Overview

Welcome to **SML 480 – Pedagogy of Data Science**! In this seminar course, we will explore the pedagogy of introductory data science. Data science is a burgeoning new field at the intersection of computer science and statistics. In recent years, *Introduction to Data Science* (I2DS) courses aimed at first and second year university students appeared in many universities. Unlike most introduction to programming courses, I2DS courses are often taught using the functional programming paradigm, influeced by the design of packages such the `tidyverse` in R and `pandas` in Python. Unlike virtually all introductory statistics course, I2DS courses use the computational skills students acquire to introduce simulation-based inference. Teaching programming to a broad audience using the functional paradigm in the context of data science tasks and teaching statistics with a focus on simulation-based inference will be the focus of the seminar.

During the seminar, we will explore the topics discussed in *I2DS* in more depth and read the literature in the pedagogy of computer science and statistics when discussing teaching techniques applicable to data science. Students will participate in the delivery of *SML 201 – Introduction to Data Science* as lab assistants.

Seminar students will complete the weekly precept exercises from *SML 201* before they are assigned to *SML 201* students, analyze them to obtain a list of learning goals, and discuss relevant pedagogical approaches during the seminar. In addition, students will complete homework assignments designed to expand their domain knowledge in the relevant areas of data science and functional programming. Seminar students will work closely with two or three students in *SML 201*, and write a final paper describing the progress of their students, drawing on their experience teaching the students and on their understanding of the pedagogical literature discussed in the seminar.

Students will gain teaching skills and pedagogical content knowledge as well as an in-depth understanding of the foundations of data science.

The academic work in *SML 480* will be done *in addition to* paid Undergraduate Course Assistant work in *SML 201*. Work done as a UCA for *SML 201* should not be regarded as part of the work in this course.

Website & Forum

Website: `http://guerzhoy.princeton.edu/480s20/`
Forum: `https://piazza.com/princeton/sml480/`

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	TBA

Grading

The grading scheme for the course is as follows.

	Worth	Due
Precept exercise analysis + domain knowledge exercises	50%	10 assignments throughout the term
Initial version of paper	10%	Midterm
Final version of paper	20%	Dean's date
Seminar participation	20%	

Course topics

Each week, we will discuss the topics that are coming up in *SML 201* and how to teach them; review teaching strategies from the previous week; and discuss the topics that are coming up in *SML 201* from a "higher-level" perspective.

1. Teaching programming via the functional programming paradigm (Weeks 1-3)
2. Interacting with students in the lab (Weeks 1-2)
3. Tidy data. `dplyr`: a programming languages perspective (Week 2)
4. Predictive modelling and cost functions: an elementary perspective. GLMs, probability models, and link functions, and their connection to predictive modelling via cost functions (Week 3)
5. The Grammar of Graphics and `ggplot`. Cross-validation (Week 4)
6. Teaching linear regression. Linear regression coefficients. Causal infrence (Week 5)
7. Teaching elementary probability. Probability: "elementary" approach vs. the axiomatic approach (Week 6)
8. P-values. Teaching p-values via simulation (Week 7)
9. Computing p-values via approximation. P-values: simulation-based approaches vs. approximation-based approaches (Week 8)
10. Hypothesis testing, and teaching hypothesis testing. Causal inference. (Week 9)
11. Hypothesis testing and confidence intervals. Teaching hypothesis testing and CIs (Week 10)
12. Teaching inference with linear regression: forming hypotheses and checking assumptions. (Week 11)
13. Closing discussion. Approaches to teaching. (Week 12)

References

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

- Andrew Gelman and Jennifer Hill. **Data Analysis using Regression and Multilevel/Hierarchical Models**. Cambridge University Press, 2006
- Cosma Rohilla Shalizi. **Advanced Data Analysis from an Elementary Point of View**. Cambridge University Press (forthcoming). Preprint available at `https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/`
- Ani Adhikari and John DeNero. **Computational and Inferential Thinking**. Online textbook. available at `https://www.inferentialthinking.com`
- David Williams. **Weighing the Odds: A Course in Probability and Statistics**. Cambridge University Press, 2001

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

- Hadley Wickham, **Advanced R**, 2nd ed. Chapman and Hall/CRC, 2019
- Matthias Felleisen, Robert Bruce Findler, Matthew Flatt, and Shriram Krishnamurthi. **How to Design Programs**, 2nd ed. The MIT Press, 2019

The following textbooks are required in *SML 201 – Introduction to Data Science*. We will refer to them as needed

- 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

Seminar format

Seminars will generally consist of three parts

- Discussion of the previous precept. What are students' difficulties? What strategies worked? What strategies didn't work?
- Discussion of the upcoming precept. What difficulties do we anticipate? How can we address them? Essential domain knowledge will be reinforced as needed.
- Discussion of upcoming content at a level higher than *SML 201*. For example, when students in *SML 201* learn the basics of `tidyverse`, the seminar will take a bird's-eye view of `tidyverse` in the context of functional programming for data science. When basic simulation-based inference is discussed in *SML 201*, there will be a more advanced and complete version presented at the seminar.

Relevant pedagogical research literature will be discussed as the need arises.

Final paper

Each student in the seminar will follow two or three *SML 201* students who entered *SML 201* with no prior background in programming or statistics. The seminar student will keep track of which learning goals their students are meeting, and will keep a record of their teaching approaches when teaching the student.

The final paper will include a progress report on the students and an analysis of the progress. The paper should link the seminar student's observations to at least two pieces of pedagogical research literature.

We expect, but cannot guarantee, that there will be a sufficient number of *SML 201* students who will consent for their progress in the course to be documented and analyzed as part of a final paper for *SML 480*. In the event that there is not a sufficient number of *SML 201* students who consent for their work to be documented and analyzed, students in *SML 480* will write a final paper that does not require analyzing student work.

Prerequisites

Students will be accepted to the seminar by interview. Demonstrated proficiency in the foundations of data science and in programming are required. Students whose performance in *SML 201* was excellent and who have not taken follow-up courses may be considered.

McGraw Center

UCAs in *SML 201* are required to participate in a training session run by the McGraw Center, and are encouraged to engage with the McGraw Center throughout the semester. Work in *SML 480* will be different from the offerings of the McGraw Center. .