Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

Oregon State University: M.S. and Graduate Certificate in Data Analytics

1. Program Descriptions

a. CIP number will be provided by registrar

b. Overview of the proposed programs

The terms “big data” and “analytics” are now commonly used to describe the vast amounts of data currently being generated and analyzed in many areas of science, education, business and government. The analysis of the information in these data sets has already led to major advances in areas ranging from genomics to astronomy and high-energy physics and to development of new information-based industries. Until recently, methods of analysis of these data had largely assumed that traditional computing and inferential tools within the analyst’s own computing environment are adequate to summarize and make informative inferences from the data. With the increased capacity to collect, store and summarize data, and with the growth of the term “analytics,” it is clear that there is a demand for more training in data management and data analytic skills.

A widely circulated report by McKinsey Global Institute estimates that by 2018, the US could face a shortage of 140,000 to 180,000 people with deep analytic skills, and those 1.5 million managers and analysts will be needed with the know-how to use the analysis of available data to make effective decisions. The “deep analytic skills” referred to above are founded primarily in the disciplines of Statistics and Computer Science, and the proposed programs draw heavily on these two disciplines, with the objective of producing trained professionals with these skills. The core courses of the M.S. program come from statistics and computer science, but the program also recognizes that some of the demand for big data skills comes from biology and the health professions; the program offers a concentration in this area. The Graduate Certificate in Data Analytics will contain courses from Statistics only.

c. Course of study - proposed curriculum, including course numbers, titles, and credit hours

The proposed M.S. degree requires a total of 45 credit hours; concentrations will be offered in Health Analytics and Statistics, but a concentration is not required. To be admitted to the program, a student must meet the minimum requirements for admission to the Graduate School and must have successfully completed an undergraduate statistics course at the level of ST351. Mathematics to the level of calculus is recommended but not required. The curriculum includes (courses are 3 credits unless otherwise indicated):

- 7 core courses in statistics, for a total of 24 credit hours.
- 3 core courses in computer science, for a total of 12 credit hours.
- Concentrations:
  - **Concentration in Health Analytics**: 3 approved courses in analytics for health/biology/genetics data, for a total of 9 credit hours;
Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

Concentration in Statistics: 3 approved courses in Statistics, for a total of 9 credit hours;
Degree without concentration: 3 approved courses in Health Analytics or Statistics, for a total of 9 credit hours.

(A future track in Business is a possibility, subject to the availability of suitable online course offerings in the College of Business. Currently there are four “background business” courses offered online in COB (BA 513, 514, 515, 516). These courses, available to those with B. S. degrees, could prepare students for more advanced business offerings in data analytics, when available.)

The courses comprising the M.S. in Data Analytics would be available to students not enrolled in the degree program, subject to completion of appropriate prerequisites and space availability.

Core courses in statistics:

ST5XX Foundations of Data Analytics. 4 credits. Probability distributions; basic principles of estimation and hypothesis testing; methods for one- and two-sample problems; introduction to simulation. Computer lab using R. Prerequisite: ST351 or equivalent.

ST5XX Data Analytics I. 4 credits. Linear regression – model fitting, evaluation, and interpretation; prediction; missing data and data imputation; mixed models; scaling up to large datasets. Computer lab using R. Prerequisite: ST5XX.

ST5XX Data Analytics II. 4 credits. Methods for categorical data – tables of counts, logistic regression, log-linear regression, prediction, mixed models; scaling up to large datasets. Computer lab using R. Prerequisite: ST5XX.

ST5XX Time Series Analytics. 3 credits. Methods for analyzing and forecasting time series data; serial correlation; smoothing; Autoregressive moving average models; spectral analysis; State-space models; Kalman Filter. Prerequisite: ST5XX.

ST5XX Multivariate Analytics. 3 credits. Basics of matrix algebra, principal components analysis, cluster analysis, factor analysis, multidimensional scaling. Prerequisite: ST5XX.

ST5XX Data Mining. 3 credits. Methods for exploring and describing features of, and patterns in large, complex datasets. Recursive splitting, neural networks, discriminant analysis. SAS Enterprise Miner will be used.

ST5XX Capstone project. 3 credits. Under the direction of an advisor, a team of students will tackle a complex data problem to address an important question of interest, culminating in a written report and oral presentation.
Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

These courses are all under development, but large pieces of them will be adapted to the big data context from existing courses in the Statistics Department catalog. Most of the ST5XX material is in current versions of ST521 and ST511; most of ST5XX and ST 518 is in the current versions of ST512 and ST513; and Time Series Analytics and Multivariate Analytics will be adapted from existing courses in Time Series (ST565) and Multivariate Analysis (ST557).

Core courses in computer science:

**CS5XX Programming Concepts for Non-majors. 4 credits.** Topics include: variables, control structures, and functions; Big-O notation and concept; linear data structures (arrays, vectors, lists, stacks, queues, hash tables); binary search trees; heaps; searching and sorting. Computer lab is included.

**CS4XX/5XX Big Data Management. 4 credits.** Topics include: accessing and distributing data online; non-relational databases; map reduction; multi-node data processing; load balancing; types of data-stores. Computer lab is included.

**CS5XX Applied Machine Learning. 4 credits.** Supervised learning: algorithms (decision trees, naive Bayes, support vector machines); ensemble methods (boosting, bagging, random forests); discriminative vs. generative methods; model selection and evaluation for supervised learning; feature selection. Unsupervised learning: algorithms (single-linkage clustering, k-means, Gaussian mixture models, spectral clustering; model selection and evaluation for unsupervised learning. Semi-supervised leaning: EM; graph-based methods; multi-view methods. Prerequisite: ST5XX or co-registration in ST5XX. Computer lab is included.

Approved courses in the Health Analytics track:

ST5XX Applied Survival Analysis, 3 credits.
ST5XX Statistical Methods for Genomic Research, 3 credits.
ST5XX Introduction to Quantitative Genomics, 3 credits.

Approved courses in the Statistics track:

ST515 Design and Analysis of Planned Experiments, 3 credits.
ST539 Survey Methods, 3 credits.
ST5XX Data Visualization, 3 credits.
Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

ST5XX Modern Analytical Methods for Large and Complex Datasets, 3 credits.

ST539 will be taught for the first time through Ecampus in Spring 2015. ST515 has been developed with Ecampus for on-line delivery. The Data Visualization course was taught in the Summer of 2014 on campus, and will be adapted for Ecampus. The course for large and complex datasets will be taught for the first time on campus in Spring of 2015, and thereafter developed for Ecampus delivery.

The Graduate Certificate in Data Analytics will include ST5XX, ST5XX and ST5XX, as well as ST5XX Time Series Analytics and ST5XX Multivariate Analytics for a total of 18 credits. The requirements for admission to the Certificate program will be the same as those to the M.S. program.

d. **Manner in which the programs will be delivered**

   All courses will be offered online, in cooperation with Ecampus. To give some context to this decision, we note that a 2012 Custom Research Brief on Graduate Degrees and Certificates in Analytics prepared by the Education Advisory Board examined 20 graduate programs in analytics. Eleven of these programs were established in the 2010-11 timeframe, seven began in 2012, and two were scheduled to begin in 2013. Of those programs, four offer both online and in-person formats, one program is hybrid and three programs are offered exclusively online. In our case, we expect that the audience for our programs will be drawn largely from pool of industry and government professionals (mostly outside Corvallis) in positions that require statistical or analytical skills; students from this population would likely continue to work while pursuing an advanced degree, making the online offering quite appealing. It is possible that the programs, especially the Certificate program, will also attract students from Oregon and beyond who are not working, but who might not want to incur the added expense of moving to and living in Corvallis. These factors suggest to us that an online format is more likely to be successful.

   OSU has a strong Ecampus presence with a marketing apparatus already in place, including mechanisms and resources to assist with course/program development. In addition, the Department of Statistics already has a fairly strong Ecampus presence and experience in online course development.

e. **Ways in which the programs will seek to assure quality, access, and diversity.**

   All applications will be reviewed by the Director of the M.S. in Data Analytics program. After a preliminary screening, additional admissions committee members will review applications. Applicants must meet the minimum requirements for admission to the Graduate School, and they must have at least one undergraduate course in Statistics, at the level of ST351. The applicants’ statement of intent and reference letters will be
evaluated to ensure that the applicants have the dedication required to complete a rigorous online program.

Given that the programs will be delivered exclusively online, we will work with Ecampus as we develop new courses to ensure ‘best practices’ and accessibility of all courses. We will also market the program broadly to help ensure diverse cohorts of students. Because of our existing M.S. program in Statistics, we are well connected with Universities around the world. We will make additional efforts to market the program in the private sector.

f. **Anticipated fall term headcount and FTE enrollment over each of the next five years.**

An initial cohort of approximately 10 students is anticipated for Fall 2015 in the M.S. program, when the program will commence. For the 2016-17 year we expect total enrollment of about 20 in the M.S. program, increasing to 30-40 by 2019-2020. Numbers in the Graduate Certificate program may be higher in the long run, though we anticipate smaller numbers in this program at the outset—5 students in Fall 2015, increasing to 15-20 by 2019-20.

g. **Expected degrees/certificates produced over the next five years.**

Beginning in 2017, we expect about 10 M.S. degrees and 5 Graduate Certificates per year, increasing to 20 and 15 per year, respectively, over a five-year period.

h. **Characteristics of students to be served**

As these are online programs, many if not most of the students are likely to be outside the Corvallis area, and some may be outside the state. We anticipate that most of our students will be professionals who are looking to add more data analytic tools to their workplace and/or who are seeking advancement or transition to a new functional area. We expect to have both full- and part-time students in the M.S. program; there is not a full-time option for the Graduate Certificate as the 516-17-18 sequence must be taken before the Time Series Analytics and Multivariate Analytics courses.

i. **Adequacy and quality of faculty delivering the programs**

All courses in the programs will be developed and taught by full-time OSU faculty members, recruited according to the same standards as all OSU faculty. Careful vetting will be carried out for any part-time faculty, though none are anticipated at this time. Faculty in the Statistics Department involved with developing and teaching these courses are:

<table>
<thead>
<tr>
<th>Faculty</th>
<th>Degree</th>
<th>Rank</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharmodeep Bhattacharyya</td>
<td>PhD Statistics, Univ of Cal,</td>
<td>Assistant Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
</tbody>
</table>
### Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Title</th>
<th>FTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yanming Di</td>
<td>Berkeley</td>
<td>Assistant Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Sarah Emerson</td>
<td>Berkeley</td>
<td>Assistant Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Alix Gitelman</td>
<td>Stanford University, Carnegie Mellon University</td>
<td>Associate Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Yuan Jiang</td>
<td>Berkeley</td>
<td>Assistant Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Virginia Lesser</td>
<td>University of North Carolina</td>
<td>Professor</td>
<td>12 month, 1.0 FTE</td>
</tr>
<tr>
<td>Debashis Mondal</td>
<td>Berkeley</td>
<td>Assistant Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Tom Sharpton</td>
<td>University of California</td>
<td>Assistant Professor</td>
<td>9 month, 0.3 FTE in Statistics</td>
</tr>
<tr>
<td>Charlotte Wickham</td>
<td>Berkeley</td>
<td>Assistant Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Lan Xue</td>
<td>California</td>
<td>Associate Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Justin Wolford</td>
<td>Oregon State University</td>
<td>Instructor</td>
<td>9 month, 1.0 FTE</td>
</tr>
<tr>
<td>Weng-Keen Wong</td>
<td>Carnegie Mellon University</td>
<td>Associate Professor</td>
<td>9 month, 1.0 FTE</td>
</tr>
</tbody>
</table>

### Faculty resources
Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

All faculty in Statistics are full-time faculty on 9- or 12-month appointments. Wolford in Computer Science is a full-time instructor, and Wong is a 9-month, full-time Associate Professor. The Department of Statistics will have one new faculty member in Fall 2015 who will handle some of the teaching needs for this program, and possibly the Directorship of these programs.

k. Other staff

Clerical support (0.3 FTE) will be needed to help administer the programs. GTA support will be needed as the programs grow.

l. Facilities, library and other resources

A small amount of funding will be needed for student recruitment, in partnership with Ecampus marketing. Existing staff and computing infrastructure will be able to handle the increased student load. The Department is in the process of investing in additional computing resources for its faculty and students.

m. Anticipated start date

September 2015.

2. Relationship to mission and goals

a. The OSU mission statement reads:

“As a land grant institution committed to teaching, research, and outreach and engagement, Oregon State University promotes economic, social, cultural and environmental progress for the people of Oregon, the nation, and the world. This mission is achieved by producing graduates competitive in the global economy, supporting a continuous search for new knowledge and solutions, and maintaining a rigorous focus on academic excellence…….”

Private companies, government agencies, and academic researchers all now deal with huge data streams on an everyday basis. Teaching and research aimed at improving the ability of these groups to analyze big data clearly contributes to economic, social, cultural, and environmental progress by promoting deeper, more accurate, and more effective utilization of the data essential to their functioning. The proposed programs, that will be unique in the state, will serve the needs of both public and private interests seeking to make more effective use of their data resources.

b. Connection to strategic priorities and signature areas of focus

One of three focus areas in OSU’s Strategic Plan is Promoting Economic Growth and Social Progress, and the goal includes “an expanding institutional culture of innovation
Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

and collaboration.” Making effective use of big data is essential not only to academic researchers and industry, but also to health providers, managers of ecological resources, transportation experts, and professionals in countless other fields. The proposed programs are cutting-edge programs in data analytics, and the multidisciplinary nature of the M.S. program in particular will help to advance OSU’s burgeoning culture of collaboration.

c. Contribution to OUS goals for access, etc.

Since there is no OUS, we will presume that the new structure involving the Higher Education Coordinating Commission will adopt goals for access similar to those of OUS. The OUS goals for access include educational opportunities for non-traditional students. We believe that these programs will appeal mainly to working professionals in positions that increasingly demand big data analytic skills. By offering the programs online, we will provide an opportunity for students not necessarily located in Corvallis, or even in Oregon, to deepen their data analytic skills. The design and execution of courses by OSU faculty members will ensure quality products; as this is a very new area of academic study, we anticipate many opportunities for knowledge creation and innovation as the programs develop.

d. Meeting statewide needs

The need for collecting and analyzing big data doesn’t know state boundaries; one could argue that in certain areas, such as environmental monitoring, Oregon’s needs are greater than those of most states. Yet to date there is no program in Oregon aimed at preparing data analytic professionals to cope with the big data onslaught. Areas such as health services, resource management, and emergency preparedness need to gather and analyze huge data sets; so do banks, insurance companies, and marketing specialists. The proposed online program would be a resource for the entire state and for the Northwest region.

3. Accreditation

At present there is no recognized accrediting body for programs of this sort. Data analytics programs are quite new; they come in many different varieties, and there is no consensus as yet regarding what these programs should look like.

4. Need

a. Evidence of market demand

Evidence from many sources indicates that the proposed program is very timely. The McKinsey study alluded to above attests to the incipient demand for trained professionals in data analytics. Another telling sign is the nationwide proliferation of graduate programs, primarily at the Master’s level, in ”data analytics”, “data science,”
Proposal: M.S. in Data Analytics and Graduate Certificate in Data Analytics

and programs with other titles aimed at preparing professionals to deal with huge data sets. Some of these programs have acceptance rates below 15%, and starting salary offers to graduates of the NC State Master’s program averaged over $96,000.

Our own experience provides direct evidence of interest in big data. In 2013, the Department of Statistics introduced a course, “Statistical Methods for Genomic Data”, which enrolled 32 students, many from outside the department. In 2014, the course “Statistical Computing and Big Data” was offered for the first time; the course was capped at 30, and 30 students were enrolled. In addition, recent conversations involving the OSU Departments of Statistics and Electrical Engineering and Computer Science and business leaders in the state have indicated a keen interest in a program of this sort.

Finally, it is relevant to mention a market analysis carried out in 2013 for a certificate program proposed by the Department of Statistics in Applied Statistics/Data Analytics. This analysis found that about 90% of students in similar certificate programs were currently employed, with many aiming to acquire the skills to transition into a new functional area. The leading employers of these students were professional, scientific, and technical services; educational services; insurance carriers; and hospitals. The analysis noted that “industry professionals with 5 or more years of experience in positions that require statistical or analytical skills and who seek advancement or transition into new functional areas enroll in applied statistics/analytics programs.” Significantly, students preferentially sought “data analytics” certificates; at Colorado State University, for example, there were twice as many inquiries regarding “data analytics” as for “applied statistics.” It seems reasonable that at least some of this analysis would be relevant for our proposed degree programs, with the audience for our degree being drawn in part from this same population.

b. Not applicable.

c. Manner in which the program would serve the need for improved educational attainment in the region and state.

One of the potentially great advantages of distance learning is that sparsely populated states (which Oregon essentially is) can participate electronically in new educational endeavors. By delivering this program through distance learning, opportunities would be enhanced for educational attainment throughout the state and region in an area rapidly becoming critical to our economic and scientific future.

To date there are no similar programs in Oregon (or Idaho, to our knowledge). The program in Data Science at the University of Washington in Seattle is run by their Department of Computer Science and appears to have more emphasis on computer science and less on statistical analysis than the programs we propose.

d. Manner in which the program would address the civic and cultural demands of citizenship
Meaningful participation in citizenship requires an informed citizenry, and with the amount of data now at our fingertips, effective ways of extracting information from data and summarizing data efficiently are critical to our civic and cultural development. The proposed program would contribute to the public’s ability to develop critical thinking skills and to approach data thoughtfully.

5. Outcomes and quality assessment

a. Expected learning outcomes of the program

Learning outcomes for each of the individual courses will be provided by the instructors of the course, who will have the primary responsibility to see that these outcomes are achieved. The overall goal of the Data Analytics programs is to equip quantitative professionals with the tools to gather, analyze and interpret data collected on scales up to terabytes \((10^{12}\) bytes\) or even petabytes \((10^{15}\) bytes\). Upon completion of the M.S. program, students will be able to:

(i) Describe and apply the basic principles of statistical inference and commonly used statistical models, machine-learning tools, and database tools.

(ii) Select appropriate statistical and computational methods and apply analytical skills to effectively summarize, visualize, and make valid inferences from data.

(iii) Fit and evaluate simple and multivariate linear regression models and generalized linear models.

(iv) Perform appropriate analyses on time-series data and multidimensional data.

(v) Describe and implement data-mining tools.

(vi) Show proficiency in R and database programming.

(vii) Communicate quantitative results to individuals who may not have expertise in either statistics or computer science.

(viii) Apply essential database skills such as storing, organizing and manipulating large amounts of data contained in databases.

(ix) Perform data cleaning and data imputation.

(x) Articulate principles of statistical and data ethics.
Learning outcomes for the Graduate Certificate program are items (i), (ii), (iii), (iv), (vii), and (x) from the list above.

b. Methods by which the learning outcomes will be assessed and used to improve curriculum and instruction

Learning outcomes will be assessed in course assignments, course exams, and the final capstone project. Exit surveys will be carried out with all students upon finishing the program.

c. Program performance indicators, including prospects for success of program graduates

Students will be surveyed at intervals after leaving the programs. Employment placement of students and their rate of compensation will be a key indicator of program quality. Employers will be surveyed periodically to ascertain their views of the programs and to aid in determining what skills should be emphasized (or not) in the curriculum.

d. Nature and level of research and/or scholarly work expected of program faculty

Nearly all program faculty will be full-time OSU faculty members, so the expectations will be the same as those for other OSU faculty. Therefore, publication of research in peer-reviewed journals will be the principal indicator of success.

6. Program Integration and Collaboration

To the best of our knowledge, there are no closely related programs in other Oregon universities or Oregon private institutions. As outlined above, the program represents a collaboration with the Department of EECS, which will provide course offerings in the program. Although the Department of EECS is developing a suite of courses specific to the M.S. program, we would not expect any substantial impact on their budget or enrollments, apart from the teaching resources devoted to these three courses.

7. Financial Sustainability

a. Business plan for the program that anticipates and provides for its long-term financial viability.

Proposed classes will be developed and taught by existing faculty. A number of these faculty members were recently hired in the past year and another search is underway this year. We anticipate revenue generated from tuition would provide support for the GTAs that may be needed if larger classes require the need for grading assistance. For the
initial program, current classified staff can assist in the program management. As the program grows, revenue generated from tuition would support additional staff needs if required. There is no anticipated need for library resources.

b. Plans for development and maintenance of unique resources necessary to offer a quality program in the field.

Since the program is online, no on-campus unique resources are needed. The department is in the process of investing departmental resources to increase computer capabilities for both on-campus and the anticipated students in this program.

c. Targeted student/faculty ratio.

Since the department recently hired three and soon four new faculty, we anticipate that there would be no negative changes to our current student/faculty ratio.

d. Resources to be devoted to student recruitment.

Since this in a new online program, we are partnering with Oregon State Ecampus in the development and delivery of the M.S. Data Analytics and Graduate Certificate in Data Analytics, marketing and student recruitment will be led by the Ecampus Marketing and Enrollment Services team. We are in the process of developing a Memorandum of Understanding with Oregon State Ecampus. This will include market research, marketing planning, marketing implementation and enrollment management. The department will provide funds from our foundation accounts to provide travel for faculty to attend key national meetings to advertise the program and recruit students.

8. External review. It is our understanding that this will occur after the proposal has made it through the OSU process. The Department had one visitor in fall 2014 who discussed the M.S. program in analytics developed at his University and provided input on our program.