Data Science for Psych & Neuro

CMU 85-432 & 85-732

Instructor Info

Roberto Vargas, PhD

Virtual Office Hrs: Mondays 1-2pm

Zoom

tinyurl.com/braindatascience

robertov@andrew.cmu.edu

Course Info

Prereq: 85-309 or 36-309

Mon & Wed

2:00-3:20 EST

BH336B & Remote

Zoom

TA Info

Larry Jiang

Virtual Office Hrs: Thurs 11a-12pm

Zoom

ruitongj@andrew.cmu.edu

Overview

Data science is a form of quantitative epistemology. As scientists, we use models of empirical observations in order to infer meaning from the data that we use to advance our knowledge and understanding of specific parts of the world.

To help you learn this process, this class will cover topics in philosophy, scientific theory, machine learning, data management, and statistics, specifically focused for applied research in modern psychology and neuroscience. Emphasis will be placed on fundamental data science theory that can support learning more complex analytical methods, as well as basic applied skills for performing data analysis in a research context.

Topics include (but are not limited to):

- Github and version control
- Juypter notebooks & markdown
- Pipeline & analysis design
- Data organization & archiving
- Data visualization
- Linear regression models
- Data cleansing
- Reducible vs. irreducible error
- Logistic regression
- Linear/Quadratic discriminant analysis
- K-Nearest Neighbors
- Cross validation
- Bootstrapping
- Model selection
- LASSO & Ridge regression
- Feature selection
- Bayes factors
- Interpretation & inference

Learning Objectives

This is a flipped class. All lectures are pre-recorded. Students are expected to have viewed each lecture and finish the assigned readings prior to each class. Class time will consist of a structured Q&A period followed by either 1) small group discussions, designed to push the depth of particular topics, or 2) lab tutorials. After class, students are expected to submit answers to the assigned discussion question and homework answers (if assigned).

Successfully meeting the objectives of this course will allow you to:

1. understand basic principles of statistical theory, measurement, and experimental design;
2. be able to clean and organize data efficiently;
3. be well versed in the execution and interpretation of data analysis;
4. use information resources to find appropriate statistical tools;
5. communicate statistical results effectively in multiple modalities;
6. be a critical consumer of data science techniques and their application in empirical research.

FAQs

Do I need to know how to program?

Some experience with R is helpful, but otherwise you'll learn the coding you need along the way.

Can I use my own data?

The final project is meant to be something that interests you. So please feel free to use your own data (so long as you have the appropriate permissions to use the data).

What is the best statistic to learn?

There is no single best statistic to know. Each method is a specific tool for looking at your data a specific way. If you know the best form of your hypothesis, you'll learn to see the tools that best address it.

What is 'data science' anyway?

It's a hydra. A mythical creature formed out of the parts of other fields. The head is statistics. The body is data visualization. The hooves are philosophy. The tail is computer science.

Materials

Class Github Repository

`https://github.com/CoAxLab/DataSciencePsychNeuro`

Data Explorations (Jupyter book)

`https://coaxlab.github.io/Data-Explorations/intro.html`

Required Textbook

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R, 2nd Edition (Vol. 6). New York: Springer. (`https://www.statlearning.com/`)

Recommended Textbook

Hadley Wickham & Garrett Grolemund (2016). R for Data Science. O'Reilly. (`https://r4ds.had.co.nz/`)

Other Readings

Any required journal articles and book chapters will be provided on Canvas.

Lectures

All lectures will be posted on Canvas ahead of class discussions.

Necessary Tools

- A Github Account
- R/R Studio (version 4.0.3)
- Jupyter notebooks

Grading

40%	Exercises
30%	Discussion Questions
30%	Final Project

Exercises

After most tutorials in the Jupyter Book there are a short set of exercises meant to build on what was discussed in class, the readings, and the tutorial notebook. Your answers should be submitted as link to a filled out notebook hosted on the personal GitHub account you set up for the class. These homework assignments are due no later than 1 week after the associated class.

Discussion Questions

On some classes we will have break out group discussions on the topic of the day. At the beginning of the discussion, the group should select one member to summarize the group's discussion to the rest of the class. Otherwise, students will be assigned a discussion question to consider and answer on their own after class. You are responsible for submitting a short (approximately 1 paragraph) summary of the discussion over Canvas. *Students who cannot make the in-person class discussion must choose one discussion questions for that class and submit a short summary over Canvas no later than 48hrs after the missed class to receive full credit.*

Final Project

At the end of the semester, you will submit a final project consisting of a summary of a data set of your choosing. In most cases this will be data relevant to your research projects outside of class. All analysis, summaries, and visualizations will be submitted as Jupyter notebooks accessible on GitHub. Final projects must be submitted by 5 pm EST on May 1st, 2025.

The data set you wish to use for your final project must be approved by the instructor no later than March 1st, 2025. If you do not have access to a relevant data set from your research, please contact the instructor to find a reliable public data set for your project.

Late Work Policy

There is a 10% penalty per week for late homework assignments (e.g., 2 weeks late means 20% penalty). Homework submitted 3 weeks after original deadline or after the last day of the semester (whichever comes first) will not be accepted. Discussion questions not submitted within 48hrs of the in-person class time will not receive credit. Final projects submitted after the deadline will receive a 10% penalty and any project not submitted within 24hrs of the deadline will not receive any credit.

Academic Integrity

Cheating and plagiarism are defined in the CMU Student Handbook, and include (1) submitting work that is not your own for papers, assignments, or exams; (2) copying ideas, words, or graphics from a published or unpublished source without appropriate citation; (3) submitting or using falsified data; and (4) submitting the same work for credit in two courses without prior consent of both instructors. Any student who is found cheating or plagiarizing on any work for this course will receive a failing grade for that work. Further action may be taken, including a report to the dean.

Now what about the use of AI technology such as large language models (e.g., ChatGPT, Copilot, Bard)? I encourage students to use such models in the narrow context of assisting with their code development, if they find it helpful. Please declare such assistance at the bottom of each exercise assignment, where you are asked who you worked with. You will be fully responsible for any errors or mistakes that may occur from use of such models. Otherwise, using AI models to write any non-coding text for this class, beyond assistance with grammar and syntax checking, is strictly forbidden. Any discovered use of AI models for text content generation in the class will be treated as plagiarism as defined in the preceding paragraph.

Equal Opportunity Accommodations

All efforts will be made to minimize conflict with students' religious schedules (e.g., holidays, prayer services, etc.) and/or any disabilities. Students should consult with the Equal Opportunity Services (EOS) office at the beginning of the semester in order to setup any necessary accommodations for the class.

Respect in the Classroom

It is my intent to present materials and activities that are respectful to the diverse backgrounds and perspectives of students in the classroom. You may feel free to let me know ways to improve the effectiveness of the course for you personally or for other students or student groups. If you feel uncomfortable discussing this with me or your TA, you may voice your concerns to the Chair of the Department of Psychology Diversity and Inclusion Committee, Kody Manke *kmanke@andrew.cmu.edu*.

Self Care

Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep, and taking some time to relax. This will help you achieve your goals and cope with stress.

- All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.
- If you or someone you know experiences academic stress, difficult life events, or feelings of anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at `http://www.cmu.edu/counseling/`. Consider reaching out to a friend, faculty, or family member you trust for help getting connected to the support that can help.

Class Schedule

PHASE 1: Information & meaning

Week 1	Quantitative epistemology	Dretske, F. I. (1983). Précis of Knowledge and the Flow of Information. Behavioral and Brain Sciences, 6(1), 55-63. Vlastelica M. (2019). Learning Theory: Empirical Risk Minimization. Towards Data Science.
	The value of openness	Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean?. Science translational medicine, 8(341), 341ps12-341ps12. Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. PLoS Comput Biol, 9(10), e1003285.
Week 2	MLK Day	No Class
	What is a theory?	van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. Perspectives on Psychological Science. Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. Perspectives on Psychological Science, 16(4), 789-802.
Week 3	Models as testable hypotheses	A Survey of Some Fundamental Problems. In Popper, K. (1959). The logic of scientific discovery. Routledge. Guest, O., & Martin, A. E. (2021). On logical inference over brains, behaviour, and artificial neural networks. PsyArXiv
	Data as objects & architectures	Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1-23. Gorgolewski, et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific data, 3(1), 1-9.
Week 4	Techniques for data cleansing	Müller, H., & Freytag, J. C. (2003). Problems, Methods, and Challenges in Data Cleansing. Berlin: HUB-IB-164.
	Visualization as analysis	Cairo, A. (2012). The Functional Art: An introduction to information graphics and visualization. New Riders. Chapters 1 & 3. Yanai, I., & Lercher, M. (2020). A hypothesis is a liability. Genome Biology, 21, 1.
Week 5	Visualization through human eyes	Franconeri, S., L. Padilla, P. Shah, J. Zacks, & J. Hullman (2021). The science of visual data communication: What works. Psychological Science in the Public Interest.

PHASE 2: Knowledge

	The bias-variance tradeoff	James et al. Chapters 1 & 2
Week 6	Linear models	James et al. Chapter 3

	The ordinary least squares solution	James et al. Chapter 3
Week 7	Limits & variations of linear regression	James et al. Chapter 3
	Classifiers	James et al. Chapter 4
	Spring Break: No Classes	
Week 8	Mixed effects models	Bates, Douglas M. "lme4: Mixed-effects modeling with R." (2010): 470-474. Chapter 1 Yarkoni, T. (2022). The generalizability crisis. Behavioral and Brain Sciences, 45, e1.
	The beauty of kNN	James et al. Chapters 2 & 4
Week 9	Cross validation	James et al. Chapter 5
	Resampling methods	James et al. Chapter 5
Week 10	Mediation & moderation	Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior research methods, 40(3), 879-891.
	Power analyses via simulations	Beaujean, A. A. (2014). Sample size determination for regression models using Monte Carlo methods in R. Practical Assessment, Research, and Evaluation, 19(1), 12.
Week 11	Selecting the “best" model	James et al. Chapter 6
	Regularized regression	James et al. Chapter 6.
Week 12	Principal component methods	James et al. Chapter 6
PHASE 3: Understanding		
	Reconsidering the p-value	Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$". The American Statistician, 73(S1), 1-19.
Week 13	Bayes factor	Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. Psychonomic bulletin & review, 14(5), 779-804.
	Errors and inferences	Mayo, D. G. (1997). Error statistics and learning from error: making a virtue of necessity. Philosophy of Science, 64(S4), S195-S212.
Week 14	Telling your data story	Mensh, B., & Kording, K. (2017). Ten simple rules for structuring papers. PLoS Comput Biol, 13(9), e1005619.
	Theories as social constructs	De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. Synthese, 144(1), 137-170.
	Alternative: PhD to Industry Discussion Panel	Have tech industry people lead discussion about translating PhD skills to industry environments.