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

February 23, 2023

Abstract

The text of your abstract. 200 or fewer words.

Keywords: Data Science, Curriculum, Pedagogy

1 Audience

Audience - new to teaching / teachers increasing course size (mention this as a draw in because this is happening). This could be a hook.

I think this needs to be further brought out in the beginning of the intro when formalling writing….

2 Introduction

Reader should want to continue reading by thinking "this is something that's interesting"

– **Why** we are writing this paper

There is a demand for data scientists. 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 (BLS 2022). **As demand increases, class sizes to best prepare students for these positions increases as well**. 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. Despite the demand, colleges are still struggling with what a modern data science curriculum should look like (Schwab-McCoy et al. 2021), and how it can be effectively taught to a large student audience. 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 should be.

What people have been investigating / what's recommended

Curricular

- College of Charleston, South Carolina (Anderson et al. 2014)
- Intro to Data Science implementation (Asamoah et al. 2015)
- Curriculum Guidelines for Undergraduate Programs in Data Science (De Veaux et al. 2017)
- 50 Years of Data Science (Donoho 2017*a*)
- Association for Computing Machinery (ACM) Education Council (Danyluk et al. 2021)

Pedagogical

- Computational and Tech recommendations in curricula (Donoho 2017*a*, Beckman et al. 2021)
- Active Learning / Other recommendations (Dogucu & Çetinkaya Rundel 2022, Çetinkaya Rundel et al. 2022) + GAISE (Carver, 2016); Conference Board of Mathematical Sciences, 2016).

But...

However, the majority of context across the landscape of data science curricula largely focuses on how to model data (Donoho 2017*b*). 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.

So

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.

The purpose of this paper is to offer valuable structure, while providing experiences from our perspective. It is of importance that we stress the amount of flexibility and strength in individuality in all aspects when creating, designing, and implementing an introductory data science course. (I don't know where this goes yet; Mine brings up a good point to make sure this is somewhere highlighted at the forefront)

In this paper, we discuss the integration of technology in our Introductory to Data Science course, and how these choices have helped shape our curricular and pedagogical decisions made. 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.

3 The Course

Reader should want to continue reading by thinking "this is something that I would like to teach"

Hook Students enroll in a large class capacity with little to no statistics, data science, or coding experience. These factors are two common hurdles identified by faculty when trying to implement a data science course (Schwab-McCoy et al. 2021, Kokkelenberg et al. 2008). By the end of this course, students are able to use R, R-studio, and GitHub 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).

– gain experience in data wrangling, exploratory data analysis, predictive modeling, and data visualization

– The Kaplan Way is a learning model "that combines a scientific, evidence-based design philosophy with a straightforward educational approach to learning" (Schweser 2023)

– Dedicated active learning approach

– have a section on the content we teach and cite a paper (maybe goes here)

In the following sections, we first detail the teaching team used to instruct Introduction to Data Science. Next, we detail the technology we've chosen to use when creating and facilitating our introductory data science course. Then, we discuss our pedagogical choices that go into a typical week of our Introduction to Data Science Course. Finally, we discuss

the preparation work needed to teach our Introduction to Data Science, informing those on how to get started when creating and modifying their own introductory data science course

4 Teaching Team

– What is needed

– What are their responsibilities

– Instructor; Head-TAs; Course Organizers; Lab Leaders; Lab Helpers

- Preparation: How to
- Communication: How to + Recommendations / strategies

5 Technology

Reader should want to continue reading by thinking "so how do I use this"

Next, we discuss our choices of technology…..

- Github: Why version control; what it is; how we use it at the beginnings of the course

– Collecting student information

– Setting up a GitHub organization

- R & R-studio; Why; what is is; how students get set up with it

– Creating Duke Containers

(Mention alternative R-studio Cloud in discussion?)

Each of these types of technology aids in the creation and implementation our data science pedagogy. This includes in-class application exercise (AEs), lectures, labs, and assessments within the Kaplan model of learning.

INSERT IMAGE HERE

Unpack the image below

6 Pedagogy

– Reader should want to continue reading by thinking "so how do I teach this"

In this course, we have chosen a combination of teaching methods, interactive activities, and learning assessments to help best prepare introductory data science students the tools they need to be successful outside of university or in future coursework. Our pedagogy includes facilitating in-class AEs, facilitating lectures, running a Lab, and assigning assessments to provide students an opportunity to show what they've learned.

6.1 Application Exercies (AEs)

A majority of the time in class will be dedicated to working on AEs. These exercises are live-coding opportunities to practice applying data science concepts and code introduced through preparation materials. AEs are no-stakes assessments, often graded on completion, where students have the ability to write and edit code while asking questions at the student or class level and receive immediate feedback.

– How GitHub and R are used to create AEs for students

It is up to the discretion of the instructor on the content that goes into the AEs. This can in-

clude having students write code themselves, fill in the blank coding exercises, commenting on complete code, or a combination of such questions.

– Kaplan Way (Practice) / Active learning implementation of AEs (typical class day strategies / discussion)

6.2 Lecture

A typical week yields 75-minute lectures on Mondays and Wednesdays. These lectures take up part of the class time, and are designed to introduce new concepts or review topics from the preparation videos in a more traditional format.

– Creating slides using Quarto in R (?)

– Kaplan Way (Prepare) / Active learning implementation during lecture (typical class day strategies / discussion)

Lectures are recorded and made available to students with an excused absence upon request.

6.3 Labs

– apply the concepts discussed in lecture to various data analysis scenarios, with a focus on the computation

– Individual and team based

– Repo creation

– Kaplan Way (Perform) (typical class day strategies / discussion)

6.4 Assessments

– Kaplan Way (Perform)

– HW

- apply what you've learned during lecture and lab to complete data analysis tasks

– Lab

- apply the concepts discussed in lecture to various data analysis scenarios, with a focus on the computation
- Individual and team based
- Repo creation

– Exams

- opportunity to demonstrate what you've learned in the course thus far
- take home exams

– Project

- analyze an interesting data-driven research question
- group project
- describe the process / expectations of the project

$\widetilde{\frown}$ The writing above should be in a form where there is practical / useful information for the reader. If they were designing the course, do they better understand the assessment structure of an Intro to Data Science course?

7 Discussion

– Pros (What works well)

Ex. Live coding good

– Cons (What could be improved)

Ex. Students can fall behind during live coding sessions

– Experiences

Examples include

– Feasible to adopt to teach for multiple semesters as data science continues to evolve. There is value in the investment.

– Computational Resources

– Human Resources

References

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.

Beckman, M. D., Çetinkaya Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J. & Tackett, M. (2021), 'Implementing version control with git and github as a learning objective in statistics and data science courses', *Journal of Statistics and Data Science Education* **29**(sup1), S132–S144.

URL: *https://doi.org/10.1080/10691898.2020.1848485*

BLS (2022), 'Occupational outlook handbook'.

URL: *https://www.bls.gov/ooh/math/data-scientists.htm*

Danyluk, A., Leidig, P., McGettrick, A., Cassel, L., Doyle, M., Servin, C., Schmitt, K. & Stefik, A. (2021), Computing competencies for undergraduate data science programs: An acm task force final report, *in* 'Proceedings of the 52nd ACM Technical Symposium on Computer Science Education', SIGCSE '21, Association for Computing Machinery, New York, NY, USA, p. 1119–1120.

URL: *https://doi.org/10.1145/3408877.3432586*

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruviluamala,

N., Uhlig, P. X., Washington, T. M., Wesley, C. L., White, D. & Ye, P. (2017), 'Curriculum guidelines for undergraduate programs in data science', *Annual Review of Statistics and Its Application* **4**(1), 15–30.

URL: *https://doi.org/10.1146/annurev-statistics-060116-053930*

Dogucu, M. & Çetinkaya Rundel, M. (2022), 'Tools and recommendations for reproducible teaching', *Journal of Statistics and Data Science Education* **30**(3), 251–260.

URL: *https://doi.org/10.1080/26939169.2022.2138645*

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

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

Donoho, D. (2017*b*), '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/*

Kokkelenberg, E. C., Dillon, M. & Christy, S. M. (2008), 'The effects of class size on student grades at a public university', *Economics of Education Review* **27**(2), 221–233.

URL: *https://www.sciencedirect.com/science/article/pii/S0272775707000271*

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*

Çetinkaya Rundel, M., Dogucu, M. & Rummerfield, W. (2022), '5ws and 1h of term projects in the introductory data science classroom', *STATISTICS EDUCATION RESEARCH JOURNAL* .