Proposal for PhD Option
in “Advanced Data Science”

April 23, 2015

Overview

Summary Description

A fourth paradigm of science has recently emerged: scientific discovery is now driven in part by the analysis of large datasets. In order for the University of Washington to remain a leader in domain sciences and engineering, we need to educate our students in data science, which is the ability to extract knowledge from large datasets and use that knowledge to advance scientific discovery. Furthermore, our students should be leaders not only in the use of data science tools but in the development of such tools.

The proposed “Advanced Data Science” option aims to educate the next generation of thought leaders who will both build and apply new methods for data science. This option will help to educate and recognize PhD students whose thesis work focuses specifically on building and using advanced data science tools. ***The goal of this option is not to educate all students in the foundations of data science but rather to provide advanced education to the students who will push the state-of-the-art in data science methods in their domain***.

Six departments have come together to offer such an advanced data science option in their PhD programs:

1. Astronomy
2. Chemical Engineering
3. Computer Science & Engineering
4. Genome Sciences
5. Oceanography
6. Statistics

An important characteristic of this advanced data science option is that, independent of their home department, students will complete the same set of core data science courses. This shared core curriculum will ensure that students are not only knowledgeable in data science but that they also had the opportunity to interact with each other and form inter-disciplinary cohorts.

To complete the “Advanced Data Science” option, students will take three out of the following four courses:

- **Data Management**: CSE 544.

- **Machine Learning**, CSE 546 or STAT 535.

- **Data Visualization**: CSE 512.

- **Statistics**: STAT 509 or STAT 512-513.

Additionally, to further expand students' education and create a campus-wide community, students will register for **at least 4 quarters** in the weekly **eScience Community Seminar**.

Finally, all six departments already have approved "Big Data Tracks" in their PhD programs, which enable students to take the required courses for this new Advanced Data Science option and count the credits toward their PhD degrees.

Three departments have a small set of extra requirements for students in their department in order to complete the Big Data track and this Advanced Data Science Option. These departments are Computer Science & Engineering, Genome Sciences, and Statistics. We describe these extra requirements below and attach in appendix the full description of the Big Data Tracks in each department.

Rationale

The path to deep scientific discoveries is changing rapidly. Most disciplines, from physical to life sciences, have entered an era where discovery is no longer limited by the collection and processing of data, but by the management, analysis, and visualization of this information. Novel developments in instrumentation have lead to a tremendous increase in the magnitude of this data, forcing scientists to perform analyses on data that is too big for standard desktop computing tools, i.e., leading to a focus on big data. To harness the opportunities that big data brings, the next generation of scientists needs education both in a domain science and in methods for data management, analysis, and visualization. Students further need an education that teaches them how to build the next generation of data science tools in addition to the knowledge in the application of these tools.

To address the above growing needs, six departments have put in place "Big Data tracks". These tracks enable students in these departments to take advanced data science classes and have these classes count toward their PhD degrees. The goal is to recognize students on these big data tracks with a transcriptable option they can list on their CVs.

Administrative Location

The following participating departments will each offer this same "Advanced Data Science option". Each department will administer the option for its own students. That is, each department will have a new code established under its PhD program.

1. Astronomy
2. Chemical Engineering
3. Computer Science & Engineering
4. Genome Sciences
5. Oceanography
6. Statistics

Program Option Name

The program option name is “Advanced Data Science”

Related Programs

While we are putting in place this PhD program in data science, we are also working toward a professional master’s program in data science. The latter covers an overlapping set of departments (Computer Science & Engineering, Statistics, HCDE, Biostats, iSchool, and Applied Math). The two programs are complementary in nature. The PhD option described in this document will serve the needs of *full time PhD students*. The professional master’s program will serve the needs of people working full time in industry who want to acquire basic data science skills through an evening program, which is a terminal master’s degree. The detailed program for the master’s degree is currently being developed.

Timeline for Implementation

Since all participating departments already have a big data track in place, we are ready to implement the option as soon as we have approval. Ideally, we would like to start recognizing students in Spring 2015.

Relationship to Institutional Role, Mission, and Academic Unit Priorities:

The role of the six departments involved in the proposed “Advanced Data Science” option is to educate the next generation of scientists and thought leaders in disciplines that are both developers and consumers of new data science methods. To continue to be world leaders in their respective disciplines, students in those departments need to become leaders in data science. It is thus an utmost priority for the academic units involved to provide advanced education in data science to their graduate students. The proposed option will accomplish this goal.

Documentation of Need for Program

Our society's ability to generate data is growing at an unprecedented scale and rate. Researchers estimate the size of the Web to be over a trillion Webpages. The size of Twitter has already exceeded the Library of Congress. The next generation of telescopic sky surveys such as the Large Synoptic Survey Telescope (LSST) will generate 10s to 100s of petabytes a year of imagery and derived data. The Earth Microbiome Project expects to produce 2.4 petabases in their metagenomics effort.

As a result, the ability to analyze massive-scale datasets has emerged as a critical tool both in industry and in the sciences.

In the current graduate education, however, the ability to efficiently mine the numerous and diverse data sets is still lacking. The effect of this gap on science is difficult to underestimate. At the UW, the eScience Institute routinely engages with researchers who tell stories of the form: “In my dissertation, I analyzed N data points over five years. Last week, I collected 100N points.” At this rate, it is not just the algorithms that need to scale, it is also the skills and tools that researchers have access to and in which they have expertise. Graduate students in domain sciences must become proficient in computational thinking – they must internalize the basic strategies needed for problem solving using algorithms, large computers, and large data sets. To be successful, the next generation of scientists thus needs to be deep in both their own field as well as computer science and statistics. Similarly, the next generation of computer scientists and statisticians needs to understand deeply the real needs of domain scientists. Students can no longer develop tools and models in isolation, because the resulting “hammers” fail to meet the growing needs of the data-enabled sciences.

The “Advanced Data Science” option will address this need through a shared, core data science curriculum. The goal of the option is to focus specifically on the students who will advance the state-of-the-art data science tools and methods in their domain and thus need advanced data science education.

Curriculum

Curriculum / Courses

To complete the Advanced Data Science option, students must take *three out of four* of the following core courses. All these courses are ***existing courses***. While the courses focus on methods, they commonly use a variety of scientific applications and datasets as examples in lectures and assignments. All outcomes will be evaluated in the context of the corresponding course.

Data Management: CSE 544 Database Systems Internals (4 credits)

This course covers the principles of data management. The course includes topics related to effectively using a data management system (locally or in a public cloud), building applications on top of such a system, and building the internals of such a system. Detailed topics include: the relational data model, the relational algebra, and SQL, query execution and optimization, transaction processing and recovery, views, data integration, ETL, OLAP, warehousing, big data management and analytics, parallel databases (shared-nothing architectures, parallel query processing, fault-tolerance, skew), modern systems (main memory databases, key-value stores, NoSQL, column-oriented databases), non-relational data models (key-value, trees, graphs, arrays, streams), general principles concerning installation and tuning of a database system and using a data management system in a public cloud.

Expected learning outcomes:

- Knowledge and hands-on experience using a data management system.

- Knowledge and hands-on experience analyzing large datasets with a cloud data management service.
- Knowledge and hands-on experience with database system internals, including knowledge of state-of-the-art architectures and algorithms.
- Knowledge of important data management principles: physical data independence, declarative query languages, query optimization, parallel query processing, etc.

Machine Learning: CSE 546 Machine Learning (4 credits) or STAT 535 Statistical Learning: Modeling, Prediction, and Computing (3 credits).

Practical methods for identifying valid, novel, useful, and understandable patterns in data. Basic statistics. Induction of predictive models from data: classification, regression, probability estimation. Discovery of clusters and association rules. Methods for designing systems that learn from data and improve with experience. Supervised learning and predictive modeling: decision trees, rule induction, nearest neighbors, Bayesian methods, neural networks, support vector machines, and model ensembles. Unsupervised learning and clustering. Emphasis will be on the ability to use and understand at a high level various machine learning methods rather than an in-depth study of theoretical considerations.

Expected learning outcomes:

- Knowledge of core machine learning principles, including data, features, models, loss-functions, parameters, complexity, bias-variance tradeoff, and evaluation.
- Knowledge of the theoretic foundations and practical applications of the core ML tasks, including classification, regression, clustering, reinforcement learning.
- Ability to derive and prove basic properties of certain methods, including convergence and sample complexity.
- Ability to choose an appropriate method for a practical machine learning task, and implement and evaluate the method.

Data Visualization: CSE 512 Data Visualization (4 credits)

Techniques and algorithms for creating effective visual displays of information based on principles from graphic design, perceptual psychology, cognitive science and statistics. Topics include data and image models, visual encoding methods, graphical perception, color, animation, interaction techniques, graph layout, and automated design. Methods of presenting complex information to enhance comprehension and analysis. Incorporating visualization techniques into human-computer interfaces.

Expected learning outcomes:

- An understanding of the key techniques and theories used in visualization, including data models, graphical perception and techniques for visual encoding and interaction.
- Ability to read and discuss research papers from the visualization literature.
- Knowledge of visualization methods and issues for common data domains and analysis tasks, including multivariate, network, text and geographic data.

- Familiarity with state-of-the-art data visualization tools.
- Practical experience designing, building and evaluating visualization systems.

Statistics: STAT 509 Introduction to Mathematical Statistics: Econometrics I (5 credits)

Examines methods, tools, and theory of mathematical statistics. Covers, probability densities, transformations, moment generating functions, conditional expectation. Bayesian analysis with conjugate priors, hypothesis tests, the Neyman-Pearson Lemma. Likelihood ratio tests, confidence intervals, maximum likelihood estimation, Central limit theorem, Slutsky Theorems, and the delta-method.

Review of random variables; transformations, conditional expectation, moment generating functions, convergence, limit theorems, estimation; Cramer-Rao lower bound, maximum likelihood estimation, sufficiency, ancillarity, completeness. Rao-Blackwell theorem. Hypothesis testing: Neyman-Pearson lemma, monotone likelihood ratio, likelihood-ratio tests, large-sample theory. Contingency tables, confidence intervals, invariance. Introduction to decision theory.

Expected learning outcomes:

- Knowledge of the mathematical theory of probability at the calculus level, including marginal and conditional densities, moment generating functions, transformations of random variables.

- Simple approaches to prediction based on the conditional expectation function and best linear predictor.

- Basic knowledge of Bayesian statistical inference: the normal and beta-binomal models. Awareness of basic Monte Carlo (and Markov chain Monte Carlo) approaches to integration.

- Knowledge of the frequentist approach to inference, including point estimators, hypothesis testing and confidence intervals. Properties of estimators including unbiasedness, efficiency, consistency. Likelihood ratio tests.

- Familiarity with simple asymptotic theory: the weak law of large numbers, central limit theorem, Slutsky theorems and the delta-method.

- The ability to derive the limiting distributions of common statistics.

- Ability to derive maximum likelihood estimators and construct asymptotic confidence intervals.

Or STAT 512-513 Statistical Inference (4 credits each).

STAT 512-513 covers much of the above, though provides a bit more rigorous treatment and goes more in depth on some of the theoretical concepts. It does not cover Bayesian statistical inference (3rd bullet above).

CHEME 599F eScience Community Seminar (1 credit)

The eScience Community Seminar is open to all. Students in the Advanced Data Science option must register for ***at least 4 quarters***. The seminar serves as an informal environment for presentations and discussions on research that is relevant to all data science researchers around campus. The seminar takes place in the new **Data Science Studio**, located in the physics and astronomy building. Several times during the quarter, the seminar is replaced by the UW Data Science Seminar, which highlights external speakers from other research institutions and industry. Finally, we also engage in at least one discussion session per quarter focused on ethical issues around Big Data & Data Science. In this seminar, the students taking the proposed advanced data science option will periodically give talks about their research in addition to other data science researchers on campus.

Expected learning outcomes:

- Breadth of knowledge in data science research and activities around campus.
- Breadth of knowledge in data science research beyond the University of Washington.
- Practice in public speaking about own research to an inter-disciplinary audience.
- Knowledge of ethical issues related to data science research.

Total credits

The total credit requirement is thus 11 credits in courses (2 courses @ 4 credits and 1 course @ 3 credits) and 4 seminar credits. Note that all 11 credits from courses will serve directly as credits in the PhD program of the student's home department. The seminar credits will be extra credits.

Timing and Location

All courses are offered during the day. Each course is one quarter in length with two, 80-min, in-person lectures per week on the main Seattle campus.

The eScience Community Seminar meets once a week for one hour during the day on the Seattle campus.

Relationship of the proposed option to existing degree programs

All participating departments have already approved a Big Data track. In most departments, the requirements for the Big Data track are identical to the requirements described above.

Three departments have a small number of additional requirements. In order to be recognized with an “Advanced Data Science” option, students in those departments ***will also have to satisfy these extra requirements***:

1. In Computer Science & Engineering there is an additional post-quals requirement: *Satisfactorily complete one additional course* with explicit emphasis on advanced “Big Data” techniques:
    - A fourth core course, *unless* already used as one of the 6 quals courses.
    - CSE 547 / STAT 548 - Machine Learning for Big Data, *unless* already used as one of the 6 quals courses.
    - STAT 513 - Statistical Inference.
    - A new Big Data Management course planned for Winter 2016 (Magda Balazinska)
    - EE 578 - Convex Optimization
    - STAT 527 - Nonparametric Regression and Classification
    - STAT 538 - Advanced Statistical Learning
    - CSE 552 - Distributed and Parallel Systems Data, *unless* already used as one of the 6 quals courses.

2. In Genome Sciences, students must also complete Genome 540: Computational Molecular Biology.

3. In Statistics, the Big Data Track has these additional elements:
    1. Statistics Core: STAT 570, STAT 581, STAT 582
    2. ML/BD Core:
        (i) One advanced ML course: STAT 538 or STAT 548
        (ii) One elective beyond the core requirements:
            * Advanced Statistical Learning (STAT 538)
            * Machine Learning for Big Data (STAT 548)
            * Graphical Models (CSE 515)
            * Visualization (CSE 512)
            * Databases (CSE 544)
            * Convex Optimization (EE 578)
    3. All other statistics PhD requirements hold, except:
        * STAT 571 may be used in place of the consulting project course
    4. Maintain an average grade of 3.3 in all ML/BD core courses.

    Prerequisites
    Students must have taken STAT 534, or equivalent, prior to entering the track.

Appendix: We attached the detailed description of the Big Data tracks in each department. The descriptions can also be found online for many departments:

- Astronomy Big Data Track
- ChemE Big Data Track
- CSE Big Data Track
  http://www.cs.washington.edu/students/grad/specializedtracks/bigdata
- Genome Science Big Data Track
  http://www.gs.washington.edu/academics/gradprogram/handbook/first/bigdata.htm
- Oceanography Big Data Track

- Statistics Big Data Track
  http://www.stat.washington.edu/graduate/programs/machinelearning/

Infrastructure Requirements

Because the option builds on existing courses, no special infrastructure is required. The eScience community seminars are offered in the new, Data Science Studio that opened this Fall 2014.

Faculty

The option builds on existing courses that are already being offered in the Computer Science and Statistics departments:

1. CSE 544 is regularly taught by Profs Dan Suciu (Professor of Computer Science & Engineering), Magdalena Balazinska (Associate Professor of Computer Science & Engineering), and Alvin Cheung (Assistant Professor of Computer Science & Engineering).
2. CSE 546 is regularly taught by Prof Carlos Guestrin (Associate Professor of Computer Science & Engineering) with Profs Pedro Domingos (Professor of Computer Science & Engineering) and Luke Zettlemoyer (Assistant Professor of Computer Science & Engineering) as potential instructors as they have taught it previously; and STAT 535 is regularly taught by Prof Marina Meila (Associate Professor of Statistics).
3. CSE 512 is regularly taught by Prof Jeff Heer (Associate Professor of Computer Science & Engineering).
4. STAT 509 is regularly taught by Prof Caren Marzban (Part-time Lecturer in Statistics); STAT 512-513 are both regularly taught by Prof Michael Perlman (Professor of Statistics).
5. The eScience Community Seminar is coordinated by Prof David A. C. Beck (Research Assistant Professor in Chemical Engineering and Director of Research - Life Sciences for the eScience Institute).

For the purpose of the option, the faculty will teach these same courses with the same frequency as without the option. Because the “Advanced Data Science” option is reserved for the most advanced students who focus their thesis on developing data science infrastructure, the enrollment in these existing courses is not expected to increase significantly. These courses already have high enrollments. For example, nearly 40 graduate students took CSE 544 in Winter 2015, 5 of which were taking the Big Data track in a department other than CSE. The advanced data science option thus raises course sizes by only a small percentage.

Program Oversight

The oversight of the program will be delegated to a steering committee. The committee will comprise one faculty from each participating department. A program director will be selected from among the steering committee and by that committee. The committee will review the selection of the director every two years.

We already have such a committee in the context of the NSF IGERT program upon which we are building this new PhD option. This existing committee will become the steering committee for the Advanced Data Science Option with the same membership.

Prof. Magdalena Balazinska is the current director and will become the first director for the Advanced Data Science option. Prof. Andrew Connolly (Astronomy), Ginger Armbrust (Oceanography), Emily Fox (Statistics), David Beck (Chemical Engineering), Bill Noble (Genome Sciences), and Carlos Guestrin (Computer Science & Engineering) form the committee and will remain on the committee for the Advanced Data Science Option. Jennifer Worrell (Computer Science & Engineering) is the administrative program manager.

The steering committee will meet at least once per quarter to review the program and make necessary adjustments to the curriculum, requirements, admission process or other. As a result, we expect the overall workload of the committee to be light.

To apply for the Data Science option, a student must be a full-time PhD student in one of the participating departments. The student must send an email to the graduate advisor in their department and declare interest in pursuing the Advanced Data Science option. The research advisor of the graduate student must approve the application. If the research advisor approves the application, the graduate advisor will register the interest by coding the student into the PhD program option code. All interested students whose advisors approve their application into the program will be allowed to pursue the Advanced Data Science option. There will be no limit on enrollment. However, the core courses in the option have pre-requisites that students must be willing to fulfill. The advanced nature of these courses and pre-requisites will naturally limit enrollment.

At the end of each quarter, the academic advisor from each department will send a summary list of all the students in the department who are pursuing the Advanced Data Science option and those who have completed the option in the last 12 months to the steering committee.

Administration

Because the oversight required for students to participate in the option is minimal, we will not require any administrative resources or other support services beyond the existing graduate advising staff in the participating departments.

Students

The option will serve the needs of existing PhD students in the participating departments. Students will thus first be admitted into the PhD programs of the participating departments. Once admitted, students will be eligible to apply for the Advanced Data Science option following the process described above.

In the rare case where a graduate student does not complete a PhD but instead leaves only after completing the master’s part of the PhD program, the student will retain the “Advanced Data Science option” recognition on their transcript if the student finishes all the requirements of the option before leaving the PhD program. This point is important because an Advanced Data Science option opens the door to a larger number of employment opportunities in industry.

All participating departments will advertise the availability of the option by posting information about the option on their department websites.

Diversity

Only students from participating departments will be able to register for the “Advanced Data Science Option”. As a result, the option will directly leverage the departments’ efforts to increase diversity.

Program Assessment

The steering committee will collect feedback from the students and their advisors at least once per year to get their input on the program. The feedback will take the form of online surveys. We have a formal evaluation process in place in the context of the IGERT program. For the first two years of the option, we will use the IGERT evaluation as the program evaluation. We will continue a similar, though lighter weight, process beyond the lifetime of the IGERT grant. In the context of the IGERT, the evaluation includes in-person interviews, social network analysis, and surveys. Beyond the IGERT grant, we will focus the evaluation on online surveys. The steering committee will recommend changes to the program based on the feedback from the students and faculty.

Budget

Because the option builds on existing courses and existing PhD programs, the option will be revenue neutral. Existing unit staffing will support the option in terms of oversight.

Unit and College/School/Campus Approval

Please see attached.