Highlights of the National Academies Report on “Undergraduate Data Science: Opportunities and Options”

Laura Haas, Alfred Hero, Robert A. Lue

Published on: Jun 22, 2019

Updated on: Oct 04, 2019

DOI: [10.1162/99608f92.38f16b68](https://doi.org/10.1162/99608f92.38f16b68)

ABSTRACT

In 2016, the National Academies of Sciences, Engineering, and Medicine of the United States established the Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective. The committee issued a 120-page report in 2018, setting forth a vision for the undergraduate education in data science. This interview of the co-chairs of the committee, Laura Haas, the Dean of the College of Information and Computer Sciences at University of Massachusetts at Amherst, and Alfred Hero, the co-director of the Michigan Institute for Data Science, is conducted by Rob Lue, the co-editor of data science education for *Harvard Data Science Review*. It provides a succinct summary of the key findings and recommendations of the Report, highlighting the call to equip students with “data acumen.” It also probes beyond the Report, including the possible roles of data science for reimagining liberal arts education in the digital age.

Keywords

Data Acumen, Data Ethics, Data Science Curriculum, Data Science Programs, Digital-age Education, Liberal Arts Education

Rob: Laura, let me start with you. You and Al chaired the committee that wrote the "Undergraduate Data Science" report. What were you hoping to accomplish when you took on this role?

Laura: I hoped that we could clarify for academic institutions, industry, and students alike what data scientists of the future would need to know, give them some insight into the emerging landscape of academic programs, and help academic institutions, in particular, understand the key considerations in designing an undergraduate program. Data science is a new field, and as a result, there is a lot of confusion about what it is, what skills and knowledge are needed (or should be expected). Many academic institutions, of all sorts and all levels, are starting programs in data science, or are already offering them. But those programs cover different topics and emphasize different skills today. Meanwhile, there is huge demand for “data scientists” – but different employers may be looking for different skill sets. I hoped we could provide guidance on what programs should offer and employers could then look for, as well as examples of what different institutions were providing. This study also offered an opportunity to further define the field, by defining the knowledge practicing data scientists would need.

Rob: Can you be specific? What knowledge is needed, and how does this differ from traditional undergraduate programs (e.g., computer science or statistics)?

Laura: The key element is what we call “data acumen.” Getting meaningful, correct, and useful answers from data requires skills that are typically not fully developed in traditional mathematics, statistics, and computer science (CS) courses. Students need exposure to a range of concepts, including mathematical, statistical, and computational foundations, but also such topics as data description and visualization, data modeling and assessment, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving. They particularly need to spend time working with real-world data, and with problems that can reinforce the limitations of tools, and demonstrate the ethical considerations that permeate many applications.

Rob: How did you arrive at your conclusions? Who was on the committee, and what process did you follow?

Al: We benefited from a diverse core committee of 17 members that represented a range of organizations including educational institutions, foundations, and industry. The educators on the committee were from a mix of public and private research universities and four-year colleges. All of the committee members had been involved in either setting up data science programs, developing data science courses, or hiring data scientists – so we collectively covered most of the pipeline, from incoming students to outgoing graduates to the workforce. The committee had a remarkable breadth of disciplinary backgrounds including computer science, statistics, engineering, natural science, social science, and the humanities. Thus, discussions were lively and often took unexpected turns that gave insights. For example, reflections by a mathematical ecologist on the committee resulted in shifting the concept of an educational “pipeline” to ‘a watershed,’ providing a better metaphor given the varied motivations and backgrounds of students of data science. I think that everyone on the committee emerged from these discussions with a richer perspective on data science—which is reflected in the report.

In terms of process and arriving at our conclusions, over a period of about 18 months we organized our activities into four stages, each involving collecting and distilling a lot of data from data science leaders and public records. In the first half of the study, we concentrated on finding out about current programs by meeting with people leading the programs, collecting preliminary findings, and framing the questions we wanted to address in the remainder of the study. These questions were incorporated into an interim report, which is available on the National Academies website, and included topics such as how to instill data acumen, how to enhance interaction with two-year colleges, and how to foster diversity, inclusion, and increasing participation. Next, we organized an interactive webinar series to solicit responses from the wider public on these questions. Each of the nine one-hour webinars was moderated by a committee member and featured two expert presenters from outside the committee. These webinars were quite successful, attracting hundreds of attendees and resulting in interesting discussion during the Q&A. Finally, we reconvened the committee to brainstorm on the information

collected in the interim report and from the webinars, and met one final time to evolve our collective thinking, come up with recommendations, and draft an outline of the report, which was finished a few months later. Overall, the process ran very smoothly thanks to the National Academies' excellent staff who provided crucial support.

Rob: Why should someone read the report? What will they learn?

Al: The report should be of interest to anyone who wants to learn about the evolving and diverse profiles of data scientists, wants to understand the term "data acumen," mentioned earlier by Laura, or wants to launch a data science program. I believe that for the informed reader, many of the findings and recommendations will confirm their beliefs, like the need for students to see that data science is a general approach to problem solving and not just a set of tools or methods to be applied out of the box. Other findings, like the need to focus on data science ethics and on broadening participation in the field (and classroom) may be less familiar. I think that the reader will appreciate the wealth of data provided in the report including many reference citations, concrete examples, tabulations, and summarizations. The report addresses the different modalities of undergraduate data science education, including majors, minors, two-year degrees, massive open online courses (MOOCs), boot camps, and summer camps—with thoughts on the role that these may play in the future of data science education. You will see that the report raises more questions than it answers, but they are interesting questions that will promote additional discussion.

Rob: What were the key findings of the report?

Laura: The central finding, Finding 2.3, is a definition of the aforementioned data acumen. Data acumen is at the heart of what a practicing data scientist must know, and therefore, what a data science program must teach. While this is the central finding, others are also important. We found that many different forms of data science education and multiple pathways will help students from a variety of backgrounds—educational and demographic—to succeed at levels ranging from basic to expert. The many types of problems to which data science is applicable will be appealing to diverse students, and this range of educational opportunities will be critical in engaging them, and beneficial to the industry as a whole. But we also found that instructional flexibility will be critical in putting together programs, as existing faculty may lack the broad expertise to cover the varied aspects of data science, and existing courses may not fully cover the material needed - or may cover it with too much overlap or from different disciplinary perspectives. Hence, in the short term, it may be necessary to team-teach classes, or to piece together courses to form a coherent program until new classes can be developed and faculty hired or trained. We found that programs will need to be nimble and evolve rapidly over time as the field evolves, and we learn what educational approaches are most effective.

Finally, we found that data science can be a powerful tool in evaluating and guiding the evolution of data science programs. Administrative records, demographic information, and economic and survey data can be analyzed both for individual programs and comparatively across programs to transform program elements or entire programs to better serve their students. Thus, evaluation should drive evolution, in an iterative process. Professional societies can also play a strong role in enabling faster progress in both data science and data science education.

Rob: What challenges were identified that are faced by those trying to launch data science programs?

Al: As someone who helped launch the programs at the University of Michigan, I am quite familiar with the challenges and also with some of the ways to overcome them. First of all, it is important to realize that there is tremendous demand for programs in data science at all levels – students are attracted by the balanced multidisciplinary curriculum with its fresh approach to integrating fundamental concepts and practical applications. As a result, we thought that those running existing statistics or computer science programs may view data science as a threat to their own pipelines. However, a more serious issue is that the faculty in statistics and computer science are often stretched too thin to devote time to outside interests like a putting together a course or program in data science. University leadership may also be reluctant to give faculty release time or to hire faculty for a new untested program. A related challenge is that a new program may initially have to rely on other departments to provide some of the courses. Data science students are motivated and smart—they can handle core courses in CS or statistics—so that is not the problem. However, some of the most popular courses, like machine learning, often have caps that restrict the number of outside students. Hence, it is important to get new versions of these courses created and rolled out quickly in the new program—or else work upfront with the university leadership to agree to relaxing enrollment caps. Other challenges are mentioned in the report, in particular the need to develop curricula and teaching styles that make data science accessible to a broader student population. Data science has tremendous potential to become a uniquely inclusive discipline with wide appeal.

Rob: What challenges were identified in recruiting and retaining students to these programs?

Laura: The challenges are very similar to those faced by any STEM discipline, with the additional challenge that data science is a new and not well-known field. Many students who do know of it are scared of the strong mathematical and computing competencies needed, or the breadth of knowledge and skills required; others are concerned about what types of positions they could get when they complete the program; still others just haven't been exposed to what data science is and what impact it can have. That said, data science offers real opportunities to recruit a large and diverse student body. Once students understand what it is about, they often respond very positively to the potential for impact, and broad "cultural relevance" that data science offers. Programs that highlight the

applications and achievements have attracted a good mix of students, and in fact, some of the courses of study have experienced dramatic growth in enrollments in a very short time. The best practices include designing the material to lower the barriers to entry; choosing project topics to be of broad interest, in particular emphasizing applications across humanities, social sciences, business and the arts; providing teams of tutors and laboratory assistants to help students needing assistance; emphasizing teamwork; and presenting data science as a life skill and cultural pursuit, not just as a science or engineering discipline.

Rob: Are there particular recommendations you would like to highlight?

Al: As mentioned previously, data science education is truly a multi-disciplinary endeavor and will require broad-based support from several departments. These departments all have different ways of thinking about data science, different priorities, and unique perspectives born out of decades of disciplinary history. I view one of the report's recommendations as especially important in this respect: the recommendation that the professional societies of the core data disciplines convene regularly as a group to share ideas, best practices and data. This recommendation has already taken flight—a recent leadership summit I attended at the Alan Turing Institute in August 2018 adopted this recommendation as an action item and the National Academies is planning a kind of summit meeting of the leading professional societies in statistics, computer science, applied math, engineering, and social sciences, among others.

Laura: The committee felt strongly that not only should all data science students be exposed to ethics, but that it should be learned and practiced throughout the curriculum. In fact, a key recommendation was that the data science community should adopt a code of ethics which should be affirmed by our professional societies and conveyed through both professional development and educational programs. I'm pleased to say that efforts along these lines are also already in progress by several professional societies and organizations of data scientists.

Rob: Given the all-encompassing nature of data science, does your committee see an opportunity here to use data science as an overarching theme to organize liberal arts education in the digital age?

Laura: We didn't really discuss this— it's a fascinating idea, though.

Al: As Laura points out, the report does not address this question, but in our discussions the committee did support incorporating data literacy as a general requirement for all undergraduate students, supplementing writing, communication, and numeracy requirements that are common today. The data scientist of the future (2040) is wistfully imagined in Chapter 1 and is painted as having a well-rounded education, exposed to concepts through diverse motivating applications and expansive

experiences. Instilled with an appreciation of ethics, the humanities, music, and other arts, the graduate in 2040 will be a skilled communicator and creative problem solver. I personally resonate with the concept that all students, regardless of major, should have some degree of data acumen, a healthy skepticism about conclusions drawn from data, the ability to recognize good data from bad, an awareness of the need for data fairness and privacy, and familiarity with some of the tools, e.g., elementary exploratory data analysis and visualization.

The committee recognized the high potential of data science to attract broad and diverse students into the field. As I mentioned above, one of our webinars was on this topic and pointed out how data science may need to be framed differently to reach a broader audience. For example, one of the webinar speakers emphasized that learning has both a cognitive and a non-cognitive component. Often it is the non-cognitive component that enables student success: instilling that magical ‘aha’ moment can make all the difference to student motivation and learning. I think data science is a prime proving ground for developing holistic ways of teaching that optimize learning of concepts, practices, and consequences, instead of simply focusing on mastering the tools. I have observed that many of my own data science students are interested in ‘empathetic problem solving’—formulating approaches that incorporate understanding of the impact of solutions on society and on people's lives. Data science gives us the opportunity to develop a different approach to teaching that goes beyond skill-building in computing, databases or statistics.

Rob: We often hear from employers (e.g., industry, governmental agencies, NGOs) that they wish the students they hire had better soft skills, such as communication skills, interpersonal skills, stress management skills, etc. What does your committee see as effective ways to integrate this kind of professional development into the data science curriculum?

Laura: As we looked at emerging programs, we saw a number of options. Some programs had required courses in communications, writing or even ‘professionalism.’ Many programs featured a capstone project that gave the student an opportunity to follow the whole process from requirements gathering to results presentation, often with a real customer—an industry partner, a government agency, or a professor or staff group on campus. I think in an ideal world, students should have both a course and an experiential opportunity to best hone their skills.

Rob: This has been a great discussion. Any final thoughts for our readers?

Al: The ground is fertile for sustained growth of data science into a thriving discipline. I think that data science programs will become a big melting pot, attracting a more diverse body of students into the sciences and into an expanding job market. The report gives a starting point for thinking about these opportunities but it really has only touched the surface, raising many more questions than answers. The data we uncovered and the broad tent of energized people I met while co-chairing this

study makes me very optimistic for the future of data science. I wish to close my comments with my gratitude to the staff at the National Academies, especially Michelle Schwalbe and her team.

Laura: I agree with Al that I expect that data science will continue to evolve and thrive as it attracts more diverse students and faculty. While our report outlines many of the challenges ahead, the responses to those challenges will be shaped by academicians and the market for skills. I hope the emerging field will continue to share information and take a data-driven approach to its evolution. Like Al, I am profoundly grateful to Michelle Schwalbe and the staff at the National Academies for their heroic efforts on this complex study.

Rob: I am sure our readers will share my deep appreciation of your great leadership in, insights of, and visions for data science education, in addition to your great accomplishments as research scholars. I look forward to a future dialogue where we can assess the impact of the Report and the lessons learned about promoting its key concept of “data acumen.” Until then, thank you, and let’s get busy!

Acknowledgement

The National Academies Study on undergraduate data science education, co-chaired by Laura Haas and Alfred Hero, was supported by Award No. 1626983 from the National Science Foundation (Directorate for Computer and Information Science and Engineering; Directorate for Education and Human Resources; Directorate for Mathematical and Physical Sciences/Division of Mathematical Sciences; and Directorate for Social, Behavioral and Economic Sciences).

This article is © 2019 by Laura Haas, Alfred Hero, and Robert Lue. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (<https://creativecommons.org/licenses/by/4.0/legalcode>), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the authors identified above.