

Dreaming of Big Data

A method for evaluating the use of data
in the undergraduate mathematics curriculum.

Interactive Qualifying Project Report completed
in partial fulfillment of the Bachelor of Science degree
at Worcester Polytechnic Institute, Worcester, MA

Submitted to:

Professor Randy C. Paffenroth, PhD *advisor*

Professor Gu Wang, PhD *co-advisor*

Student Signature: _____
MaryAnn E. VanValkenburg

Advisor Signature: _____
Randy C. Paffenroth, PhD

Co-Advisor Signature: _____
Gu Wang, PhD

Abstract

This report evaluates the current undergraduate curriculum at WPI in the face of advances in data storage and processing capability. Large quantities of data are now widely accessible in nearly every industry, and we predict that all students will require data analysis skills to succeed in their future careers. In this report, we investigate the educational needs of all students concerning data. We consider the relationship between data and mathematics to determine if the Mathematical Sciences department should be responsible for providing students with a data education.

Executive Summary

Recent advancements in data storage and processing technology have increased the amounts and diversity of types of data captured in seemingly every industry. In several cases, learning from this data has proven to be very lucrative and led to an increased demand for people who can harness “Big Data.” Some courses now exist at the undergraduate and graduate level that seek to teach students how to work with modern data. We believe similar efforts are possible at WPI. Modifying the undergraduate curriculum to feature data analysis and processing methods will benefit aspiring data scientists who will work with data in their careers. Additionally, new rich sources of data can replace trivial or outdated classroom examples and make the class material more engaging. Despite these positive effects, limited time and resources mean that teaching data comes at the cost of teaching other essential topics. Thus, we need to consider the entire impact of increasing the data education. This paper seeks to understand the effects of teaching data at the undergraduate level and to determine if modifying the curriculum to include data would benefit undergraduates at WPI.

We define *data education* to be the processing and analysis techniques that students will use when working with data in their future careers. Mathematics, especially statistics, has historically been the field responsible for teaching data skills. At WPI, the Mathematical Sciences department serves two distinct populations. There are mathematics majors; students who take advanced courses in mathematics in connection with their degree. This population includes students of interdisciplinary programs such as those in the data science program who take advanced courses in mathematics, computer science, and business to achieve their goals. However, there are also students who take mathematics courses as prerequisites for their engineering or science degrees. For non-mathematics majors, mathematics courses serve to teach practical analysis and reasoning skills that they will use in advanced major-specific courses.

In this report, we interview experts who represent these two different points of view. We

first interview professors of mathematics to gather their perspective as to the purpose of the current mathematics curriculum and to understand current practices in teaching data education. We also contact mathematics education experts outside of WPI who are well known for their contributions to mathematics education to understand the connection between mathematics and data. Next, we interview professors of non-mathematics fields at WPI to get an opposing view of the purpose of the Mathematical Sciences department. These professors see the mathematics program as responsible for teaching prerequisite skills so that students can succeed in advanced analytic courses within their own major. They have opinions as to what students should learn in mathematics before taking their classes. Finally, we interview the head of the Mathematical Sciences department to acquire the authoritative vision of the mathematics program. From these interviews, we hope to determine if data should be taught at the undergraduate level and whether the Mathematical Sciences department is the right place to do so.

By surveying professors in the Mathematical Sciences department at WPI, we find that professors are actively trying to incorporate modern data applications in their classes. However, they are constrained by the brevity of 7-week terms and by requirements to teach other mathematics topics. Also, processing more than trivially small amounts of data requires some background in computer science, and professors who try to incorporate a data analysis project in their classes find they must first teach computer programming. Finally, mathematics professors would like to address the major-specific data needs of their students, but often it is too challenging to tailor the course curriculum to the extent desired because classes are composed of students from many different fields of study.

Among the majority of mathematics professors, there was a mutual feeling that data education fits best in linear algebra courses.

In our interviews with non-mathematics professors, we find that students are not adequately prepared to manage the advanced material. From their perspective, mathematics courses spend too much time teaching theoretical concepts and do not give students enough practice with data.

Generally, there was not a consensus about what data was and how to incorporate it. A lack of formal language about data education means that different professors across depart-

ments have wildly different ideas of what is sufficient education.

By reviewing descriptions and syllabi for courses that cover data collection and processing, it is clear that different fields of study have different data education needs. The prerequisite knowledge for these courses is not well covered by the mathematics foundation, as evidenced by professors' sentiment of ill-prepared students.

To address the nuanced requirements of data education from each major, we suggest that there exist more communication between the Mathematical Sciences department and the other departments on campus for the formation of major-specific data education requirements. These requirements could guide the formation of an additional mathematics practicum that meets the needs of student's major field of study and acquaints students with the types of data and analysis they are likely to encounter in their future careers.

Table of Contents

Abstract	i
Executive Summary	ii
List of Figures	vii
1 Introduction	1
2 Background	5
2.1 Existing academic programs for data	5
RPI’s Data Analytics Through-Out Undergraduate Mathematics program . .	6
Northeastern University’s Humanities Data Analysis course	7
2.2 The need for data education	8
What is data education?	8
Data literacy	8
Data education needs	9
Mathematics as a possible place to fit a data education	12
2.3 Summary	13
3 Methodology	15
3.1 Goal 1: Evaluate the use of data in undergraduate mathematics	16
Objective: Interview mathematics professors at WPI	16
Objective: Interview mathematics experts outside of WPI	18
3.2 Goal 2: Data education requirements at WPI	19
Objective: Interview non-mathematics professors at WPI	19
Objective: Interview the head of the Mathematical Sciences department . . .	20

4 Results	22
4.1 The current mathematics curriculum according to mathematics professors. . .	22
The objective of the mathematics curriculum is to teach problem-solving and	
to give a background in mathematics.	24
Challenges for students	25
Using data to visualize theoretical concepts can help students get an intuitive	
understanding.	30
Using Big Data can help students get a deeper understanding of mathematics	
concepts.	32
Using data in class requires programming.	34
Linear algebra is a good place to introduce students to data education. . . .	36
Summary of interviewing mathematics professors.	37
4.2 Undergraduate mathematics according to experts in mathematics education	38
4.3 Data education requirements according to professors outside of Mathematical	
Sciences	40
Experimental Analysis in Chemical Engineering	41
Experimental Analysis in Aerospace/Mechanical Engineering	42
Summary of interviews with non-mathematics professors	43
4.4 Data education responsibilities according to the head of the Mathematical	
Sciences department	43
Mission of the Mathematical Sciences department	45
Concern of non-mathematics professor: lack of practical application	46
Current revisions to the mathematics curriculum	46
Summary of interview with mathematics director	47
5 Conclusions	48
References	50
Appendix A: IRB Approval	54

List of Figures

1	Comparison of memory sizes and examples of storage capacity.	1
2	Comparison of RPI's data science course and WPI's data science courses. . .	7
3	Questions for mathematics Professors at WPI	17
4	Questions for mathematics Education Experts outside WPI	18
5	Questions for Professors of Data Analysis Courses at WPI	20
6	Questions for Head of Mathematical Sciences Department at WPI	21
7	Using a matrix to encode coefficients for a series of linear equations.	36
8	Plotting values from a matrix.	37

1 Introduction

Since the first digital storage device in 1946, the capacity for data storage has increased massively. The first device held 10 Kilobytes; the size of the first chapter of Moby Dick in plain text (Napper, 1999). In 2011, there were an estimated 295 Exabytes (equivalent to 295 billion copies of Moby Dick) of stored data in the world (Nguyen, 2011).

Memory size	What can be stored?
1 byte	the letter 'a'
1 kilobyte = 1024 B	first 190 words in Moby Dick
1 megabyte = 1024 KB	the entire novel Moby Dick
1 gigabyte = 1024 MB	Moby Dick recorded as an audiobook (estimated)
1 terabyte = 1024 GB	all books published before 1850 (estimated)
1 petabyte = 1024 TB	all books published ever (estimated)
1 exabyte = 1024 PB	all books published ever as audiobooks (estimated)



Figure 1: Comparison of memory sizes and examples of storage capacity. Multiple audiobooks of Moby Dick exist, so we chose a rough estimation of size. We used WorldCat to search for the number of books published before 1850. We calculated the Petabyte example by approximating that all audiobooks were the same size as Moby Dick. To approximate the size of all books, we used the number of books estimated to exist according to the Google Books project (Taycher, 2010). Free clip art source (*Clipart of sperm whale*, 2017).

Simultaneously, the processing power of computers has eclipsed itself many times over. Whereas in 1966 it took the world's fastest computer 28 days to calculate 250,000 digits of pi, a modern home computer can perform the feat in a single second.¹ Naturally, as data storage and processing capacities have exponentially increased, so has the amount and types

¹I calculated this using the 2009 pi calculation world record with an Intel i7 2.93 GHz processor. A comparable computer with a 1 Terabyte hard drive is currently available for less than \$1000 at Best Buy. (*Chronology of computation of pi*, 2017)

of data that we store.

In the 1990s, the term “Big Data” was coined to refer to a new type of data. Compared to traditional forms of data, such as experimental measurements from an instrument or entries in a financial ledger, “Big Data” referred to amassed collections of data from diverse sources and in large quantities. The first use of the term is informally attributed by a New York Times article to John Mashey, the chief scientist at Silicon Graphics (Lohr, 2013). In a 1998 presentation entitled “Big Data... and the Next Wave of InfraStress”, Mr. Mashey states that “rapidly increasing storage and user demand for data (images, audio, and video) via the Internet has resulted in an explosion of widely accessible data” (Mashey, 1998). A working paper by Professor Francis Diebold at the University of Pennsylvania also credits computer scientists Weiss and Indurkha with use of the term in 1998 to refer to centralized data collections that allowed analysts to examine trends in the large (Diebold, 2012).

Today, large quantities of data are freely available for download and use on sites such as Kaggle². In their 13,000 publicly available projects, Kaggle includes data from sources such as linguistics, the Internet, finance, politics, demographics, crime, business, languages, healthcare, and sports (*Kaggle Datasets*, 2018). The largest of these datasets is 30 Terabytes and claims to be the most comprehensive repository of human society for the past 215 years containing news articles, images, videos, social media, quotes, names, “and more” (*The GDELT Project*, 2018).

The rise of the Big Data paradigm is marked not only by the increased capacity to store and process data, but also the availability of data and the various potentials for its application. A 2016 article by Schaefer Marketing Solutions lists 37 organizations and businesses that have capitalized on Big Data (Petersen, 2016). One such business is The Weather Channel, who has leveraged weather sensors and users’ mobile data to sell targeted advertising for products such as frizz-reducing shampoo. Another organization is Next Big Sound: a group that predicts the success of current artists and musicians via activity on Pandora, Facebook, Twitter, YouTube, and Wikipedia to assist musicians, label companies, and marketers understand music listeners. The Internal Revenue Service (IRS) has also utilized Big Data to profile taxpayers for the prevention of identity theft. From the examples in this article,

²<https://www.kaggle.com/>

there are data applications in marketing, government, insurance, banking, transportation, engineering, and entertainment.

There are also applications for data in biology and healthcare. One of the most famous works in Big Data is the Human Genome Project: an initiative started by the National Institutes of Health in 1990 to sequence the entire human genome (*All About The Human Genome Project (HGP)*, 2015; joannefox, 2006). The project was completed in 2003 and is freely available on the National Center for Biotechnology Information (NCBI) website. Since that time, there has been the development of a whole new branch of biology: genomics, or the study of the human genome. Subsequent sequencing projects of model organisms, like the mouse, are used by biologists and biomedical engineers for cloning and gene sequencing. These technologies have been used to trace ancestral lineage, screen for precursors to genetic diseases, and provide personalized treatments.

The commercial success of Big Data is evident in the increased job demand for data scientists. The Bureau of Labor Statistics cites two professions related to analyzing data: Computer and Information Research Scientists, and Mathematicians and Statisticians. Computer and Information Research Scientists “study and solve complex problems in computing for business, medicine, science, and other fields”, and employment is expected to grow by 19% (*Computer and Information Research Scientists*, 2018). Mathematicians and Statisticians “analyze data and apply mathematical and statistical techniques to help solve real-world problems in business, engineering, healthcare, or other fields”, and their employment is expected to grow by 33%. Both of these are exceptionally high rates of growth compared to the average of 5-8%.

In this paper, we are interested in understanding how the Big Data paradigm impacts undergraduate education. There is a bi-directional relationship between data and education. On one side, there is the potential for universities to teach modern forms of data analysis and thus prepare aspiring data scientists to work with Big Data. On the other, there is the potential to use Big Data applications to engage students in otherwise abstract topics and thus improve their educational experience. We explore both of these ideas to fully understand the impact that Big Data can have at the undergraduate level.

At Rensselaer Polytechnic Institute (RPI), data science professors have made data analy-

sis a requirement for all undergraduates with the Data Analytics Through-out Undergraduate Mathematics (DATUM) Program (Martialay, 2018). In the program, students take an Introduction to Data Mathematics course and complete a capstone project in their field of study. RPI's program is the first of its kind to make data analysis a universal requirement for undergraduates.

At WPI, traditional forms of data analysis are taught by the Mathematical Sciences department in courses such as statistics. One idea would be to update these courses to include teaching modern data analysis techniques such as used by data scientists in the applications mentioned above. However, these statistics courses are commonly prerequisite for other fields of study as well such as engineering, biology, chemistry whose upper-level courses involve interpreting laboratory or experimental data. Thus, we seek to identify the data education requirements of these student populations to determine if they would benefit from learning modern data analysis methods. In particular, we are interested in determining if these requirements are different from those of a more typical “data science” student.

Some upper-level mathematics courses at WPI have used Big Data to motivate topics in class. One such class is *MA 463X: Data Analytics and Statistical Learning* in which students use the methods they have learned to analyze a publicly available dataset (*Mathematical Sciences*, 2018). One could imagine, for example, using a similar approach to show a biology student how Calculus will help them in their future career. We are interested in finding such opportunities to make topics more engaging and personal for students.

This project will examine the impact of data on undergraduate education at WPI. This investigation should help the university to develop a curriculum that makes use of Big Data sources and teaches modern analysis techniques. We will interview professors at WPI and experts in undergraduate education outside of WPI to understand the data education requirements of all students and to identify opportunities to improve the undergraduate experience with data applications.

2 Background

Recent advances in storage and processing technology have fundamentally changed how we work with data. More so than ever before, we rely on Big Data to make informed decisions, and there is a demand for people who can process data for this purpose. Because of the diversity of available data and the numerous success stories of learning from data, we believe that nearly every industry can benefit from using data. Indeed, we predict that data science skills will be valuable for professionals in every industry as they advance their fields amidst this new data capability.

Universities have a responsibility to prepare their students to be successful in their careers and are thus responsible for providing a data education. We seek to understand the data education requirements of the undergraduate students at WPI and to determine if existing courses are meeting these needs. Mathematics, especially statistics, continues to play a profound role in the manipulation of data, and the teaching of methods to deal with data has already appeared at the undergraduate level. We investigate the existing relationship between mathematics and data to determine if the mathematics program is the right place to introduce a formal data education.

2.1 Existing academic programs for data

Academic programs now exist at the graduate and undergraduate level designed to educate data scientists. We profile two such programs. The first is an initiative at Rensselaer Polytechnic Institute (RPI) for an undergraduate data science program. It makes introductory data science courses a requirement for all students and offers upper-level courses for students interested in data science as a career. The second is a graduate course at Northeastern University that teaches analysis methods to humanities students. It is unique in that it teaches data science as a specialty for humanities researchers.

WPI also provides a graduate data science program and data science courses for under-

graduates. However, we believe there is potential to increase the integration of data within the current curriculum to accommodate data scientists and to engage students within and outside of the Mathematical Sciences department. We use these two examples to show how this can be done.

RPI’s Data Analytics Through-Out Undergraduate Mathematics program

RPI has introduced a new undergraduate program designed to give students “data dexterity”, a term meaning proficiency in analyzing diverse datasets (Martialay, 2018). Data education is a core requirement for all undergraduates. All students take an introductory course on data modeling and analysis and a second course within their primary field of interest in which they complete a capstone project. Undergraduates who enjoy these courses can participate in a summer research experience to practice what they learned in class. If students desire to follow a data science path, they can continue to take data science courses such as a laboratory course in which they analyze a dataset in collaboration with industry partners.

WPI currently has two undergraduate data science courses, and they cover the same material as the first required data course at RPI. RPI’s *Introduction to Data Mathematics* states in the course description to feature “basic data analysis techniques for data visualization, classification, clustering, and ridge regression with case studies to understand high-dimensional data.” (*MATP 4400 - Introduction to Data Mathematics*, 2017). WPI’s *DS 3001: Foundations of Data Science* and *MA 463X: Data Analytics and Statistical Learning* respectively cover “data collection, data management, statistical learning, data mining, data visualization, cloud computing, and business intelligence” (*Undergraduate Catalog 2017-18*, 2017) and “regression, classification/clustering, sampling methods (bootstrap and cross-validation), and decision tree learning” (*Mathematical Sciences*, 2018). Figure 2 shows the overlap in these three courses.

RPI’s data program extends over that offered at WPI in that the data science course is required, and students must complete a data science capstone in their major. In contrast,

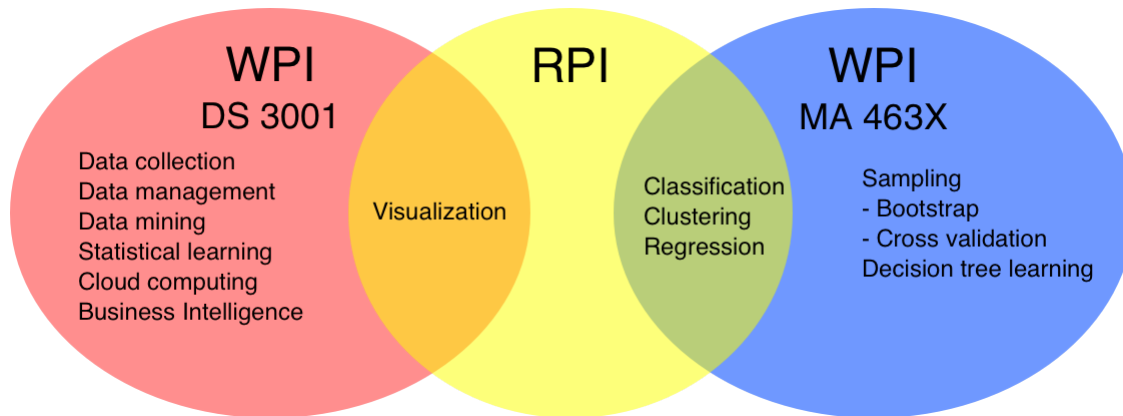


Figure 2: Comparison of RPI’s data science course and WPI’s data science courses.

the data courses at WPI are generally designed for students interested in data science and are not requirements of other majors.

Northeastern University’s Humanities Data Analysis course

Humanities Data Analysis is a graduate-level humanities course that teaches students a variety of computational methods that can be used with digitized historical artifacts (Cordell, 2017). Students work primarily with R, an open-source programming language that supports statistical computing and visualization, to clean and encode text artifacts, tabulate and plot word frequencies, identify literary patterns with topic modeling, use spatial mapping to visualize geographic regions described in literature, and plot the evolution of vocabulary use with vector space models¹. This course does not feature a capstone data project; instead, students read articles from humanities research journals that show how the methods they learn in class apply to research. This approach allows students to learn a range of humanities data analysis methods without requiring them to learn general-purpose analysis methods.

Northeastern University’s course is notable because it teaches data science to a very specialized population. Compared to the undergraduate population at RPI, the students who take this course know the specific application for which they will use data science. As a result, the course can teach more relevant topics such as spatial mapping and vector space models because it does not need to satisfy a general audience.

Innovations in digital storage technology have led to the rise of Big Data. Encouraged

¹R packages used: tidytext, tidyverse, tokenizers, dplyr, ggplot2, mallet, ggmap, wordVectors

by the widespread interest in data science, there is an opportunity for universities to adapt their curricula to be more data-focused. Some programs have already started to offer undergraduate data education requirements and domain-specific graduate programs. We believe that similar efforts can occur at WPI. In the next section, we will emphasize the need for data education to prepare aspiring data scientists and to modify the standard undergraduate curriculum to be more engaging and modern.

2.2 The need for data education

In this section, we define data education and discuss articles related to providing a data education at the undergraduate level. Most of the data-related courses are mathematics courses. We believe this is evidence of a close relationship between data and math. Later in this report, we will revisit this relationship to determine if the Mathematical Sciences department should be responsible for providing a data education for all students.

What is data education?

In this report, we define *data education* to be any curricula or topic related to data analysis that a student needs to know for their future career. This definition is purposefully broad; data education can manifest differently depending on a student’s field of study and personal interests. RPI uses the term *data dexterity* to mean “proficiency in analyzing diverse datasets” (as discussed in Section 2.1); thus, data education is the means by which students attain data dexterity.

Data literacy

Data literacy is **not** data dexterity. It is the skill of understanding a data analysis to the point that one is not susceptible to being “taken” by misleading conclusions². While data literacy is a critical skill for all who read the news, see advertisements, or otherwise use statistics generated by others, it is out of the scope of this paper. Our paper focuses

²This article gives an introduction to data literacy and shows several examples of misleading graphs: <https://venngage.com/blog/misleading-graphs/>

instead on *data education*, the method by which aspiring professionals learn specific tools and techniques to deal with data.

Data education needs

Several articles have been written talking about ways to incorporate a data education. We present four perspectives regarding data education and highlight the methods by which they propose changing the curriculum.

Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward. This first article was written in *The American Statistician*, a publication by the American Statistical Association, states that the standard statistical (data) education is insufficient. Learning from data, they write, includes understanding variability and bias, which analytic methods to use, rational interpretation, and accounting for uncertainty. Students taking traditional statistics courses often see statistics as “memorization of the use of ‘cookbook’ methods and the postulation and rote application of null hypothesis statistical procedures” (Horton, 2015).

The article sets three goals for improving the data curriculum: use more multivariate methods, develop data skills early, and expand the role of simulation and computation. For the first goal, use more multivariate methods, they observe that introductory statistics courses teach students that data is clean where in truth, data is messy. They say students need to develop an open mind to recognize confounding variables, factors not explicitly represented in the data that an untrained eye will wrongly misattribute to something else. Statistics courses need to use appropriate tools that show complex, multivariate relationships to teach students how to recognize confounding variables and other characteristics of messy, real-world data. The danger in not doing so is that when students encounter messy or complicated data, they will see statistics as an antiquated discipline incapable of dealing with anything but the most constrained scenarios.

The second goal, to develop data skills early, means that students need to wrestle with datasets of all shapes and sizes starting early in their education. Professors need to use dataset formats that go beyond the standard Excel spreadsheet. SQL databases, outputs from web scraping, and XML are examples of modern and commonly used dataset formats.

Using these kinds of datasets will likely require computational and visualization skills. The authors suggest teaching these skills in an undergraduate capstone experience after one or two introductory courses in statistics.

The third goal is to expand the role of simulation and computation in teaching data. This goal diverges from the second in that simulation and computation can be used for problem-solving outside of a data analysis setting. For example, students can use simulation to solve complex polynomial equations much easier than by solving by hand. These computational skills can grow problem-solving capabilities while at the same time encouraging analytical thinking.

In summary, this article says that traditional statistics courses do not represent modern Big Data challenges and that the statistics course should be updated to include more multivariate methods, datasets of all shapes and sizes, and an increased presence of simulation and computation.

Ensuring That Mathematics is Relevant in a World of Data Science. Another article agrees that the existing undergraduate curriculum for data education is insufficient, but that the cause is a cumbersome introductory mathematics sequence that precludes students from working with modern data (Hardin & Horton, 2017).

According to this article, many data science programs now exist at the undergraduate level, and it is with a consensus that data science students should be taught calculus, linear algebra, discrete math, probability, and statistics. The issue in this sequence is the length; an undergraduate data scientist spends a majority of their time in these general prerequisites where instead these introductory courses should change to show the “power of mathematics and statistics” (Hardin & Horton, 2017).

They suggest the formation of two new courses, one in discrete mathematics and another in continuous mathematics, to take the place of the standard introductory sequence. The discrete mathematics course should model real-world data and promote problem-solving. In particular, this class should teach linear algebra (independence/invertibility, Markov models, and eigenvalues), counting principles (first principles of randomness), computational (discrete) simulations of continuous models, and graph theory (understanding confounding, causal inference, and analysis of network data) (Hardin & Horton, 2017). The continuous

mathematics course should cover functions, basic mathematical logic, enough calculus to understand partial derivatives, Taylor expansion method for approximating functions, probability as area/integration, and multivariate thinking (functions, optimization, integration) (Hardin & Horton, 2017). They suggest that these two courses cover the necessary topics otherwise taught in standard sequence (calculus, linear algebra, discrete math, probability, and statistics). Besides these topics, these courses should emphasize computing for data analysis, simulation, and modeling. “Using computational skills to simulate produces a deeper understanding of the model” (Hardin & Horton, 2017).

The preceding two articles talk about data education needs from the perspective of the statistician and mathematician. The remaining two articles talk about data education from the perspective of the scientist and engineer and is related primarily to laboratory-based courses.

The Role of the Laboratory in Undergraduate Engineering Education. This article says that engineering students need to focus more on practical application rather than theoretical understanding (Feisel & Rosa, 2005). Modern advancements in the laboratory include the advent of digital computers. As a result, objectives of a modern engineering lab course include (1) applying the right sensors, instruments, and software to measure physical quantities, and (2) collecting, analyzing, and interpreting data and form conclusions. Students in lab-based classes need a basic understanding of statistics and possibly programming, but they do not need training in theoretical methods.

Statistical Treatment of Experimental Data. The belief that practical experience outweighs computational methods seems unchanged from the opinion expressed by this 1962 engineering reference (Young, 1962). The book tells that science and engineering students deal with experimental data, so these populations need elementary statistical knowledge. The students who come into these higher level laboratory courses are often ill-prepared in mathematics. In the author’s opinion, students should not prolong their lab education to get a rigorous (i.e., theoretical) understanding of statistics. Instead, the book provides short, intuitive derivations of useful formulas that will help students succeed in laboratory courses.

These four articles demonstrate ways in which the undergraduate curriculum could change to accommodate data education. The first article suggests modifying existing statistics

courses to feature more modern data techniques such as multivariate modeling, challenging dataset formats, and simulation and computation. The second suggests condensing the standard introductory mathematics sequence into two courses to spend more time instead teaching computation and working with data. The third and fourth articles, however, take a step back from new Big Data applications. Instead, they stress that practical data experience is more valuable than theoretical knowledge and that students should only acquire basic knowledge in statistics and programming and not waste time learning theoretical mathematics before starting in the lab.

One insight from these articles is that they view data education as a mix of mathematics and computation. Another is that incorporating data comes at the cost of some other instruction such as offered in a full introductory mathematics sequence or a more rigorous statistics course. These two ideas, that data is a mix of mathematics and computation, and that teaching data comes at the cost of another instruction, will resurface as we evaluate data education at WPI.

Mathematics as a possible place to fit a data education

At a 2009 workshop entitled *Teaching Undergraduates Mathematics*, professors from many different disciplines met together to discuss their demands of the Mathematical Sciences department. The following are the data-related demands mentioned in their workshop report (Lai, 2009).

The first demand originated from a survey of faculty from math-prerequisite fields such as engineering, physics, chemistry, computer science, economics, business, and biology. According to these experts, mathematics courses should emphasize “approximation and estimation,... the use of spreadsheets in place of graphing calculators,... facility with applications, and fluency with symbols and graphs as a language tool“ (Lai, 2009). In the context of calculus derivatives, students should spend less time learning how to take complicated derivatives and more time working with the abstract concept of “the derivative“ and that it represents slope.

A second demand, specifically from a biology professor, was that mathematics should be immediately relevant to biology. As an example, mathematics courses should focus on quickly

identifying and understanding various function curves that frequently appear in biology, such as linear, log-linear, Gaussian, logistic, and chaotic. Another example is teaching graphs since graphs are frequently used in biology to represent food webs and pedigrees (Lai, 2009)..

These two demands suggest that these faculties consider data education to be the responsibility of the Mathematical Sciences department. Mathematics courses should be tailored to fit the student's primary field of interest, and mathematics topics should be taught showing industry-specific applications.

2.3 Summary

Changes in data storage and processing capacity have led to the rise of the Big Data paradigm. The commercial success of Big Data and increased market demand for data scientists have resulted in the formation of new data science programs at the graduate and undergraduate level. In Section 2.1, we looked at two such programs. The first was an initiative to make data science a requirement for all students and for students to complete a capstone project in their major. The second was a highly specialized data science course that catered to humanities researchers. These programs suggest some ways in which WPI's data offerings can be changed to adapt to Big Data.

In Section 2.2, we looked at articles that said the undergraduate curriculum should change to incorporate a data education. Data education is a mix of mathematics and computation, and we define it as any data analysis topic a student might need in their future career. For statisticians, data education might mean fluency with multivariate models and ability to make computer simulations. For engineers, data education might mean experience with specific instruments and familiarity with experimental data.

In Section 2.2, we hinted at a relationship between mathematics, especially statistics, and data education. The mathematics program seems like a reasonable place to fit data because of this relationship as well as the fact that most students are already required to take mathematics courses as prerequisites for their primary field of study. However, modern forms of data require some amount of programming, a skill more commonly taught in a computer science course. Additionally, the articles above suggest that a modern data

education does not require instruction in theoretical mathematics topics. Because of this, teaching a data education may be contrary to the mission of the mathematics program to teach rigorous mathematics. If so, incorporating data education in mathematics courses would place an unfair responsibility on the Mathematical Sciences department to prepare students for courses outside of mathematics without simultaneously contributing to their objectives.

3 Methodology

In the Background, we showed that undergraduate mathematics plays a significant role in teaching methods to deal with traditional forms of data. Because of this existing relationship between mathematics and data, it makes sense that the Mathematical Sciences department should now be responsible for teaching modern analytic skills. Besides, the Mathematical Sciences department already reaches nearly every student at WPI because of prerequisite mathematics requirements for most majors. Thus, modifying the curriculum to teach data methods to all students would require minimal effort compared to making individual changes in each department at WPI. On the other hand, working with Big Data requires new techniques not primarily taught in mathematics, such as computer programming. Additionally, the increased diversity of available data means that there is the potential to use Big Data in nearly every field of study. These changes provide a rich opportunity to engage students with applications within their field of interest and to teach them data techniques that will serve them best in their future careers. However, this customized approach to data education requires educators to have some domain knowledge in every field which may be an undue burden on the Mathematical Sciences department. Thus, it is unclear whether a modern data education should be the sole responsibility of the Mathematical Sciences department.

We set out two goals for this project. The first goal was to evaluate the use of data in the current mathematics curriculum. Specifically, we were interested in understanding the objectives of the undergraduate mathematics program and determining if the program would benefit from adding data. The second goal was to identify the student populations at WPI who were in need of a data education and to determine what specifically they needed to succeed. In so doing, we hoped to decide whether administering a data education to these students was the responsibility of the Mathematical Sciences department.

3.1 Goal 1: Evaluate the use of data in undergraduate mathematics

The first goal of our project was to evaluate the use of data in the undergraduate mathematics curriculum. Mathematics has remained conceptually the same for centuries, and the teaching of introductory mathematics courses is well understood through many iterations of teaching the same material. However, the Big Data paradigm has had such an impact on practically every industry that we believe data could have a similar impact on mathematics.

To achieve our goal, we first set out to evaluate current practices in undergraduate mathematics. Inherent in these practices is a motivating theory of the purpose of teaching mathematics to undergraduates. We identify this purpose to gain a point of reference for what data should hope to accomplish. Additionally, we gathered opinions from experts in mathematics education as to the potential for using data in the classroom.

Objective: Interview mathematics professors at WPI

The first group to interview regarding the current state of mathematics education was mathematics professors at WPI. These professors have a valuable perspective because they directly interact with students in mathematics classes. Since these professors are responsible for teaching mathematics, we expected them to have well-formed opinions as to what undergraduate mathematics should strive to achieve. We also expected them to be able to identify mathematics concepts students find difficult. These concepts gave us an idea of the current challenges in mathematics education that may be helped or hindered by data. Finally, we asked for their opinions as to data in undergraduate math.

Since this is the first group we planned to interview, we administered interviews in a way that allowed for open discussion. Based on the responses from these professors, we had a better idea of how to frame questions when talking to other groups. To allow for open and free conversation, we elected to administer in-person interviews with open-ended questions. We expected these conversations to take a half-hour to an hour to complete. We tape-recorded the interviews to allow us to focus our attention on the conversation and to

think about follow-up questions.

To get a diverse perspective on the Department of Mathematical Sciences, we targeted professors across the entire Mathematical Sciences department. We interviewed professors of introductory courses like Calculus as well as upper-level courses like Numerical Methods; seasoned WPI professors as well as new-hires; and professors from different specialties such as differential equations, finance, and abstract algebra.

These are the questions we prepared for the interviews, but we expected the professors to direct the conversation to what they felt was important:

- Questions for mathematics Professors at WPI

 1. What is the objective of undergraduate mathematics? What should students know?
 2. What are the concepts that are hardest for your students to grasp?
 3. Do you think data can improve mathematics courses?
 - a. If yes, why and how?
 - b. If no, why not?
 4. Have you incorporated data into your courses?
 - a. If yes:
 - i. What was your methodology for choosing data?
 - ii. From where do you get your data? Who collects it?
 - iii. How do you find data most useful/relevant for students?
 - iv. What is the hardest part of getting data ready to use?
 - v. What is the thing you find easiest/hardest about using data in classes?
 - b. If no:
 - i. Why not? (doesn't exist? not convenient?)
 - ii. If a relevant dataset did exist and was convenient (already clean), would you consider using it?

Figure 3: Questions for mathematics Professors at WPI

Objective: Interview mathematics experts outside of WPI

Two significant findings from interviewing mathematics professors that we foreshadow here are that (1) the linear algebra course is a natural place to add data and (2) skills in programming are necessary to be able to use data in class. This first finding was particularly novel to us, so we chose to verify it by asking the opinions of experts in the field of linear algebra education. We defined experts as those who had published textbooks for undergraduate linear algebra courses.

We identified experts to contact by asking for recommendations by mathematics professors at WPI and by leveraging professional connections of our research advisors with these individuals. Since this group consisted of members of the mathematics community located outside of WPI, we elected to send them questions by email. We restricted the number of our questions to three to increase the likelihood that these experts would be willing to respond to our questions.

Because of the relationship between data and linear algebra, we chose to reach out to known figures in linear algebra education. When interviewing these experts, we asked specifically about using data in linear algebra. In response to this second finding, we asked these experts their opinions about programming in math. We sought to verify that programming was a necessary skill and not just an artifact of WPI's technical culture.

- | |
|--|
| <p>Questions for mathematics Education Experts outside WPI</p> <ol style="list-style-type: none">1. Do you think that data should be used to support mathematics education? If so, how?2. In what ways (if any) does undergraduate mathematics fall short in preparing students for a data-driven workplace?3. (If a professor of the specific field, X) What do you think about using X courses to introduce data analysis skills?4. Do you believe that programming is prerequisite for using data in courses? If so, does it impede the incorporation of data into the curriculum? |
|--|

Figure 4: Questions for mathematics Education Experts outside WPI

3.2 Goal 2: Data education requirements at WPI

The second goal of our project was to evaluate the data education requirements of student populations at WPI. We hypothesized that different majors across WPI, such as chemistry, engineering, and the humanities, had distinctly different needs concerning data analytic skills. Thus, we planned to interview professors from diverse fields to talk about their domain-specific data applications.

The populations interviewed in conjunction with the first goal were necessarily similar; all of them were directly involved in mathematics education and had a mathematic perspective of data. With this second goal, we targeted individuals with alternative perspectives regarding the role of data in undergraduate math.

Objective: Interview non-mathematics professors at WPI

The first group we chose to interview was professors of non-mathematics disciplines within WPI who could speak about data education needs. Specifically, we targeted professors of courses offered in the past year whose course descriptions included keywords such as “data”, “analysis”, and “experiment.” Instead of asking about challenging mathematics topics for students, we asked these professors if students taking their courses were well prepared to deal with data.

Professors in this group were large in quantity and spread across WPI’s campus, which made it difficult to talk to each of them in person. We chose to contact these professors via email and send them our list of questions. Similar to our approach contacting mathematics education experts, we chose to limit the number of questions to three to increase the likelihood that the professors would respond.

Since these professors taught classes advertising “data”, “analysis”, and “experiment”, we asked questions about the use of data in their courses. We wanted to know what about using data was challenging for students, and if these difficulties resulted from a deficiency in a specific mathematics topic. If professors of different departments identified the same mathematics topics, this would mean that data needs of different students could be met

with a shared curriculum, such as one prepared by the Mathematical Sciences department. Finally, because of our unfamiliarity with all departments across campus, we asked these professors to direct us to others in their department who could speak to the data education needs of their field. These are the questions we asked professors of data analysis courses:

- | |
|---|
| <p>Questions for Professors of Data Analysis Courses at WPI</p> <ol style="list-style-type: none">1. How did students interact with data in your course?2. Were students well prepared to work with data?<ol style="list-style-type: none">a. If no, are there any mathematics topics you wish students were more comfortable with that would have helped them succeed?3. Is there anyone else in your department you would suggest I contact next? |
|---|

Figure 5: Questions for Professors of Data Analysis Courses at WPI

Objective: Interview the head of the Mathematical Sciences department

In interviewing professors both within and outside of the Mathematical Sciences department, we identified conflicting data education requirements. This conflict led to a question as to which population should be responsible for overseeing a data education. We thus decided to approach the head of the Mathematical Sciences department at WPI. We hoped to get his opinion as to the mission of mathematics and whether or not teaching a data education aligned with that mission.

Because of the number of the questions we planned to ask, we requested an in-person interview with the head of the Mathematical Sciences department. As with the mathematics professors, we tape-recorded this interview to accurately capture his responses and so that we could stay focused on maintaining the conversation.

Questions for Head of Mathematical Sciences Department at WPI

1. What is the mission of the Mathematical Sciences department?
2. What are the causes of the challenges identified by mathematics professors?
3. What responsibilities does the Mathematical Sciences department have for preparing students for other majors?
4. What room is there, if any, for changing the undergraduate mathematics curriculum?

Figure 6: Questions for Head of Mathematical Sciences Department at WPI

4 Results

In this chapter, we look at the results of interviewing professors and experts to determine how Big Data should appear in the undergraduate curriculum at WPI. We set two goals for our project: the first was to evaluate the current role of data in existing mathematics courses to see if the mathematics program was a reasonable place to introduce modern data analysis methods. The second was to identify the specific data education needs of students at WPI and use this information to suggest approaches to teaching data.

As we will show, incorporating data into existing mathematics classes has the potential to help students overcome current challenges related to learning mathematics at the undergraduate level. There seems to be a clear link between data and linear algebra, suggesting that this course would be a reasonable place to teach modern data analytic techniques.

Professors inside and outside of the Mathematical Sciences department disagree about what courses should teach students. Professors of data-related classes outside of mathematics feel that the mathematics program should stress practical application and the use of real data more than teach theoretical mathematics. In contrast, professors within the Mathematical Sciences department feel practical application and theoretical understanding are equally valid, and they feel it is wrong to change the current curriculum to accommodate data.

The potential of data education in undergraduate math.

4.1 The current mathematics curriculum according to mathematics professors.

The purpose of evaluating the mathematics curriculum at WPI was to understand what the WPI professors are currently trying to teach undergraduates. In particular, mathematics professors at WPI are trying to achieve several goals: preparing students for careers outside of mathematics, preparing them for careers inside, and helping students become mathe-

matically literate to further societal goals. In this section, we identify challenges professors currently face in the furtherance of these goals. One of our focuses is how the mathematics curriculum can be more data-focused and thereby improve the goals of the Mathematical Sciences department.

To this end, we had in-person interviews with 6 of the 46 mathematics professors at WPI. These mathematics professors were chosen to cover a range of the mathematics interests and curricula at WPI. We interviewed Professor Weekes, who focuses on numerical methods and has been a member of the WPI mathematics faculty for over 20 years. Professor Weekes was instrumental in the formation of computer lab sessions teaching Maple and Matlab for the calculus sequence. We interviewed her to understand the history of change in the mathematics program at WPI and to get her perspective on how to implement further changes in the mathematics program. On the other end, we also interviewed Professor O’Cathain who teaches abstract algebra. He is a new faculty member at WPI and brings a unique perspective from his education and teaching tenure in Ireland. We were especially interested in his perspective of the mathematics program as it compares to other schools and styles of teaching.

In addition to Professor Weekes and Professor O’Cathain, the other professors I interviewed are Professor Olson in numerical methods, Professor Sarkis in numerical methods and calculus, Professor Sturm in financial and actuarial math, and Professor Tilley in differential equations. Together, these professors cover a range of courses that are representative of those offered by the Mathematical Sciences department. These courses include numerical methods, both calculus and differential equations and linear and nonlinear systems, numerical analysis, undergraduate group theory, discrete math, differential equations, actuarial math, and numerical optimization.

The following subsections contain the main findings of our interviews with the professors. First, the professors identified several objectives of the mathematics program. Second, the professors listed three challenges common to undergraduates in math. These objectives and challenges serve as a guidepost for what data should hope to accomplish in the undergraduate curriculum. Finally, the professors discussed their difficulties with incorporating data into their own classes. We found that each professor interpreted the word “data” differently and

that these different interpretations led to widely different answers to our questions.

The objective of the mathematics curriculum is to teach problem-solving and to give a background in mathematics.

Mathematics professors identified two primary objectives of the undergraduate mathematics program: first, to teach students problem-solving as a skill, and second, to give students a foundational background in mathematics. Solving problems in mathematics means thinking rigorously and using proof-based approaches to find solutions. “The whole idea of a mathematics curriculum is to look at problems very differently than other areas of engineering or science. There’s no hypothesis testing in mathematics. You get a theorem, you prove it,” answers one professor. “Students have to be able to think independently and think clearly. You want to have some sort of environment or situation where students develop their own confidence and skill to say ‘this is how we look at these different problems and solve them.’”

The first identified objective of the curriculum is to introduce students to the process of solving problems. We consider the calculus sequence’s progression from easily visualized mathematics topics to abstract concepts as an archetype for this process. Some introductory mathematics topics can be understood visually. For example, when taking the derivative of a function, students can verify their solutions by comparing plots of the function before and after derivation. These topics help students gain confidence in their mathematics abilities and to look for alternative ways to verify their solutions. However, other topics, such as determining the convergence of an infinite series, cannot be easily visualized and require students to rely on proofs and formulas to come to an answer.

Progressions like this are opportunities for students to practice thinking abstractly and using logical proofs to verify the correctness of their solution. Mathematics at increasing levels of abstraction helps students move away from only trusting visually verifiable solutions to instead using water-tight logic to ensure correct answers.

Second, the department is tasked with providing nearly every student at WPI with a foundational background in mathematics. “We teach people who are interested in other ma-

jors but who need the basic math.” These are courses like Calculus, Differential Equations, Probability for Applications, and Statistics. Despite their pedantic nature, these classes are intended to teach students how to use a variety of mathematics concepts to solve novel problems. One professor claims, “The biggest thing is training [students] to know how to tackle a problem, whether it is with high-level mathematics and theorems and proofs or whether it is taking some methodology or framework that we have been talking about in class and performing some computation.”

It may not be entirely obvious, but these are two very different objectives. The first is agnostic to the application and is concerned with the formation of rigorous thought for problem-solving. This objective can be thought of as the “theory” aspect of mathematics. In the extreme, a course focused exclusively on teaching rigorous thought does little to let students achieve mastery of a particular method or formula. In contrast, the second objective gives students precisely the formulas they need to solve problems and is thus the “practical” aspect. However, students trained only in applied mathematics will lack the theoretical background necessary to spot fallacies in their reasoning or mistakes in their equations. It is unlikely that any student at WPI will graduate without first attaining some experience in both theoretical and applied mathematics. However, the challenge is in finding the right balance of the two that will best serve each student in their future career.

Foreshadowing the conclusion, we will claim that using data in introductory courses is useful because it shows students how the problem-solving techniques they learn in class have real application. There is a trade-off between bulletproof theorem-proving and rote memorization of useful formulas; there always has been. We will suggest, however, that data can make the best use of time spent teaching applications.

Challenges for students

In the following three sections, we detail three challenges of mathematics that mathematics professors identified as common amongst undergraduates. These challenges serve to explain what the mathematics professors at WPI are currently trying to do to improve mathematics education. By understanding these challenges, we can consider ways to use data applications to improve mathematics education.

Challenges for students 1: Moving beyond the trivial in-class example

Several professors agree that a challenging aspect of undergraduate mathematics is extending the fundamental concepts taught in class. A trivial in-class example helps students see the mechanics of a method, but each student eventually needs to “take the leap” in comprehension to be able to successfully utilize a new concept.

When asked to identify challenging topics in the calculus sequence, one professor volunteered epsilon. Students first learn that epsilon is a small positive number. As an illustration, imagine polishing a rock in a rock tumbler. The roughness of the rock is epsilon. As the rock tumbles, the surface becomes more smooth, and epsilon goes down. Hypothetically, one could polish a rock to the point that it is entirely smooth, where epsilon becomes zero. In reality, one can never achieve an epsilon of zero despite getting very close. In calculus, epsilons are used to account for small errors, such as what one gets when approximating a value by taking the limit. Since epsilons are infinitesimally small, students may tend to ignore them or instead rely on visual cues to solve problems. In this way, they can survive the standard calculus sequence without ever understanding epsilon. This professor desired some way to strengthen a vague definition of epsilon using some models or videos in class that appeal to student’s desire for visual understanding.

In group theory, another professor stated the hardest thing was probably the quotient group. “You define your operation on sets, and then somehow you define an equivalence operation. So, now you’ve got the equivalence classes, which are subsets of your set, and you define a new binary operation on these.” Simply, a quotient group can be thought of as a subset of things that share some characteristic. For example, if you had a basket of fruit and you wanted to organize it by color, a quotient group could be all the fruits that are yellow. Another group would be the fruits that are red. In group theory, you can define an operation such that two different colored fruits are equivalent and, as a result, have a subset containing one of each colored fruit that are all the same. What happens if you add two fruits together? Group theory allows you to define and perform this operation in a systematic way. It extends beyond what can be physically represented with fruit, and the professor felt students got lost transitioning to purely abstract thought. “The goal is to get

[students] to view all of these things in a unified way.”

One result of not developing a comprehensive understanding of mathematics topics is that students are less prepared for courses such as numerical methods. In this type, of course, students use advanced methods derived from elementary principles to approximate solutions for otherwise unsolvable problems. Students who cannot move beyond the trivial example will not understand the derivations and will be lost in this type of course. States one professor, “Numerical methods is good for people who have a good background. If a student has a good background in calculus, linear algebra, and differential equations, they have a big advantage on those courses.”

Generally, there exist topics within introductory courses that are difficult for students to generalize. In calculus, a minuscule value like epsilon may be difficult to appreciate since it doesn’t always appear to help solve the problem. In group theory, mind-bending ideas like adding fruits (operations on quotient groups) may be hard to grasp and even harder to apply. Since higher-level courses require students to use concepts from courses like calculus, linear algebra, and differential equations, more time should be spent ensuring that students thoroughly understand the material.

Challenges for students 2: Not seeing the connections between mathematics concepts

Mathematics is a discipline with diverse approaches to problem-solving. On one end, there are mathematical ideas related to shapes, such as geometry or topology. On the other, there are mathematical ideas based on counting, such as combinatorics and number theory. Even more, there are continuous mathematics such as calculus. Despite their apparent differences, different branches of mathematics can be used together to solve interesting problems. As an example, Andrew Stuart, a mathematics professor at CalTech, has used the Bayes theorem, the underlying theorem of probability, to solve partial differential equations (Stuart, 2015). At the undergraduate level, students use algebra and geometry to solve integrals in calculus and calculus to derive density functions in probability.

One systemic problem cited by multiple professors was that students did not see the connection between topics in different mathematics classes. “Mathematics is complicated for

some people, but it is complicated because they do not see the geometrical idea behind. They just see calculations,” says one professor. Another professor adds, “for a lot of mathematics, and certainly, in the freshman calculus sequence, students are used to seeing a new concept that lasts one lecture long. And then they eventually get to the point where, in Calculus 3, the new concept actually requires like 5 of these other lectures from 5 different places in calculus to be synthesized together and come up with a new thing. We think that’s where students find the most challenge; when they have to pull different bits of information from their past and synthesize it in a way that’s new and different.” In reference to higher-level classes, one professor said, “The theoretical concepts are sometimes hard to grasp, and then we pull on things from some linear algebra high-level concepts, and then we pull on lots of different things to kind of prove what we need to.”

One cause may be that too much or too little time passes between when students take mathematics courses. As an example, WPI segments the calculus material into four courses. Since these courses are offered multiple times throughout the academic year, students may opt to take all calculus classes in their freshman year or spread them across multiple years. Either approach has its drawbacks. By taking all classes in quick succession, students may not have enough time to reflect and process new information. Additionally, it may not be possible for students to schedule courses in this way due to other course requirements or off-campus projects. By spreading courses out, months or years may pass where students are not taking any mathematics at all and they might forget specific details. This problem may be exacerbated between mathematics courses with different names, such as calculus and differential equations. Students who take these two courses with years in between might not remember how to use the Taylor series approximation they learned in calculus and will have to re-learn it to solve a differential equation. Either way, students are likely to forget some material from class and will have to revisit topics multiple times before they understand it well.

Challenges for students 3: Trusting formulas

One professor reminisced about a formative moment in their undergraduate experience. “In my freshman year in Chemistry, we had to do a chemistry experiment. And the chemistry

experiment was to figure out Avogadro's number, which is $6 * 10^{23}$. So, everyone had their little recipe, the algorithm, and the experimental protocol to go through as a nice ten steps. So, I went through and did this, and I was incredibly off. Astronomically off. But then I stopped and looked through the protocol and said 'oh, step 3, I'll multiply this by the 8th root of 2', and then I was off by 5 percent. You can easily come up with results that have no meaning. There is no rationale for multiplying by that, no chemical reason. It's purely a way of adapting the data to getting what you want." A common pedagogical practice when teaching new concepts is to give students a formula and let them practice solving problems with it. These formulas often work only when specific conditions are met. As an example, consider the "integration by parts" formula taught in integral calculus. To use, one simply separates the integrand (the stuff inside the integral) into pieces and matches the pieces to variables in the formula. Students are given a set of homework problems in which they practice applying the integration by parts formula. Eventually, however, students learn that integration by parts only works if the function is differentiable, meaning the function is a smooth line. Students that try to apply the integration by parts formula to a discontinuous function may reach an impossible answer, such as $1 = 0$, or a not-impossible but incorrect answer. Thus, it is important for students to understand that formulas may have limitations and that reaching a not-impossible solution does not guarantee that it is the correct solution.

One professor expressed a similar sentiment that students should question the results of their formulas, especially when using computer programs. "I like to stress interpreting the solution as well as not trusting where things have come. Say you put something into some program. Do you believe the output? Anybody who knows a little about the mathematics and the method should look at this skeptically." Using a mathematics software such as Maple or Matlab can help students avoid messy and error-prone calculations thus allowing them more time to focus on high-level comprehension. However, it also distances students from the formulas they are using which makes it more likely that a formula is improperly used.

Addressing uncertainty and troubleshooting mathematical formulas requires a significant investment in time. Owing to the brevity of the 7-week term, spending time teaching these skills comes at the expense of learning a breadth of mathematics topics. Says one professor,

“In terms of ethical questions, those are good questions for students to see and to deal with. But whether there’s time in a mathematics class to address those issues, probably not.”

In summary, mathematics professors identified three problems common in undergraduate math: moving beyond the trivial in-class example, remembering concepts from introductory sequences and knowing how to use them, and developing a healthy skepticism of formulas. Two common issues in these problems are that (1) students do not have enough time to internalize new material and (2) students are not getting enough practice with problems that challenge their assumptions. This suggests that students would benefit from additional time in class spent on practical application. We hypothesize that working with data can efficiently help students deepen their understanding of concepts. The following sections detail the professors’ thoughts as to applying data in their own courses as a way to address these challenges.

Using data to visualize theoretical concepts can help students get an intuitive understanding.

When asked how data can help address some challenges in class, some professors interpreted data to mean pictures explaining some topic. According to one professor, “Data could be a collection of pictures, or visualizations. Basically, data is about examples and counterexamples.” Seeing these pictures could help students think about concepts visually and supplement the theory they learned in lecture. This would address Problem 2, remembering and understanding concepts, from above. Seeing counterexamples would help the student know when a method was not appropriate. This could help address Problem 3, developing a healthy skepticism. When asked to give a specific example of using data in this manner, the professor volunteered visualizing the connection between continuous and differentiable functions in calculus. “You want to know if the function is continuous, if it’s differentiable, [and] if [one] can switch $f(x(y))$ with $f(y(x))$.”

This professor felt the *instructor* was responsible for preparing data examples to be used in class. When asked about assigning students datasets on which to practice, the professor was opposed. “You spend too much time preparing to deal with a dataset and a few years

later everything is different. The time that you spend trying to see how to use that data, it changes all the time.” Even if the dataset didn’t change every few years, the rapid change of technology might require instructors to spend large amounts of time updating assignments to work with cutting-edge software.

Several professors agreed that data had the potential to improve classroom instruction. “[Data is] a really good pedagogical tool. I think the students walk away saying ‘ok, now I can see why if I do the math, something that looks intuitively fair actually isn’t.’” When well prepared, data visualizations in class can help students see the connections between what they’ve learned in other mathematics classes and a new topic. Adds another professor, “It would be nice to have a more clear relationship between the techniques that you use in statistics to find relationships and the linear algebra theory that backs up a lot of that stuff.”

In group theory, visualizations can be used to explain symmetries in a group. “An example is a sheet of paper folded into a square [that has] some markings on the corners. With a square you can rotate 90 degrees or 180 degrees, you can flip it over, it’s still a square.” One can repeat this experiment with a rectangle to visually understand how groups can be different. In this scenario, the sheet of paper is the data; the visual evidence that supports an otherwise abstract lesson.

Another professor agreed that data could be used in trivial or contrived examples to explain a concept, but he qualified that it depended on the specific topic. “It’s just a question of what level of sophistication you can do that. So, the benefit of calculus or differential equations is that they are at a level, mathematically, that you can pull up physical examples and have people be comfortable with it. If you go back to long division of polynomials, where would I get data to describe that algorithm? And that would be an example where data would not be helpful. Essentially, you’re using this as a tool to take one form and write it as another form mathematically. Those are probably concepts where having data may not be easy to determine.”

A third professor suggested that data visualizations could be useful if only there was more time in class. “I think visualization is important. I made an effort my first few years to show lots of movies and things, but then I realized that takes five minutes away from my class, and then I don’t get to finish what I need to teach.” The fact that there’s a limited

amount of time to teach and that data takes away from important instruction is a recurring theme. This professor admitted that data could be useful, but that it should be judiciously used so as to not detract from the lesson.

If you interpret data to mean visualizations that serve as evidence for theoretical topics, then it can be useful. Regarding the three problems identified above, using data in this way can help students visually understand concepts and show the limitations in intuitive thinking. However, this definition of data does not address the first problem: moving past the in-class example. Data can, and maybe should, be used in more ways than just visualization. Data, such as Big Data from real sources, can be used in projects to give students the opportunity to apply new concepts in a non-trivial way.

Using Big Data can help students get a deeper understanding of mathematics concepts.

When asked to consider alternative definitions of data for use in the classroom, some professors came to the topic of Big Data¹. Some professors were open to the idea of using Big Data in their class and gave examples of data-themed projects to use in their courses. We found that these examples fit into three distinct categories: real-world applications, student-centered data, and student-generated data.

First, there were the projects that showed how methods in class directly applied to a “real-world” application and could be used to motivate concepts for students within a particular field of study. One example from a linear algebra professor was for students to process data from a Magnetic Resonance Imaging (MRI) scan. By simply applying a Fourier transform to the data and plotting the results, students can transform a matrix of 1’s and 0’s into a black-and-white projection of a human skull. Another example was a past project from a numerical methods class. The professor invited industry professionals to visit the class and give the students raw data. “The students had to come up with a mathematical model and a representation of that data, or a way of describing how that data works,” says the professor. These types of projects would likely work well in an upper-level course where students are

¹One professor distinguished big data from the data visualizations mentioned above by calling it “Randy Big Data”, a reference to a well-known data science professor at WPI.

more likely to be in the same field of study, but they are valuable in that they provide an accurate view of real data applications.

Next, there were projects that used data about students themselves, such as gathered from entities on WPI's campus. To give students practice with Markov Chains, one professor of Probabilistic Methods in Operations Research created a 95,000 line file about students who take mathematics at WPI. "Students find it an interesting application since the data is about themselves and the campus environment," says the professor. In future iterations of the project, the professor hopes to gather data from the help desk or the Career Development Center. He is even trying to gather data from the cashiers in the Campus Center so that students can practice queuing theory. The professor believes the challenge is not in using data but finding an relevant dataset to engage students. "The ideas are standard but the question is how to implement it," he says. The value in this type of project is that it inherently appeals to all students and can be used in a class where students are from different fields of study.

Third, there were projects that had students generate their own data and, by so doing, gather evidence to support an otherwise unintuitive or abstract concept. One such example is a Probability game in which two people take turns flipping a coin and the first one to get heads wins. One might think that the game is fair and that each player has a 50% chance of winning. Actually, the game favors the first player such that they are $2/3$ more likely to win. One professor had students play this game for several rounds and collect data on their flips. They then pooled their data across the class and approximated the likelihood that the first player would win. After this experiment, the professor showed how adding all the possible outcomes together created a geometric series that summed to $2/3$.

These preceding examples show the diversity in ways that data can be used to aid in learning. As demonstrated in the examples, using Big Data can do more than just solidify understanding of classroom material. Using real data such as from an MRI scan can show students what data analysis looks like in the real world. When students generate their own data, like with the coin flipping game, or analyze data related to them, like queuing patterns at the campus center, they might be more motivated to learn. However, using data in a class is not simple. In the next section, professors identified computation and programming as necessary skills to use data effectively.

Using data in class requires programming.

One insight from mathematics professors was that computational skills were necessary to use data. This often meant that professors felt that using data in class was infeasible since it would require them to teach a programming language in addition to the standard topics. However, some professors were open to the idea, justifying it by saying that mathematics in the modern era is necessarily tied to programming.

For professors of subjects that haven't changed in centuries, mathematics programming seems to evolve too quickly. "I'm not so fond about those calculus with Matlab [labs], because every year it changes. You have to keep learning the new technology and sometimes it's good in certain senses but not in other senses. The teacher has to spend time learning how to use those tools. It is, in some sense, not related to math. It's related to how to use the manual, and this might change every 2-3 years," says one professor. The reality is that professors don't have unlimited time to prepare their curriculum, and time spent preparing to use a programming language takes away from time spent preparing lectures.

One professor considered the feasibility of assigning a big data project to students in one of the introductory sequences. He felt that the overhead cost of setting up the project was directly proportional to the number of students. "The challenge is to scale examples up to a class of 150-200 students and have every student have that same experience," adds one professor. "You'd have to find a data example that somehow clarifies how that tool works, and it doesn't tell you about the example, it tells you about the tool." The amount of technical overhead that comes along with mathematics programming may make it inaccessible in a large class such as this.

Programming, mathematical or otherwise, is a topic more commonly associated with computer science. In an introductory computer science course, students learn a single programming language and use that language to learn the basic principles of computer science. The introductory computer science course at WPI, *CS 1101*, teaches a language called Racket, but the purpose of the class is to teach data structures, functions, and how to maintain a consistent programming style. In this course, students learn how to write functions using numbers, strings (words made up of 'abc' characters), Boolean operators (greater than,

less than, equal to), if/else syntax (if this is true, do this, if not, do something else), and recursion all using Racket. The topics covered in the course are not specific to Racket; all programming languages support at least some of these basic principles. Thus, students who take this course learn universal computer science principles which carry over into all other programming languages.

One additional skill students learn in this and other computer science courses is how to debug code. Mistakes and typos are inevitable, and the error messages from faulty code can be cryptic. Students in a course like CS 1101 become familiar with these error messages and become proficient at tracking down and fixing errors.

A mathematics programming language is somewhat different from a traditional programming language such as Racket, and this change may necessitate a different teaching style. In a mathematics language such as Maple, the syntax is not centered around writing functions but instead writing scripts that use existing functions to solve a mathematics equation. In the calculus sequence, students use functions like *subs* or *eval*, which takes an expression $(x + 3)$ and a variable $(x = 2)$ and solves for the variable (out: 5). While this style of language can be used extensively without writing one's own functions, students will eventually need to learn how to write their own functions when they get to problems that require them to perform a specialized action. When this occurs, students who have taken a computer science course are better prepared because they understand how functions work, have a good idea of what the function is capable of doing, and are not afraid of the error messages.

Despite the challenges that come with teaching programming, one professor felt it was a necessary burden. "I don't know how to manipulate data or do something fun with it without having a computer do it," says the professor. "When I teach senior-level grad courses, I try to get everyone an account on the clusters on campus and have them play around with matrices that are actually a million by a million, or use a GPU. But the fact that I even advertise that in the course [description] then scares away 90% of the population from taking it. So, getting the data is not hard. It's just how to use it without overwhelming the students."

Overall, professors were open to the idea of using data in their classes, but they were averse to assigning projects that required a background in programming. One question that arises is whether mathematics as a field is likely to remain tied to computer programming.

If so, it may be necessary for mathematics to assume a formal responsibility to teach programming to all its students. If not, students may be better served taking an introductory computer science class or a programming class within their field of study.

Linear algebra is a good place to introduce students to data education.

Several professors brought up the idea that big data can be used in linear algebra. Linear algebra is the branch of mathematics surrounding linear equations, equations where the variables don't have exponents, such as $4x + 3y + 2z = 10$. One might recall solving for systems of linear equations in algebra; given $2x = y$ and $y = 4$, can you solve for x ? In the undergraduate linear algebra sequence, students rewrite these systems of linear equations in a matrix, a table where each row represents an equation and each column represents the coefficient of a variable. Figure 7 shows how to rewrite these particular equations in matrix form. Students can then use the properties of matrices to solve the system of linear equations.

$$\begin{array}{l}
 4x + 3y + 2z = 10 \\
 2x = y \\
 y = 4
 \end{array}
 \longrightarrow
 \begin{array}{c}
 _x + _y + _z = _ \\
 \begin{array}{|c|c|c|c|}
 \hline
 4 & 3 & 2 & 10 \\
 \hline
 2 & -1 & 0 & 0 \\
 \hline
 0 & 1 & 0 & 4 \\
 \hline
 \end{array}
 \end{array}$$

Figure 7: Using a matrix to encode coefficients for a series of linear equations. The three equations on the left are rewritten to fit into the coefficient matrix on the right. From there, matrix operations are used to solve for the values of x , y , and z .

In addition to viewing a matrix as a way of organizing equations, one can imagine each row of a matrix as coordinates for some multidimensional point. Figure 8 shows a simple example of this. Students can plot these points to visualize matrix operations such as row reduction and can intuit ideas like matrix subspaces, rotations, and reflections. “I think linear algebra totally loses the fact that basically we envision data as matrices. Everything is a matrix! Everything! I think the tendency is to get lost in the theory. Even if we mention [application] at the beginning, I think a lot of times we don't do a good job of finishing [a

lecture] and saying again ‘this is why this is important. Here’s a little snapshot video of something that explains why this is important’.”

Quantity and types of pets in US households

obs	# cats	# fish
1	0	10
2	1	9
3	1	7
4	2	3
5	2	2
6	3	0
7	4	0
...		

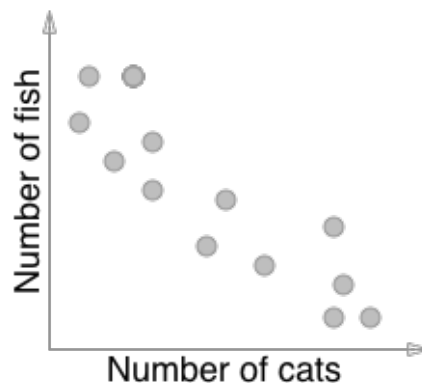


Figure 8: Plotting values from a matrix. Each row in the matrix represents one point in the graph.

Summary of interviewing mathematics professors.

Mathematics professors at WPI identified two objectives for the mathematics program: teaching students how to approach and solve problems and giving them a working set of formulas and techniques which can be applied to solve problems. These can be thought of separately as the theoretical and applied objectives of the department.

Data can be used in undergraduate classes to help visualize new concepts and strengthen previous understanding. Defined as visualizations, data can elucidate difficult theoretical concepts like epsilon in an integral or symmetries in a group. Defined as Big Data, data can show how linear transformations are used in MRI scans or why a coin flipping game is not fair. Ultimately, working with data requires more time in the classroom, but the additional understanding brought about by data are likely to significantly improve students experience elsewhere.

Programming experience seems necessary to use data in the classroom. It is unclear whether teaching programming skills should be the responsibility of the Mathematical Sciences department, the Computer Science department, or the department in charge of the student’s major field of study. One benefit of offering programming as a mathematics or

computer science course is that students from all fields of study can take the same course and have a standard shared base of knowledge upon which future data projects can build. However, specific programming languages or tools used in each field may mean that each department should be responsible for preparing its students to work with data in the way they will use it in their future careers.

Finally, several professors identified a relationship between the undergraduate linear algebra sequence and big data. Linear algebra deals with matrices and systems of linear equations, both of which are useful for representing datasets. Linear algebra may be a reasonable place to introduce students to working with data because the matrix structure is already frequently used for working with data.

One common theme we identified while interviewing mathematics professors was a lack of time for teaching data in addition to regular course material. Because these professors value a theoretical understanding of mathematics in addition to practical application, they are reluctant to modify their curricula in a way that reduces the rigor of the mathematics. We will find in the following sections that professors of non-mathematics classes highly value practical application over theoretical understanding, and so there is disagreement between these two groups as to what students should be learning in mathematics courses.

4.2 Undergraduate mathematics according to experts in mathematics education

In the previous section, one of our findings was that the linear algebra course seemed like a reasonable place to introduce Big Data. Linear algebra is a prerequisite for many of the engineering and science concentrations at WPI because it teaches students how to solve systems of linear equations and how to apply linear transformations. We were surprised by this connection between linear algebra and Big Data, so we reached out to experts in undergraduate linear algebra education who could corroborate this connection.

We sent interview emails to two such experts: Stephen Boyd and Gilbert Strang. Professor Boyd is a professor at Stanford University and is the author of the recently published

textbook *Introduction to Applied Linear Algebra* (Boyd & Vandenberghe, 2018). Professor Strang is also well known in linear algebra pedagogy. In addition to currently teaching linear algebra courses at Massachusetts Institute of Technology (MIT), Professor Strang is currently working on a new textbook, *Linear Algebra and Learning from Data*².

Both Professors Boyd and Strang agreed that data fit into the purpose of undergraduate mathematics and that courses could easily be changed to accommodate data. At MIT, Professor Strang notes that the Mathematical Sciences department has already started offering new statistics courses that include data. He also notes, however, that before these courses were available, students had to take courses within the Computer Science and Electrical Engineering departments to learn necessary data analytic skills. He speculates that other programs that are just beginning to adopt data into their curriculum will have a similar challenge.

Professor Boyd felt certain that using data in mathematics classes would help prepare students for their future careers. He wrote, “The standard curriculum is based on 19th-century physics; it covers the mathematics of the physical world (mechanics, fluids, electromagnetics). It is very important and has been used for many applications, but more modern introductory mathematics should focus on linear algebra, which is more appropriate for applications in the information world, such as fitting models to data, making predictions, and so on. Many applications are in the social sciences, economics, and other areas.”

Both Professors Strang and Boyd also supported the idea that linear algebra was an appropriate class to teach data. “It is the perfect vehicle for this,” says Professor Boyd.

From these interviews, it appears that the mathematics program is very capable of adapting to teach modern data analysis techniques. Professor Strang briefly mentioned that he felt programming was necessary to use data, but he said it in the context that the instructor needed to be able to program to show examples using data in class. Both experts felt that linear algebra was a suitable place to teach data, and both of their textbooks give practical instruction on how this can be done.

In talking with mathematics professors at WPI and notable experts outside of WPI, we found that there was a high potential for the use of data in undergraduate mathematics.

²This book is not yet published, but sections are freely available at math.mit.edu/learningfromdata

Teaching practical data applications is well aligned with the Mathematical Sciences department's objective to give students a useful background in mathematics for their major-specific courses and in their future careers. Multiple professors at WPI concluded that the linear algebra course was good place to teach data, and this was emphatically supported by experts in linear algebra education.

Some professors felt that, despite good intentions, using data in class was infeasible. Professors had to devote additional time preparing data examples and, as necessary, learning new mathematics software. Should the professor attempt to have students work with the data themselves, additional time would be spent acquainting students with programming. Most of the introductory mathematics classes are taught in lecture halls and serve hundreds of students at a time. This makes it even more difficult for professors to ensure that all students are getting the same data experience. For these reasons, the professors felt that there was not enough time to teach data and also teach the other course material.

Foreshadowing the conclusion, we believe that data is still an important part of a modern mathematics program. For these reasons, we believe students should be taught computer programming to the extent that they will need it in their field of study. Because of time and resource constraints on the current mathematics curriculum, we suggest that programming and data analysis should be taught in a separate mathematics course and should be co-directed with other departments at WPI. This would expose students to necessary modern skills without impinging on the other objectives of the Mathematical Sciences department.

4.3 Data education requirements according to professors outside of Mathematical Sciences

The second goal of our project was to determine the data education needs of all students at WPI. By defining these needs, we further understand what needs to be done to accommodate all students. After interviewing professors within the field of mathematics, we interviewed professors outside of mathematics who taught courses requiring data analysis. Our objective was to capture these professor's impression of the Mathematical Sciences department and

whether the Mathematical Sciences department was doing a sufficient job preparing students for their courses.

Of all undergraduate courses across all majors, we identified 48 courses that mention “data” or “data analysis” in their course descriptions (not including courses that only mention “databases” or “data transmission”). Professors of these classes have insight as to the data education requirements of their respective fields and can speak as to the efficacy of the current WPI mathematics curriculum.

We wanted to capture as many perspectives as possible from this population, so we chose to email questionnaires to these professors for them to answer at their own time. To increase the likelihood that our emails were answered, we initially sent emails only to professors who had taught one of these data classes within the past year. The number of questions we asked was kept at three to increase the likelihood that a professor would have time to respond. The third question in our email was to help us find other professors who would be willing to talk to us about data in their classes. Of the 8 professors whom we initially emailed, two responded.

Experimental Analysis in Chemical Engineering

The first response was from Professor Abu-Lail, an assistant teaching professor in the Chemical Engineering department. Professor Abu-Lail had recently taught *CHE 4401: Unit Operations in Chemical Engineering I*. This is a laboratory-based course in which students solve problems related to large-scale chemical processing phenomena. Students plan and execute their own experiments and then report their findings in written and oral presentations.

Professor Abu-Lail felt that students in her course did well with data, but they had some room for improvement. In terms of presenting their data, she felt they could improve in finding the best way to plot the results. With regards to data analysis, she felt they could improve in their error analysis and in attributing the right amount of meaning to the data. She wished the students had more experience using significant figures.

Experimental Analysis in Aerospace/Mechanical Engineering

The second response was from Professor Scarpino who teaches *AE 3901/ME 3901: Engineering Experimentation*. This course, offered both in the Aerospace Engineering and Mechanical Engineering majors, relates to the collection and proper analysis of data from electronic instruments. In the laboratory part of the course, students work with real equipment and collect mechanical data such as heat transfer, force/torque, and motion/vibration and materials data such as temperature and pressure. The purpose of these labs is to acquaint students with standard industry tools to understand their capabilities and limitations.

Professor Scarpino was very interested in talking to us about this project as he was in the process of submitting a proposal to initiate discussion about a similar topic. From his perspective, students are ill-prepared to deal with industry data. When they get to his Engineering Experimentation course, students have limited to no experience with real equipment. Using equipment is not only about how to interface with the sensor, students must also know the sensitivity of the equipment, to how many decimal places is it accurate, and even how to collect data.

One of the labs in this course is the temperature lab. Students use LabView, a graphical programming language, to write a program that takes temperature sensor readings. The program runs in a while loop; a programming construct that allows for the same block of code to be run repeatedly until the user decides to break the loop. Every time the while loop runs, another measurement is taken from the thermometer. The first challenge is knowing how many temperature measurements are necessary to get an accurate reading. Not only the quantity, but students must also decide at what frequency to run the loop. The while loop can run quickly and sample hundreds of times in a second, or it can run slowly and only loop once a second. “How much data [points] can we really afford to take? Millions? Just one? Just two? – It matters because it affects uncertainty in data,” says Professor Scarpino.

One might think that rapidly collecting data results in higher accuracy, but this is not necessarily the case.

After students collect their data, they have to analyze it. The sensors are physical instruments and are subject to inconsistencies. Professor Scarpino says students are surprised

when they get results they did not expect. This is a result, he says, of a lack of practical training. “It’s very hard to create an undergraduate curriculum that’s all about theory and then go to the real world and realize it looks like all this,” he says. Professor Scarpino cited that 75% of students in his course would go into industry where they will work with real devices. Despite this majority, most students in his class do not have sufficient practical experience.

Summary of interviews with non-mathematics professors

One distinguishing quality of non-mathematics professors is that they feel it is their responsibility to prepare their students to work in Industry. Because of this, these professors are very open to using industry data in their classes. In summary, these two professors both said their students had problems understanding the limitations of data. When it comes from real sensors, data carries with it some amount of error and uncertainty. The process in which one collects data can contribute to this uncertainty as well; collecting more data is not necessarily better. Professor Abu-Lail also mentioned room for improvement in displaying data in plots.

4.4 Data education responsibilities according to the head of the Mathematical Sciences department

After hearing the perspectives of professors both within and outside of the Mathematical Sciences department, we realized there was a bit of a conflict between the two. Mathematics professors highly regarded the theory of mathematics. They viewed data as a means of accomplishing their goal of teaching better mathematics. This perspective is inherent in their job as mathematics professors. However, their perspectives regarding the purpose of undergraduate mathematics are in contrast with those of professors outside of the Mathematical Sciences department. These non-mathematics professors must prepare students to be capable in industry careers and, as such, value the applied aspects of mathematics more than the theoretical. Professors in this capacity are well aware of their domain’s data and are

actively working to incorporate it into the curriculum. From their perspective, mathematics is the means by which one processes data, and thus theory takes a back seat to application.

The question that arises is which population should be principally responsible for teaching data education to undergraduates. Indeed, both mathematics and non-mathematics professors already teach some aspects of a data education, but we expect one population to assume responsibility for setting data education requirements and ensuring that they are met. If we assume that all students at WPI require some data education, then mathematics professors are well poised to teach it. Professors of introductory mathematics courses interact with almost all students at WPI, so it makes sense to adapt this curriculum to teaching data. However, we have shown in our interviews with mathematics and non-mathematics professors that the data education requirements are significantly different in each domain. Requiring mathematics professors to prepare data examples for each type of student seems an undue burden. Certainly, each department would know best what their industry requires of their students and could better prepare these examples. However, assigning data responsibilities to non-mathematics professors would practically mean assigning such responsibilities to every other department at WPI. It is almost guaranteed that some data education needs are similar across departments, and there would likely be some redundancy when preparing the data curriculum.

One person at WPI is especially knowledgeable about this problem; the director of the Mathematical Sciences department. As department head for the past 5 years, Professor Luca Capogna has overseen the development and change of the mathematics curriculum. If the responsibility for teaching a data education was given to mathematics professors, Professor Capogna would likely be the person to orchestrate the change in curriculum. Even if the responsibility was given to non-mathematics professors, Professor Capogna would serve as the liaison between Mathematical Sciences and these other departments ensuring that a data education was taught in a logical and continuous manner.

We approached Professor Capogna for an in-person interview after talking with professors inside and outside of the Mathematical Sciences department. We were interested in his opinions regarding the differing viewpoints of these professors and whether or not he thought a data education fell within the responsibilities of the Mathematical Sciences department.

Similar to our interviews with mathematics professors, we asked him the mission of the Mathematical Sciences department and what he thought of using data in math. However, as a result of our prior interviews, we also asked him about the specific challenges identified by professors. The following sections follow the major talking points that came up during our interview.

Mission of the Mathematical Sciences department

Professor Capogna hinted that the mission of the Mathematical Sciences department was not yet completely understood. This is because the department serves multiple populations within the field of mathematics itself. For example, there are two mathematics programs offered at WPI: Actuarial Mathematics and Mathematical Sciences. Within the Mathematical Sciences degree, there is also specializations in statistics. In addition to these, there are also interdisciplinary programs within math, such as the Data Science minor and the Bioinformatics and Computational Biology program. Because of the diverse nature of the Mathematical Sciences department, Professor Capogna felt it was difficult to define a unifying mission statement for the Mathematical Sciences department.

Professor Capogna pointed out, however, that the lower level mathematics courses were considered separate from the overall mission of the Mathematical Sciences department. He referred to these courses as “service courses,” meaning that they provide a service to other departments on campus. In these courses, however, the Mathematical Sciences department does try to tailor the material to appeal to the students in the class. “This is an ongoing process that changes from year to year, from professor to professor, and it changes also depending on the needs of the students. If we have an advanced course in Complex Analysis and there are, maybe, 80% who are students in Mechanical Engineering, then the focus is going to be different than if the 80% of the students were in Complex Analysis. We have to retain some of this flexibility, just to make things work,” says Professor Capogna.

Concern of non-mathematics professor: lack of practical application

We asked Professor Capogna about the Mechanical Engineering professor's concern that the students coming into his higher level lab-based courses did not have the right background. Professor Capogna stated that it was not a lack of practical mathematics education that led to this problem, but more that too much time had passed since the students took the prerequisite classes. "Most of the concerns I get from colleagues is that the students do not know the mathematics that is needed because they took calculus 3 years before and they have completely forgotten everything. And I don't know how to fix that kind of problem."

Current revisions to the mathematics curriculum

Professor Capogna did feel that the Mathematical Sciences department was actively trying to improve the curriculum. He cited two main efforts on which the department was currently working. The first was to increase the presence of computation in all mathematics courses. By doing this, students would accumulate computational skills over their entire undergraduate experience and would have the necessary analytic tools by the time they graduate. The same effort would also add a data analysis component to more introductory courses to achieve similar results to the data program at Rensselaer Polytechnic Institute. He even suggested that the Mathematical Sciences department was planning new data science courses.

The second effort was to increase the level of mathematical rigor given in mathematics courses, especially the advanced courses. Professor Capogna felt that the rigor of these classes had dropped because of attempts to accommodate students from mixed backgrounds. Within the Mathematical Sciences department, he said, some professors felt there should be less compromise in the future. "Students will be asked to step up if they want to take a course because otherwise, we are not serving our purpose within the university. Mathematics is very powerful when it is rigorous. If we do not teach that kind of rigor to our students, they will miss it when they go out in the job market. And so we believe it will help them."

Summary of interview with mathematics director

Communication and transition of mathematics to other majors is a real problem that the department is actively trying to fix. Professor Capogna feels the department can change with response to data, and changing the curriculum to teach data would be similar to previous changes to teach computation in math. In response to concerns that students do not get enough practical application, it is unsure how to fix this. The mathematics professors, in general, agree that the curriculum should be more rigorous, which directly conflicts with the non-mathematics professor's desire for mathematics to be more practical.

5 Conclusions

The first and most overarching conclusion of this project is that data can indeed be used to improve the mathematics curriculum. Mathematics professors at WPI gave numerous examples of data projects that can be used to explain abstract concepts or show real applications of mathematics concepts. Other professors at WPI felt that the Mathematical Sciences department should do more to give students chances to practice with real data.

The next conclusion is that the introductory mathematics sequence is already too packed with material to add additional data exercises. Professors in these courses are already concerned with the amount of material they need to cover in a short period, and using data comes with the extra burden of programming. Not only does this take away from material taught in class, but professors expressed a lack of desire to spend time preparing new data examples and teaching new programming techniques. Secondary to this, Professor Capogna cited that some mathematics professors felt mathematics courses should be becoming more rigorous. For this to occur, courses would have less time to spend teaching additional topics such as computer programming.

From our interviews with the professors in Chemical Engineering and Mechanical Engineering, we believe the data education needs of different student populations are diverse. Thus, we hypothesize that a single “data” course would be insufficient to meet the data needs of all students on campus. One limitation of this report is that we were unable to contact more professors outside of the Mathematical Sciences department. These professors would likely have valuable insights as to the requirements of their industry, and their opinions would likely impact our conclusion regarding a “data education” course.

We recommend that there should exist a data education capstone course situated between the introductory mathematics sequence and higher-level major-specific data analysis courses. This course would build on the material taught in otherwise unrelated introductory courses and, hopefully, would be taught in a sequence that would help students see the connections between mathematics topics. This course would focus on practicing data applications.

The course would not assume any programming experience, though basic familiarity with a computer could be ascertained from knowledge of the computational skills covered in the introductory mathematics courses. Thus, the course would be additionally responsible for teaching a programming language that is likely to be used in the student's chosen career. We suggest that students work with a dataset related to their major and practice applying skills they have acquired before. More so than familiarizing students with the specific hardware of their major, this course is meant to practice a data analysis approach and help students become comfortable working with data and using a computer to perform mathematics.

More work needs to be done to determine the number and types of courses necessary to meet each student population. We expect there to be separate data courses for the following departments: Mathematical Sciences; Computer Science, Electrical Engineering, and Robotics; Biology, Biomedical Engineering, and Chemical Engineering; Physics and Mechanical Engineering; and Humanities and Arts. The Mathematical Sciences department should be responsible for the practicums and for determining how to group student populations. However, all of the departments that need data education as part of their program need to be actively involved in developing each practicum. These departments and the Mathematical Sciences department can hopefully come to some consensus regarding the programming language to use and a relevant dataset.

References

- All about the human genome project (hgp)*. (2015). Retrieved from <https://www.genome.gov/10001772/All-About-The--Human-Genome-Project-HGP>
- Boyd, S., & Vandenberghe, L. (2018). *Introduction to applied linear algebra*. Cambridge University Press.
- Chronology of computation of pi*. (2017, December). Retrieved 2018-01-23, from https://en.wikipedia.org/w/index.php?title=Chronology_of_computation_of_pi&oldid=814544554 (Page Version ID: 814544554)
- Clipart of sperm whale*. (2017). Retrieved from http://www.clipartpanda.com/clipart_images/sperm-whale-clip-art-6086516
- Computer and information research scientists*. (2018). Retrieved from <https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm>
- Cordell, R. (2017). *Humanities data analysis, spring 2017*. Retrieved from <http://s17hda.ryancordell.org/schedule/>
- Diebold, F. X. (2012, September). *On the Origin(s) and Development of the Term "Big Data"*. University of Pennsylvania. Retrieved 2018-01-30, from <https://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf>
- Feisel, L. D., & Rosa, A. J. (2005, January). The Role of the Laboratory in Undergraduate Engineering Education. *Journal of Engineering Education*.
- The gdelt project*. (2018). Retrieved from <https://www.kaggle.com/gdelt/gdelt>
- Hardin, J. S., & Horton, N. J. (2017, October). Ensuring That Mathematics is Relevant in a World of Data Science. *Notices of the AMS*, 64(9), 986–990.
- Horton, N. J. (2015, May). Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward. *The American Statistician*, 69(2),

- 138–145. Retrieved 2018-01-18, from <https://doi.org/10.1080/00031305.2015.1032435>
doi: 10.1080/00031305.2015.1032435
- joannefox. (2006, August). *THE HUMAN GENOME PROJECT: THE IMPACT OF GENOME SEQUENCING TECHNOLOGY ON HUMAN HEALTH*. Retrieved 2018-01-23, from <https://www.scq.ubc.ca/the-human-genome-project-the-impact-of-genome-sequencing-technology-on-human-health/>
- Kaggle datasets*. (2018). Retrieved from <https://www.kaggle.com/datasets>
- Lai, Y. (2009). *Teaching Undergraduates Mathematics* (Tech. Rep.). Berkeley, CA: Mathematical Sciences Research Institute. Retrieved 2017-09-05, from <http://library.msri.org/cime/CIME-v5-TUM-booklet.pdf>
- Lohr, S. (2013, February). The Origins of 'Big Data': An Etymological Detective Story. *The New York Times*.
- Martialay, M. (2018). *Rensselaer introduces first in the nation “data dexterity” requirement for all undergraduate students*. Retrieved from <https://news.rpi.edu/content/2018/03/22/rensselaer-introduces-first-nation-OT1-textquotedbleftdata-dexterityOT1-textquotedbright-requirement-all-undergraduate>
- Mashey, J. R. (1998, April). *Big Data... and the Next Wave of InfraStress*. Silicon Graphics Inc..
- Mathematical sciences*. (2018). Retrieved from <https://www.wpi.edu/academics/calendar-courses/course-descriptions/mathematical-sciences>
- Matp 4400 - introduction to data mathematics*. (2017). Retrieved from http://catalog.rpi.edu/preview_course_nopop.php?catoid=15&coid=28411
- Napper, B. (1999). *The Manchester Mark I*. Retrieved 2018-01-23, from <https://web.archive.org/web/20140209155638/http://www.computer50.org/mark1/MM1.html>
- Nguyen, T. (2011, February). *What is the world’s data storage capacity?* Retrieved 2018-01-23, from <http://www.zdnet.com/article/what-is-the-worlds-data-storage-capacity/>
- Petersen, R. (2016, December). *37 Big Data Case Studies with Big Results - Schaefer*

- Marketing Solutions: We Help Businesses {grow}*. Retrieved 2018-01-30, from <https://www.businessesgrow.com/2016/12/06/big-data-case-studies/>
- Stuart, A. (2015, Dec). Blending mathematical models and data: Algorithms, analysis and applications.. Retrieved from http://www.turing-gateway.cam.ac.uk/sites/default/files/asset/doc/1606/Andrew%20Stuart%20Cam_PDE.pdf
- Taycher, L. (2010, Aug). *Books of the world, stand up and be counted! all 129,864,880 of you*. Retrieved from <http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html>
- Undergraduate Catalog 2017-18*. (2017). Worcester Polytechnic Institute. Retrieved from https://www.wpi.edu/sites/default/files/docs/Academic-Resources/Academic-Catalogs/WPI_UGCat17-18.pdf
- Young, H. (1962). *Statistical Treatment of Experimental Data*. USA: McGraw-Hill Book Company Inc.

Appendix A: IRB Approval

WORCESTER POLYTECHNIC INSTITUTE

Worcester Polytechnic Institute IRB# 1
HHS IRB # 00007374

28 November 2017
File:18-0129

Re: IRB Application for Exemption File:18-0129 "Data in Undergraduate Mathematics"

Dear Prof. Paffenroth,

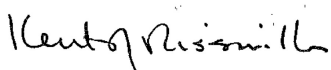
The WPI Institutional Review Committee (IRB) has reviewed the materials submitted in regards to the above mentioned study and has determined that this research is exempt from further IRB review and supervision under 45 CFR 46.101(b): (2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

This exemption covers any research and data collected under your protocol from 28 November 2017 until 27 November 2018, unless terminated sooner (in writing) by yourself or the WPI IRB. Amendments or changes to the research that might alter this specific exemption must be submitted to the WPI IRB for review and may require a full IRB application in order for the research to continue.

Please contact the undersigned if you have any questions about the terms of this exemption.

Thank you for your cooperation with the WPI IRB.

Sincerely,



Kent Rissmiller
WPI IRB Chair

