New Models and Methods in the Social Sciences
SOCIOLOGY 384
Summer Quarter, 2017

Dates: July 24 – August 4 (1:30pm-2:50pm)
Instructors: Vinodkumar Prabhakaran, Emily Putnam-Hornstein, Jeff Lewis, Jeremy Freese, Tomás Jiménez, Bill Behrman, Jon Krosnick, Matt Jackson, Sharad Goel, Cristobal Young
Location: Wallenberg Hall, Bldg. 160, 450 Serra Mall, Rm. 318

Convener: David B. Grusky
Office hours: Thursdays (please email Danielle Choi, dechoi@stanford.edu, to make an appointment)

Teaching assistants: Josh Gagne (joshpgagne@gmail.com) & Gabriel Chiu (gzzhao@stanford.edu)
Gabriel Chiu's office hours: Immediately after class (i.e., 3pm-4:30pm), Monday-Friday, Bldg. 120, Office 029)
Josh Gagne's office hours: 4pm-5:30pm, Monday-Friday, CERAS 406
NOTE: On Wednesdays, Gabriel and Josh will switch office hours, with Gabriel meeting in his office at 4pm, and Josh meeting at his office at 3pm.

Throughout the social sciences, there has been an explosion of new research methods and statistics, with the result that standard graduate course sequences often cover a declining proportion of the methods and statistics that actually appear in leading journals. Because of this development, many graduate students are not as well prepared as they once were to read the journals, to serve as research assistants, and to develop methodologically informed research programs of their own.

The purpose of Sociology 384 is to address this explosion of new methods and statistics by building a modular course featuring leading experts. The resulting course, which is offered as a concentrated two-week summer program, is designed for students who have already had initial exposure to the general linear model. It is taught in the form of ten modules that introduce the new approaches. For the 2017 course, the following topics will be offered: Natural language processing, predictive analytics, web scraping methods, contemporary methods for ensuring reproducibility, new developments in qualitative methods, state-of-the-art graphics and visualization, new survey methods, contemporary methods for analyzing networks, machine learning and related methods in computational social science, and contemporary methods for analyzing administrative and big data.

It is not of course possible in the context of a modular course of this sort to become fully proficient in any of the methods that are introduced. Rather, the objective is to familiarize students with the range of methods and statistics on offer, thereby allowing them (a) to read the journals with enough facility to understand research that uses these methods and statistics, (b) to make informed decisions about how their own research agendas might best be designed

and pursued, and (c) to know enough about particular methods and statistics to decide whether a more intensive course would serve them well.

Assignments: The purpose of this course is to encourage students to think about how these new methods and statistics might be harnessed for the purposes of their own research. The only assignment, therefore, will be to submit an essay for each module that lays out how the methods introduced in that module might be used to address an important research problem. This essay, which should be no longer than one page (single spaced), should discuss (a) the research hypothesis that will be addressed; (b) the data that might be used; (c) how the data will be analyzed (using the technique at hand); and (d) how the proposed research will advance the literature and thus constitute a contribution. The research proposals should not be fantasies but rather projects that are viable given the real constraints of money, time, and data availability under which graduate students are typically operating. The objective, in other words, is to assist in the development of proposals that might actually be carried out (for research papers or dissertations). Because students taking this intensive course will be working under extraordinary time constraints, it cannot of course be expected that proposals will be developed with a full appreciation of the relevant research literatures, but nonetheless some initial effort should be made to acquaint oneself with those literatures.

Due date: The assignments will be due by the start of the next class (with the last assignment due on Monday, 1:30pm, August 7). Please submit the assignments via Canvas. In recognition of the intensity of this course, all students will be allowed to omit any two of the ten research proposals. There is no need to indicate in advance which two you will be omitting. It is fine to simply turn in any 8 of the 10 assignments.

Extensions: If you miss a deadline (as defined above) for any of the assignments, it automatically counts as one of your two allowed omissions. There are no additional extensions available.

Grading: Each assignment will be graded on a scale ranging from 0 to 100. The course grade is based on the average across the total of 8 assignments. Students may choose to take the course for a letter grade or for credit/no credit. The grading scheme, for those opting for a letter grade, is as follows:

97-100 points: A+
93-96 points: A
90-92 points: A-
87-89 points: B+
83-86 points: B
80-82 points: B-
77-79 points: C+
73-76 points: C
70-72 points: C-
67-69 points: D+
63-66 points: D
Under 63 points: F

For those opting for the credit/no credit option, credit is given whenever the total score exceeds 63 points.

Readings: We have sought to keep required readings to a manageable minimum. We sometimes provide supplementary readings that may be consulted by students who wish to learn more. All readings are available on Canvas.

Units: We offer a range of units (3-5) to accommodate graduate students with different needs. The formal requirements don't vary with the number of units, but those taking more units should commit to writing more thoughtful and considered research plans.

Survey: Please take this one-minute survey before beginning the class: https://goo.gl/KJyZ4t. The survey will be used by Bill Behrman to make sure he's pitching the class at the right level. Thanks!

Schedule of sessions

Module 1 (Monday, July 24)

Topic: Natural Language Processing
Instructor: Vinodkumar Prabhakaran

Required readings

Prabhakaran, Vinodkumar; Reid, Emily; Rambow, Owen. "Gender and Power: How Gender and Gender Environment Affect Manifestations of Power." Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP). October, 2014. Doha, Qatar. http://www.anthology.aclweb.org/D/D14/D14-1211.pdf

Rob Voigt, Nick Camp, Vinodkumar Prabhakaran, Will Hamilton, Rebecca Hetey, Camilla Griffiths, David Jurgens, Dan Jurafsky, and Jennifer Eberhardt. 2017. "Body Camera Footage Captures Racial Disparities in Officer Respect." Proceedings of the National Academy of Sciences. http://www.pnas.org/content/114/25/6521.abstract

Module 2 (Tuesday, July 25)

Topic: Predictive Analytics
Instructor: Emily Putnam-Hornstein

Required readings (available on Canvas)

Shmueli, G. 2010. To Explain or to Predict? *Statistical Science, 25*(3), 289-310.

Panattoni L.E., R. Vaithianathan, T. Ashton, & G.H. Lewis. 2011. Predictive Risk Modeling in Health: Options for New Zealand and Australia. *Australia Health Review, 35*, 45-51.

Berk R., H. Heidari, S. Jabbari, M. Kearns, & A. Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. *University of Pennsylvania*.

Billings J., J. Dixon, T. Mijanovich, & D. Wennberg. 2006. Case Finding for Patients at Risk of Readmission to Hospital: Development of Algorithm to Identify High Risk Patients. *BMJ,* 333(7563), 327.

Suggested materials

Lewis G., N. Curry, & M. Bardsley. 2011. Choosing a Predictive Risk Model: A Guide for Commissioners in England. Nuffield Trust. https://www.nuffieldtrust.org.uk/files/2017-01/choosing-predictive-risk-model-guide-for-commissioners-web-final.pdf

Module 3 (Wednesday, July 26)

Topic: Web Scraping
Instructor: Jeffrey Lewis

Required software installations

Please bring a laptop to class with the following software installed:
RStudio (NOTE: If you prefer, you can use regular R.)
Chrome web browser

Please install the following plugins for Chrome:
selectorgadget
XPath Helper

Please install the following R packages:
ggplot2
rvest
Rselenium

Tutorials

If you are unfamiliar with the idea of webscraping or with web basics like HTML, CSS, or javascript, you should school yourself on them in advance, perhaps with this course.

Because all scraping will be done in R using the rvest package, you should also study the following on rvest:
http://cpsievert.github.io/slides/web-scraping
http://www.r-bloggers.com/rvest-easy-web-scraping-with-r/
http://www.r-bloggers.com/migrating-table-oriented-web-scraping-code-to-rvest-wxpath-css-selector-example
http://www.computerworld.com/article/2909560/business-intelligence/web-scraping-with-r-and-rvest-includes-video-code.html
http://codeforsacramento.org/blog/tutorial/2015/01/31/webscraping-with-r.html

It would also be helpful to familiarize yourself with CSS selectors and Xpath (which will be used to identify information that we would like to extract from a webpage):
http://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html
http://www.liquid-technologies.com/xpath-tutorial.aspx

Finally, if you have never encountered regular expressions (which are used to extract information from text), please examine this tutorial.

Module 4 (Thursday, July 27)

Topic: Contemporary Methods for Ensuring Reproducibility
Instructor: Jeremy Freese

Required reading

Gentzkow, Matthew, and Jesse M. Shapiro. "Code and Data for the Social Sciences: A Practitioner's Guide. https://www.brown.edu/Research/Shapiro/pdfs/CodeAndData.pdf

Module 5 (Friday, July 28)

Topic: New Developments in Qualitative Methods
Instructor: Tomás Jimenéz

Required readings (available on Canvas)

Small, Mario Luis. 2011. "How to Conduct a Mixed Methods Study: Recent Trends in a Rapidly Growing Literature." *Annual Review of Sociology* 37:57-86.

Suggested materials

Linton, April and Tomás R. Jiménez. 2009. "Contexts for Bilingualism among U.S.-Born Latinos." *Ethnic and Racial Studies* 32:967-95.

Mahoney, James and Gary Goertz. 2006. "A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research." *Political Analysis* 14:227-49.

Small, Mario L. 2009. "'How Many Cases do I Need?' on Science and the Logic of Case Selection in Field-Based Research." *Ethnography* 10:5-38.

Jerolmack, Colin and Shamus Khan. 2014. "Talk is Cheap: Ethnography and the Attitudinal Fallacy." *Sociological Methods & Research*:1-32.

Small, Mario L. 2009. *Unanticipated Gains: Origins of Network Inequality in Everyday Life.* New York: Oxford University Press.

Module 6 (Monday, July 31)

Topic: Data Manipulation and Visualization Using the tidyverse
Instructor: Bill Behrman

This is be a hands-on workshop in which students work on data manipulation and visualization projects. Before coming to the workshop, be sure to (a) install the necessary software, (b) complete the required readings below, and (c) download the three projects on which you will work during the class.

Required readings

Wickham H, Grolemund G. R for Data Science. O'Reilly, 2017.

- Chapter 3, Data visualization.
- Chapter 5, Data transformation.

Suggested materials

Wickham H. The tidyverse style guide.

Module 7 (Tuesday, Aug. 1)

Topic: New Developments in Survey Methods
Instructor: Jon Krosnick

Required reading (available on Canvas)

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Touranguea. 2004. *Survey Methodology*, Wiley-Interscience. Chs. 1-6.

Suggested materials

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Touranguea. 2004. *Survey Methodology*, Wiley-Interscience. All other chapters.

Module 8 (Wednesday, Aug. 2)

Topic: Contemporary Methods for Analyzing Networks
Instructor: Matt Jackson

Required reading

Banerjee, Abhijit, Arun Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2013. ``The Diffusion of Microfinance,'' *Science* Vol. 341 no. 6144, DOI: 10.1126/science.1236498, 26 July 2013. Also see Supplementary Materials and related data.

Suggested material

Jackson, Matthew O. 2014. ``Networks in the Understanding of Economic Behaviors.'' *Journal of Economic Perspectives,* Vol. 28, No. 4, pp. 3-22, doi: 10.1257/jep.28.4.3, 2014.

Module 9 (Thursday, Aug. 3)

Topic: Machine Learning and Related Methods in Computational Social Science
Instructor: Sharad Goel

Required readings

Dean, Jeffrey, and Sanjay Ghemawat. 2004. "MapReduce: Simplified Data Processing on Large Clusters" (available on Canvas)

Bottou, Léon. "Large-Scale Machine Learning with Stochastic Gradient Descent." http://leon.bottou.org/publications/pdf/compstat-2010.pdf

Suggested materials

Goel, Sharad, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2016. "The Structural Virality of Online Diffusion." *Management Science* 62 (January), pp. 180-96. https://5harad.com/papers/twiral.pdf

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy." *The Annals of Applied Statistics* 10, pp. 365-94. https://5harad.com/papers/frisky.pdf

Module 10 (Friday, Aug. 4)

Topic: Big Data and Administrative Data
Instructor: Cristobal Young

Assigned reading

Varian, Hal. 2014. "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, Volume 28, Number 2, pp. 3–28. http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.3

Suggested materials

Boyd, Danah, and Kate Crawford. 2012. "Critical Questions for Big Data," *Information, Communication, and Society* 15:5, pp. 662-679. http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878

Youyou, W., M. Kosinski, D. Stillwell. 2015. “Computer-based Personality Judgments are More Accurate than those Made by Humans.” *Proceedings Of The National Academy Of Sciences*. http://www.pnas.org/content/112/4/1036.full.pdf

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. “Detecting Influenza Epidemics using Search Engine Query Data.” *Nature*. http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html?foxtrotcallback=true

Butler, Declan. 2013. “When Google Got Flu Wrong.” *Nature*. http://www.nature.com/news/when-google-got-flu-wrong-1.12413

Honor Code

Go here for statement of the Honor Code and Fundamental Standard.

Students with Documented Disabilities

Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request with required documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty dated in the current quarter in which the request is being made. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations. The OAE is located at 563 Salvatierra Walk (phone: 723-1066, URL: http://studentaffairs.stanford.edu/oae).