TO: SEAS Deans and Faculty & Statistics Faculty  
From: David Parkes, Area Dean, Computer Science, and Neil Shephard, chair, Statistics Department  
RE: New Terminal On-Campus Data Science Master’s Degree Program (SM) Proposal  
This draft: 28th March, 2016

1. Rationale and Vision

A. Overview

This proposal sets forth a plan for a Data Science master’s degree (SM) to be established by Harvard’s FAS, under the joint academic leadership of the Computer Science and Statistics faculties, and administered through the Institute for Applied Computational Science (IACS) at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS). The authors propose a three-semester degree program and the restructuring of existing courses to provide support for the many disciplines in FAS and across the entire Harvard campus that use data science. The chief objectives addressed by this proposal are:

1. Harvard can expand its impact on society by leading in training talented students in this fast changing, fundamental field, aiding their careers in research, business, public policy, and other important endeavors.
2. Establishing a coherent core of courses for this Data Science master’s will serve two additional purposes:
   - Students in FAS departments and professional schools that lack a critical mass in data science faculty need access to high-quality foundation stones and targeted courses to effectively create their own data science tracks within existing degree programs.
   - Ph.D. students in varied fields can further their own research by taking some of the data science courses, either by completing a Secondary Field in Data Science or by enrolling on a course-by-course basis.
3. Providing an academic point of focus for data science education at Harvard will encourage the methodologists in many different departments to work together more effectively, reducing confusing overlaps in course content and fostering the existing broad ecosystem of data scientists across Harvard’s disciplines. In particular, the proposed curriculum and degree program will provide one of a number of institutional linkages between Computer Science (CS) and Statistics as both faculties advance knowledge in this fast-moving interdisciplinary field.
4. By appropriately organizing the teaching of data science, this structure can respond to public demand to seed coherent and influential courses for Harvard’s Extension and Summer Schools as well as HarvardX.

In addition, we propose the establishment of a “Data Science Lab” to coordinate and support the work of postdoctoral researchers, who will assist faculty in developing and teaching the new courses, sponsor research seminars and lunchtime workshops, and build a critical mass of researchers working on the science of data. This would be joint between Statistics and Computer Science (housed within those groups), but would be administered by IACS to minimize marginal administrative costs. We would expect the Lab would be folded into the Data Science Institute if it is eventually founded.
B. What is Data Science?

Data science (DS) lies at the intersection of statistical methodology, computational sciences, and an array of areas of application. By placing the generation and analysis of data at the center of modern analytics, data science is substantially impacting almost every empirical scholarly discipline and is having a profound impact on business, government and politics, pure and applied science and engineering, medicine and public health, journalism, sports, law, and education.

Given below are a few comments from various public sector leaders on the importance of this subject:

Driven by the declining cost of data collection, storage, and processing; fueled by new online and real-world sources of data, including sensors, cameras, and geospatial technologies; and analyzed using a suite of creative and powerful new methods, big data is fundamentally reshaping how Americans and people around the world live, work, and communicate. It is enabling important discoveries and innovations in public safety, health care, medicine, education, energy use, agriculture, and a host of other areas.


Harnessing big data has the potential for big impact in all areas of science and engineering, driving new innovations, and addressing some of today's most pressing challenges—enhancing our quality of life.


The era of “Big Data” has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information.


To respond to the pressing demand for education in this important area and fully prepare students for roles in the vanguard of the data science revolution, Harvard needs a degree program that provides both technical depth and a sophisticated understanding of the issues that this revolution poses to society. The data science program will roundly educate the next generation of leaders to be knowledgeable not just in the basics of data handling and calculation. Students will learn the principles of experimental design, visualization, massive and distributed computing, and software engineering. They will fully comprehend the implications of dependencies within databases, appreciating the difference between cause and correlation while gaining an understanding of the legal, ethical and computational issues associated with data privacy and security.

C. Achieving our goals

Leaders in the field of data science come from many intellectual backgrounds. Given its widespread application across fields, it is no surprise that data science attracts a highly diverse population of students. Harvard’s program must therefore be well rounded and technically robust at its core while also being
malleable enough to give students diverse routes to meet their educational needs, providing them the tools to adequately fill gaps in their previous learning. Keeping these needs in mind, we have taken strides to work collaboratively across the University over the last year to establish a set of core data science courses. This degree will use core courses as its foundation stones. Other FAS Departments and professional schools may choose to adopt certain core courses to create discipline-specific data science tracks within their disciplines.

Existing AM in Statistics and SM in Computational Science and Engineering (CSE) programs require eight courses and span two semesters. After actively debating whether the new degree program should be two or three semesters long, we have settled on three semesters—a requirement of 12 courses—for the following reasons:

1. Admission to a two-semester program would be possible only for students meeting very high prerequisites in both CS and Statistics. These requirements would exclude some students who are highly gifted but have less technical preparation, making it vastly harder to offer core DS courses across schools and provide Ph.D. students in diverse disciplines access to modern DS.
2. A three-semester program allows students to have a summer of work or research experience to leverage as they move on to the job market or further education.
3. Extending the program reduces the pressure on master’s students to launch their job search immediately on arriving at Harvard—pressure that may cause some to miss opportunities for deeply committing to their educational experience.
4. DS cuts across multiple domains. It is beneficial for students to have the time to be contemplative and explore potential points of focus and application in their third semester.
5. The third semester will give students the opportunity to write a substantial research paper or complete a capstone project in some area of DS or application.

Students taking this degree may wish to stretch their 12-course load over four semesters (e.g., to carry out a more extensive research project or serve as a teaching fellow). We have in mind that this option will be available with the support of the appropriate Director of Graduate Studies.

2. Benchmarking with other universities

This is given in the Appendix.

3. Degree requirements

The design of the proposed DS degree program revolves around achieving a set of learning outcomes for each student, regardless of their specific starting point. To meet this goal, prerequisites, core course requirements, project work, and electives must dovetail with existing courses and with degree programs having a data science component across Harvard. While acknowledging that this will require continuing faculty conversation, we present this proposal as a starting point.

We designed the course requirements to be flexible enough to address the diversity of their intellectual backgrounds. Ph.D. students across campus will be able to complete a Secondary degree in Data Science during their time at Harvard, details of this are noted after the curriculum outline.
A. Learning outcomes

A graduate of the Harvard SM in Data Science program should be able to:

1. Build statistical models and understand their power and limitations
2. Design an experiment
3. Use machine learning and optimization to make decisions
4. Acquire, clean and manage data
5. Visualize data for exploration, analysis, and communication
6. Collaborate within teams
7. Deliver reproducible data analysis
8. Manage and analyze massive data sets
9. Assemble computational pipelines to support data science from widely available tools
10. Conduct data science activities aware of and according to policy, privacy, security and ethical considerations
11. Apply problem solving strategies to open ended questions

B. Prerequisites

Prerequisites for admission to the SM in Data Science have collectively been agreed upon as:

- Computer programming: CS50 or equivalent. Our courses will center on the use of R and Python.
- Probability and statistical inference: Stat110 and Stat111 or equivalent.
- Mathematics: Linear algebra and multivariate calculus. Math 19a and 19b or equivalent.

C. Required courses

Students would typically take four courses a semester for three semesters. Of these, six would be required courses: three in the first semester, two in the second and one in the third.

Core courses: Minimum 6, Maximum 6

The core breaks down into a four course technical core, which we have labeled M1, M2, M4, and M5, and two other courses, M3 and M6.

The core has been designed carefully following numerous discussions with colleagues across schools. It has elements which can be used by other schools as a backbone for a Data Science offering in their own educational programs.

The following table spells out the requirements for the degree. Italics denote courses that are electives. Roman font denotes the six core courses.

<table>
<thead>
<tr>
<th>SM requirements</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Technical core courses (M1, M2, M4, and M5)</td>
<td>4*</td>
<td>4</td>
</tr>
<tr>
<td>2. Critical thinking and Data Science (M3)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3. Research experience (M6)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SM requirements</td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>------------------------------------------------------------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>4. Statistical electives</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5. Computer Science electives</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6. Other data science electives (FAS, HSPH, HKS, HBS)</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>7. Seminar courses</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8. Data Science independent study research course</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>12</strong></td>
</tr>
</tbody>
</table>

* Students could request to place out of a technical core course on the grounds they have already achieved the relevant learning outcomes. The Director of Graduate Studies would need to give permission for this.

Below we list the six core courses and which semesters they fall into. Four are technical courses, one is critical thinking, and the sixth is a research or Capstone project. Following this we will discuss potential electives.

**Semester 1’s core courses**

**M1: Data Science 1: An introduction to Data Science** - This course is designed for a wide audience, and could be used by all of Harvard’s professional schools as an introduction to data science when their students are not highly prepared. (Python centric. All concepts taught through data analysis examples. Computing required: understanding of basic concepts such as variables, conditionals, loops, and functions. HarvardX can provide background before course if missing.)

1. Data import
   a. Basic data wrangling
   b. Preparing data for analysis
2. Exploratory Data Analysis (EDA)
   a. Histograms
   b. qq-plots
   c. Boxplots
   d. Scatterplots
3. Basic Inference:
   a. Distributions
   b. Random variables
   c. p-values
   d. Confidence intervals
4. Monte Carlo Simulation
5. Introduction to Linear Algebra
   a. Matrix Algebra
   b. Transformation
   c. Projections
6. Regression and ANOVA
7. Dimension reduction
   a. Distance
   b. Principle component analysis (PCA)
   c. Singular value decomposition (SVD)
   d. EDA with PCA
8. Statistical models
   a. Basic parametric models
   b. Maximum likelihood estimation
   c. Bayes theorem
   d. Multilevel models
   e. Nonparametrics via bootstrap
9. Introduction to Machine Learning
   a. Test and training sets
   b. Confusion matrix
   c. kNN
   d. Cross validation
10. Multiple comparison problem
11. Data products
   a. Based on a data analysis project
   b. Communicating results

12. Reproducible research
   a. Organizing data
   b. ipython notebooks

M2: Computing for Data Science 1 - This course is designed for DS specialists. It will be highly attractive for Ph.D. students at Harvard who need to operationalize DS in their research. (Python centric. Computing required: CS50.) M2 has some overlap with CS 207, Systems Development for Computational Science, but is more focused on software. We predict it would be a very popular and intellectually important course at Harvard.

1. Agile development
   a. SCRUM
   b. git / github
   c. Doctests and debugging

2. Basic programming skills
   a. Unix scripting
   b. Regular expressions
   c. Library development (CLIs and setup.py)

3. Testing
   a. Unit testing
   b. Integration testing

4. Modern language features
   a. Modularity / OO
   b. Templates / Generics
   c. Multimethods

5. Performance and re-factoring
   a. Profiling
   b. Scaling of algorithms

6. Concurrency
   a. Multithreading
   b. Asynchronous execution

7. Language
   a. Sequences and memory protocols
   b. Streams, iterators, and generators
   c. Dictionaries, sets, and objects

8. Data structures
   a. Lists, linked lists, collections
   b. Hashing
   c. Trees

9. Semester-long data analysis project on large data set

M3: Critical thinking and Data Science - This course is designed for a wide audience to harness Harvard’s exceptional knowledge of the non-technical aspects of the subject and how it impacts the world. This unique course would give students the opportunity to think about DS from a wider viewpoint, to aid their development as future leaders in the field. (No coding or high school algebra. The course would use case studies and be made up of a series of short modules. Would involve outside speakers, e.g. General Counsel to a security agency, a data science entrepreneur.)

1. Legal & ethical aspects of privacy
2. Data security
3. Visualization and communication
4. Cause/correlation/experiments
5. DS impacting the world:
   a. DS and science
   b. DS and medicine
   c. DS and public policy
   d. DS and business

Semester 2’s core courses

M4: Data Science 2: The second step in Data Science – This will be a deepening of material in M1. (R+Python centric, Computing required: Assumes M1+M2.)
1. Statistical models and methods
   a. maximum likelihood, Bayes
   b. bias, bias-variance tradeoff
   c. Error rate controlling procedures and false discover rates
   d. Logistic regression
   e. Bootstrap
   f. Cross-validation
   g. Causal inference
   h. Missing data
2. Visualization and communication
3. Data Acquisition
   a. Web scraping
   b. Data cleanup
4. Data Management
5. Running experiments
6. Machine learning
   a. Classification, SVM, decision trees, random forests
   b. Clustering
   c. Dimension reduction
   d. L1 type regularization (e.g., Lasso)
   e. Gaussian processes
   f. Nonparametric Bayes
7. Reproducible research
8. Semester-long data analysis project on large dataset

M5: Computing for Data Science 2
1. Basic Monte Carlo methods
   a. Rejection and importance sampling,
   b. Variance reduction methods;
2. Bayes formalism and sampling
   a. Bayesian modeling, Markov chains,
   b. Metropolis-Hastings
   c. MCMC convergence analysis
   d. Hierarchical Bayes
   e. Gibbs sampling
3. Optimization
   a. Simulated annealing
4. Dynamic systems
   a. Time series analysis,
   b. Hidden Markov models
   c. Sequential Monte Carlo
   d. Gaussian processes
5. Advanced sampling methods
   a. Slice sampling,
   b. Hamiltonian MC
   c. Parallel tempering, Emcee

Currently CS109/Stat121, which is called Data Science, is the closest we have to M1 and M4. As this course has a large student enrollment, the relevant teaching faculty wishes to split it into two, unique courses: Data Science 1 and Data Science 2. It is believed this might be achieved for 2017-18. The structure would be similar to the above M1 and M4. M5 has the same structure as the current AM207.

Semester 3’s core courses

M6: Data Science Research Experience

We are modeling our Capstone Project Course on AC297r, as detailed below. We leave open the option to merge the Computational Science and Engineering (CSE) capstone with this course or to run it as a separate course.
AC297r, Capstone Project Course

The Capstone Project course integrates and applies the skills and ideas in computational and data science that students acquire in other courses, such as data management, machine learning, statistics, and visualization.

Requiring students to complete a substantial and challenging collaborative project, the Capstone course ensures that students are trained to conduct research and are prepared for the professional world. Projects are selected to combine the statistical, computational, and engineering challenges and social issues involved in solving complex real-world problems.

Students are placed in groups of three to four, and each group works with their instructor, mentors, and partners to identify a complex, open-ended real-world problem. Partners are from academia, government, and the e-commerce, medical, and financial industries. Student groups understand and define the overall problem and propose a solution. Solutions are in the form of a software package with documentation, a set of recommendations in a report, or a research paper.

Some students may team up with a Harvard faculty member during their time here, and we wish to leave open the option of replacing the Capstone Project by working on a research project instead. A project report or research paper would be assessed from this research project. Note: students can go deeper in terms of research by also taking a Data Science independent study research course.

Core courses overview

Courses M1 and M3 are not highly technical, so they would be appropriate for DS specialists in the MBA program or potentially in HKS or GSE. M2 and M4 require more technical skills but could be useful for specialists in HCSPH. M2 and M5 may be very helpful for many Ph.D. students.

M3 would reflect Harvard’s campuswide strengths in law, public policy, business, and ethics. We have begun talks of collaborative efforts with our peers in Law, Government and Business so that we may develop M3.

Expected schedule of courses

<table>
<thead>
<tr>
<th>Fall 1</th>
<th>Spring 1</th>
<th>Fall 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1. Data Science 1</td>
<td>M4. Data Science 2</td>
<td>M6. Capstone course</td>
</tr>
<tr>
<td>M2. Computing for DS 1</td>
<td>M5. Computing for DS 2</td>
<td>elective</td>
</tr>
<tr>
<td>M3. Critical Thinking and DS elective</td>
<td>elective</td>
<td>elective</td>
</tr>
<tr>
<td>elective</td>
<td>elective</td>
<td>elective</td>
</tr>
</tbody>
</table>

D. Preliminary List of Electives

Data science is primarily the intellectual child of CS and Statistics. Most of Harvard’s CS and Statistics courses are immediately relevant to the DS master degree.

Statistical electives: min 1, max 5
Stat131, Time series and prediction
Stat139, Statistical sleuthing through linear models
Stat140, Design of experiments
Stat149, Generalized linear models

Stat160, Design and analysis of sample surveys
Stat171, Introduction to stochastic processes
Stat186, Causal inference
Stat210, Probability I
Stat212, Probability II
Stat211, Statistical inference I
Stat213, Statistical inference II
Stat220, Bayesian data analysis
Stat221, Statistical computing and learning
Stat225, Spatial statistics
Stat230, Multivariate statistical analysis
Stat232r, Topics in missing data
Stat240, Matched sampling and study design
Stat244, Linear and generalized linear models

Computer science electives: min 1, max 5
CS51: Introduction to computer science II
CS105, Privacy and technology
CS124: Data structures and algorithms
CS125, Algorithms and complexity
CS134, Networks
CS165, Data systems
CS171, Visualization
CS181, Machine learning
CS187, Computational linguistics
CS205, Computational foundations of computational science
CS207, Systems development for computational science
CS281, Computability and complexity
CS222, Algorithms at the ends of the wire
CS224, Advanced algorithms
CS265, Big data systems
CS281 Advanced machine learning

If resources permit, we would like to build a new elective, a Data Science Clinic. This would be a data science consultancy for the Harvard community where people bring data science problems for advice. Statistics Department Ph.D. students also request setting up a service like this, and the two could be folded together. Potentially it would be jointly listed between CS and Statistics.

Other Data Science Electives Across Campus
One of Harvard’s key advantages is that the courses given by methodologists across our campus would be available to those enrolled. We have organized a sampling of courses below using Schools and Departments as headers.

FAS
Mathematics and Applied Mathematics
Math116, Convexity and optimization with applications
AM120, Applicable linear algebra
AM121, Introduction to optimization: models and methods
AM107, Graph theory and combinatorics
AM221, Advanced optimization
Math152, Discrete mathematics

Economics
Econ1126, Quantitative methods in economics
Econ2150, Big data
Econ2142, Time series analysis

Government
Gov2001, Advanced quantitative research methodology

HBS
1955, Big data and marketing
2107, Commercializing science: technology strategy and business models for science-based enterprises

HCSPH
Biostatistics
BIO XXX, Health data science practice
BIO 210, Analysis of rates and proportions
BIO 212, Survey research methods in community health
BIO 223, Applied survival analysis
BIO 226, Applied longitudinal analysis
BIO 257, Advanced statistical genetics

BIO 283, spatial statistics for health research and social inquiry
BIO 287, Public health surveillance
BIO 504, Introduction to geographical information systems using ArcGIS
BIO 512, Introduction to computational biology and bioinformatics
BIO 513, Advanced computational biology and bioinformatics
BIO 521, Introduction to social and biological networks

Epidemiology
EPI 201, Introduction to epidemiology methods
EPI 202, Elements of epidemiologic research: Methods 2
EPI 203, Study design in epidemiologic research
EPI 204, Analysis of case-control and cohort studies

HMS

Biomedical Informatics
BMI XXX, Precision medicine I: integrating clinical and genomics data
BMI XXX, Precision medicine II: genomic medicine
BMI 701, Introduction to biomedical informatics
BMI 726, Big data innovations in population science

Workshop elective: min 0, max 1

Harnessing the research culture within Harvard, we would allow students to have the option to participate in relevant research seminars. Examples of this are:

AC 298r, Interdisciplinary Seminar in Computational Science & Engineering
This course, centered on the Institute for Applied Computational Science (IACS) seminar series, provides broad exposure to cutting-edge topics, applications, and unifying concepts in Computational Science & Engineering. Students read, present and discuss journal articles related to IACS talks, attend the seminars and meet with visiting speakers. Topics covered include scientific visualization, computational approaches to disease, mathematical neuroscience, computational archaeology, and computational finance.

Gov3009r, Applied Statistics
This course, centered on the Institute for Quantitative Social Science (IQSS) Applied Statistics seminar series, provides broad exposure to cutting-edge topics, applications, and unifying concepts in the application of statistics and data science. Students read, present and discuss journal articles related to IQSS talks, attend the seminars, and meet with visiting speakers.
E. Mapping learning outcomes to courses

<table>
<thead>
<tr>
<th>Outcome</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>Electives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Build statistical models</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Stat139 Stat149 Stat220</td>
</tr>
<tr>
<td>Design experiments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>Stat140 Stat240</td>
</tr>
<tr>
<td>Machine learning &amp; optimization</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>CS181 CS281</td>
</tr>
<tr>
<td>Acquire, clean &amp; manage data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>AM205 AC209</td>
</tr>
<tr>
<td>Visualized data for understanding</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>CS171</td>
</tr>
<tr>
<td>Collaborate within teams</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Reproducible data analysis</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational infrastructure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Handle large data</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CS205 AC209</td>
</tr>
<tr>
<td>Ethics/legal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Problem solving</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

F. Thesis Option

Some students may wish to make a deeper dive into a research topic in data science. For these students we propose an option for completing the master of science degree with a master’s thesis. Students pursuing the thesis option will have the opportunity to work closely with a faculty member on a substantial research project. We believe this will be a particularly attractive option for students planning to apply to doctoral programs after completing the Data Science master’s degree.

Students interested in pursuing the thesis option will need to find a faculty advisor and submit a thesis proposal for approval prior to the start of their third semester in the program. If the thesis proposal is approved, the thesis would satisfy the Data Science Research Experience (or Capstone) requirement (M6). Students exercising the thesis option will be allowed to count up to 3 300-level reading and research courses towards their degree requirements (see chart below). Some students will be able to successfully complete their theses by the end of the third semester, especially if they spend time during the summer getting started on their research project. Others will choose to stay for a fourth semester to finish their thesis project.

Expected schedule of courses (SM with thesis):

<table>
<thead>
<tr>
<th>Fall 1</th>
<th>Spring 1</th>
<th>Fall 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1. Data Science 1</td>
<td>M4. Data Science 2</td>
<td>elective</td>
</tr>
<tr>
<td>M2. Computing for DS 1</td>
<td>M5. Computing for DS 2</td>
<td>Reading and research course</td>
</tr>
<tr>
<td>M3. Critical Thinking and DS</td>
<td>elective</td>
<td>Reading and research course</td>
</tr>
<tr>
<td>elective</td>
<td>elective</td>
<td>Reading and research course</td>
</tr>
</tbody>
</table>
3. Secondary Field in Data Science for those doing a Ph.D. at Harvard

One of our priorities in developing this proposal is to allow Ph.D. students the opportunity to demonstrate their advanced knowledge of data science through gaining a secondary credential in Data Science at Harvard. In concert with the SM in Data Science, we propose to establish a graduate secondary field in Data Science. The secondary field will require 5 courses in Data Science. These courses will be a subset of the master’s degree requirements (see details below). This will mirror the relationship between the current master’s degrees and secondary field in Computational Science and Engineering (CSE). After completing the necessary courses, students will be required to demonstrate mastery of the programs learning outcomes by giving a final presentation on a data science project to a faculty committee.

The secondary field will provide the opportunity for doctoral students across GSAS to deepen their knowledge of data science and interact with likeminded students and faculty, without requiring them to take on too many courses outside of their field of study. Some students may, after completing the requirements of the secondary field, wish to devote more of their time to a deeper study of data science. These students can request permission from the Director of Graduate Studies to continue on towards the SM in Data Science. These students must propose a suggested plan of study for completing the master’s degree requirements. Three letters of recommendation will be required, at least one of which will be from an advisor in the student's home department. Approval is not automatic, and the Course Director may suggest some changes to the plan of study. It is important to note that courses which are a requirement for the student’s Ph.D. normally cannot be double-counted towards the SM degree. Upon completion of the 12 courses the student will file a degree application with the Registrar's Office. This degree application requires the signature of the course director. The SM degree will then be awarded.

Secondary Field Requirements: 5 courses

<table>
<thead>
<tr>
<th>SF requirements</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Technical core courses (M1, M2, M4, and M5)</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Critical Thinking and Data Science (M3)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3. Statistical electives</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4. Computer Science electives</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

4. Anticipated enrollment, student body and logistics

Many will come to this program from the fields of statistics, CS, biostatistics, econometrics or math; others will see this program as an avenue for career growth or to advance challenging research agendas. This diversity presents a tremendous opportunity for students to learn from each other and establish networks and collaborations that will advance the field and their careers. Serving the teaching, advising, and career planning needs of such an intellectually diverse, high-potential group of students will present
special challenges. An additional assumption is that the admitted students will be a blend of mid-career professionals, traditional post-ABs, and Ph.D. students wanting to complete a secondary field in DS. The services needed by the first two groups of students are expected to include career and financial aid services as well as advising. Financial support may include fellowships, sponsored research, and teaching fellow appointments.

In the inaugural year we envision a class size of roughly 40 students. Some of these students will come from reducing the class size of the CSE degree program (which will be discussed in a moment), where a significant number of students already specialize in aspects of DS. We believe that, in time, this program could grow to 100 students in a steady state. If the Data Science Lab is incorporated into a Data Science Institute in Allston, these students would have the Institute as their physical and intellectual home.

This heterogeneous population of students is expected to focus on a blend of

- The “data” side of data science (e.g., accessing, assembling, manipulating, and processing data). This will be crystallized in our core course M2.
- The “science” side of data science (e.g., describing, predicting, and gaining understanding from data). This will be represented by our core courses M1 and M4.

Of course there is great interplay between the two topic groupings, with no firm boundaries. Students going into or coming from industry will typically delve deeply into one of these but with a broad knowledge of both. The degree highlighted here is flexible, allowing for appropriate specialization as needed by students. A small fraction of the DS master’s level students will take this degree as a preparatory part of training to become Ph.D.-level researchers; for them also, the intended area of application will determine the balance of courses chosen.

5. Overlaps and complementarity with existing Harvard programs

SM and ME in Computational Science and Engineering (CSE)
The design of this program borrows from the experience of the one-year SM and two-year ME programs launched in 2013 and 2014, respectively, by SEAS and administered by the Institute for Applied Computational Science (IACS). These programs provide rigorous training in the mathematical and computing foundations of CSE and have served as Harvard’s master’s-level offerings in DS as the field has emerged. The program’s curriculum is built around four core courses in Applied Math and Computer Science that train students in methods and techniques from these disciplines. The core courses are complemented by additional elective courses in Applied Math, Computer Science, and Statistics. Independent research projects and elective courses focusing on the application of computation to one or more domains complement the foundational coursework.

The master of science (SM) degree is currently awarded for the successful completion of eight semester-length courses at Harvard. A master of engineering (ME) degree is awarded for the successful completion of eight courses (including the same core courses required by the SM degree) as well as a research thesis during a second year.

SM/ME course requirements at a glance:

<table>
<thead>
<tr>
<th>SM requirements</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Core</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Applied Math electives</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
AC denotes Applied Computation, AM denotes Applied Math, CS denotes Computer Science.

Core courses:
1. AM 205 - Advanced Scientific Computing: Numerical Methods
2. CS 205 - Computing Foundations for Computational Science
3. AM 207 - Advanced Scientific Computing: Stochastic Optimization Methods
4. CS 207 - Systems Development for Computational Science

Applied Math electives
1. AM 201 Physical mathematics I
2. AM 202 Physical mathematics II
3. AM 221 Advanced optimization
4. AC274, Computational modeling of fluids and soft matter
5. AC275, Computational design of materials
6. Stat139, Statistical sleuthing
7. Stat149, Generalized linear models
8. Stat170, Quantitative analysis of capital markets
9. Stat210, Probability I
10. Stat220, Bayesian data analysis
11. Stat221, Statistical computing and learning
12. Stat225, Spatial statistics
13. Stat285r, Statistical machine learning

Computer Science electives
1. AC 209 Data science
2. CS 124 Data structures and algorithms
3. CS 165 Data systems
4. CS 171 Visualization
5. CS 181 Machine learning
6. CS 221 Computational complexity
7. CS 222 Algorithms at the ends of the wire
8. CS 226r Efficient algorithms
9. CS 228 Computational learning theory
10. CS 246 Advanced computer architecture
11. CS 262 Distributed computing
12. CS 281 Advanced machine learning

Program admissions and enrollment to date:

<table>
<thead>
<tr>
<th>Matriculation year</th>
<th>Applications</th>
<th>Offers</th>
<th>Admit rate</th>
<th>Class size</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SM</td>
<td>ME</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016/17</td>
<td>374</td>
<td>181</td>
<td>555</td>
<td>70</td>
<td>13%</td>
</tr>
<tr>
<td>2015/16</td>
<td>236</td>
<td>119</td>
<td>355</td>
<td>61</td>
<td>17%</td>
</tr>
<tr>
<td>2014/15</td>
<td>98</td>
<td>65</td>
<td>163</td>
<td>51</td>
<td>31%</td>
</tr>
<tr>
<td>2013/14</td>
<td>146</td>
<td>N/A</td>
<td>148</td>
<td>40</td>
<td>27%</td>
</tr>
</tbody>
</table>

Admit rate is the percentage of applicants who received offers for the SM or ME in Computational Science and Engineering (CSE). Yield is the percentage of offers who joined the class.
The AM Program in Statistics

The Statistics Department has run an intensive year of coursework leading to the Master of Arts. On average one student a year progresses from the AM program to the department’s Ph.D. program; roughly a quarter go on to other Ph.D. courses elsewhere. In recent years some high-performing undergraduate non-statistical concentrators have taken sufficient courses and been awarded an AM with permission of the Department. Additionally, each year a couple of Ph.D. students in other departments get a secondary AM degree to bolster their statistical education and academic record.

Requirements:

Numbering for Statistics courses begins at the 100 level, denoting courses primarily focused on undergraduates. 200 level courses are lecture-based courses primarily for Ph.D. students, and courses numbered 300 and above are seminar-based courses primarily for Ph.D. students.

Requirements for the AM in Statistics are as follows:

- The satisfactory completion of 8 letter-graded semester-length courses taken within the Department of Statistics and approved by the department.
- 8 courses must be at the level of Statistics 110 or above. The actual courses will vary according to the student’s interest and preparation and will be determined in consultation with the student’s advisor.
- Statistics 110 (or 210a) and Statistics 111 are required.
- At least one 200-level course is required.
- At least 2 courses at the interface of theory and applications are required. These include: Stat 115, 121, 131/231, 139, 140, 149, 160/260, 170, 183, 186, 220, 221, 225, 230, 232r, 240, 244, and 245.
- Terminal AM students must earn a B average in their Statistics courses and no more than one C.
- Terminal AM students may take 300-level courses; however, only one 300-level course will count toward the required 8 letter-graded half-courses, and prior approval by the advisor and the Director of Graduate Studies is required for the 300-level course to count. The 300-level course would take the place of one of the non-200-level courses.

Generalist courses:

1. Stat 110, Introduction to probability
2. Stat 111, Introduction to theoretical statistics
3. Stat 121, Data science
4. Stat 131, Time series and prediction
5. Stat 139, Statistical sleuthing
6. Stat 140, Design of experiments
7. Stat 149, Generalized linear models
8. Stat 160, Design and analysis of sample surveys
10. Stat 183, Learning from big data
11. Stat 186, Causal inference
12. Stat 220, Bayesian data analysis
13. Stat 221, Statistical computing and learning
14. Stat 225, Spatial statistics
15. Stat 230, Multivariate statistical analysis
16. Stat 232r, Topics in missing data
17. Stat 240, Matched sampling and study design
18. Stat 244, Linear and generalized linear models

Domain-focused courses:

1. Stat 115, Introduction to computational biology and bioinformatics
2. Stat 170, Quantitative analysis of capital markets
3. Stat 245, Statistics and litigation
Program admissions and enrollment in recent years:

<table>
<thead>
<tr>
<th>Matriculation year</th>
<th>Applications</th>
<th>Offers</th>
<th>Admit rate</th>
<th>Class size</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016/17</td>
<td>305</td>
<td>15</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015/16</td>
<td>283</td>
<td>27</td>
<td>9%</td>
<td>21</td>
<td>78%</td>
</tr>
<tr>
<td>2014/15</td>
<td>292</td>
<td>25</td>
<td>9%</td>
<td>15</td>
<td>77%</td>
</tr>
<tr>
<td>2013/14</td>
<td>280</td>
<td>24</td>
<td>9%</td>
<td>15</td>
<td>71%</td>
</tr>
<tr>
<td>2012/13</td>
<td>236</td>
<td>24</td>
<td>10%</td>
<td>18</td>
<td>75%</td>
</tr>
<tr>
<td>2011/12</td>
<td>230</td>
<td>24</td>
<td>10%</td>
<td>17</td>
<td>63%</td>
</tr>
<tr>
<td>2010/11</td>
<td>127</td>
<td>23</td>
<td>17%</td>
<td>17</td>
<td>60%</td>
</tr>
</tbody>
</table>

Admit rate is percentage of applications who received offers for the AM Program in Statistics. Yield is the percentage of offers that joined the class. Source: Graduate School of Arts and Sciences. Prepared by: Office of Institutional Research (oir@harvard.edu).

**Secondary AM in Statistics for those doing a Ph.D.**

Requirements: 8 courses offered by the Statistics Department at the level of Stat 110 and above. (Similar to the AM above.)

Process: Take at least 3 qualifying courses, including 2 from the Core. Upon completing the qualifying course, the prospective students email the Student Coordinator requesting approval to pursue the SM. Applicants must provide a list of the completed courses (and grades earned) taken thus far and suggest a plan of study for completing the requirements. Three letters of recommendation are required, at least one of which will be from an advisor in the student’s home department. Approval is not automatic, and the course director may suggest some changes to the plan of study. It is important to note that courses which are a requirement for the student's Ph.D. normally cannot be double-counted towards the SM degree. Upon completion of the 8 courses the student will file a degree application with the Registrar's Office. This degree application requires the signature of the course director. The SM degree is then awarded.

6. **Overlaps with other Harvard programs being developed**

**Harvard T. H. Chan School of Public Health—Health Data Science Degree Program**

The Harvard T. H. Chan School of Public Health has been working on a new 16-month coursework-based master of science degree in “Health Data Science,” which would be run by the Department of Biostatistics. This master’s degree will be offered in parallel with their existing Biostatistics master’s. The expected number of students is 15–20. The proposal is as follows.

60-credit degree (5 credits is equivalent to a semester-length FAS course)

7 compulsory courses (25 credits in total)

1. BIO 260, Introduction to data science (5 credits)
2. BIO 261, Data science II (2.5 credits)
3. BIO 222, Basics of statistical inference (5 credits)
4. BIO XXX, Applied machine learning (5 credits)
5. BIO XXX, Computing for big data (5 credits)
6. EPI 201, Introduction to epidemiology methods: 1 (2.5 credits)
7. BIO XXX, Computing for big data (5 credits)
A single 5 credit course in an aspect of “computer science” from the list

1. BIO 505 Database design and use for health research
2. BIO 514 Data structures
3. BMI 713 Computational statistics for biomedical science
4. CS 105, Privacy and technology
5. CS 124, Data structures and algorithms
6. CS 164, Software engineering
7. CS 165, Data systems
8. CS 171, Visualization
9. CS 181, Machine learning
10. CS 187, Computational linguistics
11. Stat 135, Statistical computing software (note: this will be relabeled Stat109 in 2016/17)
12. Stat 183, Learning from big data
13. Stat 171, Introduction to stochastic processes

A cornerstone project

BIO XXX, Health data science practice (7.5 credits)

Finally, 22.5 units need to be gained from the following courses:

1. BIO 223, Applied survival analysis
2. BIO 226, Applied longitudinal analysis
3. BIO 257, Advanced statistical genetics
4. BIO 512, Introduction to computational biology and bioinformatics
5. BIO 513, Advanced computational biology and bioinformatics
6. BIO 521, Introduction to social and biological networks
7. EPI 202, Elements of epidemiologic research: methods 2
8. EPI 203, Study design in epidemiologic research
9. EPI 204, Analysis of case-control and cohort studies
10. EPI 271, Propensity score analysis
11. EPI 288, Data mining and prediction
12. EPI 289, Causal inference
13. EPI 515, Measurement error and misclassification
14. ID 271, Advanced regression for environmental epidemiology
15. BMI XXX, Data visualization
16. BMI XXX, Precision medicine I: Integrating clinical and genomics data
17. BMI XXX, Precision medicine II: Genomic medicine
18. BMI 701, Introduction to biomedical informatics
19. BMI 726, Big data innovations in population science
20. ME 530, Clinical informatics

Harvard Medical School—Degree in Biomedical Informatics
Harvard Medical School has been working on a new 36-credit Master of Biomedical Informatics program that will replace the current two-year post-doctoral MMSc program. It has been configured as a full-time (1-year) program with the possibility of having the option for a 2-year, part-time option. This degree program would be available for post-baccalaureate students with strong quantitative backgrounds, medical students, postdoctoral students, and MDs needing didactic training to sit for the subspecialty board exam in clinical informatics.

The program, in order to ensure all incoming students have an adequate quantitative foundation for the program, will include an intensive “boot camp” in quantitative methods and skills. Weekly informatics seminars will be held to expose students to the broad range of research investigations and applications
that comprise the field of bioinformatics. The proposed curriculum for the academic year 2016–17 (pending approval) is:

<table>
<thead>
<tr>
<th>Course Name</th>
<th>HMS Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>August</strong></td>
<td></td>
</tr>
<tr>
<td>Bootcamp in Quantitative Methods</td>
<td>2</td>
</tr>
<tr>
<td><strong>Fall</strong></td>
<td></td>
</tr>
<tr>
<td>Foundations of Biomedical Informatics I</td>
<td>4</td>
</tr>
<tr>
<td>Computational Statistics for Biomedical Sciences</td>
<td>4</td>
</tr>
<tr>
<td>Health Information Technology: From Ideation to Implementation OR Quantitative Genomics</td>
<td>4</td>
</tr>
<tr>
<td><strong>Fall 1st Half</strong></td>
<td></td>
</tr>
<tr>
<td>Data Visualization</td>
<td>2</td>
</tr>
<tr>
<td><strong>Fall 2nd Half</strong></td>
<td></td>
</tr>
<tr>
<td>Precision Medicine I: Integrating Clinical and Genomic Data</td>
<td>2</td>
</tr>
<tr>
<td>Informatics Seminar I</td>
<td></td>
</tr>
<tr>
<td><strong>Spring</strong></td>
<td></td>
</tr>
<tr>
<td>Clinical Informatics OR Computational and Functional Genomics</td>
<td>4</td>
</tr>
<tr>
<td>Elective Course</td>
<td>4</td>
</tr>
<tr>
<td>Foundations of Biomedical Informatics II (Capstone)</td>
<td>4</td>
</tr>
<tr>
<td><strong>Spring 1st Half</strong></td>
<td></td>
</tr>
<tr>
<td>Data Science in Population Health</td>
<td>2</td>
</tr>
<tr>
<td><strong>Spring 2nd Half</strong></td>
<td></td>
</tr>
<tr>
<td>Precision Medicine II: Genomic Medicine</td>
<td>2</td>
</tr>
<tr>
<td>Informatics Seminar II</td>
<td></td>
</tr>
<tr>
<td><strong>Total HMS Credits</strong></td>
<td><strong>36</strong></td>
</tr>
</tbody>
</table>

7. **Inter-school collaboration**

This proposal is deliberately constructed to address the needs of at least two sets of colleagues in other schools. Coordinated planning with other schools has been aided by the work of Professors David Parkes, Hanspeter Pfister, and Neil Shephard, who have been contributing to the campus-wide Data Science Initiative discussion over the last 18 months, convened by Vice-Provost for Research Professor Richard McCullough.

At the moment HMS and HCSPH have DS-driven proposals that they would like to launch in 2017. These are detailed above. The DS program described here is designed so that these schools could opt into various parts of the foundation stone courses M1–M5 if they find that convenient. In particular, M1 and M2 were designed in consultation with colleagues in HCSPH.

The second set of colleagues we would be helping are those in professional schools that do not have sufficient critical mass to run their own data science sequences, although they have a subgroup of students who want to use the tools of this area. Examples are HBS, GSE, and HLS.

The students in our master’s program likewise may benefit from access to courses in professional schools which have a more specialist nature, e.g., those focusing on DS and genetics or DS and business.

Our core course M3 needs contributions from colleagues who are primarily in HLS, HKS, and HBS.
8. **Marketing and admissions**

Marketing and admissions will be handled by SEAS and GSAS, as IACS has done successfully with the existing CSE programs. Specifically, marketing communications to leading universities, attendance at graduate school fairs, and alumni network development will be part of the marketing effort.

Admissions to the master’s program will be overseen by the Director of Graduate Studies with review by the degree faculty Program Committee, and with assistance from the FAS Registrar and SEAS Office of Student Affairs.

9. **Administration and staff needs**

A. **Academic program leadership**

Primary responsibility for the delivery of this program will be jointly held by the SEAS Institute for Applied Computational Science (IACS) and the Department of Statistics. The Director of Graduate Studies for the Master’s in DS will have day-to-day responsibility for running the degree. The Program Committee will be responsible for curriculum, planning, assessment, admissions, and evaluation. The chair of this committee and the Director of Studies for the master’s degree will be jointly appointed by the Head of Statistics and the Faculty Director of IACS. Members of the committee will be drawn from SEAS and Statistics, but also from faculty in other departments at Harvard who are actively involved in the degree program (e.g., Math, IQSS, professional schools). Its constituency should be such that neither CS nor Statistics has a majority on it.

B. **Ongoing program evaluation**

IACS has developed a methodology for ongoing outcomes-based program evaluation. The IACS Advisory Board, comprised primarily of Harvard faculty and industry representatives, meets biennially to review curriculum and the performance of the SM, ME, and Secondary Field programs in CSE and to update the learning outcomes to ensure that graduates are well prepared for leadership in a rapidly changing field. This board includes many leaders in the DS field. We propose to schedule assessments and evaluations of the DS program so that they can be coordinated with the CSE reviews and the existing Advisory Board reorganized as appropriate to this expanded purview.

C. **Administrative implementation**

Implementation of the program, including mentoring, career support and placement, computing resources, financial aid (fellowships, sponsored research, grants), and industry engagement will be the responsibility of the Institute for Applied Computational Science and the SEAS Student Affairs Office. Specifically, Assistant Directors of Graduate Studies (ADGS) will support the master’s of data science program (SM) as well as the master’s in computational science and engineering programs (SM and ME). Students will be enrolled as graduate students within the Graduate School of Arts and Sciences (GSAS), and the SEAS Student Affairs Office will provide administrative support.
10. Advising structure

The Director of Graduate Studies will take formal responsibility for advising students in creating a meaningful program sensitive to the student’s needs. Each master’s student will be assigned an advisor from the faculty program committee and be expected to meet with that advisor periodically. Front-line advising will be delegated to the two Assistant Directors of Graduate Studies, who will actively work to develop independent research projects and external research opportunities for the master’s students and those enrolled in the secondary field program.

11. Benefits to SEAS and Statistics

A strategic priority for both CS and Statistics is to enhance our research and teaching in the area of data science and statistics and computing. It is important that we establish a critical mass of faculty and research staff to do so.

To spark this drive we propose to set up a “Data Science Lab”, which would sit physically in Statistics and CS but be administered as part of IACS. The Lab would have faculty and student affiliates from CS, Statistics, and other departments, such as Mathematics, who have research interests in big data. The goal would be to accelerate work in DS without adding to administrative overhead and strongly signal that we are working together. The Lab would serve as a much needed bridge between the faculties of Statistics and Computer Science, sponsoring seminars and lunch-time workshops. In addition to helping develop productive research collaborations, this tangible connection between the two departments would be a highly valuable recruiting tool for new faculty hires working at the CS/Statistics interface. If the Data Science Institute is founded, we would imagine this center would be folded into the Institute.

The Lab’s chief role will be to support a corps of postdoctoral researchers who will also have teaching responsibilities for the data science courses. A significant majority of these postdocs would be appointed to work with the faculty members who are teaching core courses in the master’s program, while the Lab could also fund Ph.D. students working with those faculty members.

The new master’s of data science program clearly defines an emerging academic area, integrates content across disciplines (statistics, computer science, applied computation and applied math), and positions Harvard FAS/SEAS as an intellectual leader in modern statistics and data science.

12. Incentives and benefits to ladder faculty, Areas, and Departments

We suggest following the IACS approach to their CSE degree and support faculty that teach on this degree by supporting their research. This would be carried out by funding postdocs but could be given as payments into research accounts as well. A significant focus is to minimize the generation of new non-research overhead.
13. **Risks to SEAS and Statistics**

Some Statistics Department faculty have concerns that data science may overwhelm other important aspects of statistics that do not focus on big data. Important examples include experimental design and causal methodology. On the other hand, they recognize that data science is a highly active area and that statistics needs to play a central role or risks being marginalized. The appropriate response to the above concerns about experiments and causality is to invest in those areas as the department grows.

Some 9-month master’s degrees in DS at leading universities do not have a strong reputation. We propose something very different: a challenging, three-semester-long program with the potential to be a transformative experience. In addition, we have crafted elements such as M4, which uniquely brings into the degree program some of Harvard’s unique strengths in building leaders.

14. **Industrial relations and career development**

Building on existing industrial relations and career development efforts in place at the IACS, the launch of the new master’s program would provide an opportunity to engage further with companies and organizations directly around data science.

Specifically, the institute would develop/evolve its current advisory board to advise on key educational efforts of the master’s programs, as well as to provide context and feedback on current data science and computational science. In addition, there would be exploration of industrial-funded fellowships, an industrial membership consortium, and continued sponsorship opportunities of public events.

For career development, the institute plans to continue to provide career counseling by working with Harvard’s Office of Career Services and continuing to develop a robust recruiting program with companies, surveying alumni and students, building alumni networks, and organizing company site visits. Professional resume books and access to students will be provided to selected companies to support the recruiting effort.

15. **Community building**

The IACS has had a successful track record in the last 5 years building community across Harvard and beyond with a bi-weekly seminar series, a two-week-long Computefest and annual symposium, cross-University collaborations with organizations like the i-lab and IQSS, and with undergraduates interested in data science, and very early collaborations with Harvard Business School (digital initiative) and Harvard Kennedy School (tech in gov). Outside Harvard, in addition to industry relations and career development for students, the IACS has begun to work with state, city, and national governments. The IACS holds workshops and special seminars and is developing short-term exchange and research programs with partners at other academic institutions internationally.

It is anticipated that the IACS would continue and expand these efforts to help support the launch of the data science master’s. The Data Science Lab would tap into and contribute to the seminars and symposia and provide a nexus for seeding collaborative research. Convening spaces and programmatic initiatives at the new Data Science Institute will allow this growing community to flourish and continue to catalyze
interactions among Harvard schools, industry, government, and the larger academic community interested in data science and its remarkable applications.

16. Financial model

Tuition assumptions for 2016/17

<table>
<thead>
<tr>
<th>Subject</th>
<th>Months of program</th>
<th>Months of courses</th>
<th>Total tuition</th>
<th>Tuition per month on campus</th>
<th>Annualized tuition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford Statistics: DS</td>
<td>9</td>
<td>8</td>
<td>46,000</td>
<td>5,750</td>
<td>46,000</td>
</tr>
<tr>
<td>Columbia DS</td>
<td>9</td>
<td>8</td>
<td>51,300</td>
<td>6,412</td>
<td>51,300</td>
</tr>
<tr>
<td>Stanford Comp Sci &amp; Eng</td>
<td>9</td>
<td>8</td>
<td>50,990</td>
<td>5,665</td>
<td>50,990</td>
</tr>
<tr>
<td>Berkeley Eng in DS</td>
<td>9</td>
<td>8</td>
<td>55,954</td>
<td>6,994</td>
<td>55,954</td>
</tr>
<tr>
<td>Harvard GSAS Statistics</td>
<td>9</td>
<td>8</td>
<td>41,832</td>
<td>5,229</td>
<td>41,832</td>
</tr>
<tr>
<td>Harvard Chan SPH Health DS</td>
<td>16</td>
<td>12</td>
<td>62,400</td>
<td>5,200</td>
<td>41,200</td>
</tr>
<tr>
<td>Harvard FAS Proposed DS</td>
<td>16</td>
<td>12</td>
<td>74,500</td>
<td>6,208</td>
<td>49,500</td>
</tr>
</tbody>
</table>

The total tuition we are proposing is higher for the proposed Harvard DS program than is typical for residential peer DS degrees, but that is because our degree would be longer than is usual. The tuition per month comes in as pretty typical, though.

Below is an initial three-year financial model to show the financial impact of the data science master’s degree program based on the program tuition of $74,500 and an initial cohort size of 40 students, and estimated direct expense associated with administering the program. Additional assumptions are that there would be revenue from industrial sponsorships, limited fellowship/financial aid funding, and some grants.

This model is based on initial experience administering the CSE master’s program at SEAS.
Three-year financial model

3/14/2016

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of students</td>
<td>Fall 40</td>
<td>Spring 40</td>
<td>Fall 80</td>
</tr>
<tr>
<td>tuition for 3-semester program</td>
<td>$74,500</td>
<td>$74,500</td>
<td>$75,000</td>
</tr>
<tr>
<td>tuition rate per semester</td>
<td>$24,835</td>
<td>$24,835</td>
<td>$25,704</td>
</tr>
<tr>
<td>annual tuition rate increase</td>
<td>3.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUITION REVENUE</td>
<td>$993,400</td>
<td>$993,400</td>
<td>$1,028,169</td>
</tr>
<tr>
<td>Grants</td>
<td>$25,000</td>
<td>$25,000</td>
<td>$40,000</td>
</tr>
<tr>
<td>Industry sponsorships</td>
<td>$50,000</td>
<td>$50,000</td>
<td>$75,000</td>
</tr>
<tr>
<td>OTHER REVENUE</td>
<td>$75,000</td>
<td>$75,000</td>
<td>$115,000</td>
</tr>
<tr>
<td>TOTAL REVENUE</td>
<td>$1,068,400</td>
<td>$1,068,400</td>
<td>$2,171,338</td>
</tr>
</tbody>
</table>

Master's Program direct expense

<table>
<thead>
<tr>
<th>Staff:</th>
<th>TOTAL 400,000</th>
<th>400,000</th>
<th>408,000</th>
<th>408,000</th>
<th>416,160</th>
<th>416,160</th>
</tr>
</thead>
<tbody>
<tr>
<td>Director of Master's Program</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrator/career services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial Outreach/research projects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assistant Director Graduate Studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecturers (1.5)/visiting faculty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other: marketing, travel, events TOTAL</td>
<td>$50,000</td>
<td>$40,000</td>
<td>$50,000</td>
<td>$40,000</td>
<td>$50,000</td>
<td>$40,000</td>
</tr>
<tr>
<td>Data Science Lab (20% of tuition)</td>
<td>$198,680</td>
<td>$198,680</td>
<td>$411,268</td>
<td>$205,634</td>
<td>$425,662</td>
<td>$212,831</td>
</tr>
</tbody>
</table>

EXPENSE

| EXPENSE | 648,680 | 638,680 | 869,268 | 653,634 | 891,822 | 668,991 |

NET REVENUE Before FAS/SEAS overhead

| NET REVENUE Before FAS/SEAS overhead | $419,720 | $429,720 | $1,302,070 | $489,535 | $1,386,488 | $545,164 |

Other: Proposed Fellowship funds/financial aid

| Other: Proposed Fellowship funds/financial aid | $56,000 | $56,000 | $115,000 | $115,000 | $200,000 | $200,000 |

The Data Science Lab spending is modeled on 20% of annual tuition revenue. The funds would support post-docs, research accounts of faculty teaching on the degree, and seminars and workshops. Each of these items will foster faculty research and also improve the quality of the master’s degree.
17. Conclusion

This proposal is part of a larger move to establish the architecture necessary for Harvard to build human and social capital in data science and to excel in research through innovations in, and use of, data science. Other reforms being built at the moment include a DS track within the Statistics Concentration, a proposed Secondary Field in Data Science for Harvard College students, and potentially the building of a Data Science Institute to provide a campuswide focal point for the subject. In the meantime, this proposal also advocates the creation of a Data Science Lab, which would encourage cross-departmental research cooperation within the FAS in this area while avoiding the generation of new administrative overhead.

This master’s degree has been carefully designed to allow students the opportunity to build technical, teamwork and leadership skills. Its three semester length is designed to allow the students to have time to explore different aspects of the subject and work on research projects or pursue internships.

Appendix

<table>
<thead>
<tr>
<th>School</th>
<th>Degree</th>
<th>Length (Years)</th>
<th>First Year</th>
<th>Total Program</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnegie Mellon University</td>
<td>MS Comp'l Finance</td>
<td>1.5</td>
<td>52,600</td>
<td>78,900</td>
<td>Degree granted by Business School</td>
</tr>
<tr>
<td>UCBerkeley</td>
<td>M. Engineering in Data Science</td>
<td>1</td>
<td>52,000</td>
<td>52,000</td>
<td></td>
</tr>
<tr>
<td>Columbia</td>
<td>MS Data Science</td>
<td>1.25</td>
<td>51,300</td>
<td>64,125</td>
<td>A few students finish in one year</td>
</tr>
<tr>
<td>Columbia</td>
<td>MS Computer Science</td>
<td>1.25</td>
<td>49,200</td>
<td>61,500</td>
<td>A few students finish in one year</td>
</tr>
<tr>
<td>Stanford</td>
<td>MS Comp'l and Mth. Eng'Ing</td>
<td>1</td>
<td>47,600</td>
<td>47,600</td>
<td></td>
</tr>
<tr>
<td>MIT</td>
<td>MS in Com. Design and Optimiz.</td>
<td>1</td>
<td>43,210</td>
<td>43,210</td>
<td>Academic focused (vs. industry) program</td>
</tr>
<tr>
<td>Harvard SEAS</td>
<td>SM Comp'l Science and Eng'Ing</td>
<td>1</td>
<td>40,416</td>
<td>40,416</td>
<td></td>
</tr>
<tr>
<td>Harvard SEAS</td>
<td>ME Comp'l Science and Eng'Ing</td>
<td>2</td>
<td>40,416</td>
<td>80,832</td>
<td>Second year is research only (full tuition)</td>
</tr>
<tr>
<td>Georgia Tech</td>
<td>MS Comp'l and Eng'Ing</td>
<td>1.5</td>
<td>25,170</td>
<td>37,755</td>
<td></td>
</tr>
</tbody>
</table>

*Residential Programs only. Data as of Fall 2014 except Columbia MS Data Science which is as of Spring 2015.

<table>
<thead>
<tr>
<th>School</th>
<th>Degree</th>
<th>Length (Years)</th>
<th>First Year</th>
<th>Total Program</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnegie Mellon University</td>
<td>MS Comp'l Finance</td>
<td>1.5</td>
<td>56,346</td>
<td>84,520</td>
<td>Degree granted by Business School</td>
</tr>
<tr>
<td>UCBerkeley</td>
<td>M. Engineering in Data Science</td>
<td>1</td>
<td>55,704</td>
<td>55,704</td>
<td></td>
</tr>
<tr>
<td>Columbia</td>
<td>MS Data Science</td>
<td>1.25</td>
<td>54,954</td>
<td>68,692</td>
<td>A few students finish in one year</td>
</tr>
<tr>
<td>Columbia</td>
<td>MS Computer Science</td>
<td>1.25</td>
<td>52,704</td>
<td>65,880</td>
<td>A few students finish in one year</td>
</tr>
<tr>
<td>Stanford</td>
<td>MS Comp'l and Mth. Eng'Ing</td>
<td>1</td>
<td>50,990</td>
<td>50,990</td>
<td></td>
</tr>
<tr>
<td>Harvard SEAS PROPOSED</td>
<td>SM Comp'l Science and Eng'Ing</td>
<td>1</td>
<td>49,500</td>
<td>49,500</td>
<td></td>
</tr>
<tr>
<td>Harvard SEAS PROPOSED</td>
<td>ME Comp'l Science and Eng'Ing</td>
<td>2</td>
<td>49,500</td>
<td>74,250</td>
<td>Second year is research only (half tuition)</td>
</tr>
<tr>
<td>MIT</td>
<td>MS in Com. Design and Optimiz.</td>
<td>1</td>
<td>46,288</td>
<td>46,288</td>
<td>Academic focused (vs. industry) program</td>
</tr>
<tr>
<td>Georgia Tech</td>
<td>MS Comp'l Science and Eng'Ing</td>
<td>1.5</td>
<td>26,963</td>
<td>40,444</td>
<td></td>
</tr>
</tbody>
</table>

*Residential Programs only. Original data as of Fall 2014 except Columbia MS Data Science which is as of Spring 2015 has been increased by 3.5% per year for all peers.
A. Peer Benchmarking

The information given below is taken from the relevant institutions’ websites. It is a small subset of all data science degrees currently on offer. Our list focuses on the some of the best known. We also have a more extensive list of degrees offered at other institutions which is available on request.

Carnegie Mellon University
Computational Data Science (12-, 16-, or 21-month program)

The MCDS program is designed for students with a degree in computer science, computer engineering or a related degree from a highly ranked university. Students can pick from two majors in the MCDS program: Systems and Analytics. Both require the same total number of course credits split among required core courses, electives, data science seminar and capstone courses specifically defined for each major. The degree can also be earned four different ways, depending on the length of time students wish to spend working on it. Regardless of the timing option, all MCDS students must complete a minimum of 144 units to graduate.

Options available:

- **Part-Time Timing** — a degree of variable time, usually entailing several years.
- **Short Timing** — a 12-month degree consisting of fall, spring and summer semesters of study. Each semester comprises a minimum of 48 units. This timing is limited to students who have several previous internships. Students graduate in August.
- **Medium Timing** — a 16-month degree consisting of study for fall and spring semesters, a summer internship, and fall semester of study. Each semester comprises a minimum of 48 units. This timing is typical for most students. Students graduate in December.
- **Long Timing** — a 20-month degree consisting of study for fall and spring semesters, a summer internship, and a second year of fall and spring study. Each semester comprises a minimum of 36 units. Students graduate in May.

To earn an MCDS degree, students must pass courses in the core curriculum, the MCDS seminar, a concentration area and electives. They must also complete a capstone project in which you work on a research project at CMU or on an industry-sponsored project. In total, they will complete 144 eligible units of study, including eight 12-unit courses, two 12-unit seminar courses and one 24-unit capstone course. Students must take five core courses. The remainder of the 12-unit courses with course numbers 600 or greater can be electives chosen from the SCS course catalog. Any additional non-prerequisite units taken beyond the 144 units are also considered electives.

**Systems Major**

1. Core Curriculum. Pick five, with at least three project (*) courses.

   i. Operating Systems Implementation(*)
   ii. Database Applications
   iii. Parallel Computer Architecture & Programming (*)
   iv. Cloud Computing (*)
   v. Distributed Systems (*)
   vi. Advanced Cloud Computing (*)
   vii. Advanced Databases (*)
   viii. Storage Systems (*)
   ix. Multimedia Databases and Data Mining

2. Seminar in Data Systems (15-649 A in the fall and 15-649 B in the spring)
3. Capstone project (15-649 C, D or E) in the second fall semester
4. Three electives (any graduate-level course 600 and above in the School of Computer Science)

**Analytics Major**

1. Core Curriculum (five courses)
   a. Choose two courses in Machine Learning/Statistics:
      i. Machine Learning
      ii. Machine Learning for Text Mining
      iii. Advanced Machine Learning
      iv. Machine Learning with Big Data Sets
   b. Choose two courses in Software Systems:
      i. Design and Engineering of Intelligent Info Systems
      ii. Cloud Computing
      iii. Information Systems Project
      iv. Search Engines
   c. Choose one course with a focus on Big Data:
      i. Multimedia Databases and Data Mining
      ii. Machine Learning with Big Data Sets
      iii. Big Data Analytics
      iv. Large-Scale Multimedia Analysis

2. Data Science Seminar in the first fall semester
3. Capstone Planning Seminar in the first spring semester
4. Data Science Analytics Capstone in the second fall semester
5. Three electives — any graduate level course 600 and above in SCS

Every student must complete a capstone project that integrates classroom experience with hands-on research. Working alone or as part of a team, they'll solve a research problem with either a Carnegie Mellon or industry partner.

**Columbia University**
**Master of Science in Data Science**

The Master of Science in Data Science allows students to apply data science techniques to their field of interest, building on four foundational courses offered in our Certification of Professional Achievement in Data Sciences program. Students have the opportunity to conduct original research, included in a capstone project, and interact with our industry partners and faculty. Students may also choose an elective track focused on entrepreneurship or a subject area covered by one of our six centers.

**ELIGIBILITY REQUIREMENTS**

- Undergraduate degree
- Prior quantitative coursework (calculus, linear algebra, etc.)
- Prior introductory computer programming coursework
REQUIRED COURSES

Students are required to complete a minimum of 30 credits, including 21 credits of required/core courses and 9 credits of electives.

REQUIRED / CORE COURSES

1. STAT W4105 Probability
2. CSOR W4246 Algorithms for Data Science
3. STAT W4702 Statistical Inference and Modeling
4. COMS W4121 Computer Systems for Data Science
5. COMS W4721 Machine Learning for Data Science
6. STAT W4701 Exploratory Data Analysis and Visualization
7. ENGI E4800 Data Science Capstone and Ethics

New York University

Master of Science in Data Science

It is a two-year program (either 3 or 4 semesters in length) where one needs to complete 36 credits; half of which are required courses and half of which are electives. The inaugural class launched in the fall of 2013. Admission to the Master of Science in Data Science requires substantial but specific mathematical competencies, typical of a major in mathematics, statistics, engineering, physics, theoretical economics, and computer science with sufficient mathematical training. In addition, applicants are required to have some training in programming and basic computer science.

To be considered for the program, students are required to have taken:

- Calculus I
- Linear Algebra
- Introduction to Computer Science (or an equivalent “CS-101” programming course)
- One of Calculus II, Probability, Statistics, or an advanced physics, engineering, or econometrics course with heavy mathematical content

Preference is given to applicants with prior exposure to machine learning, computational statistics, data mining, large-scale scientific computing, operations research (either in an academic or professional context), as well as to applicants with significantly more mathematical and/or computer science training than the minimum requirements.

The MS in Data Science curriculum requires a capstone project that makes the theoretical knowledge students gain in the program operational in realistic settings. During the project, students will go through the entire process of solving a real-world problem: from collecting and processing real-world data, to designing the best method to solve the problem, and finally, to implementing a solution. The problems and datasets engaged with will come from real-world settings identical to what they might encounter in industry, academia, or government.

REQUIRED COURSES

1. DS-GA-1001: Introduction to Data Science
2. DS-GA-1002: Statistical and Mathematical Methods
3. DS-GA-1003: Machine Learning and Computational Statistics
4. DS-GA-1004: Big Data
5. DS-GA-1005: Inference and Representation
6. DS-GA-1006: Capstone Project and Presentation in Data Science

Stanford University (3 degree programs)

Biomedical Informatics (within the Stanford School of Medicine)

Students in the BMI program may focus on any aspect of information management and analysis relevant to biomedical research. The BMI core curriculum requires training in 4 areas:

- Core Biomedical Informatics - Students are required to complete 17 units in the area of Biomedical Informatics.
- Computer Science, Statistics, Mathematics & Engineering - Students are required to complete 18 units in the area of Computer Science, Statistics, Mathematics & Engineering. They are expected to create a program of study with a mixture of graduate-level courses in computer science, statistics or other technical informatics-related disciplines that allows them to achieve in-depth mastery of these areas.
- Social and Ethical Issues - Students are required to complete 4 units in the area of Social and Ethical Issues. These courses are designed to familiarize students with issues regarding ethical, legal, social, organizational and behavioral aspects of the impact of biomedical informatics technologies on society in general.
- Unrestricted Electives - Students may fulfill this requirement with any Stanford course, including courses taken to satisfy core curriculum prerequisites.

Part-time students take an average of 3 to 5 years to complete the 45 unit requirement. Students must complete a Master’s degree within 5 years of starting the program. Courses are offered online and on campus.

Pre-requisites/Program timeline

Applicants should have a strong background in calculus, computer programming, statistics, linear algebra, biology, and physiology. To be eligible for the online distance professional MS program, students must be employed full-time and remain employed throughout the duration of the degree program.

Master’s Degree in Computer Science, specialization in Information Management and Analytics

Students must choose one of ten predefined specializations, or a combination of any two specializations (dual depth) for this degree. Information Management and Analytics provides coverage of the principles underlying modern database and information management systems, as well as methods for mining massive data sets. The track spans topics ranging from developing applications for database and information systems; to system design, architecture, and management; to applying algorithms and techniques from data mining and machine learning to perform analyses over massive data sets. Related topics include distributed systems, networking, and security on the system side, as well as text mining, bioinformatics, web search, and social media on the applications side.

MS in Statistics: Data Science

The focused MSc track is developed within the structure of the current MS in Statistics and the MSc program in ICME (Institute for Computational and Mathematical Engineering). Students in the program will develop strong mathematical, statistical, computational and programming skills through the MS
requirements, and they will gain a fundamental data science education by focusing 18 units of elective courses in the area of data science and related courses.

Admission to the MS program is made by the Statistics admission committee, which has representation from the Data Science track steering committee. The total number of units in the degree is 45, 36 of which must be taken for a letter grade. The degree typically spans 5 quarters over 2 years.

Degree Requirements

Mathematical Core (12 units), approved courses include

1. CME302, Linear Algebra
2. CME304, Numerical Optimization or CME364A, Convex Optimization
3. CME305, Discrete Mathematics
4. CME308, Stochastic Methods in Engineering

Advanced Scientific Programming and High Performance Computing Core (6 units, at least 3 in parallel computing)

Approved courses include:

1. CME212 Advanced Programming for Scientists and Engineering
2. CME214 Software Design in Modern Fortran for Scientists and Engineering
3. CS107, Computer Organization and Systems
4. CS249B, Large Scale Software Development

And for parallel/HPC (at least 3 units required):

1. CME213 Introduction to Parallel Computing using MPI, openMP and CUDA
2. CME342, Parallel Methods in Numerical Analysis
3. CS149, Parallel Computing
4. CS315A, Parallel Computer Architecture and Programming
5. CS315B, Parallel Computing Research Project
6. CS316, Advanced Multi-core Systems
7. CS344C, Cloud Simulation Systems

Students who do not start the program with a strong computational and/or programming background must take an extra 3 units to prepare themselves adequately for their courses.

Statistics Core (12 units), approved courses include

1. STATS200, Introduction to Statistical Inference
2. STATS203/305, Regression Models / Statistical Modeling
4. STATS315B, Modern Applied Statistics: Data Mining

Domain Specialization or preparatory courses (9 units)
One or two of these courses may be used by the students that enter the program with insufficient linear algebra or programming experience to prepare for the core requirements in the M.S. track. Specialized courses include courses that further deepen the data science core. Some possibilities include:

1. CS347, Parallel and Distributed Data Management
2. STATS290
3. CS448, Topics in Computer Graphics
4. CS224W, Social and Information Network Analysis
5. STATS366/BIOS221, Modern Statistics for Modern Biology, Holmes/Martin (Summer)
6. Psych204A, Human NeuroImaging Methods, Wandell/Dougherty (Autumn)
7. Psych303, Human and Machine Learning (not given this year)
8. OIT367, Analytics from Big Data, Bayati (Winter)
9. BioMedin215, Data Driven Medicine, Shah (Autumn)
10. Energy240, Geostatistics, tbd (Spring)
11. BIOE214, Representations and Algorithms for Computational Molecular Biology, Altman (Autumn)

Practical component (6 units)

1. A capstone project, supervised by a faculty member and approved by the steering committee: the capstone project should be computational in nature; students should submit a one-page proposal, supported by the faculty member, to the steering committee for approval.
2. Clinics, such as the new Data Science Clinic offered by ICME starting Fall 2013.
3. Other courses that have a strong hands-on and practical component, such as STATS390 (Statistical Consulting).