The Institute for Data Science

APPROVED BY THE BOARD OF TRUSTEES

OCTOBER 11, 2013

UNIVERSITY of ROCHESTER

The Institute for Data Science

The signature project of the 2013–2018 University Strategic Plan will be the creation of a University-wide Institute for Data Science.

Data science has emerged as one of the defining disciplines of the 21st century. We intend the University of Rochester to be among the world's leading institutions in this burgeoning new discipline.

We build upon enormous strengths in data science including the Health Science Center for Computational Innovation, the Department of Computer Science in Arts, Sciences & Engineering, Biostatistics in the School of Medicine and Dentistry, and historic achievements in machine learning and artificial intelligence. In the past five years, the University has spent more than $50 million to support new faculty, staff, and computing infrastructure. The results of this earlier investment are palpable. More than 100 principal investigators have been awarded a total of $307 million in research relying in part upon high performance computation during the past three years.

We now seek an additional $50 million in endowed funds to make our Institute for Data Science one of the nation's leaders. Our vision is for a $50 million initiative, comprised of a distinctive new building that will house the new Institute for Data Science and the recruitment of 20 additional outstanding faculty members to complement the great work that is already underway. The Institute will serve as an anchor building to a new Science and Engineering Quadrangle for the Hajim School of Engineering & Applied Sciences and the School of Arts & Sciences. The Institute will be a magnet to the existing and future faculty who are doing outstanding work in this domain and place the University of Rochester in a leadership role both for computational science and applications of data science for decades to come.

A New Architectural Landmark

The Institute for Data Science will be housed in a new, state-of-the art building located adjacent to Hopeman Hall. This location is ideal because it will allow us to accomplish a number of important goals, including renovating portions of Hopeman; joining the renovated and newly constructed spaces; co-locating the Department of Earth and Environmental Sciences, the new Center for Earth and the Environment, and our computer science faculty, who are currently located in the Computer Sciences Building (CSB); and moving the Electrical and Computer Engineering faculty, currently located in Hopeman, to the vacated space in the CSB. Above all, creating these proximities between formerly dispersed entities will encourage productive collaboration to address critical domains such as climate change.

Conceptual rendering showing the proposed new Institute for Data Science at the Hajim School.

The Science and Engineering Quadrangle

The new Data Science Building will complete the Science and Engineering Quadrangle, which is shaped by Hutchison Hall, Robert B. Goergen Hall, and Carlson Library and extend the collegiate character of Eastman Quad to a unique portion of the campus. Only on the Quadrangle can someone walk five minutes in any direction and engage with faculty members in medicine, the humanities, education, and business.

The Data Science Building will also provide greater definition to the Hajim School and the adjacent programs in Arts and Sciences.

Rendering showing the proposed site for the Institute for Data Science in relation to the Science and Engineering Quadrangle at the Hajim School.

Empowering Collaboration

The creation of The Institute for Data Science is the highest University-wide priority for the balance of *The Meliora Challenge*: The Campaign for the University of Rochester.

The University's expertise in data science is currently dispersed across many different departments and divisions. While there are collaborations between individual groups of researchers, there is no umbrella organization that brings them all together.

Many fundamental advances in data science, however, have resulted from collaboration between scientists in seemingly disparate disciplines. For example, collaboration between physicists, mathematicians, electrical and computer engineers, and computer scientists created concepts and tools that are widely used in analyzing complex network systems, ranging from social networks to cellular signaling systems. Such cross-disciplinary training in data science will be vital for strengthening graduate research and education in this rapidly evolving field, and opportunities for supporting such activities exist.

The distinct advantage of creating this institute is that it will enable the coalescence of multiple individual centers in data science that are emerging from domain specific applications. To support the Institute, faculty growth will be critical, as well as the need for space properly configured to support research, visiting faculty and postdocs, and a small administrative office. To leverage replacement faculty searches, emphasis for hiring will be placed upon data science in discipline specific domains. The objective is to recruit a cohort of energetic new faculty members in a variety of departments for whom data science is a critical component of their work, either as developers or users.

Using Big Data to Answer Big Questions

Big Data

According to IBM, every day we create 2.5 quintillion bytes of data—so much that 90 percent of the data in the world today have been created in the last two years alone. These data come from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, financial transaction records, and cell phone GPS signals to name a few. This is known as "big data."

IBM Blue Gene/Q

Beyond the everyday application of big data, its possibilities are profound for science, medicine, and other domains of research. Rochester researchers are using big data to model and predict the spread of infectious diseases; track the popularity of political ideas; understand consumer preferences; predict the existence of planets; understand human origins; conserve resources; and tackle issues that were previously too difficult to address because of the lack of processes to collect, manage, sort, and analyze enormous data sets.

Data Science

Data science is defined as the concepts, methods, and applications for extracting meaning from large-scale data. It is the enabling technology of 21st century commerce, as exemplified by Google, Amazon, Walmart, and SAP (and essentially all technology, financial, and service companies). It is also the foundation of the next health care revolution, supporting data-informed, personalized medicine. Data science is crucial to national security and defense and is a high priority for DARPA, NSF, and other federal agencies. Through high-performance computing, creative computer science, and collaboration across disciplines, researchers are finding themselves on the brink of a new era of data-driven scientific investigation—one that *The New York Times* has said is comparable to the advent of the microscope and telescope.

The field has evolved as a hybrid of research in applied mathematics, statistics, electrical and computer engineering, computer science, and scientific computing, driven largely by the proliferation of data in the digital age. Roughly speaking, applied math and statistics provide mathematical foundations; computer science and electrical and computer engineering provide algorithms and infrastructure; and scientific computing provides methods for numerical simulation, model fitting, and optimization that have been developed across the sciences, engineering and medicine.

Data science can also be characterized by the research goals that mark the leading edge of the field.

- **Developing methods for discovering, or deriving, models of phenomena from large data sets**. This goes beyond hypothesis testing, or running simulations to discover the implications of a model; an ultimate goal of data science is to automate the process of hypothesis formation itself.
- **Developing methods for integrating and interpreting highly heterogeneous and noisy data**. While science traditionally focuses on analyzing experimental data, gathered under carefully controlled conditions, leading-edge work in data science often focuses on finding patterns in unstructured information that was created for other purposes—for example, natural language texts or photographs—and with integrating structured and unstructured data.
- **Managing the movement and analysis of data between a multiplicity of sensing and computing platforms**. Technology to produce and store data has outstripped our ability to transport or analyze data by orders of magnitude. The

multicore revolution is turning ordinary workstations into massively parallel supercomputers, but software tools and methodologies for harnessing this power lag far behind. A key research goal is to develop an infrastructure for big data analytics that seamlessly ties together sensing and computing platforms ranging from supercomputers to personal mobile devices.

The demand for training and talent in this domain of study is growing at a rapid pace, with an expected need for approximately 200,000 technical specialists and a greater need for managers versed in the field of data science. The University of Rochester has already customized a program of study for Xerox engineers to pursue an MS degree in data science, and work is under way to tailor an undergraduate program. Computer science alumni at companies such as 1010data, a data science company based in New York City, which handles data for the New York Stock Exchange, are actively seeking relationships to foster research collaboration and create a pipeline for internships and recruitment. Google, Apple, Microsoft, Facebook, and nearly every other major developer actively recruit students from computer science here at Rochester. Rapid development of other strengths in research and education in this domain will further advance our relationships with these major companies.

Initial Domains of Excellence

Three domains of initial research focus have been identified for the Institute for Data Science: Predictive Health Analytics, Cognitive Systems, and Analytics on Demand.

Predictive Health Analytics

The biggest health care breakthroughs of the next century may not come from the development of particular "magic bullet" drugs, but from an increase in our ability to predict individual health outcomes on the basis of treatments, genomics, and lifestyle and behavioral factors. Making discoveries in this way will require integrating a wide variety of heterogeneous data, including individual genomic data, patient outcome data, and informal natural language records and reports. In-situ behavioral data gathered by and/or about particular subjects through smartphone health applications, wearable sensors, and social media is an increasingly important part of a health profile.

Research in predictive health analytics brings together medicine, computer science, electrical and computer engineering, psychology, and other disciplines. Here are some categories for expanded research in predictive health analytics:

- **Infectious disease tracking and predicting**. The University of Rochester is a leader in tracking and developing methods to control the spread of infectious disease. For example, the University's Respiratory Pathogens Research Center, the only center of its kind in the nation, provides the NIAID Division of Microbiology and Infectious Diseases with the capability of conducting translational and clinical research focused on the development and optimization of control measures for viral and bacterial respiratory pathogens. Researchers in Computer Science and the University of Rochester Medical Center (URMC) have received national attention for their work on using social media to predict disease using social media reports. The URMC's recent application for its Influenza Center of Excellence included proposed work on integrating clinical data with informal social media data.

- **Chronic heart disease care**. The URMC is home to a world center for collection and analysis of cardiac data (the Telemetric and Holter ECG Warehouse). We can further leverage this resource by integrating other data sources to accelerate discovery and quantification of (more) health factors for chronic heart disease patients.
- **Predicting cancer treatment outcomes**. Understanding how to use individual genomic data to predict the outcomes of different treatments for cancer patients is a major focus of the URMC's research. Cancer research at the University involves four major focus areas, each involving between twenty and fifty faculty members from seven or more different departments and centers.
- **Suicide prevention**. Prevention can be enhanced by integrating online social media data, individual behavioral data, and medical records.

Cognitive Systems

One of the most ambitious and exciting domains in data science is to model and/or replicate human perception and cognition. Rochester is uniquely positioned to address this challenge, as it is home to internationally recognized research in cognitive science and artificial intelligence (AI). Increases in our understanding of how the brain makes sense of the world can lead to new algorithms for practical problems in machine vision, machine audition, computational linguistics, and automated reasoning. Likewise, engineering advances in each of those problems can suggest new hypothetical mechanisms to brain scientists.

We use the phrase "cognitive systems" for this section to indicate a broad vision that encompasses (at least) researchers in brain and cognitive science; clinical and social sciences in psychology; neurology; computer science; electrical and computer engineering; the Center for Visual Science; linguistics; and the Eastman School of Music. Opportunities for interdisciplinary collaborations in this area are almost limitless, and the problems to be examined are a fundamental part of the BRAIN Initiative recently announced by the White House as a societal Grand Challenge.

While cognitive systems is an active field of study, we are exceptionally competitive in this field with over 30 faculty members from 7 different departments participating in living cognitive systems. To retain that edge, it is critical that we build on our strengths with investments in faculty positions in AI, including machine learning, computer vision and audition, and computational linguistics, as well as cognitive science.

Areas in which we can build significant strength are those related to language sciences and linguistics. Linguistics plays a key role in tying together computer science and cognitive science research in language processing. Enrollments in our classes in linguistics are at record levels, as are job opportunities for graduates who know both linguistics and computing. Enlarging the faculty in relevant areas of linguistics and language sciences would also allow the University to offer an MS program in computational linguistics—a degree program that has proved to be highly popular at several of our peer universities.

Analytics on Demand

A few years ago, the choice of computing platform for data analytics was relatively simple: a workstation, cluster, or supercomputer, depending upon the scale and nature of the problem. The picture is far more complex today. One reason is the growing scale of data that can be easily captured and stored—terabytes (1000 gigabytes) or petabytes (1000 terabytes) are not unusual. (One gigabyte provides sufficient storage for approximately 150 mp3 music files or approximately 300 digital images from your camera.) The time required to transmit data between the point of capture or storage and the point of analysis can become a limiting factor. For example, if it were going to take too much time to transmit the raw data to a centralized server, it could be necessary to perform analysis and data-reduction at the point of capture. Another reason is the proliferation of new platforms (e.g., smartphones) and potentially disruptive computing hardware with thousands of central processing units.

A programmer or scientist today must take the characteristics of the target data-capture, storage, and analysis chain and computing platforms into account to create an efficient and scalable system. If this is not done with care and expertise, most of the resources can go to waste. For example, a task may run more slowly on a supercomputer than on a cell phone. The growing ubiquity of parallel-processing systems exacerbates the problem because writing efficient code to run in parallel on computing systems is far more challenging that writing code for a single system.

We use the phrase "analytics on demand" to refer to the challenge of creating tools and systems for large-scale data analytics that relieve the end-user—the scientist or programmer—from the need to understand the details of particular platforms and chains of platforms. The scientist writes a program that describes the basic algorithm for the task; the system then takes over, and determines how to divide and parallelize the work in order to make optimal use of resources. Faculty members in computer science and in electrical and computer engineering are already working on parts of this vision.

Realizing the Vision

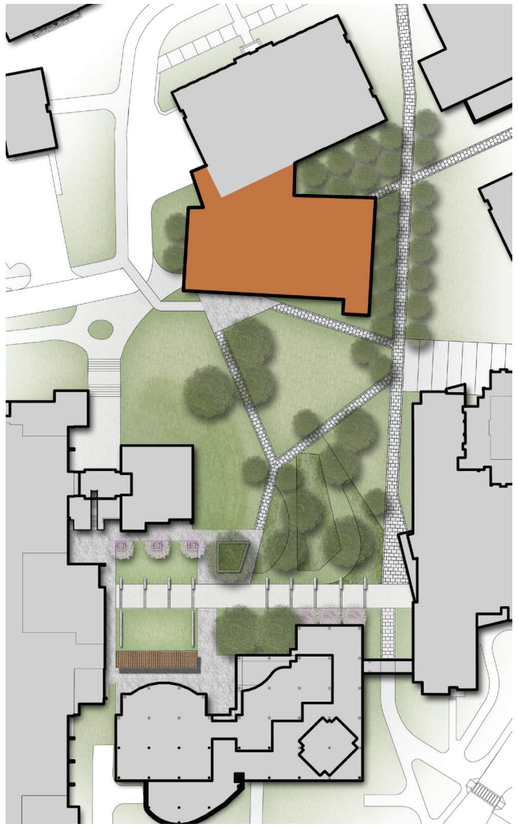
To achieve the full potential of the Data Science Initiative, we seek to create a $50 million endowed fund:

- **The Institute for Data Science and New Facility**—This fund will provide debt coverage for the $25 million cost of our new Data Science facility. After the debt on this facility is paid, these endowed funds will permanently support The Institute for Data Science. The new facility will be a state-of-the-art, 50,000-gross-square-foot building, located immediately adjacent to Hopeman Hall. There will be naming opportunities for The Institute for Data Science and for the new building.

Conceptual rendering showing potential connector between Hopeman Hall and the Institute for Data Science.

- **Professorships for New Faculty Lines in Data Science**—We seek to hire 20 new exceptional faculty members over time. Commitments ranging from $1.5 million to $2 million will provide the foundational support for recruiting an outstanding new faculty.

- **The Center for Energy and Environment**—We will seek $5 million in endowed funds for a critical application of data science in the 21st century, the Center for Energy and Environment and provide an opportunity to name this key center.
- **Data Science Research Funds**—We will seek support for a range of funds, spurring research in specific areas of inquiry.
- **Directorship for the Institute**—Funds will be critical in recruiting an exceptional leader for the new Institute.
- **New Science and Engineering Quadrangle**—We will seek funding and ongoing endowment to finish the landscaping of the quadrangle and ensure its continued maintenance. We will provide an opportunity to name the quad.



Conceptual plan for the new
Science and Engineering Quadrangle.

The University of Rochester is building upon its strengths in data science and a campus culture that attracts quantitative students to secure leadership in this rapidly evolving field. Committing to new faculty lines and signature spaces and programs will enable us to be successful in this endeavor. It is an aspiration we must fully commit to, and if realized, we will decisively move toward making the University of Rochester an international center for data science.


1850
MELIORA
UNIVERSITY of
ROCHESTER