Module Name: (B.5) Data Science and Analytics

Aim

Data Science and Data Analytics are both huge fields, but in this course we aim to cover a broad spectrum of its fundamentals. Should we assume a simplified division of Data Analysis into creating hypotheses and testing hypotheses, then, definitely this course aims at the former part: creating hypotheses or in other words, exploring data.

Learning Objectives

Through this course we will learn how to analyze data in order to support the difficult research processes and workflows, to support researchers by guiding them towards data understanding and eventually to formulating effective research questions. Data Science can hardly be considered independently of the related technologies. Therefore, in this course, all lecture will be realized by presenting the theoretical aspects in tandem with corresponding applications in the programming language R.

Learning Outcomes

Upon successful completion of the course, students should be able to:

- Reveal hidden yet important patterns through data sets after storing it in a consistent form that matches the semantics of the dataset, and visualizing the results.
- To create hypotheses for various research questions and explore data to test and validate them.
- To transform data including filtering, creating new variables that are functions of existing variables, and calculating a set of summary statistics.
- To discover relations among variables.
- To apply basic data analysis techniques like regression, classification, and clustering.
- To look for commands and packages of R to solve problems in various domains.
- Use a selection of R programming tools to combine with the data science principles to tackle interesting modelling problems.

Bibliography

[1] The pedagogy of the course is majorly based on the book: Dimitris Bertsimas, Allison O'Hair and Bill Pulleyblank, The Analytics Edge, Dynamic Ideas, 2016. ISBN: 978-0989910897

[2] Another excellent book that describes most of the techniques we will discuss in an intuitive way is: Evans, J. R. (2016). Business analytics. Pearson Higher Ed.[1]

[3] A more manager-oriented approach can be found at the (free or donate) book: Caffo, B., Peng, R. D., & Leek, R. H. (2016). *Executive data science: A guide to training and managing the best data scientists*. Leanpub https://leanpub.com/eds

[1] Of course, for each technique (Linear Regression, Logistic Regression, Trees, Clustering, etc.) there is a plethora of dedicated textbooks, but their focus is out of scope for this class…

[4] If you've never programmed before, you might find Hands on Programming with R by Garrett (https://rstudio-education.github.io/hopr/) to be a useful adjunct to this course.

[5] If you get stuck in particular with R, start with Google. Typically adding "R" to a query is enough to restrict it to relevant results: if the search isn't useful, it often means that there aren't any R-specific results available. Google is particularly useful for error messages. If you get an error message and you have no idea what it means, try googling it! Chances are that someone else has been confused by it in the past, and there will be help somewhere on the web. If Google doesn't help, try stackoverflow. Start by spending a little time searching for an existing answer, including [R] to restrict your search to questions and answers that use R.