Course Syllabus
PSY 417: Big Data in Psychology
Connecticut College, Psychology Department, Fall 2023

Class time:
Lecture: Bill Hall 403, Tuesday/Thursday 10:25 – 11:40 AM
Lab: Charles E Shain Library DAVIS, Friday 9:00 – 11:45 AM

Instructor
Dr. Hyun Joon Park, hpark3@conncoll.edu, TEL: 860-439-2336
Office hours: Wednesday 9 AM – 11 AM in Bill Hall 315-1 or by appointment. You do not need to have specific questions to come to office hours. Feel free to come to the office hours to discuss any general concerns/questions about the course or just to say hi!

Communication
You can also contact Hannah and me through email with any questions/concerns regarding the course. I will answer your emails within two business days (at the latest). Please feel free to send me a reminder/follow-up email if I do not respond within this time frame. **Please include the course number (i.e., "PSY 417") in the title of your email so I can prioritize your email.** Again, please do not hesitate to reach out. There are no obvious or trivial questions/concerns, and we will try our best to address them!

Course Description

This is the course designed to introduce computational methods to conduct social science research. This course will help you acquire the basic quantitative and computational skills to research human behavior and health. At the end of this course, you should have the basic knowledge to 1) manage your data (e.g., wrangling and managing big data), 2) access data sets from web and public data sources (e.g., using Application programming interface [API] to access social media and to access available big datasets) and 3) use computational methods to test research ideas (e.g., topic modeling, word embedding and using some machine learning algorithms). At the end of the course, you will be well-positioned to study more advanced topics in computational methods and different computational languages to study human behavior and health. The focus of this course is largely conceptual rather than statistical/mathematical. The goal is for you to become skilled and thoughtful data analysts utilizing computational methods. To the extent that we focus on formulas, it will be to serve your conceptual understanding and application of the methods to apply in your own line of research. At the end of this class, I expect you to have a research project that you can continue to pursue as a paper or manuscript.

Pre-requisites
This course is designed for students who do not have experience using computational methods in social science. However, prior experience using R, R studio, and R Markdown should be helpful. **Students without those experiences will be encouraged to undertake independent learning before and during the early phases of the course.**

Course Materials

Textbooks: In this course, we will be using few textbooks and you can access them online.

- Wickham, H., & Grolemund, G. (2023). R for Data Science (2e). O'Reilly Media.
    - The access to this book is free and you can access it at: https://r4ds.hadley.nz/data-transform

- Bauer, C.P., Landesvatter, C. & Behrens, L (2023). APIs for social scientists: A collaborative review
    - The access to this book is free and you can access it at: https://bookdown.org/paul/apis_for_social_scientists/

- Matz, S. C. (2022). The psychology of technology: Social science research in the age of Big Data (pp. xvi-451). American Psychological Association.
    - Available online through One Search

- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2020). Big data and social science: data science methods and tools for research and practice. CRC Press.
    - The access to this book is free and you can access it at: https://textbook.coleridgeinitiative.org/

Software: As briefly noted above, in this course, we will be using R, a widely used language for data analysis, as the main computing language. It is free and open-source. You will have to install the programs noted below.

- R: computational language/platform for data analysis. (cran.r-project.org)
- R studio: Open-source integrated development environment (IDE) for R (www.rstudio.com)
- R will be available on almost all computers in the labs. Please see instruction on how to install R and R studio (it is a program that makes using R more user friendly) here.

Helpful Resources:

- U Penn A guide to mining tools and methods: **https://guides.library.upenn.edu/penntdm/r**
- Text Mining with R: A Tidy Approach: **https://www.tidytextmining.com/**

Class Requirements and Activities

Class Attendance/Participation (10%): Class attendance will be expected from you throughout the semester. I understand things can happen in your life! You can miss two classes without any excuses. However, if you miss any class without excuse after missing two classes, 10% of the total attendance/participation points will be taken off for each day of missingness. If your family or personal emergency or COVID-19 quarantine or illness causes you to miss class, please let me know in advance as soon as possible so that your absence will be excused. If you miss a class (for any reason), please make sure you catch up on the materials you missed (e.g., by getting notes from a classmate), and please feel free to contact me if you have any questions about the things you missed.

Lab Participation (30%): Participating in the lab is a critical part of this class. Every Friday, we will meet at the lab and go through a problem set together during the lab session. You are asked to submit the assignment before you leave the lab. Missing policy for lab is same as class attendance policy. If you miss the lab, you will be asked to make up and submit the problem sets. Even if you missed the lab with an excuse, if you do not submit the problem set we completed during the lab your participation points will be taken off. The 10% of total lab participation will be taken off if you fail to submit the problem set.

Problem Sets outside of Lab (10%): Aside from attending the lab, you will be asked to complete problem sets. Problem sets are designed to give you practice in analyzing data and applying statistical knowledge you have learned in your course. That is, problem sets will often ask you to run statistical analyses using R. For full credit, you are asked to transparently show the process of your work (i.e., codes/syntax) and their outputs (e.g., usage of R markdown is encouraged).

Class Project (50%)

In this course, I will ask you to come up with a testable research question/hypothesis and test it with the methods you learned in the course. The midterm and final exams will be substituted by this class project.

Final Project Proposal (15%): Instead of midterm, you will be asked to submit a summary of research ideas you plan to test in this course (i.e., 5-page double spaced). This summary should include: 1) research question, 2) rationale for this research question (i.e., why do you need to study this question?), 3) proposed methods to test this research question, and 4) expected results. This is due at the end of Week 9 (Fri 10/27).

Final Project (25%): Then, for the final project, you will be asked to test this research idea using the methods you learned in this course. You will be asked provide a full write-up of the research including the results and discussion write-up. You can build-up this final project by elaborating on the final project proposal that you have made.

Final Project Presentation (15%): Further, you will be asked to give a 4-minute presentation at the end of the course and present and discuss your findings.

APA format: When you write your results in your problem sets, please follow APA style (7th edition). Used versions of the manual can be ordered online. There are also helpful online resources.

Individual and Group Work

You are definitely encouraged to work in groups for the problem sets. However, you must submit your own individual write-ups for each problem set. You are not permitted to collaborate on the exams.

AI Guideline

You may use AI programs e.g., ChatGPT to help generate ideas and brainstorm. However, you should note that the material generated by these programs may be inaccurate, incomplete, or otherwise problematic. Beware that use may also stifle your own independent thinking and creativity.

You may not submit any work generated by an AI program as your own. If you include material generated by an AI program, it should be cited like any other reference material (with due consideration for the quality of the reference, which may be poor).

Any plagiarism or other form of cheating will be dealt with severely under relevant Connecticut College policies.

Course Schedule

Week 1: Introduction to Big Data/R & R Studio (Tue 8/29, Thu 8/31, Fri 9/1)

Tuesday: Introduction

Thursday: Discussing about Bigdata + R Basics I

Readings

- Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor et al. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060–1062.
- Oswald, F. L. (2020). Future research directions for big data in psychology. In S. E. Woo, L. Tay, & R. Proctor (Eds.). Big data in psychological research (pp. 427-441). Washington, DC: APA Books
- R for Data Science (2e): Read 1. Introduction – 5. Workflow

Lab1: Installing R and R studio on laptop // R Basics I

- We will install R and R studio together in lab
- We will go over though materials we learned in Read 1. Introduction – 5. Workflow

Week 2: R Basics I & II(Tue 9/5, Thu 9/7, Fri 9/8)

Tuesday: R Basics I

Readings

- Lawson, M. A., & Matz, S. C. (2022). Saying more than we know: How language provides a window into the human psyche. In S. C. Matz (Ed.), **The psychology of technology: Social science research in the age of Big Data** (pp. 45–85). American Psychological Association. https://doi.org/10.1037/0000290-003
- R for Data Science (2e): Read 6. Data tidying – 9. Workflow: Getting help

Thursday: Discussing about Bigdata + R Basics II

Readings

- R for Data Science (2e): Read 13. Data tidying – 16. Regular expressions,

Lab2: R Basic I + II

- We will go over though materials we learned this week!

Week 3: R Basics III & IV (Tue 9/12, Thu 9/14, Fri 9/15)

Tuesday: R Basic III

Readings

- R for Data Science (2e): Read 20. Joins & 25. Web Scraping
- Khambatta, P. (2022). What Our Data Reveal About Our Minds: Predicting Psychological Characteristics From Digital Footprints. In S. C. MATZ (Ed.), **The Psychology of Technology: Social Science Research in the Age of Big Data** (pp. 9–44). American Psychological Association. http://www.jstor.org/stable/j.ctv2n4w5cj.6

Thursday: R Basic III

Readings

- R for Data Science (2e): Read 26. Functions & 27. Iteration

Lab3: R Basic III

- We will go over though materials we learned this week!

Week 4: R Basics Final (Tue 9/19, Thu 9/21, Fri 9/22)

Tuesday: R Basic IV (Work flow, Logical vectors, dealing with numbers, strings)

Readings

- Chapters 13 - 15

Thursday: Hierarchical Data

- Chapter 24

Lab4: Web scraping!

- Chapter 25

Week 5: Accessing Project Implicit data + Census data (Tue 9/26, Thu 9/28, Fri 9/29)

Tuesday: Accessing Project Implicit data

Readings

- Charlesworth, T. E., & Banaji, M. R. (2023). Evidence of Covariation Between Regional Implicit Bias and Socially Significant Outcomes in Healthcare, Education, and Law Enforcement. In *Handbook on Economics of Discrimination and Affirmative Action* (pp. 593-613). Singapore: Springer Nature Singapore.
- Stelter, M., Essien, I., Sander, C., & Degner, J. (2022). Racial bias in police traffic stops: White residents' county-level prejudice and stereotypes are related to disproportionate stopping of Black drivers. *Psychological science*, *33*(4), 483-496.
- Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Blacks' death rate due to circulatory diseases is positively related to whites' explicit racial bias: A nationwide investigation using project implicit. *Psychological science*, 27(10), 1299-1311.
- Project implicit (link)

Thursday: Census data

- R package censusapi (https://cran.r-project.org/web/packages/censusapi/vignettes/getting-started.html)

Lab4: Accessing Project Implicit data + Census data

Week 6: BRFSS and YRBSS datasets (Tue 10/3, Thu 10/5, Fri 10/6)

Tuesday: Introduction to Reddit Scraping

Readings

- APIs for social scientists: A collaborative review: Ch. 19 Reddit API
- Proferes, Nicholas, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics." *Social Media + Society* 7 (2): 205630512110190. https://doi.org/10.1177/20563051211019004.

Thursday: BRFSS (Behavioral Risk Factor Surveillance System) data & YRBSS (Youth Risk Behavior Surveillance System) data

Readings

- Park, H. J., Francisco, S., Pang, R., Peng, L., Chi, G. (2023). Exposure to Anti-Black Lives Matter Movement and Obesity of the Black Population. Social Science & Medicine. 316, 114265. doi.org/10.1016/j.socscimed.2021.114265.
- BRFSS (link)

- Galán, C. A., Stokes, L. R., Szoko, N., Abebe, K. Z., & Culyba, A. J. (2021). Exploration of experiences and perpetration of identity-based bullying among adolescents by race/ethnicity and other marginalized identities. JAMA network open, 4(7), e2116364-e2116364.
- YRBSS (link)

Lab5: Reddit data!

Readings

- Gaudette, Tiana, Ryan Scrivens, Garth Davies, and Richard Frank. 2021. "Upvoting Extremism: Collective Identity Formation and the Extreme Right on Reddit." *New Media & Society* 23 (12): 3491–3508. https://doi.org/10.1177/1461444820958123.

Week 7: API for Youtube (Tue 10/10, Thu 10/12, Fri 10/13)

Tuesday: Youtube I

Readings

- APIs for social scientists: A collaborative review: Ch. 23 Youtube API
- Obadimu, Mead, A. 2019. "Identifying Toxicity Within Youtube Video Comment." In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation.

Thursday: Youtube II

Readings

- Sangyeon Kim, Omer Yalcin, Sam Bestvater, Kevin Munger, Burt Monroe and Bruce Desmarais, 2020, "The Effects of an Informational Intervention on Attention to Anti-Vaccination Content on YouTube." In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp.949-953)

Lab6: Scraping Youtube posts!

Week 8: Scraping Twitter (or X) (Tue 10/17- No class, Thu 10/19, Fri 10/20)

Tuesday: Twitter

- TBD

Week 9: Cleaning Text data/Sentiment analysis (Tue 10/24, Thu 10/26, Fri 10/27)

Tuesday: How to clean data/Basics of text analysis

- Introduction to regular expressions and cleaning text (i.e., removing punctuation, number, symbols, stopwords, emojis)
- Text as Data Methods in R - Applications for Automated Analyses of News Content. Read Ch. 9 - 11 (https://bookdown.org/valerie_hase/TextasData_HS2021/)

Thursday: Sentiment analysis

- R package stringr (https://stringr.tidyverse.org/)
- R package SentimentAnalysis (https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html)
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

Lab7: Cleaning Text data/Sentiment analysis!

Final Project Proposal (15%) due!

Week 10: Topic modelling (Tue 10/31, Thu 11/2 – No class, Fri 11/3)

Tuesday: Topic modelling

- Introduction to regular expressions and cleaning text (i.e., removing punctuation, number, symbols, stopwords, emojis)
- Text as Data Methods in R - Applications for Automated Analyses of News Content. Read Ch. 13 (https://bookdown.org/valerie_hase/TextasData_HS2021/)
- Lee, C. S., & Jang, A. (2021). Questing for justice on Twitter: Topic modeling of# StopAsianHate discourses in the wake of Atlanta shooting. Crime & Delinquency, 00111287211057855.

Thursday: No Class

Friday: Practicing topic modeling!

Week 11: Word Embedding (Tue 11/7, Thu 11/9, Fri 11/10)

Tuesday: Words Embedding

- Introduction to word embedding (i.e., technique for identifying similarities between words in a corpus to predict the co-occurrence of words)
- Visualizing the results of word embedding

Readings/Materials

- Supervised Machine Learning for Text Analysis in R: https://smltar.com/embeddings, Read chapter 5.
- R package word2vec (https://cran.r-project.org/web/packages/word2vec/readme/README.html)
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644.

Thursday: Words Embedding II

Readings/Materials

- Xu, H., Zhang, Z., Wu, L., & Wang, C. J. (2019). The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. PloS one, 14(11), e0225385.

Friday: Practicing word embedding!

Week 12: Machine Learning Algorithms (Tue 11/14, Thu 11/16, Fri 11/17)

Tuesday: Machine Learning Algorithms

Readings/Materials

- TBD

Thursday: Machine Learning Algorithms II

Readings/Materials

- TBD

Friday: Machine Learning Algorithms

Week 13: Chat GPT (Tue 11/21, Thu 11/23- No class, Fri 11/24- No class)

Tuesday: Using Chat GPT to classify data

Readings/Materials

- Tutorials**:** https://www.youtube.com/watch?v=Mm3uoK4Fogc&t=26s

- Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis.

Thursday: No class
Friday: Practice chat GPT

Week 14: Catch up week (Tue 11/28, Thu 11/30, Fri 12/1)
Tuesday: TBD
Readings/Materials
Thursday: TBD
Friday: TBD

Week 15: Catch up week/Presentations (Tue 12/5, Thu 12/7, Fri 12/8)
Tuesday: TBD - Catch up week
Readings/Materials
Thursday: TBD - Catch up week
Friday: Presentations

Week 16-17: Exam period (12/12 Tues -12/13 Weds, 12/18 class ends).