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:

Quiz	15%
Reading	20%
Homework	25%
Project	40%

For each assigned reading, you will submit a short ($\sim$ 2 paragraphs) report.
Quizzes will be held during class or lab unannounced. We will drop your two lowest scores.

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 is recommend, but it is expected that you will be using your laptop to type along with the lecture.

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. Grades of Incomplete will be granted only for dire medical or personal emergencies that cause you to miss the final project presentation, and only if your work up to that point has been satisfactory.

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: We will be using Piazza for class discussion. Rather than emailing questions to the teaching staff, you should post your questions on Piazza.
To help your fellow students, the GSI, and me in responding to your questions, you should review 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/statistics159/home`

Students with disabilities: If you need accommodations for any physical, psychological, or learning disability, please speak to me after class or during office hours so that we can make the necessary arrangements.

Important Dates:

Form teams	Sept. 17
Homework 1	Sept. 24
Project proposal	Oct. 1
Homework 2	Oct. 29
Progress report	Nov. 3 & 5
Homework 3	Nov. 19
Project presentation	Dec. 1 & 3
Project report	Dec. 18

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

Week	Content
Week 1	• Course overview, introduction to Unix
Week 2	• Basic Git and documentation tools • **Reading 1**: L Preeyanon, AB Pyrkosz, and CT Brown. "Reproducible bioinformatics research for biologists." (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) • **Form teams**
Week 5	• Scientific computing with Python I • **Homework 1**: Twitter
Week 6	• Scientific computing with Python II • **Project proposal**
Week 7	• Collaborative workflow with Git • **Reading 4**: KJ Millman and F Pérez. "Developing open source scientific practice." (2014)
Week 8	• Project organization, process automation • **Reading 5**: V Stodden. "What computational scientists need to know about intellectual property law: A primer." (2014)
Week 9	• Exploratory data analysis • **Reading 6**: DA Freedman. "Statistics and the scientific method." (1985)
Week 10	• Statistical analysis I • **Homework 2**: Potti et al. cancer data
Week 11	• **Project progress report**
Week 12	• Statistical analysis II • **Reading 7**: JB Buckheit and DL Donoho. "Wavelab and reproducible research." 1995.
Week 13	• Model selection / validation ? • **Homework 3**: fMRI related?
Week 14	• Final thoughts/Selective inference, Thanksgiving • **Reading 8 (Extra credit)**: JW Tukey. "The future of data analysis." (1962)
Week 15	• **Project presentation**