Big Data – a threat or a chance?

Helwig Hauser
University of Bergen, Dept. of Informatics

Big Data

What is “Big Data”?  
– well, lots of data, right? ... we come back to this in a moment.
– certainly, a buzz-word... ... but a relevant one!

Examples
– big data from numerous sensors (Internet of Things, ...)
– bid data in large social networks (Facebook, Twitter, ...)

Broadly used definition
– 3V-def.: “Big data” is high-volume, -velocity & -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [Doug Laney, 2001 / Gartner]
Certainly, Big Data (usually) refers to lots of data!

“Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

[McKinsey Global Institute 2011]

Available data grows exponentially

- Exabytes of data available world-wide
  - 1 EB = 1000 PB = 1 million TB = 1 billion GB
  - hundreds of EB transferred via the Internet, annually
  - EB of new information stored, annually

Big Data, V#2: Variety

Big Data beyond numbers

- text, images & sound, relational data, ...
  - unstructured data
- 30 billion pieces of information on Facebook per month!
  - 400 million tweets per day
  - 4 billion hours of videos are watched on YouTube / month
  - >400 million wearable, wireless health monitors
- Daniel Keim, 2007: 100 million FedEx transactions per day,
  - 150 million VISA credit card transactionen per day,
  - 300 million long distance calls in ATT’s network per day,
  - 50 billion e-mails worldwide per day,
  - 600 billion IP packets per day DE-CIX backbone

Dark Data: available, but unused data
Big Data, V#3: Velocity

Real-time Big Data / Streaming Data Analysis, but also

– rapidly changing data
– data at different speeds and uneven rates (bursts)

Big Data – a moving target!

– lots of generated information cannot be stored!
  • 90% of health care data is discarded (videos, etc.)

Big Data, V#4(?): Veracity [IBM, ...]

Uncertain / low-quality data

– >$3 trillion loss to US economy due to bad data quality
– high degree of uncertainty

D. Laney blogs:

– Batman on Big Data:

Even more Vs: [K. Normandeau]

– validity: the right data for the right decisions?
– volatility: when valid, storing for how long, etc.?
Big Data in Practice

Big data is

- generated, aggregated, analyzed, and consumed
- sensed, collected (networks), stored (cloud), and analyzed (machine learning)
- process-mediated ("nicer" data), machine-generated (Internet of Things), human-sourced (from messages to videos)

2013 Corporate Culture Distribution by Geographic Region

Big Data Technology – selection

Conceptual

- MapReduce [Google, 2004]
  - **map**: distribution of queries to many nodes
  - **reduce**: gathering of results and delivery
- NoSQL ("not only SQL"), for ex. Cassandra (key-value)

Software

- Hadoop [Apache], MongoDB

Analytics Technologies

- A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualisation [McKinsey, 2011]
Big Data and Visual Analytics

Visual Analytics

– Illuminating the Path book: 2005
– VisMaster book: 2010

Big Data – Challenges & Opportunities

Selected Challenges

– shortage of Big Data talent (up to 200,000 needed in the US plus 1.5 million «data-savvy» managers)
– contextualization of Big Data – Big Data needs to be complimented by Big Judgment [Harvard Business Review]
– prediction difficult without theory

Selected Opportunities

– annually $300 billion to the US health care system, incl. cost savings up to 8%
– annually $250 billion to the European public sector adm.
– job opportunity (analysts, managers, et al.)!
Big Data in Business

Five opportunities according to McKinsey GI, 2011:

– reduced searching & processing time, e.g., in the public administration sector, as well as concurrent engineering in manufacturing due to accessible Big Data
– enabling experimentation to discover needs, expose variability, and improve performance
– segmenting populations to customize actions
– replacing/supporting human decision making with automated algorithms based on Big Data Analytics
– innovating new business models, products, and services

Active enterprises include:
– eBay, Amazon, Walmart, Facebook, in finance, real estate, ...

Big Data and Privacy Concerns

Snowden informed about NSA...

As data get large, networked, reside in the cloud, we fear
– unauthorized access
– data misuse
– identity theft

Examples:
– leaked health data
– credit card fraud
– monitored privacy
But let’s talk about science a little..

Big Data and the Fourth Paradigm

2009, Microsoft: the 4th paradigm: data-intensive scientific discovery

– refers to the last talk by Jim Gray, 2007, “A Transformed Scientific Method”

– from empirical (initially), via theoretical (modern times), and computational science (last decades) to data-intense science (now)

– eScience: capture, curation, analysis, vis.

– needle-in-a-haystack problems comparably “easy” (Higgs)

– more difficult: trends, clusters, patterns (N², or more)
Big Data in Science

Sources of Big Data
– meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research
– mobile phones, remote sensing, logs, cameras & microphones, RFID sensors & sensor networks

Big Science Examples
– The Large Hadron Collider experiments:
  • about 150 million sensors
  • delivering about 40 millions times per second (!!!)
– Sloan Digital Sky Survey (since 2000)
  • more data in a few weeks than all of astronomy so far
  • about 200 GB per night, now >140TB of data

Big Data in Medicine

P4 medicine [Leroy Hood]
– predictive, preventive, personalized, and participatory

Computational Medicine [Arvid Lundervold, 2014]
– embracing IT, bioinformatics, etc., for “systems medicine”

Examples:
– predictive medicine
– large-scale cohort studies

Case: [EMA 2013 Operationalizing the Buzz]
– Brigham and Women’s Hospital: improved drug risk awareness due to Big Data (much fast results)
Big Problems with Small Data

Christian Chabot (CEO of Tableau), 2008:

Who can Visual Analytics help?

**Everybody with data that is not getting answers**

Conclusions

Big Data is maturing, it’s unavoidable

EMA 2013: the next Big Data challenge: Ethics!

Big Data is transforming Science (4th paradigm, etc.)
  – Chris Anderson, Wired, 2008: The End of Theory

New opportunities, new challenges
  – big business, P4 medicine
  – “the other” Vs, dark data
  – how to turn data into knowledge?
  – technological challenges, new ways of thinking
  – it’s – not at the least – also an educational challenge!
Acknowledgements

You! 😊

Questions?

Stefan Bruckner
Arvid Lundervold
Lots of references...