Nicholas Eubank
Assistant Research Professor
`https://duke.zoom.us/my/nickeubank`
Office Hours: Monday 12:45-1:45

IDS 701
Spring 2024
Tuesday / Thursday 1:45-3:00

Unifying Data Science: Using Data Science to Solve Problems

1 Course Description

All too often, students learn data science by taking a course on machine learning from a computer scientist, a course on statistical modeling from a statistician, and a course on causal inference from a social scientist. As a result, graduating students find themselves with a toolbox of techniques, but no clear idea of how to use them to solve problems.

The aim of this course is to overcome this fragmentation and to provide students with a unified approach to using data science to solve real world problems. To that end, we will introduce a question-first, backwards design framework for systematically designing a data science project. Through exercises, students will practice each step of this approach, from working with stakeholders to properly articulate the problem they are seeking to address, to picking a question (which, if answered, will help the stakeholder solve their problem), selecting the appropriate methodological approach to answering that question, and developing a concrete strategy for generating an answer.

Having established this framework for solving data science problems, the class will then pivot to providing an application-focused introduction to *causal inference*, the art and science of using statistical data to make causal statements about the world. Our approach will be rooted in the potential outcomes framework, and will cover a range of methods of statistical inference including randomized experiments, pre-post analysis, differences-in-differences, and instrumental variables. In addition, we will also discuss concepts like the distinction between internal and external validity, and the limitations of estimating Average Treatment Effects.

Finally, towards the end of the semester—once we have covered causal inference in this class and MIDS students have covered machine learning in detail in IDS 705—we will return to our more general investigation of how best to use data science to solve problems, now with a focus on when (supervised) machine learning approaches are appropriate and when causal approaches are preferable.

In addition to completing a number of exercises related to project design, over the semester students will conduct a complete data science project themselves. Data science is a fundamentally applied field, and there is no substitute for learning to put these project design principles into action through practice. These projects will be developed incrementally over the course of the semester with instructor guidance. By the end of the semester, students will have picked a topic area, developed a (tractable) question, decided what an answer to that question would actually look like, developed a work plan for generating that answer, and executed and presented their project, and then iterated the project based on feedback from their initial presentation. For MIDS students, this will serve as a "capstone-project with training wheels" to prepare students for their second-year Capstone projects with external partners. And this project should provide all students with a portfolio piece they can present to potential future employers.

Throughout the course, we will also be consistently returning to a few themes, chief among them the importance of developing a skeptical mindset. This is a core data science skill, but one that students do not always have the opportunity to practice. In this course, we will discuss *and practice* approaching our data, our code, our statistical models, our problem statements, and the work of others from a constructive but skeptical perspective.

1.1 Pre-Requisites for Non-MIDS Students

This course is primarily designed for students in the Duke Masters in Interdisciplinary Data Science (MIDS) program, but students from other programs are more than welcome if they have the appropriate pre-requisite training. Data Science is a fundamentally interdisciplinary field, so the more perspectives we have represented in the classroom the better!

This course will assume that enrolled students have a good grasp of inferential statistics and statistical modelling (e.g. a course in linear models), though no prior experience with causal inference is expected. In addition, MIDS students will be taking a concurrent course in applied machine learning, and so incoming students will also be expected to have some basic experience with machine learning, or be concurrently enrolled in an applied machine learning course.

This course will also assume students are comfortable manipulating real-world data in either Python or R. The substantive content of this course is language-independent, but because students will be required to work on their projects in teams, comfort with one of these two languages will be required to facilitate collaboration (MIDS students are, generally, "bilingual" in R and Python). Where code examples are provided in class, they will use Python (‘pandas‘), but both the instructor and our TAs are also capable of providing support in R.

Finally, students will also be expected to be comfortable collaborating using git and github. If you meet the other requirements for this course but are not familiar with git and github, this is a skill you should be able to pickup on your own in advance of the course without too much difficulty. You can read more about git and github here. The Duke Center for Data and Visualization Science also hosts git and github workshops for Duke students.

2 Assignments & Grading

2.1 Participation (20% of Grade)

A major component of good participation is good *preparation*. Because we will often use class time for exercises, it is absolutely critical that students do their assigned readings before *every* class. Students who do not work through the instructional materials they have been assigned before class will not only get very little out of in-class exercises designed to reinforce the assigned materials, but they will also undermine the learning of the students they are asked to work with. With that in mind, students who do not complete their assigned readings before every class should be expected to see this reflected in their participation grades.

Cold calling: In the interest of creating an interactive learning experience, I will often "cold call" students with questions about the material we are discussing. To be clear my goal with cold calling is not to "catch" students who haven't done the reading, but rather to ensure that everyone is getting an opportunity to participate in the discussion. However, students who regularly demonstrate *unfamiliarity with readings* can expect to receive lower participation scores (not having the right answer will not get you a low score, to be clear! The material in this course is difficult, so I don't always expect everyone to have the right answers on the tip of their tongue, but it's pretty easy for an instructor to recognize the difference between somebody who is wrestling with the material and a student who just hasn't done the reading).

Partner Surveys: When working on exercises in teams, students will be asked to provide feedback on some basic questions about collaboration with their partners. These responses will contribute to student participation scores, and will also be returned to students in a pseudonymous manner — from time to time, responses will be collated, names will be removed, and the ratings will be returned to students so that students can learn how others view their collaboration skills. Because ratings and comments will be returned in batches of 3 or 4 responses and without names, the authors of each response won't be obviously identifiable, but recipients may be able to reason about authors, so one should always be thoughtful in the feedback one provides.

Participation will be graded as follows:

A range. You are fully *and consistently* engaged in class discussions and exercises. You both listen and contribute actively. You are well-prepared for class. Having done more than merely read the material, you have spent time thinking *carefully and deeply* about the material's relationship to other materials and ideas presented in previous classes. You are not only able to answer questions about the material, but also come to class with thoughtful questions. When working in teams, you work *with* your partner. If your partner is struggling with an exercise, you help them understand the material rather than just completing the material on your own. If you are struggling with material, you ask for help (both from the instructor — in class and in office hours — and your fellow students) and do not simply lean on your partner to complete the exercise.

B range. You are engaged in class discussions and exercises. You listen and contribute regularly. You come well-prepared to class having read the material and your contributions show your familiarity, but your level of engagement lacks the depth accumulated through extra time spent thinking about the material. When working in teams, you work *with* your partner when they have a similar level of understanding, but do not always invest in helping a struggling partner to understand the material. You often ask for help when you are struggling, but other times you let your partner just complete the exercise, and don't attend office hours regularly when struggling.

C range. You have met the minimum requirements of participation. You are usually, but not always prepared. You participate sometimes, but not regularly. The comments that you offer show a basic familiarity with the materials but do not help to build a coherent or productive discussion. When working in teams, you only sometimes work *with* your partner. When your partner is struggling, you often just do the exercise yourself. If you are struggling, you often do not ask for help or attend office hours and allow your partner to take over the exercise.

D range. You have not met the minimum requirements of participation. You are unprepared for class. You have not read the material with sufficient engagement to know even the most basic elements. When working in teams, you do not attempt to work *with* your partner. When your partner is struggling, you just do the exercise yourself. If you are struggling, you do not ask for help and allow your partner to take over the exercise.

As should be clear from this rubric, above all it is important to emphasize that participation is evaluated on the basis of *quality* and *consistently*, *not* quantity. Moreover, when completing in-class exercises, good participation is not about finishing first or without ever asking for help; good participation in in-class exercises is about helping your partner understand the material, and asking for help when you need it.

2.2 Causal Inference Mid-Term (20% of Grade)

After the causal inference portion of our course we will have a mid-term exam.

2.3 Exercises (20% of Grade)

Over the course of the semester, students will be asked to complete a number of small exercise assignments as homework. These exercises will, in total, be worth 20% of student grades.

2.4 Reading Reflections (20% of Grade)

Both to provide the instructor and teaching assistants with information about what topics students have found difficult, and also to ensure that students are doing the required readings (a necessity for a flipped classroom designed to be effective), students will be required to submit answers to a set of prompts about the required readings by 9am on the morning of each class.

2.5 Team Data Science Project (20% of Grade)

Over the course of the semester, you and your team will develop a full data science project—from conception to execution and presentation. Your scores on the various components of this project—including graded drafts, intermediate work, teamwork, and project management skills—will jointly constitute 20% of your overall grade.

2.6 Late Assignments, Make Up Exams and Extra Credit

Late Assignment

All late assignments will be penalized 10% per day the assignment is late, up to a maximum penalty of 50%.

The final deadline for accepting assignments that are more than one week late is at the discretion of the instructor and may vary by assignment.

Exceptions to these late penalties may be made for students dealing with exceptional circumstances (illness for themselves or family, etc.) — if you are dealing with a difficult situation, please feel free to contact me to discuss your situation.

Dropping Lowest Scores

To accommodate the fact that life happens, at the end of the semester, I will drop each student's lowest Reading Reflection *and* lowest Exercise **that has been completed.** Essentially, this is a free pass for one exercise and one Reading Reflection you totally whiff or submit very late. But it is *not* a free pass to *skip* an Exercise or Reading Reflection — uncompleted Exercises or Reading Reflections are not eligible for being dropped.

3 Honor Policy

Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and nonacademic endeavors and to protect and promote a culture of integrity.

Remember the Duke Community Standard that you have agreed to abide by:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Cheating on exams or plagiarism on homework assignments, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved. Additionally, there may be penalties to your final class grade along with being reported to the MIDS program directors.

3.1 chatGPT & Reading Reflections

Text generated by chatGPT should *never* be submitted as part of your Reading Reflections. I'm not an anti-chatGPT absolutist, and I welcome its use to aid your programming. However, it is very easy for chatGPT to provide answers to reading reflection questions in a way that short-circuits the learning process. Research is very clear that the process of summarizing an idea and putting it in your own words is a critical part of actually learning and understanding an idea. As such, any assistance in that process — *at least in this educational context* — is likely to be detrimental to your learning.

This prohibition applies both to asking chatGPT for the answer to a Reading Reflection question and pasting the answers into Gradescope *and to writing a draft of your answer and then asking chatGPT to refine it.* That's because while the refinement of you're writing *may* not compromise the learning process, it does make it impossible for me to use software for detecting Large Language Model (LLM) software reliably to identify cases where the use of chatGPT *is* clearly undermining learning.

Students whose Reading Responses are flagged as likely being the output of an LLM will receive one warning, after which further high scores that cannot be readily explained may be treated as honor code violations.

4 Disability Policy

In an effort to prevent students with disabilities from having to explain and justify their condition separately to each of their various instructors, Duke has centralized disability management in the Student Disabilities Access Office. If you think there is a possibility you may need an accommodation during this course, please reach out to their office as soon as possible (processing can take a little time).

Medical information shared with the SDAO is strictly confidential, and if SDAO determines an accommodation is appropriate, faculty members will simply be informed of the accommodation they are required to provide, not the underlying medical reason for the accommodation.

If you have any problems with SDAO, please let me know as soon as possible.

5 Final

While this course does not have a final exam, we may use our "final" time slot for group presentations, so please keep it open.

6 Mental Health and Wellness

Mental health and wellness are of primary importance at Duke, and the university offers resources to support students in managing daily stress and self-care. Duke offers several resources for students to seek assistance on coursework and to nurture daily habits that support overall well-being, some of which are listed below:

- The Academic Resource Center: (919) 684-5917, the ARC@duke.edu, or arc.duke.edu
- DuWell: (919) 681-8421, provides Moments of Mindfulness (stress management and resilience building) and Koru (meditation) programming to assist students in developing a daily emotional well-being practice. To see schedules for programs please see https://studentaffairs.duke.edu/duwell. All are welcome and no experience is necessary. duwell@studentaffairs.duke.edu, or https://studentaffairs.duke.edu/duwell

If your mental health concerns and/or stressful events negatively affect your daily emotional state, academic performance, or ability to participate in your daily activities, many resources are available to help you through difficult times. Duke encourages all students to access these resources.

- **DukeReach**.Provides comprehensive outreach services to identify and support students in managing all aspects of well-being. If you have concerns about a student's behavior or health visit the website for resources and assistance.http://studentaffairs.duke.edu/dukereach
- **Counseling and Psychological Services (CAPS)**. CAPS services include individual, group, and couples counseling services, health coaching, psychiatric services, and work-

shops and discussions.CAPS also provides referrals to off-campus resources for specialized care.(919) 660-1000. https://studentaffairs.duke.edu/caps

- **Blue Devils Care**. A convenient, confidential, and free way for Duke students to receive 24/7 mental health support through TalkNow and scheduled counseling. bluedevilscare.duke.edu
- **Two-Click Support**. Duke Student Government and DukeReach partnership that connects students to help in just two clicks. https://bit.ly/TwoClickSupport

7 Student Signature

I have read and understand this syllabus.

Name:

Signature: