Harvard Data Science Review • Issue 3.1, Winter 2021

How Can We Train Data Scientists When We Can't Agree on Who They Are?

Usama Fayyad[1,2] Hamit Hamutcu[3]

[1]Applied Machine Learning Research Group, Institute for Experiential Artificial Intelligence, Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, United States of America,

[2]International Secure Systems Lab, Northeastern University, Boston, Massachusetts, United States of America,

[3]Initiative for Analytics and Data Science Standards, Institute for Experiential Artificial Intelligence, Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, United States of America

Published on: Feb 25, 2021

DOI: https://doi.org/10.1162/99608f92.0136867f

As the demand for data science talent has exploded, so have the efforts to train data science professionals. There are many programs and formats for training in data science, ranging from short online courses to full-time undergraduate and graduate degree programs. The article "Statistics Practicum: Placing 'Practice' at the Center of Data Science Education" by Kolaczyk et al. (2020, this issue) presents a great deal of insight into the challenge of designing such a program at one of the prominent academic institutions in the United States. In our opinion, the article makes some distinctions about program design that will surely prove to be very useful for others who are on the journey to building or enhancing their own, particularly the importance of a practicum-type training in statistics education anchored to actual consulting services with real customers and real data.

As the title of this discussion article suggests, we will expand on this challenge by posing a critical question that should be top-of-mind in designing education and training programs for data science. As there is not yet an agreed-upon definition of who data scientists are and which skills and knowledge they need to have, designing programs or developing curricula is challenging. On the other hand, organizations in industry are often not able to articulate their expectations from data science talent clearly, which in turn makes hiring, managing, and developing data professionals mostly inefficient and ineffective.

In order to start providing answers to the question in the title, we will rely heavily on our work and research at Initiative for Analytics and Data Science Standards[1] (IADSS) and insights from a workshop organized by IADSS at the Knowledge Discovery and Data Mining (KDD) 2020 conference that focused on this exact challenge of training data science professionals. In the second section of the discussion, we will argue that the practicum idea can be even further expanded to a residency-type program with intensive and immersive work on real and current problems, inclusive of problematic data challenges and issues with access and completeness of data. Although our thoughts and findings are driven primarily from an industry perspective, we will try to provide 'student perspective' at the end as we see it is equally challenging for people who are interested in developing their knowledge and skills in data science to find the best path for their learning and career goals.

Understanding Roles and Skills in Data Science

So really, who are data scientists and what are they expected to know?

In the first article (Fayyad & Hamutcu, 2020) authored as part of our IADSS activities to present a framework for the knowledge and skills required in data science, we note that "although 'data scientist' has emerged as a job title, every industry, function, and business appear to be looking for their definition of the role and that universities have responded to the demand for data scientists by creating schools, institutes, and centers and establishing degree programs for relevant disciplines. These suffer from the same confusion—some are housed in business schools and others are established within computer science departments, some are cross-disciplinary, and others are considered specializations of more established disciplines. Some institutes hire a

dedicated faculty and admit their own students; others provide joint appointments to existing faculty members of traditional departments (Gorman & Klimberg, 2014)." It was refreshing to see that Kolaczyk et al. underline that the program "was designed from the ground up—starting tabula rasa—rather than as a modification or perturbation of the traditional M.A. in Statistics program we had already had for roughly 30 years." We believe this is a significantly more effective way of responding to the unique challenges raised by the field of data science rather than creating a mix of existing courses from relevant departments such as statistics and computer science.

The most apparent of these unique challenges is the undisputed multidisciplinary nature of data science. So much so that the industry came to believe and accept that it takes a 'unicorn' to master all necessary knowledge and skills relevant to data science practice. The range of topics covered or touched upon by the Statistics Practicum program in the article is a good example: computing, methods and modeling, statistical theory, data types and sources, communication (writing, speaking, and emails), data provenance/cleaning/manipulation, data visualization, statistical modeling and inference, data confidentiality/privacy/security, big data, data engineering, systems for automated data collection and management, predictive analytics, and artificial intelligence are listed as topics covered in the program. In the industry, realistically, such a broad set of skills can only be acquired at a generic level, or they can be expected from a single person only in very small organizations (Garten, 2018; Jung, 2020). Similar challenges exist for educational institutions; how can a program cover this broad range of topics at a reasonable level of depth, in a limited amount of time? This challenge is even greater for programs with shorter durations.

Given how popular data science is, the unicorn idea poses a great challenge for the community: unicorns are rare, in fact, they do not exist, as Davenport (2020) puts it. We strongly embrace the thinking that data science is an umbrella term and consists of activities more complex than a single professional, the data scientist, can perform (Irizarry, 2020). This line of thinking quickly brings us to a 'data science team,' made up of individuals with clear role definitions and specialties who collectively meet the skill and knowledge requirements of the organization. Our forthcoming article focuses on breaking down our Knowledge Framework into clearly defined individual roles such as 'Data Analyst,' 'Data Scientist,' or 'Data Engineer.' We are also planning to provide a guide to building specific roles in an organization based on these key definitions, creating a more standardized professional field.

This way of thinking about distinct roles in the industry with distinct sets of skills and knowledge might also be appropriate for training programs to adopt. While it would certainly be useful for a data science student to understand the entire data science lifecycle to start with, it would be more effective to then pick and choose certain areas of specialty to gain a level of expertise that would be useful to employers as they recruit data science talent. This would enable them to match the right person to the right job with potentially a quicker transition into the role. For example, in the Statistics Practicum program, each project team is made up of 10 to 15 students, and it might be possible to organize the project team around core roles that are typically observed

in the industry based on students' interests and support them with knowledge and skills most relevant to perform that role within the team. One might venture to guess that this might be occurring naturally within the student teams ,with each student working on a part of the problem most consistent with their skillset or area of interest.

Another important role of formal education in data science should be ensuring graduates possess certain fundamental skills that are necessary to be an effective data science professional, regardless of role. These include core skills in math and statistics (such as descriptive statistics, probability basics), computer science (such as data structures and databases), and scientific method (such as formulating a problem, research methods). In the article, the authors mention the Statistics Practicum program requires incoming students to have completed "two semesters of calculus, a semester or more of programming, and at least one of training in statistics." Students are also "expected to have had at least one of either a semester of discrete probability or a semester of linear algebra." The program also features a "2-week boot camp, focused largely on helping establish a relatively level playing field in terms of background in mathematics, statistics, and computing." While it would be difficult to ensure competency in these areas for incoming students without a thorough assessment, this set of prerequisites still enables designing a curriculum that can focus on more advanced topics and the practical application of knowledge to problems. We believe, through our interactions with data science employers, that these foundational skills are also the hardest to learn on the job, and there are significant risks to organizations if data science professionals without rigorous understanding of these fundamentals engage in data science work to drive decisions.

The KDD-2020 IADSS Workshop and Thoughts on Apprenticeship Training

In a KDD 2020 conference workshop organized by IADSS, titled "2nd Workshop at KDD on Data Science Standards—What do you need to know as a Data Scientist? Training Data Scientists of the Future," participants discussed at length the best ways of equipping data science students with the knowledge and skills that the industry needs. Co-chaired by Usama Fayyad and Xiao-Li Meng (Professor at Harvard University and Founding Editor-in-Chief at *Harvard Data Science Review*), the workshop covered several subtopics from curriculum design to academic–corporate partnership models. Panel participants Jeannette Wing (Director of the Data Science Institute at Columbia University), Pavlos Protopapas (Scientific Program Director at Harvard Institute of Applied Computational Science), and Kjersten Moody (Chief Data Officer at Prudential Financial) quickly converged on the necessity of rigor in designing a strong academic program. There was agreement that presenting the full data science picture would only be possible in longer duration programs, such as master's or undergraduate degree programs, and that shorter programs such as boot camps might be good options to develop skills in specific areas within data science. Another point brought up by Wing was that if companies sponsor students, then they can work on real-life cases throughout the year and not just in the capstone project, similar to the core idea in the Statistics Practicum program. In a short presentation, Rayid Ghani (Professor at

Carnegie Mellon University) presented a 'data science residency' approach, similar to the medical profession, where the "masters program is structured around a core of applied project work supported by lectures and workshops," again supporting the benefits of not limiting real case work to just a final project at the end of the program.

The workshop seemed to have reached a unanimous conclusion on the benefits of an apprenticeship-based program, taking the practicum idea even further. As the article by Kolaczyk et al. outlines, practicum is already a standard approach in "education, psychology, public health, social work, and others." We furthermore believe that adding expert practitioners to guide students as mentors through this residence-like process would enable students to gain an appreciation for real-world issues such as limitations of learning algorithms and oversensitivity of algorithms to data quality.

One note we would like to add is related to the discussion of "Assessing Practicum Success" in Section 6.2 of the article. Measuring the success of the core enabler of the program, that is, the consulting service, is an innovative approach. The success of the consulting service/projects can be measured objectively by tracking customer renewal rates and appetite of customers to pay. Also, we would suggest that the metrics mentioned, for example, placement rate of students being 95% after 6 months, needs to be normalized against typical statistics graduates' placement rate as well as difference in compensation of the practicum graduates against the baseline.

Comparing Programs and Student Perspective

In another short presentation at the KDD workshop, Tom Davenport (Professor at Babson College) underlined that the confusion and variety in the contents of training programs make it hard for students to even know which one to apply to.

This final point on the significant variety in the way programs are designed was of interest to us at IADSS as well over the last year. It will become obvious to anyone who reviews publicly available information on just a few programs that comparing the curriculum of one program to another is very difficult. Even when you pick a specific type of program, say a one-year graduate degree in data science, each program will have its own approach to what to teach, making it challenging for candidates to understand which skills and knowledge they will exactly gain by attending one. Course names can be arbitrary and only a detailed analysis of syllabi for all the mandatory and elective courses (which can amount to a significant number spread across multiple schools and departments) might reveal what the program covers. Even if you had this information readily (program websites rarely have this level of detail), it is a major challenge for someone who is just starting on the journey to become a data science professional to maneuver through this maze of hard-to-decipher information. We should note that this situation is equally frustrating for employers for the same reason: Unless you spend a significant amount of time understanding the particulars of each program, you would not know what kind of skills and knowledge a job candidate might offer an organization upon graduating from a specific program.

Naming of programs certainly does not make this confusion any less. Two programs carrying the same name might offer fairly different curricula, or you might find similar curricula in two programs that are named completely differently and housed in different schools. The Statistics Practicum program detailed in the article is a good example, since it might not be obvious to a potential student or an employer that the program is in fact entirely focused on data science.

There are discussions in academia about standardizing data science curriculum, but these are mostly in smaller groups and in very early stages. However, they are certainly useful for sharing best practices in curriculum design and benefit from other institutions' experiences in this rapidly evolving space.

In order to better understand the student perspective and develop a potential proposal to make comparing programs an easier task, we ran two limited-scale research studies in the weeks leading up to our KDD workshop. In one study, we asked more than 150 recent graduates of data science training programs, ranging from boot camps to degree programs, about their experience before, during, and after training. Findings confirm the difficulty in obtaining information about programs, which seems to lead candidates to pick a program mainly based on the institution's reputation. Surveys also included questions about their satisfaction with the program and how it contributed to their subsequent job search. In the second study we asked program administrators to map their curriculum to the IADSS Knowledge Framework and mark each knowledge area with the level of coverage in the program. In summary we asked them to tell us whether a particular subject, for example, Bayesian statistics, was covered as a dedicated mandatory course, one of many topics in a mandatory course, dedicated elective course, one of many topics in an elective course, or not covered at all. We received 17 responses that, although not sufficient to draw any general conclusions, enabled us to create a heat map that provides quick insight into the commonalities and differences between programs. We believe this exercise of 'normalizing' curriculum against a standardized framework of knowledge areas would make understanding focus areas of different programs a much easier exercise and could benefit students and curriculum designers as well as organizations looking to hire graduates from these training programs. We plan to scale both research studies in 2021 and will share results with the data science community in due course.

Efforts to educate data scientists of the future will certainly evolve with the field over the coming years, and communication and collaboration between academia and industry will be important. We also believe standardization of industry roles and training curricula are critical enablers of the goal to ensure that the growing global need for data science professionals can be met effectively and efficiently.

Disclosure Statement

Usama Fayyad and Hamit Hamutcu have no financial or non-financial disclosures to share for this article.

References

Davenport, T. (2020). Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review*, *2*(2). https://doi.org/10.1162/99608f92.55546b4a

Fayyad, U., & Hamutcu, H. (2020). Toward foundations for data science and analytics: A knowledge framework for professional standards. *Harvard Data Science Review*, *2*(2). https://doi.org/10.1162/99608f92.1a99e67a

Garten, Y. (2018, November 6). The kinds of data scientist. *Harvard Business Review*. https://hbr.org/2018/11/the-kinds-of-data-scientist

Gorman, M. F., & Klimberg, R. K. (2014). Benchmarking academic programs in business analytics. *Interfaces*, *44*(3), 329–341. https://doi.org/10.1287/inte.2014.0739

Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.dd363929

Jung, J. (2020, May 26). Machine learning engineer vs data scientist (Is data science over?): Vs data analyst vs research scientist vs applied scientist vs…*Medium*. https://towardsdatascience.com/mlevsds-3c89425baabb

Kolaczyk, E., Wright, H., & Yajima, M. (2020). Statistics practicum: Placing "practice" at the center of data science education. *Harvard Data Science Review*, *3*(1). https://doi.org/10.1162/99608f92.2d65fc70

Footnotes

1. The reader can find more information at www.iadss.org. ↩