# Big Data Pipeline
## Group 40 - IMDB

Casper Dahmen, Julian El-Fasih, Kevin Tran, Adam Mehdi Arafan, and Maythe Dirkzwager

**IMDb**

## 1. Joining Data  (Pandas/DuckDB)
- CSV to Parquet
  - Parquet: efficiency, size reduction
- Using DuckDB with Parquet files
  - DuckDB + Parquet files = speed + scalability
- Joined training data files with writers/directors

## 2. Exploratory Data Analysis
- Distribution of features
- Possible outliers

## 3. Preprocessing (Pandas)
- Numerical columns to type int
- endYear + startYear = publishYear

## 4. Data cleaning (Pandas)
- Handling missing values
  - Median/mode per column
- Unidecode
  - Émpłóyéés' Éntráncé --> Employees' Entrance
- Lowercase, removal of punctuation and space
  - Employees' Entrance --> employeesentrance

**Logistic Regression (Spark) :**
Votes, runtime, year

**57%**

## 5. Writer/director scores (Spark)
- Average # votes the writers/directors of the movie have received on their movies
- Average # movies the writers/directors of the movie have worked on

**+ 5%**

## 6. New ML model (Spark):
Random Forest

**+ 4%**

## 7. Additional data sets
- Revenue, user ratings, genres
- Oscar wins/nominations

**+ 12%**

## 8. Upgraded imputation
- Mean Absolute Deviation per year
  - Votes and ratings

**+ 3.5%**

## 81.5% Validation Accuracy

# Final Pipeline

Joining Data → Preprocessing → Cleaning → Feature Engineering → Predicting

## Resources
- MovieLens dataset: https://grouplens.org/datasets/movielens/
- MovieLens revenue dataset: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
- Oscar dataset: https://github.com/ernieja/oscar-nominees

**pandas**   **Spark**   **DuckDB**