ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

ROLE OF DATA SCIENCE IN EDUCATION

GADIREDDY RAJASEKHAR REDDY
Assistant Professor
Dept. of Computer Science and Engineering
V.K.R, V.N.B & A.G.K College of Engineering
GUDIVADA
Krishna Dt. Andhra Pradesh

G.V.N.KISHORE
Associate Professor
Dept. of Computer Science and Engineering
V.K.R, V.N.B & A.G.K College of Engineering
GUDIVADA
Krishna Dt. Andhra Pradesh

Abstract
Data science has been growing in prominence across both academia and industry, but there is still little formal consensus about how to teach it. Many people who currently teach data science are practitioners such as computational researchers in academia or data scientists in industry. To understand how these practitioner-instructors pass their knowledge onto novices and how that contrasts with teaching more traditional forms of programming, we interviewed 20 data scientists who teach in settings ranging from small-group workshops to large online courses. We found that: 1) they must empathize with a diverse array of student backgrounds and expectations, 2) they teach technical workflows that integrate authentic practices surrounding code, data, and communication, 3) they face challenges involving authenticity versus abstraction in software setup, finding and curating pedagogically-relevant datasets, and acclimating students to live with uncertainty in data analysis. These findings can point the way toward better tools for data science education and help bring data literacy to more people around the world.

Introduction
People across a wide range of professions now write code as part of their jobs, but the purpose of their code is often to obtain insights from data rather than to build software artifacts such as web or mobile apps. Although programmers have been analyzing data for decades, in recent years the popular term *data science* has emerged to encapsulate this kind of activity. Data scientists are now pervasive throughout both industry and academia: In industry, it is a fast-growing job title across many sectors ranging from technology to healthcare to public policy. In academia, data scientists are often STEM graduate students, postdocs, and technical staff who write code to make research discoveries.

Despite its blossoming across many fields of practice, data science has only recently begun to formalize as an academic discipline, so there is still little consensus on what should go into a data science curriculum. Many novice data scientists are currently learning their craft and associated technology stacks on the job from expert practitioners rather than from full-time teachers. To understand how these *practitioner-instructors* pass their knowledge onto novices and what challenges they face, we conducted interviews with 20 data scientists (five men and fifteen women) who teach in both industry and academic settings ranging from small-group workshops to large online courses. Our participants come from backgrounds

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

ranging from the life sciences to the behavioral sciences to the humanities; none have formal degrees in computer science.

We chose to study practitioner-instructors because they are the ones defining both the technical and cultural norms of this emerging professional community. Their insights can inform the design of new programming tools and curricula to train this growing population of diverse professionals who are responsible for making advances across science, technology, commerce, healthcare, journalism, and policy. While prior work has studied what data science practitioners do on the job to our knowledge, we are the first to systematically investigate how they teach their craft to junior colleagues and students.

Our study extends the rich lineage of HCI research on how people learn programming to pursue different career goals. On one end, there is a long history of studies on teaching computer science and engineering skills to those who aspire to become professional software engineers; on the other end, there is a parallel literature on the learning needs of end-user programmers. Data scientists are a distinct and so-far understudied population in between those two extremes: They share similarities with both software engineers (they aspire to write reusable analysis code to share with their colleagues) and end-user programmers (they view coding as a means to an end to gain insights from data).

We found that data science instructors must empathize with a diverse array of student backgrounds and expectations. Also, despite many of their students viewing coding as merely a means to an end, they still strive to teach disciplined workflows that integrate authentic practices surrounding code, data, and communication. Finally, they face challenges involving authenticity versus abstraction in software setup, finding and curating pedagogically-relevant datasets, and acclimating students to cope with uncertainty in data analysis.

These findings can point the way toward the design of specialized tools for data science education, such as block-based programming environments, better ways to find and synthesize datasets that are suitable for teaching, and fostering discussions around data ethics and bias.

In sum, this paper's contributions to HCI are:

- A synthesis of the technical workflows that data science practitioners teach to novices, along with challenges they face in teaching. These findings advance our understanding of a growing yet understudied population in between end-user programmers and professional software engineers.
- Design implications for specialized tools to facilitate data science education.

Our study was inspired by prior work in end-user programming, teaching data science, practitioners as instructors, and broadening computing education to learners who do not self-identify as programmers.

Data Science and End-user Programming

Data science is a broad term that encompasses a wide variety of activities related to acquiring, cleaning, processing, modeling, visualizing, and presenting data. Although data visualization is a highly active area of HCI research, what is more relevant to our study is prior HCI research on programming as performed by non-professional programmers.

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

Kandel et al. found great variation in levels of programming ability amongst data scientists. Many of them write code in languages such as Python and R, but they are not professional software engineers; moreover, many do not even have formal training in computer science. Much of data scientists' coding activities can be considered end-user programming since they often write code for themselves as a means to gain insights from data rather than intending to produce reusable software artifacts. Related terms for this type of insight-driven coding activity include exploratory programming (from Kery et al. and research programming (from Guo's dissertation).

However, as we discovered in our study, modern data scientists are not merely writing ad-hoc prototype code. They are now developing increasingly mature technology stacks for writing modular and reusable software. In the terminology of Ko et al., they are now engaging in *end-user software engineering* with more of an emphasis on code quality and reuse; in Segal's related terminology, data scientists are now becoming *professional end-user developers*. Along these lines, software engineering researchers such as Kim et al. have studied the role of data scientists within industry engineering teams.

In contrast to prior HCI work that focuses on what data science practitioners do on the job, our study instead focuses on how they pass on those skills to novices via teaching.

Teaching Data Science

Data science is now a highly in-demand subject within both academia and industry: Many universities are launching new data science majors, research labs are organizing hands-on workshops, and MOOCs and coding bootcamps focused on data science are some of the most popular offerings. But despite this growing interest over the past few years, there is still little agreement on what a data science curriculum should contain.

To our knowledge, there does not yet exist a systematic research study on how data science is currently being taught. The only publications on this topic are course design guides and experience reports of how instructors have taught *specific* courses within their own fields. These papers fall into two categories: descriptions of courses taught by computer science (CS) faculty, and those taught by faculty in other disciplines. CS faculty have written about their experiences teaching data science both to enrich introductory computing courses with data-oriented applications and in courses intended to serve non-CS-majors. And faculty in fields ranging from bioinformatics, business, and statistics have written field guides on teaching data science in their respective majors. In particular, data science within statistics curricula places more of an emphasis on computational workflows and tools rather than on theoretical aspects of the underlying mathematics.

Outside the classroom, instructors have also documented their experiences teaching in informal settings. For instance, the Software Carpentry and Data Carpentry organizations hold workshops to teach computing and data analysis to academic researchers; they also publish course design guides. Related groups have organized data-oriented hackathons, hack weeks, and apprenticeships to train academic researchers in data science best practices.
In contrast to the aforementioned experience reports, to our knowledge, ours is the first academic research study that attempts to provide a broad overview of how modern data science is taught by practitioner-instructors across both industry and academia—synthesizing findings in a way that transcends anecdotal experiences within individual courses.

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

Practitioner-instructors

Most of our participants were practitioner-instructors: data science practitioners who also teach students. Practitioner-instructors are often found in settings such as medical schools (clinical faculty), art schools, business schools, and law schools, where they are sometimes known as professors of the practice. Two noted benefits of learning from practitioners are that they are likely up-to-date on the latest tools in their field and that they are more direct members of the *community of practice* that their students aspire to join. However, they often lack formal pedagogical training: Wilson refers to them as *end-user teachers* (as an analogue to end-user programmers) since they teach but are not formally trained as professional teachers. To our knowledge, researchers have not yet studied practitioner-instructors in computing-related settings such as data science.

Computing Education for Broader Populations

Our study contributes to the growing body of HCI and computing education work on teaching programming to broader learner populations. Specifically, it extends prior work that target people who do not self-identify as programmers.

Although much of computing education research targets learners who aspire to become computer science majors or professional programmers, there is a growing body of studies on learners with other professional identities. For instance, Ni et al. studied the challenges faced by high school teachers who are learning programming in order to become CS teachers. Dorn et al. studied graphics and web designers who identify more as artists. Chilana et al. studied industry professionals in non-programming roles (e.g., sales, marketing, product management) who try to learn programming to communicate better with their engineering colleagues. Dasgupta and Mako Hill extended the Scratch blocks-based programming environment to enable K-12 children to perform analysis on data generated by members of the Scratch online community; although children do not yet have professional identities, they are able to use Scratch programming as a conduit to develop computational and data-oriented thinking skills. What all of this work has in common is that it focuses on teaching programming to learners who do not self-identify as programmers.

Along similar lines, the instructors we interviewed self-identified as data scientists, data analysts, researchers, or more generally, the umbrella term “scientist”; since their students are junior members of their peer groups, they would also likely identify as such. To our knowledge, we are the first to characterize the challenges involved in teaching the topic of data science in diverse professional settings. Some of our findings corroborate those of prior work on how programming is perceived as a means to an end rather than as something to be intrinsically enjoyed for its own sake.

Teaching Data-Analytic Workflows

Although many students viewed coding as a means to an end (see prior section), nonetheless instructors emphasized teaching a more disciplined data-analytic workflow using a modern stack of open-source tools (e.g., Figure **1**). In other words, they did not simply want students to create one-off scripts but rather wanted to provide them with the skills to write more robust and reproducible scientific code.

As instructors walked through the technical contents of what they taught, we noted the most salient points they raised that differed from what is typically taught in CS-oriented programming courses. Most notably, these instructors emphasized workflows that centered

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

on the integration of code, data, and communication rather than on the more algorithmic foundations of computing.

Undergraduate data science education is currently offered in many forms, and this variability is expected to continue in the near future. Common modalities include the following:

a) Introductory exposure to data science, through a single inspirational course that could satisfy a general education requirement;
b) Major in data science, including advanced skills, as the primary field of study;
c) Minor or track in data science, where intermediate skills are connected to the major field of study;
d) Two-year degrees and certificates;
e) Other certificates;
f) Massive open online courses (MOOCs), which can engage large numbers of students at a variety of levels; and
g) Summer programs and boot camps, which can serve to supplement academic or on-the-job training.

As academic institutions add courses and programming around undergraduate data science, they will need to decide what modalities are institutionally appropriate, considering many factors such as student demand, faculty and institutional strengths and resources, and curricular fit. These choices may also be influenced by the existence of graduate programs in data science at the institution. Each of these modalities—with its strengths, limitations, and possible areas for improvement—is discussed in more detail in the following sections.

Major in Data Science

Data science majors are emerging across academic institutions and will continue to do so in years to come. Similar to introductory data science experiences, there is significant variation in program structure, goals, and content in these majors as well. Some data science majors are emerging as independent programs that interface with specific domain areas, while others are emerging as specializations within a given domain area.

The most common features of current data science majors include required courses in mathematics, statistics, and computer science. Within mathematics departments, requirements often include the following:

- Mathematics courses on linear algebra, calculus, and discrete structures;
- Statistics courses on introductory statistics, probability, and various kinds of applied statistics; and
- Computer science courses on database systems, programming, data structures, algorithms, and machine learning.

Some majors have courses listed as “data science,” including cross-listed courses with statistics and computer science, while other data science majors draw entirely upon courses from connected departments. Common topics taught under the “data science” listing include advanced data analytics, big data, data mining, simulation modeling, and computational thinking. Many data science majors include required or elective courses from outside the core departments—commonly economics, business, psychology, biology, and geography or geosciences. Many data science majors also require a hands-on practicum or capstone course to help reinforce skills.

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

Conclusion

Currently, many 4-year majors fall into one of three categories: (1) data science majors housed within a college or school of business (i.e., programs in business analytics, which usually involve more marketing and finance classes and fewer computational and mathematics courses); (2) data science/analytics majors housed in a mathematics or statistics department (i.e., above-average mathematics or statistics requirements with fewer "core" computational courses); and (3) data science programs housed in a computer science department as either a stand-alone major or as a concentration to information technology (i.e., more computational courses but potentially fewer "core" mathematics courses). Variations in courses offered and required within similarly labeled majors at different institutions are notable. A few 4-year undergraduate data science majors are hybrids of these three models, being administered jointly by multiple departments.

References

1. Adhikari A, DeNero J. Computational and Inferential Thinking: The Foundations of Data Science. 2018. https://www.amherst.edu/academiclife/departments/courses/1718F/STAT/STAT-231-1718F.
2. Bay-Williams J, Duffett A, Griffith D. "Common Core Math in the K-8 Classroom: Results from a National Teacher Survey.". 2016. [March 29, 2018]. https://eric.ed.gov/?id=ED570138.
3. CCAC (Community College of Allegheny County). "Data Analytics Technology (788): Associate of Science.". 2018. [March 29, 2018]. https://www.ccac.edu/Data_Analytics_Technology.aspx.
4. Cha SH. Exploring disparities in taking high level math courses in public high schools. KEDI Journal of Educational Policy. 2015;12(1):3–17.
5. Chuang I, Ho A. "HarvardX and MITx: Four Years of Open Online Courses—Fall 2012-Summer 2016.". 2016. [April 1, 2018]. http://dx.doi.org/10.2139/ssrn.2889436.
6. Dondero M, Muller C. School stratification in new and established Latino destinations. Social Forces. 2012;91(2):477–502. [PMC free article] [PubMed]
7. Feldon DF, Jeong S, Peugh J, Roksa J, Maahs-Fladung C, Shenoy A, Oliva M. Null effects of boot camps and short-format training for PhD students in life sciences. Proceedings of the National Academy of Sciences. 2017;114(37):9854–9858. [PMC free article] [PubMed]
8. Fine E, Handelsman J. Brochure prepared for the Women in Science and Engineering Leadership Institute. 2010. "Benefits and Challenges of Diversity in Academic Settings." http://wiseli.engr.wisc.edu/docs/Benefits_Challenges.pdf.
9. Finzer W. The data science education dilemma. Technology Innovations in Statistics Education. 2013;7(2):1–9.
10. Gamoran A. Tracking and inequality: New directions for research and practice. In: Apple MW, Ball SJ, Gandin LA, editors. The Routledge International Handbook of the Sociology of Education. New York: Routledge; 2009. pp. 213–228.
11. Jones C. EdSource. Feb 19, 2018. [March 22, 2018]. "Big data" classes a big hit in California high schools. https://edsource.org/2018/big-data-classes-a-big-hit-in-california-highschools/593838.
12. Lleras C. Race, racial concentration, and the dynamics of educational inequality across urban and suburban schools. American Educational Research Journal. 2008;45(4):223–233.

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

13. Lucas SR. Tracking Inequality: Stratification and Mobility in American High Schools. New York: Teacher's College Press; 1999.
14. Lucas SR, Berends M. Race and track location in U.S. public schools. Research in Social Stratification and Mobility. 2002;25:169–187.
15. Montgomery College. “Data Science Certificate: 256.”. 2018. [January 25, 2018]. http://catalog.montgomerycollege.edu/preview_program.php?catoid=8&poid=1877&returnto=1322.
16. Nashua Community College. “Why Choose Foundations in Data Analytics?”. 2018. [March 29, 2018]. http://www.nashuacc.edu/academics/associate-degrees/stem-and-advancedmanufacturing/398-foundations-in-data-analytics.
17. Oakes J. Keeping Track: How Schools Structure Inequality. New Haven, Conn.: Yale University Press; 2005.
18. UC Berkeley (University of California, Berkeley). “Data 8: Foundations of Data Science.”. 2018. [January 25, 2018]. http://data8.org.
19. UC San Diego (University of California, San Diego). “Data Science Undergraduate Program.”. 2017. [January 25, 2018]. http://dsc.ucsd.edu.
20. UW (University of Washington). “Calling Bullshit” makes an impact at schools across the country. 2017. [February 22, 2018]. https://ischool.uw.edu/news/2017/10/calling-bullshit-makes-impactschools-across-country.