Practical Approaches to Big Data Privacy Over Time[1]

Micah Altman,[2] Alexandra Wood,[3] David R. O'Brien[4] & Urs Gasser[5]

DRAFT (December 23, 2016)

Abstract. Increasingly, governments and businesses are collecting, analyzing, and sharing detailed information about individuals over long periods of time. Vast quantities of data from new sources and novel methods for large-scale data analysis promise to yield deeper understanding of human characteristics, behavior, and relationships and advance the state of science, public policy, and innovation. At the same time, the collection and use of fine-grained personal data over time is associated with significant risks to individuals, groups, and society at large. In this article, we examine a range of long-term data collections, conducted by researchers in social science, in order to identify the characteristics of these programs that drive their unique sets of risks and benefits. We also examine the practices that have been established by social scientists to protect the privacy of data subjects in light of the challenges presented in long-term studies. We argue that many uses of big data, across academic, government, and industry settings, have characteristics similar to those of traditional long-term research studies. In this article, we discuss the lessons that can be learned from longstanding data management practices in research and potentially applied in the context of newly emerging data sources and uses.

1. Corporations and governments are collecting data more frequently, and collecting, storing, and using it for longer periods.

Commercial and government actors are collecting, storing, analyzing, and sharing increasingly greater quantities of personal information about individuals over progressively long periods of time. Advances in technology, such as the proliferation of GPS receivers and highly-accurate sensors embedded in consumer devices, are leading to new sources of data that offer data at more frequent intervals and at finer levels of detail. New methods of data storage such as cloud storage processes are more efficient and less costly than previous technologies and are contributing to large amounts of data being retained for longer periods

[1] The authors describe contributions to this essay using a standard taxonomy. *See* Liz Allen, Jo Scott, Amy Brand, Marjorie Hlava & Micah Altman, *Publishing: Credit Where Credit Is Due*, 508 Nature 312, 312–13 (2014). Altman provided the core formulation of the essay's goals and aims, and Wood led the writing of the original manuscript. All authors contributed to conceptualization through additional ideas and through commentary, review, editing, and revision. This material is based upon work supported by the National Science Foundation under Grant No. 1237235 and the Alfred P. Sloan Foundation. The manuscript was prepared for the *Identifiability: Policy and Practical Solutions for Anonymization and Pseudonymization* workshop, hosted by the Brussels Privacy Hub of the Vrije Universiteit Brussel and the Future of Privacy Forum in Brussels, Belgium, on November 8, 2016. The authors wish to thank their colleagues through the Privacy Tools for Sharing Research Data project at Harvard University for articulating ideas that underlie many of the conclusions drawn in this essay.

[2] MIT Libraries, Massachusetts Institute for Technology, escience@mit.edu

[3] Berkman Klein Center for Internet & Society, Harvard University, awood@cyber.harvard.edu

[4] Berkman Klein Center for Internet & Society, Harvard University, dobrien@cyber.harvard.edu

[5] Berkman Klein Center for Internet & Society, Harvard University, ugasser@cyber.harvard.edu

of time.[6] Powerful analytical capabilities, including emerging machine learning techniques, are enabling the mining of large-scale datasets to infer new insights about human characteristics and behaviors and driving demand for large-scale data sets. Enabled by these technological developments, data related to human activity are measured at more frequent intervals, the personal data being collected and stored increasingly describe longer periods of activity, and the length of time that has elapsed between the collection and analysis of personal data can vary significantly. In addition, databases are becoming increasingly high-dimensional, meaning the number of pieces of information collected and associated with the record for each individual in a database is growing. Moreover, the analytic uses of data and samples sizes are expanding with emerging big data techniques. Taken together, these factors are leading organizations to collect, store, and use more data about individuals than ever before, and are putting pressure on traditional measures for protecting the privacy of data subjects across different contexts.

1.1. Long-term collections of highly-detailed data about individuals create immense opportunities for scientific research and innovation.

Massive data from new commercial and government sources, as well as novel methods for large-scale data analysis, promise to yield a deeper understanding of human characteristics and behavior. Characteristics of the commercial and government data that are accumulating over time make it possible to paint an incredibly detailed portrait of an individual's life, making such data highly valuable not only to the organizations collecting the data but to secondary users as well. Data from these new sources are increasingly being made available to researchers, policymakers, and entrepreneurs, with the expectation that they will help support rapid advances in the progress of scientific research, public policy, and innovation.[7]

Across commercial and government settings, the data that are being collected, stored, and analyzed represent increasingly long periods of time and contain observations collected at increasingly frequent intervals, due to the powerful insights that can be derived from large, fine-grained data sets that are linked over time. Commercial big data are generated and used to provide goods and services to customers, as well as to support the analytics businesses use to improve their services based on the needs of their customers and make investment and other business decisions.[8] A wide range of companies, including telecommunications providers, mobile operating systems, social media platforms, and retailers, often collect, store, and analyze large quantities of data about customers' locations, transactions, usage patterns, interests, demographics, and more. In particular, many businesses use highly detailed personal information to provide targeted services, advertisements, and offers to existing and prospective customers. Governments are also experimenting with collecting increasingly detailed information in order to monitor the needs of their communities, from pothole and noise complaints to crime reports and

---

[6] *See* President's Council of Advisors on Science and Technology, Big Data and Privacy: A Technological Perspective, Report to the President (May 2014), https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

[7] *See* Executive Office of the President, Big Data: Seizing Opportunities, Preserving Values (May 2014), https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[8] *See generally* President's Council of Advisors on Science and Technology, Big Data and Privacy: A Technological Perspective, Report to the President (May 2014), https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

building inspection records, and to improve their responsiveness and delivery of constituent services.[9]

Commercial big data and government open data increasingly being put to new analytic uses as well. Long-term big data promise to yield significant gains in the commercial and government sectors, much like long-term longitudinal data collection has transformed research in the social and biomedical sciences. For example, one of the longest running longitudinal studies, the Framingham Heart Study, precipitated the discovery of risk factors for heart disease and many other groundbreaking advances in cardiovascular research.[10] Other longitudinal studies have also had profound impacts on scientific understanding in fields such as psychology, education, sociology, and economics.[11] The combination of longitudinal data, large-scale data from commercial and government sources, and big data analysis techniques, such as newly emerging machine learning approaches, promises to similarly shift the evidence base in other areas, including various fields of social science, in unforeseeable ways.[12]

1.2. Long-term data collections by corporations and governments are associated with many informational risks, and potentially a wider set of risks than those presented by traditional research data activities.

The collection, storage, and use of large quantities of personal data for extended periods of time has recently been the subject of legal and policy debates regarding the future of privacy in the big data era. However, these activities have long been occurring at a smaller scale in research settings—and in ways that have been closely studied, reviewed, and controlled. The Framingham Heart Study, for example, is a long-term longitudinal research study that shares many of the key features found in recent commercial big data and government open data programs, in terms of the extended period of data collection, retention, and use, the large number of attributes measured, and the large number of participants involved.

Long-term research studies into human characteristics and behaviors are associated with significant privacy-related harms due to the collection of a large volume of highly-sensitive personal information about individuals. These harms are arguably not very well understood on a study-specific basis, particularly as the harms and risks are expected to evolve in unanticipated ways over the course of a long-term study. In addition, the legal remedies available to victims are limited.[13] Despite such challenges, institutional review boards recognize that the nature of the personal information collected in a longitudinal study may be associated with a range of harms including psychological, social, economic, legal, and even physical harms. For instance, an individual who participates in a longitudinal study could be subject to harms such as a loss of employability, loss of insurability, price discrimination in the marketplace, embarrassment or other emotional distress, reputational losses among family, friends and colleagues, or even civil or criminal liability if that individual's sensitive information from the study is

[9] *See generally* Stephen Goldsmith & Susan Crawford, The Responsive City: Engaging Communities Through Data-Smart Governance (2014).
[10] Ralph B. D'Agostino, Sr., et al., *General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study*, Circulation (2008).
[11] Erin Phelps, Frank F. Furstenberg, & Anne Colby, Looking at Lives: American Longitudinal Studies of the Twentieth Century (2002).
[12] *See* David Lazer et al., *Computational Social Science*, 323 Science 721 (2009).
[13] *See* M. Ryan Calo, *The Boundaries of Privacy Harm*, 86 Ind. L.J. 1131 (2011).

disclosed. The harms that are of greatest concern in an ethics review are well-documented in the policies of institutional review boards and the human subjects review literature.

There are a number of key similarities in the characteristics of data managed across longitudinal research data, commercial big data, and government open data settings. For instance, there are a number of similarities in the general populations covered. The motivation driving privacy law, policy, and practice is generally the need to protect individuals and groups from informational harms related to measuring and sharing information about them. Long-term longitudinal research studies such the Framingham Heart Study, the Panel Study of Income Dynamics, Project Talent, and the Health and Retirement Study, among others, are designed to be representative of or more generally, to draw inferences about the general population in the geographic region covered. In other words, researchers are collecting and developing processes to protect information about the same people (sometimes literally) that are described in corporate and government databases.

Additionally, there are similarities in the measurements and methods employed. Research methodology involves the use of questionnaires, observations of behavior, and experiments and other categories of interventions. Similarly, corporations and governments use many of these same methods to generate data about their customers and constituents. For instance, while corporations may have employed interventions strategies infrequently in the past, the prevalence of such activities have increased through the reliance on A/B testing, microtargeting, and individual service customization in the context of commercial big data analytics. Similarities are also found in the types of personal attributes studied. For instance, researchers, companies, and governments collect data in substantially overlapping domains, from demographics, to opinions and attitudes, to readily observable behavior, including geolocation, spatiotemporal movement, and economic activity. Much like a longitudinal research study, companies that maintain large databases of information about individuals collect personal information about individuals such as their names, location and travel history, political preferences, hobbies, relationships to other individuals, purchase history, and more, and maintain individual records that enable the tracking of individual behavior and characteristics over time. Commercial data programs may have historically involved fewer direct measures of sensitive social topics such as sex, drugs, criminal behavior, and political views, compared to research studies investigating these topics from a social science perspective. However, increasingly personal information involving sensitive topics is collected in the commercial setting through web-based questionnaires and prompts, such as those utilized by online dating services or social media platforms. In addition, such information can often be readily inferred indirectly from the data being collected, including search queries, social media postings, or purchase histories.[14]

Uses across research, commercial, and government settings also involve similar domains and objects of inference. While it may have been once the case that commercial data involved a smaller domain of inference, for instance by focusing on consumer purchase behavior and its direct drivers, commercial data are now being used to make inferences about a practically unrestricted range of economic, social, and health factors and potential correlations with an individual's increased use of a product or service. In this

[14] *See, e.g.*, Michael J. Paul & Mark Dredze, You Are What You Tweet: Analyzing Twitter for Public Health, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011); Wu Youyou, Michal Kosinski, & David Stillwell, Computer-based personality judgments are more accurate than those made by humans, 112 Proceedings of the National Academy of Sciences 1036 (2014).

way, commercial data activities are increasingly resembling research uses of data. The object of inference is also shared across these settings. Most research data, including data from social science and biomedical research, is used for statistical purposes. That is, the research is used to make inferences about the population studied, rather than to make inferences about specific individuals. Corporations and governments frequently make inferences such as these as well, most notably for the purposes of market analysis, but also for developing models of consumer behavior that are used in targeted advertising and risk scoring.[15] Uses of personal information in the public sector similarly are used to create models that are used in predictive policing, for example.

At the same time, there are key differences in the risk factors and interventions at play. Corporate data in many cases present a wider array of risks than longitudinal research data, due to greater frequency of data collection, or the increased scope of data uses, among other factors, as discussed in detail below in Section 3.1. Despite the wider set of risks entailed, commercial data are subject to a set of privacy controls that is substantially different from—and arguably less finely tailored than—the controls used in research settings, as discussed in Section 2.2.

1.3. Long-term data activities generally increase identifiability and expand the range of harms to which individuals are exposed, and key driving characteristics are age, period, and frequency.

While long-term data activities have the potential to bring tremendous societal benefits, they also expose individuals, groups, and society at large to greater risks that their personal information will be disclosed, misused, or used in ways that will adversely affect them in the future. Growth in the time and scale dimensions of the data being collected and used can in many cases amplify the magnitude of potential informational harms.

Examples of longitudinal data collection and use in the research setting help illustrate ways in which long-term data activities drive heightened risks for data subjects. Despite variations in populations, types of information, and methods used in long-term research studies, such studies share certain defining characteristics, such as data collection activities that take place over a long period, with repeated measurements of the same set of individuals being linked throughout the length of the study period. Long-term research studies are generally associated with large numbers of observations about each data subject, and these data are frequently quite rich and complex. They may contain qualitative data, such as video or audio recordings, or a lengthy narrative account or other extended textual data. The richness of the data drives societal benefits, enabling scientific research into questions that cannot be explored using cross-sectional data. At the same time, data containing fine-grained records about individuals, such as their spatiotemporal location, internet browsing activity, or retail purchase history, are likely to be associated with sensitive information, such as a person's location and behavior, health history, interests, and associations, relationships, and activities with other people. Data are maintained at the level of an individual subject and used to track changes in each individual's health, socioeconomic, and behavioral

[15] *See, e.g.,* Pam Dixon & Robert Gellman, The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future (2014), http://www.worldprivacyforum.org/wp-content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf.

characteristics over an extended timeframe. Collection of such detailed information exposes individuals to potential harm to their reputations and personal relationships, risk of future loss of employability and insurability, risks of financial loss and identity theft, and potential civil or criminal liability, among other harms.

[Defined identifiability and sensitivity components -- referencing the Berkely paper. Forward reference to section 3.3 to managing through controls on object of inference, questions, use of answers…]

As discussed in detail in Section 3.1 below, a review of information privacy risks in longitudinal research suggest that the following three data characteristics related to time increase informational risks:

- *Age*, which is defined as the amount of time that has elapsed between the original data collection and a given data transformation or analysis. For instance, data may be analyzed shortly after collection, as in the case of a mobile app that targets an advertisement based on the user's current location, or data may be analyzed years after collection, as in the case of certain census records that are protected from disclosure for many decades.

- *Period*, referring to the length of the time interval within which subjects are repeatedly measured. Some data collection programs collect information at a single point in time, or over the period of a day, such as a cross-sectional statistical survey, while others may engage in collection over decades and even generations, such as a long-term longitudinal research study or a long-established social networking service.

- *Frequency*, or the interval between repeated measures on the same subject. Examples of high-frequency data collection include mobile health apps and devices that continuously track data points such as location and heart rate. On the other end of the spectrum, some health studies may collect data from participants once a year, or once every several years.

- *Dimensionality*, or the number of independent attributes measured for each data subject. Examples of high-dimensional data include commercial big datasets of thousands of attributes that are maintained by data brokers and social media companies, while low-dimensional data may include administrative datasets maintained by government agencies including just a handful of attributes within each record.

- *Analytic use,* or the mode of analysis the data is intended to support. Research studies are almost universally designed to support descriptive or causal inferences about populations. In contrast, commercial and government entities often have broader purposes, such as making inferences about individuals themselves, or performing interventions such as recommending products.

- *Sample size*, referring generally to the number of people included in the set of data, can act to increase some privacy risks by increasing both identifiability and the likelihood that some particularly vulnerable people are included in the data. Big data collection in corporate and government settings typically have much larger sample sizes than traditional longitudinal

research studies, and the samples typically constitute a substantially larger proportion of the population from which the sample is drawn.

- *Population characteristics* refer to the size and diversity of the population from which observations in the data are drawn. Most longitudinal research studies are drawn from national populations or identified subpopulations. Increasingly big data in corporate settings describe multinational or global populations. The diversity of the population included in the data increases privacy risk by expanding the range of threats that are relevant to the subjects described in the data.

The composition of even simple, seemingly benign measures over time can create unique patterns, or "fingerprints" of behavior that can be used to associate a named individual with a distinct record in the data, thereby increasing the identifiability of the data. With high-dimensional datasets, where many independent measures are associated with a single individual, an individual's record is likely to contain patterns that are unique to that individual, making the record identifiable. Datasets containing a large number of attributes about each individual are also likely to contain sensitive information. Examples of high-dimensional datasets include records held by data brokers like Acxiom, which may contain thousands of independent attributes about an individual's demographic characteristics, retail purchases, and interests, and social media providers like Facebook, which reveal details about individuals' personal relationships, sexual preferences, political and religious views, location, activities, and interests.

With the move towards long-term data collection, storage, and analysis in the commercial and government sectors, databases tend to grow along each of these dimensions identified. Moreover, risks grow with long-term storage and use of the information, as information held in storage for extended durations will be increasingly vulnerable to intentional and accidental data breaches. In addition, analytical capabilities for re-identification or learning sensitive information about individuals in a database will improve over time. Such risks are to a large extent unforeseeable, particularly as the timescale over which data will persist after they are collected is indefinite. Over the full course of the collection, use, and retention of this personal information, how the information will be used will evolve in unpredictable ways. As discussed below, the current state of the practice for privacy protection addresses (and also fails to address) these developments in important ways, and the risks that remain can be instructive for understanding where new interventions should be employed in the future.

2. Current accepted practices for protecting privacy in long-term data are highly varied across research, commercial, and government contexts.

Businesses, government agencies, and research institutions have adopted various approaches in order to protect privacy when collecting, storing, using, and disclosing data about individuals. While practices vary widely across organizations, among the most common approaches used in commercial and government contexts are notice and consent mechanisms and de-identification techniques, and these approaches often employed without further review or restriction on use of data that have been obtained according to a terms of service agreement or that have been nominally de-identified. This reliance on a narrow set of controls, without continuing review and use of additional privacy interventions, stands in contrast to the longstanding data management practice in research settings. For decades, researchers and

institutional review boards have intensively studied the long-term risks of collecting and managing personal data about research subjects, and have developed extensive review processes and a large collection of techniques for protecting the privacy of research subjects. When compared to the best practices that have been developed to safeguard research data, common corporate and government data management practices represent a narrow subset of the interventions that have been used to address challenges associated with managing privacy in long-term data activities in the research setting.

2.1. Privacy practices in long-term research studies incorporate multiple layers of protection, including explicit consent, systematic review, statistical disclosure control, and procedural controls.

Long-term research studies are carefully designed by their investigators and reviewed by an institutional review board (IRB). Among other requirements, the policies and practices implemented in a study are chosen to limit the collection of data to that which is necessary to achieve the purposes of the research, to restrict future uses of the data to those specified at the outset of the study, and to minimize the disclosure of personal information collected over the course of the study. In order to satisfy these and related requirements, researchers utilize a wide range of procedural, legal, technical, and educational controls, including explicit and informed consent, systematic review over time, statistical disclosure control techniques for limiting learning information specific to individuals in the data, legal instruments such as data use agreements, and various procedural controls for limiting access to and use of personal data.

2.1.1. Legal and regulatory frameworks for the oversight of human subjects research, in combination with sector-specific information privacy laws, leads to systematic design and review of longitudinal research studies and management of the research data produced.

Ethical and legal frameworks have been developed and adapted over time to provide strong privacy protection for human subjects participating in research studies. Researchers are generally subject to the federal regulations for human subjects research protection, known as the Common Rule,[16] which apply to research that is funded by one of the federal agencies that have subscribed to the rule or that is conducted at an institution that has agreed to comply with the regulations. Researchers are often also governed by state laws protecting the rights of human research subjects,[17] or by the research data policies of their home institutions, sponsors, and prospective journals. As a consequence, researchers conducting research involving human subjects typically must submit their project proposals for review by an IRB and follow certain consent and disclosure limitation procedures. During the review process, members of the IRB inspect materials such as a description of the proposed research study, copies of the surveys or other documents to be used to collect information from subjects, documentation of subject recruitment procedures, copies of consent forms, documentation of de-identification protocols, and documentation of any HIPAA authorizations or waivers obtained, among other documents requested. In addition, researchers must demonstrate that research subjects will be informed of the nature, scope, and purpose of the study; the types of personal information to be collected; the research and data management procedures

[16] 45 C.F.R. Part 46.
[17] *See, e.g.,* California Health and Safety Code § 24170 et seq.

to be followed; the steps to be taken to preserve confidentiality; and any risks and benefits related to participation in the study. On the basis of the submitted materials, the members of the IRB evaluate the informed consent procedures and potential risks and benefits to the research subjects. They also consider whether the subjects are adequately informed of the potential risks and whether the benefits outweigh the risks. At the conclusion of the review, the IRB makes a determination regarding the proposed study's adherence to applicable statutes and regulations, institutional policies and commitments, best practices, and ethical obligations. In a long-term research study, IRBs conduct continuing review, with reviews conducted on an annual or more frequent basis, to consider relevant information received since the previous review, to the extent it is related, for example, to risk assessment and monitoring, the adequacy of the process for obtaining informed consent, developments regarding the investigators and their home institutions, and evaluations of the progress of the research with respect to, for example, the submitted research protocol, and total subject enrollment and subject withdrawals.[18]

Because research studies often explore sensitive topics related to the development, behavior, and health of individuals, the data collected often fall into a category of information protected by at least one of the various information privacy laws in place. A large number of federal and state laws in the United States protect data privacy, though the rules vary substantially across sectors and jurisdictions depending on the actors, funding sources, types of information, and uses involved. For example, where researchers seek to obtain high school and post-secondary transcripts, medical records, or substance abuse treatment records, different rules come into play under the Family Educational Rights and Privacy Act (FERPA),[19] the Health Insurance Portability and Accountability Act (HIPAA),[20] and the federal alcohol and drug abuse confidentiality regulations,[21] respectively. In addition, researchers collecting data on behalf of federal agencies are in some cases required to establish certain privacy safeguards in order to comply with laws such as the Privacy Act of 1974[22] or the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[23]

2.1.2. Longitudinal research studies rely on a wide range of legal and procedural controls for the protection of human subjects.

A wide range of legal, procedural, and technical controls are employed at various stages of a long-term research study. Legal instruments, especially, play a significant role in governing human subjects research and data management, particularly in the areas of consent, access, use, and redisclosure. Most notably, consent from research subjects is preserved in the form of a written, legally enforceable contract that documents the scope of the study authorized. Future uses of the data are limited to the research purposes and scientific questions described in the consent form, and the language of the consent form also limits the scope of a researcher's interaction or intervention with a subject. Over the course of a long-term study, circumstances are likely to evolve in ways that are in tension with the consent previously obtained from the research subjects. For example, when the scope of information collected or analyzed is

[18] 45 CFR 46.109(e).
[19] 20 U.S.C. § 1232g; 34 C.F.R. Part 99.
[20] 45 CFR Part 160 and Subparts A, C, and E of Part 164.
[21] 42 C.F.R. Part 2.
[22] 5 U.S.C. § 552a.
[23] Pub. L. 107-347, Title V; 44 U.S.C. § 3501 note.

expanded, the study extends beyond the timeframe initially disclosed to the study participants, or the membership of the population being studied changes in some way, the investigators must obtain new consent from the participants.[24] Other measures are also put in place to address such changes over an extended period of data collection and use; for instance, some long-term studies direct participants to designate a proxy who is authorized to provide consent to new research uses in the event of future lack of capacity to provide such consent.

Modifications to research plans and consent forms are common over the course of a long-term research study. Such modifications may be made as new research questions emerge and stimulate interest in new categories of data uses and collections. For example, approximately 35 years into the Framingham Heart Study, researchers began collecting and analyzing DNA from participants' blood samples and immortalized cell lines, due to a growing interest within the scientific community at the time for exploring the genetic factors underlying cardiovascular disease.[25] Moreover, data collected in a long-term research study are often rich enough to support analysis methods and research questions not originally envisioned at the time that the original research proposal was drafted. To manage consent over time, researchers maintain a detailed record of consent for each subject and, upon each new data collection or use activity, research confirm whether the consent on file authorizes it. For instance, when Framingham Heart Study participants complete their consent forms, they are asked to mark checkboxes to provide separate authorization for each of a large number of potential research activities, including cell line creation, sharing of genetic data with researchers, and sharing of genetic data with private companies.[26] Consent forms also enable participants to authorize use for specific research purposes, with checkboxes for categories such as research into heart and blood diseases, research related to other diseases and conditions, and potentially sensitive research involving reproductive health, mental health, and alcohol use.[27] Research staff encode the individual responses from the subjects, maintain them in a database, and refer to the permissions before each interaction with one of the subjects or their data.[28]

Data use agreements limiting access to and use of a given dataset are very common, and have been widely adopted by academic institutions, data repositories, and data enclaves. Data use agreements typically describe the contents and sensitivity of the data; the restrictions on access, use, and disclosure; the data provider's rights and responsibilities; the data confidentiality, security, and retention procedures to be followed; the assignment of liability between the parties; and relevant enforcement procedures and

---

[24] National Bioethics Advisory Commission, Ethical and policy issues in research involving human participants. Report and recommendations of the National Bioethics Advisory Commission (2001), https://bioethicsarchive.georgetown.edu/nbac/human/overvol1.pdf; National Bioethics Advisory Commission, Research involving human biological materials: Ethical issues and policy guidance. Report and recommendations of the National Bioethics Advisory Commission (1999), https://bioethicsarchive.georgetown.edu/nbac/hbm.pdf.

[25] Diddahally R. Govindaraju et al., *Genetics of the Framingham Heart Study Population*, 62 Advances in Genetics 33 (2008), http://www.ncbi.nlm.nih.gov/pubmed/19010253.

[26] Daniel Levy et al., *Consent for genetic research in the Framingham Heart Study*. 152A American Journal of Medical Genetics Part A 1250 (2010), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2923558.

[27] Daniel Levy et al., *Consent for genetic research in the Framingham Heart Study*. 152A American Journal of Medical Genetics Part A 1250 (2010), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2923558.

[28] Greta Lee Splansky, Sue Blease, & Moira M. Pryde, *Tracking individual consent histories for participants in research*, Boston University Medical Center Clinical Research Times (April 2012), https://wwwapp1.bumc.bu.edu/ocr/ClinicalResearchNewsletter/article.aspx?article=394.

penalties. They are used to set forth obligations, ascribe liability and other responsibilities, and provide a means of recourse if a violation occurs. Common contractual approaches to enforcement include sanctions such as denial of future access to data files, reporting of violations to federal agencies or a researcher's institution or funders, and fines and other statutory penalties. However, oversight and enforcement of the terms of such agreements is a persistent challenge.

2.1.3. Longitudinal research studies rely on technical controls, such as statistical disclosure limitation techniques, synthetic data, differential privacy tools, and secure data enclaves, for protecting data collected from human subjects.

Technical approaches, such as the use of data security measures and statistical disclosure limitation techniques, are often used in the long-term research setting, though there is significant variation in practice. Among the most common practices is for research institutions to implement security plans and confidentiality training programs, given the large number of individuals who will likely have access to some personal data over the course of the study. In addition, best practices for data security are generally mandated by sponsors of research, such as government agencies, academic institutions, and foundations, and such institutions may prescribe specific guidelines for researchers to follow. Researchers often transform data at the collection and retention stages using techniques such as encrypting, hashing, or re-coding of personal identifiers to limit disclosure when linking and storing data between waves of data collection in a longitudinal study, while preserving the ability of certain researchers to access the personal identifiers when needed.

Multiple disclosure limitation techniques are often used in combination to protect the privacy of research subjects when sharing long-term research data. A common approach is to use tiered access to make a de-identified dataset available to the public, while offering a separate restricted-use dataset to trusted researchers upon application. In many cases, the sensitive nature of the data and the potential to draw linkages to external sources lead to such high risk that it precludes the dissemination of the data in raw, identifiable form. In such cases, researchers use a variety of statistical disclosure limitation techniques, such as aggregation, suppression, and perturbation, to produce a de-identified public-use dataset.[29] However, such techniques are less desirable for longitudinal data. Because longitudinal studies collect data at the level of an individual human subject for the purposes of studying patterns in individual behavior, data that have been aggregated or summarized in, for example, cross-tabulation tables, are often not suitable for analyses not anticipated by the researcher who produced the aggregate dataset. For this reason, researchers often prefer to receive data in individual-level, rather than aggregate, form. Individual-level data can better answer complex questions, test alternative hypotheses, calculate marginal effects of changes over time, identify errors in the data, and replicate results by other researchers.[30] As a

[29] Aggregation involves rounding and top-coding certain values to make them less precise; suppression entails removing some of the most sensitive data from a dataset before sharing it with others; and perturbing means altering some of the data, such as by introducing noise or by swapping some of the values. *See, e.g.,* Bureau of Labor Statistics, National Longitudinal Surveys: Frequently asked questions (2014), from http://www.bls.gov/nls/nlsfaqs.htm. For a survey of commonly-used statistical disclosure limitation methods, see Federal Committee on Statistical Methodology, Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22 (2005), https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf.

[30] *See* Judith D. Singer & John B. High dim, Applied Longitudinal Data Analysis (2003).

result, investigators often share data from long-term longitudinal studies in files containing linked individual-level records.

De-identification approaches, such as aggregation, suppression, and perturbation, address some risks, but there is a growing recognition that such techniques will provide only limited protection over the long term.[31] Traditional de-identification techniques are generally designed to address only certain types of attacks such as record linkage attacks using known sources of auxiliary information, leaving data vulnerable to other types of attacks. Common de-identification approaches are also likely to result in the redaction or withholding of useful information. Moreover, the tradeoff between data privacy and data utility, or the analytic value of the data, is generally more acute with respect to long-term longitudinal data. Models for assessing disclosure risk have typically been developed with cross-sectional data, i.e., data collected at one point in time or without regard to differences in time, in mind, and therefore are poorly suited for addressing longitudinal data privacy risks.[32] As a consequence, traditional statistical disclosure limitation techniques that are effective for cross-sectional datasets often result in either weaker privacy protections or a greater reduction in data utility when applied to longitudinal data.[33] These techniques are likely to change the structure of longitudinal data in ways that sharply influence the statistical models and inferences made by an analyst and may make certain types of modeling or analysis impossible. Compounding this problem is the fact that the types of transformations that are made to a dataset for the purposes of limiting disclosures of information specific to individuals in the data are not always disclosed to the public. Secondary researchers sometimes unwittingly treat a sanitized data set as an unmodified data set, leading to unanticipated and unacknowledged effects on the results of their analyses.

Newly-emerging technical approaches, such as synthetic data generation and differential privacy, are less widely utilized to protect the confidentiality of long-term longitudinal research data, though there is reason to believe such techniques may be developed to better preserve the complex relationships between variables. Differential privacy, for instance, has typically been studied in the context of a dataset that has been gathered and released, either as a single publication or interactively in response to queries from users. To date, there are few differentially private algorithmic results that apply to the setting that is typical to longitudinal research studies, in which datasets are collected over time and analyzed as the data are acquired. Although a similar model, the continual observation model,[34] is flexible enough to describe longitudinal studies, research to date has typically assumed that one person's information affects only a

[31] *See* Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, Proceedings of the 2008 IEEE Symposium on Security and Privacy 111 (2008), http://dl.acm.org/citation.cfm?id=1398064.
[32] *See* Lawrence H. Cox, Alan F. Karr, & Satkartar K. Kinney, *Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act*, 79 International Statistical Review 160 (2011), http://www.niss.org/sites/default/files/tr179_final.pdf.
[33] *See* Khaled El Emam & Luk Arbuckle, *Longitudinal discharge abstract data: State inpatient databases*, in Anonymizing health data: Case studies and methods to get you started (2013); Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, & Philip S. Yu, Introduction to privacy-preserving data publishing: Concepts and techniques (2010).
[34] Cynthia Dwork, Moni Naor, Toniann Pitassi, & Guy N. Rothblum, *Differential privacy under continual observation*, Proceedings of the forty-second ACM symposium on Theory of computing 715 (2010), http://dl.acm.org/citation.cfm?id=1806787.

limited number of stages of the study.[35] Existing literature does not directly address the design of methods that release accurate statistics about each stage of a long-term study nor detailed information about how the statistics are evolving over time. Despite these challenges, there are promising ways in which such advanced techniques could potentially be applied to releases of longitudinal data. By releasing synthetic data, or simulated microdata, researchers may be able to reduce disclosure risks while retaining validity for certain inferences that are consistent with the model used for synthesis.[36]

Due to the difficulty of anticipating and mitigating the risks that will be associated with future releases of data over the course of a long-term study, researchers often respond by implementing restrictive access controls. Restricted access regimes, such as data repositories or secure enclaves that limit access and use according to terms of use or data use agreements, are used to manage access rights and conditions when sharing data with project collaborators and secondary researchers. Data holders may require secondary researchers to submit applications requesting access to the data, and limit access to certain classes of researchers, such as faculty-level researchers or researchers working under a federal pledge of confidentiality. In addition, researchers may be required to agree to participate in confidentiality training or to implement and demonstrate compliance with a data security plan. In some cases, particularly for large research projects, researchers may be granted access to data only through physical or virtual data enclaves, which restrict and monitor uses of data in a controlled setting.

In combination, the regulations and ethics frameworks for human subjects research protection; the range of information privacy laws governing the collection, use, and disclosure of personal information in various contexts; the guidelines from IRB, university data management, funding agency, and journal policies; the oversight of IRBs and other ethics bodies; and best practices for the use of legal and technical tools for protecting the privacy of research subjects work in concert to ensure researchers carefully consider the risks and benefits of their research activities to individual subjects as well as to society, and choose among a wide variety of interventions to protect subjects appropriately. These regulations, policies, and practices explicitly recognize and seek to address, through continuing review and curation, the substantial risks associated with long-term data activities.

2.2. Industry and government actors rely on a narrow subset of the privacy controls used in research, with a notable emphasis on notice and consent mechanisms and de-identification techniques.

The review processes and safeguards employed for long-term data collection and linkage activities in commercial and government settings differ from those used in the research context in a number of key respects. For instance, businesses and government agencies generally consider privacy risks at the time they initiate a data collection program, but in most cases they do not engage in systematic and continual

[35] *See, e.g.,* T.-H. Hubert Chan, Elaine Shi, & Dawn Song, *Private and continual release of statistics*, 5 ACM Transactions on Information and System Security A:1 (2011), https://eprint.iacr.org/2010/076.pdf; Prateek Jain, Pravesh Kothari, & Abhradeep Thakurta, *Differentially private online learning*, 23 Proceedings of the 25th Conference on Learning Theory 24.1 (2012), http://www.jmlr.org/proceedings/papers/v23/jain12/jain12.pdf.
[36] *See* Ashwin Machanavajjhala et al., *Privacy: Theory meets practice on the map*, Proceedings of the 2008 IEEE 24th International Conference on Data Engineering 277 (2008), http://www.cse.psu.edu/~dkifer/papers/PrivacyOnTheMap.pdf.

review with long-term risks in mind. In addition, commercial and government actors often rely heavily on certain approaches, such as notice and consent or de-identification, rather than drawing from the wider range of privacy interventions that are available and applying combinations of tailored privacy controls at each stage of the information lifecycle, from collection, to retention, analysis, release, and post-release.

2.2.1. Long-term data activities in industry and government settings are often subject to less comprehensive and detailed regulatory requirements than those conducted in research.

Practices across research, industry, and government contexts emerged and evolved under very different regulatory and policy conditions. Variations in practice are accordingly due in large part to differences in the legal frameworks and institutional constraints that apply in these different settings. Sector-specific information privacy laws such as FERPA and HIPAA, among others, which play a significant role in protecting research data, apply directly to only a small portion of commercial and government data activities. However, some companies and government agencies elect to adopt as a best practice some of the safeguards required by such laws. In addition, some commercial and government data activities are governed by other laws that rarely apply to researchers, such as the Fair Credit Reporting Act,[37] the Children's Online Privacy Protection Act,[38] and Federal Trade Commission (FTC) enforcement under Section 5 of the FTC Act.[39] Laws such as the Privacy Act of 1974,[40] the Confidential Information and Statistical Efficiency Act,[41] and the Freedom of Information Act,[42] as well as corresponding laws governing public records at the state level, govern data collection, storage, and release activities involving certain categories of information maintained by federal and state agencies. Individual omnibus state privacy laws, such as the various data breach notification laws in place across the country,[43] may also be applicable to commercial actors. Such laws typically require limited safeguards, by restricting the collection, storage, and disclosure of certain pieces of directly identifying information such as names and Social Security numbers, and may apply to commercial or government entities, or both. Agencies also implement data security standards, such as those required by the Federal Information Security Management Act (FISMA),[44] and established by bodies such as the National Institute of Standards and Technology.[45] These laws and standards grant substantial discretionary authority to agencies, leading to wide variations in practice.

Generally, large commercial actors have adopted and implemented reasonably complex procedural and technical data security practices when required by law or industry standard. Industry best practices, such

[37] 15 U.S.C. § 1681.
[38] 15 U.S.C. §§ 6501–6506.
[39] 15 U.S.C. § 45.
[40] 5 U.S.C. § 552a.
[41] Pub. L. 107-347, Title V; 44 U.S.C. § 3501 note.
[42] *See* Freedom of Information Act, 5 U.S.C. § 552, and corresponding sunshine laws at the state level.
[43] *See, e.g.,* Cal. Civ. Code §§ 1798.29, 1798.80 et seq.; Mass. Gen. Laws § 93H-1 et seq.; N.Y. Gen. Bus. Law § 899-aa, N.Y. State Tech. Law 208.
[44] 44 U.S.C. §§ 3541 et seq.
[45] *See, e.g.,* NIST, FIPS Pub. 199, Standards for Security Categorization of Federal Information and Information Systems (2004).

as the Payment Card Industry Data Security Standard,[46] and the Health Information Trust Alliance (HITRUST) framework,[47] are widely applied, though they focus almost exclusively on data security requirements rather than privacy protections, meaning they are designed to restrict access to personal information rather than limit what can be learned about individuals from their data once such access has been granted. In addition, industry actors frequently rely on the fair information practice principles for guidance.[48] These principles are often referenced at a high-level, rather than establishing common practices through detailed requirements. The most extensive implementation of these principles is likely found in credit reporting, an industry with a long history and significant experience with handling large quantities of highly-sensitive information about individuals, and is governed by regulations focusing on the inspection of personal data and restrictions on disclosure. These principles have also guided some specific practices in some sectors, such as the hashing of IP addresses and device identifiers in the context of mobile device privacy,[49] and general policies such as the National Information Standards Organization (NISO) privacy principles for libraries (which have not yet been implemented in practice, to our knowledge).[50] While these general principles are widely referenced in many contexts, they are often not clearly implemented with specific and strong policies and controls.

2.2.2. Long-term data activities in industry and government settings rely on less systematic reviews of privacy risks and a narrow subset of privacy controls compared to the practices found in research settings.

The regulatory framework has not yet evolved to address the privacy and ethical challenges presented by the rise of big data, and commercial big data activities are conducted within the gaps that exist. Overall, review processes are sparsely documented, and the documentation that has been made publicly available is typically written at a high level that does not reveal specific details about how risks are evaluated against benefits in practice, in contrast to the extensive documentation and guidance that has been developed for the oversight of human subjects research. Moreover, although commercial firms and government agencies have implemented measures to address data privacy risks, approaches in widespread use represent a limited subset of the privacy protection techniques that are available. For instance, it is a common practice when collecting, storing, and sharing data about individuals to protect privacy by de-identifying data through the removal of pieces of information considered to be personally identifiable, such as names, addresses, and Social Security numbers.

---

[46] *See* Payment Card Industry (PCI) Data Security Standard, Requirements and Security Assessment Procedures, Version 3.2 (April 2016), https://www.pcisecuritystandards.org/documents/PCI_DSS_v3-2.pdf.
[47] The Health Information Trust Alliance (HITRUST) Common Security Framework (2016).
[48] *See, e.g.,* Testimony of Jeremy Cerasale, Senior Vice President of Government Affairs, Direct Marketing Association, Senate Committee on Commerce, Science, & Transportation Hearing on "What Information Do Data Brokers Have on Consumers, and How Do They Use It?" (Dec. 18, 2013), https://www.commerce.senate.gov/public/_cache/files/2432676f-c42e-45ae-be58-6b3b87f1cab3/08578EF6CD1EEE441ED608FD70EFF46A.cerasale-testimony.pdf.
[49] *See* materials from the Federal Trade Commission workshop "Spring Privacy Series: Mobile Device Tracking" (Feb. 19, 2014), https://www.ftc.gov/news-events/events-calendar/2014/02/spring-privacy-series-mobile-device-tracking.
[50] *See* NISO Consensus Principles on User's Digital Privacy in Library, Publisher, and Software-Provider Systems (NISO Privacy Principles) (Dec. 10, 2015), http://www.niso.org/apps/group_public/download.php/16064/NISO%20Privacy%20Principles.pdf

The legal and ethical framework is not very well-defined for many commercial big data activities, and, as a result, commercial data collection, use, and disclosure are relatively less regulated than activities involving human subjects research data, for which such frameworks are well-established. The informed consent and IRB review processes required by the Common Rule do not apply to corporate entities, except in cases where they are engaging in generalizable research that is funded by a federal agency that has subscribed to the regulations. The lack of formal review and oversight by an external ethics board such as an IRB, and the absence of other governance mechanisms to serve such a role, results in less of an emphasis on informing subjects of risks and benefits, minimizing data collection and disclosure, and implementing controls to address long-term risks in commercial settings, relative to research settings. Unlike research subjects, the subjects of commercial data collection often lack an understanding of the full extent to which data about them are collected, linked, analyzed, shared with, and re-used by third parties. Data collected from research subjects are typically not linked with data from other sources, except in limited ways that are specified at the initial design stage of the study, reviewed and approved by an independent ethics board, and disclosed and consented to by the individual subjects. However, in commercial contexts, data are frequently linked and combined with data from other sources and often redisclosed for use by third parties. These practices are typically authorized by the data subjects through privacy policies or terms of service agreements, which contain broad language regarding the collection, use, and disclosure of personal information, and these terms are often not reviewed closely, if at all, by consumers, which stands in contrast to the informed consent process used in research.

In the government context, agencies are generally directed to take into account the legal and ethical implications of disclosing information about individuals, and to review their data collection, storage, and disclosure practices and to implement appropriate privacy safeguards. However, applicable laws are context-specific and limited in scope, and lack specificity regarding the application of appropriate privacy and security measures in a particular setting.[51] There are some cases, such as data collection programs initiated by statistical agencies under a pledge of confidentiality or programs which require a privacy impact assessment to be completed, where a review of informational risks and benefits may be performed, individuals may be informed of risks and benefits in advance of data collection, and processes for minimizing data collection and controlling data uses and disclosures may be implemented. However, many other agency activities using personal data, particularly open data initiatives which call for open access to be the "default state" for information and instruct agencies to proactively release data to the extent the law allows,[52] involve less systematic consideration of long-term risks and may lead to the public release of data collected for one purpose, such as delivering constituent or emergency services in the case of 311 or 911 call data, opening it up to use by the public for any purpose. These policies are, in large part, enabled by sunshine laws at the federal and state levels, including the Freedom of Information Act,[53] which require disclosures in response to public records requests provided that no law prohibits the

[51] For a discussion of the limited nature of the practical guidance on applying privacy protections that is provided to government agencies, see Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 Berkeley Tech. L.J. 1967 (2015).
[52] *See, e.g.,* Exec. Order No. 13,642, 3 C.F.R. 244 (2014) (Making Open and Machine Readable the New Default for Government Information), https://www.gpo.gov/fdsys/pkg/CFR-2014-title3-vol1/pdf/CFR-2014-title3- vol1-eo13642.pdf.
[53] *See* 5 U.S.C. § 552.

release. Agencies are granted significant discretionary authority to withhold or redact records that implicate one of a limited set of concerns such as privacy, and they typically do so by redacting records of direct identifiers such as names, addresses, dates of birth, and Social Security numbers. However, due in part to the lack of detailed guidance, redaction processes are typically performed in an ad hoc fashion and practices vary significantly between agencies. Moreover, in freedom of information and open data releases, the extent to which individuals are informed of the benefits and risks associated with the inclusion of their information in the data prior to release is often unclear.

Commercial and government data collection activities are generally not designed with long-term future uses of the data taken into consideration. In practice, when initiating data collections, most companies and government agencies do not implement procedures for long-term review of risks and measures to mitigate risks associated with long-term storage, use, and disclosure. In most cases, companies begin by collecting some data, and the data accumulate and are combined and linked over time. While policies at large data companies are evolving to address these concerns, common practices in this area generally fall short of the study design and independent ethics review that are characteristic of longitudinal studies in the research setting. The harms associated with commercial big data collection and analysis programs have arguably not been studied and debated to the same extent that the harms associated with long-term longitudinal research studies have been considered by institutional review boards and principal investigators. Commercial and government agencies often emphasize practices such as data destruction as a risk mitigation technique. Although many data management plans rely on data destruction as a technique for protecting privacy, this approach should not be considered sufficient for eliminating privacy risks. While data destruction should be considered as part of a data management program, it should not be considered an effective strategy for eliminating risk, as deleting data does not mitigate all risks if the data have previously been used or shared.

Industry and government actors also often rely on notice and consent and contractual approaches to privacy. However, there is a growing recognition that a reliance on notice and consent is inadequate, as individuals often do not read or understand the privacy policies of the companies with which they interact, and the disclosures made in such policies are often written in language that is so broad and vague as to not fully inform individuals who do read them. Contractual approaches to data sharing and use in the commercial context generally contain terms assigning liability and penalties in the event of a breach. Many state laws and some federal statutes have also established notice duties in the event of data breaches, and this legislative activity has helped to raise public awareness of data breaches in the commercial sector. However, the ability to recover damages through a lawsuit remains limited due to the burden of showing that an actual harm has occurred as a result of a breach, though many cases settle before reaching the merits. As a matter of jurisprudence, many courts are reluctant to award damages in cases where the injury is merely an increase in the risk that a future harm might occur, finding that the harms are too speculative or hypothetical. The harm must be determined to be one that the law recognizes as worthy of redress, deterrence, or punishment, such as a concrete financial loss that has been incurred. In many cases, it can be difficult for a victim to prove that a disclosure incident directly caused a particular harm.

A number of companies that manage large-scale datasets containing personal information about individuals, particularly those that have been at the center of privacy-related controversies in the recent

past, are beginning to implement internal ethical review processes. [54] Practices are evolving, particularly in response to high-profile data breach incidents that are occurring with increasing frequency. As they adopt data-driven business models, companies are increasingly considering the risks of maintaining large quantities of personal data over the long-term, and are incorporating risk assessments and data security and privacy safeguards into their processes. Facebook has established an ethics review process for research based on the user data it maintains. Companies like Acxiom have made efforts to enable individuals to opt out of data collection and have made some portions of their data inspectable and correctable by data subjects,[55] though they have made only a small subset of the attributes they hold viewable—a few dozen out of the over 1,500 attributes they collect. Some companies are also employing advanced computational approaches to limit their collection and use of personal data, in the interest of providing strong privacy protections for users, as demonstrated by Google's and Apple's implementations of formal privacy models like differential privacy in their data collection activities.[56] In addition, the potential for bias or discrimination in the use of personal data is a concern that is receiving growing attention. Companies such as Airbnb, in response to reports and research findings of discriminatory practices of their users,[57] are making efforts to restrict the flow of personal data they hold and encourage uses of the site that rely less on the viewing of personal information of other users.[58]

Facebook is a prominent example of a company that, in response to negative publicity regarding the ethical implications of research uses of and interventions with its data,[59] established an internal research ethics review process.[60] The Facebook ethics review processes and systems is designed to meet the ethical principles for reviewing computer and information security research proposed in the Department of Homeland Security's Menlo Report, which is based in turn on the Belmont Report that guided the development of the Common Rule.[61] This process involves training for employees on privacy and research ethics, review by a senior manager with substantive expertise in the area of proposed research, and, where needed, an extended review by a standing committee of five individuals, including substantive

---

[54] *See* Molly Jackman & Lauri Kanerva, *Evolving the IRB: Building Robust Review for Industry Research*, 72 Wash. & Lee L. Rev. Online 442 (2016).
[55] *See* Natasha Singer, *Acxiom Lets Consumers See Data It Collects*, N.Y. Times, Sept. 4, 2013.
[56] *See* Úlfar Erlingsson, Vasyl Pihur, & Aleksandra Korolova, *RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response*, Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (2014), http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42852.pdf; Andy Greenberg, *Apple's 'Differential Privacy' Is About Collecting Your Data--But Not* Your *Data*, Wired, June 13, 2016, https://www.wired.com/2016/06/apples-differential-privacy-collecting-data.
[57] *See, e.g.,* Benjamin Edelman, Michael Luca, & Dan Svirsky, *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment*, American Economic Journal: Applied Economics (forthcoming), http://www.benedelman.org/publications/airbnb-guest-discrimination-2016-09-16.pdf.
[58] *See* Katie Benner, *Airbnb Adopts Rules to Fight Discrimination by Its Hosts*, The New York Times, Sept. 8, 2016, http://www.nytimes.com/2016/09/09/technology/airbnb-anti-discrimination-rules.html.
[59] *See, e.g.,* Charles Arthur, *Facebook emotion study breached ethical guidelines, researchers say*, The Guardian, June 30, 2014, https://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say.
[60] *See* Molly Jackman & Lauri Kanerva, *Evolving the IRB: Building Robust Review for Industry Research*, 72 Wash. & Lee L. Rev. Online 442 (2016).
[61] *See id.* at 448.

area experts as well as experts in law, ethics, communications, and policy.[62] A substantive area expert may exercise discretion in expediting review of certain small product tests, or consult one of the experts in law, ethics, communications, or policy as needed. When an extended review is triggered, consensus of the five member group is required. This group considers the benefits of the research, including its value to Facebook, its value to the Facebook community and society at large, its contributions to general knowledge, and other "positive externalities and implications for society."[63] Against these benefits, the committee considers the potential adverse consequences from the research, especially with respect to research involving vulnerable populations or sensitive topics, and "whether every effort has been taken to minimize them."[64] This group also considers "whether the research is consistent with people's expectations" regarding how their personal information is collected, stored, and shared, taking into account research and recommendations by ethicists, advocates, and academics.[65] The group consults other experts at Facebook, as well as outside experts, as needed.[66]

Another notable example is Acxiom, which holds what is by some measures the largest commercial database on consumers in the world.[67] The company collects, combines, analyzes, and sells sensitive personal data from a number of sources including census data and other public records, surveys and questionnaires, retail purchases, web browsing cookies, and social media postings. To protect the sensitive data it holders, Acxiom complies with a number of regulatory and industry standards, including those found in HIPAA, HITRUST, NIST, and PCI frameworks. The company also engages in a "very rigorous" privacy impact assessment program with a focus on ethical use of data, which involves a stakeholder analysis applying "ethical judgment" to produce a "very carefully curated dataset," with "every piece of data" and "every model" created having "some sort of regulation or constriction, permission or prohibition on it."[68] While this overview describes a robust, careful, and systematic ethical review process, there are indications this process is in practice applied in an ad hoc fashion. Consider the following anecdote of a data use decision made by the Acxiom leadership team. The company's analytics team had developed a model of "10,000 audience propensities," including personal scores for a number of sensitive attributes such as "vaginal itch scores" and "erectile dysfunction scores."[69] When the leadership team met to discuss whether the use of such scores would be perceived as too invasive, one member of the team came prepared to read the actual scores on these sensitive topics for each of the individuals in the room. When confronted with the problem in this direct, personal way, the leadership team decided that certain scores were "too sensitive" and should not be made available as a product to its customers.[70] This example illustrates how, despite commitments to rigorous processes, in practice decisions may be

[62] *Id.* at 451-53.
[63] *Id.* at 452-53.
[64] *Id.* at 455.
[65] *See id.* at 455.
[66] *See id.* at 453.
[67] *See* Natasha Singer, *Mapping, and Sharing, the Consumer Genome*, The New York Times, June 16, 2012, http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html.
[68] *See* Testimony of Sheila Colclasure, Global Public Policy and Privacy - Americas Officer for Acxiom, National Committee on Vital and Health Statistics Hearing on De-identification and the Health Insurance Portability and Accountability Act (HIPAA) (May 25, 2016), http://www.ncvhs.hhs.gov/transcripts-minutes/transcript-of-the-may-25-2016-ncvhs-subcommittee-on-privacy-confidentiality-security-hearing.
[69] *See id.*
[70] *See id.*

based on ad hoc, gut judgments by a small number of decisionmakers and their opinions about unspecified "social norms."[71] The employee in this example explained that had she not brought to light her concerns in such a compelling way, the leadership team likely would have made a different decision regarding the use of the scores at issue. The reliance on the judgment of an individual, or a small group of individuals, regarding ethical use of data is likely to lead to inconsistent practices in the absence of a larger guiding framework. Indeed, other companies have reached very different conclusions regarding the appropriateness of selling similar types of highly sensitive information about individuals. For instance, other data brokers have made decisions to sell lists of names of rape victims, addresses of domestic violence shelters, and names of individuals suffering from various health conditions, including genetic diseases, dementia, HIV/AIDS.[72]

3. The expanding scale of data and new commercial uses are increasing risks and decreasing the effectiveness of commonly used controls.

The expanding timescale of data activities is widening gaps between the state of the practice and the state of the art for privacy protection. The increasing scale of commercial and government data programs, including the collection of data at more frequent intervals, the extended period of data collection, and the amount of time that has elapsed between collection and use affect privacy risk in different ways and threaten to further erode the effectiveness of traditional approaches to privacy. For instance, the accumulation of a large number of observations, linked at the level of an individual, over time forms a unique pattern of activity for each individual in a set of data, increasing risks that an individual in the data can be identified, or that information can be learned about individuals based on the inclusion of their information in the data. These shifts are putting pressure on current practices for privacy protection, and scholars and practitioners are now exploring new technical, procedural, and legal interventions for managing data privacy that can complement traditional approaches and better address the challenges raised by long-term data activities.

3.1. Key drivers of risk in long-term data activities include age, period, and frequency of data collection.

The effect of time on privacy risk is complex, and has traditionally not been well understood. This section aims to separate the characteristics of long-term data collections from other privacy-related factors, in order to discuss how each factor may independently affect risk.

Many concepts are embedded in a notion of privacy risk, and decomposing the relevant dimensions of privacy risk and analyzing them separately can inform the selection among interventions that address different drivers of risk. One possible framework is to consider privacy risks as a function of three separate dimensions—identifiability, threats, and vulnerabilities—where threats and vulnerabilities are

[71] *See id.*

[72] *See* Testimony of Pam Dixon, Executive Director of the World Privacy Forum, Before the Senate Committee on Commerce, Science, and Transportation, Hearing on "What Information Do Data Brokers Have on Consumers, and How Do They Use It?" (Dec. 18, 2013), https://www.commerce.senate.gov/public/_cache/files/e290bd4e-66e4-42ad-94c5-fcd4f9987781/BF22BC3239AE8F1E971B5FB40FFEA8DD.dixon-testimony.pdf.

often bundled together in discussions of the level of the sensitivity of personal information.[73] In this analysis, it is important to note that privacy risk is not simply an additive function of these various components; rather, identifiability and sensitivity may be better modeled as multiplicative factors. Moreover, the independent analysis for each component can be quite challenging. For instance, vulnerabilities are not evenly distributed in the population, as some members may be more vulnerable to particular threats.

A review of the various literatures guiding institutional review board practice, describing the methodology and practice of data management in longitudinal research studies, and presenting findings from the scientific study of privacy, taken together, seem to suggest at least three characteristics related to time as components that influence data privacy risk: age, period, and frequency. Table 1 summarizes the relationship between these three characteristics and the components of privacy risk (identifiability, threats, and vulnerabilities). With this table, the complex relationship between the characteristics of big data and the dimensions of privacy risk become more clear. For instance, an increase in age can, at the same time, decrease identifiability but increase sensitivity.

	Identifiability	**Threats (sensitivity)**	**Vulnerabilities (sensitivity)**
Age	Small decrease	Moderate increase	Moderate decrease
Period	Small increase	Moderate increase	No substantial evidence of effect
Frequency	Large increase	Small increase	No substantial evidence of effect

Table 1. Key risk drivers for big data over time and their effects on privacy risk components.

Each of these risk factors and the ways in which they influence components of privacy risk is discussed, in turn, in the sections below.

3.1.1. The age of the data, or the duration of storage and use of personal data, over long periods of time alters privacy risks.

The age of data is often argued to reduce the risk of identifiability, due to the fact that individuals' observable characteristics generally change over time. For instance, an individual who currently has red hair may not possess this attribute thirty years later, making this attribute less identifying with time. Additionally, the availability and accuracy of data have historically decreased with time. Both of these

[73] For an extended discussion of these terms and how they are applied, see Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 Berkeley Tech. L.J. 1967 (2015).

factors make identification based on older measurements less reliable. Arguably, this is a weak reduction, as some characteristics, such as DNA, do not appreciably change over time. Further, there are notable exceptions to data becoming less available over time. When information that was not available in digital form is digitized and disseminated it becomes more widely and persistently available than it had been in the past, as seen with the online publication of arrest records and mugshots. In addition, real estate or criminal records created decades ago are now being digitized and made publicly available online, lowering the barrier to access and enabling commercial uses far removed from the contexts that the data subjects likely envisioned at the time the data were created.[74]

Increased age of the data may also lead to increased threats related to data retention. As the number of observations collected about individuals expands rapidly and the demand to use data covering longer periods of time grows, commercial and government actors are implementing procedures for storing this information over extended timeframes. This expansion in the scope of data retained in their information systems necessarily increases the risk that a data breach will occur, and that such a breach will lead to harm to the individuals in the data. The retention of accumulations of large quantities of individual-level information makes it more likely for these information systems to be the targets of hackers and others who would seek unauthorized access to the data. The large quantity of personal information that would be disclosed in the event of a breach increases the likelihood that individuals will incur harms. To address these risks, industry standards often require the encryption of data in storage, and some laws, particularly state-level data security laws, require encryption where it can be reasonably implemented.

In addition, as the time between collection and use increases, the potential for applying the data to research uses that could not be anticipated at the time of collection grows. Increased age leads to increased threats from data use. The types of data, the uses and analytical methods applied, the expected disclosure risks, public expectations and perceptions, and applicable laws and policies involved can all change in unpredictable ways throughout the extended timeframe of a long-term data management program. For example, expected risks are likely to grow over time through the emergence of new analytical learning techniques. New analytical methods, such as machine learning and social network analyses, can unlock new uses of information that were originally collected for different purposes. For instance, social media postings are being used to track the spread of illnesses, measure behavioral risk factors, and infer individuals' personality traits.[75]

The longer the study, the greater the likelihood that circumstances will change as data are collected, and the more likely it is that a researcher will encounter challenges related to obtaining consent, complying with privacy laws and regulations, and disclosing the risks of participation to research subjects. For example, these issues put pressure on consent mechanisms, as they challenge the notion that research subjects were adequately informed of the risks of their participation, the intended uses of their information, and the effectiveness of confidentiality protections put in place. These changes can also lead to new, unanticipated harms, or, alternatively, lead to a decrease in the potential for some harm as events

[74] *See generally* Federal Trade Commission, Data Brokers: A Call for Transparency and Accountability (May 2014).
[75] *See, e.g.,* Michael J. Paul & Mark Dredze, You Are What You Tweet: Analyzing Twitter for Public Health, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011); Wu Youyou, Michal Kosinski, & David Stillwell, Computer-based personality judgments are more accurate than those made by humans, 112 Proceedings of the National Academy of Sciences 1036 (2014).

to which information refers may become less relevant or more obscure.

Increased age of the data is also recognized as decreasing the vulnerability of some of the individuals in the data to privacy-related harms. Information in the distant past may be commonly viewed as less relevant and less likely to cause embarrassment and other harms such as loss of insurability. Consider, for instance, the magnitude of potential harm from a release of a high school student's grade at a time when the subject is a high school student or a recent graduate, versus such a disclosure thirty years later. Taken further, when the age of the data is great enough, the subjects measured will be deceased and unaffected by many of the consequences of personal information disclosure. For instance, the US Census Bureau and National Archives and Records Administration follow a rule restricting access to individually identifying information from decennial census records for a period of 72 years from the date of collection, after which risk is considered to be low enough to permit the release of individual-level data to the public.[76] In addition, laws such as the Common Rule are defined such that their rules apply only to data about living individuals.[77] Individual research studies may also be designed to trigger the release of data after a substantial period of time has passed. For instance, the Grant Study of Adult Development at Harvard Medical School, a 75-year longitudinal study of Harvard college sophomores from the classes of 1939-1994, has been established such that "[a]fter Jan 1, 2028, a sufficient length of time will have passed since the collection of the information so that the data can be considered 'historic'. At that time, the data will be made available to the public without restriction."[78] It is important to note, however, that even deceased individuals may have a legitimate interest in the way their private information reflects on their children or on groups to which they belong.

3.1.2. Long periods of data collection, i.e., data that describe trends, create additional privacy risks.

Increases in the period of collection may result in increased threats as the data enable analysis of trends over time that reveal sensitive characteristics related to health, behavior, and interests. An illustrative example of a program with an extended data collection period is the Framingham Heart Study, in which investigators have been continuously collecting data from study participants and their descendents since 1948.[79] The data collected over the course of this long-term study reveal information about an individual's development of risk factors for or progression of heart disease, diabetes, and Alzheimer's disease and dementia, among other sensitive attributes.

There are additional, weaker risks related to longer periods of coverage. First, long periods of data collection are correlated with greater age of the data, as age must be at least as large as the period, and age increases privacy threats. Second, because human behavior exhibits patterns at multiple temporal scales, the interaction of extended period of collection and high frequency may enable increased detection of

[76] 44 U.S.C. § 2108(b).
[77] *See* 45 C.F.R. § 46.102.
[78] *See* Grant Study of Adult Development, 1938-2000, Data Contribution Terms and Conditions, Murray Research Archive Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/00290.
[79] *See* Framingham Heart Study, History of the Framingham Heart Study, https://www.framinghamheartstudy.org/about-fhs/history.php.

trends, further increasing threats and enabling stronger behavioral fingerprinting, thereby increasing the identifiability of the data.[80]

3.1.3. High-frequency data collections pose a significant challenge to traditional privacy approaches such as de-identification.

In many cases, commercial and government big data collection leads to much more frequent observations than those collected in the research setting. For example, the microphone, camera, accelerometer, GPS receiver, and other sensors embedded in a mobile device can generate fine-grained data, capture variations microsecond by microsecond, and transmit the data to the cloud for long-term storage and analysis. Data points collected at frequent intervals can also reveal identifiable or sensitive details about individuals. For instance, mobile health apps and devices use sensors to continuously monitor and record features related to an individual's health and behavior, which can reveal sensitive facts about an individual's health status. High-frequency data dramatically increases identifiability, as researchers have demonstrated that just four data points about an individual's spatiotemporal location or retail purchases can be sufficient to uniquely identify her records in a dataset.[81]

As discussed below, commercial and government big data activities also collect data from a wider number of subjects than that of a traditional research study which may be limited to hundreds of participants due to resource constraints. In addition, the limitations of traditional research methods have led to the historical capture and analyze of fewer data points in longitudinal research studies than is found in a large-scale commercial or government dataset today. However, big data collection and analysis methods are increasingly being introduced into the research setting through the use of new electronic devices, such as dedicated electronic devices for adherence monitoring, or mobile apps for tracking health measurements over time. These factors related to the growing number of observations collected often lead to greater harm to individuals, and harm to a greater number of individuals, should the data be exposed. There are also weaker implications associated with the frequency of data collection; for example, as above high frequency data collections may interact with period, increasing the threats from data release.

3.2. Additional risk factors not specific to the time dimension, such as size and diversity of the sample, also increase privacy risks.

Some of the gaps between the current practice for managing privacy in commercial big data and open government data and the state of the art for privacy protection are independent of the extended timescale associated with the increasing reliance on long-term data activities in these contexts. For instance, many of the canonical examples of big data privacy risks were enabled by features of the data not directly correlated with time. The identification of individuals in the release of AOL search query records was enabled by the association of a large number of search queries with the record of an individual in the

[80] *See, e.g.,* Nathan Eagle & Alex (Sandy) Pentland, *Reality Mining: Sensing Complex Social Systems*, 10 J. of Personal and Ubiquitous Computing 255 (2006); Nathan Eagle & Alex (Sandy) Pentland, Eigenbehaviors: Identifying Structure in Routine, 63 Behavioral Ecology and Sociobiology 1057 (2009).
[81] *See* Yves--Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 Science 536 (2015); Yves--Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 Nature Sci. Rep. 1376 (2013).

released data;[82] although names were not provided for each record, a large set of an individual's search queries alone were found to be revealing. Individual privacy risks were identified in a release of data from Facebook profiles used in a research study, based on the individual-level attributes disclosed in the data and information about the school revealed in the codebook that led to the identification of Harvard College students as the population being studied.[83] The release of a dataset containing film ratings by Netflix users was vulnerable to re-identification attacks due to the number of attributes provided within each record making many of the records in the dataset unique and potentially identifiable when cross-referenced with relevant auxiliary information.[84] Moving beyond risks of re-identification, an analysis of public information posted by one's friends on the Facebook platform can be used to predict personal characteristics, such as an individual's sexual preference.[85] Furthermore, examples of algorithmic discrimination raise questions about how personal information is used to build and apply models to classify individuals in ways that cause harm to individuals and groups.[86]

Attempts to apply traditional statistical disclosure limitation techniques to big data, as well as scientific research into new formal models of privacy, suggest that a number of additional characteristics increase privacy risk by increasing identifiability, threats, or vulnerabilities. For instance, traditional approaches to de-identification have been demonstrated to fail when applied to large datasets such as the dataset containing Netflix users' film ratings.[87] More generally, as discussed below in Section 3.2.1, for a number of reasons, de-identification techniques often fail when applied to high-dimensional data. In addition, data releases may enable new categories of attacks, such as stylometric attacks that enable predictions of authorship based on the writing style of a sample text and new risks to populations such as discriminatory uses of big data.

		Identifiability	**Threats (sensitivity)**	**Vulnerabilities (sensitivity)**
	Population Diversity	Small decrease	Moderate increase	Small increase

[82] *See* Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. Times (Aug. 9, 2006).
[83] Michael Zimmer, *''But the data is already public'': On the ethics of research in Facebook*, 12 Ethics and Information Technology 313 (2010)
[84] *See* Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, Proceedings of the 2008 IEEE Symposium on Security and Privacy 111 (2008), http://dl.acm.org/citation.cfm?id=1398064.
[85] *See* Carter Jernigan & Behram F.T. Mistree, *Gaydar: Facebook friendships expose sexual orientation*, 14 First Monday 10 (2009).
[86] *See, e.g.*, Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 Communications of the ACM 44 (2013); Julia Angwin, Jeff Larson, Surya Mattau, & Lauren Kirchner, *Machine Bias*, ProPublica, May 23, 2016; Alessandro Acquisti, Laura Brandimarte, & George Loewenstein, *Privacy and Human Behavior in the Age of Information*, 347 Science 509 (2015).
[87] *See* Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, Proceedings of the 2008 IEEE Symposium on Security and Privacy 111 (2008), http://dl.acm.org/citation.cfm?id=1398064.

	Sample Size	Small increase	No substantial evidence of effect	Moderate increase
	Dimensionality	Moderate increase	Moderate increase	No substantial evidence of effect
	Broader Analytic Use	Large increase	Moderate increase	Large increase

Table 2. Non-temporal characteristics of big data that drive privacy risks and their effects on components of privacy risk.

3.2.1. High-dimensional data pose challenges for traditional privacy approaches such as de-identification, and increase the difficulty of predicting future data uses.

During any single interaction with an individual, a company or government agency may record just a few data points. However, the information collected that is collected is likely to be linked with data from other sources at a later time. In the commercial big data context, there are fewer constraints on linking data, compared to those that are imposed in the research context. In fact, companies often have an incentive to combine as much data as possible, and draw linkages at the individual level, in order to assemble more accurate behavioral profiles for the individuals in their databases. A prominent example is illustrated by data brokers like Acxiom, which seek to accumulate and link data about the same individuals from many different sources, including administrative records from multiple government agencies as well as data collected from other commercial sources. In turn, the data profiles compiled by Acxiom are sold to other companies, including banks, automotive companies, and department stores,[88] and linked to even more data sources. The profiles that are assembled on individuals contain a large number of dimensions, including names, addresses, phone numbers, Social Security numbers, date of birth, height and weight, race and ethnicity, marital status, occupation, religious affiliation, voting registration and party affiliation, criminal offenses and convictions, product purchase histories, social media friend connections, level of Internet usage, home market value, participation in hobbies and other activities, movie and music preferences, charitable giving, credit worthiness, vehicle ownership, travel purchases, tobacco usage, and medical ailment and prescription online search propensity, among countless other data points.[89] As of 2012, Acxiom purportedly held data on approximately 500 million individuals, including about 1,500 pieces of information about each person,[90] and these figures are likely much higher today. Acxiom also

[88] Natasha Singer, *Mapping, and Sharing, the Consumer Genome*, New York Times, June 16, 2012, http://www.nytimes.com/2012/06/17/technology/acxiom-the- quiet-giant-of-consumer-database-marketing.html.
[89] Federal Trade Commission, Data Brokers: A Call for Transparency and Accountability (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.
[90] Natasha Singer, *Mapping, and Sharing, the Consumer Genome*, New York Times, June 16, 2012, http://www.nytimes.com/2012/06/17/technology/acxiom-the- quiet-giant-of-consumer-database-marketing.html.

offers tools to its advertiser clients to help them link data contained in different databases, even where there are discrepancies in the information about an individual in different sources, such as those due to a change of name or address.[91]

Data such as those held in Acxiom's databases are high-dimensional, which presents a significant challenge for traditional approaches to de-identification. Statistical disclosure limitation techniques such as aggregation and generalization have been shown to fail when applied to data with a high dimension of observable characteristics.[92] High-dimensional data can also support a wide range of future uses, which may be difficult to anticipate at the time of collection. For example, high-dimensional data can enable unanticipated analyses such as the prediction of authorship of text based on writing style, detection of Parkinson's disease symptoms, and discoveries of an individual's predisposition for various health conditions.

The dimensionality of big data may be even higher than it appears at first observation, as richer data types, such as network data, unstructured text, audio, and video, are subject to multiple independent measurements and thus may be thought of as having multiple embedded dimensions or signals. For example, informational risks from social network analyses are a function not only of the nodes, as the structure of the network connections carries information as well.[93] Pieces of text may be associated with metadata (e.g., Twitter posts may have embedded geolocation codes), may embed direct identifiers such as names (e.g., medical records often contain names, dates, and addresses), and may also be linkable to identities through stylometric analysis.[94] Motion data, such as those collected by wearable fitness trackers, may reveal private types of activity. Video and audio data generate a range of unexpected signals; for example, indicators of Parkinson's disease have been detected based on voice recordings. Heart rate can be detected using iPhone video cameras. In addition, research has shown it is possible to extract conversations from video recorded images of vibrations on surrounding materials, and the use of WiFi signal strength may be used to determine the occupancy of a room. The potential for unexpected uses of data can be expected to grow with advances in technology, and high-dimensional data especially can support new uses of data that were unforeseen at the time of collection.

3.2.2. Broader analytic uses affect the identifiability, threats, and vulnerability components of privacy risk.

Both traditional de-identification techniques and emerging tools based on formal privacy models such as differential privacy are designed to enable accurate estimations of population parameters. Traditional, as well as modern, approaches to deidentification are not effective against harms such as algorithmic

[91] Jim Edwards, *Facebook's Big Data Partner Knows Who You Are Even When You Use a Different Name on the Web*, Business Insider, Sept. 26, 2013, http://www.businessinsider.com/facebook-and-acxioms-big-data-partnership-2013-9.
[92] *See, e.g.,* Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, Proceedings of the 2008 IEEE Symposium on Security and Privacy 111 (2008), http://dl.acm.org/citation.cfm?id=1398064.
[93] *See, e.g.*, Lars Backstrom, Cynthia Dwork, & Jon Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, Proceedings of WWW 2007 (2007).
[94] *See, e.g.*, Ahmed Al Faresi, Ahmed Alazzawe, & Anis Alazzawe, *Privacy Leakage in Health Social Networks*, 30 Computational Intelligence 514 (2014).

discrimination, which is often viewed as a misuse of personal data. Neither category of tools protects against learning facts about populations that could be used to discriminate. These techniques also cannot be used when the goal is explicitly to make inferences about or intervene with individuals.

Examples of algorithmic discrimination fall along a spectrum. At one end of the spectrum, for example, there may be differential pricing, whereby firms aim to generate individualized predictions or interventions based on their information but such outcomes are not essential, as they still derive utility from fitting models to group data and applying models to individuals based on their group attributes. Towards the center of the spectrum, there may be activities such as personalized medicine, in which models may be constructed based on aggregated data, but personalized treatment requires inference based on an individual's actual information. At the other end of the spectrum, there may be fraud detection, in which the goal is inherently to make inference about the predicted behavior of a specific individual. In addition, predicting and intervening with individuals expands the set of threats that must be considered beyond those that arise from population estimates, for examples, through differential pricing, redlining, recidivism scores, microtargeted advertising.

3.2.3. Increasing sample size and population diversity also lead to heightened privacy risks.

Increases in the sample size covered in a set of data is associated with an increase in identifiability. As a sample grows to represent a larger fraction of population, one can be more confident that any particular target individual is in the sample. This makes re-identification easier in the same way that having the list of all those who participated in the database makes re-identification easier. Sample size is also associated with an increase in vulnerability. As samples grow to be very large, it is quite likely to include individuals who are particularly sensitive, such as members of vulnerable populations, to whatever potential threats the data pose.

Broadening of population diversity also increases the range of threats that may be associated with a set of data. Covering a broader population in a set of data increases the range of threats that are relevant to at least some member of the population. For instance, disclosure of political party affiliation generally does not pose a large threat to a set of data containing individuals from the US population, though it would pose a large threat of retribution in a broader population that also includes individuals living under non-democratic regimes. In a diverse enough population, all plausible threats would be implicated. Further diversification can be thought of as increasing the likelihood that one has to consider a wider range of vulnerabilities to these threats that are specific to the context of the subject.

3.3. The key risk factors identified in long-term data activities change the surface of suitable privacy controls.

As discussed above, characteristics of big data activities are associated with increased privacy risks in a number of ways. Debates taking place today regarding the future of privacy in light of these concerns about long-term big data programs are reminiscent of earlier deliberations regarding to the long-term collection, storage, use, and retention of research data. Researchers have long been collecting information about human subjects over extended timeframes, and there are notable similarities in the populations

involved, the attributes measured, and thus the potential privacy risks and challenges posed across long-term data activities in the research, commercial, and government contexts. Much like commercial and government data programs, longitudinal research studies are a rich source of data for exploring economic, sociological, and epidemiological trends over the lifetimes of individuals and their descendents. Data collected throughout the course of such studies are in many cases highly specific, identifiable, and sensitive, and carry risks that are similar to those associated with personal data held by corporations and governments.

Despite these similarities, the management of the privacy of personal data across these various contexts varies markedly. In order to ensure robust protection of privacy, similar privacy risks should be addressed similarly, which requires applying principles for balancing privacy and utility in data releases more systematically. In order to do so, the risk-benefit analyses and best practices established by the research community can be instructive for privacy management with respect to the long-term collection and use of personal data by commercial and government organizations. Corporations and governments may consider adopting review processes similar to those that have been established at research institutions to continually analyze the risks and benefits associated with data collection, retention, use, and disclosure over time.

As presented in prior work, a systematic analysis of the threats, vulnerabilities, and intended uses associated with a set of data can be used to help guide the selection of appropriate sets of privacy and security controls, much like the review processes employed in the research context. Figure 1 below provides a partial conceptualization of the relationship between identifiability, sensitivity, and the suitability of selected procedural, legal, and technical controls at the collection and release stages of the information lifecycle. For the purposes of this conceptual illustration, Figure 1 focuses on a small subset of tools from the wide range of procedural, economic, educational, legal, and technical interventions that are available to data managers. A real-world data management program should be designed to utilize appropriate tools from the full selection of interventions available and to incorporate them at each stage of the information lifecycle, from collection, to transformation, retention, release, and post-release.

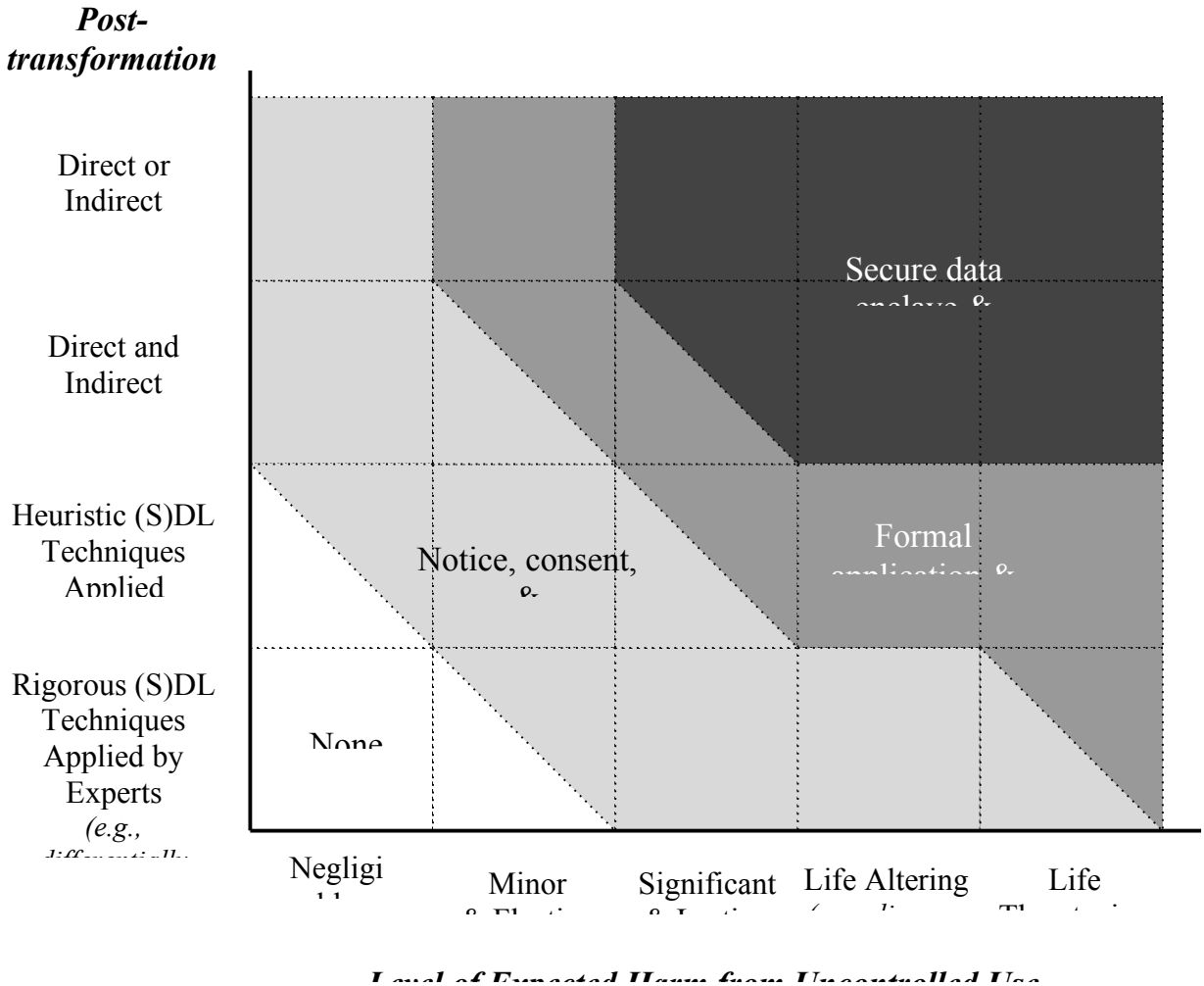
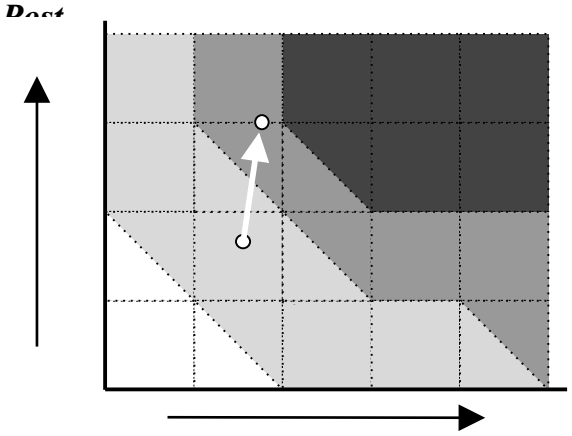


Figure 1. How identifiability and sensitivity guide the recommendations of sets of privacy and security controls.[95]

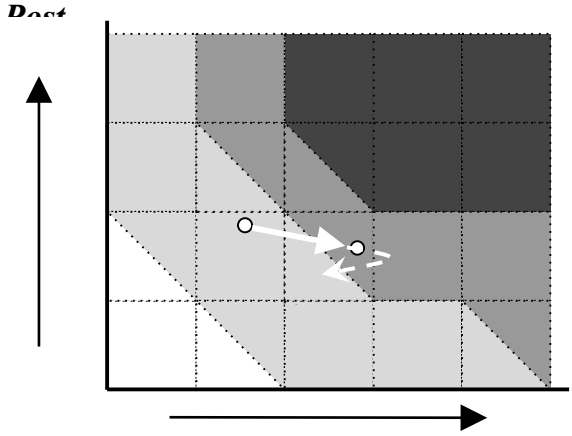
As partially illustrated in Figure 1, rather than relying on a single intervention such as de-identification or consent, corporate and government actors may weigh the suitability of combinations of interventions from the wide range of privacy and security controls that are available. There is a growing recognition that de--identification alone is not sufficient to be used as a general standard for privacy protection. Robust privacy protection requires a systematic analysis of informational risks and intended uses, and, in many cases, the implementation of additional privacy and security controls in combination with de-identification techniques. New procedural, legal, and technical tools for evaluating and mitigating risk, balancing privacy and utility, and providing enhanced transparency, review, and accountability, are being investigated and some are beginning to be deployed as part of comprehensive research data management plans. The suitability of new tools, especially formal privacy models such as differential privacy, should also be explored within the context of corporate and open government data.

[95] This diagram originally appeared in Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 Berkeley Tech. L.J. 1967 (2015).

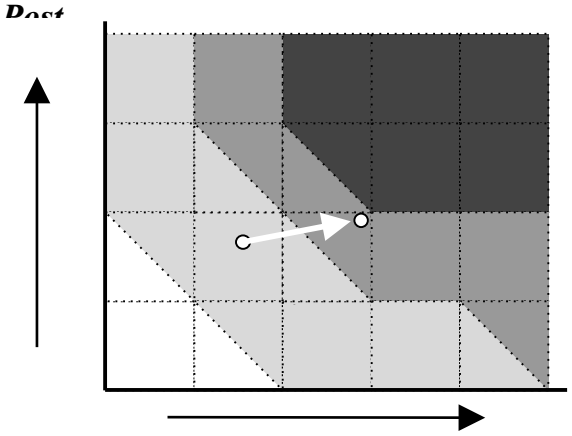
In particular, practical data sharing models can combine various legal and technical approaches. For instance, a data release could be designed to provide public access to some data without restriction after robust disclosure limitation techniques have transformed the data into, for example, differentially private statistics. Data users who intend to perform analyses that require the full dataset, including direct and indirect identifiers, could be instructed to submit an application to a review board, and their use of the data would be restricted by the terms of a data use agreement and, in some cases, accessed only through a secure data enclave. In this way, data release mechanisms can be tailored to the threats and vulnerabilities associated with a given set of data, as well as the uses desired by different users.



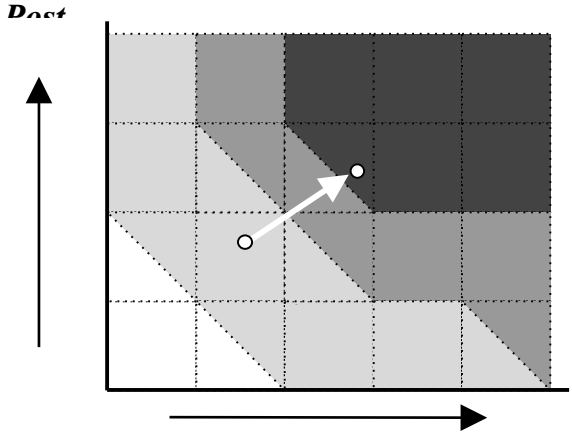
(a) A shift from one-time data collection to high-frequency data collection (e.g., a mobile app developer makes a decision to collect continuous geolocation data from the mobile device, rather than relying on the user's self-reported location).



(b) A shift to analysis of data of greater age (e.g., a web site begins offering users access to archives for analyzing older stored data).



(c) A shift to a longer period of data collection (e.g., a search engine makes a decision to collect data over a



(d) A shift from low-dimensional data collection to high-dimensional data collection (e.g., a web site

longer period of time and to customize its search results for a user based on the user's full search history, rather than using only the data collected during the user's current session).

operator makes a decision to require users to log in using their Facebook accounts, which grants the operator access to the user's Facebook account data).

Figures 2a-d. How big data characteristics shift the recommendations of sets of privacy and security controls.

As illustrated in Figures 2a-d, the characteristics of long-term data programs, such as the increasing frequency of collection, shift the recommended sets of controls. For instance, de-identifying longitudinal data effectively is a significant challenge, due to the frequency of data collection, the extended duration of data collection, and the linkage of observations at the individual level. These factors substantially increase the dimensionality of the data, and, correspondingly, the likelihood that records in a set of data will be unique for each individual. The uniqueness of the records, in turn, makes it more difficult to effectively de-identify the data through simple redaction techniques. To address privacy with respect to high-dimensional datasets, researchers often employ additional controls such as data use agreements when sharing data, rather than relying on de-identification alone. In addition, data sharing models that employ formal privacy guarantees hold promise for providing strong privacy protection while enabling analysis of high-dimensional data.

Figures 2a-d illustrate how different features of big data correspond to shifts in identifiability and harm, through a series of examples. In the first example, Figure 2a, the developer of a weather app for mobile devices makes a decision to collect coarse geolocation data on an hourly basis from the mobile device as the app runs in the background, rather than relying on the user to self-report updates on her current location. This represents a shift from one-time data collection to high-frequency data collection. This change is likely to substantially increase the identifiability of the data collected. Geolocation data collected on an hourly basis, within a few days, will likely reveal a unique pattern of behavior for an individual that can be used to determine the identity of the individual described in the record. This change is likely to have a smaller effect on the level of expected harm from the data, however, as higher frequency data collection is expected to reveal sensitive attributes about individuals only in rare circumstances.

In Figures 2a-d, an increase in expected harm implies that stronger privacy controls should be implemented. As the frequency of data collection increases, leading to a greater number of observations being collected for each individual in a set of data, the level of expected harm also increases. Moreover, potential uses of personal data that are collected are likely to evolve over the course of a long-term data program. These changes over time are likely to increase the potential for harm to individuals. This is especially likely where there are substantial deviations between the expectations of individuals and the actual uses of their information. Solutions such as granular user controls and dynamic consent are being designed to address these concerns in the long-term research context.

Upon determining that a particular risk factor increases, one is naturally tempted to mitigate this factor. Note, however, independent of other considerations, this may not be the most effective approach. For example, consider data from GPS sensors, where the high frequency of the data points from the sensors

are driving significant privacy risks for individuals whose activities are implicated in the data. The most direct approach to addressing this risk may appear to be reducing the frequency of the data. Yet, one must also consider that this approach also directly affects uses of the data. Reducing the frequency of the sensor data reduces the detail and will hinder future efforts by analysts to build models and discover fine-grained patterns using the data. Alternatively, one could attempt to reduce the frequency of the data at later lifecycle stages; for instance, data could be collected at a high frequency, but only infrequent samples would be stored. This can still have large effects on the utility of the data collection. Consider, for instance, wearable fitness trackers from which high-frequency data are especially valuable to users. One could also reduce frequency at an even later lifecycle stage, by transforming the data to be less identifiable. For high-frequency data, this requires the implementation of experimental de-identification techniques,[96] is computationally costly, and substantially reduces the analytic utility of the data.

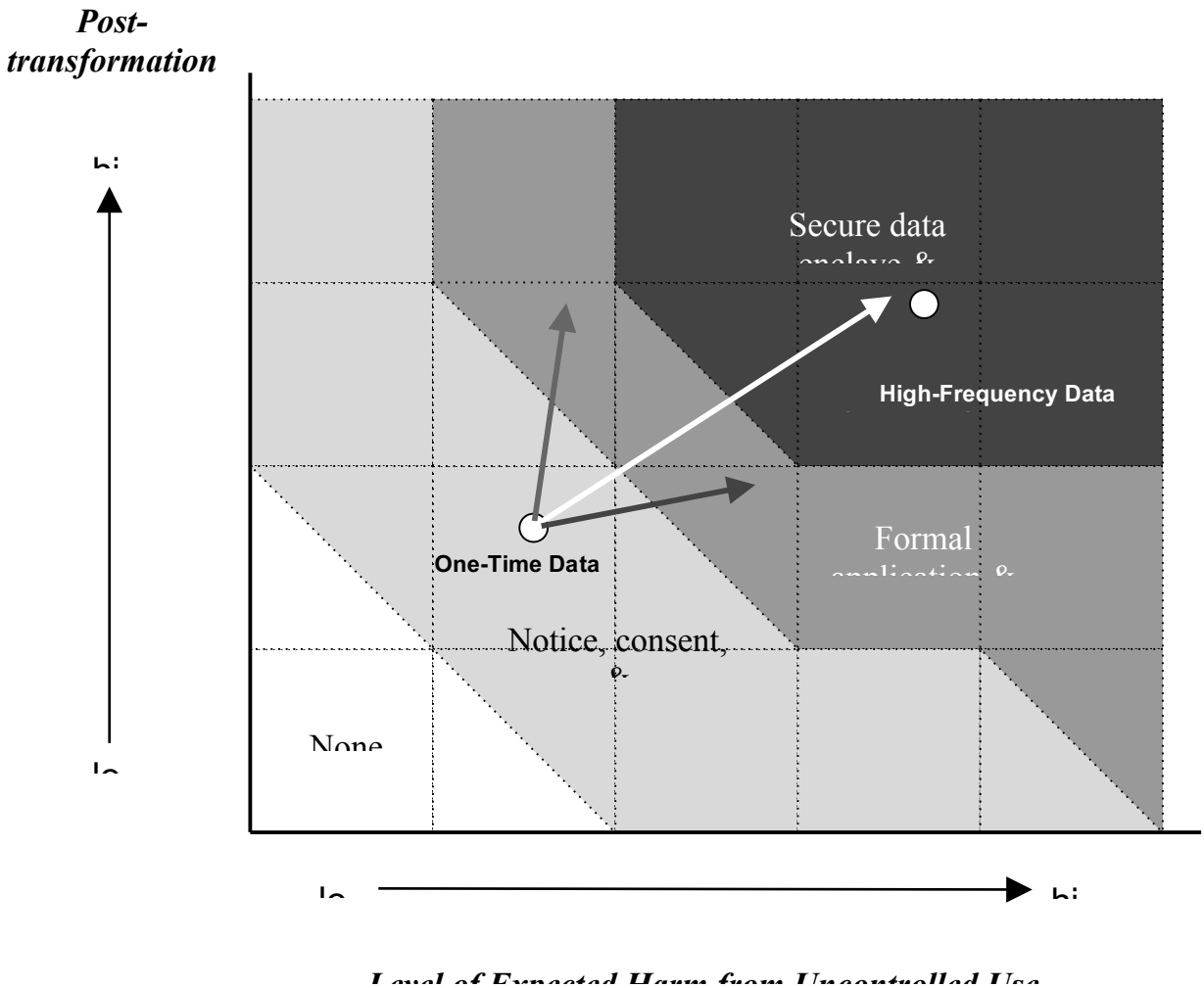
Different characteristics of data may combine to create privacy risks, and the effects are often cumulative and super-linear. For example, the privacy risks of combining increasing period and increasing dimensionality is likely to be cumulative -- both broaden the range of inferences possible from the data and create new threats. It is also possible that the range of inferences and threats grows rapidly as multiple factors change: For example, as described in section 3.1.2, the combination of more increasing periods data of collection and higher frequency of measurement creates additional risks to privacy because period and frequency interact -- when data is collected over longer terms and at higher frequency, it allows detection of behavior that manifest at many additional scales of time.

To fix ideas, consider the privacy risks of increasing the frequency of data collection for the case of the weather app (described above in Figure 2a) if its creators updated their software to support real-time weather alerts by collecting data continuously in the background. The risks from this scenario are illustrated in Figure 3 below. As frequency increases (e.g. from hourly samples to minutes) identifiability also increases (as detailed in 3.1.3) in that users become behaviorally identifiable in shorter period of time. However, as a practical matter, the change in frequency may not substantially increase the overall proportion of users identified -- even using hourly samples, the vast majority of users can be identified by their behavioral fingerprints.

However increasing the frequency of collection (while keeping the collection period constant) enables analysis of behavior that is expressed at additional time scales. In this case, the new data may enable inferences about the users walking and visiting patterns, enabling limited inferences over their fitness, excercise habits, or shopping habits. Thus the interaction of these factors creates new privacy threats. Furthermore, changing additional characteristics of the data can cause additional nonlinear increases in privacy risk: For example, if he the weather application above became ubiquitous (perhaps through a deal with major cell network providers), the sample of users included would be a large fraction of the

[96] For a survey of emerging approaches to de-identifying spatiotemporal data, see Michael Herrmann, Mireille Hildebrandt, Laura Tielemans, & Claudia Diaz, *Privacy in Location-based Service: An Interdisciplinary Approach*, Dissertation (2016); Benjamin C.M. Fung, Ke Wang, Rui Chen, & Philip S. Yu, *Privacy-Preserving Data Publishing: A Survey of Recent Developments*, 42 ACM Computing Surveys 14 (2010); Dale L. Zimmerman & Claire Pavlik, *Quantifying the Effects of Mask Metadata Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data*, 40 Geographical Analysis 52 (2008); John Krumm, A Survey of Computational Location Privacy, Personal and Ubiquitous Computing (2008).

population. One could then use derive colocation to infer, with reasonable reliability, connections between people, and even group action -- creating threats of surveillance.



Figures 3. How multiple risks combine.

A more efficient way of providing appropriate privacy protections while maintaining analytic utility is to use a set of controls in concert. The discussions to follow in Sections 3.3.1 and 3.3.2 illustrate different sets of controls and their relationship to the big data risk factors identified above in Section 2.

3.3.1. Privacy and security controls can be combined to address identifiability in large-scale longitudinal data.

Regardless of the driver of risk, limiting identifiability, where it can be accomplished without drastically reducing utility, reduces overall risk. This is a generally applicable principle, regardless of the specific

risk factor. For example, although increasing age of data does not increase identifiability (in fact, it likely weakly decreases it), the risks that age of data presents through expanded potential uses and thus potential threats would still be mitigated if adversaries were not able to learn about individuals from the data.

Commercial entities most often limit identifiability of data at the stage of publication, through "de-identifying" the data prior to release. However, identifiability can be controlled at other stages of the information life-cycle:[97] For example: identifiability may be limited at the collection stage, through designed information minimization; or at the analysis stage by providing query tools that dynamically add noise to query results.

Commercial entities apply a range of tools for static de-identifyication of data. For example, they may mask street numbers in addresses, or remove names in order to satisfy certain state-level data security laws. De-identification in accordance with the HIPAA Privacy Rule safe harbor standard can be automated for large amounts of text with sophisticated entity matching. However, longitudinal data and big data pose challenges to the current practice of de-identification.

Advances in the scientific understanding of privacy have demonstrated that privacy measures in common use, such as static de-identification, have significant limitations. De-identification using common approaches such as simple redaction of pieces of information deemed to be identifying often does not prevent many types of learning about individuals in a set of data.[98] As a result, de-identification techniques, while reducing some risks, often do not mitigate all privacy risks to individuals nor protect personal information in the manner or to the extent individual subjects would expect. De--identified data can, in many cases, be re--identified easily. For instance, numerous re--identification attacks have demonstrated that it is often possible to identify individuals in data that have been stripped of information deemed to be directly or indirectly identifying.[99] It has been shown more generally that very few pieces of information can be used to uniquely identify an individual in a released set of data.[100]

The risks of re-identification or other types of learning about individuals from data are growing over time, and these problems are particularly challenging when it comes to long-term data containing a high number of observations linked at the individual-level. De-identification techniques raise other concerns as well. For instance, de-identification is often achieved via the redaction of information that could otherwise prove useful. Another concern related to traditional de-identification standards and techniques is that they presuppose a particular privacy goal, such as the absence of certain pieces of information deemed to be identifying, which often fails to provide strong privacy protection for individuals, when compared to techniques that are designed to satisfy a more general privacy goal, such as a formal privacy

---

[97] For an extended discussion see Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 Berkeley Tech. L.J. 1967 (2015).

[98] *See* Arvind Narayanan & Edward W. Felten, *No Silver Bullet: De-identification Still Doesn't Work* (2014), http://www.randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

[99] *See* Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J. L., MED., & Ethics 98; Latanya Sweeney, *Uniqueness of Simple Demographics in the US Population*, Data Privacy Lab Technical Report (2000).

[100] *See* Yves--Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 Science 536 (2015); Yves--Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 Nature Sci. Rep. 1376 (2013).

definition.

A number of privacy controls are available for addressing risks associated with identifiability, including simple redaction approaches and tools, heuristic statistical disclosure limitation techniques (such as approaches for aggregating or generalizing individual-level data), and robust disclosure limitation techniques and tools that provide formal privacy guarantees like differential privacy. However, some big data risk factors make efficient identifiability limitation quite challenging. For example, with respect to high-frequency data, traditional de-identification techniques can yield misleading results if spatiotemporal-related data points are grouped together as a single dimension in a table. When applied to individual observations, it may appear by certain measures that privacy risks have been reduced. For instance, such an approach may be sufficient to meet a privacy standard such as k-anonymity. However, people may be readily identified from their longitudinal spatial trace.[101] With respect to high-dimensional data that are less-structured, such as text or video, however, one is especially unlikely to be aware of all of the dimensions. A single tweet, for example, has many signals, such as the content of the tweet, the names that may be mentioned in it, geolocation information attached to it, and even the unique writing style embedded in the content.[102] In addition, with high-dimensional data, traditional de-identification techniques severely reduce utility. For example, researchers have concluded that, because of the high-dimensionality and sparseness of large-scale datasets containing individual Netflix users' ratings for each of the films they have watched, traditional approaches to de-identification, such as suppression and generalization, fail to provide meaningful privacy protection and also destroy the utility of the data.[103] To address identifiability in high-dimensional data, advanced tools such as synthetic data and differential privacy, are beginning to be implemented. For example, the Census Bureau has also experimented with releasing data using synthetic data models, including the 2011 release of the synthetic Longitudinal Business Database,[104] which has been shown to meet a variant of differential privacy.[105] There are several practical implementations of differential privacy, though there are no off-the-shelf tools that can be applied without expertise. However, such off-the-shelf tools are beginning to emerge.[106]

---

[101] *See* Benjamin C.M. Fung, Ke Wang, Rui Chen, & Philip S. Yu, *Privacy-Preserving Data Publishing: A Survey of Recent Developments*, 42 ACM Computing Surveys 14 (2010); Khaled El Emam, The De-identification of Longitudinal and Geospatial Data, Presentation (July 20, 2011), http://www.ehealthinformation.ca/media/events/webinars/the-de-identification-of-longitudinal-and-geospatial-data/.

[102] *See* Mudit Bhargava, Pulkit Mehndiratta, & Krishna Asawa, *Stylometric Analysis for Authorship Attribution on Twitter*, in Big Data Analytics (Vasudha Bhatnagar & Srinath Srinivasa, eds.) (2013).

[103] Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, Proceedings of the 2008 IEEE Symposium on Security and Privacy 111 (2008), http://dl.acm.org/citation.cfm?id=1398064.

[104] *See* Satkartar K. Kinney et al., *Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database*. 79 International Statistical Review 362 (2011), http://hdl.handle.net/10.1111/j.1751-5823.2011.00153.x.

[105] *See id.*

[106] *See, e.g.*, Frank McSherry, *Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis*, Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (2009), https://www.microsoft.com/en-us/research/wp-content/uploads/2009/06/sigmod115-mcsherry.pdf; Prashanth Mohan et al., *GUPT: Privacy-Preserving Data Analysis Made Easy*, Proceedings of 2012 ACM SIGMOD International Conference on Management of Data (2012), http://www.dhosa.org/wp-content/uploads/2012/08/gupt.pdf; Marco Gaboardi et al., *PSI: A Private data Sharing Interface*, Working Paper (2016), http://privacytools.seas.harvard.edu/files/privacytools/files/psi-privacy-tool.pdf.

With broadening of analytic uses, certain types of new analytic uses, including individualized analysis and interventions, ranging from differential pricing, to personalized medicine, to fraud detection, may be in direct tension with de-identification approaches. New approaches, including formal privacy models like differential privacy and tiered access models incorporating a combination of different types of legal, computational, and procedural controls that are tailored to the risks and intended uses involved, can offer more effective risk reduction from learning about individuals. This is particularly relevant to the types of problems in which traditional de-identification techniques can yield misleading results. Differential privacy, for example, can meet some of such challenges presented by high-dimensional data. New techniques for spatial trajectory de-identification may also address some of the challenges researchers have encountered when applying traditional de-identification techniques.

3.3.2. Privacy and security controls can be combined to address sensitivity in large-scale longitudinal data.

All of the methods for limiting identifiability from simple redaction to differential privacy, have a similar goal -- to limit the *unit* of inference -- all of these methods aim to reduce the ability of an information user to make inferences about individuals (or small groups of individuals) from computing over the data. Limiting identifiability is not the only way to limit risk. Traditionally, commercial entities have limited *who* can make computations, and *the systems that are allowed to make these computations.* New computational methods, such as secure multiparty computation, blockchain, executable policies enable one to (respectively) put formal limits on *the computations that can be successfully perfomed, the visibility of those computations, and the uses to which the results of those computations are put.*

The primary way in which computations are currently limited it through the use of secure enclaves with embedded auditing procedures. For example, federal statistical research data centers operate across the country, in partnership between federal statistical agencies and other research institutions, and some large research universities operate secure enclaves as well, such as the NORC Data Enclave at the University of Chicago. Such systems employ strong data security measures, such as those required by the Federal Information Security Management Act (FISMA),[107] maintain operational logs, incorporate vetting of individual researchers who seek access, engage in disclosure review of outputs before data release and publication, and follow strict requirements for data retention and destruction. Challenges that can be addressed using secure data enclaves include large sample sizes and high-dimensionality, which make it difficult to store the data in a single location at a conventional facility. High-dimensionality and potential expansions in future analytic uses create tensions for individually vetting results before publication. For longitudinal data analyses, period and age often drive utility, and data destruction would have a high utility cost.

Emerging approaches to address challenges such as these include secure multiparty computation, computable policies, and personal data stores. While secure multiparty computation does not directly limit the ability to infer sensitive attributes about individuals, it can be used to restrict the set of computations that are permissible on the data and make these computations auditable. Computable

[107] *See, e.g.,* NIST, FIPS Pub. 199, Standards for Security Categorization of Federal Information and Information Systems (2004).

policies do not restrict inference but may be used to restrict domains of use, or classes of authorized users, and enable further auditability.[108] Personal data stores can be used to grant individuals with fine-grained control over access and use of their information and provide audit and accountability functions as well.[109]

Notice, consent, and terms of service are used to disclose to individuals how data about them will be collected, stored, used, and shared. High-dimensional data poses challenges for the effectiveness of notice because use of such data make it difficult to anticipate, and therefore provide notice of, all potential future uses. Moreover, providing control over each measure or use quickly leads to information overload for data subjects.[110] Emerging approaches such as secure multiparty computation techniques, personal data stores, blockchain tools,[111] and privacy icons,[112] can be used to grant greater control or improved forms of notice to users.

Formal application and review by an ethics board, such as an institutional review board, in combination with a data use agreement prescribing future uses and redisclosures of the data, as well as data privacy and security requirements for handling the data, can be used to address many of these concerns. With higher dimensional data and growing populations, data use agreements are becoming increasingly complex, and there are growing possibilities of incompatibility across data use agreements, institutional policies, and individual data sources. Emerging solutions include the creation of new ethics review processes, as well as modular license generators to simplify the drafting of data use agreements. New review bodies, such as consumer review boards,[113] participant-led review boards,[114] and personal data cooperatives,[115] can be formed to ensure data subjects are informed of risks and such risks are outweighed by the benefits of the data activities. Companies such as Facebook have begun to implement data privacy and ethics review boards, to provide more systematic and regular review of privacy risks and appropriate practices.

4. Analysis of the characteristics of long-term big data that drive increased privacy risks can inform recommendations for the use of privacy and security controls in specific cases.

Corporations and governments are collecting and managing personal data over increasingly long periods

[108] *See, e.g.*, Lalana Kagal & Joe Pato, *Preserving Privacy Based on Semantic Policy Tools*, 8 IEEE Security & Privacy 25 (2010).
[109] *See, e.g.*, Yves-Alexandre de Montjoye et al., On the Trusted Use of Large-Scale Personal Data, 35 IEEE DATA ENG. BULL. 5 (2013).
[110] *See, e.g.*, Aleecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, I/S: A Journal of Law and Policy for the Information Society (2008).
[111] *See, e.g.*, Guy Zyskind, Oz Nathan, & Alex "Sandy" Pentland, *Enigma: Decentralized Computation Platform with Guaranteed Privacy* (2015), http://enigma.media.mit.edu/enigma_full.pdf.
[112] *See, e.g.*, Patrick Gage Kelley et al., A "Nutrition Label" for Privacy, 5 SYMP. ON USABLE PRIVACY & SECURITY, Article No. 4 (2009).
[113] *See* M. Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 Stan. L. Rev. Online 97, 101–02 (2013).
[114] *See* Effy Vayena & John Tasioulas, *Adapting Standards: Ethical Oversight of Participant-Led Health Research*, 10 PLoS Med. e1001402 (2013).
[115] *See* Ernst Hafen, Donald Kossmann & Angela Brand, *Health Data Cooperatives—Citizen Empowerment*, 53 Methods Info. Med. 82, 84 (2014); *see also* Effy Vayena & Urs Gasser, *Between Openness and Privacy in Genomics*, 13 PLoS Med. e1001937 (2016).

of time, which is creating heightened privacy risks for individuals and groups. A decomposition of the component risk factors can inform an analysis of the effects of the time dimension on big data risks, and determination of which interventions could mitigate these effects. As identified above, key risk drivers for big data that are related to the time dimension include the age of the data, the period of collection, and the frequency of collection. Other factors interacting with these characteristics, but not directly correlated with time, include the dimensionality of the data, the potential for broader analytic uses, the sample size, and the diversity of the population studied. An analysis of these factors reveals that commercial big data and government open data activities share many of the characteristics driving the privacy risks that have been studied with respect to long-term longitudinal research. However, the most commonly used privacy measures in commercial and government contexts, such as relying solely on notice and consent or de-identification, represent a limited subset of those the interventions available and are significantly different from the controls used in long-term research.

Compliance with existing regulatory requirements and implementation of commonly used privacy practices are arguably not sufficient to address the increased privacy risks associated with big data activities. For instance, traditional legal approaches for protecting privacy in corporate and government settings when transferring data, making data release decisions, and drafting data use agreements are time-intensive and not readily scalable to big data contexts. Technical approaches to de-identification in wide use are ineffective for addressing big data privacy risks. However, combining these approaches with additional controls based on exemplar practices in longitudinal research, and methods emerging from computational research, can offer robust privacy protection for individuals. Adopting new technological solutions to privacy can help ensure stronger privacy protection for individuals and adaptability to respond to new and sophisticated attacks, such as statistical inference attacks, that were unforeseen by regulators at the time that legal standards were drafted. New privacy technologies can also provide more universal and consistent privacy protection for individuals, compared to traditional approaches that can vary substantially based on the jurisdictions, industry sectors, actors, and categories of information involved. Technological approaches can be designed to comply with legal standards and practices, while also helping to automate data sharing decisions and ensure consistent and robust privacy protection at a massive scale.

Current practice for privacy protection in the human subjects research setting, as well as the Common Rule and policies governing IRB review processes, have shortcomings as well. However, there are opportunities for commercial actors to leap ahead of current privacy practice in research. In fact, some of the first implementations of advanced data sharing models providing formal privacy guarantees satisfying the differential privacy standard have been created by the government and industry. For instance, companies such as Google and Apple have begun deploying implementations of formal privacy models such as differential privacy within tools for gathering statistics about consumer behavior while protecting privacy. To support the implementation of new computational tools providing formal privacy guarantees, privacy regulations and policies should establish a privacy goal, rather than mandating the means by which privacy should be protected. It is important to consider what an organization aims to protect and to choose a specific privacy goal, against which various tools can be evaluated against and tailored to satisfy, rather than to adopt a privacy measure that offers a relatively limited type of privacy protection that may not be well-suited to the ultimate goal.

Different risk factors from big data can reduce the effectiveness of standard controls, and addressing each risk factor directly may not be efficient or effective. Addressing these risk factors may require emphasizing compensating controls or adopting emerging methods. Where several factors are working in concert to increase privacy risk, such as the combination of high-frequency, high-dimensional data with broader analytic uses, there are many unresolved challenges for existing controls and emerging methods. For instance, such contexts may limit the ability of individual data subjects to have meaningful understanding and control of data collections, uses, and disclosures, and make it difficult to prevent algorithmic discrimination. In these areas, it is especially important to continually review and adapt practices to address new risks and new analytic methodologies. Using a combination of controls to manage the overall risk resulting from identifiability, threats and vulnerabilities, is recommended. Several clusters of controls for addressing identifiability and sensitivity can be implemented, such as notice, consent, and terms of service mechanisms in combination with robust technical disclosure limitation techniques, formal application and review in combination with data use agreements and disclosure limitation techniques, and secure data enclaves with auditing procedures.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="4">Big Data Risk Drivers
Lower Risk → Higher Risk</td></tr>
<tr><td>Age, Period, Sample Size, Population Diversity</td><td>High Dimensional</td><td>High Frequency</td><td>High Dimensional & High Frequency</td></tr>
<tr><td rowspan="2">Intended Mode of Analysis</td><td>Statistical Analysis</td><td rowspan="2">Notice, Consent, Terms of Service; Formal Oversight</td><td colspan="2">Differential Privacy; Formal Oversight</td><td rowspan="2">Secure Data Enclave/Model Server; Restricted Access; Formal Oversight</td></tr>
<tr><td>Individual Analytics</td><td colspan="2">Personal Data Stores; Blockchain Audit Logs; Secure Multiparty Computation; Formal Oversight</td></tr>
</table>

Table 3. Examples of feasible privacy and security controls based on the risk drivers and intended mode of analysis identified in a big data use case.

With long-lived large-scale data collections, threats and vulnerabilities from data management continuing to evolve, and with them the privacy risks posed to individuals. At the same time, new technical approaches are emerging that provide better privacy and utility for big data. Thus there is a need for periodic review and evaluation of these activities, rather than review decisions that are made only at the beginning of a data collection program. It is important to adopt dynamic processes for continually reviewing and adapting decisions throughout the life of a data management program. Continual review and adjustment based on newly discovered risks and intended uses, has the potential to bring substantial

benefits for both privacy and utility.