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

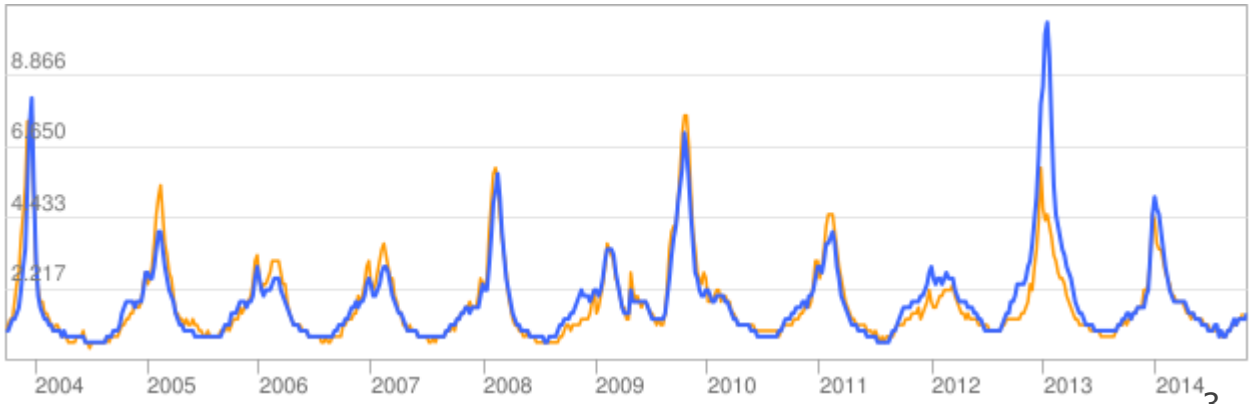
Big data


big data
information
search
management
science
analysis
internet
petabytes
storage
large
software
future
billion
sets
research
shutterstock


BIG
DATA
STORAGE
COMPLEX
LARGE
PROCESSING
PRACTITIONERS
BIOLOGICAL
STORE
ONE
DISTRIBUTED
RESEARCH
GENOMICS APPLIED
EXAMPLES
PETABYTES
DATABASES
SHARED
TARGET
CITATION
MPP
CAPTURE
SOCIAL
NETWORKS
SOFTWARE
SAN
NEEDED
HUNDREDS
TERABYTES
TECHNOLOGIES
SETS
INCLUDE
TIME
PARALLEL
EVERY
INTERNET
SYSTEMS
INFORMATION
RECORDS
SEARCH
ANALYTICS
SIZE
BUSINESS
MASSIVELY
ABILITY

Google Flu trends

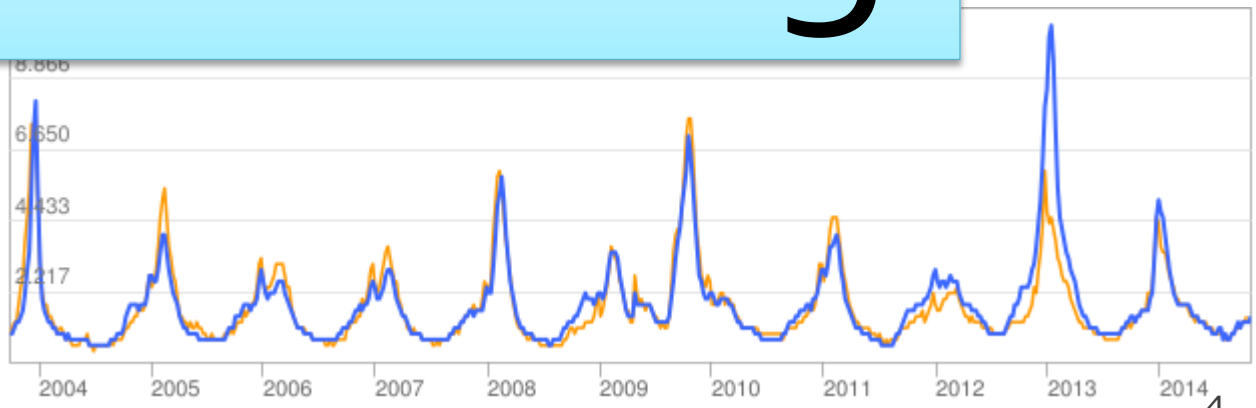
- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and Prevention – U.S.A)
  - Based on the analysis of Google search queries



Google Flu trends

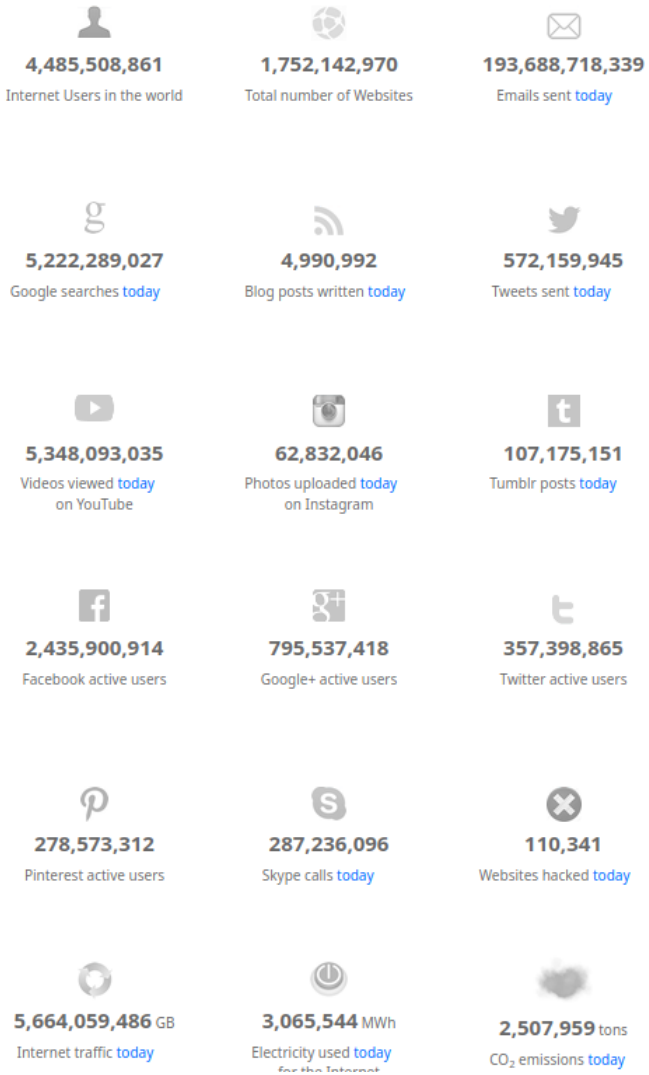
- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and
    .A)
  - lysis of
    ueries

Nowcasting



Data on the Internet…

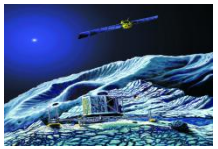
- Internet live stats
  - http://www.internetlivestats.com/



Who generates big data?

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

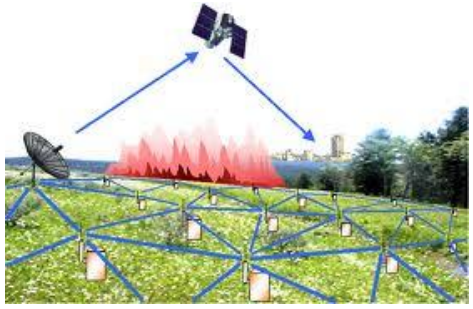
- Health and scientific computing



Who generates big data?

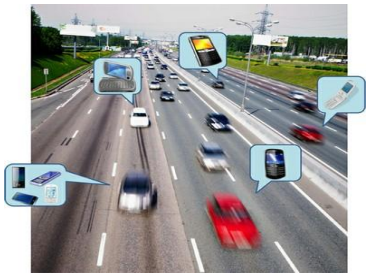
- Log files
  - Web server log files, machine system log files

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



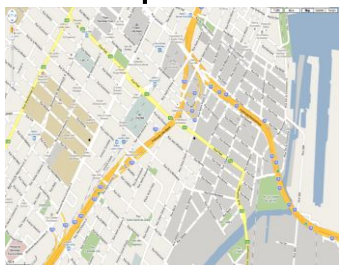
An example of Big data at work

Crowdsourcing



Sensing

Map data



Computing

Real time traffic info

Travel time forecast/nowcast

What is big data?

- 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?

- 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?

- 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"

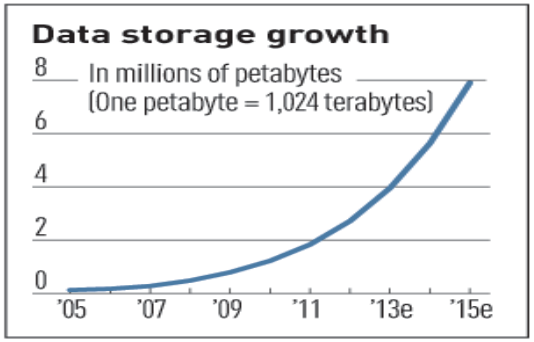
The Vs of big data

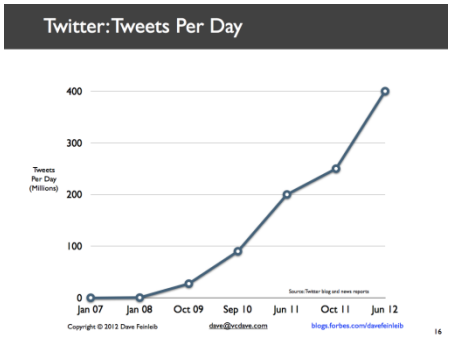
- The 3Vs of big data
  - Volume: scale of data
  - Variety: different forms of data
  - Velocity: analysis of streaming data
- … but also
  - Veracity: uncertainty of data
  - Value: exploit information provided by data

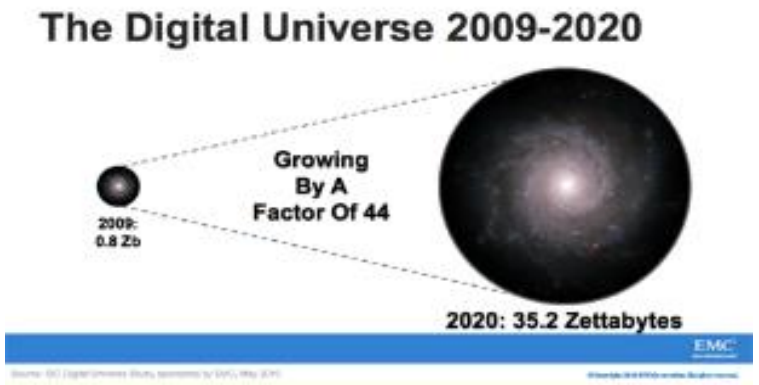
The Vs of big data

- Volume

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

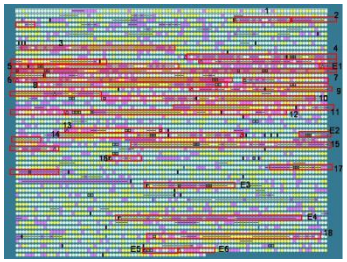


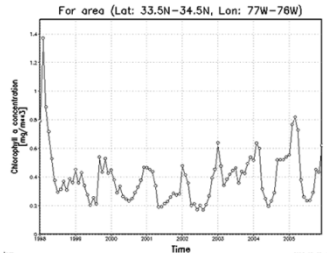


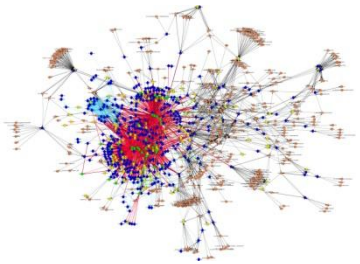


The Vs of big data

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



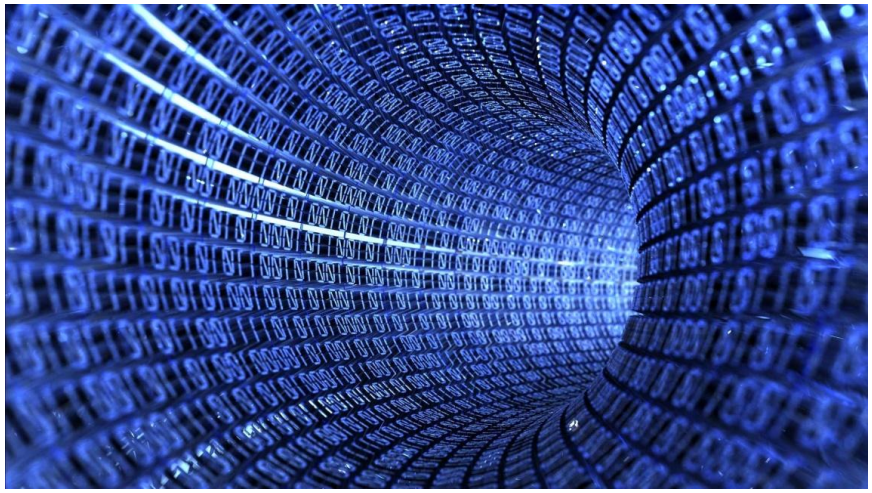




  - A single application may generate many different formats
    - Heterogeneous data
    - Complex data integration problem

The Vs of big data

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



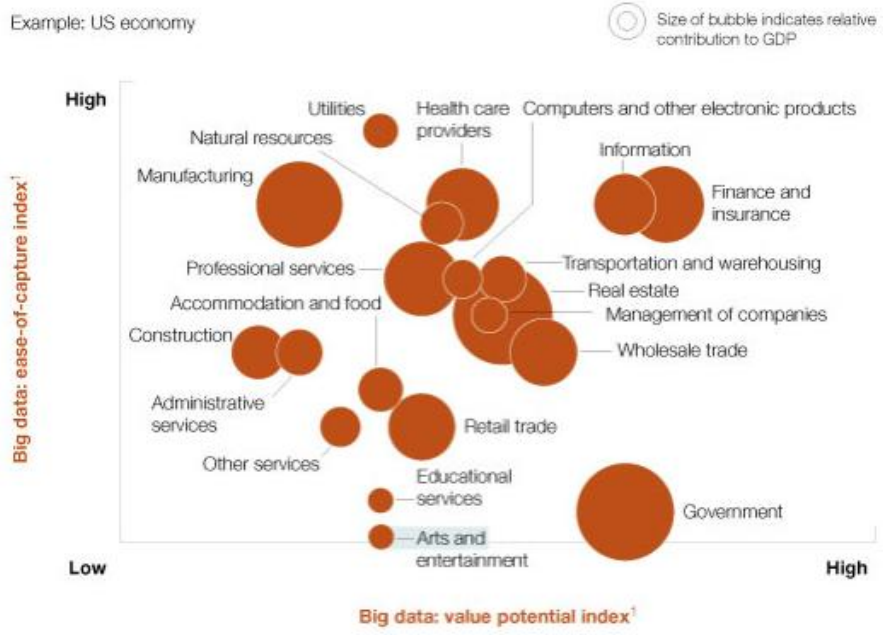
The Vs of big data

- Veracity
  - Data quality

Reliability
Accuracy
Format
Sufficiency
Flexibility
Conciseness
Timeliness
Currency
Comparability
Scope
Level-of-detail
Precision
Relevance
Completeness
Effeciency
Quantitativeness
Interpretability
Understandability
Usefulness
Usableness
Consistency
Informativeness
Clarity
Content
Importance
Freedom from bias

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

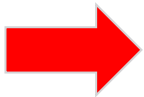
Big data value chain

- 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

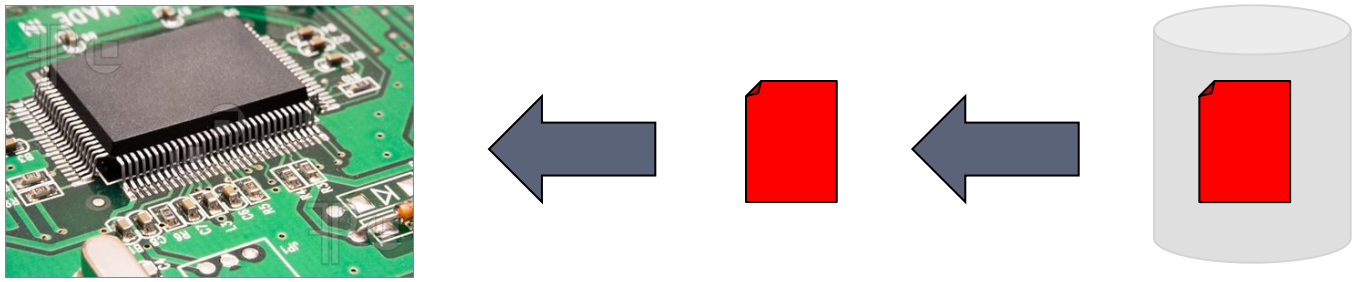
- Analysis
  - Objectives
    - Descriptive analytics, predictive analytics, prescriptive analytics
  - Methods
    - Statistical analysis, data mining, text mining, network and graph data mining
    - Clustering, classification and regression, association analysis
  - Diverse domains call for customized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