Big Data Analytics

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Outline

1. What is Big Data?
2. Overview
3. Organizational Stuff

Outline

What is Big Data?

What is Big Data?

"Big data is like teenage sex:

— Dan Ariely

What is Big Data?

"Big data is like teenage sex:
everyone talks about it,

— Dan Ariely

What is Big Data?

"Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,

— Dan Ariely

What is Big Data?

"Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,

— Dan Ariely

What is Big Data?

"Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it."

— Dan Ariely

What is Big Data?

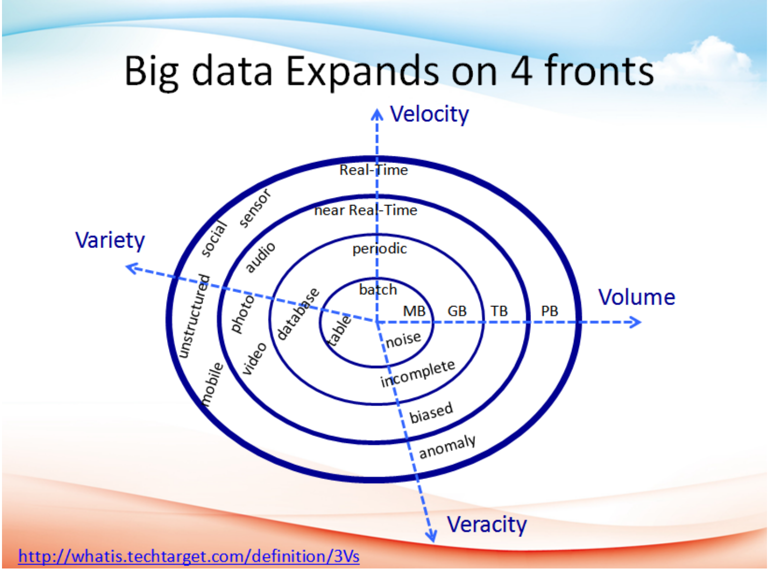
Some definitions:

- "data sets that are so **voluminous and complex** that traditional data processing application software are inadequate to deal with them."
  [http://en.wikipedia.org/wiki/Big_data]

- "Big data is **high-volume, high-velocity** and/or **high-variety** information assets that demand cost-effective, innovative forms of information processing for **enhanced insight and decision making**."
  [www.gartner.com/it-glossary/big-data/]

Note: The "3 Vs" go back to Laney [2001]. Often a 4th V "veracity" and a 5th V "value" is used.

Big Data — Dimensions (the "4 Vs")



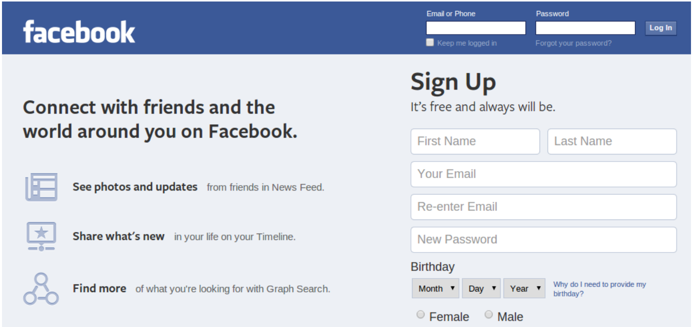
What is Big Data?

Big Data is about:

- Storing and accessing
  - large amounts of
  - (complex/unstructured) data
- Processing high volume data streams
- Making sense of the data
- Making predictions based on the data

Note: Unstructured data in this context means, data that is not already a vector. Some of this data, e.g., relational data and graph data, confusingly often also is called "structured".

Where to find Big Data?



- 1.52 billion daily active users
  (2.32 billion monthly active users, Dec. 2018)
- size of user data stored by Facebook: 300 Petabytes
- average amount of data that Facebook takes in daily: 600 terabytes
- size of Facebook's graph search database: 700 Terabytes

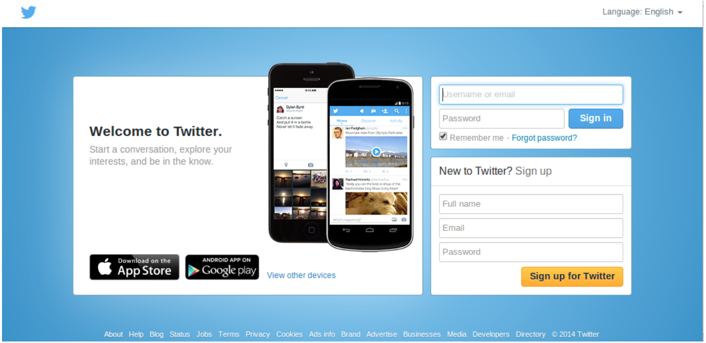
[source: https://newsroom.fb.com/company-info/; online source for points 2–4 vanished]

Where to find Big Data?

- 3.3 billion searches per day (on average)[1]
- 30 trillion unique URLs identified on the Web[1]
- 20 billion sites crawled a day[1]
- In 2008 Google processed more than 20 Petabytes of data per day[2]

[1] http://searchengineland.com/google-search-press-129925 (2012)
[2] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.

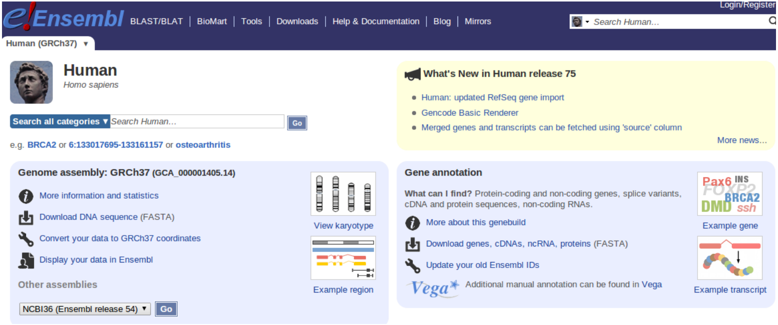
Where to find Big Data?



- tweets per day: 58 million[1]
- Twitter search engine queries per day: 2.1 billion[1]
- registered/active Twitter users: 695 million / 342 million[1]

[[1] http://www.statisticbrain.com/twitter-statistics/] (9/2016)

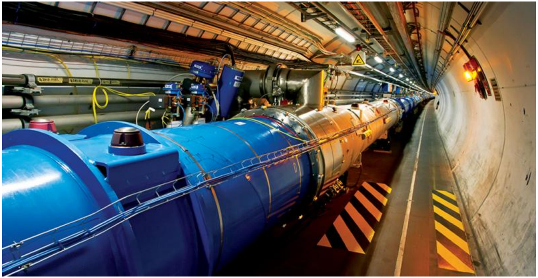
Where to find Big Data?



- Ensembl database contains the genome of humans and 50 other species
- "only" 250 GB[1]

[1] `http://www.ensembl.org/`

Where to find Big Data?



- CERN Large Hadron Collider has collected data from over 300 trillion proton-proton collisions
- Approx. 25 Petabytes per year

Big Data — Public Datasets

1000 Genomes Project	DNA of 1700 humans	200 TB
Common Crawl Corpus	5G web pages	81 TB
Wikipedia / Freebase	1.9G subject/predicate/object triples	250 GB
Million Song Dataset	audio features of 1M songs	280 GB
OpenStreetMap	a map of earth	90 GB
2000 US Census	US census data	200 GB
PubChem library	biological activities of small molecules	230 GB
NCDC weather data	daily measurements from 9000 stations	20 GB
Open Library	metadata of 20M books	7 GB
Twitter	1.6G tweets	0.6 GB

CD 700 MB, DVD 4.7–17 GB, Blu-ray 25–100 GB, hard disc: 10 TB.

How Large is 1 Petabyte

- 1 PB = 1000 TB = 10^{15} B

How Large is 1 Petabyte

- 1 PB $=$ 1000 TB $= 10^{15}$ B
- 35.7M years counted one byte per second,

How Large is 1 Petabyte

- 1 PB $=$ 1000 TB $=$ 10^{15} B
- 35.7M years counted one byte per second,
- 254 years listening to music stored in CD quality (500MB/h)

How Large is 1 Petabyte

- 1 PB = 1000 TB = 10^{15} B
- 35.7M years counted one byte per second,
- 254 years listening to music stored in CD quality (500MB/h)
- 25 years watching DVDs
  - 223.000 DVDs (a 4.7 GB)

How Large is 1 Petabyte

- 1 PB $= 1000$ TB $= 10^{15}$ B
- 35.7M years counted one byte per second,
- 254 years listening to music stored in CD quality (500MB/h)
- 25 years watching DVDs
  - 223.000 DVDs (a 4.7 GB)
  - but there are only 74.000 TV movies on IMDB !
    (1.8M including TV episodes)

How Large is 1 Petabyte

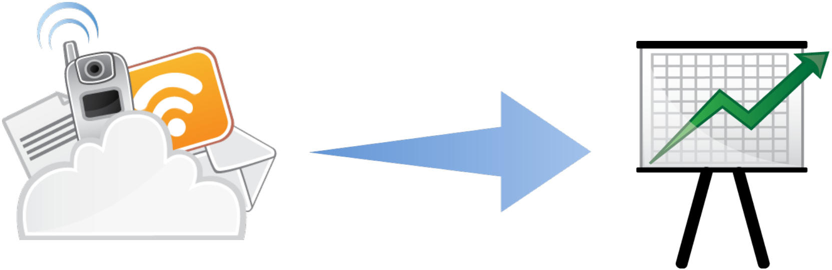
- 1 PB = 1000 TB = 10^{15} B
- 35.7M years counted one byte per second,
- 254 years listening to music stored in CD quality (500MB/h)
- 25 years watching DVDs
  - 223.000 DVDs (a 4.7 GB)
  - but there are only 74.000 TV movies on IMDB !
    (1.8M including TV episodes)
- can be stored on 100 harddisks à 10 TB/300 € (30,000 €)

How Large is 1 Petabyte

- 1 PB = 1000 TB = 10^{15} B
- 35.7M years counted one byte per second,
- 254 years listening to music stored in CD quality (500MB/h)
- 25 years watching DVDs
  - 223.000 DVDs (a 4.7 GB)
  - but there are only 74.000 TV movies on IMDB !
    (1.8M including TV episodes)
- can be stored on 100 harddisks à 10 TB/300 € (30,000 €)
- 96 days to read from standard harddisks sequentially (1030 MBits/s)

What to do with Big Data?

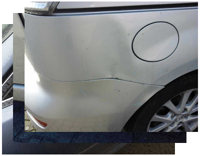
We do not want to know things but to understand them!

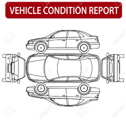


What to do with Big Data? Leasing Vehicle Return @VWFS

initial data generating process:

expert

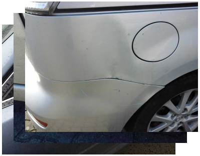




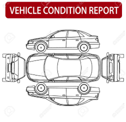
240 €

target process:

customer



automatic service

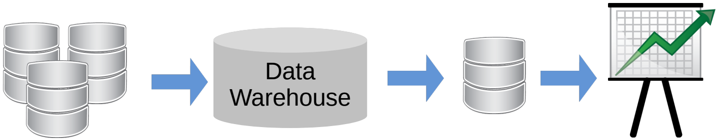


240 €

What to do with Big Data? - Case Studies

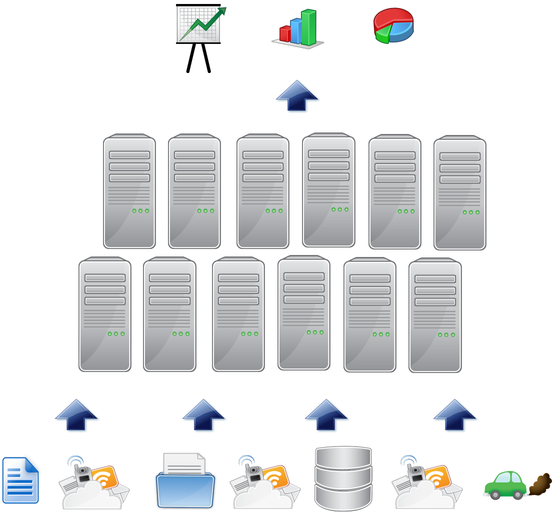
- **T-Mobile USA:**
  - Integrated Big Data across multiple IT systems to combine customer transaction and interaction data in order to better predict customer defections
  - By leveraging social media data along with transaction data from CRM and billing systems, customer defections have been cut in half in a single quarter.
- **US Xpress:**
  - Collects data elements ranging from fuel usage to tire condition to truck engine operations to GPS information
  - Optimal fleet management
- **McLaren's Formula One racing team**:
  - Real-time car sensor data during car races
  - Real-time identification of issues with its racing cars

What to do with Big Data? - The BI Approach



- Static databases
- Structured data
- Centralized approaches

What to do with Big Data?



- Heterogeneous data sources
- Unstructured data
- Data streams
- Massive Parallelism

What to do with Big Data?

Application examples:

- Online personalized advertising
- Sentiment analysis and behavior prediction
- Detecting adverse events and predicting their impact
- Automatic Translation
- Image Classification and object recognition
- Intelligent public services

Challenges Posed By Big Data

- Effectively **store and query** large amounts of data in a distributed environment
- **Parallel and distributed execution environments** / programming models
- **Distributed and scalable machine learning techniques** to learn from the data
- **Distributed and scalable data visualization techniques**

Outline

Technology Stack

Part D

Distributed Machine Learning Algorithms

Part C

Distributed Execution Environments

Part B

Distributed Storage

Part A

Parallel/Distributed Computing

Technology Stack

Part D

Distributed Machine Learning Algorithms

Part C

Distributed Execution Environments

Part B

Distributed Storage

Part A

Parallel/Distributed Computing

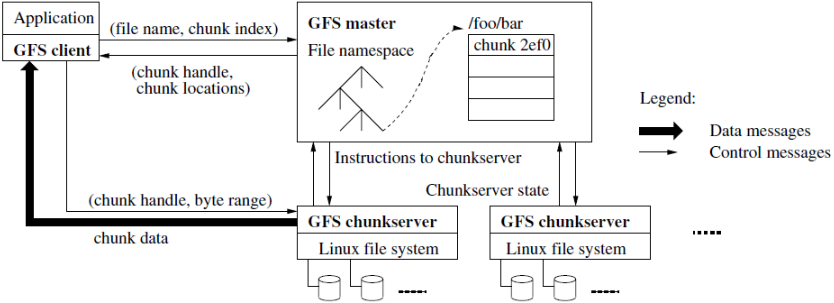
Storing

In a distributed environment the data storing mechanisms should address the following issues

- Parallel Reading and Writing
- Data node Failures
- High Availability

Distributed File Systems

The Google File System Architecture



Databases

Databases are needed for

- Querying and indexing
- transaction processing

State-of-the-art: Relational Databases

For processing big data one needs a database which:

- Supports high level of parallelism
- Supports analytical processing
- Has a flexible data model to deal with unstructured data sources

Databases for Big Data - NoSQL

NoSQL - “Not only SQL”

- Wide variety of database technologies
- Dynamic Schema
- sharded indexing
- horizontal scaling
- support columnar storage

NoSQL Databases

Technology Stack

Part D

Distributed Machine Learning Algorithms

Part C

Distributed Execution Environments

Part B

Distributed Storage

Part A

Parallel/Distributed Computing

Accessing

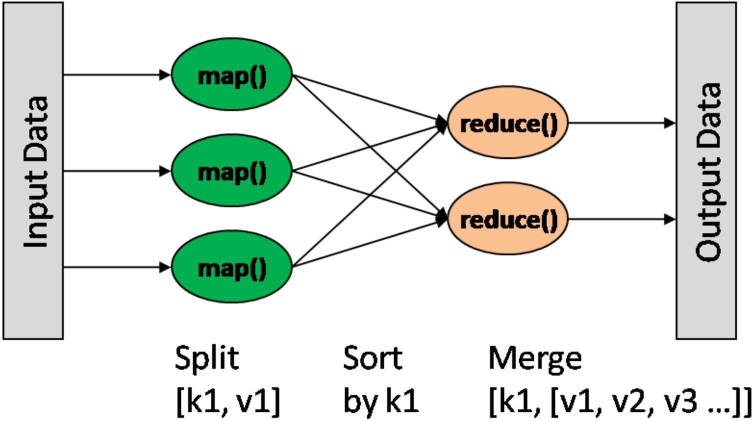
A **distributed execution environment** / **computational model** is needed to:

- Provide a set of useful computational primitives
- Hide the complexity of distributed and parallel programming
- Ensure Fault Tolerance

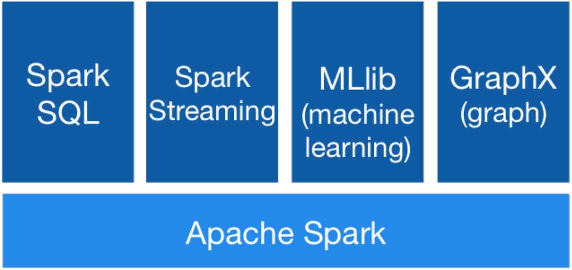
Examples:

- MapReduce
- GraphLab
- Pregel
- Apache Spark
- TensorFlow

MapReduce


Input Data
map()
map()
map()
reduce()
reduce()
Output Data
Split
[k1, v1]
Sort
by k1
Merge
[k1, [v1, v2, v3 ...]]

Apache Spark



Technology Stack

Part D

Distributed Machine Learning Algorithms

Part C

Distributed Execution Environments

Part B

Distributed Storage

Part A

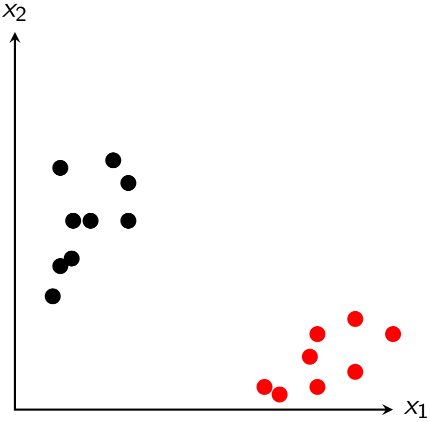
Parallel/Distributed Computing

Learning from the data

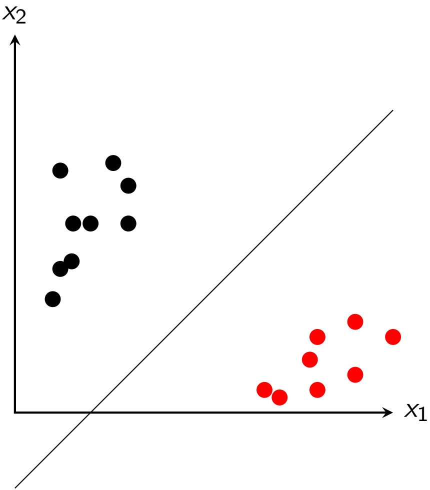
- Linear and Non Linear Models for classification and regression
  - Scalable learning algorithms (e.g. Stochastic Gradient Descent)
  - Distributed Learning Algorithms (e.g. ADMM)

- Models for Link Prediction and link analysis
  - Factorization models
  - Distributed Learning Schemes (e.g. NOMAD, FPSGD)

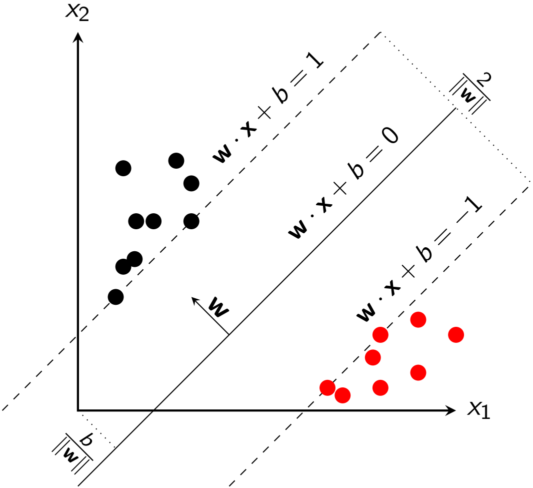
Classification



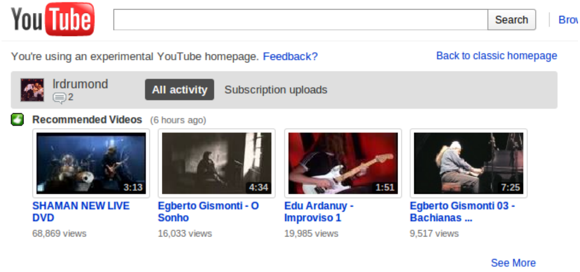
Classification



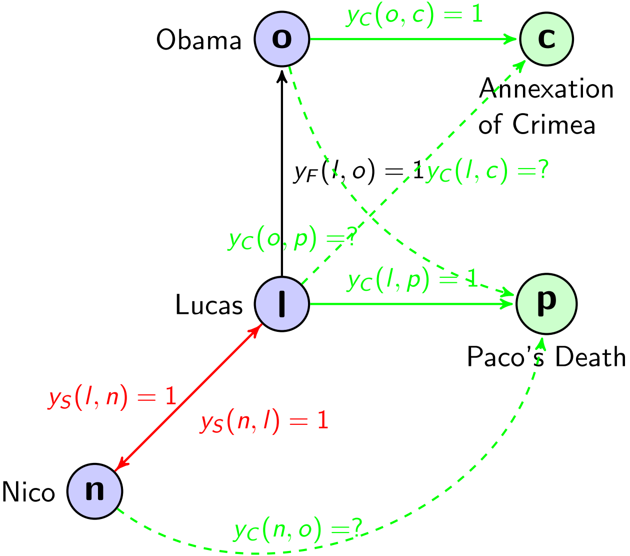
Classification



Recommender Systems



Graph Analysis



Syllabus

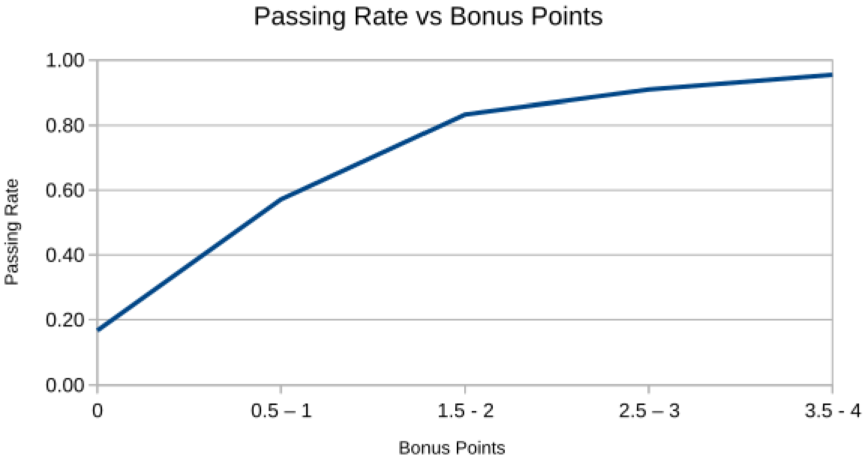
Tue.	9.4.	(1)	0. Introduction
			A. Parallel Computing
Tue.	16.4.	(2)	A.1 Threads
Tue.	23.4.	(3)	A.2 Message Passing Interface (MPI)
Tue.	30.4.	(4)	A.3 Graphical Processing Units (GPUs)
			B. Distributed Storage
Tue.	7.5.	(5)	B.1 Distributed File Systems
Tue.	14.5.	(6)	B.2 Partioning of Relational Databases
Tue.	21.5.	(7)	B.3 NoSQL Databases
			C. Distributed Computing Environments
Tue.	28.5.	(8)	C.1 Map-Reduce
Tue.	4.6.	—	— *Pentecoste Break* —
Tue.	11.6.	(9)	C.2 Resilient Distributed Datasets (Spark)
Tue.	18.6.	(10)	C.3 Computational Graphs (TensorFlow)
			D. Distributed Machine Learning Algorithms
Tue.	25.6.	(11)	D.1 Distributed Stochastic Gradient Descent
Tue.	2.7.	(12)	D.2 Distributed Matrix Factorization
Tue.	9.7.	(13)	Questions and Answers

Outline

Exercises and tutorials

- There will be a weekly sheet with two exercises handed out **each Tuesday, 12pm** after the lecture.
  - 1st sheet will be handed out Thur. 11.4. (exception), 2nd on Tue. 23.4.
- Solutions to the exercises can be submitted until **next Monday 8:00 am**.
  - 1st sheet is due Thur. 18.4., 8 am (exception), 2nd on Mon. 29.4., 8 am.
- Exercises will be corrected
- Tutorials
  - each **Tuesday 8-10** (beginners). — 1st tutorial at Tuesday 16.4.
  - each **Thursday 14-16** (advanced). — 1st tutorial at Thursday 18.4.
- Successful participation in the tutorial gives up to 10% bonus points for the exam.
  - group submissions are OK (but will yield no bonus points)

Attend and Work the Tutorials!



[ML exam 2018/19]

Exams and credit points

- There will be a written exam at the end of the term
  - 2h, 4 problems
- The course gives 6 ECTS
- The course can be used in
  - International Master in Data Analytics / Obligatory
  - Angewandte Informatik MSc. / Informatik / Gebiet KI & ML
  - IMIT MSc. / Informatik / Gebiet KI & ML
  - Wirtschaftsinformatik MSc / Informatik / Gebiet Business Intelligence

Some books

- Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman (2014):
  Mining of massive datasets,
  Cambridge University Press.
  Available online:
  `http://infolab.stanford.edu/~ullman/mmds.html`

- Gautam Shroff (2014):
  The Intelligent Web: Search, smart algorithms, and big data,
  Oxford University Press.

References

Doug Laney. 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 2001.