Overview

Welcome to **SML 201 – Introduction to Data Science**! 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.

Website & Forum

Website: `http://guerzhoy.princeton.edu/201s20/`

Forum: `https://piazza.com/princeton/Spring2020/sml201/`

All course handouts will be posted on the course website. *Students are responsible for reading all announcements on the course forum on Piazza.*

Instructor

Instructor	Email	Office	Office Hours
Michael Guerzhoy	`guerzhoy@princeton.edu`	CSML 202 26 Prospect Ave.	Mon. & Fri., 2:30p.m.-3:30p.m.

Grading

The grading scheme for the course is as follows.

	Worth
3 Projects	35%
2 Problem sets	60%
In-class quizzes	5%
In-precept assignments	22%
Two term tests	32%

Getting help

Support is provided by the instructor, the preceptors, and the Undergraduate Course Assistants (“”Lab TAs”).

1. **Piazza.** Please sign up at `https://piazza.com/princeton/spring2020/sml201/` . Please ask questions on Piazza if the could be relevant to other students.

2. **Office hours.** Office hours are offered by the instructor, as well as by the preceptors and the undergraduate course assistants. You may seek help from any of us at any time.

References

We will refer to the following books for domain knowledge and pedagogical insight on statistics.

- Russell Poldrack. **Statistical Thinking for the 21st Century**. Online textbook, 2020.
- Kieran Healy, **Data Visualization: A Practical Introduction**. Princeton University Press, 2018
- Garrett Grolemund and Hadley Wickham, **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. O’Reilly Media, 2016

Course outline

We will be aiming to discuss the following topics.

1. Fundamentals of programming using R.
2. The `tidyverse` library for data wrangling and visualization in R.
3. Principles of data visualization. Practical visualization using `ggplot`.
4. Statistical inference and statistical tests.
5. Inference using linear regression and logistic regression.
6. Evaluating statistical models.
7. (Time permitting) Introduction to machine learning.
8. (Time permitting) Science-wide false discovery rates

Academic Integerity

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.