Big Data Project - Logistics (1)

- Group project for **teams of 5 students**
- **Activities:**
  - **Design, implement and evaluate an ML pipeline for a binary classification problem on dirty, erroneous data**
  - **Submit the predictions** of your pipeline **online**
  - Create a **poster** (and a **pitch**) to present your project findings
- Final day to submit predictions to the submission server: **March 25th**
- **Poster Session: March 29th**
- There is **no final report**!

Big Data Project - Logistics (2)

- Choose one of three projects:
  - **IMDB Project** - learn to identify highly rated movies
  - **Reviews Project** - learn to identify helpful product reviews
  - **DBLP Project** - learn to identify duplicate entries in a bibliography
- Data for the projects available at https://github.com/schelterlabs/big-data-course-2022-projects
- Mimics real world setting (as a preparation for an industry job in data science): **data is disaggregated** over different files, **contains many errors** (synthetically generated missing data, typos, …)

Big Data Project - Logistics (3)

- Free to **use any programming language / ML library** you like
- You can use **additional data** (except for the original data source from which we generated the project data)
- **Submit predictions for the validation and test set online** at http://big-data-competitions.westeurope.cloudapp.azure.com:8080/
- Passwords will be given out by the TAs in the lab sessions
- TAs will support project work in the coming lab sessions, **each team should check-in with one TA each week**
- Each team can submit **up to 5 times per day**, server displays a **leaderboard for the accuracy on the validation set**
- **Test set score hidden, will be used as the final score (based on your last submission!)**

Big Data Project - Grading

- **Grade** will be **based on poster, pitch and discussion** during the poster session, as well as on **scores** on the submission server
- Each project will be **graded by two randomly assigned TAs**, final grade is the average of their assigned grades
- Focus on **innovation & data processing** (not on the ML model)
- **Four equally weighted sub-grades** (as outlined in syllabus):
  - **Innovation**
  - **Pitch & Poster Design**
  - **Pipeline Design**
  - **Analysis**
- Grading **rubric available** on the project page on canvas

Grading - Innovation / Pitch & Poster Design

- **Innovation: What is novel or interesting?**
  - What are **interesting and novel ideas** that you used in your project?
  - How do they relate to the course?

- **Pitch & Poster Design: Clear pitch? Helpful poster design?**
  - Poster should be easy-to-follow and clearly communicate your findings
  - Poster should be helpful for your pitch and discussion with the TAs

Grading - Pipeline Design

- **How reusable is your data pipeline?**
    - A well-written ML pipeline should implement a sequence of data processing operations to consume the input data, train the model and output predictions for the validation and test data
    - **Visualise your pipeline and its operations on the data with a diagram**
    - How much (manual) effort would it be to update your pipeline if the input data (or even its schema) changed?
- **How did you decide which parts of the pipeline to run in DuckDB / PySpark?**
    - Your pipeline should **use DuckDB and/or (Py)Spark in appropriate parts**
    - You should be able to **explain why (or why not) it makes sense to use DuckDB and/or (Py)Spark in a given part** (e.g., based on learnings from the course)

Grading - Analysis

- **How innovative/efficient/stable are your data integration, cleaning and preparation operations?**
    - **Which errors did you find in the data?** How does your pipeline **detect and fix** them?
    - Do you use **additional data** in your pipeline?
    - Present **experimental evidence** on how your data integration, cleaning and preparation techniques impact the data quality, stability or predictive performance of your pipeline

- **How good is your learning performance?**
    - Each leaderboard contains a submission based on random guessing and submission from a minimal pipeline created by one of our TAs
    - A good solution should **outperform random guessing and the TA baseline**
    - A good solution should **not overfit to the validation set** (e.g., validation and test accuracy should be close)