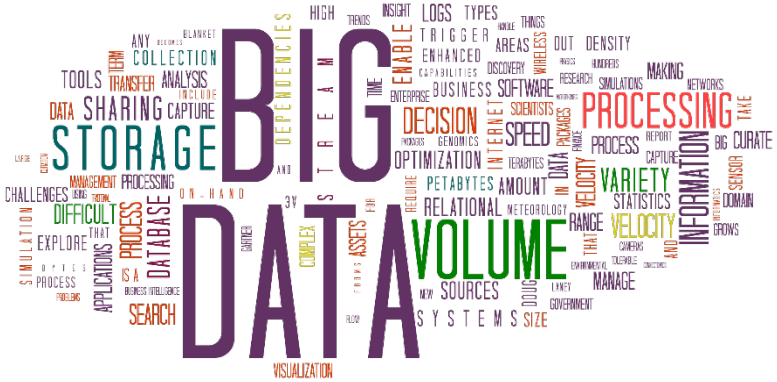
Basi Di Dati e di conoscenza

Basi di Dati e Big Data

Contenuti della lezione

- Introduzione ai Big Data
- Definizione dei Big Data
- Le 3V+2V dei Big Data
- Architetture e problematiche
- Memorizzare e gestire I Big Data
  - Hadoop & Map-reduce
  - Cloud computing
  - NoSQL DBMS

Big Data


BIG
DATA
STORAGE
PROCESSING
VOLUME
VARIETY
VELOCITY
INFORMATION
DECISION
SHARING
DATABASE
COLLECTION
ANALYSIS
CHALLENGES
DIFFICULT
EXPLORE
SEARCH
VISUALIZATION
STREAM
DEPENDENCIES
ENABLE
LOGS
TYPES
TRIGGER
ENHANCED
AREAS
DENSITY
BUSINESS
SOFTWARE
SPEED
OPTIMIZATION
PETABYTES
AMOUNT
RELATIONAL
SOURCES
SYSTEMS
SIZE
MANAGE
STATISTICS
CURATE
SENSOR
TOOLS
TRANSFER
CAPTURE
PROCESS
APPLICATIONS
SIMULATION
ASSETS
COMPLEX
RANGE
MAKING
INTERNET
SCIENTISTS
PROCESS
CAPTURE
GROWS
DOMAIN


BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

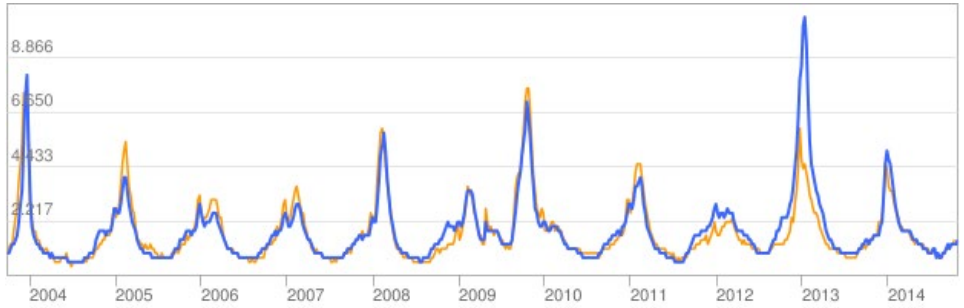
Goals

- Show state-of-the-art techniques for dealing with collections of **unstructured data whose size exceeds the capacity of storage, management, and analysis typical for traditional (relational) database systems**
- In particular:
  - Requirements for modern applications
  - Problems with big data
  - Available hardware/software solutions

Esempi di applicazioni

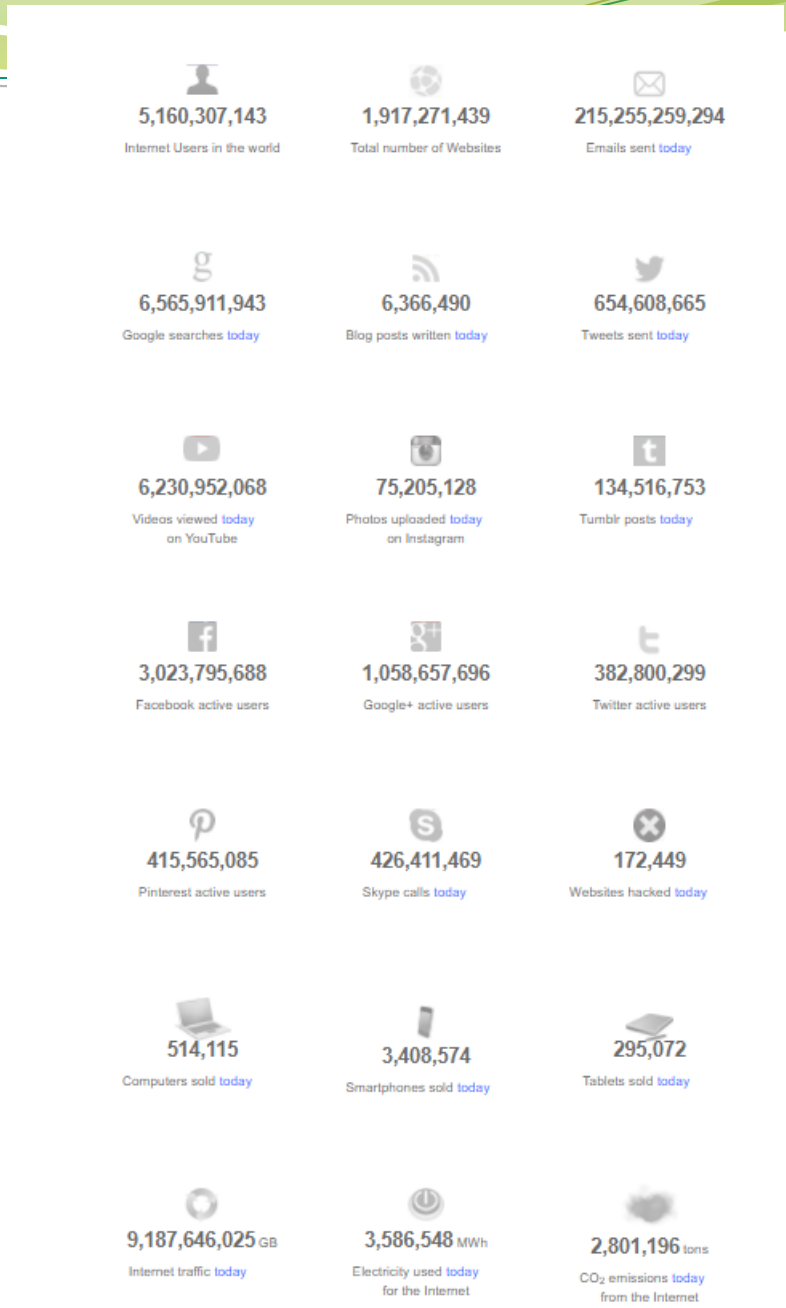
Google Flu trends

- 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**



Data on the Internet

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



Cosa Sono i big Data?

Big Data is any thing which is crash Excel.

Small Data is when is fit in RAM.
Big Data is when is crash because is not fit in RAM.

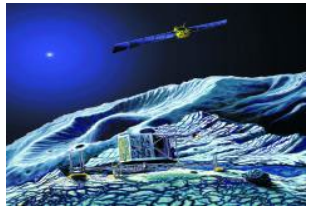
Or, in other words, Big Data is data in volumes too great to process by traditional methods.

https://twitter.com/devops_borat

Who generates big data

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

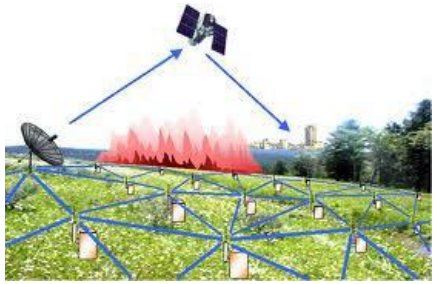
- Health and scientific computing

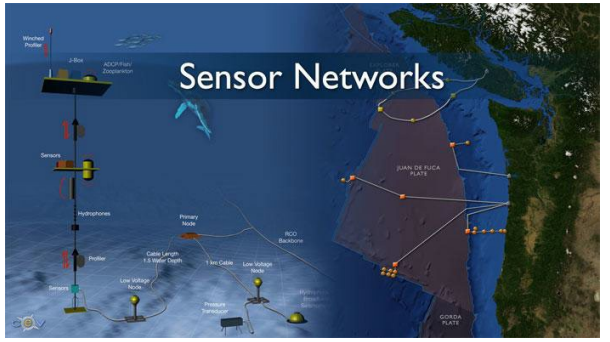


Who generates big data

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

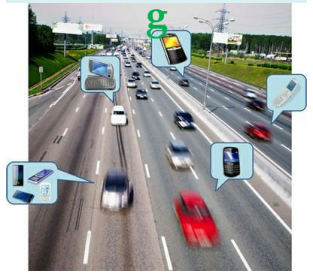
- Internet Of Things (IoT)
  - Sensor networks, RFIDs, smart meters





An example of Big Data at work

Crowdsourcing



Sensing

Map Data

Computing

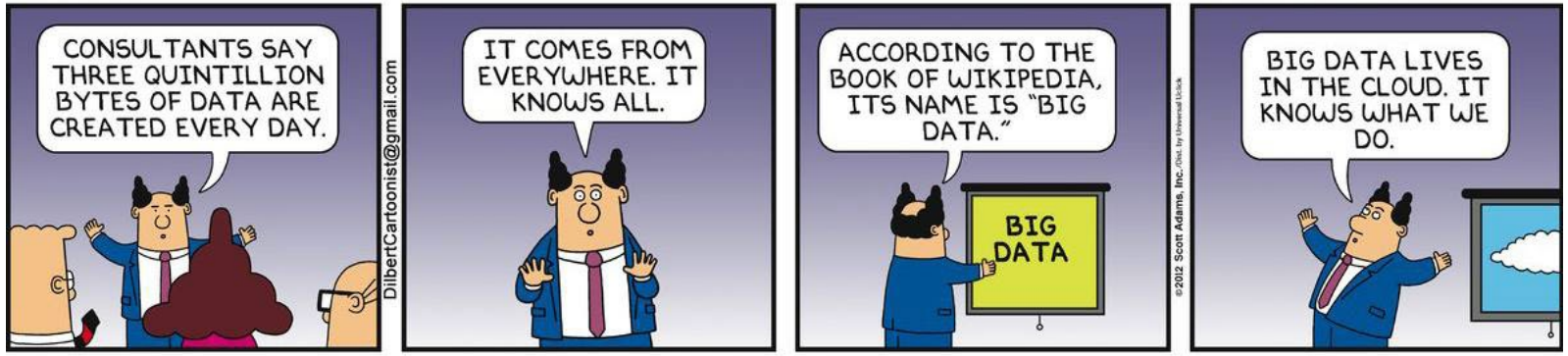
- Real time traffic info
- Travel time forecast/nowcast

Big Data? Different definitions!

- “Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population.” (Teradata Magazine article, 2011)
- “Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.” (The McKinsey Global Institute, 2012)
- “Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate.” (Wikipedia, 2016)

What is big data?

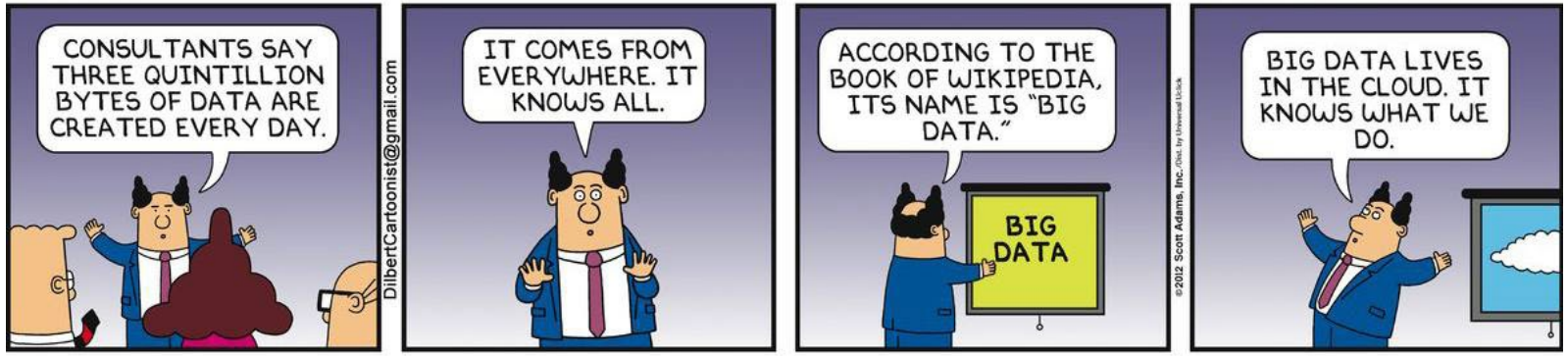
Many different definitions



"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?

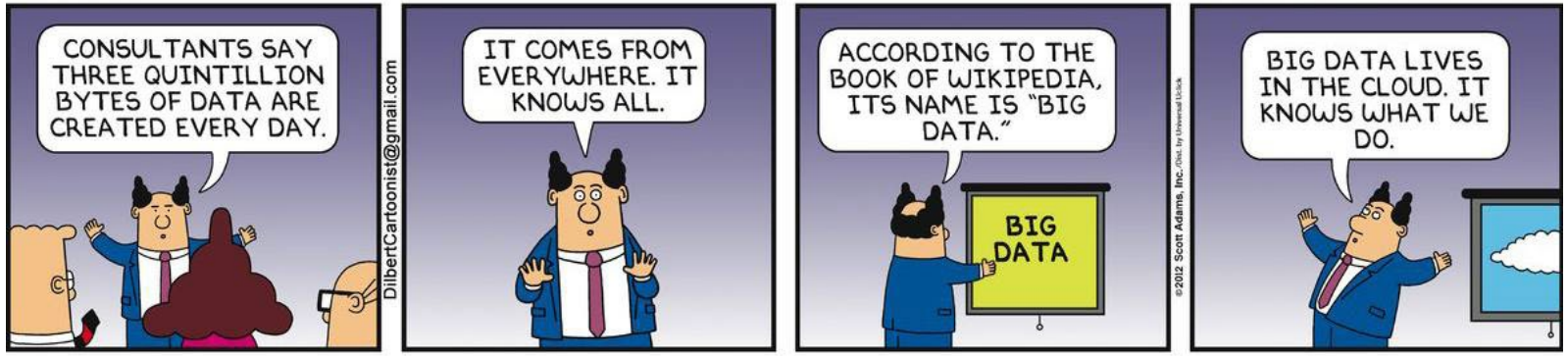
Many different definitions



"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



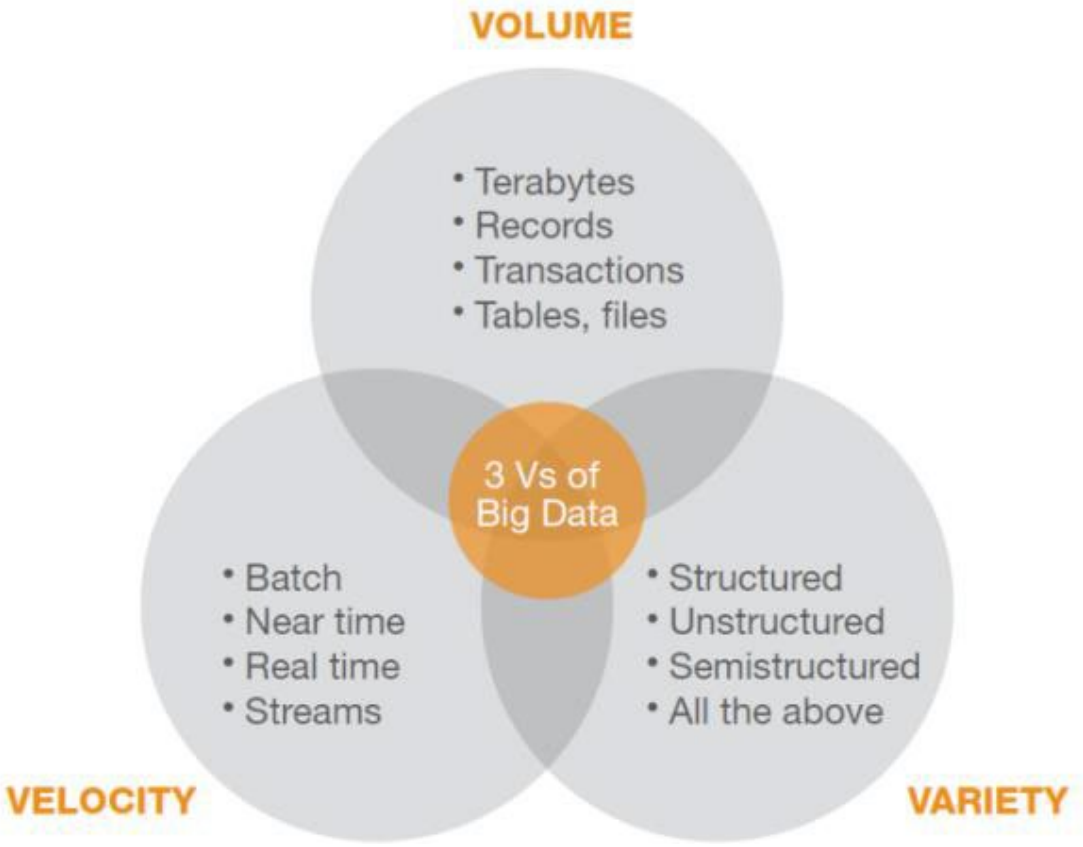
"Data whose scale, diversity and complexity require new **architectures**, **techniques**, **algorithms** and **analytics** to manage it and extract value and hidden knowledge from it"

Some numbers

- **How many data in the world?**
  - 800 Terabytes, **2000**
  - 160 Exabytes, **2006** (1EB = 1018B)
  - 500 Exabytes, **2009**
  - 2.7 Zettabytes, **2012** (1ZB = 1021B)
  - 35 Zettabytes by **2020**
- **How much is a zettabyte?**
  - 1,000,000,000,000,000,000,000 bytes
  - A stack of 1TB hard disks that is 25,400 km high
- **How many data in a day?**
  - 7 TB, Twitter
  - 10 TB, Facebook
- **90% of world's data:**
  - generated over last two years

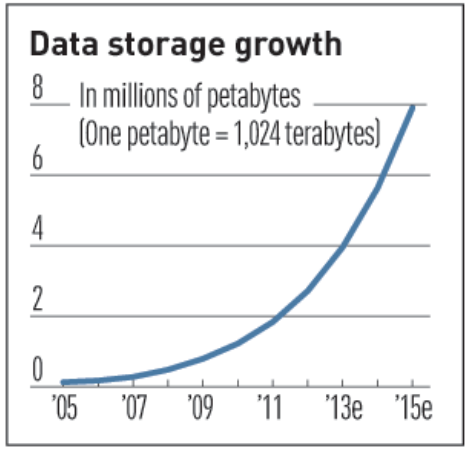
The Vs of big data: From WWW to VVV

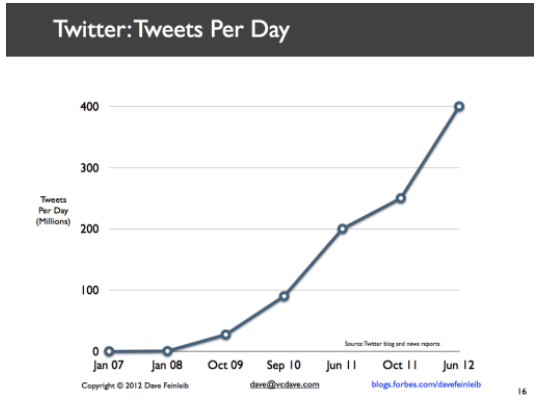
- The 3Vs of big data
  - Volume: **scale of data**
    - data volumes are becoming unmanageable
  - **Variety**: **different forms of data**
    - data complexity is growing
    - more types of data captured than previously
  - **Velocity**: **analysis of streaming data**
    - some data is arriving so rapidly that it must either be processed instantly, or lost
    - this is a whole subfield called "**stream processing**"
- … but also
  - **Veracity**: uncertainty of data
  - **Value**: exploit information provided by data

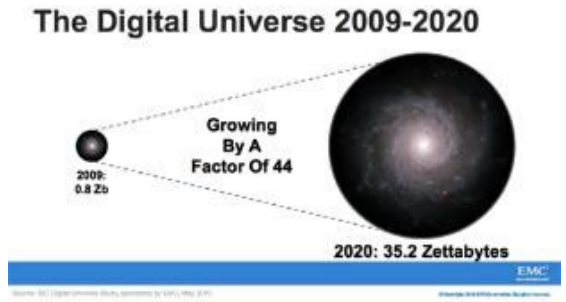


The Vs of big data: Volume

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

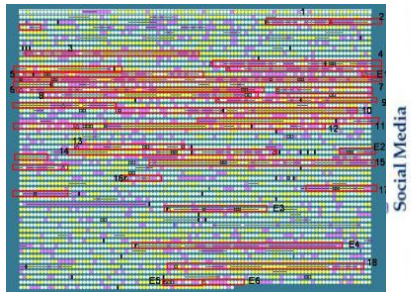


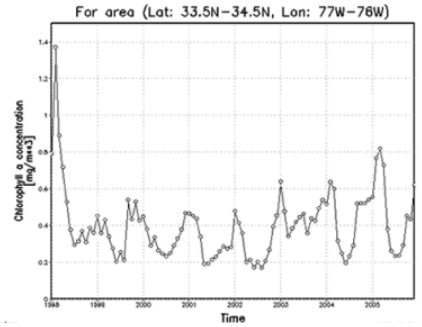


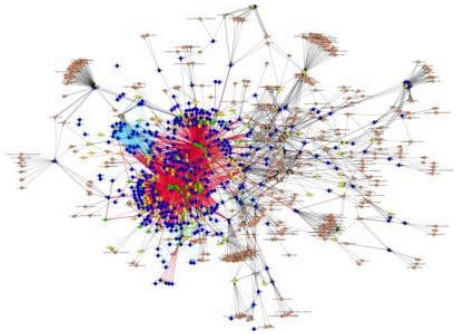


The Vs of big data: Variety

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



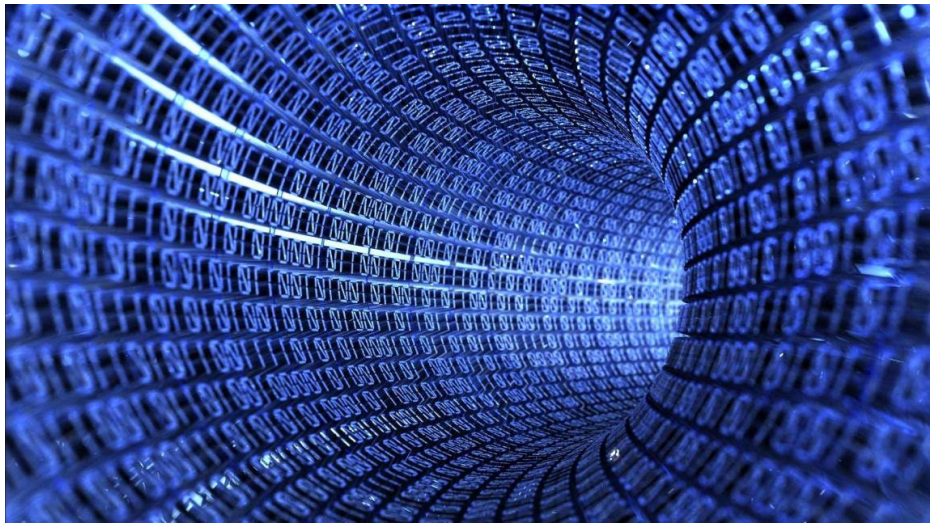




  - A single application may generate many different formats
    - Heterogeneous data
    - Complex data integration problem

The Vs of big data: Velocity

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

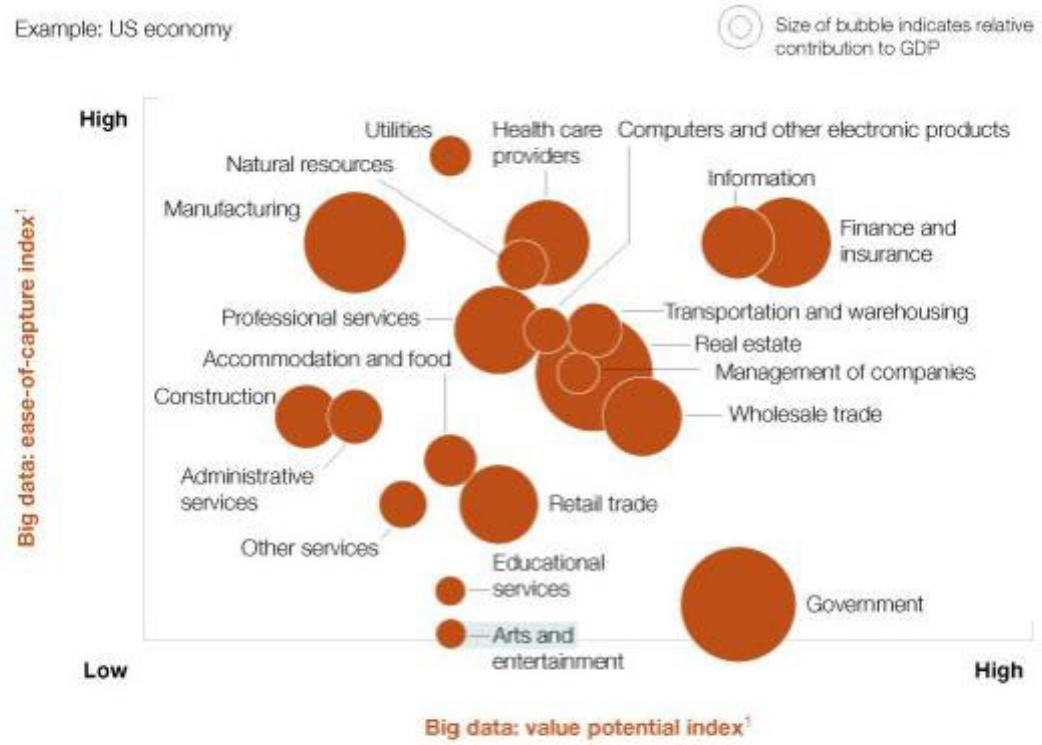


The Vs of big data: Veracity

- **Veracity**
  - Data quality

The Vs of big data: Value

- **Value**
  - they can generate huge competitive advantages!
  - 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.

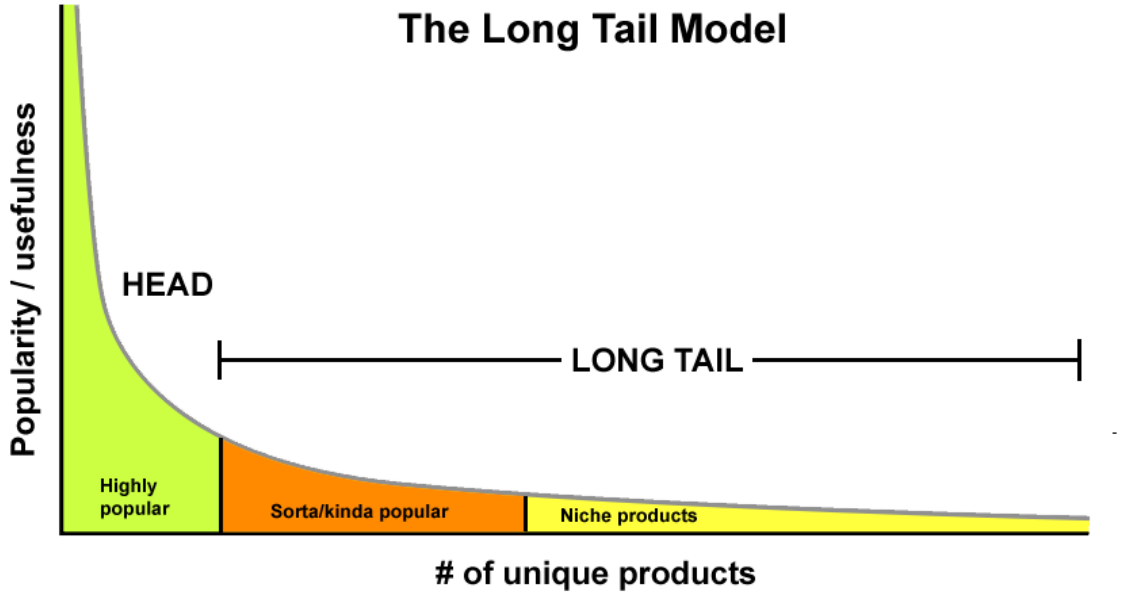
Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

The promise of Big Data

- Data contains information of great business value
- If you can extract those insights, you can make far better decisions
- ...but is data really that valuable?

Bigger=Smarter?

- Yes!
  - tolerate errors
  - discover the "long tail" and "corner cases"
  - algorithms work much better
- BUT:
  - more heterogeneity
  - data grows faster than energy on chip
  - still need humans to ask right questions



"We sold more books today that didn't sell at all yesterday than we sold today of all the books that did sell yesterday" (Amazon employee)

nature International weekly journal of science

NATURE | NEWS

Drug data reveal sneaky side effects

Mining of surveillance data highlights thousands of previously unknown consequences when drugs are taken together.

Heidi Ledford

14 March 2012

An algorithm designed by US scientists to trawl through a plethora of drug interactions has yielded thousands of previously unknown side effects caused by taking drugs in combination.

The work, published today in *Science Translational Medicine*[1], provides a way to sort through the hundreds of thousands of 'adverse events' reported to the US Food and Drug Administration (FDA) each year. "It's a step in the direction of a complete catalogue of drug–drug interactions," says the study's lead author, Russ Altman, a bioengineer at Stanford University in California.

Although clinical trials are often designed to assess the safety of a drug in addition to how well it works, the size of the trials needed to detect the full range of drug interactions would surpass even the large, late-stage clinical trials sometimes required for drug approval. Furthermore, clinical trials are often done in controlled settings, using carefully defined criteria to determine which patients are eligible for enrolment — including other conditions they might have and which medicines they can take alongside the trial drug.

A program predicts the potential side-effects of mixing different pills.

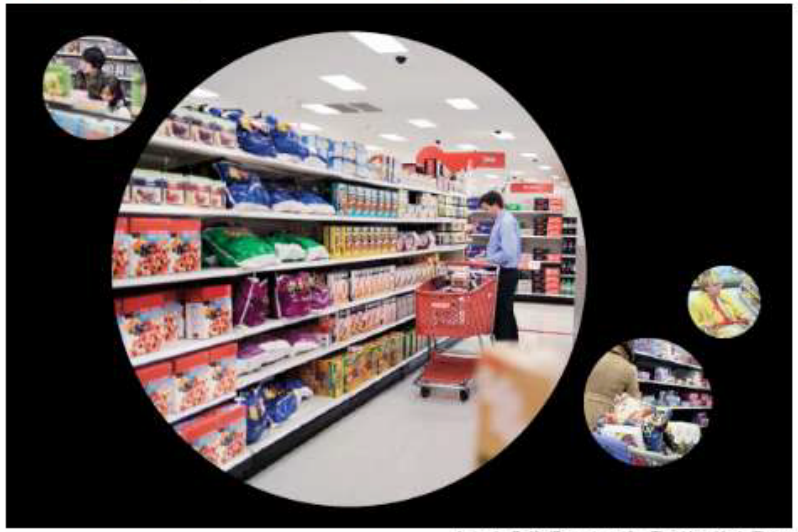
DWIMAGES/ALAMY

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | Global Edition

The New York Times | International Herald Tribune — GLOBAL EDITION Magazine

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SP

How Companies Learn Your Secrets



Antonio Bolfo/Reportage for The New York Times

By CHARLES DUHIGG
Published: February 16, 2012 | 570 Comments

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "

FACEBOOK
TWITTER
GOOGLE+
E-MAIL
SHARE
PRINT

About a year after Pole created his pregnancy-prediction model, a man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter, and he was angry, according to an employee who participated in the conversation.

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.

On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

Some more examples

- Sports
  - basketball increasingly driven by data analytics
  - soccer beginning to follow
- Entertainment
  - House of Cards designed based on data analysis
  - increasing use of similar tools in Hollywood
- "Visa Says Big Data Identifies Billions of Dollars in Fraud"
  - new Big Data analytics platform on Hadoop
- "Facebook is about to launch Big Data play"
  - starting to connect Facebook with real life

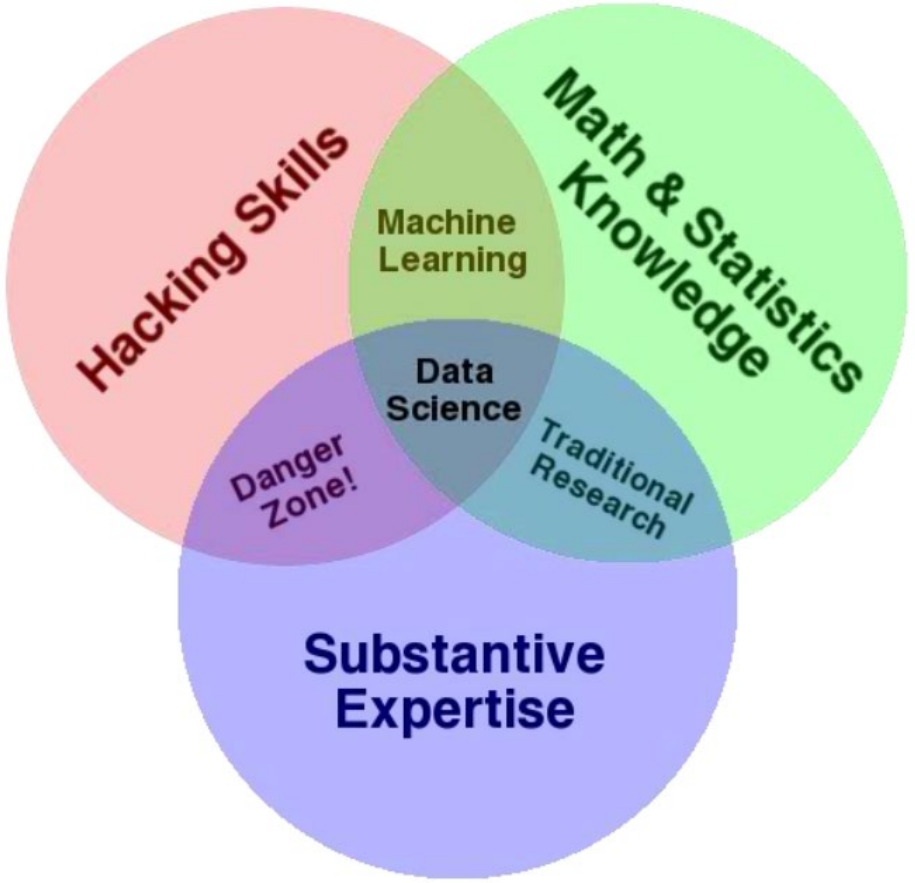
The risks of Big Data

- Data grows faster than energy on chip
  - Efficiency
  - Effectiveness
  - Scalability
- Costs
- Privacy

Le sfide dei big Data

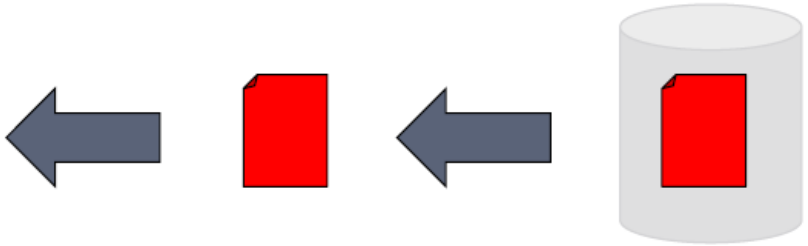
- **Technology and infrastructure**
  - New architectures, programming paradigms and techniques are needed
  - Data management and analysis
- **New emphasis on "data"**
  - **Data science**

Data Science



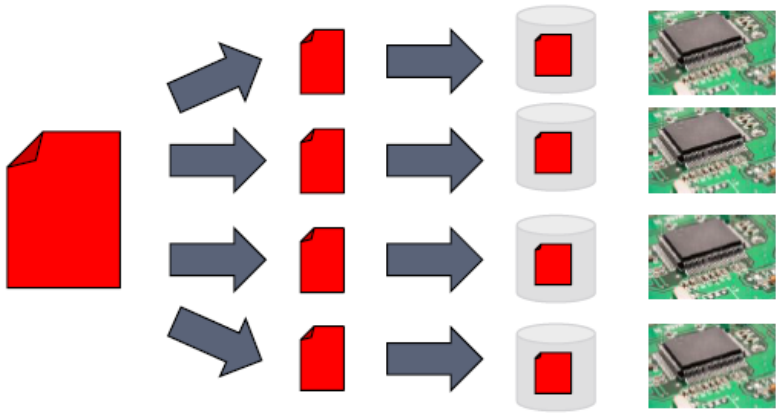
Collo di bottiglia

- 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



Data quality

- "The greatest enemy of knowledge is not ignorance; it is the illusion of knowledge."

  Daniel Borstin, in The Discoverers (1983)

95% of time, when is clean Big Data is getting Little Data

Data quality

- "A huge problem in practice
  - any manually entered data is suspect
  - most data sets are in practice deeply problematic
- Even automatically gathered data can be a problem
  - systematic problems with sensors
  - errors causing data loss
  - incorrect metadata about the sensor
- **Never, never, never trust the data without checking it!**
  - garbage in, garbage out, etc