



JACOBS
UNIVERSITY

Big Data

Instructors: Peter Baumann

email: p.baumann@jacobs-university.de

tel: -3178

office: room 60, Research 1

- „It is estimated that a week's work at the **New York Times** contains more information than a person in the 18th Century would encounter in their entire lifetime and the thought is that within 10 years the rate of information doubling will occur every 72 hours.“
 - -- P. „Bud“ Peterson, U Colorado
- “global **mobile data traffic** 597 petabytes per month in 2011
 - 8x the size of the entire global Internet in 2000
 - estimated to grow to 6,254 petabytes per month by 2015”
 - -- Forbes, June 2012

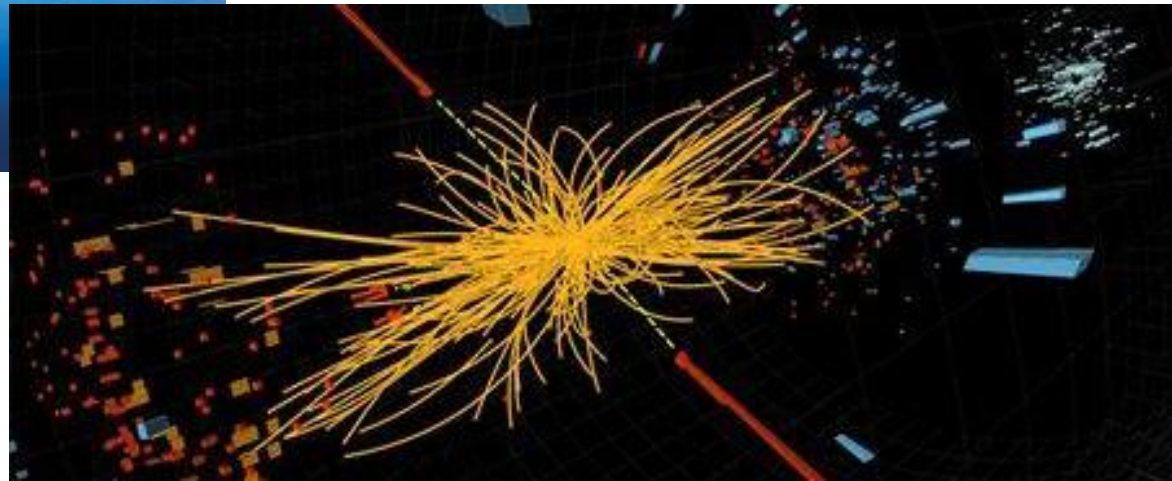
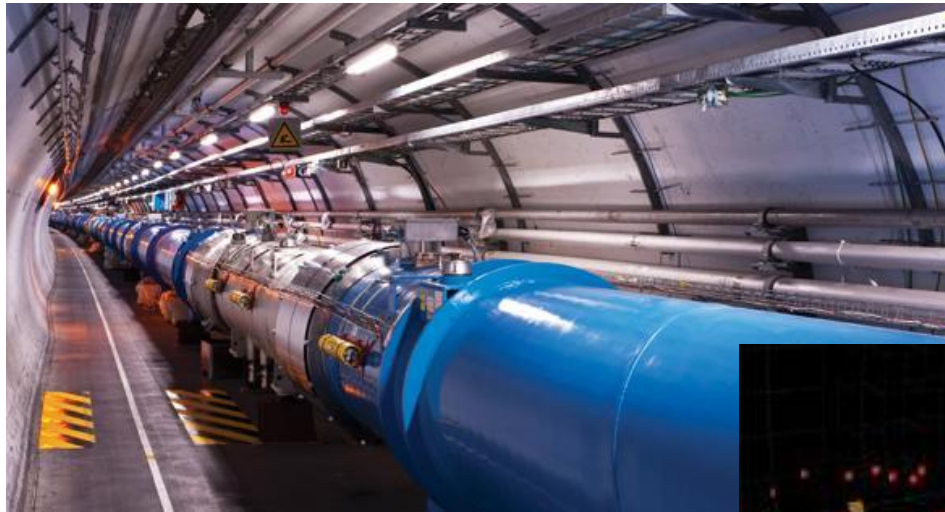
- Internet: the unprecedented information collector
 - May 2012: 200m Web servers [Yahoo]
 - estd 50+b static pages [Yahoo]
 - 2012: 31b searches / month [Google]
 - Wayback Machine: 240 billion web pages archived from 1996
- Typical Big Data:
 - Social networks - facebook, twitter, GPS, ...
 - Business: Data Warehousing
 - Geo: Satellite imagery, weather data, ...
 - Petrol industry:
„more bytes than barrels“
 - ...plus „Deep Web“



<http://www.sgi.com/go/twitter/#heatmaps>

Big Data in High Energy Physics

- CERN, Large Hadron Collider:
13 PB in 2010

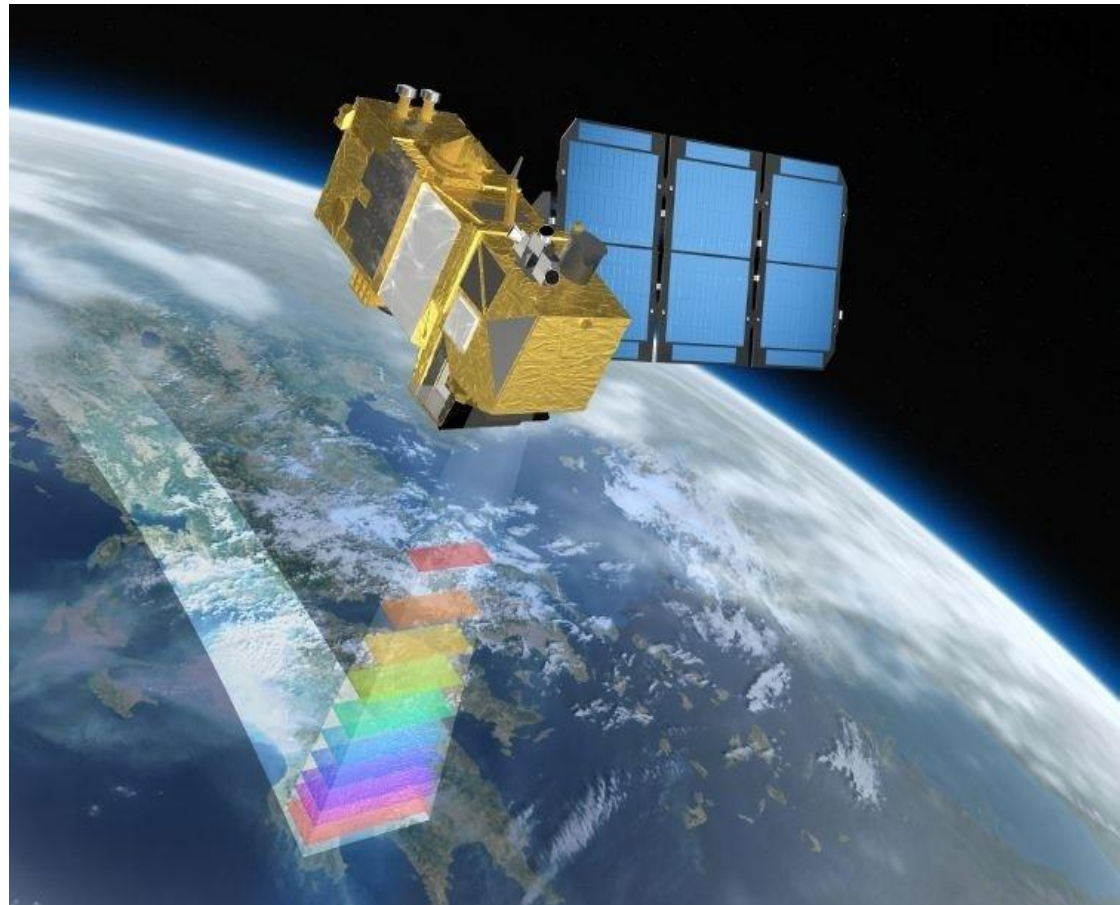


[CERN]

- Data aggregation & integration → cost effective and improved patient care
 - biological & biomedical research: next-generation sequencing (TB of raw data)
 - How to store, achieve, index, manage, learn, mine, visualize those data?
- 23andme.com: „Discover your ancestral origins and lineage with a personalized analysis of your DNA”
 - “Learn what percent of your DNA is from populations around the world.”
 - “I understand that 23andMe only sells ancestry reports and raw genetic data at this time. I understand 23andMe will not provide health-related reports. However, 23andMe may provide health-related results in the future, dependent upon FDA marketing authorization. “ [23andme.com, 2013-12-15]

Big Data in Earth Observation

- „Exaflood“: 100s of Exabytes in 2020 expected [Climate WS 2011]
 - Spectral bands:
from 5 (Landsat)
to 250 (ALI/Hyperion)
 - Resolution: few meters
- Sentinel-2 (ESA):
2.4 TB / d \rightarrow 876 TB / y
 - 10-20m ground resolution
 - 13 bands
 - One of 5 Sentinels



Big Data in Astronomy: LOFAR

- **Sloan Digital Sky Survey**: first few weeks in 2000, more data than all collected in history of astronomy
 - 200 GB per night, 140+ TB now
- **LOFAR** (Low frequency phase-coupled array)
 - Distributed radio telescope
 - Processing output <50 gbps (0.5 PB/d)
 - Long term: 2.5 PB/y
- Analytics also on Long-Term Archive
 - Ex: 10,000 x 10,000 FFT



[W. Reich, MPIfR]



[Wikipedia]

- business data worldwide, across all companies, double every 1.2 years, according to estimates
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
- Walmart:
 - 1+ million customer transactions every hour
 - imported into databases, estimated 2.5+ petabytes of data
 - =167 times all books in the US Library of Congress
- London bus networks not known in completeness; reconstructed (also) using pickpocket statistics [gossip]

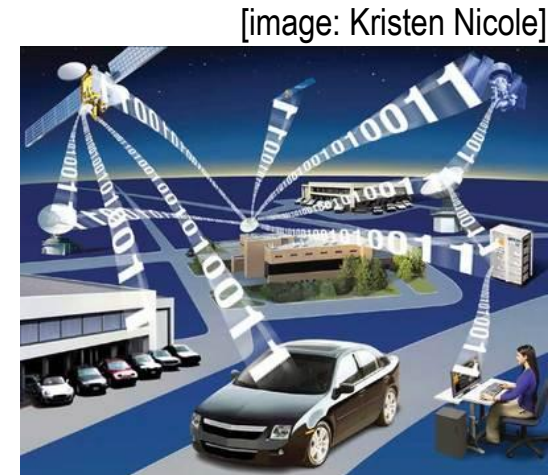
Big Data in Industry

- **Industry 4.0:** integration of production & ICT

- Optimization of value chain & life cycle

- **Automotive**

- Typical upper-class car: ~100m lines of code
- Getting networked with traffic, lights,...
 - 2.8 ZB in 2012, plus 2.5 PB / day [Computerwoche]



- **Aircrafts:**

- A380: 1 billion lines of code
- Per engine: 1 TB / 3 min
 - LHR → JFK = 640 TB



Big Code – Lines of Code

Average iPhone app

Hubble Space Telescope

Windows 3.1 (1992)

Control software for US military drone

Windows NT 3.1 (1993)

HD DVD Player Xbox

World of Warcraft Server

Google Chrome

Windows NT 4 (1996)

MySQL

Boing 787 Flight Software

F35 Fighter jet

Microsoft Office 2013

Large Hadron Collider

Facebook

US Army Future Combat System

MacOS X 4.1 Tiger

Average high-end car

1.3+ million iPhone apps,

1.3+ million Android apps

= 50.000 lines

= 2 million lines

= 2.5 million lines

= 3.5 million lines

= 4.5 million lines

= 4.5 million lines

= 5.5 million lines

= 6.5 million lines

= 11 million lines

= 12 million lines

= 14 million lines

= 23 million lines

= 44 million lines

= 50 million lines

= 61 million lines

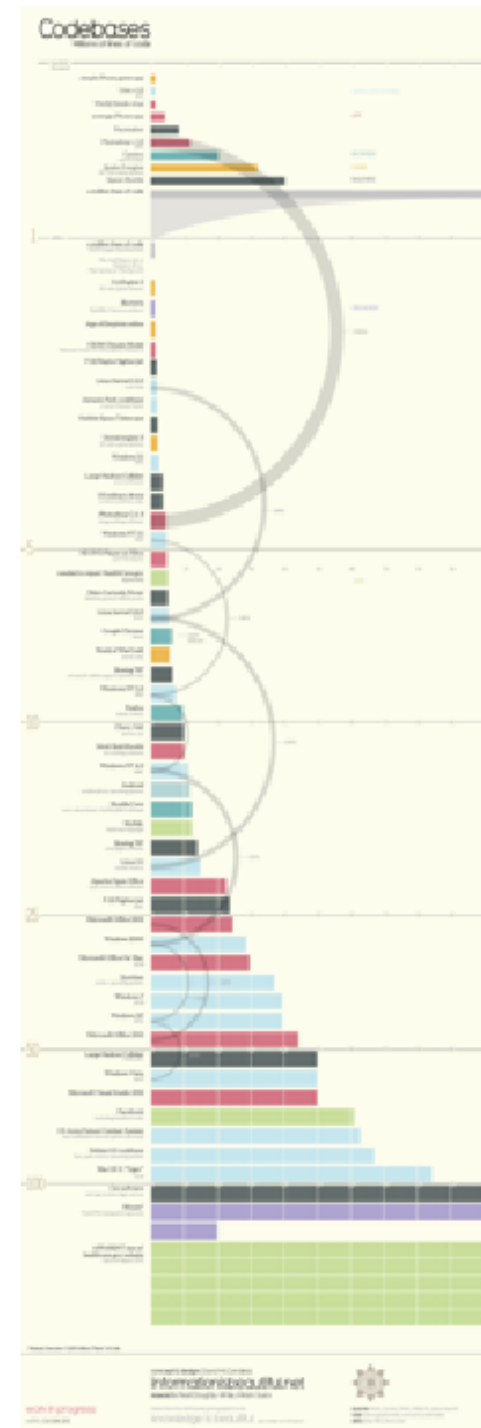
= 63 million lines

= 85 million lines

= 100 million lines

= 170 billion lines

source: <http://www.informationisbeautiful.net/visualizations/million-lines-of-code/>

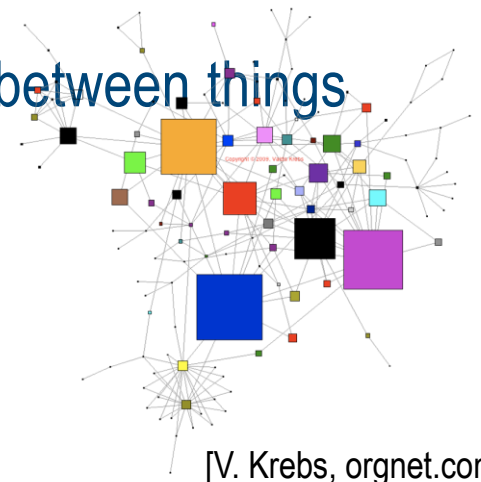


Big Data in Social Networks

- Facebook: 1m users 2004, 1.11b in 2013 [Fb]
 - 40b user photos [Wikipedia]
- Chats via MS Messenger [Leskovec, WWW 2008]
 - 30b chats between 240m participants
 - communication graph with 180m nodes, 1.3b undirected edges
 - Result : everybody knows everybody over at most 7 edges
- **Social Network Analysis** (SNA): map & measure links between things
 - How highly connected is an entity within a network?
 - What is an entity's overall importance in a network?
 - How central is an entity within a network?
 - How does information flow within a network?



[M. Rodriguez, Aurelius]



[V. Krebs, orgnet.com]

Internet of Things (IoT)

- **Every (physical) thing** is connected to the Internet
 - „the Internet“ knows state of physical world – more and more comprehensively
- Not new on principle
 - Anti-blocking brakes, engine emergency shutoff, RFIDs in car & discos, ...
- New: extent, integration, comprehensive evaluation ...in real-time
 - T-Shirt, fridge, beer bottle, fitbit, car, family, neighbours, insurance, boss, ...
- Data protection, data security?
 - Known issues, novel dimension



[Shutterstock, Forbes]

2012



Reading: „The 4th Paradigm“

- eScience: computationally intensive science
 - Complex computing and/or immense data
 - “where IT meets scientists”
- Experimental Budgets 1/4 to 1/2 Software
 - Sloan Digital Sky Survey (SDSS): Telescopes 15 ~ 20 m US\$, but software dominates
 - Neptune ocean observatory: 30% of 350m US\$ budget for cyberinfrastructure = 100m US\$
- Joint effort of various CS domains
 - databases, data mining, workflow management, visualization, cloud computing, ...



Data Scientist



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Tony Hey, Stewart Tansley, Kristin Tolle (eds.)

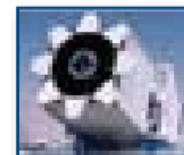
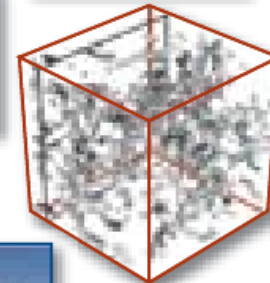
Reading: „The 4th Paradigm“

Science Paradigms

- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience)
unify theory, experiment, and simulation
 - Data captured by instruments
or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files
using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



„Big Data“: Definition

- 4V definition [Doug Laney / Gartner & IBM]:
 - Volume
 - Velocity
 - Variety
 - Veracity
- plus more in blogs: Value, Verisimilitude, Variability, Visualization, ...
- ...or simply: „Data too big to transport“

The 7 Computational Giants of Massive Data Analysis

- Basic Statistics
- Generalized N-Body Problems
- Graph-Theoretic Computations
- Linear Algebraic Computations
- Optimizations
- Integration
- Alignment Problems

[Frontiers in Massive Data Analysis.
US National Research Council, 2013]

Prominent Big Data Technologies

- MapReduce = programming model for processing large data sets with a parallel, distributed algorithm on a cluster
 - MapReduce program = Map() + Filter()
 - MapReduce system orchestrates parallelization
 - Most popular implementations: Hadoop, Spark
 - “MapReduce” originally referring to proprietary Google technology, now generic name
- TopTen Databases Program 2005 [www.wintercorp.com]:
 - size of production databases tripled since 2003; 100 TB landmark in 2005
 - Yahoo! database first production data warehouse >100 TB (100.4 TB; Unix; Oracle)
 - largest Windows database: 19.5 TB (2x over 2003)
 - highest throughput: 1.1m SQL statements per hour (z/OS, IBM UDB DB2)

Big Data Initiatives

- Research Data Alliance – www.rd-alliance.org
- NIST Big Data Initiative – bigdatawg.nist.gov
- ISO /IEC JTC1 SC32 Big Data Analytics
- OGC Big Data WG – external.opengeospatial.org/twiki_public/BigDataDwg
- all remotely data oriented conferences tackle Big Data
 - Core DB conference: VLDB

Big Data Initiatives / contd.

- United Nations and Governments Initiatives
 - United Nations: [Global Pulse](#)
 - United States: ["BIG DATA" Initiative](#) (\$200m US\$), March 29, 2014
 - European Union: [Big Data at your service](#), July 25, 2014
- Industry Initiatives
 - [IBM Big Data](#);
 - [SAS Big Data](#)
 - [Oracle Big Data](#)
 - [Google BigQuery](#)
 - [Microsoft Big Data](#)

Big Data Buzzwords

- Big Data Architecture
- Big Data Modeling
- Big Data As A Service
- Big Data for Vertical Industries (Government, Healthcare, etc.)
- **Big Data Analytics**
- Big Data Toolkits
- Big Data Open Platforms
- Economic Analysis
- Big Data for Enterprise Transformation
- Big Data in Business Performance Management
- Big Data for Business Model Innovations and Analytics
- Big Data in Enterprise Management Models and Practices
- Big Data in Government Management Models and Practices
- Big Data in Smart Planet Solutions

[IEEE Big Data Conf.]

Big Data Requires Many Disciplines

Using techniques from:

- Databases
- Supercomputing
- Data Mining
 - Artificial Intelligence
 - Machine Learning
 - Statistics
- Natural language processing
- Visualization

Domains:

- Business Intelligence
- Social networks
- Online trading
- Geospatial (& temporal) data
- + *many more...*

**Caveat: not a strict definition;
see also this discussion:
<http://wmbriggs.com/blog/?p=6465>**

Impact of „Big Data“

- New job profile: **Data Scientist**
 - CS (databases, data mining, visualization, HPC, ...) + statistics + sci domain
- New **data management & analytics paradigms**
 - MapReduce, No/NewSQL, ... far from consolidated
- New **ethical dilemmas**
 - NSA spying of Chancellor Merkel phone & other incidents

- Science, and even society, more and more **data driven**
 - „drowning in data, starving for information“
- **Big Data** = summary term for **data too big / complex** to transport, to analyze
 - Internet of Things; sensors; social networks; business data; science data; network traffic; ...
 - Some say: Big Data = Big Hype
 - But Vs leading to clarification of issues
- *“Big Data is a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.” [ACM 2013]*