Taking Our Medicine: Standardizing Data Science Education With Practice at the Core

Kit T. Rodolfa^{1,2} Rayid Ghani^{1,2,3}

¹Center for Data Science and Public Policy, Computer Science Department and Harris School of Public Policy, University of Chicago, Chicago, Illinois, United States of America,

²Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America,

³Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

Published on: Feb 25, 2021

DOI: <https://doi.org/10.1162/99608f92.64bbf5af>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

There is a parallel between medicine and data science: where medicine reflects the practical application of more fundamental and theoretical work in biology, physiology, and anatomy, data science is at its core the practice of applying theories and methods from computer science, machine learning, and statistics. Yet, practice is too often absent (or at best a limited after-thought) in our approach to educating data scientists. Before standardization and introduction of the current residency model, early medical education suffered from a similar problem: the amount of practical experience (if any) was entirely at the whim of students' mentors even to the extent that, according to (Long, 2000), "young surgeons might never have performed an operation."

In their new article, Kolaczyk and colleagues ("Statistics Practicum: Placing 'Practice' at the Center of Data Science Education," this issue) describe a program they have been developing at Boston University over the last several years that seeks to improve data science education by taking a practice-centric focus. Making data science education more effective is something we're both passionate about, and in fact we recently proposed a degree program with many aspects that align strongly with the program described here (Rodolfa et al., 2019). Effective data science practitioners need to develop competencies across several domains: problem definition, data preparation, analysis, evaluation, ethics, and communication. Unfortunately, 'standard' programs that focus on theory, methods, and technical details are not enough and poorly suited to building many of these competencies. By placing practical experience and real-world problem-solving at the center of their program, the Boston University team is creating a more holistic, well-rounded education that seems very likely to greatly benefit their students as they transition into their careers.

Several strengths of their program seem worth emulating for other data science educators. The practice-first focus doesn't just mean having students work on real projects from early in the program, but informs every aspect of the program itself, from the nature, structure, and sequence of coursework to how students interact with one another as well as with advisors, faculty, and external partners. Organizing the entire program around the data science project lifecycle seems particularly key here, and certainly corresponds to our own observations that concepts seem much more likely to stick with students when they are delivered in a 'just-in-time' manner that is well-aligned with students encountering those same issues in their own project work. That such an approach can require an iterative introduction and revisiting of concepts, tools, and frameworks (reflecting the iterative nature of projects themselves) seems clear, yet can be ill-suited to the traditional structure of distinct, concept-focused, semester- or quarter-long courses with a linear progression. Breaking with coursework-centric paradigm may be challenging in some settings, but strikes us as potentially one of the most important and impactful innovations presented here. Such innovations can often only be achieved when creating a new program, rather than tweaking existing ones. The authors are also well served by seeking out and embracing the diversity of backgrounds among their students, composing teams that balance different strengths and levels of expertise to improve their effectiveness at addressing the varied aspects of real projects. Finally, it's notable that Kolaczyk and colleagues use the term "partners" rather than "clients" to describe the relationship with the organizations that provide access to problems and data. We've been following the same

practice in our own programs for many years, and see this as a small but very important distinction as experts in the partner organizations serve a critical and very active role in the educational experience.

A key set of decisions data science master's programs struggle with is how long to make the program, what to require as prerequisites for admission into the program, what skills to teach in the program, and what to prepare the students to be able to do on completion. These tradeoffs are difficult to manage and require explicitly prioritizing a program's goals. For the program at Boston University, they chose to make the length of the program nine months and decided to have a fairly minimal set of prerequisites for students to enter the programs. Those two design choices raise some important questions about how well the program is able to provide an end-to-end project experience, cover certain topics in depth, and prepare students for a wide variety of careers. It may not be feasible to keep the program as short as 9 months without strong entrance requirements while a longer program can accommodate a wider variety of backgrounds and career aspirations.

Since there is considerable diversity in the fields to which data science students might end up applying their skills, programs must strike a balance between developing students' ability to understand and incorporate domain expertise in one particular area and developing a technical toolkit that cuts across applied domains. The program described by the authors seems to lean toward the latter, with students working on several projects in the course of their degree. The strengths of such an approach would seem to be exposure to a greater diversity of questions and methods and (hopefully) giving students a chance to start to see patterns in how similar questions might arise in widely different contexts. However, there remains a critical role for acquiring and using domain knowledge in the effective application of data science: understanding the right questions to ask through project scoping and formulation (and the impact they can have on desired outcomes), making sense of the available data through data exploration, gut-checking the reasonableness of analytical results and iteratively debugging, defining and measuring algorithmic fairness, and effectively deploying a system and integrating with existing stakeholders and decision-making practices all rely heavily on a deep understanding of the context and nature of the problem. How best to jointly develop both technical data science skills and the ability to acquire and apply domain expertise—particularly in the context of a master's program spanning only a few semesters—is no easy task. Perhaps exploring different program tracks would be fruitful here (that is, for students interested in working in different areas of industry, public service, or academic disciplines) but may also involve considerable logistical overhead and very close partnerships with other academic departments.

The relatively short length of both the overall program as well as each project here raises another question about the degree to which students here are exposed to the full data science lifecycle—from initial scoping and data acquisition through analysis to field validation, deployment, and monitoring. In our experience, developing and running the [Data Science for Social Good summer fellowship](#) at various universities, we have found the 3 months of intensive project work very valuable for students, but the relatively short timeframe of their involvement means projects must be well-scoped and data available before students arrive and they have little opportunity to see the projects through a full pilot or field trial, let alone deployment. The pairing of

external partner work with smaller consulting projects described by the authors seems to be an interesting approach to attempt to balance breadth and depth, but we wonder if there might still be some room to explore the right balance between the two types of practical experience.

As the authors point out, assessment seems to be particularly challenging here. While it does seem somewhat surprising that there doesn't appear to be any cumulative exercise (such as a thesis or oral defense), we can certainly appreciate the difficulty of developing a uniform mechanism of assessing student learning and mastery given the diversity of projects and questions they will participate in during a practice-centered program. Assessment of the program as a whole is likewise challenging. And, although the authors cite some encouraging anecdotal evidence around job placement, self-reported satisfaction, and alumni involvement, we still know little about how effective this program is with respect to the counterfactual of a more 'typical' data science master's program. We are, of course, strongly inclined to believe in the value of their approach given our own anecdotal experience, but certainly also hope that as a growing number of universities adopt more practice-centered approaches in their data science degree programs, that the field will find better opportunities for measuring and understanding what works well at fostering effective data scientists.

Overall, the program developed by Kolaczyk and colleagues is an encouraging step toward practice-first education in data science and worth building upon. Like the early days of medical education, approaches to data science training currently feel far too haphazard and disjointed. Our field is in need of better training approaches and increased standardization, perhaps even through consolidated efforts to set standards and develop a system of accreditation to ensure adherence. There are still many questions to answer to understand the impact of different design choices around length, format, scope, and content, and we hope that the recognition that deep practical experience is critical in data education can create an opening for a coalition of interested programs to come together, learn from these experiences, and systematically explore these questions. By putting practice at the core of their experience, these programs can lead the way in defining the training standards and illustrating best practices for educating effective data science practitioners.

Disclosure Statement

Kit T. Rodolfa and Rayid Ghani have no financial or non-financial disclosures to share for this article.

References

- Long, D. M. (2000). Competency-based residency training. *Academic Medicine*, 75(12), 1178–1183. https://journals.lww.com/academicmedicine/Fulltext/2000/12000/Competency_based_Residency_Training_The_Next.9.aspx
- Rodolfa K. T., De Unanue, A., Gee, M., & Ghani, R. (2019). An experience-centered approach to training effective data scientists. *Big Data*, 7(4), 249–261. <https://doi.org/10.1089/big.2019.0100>

©2021 Kit T. Rodolfa and Rayid Ghani. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.