Big Data in the Campus Landscape

Security and Privacy

ECAR Working Group Paper | July 2, 2015

This paper is a publication of the ECAR Campus Cyberinfrastructure (ECAR-CCI) Working Group. It is the third paper in a four-part series focusing on Big Data in the Campus Landscape. ECAR working groups bring together higher education IT leaders to address core technology challenges. Individuals at EDUCAUSE member institutions are invited to collaborate on projects that advance emerging technologies important to colleges and universities. ECAR-CCI helps educational institutions develop institutional strategies and plan resource deployment and helps users harness and optimize the power and capabilities of these new integrated IT tools and systems for educational and research applications in higher education.

Introduction

Big data is often characterized by the amount of data involved: data sets so large they cannot be manipulated by traditional database techniques. But several other characteristics are equally important: the number of data sets being aggregated, how different they are from one another, how quickly they change, and whether data quality is similar across the aggregation. Different types of big data being created and used by research projects in higher education will implicate different concerns. For example, the enormous streams of data being generated by a telescope—as a single data set with uniform quality, single ownership, and invariance (once captured)—will be concerned with security and potentially also with distributed storage and processing. However, combining data from multiple telescopes around the world will likely need to occur virtually, as it may not be practical or desirable to aggregate or replicate such data into a single physical repository. Conversely, an institution may want to share a gigantic data set with collaborators globally. These scenarios will have additional technical implications such as distributed access controls for those responsible for computational infrastructure.

Key Definitions

Big data is a "term used to describe data sets so large and complex that they become difficult to process and store using [traditional] data management tools."[1]

Information security is often defined as ensuring the confidentiality, integrity, and availability (CIA) of information. *IT security* is the part of information security that relates to the protection of technical infrastructure, including hardware, data, and software.

Privacy is a simple term for two complicated concepts. At its most basic layer, privacy is always about people and their control of their personal information. *Information privacy* protects individuals by protecting information about them from unauthorized disclosure (think compliance with FERPA, HIPAA, IRB regulations, or state breach notification laws). *Autonomy privacy* protects individuals by protecting them from unwanted surveillance (the monitoring of behavior, profiling), an underpinning of academic freedom and freedom of speech.

1. See the *Data Management Glossary*. Many other definitions of big data are available; Doug Laney described it as a series of V's in his 2001 Gartner research note "3D Data Management: Controlling Data Volume, Velocity, and Variety." Since that time, others have added more V's. See "Deja VVVu: Others Claiming Gartner's Construct for Big Data."

At the same time, an aggregation of student records, results from a student survey on sexual habits, and medical records—all coming from multiple institutions, hospitals, and organizations—will have a different set of concerns: privacy of human subjects, legal compliance, and ethical conduct (e.g., through misuse of these data or by unintended consequences of the different levels of quality of the data sets). These concerns will multiply in complexity if data cross legal jurisdictions. But such concerns are largely nontechnical in nature, which could require that people responsible for computational infrastructure reach out to, and partner with, other functional offices to provide new types of support needed for research big data.

Institutions will need to begin thinking about a set of overarching privacy and security issues. There are no right answers for these issues; they depend in large part on each institution's culture, appetite for risk, resources, and environment. In this paper, we discuss information security, privacy, compliance, and institutional considerations that surround the research uses of big data at institutions of higher education.

Information Security

Findings

- An informed risk-assessment approach provides the security architect with the guidance needed. A risk-based security analysis of big data may be the most effective approach.
- Effective tools for federated authentication and authorization should be applied to manage the security and privacy of research big data sets.
- Research big data security concerns drive new or strengthened partnerships among data owners/providers, researchers, and technical staff.

Research—particularly where big data is a result—is driving new partnerships and bringing new challenges in understanding how to protect personally identifiable information. In terms of information security, the scale and complexity of the data pose challenges for data management.[1] Because these data have typically been collected by different stakeholders following different rules, finding ways to collect and map them to each other is an unsolved problem. Many higher education institutions do not have an infrastructure environment that can scale to effectively compute on these data, much less follow granular provenance rules.[2] Academic computing organizations at universities or in public clouds are being recruited to support analysis of sensitive data at scale, yet the legal and risk frameworks to support such analyses are rarely in place.

Risk-Assessment Approach

A risk-assessment approach is critical for securing big data, but risk analyses need to incorporate novel factors, such as single risks that can simultaneously affect multiple, previously independent, data sets.

At first glance, the challenges of securing research big data appear daunting. Fortunately, good risk management provides the security architect with the guidance needed. Although legal requirements may dictate a minimum set of security controls, even with those constraints in place a risk-based approach to implementing security controls will ensure more effective security.

A risk management approach helps an institution identify the many different risks that it faces with respect to big data and prioritize them according to likelihood of occurrence and possible damage. From that initial assessment, an institution can decide how to address those risks in a way that makes sense according to its business practices.[3]

This type of approach is crucial because possible risks may differ from data set to data set—there isn't a single model to secure all research big data that comes from disparate sources and comprises different data types (complexity and variety typically being bigger issues than the size of the data). Many of the risks and concerns cannot be appropriately viewed through the lens of operational security. By focusing systematically on the risks that cross multiple data sets (for instance, aging systems, disgruntled employees, or agents working to breach system or data security), a security architect can prioritize the controls necessary to address understood vulnerabilities and exposures.[4]

Federated Authentication and Authorization

Effective tools for federated authentication and authorization should be applied to manage the security and privacy of research big data sets.

It is unrealistic to replicate these (often dynamic) data sets, so access management needs to include federated authentication and authorization methods as well as data-transfer mechanisms that are secure and mutually trusted.

Research big data—what may look like or be thought of as a large data set—may in fact be a compilation of a number of small data sets, some of which may have restricted access, different owners, and different consent policies. Someone may have access rights to some parts and not to others—the aggregation makes managing these access rights difficult. This implies the need to have different access rules for each set, which may make some queries difficult, but removing all controls across the board isn't the answer. As complex intellectual property rights can derive from big data, it might become more useful, as was done with the HIPAA legislation, to think of big data as having stakeholders rather than owners. Leveraging federated authentication systems, such as SAML-based ones, can help maintain inter-institutional trust fabrics necessary for ongoing management of complex data relationships.

Security Concerns Drive Partnerships

Security concerns drive new or strengthened partnerships among data owners/providers, researchers, and technical staff.

Regardless of the specific methodology used for a risk assessment, it is essential that researchers recruit security, compliance, and legal partners to work collaboratively to address security issues. The security analysts will not be able to completely identify potential risks and their drivers without the researchers. Nor will the research staff be made aware of how the project may need to adapt to lend itself to successfully manage risk through security without these other partners. Proactively developed security policies can decrease effort by providing general guidelines.

Privacy

Findings

- Aggregating disparate data may result in a new data set that permits identification of individuals.
- Big data is in tension with Fair Information Practice Principles (FIPPs), and we need to be cognizant of how that may affect privacy concerns.

Privacy always has to do with people. In the traditional research context, areas such as the social sciences and health care that involve human subjects have worked with institutional review boards (IRBs) to protect individuals' privacy.

The big data context brings new privacy challenges to the fore, though they do not typically implicate technical infrastructure. Big data can also make existing challenges more complicated; for example, how do the privacy findings identified here affect our ability to comply with federal funding agency requirements for data management plans or data sharing plans? Consider reaching out to your institution's privacy officer or other individual responsible for this area to see if that person can provide insight or other assistance.

There are really no solutions to give because of the inherent tension between data collection and privacy; this is a decades-long issue with no pat answers. With the emergence of big data, some frameworks are starting to emerge, but balancing collection and privacy will likely be an ongoing issue that requires attention for some time.

Aggregation Results in Identification

Aggregating disparate data may result in a new data set that permits identification of individuals.

Over the past several years, anonymization and de-identification techniques have been broken by security researchers with some regularity, leading some privacy professionals to conclude there may be a day when we have to assume no such technique will work for long. (While this may be caused less by the effectiveness of the techniques than the rigor with which they are implemented, the consequences are still the same.)[5] Big data can dramatically sharpen this challenge. Even if individual data sets were anonymized or had no underlying identifiable elements, aggregating them may result in a new data set that has sufficient information to (re)identify individuals. A January 2015 paper from MIT discusses research showing that "someone with copies of just three of your recent receipts—or one receipt, one Instagram photo of you having coffee with friends, and one tweet about the phone you just bought—would have a 94 percent chance of extracting your credit card records from those of a million other people."[6]

One consequence is that releasing data for research purposes becomes trickier than ever. Consider an early example from 2006, when AOL released the search terms of over 650,000 users over a three-month period for research purposes. The action backfired when others quickly began identifying the users who had made the queries.[7] Like many issues raised in this section, there may be no pat answer to give. At the same time, researchers involved should be aware of the standards of their disciplines and of other resources they can tap; so maybe it's another matter of close partnership between researchers and infrastructure.

Big Data and FIPPs

Big data is in tension with Fair Information Practice Principles, and we need to be cognizant of how that may affect privacy concerns.

One of the fundamental tenets of the FIPPs[8]—the principles that underlie most of the federal privacy laws in the United States—is to provide notice to individuals regarding what information about them is being collected and for what purpose so that they can make informed decisions about the risks of participation. Big data often means collecting, aggregating, and retaining data in anticipation of new ideas for its (re)use, which will require a new model for providing transparency to permit individuals to make informed choices about giving consent. Aggregating data and retaining it indefinitely for new uses means it is difficult to articulate—other than in very general terms—how an individual's data might be used, reused, shared with others, combined with other data, or result in unanticipated research outcomes. In turn, this makes it difficult for individuals to meaningfully make informed choices and provide consent.

This fundamental tension between FIPPs and big data can be seen in other principles. For example, some types of research benefit from mass data, with the value of the data to the research increasing as more data are aggregated. This directly conflicts with the data-minimization principle, which states that only the minimum data necessary for the specified purpose should be collected in order to protect privacy.

Compliance

Findings

- Compliance complexity can rapidly increase as data sets of different provenance and type are combined.
- Compliance complexity can rapidly increase as data are shared among entities subject to different laws.
- New or strengthened legislation that could impact research big data should be expected and will require institutional action.

Research data—no matter the size—are already subject to federal and state regulations. However, it may be less clear how to comply with existing laws when large, aggregated data sets are involved, and further legislation to clarify these issues can be expected.

Combining Data Sets Can Increase Compliance Complexity

Compliance complexity can rapidly increase as data sets of different provenance and type are combined.

Just as the privacy characteristics of an aggregated data set may differ from those of its underlying components, aggregation can trigger legal requirements not relevant to the underlying data. Consult with your institution's legal counsel for advice. Protected health information (PHI) can be particularly complicated and fraught with consequences, so if it is involved in any manner, your institution's HIPAA privacy and security officers (or equivalent personnel) should be consulted.

Sharing Data Can Increase Compliance Complexity

Compliance complexity can rapidly increase as data are shared among entities subject to different laws.

Researchers working with data sets that include personally identifiable information must consider the legal implications of sharing data, whether between public and private institutions, institutions in different states, or institutions in different countries, as well as the implications of placing research data in public cloud storage environments of uncertain location. Researchers must take into consideration, for instance, the varying data protection laws across states and countries, as well as the fact that public institutions are subject to open-records laws. The same is true of data involving intellectual property. Consult with your institution's legal counsel for advice.

New Legislation Could Impact Research Big Data

New or strengthened legislation that could impact research big data should be expected and will require institutional action.

Compliance with existing requirements—of the Federal Information Security Management Act of 2002 (FISMA; see sidebar) for information security, for example, or of various data-breach laws that speak to information privacy and the safeguarding of information about individuals—can already require a tremendous institutional investment of time, effort, and resources. The ongoing string of megabreaches is only sharpening state and federal focus in this area, and further obligations are likely to result.[9] Those responsible for computational infrastructure will need to have strong partnerships with those responsible for institutional policy, privacy, and security to be able to achieve compliance and adequately manage the risk to sensitive data.

Federal Regulations and Research Big Data Privacy

Legislation such as the Federal Information Security Management Act of 2002 (FISMA) applies to institutions doing research under contract to the federal government. An outcome of federal regulations in this space is the requirement for a risk-based approach to the security of electronic protected health information (ePHI) and other sensitive data. Such an approach provides institutions with a framework for a long-term, sustainable strategy toward reasonably secure data management.

Institutional Considerations

Findings

- Institutional analysis involving big data looks more and more like traditional research but without the IRB protections for human subjects.
- Data are institutional assets.

Big data raises issues sufficiently new that institutions have not yet converged on a common set of expectations. Thoughtful deliberation is needed by each institution to arrive at an institutional position on these issues, even amidst evolving legal requirements and social norms. Governance committees may need to reach out for expertise to help inform these discussions.

IRB Protections and Big Data

Institutional analysis involving big data looks more and more like traditional research but without the IRB protections for human subjects.

Institutions are increasingly engaging in predictive analytics for internal purposes that are not considered human-subjects research. This is particularly true in the area of learning and student success, where institutions are experimenting with a wide variety of approaches.[10] Some of these approaches build detailed profiles of student behavior in order to forecast a student's success and/or to prescribe interventions, raising serious concerns about students' autonomy and privacy, about the institutional view of *in loco parentis*, and about inadvertent harm should predictive or prescriptive models turn out to be wrong or have unintended consequences. How much of this kind of activity is permissible under FERPA is also an open question. Other concerns are related to the assumption that "hard data" are inherently objective, when in fact both data and the algorithms that manipulate them can be biased. Parallel concerns outside the academy can be found in many quarters, including the Federal Trade Commission's 2014 workshop "Big Data: A Tool for Inclusion or Exclusion?"[11]

In the absence of IRB involvement, how do we ensure that institutional actions follow the basic ethical principles—or their equivalents in the big data context—that an IRB would require of human-subjects research?[12] Ultimately, a common set of expectations about the appropriate and ethical use of data, and a governance structure to ensure these expectations are met, will provide an underpinning for community trust.

Data Are Institutional Assets

Data are institutional assets, just as buildings or trademarks are institutional assets that are understood to have value to the institution and thus are surrounded by laws and policies that protect them and articulate appropriate use. Seeing data through this lens is crucial when data are shared across entities, all the more likely with the potpourri of data sets that big data can represent.

First, issues of ownership can be muddied with big data because data can come from many sources. For example, outside the traditional research realm, in the learning analytics space there is a question about whether students should have a say in the data that are being gathered about them, if not own those data outright—benefits and risks both accrue with ownership.

Second, sharing data requires trust that your collaborator will safeguard and use data as you would. This is often implemented contractually as a set of security and privacy requirements between entities, but trust becomes more challenging between organizations whose missions differ fundamentally (consider the broad distrust that higher education has that a private cloud storage provider won't misuse students' data even with agreements in place). Even when organizations share similar missions—such as collaborations between public research universities—an unforeseen change in circumstances can affect all collaborators.

Third, outside the traditional research context but in the realm of analytics, data we have about our students, faculty, patients, donors, and other extended members of our community have enormous value to private-sector companies. There are endless opportunities to partner with companies to analyze our data on our behalf—but it's important to be mindful they are also gaining knowledge about what is important to higher education that they can then sell back and will be directly shaping the services we will live with. These companies may not conduct themselves by the same ethical standards we do regarding their treatment of data—are we culpable in providing it to them? The situation is analogous to academic publishing, where publishers have aggregated books or journal articles and then charged institutions for access to the intellectual property and aggregated data products that academics had originally created.

Conclusion

Overall Findings

- Volume may be a primary characteristic defining big data, but other characteristics can create more complex challenges without volume.
- Research big data often combines data that are not otherwise gathered together, and that brings additional scrutiny as well as security and privacy risks.
- There is an increased need for awareness about legal, policy, and ethical issues surrounding research big data.
- Security and privacy go hand in hand, and a risk management approach that considers both issues in concert will likely yield the best results.

Existing security and privacy tools and practices should be applied to big data that result from research in higher education. However, the unique nature of research big data will benefit from an institutional, strategic, risk-based approach to protecting the data.

Perhaps the most effective approach to securing these data and ensuring that privacy is protected is to be proactive in addressing these concerns. The IT department should consult with the institution’s privacy officer to help identify concerns and to understand what assistance the officer may be able to provide. Moreover, your institution may want to consider engaging collaborative teams from legal, compliance, research, security, and privacy communities to proactively develop guidelines for managing security and privacy for research big data.

Finally, in this rapidly growing and changing environment, higher education institutions should find like partners to encourage information sharing and collaboration. Sharing current practices and identifying areas where common best practices may be developed for the community will help address common concerns while decreasing the potential for problems in this space.[13]

Authors

William Barnett
Chief Research Informatics Officer
Indiana CTSI and Regenstrief Institute

Mike Corn
Deputy CIO and CISO
Brandeis University

Curt Hillegas
Associate CIO for Research Computing
Princeton University

Kent Wada
Director, Strategic IT Policy and Chief Privacy Officer
UCLA

Citation for This Work

Barnett, William, Mike Corn, Curt Hillegas, and Kent Wada. *Big Data in the Campus Landscape: Security and Privacy*. ECAR working group paper. Louisville, CO: ECAR, July 2, 2015.

Notes

1. Though physical security is required for big data, it is not differentiated from regular research data. Once data span multiple data centers, the additional or differentiated security concerns may be connected with transport more than actual physical security.
2. For more about how research big data impacts infrastructure, see the earlier publication in this series, *Big Data in the Campus Landscape: Basic Infrastructure Support*, available from the Big Data in the Campus Landscape page.
3. Numerous resources exist on information security risk management practices. For resources prepared by higher education IT practitioners, view the Risk Management chapter of the HEISC *Information Security Guide*.
4. For an example of one prioritization approach, see Intel, *Prioritizing Information Security Risks with Threat Agent Risk Assessment*, IT@Intel white paper (December 2009).
5. For more information, see the EDUCAUSE HEISC resource "Guidelines for Data De-Identification or Anonymization."
6. Larry Hardesty, "Privacy Challenges," *MIT News*, January 29, 2015.
7. Michael Arrington, "AOL Proudly Releases Massive Amounts of Private Data," *TechCrunch*, August 6, 2006.
8. National Institute of Standards and Technology, "Appendix A—Fair Information Practice Principles (FIPPs)."
9. See, for example, the attacks on Sony, Target, and the health insurer Anthem.
10. See, for instance, James Willis III, John Campbell, and Matthew Pistilli, "Ethics, Big Data, and Analytics: A Model for Application," *EDUCAUSE Review*, May 6, 2013.
11. Federal Trade Commission, "Big Data: A Tool for Inclusion or Exclusion?," FTC workshop, Washington, D.C., September 15, 2014.
12. U.S. Department of Health and Human Services, Office of the Secretary, "The Belmont Report," April 18, 1979.
13. Some forums where these discussions might occur include the EDUCAUSE Higher Education Information Security Council (HEISC), the REN-ISAC, and the CASC Regulated Data Committee.