# Designing Data Science Workshops for Data-Intensive Environmental Science Research

Anonymous

**Abstract**

Over the last 20 years, statistics preparation has become vital for a broad range of scientific fields, and statistics coursework has been readily incorporated into undergraduate and graduate programs. However, a gap remains between the computational skills taught in statistics service courses and those required for the use of statistics in scientific research. Ten years after the publication of "Computing in the Statistics Curriculum," the nature of statistics continues to change, and computing skills are more necessary than ever for modern scientific researchers. In this paper, we describe research on the design and implementation of a suite of data science workshops for environmental science graduate students, providing students with the skills necessary to retrieve, view, wrangle, visualize, and analyze their data using reproducible tools. These workshops help to bridge the gap between the computing skills necessary for scientific research and the computing skills students leave their statistics service courses with. Open to faculty, staff, and the larger community, these workshops promote continued learning of the tools necessary for working with data and provide additional resources for incorporating data science into the classroom.

*Keywords:* data science, data visualization, data wrangling, `R`, environmental science, workshops, reproducible research

# 1 Introduction

Scientific fields have seen profound increases in the volume and variety of data available for analysis. Matched with the growth in computational power, today's scientific researchers are faced with computational and statistical expectations beyond those of the coursework dictated by their curriculum. In the environmental sciences, though statistics courses have been readily incorporated into undergraduate and graduate curricula, an abundance of literature suggests that these curricula fail to equip graduate students with the computing skills necessary for research in their field (Andelman et al., 2004; Green et al., 2005; Hampton et al., 2017; Hernandez et al., 2012, Mislan, Heer, & White, 2016; Teal et al., 2015; Theobold and Hancock, 2019). Only one of these studies (Theobold and Hancock, 2019), however, acknowledges the substantial role statistics courses could potentially play in students' acquisition of computational skills.

Over the last 10 years, a large number of statistics educators have echoed Nolan and Temple Lang's call to "embrace computing and integrate it fully into statistics undergraduate major and graduate programs" (Nolan and Temple Lang, 2010, p. 97; Baumer, 2015; Baumer, Horton, & Wickham, 2015; Cetinkaya-Rundel and Rundel, 2018; Cobb, 2015; Hardin et al., 2015; Horton and Hardin, 2015; Kaplan, 2018; McNamara and Horton, 2018). Indeed, the American Statistical Association Curriculum Guidelines for Undergraduate Programs in Statistical Science (2014) reflect the increasing importance of data science skills. Despite this campaign for computing in the statistics classroom, graduate-level statistics service courses have largely been overlooked, even though their potential impact is substantial. Unlike courses designed for an undergraduate or graduate program in Statistics, these service courses often act as the sole exposure to computing with data prior to the start of a student's independent research.

The intention of this research is to (1) describe the computing skills necessary for research in the environmental sciences, (2) investigate how these skills can be infused into currently existing extracurricular workshops, and (3) understand the experiences of attendees of these workshops. We investigated these areas of interest using a three-phase design-based implementation research model (Fishman et al., 2013). In the first phase, we conducted in-depth interviews with faculty from environmental science fields regarding the computational

skills they believe are necessary for graduate students to succeed in their research. Phase two then focused on adapting currently existing workshop resources to design of a series of data science workshops targeting the key computational skills distilled from these interviews. The final phase consisted of implementing the workshops and collecting survey responses from the workshop attendees regarding their experiences participating in each workshop.

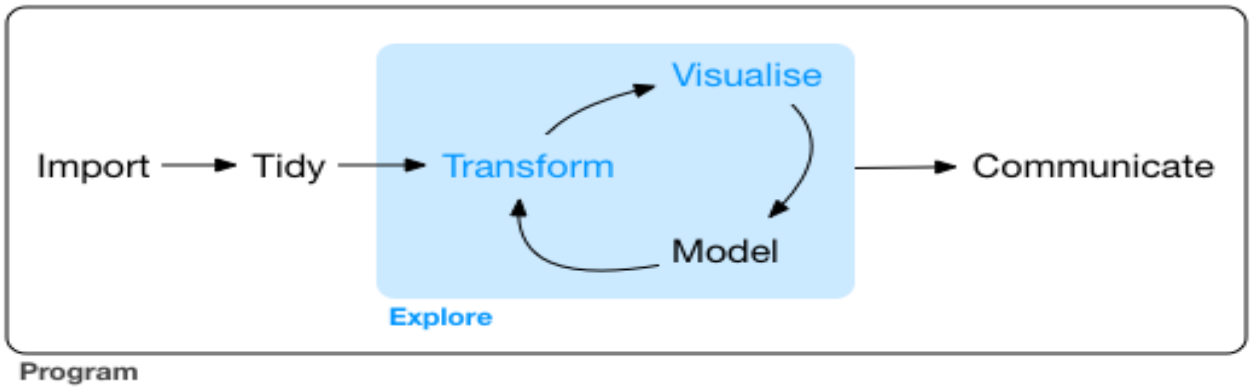


Figure 1: Data Analysis Cycle, Wickham, H. & Grolemund, G. (2017) *R for Data Science*. Sebastopol, California: O'Reilly.

For this research, the collection of disciplines who perform research across a variety of environmental science fields are captured under the term "environmental science." At our institution, these are the fields of Ecology, Land Resources nd Environmental Sciences, Plant Sciences and Plant Pathology, and Animal and Range Sciences, whose students are required or highly recommended to complete graduate-level statistics coursework for a masters or doctoral degree. In this paper, the "data analysis cycle" consists of all stages in the data analysis process, from data importation to data exploration to the communication of results (Figure 1), where data modeling is but one component. The "data science skills" necessary to engage in this cycle may include general programming concepts such as loops, user-defined functions, or conditional statements. However, the cornerstone of data science skills differs fundamentally from general programming skills, with a focus on data rather than computer architecture, design, and applications.

We begin by outlining areas of research that address the computational and statistical training of graduate students in the environmental sciences and the potential for extracur-

ricular workshops to fill gaps in students' computational preparation. Next, we outline methodology used to design and implement a suite of data science workshops tailored to environmental science graduate students. Section 4 summarizes the first phase of research, outlining the computational skills faculty members identified as necessary for graduate students to succeed in their independent research. Next, Section 5 discusses how these identified skills were interwoven into existing data science workshop materials for researchers in the environmental sciences. Section 6 summarizes the backgrounds and experiences of the workshop attendees during the 2018-2019 academic year, and describes the research conducted on the implementation of the workshops. The resources used to facilitate the sustainability of these workshops is outlined in section 7, alongside possibilities for formally integrating these workshops into the university curricula. Section 8 outline future research plans for a second iteration of this design work. To close, we revisit the current climate of computing in the statistics curriculum for service courses and describe how these types of extracurricular workshops can assist in further integration of computing into these classrooms.

# 2 The Current Climate of Statistics and Computing in the Environmental Sciences

Due to the substantial growth in the volume and variety of available data over the last two decades, the practice of environmental science has changed dramatically. Advances in technology have made computationally heavy applications of data science techniques—such as management and coalition of large data sets, high frequency spatial and temporal data visualization, and hierarchical Bayesian modeling—essential understandings for environmental science research. This flood of data has "challenged the research community's capacity to readily learn and implement the concepts, techniques, and tools" (Hampton et al., 2017, p. 546) necessary for data-intensive environmental science research, creating a crucial need to re-evaluate how our educational system can better prepare current and future generations of researchers (Green et al., 2005; Hampton et al., 2017).

## 2.1 Computing in the Environmental Sciences Curriculum

Arising from a decade of mumblings about the importance of computing to research in the environmental sciences (Andelman et al., 2004; Dodds et al., 2007, 2008; Eglen, 2009; Green et al., 2005; Hastings et al., 2005; Kelling et al., 2009; Wilson, 2006; Wilson et al., 2008; Wing, 2006), 2010 brought two studies on the computational ill-preparation of environmental students by their curriculum. In the first large scale study of ecology instructors, Strasser and Hampton found that undergraduate students were not being prepared with the data management tools necessary to engage in environmental science research. Across 51 different institutions, despite largely affirming the importance of data management skills, fewer than 20% of instructors reported data management topics in their courses. That same year, an environmental science graduate student led a large scale study of the computational experiences of future environmental scientists (Hernandez et al., 2012, p. 1068). In a survey of environmental science graduate students across the United States, the authors found that over 74% of the students surveyed reported they had no skills in any programming language—including `R`—and only 17% reported basic skill levels in any programming language. Hence, a large number of students may be leaving their undergraduate and graduate programs without the data science skills necessary for data-intensive research in their field. Hernandez and colleagues, however, noted that student-focused workshops could work toward bridging this gap, by "providing intensive environments" where students could learn "particular methods or technologies" (p. 1075).

Due to the lack of "training in data and computing skills" (Teal et al., 2015, p. 136) in undergraduate and graduate programs in the environmental sciences, external learning opportunities are necessary to prevent researchers from continuing to teach themselves or each other everything they know about data management and analysis. Out of these need for high-quality resources for learning scientific computing emerged The Carpentries project (2019). The Carpentries focuses on teaching "foundational computational and data science skills to researchers" through in-person, hands-on, domain-specific workshops. As part of their educational mission, The Carpentries collaboratively develops publicly available lessons for populations of researchers, which do not assume that attendees have any prior knowledge before attending the workshops. Teal and colleagues acknowledge that, while these work-

shops "will not be able to teach researchers all of the skills they need in two days," workshops "are a way to get started," lowering the activation energy required to begin acquiring computing skills and empowering researchers "to be able to conduct the analyses necessary for their work in an effective and reproducible way" (p. 143).

Over the last 20 years, statistical preparation in the environmental sciences has grown to be considered vital (Hampton et al., 2017), and statistics coursework has been integrated into graduate programs across the nation. Yet, none of these conversations have acknowledged the substantial role students' statistics education potentially plays in their attainment of the data science skills necessary for research. Today, throughout their research, the majority of environmental science graduate students are required to produce code as part of their data analysis process (Mislan et al., 2016). To compound the difficulties these graduate students face, over this same period, the software used by environmental science researchers has shifted, with an increase of over 45% in the use of `R` in environmental science publications since 2008 (Lai et al., 2019). The clear need for data science proficiency in environmental science research necessitates a transformation of the environmental science curriculum similar to that which infused statistics preparation into the required graduate coursework.

## 2.2 Computing in the Statistics Curriculum

Changes in the digital age have also had "a profound impact on statistics and the nature of data analysis" (Nolan and Temple Lang, 2010, p. 97), with today's skills differing substantially from what was needed but five to ten years ago. In the year following the publication of "Computing in the Statistics Curriculum" (Nolan and Temple Lang, 2010), the Mckinsey Report (Manyika et al., 2011) was published. The McKinsey report stated that, by 2018, "the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions" (p. 3). With calls to transform the undergraduate statistics curriculum resounding nationally, the 2014 American Statistical Association (ASA) President, Nathaniel Schenker, convened a workgroup to update the association's guidelines for undergraduate programs. These new guidelines included an increased emphasis on data science skills and real applications, specifically students' ability to "access and manipulate data in various ways, use a variety of computational approaches to extract

meaning from data, program in higher-level languages" (American Statistical Association Undergraduate Guidelines Workgroup, 2014, p. 7).

With this curricular momentum, in 2015, *The American Statistician* produced a special issue on "Statistics and the Undergraduate Curriculum," to encourage submissions of broader topics in the statistics curriculum. Articles in the special issue ranged from detailing how computing should be included throughout the Statistics curriculum (Green and Blankenship, 2015; Tintle et al., 2015; Hesterberg, 2015), to presenting thoughts on how data science topics should be integrated into undergraduate statistics courses, (Nolan and Temple Lang, 2015; Grimshaw, 2015; Baumer, 2015; Hardin et al., 2015). In the same issue, George Cobb provocatively stated that the statistics curriculum needed to be rebuilt "from the ground up" (2015), as "what we teach lags decades behind what we practice" and "the gap between our half-century-old curriculum and our contemporary statistical practice continues to widen" (p. 268). Moreover, despite the issue's focus on the broader statistics curriculum, statistics educators continued to lament that the current Introductory Statistics curriculum teaches a snapshot of the entire data analysis cycle, "wherein challenges with data computational methods, and visualization and presentation are typically elided" (Baumer, 2015, p. 336).

The following year, however, brought the revised GAISE college report (Revision Committee, 2014), creating a push for reform in the Introductory Statistics curriculum. The six recommendations originally outlined by the committee in 2005 continued, but the authors suggested two new emphases for the first recommendation (teach statistical thinking), which better reflect the modern practice of statistics. First, statistics educators should "teach statistics as an investigative process of problem-solving and decision making," and should "give students experience with multivariable thinking" (2014, p. 3). These recommendations reiterate the sentiments heard throughout the statistics community, that students should emerge from our courses with the understanding that data analysis "isn't just inference and modeling, it's also data importing, cleaning, preparation, exploration, and visualization" (Cetinkaya-Rundel, 2018). Yet, the inclusion of these topics in the Introductory Statistics curriculum is still a heated discussion. Many educators continue to believe (1) that it is not possible to teach statistical concepts and programming in just one course, (2) that teaching programming takes up valuable time which could be used towards teaching important statistical concepts, or (3) students are not interested in learning to program (Cetinkaya-Rundel,

2018). Thus, despite charges for the statistics community to "treat computing as fundamental as basic mathematics and writing" (Nolan and Temple Lang, 2015, p. 298), many students leave their Introductory Statistics course without "a set of practices and attitudes about data that are immediately applicable to their lives" (Gould, 2010, p. 309).

Amidst these conversations, `R` packages were being created, which would fundamentally changing how users interact with `R`. These `R` packages, universally known as the "`tidyverse`," have created user friendly `R` tools which "share an underlying design philosophy, grammar, and data structures" (Wickham, 2017). Statistics educators have begun to leverage these tools in the Introductory Statistics classroom to teach reproducibility (Baumer et al., 2014), data management (Baumer et al., 2015), dynamic data (Hardin, 2018), and big data (Wang et al., 2017). While there exists a growing momentum to incorporate these new `R` tools into the Introductory Statistics classroom, attention has yet to be paid to alternative statistics service courses, such as those taken by environmental science graduate students. These courses, like Introductory Statistics, serve graduate students from a variety of scientific backgrounds. However, unlike an undergraduate Introductory Statistics course, students are expected to emerge from their statistics coursework with the statistical and data science skills necessary for their research.

The frustrations echoed by environmental science educators (Hampton et al., 2017; Teal et al., 2015) suggest that, despite the inclusion of statistics coursework into these graduate programs, students continue to leave the statistics classroom without the data science skills necessary to participate in the data analysis cycle in their own research. The fundamental question raised ten years ago by Nolan and Temple Lang still applies today: do our students leave the statistics classroom able to "compute confidently, reliably, and efficiently?" (2010, p. 100). An in-depth study of environmental science graduate students' experiences acquiring the computing knowledge necessary for their research answered this question with a resounding 'no' (Theobold and Hancock, 2019). Like the hypothesis of Teal and colleagues (2015), these students did not attribute their acquisition of the data science skills necessary for their research to the statistics courses they took for their degree. Rather, students gained the data science skills necessary to engage in the entire data analysis cycle through independent research experiences, an "all-knowing" past or current graduate students, and peer networks. Ten years after the publication of "Computing in the Statistics Curriculum," we

continue to assume that "students will 'pick up' the skills they need" to participate in the data analysis cycle outside of their statistics coursework (Gould, 2010, p. 309).

## 2.3 Extracurricular Workshops to Bridge the Gap

Reiterated by both statistics education and environmental science researchers alike (Nolan and Temple Lang, 2010; Teal et al., 2015), this lack of training in computational skills impedes the progress of scientific research, sends the signal to students that computing is not of intellectual importance, and is laden with hidden costs. Students may pick up bad habits, misunderstandings, or the wrong concepts, learn just enough to get what they need done, spend weeks or months on tasks that could be done in hours or days, and they may be unaware of the reliability and reproducibility—or lack there of—of their results (Nolan and Temple Lang, 2010, p. 100; Teal et al., 2015, p. 136). But why are these skills still so rarely included in these service courses when the need for them is widely recognized?

Environmental science educators have reiterated the challenges in integrating computing into the curriculum outlined by Nolan and Temple Lang. These barriers can be boiled down to "attempting to fit more material into already-full courses and curriculum, which are taught by people who do not feel prepared to address topics relevant to big data and data-intensive research" (Hampton et al., 2017, p. 547). These hurdles are potentially even greater for graduate-level statistics service courses. Instructors of these courses are often explicitly told the statistical content students are expected to learn, while it is implicitly assumed they are also teaching students the data science skills necessary for them to participate in the entire data analysis cycle. Claiming these graduate students ought to take additional, data science specific courses to obtain these skills is infeasible for many, as graduate programs frequently leave little room for additional coursework.

Until computing has been meaningfully integrated into these service courses, extracurricular workshops hold the potential to address the gap between the computational preparation of students by their coursework and the computational requirements of their research. Although data science skills can potentially be acquired from the drove of currently available online resources, such as online lessons, MOOCs and books, none of these resources provide researchers with the ability to pose their questions directly to an instructor or to learn from others. Moreover, this drove of online materials, poses a "significant challenge in being able

to discover relevant and high-quality materials," for researchers with limited time.

As, reiterated by Nolan and Temple Lang (2015), extracurricular learning opportunities are not a direct substitute for the prolonged instruction of these skills that occurs in a course. However, this is not the goal of these learning opportunities. Instead, short, intensive workshops, such as those provided by The Carpentries, are able to teach immediately useful skills that can be taught and learned quickly, keep learners active by using live coding and formative assessment, work with a learners from a variety of backgrounds, and build learners' self efficacy (Word et al., 2017), so that attendees "learn the computational aspects as part of an interesting, challenging, and confidence-building process" (Nolan and Temple Lang, 2010, p. 101).

# 3 Methodology

Improving environmental science graduate students' access to "powerful, effective learning opportunities" (Fishman et al., 2013, p. 137) necessitates understanding the skills required for these students to be successful in their research. Design-based implementation research (DBIR) (Cobb et al., 2003; Fishman et al., 2013; O'Neill, 2012) "offers a model for the design and testing of innovations within the crucible of classrooms and other contexts for learning" (Fishman et al., 2013, p. 140). Rather than creating workshops covering content outside parties believe are important, DBIR uses collaboration with members of the community to develop "evidence-based improvements" (p. 143) to teaching innovations—situating community members as "co-designers of solutions to problems" (Fishman et al., 2013, p. 140) rather than bystanders.

This paper describes the results of the first iteration of this DBIR, consisting of three phases. Section 4 summarizes phase one which investigated the computing skills necessary for graduate-level environmental science research. Phase two of this research, described in Section 5, details how the skills identified during phase one were used to tailor currently existing Data Carpentry 2020 and Software Carpentry 2020 lessons to meet the needs of environmental science graduate students. Finally, Section 6 chronicles the third phase of this research, implementing and evaluating these workshops. This final evaluation phase focuses on the survey results of the backgrounds and experiences of workshop attendees,

rather than the workshop content or learning outcomes of attendees, which are described as directions for future research.

# 4 Outlining the Computing Skills Necessary for Environmental Science Research

As the direct supervisors of graduate students, environmental science faculty members are potentially aware of the computing skills that are vital to researchers in their respective fields. Thus, interviews with faculty members from these fields allow for us to gain an understanding of the essential skills required of environmental science graduate students.

In the spring of 2017 and fall of 2018, every faculty member in the Ecology, Land Resources and Environmental Sciences, Animal and Range Sciences, and Plant Sciences and Plant Pathology departments, currently overseeing a graduate student were emailed requesting their participation in this research. While some faculty enthusiastically agreed to participate, others declined for three main reasons—they hadn't directly overseen a graduate student recently, they deemed themselves to be weak in statistics, or they were unavailable to meet. Table 1 outlines the number of faculty requested for participation and the number of faculty interviewed, by departmental affiliation.

| Department | Faculty Invited | Faculty Interviewed |
|---|---|---|
| Animal & Range Sciences | 7 | 2 |
| Ecology | 15 | 8 |
| Land Resources and Environmental Sciences | 24 | 8 |
| Plant Sciences & Plant Pathology | 15 | 5 |

Table 1: Number of faculty members requested for participation and interviewed, by department.

## 4.1 Data Collection

The faculty members who agreed to participate, engaged in a one-hour interview regarding (1) the computational skills they believe are necessary for masters and doctoral students to implement statistics for research in their field, and (2) how they believe graduate students

acquire these necessary skills. The full interview protocol as supplementary material [1].

## 4.2 Data Analysis

The primary author led a three-stage data analysis process (Miles, Huberman, Saladaña, 2014). During the first stage, the interviews for every faculty member were transcribed verbatim. Following this process, the primary author read the transcripts independently, highlighting excerpts where computing skills were discussed. The author then created descriptive codes for the skills faculty identified as necessary in each of these excerpts. At the close of this stage, the author examined these codes for specific references to computing skills currently addressed in Data Carpentry's *Data Analysis and Visualization in* `R` *for Ecologists* lesson (Michonneau et al., 2019).

Following this process, the primary author began the second stage of analytical coding, synthesizing descriptive summaries into instances of a general concept (Miles et al., 2014, p. 95). During this stage, skills were linked thematically, and themes that held across multiple interviews were retained. Next, the author searched through these themes to uncover how each theme related to the others. Through this process it was determined that certain themes captured similar constructs, and were merged into a single theme, whereas other constructs were voiced independently, and separate themes were formed. For example, while every faculty voiced students' need to work with data in `R`, these sentiments were voiced alongside students' need to perform other data wrangling operations, such as reorganize data, filtering out rows of data, selecting columns, creating new variables, or modifying existing variables. Hence, the themes of "working with data" and "data wrangling" were merged into the single theme of "working with data." Alternatively, while reproducibility is a key aspect to working with data in `R`, faculty sentiments regarding the need for students' work to be reproducible was not voiced alongside a specific software.

In the final stage of the analysis, the primary author searched the faculty transcripts for evidence supporting the emerging themes, scrutinizing whether each identified skill fit into the identified themes. Following this validation process, the first and second authors met to discuss the rationale for each code and inspect the skills identified by faculty in the

[1] Materials associated with this manuscript are available at https://github.com/atheobold/data-science-workshops-jse

context of the emergent themes. These final themes provide evidence for designing data science workshops for environmental science graduate students.

## 4.3 Skills Identified by Environmental Science Faculty

While some faculty had difficulties disentangling the statistical methods students use in research from the computing required to implement those methods, many were able to express the computing skills necessary for graduate students in their field to engage in the entire data analysis process. A substantial overlap was seen between faculty expectations and the "data acumen" outlined by NAS (National Academies of Sciences, Engineering, and Medicine, 2018), falling into three categories: (1) working with and wrangling data, (2) data visualization, and (3) reproducibilty.

### 4.3.1 Working with Data

Every faculty member interviewed believed that students' experiences in the statistics classroom do not adequately prepare students to work with and organize large, messy datasets. As graduate students perform their research, they are required to think about "storing data, managing data, matching data, and collating data," into a meaningful datasets for analysis. Some faculty, aware of the different types of data their students work with, reflected that it is "not uncommon to be analyzing half a million records, but I think it's uncommon to be doing it effectively or efficiently."

These skills for working with data ranged from students' ability to "organize their data and get it in a way that can be used by `R`" to tasks that required reorganizing data formats from wide to long or vice versa—a skill which every faculty member griped is not acquired through the standard curriculum. A faculty member bemoaned that standard examples in statistics courses provide students with data which are the product of cross-tabulation, so students are never forced "to figure out how to get the cross-tabulation [they] need, so that [they] can bring it into `R` and do [their] regression." These concerns reiterate the importance of "data management and curation" detailed by NAS, who stated that "at the heart of data science is the storage, preparation, and accessing of data" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 26).

### 4.3.2 Data Visualization

The importance data visualization plays in every stage of students' research was emphasized by every faculty member interviewed. Faculty affirmed that students should possess the ability to create visualizations of their data early and often. These expectations align with the the facility outlined by NAS (2018), who stated that students need to have the ability to "present data in a clear and compelling fashion" (p. 26). One faculty member declared that students' ability to look at their data in different ways dramatically shapes their research potential, and the tools available today allow for researchers to create visualizations precisely tailored for each investigation. Many faculty voiced the usefulness of the `ggplot2` package (Wickham, 2016) lowering the barriers for students to learn "how to visualize [their] data to explore and understand it."

### 4.3.3 Reproducibility

Every faculty member emphasized the usefulness of "manipulating data in ways that are repeatable," through scripted programs such as `R`. Across environmental science disciplines, faculty concurred that many students do not use `R` for data wrangling, and instead rely on Excel because "the code (`R`) is kind of a black box" and when they "don't have that instant connection with [their] data, I think it fundamentally boils down to fear." Concerns were raised for the students using unreproducible tools to wrangle their data, as "they would never find [their] way back to what the original data set would have been" and their advisers would have no way to understand why certain data are missing. While many advisers stated that they encourage students to avoid these brute force Excel manipulations, they reflected that students may not have the computing skills necessary to perform the same data wrangling task in a scripted and reproducible manner. These faculty concerns parallel the "workflow and reproducibility" acumen outlined by NAS, who stated that students need to "be exposed to the concept of workflows" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 28).

### 4.3.4 How Students Gain Computational Skills

Across environmental science disciplines, faculty stated that they assume students are acquiring the computing skills necessary to analyze their data either in their required statistics coursework or on their own. When asked why students are not acquiring computing skills in their field-specific courses, a faculty member stated, "We don't really have anyone to teach that. It's not that it isn't valuable, but there is no one to teach it." Some faculty believed "most graduate students come in knowing more about the tools one might use to manipulate data than their advisers do," while others lamented the gaps between the computational skills of their graduate students and their own training, feeling "personally out of touch [with students] because I haven't taken the time to learn `R`, because of my training and my age."These gaps impact the assistance faculty can provide to their students, as "increasingly faculty feel that they're not at the forefront of their programming abilities, so their students are being self-taught and are often computationally ahead of them."

# 5 Designing Data Science Workshops for Graduate Students in the Environmental Sciences

The second phase of this research attended to the development of a suite of data science workshops targeted to graduate students in the environmental sciences. Skills identified through faculty interviews were incorporated into a set of four 3-hour workshops covering (1) the basics of programming in `R`, (2) intermediate programming tasks in `R`, (3) creating appropriate and effective data visualizations, and (4) cleaning and merging data in preparation for analysis and visualization, all using reproducible tools.

The materials for these workshops were adapted from Data Carpentry's *Data Analysis and Visualization in `R` for Ecologists* lesson (Michonneau et al., 2019), a curriculum maintained by experienced researchers in ecological fields, which uses ecology-specific data contexts [2]. Importantly, the first workshop in *Data Analysis and Visualization in `R` for*

[2]This work is a derivative of *Data Analysis and Visualization in `R` for Ecologists* (https://datacarpentry.org/R-ecology-lesson/) by Data Carpentry, used under CC-BY (https://creativecommons.org/licenses/by/4.0/).

*Ecologists* does not assume that attendees have any previous experience working in R, and each workshop builds on knowledge acquired at previous workshop(s), without the expectation that attendees have acquired additional knowledge or skills between workshops. The workshop materials developed for this research are available through GitHub [3], with video tutorials recorded and available through our institution's library [4].

## 5.1 Data Context

Emphasized by both the NAS and these faculty, "effective application of data science to a domain requires knowledge of that domain" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 29). Hence, data science instruction ought to be grounded in "substantive contextual examples," to "ensure that data scientists develop the capacity to pose and answer questions" with data relevant to them (2018, p. 30). Therefore, ecological data are used for these workshops, originating from Montana Fish, Wildlife, and Parks and the Portal Project Teaching Database (Ernest et al., 2018). These data highlight a variety of aspects that commonly occur in ecological data, including multiple sampling instances, mark-recapture, biological measurements, and meta- and micro-level data.

## 5.2 Computing Tools for Environmental Science Research

The structure and context of these workshops include a statistical programming language used extensively throughout environmental science research (R), environments which facilitate the learning of R (RStudio and RStudio Cloud), and tools that promote reproducibility throughout the entire data analysis cycle (R Markdown).

### 5.2.1 Why R?

The use of R is widespread throughout the environmental science research community, a dramatic change over the last decade (Lai et al., 2019). Furthermore, with the invention of the RStudio IDE (RStudio Team, 2015b), this user-ship continues to increase, as R includes over 100 packages frequently used in ecological data analysis (https://CRAN.R-

[3] GitHub repository (link)

[4] MSU Library videos (link)

project.org/view=Environmetrics). R is free and open source, so attendees learn a statistical programming language that will be accessible to them throughout their careers. Unlike MARK, VORTEX, or RMAS, with R, attendees results do not depend on remembering the sequence of buttons they clicked. With the shocking realization that large numbers of modern scientific findings cannot be replicated (The Economist Editorial, 2013; Johnson, 2014) and the growing appreciation for reproducible data analysis methods in ecological research (Cassey and Blackburn, 2006; Ellison, 2010; Morrison et al., 2016; Powers and Hampton, 2019), today's researchers in scientific fields are becoming more aware of the need for a reproducible data analysis workflow.

### 5.2.2 Why RStudio?

RStudio is a free computer application that allows you access to the resources of R, while providing you with a comfortable working environment (RStudio Team, 2015b). The RStudio IDE "makes [programming] less intimidating than the bare R shell" (Cetinkaya-Rundel and Rundel, 2018, p. 59). Additionally, the RStudio environment is consistent across operating systems, which is not the case for other statistical software packages. Moreover, because RStudio is an IDE, it includes integrated help files, intelligent code completion, and syntax highlighting—all of which help to reduce the learning curve.

### 5.2.3 Why RStudio Cloud?

The RStudio Cloud was created as a platform to make it easy to do, share, teach, and learn data science using R (RStudio Team, 2015a). Through the Cloud, attendees are able to access publicly available workshop materials, without worrying about software installation, package installation, or transferring data. Workshop participants interact with the workshop's materials in the same manner as a locally installed version of RStudio, as seen in Figure 2, and an organized RStudio project directory exposes attendees to best practices for reproducible project construction.

### 5.2.4 Why R Markdown Documents?

R Markdown documents provide an easy-to-understand framework to combine statistical computing and written analysis in a single document, helping to break the copy-paste

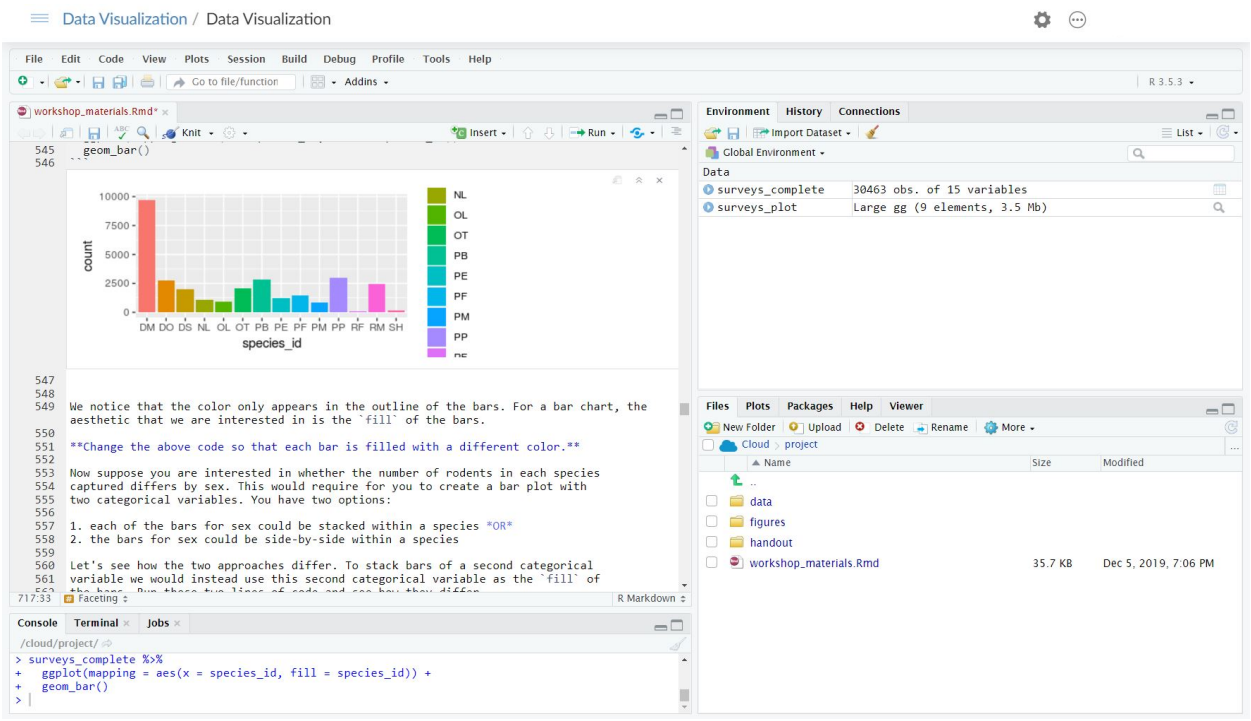


Figure 2: RStudio Cloud workspace environment for *Data Visualization with* `ggplot2` workshop. Every workshop works in an RStudio project, containing a master R Markdown file, a data folder containing the data used in the workshop, and the handout produced for attendees.

paradigm for generating statistical reports (Baumer et al., 2014). During the workshop, R Markdown documents allow for attendees to keep their code organized and their workspace clean, which is unnatural for new learners. Each workshop's master R Markdown document contains blocks of code and descriptions for every topic covered, allowing for participants' exploratory work to be saved within a topic. For additional information on R Markdown documents see Baumer et al. (2014).

## 5.3 Workshop Content

### 5.3.1 Introduction to `R`

This first workshop in the series covers the basics of learning to program in `R`. The workshop first introduces the RStudio environment and project work flow in RStudio, discussing project

working directories and relative paths. Next, the workshop progresses through tools for working with vectors and lists of different data types, motivating methods for working with dataframes. After learning how to import data into `R`, the workshop proceeds through inspecting data, extracting data, and changing data types. Motivated by working with missing data, the workshop introduces `R` help files to inspect function arguments and their default values. These help files are called upon as participants make use of base `R` functions to create data summaries, perform simple data cleaning, and produce both univariate and bivariate visualizations of the data.

### 5.3.2 Intermediate `R`

This second workshop covers coding skills to modularize `R` code. The content in this workshop, excluding relational statements, is not included in Data Carpentry's *Data Analysis and Visualization in `R` for Ecologists* lesson. Instead, many of these concepts are covered in Software Carpentry's *`R` for Reproducible Scientific Analysis* lesson. Yet, conditional statements, for-loops, and user-defined functions are skills that many environmental science faculty asserted were necessary for graduate students to possess as they perform independent research.

First, the workshop then progresses through the use of relational statements and how to link these statements using and (`&`), or (`|`), and not (`!`) conjunctions. Next, the workshop dives into the use of conditional statements, stepping from `if`, to `if else`, to `else if` statements. The second half of the workshop covers methods to iterate or replicate a set of instructions many times. Looping, specifically `for()` loops, are introduced as a popular way to iterate or replicate the same set of instructions. Working through exercises which repeat operations on a dataset using both a `for()` loop and a recursive `for()` loop, motivate a discussion of why `R` users recommend the use of vectorization for non-recursive `for()` loops. To conclude, functions are presented as an approach to replicate the same set of instructions in multiple locations throughout your code. Motivated by a script which copies and pastes the same process multiple times, attendees understand why this is an undesirable practice. Attendees are then tasked with transforming the copy-paste-modify process into a function. By parsing out the function writing process into a set of steps that should be used when you have copied and pasted your code multiple times, participants leave with a foundational understanding of why functions are useful and practical approaches for implementing them

in their own code.

### 5.3.3 Data Wrangling with `dplyr` and `tidyr`

The *Data Wrangling* workshop introduces common data wrangling issues faced by environmental science researchers. Inspired by the difficulty of reading bracket subsetting and how cumbersome it can be to remember the different base `R` functions and formats to wrangle your data, this workshop introduces the `dplyr` (Wickham et al., 2018) and `tidyr` (Wickham, 2014) packages. Much of `R`'s language has not changed over the last 20 years, which leaves the desire for a "smoother, more efficient, and more readable pipeline for modern `R` workflows" (Ross, Wickham, & Robinson, 2017, p. 19). The `tidyverse` packages share common interfaces and data structures that make it simpler to learn data wrangling tasks and allow for the process to flow naturally from one step to the next.

The workshop begins by outlining six of the common "verbs" that handle common data wrangling challenges, included in the `dplyr` package: `select()`, `filter()`, `mutate()`, `group_by()`, `summarise()`, and `arrange()`. Prompted by the need to perform a sequence of multiple data wrangling operations, participants learn how to connect each of these data wrangling verbs using the pipe operator (`%>%`). Next, motivated by the need to integrate additional data files for analysis, the concept of relational data is outlined. After an introduction to key-value pairs, attendees make use of the `left_join()` and `right_join()` functions to join these additional data files.

The final topic of the workshop involves the issue of data reorganization. Until now, participants have been presented with "tidy" data, where every observation is one row, each variable has a column, and every value has one cell. This concept of "tidy" data is used to describe 'long' and 'wide' data formats. The `tidyr` package is introduced to alleviate the burden of data reorganizations, when transforming data from one layout to another. In groups, participants work through a final exercise summarizing groups, using `pivot_wider()` to spread these values across multiple columns, and finally using `pivot_longer()` to gather these multiple columns into a single column.

### 5.3.4 Data Visualization with `ggplot2`

The final workshop in the series dives into creating data visualizations using the `ggplot2` package (Wickham, 2016). Rather than remembering a list of functions that make different visualizations, each with its own unique syntax, arguments, inputs, and outputs, `ggplot2` creates a uniform interface with functions that each solve a particular class of problems. This uniform syntax and "vocabulary for describing the elements of a statistical plot" (Nolan and Perrett, 2016, p. 261), allows participants to create more dynamic visualizations out of the gate.

Using the joined data from the close of the *Data Wrangling* workshop, a scatterplot is used to illuminate a discussion of the `ggplot()` syntax. Participants learn about the `mapping` argument for specifying aesthetics (`aes`) for the plot and the set of `geom` functions which define the type of plot you produce. By making explicit connections between the addition operator (`+`) and the pipe operator, participants understand addition to be an intuitive metaphor for adding layers to a plot. Next, the workshop examines how to modify the `ggplot()` aesthetics and geoms to create violin plots, density plots, bar charts, and line plots, allowing for participants to explore the `geom` functions and aesthetics that pair with each plot. A conversation is had about the importance of plotting raw data rather than simply aggregate measures of the data, and the difficulties that might arise. Similar to the advise of Nolan and Perrett (2016), adding a `geom_point()` or `geom_jitter()` layer to a visualization highlights tools that can be used so graph elements don't interfere with the data (e.g. jittering, transparency). Finally, faceting, using `facet_wrap()` and `facet_grid()`, is introduced as an additional visualization tool to facilitate multivariate comparisons (Nolan and Perrett, 2016, p. 261).

By this point in the workshop, participants have posed many questions on how to modify aspects of a plot that don't depend on the geom. For the final section of the workshop, the group walks through different customizations one can make to each `ggplot` object, to add clarity and information to the plot. Participants learn how to flip a plot's coordinate, how to make customizations of the plot's labels, the size of the points, the thickness of lines, the appearance of the plotting window, and the color scheme used. Each of these customizations continue to emphasize to participants the iterative nature of creating data visualizations,

transforming a simple plot step-by-step "into a graph that is data rich and presents a clear vision of the important features of the data" (Nolan and Perrett, 2016, p. 262).

# 6 Evaluating Data Science Workshops

During the 2018-2019 academic year, a total of 202 students, faculty, and staff attended at least one of the workshops, and we obtained 121 and 56 complete pre- and post-workshop survey responses, respectively. During the fall and spring semesters, a total of 84 individuals attended the *Introduction to* `R` workshop, 74 attended *Intermediate* `R`, 20 attended *Data Wrangling*, and 24 attended the *Data Visualization* workshop. The *Introduction to* `R` and *Intermediate* `R` workshops were offered twice during the fall semester, and once during the spring semester. The *Data Wrangling* and *Data Visualization* workshops were only offered once during the spring semester. The first workshop was offered two weeks after the start of the semester, with three week breaks taken between each of the subsequent workshops. Each workshop lasted a total of three hours and was taught by one lead instructor with two to three workshop assistants.

In the week prior to each workshop, a pre-workshop survey is sent out to those registered through a Google Form. This survey details individuals' demographics and backgrounds prior to the workshop. In the week following each workshop, attendees are asked to complete a post-workshop survey, detailing their experiences in the workshop. The content of these surveys was informed from the pre- and post-workshop surveys developed by The Carpentries [5]. The full pre- and post-workshop surveys are included as supplementary materials.

## 6.1 Backgrounds of Workshop Participants

The majority of the workshop attendees were from environmental science fields—from departments such as Land Resources and Environmental Sciences (LRES), Ecology, Plant

[5]This work is a derivative of The Carpentries pre- and post-workshop survey materials (https://github.com/carpentries/assessment/), used under CC-BY (https://creativecommons.org/licenses/by/4.0/), with revisions to the disciplines and occupations, and removal of questions regarding the degree of agreement with statements provided

Sciences and Plant Pathology, Biochemistry or Microbiology, Animal and Range Sciences, and Earth Sciences. Additionally, over 60% of workshop attendees were masters and doctoral students. It is worth noting, however, that 18 faculty, staff, and postdocs also attended these workshops. Figure 3 displays the department affiliations of the workshop attendees and their current occupation.

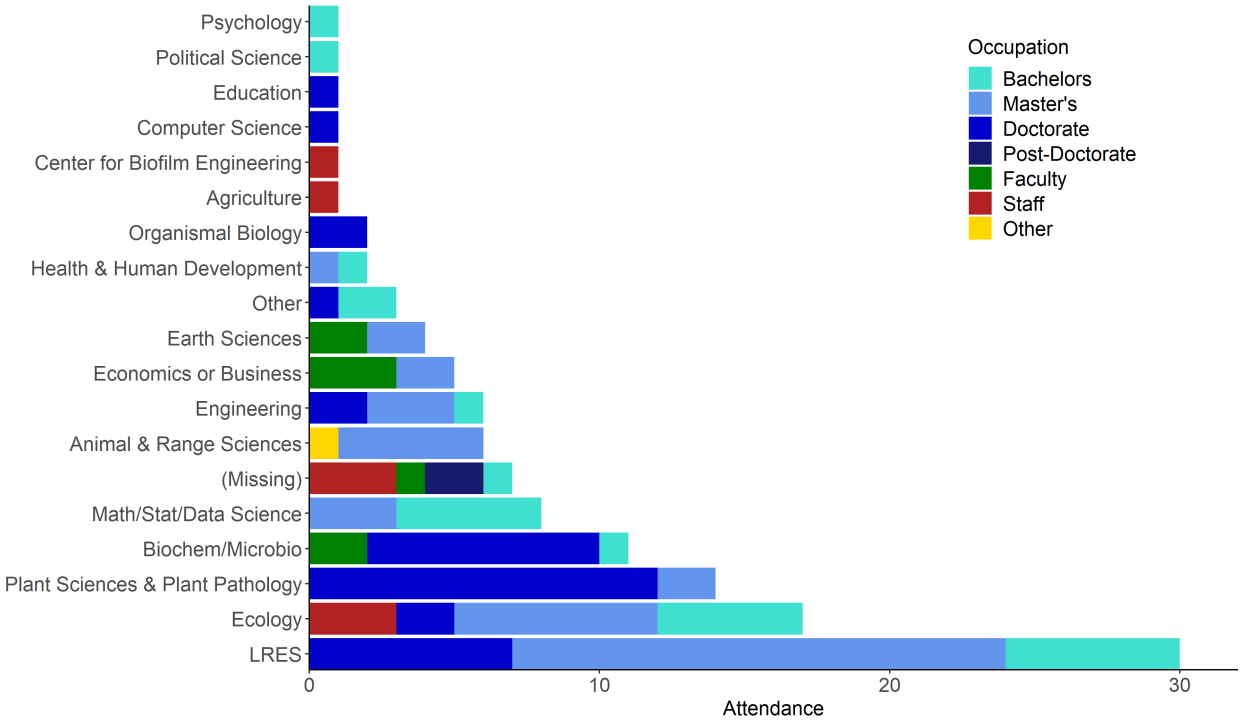


Figure 3: Number of attendees by department and current occupation, selected from an itemized list of campus departments and positions.

Consistent with the environmental science literature (Andelman et al., 2004; Hampton et al., 2017; Hernandez et al., 2012; Teal et al., 2015), a large number of workshop participants were either unfamiliar with the concept of a programming language or had no experience with any programming languages. Nearly 60% of attendees reported no experiences with any programming languages, 20% reported experiences working in `R`, and 30% reported experiences with other programming languages (e.g. MatLab, SQL, Java, C).

Many attendees, however, stated that they had taken courses in statistics. The majority of participants reported having either undergraduate or graduate experiences with an introductory level statistics course. Notably, over 15% of attendees reported having no

formal statistical training. Most graduate students had enrolled in discipline-specific introductory statistics courses in their own department or a graduate-level applied statistics course offered by the Department of Mathematical Sciences. Table 2 consolidates themes of workshop participants' previous statistical experiences, when asked to report the statistics courses they have taken over the course of their education.

| Stat Courses | Participants |
|---|---|
| Introductory Statistics | 46 |
| Applied Statistics | 42 |
| None | 24 |
| Discipline Specific Introductory Statistics | 20 |
| Intermediate Statistics | 10 |
| Experimental Design | 8 |
| Probability Theory | 6 |
| Statistical Computing | 3 |
| Sampling | 3 |
| Biostatistics | 2 |
| Spatial Analysis | 2 |
| Econometrics | 1 |
| Time Series Analysis | 1 |

Table 2: Workshop attendees' responses to the question "What are your previous statistical experience(s)? List course names.," thematically organized based on content of the course.

## 6.2 Motivation for Attending

As expected from the prevalence of the use of `R` in environmental science research (Lai et al., 2019; Mislan et al., 2016), over half of the master's, doctoral, and post-doc workshop participants attended for assistance with their research. Others were seeking additional assistance for learning the `R` skills necessary for their coursework, refreshing or updating their `R` skills to include new tools they were unfamiliar with(e.g. `ggplot`, `dplyr`), or undergraduates preparing for graduate school.

As echoed by previous studies of environmental science graduate students (Teal et al., 2015; Theobold and Hancock, 2019), attendees overwhelmingly stated that they primarily use the internet (27%), their peers (21%), or their lab mates (15%) when learning `R`. Based on the statistical backgrounds of these participants and the statistics education literature on computing in the statistics classroom, it is not surprising that nearly two-thirds of these

individuals reported using resources other than course materials as their main resource for learning R.

### 6.3 Reflections of Workshop Participants

The percentage of individuals reporting that all of the information presented was new to them differed by workshop, with 40% of *Introduction to R* participants, 30% of *Intermediate R* participants, 80% of *Data Wrangling* participants, and 50% of *Data Visualization* participants stating the information was new to them. Across every workshop, nearly all participants stated that they "strongly agreed" that they "learned skills that [they] will be able to use in [their] research/work." Over 75% of the workshop participants reported that they would use the skills they learned in their research immediately or in the next 30 days.

The themes which emerged from these attendees' reflections to what they enjoyed most about the workshop were hands-on learning, workshop atmosphere, instructor attributes, and confidence. Many attendees felt walking through the code step-by-step and the hands-on exercises "fostering a much greater level of understanding" and left them feeling more "confident figuring things out on my own." Furthermore, these attendees voiced that the workshop left them feeling more independent, because "I have a better understanding of how to read code, what certain symbols/terms/etc mean and how they work." Individuals who reported using the internet as a resource to learn R stated that "it's easy to walk away from R workshops wondering if anything was learned, however the exercises were a clear tool which allow me to see what I gained."

## 7 Sustainability of Workshops

To facilitate the sustainability of these workshops, we forged a partnership between our institution's library and the Department of Mathematical Science's Statistical Consulting and Research Services (SCRS). We believe a university's library is an optimal unit for offering these workshops, as it is both department-agnostic and a central hub for the entire university community. Furthermore, by partnering with a organization that provides statistical consulting, workshop participants are provided with a potential avenue if difficulties or additional questions arise—so the peer network is not shifted onto workshop instructors.

A data-engagement grant from the National Network of Libraries of Medicine during the 2018-2019 academic year supported the primary author in leading the workshops, becoming a Carpentries certified instructor, and incorporating the results of this research into the broader Data and Software Carpentry curricula. A $5,000 faculty excellence grant during the 2019-2020 academic year, facilitated the implementation of a "train-the-trainer" model, training two future graduate student instructors. Students were recruited from the masters and doctoral programs in statistics, but because of the widespread use of `R` across scientific fields, students from a variety of backgrounds hold the potential to be effective instructors. Both semesters, the authors met with these students for one hour a week to build students' facilities and confidence instructing each workshop. Each of these semesters, students taught different 30 to 45 minute portions of each workshop during, and acted as assistants for the remainder for the workshop.

The Carpentries does not require training for instructors to use their content, as The Carpentries materials are publicly available for use and adaptation (with acknowledgement). However, if the instructor or institution desires to advertise their workshops as Carpentries workshops there are two options: (1) the lead instructor is a Carpentries certified instructor, or (2) the institution requests a workshop through The Carpentries, who then recruits instructors for a fee. Through the primary author completing The Carpentries instructor training, we were able to offer self-organized workshops interweaving the content from both Data and Software Carpentry workshops. Additionally, through this experience the primary author was able to guide the future workshop instructors through the process of becoming Carpentries certified instructors.

Similar to the Explorations in Statistics Research workshop model (Nolan and Temple Lang, 2015), the "standard" Carpentries workshop format takes place over an intensive two days. Self-organized workshops allow for the added flexibility of tailoring this format to be more conducive for busy students, faculty, and staff. This revised format has both benefits and costs. The additional time between each workshop helps to alleviate the brain fatigue often experienced in intensive workshops, and allows for participants to attend the workshops that are relevant to the skills they wish to acquire. However, in this extended format, workshops after *Introduction to* `R` are potentially considered "specialized" workshops and experience lower attendance. At an academic institution, there is the possibility of in-

tegrating this type of workshop series into a single credit course. When considering this as an option, however, institutions should think critically about how faculty and staff can still participate in these critical learning opportunities. Alternatively, institutions could offer undergraduate students the option of assisting in the implementation of these workshops for course credit, and allow for the possibility of students becoming lead or co-instructors as they progress through their program.

# 8 Limitations & Future Research

The sentiments heard by faculty members in this research, unearth the possibility that many faculty may be unaware of the computing skills necessary for their graduate students to participate in the entire data analysis cycle. Instead, students may have more relevant knowledge regarding the data science skills that are necessary for their research. Hence, the next iteration of this design work will be informed by the collection of the research (`R`) code produced by environmental science graduate students. Graduate students' research code acts as artifacts of their research experience, providing "mute evidence" (Hodder, 1994) of the data science skills necessary throughout the data analysis cycle. The skills outlined by this research aid in reevaluating the content of these workshops, to ensure they cover the skills necessary for graduate-level environmental science research.

Additionally, the attendance of these tailored workshops by students, faculty, and staff from disciplines outside of the environmental sciences brings to question whether this type of tailored design work is necessary. Over a third of the workshop attendees came from disciplines outside of the environmental sciences, and, strikingly, these attendees reported similar workshop experiences to attendees from these targeted disciplines. This brings to question if there are common computational understandings necessary for research in *any* scientific field, which should be infused into *every* statistics and data science course. Alternatively, we saw a greater persistence across workshops by attendees from environmental science fields. This made us wonder, what are the drivers behind these individuals' continued attendance? Future research investigating the learning outcomes of workshop attendees holds the potential to provide fruitful insight on the necessity of discipline-specific learning opportunities.

Finally, despite the increasing availability of extracurricular workshops, research has yet to investigate the consistency or drift of these workshops. In this research, because of the large attendance at many *Introduction to R* workshops, a large number of questions would arise over the course of the afternoon. This lead to an inability to cover some of the workshop content in as much depth as hoped, yet some attendees remarked that "with so many people, [the workshop] had better discussions." A large scale analysis of the content covered by these workshops could unearth common questions or misunderstandings, aiding in the reconstruction of lessons to better scaffold learning.

# 9 Conclusion

Ten years ago, Nolan and Temple Lang declared that "modernizing the statistics curricula to include computing [...] is an issue that deserves widespread attention and action" (p. 106). Over the last ten years, we have seen both small (Revision Committee, 2014) and large (American Statistical Association Undergraduate Guidelines Workgroup, 2014) changes advocated to the statistics curriculum. Unfortunately, changes to graduate-level statistics service courses has received less attention and poses different issues.

Statistics courses that serve a variety of students (undergraduate, graduate, statistics major, non-major) reflect a snapshot of the statistics curriculum, but often act as many students' sole statistics course prior to conducting scientific research. Instructors of these courses thus grapple with difficult decisions of how they can ensure their students have both the statistical and "computational understanding, skills, and confidence needed to actively and wholeheartedly participate" in the scientific research arena (Nolan and Temple Lang, 2010, p. 106). For instructors unfamiliar with students' scientific disciplines, it can be difficult to "be bold and design curricula from scratch" (Nolan and Temple Lang, 2010, p. 106). The topics suggested by Nolan and Temple Lang (2010) represent a starting point toward building a taxonomy for computing in statistics for undergraduate and graduate statistics programs. These topics, however, may not be relevant to or emphasized by other scientific disciplines whose students enroll in graduate-level statistics service courses. In our research, we found that environmental science faculty stressed the importance of graduate students developing skills surrounding the fundamentals of working with data in `R`, software skills for

data processing and preparation, creation of data visualizations, and usage of reproducible work flows.

The time is ripe for us to "update the foundational concepts and infrastructure" (He et al., 2019, p. 5) included in statistics service courses, in the new era of data science. As we work toward a more thorough integration of computing into these courses, this research offers a model for facilitating external workshops, which hold the potential to fill a critical hole in the curriculum of many college programs. External workshops hold the opportunity for co-curricular learning, when paired with statistics service courses, so students leave their statistics service course with the computing skills necessary to engage in the entire data analysis cycle. Moreover, these workshops support university-wide data science literacy, facilitating avenues for faculty to acquire data science knowledge and skills which "they have not had the opportunity to learn well" (Nolan and Temple Lang, 2010, p. 106), and providing resources for instructors to meaningfully integrate discipline-specific computing skills into their classroom.

# 10 Acknowledgements

We would like to specially thank the participants from this study, without whom this research would not have been possible. We would also like to thank the workshop helpers for their time and assistance, helping to grow the data literacy across our campus. Lastly, we thank Mary Alice Carlson, Jennifer Green, Mark Greenwood, Megan Wickstrom, editor Johanna Hardin, and the reviewers for their insightful comments on this paper.

# References

American Statistical Association Undergraduate Guidelines Workgroup (2014). *2014 curriculum guidelines for undergraduate programs in statistical science*. American Statistical Association, Alexandria, VA.

Andelman, S. J., Bowles, C. M., Willig, M. R., and Waide, R. B. (2004). Understanding environmental complexity through a distributed knowledge network. *BioScience*, 54(3):240–246.

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician*, 69(4):334–342.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014). R markdown: Integrating a reproducible analysis tool into introductory statistics. *Technology Innovations in Statistics Education*, 8(1):1–22.

Baumer, B. S., Horton, N. J., and Wickham, H. (2015). Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics. *CHANCE*, 28(2):40–50.

Cassey, P. and Blackburn, T. M. (2006). Reproducibility and repeatability in ecology. *BioScience*, 56(12):98.

Cetinkaya-Rundel, M. (2018). Intro stats, intro data science: Do we need both? Presented at the 2018 Joint Statistical Meetings.

Cetinkaya-Rundel, M. and Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician*, 72(1):58–65.

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, 69(4):266–282.

Cobb, P. A., Confrey, J., diSessa, A. A., Lehrer, R., and Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1):9–13.

Data Carpentry (2020). https://datacarpentry.org/.

Dodds, Z., Alvarado, C., Kuenning, G., and Libeskind-Hadas, R. (2007). Breadth-first CS 1 for scientists. In *Proceedings of the 2007 ITiCSE*. ACM.

Dodds, Z., Libeskind-Hadas, R., Alvarado, C., and Kuenning, G. (2008). Evaluating a breadth-first CS 1 for scientists. In *Proceedings of the 2008 SIGCSE*. ACM.

Eglen, S. J. (2009). A quick guide to teaching R programming to computational biology students. *PLOS Computational Biology*, 5(8):1–4.

Ellison, A. M. (2010). Repeatability and transparency in ecological research. *Ecology*, 91(9):2536–2539.

Ernest, M., Brown, J., Valone, T., and White, E. P. (2018). Portal project teaching database.

Fishman, B. J., Penuel, W. R., Allen, A.-R., Cheng, B. H., and Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. In Fishman, B. J. and Penuel, W. R., editors, *Design Based Implementation Research*, volume 112, pages 136–156. National Society for the Study of Education.

Gould, R. (2010). Statistics and the modern student. *International Statistics Reveiw*, 78(2):297–315.

Green, J. L. and Blankenship, E. E. (2015). Fostering conceptual understanding in mathematical statistics. *The American Statistician*, 69(4):315–325.

Green, J. L., Hastings, A., Arzberger, P., Ayala, F. J., Cottingham, K. L., Cuddington, K., Davis, F., Dunne, J. A., Fortin, M.-J., Gerber, L., and Neubert, M. (2005). Complexity in ecology and conservation: Mathematical, statistical, and computational challenges. *BioScience*, 55(6):501–510.

Grimshaw, S. D. (2015). A framework for infusing authentic data experiences within statistics courses. *The American Statistician*, 69(4):307–314.

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R. R., Boettiger, C., Collins, S. L., Gross, L. J., Fernndez, D. S., Budden, A., White, E. P., Teal, T. K., Labou, S. G., and Aukema, J. E. (2017). Skills and knowledge for data-intensive environmental research. *BioScience*, 67(6):546–557.

Hardin, J. (2018). Dynamic data in the statistics classroom. *Technology Innovations in Statistics Education*, 11(1):1–22.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., and Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "think with data". *The American Statistician*, 69(4):343–353.

Hastings, A., Arzberger, P., Bolker, B., Collins, S., Ives, Anthony, R., Johnson, N. A., and Palmer, M. A. (2005). Quantitative bioscience for the 21st century. *BioScience*, 55(6):511–517.

He, X., Madigan, D., Yu, B., and Wellner, J. (2019). Statistics at a crossroads: who is for the challenge. Technical report, The National Science Foundation.

Hernandez, R. R., Mayernik, M. S., Murphy-Mariscal, M. L., and Allen, M. F. (2012). Advanced technologies and data management practices in environmental science: Lessons from academia. *BioScience*, 62(12):1067–1076.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386.

Hodder, I. (1994). The interpretation of documents and material culture. In Denzin, N. K. and Lincoln, Y. S., editors, *Handbook of qualitative research*, pages 393–402. Sage Publications, Inc., Thousand Oaks, California.

Horton, N. J. and Hardin, J. S. (2015). Teaching the next generation of statistics students to "think with data": Special issue on statistics and the undergraduate curriculum. *The American Statistician*, 69(4):259–265.

Johnson, G. (2014). "new truths that only one can see". *The New York Times*.

Kaplan, D. (2018). Teaching stats for data science. *The American Statistician*, 72(1):89–96.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, (59):613–620.

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., and Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, 10(1).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute.

McNamara, A. and Horton, N. (2018). Wrangling categorical data in R. *The American Statistician*, 72(1):97–104.

Michonneau, F., Teal, T., Fournier, A., Seok, B., Obeng, A., Pawlik, A. N., Conrado, A. C., Woo, K., Lijnzaad, P., Hart, T., White, E. P., Marwick, B., Bolker, B., Jordan, K. L., Ashander, J., Dashnow, H., Hertweck, K., Cuesta, S. M., Becker, E. A., Guillou, S., Shiklomanov, A., Klinges, D., Odom, G. J., Jean, M., Mislan, K. A. S., Johnson, K., Jahn, N., Mannheimer, S., Pederson, S., Pletzer, A., Fouilloux, A., Switzer, C., Bahlai, C., Li, D., Kerchner, D., Rodriguez-Sanchez, F., Rajeg, G. P. W., Ye, H., Tavares, H., Leinweber, K., Peck, K., Lepore, M. L., Hancock, S., Sandmann, T., Hodges, T., Tirok, K., Jean, M., Bailey, A., von Hardenberg, A., Theobold, A., Wright, A., Basu, A., Johnson, C., Voter, C., Hulshof, C., Bouquin, D., Quinn, D., Vanichkina, D., Wilson, E., Strauss, E., Bledsoe, E., Gan, E., Fishman, D., Boehm, F., Daskalova, G., Tavares, H., Kaupp, J., Dunic, J., Keane, J., Stachelek, J., Herr, J. R., Millar, J., Lotterhos, K., Cranston, K., Direk, K., Tyln, K., Chatzidimitriou, K., Deer, L., Tarkowski, L., Chiapello, M., Burle, M.-H., Ankenbrand, M., Czapanskiy, M., Moreno, M., Culshaw-Maurer, M., Koontz, M., Weisner, M., Johnston, M., Carchedi, N., Burge, O. R., Harrison, P., Humburg, P., Pauloo, R., Peek, R., Elahi, R., Cortijo, S., Umashankar, S., Goswami, S., Yanco, S., Webster, T., Reiter, T., Pearse, W., and Li, Y. (2019). datacarpentry/r-ecology-lesson: Data carpentry: Data analysis and visualization in R for ecologists, July 2019.

Miles, M. B., Huberman, A. M., and Saldana, J. (2014). *Qualitative Data Analysis, A Methods Sourcebook*. SAGE, Thousand Oaks, CA, 3rd edition.

Mislan, K., Heer, J. M., and White, E. P. (2016). Elevating the status of code in ecology. *Trends in Ecology & Evolution*, 31(1):4–7.

Morrison, C., Wardle, C., and Castley, J. (2016). Repeatability and reproducibility of population viability analysis (pva) and the implications for threatened species management. *Frontiers in Ecology and Evolution*, 4:98.

National Academies of Sciences, Engineering, and Medicine (2018). *Data Science for Undergraduates: Opportunities and Options*. The National Academies Press, Washington, DC.

Nolan, D. and Perrett, J. (2016). Teaching and learning data visualization: Ideas and assignments. *The American Statistician*, 70(3):260–269.

Nolan, D. and Temple Lang, D. (2010). Computing in the statistics curriculum. *The American Statistician*, 64(2):97–107.

Nolan, D. and Temple Lang, D. (2015). Explorations in statistics research: An approach to expose undergraduates to authentic data analysis. *The American Statistician*, 69(4):292–299.

O'Neill, D. K. (2012). Designs that fly: What the history of aeronautics tells us about the future of design-based research in education. *International Journal of Research and Method in Education*, 35(2):119–140.

Powers, S. M. and Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, 29(1).

Revision Committee, A. (2014). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. American Statistical Association, Alexandria, VA.

Ross, Z., Wickham, H., and Robinson, D. (2017). Declutter your R workflow with tidy tools. Technical report, PeerJ Preprints.

RStudio Team (2015a). *RStudio Cloud*. RStudio, Inc., Boston, MA.

RStudio Team (2015b). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Software Carpentry (2020). https://software-carpentry.org/.

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., and Pawlik, A. (2015). Data carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation*, 10(1):343–353.

The Carpentries (2019). https://carpentries.org/.

The Economist Editorial (2013). *"Trouble at the lab. (Cover story)"*. .

Theobold, A. and Hancock, S. (2019). How environmental science graduate students acquire statistical computing skills. *Statistics Education Research Journal*, 18(2):68–85.

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., and VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. *The American Statistician*, 69(4):362–370.

Wang, X., Rush, C., and Horton, N. J. (2017). Data visualization on day one: Bringing big ideas into intro stats early and often. *Technology Innovations in Statistics Education*, 10(1):1–22.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wickham, H., Franois, R., Henry, L., and Mller, K. (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6.

Wilson, G. (2006). Software carpentry: Getting scientists to write better code by making them more productive. *Computing in Science & Engineering*, 8(6):66–69.

Wilson, G., Alvarado, C., Campbell, J., Landau, R., and Sedgewich, R. (2008). CS-1 for scientists. In *Technical Symposium with Computer Science Education*, pages 36–37. ACM.

Wing, J. (2006). Computational thinking. *Communications of ACM*, 49(3):33–35.

Word, K. R., Jordan, K., Becker, E., Williams, J., Reynolds, P., Hodge, A., Belkin, M., Marwick, B., and Teal, T. (2017). When do workshops work? a response to the 'null effects' paper from Feldon et al. Technical report, Software Carpentry.