Practical Big Data Science

Max Berrendorf  Felix Borutta  Evgeniy Faerman
Prof. Dr. Thomas Seidl

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

12.04.2018
Agenda

Organisation

Goals

Schedule

Topics

Gitlab Introduction

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*)

- 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/](https://zentrum-digitalisierung.bayern/massnahmen-alt/innovationslabore-fuer-studierende/)

\(^2\) [http://www.dbs.ifi.lmu.de](http://www.dbs.ifi.lmu.de)

\(^3\) [http://www.compstat.statistik.uni-muenchen.de/](http://www.compstat.statistik.uni-muenchen.de/)

\(^4\) [http://www.nm.ifi.lmu.de](http://www.nm.ifi.lmu.de)
Lab Organisation

Supervisors

<table>
<thead>
<tr>
<th>Name</th>
<th>Mail</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Berrendorf</td>
<td><a href="mailto:berrendorf@dbs.ifi.lmu.de">berrendorf@dbs.ifi.lmu.de</a></td>
<td>F110</td>
</tr>
<tr>
<td>Felix Borutta</td>
<td><a href="mailto:borutta@dbs.ifi.lmu.de">borutta@dbs.ifi.lmu.de</a></td>
<td>156</td>
</tr>
<tr>
<td>Evgeniy Faerman</td>
<td><a href="mailto:faerman@dbs.ifi.lmu.de">faerman@dbs.ifi.lmu.de</a></td>
<td>F109</td>
</tr>
<tr>
<td>Dave Chen</td>
<td><a href="mailto:davech2y@outlook.com">davech2y@outlook.com</a></td>
<td></td>
</tr>
<tr>
<td>Robert Müller</td>
<td><a href="mailto:robert.mueller@campus.lmu.de">robert.mueller@campus.lmu.de</a></td>
<td></td>
</tr>
</tbody>
</table>

Website

- Time schedule and material
- Check regularly for updates and announcements
Lab Organisation

Process

- We assign students to groups of 5-6 students
- Each group can specify preferences for 5 different topics
- We assign the groups to the topics
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.
Project Management

GitLab

Compute Cloud

Open Nebula

Room

- Room U 151, Thursday, 14:00 - 18:00, exclusive usage
- The room is equipped with CIP-terminals, beamers and whiteboards
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
- *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:
  - Familiarization with a research direction
  - Application of the Data Science process
- In-depth experience with machine learning platform TensorFlow
- Working with a cloud computing system: OpenNebula
- 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: 25h/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

<table>
<thead>
<tr>
<th></th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>S₄</th>
<th>S₅</th>
<th>S₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kickoff</td>
<td>12.04.</td>
<td>03.05.</td>
<td>17.05.</td>
<td>31.05.</td>
<td>14.06.</td>
<td>28.06.</td>
</tr>
<tr>
<td>Final Presentations</td>
<td>12.07.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Times

- Thur., 14:00-16:00: Scrum Meetings
- Thur., 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
- Pays for the project *execution* first
- Optionally acquires rights of use (exclusive or non-exclusive)

Students

- Sign contract with the university
- If necessary sign NDA (and take it seriously)
- Execute project
- Get money *if* the company acquires rights of use
  - $x$ € for the team for non-exclusive rights of use
  - $y$ € for the team for exclusive rights of use
1. Company X (industry)
Spatio-temporal signal interpolation
Spatio-temporal signal interpolation

Problem

▶ Measure stations spatially distributed
▶ Input:
  ▶ Historic data for each station
  ▶ Future prediction for few stations
▶ Output: Predictions for all other stations

What will you learn

▶ Work on real-life project
▶ Experience with state-of-the-art Deep Learning methods:
  ▶ Recurrent networks
  ▶ Graph Neural Networks (Attention)
  ▶ Integration of different information sources
2. Harman (industry)
Active Learning for Object Detection (industry)

Street Scenes Data

Image Source: http://cbcl.mit.edu/softwareatasets/streetscenes/
Active Learning for Object Detection (industry)

Basic Idea: Creating a support system for labeling

- Data: Street scenes images
- Problem: The set of labels is going to be very sparse
- Goal: Integrating user expertise into semi-automated labeling process
- Active Learning approaches to solve two problems
  1. Object Detection
  2. Object Labeling
- Tasks:
  - Identification and Implementation of suitable algorithms
  - Join two active learning steps within one framework
  - Integration into existing UI
- Profit: Learn fundamental AI concepts that are already established in the area of ML
3. Movie Rating Prediction
Movie Rating Prediction

Task

- Predict the average IMDb rating for new movies based on meta data (e.g., actors, directors, posters, …)
- As data sources, you may use all freely available resources (e.g., IMDb, Wikipedia, OMDB, …)

Goal

- Develop a website where the user can input meta information concerning a specific movie
- AI backend should provide an accurate prediction of the average IMDb rating
Movie Rating Prediction

Challenges

- Heterogeneous data sources
- Cope with missing meta-data

Profit

- Choose data sources by yourself
- Evaluate ML algorithms w.r.t. to heterogeneous data sources
- Find out if a new movie is worth watching
4. Air pollution prediction (KDD CUP of Fresh Air)
Air pollution prediction

Task

- Predict air pollutants concentration for future
- Data: historical pollution and weather data from different sources
  - 35 stations in Beijing and 13 in London
  - Data from KDD Cup 2018

Goal

- Develop a system for air pollutant prediction
- Include additional information (e.g. distance between stations, etc.)
5. Explainable AI
Explainable AI for CNNs

Inception Activations

Image  "Colour"  "Texture"  "Shape"

Berrendorf, Borutta, Faerman (LMU)
Explainable AI for CNNs

Goal
Open black-box of CNNs

Activation Maximisation

Data Space

Data Set

Image Source: https://distill.pub/2017/feature-visualization
Explainable AI for CNNs

**Task**
- Explorative Analysis of CNN activations for full Imagenet

**Goal**
- Determine role of neurons ("Explanation by Example")
- Identify important neurons
- Similarity Search based upon different Feature Representations
Explainable AI for CNNs

Challenges

- Huge data (for 1.2M images approx. 16 TiB raw data)
- Many possible queries (top-k retrieval, correlations, clustering, …)
- For explorative analysis: near realtime processing

Profit

- Develop a system for big data analysis (backend + frontend)
- Deepen understanding of the inner workings of CNN
- Improve CNN structure?
Gitlab Introduction
Gitlab Introduction

GitLab

- https://gitlab.lrz.de
- Sign in with LRZ-ID
- How to create a group?
- How to create a project?
- Issues & Milestones

---

The LRZ-ID can be found at https://www.portal.uni-muenchen.de/benutzerkonto/index.html

Berrendorf, Borutta, Faerman (LMU)
Group Assignment
(removed for privacy reasons)
Homework (until tomorrow)

- Get together with your group
- Decide for a group name 1h
- Decide on a ranking for the topics with your group 1h
  - Send us an e-mail until Friday, 13.04., 15:00
  - We will match the groups to the topics based upon this rankings
- In LRZ-Gitlab\(^7\) 1h
  - Create a group named as your group; invite all three supervisors and both Hiwis.
  - Create a project within this group
  - (More information about Gitlab later)

\(^7\)https://gitlab.lrz.de/
Homework (until next week)

Get familiar with:

- Python
- numpy
- TensorFlow
- OpenNebula
- Git
- Scrum
- GitLab Issues/Milestones

22h
Useful References

Related Lectures

- Knowledge Discovery in Databases I (KDD I)
- Knowledge Discovery in Databases II (KDD 2)
- Big Data Management and Analytics
- Machine Learning

OpenNebula

- Info LRZ
- Tutorials
Useful References

TensorFlow

- Get Started With TensorFlow

Git

- Basics
- Branching
- Feature/Development/Master Branch (by Atlassian)
Useful References

GitLab

- LRZ GitLab
- Workflow Overview

SCRUM

- Scrum Overview (Atlassian)