17.835: Machine Learning and Data Science in Politics

Fall 2023

Instructor: Professor In Song Kim
TAs: Raymond Wang, Serene Ho

Department of Political Science
MIT

1 Contact Information

	In Song Kim	Raymond Wang	Serene Ho
Email:	insong@mit.edu	raywyf@mit.edu	sereneho@mit.edu
Office Hours:	Fri 4-5pm (By app)	Wed 4-5pm (66-154)	Wed 11-12pm (66-156)

2 Logistics

- Lectures: Tuesdays and Thursdays 3:00pm –4:30pm, 3-133
- Recitations: Mondays 11-12pm (66-156); 4-5pm (66-154).
  - There will be weekly recitation sessions. We will cover a review of the theoretical material and also provide help with computing issues. The teaching assistants will run the sessions and can give more details. Attendance is strongly encouraged.

Note that the first class meets on September 7. No class will be held on October 10, November 23, 24. Last day of class is December 12.

3 Course Description

Empirical studies in political science is entering a new era of "Big Data" where a diverse range of data sources have become available to researchers. Examples include network data from political campaigns, data from social media generated by individuals, campaign contribution and lobbying expenditure made by firms and individuals, and massive amount of international trade flows data. How can we take advantage of these new data sources and improve our understanding of politics? This course introduces various machine learning methods and their applications in political science research. Students will:

1. Be introduced to various quantitative political science research topics in its four subfields: American Politics, International Relations, Comparative Politics, and Political Methodology.
2. Learn basic machine learning algorithms and data science tools that are applied in political science research

3. Apply data analysis tools using R programming language through problem sets.

4. Collect and analyze data to learn substantive topics of own interest.

5. Learn how to communicate data-driven findings and insights.

Note: the topics covered in this class represent only a very small subset of political science research. If you enjoy this class, please consider a HASS concentration in Political Science. We also offer a major and a minor in Political Science, as well as a minor in Public Policy and a minor in Applied International Studies. Internships and research opportunities too. Check out these programs and more at: `https://polisci.mit.edu/undergraduate`.

4 Prerequisites

This class will assume that you do not have any prior exposure to political science and machine learning. One prerequisite for this course is basic programming skills in at least one language (e.g., Python). Students who have taken **6.100A: Introduction to Computer Science and Programming in Python** or the equivalent are ready to take this course. If you have any questions about whether you are prepared for this course, please talk to the instructor.

5 Notes on Computing

In this course we use R, an open-source statistical computing environment that is very widely used in statistics and data science. (If you are already well versed in another statistical software, you are free to use it, but you will be on your own). We will begin the course with an introduction to R and no prior exposure to the programming language is required. Each problem set will contain computing and/or data analysis exercises which can be solved with R but often require going beyond canned functions to write your own program.

To use R, install it by visiting https://cran.r-project.org/ and clicking the appropriate link for your operating system. *After installing R*, we strongly recommend you also install RStudio, a tremendously useful interface to work with R. To install the free RStudio Desktop, visit this webpage.

6 Policy on the Use of Generative AI

The use of generative AI tools such as ChatGPT is allowed for all assignments (problem sets and group project) in this class. However, a central goal of the class is to help you become independent and critical thinkers, so we discourage you from the *extensive* use of generative AI tools for writing code or plotting graphs in your work.

If you do use AI-generated content in your assignments, you must clearly indicate what work is yours and what part is AI-generated through proper attribution. We also ask you provide a short one-paragraph summary at the beginning of the assignment on how you used AI tools. Please consult this APA post on how to cite AI tools. Failure to do so will be considered plagiarism, and MIT's Academic Integrity policies will be followed.

7 Course Requirements

The final grades are based on the following items:

- **Problem sets** (40%): **Five** problem sets will be given throughout the semester. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will contribute equally toward the calculation of the final grade and graded on a 5-point scale. The following instructions will apply to all problem sets unless otherwise noted.

  - *Late submission will not be accepted* unless you ask for special permission from the instructor in advance (Permission may be granted or not granted, with or without penalty, depending on the specific circumstances).
  - Working in groups is encouraged, but each student must submit their own writeup of the solutions. In particular, you should not copy someone else's answers or computer code. We also ask you to write down the names of the other students with whom you solved the problems together on the first sheet of your solutions.
  - For analytical questions, you should include your intermediate steps, as well as comments on those steps when appropriate. For data analysis questions, include annotated code as part of your answers. All results should be presented so that they can be easily understood.
  - All answers should be typed. Students are strongly encouraged to use a typesetting system such as LaTeX.
  - As mentioned in Section 6, if you do choose to use generative AI tools, please include a one-paragraph summary in each assignment detailing how you used these tools.

- **Final group project** (40%): Students are expected to form a group of between 3 to 5 people. Each group will apply methods they learned in this course to an empirical problem of their own substantive interest. The final deliverable is a **10-page paper and an in-class presentation**. The paper is due on December 12 and should include a research question, a description of the dataset, descriptive analysis, data analysis and results. You will be evaluated on the clarity of the research question, the appropriateness of the chosen empirical methodology and its execution, and your presentation. You *will not* be graded on intermediate deliverables (one and five-page reports, see below). The intermediate deliverables are an opportunity for the teaching team to offer valuable feedback and comments on your project progress. We provide more details on each component below:

  - **Four Components**
    1. Research Question: What political or social phenomenon are you interested in explaining? Why is this phenomenon worth explaining? What are the dependent and independent variables? What hypotheses do you aim to test?
    2. Data Collection and Cleaning : Collecting political science data. Final output: (1) dataset and (2) replication codes.
       * Your group will either engage in your own data collection (e.g., using web-scraping) or utilize multiple existing datasets in political science. The dataset should be submitted in a standard data format (e.g., `.csv`, `sql`, `.json`) as an output of this task.

- Collecting new data: Your team will choose a topic of interest related to various subfields in political science such as American Politics, International Relations, Comparative Politics, and Political Economy. The instructor will provide guidance to identify potential sources for novel data collection.
- Utilizing existing datasets: You should merge various datasets available in political science research. The instructor will also make two of his own databases available: (1) Money in Politics Database (see `www.LobbyView.org`), and (2) International Trade data.
- A one-page report should be submitted summarizing the data collection plan and research design by October 5.

3. Descriptive Analysis: Your group will then conduct descriptive data analysis. You should submit tables and figures that effectively illustrate key patterns in your data. A five-page report should be submitted as an output of this task by November 21.
4. Data Analysis/Presentation : You will utilize various tools that you learn from the course to conduct an in-depth data analysis. Each group will give 10 minutes in-class presentation of their main findings in the last weeks of the semester.

– **Deadlines:** Please be aware of the following deadlines. Late submission will *not* be accepted. You are welcome to arrange a meeting (during the office hours) with the instructor and the TAs as you make a progress over the semester.

* **September 21**: By this date, please form your team. A group should have at least three students and consist of no more than 5 students. Your team should arrange a meeting with the recitation TA within a week after this date.
* **October 5**: By this date, your team should identify the dataset to analyze. Please submit one-page description of your project that explains (a) the specific dataset that your team is going to collect/analyze, (b) the main puzzle/problem that your team plans to study.
* **November 21**: By this date, your team should submit a five-page long report summarizing the results from your descriptive data analysis. The report should have 1-inch margins with double-spaced 12 point font text. Please submit a document with at most 5 figures or tables that summarize your data with informative caption for each.
* **December 12**: By this date, your team should incorporate the feedback from your in-class presentation and submit a 10-page long paper on your final project. The paper should have 1-inch margins with double-spaced 12 point font text. Please submit a document with at most 5 figures or tables that summarize your data with informative caption for each.

- **Participation & Attendance** (20%): Students are expected to attend all classes and actively participate in their group project. Students are strongly encouraged to ask questions and participate in discussions during lectures and recitation sessions. In addition, there will be required readings for each section of the course which students are expected to complete *prior to* the lectures.

8 Course Website

You can find the Canvas website for this course at:

https://canvas.mit.edu/courses/21355

We will distribute course materials, including readings, lecture slides and problem sets, on this website. All the assignments should be submitted to the course webpage.

9 Questions about Course Materials

In this course, we will utilize an online discussion board called *Piazza*. In addition to recitation sessions and office hours, please use the Piazza Q&A board when asking questions about lectures, problem sets, and other course materials. You can access the Piazza course page either directly from the below address or the link posted on the Stellar course website:

https://piazza.com/mit/fall2023/17835

Using Piazza will allow students to see other students' questions and learn from them. Both the TA and the instructor will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion. A student's respectful and constructive participation on the forum will count toward his/her class participation grade. *Do not email your questions directly to the instructor or TA* (unless they are of a personal nature)— we will not answer them!

10 Books

- Recommended books: We will read chapters from these books throughout the course. We strongly recommend that you at least purchase "Quantitative Social Science An Introduction." (QSS) These books will be available for purchase at COOP and online bookstores (e.g. Amazon) and on reserve in the library.
  - Imai, Kosuke. 2017 *Quantitative Social Science An Introduction*. Princeton University Press.
  - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer.
  - Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning*, Springer (A great introduction to machine learning).

11 Tentative Course Outline

11.1 Introduction

- Machine Learning and Data Science in Political Science

  Recommended Reading:
  - Justin Grimmer. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." Available at `https://stanford.edu/~jgrimmer/bd_2.pdf`
  - Tulchinsky, Theodore H. "John Snow, Cholera, the Broad Street Pump; Waterborne Diseases Then and Now." *Case Studies in Public Health* (2018): 77.

- Introduction to R Programming Language

  Required Reading:

  - Tomz, Michael, Judith L. Goldstein, and Douglas Rivers. "Do we really know that the WTO increases trade? Comment." American Economic Review 97, no. 5 (2007): 2005-2018.
  - Athey, Susan. "Beyond prediction: Using big data for policy problems." *Science* 355 (6324): 483–485.

11.2 Causality

- Causal Inference
- Average Treatment Effect (ATE) and Average Treatment Effect for the Treated (ATT)

  Required Reading:

  - Fowler, James. 2008. "The Colbert Bump in Campaign Donations: More Truthful than Truthy." *PS: Political Science & Politics* 41(3): 533–539
  - Hersh , Eitan D. 2013. "Long-Term Effect of September 11 on the Political Behavior of Victims' Families and Neighbors." *Proceedings of the National Academy of Sciences* 110 (52): 20959 —63.
  - Gerber and Green. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-scale Field Experiment." *American Political Science Review.* 102(1): 33–48

  Recommended Reading:

  - QSS: Chapter 2, available from `https://assets.press.princeton.edu/chapters/s2-11025.pdf`

11.3 Linear Regression

- OLS (Ordinary Least Squares)
- Difference in means estimator
- Regression and Causation

  Required Reading:

  - Chapter 4.2 (First Week)
  - Chapter 4.3 (Second Week)
  - Wand, Jonathan N and Shotts, Kenneth W and Sekhon, Jasjeet S and Mebane, Walter R and Herron, Michael C and Brady, Henry E. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review.* Vol 95. No. : 793–810
  - Stephens-Davidowitz , Seth I . 2014 a. "The Cost of Racial Animus on a Black Presidential Candidate: Evidence Using Google Search Data." Journal of Public Economics. 118 : 26–40

  Recommended Reading:

  - Eggers, Andrew C., and Jens Hainmueller. "MPs for sale? Returns to office in postwar British politics." *American Political Science Review* 103, no. 4 (2009): 513-533.

11.4 Supervised Learning

- Introduction to Supervised Learning
- K-Nearest-Neighbor (KNN) Classifier
- Support Vector Machine (SVM)
- Over fitting
- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)

Required Reading:

- Francisco Cantú and Sebastián M. Saiegh. 2011 "Fraudulent Democracy? An Analysis of Argentina's Infamous Decade Using Supervised Machine Learning." *Political Analysis*. 19: 409–433
- Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran et al. "A large-scale analysis of racial disparities in police stops across the United States." *Nature Human Behaviour* (2020): 1-10.
- Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa." *Nature Communications* 11, no. 1 (2020): 1-11.

Recommended Reading:

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 3.1–3.4, Ch 7.

11.5 Unsupervised Learning Methods

- Principal Component Analysis (PCA)
- Clustering Algorithm

  Required Reading:

  - Pan, Jennifer, and Yiqing Xu. "China's ideological spectrum." The Journal of Politics 80, no. 1 (2018): 254-273.
  - In Song Kim, Steven Liao, and Kosuke Imai. "Measuring Trade Profile with Two Billion Observations of Product Trade." *American Journal of Political Science*, (2020), Vol 64, No. 1, pp. 102–117.

  Recommended Reading:

  - QSS: Chapter 3.7

- Mixture Models and EM Algorithm

  Required Reading:

- Bishop Ch.9
- In Song Kim, Steven Liao, and Kosuke Imai. "Measuring Trade Profile with Two Billion Observations of Product Trade." *American Journal of Political Science*, (2020), Vol 64, No. 1, pp. 102–117.

11.6 Text Analysis

- Introduction to Text Analysis

  Required Reading:

  - Gary King, Jennifer Pan, and Margaret E Roberts. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review*, 107.2: 326-343.
  - Maya Berinzon and Ryan Briggs, "60 years later, are colonial-era laws holding Africa back?" *The Washington Post*, available here.
  - Rodman, Emma. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* 28, no. 1 (2020): 87-111.

  Recommended Reading:

  - QSS: Chapter 5.1
  - Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* (2013): 28.

- Latent Dirichlet Analysis (LDA)

  Reading:

  - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3 (2003): 993-1022.
  - Roberts, Margaret E., et al. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* (2014).

- Word Embeddings

  Recommended Reading:

  - Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey. "Efficient estimation of word representations in vector space." 2013. `arXivpreprintarXiv:1301.3781`.
  - Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff. "Distributed representations of words and phrases and their compositionality". 2013. *Advances in neural information processing systems*. pp. 3111–3119.
  - Ludovic Rheault and Christopher Cochrane. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora" *Political Analysis*. 2020. Vol. 28, No. 1, pp. 112-133.

11.7 Network Analysis

- Network Analysis

  Required Reading:

  - Fowler, James H. "Connecting the Congress: A study of cosponsorship networks." *Political Analysis*. (2006): 456-487.
  - Kim, In Song and Dmitriy Kunisky. "Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics." *Political Analysis* (2020). `http://web.mit.edu/insong/www/pdf/network.pdf`

  Recommended Reading:

  - QSS: Chapter 5.2

11.8 Applications in Political Science

- International Trade with Big Data

  Reading:

  - C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann. "The Product Space Conditions the Development of Nations." *Science* 317.5837 (2007): 482-487

- Lobbying and Campaign Contribution

  Reading:

  - In Song Kim. "Political Cleavages within Industry: Firm-level Lobbying for Trade Liberalization." *American Political Science Review*, 111.1: 1-20.
  - Stephen Ansolabehere, John M. de Figueiredo, and James M. Snyder. "Why is There so Little Money in U.S. Politics?" *Journal of Economic Perspectives*, 17.1 (2003): 105-130

- Identifying Behavioral Patterns using Massive Data

  Reading:

  - Gary King, Jennifer Pan, and Margaret E Roberts. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review*, 107.2: 326-343.
  - Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Ramachandran, V., Phillips, C., and Goel, S. (2017). "A large-scale Analysis of Racial Disparities in Police Stops across the United States." arXiv preprint arXiv:1706.05678.

- Measuring Ideological and Political Preferences using Social Network Data

  Reading:

  - Robert Bond and Solomon Messing. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109.1 (2015): 62-78.
  - Pablo Barberá "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23.1 (2014): 76-91

- What do Politicians Do?

  Reading:

  - Justin Grimmer, Solomon Messing, and Sean Westwood. "How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation." *American Political Science Review*, 106.4 (2012), 703-719
  - Justin Grimmer. "Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science*, 57.3 (2013), 624-642

- Big Administrative Data: Promises and Pitfalls

  Reading:

  - Connelly, R., Playford, C.J., Gayle, V., Dibben, C., 2016. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." *Social Science Research*, Special issue on Big Data in the Social Sciences 59, 1–12
  - Kopczuk, W., Saez, E., Song, J., 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937." *The Quarterly Journal of Economics* 125, 91–128.
  - Jens Hainmueller and Dominik Hangartner, 2013. "Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination." *American Political Science Review* 107.1, 159–187.

- Machine Learning Algorithms in Society

  Reading:

  - Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1):237–93