Advanced Computational Methods for Social Media and Text Data

Christopher Fariss (cjfariss@umich.edu)

click here for the most up to date version of this document

Introduction to the Course

Course Content

This course focuses on the research design and data analysis tools used to explore and understand social media and text data. The fundamentals of research design are the same throughout the social sciences, however the topical focus of this class is on computationally intensive data generating processes and the research designs used to understand and manipulate such data at scale.

By massive or large scale, I mean that there are lots of subjects/connections/units/rows in the data (e.g., social network data like the kind available from twitter), or there are lots of variables/items/columns in the data (e.g., image or text data with many thousands of columns that represent the words in the document corpus), or the selected analytical tool is a computationally complex algorithm (e.g., a Bayesian simulation for modeling a latent variable, a random forest model for exploratory data analysis, or a neural network for automatically classifying new observations), or finally some combination of these three issues. The course will provide students with the tools to design observational studies and experimental interventions into large and unstructured data sets at increasingly massive scales and at different degrees of computational complexity.

How will we go about learning these tools? In this class, we will learn to program and program to learn. What do I mean? First, we will use the R program environment to learn the building blocks of programming. These skills are essential for managing the increasingly large and complex datasets of interest to social scientists (e.g., image data, text data).

As we develop programming skills in R, we will use them to help us understand how different types of data analysis tools work. For example, by the end of the course, students will be able to program and evaluate their own neural network or structural topic model from scratch.

We will start very small and learn how to scale up. In the beginning of the course, we will not make use of many packages other than the base packages available by default in R. As we proceed, we will learn how models for data work before then investigating the functions that exist in the large, always increasing catalogue of packages available for you to use in R. The development of new functions in R is advancing rapidly. The tools you learn in this class will help you improve as a programmer and a data scientists but learning how to program and using your programming skills to learn how to analyze data.

Course Objectives

Students will learn how to design models for data that take advantage of the wealth of information contained in new massive scale online datasets such as data available from twitter, images, and the many newly digitized document corpuses now available online. The focus of the course is on learning to program in R with special attention paid to designing studies in such a way as to maximize the validity of inferences obtained from these complex datasets.

- Learn to program models in R at a small scale using the base package and a minimal number of other packages
- Use the tools from research design to assist in model development
- Validate models of observational data in comparison to an appropriate baseline model
- Develop simulation based models for large scale, observational data
- Develop and validate measurement (e.g., latent variable models, structural topic models) and classification models (e.g., neural networks) of text and image based data

Course Prerequisites

Students should have some familiarity with concepts from research design and statistics. Generally, exposure to these concepts occurs during the first year course at a typical PhD program in political science. Students should also have familiarity with the R computing environment. The more familiarity with R the better.

Course Details

- We will begin each class period with a "programming challenge" (approximately 20-25 minutes).
- I will then give a short lecture over the class material (approximately 30-45 minutes).
- The remaining portion of class (approximately 1.5-2 hours) will be devoted to hands on learning with R, simulated data, and the large scale datasets provided by the instructor.
- The course schedule section, which is below, provides even more details about the topic of the lecture for each class day, citations for the discussion readings, and chapter entries from the text books for the programming and data analysis tools covered in the class.

Required Readings (Books)

1. Efron, Bradley and Trevor Hastie. 2016. *Computer Age Statistical Inference* Cambridge University Press. https://web.stanford.edu/~hastie/CASI/

2. Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

3. Matloff, Norman. 2011. *Art of R Programming: A Tour of Statistical Software Design*. no starch press.

Suggested Readings (Books)

4. Bolker, Ben. 2007. Ecological Models and Data in R. Princeton NJ: Princeton University Press.

5. Stan Development Team. 2018. "Stan Modeling Language: User's Guide and Reference Manual. Version 2.19." http://mc-stan.org/manual.html

6. Wickham, Hadley. "The tidyverse style guide" https://style.tidyverse.org

Additional required and suggested articles are listed below in the course schedule.

Day 0: Background Reading

Suggested Readings:

1. Efron and Hastie (Ch.2. "Frequentist Inference").
2. Efron and Hastie (Ch.3. "Bayesian Inference").
3. Efron and Hastie (Ch.4. "Fisherian Inference and Maximum Likelihood Estimation").
4. Efron and Hastie (Ch.5. "Parametric Models and Exponential Families").
5. Lazer, David and Jason Radford. 2017. "Data ex Machina: Introduction to Big Data" *Annual Review of Sociology* 43:19-39. http://doi.org/10.1146/annurev-soc-060116-053457
6. Lazer, David, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James H. Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, Marshall Van Alstyne 2009. "Computational Social Science." *Science* 323(5919): 721-723. https://doi.org/10.1126/science.1167742

Day 1: Introduction to Inference and Programming

Introduction to Research Design and Data Analysis using Programming and Simulation Based Methods as Learning Tools:

Introduction to the research design and data analysis tools used to explore and understand social media and text data.

Readings:

1. Matlofff 2011 (Ch.1: "Getting Started").
2. Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808-840. https://doi.org/10.1214/08-AOAS187

Suggested Readings:

3. Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics" *Econometrica* 12:1-115. https://doi.org/10.2307/1906935
4. Lin, Winston, Donald P. Green, and Alexander Coppock. "Standard operating procedures for Don Greens lab at Columbia." Version 1.05: June 7, 2016. https://github.com/acoppock/Green-Lab-SOP
5. Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25(3):289-310. http://dx.doi.org/10.1214/10-STS330
6. Shadish, William R. 2010. "Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings." *Psychological Methods* 15(1):3-17. https://doi.org/10.1037/a0015916
7. Tukey, John W. 1980. "We Need Both Exploratory and Confirmatory" *The American Statistician* 34(1):23-25. https://doi.org/10.2307/2682991

Day 2: Data Management and Analysis for Large Scale Data Structures

Introduction to Parallel Programming in R for Analysis and Exploration of Social Media Data:

Introduction to the R programming environment. We will review the various data and programming structures that are available for use in R. We will pay particular attention to vectorization and parallelization. Though we will begin with very small programs for learning, we also need to remember that the massive scale datasets that are increasingly available need optimized programs designed to manage and analyze these massive scale data structures.

Readings:

1. Matlofff 2011 (Ch.2: "Vectors").
2. Matlofff 2011 (Ch.3: "Matrices and Arrays").
3. Matlofff 2011 (Ch.4: "Lists").
4. Matlofff 2011 (Ch.5: "Data Frames").
5. Matlofff 2011 (Ch.6: "Factors and Tables").
6. Matlofff 2011 (Ch.7: "R Programming Structures").
7. Matlofff 2011 (Ch.14: "Performance Enhancement: Speed and Memory").
8. Matlofff 2011 (Ch.16: "Parallel R").

Suggested Readings:

9. Bolker (Ch.2, "Exploratory data analysis and graphics")
10. Matlofff 2011 (Ch.9: "Object-Oriented Programming").
11. Matlofff 2011 (Ch.13: "Debugging").
12. Wickham, Hadley. "The tidyverse style guide" https://style.tidyverse.org
13. R graph gallery. http://r-graph-gallery.com/

Day 3: Simulation Based Programming and Inference

Introduction to simulations in R:

We will develop and implement simulations in R to accomplish two learning goals. For the first learning goal, we will develop simulations in R to help us practice all of the programming and data structures that are available in R. For the second learning goal,

Readings:

1. Efron and Hastie (Ch.1. "Algorithms and Inference").
2. Efron and Hastie (Ch.3. "Bayesian Inference").
3. Efron and Hastie (Ch.4, "Fisherian Inference and Maximum Likelihood Estimation").
4. Gelman and Hill (Ch.7, "Simulation of probability models and statistical inferences").
5. Gelman and Hill (Ch.8, "Simulation for checking statistical procedures and model fits").
6. Matlofff 2011 (Ch.8: "Doing Math and Simulations in R").

Suggested Readings:

7. Bolker (Ch.5, "Stochastic simulation and power analysis")
8. Efron and Hastie (Ch.8, "Generalized Linear Models and Regression Trees").
9. Efron and Hastie (Ch.9, "Survival Analysis and the EM Algorithm").
10. Gelman and Hill (Ch.2, "Concepts and methods from basic probability and statistics").

Day 4: Simulation Based Programming for Model Comparison and Selection

Evaluating Model Performance Using In-sample and Out-of-sample Data:

In this section, we will program and learn several simulation based validation tools for assessing a model of observational data in comparison to an appropriate baseline model.

Readings:

1. Efron and Hastie (Ch.10, "The Jackknife and Bootstrap")
2. Efron and Hastie (Ch.12, "Cross-Validation and C_p Estimate of Prediction Error")

Suggested Readings:

3. Efron and Hastie (Ch.7, "James-Stein Estimation and Ridge Regression").
4. Efron and Hastie (Ch.11, "Bootstrap Confidence Intervals").
5. Fariss, Christopher J. and Zachary M. Jones. "Enhancing Validity in Observational Settings When Replication is Not Possible" *Political Science Research and Methods* 6(2):365-380. https://doi.org/10.1017/psrm.2017.5

Day 5: Measurement and Validation using Latent Variable Models

Measurement Models for Reducing High Dimensional Data for Visualization and Analysis with Applications to Text Data and Star Trek:

Measurement, as a scientific process, should accomplish two related tasks. First, link a theoretical concept to a data generating procedure. Second, link the data generating procedure to observed information. In this section we continue our focus on text as data, which is almost always found and not generated explicitly by a researcher. This means that we need to consider both the concept and the data generating process as part of our research program. What does this mean?

Political scientists are often interested in explaining concepts that are difficult or impossible to observe. Examples of unobservable concepts include political knowledge, political ideology, democracy, respect for human rights, or inequality. Even concepts that are based on directly observable information such as the number of individuals that reside in a state, the number of individuals killed during a conflict, or the level of economic output are often not easily observed. A key challenge for political scientists and social scientists generally, is creating models that can measure these concepts while also capturing the uncertainty associated with the processes by which they are measured.

This sections provide an introduction to measurement models generally with specific focus on Bayesian measurement models and measurement models that make use of text data. We will emphasize the use of construct validity to assess new and existing measures in applied research. We will motivate the development of these models with a discussion of the Bayesian perspective on the relationship between data and model parameters. This perspective is useful because it shifts the burden of validity from the primary source documentation and raw data to the model parameters that bind these diverse pieces of information together.

Readings:

1. Jackman, Simon. 2008. "Measurement." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford University Press.

2. Roberts, Margaret E, Brandon Stewart, and Dustin Tingley. "Navigating the Local Modes of Big Data: The Case of Topic Models." In *Data Analytics in Social Science, Government, and Industry*, New York: Cambridge University Press.

Suggested Readings:

3. Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529-546. https://doi.org/10.1017/S0003055401003100

4. Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76-91. https://doi.org/10.1093/pan/mpu011

5. Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. (2015). "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26(10):1531-1542. https://doi.org/10.1177/0956797615594620

6. Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Mrubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1). http://dx.doi.org/10.18637/jss.v076.i01

7. Efron and Hastie (Ch.13, "Objective Bayes Inference and MCMC").

8. Gelman and Hill (Ch.13, "Multilevel linear models: varying slopes, non-nested models, and other complexities").

9. Gelman and Hill (Ch.14, "Multilevel logistic regression").

10. Gelman and Hill (Ch.16, "Multilevel modeling in Bugs and R: the basics").

11. Gelman and Hill (Ch.25, "Missing-data imputation").

12. Hand, D. J., 1996. "Statistics and the Theory of Measurement." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*. 159(3):445-492. https://doi.org/10.2307/2983326

13. Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data" *American Political Science Review* 110(4):631-656. https://doi.org/10.1017/S000305541600037X

14. Stevens, S.S. 1946. "On the Theory of Scales of Measurement" *Science* 103(2684):677-680. https://doi.org/10.1126/science.103.2684.677

Day 6: Text as Data

Introduction to Regular Expressions, Document-by-Term matrices, and Construct Validity:

We will continue to use simple R programs to help us understand some of the common and important text processing steps. Much of the work involved in using text as data is in the processing of the character/string information. Regular expressions are the key functions that we will use. These key functions are embedded in many many R packages. We will start with the basics before move to more efficient libraries, again with the idea that we will learn the nuts and bolts of these critical tools.

Readings:

1. Matlofff 2011 (Ch.10: "Input/Output").
2. Matlofff 2011 (Ch.11: "String Manipulation").
3. Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267-297. https://doi.org/10.1093/pan/mps028

Suggested Readings:

1. Barberá, Pablo and Thomas Zeitzoff. 2017. "The New Public Address System: Why Do World Leaders Adopt Social Media?" *International Studies Quarterly* 62(1):121-130. https://doi.org/10.1093/isq/sqx047.
2. Steinert-Threlkeld, Zachary C. 2018. *Twitter as Data*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108529327.

Day 7: Automating Classification with Neural Networks

Introduction to Feed-forward Neural Networks and Back-propagation using Gradient Descent:

Neural Networks are a powerful tool for automated classification and any other predictive task. The main hurdle to understanding how these models work is the terminology associated with their implementation and use. During this section, we will see that a neural network is simply a linear model at its core and a set of liner transformations in the more exotic varieties of these models. Once we understand the basic building blocks of these models, we will see that neural networks are straight forward to implement but computationally quite expensive when implementing them on massive scale datasets. As usual, we will start with very simple models to learn. We will then apply this learning to more complex implementations of these models. We will use all of the tools developed during prior class periods to evaluate the performance of these models relative to a substantively meaningful baseline model.

Readings:

1. Efron and Hastie (Ch.18, "Neural Networks and Deep Learning").

Suggested Readings:

2. Bolker (Ch.7, "Optimization and all that")
3. Cantú, Francisco. 2019. "The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election" *American Political Science Review* 113(3):710-726. https:doi.org/10.1017/S0003055419000285.
4. Efron and Hastie (Ch.21, "Empirical Bayes Estimation Strategies").

Day 8: False Discovery and Inference After the Design and Selection of a Model

Techniques for selecting a substantively meaningful model:

When should we believe that the pattern we discovered using our model is a meaningful one? This is a question that we have already begun to develop tools for answering during this course.

Readings:

1. Efron and Hastie (Ch.15, "Large-Scale Hypothesis Testing and FDRs").
2. Efron and Hastie (Ch.20, "Inference After Model Selection").

Suggested Readings:

3. Imai, Kosuke, Luke J. Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765-789.
4. Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25(3):289-310. http://dx.doi.org/10.1214/10-STS330

Day 9: Putting everything together and extensions

Implementing and evaluating the wide variety of neural networks and other models for automating classification and prediction:

We will work on bringing all of the course material together during this penultimate class. It is also an opportunity to prepare for the optional final examination of the course, which takes place in the evening. The suggested readings also cover additional models that we were not able to cover in class. There are optional R programs that demonstrate the implementation of each of these models. These models can be implemented in predictive or categorization tasks and compared against the neural network from yesterday using one of the model comparison metrics that we covered earlier in the course.

Readings:

1. Review readings from prior class periods.

Suggested Readings:

2. Efron and Hastie (Ch.16, "Sparse Modeling and the Lasso").
3. Efron and Hastie (Ch.17, "Random Forests and Boosting").
4. Efron and Hastie (Ch.19, "Support-Vector Machines and Kernel Methods").

Day 10: Ethical Responsibilities for the Social Data Scientist

Issues relating to transparency and research ethics:

Whenever we are using social media data, no matter how aggregated, the privacy of the individual's personal data is an important consideration when designing a study.

Readings:

1. Driscoll, Jesse. 2016. "Prison States & Games of Chicken" in S. Desposato, *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, Taylor and Francis.

2. Margetts, Helen. 2017. "Political Behaviour and the Acoustics of Social Media" *Nature Human Behaviour* 1 (0086). https://doi.org/10.1038/s41562-017-0086

Suggested Readings:

3. Adam D. I. Kramer, Jamie E. Guillory, Jeffrey T. Hancock. 2014. "Emotional contagion through social networks" *Proceedings of the National Academy of Sciences* 111(24):8788-8790. https://doi.org/10.1073/pnas.1320040111

4. Lorenzo Coviello, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, James H. Fowler. 2014. "Detecting Emotional Contagion in Massive Social Networks" *PLOS ONE* 9(3):e90315. https://doi.org/10.1371/journal.pone.0090315

Day ++: Next Steps

Additional readings and training that will help to augment the material from this course:

In this course, we have focused exclusively on the research design tools and analysis techniques necessary for working with and understanding large scale social media and text datasets. We have not considered tools designed to gather and maintain large scale datasets. R is a useful tool for some of these tasks but other more powerful tools exist. Students should consider devoting additional energy to learning tools such as the python, json, and java programming languages and other tools for database management such as hive, hadoop, and SQL (Structured Query Language).

There is also still debate about the relative placement of the tools we have covered in this course we respect to other social science approaches. Below are links to entries from a recent symposium that debates this issue. Efron and Hastie also discuss throughout their book and especially in the "Epilogue", the relative development of statistical theory and justification for the new computationally intensive models that computer scientists and social scientists have been developing for particular applications. This continues to be an active area of scholarship.

Suggested Readings:

1. Efron and Hastie (Ch14. "Postwar Statistical Inference and Methodology").
2. Efron and Hastie ("Epilogue").
3. Matlofff 2011 (Ch.15: "Interfacing R to Other Languages").

Symposium: Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?

4. Clark, William Roberts and Matt Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?: Introduction" *PS: Political Science & Politics* 48(1):65-70. https://doi.org/10.1017/S1049096514001759
5. Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen, Betsy Sinclair. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):71-74. https://doi.org/10.1017/S1049096514001760
6. Titiunik, Rocío. 2015. "Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):75-79. https://doi.org/10.1017/S1049096514001772
7. Grimmer, Justin 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):80-83. https://doi.org/10.1017/S1049096514001784
8. Nagler, Jonathan and Joshua A. Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):84-88. https://doi.org/10.1017/S1049096514001796

9. Ashworth, Scott, Christopher R. Berry, Ethan Bueno de Mesquita. 2015. "All Else Equal in Theory and Data (Big or Small)" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):89-94. AllElseEqualinTheoryandData(BigorSmall)

10. Patty, John W., Elizabeth Maggie Penn. 2015. "Analyzing Big Data: Social Choice and Measurement" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):95-101. https://doi.org/10.1017/S1049096514001814

11. Keele, Luke. 2015. "The Discipline of Identification" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):102-106. https://doi.org/10.1017/S1049096514001826

Additional Course Information

Biographical Details

I am currently an Assistant Professor in the Department of Political Science and Faculty Associate in the Center for Political Studies at the University of Michigan. Prior to beginning these appointments, I was the Jeffrey L. Hyde and Sharon D. Hyde and Political Science Board of Visitors Early Career Professor in Political Science in the Department of Political Science at Penn State University. I am also an Affiliated Scholar at the Security and Political Economy (SPEC) Lab at the University of Southern California. In June 2013, I graduated with a Ph.D. in political science from the University of California, San Diego. I also studied at the University of North Texas, where I graduated with an M.S. in political science (2007), a B.F.A in drawing and painting (2005), and a B.A. in political science (2005).

My core research focuses on the politics and measurement of human rights, discrimination, violence, and repression. I use computational methods to understand why governments around the world torture, maim, and kill individuals within their jurisdiction and the processes monitors use to observe and document these abuses. Other projects cover a broad array of themes but share a focus on computationally intensive methods and research design. These methodological tools, essential for analyzing data at massive scale, open up new insights into the micro-foundations of state repression and the politics of measurement.

Resources for Harassment

Title IX makes it clear that violence and harassment based on sex and gender, including violence and harassment based on sexual orientation, are a Civil Rights offense subject to the same kinds of accountability and the same kinds of support applied to offenses against other protected categories such as race, national origin, etc. If you or someone you know has been harassed or assaulted, you can find the appropriate resources here: www.bw.edu/resources/hr/harass/policy.pdf

Language and Gender

"Language is gender-inclusive and non-sexist when we use words that affirm and respect how people describe, express, and experience their gender. Just as sexist language excludes womens experiences, non-gender-inclusive language excludes the experiences of individuals whose identities may not fit the gender binary, and/or who may not identify with the sex they were assigned at birth. Identities including trans, intersex, and genderqueer reflect personal descriptions, expressions, and experiences. Gender-inclusive/non-sexist language acknowledges people of any gender (for example, first year student versus freshman, chair versus chairman, humankind versus mankind, etc.). It also affirms non-binary gender identifications, and recognizes the difference between biological sex and gender expression. Teachers and students should use gender-inclusive words and language whenever possible in the classroom and in writing. *Students, faculty, and staff may share their preferred pronouns and names, either to the class or privately to the professor, and these gender identities and gender expressions should be honored.*." For more information:
www.wstudies.pitt.edu/faculty/gender-inclusivenon-sexist-language-syllabi-statement.

Syllabus Acknowledgments

This syllabus is based on several courses that I have taken and designed over the last several years. Some of the material is based on the Research Design (PL SC 501) course that I developed at Pennsylvania State University when I began teaching there in the fall of 2013, which itself is based on similar course developed by David Lake and Mathew McCubbins at the University of California, San Diego. It is also based on material that I developed for a graduate measurement theory class (PL SC 597) and undergraduate Social Data Analysis and Design class (SO DA 308) that I also developed at Pennsylvania State University. Elements of the syllabus and other class materials created for this class are also based in part on the Bayesian Statistics class offered by Seth Hill at University of California, San Diego and the Measurement class offered by Keith Poole at UCSD and now the University of Georgia. Additional information about these courses is available in my teaching and mentoring statement http://cfariss.com/documents/Fariss_teaching_mentoring_statement.pdf.

Links to Earlier Versions of the Course Syllabi

Advanced Methods for Social Media and Textual Data, The 51st, 52nd, and 53rd Essex Summer School in Social Science Data Analysis

- Summer 2020 Syllabus
- Summer 2019 Syllabus
- Summer 2018 Syllabus

Exploration and Analysis of Social Media Data, The 49th, and 50th, Essex Summer School in Social Science Data Analysis

- Summer 2017 Syllabus
- Summer 2016 Syllabus

Analyzing Big Data, The 48th Essex Summer School in Social Science Data Analysis

- Summer 2015 Syllabus