Overview of state of art in Data management

Stefano Cozzini

CNR/IOM and eXact lab srl

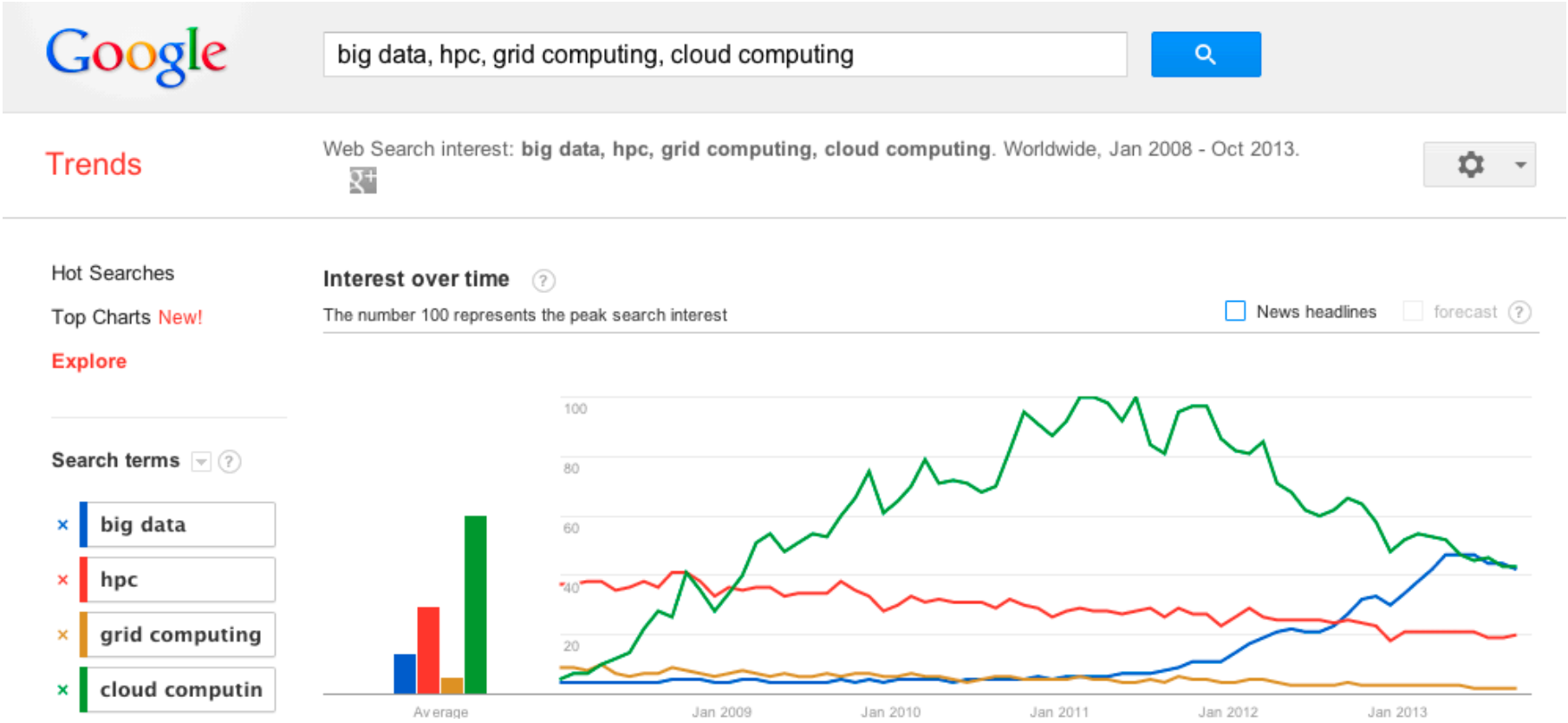
AIM of this short talk

- Frame the problem and the discussion around DATA:
  - What are big data ?
  - Which kind of challenges ahead of us ?

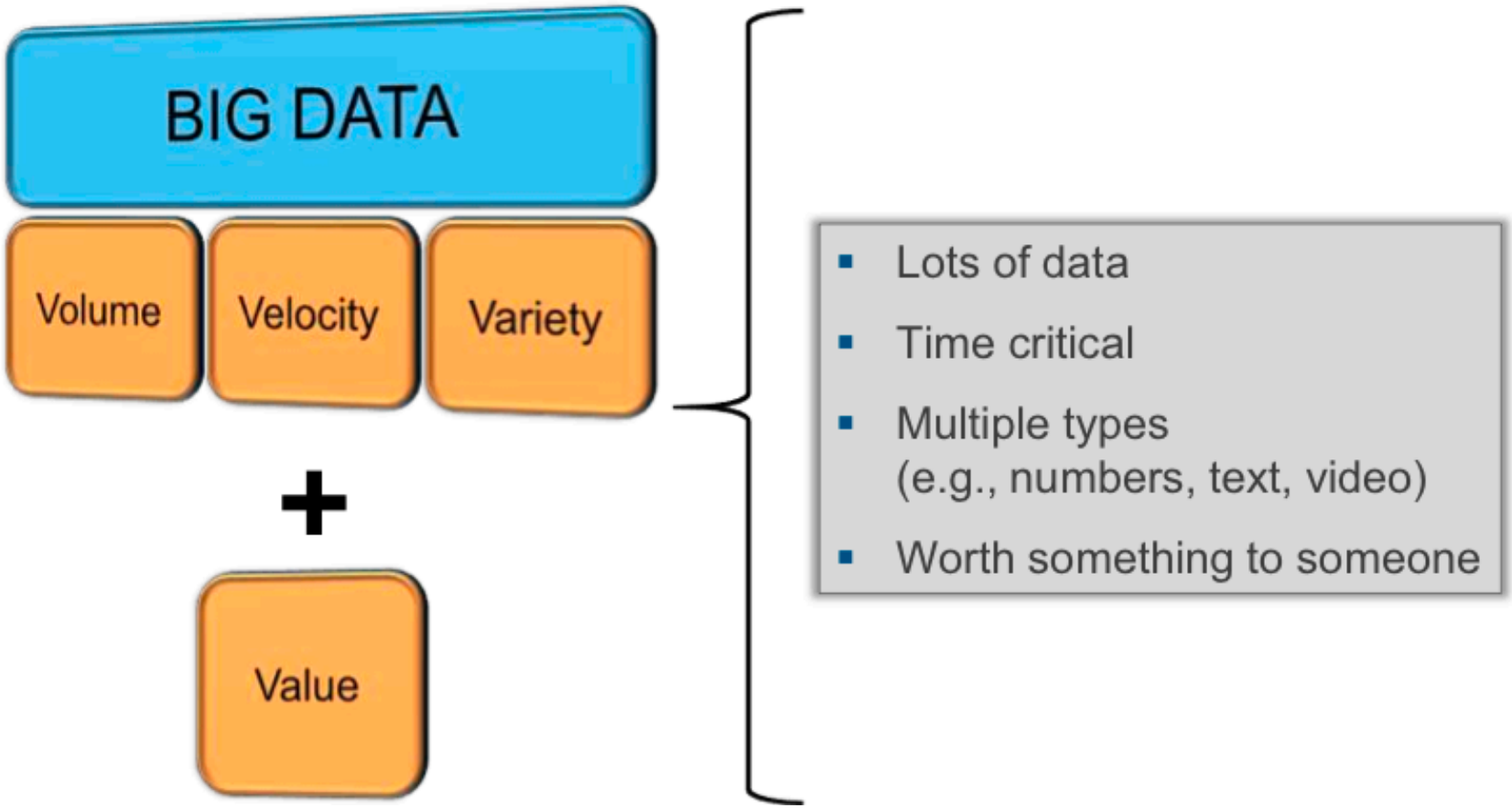
Disclaimer:

Slides and numbers are taken around: there are a lot of data discussing big data ☺

Big Data: a buzzword..



Big Data: general definition



The 3 V's of big data..

- Velocity

  Data are produced at speed higher than the speed you are able to move/analyze and understand them..

- Variety
  - Data range from simulation to remote sensing information, from instruments to market analysis etc..
  - datasets come in a variety of data formats and span a variety of metadata standards
- Volume
  - The amount of data will increase of factor 61 in the next 10 years
  - The amount of data is estimated to exceed the size of available data infrastructure to store them by 60%.

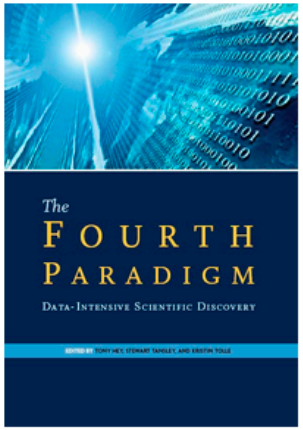
[from The 2011 IDC Digital Universe Study]

Data-intensive science

- A "fourth paradigm" after experiment, theory, and computation..

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science



Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Critical praise for *The Fourth Paradigm*

Download

- Full text, low resolution (6 MB)
- Full text, high resolution (93 MB)
- By chapter and essay

Purchase from Amazon.com

- Paperback
- Kindle version

In the news

- Sailing on an Ocean of 0s and 1s *(Science* Magazine)
- A Deluge of Data Shapes a New Era in Computing *(New York Times)*
- A Guide to the Day of Big Data *(Nature)*

It involves collecting, exploring, visualizing, combining, subsetting, analyzing, and using huge data collections

Challenges & Requirements

Challenges:

- Deluge of observational data, "exaflood" of simulation model outputs
- Need for collaboration among groups, disciplines, communities
- Finding insights and discoveries in a "Sea of Data"

Requirements:

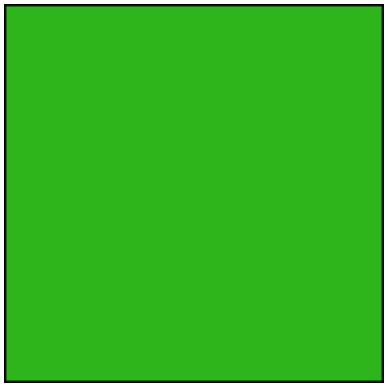
- New tools, techniques, and infrastructure
- Standards for interoperability
- Institutional support for data stewardship, curation

IOPS vs FLOPS

- HPC is today *compute-centric*
- scientific computing needs data accessibility rather than computing speed

computing 1 calculation
≈ 1 picojoule

moving 1 calculation
≈ 100 picojoule



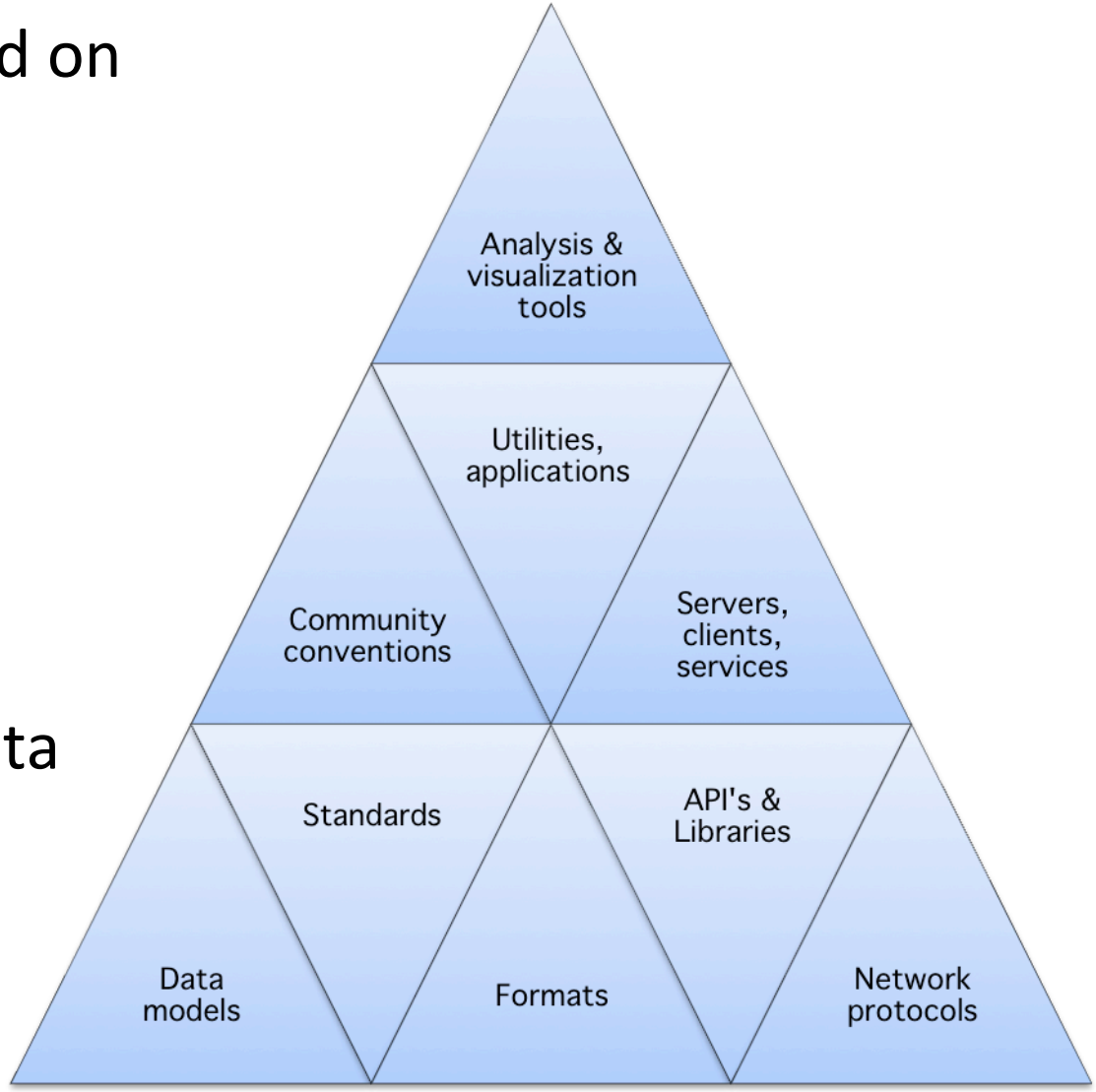
Source: IDC Direction 2013

Roles in Data-intensive Science

- **Scientists/researchers:** acquire, generate, analyze, check, organize, format, document, share, publish research data
- **Data users:** access, understand, integrate, visualize, analyze, subset, and combine data
- **Data scientists:** develop infrastructure, standards, conventions, frameworks, data models, Web-based technologies
- **Software developers:** develop tools, formats, interfaces, libraries, services
- **Data curators:** preserve data content and integrity of science data and metadata in archives
- **Research funding agencies, professional societies, governments:** encourage free and open access to research data, advocate elimination of most access restrictions

Infrastructure for sharing scientific Data

- Applications depend on lower layers
- Sharing requires agreements
  - formats
  - protocols
  - conventions
- Data needs metadata



Data infrastructure crucial aspects

- Large datasets:
  - Deal with large datasets and large data rates from experiments
- Reliability :
  - Increase the level of QoS and SLA, e.g. increasing reliability by replicating data sources and increasing accessibility by copying source to several places
- Accounting
  - Allow monitoring and checking of resource usage Integration Provide the same set of services that are understandable (compatible) between domains
- Interoperation
  - Interoperation through common standard schemes
- Access
  - Broadband data access Allow transparent and secure remote access to data
- Data preservation
  - Allow long-term availability of data
- High quality
  - Quality of data to enable advanced and cross-disciplinary access and enrichment operations
- Economic justification
  - As the scientific community is operating on increasingly larger datasets and want to preserve the information concerned, the infrastructure provided should have a clear roadmap of technology exchange and backwards compatibility.
- Access control
  - Provide the infrastructure to allow for fine-grained access control

Source: e-irg blue paper on data management 2012

http://www.e-irg.eu/images/stories/dissemination/e-irg-blue_paper_on_data_management_v_final.pdf

Data are not just for science..

- High-end commercial analytics pushing up into academia.
- The journey from science data to industry& commerce can be relatively short…
- ..and plenty of ethical legal and societal implications

IEEE SECURITY & PRIVACY

Big Data, Big Brother, Big Money

July-Aug. 2013 (vol. 11 no. 4)

pp. 85-89

Michael Lesk, Rutgers University

DOI Bookmark: http://doi.ieeecomputersociety.org/10.1109/MSP.2013.81

ABSTRACT

Government snooping, recently publicized, is now using the same data sources that corporations use to watch us. The same records used by federal agencies to search for terrorists are used, with fewer controls, by corporations searching for customers.

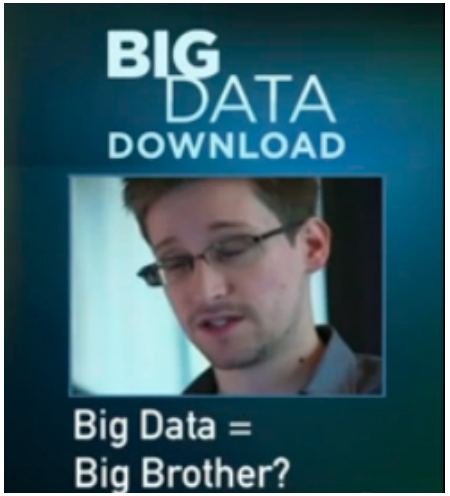
ADDITIONAL INFORMATION

Index Terms:

Surveillance,Government agencies,Surveillance,Data mining,Marketing and sales,Analytical models,marketing analytics,surveillance,data mining

Citation:

Michael Lesk, "Big Data, Big Brother, Big Money," *IEEE Security & Privacy*, vol. 11, no. 4, pp. 85-89, July-Aug. 2013, doi:10.1109/MSP.2013.81



August 06, 2013

Big Data Meets Big Brother: Inside the Utah Data Center

Thomas Parent

The widely publicized leaks from the National Security Agency (NSA) have provided a fascinating glimpse into the covert data collection activities of the U.S. Government. Among other things, these leaks have revealed the enormous volume of phone records currently being collected, stored and analyzed.

This begs the question: "Where does all of this Big Data go?" Starting this fall, it's likely that a good chunk of it will be going to Utah!

While the NSA operates many data facilities, none compare to the new $1.5 Billion data center scheduled to open this September in Bluffdale, Utah. Vanee Vines, an NSA spokeswoman, has provided the following information: