View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Biodiversity Informatics

Developing Data Science at UNL: Progress, Challenges, and Opportunities for Research

Jennifer L. Clarke, PhD, University of Nebraska

Over the past several years we have seen a groundswell of interest and investment in what is commonly referred to as `data science'. As a statistician, I am biased to consider data science as simply what statisticians have been doing for decades. However, I have come to appreciate that data science can be a broader and more encompassing endeavor, one that encourages interdisciplinary research.

I was hired in 2013 by the University of Nebraska-Lincoln as a faculty member in the Department of Statistics and the Department of Food Science and Technology. My primary role on campus is as Director of the Quantitative Life Sciences Initiative (QLSI), a program of excellence whose mission is to develop expertise and resources in data science and `Big Data' for disciplines in the Life Sciences. I advocate for resources and expertise related to turning data into knowledge, e.g., develop graduate and undergraduate curricula on data topics in the life sciences, serve as a liaison between UNL and stakeholders with interests in Big Data, and enable research in data and the life sciences. I report to the Dean of Agricultural Research within the Institute for Agriculture and Natural Resources, the UNL Vice Chancellor for Research and Economic Development, and to the QLSI faculty advisory committee.

Why QLSI? Over the past 20 years, and certainly over the last few years, the ways in which we analyze, process, store and interact with data have been rapidly changing (and there is no indication that this process is slowing). The pace of change can be quickly exemplified by a quick look at the types of media and communication devices we use today (MP3 players, BlueRay discs, smart phones) compared to a few decades ago (VHS tapes, floppy discs, answering machines)[see Figure 1]. Advances in computing have brought us the era of `Big Data', a term with different meanings to different constituencies. A good working definition, albeit relative to each individual, is more data than one is accustomed to or more than one can manage. Experts continue to discuss what aspects of data define `Big Data' [see Figure 2]; four common attributes of Big Data are

- *Volume* or scale of data, e.g., in petabytes or exabytes
- *Velocity* or speed of data, e.g., streaming data from sensors
- *Variety* or different types of data, e.g., text and images and GPS-tagged locations, and
- *Veracity* or level of uncertainty, e.g., missing or inaccurate data.

The last attribute, veracity, applies to any type of data and hence is not exclusive to Big Data (see [1] for an ongoing discussion of Big Data). However, it is an important attribute to keep in mind as a reminder that the amount of useful information in data may not scale with data volume. The discussions around Big Data are happen-

ing now because (1) academic disciplines are becoming more quantitative; (2) data collection is becoming easier and less expensive; and (3) there is enough available computing power to analyze larger amounts of data than has previously been possible [2].

This brings us to one of the challenges of 21st century science: How to get from data to information to knowledge when data are large, noisy, and complex. The data-to-knowledge process requires a diverse skill set that draws upon expertise from multiple disciplines. For example, a key societal challenge is feeding a growing global population in a manner that is resource efficient and environmentally sustainable. The development of such a process will involve improvements in weather prediction, farm management practices, plant and animal breeding, and food storage and transportation, as well as reductions in food waste. Each of these improvements can only be achieved with the collection and analysis of data by scientists with domain knowledge as well as advanced data management, analysis, and communication skills. This combination of skills is relatively unusual and requires considerable education and training to achieve.

We need to rethink undergraduate, graduate, and continuing education if the academic community is going to fulfill the national workforce needs in data science. When I arrived at UNL I spent a lot of time learning about the campus, identifying a set of initial goals and plans for evaluation for QLSI, building connections with external partners, and developing buy-in among faculty and administrators. This process revealed several opportunities for the development of data science for the life sciences that would benefit both the campus and its stakeholders. One idea that I pursued in Spring of 2014 was an undergraduate major in data science/informatics that would provide students with a coherent set of curricula in the information sciences. Although this idea garnered strong support from several constituencies on campus, the academic administration decided that the university should increase undergraduate enrollment in order to accommodate an additional major.

We decided to develop an interdisciplinary doctoral program in Complex Biosystems [3]. This program was primarily driven by junior faculty from three separate colleges whose research programs required access to graduate students with training in the quantitative life sciences. All students participate in an initial year of core training before selecting advisors and a program specialization; the current specializations are microbial interactions, integrated plant sciences, systems analysis, pathobiology and biomedical sciences, and computational organismal biology, ecology, and evolution (COBEE). Qualified faculty can participate in one or more specializations. We also co-host a graduate student recruitment event each year with the Office of Graduate Studies and existing graduate programs in the life sciences. This event is very popular and has increased both program awareness and recruitment success rates.

QLSI has active research and/or educational partnerships with local, regional, national, and international organizations. These include the Midwest Big Data Hub (midwestbigdatahub.org/), the North American Plant Phenotyping Network (http://nappn.plant-phenotyping.org/),

the Nebraska Food for Health Center (http://foodforhealth.unl.edu/), the Fraunhofer Institute for Integrated Circuits (https://www.iis.fraunhofer.de/en.html), the Great Plains Network (https://www.greatplains.net/), and CyVerse (http://www.cyverse.org/). These partnerships are critical to the success of the Initiative for several reasons. First, data science is a rapidly evolving discipline and partnerships are an effective way to become aware of the latest developments. Second, these partnerships provide opportunities for cutting-edge graduate training experiences. Finally, the research reputation of Nebraska and the UNL research funding portfolio both benefit from such collaborations.

A recent area of emphasis for QLSI is reproducible research and Big Data management and analysis. We have partnered with UNL Libraries and Office of Graduate Studies to support and promote the use of ORCiD (https://orcid.org/) and common metadata standards. We also encourage the use of shared research infrastructure such as NSF XSEDE, the Open Science Grid (OSG), Galaxy (https://galaxyproject.org/), and CyVerse [see Figure 3]. These activities are of particular interest to faculty associated with federally supported research centers who are obligated to comply with federal data sharing standards and expectations. Sharing and hosting large amounts of research data can be both time consuming and costly, while universities that receive public research funding have an obligation to conduct `open science' and share their research products. How to finance the maintenance and effective sharing of data in the era of Big Data and the Internet of Things (IoT) remains an open challenge [4], and one we must surmount if we are to remain the stewards of research for public benefit.

Fig. 1. A graphic example of how our relationships with data and modes of communication have changed rapidly over the past few decades with advances in computing and information technology

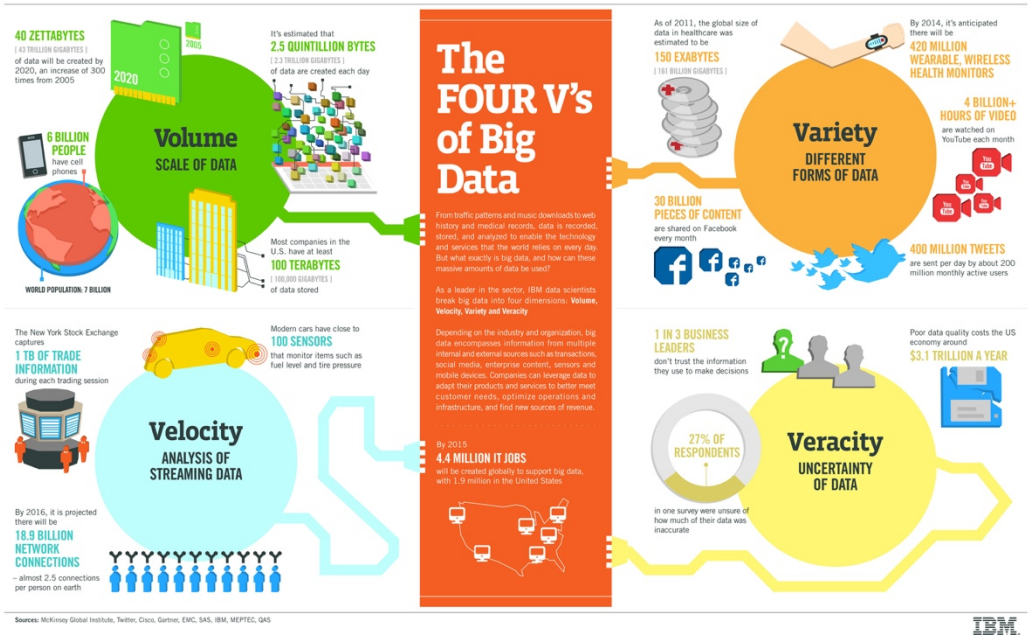

Fig. 2. The Four V's of Big Data as an infographic from IBM. The debate over the 'V's of Big Data continues, with some favoring only 3 'V's (without veracity) and others advocating for 5 'V's (including value).

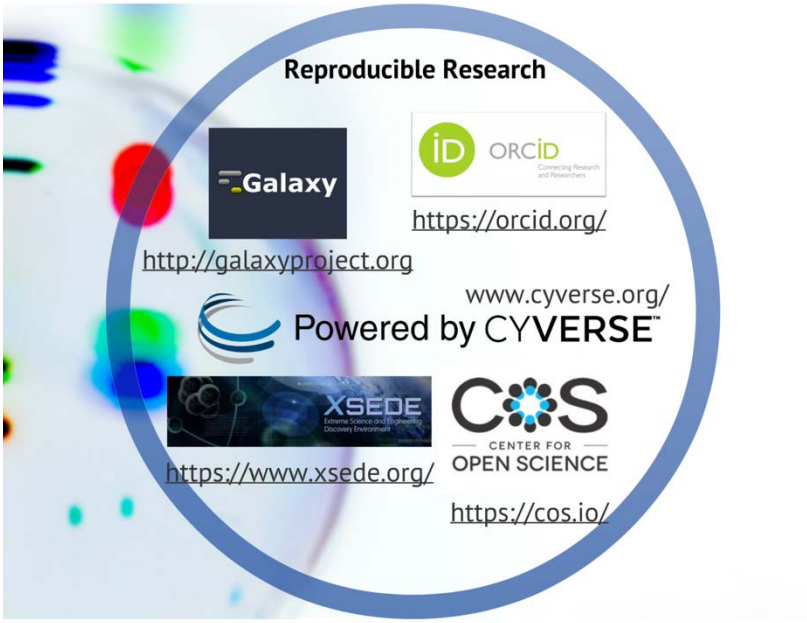

Fig. 3. Several examples of resources that enable reproducible research. These include tools for researcher disambiguation, data analysis, distributed computing, and open science.

Works Cited

1. Laney, Doug. Batman on Big Data. Post to the Gartner Blog Network, November 13, 2013. http://blogs.gartner.com/doug-laney/batman-on-big-data/
2. King, Gary. Big Data is not about the data! Presentation at Shanghai Jiao Tong University, January 4, 2017. https://gking.harvard.edu
3. Schrage, Scott. New doctoral program links life sciences with big data. Nebraska Today, October 28, 2016. http://news.unl.edu/newsrooms/today/article/new-doctoral-program-links-life-sciences-with-big-data/
4. CDWVoice. The future is data-driven, but IoT has its challenges. Forbes BrandVoice, May 2, 2017. https://www.forbes.com/sites/cdw/2017/05/02/the-future-is-data-driven-but-iot-has-its-challenges/#52abf1511459