Big Data Week is one of the most unique global platforms of interconnected community events focusing on the social, political, technological and commercial impacts of Big Data

Follow all the events at

bigdataweek.com

Official Event Hashtag #bdw13

BDW13: Entender el Big Data

BIG_DATA_WEEK_2013

Óscar Marín Miró
@oscarmarinmiro
oscar@outliers.es

Outliers
Because differences matter.

CONTENIDOS

¿QUÉ ES BIG DATA?

RECORRIDO HISTÓRICO

ECOSISTEMA

EL VALOR DE BIG DATA

ESCENARIOS PRÁCTICOS

EL FUTURO DE BIG DATA

BIG DATA EN NUESTRAS VIDAS

Material del curso en http://assets.outliers.es/bdw13/bigdata

¿QUÉ ES BIG DATA?

DEFINICIONES

"Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"

http://en.wikipedia.org/wiki/Big_data

DEFINICIONES

- Big Data y las 3 ‘V’
  - Velocidad
  - Volumen
  - Variedad

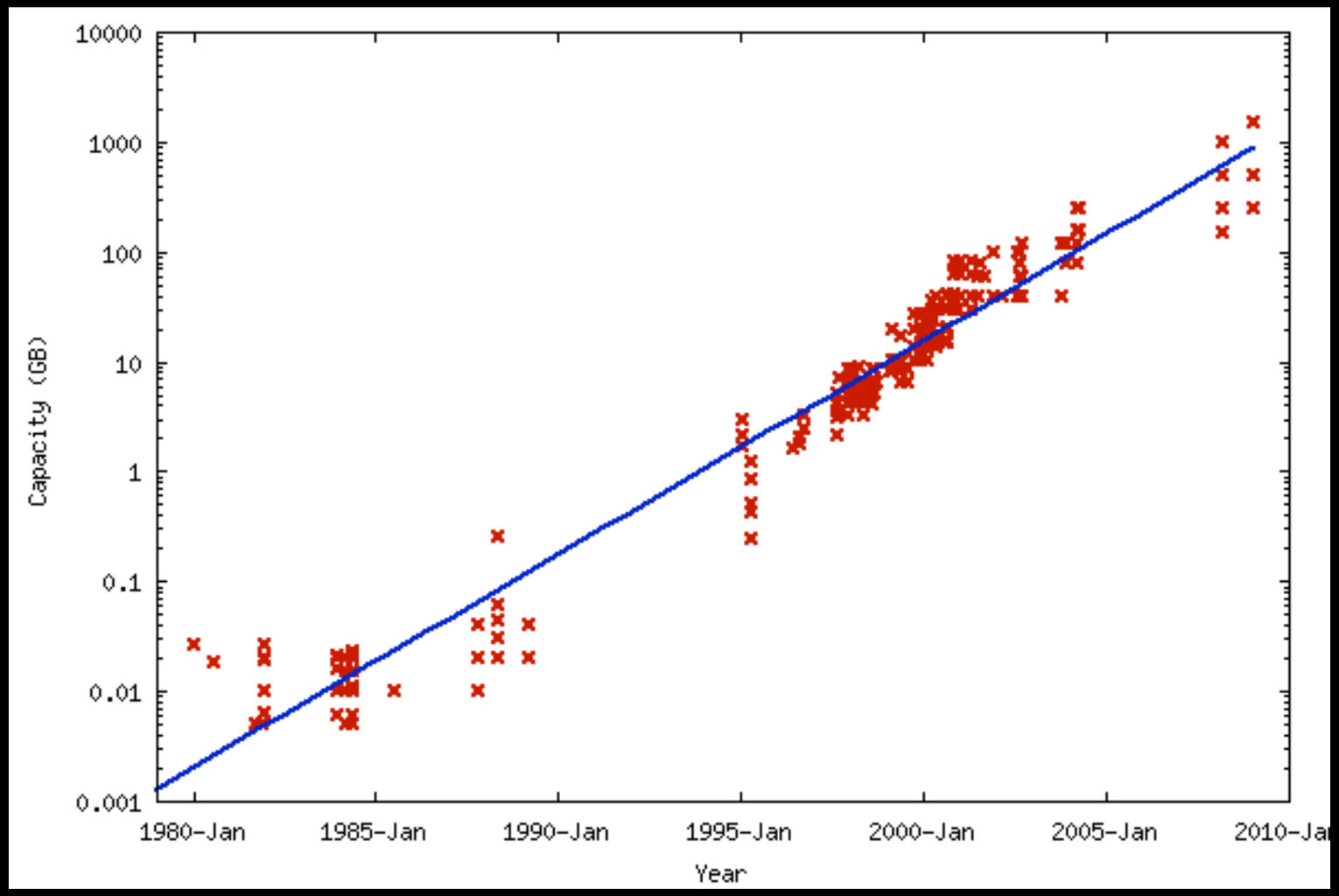
http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

LOS PROBLEMAS: EL VOLUMEN

http://royal.pingdom.com/2010/02/18/amazing-facts-and-figures-about-the-evolution-of-hard-disk-drives/

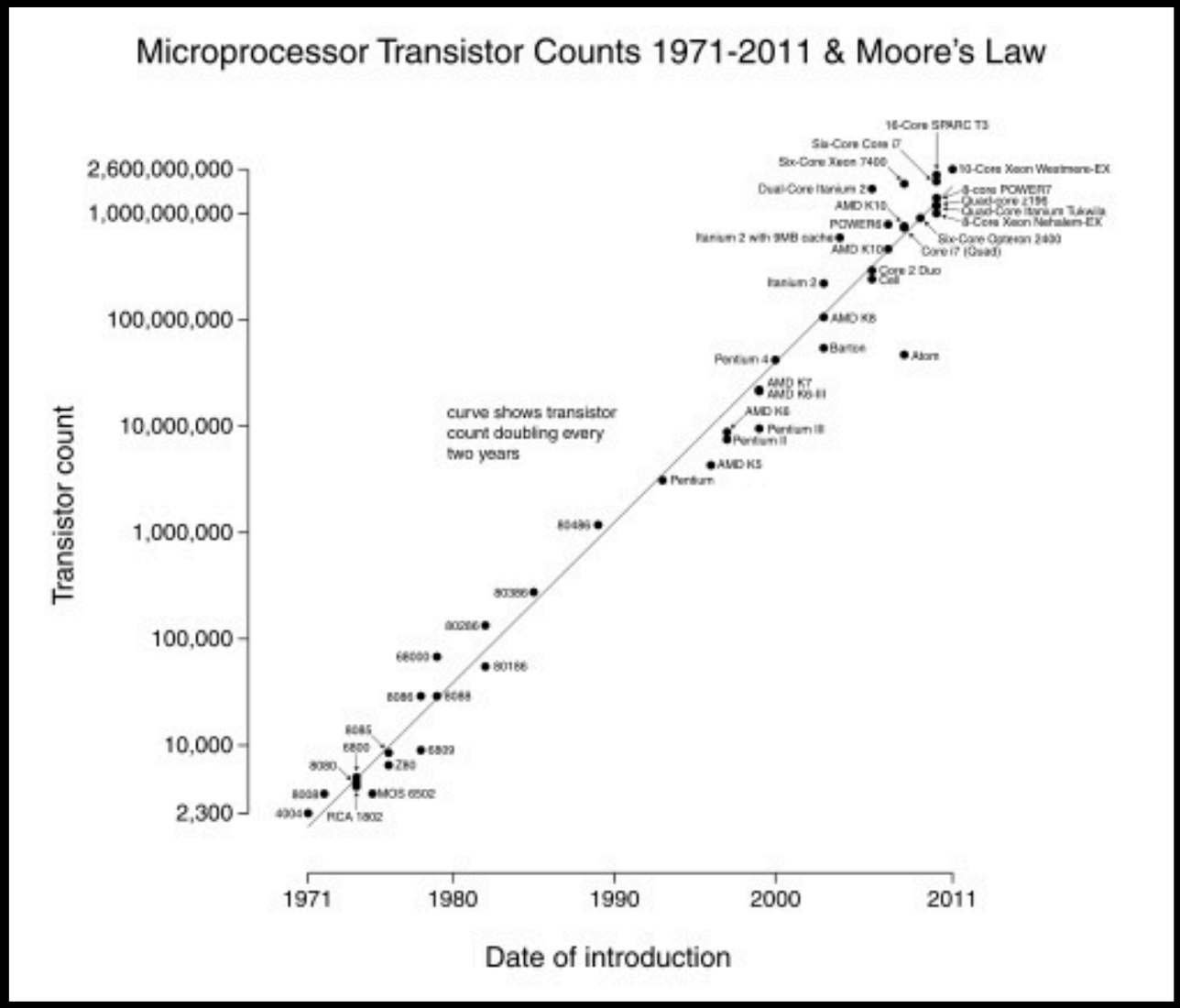
LOS PROBLEMAS: EL VOLUMEN

¿Qué hacemos cuando los datos superan con creces el tamaño de un disco duro?



http://en.wikipedia.org/wiki/File:Hard_drive_capacity_over_time.png

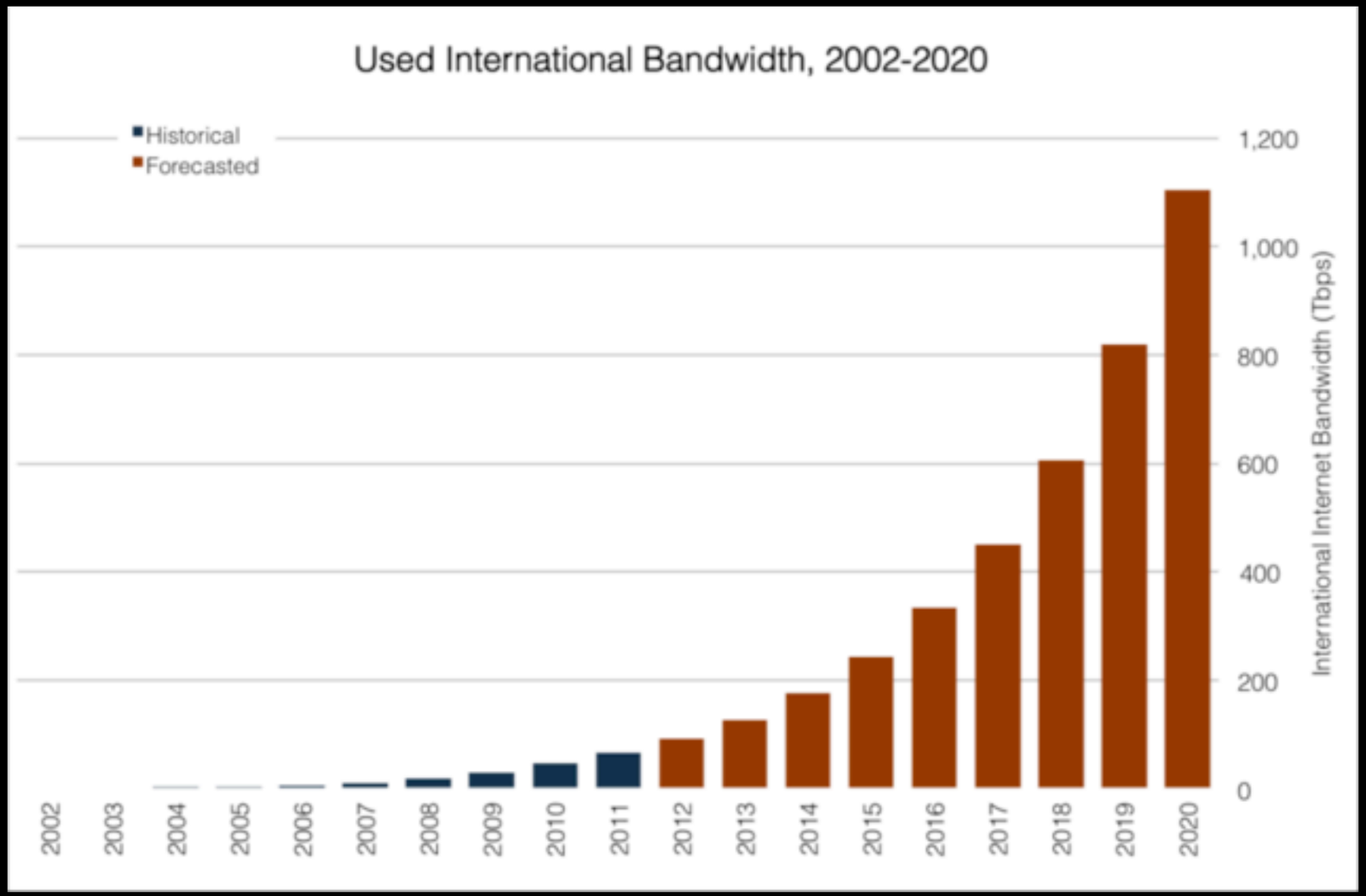
LOS PROBLEMAS: LA VELOCIDAD



¿Qué hacemos cuando los datos llegan a un ritmo superior al que pueden ser analizados?

http://en.wikipedia.org/wiki/Moore's_law

LOS PROBLEMAS: LA VELOCIDAD

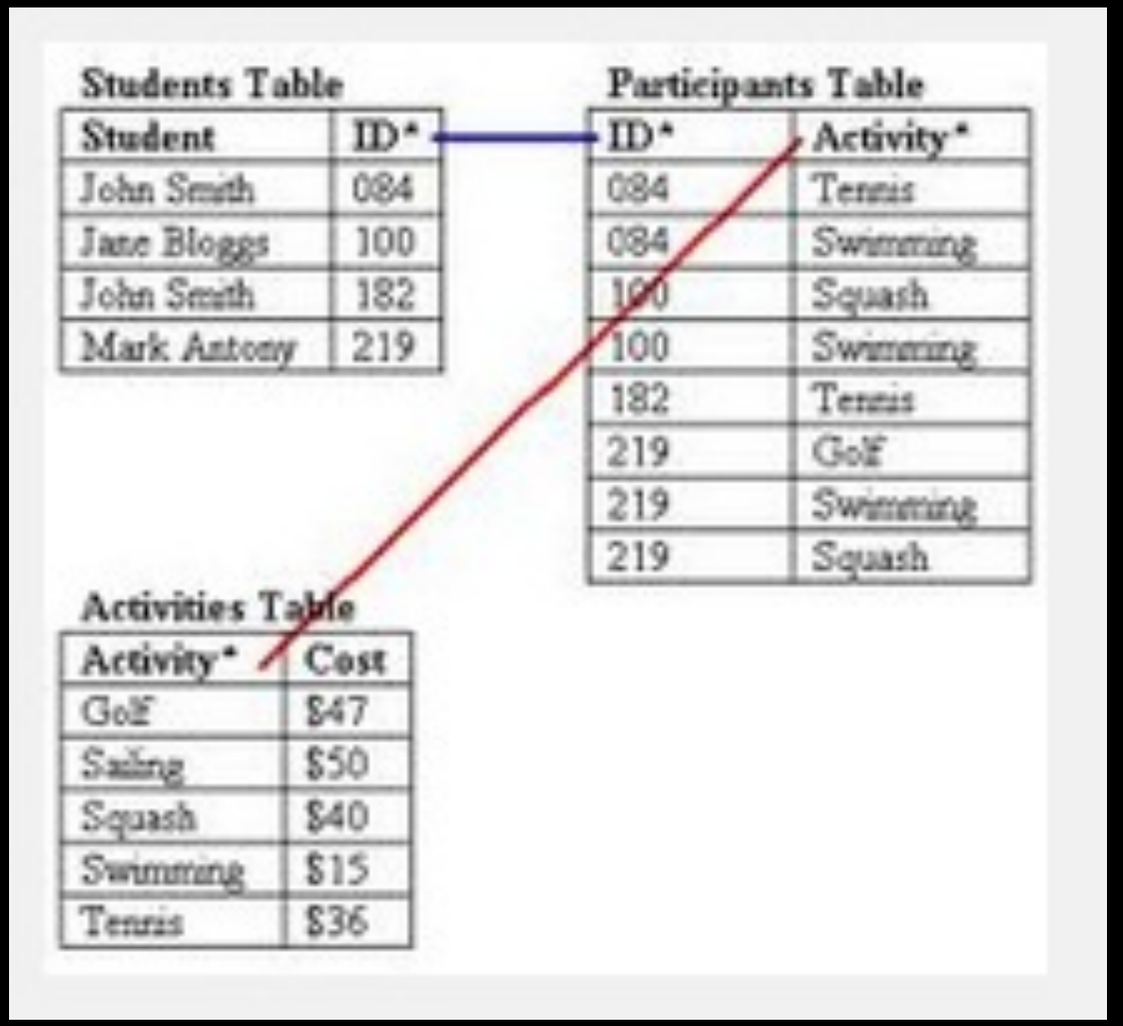


¿Y si no nos llega el ancho de banda?

http://arstechnica.com/business/2012/05/bandwidth-explosion-as-internet-use-soars-can-bottlenecks-be-averted/

LOS PROBLEMAS: LA VARIEDAD

El problema del join en RDMS



Students Table

Student	ID*
John Smith	084
Jane Bloggs	100
John Smith	182
Mark Antony	219

Participants Table

ID*	Activity*
084	Tennis
084	Swimming
100	Squash
100	Swimming
182	Tennis
219	Golf
219	Swimming
219	Squash

Activities Table

Activity*	Cost
Golf	$47
Sailing	$50
Squash	$40
Swimming	$15
Tennis	$36

LOS PROBLEMAS: LA VARIEDAD

- El problema del ‘join’ en Big Data
- ‘Wide Data’: Datos con esquema variable o sin esquema

BIG DATA: SOLUCIONES

- Volumen: Sistemas de Ficheros Distribuidos (HDFS)
- Velocidad/Caudal: Sistemas de Distribución de Procesos (MapReduce)
- Variedad: BBDD No relacionales (NoSQL)

BIG DATA: EJEMPLOS

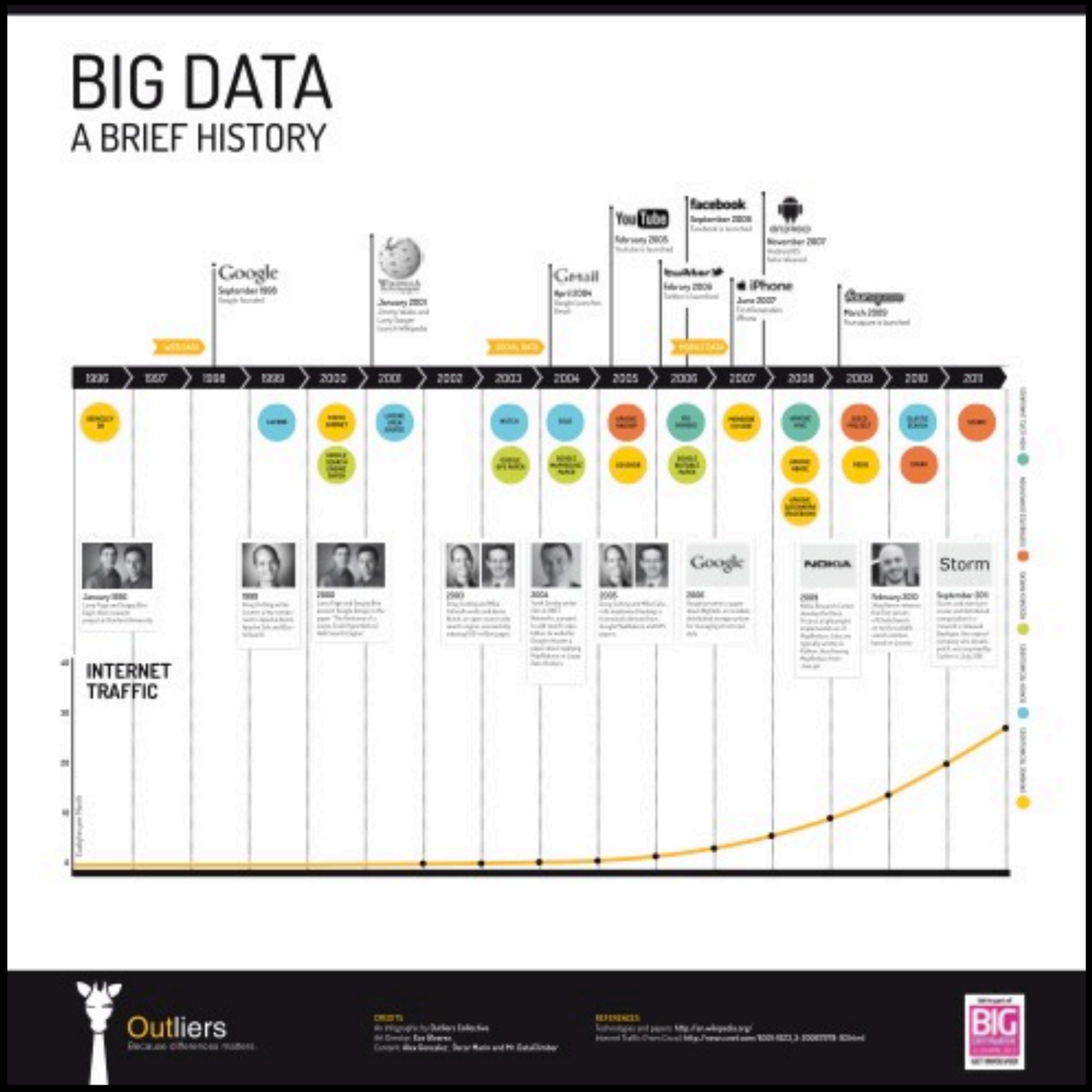
- Twitter: 340 millones de tweets diarios (~= 1TB/día)
- Facebook: 800 millones de status diarios
- Google: 1000 millones de consultas diarias

http://www.slideshare.net/gigaom/the-3vs-of-big-data-variety-velocity-and-volume-from-structuredata-2012

‘DATA NEVER SLEEPS’

http://7.mshcdn.com/wp-content/uploads/2012/06/DataNeverSleeps.jpg

BIG DATA: RECORRIDO HISTÓRICO



http://assets.outliers.es/infographics/BigData_A_Brief_History.pdf

1996

- Nace BerkeleyDB: BBDD clave-valor, posteriormente muy usada en sistemas de búsqueda
- Page & Brin comienzan a investigar en Stanford. Google comenzaría en 1998

1999

- Doug Cutting escribe la librería de indexación y búsqueda de texto Lucene

2000

- Nace ‘Tokyo Cabinet’: BBDD clave-valor de objetos
- Page & Brin presentan ‘The anatomy of a large-scale Hypertextual Search Engine’

2003

- Nace ‘Nutch’: Motor de búsqueda Web (un ‘Google’ Open Source)
- Google presenta su paper sobre su File System Distribuido: Google File System

2004

- Nace Solr, sobre la librería Lucene, un servicio de búsqueda orientado a sitios web
- Google presenta su paper hablando del paradigma MapReduce aplicado al procesado de grandes datos en clusters

2005

- Nace couchDB, BBDD no relacional, distribuida y orientada a documento
- Nace Hadoop de la mano de Doug Cutting y Mike Cafarella, framework Open Source basado en GFS y Google MapReduce

2006

- Google presenta su paper sobre BigTable, un sistema de persistencia escalable y distribuido para manejar datos de una manera estructurada

2007

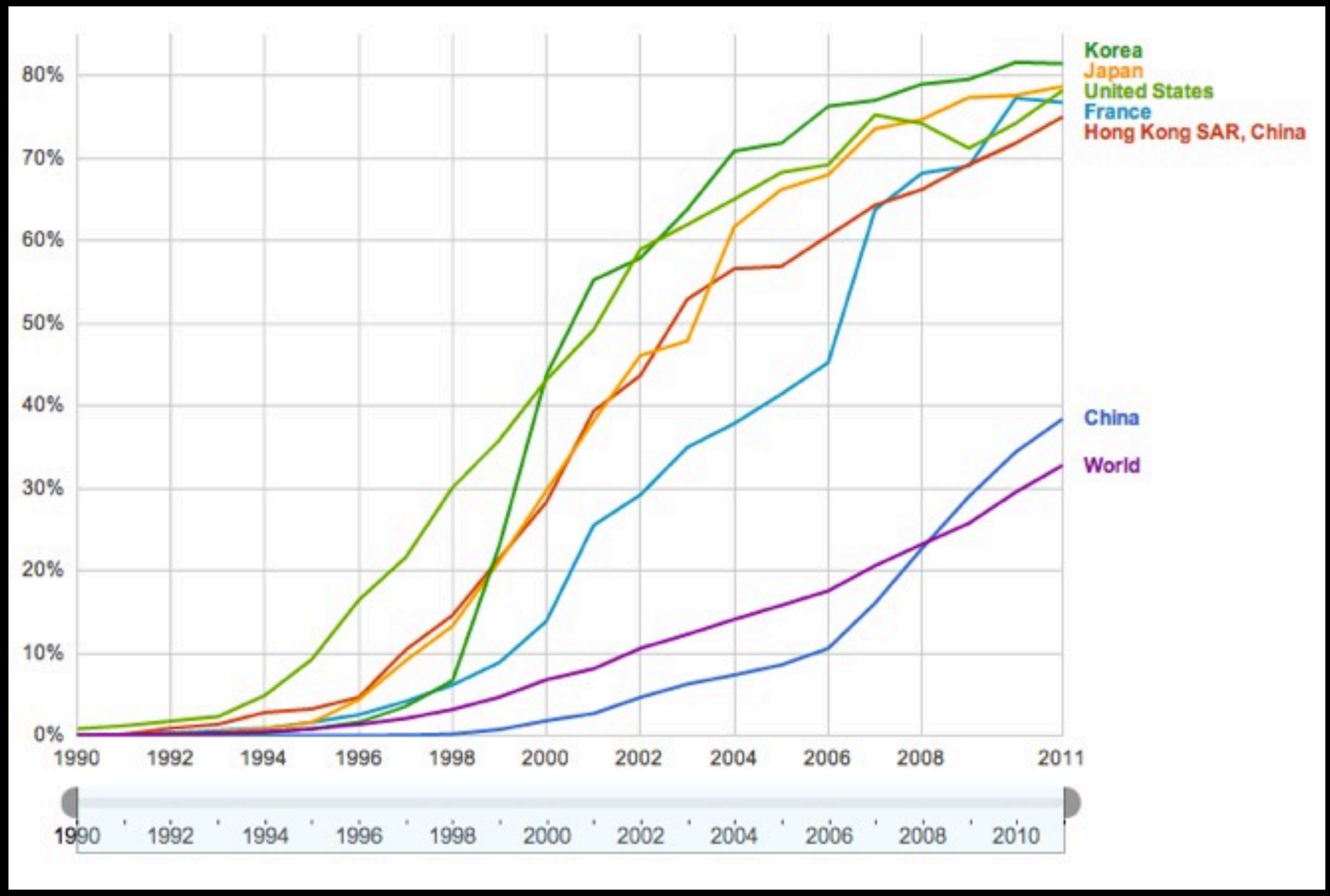
- Nace MongoDB, BBDD no relacional, orientada a documento y de fácil uso e instalación

2008

- Nace Apache HBase, BBDD Open Source, distribuida, orientada a columnas y basada en BigTable de Google
- Nace Apache Cassandra, BBDD Open Source, orientada a columnas y diseñada para escalabilidad lineal. Surge en Facebook

OLAS DETRÁS DE ‘BIG DATA’

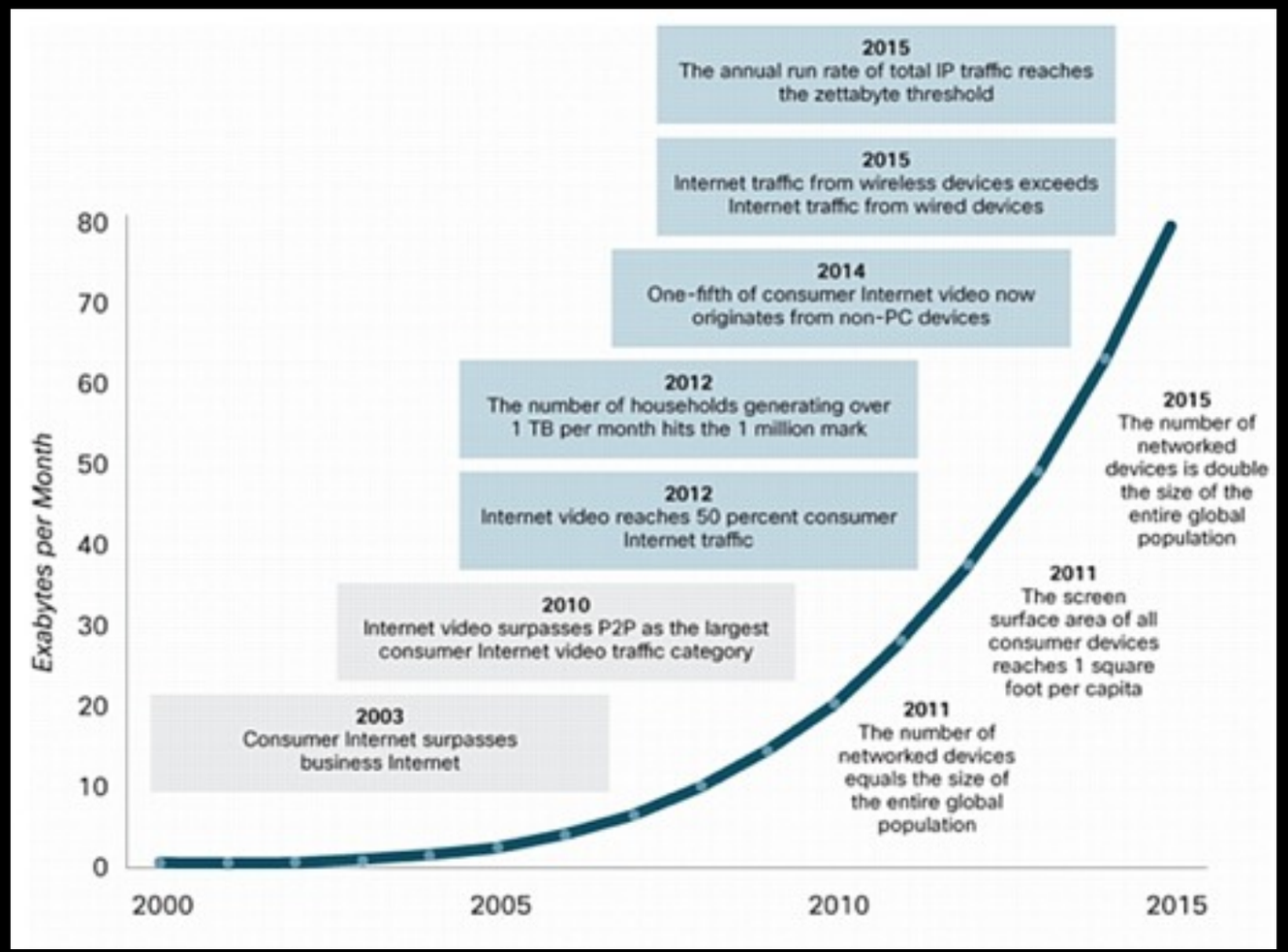
‘Adopción Masiva’: Big Data es el resultado del crecimiento exponencial del uso de Internet



http://www.imweiwei.com/wp-content/uploads/2012/12/2011-internet-users-as-percentage-of-population.png

OLAS DETRÁS DE 'BIG DATA'

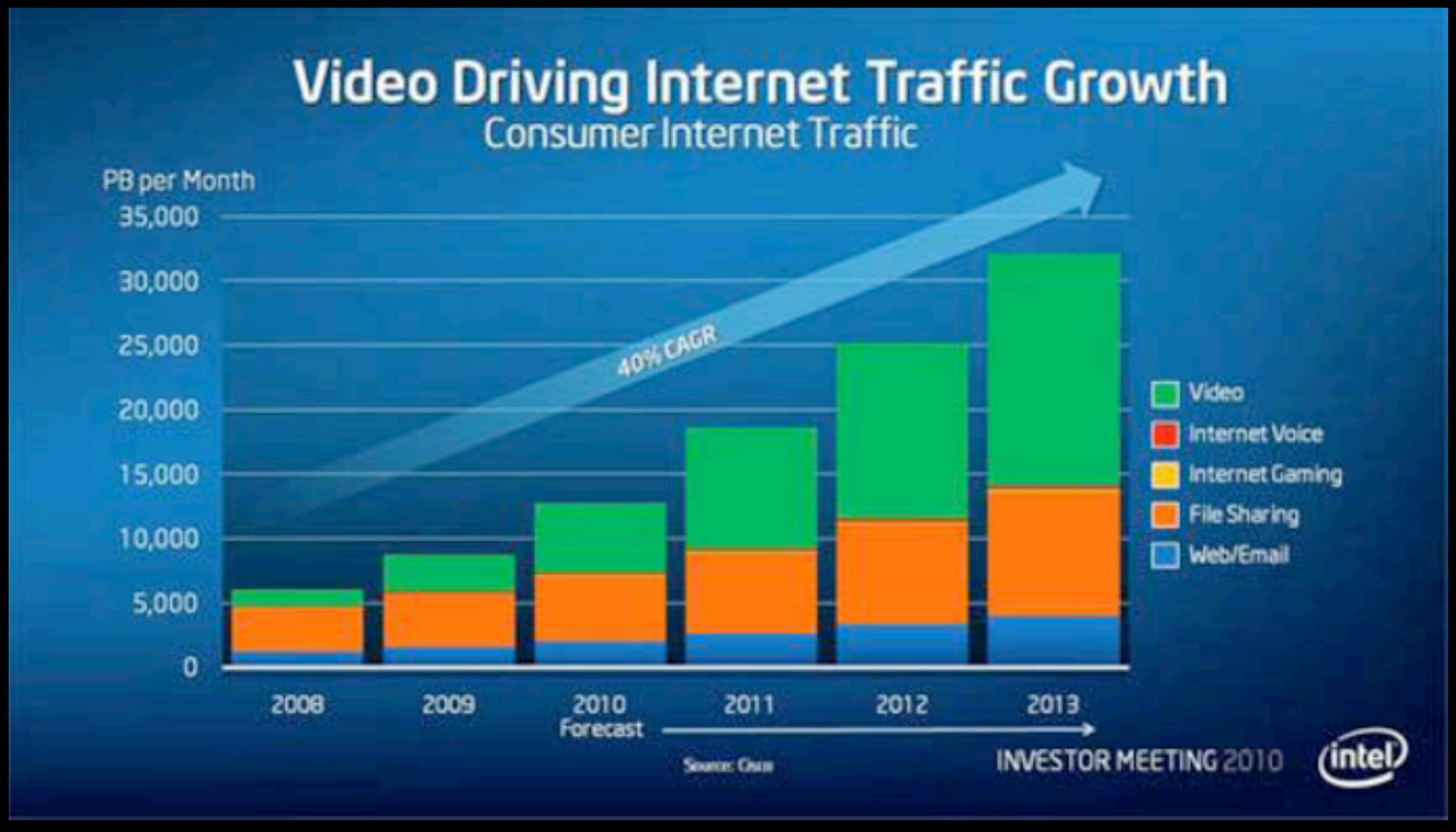
'Adopción Masiva': Big Data es el resultado del crecimiento exponencial del uso de Internet



http://news.cnet.com/8301-1023_3-20067979-93.html

OLAS DETRÁS DE 'BIG DATA'

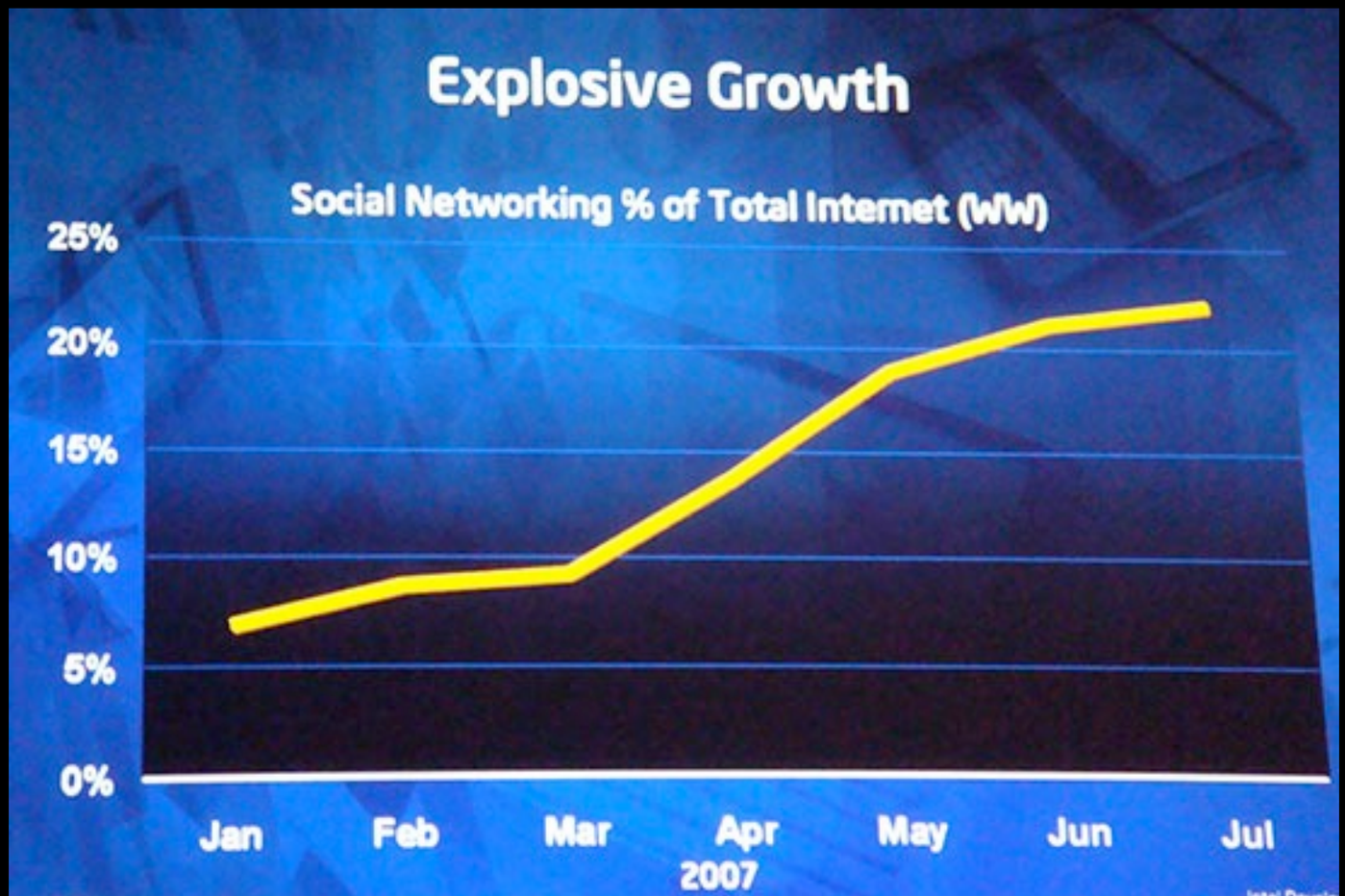
'Adopción Masiva': Big Data es el resultado del crecimiento exponencial del uso de Internet



http://www.louisesteiner.com/7-reasons-need-video-email-software-business/

OLAS DETRÁS DE ‘BIG DATA’

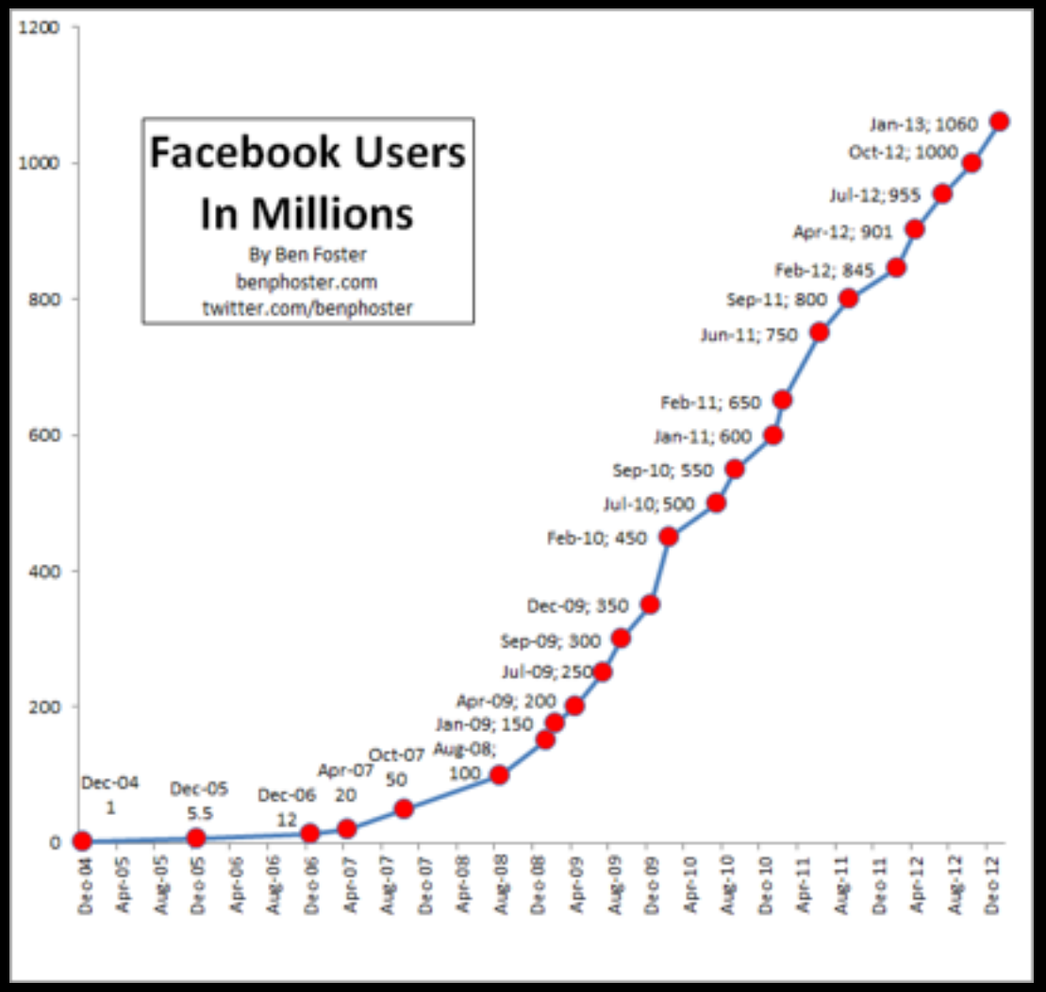
‘Socialización Masiva’: Big Data es el resultado de volcar nuestra vida en la nube



http://www.xbitlabs.com/articles/other/display/idf-f2007-2_5.html

OLAS DETRÁS DE ‘BIG DATA’

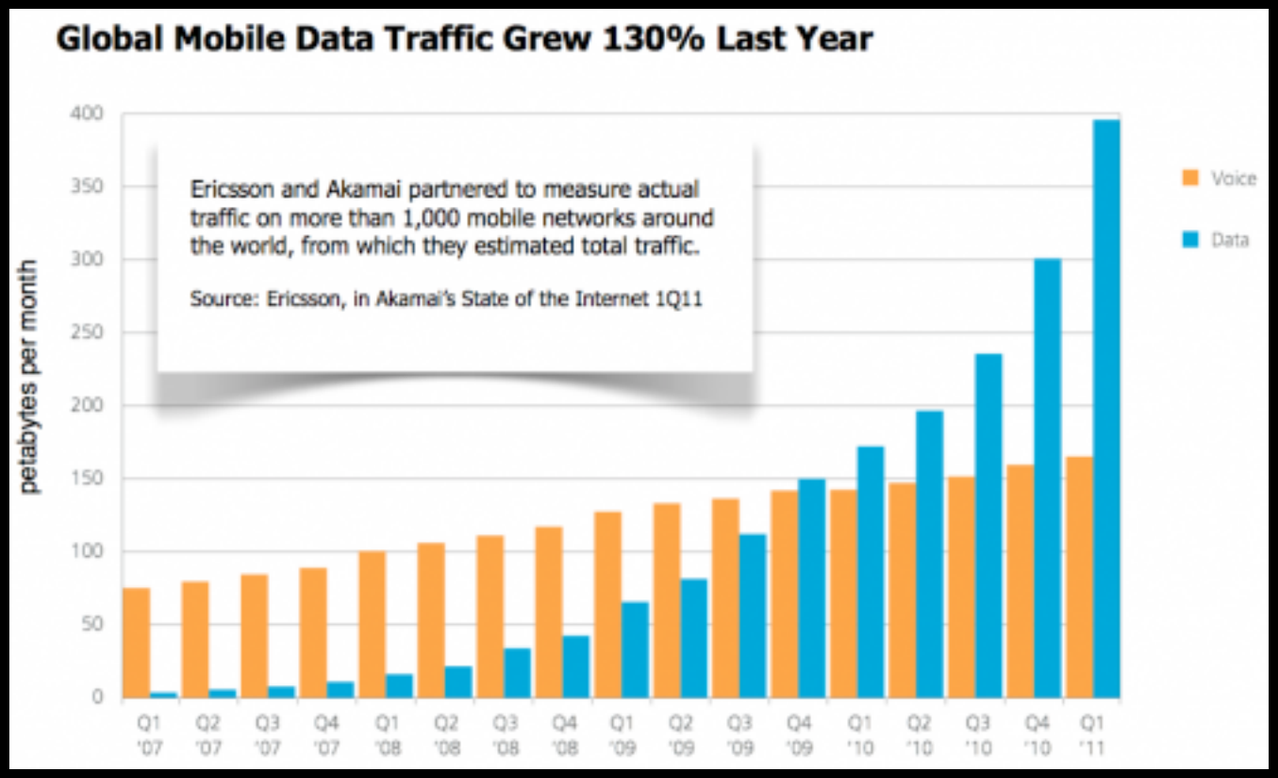
‘Socialización Masiva’: Big Data es el resultado de volcar nuestra vida en la nube



http://www.benphoster.com/facebook-user-growth-chart-2004-2010/

OLAS DETRÁS DE ‘BIG DATA’

‘Sensorización Masiva’: Big Data es el resultado de la computación ubícua



http://www.forbes.com/sites/bretswanson/2011/09/19/damming-the-digital-river/

OLAS DETRÁS DE 'BIG DATA'

- 'Socialización' masiva
- 'Adopción' masiva
- 'Sensorización' masiva

BIG DATA: ECOSISTEMA

EL ECOSISTEMA HADOOP

¿Qué es un Cluster Hadoop?
Un conjunto de servidores (nodos), sobre el que se ejecutan procesos MapReduce y que comparten datos mediante HDFS (Hadoop Distributed File System

MAPREDUCE: Divide y vencerás

MAP:
Función de procesado.
Los datos se particionan y se pasa cada ‘trozo’ a una función ‘map’
La función ‘map’ es sin estado

REDUCE
Función de reducción
La salida del map es la entrada del reduce
Se usa para consolidar y eliminar redundancias

MAPREDUCE: Contar las palabras de un fichero

MAP:
< Hello, 1>
< World, 1>

< Bye, 1>
< World, 1>

REDUCE:
< Hello, 1>
< World, 2>
< Bye, 1>

Fichero:
“Hello World
Bye World”

MAPREDUCE: Control

- Se ejecutan ‘jobs’ que el framework divide en ‘tasks’
- Master JobTracker
- TaskTracker por nodo

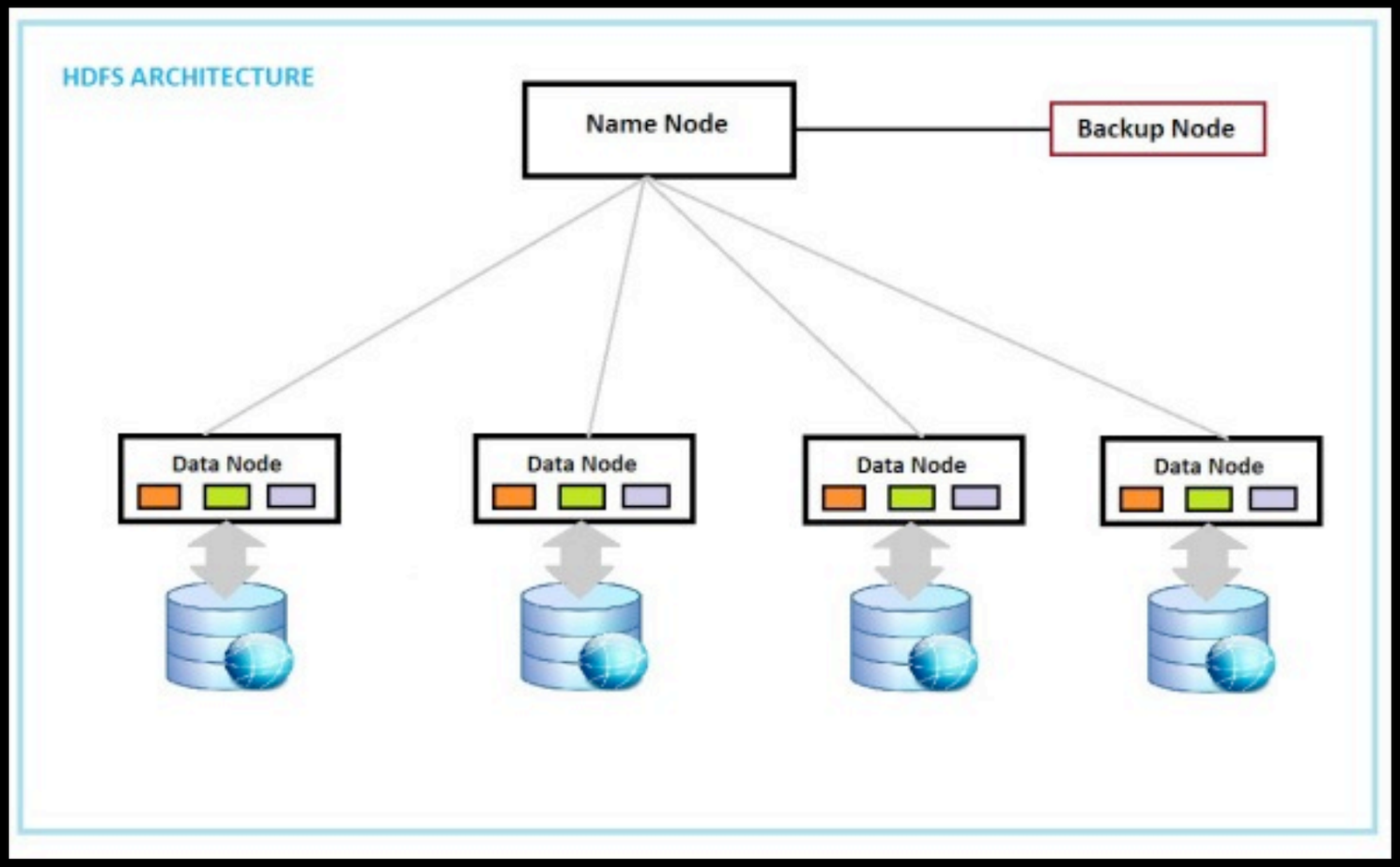
EL ECOSISTEMA HADOOP

HDFS

- Implementa un único sistema de ficheros ‘juntando las capacidades’ de todos los nodos
- Es transparente para el programador
- Se implementa tolerancia a fallos con nodos de ‘backup’

EL ECOSISTEMA HADOOP

HDFS



http://codemphasis.wordpress.com/2012/09/27/big-data-hadoop-hdfs-and-mapreduce/

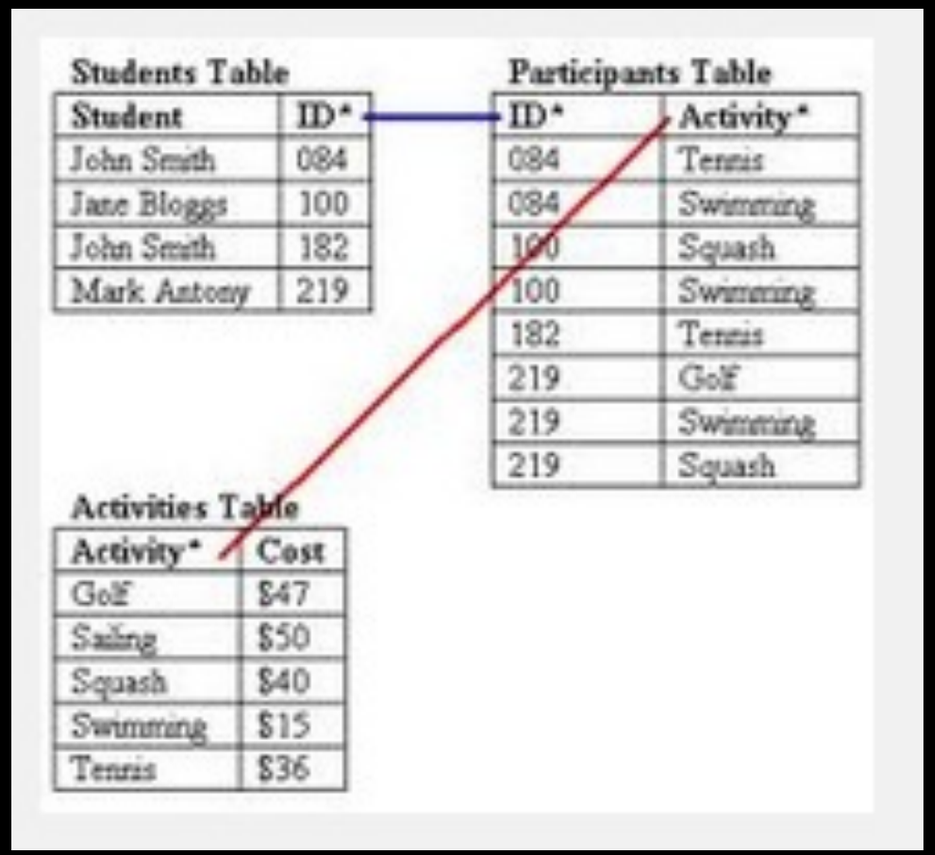
EL ECOSISTEMA HADOOP

The Hadoop Bestiary

Ambari	Deployment, configuration and monitoring
Flume	Collection and import of log and event data
HBase	Column-oriented database scaling to billions of rows
HCatalog	Schema and data type sharing over Pig, Hive and MapReduce
HDFS	Distributed redundant filesystem for Hadoop
Hive	Data warehouse with SQL-like access
Mahout	Library of machine learning and data mining algorithms
MapReduce	Parallel computation on server clusters
Pig	High-level programming language for Hadoop computations
Oozie	Orchestration and workflow management
Sqoop	Imports data from relational databases
Whirr	Cloud-agnostic deployment of clusters
Zookeeper	Configuration management and coordination

http://oreilly.com/data/radarreports/planning-for-big-data.csp

NOSQL



Students Table

Student	ID*
John Smith	084
Jane Bloggs	100
John Smith	182
Mark Antony	219

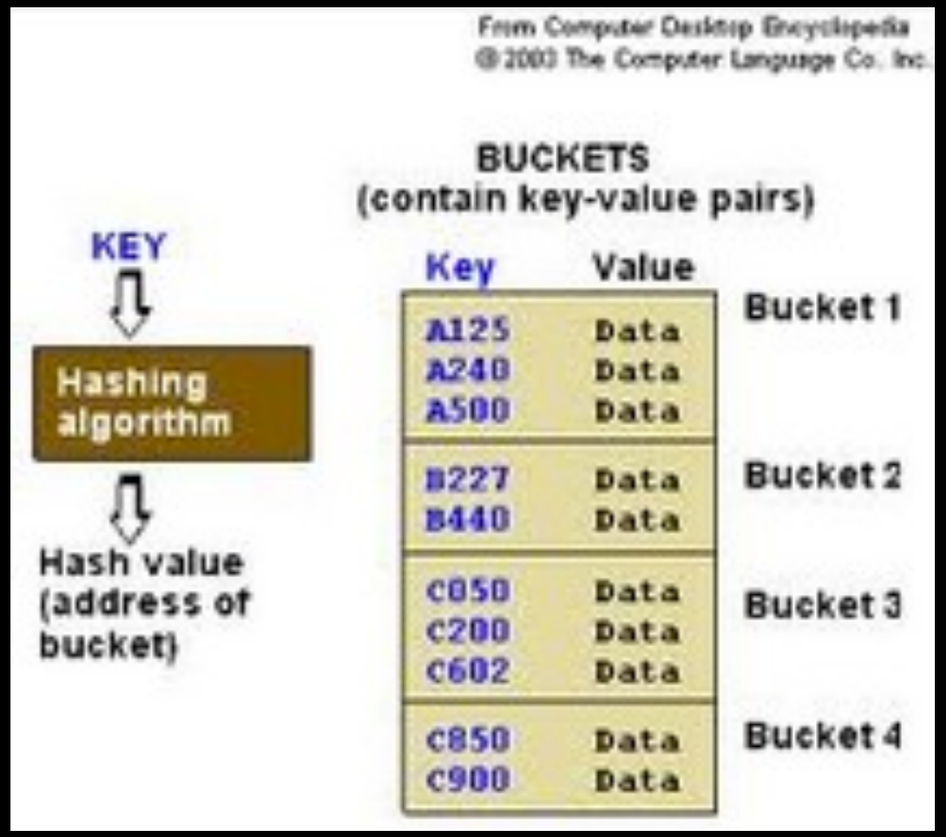
Participants Table

ID*	Activity*
084	Tennis
084	Swimming
100	Squash
100	Swimming
182	Tennis
219	Golf
219	Swimming
219	Squash

Activities Table

Activity*	Cost
Golf	$47
Sailing	$50
Squash	$40
Swimming	$15
Tennis	$36

VS



NOSQL (MONGODB)

#7656 Update | Delete | New Field | Duplicate | Refresh | Text | Collapse


```
{
   "_id": ObjectId("5152303ebc250e1d2d001de7"),
   "contributors": null,
   "truncated": false,
   "text": "RT @SurSiendo: política p2p cambiará las reglas del juego político para siempre, x @bernardosampa: http:
\/\/t.co\/id3VmgKLmm #globalp2p",
   "in_reply_to_status_id": null,
   "id": 3.1669416112439e+17,
   "favorite_count": 0,
   "source": "<a href=\"http: \/\/twitter.com\/download\/iphone\" rel=\"nofollow\">Twitter for iPhone<\/a>",
   "retweeted": false,
   "coordinates": null,
   "created_at_other": "20130326233226",
   "entities": {
     "user_mentions": {
       "0": {
         "id": 469501029,
         "indices": {
           "0": 3,
           "1": 13
       },
         "id_str": "469501029",
         "screen_name": "SurSiendo",
         "name": "SurSiendo"
     },
       "1": {
         "id": 87920683,
         "indices": {
           "0": 84,
           "1": 98
       },
         "id_str": "87920683",
         "screen_name": "bernardosampa",
         "name": "Bernardo Gutiérrez"
     }
```

Problemas de las BBDD relacionales

- Leer datos completos es costoso (“joins”)
- Transacciones (“integridad”)
- Escalabilidad
- Cambio del modelo de datos (migraciones)

Soluciones NoSQL

- Almacenes clave-valor (Redis, BerkeleyDB, Tokyo Cabinet)
- Orientadas a documento (MongoDB, CouchDB)
- Orientadas a columnas (Cassandra, HBase, BigTable)

EL VALOR DE BIGDATA

APLICACIONES

ADVERTISING

- Recomendación
- Segmentación
- Tecnologías: Apache Mahout

APLICACIONES

CONTENIDO

- Búsqueda
- Documentos relacionados
- Tecnologías: Apache Solr, ElasticSearch

TEXTO Y LENGUAJE

- Frecuencias
- Identificación de lenguaje
- Tecnologías: MapReduce/Hadoop

APLICACIONES

REDES SOCIALES

- Recomendaciones de amigos
- Recomendaciones sociales
- Métricas de red
- Tecnologías: Neo4j

APLICACIONES

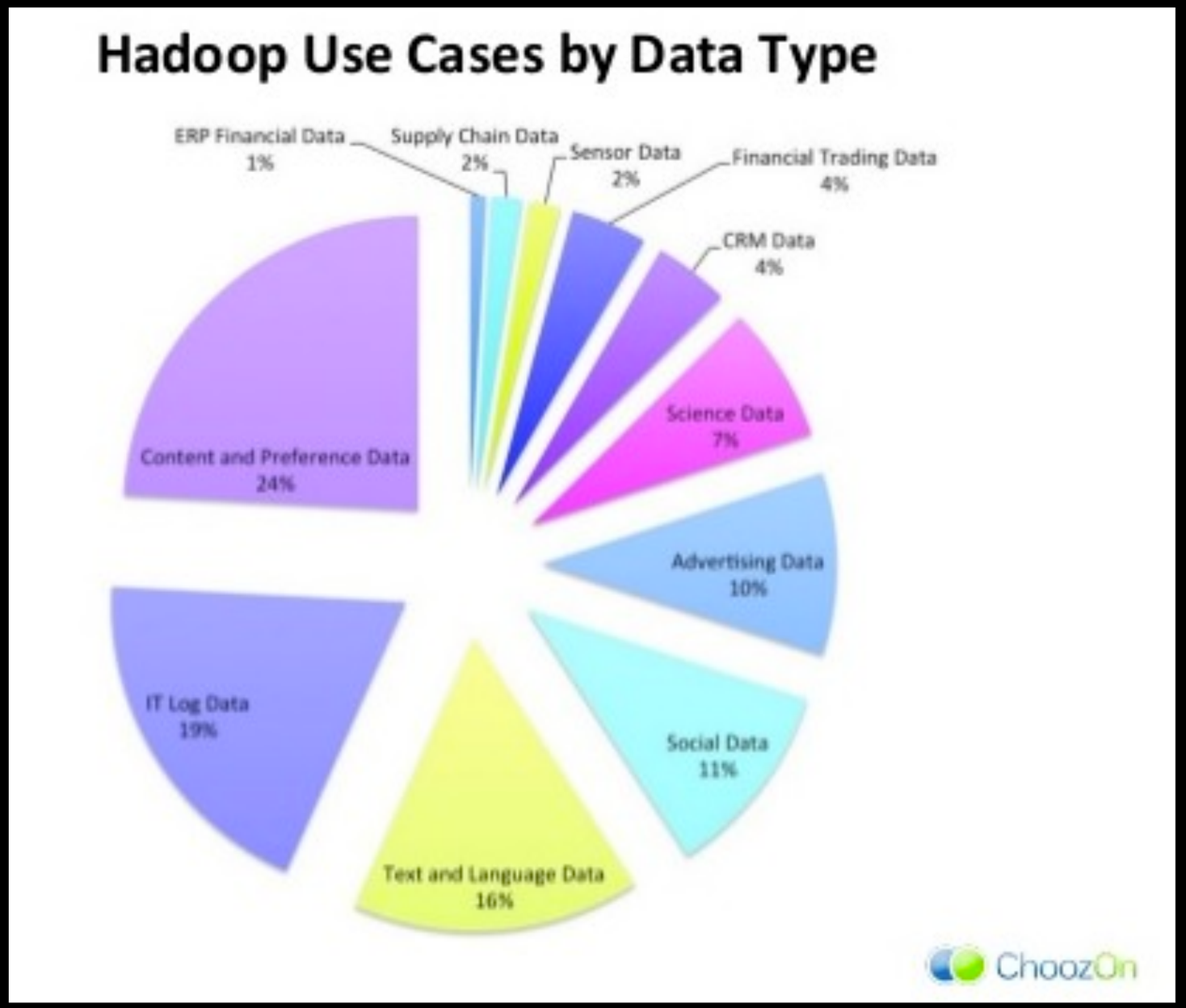
ANALYTICS

- Logs
- ‘Clickstreams’
- Tecnologías: Apache Pig

ESCALABILIDAD WEB

- Bajo tiempo de consulta
- Elasticidad y escalabilidad
- Tecnologías: MongoDB, Cassandra, HBase

APLICACIONES



http://www.slideshare.net/gigaom/the-3vs-of-big-data-variety-velocity-and-volume-from-structuredata-2012

OTRAS CONSIDERACIONES

- La distinción entre ‘Data’ y ‘Big Data’ está desapareciendo
- La Tercera V: Variedad aplica en ‘Data’
  - Escalabilidad
  - Flexibilidad
- El texto sigue siendo el ‘driver’ del Big Data

OTRAS CONSIDERACIONES

- ¿Cuándo usar Hadoop?
  - Big Data != Hadoop !!!!
  - MapReduce es útil cuando se ejecuta una operación simple sobre un gran conjunto de datos
  - MapReduce NO es útil cuando el trabajo estadístico es intensivo

BIG DATA: ESCENARIOS PRÁCTICOS

ESCENARIO PRÁCTICO I

- Reputación + Sentiment On-line
- Volumen: ~= millón de menciones diarias
- Problema: Mostrar la evolución diaria cuantitativa y cualitativa de cada marca
- Solución:
- Apache Solr
  - Escalable
  - Facetado
  - Indexación de campos de consulta

ESCENARIO PRÁCTICO II

- Agregación de noticias de diarios
- Near-real time
- Volumen: Poco. Velocidad: Mucha. Variedad: No
- Solución : Apache Nutch + Hadoop

ESCENARIO PRÁCTICO III

- Buscador local
- Problema: Cálculo de POI cercanos en tiempo real
- Volumen: Sí. Velocidad: Sí. Variedad: No
- Solución: Hadoop+HBase

ESCENARIO PRÁCTICO IV

- Newsletter de ‘deals’ diaria con recomendaciones en función de tu clickstream
- Volumen: Sí. Velocidad: Sí. Variedad: No (pero...)
- Solución: Hadoop + Pig + Mahout + MongoDB

EL FUTURO DE BIG DATA

EL FUTURO DE BIG DATA

SMART DATA

- Multidisciplinariedad
- Reducción de complejidad
- ROI en el análisis
- No es ‘Big Data’: Es Smart Data scalable
- No confundir la herramienta con la utilidad
- Volumen != sofisticación
- Contexto y conocimiento del mundo

EL FUTURO DE BIG DATA

OTROS

- Mercados de Datos
- Streaming Data vs Batch Processing
- Quantified Self
- La ciudad como productora de datos

EL BIG DATA EN NUESTRAS VIDAS

EL BIG DATA EN NUESTRAS VIDAS

DOWNSIDES

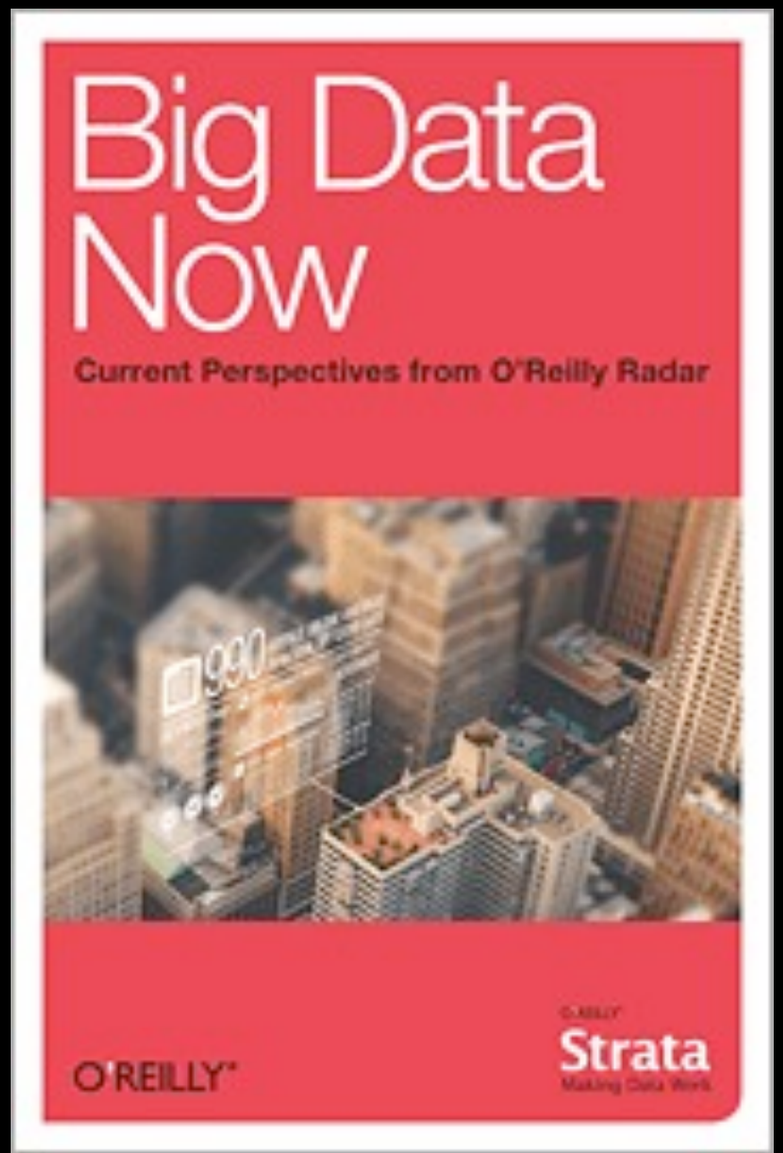
- Solucionismo
- Algorithms vs Human Intervention
- Privacidad y control
- Economía de la atención
- Big Data is great but don’t forget intuition

EL BIG DATA EN NUESTRAS VIDAS

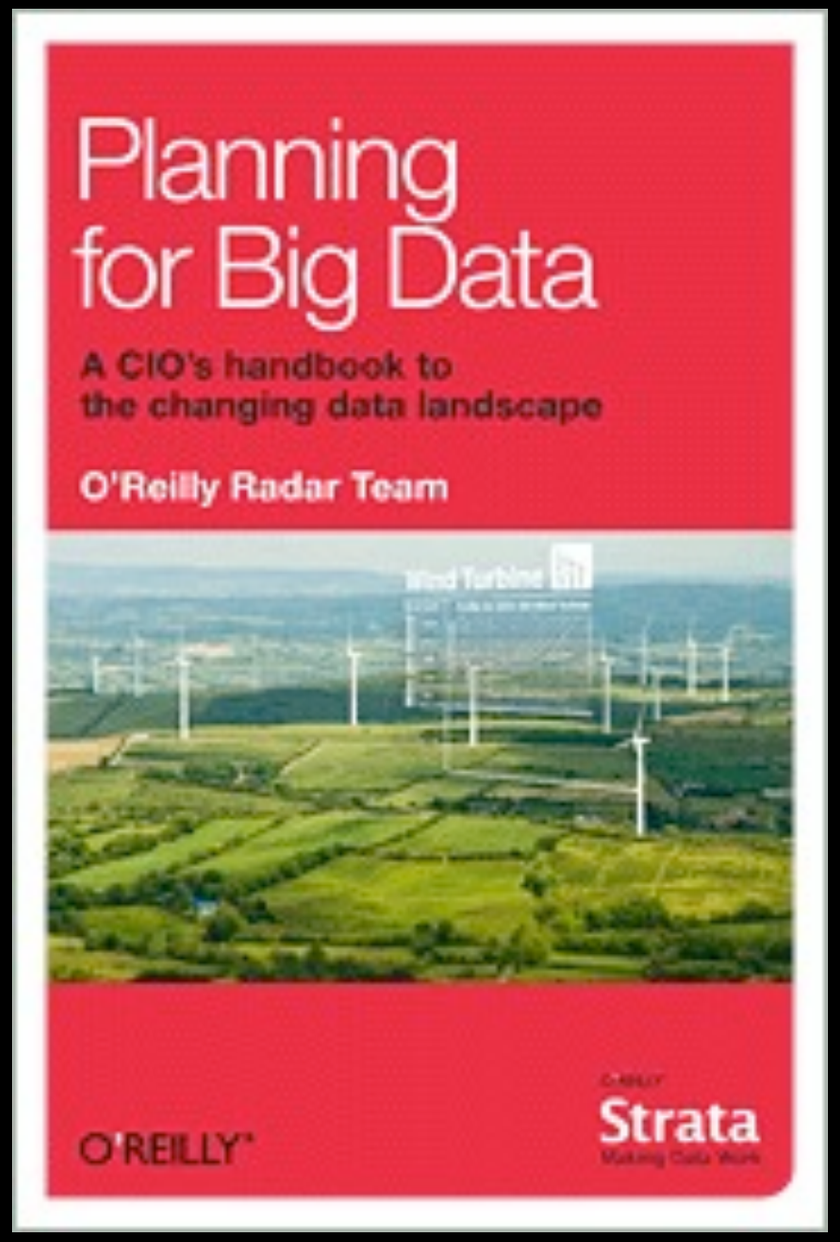
UPSIDES

- Open Data
- eHealth
- Soluciones Bottom-up y Crowdsourcing

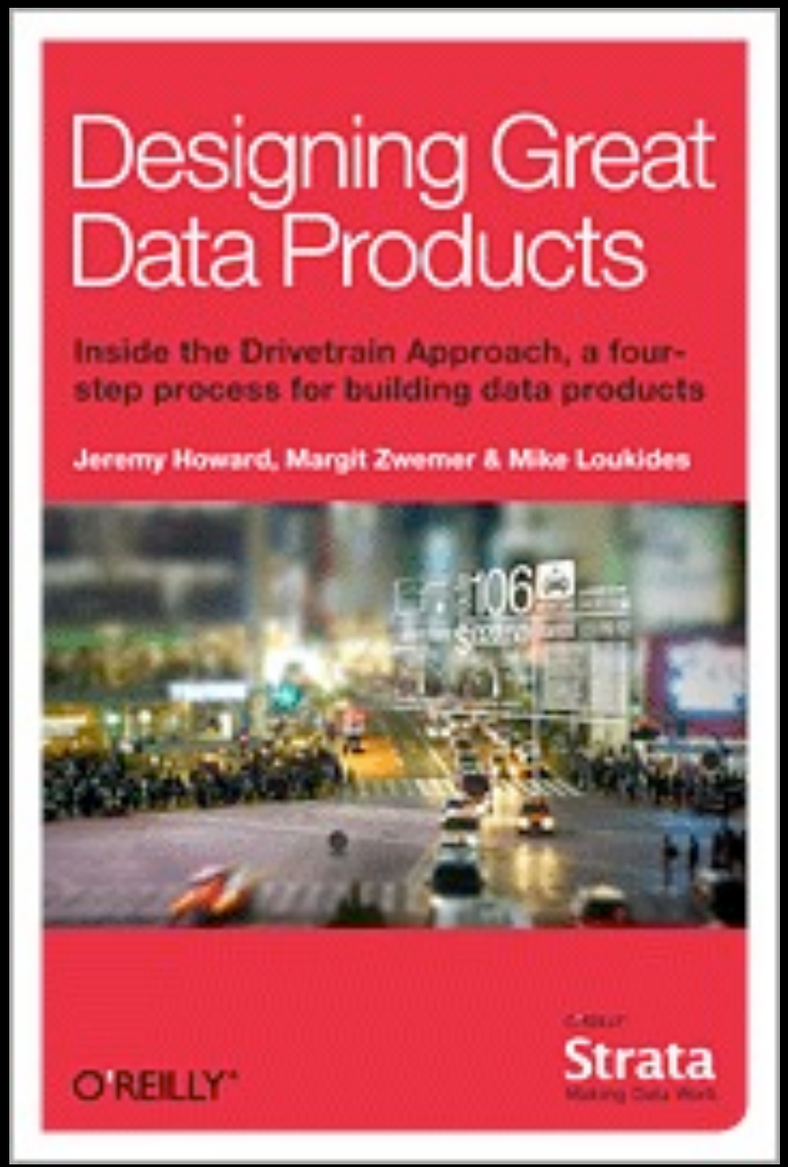
REFERENCIAS

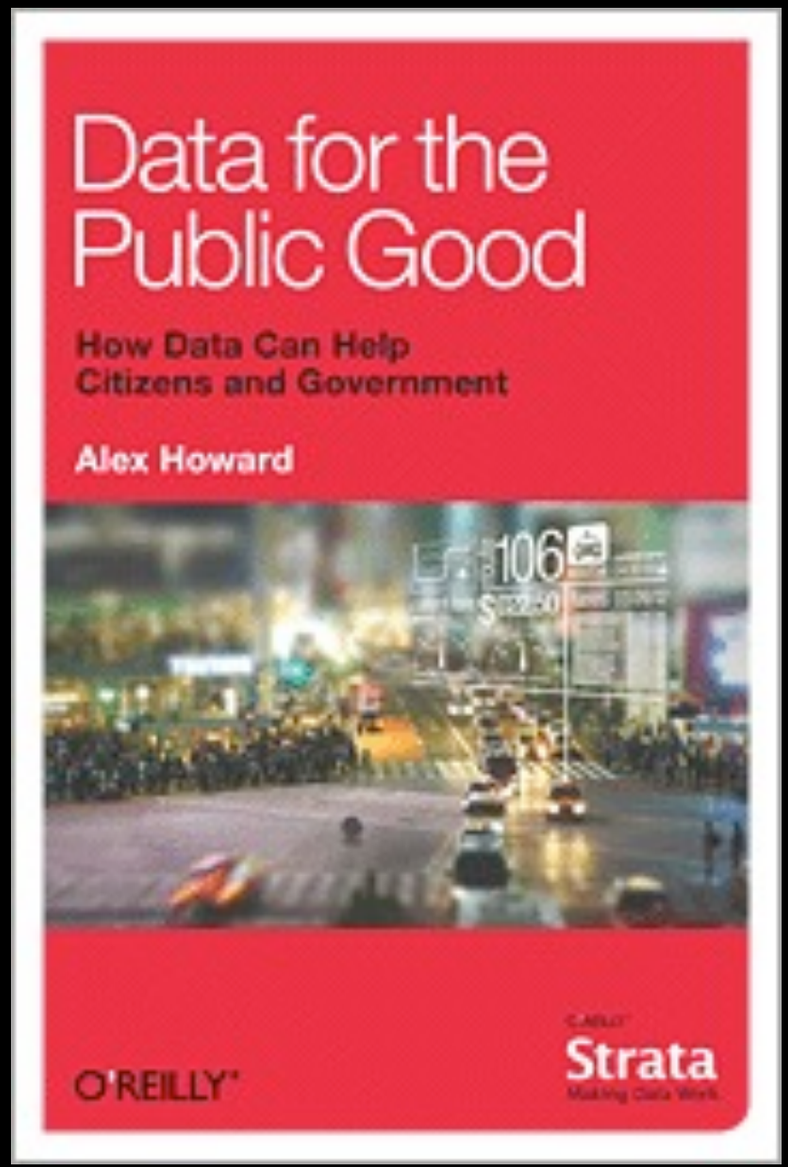


http://shop.oreilly.com/product/0636920022640.do
http://oreilly.com/data/radarreports/big-data-now-2012.csp



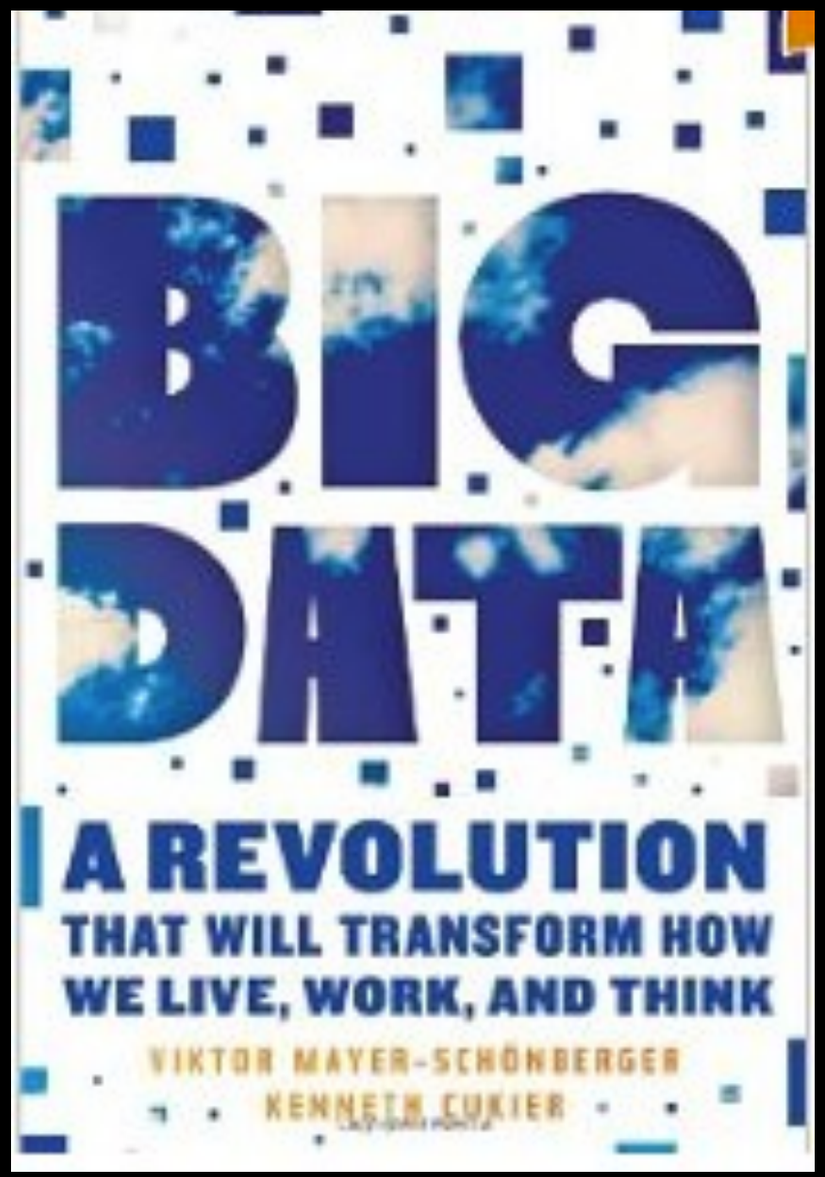
http://oreilly.com/data/radarreports/planning-for-big-data.csp



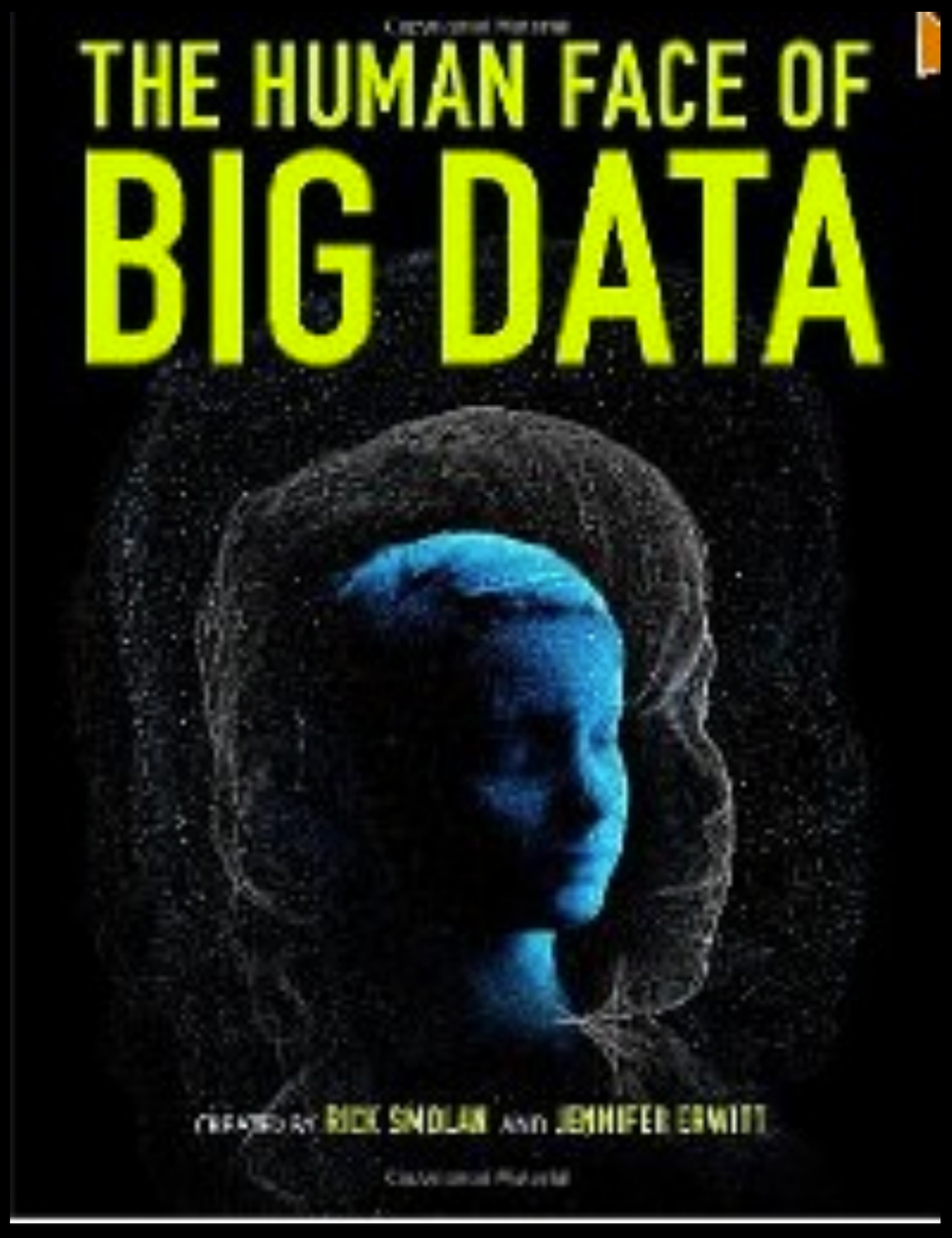


http://shop.oreilly.com/product/0636920026082.do

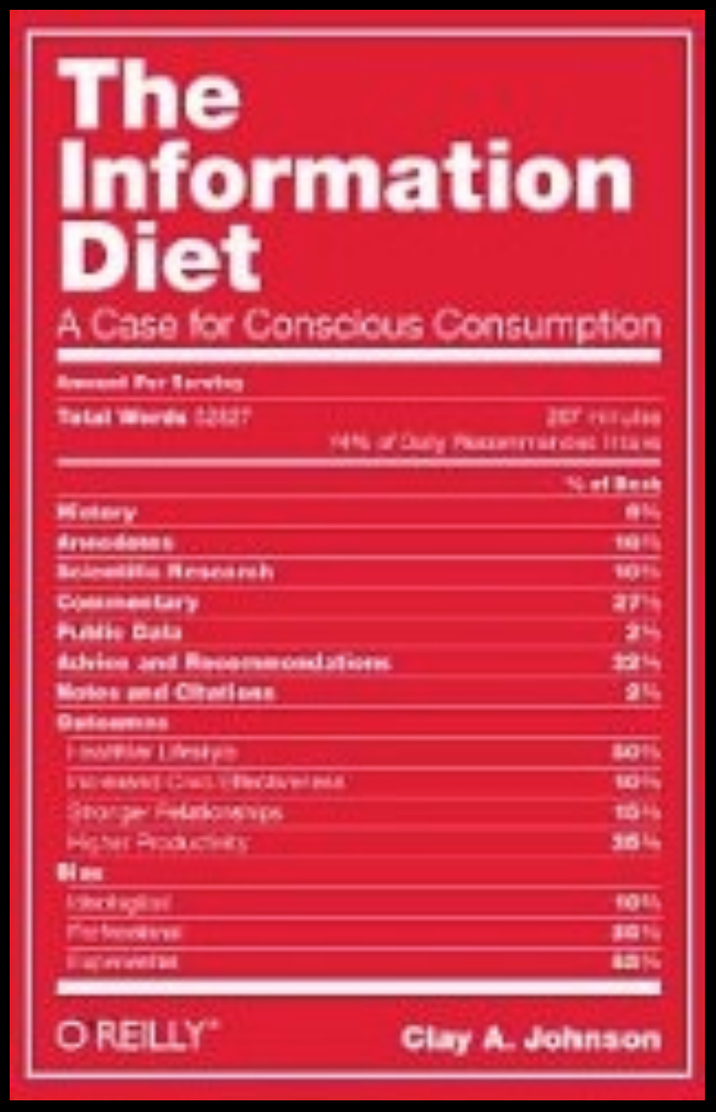
http://shop.oreilly.com/product/0636920025580.do



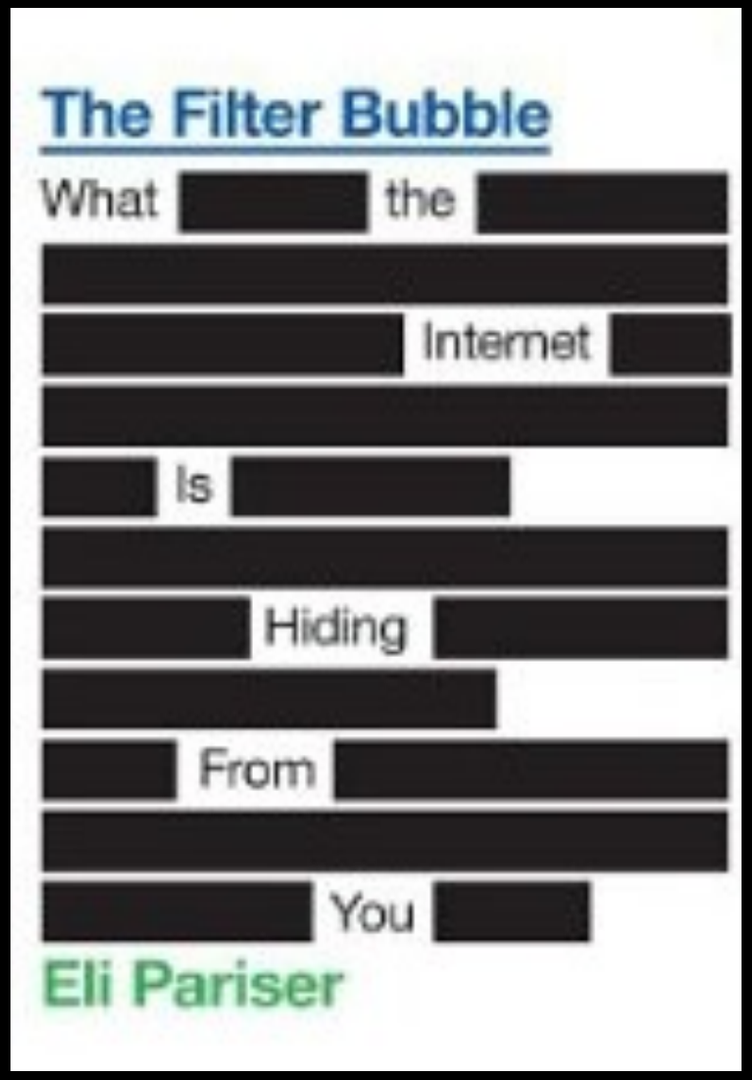
http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695/



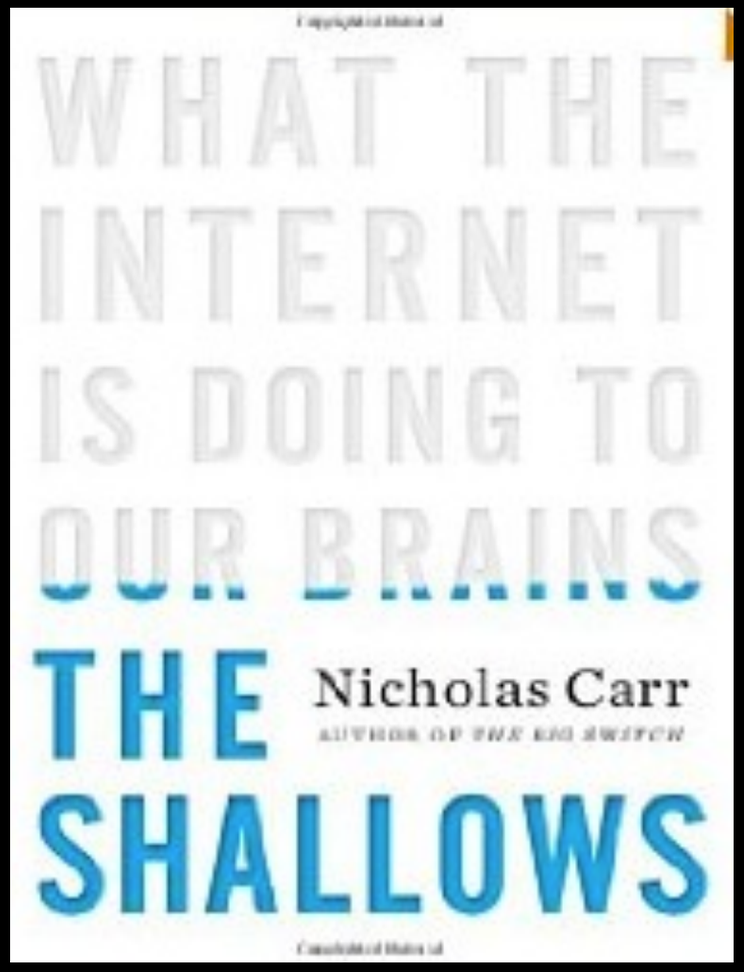
http://www.amazon.com/The-Human-Face-Big-Data/dp/1454908270/



http://www.amazon.com/The-Information-Diet-Conscious-Consumption/dp/1449304680

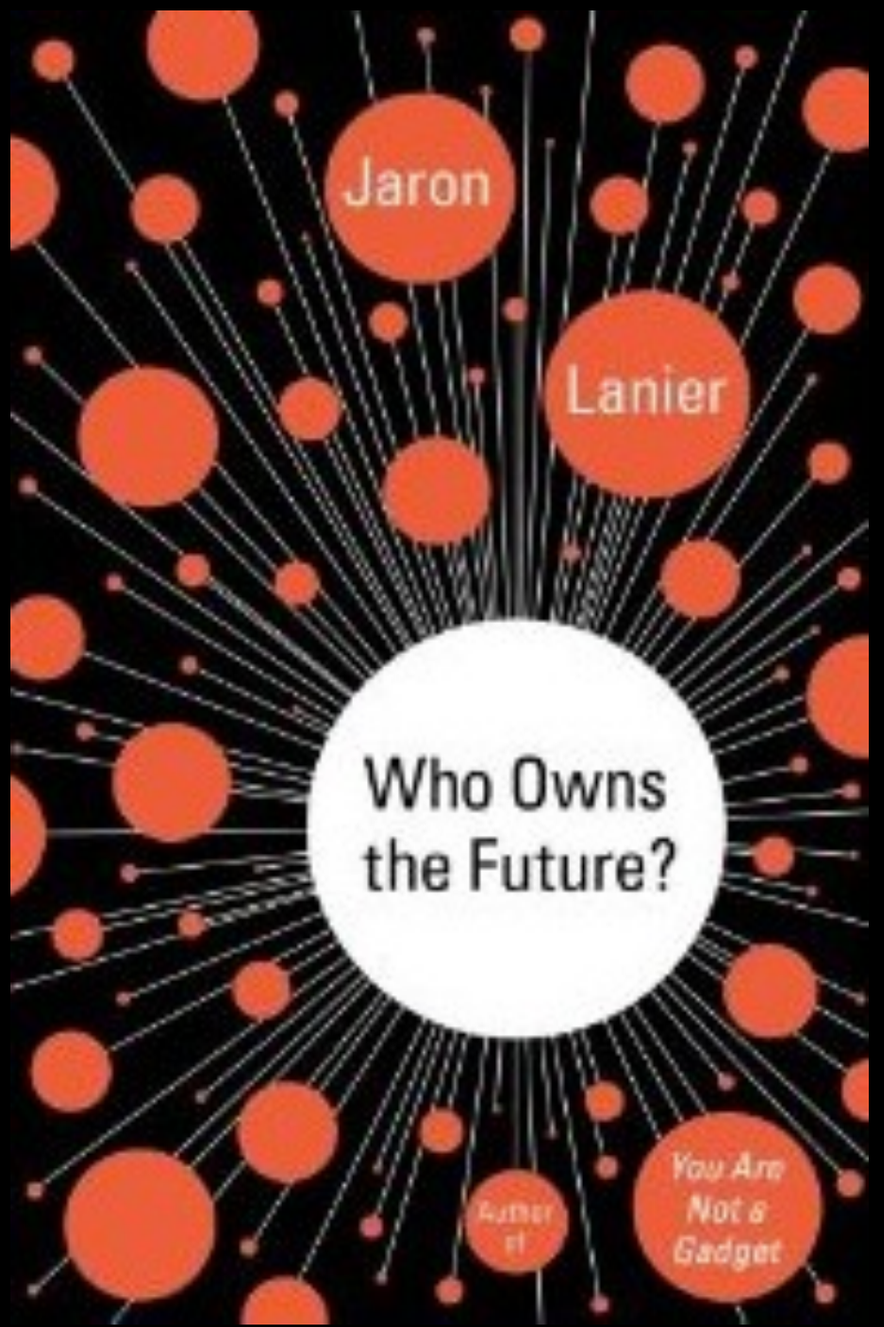


http://www.amazon.com/The-Filter-Bubble-Internet-Hiding/dp/1594203008/



http://www.amazon.com/The-Shallows-Internet-Doing-Brains/dp/0393072223/

http://www.amazon.com/Save-Everything-Click-Here-Technological/dp/1610391381



http://www.amazon.com/Who-Owns-Future-Jaron-Lanier/dp/1451654960/

Outliers

Because differences matter.

www.outliers.es
@outliers_es