Introduction to Big Data

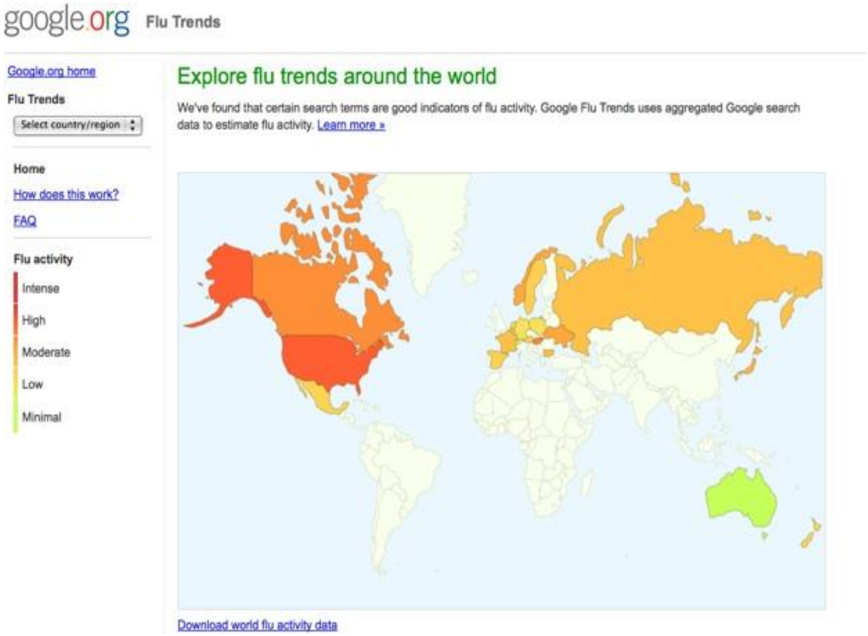
Based on “Big Data: Hype or Hallelujah?” by Elena Baralis
http://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/BigData_2015_2x.pdf

Big data

big data
information
search
petabytes
internet
management
analysis
science
storage
large
software
future
research
sets
new
billion
real
process
technology
volume
exabytes
size
world
social
million time
approaches
development
business
databases
terms
sensor
based
used
uses
users
many
higher
out
announced
decision
processing
online institute
massive
use
national systems
year
critiques
various
countries
mapreduce
every
initiative
technologies
government
learning
years
less
parallel
san
nas
high
applications

BIG
DATA
STORAGE
COMPLEX
ANALYTICS
TECHNOLOGIES
INFORMATION
DATABASES
RESEARCH
EXAMPLES
PETABYTES
SHARED
TARGET
CAPTURE
SOCIAL
LARGER
NEEDED
HUNDREDS
SAN
TERABYTES
INCLUDE
TIME
PARALLEL
SYSTEMS
CASE
RECORDS
SEARCH
LOGS
USING
INTERNET
MANAGEMENT
SETS
EVERY
MANAGE
CURRENT
CONTINUES
WITHIN
DIFFICULTY
SIZE
ALSO
BUSINESS
ABILITY
MASSIVELY
NETWORKS
SOFTWARE
DISK
CAPACITY
GARTNER
CITATION
MPP
RELATIONAL
TOOLS
MOVING
SET
COST
PERFORMANCE
LARGE
PROCESSING
PROCESS
DEFINITION
WORKING
MAY
PRACTITIONERS
BIOLOGICAL
ONE
STORE
DISTRIBUTED
GENOMICS APPLIED
USE

Google Flu trends

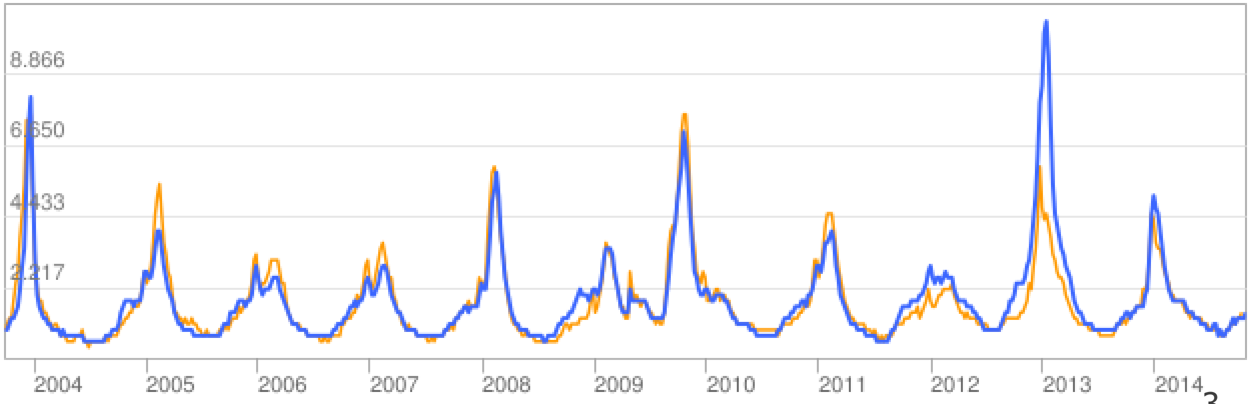


Explore flu trends around the world

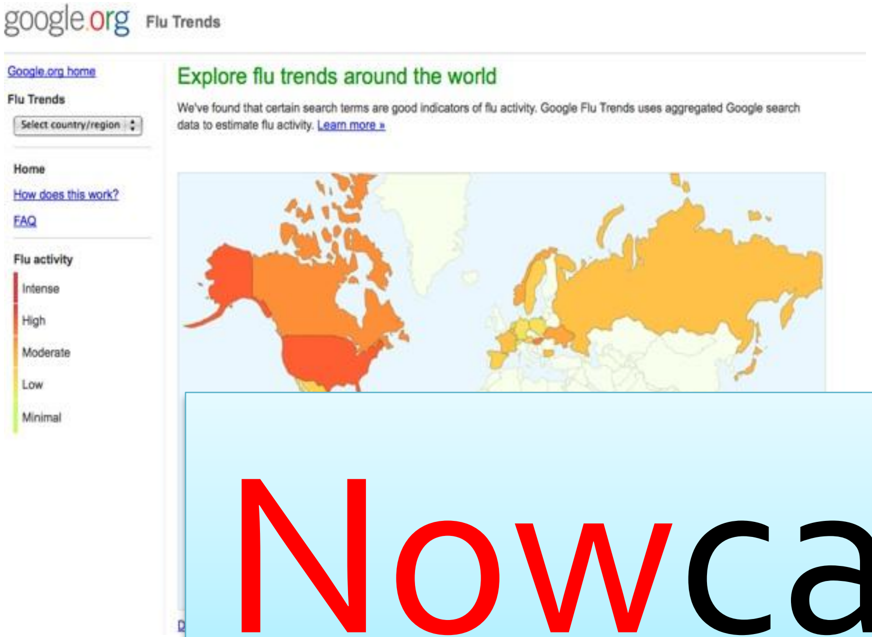
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Download world flu activity data

- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and Prevention – U.S.A)
  - Based on the analysis of Google search queries



Google Flu trends



google.org Flu Trends

Google.org home

Flu Trends

Select country/region

Home

How does this work?

FAQ

Flu activity

Intense

High

Moderate

Low

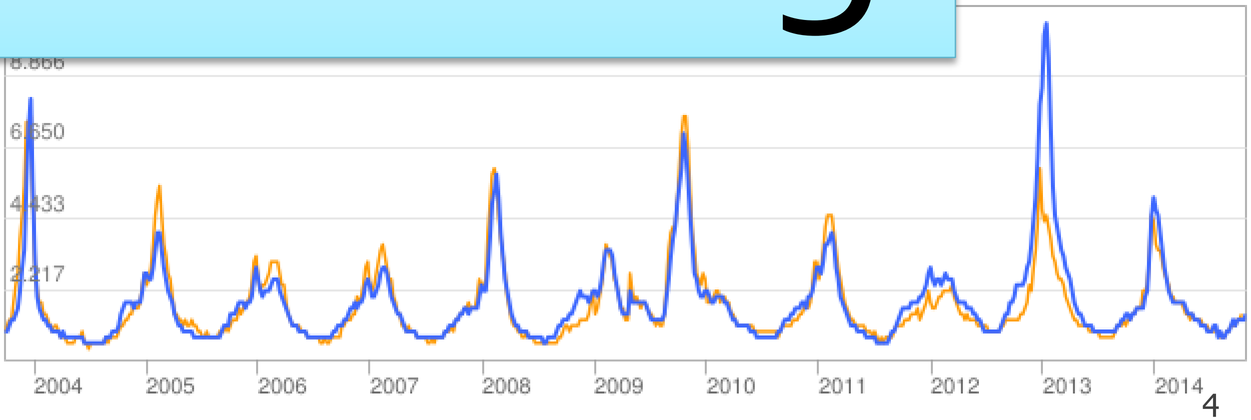
Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

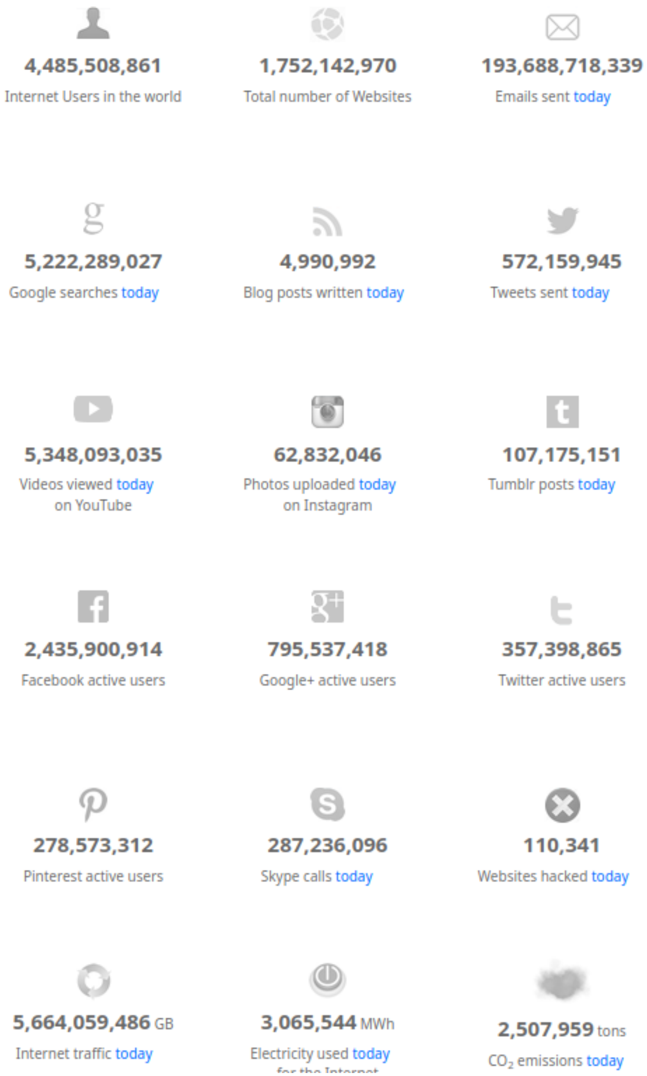
- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and
    .A)
  - alysis of
    ueries

Nowcasting



Data on the Internet…

- Internet live stats
  - http://www.internetlivestats.com/



4,485,508,861
Internet Users in the world

1,752,142,970
Total number of Websites

193,688,718,339
Emails sent today

5,222,289,027
Google searches today

4,990,992
Blog posts written today

572,159,945
Tweets sent today

5,348,093,035
Videos viewed today on YouTube

62,832,046
Photos uploaded today on Instagram

107,175,151
Tumblr posts today

2,435,900,914
Facebook active users

795,537,418
Google+ active users

357,398,865
Twitter active users

278,573,312
Pinterest active users

287,236,096
Skype calls today

110,341
Websites hacked today

5,664,059,486 GB
Internet traffic today

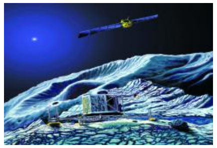
3,065,544 MWh
Electricity used today for the Internet

2,507,959 tons
CO_2 emissions today

Who generates big data?

- User Generated Content (Web & Mobile)
  - E.g., Facebook, Instagram, Yelp, TripAdvisor, Twitter, YouTube

- Health and scientific computing

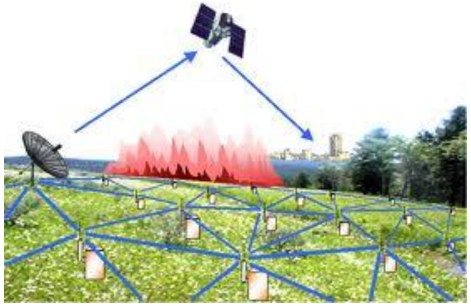


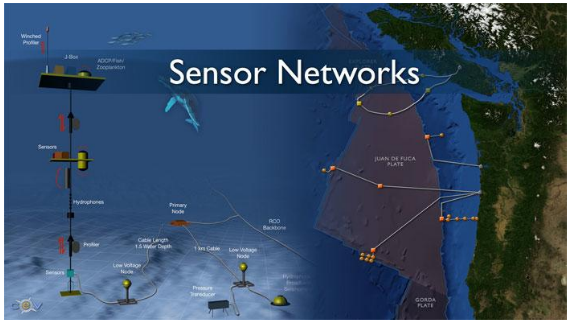
Who generates big data?

- Log files
  - Web server log files, machine system log files

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
aliquam massa posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobortis lorem eget vi
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
aliquam massa posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobortis lorem eget vi
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
aliquam massa posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobortis lorem eget vi
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
aliquam massa posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobortis lorem eget vi
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nislva posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobo
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
aliquam massa posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobortis lorem eget vi
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
aliquam massa posuere.Fusce laoreet risus et dui sodales et fringilla lorem pellentesque. Nulla sagittis lobortis lorem eget vi
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean at urna at eros dignissim adipiscing. Sed ullamcorper lectus a
Maecenas dapibus arcu ut nisl eleifend elementum. Morbi turpis magna, placerat ut ultricies vel, malesuada non massa. Class apt
Proin quis porttitor nulla. Maecenas in magna risus. Nulla enim urna, sagittis dignissim egestas ac, sodales ac nulla. Lorem ip
corper tortor tincidunt eu.
vestibulum vel facilisis nulla. Pellentesque eget neque non augue condimentum posuere a ac augue. Praesent eu dolor orci. Null
```

- Internet Of Things (IoT)
  - Sensor networks, RFIDs, smart meters





An example of Big data at work

Crowdsourcing

Sensing

Map data

Computing

Real time traffic info

Travel time forecast/nowcast

What is big data?

CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

DilbertCartoonist@gmail.com

©2012 Scott Adams, Inc./Dist. by Universal Uclick

- Many different definitions
  - "Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"

What is big data?

CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc./Dist. by Universal Uclick
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

- Many different definitions
  - "Data whose **scale**, **diversity** and **complexity** require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"

What is big data?

CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc./Dist. by Universal Uclick
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

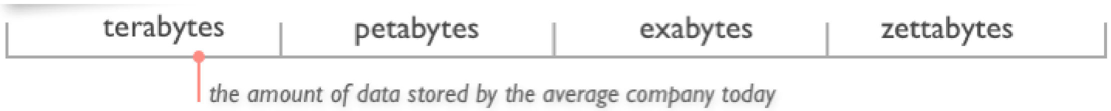
- Many different definitions
  - "Data whose scale, diversity and complexity require new **architectures**, **techniques**, **algorithms** and **analytics** to manage it and extract value and hidden knowledge from it"

The Vs of big data

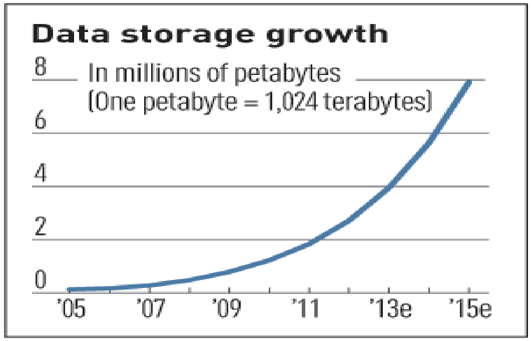
- The 3Vs of big data
  - Volume: scale of data
  - Variety: different forms of data
  - Velocity: analysis of streaming data
- … but also
  - Veracity: uncertainty of data
  - Value: exploit information provided by data

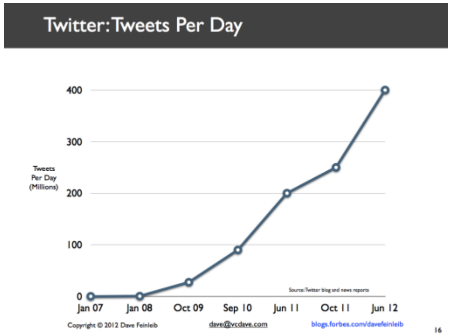
The Vs of big data

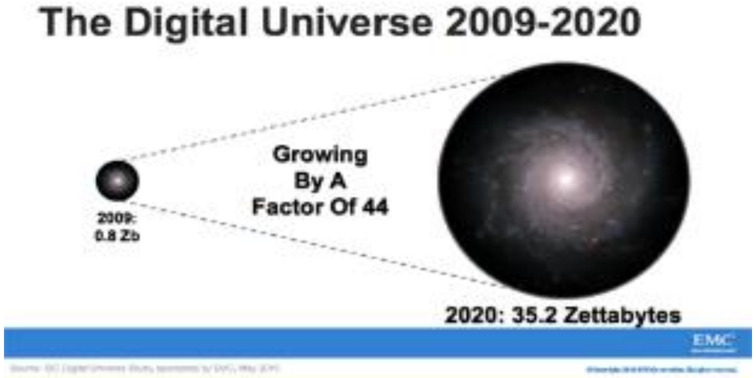
- Volume



  - Data volume increases exponentially over time
  - 44x increase from 2009 to 2020
    - Digital data 35 ZB in 2020

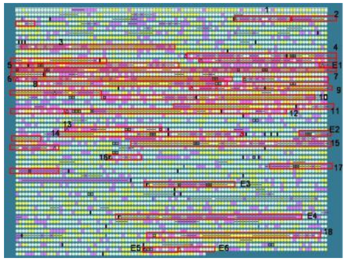


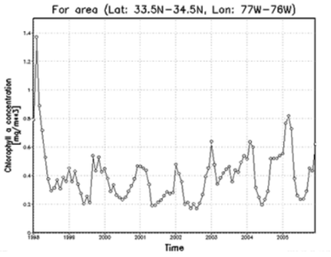


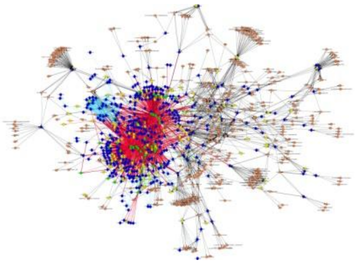

The Digital Universe 2009-2020
Growing By A Factor Of 44
2009: 0.8 Zb
2020: 35.2 Zettabytes
EMC
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

The Vs of big data

- Variety
  - Various formats, types and structures
    - Numerical data, image data, audio, video, text, time series



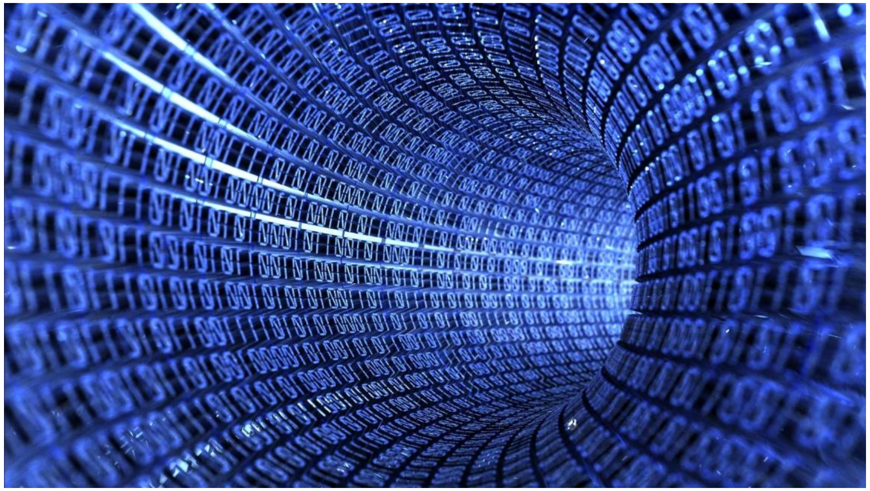

For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m**3]
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1998
1999
2000
2001
2002
2003
2004
2005
Time



  - A single application may generate many different formats
    - Heterogeneous data
    - Complex data integration problem

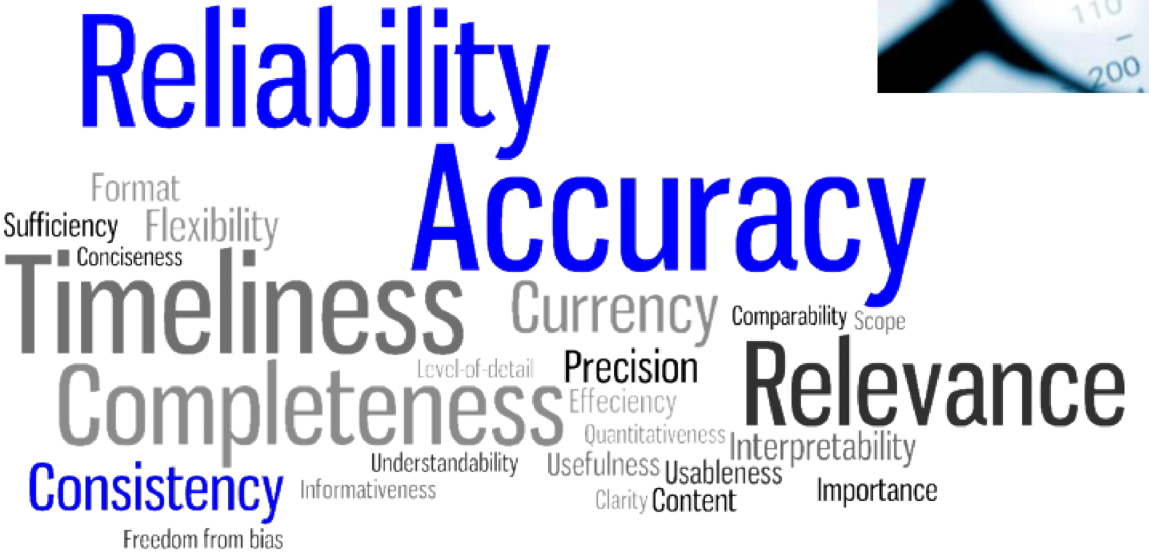
The Vs of big data

- Velocity
  - Fast data generation rate
    - Streaming data
  - Very fast data processing to ensure timeliness



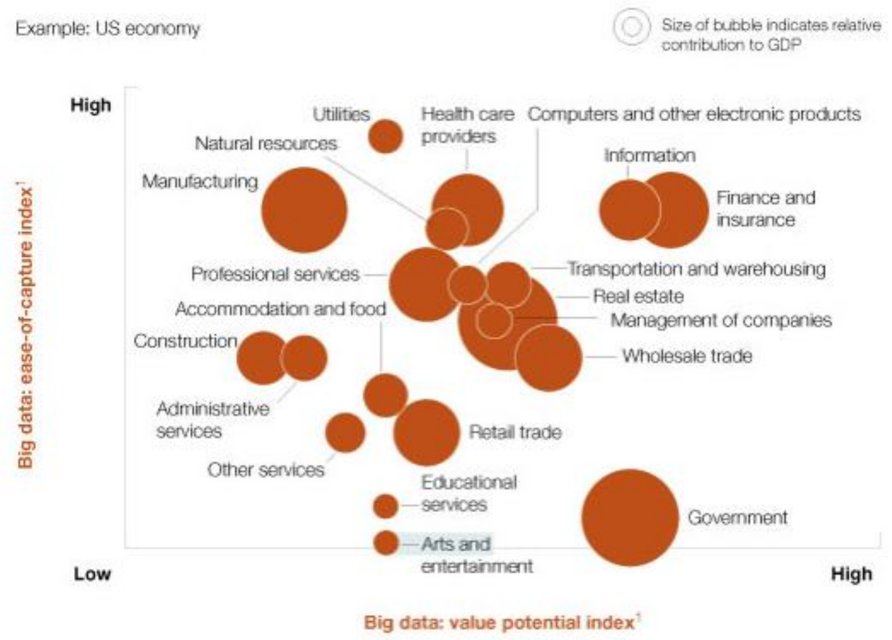
The Vs of big data

- Veracity
  - Data quality



The Vs of big data

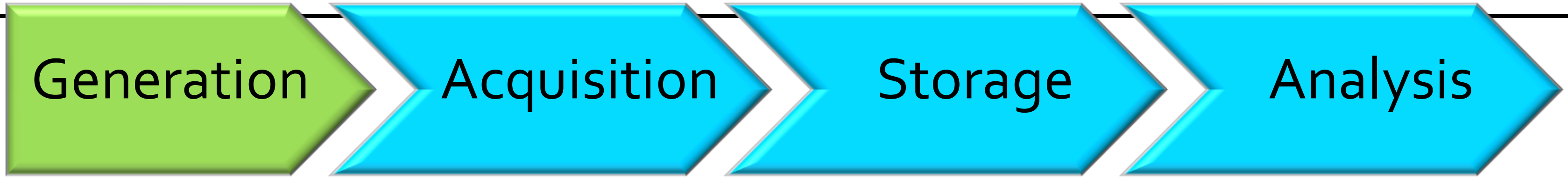
- Value
  - Translate data into business advantage



[1] For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

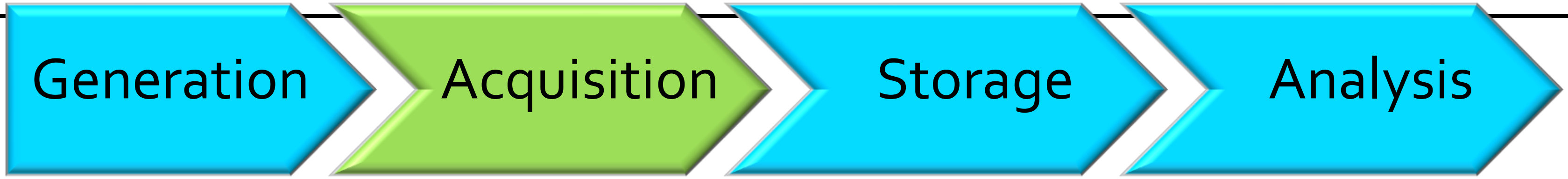
Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Big data value chain



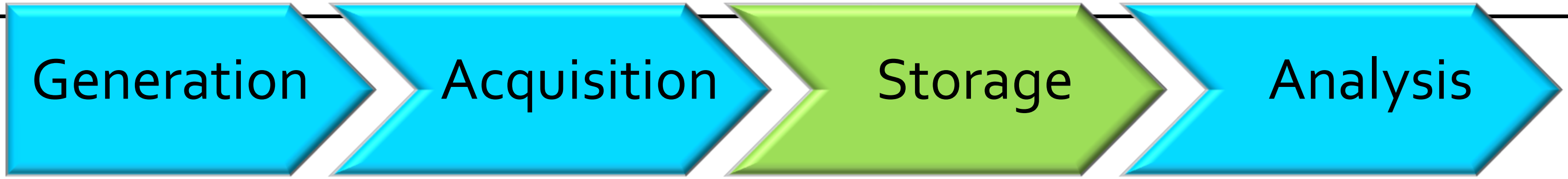
- Generation
  - Passive recording
    - Typically structured data
    - Bank trading transactions, shopping records, government sector archives
  - Active generation
    - Semistructured or unstructured data
    - User-generated content, e.g., social networks
  - Automatic production
    - Location-aware, context-dependent, highly mobile data
    - Sensor-based Internet-enabled devices

Big data value chain



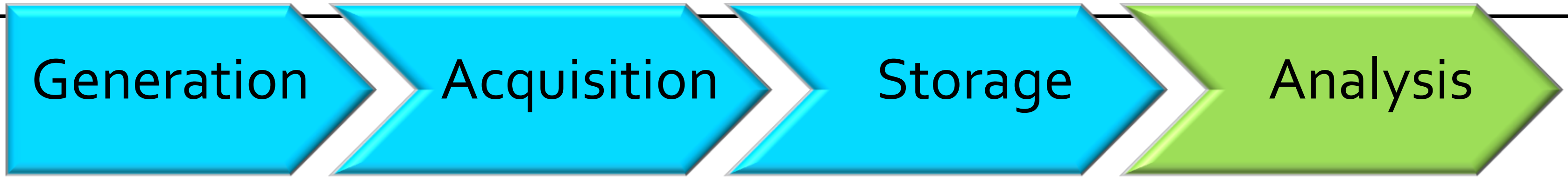
- Acquisition
  - Collection
    - Pull-based, e.g., web crawler
    - Push-based, e.g., video surveillance, click stream
  - Transmission
    - Transfer to data center over high capacity links
  - Preprocessing
    - Integration, cleaning, redundancy elimination

Big data value chain



- Storage
  - Storage infrastructure
    - Storage technology, e.g., HDD, SSD
    - Networking architecture, e.g., DAS, NAS, SAN
  - Data management
    - File systems (HDFS), key-value stores (Memcached), column-oriented databases (Cassandra), document databases (MongoDB)
  - Programming models
    - Map reduce, stream processing, graph processing

Big data value chain



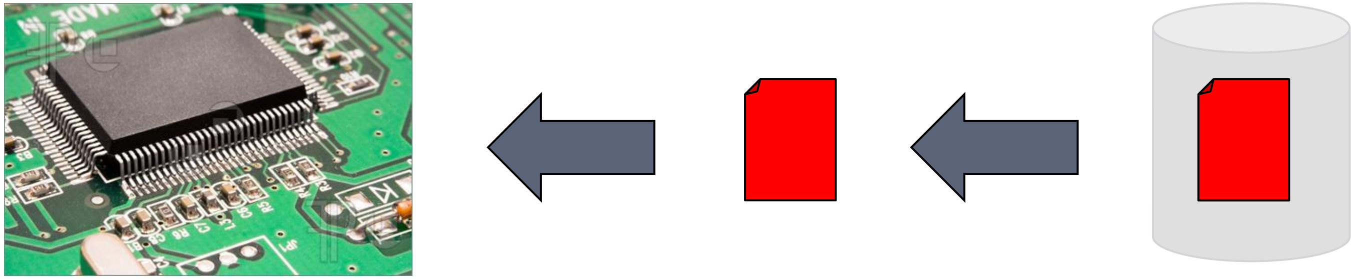
- Analysis
  - Objectives
    - Descriptive analytics, predictive analytics, prescriptive analytics
  - Methods
    - Statistical analysis, data mining, text mining, network and graph data mining
    - Clustering, classification and regression, association analysis
  - Diverse domains call for customized techniques

Big data challenges

- Technology and infrastructure
  - New architectures, programming paradigms and techniques are needed
- Data management and analysis
  - New emphasis on "data"
  - ➡ **Data science**

The bottleneck

- Processors process data
- Hard drives store data
- We need to transfer data from the disk to the processor



The solution

- **Transfer the processing power to the data**
- Multiple distributed disks
  - Each one holding a portion of a large dataset
- Process in parallel different file portions from different disks