Nicholas Eubank

Assistant Research Professor

Office Hours: TBD

Gross Hall, Room TBD

# Unifying Data Science

## 1 Course Description

The aim of the course is 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.

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.

In addition to learning to how the range of tools available to the modern data scientist relate, over the course of the semester students will have the opportunity to 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 finally executed and presented their project.

### 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 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).

**The Big Idea**

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

[Coming soon]

## 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 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."

**The Big Idea**

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.

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    Class Organization

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

# 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 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.