Statistics 159 & 259 — Fall 2017 Syllabus
Reproducible and Collaborative Statistical Data Science

Class meets TuTh 9:30–11A in 3106 Etcheverry
Lab meets W 9–11A or 11–1P in 330 Evans

Professor: Fernando Pérez, http://fperez.org
Email: fernando.perez@berkeley.edu
Office Location: 419 Evans Hall
Office Hours: Tu 11A and W3P in 419 Evans
GSI: Elijahu Ben-Michael
GSI Office Hours: Mon 9-11A, Fri 3-5P, both in 342 Evans
Email: ebenmichael@berkeley.edu

I reserve the right to make changes to the syllabus.

Course Description: A project-based introduction to statistical data science. Through lectures, computational laboratories, readings, homeworks, and a group project, you will learn practical techniques and tools for producing statistically sound and appropriate, reproducible, and verifiable computational answers to scientific questions. The course emphasizes version control, testing, process automation, code review, and collaborative programming. Software tools include Bash, Git, Python, Jupyter and LaTeX.
Prerequisites: Statistics 133, Statistics 134, and Statistics 135 (or equivalent). Graduate standing is required to register for Statistics 259.
Credit Hours: 4

Text(s): Readings will be assigned weekly and will mostly consist of articles and tutorials.

Course Objectives:
At the completion of this course, students will:

1. understand the issues regarding reproducible research in modern scientific practice, including the definitions of key concepts and the different challenges that exist across disciplines
2. understand the computational and statistical issues involved with reproducibility
3. be proficient at the Unix commandline
4. be proficient at version control with Git
5. be able to write documents in Markdown or LaTeX (including using pandoc)
6. be familiar with scientific computing in Python

Grading:

Reading	10%
Quiz	10%
Homework	10%
Project 1	10%
Project 2	20%
Project 3	40%

Readings: For each assigned reading, you will submit a 2 paragraph report by 21:00 on the Thursday it is due. The first paragraph should summarize the reading. The second paragraph should briefly explore something that interested you (e.g., you may wish to focus on one aspect of the paper in more depth, you may wish to discuss something in the reading that you disagree with).

Quizzes: Quizzes will be held during class or lab. I will drop your lowest score.

Homework: There will be small individual homework assignments (to be submitted by 21:00 on the Thursday it is due) as well as three team projects. You may discuss the homework assignments with classmates, but you will be required to work on the homework independently and prepare an individual submission.

Projects: During the course, you will work on three, increasingly complex projects:
1. In pairs, you will attempt to replicate a paper from the scientific literature, will document your process, and will write a short report on your process and findings. I will provide you with suggestions for what journals and papers to consider, but you will have freedom to choose one that interests you.
2. In teams of three, you will complete the analysis of a dataset and will produce both a written report as well as a set of artifacts (code, figures, etc.) that should be fully reproducible. You will choose which dataset to work on from a few options given by the instructor.
3. Final project: in teams of five, you will work on an in-depth replication and post-publication peer-review of an existing paper in the literature. All teams will work on the same paper, but independently of each other.
For the first project you can choose your own work partner; for the second and third, I will make the team assignments. All members of any team must be familiar with the entirety of the work, I may call upon any member to discuss any aspect of the work and you should be able to demonstrate reasonable familiarity with areas you didn't work on, as well as complete expertise on the aspects you contriubted to.

Course Policies:

Attendance and behavior in class: You are expected to attend all lectures and labs. Any known or potential extracurricular conflicts should be discussed in person with me during the first two weeks of the semester, or as soon as they arise. **Cellphones** are to be silenced during class time and should not be used at all (if you have an emergency, step outside of the classroom to handle it and notify me afterwards). **Laptop** use during class will often be required, but should be used for course work only (i.e., not for surfing the web).

Submission of assignments: Assignments will be accepted by electronic submission to GitHub only. There will be no makeup quizzes. No late reading reports or homeworks will be accepted.

Academic integrity: Any test, paper, or report submitted by you is presumed to be your own original work that has not previously been submitted for credit in another course. While you are encouraged to work together on homework assignments, the work and writeup must be your own. For example, suggesting a function to another student is acceptable, whereas simply giving him or her your own code is not. If you are not clear about the expectations for completing an assignment or taking a quiz, be sure to seek clarification from me or the GSI beforehand. Any evidence of

cheating and plagiarism will be subject to disciplinary action. Please read the (brief) Honor Code (http://teaching.berkeley.edu/berkeley-honor-code) and its accompanying discussion carefully. Ask me if you have any questions.

Class discussion: Rather than emailing questions to the teaching staff, you should post your questions on Piazza (the class page is at: https://piazza.com/berkeley/fall2017/stat159259/home). When asking questions, especially regarding problems with code, make every effort to be clear and to provide sufficient information for others to be able to understand the problem you are having. This may include providing a minimal amount of code to replicate your problem, or copies, including screenshots, of the results you are getting. You may find this document useful, it contains tips on how to phrase questions regarding code and computation in an effective manner: https://www.mikeash.com/getting_answers.html.

Students with disabilities: If you need accommodations, please make arrangements in at timely manner through DSP. If your DSP oficer has already communicated your accommodation letter to me, I will have it in my files. I am happy to talk privately to you to work out the specifics of your situation, to ensure you have the support you need.

A note on Hurricane Harvey: if you or your loved ones have been affected by the situation unfolding in Texas and Louisiana, I am happy to make necessary accommodations. Contact me for an appointment and we can discuss the matter privately.