# The Future of Statistics and Machine Learning at Princeton University

Prepared by the Task Force on
Statistics and Machine Learning
2014-2015

PRINCETON
UNIVERSITY

## TASK FORCE MEMBERS

**Chair**
John D. Storey, *William R. Harman '63 and Mary-Love Harman Professor in Genomics; Professor in the Lewis-Sigler Institute for Integrative Genomics; Director, Center for Statistics and Machine Learning*

**Faculty Members**
Jonathan Cohen, *Robert Bendheim and Lynn Bendheim Thoman Professor in Neuroscience; Professor of Psychology and the Princeton Neuroscience Institute; Co-Director, Princeton Neuroscience Institute*

Jay Dominick, *Vice President for Information Technology and Chief Information Officer*

Jianqing Fan, *Frederick L. Moore, Class of 1918, Professor in Finance; Professor of Operations Research and Financial Engineering; Chair, Committee for Statistical Studies*

Kosuke Imai, *Professor of Politics; Director, Undergraduate Certificate Program in Statistics and Machine Learning*

Matthew Salganik, *Professor of Sociology*

Christopher Sims, *John J. F. Sherrerd '52 University Professor of Economics*

James Stone, *Professor of Astrophysical Sciences and Applied and Computational Mathematics; Director, Princeton Institute for Computational Science and Engineering; Director, Fund for Canadian Studies*

Olga Troyanskaya, *Professor of Computer Science and the Lewis-Sigler Institute for Integrative Genomics*

**Staff Members**
Kara Dolinski, *Assistant Director, Lewis-Sigler Institute for Integrative Genomics*

# TABLE OF CONTENTS

# EXECUTIVE SUMMARY

Data have become the driving force behind modern research in nearly every field. From English scholars working in the digital humanities to neuroscientists tracking the circuitry of the brain and sociologists scraping billions of data points from social networks, scholars must develop and use sophisticated tools to make meaningful, data-driven discoveries. Underlying these research areas is the central question: *How do we learn from data?*

In response to the impacts of significant technological progress and the "big data" movement, scientific fields and their approaches to data analysis have evolved, converging at the interface of statistics and machine learning. There, researchers are leveraging the data analysis tools of statistics and the computational innovations from machine learning, to enable the analysis of massive data sets in ways that have never before been possible. The new field of "data science", which includes at its core statistics and machine learning, has the potential for transformative impact in a wide range of human endeavors, from advances in precision medicine that would personalize medical treatments of individual patients to developments in deep learning that could help solve real-world problems ranging from automatic image recognition to drug discovery.

In July 2014, Princeton University took a critical step forward with the establishment of the Center for Statistics and Machine Learning (CSML) as the intellectual hub for education and research activities in these areas. Recognizing the University's unique opportunity to assume a leadership position in the field, a task force was established to develop a strategic plan for the Center that would help to create a vibrant and cohesive community in statistics and machine learning (SML) while also nurturing the field's connections to research and teaching throughout the University.

This report presents a vision and strategic plan for Princeton's Center for Statistics and Machine Learning and puts forth a series of recommendations intended to support, enable, and enhance the University's excellence in the evolving fields of statistics and machine learning. These recommendations in order of priority are:

1. Appoint Existing Faculty Members and Recruit New Faculty Members to CSML
2. Create a New Graduate Program in SML
3. Provide Research Infrastructure for CSML
4. Improve the University's Undergraduate Curriculum

The most important first step is to create faculty positions for CSML. We have the opportunity to build a world-class faculty in SML given our interdisciplinary, mathematical, and computational strengths at Princeton University. The ability to make 100% effort appointments in CSML is essential for recruiting the full range of faculty members that we require. In order to retain and recruit the best faculty members in SML, we will need to introduce a first rate graduate program and provide a modern research infrastructure for these individuals. Finally, while CSML already has a thriving undergraduate certificate program, there is room for improvement in the design and breadth of our undergraduate course offerings.

# VISION

### Why Statistics and Machine Learning Together in One Center?

Statistics and machine learning ask the same fundamental question: *How do we learn from data?* Statistics has existed for a much longer period of time than machine learning, growing out of both science and engineering environments. In particular, statistics plays a central role in almost all empirical sciences. Machine learning involves topics from the intersection of statistics and computer science. It is specifically concerned with computational statistics and automated methods for extracting information from data. Machine learning primarily grew out of an engineering environment. However, over time the theoretical and applied problems central to these two fields have evolved so that their commonalities far outweigh their differences.

There is a well-known quotation in statistics (based on another well-known quotation): "Those who ignore statistics are condemned to reinvent it." This applies to machine learning as well, and the machine learning community has embraced the idea that a well-trained machine learning researcher must also be well-trained in statistics. On the other hand, statisticians have also recognized that much of what has traditionally consumed its field was introduced before modern computing and massive data sets became so common in science and engineering. Therefore, statistics has shifted away from developing inference methods for small data set problems and has become increasingly focused on massive data set problems. Statisticians recognize that machine learning provides fundamental computational algorithms for the analysis of large-scale data. Therefore, statisticians increasingly recognize that the modern statistician must also be well-trained in machine learning.

Machine learning faculty members have been traditionally appointed in Departments of Computer Science, and statistics faculty have traditionally been appointed in Departments of Statistics or Biostatistics. If one were given the opportunity to make those appointments today without these traditions in place, then the machine learning and statistics faculty members may very well be placed in the same unit. Princeton University is in the unique position that we can do exactly this since there is no

Department of Statistics and there is not yet a critical mass of machine learning faculty in the Department of Computer Science.

### Defining the Core of CSML

CSML's critical mass will be built around faculty members who work at the interface of statistics and machine learning. These will be individuals focused on learning from large, complex data sets where both statistical model fitting, computational efficiency, and automated discovery are key challenges. Underlying this critical mass will be a focus on interdisciplinary research activities to maintain close ties to the problem areas that drive the most relevant research problems in statistics and machine learning. This environment will be appealing to students, who will be receiving the most relevant training needed today. It will also be appealing to faculty members working on other important areas in statistics and machine learning, as the theory and methods provided by the intersection of statistics and machine learning provide a fertile intellectual basis for working on broader research topics.

### Impact

The number of faculty currently at Princeton who primarily work in statistics and machine learning (SML) is small; however, the research breadth, depth, and impact of these faculty members are exceptional. These faculty members have close relationships with many units on campus. This interdisciplinary framework is a defining feature of the Center, which distinguishes it from traditional statistics and machine learning programs. With this intellectual leadership in place, Princeton has the foundation needed to grow the Center into one of the most influential research institutions in this field, but only if we act now. We will be able to attract the best students to our programs and the best scholars to our faculty. The merging of statistics and machine learning with an emphasis on real-world, interdisciplinary problems is novel as of today. CSML is positioned to be path breaking in this regard if we move forward quickly with this strategic plan.

CSML will impact both research and education at the University, and bring outside recognition as a top ranked academic program. The Center will provide a coherent curriculum in the data sciences for students, and a core of faculty who will be essential collaborators to other researchers on campus. In addition to the traditional disciplines now consumed by data (from physics to sociology), the University has recently made major investments in new scientific directions, such as genomics, neuroscience, and energy development in which collecting and understanding data are central. Success in data-driven research will be a significant area of focus for Princeton University, and its centrality to our teaching and research will only continue to grow.

The Center will quickly grow to encompass a top PhD program, a cohesive undergraduate curriculum, a data science computing infrastructure, and a thriving intellectual community of world-class students and faculty. There is widespread support on campus for the Center's mission as well as strong potential for substantial funding from outside sources. The Center will become a pillar of the data sciences both on campus and abroad. The contributions from its graduates and faculty will have a major impact on science, industry, and society.

## CURRENT STATE OF SML AT PRINCETON

### History

Princeton University has a rich and influential history in statistics and machine learning. Samuel Wilks (1930s-50s) and John Tukey (1950s-70s) both played leading roles in developing statistical theory and methods that have had profound impacts on science and technology. William Feller (1950s-60s) established probability theory as a branch of mathematics in the United States. Gilbert Hunt (1950s) made fundamental contributions to the theory of stochastic processes. These Princeton faculty members are among the greatest statisticians and probability theorists of the last century. At the same time, Princeton also played a major role in developing modern computational science. Alonzo Church and Princeton graduate alumnus Alan Turing independently conceptualized computing machines (1930s), while John von Neumann (1930s-50s) later envisioned how modern computers could solve complex problems.

A Department of Statistics existed at Princeton University during 1965-1985, with John Tukey being its inaugural chair. Princeton educated many highly influential leaders in statistics. Examples include David Brillinger (Berkeley), Arthur Dempster (Harvard), David Donoho (Stanford), Bill Eddy (CMU), Don Frazer (Toronto), David Freedman (Berkeley), John Hartingan (Yale), and Don Rubin (Harvard).

Since the Department of Statistics closed, statistics faculty at Princeton have been appointed in a wide range of departments, resulting in an interdisciplinary group of faculty. Very few faculty members were hired in statistics and machine learning in the 1990's and early 2000's. More recently, with the increasing role that data-driven discovery plays in many fields, more statistics and machine learning faculty have been hired in several departments and institutes at Princeton. Today's state of faculty in SML is described in detail below.

The current Princeton faculty members working in statistics and machine learning have not yet reached a critical mass, but will be able to do so once CSML begins hiring outside faculty with appointments in CSML.

**Strengths**

*High-Impact Faculty.* Princeton currently has about six faculty members who work primarily in statistics and machine learning. These SML faculty members have excellent research and teaching records, carrying out well-funded and highly cited research. Princeton has a much larger number of faculty members who primarily work in other fields, but whose work nontrivially involves SML to varying degrees.

Two faculty members at Princeton are COPSS Presidents' Award winners, which is arguably the most prestigious award given in statistics. Princeton University was recently recognized as having the third highest citation impact in probability and statistics journals (http://archive.sciencewatch.com/dr/sci/11/jan2-11_2/), behind only Stanford University and Johns Hopkins University that have both invested substantial resources in their programs. A recent study also concluded that two of the top 10 most cited mathematicians in the world are SML faculty at Princeton (http://www.in-cites.com/top/2007/fourth07-math.html). The SML faculty are highly active in the fields of statistics and machine learning, from editing top journals such as the *Annals of Statistics* and the *Journal of Econometrics* to presiding over academic societies such as the Institute of Mathematical Statistics and sitting on the editorial boards of the *Journal of*

*American Statistical Association*, the *Journal of Machine Learning Research*, and organizing committees of the top conferences in machine learning.

*Interdisciplinary Research.* Princeton has particularly strong interdisciplinary research in statistics and machine learning, in fields such as economics, finance, genomics, neuroscience, political science, and sociology. This diversity and depth of interdisciplinary research is noteworthy compared to other universities, and it is a strength on which we should build.

*Interest from Students.* The level of interest in statistics and machine learning from the undergraduate and graduate students is likely at an all-time high. The enrollment numbers in our undergraduate fundamentals courses in statistics and machine learning have exploded over the last seven years.

We also just recently started an undergraduate certificate in SML in the 2013-2014 academic year (see section below, **UNDERGRADUATE CURRICULUM**). It is already one of the most popular certificate programs in the University with 68 students enrolled this year. This compares in size with the neuroscience certificate program, which just recently became an undergraduate concentration.

### Weaknesses

*Faculty Appointments, Size, and Retention.* Faculty members who primarily work in SML are appointed in a wide range of units on campus. These faculty members are physically and administratively separated from one another, which significantly limits their interactions and ability to carry out the SML mission. Their teaching activities are determined by requirements from their home departments, so they are usually unable to teach core SML courses. This has lead to a lack of coherent course offerings in SML and a lack of a much-needed critical mass of SML faculty. The number of core SML faculty is currently small, limited to about six as of today.

Faculty retention has been a significant issue, as other universities have been growing their programs in statistics, machine learning, and data science at a rapid rate. In order to reverse this trend, CSML should begin to hire core faculty members and lecturers as soon as possible.

*Course and Degree Offerings.* There are no graduate or undergraduate concentration degree programs in SML at Princeton University. As mentioned above, there is a recently established undergraduate certificate program in SML. A major hurdle in establishing a PhD program and strengthening the undergraduate curriculum is the lack of course offerings in SML. The courses that are available are spread out over many different departments and often overlap significantly with each other, which have lead to an inefficient use of teaching resources and a lack of other key SML courses. As an example, there are currently seven introductory statistics courses at Princeton, which are taught by seven different departments. (The seven courses are ECO202, EEB/MOL355, POL345, WWS200, ORF245, PSY201, and SOC301.) The names of these courses are varied and often confusing to undergraduates, from "Quantitative Analysis and Politics" to "Fundamentals of Engineering Statistics". However, they are all introductory statistics courses that cover essentially the same topics. There is a lack of undergraduate courses that cover material in the core areas of SML that follow these introductory courses. There are very few offerings at the graduate level in the core areas of SML. As stated above, these issues are due to the fact that SML faculty are appointed in other departments, and their teaching activities are determined by those appointments. This makes it infeasible both to coordinate courses and to insure that they are regularly offered.

## SML at Comparable Universities

We plan to build CSML to be path breaking and unique compared to existing programs; this is possible as long as we move quickly to do so. Departments of Statistics and Departments of Biostatistics are very common at research universities in the United States. Departments of Machine Learning are rare, with just several in the United States in existence, including one at Carnegie Mellon University. Machine Learning faculty members are typically appointed in Departments of Computer Science or Departments of Statistics.

To characterize how SML is organized outside of Princeton, we considered Departments of Statistics, Biostatistics, and Machine Learning at the following high-impact research universities: Carnegie Mellon University, Harvard University, Johns Hopkins University, Stanford University, University of California at Berkeley, University of Chicago, University of Pennsylvania, and Yale University. We found that the median number of faculty members in these departments is 23; the smallest number of faculty members in any department is 14 (which is considered very small for a Department of Statistics). The median number of PhD students is 46 in the graduate programs. As described in more detail in the **FACULTY APPOINTMENTS & RECRUITING** section below, we aim to have ~20 active faculty members (mostly joint appointments, which will make us unique compared to existing programs) and a PhD program of 40-50 students to provide adequate intellectual depth and breath, as well as teaching coverage, which is in line with the median size of these other departments at our peer institutions. We also noted that among the top departments in statistics or biostatistics, none of these is administered jointly with related fields such as mathematics, operations research, or computer science. This supports the decision to establish CSML as a separate entity from these departments; although, as outlined below, there should be close partnerships with these and other departments.

There are very few joint programs in statistics and machine learning. Over time the problems of interest, the intellectual content, and the techniques in statistics and machine learning have become very similar to the point where it can be difficult to distinguish the two fields. Many highly successful researchers are active in both fields. However, universities have been slow to merge these two fields into a single organization. There is no obvious advantage for this separation, so we conclude it is likely due to historical and administrative barriers.

The separation of statistics and machine learning is arguably artificial and unnecessary today, and it may be detrimental to both fields. The lack of such a separation at Princeton will put us at an advantage over other universities, but only if we act quickly to properly develop CSML. There are other universities, such as MIT, that have recently announced similar initiatives. Princeton University has embraced its unique opportunity to combine statistics and machine learning into one program. This will be a highly attractive feature to faculty members and students from outside of the university whom we would like to recruit. This will also be valuable for obtaining research and training support from funding agencies.

## RESEARCH PRIORITIES

The Center for Statistics and Machine Learning is an inherently interdisciplinary entity whose research will be focused around methodological challenges of data sciences at the intersection of statistics and machine learning. These challenges and the methods that address them are pervasive and permeate methodological disciplines such as statistics and computer science as well as application areas from the natural and social sciences. The key challenge in defining the Center's research mission is balancing the core methodological research in statistics and machine learning with the applications of these methods to diverse data in the natural and social sciences.

It is clear that the Center must possess a clear intellectual and scholarly core focused on statistical and machine learning methodological developments, including statistics and machine learning theory, probabilistic modeling, and applied statistics and machine learning. In fact, several core members of the Center are already research leaders in this area, and both the CSML graduate and undergraduate programs will provide core courses covering these subjects. This methodological development in statistics and machine learning should remain the key defining aspect of the Center's research and its participants. It will clearly define the intellectual and scholarly contributions of CSML both within the University and to the outside world.

However, it is also critical that CSML sustains deep connections with real-world application areas, providing focus and application prominence for the SML research. These more applied directions are fundamental to the mission and success of the Center for several reasons: (1) applications of statistics and machine learning methods to real-world data are critical for wide impact and identification of key next challenges for methods development; (2) top-rate applied statistics and machine learning contributions are a strength of Princeton, both in terms of existing faculty and in taking advantage of the size and collaborative atmosphere of the University; (3) for the Center to thrive, both within the University and in achieving world-class eminence, the Center must engage and leverage University departments.

Thus, CSML will include core methodological development faculty, as well as core members who are more applied and may be jointly appointed with any of the participating natural and social sciences department. The key characteristic of all these members will be intellectual and scholarly contributions to development and application of statistical and machine learning methods.

## RECOMMENDATION: UNDERGRADUATE CURRICULUM

The Center will work with departments across campus to improve our undergraduate course offerings in statistics and machine learning. Given the growing importance and prominence of statistics and machine learning in both industry and academia across a wide spectrum of fields, it is crucial for Princeton undergraduates to have access to a first rate education in these areas.

In addition to the certificate program that we already established, our plan for undergraduate education includes a more comprehensive and well-organized catalog of SML courses that fills in some existing gaps. There are currently several courses available on campus in statistics and machine learning. However, as discussed above in regards to the seven introductory statistics courses, the courses as a whole can be difficult to identify as SML-related based on their titles, they have overlapping material, and there are key areas in SML where no course is offered. To streamline and expand our offerings, we will work to consolidate several courses, offer a course accessible to all undergraduates, and introduce new courses.

The task force discussed the following proposal to offer new courses, which was not universally supported among all task force members. Therefore, the following proposal comes with the disclaimer that we believe that further deliberation is necessary before any permanent changes are made. Specifically, we propose to form an organizing committee subsequent to this task force that will further study and work with existing departments to determine the best course of action for the future in regards to undergraduate course offerings. The following proposal may be a useful starting point.

**New Courses**

We discussed a proposal to establish several new courses in the first few years of CSML. In particular, this plan introduces a course that is open to any undergraduate at Princeton, regardless of her or his background. This fills an extremely important hole in undergraduate education at Princeton: a course that provides basic literacy in statistics, machine learning, and data science. The proposed title of this course is "Reasoning with Data." The course will cover the challenging concepts and strategies that one must understand in order to make sound conclusions from data. The course will not focus on technical matters and mathematics, but it will involve substantial data analysis.

As a second, more technical new course, this plan includes attempting a new approach to getting students interested in SML by developing a course called "Introduction to Data Science." The traditional introductory statistics course has remained unchanged for decades. This style of the traditional introductory course is often considered to be dry and not engaging (irrespective of the university). "Introduction to Data Science" will give students (with some necessary mathematical and computing background) the opportunity to collect big data sets, clean them, and analyze them in their first SML course at Princeton. The course will provide a set of skills that will be useful to the student regardless of their ultimate career path. This course can be organized in such a way that it complements existing introductory courses, or so that it provides students with an alternative means to initiate their education in SML.

The proposal also organizes follow-up courses that dig deeper into statistics and machine learning, and that move beyond introductory level courses. An example set of courses is the following:

SML 101: Reasoning with Data
SML 201: Introduction to Data Science
SML 301: Fundamentals of Statistical Inference
SML 302: Fundamentals of Machine Learning

SML 401: Models and Methods for Data Analysis

We note that Computer Science currently teaches an undergraduate course titled "Artificial Intelligence" and Operations Research and Financial Engineering teaches a course titled "Analysis of Big Data." Both of these complement the above core courses and will likely be cross-listed with SML. COS 424, Fundamentals of Machine Learning, already exists and is currently cross-listed as SML 302.

After organizing a core set of courses in collaboration with other departments, we propose to introduce courses in other key areas of SML. Topics include nonparametric statistics, applied machine learning, and statistical computing. There is no immediate need for CSML to develop courses focused on the application of statistics and machine learning to a specific application area (e.g., economics, neuroscience, or political science). These exist in abundance due to the interdisciplinary manner in which SML has existed at Princeton so far.

### Undergraduate Certificate Program

CSML introduced an undergraduate certificate program in SML last academic year. As mentioned above, this program is already one of the most popular certificate programs at Princeton. The exact details of the program are given in **APPENDIX: EXISTING UNDERGRADUATE CERTIFICATE PROGRAM**; briefly, the certificate requires the students to take five courses, one of which must be a "Fundamental of Statistics" course and the other a "Fundamentals of Machine Learning" course. A quick overview of the courses available to certificate students demonstrates the point made above about the significant overlap among existing courses and the lack of other core material in SML.

### Coordination with Other Departments

One important mandate of the undergraduate SML certificate program is the possible consolidation of introductory statistics courses offered in many departments. There are currently at least seven entry level statistics courses (ECO202, EEB355, POL345, WWS200, ORF245, PSY201, and SOC301) that have significant overlap in their content. Each year, hundreds of students enroll in these courses. Some departments have difficulty staffing these courses every year, so there is an important need for coordination. CSML should work closely with the relevant departments to develop the above courses so that they complement and meet the needs of departments across campus. Existing introductory statistics courses often provide students with the data analysis skills necessary for their junior papers and senior theses. Close coordination between CSML and these departments will be critical. The proposed SML courses from above are a good starting place to begin to coordinate with departments across campus. There are at least three potential paths that this coordination could take. One path is that the introductory courses are consolidated into a fewer number of introductory courses that are coordinated between CSML and the departments. A second path is that departments may wish to allow their students to take CSML courses in place of their own courses. A third path is that these introductory course remain as is, but they are cross-listed according to coordinated SML course numbers to provide guidance to students, and to make the introductory courses more interchangeable. These paths are neither mutually exclusive nor exhaustive.

### Undergraduate Concentration in SML

There is likely enough interest among undergraduates to justify implementing a concentration in SML. However, given that we are essentially starting from the beginning in terms of courses, we believe it makes more sense to gradually introduce a comprehensive set of courses over the next five years, and then revisit whether an undergraduate concentration is warranted based on the enrollment levels and success of these courses, the evolution of the field, and the emerging interests of the student population.

## RECOMMENDATION: GRADUATE CURRICULUM

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 PhD 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 establishment of a PhD program is important to both create a fertile research environment in CSML and to also properly train students in this field. Although there are currently PhD students in other Princeton PhD programs whose thesis topics are primarily in the SML fields, achieving the proper breadth and depth of training is extremely difficult for those students. It also puts the students at a disadvantage when seeking positions after training at Princeton because they are competing with others who were trained in proper statistics or machine learning programs.

**Core Courses**

We propose to create a program to provide our graduate students with a unique multi-disciplinary education that will train them in both statistics and machine learning. We intend to do this with a balanced curriculum of courses in both areas, and courses that are both theoretical and applied. A regular seminar series will also be a central component of the program.

We will require all of our graduate students to take at least six courses, chosen to reflect the interdisciplinary breadth of the program. This program of courses must satisfy the following requirements:

- Two courses in machine learning;
- Two courses in statistics;
- One course in probability (e.g., ORF 526 Probability Theory).

Additionally, students will be required to complete a responsible conduct in research course that CSML will organize. We will establish the following new graduate courses to serve as the core courses in the SML PhD program:

SML 501: Applied Statistics
SML 502: Applied Machine Learning
SML 511: Theoretical Statistics
SML 512: Theoretical Machine Learning
SML 520: Modern Computing for Data Science

Each of these courses will be offered once a year. The specifics of how the PhD program will work are given in **APPENDIX: GRADUATE PROGRAM LOGISTICS**.

### Elective Courses

We will also offer a rich variety of specialized courses, focusing on the state of the art in modern statistics and machine learning. These will be an essential component for fostering PhD level research. Some example topics are the following.

*Machine Learning*

- Boosting: Foundations and Algorithms
- Online Learning
- Reinforcement Learning

*Modern Applied Statistics*

- Nonparametric Statistics (Bayesian and Frequentist)
- Bootstrap and Other Resampling Methods
- Large Scale Inference
- Hierarchical Bayesian Modeling
- Numerically Intensive Methods for Bayesian Inference
- Truth in Data: Inference from Observational Data
- Causal Inference

*Theoretical Statistics*

- Asymptotic Theory
- Bayesian Nonparametrics
- Causal Inference
- Statistical Learning Theory
- Functional Data analysis

*Interdisciplinary Applications*

- Prediction in Financial Data
- Detecting Systemic Risk in Financial Markets
- Statistics and Machine Learning in Genomics

- Quantitative Methods in Neuroscience
- Large Scale Survey Analysis for Social Science
- Methods in Natural Language Processing
- Learning and Robotics
- Social Networks

### Size of Graduate Program

The normal length to get a Ph.D. degree is five years. In its steady state, we aim to have about three PhD students per CSML faculty member. This would involve successfully recruiting a number of new PhD students per year slightly smaller than the number of CSML faculty members.  We will work with the Dean of the Graduate School to identify the source of these required graduate student slots.

### Interactions with Other PhD Programs at Princeton University

Beyond the Ph.D. program, we aim to serve students across almost all disciplines including the natural and social sciences.  We expect that many SML courses will be regularly taken by graduate students from other programs.

We plan to create a **Graduate Certificate in Data Science**.  This program will require four courses as follows:

- Two courses chosen from among SML 501, 502, 511, 512
- SML 520
- APC 524: Software Engineering for Scientific Computing

We expect that this graduate certificate program will be both very helpful and popular among graduate students from other PhD programs.

## RECOMMENDATION: FACULTY APPOINTMENTS & RECRUITING

### Configuration of Appointments

As we have emphasized in this report, Princeton is unique in that SML faculty and research here are strongly interdisciplinary.  Therefore, CSML will make it a priority to hire faculty members that are jointly appointed with an appropriate partnering department, institute, center, or program. These appointments would typically be 50% effort in CSML, which would imply that the jointly appointed faculty member would teach about one course per year for CSML. Like other institutes on campus that are interdisciplinary (e.g., Lewis-Sigler Institute for Integrative Genomics and Princeton Neuroscience Institute), we will hire jointly appointed CSML faculty with the intention that the faculty member and her or his research group will reside in CSML. We also strongly suggest that 50% effort appointments made jointly with other departments be under the control of CSML so that faculty searches are not limited to a single partnering department and so that the 50% effort remains tied to a joint CSML position should a jointly appointed faculty member leave the University.

CSML should not be limited to joint appointments, however. Departments of Statistics exist at most research universities. Faculty members in these departments may very well not have just one area in which they consider applications. Some of the best SML faculty members in the world do interdisciplinary research in several areas, and some of the best in the world do not do any interdisciplinary research. We believe it would be a critical mistake for CSML to be limited to joint appointments, so we strongly suggest that CSML be permitted to make 100% effort appointments when necessary. The list of world-renowned SML faculty at other universities who would be in this scenario is substantial. We also argue that joint appointments with CSML should not be limited to traditional departments. This would prevent us from making joint appointments with some of the most data-intensive institutes on campus, such as Lewis-Sigler Institute for Integrative Genomics, Princeton Neuroscience Institute, or the Andlinger Center for Energy and the Environment.

**Number of Faculty**

A key question is how many faculty members should ultimately be appointed to CSML. We consider this question in two ways, resulting in similar answers. First, we must consider the size of SML departments or centers at comparable universities. As described in **SML AT COMPARABLE UNIVERSITIES**, we found that the median number of full time faculty in Departments of Statistics, Machine Learning, or Biostatistics is 23. The smallest department (Harvard) has 14 full time faculty members, and the next smallest department has 20 faculty members.

Second, we should consider the teaching mission of CSML. The following is an example proposed course load for CSML when it is at full capacity:

| Course | Sections Taught Per Year |
|---|---|
| SML 101 | 6 |
| SML 201 | 6 |
| SML 301 | 2 |
| SML 302 | 2 |
| SML 401 | 1 |
| SML 501 | 1 |
| SML 502 | 1 |
| SML 511 | 1 |
| SML 512 | 1 |
| SML 520 | 1 |
| SML UG Electives | 4 |
| SML GR Electives | 6 |
| Total | 32 |

We also propose that CSML explores hiring two fulltime, stable, and highly experienced lecturers to mainly teach SML 101 and 201, and to cover faculty sabbaticals and other forms of faculty leave. The task force recognizes that important introductory courses should not be outsourced to temporary lecturers; instead these courses should be taught by experienced, long-term senior lecturers or by faculty members. Full time lecturers teach 6 courses per year. This leaves ~20 courses for the CSML faculty to teach.

Taking these two analyses together, **we propose that CSML grows to include at least 10 whole (full time effort) faculty positions**. Given our goal to primarily make joint appointments, an example configuration is 14 faculty who are appointed at 50% effort in CSML and three faculty who are appointed at 100% effort, for a total of 17 faculty members. Even at this level, we would still have four fewer full time positions than the smallest department we identified.

### Current Faculty

There are currently about six faculty members who should be given the opportunity to move one-half of their faculty positions (50% effort) to CSML.

**Recruiting New Faculty**

Given the small number of core SML faculty at Princeton, there are a number of crucial research areas that are not represented. Important areas that should guide how we recruit faculty include the following:

- Bayesian statistics
- Computing on massive data sets
- Data visualization
- Network modeling
- Reinforcement learning
- Spatial and temporal modeling
- Statistical theory

**Relationship with Other Departments**

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, PhD 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.

There are many departments, centers, institutes, or programs on campus where current faculty members have SML as primary or secondary interests. These departments are strong possibilities for joint faculty appointments with CSML in the future. Examples include Astrophysical Sciences, Computer Science, Ecology and Evolutionary Biology, Economics, Electrical Engineering, the Lewis-Sigler Institute for Integrative Genomics, Mathematics, the Office of Population Research, Psychology, Operations Research and

Financial Engineering, Politics, Princeton Neuroscience Institute, Program in Applied and Computational Mathematics, Sociology, and the Woodrow Wilson School of Public and International Affairs. This list is not exhaustive, but rather gives a sense of the breadth of possible joint faculty appointments.

CSML will work to engage in “big data” applications with partnering departments. This could be implemented through joint graduate students working on projects directly with application scientists and core scientists at CSML, joint faculty appointments with application departments, and course work directed towards students in application areas.

## RECOMMENDATION: RESEARCH INFRASTRUCTURE

CSML will need significant research infrastructure to be successful, in particular, space, research personnel, and computing hardware.

### Space

Because CSML is inherently multidisciplinary, having space designated for our diverse group to work is essential to its success. Space on campus will foster collaborative research and the exchange of ideas, and it will provide a focal point for the newly formed educational program. We envision a place for administrative offices, seminar and meeting rooms, graduate student and postdoctoral fellow offices, faculty offices, and a shared workspace for visitors and undergraduates. Ideally, the location will be central on campus, making it close to the home departments of its related disciplines.

### Data Science Core

We propose to create a *Data Science Core*, which will be composed of highly skilled staff to propel the University’s data intensive research and to support the CSML faculty’s computational work. This research core will be an extremely valuable asset to the University and an attractive feature to faculty that we recruit.

The Data Science Core will be composed of about eight highly skilled staff members who can assist the researchers in exploiting novel methods on special purpose computing systems. Faculty, postdoctoral fellows, and graduate students must be able to rely on technical staff who are able to assist code development, profile and optimize existing codes, create data architectures that support the research, and provide the basic technical support. We have identified several kinds of experts that will need to support CSML, but the ultimate qualifications of the staff will be determined by the particular research programs of the Center.

*Scientific Programmers*. These are experts in modern software development practices, especially High Performance Computing. Typically holding PhD degrees in subjects aligned with the research, these people will help write the software that supports the research activities of the faculty. The best Scientific Programmers are difficult to successfully recruit and retain; they are often on their way to another position or the continuation of their academic trajectory, or they will have higher paying positions elsewhere. Relative to each faculty member supporting their own Scientific Programmers, supporting a group of Scientific Programmers within CSML increases the likelihood of recruiting and retaining the most capable individuals.

*Data Scientists*. These are experts in collecting, managing and manipulating large, complex data sets that support the research activities of the faculty. For “big data” science, collecting the data is a non-trivial task that often includes a deep understanding of the scientific domain and of technical methods for data manipulation, storage and archival. They will work closely with the Scientific Programmers and researchers in understanding the data, and they will typically have PhD degrees from data science areas.

*Data Administrators*. Data Administrators will be responsible for supporting the long term archival, management, and access to research data and research product. They will support open access initiatives, manage data sharing and partnership agreements, and provide the expertise to manage the data and published research product in a manner consistent with funder obligations, such as those required by the National Institutes of Health. These people would potentially be a highly sought-after resource on campus and would, in conjunction with the library, provide leadership in the Open Access space.

*Software Developers*. Skilled in writing code to support data collection, manipulation and visualization, these staff members work under the direction of the Data Scientists and the Scientific Programmers to accomplish the research programs of the faculty.

The Center will depend on resources in OIT's Research Computing group for hardware and systems administration support, in the same manner as does PICSciE. The preferred funding model for the Data Science Core will be for the majority of funds to come from a stable source, such as the University or an endowment. A model involving external research funding contributing to this group is possible, but measures would have to be taken to make it robust to the unpredictable nature of external funding.

**Hardware**

The Center is likely to need computational resources in addition to the existing resources available through PICSciE and that currently exist at the HPCRC. SML problems are more likely to need computers with very large amounts of memory and specially configured, high-capacity storage. The University has room for expansion in HPCRC where existing compute resources are located.

The expected hardware requirements will be a natural complement to the resources already provided through PICSciE. The latter has focused on high compute-performance machines (e.g., for simulations), whereas CSML is likely to need high memory-performance machines for data managements and analyses. CSML will require machines with large shared memory space and potentially exabyte data repositories connected with fast, low-latency networking. A balanced mixture of both architectures (both big-computing and big-data machines) is crucial to support the overall research computing environment at Princeton, and will give the University a competitive edge in recruiting and retaining faculty.

Any new machines acquired to support research in the CSML should be physically located at the HPCRC, and supported by the same groups in OIT that support our existing infrastructure. There should only be one unified research computing infrastructure at Princeton available to all, rather than "PICSciE machines" and "CSML machines."

It is essential that the Center as a whole, as well as individual faculty within the Center, are allowed to develop a computing environment that meets their specific needs. There is no interactive computing environment provided through PICSciE, which is an essential environment for timely data science research. Second, some funding agencies -- the National Institutes of Health (NIH) in particular -- prohibit funds from being spent on equipment that is not for the dedicated use of the project. This policy is non-negotiable and strictly enforced by the NIH. The current PICSciE policy is that funding for new nodes goes to the general pool of cluster nodes.

In order to make the PICSciE model feasible for CSML and to ensure it remains the *de facto* big data cluster on campus, we propose the following policy modifications. First, we propose that a parallel cluster be developed for CSML with the storage and compute requirements listed above. This cluster should have a limited number of interactive nodes to allow iterative real-time application and testing of software for data analyses. Second, priority buy-in for individual faculty members should be allowed to enable the spending of research funds to support the cluster in accordance with the research funder's policies. Finally, the charge applied to faculty for compute and storage should not be on a yearly basis because of how funding cycles work.

If these systems are in place, the compute requirement will be approximate to that which is already provided to PICSciE.  Along with PICSciE, the success of the CSML will depend on having the funding available to ensure that hardware resources are replaced/upgraded on a reasonable cycle.  This funding needs to be sustained and predictable (e.g., via the University or an endowment), but provided in a manner that still encourages faculty to seek hardware-oriented grants and gifts.

It is likely that the storage system needed to support the Center will exceed the current capacity or technology of the existing PICSciE infrastructure by a wide margin. Based on current pricing, a storage infrastructure for SML would cost about $1M for 10 petabytes of data. The design of that system will depend on the nature of the research being done at CSML and can be phased in as CSML grows.

### Data Acquisition

Data is obviously an essential resource for CSML. The Center may require funds for acquiring non-public data sets. While data will be sourced through publicly available resources whenever possible, market forces in the big data arena are driving towards proprietary collections in addition to and, in some instances, superseding publicly available collections.

### Networking

The Data Cluster and new storage infrastructure will drive use of the campus network in ways that it has not been driven in the past. This will likely require a higher level of investment by the University in connectivity on campus, out to the HPCRC, and to the national and global infrastructure. It is assumed that the University will have both on-campus and global bandwidth available to support the data movement required by the research.

## Staff

The staffing needs will involve administrative staff and research staff. The Data Science Core research staff members were discussed above in **Research Infrastructure**.

The administrative staff will be in line with the usual needs for a Princeton organization that involves faculty members, postdoctoral fellows, PhD students, undergraduate researchers, and undergraduate certificate students. Specifically, CSML will require the following administrative staff members:

- Assistant Director (administrative)
- Grants / Finance Manager
- Assistant Grants Manager
- Administrative Assistant to the Director
- Administrative Assistants (number commensurate with CSML size)

- Student Coordinator
- IT Manager
- IT Support Staff (number commensurate with CSML size)

## MEASURES OF SUCCESS

We are entering the age of big data, which has become important in almost every discipline and society at large. With CSML, Princeton will emerge as a leader in this vital new field whose impact could persist for decades. CSML 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.

In the first five years of the Center's existence, we aim to accomplish a number of goals summarized in **APPENDIX: PROPOSED TIMELINE AND GOALS**. We will have a successful and popular undergraduate curriculum, including the existing certificate program. We will have a top-ranked PhD program in statistics and machine learning, which will be considered a path-breaking interdisciplinary program. We will have established a world-class data management and analysis infrastructure that supports the research and educational activities of CSML. 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.

To evaluate the success of CSML efforts in education, we should focus on three keys areas: the undergraduate certificate program in SML, the Ph.D. program in SML, and the graduate certificate in Data Science, which will be open to Ph.D. students across the University.

For the undergraduate program, the two most direct signs of a healthy program will be course enrollments and certificate completion. Because we believe that the training offered by the Center is critical for success in modern times, we will strive to ensure that the students we serve – in our courses and our certificate program – are representative of the entire student body. In particular, we will strive for a balanced representation of women, historically under-represented minority groups, and first-generation college students. Further, the Center's robust and reliable course offerings will improve the undergraduate training in departments across campus by reducing redundancy and freeing faculty to offer complementary courses.

For the Ph.D. program, the most direct signs of a healthy program are number of applications, yield on admitted graduate students, and placement record. Further, we believe that the most critical measure of the success of the program is the impact of the research produced by our graduate students. As with the undergraduate program, we will strive increase the diversity of the SML community by attracting and training students from under-represented groups in coordination with the university-wide effort.

The final component of our educational offerings is the graduate certificate in Data Science, which would be open to Ph.D. students across the University. This certificate program will provide a valuable service to graduate students by offering cutting-edge training to students in other departments. Further, the certificate program will promote creative, interdisciplinary research by bringing together students who might not ordinarily interact, for example, a biology student researching protein interaction networks and sociology student researching social networks. The most direct signs of a successful graduate certificate program are course enrollments, certificate completion rate, and the impact the program has on participants' research and career path.

In addition to its educational mission, the CSML faculty will also play an integral role in carrying out research on some of the most important problems we face today. The most direct measures of research impact are papers published, citation counts, research funding by government and industry, awards won by the faculty, and so on. But, in addition to those measures that can be entered into a spreadsheet, we believe that the ultimate measure of success for CSML will be in its potentially profound impact on how people understand and utilize the enormous amounts of data that we have today, and the new scientific insights and technologies to which these can give rise.

## APPENDIX: PROPOSED TIMELINE AND GOALS

The following is a summary of the timeline and goals that we propose in this strategic plan.

| Academic Year | Goals |
|---|---|
| 2014 – 2015 | • Establish the Center for Statistics and Machine Learning ✓<br>• Appoint first director ✓<br>• Establish undergraduate certificate program ✓<br>• Implement inaugural seminar series consisting of world leaders in SML ✓<br>• Write strategic plan ✓<br>• Initial development of undergraduate and graduate curricula ✓<br>• Move into temporary space ✓<br>• Hire support staff ✓ |
| 2015 – 2016 | • Appoint ~6 existing faculty members to CSML by moving 0.5 full time effort positions into CSML for each individual<br>• Introduce new undergraduate courses SML 101 and 201<br>• Establish PhD program and carry out first year admissions<br>• Hire 2 full time lecturers<br>• Hire 2 new faculty members<br>• Establish regular seminar series<br>• Hire director of Data Sciences Core |
| 2016 – 2017 | • Move into permanent space during Summer 2016<br>• Enroll first PhD program cohort<br>• Hire 2 new faculty members<br>• Introduce new undergraduate courses SML 301, 302, and 401<br>• Introduce new graduate courses SML 501, 502, 511, and 512<br>• Formally establish and hire staff for Data Sciences Core |
| 2017 – 2018 | • Hire 2 new faculty members<br>• Introduce workshops on big data analysis for researchers on campus lead by Data Sciences Core<br>• Introduce new graduate course SML 520<br>• Introduce several new undergraduate and graduate SML elective courses |
| 2018 – 2019 | • Hire 2 new faculty members<br>• Carry out self-study to assess previous 5 years<br>• Investigate whether an undergraduate concentration in SML is appropriate<br>• Introduce several new undergraduate and graduate SML elective courses |

✓ = completed

## APPENDIX: EXISTING UNDERGRADUATE CERTIFICATE PROGRAM

### Overview

The Program in Statistics and Machine Learning is offered by the Center for Statistics and Machine Learning. The program is designed for students, concentrating in any department, who have a strong interest in data analysis and its application across disciplines. Statistics and machine learning, the academic disciplines centered around developing and understanding data analysis tools, play an essential role in various scientific fields including biology, engineering, and the social sciences. This new field of “data science” is interdisciplinary, merging contributions from computer science and statistics, and addressing numerous applied problems. Examples of data analysis problems include analyzing massive quantities of text and images, modeling cell-biological processes, pricing financial assets, evaluating the efficacy of public policy programs, and forecasting election outcomes. In addition to its importance in scientific research and policy making, the study of data analysis comes with its own theoretical challenges, such as the development of methods and algorithms for making reliable inferences from high-dimensional and heterogeneous data. This program provides students with a set of tools required for addressing these emerging challenges. Through the program, students will learn basic theoretical frameworks and apply statistics and machine learning methods to many problems of interest.

### Enrollment to the Program

Students are admitted to the program after they have chosen a concentration, generally by the beginning of their junior year. At that time, students must have prepared a tentative plan and timeline for completing all of the requirements of the program, including required courses and independent work (as outlined below), as well as any prerequisites for the selected courses.

### Program of Study

Students are required to take a total of five courses and earn at least B- for each course: one of the “Foundations of Statistics” courses, one of the “Foundations of Machine Learning” courses, and three elective courses. With all necessary permissions, advanced students may also take approved graduate-level courses. Students may count at most two courses from another degree program (departmental concentration or another certificate program) towards this certificate program.

Students are also required to complete a thesis or at least one semester of independent work in their junior or senior year on a topic that makes substantial application or study of machine learning or statistics. This work may be used to satisfy the requirements of both the program and the student's department of concentration. Submission is due on the same date as your department deadline for thesis or junior independent work. All work will be reviewed by the Statistics and Machine Learning Certificate committee. At the end of each year, there will be a public poster session at which students are required to present their work to each other, to other students, and to the faculty.

Finally, students are encouraged to attend one of the Statistics and Machine Learning colloquia on campus. These include the Wilks Statistics Seminar, the Machine Learning Seminar, the Political Methodology Seminar, or the Quantitative and Computational Biology Seminar.

**Certificate of Proficiency**

Students who fulfill the program requirements receive a certificate upon graduation.

Courses

One of the following courses (“Fundamentals of Statistics”):

| | |
|---|---|
| ECO 202 | Statistics & Data Analysis for Economics |
| EEE 355 | Introduction to Statistics for Biology (also MOL 355) |
| ORF 245 | Fundamentals of Engineering Statistics |
| POL 345 | Quantitative Analysis and Politics |
| PSY 251 | Quantitative Methods |
| WWS 200 | Statistics for Social Science |

One of the following courses (“Fundamentals of Machine Learning”):

| | |
|---|---|
| COS 424 | Fundamentals of Machine Learning |
| ORF 350 | Analysis of Big Data |

Three of the following courses (including those above, with permission):

| | |
|---|---|
| COS 402 | Artificial Intelligence |
| ECO 302 | Econometrics |
| ECO 312 | Econometrics: A Mathematical Approach |
| ECO 313 | Econometric Applications |
| ELE 486 | Compression and Transmission of Information |
| GEO 422 | Data, Models, and Uncertainty in the Natural Sciences |
| MAT 385 | Probability Theory |
| MOL 436 | Statistical Methods for Genomic Data |
| ORF 309 | Probability and Stochastic Systems |
| ORF 405 | Regression and Time Series |
| ORF 418 | Optimal Learning |
| POL 346 | Applied Quantitative Analysis |

**Example Paths for SML Certificate**

Computer Science, Mathematics, or Engineering Student

- ORF 245 Fundamentals of Engineering Statistics
- COS 424 Fundamentals of Machine Learning
- ORF 309 Probability and Stochastic Systems
- ORF 350 Analysis of Big Data
- COS 402 Artificial Intelligence

Economics or Finance Student

- ORF 245 Fundamentals of Engineering Statistics
- COS 424 Fundamentals of Machine Learning
- ECO 312 Econometrics: A Mathematical Approach
- ORF 350 Analysis of Big Data
- ELE 486 Compression and Transmission of Information

Life Sciences Student

- MOL 355 Introduction to Statistics for Biology
- COS 424 Fundamentals of Machine Learning
- ORF 309 Probability and Stochastic Systems

- GEO 422 Data, Models, and Uncertainty in the Natural Sciences
- MOL 436 Statistical Methods for Genomic Data

Social Sciences Student

- POL 345 Quantitative Analysis and Politics
- COS 424 Fundamentals of Machine Learning
- ECO 312 Econometrics: A Mathematical Approach
- ECO 313 Econometric Applications
- POL 346 Applied Quantitative Analysis

## APPENDIX: GRADUATE PROGRAM LOGISTICS

### Coursework

The CSML faculty will maintain a list of courses approved for the purposes of fulfilling the requirements. Each of these courses will also be marked as to the requirements that the course fulfills, with the same course often fulfilling more than one requirement.

Students must get a grade of at least A- in four of the six courses, and a grade of at least B+ in all six of them. Three of the six courses must be completed by the end of the first year, five by the end of the second year, and all six by the end of the third year.

At least initially, we do not anticipate a facility for granting credit to students who have taken similar courses elsewhere. Rather, such students will be encouraged to take other courses, or courses at a more advanced level.

### Financial Support

As is standard for other PhD programs, we request that the University provide a one-year fellowship to all graduate students. We expect that this will typically be taken in the first year of graduate studies, although in certain circumstances, some flexibility in the timing of when the fellowship is taken might benefit all involved (for instance, if the advisor has external funding that would otherwise expire if not used in the student's first year).

In their second year, we expect that students will receive teaching assistantships. In the years that follow, students will be supported with research assistantships, fellowships, or teaching assistantships, to be arranged by the student's advisor. As is the case in the departments involved, we expect to easily secure support for our students.

### Exams

We will implement the core exam to identify issues with students early-on in their career. If a student gets A- or better in at least three core courses at the end of the second semester then the student passes the core exam. Otherwise, core exams for the student will be arranged at the beginning of the third semester to be conducted by a committee appointed by the Director of Graduate Studies.

The general exam for this program, typically taken at the end of the second year, will be modeled on the process used in Computer Science. The exam committee will consist of the advisor and two other faculty, at least one of whom is a faculty member is CSML. The composition of the committee, which must be approved by the Director of Graduate Studies, will reflect the breadth of disciplines that are central to the program.

Several months before the exam, students will agree with their committee on a reading list of about a dozen papers and books. At the exam, the student will give a public research presentation lasting about an hour, and will answer questions from the committee and other audience members. In the second part of the exam, the student will be questioned privately by the exam committee on material contained in the pre-selected reading list. Students may also be questioned on material in any core course in which they received a grade below A-.

For both the core exam and general exam, the CSML faculty will decide whether a student should continue her/his Ph.D based on the recommendation of the oral examination committee.