17.831 - Data and Politics

Daniel Hidalgo

September 30, 2019

Instructor: F. Daniel Hidalgo (dhidalgo@mit.edu)

Time: Monday and Wednesday 1:00-2:30, Spring 2019

Location: 56-169

Office Hours : Thursday 3-5pm in E53-402

Course Description

After the 2012 re-election of Barack Obama, Time Magazine proclaimed that data "played a huge role in creating a second term for the 44th President". According to Time, traditional campaign professionals are being replaced “by the work of quants and computer coders who can crack massive data sets for insight”. Others, like political scientists John Sides and Lynn Vavreck are more skeptical, arguing that data crunching “did not win the election”. Can the analysis of huge datasets help win elections? More broadly, how has “big data” affected how citizens interact with parties and politicians?

Whatever its impact may be, the growing availability of data and the development of new technologies to analyze it has started to change the practice of electoral politics. Political candidates and parties now spend large sums of money compiling huge datasets, hiring programmers, and building teams of social scientists to maintain an edge in hard fought elections. Large scale data analysis has been widely used in business and other fields, but its use in politics is relatively new. The marriage of "big data" and old fashioned retail politics is, according to some, altering how citizens are represented by politicians, changing the composition of the electorate, and transforming how campaigns are now run.

Despite the hype, data does not speak for itself. The proper use of data for decision-making in politics (or any field) rests on basic statistical and social scientific principles. Three foundational concepts for the successful analysis of data are *sampling*, *causal inference*, and *predictive inference*. The basic principle underpinning sampling and causal inference is that descriptive or causal conclusions require an understanding of how the data was generated. When data is combined with a detailed understanding of how the sample was created, powerful insights about the nature and causes of social behavior are possible. For prediction, statistical learning theory (or "machine learning") provides a framework for combining algorithms and data on past behavior that can be useful for predicting future behavior. All three approaches to learning from data are now heavily used in electoral politics, business, and even the nonprofit sector.

In this course, students will both learn how statistics are changing elections and how to use statistics to analyze political data. While the substantive focus will be on elections, the principles and methods learned in this course have broad applicability to the decision-making in a broad variety of fields. The course will be roughly divided into 4 sections organized around a different methodological topic, with an application to an electoral phenomenon. For each section, students will work with the professor on analyzing a unique dataset related to electoral politics. The first section will focus on data description and dimension reduction. The second section will involve the analysis of survey data on electoral behavior. The third section will use statistical models to predict electoral behavior using large datasets. The fourth section will focus on the design and implementation of original experiments in order to study political attitudes and behaviors.

Course Objectives

By the end of this course, students will be able to:

- Describe why and how the use of data and statistical methods is influencing decision-making in elections.
- Understand the basic principles of social science statistics.
- Analyze data using modern statistical computing tools, in particular, the statistical programming language `R`.
- Complete original projects that will involve collection, analysis, and interpretation of data used in campaigns today.

Books and Computation

Books

The book you are required to purchase are:

- Nolan McCarty (June 4, 2019). *Polarization: What Everyone Needs to Know*. Oxford University Press. 276 pp.

Books that you *may* purchase, but are also freely available online are:

- Hadley Wickham and Garrett Grolemund (Jan. 10, 2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1 edition. Sebastopol, CA: O'Reilly Media. 522 pp.
- Gareth James et al. (Sept. 1, 2017). *An Introduction to Statistical Learning: With Applications in R*. 1st ed. 2013, Corr. 7th printing 2017 edition. New York: Springer. 426 pp.

Computation

The assignments in this course will require the use of `R`, a programming language and software environment for statistical computing that is heavily used in statistics and related fields.

- R is free and can be downloaded and installed from CRAN, the Comprehensive R Archive Network.
- As an interface to R, I strongly recommend that you use RStudio, a powerful integrated development environment (IDE) for `R`.

R is free, but sometimes configuration and installation can be non-trivial. Particularly at the beginning of the course, you may want to use the free RStudio Cloud service, which allows you to run RStudio and R in your browser. RStudio Cloud can be slow with larger datasets, however, so eventually you will want to install R and RStudio on your own laptop.

Please bring your laptops to class, as we will frequently be doing in-class activities that require R.

Assignments and Grading

The grade for this course will be based on the following components:

- **Biweekly homework assignments** (50%)
  - These biweekly assignments are generally due on Wednesday before midnight. Weekly assignments will typically take the form of a single **R** Markdown text file, which is a document format that allows for code and text to be interspersed in the same document. You may work with others on your homework, but your writeup must be your own. Before you turn in your homework, please be sure that your document compiles.
  - Homework will be graded on a 10 point scale. Your homework with the lowest grade at the end of the semester will be dropped. Late homework will not be accepted without permission from the instructor.
  - Please turn in your homework via the class website.
- **Research project** (40%)
  - This is a group project where students will apply the methods developed in this course to a an empirical problem that is of substantive interest to the group. The project will have multiple components:
    1. Data collection. (10%). Your group will collect some original data. A description of the data and evidence that you can obtain it should be submitted by 10/14.
    2. Descriptive Analysis (15%): Your group will conduct a descriptive data analysis. You should submit tables and figures that illustrate the basic patterns in your data. A 4 page report (figures and tables with informative captions included) should be submitted by 11/18.
    3. Data Analysis (15%): You will use the tools you learned in the course to conduct an in-depth data analysis of your data. Each group will give a 10 minute in-class presentation in the last week of the semester. A poster should be submitted as a final output on the last day of the class.
- **Class Participation** (10%)
  - Class time will be a mix of lecture and active learning. In a typical class, I will introduce the basic concept in lecture and then students will apply the concept through coding exercises. Your participation grade will reflect effort applied during these exercises.

Class Website

The course website is http://stellar.mit.edu/S/course/17/fa19/17.831/

In addition to readings, slides, and this syllabus, all problem sets will be distributed there.

Office Hours and Getting Help

My office hours are on Tuesdays 3:00pm to 5:00pm. Please sign up for time slots via this website: `https://calendly.com/fdhidalgo/office-hours`

If you run into problems that you can't solve on your own, please use the class website on Piazza to post questions. I strongly encourage students to assist in answering questions as they come up. Unless absolutely necessary, it is better to post on Piazza than email me as everyone can benefit from the posted responses. I will monitor Piazza and try to answer within 24 hours. The site is: `https://piazza.com/mit/fall2019/17831`

Class Schedule

Lecture Schedule

Lecture	Date	Topic
1	9/4	Introduction to the Course
2	9/9	Visualization and Data Wranging in `R`
3	9/11	Measuring and Describing Political Polarization
4	9/16	Dimension Reduction, Working with Campaign Finance Data
5	9/18	Dimension Reduction, Polarization
6	9/23	Dimension Reduction, Correspondence Analysis
7	9/30	Survey Sampling in Politics
8	10/2	The Statistics of Surveys
9	10/7	Asking Questions in Political Surveys
10	10/9	Non-Response and Measurement Error in Political Surveys
11	10/16	Political Polarization in the Electorate
12	10/21	Data Analytics in Campaigns
13	10/23	Predictive Modeling and Over-fitting
14	10/28	Predictive Modeling with Linear Models
15	11/4	Predictive Modeling with Voter Files in Elections
16	11/6	Predictive Modeling with Regression Trees
17	11/13	Election Forecasting
18	11/18	Causation and Experiments
19	11/20	Uncertainty in Experiments
20	11/25	Survey Experiments in Politics
21	11/27	*No Class*
22	12/2	Designing Survey Experiments
23	12/4	Voter Mobilization and Persuasion Experiments
25	12/9	Class Presentations
25	12/11	Class Presentations

Assignment Schedule

Due Date	Assignment
9/18	Problem Set 1
10/2	Problem Set 2
10/14	Project Data Description
10/16	Problem Set 3
10/30	Problem Set 4
11/13	Problem Set 5
11/18	Project Descriptive Analysis
11/27	Problem Set 6
12/9	Class Presentations
12/11	Class Presentations
12/11	Project Data Analysis Due

Lecture Readings

Introduction to the Course

- RFDS Chapter 1

Visualization and Data Wranging in `R`

- *Topics*: visualization, data transformation
- *Reading*: RFDS Chapters 3-5

Measuring and Describing Political Polarization

- *Topics*: political polarization, exploratory data analysis
- *Reading*: Nolan Chapter 2-3; RFDS Chapter 7

Dimension Reduction, Working with Campaign Finance Data

- *Topics*: money in politics, principal components analysis, data frames, variable types
- *Readings:*
  - RFDS Chapters 10, 13, 15
  - Adam Bonica (Dec. 2016). “Avenues of Influence: On the Political Expenditures of Corporations and Their Directors and Executives”. In: *Business and Politics* 18.4, pp. 367–394. URL: `https://www.cambridge.org/core/journals/business-and-politics/article/avenues-of-influence-on-the-political-expenditures-of-corporations-and-their-directors-and-executives/F9EF7632D19B480BACEA337FF0516ADE` (visited on 08/19/2019)
  - Maggie Koerth-Baker, How Money Affects Elections

Dimension Reduction, Polarization in Social Media

- *Topics*: social media, principal components analysis, functions in `R`

- *Readings*:
  - ISL Sections 10.1-10.2
  - RFDS Chapter 19

Survey Sampling in Politics

- *Topics*: surveys, sampling, loops, functionals
- *Readings*:
  - Jill Lepore, "Politics and the New Machine"
  - Chapter 2 in Robert M. Groves et al. (July 14, 2009). *Survey Methodology*. 2 edition. Hoboken, N.J: Wiley. 488 pp.
  - RFDS Chapter 21

The Statistics of Surveys

- *Topics*: unbiasedness, sampling error
- *Readings*:
  - Chapter 4 in Robert M. Groves et al. (July 14, 2009). *Survey Methodology*. 2 edition. Hoboken, N.J: Wiley. 488 pp.
  - Will Jennings and Christopher Wlezien (Apr. 2018). “Election Polling Errors across Time and Space”. In: *Nature Human Behaviour* 2.4, pp. 276–283. URL: `https://www.nature.com/articles/s41562-018-0315-6` (visited on 08/21/2019)

Asking Questions in Political Surveys

- *Topics*: question design, measurement error
- *Readings*:
  - Nora Cate Schaeffer and Stanley Presser (2003). “The Science of Asking Questions”. In: *Annual Review of Sociology* 29.1, pp. 65–88. URL: `https://doi.org/10.1146/annurev.soc.29.110702.110112` (visited on 08/15/2019)
  - Maggie Koerth-Baker, "The Tangled Story Behind Trump’s False Claims of Voter Fraud"

Non-Response and Measurement Error in Political Surveys

- *Topics*: non-response bias, post-stratification, raking
- *Readings*:
  - Scott Keeter et al. (2017). *What Low Response Rates Mean for Telephone Surveys*, pp. 1–39. URL: `https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/`

Political Polarization in the Electorate

- *Topics*: polarization, measuring ideology, consistency
- *Readings*:

  - Nolan Chapter 4
  - David E. Broockman (2016). "Approaches to Studying Policy Representation". In: *Legislative Studies Quarterly* 41.1, pp. 181–215. URL: `http://onlinelibrary.wiley.com/doi/abs/10.1111/lsq.12110` (visited on 08/16/2019)

Data Analytics in Campaigns

- *Topics*: turnout and persuasion scores, voter files
- *Readings*: Chapters 1,2, XX, XX in Eitan Hersh (June 9, 2015). *Hacking the Electorate: How Campaigns Perceive Voters*. New York, NY: Cambridge University Press. 272 pp.

Predictive Modeling and Over-fitting

- *Topics*: regression, over-fitting, bias / variance tradeoff
- *Readings*: ISL Chapter 1

Predictive Modeling with Linear Models

- *Topics*: OLS, linear models in `R`
- *Readings*: ISL Chapter 3, RFDS Chapter 23

Predictive Modeling with Voter Files in Elections

- *Topics*: lasso models
- *Readings*:
  - ISL Chapter 4.1-4.3
  - David W. Nickerson and Todd Rogers (May 2014). "Political Campaigns and Big Data". In: *Journal of Economic Perspectives* 28.2, pp. 51–74. URL: `https://www.aeaweb.org/articles?id=10.1257/jep.28.2.51` (visited on 08/17/2019)

Predictive Modeling with Regression Trees

- *Topics*: classification and regression trees
- *Readings*: ISL Chapter 8

Election Forecasting

- *Topics*: election forecasting, variable selection
- *Readings*:
  - Benjamin E. Lauderdale and Drew Linzer (July 1, 2015). "Under-Performing, over-Performing, or Just Performing? The Limitations of Fundamentals-Based Presidential Election Forecasting". In: *International Journal of Forecasting* 31.3, pp. 965–979. URL: `http://www.sciencedirect.com/science/article/pii/S0169207015000102` (visited on 08/21/2019)
  - Nate Silver, How FivethirtyEight's House, Senate, and Governor Models Work

Causation and Experiments

- *Topics*: causation, potential outcomes, experiments
- *Readings*:
  - Chapter 1 in Joshua D. Angrist and Jörn-Steffen Pischke (Dec. 21, 2014). *Mastering 'Metrics: The Path from Cause to Effect*. with French flaps edition. Princeton ; Oxford: Princeton University Press. 304 pp.

Uncertainty in Experiments

- *Topics*: sampling distribution, hypothesis testing
- *Readings*: TBD

Survey Experiments in Politics

- *Topics*: survey experiments, list experiments, randomized response
- *Readings*:
  - Regina Bateson (June 30, 2019). *Strategic Discrimination*. SSRN Scholarly Paper ID 3412626. Rochester, NY: Social Science Research Network. URL: `https://papers.ssrn.com/abstract=3412626` (visited on 08/19/2019)
  - Bryn Rosenfeld, Kosuke Imai, and Jacob N. Shapiro (2016). "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions". In: *American Journal of Political Science* 60.3, pp. 783–802. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12205` (visited on 08/19/2019)

Designing Survey Experiments

- *Topics*: survey experiments, survey platforms
- *Readings*: TBD

Voter Mobilization Experiments

- *Topics*: turnout experiments, social pressure
- *Readings*:
  - Chapter 3 and Chapter 7 in Sasha Issenberg (Sept. 17, 2013). *The Victory Lab: The Secret Science of Winning Campaigns*. Reprint edition. New York: Broadway Books. 416 pp.
  - Chapter 12 in Donald P. Green and Alan S. Gerber (Aug. 27, 2019). *Get Out the Vote: How to Increase Voter Turnout*. 4 edition. Brookings Institution Press. 260 pp.

Voter Persuasion Experiments

- *Topics*: persuasion, accountability, meta-analysis
- *Readings*:
  - Joshua L. Kalla and David E. Broockman (Feb. 2018). "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments". In:

American Political Science Review 112.1, pp. 148–166. URL: https://www.cambridge.org/core/journals/american-political-science-review/article/minimal-persuasive-effects-of-campaign-contact-in-general-elections-evidence-from-49-field-experiments/753665A313C4AB433DBF7110299B7433 (visited on 08/19/2019)

- Thad Dunning et al. (July 1, 2019). “Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials”. In: *Science Advances* 5.7, eaaw2612. URL: https://advances.sciencemag.org/content/5/7/eaaw2612 (visited on 08/19/2019)

Designing Field Experiments

- *Topics:*
- *Readings*: None