Statistics 159 & 259 — Fall 2015 Syllabus
Reproducible and Collaborative Statistical Data Science

CCN: 87680 (Stat 159) and 87812 (Stat 259)
Class meets TuTh 9:30–11A in 150 GSPP
Lab meets M 10–12P or 12–2P in 340 EVANS

K. Jarrod Millman
http://www.jarrodmillman.com
Office Location: 210 Barker Hall
Office Hours: TBD

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. Course emphasizes version control, testing, process automation, code review, and collaborative programming. Software tools include Bash, Git, Python, 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. be proficient at the Unix commandline
2. be expert at version control with Git
3. be able to write documents in Markdown or LaTeX(including using pandoc)
4. be familiar with scientific computing in Python
5. understand the computational and statistical issues involved with reproducibility
6. be familiar with computational issues in modern statistical data analysis through hands-on analysis of functional MRI data

Grading:

Reading	10%
Quiz	15%
Homework	20%
Project	55%

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 will be held during class or lab. I will drop your lowest score.
There will be 2 homeworks (to be submitted by 21:00 on the Thursday it is due), which will involve a substantial amount of effort. You may discuss the homework with your classmates, but you will be required to work on the homework independently.
The majority of the class focuses on a final group project, which is explained in more detail here: `http://www.jarrodmillman.com/stat159-fall2015/pages/project.html`

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 turned off during class time. **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 GSI beforehand. Any evidence of cheating and plagiarism will be subject to disciplinary action. Please read the Honor Code (`http://asuc.org/honorcode/index.php`) carefully.

Class discussion: Rather than emailing questions to the teaching staff, you should post your questions on Piazza. Please read Eric Raymond's "How To Ask Questions The Smart Way" (`http://www.catb.org/esr/faqs/smart-questions.html`).
Find our class page at: `https://piazza.com/berkeley/fall2015/stat159/home`

Students with disabilities: If you need accommodations, please make arrangements in at timely manner through DSP.

Important Dates:

Form teams	Sept. 22
Homework 1	Sept. 24
Project proposal	Oct. 1
Homework 2	Oct. 26
Progress presentation	Nov. 12
Draft report	Nov. 12
BIC tour	Nov. 23
Project presentation	Dec. 1 & 3
Final report	Dec. 14

Tentative Course Outline:
The weekly coverage might change as it depends on the progress of the class.

Week	Content
Week 1	• Course introduction
Week 2	• Introduction to Git; Using the bash shell • **Reading 1**: L Preeyanon, AB Pyrkosz, and CT Brown. "Reproducible bioinformatics research for biologists." Implementing Reproducible Research (2014)
Week 3	• Statistical analysis of fMRI • **Reading 2**: MA Lindquist. "The statistical analysis of fMRI data." (2008)
Week 4	• Introduction to Python • **Reading 3**: F Pérez, BE Granger, and JD Hunter. "Python: an ecosystem for scientific computing." (2011)
Week 5	• Scientific computing with Python • **Form teams** • **Homework 1**
Week 6	• Collaborative workflow with Git • **Project proposal**
Week 7	• Exploratory data analysis • **Reading 4**: JB Buckheit and DL Donoho. "Wavelab and reproducible research." (1995)
Week 8	• Project organization, process automation • **Reading 5**: KJ Millman and F Pérez. "Developing open source scientific practice." (2014)
Week 9	• Statistical analysis • **Homework 2**
Week 10	• TBD
Week 11	• **Project progress presentation**
Week 12	• TBD • **Draft report**
Week 13	• TBD
Week 14	• TBD
Week 15	• **Project presentation**
Week 16	• **RR Week**
Week 17	• **Final report due Monday**