

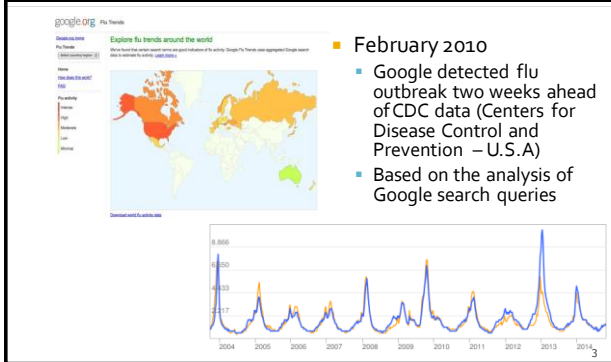
Introduction to Big Data

Based on "Big Data: Hype or Hallelujah?" by Elena Baralis
http://dbdm.polito.it/wordpress/wp-content/uploads/2010/12/BioData_2015_2x.pdf

Big data



Google Flu trends



Google Flu trends



Data on the Internet...

Internet live stats

- <http://www.internetlivestats.com>

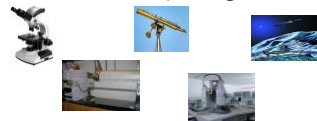


Who generates big data?

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



- Health and scientific computing



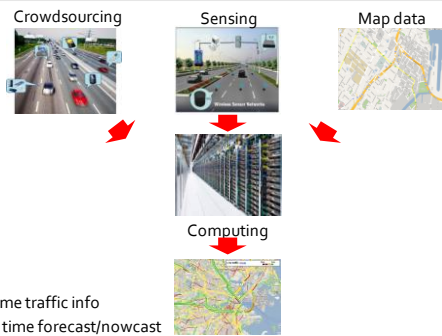
Who generates big data?

- Log files
 - Web server log files, machine system log files
- Internet Of Things (IoT)
 - Sensor networks, RFIDs, smart meters



7

An example of Big data at work



8

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"

9

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"

10

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"

11

The Vs of big data

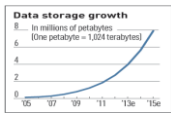
- The 3Vs of big data
 - V**olume: scale of data
 - V**ariety: different forms of data
 - V**elocity: analysis of streaming data
- ... but also
 - V**eracity: uncertainty of data
 - V**alue: exploit information provided by data

12

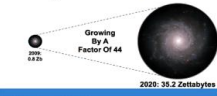
The Vs of big data

Volume

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



The Digital Universe 2009-2020

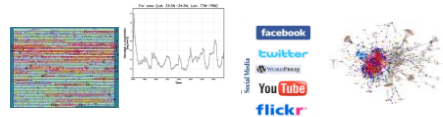


13

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

14

The Vs of big data

Velocity

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



15

The Vs of big data

Veracity

- Data quality

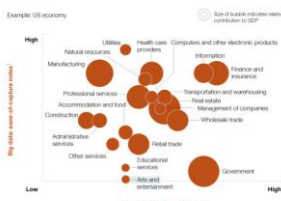


16

The Vs of big data

Value

- Translate data into business advantage



The detailed replication of events, as opposed to McKinsey Global Institute's full report Big data: The next frontier for innovation, competition, and productivity, available for download at mckinsey.com/industries/big-data. Source: McKinsey & Company, McKinsey Global Institute analysis.

17

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

18

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

19

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

20

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

21

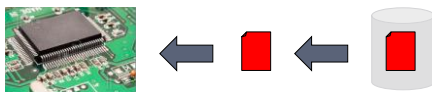
Big data challenges

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

22

The bottleneck

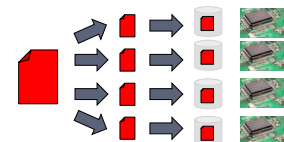
- Processors process data
- Hard drives store data
- We need to transfer data from the disk to the processor



23

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



24