94-889: Machine Learning for Public Policy Lab
Fall 2021
Syllabus

Instructor Information

Rayid Ghani
rayid@cmu.edu, GHC 8023
Office Hours: Tue 12-1, Wed 2-3

Kit Rodolfa
krodolfa@cmu.edu, GHC 8018
Office Hours: Wed 11-12, Thu 12-1

Teaching Assistants

Riyaz Panjwani
rpanjwan@andrew.cmu.edu
Office Hours: Mon 12-1, Fri 10-11 (by GHC 8th Floor Printer)

Abhishek Parikh
akparikh@andrew.cmu.edu
Office Hours: Mon 11-12, Fri 2-3 (by GHC 8th Floor Printer)

Class Information

Lecture Time & Location: Tuesday and Thursday, 3:05-4:25pm, HBH 2008
Lab Time & Location: Wednesday, 6:20-8:00pm, HBH 1202
Website: https://github.com/dssg/mlforpublicpolicylab

Course Description

This is a project-based course designed to provide students training and experience in solving real-world problems using machine learning, with a focus on problems from public policy and social good.

Through lectures, discussions, readings, and project assignments, students will learn about and experience building end-to-end machine learning systems, starting from project definition and scoping, through modeling, to field validation and turning their analysis into action. Through the course, students will develop skills in problem formulation, working with messy data, communicating about machine learning with non-technical stakeholders, model interpretability, understanding and mitigating algorithmic bias & disparities, and evaluating the impact of deployed models.

Students will be expected to know python, and have prior coursework in machine learning.

Textbook & Software

Textbook: The course will rely on selected readings from various sources and has no required textbook – each week, we'll have selected readings from a variety of sources, listed below.

Software: For project work, we will provide students with access to a shared data and ML infrastructure. Data will be available in a postgreSQL database and SQL and python will be used throughout the course. Students will be expected to store project code in a shared github repository, so you should create an account if you do not already have one (github.com). Additionally, we will be making use of the machine learning pipeline package triage for modeling. **More Details to follow.**

Phone, Laptop, and Device Policy

Because much of the work in this course involves group discussions and responding thoughtfully to your colleagues' progress reports, mobile devices (including laptops, smartphones, tablets, blackberries, palm pilots, apple newton, and tamagotchi) are not permitted for use during the class. If you have a disability or other reason that necessitates the use of a mobile device, please speak to one of the instructors or teaching assistants.

Grading

Throughout the semester, students will work together in small groups on an applied machine learning project that will illustrate the concepts discussed in class and readings.

Graded components will include:

- Data loading exercise (5%)
- Written scope and proposal for their project work (10%)
- Peer reviews of three peer project proposals (2.5%)
- Midterm project update presentation (7.5%)
- Brief project progress update assignments (20%)
- Final group presentation of results targeted towards policy stakeholders (10%)
- Written final project report and code (20%)
- Quizzes on readings and lecture videos (5%)
- Class attendance and participation (15%)
- Submitting weekly check-in and feedback forms (5%)

The data used for the course projects should be considered sensitive and private and must remain in the secure computing environment provided for the course. **Any attempt to download any portion of the project data to a machine outside this environment will result in automatic failure**

of the class. Note that you may use tools like SQL clients, jupyter notebooks, etc. to interact with the data on the remote servers, but may not save the dataset (or a portion of it) to disk on a local machine.

Applied ML Project

Beginning in the second week of class, groups of 4 or 5 students will work together on a machine learning project throughout the semester with one of several real-world public policy problems. Each week, every group will be expected to provide an update on their current status. In addition to helping connect readings and discussion topics to an applied domain, these updates and discussions will give you a chance to elicit input and feedback from your classmates about challenges you're facing (and they likely are too!) in your analyses.

Throughout the semester, students will be responsible for several intermediate deliverables as they work on their group projects:

- An initial project proposal, submitted as a group, including the project scope and preliminary descriptive statistics about the entities in their dataset. The proposal should be 4-5 pages in length, not including figures, tables, or references and should include the provided scoping sheet as an appendix.
- A technical ML plan, submitted as a group, detailing how the scope described in their proposal can be formulated as a machine learning problem and the elements of the pipeline the group will be building.
- A description of specific features to be built for the modeling project, submitted as a group and describing the underlying data, level at which information is available, aggregation strategies (e.g., over time or geography), and plan for handling missing values.
- An in-class project mid-term update presentation (approximately 7 minutes in length plus 3 minutes for questions), describing the problem setting, approach, pipeline, and initial results.
- Brief weekly update assignments to guide our check-in discussions. These typically take the form of filling results or modeling details into a handful of template slides. These updates will be graded for completeness and correctness, however we expect this work to be iterative and errors identified in one week's update that are corrected by the next week can result in revision of the previous score up to 80% of the total possible.

At the end of the semester, each group will be responsible for a final presentation (10 minutes in length plus 3 minutes for questions). While the deep dive presentations should be more technical in nature, the final presentation should be geared towards the relevant decision makers for your project, including an overview of the problem and approach, your results, policy recommendations, and limitations of the work.

Accompanying the final presentation is a written report, approximately 15 pages in length, which should include:

- An executive summary not to exceed 1 page that succinctly describes the project, results, and recommendations.
- An overview of the problem, its significance, and the scope and goals of the current work.
- A description of the methodology and results of the analysis. The report should also provide a link to well-documented code in your group's course github repository.
- Brief (1-2 paragraph) design of a field trial to evaluate the accuracy of the resulting model in practice as well as its ability to help the organization achieve its goals.
- Concluding lessons and recommendations for the partner organization.
- Optionally, you may also wish to include a proposal for future avenues of research beyond the scope of this work, for instance on novel machine learning methods to improve on the current work, new policy interventions to evaluate or explore, or other related research opportunities.

Tentative Schedule

In general, the course will be structured around three sessions each week:

- During the Tuesday sessions, we'll focus on structured lectures and discussions of the weekly topic (including a mix of live lectures and discussions of pre-recorded content throughout the semester).
- During the Wednesday lab/recitation sessions, we'll discuss technical skills and tools you'll need for the project work early in the semester and then shift to check-ins with each team to discuss the status of their project work, generally surround short update assignments due on Monday (each team should review the updates of all teams working on the same project and the discussion will involve feedback from your peers and the instructors).
- Early in the semester, Thursday sessions will also focus on lectures and discussions, but once the projects are underway, most weeks will reserve this time for group meetings and project work (note that attendance on zoom is still mandatory at this time – one piece of feedback we received in the last iteration of the course was that many groups had trouble coordinating regular meeting times, so we wanted to find a way to dedicate some class time to help resolve this challenge).

  Although we're dedicating some time in class to work with your group, please note that successfully completing the project will require considerable work outside of class time as well and will constitute the majority of the "homework" for the course.

Below is a preliminary schedule of the course, including the readings that will be assigned for that week. Please be sure to have read and be prepared to discuss the readings before the specified class session. Most of these topics can be (and often are) the focus of entire courses and generally we'll only scratch the surface, but hopefully inspire you to delve deeper into areas that interest you (and you'll find plenty of open research questions in each). Optional readings are also listed for most sessions which may be of interest to students who wish to delve deeper in a given area as well as provide additional context for your related project work.

- **Week 1 (Aug 31, Sep 2): Introduction and Project Scoping**
  On Tuesday, we'll provide an introduction to the class, its goals, and an overview of the project options to help you decide what you're interested in working on for the remainder of the semester.

  During the Wednesday session, we'll help ensure everyone is set up to access the class technical resources.

  On Thursday, we'll talk about scoping, problem definition, and understanding and balancing organizational goals. Well before the outset of technical work, a decision needs to be made about whether a given policy problem can and should be addressed with machine learning: is the problem significant, feasible to solve with a technical approach, and of sufficient importance to policy makers that they will devote resources to implementing the solution? How will success be measured? How will (often competing) goals of efficiency, effectiveness, and equity be balanced?

  Required Readings for Thursday:

  - *Data Science Project Scoping Guide* Available Online
  - *Using Machine Learning to Assess the Risk of and Prevent Water Main Breaks* by Kumar, A, Rizvi, SAA, et al. KDD 2018. Available Online

  Optional Reading:

  - *Deconstructing Statistical Questions* by Hand, D.J. J. Royal Stat Soc. A 157(3) 1994. Available Online

- **Week 2 (Sep 7,9): Case Studies and Acquiring Data**
  This week, we'll organize groups and begin project work

  Due Tuesday, Sep 7: Data loading exercise with ACS data.

  Practical examples can provide a great way to gain an understanding of the nuance of applying machine learning to policy problems, so Tuesday will focus on a class discussion of a case study of a recent application, scoping the case together in breakout sessions.

  Required Reading for Tuesday:

  - *Fine-grained dengue forecasting using telephone triage services* by Rehman, NA, et al. Sci. Adv. 2016. Available Online

  During the Wednesday session, the we will lead a tutorial on using remote workflow tools for your class project.

  On Thursday, we'll delve into some of the details of acquiring data, protecting privacy, and linking records across data sources. Acquiring data from a project partner is often an involved process with a number of legal and technical aspects. Researchers need to understand how the data acquired may and may not be used (typically formalized in a data use agreement as well as underlying law) and ensure that the privacy of individuals in the dataset is protected (potentially both through access restrictions and techniques like anonymization). Once data has been acquired, it often needs to be transformed to ingest into the system used for analysis, records from multiple data sources linked, and data structured for further analysis.

During class on Thursday, we'll also talk a little bit about working together with your project team.

Optional Readings:

- *Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning* by Potash, E, et al. KDD 2015. Available Online
- *What Happens When an Algorithm Cuts Your Health Care* by Lecher, C. 2018. (The Verge) Available Online
- *Broken Promises of Privacy* by Ohm, P. UCLA Law Review. 2009. Introduction and Section 1. Available Online
- *Data Matching* by Christen, P. Springer (2012). Chapter 2: The Data Matching Process Available Online
- *Big Data and Social Science* edited by Foster, Ghani, et al. Chapter 4: Databases.

- **Week 3 (Sep 14,16): Data Exploration, Analytical Formulation, and Baselines**
  Work on your project during this week should include continuing to develop and refine your scope as you begin to explore the data. You'll also need to prepare and load some data into a database in order to make use of it in your modeling.

  Tuesday of this week will provide a crash course in exploratory data analysis. Data exploration is fundamental to developing an understanding of the nuances of the data and how the policy problem you initially scoped can be specifically formulated as a machine learning problem. This process involves generating and plotting summary statistics, exploring trends over time and understanding rapid changes in distributions, as well as identifying missing data and outliers. Typically, data exploration should involve considerable input from domain experts as you develop an understanding of how the data relates to the underlying generative process, as well as its idiosyncrasies and limitations.

  We'll also dedicate about 30 minutes during class on Tuesday for you to meet with your project teams and discuss your project scope.

  During the Wednesday session, we'll lead a tutorial about using GitHub and SQL for your project.

  On Thursday, we'll discuss analytical formulation of policy projects. Distinct from the initial scoping, a true analytical formulation of your policy problem can only come after you have developed an understanding of the data at hand, which in turn will often result in a greater understanding of the problem itself. Here, you'll ask how specifically your target variable (if relevant) is defined in the data, what types of information are available as predictors, and what baseline you'll be measure performance against. Very rarely is the appropriate baseline as simple as "random choice" or the population prevalence. Rather, it should reflect what would be expected to happen otherwise: perhaps a simple decision rule that an expert would come up with or even a pre-existing statistical model that the current effort is seeking to replace.

Required Readings for Thursday:

- *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations* by Obermeyer, Z., Powers, B., et al. Science. 2019. Available Online
- *Problem Formulation and Fairness* by Passi and Barocas. FAT* 2019. Available Online

Optional Readings:

- *Always Start with a Stupid Model, No Exceptions* by Ameisen, E. Medium. Available Online
- *Data Analysis, Exploratory* by Brillinger. Available Online
- *Create a Common-Sense Baseline First* by Ramakrishnan. Medium. Available Online
- *Data Science for Business* by Provost and Fawcett. O'Reilly. 2013. Chapter 2: Business Problems and Data Science Available Online

- **Week 4 (Sep 21,23): Machine Learning Pipeline Overview**
  At this point in your project work, you should be developing your initial end-to-end pipeline.

  Due Friday, Sep 24: Project proposal with scope and descriptive statistics.

  On Tuesday, we'll describe the components of typical machine learning pipelines. End-to-end ML Pipelines can quickly become unwieldy with several moving pieces and well-structured, modular code is often critical to detecting and fixing bugs in the process. This session will provide an overview of the pipeline, each underlying element, and some best practices for building them.

  Required Reading for Tuesday:

  - Review the lecture slides before class: Online

  During the Wednesday session, we'll talk about using `triage`, the machine learning pipeline toolkit we will use for the class project.

  On Thursday, you'll have time to work with your group on the proposal due this week as well as your initial pipeline.

  Optional Readings:

  - *Architecting a Machine Learning Pipeline* by Koen, S. (Medium) Available Online
  - *Meet Michelangelo: Uber's Machine Learning Platform* by Hermann, J and Del Balso, M. Available Online

- **Week 5 (Sep 28, Sep 30): Choosing Performance Metrics & Evaluating Classifiers, Part I**
  Pipeline development should be continuing in your project, with a focus on producing the simplest-possible version of the full system.

  Due Friday, Oct 1: Peer reviews of three project proposals.

In most cases, a vast array of methods — each with a number of tunable hyperparameters — can be brought to bear on your modeling question. How do you decide which models are better than others and how can you be confident this decision will carry forward into the future when the model is deployed? How should you balance considerations of performance and fairness when making these decisions? Are models that are performing similarly well giving similar predictions? What should you do if they are not? In this week, we'll begin to answer these questions, focusing on the choice of performance metrics.

Required Readings for Tuesday:

- *Transductive Optimization of Top k Precision* by Liu, LP, Dietterich, TG, et al. IJCAI 2016. Available Online

On Wednesday, we will lead tech sessions on using Python and SQL together.

Optional Readings:

- *Evaluating and Comparing Classifiers* by Stapor, K. CORES 2017. Available Online

- **Week 6 (Oct 5,7): Choosing Performance Metrics & Evaluating Classifiers, Part II**
By this week, your group should have a very simple version of an end-to-end pipeline with preliminary results for a single model specification.

Due Friday, Oct 8: Skeleton pipeline code/triage configuration file, one-sentence analytical formulation, and baselines.

This week, we'll continue our discussion from the previous week, focusing specifically on validation strategies that reflect how you want your model to generalize. In particular, we'll focus on the common case of modeling contexts with a strong temporal component where predicting into the future is desired, exploring how your choice of training and validation sets can reflect this context.

Required Readings for Tuesday:

- *Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure* by Roberts, DR, Bahn, V, et al. Ecography 40:2017. Available Online

On Wednesday, we'll start our regular group check-ins to provide feedback on your project progress and **on Thursday, we'll meet together as a class** to do a deep dive on temporal validation through a few class project examples.

Optional Readings:

- *Time Series Nested Cross-Validation* by Cochrane, C. Medium. Available Online
- *The Secrets of Machine Learning* by Rudin, C. and Carlson, D. arXiv preprint: 1906.01998. 2019. Available Online
- *Big Data and Social Science (2nd edition)* edited by Foster, Ghani, et al. Chapter 7: Machine Learning. Available Online

- **Week 7 (Oct 12,14): Feature Engineering and Imputation**
  Note: No classes on Thursday, October 14 for a mid-semester break.

  By this week, your group should have a very simple version of an end-to-end pipeline with preliminary results for a single model specification.

  In many real-world contexts, expressing domain expertise through thoughtful feature engineering can dramatically improve model performance by understanding what underlying factors are likely to be predictive and helping the model find these relationships. Likewise, most data sets you'll encounter in practice are littered with outliers, inconsistencies, and missingness. Handling these data issues in a smart way can be critical to a project's success.

  Required Reading/Watching for Tuesday:

  - Short Video Lecture and corresponding slides

  On Wednesday, we'll continue our group check-ins.

  Optional Readings:

  - *Missing Data Conundrum* by Akinfaderin, W. Medium. Available Online
  - *Feature Engineering for Machine Learning* by Zhang, A. and Casari, A. O'Reilly. 2018. Chapter 2: Fancy Tricks with Simple Numbers Available Online
  - *Missing-data imputation* by Gelman, A. Available Online

- **Week 8 (Oct 19,21): ML Modeling in Practice**
  During this week, your pipeline development and refinement should continue with a widening set of model specifications and features to explore.

  Due Monday, Oct 18: Technical modeling plan and detailed feature list

  On Tuesday, we'll cover some practical tips about building machine learning models for real-world projects: how should you think about what types of models to build? What hyperparameters should you explore and how do you design a hyperparameter grid?

  On Wednesday, we'll continue our group check-ins and on Thursday, you'll time to work with your project group.

  Required Readings:

  - *Three Pitfalls to Avoid in Machine Learning* by Riley, P. Nature. 527. 2019 (Comment) Available Online
  - *Top 10 ways your Machine Learning models may have leakage* by Ghani, R. et al. DSSG Blog. Available Online

  Optional Readings:

  - *Data Science for Business* by Provost and Fawcett. O'Reilly. 2013. Chapter 5: Overfitting and Its Avoidance Available Online

  - *Leakage in Data Mining* by Kaufman, S., Rosset, S., et al. TKDD. 2011. Available Online
  - *Why is Machine Learning Deployment Hard?* by Gonfalonieri, A. Medium. Available Online
  - *Overview of Different Approaches to Deploying Machine Learning Models in Production* by Kervizic, J. KDnuggets. Available Online

- **Week 9 (Oct 26,28): Choosing Performance Metrics & Evaluating Classifiers, Part III**
  At this point, your group should be continuing to refine and expand on your preliminary modeling results.

  Due Monday, Oct 25: Weekly project update with updated validation splits, features, and “version 0” baseline results.

  This week, we’ll return to our discussion of model selection, delving into the details of winnowing down a large number of model specifications to one or a handful that perform "best" for some definition of "best". In particular, we’ll focus on the common case of machine learning problems with a strong time series component and the desire to balance performance and stability in model selection.

  On Wednesday, we’ll continue our group check-ins and on Thursday, you’ll time to work with your project group.

- **Week 10 (Nov 2,4): Project Update Presentations**
  By this week, your group should have a preliminary set of “correct but crappy” results reflecting a relatively simple model grid and the features you prioritized to build as a first pass.

  Due Monday, Nov 1: Weekly project update with initial “version 0” results and a list of models and hyperparameters you’ll be running.

  This week, each group will give a presentation about the current status of their project, covering the problem setting, approach, and initial results. The presentations will be split between the Tuesday and Thursday Sessions. On Wednesday, we’ll continue our group check-ins and sometime this week we’ll hold an additional tech session with a deep dive on building modeling pipelines.

- **Week 11 (Nov 9,11): Model Interpretability**
  By this week, project work should be beginning to focus more heavily on evaluation, model selection, and interpretation.

  Due Monday, Nov 8: Weekly project update.

  Model interpretability can be thought of at two levels: global (how the model works in aggregate) and local (why an individual prediction came out as it did). This week, we’ll focus on some practical aspects and applications of interpretability at the two levels: understanding how a model is performing globally, what it means to compare this performance across model specifications, how these methods can help researchers debug and improve their models, build trust among stakeholders (including a growing legal movement towards a "right to explanation"), help those acting on model predictions understand when they should override the model with their judgement, and importantly help those actors decide not only on whom to intervene but suggest what sort of intervention to take.

Required Readings for Tuesday:

- *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission* by Caruana, R, et al. KDD 2015. Available Online
- *Why Should I Trust You? Explaining the Predictions of any Classifier* by Ribeiro, MT, Singh, S, and Guestring, C. KDD 2016. Available Online
- *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery* by Lundberg, SM, Nair, B, et al. Nature Biomed. Eng. 2018. Available Online

On Wednesday, we'll continue our group check-ins and on Thursday, you'll have time to work with your project group.

Optional Readings:

- *Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice* by Rudin, C, and Usutn, B. INFORMS Journal on Applied Analytics. 2018. Available Online
- TBD
- *Interpretable Classification Models for Recidivism Prediction* by Zeng, J, Ustun, B, and Rudin, C. J. Royal Stat. Soc. A. 2016. Available Online
- *Model Agnostic Supervised Local Explanations* by Plumb, G, Molitor, D, and Talwalkar, AS. NIPS 2018. Available Online
- *A Unified Approach to Interpreting Model Predictions* by Lundberg, SM and Lee, S. NIPS 2017. Available Online
- *Explainable AI for Trees* by Lundberg, SM, Erion, G, et al. arXiv preprint: arxiv/1905.04610. Available Online

- **Week 12 (Nov 16,18): Algorithmic Bias and Fairness, Part I**
  By this week, you should be finalizing your modeling results and beginning to look at bias and disparities in your models.

  Due Monday, Nov 15: Weekly project update.

  Just as important as assessing whether your model is making accurate predictions is determining whether it is doing so in a fair manner. But, what do we mean by fairness? How can you measure it and what can you do to mitigate any disparities you might find? Where in your pipeline can bias be introduced? (spoiler: everywhere). This week will provide a very brief introduction to the expansive field of algorithmic fairness.

  Required Readings for Tuesday:

  - *Fairness Definitions Explained* by Verma, S and Rubin, J. Available Online
  - *A Theory of Justice* by Rawls, J. 1971. Chapter 1: Justice as Fairness, pp. 1-19. Available Online

  - *Racial Equity in Algorithmic Criminal Justice* by Huq, A. Duke Law Journal. 2018. Available Online [Focus on sections: I.B.2, all of section II, III introduction, III.B, and III.D.3]

  On Wednesday, we'll continue our group check-ins and on Thursday, you'll time to work with your project group.

  Optional Readings:

  - *Is Algorithmic Affirmative Action Legal?* by Bent, JR. Georgetown Law Journal. 2019. Available Online
  - *Does Mitigating ML's Impact Disparity Require Treatment Disparity?* by Lipton, Z, McAuley, J, and Chouldechova, A. NIPS 2018. Available Online
  - *Equality of Opportunity* by Roemer, JE and Trannoy, A. 2013. Available Online

- **Week 13 (Nov 23): Algorithmic Bias and Fairness, Part II**
  Note: No classes on Wednesday, Nov 25, or Thursday, Nov 26, for Thanksgiving.

  During this week, your group should be continuing to investigate any disparities in your model results as well as performing any other necessary post-modeling analyses.

  Due Monday, Nov 22: Weekly project update.

  This week, we'll continue our discussion of bias and fairness with a very brief survey of practical considerations and open research questions in the rapidly-developing field.

  Required Readings for Tuesday:

  - *A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions* by Chouldechova, A, Putnam-Hornstein, E, et al. PMLR. 2018. Available Online
  - TBD

  On Wednesday, we'll continue our group check-ins and on Thursday, you'll time to work with your project group.

  Optional Readings:

  - *Equality of Opportunity in Supervised Learning* by Hardt, M. and Price, E. NIPS 2016. Available Online
  - *Classification with fairness constraints: A meta-algorithm with provable guarantees* by Celis, E, Huang, L, et al. FAT* 2019. Available Online
  - *Fairness Through Awareness* by Dwork, C, Hardt, M, et al. ITCS 2012. Available Online
  - *Fairness Constraints: Mechanisms for Fair Classification* Zafar, M, Valera I, et al. PMLR 2017. Available Online
  - *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments* by Chouldechova, A. Big Data. 2017. Available Online

- **Week 14 (Nov 30, Dec 2): Final Presentations**
  During this final week of classes, each group will give a presentation about their applied ML

project as described above. On Thursday, we'll use a little time to wrap up the class and briefly touch on some of the topics we didn't have time to cover (including field trials, model deployment, and monitoring/maintaining the system over time)

- **Finals Week (Date TBA): Final Report Due**
  Incorporating the results of your project work throughout the semester as well as feedback from your final presentation, each group will write a final project report (due date TBA) as described above.

More Resources

You may find a number of books useful as general background reading, but these are by no means required texts for the course:

- *Data Science for Business* by Provost and Fawcett
- *Big Data and Social Science* edited by Foster, Ghani, et al. Available Online
- *Practical Fairness: Achieving Fair and Secure Data Models* by Nielsen
- *Fairness and Machine Learning* by Barocas, Hardt, and Narayana
- *Weapons of Math Destruction* by O'Neil
- *Exploratory Data Analysis* by Tukey

Additionally, the Global Communication Center (GCC) can provide assistance with the written or oral communication assignments in this class. The GCC is a free service, open to all students, and located in Hunt Library. You can learn more on the GCC website: cmu.edu/gcc.

Your Responsibilities

Attendance: Because much of this course is focused on discussion with your classmates, attending each session is important to both your ability to learn from the course and to contribute to what others get out of it as well. As such, you'll be expected to attend every session and your participation will factor into your grade as described above. Should anything come up will require you to miss a class (illness, conferences, etc), please let one of the course staff know in advance.

Academic Integrity: Violations of class and university academic integrity policies will not be tolerated. Any instances of copying, cheating, plagiarism, or other academic integrity violations will be reported to your advisor and the dean of students in addition to resulting in an immediate failure of the course.

Data Security: As noted above, the data used for the project work in this course should be considered sensitive and care must be taken to protect the privacy of those in the dataset. The data must remain on the computing environment provided for the class and attempts to download it to any other machine will result in failure of the course.

Additionally, care must be taken to avoid accidentally committing any raw data, queries containing identifiable information, or secrets (key files, database passwords, etc) to github. Should this occur, or should you have any reason to believe your personal computer or private key has been compromised, you must immediately notify the course staff of the issue.

Resources

Students with Disabilities: We value inclusion and will work to ensure that all students have the resources they need to fully participate in our course. Please use the Office of Disability Resource's online system to notify us of any necessary accommodations as early in the semester as possible. If you suspect that you have a disability but are not yet registered with the Office of Disability Resources, you can contact them at access@andrew.cmu.edu

Health and Wellness: As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work.

All of us benefit from support during times of struggle. 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 almost always 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 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.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night:

CaPS: 412-268-2922
Re:solve Crisis Network: 888-796-8226
If the situation is life threatening, call the police
On campus: CMU Police: 412-268-2323
Off campus: 911

Discrimination and Harassment: Everyone has a right to feel safe and respected on campus. If you or someone you know has been impacted by sexual harassment, assault, or discrimination, resources are available to help. You can make a report by contacting the University's Office of Title IX Initiatives by email (tix@andrew.cmu.edu) or phone (412-268-7125).

Confidential reporting services are available through the Counseling and Psychological Services and University Health Center, as well as the Ethics Reporting Hotline at 877-700-7050 or www.reportit.net (user name: tartans; password: plaid).

You can learn more about these options, policies, and resources by visiting the University's Title IX Office webpage at https://www.cmu.edu/title-ix/index.html

In case of an emergency, contact University Police 412-268-2323 on campus or call 911 off campus.

Student Academic Success Center (SASC)

SASC focuses on creating spaces for students to engage in their coursework and approach learning through a variety of group and individual tutoring options. They offer many opportunities for students to deepen their understanding of who they are as learners, communicators, and scholars. Their workshops are free to the CMU community and meet the needs of all disciplines and levels of study. SASC programs to support student learning include the following (program titles link to webpages):

- Academic Coaching – This program provides holistic, one-on-one peer support and group workshops to help undergraduate and graduate students implement habits for success. Academic Coaching assists students with time management, productive learning and study habits, organization, stress management, and other skills. Request an initial consultation here.

- Peer Tutoring – Peer Tutoring is offered in two formats for students seeking support related to their coursework. Drop-In tutoring targets our highest demand courses through regularly scheduled open tutoring sessions during the fall and spring semesters. Tutoring by appointment consists of ongoing individualized and small group sessions.You can utilize tutoring to discuss course related content, clarify and ask questions, and work through practice problems. Visit the webpage to see courses currently being supported by Peer Tutoring.

- Communication Support – Communication Support offers free one-on-one communication consulting as well as group workshops to support strong written, oral, and visual communication in texts including IMRaD and thesis-driven essays, data-driven reports, oral presentations, posters and visual design, advanced research, application materials, grant proposals, business and public policy documents, data visualisation, and team projects. Appointments are available to undergraduate and graduate students from any discipline at CMU. Schedule an appointment on their website (in-person, zoom synchronous, or recorded video), attend a workshop, or consult handouts or videos to strengthen communication skills.

- Language and Cross-Cultural Support – This program supports students seeking help with language and cross-cultural skills for academic and professional success through individual and group sessions. Students can get assistance with writing academic emails, learning expectations and strategies for clear academic writing, pronunciation, grammar, fluency, and more. Make an appointment with a Language Development Specialist to get individualized coaching.

- Supplemental Instruction (SI) – This program offers a non-remedial approach to learning in historically difficult courses at CMU. It utilizes a peer-led collaborative group study approach to help students succeed and is facilitated by an SI leader, a CMU student who has successfully completed the course. SI offers a way to connect with other students studying the same course, a guaranteed weekly study time that reinforces learning and retention of information, as well as a place to learn and integrate study tools and exam techniques specific to a course. Visit the website to see courses with SI available here.