GradEdge

COUNCIL OF GRADUATE SCHOOLS

Insights and Research on Graduate Education

Volume 4, Number 9 • November 2015

Big Questions about “Big Data”: Perspectives from the 2015 Global Summit on Graduate Education

Julia Kent, Assistant Vice President, Communications, Advancement and Best Practices

"Big Data" has been broadly defined as "the collection, aggregation...and analysis of vast amounts of increasingly granular data” (Cate, 2014). Contemporary debates about big data have raised both interest and concern in the graduate community.

On the one hand, graduate leaders are accustomed to using data to inform decision- making and have expressed curiosity about the potential of big data experiments in graduate education, such as the collection of data on student learning in large online courses. On the other hand, big data have been associated with a number of problems that directly concern graduate leaders, posing a number of challenges and questions:

- How should large amounts of data be managed and stored?
- What methods should be used for analysis and interpretation?
- How do we think about informed consent and privacy rights in a big data context?
- How should we be preparing the next generation of graduate degree recipients to manage the world of big data?

CGS set out to answer these questions at the Ninth Annual Strategic Leaders Global Summit, this year a collaboration with the National University of Singapore (NUS), a CGS international member. With sponsorship support from Educational Testing Service (ETS) and ProQuest, we convened graduate deans and individuals with similar university roles from 14 countries, coming together on the NUS campus from September 27 to 29.

The topic for this year’s summit was particularly well-suited to an international meeting. The scientific trends driving big data are typically global phenomena, supported by advancements in research and development that have broad implications for international research networks both inside and outside universities. Yet the laws that govern the collection and use of data vary widely by country, creating new risks as datasets are merged, exchanged and used across national boundaries.

National Perspectives on the Benefits and Challenges of Big Data

The two opening sessions of the Global Summit provided broad frameworks for understanding these differences in national and institutional contexts. One major theme of the discussion was how to define big data, a concept that was first used by computer scientists but which has now been adopted for use (and potential misuse) in everyday language. A definition cited by many participants is based on three "V's": Volume, Velocity and Variety; *Volume* describes the growing amounts of data collected, *Velocity* the speeds at which they are collected, and *Variety* the diversity of their sources.1 This definition complicates our impulse to think of "big" datasets in terms of size alone. Mohan Kankanhalli of National University of Singapore put the problem this way: "The amount of data doesn't matter. A lot of graduate student data is network data, which creates interconnection issues." Bernadette Franco of the University of São Paolo gave further weight to this idea when she observed that it is the variety of data, not the volume that presents challenges to her university. Finding appropriate ways to analyze and synthesize data that are generated from different sources was repeatedly cited as a challenge.

A second and related theme concerned what Hans Bungartz of Technische Universität München called a "missing tradition of data support management." Bungartz, a computer scientist and graduate dean, noted that universities in particular often lack this capacity, and observed that graduate institutions will require structural changes in technology research and administration to manage data sets that are growing larger and more complex. Barbara Knuth of Cornell University echoed this concern, noting that universities often struggle with a lack of access to staff with the ability to analyze large data sets or the infrastructure needed to support this work.

Finally, participants highlighted a common goal for international universities— the use of big data to "personalize" graduate education. Presentations offered several striking examples of this trend: individualizing a student's online curriculum based on his or her interests and strengths, tracking a student's participation in various elective activities in order to better understand individual educational outcomes, and using data from multiple sources to identify students in academic trouble.

Alongside voices of optimism about using big data to individualize education, we also heard notes of caution, including the view that big data could make education less personalized and creative. As Bungartz observed, "Any kind of data analytics has a tendency to look at the overall experience of many, not just the individual case." Big data present opportunities to notice patterns in student experiences, in other words, but they also challenge us to identify the right patterns and to interpret them in meaningful ways.

Weighing the Costs: Resources and Ethical Issues

Summit participants delved deeper into the issues surrounding big data in conversations about two key topics: the resources required to collect, analyze and store large and complex datasets, and the legal and ethical issues raised by big data.

In discussions about resources, we heard about a number of complex and sophisticated systems developed to synchronize different data systems on campuses. For example, Karen Butler-Purry of Texas A&M University described the development of a new graduate student portal that will accumulate a variety of student data, drawing from student input, the university student information system and graduate student ORCID records. The tool will improve efficiency in graduate school processes by allowing online processing of graduate academic requirements. A number of other institutions mentioned that they outsource some of their data analysis to companies such as Academic Analytics, a provider of business intelligence data for research institutions.

Unsurprisingly, universities are often called upon to weigh investments in such systems against the potential benefits in terms of saved time and resources. Complex tools for data collection and processing may also be out of the reach of smaller graduate institutions, as was pointed out by Kevin Vessey of Saint Mary's University and Magnús Lyngdal Magnússon of the University of Iceland. Vessey noted that small institutions may not be able to afford software to provide analytics in areas such as recruitment, student monitoring and advising, and analysis of research performance.

Compared with resource issues, legal and ethical issues raised by big data present even greater limitations and risks to universities. One reason that big data are associated with thorny ethical problems, those concerning privacy in particular, is that they may involve the merging of datasets that were collected under different privacy protections. Some of the broad legal and ethical issues raised by participants included the legal collection of data, anonymization and encryption, secure storage, controlling access, data leakage and Intellectual Property (IP). Participants agreed that universities must remain aware of these issues as they collect, share and analyze data, and they must also prepare their current graduate students to navigate them.

Enhancing Learning and Student Success

Summit presentations offered a number of important insights from the growing body of research and data on student learning processes. Martin Gersch of Freie Universität Berlin pointed out that data-based learning analytics provide ways to collect and interpret data in online education, including network analysis and web and text mining. A pertinent example was offered by Y. Narahari of the Indian Institute of Technology, Bangalore, who outlined an experiment in which data from online courses were analyzed to determine how students respond to various incentives for course participation. Such developments are based on increasingly sophisticated analytical models. As David Payne of Educational Testing Service explained in a paper on big data and learning assessment, the emerging field of computational psychometrics merges data mining methods and machine learning algorithms with psychometric models; the result is more individualized learning experiences.

Preparing the Next Generation of Experts

Many summit participants indicated a lack of data scientists in their national contexts, raising the question of what graduate institutions need to do to help societies and economies meet this workforce need. But another issue raised in the discussion was the preparation needed for the majority of graduate students who will work in other fields. Not every graduate degree holder needs to be an "expert," participants agreed, but all will be impacted by, and need some degree of, big data literacy.

Two concrete ideas for preparing master's and doctoral students in this area emerged. First, many noted that big data approaches to research are also interdisciplinary or multidisciplinary, and recommended that institutions integrate lessons on big data into existing interdisciplinary learning opportunities. Shiyi Chen of the South University of Science and Technology of China explained that Peking University already offers such opportunities in a program that draws from the fields of mathematics, statistics, computer science, sociology and biomedical informatics.

Second, there is a strong need to fill existing gaps in graduate student training in the Responsible Conduct of Research (RCR) and research ethics. CGS President Suzanne Ortega identified three potential areas where curricula might be examined and strengthened: 1) shifting definitions of informed consent in big data contexts; 2) the ethical implications of predictive analytics; and 3) understanding one's responsibilities for data management and curation in a world of open access.

A challenge that lingers is that we don't necessarily have a clear picture of the big data contexts that students are being prepared to navigate. Graduate programs will need to remain flexible and attune to the skills and knowledge that their students will need as trends in data collection and analysis evolve.

Research Collaboration and Productivity

Throughout the summit, we heard ambitious big data research projects and collaborations that have emerged in recent years. These examples prompted reflections about how universities, the commercial sector and other entities might better support the development of researchers and research.

Both Paul Burnett of Queensland University of Technology and Niels Dam of ProQuest pointed out that respecting IP rights is becoming an important mandate for universities and companies as research becomes more collaborative. Also posing challenges are unprotected data environments such as social media, where individuals may make publicly available their data without formally consenting to their mining and use. In the wake of these developments, it is all the more important for universities to develop clear policies on the collection, storage and management of data by their researchers, including graduate students.

Next Steps

At the conclusion of the meeting, summit participants developed “A Proposal for Further Action” designed to help graduate education leaders better understand and manage big data issues. These recommended actions are intended to serve as a menu of options for graduate institutions, government agencies, non- profit, and commercial actors seeking to better prepare institutions and their students for big data concerns. For each proposed action, potential actors and collaborators are indicated.

Some of these actions are considered particularly high priority, and are listed below:

1. Develop a process to generate a data dictionary of key graduate outcomes and metrics useful for purposes of national and international benchmarking. This infrastructure will be needed to generate truly big data on graduate education.
    i. *Associations and consortia of graduate institutions.*
2. Promote the development and sharing of open-access analytic software tools customized for use in the administration of graduate schools. These would include standard formats for integrating data from applications, student experiences and milestones, etc. for institutional and cross- institutional comparison.
    i. *Associations and consortia of graduate institutions; non-profit and commercial partners.*
3. Develop best practices and case studies of 'big data' education across disciplines. These would address requirements for infrastructure support, ethics training, and thesis supervision.
    i. *Associations and consortia of graduate institutions.*
4. In addition to technical and statistical skills associated with big data analytics, identify what other skills and knowledge students will need to succeed in a world of big data. Determine how graduate programs can provide this preparation.
    i. *Associations and consortia of graduate institutions; individual graduate institutions.*
5. Consider whether responsible conduct of research training is sufficient to address issues related to big data use by graduate students in their research. Give particular attention to privacy issues, legal issues, and to special challenges in interpreting large data sets.
    i. *Associations and consortia of graduate institutions; individual graduate institutions.*

In the full document, additional actions are listed in the areas of improving data-based decision-making, preparing the next generation of experts, ethical issues, and supporting research using big data. Many of these are geared toward individual universities seeking ways to address gaps in preparation to manage big data issues.

As in past years, CGS will ensure that the insights developed at the Global Summit are shared broadly with the entire CGS membership. Before the end of the year, we will publish the online proceedings of the event, making available the brief papers presented by the 32 summit participants. These papers include summaries of current big data issues experienced by universities around the world, including the US and Canada, and describe resources developed by CGS member institutions. In addition to accessing these electronic proceedings and sharing them with colleagues on campus, we hope that CGS members will join us for a special session devoted to the summit's outcomes, *Implications of Big Data for Graduate Education: A Global Conversation*, at the 2015 CGS Annual Meeting in Seattle.

Endnote

[1] Lane and Finsel's introductory chapter to *Building a Smarter University* provides a concise overview of the "V's" often used to characterize big data and its outputs. See pages 6-8.

References

Cate, F.H. (14 November 2014). The big data debate, *Science* 346(6211), 818.

Lane, J.E. and Finsel, B.A. (2014). Fostering smarter colleges and universities: Data, big data and analytics. In Lane, J.E., B*uilding a smarter university: Big data, innovation and analytics*. Albany: State University of New York Press.