Tools for Reproducible and Collaborative Science

PSYC GR6030

Fall 2019 (3 points)
Fridays, 4:10pm-6:00pm, Room 200C

Instructors: Christopher Baldassano [c.baldassano@columbia.edu]
Agnes Chang [agneshchang@gmail.com]

Prerequisites: Instructor's permission and basic programming proficiency (i.e., understanding of variables, functions, and text input/output, preferably in python and/or R). Enrollment will be limited to a maximum of 12 students, with priority given to Psychology graduate students and senior graduate students.

Course description: This seminar provides a hands-on survey of recently-developed tools for reporting scientific results accurately, reproducibly, and collaboratively.

Detailed description: Traditionally, the goal of a scientific project has been to produce a static document with text and images describing and visualizing the data that was collected. This type of article, however, lacks any documentation about the chain of analyses that were conducted to go from the original raw data to the plots and statistical tests, and it can be nearly impossible for other researchers or even the original authors to precisely reconstruct the reported results. Students in this seminar will gain hands-on experience with new software tools that allow for the creation of analysis pipelines that can be verified, reproduced, and shared. Each week will include both a discussion of specific tools for reproducible research and then a tutorial session in which small groups of students will experiment with these tools. Every student will conduct a novel analysis on a previously-collected dataset and produce a fully reproducible manuscript as a final project, which will be presented during the final class session.

Course goals and learning objectives: This seminar will involve building practical and technical skills and will also provide opportunities for discussion about what research and publications could or should look like in the near future.

- A primary focus of the course will be to build proficiency with specific tools that have been developed in the past several years for reproducible analysis, especially those most relevant to psychology and neuroscience.
- In-class lectures will describe the motivations behind the design of these tools and highlight the ways in which they are currently being used by scientists in the field.
- Through in-class discussions, this course will help students brainstorm about the ways in which these tools could be applied to their own research projects, and to identify ways in which they can help lead the field toward better practices
- The final project will help students integrate the topics of the class in the context of a real data analysis, and provide opportunity for feedback from their peers

Grading:

- **25% Attendance and participation:** You are expected to come to class having read and thought about the week's assigned readings, and you should contribute to class discussions each day. If you are unable to attend class (for example, due to a religious holiday or illness), you should email me **before** class to avoid losing participation points for that session. To avoid distraction for both yourself and your fellow students, cell phones should not be used during class. Laptops are recommended for viewing papers, taking notes, and completing the tutorials, but should not be used for messaging or social networking.

- **15% In-class tutorials:** Each week, students will spend the second half of the seminar time working on a tutorial in small groups (randomly assigned by the instructor each week). These tutorials will focus on a different tool each week, and will involve both directed exercises and open experimentation. Students are expected to complete these tutorials, and assist other group members in learning how to use and apply these software tools.

- **60% Course project:** Throughout the course, each student will work on a research project using the tools being discussed. The specific content of this project can vary, but should make use of a dataset in psychology or neuroscience (publicly available, or from their own research) and should involve some novel and non-trivial analyses. The components of the final project grade are:
    - **10% Project proposal, *due 9/20/19***: The project proposal should be a one-page document describing the dataset that will be used, the scientific question that the final project will address, and the scope of the planned analyses. Each student will receive detailed feedback on their proposal, including suggestions on which specific tools relevant to the seminar could be relevant to completing their project.
    - **10% Interim project report, *due 10/25/19*:** Approximately midway through the seminar, each student will submit a two-page document describing their progress toward their project goals. This report should describe what has been accomplished so far (including writing text for the final project, conducting analyses, and creating visualizations) and lay out a plan for completing the project on time. Each student will have a 30-minute one-on-one meeting with the instructor after submitting this report, to get feedback on their analyses and project plan.
    - **30% Final project, *due 12/6/19*:** The final project should consist of an approximately 10-page report detailing the goals and results of the research study, and more importantly a *reproducible pipeline for generating this report*. The precise format of this pipeline can be chosen by the student (e.g. a CodeOcean capsule, a docker container, a GitHub repository with an environment specification) but should allow the instructor to easily regenerate the report pdf from the original data. Projects will be evaluated on:
        - Description of the scientific problem being addressed and the motivation for choosing this dataset for analysis
        - Design and correctness of the analyses conducted

        - Quality and clarity of the plots and visualizations in the report (which should all be exactly reproducible without manual intervention)
        - Description of conclusions from the analysis and connections to past or future work
        - Organization, documentation, and code quality of the project
        - Design of reproducible pipeline and ease of reproducing results
    - **10% Final project presentation, *12/6/19*:** Each student will give a presentation of approximately 10 minutes during the final class session, describing both the scientific and technical aspects of their project. Presentations will be graded based on: coverage of the necessary scientific background, description of the goals of the analysis, explanation of the technical design decisions made in the project, and clarity of the presentation of the final results.

Course rationale: This seminar is designed to train graduate students in recent techniques for creating and sharing data analyses. Given the rapidly-changing norms in psychology regarding verification and reproducibility, having hands-on knowledge of these tools is becoming increasingly important for conducting research. Discussing and experimenting with these approaches in groups will allow students to be at the forefront of ongoing methodological improvements in the field.

Academic integrity: Maintaining academic integrity is a critical responsibility of all Columbia students. Academic dishonesty includes (but is not limited to): plagiarism (using another person's work without attribution), misrepresentation of authorship (e.g., having work prepared by or purchased from someone else), and lying about completion of work (e.g., claiming that you submitted a post when you did not, or purposefully submitting a corrupted file to obtain more time to complete an assignment). Violations of the Honor Code may not only result in a failing grade for this course, but can also lead to serious disciplinary actions from the University, including expulsion. If you are falling behind in the course, know that you will be unable to finish work on time, or otherwise feel that you cannot complete your work, please talk to me as soon as possible to make a plan, rather than taking actions that will jeopardize your entire academic career.

Students with disabilities: In order to receive disability-related academic accommodations, students must first be registered with Disability Services (DS). More information on the DS registration process is available online at www.health.columbia.edu/ods. I must be notified of registered students' accommodations before exam or other accommodations will be provided. Students who have, or think they may have, a disability are invited to contact DS for a confidential discussion at (212) 854-2388 (Voice/TTY) or by email at disability@columbia.edu.

Weekly topics and readings

Date	Topic	Reading	Due
9/6/19	Introduction	Parker, H. (2017). *Opinionated analysis development* (No. e3210v1). https://doi.org/10.7287/peerj.preprints.3210v1	
9/13/19	Notebooks for python and R *Jupyter, Jupyterlab*	Pryke, B. (2018). *Jupyter Notebook for Beginners: A Tutorial*.	
9/20/19	Version control, collaboration *git, GitHub*	GitHub (2017). *Git Handbook*.	Project proposal
9/27/19	Reproducible plots in R *ggplot*	Wickham, H. (2010). *A Layered Grammar of Graphics*. Journal of Computational and Graphical Statistics *9*(1), 3–28.	
10/4/19	Reproducible plots in python *matplotlib, seaborn, plot*	Hamrick, J. (2016). *Reproducible, Publication-Quality Plots With Matplotlib and Seaborn*.	
10/11/19	Text editors and IDEs *Atom, PyCharm*	Kroger, P. (2015). *Running, Debugging, and Testing*. In Modern Python Development with PyCharm.	
10/18/19	Reproducible reports *RMarkdown, Pweave*	Turner, S. and Nagraj, V.P. (2018). *Reproducible Reporting with RMarkdown*. In UVA HSL Workshop Series.	
10/25/19	Reproducible manuscripts *Pandoc, make*	Barugahare, A., Harrison, P., and Tyagi, S. (2017). *Automation with makefiles*. In Using Open Science in Bioinformatics Training workshop.	Interim project report
11/1/19	Verification through simulation *psych, fmrisim*	Lotterhos, K. E., Moore, J. H., & Stapleton, A. E. (2018). *Analysis validation has been neglected in the Age of Reproducibility*. PLoS Biology, 16(12), e3000070.	
11/8/19	Verification through testing *pytest, testthat*	Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., … Wilson, P. (2014). *Best practices for scientific computing*. PLoS Biology, 12(1), e1001745.	
11/15/19	Reproducible containers *Docker*	Kasireddy, P. (2016). *A Beginner-Friendly Introduction to Containers, VMs and Docker*.	
11/22/19	Sharing reproducible pipelines *Dockerhub, CodeOcean*	Clyburne-Sherin, A., Fei, X., & Green, S. A. (2018). *Computational Reproducibility via Containers in Social Psychology*. https://doi.org/10.31234/osf.io/mf82t	
12/6/19	Final project presentations		Final project