Data Science **Winter 2022/23**

Tom Zimmermann, Ph.D.
University of Cologne

Description

The goal of this course is to familiarize students with different components of the data science workflow. This workflow is represented differently by different people but we will focus on the following four aspects: Data acquisition, Data processing and Data Integration, Data exploration and Data Analysis. The course combines classical lectures with flipped classroom exercise sessions. Applications will cover a variety of fields, from labor economics to development economics to finance. The course will involve a fair amount of programming in R.

Schedule

We will meet once a week for 3 hours. Sessions will consist of lectures and exercises with the exact split between the two potentially varying between different sessions.

- Wednesdays, 1230 – 330pm
- Location: Hauptgebäude, Hörsaal XVIII
- Sessions are **cancelled** for Nov 16 (no room) and during the winter break

Session	Date	Topics
1	Oct 12	Intro / What is data? / R refresher
2	Oct 19	Data wrangling / The tidyverse
3	Oct 26	Data Acquisition / APIs and SQL
4	Nov 2	Exploratory Data Analysis and Data Visualization
5	Nov 9	Linear Regression I
6	Nov 23	Linear Regression II
7	Nov 30	ML Basics / Regularization and Cross-Validation
8	Dec 7	Regression Trees
9	Dec 14	Random Forest and Gradient Boosting
10	Dec 21	Neural Nets / TensorFlow
11	Jan 11	Classification
12	Jan 18	Text Analysis
13	Jan 25	Review
14	Feb 1	Exam

Literature

Main textbooks

- **[DABEP]** Békés and Kézdi: Data Analysis for Business, Economics, and Policy

Supplementary textbooks

- **[ISL]** James, Witten, Hastie and Tibshirani: An Introduction to Statistical Learning (available online)
- **[BDS]** Taddy: Business Data Science

Grading

Your grade will be based on

- A data presentation that you give in class (10%)
- Labs that you turn in (40%)
- A term paper at the end of the course (50%)

Data presentation: You can't spell big data without data. To get you up to speed with high-quality datasets that are publicly available and that can be used for research, you will be asked to prepare a short presentation on a dataset (in a group of 2-3 students). Issues that I expect to be covered are: What are these data? Where have they been used? How can I process them in R or Python easily? Can you provide a simple (descriptive) usage example?

Labs: Labs give you an opportunity to practice important concepts discussed during lectures and to apply them to different datasets. There will be 6 labs in total, and you will have to turn in at least 4 of them to pass this class (if you turn in more than 4 labs, we will count the best 4 towards your final grade). You should work on the labs in groups of 2-3 students, and turn your solutions in via email on the day before they are discussed in the section. A complete solution includes a short write-up and code.

Term paper: For the term paper, you will use your newly acquired skills on a real-world prediction challenge. The challenge and data will be described towards the end of the course. You will be asked to discuss your approach, your model and your results in at most five pages. Work in teams of 2-3 students.

Office hours

I will hold office hours on Thursdays **from 1:20-3pm** via Zoom.

You can sign up here in advance: https://calendly.com/tomzimmermann

I am also happy to chat before and after class.