Preparing Graduate Students for the Ethical Challenges of Big Data

A Report from the Council of Graduate Schools (CGS) &
Pervasive Data Ethics for Computational Systems (PERVADE)

Preparing Graduate Students for the Ethical Challenges of Big Data

A Report from the Council of Graduate Schools (CGS) &
Pervasive Data Ethics for Computational Systems (PERVADE)

By Jeffrey Engler and Katie Shilton

March 2022

Preparing Graduate Students for the Ethical Challenges of Big Data: A Report from the Council of Graduate Schools and PERVADE

ISBN-13: 978-1-933042-62-6
ISBN-10: 1-933042-62-1

Suggested Citation: Engler, J. and Shilton, K. (2022). Preparing Graduate Students for the Ethical Challenges of Big Data: A Report from the Council of Graduate Schools and PERVADE. The Authors.

ABOUT THE PROJECT

Preparing Graduate Students for the Ethical Challenges of Big Data is based on a project designed to better understand challenges and opportunities in preparing graduate students for the ethical challenges of research using large datasets. In collaboration with graduate deans, experts in big data, and practitioners of research integrity education, CGS and PERVADE analyzed current perceptions of need and identified action steps with the potential to expand current approaches to research training. The project was made possible with generous funding from the Office of Research Integrity (ORI) and Elsevier.

ABOUT CGS

CGS is an association of 500 graduate institutions that is dedicated to the advancement of master's and doctoral education and research. Working closely with our members, we advocate for graduate education, develop innovative research, and identify best practices. Our projects generate information and data that help graduate deans and their institutions better support graduate students and programs. CGS has engaged in projects that advance the professional development and career success of graduate students and support comprehensive approaches to research and scholarly integrity and the responsible conduct of research (RCR).

ABOUT PERVADE

Pervasive Data Ethics for Computational Systems (PERVADE) is a cooperative research project formed in 2017 with funding from the National Science Foundation. Uniting researchers across six institutions, PERVADE research seeks to answer open empirical questions in big data ethics such as: how do we quantify the risks to individuals and groups in the use of pervasive data? How do people experience the reuse of their personal data? And how should existing ethical codes be adapted and adopted for computational research? PERVADE's work is supported by the National Science Foundation under Grant IIS-1704369. *Any opinions, findings, and conclusions or recommendations expressed in this report do not necessarily reflect the views of the National Science Foundation.*

Contents

Acknowledgments

This work was supported by a grant from the Office of Research Integrity of the U.S. Department of Health and Human Services (grant ORI IR190054-01-00) and by funding support from Elsevier. PERVADE is supported by NSF grant IIS-1704369 (Katie Shilton, Project Director). CGS and PERVADE are grateful to Suzanne Ortega, Julia Kent, Michael Zimmer, Casey Fiesler, and Thomas Jeitschko for helpful guidance and advice in planning the content for the project workshop in April 2021, and to the workshop participants for their thoughtful contributions and insights. We also thank Matthew Linton, Heidi Shank, and Katherine Hazelrigg for providing technical support during the workshop, and Katherine Hazelrigg for editorial assistance in the preparation of this report.

This publication was supported by Award No. 6 ORIIR190054-01-02 from the Office of Research Integrity. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of ORI.

Executive Summary

Large datasets present exciting new opportunities for the U.S. and global research enterprise. Indeed, "big data" approaches to research have the potential to develop new knowledge and innovations across nearly every broad field of study, particularly in the biomedical sciences, computer science, engineering, and the social sciences. Yet the methods used to assemble large datasets, and their applications in decision-making contexts, challenge existing ethical paradigms for data management, data integrity, human subject protections, and data use. In many fields, for example, aggregating data from different sources can make privacy protections for human subjects more complex and raise questions about data ownership. In others, the use of algorithms and predictive analytics may lead researchers to influence—not simply predict—human behaviors. Unfortunately, current attempts to identify and address these challenges are often focused within specific disciplines or corporate settings and offer little opportunity to integrate these evolving ethical concerns within graduate programs preparing the next generation of researchers.

Graduate deans often oversee professional development and responsible conduct of research (RCR) training curricula and are uniquely positioned to present the ethical concerns of big data research to their university communities and to bridge potential silos that impede sharing best practices to address these evolving challenges. To address this gap in graduate student preparation, the Council of Graduate Schools (CGS) and PERVADE (Pervasive Data Ethics for Computational Systems) embarked on a project to better understand the challenges and opportunities universities face in preparing graduate students in the ethical use of big data. Our goals were to identify both broad and specific ethical challenges that arise from the use of big data resources in graduate student research; to discuss and evaluate existing resources for training in the ethical use of big data; to identify potential levers for introducing and discussing these challenges, and for engaging Principal Investigators (PIs) and advisors in helping students prepare for them; and to formulate potential strategies for deploying and embedding resources for big data ethics within academic programs, professional development opportunities, and RCR training.

With generous funding from the Office of Research Integrity (ORI) and Elsevier, CGS and PERVADE convened a diverse group of graduate education leaders around these topics. The virtual event, held in April 2021, brought together graduate deans, experts in the ethics of big data research, and representatives from disciplinary societies and other organizations. This report synthesizes lessons learned from this event with the goal of informing and strengthening efforts to prepare graduate students for the challenges of big data research.

The five major conclusions and recommendations from this collective work are intended to stimulate further action and reflection in the research and graduate education communities.

CONCLUSIONS AND RECOMMENDATIONS:

1. **The ethical challenges of research involving big data are relevant to a large population of master's and doctoral students and should be broadly integrated into graduate research training.** While big data methodologies are sometimes seen as a hot topic or novel innovation, they should not be relegated to specialized training programs or courses.

2. **The research and graduate education communities should evaluate current RCR curricula and ensure that they address challenges in big data.** Research with large datasets is changing the way we need to teach several categories of RCR training recognized by the federal government, in particular, collaborative science; data acquisition, management, sharing and ownership; and human research protections.

3. **Plans to expand graduate research training to address ethical issues in big data research should include the participation of a broad range of stakeholders.** Universities should involve faculty, students, IRB review boards, vice provosts for research, IT staff, and others, from many different disciplinary training programs in their efforts. Students benefit when different groups on campus communicate and collaborate on a coordinated approach.

4. **Graduate deans, as the individuals with the broadest responsibility for the quality of graduate student research training, should play a lead role in supporting and facilitating institution-wide collaborations.** The graduate dean community has a strong track record of supporting communication and collaboration across campus with the goal of improving and expanding student learning and professional development.

5. **Universities, organizations that support graduate education, and funders should increase their efforts to develop resources that prepare graduate students for the ethical challenges of research using large datasets.** As resources such as curricula and case studies become available, these groups should work together to make them centrally available.

Part I: The Challenge

"Big data" are often characterized by one or more of the following three characteristic "V's": *volume* (data come from many sources), *velocity* (data are generated and can accumulate rapidly), and *variety* (data come in many different formats that must be aligned and annotated) (1). The adoption of big data in research has strained our existing paradigms about the ethical use of data in the areas of data management, data integrity, personal privacy, and cybersecurity. These challenges affect not only the research community but also many other political, economic, and social interactions. For example, a report from the National Academies (2) cites the possibility of attacks on big data repositories and the corruption or distortion of the data stored within them ("intentional data alteration"). Public reports of hacking of consumer data and the inappropriate use of the Facebook platform by Cambridge Analytica reinforce the need to train future users and developers in the ethical use of big data. However, big data also present ethical challenges to researchers in more typical contexts, creating confusion, for example, about ethical principles for using social media data in research (3, 4).

Graduate training presents an ideal opportunity to expose students to ethical challenges in research and to provide them with opportunities to navigate those challenges. Graduate students are in the process of developing their habits and professional identities as researchers and scholars (5). However, as has been noted in previous studies, minimalist or compliance-oriented approaches to ethics training may lead to important gaps in training (6). In the context of big data, one example of such a gap is in big data and artificial intelligence tools. In some training programs, particularly those associated with data or computer science, student researchers may learn about the risks of using AI tools that draw from large datasets. However, more and more disciplines are using such tools without a clear understanding of the potential for harm to individuals, for example in studies using anonymized big data in sociology, public health, and medical research.

There has been some previous discussion of these issues within the graduate education community. In 2015, a Global Summit hosted by the Council of Graduate Schools and the National University of Singapore (NUS) on the implications of big data in graduate education shed light on several challenges for research ethics education. One of these was the interdisciplinary nature of many projects involving large datasets. Interdisciplinary and multidisciplinary research may present additional complexity to research endeavors, requiring a student to navigate different research cultures and norms and to use a variety of research methodologies (7). This, too, may create disagreement or ambiguity about what constitutes ethical research practice for students beginning their research careers.

The Global Summit resulted in a set of recommendations for addressing this and other issues in graduate education and big data research. The summit participants, representing 15 countries, made a number of recommendations related to graduate training and big data research. These included: developing best practices and case studies of 'big data' education across disciplines; identifying what skills and knowledge—beyond technical and statistical skills—students need to succeed in a world of big data; and determining whether responsible conduct of research training will adequately address such issues (8).

Since 2015, with the sheer amount of data generated for research use growing exponentially, the pressure to address these issues has intensified. While there have been a number of systematic efforts to improve research and scholarly integrity in graduate education, none of these efforts has focused on how universities are handling big data ethics education broadly within graduate education and at an institutional level. The specific challenges of big data and its impact on a broad range of STEM research fields are topics not often included in this training, even as the use of large, aggregated databases and new methods to mine them expands in many biomedical, STEM, and humanities disciplines.

Part II: Potential Solutions

To help graduate deans further understand issues and challenges in the ethical use of big data, CGS partnered with PERVADE, an NSF-funded project to explore the potential risks posed by using big data in research and in technology. This partnership pooled our expertise with the goal of strengthening research education as these developing research risks arise. Through the conversations held in a workshop setting, participants explored the ethical issues and resources needed to expand ethics education in this area.

A majority of participants were graduate deans, university leaders who play a key role in the professional development of graduate students. Graduate deans can help graduate programs identify student needs and help scale promising ethical practices across many programs of study. In many cases, deans and their staff may also advise graduate programs and other university leaders as they adjudicate allegations of ethical or academic misconduct.

The virtual workshop, which took place over three afternoons (April 12-14, 2021), sought to engage graduate deans and other experts in developing strategies for effective training in the scholarly and ethical challenges of big data research. This workshop brought together thought leaders from the big data ethics community with graduate deans from STEM-research-intensive institutions to identify those areas where graduate deans are best positioned to integrate the unique ethical concerns of, and approaches to, big data in research into graduate training.

More specifically, within the workshop program, participants sought to:

1. Identify both broad and specific ethical challenges that arise from the use of big data resources in graduate student research;

2. Critique existing resources for training in the ethical use of big data;

3. Identify potential levers for introducing and discussing these challenges, and for influencing ethical cultures within an institution; and

4. Formulate potential strategies for deploying and embedding resources for big data ethics within the RCR training curriculum.

WORKSHOP DAY 1: "THE CHALLENGE OF TRAINING IN BIG DATA ETHICS: WHY IS IT HARD?"

The goal for day one of the workshop was to identify both broad and specific ethical challenges that arise from the use of big data methods and resources in graduate student research.

Michael Zimmer, PhD, an associate professor at Marquette University who specializes in data ethics and privacy issues, framed the discussion. Zimmer began by underlining the enormous power of big data in research: large databases and sophisticated data analytics have allowed detailed insight into a substantial variety of human activities, and the adoption of these approaches has become widespread among both commercial and academic research enterprises. In this presentation, Dr. Zimmer points out that accumulating data about people is easy and pervasive, yet there are no widely-used standards to ensure individual privacy, consent for human subjects research, or assessment of the potential harms that may result. He also provided some potential targets, such as embedded ethics training within the curriculum and updated RCR resources, for better training of graduate students and other researchers in the ethical use of these methods and

resources. In addition, he advocated for conversations with industry and government stakeholders to expand these discussions and concerns to a broader audience.

Summary of group discussions

After the keynote, participants broke into smaller groups to discuss a series of questions raised in Dr. Zimmer's talk.

Question 1: How aware are you of any big data research taking place at your institution that might spark some of the ethical challenges discussed in Zimmer's presentation? In which departments would these most likely take place and in what form?

Participants noted that departments that focus on computer technology and data analysis are the easiest to identify as utilizing big data: computer science and information science engage frequently in this kind of research. But big data research methods are often used across disciplines - for example, in learning science. Disciplinary silos can make it hard to identify where research using big data may be employed at the larger university level. In discussion, it was observed that graduate students are an important source of information about training needs. While IRBs might help universities identify where large datasets are being used, they frequently don't require formal approvals for the use of "public" datasets and may not be aware of the range of research using these resources. Funders and publishers also have insights into the use of big data and might be checkpoints for ethics requirements; however, publishers generally do not inform the university when they discuss research manuscript issues with authors. CITI modules may provide opportunities for distribution because so many universities use them for RCR education. All of these options were seen as preferable to no action. The worst-case scenario, participants agreed, is when the university only notices big data research ethics problems because students are charged with violations of academic integrity or when they arise within other misconduct processes.

Question 2: How prepared do you think your institution is to address the emerging ethical issues surrounding big data research? Does your RCR training address these concerns? Do you feel your faculty are prepared to guide students properly and your IRB is equipped to handle such protocols?

Discussion participants noted that one challenge is the preparation of senior faculty to deal with big data ethics. Senior faculty train a large number of graduate students through research mentoring, but if they don't possess adequate knowledge of these issues, their students might not be well trained in the ethical use of these methods either.

It was also agreed that data ethics should be expanded within RCR courses. Current RCR training in these issues is minimal. Some institutions have made substantial progress in scalable team education for RCR, and their approach may be a model for other institutions. Funding agencies could also play a role in strengthening training requirements by issuing new RCR guidance/requirements (or for training grants to develop guidance). There may also be issues with ethics training for values/ideas not covered well by current RCR and IRB training resources, such as the risks of de-identification of participants and the potential for harm to those individuals.

Participants also raised the question of whether training would be more effective if research ethics instruction were tied to the use of particular research tools during graduate training. IRB staff may be natural allies in this effort because they evaluate research plans before data are collected. To be most effective in detecting lapses in the use of big data methods, IRB reviewers and staff may need further training in these ethical issues.

Question 3: What kind of resources would be most helpful to ensure your graduate students are properly trained about the challenges of big data ethics? What is the best way to engage and to ensure the full complexity and diversity of ethical concerns are properly communicated?

Institutions aspire to build cultures of ethical and academic integrity. As part of that social mission, graduate faculty, disciplinary societies, IRB offices, and research oversight and compliance offices all bear the responsibility of assuring that ethics training occurs and is robust. A shift in mindset beyond compliance is needed to ensure that research ethics are considered integral to the rigorous use of data analysis methods.

Participants also noted that training modules or course modules in big data ethics that are aimed at students are needed. Training should focus on their responsibility as researchers and what they can do to anticipate and avoid ethical issues in data science. A directory of data ethics resources (both existing resources as well as new examples and case studies that can be easily incorporated into courses or methods training) would be a good first step.

WORKSHOP DAY 2: "THE CHALLENGE OF TRAINING IN BIG DATA ETHICS: WHAT RESOURCES ARE AVAILABLE OR ARE NEEDED?"

The goal for day two of the workshop was to survey what existing resources are available for training in the ethical use of big data, to identify new resources that may be needed, and to strategize about ways to engage students in their use.

The second keynote presentation was delivered by Casey Fiesler, PhD, an assistant professor at the University of Colorado Boulder and an expert in technology policy and law. Fiesler described available resources and the challenges to introducing them effectively within disciplinary training. Her central point was that embedding these issues within a disciplinary context can prevent the silo-ing of "ethics" from research. Fiesler suggested a set of priorities for faculty and other leaders as they identify opportunities within the graduate training curriculum to engage students in ethical concerns. These priorities include: (1) integrating ethics content into the curriculum so that they are gaining exposure to these issues at the same time they are learning the methods, (2) creating shared resources between disciplines, (3) facilitating conversations among stakeholders, and (4) making experts in the field available during these conversations.

Summary of group discussions

Question 1: Where do students at your university typically receive training for research ethics? These might be specific classes in different departments, general training materials (e.g., through an IRB), or even outside resources.

Participants in the discussion noted that CITI training provides the most frequently-cited resources for research ethics training at a campus level. However, some departments provide their own RCR training programs. One avenue for increasing attention to big data might be to target departments already doing their own RCR work, as well as the increasing number of institutes and programs on campuses devoted to big data research. Providing opportunities for this ethics training within laboratory groups might be another venue for this instruction, because so much graduate student training occurs within research teams. There may also be opportunities to provide micro-credentials documenting the training in the ethical use of big data methods as part of a student's professional development.

Informal opportunities for ethics training were also considered. It was noted that Cornell University offers an "Ethics Cafe," a monthly discussion space for graduate students to raise issues they may be facing. Another suggestion was to provide "Ethics Dinners" with deans to signal the importance of the topic to institutional leadership.

Question 2: Where do students at your university typically receive training in data science or big data research methods? Is there overlap between the answer to this question and the previous one, for example in classes?

Discussion participants observed that there are many new efforts on campuses for data science training, many focused on technical tools and programming skills. Including ethical issues within the context of those skill development projects could further engage students in this concern. Courses in data science methods might be another place to integrate ethics training into professional skill development.

Question 3: If your university were going to improve education in big data ethics (whatever you might imagine that means), what kinds of resources do you think would help?

Several suggestions were made about resources that CGS and other external partners might provide:

a. Providers of online professional development could identify or develop basic modules that provide general knowledge and a range of case studies in big data ethics.

b. Such providers (or faculty working in consultation with them) could also present examples of ethical dilemmas that offer students opportunities to practice decision-making in the context of big data research methods. Disciplinary case studies could be particularly helpful. Materials should emphasize data ethics rather than big data ethics. Ethical issues emerge even when dealing with "small" datasets, and researchers might not recognize their data as "big" within their own discipline/area.

c. CGS could take a role in aggregating resources and making them available so that programs can avoid duplicating effort.

Participants also offered ideas about university-developed resources and policies:

a. Degree program approvals at the campus level might be a lever for increasing attention to ethics. Campus leaders could require ethics components for new program approvals.

b. Campuses could host seminars with experts doing research on ethics in particular data areas (health, political science, etc.).

c. Universities and their partners should consider research areas that have been successful in advancing cultural change around research ethics, such as gene therapy research and research using animals.

Looking ahead to the development of new resources, participants agreed that funding and faculty time are a challenge. In rapidly growing units such as data science, faculty are frequently already over-committed. Yet another challenge is the relative lack of agreed-upon norms in the big data space, although this is changing. Illustrations of potential harms can help with norm-setting.

WORKSHOP DAY 3: "HOW CAN GRADUATE DEANS ENGAGE WITH FACULTY TO EXPAND RCR TRAINING AND CURRICULA REGARDING BIG DATA?"

The goal for the final session was to identify potential levers for introducing and discussing the ethical challenges of big data with stakeholders and for influencing ethical cultures within an institution. We also sought to formulate potential strategies for deploying and embedding resources for big data ethics within the graduate training curriculum.

The final keynote presentation was delivered by Thomas Jeitschko, PhD, dean of the Graduate School at Michigan State University. Jeitschko began by reaffirming a message heard in previous presentations and discussions: research integrity is fundamental to the quality of scholarship and creative work. Training in academic integrity should follow best practices in research, not just compliance and rule following. In line with that principle, ethics training in big data should include content and examples that embed the training within the professional standards of the discipline. Big data are increasingly used in many disciplines, and the ethical use of these methods present the additional challenges of biased data, unanticipated harms that may arise from data release, and undercutting the privacy concerns of individuals. Dr. Jeitschko noted that it is important to engage faculty and other leaders in discussing how to address these issues in graduate student training both within and across disciplines.

Summary of group discussions

Question 1: Given that many of the ethical questions involved in big data research are not well understood and often not appreciated, what are possible means to raise awareness of big data research ethics questions across disciplines?

Several discussion participants observed that training materials should focus on helping students make connections between their data, which can often seem disembodied or decontextualized, and the people they represent. This can be a matter of language: instead of always referring to "samples," we might substitute the word "biopsies;" "scraping data from Twitter" can be referred to as "collecting opinions or data on human interactions."

Question 2: What resources should be centrally available to support programs and students, and which resources would be standing resources, ad hoc resources, or on demand?

Ideas generated in response to question two included:

- Creating a bank of ethical dilemmas to discuss in the context of disciplinary and ethical standards and values.
- Considering a continuing education model— for example, offering continuing credits/requirements for researchers using big data research methodologies as these approaches change over time. Micro-credentialing and badging could be one means to demonstrate completion of ethical training by students to document skills and knowledge for future employers.
- Finding local resources and experts on each campus, including identifying which individuals and units might serve as resources/knowledge repositories for local efforts.

- Providing templates for language detailing professional standards for inclusion in graduate handbooks. This is a little-used resource for norm-setting in programs.

- Ensuring that ethical behavior is modeled for graduate students and other researchers at all levels of leadership within the institution. The message should be that research ethics are not merely about compliance, but about professional development and academic excellence.

Question 3: How can faculty be engaged to facilitate discussions with students about big data ethics?

For this question, participants identified two potential tactics:

- Taking advantage of opportunities to seed discussion between mentors and students is a key challenge but also an opportunity. Graduate student handbooks should cue students to ask faculty about research ethics. Contracts and formal expectations for graduate student researchers might also ask students to engage faculty members on these issues.

- To engage faculty in addressing these ethical issues, CGS and others might convene a panel designed to allow faculty and students to anonymously raise questions about ethical challenges in big data.

CONCLUSION

By the conclusion of the workshop, there was general agreement that we need to better utilize existing resources and develop new ones that illustrate best practices in the use of big data methods and databases.

Graduate deans can assist and support institution-wide discussions of expanded training opportunities in the ethical use of these methods; these discussions should include a wide range of stakeholders, including faculty, students, IT staff, IRB review boards, and others. Discussions and training opportunities have the potential to strengthen the culture of research and academic integrity within institutions.

CGS has had a long-term commitment to supporting graduate deans by providing resources and identifying best practices in all areas of graduate training, including in academic integrity and the responsible conduct of research. The discussions in this workshop provided ideas and resources that have the potential to help the graduate education community expand training in the ethics of big data research. CGS will continue this conversation in its future communications with graduate deans and their staff, and through presentations at future meetings.

Research using big data has enormous potential to help us address pressing problems in and between a broad range of fields, but we must work together to ensure that researchers are prepared for both the risks and rewards of this work. As we look to the future, CGS and PERVADE welcome the insights and support of members of the broader research community.

Works Cited

1. SAS. (n.d.). *Big data: What it is and why it matters.* https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

2. National Academies of Sciences, Engineering, and Medicine. (2015). Appendix A: Categorized list of cybersecurity threats. *Guidebook on best practices for airport security.* Washington, DC: The National Academies Press. https://www.nap.edu/read/22116/chapter/12

3. Golder, S., Ahmed, S., Norman, G., & Booth, A. (2017). Attitudes toward the ethics of research using social media: A systematic review. *Journal of Medical Internet Research, 19(6),* e195. https://doi.org/10.2196/jmir.7082

4. Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2016). Presented at the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2016), San Francisco, CA.

5. Walker, G.E., Golde, C.M., Jones, L., Bueschel, A.C., & Hutchings, P. (2008). *The Formation of scholars: Rethinking doctoral education for the twenty-first century.* San Francisco, CA: Jossey-Bass.

6. Council of Graduate Schools. (2012). *Research and scholarly integrity in graduate education: A comprehensive approach.* Washington, DC: Author.

7. Kent, J. (2015, November). "Big questions about 'big data:' Perspectives from the 2015 global summit on graduate education." *GradEdge* vol 4 (9). Accessed at https://cgsnet.org/ckfinder/userfiles/files/2015%20Global%20Summit%20Event%20Summary_Kent(1).pdf.

8. Council of Graduate Schools. (2015). *The implications of big data for graduate education: A proposal for future action: Developed by and for graduate institutions.* Washington, DC: Author. Accessed at https://cgsnet.org/ckfinder/userfiles/files/WebVersion_Practical_Actions_2015_final2.pdf.