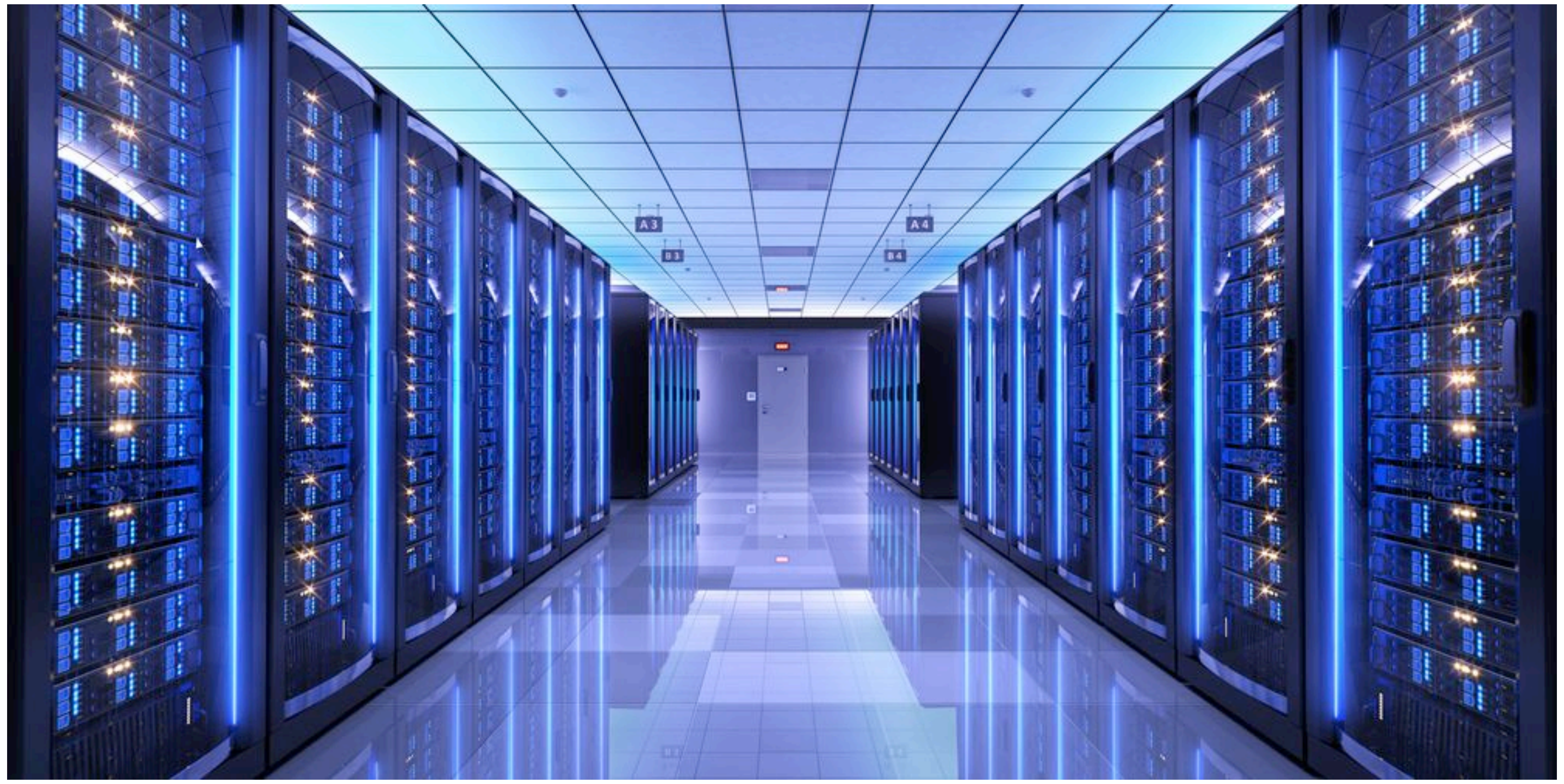
Big Data 2022
Course Syllabus

Sebastian Schelter



Version 2022/01/20

Staff

1. **Coordinator and main lecturer**
   Sebastian Schelter

2. **Senior Teaching Assistants**
   Ashish Sai, João Lebre Magalhães Pereira

3. **Teaching Assistants**
   Stefan Grafberger, Barrie Kersbergen, Shubha Guha, Mayesha Tasnim

Course Content

Data is at the center of modern businesses and institutions, and we are experiencing a big shift towards predictive, data-driven decision making in recent years. This development has given rise to Big Data, a novel set of data storage and processing methods, accompanied by a new software stack that serves as the foundation for the ongoing AI revolution. Pioneered by companies building web-scale search engines such as Google, Big Data technologies are used across all industries nowadays and are a major component of modern cloud infrastructure.

In this course, we study various abstractions, processing techniques and Big Data systems for working with large collections of data, including relational database management systems, MapReduce and Apache Spark. We address the foundations of Big Data applications, as well as topics like choosing a suitable data representation and implementing distributed processing operations. We review and focus on the foundations of distributed data processing and programming models for parallelisable programs, as well as on data quality and responsible data management.

You will learn theoretical concepts for data processing during the lectures. In the lab you will gain hands-on experience through a number of coding assignments and by participating in a Kaggle-like Big Data competition. In addition, we offer online tutorials for hands-on experience with established industry software packages. Finally, experts from the field (both academic as well as industry colleagues) are invited for giving guest lectures.

Intended Learning Outcomes

After completing the course, students should be able to:

1. Explain the high impact and potential of big data technology
2. Create a scalable Big Data application for a given scenario
3. Analyse data schemas and distributed data processing strategies
4. Design a scalable Big Data application for a machine learning task and present findings in a poster session
5. Explain what relates and differentiates relational data processing, MapReduce and Resilient Distributed Datasets
6. Describe common data quality issues and apply error detection and data cleaning methods
7. Program key Big Data systems

Prior knowledge, additional materials, course materials

Recommended prior knowledge:

- Basic programming skills and a basic understanding of machine learning are required.

Additional information:

- The course comes with mandatory practical labs, where students will implement Big Data tasks in Python, and design a data-centric solution for a Kaggle-like task.

Course materials:

- Book chapters
- Videos / Tutorials
- Papers
- Syllabus

Assessment

The final grade is a weighted combination of:

1. **Written exam (55%)**
    a. Technical questions
    b. Open / insight questions
2. **Project (40%)**
    c. Kaggle-like task in class
    d. Innovation & performance
    e. Poster presentation
3. **Assignments (5%)**
    f. 2 Lab assignments;
    g. 1 Open assignment;
    h. Each graded fail/pass only;
    i. Late submission = fail;
    j. Grade: 10-(2.5 * Failed)

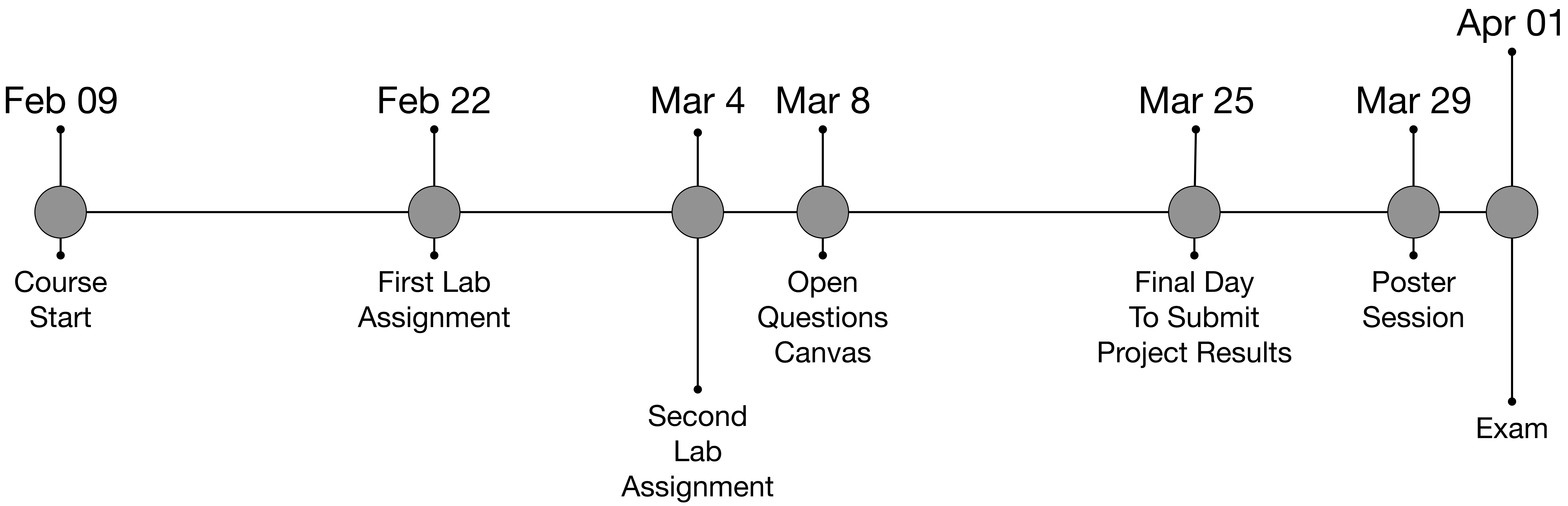
To pass the course, **all** parts should be passed (i.e. minimum grade of 5.5), we will have firm deadlines. No substitution between parts.

Course Structure 2022

	CW	Lecture	Lab	Tech Tutorials [Online + Q&A; voluntary]
1	6	Intro & Foundations	Intro & Setup	-
2	7	Relational Data Processing	**Lab Assignment 1 [g]**: Pandas/SQL	Version Control with Git
3	8	MapReduce	**Lab Assignment 2 [g]**: MapReduce/Spark	DuckDB
4	9	Resilient Distributed Datasets	Kaggle-Project week 1 **Open questions [g]**	Apache Spark Deep Dive
5	10	Data Cleaning	Kaggle-Project week 2	Great Expectations
6	11	Responsible Data Management	Kaggle-Project week 3	AIF 360
7	12	Big Data at bol.com	Kaggle-Project week 4	Kubernetes
8	13	**Exam [g]**	**Poster Session [g]**	-

[g] indicates a graded activity

Deliverables Timeline



Big Data Project

Kaggle-like project with a focus on data

- Automatic evaluation
- Leaderboard with classmates
- Preparation for real-world data science work

- Run and submit a Kaggle-like project
- Integrate, clean and prepare data to train an ML model
- **Focus on innovation & data integration, cleaning and preparation (not on ML model)**

- Free to choose from 3 projects
- Work in groups of 5
- Should use DuckDB and/or PySpark for suitable parts of the data integration, cleaning and preparation code
- You can use any ML model and leverage additional data (except for the original data)

- **Poster Session**
  - Poster should help tell your story
  - 2 minute elevator pitch highlighting your work; afterwards expect any questions about your work
  - In principle, all members should present
  - Might be online, details in first lecture
- **There is no final report**

The Projects

1. **Movie Genre Classification**
   - Learn to distinguish between “Drama” and “Thriller” movies from the IMDB movie database

2. **Bibliography Deduplication**
   - Learn to identify duplicate entries in the DBLP research paper database

3. **Product Review Classification**
   - Learn to identify “helpful” reviews from a multilingual dataset of Amazon product reviews

The data for all projects is spread over multiple files and comes with synthetic errors!

Project Grading

1. **Innovation**
   a. What is novel or interesting?

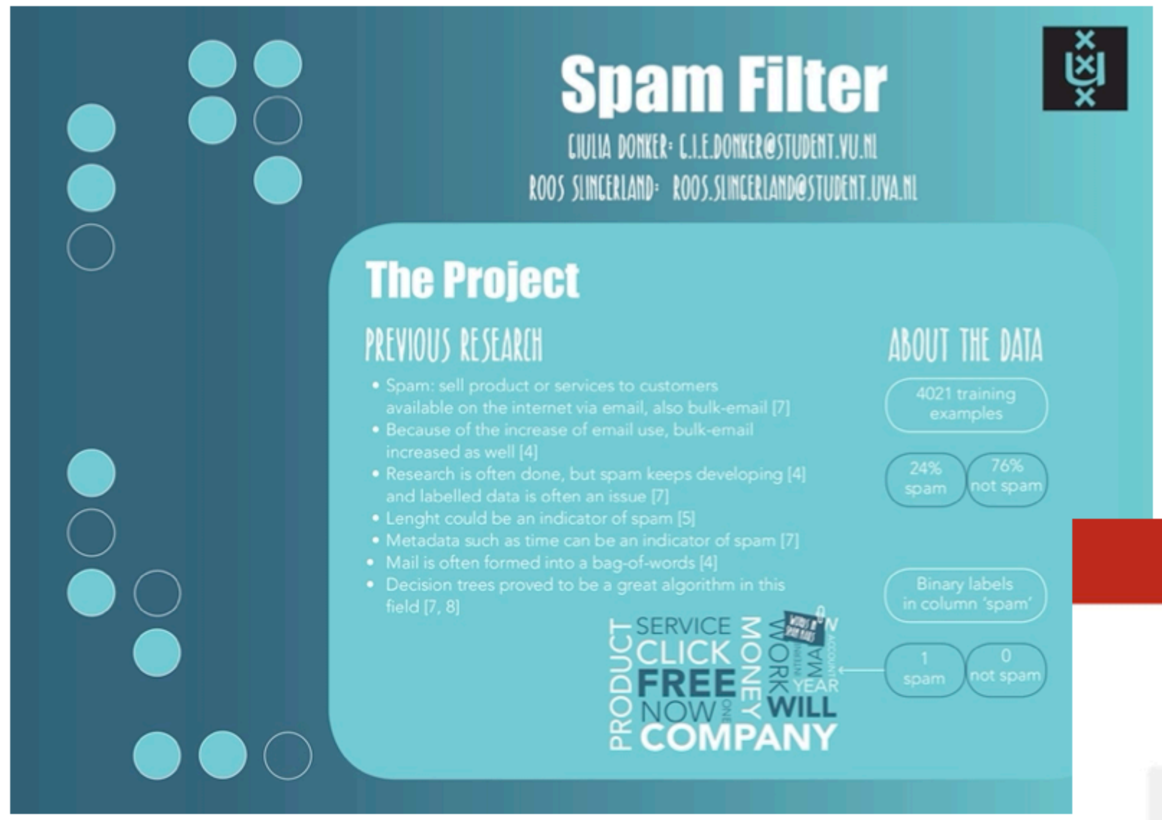
2. **Pipeline Design**
   a. How reusable is your data pipeline?
   b. How did you decide which parts of the pipeline to run in DuckDB / PySpark?

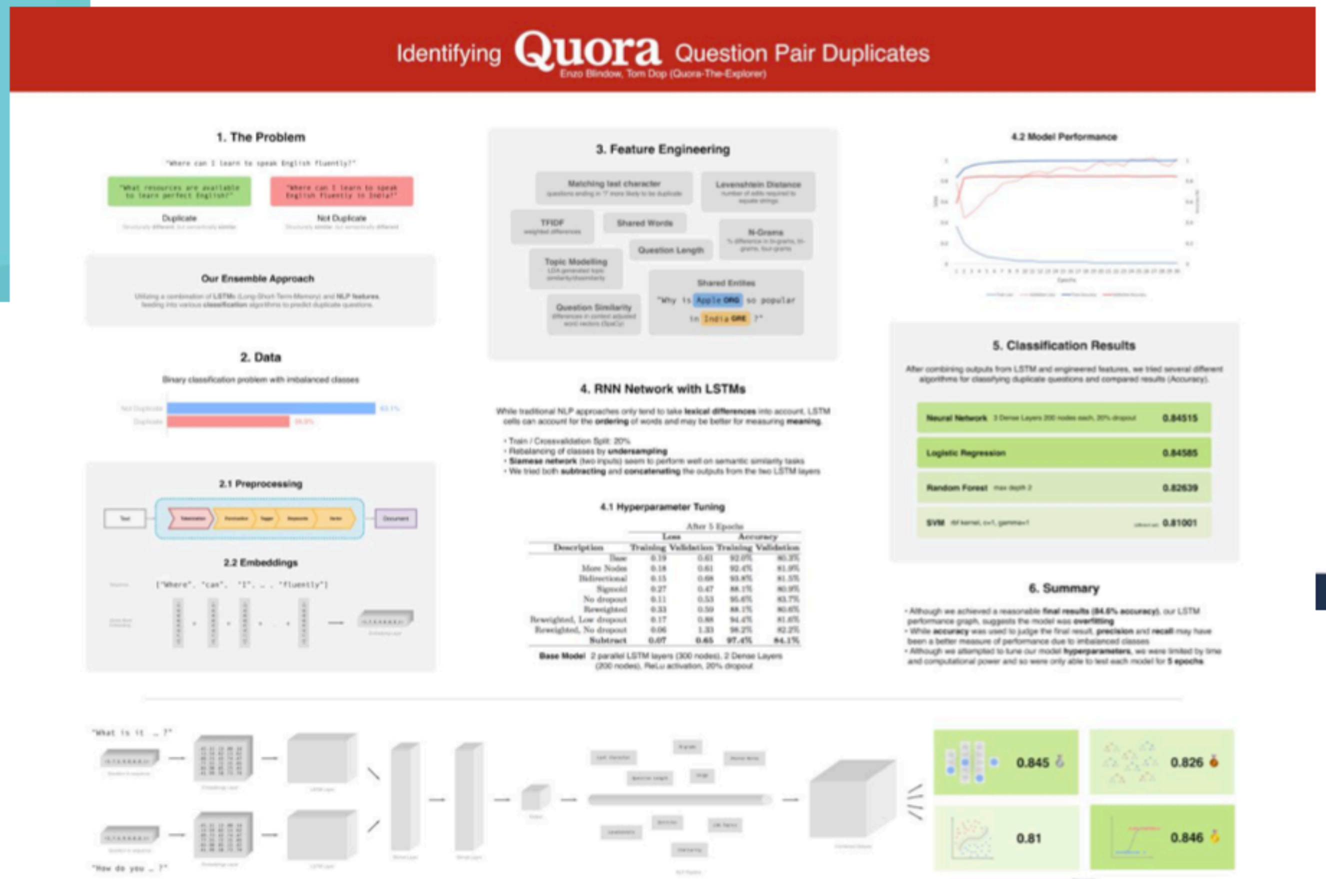
3. **Analysis**
   a. How innovative/efficient/stable are your data integration, cleaning and preparation operations?
   b. How good is your learning performance?

4. **Pitch and Poster Design**
   - Clear pitch? Helpful poster design?

All components weight equally (25%).
Each project will be graded by 2 or 3 TAs.

Related Posters from the Applied ML Course





Exam

Covers all lectures and required reading materials.

Consists of theoretical questions (to test knowledge) and open questions (to test insight).

In case of a resit, the resit grade is used.

Weekly Materials

- Every lecture accompanied by related book chapters and/or papers as well as online videos
- Links to materials available in Canvas
- **Read and watch the materials** as a preparation **before** the weekly lecture, the lab assignments and the exam

Week 1:

Intro & Foundations

Reading

- Halevy et al.: The Unreasonable Effectiveness of Data, IEEE Intelligent Systems 2009

Videos

- Joe Hellerstein: Big Shifts in Data and Analytics
- Daniel Pearl: Volume, Velocity, and Variety of Big Data
- Ted Dunning: Why Hadoop Works

Week 2:

Relational Data Processing

Reading

- Garcia-Molina et al.: Database Systems, The Complete Book, Chapters 1.1, 2.2, 2.3, 6.1, 6.2, 6.4

Videos

- CS186Berkeley - SQL Tutorial Videos 1.2, 1.3, 1.4, 1.5, 1.6, 2.1, 2.6

Week 3:

MapReduce

Reading

- Leskovec et al.: Mining Massive Datasets: Chapter 2.1, 2.2
- Dean et al.: MapReduce: Simplified Data Processing on Large Clusters, OSDI'04

Videos

- Mining Massive Datasets: Videos 1, 2 for Chapter 2

Week 4:

Resilient Distributed Datasets

Reading

- Zaharia et al.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing, NSDI'12
- Learning Spark for Lightning Fast Big Data Analysis: Chapter 3 - Programming with RDDs

Videos

- Matei Zaharia: What is Apache Spark?
- Matei Zaharia: Spark RDD: A Fault-Tolerant Abstraction for In-Memory Cluster Computing

Week 5:

Data Cleaning

Reading

- Hellerstein: Quantitative Data Cleaning for Large Databases: Chapter 1 + Chapter 2 until 2.7
- Biessmann et al.: “Deep” Learning for Missing Value Imputation in Tables with Non-Numerical Data, CIKM'18
- Forbes: Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Videos

- Mike Stonebraker: Why Is Enterprise Data Integration So Challenging?

Week 6:

Responsible Data Management

Reading

- Stoyanovich et al.: Responsible Data Management, VLDB'21
- Khan et al.: Data, Responsibly (Vol. 2) Fairness and Friends

Videos

- Joy Buolamwini: How I am fighting bias in algorithms
- Julia Stoyanovich: Building Data Equity Systems

Week 7:

Big Data at bol.com

No reading required

Tech Tutorials

- Practical introduction to state-of-the-art Big Data technologies
- **Not graded, participation voluntary**
- Opportunity to get additional practical training

- Tutorials for important data science libraries
  - Version control with git
  - DuckDB
  - Apache Spark Deep Dive
  - Great Expectations
  - AIF 360
  - Kubernetes

- Each tutorial consists of the following (Zoom invites will be provided via canvas)
  - Introductory video
  - Tutorial task video
  - Online Q&A session