Data Analytics & Security

INTA 6450

Instructor Info

David Muchlinski

M-W 1:00-3:00 via Bluejeans

Habersham 147

www.davidmuchlinski.com

david.muchlinski@inta.gatech.edu

Course Info

Preq: A course on regression modeling covering AT LEAST up to OLS strongly encouraged.

Hybrid

MW 11:00-12:15

Online with Touchpoints

About

Computational social science is the fastest developing field of social science. Methodologically, it encompasses the use of semi-parametric and non-parametric algorithmic statistical models for the purposes of prediction. It has also pioneered the usage of novel sources of data including text and images that cannot be analyzed by traditional regression models. It is still an emerging field, with unsettled theoretical debates. We will approach these issues critically and thoughtfully as we become introduced to this growing field of study.

Overview

This course serves as an introduction to the principles of the field of computational social science with application to the study of international relations and comparative politics. We will discover how to apply both supervised and unsupervised statistical learning methods to perennial questions in the field including conflict, development, and democracy. Using `caret` and other Libraries in `R`, we will learn to develop, interpret, visualize, and criticize widely-used supervised methods including penalized regressions, decision trees, support vector machines, and neural networks. We will also learn how to analyze text as data through unsupervised learning methods including clustering, topic modeling, and sentiment analysis. Students will become familiar with the main theoretical and methodological debates in the field, and will become proficient in the use of various statistical learning methods for the analysis of structured and unstructured data.

Material

The following books are required reading for this course. Students with multiple semesters of experience with regression modeling, including OLS and MLE, should challenge themselves with the recommended texts. The James et al. text will be the main workhorse text for the class. The Silge and Robinson, Wickham and Grolemund, and Xie et al texts serve as reference guides to the `Tidyverse`, `Tidytext`, and `R Markdown` Libraries of which we will make extensive use.

Required Texts

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. New York: Springer. Available online at `http://faculty.marshall.usc.edu/gareth-james/ISL/`

Silge, J., Robinson, D. (2017). Text Mining with R: A Tidy Approach. O'Reilly Media, Inc. Available online at `https://www.tidytextmining.com/`

Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc. Available online at `https://r4ds.had.co.nz/`

Xie, Y., Allaire, J. J., Grolemund, G. (2018). R Markdown: The Definitive Guide. CRC Press. Available Online at `https://bookdown.org/yihui/rmarkdown/`

Recommended Additional Texts

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer. Available online at `https://web.stanford.edu/~hastie/ElemStatLearn/`

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT press.

Murphy, K. P. (2012). Machine Learning: a Probabilistic Perspective. MIT press.

Other

Any required reading not explicitly covered in Witten et al, (2013), including articles and book chapters will be provided on Canvas.

Class DataCamp Site

There is a class DataCamp site which students may utilize for free to improve their skills with certain applications in `R`. I have sent each student an invitation which can be used to access videos reviewing important topics. Each short course is about 4 hours in length, making each possible to accomplish within a week. If students encounter difficulties with `R`, I strongly recommend they utilize this resource.

FAQs

How will communication be handled?

Important announcements will be posted to Canvas. You may reach me via email using Canvas. Office hours will be held virtually via Bluejeans from 1-3 on Tuesdays and Thursdays. Feel free to drop in during those hours or make an appointment via email. I will respond to student email questions and requests from 10:00-12:00 M-F.

What does Hybrid Mean?

This is a hybrid class; most of our interaction will be online to ensure everyone's safety. Attendance will be recorded, and all learning will be synchronous. All lectures, lecture slides, R Lab documents, and other material are accessible via the course's Canvas site. Please keep in mind that this is an advanced course, especially for social science undergraduates, so attending lectures will provide the greatest benefit. We will be meeting in person periodically throughout the semester for days indicated on the class schedule. These in-person meetings are essential to facilitate learning, so please make them if at all possible.

What if we go Remote?

There are contingency plans for all courses, including this course, to go fully remote if needed. I will release those plans to Canvas, and to you all via email, if and when needed.

Grading Scheme

15 (30)%	Homework (3 assignments)
25 (15)%	Midterm Exam
0 (20)%	Quizzes and Assessments
30 (20)%	Cumulative Final Exam
30 (15)%	Research Project

Grades will follow the standard scale: A = 89.5-100; B = 79.5-89.4; C = 69.5-79.4; D = 60-69.4; F <60. Amounts in parentheses indicate undergraduate grade weighting. Curving will be applied at the discretion of the professor. Rounding will be applied only if students have completed all homework assignments with at least an 80% average, and only for undergraduate students. Rounding will be applied only for 1% point towards the higher grade.

Course Structure

This course is divided, roughly, into three parts.

- The theory of machine learning as a statistical field and how the fundamentals of machine learning differ from the fundamental assumptions of the more familiar parametric models. Students in this part will be introduced to what machine learning is as well as its relationship to and purpose in the field of quantitative (i.e. statistical) social science. Readings will focus on "big picture" topics including how machine learning is used in political science, how it differs from parametric regression based research designs, how machine learning is used to understand social processes, and why predictive accuracy is a valid criterion for model validation.

- The mechanics of supervised learning. This section of the course will introduce students to the technical aspects of supervised machine learning analytics in the R programming language including: cross-validation, out-of-sample prediction, model selection and validation, as well as methods of data analysis including penalized regression models, tree-based models, support vector machines, and neural networks. Students will make extensive use of the `caret` (Classification And REgression Training) library, one of the most powerful machine learning libraries in R.

- Unsupervised learning. This section of the course will introduce topics including clustering, topic modeling, text as data, and sentiment analysis. Students will also be introduced to the Kreas Library to construct deep learning neural networks in R.

Learning Objectives

- Students will demonstrate methodological literacy to analyze international political phenomena
- Students will demonstrate facility with the research process from problem identification to analysis of findings.
- Students will use written communication to demonstrate knowledge and to make cogent arguments in international affairs.
- Students will demonstrate knowledge to address theoretical and applied issues in international affairs.

Hybrid Lectures

Lectures will convene synchronously with the course meeting date and time as specified in the syllabus and OSCAR. Lectures will be recorded at point of delivery. All lecture materials, including slides, will be posted before each lecture so students may follow along. The first lecture will cover the required reading for each week, the second lecture will be the R Lab for the week and will consist of the professor guiding students through the implementation of that week's topics using the software.

Research Project

Students will complete a semester-long research project culminating in a research note ($\sim$ 4k-6k words for graduate students, $\sim$ 3-5k words for undergraduates) on a topic of their choosing. Students must submit to me by week 4 a detailed prospectus detailing their initial research topic. Students may collect their own data or utilize existing data in a novel way. As collaboration is extremely common in this field, students may form co-author teams of between 2-3 individuals to co-author this research paper.

Midterm and Final Exam

Students will be assigned one midterm and one final examination which they will have two each weeks to complete. These exams will be conducted outside of class on a "take home" basis and will involve conceptual and theoretical as well as applied sections. Students will provide answers to the applied sections using R Markdown, and can submit answers to other sections in other formats including Word or LaTeX. More detailed instructions will follow at the appropriate time. These exams are to be completed individually.

Assessment

There are three homework assignments that will be assigned. Each will be due on the conclusion of each section of the course as described above. Students are encouraged to work continuously on all homework assignments as they are challenging. All homework assignments are to be submitted on time in .html format using R Markdown. Late homework will be penalized at a rate of one letter grade per day late for undergraduates. Graduate students submitting late work will receive no credit. All homework assignments must be completed individually, but collaboration for the purposes of problem solving is encouraged.

There will be a series of graded quizzes due throughout the semester to evaluate student learning for undergraduate students only. There will be five quizzes, and I will drop the lowest quiz grade for each undergraduate. Our touchpoint meetings will be spent discussing and reviewing these quizzes, so please come to these meetings with questions pre-prepared.

Make-up Policy

Make up exams will be allowed if a student becomes ill due to COVID. No documentation will be necessary. However, the version of the make up exam will be different from the official exam previously distributed.

Reading

All students are responsible for all readings marked as Required. These readings must be completed prior to each course meeting. Graduate students are required to read all Supplementary readings in addition to all Required reading. Recommended readings may be read at one's discretion. These are foundations readings if one wishes to become more knowledgeable about the broader literature and may prepare Ph.D. students well for comprehensive exams.

Diversity and Inclusivity Statement

The Institute does not discriminate against individuals on the basis of race, color, religion, sex, national origin, age, disability, sexual orientation, gender identity, or veteran status in the administration of admissions policies, educational policies, employment policies, or any other Institute governed programs and activities. The Institute's equal opportunity and non-discrimination policy applies to every member of the Institute community. The Institute's affirmative action program, Title IX program, and related policies are developed in compliance with applicable law. Pursuant to Title IX, the Institute does not discriminate on the basis of sex in its education programs and activities. As such, the Institute does not tolerate any kind of gender-based discrimination or harassment, which includes sexual violence, sexual harassment, and gender-based harassment. Inquiries concerning the Institute's application of or compliance with Title IX may be directed to the Title IX Coordinator, Burns Newsome, burnsnewsome@gatech.edu, 404-385-5151. Additionally, inquiries concerning the application of applicable federal laws, statutes, and regulations (such as Title VI, Title IX, and Section 504) may be directed to the U.S. Department of Education's Office of Civil Rights at www2.ed.gov/ocr.

Accommodations for Students with Disabilities

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students must register with the Office of Disability Services at Suite 123, Smithgall Student Services Building, 353 Ferst Drive, 404-894-2563 (Voice); 404-894-1664 (TDD). For more information on Georgia Tech's policy on working with students with disabilities, please see review the Office of Disability Service's web page at https://policies.ncsu.edu/regulation/reg-02-20-01/. The Office of Disability Services collaborates with students, faculty, and staff to create a campus environment that is usable, equitable, sustainable and inclusive of all members of the Georgia Tech community. Disability as an aspect of diversity that is integral to society and Georgia Tech. If students encounter academic, physical, technological, or other barriers on campus, the Disability Services team is available to collaboratively find creative solutions and implement reasonable accommodations.

Academic Integrity

Academic dishonesty in the form of cheating or plagiarism will not be tolerated. In brief, plagiarism is defined, for the purposes of this class, as: copying, borrowing, or appropriating another person's work and presenting it as your own in a paper or oral presentation, deliberately or by accident. Acts of plagiarism will be reported in accordance with the Honor Code. In order to avoid being charged with plagiarism, if you use the words, ideas, phrasing, charts, graphs, or data of another person or from published material, then you must either: 1) use quotation marks around the words and cite the source, or 2) paraphrase or summarize acceptably using your own words and cite the source. The plagiarism policy is not restricted to books, but also applies to video and audio content, websites, blogs, wiki's, and podcasts. Plagiarism includes putting your name on a group project to which you have minimally contributed. For information on Georgia Tech's Academic Honor Code, please visit http://www.catalog.gatech.edu/policies/honor-code/ or http://www.catalog.gatech.edu/rules/18/. Any student suspected of cheating or plagiarizing on a assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations. The student will also receive a grade of zero on the assignment at the professor's discretion.

Class Schedule

SECTION 1: Review and Introduction to Computational Social Science

Week 1: M Intro to Comp. Social Science

REQUIRED READING

James et al. Chs. 1-2

SUPPLEMENTARY READING

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Jebara, T. (2009). Computational Social Science. *Science*, 323(5915), 721-723.

Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, 43, 19-39.

RECOMMENDED READING

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, 40, 129-152.

Grimmer, J. (2015). We are all Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, 48(1), 80-83.

Murphy (2012) Ch. 1

Hastie et al. (2009) Ch. 2

Week 1: W Review of Parametric Models

REQUIRED READING

James et al. Ch. 3

SUPPLEMENTARY READING

Freedman, D. (1989). Statistical Models and Shoe Leather. *Mathematical Social Sciences*, 18(2), 192-192.

RECOMMENDED READING

Nagler, J. (1995). Coding Style and Good Computing Practices. *PS: Political Science & Politics*, 28(3), 488-492.

Salganik, M. (2019). Bit by bit: Social Research in the Digital Age. Princeton University Press. Chapter 6.

Skills Development: Introduction to `Caret`, `Tidyverse`, `R Markdown`, and LaTeX.

DataCamp Review Courses: Data Manipulation with R, Importing and Cleaning Data with R, Data Visualization with R, R Programming

Week 2: M Categorical and Limited Dependent Variables I

Class Meeting in Person: Welcome & Intro to Computational Social Science

REQUIRED READING

James et al. Ch. 4

SUPPLEMENTARY READING

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 137-163.

RECOMMENDED READING

King, G. (1998). Unifying Political Methodology: The Likelihood Theory of Statistical Inference. University of Michigan Press. Ch. 5

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables (Vol. 7). Advanced Quantitative Techniques in the Social Sciences. Chs. 3 & 5

Week 2: W Categorical and Limited Dependent Variables II

REQUIRED READING

Collier, P., & Hoeffler, A. (2004). Greed and Grievance in Civil War. *Oxford Economic Papers*, 56(4), 563-595.

SUPPLEMENTARY READING

Menard, S. (1995). Applied Logistic Regression Analysis, Quantitative Applications in the Social Sciences, 106. London: Sage. Chs. 1-3.

Skills Development: Generalized Linear Models (`glm`) in `R`. Model validation and interpretation.

DataCamp Review Courses: Data Manipulation with R, Data Visualization with R, Statistics Fundamentals with R

Quiz 1 Distributed

Week 3: M Philosophy of Statistical Learning I

REQUIRED READING

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199-231.

SUPPLEMENTARY READING

Shmueli, G. (2010). To Explain or to Predict?. *Statistical Science*, 25(3), 289-310.

Titiunik, R. (2015). Can Big Data Solve the Fundamental Problem of Causal Inference?. *PS: Political Science & Politics*, 48(1), 75-79.

RECOMMENDED READING

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons from Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100-1122.

Tu, Y. K., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's Paradox, Lord's Paradox, and Suppression Effects are the Same Phenomenon–the Reversal Paradox. *Emerging Themes in Epidemiology*, 5(1), 2.

Gelman, A. (2009). Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics. *Statistical Science*, 24(2), 176-178.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the Practice of Bayesian Statistics. B*ritish Journal of Mathematical and Statistical Psychology*, 66(1), 8-38.

Week 3: W Philosophy of Statistical Learning II

REQUIRED READING

Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The Perils of Policy by p-Value: Predicting Civil Conflicts. *Journal of Peace Research*, 47(4), 363-375.

SUPPLEMENTARY READING

Schrodt, P. A. (2014). Seven Deadly Sins of Contemporary Quantitative Political Analysis. *Journal of Peace Research*, 51(2), 287-300.

Schrodt, P. A., Yonamine, J., & Bagozzi, B. E. (2013). Data-based Computational Approaches to Forecasting Political Violence. In Handbook of Computational Approaches to Counterterrorism (pp. 129-162). Springer, New York, NY.

Skills Development: `Tidyverse` and associated `R` Libraries for Data Science.

DataCamp Review Courses: Data Manipulation with R, Data Visualization with R, Statistics Fundamentals with R

RECOMMENDED READING

Gill, J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly*, 52(3), 647-674.

Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015). No! Formal Theory, Causal Inference, and Big Data are not Contradictory Trends in Political Science. *PS: Political Science & Politics*, 48(1), 71-74.

Mearsheimer, J. J., & Walt, S. M. (2013). Leaving Theory Behind: Why Simplistic Hypothesis Testing is Bad for International Relations. *European Journal of International Relations*, 19(3), 427-457.

Ward, M. D. (2016). Can we Predict Politics? Toward what End?. *Journal of Global Security Studies*, 1(1), 80-91.

Week 4: M Bias-Variance Tradeoff and Resampling Methods I

REQUIRED READING

James et al. Ch. 5

SUPPLEMENTARY READING

Friedman, J. H. (1997). On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, 1(1), 55-77.

RECOMMENDED READING

Hastie et al. (2009) Ch. 7

Neunhoeffer, M., & Sternberg, S. (2019). How Cross-Validation can go Wrong and what to do about it. *Political Analysis*, 27(1), 101-106.

Muchlinski, D. A., Siroky, D., He, J., & Kocher, M. A. (2019). Seeing the Forest through the Trees. *Political Analysis*, 27(1), 111-113.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine Learning Algorithm Validation with a Limited Sample Size. *PloS One*, 14(11).

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot is more Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. *PloS One*, 10(3).

Quiz 2 Distributed

Week 4: W Bias-Variance Tradeoff and Reampling Methods II

Class Meeting in Person: Data Analytics Review

REQUIRED READING

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8), 861-874.

SUPPLEMENTARY READING

Greenhill, B., Ward, M. D., & Sacks, A. (2011). The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models. *American Journal of Political Science*, 55(4), 991-1002.

Hill, D. W., & Jones, Z. M. (2014). An Empirical Evaluation of Explanations for State Repression. *American Political Science Review*, 108(3), 661-687.

Skills Development Data Splitting, Cross-Validation, the Bootstrap

DataCamp Review Courses: Data Manipulation with R, Data Visualization with R, Intermediate Tidyverse Toolbox

Homework: #1 Due Today

Section 2: Supervised Learning

Week 5: M Linear Model Selection and Regularization I

REQUIRED READING

James et al. (2012) Ch. 6

SUPPLEMENTARY READING

Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 48-62.

RECOMMENDED READING

Hastie et al. (2009) Chs. 3-4

Murphy (2012) Ch. 13

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Week 5: W Linear Model Selection and Regularization II

REQUIRED READING

Beauchamp, N. (2017). Predicting and Interpolating State-Level Polls using Twitter Textual Data. *American Journal of Political Science*, 61(2), 490-503.

SUPPLEMENTARY READING

Bucca, M., & Urbina, D. R. (2019). Lasso Regularization for Selection of Log-linear Models: An Application to Educational Assortative Mating. *Sociological Methods & Research*.

Skills Development Penalized regression models and model selection/validation metrics

DataCamp Review Courses: Machine Learning Fundamentals in R, Supervised Machine Learning in R

Week 6: M Moving Beyond Linearity: GAMs I

REQUIRED READING

ISL Ch. 7

SUPPLEMENTARY READING

Week 6: W Moving Beyond Linearity: GAMs II

Goldsmith, B. E., & Butcher, C. (2018). Genocide Forecasting: Past Accuracy and New Forecasts to 2020. *Journal of Genocide Research*, 20(1), 90-107.

Beck, N., & Jackman, S. (1998). Beyond Linearity by Default: Generalized Additive Models. *American Journal of Political Science*, 596-627.

Carter, D. B., & Signorino, C. S. (2010). Back to the Future: Modeling Time Dependence in Binary Data. *Political Analysis*, 18(3), 271-292.

Skills Development: GAMS and Regression Splines

DataCamp Review Courses: Machine Learning Fundamentals in R, Supervised Machine Learning in R

Week 7: M Tree-Based Methods I

REQUIRED READING

ISL Ch. 8

Week 7: W Tree-Based Methods II

REQUIRED READING

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1), 87-103.

SUPPLEMENTARY READING

Montgomery, J. M., & Olivella, S. (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science*, 62(3), 729-744.

Kaufman, A. R., Kraft, P., & Sen, M. (2019). Improving Supreme Court Forecasting Using Boosted Decision Trees. *Political Analysis*, 27(3), 381-387.

Skills Development Tree-Based Methods

DataCamp Review Courses: Machine Learning Fundamentals in R, Supervised Machine Learning in R

RECOMMENDED READING

Siroky, D. S. (2009). Navigating Random Forests and Related Advances in Algorithmic Modeling. *Statistics Surveys*, 3, 147-163.

Metternich, N. W., Çiflikli, G., & Ali, A. (2019). Predicting the Severity of Civil Wars: An Actor-Centric Approach. SocArXiv. March, 28.

Jones, Z., & Linder, F. (2015, April). Exploratory Data Analysis using Random Forests. In Prepared for the 73rd annual MPSA conference.

Berk, R. A. (2006). An Introduction to Ensemble Methods for Data Analysis. *Sociological Methods & Research*, 34(3), 263-295.

Berk, R. (2012). Criminal Justice Forecasts of Risk: A Machine Learning Approach. Springer Science & Business Media.

Quiz 3 Distributed

Week 8: M Support Vector Machines I

REQUIRED READING

ISL Ch. 9

Week 8: W Support Vector Machines II

REQUIRED READING

D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the Wheat from the Chaff: Applications of Automated Document Classification using Support Vector Machines. *Political Analysis*, 22(2), 224-242.

SUPPLEMENTARY READING

Scanlon, J. R., & Gerber, M. S. (2014). Automatic Detection of Cyber-Recruitment by Violent Extremists. *Security Informatics*, 3(1), 5.

Skills Development Support Vector Machines

DataCamp Review Courses: Machine Learning Fundamentals in R, Supervised Machine Learning in R

RECOMMENDED READING

Diermeier, D., Godbout, J. F., Yu, B., & Kaufmann, S. (2012). Language and Ideology in Congress. *British Journal of Political Science*, 42(1), 31-55.

SUPPLEMENTARY READING

Minhas, S., Ulfelder, J., & Ward, M. D. (2015). Mining Texts to Efficiently Generate Global Data on Political Regime Types. *Research & Politics*, 2(3), 1-8.

Gründler, K., & Krieger, T. (2015). Using Support Vector Machines for Measuring Democracy (No. 130). Discussion Paper Series.

Week 9: M Neural Networks I

REQUIRED READING

Zeng, L. (1999). Prediction and Classification with Neural Network Models. *Sociological Methods & Research*, 27(4), 499-524.

SUPPLEMENTARY READING

Hastie et al. Elements of Statistical Learning, Ch. 11

Week 9: W Neural Networks II

REQUIRED READING

Beck, N., King, G., & Zeng, L. (2000). Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review*, 94(1), 21-35.

SUPPLEMENTARY READING

De Marchi, S., Gelpi, C., & Grynaviski, J. D. (2004). Untangling Neural Nets. *American Political Science Review*, 98(2), 371-378.

Skills Development Single-layer, feed-forward, Neural Networks

DataCamp Review Courses: Machine Learning Fundamentals in R, Supervised Machine Learning in R

Week 10: M Ensemble Learning I

REQUIRED READING

Montgomery, J. M., & Nyhan, B. (2010). Bayesian Model Averaging: Theoretical Developments and Practical Applications. *Political Analysis*, 18(2), 245-270.

SUPPLEMENTARY READING

Hastie, et al. (2009) Ch. 8

Sagi, O., & Rokach, L. (2018). Ensemble Learning: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), 1-18

Quiz 4 Distributed

Week 10: W Ensemble Learning II

Class Meeting in Person: Data Analytics Review

REQUIRED READING

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving predictions using ensemble Bayesian model averaging. Political Analysis, 20(3), 271-291.

SUPPLEMENTARY READING

Hastie, et al. (2009) Ch. 16

Muchlinski, D. (2020) Hail Hydra: Ensembles of Machine Learning Models to Forecast Political Violence. *Working Paper*

Skills Development Ensemble Learning, Bayesian Model Averaging, and Stacking

DataCamp Review Courses: Machine Learning Fundamentals in R, Supervised Machine Learning in R

Homework # 2 Due Today

RECOMMENDED READING

Clarke, B. (2003). Comparing Bayes Model Averaging and Stacking when Model Approximation Error cannot be Ignored. *Journal of Machine Learning Research*, 4, 683-712.

Domingos, P. (2000, June). Bayesian Averaging of Classifiers and the Overfitting Problem. In ICML (Vol. 2000, pp. 223-230).

Minka, T. P. (2000). Bayesian Model Averaging is not Model Combination. Available electronically at http://www. stat. cmu. edu/minka/papers/bma. html, 1-2.

Le, T., & Clarke, B. (2017). A Bayes Interpretation of Stacking for $\mathcal{M}$-Complete and $\mathcal{M}$-Open Settings. *Bayesian Analysis*, 12(3), 807-829.

Section 3: Unsupervised Learning

Week 11: M Unsupervised Learning I

REQUIRED READING

James et al. Ch. 10

SUPPLEMENTARY READING

Ahlquist, J. S., & Breunig, C. (2012). Model-based Clustering and Typologies in the Social Sciences. *Political Analysis*, 20(1), 92-112.

Week 11: W Unsupervised Learning II

REQUIRED READING

Jang, J., & Hitchcock, D. B. (2012). Model-Based Cluster Analysis of Democracies. *Journal of Data Science*, 10.

SUPPLEMENTARY READING

Hastie, Tibshirani, and Friedman (2009) Ch. 14

RECOMMENDED READING

Obinger, H., & Wagschal, U. (2001). Families of Nations and Public Policy. *West European Politics*, 24(1), 99-114.

Grimmer, J., & King, G. (2011). General Purpose Computer-Assisted Clustering and Conceptualization. *Proceedings of the National Academy of Sciences*, 108(7), 2643-2650.

Skills Development Clustering algorithms

DataCamp Review Courses: Unsupervised Machine Learning in R

RECOMMENDED READING

Schrodt, P. A. (2015). Comparing Methods for Generating Large Scale Political Event Data Sets. In Text as Data Meetings, New York University, 16–17, 2015 (pp. 1-32).

Gerner, D. J., Schrodt, P. A., Yilmaz, O., & Abu-Jabr, R. (2002). Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions. International Studies Association, New Orleans.

Kim, H., D'Orazio, V., Brandt, P., Looper, J., Salam, S., Khan, L., & Shoemate, M. (2019). UTDEventData: An R Package to Access Political Event Data. *Journal of Open Source Software*, 4(36), 1322.

Halterman, A., Irvine, J., Landis, M., Jalla, P., Liang, Y., Grant, C., & Solaimani, M. (2017). Adaptive Scalable Pipelines for Political Event Data Generation. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2879-2883). IEEE.

D'Orazio, V., Deng, M., & Shoemate, M. (2018). TwoRavens for Event Data. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 394-401). IEEE.

Week 12: M Text as Data I: Getting Started with Text

REQUIRED READING

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267-297.

SUPPLEMENTARY READING

Wilkerson, J., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, 20, 529-544.

Week 12: W Text as Data II: Preparing Text for Analysis

REQUIRED READING

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to do about it. *Political Analysis*, 26(2), 168-189.

SUPPLEMENTARY READING

Fariss, C. J., Linder, F. J., Jones, Z. M., Crabtree, C. D., Biek, M. A., Ross, A. S. M., ... & Tsai, M. (2015). Human Rights Texts: Converting Human Rights Primary Source Documents into Data. *PloS One*, 10(9).

Skills Development Using Text as Data: the `Tidy` Way

DataCamp Review Courses: Unsupervised Machine Learning in R, Text Mining with R

RECOMMENDED READING

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding Bag-of-Words Model: a Statistical Framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43-52.

Park, B., Colaresi, M., & Greene, K. (2018). Beyond a Bag of Words: Using PULSAR to Extract Judgments on Specific Human Rights at Scale. *Peace Economics, Peace Science and Public Policy*, 24(4).

Slapin, J. B., & Proksch, S. O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3), 705-722.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts using Words as Data. *American Political Science Review*, 97(2), 311-331.

Week 13: M Text as Data III: Topic Modeling

REQUIRED READING

A Beginner's Guide to Latent Dirichlet Allocation (LDA) `https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2`

Intuitive Guide to Latent Dirichlet Allocation `https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-`

SUPPLEMENTARY READING

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Lo, J., Proksch, S. O., & Slapin, J. B. (2016). Ideological Clarity in Multiparty Competition: A New Measure and Test using Election Manifestos. *British Journal of Political Science*, 46(3), 591-610.

Cordell, R., Clay, K. C., Fariss, C. J., Wood, R. M., & Wright, T. M. (2020). Changing Standards or Political Whim? Evaluating Changes in the Content of US State Department Human Rights Reports following Presidential Transitions. *Journal of Human Rights*, 19(1), 3-18.

Week 13: W Text as Data IV: Sentiment Analysis

REQUIRED READING

Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting Linguistic Markers for Radical Violence in Social Media. *Terrorism and Political Violence*, 26(1), 246-256.

SUPPLEMENTARY READING

Öztürk, N., & Ayvaz, S. (2018). Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis. *Telematics and Informatics*, 35(1), 136-147.

Skills Development Topic Modeling, Scaling, and Sentiment Analysis

DataCamp Review Courses: Unsupervised Machine Learning in R, Text Mining with R

RECOMMENDED READING

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3(Feb), 1137-1155.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12(Aug), 2493-2537.

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, 26(4), 417-430.

Dietrich, B. J., Enos, R. D., & Sen, M. (2019). Emotional Arousal Predicts Voting on the US Supreme Court. *Political Analysis*, 27(2), 237-243.

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

Lauderdale, B. E., & Herzog, A. (2016). Measuring Political Positions from Legislative Speech. *Political Analysis*, 24(3), 374-394.

		Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2), 254-277. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems (pp. 3111-3119). Quiz 5 Distributed
Week 14: M	Deep Learning I	REQUIRED READING Muchlinski, D., Yang, Z. Brich, S. Macdonald, C., & Ounis, I. (2020) We Need to Go Deeper: Measuring Electoral Violence using Convolutional Neural Networks and Social Media. *Political Science Research and Methods*, Online First SUPPLEMENTARY READING Zhang, H., & Pan, J. (2019). CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. *Sociological Methodology*, 49(1), 1-57. Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017, October). Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 786-794).
Week 14: W	Deep Learning II	Class Meeting in Person: Data Analytics Review REQUIRED READING Skills Development: `Kreas` for Deep Learning Homework # 3 Due Today RECOMMENDED READING Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. Journal of Artificial Intelligence Research, 57, 345-420. Peirson, V., Abel, L., & Tolunay, E. M. (2018). Dank Learning: Generating Memes using Deep Neural Networks. arXiv preprint arXiv:1806.04510.
TBD	RESEARCH NOTE	Due 22:00
TBD	FINAL EXAM	Due 22:00