Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years

Aimee Schwab-McCoy, Catherine M. Baker, and Rebecca E. Gasper

1. Introduction

In 2010, Nolan and Temple Lang (2010) recommended incorporating six computational topics into undergraduate statistics education:

1. Fundamentals in scientific computing with data
2. Information technologies
3. Computational statistics (e.g., numerical algorithms) for implementing statistical methods
4. Advanced statistical computing
5. Data visualization
6. Integrated development environments (IDEs)

Since then, the academic landscape has been changed by the rise of data science. College faculty and administrators are grappling with the question of what a data science curriculum should look like and how much computing and statistics should be involved at each point in the design process as new academic programs, courses, and departments are created. Nolan and Temple Lang envisioned a future for statistics with a dramatically increased computational presence, however, the unprecedented growth of data science has shifted some of the urgency and emphasis on computing to new data science courses. In Fall 2019, 69 college and university faculty teaching data science courses and developing data science curricula were surveyed to learn about their curricula, computing tools, and challenges they face in their classrooms. Faculty reported teaching a variety of computing skills in introductory data science (albeit fewer computing topics than statistics topics), and that one of the biggest challenges they face is teaching computing to a diverse audience with varying preparation. The ever-evolving nature of data science is a major hurdle for faculty teaching data science courses, and a call for more data science teaching resources was echoed in many responses.

2. Related Work

2.1. Data Science Case Studies

As data science is still a relatively young discipline, so is data science education. Case studies are a central component of the data science education body of literature, often exploring a single institution's approach to teaching data science. One such example in The American Statistician described "SDS 292: Data Science" at Smith College, a course that covers data visualization, data wrangling, statistical modeling, and machine learning (Baumer 2015). From the Computer Science perspective, Brunner and Kim designed their "Introduction to Data Science" course at the University of Illinois at Urbana-Champaign to cover topics from data storage, data visualization, probability and statistical methods, through to high-performance computing (Brunner and Kim 2016). Another notable example is UC Berkeley's course “Data 8: The Foundations of Data Science.” First offered in Fall 2015, Data 8 introduces students to data science through three lenses: inferential thinking, computational thinking, and real-world applications. Like many data science courses, Data 8 is designed for students without a background in computer science or statistics. (The complete course repository is available at: http://data8.org/) Other examples of data science course and program case studies come from institutions such as Bowling Green (Albert and Rizzo 2016), the University of Nebraska at Omaha (Majumder and Cheng 2017), Wright State University (Asamoah, Doran, and Schiller 2020), Denison University (White 2018), University of Massachusetts, Dartmouth (Yan and Davis 2019), Harvard University (Hicks and Irizarry 2018), Columbia University (Zheng 2017), and Purdue University (Yavuz and Ward 2020). Many of these case studies echo similar course content and sequencing: data wrangling,
data visualization, and some basic statistical modeling. Hardin, Hoerl, and others describe how data science has been integrated into six existing statistics programs around the world (Hardin et al. 2015). They divide the data science curriculum into three major areas: programming, data technologies and formats, and statistical topics.

### 2.2. Data Science Curriculum Content

While case studies are useful, especially for instructors developing their own data science courses, they only provide a snapshot of a small part of the data science education landscape. As a discipline data science has yet to establish a “consensus curriculum” in the introductory course. Part of this challenge lies in the very definition: what is data science, and what should a data science student know? As an early step toward differentiating the data science student, Dichev and Dicheva (2017) proposed data science literacy as a framework requiring students to be competent in computation, statistics, machine learning, visualization, and ethics.

The common quest to define competence and literacy has also attracted the attention of professional organizations. The Park City Math Institute published a report in 2017, “Curriculum Guidelines for Undergraduate Programs in Data Science,” providing their vision for the competencies needed for a data science major and outlining suggested course divisions that would ensure that the competencies are covered (De Veaux et al. 2017). In 2018, the National Science Foundation held a Data Science Leadership Summit which made several recommendations regarding data science education. The Leadership Summit placed less focus on trying to create a core curriculum that should be included in data science, recognizing that data science degrees differ based on their placement in an institution (within an existing department, creating a new department, etc.) and the various paths that a data scientist can take (Wing et al. 2018). As a group, they made nine recommendations for future research and workshops, including hosting additional workshops with more inclusion from different colleges, universities, and academic disciplines. The EDISON Data Science Framework, published in 2017, outlined a data science competence framework, data science body of knowledge, and examples of a data science model curriculum (Demchenko, Belloum, and Wiktorski 2017). The National Academy of Sciences, Engineering, and Medicine report “Data Science for Undergraduates: Opportunities and Options” recommends that academic institutions encourage all students to develop a basic understanding of data science, and that colleges and universities should recognize and embrace data science as a new and evolving field and provides detailed recommendations regarding the curriculum and teaching approaches for data science (NASEM 2018). The NASEM report stated that a crucial role of data science education is to develop data acumen: “the ability to understand data, make good judgments about and good decisions with data, and use data analysis tools responsibly and effectively.” The National Academy also hosted a Roundtable on Postsecondary Data Science Education in 2018, which focused on developing PhD programs in data science and building models for faculty development. The Association for Computing Machinery takes a different perspective in their 2019 report: “Computing Competencies for Undergraduate Data Science Curricula.” Instead of trying to identify all the objectives that should be in a data science education, they instead only focus on the computer science objectives that would be needed in the interdisciplinary field that is data science (Danyluk et al. 2019).

Leaders in the field of data science have made their own recommendations as to what should be covered in a data science course or curriculum. In “50 Years of Data Science,” David Donoho proposes that there are 6 divisions of data science that curricula should focus on: data gathering, preparation, and exploration, data representation and transformation; computing with data; data modeling; data visualization and presentation; and science about data science. However, it is unclear whether current programs have the breadth suggested by these six categories (Donoho 2017). Hofmann and VanderPlas (2017) instead proposed that curricula should focus on the 6 steps of data analysis: data provenance, data exploration and preparation, data representation and transformation, computing with data, data modeling, and communications of results. As another example, Dhar (2013) focused on the key skills that a data scientist needs: text processing/mining, statistics, representing and manipulating data, and formulating and analyzing problems. In their 2018 report, the National Academy suggested concepts that should be taught to students to further develop their data acumen, including mathematical, computational, and statistical foundations, data management and curation, data description and visualization, modeling, workflow and reproducibility, communication, domain-specific knowledge, and ethics in problem solving (NASEM 2018). Finally, Hicks and Irizarry (2018) proposed five guiding principles for teaching data science: organize the course around a set of diverse case studies, integrate computing into every aspect of the course, teach abstraction, but minimize reliance on mathematical notation, structure course activities to realistically mimic a data scientist’s experience, and demonstrate the importance of critical thinking/skepticism through examples. Clearly the design approach to building data science courses and programs depends greatly on the instructor’s point of view and priorities. Throughout these skills, steps, principles, and divisions, there are several similar ideas and concepts: data exploration and computing, data wrangling, data visualization, data modeling, and data communication.

Though the established research and guidelines above produce recommendations for what should be covered, it is important to understand current practice in data science programs and courses. Our review identified three papers that have looked at which concepts are covered in existing data science programs, using different approaches to analyze the data. Tang and Sae-Lim (2016) looked at 30 randomly selected Data Science programs and focused on which high levels skills were covered (e.g., Communication, Mathematics, Information, Visualization) and when these skills were addressed in the curriculum. Song and Zhu (2016) looked at programs in the United States and analyzed the topic coverage based on the courses that were required within a program. West (2018) took an algorithmic approach to studying data science curricula by using word frequency analysis and clustering to analyze data science program summaries across the globe. West found five major categories in data science program descriptions: statistics, computer coding, visualization, machine learning, and applications or innovations.
2.3. Computing in Statistics… or Data Science?

Nolan and Temple Lang’s (2010) call for increased computational skills in the statistics curriculum was just one of the first: many other authors have followed suit in both statistics and data science. However, in recent years the lines between statistics education and data science education have become increasingly blurry. During the last ten years the need to help statistics, data science, and computer science educators incorporate modern programming and software resources such as R/RSstudio/RMarkdown, Python/Jupyter, and GitHub into their classrooms has been partially understood and met (Dichev and Dicheva 2017; Stander and Dalla Valle 2017; Çetinkaya-Rundel and Rundel 2018; Hicks and Irizarry 2018; Broatch, Dietrich, and Goelman 2019; Fiksel et al. 2019) and lesson plans or case studies used in data science programs are often shared with the wider community (Loy, Kuiper, and Chihara 2019).

Given the recent growth of data science programs, and the increased emphasis on computing that has come with it, a discussion of the impact Nolan and Temple Lang’s article has had on the field of statistics education would be incomplete without a comparison to the “new kid on the block.” In this article, we present results from a survey of data science faculty in multiple disciplines. Faculty were asked to provide insights on the challenges they’ve faced building their data science curriculum, the resources that would help them be a more successful data science instructor, the computing resources they use in the classroom, and the content of their data science courses and programs. Through this data, we hope to create a richer picture of the current “Data Science 101” curriculum and highlight how Nolan and Temple Lang’s recommendations for increased computing in the classroom have come to life in data science.

3. Methods

Faculty were invited to participate in the Data Science Faculty Survey through various mathematics, statistics, and computer science E-mail lists (ASA Section on Statistics and Data Science Education; ASA Section on Teaching Statistics in the Health Sciences; Isolated Statisticians; ACM Special Interest Group on Computer Science Education; SIAM Activity Group on Data Mining and Analytics; SIAM Activity Group on Applied Mathematics Education; Business, Industry, and Government SIGMAA; Project-NExT), social media channels, and personal invitations in early fall 2019. Survey questions included faculty background and experience teaching data science, courses and degrees offered at the faculty member’s institution, primary audience for and types of departments offering introductory data science, and course enrollment (actual or anticipated).

After providing information about the programmatic context for introductory data science, faculty were presented with 34 knowledge or topic areas, and asked to select whether each one was: (1) covered in introductory data science; (2) not covered in introductory data science, but covered elsewhere in their data science degree program(s); or (3) not covered in their curriculum (a fourth, “unknown” option was also provided). The list of topic areas was constructed based on the EDISON Data Science Framework (Demchenko, Belloum, and Wiktorski 2017), Curriculum Guidelines for Undergraduate Programs in Data Science (De Veaux et al. 2017), and the ACM Task Force on Data Science Education Draft Report (Danyluk et al. 2019). Instructors were also asked about software tools and programming languages used in their courses, assessment strategies, challenges they have faced teaching data science, and resources that could help them become a more effective data science instructor. The goals of this survey were to better understand the current state of data science education and identify new directions for the emerging field of data science education.

3.1. Respondent Demographics

We received 69 survey responses from faculty either currently teaching data science or planning to sometime in the next two years. The most common home department listed by faculty respondents was Mathematics (26), followed by Statistics (18), then Computer Science (9). Seven respondents indicated that they were from a department representing multiple disciplines, such as a Department of Mathematics and Statistics. The remaining nine respondents were from departments such as Business Analytics (2), Political Science (1), and Biology (1).

We also asked each respondent which of the following programs or courses were offered by their department:

- Introductory Data Science at the undergraduate level
- A bachelor’s degree in something else with an emphasis or concentration in Data Science
- An undergraduate minor in Data Science
- An undergraduate major in Data Science
- Introductory Data Science at the graduate level
- A master’s degree in something else with an emphasis or concentration in Data Science
- A master’s degree in Data Science

The distribution of offerings is shown in Figure 1. In our survey, none of the respondents indicated having a PhD program in Data Science. In all disciplines, the most common undergraduate offerings were the Introduction to Data Science course, followed by a minor in Data Science. Undergraduate degree programs such as a major in Data Science or an emphasis/concentration in Data Science were not uncommon, but not
offered by the majority of respondents’ schools. At the graduate level, the most common offerings in all disciplines were an Introduction to Data Science course. Graduate programs such as a master’s degree were mostly offered in Statistics or Computer Science departments, however, some of the departments in the other category were also offering master’s degrees. This speaks to the variety of programs and offerings available to future data scientists, and the interdisciplinary nature of data science.

Most faculty respondents were relatively new to teaching data science and had been teaching the course for two years or less (Figure 2). There were also some seasoned data science instructors, who had been teaching data science courses for five or more years, as well as some future instructors who planned to teach data science for the first time in the next two years.

4. Results

Data science instructors were asked about the biggest challenges they faced teaching data science, the resources that would help them become a better data science instructor, the computing languages and tools they used to teach data science, and the content of their data science courses and programs. Open-ended questions were analyzed using two-step axial coding, and the results are visualized in Figure 3.

4.1. Challenges Facing Data Science Instructors

When asked “What has been or is most challenging for you about teaching data science?,” the most common response category was the scope of the course. During coding, we defined “scope” as representing the “breadth” of the course, including topic standardization, course pacing, and building unified themes throughout the introductory data science course. Some instructors described difficulty “[n]arrowing down the vast amount of information into one semester introductory course.” Others mentioned differentiating introductory data science courses from introductory statistics courses, and frustration that a clear “consensus curriculum” has not yet been established: “Obtaining a clear definition of DS and what is required and what is recommended.” Instructors shared concerns about the time-intensive nature of teaching a new course in a new discipline, with several reporting that there isn’t enough time to cover everything they would like to in a data science course.

Another repeated theme in the faculty responses was technology. This encompassed concerns about local IT issues and using new technologies such as GitHub, but the most common concern was choosing an appropriate language for teaching data science (most commonly R or Python), and teaching coding. Instructors also described concern about the depth of topics in introductory data science courses (Depth) and stimulating critical thinking about data science. One respondent wrote that their biggest challenge was “[s]electing cases/assignments at the right level of difficulty. It’s easy to find trivial artificial problems and problems with insurmountable complexity.”

Instructors also listed student background as a challenge: specifically, that students were either unprepared for an introductory data science course in mathematics, statistics, or programming, or that students came from a diverse background in terms of home majors and academic preparation. One instructor wrote: “Teaching in a liberal arts setting, I can’t expect too
much by way of prereqs. Students come in with vastly different backgrounds.” Students were not the only ones inadequately prepared for introductory data science—many instructors had concerns about their continuing education, ability to learn new theory or software, or keep up to date with developments in data science. Instructors were also concerned about student attitudes, or affect, toward data science, in particular whether students were engaged in the course content or found it meaningful and relevant.

There were some differences in perceived challenges by respondent department. Mathematicians were more likely to indicate a desire for instructor continuing education, that the scope of the course was a challenge, and difficulty working with college and university shareholders (e.g., Curriculum committees). Statisticians indicated that teaching technology, course content, and student affect were struggles. Computer science faculty and members of combined departments were most likely to find differing student backgrounds and the scope of the introductory data science course to be challenging.

Four future data science instructors wrote about the challenges they face while developing the new course. All four of these instructors mentioned the time needed to create a new course. Other new instructors reported difficulty getting administrative approval to teach the course or approval from a college curriculum committee (shareholders) and that the “current nonstandardized nature of the content” (scope) was a challenge. Some new instructors reported that they were still learning the material themselves (instructor continuing education) or were preparing to teach it in the future (new course).

### 4.2. Resources Needed to Teach Data Science Effectively

As data science grows and expands, understanding the needs of data science instructors is important for prioritizing resource development. Instructors were asked “What resources might help you become a more effective data science instructor?” The most common theme that emerged from instructor responses was a call for more teaching resources like an activities data base or textbooks. Instructors were interested in online resources with exploration exercises, notebooks for coding, classroom activities, and engaging and relevant databases for use in the classroom. For example, incorporating active learning into the classroom through interactive, online tools was a common request. Instructors also stated that they would like to see more printed or digital textbooks that can be used as complete references for students. Some instructor responses mentioned existing resources for teaching statistics, computing, or data science such as CAUSEweb.org, the data8 group at UC Berkeley, and the teach data science blog (teachdatascience.com). One instructor wrote:

> I would love to find a big repository of messy, gritty case studies. It would be really interesting if there was a debrief instrument to use afterwards to show what sorts of results could come out of several representative approaches. My students learned best when I could give them open-ended problems and questions.

There was also a need described for continuing education resources geared toward instructors, as well as additional standards for introductory data science. Professional development workshops and journal articles were mentioned as steps that instructors were currently taking to improve their data science acumen. Several instructors said that they would like to see “[d]iscussion groups with other instructors,” a need that could be met with a formal forum or listserv for data science instructors in multiple disciplines. There was also a consistent need for additional support staff in the form of student TAs and new faculty hires. Other requested resources included help developing an accessible course for students with disabilities, establishing business and external partnerships, seeking grants to support program development, flexibility in course design, and structural changes to the curriculum.

The most common theme for computer science faculty was the desire for an activities database or resources, which was also a popular option for faculty in mathematics, statistics, and combined departments. Mathematicians were the most vocal about a set of common standards for introductory data science courses. Both mathematicians and statisticians felt that instructor continuing education would help them become better data science instructors.

Only three future data science instructors responded about the resources that would make them better data science instructors. For all three, having set standards for the introductory data science course would be a helpful resource. Two of the three mentioned a need for continuing instructor education and an activities or dataset database.

### 4.3. Computing: Programming Languages and Software Tools

Survey respondents were asked what software tools, programming languages, or other tools such as textbooks or websites they are currently using or plan to use to teach introductory data science. Responses were categorized into programming languages and software tools (see Figure 5). For programming languages, R was the most common response followed by Python.
Table 1. Programming languages used in introductory data science courses by discipline.

| Language | Mathematics | Computer Science | Statistics | Combined | Other | Total |
|----------|-------------|------------------|------------|----------|-------|-------|
| Java     | 0           | 1                | 0          | 0        | 0     | 1     |
| Julia    | 0           | 0                | 1          | 0        |       |       |
| Python   | 7           | 5                | 6          | 1        | 5     | 24    |
| R        | 14          | 1                | 11         | 5        | 5     | 36    |
| SQL      | 2           | 1                | 2          | 0        | 0     | 3     |

Choice of programming language was not dichotomous: thirteen respondents indicated using multiple programming languages to teach Introduction to Data Science, usually both R and Python. Not surprisingly, statisticians favored using R and computer scientists favored using Python (Table 1). Faculty in mathematics departments were twice as likely to use R than Python in the data science classroom, however, this may be because at smaller institutions statisticians tend to be housed in mathematics departments. Julia and Java were both used in the introductory course, however, neither was used alone. SQL was used as the sole programming language for a single course and used along with R in two others.

There was more variety in the software tools used in the data science classroom. RStudio and Jupyter were two popular options for IDEs for programming in data science. Excel was also a common choice, used in eight courses. Each of the other responses occurred in three or fewer courses, illustrating the diversity in software used in the data science classroom. Some instructors were very detailed with their feedback, mentioning specific sets of R packages such as the "tidyverse" (Wickham et al. 2019), or development environments such as RStudio, Jupyter, and Colab. Several mentioned using tools from the Google infrastructure like Google Classroom, Google Sheets, and Google Forms, which were grouped as "Google" in Figure 5.

4.4. Other Recommended Resources

Faculty also suggested a range of textbook resources, both online and in hard copy, including (in alphabetical order by title):

- *The Data Science Design Manual* (1st Edition) by Steven Skiena
- *Data Science and Big Data Analytics* by EMC Education Services
- *Introduction to Data Mining* (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpakne, and Vipin Kumar
- *Introduction to Statistical Investigations* by Nathan Tintle, Beth Chance, George Cobb, Allan Rossman, Soma Roy, Todd Swanson, and Jill VanderStoep
- *Modern Data Science with R* by Ben Baumer, Danny Kaplan, and Nick Horton
- *Naked Statistics* by Charles Wheelan
- *Openintro Statistics* by David Diez, Mine Çetinkaya-Rundel, and Christopher Barr (openintro.org)
- *Python Data Analysis* by Ivan Idris
- *The Python and Pandas Field Guide: An Introduction to Computer and Data Science* by Mark Liffiton and Brad Sheece (snakebear.science)
- *R for Data Science* by Garrett Grolemund and Hadley Wickham (r4ds.had.co.nz)
- *R Programming for Data Science* by Roger Peng (https://bookdown.org/rdpeng/rprogdatascience/)

None of the textbooks listed here were mentioned more than twice, suggesting there is not yet a "best" or widely recommended textbook for introductory data science courses. Other online resources suggested were (in alphabetical order by title):

- DataCamp: a set of online, interactive lessons for learning data science in R, Python, SQL, and other tools (https://www.datacamp.com/)
- Data Science for Statistics: a set of tutorials and case studies for introducing data science concepts in a variety of courses (available via GitHub: https://github.com/ds4stats)
- Kaggle: an online community for sharing datasets, hosting and entering data science challenge competitions, and job seeking (https://www.kaggle.com/)
- Stat2Labs: a webpage providing web applets for "more advanced" topics in statistics (http://web.grinnell.edu/individuals/kuipers/stat2labs/Labs.html)
4.5. Topics Covered in Data Science Courses and Programs

Survey respondents were asked to place each of the 34 selected topic areas into one of four categories: covered in the introductory data science course offered (or planned) at their institution, covered elsewhere in the data science curriculum at their institution, not covered at all, or unknown. One respondent skipped this section, so the maximum number of raters per topic possible is 68, and not all respondents rated every topic. Data visualization was the most commonly covered topic in introductory data science, taught by 82% of responding data science faculty (56 out of 68 respondents). Data cleaning was the next most common introductory topic, included in 75% of courses (Table 2). The next most common topics in the introductory course were professional ethics in data science, the data science lifecycle, reproducible research, and regression models. At least one of machine learning, advanced visualization, and statistical inference are discussed in about half of introductory data science courses taught by faculty responding to this survey. Data ethics and responsible data science, data curation, regression models, reproducible research, and the data lifecycle rounded out the list of topics taught by over half of the faculty respondents. Instructors were also asked to indicate topics that were not taught in their data science curriculum, either in the introductory course or in upper-level courses required for the data science degree. Four of the five most often omitted topics were related to advanced computing skills: big data infrastructures and technologies such as high-performance networks, batch and parallel processing, cloud computing, and systems engineering. Issues of data management, data security, and data storage were also often omitted from the curriculum. Additionally, the survey responses revealed several topic areas that instructors felt should be included in the data science curriculum, though not necessarily in the first course. Many computing and advanced mathematics topics fell into this category. For a full list of topics and counts, see the Appendix.

| Category                        | Description                                                                 | Frequency of inclusion |
|---------------------------------|-----------------------------------------------------------------------------|------------------------|
| Data visualization              | Exploratory data analysis: multivariate plots and graphs, summary statistics. Data cleaning and wrapping (i.e., restructuring data tables, creating new variables). | 56                     |
| Professional practices          | Data ethics and responsible data use.                                        | 43                     |
| Data management                 | Data curation and data quality.                                              | 36                     |
| Statistical methods             | Regression models: simple linear regression, least squares regression, logistic regression. | 36                     |
| Professional practices          | Reproducible research: computational reproducibility, empirical reproducibility, statistical reproducibility. | 35                     |
| Professional practices          | Data lifecycle and data collection, data quality evaluation.                | 34                     |
| Statistical methods             | Research methods: research cycle, hypothesis definition and testing.        | 28                     |
| Data architecture               | Data architecture, data types, and data formats.                            | 27                     |
| Machine learning                | Text mining: natural language processing, topic modeling, text visualization (i.e., word clouds, frequency plots). | 27                     |
| Visualization                   | Customizing data visualizations: color and composition, accessibility, "grammar of graphics". | 27                     |
| Machine learning                | Supervised learning: decision trees, Naive Bayes classification, neural networks, support vector machines (SVM), ensemble methods. | 26                     |

Table 3. Topics that are most often omitted from a data science curriculum for all faculty respondents.

| Category                        | Description                                                                 | Frequency of omission |
|---------------------------------|-----------------------------------------------------------------------------|-----------------------|
| Advanced computing              | Big data infrastructures: high-performance networks, infrastructure and services management and operations. | 30                    |
| Machine learning                | Reinforced learning: Q-learning.                                            | 25                    |
| Advanced computing              | Big data technologies: batch and parallel processing, streaming systems.    | 24                    |
| Advanced computing              | Cloud computing.                                                            | 24                    |
| Data management                 | Data governance: overall management of the availability, usability, integrity and security of data used in an enterprise. | 21                    |
| Information technologies        | Systems and data security: data anonymization, federated access control.    | 21                    |
| Data storage                    | Data storage systems, archive services, digital libraries, and their operational models. | 20                    |

4.6. Prerequisites

The prerequisite courses for introductory data science can help shine a light on the omitted topics: was a topic left out because it’s not covered in a data science program, or because it’s in the curriculum before introduction to data science? To learn more about prerequisites and co-requisites, we asked faculty members in our survey about the required courses students at their institutions should take before introductory data science. About 28% (19/69) of the faculty members we surveyed indicated that there was either a required (17 instructors) or suggested (2 instructors) computing prerequisite before taking introductory data science, and about 25% (17/69) had an introductory statistics prerequisite. This was not a disjoint group—nine faculty members indicated that both introductory computing and introductory statistics are required before taking the first statistics course. This suggests that, not only are there a variety of data science courses, but there are a variety of levels of student preparation before taking data science courses.

5. Discussion

Computing in introductory data science programs and courses takes many different forms, however, there are some common threads. Over 75% of faculty respondents surveyed indicated using at least one programming language in their data science programs: usually (but not always) Python or R. This is in-line with many industry surveys, such as the 2018 Kaggle Machine Learning and Data Science Survey, which found that 83% of Kaggle users used Python on a regular basis, and 36% used R (Mitchell 2019). In our survey, faculty respondents were more
likely to be housed in departments of mathematics or statistics, which may explain the slight preference for R in our data. We also found that many students are computing in a reproducible way through the use of user-friendly IDEs such as Jupyter for Python or RStudio for R and version-control tools like GitHub. The sheer variety of programming languages and software tools represented in Figure 5 shows that computing in the data science classroom is far from standard, much like computing practice in industry.

Computing was also often mentioned as a challenge for data science instructors, and an area where additional resources are needed. Faculty concerns about incorporating technology often surrounded language selection (perhaps the eternal debate in data science), and teaching coding. Many instructors, especially those outside of computer science, were adapting their pedagogy to teach coding for the first time. Professional development workshops like those held at the U.S. Conference on Teaching Statistics (USCOTS) and the Special Interest Group on Computer Science Education (SIGSCE) Technical Symposium are one way to address the challenge of teaching computing in data science for the first time and were often mentioned by instructors as a resource that would be welcomed. However, of course not every data science instructor attends conferences or has the resources or time availability needed to travel, so local or virtual workshops would likely be welcomed by the data science community. Instructor continuing education, especially centered around computing, is another resource that respondents indicated would help their development as data science instructors. Finally, an activities base was the top desired resource that would be useful for new and developing data science instructors. This is not a new call—in their 2018 report the National Academy recommended establishing spaces for data science instructors across institutions to share ideas (Recommendation 3.1) and creating “flexibility and incentives” to share course materials and faculty resources with the broader community (Recommendations 5.1 and 5.2). To that end, some activity data bases and resources for new instructors are already available, such as UC Berkeley’s Data Science Academic Resource Kit (https://data.berkeley.edu/education/ark) or Data Science in a Box (https://datasciencebox.org/), and as the field data science grows and expands, we can expect more of these resources to develop.

In 2010, Nolan and Temple Lang recommended three fundamental changes to the practice of statistical education:

1. Broaden statistical computing. “…statisticians must access and integrate large amounts of data via Web services and databases, manipulate complex data (e.g., text, network graphs) into forms more conducive to statistical analysis, and produce interesting statistical presentations of data.”

2. Deepen computational reasoning and literacy. “…must be able to express themselves through computations, understand the fundamental concepts common to programming languages, and discuss and reason about computational problems precisely and clearly.”

3. Compute with data in the practice of statistics. “…statistical computing… should be taught in the context of statistical practice to give students both the motivation to interact with data and the experience needed to be successful in their future statistical endeavors. The nature of computing with data needs to be addressed by data acquisition, statistical analysis, and reporting.”

There can be no question that data science courses, and statistics courses themselves, have expanded to broaden the role of statistical computing in the classroom through introducing modern technologies (R, Python, GitHub) and techniques (data wrangling, text analysis, simulation-based inference). Moreover, for the courses that we reviewed the instructors we surveyed all recommended using real data in context, or discussed the need for additional infrastructure and resources to aid introducing such examples. It is unclear the extent to which computational reasoning and literacy have been incorporated into data science courses and programs. The nature of doing computation in the classroom requires students to be familiar with concepts like debugging, code formatting, and reproducible programming. However, are we truly developing students who understand R, Python, or any of the other computing languages used to teach data science “think”?

Ten years ago, data science was a blip on the radar for many statistics instructors. The only place the phrase “data science” appears in Nolan and Temple Lang is in a caption of a figure. In Figure 1 of their article, the authors arrange computational topics relevant to statistical practice into six major topic areas:

1. Fundamentals in scientific computing with data
2. Information technologies
3. Computational statistics (e.g., numerical algorithms) for implementing statistical methods
4. Advanced statistical computing
5. Data visualization
6. IDEs

Figure 1 from Nolan and Temple Lang is reproduced below, with some new additions (Figure 6). In the revised figure, topics that were covered in more than 75% of data science curricula, either in the introductory course or elsewhere in the program, are boxed in blue, while topics covered in more than 50% of data science curricula boxed in orange. Topics that were omitted from our survey are grayed out.

It is perhaps not surprising to see that most topics recommended by Nolan and Temple Lang are taught in many currently existing data science courses and programs. For example, nearly all computing topics in data visualization were covered in at least 50% of the data science courses and programs in our survey. Computational statistics was also heavily represented, although one should note that Bayesian computation and representation of numbers were not included in our list of topic areas. We believe that Bayesian statistics, at least at a cursory level, may be included in several data science programs. IDEs are also discussed in data science curricula, especially version control, authoring tools such as RMarkdown, and reproducible computation. Other topic areas such as advanced computing and web technologies seemed to receive little coverage.

There are some areas of Nolan and Temple Lang’s diagram that are notably lacking. Based on our survey, we were unable to explicitly measure the coverage of web technologies such as Flash, HTTP, and XML in the data science courses and curricula.
taught by our survey respondents. Our survey did include “Data acquisition through web scraping and/or API calls,” which was one of the most popular topics listed by our respondents. However, with the introduction of software like the “rvest” package in R (Wickham 2020) to make web scraping more easily accessible to novice programmers, we hesitate to make conclusions about the coverage of web technologies in data science courses. Other topic areas like parallel and distributed computing, which were included in the survey, were often not included in courses mentioned. This is likely because these are topics often reserved for upper-division computer science courses, and most of the programs and courses we collected data on consisted of an introductory course or a minor in data science. Several topics in the diagram, such as debugging and programming scope, were not explicit learning objectives in our survey but are typically addressed “along the way” in a data science course.

Nolan and Temple Lang asked, “how well are we, as a community of statistics educators, preparing our students for this modern era of statistics research and practice?” As our community expands to include not just statistics, but data science, we should understand the current state of practice in existing courses and programs, and the challenges students face. In many ways, data science’s diversity means we are probably far from establishing a consensus curriculum for “Data Science 101;” however, a set of common topics including data visualization, data wrangling, the essentials of statistical modeling and machine learning, and reproducible research and programming appears to have emerged at many colleges and universities.

One question left unanswered is this: How does the computational training needed for statisticians differ from the computational training needed for data scientists? Does it? In our survey, we saw a range of course and program descriptions ranging from computationally infused statistics to computer science with a dash of statistics. Many data scientists would agree that data science as a discipline falls somewhere in between, but where is not yet clear.
There are some limitations of this work. Data were collected through invitations shared by e-mail and on the authors’ social media and professional networking websites and is not a representative sample of all data science faculty. Furthermore, since there is no designated professional network or accrediting body for data science programs and curricula, we cannot know for certain which faculty or departments are underrepresented or overrepresented. Many of our respondents were housed in departments of mathematics, however, at many smaller colleges and universities computer science and statistics faculty are still located in mathematics departments, so this may not be a misrepresentation. As a reference, only nine of our respondents were in a computer science department, two were in a department of business analytics, and six others were in other disciplines from the arts and the sciences. While this speaks to the diversity of programs and disciplines involved in teaching data science, a more comprehensive picture of the discipline would emerge from a larger sample. Since data science is still in its youth as a discipline, we cannot reliably use a faculty member’s educational background as a measure of their preparation for teaching data science. In academia and in industry, data scientists often refer to massive open online courses (MOOCs), textbooks, or websites to get their first experience with what data science is, and the techniques a data scientist uses.

Appendix

Number of respondents covering selected topics in introductory data science courses, later within data science programs, or not covered in the data science curriculum for all disciplines.

| Description                                                                 | Intro covered | Program covered | Not covered | Unknown | N | Percent covered in intro | Percent covered in Program |
|-----------------------------------------------------------------------------|---------------|-----------------|-------------|---------|---|--------------------------|---------------------------|
| Regression models: simple linear regression, least squares regression, logistic regression. | 36            | 27              | 1           | 2       | 66 | 95%                      | 55%                       |
| Exploratory data analysis: multivariate plots and graphs, summary statistics. | 57            | 5               | 2           | 2       | 66 | 94%                      | 86%                       |
| Data cleaning and wrangling (i.e., restructuring data tables, creating new variables). | 52            | 7               | 6           | 0       | 65 | 91%                      | 80%                       |
| Research methods: research cycle, hypothesis definition and testing.         | 28            | 33              | 3           | 3       | 67 | 91%                      | 42%                       |
| Traditional statistical inference: t/z-based hypothesis tests, confidence intervals. | 16            | 45              | 4           | 3       | 68 | 90%                      | 24%                       |
| Data ethics and responsible data use.                                        | 44            | 15              | 3           | 4       | 66 | 89%                      | 67%                       |
| Linear algebra: matrix manipulation, eigenvalues, singularity.               | 8             | 50              | 8           | 2       | 68 | 85%                      | 12%                       |
| Simulation-based inference: bootstrapping, randomization tests.              | 22            | 36              | 7           | 3       | 68 | 85%                      | 32%                       |
| Data architecture, data types, and data formats.                             | 28            | 28              | 9           | 2       | 67 | 84%                      | 42%                       |
| Reproducible research: computational reproducibility, empirical reproducibility, statistical reproducibility. | 36            | 18              | 5           | 7       | 66 | 82%                      | 55%                       |
| Experimental design, modeling, and planning.                                 | 17            | 39              | 9           | 3       | 68 | 82%                      | 25%                       |
| Machine learning (supervised): decision trees, Naive Bayes classification, neural networks, support vector machines (SVM), ensemble methods. | 27            | 28              | 9           | 4       | 68 | 81%                      | 40%                       |
| Relational and non-relational databases; SQL, NoSQL, data warehouse solutions, ETL, OLTP, OLAP. | 12            | 40              | 10          | 4       | 66 | 79%                      | 18%                       |
| Customizing data visualizations: color and composition, accessibility, “grammar of graphics.” | 28            | 25              | 8           | 6       | 67 | 79%                      | 42%                       |
| Data curation and data quality.                                              | 36            | 15              | 7           | 7       | 65 | 78%                      | 55%                       |
| Machine learning (unsupervised): clustering algorithms, principal components analysis (PCA), singular value decomposition (SVD), independent components analysis (ICA). | 20            | 32              | 14          | 2       | 68 | 76%                      | 29%                       |
| Text mining: natural language processing, topic modeling, text visualization (i.e., word clouds, frequency plots). | 28            | 23              | 15          | 2       | 68 | 75%                      | 41%                       |
| Advanced data visualizations: creating animated or dynamic visualizations, maps. | 22            | 27              | 11          | 6       | 66 | 74%                      | 33%                       |
| Data acquisition through web scraping and/or API calls.                      | 24            | 25              | 9           | 9       | 67 | 73%                      | 36%                       |
| Applications software requirements and design.                               | 10            | 38              | 11          | 7       | 66 | 73%                      | 15%                       |
| Graph analysis: paths, social networks.                                      | 15            | 33              | 12          | 8       | 68 | 71%                      | 22%                       |
| Data lifecycle and data collection, data quality evaluation.                 | 35            | 11              | 12          | 8       | 66 | 70%                      | 53%                       |
| Optimization and numerical algorithms (EM).                                   | 10            | 36              | 17          | 5       | 68 | 68%                      | 15%                       |
| Data protection, backup, and privacy.                                        | 24            | 18              | 14          | 10      | 66 | 64%                      | 36%                       |
| Information systems and collaborative systems.                               | 7             | 33              | 16          | 10      | 66 | 61%                      | 11%                       |
| Systems engineering and software engineering principles.                    | 4             | 35              | 21          | 7       | 67 | 58%                      | 6%                        |
| Systems and data security: data anonymization, federated access control.     | 13            | 26              | 21          | 7       | 67 | 58%                      | 19%                       |
| MCMC simulation.                                                             | 5             | 32              | 20          | 11      | 68 | 54%                      | 7%                        |
| Cloud computing.                                                             | 6             | 28              | 24          | 9       | 67 | 51%                      | 9%                        |
| Data storage systems, archive services, digital libraries, and their operational models. | 8             | 25              | 20          | 12      | 65 | 51%                      | 12%                       |
| Big data technologies: batch and parallel processing, streaming systems.     | 2             | 31              | 25          | 9       | 67 | 49%                      | 3%                        |
Acknowledgments

The author team thanks the anonymous peer reviewers for their thoughtful comments and suggestions.

References

Albert, J., and Rizzo, M. (2016), “An Undergraduate Data Science Program,” in *JSM 2016 Proceedings*, pp. 1135–1147.

Asamoah, D. A., Doran, D., and Schiller, S. (2020), “Interdisciplinarity in Data Science Pedagogy: A Foundational Design,” *Journal of Computer Information Systems*, 60, 370–377, DOI: 10.1080/08874417.2018.1496803.

Baumer, B. (2015), “A Data Science Course for Undergraduates: Thinking With Data,” *The American Statistician*, 69, 334–342, DOI: 10.1080/00031305.2015.1081105.

Broatch, J. E., Dietrich, S., and Goelman, D. (2019), “Introducing Data Science Techniques by Connecting Database Concepts and dplyr,” *Journal of Statistics Education*, 27, 147–153, DOI: 10.1080/10691898.2019.1647768.

Danyluk, A., Leidig, P., Buck, S., Cassell, L., Doyle, M., Ho, T. K., McGettrick, D., Donoho, D. (2017), “50 Years of Data Science,” *Annual Review of Statistics and Its Application*, 4, 309–339, DOI: 10.1146/annurev-statistics-060118-045239.

Četinkaya-Rundel, M., and Rundel, C. (2018), “Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum,” *The American Statistician*, 72, 58–65, DOI: 10.1080/00031305.2017.1397549.

Danyluk, A., Leidig, P., Buck, S., Cassell, L., Doyle, M., Ho, T. K., McGettrick, A., McIntosh, S., Qian, W., Schmitt, K., Servin, C., Stefik, A., Wang, H., and Wittenbach, J. (2019), “ACM Data Science Task Force Draft Report,” available at [https://dsf.acm.org/](https://dsf.acm.org/).

Donoho, D. (2017), “50 Years of Data Science,” *Journal of Computational and Graphical Statistics*, 26, 745–766, DOI: 10.1080/10618600.2017.1384734.

De Vaau, R. D., Agarwal, M., Arerrett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., and Kim, A. Y. (2017), “Curriculum Guidelines for Undergraduate Programs in Data Science,” *Annual Reviews of Statistics and Its Application*, 4, 15–30, DOI: 10.1146/annurev-statistics-060116-053930.

Demchenko, Y., Belloum, A., and Wiktorski, T. (2017), “Edison Data Science Framework: Part I. Data Science Competence Framework,” available at [http://edison-project.eu/data-science-competence-framework-ctds](http://edison-project.eu/data-science-competence-framework-ctds).

Dhar, V. (2013), “Data Science and Prediction,” *Communications of the ACM*, 56, 64–73, DOI: 10.1145/2500499.

Dichev, C., and Dicheva, D. (2017), “Towards Data Science Literacy,” *Procedia Computer Science*, 108, 2151–2160, DOI: 10.1016/j.procs.2017.05.240.

Fiksel, J., Jager, L. R., Hardin, J. S., and Taub, M. A. (2019), “Using GitHub Classroom to Teach Statistics,” *Journal of Statistics Education*, 27, 110–119, DOI: 10.1080/10691898.2019.1617089.

Hicks, S. C., and Irizarry, R. A. (2018), “A Guide to Teaching Data Science,” *American Statistician*, 72, 382–391, DOI: 10.1080/00031305.2017.1356747.

Hofmann, H., and VanderPlas, S. (2017), “All of This Has Happened Before. All of This Will Happen Again: Data Science,” *Journal of Computational and Graphical Statistics*, 26, 775–778, DOI: 10.1080/10618600.2017.1385474.

Loy, A., Kuiper, S., and Chihara, L. (2019), “Supporting Data Science in the Statistics Curriculum,” *Journal of Statistics Education*, 27, 2–11, DOI: 10.1080/10691898.2018.1564638.

Majumder, M., and Cheng, X. (2017), “Focusing on the Needs: Experiences of Developing a Data Science Program,” *Journal of Computational and Graphical Statistics*, 26, 779–780, DOI: 10.1080/10618600.2017.1385475.

Mitchell, M. (2019), “Programming Languages for Data Scientists,” available at [https://towardsdatascience.com/programming-languages-for-data-scientists-adfe2af5cc5](https://towardsdatascience.com/programming-languages-for-data-scientists-adfe2af5cc5).

National Academies of Sciences, Engineering, and Medicine (2018), “Data Science for Undergraduates: Opportunities and Options,” The National Academies Press website, available at [https://doi.org/10.17226/25104](https://doi.org/10.17226/25104).

Nolan, D., and Temple Lang, D. (2010), “Computing in the Statistics Curriculum,” *The American Statistician*, 64, 97–107, DOI: 10.1198/tast.2010.09132.

Song, I.-Y., and Zhu, Y. (2016), “Big Data and Data Science: What Should We Teach?” *Expert Systems*, 33, 364–373, DOI: 10.1111/exsy.12130.

Stander, J., and Dalla Valle, L. (2017), “On Enthusing Students About Big Data and Social Media Visualization and Analysis Using R, RStudio, and RMarkdown,” *Journal of Statistics Education*, 25, 60–67, DOI: 10.1080/10691898.2017.1322474.

Tang, R., and Sae-Lim, W. (2016), “Data Science Programs in U.S. Higher Education: An Exploratory Content Analysis of Program Description, Curriculum Structure, and Course Focus,” *Education for Information*, 32, 269–290, DOI: 10.3233/EFI-160977.

West, J. (2018), “Teaching Data Science: An Objective Approach to Curriculum Validation,” *Computer Science Education*, 28, 136–157, DOI: 10.1080/08993408.2018.1486120.

White, D. (2018), “A Project Based Approach to Statistics and Data Science,” *Primus*, 79, 1–25, DOI: 10.1080/10511970.2018.1488781.

Wickham, H. (2020), “rvest: Easily Harvest (Scrape) Web Pages,” R Package Version 0.3.6, available at [https://CRAN.R-project.org/package=rvest](https://CRAN.R-project.org/package=rvest).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., and Kuhn, M. (2019), “Welcome to the Tidyverse,” *Journal of Open Source Software*, 4, 1686, DOI: 10.21105/joss.01686.

Wing, J. M., Janeja, V. P., Kloetkorn, T., and Erickson, L. C. (2018), “Data Science Leadership Summit: Summary Report,” National Science Foundation, USA, available at [https://dl.acm.org/doi/book/10.5555/3293458](https://dl.acm.org/doi/book/10.5555/3293458).

Yan, D., and Davis, G. E. (2019), “A First Course in Data Science,” *Journal of Statistics Education*, 27, 99–109, DOI: 10.1080/10691898.2019.1623136.

Yavuz, F. G., and Ward, M. D. (2020), “Fostering Undergraduate Data Science,” *American Statistician*, 74, 8–16, DOI: 10.1080/00031305.2017.1407360.

Zheng, T. (2017), “Teaching Data Science in a Statistical Curriculum: Can We Teach More by Teaching Less?” *Journal of Computational and Graphical Statistics*, 26, 772–774, DOI: 10.1080/10618600.2017.1385473.