SOCIOL 273M: Computational Social Science, Part B
UC Berkeley
Spring 2021
Instructors: David Harding (Sociology) and Aniket Kesari (D-Lab)

Lecture: Tuesdays 10am-noon (online via Zoom)
Lab: Thursdays 10am-noon (online via Zoom)
Instructor Office Hours:
Harding (dharding@berkeley.edu): Tuesdays 1-3pm (sign-up: https://tinyurl.com/HardingOH)
Kesari (akesari@berkeley.edu): Thursdays 3-4 and Fridays 10-11, or by appointment (sign-up link)

Course bCourses Site: https://bcourses.berkeley.edu/courses/1500983
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 (csstp.berkeley.edu). Enrollment is open to doctoral students from any department. Students who have not taken SOCIOL 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 correctly interpret results
- Familiarity with key empirical papers that apply computational social science methods to research

Prerequisites: SOCIOL 273L: Computational Social Science, Part A (or equivalent knowledge). 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 6th lab.

Instructional Resilience and Remote Instruction: All "lecture" and lab meetings will be held via Zoom (see bcourses site for links). Lectures will be provided via pre-recorded videos, with timestamps for key sections and "self-quiz" questions to focus viewing of the videos. Students should view the lectures and do the readings BEFORE each week's lecture session. Class "lecture" meeting times will be used to answer questions about the lectures and discuss the readings. 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: 10%
- Project #3: 10%
- Project #4: 10%
- Project #5: 10%

Course Schedule

Module 1: Natural Language Processing

Week 1 (Jan 19/21): Introduction to NLP

Readings:
- Dan Jurafsky and James H. Martin. Speech and Language Processing, 2nd Edition. (Introduction)
- James A. Evans and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology*
- Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 2019, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

Week 2 (Jan 26/28): Exploratory/Unsupervised Methods, Part 1

Readings:
- Dan Jurafsky and James H. Martin. *Speech and Language Processing*, 3rd Edition. (selections)

Week 3 (Feb 2/4): Exploratory/Unsupervised Methods, Part 2

Readings:
- Dan Jurafsky and James H. Martin. *Speech and Language Processing*, 3rd Edition. (selections)

- Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. How to Analyze Political Attention with Minimal Assumptions and Costs. American Journal of Political Science Vol. 54, No. 1 (Jan., 2010), pp. 209-228.

Week 4 (Feb 9/11): Classification

Readings:

- Dan Jurafsky and James H. Martin. *Speech and Language Processing*, 3rd Edition. (selections)
- Andrew Peterson and Arthur Spirling. Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. Political Analysis (2018) vol. 26:120–128
- OPTIONAL:
    - David E. Pozen, Eric L. Talley, and Julian Nyarko. A Computational Analysis of Constitutional Polarization. Cornell Law Review, Vol. 105, pp. 1-84, 2019 (especially sections 3-4)
    - Matthew Gentzkow, Jesse M. Shapiro, Matt Taddy. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. ECONOMETRICS: Jul 2019, Volume 87, Issue 4

Week 5 (Feb 16/18): Vector Models

Readings:

- Dan Jurafsky and James H. Martin. *Speech and Language Processing*, 3rd Edition. (selections)
- Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review*

**** Project #1 (NLP) Due Feb 26 ****

Module 2: Introduction to Causal Inference

Week 6 (Feb 23/25): Introduction to Causal Inference (R Refresher in Lab)

Readings:

- Elwert, Felix. 2013. "Graphical Causal Models." Pp. 245–73 in *Handbook of Causal Analysis for Social Research*, S. Morgan (ed.). Springer.
- Imbens, G. W. and Rubin, D. B (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press (Chapters 1, 3[skim])

Week 7 (March 2/4): 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
- Raj Chetty, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review*

Week 8 (March 9/11): Matching Methods

Readings:

- Imbens, G. W. and Rubin, D. B (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press (Sections 12.1, 12.3-12.7, 13.1, 13.5, 13.7, 14.1-14.3, 15.1, 15.3, Chapter 18)

- Jennie Brand and Yu Xie. 2010. "Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review.*

**** Project #2 (Matching) Due March 19 ****

Module 3: Natural Experiments and Sensitivity Analysis

Week 9 (March 16/18): Natural Experiments and Instrumental Variables

Readings:

- Imbens, G. W. and Rubin, D. B (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press (Chapters 23-25 [subsections to come])
- Guy Grossman, Oren Gazal-Ayal, Samuel D. Pimentel, Jeremy M. Weinstein.2015. "Descriptive Representation and Judicial Outcomes in Multiethnic Societies." *American Journal of Political Science.* (replication data)

No class March 23/25 -- Spring Break

Week 10 (March 30/April 1): Diff-in-Diff and Synthetic Controls

Readings:

- Alene Kennedy-Hendricks, Matthew Richey, Emma E. McGinty, Elizabeth A. Stuart, Colleen L. Barry, and Daniel W. Webster, 2016. "Opioid Overdose Deaths and Florida's Crackdown on Pill Mills." *American Journal of Public Health* 106, 291-297, https://doi.org/10.2105/AJPH.2015.302953
- Stuart, Elizabeth A., Haiden A. Huskamp, Kenneth Duckworth, Jeffrey Simmons, Zirui Song, Michael E. Chernew, and Colleen L. Barry. 2014. "Using Propensity Scores in Difference-in-Differences Models to Estimate the Effects of a Policy Change." *Health Services and Outcomes Research Methodology* 14 (4): 166–82. doi:10.1007/s10742-014-0123-z.
- Alberto Abadie, Alexis Diamond & Jens Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association.* https://doi.org/10.1198/jasa.2009.ap08746

**** Project #3 (Natural Experiments) Due April 9 ****

Week 11 (April 6/8): Regression Discontinuity

Readings:

- Russell C. Callaghan, Marcos Sanches. Jodi M. Gatley, and James K. Cunningham. 2013. Effects of the Minimum Legal Drinking Age on Alcohol-Related Health Service Use in Hospital Settings in Ontario: A Regression–Discontinuity Approach. *American Journal of Public Health*
- Kalena E. Cortes, Joshua S. Goodman and Takako Nomi. 2015. Intensive Math Instruction and Educational Attainment Long-Run Impacts of Double-Dose Algebra. *Journal of Human Resources*

Week 12 (April 13/15): Sensitivity Analysis

Readings:

- Imbens, G. W. and Rubin, D. B (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press (Chapters 21 and 22)

- Miratrix, L., Furey, J., Feller, A., Grindal, T., and L. Page (2017). "Bounding, an accessible method for estimating principal causal effects, examined and explained," *Journal of Research on Educational Effectiveness*, 11(1): 133–162

**** Project #4 (RD, Diff-in-Diff, Synthetic Controls) Due April 23 ****

Module 4: Longitudinal Data and Time-Dependent Confounding

Week 13 (April 20/22): Longitudinal Data and Time-Dependent Confounding (Part A)

Readings:

- Wodtke, Geoffrey T., David J. Harding, and Felix Elwert. 2011. "Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentrated Disadvantage on High School Graduation." American Sociological Review 76:713-736. [Link to PDF on research gate]
- Sharkey, Patrick and Felix Elwert. 2011. "The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability." American Journal of Sociology 116: 1934-1981. [link to text in PUBMED]

Week 14 (April 27/29): Longitudinal Data and Time-Dependent Confounding (Part B)

Readings:

- Tran L, Yiannoutsos CT, Musick BS, et al. Evaluating the Impact of a HIV Low-Risk Express Care Task-Shifting Program: A Case Study of the Targeted Learning Roadmap. *Epidemiol Methods*. 2016;5(1):69-91. doi:10.1515/em-2016-0004 link to pdf in Pubmed

**** Project #5 (Longitudinal Data) Due May 14 ****