If a Tree Fell in a 300 Million-year Old Forest, Did it Leave a Data Trail? [1]

Joseph A. Heppert, Ph.D., Vice President for Research
Texas Tech University

You have just checked into a deluxe ski resort. You grab your skis, take several lifts to get to the top of the mountain, and arrive at an area with a breathtaking view, but startlingly few staff to direct you to the runs. In fact, you are concerned that none of the slopes seem to be formally marked. Nevertheless, deducing that there must be some way off the mountain, you find a slope that shows signs of use, seems free of debris, and is about your skill level. You start down the trail. Switching to another trail, you encounter other skiers and, feeling more at ease, make it to the bottom of the second run. There, you happen across an individual who, based on their clothing, seems to be part of the ski patrol. You politely inquire about the state of the signage on the slopes. "Oh," the individual remarks in an off-hand manner, "We've never bothered to mark or groom the slopes at this resort." As this response is sinking in, you hear and feel a low frequency rumbling. You look up and see a 20-foot wall of snow hurtling down the mountain in your direction…

Of course, this is an allegory, not something that would actually occur in today's ski industry. But it is a fair description of a situation unfolding today in aspects of data management within numerous U.S. research universities.

Researchers in many different fields—genomics, bioinformatics, climate science, fluid dynamics, economics and marketing, etc.-- are creating masses of data at a rate unprecedented in the history of the age of scientific discovery. It is estimated that the entirety of the dataverse reached 1.8 zettabytes (1 zettabyte = 1 trillion gigabytes) in 2011[2]. That nearly doubled by 2012, and is expected to reach 160 zettabytes by 2025 [3]. The developing ability of artists, scientists, and scholars to create massive datasets that are integral to their scholarly work has far outstripped the rate at which university data curation systems and policies adapt to the new condition. This paper outlines some of the origins of this explosion in data volume, describes some of the specific challenges posed by the accelerating pace of data creation, and examines strategies that university systems are considering for managing the unfolding Data Age.

The proliferation of big data in research

In part, the explosion of research problems employing big data sets has been driven by the remarkable technological advancements in high performance computing and networking. Many of today's researchers have never experienced the challenges associated with the early days of computing, when computer code and, often, data sets had to be entered by hand onto punch cards or paper tape, and when CPU register space was severely limited. Recent estimates suggest that compute capacity per dollar has increased by a factor as large as 10^{15} over the past 60 years. [4] At the same time, the decreasing price of data storage capacity, the invention of the Internet in the late 1960's, and the recent move toward 100 Gbps Ethernet capacity has made it more feasible to bring large data sets to the compute capacity together.

These dramatic technological enhancements have allowed investigators to analyze more comprehensive and, presumably, more realistic data sets in a range of computational and modeling applications. Arguably, the signature big data project of the last century was the mapping of the human genome. But today, evolutionary biologists regularly use datasets ranging up to 10 Tb to map previously unknown genomes using de novo assembly strategies. Similar types of big data applications are playing out in analyses of climate data, educational testing and evaluation, modeling of turbulence around aircraft and wind turbines, business analytics, and countless other fields.

How universities are addressing the data challenge

One clear trend among institutions supporting research using big data sets is investment in high-quality high-performance research computing (HPC). The University of Kansas (KU) has only had a centralized HPC strategy for five years, and is therefore a late entrant into this arena. The factors that motivated us to provide centralized support for HPC probably mirror those of many public research universities:

- Saving resources by reducing the proliferation of cold rooms for housing servers.
- Minimizing duplication and underutilization of IT staff.
- Allowing researchers and students to focus on research, not on attempting to maintain clusters and servers.
- Creating an efficient HPC computing and networking environment.
- Providing data management capabilities in response to federal sponsors.
- Minimizing threats to data security.

KU chose to initially undertake a subsidized "condominium" model for support of HPC, with investigators storing hardware purchased for their programs in a common cold room environment. This program was modeled on the "community cluster" model Purdue University uses to support HPC. Many other institutions purchase or lease large amounts of compute and storage capacity and charge investigator grants for use of the resource, while some provide free access to institutional computing resources to all investigators. There is no single model for how major universities offer similar resources to support HPC-based research with large datasets.

I have often told my colleagues and staff, '…the minute it becomes economical, secure, and efficient to off-load HPC computing and storage capacity to the cloud, we want to be out of the business of owning and leasing "enterprise" HPC hardware.' By "enterprise" HPC hardware, I mean multiprocessor cluster units (containing either standard processors or GPUs) and standard spinning storage media that can support 85% or more of common needs for HPC research computing applications. (In contrast, I note that there will always be a need for universities to host hardware for researchers investigating new technologies and configurations of computing, networking and storage hardware.) Universities generally are not adapted to the mission of optimizing technology business processes, nor do our individual operations create the financial efficiencies of operations run by Amazon©, Google©, or Microsoft©. Owned or leased computer hardware also does not support the degree of scalability offered by major cloud computing providers.

Unfortunately, most contemporary analyses of cloud computing services still do not support fully moving university enterprise HPC to a cloud platform. In part, this is because the cost of using dynamic storage in a cloud computing environment is still prohibitively expensive. But the landscape is constantly changing. Most larger university compute customers have transitioned from outright hardware ownership to designed multiple year leases of hardware. This provides a reliable method for maintenance and upgrade of computational hardware at a predictable annual cost that is a fraction of the lump sum investment in a hardware purchase. Many universities are using cloud HPC services when their compute demand bursts over on-campus capacity. Several have experimented with the services of a computing broker to advise them on diversifying their monetary investment in HPC to maximize their computing power.

The current economics of storage technology represents more of a mixed picture. Dynamic cloud storage is currently far more expensive than on-campus spinning media. This is driven by data access costs companies add to the fees for storage capacity. However, glacial cloud storage, which is used for long-term archiving of infrequently accessed data, does provide a cost advantage over even low-tier on-site storage. Many big data research file storage systems are now featuring a seamless interface between dynamic on-site spinning media and glacial cloud storage. Part of the trade-off for employing cloud storage is the advantage of having scalable, immediately accessible storage resources at the expense of giving up some control over institutional data integrity.

Key challenges facing universities and researchers

Though the cost of computing, networking and storage capacity have all exponentially declined over our lifetime, one of the challenges facing universities is that the increase in size of many research data sets has balanced and possibly outstripped the rate of these cost savings. Investigators have noticed this, and, in the absence of long-term institutional strategies for high quality data curation, have flocked to low-cost and sometimes low-quality technologies for data storage. Funding agencies, in turn, have noticed this trend, and have begun to intervene out of concern for data integrity and accountability to taxpayers. Federal agencies funding research have instituted requirements in research proposals for data management plans. NIH has notably created publication and data repositories for investigators funded by the Department of Health and Human Services. In 2013, the Office of Science and Technology

(OSTP) instituted a policy aimed at making the data collected through federal research funding available to the public. [5] This policy provided no hint of how universities were supposed to (a) curate such enormous volumes of research data or (b) fund this policy mandate.

The OSTP mandate has created a dilemma for research universities. Few university IT systems have been engineered with the bandwidth or server capacity to provide public access to research data. In fact, one could argue that security concerns are driving most universities to isolate their research data platforms from public access. Numerous research universities have moved toward creating research DMZ's (sequestered research computing, storage and networking zones) to separate research data flow from business enterprise, education, entertainment and social media traffic, which often dominate day-to-day activity on university networks. But the question facing universities is: Should each research university create its own public research data portal in order to provide access to data from federally funded projects? Six hundred forty universities are listed in the most recent NSF Higher Education Research and Development (HERD) survey. [6] It seems apparent that scattering unrelated disciplinary data among 640 separate data archives, with the accompanying potential for devolution of technology and interface compatibility, might satisfy the letter of the OSTP directive; but would not provide the public with transparent accessibility and use of the archived data.

Another challenge facing research universities is policy makers' concerns about the leakage of technologies (including physical artifacts, information, and tacit knowledge and skills) that contribute to U.S. competitiveness to foreign governments and companies. Though some of this leakage is unintentional, cases of espionage aimed at industrial and academic research and development activities are well documented. According to most security experts, such incidents are increasing in frequency and intensity. This activity is not solely aimed at classified and dual-use technologies, but also at technologies that fall into the category of controlled unclassified information. Controlled unclassified information (CUI) is defined and categorized by the National Archives, which is also responsible for creating standards for its protection. [7] CUI can include an astounding range of data types, including:

- Patent data/proprietary process information.
- Confidential technical information.
- Export controlled information and technology.
- Personal health and financial information.
- Law enforcement data.
- Critical infrastructure specifications.

This area of regulatory controls represents a new world and new challenges for research universities. Not only will CUI regulation affect the nature and scope of some technological information we are can openly publish, but deemed export controls are likely to affect which subjects can be studied by certain groups of foreign nationals.

Key principles for a saner, more useful, and secure data landscape

Resolutions to the issues outlined above will require continued dialog among the academy, and the federal agencies that fund and regulate research. A diverse group of stakeholders in research universities must be engaged in developing proposals to address these concerns. University leaders must engage faculty in developing funding plans and policies to support the efficient management of research data. Representatives of scholarly disciplines must define best practices for data utilization in a way that serves the needs of modern research in their fields. Librarians must examine methods for the efficient curation of scholarly data, and collaborate with disciplinary representatives to create interfaces that support the research culture of the discipline. Information Technology specialists must ensure that systems are resilient, affordable and broadly compatible with the needs of the vast array of disciplines represented in the academy.

At a recent symposium examining the challenges of supporting research using big data sets, there was a discussion of the mandate for public access to data from federally sponsored research. Several thought leaders proposed an alternative to the creation of individual institutional data archives. Following on the pattern of NIH's creation of PubMed [8], and the Census Bureau's development of Federal Statistical Research Data Centers [9], a suggestion was made to gather related data into disciplinary data archives. This solution would create centralized repositories of research data commonly used in similar disciplines. In spite of the potential advantages of this approach, several concerns were discussed. Some of the most interesting research questions derived from big data may stem from finding correlations between datasets not commonly associated with a single discipline. By creating silos composed of commonly associated data sets, we might inadvertently impede interdisciplinary inquiry. But perhaps the larger conundrum with this proposal is that disciplinary societies and associations seem unlikely to volunteer to host and fund continuously growing data archives, and federal agencies have not stepped forward to offer ongoing funding to support their formation. Continued examination of these issues must continue to engage all stakeholders.

In order to promote a healthy data culture in higher education, the following principles seem to form a reasonable framework for future action by research universities:

- Provide economical access to high quality, professionally maintained computing capacity and archival storage.
- Wherever possible give up ownership of computational and storage hardware to commercial vendors who maximize value for compute and storage spending.
- Facilitate, as appropriate, the transition from paper-based research records to electronic records; and streamline the association of various records, research products, and data sets associated with a particular project.
- Standardize meta-data to identify data sources and ownership, associate data sets with original funding sources and publications, and define keywords and data

tags to provide consistency in data curation and searching.

- Create internal data management training and policies that minimize the volume of data retained for long periods of time.
- Engage disciplinary experts to incorporate data management best practices into curricula, create norms for data lifecycles, and develop strategies for centrally storing and curating related data resources.
- Develop shared application interfaces to bring computing tasks to large data sets.
- Create institutional capacity to ensure compliance with CUI controls.
- Continue the dialog with funding agencies about sustainable support for research data archives.

Notes and references

[1] The answer, of course, is: "Yes, assuming someone chooses to include it in a research study."

[2] Webopedia Staff, "How much data is out there?" Webopedia, March 2014, http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html

[3] Andrew Cave, "What will we do when the world's data hits 163 zettabytes in 2025?" Forbes, April 2017, https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#f7788a3349ab

[4] Hardware and AI Timelines, "Trends in the cost of computing." March 2015, https://aiimpacts.org/trends-in-the-cost-of-computing/

[5} OSTP, "Increasing access to the results of federally funded research." February 2013, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[6] National Science Foundation, "Higher Education Research and Development Survey: Fiscal Year 2015." March 2017, https://ncsesdata.nsf.gov/herd/2015/.

[7] National Archives, "Controlled Unclassified Information (CUI)." October 2017, https://www.archives.gov/cui.

[8] NIH, "PubMed." https://www.ncbi.nlm.nih.gov/pubmed/.

[9] U.S. Census Bureau, "Federal Statistical Research Data Centers." https://www.census.gov/about/adrm/fsrdc/locations.html.