Big Data Management and Analytics

Lecture Notes

Winter semester 2016 / 2017
Ludwig-Maximilians-University Munich

Based on lectures by
Donald Kossmann (ETH Zürich), as well as
Jure Leskovec, Anand Rajaraman, and Jeff Ullman (Stanford University)

Course Logistics

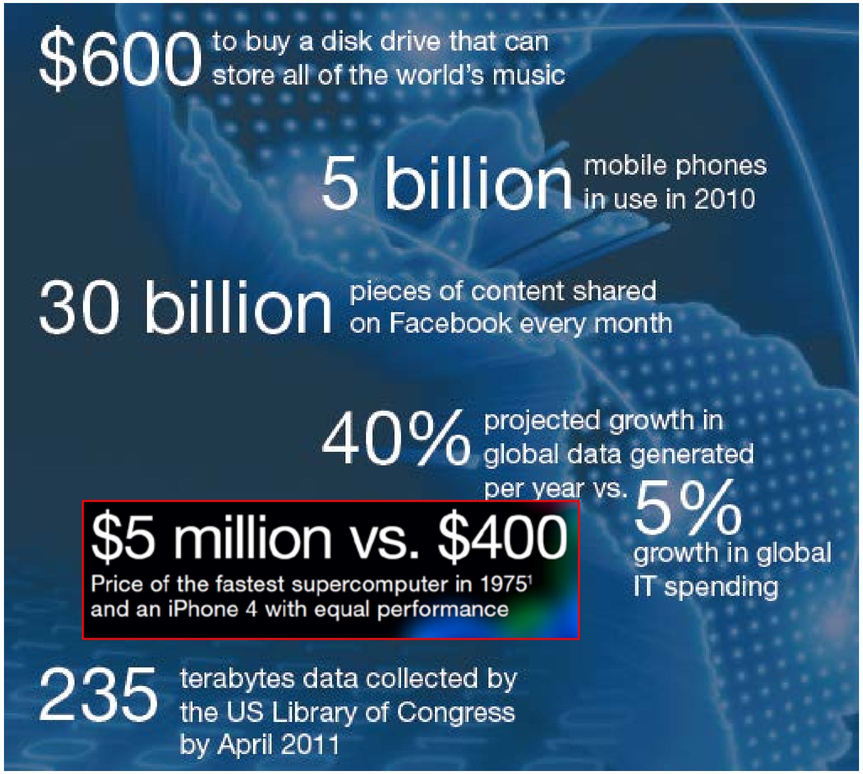
- Course website:
  - http://www.dbs.ifi.lmu.de/cms/Big_Data_Management_and_Analytics
  - Registration for this lecture is now open via Uniworx
  - Registration required to attend the exams!!!
- Organization:
  - Load: 3+2 hours weekly
  - Required: Lecture "Database Systems I" or equivalent
  - Beneficial: Lecture "Knowledge Discovery in Databases I" or equivalent
  - Lecture: Prof. Dr. Matthias Schubert
  - Assisting: Daniyal Kazempour

Why this course?

- **Big Data is big**
  - $ and science: choose your poison

We are drowning in data … but starving for information

- Exponential grows in data



$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

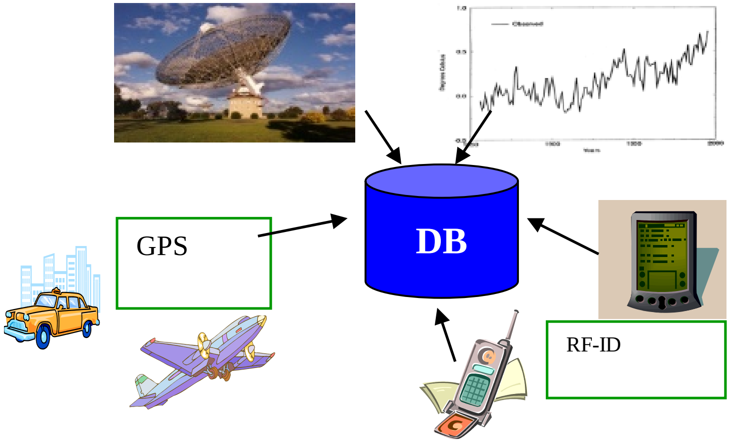
30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

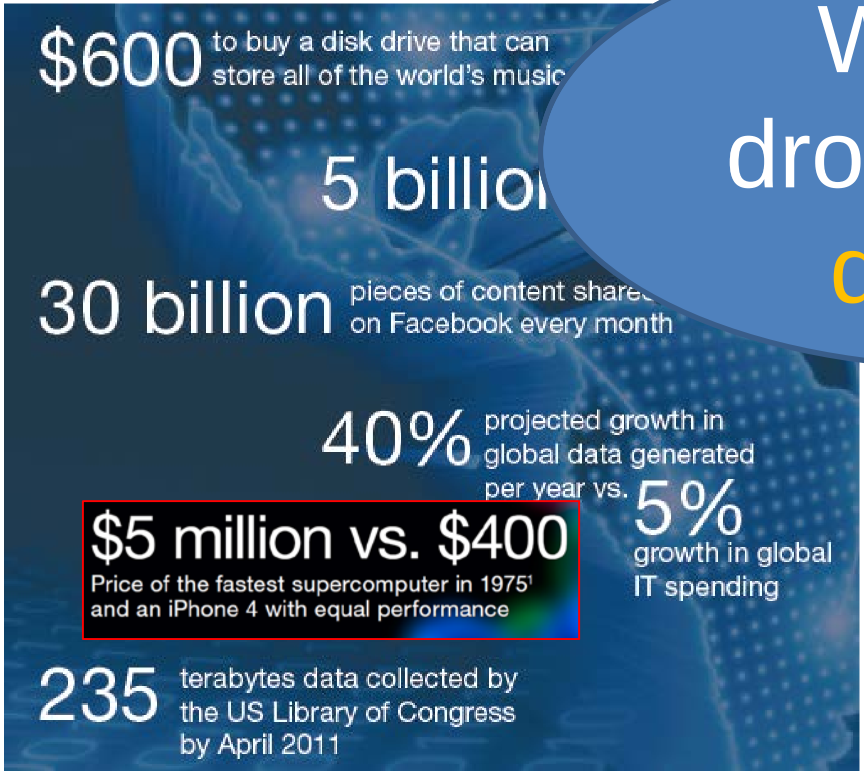


http://www.popsci.com/announcements/article/2011-10/november-2011-data-power

- Data contains value and knowledge

We are drowning in data … but starving for information

- Exponential grows in data



$600 to buy a disk drive that can store all of the world's music

5 billio

30 billion pieces of content share on Facebook every month

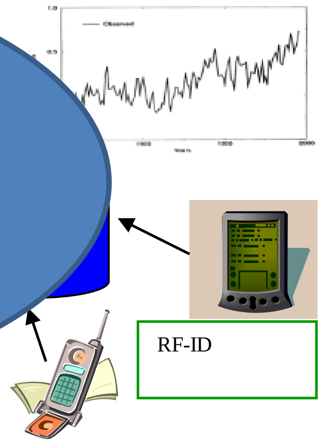
40% projected growth in global data generated per year vs. 5% growth in global IT spending

$5 million vs. $400 Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

We are drowning in data…



http://www.popsci.com/announcements/article/2011-10/november-2011-data-power

- Data contains value and knowledge

We are drowning in data … but starving for information

- Exponential grows in data

$600 to buy a disk drive that can store all of the world's music

5 billio

30 billion pieces of content share on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in glob IT

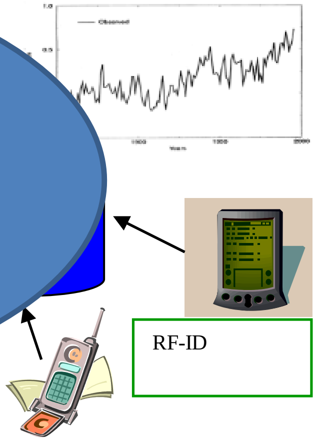
$5 million vs. $400

075¹

Datasets,

We are drowning in data…

…but starving for information



http://www.popsci.com/announcements/article/2011-10/november-2011-data-power

- Data co ue and knowledge

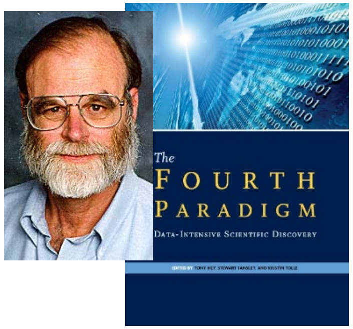
Why this course?

- **Big Data is big**
  - $ and science: choose your poison
  - Big Data approaches required for Data Science
    “move data from raw to relevant”

Data Science (~eScience/Industry 4.0)

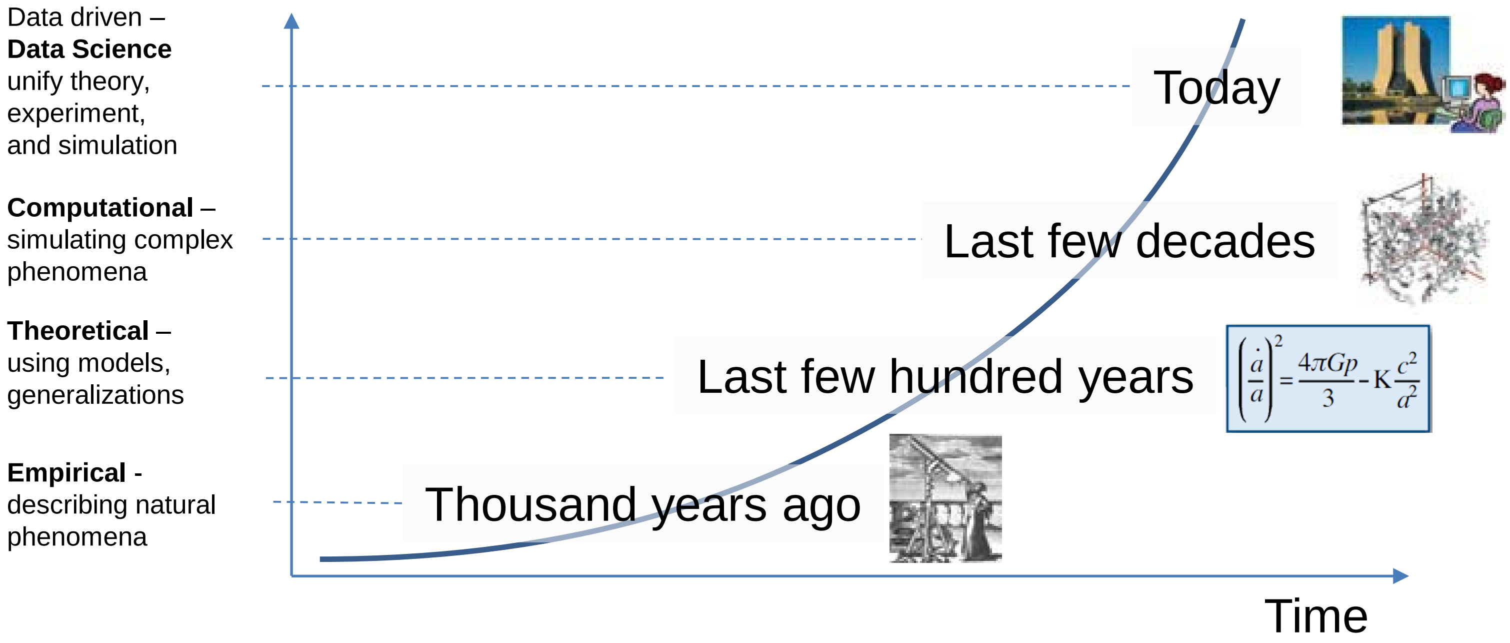
- **The Fourth Paradigm:**
  Age of data driven exploration
  → **Data Science** (eScience / Industry 4.0)

[Informatik Pionier Jim Gray]



[Hey, Tansley, Tolle: Fourth Paradigm, 2009]

- **Science Paradigms**



Data driven – **Data Science** unify theory, experiment, and simulation

Computational – simulating complex phenomena

Theoretical – using models, generalizations

Empirical - describing natural phenomena

Today

Last few decades

Last few hundred years

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

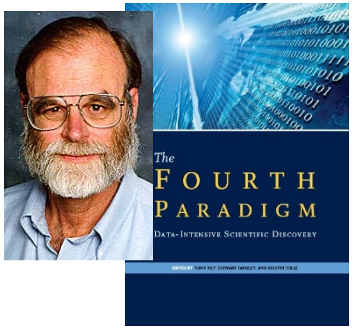
Thousand years ago

Time

Data Science (~eScience/Industry 4.0)

- **The Fourth Paradigm:**
  Age of data driven exploration
  → **Data Science** (eScience / Industry 4.0)

[Informatik Pionier Jim Gray]



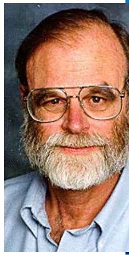
[Hey, Tansley, Tolle: Fourth Paradigm, 2009]

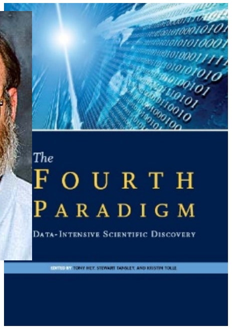
- **Data Science**
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist/Analyst analyzes database / files using data management and statistics

Data Science (~eScience/Industry 4.0)

- **The Fourth Paradigm:**
  Age of data driven exploration
  → **Data Science** (eScience / Industry 4.0)

[Informatik Pionier Jim Gray]





[Hey, Tansley, Tolle: Fourth Paradigm, 2009]

- **Data Science**
  - Data
    gene
  - Proc
  - Information/knowledge stored in computer
  - Scientist/Analyst analyzes database / files using data management and statistics

> *“Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data.”*
>
> [Hey, Tansley, Tolle: Fourth Paradigm, 2009]

Why this course?

- **Big Data is big**
  - $ and science: choose your poison
  - Big Data approaches required for Data Science "move data from raw to relevant"

- **Big Data is exciting**
  - gives a new twist to almost everything
  - allows you to reinvent the wheel

Why this course?

- **Big Data is big**
  - $ and science: choose your poison
  - Big Data approaches required for Data Science "move data from raw to relevant"
- **Big Data is exciting**
  - gives a new twist to almost everything
  - allows you to reinvent the wheel
- **Big data is old**
  - opportunity to teach you some fundamental technology

Outline of this course

- Introduction (Motivation and Overview)
- Introduction to Big Data — the four V's
- NoSQL
- Hadoop / HDFS / MapReduce & Applications
- Spark
- Data Stream Processing & Applications & Algorithms
- High-Dimensional Data
- Graph Data Processing
  (Link Analysis, Page Rank, Community Detection)

Literature

- This course is mainly based on a mixture of existing external lectures, Surveys, Papers and Reports on Big Data
- There is NO, or better, I'm not aware of a single book or script that is equivalent to this course (and addresses all issues discussed in this course)
- Since Big Data is a quite new and hot topic, standards and basic concepts are quite dynamic => The Web is a very appropriate source of relevant information
- External lectures basically used for this course:
  - Big Data: Donald Kossmann & Nesime Tatbul, Systems Group ETH Zurich - http://www.systems.ethz.ch/node/217
  - Mining of Massive Datasets: Jure Leskovec, Anand Rajaraman, Jeff Ullman, Stanford University - http://www.mmds.org
- Further material will appear at our web page
  (check for updates during the course / open to further suggestions!)