Big Data
Fluch oder Segen?

Dr. Rainer Schmidt
Scientist
AIT Austrian Institute of Technology GmbH

Energiesysteme im Umbruch VI
OVE Österreichischer Verband für Elektrotechnik
4.10.2017

Faster and Faster Growing Volumes of Data

- 90% of the world‘s data was produced in just the past two years.
  - networks, sensors, IoT
- NASA Solar Dynamics Observatory
  - 4 telescopes gathering 8 images of the sun every 12 sec. In Jan 2015, 100 million images
- Square Kilometre Array radio telescope project
  - Expected to produce 2.8 Gbytes of data / second to create astronomic maps of the universe.
- CERN’s Large Hadron Collider
  - 150 million sensors capturing data about nearly 600 million collisions per second.
- WWW and Social Media
  - Facebook users add 300 million new photos a day
  - 300 million Instagram users share 60 million photos every day
  - More than 100 hours of video uploaded to YouTuber / minute

Data-Driven Scientific Discovery

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

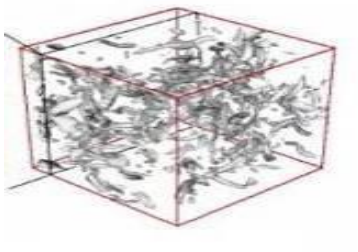
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

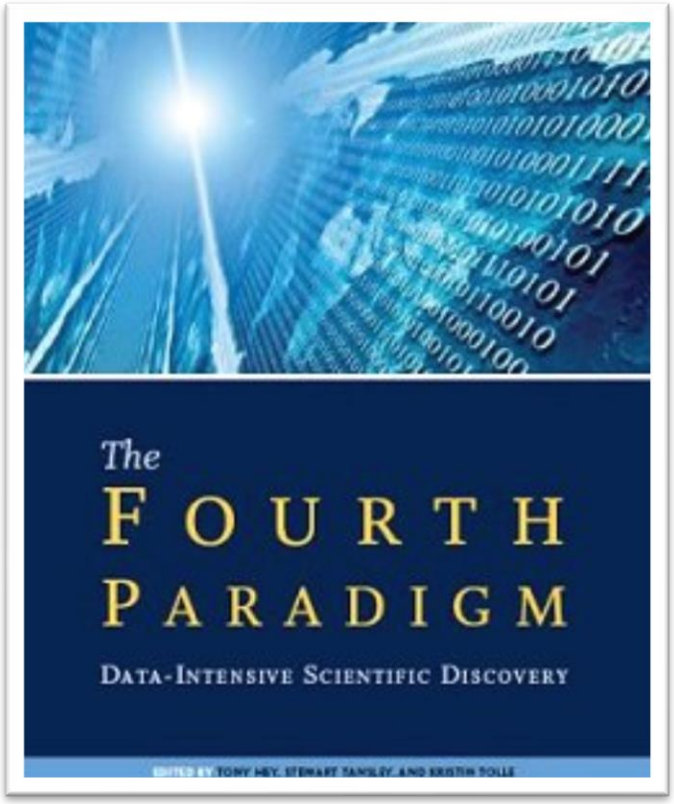
- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

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



Taken from Tony Hey: The Fourth Paradigm: Data-Intensive Scientific Discovery and Open Science

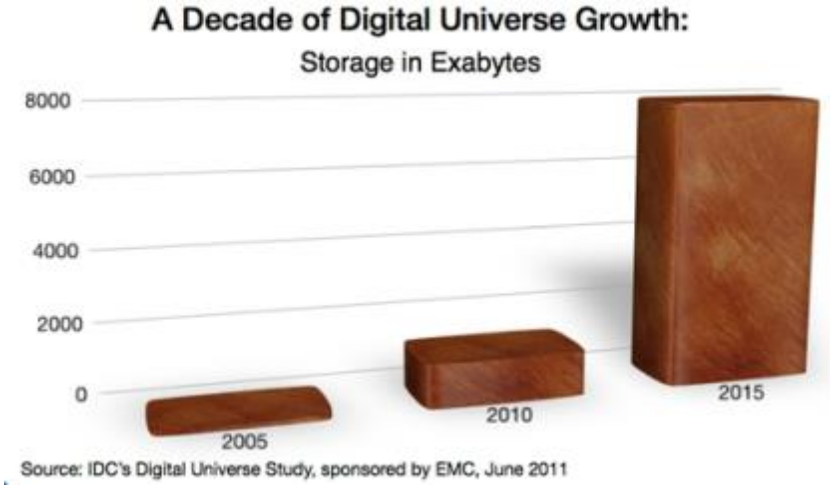
Data-Intensive Scientific Discovery

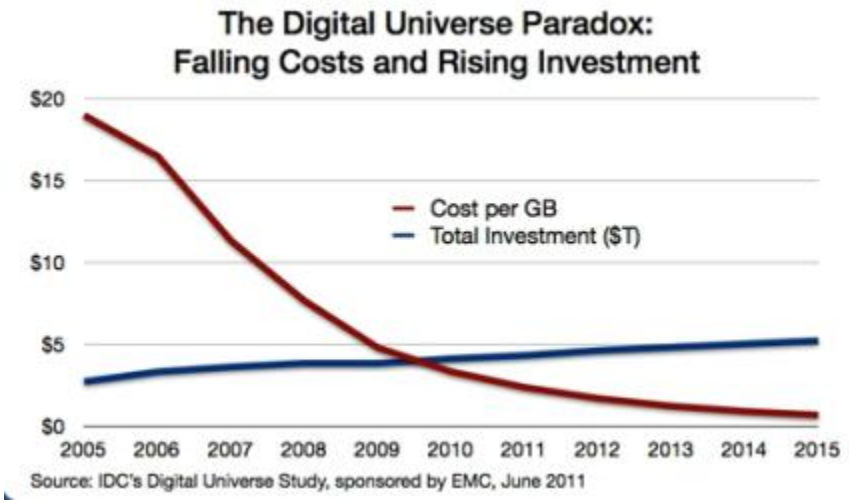


Published 2009 under Creative Commons License and available online from The Fourth Paradigm and Science@Microsoft at http://research.microsoft.com and on Amazon.com

Growing Volumes of Data

- Drivers
  - Internet of Services
  - Cloud Computing
  - Internet of Things
  - Cyberphysical Systems
- Resources
  - Video Streams
  - Audio Streams
  - Sensor Data
  - Web Archives
  - Open Data Sets
  - Generated Technical Data
  - Social Media



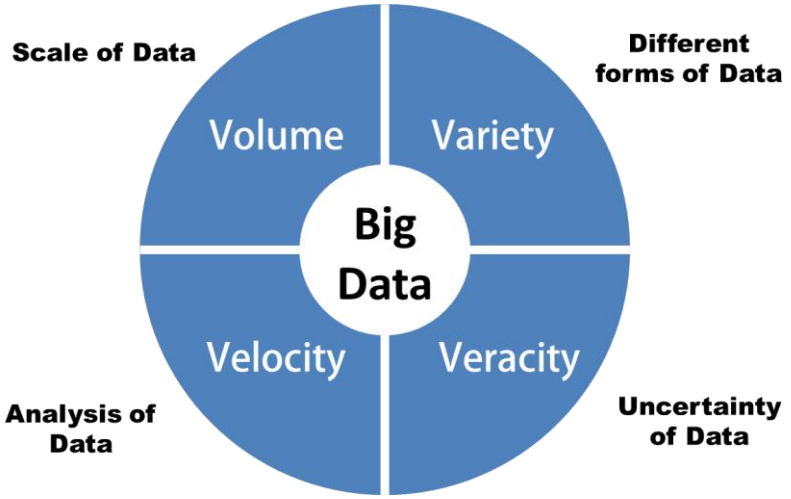


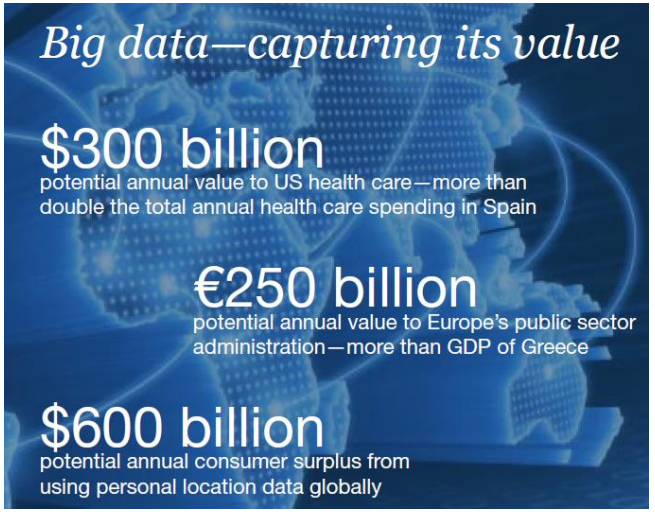
Data-Driven Applications

- Transportation Industry
  - Lifecycle Management
  - Predictive Maintenance
- Traffic Management
  - Tracking and Control
- Energy Management
  - Grid Management Systems
  - IoT, Complex Event Processing
- Water Management
  - Network Mgnt.
  - E.g. Rain-data sensors
- Health
  - Medical Imaging
  - Patient Data Management
- Market Research
  - Profiling and Predicting
- E-Commerce
  - Search, Ranking, Ordering

Big Data – What is it?

- No clear origin for the term, despite a number of claimants*
- Recently popularized in “Big Data: A Revolution That Will Transform How We Live, Work and Think” by Cukier and Mayer-Schönberger (2013)
- Defined as too large and too complex to capture, process, and analyze using current computing infrastructure
- Often described in terms of the 4Vs
  - Originally three (Doug Laney/Gartner, 2001), then added Veracity
  - Now seven (adding Visualization, Variability, and Value)
- Huge investments by science and commercial sector (Internet), believed to have lot of value





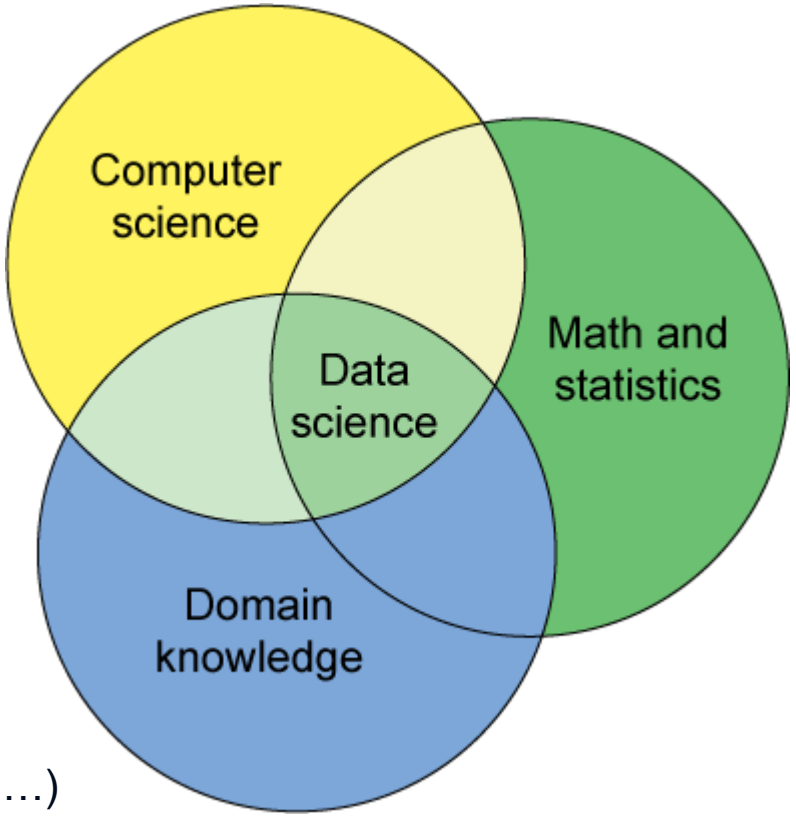
**

* http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/?_r=0

** http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Data Science

- The Statistical Trilogy (1997, Jeff Wu)
  - Data Collection
  - Data Modeling and Analysis
  - Problem Solving & Decision Making



- **Data Science** is the process of extracting knowledge or insight from large volumes of structured or unstructured heterogeneous data

- Data Scientist
  - Various Domains (Engineering, Physics Logistics, Medicine, Image Processing, …)
  - Complex Computer Systems (Scale Out Architecture, Distributed Memory, Databases, Optimization, …)
  - Statistics (ML, CNN, Dimensionality Reduction, Linear Algebra, Regression, Mathematical Programming, …)

What is a Data Scientist?

Data Engineer

People who are expert at

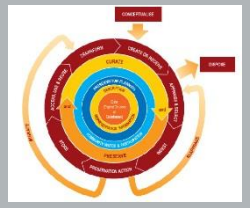
- Operating at low levels close to the data, write code that manipulates
- They may have some machine learning background.
- Large companies may have teams of them in-house or they may look to third party specialists to do the work.

Data Analyst

People who explore data through statistical and analytical methods

- They may know programming; May be an spreadsheet wizard.
- Either way, they can build models based on low-level data.
- They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.

Data Steward



People who think to managing, curating, and preserving data.

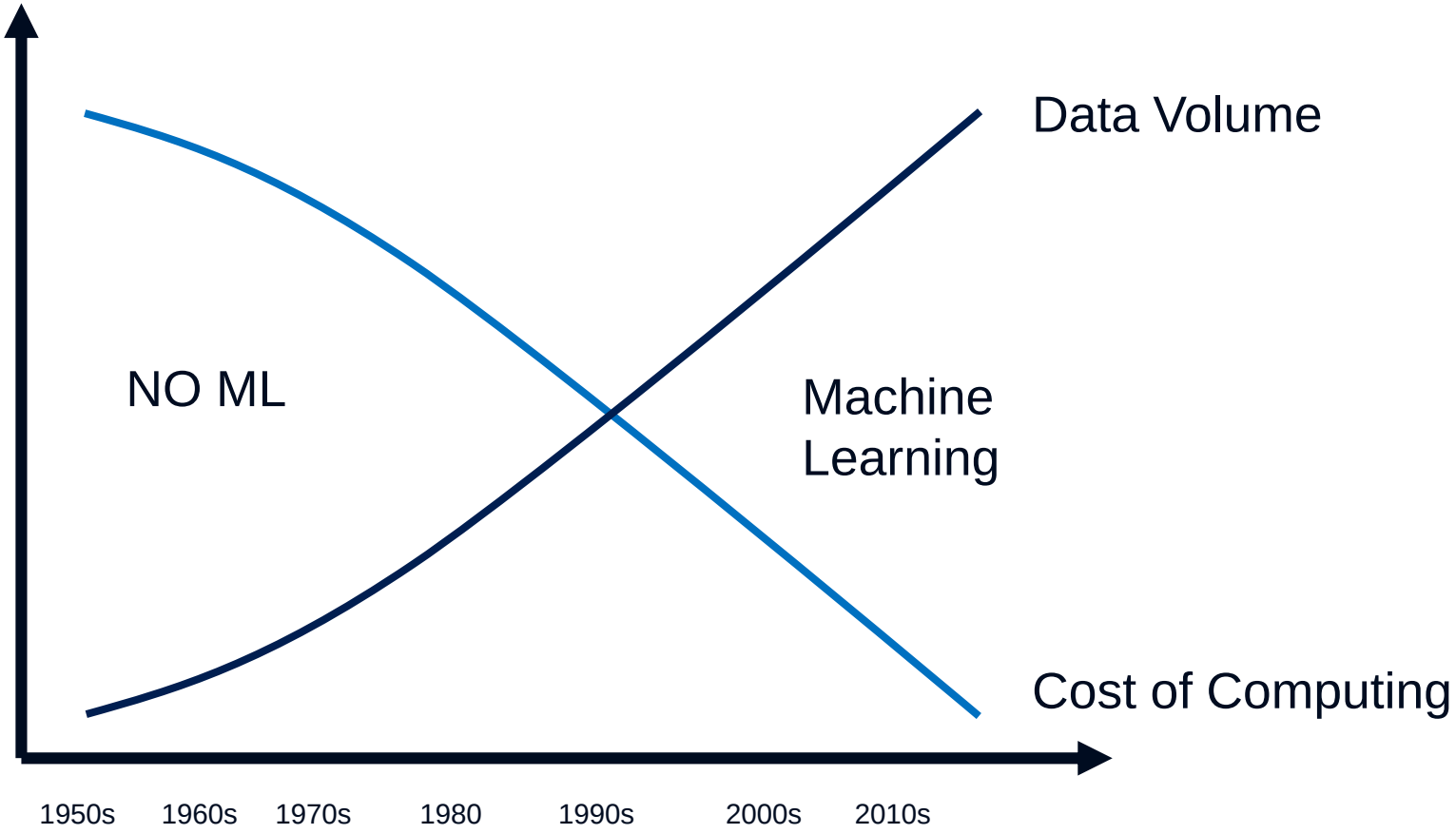
- They are information specialists, archivists, librarians and compliance officers.
- This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

What is a data scientist? Microsoft UK Enterprise Insights Blog, Kenji Takeda
http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/01/31/what-is-a-data-scientist.aspx

Complex Data Analytics of Big Data

Based on Volker Markl – Deep Analysis of Big Data

Big Data and Machine Learning

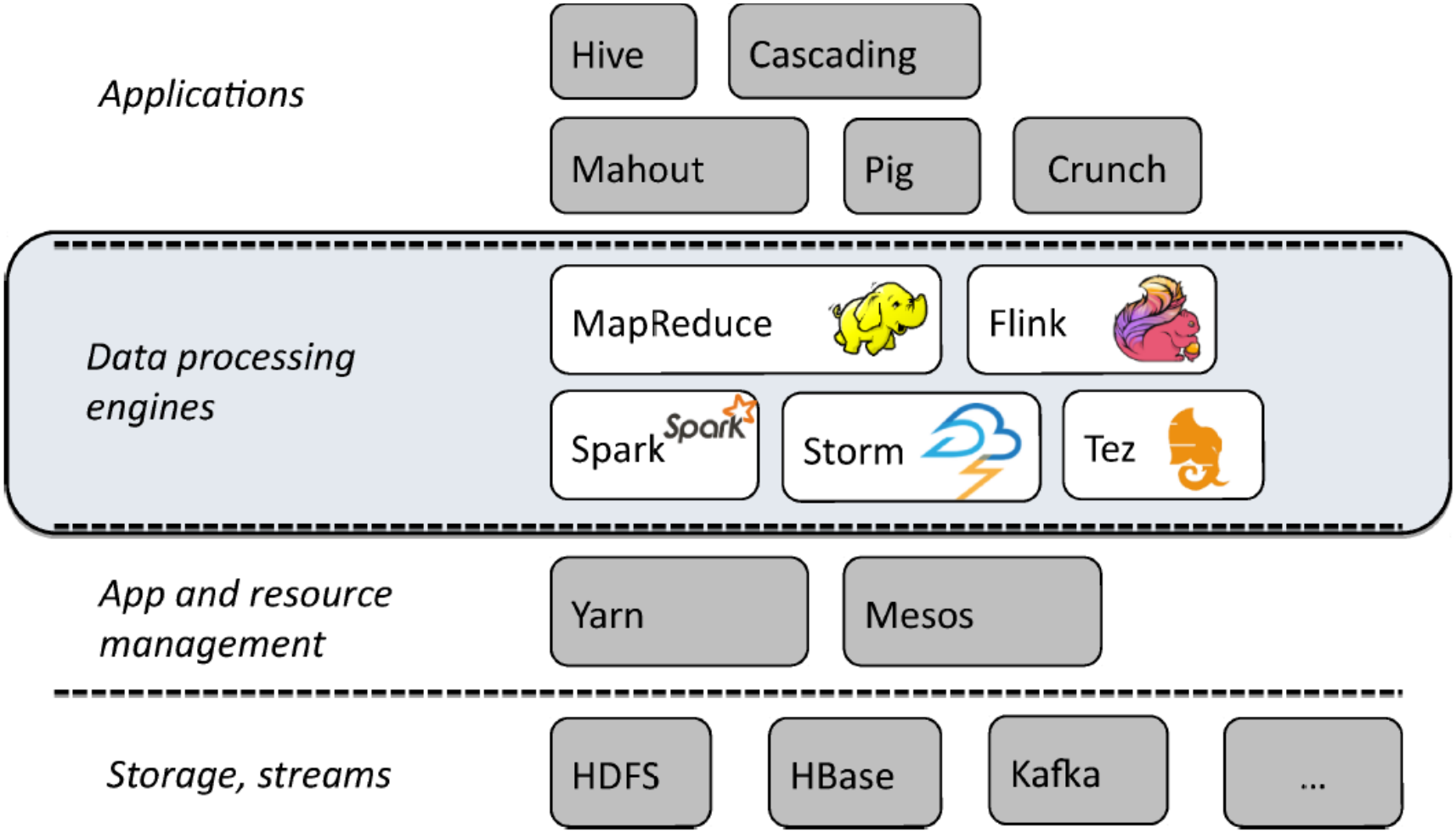


Based on Mike Olson (Cloudera) 2017 keynote talk – the Machine Learning Renaissance

- Open Source Apache Project
- Derived from publications Google File System and MapReduce publications (2003, 2008).
- Supported by Yahoo!, Facebook, ….
- Hadoop Core includes:
  - Distributed File System - distributes data
  - Map/Reduce - distributes application
- Runs on commodity hardware
- 2008, fastest sort of a TB of data, 3.5min, over 910 nodes.
- Today, part of a rich data-intensive computing eco-system

Based on Owen O'Malley

Open Source Processing Landscape



Key Functionalities

- Batch Processing
  - Loading and processing large volumes of data.
- Interactive SQL
  - Self-service for BI analysts
- Search
  - User-friendly search for text
- NoSQL
  - Real-time single event querying
- Stream Processing
  - Real-time querying on collection of events
- Machine Learning
  - Data preparation, model application for data scientists and advanced analytics

Big Data – for Better or Worse

- We are constantly producing data about ourselves
  - Passively and actively; in reality and in cyberspace
- Internet Companies have access to large volumes of personalized data that can be grouped and analyzed.
  - Also run large research centers.
  - Data is a big advantage for those companies.
- Big Data can tell us a great deal about developments in the world in many different areas.
  - How trustworthy?, Biased?, Social/online media?
- Big Data can become a problem
  - Combining sources of data (for commercial use).
  - Tax, health, insurance registers, financial records
  - Vulnerability and data protection are key issues

Based on Åse Dragland, Big Data, for better or worse: 90% of world's data generated over last two years, SINTEF, May 2013.

Some Data Sets from Projects at AIT

- **Audio- and Video from CCTV and mobile devices**
  - footage of terror attacks
- **Bitcoin transaction graph**
  - illegal transactions, money laundering, …
- **Web archives and snapshots of public domains**
  - .dk, .it, .eu, or gov.uk domains
- **Millions of digitized historic materials**
  - newspapers, posters, and books (up to 230MB/object)
- **TBs of broadcast radio and TV output**
  - up to 73GB/object
- **Many hundreds of thousands of scientific data sets**
  - Generated by scientific instruments like synchrotrons
- **Ten thousands items from scientific publications**
  - open access journal articles

Summary

- Big Data technology has its origins in the Big Sciences
  - High-Energy Physics, Astronomy
- Very successful projects developed by large Internet companies
  - Google, Yahoo, Twitter, Facebook, …
- As overall data grows, applicable to many more domains
  - Critical Infrastructure, Law Enforcement, Financial Sector, Industry
- Technically, Big Data is tough to handle
  - Requires access to infrastructure
  - Often hard to tackle common problems at large
- However, new tools and infrastructures available to data scientists
  - Cloud computing, new platforms, high-level machine learning, …
- These opportunities come also with massive legal and ethical issues
  - Storing, tracking, profiling, machine-based decisions …
- Overall, the big data hype is fading
  - Technologies available and evolving at great pace
  - but potential has yet to be realized in many fields

Danke für die Aufmerksamkeit!