Discipline: Information Systems

1 Language

English

2 Title

Data Science as a Research Method

3 Lecturer

Professor Dr. Oliver Müller, Universität Paderborn
https://wiwi.uni-paderborn.de/dep3/mueller/team/

oliver.mueller@uni-paderborn.de

4 Date and Location

25.-28.10.2021

Universität Paderborn
LS für Wirtschaftsinformatik, insb. Data Analysis
Warburger Str. 100
33098 Paderborn

5 Course Description

5.1 Abstract and Learning Objectives

The course is targeted at PhD students and young researchers who want to apply data science methods in their research. It covers various data preparation, statistical modeling, and visualization techniques for extracting knowledge from the vast and complex data sets that have emerged in business over the past years. The learning objective of the course is to enable participants to apply these techniques in design-oriented and/or quantitative empirical research projects.

5.2 Content

Topics include supervised machine learning for regression and classification (linear and non-linear models), unsupervised machine learning (e.g., clustering, dimensionality reduction), ensemble models, AutoML, interpretability of machine learning, and natural language processing (e.g., dictionary-based methods, supervised text classification, topic modeling).

5.3 Schedule

Day I

- 10:00 – 10:30: Welcome
- 10:30 – 12:00: Introduction to data science
- 12:00 – 13:00: Lunch
- 13:00 – 14:00: Introduction to R
- 14:00 – 15:00: Supervised learning with linear models (i.e., linear and logistic regression)
- 15:00 – 16:00: Exercises with R
- 16:00 – 17:00: Assumptions and extensions of the linear model
- 17:00 – 18:00: Exercises with R

Day II

- 09:00 – 10:30: Tree-based models (incl. bagging, random forest, and boosting)
- 10:30 – 12:00: Exercises with R
- 12:00 – 13:00: Lunch
- 13:00 – 13:45: Ensemble models and AutoML
- 13:45 – 14:30: Interpretable machine learning
- 14:30 – 15:15: Exercises with R
- 15:15 – 16:15: Unsupervised learning with clustering and dimensionality reduction
- 16:15 – 17:00: Exercises with R

Day III

- 09:00 – 10:30: Natural language processing with dictionaries and the bag-of-words model
- 10:30 – 12:00: Exercises with R
- 12:00 – 13:00: Lunch
- 13:00 – 14:00: Natural language processing with probabilistic topic models
- 14:00 – 15:00: Exercises with R
- 15:00 – 16:00: Natural language processing with neural networks
- 16:00 – 17:00: Exercises with R

Day IV

- 09:00 – 12:00: Group work on data science mini-project (Hackathon)
- 12:00 – 13:00: Lunch
- 13:00 – 14:00: Group work on data science mini-project (Hackathon)
- 14:00 – 15:30: Presentation of data science mini-project
- 15:30 – 16:00: Wrap-up and feedback

5.4 Course format

The course consists of a mix of conceptual lectures, hands-on exercises with R, and discussions of research papers. Participants get the chance to directly apply the methods taught in the lectures to real-life datasets. In addition, we will discuss a number of papers applying data science methods. Hence, we strongly recommend participants to familiarize themselves with the papers listed in the reading list. The course language is English.

6 Preparation and Literature

6.1 Prerequisites

Master-level education in Business, Economics, Computer Science, Engineering or a related field.

6.2 Essential Reading Material

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning – with Applications in R. Springer.
  Free PDF version available at: http://www-bcf.usc.edu/~gareth/ISL/
  Relevant chapters: 1, 2, 3, 4, 5, 8, 10
- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 16(7), 16-07. Available at https://www.wired.com/2008/06/pb-theory/
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205.
- Müller, O., Junglas, I., vom Brocke, J., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems, 25(4), 289-302.
- Debortoli, S., Junglas, I., Müller, O., & vom Brocke, J. (2016). Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. Communications of the Association for Information Systems, 39(1), 7.

6.3 To prepare

All participants are required to read the essential reading material prior to the course. In addition, participants will receive a list of R programming assignments that have to be completed before the course.

7 Administration

7.1 Max. number of participants

The number of participants is limited to 20.

7.2 Assignments

The course will include a number of in-class assignments. These assignments will not be graded.

7.3 Exam

After the course, students are required to finalize and write-up the results of their Data Science Mini-Project (Day IV) in a working paper (10 pages). The working paper will be graded.

7.4 Credits

The course is eligible for 6 ECTS.

8 Working Hours

Working Hours	**Hours**
Mandatory readings *R programming assignments (to be completed before the course)* *Active participation in class* *Final exam (working paper to be written after the course)*	40 h 40 h 30 h 70 h
TOTAL	**180 h**
ECTS: 6	