Introduction to Data Science for Public Policy and Management

Jae Yeon Kim

Updated: 2021-04-29

Overview

What is data science, and how can you apply it to solve social problems? This course is for students with a serious interest in using data-driven approaches to tackle pressing social problems. The course will provide students with a mental framework to think through the critical decisions they need to make in each step of the data science workflow.

The course is divided into two main sections: fundamentals and applications. As for fundamentals, we will first learn what data science is and what it is not (spoiler: it is not all about big data!) and how to design a data science project. We will then learn the framework of computational social science: combining data science tools and techniques with social science theories and research methods.

As for applications, we will learn how to answer descriptive, inferential, and predictive questions using administrative, survey, experimental, and digital trace data. Each class session is built around discussions on data science applications to social problems, ranging from detecting poverty to reducing pollution.

Note that the course is not about learning programming. The emphasis is on navigating various ways of combining data, computing, and social science to solve real-world problems.

Objectives

By the end of the semester, students will be capable of critiquing and designing a data science project. They will learn what aspects of social problems they can or can't solve using data-driven approaches.

- Students will *describe* what data science is and what data science is not.
- Students will *understand* the basic data science workflow.
- Students will *understand* the computational social science framework.
- Students will *evaluate* various applications of computational social science research.
- Students will *design* a data science project applied to social problems.

Logistics

Contributors

- Instructor: Jae Yeon Kim

Time and location

Office hours

Office hours will be held ...

Slack & GitHub

- Slack for communication (announcements and questions). It would be best if you asked questions about class material and assignments through the Slack channels so that everyone can benefit from the discussion. We encourage you to respond to each other's questions as well.
- GitHub for course materials. All course materials will be posted on GitHub at ..., including lecture notes, code demonstrations, sample data, and assignments.

Accessibility

This class is committed to creating a safe and inclusive environment in which everyone can participate. If you have a particular concern (e.g., disability), please come to me as soon as possible (ideally within the first two weeks) so that I can make special arrangements.

It's Okay Not to Know

Asking questions is your privilege in an academic environment (we are all here to learn). There is no such thing as a stupid question.

Audting

No auditing is permitted.

Late policy

No late assignments are accepted.

Course requirements and grades

Students will form a team of 3-5 people. Note that all of the following activities are **team-based.** I will assign you to teams based on the responses to the survey, which I will circulate at the beginning of the semester. The assignment will be based on an algorithm that optimizes diversity among team members. You can change your team **once** before the submission of the extended research proposal.

This is a graded class based on the following:

- Participation (20%)
- Short research proposal (10%)
- Extended research proposal (20%)
- Pitch (10%)
- Final project (50%)

Class participation The class participation portion of the grade can be satisfied in one or more of the following ways:

- attending the lectures
- asking and answering questions in class
- contributing to class discussion through the Slack workspace

Research proposals and final projects Note that these are all team-based projects.

1. Short research proposal (due **week 5**): Teams will submit a two-paragraph description of their final project and why it matters. For more information, see this guideline.

2. Extended research proposal (due **week 7**): Teams will submit a two-page description of their project, why it matters, how it will be conducted and evaluated. For more information, see this guideline.
3. Lightning talk (elevator pitch) (due **week 8-10**): Teams present their projects in a maximum of 5-minute talk, with 5 minutes for class Q&A. When you prepare slides, think about spending one minute for each slide. For more information, see this guideline.
4. Final project (due **week 12**): Teams submit the final project using the framework we learned in class on your own problem of interest.

Computer requirements We do not learn to program in this course, but the course materials are based on code examples. Therefore, for replications, you are required to install the following programs:

- Access to the UNIX command line (e.g., a Mac laptop, a Bash wrapper on Windows)
- Git
- R and RStudio (latest versions)
- Anaconda and Python 3 (latest versions)
- Pandoc and LaTeX

All of the required software is **free** and should be installed in the first week of the class.

See install.md for more information.

Course schedule

- Week 1: Demythifying Data Science and Its Workflow
- Week 2: Breaking Down Social Problems Into Research Questions
- Week 3: Descriptive, Inferential, and Causal Questions
- Week 4: Best Practices in Creating and Managing Spreadsheets
- Week 5: Designing, Analyzing, and Optimizing Surveys and Experiments
- Week 6: Administrative Data: What It Can Tell and It Can't Tell
- Week 7: Digital Trace Data: Larger, Faster, but Better?
- Week 8: High-dimensional Data: The Current and Next Frontiers
- Week 9: Prediction Models: The Revolution Hasn't Happened Yet
- Week 10: What Should I Optimize? Data Science, Profits, and Ethics
- Week 11: Reading Week
- Week 12: Final Project Due

Questions, comments, or suggestions

Please create issues if you have questions, comments, or suggestions.