Course Manual
Introduction to Applied Data Science
2022-2023

Bas Machielsen
a.h.machielsen@uu.nl

April 2023

Course Description

This course will introduce you to the modern data science toolkit. Data Science is an emergent discipline growing faster and faster due to the large amounts of data that we have at our disposal, and increases in computational power and flexibility. In the words of Vanderplas, data science is:

> fundamentally an interdisciplinary subject. Data science comprises three distinct and overlapping areas: the skills of a statistician who knows how to model and summarize datasets (which are growing ever larger); the skills of a computer scientist who can design and use algorithms to efficiently store, process, and visualize this data; and the domain expertise - what we might think of as "classical" training in a subject—necessary both to formulate the right questions and to put their answers in context.

The course will provide a **practical introduction** to the tools and techniques that are at the heart of modern data science. We will focus on aspects such as:

- **Data acquisition** by means of e.g. text mining, querying relational database, and web scraping.
- **Data wrangling and cleaning** to turn messy, disorganized data into tidy data that can be analyzed.
- Techniques of data analysis that are traditionally excluded from econometrics courses such as **spatial analysis and visualization**.
- Topics that teach you to have an effective workflow, like **version control and effective project management**.

The aim of the course is to be *broad* rather than *deep*, giving a broad overview of the tools available. More in-depth courses on programming, deep learning, and econometrics will be given in later courses. However, this course does plan to familiarize you with the *basic* aspects of these techniques, so as to facilitate the development of your skills.

Coming back to the *domain-specific* expertise: most of the applications and assignments in this course ask you to answer concrete economic questions. The philosophy behind these assignments is that you answer questions from the ground up, just like researchers do, and just like you will have to do at a later stage of your study. As such, this course also gives you an introduction into what economists do when they conduct *empirical research*.

Finally, what is left out of this course? We don't pay attention to two more advanced topics: cloud computation and parallel computing. Cloud computation concerns the outsourcing of your computational work to more powerful computers, often using commercial services. Parallel computing focuses on paralleling your computational tasks over different processor cores, often giving enormous speed improvements. This is also not a deep learning course. The course does not focus on neural networks and their derivatives to leverage artificial intelligence for economic goals.

In this course, we develop these skills using the R programming language, which has many advantages for Data Science. Firstly, it is a free open-source programming language with a large community of data scientists, economists and statisticians working together to develop state-of-the-art algorithms and packages. Secondly, there is a large market of companies and universities that require knowledge of R. Thirdly, **it is easy and logical**.

- This course features one weekly lecture (2 contact hours), and 1 tutorial (2 contact hours), and office hours, during which students can ask questions Thursday 16:00 - 17:00. You can also ask questions by e-mail.

Overview

Code	ECB1ID
Period	4
Timeslot	B (Tuesday Morning, Thursday Afternoon)
Level	1
ECTS	7.5
Course Type	Optimal Minor Course
Programme	BSc Economics & Business Economics
Department	U.S.E., Applied Economics
Coordinator/Lecturer	Bas Machielsen
Tutorial Teachers	Tba
Language	English

Course Materials

You don't need to buy any books for this course, but we use a couple of resources that you should read as a preparation for lectures/assignments. These are references materials that are regularly updated following the newest changes in the R community.

- R for Data Science: This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualize it and model it.
- RMarkdown Cookbook, which is designed to provide a range of examples on how to extend the functionality of your R Markdown documents.
- Happy Git With R: Happy Git provides opinionated instructions on how to install Git and get it working smoothly with GitHub, in the shell and in the RStudio IDE. It also contains a few key workflows that cover your most common tasks, and how to integrate Git and GitHub into your daily work with R and R Markdown.
- **Lecture Slides and Assignments** are available on Blackboard, but also here

Schedule and Syllabus

Event	Date	Subject	Materials
Lecture	1	Introduction to Data Science and Big Data	
Working Group	1	Setting up RStudio	
Lecture	2	Programming Basics	
Working Group	2	Using Tidyverse	
Lecture	3	Tables and Graphs	
Working Group	3	Getting Data, Visualizing and Summarizing	
Lecture	4	Programming Flow & Algorithms	
Working Group	4	Writing your own gradient descent	
Lecture	5	Debugging & Handling Errors	
Working Group	5	Debugging Strategies	
Lecture	6	Writing Reports in RMarkdown	
Working Group	6	Doing Data Analysis	
Lecture	7	Collaborating Effectively with Version Control	
Working Group	7	Creating and Changing a Github Repository	
Lecture	8	The Ethics of AI	
Working Group	8	Ethical Dilemmas	

Assignments

This course has three intermediate assignments. These will be uploaded to Blackboard, but they are also available [here]:

Assignment	Deadline
Assignment 1: The Causes of Economic Growth	tba
Assignment 2: Predicting Poverty	tba
Assignment 3: Collaborating Together	tba

Prerequisites

Mathematics, Statistics and introductory economics courses. This course is part of the Dedicated Minor in Applied Data Science for Economists. This is the first course, following which you will learn about:

- Introduction to Programming in R (ECB2PR, year 2, period 3)
- Data Analysis & Visualization I - Supervised Learning (ECB2ADAVE, year 2, period 4)
- Data Analysis & Visualization II - Unsupervised Learning (ECB3ADAVE2, year 3, period 1)
- Applied Microeconometric Techniques (year 3, period 2)
- Data Science Lab for Economics (year 3, period 3)

Learning Objectives

On effective completion of the course, students should:

- Understand principles of programming on an applied level
- Particularly, understand and be able to independently produce R code solving applied problems
- Being able to extrapolate the knowledge to other programming languages
- Gather, manipulate and wrangle untidy datasets
- Understand how to deal with errors and how to debug code

- Be able to read and exploit to their advantage code, package & function documentation
- Implement several elementary algorithms to solve concrete problems
- Be able to analyze and show results in a tidy and well-organized way
- Effectively collaborate together using version control
- Be able to reflect on the use of AI and Big Data in society

On successful attendance of the lectures, students should:

- have knowledge of the importance of Data Science and Machine learning.
- understand algorithmic thinking.
- understand the difference between causation and correlation
- be able to distinguish between descriptive, predictive, prescriptive and causal analysis.
- have awareness of the role of Data Science for Economics and its role in society.

Grading and Inspection

- The course will feature three individual assignments and one final exam. All assignments have to be handed in. The assignments have a weight of 60% and the final exam weight of 40% for the final grade. After the exam, students have the right to inspect their exam and assignment. Information about this will be announced in due time on Blackboard.

- All grades are rounded upward to two decimal places. Examples: 5.493 becomes 5.50, meaning *pass*, 5.490 becomes 5.49 meaning *fail*.

- If the final grade is below 5.50, there is a possibility of a resit, but only if the effort requirement is satisfied. No resit opportunity is possible for people obtained grades higher than 5.50.