Syllabus Fall 2021

Basic information

DEPARTMENT: Biostatistics and Bioinformatics

COURSE NUMBER: INFO550

COURSE TITLE: Data Science Toolkit

CREDIT HOURS: 2

SEMESTER: Fall 2022

CLASS HOURS AND LOCATION: Thursdays 3:00PM - 4:50PM 1525 Clifton Building, Alperin Auditorium

INSTRUCTOR NAME: David Benkeser

INSTRUCTOR CONTACT INFORMATION * **EMAIL**: benkeser@emory.edu * **PHONE**: (404)712-9975 * **SCHOOL ADDRESS OR MAILBOX LOCATION**: GCR 1518-002-3AA * **OFFICE HOURS**: TBD (location: GCR 322 or Zoom)

TEACHING ASSISTANT INFORMATION: * **NAME**: Mingze Liu * **EMAIL**: mingze.liu@emory.edu * **OFFICE HOURS**: TBD (location: TBD)

COURSE DESCRIPTION

This course is an elective for Masters and PhD students interested in learning several fundamental tools used in modern data science. Together, the tools covered in the course will provide the ability to develop fully reproducible pipelines for data analysis, from data processing and cleaning to analysis to result tables and summaries. By the end of the course students will have learned the tools necessary to: * develop reproducible workflows collaboratively using version control based on Git/GitHub; * execute workflows on a local computer using command line operations, RMarkdown, and GNU Makefiles; * execute workflows in a containerized environment using Docker; * execute the workflow in a cloud environment using Amazon Web Services EC2 and S3 services; * have confidence in seeking out help online to understand new data science tools.

Time-permitting, we may cover a additional tools for data science including basic python, Jupyter notebooks, cluster computing, R package development, or other related topics.

PRE-REQUISITES

Many topics covered will involve the R programming language and so familiarity with R is needed (e.g., BIOS 544/545 or similar level of competency).

Necessary skills include: reading data into R, basic data cleaning in R (e.g., subsetting data, finding missing values, merging data), operating on data.frames (e.g., changing column names, row names, summarizing rows/columns of data using simple statistics), basic graphics (e.g., plot or ggplot2).

Given the similarities between python and R, students with a background in python programming should also be equipped to succeed in the class, but will possibly require more effort to get up to speed with R.

COURSE LEARNING OBJECTIVES

- Understand why automation is a key element of reproducible data science.
- Operate capably and comfortably at the command line.
- Implement best practices for version control and open source projects.
- Produce reproducible workflows for data cleaning, analysis and report generation using the suite of tools learned in the class.
- Create simple R packages.
- Build Docker containers and use them to develop containerized workflows.
- Understand basic uses of bash.
- Utilize AWS cloud computing services for computation and storage.

CONCENTRATION COMPETENCIES:

- Develop public health information systems to support public health efforts
- Identify software for the interface of data entry and statistical analysis
- Assess individual data elements and display results effectively and appropriately
- Apply standard statistical methods in the analysis of public health information

ATTENDANCE POLICY

The class will meet in-person each week.

It is expected that students will attend the vast majority of classes synchronously and in-person. In general, there will not be a Zoom link available for in-class sessions. If you have exentuating circumstances that prevent synchronous attendance, please discuss with the professor ahead of time so temporary Zoom accommodations can be made.

If there is no way to attend synchronously, it is the student's responsibility to contact the professor to develop a plan for making up in-class assignments.

EVALUATION

Students will be evaluated based weekly assignments and a semester-long project. Submission of homeworks and assignments will occur via Canvas.

Weekly assignments

The weekly assignments will consist of two types of exercise.

1. Pre-read exercises: these are short, low stakes assignments designed to test very basic understanding of the preparation material each week. These assignments will generally be due before class begins each week. **I suggest reading through the pre-read exercises prior to going through the prepartion material**.
2. In class exercises: these will generally be group assignments that happen in-class and should be turned in during class time. These will usually be graded for completion, rather than correctness.

Final project

Students will develop a single project over the course of the semester. Each student should pick an analysis or workflow that they want to develop into a reproducible pipeline. The assignments will be graded week-by-week and will develop into a fully reproducible pipeline that can be executed in a containerized environment both locally and on the cloud.

Ideally, students enrolled in the course will have an existing analysis or data cleaning/visualization problem that they can use for the course (e.g., a thesis or dissertation project), so that the developments throughout the semester are specifically relevant to each student.

Project requirements Each project must involve the following three components (examples follow each separate component):

- Data manipulation – e.g., reading data from a file and cleaning the data; downloading data from the internet in an automated way; splitting data into training and test sets.
- Data analysis – e.g., making a Table 1 with descriptive summaries of key elements of the data; fitting regression models that answer scientific questions of interest; training machine learning algorithm and summarizing performance.
- Report writing – a report must be generated in the end that includes at least one table and at least one figure, along with writing.

FAQ on project selection: **Q**: *What if I don't have a data set/workflow to use?*

A: We will work with you to identify one. Possible examples could be based on a publicly available data set from an online data repository (e.g., UCI ML library or Kaggle). The workflow could involve producing a report that summarizes descriptive statistics and produces basic visualizations of the data, or one that involves fitting a ML algorithm to the data and summarizing its predictive performance.

Q: What if I have data, but I can't make it public?

A: You can create a mock data set with similar format to your real data, but with made up values. You will produce a pipeline for analyzing the mock data and, since everything we do is reproducible, you can swap in your real data at a later date and still take advantage of the pipeline.

Q: Does the analysis need to involve complicated statistical methods?

A: No. Basic analyses are fine. The goal is not to evaluate the contents of the analysis itself, but rather whether the analysis, whatever it may be, can be executed in a reproducible way.

If you have questions about the suitability of a project, please contact the professor.

Final project grading The final project will be developed over the course of five separate assignments throughout the semester.

Students will be responsible for producing:

1. (15 points) An R Markdown document that successfully compiles locally.
2. (15 points) A well-organized and documented GitHub repository that hosts the project
3. (15 points) A Make file that documents and can be used to compile the report
4. (15 points) An renv lockfile that handles package management for the project
5. (40 points) A publicly available Docker container that executes the analysis

Often, peer grading will be used to evaluate submissions and provide comments/feedback on its usability. Remember that the overarching goal is to develop workflows that are reproducible and easy for another researcher (or you, at a future date) to use. As such, you will be evaluating your peers on the usability of their submission each week. A standardized rubric for peer review will be provided each week as necessary. Students will receive points towards their own assignment by completing the peer grade.

Grade composition

Your grade will be determined as follows:

- 50% final project
- 25% pre-read assignments
- 25% in-class assignments

Grade scale

- A = 93 – 100%
- A- = 90 – 93%
- B+ = 87 – 90%
- B = 83 – 87%
- B- = 80 – 83%
- C = 65 – 80%
- F = <65%

Final grades are not rounded and the lower limit of each letter grade is inclusive, so e.g., 93.0 is an A, while 92.9 is an A-.

COURSE STRUCTURE

This semester, I will attempt to implement a flipped classroom. That is, much of the traditional, lecture-based content will be delivered outside of regular class hours, while in-class time will be devoted in large part to group exercises. Students will have weekly preparation to complete that may include watching videos and completing practice exercises. It is expected that students come to class prepared to engage in the in-class exercises. Because of the emphasis on in-class activities, synchronous attendance is required for success in this course.

Example assignments are shown below along with the associated core competency.

Assignment	Competency
Properly install software for reproducible research: R, R Studio, homebrew (Mac), Git, text editor (e.g., Notepad++ or Sublime Text), LaTex, GNU make	Identify software for the interface of data entry and statistical analysis
Build a GitHub repository, tag a release of your code, and write a proper README file	
Develop public health information systems to support public health efforts	
Develop an R Markdown document performing a reproducible statistical analysis and summarizing results visually	
Apply standard statistical methods in the analysis of public health information Assess individual data elements and display results effectively and appropriately	
Build a Docker container for running a fully reproducible analysis pipeline	Develop public health information systems to support public health efforts
Execute basic python commands in a Jupyter notebook.	

Assignment	Competency
Identify software for the interface of data entry and statistical analysis	
Set up an EC2 instance and S3 bucket on AWS suitable for a reproducible data analysis in the cloud.	
Develop public health information systems to support public health efforts	

COURSE POLICIES

Students are expected to attend lectures synchronously if able.

As the instructor of this course I endeavor to provide an inclusive learning environment. However, if you experience barriers to learning in this course, do not hesitate to discuss them with me and the Office for Equity and Inclusion, 404-727-9877.

Teaching and learning during the pandemic

Some of you may need to be off campus for some portion of the semester. I want our classroom community to thrive no matter the classroom delivery method or your individual methods of participating in class. I cannot guarantee an identical experience for students who cannot be physically in the classroom or an experience that is identical to pre-pandemic semesters, but my goal is to treat all students equitably and to ensure grading is clear, consistent, and fair for all of you.

As always, communication is important. I commit to responding to emails within 48 hours of receipt, and my intention to respond faster than that most of the time. I may be slower on weekends and over holidays. Likewise, if your situation changes regarding health, housing, or in any other regard with respect to your ability to participate in the class, please contact the appropriate Emory student support organization first and then me as soon as feasible. It is easier for me to address your needs if I know about them as soon as they arise. This does not mean I can successfully respond to every request for consideration, but I emphasize that my goal is to treat you all equitably and do what I can to help you succeed in this course.

Health considerations

At the very first sign of not feeling well, stay at home and reach out for a health consultation. Please consult the campus FAQ for how to get the health consultation. As you know, Emory does contact tracing if someone has been diagnosed with COVID-19. A close contact is defined as someone you spend more than 15 minutes with, at a distance less than 6 feet, not wearing facial coverings. This typically means your roommates, for example. However, your classmates are not close contacts as long as we are following the personal protective equipment protocols in the classroom: wearing facial coverings, staying six feet apart.

RSPH POLICIES

Accessibility and Accommodations

Accessibility Services works with students who have disabilities to provide reasonable accommodations. In order to receive consideration for reasonable accommodations, you must contact the Office of Accessibility Services (OAS). It is the responsibility of the student to register with OAS. Please note that accommodations are not retroactive and that disability accommodations are not provided until an accommodation letter has been processed.

Students who registered with OAS and have a letter outlining their academic accommodations are strongly encouraged to coordinate a meeting time with me to discuss a protocol to implement the accommodations as needed throughout the semester. This meeting should occur as early in the semester as possible.

Contact Accessibility Services for more information at (404) 727-9877 or accessibility@emory.edu. Additional information is available at the OAS website at http://equityandinclusion.emory.edu/access/students/index.html

Honor Code

You are bound by Emory University's Student Honor and Conduct Code. RSPH requires that all material submitted by a student fulfilling his or her academic course of study must be the original work of the student. Violations of academic honor include any action by a student indicating dishonesty or a lack of integrity in academic ethics. Academic dishonesty refers to cheating, plagiarizing, assisting other students without authorization, lying, tampering, or stealing in performing any academic work, and will not be tolerated under any circumstances.

The RSPH Honor Code states: "*Plagiarism is the act of presenting as one's own work the expression, words, or ideas of another person whether published or unpublished (including the work of another student). A writer's work should be regarded as his/her own property.*"

COURSE CALENDAR AND OUTLINE

It is difficult to provide an exact schedule, as we will modify pace/content to reflect the needs of the class. Here is a rough outline of topics you might expect and when to expect them.

Date	Topics
8/26	Motivation, text editor
9/2-9	Command line basics
9/9-16	R Markdown
9/23-30	Version control using Git/GitHub
9/30-10/7	GNU Makefiles
10/7-14	Best coding practices/project organization
10/14-21	Amazon web services
10/28-11/4	Containerization
11/11-18	Introduction to python
11/18	Jupyter notebooks
12/2	Additional topics