The New Computational and Data Sciences Undergraduate Program at George Mason University

Kirk Borne, John Wallin, and Robert Weigel

Computational and Data Sciences, George Mason University,
Fairfax, VA 22030, USA

Abstract. We describe the new undergraduate science degree program in Computational and Data Sciences (CDS) at George Mason University (Mason), which began offering courses for both major (B.S.) and minor degrees in Spring 2008. The overarching theme and goal of the program are to train the next-generation scientists in the tools and techniques of *cyber-enabled science (e-Science)* to prepare them to confront the emerging petascale challenges of data-intensive science. The Mason CDS program has a significantly stronger focus on data-oriented approaches to science than do most computational science and engineering programs. The program has been designed specifically to focus both on simulation (Computational Science) and on data-intensive applications (Data Science). New courses include Introduction to Computational & Data Sciences, Scientific Data and Databases, Scientific Data & Information Visualization, Scientific Data Mining, and Scientific Modeling & Simulation. This is an *interdisciplinary science* program, drawing examples, classroom materials, and student activities from a broad range of physical and biological sciences. We will describe some of the motivations and early results from the program[1]. More information is available at http://cds.gmu.edu/.

1 Data-Intensive Science: A New Vision for Science Education

The development of models to describe and understand scientific phenomena has historically proceeded at a pace driven by new data. The more we know, the more we are driven to tweak or to revolutionize our models, thereby advancing our scientific understanding. This data-driven modeling and discovery linkage has entered a new paradigm [1]. The acquisition of scientific data in all disciplines is now accelerating and causing a nearly insurmountable data avalanche [2]. In astronomy in particular, rapid advances in three technology areas (telescopes, detectors, and computation) have continued unabated [3] – all of these advances

[1] The development of the Mason Computational and Data Sciences undergraduate program is sponsored by the NSF CCLI (Course, Curriculum, and Laboratory Improvement) program, through award # 0737091.

G. Allen et al. (Eds.): ICCS 2009, Part II, LNCS 5545, pp. 74–83, 2009.

lead to more and more data [4]. With this accelerated advance in data generation capabilities over the coming years, we will require an increasingly skilled workforce in the areas of computational and data sciences in order to confront these challenges. Such skills are more critical than ever since modern science, which has always been data-driven, will become even more data-intensive in the coming decade [4,5]. Increasingly sophisticated computational and data science approaches will be required to discover the wealth of new scientific knowledge hidden within these new massive scientific data collections [6,7].

We live in an information age in which we are inundated with facts, tending toward information overload. Though the data glut problem is not limited to science, science is first and foremost a forensic discipline – we gather evidence, first to develop a hypothesis, then to test our hypothesis, and finally to vindicate or else to invalidate the hypothesis, at which point we gather more evidence, and the process continues. We must educate the next generation scientists, if not all citizens, in the principles of evidence-based reasoning, fact-based induction, and data-oriented science. In particular, we must muster educational resources to train a skilled data-savvy workforce: one that knows how to find facts (i.e., data, or evidence), access them, assess them, organize them, synthesize them, look at them critically, mine them, and analyze them.

2 Background and Motivation

The growth of data volumes in nearly all scientific disciplines, business sectors, and federal agencies is reaching epidemic proportions. This epidemic is characterized roughly by a doubling of data each year. It has been said that “while data doubles every year, useful information seems to be decreasing” [8], and “there is a growing gap between the generation of data and our understanding of it” [9]. In an information society with an increasingly knowledge-based economy, it is imperative that the workforce of today and especially tomorrow be equipped to understand data. This understanding includes knowing how to access, retrieve, interpret, analyze, mine, and integrate data from disparate sources. This is emphatically true in the sciences. The nature of scientific instrumentation, which is becoming more microprocessor-based, is that the scale of data-capturing capabilities grows at least as fast as the underlying computational-based measurement system[10]. For example, in astronomy, the fast growth in CCD detector size and sensitivity has seen the average size of a typical large astronomy sky survey project grow from hundreds of gigabytes 10 years ago (e.g., the MACHO survey), to tens of terabytes today (e.g., 2MASS[2] and Sloan Digital Sky Survey[3] [3], up to a projected size of tens of petabytes 10 years from now (e.g., LSST, the Large Synoptic Survey Telescope[4] [4]). LSST will produce one 56K x 56K (3-Gigapixel) image of the sky every 20 seconds, generating nearly 30 TB of

[2] http://www.ipac.caltech.edu/2mass/

[3] http://www.sdss.org

[4] http://www.lsst.org

data daily for 10 years. In the field of Space Weather and Solar Physics, NASA announced in 2008 a science data center specifically for the SDO (Solar Dynamics Observatory). The SDO will obtain one 4K x 4K solar image every 10 seconds, generating 1 TB of data per day. NASA recognizes that previous approaches to scientific data management, analysis, and mining will simply not work. Consequently, we see the floodgates of data opening wide in astronomy, high-energy physics, bioinformatics, numerical simulation research, geosciences, climate monitoring and modeling, and more. Outside of the sciences, it is widely documented that the data flood is in full force in banking, healthcare, homeland security, drug discovery, medical research, insurance, and (as we all have seen) e-mail. The application of data mining, knowledge discovery, text mining, and e-discovery tools to these growing data repositories is essential to the success of agencies, economies, and scientific disciplines.

2.1 Data Sciences: A National Imperative

The article “Agencies Join Forces to Share Data” calls for more training in data skills [11]. This article describes a new Interagency Working Group on Digital Data, representing 22 federal agencies in the U.S., including the NSF, NASA, DOE, and more. The group plans to set up a robust public infrastructure so that all researchers have a permanent home for their data. One option is to create a national network of online data repositories, funded by the government and staffed by dedicated computing and archiving professionals. Who are these computing and archiving professionals? We believe that this professional workforce must be trained in the disciplines of computational and data sciences. We are addressing this societal need through the new CDS curriculum in the Mason Department of Computational and Data Sciences (CDS).

Within the scientific domain, Data Sciences is becoming a recognized academic discipline. In a recent Data Sciences Journal article [12], it is argued that now is the time for Data Sciences curricula. In another article [13], Data Science is again promoted as a rigorous academic discipline. Further, there was a 2007 NSF-cosponsored workshop on Data Repositories, which included a track on data-centric scholarship, where they explicitly state what we now believe: “Data-driven science is becoming a new scientific paradigm – ranking with theory, experimentation, and computational science” [14]. Another Data Sciences Journal article states: “Without proper management of continuously-produced important data and without the productivity of new disciplines based on data, we cannot solve important problems of the world” [15]. An excellent article recently described Informatics, the new paradigm of data-intensive science, as “the use of digital data, information, and related services for research and knowledge generation” [16]. Consequently, many scientific disciplines are developing subdisciplines that are information-rich and data-based, to such an extent that these are now becoming (or have already become) recognized stand-alone research disciplines and academic programs on their own merits. The latter include bioinformatics and geo-informatics, but will soon include astroinformatics, e-Science, medical informatics, and data science. Several national study groups have issued reports on the urgency of establishing scientific and educational programs to face the data flood challenges:

1. National Academy of Sciences report: “Bits of Power: Issues in Global Access to Scientific Data” (1997) [17];
2. NSF report on “Knowledge Lost in Information: Report of the NSF Workshop on Research Directions for Digital Libraries” (2003) [18];
3. NSB (National Science Board) report on “Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century” (2005);
4. NSF “Atkins Report” on “Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure” (2005) [19];
5. NSF-sponsored report with the Computing Research Association on “Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda” (2005);
6. NSF report on “Cyberinfrastructure Vision for 21st Century Discovery” (2007) [20];
7. JISC/NSF Workshop on Data-Driven Science & Repositories (2007) [14].

Each of these reports issues a call to action in response to the data avalanche in science, engineering, and the global scholarly environment. For example, the NAS “Bits of Power” report lists 5 major recommendations, one of which includes: “Improve science education in the area of scientific data management” [17]. More recently, the Atkins Report stated that skills in digital libraries, metadata standards, digital classification, and data mining are critical [19].

3 Computational and Data Sciences at Mason: CUPIDS

The new Computational & Data Sciences curriculum at Mason uniquely responds to the recommendations of these national studies and reports. The urgent need for such a curriculum cannot be overstated, as the Atkins Report has said: “The importance of data in science and engineering continues on a path of exponential growth; some even assert that the leading science driver of high-end computing will soon be data rather than processing cycles. Thus it is crucial to provide major new resources for handling and understanding data” [19]. The core and most basic resource is the human expert, trained in key data science skills. As stated in the 2003 NSF “Knowledge Lost in Information” report, human cognition and human capabilities are fundamental to successful leveraging of cyberinfrastructure, digital libraries, and national data resources [18].

CUPIDS[5] is the NSF-supported “Curriculum for an Undergraduate Program in Data Sciences” at Mason in the CDS Department. The central goal for the CUPIDS project is *to increase student’s understanding of the role that data plays across the sciences as well as to increase the student’s ability to use the technologies associated with data acquisition, mining, analysis, and visualization.* We have five objectives for this project:

[5] Funded through NSF Award # 0737091.

1. To teach students what Data Science is and how it is changing the way science is being done across the disciplines
2. To change student's attitudes about and improve their confidence in using computers to address scientific data problems
3. To increase student's abilities to use visualization for generating and addressing scientific questions
4. To increase student's abilities to use databases for scientific inquiry
5. To increase student's abilities to acquire, process, and explore experimental data with the use of a computer

As evident from this list, the goals of the CUPIDS project are focused primarily on the Data Sciences, which are a subset of the educational goals and programs within CDS. We describe several aspects of these programs below.

3.1 The CDS Degree Program and Curriculum

Students are required to complete a total of 18 credits (6 core courses) in computational and data sciences (CDS), 15 credits in computer science, 23 credits in mathematics, 6 credits in statistics, 21-25 credits in a science concentration, and 3-9 credits in CDS electives. Three concentration areas are currently offered: Physics, Chemistry, and Biology. Additional concentrations may be added to the program in the future (perhaps astronomy, materials science, and geosciences). For a given concentration, the 21-25 credits that a student must take in that science discipline include the core courses that are required for majors matriculating in those programs. As much as possible, the core CDS courses include scientific examples and applications from all of the science concentrations. Of course, this implies a certain degree of heterogeneity in the scientific knowledge of the students who come from different fields. Consequently, the primary focus of the CDS courses are on the techniques of computational and data sciences, and not on the specific experimental and theoretical bases of the various science disciplines. As part of their education, students are encouraged to undertake an optional research project that allows them to gain useful experience in the development of simulations and other aspects of computational science. To facilitate this interdisciplinary science environment, the CDS faculty have degrees in the science disciplines (including Astronomy & Astrophysics, Space Weather, Physics, Computational Fluids, Materials Science, Computational Chemistry, Computational Statistics, Applied Mathematics, and more) and many of the faculty have affiliations (or joint appointments) within those other departments.

The six required computational and data sciences core courses are:

1. *CDS 101 Introduction to Computational & Data Sciences* - Introduces the use of computers in scientific discovery via simulations and data analysis.
2. *CDS 301 Scientific Information and Data Visualization* - The techniques and software used to visualize scientific simulations, complex information, and data visualization for knowledge discovery.
3. *CDS 302 Scientific Data and Databases* - Data and databases used by scientists, including data types, database queries, and distributed data systems.

4. *CDS 401 Scientific Data Mining* - Data mining techniques from statistics, machine learning, and visualization applied to scientific knowledge discovery.
5. *CDS 410 Modeling and Simulations I* - Numerical differentiation and integration, initial-value and boundary-value problems for ordinary differential equations, methods of solution of partial differential equations, iterative methods of solution of nonlinear systems, and approximation theory.
6. *CDS 411 Modeling and Simulation II* - The application of modeling and simulation methods to various scientific applications, including fluid dynamics, solid mechanics, materials science, molecular mechanics, and astrophysics.

We describe below two of the core courses in more detail and we enumerate the desirable skills that we expect for students who complete the program of study.

3.2 Course: Introduction to Computational and Data Sciences

This course provides an interdisciplinary introduction to the tools, techniques, methods, and cutting edge results from across the Computational and Data Sciences. Students are shown how computational tools are fundamentally changing our approach in the experimental, observational, and theoretical sciences through the use of data and modeling systems. No mathematical background is assumed, other than high school algebra. Qualitative results are emphasized, to show the problems and challenges facing researchers today. Examples are drawn from both the "real world" familiar to students and also from the frontiers of science where these techniques are being used to solve complex problems.

Upon completion of the course, students should be able to:

1. Describe how data are represented within a computer, from binary numbers to arrays and databases.
2. Explain how scientific data are acquired, processed, stored, reduced, and analyzed using computers.
3. Express how we create knowledge from data and information using visualization and data mining.
4. Create effective ways to visualize simple data sets.
5. Conduct and explain simple simulations of complex phenomena.
6. Express how changing computing technologies further scientific research, and how the technological and scientific progress are tied together.

Lecture topics in the course include: the scientific method; computer internals (binary numbers and logic circuits); computer algorithms and tools (Matlab introduction); data acquisition; signal processing (understanding noise and error); scientific databases; data reduction and analysis; data mining; computer modeling; numerical simulations; visualization; high-performance computing; and future directions in computational science.

3.3 Course: Scientific Data Mining

This course provides a broad overview of the knowledge discovery (data mining) process, as applied to scientific research. Data mining is the search for hidden meaningful patterns in large databases (e.g., *find the one gene sequence in a large genome DNA database that always associates with a specific cancer*). These patterns and relationships are often expressed as rules (e.g., *if a blue star-like object is found next to a faint unusual-shaped galaxy in a large astronomy database, then the blue object might be a distant quasar whose outburst in being triggered by a collision with that galaxy*). Consequently, data mining is sometimes referred to as the process of converting information from a database format into a knowledge-based rule format. Identifying these patterns and rules from enormous data repositories can provide significant competitive advantage to scientific research projects and in other career settings.

Data mining is motivated and analyzed in this course as the "killer app" for large scientific databases (i.e., a key enabler of scientific discovery). Data mining techniques, algorithms, and applications are covered, as well as the key concepts of machine learning, data types, noise handling, feature selection, data transformation, and similarity/distance metrics. Techniques are analyzed specifically in terms of their application to scientific research problems. Several scientific case studies are drawn from the science research literature, including astronomy, space weather, geosciences, climatology, bioinformatics, numerical simulation research, drug discovery, health informatics, combinatorial chemistry, digital libraries, and virtual observatories. Prerequisites for this course include the undergraduate Scientific Data and Databases course and mathematics/statistics courses.

Upon completion of the course, students should be able to:

1. Express the role of data mining within scientific knowledge discovery.
2. Express the most well known data mining algorithms and correctly use data mining terminology.
3. Express the application of statistics, similarity measures, and indexing to data mining tasks.
4. Identify appropriate techniques for classification and clustering applications.
5. Determine approaches used for mining large scientific databases (e.g., genomics, virtual observatories).
6. Recognize techniques used for spatial and temporal data mining applications.
7. Express the steps in a data mining project (e.g., cleaning, transforming, indexing, mining, analysis).
8. Analyze classic examples of data mining and their techniques.
9. Effectively prepare data for mining and use data mining software packages.

Lecture topics in the course include: scientific motivation for data mining; quantitative and statistical concepts; software packages; data preparation (previewing, cleaning dirty data, normalization, transformation); distance and similarity metrics for clustering and classification; supervised learning methods; unsupervised learning methods; scientific data mining case studies; and special topics (time series, image mining, spatial data, and outlier/event/anomaly detection).

3.4 Results from the Program and Future Work

Early results from the program include the following: (a) the B.S. (CDS major) was approved in 2007; (b) the minor was approved in 2008; (c) the first courses

(CDS 101 and 302) were offered in Spring 2008; (d) additional courses (CDS 301 and 401) were offered in Fall 2008; (e) ∼10 students are currently majoring in the program; and (f) one new course has been approved and will be offered in the coming year: *CDS 151 Data Ethics in an Information Society.*

We have identified a need for additional courses in order to retain students in the program. The initial selection of core courses includes only one course before the Junior year: CDS 101. This was not a problem with the first wave of students majoring in the program, all of whom were Junior-level transfer students. To address the retention problem for the newest students, we are developing additional courses, which may include courses in computational tools for scientists and discipline-specific topics. We have received a small "pedagogy" grant from the Mason College of Science to develop a new Gen Ed course for science majors: *CDS 120 Computational and Data Tools for Scientists.* This course will employ novel student-led peer instruction approaches, to enable the course to scale to large numbers of students. The goal is to make this course the default computational tools course for all science majors at Mason. At present, these students take a computer languages course from the I.T. school, which does not have a science focus nor does it include scientific applications. CDS 120 will cover presentation tools (Powerpoint, HTML), analysis tools (spreadsheets), databases, search methods, basic programming in Matlab, overview of data acquisition and signal processing, and numerical simulations (verification and validation).

In the area of discipline-specific courses, we are considering developing a basic course in tools for computational biology and bioinformatics, to meet the specific needs of students in Biology. Though this course is biology-specific, it is consistent with another discipline-specific course that is now under review: "Materials Science with Applications to Renewable Energy." We are also discussing with the Mason Physics & Astronomy department a concept for developing a course in Astroinformatics. In these cases, we are drawing upon the considerable expertise of the CDS faculty in specific science areas to develop and to offer computational and data science courses for majors in those scientific disciplines. Such courses would not be required for all CDS majors, but they will be appropriate electives for CDS students in the corresponding science concentration.

3.5 Similar Programs at Other Universities

We are aware of a few similar programs at other universities. In nearly all of these cases, the focus is either on (i) computational science (with little attention to data sciences), (ii) data sciences (generically, not within the context of teaching science), or (iii) data sciences within the context of a single specific science. Type (i) progams include the CACR (Center for Advanced Computation and Research) at Caltech (who do have a strong connection with Astronomy and therefore are moving toward a focus on cyber-enabled data sciences), and many other computational science programs (e.g., LSU's CCT, U. Texas' TACC). Type (ii) programs include the Discovery Informatics program at the College of Charleston. Type (iii) programs include the POCA (Partnership in Observational and Computational Astronomy) at SCSU and Clemson University, Purdue's Discovery Informatics program, the emerging joint programs between CS and astronomy departments at Notre Dame, and similarly at U. Michigan. Beyond these are the science programs in data sciences, including Cornell's new DISCOVER data-driven science program[6] (with a focus on astronomy plus other disciplines) and the new e-Science Institute at U. Washington[7] (focusing on oceanography, environmental sciences, and astronomy) – these two programs appear to be most similar to the Mason CDS program.

4 Conclusions and Summary Remarks

Computational and Data Sciences are emerging fields involving applications of sophisticated simulation and data-oriented methods to build models and solve problems in science and engineering. Recently emerged interdisciplinary areas in the chemical, physical and biological sciences (such as biotechnology, nanotechnology, molecular electronics, photonics in nanoscale systems, and energetics of DNA/protein binding) require highly-qualified professionals with strong computational skills to work closely with experimentalists in solving complex scientific and engineering problems. Emerging data-intensive science fields (such as Geoinformatics, Bioinformatics, Astroinformatics, and Materials Informatics) require specially trained professionals with strong data skills to address ubiquitous data-intensive applications in science, industry, and government. Computational and data sciences complement existing theoretical and experimental science approaches and may be thought of as a new mode of scientific inquiry.

The new CDS undergraduate *science* program at Mason complements the existing graduate program in Computational Science & Informatics (CSI), which has existed since 1992, having graduated 175 PhDs to-date. There are over 90 graduate courses in the Mason CSI program, covering many science disciplines. In conclusion, we summarize the key features of the CDS program:

– *Who?* – Students with a broad interest in computers and sciences will benefit from the program.
– *Why?* – Students graduating with a traditional discipline-based bachelors degree in biology, chemistry, or physics generally do not have the required computational background necessary to participate as productive members of modern interdisciplinary scientific research teams, which are becoming increasingly computational- and data-intensive. The motivating theme and goal of the CDS program are to train the next-generation scientists in the tools and techniques of *cyber-enabled science (e-Science)* to prepare them to confront the emerging petascale challenges of data-intensive science.
– *What?* – The BS program in CDS provides science students with a variety of opportunities to become research professionals possessing interdisciplinary knowledge, including sciences and applied mathematics, augmented with strong computational and data-oriented skills. This program has a significantly stronger focus on data-oriented approaches to science than do most

[6] http://arecibo.tc.cornell.edu/DRSG/Links.aspx

[7] http://escience.washington.edu/

Computational Science and Engineering (CSE) programs. Graduates from this program will acquire interdisciplinary knowledge and will be able to apply scientific principles in solving complex real-world problems.

- *How?* – Students in this program are exposed to a wide range of computational and data science applications, and will learn computational science tools, high-performance computing, applied and theoretical computational techniques, modeling and simulation, statistical analysis, optimization, data & information visualization, scientific database applications, scientific data mining & knowledge discovery in databases (KDD), and data-intensive science research methods. The CDS program has been designed specifically to focus both on simulation (Computational Science) and on data-intensive applications (Data Science) within an interdisciplinary science environment.

References

1. Mahootian, F., Eastman, T.: Complementary Frameworks of Scientific Inquiry. World Futures journal (2009) (in press)
2. Bell, G., Gray, J., Szalay, A.: arxiv.org/abs/cs/0701165 (2005)
3. Gray, J., Szalay, A.: Microsoft technical report MSR-TR-2004-110 (2004)
4. Becla, J., et al.: arxiv.org/abs/cs/0604112 (2006)
5. Szalay, A.S., Gray, J., VandenBerg, J.: arxiv.org/abs/cs/0208013 (2002)
6. Gray, J., et al.: arxiv.org/abs/cs/0202014 (2002)
7. Borne, K.D.: Data-Driven Discovery through e-Science Technologies. In: 2nd IEEE Conference on Space Mission Challenges for Information Technology (2006)
8. Dunham, M.: Data Mining Introductory and Advanced Topics. Prentice-Hall, New Jersey (2002)
9. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)
10. Gray, J., et al.: Scientific Data Management in the Coming Decade, arxiv.org/abs/cs/0502008 (2005)
11. Butler, D.: Agencies Join Forces to Share Data. Nature 446, 354 (2007)
12. Smith, F.: Data Science as an Academic Discipline. Data Science Journal 5, 163 (2006)
13. Cleveland, W.S.: Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistics Review 69, 21 (2007)
14. NSF/JISC Repositories Workshop (2007), `http://www.sis.pitt.edu/~repwkshop/`
15. Iwata, S.: Scientific "Agenda" of Data Science. Data Science Journal 7, 54 (2008)
16. Baker, D.N.: Informatics and the 2007-2008 Electronic Geophysical Year. EOS 89, 485 (2008)
17. Bits of Power: Issues in Global Access to Scientific Data, `http://www.nap.edu/catalog.php?record_id=5504`
18. Knowledge Lost in Information: Report of the NSF Workshop on Research Directions for Digital Libraries, `http://www.sis.pitt.edu/~dlwkshop/report.pdf`
19. Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure, `http://www.nsf.gov/od/oci/reports/atkins.pdf`
20. Cyberinfrastructure Vision for 21st Century Discovery, `http://www.nsf.gov/pubs/2007/nsf0728/index.jsp`