Title: IGERT: Developing Data Scientists for Data Driven Science

Principal Investigator: Arcot Rajasekar, Professor, School of Information and Library Science (Co-director, Data Intensive Cyber-Environments Center) and Chief Scientist, Renaissance Computing Institute

Co-Investigators:

- Helen Tibbo, Alumni Distinguished Professor, School of Information and Library Science
- Stanley Ahalt, Professor, Computer Science and Director, Renaissance Computing Institute
- Anselmo Lastra, Professor and Chair Computer Science Department
- Lawrence Band, Voit Gilmore Distinguished Professor Department of Geography, Director UNC Institute for the Environment

Vision, Goal and Thematic Basis:

With the explosion in scale of scientific data [1-6], scientists must spend more time as data managers. NSF, DOD, DARPA, USGD and NIH [7,8] have initiated programs in "Big Data" to develop best practices in data management, analysis and curation. NSF's Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) [9] recognizes the need to educate and support "analyzing and dealing with challenging computational and data enabled science and engineering (CDS&E) problems" as crucial for innovation in science and engineering. We are proposing the Big Data Science (BDS) program, which will prepare emerging scholars to participate in data-driven science initiatives. It will support education of a new generation of data scientists through integrative courses and interdisciplinary training in data-driven science domains. The program aims to fill two major educational gaps: a lack of awareness by information scientists of large data problems in science and engineering domains, and a complementary lack of awareness by domain scientists of best practices and tools. We will address these gaps in two fundamental ways:

1. augment the training of data scientists (PhD students in information science) with a curriculum on large-scale data management combined with interdisciplinary field training in data-intensive domain sciences, and

2. equip domain scientists (PhD students in domain sciences such as environmental science, biology, astronomy and physics) with tools in advanced data analysis and large-scale data handling that directly supports their disciplinary research goals.

The BDS program will be built on four elements:

a. a set of courses specializing in big data,
b. field experience rotations in multiple domains,
c. a virtual lab with access to large data collections and computational platforms, and
d. seminars and field experiences to expose students to emerging expertise.

The plan of study will prepare both information and domain scientists in large-scale data advanced data analysis and management. The "Data Science" curriculum will cover emerging areas of policy-based data management, digital curation and forensics, cloud computing and scientific workflows, computer security and privacy, graphics and visualization, and foundational information science courses that catalyze cross-disciplinary collaboration. The curriculum will evolve, adopting cutting-edge technological advances and providing training and experience in state-of-the-art tools and applications. The IGERT scholars from both information science and disciplinary sciences will receive a "Data Scientist" certificate to highlight their data-intensive training in the BDS program.

Field experience rotations will allow information and domain science students to study problems faced in multiple domains and to learn strategies, tools and methods for solving them. Field experience rotations will provide direct experience with current and emerging tools, and infrastructure for large-scale data handling and analysis.

The field experiences will be supported through a virtual laboratory to be implemented as part of the IGERT program and will provide researchers with access to large data collections from multiple science and engineering domains, as well as open source software tools and technologies for analyzing and processing large-scale data. The laboratory will also provide access to computational platforms to run large-scale data-intensive tasks. The facility will provide access to cloud and cluster computing, through the RENCI infrastructure, as well as access to national super computer centers such as TACC, SDSC and UIUC. A seminar series will involve leading experts who generate, provide, apply and manage large data collections.

Scientific research increasingly requires collaboration across multi-disciplinary sciences, and hence a scientist needs to be aware of the data practices in multiple domains. This includes not just how data are organized, but also the metadata and ontologies that are part of each domain's practice. Many funding agencies, including NSF and NIH, require data management plans as part of grant proposals. Hence handling large-scale data in multiple disciplines, in large collaboratory projects, with long-term data management is becoming the norm.

Data scientists must be well-versed in different data handling, modeling and analysis tools and practices, and their advantages and short comings. The data scientist must also understand the legal, ethical, and community practice policies related to the entire data life cycle. To our knowledge training in multi-disciplinary Big Data is not available in any university. The concept of rotations to obtain multi-disciplinary field experiences also provides a novel innovation. UNC is well-positioned to provide this training given the strong existing programs in data management and the critical mass of researchers and teachers involved in national and international data curation and data science projects.

Program Elements:

The goal of the BDS program is to promote expertise in Big Data management that improves the ability of researchers to collaborate on science and engineering projects. This includes the ability to manage all phases of the data life cycle, from the conception of a study, to data collection, organizing data into collections, sharing data within data grids, publishing data in digital libraries, analyzing data in processing pipelines, to curating data within preservation archives. Throughout this cycle, the researcher should maintain security and privacy of data, and be aware of workflow environments, diverse computing models, and presentation and visualization approaches. We will provide a well-rounded program to achieve these aims through faculty partnerships, an extended curriculum with hands-on domain field experience, access to large-scale data through a virtual laboratory, embedded field experience in working data centers, and exposure to new and emerging technologies and practices through seminars.

Partners: The BDS program harnesses and integrates the expertise from multiple faculty and national projects across diverse fields situated within UNC into a program that provides training in Big Data. It combines faculty from the School of Information and Library Science (SILS) , the Department of Computer Science, and the Odum Institute for Social Science Research, with disciplinary areas in sciences and engineering including the Department of Physics and Astronomy, Department of Geography, the Institute for the Environment, and Department of Statistics and Operations Research. We plan to reach out to other science and engineering departments for inclusion in our program. We are also looking at North Carolina Central University as a possible partner for master's students at NCCU to train in Big Data and continue with a doctoral program at UNC. Appendix 1 lists the personnel involved in the partnership BDS.

Faculty Strengths: Moore is the PI for Datanet Federation Consortium (datafed.org), an NSF funded-project that is developing a network for sharing and managing multi-disciplinary data that federates several NSF-funded national-scale projects. Moore teaches large-scale data management courses at SILS. Rajasekar leads the NSF's SDCI (irods.org) project for the design, research and development of data grids for policy-oriented data management for large-scale data collections. The middleware iRODS is used by several projects around the world for managing distributed multi-disciplinary data. Tibbo and Lee have led a series of digital curation education and professional engagement projects funded by the Institute for Museum and Library Services (IMLS): DigCCurr, DigCCurr II, Closing the Digital Curation Gap, Educating Stewards of Public Information in the 21st Century and Educating Stewards of the Public Information Infrastructure. Lee is leading a project called BitCurator, funded by the Andrew W. Mellon Foundation, which is applying digital forensics tools and methods to large data files. Marchionini has led projects funded by the NSF and Library of Congress to develop digital libraries for video collections. Hemminger was involved in developing the inter-disciplinary curriculum for the successful Bioinformatics and Computational Biology (BCB, bcb.unc.edu) program at UNC. Our BDS program has elements based on the BCB program. Greenberg is a co-PI of the Dryad (datadryad.org) team, an international repository of data, as well as the DataONE project (dataone.org), an NSF-funded project federating large-scale earth observational data. She is also SILS PI for NESCent (National Evolutionary Synthesis Center), an NSF center of excellence focused on synthesis science. Pomerantz has been part of the NSF National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) project. Ahalt has been involved in leading two large super-computing and data centers (renci.org, osc.edu); he is also teaches advanced computational courses in the CS department and is in the

process of developing a data science research center at UNC. Lastra is a computer graphics researcher and teacher with expertise in general-purpose computation on graphics. Band is involved with the CUAHSI project and leads data-intensive modeling of watersheds. Crabtree manages large-scale statistical data collections for the Odum Institute using the Dataverse network (thedata.org). Liu is member of several institutes at UNC and is coordinating the program on statistical and computational methodology for massive datasets at SAMSI (samsi.info). Cecil is an astronomer whose novel instrument on the SOAR telescope (soartelescope.org) will soon be generating extensive galaxy spectra. Kannappan is involved in multi-wavelength astronomical surveys and is director of the Computational Astronomy & Physics (CAP) REU program.

Curriculum & Field Experience: The BDS program will be based on a plan of study that educates PhD students in a new information and computational science sub-discipline on the management of Big Data. Table 1 shows a representative list of possible courses that will be considered in developing the BDS plan of study. Additional courses will be defined to cover gaps when developing this plan of study.

Table 1 Courses

Course Number	Course Title
School of Information and Library Science courses	
INLS 465	Understanding Information Technology for Managing Digital Collections
INLS 490-187	Issues in Cloud Computing
INLS 490-046	Data Management and Curation
INLS 490-163	Large-scale Databases for Social Networking Services
INLS 523	Database Management
INLS 541	Information Visualization
INLS 582	Systems Analysis
INLS 624	Policy-based Data Management
INLS 706	Bioinformatics Research
INLS 720	Metadata Architectures and Applications
INLS 723	Database Systems III: Advanced Databases
INLS 740	Digital Libraries
INLS 818	Human-Computer Interaction
Department of Computer Science Courses	
CS 535	Introduction to Computer Security
CS 555	Bioalgorithms
CS 633	Parallel and Distributed Computing
CS 655	Cryptography
CS 662	Scientific Computation II
CS 715	Visualization in the Sciences
CS 722	Data Mining
CS 735	Distributed and Concurrent Algorithms
CS 750	Algorithm Analysis
CS 770	Computer Graphics

The Odum Institute for Social Sciences	
Short Courses	Several courses in research analytics, methodologies and data management

For field experiences, the BDS program, with faculty in domain sciences, will develop rotation modules in each domain. These will be short courses that use the virtual laboratory to provide hands-on experience in domain-centric solutions as well as generic and emerging solutions for large data problems. The idea is to provide a rich set of experiences in diverse (but real-life) problem settings and the solutions that solve problems in these settings. Field experiences will also be developed with faculty from information, computer and computational sciences to provide exposure to generic and emerging tools for large data handling and analysis. Table 2 provides a list of representative field experience rotations.

Table 2 Field Experience Rotations

Domain Science	Description
Hydrology	Working with large-scale workflows and data from CUAHSI
Climate Science	Working with data sets from National Climatic Data Center (NCDC)
Astronomy	Working on data from Southern Astrophysical Research (SOAR) Telescope
Environmental Science	Working on Watershed management data
Marine Science	Working on real-time data streams
Plant Biology	Working with iPlant Collaboratory
Computational Social Science	Working with population, demographics and administrative data
Information Science	Working on collection metadata, provenance and standards, TREC test sets
Information Science	Working on data life cycle, digital forensics and curation
Information Science	Working on large NOSQL databases
Computer Science	Working on scientific computation models in cloud computing
Computer Science	Working on Cyber Security for Large-scale Data
Computer Science	Working on Graphics and Visualization mechanisms

Hands-on, participatory learning is the keystone of these field experiences. Students will participate in science and engineering projects as team members who help to identify and then implement data governance policies and procedures that are appropriate to the given project.

In consultation with their academic advisors, IGERT scholars will develop flexible individualized programs from the curriculum, accommodating a broad range of subjects and disciplinary experiences. The BDS program will apply its 'Competitive Innovation Incentive Fund' to develop integrated interdisciplinary courses by IGERT scholar teams to develop and apply large-scale data management practices and tools that will help their disciplines.

Virtual Lab and Facilities: The BDS program will provide access to large-scale data collections that are available through several NSF funded national-scale projects (including DataONE, TerraPop, Datanet Federation Consortium, Ocean Observatory Initiative, iPlant Consortium) and through National Agencies (including the NASA Center for Climate Simulation (NCCS) and NOAA National Climate Data Center (NCDC)). Through these national collaborations we will provide access to data in multiple scientific domains including Climate Sciences, Hydrology, Environmental Sciences, Marine Sciences, Plant Biology, Sociology and Cognitive Science. Computational platforms and data collections are provided through the facilities offered by the Renaissance Computing Institute (RENCI) and the Data Intensive Cyber Environments (DICE) Center.

Seminars, Conferences & Apprenticeships: The BDS program will draw on world-class researchers and information scientists and engineers to come and provide information about emerging and novel ideas in the field

of large-scale data management. Students will organize and lead a monthly seminar in which they review and critique each other's presentations and papers. Diversity of disciplinary orientation, research questions, site and situation, and skills and experience will enliven the monthly seminar and make it a critical socializing environment in which students will teach and mentor each other. Two existing forums at SILS – the weekly CRADLE seminar series and the Curation and Archives Research (CAR) group -- are examples of successful socialization environments for doctoral students. The program will also provide some travel support for students who are reporting results of their research at conferences.

The program will also provide on-site experience to students to travel to data centers and national laboratories to gain experience in practical details in handling and analyzing large data. In some cases, students may be embedded with such groups for short periods of time (with credit for field experience). We will approach NCDC, NASA/NCCS and other agencies for such field experience. We will also approach industrial partnerships to provide training. Possible companies include RTI, SAS, IBM, EMC and RedHat, all located in the Research Triangle Park.

International Training: The BDS program will also include training in international projects. Several faculty in the BDS program are involved in international research initiatives and students will directly benefit from these connections. An example is the EUDAT program to build national data infrastructure for the European Union. An exchange program will be established to enable IGERT data scientists to participate in similar data management initiatives in Europe. EUDAT has a memorandum of understanding for mutual research support with RENCI and the DICE Center.

Recruitment and Retention: Any student accepted into or enrolled in a PhD program at the UNC-CH will be eligible to apply to the IGERT training program in Data Science. Students who complete all departmental requirements for their specific PhD can also receive a certificate in Big Data. Stipend support will be limited to students within the first four years of their graduate studies. IGERT funding will be awarded for a two year period. Review of current and prospective IGERT data scientists will be conducted annually by a multidisciplinary training committee consisting of six members, and will be based on GRE scores, transcripts, departmental evaluations, trainee statements, and faculty evaluations. Committee members consider academic excellence, research productivity, project relevance, creativity, participation, and timeliness of progress toward degree to rate and rank applicants. Approved applicants from underrepresented groups will be given priority.

Minority Recruitment and Retention: Special efforts will be made to recruit underrepresented minorities. SILS will provide an advanced education path for qualified students who have completed a Masters at North Carolina Central University, a minority serving institution. Qualified students will be admitted to the PhD program at SILS. Collaborations to broaden the program impact will be initiated with universities that have existing EPSCOR programs. An example is the Clemson University EPSCOR program which is establishing an iRODS data grid to enable collaborative research for students from 28 EPSCOR states.

Outreach and Placement: The BDS program will also actively coordinate placement for its students. This will include informational advertisements, tailored campus interviews and placement apprenticeships.

Conclusion

The primary elements of the IGERT BDS program are: formal multi-departmental coursework on Big Data management, analysis, and curation; an interdisciplinary faculty-mentored research field training rotation in which students participate on science and engineering projects and manage and analyze large-scale research collections; a student-coordinated interdisciplinary seminar in which students document their work experiences; and field experience apprenticeships with national scale initiatives and data-centric industries. Participation in the program will lead to the development of domain scientists who are well-versed in handling large data collections. We believe that beyond the IGERT funding period, the BDS program will be self-sustaining. After a successful implementation, we will make all IGERT-developed curriculum materials freely available, as open courseware, for other universities to offer as data-intensive inter-disciplinary training in their doctoral programs.

The BDS team from UNC is well placed to develop a world-class Big Data program, bringing together renowned researchers and faculty from multiple departments, providing access to large-scale, multi-disciplinary data collections and hands-on training through its existing participation in collaborative national projects, and bringing experience in developing well-rounded curricula in emerging areas.

References:

1. Science. (2011). *Dealing with Data* (Vol. 331).

2. Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. *science, 331*(6018), 700-700.

3. National Science Foundation. (2011). Changing the Conduct of Science in the Information Age *Summary Report of Workshop Held on November 12, 2010*: National Science Foundation.

4. Maniyka, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute.

5. Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm : data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research.

6. NIH. (2011). Meeting the Challenge of Big Data in Biomedical and Translational Science. Retrieved from http://commonfund.nih.gov/InnovationBrainstorm/post/Meeting-the-Challenge-of-Big-Data-in-Biomedical-and-Translational-Science-(see-e2809cCross-Cutting-Issues-in-Computation-and-Informaticse2809d-in-Innovation-Brainstorm-ideas).aspx

7. White House BigData Initiative
http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf

8. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), NSF Program Solicitation, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767

9. Cyberinfrastructure Framework for 21st Century Science and Engineering, http://www.nsf.gov/cif21

Appendix 1. Agreed Faculty List (as of 5/1)

Faculty	**Specialization**
School of Information and Library Science	
Jane Greenberg	Metadata
Brad Hemminger	Database and Health Informatics
Cal Lee	Digital Preservation and Digital Forensics
Gary Marchionini	Human Computer Interaction (Dean, SILS)
Reagan Moore	Large-scale Data Management (also Director, DICE Center)
Jeff Pomerantz	Digital Libraries
Arcot Rajasekar	Large-scale Data Management
Helen Tibbo	Data Management and Curation
Department of Computer Science	
Ron Alterovitz	Robotics and Large-scale 3D-Databases
Anselmo Lastra	Computer Graphics (Chair, CS)
Ming Lin	Distributed interactive simulation and many-core computing
Fabian Monrose	Computer Security
Diane Pozefsky	Software Engineering and Ethics in Information Technology
Jan Prins	Scientific and Parallel Computing and Bio Informatics
Michael Reiter	Network Security and Cryptography
Russ Taylor	Visualization
Wei Wang	Data Mining, Databases and Bio Informatics
Renaissance Computing Institute (RENCI)	
Stanley Ahalt	High Performance Computing (Director, RENCI, also Professor, CS)
Hye-Chung Kum	Federated Data Infrastructure, Privacy, and Computational social science
Charles Schmitt	Computational Genomics
Department of Geography	
Lawrence Band	Hydrology and Remote sensing (Director, Institute for the Environments)
Odum Institute for Social Sciences	
Jon Crabtree	Social science databases
Department of Physics and Astronomy	
Gerald Cecil	Astrophysics/SOAR
Sheila Kannappan	Astrophysics
Department of Statistics and Operations Research	
Yufeng Liu	Statistical Machine Learning and Data Mining, Bioinformatics

Appendix 2: Potential External Collaborators:

Organization	**Role**
National Climatic Data Center (NCDC)	Access to data and possible field apprenticeship
NASA Center for Climate Simulation (NCCS)	Access to data, analysis and possible field apprenticeship
iPlant Collaborative	Access to data and analysis software and possible apprenticeship
Ocean Observatory Initiative (OOI)	Access to real-time data and applications
CUAHSI	Access to hydrology data
DataONE	Access to earth observational data
TerraPopulus	Access to population and environmental data
EMC	Possible field apprenticeship at their data centers
RTI	Possible field apprenticeship
IBM	Possible field apprenticeship
SAS	Possible field apprenticeship
Oracle	Possible field apprenticeship
Microsoft	Possible field apprenticeship

Provisional Budget:

	Year 1	Year 2	Year 3	Year 4	Year 5	Total
PARTICIPANT SUPPORT:						
Scholarships	300,000	300,000	300,000	300,000	300,000	1,500,000
Tuition and Fees	105,000	105,000	105,000	105,000	105,000	525,000
Research Innovation Incentive funds	40,000	40,000	40,000	40,000	40,000	200,000
International Collaboration funds	40,000	40,000	40,000	40,000	40,000	200,000
OTHER PROJECT COSTS:						
Salaries & Fringes	80,000	80,000	80,000	80,000	80,000	400,000
Seminars & Conferences	5,000	5,000	5,000	5,000	5,000	25,000
Apprenticeships	27,000	27,000	27,000	27,000	27,000	135,000
Minority Recruitment	30,000	30,000	30,000	30,000	30,000	150,000
Evaluation	3,000	3,000	3,000	3,000	3,000	15,000
Total	**630,000**	**630,000**	**630,000**	**630,000**	**630,000**	**3,150,000**
Indirect (48%)	69,600	69,600	69,600	69,600	69,600	348,000
Total Costs	**699,600**	**699,600**	**699,600**	**699,600**	**699,600**	**3,498,000**

Advocates:

1. Javed Mostafa
2. Ruth Marinshaw
3. Sarah Michalak