

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

March 8, 2023

Abstract

The text of your abstract. 200 or fewer words.

Keywords: Data Science, Curriculum, Pedagogy

1 The Course

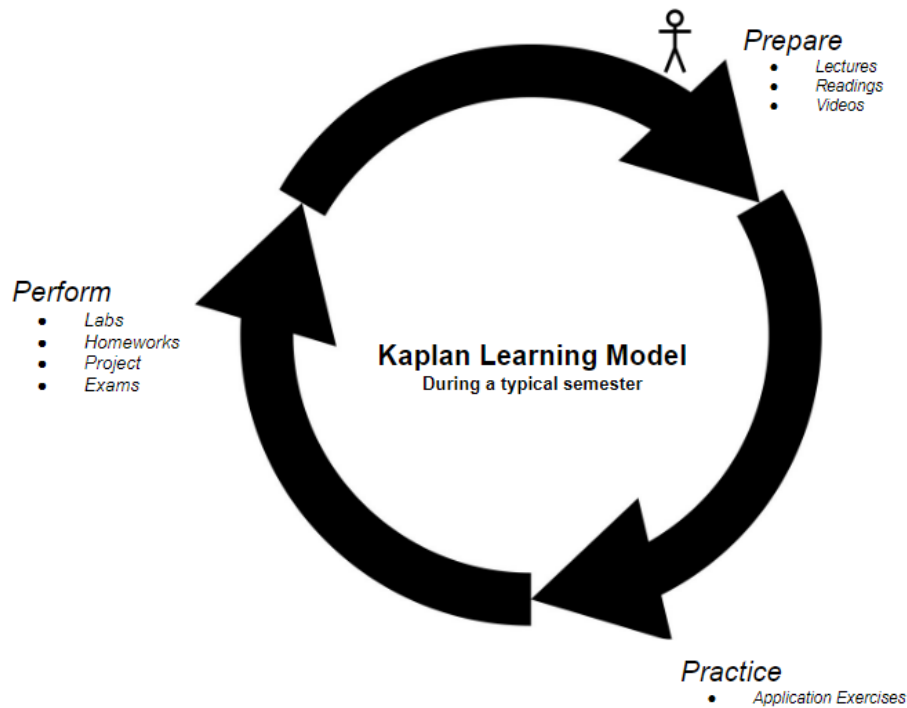
Reader should want to continue reading by thinking “this is something that I would like to teach”

In the following sections, we describe our introductory data science course offered to predominantly freshman level students at Duke University called Introduction to Data Science and Statistical Thinking (STA199). Often, these students are (interested in / major + minor in / description of typical student demographic). Students enrolling in STA199 commonly have little to no statistics, data science, or coding experience. In a typical semester, this course seats roughly 179 students, which is considered to be large by all measures. Both the lack of experience and size of class are identified as two common hurdles by faculty when trying to implement an introductory data science course ([Schwab-McCoy et al. 2021](#), [Kokkelenberg et al. 2008](#)). However, by the end of this course, students are able to use data both R and GitHub to understand natural phenomena, investigate patterns, and model outcomes in a reproducible format. \

This course is built on four large-scale learning objectives: learn to explore, visualize, and analyze data in a reproducible and shareable manner; gain experience in data wrangling and munging, exploratory data analysis, predictive modeling, and data visualization; work on problems and case studies inspired by and based on real-world questions and data; learn to effectively communicate results through written assignments and project presentation. These objectives are accomplished through interactive lectures and labs that present content, problems, and case studies inspired by and based on real-world questions and data.

When teaching, instructors are committed to the Kaplan Way learning model “that combines a scientific, evidence-based design philosophy with a straightforward educational

approach to learning” (Schweser 2023, pg. 3).



This model is composed of three phases: prepare, practice, and perform. All three phases are designed to guide the instructor in facilitating an overall quality learning experience for the student. Each of these phases are performed during a typical week within a semester, including through preparation material, lectures, in-class application exercises, labs, and assessments such as homework.

During the prepare phases, students are completing readings, watching videos, and listening to lecture that ultimately builds upon a new foundation of data science concepts being learned. The goal of preparing students is to put them in a position where they can build upon what they are learning, and create new knowledge through the practice phase. The practice phase is designed to be an opportunity for students to reinforce the new information gained, as well as uncover new concepts in data science. This is achieved through the use of interactive application exercises (AEs) in class where students work alone, in groups, or

with the instructor in live coding sessions. Finally, students enter the perform phase to show their progress made in the previous two phases. This is typically done through assessments such as homework, exams, and projects. Additionally, students perform individually, or with a group, during weekly labs that tend to focus more on computation. This learning model is a continuous cycle throughout the semester as new topics are introduced.

Topics taught in STA199 fall under two major units: Unit 1 - Exploring data; Unit 2 - Making rigorous conclusions. In Unit 1, students become first introduced to R, R-studio, and Github. During this unit, students start to create data visualizations and learn how to both import and manipulate data to be better suited for analysis. In unit 2, students extend their investigations with data to include modeling. Specifically, students fit a variety of models (simple linear regression, multiple linear regression, logistic regression), and learn the fundamentals of hypothesis testing. For a more complete description about the topics taught and data sets used in creating these lessons, please see **A fresh look at introductory data science (cite)**.

In this paper, we describe the preparation and implementation process of STA199 in its entirety. This includes details of the teaching team used to instruct STA199, technology chosen to use when creating and facilitating our introductory data science course, and the pedagogical choices that go into a typical week of teaching. This includes a comprehensive description of a flexible framework on how to create, set up, and implement an introduction to data science course similar to STA199. When describing this framework, we articulate first hand experiences and suggestions surrounding some of the choices made to create and instruct STA199.

2 Teaching Team

We define a teaching team as a group consisting of one instructor and multiple teaching assistants (TAs). The assignment of any teaching assistant is to both support the instructor in charge of the class, and support the students in the classroom. These TAs range from undergraduate to PH.D. level students, and vary in teaching experience. (Writing on TA selection process). Once selected... (writing on TA training).

Once training is complete, students are assigned roles that indicate their responsibilities during the semester. These roles include *course organizer*, *head TA*, *lab leader*, and *lab helper*. Often, these roles are given based on the level of student, with more academically experienced TAs taking on the roles of course organizer, head TA, and lab leader, where as students with less experience (i.e. undergraduate students) take on the role of lab helper.

Lab sections are held once a week, and are facilitated in person by both a lab leader and lab helper. The responsibilities of a lab helper are supporting both the students and lab leader as they see fit. Examples of this may include setting up the classroom before class, or conducting small group conversations when students have questions about the material. The lab leader is responsible for facilitating the lab. This involves working through a pre-made lab to ensure they can help students apply concepts discussed in lecture to what's being assigned. In addition, both must hold two hours of office hours each week and have grading responsibilities assigned throughout the semester.

Head TA responsibilities can generally be categorized as one of the following: Administrative or Pedagogical. Administrative responsibilities include the organization and distribution of TA responsibilities throughout the semester. It is imperative that the head TA and instructor clearly communicate expectations with each other to establish exactly how rules

and responsibilities that are given to the TAs are assigned. Administrative duties include reminding other TAs about bi-weekly payroll deadlines and ensure TAs are working their allotted hours per week (and not more). Within these allotted hours, a head TA distributes grading assignments and deadlines to both lab helpers and leaders per week. They make sure all TAs complete grading within a week and spot check the grading accuracy and quality of all written feedback given. (insert 1-2 sentences about experience). Pedagogically, head TAs are responsible for creating or reviewing answer keys and grading rubrics for homework and lab assessments as the instructor sees fit. Each head TA is also assigned to instruct one lab section during the semester. Before becoming a Head TA, there are additional

The course organizer is expected to work across each section of STA199, instead of working with a single instructor. Their responsibilities include creating rubrics for and working through homework and lab assignments. Additionally the course organizer, along with the instructor, answers real time questions virtually during labs from lab leaders. Questions often range from content related to technical questions about GitHub and R. Finally, the course organizer is responsible for handling all requested assignment extension requests from students. This includes filing away student exemptions, providing extensions for extreme circumstances, and enforcing the late work policy outlined in the syllabus when necessary.

(Paragraph on flexibility in team structure)

Through my experience as an instructor working with this designed team, it is imperative that everyone is communicating with each other. A team with many different roles poses risk for the instructor to be unaware of how or what decisions are being enacted at the grading and lab levels of the course. Thus, it is recommended that the instructor trains everyone on the teaching team to use a communication system that allows every member

to communicate any questions they may have, or decisions they make to the entire team. In the past, we have used the software *Slack*, with appropriately named channels such as *grading-questions* where TAs can post examples and questions about grading and the instructor can clearly state their expectations. Further, it should be noted that the head TA should not be treated as a “bridge of communication” for the instructor to the rest of the teaching team. It is critical that the instructor is in consistent contact with all members of the teaching team in making sure all lab leaders and helpers understand the course content, upcoming assignments, and know what’s expected of them in their assigned role. We recommend holding a weekly meeting with all members of the teaching team to ensure this. When members are unable to come, it is an expectation that they watch a recorded video of the meeting and reach out if they have any questions about what was discussed.

3 Technology

In this section we will detail the computing infrastructure used in STA199 used to create and a course such as STA199. This includes details on how to use R, R-studio, and GitHub’s from the instructor’s perspective to set up lectures, AEs, labs and homework. First, we start with the student information necessary to collect and steps that must be taken before creation can take place.

3.1 Setting up a GitHub account

We can use R, R-studio, and GitHub to create interactive lessons, and assign pre-created assessments to individual or groups of students. To do so, we must instruct students to create a GitHub account. This is done on the first day of class, and often, students are given

time during class to sign up. Following tips from “Happy Git with R” (cite), we suggest students do the following when creating their name:

- Incorporate your actual name
- Reuse your username from other contexts if you can
- Pick a username you will be comfortable revealing to your future boss
- Be as unique as possible in as few characters as possible. Shorter is better than longer
- Avoid words with special meaning in programming (i.e., NA)

Once students create their account, we suggest getting this information from them in a survey. This normally can be done through your learning management system. It is critical to reiterate to students that spelling and capitalization matter when answer this question.

We suggest asking the question as follows:

What is your GitHub username?

Answer this question by ONLY typing your GitHub name and nothing else.

Make sure you triple-check the spelling so I don't add random strangers to our course organization on GitHub.

If you, as the instructor, do not have a GitHub account, you will need to create one as well. This student information will be needed to create your GitHub organization for your course.

3.2 GitHub organization

GitHub organizations are shared accounts where instructors and students can collaborate across many projects at once. Using your account, you can create a new GitHub organization by clicking on your profile icon in the upper right hand corner, go to *Settings*,

Access, New Organization. It's suggested to name this organization the name of your class and the current semester you teaching in (i.e., sta199-s23). One of the many benefits of teaching through GitHub is the ability to re-use what you currently create as a template for subsequent semesters. Once your organization is created we can use packages within R and R-studio to efficiently invite students and create assessments for them.

3.3 Setting Up R & R-Studio

R is a statistical programming language for computing, modeling, and data visualization, while R-studio is an integrated development environment for R (cite). Both are freely available to download and use. In an introductory course, it is recommended to minimize student frustration and distraction through the use of pre-packaged computational infrastructure (Çetinkaya-Rundel and Rundel, 2018). Per this recommendation, STA199 has students use R and R-studio through a *Duke Container*. Duke containers provides instructors at Duke university the opportunity to facilitate the use of different software, such as R and R-studio, through an online container instead of needing students to locally download both programs. Additionally, instructors have the ability to manage and install packages students will need for the semester, helping provide a neat and well organized starting experience with a new statistical language. (insert discussion on how this is done).

(Thoughts: Transition to R-studio cloud discussion? What alternatives can instructors use to imitate Duke containers if this is something they do not have access to at their university.)

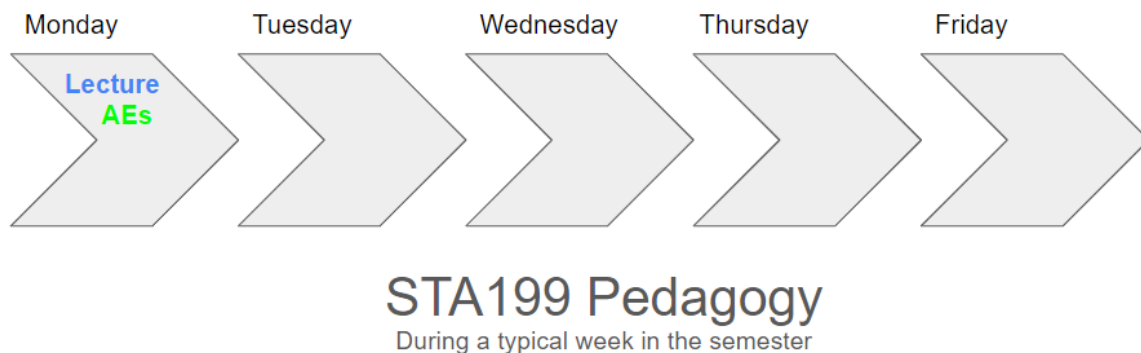
3.4 Using R & R-Studio to manage GitHub organization

(insert discussion / how-to on ghclass): includes invitations.

(set the stage for writing on how to create repos for students in subsequent sections *which will be a good segway*)

...

R, R-studio, and GitHub aids in the creation and implementation of our data science pedagogy. This includes the creation and distribution of in-class AEs, lectures, labs, and assessments. In the following sections we describe in detail both how to create, structure, and distribute materials used in STA199.



4 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.

4.1 Application Exercises (AEs)

4.2 Lecture

4.3 Labs

4.4 Assessments

Discussion

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>

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>