Data@Carolina

Workshop 2015 Report

Kevin Jeffay
Chair, Computer Science

Stan Ahalt
Director, RENCI

Thomas Carsey
Director, Odum Institute

The University of North Carolina Department of Computer Science

renci
RESEARCH ⟍ ENGAGEMENT ⟍ INNOVATION

UNC THE ODUM INSTITUTE

The Data@Carolina Workshop 2015, led by RENCI, the Howard W. Odum Institute for Research in Social Science, and the Department of Computer Science, was held on September 25, 2015 at RENCI's Europa Drive campus at the University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

This report was prepared by Stan Ahalt, director of RENCI, Tom Carsey, a political scientist and director of the Odum Institute, and Kevin Jeffay, chair of the computer science department.

Contents

Executive Summary

We are awash in data streaming from our computers, phones, cars, home appliances, portable medical devices, social media outlets, workplaces, and countless other sources. In an information economy, data serves as the new currency. U.S. industries are moving to capitalize on the "datification" of society, but full realization of the potential economic, health, and scientific benefits of today's information-rich and data-driven society requires a highly skilled workforce complimented by broad data literacy across all corners of society.

In the academy, "big data" is the new frontier in interdisciplinary, collaborative research. The centrality of data is sparking debate about what it means to be an educated person in the early 21st Century.

> ***Academic institutions across the country are making huge investments in data-related research and education in order to meet these growing needs. Carolina must do the same – starting right now.***

To initiate a cross-campus dialogue on this pressing issue, the Data@Carolina Working Group organized and hosted "Data@Carolina Workshop 2015", sponsored by RENCI and the Howard W. Odum Institute for Research in Social Science and held on September 25, 2015.

The workshop was attended by 53 faculty and staff members drawn from numerous academic units, including the College of Arts & Sciences, School of Information and Library Science, Kenan-Flagler Business School, School of Social Work, Eshelman School of Pharmacy, School of Public Health, School of Dentistry, UNC Libraries, the Graduate School, and several centers and institutes, including RENCI, Odum Institute, Carolina Population Center, Cecil G. Sheps Center for Health Services Research, and Center for Galapagos Studies. The workshop included eleven brief (7-minute) presentations by attendees on current data initiatives at Carolina. The workshop also included three breakout sessions during which participants discussed needs and opportunities for research, graduate education, and undergraduate education at Carolina. A final large group discussion included reports from each breakout session and closing remarks by the working group.

Workshop participants identified several opportunities for data-related research, graduate education, and undergraduate education at Carolina.

- **Opportunities for Research** include expanding and staffing data-intensive capabilities for: secure, policy-based data sharing, computing, and storage capabilities at Carolina; technical networking capabilities to identify and access both local and remote data repositories; and social networking capabilities to foster interdisciplinary, team-based research and shared resources.
- **Opportunities for Graduate Education** include the establishment of: a fast-track Professional Science Master's (PSM) program in Data Science; one or more graduate certificate programs in Data Science/Data Studies; a core curriculum for graduate students in data-intensive and -enabled fields; innovative modular courses and workshops to provide educational opportunities outside the traditional semester course format.
- **Opportunities for Undergraduate Education** include the implementation of: a data literacy program adaptable to every major on campus; an undergraduate minor in Data Science/Data Studies; course development grants and other support to launch such courses, possibly in modular or team-teaching formats.

The Data@Carolina workshop revealed substantial interest and urgency among participants. Much remains to be done, but two point of consensus emerged from the workshop: 1) Carolina should think big as we launch our efforts in data, and 2) There are several concrete actions we can and should take immediately. The Data@Carolina Working Group expects to build on this enthusiasm as it continues to work with administration, faculty, and staff from across campus to establish Carolina as an international leader in data-related research, graduate education, and undergraduate education.

1. Overview

The "datafication" of society has left businesses, health care providers, governments, and universities awash in data – a reality that will only continue to grow. For example, there are now more objects and devices connected to the Internet than people.[1] Current estimates indicate that today's 25 billion Internet-connected devices will double to 50 billion by 2020 (i.e., >6 devices/person worldwide). U.S. industries are quickly realizing the value of these data sources. For example, a 2015 McKinsey report projects up to $11.1 trillion in 2025 economic yield from so-called "Internet of Things" applications, including individuals, homes, offices, worksites, retail environments, hospitals and health care clinics, factories, vehicles, cities, and the outside (non-urban) environments.[2]

However, society remains under-equipped to realize the full economic, health, and scientific benefits of this explosion of data. Gartner, for instance, estimates that as many as 2/3 of all available jobs in information technology will not be filled because of a lack of talent.[3] McKinsey projects that by 2018, the U.S. will have a shortfall of 140,000 to 190,000 "deep" analysts and an additional shortfall of 1.5 million general analysts and managers with analytical expertise.[4] While critical, these estimates miss a broader point – the "datafication" of society affects every corner of the economy and society, meaning that everyone needs to become data literate.

Academic institutions across the country are making huge investments in data-related research and education. For example, Georgia Tech is investing ~$300 million in data science. The University of California at San Francisco recently established the Institute for Computational Health Sciences. The University of Michigan just announced plans to spend $100 million on data science, with plans to hire 35 new faculty members. Ohio State recently launched a plan to hire 100 faculty members as part of an interdisciplinary Big Data initiative. In North Carolina, NCSU partnered with SAS in 2007 to establish a highly innovative, successful, and self-sustaining 10-month Master of Data Analytics program. The program accepts only 1 in 8 applicants but boasts a 100% job placement rate at graduation. UNC Charlotte offers a Professional Science Master's degree and a graduate certificate in Data Science and Business Analytics through its Data Science Initiative.

Clearly, Carolina is falling behind the country in research and education in Big Data.

To initiate a cross-campus dialogue on this pressing issue, the Data@Carolina Working Group organized and hosted "Data@Carolina Workshop 2015". This report provides an overview of the workshop proceedings, and the discussions and consensus reached by participants on Big Data needs in research, graduate education, and undergraduate education at Carolina.

---

[1] Evans, D., on behalf of Cisco Internet Business Solutions Group (IBSG). (April 2011). The Internet of Things. How the Next Evolution of the Internet is Changing Everything. http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf.

[2] Manyika, J., Chu, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J., & Aharon, D. (2015). The Internet of Things: mapping the value beyond the hype. *McKinsey Global Institute Report.* McKinsey & Company. http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Business%20Technology/Unlocking%20the%20potential%20of%20the%20Internet%20of%20Things/Unlocking_the_potential_of_the_Internet_of_Things_Executive_summary.ashx.

[3] http://www.gartner.com/newsroom/id/2207915.

[4] Manyika, 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 Report.* McKinsey & Company. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

2. Summary of Workshop Proceedings

The Data@Carolina Workshop 2015 was led by RENCI, the Howard W. Odum Institute for Research in Social Science, and the Department of Computer Science. A Data@Carolina Workshop 2015 Website (http://data.web.unc.edu/) was created to inform the broader Carolina campus about the event. Email invitations were distributed by the working group to nearly 100 faculty and staff members at each major academic unit at Carolina. The final participant list included 53 persons with representation from the College of Arts & Sciences, School of Information and Library Science, Kenan-Flagler Business School, School of Social Work, Eshelman School of Pharmacy, School of Public Health, School of Dentistry, UNC Libraries, the Graduate School and several centers and institutes, including RENCI, Odum Institute, Carolina Population Center, Cecil G. Sheps Center for Health Services Research, and Center for Galapagos Studies (see Appendix A for a list of participants).

The complete workshop agenda can be found in Appendix B. The workshop began with a welcome breakfast and registration and introductions by select members of the Data@Carolina Working Group. Stanley Ahalt, Kevin Jeffay, and Tom Carsey then provided an overview of the genesis of the workshop. The group was asked to consider two questions over the course of the workshop:

(1) What is your vision of a major "data X" initiative for the campus?
(2) What would it take for Carolina to become a world leader in data X in your area/arena?

Eleven speakers, selected by the working group then provided brief (7-minute) presentations on ongoing and planned Big Data–related research and education programs. Presentations came from all corners of the campus and were focused on sharing information and identifying future directions. (All presentation slides are posted at https://data.web.unc.edu/datacarolina-workshop-2015/agenda/) The workshop also featured three 1-hour breakout sessions on research, graduate education, and undergraduate education; these were held simultaneously and led by facilitators. Workshop participants self-selected one breakout session to attend. The main charge of participants was to discuss current data-related needs and opportunities in each of the three areas. A final plenary session provided an opportunity to report back on the breakout session discussions. The workshop closed with final remarks and a luncheon.

3. Big Data Needs in Research, Graduate Education, and Undergraduate Education at Carolina

Research

Workshop participants agreed that Carolina has many highly successful research programs that involve data, including those discussed above. However, workshop participants concurred that Carolina must improve its research cyberinfrastructure, build a pool of data-related research staff, and further stimulate a culture of communication between units in order for faculty members to maintain leadership in their fields, remain competitive in efforts to obtain external research funding, and be successful in recruiting and retaining top-notch graduate students. Specific challenges include the need for:

- secure, policy-based data sharing, computing, and storage capabilities at Carolina available to both campus researchers and their collaborators across the globe;
- a pool of data-related research staff and expertise that can be flexibly allocated to emergent projects;
- technical networking capabilities to identify and access data repositories both on campus and elsewhere; and
- social networking capabilities to foster interdisciplinary, team-based research and shared resources.

Workshop participants recognized that there are existing solutions to these challenges at Carolina, but they acknowledged that current solutions are insufficient to position Carolina as a leader in Big Data research. Participants also acknowledged that solutions to these challenges may be best provided at Carolina or may involve new partnerships with commercial providers and outsourced human resources, especially for projects that require significant upfront investment but little downstream time or money. Participants agreed that incentives for interdisciplinary, team-based research and shared resources are lacking and could be fostered through the development of "learning communities" comprised of established and junior faculty members, as well as students, from across campus, all with shared data needs. Participants also discussed the possibility of creating a comprehensive domain-agnostic "data engine" to foster the movement of data from knowledge to wisdom by making better use of both researchers and research data.

Graduate Education

Workshop participants assented that Carolina offers numerous data-related graduate training opportunities and programs. Nonetheless, participants also felt strongly that Carolina is behind other academic institutions in offering graduate and professional training specifically focused on Big Data and intended to fill current industry needs for a highly skilled data-literate workforce. Specific mechanisms to address this need include:

- a fast-track Professional Science Master's (PSM) program in Data Science;
- one or more graduate certificate programs in Data Science/Data Studies; and
- a core curriculum for graduate students in data-intensive and data-enabled fields.

Each program would be flexible in allowing units at Carolina to tailor the program to their students' needs. All programs would feature interdisciplinary, team-based training with a core curriculum featuring key data-related topics (i.e., data, storage-curation, ethics, privacy, analysis, computation, visualization), specialty training specific to a graduate student's domain or a professional's job demands, and training in professional skills. Where appropriate, these courses should be modular, as opposed to semester-based. This will facilitate program flexibility and enable students to acquire a broader skill set. Academic-industry partnerships will be essential, particularly for the PSM program, and a tuition-based revenue model would ensure sustainability. All efforts to promote new data-intensive graduate programs should be guided by a well-balanced committee, with representation from across academic units at Carolina.

Undergraduate Education

Workshop participants largely agreed that Carolina's current undergraduate programs are insufficient in addressing the need for data literacy among undergraduates in order to equip them with the skills required

to be competitive in today's marketplace. Specific mechanisms to drive data literacy at Carolina and educate a competitive workforce could include:

- two required undergraduate courses to enable data literacy, one general and one specialized to a student's major;
- a two-year data literacy program; and
- an undergraduate minor in Data Science/Data Studies.

New undergraduate data literacy programs could leverage existing resources across campus. Data-intensive undergraduate courses and programs could be launched relatively quickly, with little cost or need for a sustainability plan, although workshop participants recognized that a change in undergraduate requirements would require broad support from all 16 North Carolina schools within the UNC system.

4. Concluding Summary & Remarks

The current and future role of data in our society is undeniable. To remain a world-class public university requires that Carolina develop a world-class response. We must seize the opportunity and put a uniquely Carolina focus on our efforts, potentially using the twin goals of achieving data literacy for every Carolina graduate and leveraging data for the public good. Other academic institutions across the country are making huge investments in data-intensive research and education. Carolina must do the same – starting right now.

Acknowledgements

The Data@Carolina Working Group extends appreciation to the following persons for organizing the workshop, managing all workshop logistics, taking notes during the workshop, and helping write this report: Dawn Carsey, Project Coordinator, NCDS and iRODS Consortium; Karamarie Fecho, Ph.D., Scientific Writer; Annie Goessling, RENCI Intern; and Asia Mieczkowska, Project Coordinator, RENCI.

Appendix A: Workshop Participants

Name	Department/Center/Institute/School
David Adalsteinsson	Mathematics
Stanley Ahalt	RENCI
Jay Aikat	Computer Science/RENCI
Deb Aikat	School of Media and Journalism
Roger Akers	Sheps Center
Robert Allen	American Studies/Digital Innovation Lab
Hyowon An	STOR/A&S
Dan Anderson	CDHI
Ruth Anderson	Nursing
Sridhar Balasubramanian	Kenan-Flagler Business School
Brian Blanton	RENCI
William Bosley	Digital Innovation Lab
Moorek Bryant	UNC Chapel Hill
Amarjit Budhiraja	STOR/A&S
Neal Caren	Sociology
Thomas Carsey	Odum Institute
Dawn Carsey	RENCI
Judith Cone	InnovateCarolina
Jennifer Conrad	Kenan-Flagler business School
Patrick Cownay	Economics, CAS
Patrick Curran	Psychology and Neuroscience
Nancy Dole	Carolina Population Center
Greg Forest	Mathematics/BME/APSc
Mark Fraser	School of Social Work
Stephen Gent	Political Science
Ann Goessling	Renci Intern

Name	Department/Center/Institute/School
David Gotz	SILS
Heidi Harkins	The Graduate School
Michele Hayslett	University Libraries
Carol Hunter	University Library
Kevin Jeffay	Computer Science
Rebecca Kitzmiller	UNC School of Nursing
Ashok Krishnamurthy	RENCI
Gary Marchionini	SILS
J. S. (Steve) Marron	STOR/A&S
Rich McLaughlin	Mathematics
Jacqui McLaughlin	UNC Eshelman School of Pharmacy
Sarah Michalak	University Library
Asia Mieczkowska	RENCI
Piotr Mieczkowski	Genetics
Danianne Mizzy	Kenan Science Library
Peter Mucha	Mathematics
Philip Page	Center for Galapagos Studies
Diane Pozefsky	computer science
Jan Prins	Computer Science
Arcot Rajasekar	SILS/UNC
David Stotts	Computer Science
Helen Tibbo	SILS
Alex Tropsha	Pharmacy
Todd Vision	Biology
Steve Walsh	Center for Galapagos Studies
Barbara Wildemuth	SILS
Di Wu	School of Dentistry
Donglin Zeng	Biostatistics/Public Health

Appendix B: Workshop Agenda

Friday, September 25, 2015

8:00 – 8:20 Breakfast and registration

8:20 – 8:40 Introduction

8:40 – 10:10 Talks (7 mins each) by representatives of various campus units

- *Stan Ahalt (Renaissance Computing Institute)*
- *Roger Akers (Cecil G. Sheps Center for Health Services Research)*
- *Bobby Allen & Will Bosley (Digital Innovation Lab)*
- *Dan Anderson (Carolina Digital Humanities Initiative)*
- *Sridhar Balasubramanian (Kenan-Flagler Business School)*
- *Tom Carsey (Odum Institute)*
- *Patrick Curran (Psychology & Neuroscience)*
- *Steve Marron (Statistics & Operation Research)*
- *Alex Tropsha (Eshelman School of Pharmacy)*
- *Barbara Wildemuth (School of Information and Library Science)*
- *Donglin Zeng (School of Public Health)*

10:10 – 10:25 Intro by the three session leaders – i.e. graduate education, undergraduate education, and research

10:30 – 11:30 Breakout sessions for graduate education, undergraduate education, and research

11:30 – 12:00 Plenary – Report back (10 mins each) from the three breakout groups

12:00 – 1:00 Lunch