Nicholas Eubank
Assistant Research Professor
Office Hours: TBD
Gross Hall, Room TBD

IDS 690-04
Spring 2019
Monday / Wednesday 10:05-11:20
Gross Hall, Room 230

Unifying Data Science

1 Course Description

The aim of the course is to two-fold. First, it aims to provide students with a conceptual framework for understanding the relationship between the many tools that are currently taught under the "data science" umbrella. This course takes the view that data science is fundamentally about answering questions with data, and so is organized around helping students identify different classes of questions (descriptive, causal, and predictive). Over the course of the semester, we will explore each of these types of questions in turn, learning which tools are appropriate for each, and what what pitfalls are common to efforts to answer each type of question.

Second, it aims to provide students with experience both developing and actually answering real questions. Data science is a fundamentally applied field, and so while it is important to have the conceptual framework described above to aid your work, there is no substitute for learning to put these principles into action through practice.

To achieve this second learning goal, over the course of the semester students will develop their own data science projects in small teams. 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.

As this course is primarily designed for students in the MIDS program, it will assume familiarity with statistical modeling (basic statistics, linear regression, logistitic regression, model selection) and the basics of both supervised and unsupervised machine learning. The goal of this course will not be to teach these topics, but rather to help *contextualize* them.

Of the three types of questions we will cover, methods for answering causal questions will receive the greatest attention. This course assumes no familiarity with causal inference, and will cover everything from the basic problem of causal inference to experiments, and to the range of tools available for making causal inferences from observational data.

1.1 Pre-Requisites

This course will assume that enrolled students have a good grasp of inferential statistics, statistical modeling, and have 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 TA 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 Types of Questions: The Big Ideas

The instructional material for this course will be organized around a three-fold taxonomy of questions one may seek to answer as a data scientist: Descriptive Questions, Causal Questions, and Predictive Questions.

2.1 Descriptive Questions

Descriptive questions are often the least respected in the data science realm, but in my view good descriptive analyses are both one of the hardest things to do well, and also are often the most important to generating new knowledge.

In this course, we will discuss a range of different methods for descriptive analysis, ranging from summary statistics (means, medians, standard deviations), to data visualization, and to unsupervised machine learning algorithms (such as tools clustering and dimensionality reduction).

As we explore these tools, we will continually come back to the fundamental problem of descriptive analysis: descriptive analysis is about summarizing data, but the process of summarization requires discarding information, and it is *always* up to the data scientist to determine what information can be discarded as extraneous, and what data cannot. Descriptive analysis tools will always provide "an answer," but it is up to the data scientist to know if that answer is a faithful representation of the structure of the data.

2.2 Causal Questions

[Note that this section is stolen shamelessly from the syllabus of Adriane Fresh, whose research design course provides the basis of the organization of this portion of the class.]

A few big ideas and enduring understandings will facilitate our exploration of causal inference in this course. The first big idea is that *causality is fundamentally unknowable, and we must therefore* approximate *an unobserved and unobserv*able *outcome* in order to draw causal inferences. This is sometimes referred to as the fundamental problem of causal inference — the idea that an understanding of cause and effect depends on our ability to understand outcomes that never happened. In this way, I (personally) think the study and application of causality is about as close as science gets to magical thinking, which, when you think about it deeply, is pretty

cool. Thus, our work as empirical social scientists is in how we *approximate* the unobserved counterfactual outcome, with the understanding that it will *always be* an approximation.

The second big idea is that *causal inference is a* logical *process of making comparisons, not a quantitative nor a qualitative process.* Regardless of which type of evidence you are utilizing, causality depends on making thoughtful comparisons that allow you the best opportunity to approximate an unobservable occurrence. In understanding what this means, we will hopefully dispel (or ward off) notions that one type of research is inherently better than another, and see that the choice of evidentiary approach, and research design more broadly, are question-driven rather than ideological.

Finally, the third big idea is that *causal inference always requires assumptions, and evaluating those assumptions requires judgement.* The extent to which we think the necessary assumptions are *true* will dictate how confident we are in a given causal inference. But we will always need assumptions, because our approaches are only approximating the counterfactual. Yet evaluating those assumptions is an inherently subjective process — there is no test, no summary statistic, no one number or set of numbers that will tell you exactly whether an assumption is true or not.[1] And although some of the assumptions we'll study have statistical analogues, we'll the learn the importance of recognizing (in the vein of the big idea above) that assumptions for causal inference are logical assumptions, not necessarily statistical assumptions.

2.3 Predictive Questions

Making predictions is perhaps the hottest corner of data science today. Supervised machine learning – in which one feeds an algorithm examples of the predictive behavior one wishes the algorithm to emulate, then points the algorithm at new sources of data and asks it to make novel predictions – is viewed by some as synonymous with "data science."

(Note that throughout this course, I will use the term "predictive" not to refer specifically to trying to make a guess about what will happen *in the future*, but rather as a general term for the behavior that we ask a supervised machine learning model to exercise. That is because on some level, what a machine learning model is trying to do is predict *what the agent that generated its training data (usually a person) would do if the agent were do it itself.* For example, if you wanted a machine learning algorithm to identify cats in pictures, you would likely start by training the algorithm by showing it lots of pictures that have already been labeled as containing cats or not by humans. The algorithm would then do its best to build a model that accurately predicts, for a new set of pictures that haven't been labeled by humans, what label a human would apply if humans were to also label the new pictures.)

As we will discuss in this portion of the class, however, the scope for supervised machine learning is often much more narrow than is generally assumed, and *mis-application* of machine learning can have disastrous (and often extremely discriminatory) results.

[1] As an aside, this is the same as statistical inference in the frequentist tradition of statistics. The choice, for instance, of which *p-value* one is willing to accept in order to believe that an estimate of a parameter is different from some null hypothesis value is a personal, subjective choice. It is a choice that balances the probability of two types of opposing errors — Type I and Type II. While there may be *norms* that see groups converge on being willing to accept a particular value — say, 5% — there is no objective truth that makes that choice of a value "better" than another.

With that in mind, we will split our discussion of predictive questions into two halves: predictive questions in stable contexts and predictive questions in *unstable* contexts.

Stable contexts are situations where we plan to make predictions in situations where the behavior observed during training is nearly identical to the context in which we will apply our algorithm. For example, a stable context is one in which we might use machine learning to predict the likely future value of new customers at a big box store (like Target) on the basis of the behavior of current customers. In these contexts, supervised machine learning algorithms can be very helpful.

In unstable contexts, by contrast, supervised machine learning algorithms struggle, and better predictions may often come from more robust causal analyses. For example, we if wanted to plan a major change to US insurance subsidies, it is unlikely that a machine algorithm would be able to predict how Americans would respond to this kind of novel change.

Finally, we will also discuss what makes a context stable or unstable, which is not always obvious. One major problem with the use of machine learning algorithms, for example, is that they sometimes fail not because the context in which agents operate is unstable, but because the subjects of machine learning algorithms (i.e. people) may change their behavior once they are aware that they are interacting with algorithms (so called "adversarial users"), a phenomenon that comes up not only in information security, but also when algorithms are used to grade elementary student essays.

3 Developing and Answering Questions: The Big Ideas

Perhaps the hardest job of a data scientist is not answering a question, but coming up the with right question to ask in the first place. And so over the course of the semester, students will be repeatedly asked to practice generating tractable questions (and plans for answering those questions).

But wait, I thought the job of a data scientist was to make recommendations?

If you are working in the public policy or private sector realms, making a recommendation is often the last step of your data science project. However, I will posit that the key to making a good recommendation is to ask yourself:

What question, if answered, would make deciding what to do next easy?

Suppose, for example, you are hired by a company that gives you its customer data and says "we want you to use machine learning to help us target customers." That certainly doesn't sound like a question. But if you think about it for a little while, you come to realize that there is a question implicit in the task you've been given: *Which of our customers are most likely to respond positively to targeting?*

A core idea of this course is that converting your task into a question should be both the first step of your work, and perhaps the most important. The question you seek to answer is what motivates everything you do with your data. And if you haven't clearly defined the question you're seeking to answer, I can absolutely guarantee that when you turn to your data, you will find yourself unsure of what to look at, and you will not be able to use your time efficiently.

With that in mind, in this course we will focus on using *backwards design* to develop your data science projects, in which you:

- begin by identifying a topic area of interest (if you're a researcher) or a problem to be solved (if you're of a more applied mindset),
- decide what question, if answered, would either improve your understanding of the topic that motivates you (if you're a researcher) or allow you to solve the problem you identified (if you're in an applied setting),
- decide what an answer to your question would look like in concrete terms (what figure, regression, or machine learning output would constitute an "answer" to your question),
- figure out what data you need to generate that answer, and finally
- figure out where you will find the data you need and merge / manipulate it to allow you to generate your answer.

4 Class Organization

Data science is an applied discipline, and so this will be an intensely applied class with *lots* of hands-on exercises.

5 Laptop Policy

The causal evidence from the teaching and learning field clearly shows that we retain information better when we write with paper and a pen, rather than type our notes. This is primarily a function of the information synthesis that has to occur when we write 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.

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

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).

6 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.

7 Disability Statement

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 are 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.