Statistics and Machine Learning at Princeton University

A statement by
President Christopher L. Eisgruber and Dean of the Faculty Deborah Prentice

August 22, 2016

Princeton University has a long history of excellence and innovation in the field of statistics. In the early part of the 20th century, that excellence was nurtured in the Department of Mathematics. In 1965, the statisticians split off from the math department and for two decades, Princeton had a Department of Statistics. When that department closed in 1985, the University made a conscious decision not to have a centralized statistics unit but instead to allow departments in the natural and social sciences and engineering to invest in statistics as befit their scholarly and curricular needs and opportunities. In the years since, many departments hired statisticians and, more recently, experts in machine learning onto the faculty and established statistics and machine learning courses at the undergraduate and graduate levels. The result is that Princeton continues to have excellent, innovative faculty working in this area, but has not fully leveraged their potential as a community of scholars and teachers with shared interests and objectives.

The impetus to bring this community together began in 2011, with a workshop that convened a dozen faculty members and another 60 students and staff to learn about each other's research in statistics and machine learning. Out of that workshop came a proposal in 2012 for the creation of the Center for Statistics and Machine Learning (CSML). Co-authored by Professors David Blei, Jianqing Fan, Robert Schapire, and John Storey, the proposal articulated the five-year plan for CSML as follows:

> *We will have a successful and popular undergraduate certificate program. We will have a top-ranked Ph.D. program in statistics and machine learning, which will be considered a path-breaking interdisciplinary program. We will have grown and broadened through the recruitment of several top-notch faculty members. We will have built a symbiotic relationship on campus with departments that rely heavily on the data sciences. We will have established a reputation in industry and academia for educating our students on the most cutting-edge, relevant aspects to the data sciences.*
>
> *We are entering the age of big data, which has become important in almost every discipline. With the center, Princeton will emerge as a leader in this vital new field whose impact could persist for decades.*
>
> *A Center for Statistics and Machine Learning will be a cornerstone on campus for the data sciences, and this will transform the university. It will be a place for graduates and undergraduates of all disciplines to study how to think about and analyze data. And it will spawn interdisciplinary faculty collaborations that launch scientific progress, harnessing the synergy between modern science and modern data analysis.*

The administration found this vision for CSML very compelling; it built on the strengths of Princeton's distributed model of statistics teaching and training while addressing some of the weaknesses of that model. It was broad and inclusive, and had widespread support from the data science community on campus. Although the subsequent departure of key faculty members (including Blei and Schapire) delayed adoption of the proposal's recommendations, the University established, in July 2014, the Center for Statistics and Machine Learning, naming Professor John Storey as the inaugural director.

At the same time, the president established the Task Force on Statistics and Machine Learning and charged it with developing a strategic plan for the new CSML. The report of the task force does just that, outlining an ambitious plan that draws heavily on the proposal that guided the creation of CSML but that also goes beyond it in significant ways. We are grateful to the task force for its broad-based examination of the state of statistics and machine learning at Princeton and other institutions and for its thoughtful recommendations on how to move forward. In this memo, we respond to these recommendations with an indication of those on which we will take immediate action, those that need additional development, and those that are of lower priority.

The task force report outlines four major recommendations: the appointment of existing faculty members and recruitment of new faculty to CSML, a new graduate program in statistics and machine learning, research infrastructure for CSML, and improvements to the University's undergraduate curriculum. The recommendation regarding faculty appointments was the top priority for the task force and the target of most of the feedback we received from the campus community. We accordingly begin our response with it.

Faculty

The faculty are the heart of any academic endeavor. Recognizing this, the task force focused its top-priority recommendation on building a faculty within CSML. The report recommends that the Center hold at least 10 FTEs, with the goal of making 14 joint appointments with other departments and 3 full-time appointments within CSML. It identifies just 6 current faculty members as candidates for joint appointments, with the remaining CSML faculty to be recruited from outside the University. Underlying this recommendation is an assumption about institutional architecture that was not present in the initial proposal for CSML. As the task force report states:

> *CSML will place a strong emphasis on interdisciplinary research, which implies that it must maintain close intellectual ties to other departments, including making joint appointments with other departments when appropriate. At the same time, statistics and machine learning are well defined disciplines with their own journals, professional societies, conferences, Ph.D. degree programs, and departments established throughout the world. None of the top 15 ranked Departments of Statistics in the United States is jointly administered with another discipline such as applied mathematics, operations research, or computer science.* ***Therefore, it is crucial that CSML exists with independence from other departments at Princeton*** [emphasis added].

We received a great deal of feedback about how the CSML faculty should be constituted. Some of the feedback supported the task force's recommendation, but most of it did not. We heard from individual faculty members who wanted to participate in CSML's mission but either did not want joint appointments or were not among the six that would be offered appointments. We heard concern from faculty members and departments that had made significant investments in statistics and wondered how those investments would relate to CSML and its activities. We heard from a few colleagues who were delighted that Princeton was finally moving toward establishing a statistics department and from many more who were alarmed at this development. It is important to note that everyone who weighed in was supportive of the creation of CSML and eager to see growth in the statistics and machine learning faculty on campus. Most were agnostic about how faculty should be affiliated with the Center. They simply wanted stronger connections and more flexible boundaries between CSML and other academic units than the task force report envisions.

In light of this feedback, we conclude that it is premature to settle on a faculty appointment structure for CSML. The University has many successful centers and other interdisciplinary units that operate on different models. Some offer undergraduate and graduate programs but make no faculty appointments (e.g., the Bendheim Center for Finance, the Office of Population Research); some make a small number of faculty appointments but engage a much broader set of faculty in their leadership, governance, and programs (e.g., the University Center for Human Values, the Princeton Environmental Institute, the Princeton Institute for the Science and Technology of Materials); and some appoint most or all of their core faculty (e.g., the Lewis Sigler Institute for Integrative Genomics, the Princeton Neuroscience Institute). Right now, there is no consensus on which of these models will best serve the growth and development of CSML; indeed, within the group of a dozen or so faculty who specialize in statistics and machine learning, there are advocates of each of the three models. We view this lack of consensus as unsurprising (and possibly even healthy) at this stage of CSML's development, and for that reason will not authorize CSML to make faculty appointments at this time. The dean of the faculty will work with CSML to insure that they have the faculty participation they need to support and grow their research and teaching programs.

We are also eager to support CSML in its goal of building a world-class faculty in statistics and machine learning. To that end, the Academic Planning Group (APG) allocated two faculty FTEs to CSML last year for use as incentives to departments to hire faculty and authorized CSML to initiate faculty searches. These resources have enabled CSML to identify top faculty members who can be hired into existing departments and to contribute half of the FTE support to make those hires possible. There are currently at least two such candidates working their ways through the appointment process. In addition, a number of departments have made statistics and machine learning priority areas in their own hiring. The result is a growing community of statistics and machine learning faculty on campus, which bodes well for the future of CSML. The APG stands ready to allocate additional FTEs to CSML as needed to support faculty hiring.

Graduate program in statistics and machine learning

We now turn to the recommendation of a new graduate program in statistics and machine learning. In the words of the task force report:

> *A critical part of CSML will be a top-flight graduate program, which will train the next generation of scholars in statistics and machine learning. Our aim is to give our Ph.D. students a broad interdisciplinary perspective on both fields. Our graduates will have an understanding of the theoretical foundations and their applications in practical domains and disciplinary subjects. They will become leaders in academia, industry, and government.*

The report goes on to detail the requirements of the Ph.D. program and its parameters. It also describes plans for the creation of a graduate certificate in data science.

This is a very exciting initiative, one that draws on Princeton's distinctive model of graduate education as well as its strengths in statistics and machine learning. We are eager to see it move ahead. As the University builds the critical mass of faculty needed to sustain a Ph.D. program, we would encourage the CSML leadership to begin conversations with the dean of the Graduate School about the program parameters, funding sources, and degree requirements. Indeed, it might be possible to establish the graduate certificate program in the coming year, with the Ph.D. program to follow.

Research infrastructure for CSML

The third recommendation of the task force is a significant investment in research infrastructure. This recommendation includes three separate proposals for dedicated space, a Data Science Core, and hardware beyond what the Princeton Institute for Computational Science and Engineering (PICSciE) and the High-Performance Computing Research Center (HPCRC) currently provide. We consider each of them in turn.

Space. The task force argued persuasively that CSML could only fulfill its goals of fostering collaborative research and providing innovative educational programs if it had space at a central campus location, proximate to related departments. We fully support that recommendation. Space for the Center has been identified in the old Dial Lodge, which will be vacated by the Bendheim Center for Finance in early 2017.

Data Science Core. The Data Science Core is an especially innovative aspect of the task force's recommendations: the Core includes scientific programmers, data scientists, data administrators, and software developers whose role is to support the computational work of CSML faulty and students. We understand the value of the Data Science Core for supporting faculty and student research and for enabling innovative research collaborations. We appreciate the role it has to play in faculty recruitment. The provost's office and the dean for research are exploring how best to meet the research infrastructure needs of both CSML and the broader data science community.

Hardware. The task force report describes the computational resources needed to support the research of faculty working in statistics and machine learning. It also describes the dependence of this work on data acquisition and network infrastructure. We are well aware of these needs, which we have supported in the past through start-up packages to individual faculty members. The task force proposed a strategy to meet these needs more efficiently for the community as a whole and thereby build an infrastructure that will make Princeton a highly attractive place to do work in this field. We recognize the significant benefits of such a strategy, and the Offices of the Provost, the Dean for Research, and Information Technology are working with relevant academic units to find effective and coordinated ways to serve these needs.

Undergraduate curriculum

Finally, we turn to the recommendation of changes to the undergraduate curriculum in statistics and machine learning. CSML has already made a major contribution to undergraduate education at Princeton with the establishment of a popular undergraduate certificate in statistics and machine learning in 2013. To build on that success, the task force proposed that CSML work with departments to improve existing course offerings and possibly establish new SML courses to fill curricular gaps.

The first of these initiatives is welcome and, indeed, long overdue. The lack of coordination in the statistics course offerings across the curriculum has been a source of confusion and frustration for generations of Princeton students. In response, the University established the Committee on Statistical Studies and charged it with coordinating statistics-related course offerings across departments. Though the Committee on Statistical Studies exists to this day, it has never been able to fulfill its mandate. We are fully supportive of the proposal that CSML take up this charge and further suggest that the Committee on Statistical Studies be brought into CSML or formally disbanded.

The proposal for new SML courses to fill gaps in the undergraduate curriculum was not universally supported by the task force; it will be a topic of continued discussion by faculty in CSML. The need for new SML courses may become clearer once the work of cataloging existing courses is done. In any event, the APG allocated two lecturer FTEs to CSML last year to support new introductory-level undergraduate courses in statistics and machine learning. If CSML decides not to offer such courses through the SML designation, it can contribute these lecturer FTEs to support departmental course offerings.

Conclusion

We thank the members of the Task Force on Statistics and Machine Learning for a forward-looking and thoughtful report. We will return to the report often as we support the Center for Statistics and Machine Learning through the initial years of its development.