Lecture: Tuesdays 10am-noon (402 Social Science Building)
Lab: Thursdays 10am-noon (402 Social Science Building)
Instructor Office Hours:
  Harding (dharding@berkeley.edu): Tuesdays 1-3pm (sign-up: https://tinyurl.com/HardingOH)
  Sharma (prashant.sharma@berkeley.edu): TBD

Course bCourses Site: https://bcourses.berkeley.edu/courses/1521183/
Course Github repository: https://github.com/dlab-berkeley/Computational-Social-Science-Training-Program

Course description: This is the second semester of a two-semester course that provides a rigorous introduction to methods and tools in advanced data analytics for social science doctoral students. The goal of the course is to provide students with a strong foundation of knowledge of core methods, thereby preparing them to contribute to research teams, to conduct their own research, and to enroll in more advanced courses. The course will cover research reproducibility (fall), machine learning (fall), natural language processing (spring), and causal inference (spring). In contrast to other courses currently offered on campus, this course’s intended audience is applied researchers, typically social science doctoral students in their second or third year of graduate school. This is a required course for students in the Computational Social Science Training Program. Enrollment is open to doctoral students from any department. Students who have not taken SOCIO 273L should consult the instructor before enrolling.

The course is divided into modules, each lasting 3-5 weeks. Each module will include lectures, discussion of example research articles, lab exercises, and a group project involving Python or R programming. Projects, typically done in groups of 3-4 students, will also provide the opportunity to practice reproducibility techniques, data manipulation and transformation, and data science workflows.

Course objectives (Spring semester):
  ● Conceptual understanding of methods for extracting data from text using natural language processing
  ● Conceptual understanding of experimental design and the structural causal model framework
  ● Conceptual understanding of causal inference problems, solutions, and methods for longitudinal settings
  ● Ability to apply these concepts and execute relevant methodologies on social science data in Python and R and correctly interpret results
  ● Familiarity with key empirical papers that apply computational social science methods to research

Prerequisites: SOCIO 273L: Computational Social Science, Part A (or equivalent knowledge), and a year-long course in statistical methods for social science graduate students (or equivalent prior experience) will generally be sufficient statistical preparation. Students should have a background in multivariate regression (both linear and non-linear models), maximum likelihood estimation, and introductory causal inference (omitted variable bias, potential outcomes, average treatment effects, causal graphs). Students may consult the instructor about readings on these topics to ensure adequate preparation. In addition, this course will be taught in Python and R. Students without a background in introductory Python programming should take the D-Lab Python Fundamentals Workshop series, which is usually offered in the week before the fall semester begins. Those who need a Python refresher may wish to review the Jupyter Notebooks for D-Lab Python Fundamentals here: https://github.com/dlab-berkeley/python-fundamentals Students without a background in introductory R
programming should take the D-Lab R Fundamentals Workshop series, which is usually offered in the week before each semester begins. Those who need an R refresher may wish to review the materials for D-Lab R Fundamentals here: https://github.com/dlab-berkeley/R-Fundamentals

Instructional technology: Examples and student projects will occur in Python using Jupyter Notebooks and in R using R Studio. Students should install Anaconda before the first lab. Students should install R Studio before the 7th lab.

Instructional Resilience and the possibility of Remote Instruction: All “lecture” and lab meetings will be held in-person by we will need to be prepared to move to Zoom if conditions require (see bcourses site for links).

How we will use class time: During most weeks, we will have pre-recorded videos, typically in 10-15 minute segments. Students should view these lectures, take the self quizzes, and do the readings BEFORE each week’s lecture session. Class “lecture” meeting times will be used to review key points, answer questions and discuss the readings. Occasionally we may use half of the lecture section for lab. Each student will submit a one-page weekly reflection memo by 5pm the night before lecture (students may skip 5 weeks during the semester). Lab times will be used to work through Python or R exercises applying the week’s concepts, tools, and models to data. Group projects with rotating, randomly assigned group membership will provide students with opportunities to build a course community with fellow students. To accommodate students taking the course asynchronously in other timezones, lecture and lab will be recorded. Recordings will only be available to instructors and fellow students. By enrolling in the course, you are consenting to these recordings.

Grading:
- Lecture and Lab Participation: 25%
- Weekly reflection memos (graded credit/no credit): 20%
- Project #1: 15%
- Project #2: 15%
- Project #3: 15%
- Project #4: 10%

Course Schedule
(note #1: links to readings are on bcourses; some require UC-Berkeley CalNet login)
(note #2: once the semester begins, any changes will be reflected on the bcourses site)

Module 1: Natural Language Processing

Week 1 (Jan 17/19): Introduction to NLP
Readings:
- Justin Grimmer, Margaret E. Roberts, and Brandon m. Stewart. *Text as Data*. Preface, Chapters 1-6
- James A. Evans and Pedro Aceves. 2016. “Machine Translation: Mining Text for Social Theory,” *Annual Review of Sociology* (focus on various research questions that can be answered with NLP and how supervised and unsupervised ML are applied in NLP)

Week 2 (Jan 24/26): Exploratory/Unsupervised Methods, Part 1
Readings:
- Justin Grimmer, Margaret E. Roberts, and Brandon m. Stewart. *Text as Data*. Chapters 10-12
Week 3 (Jan 31/ Feb 2): Exploratory/Unsupervised Methods, Part 2
Readings:
- Justin Grimmer, Margaret E. Roberts, and Brandon m. Stewart. *Text as Data*. Chapters 13-14

Week 4 (Feb 7/9): Classification
Readings:
- Justin Grimmer, Margaret E. Roberts, and Brandon m. Stewart. *Text as Data*. Chapters 15-21

Week 5 (Feb 14/16): Vector Models
Readings:
- Justin Grimmer, Margaret E. Roberts, and Brandon m. Stewart. *Text as Data*. Chapters 7-9

Week 6: (Feb 21/23): Neural Nets for NLP Using GPUs and Google Collab
Readings:
- Francois Chollet, Deep Learning with Python, 2nd Edition. Chapters 3 and 11 (focus on Sections 11.3 and 11.4).

*** Project #1 (NLP) Due March 3 ***

Module 2: Introduction to Causal Inference

Week 7 (Feb 28/March 2): Introduction to Causal Inference (R Refresher in Lab)
Readings:
- Scott Cunningham. *Causal Inference: The Mixtape*, Chapters 1, 3, and Chapter 4 (sections 4.0-4.1).

Week 8 (March 7/9): Randomized Experiments
Readings:
- Gerber, Alan and Donald Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton. (Chapters 1, 4, 12) – available on bcourses
Week 9 (March 14/16): Matching Methods
Readings:

*** Project #2 (Matching) Due April 7 ***

**Module 3: Natural Experiments and Sensitivity Analysis**

Week 10 (March 21/23): Natural Experiments and Instrumental Variables
Readings:

No class March 28/30 -- Spring Break

Week 11 (April 4/6): Diff-in-Diff and Synthetic Controls
Readings:

Week 12 (April 11/13): Regression Discontinuity
Readings:

Week 13 (April 18/20): Sensitivity Analysis
Readings:

*** Project #4 (RD, Diff-in-Diff, Synthetic Controls) Due April 28 ***

**Module 4: Machine Learning for Causal Inference**
Week 14 (April 25/27): Machine Learning for Causal Inference with Longitudinal Data

Readings:


*** Project #5 (Longitudinal Data) Due May 12 ***