UNIVERSITY OF CALIFORNIA SAN DIEGO

An Examination of Master's Level Data Science Programs Across the United States

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Data Science

by

Sarah Albers Fester

Committee in charge:

Professor David Danks, Chair
Professor Justin Eldridge,
Professor Stuart Gieger

2024

The Thesis of Sarah Albers Fester is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To the Bachelor of Behavioral and Social Sciences at IE University, I have used **everything** you taught me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge David Danks. I am so grateful for all your support, you truly changed my experience here.

I would like to acknowledge Bibi and Hena, my incredibly last-minute coders. Hena you're fabulous, welcome to social sciences, and Bibi, if Gede doesn't marry you I will.

I would like to acknowledge Rafif, Neringa, and Jess from IE. You all showed me there was room for humanities and compassion in every field.

I would like to acknowledge my mom. She listened patiently to every stress phone call about this thesis and provided unconditional love and support.

ABSTRACT OF THE THESIS

An Examination of Master Level Data Science Programs Across the United States

by

Sarah Albers Fester

Master of Science in Data Science

University of California San Diego, 2024

Professor David Danks, Chair

Data science is an emerging discipline that has grown increasingly popular in the past decade. In response, numerous schools have developed and launched their own master's level data science programs. Through the development and launch of these programs, universities are selecting aspects of data science to structure their programs around, which then further develops data science as a discipline. Of the three aspects that compose data science, Theoretical Knowledge, Technical Execution, and Human Oriented Professional Skills (HOPS), universities can choose to value certain aspects over others through their program descriptions, program

prerequisites, and mandated courses. This work found that programs tend to gravitate towards favoring technical execution aspects in mandated courses, favoring theoretical knowledge in program prerequisites, while neglecting the aspect of HOPS in all areas of the program.

The consequences of this are threefold. First, when theoretical knowledge is used as a program prerequisite it can prevent students from non-STEM backgrounds from entering into data science, reducing the thought diversity in the field. Second, when HOPS are neglected, students cannot effectively access knowledge from other disciplines, closing off data science from new tools, methods, and problems. Third, when HOPS skills are neglected in favor of Technical Execution courses, students learn how to apply tools, but perhaps not judge the consequences or implications of their tool or method choices.

INTRODUCTION

Data science is an emerging discipline that is growing increasingly popular. This has caused the widespread roll out of numerous graduate level data science programs. However, the actual definition of what data science is can be difficult to capture. There are numerous opinions on what is and is not data science, what data science should and should not be, and who is and is not a data scientist. A way to attempt to define the discipline is to look to graduate level education programs that offer data science degrees.

Graduate level programs offer communal membership. They define levels of knowledge, skills, and norms that students need to know in order to become practicing members of the discipline. This work specifically examines master's of data science programs. The level of master was selected because there is a level of professionalization that occurs at the master level that is not present at the undergraduate level. In addition, master's programs have a time constraint that requires they offer what is most necessary in order to be a practitioner of the discipline. A master's program lasts typically one to two years. Curriculum designers must introduce the core requirements for discipline membership quickly as they have limited time with students, unlike an undergraduate or doctoral program where the time with students is longer. An undergraduate or doctoral program may introduce a wider variety of skills and practices that while helpful, are not strictly required in order to be a practicing data scientist, than a master's program can. For example, an undergraduate has numerous opportunities to take electives in other departments and hone domain knowledge in addition to general education requirements. A doctoral program is meant for deep advanced study in one area. A master's program may allow for a few electives, but generally, the courses are limited to the topics a practitioner of the field must know.

The aspects of data science that a program prioritizes are communicated through different ways. A program can communicate the aspects of data science it's choosing to prioritize through its program description, program prerequisites, and mandated courses.

The description of a program is the first interaction a potential student has with the program. A program description is what the program is telling prospective students it values. The program description may focus on the content covered in the program, the skills a student will gain in the program, and/or how employable the student will be after the program. While it is in the program's best interest to appear appealing to students, the program description can still reveal what a program says it values or perhaps even what the program believes it values.

The prerequisites of a program are the minimum requirements a prospective student needs in order to be considered for admission into the program. These requirements serve as gates that only certain students can pass through. Understanding these gates is important, because a discipline is not solely composed of knowledge, it's composed of people who transfer knowledge to each other. The people who are practitioners of a discipline are practitioners because they have gone through a process that deems them worthy of membership. As members of the community, they then can go on to roles where they are the ones that decide what to teach new membership hopefuls. If a select group of people are admitted into data science programs from one specific background, then they populate that field. They can set program prerequisites that beget students that are similar to themselves, and the process continues. This can lead to a lack of thought diversity in the field or isolate the discipline. Further, an examination of mandated course prerequisites is also beneficial in terms of identifying any hidden gates a student must pass through after they are admitted into the program.

The mandated courses of a program are extremely useful in understanding the aspects of data science a program values. Mandated courses demonstrate the aspects of data science that a program values because they are the courses that all students must pass through to earn their degree. The mandated courses convey the knowledge that has been deemed to be absolutely integral to becoming a practitioner of data science. In an ideal world, the aspects that a program says that they value are the aspects taught in the mandated courses. However, this may not be the case and drawing a comparison between what is taught and what is said to be taught is useful for determining how the discipline is talked about versus how the discipline is practiced.

Data science is a unique discipline in that it is young, has multidisciplinary origins, is rapidly developing, and has widespread impact. The practices and consequences of data science are incredibly far reaching because data science interacts with numerous other fields. In addition, as technology weaves its way into more and more aspects of daily life, the average person is impacted by the choices a data scientist makes. Understanding what a data scientist learns and what aspects of data science are valued by the programs that produce data scientists is incredibly important.

In order to understand how data science programs across the United States are shaping the discipline of data science, first a literature review was done to understand how the discipline has previously been defined and identify the key aspects of data science a program ought to teach. Then, using data from sixty-two master level, data science programs, program descriptions, mandated courses, program prerequisites, and mandated course prerequisites were analyzed to determine which aspects of data science they corresponded to. Program descriptions corresponded primarily to the Technical Execution aspect, however, did discuss HOPS. Program prerequisites corresponded primarily to the Theoretical Knowledge aspect. Mandated Coures

primarily corresponded to Technical Execution. HOPS was largely neglected in many programs. In addition, program descriptions were compared to mandated courses to determine the extent of correspondence. There were significant differences between the aspects discussed by the program description and the aspects covered in the mandated courses. Similarly, program prerequisites and mandated course prerequisites were compared in order to identify potential gates to program access and program success. There were not significant differences between program prerequisites and mandated course prerequisites. Finally, the paper concludes with a discussion on the current state of master's level programs and recommendations for future directions of program development.

LITERATURE REVIEW

In order to determine what should be taught in a graduate level data science program it is necessary to understand what the discipline of data science is. The field of data science has its origins in statistics, with statistician John Tukey proposing the field of data analytics in response to the emergence of big data and increasingly complex data analytics problems. While taking some inspiration from statistics, this new field is different from statistics in several key ways. In this new field, later to be called data science, Tukey (1962) outlines that the problems this field grapples with should come from real world issues that produce complex or big data. This field does not do mathematical research, it does scientific research. In terms of methodological approaches to these complex, real-world problems, analysis should be data driven. The goals of analysis and the tools of analysis should be selected after an exploration of the data. Understanding the context of the data is just as important to the analysis as the data itself. The problems examined and the methods used in data analysis are fundamentally different from the problems examined and methods used in statistics. Data analysis can adopt tools and methods from statistics, but that does not make it a form of statistics. Data analysis is more flexible in its problem approach and priority should be placed on applying statistical tools and models that work well and consistently on a variety of different data sets. Data analysis does not focus on developing optimized or perfect solutions for one particular data set. Instead, the focus is on producing tools inspired by numerous different disciplines and creating methods and tools that work well on many different datasets from many different disciplines. In summary, Tukey (1962) has outlined a discipline that is focused on the application and refinement of statistical and/or mathematical tools in order to solve complex, big data problems in other disciplines. While Tukey provided a loose plan of data analytics education, primarily composed of how

instruction of data analytics should be executed (avoidance of cookbookery, hands on approaches, no theory for theory's sake), William Cleveland developed a specific curriculum for teaching data analytics, which he refers to as data science. This marks a shift from the question Tukey raised which was, "What is Data Science?" to "What should be taught in Data Science?" A curriculum plan is important for understanding what a discipline values because it communicates what experts in the discipline deem as important for a practitioner of the discipline to know in order to be a practicing member of the discipline. A curriculum plan provides the criteria a nonmember must satisfy in order to become a member of the discipline, thereby communicating the values of the discipline.

The curriculum plan Cleveland develops heavily aligns to the values of the discipline Tukey outlined. Cleveland believes that twenty-five percent of the curriculum should be based on multidisciplinary investigations. Twenty percent of the curriculum should be in building models and methods of data analysis. Fifteen percent of the curriculum should be allocated to computing and five percent should be in tool evaluations. Fifteen percent of the curriculum should be focused on pedagogy. The last twenty percent of the program content should be composed of the theory of data science. In this plan, sixty-five percent of the program's instruction is centered on application of data analysis tools to real world problems. (Cleveland, W.S., 2001). Theory instruction is deprioritized in this curriculum outline, comprising only twenty percent of the program. This aligns with a recommendation Tukey makes, saying that mathematical theory should only be learned on an as needed basis. Theory is still taught, but in a relatively small percentage communicating it is valued less than skills related to application of methods and tools.

Cleveland built upon the value of multidisciplinary collaboration introduced in Tukey's

outlining of the field of data science. Cleveland proposes that fifteen percent of a data science curriculum should be allocated toward studying and teaching the pedagogy of data science. He reasons that data science is useful for learning about the world, so data science education should not be limited to university educational settings. (Cleveland, 2001). Nonmembers of the data science discipline should be able to access and interact with the data science discipline. This particular idea is furthered in the curriculum plan proposed by the National Academies' Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective, henceforth the NASEM Committee.

The curriculum plan provided by the NASEM Committee (2018) outlines areas that should be addressed in an undergraduate data science program. This outline states that an undergraduate program should cover the following ten areas, "mathematical foundations, computational foundations, statistical foundations, data management and curation, data description and visualization, data modeling and assessment, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving." (p. 22). This curriculum plan does elevate theory instruction slightly to thirty percent of the curriculum. Technical Execution skills are still the predominant focus of the program, comprising fifty percent of the program. However, the NASEM Committee has introduced a new value to the discipline through the explicit inclusion of communication and teamwork and ethical problem solving. Both of these skill areas convey a concern with humans involved with and impacted by the discipline of data science. Cleveland and Tukey did lay the groundwork for the inclusion of communication and teamwork in their outlines by stressing the importance of multidisciplinary collaboration. Teamwork and communication are necessary for multidisciplinary collaborations. However, the NASEM Committee is separating the value of

multidisciplinary collaboration into two separate aspects. Teamwork and communication are related to the HOPS aspect. Domain-specific considerations are a more related to Technical Execution as they are focused on doing data science with data coming from a different discipline.

In addition, the NASEM Committee is further developing the aspect of HOPS through the inclusion of ethical problem solving as a skill that should be taught to all data science students. (National Academies of Sciences, Engineering, and Medicine, 2018). This skill inclusion communicates that data scientists should not only be able to work with and communicate with others, but also understand how their actions as data scientists affect others.

Through the examination of the three frameworks discussed above, there are three distinct aspects within data science that a data science program ought to convey- Theoretical Knowledge, Technical Execution, and HOPS. Theoretical Knowledge pertains to all the foundational, mathematical skills required to understand and build data science tools. Technical Execution pertains to the ability to select the appropriate data science tool, implement it, and evaluate the result. HOPS has two parts. The first being skills typically referred to as soft skills like teamwork and communication. The second being skills that are concerned with the impact of the analyst's choices on others, like ethical thinking skills and legal/data privacy knowledge.

Examining curriculum plans is useful for understanding what values are deemed important to convey to students of a discipline. Examining educational materials is also useful for understanding what values are being conveyed to students. Introductory data science textbooks often provide a brief definition of what data science is. These definitions can be used

to understand what values of data science are being conveyed to students in the classroom setting.

Data Science (Kelleher and Tierney, 2018) defines data science by the activities done in data science. They say that data science captures, cleans, and organizes unstructured data into structured data in order to process, analyze, extract, and learn from patterns hidden with the data while being conscious of data ethics and data regulations. This definition conveys the aspect of Technical Execution by focusing on the actions that a data scientist does. This definition of data science also conveys the aspect of HOPS by asserting that the data scientist must also be aware of the ethical and legal implications of their analysis choices while executing a data science task.

Data Science: Techniques and Intelligent Applications (Chavan et al., 2023) provides a comprehensive definition of data science that differentiates between what the field of data science is composed of, what the goals of data science are, and what data science is the study of. Regarding what fields data science is composed of, Chavan et al., says, "Data science is a combination of two or more fields that uses different kinds of math and statistics, scientific methods, specialized programming, artificial intelligence, data analysis, algorithms, and systems for the extraction of knowledge from the data." (p., 3). A data scientist "prepares data for analysis, expands data science problems, makes data-driven solutions, analyzes data, and searches the high-level decisions in a broad range of application domains." (p.,3). Data science is however the study of "...the massive amount of data that includes extraction of meaningful insight from structured and unstructured data which is completed using different algorithms and scientific methods." (p.,3). In terms of value communication, again the priority is placed on Technical Execution. In describing the goals of data science, Chavan et al., describes a workflow that is a combination of Technical Execution and HOPS. Data preparation, creation of

data-driven solutions, and data analysis are all activities that require the use of different tools and methods from various disciplines mapping it to the aspect of Technical Execution. Defining and expanding data science problems and searching high-level decisions in a broad range of applications are activities that involve communication, critical thinking, and potentially teamwork skills, mapping these skills to the value of human oriented professional values. The third and final part of the definition provided by Chavan et al. focuses on how meaning is extracted from data. This directly maps to the aspect of Technical Execution, as the action of meaning extraction through the use of various tools and methods is the focus of the definition. Overall, this definition of data science communicates the aspects of Technical Execution and HOPS.

The Practitioners Guide to Data Science (Lin, H., & Li, M., 2023), defines data science through three different tracks: engineering, analysis, and modeling/inference. Data engineering curates, organizes, and formats the data. Analysis performs exploratory analysis of the data to understand the data. Modeling/Inference employs various statistical and mathematical tools for pattern detection and identification. All responsibilities of the provided roles map to Technical Execution. The responsibilities of each role are all centered around the different tools and methods they use to either prepare the data for a data science effort or to execute the analysis. It is important to note in this textbook, HOPS are discussed in other chapters as important skills a data scientist should have, however, they are not mentioned in the roles used to define data science, communicating that while these skills are important enough for inclusion, they not intrinsically linked to the definition of data science.

The Foundations of Data Science (Blum et al., 2018) differs greatly from the other textbooks described above. This textbook focuses solely on the mathematical foundations that

are required to do data science. The authors focus primarily on the relationship between data science and the discipline of computer science. This connection is largely unaddressed in the other textbooks. While not providing an explicit definition of data science like the previous textbooks, this textbook does make transparent the theoretical knowledge a practitioner of data science is expected to know. The skills outlined are geometry, linear algebra, and calculus. These theoretical skills are the basis for understanding how singular value decomposition (SVD), the perceptron algorithm, and various forms of algorithmic learning work. While the other textbooks focus on the application of various tools and methods, this textbook is concerned with building the theoretical skills necessary for understanding and potentially developing or refining data science tools. This book exclusively conveys the aspect of Theoretical Knowledge.

The educational materials used in data science curriculums do convey all three aspects of Theoretical Knowledge, Technical Execution, and HOPS. There is a distinct preference for the aspect of Technical Execution as this aspect is discussed most frequently in the education materials. Defining data science by a set of skills is convenient as Technical Execution is the most visible aspect of data science. However, as *The Foundations of Data Science* (Blum et al., 2018) points out, there is an amount of mathematical knowledge required in order to understand how to apply tools and methods correctly and appropriately. In addition, HOPS are also required for appropriate and correct technical execution as other disciplines need to be consulted and analysis choices need to be informed through the context of the data and their potential impact on others.

It is important to understand if all three aspects of data science- Theoretical Knowledge, Technical Execution, and HOPS- are conveyed in teaching institutions and to what extent each aspect is conveyed. There is a growing body of work in this area. Oliver and McNiel (2021),

examined twenty-five undergraduate data science programs in the US to determine the content being taught. They found that a majority of their sampled programs heavily favored computer science and statistics instruction and had very little priority placed on domain knowledge building courses, courses centered around data ethics, or data communication-based courses. One may argue that undergraduate education is more broad and that undergraduate programs are focused on building foundational skills, explaining why the undergraduate curriculum conveys the aspects of Theoretical Knowledge and Technical Execution predominantly. Professionalization and the human oriented aspects that come with it should occur at the graduate level.

However, Tang and Sae-Lim (2016) did not find a marked increase in human oriented courses in graduate level data science programs. In their study thirty graduate programs across the United States were examined, with the goal of understanding how the program describes itself, the structure of the curriculum, and the course content taught. They found that data science programs rarely prioritize human oriented professional skills like communication and visualization in their core classes. Communication and visualization based courses comprised thirteen and twelve percent of the core curriculum respectively of the programs they examined. Regarding the examination of HOPS, Tang and Sae-Lim (2016) were limited in the skills they coded for. While communication and visualization are important skills for data scientists, they are not solely representative of the human oriented skills a data scientist needs to have. In addition, they did not examine the relationship between how a program positions itself in its description and the content actually taught in the program.

Curriculums and education materials matter a great deal in communicating the values of a discipline, however, it is actual members of the data science discipline that perform the action of

communicating. Understanding who these communicators are is important. Their educational backgrounds and trainings affect which aspects of the data science discipline they choose to communicate. In a survey done of thirty-eight higher education institutions and two independent research facilities, Norén et al. (2019) found that overall fifty-four percent of researchers at these institutions came from a computer science, statistics, math, physics, or engineering background. When interacting with data science, these disciplines predominantly provide theoretical and technical execution based instruction. Traditionally, these disciplines do not provide training in HOPS. It is unlikely that members of these disciplines who have transitioned into teaching data science are able to pass on HOPS as their own trainings were unlikely to cover these skills. Social science and the humanities only composed seventeen percent of the researchers at these institutions. Members of these disciplines are far more likely to have received training in HOPS. When they transition into data science, they are more likely to be able to provide HOPS trainings to their students.

CURRENT PROJECT

The goal of this thesis is to understand how the three aspects of data science; Theoretical Knowledge, Technical Execution, and HOPS are communicated and executed in data science masters programs across the United States. This goal will be accomplished through the examination of program description, program prerequisites, mandated courses, mandated course prerequisites, and availability of data ethics courses. The following questions will be addressed:

1. What aspects of data science do master's programs prioritize?
    a. Are the aspects described in the program description congruent with the aspects being taught in the program?

2. What skills are being developed in the master's program and which aspects do those skills align with?
    a. Which aspects do master's program admission criteria reflect?
    b. What aspects are being conveyed in the mandated courses of the master's program?
    c. Are the aspects in the admission criteria congruent with the aspects in the prerequisites of the mandated courses?
3. How is the emerging concern of HOPS being accounted for in master's programs?
    a. Is one aspect of HOPS being valued over the other?
    b. To what extent are data science students being exposed to ethics?

METHOD

Sample Selection

Master's level data science programs from United States institutions were initially selected through online rankings. Different program rankings tended to include many of the same data science programs, so additional recommendation lists for students looking to apply to data science programs were used. Completely online programs were excluded from this study. Online programs are typically geared for the working professional and their curriculum and goals are different from an in-person program. Programs with hybrid elements that allowed the student to choose their attendance modality were included as the programs do have a fully in-person option available. In addition, programs that had separate online programs were included, but no information about the online version of the program was utilized for the study. Some programs utilized variations on the title of data science, such as data analytics. Programs that did not refer to themselves as data science were included only if their curriculum covered topics in statistics, computer science, and machine learning. Approximately fifty programs were excluded

for failing to meet this criterion. This inclusion criterion was set for comparison fairness purposes. All data science programs that referred to themselves as data science programs provided classes that covered statistics, computer science, and machine learning topics. Certain data analytics programs only covered one or two of these topics and included classes that were outside the scope of data science and that did not correspond to or build upon the theoretical skills necessary to practice data science.

In total, sixty-five programs were initially selected for this study. In order to be eligible for inclusion, all programs needed to have publicly available program description, a uniform set of mandated course requirements, mandated course descriptions complete with prerequisites, and program admission criteria. Two programs had a sub track system with no shared core requirements for each sub track and were removed for failing to meet. One program did not have mandated course descriptions publicly available. This left sixty-two programs for analysis. (Appendix I).

Data Collection

As outlined in the literature review, there are three aspects of data science- Theoretical Knowledge, Technical Execution, and HOPS. A master's program may choose to reflect all or only some of these aspects through how they publicly describe their program, what courses they mandate, the prerequisites for those required courses, the program admission requirements, and the inclusion of human oriented professional skills in the curriculum. This information is available from program websites and university course catalogs.

Program descriptions were taken from master's program websites. Information qualifying as a program description includes broad overviews of the curriculum, teaching methodology, learning outcomes, and clarifying points regarding program content under FAQ

sections. These bodies of text were typically labeled as “About” or “Our Program”. Mandated courses were taken from the requirements or curriculum sections of program websites. The mandated courses were typically designated under the headings “Mandated”, “Core”, or “Required”. Corresponding prerequisites for the mandated courses were found in the university course catalog. The most recent and publicly available university catalog was used. Course descriptions for the mandated courses were either provided directly on the program website, were located in the university course catalog, and/or found on published syllabi. Program prerequisites required for application eligibility were on the program website. They were typically found under the headings “Admission”, “Eligibility”, “Requirements”, or in the program FAQ section. Ethics classes were found in either the mandated courses section of the program website or in the offered elective section of the curriculum description.

Data Coding

In order to code the collected data for which aspects of data science are reflected, two qualitative codebooks were developed. One for the program descriptions and one for mandated classes. A more detailed description of these indicators can be found in Appendix II. Mandated courses were coded following the codebook in Appendix II. Prerequisites for both the program and the mandated classes were assigned to specific categories following the criteria outlined in Appendix II.

Program Description Indicators

Indicators for program descriptions were developed through an iterative process. Beginning with the aspects of data science- Theoretical Knowledge, Technical Execution, and HOPS - a small sample of programs was examined looking for words or phrases that provided evidence for a particular aspect. Common themes among this sample pool were identified and used as indicators. This set of indicators was applied to another small sample pool and refined.

This process was completed when no additions or adjustments were made to the sets of indicators.

Theory Indicators

The aspect of Theoretical Knowledge has few indicators. Typically, programs will make an explicit theory statement, asserting that they value **theoretical education**. Some programs will also assert that their master's program is geared towards **PhD program preparation**. Other programs will make multiple references towards providing a strong **foundational education**. As outlined in the literature review, the foundational skills for data science are predominantly skills in maths, including linear algebra, calculus, and probability and statistics. These skills are the theory oriented foundations which then allow for technical execution to occur.

Technical Execution Indicators

The aspect of Technical Execution has very overt indicators. Indicators referenced both the structure of the program and content covered in the program. Structure related indicators identified programs that served working students or professionals looking to change or advance their careers. Programs serving this group of people are focused more on technical execution because that is what is required by industry. The structural indicators are the availability of **hybrid or evening/weekend classes** and the inclusion of an **internship/practicum**. The content indicators are the references to the courses being **application oriented** and having **industry focused curriculum.**

Human Oriented Professional Skills

The aspect of HOPS is composed of two parts. These two parts are Responsible Data (HOPS: RD) and Soft Skills (HOPS: SS). These two parts each have a corresponding set of indicators.

Responsible Data

Skills corresponding responsible use of data are still human oriented skills as they are concerned with responsible custodianship of data. **Data ethics**, **data privacy/security/legal** considerations of data scientists, and **designing and evaluating** appropriate experiments and data collection methods all are related to the data science project pipeline. These skills are typically explicitly stated in program description.

Soft Skills

Soft skills are skills that facilitate collaboration, an integral part of data science. These skills include **communication**, **teamwork**, **data visualization**, and **professional skills**. Programs will list these skills specifically. Programs that value soft skills will provide career services to students in order to practice interviewing, public speaking, networking, and various other skills necessary for working in either industry or academia on a multidisciplinary team.

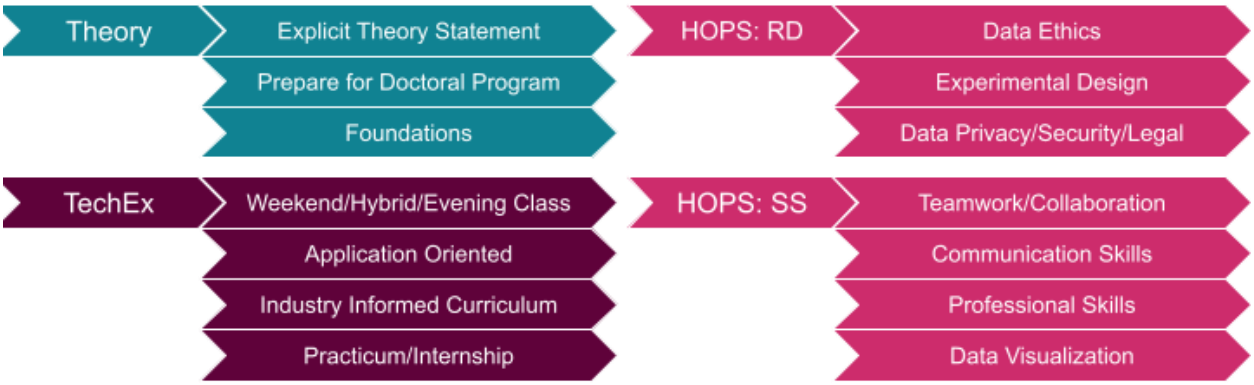

Figure 1: Aspects and Corresponding Indicators

Validation

A sample of twenty percent of the program description coding data was taken in order to perform validation coding. This sample was delivered to two independent raters along with all text used from the program websites in order to perform the original coding and the codebook.

Raters performed their coding individually. The Light's kappa coefficient was 0.721 with $p < .01$. This indicates substantial agreement between the independent raters. Only the program description data had validation coding performed because the program descriptions provided by the universities.

Mandated Courses

Course descriptions were more overt in describing the content that would be covered in the class than the program descriptions. If a class covered probability and statistics, linear algebra or another foundational mathematical skill, it was coded as a one for Theoretical Knowledge. In addition, if the course lacked any indication there were Technical Execution aspects and covered nine topics or less, or explicitly stated it covered foundations and theory, it was coded as Theoretical Knowledge. The rationale being that one course would not have the time to cover nine topics if the focus was on the theory behind those topics. However, up-to-nine topics is consistent with being theory-centric. The UCSD Machine Learning course only includes nine topics in its course description and this course was confirmed through a post-hoc evaluation to only teach theory related to machine learning, no implementation instruction was provided.

Mandated courses covering applied statistics, databases, and/or programming were coded as Technical Execution. In addition, any program discussing applying concepts to real world data sets, a lab component, data engineering, having a programming prerequisite, and/or having students perform any task similar to what they would do in industry was coded as Technical Execution. HOPS: RD were very straightforward to code for. Any course covering critical thinking/evaluation of experimental design, data ethics, and/or legal/security/privacy aspects of data science was coded as HOPS: RD. HOPS: SS were also very straightforward. Any courses

covering collaboration, teamwork, data visualization, or any form of other professional skills were coded as HOPS: SS.

Numerous programs had courses that covered more than one of these aspects. For these cases, fractions summing to one were assigned to the aspects of data science the course covered. For example, if a course description described the course as covering implementation of common deep learning frameworks and the ethics of applying deep learning in healthcare the course would be coded as <0, .5, .5, 0>. A .5 is placed in the y component of the vector corresponding to the aspect of Technical Execution because one focus of the class is on implementation. A .5 is also placed in the z component of the vector corresponding to HOPS: RD because the other focus of the class is on ethics. If a course description was to say this course is focused on solving a real world problem by designing and implement a solution, discussing the ethical impact of the proposed solution, and presenting the solution to industry partners with a team, the course would be coded as <0, ⅓, ⅓ , ⅓ >. This is because the course covers Technical Execution, HOPS: RD, and HOPS: SS.

In addition, if a course description was unclear or limited in the information it provided, the course syllabus was used to provide more clarity on the goals of the course in order to make an accurate determination of which aspect the course values. If a syllabus was used to clarify course goals, the most current copy of the syllabus locatable was used. If a syllabus could not be located, then only the information available in the university course catalog was used.

Some programs allowed mandated courses to be selected from two options. In this case, fractions were assigned to the relevant aspects represented in the course descriptions so that when added together the sum of both courses was equal to one. For example, suppose Course A and Course B can be interchanged, where Course A is coded as 0.5 for Technical Execution and

0.5 for Theoretical Knowledge, while course B is coded as 1.0 for Technical Execution. The code for the "A or B" requirement, weighing the courses evenly, is <0.25, 0.75, 0, 0>. Programs had a high variety in their course descriptions of mandatory capstone or thesis projects. If information about the expected project was given, it was coded based on the criteria previously established for aspect assignment. If there was no information available or the course description said that expectations would be set by the chosen faculty advisor the course was coded as NA. No expectations for the resultant capstone or thesis project communicates that the student may choose any aspect of data science they would like to pursue.

There were two programs that did not have any mandated courses. They instead had a pool set up, where students are given three or more courses to choose from in a certain topic area to satisfy their course requirements. While not as explicit in value communication as a distinct set of mandated classes, a pool set up still does communicate the aspects of data science a program values, through the contents of the pool. For these programs, the titles of the pools were used as the mandated class names, with a number added to the end of the pool title to denote how many courses were expected from that pool. For example, if a student needed to select two classes from a pool titled Algorithms, the mandated course names would be Algorithms 1 and Algorithms 2. The aspects these courses correspond to were determined by examining the individual classes listed in the pool to determine the predominant aspect. For example, if a majority of the courses in the pool forced on teaching Technical Execution then the mandated courses for that pool were represented by coding a one for Technical Execution.

Program Prerequisites

Program prerequisites varied in terms of specificity, some programs outlined specific skills that students needed to arrive with, and others listed topic areas students should be familiar

with. These program prerequisites were typically found in application FAQ areas or listed eligibility sections. Each program prerequisite had its own vector to map it to the aspect of data science it corresponded to. Any mathematical skills required, like linear algebra, calculus, or probability/statistics were coded as Theoretical Knowledge. Any programming or computer science oriented skill, like computing, data storage, and/or algorithms was coded as Technical Execution. Any skills related to data ethics/privacy/security or critical thinking about experimental design were coded as HOPS: RD. Skills related to teamwork, collaboration, communication, or other professional skills were coded as HOPS: SS. For example, a program may require students to have taken linear algebra and calculus in order to be considered for admission. This program would have two vectors because there are two program prerequisites, one vector for linear algebra and one vector for calculus. Linear algebra is only a Theoretical Knowledge skill. The resulting vector would be coded as <1,0,0,0>. Calculus also corresponds only to the area of theoretical knowledge. Calculus would be coded as <1,0,0,0>.

If programs listed courses like Calculus I, Calculus II, and Multivariate Calculus, then only the highest level prerequisite was used for the coding. This is because in order to have taken Multivariate Calculus, Calculus I and II would have had to be completed. This was done in order to prevent overweighting an aspect of data science. Some programs only list Multivariate Calculus and have the assumption that Calculus I and II were completed. Other programs list every calculus class that must be completed. In addition, there was only discrete coding done as the program prerequisites did not correspond to more than one aspect of data science. Attention was paid to the descriptions of the requirements, if a university clarified that the statistics they expected were more applied, like linear or logistic regression then that statistics requirement was coded as technical execution (<0,1,0,0>).

Mandated Courses Prerequisites

Mandated course prerequisites also had a high amount of variation in their formatting. Some mandated courses would use a topic area as a prerequisite, like linear algebra or calculus. Other mandated courses would use a specific class with a course code as a prerequisite. For the first case, these topic-based prerequisites were coded in accordance with the procedure established for the program prerequisites. For the second case, the course descriptions were analyzed in accordance with the procedure established for the mandated courses. In addition, each mandated course prerequisite received a Yes/No designation if the prerequisite was taught anywhere in the program curriculum, either as an elective or mandated course. Items in mandated course descriptions that we labeled as recommended preparation were not considered prerequisites.

In situations where programs had a pool set up, the courses with the least amount of prerequisites were chosen to represent the prerequisites for the pool. The least amount of prerequisites was used because those would be the minimum qualifications necessary in order to complete the mandated courses. If courses all had the same amount of prerequisites, the prerequisite that appeared the most frequently was selected. If two different prerequisites appeared the same amount of times, the prerequisite that was taught elsewhere in the program was selected. For example, a pool has four courses and two courses have one prerequisite of linear algebra and the other two courses have a prerequisite of statistics. If there is a statistics course taught in a different part of the program, either as an elective or in another pool, then statistics is selected as the prerequisite of the course.[1]

[1] 0.25 was only assigned to two universities. The Johns Hopkins Master of Data Science program. Ethics is required to graduate; however the students are required to take an online ethics class offered through Coursera. The University of Minnesota briefly covers ethics in an optional machine learning elective.

Ethics

Modalities and requirements for ethics courses varied greatly across programs. A zero to one discrete categorization scale was developed in order to assign a numerical weight to the value the program places on data ethics based on the modality of the ethics course, if an ethics course is available in the curriculum, and if there is an ethics requirement. The highest value of the scale is a one, meaning that an ethics course is offered and required as a part of the core curriculum. The lowest score is zero, indicating there is no ethics course required or offered in the program. A 0.75 is assigned if there is not a requirement for students to take ethics, but a full, standalone ethics course is provided. A 0.5 is assigned if a mandated course in the program covers ethics, but that is not the focus of the course. A 0.25 is assigned if there is a non-mandated course that contains a section on ethics, or there is an ethics requirement, but the ethics instruction is not offered at the university.

ANALYSIS AND RESULTS

All data analysis was performed in RStudio using version RStudio 2023.12.1+402.

What aspects of data science do master's programs outwardly communicate they prioritize?

In order to answer this question, program descriptions were examined. Examining the frequency an indicator is used revealed that fifty-nine of sixty-two programs mention being application oriented and forty-four mention having industry oriented curriculum. Both of these indicators correspond with the Technical Execution aspect of data science, potentially suggesting that the majority of programs outwardly project that they value Technical Execution. The next most frequently used indicators are communication skills, data visualization, and teamwork and collaboration. These indicators all correspond to HOPS: SS. The combination of the most

frequently used indicators belonging to Technical Execution and HOPS: SS may suggest that the master's programs in this sample have the larger goal of preparing students for industry roles as industry requires both strong professional and technical skills.

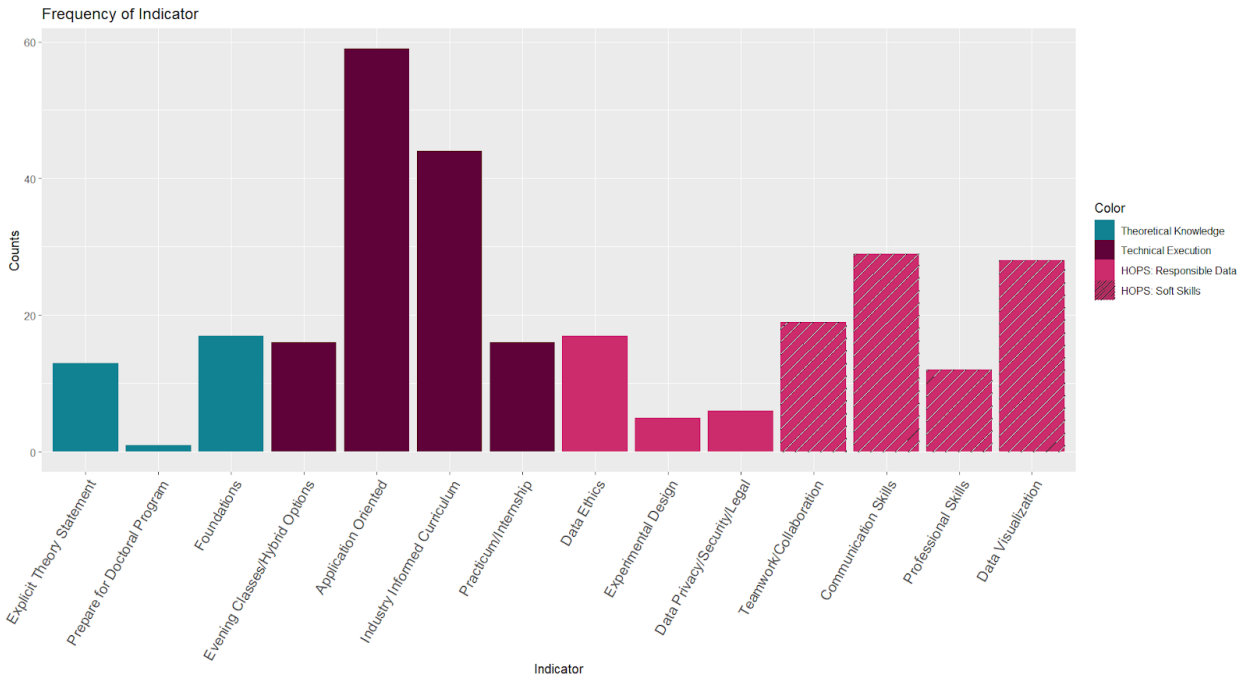

Figure 2: Bar graph depicting frequency of indicator occurrence

In order to view the universities in relation to each other, a simplex was plotted. In order to create the simplex three dimensional vectors were created by summing all the indicators that corresponded to one aspect of data science. For example, if a program was coded so there was a one in "Explicit Theory Statement" indicator, a zero in "Prepare for Doctoral program", and a one in "Foundations", the aspect value for Theoretical Knowledge would be two. The aspect values for HOPS: RD and HOPS: SS were summed to produce one value for HOPS. This produced a vector of three values corresponding to the three aspects of data science. The aspect values for Theoretical Knowledge, Technical Execution, and HOPS correspond to <x, y, z> respectively. In order to account for the unequal number of indicators for each aspect of data science, each aspect value was divided by the number of indicators for each value, three, four,

and seven respectively. This was done to avoid having one aspect overweight the other two just because it had more indicators. After this, these vectors were normalized by the sum of the vector so the vector would sum to one. For example, if a vector was <2,3,1>, the components would first be divided by the number of indicators for each aspect, making the vector <2/3,3/4,1/7>. Then, the vector would be normalized by the sum of all components to ensure the vector summed to one. This makes the vector < 56/131, 63/131, 12/131>. This normalization was done in order to establish relative importance of each aspect of data science. These normalized sum vectors were used for all statistical tests. For simplex plotting, the package MSCquartets was used to convert the vector to two dimensional coordinates.

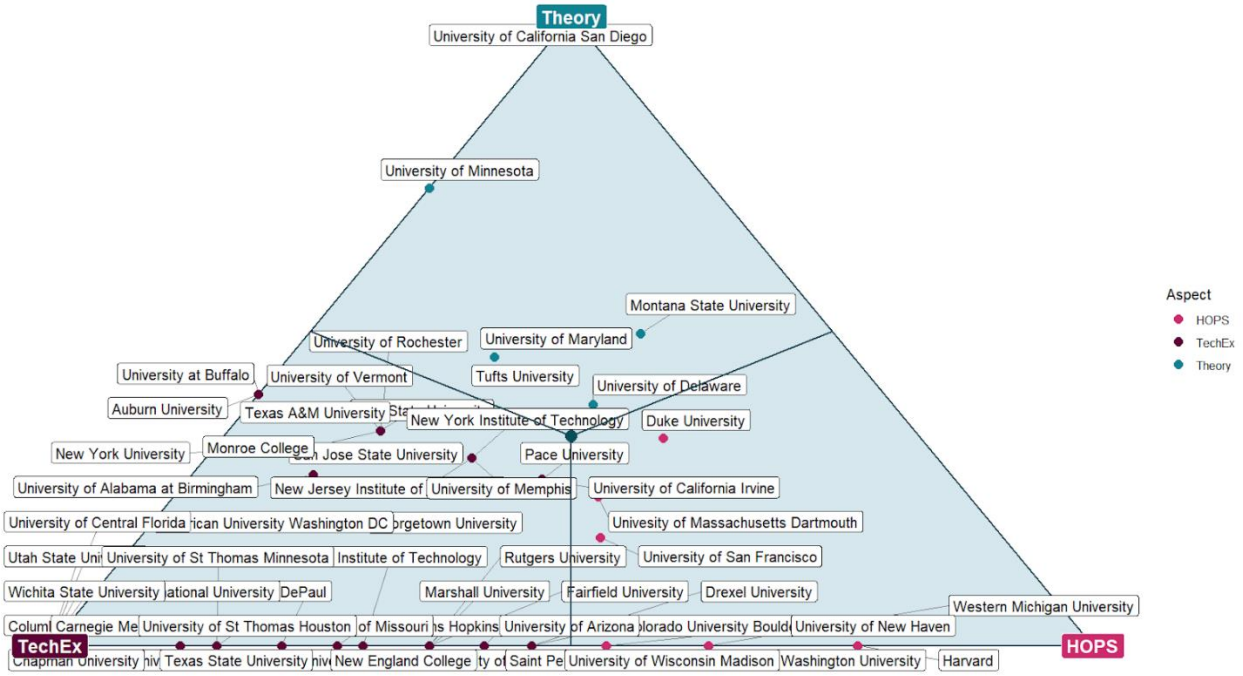

Figure 3: Simplex showing the distribution of program descriptions

Table 1: Region classification counts based on program description

Theory	TechEx	HOPS	Border
6	48	8	0

This simplex shows where programs fall in relation to the three aspects of data science. The simplex is split into three regions. Each region corresponds to an aspect of data science. If

a university falls into a particular region, it means they predominantly favor that aspect of data science. Most of the universities fall in the lower half of the simplex, heavily favoring the Technical Execution region. Few universities discuss the aspect of Theoretical Knowledge to the point where it is the predominant aspect in their program description. The University of California San Diego is the most extreme in its discussion of theory, only covering theory in their program description. A majority of the universities are spread between Technical Execution and HOPS, confirming the indication in Figure 2.

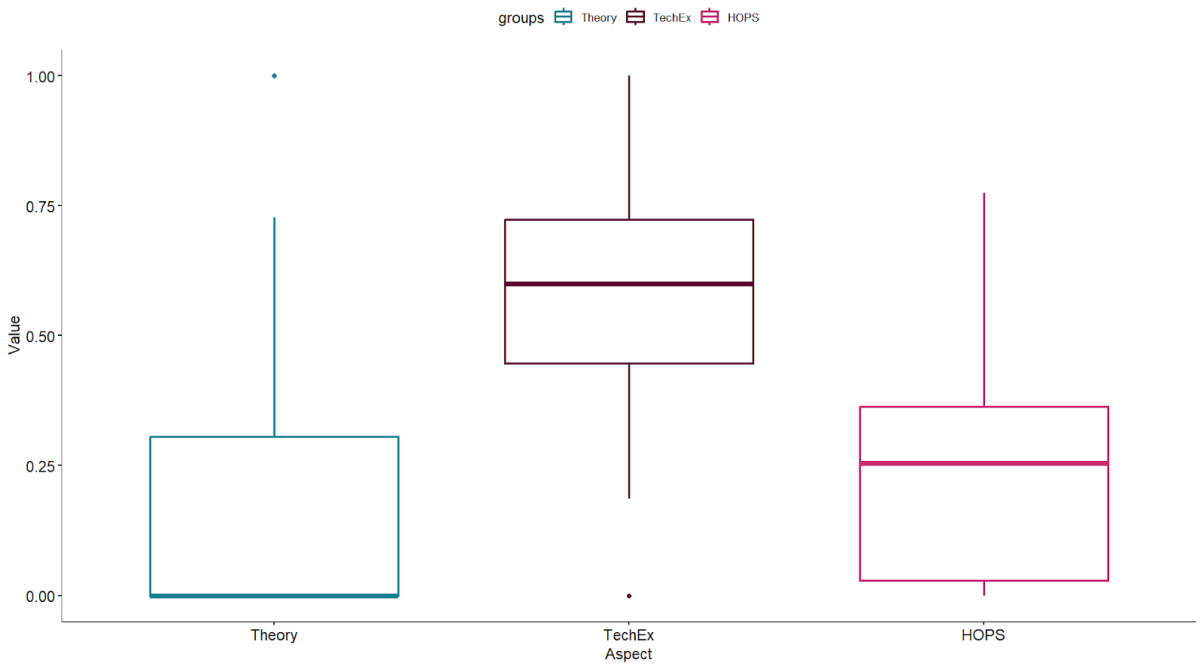

Figure 4: Boxplots of program description's normalized aspect values

Examining a boxplot of the normalized vector components confirms that Theoretical Knowledge is discussed less in program descriptions than both Technical Execution and HOPS. The average value of the Theoretical Knowledge component is 0.15. This is smaller than the average values for Technical Execution and HOPS, which are 0.60 and 0.24 respectively. Performing a Kruskal-Wallis Rank Sum Test reveals that there is a significant difference ($p <$

0.01) between the three aspects of data science. Performing a follow up Pairwise Comparisons using Wilcoxon Rank Sum Test reveals all three aspects are statistically different ($p < 0.01$) from each other.

What aspects of data science do universities teach?

In order to understand what data science programs teach, mandated courses were examined. Mandated courses were selected because these are the courses that all students must take in order to graduate. Mandating a certain class communicates that that class is integral to becoming a member of the data science community.

The mandated courses were coded, as described above, in a way where each course had its own vector corresponding to the aspects of data science. In order to produce a simplex for the mandated courses, all vectors corresponding to the mandated courses offered at a university were summed to produce one vector. The aspects of HOPS: RD and HOPS: SS were summed to produce vectors with a length of three. Using the package MSCquartets, this vector was converted to two dimensional coordinates and then plotted on the simplex shown below. Utah State University was omitted as their only mandated courses were thesis and data science seminar courses.

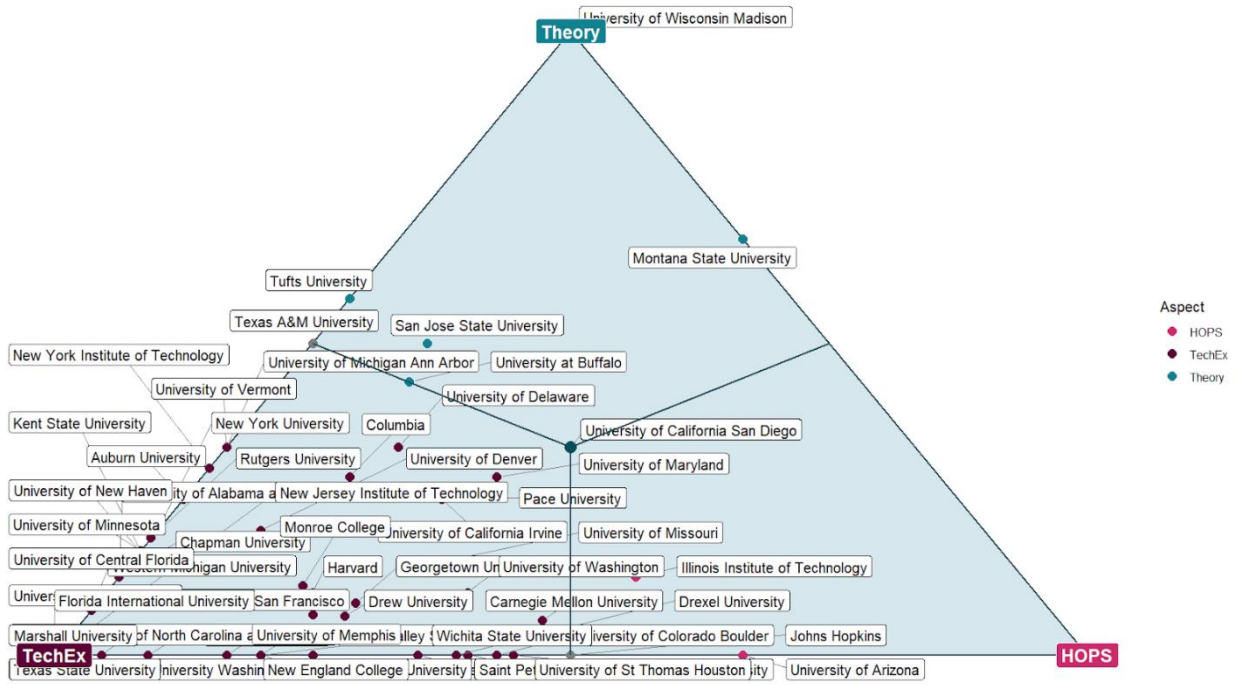

Figure 5: Simplex showing the distribution of mandated courses

Table 2: Region classification counts based on mandated courses

Theory	TechEx	HOPS	Border
6	51	2	2

This simplex shows that a majority of the universities fall in the Technical Execution region. Very few universities fall in the Theoretical Knowledge or HOPS region. This simplex appears to be a different distribution than the program description, which was more spread across the bottom of the simplex.

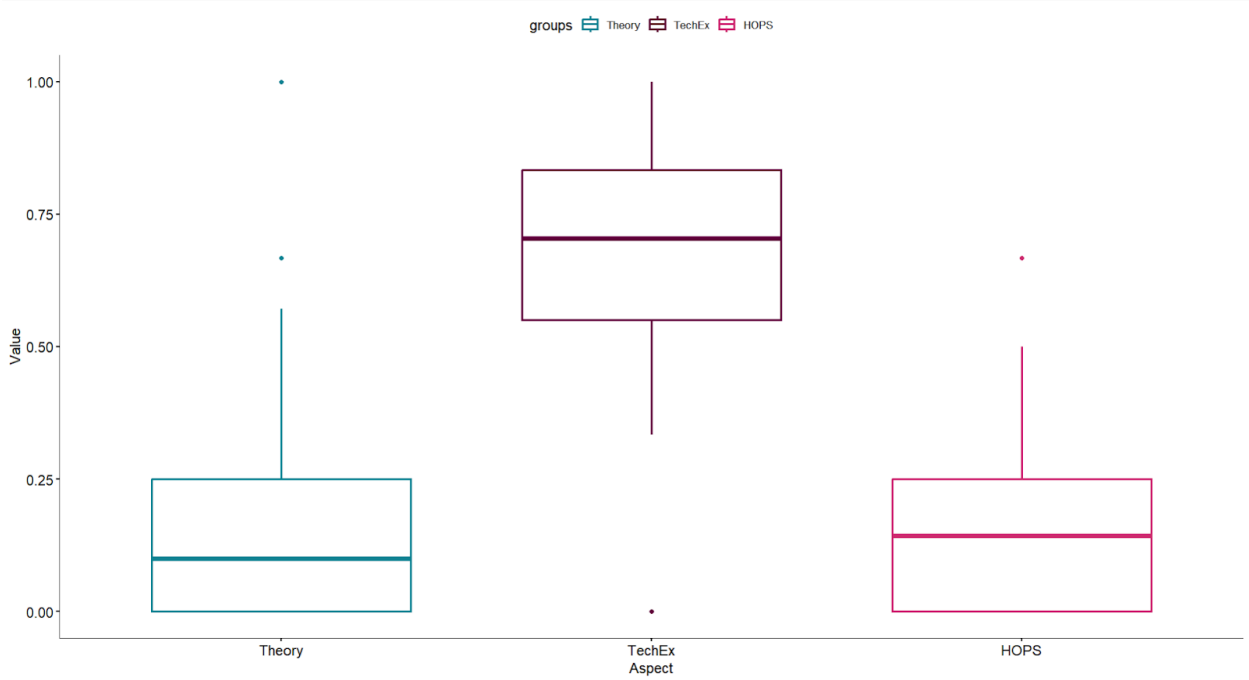

Figure 6: Boxplots of mandated course's aspect values

Technical Execution is most prevalent in the course content of the mandated classes. The average value of the Technical Execution component of the vector is 0.69. This value is higher than the vector components of the aspects of Theoretical Knowledge and HOPS, which are 0.15 and 0.16 respectively. Performing a Kruskal-Wallis Rank Sum Test reveals that there is a significant difference ($p < 0.01$) between the three aspects of data science. Performing a follow up Pairwise Comparisons using Wilcoxon Rank Sum Test reveals Technical Execution is statistically different ($p < 0.01$) from both HOPS and Theory. HOPS and Theory are not statistically different from each other.

Are the aspects described in the program description congruent with the aspects being taught in the program?

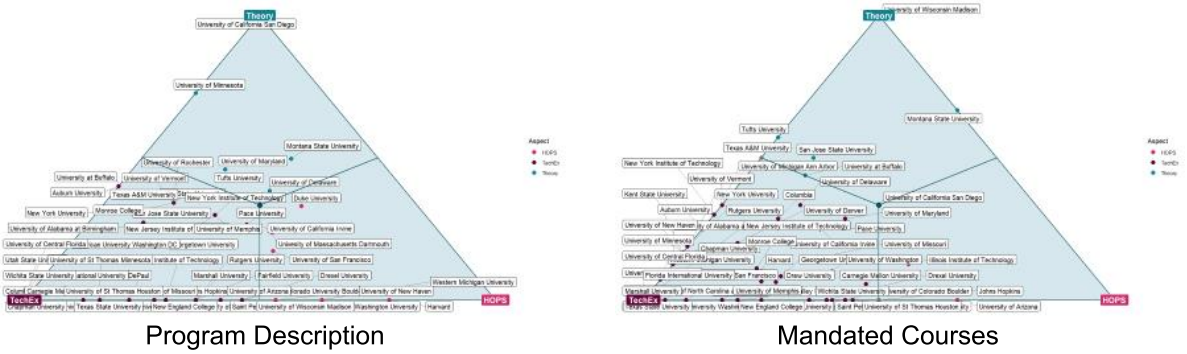

Figure 7: Simplex comparison between program description and mandated courses

Table 3: Region classification comparison between program description and mandated courses

	Theory	TechEx	HOPS	Border
Program Description	6	48	8	0
Mandated Courses[2]	6	51	2	2

In order to confirm that there is a difference between the distributions of program descriptions and mandated classes Hotelling's Two Sample T Squared Test was performed[3]. Due to the vectors being normalized by sum, the matrix required for the Hotelling's Two Sample T Squared Test was not invertible. In order to solve this, only two components from the vectors were tested at a time. This means the Hotelling's Two Sample T Squared Test was performed three times on the component pairs theoretical knowledge and technical execution, theoretical knowledge and human oriented professional skills, and technical execution and human oriented professional skills. All of Hotelling's Two Sample T Squared Tests were insignificant at $p < 0.01$, but significant at $p < 0.05$. The confidence intervals revealed that HOPS regions were

[2] Utah State was omitted as its mandated courses are seminar and thesis, making 61 programs in Mandated Courses

[3] Neither distribution satisfied the mShapiro test for normality, but the equal covariance assumption was met. The Hotelling's Two Sample T Squared Test is fairly robust to non-normality, so the test was proceeded with. Utah State was omitted from these tests, making only 61 points in each distribution.

different from each other in terms of distribution, suggesting programs are discussing these skills in their program descriptions, but not teaching these skills in their mandated courses.

Which aspects of data science do master's program prerequisites value?

The program prerequisites a program uses are important because those program prerequisites control who can access the data science program. Programs with program prerequisites gate who can and cannot participate in data science at a professional level. For example, a program that requires students to have linear algebra, calculus, and coding experience before they arrive to the program, limits what students can be admitted to the program. Students from more humanities oriented backgrounds may not be able meet the program prerequisites if the program prerequisites are all STEM based. Program prerequisites may limit the type and amount of students that can access a masters of data science program.

To create the simplex, all of the program prerequisite vectors for a university were summed to produce one vector representing the program prerequisites for the university. The aspect values of HOPS: RD and HOPS: SS were again summed. This produced a vector with a length of three to represent the three aspects of data science. These three-dimensional vectors were normalized by the sum of the vector to produce a vector with components that summed to one. The within-aspect normalization used for the program descriptions was not utilized for the program prerequisites as no aspect indicators were used for the coding of the program prerequisites. The package MSCQuarters was used to convert the vectors to two dimensional coordinates used to plot the simplex. The nineteen programs that did not have admission criteria were not included on this simplex. These programs are discussed later.

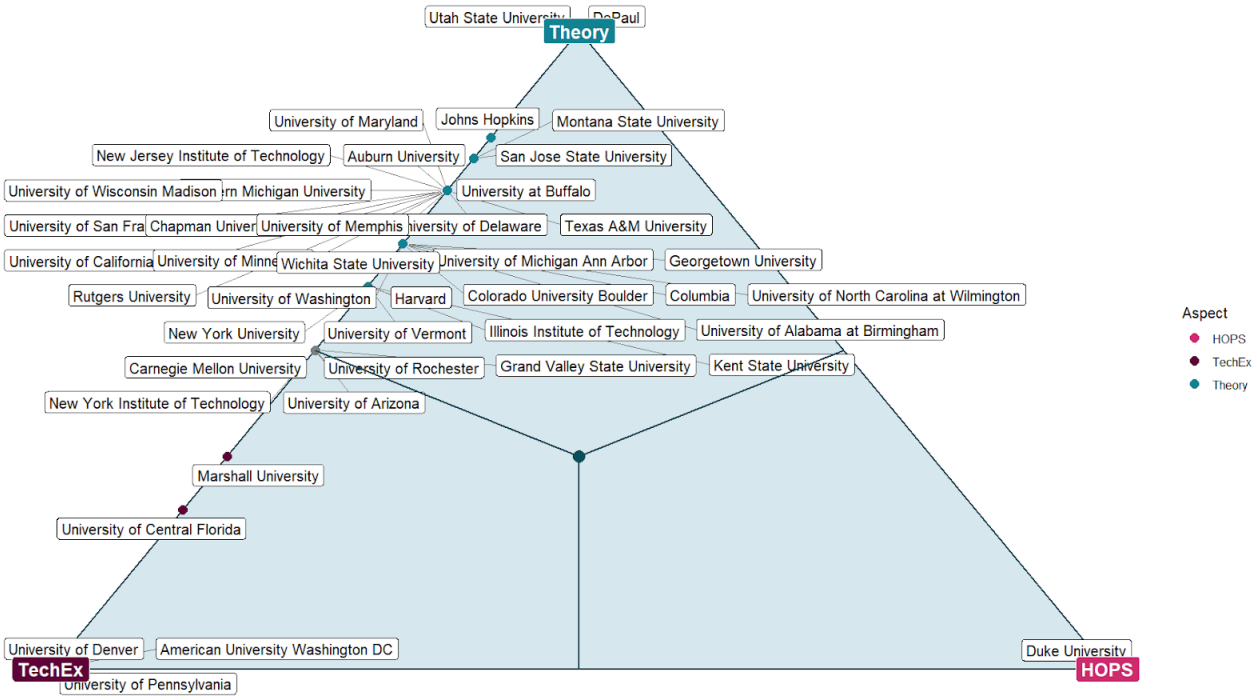

Figure 8: Simplex showing the distribution of program prerequisites

Table 4: Region classification counts based on program prerequisites

Theory	TechEx	HOPS	Border
32	5	1	5

The program prerequisites are predominantly theory oriented, however they are pulled towards the edge of the simplex indicating technical execution skills are also requested. Thirty-three of the universities fall into the Theoretical Knowledge region of the simplex. Programs tend to expect students to arrive with prior exposure to the theoretical knowledge aspect of data science and some technical execution skills. Duke University is the only university to ask that students arrive with HOPS and they only ask for HOPS.

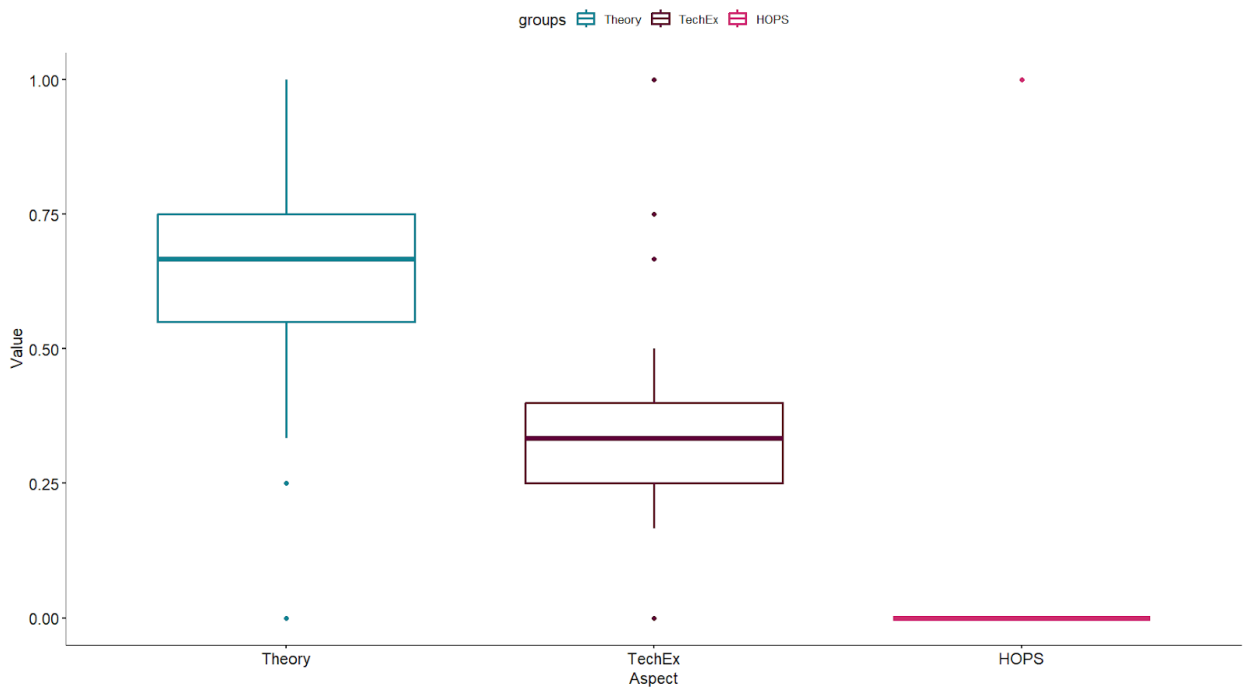

Figure 9: Boxplots of program prerequisites' aspect values

Examining a boxplot of the normalized vector components also suggests that the aspect of Theoretical Knowledge is most prevalent in the program prerequisites. The average value of the Theoretical Knowledge component of the vector is 0.61. This value is higher than the vector components of the aspects of Technical Execution and HOPS, which are 0.36 and 0.02 respectively. Performing a Kruskal-Wallis Rank Sum Test reveals that there is a significant difference ($p < 0.01$) between the three aspects of data science. Performing a follow up Pairwise Comparisons using Wilcoxon Rank Sum Test reveals that all three aspects are significantly ($p < 0.01$) different from each other.

Which aspects are being reflected in the mandated class prerequisites?

Program prerequisites should correspond to the mandated class prerequisites. Program prerequisites represent the foundational knowledge required to enter and participate in the program. Mandated course prerequisites represent the required knowledge necessary to participate in and understand the content in a mandated class. A program should set program

prerequisites that, at the minimum, include the skills required in the mandated course prerequisites. It is especially important that programs without program prerequisites provide foundational skills that satisfy the mandated course prerequisites.

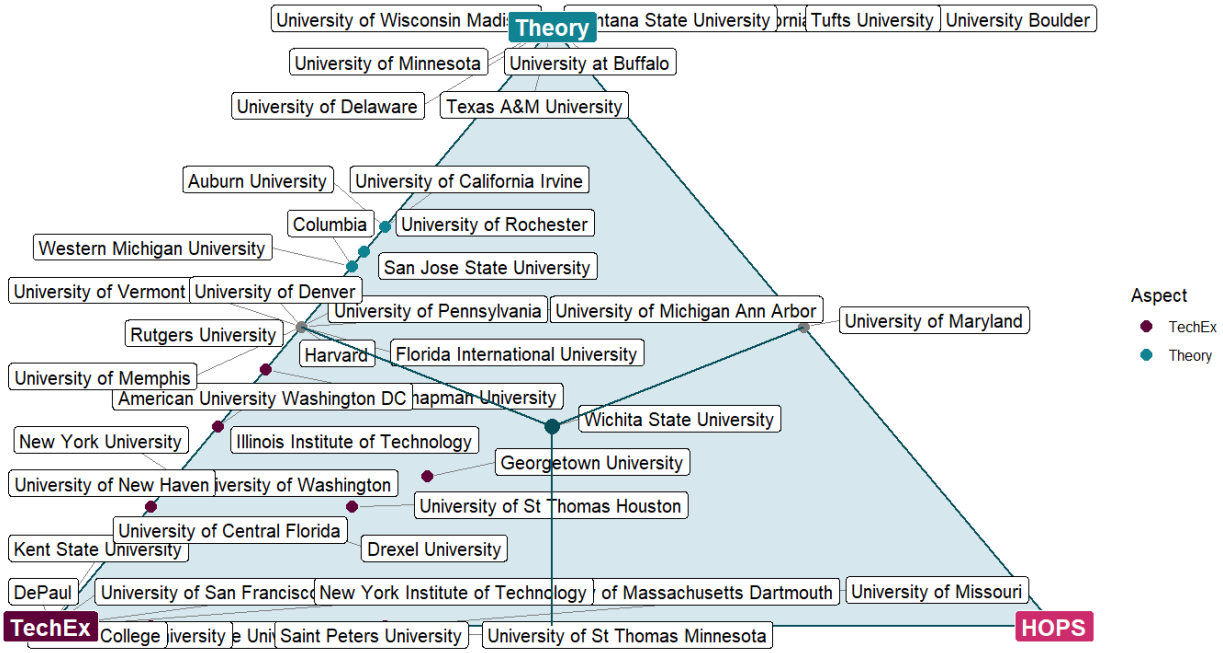

Figure 10: Simplex showing the distribution of mandated course prerequisites

Table 5: Region classification counts based on mandated course prerequisites

Theory	TechEx	HOPS	Border
16	23	0	9

The prerequisites of the mandated classes simplex is shown above. If a specific course with a course code was listed as a prerequisite it was coded per the guidelines established for the mandated courses. If the prerequisite was listed only as a topic, like linear algebra, it was coded per the guidelines established for program prerequisites. This simplex was constructed in the same fashion as the simplex for program prerequisites. The vectors were also created and normalized following the method used for the program prerequisites. Universities that do not have prerequisites for their mandated classes have been excluded from the simplex.

Course prerequisites are tending towards the technical execution aspect. However, there

is also a strong presence in the Theoretical Knowledge region. Only one university, the University of Maryland, is on the border between the Theoretical Knowledge region and the HOPS region.

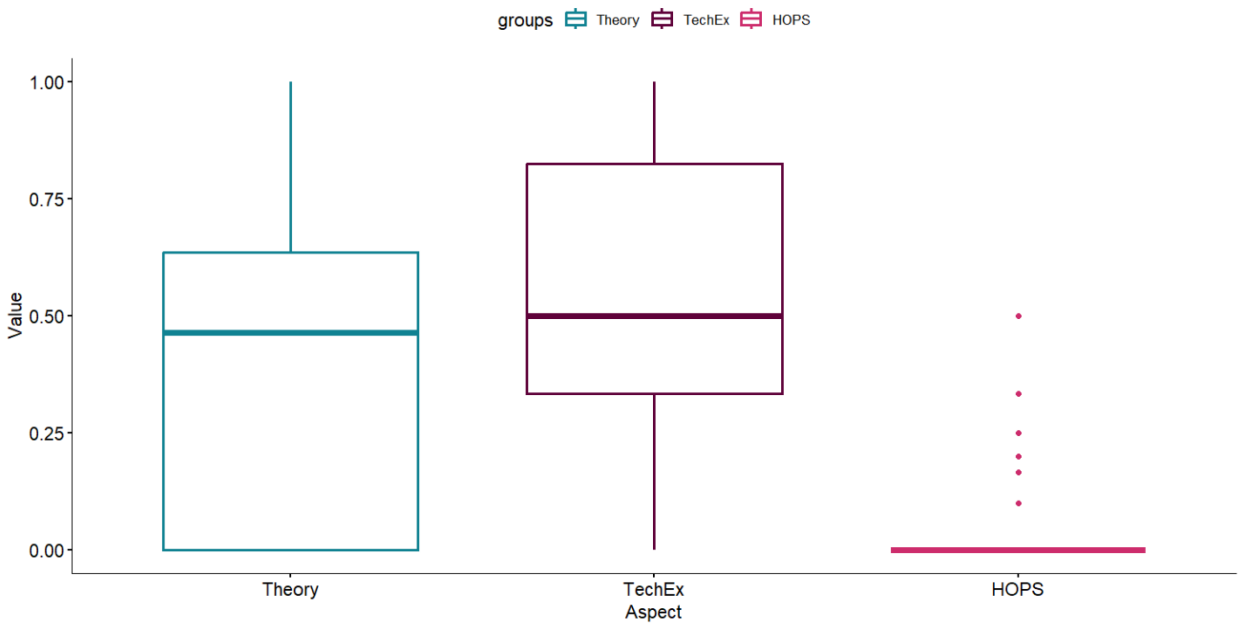

Figure 11: Boxplots of mandate course prerequisites' aspect values

Examining a boxplot of the normalized vector components also suggests that the aspect of Technical Execution is most prevalent in the program prerequisites, followed closely by Theoretical Knowledge. The average values of the Technical Execution and Theoretical Knowledge components of the vectors are 0.53 and 0.43 respectively. The average value for HOPS is 0.04, which is very low in comparison. Performing a Kruskal-Wallis Rank Sum Test reveals that there is a significant difference ($p < 0.01$) between the three aspects of data science. Performing a follow up Pairwise Comparisons using Wilcoxon Rank Sum Test reveals that HOPS and Technical Execution and HOPS and Theoretical Knowledge are significantly different at $p < 0.01$. Theoretical Knowledge and Technical Execution are not significantly different from each other.

How do mandated course prerequisites compare to program prerequisites?

In an ideal world, program prerequisites and mandated course prerequisites should align in terms of the skills they are asking students to have. Program prerequisites should be, at minimum, the aggregate of mandated course prerequisites. There should not be skills that are needed to successfully complete the mandated courses that are not addressed in the program prerequisites or taught within the program. This would not be fair to students.

In order to understand if there is coherence between program prerequisites and mandated course prerequisites, a comparison of the aspect vectors for the program prerequisites and mandated course prerequisites was done using Hotelling's Two Sample T Squared Test[4]. This comparison was made with the thirty-four universities that have both mandated course prerequisites and program prerequisites. Performing Hotelling's Two Sample T Squared Test in the same fashion as described in the description-mandated classes section revealed no significant differences in the distributions of mandated course prerequisites and program prerequisites. Program prerequisites and mandated course prerequisites value the same aspects of data science.

If a program does not teach the mandated course prerequisites, are they covered in the program prerequisites?

Thirty-four programs have both program prerequisites and mandated course prerequisites. The simplexes and statistical analysis above show that on an aggregate level the program prerequisites align with the prerequisites requested in the mandated courses. However, understanding the extent to which individual schools align in program prerequisites and mandated course prerequisites is also useful. Thirty-four programs have both program prerequisites and mandated course prerequisites. The graph below shows the counts of mandated

[4] The normality and covariance assumptions were not met, the adjusted version of the Hotelling's Two Sample T Squared Test was used to account for unequal covariance.

class prerequisites taught in the program to the counts of mandated class prerequisites not taught in the program. Twelve universities teach all of their mandated course prerequisites within their programs. The remaining twenty-two universities vary in the amount of mandated course prerequisites they teach in their programs.

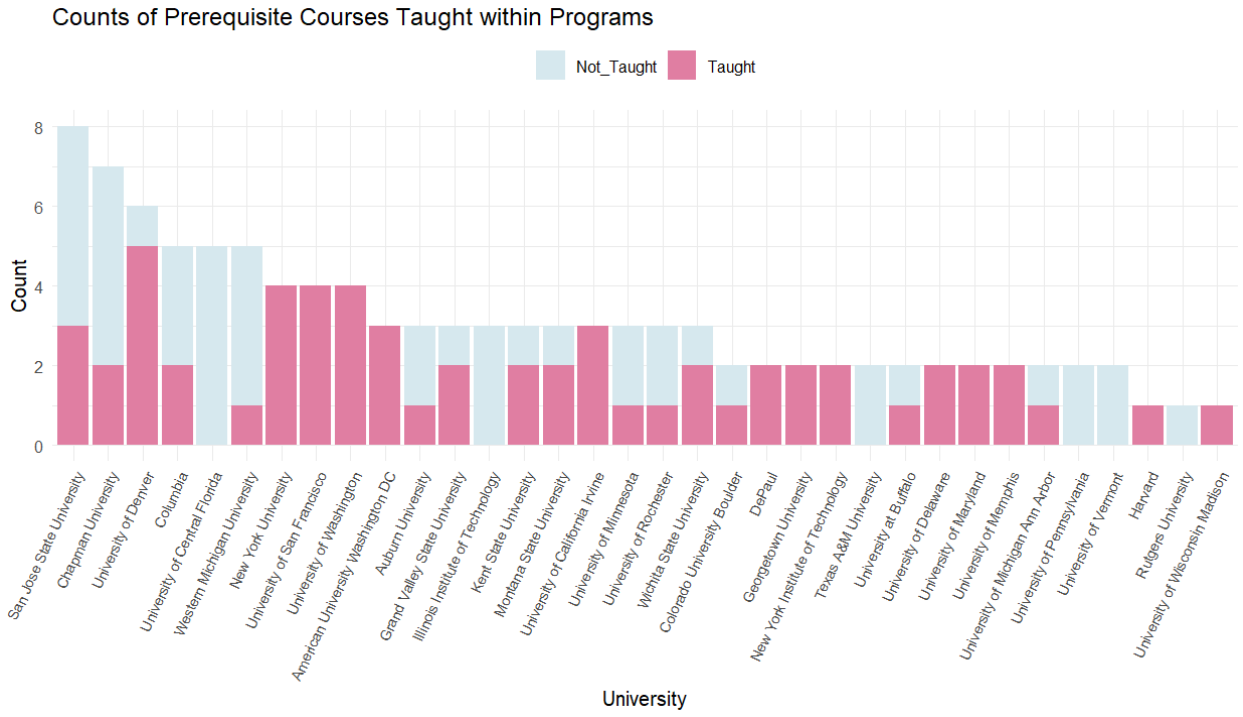

Figure 12: Counts of prerequisite courses taught within a program

Of the universities that do not teach all of their mandated course prerequisites within their program, it is important to understand if their mandated course prerequisites are accounted for in their program prerequisites. If a program does not teach all of its mandated course prerequisites it is important for these prerequisites to be accounted for within the program prerequisites to ensure that students do not arrive missing the skills necessary to complete the mandated courses.

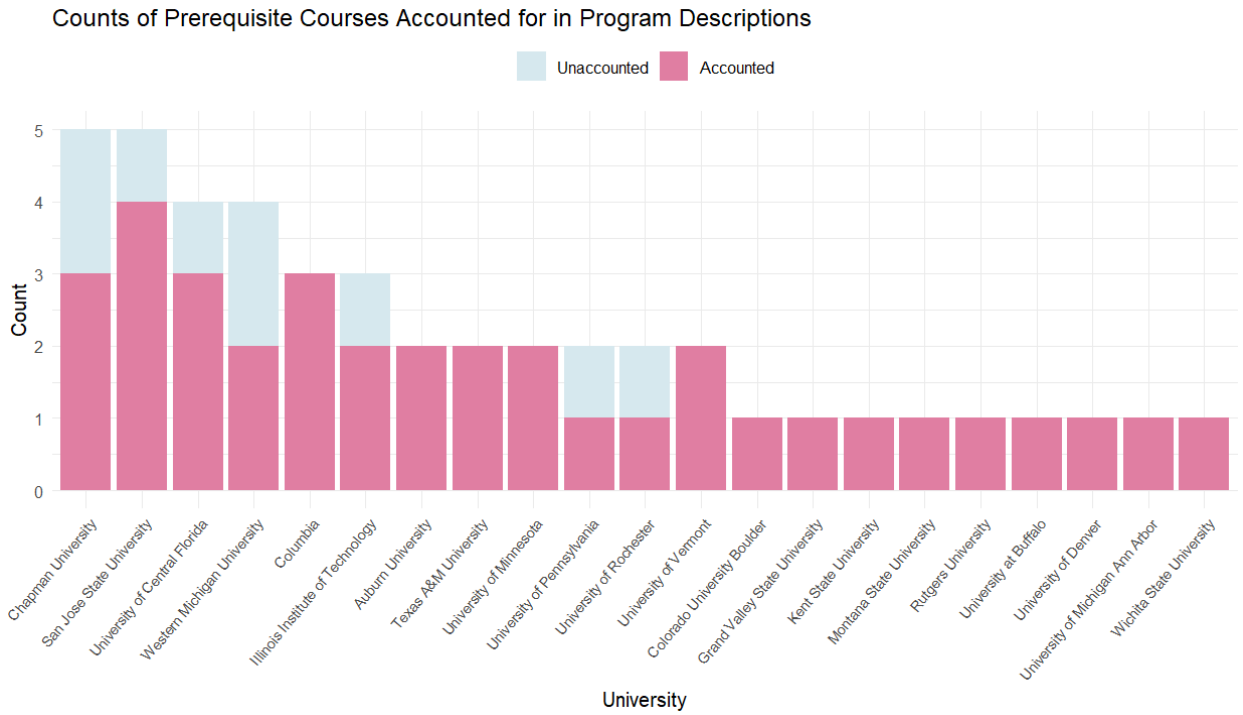

Figure 13: Counts of prerequisite courses accounted for in program prerequisites

In this set of twenty-one, the majority do account for untaught mandated course prerequisites within their program prerequisites. Five universities are only missing one prerequisite and two universities are missing two. When examining the specific mandated course prerequisites that are unaccounted for, there is no commonality between them.

Programs With No Program Prerequisites

There are nineteen programs that have no program prerequisites. These programs are important because they broaden the accessibility of data science. Programs that do not have program prerequisites are unique because they must provide all the foundational skills necessary to do data science in addition to teaching the data science skills that are expected at the graduate level. It is important to understand if the programs that do not have prerequisites teach all

required skills within their curriculums or if their mandated courses make skill assumptions that may not be met by students from a non-STEM background.

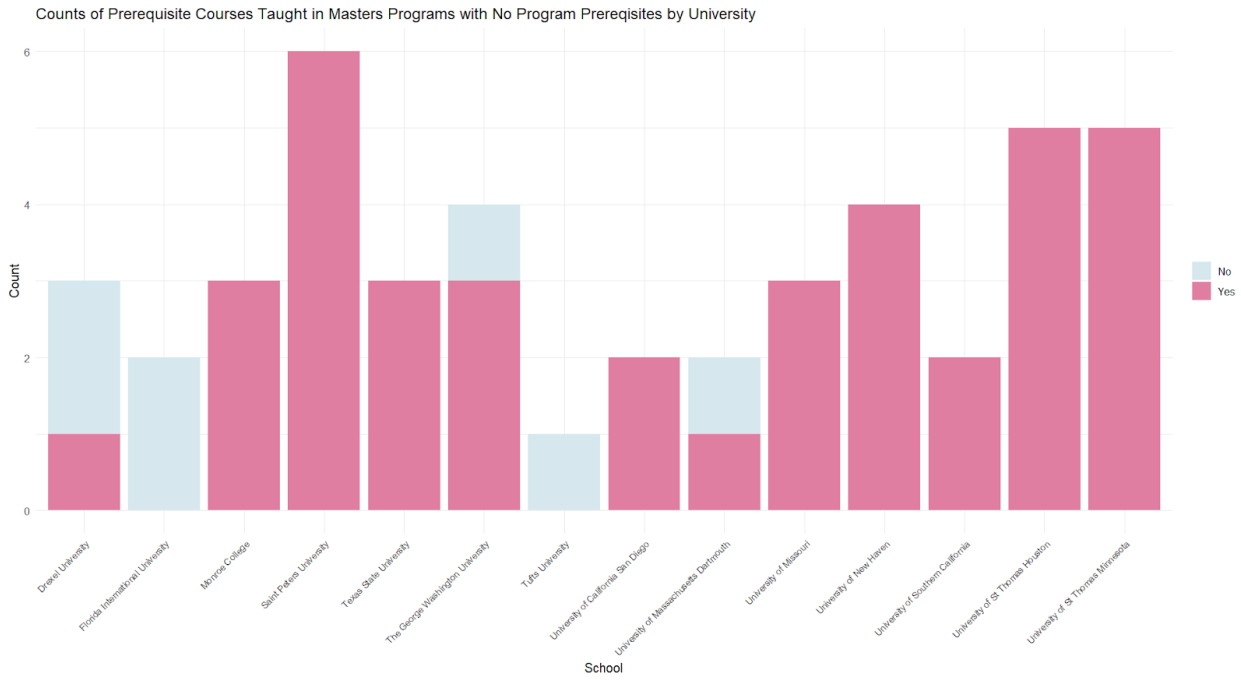

Figure 14: Counts of course prerequisites from programs without program prerequisites

The graph above shows the counts of mandated course prerequisites offered and not offered in programs that do not have program requirements. Five universities of the original nineteen universities were omitted because they did not have prerequisites for their mandated classes. Eight of the programs taught all mandated course prerequisites within their programs. Three of the programs taught some, but not all mandated course prerequisites within their programs. Two programs taught none of the mandated course prerequisites. On the whole, programs without program prerequisites are teaching their necessary prerequisites for their mandated courses.

How is the emerging concern of human oriented professional skills being accounted for in master's programs?

The aspect of HOPS is unique in that it has two sub aspects. The sub aspect of Responsible Data covers professional skills like data ethics, experimental design, and data privacy and security. These skills are human oriented because they involve the responsible and fair use of data that is either collected from humans or the design and implementation of a data science project that will affect humans. The other sub aspect of Soft Skills covers human oriented skills that are required to collaborate and communicate with others in the workplace.

Examining the original program description simplex, only eight universities fall into the HOPS region. However, when comparing the two sub aspects, it becomes clear that the Soft Skills aspect is being mentioned in program descriptions more frequently. Twenty-one programs mention at least one of the three indicators corresponding to the sub aspect of Responsible Data. Whereas forty-one programs mention at least one of the four indicators corresponding to the sub aspect of Soft Skills. When the Responsible Data Science sub aspect is removed there is an increase in program descriptions that fall into the HOPS region. In addition, eleven programs shift into border classifications located at the boundary line between Technical Execution and HOPS. While at first this may seem surprising, this shift into the HOPS region can be explained by the indicator normalization. The original HOPS aspect was divided by seven as there are seven total HOPS indicators, three for Responsible Data and four for Soft Skills. When Responsible Data is removed there are less indicators to divide by for indicator normalization, four, instead of seven. If there were few or no values in the Responsible Data Science aspect, normalizing by seven would make the total HOPS value relatively small. When Responsible Data is removed, the HOPS aspect is only divided by four, because it is only composed of the

Soft Skills aspect. If there are more values in the Soft Skills aspect, then the value for HOPS increases, pulling more universities into the HOPS region. Performing a Hotelling's Two Sample T Squared Test[5] found that there was not a significant difference ($p<0.01$) in the HOPS aspects pre and post Responsible Data removal.

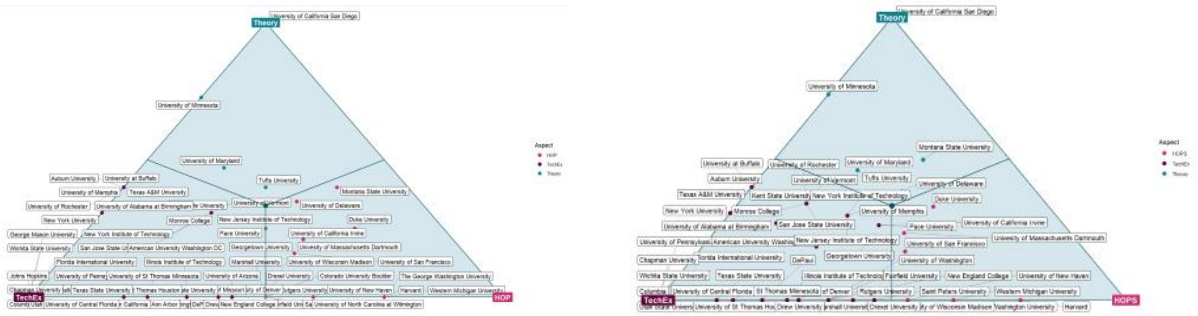

Figure 15: Simplex comparison of program distribution when HOPS: RD is removed

Table 6: Region classification comparison of program description without HOPS: RD

Program Description	Theory	TechEx	HOPS	Border
RD Removed	4	36	11	11
Original	6	48	8	0

When the sub aspect of Soft Skills is removed there is very subtle movement in region classification for program description. Three program descriptions changed region classification from HOPS into the Technical Execution or Theoretical Knowledge region.

[5] The normality assumption was not met, but the covariance assumption was, so the test was preceded with as Hotelling's Two Sample T Squared is fairly robust to normality violations

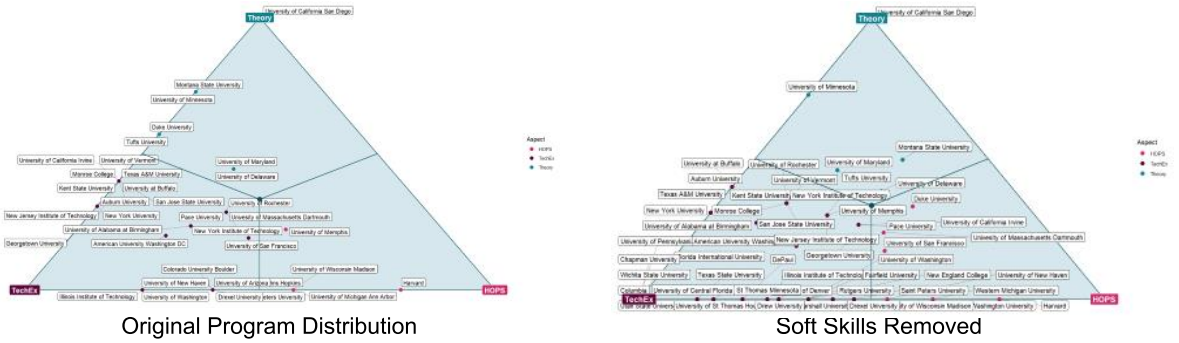

Figure 16: Simplex comparison of program distribution when HOPS: SS is removed

Table 7: Region classification comparison of program description when soft skills is removed

Program Description	Theory	TechEx	HOPS	Border
SS Removed	7	50	5	0
Original	6	48	8	0

Responsible Data is not capable of pulling programs into the HOPS region when Soft Skills are removed. Indicating that most universities fall into the HOPS region because of the presence of Soft Skills. Performing Hotelling's Two Sample T Squared Test found that there was significant difference ($p<0.01$) in the HOPS aspect pre and post Soft Skills removal.

In mandated courses, when Responsible Data is removed from the HOPS aspect, there is a significant change in distribution location according to Hotelling's Two Sample T Squared Test[6]. While all three of the Hotelling's Two Sample T Squared Test were significant ($p< 0.01$), the only difference was between the HOPS: RD removed and the original distribution. Programs shift out of the HOPS region and into Technical Execution or become border cases in the boundary between HOPS and Technical Execution.

[6] The assumption for equal covariance was unmet for mandated courses, so the adjusted Hotelling T Squared test was used.

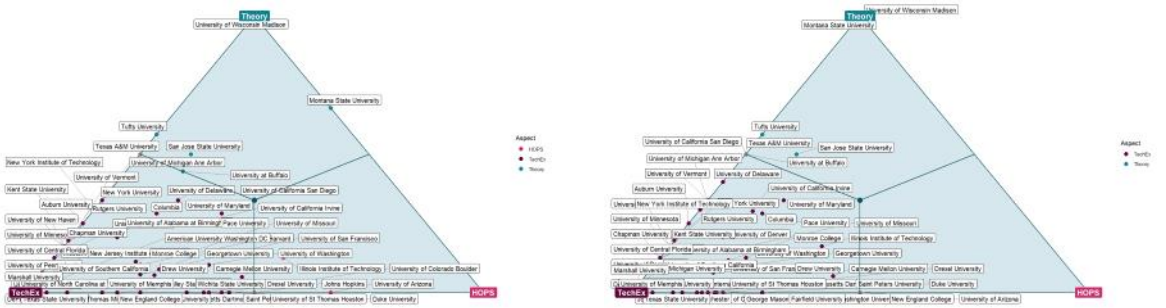

Figure 17: Simplex comparison of program distribution when HOPS: RD is removed

Table 8: Comparison for Mandated Courses

Mandated Courses	Theory	TechEx	HOPS	Border
RD Removed	4	53	0	4
Original	6	51	2	2

When Soft Skills are removed for the HOPS aspect the distribution does not change. Three Hotelling's Two Sample T Squared Tests were performed and none returned significant differences. The simplexes below only show very mild shifts in location when Soft Skills are removed, with universities formerly in the HOPS region of the simplex only moving to the border of HOPS and Technical Execution or into Technical Execution.

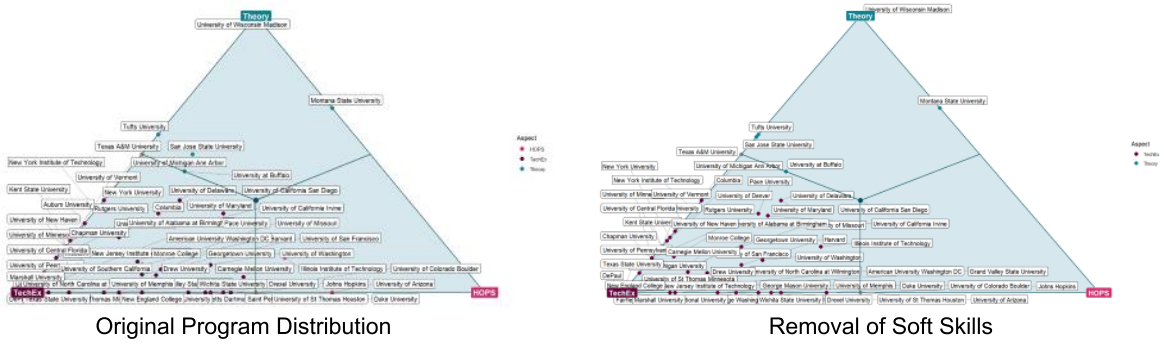

Figure 18: Simplex comparison of program distribution when HOPS: SS is removed

Table 9: Region Comparison for Mandated Courses

Mandated Courses	Theory	TechEx	HOPS	Border
SS Removed	6	52	0	3
Original Distribution	6	51	2	2

As opposed to the program descriptions where Soft Skills is the main contributor to the HOPS aspect, the main contributor to the HOPS aspect is Responsible Data. There are not courses that focus solely on soft skills like there are courses that focus solely on responsible data science practices, a few universities teach fully mandated ethics courses, but there are typically not courses fully dedicated to only building soft skills. It is possible that soft skills are expected to be built in supplement with coursework, i.e. there are group projects in the class itself that are not listed in the course description.

Ethics

It has been established that HOPS: RD is not discussed widely or thoroughly in program descriptions, but does seem to be mildly important for the locations of the mandated courses. Of the indicators used in the program description coding to identify HOPS: RD, the indicator most discussed was Data Ethics. Data Ethics was mentioned by seventeen different programs in their descriptions. The other two indicators, Experimental Design and Data Privacy/Security/Legal were only mentioned by five and six programs respectively. As data ethics is the most prevalent representative of HOPS: RD, it is important to understand how programs are delivering ethics education to students.

Ethics instruction tends to be delivered in three different ways. There can be a full ethics course that is mandatory. There can be a full ethics course that is optional. Some form of ethics instruction can also be given inside a mandated course. This is most typically seen in capstone course situations. The capstone course is a common graduation requirement and there can be a module within that course that covers data ethics and/or ethical thinking. In very rare cases, there will be an elective that covers ethics, but ethics is not the sole focus of the class, or ethics instruction will be outsourced to an online program, not hosted by the university.

The different methods of providing data ethics instruction communicate the extent to which a program values ethics instruction. A fully mandated course communicated that a program values ethics highly as they've made the ethics course a requirement to pass into the data science community. Nonmandated full courses communicate that a school values ethics, but they do not deem it a requirement to be a practicing data scientist or a member or the data science community. When ethics is contained in a different mandated course, it communicates the school is acknowledging that ethical thinking is important for students to have, but not important enough to warrant a full course. Ethics as a subtopic in a nonmandated course is the weakest offering of ethics a program can provide.

Table 10: Modality of Ethics Instruction

No Ethics	Subtopic Contained within a non-Mandated Course	Subtopic contained within Mandated Course	Ethics is a non-Mandated full course	Ethics is a mandated full course
38 (61.29%)	2 (3.23%)	5 (8.06%)	5 (8.06%)	12 (19.35%)

If ethics is offered, then a full mandated course is the most popular option. The next two common offerings are ethics instruction as a non-mandated full course and ethics instruction within a mandated course, but not the subject of the course. In situations where ethics instruction is provided as a subtopic in a different course, it is unclear how much of the course is focused on ethics instruction and what in particular is covered in the topic of ethics of data science. Overall, there is a very low priority placed on providing comprehensive ethics instruction, with only twelve programs deeming ethics training necessary to become practicing members of the data science community.

How data science programs are approaching ethics instruction seems to be different from the established ethics frameworks that have come from the Computer Science discipline. In

Computer Science, the framework ImpactCS favors a broad integration of ethics across the CS curriculum. Students are exposed early to ethics in CS concepts and those concepts should be revisited in individual courses. Other frameworks also generally advocate for the repeated discussion of ethics throughout the curriculum, but there are a few that suggest just one intensive course. (Carter & Crockett, 2019). Data science curriculums appear to be offering only one individualized ethics course, if they offer one at all. Examining mandated course descriptions, there does not appear to be frequent discussion of ethics in classes that are not specifically ethics related. It is unlikely, however, possible that ethics discussion is appearing in other courses. It is impossible to know the extent to which ethics is discussed in courses that offer ethics as a module of a course.

DISCUSSION

As outlined in the literature review there ought to be three aspects of data science - Theoretical Knowledge, Technical Execution, and HOPS. These normative considerations were used to develop coding schemes used to examine master's of data science programs across the United States. Each of these three aspects are present, but emerge in different areas of program descriptions and are unequally valued within the programs. Understanding where these values emerge and the consequences of the locations of emergence is important for understanding how universities are defining data science.

Beginning with program descriptions, the goal of a program description is twofold. First, it attempts to provide an overview of the program's content, goals, and takeaways. Second, it attempts to attract prospective students. The way programs are generally describing themselves leans heavily into the Technical Execution aspect of data science. There are a few programs that discuss the HOPS aspect, however these are the minority. Even fewer programs discuss the Theoretical Knowledge aspect, to the point where the median value of the Theoretical

Knowledge aspect is zero. This outwardly communicates to prospective students that the discipline of data science is primarily focused on Technical Execution, but the sub aspect of HOPS: SS which allows for the communication between disciplines is also valued, but to a lesser extent. This aligns well with the textbook definitions of data science that were provided in the literature review. They detail that data science is about technical problem solving for complex data problems, with the data coming from other disciplines. Priority in these definitions is placed on what a data scientist can and should do with data with a brief mention on communicating results or interfacing with a domain knowledge specialist. A consequence of this messaging is that it could be attractive to a particular set of students that come from other disciplines that also learn towards being primarily based on technical execution skills and place a lower priority on Theoretical Knowledge or HOPS.

However, there is a shift in which aspects are prioritized when examining program prerequisites. Program prerequisites outline the minimum knowledge a prospective student needs to have in order to be accepted to the master's program. In program prerequisites, Theoretical Knowledge skills are most frequently requested. Programs are requesting students arrive to the program already familiar with linear algebra, calculus, probability and statistics, etc. Technical execution skills are also requested, but in a lesser capacity, and typically the request is that students arrive with some computer science skills, like coding. While the program prerequisites seem to be at odds with the program descriptions, this mismatch can be explained. In the textbook, *The Foundations of Data Science*, data science is defined primarily by the mathematical skills needed to practice data science, not by the types of problems data science addresses. Theoretical Knowledge skills may function as a gate into data science, hence their use in program prerequisites. If programs are going to primarily focus on Technical Execution,

then Theoretical Knowledge is required first. If a student does not understand why and how a method works, they cannot apply it correctly. The aspect of Theoretical Knowledge underpins the aspect of Technical Execution.

The use of Theoretical Knowledge as a gateway into master's of data science programs is understandable. Certain skills are needed in order to progress to the skills required at the graduate level. In addition, master's programs are shorter than an undergraduate degree program. Undergraduate degrees are typically four years and master's programs are typically only one or two years. Universities may not be able to teach all the theoretical foundational skills and then have enough time to build upon those skills while adding new technical or human oriented skills necessary for those at the graduate level, if the program length is only two years. In the data set analyzed there are two approaches to program prerequisites. The first, already discussed, Theoretical Knowledge is used as program prerequisites. Second, the program has no prerequisites. Both models have benefits and drawbacks.

The benefit of using Theoretical Knowledge as a prerequisite is that it allows for more time to be dedicated to refining Technical Execution Skills and building HOPS. Students can also advance into more complex methods and study more complex problems if they arrive with a strong foundation in Theoretical Knowledge. If students arrive with foundational skills already, time doesn't need to be taken out of the program to teach Theoretical Knowledge aspect skills. However, there may be some unintended consequences of using Theoretical Knowledge as a gateway into a master's of data science program. Primarily, it limits the types of students that can access the field of data science. The Theoretical Knowledge skills required for entrance into a data science program are not necessarily covered broadly by general education requirements. This restricts the pool of applicants down to students who are able to meet theoretical knowledge

requirements. These students are likely to come from STEM oriented backgrounds. It would be unlikely, but not impossible, for students with a more humanities oriented undergraduate degree to fulfill the Theoretical Knowledge based entrance requirements. The use of Theoretical Knowledge as entrance requirements favors students coming from STEM backgrounds like mathematics, computer science, and/or engineering. This can inadvertently reduce the thought diversity within the data science program, because there is less variety in educational training and backgrounds.

In itself, this may seem fine, other disciplines, like biology or chemistry, restrict who can access graduate level training in that discipline by setting out program prerequisites that only a few groups of students can meet. Data science is not like other disciplines. The reason data science began to develop as a discipline was in response to increasingly complex data problems in numerous other disciplines that statistics alone could not solve. These problems began to become solvable when statistics, computer science, and domain knowledge were all leveraged together. The goal of data science is to interface with other disciplines to develop or adopt new methodologies to add to the data science tool kit or solve complex data problems in other disciplines. Using Theoretical Knowledge as a program prerequisite may betray the foundations of the field of data science.

The second approach to program prerequisites is that there are no program prerequisites. Any student from any disciplinary field may apply. These programs tend to build Theoretical Knowledge skills into their programs. For example, linear algebra is a mandated course in a program and serves as a prerequisite to another mandated course. The issue with this approach is that due to time constraints students may not get as much exposure to Technical Execution or HOPS because more time is spent on Theoretical Knowledge education. Another possibility is

that these students receive much less Theoretical Knowledge education and the program focuses largely on just Technical Execution, running the risk of students not fully understanding what they are doing, why a method works how it does, or the boundaries of a certain method's application. This approach, however, does provide paths into the discipline of data science for students from non-STEM backgrounds, encouraging thought diversity within the field.

In order to fulfill the maintain the multidisciplinary nature of data science, a data science student body should either come from numerous different undergraduate backgrounds which brings in the training, practices, and teachings from other fields to the data science discipline or data science programs should continue to use Theoretical Knowledge a program prerequisites, but focus on building and refining the human oriented professional skills required to interface with and appreciate the knowledge and skills that comes from other disciplines.

In regard to the first option, if a data science program chooses to forego Theoretical Knowledge based program prerequisites in favor of no program prerequisites or program prerequisites that are HOPS: SS based like Duke University's, then they need to focus on building Theoretical Knowledge and Technical Execution skills. If students are more varied in their educational backgrounds and trainings, then they increase the thought diversity within the program and HOPS may not need that much attention because those skills were taught in their undergraduate degree. However, attention needs to be paid to the mandated course prerequisites if there are no program prerequisites. If there are no program prerequisites, then mandated courses need to be accessible to students with multiple levels of Technical Execution and Theoretical Knowledge. Currently, in data science programs, mandated course prerequisites tend to be focused on Technical Execution, with Theoretical Knowledge close behind. In data science programs that do not have program prerequisites, foundational courses that build skills in

programming, linear algebra, probability and statistics, etc. are taught in the program either as mandated courses or as electives. This is one possible model. Another approach could be to shift courses into being split between Theoretical Knowledge and Technical Execution. This approach is similar to the idea Tukey (1962) raised when outlining the discipline of data analytics. He argued that math should be learned on an as needed basis. Theoretical skills absolutely necessary for Technical Execution can be taught in the same course. This approach may allow students with non-STEM backgrounds to build their Theoretical Knowledge and Technical Execution skills more quickly and avoid teaching Theoretical Knowledge skills that are not directly relevant.

If the second option is selected by a program, then there would need to be a large shift in the content of the mandated classes and the faculty present in data science programs. As it stands currently, the mandated courses inside of data science programs are primarily geared towards Technical Execution. They do not focus on building HOPS in a way that is congruent with the way that program descriptions describe the values of their programs. The sub aspect of HOPS that is represented in any capacity in mandated courses is Responsible Data, particularly data ethics. This representation is incredibly small. Program descriptions describe a myriad of soft skills they want to build in their students, yet few programs follow through on building those skills in their mandated courses. This finding aligns with what Oliver and McNiel (2021) documented at the undergraduate level and what Tang and Sae-Lim (2016) documented at the graduate level.

On its own, building Technical Execution skills is not an issue. The issue arises when Technical Execution skills are only skills that are built, and programs do not provide support for teaching HOPS. Findings from Norén, 2019, show that the faculty that primarily compose data

science departments are from CSE, math, and/or engineering backgrounds. Percentagewise, there are very few faculty from social sciences, business, or humanities backgrounds that can teach HOPS. Restricting students to particular educational backgrounds that do not normally include HOPS and then focusing primarily on the Technical Execution aspect of data science betrays the spirit of data science and isolates the discipline.

Overall, regardless of the structure of the program, HOPS: RD should be built into each program. Responsible Data skills are somewhat unique to the discipline of data science, so it is unlikely that any student is going to arrive to a masters of data science program already having those skills. Currently, there is a very low priority placed on HOPS: RD. Data science is a unique field in that it touches many other disciplines and with the saturation of technology in society, the choices a data scientist makes can have a very wide impact. Responsible Data science was not discussed when Tukey (1962) and Cleveland (2001) discussed the discipline of data science. Discussions about Responsible Data practices and how it should be taught in a data science curriculum occurred later in the establishment of data science curriculums.

Master's of data science programs in the United States are tending towards favoring the Technical Execution aspect of data science. The program descriptions and the mandated courses predominately feature Technical Execution. While Theoretical Knowledge is featured in the program prerequisites, there is not generally a continuation of theoretical education within the program. HOPS instruction is largely neglected and while HOPS: SS are discussed in program descriptions; they are not implemented in a meaningful way in the mandated courses. HOPS: RD is given more consideration, with a few universities requiring instruction in data ethics, but this consideration is not widespread enough to say that data science programs across the United States are valuing this aspect. There has been a growing awareness towards HOPS: RD in the

past years. It is possible more programs will add courses covering HOPS: RD topics. Data science emerged as a discipline with three aspects, Theoretical Knowledge, Technical Execution, and HOPS, through examining data science programs in the United States, it appears to only be two aspects, Theoretical Knowledge and Technical Execution, with a slowly developing third aspect of Human Oriented Professional Skills. If the third aspect is not developed, programs run the risk of isolating the discipline of data science, turning it into something that betrays its founding spirit.

LIMITATIONS

There are a few limitations worth discussing regarding this study. First in aspect measurement, the aspects selected were those developed through concepts found in the literature related to the development of data science as a discipline. There may be other aspects of data science that were not captured. Tang and Lim (2016) included the availability of courses that built domain knowledge in their analysis of master's of data science programs. Domain knowledge could perhaps be an additional aspect of data science and is an extension of the idea that data science should be multidisciplinary. Data science can be multidisciplinary in that data scientists can work with other people who have domain expertise or data scientists themselves can have training in another discipline, like biology, that allows them to work in that field. This work focuses more on if programs are the building the skills necessary for data scientists to work in multidisciplinary environments within their mandated courses. An area of extension would be to examine the elective offerings and the program policies regarding taking courses in other departments to understand the extent to which a program values building domain knowledge within their data science students. The use of only mandated courses in this work may limit the aspects of data science that can be studied. There may be other aspects of data science being

conveyed in nonmandated courses. In the programs examined for this work, there were programs that offered specialization tracks in other disciplines suggesting that domain specialization could potentially standalone as its own aspect. The inclusion of electives, the structure of specialization tracks, the course content, and their frequency of occurrence in data science programs would need to be examined further to understand if this is a full aspect of data science. It is also possible that these electives and specialization tracks are merely teaching Technical Execution in a specific domain. A data science student take biology classes to understand the broader context of the discipline they are working in is different from teaching a data science student the common tools of analysis used in a certain discipline or having them apply methods they already know to data from a different discipline.

In addition, it may be useful to examine the data analytics programs that were not included in this study. Data analytics programs share certain similarities with data science programs in regard to curriculum, but tend to have different requirements and different learning outcomes. In working to define what data science is, it may be useful to understand where the difference lies between data science and data analytics. Through examining differences between the two disciplines, a more rigorous definition of each field may be provided. There may be aspects of data science that are shared between the two, or there may be aspects unique to each field. Through comparison of the two disciplines, it may be possible to identify a new aspect of data science that is not obvious when only examining data science programs.

Regarding data collection, this study used only publicly available data. This is all content published by the university or a representative of the university. It is possible that a university has unknowingly misrepresented itself in its program description. There may have been a recent curriculum change that was not reflected in the program description or the program description

was created before all the courses in the program were set. Building upon this limitation for future work, engaging the university in a conversation related to how they see their program compared to how they are presenting their program may be useful. A university may be valuing certain aspects of data science internally, but not communicating those values in their program descriptions.

Using only publicly available data is also a limitation particularly when it comes to judging the content of the mandated classes. Some universities were very detailed in their course descriptions, others were not. If available, course syllabi were used to inform content and course goal determinations. However, there is no guarantee if the course description or syllabi align with what is actually taught in the classroom. In addition, depending on how the course description was written, a course may seem to lean more towards Theoretical Knowledge in description, but be taught in a way that is more Technical Execution oriented. In an ideal situation, it would be possible to attend every mandated course and determine the extent to which the course description aligned with the content taught in class. Notes could also be taken regarding how the content was delivered and which aspects of the content were prioritized. A less extreme step than attending every mandated course in the study could be to analyze the teaching materials, homework assignments, exams, and project guidelines used in the mandated courses of the study to understand what is being taught and how it is being taught.

Further, this study did not capture anything regarding the social environment of the program. There may be student or faculty lead activities on campus that develop an aspect of data science not developed in the mandated courses. A program may be primarily theory based in its mandated courses, but there could be student lead activities that practice Technical

Execution skills. There could also be opportunities through a career center on campus to build soft skills that are not taught in class.

In addition, it may be useful to capture beliefs from students and faculty on what they believe data science is. What students believe data science is may affect the courses they select to take or what additional trainings they seek out. A faculty member's belief in what data science is can guide what they include in their course content and the avenues they encourage students to explore.

It might prove helpful to collect data regarding faculty background and research interests, student background and research interests, as well data on thesis and/or capstone projects completed by the students. A program's values are not only communicated through the courses they offer and how they describe themselves, but the people they choose to bring into the program. Understanding the backgrounds of faculty is important, because faculty background controls what can and cannot be taught. If there is no faculty who has studied HOPS, then courses that convey HOPS cannot be offered at the same frequency as other more theoretical knowledge or technical execution courses can be offered. If there is an abundance of faculty who study the theoretical aspects of data science, then courses are likely to have more of a theoretical lean.

Student background is also important to understand. The types of students that are let into a data science program are the people who senior practitioners of data science deem as worthy to become future practitioners of data science. This work only examines the program prerequisites, as these requirements are publicly available. It cannot say the extent to which the program prerequisites are being followed. There may be other elements involved in the

admission process that allow students from educational backgrounds that differ from the program prerequisites to qualify for the program.

CONCLUSION

As data science is a young discipline and rapidly developing, educational institutions have a great power to shape the field. The aspects of data science that master's programs prioritize determine the skills and beliefs that are passed on to students. The discipline has roots in statistics and was developed further by advances in computer science. As the discipline evolved, three aspects of data science emerged -Theoretical Knowledge, Technical Execution, and HOPS.

These three aspects are not conveyed equally in master's of data science programs across the United States. Master's programs strongly prioritize technical execution, keeping with one of the founding ideas of data science- data science should prioritize application. Theoretical Knowledge is only prioritized in the context of prerequisites, appearing frequently in program and mandated course prerequisites. HOPS is largely unaddressed, with soft skills being referenced in program descriptions and responsible data skills being very infrequently taught in mandated courses.

Over prioritizing one aspect can create certain risks for the discipline. All aspects play a role in the process of doing data science. One of the founding ideas of data science is that the discipline was meant to be flexible and collaborate with other disciples to solve their complex data problems, using a variety of tools and methods. Solving complex data problems is not only about Technical Execution, it is about understanding how the tools and methods work, knowing when it is appropriate to apply certain tools and methods, making responsible choices with the data, and presenting findings and outcomes to other people. All aspects play a role in the process

of doing data science. Reducing educational instruction to only prioritizing only one aspect of data science can isolate the discipline, removing the collaborative element that makes the discipline unique. Technical Execution instruction is important, but it is not the only aspect of data science that matters. HOPS instruction is particularly import in ensuring that data science programs honor the founding spirit of the discipline.

REFERENCES

Blum, A., Hopcroft, J. E., & Kannan, R. (2018). *Foundations of Data Science*. Cambridge University Press.

Carter, L., & Crockett, C. (2019). An ethics curriculum for CS with flexibility and continuity. 2019 IEEE Frontiers in Education Conference (FIE). https://doi.org/10.1109/fie43999.2019.9028356

Chavan, P., Mahalle, P. N., Mangrulkar, R., & Williams, I. (2023). *Data science: Techniques and intelligent applications*. CRC Press.

Cleveland, W.S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. Statistical Analysis and Data Mining: The ASA Data Science Journal, 7.

Kelleher, J. D., & Tierney, B. (2018). *Data science*. The MIT Press.

Lin, H., & Li, M. (2023). *Practitioner's Guide to Data Science*. CRC Press.

National Academies of Sciences, Engineering, and Medicine. 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.
National Academies of Sciences, Engineering, and Medicine. 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.

Noren, L., Helfrich, G., & Yeo, S. (2019, October 30). *Who's building your ai? research brief*. Obsidian Security. https://www.obsidiansecurity.com/blog/whos-building-your-ai-research-brief/

Oliver JC, McNeil T. 2021. Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training and domain-specific context. PeerJ Computer Science 7:e441 https://doi.org/10.7717/peerj-cs.441

Tang, Rong & Lim, Watinee. (2016). Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. Education for Information. 32. 269-290. 10.3233/EFI-160977.

Tukey, J.W. (1962). The Future of Data Analysis. Annals of Mathematical Statistics, 33, 1-67.

APPENDIX 1

List of Schools in Sample

University of California San Diego
Georgetown
University of Washington
University of Michigan Ann Arbor
Harvard
Johns Hopkins
DePaul
University of California Irvine
University of San Francisco
University of Minnesota
Colorado University Boulder
Columbia
University of Pennsylvania
Duke University
University of Southern California
Carnegie Mellon University
Tufts University
New York University
University of Rochester
Rutgers University
The George Washington University
University of Massachusetts Dartmouth
University of Delaware
University at Buffalo
The University of Vermont
Illinois Institute of Technology
University of Denver
University of Missouri
George Mason University
University of Maryland
Florida International University
Kent State University
Western Michigan University
Utah State University
Fairfield University
Chapman University
University of St Thomas Minnesota
Grand Valley State University
San Jose State University
Saint Peters University
University of St Thomas Houston
University of North Carolina at Wilmington
University of Arizona
University of Alabama at Birmingham

Marshall University
Drew University
Auburn University
American University Washington DC
University of Central Florida
New Jersey Institute of Technology
Pace University
Wichita State University
University of Memphis
Texas State University
Montana State University
Texas A&M University
University of New Haven
New York Institute of Technology
University of Wisconsin Madison
Monroe College
New England College
Drexel University

APPENDIX II

Codebooks

Program Descriptions

Theory	
Explicit Theory Statement	The program says we prioritize/balance/teach theory/theoretical concepts, word theory or theoretical if the program description mentions a topic (theoretical or theory of …), the program may directly say we teach theory
Prepare for Doctoral Program	The program says something about pursuing a PhD or further studies in data science
Foundations	The program says they provide strong foundations/teach foundations
Technical Execution	
Evening Classes/Online or In Person Hybrid	The program says it offers some/all classes in the evenings/late afternoon/weekend or allows students to attend in person or online, hybrid set up, programs that make you choose all in person or all online do not qualify for this box
Application Oriented	The program will use words like hands on, applied, using skills, working on real world problems, providing real world/job/hands on experience
Industry Informed Curriculum	The program says they are preparing students for industry/careers or that the program and/or the program says if was developed with industry in mind and/or by industry professionals, the program may discuss becoming a data scientist or that the program will prepare you for working in different sectors
Practicum/Internship	The program highlights that students will complete a practicum or internship, a capstone or thesis does not qualify for this box, case studies also do not qualify for this box
Responsible Data	
Data Ethics	The program says they teach data ethics, want students to think about ethics, or mention ethics in any capacity related to data
Experimental Design	The program says it teaches students to evaluate experimental designs or select correct experimental designs based on context
Data Privacy/Security/Legal	The program says it teaches about data privacy/security/legal aspects of being a data

	scientist
Soft Skills	
Teamwork/Collaboration	The program says teamwork, says students will collaborate in teams on projects
Communication Skills	The program says it teaches data scientists how to present or communicate their findings/projects/etc
Professional Skills	Professional Skills The program mention of networking, practicing professional skills, practicing interviews, building workplace like skills, discusses a career services like feature
Data Visualization	The program says it teaches data visualization

Mandated Courses

Theory	Course covers probability and statistics, linear algebra or another foundational mathematical skill, program lacks any technical execution aspects and covers nine topics or less, or explicitly states it covers foundations and theory
Technical Execution	Courses covering applied statistics, databases, and/or programming, any course discussing applying concepts to real world data sets, a lab component, data engineering, and/or having students perform any task similar to what they would do in industry, a programing prerequisite was also examined in the context of the course description, with programming suggesting the course was focused on technical execution
HOPS: Responsible Data	Any course covering critical thinking/evaluation of experimental design, data ethics, and/or legal/security/privacy aspects of data science
HOPS: Soft Skills	Any courses covering collaboration, teamwork, data visualization, or any form of other professional skills

Program Prerequisites

Theory	Topics based on mathematics, like Linear Algebra, Probability and Statistics, Calculus, Discrete Math, anything that is proof related
Technical Execution	Computer science related topics, programing,

	algorithms, databases, linear regression, or applied statistics, any prerequisite that appears to be based on applying knowledge to problems
HOPS: Responsible Data	Ethics and anything related to data privacy/security/legal knowledge
HOPS: Soft Skills	Teamwork, presentation skills, any form of professional skills beyond what is expected at the graduate level (i.e. essay writing would not be considered a soft skill)