Development of a Computational and Data-Enabled Science and Engineering Ph.D. Program

Paul T. Bauman*, Varun Chandola †, Abani Patra‡, and Matthew Jones§

*Mechanical and Aerospace Engineering
Computational and Data-Enabled Science and Engineering
University at Buffalo, State University of New York, Buffalo, NY 14260-4400
Email: pbaumun@buffalo.edu

†Computer Science and Engineering
Computational and Data-Enabled Science and Engineering
University at Buffalo, State University of New York, Buffalo, NY 14260-2500
Email: chandola@buffalo.edu

‡Mechanical and Aerospace Engineering
Computational and Data-Enabled Science and Engineering
University at Buffalo, State University of New York, Buffalo, NY 14260-4400
Email: abani@buffalo.edu

§Center for Computational Research
University at Buffalo, State University of New York
Buffalo, New York 14203
Email: jonesm@buffalo.edu

Abstract—The previous two decades have seen the successful deployment of Computational Science programs in universities across the globe. These programs are aimed at training scientists and engineers to tackle problems requiring interdisciplinary approaches to finding solutions to scientific and engineering problems and the development of new computing, as exemplified by the co-design approach to exascale architectures and applications. Thus, the programs emphasize preparation in applied mathematics, numerical analysis, and scientific computing in addition to science and engineering work relevant to the target application. The rise of so-called “Big-Data” applications and the use of large data in business decision support and even in computational science workflows like uncertainty analysis are driving a need for training in data sciences. This paper makes the argument that, rather than treating topics in machine learning, statistics, etc. as stand-alone fields of study that students learn as electives, data-science should be an integral part of interdisciplinary training for future researchers. This approach is at the core of the newly developed Computational and Data-Enabled Science and Engineering (CDSE) Ph.D. program at the University of Buffalo. This paper describes the development of the Ph.D. program, the target student audience, and strategies for effectively executing the proposed curriculum.

I. INTRODUCTION

The onset of the “big data” focus over the past decade has made the training of data-scientists a critical endeavor in science and engineering graduate programs in universities around the country. Indeed, the McKinsey Global Institute estimates that the economy will need 140,000-190,000 additional trained personnel for “deep analytical talent positions”, and 1.5 million more “data-savvy managers” to take full advantage of big data in the United States [15]. A recent New York Times article states: “Universities can hardly turn out data scientists fast enough.” [16] It is estimated that the national shortage for such talent is at least 60% [15]. In addition to these estimates for the commercial sector, we believe that the academic and national/commercial laboratory sections will have significant demand for personnel with Ph.D. level training. Recent reports from the Department of Energy’s, Advanced Scientific Computing Advisory Communities and Science and Engineering Advisory Board provide support for this need.

Until recently, training in data sciences has been discipline-specific, typically in mathematics, statistics, or computer science departments 1. While such focus has yielded many research developments in the areas of machine learning and Bayesian statistics and produced many interesting developments in fields as diverse as hand-writing recognition [13] and the discovery of the Higgs boson, such focused, isolated training will not meet the interdisciplinary needs of future research and business activities. Indeed, we view the current trends as a parallel to those when “big computing” emerged and drove the development of computational science programs.

Since 2001 at the University at Buffalo, training in computational science and engineering has been done as part of

1A simple exercise in analyzing the course offerings at University at Buffalo reveals five separate classes on machine learning in five departments.
a certificate program offered through various participating departments (Physics, Mathematics, Chemical and Biological Engineering, Mechanical and Aerospace Engineering, and Computer Science and Engineering). During this time enrolment in the core courses for the certificate grew from single digits to more than twenty, representing all of the participating departments. The success of this effort testifies not only to the interdisciplinary nature and interest in computational science and engineering, but also has led directly to a more formal degree program at our institution.

In this paper, we make the argument that training in the data sciences should be an interdisciplinary endeavor and greatly parallels the development of computational science programs. Further, we describe the forthcoming Computational and Data-Enabled Science and Engineering (CDSE) program at the University of Buffalo that aims to balance the training in computational and data sciences while building on expertise in a core discipline. We describe how the CDSE program addresses this interdisciplinary need and will provide an example of successful interdisciplinary training in the computational and data sciences.

II. NEED FOR INTERDISCIPLINARY DATA SCIENCE TRAINING

The emergence of business based drivers of “big data” has been well documented. Another driver for interdisciplinary data science training is the emergence of statistical methods in the engineering disciplines (see for e.g. [17] which describes the interdependent use of big computing and big data analysis methods in an uncertainty quantification workflow). The previous five years have seen a concerted effort in developing algorithms and technologies to support inference on engineering systems. This is due, in large part, to the growth of computing resources, which is making the application of such approaches feasible. Previously, for calibration problems for example, even deterministic approaches required substantial computational resources [11]. Additionally, there is a growing literature on “uncertainty quantification” (UQ) [18], [14], driving the incorporation of probabilistic methods into scientific and engineering simulation. To leverage UQ analyses, probability information must be supplied for the uncertain parameters and, thus, inference algorithms are being explored to provide such information. Nevertheless, there are still many challenges in making statistical approaches practical. Furthermore, there is vast, and still growing, literature in Bayesian statistics, machine learning, data analytics, database systems, etc. Interdisciplinary training will be required for researchers in the engineering disciplines to bridge this gap in order to leverage and contribute to data science methodologies.

From the infrastructure side, there is a big disconnect between computing and data science practitioners. Traditionally, numerical simulation codes are optimized to run on high performance computing clusters. On the other hand there have been significant advances in data processing infrastructure including specialized appliances such as Cray Urika, IBM Netezza, etc., and distributed/cloud-based architectures coupled with software solutions such as Hadoop. While the computing community has become well-versed in exploiting the capabilities of the traditional HPC clusters they are not trained to utilize the data processing solutions for their needs. On the other hand, data science practitioners have not been able to develop analytic and machine learning codes that can exploit the parallelism offered by HPC clusters.

Several universities have started offering Masters program in Data Sciences. Some universities offer such programs from within the Computer Science departments (Columbia, Stanford) while others have established multi-department centers (NCSU [1], NYU [2]) to provide a multi-disciplinary flavor to the program. Most of these programs are dominated by courses in Computer Science and applied statistics. Lately, there have been few Ph.D. programs designed for the data science area. Of these, the one currently being offered at University of Washington is based within the Computer Science department. The same is true for programs still in development at University of Utah [4] and The University of Utah [4]. Thus, most existing programs lack interdisciplinary structure in terms of the curriculum, especially in the areas of applied mathematics, numerical methods, and high performance computing. Nevertheless, it must be said that the time-bound nature and technology-use focus of a Master’s program does not provide adequate time and opportunity to provide the interdisciplinary balanced training in domain, data, and computational sciences.  

2 As a senior administrator remarked – such training unsparing of breadth and depth is reminiscent of that provided in the medical sciences to M.D./Ph.D.s – whose number is necessarily limited but whose contributions as leaders of the field is of the greatest impact.
III. PARALLELS WITH COMPUTATIONAL SCIENCE PROGRAM DEVELOPMENT

The previous three decades have seen the deployment of ever growing computational resources. The advancement of these resources have enabled increasingly sophisticated modeling and simulation of scientific and engineering systems. However, in the earlier periods of supercomputing resource availability, there was much more limited interdisciplinary training of scientists and engineers. Algorithms and software deployed on these systems lacked scalability and efficiency. Such factors were recognized early and were major drivers in the developments of interdisciplinary Ph.D. programs. Indeed, such factors have motivated the support of specific fellowship programs, such as the Computational Science Graduate Fellowship (CSGF) [5].

These computational science programs emphasize interdisciplinary training needed for tackling the modeling and simulation of scientific and engineering systems on supercomputers including applied mathematics, numerical analysis, algorithms, software engineering, and domain sciences. Among such programs are the computational science, engineering, and mathematics program at The University of Texas at Austin [6], the computational and mathematical engineering program at Stanford University [7], the computational science program at Florida State University [8], the computational science and engineering program at Georgia Tech University [9], and the computational and applied mathematics program at Rice University [10]. Additionally, many Master’s degree programs have been instituted.

This interdisciplinary approach to research and education has yielded significant advancements in numerical simulation. Indeed, as presented by David Keyes, Figure 2 illustrates performance advancements in the simulation of magnetohydrodynamic flows (fusion) both according to hardware and architecture improvements and to algorithmic improvements. As can be seen, while there have been several orders of magnitude gain in hardware performance, there is as much if not more advancement due to algorithmic advancements.

While topics such as uncertainty quantification, statistical inference, Bayesian methods, etc. are becoming an active area of research in the scientific computing and computational mechanics disciplines, there is still no formal training requirements in these topics within the computational science programs, particularly at the Ph.D. level. We anticipate that the significant insights that drive advancement and innovation will come through interdisciplinary efforts and education in the era of “big data”, similar to the trends that have been observed during the era of “big computing”, where combined training in applied mathematics and science and engineering have significantly advanced modeling and simulation research.

IV. COMPUTATIONAL AND DATA-ENABLED SCIENCE AND ENGINEERING PROGRAM DESCRIPTION

To address the need for interdisciplinary training in computational and data sciences, the University of Buffalo will support a Ph.D. program in Computational and Data-Enabled Science Engineering (CDSE). The program brings together faculty from a variety of different disciplines, including computer science, engineering, finance, management, biostatistics, mathematics, and pharmacy thereby impacting several Decanal Units: The School of Engineering and Applied Sciences, The College of Arts and Sciences, The School of Pharmacy and Pharmaceutical Sciences, and The School of Management. Of these, the largest impact is expected to be in the School of Engineering and Applied Sciences and the College of Arts and Sciences, both of whom are primary partners in this program. The School of Pharmacy and Pharmaceutical Sciences and The School of Management have identified this program as a primary opportunity and are active participants. The other decanal units will be impacted indirectly; new CDSE research will cut across all department and decanal boundaries, creating new opportunities in many fields. The new data enabled computational science effectively makes a new method of inquiry.

The School of Engineering and Applied Sciences has a Department of Computer Sciences and Engineering, but its degree program is in Computer Sciences. The CDSE program has an explicit focus on integrating domain science, applied mathematics knowledge, high performance computing, and large-scale data analysis and creation of digital artifacts (software, tools for analysis of large data, etc.). In much the same way as engineering programs build upon core knowledge and conceptual frameworks from the natural
A. Program Detail

Students will enter the program with an MS from a “home” department, bringing a breadth of study in one of the typical relevant scientific fields: applied mathematics, computer science, engineering, etc. It is expected that this will be the primary source of the “domain science” that will be the application focus of the students dissertation. Students then engage in a program of study in three focus-areas (with a minimum number of course hours in each area): applied mathematics and numerical methods, high performance and data intensive computing, and data sciences. The subsequent dissertation by the student must contain elements of each of the three core areas. Figure 3 graphically depicts the program of study.

To enable effective coverage we require that the students complete 30 credit hours of study in these topics with a minimum of 9 credit hours in each. While this is a fairly high level of required classes for a Ph.D. program we provide two good arguments for this. Firstly, the balanced training we seek requires this diversity of inputs. Secondly, (at least anecdotally) it is our experience that most students entering such a Ph.D. program have had at least 6 and often 12 or 15 credit hours of our required classes during their disciplinary MS training and are thus likely to waive some of this. Note, that the topics are broad and a final set of classes needs to be customized to meet the learning needs of the individual student – shoring up deficiencies and building on strengths. The dissertation committee of the student will be engaged in this customization.

A key feature of our program is the need for a disciplinary MS program. While many comparable programs have opted for a MS in Data Sciences, we have deliberately chosen to stay away from that option. The reasoning is that the computational and data sciences training is best conducted in the context of a discipline. Deep knowledge of a discipline will provide both context for the training and the opportunities to put the acquired knowledge to profitable use (in employment opportunities and/or research topics) in a short time frame. In our experience, students with purely cross disciplinary training often struggle to find an intellectual “home” and a community of peers. We believe our design avoids these pitfalls and provides the right opportunities to the students for learning and personal growth.

Topics in the first area, applied mathematics and numerical methods, include more traditional topics in computational science: numerical linear algebra, partial differential equations, functional analysis, finite element methods, numerical analysis, optimization, game theory, etc. The second focus area, high performance and data intensive computing, again borrows elements from computational science programs, including high performance computing, parallel computing, etc., but also includes topics typically relegated to computer science programs: large-scale distributed data systems, data mining, and data intensive computing. Finally, the third focus area, data science features courses in Bayesian statistics, machine learning, stochastic processes, and information theory. Classes in the all topics listed above are offered across many departments. For example, courses with heavy focus on Bayesian statistics are offered in the Biostatistics, Computer Science and Engineering, and Mechanical and Aerospace Engineering departments. The diversity in the departmental offerings mimics the interdisciplinary nature of the program, providing different view points that students may not otherwise encounter from single-department curricula.

B. Qualifiers and Program Flow

The qualifying process also reflects the interdisciplinary approach taken by the program. Qualifying will be a two-part process. In the first part of the qualifying process, candidates have three requirements:

1) Select advisor and dissertation committee. The committee should reflect the intellectual diversity of the CDSE program. Thus, the committee should not have more than one core member and one additional member with primary academic affiliation in the same department. The core members must be CDSE affiliated faculty. The committee has to be approved by the graduate director. The core members must be CDSE affiliated faculty. The committee has to be approved by the graduate director.

2) Decide on a research topic.

3) Present a dissertation prospectus (short research proposal) that articulates the CDSE challenges and application need to their committee. Alternatively, the candidates may substitute this with an acceptable technical publication.

The student is required to complete the first stage of process by the end of their second semester in the program. While this is earlier than is typical in other traditional Ph.D. programs, it was chosen in order to ensure that...
1) the student is developing an interdisciplinary research agenda, and
2) the student is targeting an application to which the proposed theory, algorithms, tools, etc. can be applied.

While it is anticipated that students entering the program from an engineering discipline will naturally target an application of interest, students entering from mathematics or computer science may not automatically have such an application. Thus, the application requirement is meant to spur the selection of such an application. The application requirement is one of the defining features of the CDSE Ph.D. program that naturally differentiates it from more traditional natural sciences programs, such as mathematics and computer science.

The second part of the qualifying process focuses on examination and a dissertation proposal.

1) The dissertation committee will conduct an oral examination of background material needed to conduct the research outlined in the dissertation prospectus. This requirement must usually be met no later than the end of the third semester in the program for which the student is registered.
2) Candidates will complete the previously discussed core courses that will support the their dissertation research, and their subsequent CDSE career. The candidate will refine the proposal developed during his/her initial year in the program into a final dissertation research proposal. This dissertation research proposal will be presented to the candidates dissertation committee. The committee will then provide feedback in order to facilitate the completion of the students dissertation. Both the core course work and the dissertation research proposal must be completed by the end of the 5th semester in the CDSE program.

Completion of these two stages of the qualifying process will then make the student eligible to become Ph.D. candidates. Upon successful completion of the dissertation and defending the dissertation to the candidates dissertation committee, the student will be awarded a Ph.D. in Computational and Data-Enabled Science and Engineering.

C. Program Self Assessment

The self-assessment process of the Ph.D. program will proceed at three different but coordinated levels: (i) Graduate director; (ii) Steering committee; (iii) Advisory board. At the lowest level, the graduate director (who by his/her role will be the person most in contact with the students) will on an ongoing basis, in consultation with the graduate studies committee and with the instructors of the relevant courses: (a) evaluate how well individual courses fit within the planned curriculum, (b) review whether the outcomes for the courses are consistent with the desired learning outcomes, (c) evaluate whether the financial support system and overall structure of the program are effective, (d) assess whether changes need to be made for individual students and whether more systematic changes are necessary. He/she will bring any systematic issues to the attention of the CDSE steering committee. At the next level, the CDSE steering committee will meet every semester to monitor the program, discuss any issues that may arise and determine an appropriate course of action to resolve them if necessary. If more is needed, they will bring any remaining issues to the attention of the CDSE advisory board. Finally, the CDSE advisory board (which includes the Deans of the participating decanal units) will meet yearly to review the program and determine whether any high-level changes are required. The University funding being used to establish the programs has an annual evaluation before the next round if funding is made available. We have gone through one of these reviews having achieved every one of our objectives. After the program is well established we will integrate this into the University’s graduate program assessment and accreditation processes.

V. CONCLUDING REMARKS

The growing importance of data sciences in both the public and private sectors are driving efforts to develop better training for future employees and researchers. We believe this movement has many analogies to the growth of “big computing” and the emergence of computational science and engineering Ph.D. programs that brought an interdisciplinary approach to student development. Such an interdisciplinary approach is at the core of the developing CDSE Ph.D. program at the University of Buffalo. As with the development of computational science programs, the CDSE program will yield students with the interdisciplinary training necessary to bridge the gap of traditional data analytics and science and engineering applications utilizing HPC systems and statistical and informatics algorithms.

REFERENCES

[6] https://www.ices.utexas.edu/graduate-studies/.
[8] https://www.sc.fsu.edu/graduate/phd.


**APPENDIX**

### List of Core Classes

1) Data Science
   a) CSE 574: Introduction to Machine Learning
   b) EE 634: Principles of Information Theory and Coding
   c) IE 512: Decision Analysis
   d) MAE 674: Optimal Estimation Methods
   e) STA 502: Statistical Inference
   f) STA 503: Regression Analysis
   g) STA 521: Introduction to Theoretical Statistics I
   h) STA 536: Statistical Design and Analysis of Experiments
   i) STA 567: Bayesian Statistics

2) Applied and Numerical Mathematics
   a) IE 572: Linear Programming
   b) IE 575: Stochastic Methods
   c) MAE 702: Applied Functional Analysis
   d) MTH 539: Methods of Applied Mathematics I
   e) MTH 540: Methods of Applied Mathematics II
   f) MTH 543: Fundamentals of Applied Mathematics
   g) MTH 631: Analysis I
   h) MTH 649: Partial Differential Equations

3) High Performance and Data Intensive Computing
   a) CSE 562: Database Systems
   b) CSE 587: Data Intensive Computing
   c) CSE 603: Parallel and Distributed Processing
   d) CSE 999: Large-Scale Distributed Data Systems
   e) IE 551: Simulation and Stochastic Models
   f) MAE 609: High Performance Computing I
   g) MAE 610: High Performance Computing II
   h) STA 545: Data Mining I
   i) STA 546: Data Mining II