Modern Workflow in Data Science
SURV 699Y
2 credit/4 ECTS
Summer Term 2020
Sec1

Instructor
Alexandru Cernat PhD, contact@alexcernat.com

Video lecture by Alexandru Cernat PhD

Short Course Description
Large data, fast pace of production, and collaboration are hallmarks of the new data environment. In this context, researchers must have a good understanding of data workflows and they must ensure consistent and reproducible practices in order to collaborate and consistently produce insights. This course deals with some of these essential topics. We will discuss the main types of workflows in data and survey sciences and how tools such as GitHub can enhance collaboration and insure reproducibility. We will also discuss the use of reproducible documents such as Rmarkdown or Jupyter Notebooks before covering the how to work with distributed data using Spark. We will finish the course by discussing the use of dashboards and how to develop such tools using R Shiny.

Course and Learning Objectives
By the end of the course, students will…

- Understand the main types of **workflows in data and survey sciences**
- Understand the **principles of reproducible workflows**
- Know how to **use Github to support reproducible flows**
- Understand the basics of **reproducible documents**
- Learn how to **use Rmarkdown and Jupyter Notebooks**
- Learn about **the main types of storage for online data** (e.g., SQL, JSON)
- Learn how to **access distributed clusters using Spark**
- Learn **how to manage computing clusters**
- Learn the **principles of building a dashboard**
- Learn how to build a **dashboard using R Shiny**

Prerequisites

SURV665 Real World Data Management with R or a good knowledge of R base and tidyverse.

Class Structure and Course Concept

This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video-recorded lectures and reading the required literature for each unit prior to participating in mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching prerecorded videos, attending the live online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. This is a 2-credit/4-ECTS course that runs for 8 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Tuesday, 3:00 PM EDT/9:00 PM CEST, starting June 2

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Monday, 3:00 PM EDT/9:00 PM CEST

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room

to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can use other online meeting platforms such as Google Hangout or Skype.

Grading

Grading will be based on:

- Four homework assignments (worth 60% total)
- Participation in discussion during the weekly online meetings and submission of questions via e-mail (deadline: Monday, 3:00 PM EDT/9:00 PM CEST before class) demonstrating understanding of the required readings and video lectures (10% of grade)
- A final project (30% of grade)

Grades will be assigned on the following scale:

A+ 100 - 97
A 96 - 93
A- 92 - 90
B+ 89 - 87
B 86 - 83
B- 82 - 80
Etc.

The grading scale is a base scale recommended by the IPSDS. Variations for grading on a scale are at the discretion of the instructor.

Dates of when assignment will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructor.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description

Working with large datasets, presenting insights and collaborating with others are essential skills for data and survey scientists. In this course you will learn some keys skills needed in this research environment.

We will start the course by discussing different types of data workflows. This will cover typical ways in which organizations produce, manipulate and report on data. Getting an overview of these practices and understanding how other organizations work can bring important insights that can make your own work better. We will then discuss emerging practices from reproducible research. Finally, we will discuss how tools such as Docker and GitHub can help collaboration and improve reproducibility.

The second topic covered in the course will be reproducible documents. These are essential tools that can be used to create reports, research papers, books and websites. They are vital for reproducible research and collaboration as they can combine text and code while enabling version control. In this way, typical errors due to copy and pasting and imprecise language can be avoided. We will discuss how to use this efficiently to write reports, presentations books and automated reporting. We will cover mainly Rmarkdown but will also briefly discuss Jupyter notebooks.

The third topic discussed will be about working with distributed. Many organizations store data on servers due to their size and speed of production. Often you will need to be able to interact with servers directly in order to access, clean and analyze data. We will discuss the main technologies for storing data (such as SQL and JSON) and how you can use Spark and R to work with distributed data.

The final topic of the course will be dashboards. These are important tools used to present data in an interactive and easy to read fashion. They are especially useful when data is collected at high speeds and decisions need to be made based on such data. It is a very useful tool also for presenting results to clients and a lay audience. Here we will be discussing how RShiny can be used to create such dashboards.

Each topic will be covered in two weeks. The first week will cover the online course and the reading materials. In the second week students will have to prepare a project based on what they learned in the first week.

Readings

Primary readings will be from the following volumes:

Bryan, J. (early release). *Happy Git and GitHub for the user* (https://happygitwithr.com/).

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The definitive guide*. Taylor & Francis, CRC Press. (bookdown.org/yihui/rmarkdown/).

Luraschi, J. (2020). *Mastering Spark with R: The complete guide to large-scale analysis and modeling*. O'Reilly Media. (https://therinspark.com/).

Wickham , H. (early release) *Mastering Shiny*. CRC press. (mastering-shiny.org/).

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland) and

https://www.uni-mannheim.de/en/research/good-research-practice/ (University of Mannheim).

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule

Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.

Unit 1: Data workflow with Github

Video lecture: available Tuesday, May 26

Online meeting: Tuesday, June 2, 3:00 PM EDT/9:00 PM CEST

Required Readings:

Bryan, J. (early release). *Happy Git and GitHub for the user* (https://happygitwithr.com/). Sections: I, II, III, IV & V

Unit 2: Practical 1

Online meeting: Tuesday, June 9, 3:00 PM EDT/9:00 PM CEST

Assignment 1: due Thursday, June 11, 3:00 PM EDT/9:00 PM CEST

Unit 3: Reproducible documents with Rmarkdown and Jupyter Notebooks

Video lecture: available Tuesday, June 9

Online meeting: Tuesday, June 16, 3:00 PM EDT/9:00 PM CEST

Required Readings:

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The definitive guide*. Taylor & Francis, CRC Press. (bookdown.org/yihui/rmarkdown/). **Chapters 1, 2, 3 & 15 (pp. 1-90, 181-198)**

Optional Readings:

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The definitive guide*. Taylor & Francis, CRC Press. (bookdown.org/yihui/rmarkdown/). **Chapters 4 & 7 (pp. 93-113, 117-135)**

Unit 4: Practical 2

Online meeting: Tuesday, June 23, 3:00 PM EDT/9:00 PM CEST

Assignment 2: due Thursday, June 25, 3:00 PM EDT/9:00 PM CEST

***** No online meeting on Tuesday, June 30, 2020 *****

Unit 5: Accessing data online

Video lecture: available Tuesday, June 23

Online meeting: Tuesday, July 7, 3:00 PM EDT/9:00 PM CEST

Required Readings:

Luraschi, J. (2020). Mastering Spark with R: The complete guide to large-scale analysis and modeling. O'Reilly Media. (https://therinspark.com/). **Chapters 1, 2, 3, 6, 7, 8 (pp. 1-52, 93-152)**

Optional Readings:

Luraschi, J. (2020). Mastering Spark with R: The complete guide to large-scale analysis and modeling. O'Reilly Media. (https://therinspark.com/) **Chapters 4 & 5 (pp. 53-90)**

Unit 6: Practical 3

Online meeting: Tuesday, July 14, 3:00 PM EDT/9:00 PM CEST

Assignment 3: due Thursday, July 16, 3:00 PM EDT/9:00 PM CEST

Unit 7: Interactive dashboards with Shiny

Video lecture: available Tuesday, July 14

Online meeting: Tuesday, July 21, 3:00 PM EDT/9:00 PM CEST

Required Readings:

Wickham , H. (early release) Mastering Shiny. CRC press.: (mastering-shiny.org/). **Chapters 1, 2, 3, 4, 6, 8, 9, 10**

Optional video:

Shiny in production: Principles, practices, and tools - Joe Cheng - https://www.youtube.com/watch?v=Wy3TY0gOmJw

Unit 8: Practical 4

Online meeting: Tuesday, July 28, 3:00 PM EDT/9:00 PM CEST

Assignment 4: due Thursday, July 30, 3:00 PM EDT/9:00 PM CEST

Final Exam
Due: Sunday, August 9, 3:00 PM EDT/9:00 PM CEST