Reproducible Research in R (Biol 4800)

Location:

Time:

Instructor: Dr. Tad Dallas (tadallas@lsu.edu)

Office: A343 Life Sciences

Office hours: T and Th from 2:30 - 4:20pm

Course Overview:

Scientific knowledge builds upon existing scientific knowledge. This knowledge is generated through observation and experimentation, which results in scientific publications, but also in data and computer code. While the published record may persist, much of the computer code and data are no longer able to be confirmed due to a degradation of the analytical tools used (e.g., fortran punchcards) or unreadable data formats (e.g., some Microsoft Works data formats).

Only recently have scientists made their data and code available, allowing other scientists to examine the code and reproduce the results of the original paper. This ability to reproduce analyses is at the core of this course.

With the goal of doing reproducible research, we will learn many basic data science skills, including the unix command line, version control, makefiles, and the R programming language. Together, the student will gain the ability to create an analytical pipeline that can successfully run on any computer with sufficient memory.

Course Goals:

Over the course, it is expected that students gain

- a knowledge of the issues surrounding reproducibility and openness of scientific research
- an ability to design and implement reproducible scientific workflows
- experience with algorithmic thinking
- experience using the unix command line, R, and other computational tools
- an appreciation and/or deep hatred for reproducible research

Syllabus Subject to Change:

Changes to the syllabus may be made during the semester. The most up-to-date and current syllabus will always be available on the course website (on Github). Syllabus and grades will be available on Moodle (as standard).

Grading

There will be a total of 500 points, consisting of 5 assignments, some attendance/participation points, and a final project.

The final project will be performed in small groups for undergraduate students (4800), and independently for graduate students (7800).

Break down:

Item	Points	Total
Assignments	5 x 50	250
Attendance/Participation	75	75
Project	125	125
Final exam	50	50

Assignments:

There will be 5 assignments throughout the semester. *Each assignment will be worth 50 points.* The assignments will be submitted through Github, a version control platform. **Assignments are due before class on Thursday**

Attendance:

Much of the material presented will not be available if you aren't in class to hear it. Given the pace of the course, and the computational and programming hurdles you will almost certainly encounter, you should come to class. If you do not, you will likely fail. Part of the class will be discussing relevant papers. If you do not participate in paper discussions, clearly demonstrating that you have read and understand the nuances of the work, this will affect your grade. Attendance and participation will be worth 75 points.

Project:

The project will be a demonstration of your learning and allow you to use the new tools you have acquired in the course.

The final project will be an R markdown document that reads in and analyzes a data source. I can provide data, if necessary, but would prefer if data came from the student (either from their research or on something they are passionate about).

You will actively develop your work on Github, structuring your directory as we discussed in class. The final product will be an R markdown file that is entirely reproducible. This means I will clone your directory, and run your files on my local machine. **I must be able to reproduce your analyses**.

More information on the final project and its components are provided at the bottom of this document. The three core parts are 1) the proposal, 2) the implementation, and 3) the presentation

The project is worth 125 points

Final exam:

The final exam will go over high-level concepts learned in the class, and be worth 50 points.

Late Assignments:

All assignments are expected to be electronically submitted by the due date. I will not accept any assignment that is submitted after the deadline and the assignment will receive the grade of 0. I will, however, grade whatever the last committed (partial submission to the class assignment repository) that you have made prior to the due date. This grade cannot be appealed at a later date, but is better than a 0. This course is a senior level class and you are expected to work at the appropriate level and be responsible for your work. All deadlines are posted with the syllabus or any changes will be announced ahead of time.

Academic honesty

Louisiana State University adopted the Commitment to Community in 1995 to set forth guidelines for student behavior both inside and outside of the classroom. The Commitment to Community charges students to maintain high standards of academic and personal integrity. All students are expected to read and be familiar with the LSU Code of Student Conduct and Commitment to Community, found online at www.lsu.edu/saa. It is your responsibility as a student at LSU to know and understand the academic standards for our community.

Students who are suspected of violating the Code of Conduct will be referred to the office of Student Advocacy & Accountability. For undergraduate students, a first academic violation could result in a zero grade on the assignment or failing the class and disciplinary probation until graduation. For a second academic violation, the result could be suspension from LSU. For graduate students, suspension is the appropriate outcome for the first offense.

Further information is provided on the LSU website

Disability services

My goal is to help you learn. Students who have any difficulty (either permanent or temporary) that might affect their ability to perform in class can reach out to the LSU Disability Services staff.

More information on registering a disability is available at LSU Disability Services, located at 124 Johnston Hall. Contact the Center by telephone at 225-578-5919 or via email at `disability@lsu.edu`.

Schedule

Week	Topic	Assignment
1	Background	–
2	Intro to tools (git, Github, unix command line)	
3	R basics (including Rmd)	`rIntro`
4	Analytical pipeline structure (make, Travis)	`rPipeline`
5	R data manipulation	`rDataManip`
6	Unit testing and line profiling	`rTesting`
7	R, APIs, and data visualization	`rAndAPI`
8	R markdown and LaTeX	–
9	Dependency hell and containerization	–
10	Parallel computing	–
11	Final project work	–
12	Short presentations on final projects	–

Things not covered, but maybe can be covered given student interest:

- creating an R package
- code benchmarking (we cover line profiling but not benchmarking)
- handling spatial data
-

Final project

The final project will consist of 3 parts. First, you will develop a project proposal. Second, you will do the proposed project, and the final product will be a versioned and end-to-end reproducible analytical workflow that could potentially lead to a peer-reviewed publication. Lastly, you will briefly present your work to the class.

This means that I will expect not only a well-written and clear manuscript, but also documentation to easily reproduce all analyses, figures, and compile manuscript text. Some things I would like to see incorporated into the final projects:

- a clear README and informative file structure
- continuous integration or a Dockerfile
- a LICENSE file
- visualizations (reproducible plots within the R markdown file)
- version control (everything on Github, with clear commit history)
- R and R markdown
- no proprietary software or format (if you turn in a .docx file, I will be quite sad)

Final project proposal

Please prepare a short proposal on your final project idea by September 17. The proposal should include:

- Title & description of the project
- Team members names
- A description of the data required, and how it will be obtained (e.g. URL/DOI to data source)
- 3 questions / analysis tasks you will perform on the data; in the spirit of the assignments we have been doing.

Please create your proposal in a markdown file called `proposal.md` in the root directory of the final project repo.

Project Guidelines

Project questions must illustrate all of the following tasks:

- Some form of data access / reading into R
- Data manipulation
- Data visualization

- Use of GitHub
- Reproducible execution with use of Travis and/or Docker
- RMarkdown writeup, with final submission as a nicely formatted PDF document that includes code and results.
- Overall clean and clear presentation of repository, code, and explanations.

and many of the following skills

- Writing / working with custom R functions
- Creating an R package for functions used in the analysis
- Interaction with an API
- Use of regular expressions
- Use of an external relational database
- Preparing processed data for archiving / publication
- Parsing extensible data formats (JSON, XML)
- Making layout and presentation into secondary output (e.g. .pdf, website) - useful for final presentation

Final Rubric 125 pts total

proposal

- 25pts Proposal, turned in on time and with team member names, background, dataset, and outline of questions

implementation

- 25pts Polished github repository, including:
- 10pt: project runs on any machine (through Travis-CI and/or Docker file)
- 10pt: clean and well formatted repo (clear README file and file structure)
- 5pt: clear commit messages in project repo
- 50 pts Project Substance: Objectives, Code, Visualization.
- 10pt: project is correct scope, addresses the question, and has tangible output.
- 10pt: Visualizations and tables compliment the overall findings.
- 10pt: References are included.
- 10pt: The final document is in a format that can be reproduced and is versionable.

- 10pt: Code is clear, well-documented, and reproducible

presentation

- 25pts Final presentation
- 10pt: clarity of presentation
- 10pt: presentation is engaging, structured well, and highlights project findings and reproducibility
- 5pt: style