Proposed Master of Science Program in
Health Data Science

Department of Biostatistics
February 27, 2016

The following proposal was presented to, and unanimously approved by, the Harvard T.H. Chan School of Public Health faculty on December 16, 2015.

BACKGROUND AND DEVELOPMENT OF THE DEGREE

Current Master’s Programs in the Department of Biostatistics:
The Department of Biostatistics offers three Master’s degree programs in Biostatistics (45 credit, 60 credit, 80 credit) that emphasize competence in biostatistical theory and methods, and an 80 credit Master’s Program in Computational Biology and Quantitative Genomics. All of these programs have been highly successful in training students.

The health and biomedical fields beyond computational biology and genetics/genomics are being transformed by high-throughput data collection technologies as well. These technologies yield more information than can be readily examined and absorbed. Examples in public health and biomedicine include continuous time remote sensing data acquired via mobile devices, electronic medical records, and remote sensing data that records levels of environmental exposures on a continuous basis in both space and time, among others. Recently, the National Research Council (NRC, 2013) report “Frontiers in Massive Data Analysis” documented the pervasiveness of massive data in science, that this trend is only likely to increase over time, and that a key challenge now is to develop quantitative and computational tools to draw reliable inferences from all of this information.

To address these issues, we are proposing to create a new 60 credit SM in Health Data Science. The choice of a 60 credit requirement follows from our successful 60 credit Masters program in Biostatistics. This program, which will be offered through the Department of Biostatistics, will build on our existing core competencies in statistical methods, inference, and computing as well as introduce new courses in data science and machine learning. This coursework will include a project-based research course that tests all competencies taught by the program through novel data analysis and presentation of multiple real-world big data in health science.

There is a definite need for a new Master’s program that better prepares graduates to handle “Big Data” in addressing the biomedical research questions that are becoming increasingly commonplace in hospitals and universities, research organizations, and the pharmaceutical and biotechnology industries.

Summary of New Program:
• Allow admittance of students without a post-baccalaureate degree to a new Master’s program in Health Data Science (60 credits), provided they have appropriate quantitative training as undergraduates or before arrival to the program.
• We are requesting to offer the degree of Health Data Science to graduates of this new program. We wish to obtain the appropriate approvals to use this new name. In the interim
(pending approval for this new degree name), we would offer the Master’s of Science in Biostatistics, with the area of interest being Health Data Science. In the proposal below we use the Health Data Science name.

- The new program will require 60 credits of course work, with 7.5 credits of course work focusing on a culminating research experience in which students integrate the knowledge learned from all the courses they have taken to analyze five real big health data from ongoing research projects on which faculty collaborate. This research course will be designed to tests all competencies targeted by the program.
- This new program would typically be of three semesters duration.
- The program would be designed as a terminal degree, giving students essential skills for the job market. At the same time, it would provide a route to a PhD in Biostatistics or other Quantitative or Computational Science with an emphasis in data science if the student wished to pursue that route.
- No changes to our current Master’s programs in either Biostatistics or Computational Biology and Quantitative Genomics. We anticipate that this program may reduce our applications to these existing programs somewhat, due to the popularity of data science in the health sciences today. Given the currently very large pool of applicants for the biostatistics masters programs, we expect we will have strong pools for both the existing programs and the new Health Data Science MS program. We will evaluate the extent of this interplay after the first few years of the new program.

**RATIONALE/PURPOSE**

- There is a growing need to train the next generation of scientists in quantitative methods for data-driven discovery in public health and biomedical research. The NRC report “Training Students to Extract Value from Big Data” (2014) predicted an enormous shortfall of data analysts and managers knowledgeable about data and their relevance. The report states: “It is becoming increasingly important to increase the pool of qualified scientists and engineers who can extract value from big data. Training students to be capable in exploiting big data requires experience with statistical analysis, machine learning, and computational infrastructure that permits the real problems associated with massive data to be revealed and, ultimately, addressed.” Therefore, demand for such people, while high now, is likely to only grow in the future.
- It is important to note that in public health and biomedical research, studies relying on large-scale data will ultimately be concerned with population assays. This brings statistical inference to the forefront. The National Research Council’s Frontiers in Massive Data Analysis explains: “The challenges for massive data go beyond the storage, indexing, and querying that have been the province of classical database systems and instead hinge on the ambitious goal of inference.” These considerations highlight the need for a data science program specializing in health sciences, as generic data science training may not fully meet the training goals in a public health program. In particular, there is a need to integrate statistics and computer science into a program offering substantive training in public health. This program will offer students the opportunity to integrate training in data science with formal didactic and research training in the health sciences.
- The Department of Biostatistics already offers excellent courses in statistical methods, inference, and computing that would provide rigorous training in data science, with several new courses in this area planned over the next two years. Department faculty are also well
integrated in various, basic science, clinical and epidemiologic research projects at the Harvard Chan School, Harvard University, and the affiliated hospitals that will provide excellent project based training.

- This new program will also play a key role in the Harvard’s university-wide Data Science Initiative. The School plays a key role in the university-wide planning committee for this Initiative, chaired by Vice Provost Rick McCulloch, with Drs. Xihong Lin and Francesca Dominici serving on it. It could well be the case that some of the classes listed in our Core below would ultimately be integrated into University-wide “core courses” in data science (applied linear regression and machine learning, for instance) that cut across different data science programs at the University. We would welcome such University-wide coordination as planning for this common core develops. Our coherent and cultivated specialized courses in health data science could potentially attract students working in undergraduate programs or an FAS Masters degree in data science seeking applications areas in the health sciences.

**Types of Students that would enroll:**
All candidates for admission to the MS in Health Data Science program should have successfully completed calculus through partial differentiation and multivariable integration, one semester of linear algebra or matrix methods, and either a two-semester sequence in probability and statistics or a two-semester sequence in applied statistics. Students should have practical knowledge of computer scripting and programming as well as some experience with a statistical computing package such as R or Python. We expect all accepted students will meet these requirements. To ensure students are familiar with R and Python, we will make online material on R and Python available to the enrolling students the summer before students enter the program (see Sample Curriculum provided below). We will also recommend accepted students take an online course similar to CS50 (Introduction to Computer Science) before entering into the program. Applicants should also show excellence in written and spoken English. Additional research or work experience would be considered beneficial, but not required. Students who enroll in this program will come with undergraduate degrees in the mathematical sciences or from allied fields (statistics, economics, etc.) or computer science, with a strong interest in health science. Students are expected to have the goal of working as a data scientist in teaching hospitals, universities, research organizations, the pharmaceutical and biotechnology industries, or government.

**Expected Size:**
Approximately 15-20 students per year. We do not anticipate the recruitment and admission of these students will affect enrollments in existing MS programs in the department. We do note however, that current enrollment in our existing Masters programs is slightly higher than historical norms, and that this increase will likely not be sustained in future years. Specifically, this year we admitted 28 students into our Biostatistics Masters programs, as compared to our usual cohort of 18-20 students. This was due to elevated targets requested by the School to help with covering costs of launching new degree programs. We anticipate that our Biostatistics SM enrollment will return to 18-20 students next year, which will lead to an overall increase in the total number of students enrolled in our SM programs of approximately 5-12 as compared to this year’s numbers. Overall, we have always had a robust pool of applicants for our existing Masters programs in Biostatistics, and we anticipate the demand for this proposed program will be at least as much, and will probably exceed, these existing programs, as we expect this program will attract considerably more computer science students to apply. The following table presents the number of applicants, the number of accepted applicants, and the number of
enrollees in our existing Masters programs in Biostatistics for the period 2013-2016. Out of the 168 applicants we received this Winter, we estimate that between 50-60 of those were competitive applications that meet the requirements for acceptance into these programs.

| Biostatistics MS Programs: Admissions Data for Classes Entering Fall of 2013 - 2016 |
|---------------------------------|--------|--------|--------|--------|
|                                 | 2013   | 2014   | 2015   | 2016   |
| Applications                    | 131    | 127    | 130    | 168    |
| Offers                          | 33     | 42     | 51     | 33     |
| Matriculants                    | 20     | 22     | 31     | TBD    |

**Impact at the School:**

- The program will require the development of several new courses in the areas of health data science and machine learning in the Department of Biostatistics. A computing course in the will result from a revamping of existing computing courses (BIO508, BIO514), so that at least one of these courses will not continue.

- The expanded course offerings in data science can be covered by existing faculty as well as new faculty hires. The Department has received approval from the Dean’s Office to search for a Lecturer/Sr. Lecturer whose primary responsibility will be to serve as the Executive Director of this program (Professor Rafa Irizarry will serve as scientific director). The primary responsibility of this new faculty member will be running the day-to-day operations of this program, teaching at least two courses a year in the program, and enhancing collaborations between our department, HMS and the Harvard affiliated hospitals on big data projects. We are also opening searches this year for three junior faculty positions. The targeted research areas for these three positions are big data, statistical genetics and genomics, and computational biology, all of which are directly relevant to the new initiative in data science. Therefore, in total, we are initiating searches for four new faculty members, all of which will teach and mentor in the program and one of which focuses on the proposed Masters program.

- The Fall 2015 enrollments and room capacities for the two existing classes that are required are as follows:
  - BIO 222 - 60 students (Kresge 200 - capacity 64)
  - EPI 201 - 209 students (offering 2 sections this year for the first time: Kresge G1 - capacity 165 / FXB G12 - capacity 77).

Therefore, there is capacity in EPI 201 to accommodate 5-12 additional SM students. The other class, Bio 222, will be just slightly over capacity if the current enrollment stays constant. However, in addition to the expected decrease in the Biostatistics SM students, we also anticipate enrollment will likely decrease further because fewer doctoral students from other Departments will take this class due to the fact that the new PhD students will be required to take the new common PHS quantitative methods core course and will also likely take a more specialized intermediate biostatistics course (longitudinal data, survival analysis, gene mapping, data science, etc). So we should be able to accommodate 5-12 additional students in Bio 222. If not, then we will split this course into two sections.

- Any new course to be offered would also be open to students from other Departments at the School as well as the broader University community. Therefore it would allow
additional opportunities for students in FAS, the Medical School, or the T. H. Chan School to receive training in health data science. We anticipate these courses will be similar to those in our 220’s series, such as applied longitudinal and applied survival analysis, which are very popular across the School of Public Health and the Medical School.

• **Difference Between the Department’s Existing Biostatistics SM Programs and the Proposed Program:** Data science is the science of extracting knowledge from big data. It represents an interface between statistics, computer science and software engineering. Data scientists need skills including the ability to define questions in the face of massive data, identify and link big data from different sources, implement data wrangling and visualization to gain insights from such data, apply techniques from statistics and computer science, such as machine learning and theoretical understanding of algorithms, to analyze the data, and be able to implement high performance scientific computing to engineer software for reproducible analyses of such datasets. Therefore, knowledge of statistical methods for data analysis is only one component of data science. In particular, a top training program in data science requires significant training in computer science and software engineering. We have developed a program that requires basic programming training through EdX, at least 10 credits of advanced computer science training (in particular through our Computing for Big Data course with the possibility of more electives in this area), and a culminating research experience that will test training in statistical, machine learning, computer science, and software engineering by requiring students to apply it to current data science problems in population health sciences. Accordingly, this program is distinct from our existing Masters programs in Biostatistics, and it is not possible to construct a strong training program in data science within the current Masters programs.

• **Difference Between the School’s Existing Master’s Program in Computational Biology and Quantitative Genomics:** The current Master’s program in Computational Biology and Quantitative Genomics (CBQG) attracts students with different career histories and different career paths than does the proposed program. This program recruits students with prior training in biology, including molecular biology and genetics. The CBQG curriculum has a heavy emphasis on genetic and genomic courses and requires students to do a research practicum related to genetics and genomics as part of the degree. Training in the CBQG Program can include training in wet-lab techniques. In contrast, the proposed program in Health Data Science will focus on training in the intersection of statistical methods, statistical inference, and computer science, with applications in health science. Our biostatistics Master’s programs attract students with strong quantitative backgrounds. While this new program will continue to attract students with strong quantitative backgrounds, it is likely to attract students who have stronger training in computer science and with interest but perhaps little experience in statistics and biomedical research. The HDS MS program also places more emphasis on computing training and analysis of large health science data compared to the biostatistics MS program. Therefore, we believe that the students interested in this program will likely have career goals that are complementary to the existing Master’s programs and will not adversely affect enrollment in them.

• The Department of Biomedical Informatics (BMI) of Harvard Medical School is converting their two year master program of biomedical informatics to a one-year master program. Our program has a more emphasis on statistics and computing in population
science and biological science, while their program has more emphasis on informatics in clinical and biological science. The two programs are complimentary to each other and will coordinate. Our students can take courses offered by BMI as electives, e.g., the Electronic Medical Records course, and vice versa.

- The Departments of Computer Science and Statistics are developing a new master program in data science. Our program has a primary focus on health science, while their program is more general and has a primary focus on non-health science.

NEW DESCRIPTION FOR HARVARD CHAN SCHOOL CATALOG

Master of Science in Health Data Science (60-credit program)
The master’s degree program in Health Data Science will train students to extract knowledge from data and to communicate or share this knowledge. The first year will consist of case-based training in the areas of statistical inference, machine learning, and computing, as well as training in public health and biomedical sciences. Through this case-based approach, students will simultaneously learn computing skills necessary to manage and analyze massive health science data and will start gaining experience in answering scientific questions with data. These skills will be further developed during an intensive semester-long course during the third semester focusing on project-based work in health data science. Typically of 16 months duration, this program will focus on the interface of statistics and computational science, including machine learning, in the biological, population, clinical health sciences.

Further, although not a formal requirement for the degree, students will be encouraged to engage in research training in data science in the Harvard Chan School, the Medical School, or one of the affiliated hospitals, or an external research internship. Faculty mentors in the program (see Faculty Members Associated with the Program) will inform students of available opportunities in the form of a compiled list of analysis groups, labs and external internships that will engage students in a research experience in data science. This research experience will typically occur during the summer between the second and third semesters of the program, but could also occur during an academic year as well.

Applicants to the MS in Health Data Science program should have successfully completed calculus through partial differentiation and multivariable integration, one semester of linear algebra or matrix methods, and either a two-semester sequence in probability and statistics or a two-semester sequence in applied statistics, or computer science or related quantitative disciplines. Students should have practical knowledge of computer scripting and programming as well as some experience with a statistical computing package such as R or Python. Applicants should also show excellence in written and spoken English. Evidence that these requirements have been fulfilled should form part of the application. Additional research or work experience would be considered beneficial, but not required. Students who enroll in this program will come with undergraduate degrees in the mathematical sciences or from allied fields (statistics, economics, etc.) or computer science. Students are expected to have the goal of working as a data scientist in teaching hospitals, universities, research organizations, the pharmaceutical and biotechnology industries, or government.

(Minor modifications would be made in the rest of text as necessary.)
DEGREE REQUIREMENTS

Course Requirements:
A total of 60 credits of coursework are required for the MS in Health Data Science. This includes a 25 credit ordinally graded core curriculum consisting of:

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 Regression and Machine Learning (5 credits)
5. EPI 201      Introduction to Epidemiology Methods: 1 (2.5 credits)
6. BIO XXX  Computing for Big Data (5 credits)

Course 3 currently exists within the department, and will be updated to incorporate material in data science based on the continued treatment of cases originally introduced in courses 1 and 2. Initial offerings of Courses 1 and 2 will occur in the 2016 Spring and Fall semesters, respectively. Our department currently offers courses in computational science, including BIO 514 Data Structures and Algorithms and BIO 508 Genomic Data Manipulation. The new Computing for Big Data course (6) will be developed leveraging the material presented in these courses but broadening its focus to particular computing challenges in big health science data.

Students with prior equivalent background to any of these courses or strong reasons to take a different course can request permission from the director of the program for a substitution of one or more of these required courses.

The program is designed to produce strong programmers. Students will also be required to take an additional 5 credits of coursework in computer science. This feature distinguishes this degree from others in the department. The chosen course could be one offered here at the School or other departments in the University. Students much choose 5 credits from the following:

- BIO 505    Database Design and Use for Health Research
- BIO 514    Data Structures
- BMI 713    Computational Statistics for Biomedical Science
- CS 105    Privacy and Technology
- CS 124    Data Structures and Algorithms
- CS 164    Software Engineering Computer Science
- CS 165    Data Systems
- CS 171    Visualization
- CS 181    Machine Learning
- CS 187    Computational Linguistics
- STAT 183  Learning from Big Data
- STAT 171 Introduction to Stochastic Processes

The proposed master’s program will provide a culminating research experience that tests all competencies through a hands-on semester-long project-based research course (7.5 credits). This course will allow students to immerse themselves in multiple health data science projects in public health and biomedical science.
BIO XXX Health Data Science Practice (7.5 credits)

The course will be co-taught by two faculty members with complementary expertise. The projects, which will include investigations from areas such as statistical genetics and genomics, computational biology, comparative effectiveness research, analysis of electronic medical records, remote sensing environmental data, and network analysis, among others, will be developed and organized by the instructors, based on current data science projects being directed by faculty in the department. The overall objective of this course is to help the student integrate and apply data science methods presented in courses 1-6 above to real world, different massive datasets. For each project, the data are likely to be messy and students need to develop their analytic methods appropriately. With faculty input, the students will write a paper that is of the same standard as publishable papers in health science journals and present the results in class. We plan to have two instructors with complementary expertise (one statistical and one in computer science) and the faculty instructors (mentors) grade the project papers and also give feedback on how to improve presentations. Students then incorporate this feedback on paper writing and presentation into the next project, and so on.

This process will provide a comprehensive way to test competencies (listed at the end of this proposal) and teach students to conduct research in health data science. Specifically, the student will learn to be aware of problems that arise in study design, data collection, data wrangling and visualization, inference and computation, and develop practical problem solving skills. The student will learn to communicate through presentation of oral and written reports, following the standard of health science journal articles for each project completed in the class, and through student and faculty critiques of these reports at both the intermediate and final stages of projects. In summary, the course represents a research-based class experience, and will allow students to generate novel findings from current big data in health.

Students will also have the opportunity to take as electives other existing classes in the Department focusing on topics in big data, computation, or statistical inference, including courses in the existing Master of Science program in Computational Biology and Quantitative Genomics (CBQG), statistical genetics, Bayesian data analysis, applied longitudinal analysis, applied survival analysis, courses from Harvard Medical School, such as Analysis of Electronic Medical Records and clinical informatics, as well as courses from the Departments of Statistics and Computer Science. Students will also select public health and biomedical courses to enhance their knowledge in health science.

A minimum of 22.5 additional credits will come from the following list of elective courses offered by the departments of biostatistics, biomedical informatics, computer science, statistics, and substantive fields in public health, such as epidemiology, health policy, nutrition, environmental health, global health, and public health genomics. In addition to the computer science courses listed under the computing requirement, which could also be counted as electives once the 5 credit requirement has been met, students could also choose from a list that includes but is not limited to:

BIO 223 Applied Survival Analysis
BIO 226 Applied Longitudinal Analysis
BIO 257 Advanced Statistical Genetics
BIO 512 Introduction to Computational Biology and Bioinformatics
BIO 513  Advanced Computational Biology and Bioinformatics  
BIO 521  Introduction to Social and Biological Networks  
EPI 202  Elements of Epidemiologic Research: Methods 2  
EPI 288  Data Mining and Prediction  
EPI 289  Causal Inference  
EPI 515  Measurement Error and Misclassification  
ID 271  Advanced Regression for Environmental Epidemiology  
BMI XXX  Data Visualization  
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  
ME 530  Clinical Informatics  

Other courses may also be acceptable to satisfy these 22.5 additional credits. A distinguishing feature of this program is students may consider substantive classes in public health, such as in infectious epidemiology, failing health systems, or environmental exposure assessment, among others. Other courses of interest will include courses made available in the planned FAS Masters Program in Data Science, such as the critical thinking course in data science that will include training in law, ethics, privacy, and public policy issues associated with research using big data. Students are advised to consult with the Director of Master’s Studies to check prior to enrolling in the courses in question.

**Research Ethics:**  
Students must satisfy a research ethics requirement through attendance at a lecture series or satisfactory completion of a web-based training program. Students who feel they have already completed an equivalent training program must submit adequate documentation to, and receive approval from, the Director of Master’s Studies during the first semester in residence.

**SAMPLE CURRICULUM**

With 22.5 electives, the program is designed to offer a curriculum that is flexible in that it can be tailored to a student’s interests. Some students may design a training plan that emphasizes computer science and software engineering, while another may focus on a particular substantive field in public health, such as infectious epidemiology. Below is a sample curriculum for a student interested in applied statistical analysis and computer science.

Students will start the program in the Fall, with some self-guided training in computing available in the Summer before matriculation.

**Summer**  
R/Python Programming Bootcamp/EdX-on line course

**Fall Semester, first year:**  
BIO 260  Introduction to Data Science  (5 credits)  
EPI 201  Introduction to Epidemiology Methods: 1 (2.5 credits, Fall 1)  
BIO XXX  Applied Linear Regression and Machine Learning  (5 Credits)
BIO XXX       Computing for Big Data (5 Credits)
Other Electives (2.5 credits)

Winter Session first year:
EPI 288   Data Mining and Prediction (2.5 credits)

Spring Semester first year:
BIO 261   Health Data Science 2 (2.5 credits, Spring 1)
BIO 222   Basics of Statistical Inference (5 credits)
BIO 223
BIO 226   Applied Longitudinal Analysis (5 credits)
CS 124    Data Structures and Algorithms (5 credits)
Other Electives (2.5 credits)

Summer: Research Experience in Health Data Science

Fall Semester, second year:
BIO XXX   Health Data Science Practice (7.5 credits)
BIO 249   Bayesian Data Analysis (5 credits)
BMI 713    Computational Statistics for Biomedical Science (5 credits)

FINANCIAL MODEL

The Harvard T. H. Chan School’s Office of Financial Aid (OFA) develops a Cost of Attendance (COA) estimate each year to assist students with understanding and budgeting for the costs of their education. The OFA also works closely with students to help them finance the cost of their education with loans, as needed, and to identify potential scholarship opportunities.

The cost of attendance for the proposed program estimated here is based on the 2015-2016 COA for 60-credit master’s programs calculated by OFA.

<table>
<thead>
<tr>
<th>Expenses</th>
<th>13 months (September – May; September – December)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuition (60 credits x $1040)</td>
<td>$62,400</td>
</tr>
<tr>
<td>University Health Service Fee (1.5 years)</td>
<td>$ 3,585</td>
</tr>
<tr>
<td>Blue Cross Blue Shield Insurance Fee (1.5 years – may be waived)</td>
<td>$ 1,563</td>
</tr>
<tr>
<td>Books &amp; Supplies ($693/term)</td>
<td>$ 2,079</td>
</tr>
<tr>
<td>Rent/Utilities ($1320/month)</td>
<td>$17,160</td>
</tr>
<tr>
<td>Food ($350/month)</td>
<td>$4,550</td>
</tr>
<tr>
<td>Limited Personal Expenses ($400/month)</td>
<td>$5,200</td>
</tr>
<tr>
<td>Local Transport ($90/month)</td>
<td>$1,170</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$97,707</td>
</tr>
</tbody>
</table>

These are the same rates per credit hour for our existing Masters programs in Biostatistics and in Computational Biology and Quantitative Genomics, and the demand for these programs is
strong. For instance, this past admissions season, we received over 180 applications for a targeted enrollment of 18 students.

**PROCESS FOR DOCUMENTING COMPLETION OF THE DEGREE**

The Department of Biostatistics will develop a Degree Program Form for this new program. The purpose of this form is to permit the Director of Master’s Studies to verify that all degree requirements are being met. Details of the degree requirements will be given in the Graduate Student Handbook of the Department. Students are responsible for the requirements at the time they enter the program. Any change from Departmental requirements must be approved by the Director of Master’s Studies, including course waivers or substitutions.

This Degree Program Form will include core requirements, the epidemiology requirement, the research ethics requirement, and electives. It will be filled out by the student in September, reviewed and signed by the student’s academic advisor, and then final approval must be obtained from the Director of Master’s Studies. The Department will also follow student course selections and grades as students move through the program.

**DEGREE TO BE AWARDED**

Master’s of Science in Health Data Science. We wish to obtain the appropriate approvals for this new degree. In the interim (pending approval for this new degree name), we would offer the Master’s of Science in Biostatistics, with the area of interest being Health Data Science.

**DESCRIPTION OF LIKELY CAREER PATHS FOR GRADUATES**

We anticipate that graduates will take positions as a data scientist in teaching hospitals, universities, research organizations, the pharmaceutical and biotechnology industries, or government. They will be working on cutting edge research projects at the interface of biostatistics, epidemiology, and computer science. After a few years, our graduates would be expected to expand to project leadership and management and in the supervision of junior data science staff and other data analysts. Some graduates may decide to pursue a doctoral degree at some point in time after completing the SM.

**FACULTY MEMBERS ASSOCIATED WITH THE PROGRAM**

Rebecca Betensky (BIO)
Tianxi Cai (BIO)
Brent Coull (BIO)
Victor DeGruttola (BIO)
Francesca Dominici (BIO)
Sebastien Haneuse (BIO)
David Harrington (BIO)
Michael Hughes (BIO)
Curtis Huttenhower (BIO)
Rafael Irazarry (BIO)
Christoph Lange (BIO)
Liming Liang (EPI/BIO)
Other Biostatistics and Epidemiology faculty members will be involved with academic advising, course instruction, and mentoring students in this program.

COMPETENCIES

The program is designed to provide participants with targeted skills and knowledge required for work in health data science. These specific skills and knowledge domains are:

1. Apply experimental design in a research project.
2. Implement data wrangling by manually mapping data from one form to another.
3. Be able to perform data visualization/communication
4. Critically analyze and interpret data science.
5. Appropriately apply statistical inference/probability to make scientific conclusions from data.
6. Understand and employ linear models, regression and matrix algebra
7. Apply methods for high-dimensional data
8. Apply and interpret practical machine learning
9. Be able to engineer software, including reproducible research
10. Be proficient with high performance scientific computing
11. Team work / collaborative science
12. Be able to rigorously apply data science principles to address problems in the health sciences

Students will be exposed to computing skills in the programming languages R and Python throughout these courses.