MS in Data Science

proposed by

Department of Computer Science, B. Thomas Golisano College of Computing and Information Sciences
Department of Information Sciences and Technologies, B. Thomas Golisano College of Computing and Information Sciences
School of Mathematical Sciences, College of Science
Center for Quality and Applied Statistics, Kate Gleason College of Engineering

I. Justification and Program Goals

Data science, a term first coined in 2008 by data analytics leaders at Facebook and Google[1], is a new and inherently multidisciplinary field—combining computing, mathematics, statistics, and the sciences—devoted to the management and analysis of massive, mostly unstructured data. The computing technology advances of the 1990s laid a foundation for data generation and storage on a staggering scale: 10 TB of data from one aircraft engine in 30 minutes; 8 TB of Twitter data per day[2]; 72 TB per second from Australia's ASKAP radio telescope array.[3] The sheer volume of this data is one issue; arguably a bigger issue is what knowledge lies within. The first commercial sector to address this problem was financial services, where physical scientists and other high-performance computing experts versed in advanced data analysis are in high demand.[4] Variously called Business Intelligence, Data Analytics, Data Mining, and Big Data Analytics, the term of choice for the field is now **Data Science**.

The hallmark of a data scientist is the ability to design new solutions and approaches that turn the raw data deluge into actionable knowledge. This new field is at the confluence of the foundational areas such as computer science, information technology, mathematics, and statistics as well as the domain areas such as biology, bioinformatics, meteorology, astronomy, and business management. The focus of this endeavor is to coax discernible patterns, connections and structures from seemingly random information to help individuals, businesses, industries and government make decisions.

Employment opportunities for data scientists are plentiful in almost all sectors of the economy and demand is correspondingly high.[5,6] Data Science Central, a website devoted to big data practitioners, lists 6000 companies employing data scientists—a list headed by giants Microsoft, SAS, Google, Facebook, IBM, and Amazon, but also including companies in healthcare, telecommunications, education, marketing, and consulting.[7]

Presently there are several data science master's degree programs in New York State including at Syracuse University, University at Albany, RPI, Columbia University and New York University; there are many more across the country, with undoubtedly more on the horizon. MOOC provider Coursera offered its first data science class in 2013.

---

[1] Data Scientist: The Sexiest Job Of the 21st Century. By: Davenport, Thomas H., Patil, D. J., Harvard Business Review, 00178012, Oct2012, Vol. 90, Issue 10.
[2] Oracle: Big Data for the Enterprise. White paper, June 2013, p.3.
[3] http://www.atnf.csiro.au/projects/askap/data_transport.html; accessed March 2, 2014.
[4] Big Data meets HPC. By: Tracey, Suzanne. February 2014, pp. 6-11.
[5] http://www.gartner.com/newsroom/id/2207915, accessed March 2, 2014.
[6] http://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html?pagewanted=all&_r=0 accessed March 2, 2014
[7] http://www.datasciencecentral.com/profiles/blogs/6000-companies-hiring-data-scientists, accessed March 2, 2014.

The nature, focus and objective of the programs that exist around the country are as varied as the departments in which they are offered. The MS in Data Science program at RIT, a natural extension of the GCCIS Advanced Certificate in Big Data Analytics, is intended to provide a stronger computational science perspective of Data Science and Big Data Analytics to students with necessary core skills (data management, mathematical and statistical modeling, and advanced data analysis) and the flexibility to develop depth in a particular domain of their choice that might involve other programs at RIT. As a multi-college degree program, the MS in Data Science offers students a unique opportunity to pursue theoretical as well as applied study, and to work with faculty who are active researchers in the fields of data science infrastructure and domain-based analytics and can provide hands-on experience with real data and real problems.

RIT's approach to data science is distinctly different from the existing programs. Firstly, this degree is **career focused**: Most of the data science programs available across the nation are research- centric with a curriculum focusing on the fundamental theory, algorithmic and architectural support of data management and analytics in scale. The proposed M.S. degree program, on the other hand, has a strong career-oriented focus, aiming to equip students with practical skills to handle large-scale data management and analysis challenges that arise in their daily work. This degree will fit perfectly into the career-oriented education that RIT is recognized for. The career-focused degree will also significantly benefit from one of the world's largest co-op programs at RIT, which brings in practical problems, real-world data, and software tools commonly adopted in industry to enrich our curriculum.

Secondly, the program is **highly interdisciplinary and domain driven**: Most of the existing data science programs have a curriculum that teaches students the general knowledge of data management and analytics. Nonetheless, data science enjoys a wide spectrum of application fields. Data generated from different scientific, engineering and business domains may carry distinct characteristics. Furthermore, the way of managing and analyzing data could vary significantly from one domain to another. To allow students to learn practical knowledge in this field, the proposed M.S. degree focuses on domain specific problems and solutions. It also provides students the opportunities of interdisciplinary study. Important domains (e. g, biology, physics, and statistics) are judiciously selected and integrated as part of the curriculum to provide customized, domain-specific training to next-generation data scientists. The interdisciplinary nature of the degree will be well accommodated by the diverse background of faculty in the Golisano College, active involvement of the faculty in the College of Science and the Center for Quality and Applied Statistics to provide the analytical foundation and participation by the faculty in many disciplines to contribute the required domain expertise that goes beyond computing.

The inter-disciplinary M.S. Degree program in Data Science is intended to educate and train professionals in the analysis of big data and to give them the necessary tools for a successful career in this growing field. Between 2012 and 2013, there was a 13% year-over-year increase in available jobs that require big-data skills. One of the goals of the proposed program is to feed this voracious demand for highly-skilled professionals who are able to handle the increasingly complex nature of the data that surrounds us.

Given the rich and successful tradition of RIT as a technological university that answers the call of business, industry and government to prepare future leaders, professionals and decision makers, the program will be a good fit to RIT's mission and a valuable addition to RIT's portfolio of programs.

Program goals

The MS in Data Science is intended to:

- Educate students to meet the demand for data scientists in government, industry, and education.

- Develop data scientists with innovative skills in advanced data management and analytics.
- Engage students with active faculty research projects in this field.
- Leverage collaboration with colleagues across RIT to deliver inter- and multi-disciplinary educational experiences to students.

II. Program Description

Students entering this master's program must have basic programming skills and knowledge of probability and statistics. Students who do not meet the entrance requirements have a number of existing options for bridge courses to close that gap. For example, programming skills can be acquired by taking CSCI 603 Advanced C++ and Program Design or CSCI 605 Advanced Java Programming.

The curriculum provides students with the necessary core skills they need as data scientists; students can structure elective courses to develop a depth of experience and knowledge in a specific application domain. The program comprises 30 credits:

- Four core courses (12 credits)
- Two computing electives (6 credits)
- Two domain electives (6 credits)
- Thesis (6 credits).

Alternatively, students could opt for a project instead of a thesis. The credits then would be

- Four core courses (12 credits)
- Two computing Electives (6 credits)
- Three domain electives (9 credits)
- Project (3 credits)

The four core courses will provide the foundation in big data management, big data analytics, knowledge processing from unstructured data, and numerical analysis. All of these courses currently exist as shown in Table 1.

Table 1: Core Courses

Course Number and Title	Frequency
CSCI 620 Introduction to Big Data	F, Sp, Su
CSCI 720 Big Data Analytics **OR** CQAS 747 Principals of Statistical Data Mining	F, Sp, Su F, Sp
ISTE 612 Knowledge Process Technologies	F, Sp
MATH 611 Numerical Analysis	F

Students will be given a choice of CSCI 720 or CQAS 747, which cover similar topics but with different foci based on student background and interest.

The elective courses will come from the application domains including but not limited to such areas as computing, statistics, biology and astrophysics. The list of possible electives shown in Table 2 is for the purpose of illustration as this list will change over time as courses are refined or replaced; for example, topics in languages and tools such as R and Python, which are popular with data scientists, may also be incorporated into current courses. New courses in data intensive computing, scientific data modeling, mathematical modeling, and other domain areas may be developed as additional electives.

Students may also take electives from different disciplines across RIT, with the permission of the program coordinator and, if required, the relevant department. Students may select their thesis or project advisor from the participating faculty in the degree program.

Table 2: Possible Electives

	Course	Title	Frequency
Sample Electives: Computing	CSCI 621	Database System Implementation	F
	CSCI 622	Secure Data Management	Sp
	CSCI 652	Distributed Systems	F, Sp
	CSCI 654	Foundations of Parallel Computing	F
	CSCI 714	Scientific Visualization	F
	CSCI 721	Data Cleaning and Preparation	Sp
	CSCI 729	Topics in Data Management. Different seminar courses on current Big Data topics such as text mining, web mining, cloud data management, streaming data, and others	F, Sp, Su
	CSCI 736	Neural Networks and Machine Learning	Sp
	CSCI 737	Pattern Recognition	F
	ISTE 724	Data Warehousing	F
	ISTE 740	Geographical Information Science and Technology	F
	ISTE 780	Data-driven Knowledge Discovery	Sp
	ISTE 782	Visual Analytics	Sp
Sample Electives: Mathematics and Statistics	MATH 605	Stochastic Processes	F
	MATH 711	Advanced Methods in Scientific Computing	Sp
	CQAS 701	Foundations of Experimental Design	F, Sp
	CQAS 741	Regression Analysis	F, Sp
	CQAS 756	Multivariate Analysis	F, Sp
	CQAS 773	Time Series Analysis and Forecasting	F, Sp
Sample Electives: Biology	MATH 695	Statistical Models for Bioinformatics	Sp
	MATH 761	Mathematical Biology	Sp
Sample Electives: Astrophysics	ASTP 611	Statistical Methods for Astrophysics	F
	ASTP 720	Computational Methods for Astrophysics	Sp

III. Fit with RIT Academic Portfolio Blueprint Characteristics and Criteria

As a burgeoning new discipline and area of inquiry, data science clearly advances RIT's education mission and strategic direction. The program will foster faculty and student research in many areas, and provide students with relevant experiential learning through real world data and problems. The collaborative nature of this degree showcases the interdisciplinary synergy that is possible at RIT, and positions the university to be a leader in this field. As we have shown, there is substantial external demand for graduates from this type of program, and the popularity of the Big Data-related courses in GCCIS, currently the electives most in demand in the existing MS Computer Science program, are evidence for internal demand as well. The MS in Data Science will be a valuable addition to RIT's academic portfolio.

IV. Synergy with other programs

Since Data Science deals with the study of data collection, analysis of data and application of the analysis, the program will feature strong interplay among Computer Science and Information Sciences and Technologies for the technical foundation, Mathematical Sciences and Applied Statistics for the analytical foundation, and domain areas for data and application context and analysis. The domain area could be in any field that is using or could leverage the data science topics addressed within this program. For example, the programs in Bioinformatics and Astrophysical Science and Technology in the College of Science could provide the domain knowledge and the corresponding courses for students to apply their skills to analyze gene sequences and gravitational wave data. Programs in Computer Science and Information Sciences and Technologies have courses and expertise in such diverse areas as social networks, web analytics, visualization, geographical information systems, and text mining. In addition, students from industry may bring their domain background to perform data analytics in a relevant application area. Other domains (such as digital humanities, imaging, and engineering) are certainly possible.

V. Administrative structure for the new program

The M.S. in Data Science program will be administered under the Golisano College of Computing and Information Sciences (GCCIS). Faculty from Computer Science, Information Sciences and Technologies, the School of Mathematical Sciences, the Center for Quality and Applied Statistics and other participating units will be affiliated with this program and assume instructional and advising responsibilities. A Program Director will be selected from the GCCIS faculty engaged in the program to perform administrative duties and work with the participating academic units to deliver the curriculum. This director will require a course release for these duties. Depending on enrollment projections, a half-time staff position may be needed to assist the Program Director.

VI. Enrollment Management Expectations and Sustainment

To be included after discussion with Diane Ellison in Enrollment Management.

VII. Impact on Resources

The program proposes to use many courses that are already being offered in our existing programs. New courses that might be needed to complete the program would be offered as part of the regular course offering and would be part of the normal faculty workload. Additional computing resources will be necessary to support a computing cluster and labs.

VIII. Conclusion

The proposed inter-disciplinary M.S. program in Data Science adds another career-focused, cutting-edge discipline to RIT's portfolio, addressing the exponentially growing need for skilled data science professionals. The program takes advantage of the computing, mathematical, and statistical expertise that is available at RIT and offers an avenue for prospective students to enter this important career path. The growth in the program will have the beneficial impact of increasing the research activities of faculty and students in Data Science and related areas, thus enhancing our research profile and providing our students with more opportunities for research and exploration.

Subject: RE: MS in Data Science
Date: Wednesday, March 12, 2014 9:20:58 AM Eastern Daylight Time
From: Diane Ellison
To: Mohan Kumar
CC: Mihail Barbosu, Steve Zilora, Hans-Peter Bischof, James G. Miller (EMCS VP)

Mohan and all,

We have reviewed the concept paper for the MS in Data Science and the feedback you provided to my follow up questions. Following is our feedback regarding enrollment potential:

The program offers the opportunity to attract students with a range of academic programs, and to leverage the strengths and resources of a number of different departments at RIT toward what has become a high demand, interdisciplinary content area.

Given that the program is proposed as a campus-based program, we assume that the majority of students will study full-time and significant enrollment demand will come from the international student market. Adding on-line options will increase future demand even further.

Our initial estimate is that the program could enroll 20 students in the first year, and 25 students in subsequent years.

This estimate is based on a 30-hour, campus based program that provides additional bridge courses for students who lack the required programming skills, and knowledge of probability and statistics required for admission to the program. We assume that it will take most students 1 ½ years (3 semesters plus summer courses) to complete, that most will require 1 – 2 bridge courses, and courses will be offered in summer as well as fall and spring. Enabling students to commence studies in fall, spring and summer will contribute to even greater demand for entry.

Let me know if you have additional questions,

Diane Ellison
Assistant Vice President
Part-time and Graduate Enrollment Services
Rochester Institute of Technology
phone: 585-475-7284
fax: 585-475-7164
dmeges@rit.edu

From: Mohan Kumar
Sent: Monday, March 10, 2014 9:27 AM
To: Diane Ellison
Cc: Mihail Barbosu; Steve Zilora; Hans-Peter Bischof; James G. Miller (EMCS VP)
Subject: Re: MS in Data Science

Diane,

Here is the almost complete concept paper.
Please let us know if you have any comments. Let me know if you want to meet or give your comments by phone/email.