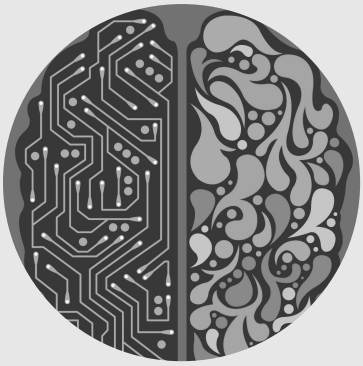

Data Science for Psych & Neuro

CMU 85-432 & 85-732

Instructor Info

Tim Verstynen PhD

Virtual Office Hrs: Thursdays 1:30-2:15pm

Zoom

tinyurl.com/braindatascience

timothyv@andrew.cmu.edu

Course Info

Prereq: 85-309 or 36-309

Tues & Thurs

2:20-3:40pm EST

Zoom

TA Info

Patience Stevens

Virtual Office Hrs: Mondays 3-4pm

Zoom

pstevens@andrew.cmu.edu

Overview

Data science is detective work. Like a good mystery, as a data scientist, you must learn the art of carefully deciphering the story that your evidence is trying to tell you.

To get you there, this class will cover topics in 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
- 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
- Overfitting
- Dimensionality reduction
- Decision trees
- Random forests
- Bagging & Boosting
- Bayes factors

Learning Objectives

This is a flipped class. All lectures are pre-recorded. Students are expected to have viewed each lecture, done the assigned readings, and finished the tutorials prior to each class.

Class time will consist of structured Q&A periods and guided small group discussions designed to push depth of knowledge on specific topics. Students requiring asynchronous learning should contact the instructor for special permission.

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`

Required Textbook

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer. (`http://www-bcf.usc.edu/~gareth/ISL/`)

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.

Tutorial notebooks

All tutorials will be provided as Jupyter notebooks on the class Github repository.

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%	Homework
20%	Discussion Questions
40%	Final Project

Discussion Questions

Each class 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 and a separate member should be responsible for submitting a short (approximately 1 paragraph) summary on the Discussion Summary Form. *Students who cannot make the in-person class discussion must choose one discussion question for that class from the course Github repository and submit a short summary to the Discussion Summary Form 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 presented, in class, as Jupyter notebooks accessible on GitHub. Final projects must be submitted by 12 pm EST on May 11th, 2021.

The data set you wish to use for your final project must be approved by the instructor no later than March 1st, 2021. 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.

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, Jessica Cantlon *jcantlon@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 anyone you know experiences any academic stress, difficult life events, or feelings like 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: Assessing the crime scene

Week 1	Art of data investigations	
	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	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. (2020). How computational modeling can force theory building in psychological science. PsyArXiv
	Constructing a testable hypotheses	van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. Perspectives on Psychological Science. Platt, J. R. (1964). Strong inference. Science, 146(3642), 347-353.
Week 3	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.
	Techniques for data cleansing	Müller, H., & Freytag, J. C. (2003). Problems, Methods, and Challenges in Data Cleansing. Berlin: HUB-IB-164.

PHASE 2: Evaluating the evidence

Week 4	No classes	
	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	The bias-variance tradeoff	James et al. Chapters 1 & 2
	Linear models	James et al. Chapter 3
Week 6	The ordinary least squares solution	James et al. Chapter 3
	Limits of linear regression	James et al. Chapter 3

Week 7	Mixed effects models	Bates, Douglas M. "lme4: Mixed-effects modeling with R." (2010): 470-474. Chapter 1 Yarkoni, T. (2019). The generalizability crisis. PsyArXiv
	Classifiers	James et al. Chapter 4
Week 8	The beauty of kNN	James et al. Chapters 2 & 4
	Cross validation	James et al. Chapter 5
Week 9	Randomization methods	James et al. Chapter 5
	Power analyses via simulations	TBD
Week 10	Mediation, moderation, & graphs	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.
	Selecting the “best" model	James et al. Chapter 6
Week 11	Regularized regression	James et al. Chapter 6.
	No classes	
Week 12	Principal component methods	James et al. Chapter 6
	Decision trees	James et al. Chapter 8
Week 13	Bagging, random forests, & boosting	James et al. Chapter 5
PHASE 3: Solving the mystery		
	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 14	Bayes factor: accepting the null	Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. Psychonomic bulletin & review, 14(5), 779-804.
	Telling your data story	Mensh, B., & Kording, K. (2017). Ten simple rules for structuring papers. PLoS Comput Biol, 13(9), e1005619.