MACS 40500: Computational Methods for American Politics

Dr. Philip Waggoner
MACSS
Office: 205 McGiffert House
https://pdwaggoner.github.io/
pdwaggoner@uchicago.edu
Office Hours: Wednesdays 10–12, or by email appt.

Location: Social Sciences Rsch Bldg 401 **Days & Time**: Tuesday / Thursday, 3:30-4:50 PM

1 Overview & Introductory Remarks

Welcome to Computational Methods for American Politics! In this class, we will be walking through a host of statistical and computational techniques aimed at exploring and better understanding substantive political phenomena in the American context. We will link substantive political and 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 primarily. There are two interrelated goals of the class: 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.

2 Course Objectives

By the end of the course, students should:

- Be aware of computational methods for explaining substantive American political concepts.
- Develop a toolbox of different algorithms and approaches to computational and statistical analysis.
- Understand when and where it makes sense to apply different computational tools for solving problems in American politics research.
- Understand how to implement, diagnose, and troubleshoot different models and methods in the `R` programming language.
- Understand how to generate reproducible research reports using `R Markdown`.
- Complete a final project that reflects all aspects of the course, from theoretical perspectives of methods to fitting and diganosing models aimed at addressing questions of interest in American politics.

Prerequisites: Though not formally required, it is highly recommended that students have some level of experience with computational methods and/or applied statistics. Such experience could come from, e.g., MACS 30100 (Perspectives on Computational Modeling), MACS 30000 (Perspectives on Computational Analysis), or other training in statistical learning. Further, some level of programming experience is expected (e.g. CAPP 30122 (Computer Science with Applications - 2) or MACS 30500 (Computing for the Social Science)). Experience with `R`

recommended, but not required.[1] I will provide "crash courses" covering the basics of `R` as well as basic statistical concepts at the outset of the quarter. If a student is unsure of their likelihood of success in the course, they should reach out to me directly to discuss and make a case. Such course petitions will be handled on a case-by-case basis.

3 Course Structure

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 unfolding techniques (optimal classification, IRT, NOMINATE, etc.). The method (unfolding) and the substantive topic (ideology) will reinforce each 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 `R`. Our goal is to learn how to think about complex political questions from a computational 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 and apply the appropriate method to address the question of interest.

Most weeks we will proceed as follows: Tuesdays will be both lecture and seminar-style, dependent on the week's readings. Then, Thursdays will typically be reserved for application, where we will *implement* the week's method using a political dataset that I will provide. The hope with this approach is that students would get a sound introduction to a more methodologically-oriented presentation of key concepts.

4 Text & Materials

Required:

1. If you haven't already, sign up for a free GitHub account at: `https://github.com`. Then, join our class organization, *macss-cmap19*. Most interaction outside of class (questions, submissions, etc.) will happen via GitHub. Submission via GitHub is described below in the "Evaluation & Assessment" section and in Figure 1.

2. Download the *free* statistical computing program, `R`:
   `https://www.r-project.org/`[2]

3. Download the *free* `R` IDE, `R Studio`:
   `https://www.rstudio.com/products/rstudio/download/`

4. Articles, book chapters, and papers assigned throughout the quarter

Recommended:

1. Quan Li. 2018. Using R for Data Analysis in Social Sciences. Oxford UP.

2. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Free PDF version:
   `https://web.stanford.edu/ hastie/ElemStatLearn/`

3. Hadley Wickham and Garrett Grolemund. 2017. R for Data Science. Free `bookdown` version: `https://r4ds.had.co.nz/`

[1] Note that I primarily teach and research in `R`. While this is the case, if students are more comfortable with other languages such as Python or Julia, you are welcome to use them in the course.

[2] Or ensure that your preferred program/IDE is downloaded on your computer, e.g., Python, Anaconda, etc.

5 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 (15%), (3) problem sets (30%), and (4) participation (15%).

Submitting Assignments via GitHub: To submit assignments (problem sets and final data papers), you will need to follow our GitHub workflow described in Figure 1. In brief, fork the repository for the corresponding assignment (e.g., problem-set-1). Upon the new repo showing up in your own GitHub profile, clone the repository to a GitHub client (e.g., RStudio, Visual Code Studio, etc.). Do your homework, respond to questions, create your `.md` file, etc. and then commit your changes once you are finished with the assignment. Then, push the changes back to your repo. Finally, open a new pull request with your assignment submission. Once merged, your assignment is considered "submitted" and you're finished. Please reach out to me, the TA, or anyone if you have questions or issues with this process.

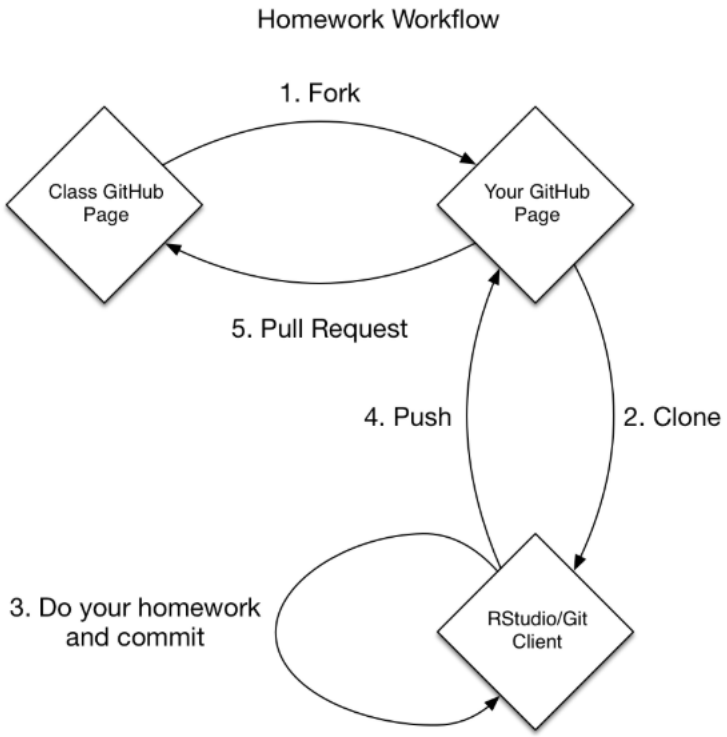

Figure 1: Assignment Submission Workflow (courtesy of Dr. Soltoff)

1. Final Data Paper (*40% of final grade*): The key assignment in this class will be a final data paper. The idea is to follow a similar format as the class, but on your own. Select a substantive concept or topic and then 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 multidimensional scaling, unfolding, natural language processing, latent variable models, or even deep learning 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** as either R markdown or a Jupyter Notebook, 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 new to most students. I will ensure you 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, 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 final paper is due on Thursday, December 12.

2. Final Presentation (*15% of final grade*): During the final week (two class periods) at the end of the quarter, students will present their final projects to the class. Though I do not have a specific set of guidelines for what your presentation should include, good presentations typically include some combination of the following:

   - Introduce the topic to a general audience with a motivating anecdote, paradox, or example
   - Summarize your process, workflow, and your main approach or method
   - Highlight the key findings
   - "Appendix" set of slides with more detail at the end in case you are questioned on specific "in-the-weeds"-type details

   The presentation should be around 6-8 minutes long, plus 2 minutes for questions. Presentation order will be determined and announced later in the quarter.

3. Problem Sets (*30% of final grade*): Throughout the quarter, students will be given 3 problem sets related to a given week's material (3 sets equally weighted at 10% each, totaling 30% of your final grade). Some problem sets will be more conceptual requiring long answers, others will be more computationally intensive, while still others may be a blend of these types. Typically, problem sets will be handed out at the beginning of a given week on Monday, and then due on Friday of the same week by 5 pm, even though we are on a Tues/Thurs schedule. The goal here is two-fold: first to give you sufficient time to complete the problem set, and second, to encourage you to be thinking about the given week's topic all week, whether in or out of class.

   Note: A completed submission will be a single rendered PDF file (from the .Rmd or Jupyter Notebook), and will include: a response to the question, presentation of all equations, code, and output (e.g., plots, tables, etc.) inserted *directly below the response to the question* (i.e., no appendices necessary for problem sets; just write and run the code, and produce the output in-line). **A single file is all that is required for submission.**

4. Participation (*15% of final grade*): As this is a largely seminar-style graduate course, whether we are discussing an article, working through code as a class, or engaging in a workshop, I expect all students to come to every class prepared to engage and discuss, having read all required materials *prior* to class and to be respectful of their peers in class. Respect includes the wide variance in opinions that is sure to exist. I don't anticipate problems, but need to underscore these expectations at the outset to ensure everyone is on the same page.

6 Diversity & Inclusion

The University of Chicago is committed to diversity and rigorous inquiry from multiple perspectives. The MAPSS, CIR, and MACSS programs share this commitment and seek to foster productive learning environments based upon inclusion, open communication, and mutual respect for a diverse range of identities, experiences, and positions.

Any suggestions for how we might further such objectives both in and outside the classroom are appreciated and will be given serious consideration. Please share your suggestions or concerns with your instructor, your preceptor, or your program's Diversity and Inclusion representatives: Darcy Heuring (MAPSS), Matthias Staisch (CIR), and Chad Cyrenne (MACSS). You are also welcome and encouraged to contact the Faculty Director of your program.

This course is open to all students who meet the academic requirements for participation. Any student who has a documented need for accommodation should contact Student Disability Services (773-702-6000 or `disabilities@uchicago.edu`) and the instructor as soon as possible.

7 Course Schedule

**Below is a tentative outline of the quarter. Any changes made will be announced with sufficient time for adequate adaptation.*

- **Week 1**: Course introduction and syllabus
  - *Tuesday, 10/1*: Syllabus; Computational Social Science and Political Research
    - **Required Reading**
      1. Syllabus
  - *Thursday, 10/3*: Theoretical Foundations
    - **Required Reading**
      1. Grimmer, Justin. 2015, "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS*
      2. King, Gary. 1995, "Replication, Replication." *PS*
      3. Lazer, David. 2015, "The rise of the social algorithm." *Science*
      4. Lazer et al. 2009, "Life in the network: the coming age of computational social science." *Science*
      5. Wallach, Hanna. 2018. "Computational social science $\neq$ computer science + social data." *Communications of the ACM*, 61(3): 42-44
      6. McKiernan et al, "How open science helps researchers succeed" (**skim this one**)
    - **Recommended Reading**
      1. Scott de Marchi 2005, "Computational and mathematical modeling in the social sciences" (book)
      2. Ram 2013, "Git can facilitate greater reproducibility and increased transparency in science"
      3. Chang et al. 2014, "Understanding the paradigm shift to computational social science in the presence of big data"
      4. Donoho 2010, "An invitation to reproducible computational research"
      5. Giles 2012, "Making the Links"
      6. Sandve et al. 2013, "Ten Simple Rules for Reproducible Computational Research"
- **Week 2**: Computational foundations
  - *Tuesday, 10/8*: Crash course on `R` and `R Markdown`
    - **Required Reading**
      1. None
    - **Recommended Reading**
      1. Xie, Allaire, and Grolemund 2019, "https://bookdown.org/yihui/rmarkdown/" (book for R markdown)
      2. Wickham and Grolemund 2017, "R for Data Science": `https://r4ds.had.co.nz/` (book for Tidyverse)
      3. Waggoner 2019, "Introduction to Computation in R": `https://github.com/pdwaggoner/Intro-to-R` (course materials and scripts from a past class I taught)

- *Thursday, 10/10*: Data Munging and Wrangling in `R`
  - **Recommended Reading**
    1. Wickham and Grolemund 2017, "R for Data Science": `https://r4ds.had.co.nz/wrangle-intro.html` (chs. 9-16)
    2. Ismay and Kim 2019, "Modern Dive": `https://moderndive.com/4-wrangling.html` (ch. 4)

- **Week 3**: Regression
  - *Tuesday, 10/15*: Ordinary Least Squares & Maximum Likelihood Estimation refresher; Data for papers
    - **Required Reading**
      1. None
    - **Recommended Reading**
      1. Long, J. Scott. 1997. "Regression Models for Categorical and Limited Dependent Variables." (chs. 1-3)
      2. Aldrich, John. 1997. "R. A. Fisher and the Making of Maximum Likelihood 1912-1922." Statistical Science 12(3): 162-176
      3. Fisher, Ronald A. 1925. "Theory of Statistical Estimation." Proceedings of the Cambridge Philosophical Society 22: 700-725
      4. Neyman, Jerze. 1934. "On the Two Different Aspects of the Representative Methods: The Method of Stratified Sampling and the Method of Purposive Selection." Journal of the Royal Statistical Society 97(4): 558-625
  - *Thursday, 10/17*: Penalized Regression (LASSO, Ridge, and Elastic-net); application in `R`
    - **Required Reading**
      1. None
    - **Recommended Reading**
      1. ESL, ch. 3.4

- **Week 4**: Cluster analysis and state legislative professionalism
  - **MONDAY: handout HW1 (due Friday, 10/25, by 5 pm)**
  - *Tuesday, 10/22*: Clustering Basics & Hierarchical Agglomerative Clustering
    - **Required Reading**
      1. Johnson, S. C. (1967). "Hierarchical clustering schemes." Psychometrika, 32(3), 241-254.
    - **Recommended Reading**
      1. Grimmer, J., and King, G. (2011). "General purpose computer-assisted clustering and conceptualization." Proceedings of the National Academy of Sciences, 108(7), 2643-2650.
      2. Ahlquist, J. S., and Breunig, C. (2012). "Model-based clustering and typologies in the social sciences." Political Analysis, 20(1), 92-112.
      3. Luo, M., Wang, L. N., and Zhang, H. G. (2003). "An unsupervised clustering-based intrusion detection method." Acta Electronica Sinica, 31(11), 1713-1716.
      4. Hastie et al. 2009. ch. 14.3, "Clustering"
      5. Fung 2001, "A Comprehensive Overview of Basic Clustering Algorithms": `https://sites.cs.ucsb.edu/ veronika/MAE/clustering_overview_2001.pdf`

- *Thursday, 10/24*: k-means, Gaussian mixture models, application in `R`
- **Required Reading**
  1. Hartigan, J. A., and Wong, M. A. (1979). "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.
  2. Reynolds, D. (2015). "Gaussian mixture models." Encyclopedia of biometrics, 827-832.
  3. Squire. 2007. "Measuring State Legislative Professionalism: The Squire Index Revisited." SPPQ, 7(2)
  4. Bowen and Greene. 2014. "Should We Measure Professionalism with an Index? A Note on Theory and Practice in State Legislative Professionalism Research." SPPQ, 14(3)

- **Week 5**: 2-Class classification problems and Supreme Court decision-making
  - *Tuesday, 10/29*: 2-Class Classification Problems: Logistic Regression and Linear Discriminant Analysis
    - **Required Reading**
      1. Xitao Fan and Lin Wang. 1999. "Comparing Linear Discriminant Function with Logistic Regression for the Two-Group Classification Problem, *Journal of Experimental Education,* 67(3): 265-286
    - **Recommended Reading**
      1. Esarey, Justin and Andrew Pierce. 2012. "Assessing Fit Quality and Testing for Misspecification in Binary-Dependent Variable Models." Political Analysis, 20(4): 480-500.
      2. Herron, Michael C. 2000. "Postestimation Uncertainty in Limited Dependent Variable Models." Political Analysis 8(1): 83-98.
  - *Thursday, 10/31*: Application: Classifying U.S. Supreme Court decisions

- **Week 6**: Unfolding binary choice data and roll call voting
  - **MONDAY: handout HW2 (due Friday, 11/8, by 5 pm)**
  - *Tuesday, 11/5*: Geometry and space scaling in political decision making
    - **Required Reading**
      1. Keith T. Poole. 2005. Spatial Models of Parliamentary Voting. (chs. 1-2)
      2. Simon Hix, Abdul Noury, and Gerard Roland. 2006. "Dimensions of Politics in the European Parliament," American Journal of Political Science 50(2)
    - **Recommended Reading**
      1. Platt, G., Poole, K. T., and Rosenthal, H. (1992). "Directional and Euclidean theories of voting behavior: a legislative comparison." Legislative Studies Quarterly, 561-572.
      2. Carroll, R., Lewis, J. B., Lo, J., Poole, K. T., and Rosenthal, H. (2009). "Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap." Political Analysis, 17(3), 261-275.
      3. Carroll, R., Lewis, J. B., Lo, J., Poole, K. T., and Rosenthal, H. (2009). "Comparing NOMINATE and IDEAL: Points of difference and Monte Carlo tests." Legislative Studies Quarterly, 34(4), 555-591.
      4. Clinton, J. D., and Jackman, S. (2009). "To simulate or NOMINATE?" Legislative Studies Quarterly, 34(4), 593-621.

5. Rosenthal, H., and Voeten, E. (2004). "Analyzing roll calls with perfect spatial voting: France 1946-1958." American Journal of Political Science, 48(3), 620-632.

- *Thursday, 11/7*: Application: NOMINATE algorithm

- **Week 7**: Latent space analysis and Supreme Court ideology
  - *Tuesday, 11/12*: IRT models and latent space analysis
    - **Required Reading**
      1. 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
      2. Simon Jackman and Shawn Treier. 2008. "Democracy as a Latent Variable," American Journal of Political Science 52(1): 201–217
      3. Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data," American Political Science Review 98: 355–70
    - **Recommended Reading**
      1. Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. Political Analysis, 11(2), 139-163.
      2. Linzer, D. A., and Staton, J. K. (2015). A global measure of judicial independence, 1948-2012. Journal of Law and Courts, 3(2), 223-256.
      3. McGann, A. J. (2014). Estimating the political center from aggregate data: an item response theory alternative to the Stimson dyad ratios algorithm. Political Analysis, 22(1), 115-129.
      4. Holman, R., and Glas, C. A. (2005). "Modelling non-ignorable missing-data mechanisms with item response theory models." British Journal of Mathematical and Statistical Psychology, 58(1), 1-17.
      5. Pemstein, D., Tzelgov, E., and Wang, Y. T. (2015). "Evaluating and improving item response theory models for cross-national expert surveys." V-Dem Working Paper, 1.
  - *Thursday, 11/14*: Application: Bayesian IRT and MCMC simulations

- **Week 8**: Natural language processing & text mining
  - **MONDAY: handout HW3 (due *next Wednesday*, 11/27, by 5 pm)**
  - *Tuesday, 11/19*: Sentiment analysis, dictionary methods, & finding discriminating words
  - *Thursday, 11/21*: Topic models (LDA/Gibbs and Structural(?))
    - **Required Reading**
      1. van Kessel. 2018. "An intro to topic models for text analysis", Pew: `https://medium.com/pew-research-center-decoded/an-intro-to-topic-models-for-text-analysis-de5aa3e72bdb`

- **Week 9**: Natural language processing & text models
  - *Tuesday, 11/26*: Advances in quantitative text analysis; ideal point estimation from text
    - **Required Reading**
      1. 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
      2. 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
      3. Klemmensen, R., Hobolt, S. B., and Hansen, M. E. (2007). "Estimating policy positions using political texts: An evaluation of the Wordscores approach." Electoral Studies, 26(4), 746-755.
    - **Recommended Reading**
      1. Laver, M., Benoit, K., and Garry, J. (2003). "Extracting policy positions from political texts using words as data." American political science review, 97(2), 311-331.
      2. 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
      3. Benoit, K., Bräuninger, T., and Debus, M. (2009). "Challenges for estimating policy preferences: Announcing an open access archive of political documents." German Politics, 18(3), 441-454.
  - **WEDNESDAY, 11/27: HW3 due by 5 pm**
  - *Thursday, 11/28*: **No class: Thanksgiving Break**

- **Week 10**: Presentations
  - *Tuesday, 12/3*: Presentations, Day 1
  - *Thursday, 12/5*: Presentations, Day 2
    - **WEDNESDAY, 12/11: Final data papers due by midnight**