STS 101: Introduction to Data Studies

Fall 2020 Course Syllabus
Lecture: Tuesdays and Thursdays 12:10-2:00PM
Professor: Lindsay Poirier, lnpoirier@ucdavis.edu
Office Hours: Mondays, 11-1, 1258 Social Sciences and Humanities
TA: Alejandro Ponce de León, poncedeleon@ucdavis.edu

What is this course about?

Data is increasingly becoming a significant resource for understanding civic life and addressing civic problems. In this course, we will examine the cultural forces that impact the availability, structure, and governance of civic data, as well as the forms of activism and advocacy that emerge around data. As we analyze civic datasets about eviction, toxics, education, crime, municipal services, civic infrastructure, and water quality, we will think through the politics shaping how data gets defined, calculated, and visualized.

Why is this course important?

Data is a powerful tool for making claims - claims that can advance equity and social justice, as well as claims that can misrepresent civic issues and marginalize certain populations. To critique claims posed from data, we need to understand where data comes from and the techniques used to analyze and visualize it. This course will provide you with an opportunity to engage in analysis of civic issues, develop skill in responsible data analysis and interpretation, and critically think through what a dataset *represents*.

What should you be able to do by the end of this course?

1. Communicate how cultural, political, and technical forces shape the collection, categorization, presentation, and publication of data
2. Apply techniques in data manipulation, analysis, and presentation to produce narratives from data
3. Critically communicate how data analyses can be interpreted in the face of biases and knowledge gaps
4. Assess the array of narratives that can be drawn from a single dataset, and the techniques diverse stakeholders may use to produce those narratives.

COURSE TEXTS

REQUIRED

All required course texts will be posted to Canvas at least one week before the assigned reading.

SUGGESTED RESOURCES

Grolemund, Garrett, and Hadley Wickham. R for Data Science. https://r4ds.had.co.nz/

Machlis, Sharon. "Learn to Use R: Your Hands-on Guide." ComputerWorld. (with 60+ links to additional resources!)

What are the course policies?

Attendance

Attendance in class and labs is required. You may miss one class session or lab without penalty. After this, points will be deducted from your grade. Exceptions may be made in the case of emergencies at my discretion.

Participation

Thoughtful participation in class discussions is a key component of this class and requires that you come prepared to discuss the week's reading. In this course, you will be graded on the *quality* of your contributions to discussion.

Late Assignments

Late assignments will receive a 10% point deduction immediately. After this, an additional 10% will be deducted for each additional day late. Exceptions may be made in the case of emergencies at my discretion.

Academic Integrity

Any time you use the ideas, images, language, etc. of another, you must cite that individual. If you use the words of another author verbatim (word-for-word), you must indicate that by putting the words in quotation marks. As a UC Davis student, you are expected to know when and how to cite and paraphrase correctly. If you do not, ask me or your TA for help.

Accommodations

Please contact the UC Davis Student Disability Center to request accommodations. https://sdc.ucdavis.edu/

How will you be graded?

CRITERIA	PERCENTAGE
R Assignments	40%
Journal Entries	20%
Final Project	30%
Attendance and Participation	10%

What are the reading expectations for this course?

My expectation is that the readings for this course should take you no longer than 3 hours per week, leaving time to apply what you are learning in the reading to writing and coding prompts. If you find the reading is taking longer than this, I encourage you to talk with me or your TA about various reading strategies.

RESOURCES FOR SUCCESS

PIAZZA

In this course, you will be individual coding exercises, but I want you to think about data analysis and programming as a collaborative activity. When you are confronting errors in R or cannot figure out the correct syntax, I encourage you to post a question in the relevant folder on the Q&A platform Piazza. You may also use this space to post relevant news articles, blog posts, videos, job/internship postings, or interesting datasets or tools. Extra credit will be awarded for considerable engagement in Piazza.

I will do my best to respond to all posts, but I will not respond to any questions regarding code within 24 hours that an assignment is due.

HOW TO SIGN UP

Navigate to piazza.com/uc_davis/FILL/sts101 and enter access code **sts101**.

Assignments & Assessments

Data Field Journal Entries

Every other Thursday, I will assign a short writing prompt related to the course's topics. You should respond to the prompt in 200-300 words. Entries should be posted to Canvas before class the following Thursday. These entries will be graded on a 4-point scale based on the development of your arguments and your engagement with the course themes.

R Assignments

Throughout the quarter, you will be analyzing the College Scorecard dataset through a series of data analysis prompts. Every other week, I will assign two prompts. For each prompt, I expect you to:

1. Produce the code for analyzing the data in an R environment.
2. List two stakeholders that would be interested in the results of the analysis, and characterize their stakes in the results.
3. Examine the College Scorecard Data Documentation for variables used in the data analysis, and describe how the values were produced and any ambiguities that might emerge in the analysis based on how the values were defined or collected.
4. Summarize what the analysis represents.

JUPYTERHUB

All assignments can be completed in UC Davis's JupyterHub - a digital environment for producing R Notebooks. We will discuss in class how to log in, use the system, and submit assignments.

Final Project

In the final project, you will think through how data analysis techniques can be used to support diverse, and sometimes conflicting claims. You will select one dataset that we have worked with over the course of the quarter. (You may select a dataset that we have not worked with, but you need to have it approved by me at least 3 weeks before the final is due.) You will then produce three data tables or visualizations in R that support a claim you wish to make from the data and describe in 200-300 words how the outputs support your claim. I would then like you to think through how someone may use the same dataset to refute your claim. Select 4 of the following prompts to write about in 200-300 words each:

- **Select:** How might someone cherry-pick variables to produce a competing claim?
- **Filter:** How might someone subset the data to produce a competing claim?
- **Not filter**: How, in not filtering the data to a certain subset, might someone gloss over issues that we can only see when the data is zoomed in?
- **Group-by**: How might someone aggregate the data into groups in order to hide details we can only see at individual record levels?
- **Ungroup**: How might someone divide grouped portions of the data into individual records to hide issues that we can only see when the data is aggregated?
- **Plot**: How might someone use specific plotting techniques to produce a data visualization that would refute your claim?

Further details will be provided on Canvas.

Course Schedule

WEEK 1: HOW WE COUNT IS HOW WE DEFINE: COLLEGE SCORECARD

10/1

- Martin, Aryn, and Michael Lynch. 2009. “Counting Things and People: The Practices and Politics of Counting.”
- NASA. 2015. “When a Definition Makes a Forest Disappear.”

10/3

- Journal Entry Due: Choose three data systems that have impacted your life in some way. (When I say data system, I'm referring to a specific dataset or a system of data collection.) For each, describe the data system, and how it has afforded/constrained possibilities for you. Who collects this data? (You may need to look this up). Can you access the data? Can you modify it? Reflect on the experience of living with this data system.

WEEK 2: INTRODUCTION TO R ENVIRONMENT

10/8

- Attend Responsible Computing Event

10/10

- R Assignment Due: Create an account on JupyterHub & Piazza

WEEK 3: AMBIGUITIES IN THE COLLECTION AND CLEANING OF DATA: EVICTION LAB

10/15

- Ribes, David, and Steven J Jackson. 2013. “Data Bite Man: The Work of Sustaining a Long-Term Study.”
- Watch: The Eviction Lab. 2018. “The Eviction Epidemic.” https://www.youtube.com/channel/UCBWad9JqRTyRwnbU13YZ6ZA (~2 min)
- The Eviction Lab. 2018. “Methodology Report.” pp 1-9.
- Aiello, Daniela, Lisa Bates, Terra Graziani, Christopher Herring, Manissa Maharawal, Erin McElroy, Pamela Phan, and Gretchen Purser. 2018. “Eviction Lab Misses the Mark.”

10/17

- Journal Entry Due: Describe a specific example of a data collection practice that involves numero-politics. For the specific example you choose, please respond to the following questions: What decisions have to be made about what to include/exclude in the count? How do the people or things being counted get defined? Who has stakes in the outcome of the count, and what sorts of influence do they exert on the count? In what ways do numbers get contested? In what ways is this act of counting a political activity?

WEEK 4: FILTERING DATA & THE ROLE OF STAKEHOLDERS: TOXIC RELEASE INVENTORY

10/22

- Fortun, Kim. 2004. “From Bhopal to the Informating of Environmentalism: Risk Communication in Historical Perspective.”
- Watch: Environmental Protection Agency. 2016. “The Power of Community Right to Know.” https://www.youtube.com/watch?time_continue=77&v=Fqjh6t6Hx6s (~2 min)
- Environmental Protection Agency. 2013. “Toxic Release Inventory in Action: Media, Government, Business, Community, and Academic Uses of TRI Data.” pp 4-10.
- Hiar, Corbin. 2012. “EPA’s Toxics Release Inventory Doesn’t Offer Full Picture of Pollution.”

10/24

- R Assignment Due

WEEK 5: GROUPING DATA WHEN THE STATS HAVE BEEN JUKED: STOP, QUESTION, AND FRISK

10/29

- “Introduction.” Muller, Jerry. 2018. *Tyranny of Metrics.*
- Smith, Chris. 2018. “The Crime-Fighting Program That Changed New York Forever.”
- Denvir, Daniel. 2015. “The Missing Ingredient in Stop-and-Frisk Reform: Open Data”

10/30

- Journal Entry: Consider a data system that classifies people into a series of categories. Describe the data system. Into what categories does the data system classify people? Who came up with these categories, and how much control do individuals have over how they get classified? Are the categories stable, or do they change over time? What causes them to change? How do the categories shape identities and life paths?

WEEK 6: PLOTTING KNOWLEDGE GAPS: SF 311

11/5

- Liboiron, Max. 2015. “Disaster Data, Data Activism: Grassroots Responses to Representing Superstorm Sandy.”
- Johnson, Steven. 2010. “What a Hundred Million Calls to 311 Reveal About New York.”
- Vo, Lam Thuy. 2018. “They Played Dominoes Outside Their Apartment For Decades. Then The White People Moved In And Police Started Showing Up.” https://www.buzzfeednews.com/article/lamvo/gentrification-complaints-311-new-york.

11/7

- R Assignment Due

WEEK 7: VISUALIZING DATA & CRITIQUING CONVENTIONS: NATIONAL BRIDGE INVENTORY

11/12

- Kennedy, Helen, Rosemary Lucy Hill, Giorgia Aiello, and William Allen. 2016. “The Work That Visualisation Conventions Do.”
- Bergstrom, Carl T., and Jevin D. West. “Why Scatter Plots Suggest Causality, and What We Can Do about It.”
- Triplett, Jason, and Paul Jordan. “Keeping Us Safe: A Day in the Life of a Bridge Inspector.” http://www.sehinc.com/news/keeping-us-safe-day-life-bridge-inspector
- Lwin, M. Myint. 2007. “Public Disclosure of National Bridge Inventory (NBI) Data.” https://www.fhwa.dot.gov/bridge/nbi/20070517.cfm

11/14

- Journal Entry: Search online for a data visualization. First, describe your first impressions of the visualization. Where is your attention drawn first? What does the visualization seem to communicate? Second, describe some of the choices that the designers had to make in order to produce this visualization. What decisions did they make about what to include/exclude in the visualization, how to scale the visualization, or how to style it? What visualization conventions did the designers use to make the visualization persuasive?

WEEK 8: INTERPRETING DATA & RISK: VIOLATIONS TO HUMAN RIGHT TO WATER

11/19

- Ottinger, Gwen, and Rachel Zurer. “New Voices, New Approaches: Drowning in Data.” . https://issues.org/ottinger/
- EPA. 2009. “Drinking Water Enforcement Response Policy.” https://www.epa.gov/sites/production/files/2015-09/documents/drinking-water-erp-2009.pdf

11/21

- R Assignment Due

WEEK 9: DATA PRIVACY

11/26	- Sweeney, Latanya. 2018. *Keynote Address*. Cambridge, MA: Women in Data Science, Harvard University. https://www.youtube.com/watch?v=eLyPxzR29eI. - Review (*skim*): California Department of Health Care Service. 2016. "Data De-identification Guidelines (DDG)." https://www.dhcs.ca.gov/dataandstats/Documents/DHCS-DDG-V2.0-120116.pdf - Identify 3 recommended strategies for de-identifying individuals in the data and be prepared to discuss them in class.	11/28	- Thanksgiving Holiday

WEEK 10: COURSE WRAP-UP

12/3	- Journal Entry Due: Be on the lookout for a news article/clip (something published within the past month) that reports on a statistical claim. How are numbers being used in this article (to legitimize a claim? ...refute a claim? ...analyze a phenomenon?) Does the article make any causal claims? - If so, what causal claims does it make, and what language do the authors use to suggest causation? What evidence are the authors basing their causal claims off of? The data alone? Other scientific studies? Historical narratives? Describe a potential confounding factor which might impact how we think about the statistic being reported. - If it doesn't make causal claims, what language is used to portray what the data is showing? What strategies do the authors use to signal to the reader how they might interpret the data? (e.g. does the article list any confounding variables that might impact what is presented in the data?)	12/5	Final Project is due

WHO IS MY INSTRUCTOR?

I am a cultural anthropologist that studies how civic data gets produced and how communities think about and interface with data infrastructure. As a professor, I aim to provide students with opportunities to practice and hone skills in critical thinking, research, data analysis, and persuasion. My classes often include many thought exercises and hands-on group activities.