

Description

Course 85-732

Lectures: Tuesdays & Thursdays, 3-4:20pm, 332P Baker Hall

Recitation: Tuesdays 4:30-5:30pm, 332P Baker Hall

Instructor

Dr. Timothy Verstynen

412-533-AXN1 (412-533-2961)

timothyv@andrew.cmu.edu

Office Hours: Thursday 2-3pm or by appt. (email to schedule), Baker Hall 335G

Teaching Assistant

Charles Wu

yunanw@andrew.cmu.edu

Recitation Session: Tuesdays 4:30-6pm

Course Overview:

Data science has sometimes been referred to as “data poetry.” It is the art of finding the story that your data tells and clearly conveying that story to others. This class will cover topics in machine learning and statistics necessary 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
- Jupyter 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

Lectures will focus heavily on theory while lab portions of the course will use the R statistical language to provide hands on data analysis experience. The goal of this class is to provide you with both the theoretical and practical knowledge for using modern data science tools in psychology research.

To meet this goal, this course is designed to balance two educational approaches:

1. Classroom instruction on statistical theory & methods.
2. Hands on experience with the statistical tools.

Learning Objectives

This course is designed to provide fundamental skill sets for use in applied psychological and neuroscientific research.

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 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.

Required Materials

Text & Materials:

- For a textbook, we will be using James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: springer. (<http://www-bcf.usc.edu/~gareth/ISL/>).
- Supplemental book: Hadley Wickham & Garrett Grolemund (2016). R for Data Science. O'Reilly (<https://r4ds.had.co.nz/>).
- Auxiliary readings will be posted on Canvas for class sections covering material not in the main textbook.

Tools:

1. **A physical notebook:** Understanding the statistical theory behind data science is a dynamic and creative process, not point and click. On many theory sections of the course, use of laptops will be discouraged unless needed for accessibility reasons, and use of hand written notes will be encourages (notes can be sketched out on tablets if that is easier too, this is the 21st Century after all).
2. **Jupyter Notebooks:** Final projects will be presented as a Jupyter notebook (<http://jupyter.org/>). It's like an electronic lab notebook and very useful for keeping track of what you're doing.

3. **Github:** You are expected to keep your final project materials organized on Github, a free version control repository. As students, you are able to register for the discounted student version which allows you to keep your repository private (see <https://education.github.com/>). The class github page is here: https://github.com/CoAxLab/DataSciencePsychNeuro_CMU85732
4. **R/R Studio:** All labs, homework, and take home tests will require analyzing data using the R statistical language (<https://www.r-project.org/>). R can be downloaded and installed for any operating system here: <https://cran.r-project.org/mirrors.html>. I recommend using R-Studio as an interactive data environment when playing with simulations.
 - It is critical that you use the latest version of R (v. 3.5). If you have an older version installed, please update.
 - R & R Studio will be installed in the Psychology Department's computer lab (332P Baker Hall) for students who do not have working laptops. This is the room we will use for all lab portions of the class.

Assessments:

- **Homework (50%):** A selection of conceptual and applied exercises will be assigned to evaluate understanding of material covered in lectures and labs. Much of this will be designed to help construct the Final Project (see below).
 - *There is a 10% penalty per week for late homework assignments.*
- **Participation (10%):** Data science is a team effort, requiring interaction with your peers and educators. As such you are regularly expected to attend classes and participate in exercises.
- **Final Project (40%):** At the end of the semester, you will present 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.
 - Half of the Final Project grade will be determined based off of the Jupyter notebook and half based off of the in-class presentation at the end of the semester. Therefore, not late assignments will be accepted after the last class of the semester.
 - The dataset you wish to use for your final project must be approved by the instructor **no later than** Feb 11th, 2020. 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.

Code of Conduct:

By enrolling this course, you agree to abide by the following codes of conduct.

- **Cheating & Plagiarism:** 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.
- Take care of yourself. 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](tel: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.