MADM

UIB Universitat de les Illes Balears

Jornadas de Análisis de Datos Masivos en Empresa

Big Data: Algoritmos, tecnología y aplicaciones

José Manuel Benítez Sánchez
Depto. Ciencias de la Computación e I.A.
Universidad de Granada

DiCITS lab Distributed Computational Intelligence and Time Series Lab

DECSAI Universidad de Granada

Contenido

- Datos
- Big Data
- Algoritmos
- Paradigma MapReduce
- Tecnologías para Big Data
- Aplicaciones

Big Data en los medios de comunicación

EL MUNDO
Líder mundial en español | Miércoles 04/09/2013. Actu

INTERNET | Campus Party Europa 201
'Es la década d
ahí vendrá la re

Alex 'Sandy' Pentland, durante su ponenci

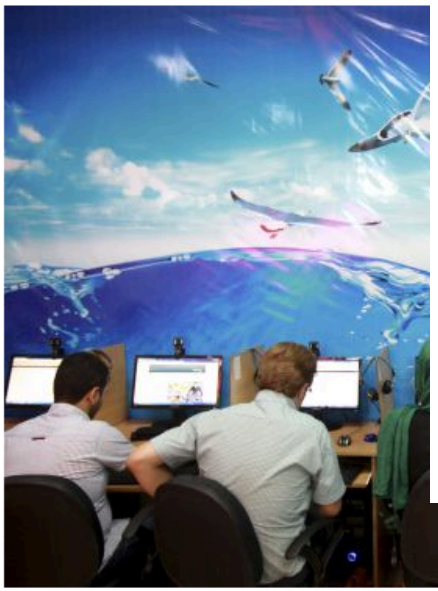
EL PAÍS
ECONOMÍA
ECONOMÍA EMPRESAS MERCADOS BOL

El maná de los datos

- La conversión de datos en información útil par millones de dólares en 2015. La herramienta '

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

Archivado en: Citigroup Cap Gemini Sogeti SA
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

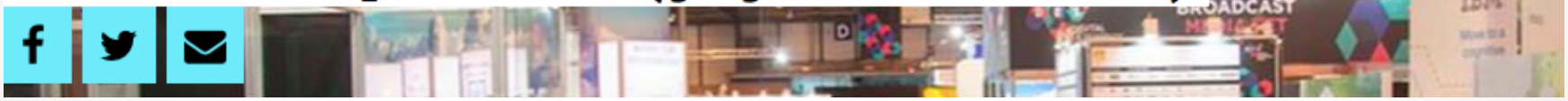
Archivado en: Bases datos Emprendedores Aplicaciones informáticas Empresas
Programas informáticos Economía Informática Industria

http://www.elmundo.es/elmur

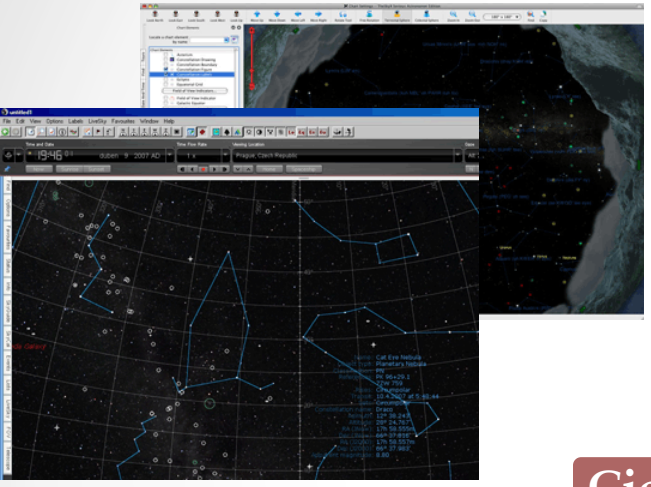
ED economía Digital

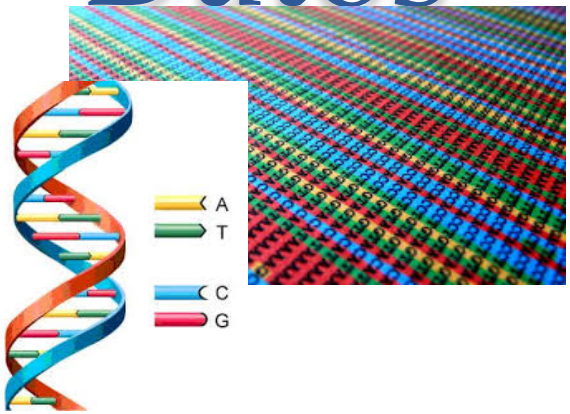
A ESTA HORA Las siete noticias necesarias antes de desconectar

Todos hablan de big data pero nadie sabe lo que es (y ya es hora)



Datos





- Ciencia
- Medicina
- Ciencias Sociales y Humanidades
- Negocio, Comercio e Industria
- Entretenimiento y Ocio

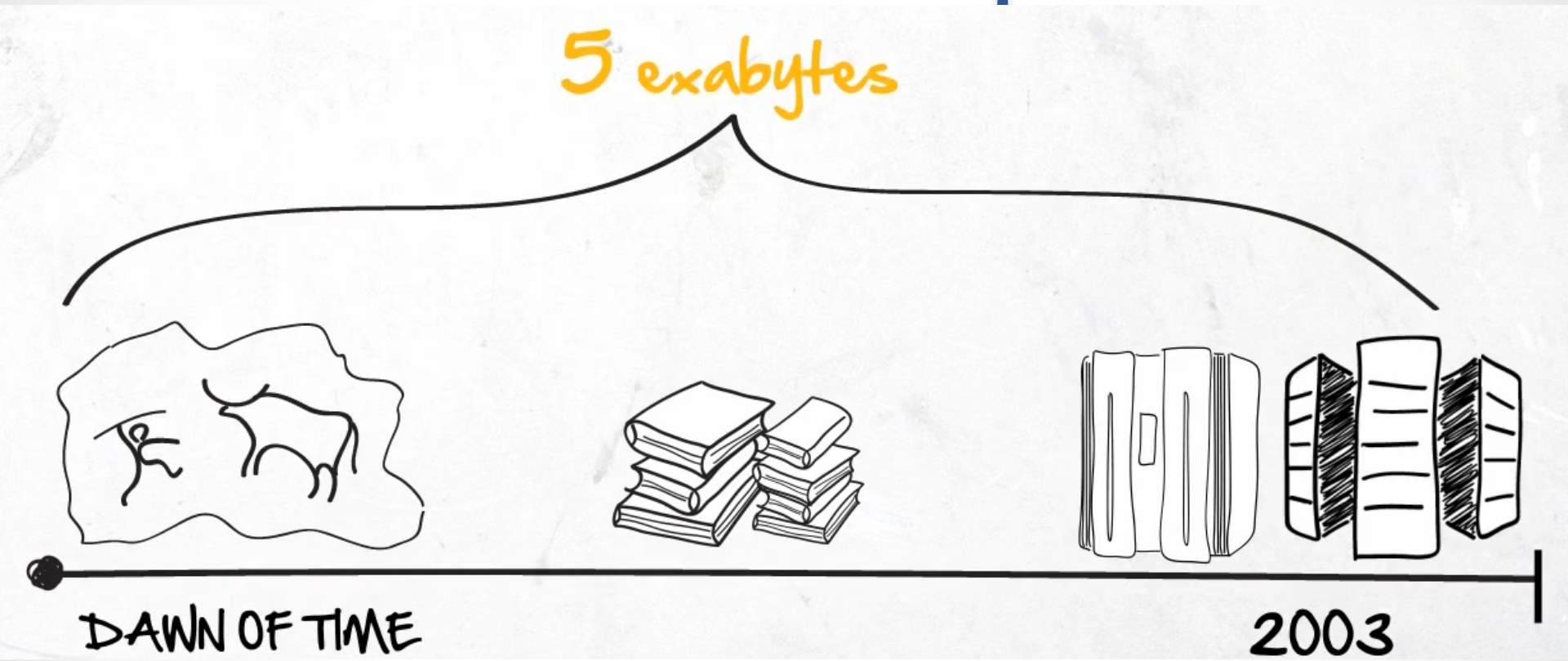
Economía basada en los datos

Los datos están en el centro de la sociedad y economía del conocimiento

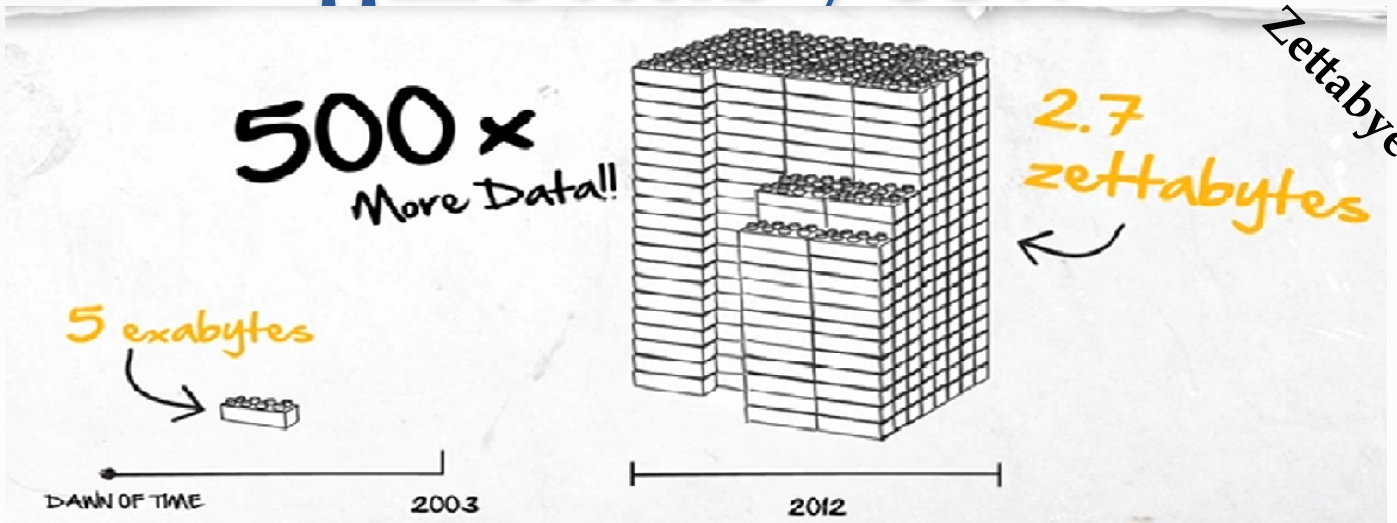
Towards a thriving data-driven economy
(Hacia una próspera economía basada en datos)
Communication SWD 214
European Comission, July 2014

https://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy

Crecimiento explosivo



¡¡Zettabyes!!

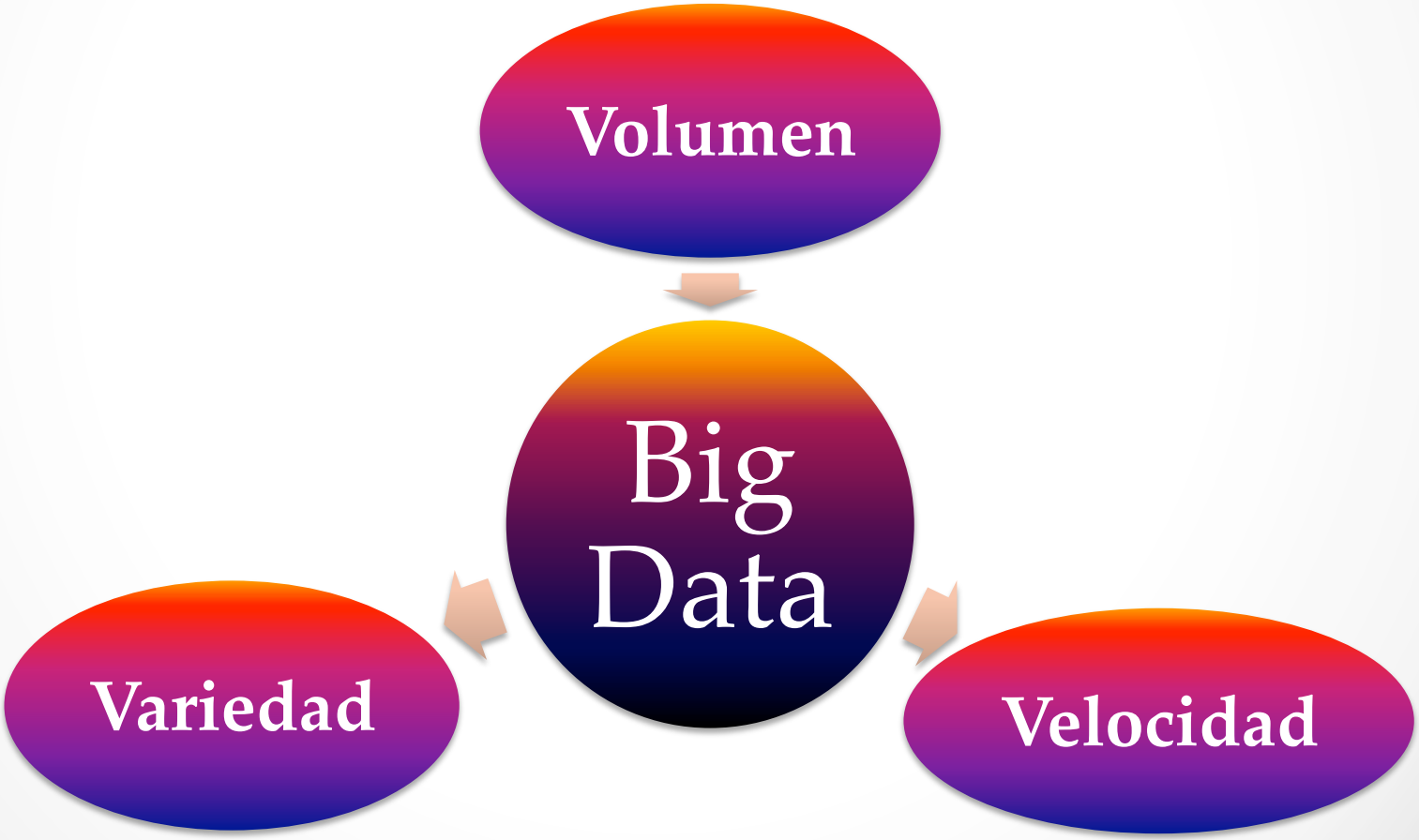


Zettabyes: 10^{21} bytes.

Unidades de información

Kilobyte = 10^3 = 1.000
Megabyte = 10^6 = 1.000.000
Gigabyte = 10^9 = 1.000.000.000
***Tera**byte = 10^{12} = 1.000.000.000.000*
***Peta**byte = 10^{15} = 1.000.000.000.000.000*
***Exa**byte = 10^{18} = 1.000.000.000.000.000.000*
***Zetta**byte = 10^{21}*
***Yotta**byte = 10^{24}*

Dimensiones V

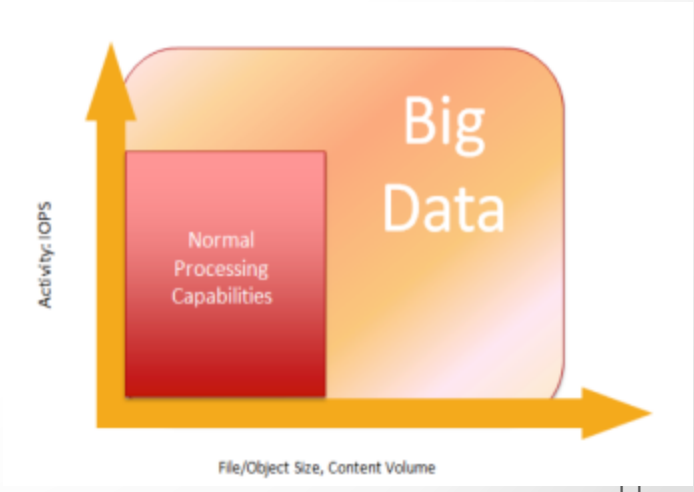


¿Qué es Big Data?

No hay una definición estándar

Big data es una colección de datos grande, complejos, **muy difícil de procesar a través de herramientas de gestión y procesamiento de datos tradicionales**

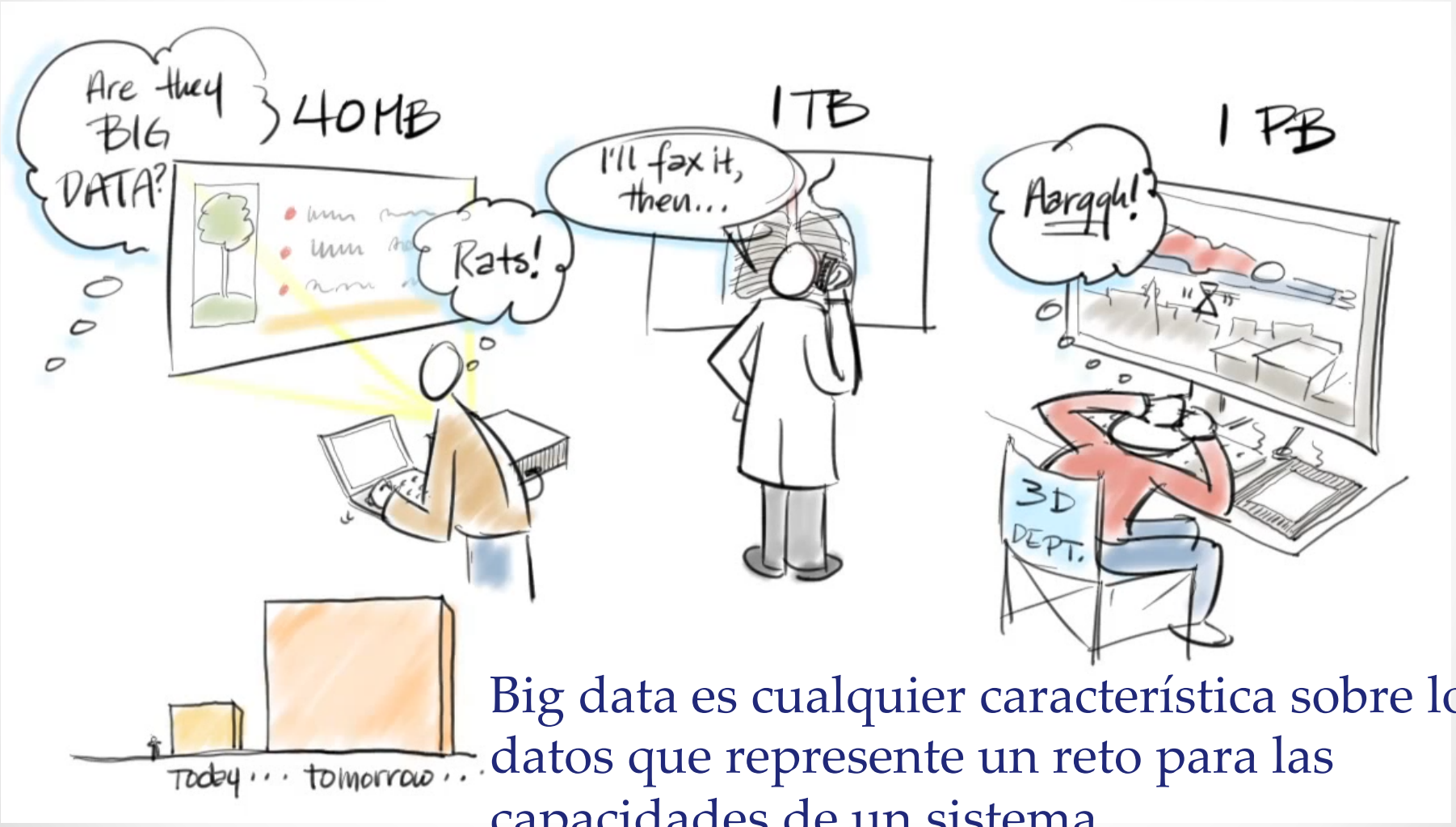
"*Big Data*" son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos ...



Algunas definiciones …

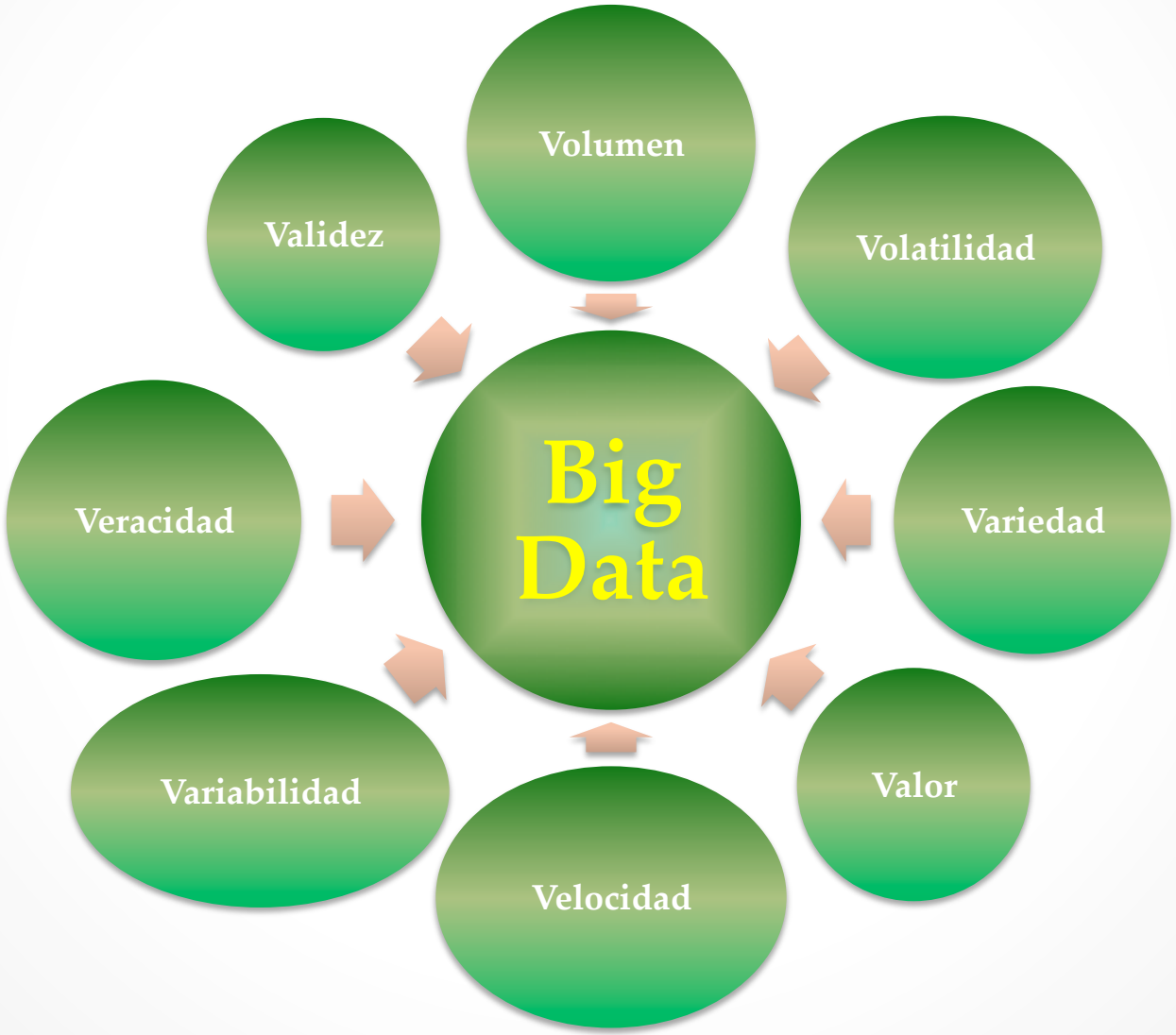
- "Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises"
- "Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapse time for its user population"
- "Disciplina que se ocupa de todas las actividades relacionadas con los sistemas que manejan **grandes** conjuntos de datos, principalmente, almacenamiento, búsqueda, compartición, análisis y visualización"

¿Qué es Big Data?



Big data es cualquier característica sobre los datos que represente un reto para las capacidades de un sistema.

Las 8 V's de Big Data



Predicción de propagación de la gripe

"Google puede predecir la propagación de la gripe (...) analizando lo que la gente busca en internet"

Más de 3.000M de búsquedas a diario

- **50 M de términos de búsqueda** más utilizados
- **Google comparó** esta lista con los datos de los CDC sobre propagación de gripe entre 2003 y 2008

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant.
Detecting influenza epidemics using search engine query data.
Nature 475 (2009) 1012-1014

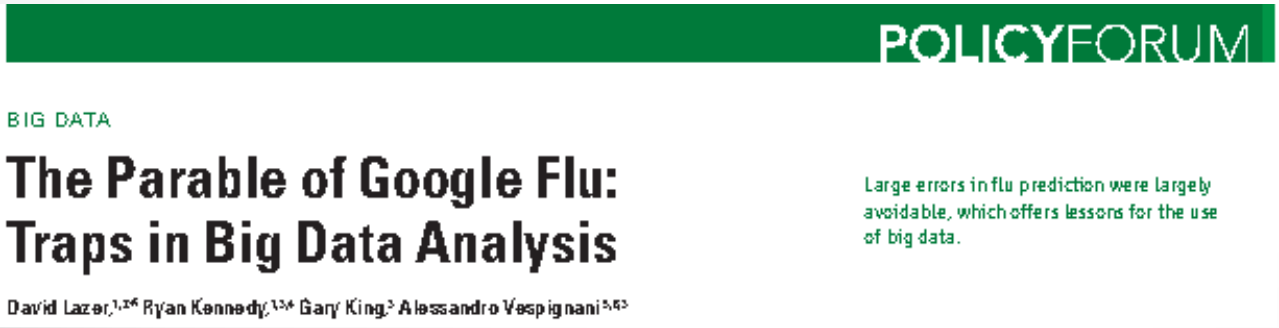
- Encontraron una **combinación de 45 términos de búsqueda** que al usarse con un modelo matemático presentaba una correlación fuerte entre su predicción y las cifras oficiales de la enfermedad

- Podían decir, como los CDC, **a dónde se había propagado la gripe pero casi en tiempo real**, no una o dos semanas después,

con un método basado en Big Data

- Se ha extendido a 29 países

- **En 2013 sobreestimó los niveles de gripe (x2 la estimación CDC)**
  - La sobreestimación puede deberse a la amplia cobertura mediática de la gripe que puede modificar comportamientos de búsqueda
  - Los modelos se van actualizando anualmente



POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3*] Gary King,[2] Alessandro Vespignani[4,5]

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

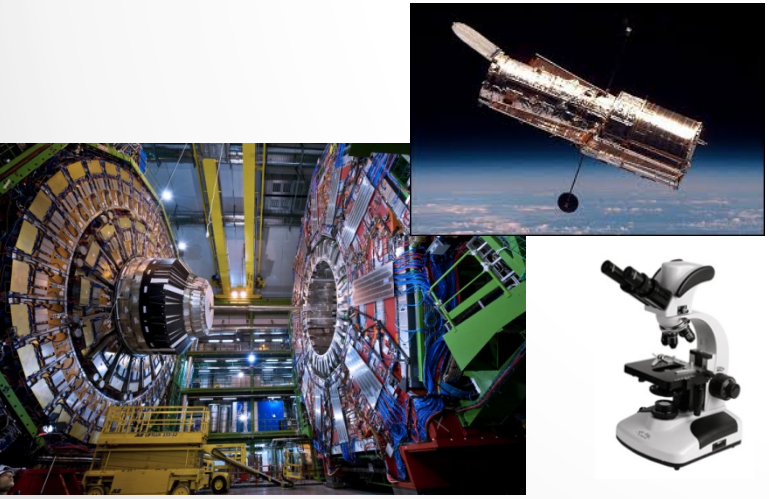
www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014

Generadores de datos masivos

Redes sociales y multimedia

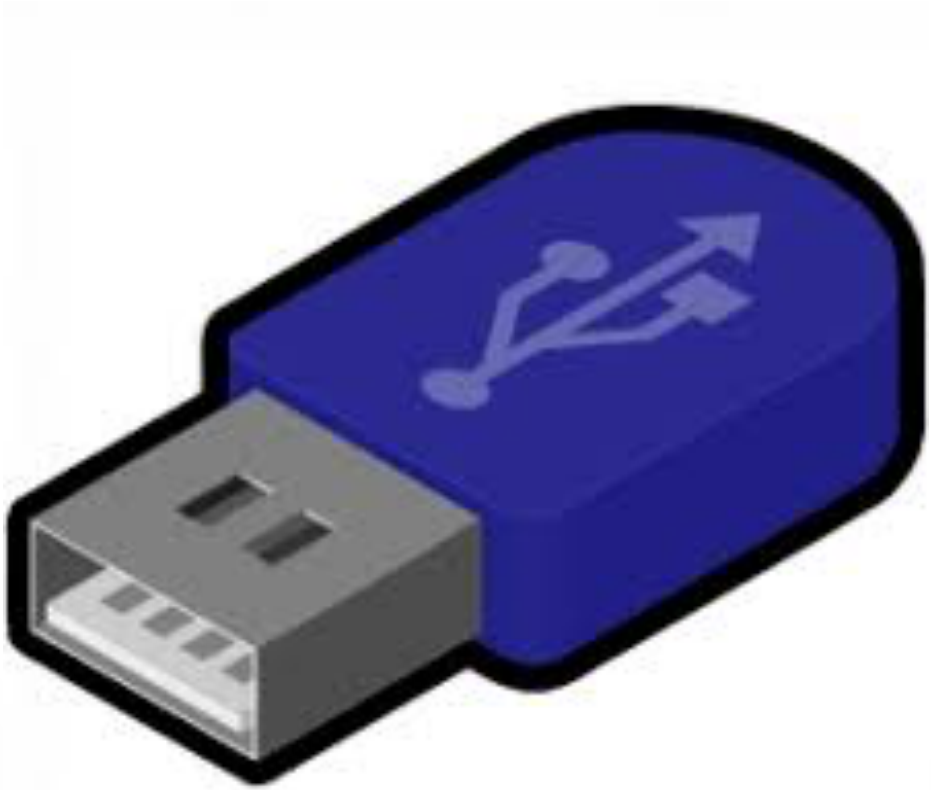
Transacciones

Dispositivos úbicuos de generación de contenido



Redes de sensores

¿Cómo de “Big”?



Ejemplo

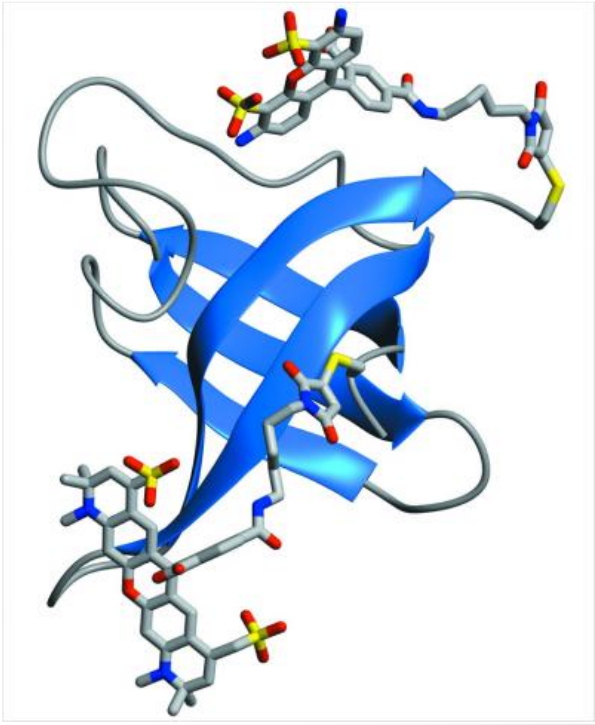
Evolutionary Computation for Big Data and Big Learning Workshop

Big Data Competition 2014: Self-deployment track

Objective: Contact map prediction

Details:

- 32 million instances
- 631 attributes (539 real & 92 nominal values)
- 2 classes
- 98% of negative examples
- About 56.7GB of disk space



Evaluation:

True positive rate · True negative rate

TPR · TNR

- http://cruncher.ncl.ac.uk/bdcomp/index.pl?action=data

Algoritmos en Big Data

Escalabilidad

Tecnología

BIG DATA

2001
3V's Gartner
Doug Laney

Google

MapReduce

Hadoop Ecosystem

2009-2013 Flink
TU Berlin
Flink Apache
(Dec. 2014) Volker Markl

2004
MapReduce
Google
Jeffrey Dean

Flink

Big Data

YAHOO!

hadoop

2008
Hadoop
Yahoo!
Doug Cutting

Spark
Lightning-Fast Cluster Computing

2010 Spark
U Berckeley
Apache Spark
Feb. 2014
Matei Zaharia

2010-2015:

Big Data Analytics: Mahout, MLLib, …

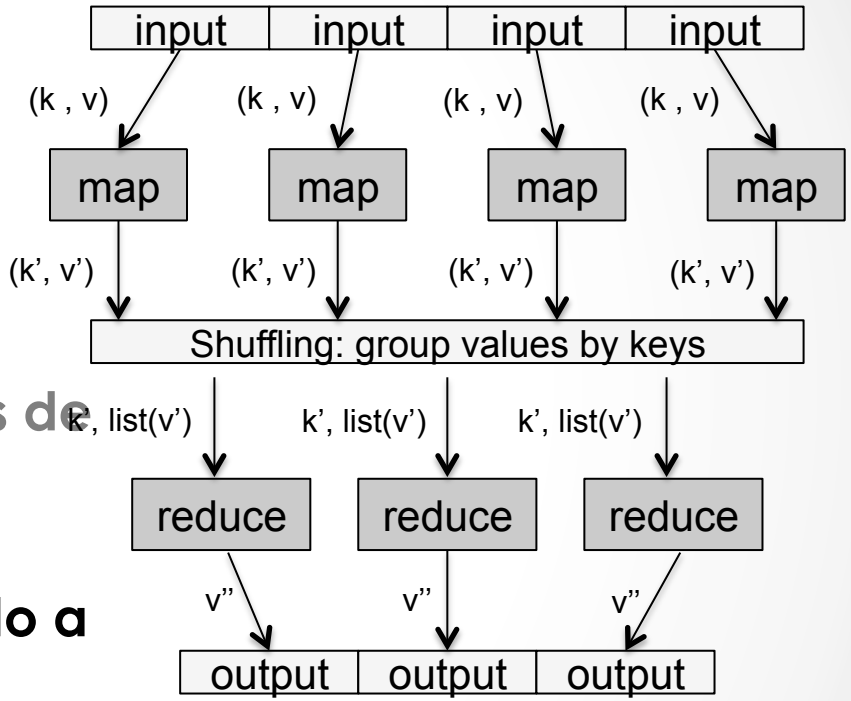
Hadoop Ecosystem

Aplicaciones

Nuevas Tecnologías

MapReduce

- MapReduce es el entorno más popular para Big Data
- Basado en la estructura {clave,valor}
- Dos operaciones:
  1. **Función Map: Procesa bloques de información**
  2. **Función Reduce: Fusiona los resultados previous de acuerdo a su clave.**
- + Una etapa intermedia de agrupamiento por clave **(Shuffling)**

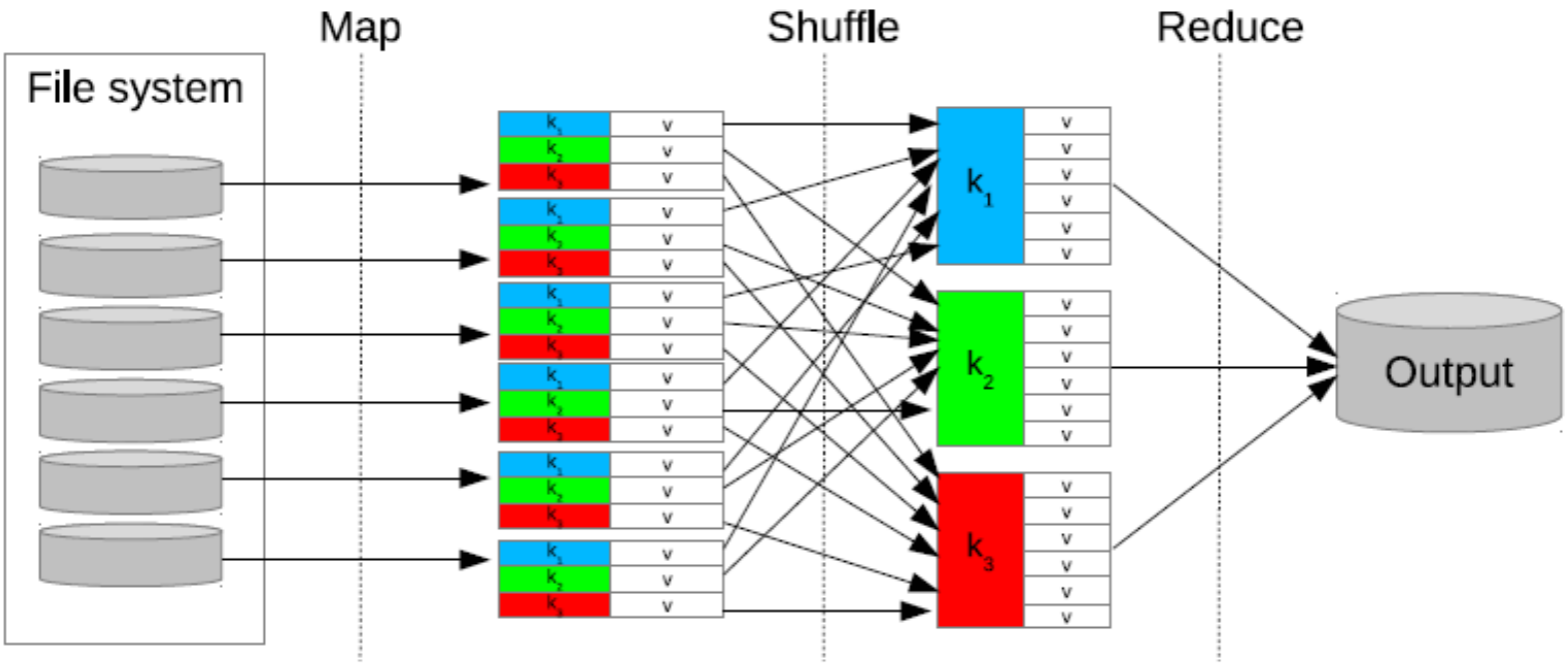


map (k, v) → list (k', v')
reduce (k', list(v')) → v''

J. Dean, S. Ghemawat, **MapReduce: Simplified data processing on large clusters**, Communications of the ACM 51 (1) (2008) 107-113.

Flujo de datos

Flujo de datos es transparente para el programador



Ficheros de entrada

Ficheros Intermedios

Ficheros salida

Características de MapReduce

- **Paralelización automática:**
  - Dependiendo del tamaño de ENTRADA DE DATOS➔ se crean múltiples tareas MAP
  - Dependiendo del número de intermedio <clave, valor> particiones ➔ se crean tareas REDUCE
- **Escalabilidad:**
  - Funciona sobre cualquier cluster de nodos/procesadores
  - Puede trabajar desde 2 a 10.000 máquinas
- **Transparencia programación**
  - Manejo de los fallos de la máquina
  - Gestión de comunicación entre máquinas

Resumiendo …

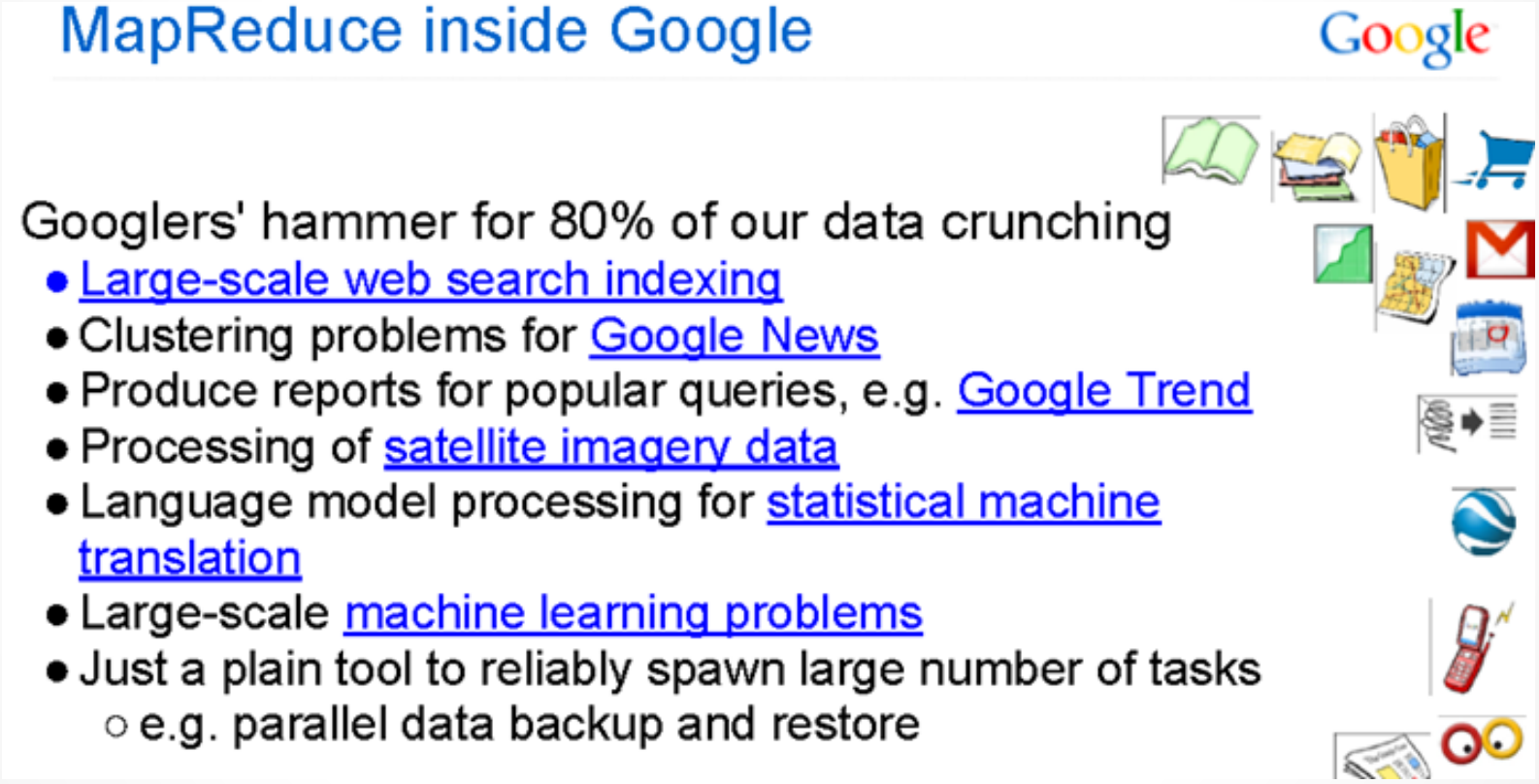
- **Ventaja frente a los modelos distribuidos clásicos:** El modelo de programación paralela de datos de MapReduce oculta la complejidad de la distribución y tolerancia a fallos

Programación distribuida para *dummies*

- **Claves de su filosofía**: Es
  - ***escalable:*** *se "olvidan" los problemas de hardware*
  - ***más barato:*** *se ahorran costes en hardware, programación y administración*

- **MapReduce no es adecuado para todos los problemas, pero cuando funciona, puede ahorrar mucho tiempo**

Google Applications: más de 10.000 herramientas



Limitaciones de MapReduce

"If all you have is a hammer, then everything looks like a nail."

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

- **Iterative Graph Algorithms: PageRank, ...**
- **Gradient Descent**
- **Expectation Maximization**

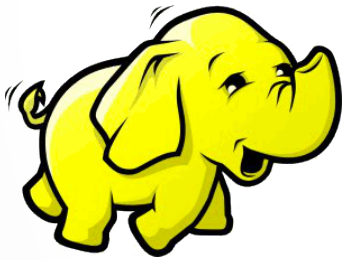
Hadoop

Hadoop es una implementación de código abierto del paradigma de programación MapReduce.

Hadoop

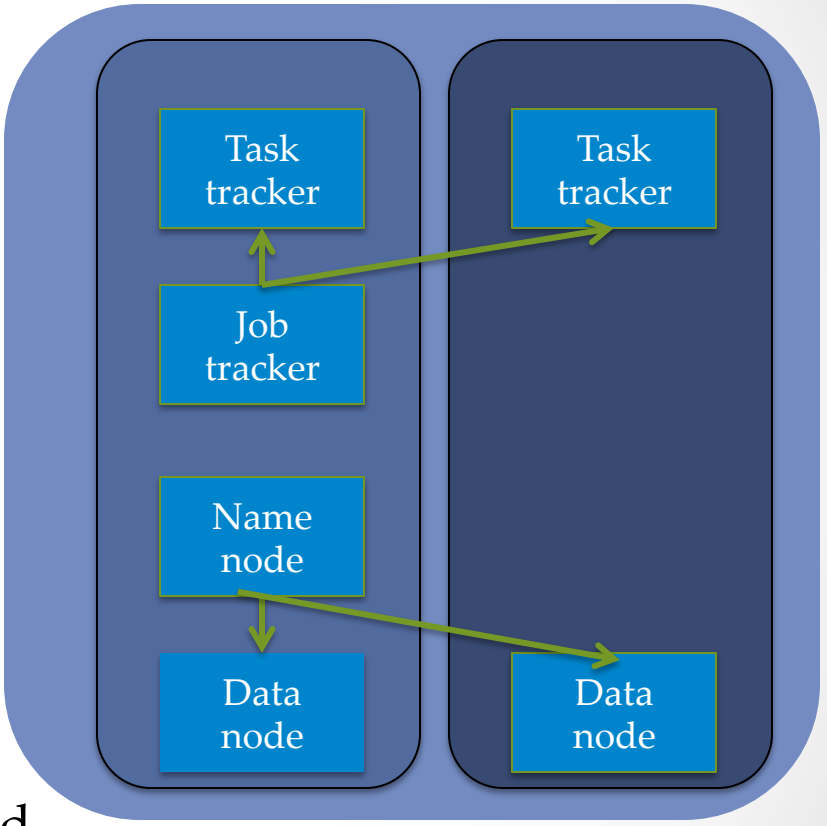
Escrito en java e inspirado en:

- MapReduce
- Google Distributed File System



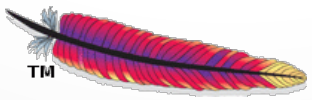
Map Reduce Layer

HDFS Layer



Creado por **Doug Cutting** (chairman of board of directors of the Apache Software Foundation, 2010)

http://hadoop.apache.org/



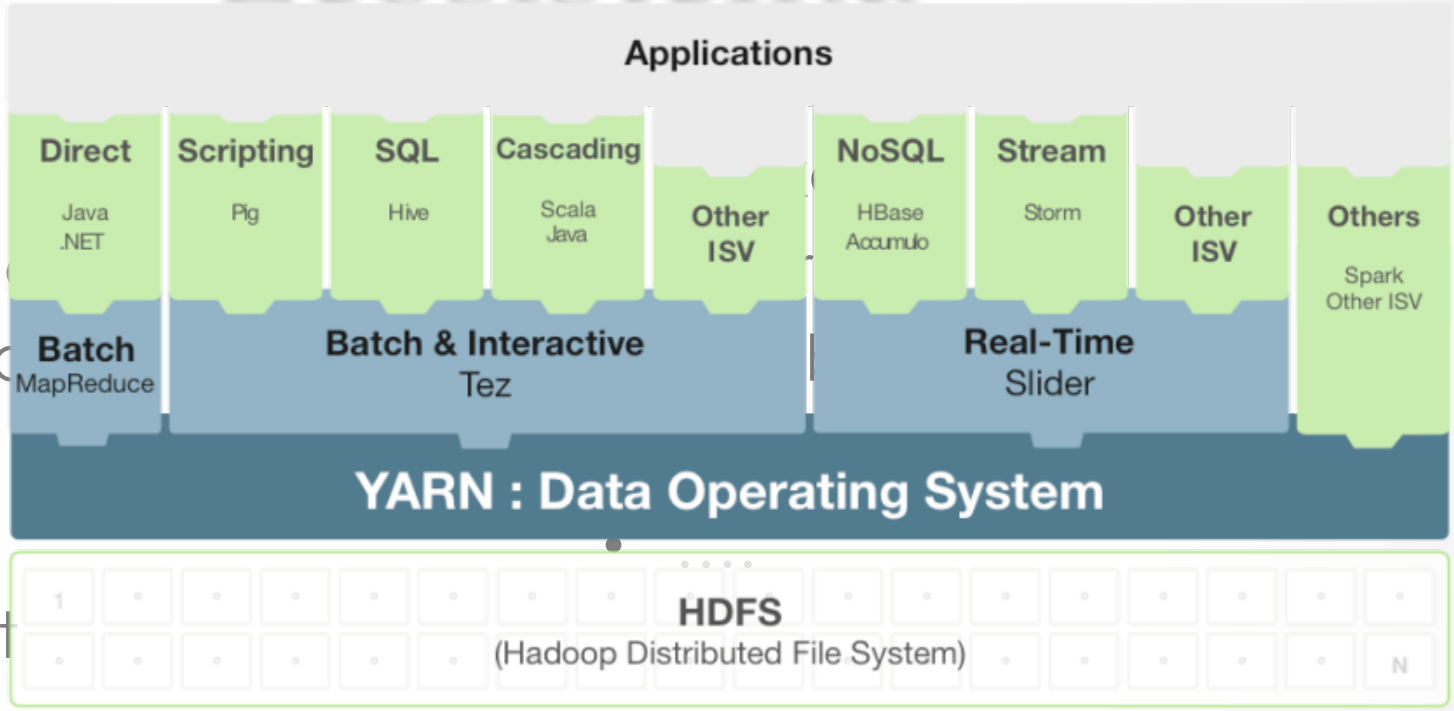
Almacenamiento y procesamiento

- Consta de dos servicios principales:
  - **Almacenamiento**: HDFS.
  - **Procesamiento**: MapReduce.

- Aporta una serie de ventajas:
  - **Bajo coste**: clústeres baratos / cloud.
  - **Facilidad de uso**.
  - **Tolerancia a fallos**.

Ecosistema

- Ambari
- Avro
- Cassan
- Chukwa
- Hbase
- Hive
- Mahout
- Pig
- Tez
- ZooKeeper

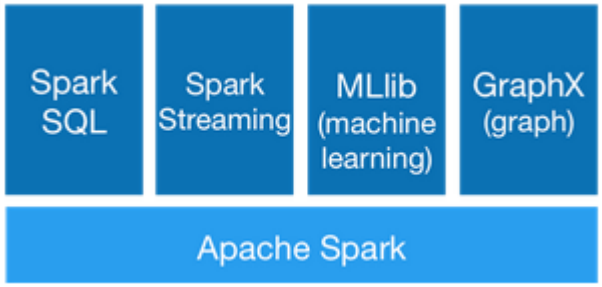


Spark

Entorno de trabajo (gestión de ejecución, API) genérico y rápido (hasta 100 veces más que Hadoop) para procesamiento de datos masivos
Centrado en una estructura de datos distribuida denominada “Resilient Distributed Dataset” (RDD)

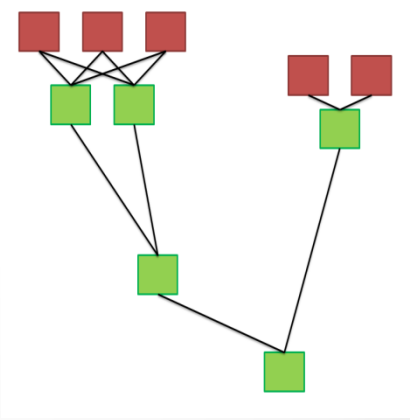
Desarrollado en Scala. Interfaces en Scala, Java, Python, …

Spark



- **Big Data "in-memory".** Spark permite realizar trabajos paralelizados totalmente en memoria:
  - Reducción de tiempos
  - Procesos iterativos

- **Esquema de computación más flexible que MapReduce.** Permite la flujos acíclicos de procesamiento de datos

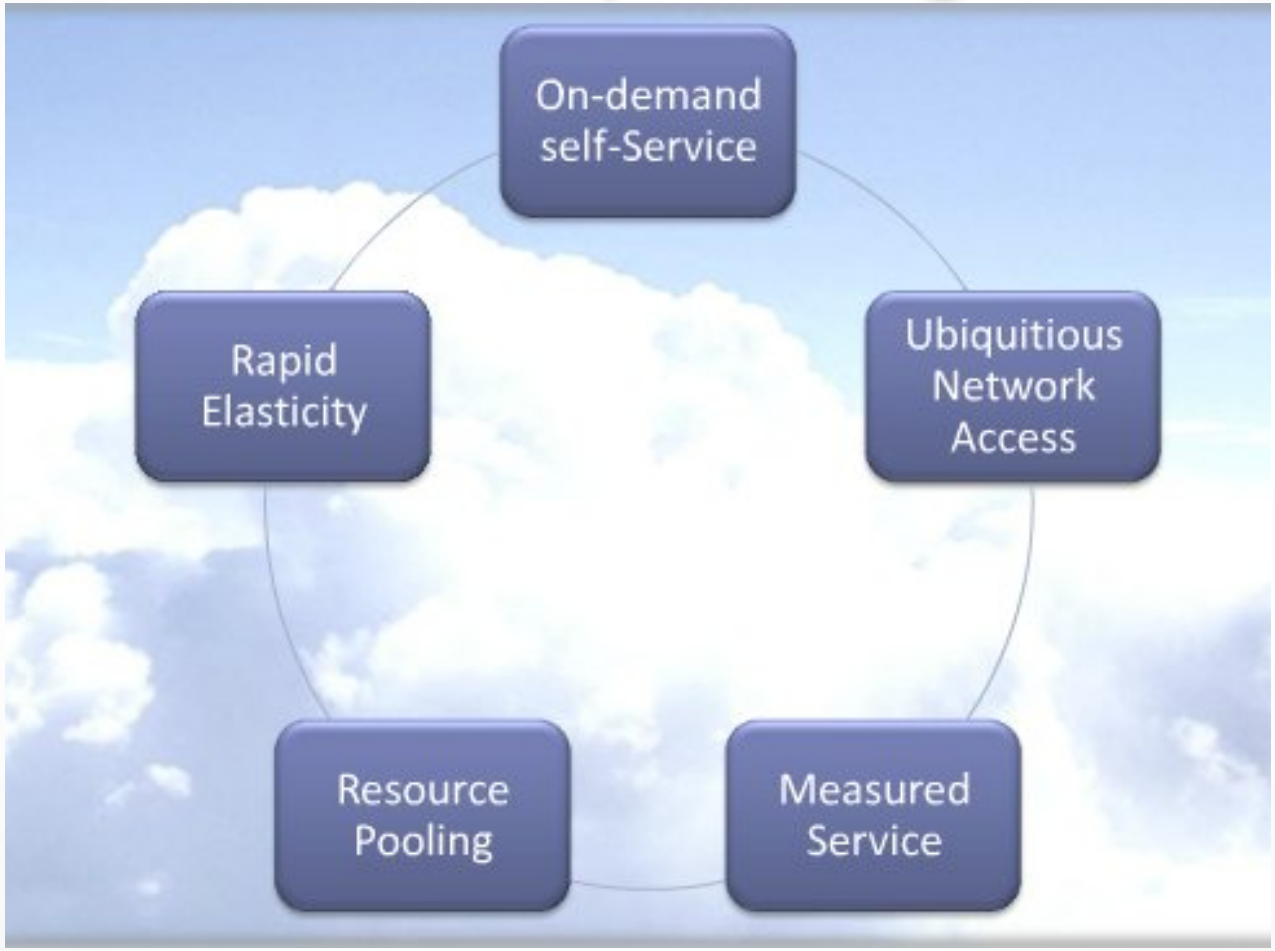


Cloud Computing

- Modelo de **prestación de servicios** de negocio y tecnología, que permite al usuario acceder a un catálogo de servicios estandarizado y responder a las necesidades del negocio, de forma **flexible** y **adaptativa**, […] **pagando** únicamente por **el consumo efectuado**.

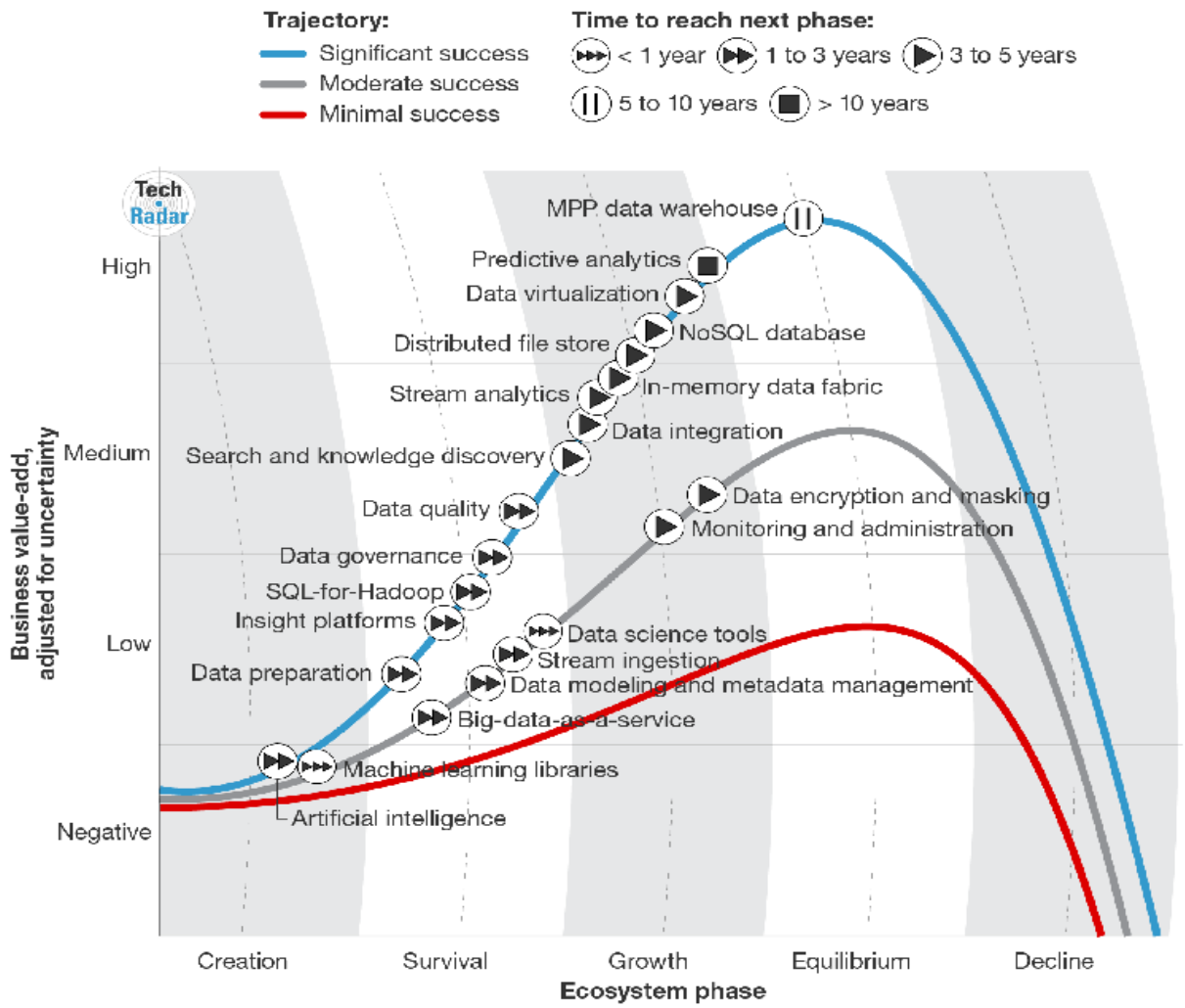
- El usuario tiene la ilusión de estar utilizando un **ordenador virtual con recursos ilimitados**

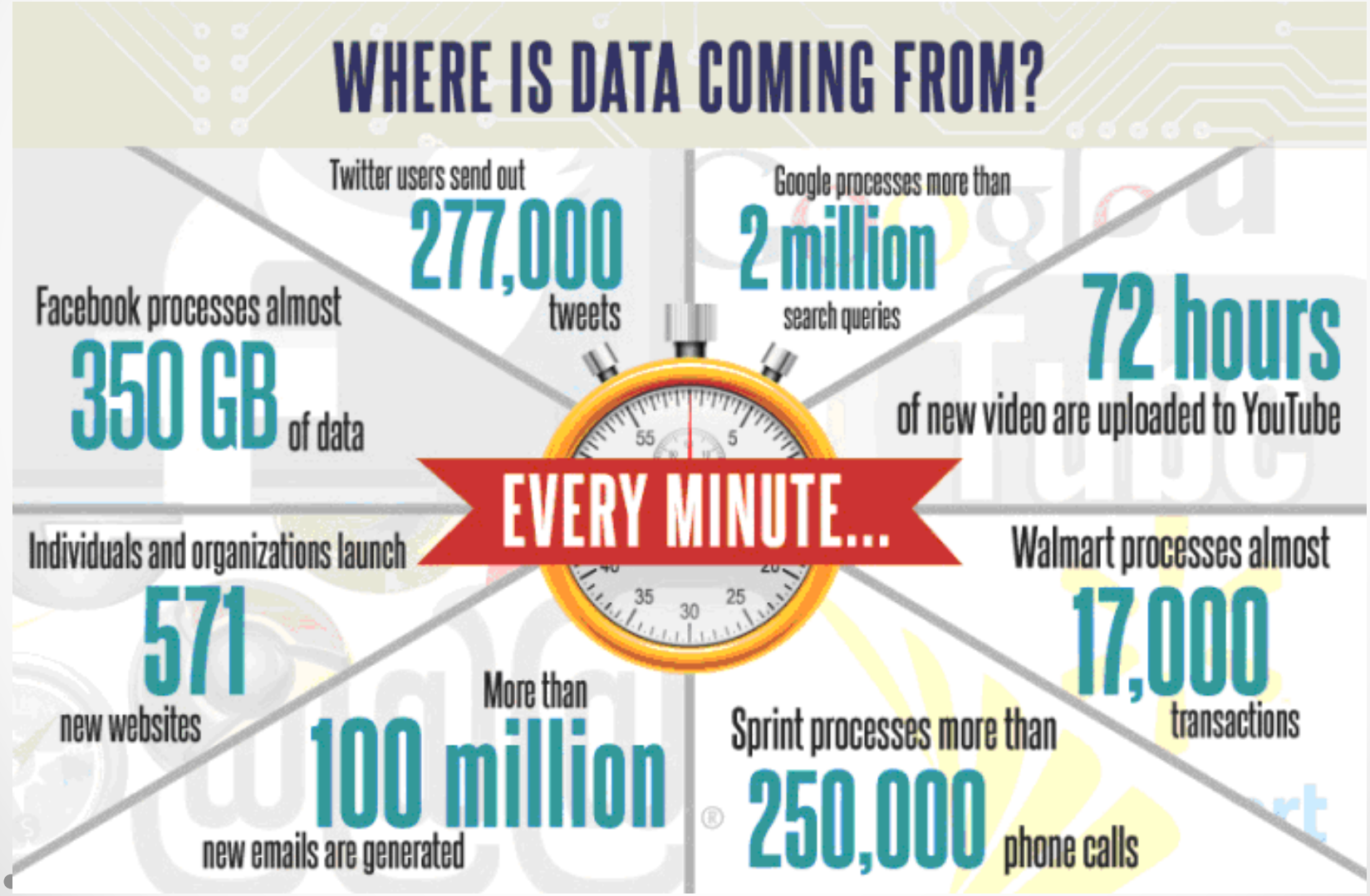
Aspectos claves de Cloud Computing



TechRadar™: Big Data, Q1 '16

TechRadar™: Big Data, Q1 2016




WHERE IS DATA COMING FROM?
Twitter users send out
277,000
tweets
Google processes more than
2 million
search queries
Facebook processes almost
350 GB
of data
72 hours
of new video are uploaded to YouTube
EVERY MINUTE...
Individuals and organizations launch
571
new websites
Walmart processes almost
17,000
transactions
More than
100 million
new emails are generated
Sprint processes more than
250,000
phone calls

¿Qué hacemos con estos datos?

- La tecnología hace salvables los obstáculos de recopilar y almacenar datos masivos, pero

Científico de datos

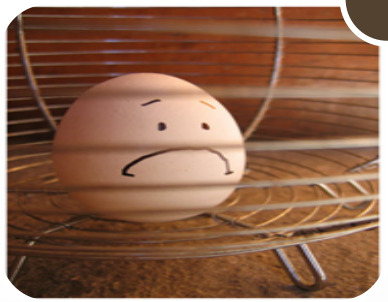
Asociaciones en transacciones de tarjetas de crédito

Asociaciones en datos de salud

Expedientes médicos varias años

Para reducir la tasa de reingresos

aflicción mental como "depresión" aumentaban los reingresos en 1 mes



Acción:

Intervención después del alta

Mejora salud mental y física

Sistemas de recomendación

Los datos incrementaron tremendamente 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. The power of habit, Charles Duhigg. Feb. 2012

Un chivo expiatorio

TARGET

Modelo de predicción de clientes embarazadas por medio de sus patrones de compra.

Descubrimiento: Cremas sin perfume al tercer mes. Dos docenas de productos

Predicción de fecha parto

Acción: Envío de cupones para cada fase del embarazo

Enfado de un padre: Su Hija recibe publicidad de productos para embarazadas

Disculpas Empresa. Disculpas del padre, confirmación del embarazo de la hija

Un chivo expiatorio …

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

Unos días después el director llamó al hombre para disculparse. Respuesta conciliadora: **"He estado hablando con mi hija –dijo el hombre – 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".**

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 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*.

España necesitará 60.000 profesionales de Big Data hasta 2015

22 octubre, 2013

Toledo.

EL PAÍS

PORTADA INTERNACIONAL POLI

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOLSA FINANZAS PERSONALES VIVIENDA TECNOLOGÍ

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 10

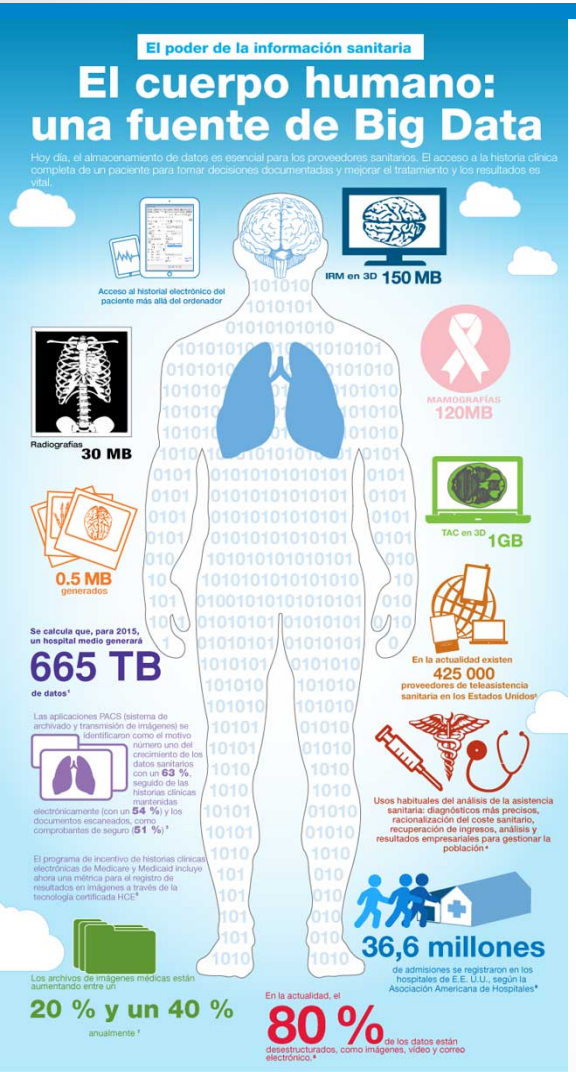
Archivado en: Citigroup Cap Gemini Sogeti SAP Oracle ING Bank BBVA Mapfre Bases datos IBM Telefónica Aplicaciones informáticas Tecnología Empresas Programas informáticos Economía

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#!

Cuidando la Salud

Las organizaciones de salud han coleccionado en los últimos años exabytes de información



Big data puede faciliar acciones para modificar los factores de riesgo que contribuyen a un porcentaje grande de las enfermedades crónicas, tales como actividad física, dieta, tabaco, exposición a la polución …

Móviles y Big data, una buena alianza: Sensores y aplicaciones en el móvil

Wearable blood pressure monitors send data to a smartphone app, then off to the doctor. (Photo by John Tlumacki/The Boston Globe via Getty Images)

http://www.pmfarma.es/noticias/16337-el-cuerpo-humano-una-fuente-de-big-data.html

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, 2014

Banca: Análisis de datos

Expansión.com
Domingo, 08.12.13. Actualizado a las 11:02

BBVA da un paso al frente en 'big data'

El banco ha actualizado todo su sistema de almacenamiento de datos corporativo para tener acceso desde un único punto a todas las 'islas de información' de la entidad en el mundo y consolidarse como un referente mundial en innovación tecnológica.

BBVA

Innova Challenge API

Available statistics services

Tendencias de negocio: Análisis de compras con tarjetas de crédito …

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

http://elpais.com/elpais/2015/01/29/ciencia/1422520042_066660.html

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

Science

AAAS

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

Yves-Alexandre de Montjoye,[1]* Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex "Sandy" Pentland[1]

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.

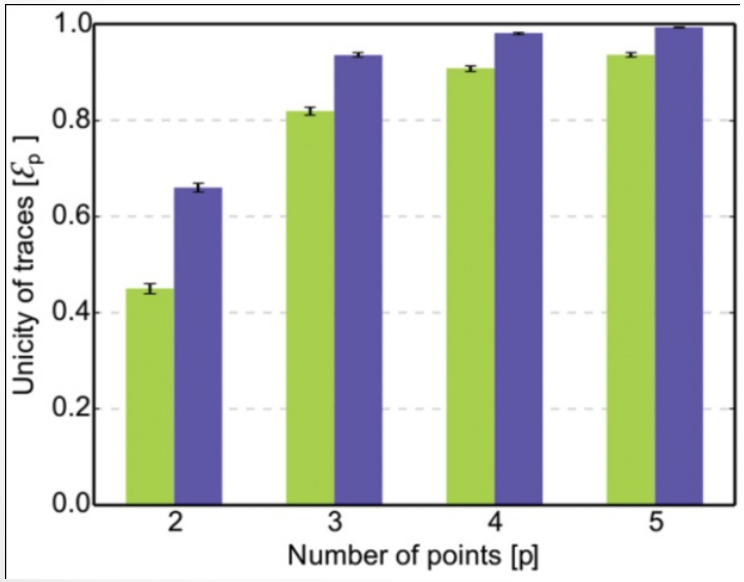
http://www.sciencemag.org/content/347/6221/536

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

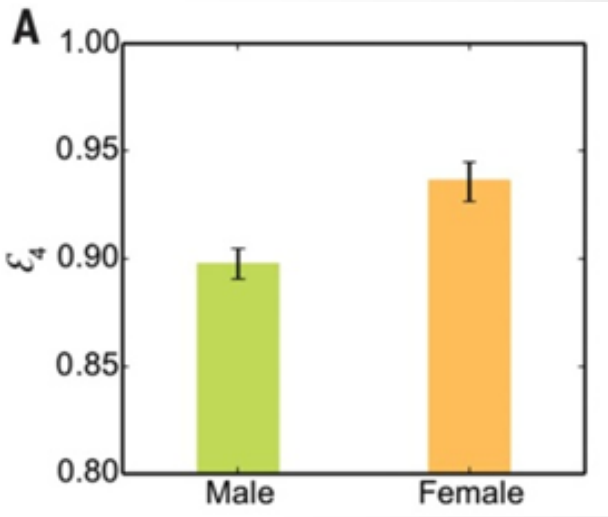
PRIVACIDAD EN INTERNET »

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

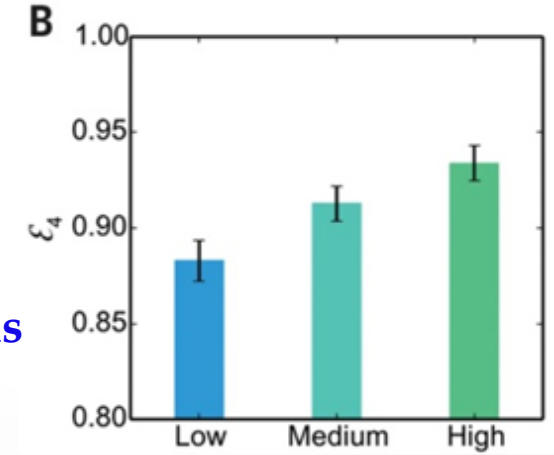
- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT



Identificación por el número de compras



Identificación por el género



Identificación por el poder adquisitivo

Industria. Automoción

EL PAÍS

COCHES SIN CONDUCTOR

Un coche autónomo recorre por primera vez las carreteras de España

- Un prototipo cubre la distancia entre Vigo y Madrid, apenas una semana desp DGT cambiara la normativa para permitir estos ensayos
- ¿Llega la hora del vehículo conectado?

J. J. GÁLVEZ | Madrid | 23 NOV 2015 - 18:49 CET

1.088 | 54 | 19 | 16

Archivado en: DGT Citroën Coches sin conductor Direcciones Generales Gru Nuevas tecnologías Coches Ministerios Vehículos Fabricantes automóviles

Interior del vehículo autónomo que ha salido este lunes a la carretera. / EFE (PROMOCIONAL)

PILOTO AUTOMÁTICO »

Conduce si te apetece

- El piloto automático está cada vez más cerca, aunque la tecnología llegará por fases
- e, el de ciudad
- nomo de Nissan

Nuevas tecnologías

gía Transporte

GOOGLE CAR »

Con flores y sin volante

- La artista española Inmaculada del Castillo gana el concurso para decorar el coche sin conductor de Google
- El accidentado estreno del coche sin conductor de Google

ROSA JIMÉNEZ CANO | San Francisco | 20 OCT 2015 - 11:52 CEST

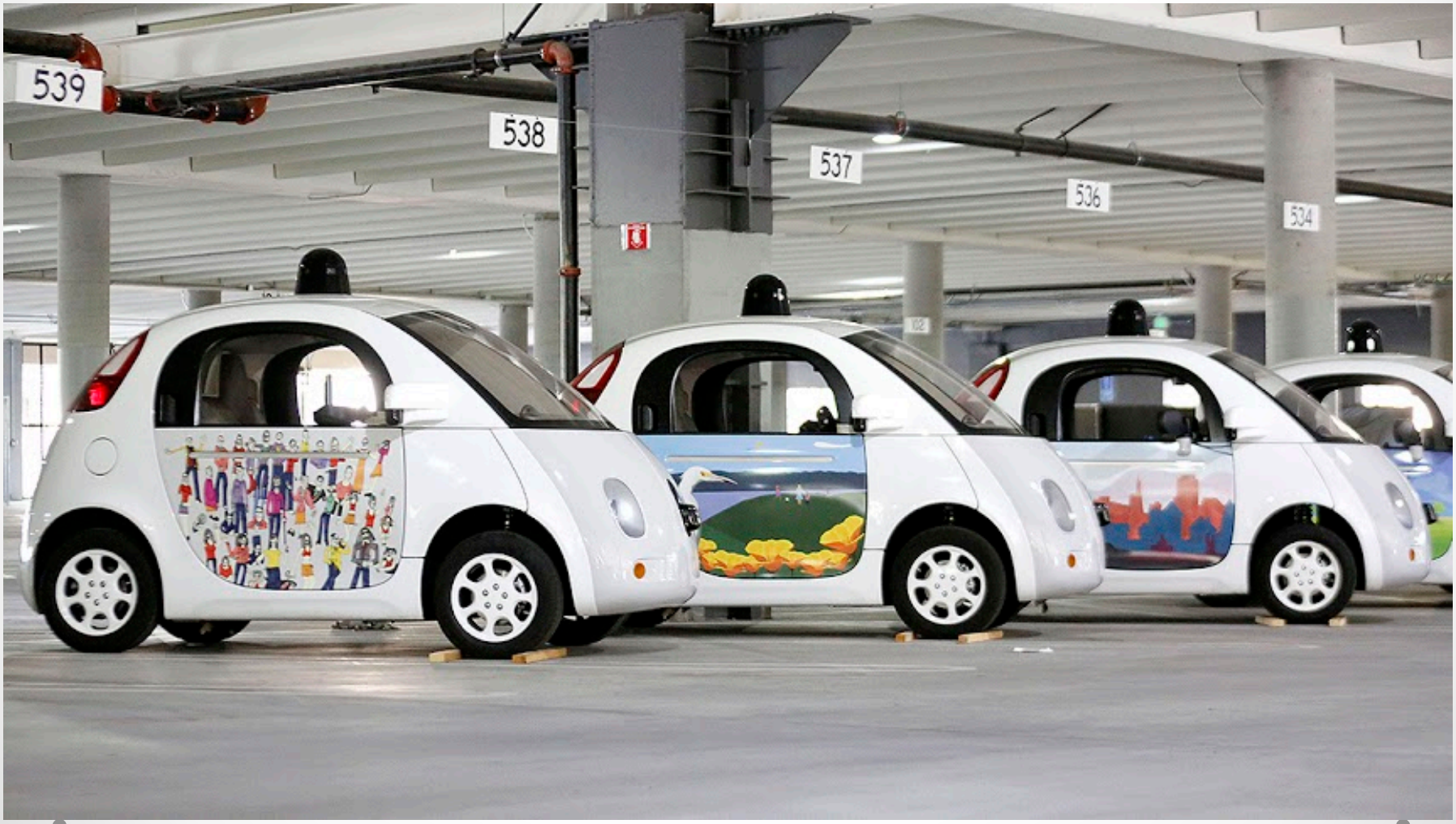
289 | 47 | 6 | 7

Archivado en: Silicon Valley Tesla Motors Google car San Francisco Coches sin conductor Google California Nuevas tecnologías Buscadores Alphabet Estados Unidos Coches

La artista Inmaculada del Castillo posa junto al coche sin condutor de Google y su diseño. / DAVID DITZEL (EL PAÍS)

http://elpais.com/tag/coches_sin_conductor/

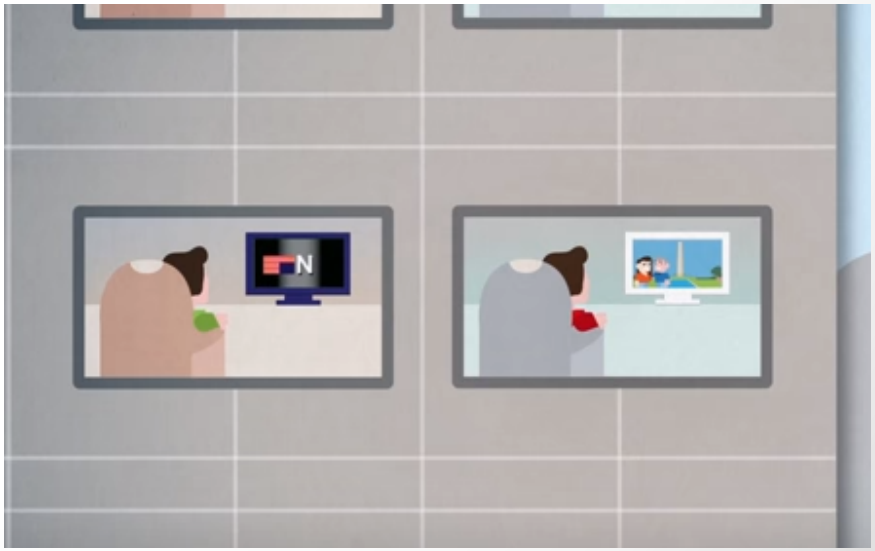
Industria. Automoción



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



Big Data para el Bien Social

UNITED NATIONS GLOBAL PULSE

Harnessing big data for development and humanitarian action

GLOBAL PULSE PROJECT SERIES

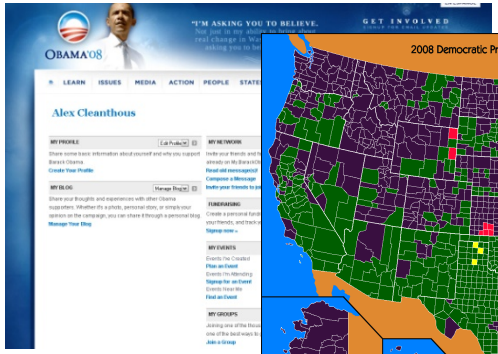
The Global Pulse Project Series is a collection of case studies of 20 data innovation projects covering global issues ranging from public health to climate change, food security to employment.

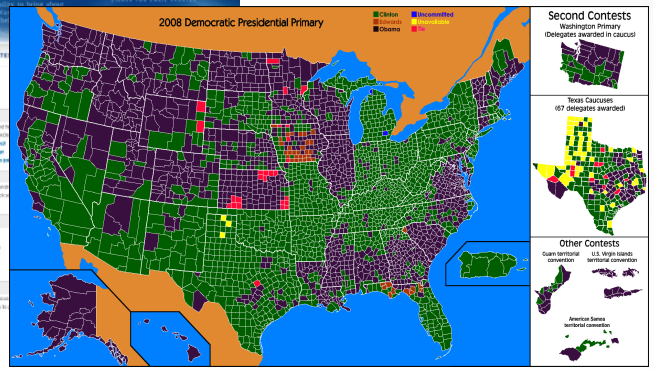
Read More /

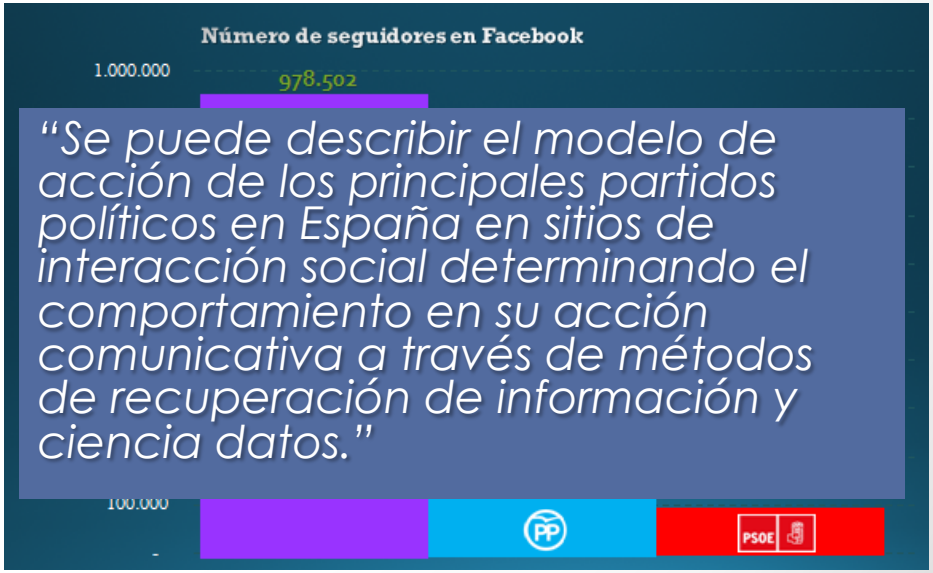
El análisis de datos también debe servir para mejoras sociales y avances en la sociedad

Big Data en la Política

26 Septiembre, 1960



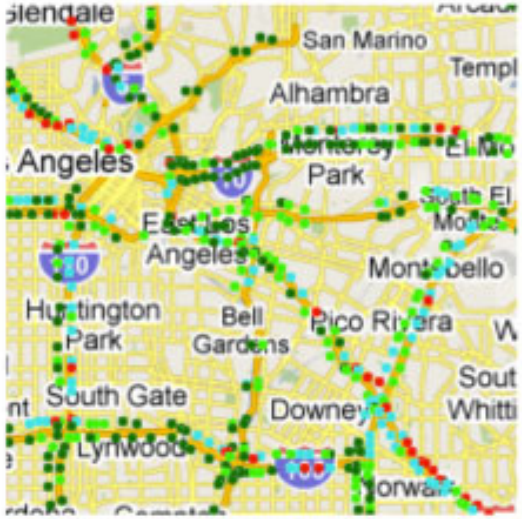




"Se puede describir el modelo de acción de los principales partidos políticos en España en sitios de interacción social determinando el comportamiento en su acción comunicativa a través de métodos de recuperación de información y ciencia datos."

Estrategias de Comunicación de los Partidos Políticos en Sitios Web de Interacción Social, Tesis Doctoral, Agustín Zubillaga, Universidad de Deusto, Noviembre, 2015. (Dirs. Pablo García e Iker Pastor)

1. Moving objects



2. Traffic sensor data

3. Road network data

4. Points-of-interest data

Conclusiones

- Existe el Big Data
- Resuelve problemas **reales**
- No todo es Big Data
- La tecnología está en desarrollo

Las grandes colecciones de datos nos conducen continuamente a innovaciones (en la industria, ciencia, sociedad)