Harvard Data Science Review • Issue 3.2, Spring 2021

The Building Blocks of Statistical Education in the Data Science Ecosystem

Alison L. Gibbs¹ Nathan Taback¹

¹Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Published on: Jun 07, 2021

DOI: <https://doi.org/10.1162/99608f92.8bb28793>

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

It Starts With a Question

The philosopher Karl Popper (2005, p. 3) contends that “the natural as well as the social sciences always start from *problems*, from the fact that something inspires *amazement* in us, as the Greek philosophers used to say.” John Tukey (1962) regarded “data analysis ... as a science ... defined by a ubiquitous problem rather than a concrete subject,” where “quantification in an ever wider variety of disciplines” is one of the major influences on the science of data analysis. Similarly, data science can be defined by the ubiquitous problem of learning from data. If you have a hypothesis that you want to pursue, you can often test it by looking at data. But you start with having a question.

The UC Berkeley undergraduate initiatives in data science education have had a far-reaching influence on undergraduate data science initiatives at other colleges and universities. Adhikari et al. (2021) have provided an engaging overview of the evolution of Berkeley’s undergraduate data science curriculum, and the important role that “computational and inferential thinking” plays in their curriculum. Inculcating students with this type of thinking brings together “complementary problem-solving styles associated with computer science and statistics.” We agree with their assertion that “data science focuses on real-world problems involving data and decisions, [and] [t]hat these problems are not generally the province of a single discipline.” But what does that mean for our traditional disciplines, especially statistics, which plays a major role in almost all data science problems?

Adhikari et al. pose the question: “Should the new data science curriculum entirely displace classical curricula in computer science and statistics, or should it simply live side by side with those curricula?” We considered this question as we renovated the undergraduate statistics curriculum over the past few years in the Department of Statistical Sciences at the University of Toronto. If classical curricula in statistics means formal statistical theories, then our answer to the question Adhikari et al. posed is that some parts associated with data science should be added and some classical parts retired. Overall, this will increase the scope of modern statistics compared with what we were used to teaching. And adding to an already large curriculum might worry readers. But such concern is not new. In fact, “Tukey anticipated such objections, by pointing out that biochemistry textbooks seem to cover much more material than statistics textbooks; he thought that once the field commits to teaching more ambitiously, it can simply ‘pick up the pace’” (Donoho, 2017). This ambition, motivated by the desire to add parts of data science to our statistics curriculum, catalyzed the recent overhaul of our undergraduate statistics program.

Students at the University of Toronto have been flocking to statistics programs in droves over the past few years. Our students enroll in their program of study at the end of their first year, so our counts include students in their second, third, and fourth year of study. In 2008 this was fewer than 400 students; by 2018 it was 4,400 students. Why the increase? Were students signing up for statistics to learn about the Neyman-Pearson lemma or ancillary statistics? As much as we might love these elegant results and realize their theoretical importance,

we fully recognized that the answer is no. Most students were enrolling in statistics programs in combination with in-depth study in another discipline, out of an interest in solving problems that involve data. They wanted to be data scientists. And we grappled with what was important in their education as a statistician, to prepare them to be functioning and contributing residents in the data science ecosystem (Meng, 2019).

The Building Blocks in the Renovation of an Undergraduate Program in Statistical Sciences

With these considerations in mind, in 2016 we set out to reconsider our undergraduate programs of study in statistics. Over many years we had tinkered with our programs in a variety of ways, adding some courses and including more authentic experiences with real data. But when we asked ourselves how we would describe our ideal graduate, who would be the statistician making significant contributions in the ecosystem of data science? and do our programs produce such graduates? we came to the conclusion that tinkering was no longer a sufficient option.

Borrowing the home renovation metaphor many have adopted, we had moved beyond a cosmetic makeover to a renovation that needed some structural work, including a new entrance and additional rooms. Nevertheless, it is a renovation as opposed to a new build, so inevitably there are limitations. But the foundation is in place.

The building blocks of our renovation are:

1. *Statistical theory, including probability.* This maintains its historical foundational role in our program, and interested students can choose to structure their program so that it is a focus, built on the foundation that all of our students experience.
2. *Methods and applications.* While this was always core to our program, and the focus of much of the home redecorations over the years, we imagined a program with a better balance of description, explanation, and prediction (Shmueli, 2010).
3. *Computing with data.* Computation is now an integral part of every course, and not just for data analysis. Statistical algorithms and their computational complexity have been brought into the discussion, so students understand how statistical output was generated. Data wrangling is intentionally included, so that data processing and variable definitions are explicit. And simulation is used for multiple purposes, including inference, evaluating methods, and exploration of theoretical concepts.
4. *Professional practice.* We want all students to develop skills in communication, teamwork and collaboration, the application of ethical principles, and to appreciate and experience how data, statistical methods, and predictive algorithms are used in society.
5. *Problem-solving.* Our students need broad skills to think critically about the strengths and limitations of their work and take what they have learned and apply it in new settings.

These building blocks reflect both historical strengths and newly important considerations and emphases. In particular, there is new emphasis that we are creating problem solvers, who are not just experts in a host of statistical methods. Our goal is to produce graduates who can adapt the expertise they acquire in statistical and computational methods to new contexts that have different problems and sources of data. However, important fundamental ideas, such as data provenance, confounding, the role of randomization, sampling, and overfitting, are not lost. Students in our statistics program cannot avoid each of these themes, but the relative emphasis may depend on choices they make depending on their personal interests.

The New Entrance

Adhikari et al. described an introductory course (Data 8) that sets the stage for their new data science curriculum. A key addition in the University of Toronto Statistics renovation was our own new version of an introductory course, STA130. Although it has some elements in common with Data 8, it was purposefully designed for students pursuing statistics programs of study, to give them a broad understanding of each of our five building blocks, and to whet their appetites for what is to come.

Some key features of STA130 include:

- It starts with a question. The first thing we do in the first class is to tell a story, which has an accompanying data set and compelling authentic questions.
- Computation is integrated throughout, in a just-in-time manner, motivated by the desire to answer those questions.
- Like Data 8, statistical thinking is introduced using computation and simulation. There are p -values and confidence intervals and linear models and classification trees, but no theoretical probability distributions.
- The course time is equally spent exploring statistical methods to answer the questions in large lecture sections, and talking or writing about explorations in small group tutorials.
- Multiple approaches are incorporated to develop problem-solving skills. For example, many methods are introduced, and students must focus on the key connecting ideas to make sense of it all. Most practice problems have a differentiating aspect from what has been previously discussed, so students must confront variations and the need to adjust their thinking. In tutorials, students are asked to articulate the process of their problem-solving. These practices are designed to support the development of adaptive expertise in our students (Gibbs & Damouras, 2019).
- All five building blocks are laid (professional practice primarily through experiential learning as discussed below) to set the foundation for the rest of the program.

In parallel with STA130, students acquire the traditional background in calculus, computer science, and linear algebra. But the students who finish STA130 are ready and raring to work on answering questions with data. Depth of understanding in the underlying theory and methods comes later. Like Tukey suggested was needed, we're being ambitious and picking up the pace. Surveys of our STA130 students have indicated a confidence

with using data to tackle questions, in contrast to our previous students, who continued to express reticence after their third year.

What about students outside of statistical and computer science programs? The Faculty of Arts and Science at the University of Toronto is currently in the process of designing three separate introductory data science courses for students in the humanities, physical and life sciences, and social sciences. These courses will be developed and co-taught by faculty in computer science, statistical sciences, and cognate disciplines such as English and geography. They will have similar building blocks to STA130, although the problems and associated methods and applications will be baked into the course via an interdisciplinary teaching model. This approach is similar in spirit to the approach of Data 8 plus a connector course and is “designed to give students a more immersive experience in the data science of a particular domain” (Adhikari et al., 2021).

Real Problems: The Need for Experiential Learning

An important thread throughout our curriculum, starting with STA130, is engagement with real problems. Similar to Kolaczyk et al. (2021), this experience is not delayed until a capstone. The efforts required for this are substantial, and well recognized. But like the changes exhorted by Cobb (2015), we strongly believe that “[o]ur job is to help [students] use data to answer a question that matters ... This may be the biggest exogenous challenge to our profession, the least explored in our undergraduate curriculum, and the most promising for rethinking what we teach.” If our students are to develop all the foundational supports we expect, particularly professional practice and problem-solving, we believe that they must engage with real problems. Authentic problems that look like real problems are not sufficient.

Experiential teaching and learning in statistics education involve using data to answer real, compelling, and critical questions. The engagement in these questions inspires students to make connections between the questions, data, analytical strategies, and communication of findings. This engagement also provides an environment for the development of problem-solving. As Tukey (1962) recognized, teaching data analysis involves “admit[ting] that it uses judgement ... we must teach an understanding of why certain sorts of techniques are indeed useful.” The instructor’s main role in experiential learning is creating a safe environment for students to make these connections and to develop this judgment.

Experiential education is not simply “learning by doing” and “simple participation in a prescribed set of learning experiences does not make something experiential” (Chapman, 1992). A student simply analyzing a set of ‘real’ data, without a meaningful question, even if it involves the entire ‘data lifecycle’ (Wing, 2019) including wrangling, modeling, and visualizing the data, and talking about the results, may not be experiential learning (Taback, 2018). Analysis of a large data set without a question is analogous to a person trying to become a physicist by simply going to the MIT Physics library. Just because it is big does not mean that all that is needed is in place for learning to happen. It is impossible to learn anything unless you have a question or framework to accompany your search, and a teacher to help you reflect on your answers.

We have incorporated experiential learning opportunities in a scaffolded way throughout our undergraduate statistics curriculum, starting with the first course. The final third of STA130 is a project: An industry or research partner introduces the project to the students and often invites some students to present their findings to colleagues in their organization. Our students have investigated whether new company practices lead to better outcomes, and if data from sensors in vehicles can capture whether the driver engages in hazardous driving practices and conditions. They are mentored on these projects in their small-group tutorials, and, to manage the scale and promote effective visual, written, and oral communication, they present their findings in a poster session.

Further experiential learning opportunities are available throughout our program, through summer opportunities, annual datathons such as [ASA DataFest](#), and courses offering micro-experiential opportunities and supported full-scale collaborations.

The Only Constant Is Change

“This river I step in is not the river I stand in.”

This phrase, attributed to the Greek philosopher Heraclitus, frames the passage of a bridge over the Don River on Queen Street East in Toronto, Canada, part of an [art installation by Eldon Garnet](#). One of us passes it regularly and often talks about it with our students. It is a reminder that change is constant, and the problems that frustrate our students are the problems that will prepare them to be able to adapt their expertise in the face of this change.

When we were considering what our statistics students needed to thrive in a data science ecosystem, it was clear to us that there is not a defined set of tools that we can provide them with that will serve their needs in every data science problem they encounter. But our building blocks are designed to give them the broad range of knowledge, skills, and experiences that they can adapt.

The data science ecosystem also includes many people who are not statisticians. And although what they need may vary in depth, rigor, and balance, we believe that our building blocks are foundational to data science education for everyone.

Disclosure Statement

Alison L. Gibbs and Nathan Taback have no financial or non-financial disclosures to share for this article.

References

Adhikari, A., DeNero, J., & Jordan, M. I. (2021). Interleaving computational and inferential thinking in an undergraduate data science curriculum. *Harvard Data Science Review*, 3(2).

<https://doi.org/10.1162/99608f92.cb0fa8d2>

Chapman, S. M. (1992). What is experiential education? *Journal of Experiential Education*, 15(2), 16–23.

<https://doi.org/10.1177/105382599201500203>

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, 69(4), 266–282. <https://doi.org/10.1080/00031305.2015.1093029>

Gibbs, A. L., & Damouras, S. (2019). Evolving statistics education for a data science world. In *Proceedings of the 62nd ISI World Statistics Congress 2019*, Kuala Lumpur, Special Topic Session (Vol. 3, pp. 37–46).

International Statistical Institute. <http://isi2019.org/proceeding/2.STS/STS%20VOL%203/index.html#p=48>

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>

Kolaczyk, E. D., Wright, H., & Yajima, M. (2021). Statistics Practicum: Placing “practice” at the center of data science education. *Harvard Data Science Review*, 3(2). <https://doi.org/10.1162/99608f92.2d65fc70>

Meng, X.-L. (2019). Data science: An artificial ecosystem. *Harvard Data Science Review*, 1(1).

<https://doi.org/10.1162/99608f92.ba20f892>

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>

Taback, N. (2018). Do you have experience? Incorporating experiential learning opportunities into statistics education is messy but important. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Proceedings of the 10th International Conference on Teaching Statistics*. International Statistical Institute. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_10A2.pdf

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1), 1–67. <https://doi.org/10.1214/aoms/1177704711>

Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review*, 1(1).

<https://doi.org/10.1162/99608f92.e26845b4>

©2021 Alison L. Gibbs and Nathan Taback. 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.