DATASCI 510 - Reasoning and Design for Data Science

Instructor: Kevin McAlister

Email: kevin.mcalister@emory.edu

Course Description: The role of a data scientist is evolving rapidly - data scientists are not only expected to manage data sources and perform analyses, but also come up with relevant questions and methods for testing those questions. These evolving demands require significant command of both technical skills (programming and implementation of methods) **and** an understanding of how various analysis methods work. In turn, data scientists are expected to be able to explain these complicated ideas to less technical audiences.

In this class, we will discuss methods for assessing questions in data science. Unlike your other classes, however, we will concentrate on the *logic* of different approaches. This emphasis means that the topics covered in this class will be question-driven: given a specific type of question, how could we answer it? A function of this approach is that the specifics of mathematics and statistical inference take a bit of a back seat - the methods discussed in this class will arise as a function of first principles related to the questions we are trying to answer. We'll break this class into 4 parts:

1) Asking questions with data and generating evidence, summarizing relationships between variables, describing variables, answering "What is?" questions

2) Answering "Why?", "How?", and "What if it changed?" questions: causal inference and counterfactual assessment, the logic of causation, bias, all-else-equal conditions

3) Randomized experiments and pseudo-randomization: randomization and design-based causal inference, "simple" randomized experiments, experimental validity, ethics of experimentation, extensions to randomized experiments, brief overview of observational methods with pseudo-randomization

4) Answering "What's next?" questions: comparing causal inference to predictive models, a theoretical introduction to statistical learning, out-of-sample assessment and truthful learning, overfitting, statistical learning for continuous outcomes, bias/variance tradeoffs, assessing OOS performance, and classification tasks. With time left over at the end of the semester, I hope to introduce the basics of generative machine learning and discuss its usage in the data science workflow.

Attendance and Communication: Communication is important. If you have a question then please email me. You are also encouraged to attend office hours. Since this class is small, I am also willing to schedule additional office hours, as needed. I will likely be slower on weekends and it is usually not a great idea to ask questions on a Friday night or right before something is due.

If your situation changes regarding health, housing, or in any other regard with respect to your ability to participate in the class, please contact the appropriate Emory student support organization first and then me as soon as feasible. It is easier for me to address your needs if I know about them as soon as they arise. This does not mean I can successfully respond to every request for consideration, but I emphasize that my goal is to treat you all equitably and do what I can to help you succeed in this course.

If you are not feeling well, please do not attend lectures in person! If you are sick, understand that I will be flexible about attendance and keeping up with work. If you expect that illness or other circumstances will prevent you from attending more than a single lecture, please make sure to email me so that we can discuss your individual circumstances. In order to remove the incentive for students to attend lectures while sick, I

will be simulcasting the lecture via the class Zoom room during the regular lecture time. I will also be recording each lecture and posting the recording to the course's Canvas site. Together, these two options should remove the incentive to come to class while sick.

However, these recordings are not intended to be a substitute for the actual lectures! Much of the content in this course centers on applying the concepts discussed in class and immersion in the class' materials for the full lecture period will greatly assist with retention and understanding. There are also parts of this content that is more likely to "land" when heard more than once. There really is no substitute for the in-person interaction that happens during this course. If I determine that attendance has dipped below an acceptable level, I will stop posting the recordings and distribute them to students by request only. If everyone uses these recordings responsibly (as I expect will be the case), then this common good will benefit the class greatly.

Textbooks: There will be two main text resources for this course:

- Huntington-Klein, N. (2021). The Effect: An Introduction to Research Design and Causality - the book is available from the author for free here
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning: With Applications in Python. - ISL is available from the authors for free here

Over the course of the semester, we will also be using a number of other resources that will posted to the course's Canvas site. **You do not need to purchase any books for this class!**

Readings for each topic will be posted on the course Canvas page. Many of the materials covered in the lectures are also covered in these books. Students are not expected to have done the reading before each lecture. However, some students may find the lectures more beneficial if the reading is completed prior to the lecture time. After the lectures, I highly recommend reading the posted chapters to fill in any gaps. When possible, I'll try to post the corresponding sections from each book.

Beyond these texts, there are numerous other references on the topics covered in this course that are freely available online. In many cases, these sites/books do a better job of explaining the material than the chosen textbooks. I highly recommend seeking out these materials to better your understanding of the topics in this course. If you find anything particularly good, please send them my way so that I can post them for all the students in the course!

Prerequisites: This class is designed to be taken simultaneously with a class on applied statistical methods (QTM 520), a class on statistical programming (QTM 530), and a class on communication (QTM 540). This course assumes basic knowledge of algebra and calculus. There is also an assumption that students have, at least, minimal exposure to probability and statistical inference (think conditional probability, expectations, etc.) and applied statistical methods. Students should also have some exposure to programming in R, Python, or another numerical computing language. However, the level of exposure to programming assumed each week will align with the topics covered in QTM 530 and the level of knowledge about applied statistical methods will align with the topics covered in QTM 520.

Course Structure: This course will be made up of lectures and discussions, problem sets, two short projects, and a final applied exercise.

1. **Synchronous Lectures and Participation** - Over the semester, there will be 28 in-person 75 minute lectures. Students are expected to attend these lectures. Periodically, there will be posted case studies that students are expected to read prior to the lecture session. Your participation in these discussions is also expected. Overall, attendance and participation in this class **will count for 10% of your final grade**

in this class. With regular attendance and an effort to participate in the lecture, I see no reason that a student will not receive full marks for this part of the final grade.

2. **Problem Sets** - There will be 4-5 problem sets **that will account for 40% of your final grade**. Problem sets will contain 1-3 exercises that are related to the course contents. Problem sets may contain derivations and proofs, coding exercises, data analysis, or writing exercises.

   Markdown or a similar typesetting method is preferred for problem set solutions submissions. In R, RMarkdown is the best option. A short introduction to RMarkdown can be found here. For Python, assignments can be typeset in a .ipynb file using Jupyter Notebooks. However, students may type up problem set answers in any form that can publish a .pdf.

   Each student must submit their own solutions document. However, **collaboration among students is strongly encouraged.** That said, please don't just copy solutions. If I find that there are identical solutions to any or all of the problems that are turned in, I will treat the submission as fraudulent and each student will receive a 0 for the assignment. In short, **don't cheat** - you're just cheating yourself at this point in your academic career.

3. **Short Projects** - Over the course of the semester, there will be two short projects that **will account for 30% (15% each) of your final grade**. The first project will relate to extensions to the standard randomized experiment with binary treatment assignment and the second project will relate to an extension to basic regression methods (linear and logistic) for building predictive models.

   The design logic for answering different types of questions with data we will cover in this class extend to many different methods. While we will cover some of the main applied methods that correspond to these logical tools, there are **many** other methods that use the discussed base logic and improve our ability to truthfully answer questions or extend it to answer slightly different types of questions. These short projects will see students apply the knowledge gained from this class (and the corresponding corequisite classes) to create materials that explain one of these approaches to a less statistically-inclined audience.

   Each short project will require two deliverables: a short blog-style document (an "explainer") that explains a method to the appropriate audience and a presentation to the class on that method. The blog-style document should explain the method to an audience that has a little exposure to the base method and present the extension in a way that someone with little knowledge could understand why it improves one's ability to answer specific questions. At a minimum, this should include:

   - A short explanation of the base method and what kind of questions it can answer
   - The problem with the base method that needs to be addressed
   - The new method and its link to the base method
   - A discussion of how the new method addresses the problem and how it allows us to answer new questions or provide a better answer to the old question
   - Trade-offs in using the new method instead of the old method (pros and cons)
   - (If appropriate) An example implementation of the new method and a visual comparison of the base and new methods

   Examples of online documents that do this well will be posted to the class Canvas site. A successful explainer will be one that could be posted online (as a part of a job portfolio, for example).

   The explainer document should be succinct without sacrificing important information, make good usage of visualizations, and be written in a professional manner. The document can be written up in any way

that allows you to do this (a well-formatted Word document, at a minimum). However, I highly recommend using this as an opportunity to learn how to create web-ready documents with Markdown. A guide for doing this with Markdown or Quarto will be posted to the course Canvas site.

Along with the explainer document, each student should prepare a **10 minute presentation** that distills the information from the explainer document. This presentation will be made to the class on the date specified in the syllabus. Your presentation should include slides which will be turned in to the course Canvas site after the in-class presentation.

The presentation should effectively explain the new method and link it to the concepts covered in class - explain how the method improves our ability to provide data-driven answers to certain types of questions. This exercise should see you take your succinct explainer document and further distill it to only the bare essentials. The goal of this presentation to is assist you in learning to explain the logic of different methods to different audiences (think a boss with an MBA who has a little exposure to randomized experiments).

Your grade for each of these assignments will be a function of the quality of the explainer document and your presentation. A more specific rubric will be posted when it is time to start working on these projects.

Example methods for experiments:

- Conjoint experiments
- Multi-armed bandit designs
- Stratified designs
- Early-stopping rules

Example methods for prediction:

- Splines for continuous outcomes
- Regularized regression for subset selection
- Random Forests
- Support Vector Machines for Classification

4. **Final Applied Exercise** - Immediately following our final class meeting on December 10th, I will post a final applied design exercise. Your solutions will be due on December 17th at 5PM EST. Your solutions for this final exercise **will account for 20% of your final grade**.

   The final exercise will attempt to replicate a realistic workflow that a data scientist might experience on the job. Much like the problem sets, there will be a series of scenarios and you will be asked to address a number of questions given some contextual information and a data set about best approaches, pros and cons of different choices, and interpretation of outcomes.

   The goal of this assignment is to tie everything from the class together into a realistic scenario which you might encounter out in the real world. Largely, you should view this assignment as a longer problem set that serves as a capstone to the first semester of the program.

Late Assignments: Any assignment turned in after the due date will receive a 5% per-day penalty. Late days are rounded up to the nearest day (an assignment due at 11:59 PM turned in at 12:01 AM on the next day counts as one late day). Saturdays and Sundays count as a single day - if an assignment is due on Friday and you turn it on the next Monday, you'll receive 10% subtracted from your final score.

I am amenable to extensions in certain scenarios. Just reach out to me to let me know in advance of the due date - the likelihood that I will grant an extension is inversely proportional to the time left before the assignment is due!

Final Grades: Final grades will be determined using the above weightings and the following (estimated) grade ranges:

- **A**: 93% – 100%
- **A-**: 90% – 92%
- **B+**: 87% – 89%
- **B**: 83% – 86%
- **B-**: 80% – 82%
- **C+ and below**: < 80%

As assignment scores are computed and I get more of a feel for the average grades on each assignment, the grading rubric may be updated. However, the grade scale will only be loosened from this initial rubric (e.g. if your final average is 93%, then you will get an A for the course regardless of rubric changes). Final grade percentages will be rounded to the nearest integer. There is no curve for final course grades - if everyone in the class scores high enough to get an A, then everyone will get an A. Hence, there is no "competition" for grades. **The assignments are not weighted on Canvas! You can always compute your final average by applying the appropriate weighting to your scores on your own.**

Grade Appeals: If you believe that your grade on any assignment is incorrect or unfair, you should submit your concerns, in writing, to me. The written appeal should fully summarize what you believe the problems are and why. Unless the appeal regards a simple addition error, please wait 48 hours before submitting a written appeal.

Accessibility and Accommodations: As the instructor of this course, I endeavor to provide an inclusive learning environment. I want every student to succeed. The Department of Accessibility Services (DAS) works with students who have disabilities to provide reasonable accommodations. It is your responsibility to request accommodations. In order to receive consideration for reasonable accommodations, you must register with the DAS at http://accessibility.emory.edu/students/. Accommodations cannot be retroactively applied so you need to contact DAS as early as possible and contact me as early as possible in the semester to discuss the plan for implementation of your accommodations. For additional information about accessibility and accommodations, please contact the Department of Accessibility Services at (404) 727-9877 or accessibility@emory.edu.

Tentative Course Schedule (Work in Progress):

Week 1 (Aug 28)

u **Topics**: Introductory Comments; Class Overview; Theory and Data Science

Week 2 (Sept 3 and 5)

u **Topics**: Asking Questions and Getting Answers; Bayes' Theorem and Evidence; Describing single features

Week 3 (Sept 10 and 12)

u **Topics**: Relationships between variables; What is your variation?; Descriptive Approaches; Answering "Why?", "How?", and "What if it changed?" questions u

Assignments: Problem Set 1 Posted

Week 4 (Sept 17 and 19)

u **Topics**: A counterfactual definition of cause and effect; Counterfactuals and the all-else-equal condition

Week 5 (Sept 24 and 26)

u **Topics**: Bias and confounding; What is the estimand?; Estimators of average treatment effects; Designbased methods u **Assignments**: Problem Set 2 Posted

Week 6 (Oct 1 and 3)

u **Topics**: Randomization; A/B Experiments; Inference for Randomized Experiments

Week 7 (Oct 8 and 10)

u **Topics**: P-values; False positives; Power; Internal Validity of Experiments u

Assignments: Problem Set 3 Posted u **Assignments**: Randomized

Experiments Explainer Assignment Posted

Week 8 (Oct 17)

u **Topics**: External Validity of Experiments and Ethics

Week 9 (Oct 22 and 24)

u **Topics**: Student Presentations; Causation vs. Predictive Models; Statistical Learning Theory u

Assignments: Student Presentations on October 22nd, Explainer and Slides Due on October 22nd

Week 10 (Oct 29 and 31)

u **Topics**: What can we do when we don't care about coefficients?; Supervised learning; In-sample vs. Out of Sample fit; Expected Prediction Error

Week 11 (Nov 5 and 7)

u **Topics**: Learning functions for continuous outcomes; Mean squared error and a neat decomposition; The bias-variance tradeoff; Closed form solutions via Optimism u **Assignments**: Problem Set 4 Posted

Week 12 (Nov 12 and 14)

u **Topics**: Validation set approaches and Cross Validation; Controlling overfitting via regularization

Week 13 (Nov 19 and 21)

u **Topics**: Intro to Classification; Bias/Variance Tradeoff?; Deep Learning as an extension to regression u **Assignments**: Problem Set 5 Posted u **Assignments**: Predictive Models Explainer Assignment Posted

Week 14 (Nov 26)

u **Topics**: Deep learning for images and text

Week 15 (Dec 3 and 5)

u **Topics**: Generative machine learning - basics of VAEs, GANs, and autoregressive models

Week 16 (Dec 10)

u **Topics**: Student presentations and Wrap-Up u **Assignments**: Student Presentations on December 10th, Explainer and Slides Due on December 10th u **Assignments**: Final Applied Exercise posted on December 10th. Due on December 17th at 5:00 PM.