SML 201 | Introduction to Data Science

Spring 2019

Overview[1][2]

This course provides an introduction to the burgeoning field of data science. Data science 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. Facility with high school-level math is strongly recommended. The course does not assume any prior experience with programming.

Instructors

Course Instructor

Michael Guerzhoy

Email: guerzhoy@princeton.edu

Office: CSML 202 (26 Prospect Ave.)

Preceptors

David Ribar, dribar@princeton.edu

Mert Al, merta@princeton.edu

Ekaterina (Katya) Chegaeva, chegaeva@princeton.edu

Stephen Keeley, skeeley@princeton.edu

Sulin Liu, sulinl@princeton.edu

[1] The course instructor reserves the right to make changes to this syllabus; this version is current as of October 2018.

[2] This syllabus is based on syllabi by John Storey and Daisy Huang

Office Hours

Refer to the course website.

Getting Help

Piazza

Please sign up at https://piazza.com/princeton/spring2019/sml201/. Please ask questions on Piazza if they could relevant to other students.

Introductory level R programming workshops

The Departments of Politics and Sociology offer introductory-level R programming workshops this semester. You can learn basic programming skills by participating in their workshops. Students can see the schedule of the workshops on the website (https://compass-workshops.github.io/info/) and to check if any change takes place.

Evaluation

- Projects: 35%
- In-class quizzes: 5%
- Precept problem sets: 30%
- Term tests: 30%

Quizzes

Unannounced closed-notes in-class quizzes will be held during lecture. Quiz problems will be about the material covered in the two lectures prior to the quiz. You will earn a grade of zero for any quiz missed due to unexcused absence. You should email the instructor ahead of time in case of illness or other personal circumstances that prevent you from attending lecture. Quizzes missed due to excused absences will not count towards your final grade. Your grade will be computed using the best 75% of quizzes, after excluding quizzes from which you were excused.

Precept

Part of the grades awarded for precept problem sets will be awarded for making a reasonable effort towards completing the problem sets in precept.

Schedule

A detailed schedule will be posted on the course website.

Topics

1. Fundmentals of R
2. The `tidyverse` libraries for data wrangling in R
3. Data vizualization
4. Predictive modelling
5. Statistical inference and statistical tests
6. Inference using linear regression and logistic regression
7. Evaluating statistical models
8. (Time permitting) Introduction to machine learning
9. (Time permitting) Science-wide false discovery rates

References

The following textbooks are recommended. They are all available in electronic format for free from the authors.

- David M Diez, Christopher D Barr, and Mine Çetinkaya-Rundel. OpenIntro Statistics (3rd ed.) OpenIntro, Inc., 2015.
- Kieran Healy. Data Visualization: A Practical Introduction. Princeton University Press, 2018.
- Hadley Wickham and Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, 2017.

Academic Integrity and Collaboration Policy

The problem sets and projects are to be done by each student or team alone. Any discussion of the assignments with other students should be about general issues only, and should not involve giving or receiving written, typed, or emailed notes. You should never show your write-up or code to other students, and you should never look at other students' write-ups and code.

You may consult any textbook or internet resource regarding general issues. Any use of a resource (apart from the course notes) should be clearly acknowledged in your write-up. For example, if you got a piece of code from a website, it should be clear from your submission that you did not author that piece of code.