Big Data Systems

Winter Semester 2019 / 2020

Introduction

Prof. Tilmann Rabl

Data Engineering Systems

Hasso-Plattner-Institut

This Lecture

1. Introduction
   - Big Data
   - Data Engineering
2. Course Organization
   - Content, Timeline, Textbooks
   - Data Engineering Curriculum
   - Exercises & Exam
   - Registration & Projects

Who am I?

Until 2011: PhD in CS at University of Passau

- Distributed databases

Until 2015: Postdoc at University of Toronto

- Big data systems / benchmarking

Until April

- Visiting Professor & Research Director at DIMA group, TU Berlin
- Deputy Director of Department IAM at DFKI
- Scientific Coordinator of the Berlin Big Data Center

Since Mai

- Professor for Data Engineering Systems, Digital Engineering Faculty, HPI, U Potsdam

Other activities

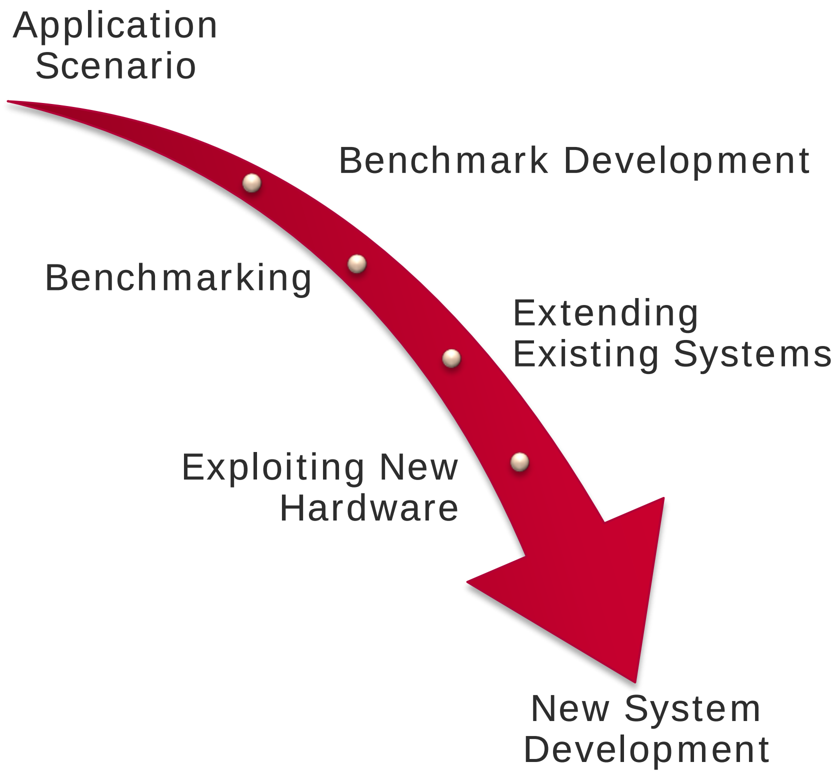
- TPC Professional Affiliate
- Co-Founder of bankmark

Research Topics

- Database Management
  - SIGMOD 17, VLDBJ 18
- Machine Learning Systems
  - PVLDB 17, SOCC 18, DAMON 18, EDBT 19
- Stream Processing
  - ICDE 18, EDBT 18, SIGMOD 19, PVLDB 19
- Benchmarking
  - SOCC 17, BeyondMR 17, ICDE 18, ICPE 18

Interested in a thesis? Write us an e-mail.

Research Approach



Big Data

Data is growing

Messages, tweets, social networks (statuses, check-ins, shared content), blogs, click streams, various logs, ...

- *Facebook: > 1,5B active users, > 60B messages/day*
- *Twitter: > 300M active users, > 500M tweets/day*

Everyone is interested!

The value of data is decreasing with its age!

What is Big Data?

- **Big data** is an *accumulation* of data that cannot be processed / handled using traditional data management processes / tools.

- A *big data management infrastructure* should ensure that the underlying hardware, software, and architecture have *the ability to enable learning (from data) using analytics*.

Big Data Characteristics (*in* ***V****s*)

- volume
  - □ “data at rest”
- the amount of data with respect to the number of observations (size of the data) and the number of variables (dimensionality of the data)

- variety
  - □ “data in many forms”
  - □ heterogeneity of data types (i.e., structured, semi-structured, unstructured)
  - □ data sources that are either private or public
  - □ examples include log files, text, web, images, video, audio

Big Data Characteristics (*in* ***V****s*) - Contd.

- velocity
  - “data in motion” or “data in transit” (aka streaming data)
  - … is concerned with the data generation rate
  - … is concerned with the rate which data arrives
  - … is concerned with the timeframe in which they must be acted upon
  - requires appropriate data handling mechanisms

- veracity
  - “data in doubt”
  - … is concerned with noise and processing errors, including the reliability (i.e., quality over time) and validity of the data

- value, variability, validity, vulnerability, volatility, visualization, …

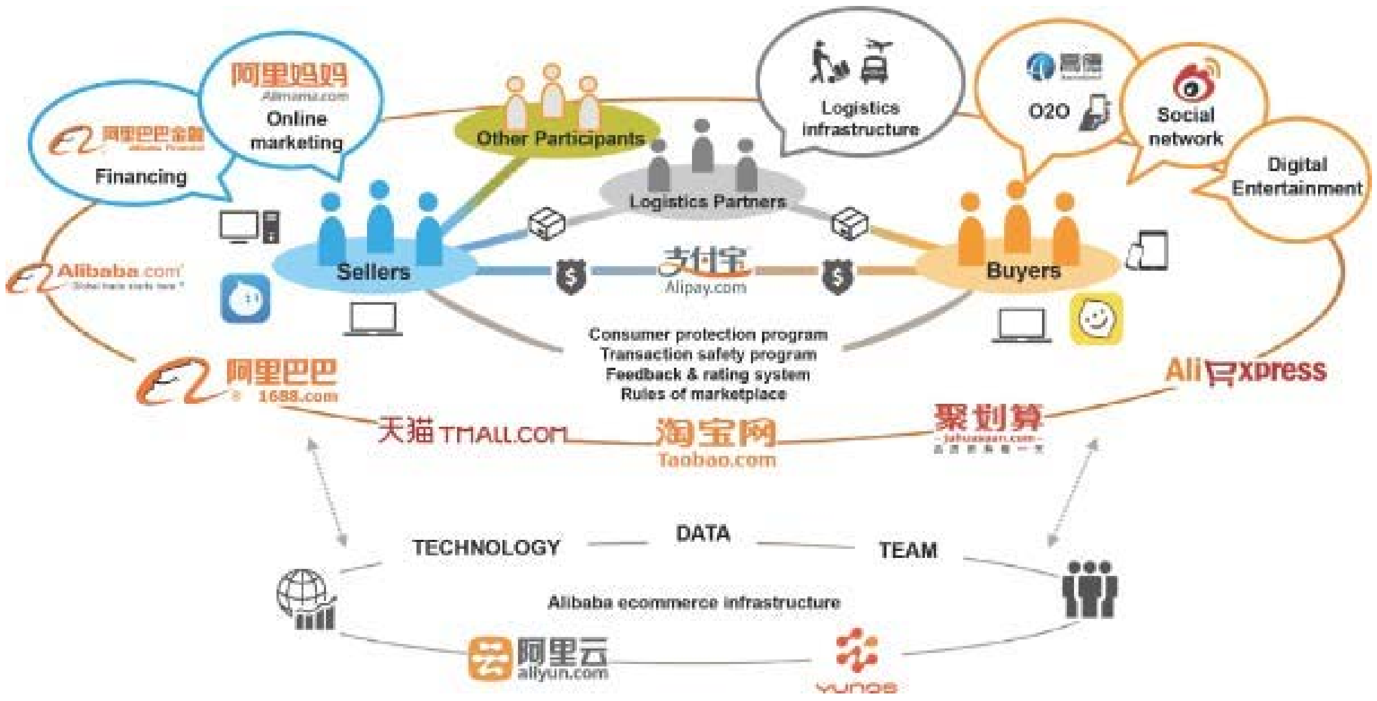
Where Does Big Data Come From?
Online Retail – Alibaba Group

Alibaba Group 阿里巴巴集团

- 443M internet users (Feb 2017)
- 493M mobile users
- 12,7B annual orders
- 255M annual active buyers
- 57M packages/day

Alibaba Singles Day 2018 (11.11.)

- USD 1B$ in <2 min
- USD 30B$ the whole day

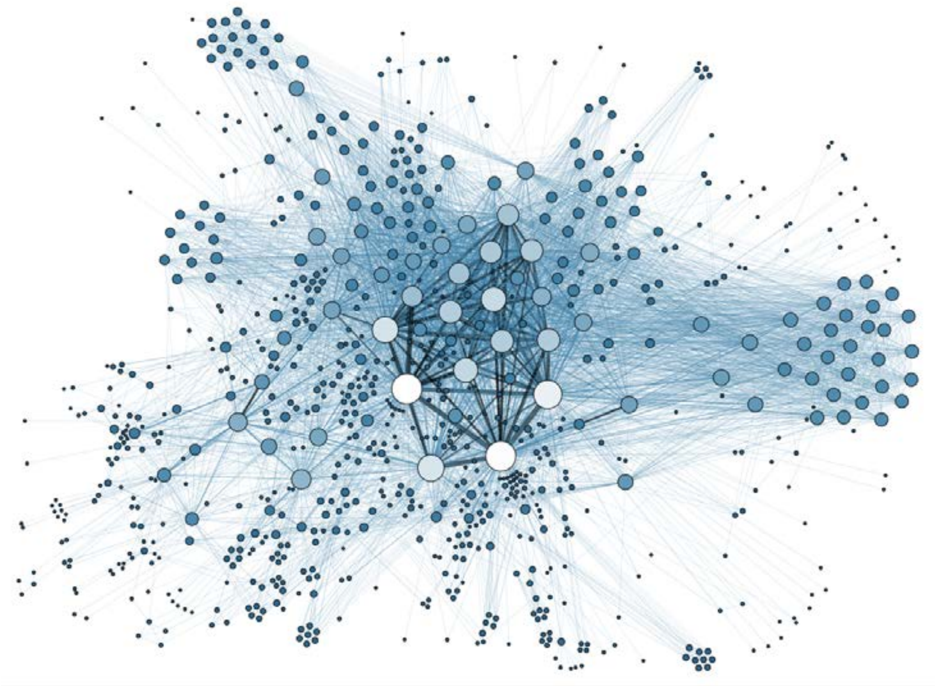


- Buyers, Sellers, Logistics, Payments, …
- Cloud computing, Big Data Technologies , Machine Learning, …

Where Does Big Data Come From?
Social Network - Twitter

- ~ 500M tweets sent per day current
- 310M monthly active users
- 1.3B accounts created
- 500M visitors monthly without logging in
- 208 the average number of followers
- Katy Perry has the most followers with over 87M
- 83% of world leaders are on Twitter

- Applications
  - □ Shortest paths between two people
  - □ Centralities analysis
  - □ Finding and ranking expertise
  - □ Social media monitoring
  - □ Security



Risks of (Big) Data

Mistakes

Spurious correlations

Bias

Simpsons paradox

Manipulation

Fit data to result

Biased training

Manipulative visualization

Misuse

Discrimination

Observation

Data privacy violations

Espionage

Data Monopoly

Circle of big companies

Lack of control

Non-equal opportunity

General problems

- Lack of knowledge
- Lack of ethics

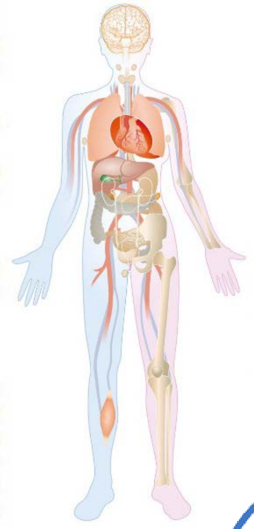
General solutions

- Education
- Regulation

But there is so much more...

Autonomous Driving

- Requires rich navigation info
- Rich data sensor readings
- 1GB data per minute per car (all sensors)[1]



E-health

- 3.2 billion base pairs of DNA (genomics)
- 10 million proteins in a person (proteomics)

Pre-processing of sensor data

- CERN experiments generate ~1PB of measurements per second.
- Unfeasible to store or process directly, fast preprocessing is a must.

[1] Cobb: http://www.hybridcars.com/tech-experts-put-the-brakes-on-autonomous-cars/

Benefits of (Big) Data

Economics	Automotive	Health	Science
Predictive maintenance	Traffic optimization	Early diagnostics	Evidence-based research
Fraud detection	Easier multi-modal routing	Personalized medicine	Fast analysis and data reuse
Capacity planning	Autonomous driving	Support for doctors	
Process optimization	Increase of safety	Reduction of costs	

General benefits

- Superior pattern recognition
- Automatic learning

Making Use of Big Data – a.k.a. Data Science

Empirical evidence

- Describe natural phenomena

Scientific theory

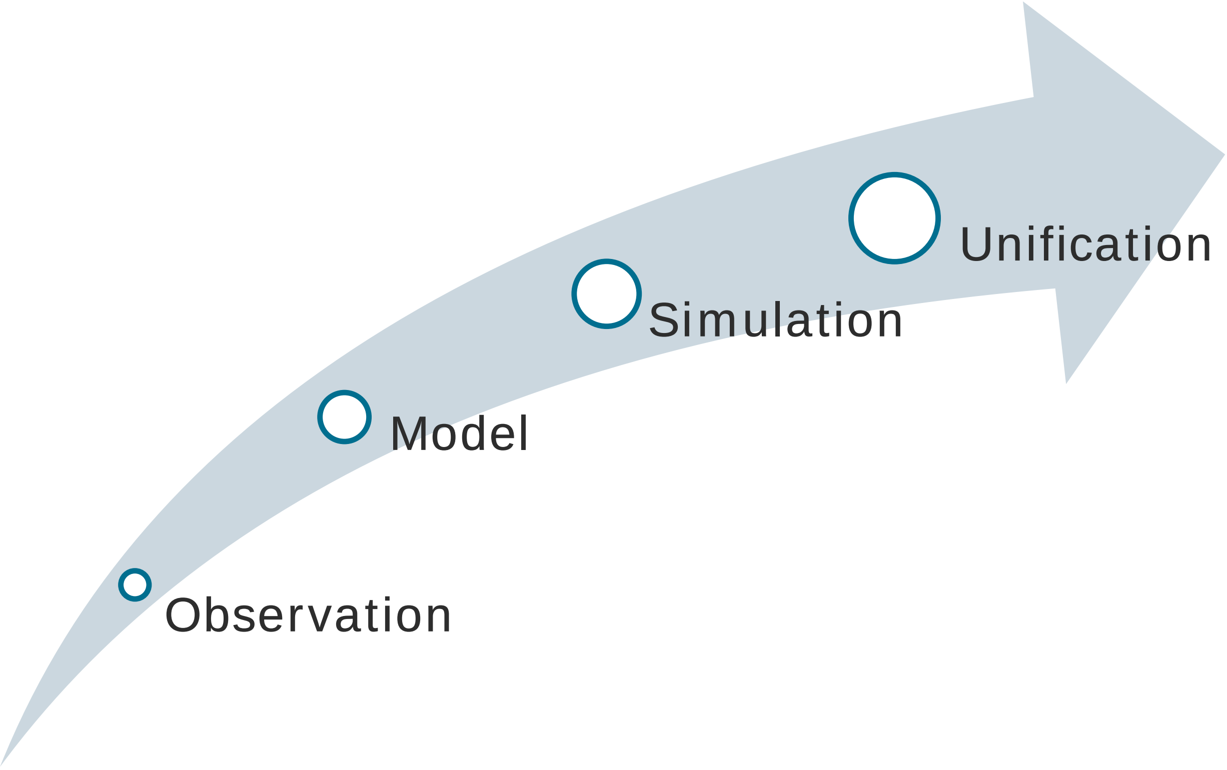
- Build theory and test it
- Relativity theory

Computational science

- Complex models and simulations
- Weather prediction

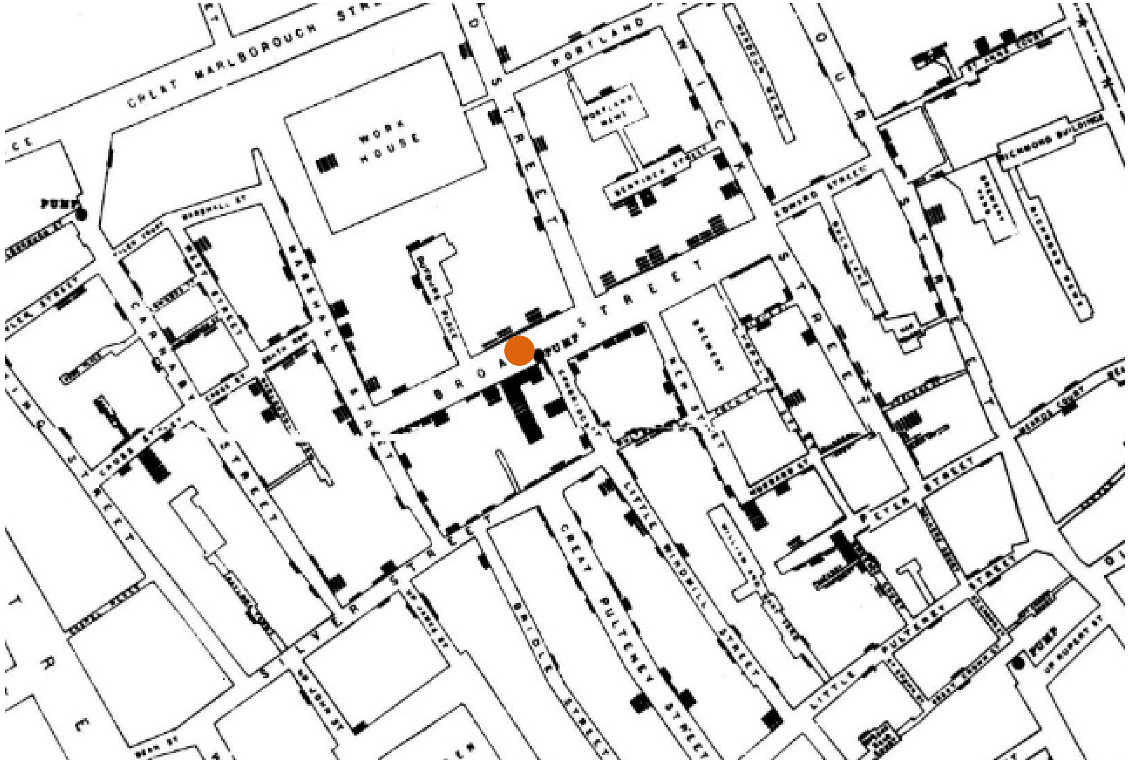
Data science

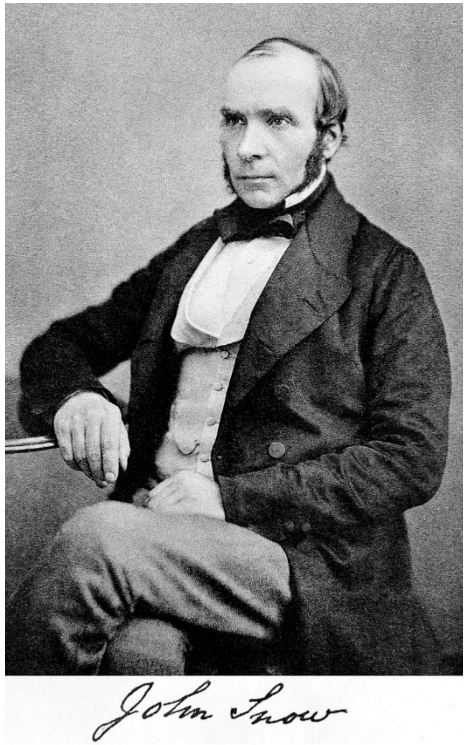
- Unify theory, experiment, and simulation
- Process big data by software and hardware



John Snow – 19th Century Data Scientist

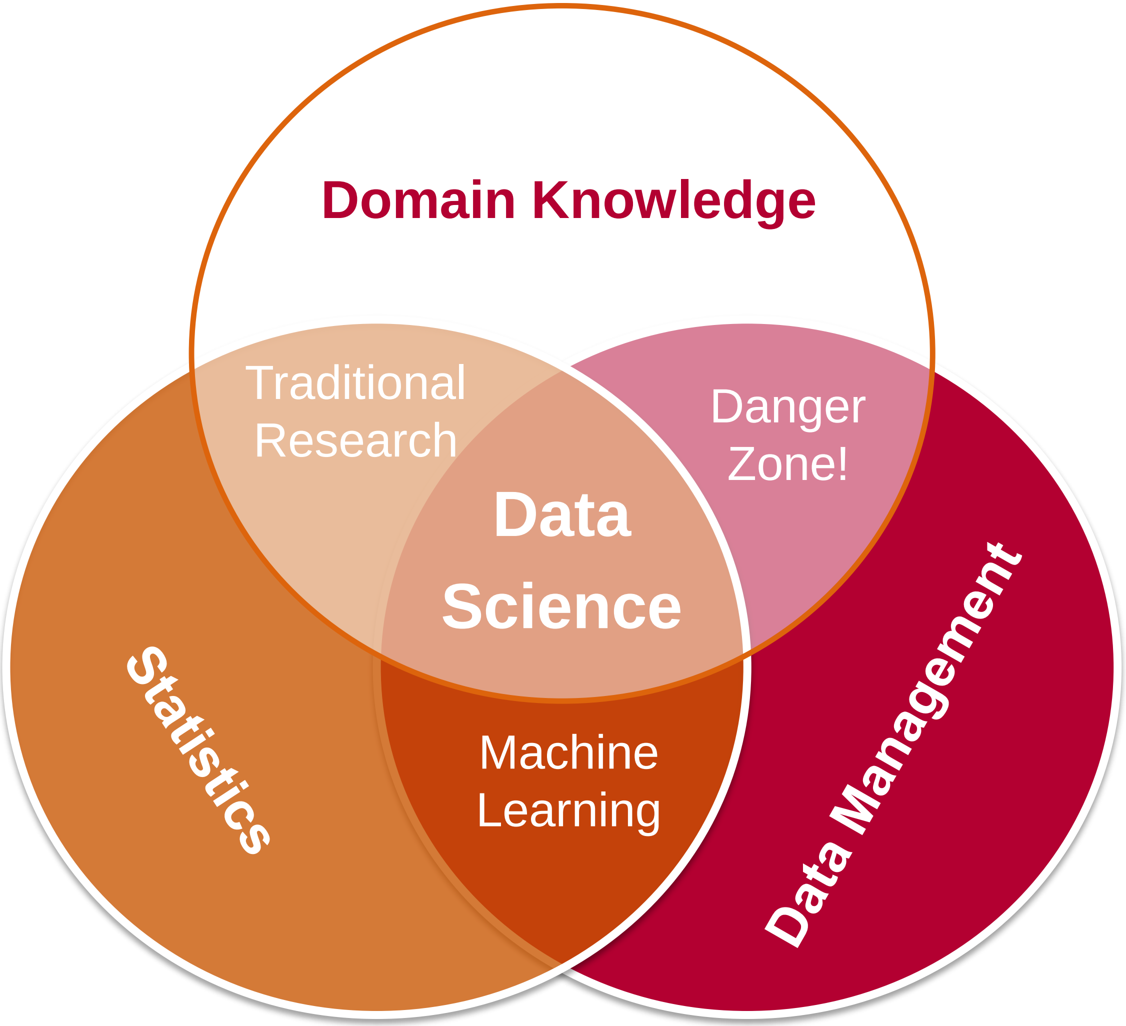
- Cholera in London 1850s
  - Miasma theory – bad air
- John Snow
  - Found connections to food and water
  - Investigated relation of water supply and Soho outbreak (1854)
  - Later discovered general problems with downstream Thames water supplies



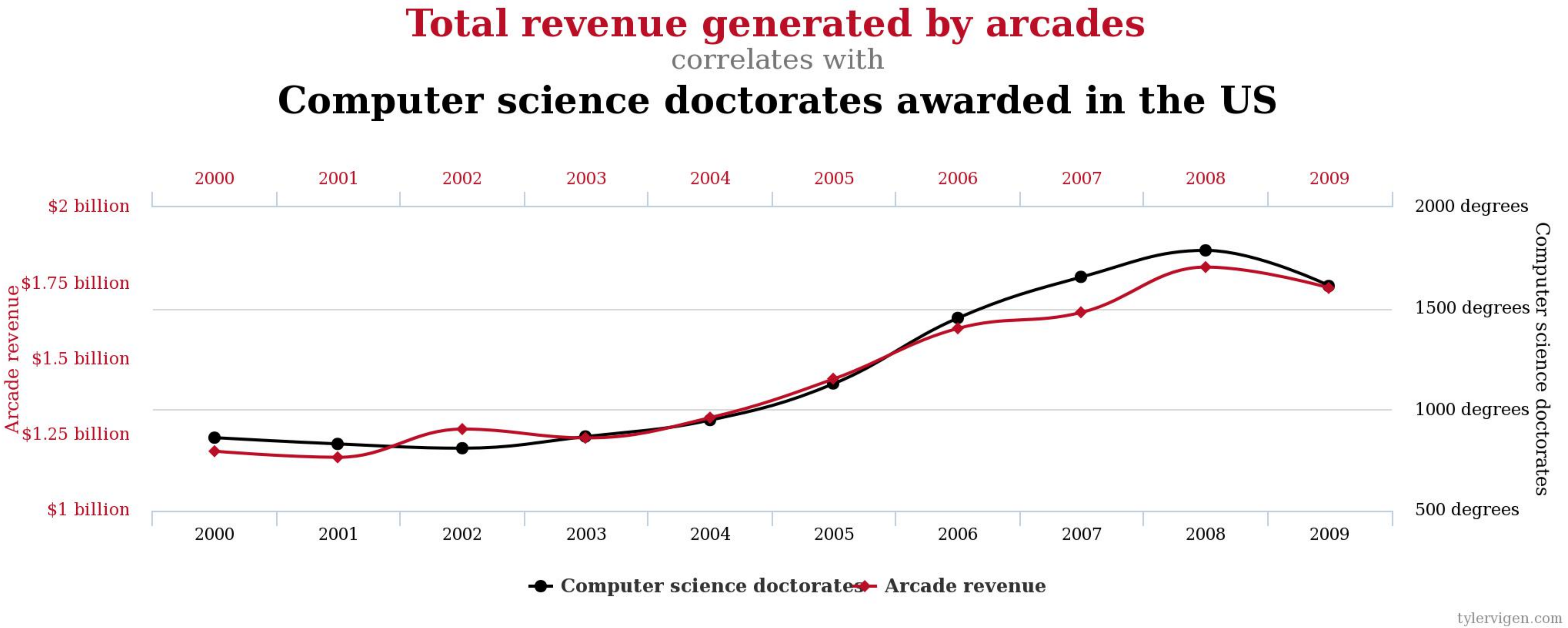

John Snow

Data Scientist Today

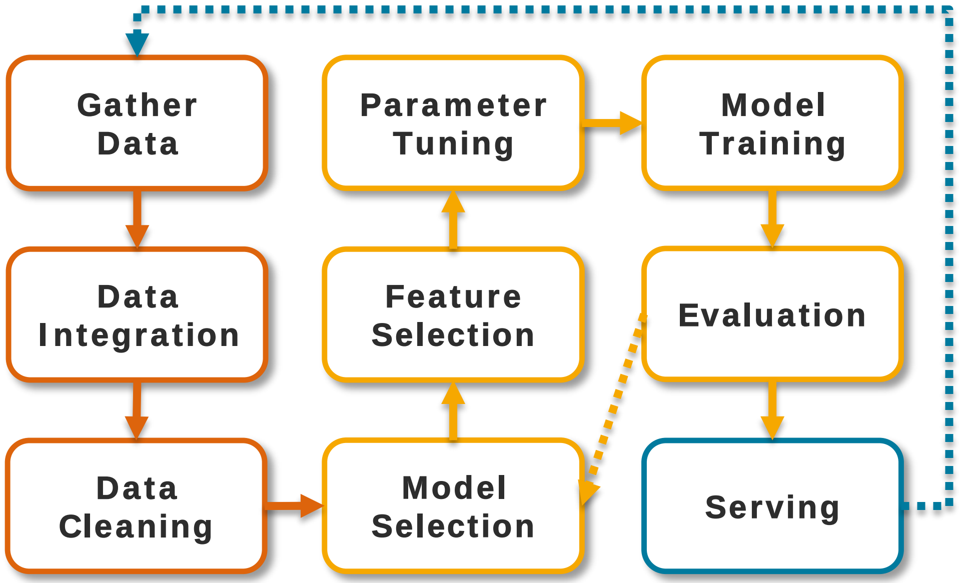
- Sexiest job of the 21st century
  - Requires rich skill set
  - Statistics, ML, CS, …
- Typical job description
  - Curate data
  - Explore data
  - Build ML models
  - Infer information
  - Solve business problem
  - Master‘s or Ph.D. in relevant field



Danger Zone!


Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$2 billion
$1.75 billion
$1.5 billion
$1.25 billion
$1 billion
Arcade revenue
2000 degrees
1500 degrees
1000 degrees
500 degrees
Computer science doctorates
Computer science doctorates
Arcade revenue
tylervigen.com

Data Science Pipelines



Gather Data

Data Integration

Data Cleaning

Model Selection

Feature Selection

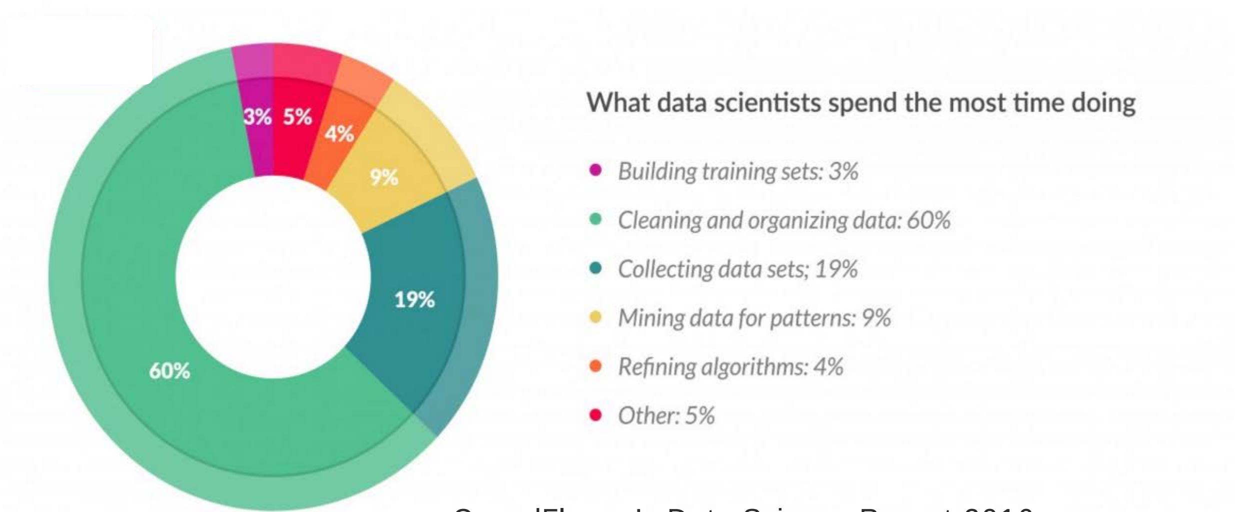
Parameter Tuning

Model Training

Evaluation

Serving

Data preparation *accounts for about 80% of the work of data scientists*



CrowdFlower's Data Science Report 2016

- Preprocessing, Training, Inference
- Highly iterative and repetitive
- One data scientist per use case
- Month(s) per use case

The Data Science Challenge

At least one data scientist per problem / use case

- Needs to understand the domain
- Needs to understand ML / programming / statistics

So many problems

- Every instance of a use case comes with different data
- Volume, variety, velocity, …

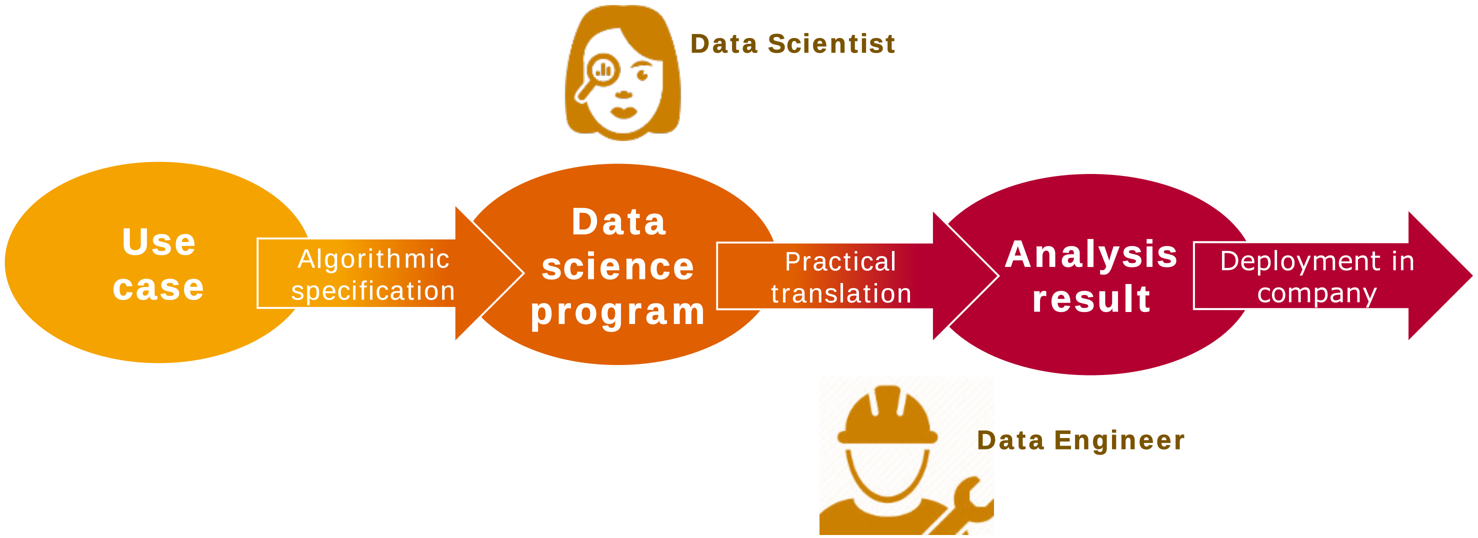
Too few people can do it!

Enter Data Engineering

- Data Science
  - □ Organize data
  - □ Analyze data
  - □ Solve business problem

- Data Engineering
  - □ Construct architecture
  - □ Maintain data pipelines
  - □ Productize data science

- Job openings in US (on LinkedIn)
  - □ Data Science 20K, Data Engineering 40K

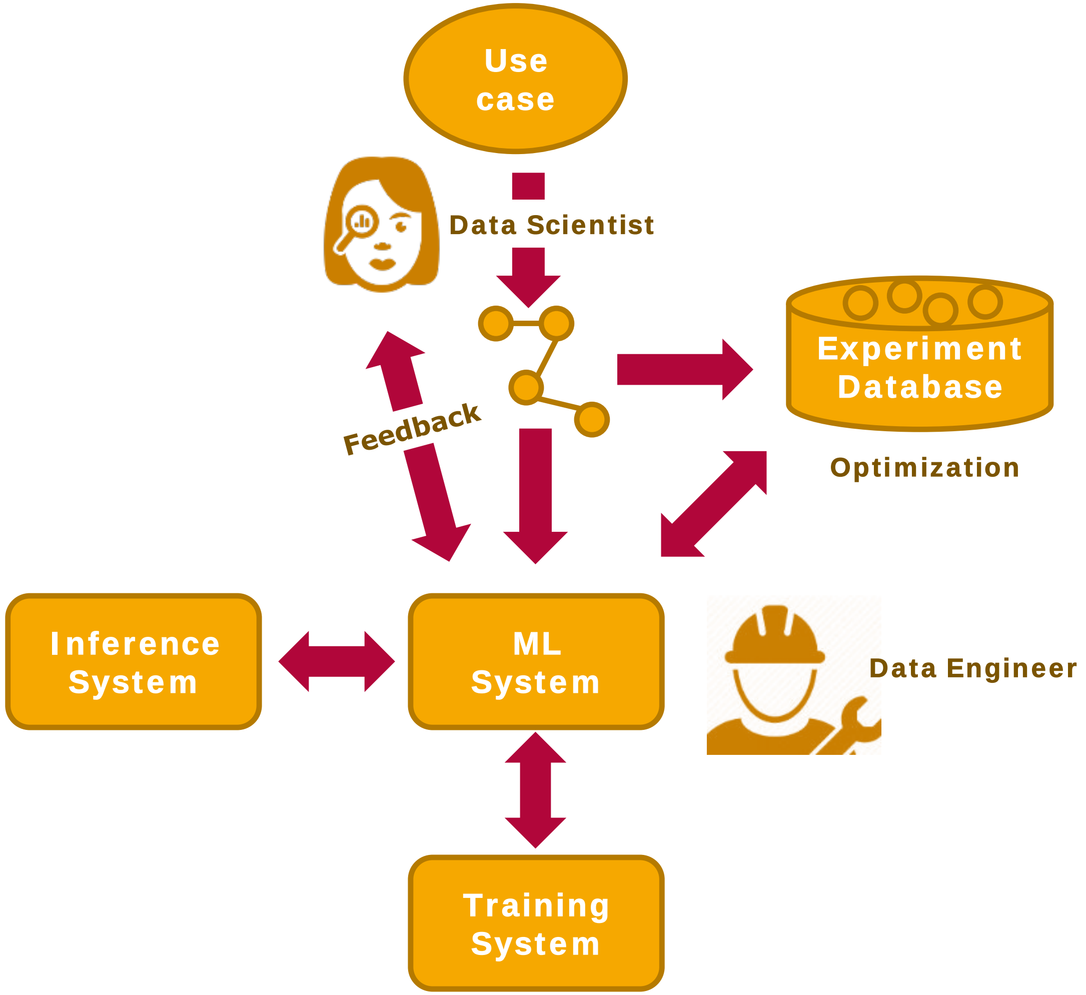


End-to-End Machine Learning (Data Engineering 2.0)

Decouple and formalize individual steps

Create common intermediate representations

- Tooling for preprocessing
- Experiment databases
  - Reuse knowledge
  - Simplify process
- Deployment systems
  - Manage models
  - Deploy training
  - Deploy inference



Data Engineering Systems

Systems that enable data engineering

- Database systems
- Stream processing systems
- Graph processing systems
- Machine learning systems
- ...

Systems that support data science

- Experiment databases
- Optimizers
- Deployment systems
- End-to-end ML

Ethical Considerations

- With great power comes great responsibility
- Not everything is a tool
  - True for systems, algorithms, data
- Everybody in/on the pipeline needs to be responsible, careful, and thoughtful
- Examples
  - Scoring / rating
  - Targeting
  - No built in safety switch

vs.

Course Logistics

DBT: Lecture

- Tuesdays, 11-12:30, HS 3
- Thursdays, 11-12:30, HS 2
  - (check full schedule for details)

Contact

- Prof. Dr. Tilmann Rabl
- e-mail: tilmann.rabl@hpi.de
- Office hours (by advance appointment via email)
  - Tuesdays, 13-14, F1.03
  - I have an open door policy (as long as I can)

Course Contents Overview

- Introduction
- Large Scale Data Processing
- Key Value Stores
- Stream Processing
- Machine Learning Systems
- Graph Processing
- Processing on Modern Hardware
- Performance Estimation/Analysis

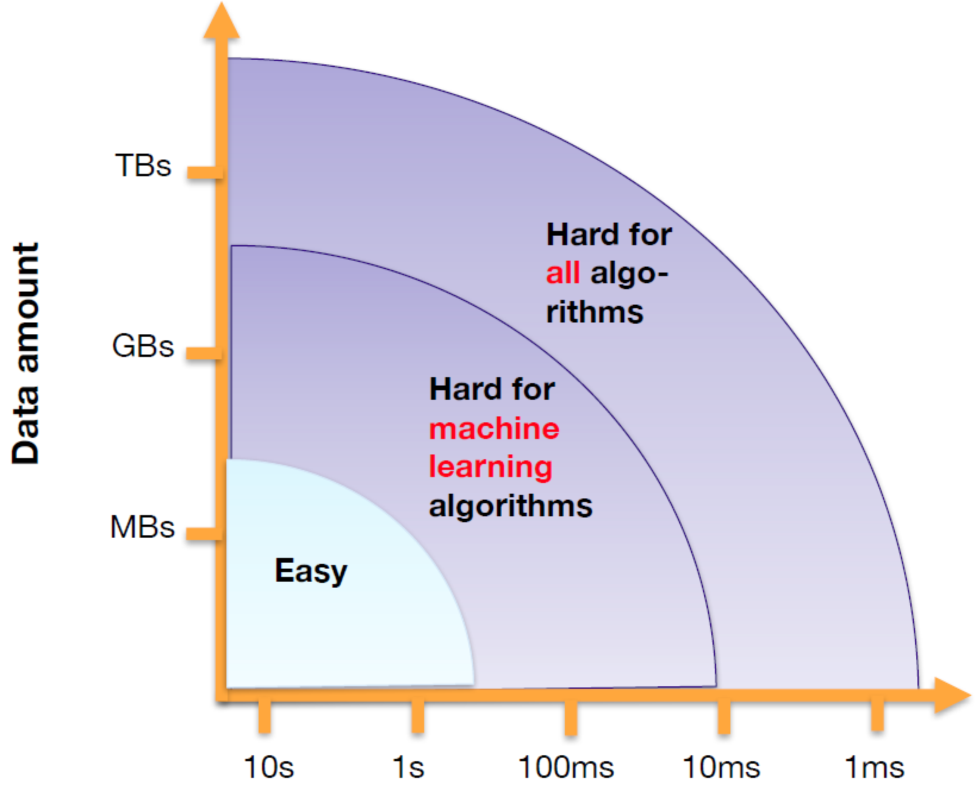


Image: Peter Pietzuch

Tentative Timeline

Date	Tuesday	Thursday
15./17.10.	Introduction	*No class*
22./24.10.	DBS Recap	DBS Recap II
29.10/31.10.	*20 Years HPI*	*Holiday*
5./7.11.	Big Data Stack	*Solution Quiz I*
12./14.11.	Benchmarking & Measurement	Cloud/Container
19./21.11.	*Modern Hardware*	File Systems
26. /28.11.	Map/Reduce	*Solution Quiz II*
3./5.12.	KV-Stores	Consistency
10./12.12.	Stream Processing	Windows
17./19.12.	Tables and State	*Solution Quiz III*

Tentative Timeline cont‘d

Date	Tuesday	Thursday
7./9.1.	Stream Optimizations	*Solution Quiz IV*
14./16.1.	ML Systems	ML Exec Strategies
21./23.1.	ML Lifecycle	Graph Processing
28./30.1.	Graph Processing II	*Solution Quiz V*
4./6.2.	Q&A	Final Exam

Grading in a Nutshell

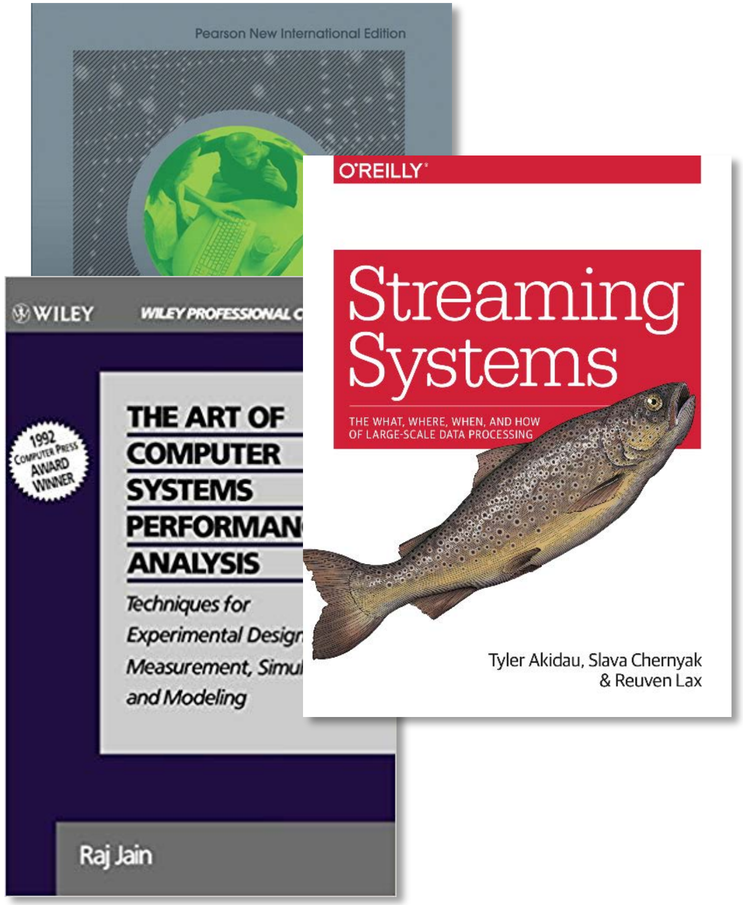
- 5 Exercise sheets
  - 1 self assessment
  - 4 graded exercises (5 points each)
  - 20% of total points

- Programming Exercises
  - November (7%)
  - January (8%)

- Exam
  - February (65%)

Literature

- Database Systems: The Complete Book, Hector Garcia-Molina, Jeff Ullman, and Jennifer Widom
- Streaming System, Akidau, Chernyak, Lax
- The Art of Computer Systems Performance Analysis, Jain
- More to be added…
- Will be available at Data Engineering Systems library soon! (Campus II, Building F, second floor)
- *My slides are usually self-containing!*


Pearson New International Edition
WILEY
WILEY PROFESSIONAL C
1992 Computer Press Award Winner
THE ART OF COMPUTER SYSTEMS PERFORMAN ANALYSIS
Techniques for Experimental Design Measurement, Simul and Modeling
Raj Jain
O'REILLY
Streaming Systems
THE WHAT, WHERE, WHEN, AND HOW OF LARGE-SCALE DATA PROCESSING
Tyler Akidau, Slava Chernyak & Reuven Lax

Slides

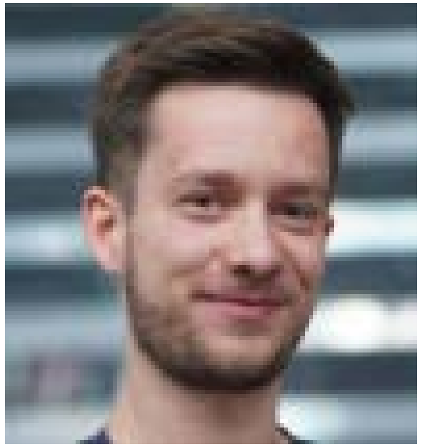
- I refine my slides until the day/hour before the lecture
- Slides are usually available shortly before the lecture (sometimes shortly after the lecture)
- If there are any errors, please send me **an e-mail**

Course Instructors

- Lawrence Benson

Additional support:

- Nina Ihde
- Maximilian Böther



Course Registration

- Lst. Friedrich Moodle
  - https://hpi.de/friedrich/moodle/course/view.php?id=66
  - Password: BigDataRocks!

- All slides, quizzes, resources
- Use the forum!

Code of Conduct

- Asking questions is greatly encouraged
  - Discuss questions with each other (except exams)
  - Submit homework individually, but feel free to discuss

- The Limits of collaboration
  - Do not just share solutions with each other - explain your solutions
  - Plagiarism, copying or other forms of dishonesty will result in failing the course

- Communication
  - Learn how to write professional emails: https://medium.com/@lportwoodstacer/how-to-email-your-professor-without-being-annoying-af-cf64ae0e4087

- Generally
  - Treat everyone with respect and consideration

- Disclaimer: This is a new course, by a new group, please be patient. ☺

Other Courses

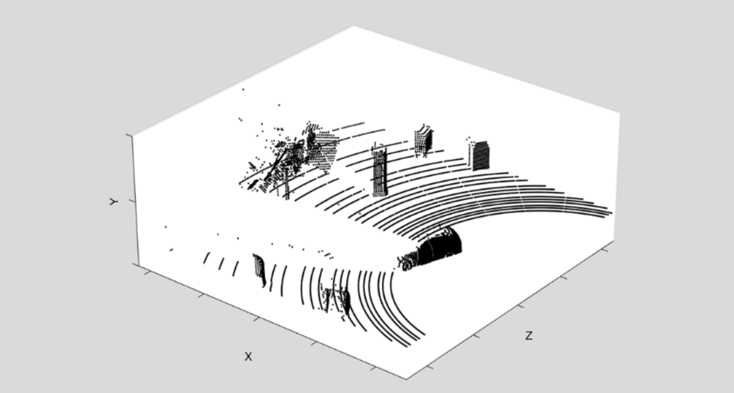
- Practical Data Engineering
  - Lecture series with practitioners and researchers
  - Tuesdays, 17:00 – 18:30, HS 2
  - Master and Bachelor, 3 ECTS

- Foundations of Database Systems
  - Seminar on basic DBS topics
  - Thursdays, 13:30 – 15:00, F-1.04
  - Bachelor, 3 ECTS

- Projects & Theses
  - Topics on Data Engineering Systems available

DEBS Grand Challenge 2020

- Event / Stream Processing Challenge
  - Since 2011, open to everyone
  - Yearly contest, https://debs.org/grand-challenges/
  - Opportunity to compete and work with interesting data sets
  - Current contest not announced yet (~ Dec)



- Interested Students
  - Can contact me for mentoring
  - Finalists attend DEBS 2020 in Montreal, Canada
    (we pay whatever is not covered by travel stipend)
  - Could be a great start into a research career and opportunity for networking

SIGMOD Programming Contest

- SIGMOD Programming Contest
  - Since 2009, student teams of degree-granting institutions
  - Yearly contest, see last year https://sigmod19contest.itu.dk/index.shtml
  - Opportunity to compete and learn DB internals (last year: sorting)
  - Current contest not announced yet (~ End Feb – End May)

- Interested Students
  - Can contact me for mentoring
  - Finalists attend SIGMOD 2020 in Portland, USA
    (we pay whatever is not covered by travel stipend)
  - Could be a great start into a research career and opportunity for networking

Thank you for your attention!

- Questions?

- Thursday no class!

- Next week:
  - DB Recap!