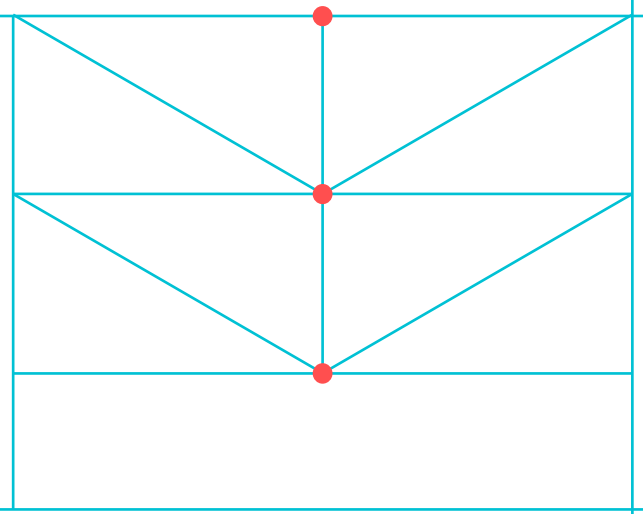
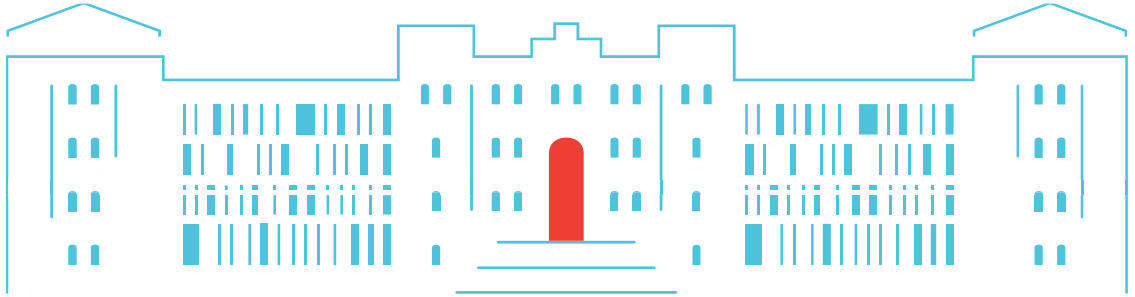
Lecture 01: Introduction
Big Data

institute for data engineering

TUHH
Hamburg University of Technology



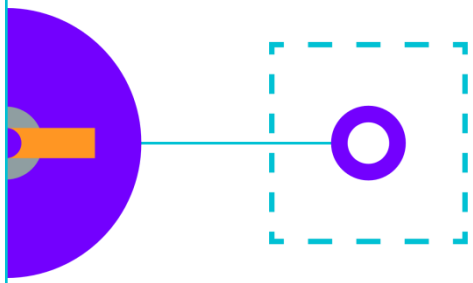


Prof. Dr.-Ing. Stefan Schulte
Institute for Data Engineering
stefan.schulte@tuhh.de

Goals of this Course

- Decribe basic concepts and characteristics of the “Big Data Stack”, including MapReduce, NoSQL Databases, Stream Processing, as well as according technologies.
- Discuss and assess the benefits and drawbacks of Big Data technologies
- Select and apply Big Data software for particular application areas
- Design and develop practical solutions

Agenda for Today



1. Lecture and Lab Overview and Organizational Information
2. Overview of the Topics
   - Course Content Overview
   - Big Data: Motivation and Basic Characteristics
   - Data Engineering

Lecturers

Team:

- Lecture: Stefan Schulte
- Lab: Nisal Hemadasa

Teaching Approach

- Language: Entirely in English
- In general, on-site lectures
  - No on-site lecture on June 27th (instead: lecture recording)
- Recordings of the lectures (but not of the exercises)
  - These may go wrong!
- Utilization of classroom response tools from time to time (Particify and/or EduVote)
- Brief break (2-3 minutes) after 45 minutes
- Interrupt me whenever you think that's a good idea

Lecture Overview (Tentative)

- Lecture 1 (Today): General Introduction
- Lecture 2: Lab kickoff, Introduction to Big Data Processing
- Lectures 3-5: MapReduce
- Lectures 5-9: Key-Value Stores / NoSQL
- Lectures 9-11: Stream Processing
- Lectures 12-13: Still open

Literature

- Unfortunately, there is no single textbook that covers the whole lecture
  - M. Kleppmann, *Designing Data-intensive Applications,* O'Reilly Media, Inc., 2017.
  - T. Akidau, S. Chernyak, and R. Lax, *Streaming Systems*, O'Reilly Media, Inc., 2018.
- Literature will be provided throughout the semester

Dates and Rooms

- Lectures: Thursday, 16:45-18:15, A-1.15
- Tutor Consultation Hours: Thursday, 11:30-13:00 (D-2.022) or 13:15-14:45 (D-2.022), starting on April 18th
- Fixed consultation hours per team, i.e., either the 11:30 or 13:15 slot on Thursday
- There are some exceptions, so please check the schedule in Stud.IP. If there are changes in the semester, there will also be an announcement.

Lab Overview

- Learning by doing
- Projects from the field of Big Data
  - This semester, focus on data stream processing
  - Topic descriptions will be released on Monday (April 8th)
- Team size: 4 students
- Expected outcomes:
  - Project solution (software prototype)
  - Two presentations

Lab Overview II

- The lab is the “PBL” part of the module, and therefore makes up 50% of the overall grade, i.e., you cannot pass the module without doing the lab
  - The other half of the grade is coming from the end-of-semester written exam
- But of course, the main aspect is to learn how to work with modern Big Data technologies
- Decide until the kickoff meeting (April 11th) if you want to do it or not
  - Check the topic descriptions provided in Stud.IP
  - But please do not drop out during the semester

Lab Overview III

- If you really want to do the lab exercise: Sign up for a consultation hour in Stud.IP
  - Start: April 7th, 9:00
  - End: April 9th, 23:59
  - Topics will be distributed at the kickoff meeting next week
- Lab meetings:
  - Kickoff meeting: Next week, as part of the lecture; if there are any things to be clarified, we'll do it there
  - Interim presentations: As part of the consultation hours. **Attendance is mandatory for members of the presenting teams.**
  - Final presentations: July 11th, 10:00 to 15:00. **Attendance is mandatory for all!**

General Information: Stud.IP

Content:

- Announcements
- Slides (“Files”)
- Dates and rooms (“Schedule”)
- Content (especially the lab topics), exam regulations and grading (to be added), references, …
- Lecture recordings (“Media Links”)
- Forum

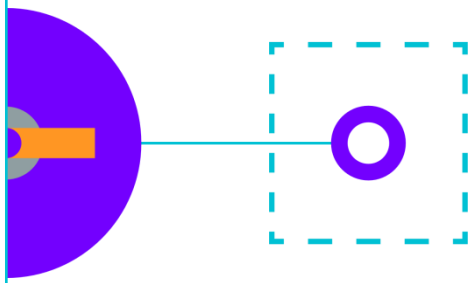
Questions/requests/comments:

- If the answer could be interesting to more students, please use the forum in Stud.IP!
- If there are further questions w.r.t. the lab, nisal.hemadasa@tuhh.de
- If there are further questions w.r.t. the lecture, contact stefan.schulte@tuhh.de

Code of Conduct

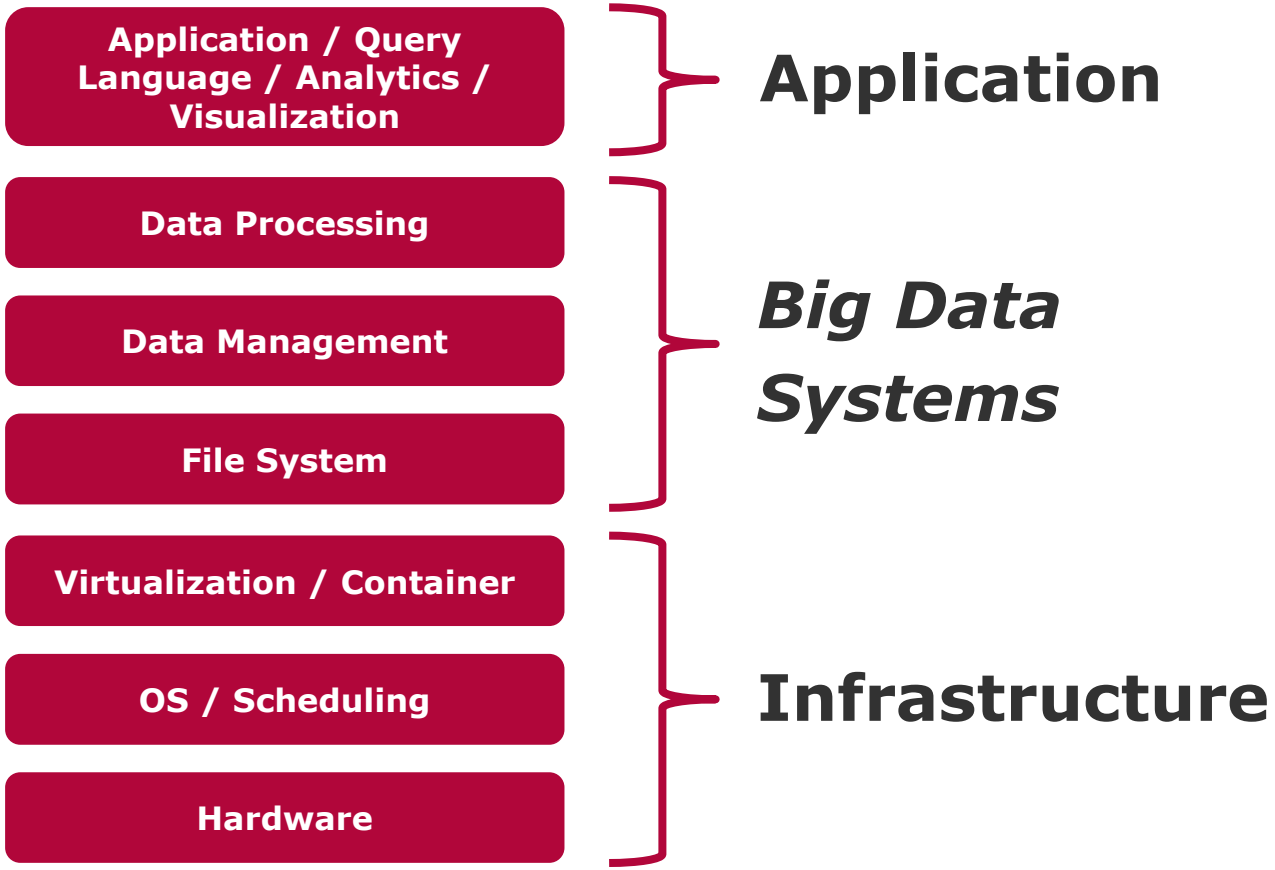
- Asking questions is greatly encouraged
  - Discuss questions with each other (except in the exams)
  - Submit lab exercises individually (i.e., per team), but feel free to discuss with other teams
- The limits of collaboration
  - Do not just share solutions with each other - explain your solutions
  - Plagiarism, copying, or other forms of dishonesty **will result in failing the course**
- Communication
  - Write professional, polite emails
  - Use netiquette in forum, email, chats, etc.

Agenda for Today



1. Lecture and Lab Overview and Organizational Information
2. Overview of the Topics
   - Course Content Overview
   - Big Data: Motivation and Basic Characteristics
   - Data Engineering

Course Contents Overview



Big Data Stack

Big Data

Data is growing
Messages, tweets, social networks (statuses, check-ins, **shared content**),
blogs, click streams, various logs, …

- *Facebook: > 3B monthly active users,*
  > 100B messages/day (Source: Statista, Meta)
- *Instagram: > 2B monthly active users, 66K photos*
  and videos shared per minute (Source: Statista, LocaliQ)

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, e.g., (object-) relational database management systems

- 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 vs. Small Data

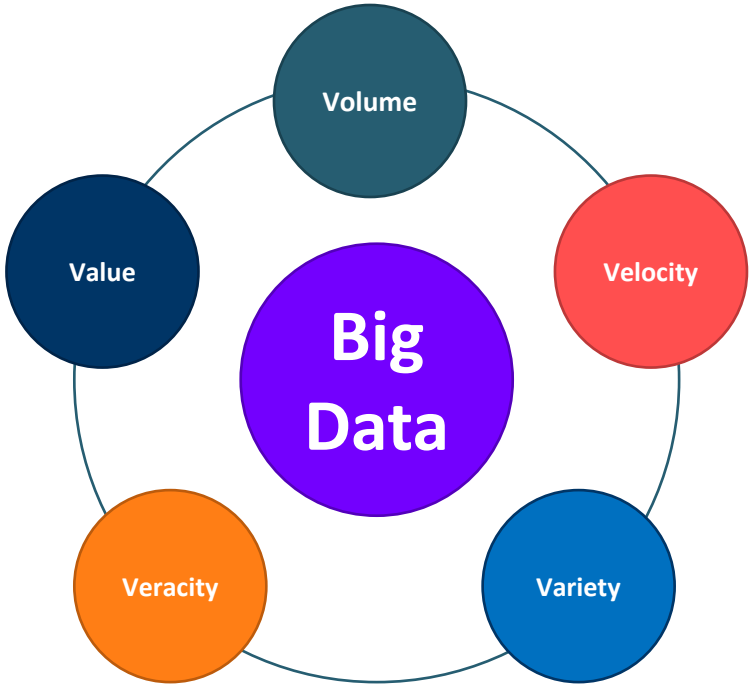
Small Data:

- Small enough for human inference
- Accumulated slowly
- Relatively consistent, and (semi-) structured
- Mostly located in storage systems within organizations and data centers

Big Data:

- Generated in huge volumes, and could be structured, semi-structured, or unstructured
- Needs processing to generate insights for human consumption
- Arrives continuously at enormous speed from multiple sources
- Comprises any form of data including video, photos, and more
- Distributed in the cloud or a server farm

Big Data: 5 Vs



Velocity

- Data that is generated very fast
  - Data generation rate
  - Data generation never stops
  - Data arrival rate
  - How quick do we have to act on the data?
- Drivers:
  - Improved connectivity and hardware
  - Rapid response times needed in many settings

Volume

- Scale of the data
- Increased amount of stored data: With respect to the number of observations (size of the data) and the number of variables (dimensionality of the data).
- Drivers:
  - Increase in data sources
  - Higher resolution sensors
  - Scalable hardware infrastructure

Variety

- Data that comes from machines, people, and processes
- “Data in many forms”: Structured, semi-structured, and unstructured data
- Drivers:
  - Mobile technologies
  - Scalable infrastructure
  - Efficient storage and retrieval

Veracity

- “Data in doubt”
- Quality, origin, and conformity of facts
- Accuracy of data: Processing errors, noise
- Drivers:
  - Robust ingestion
  - ETL mechanisms (Extract, transform, load)

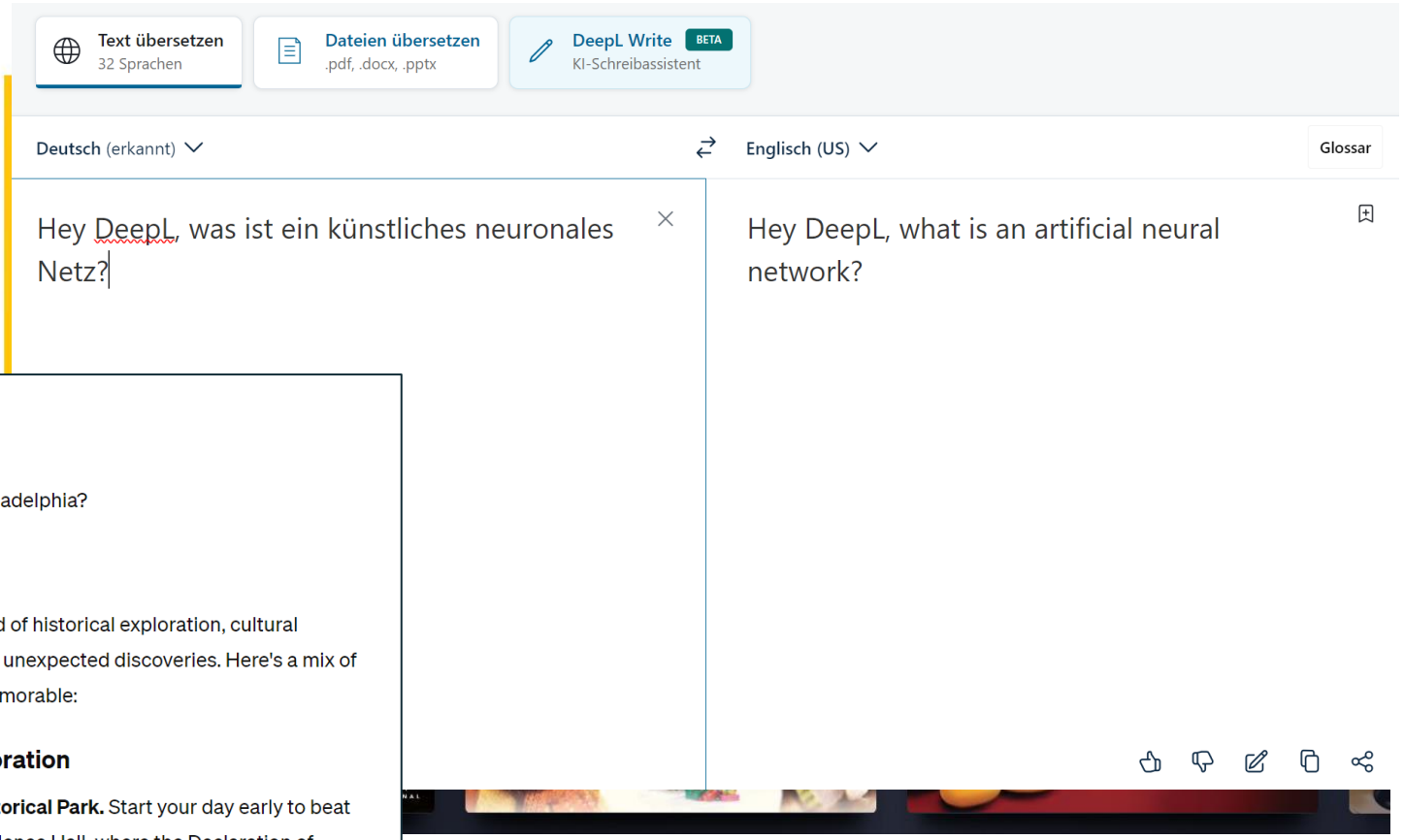
Value

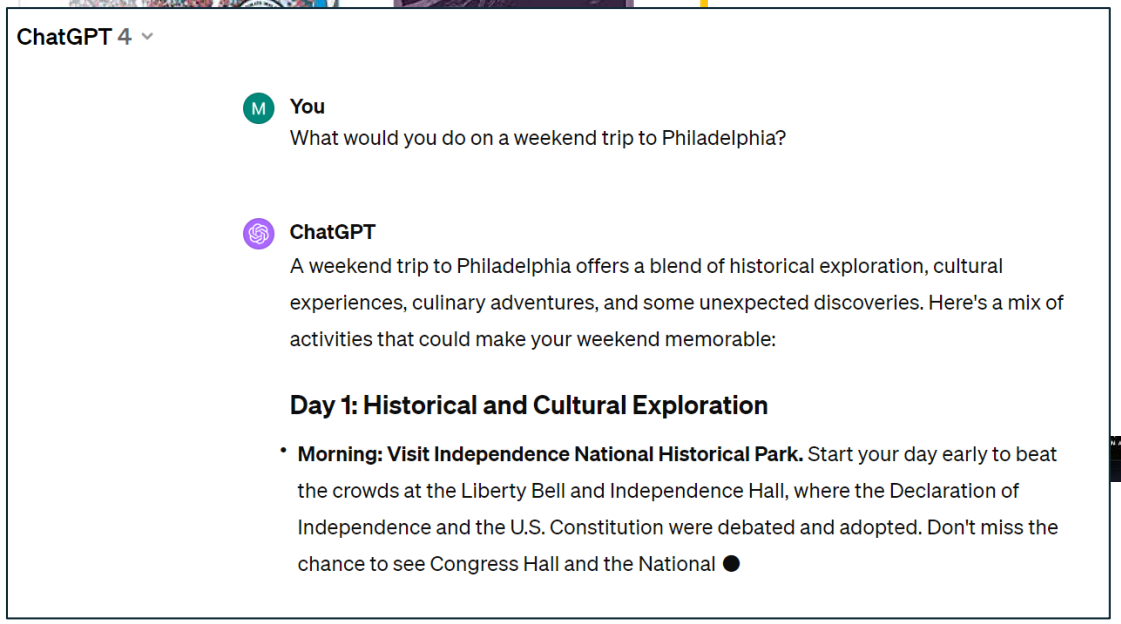
Ultimate goal of organizations: Produce value in the form of:

- Faster and smarter (business) decisions
- Increase efficient use of resources
- Discover new (market) opportunities

Impact of Big Data on Your Daily Life

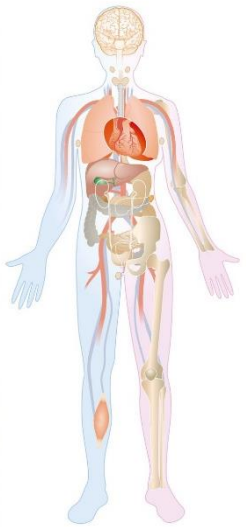
TUHH





And there is so much more…

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



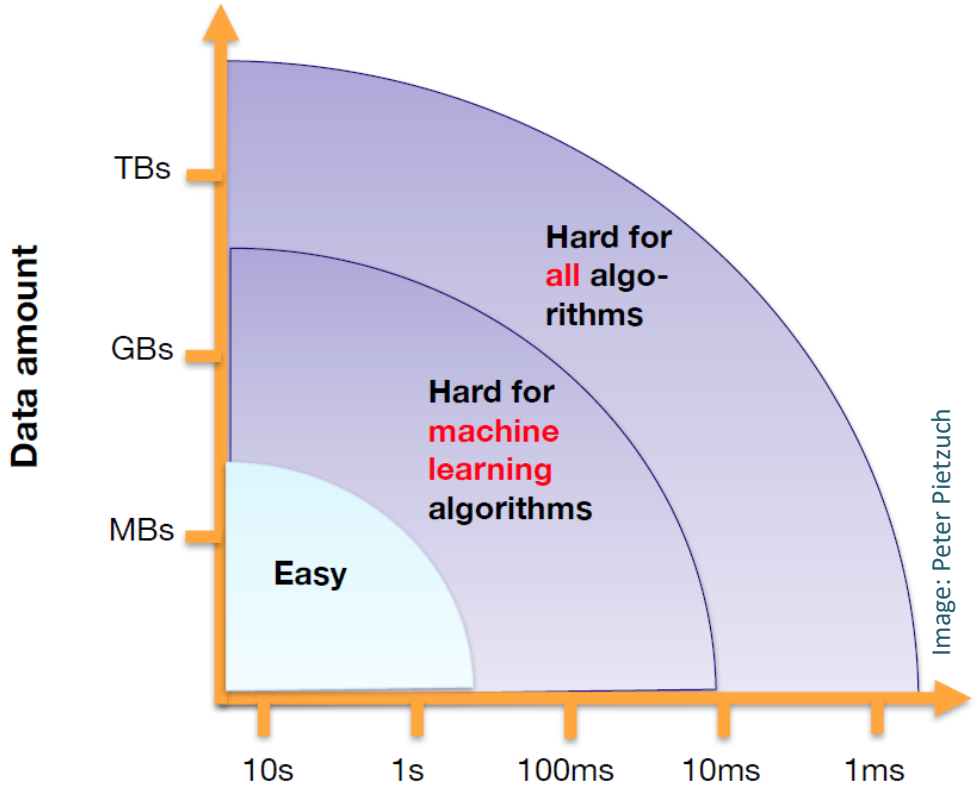
- E-health
  - 3.2 billion base pairs of DNA (genomics)
  - 10 million proteins in a person (proteomics)
- Preprocessing of sensor data
  - CERN experiments generate ~1PB of measurements per second
  - Unfeasible to store or process directly, fast preprocessing is a must

CERN

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

However – Basic Considerations First

- Do I have a big data problem?
  - My data fits in memory -> probably no



Key Takeaway Messages

- Lectures will be recorded
- Register for a lab topic (more precisely: for a certain group)
- Stud.IP should have all the necessary information – if something is missing, please let me know
- The impact of “Big Data”
- The 5 Vs: Velocity, Volume, Variety, Veracity, Value

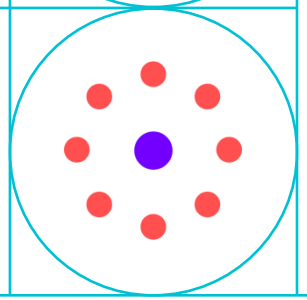
Further Readings / Acknowledgements

- A lot of these slides are (heavily) based on input by Prof. Tilman Rabl, Hasso-Plattner-Institut, Potsdam

Thank You Very Much!

Hamburg University of Technology
Stefan Schulte
Institute for Data Engineering
Blohmstr. 15
21079 Hamburg
Tel.: +49 40 42878-3027
stefan.schulte@tuhh.de

https://www.tuhh.de/ide/



TUHH
Hamburg
University of
Technology