PS 498: Data Science for International Studies

Christopher Fariss (cjfariss@umich.edu)

Office: Institute for Social Research, 4248
Zoom Office Hours: Monday 2:00pm-4:00pm,
Zoom Office Hours: Tuesday, 10:00am-2:00pm
Zoom Office Hours: Wednesday, 5:30pm-6:30pm
and by appointment.

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

Introduction

Course Content

This class will provide undergraduate students with an introduction to the scientific method and programming tools for data science. Students will learn the fundamentals of the scientific method and, through programming and research design, how to improve both causal inference and the measurement of international political phenomena.

Course Objectives

Students will learn how to design studies to take advantage of the wealth of information contained in new online datasets such as data available from twitter, newly digitized document corpuses now available online, and event-based datasets such as the ACLED (Armed Conflict Location and Event Data) dataset. The focus of the course is on building data science tools in such a way as to maximize the validity of inferences obtained from complex datasets like these.

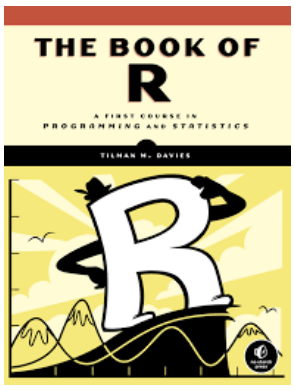
- Learn to program in R using the base package and a minimal number of other packages
- Collect and analyze international relations data using R
- Use the tools from research design to understand the what we can and cannot determine about the relationship between two or more variables

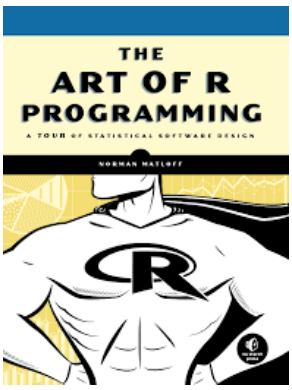
How will we go about learning these tools? In this class, we will learn to program and program to learn. What do I mean? First, we will use the R program environment to learn the building blocks of programming. As we develop programming skills in R, we will use them to help us understand how different types of data analysis tools work. For example, by the end of the course, students will be able to program and evaluate a model of data from several datasets of interest to students of world politics and international relations.

Required Books

1. Davies, Tilman M. 2016. *The Book of R: A First Course in Programming and Statistics*. no starch press.

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





Required Articles

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. Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get." *Political Analysis* 2:131-150. https://doi.org/10.1093/pan/2.1.131.

3. Hyde, Susan. 2015. "Experiments in International Relations: Lab, Survey, and Field." *Annual Review of Political Science* 18:403-424. https://doi.org/10.1146/annurev-polisci-020614-094854.

4. Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. 2010. "Introducing ACLED-Armed Conflict Location and Event Data" *Journal of Peace Research* 47(5):651-660. https://doi.org/10.1177%2F0022343310378914

5. 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

Web Documentaries and Lectures

We will also watch some short, 5-15 minutes web based documentaries and lectures about using data to understand trends around the globe.

1. Lublin, Nancy. 2012. "Analyzing text messages to save lives" (September 5, 2012)
   http://flowingdata.com/2012/09/05/analyzing-text-messages-to-save-lives/
2. Lublin, Nancy. 2015. "How data from a crisis text line is saving lives" (May, 2015)
   https://www.ted.com/talks/nancy_lublin_the_heartbreaking_text_that_inspired_a_crisis_help_line
3. Porway, Jake. 2013. "Data in the service of humanity" (September 2, 2013)
   http://flowingdata.com/2013/09/02/data-in-the-service-of-humanity/
4. Roser, Max. 2015. "Good Data will Make You an Economic Optimist"
   https://www.youtube.com/watch?v=519RSd65yFw
5. Rosling, Hans. 2006. "The best stats you've ever seen" (February 2006)
   https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen
6. Rosling, Hans. 2007. "New insights on poverty" (March 2007)
   https://www.ted.com/talks/hans_rosling_reveals_new_insights_on_poverty
7. "DNA Identifies War Victims" (September 29, 2013)
   http://www.youtube.com/watch?v=Kbk6QAfErXA
8. "International Commission on Missing Persons" (December 5, 2006)
   http://www.youtube.com/watch?v=w-Ykrhu8K78#t=386
   http://www.ic-mp.org/resources/video-material/

Data Science Projects

1. Case selection essay

The Case selection essay is a 2-4 page essay (12-point font, 1-inch margins, double space). In consultation with the instructor, select a country in the world. The country needs to be included in the ACLED dataset: https://www.acleddata.com/wp-content/uploads/dlm_uploads/2018/02/Country-and-Time-Period-coverage_updatedJune2019.pdf.

In the essay, you should provide a brief summary that describes why you choose the specific country case. That is, explain why you are curious about the region. What about the country is interesting to you? What are the features about the country that led you to select it? You will select 2 variables from https://ourworldindata.org/. How does your country compare to others for the two variables you selected?

You will use the country you select when gathering data for the ACLED and twitter data assignments, which are both described below. Additional readings suggested for this project include:

- Eck, Kristine and Christopher J. Fariss. 2018. "Ill Treatment and Torture in Sweden: A Critique of Cross-Case Comparisons" *Human Rights Quarterly* (40(3):591-604. https://doi.org/10.1353/hrq.2018.0033.
- Seawright, Jason. 2016. "The Case for Selecting Cases That Are Deviant or Extreme on the Independent Variable" *Sociological Methods & Research* 45(3):493-525. https://doi.org/10.1177/0049124116643556.

2. ACLED data collection and analysis project

Develop a visualization and statistical analysis for the daily ACLED event data associated with your selected country for the last 2 years. Additional readings suggested for this project include:

- Eck, Kristine. 2012. "In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets" *Cooperation and Conflict* 47(1):124-141. http://doi.org/10.1177/0010836711434463.
- Pierskalla, Jan H. and Florian M. Hollenbach. 2013. "Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa" *American Political Science Review* 107(2):207-224. https:doi.org/10.1017/S0003055413000075.

3. Twitter data collection and analysis project

In consultation with the instructor, select one twitter account associated with your selected country. For example, you could pick the account associated with the executive of the country (e.g., @Macky_Sall, the President of Senegal) or a prominent organization that works in the country (e.g., @AmnestyNigeria or @AmnestyWARO). You will collect the tweets associated with this account and develop a visualization and statistical analysis of the content of those tweets. Additional readings suggested for this project include:

- 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.

Active Learning and Attendance

As students, we learn more when we actively engage with material instead of passively consuming it. This insight is supported by extensive research from college-level courses.[1] Because our course does not contain an in-class passive lecture (see Class Schedule Key below), participation and therefore attendance is critical for achieving our shared learning goals and is therefore mandatory. We will program Virtual in class together each week during our 3-hour class session, every Wednesday morning (one class per week over 14 weeks). Students may miss one class period with documentation. Any missed class period without documentation or after the first documented absence will result in 8 points taken from the final grade. Missing five classes (15 hours of class time) will result in a grade of F.

Grade Values

Participation/Attendance: 40 points

Case selection essay: 10 points

ACLED data collection and analysis project: 15 points

Twitter data collection and analysis project: 15 points

Midterm Exam: 10 points

Final Exam: 10 points

Grade Ranges

A [93–100]

A- [90–93)

B+ [87–90)

B [83–87)

B- [80–83)

C+ [77–80)

C [73–77)

C- [70–73)

D [60–70)

F [0–60)

[1] See for example: Louis Deslauriers, Logan S. McCarty, Kelly Miller, Kristina Callaghan, Greg Kestin. 2019. "Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom" *Proceedings of the National Academy of Sciences* 116 (39) 19251-19257; https://doi.org/10.1073/pnas.1821936116

Class Schedule

Class Schedule Key

Virtual program challenges: These challenges are meant to be just that, challenging. They should also be fun. I encourage you to think creatively and collaboratively. Getting stuck or not finishing all the steps is expected and encouraged. This is how learning works. Always start with step (1) and then continue to each step as time permits. Don't worry about completing each step. Document your code for each step. You may wish to come back to some of the harder steps as you progress through the course. Note that some of the steps may ask you to use skills we have not yet covered in the course. Don't worry about these steps now but definitely think through the programming logic if you are stuck and make plans to come back to try them once you feel ready.

Virtual class programming lab: About an hour of class time will be devoted to programming in R together. Please make sure to read the assigned chapters and view the video lectures before each weeks classes so that you are ready to program Virtual in class.

Virtual in class reading discussion: Occasionally over the course of the term, we will read and discuss an academic article. Please read these articles prior to the class date.

Offline video: We will watch a few short videos about data science in some classes.

Out of class lecture: Each week I will post slides and a lecture video file in which I review the material contained within the slides. These weekly lectures will review key information from the weekly reading assignments. You should review these lectures after you read the assigned weekly readings and before you attend class.

Readings for the week: Assigned readings (book chapters or articles). Please make sure to read these articles prior to the Wednesday class each week.

Week 1: Introduction

Readings for the Week:

1. Davies Ch.1 "Getting Started"

Wednesday — 01/20/2020

- **Out of class video:** Porway, Jake. 2013. "Data in the service of humanity" (September 2, 2013) http://flowingdata.com/2013/09/02/data-in-the-service-of-humanity/
- **Virtual in class programming lab:** Download and install R and Rstudio

Week 2: Numbers in R

Readings for the Week:

1. Davies Ch.2 "Numerics, Arithmetic, Assignment, and Vectors
2. Matloff Ch.2 "Vectors" (optional)
3. Davies Ch.3 "Matrices and Arrays"
4. Matloff Ch.4 "Matrices and Arrays" (optional)

Wednesday — 01/27/2020

- **Out of class video:** Rosling, Hans. 2006. "The best stats you've ever seen" (February 2006) https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen
- **Virtual in class programming lab:** Introduction to programming in R
- **Virtual in class programming lab:** Making and manipulating matrices and arrays in R

Week 3: Case Selection

Readings for the Week:

1. Geddes (1990)

Wednesday — 02/03/2020

- **Out of class video:** Roser, Max. 2015. "Good Data will Make You an Economic Optimist" https://www.youtube.com/watch?v=519RSd65yFw
- **Virtual in class reading discussion:** Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get." *Political Analysis* 2:131-150.
- **Virtual in class programming lab:** Making and manipulating matrices and arrays in R (continued)

Week 4: Data inside and outside of R

Reminder: Case selection choice is due on the Canvas website.

Readings for the Week:

1. Davies Ch.8 “Lists and Data Frames”
2. Matloff Ch.4 “Lists” (optional)
3. Matloff Ch.5 “Dataframes” (optional)
4. Matloff Ch.6 “Factors and Tables” (optional)

Wednesday — 02/10/2020

- **Virtual in class programming lab:** working with lists, dataframes, and tables in R

Week 5: Programming in R

Due Date: Case selection essay is due on Canvas this week.

Readings for the Week:

1. Davies Ch.10 “Conditions and Loops”
2. Matloff Ch.7 “R Programming Structures” (optional)
3. Matloff Ch.4 “Lists”
4. Matloff Ch.5 “Dataframes”
5. Matloff Ch.6 “Factors and Tables”

Wednesday — 02/17/2020

- **Virtual in class programming lab:** Conditions and Loops in R
- **Virtual in class programming lab:** working with lists, dataframes, and tables in R (continued)

Week 6: Programming in R

Readings for the Week:

no readings this week

Wednesday — 02/24/2020

- Rest Day

Week 7: Programming in R

Readings for the Week:

1. Davies Ch.9 "Calling Functions"
2. Davies Ch.11 "Writing Functions"
3. Davies Ch.8 "Reading and Writing Files"
4. Matloff Ch.10 "Input/Output" (optional)

Wednesday — 03/03/2020

- **Virtual in class programming lab:** Using and Writing functions in R
- **Virtual in class programming lab:** Reading and writing files in R 5
- **Virtual in class programming lab:** R programming review

Week 8: Review

no readings this week

Wednesday — 03/10/2020

- **Virtual in class exam:** The R programming exam consists of two parts (1) an open book take home exam, and (2) and in-class program challenge. Each student should select one of the program challenges from the course so far and work to finish as many of the steps for the program challenge as possible. During the in-class portion, students may work collaboratively.

Week 9: Statistical and Visual Summaries of Data

Readings for the Week:

1. Davies Ch.15 "Probability"
2. Davies Ch.16 "Common Probability Distributions"
3. Davies Ch.14 "Basic Data Visualization"
4. Matloff Ch.12 "Graphics" (optional)

Wednesday — 03/17/2020

- **Virtual in class programming lab:** visualizing probability distributions in R

Week 10: Statistical and Visual Summaries of Data

Readings for the Week:

1. Davies Ch.13 "Elementary Statistics"
2. Davies Ch.17 "Sampling Distributions and Confidence"
3. Raleigh et al. (2010)

Wednesday — 03/24/2020

- **Virtual in class reading discussion:** Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. 2010. "Introducing ACLED-Armed Conflict Location and Event Data" *Journal of Peace Research* 47(5):651-660.
- **Virtual in class programming lab:** Visualizing and summarizing event count data from ACLED

Week 11: Using Strings of Text as Data

Readings for the Week:

1. Barberá and Zeitzoff (2017)
2. Matloff Ch.11 "String Manipulation"

Wednesday — 03/31/2020

- **Out of class video:** Lublin, Nancy. 2015. "How data from a crisis text line is saving lives" (May, 2015) https://www.ted.com/talks/nancy_lublin_the_heartbreaking_text_that_inspired_a_crisis_help_line http://flowingdata.com/2012/09/05/analyzing-te
- **Virtual in class reading discussion:** 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.
- **Virtual in class programming lab:** Regular expressions and twitter data in R

Week 12: Research Design and Data Analysis

Readings for the Week:

1. Davies Ch. 18 "Hypothesis Testing"
2. Davies Ch. 20 "Simple Linear Regression"

Wednesday — 04/07/2020

- **Virtual in class programming lab:** Programming statistical comparisons in R
- **Virtual in class programming lab:** Simple linear model in R

Week 13: Research Design and Data Analysis

Readings for the Week:

1. Hyde (2015)
2. Rubin (2008)
3. Matloff Ch.8 "Doing Math and Simulations in R" (optional)

Wednesday — 04/14/2020

- **Virtual in class reading discussion:** Hyde, Susan. 2015. "Experiments in International Relations: Lab, Survey, and Field." *Annual Review of Political Science* 18:403-424.
- **Virtual in class reading discussion:** Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808-840.
- **Virtual in class programming lab:** Simulating potential outcomes in R

Week 14: Graphics

Due Date: Selected data visualization (ACLED or twitter data) is due Virtual in class this week.

Readings for the Week:

1. Davies Ch.23 "Advanced Plot Customization"
2. Davies Ch.24 "Going Further with the Grammar of Graphics" (optional)
3. Davies Ch.25 "Defining Colors and Plotting in Higher Dimension" (optional)

Wednesday — 04/21/2020

- **Virtual in class programming lab:** Advanced graphics in R
- Data Visualization Critique

Finals Week

Figure 1: Artwork by @allison_horst

Additional Course Information

Student Mental Health and Wellbeing

University of Michigan is committed to advancing the mental health and wellbeing of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available.

For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and `https://caps.umich.edu/` during and after hours, on weekends and holidays, or through its counselors physically located in schools on both North and Central Campus.

You may also consult University Health Service (UHS) at (734) 764-8320 and `https://www.uhs.umich.edu/mentalhealthsvcs`, or for alcohol or drug concerns, see `www.uhs.umich.edu/aodresources`.

For a listing of other mental health resources available on and off campus, visit: `http://umich.edu/ mhealth/`.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way the course is usually taught may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; `http://ssd.umich.edu`) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.

Religious and Academic Conflicts

Although the University of Michigan, as an institution, does not observe religious holidays, it has long been the University's policy that every reasonable effort should be made to help students avoid negative academic consequences when their religious obligations conflict with academic requirements. Absence from classes or examinations for religious reasons does not relieve students from responsibility for any part of the course work required during the period of absence. Students who expect to miss classes, examinations, or other assignments as a consequence of their religious observance shall be provided with a reasonable alternative opportunity to complete such academic responsibilities.

It is the obligation of students to provide faculty with reasonable notice of the dates of religious holidays on which they will be absent. Such notice must be given by the drop/add deadline of the given term. Students who are absent on days of examinations or class assignments shall be offered an opportunity to make up the work, without penalty, unless it can be demonstrated that a make-up opportunity would interfere unreasonably with the delivery of the course. Should disagreement arise over any aspect of this policy, the parties involved should contact the Director of Undergraduate Studies/Director of Graduate Studies. Final appeals will be resolved by the Provost.

Students Representing the University of Michigan

There may be instances when students must miss class due to their commitment to officially represent the University. These students may be involved in the performing arts, scientific or artistic endeavors, or intercollegiate athletics. Absence from classes while representing the University does not relieve students from responsibility for any part of the course missed during the period of absence. Students should provide reasonable notice for dates of anticipated absences and submit an individualized class excuse form.

Academic Integrity

The LSA undergraduate academic community, like all communities, functions best when its members treat one another with honesty, fairness, respect, and trust. The College holds all members of its community to high standards of scholarship and integrity. To accomplish its mission of providing an optimal educational environment and developing leaders of society, the College promotes the assumption of personal responsibility and integrity and prohibits all forms of academic dishonesty and misconduct. Academic dishonesty may be understood as any action or attempted action that may result in creating an unfair academic advantage for oneself or an unfair academic advantage or disadvantage for any other member or members of the academic community. Conduct, without regard to motive, that violates the academic integrity and ethical standards of the College community cannot be tolerated. The College seeks vigorously to achieve compliance with its community standards of academic integrity. Violations of the standards will not be tolerated and will result in serious consequences and disciplinary action.

Grade Grievances

If you believe a grade you have received is unfair or in error, you will need to do the following:
Wait 24 hours after receiving the grade before approaching your instructor. Provide an explanation in writing for why the grade you received was unfair or in error. If you believe the instructorÕs response fails to address your claim of unfairness or error, you may petition the departmentÕs Director of Undergraduate Studies at the latest within the first five weeks of classes following the completion of the course. You must convey in writing the basis for the complaint, with specific evidence in support of the argument that the grade either was given in error or was unfairly determined. This formal complaint also should summarize the outcome of the initial inquiry to the course instructor, indicating which aspects are in dispute. Within three weeks of the receipt of the petition, the DUS will determine whether to convene the Undergraduate Affairs Committee, the student, and the instructor(s) for a formal hearing. Further details on this process are included on the department website under Advising $>$ Contesting a Grade.

Late Assignments

I will deduct one letter grade from an assignment for each week it is past due.

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 womenÕs 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.