iPython (19.5%). Hadoop(12.7%).
- **Comercial:** SAS (10.8%), XLMiner(25%), Microsoft SQL(10.5%), KNIME (15.0%), Oracle(2.2%).

A quick comparison of R and iPython Notebook

- **Dataset:** Titanic: Machine learning from disaster (www.kaggle.com). The goal is to apply the tools of ML to predict which passengers survive the tragedy.
- S(Chambers, 1983), R (Ihaka nad Gentleman, 1994), Rstudio (2011).
- Python(1989, G. Van Rossum, Netherlands), iPython (2011, F. Perez, UC Berkeley)
- Pandas (2012) is a software library written for the Python programming language for data manipulation and analysis.
- Rpython(Gil Bellosta, 2014)

Google Trends for Data Mining

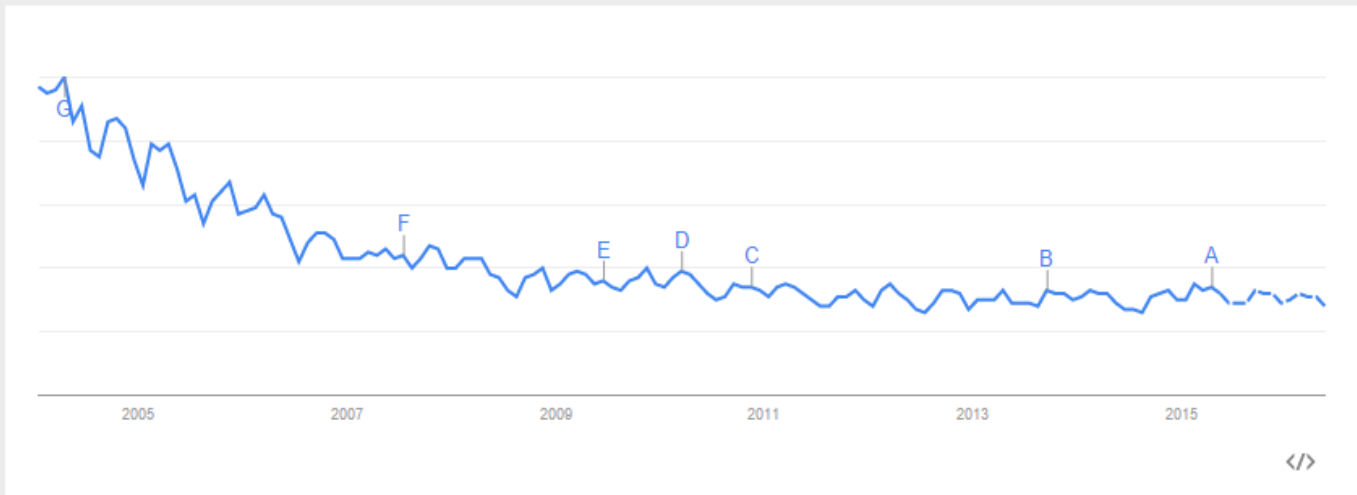
Compare Search terms ▾

Data Mining
Search term

+ Add term

Interest over time ?

News headlines Forecast ?



From STAT/DM Big Data[1]

February 1977, L. Breiman organized the conference “Analysis of large complex data sets”. Sponsored by ASA and IMS Dallas, USA.

1994 COMPSTAT, Proceedings of Computational Statistics. Part I. Treatment of “huge” Data sets.

May 1997. The 29th Symposium on the Interface (Houston, TX) “ Data Mining and the analysis of large data sets”.

October 1997, M. Cox y D. Ellsworth used the term “big data” in a conference on visualization organized by the IEEE.

1998, Workshop on Massive data sets. Committee on Applied and Theoretical Statistics. NRC, USA.

April 1998, John Massey, chief scientist of SGI, presented a paper “Big data... and the next wave of Infra-stress” in a meeting of the USENIX.

From STAT/DM to Big Data[2]

In February 2001, Doug Laney, analyst for the Meta Group, defined data challenges and opportunities as being three-dimensional: increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).

September 2008, the scientific magazine Nature published a special edition about “big data”.

May 2011, Manyika, J. et al. researchers of the McKinsey Global Institute published the report: “Big data: the next frontier for innovation, competition and productivity”.



Big Data: Examples

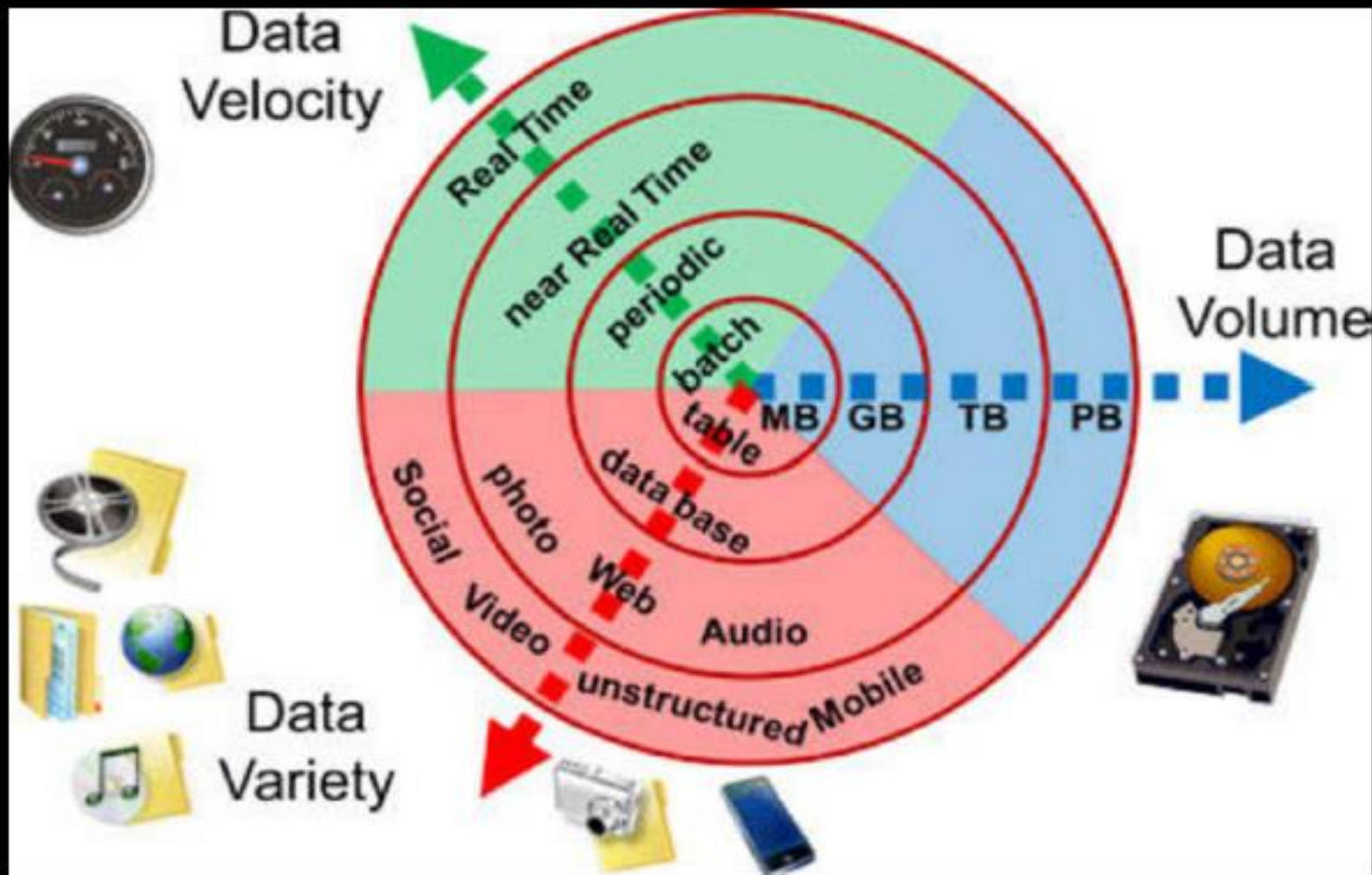
- The Large Hadron Collider (LCH) storages around 25 Petabytes of sensor data per year.
- In 2010, the ATT's database of calling records was of 323 Terabytes.
- El 2010, Walmart handled 2.5 Petabytes of transactions.
- In 2009, there was 500 exabytes of information on the internet.
- In 2011, Google searched in more than 20 billions of web pages. This represents aprox. 400 TB.
- In 2013, it was announced that the NSA's Data Center at Utah will storage up to 5 zettabytes (5,000 exabytes).

Big Data: Definition

In 2012, Doug Laney, working now at the Gartner Group, updated its definition of Big Data as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.

Big data sizes are constantly changing. In 2002, maybe 100GB, in 2012 perhaps from 10TB to many petabytes of data in a single data set.



ICSA Bulletin, Jan 2014

Google trends for Big data

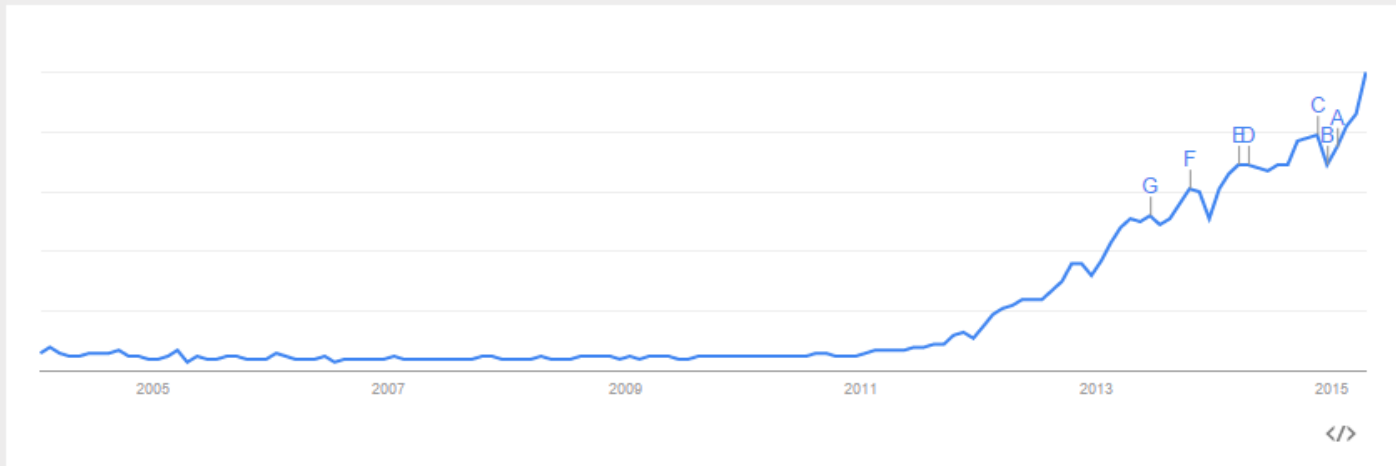
Comparar Términos de búsqueda ▾

big data
Término de búsqueda

+ Añadir término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



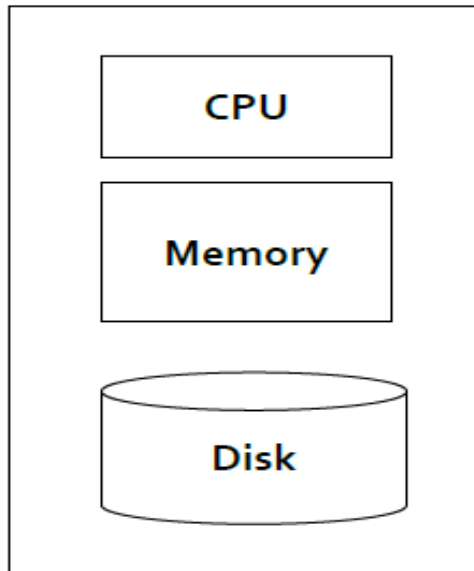
Big data is ...

Big data is like teenage sex:

everyone talks about it, nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it...

Dan Ariely (June 2013), Professor of Psychology and Behavioral Economics. Duke University.

Single Node Architecture



Machine Learning, Statistics

"Classical" Data Mining

Motivation: Google Example

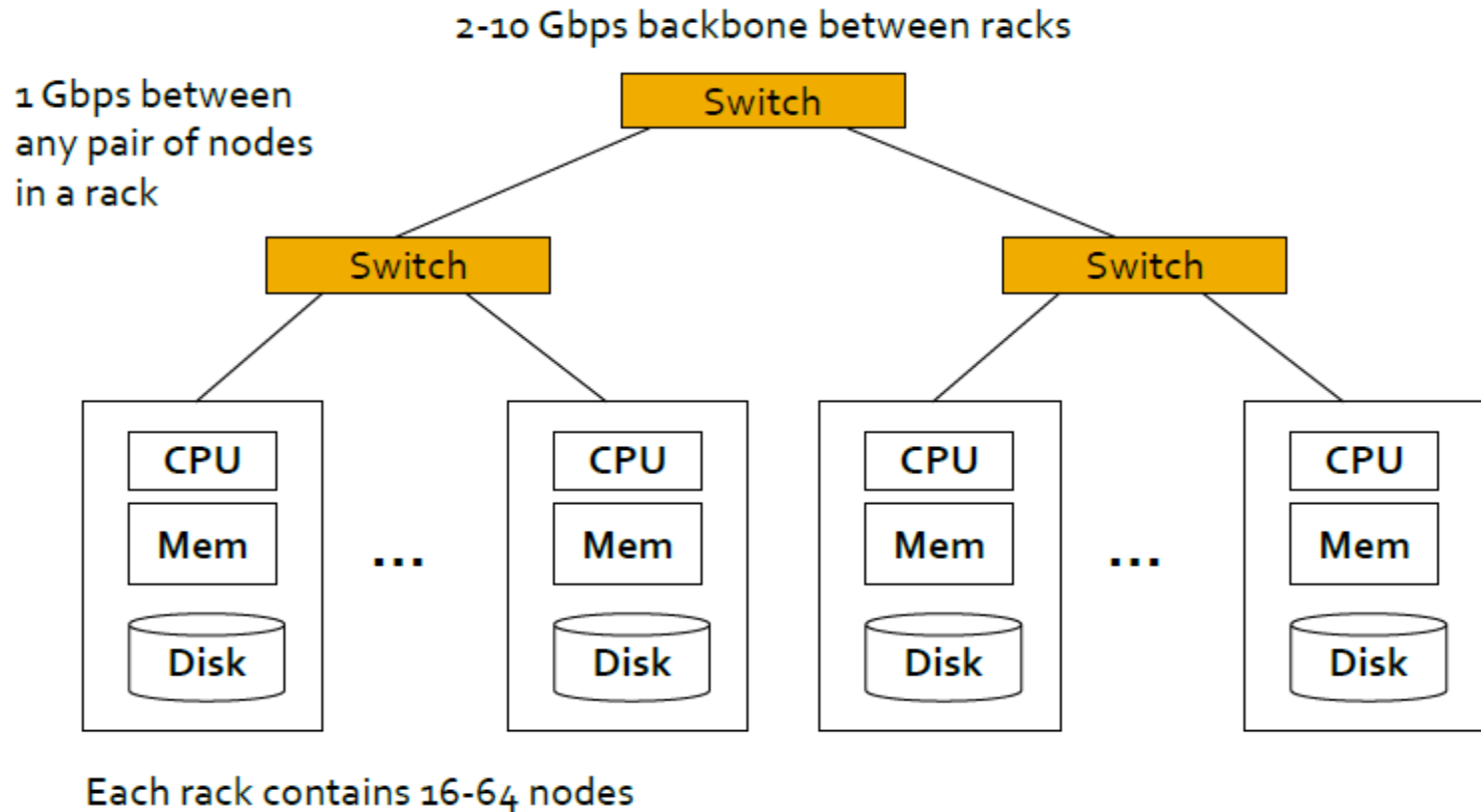
Google searches in more than 20 billion of webpages x20KB= 400+ TB. One computer reads with a speed of 30-35 MB/sec from disk. It will need approx 4 months to read the web.

It will be necessary approx 1000 hard drives to read the web
It will be necessary even more than that to analyze the data

Today a standard architecture for such problems is being used. It consists of

- a cluster of commodity Linux nodes
- commodity network (ethernet) to connect them

Cluster Architecture



Challenges in large-scale computing for data mining

How to distribute the computation?

How to write down distributed program easily?

Machines fail!.

One computer may stay up three years (1000 days)

If you have 1000 servers , expect to loose 1 per day

In 2011, it was estimated that Google has 1 million of computers, so 1000 servers can fail everyday

Using R in parallel and distributed computation

RMPI: Package to run R on MPI

Snow: Package to run R on MPI

Gputools: Package to run R on CUDA

Rhadoop: Interface between R y Hadoop. It includes rhbase, rhdfs, rmr. Developed by Revolution Analytics, a company bought by Microsoft in March 2015.

Between 2003-2006, Elio Lozano and Edgar Acuna wrote several articles on aplicacion of R in parallel and distributed computation to calculate metaclassifiers, metaclustering, boosting, bagging, kernel density estimation, ensembles and outlier detection.

What is Hadoop?

- In 2004, J. Dean and S. Ghemawat wrote a paper explaining Google's MapReduce, a programming model and a associated infrastucture for storage of large data sets (file system) called Google File System (GFS).
- GFS is not open source.
- In 2006, Doug Cutting at Yahoo! , created a open source GFS and called it Hadoop Distributed File System (HDFS). In 2009, he left to Cloudera.
- The software framework that supports **HDFS**, MapReduce and other related entities is called the project Hadoop or simply Hadoop.
- Hadoop is distributed by the Apache Software Foundation.

Hadoop

Hadoop includes:

Distributed Files System(HDFS) –distributes data
Map/Reduce-distributes application

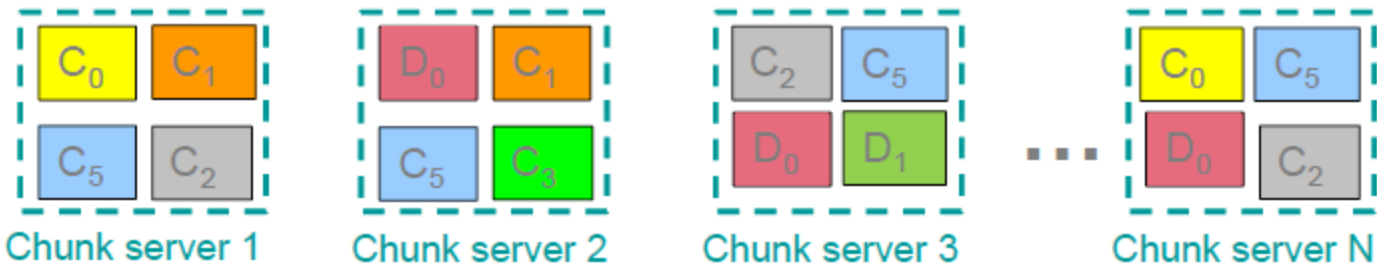
It is written in Java

Runs on

- Linux, MacOS/X, Windows, and Solaris
- Uses commodity hardware

Distributed File System

- **Reliable distributed file system**
- Data kept in “chunks” spread across machines
- Each chunk **replicated** on different machines
 - Seamless recovery from disk or machine failure



Bring computation directly to the data!

Chunk servers also serve as compute servers

HDFS Architecture

- Provides
 - Automatic Parallelization and Distribution
 - Fault Tolerance
 - I/O Scheduling
 - Monitoring and Status Updates

MapReduce

Map-Reduce is a programming model for efficient distributed computing

It works like a Unix pipeline:

-cat input |grep | sort |unique -c |cat > output

- **Input | Map | Shuffle & Sort |Reduce |Output**

Efficient because reduces seeks and the use of pipeline

The WorkFlow

- Load data into the Cluster (HDFS writes)
- Analyze the data (MapReduce)
- Store results in the Cluster (HDFS)
- Read the results from the Cluster (HDFS reads)

Hadoop Non-Java Interfaces

Hadoop streaming: C++,Python, perl,ruby.

Rhadoop (R and Hadoop),

Weka(Mark Hall is working on that),
Radoop (Rapidminer and hadoop, comercial)

Hadoop Pipes: (C++) It not recommendable

Where can I run hadoop?

In your personal computer using the single node hadoop cluster. If you have windows install a virtual machine for running Ubuntu(See Michael Noll's website)

Free:

At the Gordon cluster of the SDSC (1024 nodes, each node has 16 cores) and the Stampede cluster of TACC (6400 nodes, each node has 16 cores). Access can be gained through the XSEDE project.

Non-Free, but not too expensive

Amazon Elastic Compute Cloud (EC2)

Who is using Hadoop?

Yahoo
Facebook
Amazon
Google
IBM
Netflix
Ebay
LinkedIn
Twitter

Example of Performance:Hadoop for Feature Selection(Gomez & Acuna,2014)

The Goal is to select the best features for prediction in the Covtype dataset. Covtype has 581,102 instances and 54 predictors and 7 classes.

	Before Hadoop	After Hadoop
Time	17 hours	5m23s
Language	R and c++	Java

Mahout

Scalable Machine learning library that runs on Hadoop
Includes some algorithms for:

Recommendation mining

Clustering

Classification

Finding frequent itemsets

Some people thinks that the algorithms included in Mahout are not programmed in a optimal way.

Google Flu Trends: A Failure of Big data

In February 2009, researchers from Google tried to predict the spread of the flu epidemic building a algorithmic model. Google's algorithm mined five years of web logs, containing hundreds of billions of searches, and created a predictive model utilizing 45 search terms that proved to be a more useful and timely indicator of flu than government statistics with their natural reporting lags. (Viktor Mayer-Schönberger and Kenneth Cukier: Big Data, 2013).

During the first two years the model was given acceptable predictions. But, in 2011 the model starts to fail beginning with a 50% up to 92% of error in the predictions. (Science, March 2014). The GFT has over-estimated the prevalence of flu for 100 out of the last 108 weeks; it's been wrong since August 2011.

In May 2011, Google launched the Dengue Google Trends

Other computational tools for Big data

Spark developed at the AMP lab of the UC Berkeley. Programming must be done in Scala, Java, Python and R. It is Free. AMPLAB claims the Spark beats MapReduce on several algorithms.

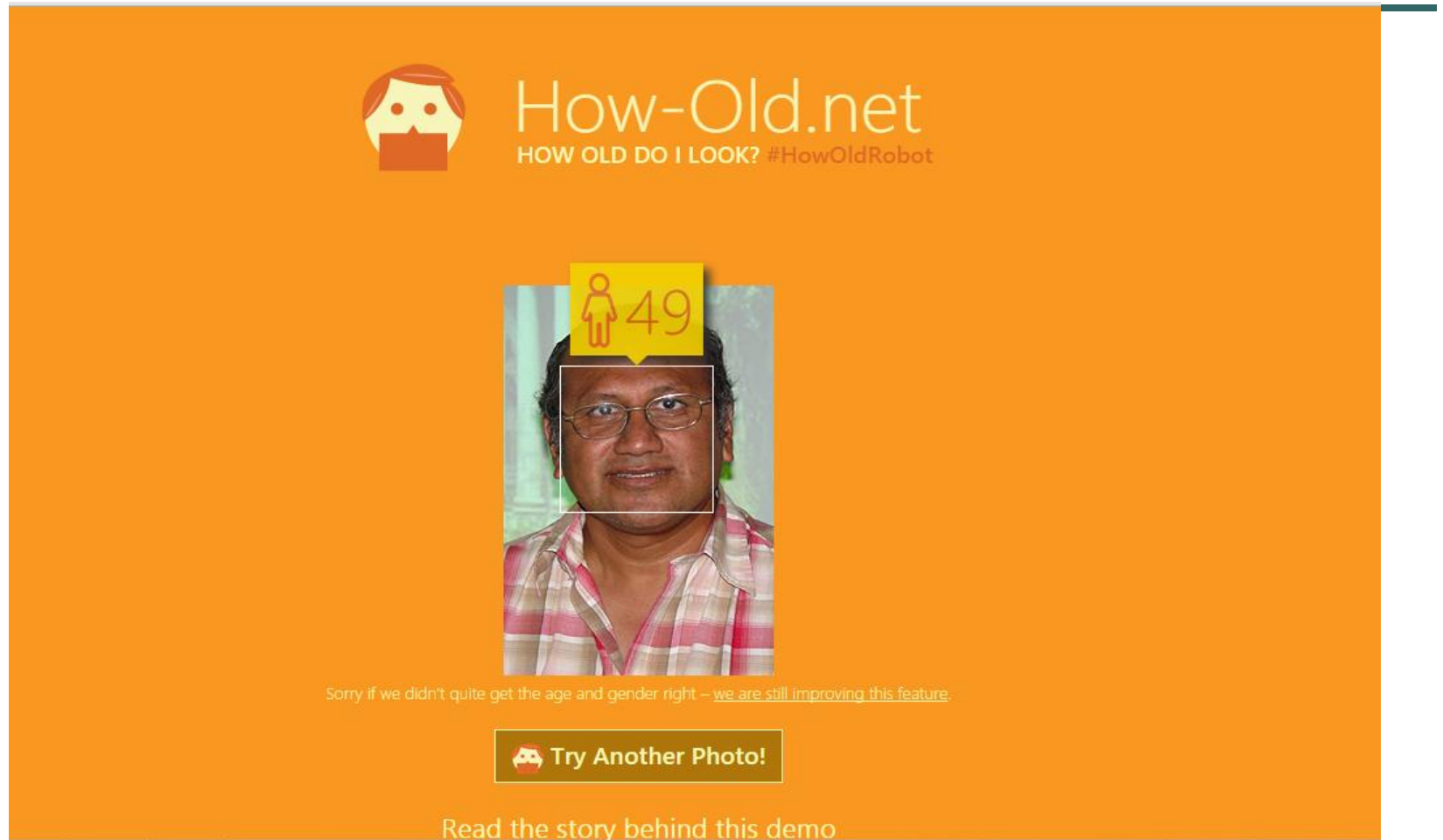
Watson Analytics from IBM (SPSS's owner). Free for academia.

Azure Machine Learning from Microsoft Research, software built on Linux. Free for academia.

According to kdnuggets, Azure is better developed than Watson

Amazon Web Services.

Face detection: Azure Machine Learning



The image shows a screenshot of the How-Old.net website. At the top left is a logo of a stylized face with a beard. To its right is the text "How-Old.net" and "HOW OLD DO I LOOK? #HowOldRobot". In the center, a photograph of a man with glasses is shown with a white bounding box around his face. A yellow speech bubble above the face contains a person icon and the number "49". Below the photo, there is a message: "Sorry if we didn't quite get the age and gender right - we are still improving this feature." At the bottom of the page, there is a button that says "Try Another Photo!" and a link that says "Read the story behind this demo".

The absence of statisticians in Big Data activities

- May 2013, White House Data Partners Workshop-0/19 statisticians.
- April 2014, National Academy of Science Big Data Workshop-2/16 speakers statisticians.
- In 2012, a NSF formed a working group of 100 members on Big data. None statistician were included.
- 2012, NIH BD2K executive committee composed of 18 members does not include any statistician.

More examples in <http://simplystatistics.org>

Layman's perception of statisticians

A statistician is someone that counts numbers (similar to an accountant).

A statistician makes tables and graphs to report and/or summarize information.

A statistician computes averages percentages and standard deviations to summarize information.

A statistician helps politicians and the government to fool the people (Lies , Dammed lies and Statistics).

More details in J. Wu's presentation(1997).

From Statistics to Data Science[1]

John Tukey: “For a long time I have thought I was a statistician interested in inferences... I have come to feel that my central interest is in data analysis “. (Annals of Mathematical Statistics, 1962).

In 1977, Tukey published “Exploratory Data Analysis”. The data is analyzed to summarize their main characteristics using mostly visual methods: Boxplots, stem-and-leaf, Parallel coordinates, Multidimensional Scaling, etc.

In 1996, the International Federation of Classification Societies (IFCS) held a conference titled “Data Science, Classification and Related Methods”.

From Statistics to Data Science [2]

J. Wu (1997). In his acceptance lecture as a chairman of the Statistics Department at University of Michigan proposed to change the name from Statistics to Data Science.

Prof. Wu also proposed to rename statisticians as data scientists [But Stigler's law will apply here].

Stigler's law of eponymy". *Transactions of the New York Academy of Sciences* **39**: 147–58(1980).

Wu also proposed that statisticians should get involved with large and complex data.

From Statistics to Data Science[3]

J. H. Friedman proposed that Statistics should embrace Data Mining as a sub-discipline.

“The role of the statistics in the data revolution”.
Stanford University Tech Report (2000).

- In 2001 appeared the Breiman’s paper. “Statistical Modelling: The two cultures”. He states; “There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.

From Statistics to Data Science[4]

The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

Algorithmic modeling can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

From Statistics to Data Science[5]

W. Cleveland (2001). Proposed the creation of the field of Data Science, similar to multi-disciplinary statistics. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics".

In 2002 starts the publication of the "Data Science Journal" and the "Journal of Data Science".

En el 2008, D.J.Patil (LinkedIn) y Jeff Hamerbacher (Facebook) introduced the term "data scientist".(Stigler's law was applied here).

Data Science as a discipline

In January 2009, Hail Varian, chief economist of Google stated “I keep saying the sexy job in the next ten years will be statisticians..” (The McKinsey Quarterly).

From 2009 through 2012, several professionals from computer science, social networks and business schools began introducing the idea that Statistics according to Varian was not the traditional one, but a new field called “Data Science”. T.H Davenport and J.H. Patil (October 2012, Harvard Business Review) “Data Scientist: The sexiest job of the 21st century”

Statistician versus Data Scientist (D. Smith Revolution Analytics, 2013)

	Statistician	Data Scientist
Imagen	Baseball/Football	HBR sexiest job of the 21 st century
Works	Solo	In a team
Data	Prepared, clean	Distributed, Messy, Unstructured
Data Size	Kilobytes	Gigabytes
Tools	SAS, Mainframe	R, Python, Hadoop, Linux
Focus	Inference(why)	Prediction(what)
Latency	weeks	seconds
Output	Report	Data Product/data App

Top 10 Data Scientist

- Hillary Mason (MS in CS, bit.ly-Accel Partners)
- John Myles White (PhD Psychology, Facebook)
- Peter Skomoroch (BS in Mathematics, LinkedIn-DataWrangling)
- Gregory Piatetsky (PhD in CS, Kdnuggets)
- D.J Patil (PhD Math, LinkedIn- White House)
- Jeff Hammerbacher (BS Mathematics, Facebook-Cloudera y Columbia University)
- David Smith (PhD Statistics, Revolution Analytics)
- Christopher D. Long (MS in Math/STAT, JMI Equity)*
- Peter Norvig (PhD in CS, Google)
- Ben Lorica (PhD Math, O'Reilly Media)

Academic Programs in Data Science

Undergraduate: University of Rochester, University of San Francisco, College of Charleston, Northern Kentucky University, Case Western University (more than a dozen in USA). Warwick University (UK).

Master: Carnegie Mellon University, Stanford University, Columbia University, Indiana University, NYU, SMU, Virginia. More than 30 in USA, Canada, UK, Spain, France, Germany and the rest of the world.

Latinoamerica: Only Mexico (ITAM, since 2013). Next August in Peru.

Doctorate: Edimburgh University (UK), Columbia. Less than one dozen.

Sample Curriculum of a MS Program in Data Science

Core Courses:

Theory of Statistics

Computational Statistics

Data Mining

Machine Learning

Databases Systems

Data Science Capstone (Data Science Practicum)

Three elective courses: Algorithms, High Performance Computing, Scientific Visualization, Regression, Big data analytics , a business course, a biology course, a chemistry course, a marine science course, etc.

Data Science segun Google Trends

Todo el mundo ▾

2004 - hoy ▾

Todas las categorías ▾

Búsqueda en la Web ▾

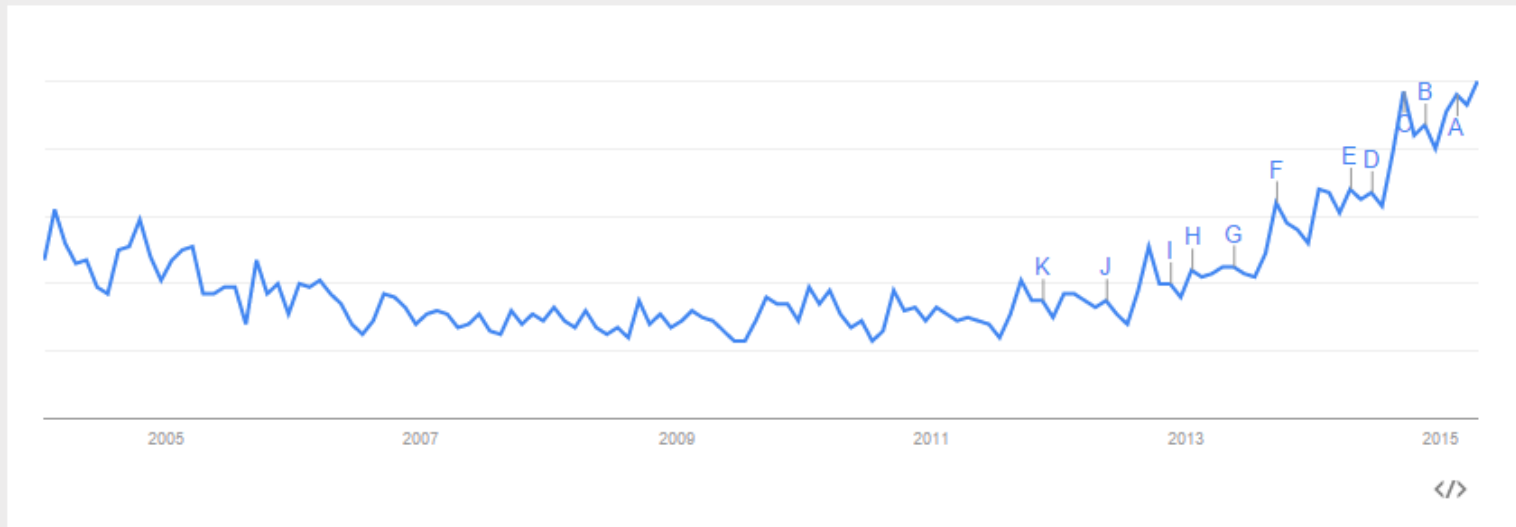
Comparar Términos de búsqueda ▾

data science
Término de búsqueda

+ Añadir término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



Comparacion de Data Mining, Big Data, Bioinformatica y Data Science segun Google Trends

Todo el mundo ▾

2004 - hoy ▾

Todas las categorías ▾

Búsqueda en la Web ▾



Comparar Términos de búsqueda ▾

data sci...

Término de bús...

Big data

Término de bús...

data mi...

Término de bús...

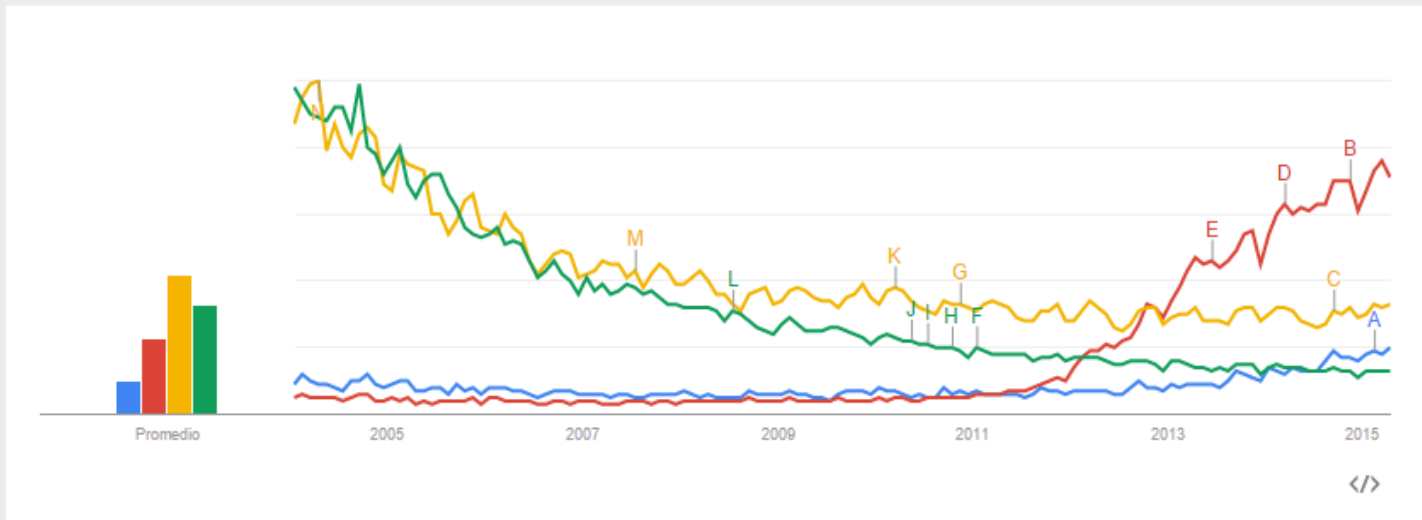
bioinfor...

Término de bús...

+ Añadir término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



On the Future of Big Data

... And soon, in a two- to five-year span, people will say, “The whole big-data thing came and went. It died. It was wrong.” I am predicting that. (Prof. Michael Jordan, EECS/STAT UC Berkeley in a interview for IEEE Spectrum, October 20, 2014)

Talking at Gartner’s Business Intelligence and Information Management summit in São Paulo, Feinberg said “The term ‘big data’ is going to disappear in the next two years, to become just ‘data’ or ‘any data.’ But, of course, the analysis of data will continue.” (dataconomy.com, May 2014)

On the Future Data Science

Is Data Science same as Statistics using more computer power?
Some people, particularly statisticians consider it so.

Is Data Science same as multidisciplinary modern applied Statistics?
Most of the people consider that this is the case.

Is Data Science the field to analyze only Big data?
Some people, particularly from the CS community consider it so.

Only two of the top ten Business School (Sloan at MIT and Kellogg at NU) do data science. Kellogg has a statistics department inside its business school. But some business schools are trying to rename “business statistics” as “data science”.