17.835: Machine Learning and Data Science in Politics

Fall 2020

Instructor: Professor In Song Kim
TAs: Jesse Clark, Sean Shiyao Liu, and Nicole Wilson

Department of Political Science
MIT

1 Contact Information

	In Song Kim	Jesse Clark	Sean Shiyao Liu	Nicole Wilson
Email:	insong@mit.edu	jtclark@mit.edu	ssliu@mit.edu	newilson@mit.edu
Office Hours:	Fri 4pm-5pm Eastern time	Tue 1pm-2pm Eastern Time	Tue 10pm-11pm Eastern time	Mon 2pm-3pm Eastern time

2 Logistics

- Lectures: Mondays and Wednesdays 11:00am – 12:30pm via Zoom
  - Each Zoom session *will be recorded* and be made available for remote access (the link and the password will be announced as soon as each session is available).
  - *Please do not download/publicly-disseminate the lectures*
- Recitations: Fridays. There will be three weekly recitation sessions. We will cover a review of the course material, problem set questions, and also provide help with computing issues. Attendance is strongly encouraged. Recitation group will be chosen randomly at the beginning of the semester (if a student has a scheduling constraint, please contact the teaching team by September 6).
  - Nicole's recitation sessions are held Friday 9am - 10am Eastern time on Zoom
  - Jesse's recitation sessions are held Friday 1pm -2pm Eastern Time on Zoom
  - Sean's recitation sessions are held Friday, 10pm - 11pm Eastern time on Zoom

Note that the first class meets on September 2. No class will be held on September 7 (Labour Day), October 12 (Columbus Day: We will meet on October 13 instead), November 11 (Veterans Day), November 23, 25 (Thanksgiving). Last day of class is December 9.

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 government responses to COVID-19, 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 Course Format

4.1 Synchronous Learning via Zoom

Each lecture (Mondays and Wednesdays 11:00am – 12:30pm) consists of two parts: (1) interactive lecture session (including Q&A), and (2) Zoom Breakout Rooms discussion session.

1. The lecture part will introduce new computational methods for social science research. We will also introduce various findings/debates in political science research.
2. The Zoom Breakout Rooms session will allow students to discuss both methodological and substantive topics of the day as a group. If the discussion is based on a specific reading, we will post an announcement on the course webpage about any required readings (also, please see *Required Reading* in Section 11).

We encourage students to attend the lecture synchronously and participate in the discussion session. Some of discussion topics will also be used as a basis for problem set questions.

4.2 Zoom Breakout Rooms Discussion

- A group of approximately 6 students will be randomly chosen for each discussion session (for about 15 minutes). Toward the end of the semester we may break out based on the groups that formed for projects. We will provide a shared Google document so that students can effectively take notes on and summarize their discussion in real time. We may ask the

members of one or two groups to share their ideas with the entire class once the breakout room session is over.

- Each student is expected to lead three discussion sessions over the course of the semester. We will distribute a sign-up sheet at the beginning of the semester so that students can choose their three preferred dates in advance. The leader will be responsible for facilitating the discussion and posting a short (no more than three paragraphs) summary report on the "Discussions" section of the course webpage after the lecture. This report should be submitted *before* the next class begins.

4.3 Asynchronous Learning

We understand that some students may not be able to attend all lectures synchronously. Students who cannot attend a lecture synchronously are expected to:

1. Watch the recorded lecture *prior to* the next class, e.g., watch Monday's lecture before 11:00am on Wednesday.
2. Complete the required readings for the class.
3. Submit a short note regarding the Zoom Breakout Rooms discussion topic and the readings for each lecture that they could not attend synchronously. This note should be submitted on the "Discussions" section of the course webpage *before* the next class begins.

5 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 **6.0001: 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.

6 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 lecture/recitation 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.

7 Course Requirements

Due to the COVID-19 outbreak, a letter grading system with extra flexibility will be in effect. The final grades that will be awarded include A, B, C, D/NE, and F/NE, where NE indicates that no

record will appear on the external transcript. First-year undergraduate students will be graded on the normal P/NR basis for all subjects in the fall semester (more details on the grading policy for the semester are available here). This final grade will be based on the following items:

- **Problem sets** (35%): **Five** problem sets will be given throughout the semester (each will be due one week). Problem sets will contain analytical, computational, and data analysis questions. Each problem set will contribute equally toward the calculation of the final grade (We will use Letter Grade grading scheme with (+/-) modifiers, e.g., A+, A, A-, B+, ...). 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 and submitted in two separate file formats on the course webpage: (1) answers in a .pdf and codes in .R, e.g., `kim_pset1.pdf`, `kim_pset1.R`. Students are strongly encouraged to use a typesetting system such as LaTeX and R Markdown. Please do not send your answers using email.

- **Final problem set** (15%): The last problems set will be a special problem set, which will be weighted more heavily toward the calculation of the final grade. *You will not be allowed to collaborate with anybody on the final problem set.* This is to check if you have developed sufficient experience to work through problems on your own. The final problem set will be distributed on November 16 and will be due December 2.

- **Final group project** (35%): Students are expected to form a group. Each group will apply methods they learned in this course to an empirical problem of own substantive interest. If your group is not sure whether a certain group project topic is appropriate for the course, please contact the teaching team by September 21. The group project will be evaluated based on the performance across the following four tasks.

  - **Four Tasks**
    1. Data Collection and Research Design (5%):
       * 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 September 30.

2. Descriptive Analysis (10%): 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 (including appendix and references) should be submitted as an output of this task by November 9.
3. Presentation (5%): 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.
4. Final Poster(15%): A poster should be submitted as a final output of the project on the last day of the class.

– **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. We will create a forum so that students in the same recitation group can find members of their final group project. 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.
* **September 30**: 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 9**: 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 9**: By this date, your team should submit the final presentation of your project.

- **Participation & Attendance** (15%): Students are expected to actively participate in the Zoom Breakout Rooms sessions. Each student will be assigned to lead three group discussion sessions over the course of the semester (see Section 4.2). Students who could not participate in the Zoom Breakout Rooms discussion synchronously should (1) watch the recorded lecture and (2) submit a short note on the discussion topic and required readings (if any) *prior to* the following class. Students are strongly encouraged to ask questions and participate in discussions during online lectures and recitation sessions.

8 Course Website

You can find the Canvas website for this course at:

https://canvas.mit.edu/courses/3574

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 Canvas course website:

https://piazza.com/mit/fall2020/17835

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 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.

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

  Reading:

  – Bishop Ch.9

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)

  Recommended 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

We will discuss various applications of machine learning and data science tools in political science throughout the semester

- 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:

  – 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