Gov 1005: Data

David Kane

Spring 2020

Description

Data matters. Learning to think critically about data is a fundamental skill. How much money is donated to political campaigns? How do polls measure public opinion? Does exposure to Spanish-speakers affect attitudes toward immigration? We need data to answer these questions – to describe, to predict, and to infer.

This course, an introduction to data science, will teach you how to *think with data*, how to gather information from a variety of sources, how to import that information into a project, how to tidy and transform the variables and observations, how to visualize, how to model relationships, how to assess uncertainty, and how to communicate your findings. Each student will complete a final project, the first entry in their professional portfolio. Our main focus is data associated with political science, but we will also use examples from education, economics, public health, sociology, sports, finance, climate and any other topic which students find interesting.

We use the R programming language, RStudio, Git and GitHub.

Prerequisites: None. You must have a laptop with R, RStudio and Git installed.

Logistics: Class meets in Tsai Auditorium from 1:30 to 2:45 on T/TH.

Figure 1: *Ulysses and the Sirens*, 1891, by John William Waterhouse. Homer's *Odyssey* recounts the decade-long journey home of Odysseus (known as Ulysses in Latin) after the Trojan War. Although Ulysses's ultimate goal is his kingdom of Ithaca, he does not shy away from adventure along the way. The Sirens use their enchanting voices to lure unwary sailors to their deaths. Ulysses wanted to hear their songs. He instructed his men to fill their ears with beeswax and to tie him to the mast.

Course Metaphor

The central metaphor for this class is *Ulysses and the Sirens*. You are Ulysses. Ithaca is the future you want. The Sirens are the many distractions of the modern world. *I am the rope.* No course at Harvard does more to increase students' chances of getting the future they want.

Course Staff

Preceptor David Kane; dkane@fas.harvard.edu; CGIS South 310; 646-644-3626; office hours Thursday from 8:30 to 11:00, generally held in Fisher Commons. Please address me as "Preceptor," not "David," nor "Preceptor Kane," nor "Professor Kane," nor "Mr. Kane," nor, worst of all, "Dr. Kane." I respond to e-mail within 24 hours. If I don't, e-mail me again.

Head Teaching Fellow: Alyssa Huberts (alyssa_huberts@g.harvard.edu). E-mail Alyssa (and cc your assigned TF and Preceptor) if you have a complaint about a grade.

Teaching Fellows: Mitchell Kilborn (mkilborn@g.harvard.edu), June Hwang (hwang@g.harvard.edu) and Kaneesha Johnson (krjohnson@g.harvard.edu).

Head Course Assistant: Claire Fridkin (clairefridkin@college.harvard.edu).

Course Assistants: Shivani Aggarwal (saggarwal@college.harvard.edu), Evelyn Cai (evelyncai@college.harvard.edu), Katie Cao (kcao@college.harvard.edu), April Chen (aprilchen@college.harvard.edu), Rucha Joshi (ruchajoshi@college.harvard.edu), Seeam Noor (seeamnoor@college.harvard.edu), Joshua Pan (joshuapan@college.harvard.edu), Jack Schroeder (jackschroeder@college.harvard.edu), Dillon Smith (dillon_smith@college.harvard.edu), Amy Tan (atan@college.harvard.edu), Yao Yu (yaodongyu@college.harvard.edu) and Roger Zhang (ruoqizhang@college.harvard.edu).

Course Philosophy

No Lectures: The worst method for transmitting information from my head to yours is for me to lecture you. There are no lectures. We work on problems together during class. You learn soccer with the ball at your feet. You learn about data with your hands on the keyboard.

Bayesian: The philosophy of this class is unapologetically Bayesian.

R Everyday: Learning a new programming language is like learning a new human language: You will practice (almost) every day.

Community: You will probably learn the names of more students in this course than in all your other courses combined. *Awkwardness in the pursuit of community is no vice.*

Professionalism: We use professional tools. Your workflow will be very similar to the workflow involved in paid employment. Your problem sets and final project will be public, the better to impress others with your abilities. High quality work will be shared with your classmates. We will learn the "full cycle" of how to draw inferences from data and communicate those inferences to others.

Cold Calling: I call on students during class. This keeps every student involved, makes for a more lively discussion and helps to prepare students for the real world, in which you can't hide in the back row. Want to be left alone? Don't take this course.

Advisory Groups: We do not have normal sections. Instead, you will spend 30 minutes each week meeting with your assigned TF in a small group. Recall what I just said about (not) being left alone in this course.

Millism: Political disputes are not the focus of this class but, when such topics arise, I will insist that we follow John Stuart Mills' advice: "He who knows only his own side of the case, knows little of that. His reasons may be good, and no one may have been able to refute them. But if he is equally unable to refute the reasons on the opposite side; if he does not so much as know what they are, he has no ground for preferring either opinion."

Teaching to Learn: My main goal is not to teach you how do X. That is easy! More importantly, in a few months, I won't be around to teach you Y. My goal is to teach you how to teach yourself, how to figure out X and Y and Z on your own. That is harder! Much of the pedagogy of the course — especially my insistence that you work on topics not covered in lecture — is driven by this goal. You will find it frustrating.

Course Policies

Book: The text for the class is *Preceptor's Primer for Bayesian Data Science*, a book-in-progress.

Workload: The course should take about 10 to 15 hours a week, outside of class meetings, exams and the final project. This is an expected average across the class as a whole. *It is not a maximum.* Some students will end up spending much less time. Others will spend **much, much more**.

Late Days: Assignments (DataCamp, Problem Sets and Final Project Milestones) are always due at 11:55 PM, unless specified otherwise. An assignment is a day late if it is turned in any time after it was due (even 5 minutes after) but within 24 hours. After that, it is two days late, and so on. You have 5 *late days* in total. Late days may be used for any assignment, except the four exams and the final project Demo Day. **You should save your late days.** If you use them early in the semester for no particularly good reason and then, later in the semester, have an actual emergency, we will not be sympathetic. We will not give you extra late days in such a situation. (That isn't fair to your classmates, and we are all about fairness.) We will just, mentally, move the late days you wasted so that they cover your actually emergency. You will now be penalized for being late earlier in the semester, when you did not have a good reason for tardiness. You may only use one late day on a given assignment. Hand it in after more than 24 hours and you get a zero on that assignment. **But you still must hand it in!** Everything must be completed. Late days accrue until you do. Each day late (beyond the five allowed) results in -1 point to your final score. This decrement is a *point*

not *percentage* penalty. In other words, each additional late day used outside of the allotted five will drop your final class grade by one point out of 100.

Submissions: All problem sets and milestones are turned in via Canvas. Late days are assigned on the basis of the official Canvas submission time. For problem sets, you submit the html file. For milestones, the submission will generally be a simple text file with some information about your project. Nothing is submitted for DataCamps. In addition to looking at your Canvas submissions, we will also grade your repos for each assignment.

Missing Class: You expect me to be present for lecture. I expect the same of you. There is nothing more embarrassing, for both us, than for me to call your name and have you not be there to answer. But, at the same time, conflicts arise. It is never a problem to miss class if, for example, you are out of town or have a health issue. Simply put an X by your name in the Google absence sheet **and** send me and your TF an e-mail. Failure to do so will decrease your participation points, as will missing too many classes, even with notification. There is no need to put a reason in the sheet. An X is enough.

Major Emergencies: We are not monsters. If you are hit with a major emergency — the sort of thing that necessitates the involvement of your Resident Dean — we will be sympathetic. We require a signed letter (a piece of paper, not an e-mail) from your Resident Dean as documentation. Speak with your TF.

Organized by House: We use geography to create a community. During class, you will sit with students from your house, grouped with other houses near yours. You will work with different peers each class.

Monologues: I give brief monologues, designed to explain specific topics that have confused students in the past. I hope to never talk for more than 5 minutes straight.

Speakers: Data scientists, from both industry and academia, will speak with us. If there is someone you would like to meet, talk to me about it and we can invite them! We follow a *No Laptop Rule* during speaker presentations. Close your laptops. Put down your phones. If you want to take notes, use a pen. We do this because we respect the speakers, want to give them our full attention, and are thankful that they have taken the time to talk with us.

No Cost: Every reading/tool we use is free. You don't have to spend any money on this class. Some activities, like DataCamp and GitHub, have paid options which provide more services, but you never have to use them. Don't give anyone your credit card number.

Remind Me: In conversations outside a class, a student will often ask an important question or raise of issue of general interest. These topics should be brought to the attention of other students. I will ask you to "Remind me" about it. This means that, in the next class, **you must raise your hand** when I ask for reminders and then remind me! Couldn't I just write it down in my notes? Perhaps. But learning how to raise your hand/voice in a big class is a useful skill. This is your opportunity to practice.

Role of Teaching Fellows: The TFs run their Office Hours, grade all assignments, keep track of late days, deal with emergencies and so on. Go to them first with any problems. You will be assigned to work closely with a specific TF — your "assigned" TF — but you may ask other TFs for help as well.

Role of Course Assistants: The CAs only run Study Halls. They can make no commitments about how the TFs will assign the final grade on a problem set, milestone or exam. *Never ask a CA a question about grading.* Instead, ask on Piazza and a TF will respond, or come to a TF privately with your question.

Exceptions: There may be a reason why you can't adhere to class policies. For example, severe social anxiety may make being cold-called problematic. A learning disability may make take-home tests unfair. Whatever the situation, please seek me out for conversation. I am sure we can work out something! I will do whatever it takes to allow every Harvard student to thrive in this class.

Use your Harvard e-mail: Please use your official Harvard e-mail address for all aspects of this class, especially things like signing up for services like DataCamp, GitHub, and so on. Doing so makes it much easier for us to figure out who is doing what. This may not be easy if you already connect with these services but, even in that case, you should be able to add your Harvard e-mail address to your account.

Piazza: All general questions — those not of a personal nature — should be posted to Piazza so that all students can benefit from both the question and the answer(s).

Plagiarism: If you plagiarize, you will fail the course. See the Harvard College Handbook for Students for details.

Working with Others: Students are free (and encouraged) to discuss problem sets and their final projects with one another. However, you must hand in your own unique code and written work in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your friend, sitting side-by-side and going through the problem set question-by-question, but you must **each type your own code**. Your answers may be similar (obviously) but they must not be identical, or even identical'ish.

Git and GitHub: Analyzing data without using source control is like writing an essay without using a word processor — possible but not professional. We will do all our work using Git/GitHub.

Readings: Assignments in a given week cover (approximately) the material that we will use that week. I will not hesitate to cold-call students with questions about the readings. Do them.

Optional Activities: The syllabus includes background readings, videos and materials which students may find interesting. You do not have to do them.

Waite Rule: We don't wear hats in the classroom. (Obviously, this prohibition does not apply to headgear of a religious nature.)

Computer Emergencies: We are not sympathetic about computer emergencies. You should keep all your work on GitHub, so it won't matter if your computer explodes. If it does explode, you will lose only the work after your last push. You can then restart your work on a public computer (the basement of CGIS Knafel has machines with R/RStudio installed) or on your roommate's computer.

Github Classroom: We use Github Classroom to distribute problem sets and exams. You will receive an e-mail with a link. Click on that link and a repo, with instructions, will be created. *Do this as soon as you receive the e-mail.* We don't want GitHub problems to arise the night before the assignment is due.

Tardiness: We begin on time and end on time. Do not start gathering your belongings until class is over, especially when we have a speaker.

Credit: Gov 1005 fulfills the QRD requirement. You may also get concentration credit. This is true, obviously, for Government. It is also true in Statistics, Psychology, Sociology, and Social Studies. I am happy to support students who want to petition other departments.

Announcements: You are responsible for any assignment/exam/deadline updates/changes which are either announced in class or promulgated via the course Canvas e-mail list. The official Preceptor's Notes on Piazza are important, but we will e-mail them to you. You are not responsible for every other random post on Piazza. In fact, you can ignore Piazza if you want.

Grading

Solo Participation: 5 points. This category relates to things you do alone in class. Missing class (without notifying us) or missing too many classes will cost you points, as will a failure to participate in class activities. We keep track of this via Google forms and sheets, so be sure to fill them out when requested. Note that I do not care if you know the answer when I cold-call you. This plays no part in your grade.

Group Participation: 5 points. This category relates to activities you do with other students. Helping your fellow students, especially on Piazza, is the best form of group participation, as is volunteering for a class role. Be a good class citizen. Help your classmates during Study Halls. Do not shirk on group projects.

DataCamp Lessons: 5 points. Grades are pass/fail only. Given the level of the questions and the hints provided, it is essentially impossible not to get full credit as long as you make an honest effort. You may use the hints and, very occasionally, just show the answer.

Problem Sets: 22 points. The first problem set is worth 1 point. The remaining 7 are worth 3 points each. Problem sets after the first are distributed on Thursday and then due the following Wednesday at 11:55 PM. You are welcome to work on them with your friends but, first, you must personally type in every character in the work you submit and, second, you must list all the people you worked with. We define "work with" very broadly, to include minor interactions. You would certainly list anyone you sat nearby during Study Hall, for example.

Exams: 35 points total. The four exams are take-home and unhackable. The first is worth 5 points and the others are each worth 10 points. They are open-book and open-web. Because students have different schedules, you can complete the exam any time within a four-day window starting after exam distribution. Late exams earn zero points. You may not seek or receive help on the exam from a person, e.g., asking a roommate or posting at RStudio Community. You may use any written materials from the class, including problem set answers. If you have a question, ask on Piazza. Teaching staff (not other students) will answer it.

Final Project: 28 points. Students will present their projects publicly at the end of the semester. They will then have the opportunity to incorporate feedback before submitting the final version. There are eight milestones for the projects, each worth one point. Demo Day (which includes a review of your code) is worth 10 points. The final project submission is worth 10 points. Follow this advice.

Calculation: Each problem set, milestone and exam is graded out of a maximum of score of 20, regardless of its weight in the final grade calculation. For example, both Exam 2 and Milestone 2 are graded out of 20, but the former is worth ten times as much to your final grade.

Final Project

Do you love soccer or wine or NYC politics? The final project provides you with an opportunity to study that topic in depth. Your final project will be, for most of you, the first item in your professional portfolio, something so impressive that you will be eager to show it to graduate schools or potential employers. You must show this work publicly, both on the web (viewable by all) and in person at our Demo Day. You will host your final project using Shiny Apps. Make use of free statistical consulting from the Harvard Statistics Department and from IQSS. Read this advice if you are working with data larger than 100 megabytes. Consider scheduling an interview with Hugh Truslow (truslow@fas.harvard.edu), Head, Social Sciences and Visualization, Harvard University. No one at Harvard knows more about potential data sources. Visualization Specialist Jessica Cohen-Tanugi (jessica_cohen-tanugi@harvard.edu) is a great person to talk to about your graphics. Explore the final projects from past semesters.

Possible Approaches

Most students will gather some data, estimate some models, and create a Shiny App. Good stuff! But there are other possible approaches:

Original Data Collection

Students interested in a topic about which there is no publicly available data are welcome to collect their own data. This must be something much more substantive than just asking 100 students outside Annenberg about their favorite salad. Two categories of data work best. First, pick a topic which you truly care about. Second, pick something Harvard-specific. This *Crimson* article and these class projects — spring 2019 and fall 2019 — are great examples of the latter.

Work with Other Classes

You are welcome to use data from your other classes in the creation of your final project. This includes thesis work. You automatically have permission from us to do this, but you must also obtain permission from the instructor of the other class.

Others?

Interested in doing a project which seems different from what we describe above? Come talk to me! The best projects involve topics which students are passionate about. If you really care about X, then we are eager to help you create a final project about X. Examples: participation in the NFL Big Data Bowl, submitting Numerai forecasts or entering a Kaggle competion.

Conclusion

If you had tried to complete a data analysis project before taking this class, you would have done X well. Now that you have taken the class – now that you know how to describe, predict and infer – you will do Y well. The success (or failure) of the class can be measured by comparing Y with X.

Organization

Everything — DataCamp (Mondays), Problem Sets (Wednesdays), Milestones (Fridays) and Exams (Sundays) — is due at 11:55 PM, unless otherwise specified.

Rhythm of the Class

The class follows a steady weekly rhythm:

Sunday 1:00 PM – 4:00, Study Hall with April Chen, Adams Dining Hall.
Sunday, 3:00 PM – 6:00, Study Hall with Seeam Noor, Eliot Dining Hall.
Sunday, 7:00 PM – 10:00, Study Hall with Jack Schroeder, Pfoho Dining Hall.
Monday 6:00 PM – 9:00 PM, Study Hall with Shivani Aggarwal, Smith Center.
Monday 8:00 PM – 11:00 PM, Study Hall with Rucha Joshi, Leverett Dining Hall.
Monday 11:55 PM, DataCamp exercises due.
Tuesday 1:30 PM – 2:45 PM, Class.
Tuesday 3:00 – 6:00 PM, Study Hall with Joshua Pan, Fisher Commons.
Tuesday 5:00 – 8:00 PM, Study Hall with Roger Zhang, Smith Center.
Tuesday 7:30 – 10:30 PM, Study Hall with Evelyn Cai, Lowell Dining Hall.
Tuesday 8:00 – 11:00 PM, Study Hall with Katie Cao, Eliot Dining Hall.
Wednesday 6:00 – 9:00 PM, Study Hall with Yao Yu, Cabot Science Library.
Wednesday 8:00 – 11:00 PM, Study Hall with Dillon Smith, Smith Center.
Wednesday 11:55 PM. Problem set due.
Thursday 8:30 – 11:00 AM. Office Hours with Preceptor, Fisher Commons.
Thursday 1:30 PM – 2:45 PM. Class.
Thursday 3:00 – 6:00 PM, Study Hall with Amy Tan, Fisher Commons.
Thursday evening. Problem set due next week will be distributed.
Friday 11:55 PM. Final project milestones are due.
Sunday 11:55 PM. Exams, if distributed, are due.

We generally pair program on Tuesdays and have speakers on Thursdays.

Key Dates

Part 1: Tools and Framework

DataCamp #1 due Monday, February 3.
Problem Set #1 due Wednesday, February 5. Completed together in class.
Final Project Milestone #1 due Friday, February 7.
DataCamp #2 due Monday, February 10.
Problem Set #2 due Wednesday, February 12.
Final Project Milestone #2 due Friday, February 14.
DataCamp #3 due Monday, February 17.
Problem Set #3 due Wednesday, February 19.
Exam #1 distributed on Thursday morning, February 20 and due Sunday, February 23.

Part 2: Sampling and Inference

DataCamp #4 due Wednesday, February 26.
Final Project Milestone #3 due Friday, February 28.
Problem Set #4 due Wednesday, March 4.
Final Project Milestone #4 due Friday, March 6.
Problem Set #5 due Wednesday, March 11.
Final Project Milestone #5 due Friday, March 13.
Spring Break. Problem Set #6 due Wednesday, March 25.
Exam #2 distributed Thursday morning, March 26 and due Sunday March 29.

Part 3: Models

Final Project Milestone #6 due Friday, April 3.
Problem Set #7 due Wednesday, April 8.
Final Project Milestone #7 due Friday, April 10.
Problem Set #8 due Wednesday, April 15.
Exam #3 distributed Thursday, April 16 and due Sunday, April 19.
Final Project Milestone #8 due Wednesday, April 22.

Part 4: Conclusion

Demo Days
Last Day of class is Tuesday, April 28.
Final project due Friday, May 1.
Exam #4 distributed Saturday, May 2 and due Sunday, May 10.

Schedule

Part 1: Tools and Framework

Data science involves both inputs and outputs. We bring in data from somewhere to analyze and, once we have some answers, distribute our results. During Part 1, we will bring in data from R packages, downloaded text files and files which reside on the web. We will distribute our results as html files to the course staff, requests for help (from strangers) using reproducible examples and animated graphics posted to the web.

Week 1: January 27

Shopping Week

You are Ulysses. I am the rope.

Install R, RStudio and Git on your laptop. Start on the DataCamp assignments. They are due on Monday, February 3 at 11:55 PM. Sign up for a meeting with a TF. This will fulfill the first milestone, due February 7, for the final project. We will use RStudio Cloud on Tuesday and individual laptops on Thursday. Although it is not officially due till Monday, please try to do Introduction to the Tidyverse before Thursday's class.

Week 2: February 3

Visualization

You can never look at your data too much. – Mark Engerman

We will learn how to create an R project in RStudio. The first problem set will be distributed on Tuesday, via Github Classroom, and completed during class. We will also learn how to recover from git mistakes.

Assignments

Readings: Chapters 1 through 3: Getting Started, Visualization and Productivity.

DataCamp #1: Introduction to the Tidyverse, Introduction to Shell for Data Science. Chapter 1, Manipulating files and directories. Introduction to Git for Data Science. Chapter 1, Basic workflow. Introduction to Data Visualization with ggplot2. Communicating with Data in the Tidyverse, Chapter 3, Introduction to RMarkdown.

Remember: DataCamp assignments are due Monday at 11:55 PM.

Problem Set #1 due February 5 at 11:55 PM. We will complete and submit this problem set in class on Tuesday. Its purpose is to ensure that everyone has a working computer, understands Git/GitHub and can compile an R Markdown document.

Final Project Milestone #1 due Friday, February 7. Speak with a Teaching Fellow about your final project. Bring your laptop. Submit information via Canvas. No need to prepare. But it is important to start thinking about what you want to do. Google Dataset Search is a good way to find data. See also these resources.

Optional: RStudio Essentials Videos. Most relevant for us are "Writing code in RStudio", "Projects in RStudio" and "Github and RStudio". Again, these are optional! But they are very useful for students who find find traditional lectures to be a helpful supplement to classroom practice. See also *GitHub Classroom Guide for Students*. If you love DataCamps, you might also enjoy Working with the RStudio IDE (Part 1) and Working with the RStudio IDE (Part 2), especially chapter 2 on version control.

Speaker: February 6: Liberty Vittert, Harvard University.

Week 3: February 10

Seeking Help

The best data science superpower is knowing how to ask a question. – Mara Averick

We will learn how to produce a **repr**oducible **ex**ample — a "reprex" — in order to help strangers to help us. We will introduce the "potential outcomes" framework and review the fundamental problem of causal inference. We will discuss the slogan "no causation without manipulation."

Assignments

Readings: Chapters 4 and 5: Wrangling and Tidy. Appendix: Rubin Causal Model.

DataCamp #2: Working with Data in the Tidyverse. Note that `pivot_wider/pivot_longer` have replaced `spread/gather` as the preferred approach to reshaping tibbles.

Problem Set #2 due Wednesday, February 12.

Final Project Milestone #2 due Friday, February 14. Github repo with Rmd (and knitted html) which discusses pros and cons two projects from past years. At least one project should be one which did extensive data gathering/cleaning. You should not select the same projects for commentary as your friends have. Students generally write about a paragraph for each project. The Rmd/html file should include the url for your repo. The only thing you are submitting is the html, via Canvas.

Optional: RStudio Webinar on Reprex. Again, these are optional! But they are very useful for students who find find traditional lectures to be a helpful supplement to classroom practice. Causality, Chapter 2 of *Quantitive Social Science* by Kosuke Imai, especially pages 46 – 63. Working with Web Data in R.

Speaker: February 13: Mara Averick, RStudio.

Week 4: February 17

Maps

Workflow: you should have one. – Jenny Bryan

Assignments

Readings: Appendix: Maps and *The Cognitive Style of Powerpoint* by Edward Tufte.

DataCamp #3: Joining Data in R with dplyr.

Problem Set #3 due Wednesday, February 19.
Exam #1 distributed on Thursday morning, February 20 and due Sunday, February 23.

Optional: Introduction to Mapping with sf and *The Unix Workbench*, chapters 1 – 6. wikitablr seems interesting.

Speaker: February 20: David Sparks, Director of Basketball Analytics for the Boston Celtics. Starts at 1:30.

Part 2: Sampling and Inference

Week 5: February 24

Probability and Bayes

I stopped teaching frequentist methods when I decided they could not be learned. – Donald Berry

We will roll some dice and write some functions.

Assignments

Readings: Chapters 6, 7 and 8: Functions, Probability and Bayes; Appendix on Shiny.

DataCamp #4: Introduction to Function Writing in R.

Because of the exam, DataCamp is not due till Wednesday at 11:55 PM.

Final Project Milestone #3 due Friday, February 28. Create a new Github repo. This is the first version of your final project. (There will be many more to come.) Write an Rmd which provides a draft of your About page. (Naming it about.Rmd is wise.) Knit that Rmd into an html and submit the html via Canvas. The Rmd should include the url to your repo, should we want to examine it. Discuss all your data sources in the Rmd. (If you are gathering Harvard data, you should have a draft of your survey questions.) With luck, you will have gathered all your data and placed it in the repo. (This will generally be done with a different Rmd, like gather.Rmd, in your repo which contains the code which actually downloads your data.) You should have processed your data. (It is OK if you have not gotten quite this far as long as you discuss your progress and your plan in the About page.) Remember: You must gather data from two or more different sources. Learning how to source, clean and combine data is one of the goals of the project. On almost any topic, there are useful tables of information on Wikipedia. See here and here for advice.

Optional: Statistical Rethinking: A Bayesian Course with Examples in R and Stan (pdf) by Richard McElreath. Chapter 1.

Speaker: February 27: Will Kurt, Data Science Lead at Hopper.

Week 6: March 2

Sampling

Lot of points were taken off for small errors that I did not see as pedagogically important. – Gov 1005 student

Assignments

Readings: Chapter 9: Sampling.

Problem Set #4 due Wednesday, March 4.

Final Project Milestone #4 due Friday, March 6. Create a rough Github repo, with at least some of your raw data *or* code which shows you working with data that is stored elsewhere *or* details of your plan to get the data. The repo must also include a reproducible html which provides a brief description of the data: where you got it, what you have done with it so far and what you plan to do. You may change your project completely, all the way until Demo Day. But you are still responsible for meeting these milestones, even if you know you are going to pivot. Your data can not be from a single source. Typing **library(fivethirtyeight)** is not enough!

Optional: RStudio Webinar titled "How to Work with List Columns" by Garrett Grolemund. Background reading about anonymous functions in R. Try Foundations of Functional Programming with purrr for more advanced usage of the **purrr** package.

Week 7: March 9

Confidence Intervals

Last week before Spring Break.

Comment as a service to the dumbest possible version of your future self. – Alex Albright

Assignments

Readings: Chapter 10: Confidence Intervals. "Causal effect of intergroup contact on exclusionary attitudes" by Ryan Enos. PNAS March 11, 2014 111 (10) 3699-3704. link

Problem Set #5 due Wednesday, March 11.

Final Project Milestone #5 due Friday, March 13. Create a beautiful graphic, using **ggplot2** or another package of your choice, which uses some of your data.

Optional: Two videos about permutation tests along with a traditional textbook treatment, pages 16-41ff. Also, this pretty animation.

Speaker: Dishant Rana, DataCEVA.

Week 8: March 23

Shiny and Animation

We will learn how to make engaging animations.

Assignments

Readings: Appendices: Animation and Shiny.

Problem Set #6 due March 25.

Exam #2 distributed Thursday morning, March 26 and due Sunday March 29.

Optional: How to Start Shiny video tutorial. "Let's Take the Con Out of Econometrics," by Edward E. Leamer. The American Economic Review, Vol. 73, No. 1 (March, 1983), pp. 31-43. link.

Part 3: Models

Week 9: March 30

Regression

Fitting is easy. Prediction is hard. – Richard McElreath

Assignments

Readings: Chapter 11: Regression.

Final Project Milestone #6 due Friday, April 3. You must have a working Shiny app, just to demonstrate that you can get something up and running. It can be a mess, but it should have at least one graphic with your data.

Optional: Shiny tutorials. Modeling with Data in the Tidyverse.

Speaker: Stefanie Costa Leabo, Chief Data Officer, City of Boston.

Week 10: April 6

Multivariate Regression

Amateurs test. Professionals summarize.

Assignments

Readings: Chapter 12: Multiple Regression.

Problem Set #7 due Wednesday, April 8.

Final Project Milestone #7 due Friday, April 10. Cleaned up Github account.

Speaker: Alex Albright, Harvard University.

Optional: "The Bayesian New Statistics" by John K. Kruschke and Torrin M. Liddel.

Week 11: April 13

Classification

Assignments

Readings: Chapter 13: Classification.

Problem Set #8 due Wednesday, April 15.

Exam #3 distributed Thursday, April 16 and due Sunday, April 19.

Optional: Video lectures of generalized linear models with binary data, parts 1, 2 and 3. Multiple and Logistic Regression in R.

Week 12: April 20

Machine Learning

Put your work on the web. – David Sparks

Assignments

Readings: Chapter 14: Machine Learning.

Final Project Milestone #8 due Wednesday, April 22. Working rough draft of your final project. Demo Day is still one week away, and you can completely pivot if you want, but you must have a fairly complete version of your current project: a Shiny app with your About page, your data and your model.

Optional: Video: Intro to Machine Learning with R and Caret. DataCamp: Machine Learning in the Tidyverse.

Part 4: Projects

Week 13: April 27

Wrap Up

A public portfolio of high quality work is better than a Harvard degree.

Last day of classes. Make memes, provide course feedback, discuss final projects and have fun!

Optional: Mastering Shiny by Hadley Wickham.

Important: Check your grades on Canvas, including your calculated late days. Any questions/complaints must be made before the last day of classes. After that, no changes will be made.

Demo Day Sessions All presentations will be in Tsai. Arrive on time. Dress is casual.

DD1: Monday, April 27 from 9:00 to 10:30 AM.
DD2: Monday, April 27 from 10:30 to 12:00 PM.
DD3: Monday, April 27 from 12:00 to 1:30 PM.
DD4: Tuesday, April 28 from 9:00 to 10:30 AM.

Class Room Seating

- Seating is organized, by campus geography, into several large "Groups" of 20 to 30 students: first years, Eliot House, Quadlings, et cetera. Details depend on enrollment.
- Students work in "Pairs" of two "Partners." Sometimes, this will be "side-by-side," each of you with a computer open, each writing code, but talking with each other throughout. Other times, we will "pair program," meaning just one computer open and both of you collaborating on a single project. You will work with a different partner every class.
- If you are the stronger student in a Pair, do not simply charge ahead. Instead, make sure that your Partner keeps up with you. Help each other! If you aren't talking with each other often, then you are doing it wrong. *There is no better way to learn than to teach.* The stronger student should type less and talk more.
- Besides your Partner, the students sitting immediately beside, behind and in front of you are members of your Circle that day. Introduce yourself to them when you/they arrive.

Record the name of your Partner in the Google sheet for the day and the names of your Circle in a different Google sheet. Each person does this, even though doing so leads to duplication. Also, in a different sheet, record the names of the students in your Circle. (Don't stress about spelling.)

Assignment Details

Participation

There are several ways to earn group participation points in class.

Imperator: Each Group will have a class Imperator, someone who helps to organize a study group, coordinate activities with other Groups, and so on. Imperators make everyone feel welcome, first by learning everyone's name and, then, by introducing classmates to each other. Imperators must be able to get to class 10 minutes early. Role of Imperators is ensuring that everyone in a Group knows each other. First years (Jessica Wu (jessicawu@college.harvard.edu) and Amy Zhou (amyzhou@college.harvard.edu)), Adams/Lowell/Quincy (Jenna Moustafa (jennamoustafa@college.harvard.edu)), and Eliot/Kirkland/Winthrop (Brendan Chapuis (chapuis@college.harvard.edu)).

Scribe: We need note-takers, four students for each day. Meet for a meal at least one week prior to your class, take a selfie and post to Piazza. They work separately, but will still be partnered with someone so they can participate in coding. After class, the four scribes get together and create one unified set of notes, which must be posted to Piazza before 11:55 PM that evening.

Welcome Committee: We organize a Welcome Committee of four students for each speaker. Meet for a meal at least one week prior to your class, take a selfie and post to Piazza. See here for other duties associated with this role.

Piazza: Answering your classmates questions on Piazza is the best way to earn participation points. Be a good class citizen! If you find a (meaningful!) typo in a problem set or exam, please post it to Piazza. The first student to do so earns many participation points.

Extra Help

Course assistants (but not teaching fellows) are available for one-on-one or small group meetings outside of their regularly scheduled Study Halls. These meetings may *not* be used to work on the next problem set. That is what regular Study Halls are for. The most common purpose of these meetings is to review the

questions/answers from past problem sets and exams, the better to set a solid foundation for students moving forward. A second purpose is to provide help for the final projects. Process:

1. E-mail a CA with whom you would like to work to see if they are available. Mention the material you hope to review and cc Preceptor, Claire Fridkin and your TF. (You only cc us on the first e-mail, not on subsequent back-and-forths.)
2. If that CA is too busy, try another CA.
3. Arrange with the CA a mutually agreeable time/location for the meeting.
4. Do not blow off the meeting! The CA will tell us and we will be upset.
5. After the meeting, the CA will e-mail us with an update on material covered.

Social Events

Socializing with students outside of class is fun. Joining/inviting me is optional and has no influence on your grade in the class, i.e., **it earns you no participation points.** The three main options are:

Restaurant Lunches: My wife and I host students in groups of 4 for lunch throughout the semester, sometimes via the Harvard Class-Room-to-Table program and sometimes on our own dime. Invitations to come. Dress is casual. Please be on time. The reservation will be under "Kane." Just go straight to the table.

House Lunches: I enjoy having lunch with you before class, either in the CGIS cafe, Annenberg or at your House.

Faculty Dinners: I enjoy attending faculty dinners, so feel free to invite me to yours. My only requirement is that you also invite the other students in the class who live in your House. It is often fun to take over a table with a group of 4 or 5 or . . . To do this, just post the fact that I am attending with you to Piazza, inviting other students in the house to e-mail you. Very few (often none) will, but I insist that you be inclusive. You then take responsibility for RSVP'ing to the House on behalf of all of us. Please send me (and all others joining us) an e-mail the morning of the dinner to confirm, including the start time.

Useful Links

Google sheets for Partners, Circle, Scribes, Welcome Committee, Final Projects, and Absences.

Overview of, and grading rubrics for, the problem sets and exams.

How we conduct Study Halls.

How to improve your Github account.

Possible data sources for final projects.

Detailed instructions for members of the Welcome Committee.

Technical advice which students should follow. Read this at least once before submitting Problem Set #2.

Follow this advice if you have computer problems.

List of my friends/acquaintences from the world of data science, all of whom are happy to talk with students in my class. Reach out to them!

Free R Books.

Acknowledgements

This course is inspired by STAT 545, created by the legendary Jenny Bryan. The pedagogical goals follow Don Rubin's vision. Some of the classroom exercises come from (*Statistical Inference via Data Science: A moderndive into R and the tidyverse*) by Chester Ismay and Albert Y. Kim. Slides were created via the R package **xaringan** by Yihui Xie. Many thanks to all the folks responsible for R, RStudio, Git, GitHub and DataCamp. This course would not be possible without their amazing contributions.