

Módulo BIG DATA

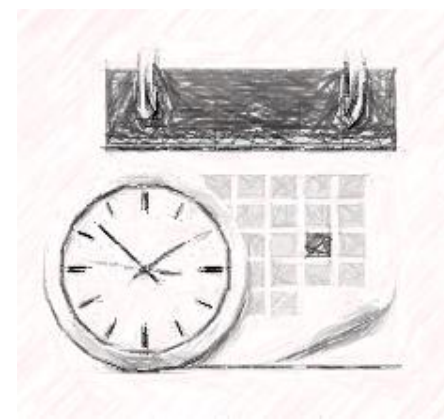
Procesamiento y Almacenamiento de Grandes Volúmenes de Datos

Introducción

Por

Ing. Elizabeth León Guzmán, Ph.D.

Agenda



- Introducción
- Definición
- Aplicaciones



<https://www.eweek.com/blogs/storage-station/toshiba-demonstrates-scale-out-big-data-storage-system>

Datos

Un dato es una representación simbólica de un atributo o variable cuantitativa o cualitativa.

Temperatura valores: 17, 28, 15

Ciudad valores: Bogotá, Cartagena

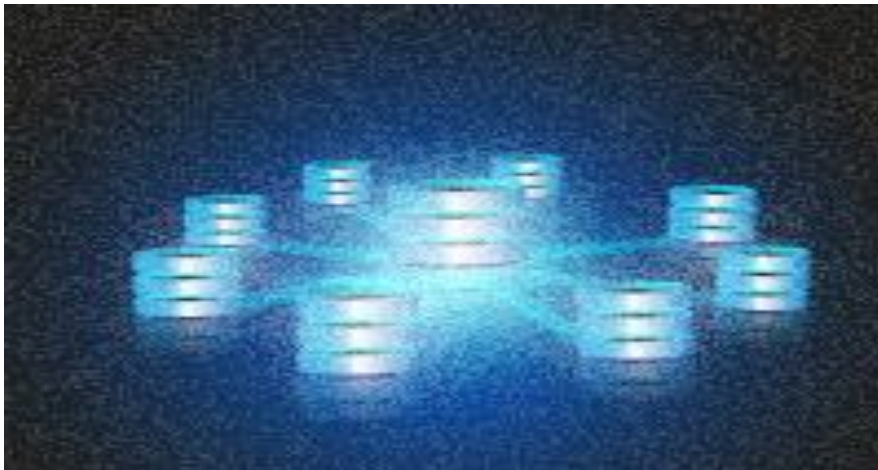
Fecha valores: Julio 20 2015, Julio 20 2016



variables

Datos

Un dato es una representación simbólica de un atributo o variable cuantitativa o cualitativa.



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

Datos almacenados
en “bases de datos”



Datos estructurados y
relacionados
“Información”

Datos

Toma de
decisiones

Conocimiento

Información

Datos

Datos

Toma de
decisiones

Conocimiento

Análisis exploratorio
Minería de Datos
OLAP

Información

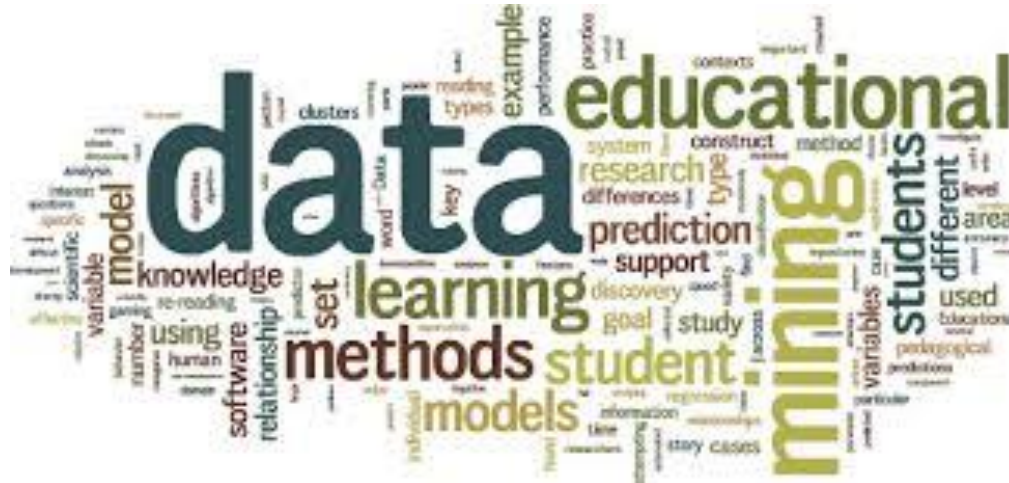
Bases de datos
Bodegas de datos

Datos

Datos



Son el centro de la futura sociedad de la economía del conocimiento



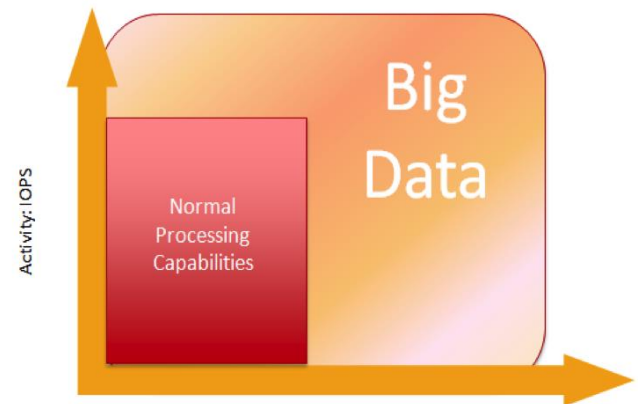
¿Qué es Big Data?

No hay una definición estándar



“*Big Data*” es una colección de datos complejos muy grande, 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 ...



File/Object Size, Content Volume

Fuente: Enterprise strategy group 2011 8

¿Qué es Big Data?

“Big Data” son todos los datos



Volumen

Peta bytes
Zetta bytes
Tablas, archivos



Variedad

- Datos Estructurados
- Datos No Estructurados
- Imágenes, texto, multimedia



Velocidad

- Flujos (streams)
- Real time
- Batch



Veracidad

- Autenticidad
- Integridad
- Integración
- Confiabilidad
- Ruido, inconsistencias, vagos, errores, etc.



Valor

Conocimiento

Análisis

5 Vs

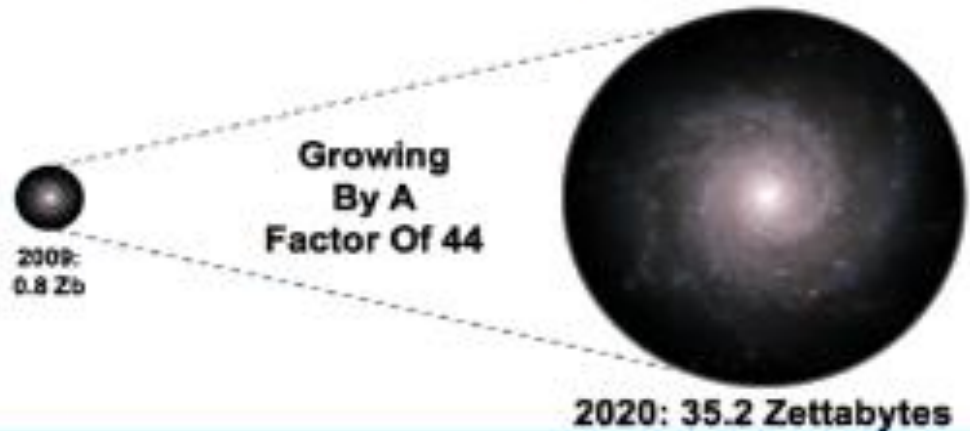
¿Qué es Big Data?

1 V : Volumen

Crecimiento
exponencial

En los últimos dos años
se ha generado el 90%
de los datos

The Digital Universe 2009-2020



Almacenamiento por compañías promedio (ahora)



Terabytes

Petabytes

Exabytes

Zettabytes

1TB = 10^{12} bytes (un billón)

1PB = 10^{15} bytes $\sim 1000^5$

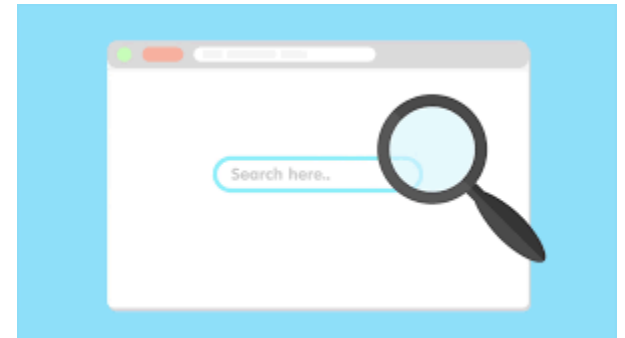
1EB = 10^{18} bytes $\sim 1024^6$

1ZB = 10^{21} bytes $\sim 1000^7$

¿Qué es Big Data?

1 V : Volumen

- Búsquedas web: más de la mitad de las búsquedas se realizan por teléfono
- 40.000 búsquedas cada Segundo (2016)
- 5 billones de búsquedas por día (todos los buscadores)



Fuente: <https://iobint.com/effective-web-search-engine-marketing-tips/>

¿Qué es Big Data?

1 V : Volumen

Según [Data Never Sleeps 5.0 report](#) cada minuto en 2017:

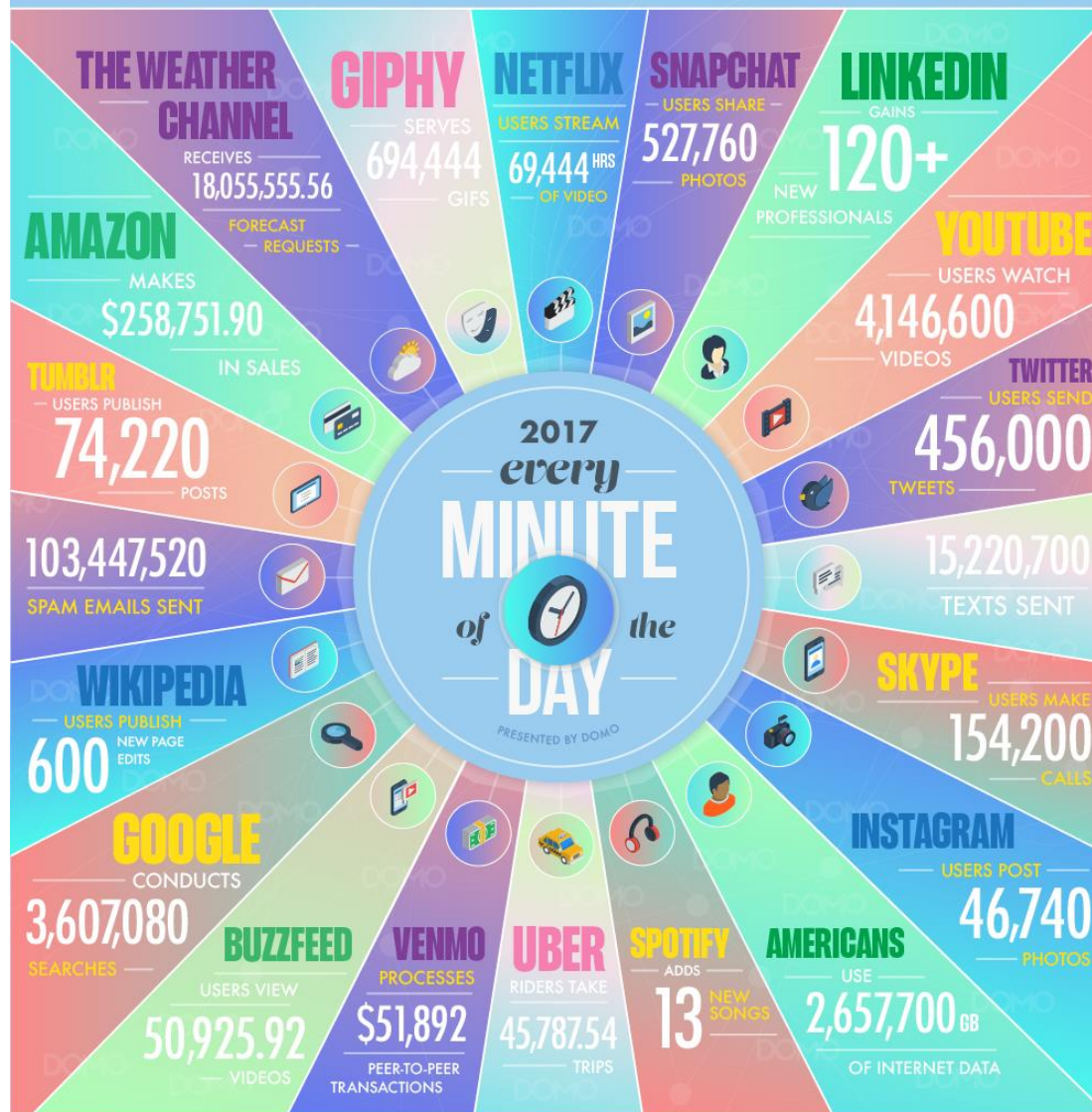
- Usuarios de Snapchat comparten 527,760 fotos
- Mas de 120 profesionales se unen a LinkedIn
- Se ven 4,146,600 videos en YouTube
- Se envían 456,000 tweets en Twitter
- Usuarios de Instagram comparten 46,740 fotos



DATA NEVER SLEEPS 5.0

How much data is generated *every minute*?

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.

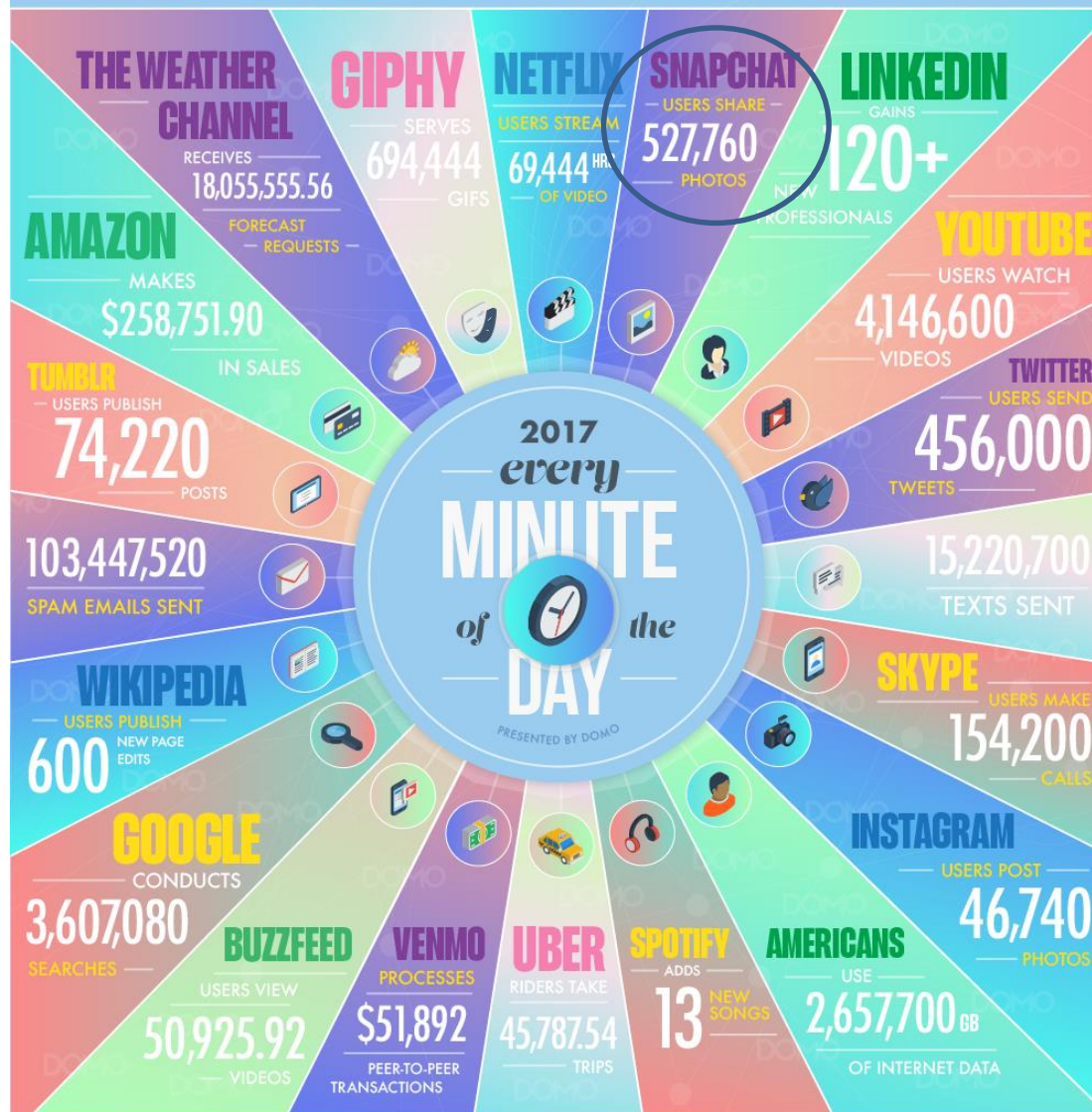




DATA NEVER SLEEPS 5.0

How much data is generated *every minute*?

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.

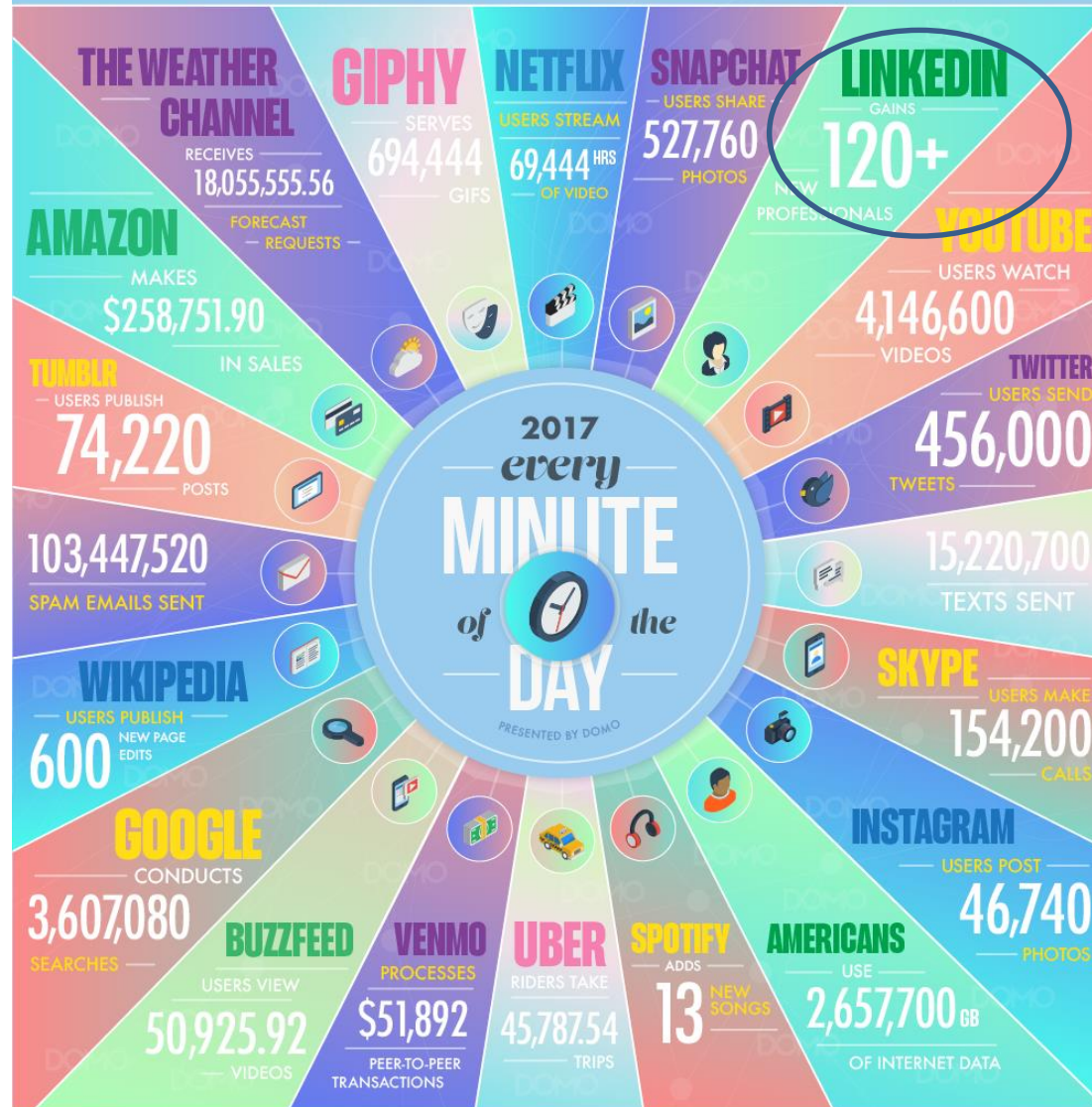




DATA NEVER SLEEPS 5.0

How much data is generated *every minute*?

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.

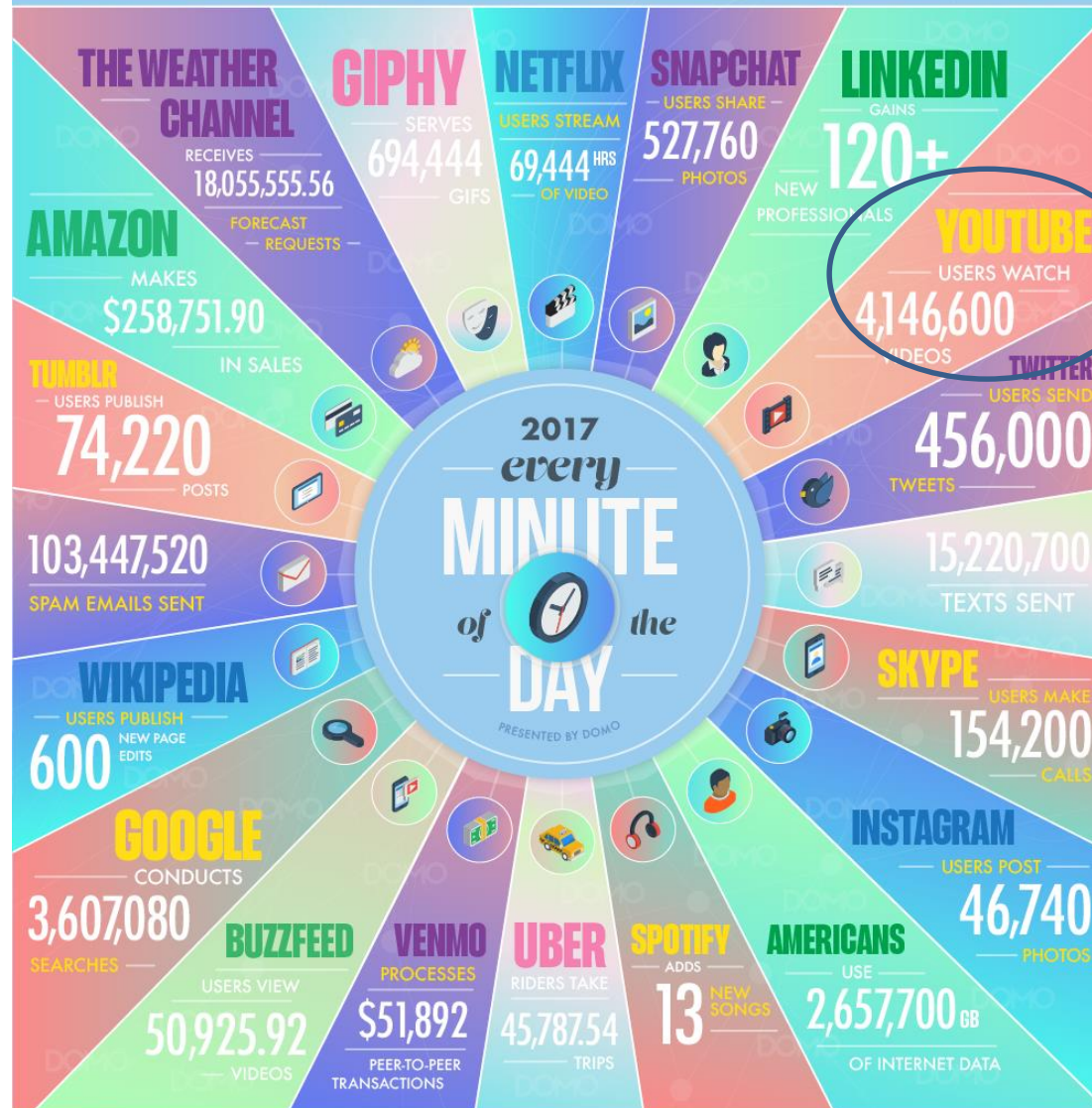




DATA NEVER SLEEPS 5.0

How much data is generated *every minute*?

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.

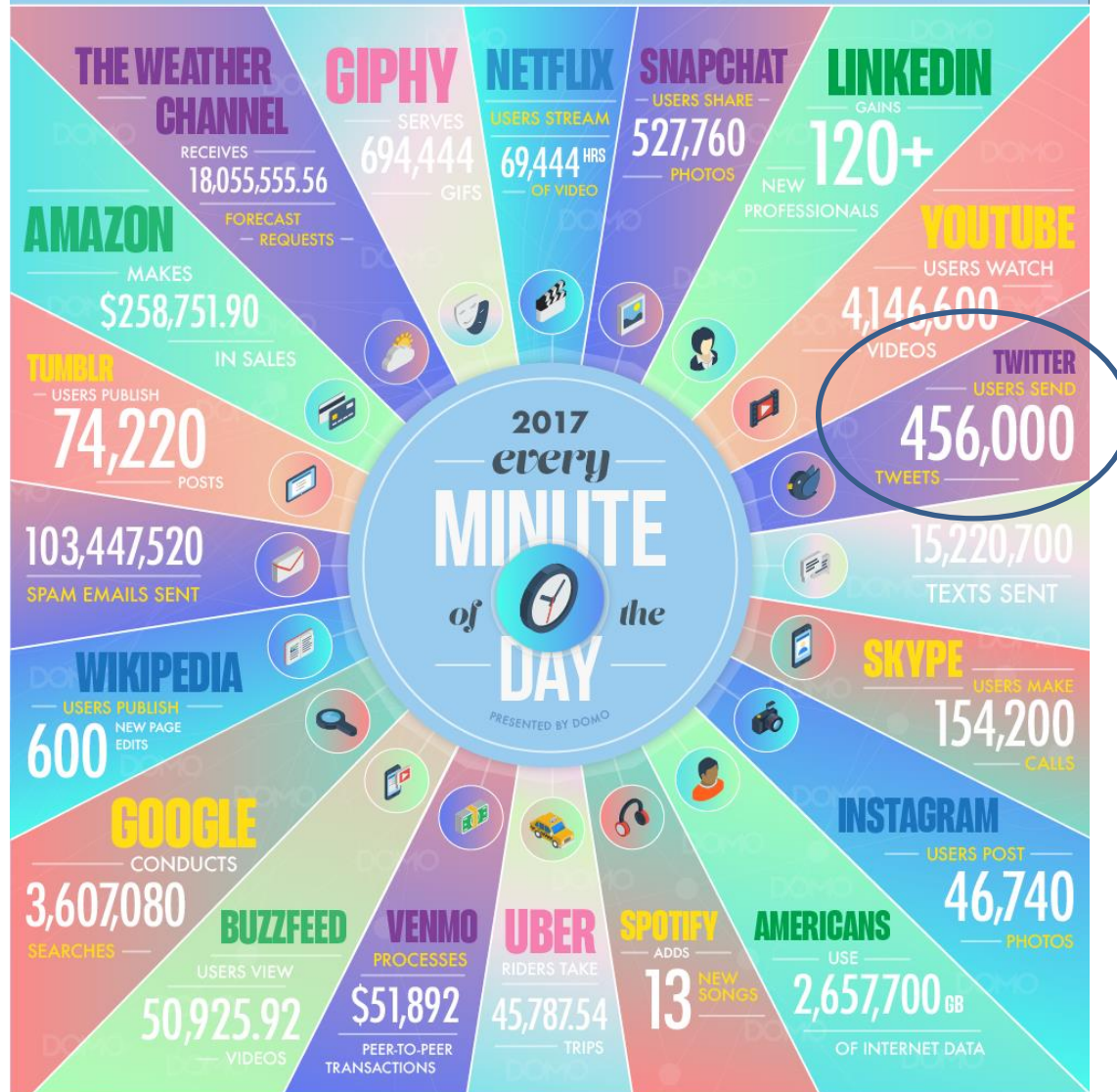




DATA NEVER SLEEPS 5.0

How much data is generated *every minute*?

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.

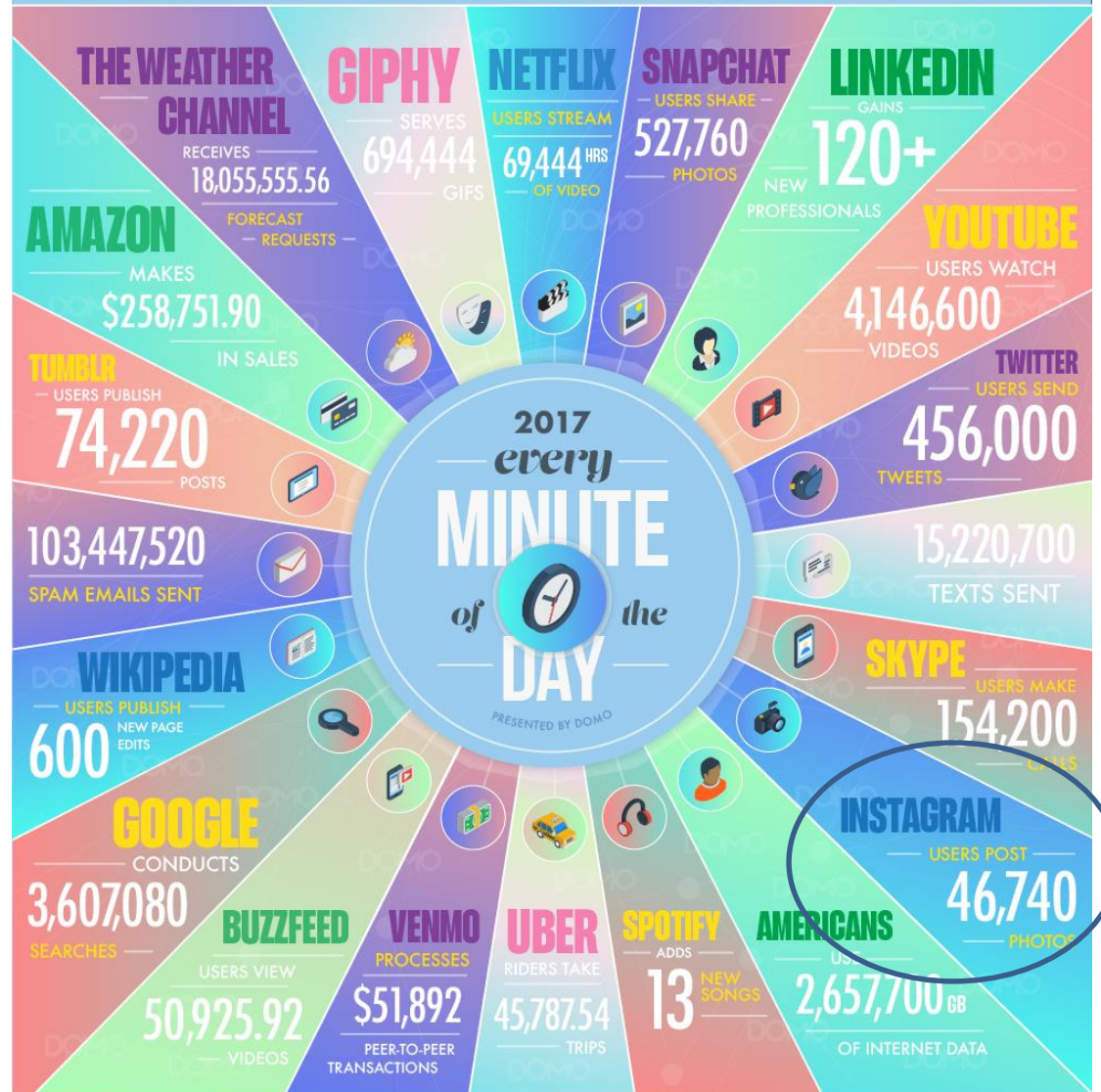




DATA NEVER SLEEPS 5.0

How much data is generated *every minute*?

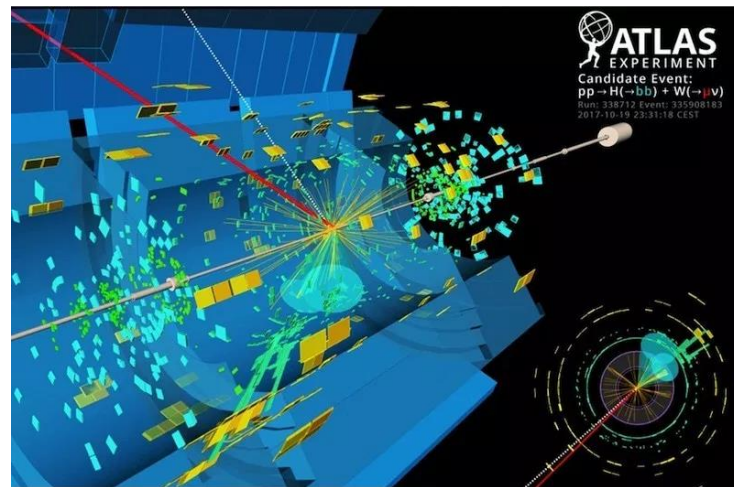
90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.



¿Qué es Big Data?

1 V : Volumen

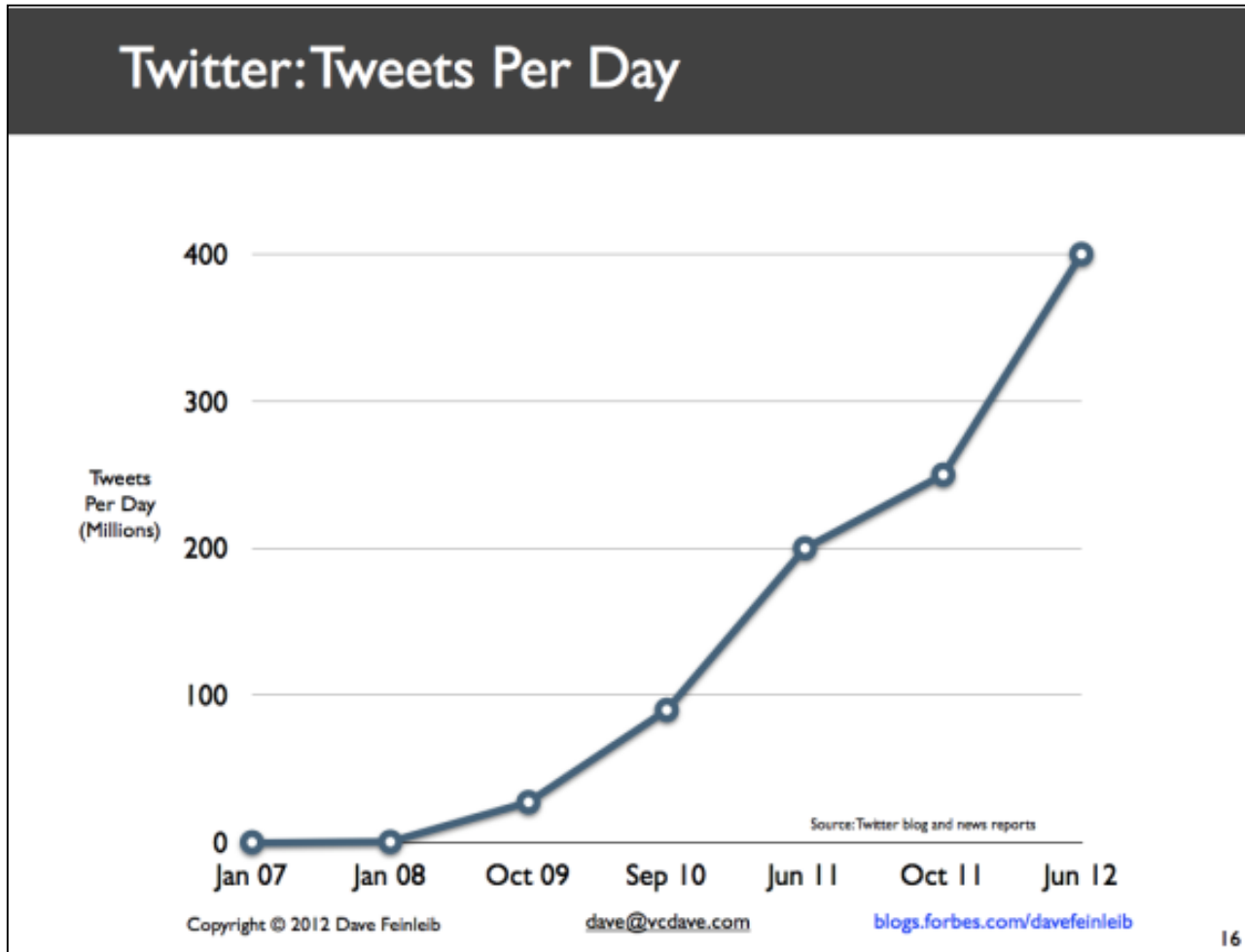
- CERN's Large Hydron Collider (LHC) alcanzó **200 petabytes** de datos almacenados en cinta (Junio de 2017)



<https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone>

¿Qué es Big Data?

1 V : Volumen



¿Qué es Big Data?

2 V : Velocidad



- Datos se generan muy rápido y necesitan ser procesados y analizados rápidamente.
- Algunas veces, no es posible almacenar los datos: “Online Data Analytics”

Decisiones tardías → oportunidades perdidas

Ejemplos:

- **E-Promociones:** Posición actual e historial de compra → envío de promociones
- **Monitoreo:** En medicina, monitoreo sensorial de las actividades del cuerpo → medida anormal requiere reacción inmediata

¿Qué es Big Data?

2 V : Velocidad

Decisiones tardías → oportunidades perdidas

Twitter



- Análisis de sentimientos en campañas presidenciales. Toma de decisiones de campaña.

Análisis de Imágenes (Solares)

- Detectar “solar loops” temprano para evitar daños en las telecomunicaciones

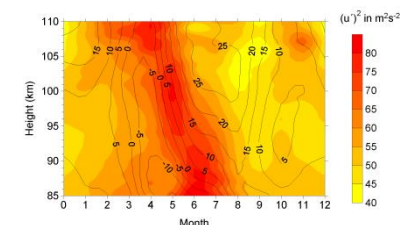
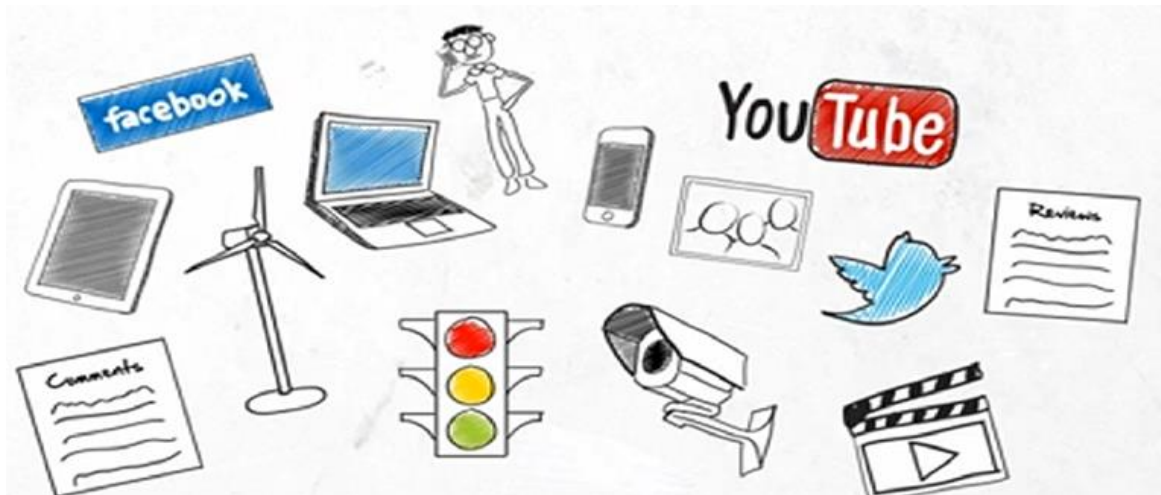
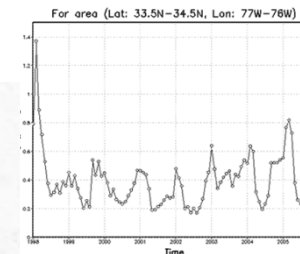
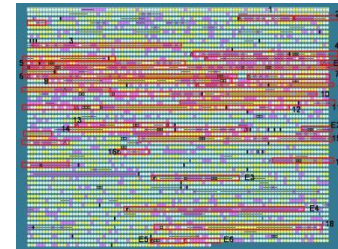


<https://www.theverge.com/2013/2/22/4017464/fiery-coronal-rain-loops-onto-suns-surface>

¿Qué es Big Data?

3 V : Variedad

- Múltiples formatos y estructuras:
Texto, numéricos, imágenes, audio, video, secuencias, grafos, series temporales, etc.



¿Qué es Big Data?

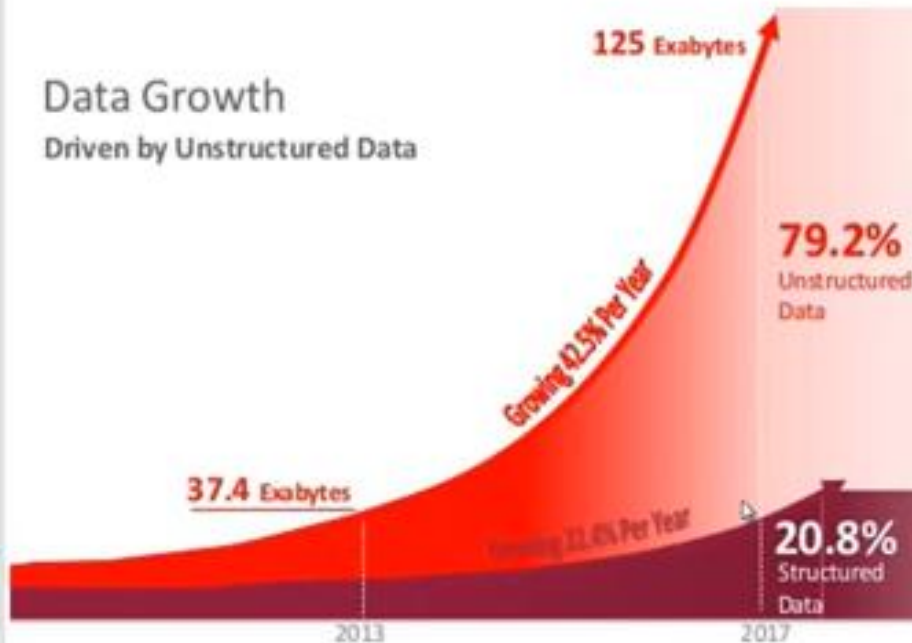
3 V : Variedad

Structured and Unstructured Data Growth

IDC Study: Structured Versus Unstructured Data: The Balance of Power Continues to Shift

Data Growth

Driven by Unstructured Data

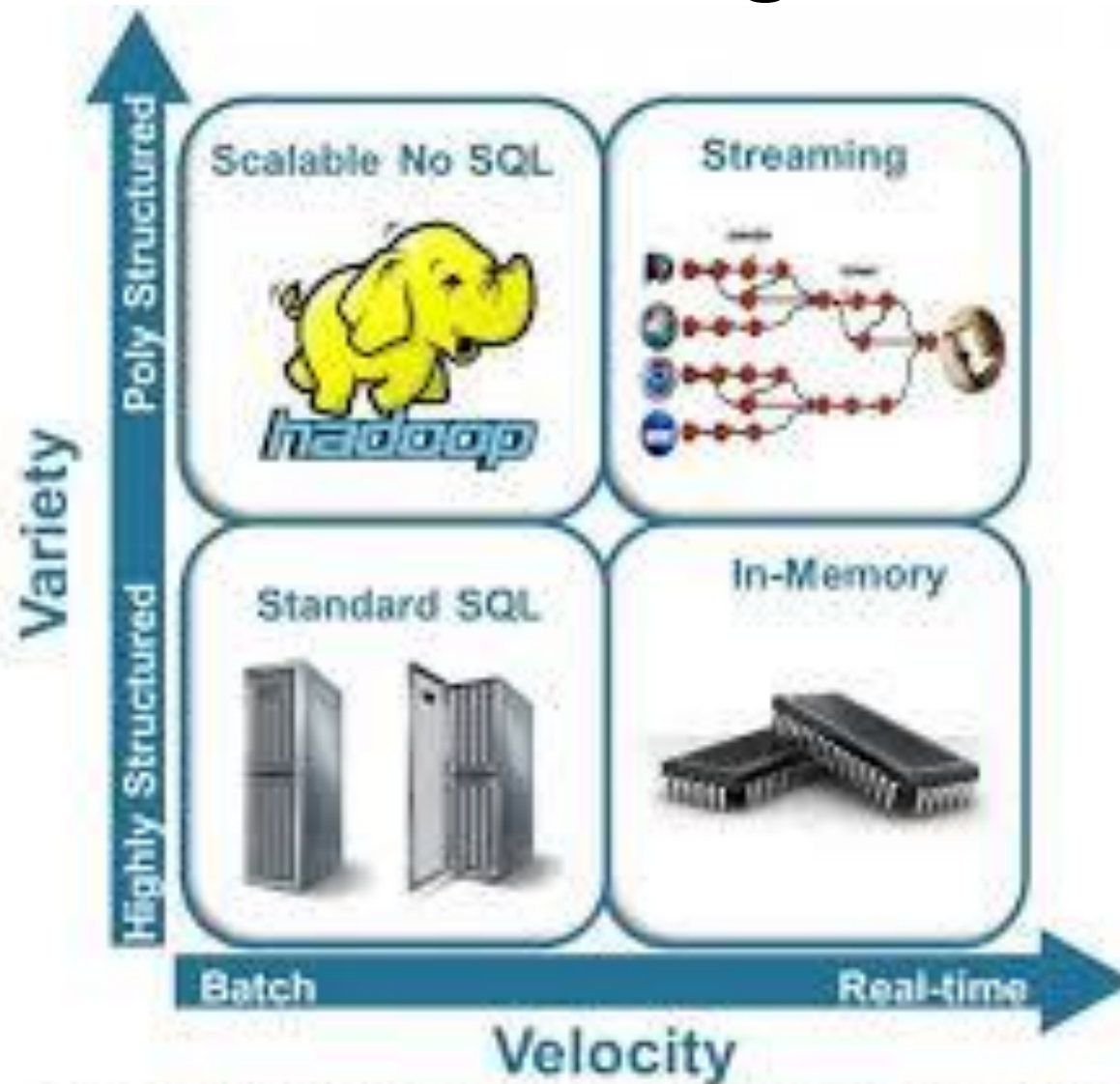


"80% of business-relevant information originates in unstructured form, primarily text."



Structured Data vs. Unstructured Data

¿Qué es Big Data?



Source: Forrester Webinar: Big Data: Gold Rush or Illusion?, Sept. 19, 2013

¿Qué es Big Data?

4 V : Veracidad

- Autenticidad
- Integración: Bodegas de datos (agiles)
- Ruido, inconsistencias, vagos, errores, etc.



¿Qué es Big Data?

5 V : Valor

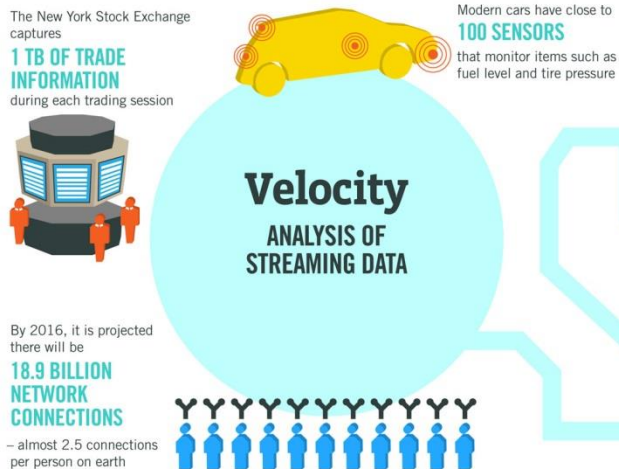
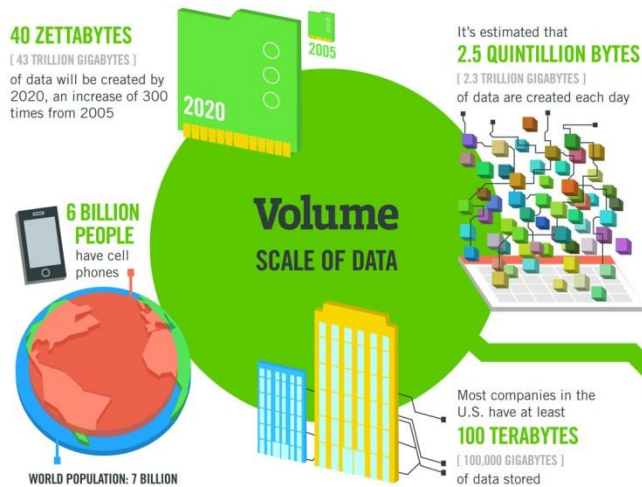


← **Conocimiento**



“Machine learning”





The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data,
with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT
are shared on Facebook every month



Variety
DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users



1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Veracity
UNCERTAINTY OF DATA

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTec, QAS

IBM

Generación de “Big data”

Grandes volúmenes de información en cortos periodos de tiempo



Redes sociales y multimedia



Smart Phones, tables



Consolas de juego, GPS



Proveedores de Internet



Instrumentos científicos



Redes de sensores IoT



Web - películas

Generación de “Big data”

Tiquetes aereos Recibo Restaurante



1	Cab Sauv Corvina	E15.25
1	Cab Sauv Glass	E4.00
20 Met Sales Total		E73.45

1	SUCK MY D. F. FACE	E0.00
1	FISH CAKES	E4.95
2	1/2 Wings Starter	E7.90
1	MELON AND PARM HAM	E3.95
1	Calamari	E4.95
1	Garlic brd starter	E2.25
3	Meatball Starter	E17.85
1	Aub & Feta Starter	E3.95
1	Chef's Special	E12.80

Facturas



Facturas de hotel

Sheraton Gateway Hotel Atlanta Airport	
1400 Shilpala Rd	
Atlanta, GA 30327	
Tel: 770-977-1100 Fax: 770-941-0800	

Room	Room	Room	Room
01-DEC-01	01-DEC-01	01-DEC-01	01-DEC-01
02-DEC-01	02-DEC-01	02-DEC-01	02-DEC-01
03-DEC-01	03-DEC-01	03-DEC-01	03-DEC-01
04-DEC-01	04-DEC-01	04-DEC-01	04-DEC-01
05-DEC-01	05-DEC-01	05-DEC-01	05-DEC-01
06-DEC-01	06-DEC-01	06-DEC-01	06-DEC-01
07-DEC-01	07-DEC-01	07-DEC-01	07-DEC-01
08-DEC-01	08-DEC-01	08-DEC-01	08-DEC-01
09-DEC-01	09-DEC-01	09-DEC-01	09-DEC-01
10-DEC-01	10-DEC-01	10-DEC-01	10-DEC-01
11-DEC-01	11-DEC-01	11-DEC-01	11-DEC-01
12-DEC-01	12-DEC-01	12-DEC-01	12-DEC-01
13-DEC-01	13-DEC-01	13-DEC-01	13-DEC-01
14-DEC-01	14-DEC-01	14-DEC-01	14-DEC-01
15-DEC-01	15-DEC-01	15-DEC-01	15-DEC-01
16-DEC-01	16-DEC-01	16-DEC-01	16-DEC-01
17-DEC-01	17-DEC-01	17-DEC-01	17-DEC-01
18-DEC-01	18-DEC-01	18-DEC-01	18-DEC-01
19-DEC-01	19-DEC-01	19-DEC-01	19-DEC-01
20-DEC-01	20-DEC-01	20-DEC-01	20-DEC-01
21-DEC-01	21-DEC-01	21-DEC-01	21-DEC-01
22-DEC-01	22-DEC-01	22-DEC-01	22-DEC-01
23-DEC-01	23-DEC-01	23-DEC-01	23-DEC-01
24-DEC-01	24-DEC-01	24-DEC-01	24-DEC-01
25-DEC-01	25-DEC-01	25-DEC-01	25-DEC-01
26-DEC-01	26-DEC-01	26-DEC-01	26-DEC-01
27-DEC-01	27-DEC-01	27-DEC-01	27-DEC-01
28-DEC-01	28-DEC-01	28-DEC-01	28-DEC-01
29-DEC-01	29-DEC-01	29-DEC-01	29-DEC-01
30-DEC-01	30-DEC-01	30-DEC-01	30-DEC-01
31-DEC-01	31-DEC-01	31-DEC-01	31-DEC-01
01-JAN-02	01-JAN-02	01-JAN-02	01-JAN-02
02-JAN-02	02-JAN-02	02-JAN-02	02-JAN-02
03-JAN-02	03-JAN-02	03-JAN-02	03-JAN-02
04-JAN-02	04-JAN-02	04-JAN-02	04-JAN-02
05-JAN-02	05-JAN-02	05-JAN-02	05-JAN-02
06-JAN-02	06-JAN-02	06-JAN-02	06-JAN-02
07-JAN-02	07-JAN-02	07-JAN-02	07-JAN-02
08-JAN-02	08-JAN-02	08-JAN-02	08-JAN-02
09-JAN-02	09-JAN-02	09-JAN-02	09-JAN-02
10-JAN-02	10-JAN-02	10-JAN-02	10-JAN-02
11-JAN-02	11-JAN-02	11-JAN-02	11-JAN-02
12-JAN-02	12-JAN-02	12-JAN-02	12-JAN-02
13-JAN-02	13-JAN-02	13-JAN-02	13-JAN-02
14-JAN-02	14-JAN-02	14-JAN-02	14-JAN-02
15-JAN-02	15-JAN-02	15-JAN-02	15-JAN-02
16-JAN-02	16-JAN-02	16-JAN-02	16-JAN-02
17-JAN-02	17-JAN-02	17-JAN-02	17-JAN-02
18-JAN-02	18-JAN-02	18-JAN-02	18-JAN-02
19-JAN-02	19-JAN-02	19-JAN-02	19-JAN-02
20-JAN-02	20-JAN-02	20-JAN-02	20-JAN-02
21-JAN-02	21-JAN-02	21-JAN-02	21-JAN-02
22-JAN-02	22-JAN-02	22-JAN-02	22-JAN-02
23-JAN-02	23-JAN-02	23-JAN-02	23-JAN-02
24-JAN-02	24-JAN-02	24-JAN-02	24-JAN-02
25-JAN-02	25-JAN-02	25-JAN-02	25-JAN-02
26-JAN-02	26-JAN-02	26-JAN-02	26-JAN-02
27-JAN-02	27-JAN-02	27-JAN-02	27-JAN-02
28-JAN-02	28-JAN-02	28-JAN-02	28-JAN-02
29-JAN-02	29-JAN-02	29-JAN-02	29-JAN-02
30-JAN-02	30-JAN-02	30-JAN-02	30-JAN-02
31-JAN-02	31-JAN-02	31-JAN-02	31-JAN-02



Compañías de tarjetas de crédito

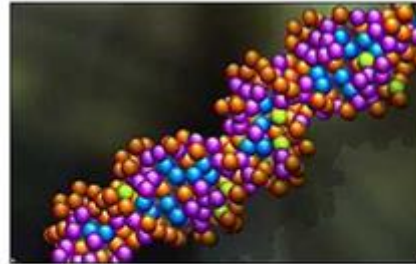
Aplicaciones

Astronomía



- Astronomical sky surveys
- 120 Gigabytes/week
- 6.5 Terabytes/year

Genómica



- 25,000 genes in human genome
- 3 billion bases
- 3 Gigabytes of genetic data

Telefonía



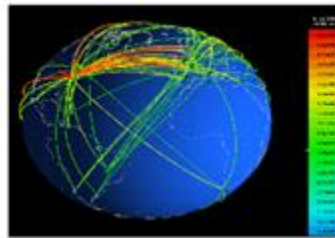
- 250M calls/day
- 60G calls/year
- 40 bytes/call
- 2.5 Terabytes/year

Transacciones de tarjetas de crédito



- 47.5 billion transactions in 2005 worldwide
- 115 Terabytes of data transmitted to VisaNet data processing center in 2004

Tráfico en Internet



Traffic in a typical router:

- 42 kB/second
- 3.5 Gigabytes/day
- 1.3 Terabytes/year

Procesamiento de información WEB



- 25 billion pages indexed
- 10kB/Page
- 250 Terabytes of indexed text data
- "Deep web" is supposedly 100 times as large

Aplicaciones

- Hospitales/medicina

Farmacias
Laboratorios
Centros de imagenes
Servicios de medicina de emergencia (EMS)
Sistemas de información de hospitales
Doc-in-a-Box
Registros medicos electrónicos
Bancos de sangre
Registros de nacimiento y muerte

- Tratamientos más efectivos para condiciones particulares
- Identificar patrones relacionados con los efectos de los fármacos
- Reducir costos
- Análisis mHealth, eHealth: registros medicos electrónicos , imagenes
- Mapeo de datos de salud con geografía (salud pública)

Aplicaciones

- Análisis de clientes
 - Comportamientos y preferencias
 - Extender datos tradicionales con datos de redes sociales, “browser logs”, análisis de texto, imágenes, datos de sensores

Aplicaciones

- Empresas de telecomunicaciones: predecir pérdida de clientes
- Compañías de seguros de carros: describir los comportamientos de conducción de los clientes
- Target (compañía de vta EU): predecir cuando los clientes van a tener bebé.

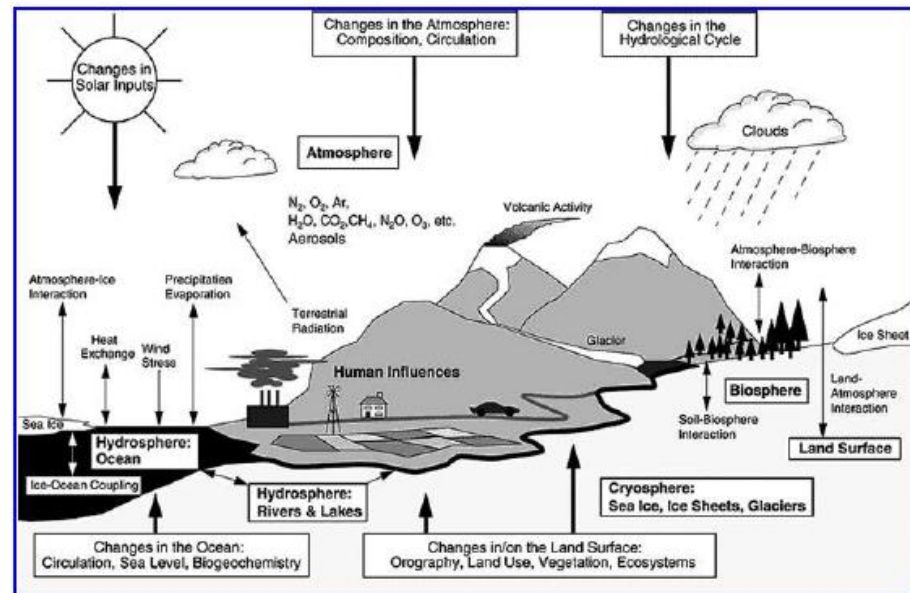
Aplicaciones



A BIG DATA GUIDE TO UNDERSTANDING CLIMATE CHANGE:

The Case for Theory-Guided Data Science

*James H. Faghmous and Vipin Kumar
Department of Computer Science and Engineering,
The University of Minnesota–Twin Cities
Minneapolis, Minnesota*



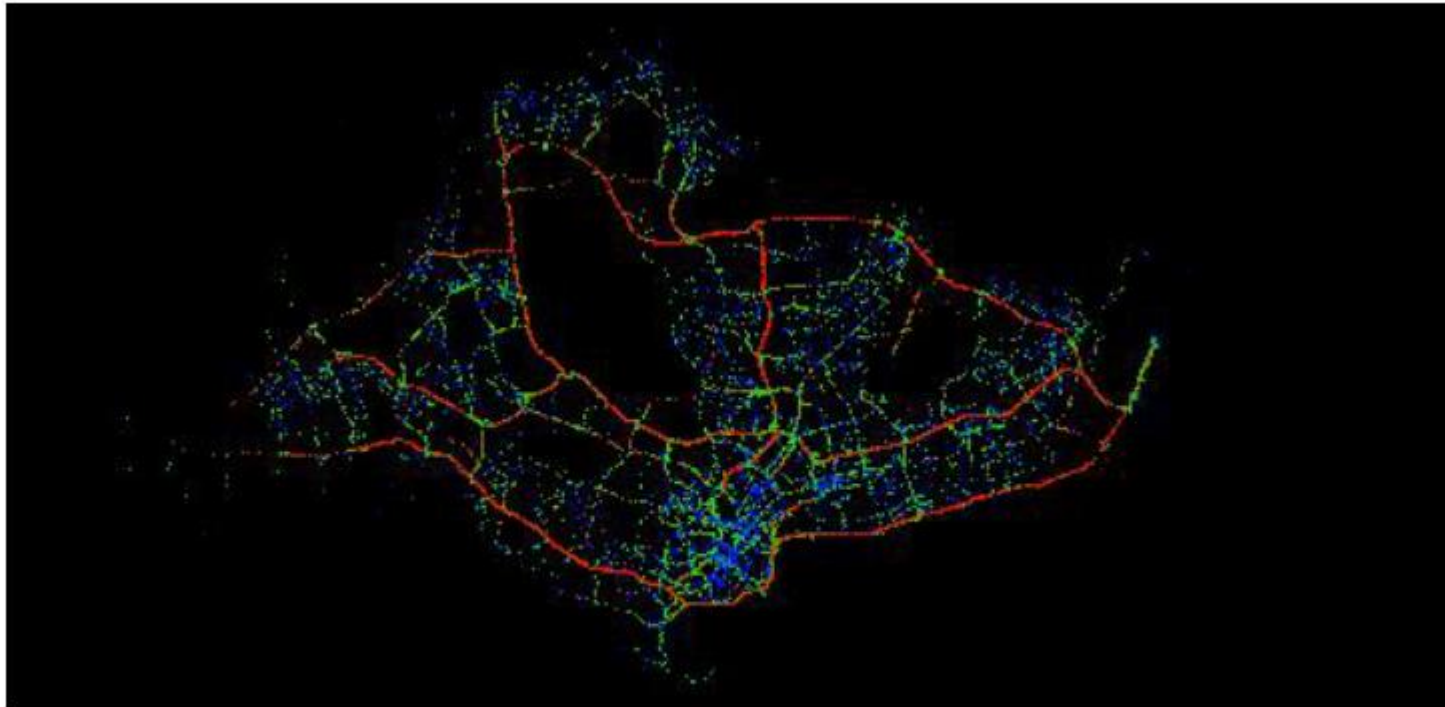
Big Data, Vol.2, 3, 2014

Aplicaciones

- Tráfico
 - Visualización del tráfico
 - Análisis de los patrones de desplazamiento
 - Rutas de congestión
 - Planificación urbana
 - Uso de las carreteras

Aplicaciones

- Visualización de tráfico



Slow

Medium

Fast

Aplicaciones

- Patrones en Tráfico
 - Flujo del tráfico (días hábiles/festivos, horas pico)
 - Lugares donde se origina el tráfico
 - Lugares destinos
 - Rutas más frecuentes
 - Planeación, asignación de recursos

Aplicaciones

Detección de tráfico



CarTel sensor node:
[Hull, Madden, Balakrishnan 2006]

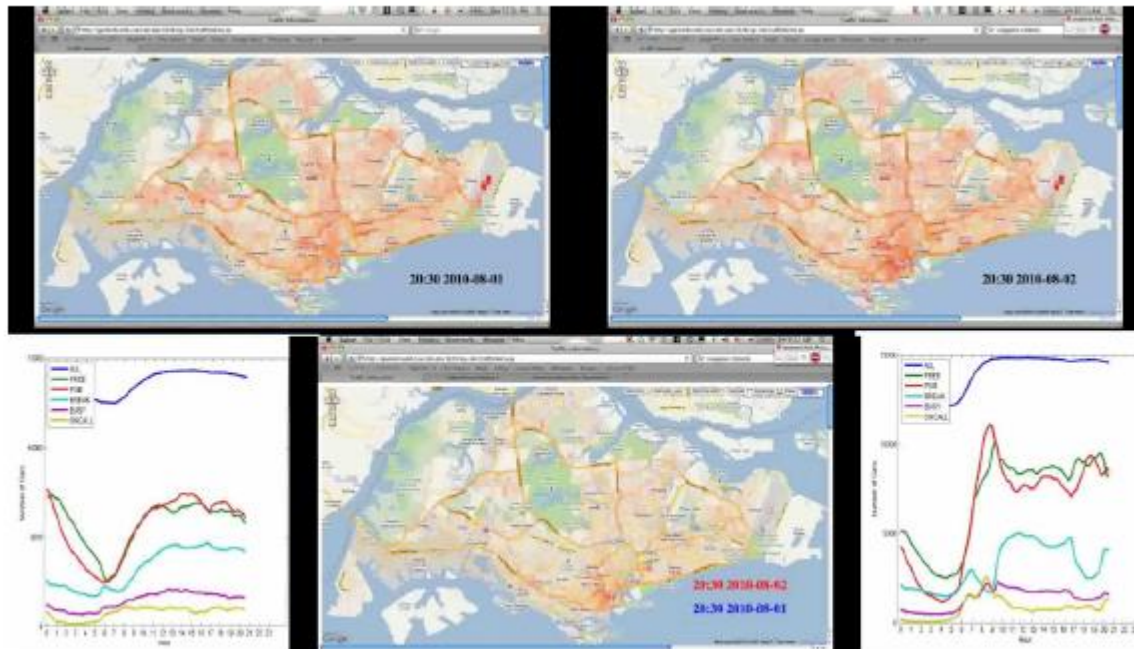


Modelamiento:

- GPS en automoviles:
 - capturan velocidad, paradas
 - solución fácil
- Sensores en el asfalto
 - Flujo
 - Solución costosa

Aplicaciones

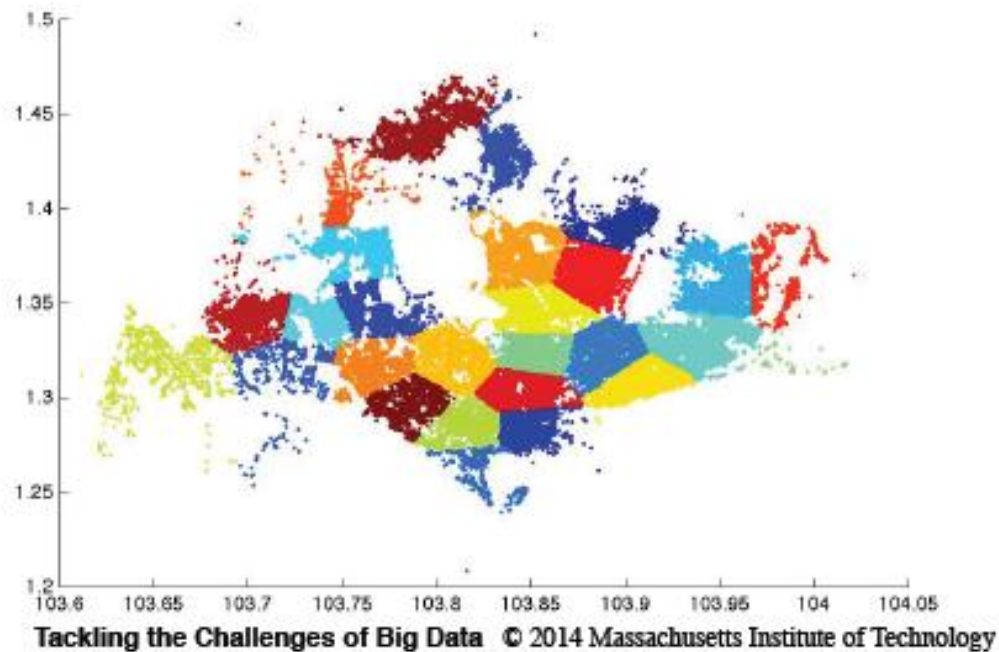
- Singapur



Visualización del tráfico

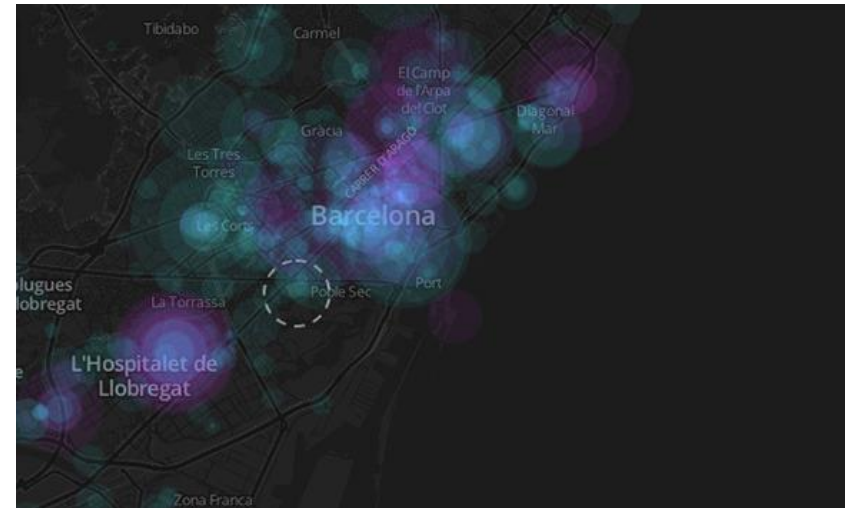
Aplicaciones

- What are origins/destinations?
 - GPS coordinates, street addresses, *regions*...



Aplicaciones

BBVA – Análisis viual de las transacciones de tarjetas de crédito



<http://www.centrodeinnovacionbbva.com/noticias/ejemplos-reales-del-uso-de-big-data>

<https://youtu.be/Zel6wych9p0>

<http://mwcimpact.com>

Google Knowledge Graph



<https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html>

Aplicaciones

- Deportes: seguimientos al rendimiento de cada jugador (análisis de videos), tecnología de sensores (cestas, mallas, etc.) usando telefonos inteligentes y servicios de la nube.
- NBA -> preparación de partidos
- NFL (National Football League)
<https://www.nfl.com/now>
<https://youtu.be/aztUUcZfXb8> (optimización de las agendas de los partidos usando IBM tool)

Aplicaciones

Análisis de Redes Sociales

- Campaña presidencial Obama – Reelección 2012
20 personas centrados en la interpretación de los datos recibidos:
 - registro (recoger datos de los votantes convencidos),
 - persuasión (dirigirse a los dudosos de una forma eficaz)
 - voto del electorado (asegurarse de que los partidarios fueran a ejercer el voto sí o sí).
- Análisis de sentimientos



Aplicaciones

A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election [6 Ceron, León]

Performance Evaluation

Overall accuracy of 92.62%, TPR of 93.02%, and FPR of 7.78%.

		Predicted	
		Spammer	Non-Spammer
Actual	Spammer	253	19
	Non-Spammer	21	249

Table: Performance evaluation on the test set

Aplicaciones

A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election

Regression

Class	Precision	Recall	F1-score
Positive	0.65	0.43	0.52
Negative	0.62	0.74	0.67
Neutral	0.56	0.55	0.55

Table: Discriminative power of the system for each class

Aplicaciones

A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election

- Voting Intention Inference

Candidate	Result	Polls	Twitter volume	Proposed method
Zuluaga	29.28%	27.53%	24.10%	29.21%
Santos	25.72%	28.99%	35.12%	28.34%
Ramírez	15.52%	9.43%	8.99%	9.23%
López	15.21%	10.56%	12.09%	10.15%
Peñalosa	8.27%	11.25%	8.13%	11.54%
<i>Blank vote</i>	5.98%	12.24%	11.65%	12.87%

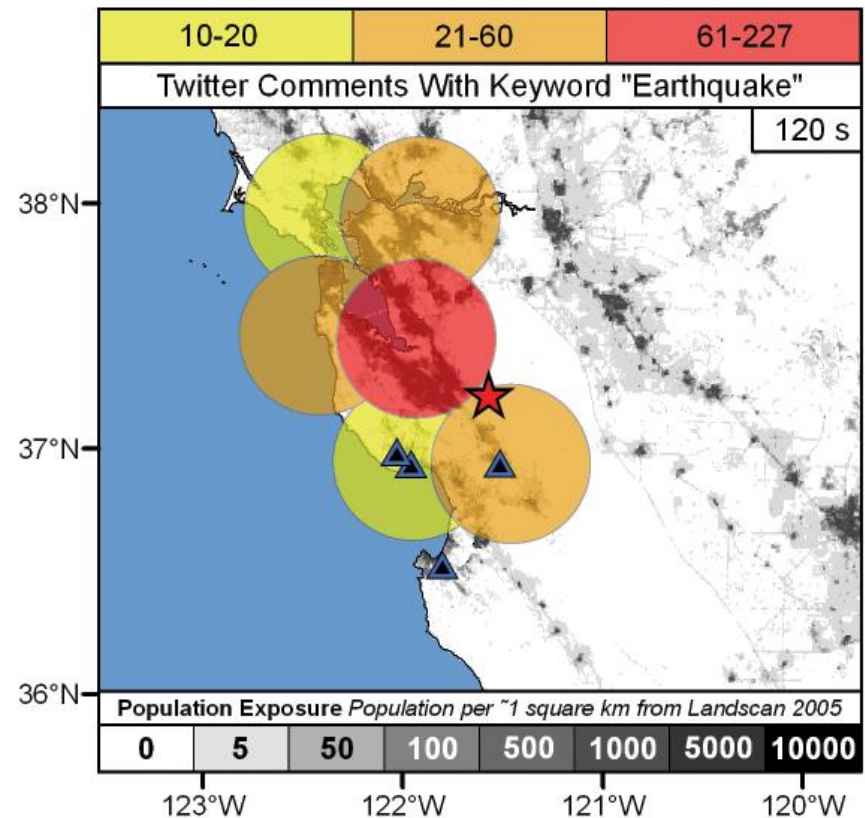
Table: Results and voting inference per method in the first round election. Numbers in bold show the inference method with the lowest absolute error that correctly ranked a candidate

Aplicaciones

Tweets usados para un sistema de advertencias de terremotos

The United States Geological Survey Twitter busca incrementar el volumen de los mensajes sobre sismos y ser capaz de localizar terremotos con un 90% de precisión.

<http://www.kurzweilai.net/tweets-used-as-earthquake-warning-system>



Big Data

- Internet fuente de datos considerada “Big Data” por lo que ha capturado la atención de la industria.



Big Data

El progreso y la innovación no se ven obstaculizados solo por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable



Big Data

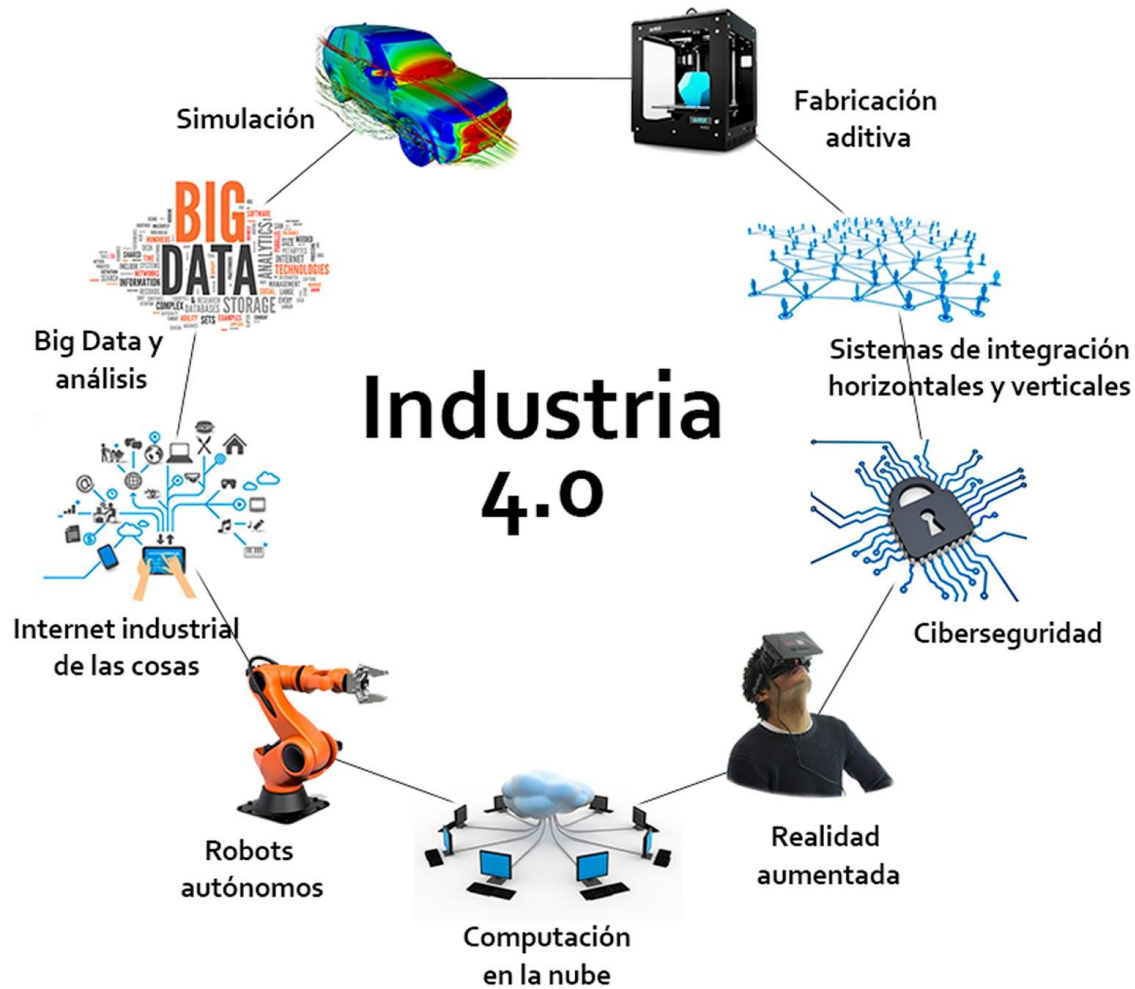
- La competitividad en la productividad de los negocios y las tecnologías seguramente van a converger a las exploraciones en Big data[1 Philip Chen].

Big Data

- Oportunidad:
 - Toma de decisiones

- Reto:
 - Manejar inconsistencias, datos incompletos, escalabilidad, flujo de datos, problemas de seguridad.
 - Se requieren nuevas tecnologías para: almacenamiento, operaciones de entrada/salida de datos, procesamiento y análisis.

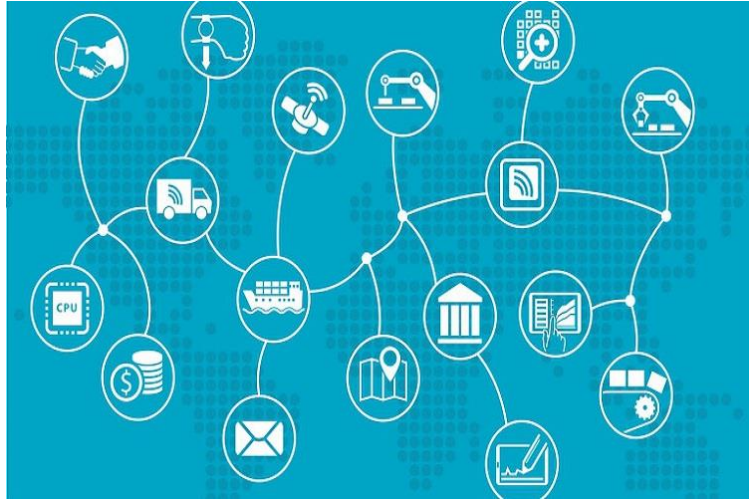




industria-4.blogspot.com

[https://www.laprensagrafica.com/__export/1508349926661/sites/prensagrafica/img/2017/10/18/innovacxn_empresarial.jpg]

Internet de las cosas



[<https://www.muycomputerpro.com/zona-transformacion-digital/iot-industrial-optimizacion-transformacion/>]

- IPv6: todo objeto conectado ($6,7 \times 10^{17}$ direcciones por milímetro cuadrado)
- Dispositivos pequeños y baratos, GPS, todos los aparatos podrán tener su propia IP

[<https://industrie4.0.gtai.de/INDUSTRIE40/Navigation/EN/Topics/The-internet-of-things/internet-ofthings-what-is-it.html>]

“Autonomous things” (Cosas automáticas):
robots, vehículos, drones, agentes.

Big Data

Obliga a:

- Trabajar con mucha información privada y romper con los enfoques clásicos de seguridad de los datos.
- Manipular enormes cantidades de datos no estructurados.
- Mucho intercambio y cooperación internacional.
- Romper con el enfoque relacional de las bases de datos.
- Buscar nuevas alternativas para el procesamiento paralelo.

Referencias

- [1] Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. C.L. Philip Chen, Chun-Yang Zhang. Information Sciences 275 (2014) 314–347
- [2] Eric Savitz, Gartner: Top 10 Strategic Technology Trends for 2013, October 2012.
<<http://www.forbes.com/sites/ericsavitz/2012/10/23/gartner-top-10-strategic-technology-trends-for-2013/>>.
- [3] Eric Savitz, Gartner: 10 Critical Tech Trends for the Next Five Years, October 2012.
<http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/>
- [4] Richard T. Kouzes, Gordon A. Anderson, Stephen T. Elbert, Ian Gorton, Deborah K. Gracio, The changing paradigm of data-intensive computing, Computer 42 (1) (2009) 26–34.
- [5] Rus, Daniela. Tackling the Challenges of Big Data. MIT. 2016
- [6] A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election. Tesis de Maestría. 2016