Introductory Data Science: A Blueprint to Navigate Curricular, Pedagogical, and Computational Challenges

Elijah Meyer
Department of Statistics, Duke University
and
Mine Çetinkaya-Rundel
Department of Statistics, Duke University

January 24, 2023

Abstract

The text of your abstract. 200 or fewer words.

Keywords: Data Science, Curriculum, Pedagogy

1 Introduction

The increasing volume of enrollment of data science students (Redmond 2022) requires that statistics and data science educators commit to developing modern curriculum in order to help students be successful. This rapid growth is largely motivated by industries evolving towards data-driven decision making, requiring that data science graduate possess the tools required to make those kinds of decisions. An estimated 11.5 million new data science jobs are projected to be created by 2026, while employment of data scientists is projected to grow by 36 percent from 2021 to 2031 (U.S. Bureau of Labor Statistics, 2023). Despite the demand, colleges are still struggling with what a modern data science curriculum should look like (Schwab-McCoy et al. 2021). To this point, much more thought, work, and discussions need to take place there before a consensus is reached on what should a modern data science curricula, to best prepare students, should be.

Some recent research on creating and examining data science curricula has been conducted. For example, in 2014, research at the College of Charleston, South Carolina, discuss their design and implementation of a four-year data science undergraduate program. They detail their success and challenges of implementation across a ten-year year span, ultimately concluding that a data science degree program can be successfully mounted at the undergraduate level through the requirement of courses across data science, computer science, and mathematics (Anderson et al. 2014). At the course level, Asamoah, Doran, & Schiller presented their experience in designing, developing and delivering an interdisciplinary introduction to data Science course across multiple colleges for an "Introudction to Data Science" course. The three course components including *Data Hacking*, *Math & Statistics Foundations*, and *Predictive Analytics*. Surveys ultimately concluded that students were satisfied with the course content (Asamoah et al. 2015).

However, the majority of context across the landscape of data science curricula largely focuses on how to model data (Donoho 2017). This, in combination when a lack of consensus on what constitutes a data science curricula, and even less research on how to develop modern data science course presents a need for a blueprint to design and implement and modernized introduction to data science course. In this paper, we attempt to add address this need by describing describing the creation and implementation of a modernized data science curricula for an introductory data science course at Duke University.

This course is designed for students with little to no statistics, data science, or coding experience, a common hurdle identified by faculty when trying to implement a data science course (Schwab-McCoy et al. 2021). By the end of this course, students are expected and able to clean, investigate, and communicate with data in a reproducible manner while answering a targeted research question. Detailed learning objectives of this course include learning to explore, visualize, and analyze data in a reproducible and shareable manner through the use of R-studio and GitHub (R Core Team 2021, github 2020). Through these programs, students gain experience in data wrangling, exploratory data analysis, predictive modeling, and data visualization. These experiences are generated through the *Kaplan Way*. The Kaplan Way is a learning model "that combines a scientific, evidence-based design philosophy with a straightforward educational approach to learning" (Schweser 2023). This learning model posits a three-phase learning strategy: Prepare, Practice, and Perform. During each of these phases, students are equipped with the appropriate tools to acquire knowledge, be given an opportunity to apply what they know, and to demonstrate mastery of the tasks at hand.

In this paper, we examine current curriculum recommendations for undergraduate programs in data science and review current pedagogical recommendations on how to teach such courses. Next, we discuss the creation and implementation of curricular and pedagogical decisions made in designing the introductory data science course at Duke University. This includes detailing the implementation of the Kaplan Way learning model, to support a large class of students with a diverse background in statistics, data science, and coding experience. Within this format, we provide examples of and describe activities and assessments given both in and outside of class. We extend discussions and provide recommendations for implementing and integrating computing tools, such as R-studio and GitHub, through our experiences in our course. Lastly, we discuss challenges, and provide insight to help faculty wanting to adopt or adapt a course similar to introductory to data science at Duke University. The purpose of this paper is to continue the discussion, and present a modernized curricula for an introductory data science course at Duke University, and the pedagogical decisions to help best equip students with the data science skills necessary for future classes.

2 Data Science: A Review

Although the definition of data science is fluid, it can be generally defined the process transforming raw data into understanding, insight, and knowledge (Wickham & Grolemund, 2022). In practice, data scientists often describe their work as a means to "gain insights" or "extract meaning" from data (Hernán … , 2019). In the following sections, we describe the current curricular and pedagogical recommendations for designing a course in data science.

2.1 Curriculum

Drivien by the high demade for data science skillsets, there has been a call to create more well-rounded curricula that better unifies methodology in statistics, mathematics, computer science, and machine learning to prepare students to understand, analyze, and communicate with data. In 2017, Curriculum Guidelines for Undergraduate Programs in Data Science provided six major recommendations as to what practitioners of data science should be competent in: Computational and statistical thinking; Mathematical foundations; Model building and assessment; Algorithms and software foundation; Data curation; Knowledge transference—communication and responsibility (Veaux, et al., 2017). These guiding pillars offer an approach for students to develop into problem solvers who interact with, investigate, and make meaning with data. Other visions of data science include more broad divisions of class activity. This includes *greater data science.* Coined by Chambers and Clevland, the activities of *greater data science* are classified into six divisions: Data Gathering, Preparation, and Exploration; Data Representation and Transformation; Computing with Data; Data Modeling; Data Visualization and Presentation; and Science about Data Science (Donoh, 2017). They suggest that these activities can be used to guide and assess if data science programs are adequately addressing the entirety of the field.

Additionally, a task force, titled the Association for Computing Machinery (ACM) Education Council, explores and expands discipline-specific conversations around the field of data science (Danyluk & Leidig, 2021). This council acknowledges that data science curricula can be flexible, but suggests that data science curricula should include applications designed towards building skills in computing, statistics, machine learning and mathematics.

Among existing curricular recommendations, there are computational and technological recommendations for data science classrooms. This includes the use of GitHub to ensure the concept of reproducibility and incorporating quantitative programming environments (such as R), to better work with data (Donoho, 2017; Beckman et al., 2021). Among all recommendations, it is suggested to immerse students in real-world data and provide open-ended projects. This allows students “early exposure to and experience with the full data science cycle”, in a practical real-world context (Lui & Huang, 2017; CETINKAYA-RUNDEL, et al., 2022, pg. 3).

2.2 Pedagogical

Although not specific to data science, national guidelines have recommended the use of more student-centered teaching techniques, such as active learning, in statistics and mathematics classrooms, to better engage students in the learning process (Carver et al., 2016; Conference Board of Mathematical Sciences, 2016). The Guidelines for Assessment and Instruction in Statistics Education College Report further stresses the importance of active learning in the mathematical sciences, encouraging instructors to use active learning to foster and enhance students’ understanding and communication of topics taught (Carver et al., 2016). In general, research has shown that the adoption, integration, and implementation of active learning teaching techniques can help promote student learning, achievement, and confidence in science courses (e.g., Freeman et al., 2014). Teaching data science courses through the use of active learning can help better facilitate a more complete understanding of working with data.

As the volume of data continues to grow, the need for those who can make meaning of data are clear. Thus, it is critical that data science education provides programs that adequately prepare and train students in the field of data science. This article adds to the existing literature by describing a modernized introductory data science course and how its taught, at Duke University.

Other potential models

This is what we do …

Kaplan Way Discussion

3 The Course

Could make learning units a section, or could talk about them here.

This is the "what" - curricular

3.1 Technology

3.1.1 GitHub

– What is this

– Why are we using this

3.1.2 R + R-studio

– What is this

– Why are we using this

3.1.3 Slack

– Emphasis on communication

– How to set up channels

– What channels are created

4 Getting Started

4.0.1 Duke Container

– How instructor sets this up

– Students reserve container / do not need local download

4.0.2 Collecting Information

– How

– What

– What

GitHub username Slack email etc.

4.0.3 GitHub Organization

– Your class can collaborate on GitHub by using an organization account, which serves as a container for your AEs, labs, hws, and projects.

– Why do we use this

– What does this allow us to do

4.0.4 Teaching Assistants

– What are they?

– Roles / Assignments

– Training

5 Pedagogy

5.1 Lectures (Prepare)

Part of the class time will be lectures that introduce new concepts or review topics from the preparation videos. Lectures will not repeat everything in the videos, they will instead highlight important and known to be complex concepts and will be supplemented with live coding activities. You are expected to attend every lecture. Lectures will be recorded and made available to students with an excused absence upon request.

5.2 Application exercises (Practice)

The purpose of application exercises (AEs) are to create interactive live-coding experiences for students in courses using the R programming language. A majority of the in-class lectures will be dedicated to working on AEs. These exercises give students an opportunity to practice apply the statistical concepts and code introduced in previous readings, lectures, and recorded videos. AEs are due within three days of the corresponding lecture period, and graded based on completion.

5.2.1 Creation

Before class, instructors must create student repositories that correspond with the upcoming class AE. To do so, instructors must first create a template repository, in GitHub, that contains the class' content. We then use `gh_class` package in R to clone and grant access for student's individual AE repository.

AEs are created using Quarto, the next-generation version of R Markdown (cite). Features of Quarto provide multiple ways for students to write code and type written answer to questions on their own, in groups, or along with the instructor during lecture.

5.2.2 Implementation

5.3 Assessments (Perform)

There are three forms of assessment in this course

5.3.1 Homework

5.3.2 Labs

5.3.3 Project

Discussion

Anderson, P. E., Bowring, J. F., McCauley, R. A., Pothering, G. J. & Starr, C. W. (2014), ‘An undergraduate degree in data science: curriculum and a decade of implementation experience’, *Proceedings of the 45th ACM technical symposium on Computer science education* .

Asamoah, D., Doran, D. & Schiller, S. (2015), Teaching the foundations of data science: An interdisciplinary approach.

Donoho, D. (2017), ‘50 years of data science’, *Journal of Computational and Graphical Statistics* **26**(4), 745–766.

URL: *https://doi.org/10.1080/10618600.2017.1384734*

github (2020), ‘Github’.

URL: *https://github.com/*

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: *https://www.R-project.org/*

Redmond, F. (2022), With a rise in computing disciplines comes a greater choice of computing degrees in higher education, *in* 'Proceedings of the 22nd Koli Calling International Conference on Computing Education Research', Koli Calling '22, Association for Computing Machinery, New York, NY, USA.

URL: *https://doi.org/10.1145/3564721.3565946*

Schwab-McCoy, A., Baker, C. M. & Gasper, R. E. (2021), 'Data science in 2020: Computing, curricula, and challenges for the next 10 years', *Journal of Statistics and Data Science Education* **29**(sup1), S40–S50.

URL: *https://doi.org/10.1080/10691898.2020.1851159*

Schweser, K. (2023), 'Our philosophy'.

URL: *https://www.schweser.com/about-kaplan/philosophy*