Proposal for Undergraduate Certificate in Large Data Analysis

To: Helena Dettmer, Associate Dean for Undergraduate Programs and Curriculum
From: Suely Oliveira (Computer Science), Kate Cowles (Statistics), Bruce Ayati (Mathematics), Isabel Darcy (Mathematics), David Stewart (Mathematics), Alberto Segre (Computer Science), Luke Tierney (Statistics), Dan Anderson (Mathematics), Oguz Durumeric (Mathematics).

The above faculty from the departments of Computer Science, Mathematics, and Statistics and Actuarial Science propose a new undergraduate certificate in Large Data Analysis.

Rationale

Recent years have witnessed an explosion of massive data in many areas of human endeavor – from satellites delivering billions of pixels in daily images of the earth's surface, to customer data collected by Amazon and Google, to weather and climate. The need for people with the quantitative and computational skills to make sense of such data is urgent and the job opportunities abundant, as indicated in such mainstream media pieces as “The Age of Big Data” by Steve Lohr, which appeared in The New York Times in 2012 [http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all].

The National Science Foundation initiative called Expeditions in Training, Research, and Education for Mathematics and Statistics through Quantitative Explorations of Data (EXTREEMS-QED) aims to address this need by funding interdisciplinary programs to prepare undergraduate students in Mathematics and Statistics for work in big data. NSF recognizes that technical expertise in big data analysis and visualization requires integrating skills and knowledge from multiple disciplines, especially Statistics, Mathematics, and Computer Science.

The certificate proposal presented here arose from discussions of a University of Iowa proposal to EXTREEMS-QED which has now been accepted. For strong majors in any of the three departments, the certificate will open doors to careers and further graduate study. Additional benefits to the College and university are promotion of interdisciplinary communication and collaboration among both students and faculty in the quantitative sciences.

It should be noted that expertise in Large Data Analysis necessarily involves skills overlapping the three Departments: statistical analysis is needed to understand if correlations seen in large data sets are significant; computational and algorithmic skills to efficiently process large data sets; and mathematical skills are needed to develop and understand the underlying algorithms for the data analysis. Few students would recognize the relevant courses without guidance arising from having a certificate; nor would students choosing such courses be rewarded with the recognition that they deserve. The proposed Certificate gives both guidance and recognition of these skills in Large Data Analysis.

Description of the certificate

The certificate would require 21 credits from the courses listed below depending on the area of interest. Course descriptions are in the appendix.
Note that no more than 6 s.h. of course work for a major will be counted towards the Certificate.

Prerequisites: (not counted as certificate courses)

MATH:1850 or equivalent  Calculus I
MATH:1860 or equivalent  Calculus II
CS:1210 or equivalent  Computer Science I: Fundamentals
MATH:2700  Introduction to Linear Algebra

Level I (6 s.h.)
STAT:2010  Statistical Methods and Computing
STAT:3200  Applied Linear Regression

Level II (9 s.h.)
MATH:3800  Elementary Numerical Analysis
Two of the following:
MSCI:3200  Database Management
CS:4700  High Performance Computing
CS:4980  Topics in Computer Science (Selected topics only)
STAT:5810  Research Data Management
or STAT:5400  or Computing in Statistics

Level III (6 s.h.)
MSCI:6421/CS:6421  Knowledge Discovery
MATH/STAT/CS:4740  Large Data Analysis (capstone course)

Calculus I and II are prerequisites for several of the listed courses, but they are required for the majors in all three departments so cause no additional burden. Computer Science I (CS:1210) is required by both Computer Science and Statistics, and many Mathematics majors take it, and so does not add much to the overall student burden. Introduction to Linear Algebra (MATH:2700) is required for Mathematics and Statistics, and is a common elective for Computer Science. The certificate coursework does not duplicate any existing major or minor, but includes courses from all three departments' majors as well as additional coursework specific to data management and big data analysis.

The capstone course is the only new courses being developed for the certificate. This will give students experience working in teams to apply the methods they have learned in earlier certificate courses to real datasets and communicating their results to others.

Affiliated faculty
Bruce Ayati (Mathematics)
Kate Cowles (Statistics)
Isabel Darcy (Mathematics)
Oguz Durumeric (Mathematics)
Suely Oliveira (Computer Science)
Alberto Segre (Computer Science)
Resources

With the exception of the new capstone course, all certificate coursework is drawn from existing courses that are offered at least every other year. The new courses will be team-taught by faculty from the three departments and thus will not overburden any single faculty member. The only resource issue is that some of the courses (22M:072, 22S:201, 06K:182) already are at capacity so additional sections will be needed to accommodate certificate students.

Starting date and Five-year enrollment projection

Ideally we would like to begin the certificate program in Fall of 2014. However, we understand if implementation must be delayed until Fall of 2015 because of the approval process for certificates.

As the certificate becomes known, we expect it to attract up to 5% of majors in Computer Science and Math and up to 20% of majors in Statistics, plus a few students from the College of Engineering. Our projections are as follows.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Mathematics</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Statistics</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>21</td>
<td>33</td>
<td>42</td>
<td>54</td>
</tr>
</tbody>
</table>

Outcomes Assessment

We will administer a voluntary anonymous online survey to the students enrolled in the certificate at the end of each spring semester. In addition, we will track their performance in all courses required for the certificate and their time to graduation. For students who receive the certificate and graduate, we will send annual mail surveys to track their career or graduate education status.

Plans for Disseminating Information to the Campus Community

We will email information to the Registrar, the Advising Center, and advisors in departments in the College of Engineering. The DEOs and/or Directors of Undergraduate Studies from Computer Science and Mathematics are among the faculty supporting the certificate.
Appendix: Course descriptions

22C:016 (CS:1210) Computer Science I: Fundamentals
General Catalog:

Introduction to programming using Python; programming constructs, data types, problem-solving strategies, data structures, object-oriented programming.

Description:
This is the introduction-to-programming course in the computer science major and minor curricula. Prior programming experience is not required, although some students will have had some previous exposure to programming. It emphasizes object-oriented programming style and methodology. The lecture is taught three times a week. The 50-minute discussion periods, led by a TA, are used to discuss programming exercises, hold quizzes, and to answer questions in a small classroom atmosphere. Concepts are presented in the context of working examples and exercises. Language syntax and computing paradigms are studied. Programming projects are used to reinforce key programming notions, including iteration, data types, functions, and objects. Projects may include graphics, string processing, and network applications. Lectures are taught by a faculty member; discussion sections are led by TAs.

22M:027 (MATH:2700) Introduction to Linear Algebra
General Catalog:

Vector algebra and geometry of three-dimensional Euclidean space and extensions to n-space and vector spaces; lines and planes, matrices, linear transformations, systems of linear equations, reduction to row echelon form, dimension, rank, determinants, eigenvalues and eigenvectors.

Description:
This course introduces students to vectors, linear transformations, and matrices with applications to the geometry of Euclidian three-space and generalizations. The course usually begins with a careful study of the solution of linear systems of equations and ends with the orthogonal diagonalization of symmetric matrices. Topics include row reduction, determinants, bases, dimension, rank, nullity, eigenvalues, and eigenvectors. The course meets four times weekly with a faculty member in small sections. Only students who did very well in Calculus I should attempt this course before completing a full year of calculus.

22S:30 (STAT:2210) Statistical Methods and Computing
General Catalog:

Methods of data description and analysis using SAS; descriptive statistics, graphical presentation, estimation, hypothesis testing, sample size, power; emphasis on learning statistical methods and concepts through hands-on experience with real data.
22S:152 (STAT:3200) **Applied Linear Regression**

General Catalog:

Regression analysis with focus on applications; model formulation, checking, selection; interpretation and presentation of analysis results; simple and multiple linear regression; logistic regression; ANOVA; hands-on data analysis with computer software.

22S:166 (STAT:5400) **Computing in Statistics**

General Catalog:

R; database management; graphical techniques; importing graphics into word-processing documents (e.g., LaTeX); creating reports in LaTeX; SAS; simulation methods (Monte Carlo studies, bootstrap, etc.).

22S:201 (STAT:5810) **Research Data Management**

General Catalog:

Overview of problems encountered in gathering and processing data from biomedical investigations; introduction to data management techniques useful in biomedical studies; introduction to Microsoft Access.

22M:072 (CS:3700) **Elementary Numerical Analysis**

General Catalog:

Computer arithmetic, root finding, polynomial approximation, numerical integration, systems of linear equations, ordinary differential equations; use of higher-level computer language such as Matlab, Maple, Mathematica.

**Description:**
This course examines numerical methods for finding solutions of nonlinear equations, polynomial interpolation and approximation of functions, numerical integration, solution of simultaneous linear algebraic equations, and solution of ordinary differential equations. Although not a required prerequisite, students would benefit from some background in linear algebra; 22M:027 or 22M:040 would be sufficient. Problem assignments require use of a computer, and this is generally done in Matlab, a system and language taught in the course.

06K:182 (MSCI:3200) **Applications Database Management Systems**

General Catalog:

Design and implementation of a database using relational DBMS; emphasis on issues of logical and physical design, database administration, concurrency control, maintenance.
Knowledge Discovery

Knowledge discovery process, including data reduction, cleansing, transformation; advanced modeling techniques from classification, prediction, clustering, association; evaluation and integration.

New course

CS/MATH/STAT:4740 Large Data Analysis (Capstone)

We will create a capstone course for the certificate program in Large Data Analysis and Visualization. All students must take this course to complete the certificate. Since it will be taken by a large number of students with different emphases, the assumed knowledge is covered by the common courses in all paths through to the certificate. This capstone course will include a project component, and this project may emphasize either data analysis or simulation, or even a fusion of the two. The course will be team-taught with instructors from the three Departments of Computer Science (Suely Oliveira), Mathematics (David Stewart), and Statistics (Kate Cowles).

Here is a brief description of a syllabus for a capstone course for the “big data” certificate.

1. Uses/Applications
   - Google PageRank
   - seismology
   - Netflix problem
   - weather forecasting & data/simulation fusion

2. Computers
   - high performance computing
   - CPU architecture & memory hierarchies
   - parallel computing: shared memory vs message passing
   - programming for parallel computers
   - MapReduce/Scan algorithms

3. Techniques
   - matrix operations
   - Fast Fourier Transforms
   - discrete and continuous simulation techniques
   - dimension reduction techniques
   - principal components analysis
   - factor analysis
• cluster analysis
• multidimensional visualization (e.g., using R+googleVis, Matlab)

4. Projects – intent is to base this on common collection of data sets
   • The intent here is to provide a small number of group projects; each project works on a given data set with the aim to extract useful information from it.