Nicholas Eubank
Assistant Research Professor
Office Hours: Thrs 9:30-10:30, Gross 231 &
https://duke.zoom.us/my/nickeubank

IDS 701
Spring 2025
Tuesday / Thursday 1:25-2:40
Gross Hall 330

Solving Problems with Data: A Unified Approach to Project Design

1 Course Description

Few fields have shown as much promise to address the world's problems as data science. At the same time, however, recent years have also made clear that today's global challenges will not be met by simply "throwing data science at the problem" and hoping things will work out. Even in business, where many assume that Artificial Intelligence is a sure ticket to profits, a major recent study found only $> 11\%$ of businesses that had piloted or employed Artificial Intelligence had reaped a sizeable return on their AI investments.

How, then, should a burgeoning data scientist approach this field full of such promise but also so many pitfalls? And why have so many data science endeavors failed to deliver on their promise?

The answer lies — at least in significant part — in a failure to provide students with a systematic approach to bringing the techniques learned in statistical modeling and machine learning courses to bear on real-world problems. Data science curricula usually begin with coding, statistics, and model evaluation techniques. All too often, however, that's where they stop. But while the hardest part of data science *classes* is often fitting a model well or getting a good AUC score, the hardest part of being an effective *professional* data scientist is ensuring that the models being fit and the results being interpreted actually solve the problem that motivated you (or your stakeholder) in the first place.

This class is designed to fill this gap. Through exercises, case studies, and projects, students will develop a *systematic* understanding of how to approach and manage data science projects from conception through delivery and adoption. It will provide a unified perspective on how the perspectives and tools learned in other courses complement one another, and *when* different approaches to data science are most appropriate.

In addition, this course will also provide an in-depth introduction into *causal inference* — the practice of answering causal questions. Given the interests of MIDS students, this introduction will focus heavily on experiments and A/B testing, but will also cover the use of observational data (data that did not come from an experiment that employed random assignment to treatment) for answering causal questions.

1.1 How to Succeed in this Class

In Duke course reviews, students are asked, "What would you like to say about this course to a student who is considering taking this course in the future?"

By far the most consistent thing past students say they would like to tell a student considering taking it in the future is to **do the readings and take them seriously**.

There is a tendency among data science students — especially those from a STEM background — to assume that readings that don't have a lot of math aren't "serious," and consequently don't require substantial focus. That's a mistake. This course is about the critical reasoning required to make the leap from the relatively clean math of statistics and machine learning to the messiness of real world problems. To help you learn how to do so, the readings are full of examples, ways to think about problems, and problem-solving frameworks to help you cross that wobbly bridge from the clean world of problem sets to the real, under- or mis-defined problems you will face when you enter the work force. But with this type of material, what you get out of it depends on what you put into it, and unlike with a theorem — which you either follow or you don't — thinking critically happens on many levels. So while it's easy to skim a reading and — because you weren't confused by any greek notation — assume you internalized it, succeeding in this class requires actively wrestling with the material, not just letting your eyes glide over it.

In other words, **take the readings for this course just as seriously as the exercises.** There is as much learning to be done by thinking deeply about the readings as there is to be gained from doing the exercises, a fact that is also reflected in how the course is graded (individual or two-person exercises make up 20

1.2 Big Ideas

This course is organized around three big ideas:

1. **Data science is about solving problems**. All too often, data scientists get lost in the technical details of models and lose sight of the bigger picture. Data science is not about maximizing accuracy or AUC scores — it's about using data and quantitative methods to solve problems, and at the end of the day the only "metric" that matters is whether your work has solved the problem you set out to address.
2. **Data scientists solve problems by answering questions, and the question you are asking determines what tool is appropriate.** At their core, all data science tools are tools for answering questions, whether you realize it or not. Learning to recognize how data scientists use questions to solve problems — and exactly what questions are being answered by the tools you use every day — is key to navigating the ambiguity of real-world problem solving.
3. **Reasoning rigorously about uncertainty and errors is what differentiates good data scientists from great data scientists.** Data science isn't just about minimizing classification errors and uncertainty — it's also about deciding how unavoidable errors should be distributed, and how to weigh the risks and trade-offs inherent in probabilistic decision-making rigorously and in a manner that takes into account the problem you are trying to solve.

1.3 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 modeling (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, 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 Python will be required to facilitate collaboration (MIDS students are, generally, "bilingual" in R and Python, but have a strong preference for Python, and it's hard to write problem sets to accommodate multiple languages).

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. 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).

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 Mid-Term (20% of Grade)

After completion of the portion of the class focused on experiments, there will be an in-class, closed-book midterm. The aim of this midterm is not to force memorization, or to create an

artificial means of grading students. This class is fundamentally about critical thinking, and to be useful it is not enough that students know the material when they have their notes in front of them; to be impactful, students must understand the concepts covered well enough to recognize their relevance in new contexts, and that requires a level of familiarity that requires a closed book 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. **Note:** because of the way students get autograder feedback before final submission and because the lowest exercise grade gets dropped, grades on exercises tend to converge towards 100%. In light of that, I will generally adjust exercise scores down 5 percentage points when doing grade calculations so that getting all available points on an exercise (reported as 100% on gradescope) maps to a 95% solid A (but not a 100%).

2.4 Reading Reflections & Quizzes (20% of Grade)

In previous iterations of this class, I had students write reading reflections prior to each class. With the rise of chatGPT, however, I have found that the share of reading reflections actually being completed by humans has fallen off dramatically, which is problematic for two reasons. The first is that the point of reading reflections is to encourage, well... *reflection* on the material you are reading (see discussion of active learning above). But if prompts are just being fed into chatGPT, then this is not taking place.

In light of that fact, I am still working through how best to ensure students give the pre-class readings the attention they deserve. At the moment, my plan is primarily to give regular (closed-book) quizzes at the start of class on topics from the readings.

But aren't closed book quizzes "inauthentic" assessments? I don't think so. Just because chatGPT can answer these prompts when these prompts are provided doesn't mean that there's no need to understand the material when you don't have access to chatGPT. To be successful professionally, you need to understand the concepts and ideas you will encounter in this class well enough to recognize their relevance *in other contexts.* Yes, chatGPT will always be available to provide you the definition of "SUTVA," but if you don't understand the idea at an intuitive level, then when you're deploying an A/B test in your job it won't occur to you that it's a problem that your platform allows users in different treatment arms to interact with one another. The rise of LLMs means that there will be lots of things data scientists no longer need to learn to the same level of detail that they needed to learn previously, but those are not the types of things on which you will be quizzed.

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.

For quizzes:

- students who are not present in class will not be allowed to make up quizzes
- except if the student has reached out to me **before** class to alert me to their absence and the reason for it. If the reason is it deemed to be grounds for an excused absence, the student will be able to take the quiz at a different time.
- students who are not physically in class are also **NOT** allowed to take quizzes remotely without express permission. If you are found to have taken a quiz when not in class, not only will you receive a zero on that quiz, but **it may be considered an honor code violation**. Note that because classes are recorded (for students with excused absences) and gradescope tracks IP addresses, this is not a hard thing to figure out.

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 (or quiz) *and* lowest Exercise **that has been completed (see notes about quizzes below).** Essentially, this is a free pass for one exercise and one Reading Reflection/quiz you totally whiff or submit very late. But it is *not* a free pass to *skip* an Exercise or Reading Reflection/quiz — uncompleted Exercises or Reading Reflections are not eligible for being dropped.

3 Laptop Policy

The causal evidence from the teaching and learning field clearly shows that learning outcomes are worse when students have laptops in the classroom. This appears to be due in part to computers allowing students to become easily distracted (e.g. Facebook), and also in part because significant information synthesis occurs when we take notes by hand because we are not fast enough writers to transcribe everything being said. Most typists, on the other hand, can conceivably transcribe word for word for what's being said, but this requires little to no mental processing of the information.[1]

Moreover, research also makes clear that the presence of laptops in the classroom undermines

[1]For excellent work on this topic as well as a nice summary of existing literature on the impact of laptops on learning outcomes, see Mueller and Oppenheimer (2014, *The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking*). Fried (2007, *In-class laptop use and its effects on student learning*) and Ravizza et. al. (2017, *Logged In and Zoned Out*) are also worth reading.

learning not only for the person with the laptop, but also for other students in the room (presumably it is distracting to have the computer next to you jumping back and forth from Instagram to TikTok).

For these reasons, during portions of the class when I am speaking and when you are not actively engaged in programming exercises, I do not allow laptops out in class. (When I use slides, those will be provided online so you can refer back to them later if you would like).

However, please **do bring your laptops every day**, as we will use them for in-class quizzes, and some in-class exercises.

Note: If you would like an exception to this rule for medical reasons or due to learning differences, please speak to me and/or follow the directions in the Disability Policy section below, and we will be sure to find an arrangement that works for you.

4 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.

5 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.

6 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.

7 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 workshops 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

8 Student Signature

I have read and understand this syllabus. No, seriously: I read the laptop policy, the quiz policy, the attendance policy, the late submission policy, all of it.

Name:

Signature: