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

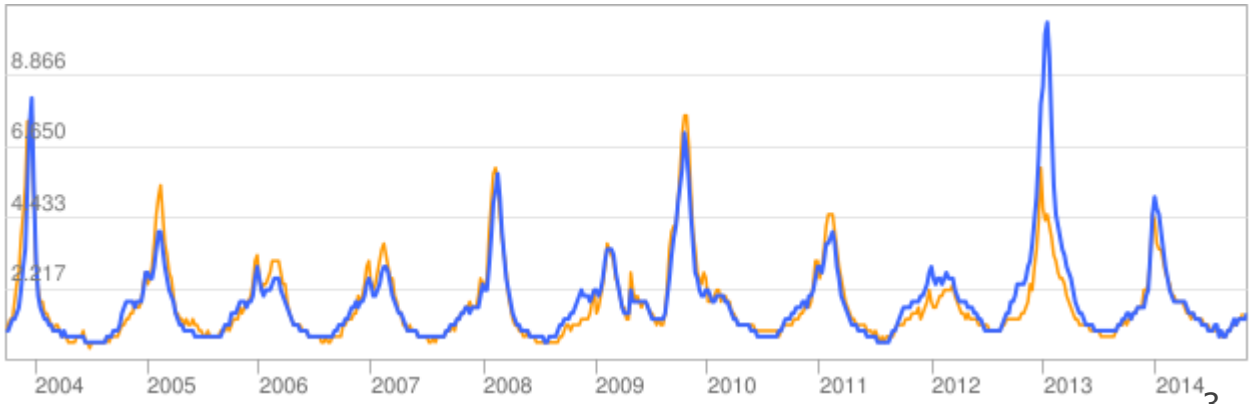
Big data


big data
information
search
management
science
analysis
based
internet
research
petabytes
storage
large
software
future
billion
use
sets
new
nas
size
volume
exabytes
technology
process
years
million
time
approaches
development
real
decision
processing
online institute
shutterstock

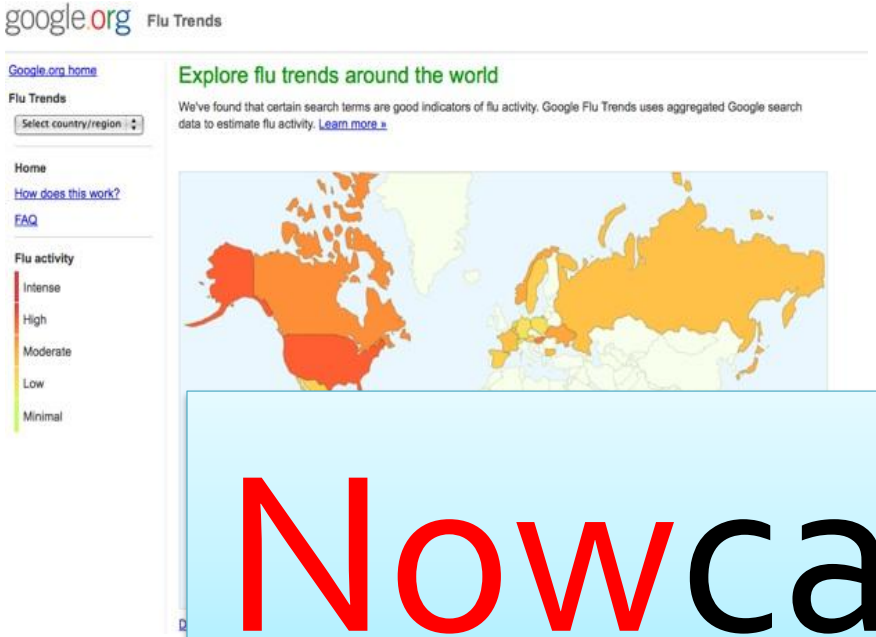

BIG
DATA
STORAGE
COMPLEX
ANALYTICS
TECHNOLOGIES
INFORMATION
RESEARCH
EXAMPLES
PETABYTES
DATABASES
SHARED
DISTRIBUTED
LARGE
PROCESSING
PRACTITIONERS
BIOLOGICAL
ONE
STORE
PROCESS
DEFINITION
WORKING
COST
SET
MOVING
TOOLS
RELATIONAL
GENOMICS APPLIED
MPP
CITATION
TARGET
USE
GARTNER
CAPACITY
CAPTURE
SOCIAL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

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

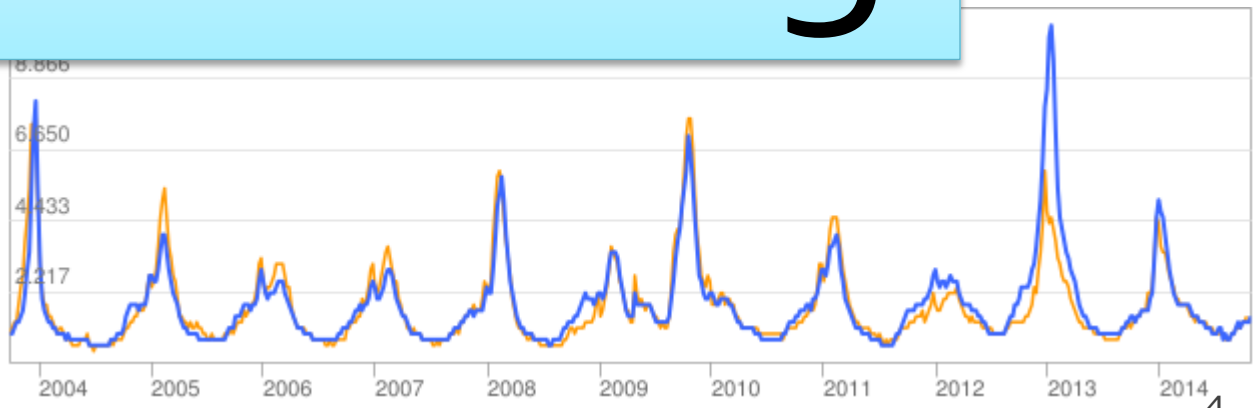


Google Flu trends



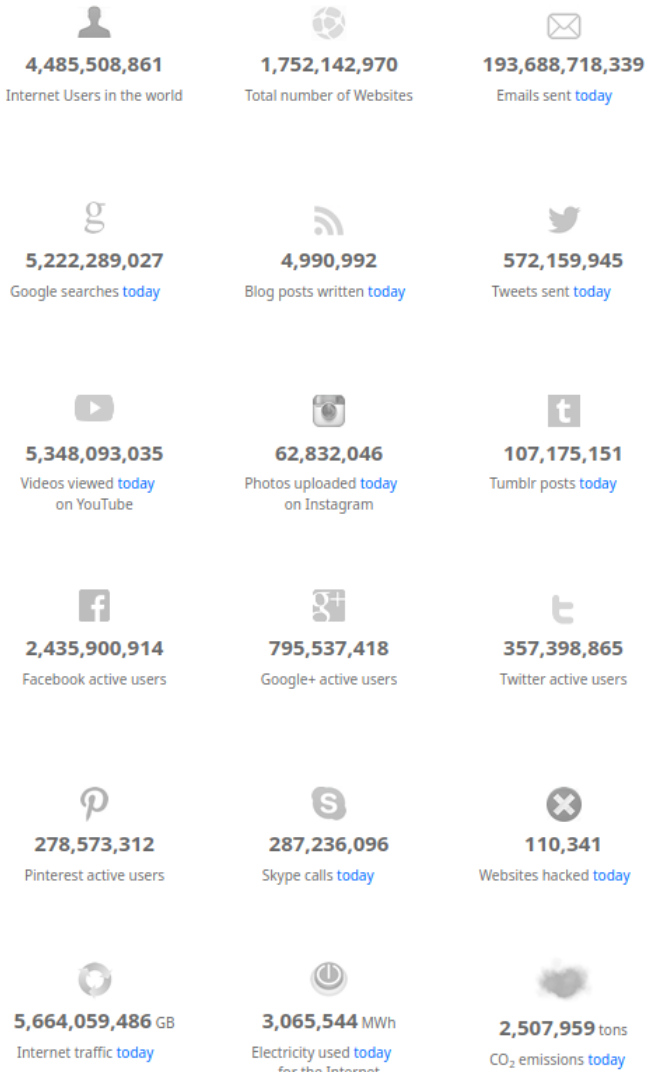
- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and
    .A)
    alysis of
    ueries

Nowcasting



Data on the Internet…

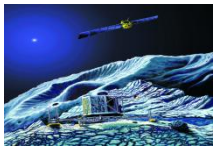
- Internet live stats
  - http://www.internetlivestats.com/



Who generates big data?

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

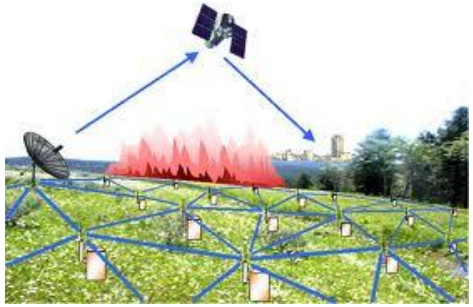
- Health and scientific computing

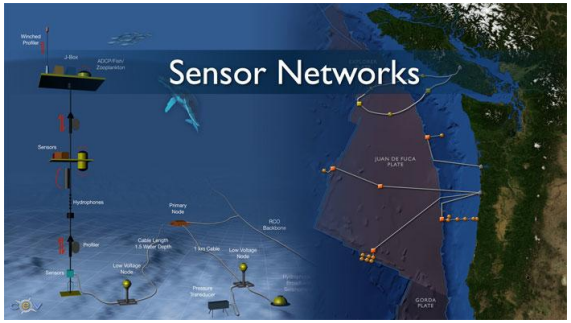


Who generates big data?

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

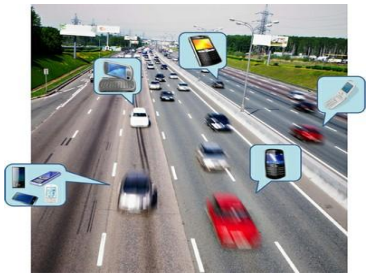
- Internet Of Things (IoT)
  - Sensor networks, RFIDs, smart meters





An example of Big data at work

Crowdsourcing



Sensing

Map data

Computing

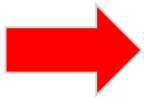
Real time traffic info

Travel time forecast/nowcast

The Vs of big data

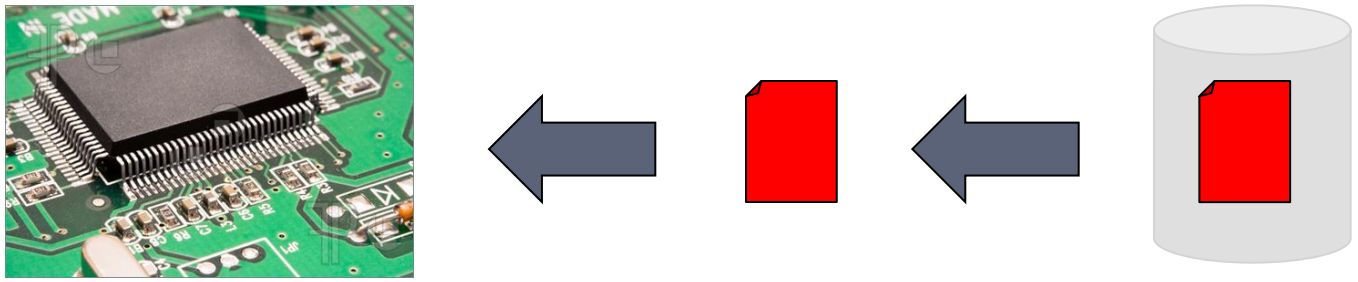
- The 3Vs of big data
  - Volume: scale of data
  - Variety: different forms of data
  - Velocity: analysis of streaming data
- … but also
  - Veracity: uncertainty of data
  - Value: exploit information provided by data

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