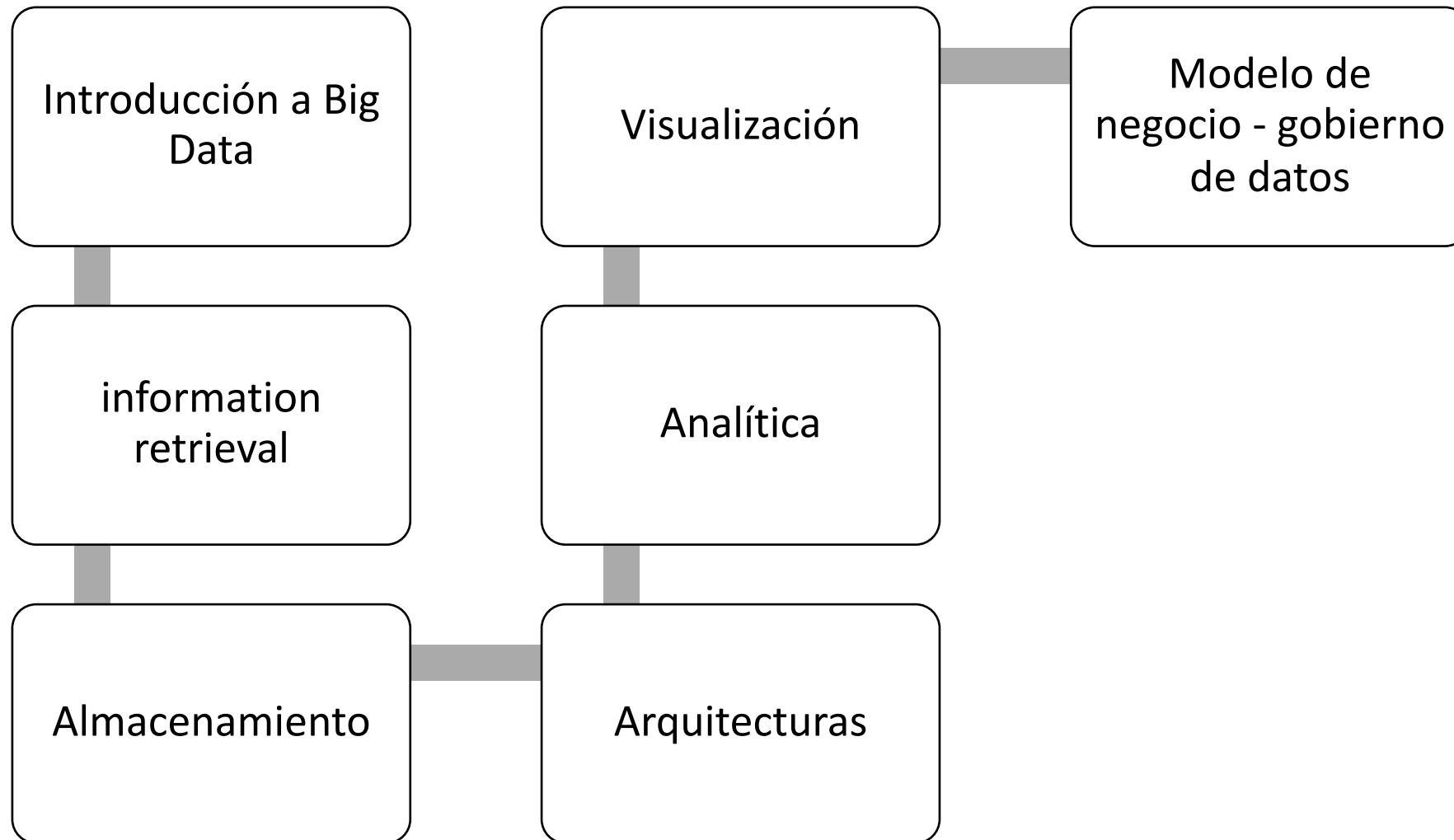


# Temas generales del curso



# Evaluación del curso

## **Primer periodo 35%**

Tareas - Talleres	10%
Exposición	15%
Evaluación No 1	10%

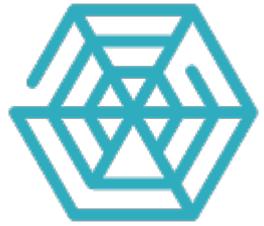
## **Segundo periodo 35%**

Tareas - Talleres	20%
Evaluación No 2	15%

## **Tercer periodo 30%**

**Proyecto final**

# Requisitos de software Básicos



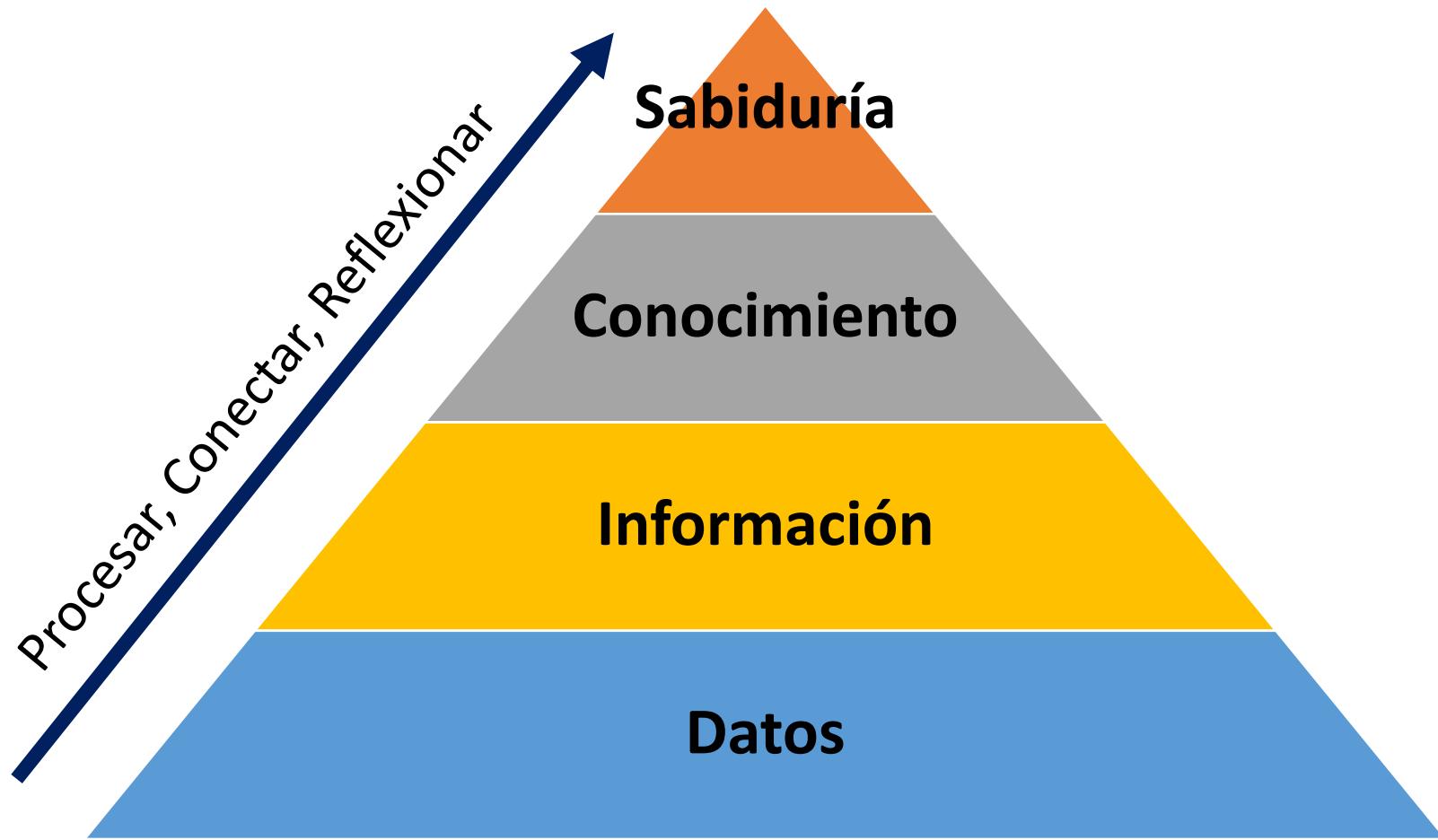
Web Scraper



Open for Innovation ®  
**KNIME**

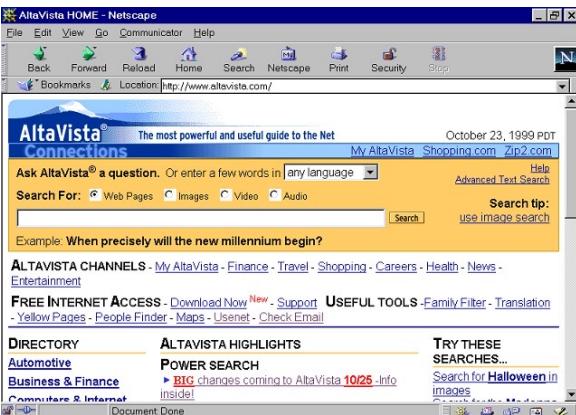


# La Información



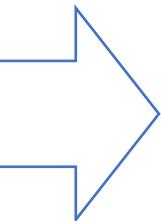
Un dato es una representación simbólica (**numérica, alfabética, algorítmica, espacial, etc.**) de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades.





## Búsquedas basados en palabras clave

En 1992 - 50 millones de páginas web, y actualmente supera 12.000 millones de sitios web, a los que diariamente se les suma a raíz de 4400 por día.



Preocupación por la escalabilidad del motor de búsqueda y las limitaciones que tienen las arquitecturas centralizadas de índices para dar unos tiempos de respuesta aceptables para el usuario.

### The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

{sergey, page}@cs.stanford.edu

Computer Science Department, Stanford University, Stanford, CA 94305

#### Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at <http://google.stanford.edu/>

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every

**1996 Sergey Brin y Lawrence Page**

- Google File System (2003)
- Framework de procesamiento de datos distribuido MapReduce (2004)

# YAHOO!

(mantenidas de forma manual)

- Kylobyte (KB) =  $10^3$  = 1,000 bytes
- Megabyte (MB) =  $10^6$  = 1,000,000 bytes
- Gigabyte (GB) =  $10^9$  = 1,000,000,000 bytes
- Terabyte (TB) =  $10^{12}$  = 1,000,000,000,000 bytes
- Petabyte (PB) =  $10^{15}$  = 1,000,000,000,000,000 bytes
- Exabyte (EB) =  $10^{18}$  = 1,000,000,000,000,000,000 bytes
- Zettabyte (ZB) =  $10^{21}$  bytes
- Yottabyte (YB) =  $10^{30}$  bytes
- Quintillón (QB)=  $10^{33}$  bytes

**Surgen nuevos tipos de datos y necesidades que actualmente los sistemas no son suficientemente buenos o adecuados para poder atacar estos problemas pues las empresas son más exigentes y buscan exprimir al máximo sus recursos para obtener el mayor beneficio.** Sería semejante a escuderías de F1 que buscan superar al rival buscando la diferencia hasta en los grados de regulación de un alerón, analizando y optimizando al mayor detalle.



# Problemas actuales ...

## Tipos de datos

### Variedad

- Han surgido nuevos tipos de datos que se quieren almacenar: datos no estructurados.
- Las BD Relacionales no pueden almacenar este tipo de datos.

## Escalabilidad

- En búsqueda de la rapidez y rendimiento en consultas o procesamiento de datos se busca escalar siempre en horizontal.

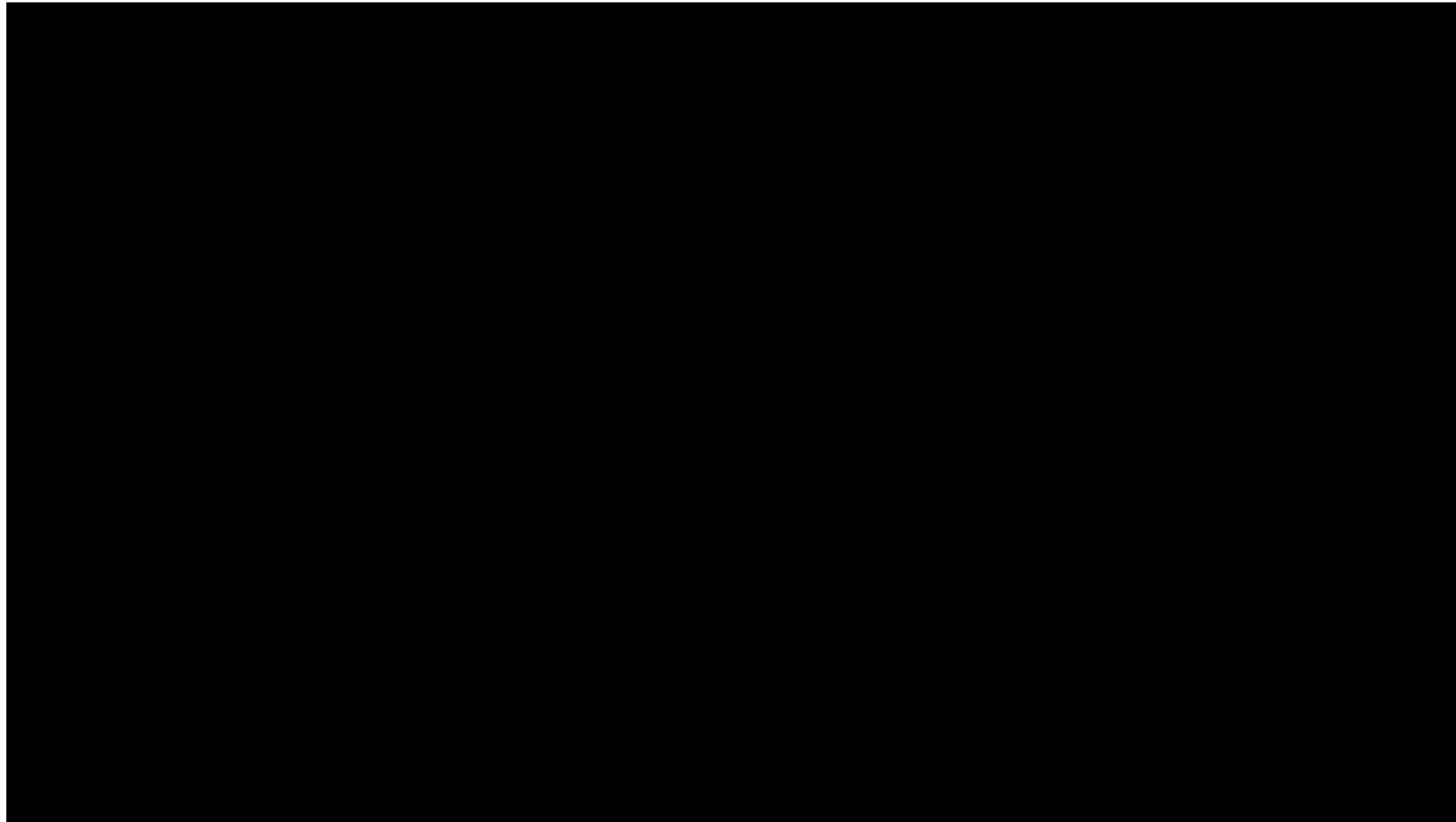
## Modelo relacional

- El modelo relacional no da soporte para todos los problemas. No podemos atacar todos los problemas con el mismo enfoque, queremos optimizar al 100% nuestro sistema y no podemos ajustar nuestros sistemas a estas BD.

## Velocidad

- Esta es una de las "3 V's" del Big Data (velocidad, variedad, volumetría). La velocidad de generación de datos hoy en día es muy elevada, simplemente hay que verlo con las redes sociales actuales, aunque las empresas medias y muchas de las grandes no se ven afectadas por ello.

# **Big Data Introducción**



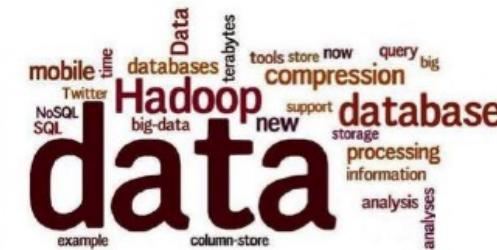
INTERNET | Campus Party Europa 2013

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



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 Chavis - Fotolia

<http://www.elmundo.es/elmundo/2013/09/03/navegante/1378243782.html>

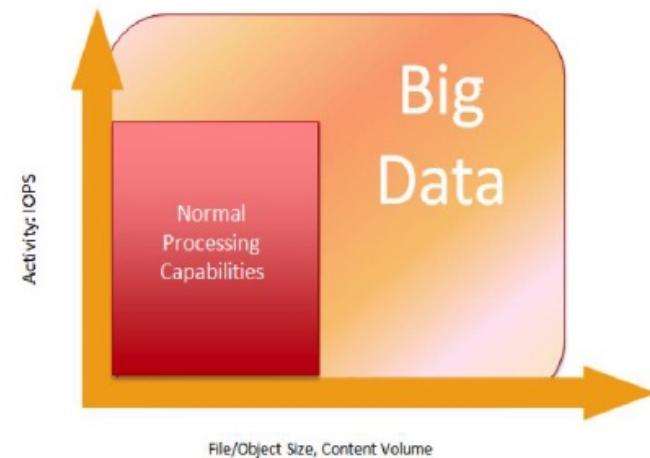
## No hay una definición estándar



**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 ...



## Definiciones ...

*"Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son demasiado grandes y difíciles de procesar con las bases de datos y el software tradicionales"* (ONU, 2012)

*"Es un conjunto de datos cuyo tamaño está más allá de la capacidad de la mayoría de los software utilizados para capturar, gestionar y procesar la información dentro de un lapso tolerable de tiempo."*

*"Datos masivos es un término que hace referencia a una cantidad de datos tal que supera la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable."*

*El término "Big Data" ha sido vapuleado en los últimos años bajo la acusación de que los de Marketing y los Analistas han estirado y comprimido el término para llevarlo a cubrir multitud de problemas, tecnología y productos. Sin embargo en esencia Big Data sigue siendo lo mismo que planteó Doug Laney en 2001, **las tres Uves, Volumen, Velocidad y Variedad** y sigue señalando desafíos que exigen recursos y procesos de computación no-habituales. (Seth Grimes, Alta Plana Corporation)*

# ¿Qué caracteriza a un sistema Big Data?

## Escalabilidad lineal

es decir, que permita aumentar la capacidad de procesamiento linealmente añadiendo nuevo hardware de forma ilimitada.

## Tolerancia a fallos

de tal forma que si uno o varios nodos se averían, el sistema siga funcionando sin pérdida de disponibilidad ni pérdida de ningún dato.

## Despliegue sobre hardware económico de propósito general

(inexpensive commodity hardware) que permita la creación de granjas de servidores con un número elevado de nodos con unos costes sostenidos. También tienen que permitir el despliegue en Cloud (cada vez más habitual sobre todo en startups).

## Procesamiento distribuido y localidad de los datos

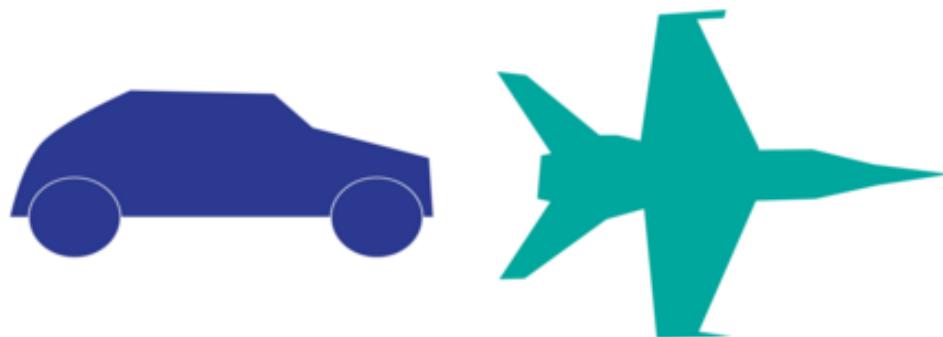
entendido como la ejecución de los procesos analíticos se realizan lo más cercanos de donde se encuentra el dato almacenado, evitando tanto el trasiego de la información como el cuello de botella que puede suponer un almacenamiento centralizado.

# Traditional vs Big Data

AMOUNT OF DATA (VOLUME)

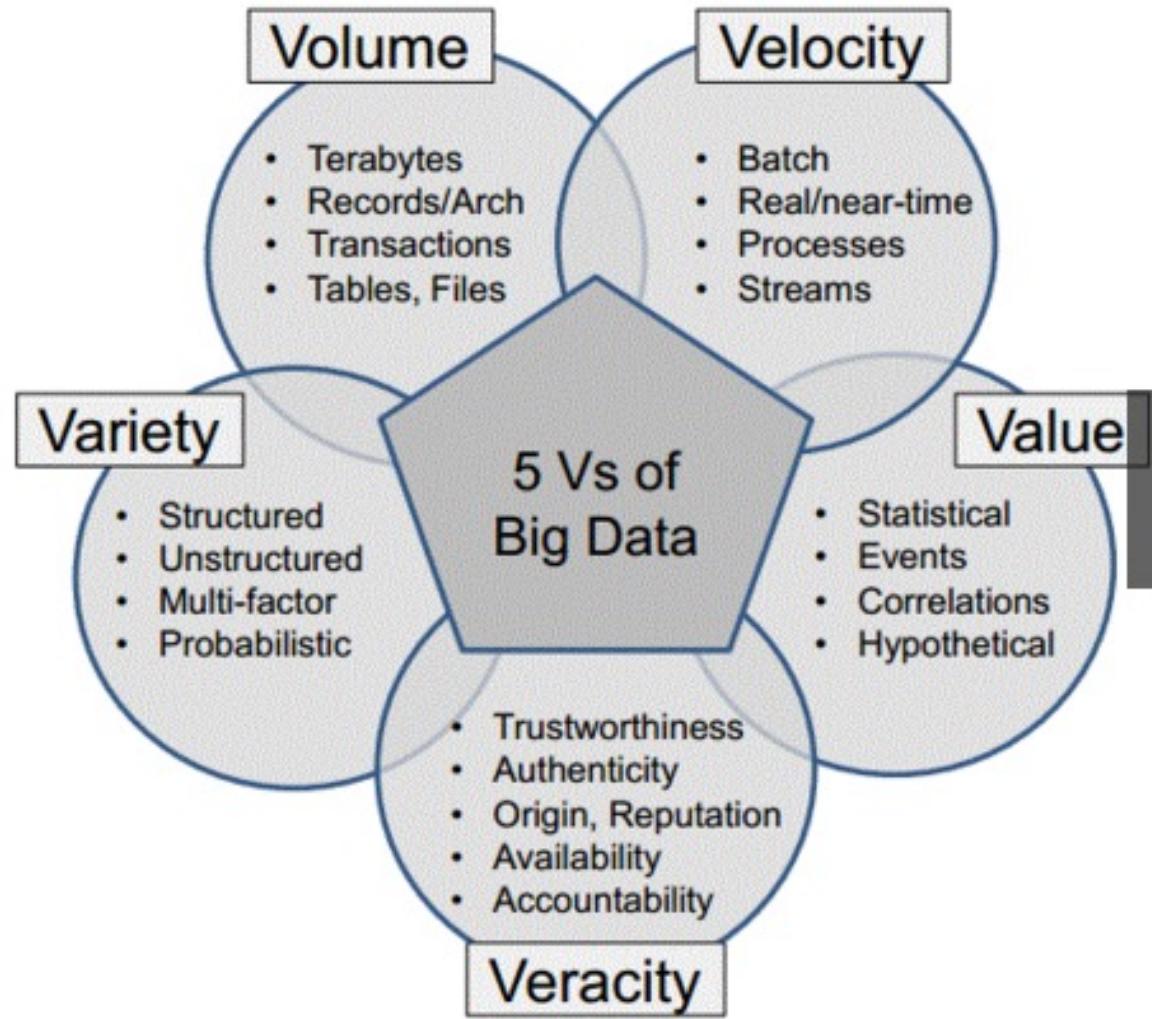


RATE OF DATA GENERATION AND TRANSMISSION (VELOCITY)



TYPES OF STRUCTURED AND UNSTRUCTURED DATA (VARIETY)





# Volume: grandes volúmenes de información

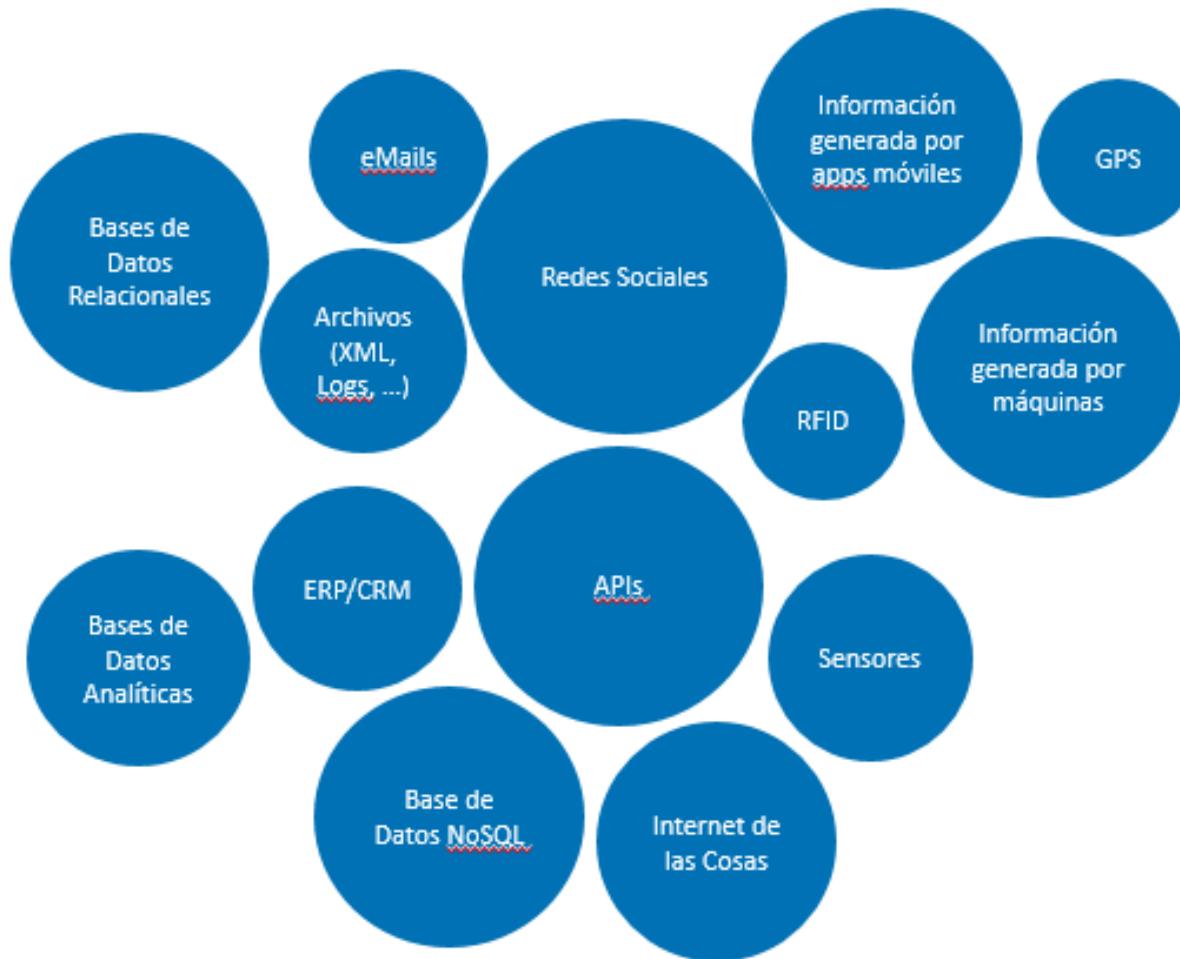
Se está pasando de hablar en Gigabytes o Terabytes a tamaños de datos de Petabytes, Exabytes o Zettabytes. Volúmenes que se nos escapan.

Unidades de información (del Byte)			
Sistema Internacional (Decimal)		ISO/IEC 80000-13 (Binario)	
Multiplo - (Símbolo)	SI	Multiplo - (Símbolo)	ISO/IEC
kilobyte (kB)	$10^3$	Kibibyte (KiB)	$2^{10}$
Megabyte (MB)	$10^6$	Mebibyte (MiB)	$2^{20}$
Gigabyte (GB)	$10^9$	Gibibyte (GiB)	$2^{30}$
Terabyte (TB)	$10^{12}$	Tebibyte (TiB)	$2^{40}$
Petabyte (PB)	$10^{15}$	Pebibyte (PiB)	$2^{50}$
Exabyte (EB)	$10^{18}$	Exbibyte (EiB)	$2^{60}$
Zettabyte (ZB)	$10^{21}$	Zebibyte (ZiB)	$2^{70}$
Yottabyte (YB)	$10^{24}$	Yobibyte (YiB)	$2^{80}$

Véase también: Nibble · Byte · Octal

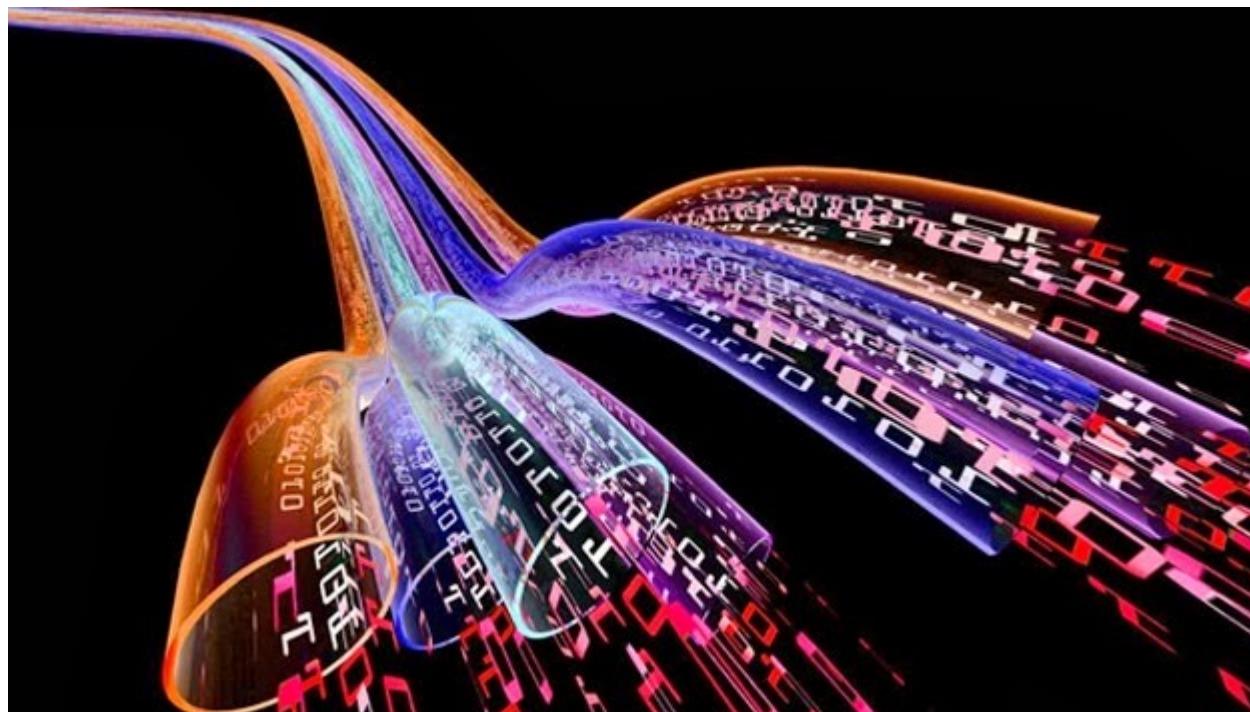
## Variety: información de tipos muy diversos

Ya no solo tenemos información estructurada en Bases de Datos o Archivos.  
Ahora empezamos a tener información con tipos diferentes y totalmente desestructurada



## **Velocity: velocidad con la que se genera la información**

La velocidad a la que se genera esta información hace imposible gestionarla con sistemas de base de datos convencionales. Las empresas y las personas ya no quieren estar al día, quieren “estar al segundo”



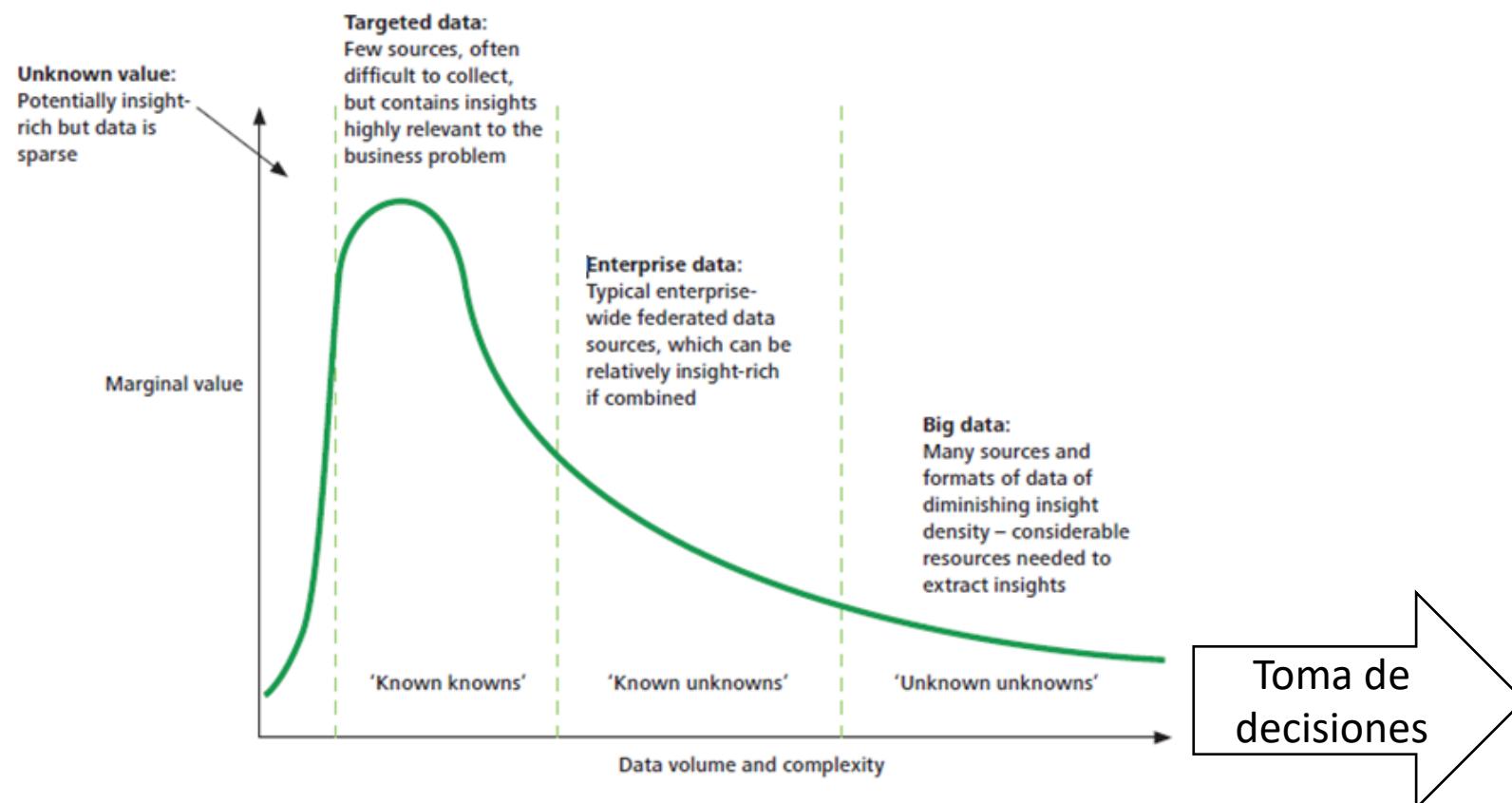
**Veracity: La veracidad puede entenderse como el grado de confianza que se establece sobre los datos a utilizar**

Dentro de la caracterización del Big Data la Veracidad determina su cuarta dimensión, y es de gran importancia para un analista de datos, ya que la veracidad de los mismos determinará la calidad de los resultados y la confianza en los mismos. Por lo tanto un alto volumen de información que crece a velocidad muy rápida y basada en datos estructurados y desestructurados y provenientes de una gran variedad fuentes, hacen inevitable dudar del grado de veracidad de los mismos.

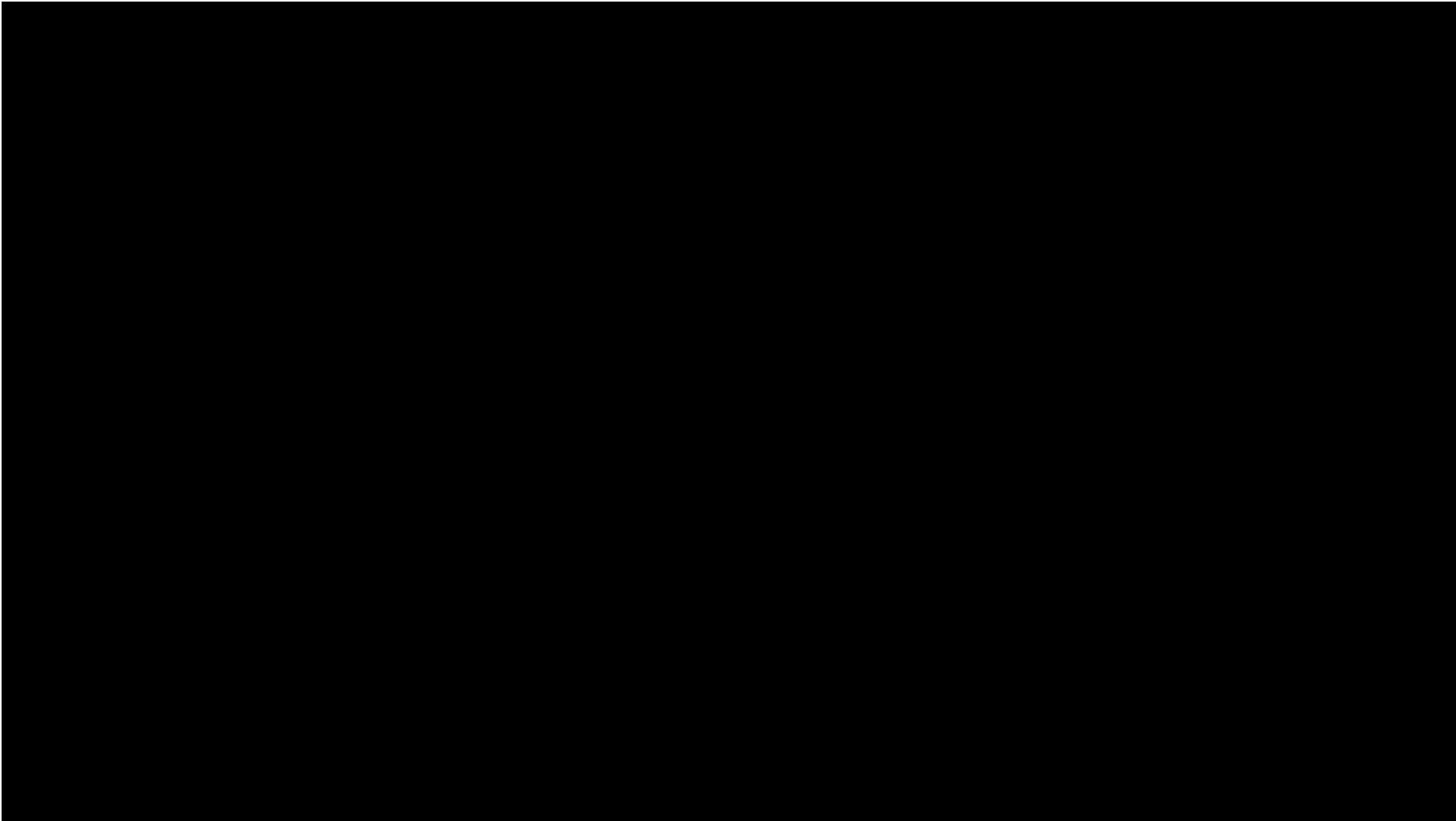


## Value: representa el aspecto más relevante del Big Data

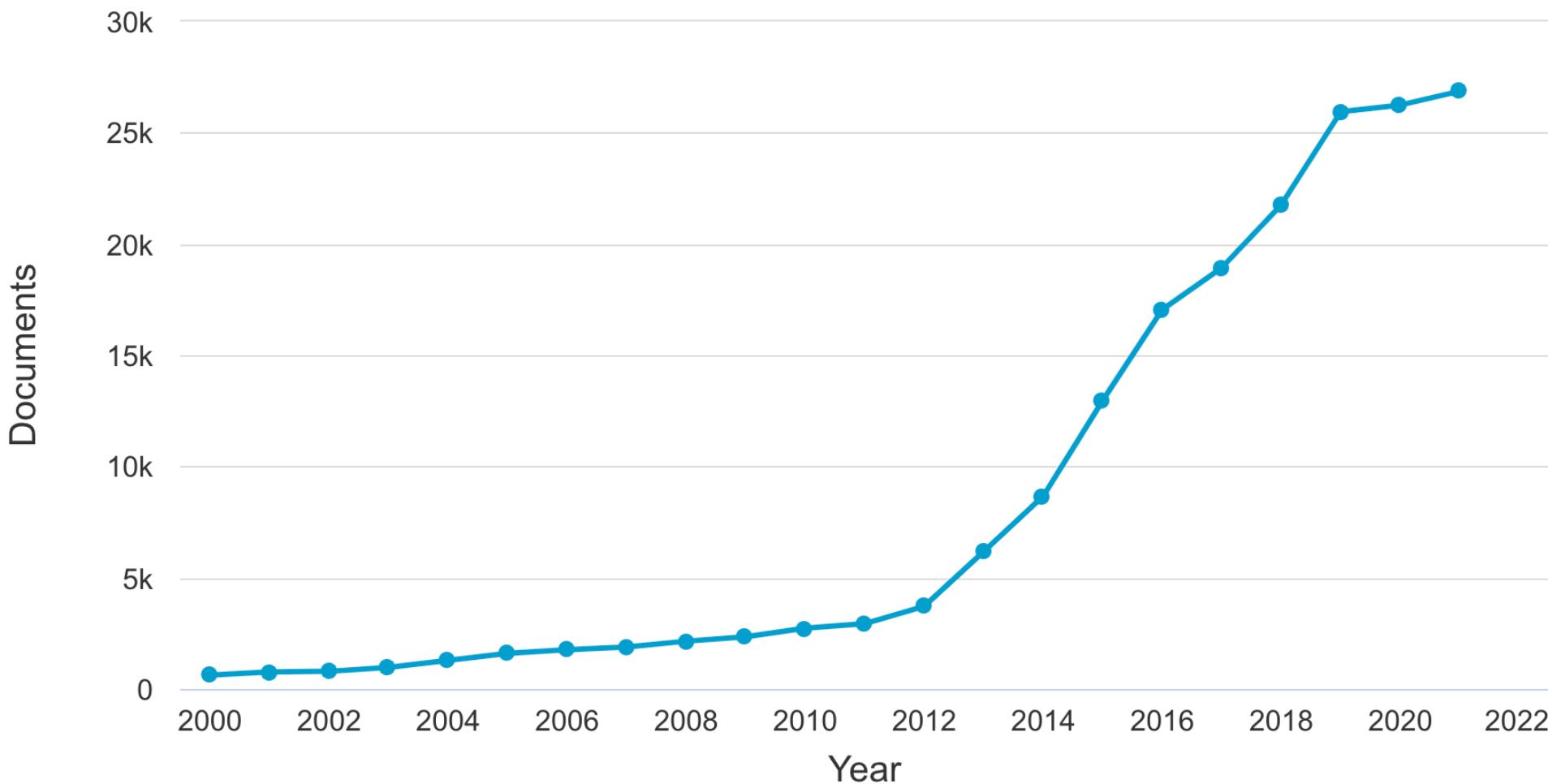
Actualmente el valor marginal de los datos se representa mediante la siguiente gráfica. En dicha gráfica se observa que a medida que aumenta el volumen y complejidad de los datos, su valor marginal disminuye considerablemente, debido a su dificultad de explotación.



# Estructura del Big Data



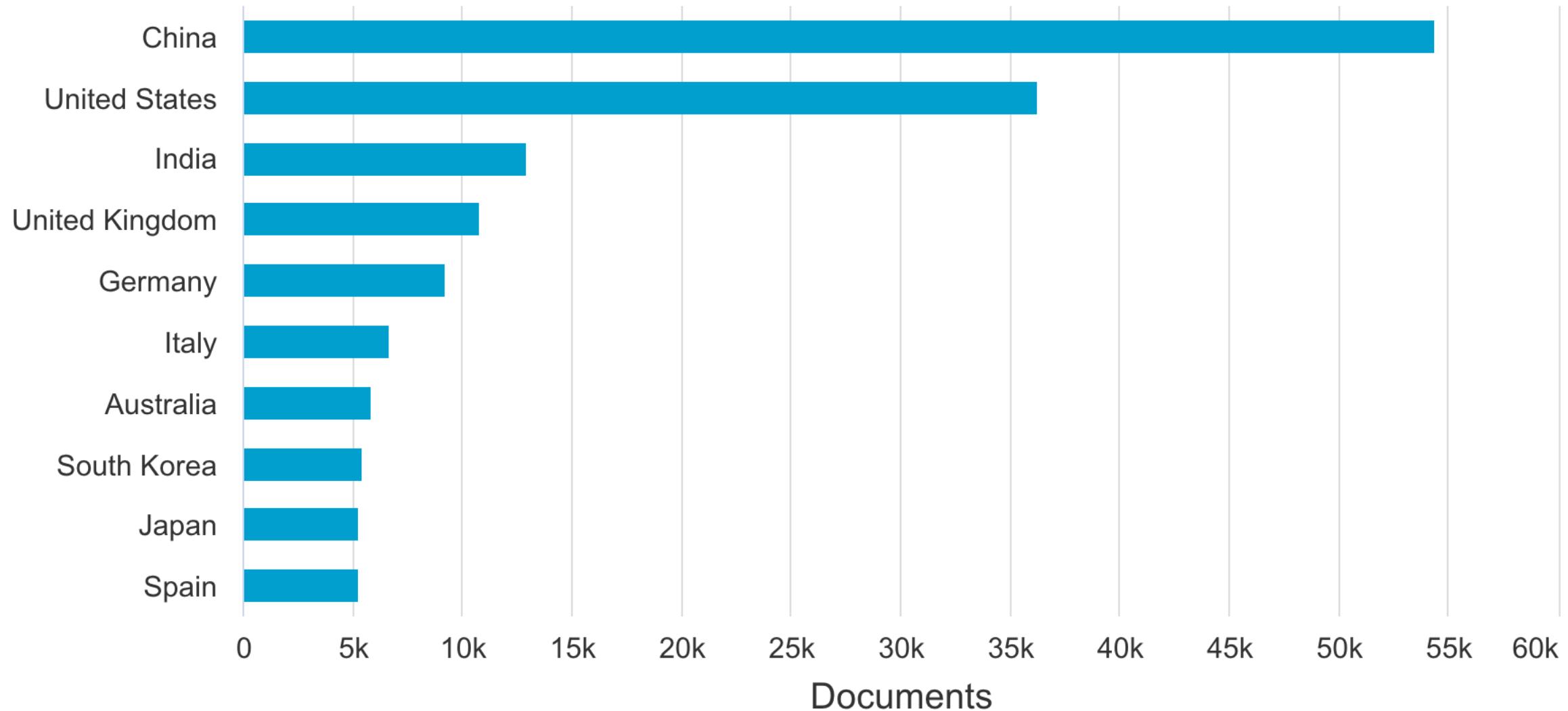
## Documents by year



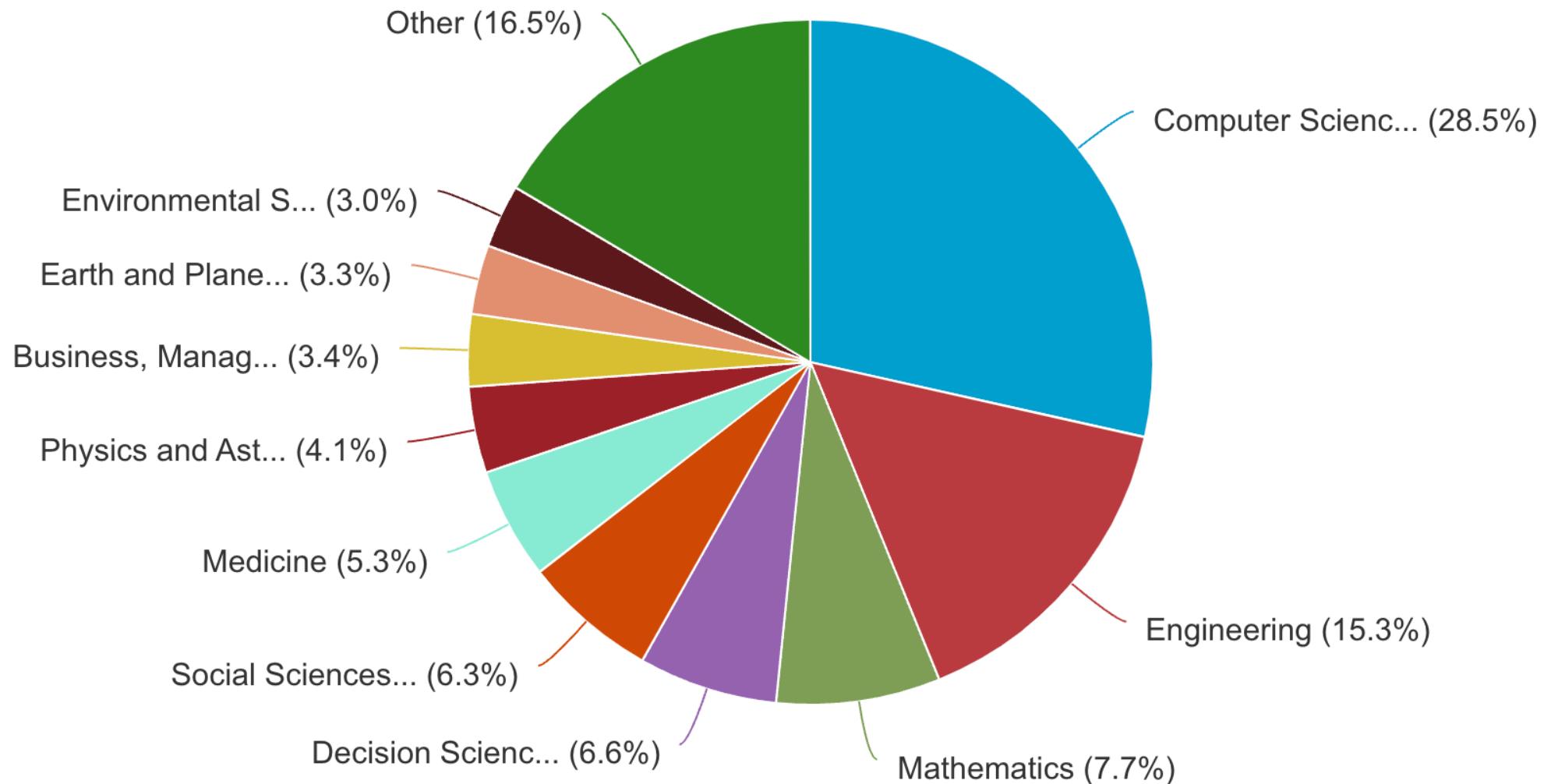
Fuente: Scopus KEY ( **big data** )

# Documents by country or territory

Compare the document counts for up to 15 countries/territories.



## Documents by subject area



- Datos de flujos de Clicks
- Feeds de Twitter
- Entradas de Facebook
- Contenido Web

Web y Medios Sociales



- Lectura de medidores inteligentes
- Lecturas RFID
- Lectura sensores de plataformas petroleras
- Señales de GPS

Maquina a Maquina



- Demandas de salud
- Llamadas de Telecomunicaciones
- Registro de detalles
- Registros de Facturación

Datos de transacciones grandes



- Reconocimiento Facial
- Genética

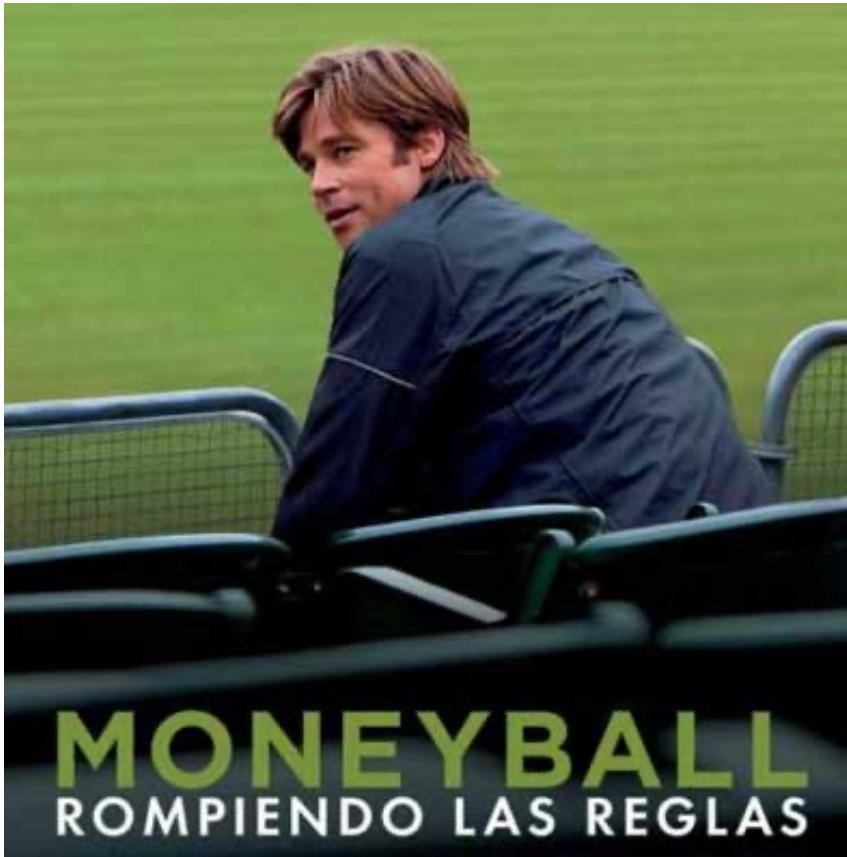
Biometría



- Registros de voz de centros de llamadas
- Correo electrónico
- Registros médicos electrónicos

Generado por los humanos





Ocurrió en la pretemporada de 2002 en Oakland Athletics de las Grandes Ligas de Béisbol de los Estados Unidos. El gerente deportivo Billy Beane, revolucionó la historia del club y posiblemente del deporte en general tras fichar a un joven economista, Peter Brand, que traía nuevas ideas. **Juntos contrataron jugadores infravalorados, pero económicamente rentables, con un criterio de selección muy diferente.** La intuición y sapiencia de los ojeadores es sustituída por las conclusiones de los análisis de estadísticas y números acumulados a la hora de establecer las necesidades del equipo y los jugadores que mejor se adaptan a éstas



**En la NFL tiene una plataforma que ayuda con sus aplicaciones a los 32 equipos a tomar las mejores decisiones en base a la analítica de datos:** desde el estado de la superficie del césped a las condiciones climatológicas, pasando por datos de la etapa universitaria de cada jugador...todo está registrado y todo puede servir para sacar conclusiones diversas, como la de prevenir lesiones en jugadores. Además, analiza las preferencias de los aficionados gracias a su aplicación NFL Now, que ofrece la posibilidad de que éstos creen su propio canal con contenido variado de la NFL: vídeos divertidos, cheerleaders preferidas, información por equipos, por jugadores, etc. También utilizan NetApp para almacenar todos estos datos. Con esto consiguen establecer las demandas de los fans y facilita las cosas a la hora de establecer acciones de marketing, expandir el mercado, encontrar los partners más apropiados, etc

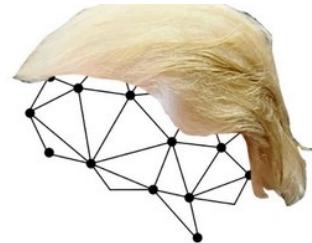




**Tras su primer mandato, el presidente de los EEUU, Barack Obama, decidió utilizar Big Data para su reelección en 2012.** Un centenar de personas trabajaron en el departamento de analítica de la campaña. 50 estaban fijos en las oficinas centrales, otros 30 se movilizaron a lo largo y ancho de las distintas sedes del país, y 20 estaban única y exclusivamente centrados en la interpretación de los datos recibidos. Tras un primer análisis, los esfuerzos de la campaña se enfocaron en tres aspectos: registro (**recoger datos de los votantes convencidos**), **persuasión (dirigirse a los dudosos de una forma eficaz)** y voto del electorado (asegurarse de que los partidarios fueran a ejercer el voto sí o sí). Y, por primera vez, los tres equipos más importantes de las campañas electorales: el de campo, el digital y el de comunicación, trabajaron con una estrategia unificada con los respectivos datos de cada uno. El motor de todo, la plataforma inteligente utilizada fue HP Vertica. Entre las acciones más efectivas que permitía esta plataforma estaban: recoger datos a pie de campo y realizar un feedback muy rápido vía notificaciones email por parte del equipo online (se mejoraba en tiempo y eficiencia); o detectar los nichos en los que funcionaría mejor la publicidad en TV cruzando datos de los votantes con otros demográficos, audiencias, precios de publicidad, programas...

BBC  
MUNDO





## Cambridge Analytica

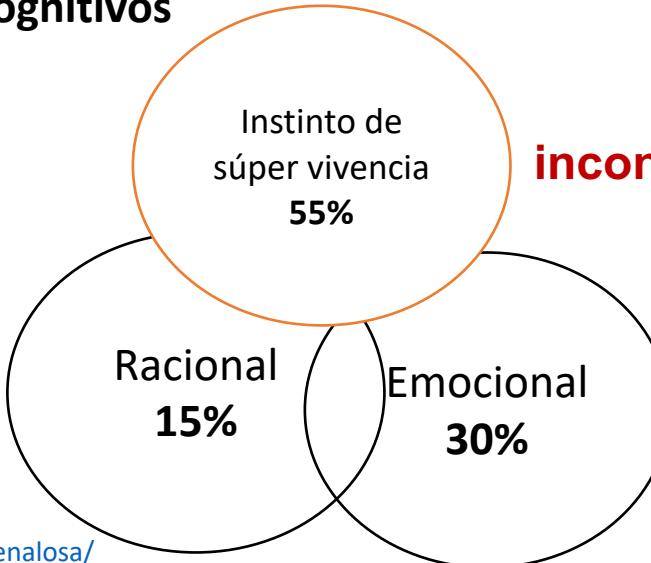


5.000 data points - 240 million Americans



## Campañas de persuasión Psicológica Sesgos cognitivos

**subconsciente**



**inconsciente**





BBVA analizó el uso de las tarjetas de crédito en España durante la Semana Santa de 2011 en cuatro sectores: mercados y alimentos, bares y restaurantes, moda y gasolineras

# BIG DATA: LA REVOLUCIÓN DE LA GESTIÓN

## 1. Liderazgo

- Crear equipos comprometidos para dar respuesta de objetivos

## 2. Gestión del talento

- Nuevos Roles (Científicos de datos )

## 3. Tecnología

- Hadoop
- NoSQL
- Cloud computing

## 4. Toma de Decisiones

- Definir objetivo a alcanzar
- Construcción de modelos predictivos

## 5. Cultura corporativa

- Desarrollar analíticas que demuestren con sencillez la evaluación del negocio
- Crear herramientas sencillas usables para cualquier funcionario
- Desarrollar capacidades necesarias para sacar mejor provecho.

# ¿Qué son los Datos Abiertos (Open Data)?

**“Los datos Abiertos (open data) son datos que pueden ser libremente utilizados, reutilizados y redistribuidos por cualquier persona” (ver Open Knowledge Foundation).**

Los datos abiertos presuponen su publicación y difusión de información en la Internet, sin limitaciones de acceso ni de uso, compartida en formato electrónico y abierto. El formato abierto permite la combinación de conjuntos de datos de diferentes orígenes, su reutilización y difusión, libremente y de forma automatizada.

# ¿Cuándo un dato es un dato abierto?

**Un dato es considerado abierto cuando existen:**

- **Disponibilidad y acceso:** el dato tiene que estar disponible en la Internet (online), integralmente, sin limitaciones de acceso.
- **Reutilización y redistribución:** el dato tiene que ser ofrecido en condiciones y en un formato conveniente, que permitan su reutilización, combinación con conjuntos de datos de diferentes orígenes, su difusión y redistribución. Esto significa que los datos deben ser preferiblemente procesable por maquinas, en formato no-proprietario, y no cubierto por licencias que puedan limitar su uso.
- **Participación universal:** el dato tiene que estar disponible sin limitaciones de uso, todos deben poder usar, reutilizar y redistribuir la información, sin discriminación con las áreas de actuación, personas o grupos.



<http://www.data.gov/opendatasites>



## **FORMATOS DE DATOS ABIERTOS**

[http://www.navarra.es/home\\_es/Open-Data/](http://www.navarra.es/home_es/Open-Data/)

### **Formatos**

La información se publica en formatos de datos estructurados para facilitar que pueda ser utilizada de forma automática por los lenguajes de programación. De esta manera, se intenta cumplir el objetivo de reutilizar al máximo la información publicada.

### **Formatos estructurados**

Estos son los formatos más utilizados para publicar los datos:

#### **XML (eXtensible Markup Language)**

Es un metalenguaje extensible de etiquetas desarrollado por el W3C que permite definir lenguajes para diferentes necesidades. Es el estándar para el intercambio de información estructurada entre diferentes plataformas.

Más información: [www.w3.org/standards/xml/core](http://www.w3.org/standards/xml/core)

#### **CSV (Comma-separated values)**

Valores separados por coma. Los ficheros CSV son un tipo de documento en formato abierto sencillo para representar datos en formato de tabla. Las columnas se separan por comas (o punto y coma) y las filas por saltos de línea.

Más información: [tools.ietf.org/html/rfc4180](http://tools.ietf.org/html/rfc4180)

## FORMATOS DE DATOS ABIERTOS

### Formatos

#### RSS (Really Simple Syndication)

Es un formato XML para la distribución de contenidos de páginas web. Facilita la publicación de información actualizada a los usuarios suscritos a la fuente RSS sin necesidad de usar un navegador, utilizando un software especializado en este formato.

Más información: <http://es.wikipedia.org/wiki/RSS>

#### SHP (Shapefile)

Shapefile es un formato propietario estándar de datos espaciales, desarrollado por la compañía ESRI, que almacena tanto la geometría como la información alfanumérica. Este formato no está preparado para almacenar información topológica.

Más información: <http://es.wikipedia.org/wiki/Shapefile>

#### XLS (Microsoft Office Excel)

Microsoft Office Excel es un formato propietario de Microsoft que muestra la información en celdas organizadas en filas y columnas, y cada celda contiene datos o fórmulas, con referencias relativas o absolutas a otras celdas.

Más información: [http://es.wikipedia.org/wiki/Microsoft\\_Excel](http://es.wikipedia.org/wiki/Microsoft_Excel)

## **Formatos**

### **JSON (JavaScript Object Notation)**

Es un formato ligero para el intercambio de datos basado en la notación literal de objetos de JavaScript. Su sintaxis es simple, por lo que facilita el tratamiento en los navegadores. Además, su concisión reduce el tamaño de flujo de datos entre cliente y servidor.

Más información: [json.org/json-es.html](http://json.org/json-es.html)

### **RDF (Resource Description Framework)**

Es una especificación del W3C para el modelado de información y la descripción de recursos, que se hace con la forma de sujeto-predicado-objeto. La combinación de RDF con otras herramientas permite añadir significado a las páginas y es una de las tecnologías esenciales para la web semántica.

Más información: [www.w3.org/standards/techs/rdf#w3c\\_all](http://www.w3.org/standards/techs/rdf#w3c_all)

## **Formatos**

### **ODS (Operational Data Store)**

Es un contenedor de datos activos, es decir operacionales que ayudan al soporte de decisiones y a la operación. Es un formato de archivo abierto y estándar para el almacenamiento de hojas de cálculo que muestra información en celdas organizadas en filas y columnas, y cada celda contiene datos o fórmulas, con referencias relativas o absolutas a otras celdas.

Más información: <http://es.wikipedia.org/wiki/ODS>

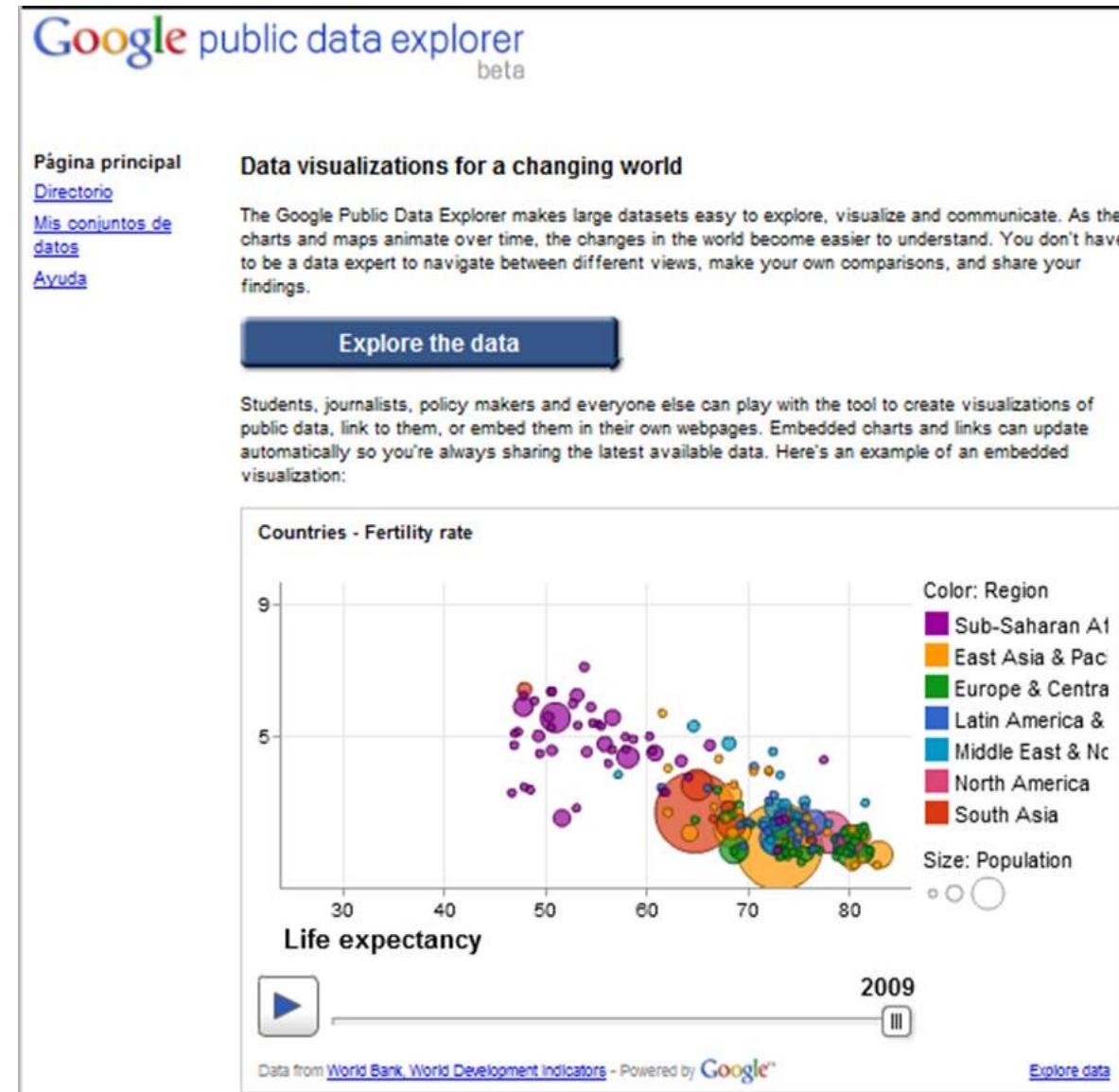
### **KML (Keyhole Markup Language)**

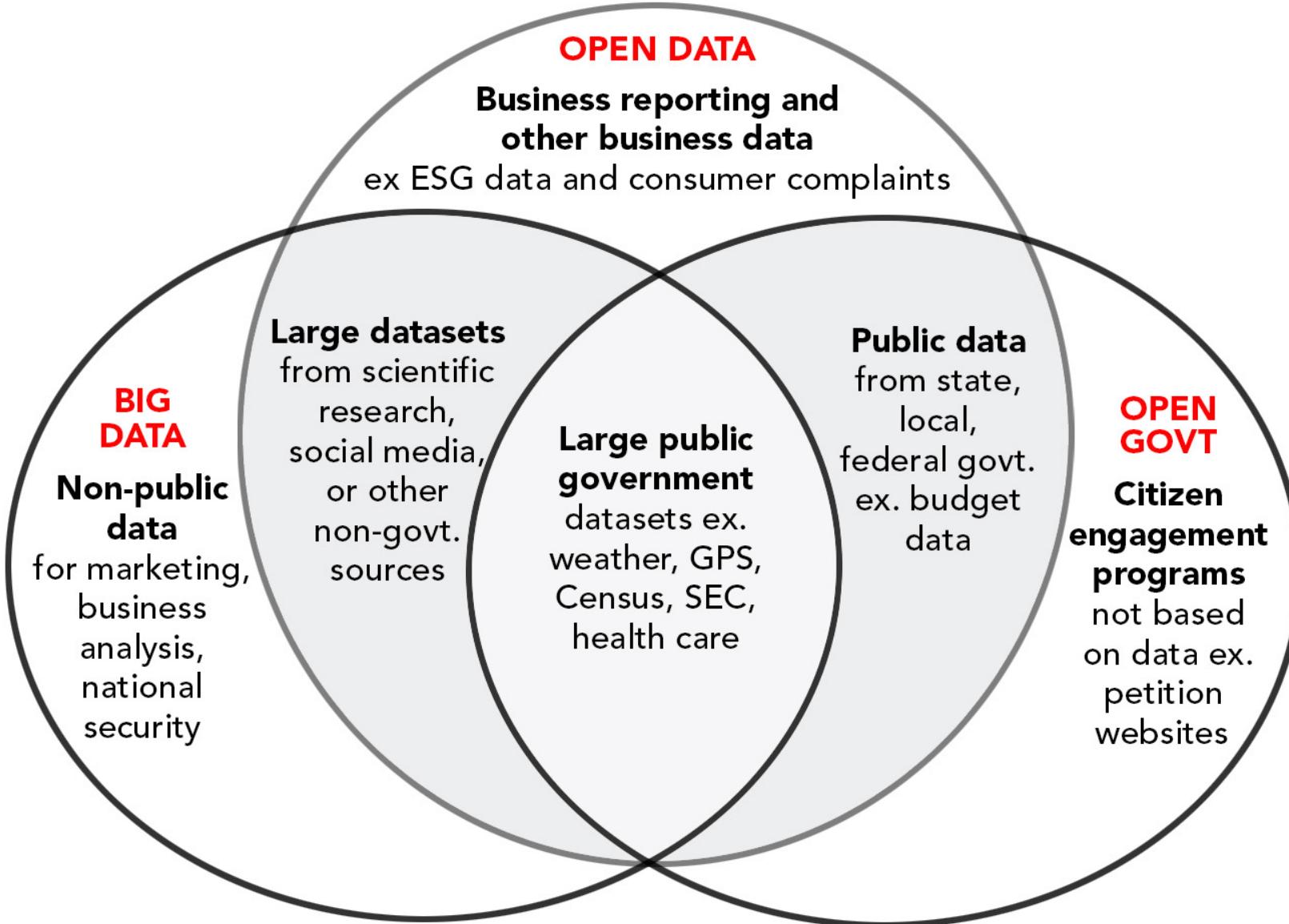
Es una gramática XML y un formato de archivo para la creación de modelos y el almacenamiento de funciones geográficas como puntos, líneas, imágenes, polígonos y modelos que se mostrarán principalmente en aplicaciones de mapas. KML es utilizado para compartir lugares e información entre aplicaciones.

Más información: <http://es.wikipedia.org/wiki/KML>

# Google public data explorer

<http://www.google.com/publicdata/home>





# NoSQL



Gaming



Social



IoT



Web



Mobile



Enterprise



Key/value store



Document database



Column family store

# SQL



Web



Mobile



Enterprise



Data mart



Relational table storage



Relationships use joins



## SQL

Cuando el volumen de mis datos no crece o lo hace poco a poco.

Cuando las necesidades de proceso se pueden asumir en un sólo servidor.

Cuando no tenemos picos de uso del sistema por parte de los usuarios más allá de los previstos.



## NoSQL

Cuando el volumen de mis datos crece muy rápidamente en momentos puntuales.

Cuando las necesidades de proceso no se pueden prever.

Cuando tenemos picos de uso del sistema por parte de los usuarios en múltiples ocasiones.

# NO SQL

- **NoSQL** – "not only SQL" – es una categoría general de sistemas de gestión de bases de datos que difiere de los RDBMS en diferente modos:
  - No tienen schemas, no permiten JOINs, no intentan garantizar ACID y escalan horizontalmente
  - Tanto las bases de datos NoSQL como las relacionales son tipos de **Almacenamiento Estructurado**.
- El término fue acuñado en 1998 por Carlo Strozzi y resucitado en 2009 por Eric Evans
  - Evans sugiere mejor referirse a esta familia de BBDD de nueva generación como "**Big Data**" mientras que Strozzi considera ahora que **NoREL** es un mejor nombre

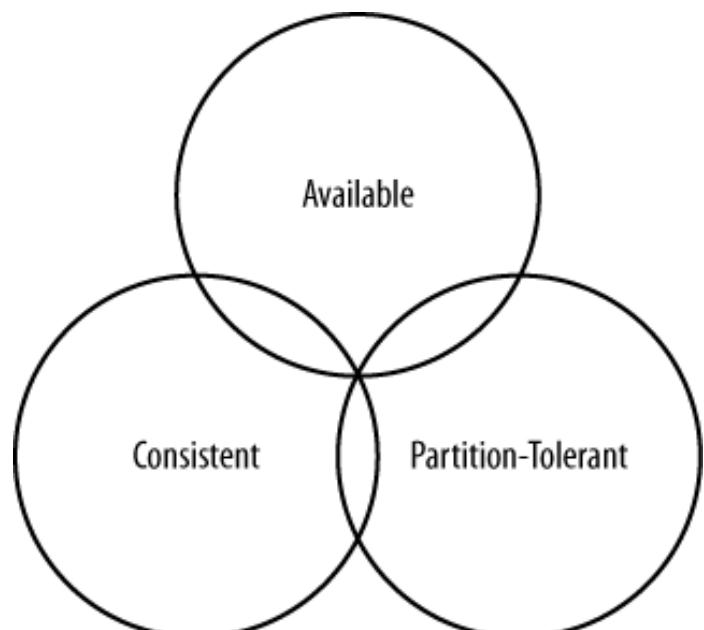
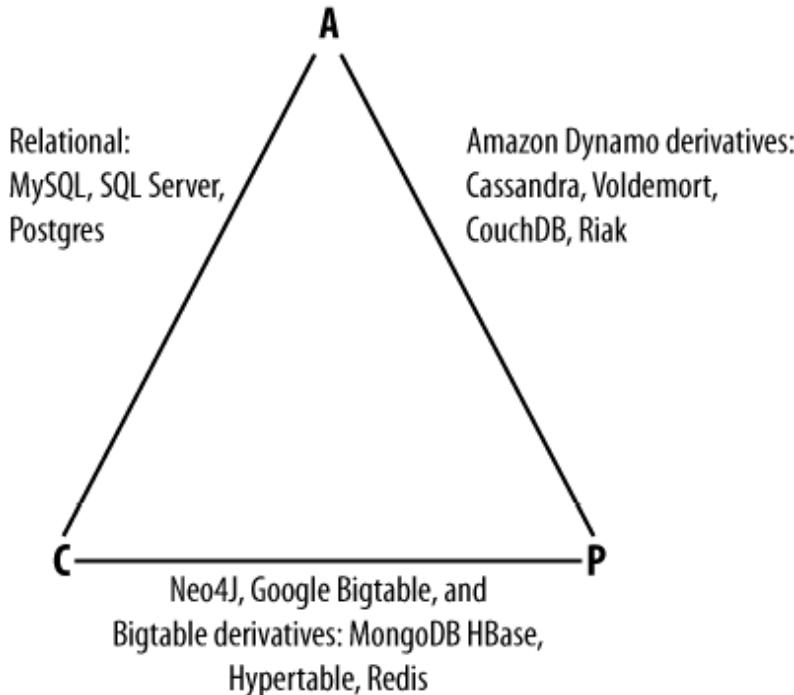
- La **principal diferencia radica en cómo guardan los datos** (por ejemplo, *almacenamiento de un recibo*):
  - En una RDBMS tendríamos que partir la información en diferentes tablas y luego usar un lenguaje de programación en la parte servidora para transformar estos datos en objetos de la vida real.
  - En NoSQL, simplemente guardas el recibo:
    - NoSQL es **libre de schemas**, tú no diseñas tus tablas y su estructura por adelantado
- **¡¡¡NoSQL no es la panacea!!!**
  - Si tus datos son relacionales, quedarte con tu RDBMS sería la opción correcta

# Características principales

NO  
SQL

- Fáciles de usar en clústers de balanceo de carga convencionales → facilitan **escalabilidad horizontal**
- Guardan **datos persistentes** (no sólo cachés)
- **No tienen esquemas fijos y permite la migración del esquema** sin tener que ser reiniciadas o paradas
- Suelen tener un **sistema de consultas propio** en vez de usar un lenguaje de consultas estándar
- Tienen propiedades ACID en un nodo del clúster y son “**eventualmente consistentes**” en el clúster

- **Teorema de Brewer:** “es imposible para un sistema computacional distribuido ofrecer simultáneamente las siguientes tres garantías”:
  - **Consistencia** – todos los nodos ven los mismos datos al mismo tiempo
  - **Disponibilidad (Availability)** – garantiza que cada petición recibe una respuesta acerca de si tuvo éxito o no
  - **Tolerancia a la partición (Partition)** – el sistema continua funcionando a pesar de la pérdida de mensajes
- Equivalente a:
  - “**You can have it good, you can have it fast, you can have it cheap: pick two.**”



# RDBMS vs. NoSQL

- Los RDBMS tradicionales nos permiten definir la estructura de un esquema que demanda reglas rígidas y garantizan ACID
- Las aplicaciones web y sistemas de información modernos presentan desafíos muy distintos a los sistemas empresariales tradicionales (e.j. sistemas bancarios):
  - Datos a escala web
  - Alta frecuencia de lecturas y escrituras
  - Cambios de esquema de datos frecuentes
  - Las aplicaciones sociales (no bancarias) no necesitan el mismo nivel de ACID
- Consecuencia → aparición de soluciones NoSQL
  - Cassandra, MongoDB, Jackrabbit , CouchDB, BigTable, Dynamo o Neo4j

# RDBMS vs. NoSQL

- Los RDBMS tradicionales nos permiten definir la estructura de un esquema que demanda reglas rígidas y garantizan ACID
- Las aplicaciones web y sistemas de información modernos presentan desafíos muy distintos a los sistemas empresariales tradicionales (e.j. sistemas bancarios):
  - Datos a escala web
  - Alta frecuencia de lecturas y escrituras
  - Cambios de esquema de datos frecuentes
  - Las aplicaciones sociales (no bancarias) no necesitan el mismo nivel de ACID
- Consecuencia → aparición de soluciones NoSQL
  - Cassandra, MongoDB, Jackrabbit , CouchDB, BigTable, Dynamo o Neo4j