**Text as Data**
**Poli Sci, Spring 2019**
Tuesday, Thursday 1200pm- 120pm
200-034

**Instructor**: Justin Grimmer, Political Science Department

Office: Encina West 416

Contact: jgrimmer@stanford.edu, 617-710-6803. Gchat; justin.grimmer@gmail.com

Office Hours: My door is almost always open during normal business hours. Please email me to setup an appointment if you need to meet at a specific time.

**TA**: Masha Krupenkin

Office Hours:

Contact:

Political campaigns, government, businesses, and social scientists increasingly use large data sets and machine learning techniques. Sometimes they are applied to accomplish the same goals the organizations have pursued for decades. For example, political campaigns use consumer data, prior voting history, and models to classify voters as likely supporters. Governments use data on demographic change, employment, and tax revenues along with predictive methods to forecast the stability of social safety net programs, like social security. Machine learning methods and massive datasets have also created new opportunities and applications. For example, businesses use information about prior purchasing behavior, website searches, and experiments along with methods to estimate heterogeneous treatment effects to target advertisements to maximize their effectiveness. And across the social sciences machine learning methods are used throughout the research process to discover new hypotheses, measure quantities of interest, infer causal effects, and predict outcomes.

Machine learning methods have proven especially effective when applied to text. Text as data is a rapidly growing area within the social sciences, in part because of the importance of documents to social interactions. Language is the medium for politics and political conflict. Candidates debate during elections. Representatives write laws. Nations negotiate peace treaties. Clerics issue Fatwas. Citizens express their opinions about politics on social media sites. These examples, and many others, suggest that to understand what politics is about, we need to know what political actors are saying and writing.

This course introduces machine learning techniques by introducing methods to collect, analyze, and utilize large collections of text for social science inferences. The ultimate goal of the course is to introduce students to machine learning methods and modern quantitative text analysis techniques, while providing the skills necessary to apply the methods in their

own research. In achieving this ultimate goal, students will also learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems. They will also have the opportunity to develop their programming abilities and develop an original research project.

# Prerequisites

At a minimum, students should have completed coursework on univariate inference and linear regression. The ideal student will have taken 450A-D. The course will develop student's programming skills. Prior experience with `R`, `Python`, or a related language is strongly recommended. If you have any questions about whether you're ready for the course, please speak with me.

# Evaluation

Students will be evaluated across three areas.

**Homework** Students will be asked to complete five homework assignments. The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for their work. Portions of the homework completed in R can be submitted using R markdown, a markup language for producing well-formated HTML documents with embedded R code and outputs. R markdown requires installation of the knitr package. We recommend using Rstudio, an IDE for R, which is set up well for the creation of R markdown documents. Python assignments (and R assignments if you choose), including code and output, should be submitted in LaTeX or similar document preparation language. We recommend downloading and installing the Anaconda python distribution and working in Jupyter notebooks for Python code.

More about `RStudio` can be found here:
http://www.rstudio.com/

`R Markdown` can be found here:
http://rmarkdown.rstudio.com/

`Jupyter` can be found here:
https://www.anaconda.com/download/

**Final Project** Students will have the opportunity to complete a final project. For the final project students will complete an original research project and (in the best case scenario) the project will contribute to completing their dissertation, field paper, or ongoing research.

Political science is an increasingly collaborative discipline. So, students will be allowed (and encouraged) to complete the final project as a two or three-person team. Students will present their final project during a class wide poster session on the final class meeting, where

all faculty and graduate students will be invited to attend. Poster sessions provide the opportunity to receive a lot of feedback from many people and (I think) are the best way to present research to receive actual feedback. After the poster session students will submit a paper. Specifics about the paper will be discussed in class.

To complete the assignment we will have several intermediate steps. Students are encouraged to regularly discuss their project with me, but they must meet the following deadlines to receive full credit for their research project.

- Initial project selection/question: April 16th.
- Data set collected, ready to analyze: May 7th
- Initial analyses/Write Up: May 16th
- Final Meeting with me to discuss project: May 28th
- Poster Session: June 4th

**Participation** Students are expected to attend each class and to ask questions regularly. This is a challenging course and I will only proceed when everyone understands the content. I can only know if you do not understand if you slow me down with questions.

# Books

There are no required books for the class. I will provide working versions of chapters from a book project I am completing with Molly Roberts and Brandon Stewart (GRS hereafter). This book is a work in progress, which means that many of the chapters will need your help. I sincerely appreciate all feedback you might offer (and pointed criticism is always helpful, but any comments are always welcome).

I have several other texts you might consider exploring further for the class.

Machine Learning

- Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Hastie, Tibshirani, and Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edition. Springer.
- McLachlan and Peel. 2000 *Finite Mixture Models* Wiley.
- McLachlan and Krishnan. 2008. *The EM Algorithm and Extensions*. 2nd Edition Wiley.

Natural Language Processing

- Manning, Raghavan, and Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
  Available at `http://nlp.stanford.edu/IR-book/information-retrieval-book.html` (hereafter MRS)
- Jurafsky, Daniel and James Martin. 2008. *Speech and Language Processing*. Prentice Hall.

# Class Outline

**4/2: Machine Learning and the Social Sciences: Discovery, Measurement, and Causal Inference**

- GRS: Chapter 2
- Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents" *Political Analysis*. 21, 3 267-297.
- Lucas, Christopher, Richard Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2014. Computer assisted text analysis for comparative politics. *Political Analysis*
- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356, no. 6334 (2017): 183-186.

**4/4: Feature Engineering: Acquiring and Manipulating Text Data**

- GRS: Chapter 3
- Denny, Matthew and Arthur Spirling. 2018. "Text Processing for Unsupervised Learning: Why It Matters, Why It Misleads, and What to Do About It" `https://www.nyu.edu/projects/spirling/documents/preprocessing.pdf`
- Schofield, Alexandra and David Mimno. "Comparing Apples to Apple: The Effects of Stemmers on Topic Models." Transactions of the Association for Computational Linguistics 4 (2016): 287-300.

## Discovery

**4/9: Regular Expressions and Vector Space Model of Text**

- Ban, Pamela; Alexander Fouirnaies; Andrew B. Hall and James Snyder. 2017. "How Newspapers Reveal Political Power". *Political Science Research and Methods.*
- Turney, Peter D; Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics" *Journal of Artificial Intelligence Research*. 37, 141-188.

**4/11: Counts, Proportions, and Distributions (Getting to Know the Dirichlet Distribution and Other Distributions on the Simplex)**

- Chapter 2 Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning* (Sections 2.1, 2.2 especially)
- Katz, Jonathan and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data" *American Political Science Review* 93, 1, 15-32.
- GRS Chapter 4.

**4/16: Clustering Methods (Fully Automated and Computer Assisted)**

- 14.3. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [Coursework]
- Chp 9. Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning* (Sections 2.1, 2.2 especially) [coursework]
- Grimmer, Justin and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization" *Proceedings of the National Academy of Sciences* 108(7), 2643-2650
- GRS Chapter 4.

**4/18: Guest Lecture Michelle Torres Assistant Professor, Rice University**

**4/23: Topic Models: Vanilla LDA for Discovery**

- Blei, David, Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation" *Journal of Machine Learning*
- Blei, David. 2012. "Probabilistic Topic Models". *Communications of the ACM*. 55, 4, 77-84
- Wallach, Hanna, David Mimno, and Andrew McCallum. "Rethinking LDA: Why Priors Matter". *Proceedings of the 23rd Annual Conference on Neural Information Processing*
- Nelson, Laura. 2017. "Computational Grounded Theory". *Sociological Methods & Research*
- GRS Chapter 4.

**4/25: Principal Components, Multi-dimensional Scaling, and Word Embeddings**

- 14.8. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [Coursework]
- Spirling, Arthur. US Treaty-Making with American Indians: Institutional Change and Relative Power 1784-1911 *American Journal of Political Science* 56, 1, 84-97.
- GRS Chapter 4.

**4/30: Fictitious Prediction Problems: Finding Discriminating Words**

- Mosteller, Frederick and David Wallace. 1963. "Inference in an Authorship Problem" *Journal of the American Statistical Association* 58, 302. 275-309

- Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". *Political Analysis* 16(4)

- Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis" *Journal of the American Statistical Association* 108, 755-770

- Grimmer, Justin. "Evaluating Model Performance in Fictitious Prediction Problems" Discussion of "Multinomial Inverse Regression for Text Analysis" by Matthew Taddy. Journal of the American Statistical Association, 2013.108 (503) 770-771

- GRS Chapter 4.

## Measurement

**5/2: Hand Coding, Codebook Creation, and Classification into Existing Categories**

- Grimmer, Justin, Gary King, and Chiara Superti. 2019. "The Unreliability of Intercoder Reliability and What to do About It". *Stanford University Mimeo*

- Dawid, Alexander Philip, and Allan M. Skene. "Maximum likelihood estimation of observer error-rates using the EM algorithm." Applied statistics (1979): 20-28. APA

**5/7: Dictionary Methods: Classification without Optimization**

- Soroka, Stuart and Lori Young. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts" *Political Communication* 29: 205-231

- Dodds, Peter and Christopher Danforth. 2009. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents". *Journal of Happiness Studies* 11, 4. 441-456

- Loughran, Tim and Bill McDonald.2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks" *Journal of Finance* 66, February 35-65

**5/9: Classification Methods Part 1: Ridge, LASSO, and Elastic Net**

- 3.4. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer.

- Bloniarz, Adam; Hanzhong Liu; Cun-Hui Zhang; Jasjeet Sekhon; Bin Yu. 2016. "Lasso adjustments of treatment effect estimates in randomized experiments" *PNAS*

**5/14: Classification Methods Part 2: Naive Bayes and ReadMe**

- Hopkins, Dan and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science" *American Journal of Political Science*, 54, 1

- King, Gary; Yin Lu. 2008. "Verbal Autopsy Methods with Multiple Causes of Death". *Statistical Science*. 23, 1. 78-91

- Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech". *Journal of Information, Technology, and Politics*. 5(1).

- King, Gary; Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression" *American Political Science Review*

**5/16: Classification Methods Part 3: Boosting, Bagging, and Ensembles via the Random Forest**

- Hillard, Dustin, Stephen Purpura and John Wilkerson. 2007. "Computer Assisted Classification for Mixed Methods Social Science Research". *Journal of Information, Technology, and Politics*.

- 7.10. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [Coursework]

- Chapter 3. Grimmer, Justin, Sean Westwood, and Solomon Messing. 2014. "The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability" *Princeton University Press*

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

- Jurka, Timothy; Loren Collingwood; Amber Boydstun; Emiliano Grossman, and Wouter van Atteveldt "RTextTools: A Supervised Learning Package for Text Classification" *The R Journal*.

**5/21: Structural Topic Models for Measurement**

- Quinn, Kevin et al. 2010 "How to Analyze Political Attention with Minimal Assumptions and Costs". *American Journal of Political Science*, 54, 1 209-228.

- Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". *Political Analysis*, 18(1), 1-35.

- Chp 5. Wallach, Hanna "Structural Topic Models for Language" `http://people.cs.umass.edu/~wallach/theses/wallach_phd_thesis.pdf`

- Roberts, Margaret E, Brandon Stewart, Dustin Tingley, Chris Lucas, Jetson Leder-Luis, Bethany Albertson, Shana Gadarian, and David Rand. 2014. Topic models for open-ended survey responses with applications to experiments. *American Journal of Political Science*

- Roberts Margaret E, Stewart Brandon M, Airoldi Edo M. A model of text for experimentation in the social sciences. 2016. *Journal of the American Statistical Association*

## Causal Inference

**5/23: No Class, Work on Posters**

**5/28: Causal Inference Refresher**

- Holland, Paul W. 1986. "Statistics and Causal Inference" *Journal of the American Statistical Association*

- Egami, Naoki; Christian Fong; Justin Grimmer; Margaret E. Roberts; and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts". *Princeton University Mimeo*

- Titiunik, Rocio. 2015. " Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science and Politics*

**5/30: Text as Intervention, Response, and Covariate**

- Fong, Christian and Justin Grimmer. 2016. "Discovery of Treatments from Text Corpora" *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany

- Fong, Christian and Justin Grimmer. 2019. "Causal Inference with Latent Treatments". *Stanford University Mimeo.*

- Egami, Naoki; Christian Fong; Justin Grimmer; Margaret E. Roberts; and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts". *Princeton University Mimeo*

- Guo, Fangjian et al. 2015. "The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation" *AISTATS*

- Roberts, Margaret E, Brandon Stewart, Dustin Tingley, Chris Lucas, Jetson Leder-Luis, Bethany Albertson, Shana Gadarian, and David Rand. 2014. Topic models for open-ended survey responses with applications to experiments. *American Journal of Political Science*

- Roberts Margaret E, Stewart Brandon M, Airoldi Edo M. A model of text for experimentation in the social sciences. 2016. *Journal of the American Statistical Association*

- Roberts, Margaret E, Brandon M. Stewart and Richard Nielsen. Matching Methods for High-Dimensional Data with Applications to Text. *University of California, San Diego. Mimeo*

**6/4: Poster Session**