LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Database Systems Group • Prof. Dr. Thomas Seidl

Praktikum Big Data Science SS 2017

Lecturer: Prof. Dr. Thomas Seidl

Assistants: Julian Busch
Evgeniy Faerman
Daniyal Kazempour
Sebastian Schmoll

Agenda

- Lab Organization
- Introduction
  - Data Science
  - Big Data
- Lab Goals
- Time Schedule
- Next Week
- References
- Topics

Lab Organization

- The lab is offered for the first time as part of the **ZD.B Innovation Lab Big Data Science**[1], coordinated by the chairs of
  - Prof. Dr. Bernd Bischl
    - http://www.compstat.statistik.uni-muenchen.de/
  - Prof. Dr. Dieter Kranzlmüller
    - http://www.nm.ifi.lmu.de
  - Prof. Dr. Thomas Seidl
    - http://www.dbs.ifi.lmu.de

- The lab will be hosted alternately at the chairs of Prof. Bischl (winter term) and **Prof. Seidl (summer term)** and is open to master students in Informatics and Statistics programmes
- Technical infrastructure for the lab is provided and maintained by the chair of Prof. Kranzlmüller and the Leibniz-Rechenzentrum (LRZ)

[1] https://zentrum-digitalisierung.bayern/massnahmen-alt/innovationslabore-fuer-studierende/

Lab Organization

- Supervisors
  - Julian Busch busch@dbs.ifi.lmu.de Room F 104
  - Evgeniy Faerman faerman@dbs.ifi.lmu.de Room F 109
  - Daniyal Kazempour kazempour@dbs.ifi.lmu.de Room F 106
  - Sebastian Schmoll schmoll@dbs.ifi.lmu.de Room F 110

- Website
  - http://www.dbs.ifi.lmu.de/cms/Praktikum_Big_Data_Science
  - Time schedule and material
  - Check regularly for updates and announcements

Lab Process

- Students will be assigned to groups of 5 students
- Each group can specify preferences for 4 different topics
- The lab is divided into two phases
  - Introductory phase
    - Prepare background, material and tools necesssary for the lab
    - Get familiar with your topic and prepare related theory
  - Project phase
    - Solve the tasks specified by your topic

Lab Process

- Each group will work on its topic following an agile scrum-like process
  - The lab is divided into sprints
    - Each sprint starts with a sprint planning session
    - „Daily“ stand-ups (2 appointments per week)
    - Each sprint ends with a sprint review and retropective
  - At the end of each sprint, the group will give a short report in the plenum
  - The group will maintain a documentation of its work
- During the last plenum session, all groups will present their results and provide a demonstration of their developed systems

Infrastructure

- For this lab, you will be provided with technical infrastructure
  - Project management
    - GitLab
    - JIRA

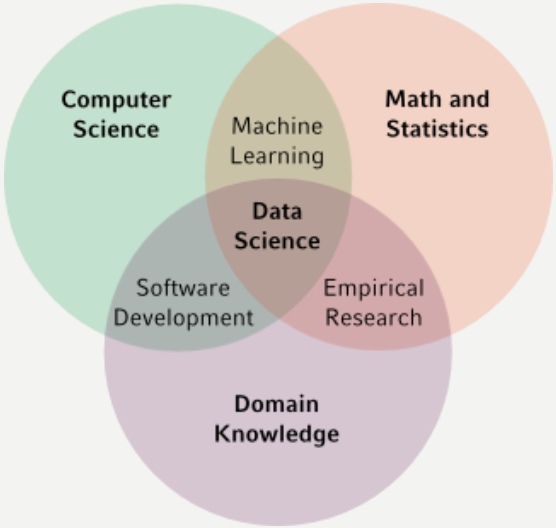
  - Compute cloud
    - OpenNebula

  - CIP Room N005/N006 (Baracke)
    - You will have exclusive access on Wednesdays, 14:00 – 18:00
    - The room is equipped with CIP-terminals, beamers and whiteboards

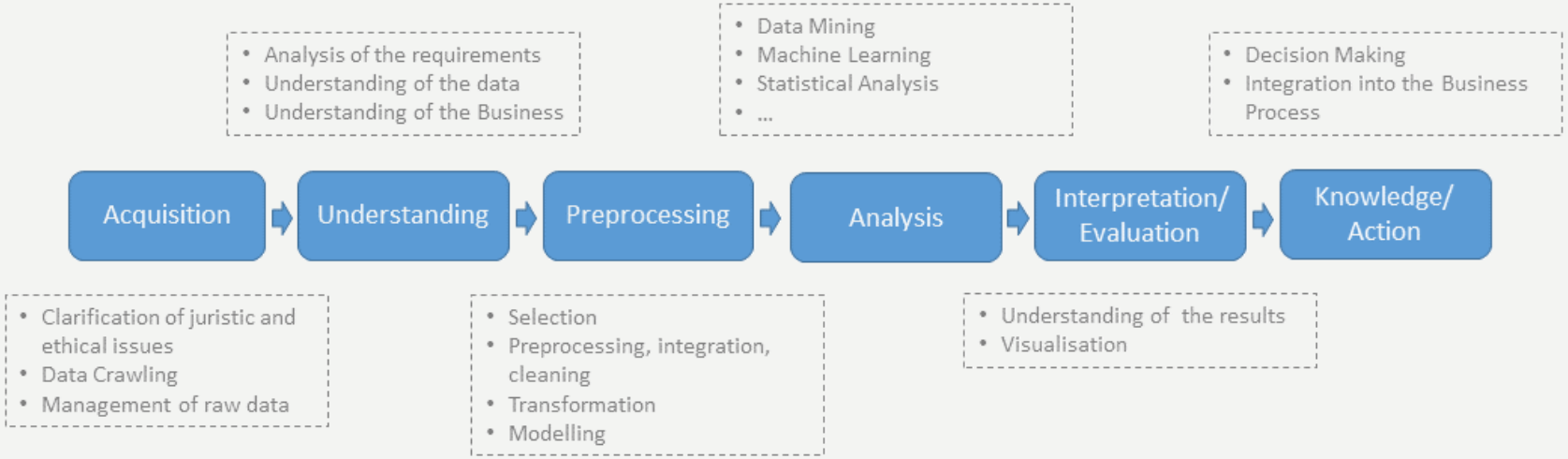
- For GitLab and the CIP-terminals, you will need your CIP-account
  - If you don't have one, you can register in Room LU113, Mon. - Fri., 14 - 17
    http://www.rz.ifi.lmu.de/FAQ/NeueKennung.faq.html

Data Science

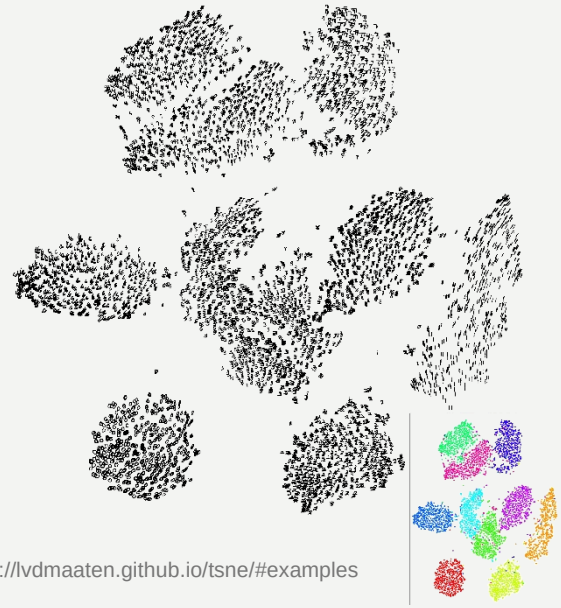
- Science of managing and analyzing data to generate knowledge



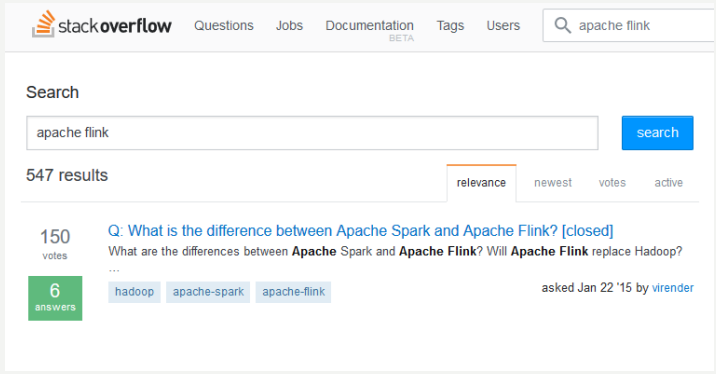
- The Data Science process
  - Requires knowledge from several domains
  - Usually consists of the following steps:

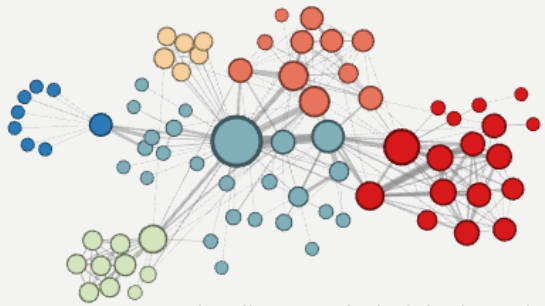


- Data Science Tasks:
  - Feature Extraction & Representation Learning
  - Clustering
  - Outlier & Trend Detection
  - Classification & Regression
  - Network Analysis & Graph Learning
  - Search & Retrieval
  - …



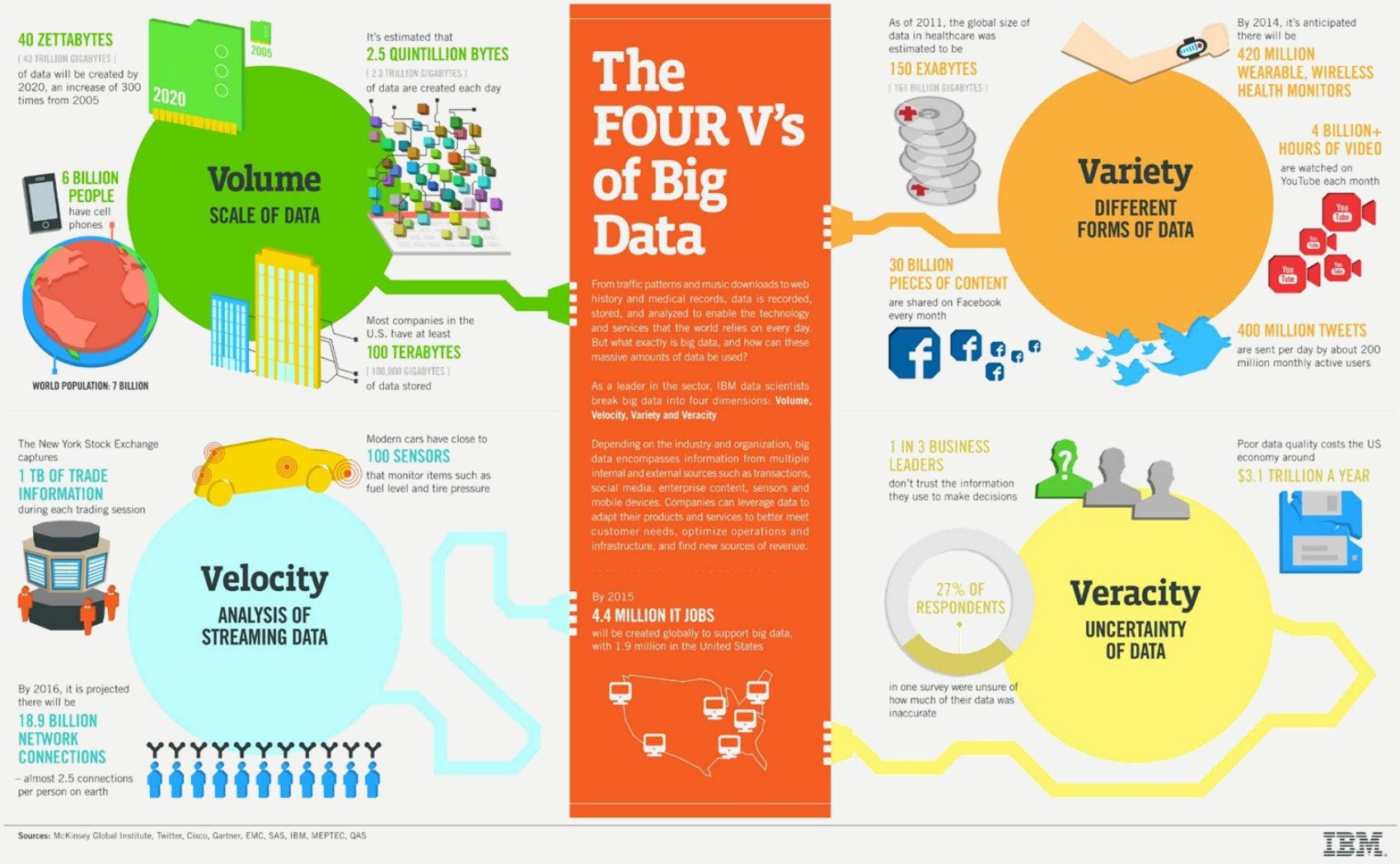
https://lvdmaaten.github.io/tsne/#examples



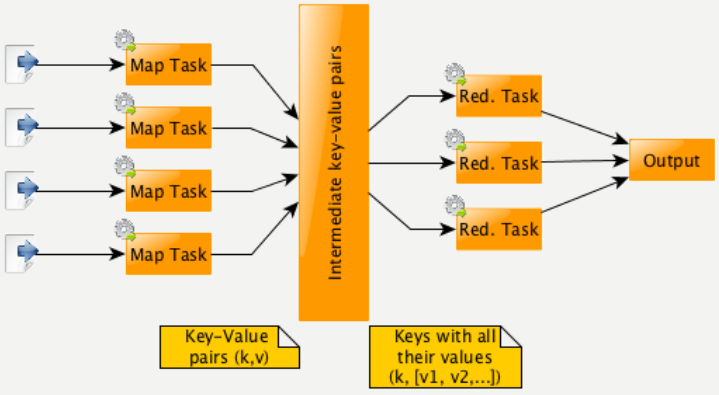


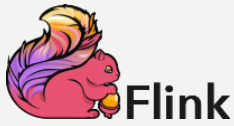
http://snap.stanford.edu/node2vec/

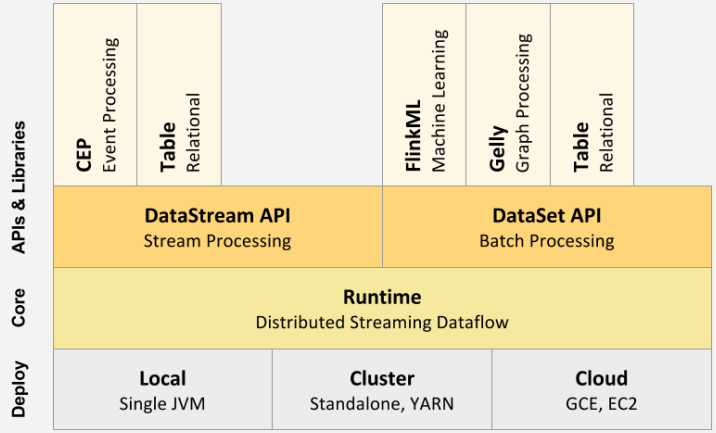
Big Data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The Four V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

- Big data requires large-scale data processing
  - In contrast to traditional grid computing, **MapReduce** offers a high-level programming interface that
    - Implicitly manages data flow
    - Partitions data to conserve network bandwidth
    - Is tolerant to hardware faults
  - **Apache Flink**: Open-source framework for batch and real-time stream processing based on MapReduce





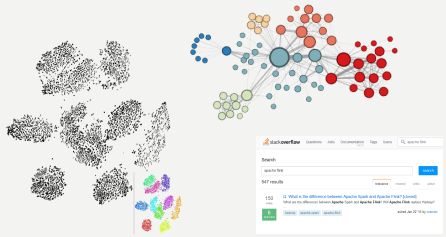


Lab Goals

- What will we do in this lab?
  - **Literature study** and familiarization with an active research direction in data science and related approaches
  - **Implementation** of state-of-the-art approaches in **Apache Flink**
  - **Application** of these approaches to a use case on real data
  - **Evaluation** of the approaches w.r.t.
    - Result quality
    - Efficiency
    - Scalability
  - Implementation of a **demo framework** for visualization and exploration
    - Integration of your implemented approaches
    - Presentation of your use case and evaluations

Lab Goals

- What will you learn?
  - Hands-on experience with a Data Science topic
    - Familiarization with a research direction
    - Application of the Data Science process



  - In-depth experience with a big data processing platform
    - Apache Flink

Flink

  - Working with a cloud computing system
    - OpenNebula

Open Nebula

  - Agile development in a team using Scrum
    - GitLab, JIRA

JIRA

Successful Participation

- In order to successfully complete the lab, you have to
  - Attend all meetings
  - Contribute actively in your group
    - As a guideline: 1 ECTS = 30 hours of work, i.e. during the 12 weeks of the lab course, you are expected to spend ≤ 30 hours per week on the lab
    - The topics are designed such that they can be flexibly rescaled if we observe that the workload is too small/large
    - Note: What counts is *what* you achieve, not how much time you need
  - Implement the backlog items specified by your topic according to their respective definitions of done
  - Maintain your group documentation and provide regular reports
  - Present your final results and your developed system
  - Participate in the discussions of other presentations

Time Schedule

- Fixed dates
  - **03.05. Kickoff-Meeting**
  - 10.05. Planning of Sprint 0
  - 24.05. End of Sprint 0, Planning of Sprint 1
  - 07.06. End of Sprint 1, Planning of Sprint 2
  - 21.06. End of Sprint 2, Planning of Sprint 3
  - 05.07. End of Sprint 3, Planning of Sprint 4
  - 19.07. End of Sprint 4, Final presentations

Introductory phase

Project phase

- Times
  - Wed. 14:00 – 16:00: Scrum Meetings
  - Wed. 16:00 – 17:00: Plenum Session
  - Stand-up meetings on appointment with your supervisor

Homework

- Homework until next week
  - Get together with your group
  - Decide for a group name
  - Decide on a ranking for the topics with your group
    - Send us an e-mail until next Monday, 08.05., 09:00
    - Will will then match the groups to the topics based on your rankings
  - Get familiar with Apache Flink
    - Get an overview and a basic understanding of the framework
    - Complete the Flink training by dataArtisans
      - http://dataartisans.github.io/flink-training/
      - Complete at least the DataStream and DataSet API parts
    - In the end, everyone should have a development environment ready
  - Get familiar with GitLab and JIRA
  - Get familiar with OpenNebula

Next Week

- Agenda for next week
  - Plenum session (14:00 – 15:00)
    - Short introduction to Scrum and how we will implement it in the lab
    - Short introduction to GitLab and JIRA and how we will use it
    - You will get your accounts for JIRA and OpenNebula
  - Planning of Sprint 0 (15:00 – 17:00)
    - Setup and configuration of GitLab and JIRA
    - Sprint goals:
      - Theoretical preparation of your topic
      - Setup of a Flink cluster in OpenNebula

References

- Useful references (not exhaustive)
  - Related lectures at DBS
    - http://www.dbs.ifi.lmu.de/cms/Big_Data_Management_and_Analytics_WS1617
    - http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)_16
    - http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II)_WS1516
    - http://www.dbs.ifi.lmu.de/cms/Maschinelles_Lernen_und_Data_Mining_16
  - Apache Flink
    - https://flink.apache.org/
    - https://mapr.com/introduction-to-apache-flink/
    - https://ci.apache.org/projects/flink/flink-docs-release-1.3/
  - GitLab, JIRA, Scrum
    - https://gitlab.cip.ifi.lmu.de/help
    - https://www.atlassian.com/software/jira
    - https://www.atlassian.com/agile/scrum
  - OpenNebula
    - https://www.lrz.de/services/compute/cloud_en/