17.831 Data and Politics

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

Time: Monday and Wednesday 11:00-12:30, Spring 2016

Location: 4-159

Office Hours : Tuesdays 3:20-5:00pm at 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".[1] 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”.[2] 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 have 3 modules organized around a different methodological topic, with an application to an electoral phenomenon. For each module, students will work with the professor on analyzing a unique dataset related to electoral politics. The first module will involve the analysis of survey data on electoral behavior. The second module will focus on the design

and implementation of original experiments in order to study political attitudes. The third module will use statistical models to predict electoral behavior using large datasets.

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: http://web.mit.edu/polisci/undergraduate/index.html.

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 books you are required to purchase are:

- Agresti, Alan. 2009. *Statistical Methods for the Social Sciences* (4th edition). Pearson Prentice Hall.
    - Note: this book is absurdly expensive to purchase outright, so I encourage you to rent it from Amazon.com or Barnes and Noble if you wish to save money.
- Hersh, Eltan. 2015. *Hacking the Electorate: How Campaigns Perceive Voters*. Cambridge University Press.
- Issenberg, Sasha. 2012. *The Victory Lab: The Secret Science of Winning Campaigns*. Broadway Books.
- Sides, John and Vavreck, Lynn. 2014. *The Gamble: Choice and Chance in the 2012 Presidential Election*. Princeton University Press.

The following required books are *free* to download (printed copied are available for purchase):

- Gareth, James et al. 2014. *An Introduction to Statistical Learning with Applications in R*. Springer. link
- Peng, Roger. 2016. *R Programming for Data Science*. Leanpub. link
- Peng, Roger. 2015. *Exploratory Data Analysis with R*. Leanpub. link

Other articles and materials will be posted on the Stellar site as needed.

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.

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

Assignment and Grading

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

- **Weekly homework assignments** (50%)
    - These weekly assignments are generally due on Wednesdays 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 following link: https://www.dropbox.com/request/thcPQiDGAgrBLXewmCXO
- **Reading Quizzes** (10%)
    - Reading quizzes are very short open book multiple choice questions that are to completed online before the Monday's of each class. The quiz will become accessible shortly after Wednesday's class and your answers are due half an hour before the next class (10:30am). A link to the quiz will be posted on the class website. Each student's lowest reading quiz grades will be dropped when calculating the final grade.
- **Research project** (30%)
    - This is a group project where students will design and implement an original survey experiment on subjects recruited via Facebook or Amazon Turk. Groups must settle on their experimental design by *March 30*. The final questionnaire is due on *April 13*. This questionnaire will be programmed into the Qualtrics platform. Data will be returned to the groups by *April 25*. Results will be written up and presented as scientific posters on *May 11*.
- **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.

Office Hours and Getting Help

My office hours are on Tuesdays 3:20pm to 5:00pm. Please sign up for time slots via this website: https://calendly.com/fdhidalgo/officehours

If you run into problems that you can't solve on your own, please use Piazza to post questions. Unless absolutely necessary, it is better to post on Piazza than email me as everyone can benefit from my answer. I will monitor Piazza and try to answer within 24 hours. The site is: https://piazza.com/class/ik31qm2zmzc1v1?cid=4#

Course Schedule

The schedule below is tentative and subject to change, depending on time and class interests. We will move at a pace dictated by class discussions and electoral politics. Please check this page often for updates.

Introduction to the Course (*February 3*)

- Introduction and orientation
- Class survey
- Introduction to R

Introduction to R (*February 8*)

- *R Programming for Data Science*: Chapters 4, 5, 6, 10

The Political Science of Elections (*February 10*)

Homework 1 Due

- Sides and Vavreck: Chapters 2, 3, and 4

Sampling and Surveys (*February 16*)

- Agresti and Finlay: Chapter 2
- Jill Lepore, "Politics and the New Machine": link
- Mark Blumenthal, "2012 Poll Accuracy: After Obama, Models and Survey Science Won the Day": link
- *R Programming for Data Science*: Chapter 11, 13

Descriptive Statistics (*February 17*)

Homework 2 Due

- Agresti and Finlay: Chapter 3
- *Exploratory Data Analysis in R*: "Principles of Analytic Graphics" and "Exploratory Graphs"

Visualization in R (*February 22*)

- Exploratory Data Analysis in R: "Plotting Systems", "Graphics Devices", "The ggplot2 Plotting System: Part 1"
- Wand et al. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida". *American Political Science Review* 95 (4). link
    - Note: Skim the methodological details and focus on how the graphical presentation of data is used to support the authors' argument, especially in the section entitled "A Natural Experiment: Florida's Election-Day and Absentee Voters in 2000".

Probability (*February 24*)

Homework 3 Due

- Agresti and Finlay: Chapter 4 (pgs. 73-85)
- Bernd Beber and Alexandra Sacco, "The Devil Is in the Digits". link
- *R Programming for Data Science*: Chapter 14

Sampling Distributions (*February 29*)

- Agresti and Finlay: Chapter 4 (pgs. 85-99)
- *R Programming for Data Science*: Chapter 15

Estimation and Confidence Intervals (*March 2*)

Homework 4 Due

- Agresti and Finlay: Chapter 5

Nonresponse in Surveys (*March 7*)

- Chapter 6 in Groves et al. 2009. *Survey Methodology, Second Edition*. John & Wiley Sons, Inc.
- Andrew Gelman. "Tracking Public Opinion with Biased Polls": link

Causation and Experiments (*March 9*)

*Homework 5 Due *

- Chapter 1 in Joshua Angrist and Jorn-Steffen Pischke. 2015. *Mastering Metrics: The Path from Cause to Effect*. Princeton University Press.

Survey Experiments in the Social Sciences (*March 14*)

- Chapters 2-4 in Diana Mutz. 2011. *Population-Based Experiments*. Princeton University Press.
- Berinsky, Adam. 2016. "Rumors and Health Care Reform: Experiments in Political Misinformation". *British Journal of Political Science*. link

Experimental Design (*March 16*)

Homework 6 Due

Hypothesis Tests (*March 28*)

- Agresti and Finlay: Chapter 6, Chapter 7 (pgs. 183-192)
- Chapter 5 in James Monogan. 2015. *Political Analysis Using R*. Springer. Link

Voter Mobilization (*March 30*)

Homework 7 Due

- *The Victory Lab*: Chapter 3, 7, 8
    - Alan Gerber, Donald Green, and Christopher Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment". *American Political Science Review* 102 (February): 33-48. Link

Data Analytics in Campaigns (*April 4*)

- *Hacking the Electorate*: Chapters 1-4.
- John Sides and Lynn Vavreck, "Obama's Not-So-Big Data," *Pacific Standard*, January 21, 2014 link

Regression 1 (*April 6*)

Homework 8 Due

- Agresti and Finlay: Chapter 9 (pgs 255-265)
- James et al: Chapter 3 (pgs. 59-71, 109-113)
- *Hacking the Electorate*: Chapter 5.

Regression 2: Multivariate Analysis (*April 11*)

- Agresti and Finlay: Chapter 11 (pgs. 321-340)
- James et al: Chapter 3 (pgs. 71-90, 113-114)
- Seth Masket. 2008. "Does Obama's Ground Game Matter? The Influence of Local Field Offices During the 2008 Presidential Election". *Public Opinion Quarterly* 75(5).:link

Regression 3: Interactions and Nonlinearities (*April 13*)

- Agresti and Finlay: Chapter 11 (pgs. 340-345), Chapter 14 (pgs 463-468)
- James et al: Chapter 3 (pgs. 82-92, 115-119)

Overfitting / Bias Variance Tradeoff (*April 20*)

Homework 9 Due

- James et al: Chapter 2 (pgs. 15-36)

Resampling Methods (*April 25*)

- James et al: Chapter 5 (pgs. 175-186, 190-194)

Variable Selection (*April 27*)

Homework 10 Due

- James et al: Chapter 3 (pgs. 203-214, 244-251)

Regularization (*May 2*)

- James et al: Chapter 6 (pgs. 214-228, 251-255)

Election Forecasting (*May 4*)

- *The Victory Lab*: Chapter 9, 10
- James Cambpell. 2012. "Forecasting the 2012 American National Elections". *PS* (October). link
- Nate Silver, "How We're Forecasting the Primaries". link
- Drew Linzer, "How It Works". link

No Class (*May 9*)

Poster Presentation (*May 11*)

---

1. Michael Scherer, “Inside the Secret World of Data Crunchers Who Helped Obama Win,” *Time*, November 7, 2012. link ↩

2. John Sides and Lynn Vavreck, “Obama's Not-So-Big Data,” *Pacific Standard*, January 21, 2014. link ↩