SML 201 | Introduction to Data Science

Fall 2017

Overview[1][2]

This course provides an introduction to the burgeoning field of data science, which is primarily concerned with data-driven discovery and utilizing data as a research and technology development tool. We cover approaches and techniques for obtaining, organizing, exploring, and analyzing data, as well as creating tools based on data. Elements of statistics, machine learning, and statistical computing form the basis of the course content. We consider applications in the natural sciences, social sciences, and engineering.

Prerequisites

There are no official prerequisites for this course. However, it is recommended that students be comfortable with the basics of command-line computer usage and undergraduate-level mathematical notation and symbols (e.g., summation and product symbols, algebraic notation). Note: COS 126 or an equivalent course is NOT a prerequisite.

Instructors

Course Instructor

Daisy Huang
Email: yanhuang@princeton.edu
Office location: Room 201 in 26 Prospect Ave

Head AI

Darl Lewis
Email: glewis@princeton.edu
Office location: 224 Roberson (different loaction for OHs)

[1] The previous version of this syllabus was modified based on the version created by John Storey for SML 201 in Spring 2016.

[2] The course instructor reserves the right to make changes to this syllabus at anytime; this version is created on Sept. 14, 2017.

AIs

Libby L. Barak
Email: lbarak@princeton.edu
Office hour location: Lounge in 26 Prospect Ave

Yeohee Im
Email: yeoheei@princeton.edu
Office location: EQuad F218

Puneet Singh
Email: puneets@princeton.edu
Office hour location: Lounge in 26 Prospect Ave

Vennela Devabhaktuni
Email: vennelad@princeton.edu

(Temporary) Office Hours

	M 12:30-2pm	T 12:30-2pm	W 3-4:30pm	Th 6-7:30pm	F 2:30-4
Instructor	Daisy	Libby	Puneet	Darl	Yeohee
Location	Room 201 in 26 Prospect Ave	Lounge in 26 Prospect Ave	Lounge in 26 Prospect Ave	Lounge in 26 Prospect Ave	EQuad F218

Getting Help

Piazza

Students can sign up at http://piazza.com/princeton/fall2017/sml201. Use Piazza to ask and answer other students' questions about the course under on the Q&A page. Questions are anonymous to other students but not to the instructors, so please apply proper etiquette when posting. Instead of emailing the instructors with questions, we prefer that you post on Piazza because other students may have similar issues or have insightful contributions.

McGraw Center

McGraw Center offers R-Programming Tutoring on Sunday and Monday nights 7:30-10:30pm in Frist.

http://mcgraw.princeton.edu/undergraduates/group-and-individual-tutoring/group-study-hall

Tutors will not help you with the problem sets or projects for the course, but will help you on more general tasks, such as, setting up R and R Studio, reading help-manual in R, understanding R codes used in lectures, or going over solutions posted by instructors for precept exercises, problem sets and projects.

Introductory level R programming workshops

Departments of Politics and Sociology offer introductory level R programming workshops this semester. You can learn basic programming skills by participating their workshops. The first workshop will be held on Thursdays from 7:30 pm to 9:00 pm at Friend Center 101. Students can check the website (https://compass-workshops.github.io/info/) to see if any change takes place.

Projects

For each project you are not allowed to seek help from the instructors or anyone else other than your group mates for the project. You should treat projects as take-home exams. General questions (such as how to read the help manual for a specific function) are acceptable but no specific question about a particular problem.

Grading

Your grade will be determined by the following:

- 4 Problem sets (10% each) 40%
- 3 Projects (mini project 10%-midterm project 15%-final project 20%) 45%
- In-class clicker quizzes 10%
- Precept participation 5%

up to 2% extra credit for providing correct answers to other students' questions on piazza; extra credit will be awarded to the top 10 students on Piazza.

Your letter grade will only be determined by your final percentage based on the above items.

Quizzes

There will be in-class clicker-format quiz questions throughout the semester in most lectures. They will be closed notes, computers, phones, etc. The quizzes may be given at any point during the lectures. The material on the quizzes will be limited to the content of the current or previous lecture. If you are attending the lectures reviewing the previous lecture and completing the reading assignments beforehand, then the quizzes will be straightforward for you. Only the top 70% of your quiz scores will be counted (See Lecture under Attendance below).

Questions about your problem set or project scores

If you do not understand why you get points off for a certain question, please contact Puneet *and* Vennela since they are the ones grading the problem sets and projects. If you and the grader cannot agree on the outcome please see Darl, the Head AI.

Attendance

Lecture

You are expected to attend every lecture. However, we understand that sometimes you might miss a class because of sickness, job interviews, conferences, athletic competitions, personal emergencies etc. This is why we only count the top 70% of your quiz scores. Note that 30% of 24 lectures is approximately 7 lectures. Therefore, unless you anticipate missing more than 7 lectures you do *not* need to notify us about your absences. In case that you anticipate missing more than 7 lectures arrangement must be made *in advance* with the course instructor. After missing a lecture, you are encouraged to get notes from your fellow classmates and seek help from one of the instructors during his/her office hours. This course moves fast so do not get behind. Also, poor time management or having to do the work on materials from another class is not an excuse to miss the lectures. **Use of cellphones or laptops is prohibited during a lecture unless instructed otherwise.**

Precept

You are free to attend *any* precept of the week. You are not obligated to attend the precept that you enroll in. Your attendance for the precept should be marked within a week; if your attendance was not marked please contact the preceptor whose precept you attended that week. In case you cannot attend any precept of the week due to sickness please contact either the dean of your residential college or the director of studies so that they can contact the course instructor to confirm your sickness.

Problem Sets and Projects

Submitting Problem Sets and Projects

You must submit your work electronically on Blackboard. **Please do not email your work to any of the instructors in any case (except if you were locked out of Blackboard)**; this will only cause confusion and delay in grading. Please be sure to read the instructions for the submission procedure. You are also required to print out the PDF version of your solutions and turn it in in the first lecture following the assignment due date. If you forget to print out the PDF you can still drop it in the mailbox (after you enter the building turn to the left and you will see a file cabinet to the left with an open slot with the label "SML 201 Homework") located in the common area of 26 Prospect Avenue by the **same day** of the lecture; note that the building might be locked after 6pm.

Except for the first problem set there will be two files due for every problem set or project: an R Markdown file (e.g., project_1.Rmd) and its compilation into either a PDF file (e.g., project_1.pdf) or an HTML file (e.g., project_1.html). We will provide a template for you to use. As part of grading your problem sets and projects, we will be recompiling the R Markdown files, so you should make sure it compiles with no problems before submitting your project.

Tentative Problem Set Due Dates (Will Be Assigned 1 Week Before Due Date)

- PS1: Monday Oct. 2
- PS2: Monday Oct. 16
- PS3: Wednesday Nov. 8
- PS4: Wednesday Dec. 6

Project Due Dates

- Project 1: Monday Oct. 23
- Project 2: Monday Nov. 27
- Project 3: Monday Jan. 15

Late Submissions

Late problem sets will not be accepted. Late projects will be penalized by 10 points (out of 100 total points) per 24 hour interval past the project due date and time.

Working with others

You are allowed to use text books and resources online. You may not ask questions from individuals on the internet (e.g., you may not ask questions on Stack Exchange or the R help discussion groups). See below regarding citing your references.

Problem sets and projects may be completed in groups of up to 3 students. It is okay to work by yourself, if this is preferable.

For problem sets you are free to work with whoever you prefer and you can work with the same partner(s) multiple times.

For projects **you may not work with a given student on more than once.** In other words, if you work with Joe and Jane on Project 2, then you cannot work with Joe and Jane on any other projects. **You must form completely new groups for every project.** We expect that the work on any given project contains approximately equal contributions from all members of the group. Failing to make contributions and then putting your name on a project will be considered a violation of the honor code.

Citing your references

In accordance with the honor code, you must cite all sources of external information used in your work. This can be a book, a web site, or an individual. Part of being a successful data scientist is having the ability to leverage existing information and techniques, so it is okay to do so in this course unless stated otherwise (specifically, the in-class quizzes and problem set 1 as detailed above).

How to succeed in this course (some suggestions from students from last semester)

> It's a good idea to read over the notes right after the lecture to make sure that you understood what's covered that day.
>
> Do not be afraid of coding. Take it step by step. Run every 4 lines and knit every paragraph. This makes debugging easier!
>
> I would tell students to pay attention in lecture and precept as looking at lecture/precept notes isn't an alternative to attending lecture.

Schedule

Please see the course web site for the most recent schedule of topics and reading assignments.

Topics

1. Applications in Data Science and statistics
2. Fundamentals of R
3. Manipulating and tidying data
4. Reproducible analyses
5. Exploring and visualizing data
6. Random variables and probability
7. Statistical inference
8. Statistical tests
9. Linear regression and modeling
10. Prediction and validation
11. (If time allows) Bootstrap plus other more advanced techniques

Recording Lectures

Audio or video recording of the lectures is prohibited without permission from the instructor.