BDPEMS
Berlin Doctoral Program in Economics and Management Science

SFB/Transregio 266
ACCOUNTING FOR TRANSPARENCY

1 Title

Statistical Programming and Open Science Methods

2 Faculty

Prof. Dr. Joachim Gassen (TRR 266 Accounting for Transparency, Humboldt-Universität zu Berlin)
https://www.wiwi.hu-berlin.de/rewe/staff/gassen
gassen@wiwi.hu-berlin.de

3 Outline

3.1 Learning Objectives

This course communicates how to develop data science applications that comply to the FAIR principles of open science. That means that they are findable, accessible, interoperable and reusable. After this course, participants should

- be able to use common collaboration tools in software development like Git and Github,
- understand how to use functional and object-oriented programming approaches to develop accessible code,
- be capable to develop test routines and debug code,
- have gained an understanding on how to profile code,
- have developed routines for standard data analysis tasks, like data scraping, cleaning and visualization, and
- have understood how to package statistical applications so that they are portable across platforms.

3.2 Course format

The course consists of two block sessions covering two days each and online tutorials, assignments and groupwork in between. Most of the class assignments will be based on R and/or Python but students are free to prepare their assignments using a statistical programming language of their choice.

4 Administration

4.1 Schedule (preliminary)

10.10.2019 (Day 1)

The development environment
Project organization
Using Git and Github

Statistical programming languages: An overview
Functional programming and object-oriented programming

11.10.2019 (Day 2)

Writing readable and reusable code
Debugging tools

The concept of normalized data
Data wrangling and data cleaning
Data visualization fundamentals

Online assignments (14.10.2018 – 14.02.2019)

Group assignment: Reproduction of a published study
Individual assignments: Development of tools for FAIR data science

17.02.2020 (date to be confirmed)

Testing
Profiling

Data scraping
Interactive visualization

18.02.2020 (date to be confirmed)

Simulations and effect size analysis
Explore your researcher degrees of freedom

Providing data access via RDBMS and APIs
Reproducibility by containerization

4.2 Location

HU Berlin, details to be announced

4.3 Application

While this course is targeted at incoming doctoral researchers of the TRR 266 "Accounting for Transparency", non TRR members at the doctoral and master level are free to attend, capacity permitting. **Please apply by September 2nd by sending an email including a brief CV and your current transcript to gassen@wiwi.hu-berlin.de**. I will inform students about their acceptance by September 4th.

5 Prerequisites

The course requires intermediate skills in statistics and econometrics.

In terms of data science experience, some knowledge of a statistical programming language (e.g., Python, R or Stata) is a plus but not a must. We will be predominantly working with Python and R during the seminar but students are free to use other languages for their assignments if they prefer. Students that are not familiar with either language are strongly encouraged to work through the opening chapters of "R for data science" prior to attending the class.

6 Preparatory Reading

Grolemund, G. and H. Wickham (2017): R for Data Science, O’Reilly: http://r4ds.had.co.nz.

7 To prepare

Students not familiar with statistical programming should work through the opening chapters (1-8) of “R for data science”. If feasible, students should bring a laptop to class with Python 3, R, RStudio, and git installed. All software packages are open source and freely available. A guide for setting this work environment up can be found at: http://happygitwithr.com.

8 Assessment

The grade will be based on the group assignments (50 %), individual assignments (25 %) and on active participation during class (25 %).

9 Credits

The course is eligible for 6 ECTS.