Practical Big Data Science

Max Berrendorf   Evgeniy Faerman   Michael Fromm   Prof. Dr. Matthias Schubert

Lehrstuhl für Datenbanksysteme und Data Mining
Ludwig-Maximilians-Universität München

24.04.2019

Agenda

Organisation

Goals

Schedule

Topics

Group Assignment

Organisation

Lab Organisation

- Offered as part of *ZD.B Innovation Lab Big Data Science*[1], coordinated by the chairs of
  - Prof. Dr. Thomas Seidl[2]
  - Prof. Dr. Bernd Bischl[3]
  - Prof. Dr. Dieter Kranzlmüller[4]
- Hosted alternately at the chairs of *Prof. Seidl (summer term)* and Prof. Bischl (winter term)
- 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/

[2] http://www.dbs.ifi.lmu.de

[3] http://www.compstat.statistik.uni-muenchen.de/

[4] http://www.nm.ifi.lmu.de

Lab Organisation

Supervisors

Name	Mail	Room
Max Berrendorf	berrendorf@dbs.ifi.lmu.de	F110
Evgeniy Faerman	faerman@dbs.ifi.lmu.de	F112
Michael Fromm	fromm@dbs.ifi.lmu.de	F110

Website

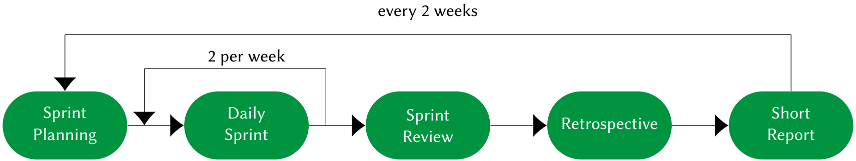
- `http://www.dbs.ifi.lmu.de/cms/studium_lehre/lehre_master/pbds19/index.html`
- Time schedule and material
- Check regularly for updates and announcements

Lab Organisation

Process

- We assign students to groups of 5 students
- Each group can specify preferences for 7 different topics
- We assign the groups to the topics

Lab Organisation



Process

- Each group will work on its topic following an agile scrum-like process
- The lab is divided into sprints
- At the end of each sprint groups report about last sprint and plans for the next
- During the last plenum session, all groups will present their results and provide a demonstration of their developed systems

Infrastructure

Project Management

GitLab

Compute Cloud

lrz

openstack®

Room

- Room 161, Wednesday, 14:00 - 18:00, exclusive usage

Goals

Lab Goals

What will you 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 TensorFlow/PyTorch
- *Application* of these approaches to a use case on real data
- *Evaluation* of the approaches w.r.t.
  - Result quality
  - Efficiency
  - Scalability

Lab Goals

What will you learn?

- Hands-on experience with a Data Science topic
- In-depth experience with machine learning platform TensorFlow/PyTorch
- Working with a cloud computing system: OpenStack
- Agile development in a team using Scrum: GitLab

Lab Goals

Successful Participation

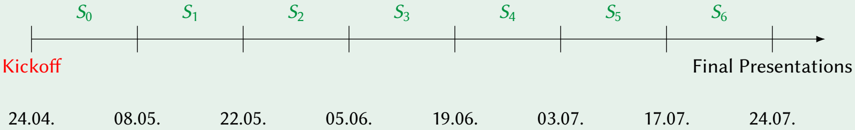
In order to successfully complete the lab, you have to

- Attend all meetings
- Contribute *actively* in your group – Guideline: 28h/week
- 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

Schedule

Time Schedule

Fixed Dates



S_0 S_1 S_2 S_3 S_4 S_5 S_6

Kickoff

Final Presentations

24.04. 08.05. 22.05. 05.06. 19.06. 03.07. 17.07. 24.07.

Times

- Wed., 14:00-16:00: Default appointment for Scrum Meetings
- Wed., 16:00-18:00: Plenum Session
- Stand-up meetings on appointment with your supervisor

Topics

Conditions for Industry Projects

Company

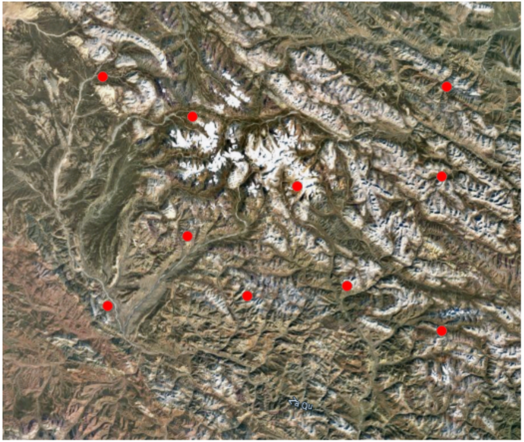
- Signs contract with the university
- Optionally acquires rights of use (exclusive or non-exclusive)

Students

- Sign contract with the university
- Execute project
- Get money *if* the company acquires rights of use
  - for the team for non-exclusive rights of use
  - for the team for exclusive rights of use

1. CompanyX (Industry)

CompanyX (Industry)



Tasks

- Spatial interpolation of measurements
- Identification of corrupt sensors
- Knowledge transfer between different regions

Profit

- Work with state-of-the-art relational (also deep neural networks) models
- Understand shortcomings of current approaches
- Adapt and extend state-of-the-art models

2. Anomaly Detection in X-Ray Images (Industry)

Anomaly Detection in X-Ray Images (Industry)

Setting

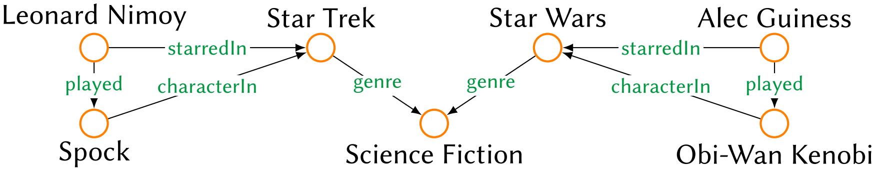
- Data: X-Ray images of hand
- Problem: Support detection of anomalies

Task

- Unsupervised learning
- Adapt and extend existing technology for MRI images

3. Link Prediction for Knowledge Graphs

Link Prediction for Knowledge Graphs



Example Graph Source: https://arxiv.org/pdf/1503.00759.pdf

Data

- A knowledge graph contains facts in the form (s, p, o)
- s, o are entities, p is a relation

Goal

Given s and p, what are the likely entities for o?

Link Prediction for Knowledge Graphs

Tasks

- Different models / different initialisations lead to different performance
- Analyse errors made by KG models
- Compose ensemble to improve performance

Profit

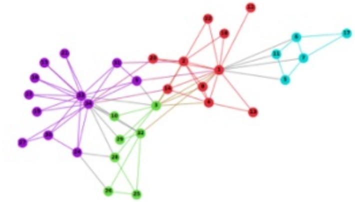
- Work with state-of-the-art relational models
- Understand shortcomings of models
- Learn different ensemble models on real-world task

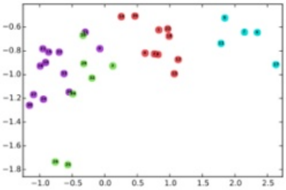
4. Entity Linking for Argument Mining

Entity Linking for Argument Mining

Documents

Knowledge Graph and Embeddings





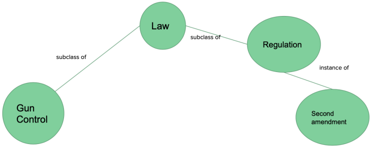
Data

- Annotated data set of arguments
- Knowledge graphs with entities and relations

Goal

Further improve the argument detection with the usage of knowledge graph information

Entity Linking for Argument Mining



Tasks

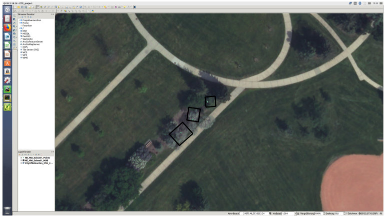
- Embed knowledge graphs
- Link words in sentences to entities in knowledge graphs
- Improve Argument Identification

Profit

- Use state-of-the-art Neural Network techniques based on RNN
- Use state-of-the-art embedding methods on knowledge graphs

5. Vegetation Registration for Environmental Monitoring

Vegetation Registration for Environmental Monitoring



QGIS 2.18.16 - ETP_project

Koordinate 29079.48,5936603.24 Maßstab 1:284 Vergrößerung 100% Drehung 0,0 Zeichnen EPSG:3776 (SRP)

Tasks

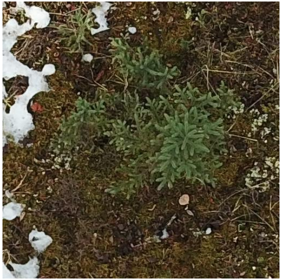
- Annotate vegetation and environmental features
- Adapt results to Geo-Information Systems
- Label augmentation
- Semisupervized meta-data generation

Profit

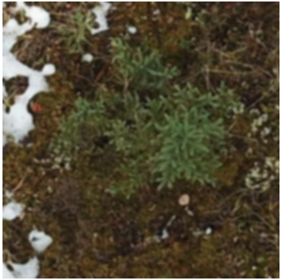
- Use state-of-the-art Imaging techniques based on CNNs (Detection, Segmentation)
- Manage Spatial data with Geo-Information Systems

6. Superresolution and Object Detection

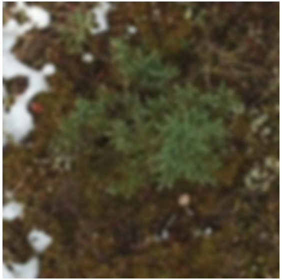
Superresolution and Object Detection



Original flight height (5m)



Down-sampled to 25m



Down-sampled to 45m

Down-sampled to 105m

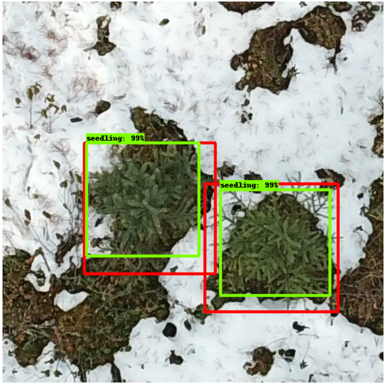
Data

- Annotated data set of seedlings.
- Images are shot at 5m flight height.

Goal

Further improve the object detection performance on the seedling

Superresolution and Object Detection



Tasks

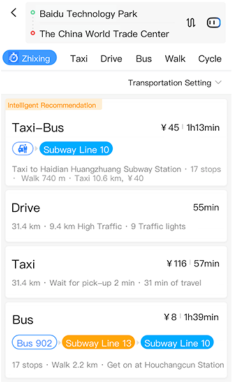
- Generate CNN based super resolution models
- Generate GAN based super resolution models
- Compare against standard methods
- Influence of super resolution networks on object detection

Profit

- Use state-of-the-art Imaging techniques based on CNNs (Detection, Super Resolution)
- Use state-of-the-art Imaging techniques based on generative models (Super Resolution)

7. KDD Cup 2019

KDD Cup 2019



Baidu Technology Park
The China World Trade Center

Zhixing Taxi Drive Bus Walk Cycle

Transportation Setting

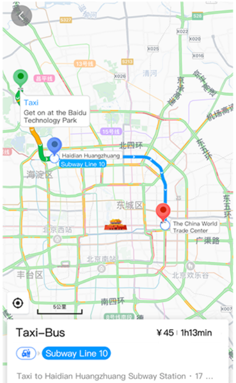
Intelligent Recommendation

Taxi–Bus ¥45 | 1h13min
Subway Line 10
Taxi to Haidian Huangzhuang Subway Station · 17 stops · Walk 740 m · Taxi 10.6 km, ¥40

Drive 55min
31.4 km · 9.4 km High Traffic · 9 Traffic lights

Taxi ¥116 | 57min
31.4 km · Wait for pick-up 2 min · 31 min of travel

Bus ¥8 | 1h39min
Bus 902 Subway Line 13 Subway Line 10
17 stops · Walk 2.2 km · Get on at Houchangcun Station



Taxi–Bus ¥45 | 1h13min
Subway Line 10
Taxi to Haidian Huangzhuang Subway Station · 17 ...

Tasks

- Context-Aware Multi-Modal Transportation Recommendation
- Context: User type, Price, Duration, ?

Profit

- Work with state-of-the-art (also deep neural networks) models
- Participation in data science competition

Homework

Homework (until tomorrow)

- Join Slack via: https://tinyurl.com/y5guhzz9
- Get together with your group (shown in two slides); 1h
  - decide for a group name
  - discuss which topics you prefer
  - afterwards fill out this survey (as a group): https://forms.gle/f6bag2hzczH9kHh99
- In LRZ-Gitlab[5] 1h
  - Create a group named as your group name; invite all three supervisors
  - Create a project within this group

[5] https://gitlab.lrz.de/

Homework

Homework (until next week)

Get familiar with:

- Python
- Numpy
- OpenStack: Link
- GitLab: Link 1 Link 2
- PyTorch: Link
- DVC: Link 1 Link 2
- MLFlow: Link 1 Link 2

Group Assignment

Group Assignment

(removed for privacy reasons)