GOVT 401: Data Science & Politics

Professor Waggoner
Department of Government
441 Tyler Hall

www.philipdwaggoner.com
pdwaggoner@wm.edu
Office Hours: Monday 2-4 pm, or by email or appt.

Location: Tyler Hall 134

Day/Time: T TR, 3:30–4:50 pm

1 Overview & Introductory Remarks

Welcome to Data Science and Politics! In this class, we will be walking through a host of statistical and data science techniques aimed at exploring and better understanding (and diagnosing) substantive political phenomena. Specifically, we will link substantive political/social concepts with targeted statistical and computational methods. Rather than focus on derivations of statistical models, the main focus of the course will be applying and diagnosing model fit, along with computation and application in the `R` computing language. The goals of the class are twofold: first, to offer students a methodological toolbox to tackle complex questions of interest in the social sciences. The second goal, then, is to prepare students for quantitative social and political research, offering data science techniques and computational training in the service of understanding and predicting human behavior in a range of contexts. In sum, we are interested in joining the data science and social science worlds to use each to more fully explain the other.

In the recently released statement, *Envisioning the Data Science Discipline: The Undergraduate Perspective*, the National Academies of Sciences, Engineering, and Medicine offered their vision (and justification) of high quality data science undergraduate education.[1] Specifically they note,

> *The need to manage, analyze, and extract knowledge from data is pervasive across industry, government, and academia. Scientists, engineers, and executives routinely encounter enormous volumes of data, and new techniques and tools are emerging to create knowledge out of these data, some of them capable of working with real-time streams of data. The nation's ability to make use of these data depends on the availability of an educated workforce with necessary expertise. With these new capabilities have come novel ethical challenges regarding the effectiveness and appropriateness of broad applications of data analyses.*

In this class, we are seeking to meet this challenge, by learning to effectively leverage methods and techniques to "extract knowledge from data." Practically, each week we will focus on and pair a different concept with an appropriate method. For example, we may focus on measuring U.S. Supreme Court ideology using Bayesian Item-Response Theory (IRT) models and Markov-Chain Monte Carlo (MCMC) methods to estimate ideal points. The method (IRT) and the substantive topic (SCOTUS ideology) will reinforce the other, where we will read some articles and discuss what has been done on both of these topics, and then walk through precisely how to apply the method in the given context in `R`. We will begin by learning how to think about complex questions from a "data science" perspective (e.g., diagnosing the "feature space," considering a paradox, evaluating several approaches that do similar things, and so on), and then how to select the appropriate method or model to address our question of interest.

[1] National Academies of Sciences, Engineering, and Medicine. 2018. Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report. Washington, DC: The National Academies Press. https://doi.org/10.17226/24886.

Prerequisite: The expectation is that students will have a firm grasp on at least *basic* statistical concepts, though ideally a solid understanding of intermediate statistics as well as computation and programming in `R` and LaTeX. Though I will provide an overview of basic features of these programs at the outset, we will jump into mechanics relatively quickly thereafter.

The course will cover a different concept and method each week. On Tuesdays, we will read and discuss several articles related to key concepts, both substantive and methodological. I will lecture on more technical aspects as needed to fill in the gaps in understanding. Then, Thursdays will be reserved for application, where we work exclusively in `R` to implement the given method using a social science dataset, which I will provide. The idea with this format is to provide a solid foundation of the concept as it is treated in the field, as well as several examples of application of the related method. The hope with this approach is that students would get a sound introduction to a more methodologically-oriented presentation of key social and data science concepts.

2 Text & Materials

Required:

1. Download the *free* statistical computing program, `R`:
   `https://www.r-project.org/`
2. Download the *free* `R` complementary platform, `R Studio`:
   `https://www.rstudio.com/products/rstudio/download/`
3. Many articles, book chapters, and papers assigned throughout the semester

3 Evaluation & Assessment

All final grades are rounded to the nearest decimal (e.g., 88.38% = 88.4%). I use the following grading scheme to determine your final grade: A (93-100), A- (90-92), B+ (88-89), B (83-87), B- (80-82), C+ (78-79), C (73-77), C- (70-72), D+ (68-69), D (63-67), D- (60-62), F (0-59). I deduct a letter grade per day any assignment is late.

There are four components to students' final grades: (1) data paper (40%), (2) presentation (10%), (3) take-home data exercises (30%), and (4) participation (20%).

1. Final Data Paper (*40% of final grade*): The key assignment in this class will be a final "data" paper. The idea is for students to follow a similar format as the class, but on their own. Select a substantive concept/topic and select an appropriate method for assessing and evaluating that concept (there could be many methods for a single concept, e.g., ideology could be assessed through scaling techniques, unfolding binary choice data, text as data methods, latent variable models, or even deep learning machine algorithms). A completed assignment will be about 12-15 pages. As this is assignment is focused on proper empirical application, there need not be a lengthy theory section ending in generation of original hypotheses, though students can certainly take this approach if they wish. In addition to the application of the method in relation to the concept of interest, students will be required to submit replication materials associated with the paper all in a **single document** submitted on the course blackboard page, including: the **final paper, all `R` code, *which should be replicable with zero errors*, and any supporting/Appendix material.** Students may consult with me as much as they wish to help hone their topics and interests. Further, I will provide a great deal of guidance on methodological options for

students as lot of these topics will be very new to many. I will ensure students have been exposed to the tools and detail necessary to effectively, efficiently, and ethically pair the proper method with a substantive topic of political interest. **A few essentials**: 12-15 pages, double-spaced, 12 point font, Times New Roman/standard font, 1 inch margins, and a properly formatted reference list *with* in-text citations. You can select any reference style you wish (Chicago, APA, MLA, etc.); just be correct and consistent.

Your data paper is due on the last day of classes, Friday 4/26, by 11:59 pm.

2. Final Presentation (*10% of final grade*): Students will present a brief, conference-style 10-15 minute presentation of their projects. Assigned presentation days, shown below, are alphabetical. This will include information on the topic, methods options, the justification behind method selection, then all of the details on the application exercise. The idea is to have a "mini-conference" where students get to show case their hard work. I am considering inviting faculty members to attend presentations to offer feedback to students, or at the very least, to demonstrate support for their projects and effort. More later.

3. Data Exercises (*30% of final grade*): Throughout the semester, students will be given a total of three (3) data exercises. These will be take-home assignments, and students will be given about a week to finish each. The basic format will be a series of questions related to concepts we have discussed to this point. Students will be asked to write and run proper code to address these questions. Though the length of these assignments will likely be around 5-7 pages, there is no explicit length required. I just expect students to take sufficient space to thoroughly, yet concisely address all parts of all questions. Importantly, students are strongly encouraged to collaborate and problem solve together as this is the way most high-level and academic research is conducted. Thus, students may work in groups of any size, if they wish to do so. However, whether working in pairs, groups, or alone, **every student is required to submit their own answers and code**. *If there is any confusion around expectations and "appropriate" levels of collaboration, students are encouraged to reach out and ask me.*

   Note: A completed assignment will look like a brief response to the question, *in addition* to presentation of all code and output (e.g., plots, tables, etc.) inserted *directly below the response to the question* all in a single document.

   Extra Credit: I am willing to offer 5 additional extra credit points per data exercise to anyone who submits their completed assignment(s) using LaTeX or `R Markdown`. We will discuss how to use these at the outset of the semester. But given the value of these document processing programs for statistical and data science fields, demonstration of your proficiency using them for your assignments in this class will result in some extra credit and ultimately serve you well moving forward in these fields.

4. Participation (*20% of final grade*): Weeks 4–14 are the meat of the class. These will generally follow a similar pattern of discussion and light lecture (only to fill in some gaps) on Tuesdays, and then application and coding on Thursdays. As this is a senior seminar, we will treat the discussion portion on Tuesdays similar to what you would find in nearly all graduate political science programs. There will be a few articles to read every Tuesday, and students will sign up to read one of these a little more closely and lead the class discussion on that piece. We will pass around sign up sheets at the beginning of the prior week for discussion leading duties for the following week. For example, if we had 3 articles

to read for week 7, three students sign up to lead discussion on the articles (1 article per student) during week 6 to allow for a week to prepare. Beyond formal discussion leading duties, all students are expected to come prepared to engage and discuss, having read all required materials *prior* to class.

4 The William & Mary Honor Code

The College of William & Mary has had an honor code since at least 1779. Academic integrity is at the heart of the university, and we all are responsible for upholding the ideals of honor and integrity. The student-led honor system is responsible for resolving any suspected violations of the Honor Code, and I will report all suspected instances of academic dishonesty to the honor system. The Student Handbook (`www.wm.edu/studenthandbook`) includes your responsibilities as a student and the full Code. Your full participation and observance of the Honor Code is expected. To read the Honor Code, see `www.wm.edu/honor`.

4.1 The W&M Pledge

As a member of the William and Mary community, I pledge on my honor not to lie, cheat, or steal, either in my academic or personal life. I understand that such acts violate the Honor Code and undermine the community of trust, of which we are all stewards.

4.2 Academic Honesty

The College defines academic dishonesty in several ways, such as plagiarism, which is the form of "deliberate" or "reckless" representation of another's words, thoughts, or ideas as one's own without appropriate attribution to the original author in connection with submission of academic work, whether graded or otherwise, is a serious breach of academic integrity demanded by the Honor Code and one of the most common forms of academic misconduct processed by the honor system. Plagiarism can take many forms and there may be a number of reasons why it occurs. For example:

- Quote and cite any words that are not your own. If you paraphrase the words of another, you must still give proper attribution. If you look it up, write it down.
- Authorized vs. Unauthorized Collaboration. All academic work in this course, including homework, quizzes, and exams, is to be your own work, unless otherwise specifically provided. It is your responsibility if you have any doubt to confirm whether or not collaboration is permitted. Whenever possible, be clear and concise. Ambiguous statements often lead to confusion.

5 Student Accessibility Services

William & Mary accommodates students with disabilities in accordance with federal laws and university policy. Any student who feels s/he may need an accommodation based on the impact of a learning, psychiatric, physical, or chronic health diagnosis should contact Student Accessibility Services staff at `757-221-2509` or at `sas@wm.edu` to determine if accommodations are warranted and to obtain an official letter of accommodation. For more information, please see `www.wm.edu/sas`.

6 Outline of Topics & Calendar

**Below is a tentative outline of the semester. I reserved the right to make changes to dates, topics, readings, etc. as needed. If changes are made, students will be notified as soon as they are made to allow for adaptation.*

UNIT 1: INTRODUCTION & ORIENTATION

- **Week 1**: Course Introduction & Syllabus
  - *Thursday, 1/17*: Course Introduction & Syllabus
    - Reading:
      - Syllabus
- **Week 2**: Data Science, Social Research, & Computational Social Science
  - *Tuesday, 1/22*: Thinking like a Data Scientist
    - Reading:
      - Grimmer, Justin. 2015 "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science and Politics*, `doi:10.1017/S1049096514001784`
      - Lazer, David. 2015. "The Rise of the Social Algorithm: Does content curation by Facebook introduce ideological bias¿' *Science*, 348(6239)
      - Metzler, Katie, David Kim, Nick Allum, and Angella Denman. "Who Is Doing Computational Social Science? Trends in Big Data Research" *Sage White Paper*, (`http://repository.essex.ac.uk/17679/1/compsocsci.pdf`)
  - *Thursday, 1/24*: Refresher on Bivariate and Multiple Regression
    - Reading:
      - Kellstedt and Whitten, chs. 8, 9, and 10 (on Blackboard)
- **Week 3**: Crash course in `R` and LaTeX
  - *Tuesday, 1/29*: Crash course in `R`
    - Reading:
      - Leeper, "Really Introductory Introduction to R"
      - Waggoner, "Introduction to R"
  - *Thursday, 1/31*: Crash course in `R` cont'd, and LaTeX (Briefly)
    - Reading:
      - Waggoner, "A Soft Introduction to the Language of LaTeX"

UNIT 2: MEASUREMENT & CLASSIFICATION

- **Week 4**: Clustering Algorithms & National Elections
  - *Tuesday, 2/5*: Hierarchical clustering, k-Means, Gaussian mixture models (if time)
    - Reading:
      - Waggoner, Philip. "Unsupervised Machine Learning for Clustering in Political & Social Research." Book in process with *Cambridge UP*
      - Filho et al. 2014. "Cluster Analysis for Political Scientists." *Applied Mathematics* (*Recommended* – no discussion)
      - Hastie et al. 2009. ch. 14.3, "Clustering" (*Recommended* – no discussion)
  - *Thursday, 2/7*: Clustering and the 2012 Presidential Election in `R`
- **Week 5**: Unfolding & roll call voting in Congress
  - *Tuesday, 2/12*: Optimal classification and NOMINATE algorithms
    - Reading:
      - Keith T. Poole. 2005. Spatial Models of Parliamentary Voting. Chs. 1–2
      - Simon Hix, Abdul Noury, and Gerard Roland. 2006. "Dimensions of Politics in the European Parliament," American Journal of Political Science 50(2)
  - *Thursday, 2/14*: OC and NOMINATE and 108th House voting in `R`
- **Week 6**: Unfolding (pt. 2) & Supreme Court ideology
  - *Tuesday, 2/19*: Bayesian Item Response Theory & Markov-Chain Monte Carlo (MCMC) methods
    - Reading:
      - Andrew D. Martin and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999," Political Analysis 10: 134-153
      - Simon Jackman and Shawn Treier. 2008. "Democracy as a Latent Variable," American Journal of Political Science 52(1): 201–217
      - Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data," American Political Science Review 98: 355–70
  - *Thursday, 2/21*: IRT & SCOTUS Voting in `R`
    - **Handout Homework 1**

- **Week 7**: Natural language processing & political speech
  - *Tuesday, 2/26*: Intro to NLP for content and descriptive analysis
    - Reading:
      - Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts" Political Analysis 1–31
      - Daniel J. Hopkins and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science," American Journal of Political Science 54(1), 229–247
      - Jonathan B. Slapin and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time Series Policy Positions from Texts," American Journal of Political Science 52(3), 705–722
  - *Thursday, 2/28*: NLP & Trump speeches and Party platforms in `R`
    - **Homework 1 Due In Class**
- **Week 8**: **NO CLASS - Spring Break, March 2–10**

UNIT 3: ANALYSIS AND INFERENCE

- **Week 9**: Regression Discontinuity Design & Senate elections
  - *Tuesday, 3/12*: RDD in multiple contexts
    - Reading:
      - de la Cuesta and Imai, 2016. "Misunderstandings About the Regression Discontinuity Design in the Study of Close Elections," Annual Review of Political Science, 19:375–96
      - Eggers et al. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races," American Journal of Political Science
      - Andrew Hall. 2015. "What Happens When Extremists Win Primaries?" APSR, 109(1)
  - *Thursday, 3/14*: Using RDD to analyze close Senate elections, 1914–2010 in `R`
- **Week 10**: Big Data & Racial bias
  - *Tuesday, 3/19*: Using "Big Data" to study racial bias and policing
    - Reading:
      - TBD
  - *Thursday, 3/21*: ***Guest Lecture and Demo: Kelsey Shoub, UVA***

- **Week 11**: Binary response models & Supreme Court decision making
  - *Tuesday, 3/26*: Logit and probit models and binary outcomes
    - Reading:
      - Scott Long, ch. 1 and 3
      - Esarey, Justin and Andrew Pierce. 2012. "Assessing Fit Quality and Testing for Misspecification in Binary-Dependent Variable Models." Political Analysis 20(4): 480–500
      - Herron, Michael C. 2000. "Postestimation Uncertainty in Limited Dependent Variable Models." Political Analysis 8(1): 83–98
  - *Thursday, 3/28*: Analyzing SCOTUS decisions using Logit and Probit in `R`
    - **Handout Homework 2**

- **Week 12**: Event count models & judicial review
  - *Tuesday, 4/2*:
    - Reading:
      - Jeremy Waldron, 2006. "The Core of the Case against Judicial Review," 115 Yale Law Journal, 1346
      - King, Gary. 1988. "Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for Exponential Poisson Regression Model." American Journal of Political Science 32(3): 838–863
  - *Thursday, 4/4*: Analyzing judicial review using event count models in `R`
    - **Homework 2 Due In Class**

- **Week 13**: Ordered response models & political preferences
  - *Tuesday, 4/9*: Ordered logit and probit models to study ordinal scales
    - Reading:
      - Jones, Bradford S. and Michael E. Sobel. 2000."Modeling Direction and Intensity in Semantically Balanced Ordinal Scales: An Assessment of Congressional Incumbent Approval." American Journal of Political Science 44(1):174-185
      - Franklin, Charles H. and Liane C. Kosaki. 1989. "Republican Schoolmaster: The U.S. Supreme Court, Public Opinion, and Abortion." American Political Science Review 83(3): 751-771
      - Espenshade, Thomas J. and Haishan Fu. 1997. "An Analysis of English Language Proficiency among U.S. Immigrants." American Sociological Review 62(2): 288-305
  - *Thursday, 4/11*: Ordered response models & preferences on immigrants' access to healthcare in `R`
    - **Handout Homework 3**

- **Week 14**: Duration analysis & experimental criminal recidivism
  - *Tuesday, 4/16*: Focus on substantive demonstration in `R`
    - Reading:
      - Jonathan Katz and Brian Sala. 1996. "Careerism, Committee Assignments, and the Electoral Connection." APSR, 90(1) (*Recommended* – no discussion)
      - McCarty, Nolan and Rose Razaghian. 1999. "Advice and Consent: Senate Responses to Executive Branch Nominations." American Journal of Political Science 43(4) (*Recommended* – no discussion)
  - *Thursday, 4/18*: **No Class: Writing Day**
- **Week 15**: Presentations and Final Week
  - *Tuesday, 4/23*: Student Presentations/Mini-Conference (Chand – Pyle)
    - **Homework 3 Due In Class**
  - *Thursday, 4/25*: Student Presentations/Mini-Conference (Saylor – Yang)
  - **Friday, 4/26: Final data papers due on Blackboard by 11:59 pm.**