Practical approaches to big data privacy over time

Micah Altman*, Alexandra Wood**, David R. O'Brien** and Urs Gasser***

Key Points

- Governments and businesses are increasingly collecting, analysing, 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, behaviour, and relationships and advance the state of science, public policy, and innovation.
- The collection and use of fine-grained personal data over time, at the same time, is associated with significant risks to individuals, groups, and society at large.
- This article examines a range of long-term research studies in order to identify the characteristics that drive their unique sets of risks and benefits and the practices established to protect research data subjects from long-term privacy risks.
- We find that many big data activities in government and industry settings have characteristics and risks similar to those of long-term research studies, but are subject to less oversight and control.
- We argue that the risks posed by big data over time can best be understood as a function of temporal factors comprising age, period, and frequency and non-temporal factors such as population diversity, sample size, dimensionality, and intended analytic use.
- Increasing complexity in any of these factors, individually or in combination, creates heightened risks that are not readily addressable through traditional de-identification and process controls.
- We provide practical recommendations for big data privacy controls based on the risk factors present in a specific case and informed by recent insights from the state of the art and practice.

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

Commercial and government actors are collecting, storing, analysing, and sharing increasingly greater quantities of personal information about individuals over progressively longer periods of time. Advances in technology, such as the proliferation of Global Positioning System (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 are more efficient and less costly than previous technologies and are contributing to large amounts of data being retained for longer

* Micah Altman, MIT Libraries, Massachusetts Institute of Technology, Cambridge, MA, USA.

** Alexandra Wood and David R. O'Brien, Berkman Klein Center for Internet & Society, Harvard University, Cambridge, MA, USA.

*** Urs Gasser, Berkman Klein Center for Internet & Society, Harvard University, Cambridge, MA, USA and Harvard Law School, Cambridge, MA, USA.

This material is based upon work supported by the National Science Foundation under Grant No. 1237235, the Alfred P. Sloan Foundation, and the John D. and Catherine T. MacArthur Foundation.

periods.[1] Powerful analytical capabilities, including emerging machine learning techniques, are enabling the mining of large-scale datasets to infer new insights about human characteristics and behaviours and driving demand for large-scale datasets. These developments make it possible to measure human activity at more frequent intervals, collect and store data describing longer periods of activity, analyse data long after the data were collected, and draw inferences from a large number of individual attributes. Moreover, analytic uses 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, putting pressure on traditional measures for protecting privacy.

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

Commercial and government data 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. These data are increasingly being made available to researchers, policymakers, and entrepreneurs, in support of rapid advances in scientific research, public policy, and innovation.[2]

These data encompass increasingly long periods of time and contain observations collected at increasingly frequent intervals, enabling insights that can only be derived from large, fine-grained datasets linked over time. Commercial big data are generated and used to provide goods and services to customers, and enable analytics to improve services and make investment or other business decisions.[3] Telecommunications providers, mobile operating systems, social media platforms, and retailers often collect, store, and analyse large quantities of data about customers' locations, transactions, usage patterns, interests, demographics, and more. Highly detailed personal information is used 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 building inspection records, and to improve their responsiveness and delivery of constituent services.[4]

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.[5] Other longitudinal studies have also had profound impacts on scientific understanding in fields such as psychology, education, sociology, and economics.[6] 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.[7]

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 is the subject of recent legal and policy debates spanning topics as varied as the right to be forgotten,[8] algorithmic discrimination,[9] and a digital dark

1 See President's Council of Advisors on Science and Technology, Big Data and Privacy: A Technological Perspective, Report to the President (May 2014), <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf> accessed 25 January 2018.

2 See Executive Office of the President, Big Data: Seizing Opportunities, Preserving Values (May 2014), <https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf> accessed 25 January 2018.

3 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://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf> accessed 25 January 2018.

4 See generally Stephen Goldsmith and Susan Crawford, *The Responsive City: Engaging Communities through Data-Smart Governance* (Jossey-Bass, San Francisco 2014).

5 Ralph B D'Agostino, Sr and others, 'General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study' (2008) 117 Circulation 743.

6 Erin Phelps, Frank F. Furstenberg and Anne Colby, *Looking at Lives: American Longitudinal Studies of the Twentieth Century* (Russell Sage Foundation, New York 2002).

7 See David Lazer and others, 'Computational Social Science' (2009) 323 Science 721.

8 See, eg Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1, Article 17 ("Right to erasure ('right to be forgotten')").

9 See, eg Megan Smith, DJ Patil and Cecilia Muñoz, 'Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights', Executive Office of the President Report (May 2016).

age.[10] These developments are likely to have far-reaching implications for the future of privacy in the big data era, and the full extent of their impact is yet to be seen.[11] However, long-term data activities have occurred for decades at a smaller scale in research settings in ways that are similar to commercial and government data activities.

Long-term human subjects research is 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 risks evolve in unanticipated ways over the course of a long-term study.[12] Nonetheless, the harms that are of greatest concern in an ethics review have been closely studied, reviewed, and controlled, and are well-documented in the policies of institutional review boards and the human subjects review literature. The literature recognizes that a disclosure of sensitive information from a study could expose an individual 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.

Data managed across longitudinal research, commercial, and government settings share a number of characteristics. Driving privacy law, policy, and practice across these settings is generally the need to protect individuals and groups from informational harms related to measuring and sharing information about them. Long-term longitudinal research is designed to draw inferences about the population being studied and involves collecting and protecting information about the same people (sometimes literally) that are described in corporate and government databases.

Corporations and governments use methods from research, such as questionnaires, observations of behaviour, and experiments, to generate data about their customers and constituents, through reliance on A/B testing, microtargeting, and individual service customization in big data analytics. Researchers, companies, and governments also collect data in substantially overlapping domains, from demographics, to opinions and attitudes, to readily observable behaviour, 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, and purchase history. Increasingly personal information involving sensitive topics is also collected in the commercial setting through web-based questionnaires and prompts, such as those utilized by online dating services or social media platforms. In some cases, such information can be readily inferred indirectly from data such as search queries, social media postings, or purchase histories.[13]

Uses across these settings also involve similar domains and objects of inference. While it may have once been the case that commercial data focused on consumer purchase behaviours and their 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. Corporations and governments frequently make population-level inferences that resemble research uses, most notably for the purposes of market analysis, but also for developing models of consumer behaviour that are used in targeted advertising and risk scoring.[14] Uses of personal information in the public sector include the creation of similar models for purposes such as predictive policing.

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 personal information will be disclosed, misused, or used in ways that will adversely affect them in the future. Examples of longitudinal research help illustrate ways in which long-term data activities drive heightened privacy risks. Long-term research studies are generally

10 See, eg Pallab Ghosh, 'Google's Vint Cerf warns of "digital Dark Age" ' (*BBC News,* 13 February 2015) <http://www.bbc.com/news/science-environment-31450389> accessed 25 January 2018.

11 See, eg John Podesta and others, 'Big Data: Seizing Opportunities, Preserving Values' Executive Office of the President Report (May 2014).

12 See M Ryan Calo, 'The Boundaries of Privacy Harm' (2011) 86 Ind LJ 1131.

13 See, eg Michael J Paul and 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 and David Stillwell, 'Computer-based Personality Judgments are more Accurate than those made by Humans' (2014) 112 Proceedings of the National Academy of Sciences 1036.

14 See, eg Pam Dixon and 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> accessed 25 January 2018.

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 lengthy narrative accounts 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, 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 often maintained at the level of an individual subject and used to track changes in each individual's health, socioeconomic, and behavioural 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.

As discussed in detail in the next section, 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 its analysis. Data may be analysed 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 analysed years after collection, including government records that are protected from disclosure for many decades.
- *Period*, referring to the length of the interval within which subjects are repeatedly measured. Some data collections make observations at a single point in time, 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-standing 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.

In addition, as discussed in the next section, four additional factors are increasing the privacy risks associated with big data in corporate and government contexts. Although these factors are not directly correlated with time, they are increasingly associated with longitudinal data collection.

- *Dimensionality*, or the number of independent attributes measured for each data subject. Examples of high-dimensional data include datasets of thousands of attributes that are maintained by data brokers and social media companies, while low-dimensional data may include official statistics that contain only several attributes within each record.
- *Analytic use*, or the mode of analysis the data are 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 or engaging in interventions such as recommending products to them.
- *Sample size*, referring generally to the number of people included in the set of data. Big data collection in corporate and government settings typically have much larger sample sizes than traditional longitudinal research studies, and such data sources typically constitute a substantially larger proportion of the population from which the sample is drawn.
- *Population characteristics*, referring 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.

Databases tend to grow along each of these dimensions, with expansions in long-term data collection, storage, and analysis. The current state of the practice for privacy protection addresses, and also fails to address, these developments in important ways. The risks that remain can be instructive for understanding where new interventions should be employed in the future.

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 handling personal data. While practices vary widely, among the most common approaches in commercial and government contexts are

notice and consent mechanisms and de-identification techniques. These approaches are often employed without further review or restrictions on use of data that have been obtained according to a terms of service agreement or been nominally de-identified. This reliance on a narrow set of controls, without continuing review and use of additional privacy interventions, differs significantly from longstanding practices in research settings.

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

Researchers and institutional review boards (IRBs) have implemented extensive review processes and a large collection of techniques for addressing the long-term risks of collecting and managing personal data about human subjects. Researchers must carefully consider the risks and benefits of their research activities to individual subjects and society at large, and choose among a wide variety of interventions to protect subjects responsibly. They must limit the collection of data to that which is necessary, restrict future uses of the data to those specified at the outset of the study, and minimize the disclosure of personal information. Researchers utilize a range of privacy controls, including explicit and informed consent, systematic review over time, statistical disclosure control techniques to limit learning about information specific to individuals, data use agreements, and procedural controls to limit data access and use.

Legal and regulatory frameworks for the oversight of human subjects research, in combination with sector-specific information privacy laws, lead 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 research participants. The Common Rule[15] applies to research funded by one of the federal agencies that have subscribed to the rule or conducted at an institution that has agreed to comply with the regulations. Researchers may also be governed by state laws protecting the rights of human research subjects,[16] or by the research data policies of their home institutions, sponsors, and prospective journals.

Researchers conducting studies involving human subjects typically must submit their proposals for review by an IRB and follow certain consent and disclosure limitation procedures. They must demonstrate that subjects will be informed of the nature, scope, and purpose of the study; specify the types of personal information to be collected, the research and data management procedures to be followed, and steps to be taken to preserve confidentiality; and describe any risks and benefits related to participation in the study. IRBs evaluate informed consent procedures and potential risks and benefits to research subjects, and determine whether subjects are adequately informed of potential risks and whether the benefits outweigh the risks. In a long-term research study, IRBs conduct continuing review, with reviews conducted on an annual or more frequent basis.[17]

Because studies often explore sensitive topics related to the development, behaviour, and health of individuals, collected data often fall into a category of information protected by law. A large number of federal and state laws protect data privacy, though the rules vary substantially depending on the actors, funding sources, types of information, and uses involved. 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),[18] the Health Insurance Portability and Accountability Act (HIPAA),[19] and the federal alcohol and drug abuse confidentiality regulations,[20] respectively. In addition, researchers collecting data on behalf of federal agencies are required to establish privacy safeguards in accordance with laws such as the Privacy Act of 1974[21] and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[22]

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. Consent from research subjects is preserved in the form of a written, legally enforceable contract that documents the scope of the study authorized. A researcher's interactions with a subject and uses of personal information are limited to the research purposes described in the consent form. When the scope of information collected or analysed is expanded, the study extends beyond the

15 45 CFR pt 46.
16 See, eg California Health and Safety Code s 24170 et seq.
17 45 CFR 46.109(e).
18 20 USC s 1232g; 34 CFR pt 99.
19 45 CFR Part 160 and Subparts A, C, and E of pt 164.
20 42 CFR pt 2.
21 5 USC s 552a.
22 Pub L 107-347, Title V; 44 USC s 3501 note.

timeframe initially disclosed, or the membership of the population being studied changes, the investigators must obtain new consent from participants.[23]

Modifications may result as new research questions emerge. Approximately 35 years into the Framingham Heart Study, researchers began collecting and analysing DNA from participants' blood samples and immortalized cell lines, due to new interest in exploring the genetic factors underlying cardiovascular disease.[24] Research data are often rich enough to support analysis methods and research questions not originally envisioned at the time of the original proposal. Researchers maintain a detailed record of consent for each subject and, upon each new data collection or use activity, confirm whether it is authorized by the consent on file. Framingham Heart Study participants provide consent authorizing various potential research activities, including cell line creation, sharing of genetic data with researchers, and sharing of genetic data with private companies.[25] Consent forms also enable participants to authorize use for specific types of research, such as heart and blood diseases, and potentially sensitive research involving reproductive health, mental health, and alcohol use.[26]

Data use agreements limiting data access and use have been widely adopted by academic institutions, data repositories, and data enclaves. 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 penalties. However, oversight and enforcement of the terms of such agreements are persistent challenges.

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 are used in the long-term research setting, though there is significant variation in practices. Research institutions often implement security plans and confidentiality training programmes for the individuals who will have access to personal data over the course of a study. 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. Researchers may transform data at collection or retention 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.

When sharing long-term research data, multiple disclosure limitation techniques are often used in combination. The sensitive nature of the data and the potential to draw linkages to external sources often precludes the dissemination of data in raw, identifiable form. Tiered access may be used to provide public access to de-identified datasets and restricted-use datasets to trusted researchers upon application. Researchers use a variety of statistical disclosure limitation techniques, such as aggregation, suppression, and perturbation, to produce a de-identified public-use dataset.[27] These techniques address some risks, but there is a growing recognition that such techniques provide only limited protection over the long term.[28] Such approaches are generally designed to address specific types of attacks such as record linkage using known sources of auxiliary information, leaving data vulnerable to other types of attacks. Traditional approaches are also likely to result in the redaction or withholding of useful information.

The tradeoff between data privacy and utility is more acute for long-term longitudinal data. Models for assessing disclosure risk have been developed with cross-sectional data, ie data collected at one point in time or without regard to differences in time, in mind, and are poorly suited for addressing longitudinal data privacy risks.[29] Techniques that are effective for cross-sectional datasets often result in either weaker privacy protections or a greater reduction in data utility when

23 See, eg 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> accessed 25 January 2018.

24 Diddahally R Govindaraju and others, 'Genetics of the Framingham Heart Study Population', (2008) 62 Advances in Genetics 33.

25 Daniel Levy and others, 'Consent for Genetic Research in the Framingham Heart Study' (2010) 152A Am J Med Genetics Part A 1250.

26 See ibid.

27 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, eg 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> accessed 25 January 2018.

28 See Arvind Narayanan and Vitaly Shmatikov, 'Robust De-anonymization of Large Sparse Datasets' Proceedings of the 2008 IEEE Symposium on Security and Privacy 111 (2008).

29 See Lawrence H Cox, Alan F Karr and Satkartar K Kinney, 'Risk-utility Paradigms for Statistical Disclosure Limitation: How to Think, But not How to Act' (2011) 79 Intl Stat Rev 160.

applied to longitudinal data.[30] Because longitudinal studies collect data at the level of an individual for the purposes of studying patterns in individual behaviour, data that have been aggregated or summarized in, eg cross-tabulation tables, are often not suitable for analyses not anticipated by the producer of the aggregate dataset. Traditional techniques are likely to change the structure of longitudinal data in ways that sharply influence future statistical models and inferences and may make certain types of analysis impossible. Moreover, privacy transformations made to a dataset are not always disclosed to the public. Secondary researchers may unwittingly treat a sanitized data set as an unmodified data set, leading to unanticipated and unacknowledged effects on their results.

Newly-emerging technical approaches, such as synthetic data generation and differential privacy, are less widely utilized to protect the confidentiality of long-term longitudinal data, though new techniques are being developed. Differential privacy, for example, has typically been studied in the context of a dataset that has been 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 studies, in which datasets are continuously collected and analysed over time. Although a similar model, the continual observation model,[31] is flexible enough to describe longitudinal studies, research to date has typically assumed that one person's information affects only a limited number of stages of the study.[32] 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.[33]

Due to the difficulty of mitigating risks associated with future releases of data over the course of a long-term study, researchers often implement restrictive access controls. Data repositories or secure enclaves, together with 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 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. Researchers may be required to participate in confidentiality training or to demonstrate compliance with a data security plan. Particularly for large institutions, 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.

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

Review processes and safeguards employed for long-term data activities in commercial and government settings differ from those used in the research context (Table 1). Businesses and governments generally consider privacy risks at the time they initiate a data collection programme, but in most cases they do not engage in systematic and continual review with long-term risks in mind.[34] They often rely heavily on certain controls, such as notice and consent or de-identification, rather than drawing from the wide range of privacy interventions that are available and applying combinations of tailored privacy controls throughout the information lifecycle.

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 emerged and evolved under very different regulatory and policy frameworks. Sector-specific information privacy laws such as FERPA and HIPAA play a significant role in protecting research data but apply directly to only a small portion of commercial and government data activities. Nonetheless, some organizations elect to

30 See, eg Benjamin CM Fung and others, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques* (Chapman & Hall/CRC Press, New York 2010).

31 Cynthia Dwork and others, 'Differential Privacy under Continual Observation' Proceedings of the 42nd ACM Symposium on Theory of Computing 715 (2010).

32 See, eg T-H Hubert Chan, Elaine Shi and Dawn Song, 'Private and Continual Release of Statistics' 5 ACM Transactions on Information and System Security A:1 (2011); Prateek Jain, Pravesh Kothari and Abhradeep Thakurta, 'Differentially Private Online Learning' 23 Proceedings of the 25th Conference on Learning Theory 24.1 (2012).

33 See Ashwin Machanavajjhala and others, 'Privacy: Theory Meets Practice on the Map' Proceedings of the 2008 IEEE 24th International Conference on Data Engineering 277 (2008).

34 For a specific example, see the discussion in the sub-Section '*A real-world example illustrates the lack of systematic evaluation of and protection against privacy risks in the commercial context*' below.

Table 1. Comparison between typical privacy controls for similar longitudinal data when managed within research setting versus commercial and government contexts

	Research settings	Commercial and government settings
Legal and ethical frameworks	• Activities governed by strict ethical and legal frameworks, including oversight by an IRB. • Clear responsibilities assigned to IRBs, hosting institutions, and principal investigators.	• Actors operate within a legal framework that has arguably been slower to evolve to address data privacy and ethical challenges. • No broadly applicable regulations governing data management are in place. • Oversight responsibility is often unspecified.
Risk assessment	• Systematic risk assessment by IRBs and curation by research investigators.	• Actors rarely engage in systematic review of privacy risks and planning for long-term review, storage, use, and disclosure.
Controls	• Researchers incorporate multiple layers of protection, including explicit consent, systematic design and review, statistical disclosure control, and legal/procedural controls.	• Actors rely on a narrower subset of privacy controls such as notice and consent and de-identification.
Approach to big data	• Risk analysis and controls are not adapted to scale of big data.	• Actors are increasingly aware of and engaged with adapting workflows and policies to big data

adopt as a best practice some of the safeguards required by such laws. Their data activities may also be governed by other laws that rarely apply to researchers, such as the Fair Credit Reporting Act,[35] the Children's Online Privacy Protection Act,[36] and Federal Trade Commission (FTC) enforcement under section 5 of the FTC Act.[37] Laws such as the Privacy Act of 1974,[38] the Confidential Information and Statistical Efficiency Act,[39] and the Freedom of Information Act,[40] as well as corresponding state laws, govern activities involving certain categories of government information. State privacy laws, such as data breach notification laws,[41] may also apply but typically require limited safeguards and a relatively narrow scope of protection covering direct identifiers such as names and Social Security numbers. Data security standards are also established by the Federal Information Security Management Act (FISMA),[42] and by bodies such as the National Institute of Standards and Technology.[43] These laws and standards grant substantial discretionary authority to agencies, leading to wide variations in practice.

Industry best practices, such as the Payment Card Industry Data Security Standard,[44] and the Health Information Trust Alliance (HITRUST) framework,[45] are widely applied, though they focus almost exclusively on data security requirements, such as encryption and access controls, rather than privacy protections limiting what can be learned about individuals once access to the data has been granted. Industry actors also refer to the fair information practice principles for guidance,[46] though these principles are 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, a highly-regulated industry with a long history and significant experience handling large quantities of very sensitive information about individuals. While these general principles are widely referenced in many

35 15 USC s 1681.

36 15 USC ss 6501–06.

37 15 USC s 45.

38 5 USC s 552a.

39 Pub L 107-347, Title V; 44 USC s 3501 note.

40 See Freedom of Information Act, 5 USC s 552, and corresponding sunshine laws at the state level.

41 See, eg Cal Civ Code ss 1798.29, 1798.80 et seq; Mass Gen Laws s 93H-1 et seq; NY Gen Bus Law s 899-aa, NY State Tech Law 208.

42 44 USC ss 3541 et seq.

43 See, eg NIST, FIPS Pub 199, Standards for Security Categorization of Federal Information and Information Systems (2004).

44 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>.

45 The Health Information Trust Alliance (HITRUST) Common Security Framework (2016).

46 See, eg Testimony of Jeremy Cerasale, 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?' (18 December 2013).

contexts, they are often not consistently implemented with specific and strong policies and controls.

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

Legal and ethical frameworks for commercial data are not as well-defined as those for human subjects research. The lack of formal review and oversight by an IRB or similar external governance body results in less emphasis on informing subjects of risks and benefits, minimizing data collection and disclosure, and implementing strong controls to address long-term risks. Consumers lack an understanding of the full extent to which their personal data are collected, linked, analysed, shared, and reused by third parties. Commercial data are frequently linked with data from other sources and disclosed to third parties, whereas research data may not be linked with other data except in limited circumstances outlined in the research proposal, reviewed and approved by an IRB, and disclosed and consented to by the individual subjects. In contrast to the informed consent process used in research, commercial practices are typically authorized through privacy policies or terms of service, which contain broad language that is not reviewed closely, if at all, by most consumers.

Consent is widely used in commercial settings, and frequent interactions with users provide opportunities to renew consent. However, there is a growing recognition that a reliance solely on notice and consent is inadequate. Consumers often do not read or understand privacy policies, and the terms of such policies are often written so broadly or vaguely as to not fully inform those who do read them. Many state laws, federal regulations, and data sharing agreements establish notice duties and penalties in the event of data breaches. 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. 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 recognized as worthy of redress, deterrence, or punishment, such as a concrete financial loss that has been incurred, and it may be difficult to prove that a disclosure directly caused a particular harm.

Government agencies are generally required to consider the legal and ethical implications of disclosing information about individuals; review their data collection, storage, and disclosure practices; and implement appropriate privacy safeguards. However, applicable laws are context-specific, are limited in scope, and lack specificity regarding how to apply appropriate privacy and security measures in a particular setting.[47] Risk–benefit assessments are performed by statistical agencies and by agencies required to conduct privacy impact assessments, but other activities require less systematic consideration of long-term risks. Open data initiatives, which call for open access to be the 'default state' for information and proactively release data to the extent the law allows,[48] lead to data collected for a specific purpose, such as delivering constituent or emergency services, being made available for use by the public for any purpose. These policies are, in large part, enabled by federal and state sunshine laws including the Freedom of Information Act,[49] which require disclosures in response to public records requests provided that no law prohibits the 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 typically do so by redacting records of direct identifiers such as names, addresses, and Social Security numbers. Due in large part to the lack of detailed guidance, redaction processes are typically performed in an ad hoc fashion and practices vary significantly between agencies. Individuals are typically not informed of the specific disclosure of their personal data, nor of the associated benefits and risks, prior to release.

Most companies and government agencies do not implement procedures for long-term review and mitigation of risks. Some corporate policies recognize the need to renew notice and consent for expanded data collections or uses. Google requires third-party developers using its services to update their privacy policies and reobtain consent from users if they plan to access or use a type of user data that was not disclosed in their privacy policy,[50] or use their data in a new way or for a different purpose than previously disclosed.[51] While policies at large data companies are evolving to address

47 See Micah Altman and others, 'Towards a Modern Approach to Privacy-Aware Government Data Releases' (2015) 30 Berkeley Tech LJ 1967.

48 See, eg Exec Order No 13,642, 3 CFR 244 (2014) (Making Open and Machine Readable the New Default for Government Information).

49 See 5 USC s 552.

50 See Google API Services: User Data Policy, <https://developers.google.com/terms/api-services-user-data-policy> accessed 19 June 2017.

51 See ibid.

these concerns, common practices in this area generally fall short of the careful design and independent ethics review processes characteristic of longitudinal studies in the research setting. The harms associated with commercial big data programmes have arguably not been studied and debated to the same extent that the harms associated with long-term research studies have long been assessed by IRBs and principal investigators.

As they adopt data-driven business models and respond to high-profile data breach incidents that are occurring with increasing frequency, corporations are increasingly incorporating risk-benefit assessments and stronger data security and privacy safeguards. A number of large companies are beginning to implement internal ethical review processes.[52] Facebook has established an ethics review process for research based on its user data. 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,[53] though they have made only a small fraction of the attributes they hold viewable. 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.[54] In addition, bias or discrimination in the use of personal data is receiving growing attention. Companies such as Airbnb, in response to reports and research findings of discrimination by their users,[55] are restricting flows of personal data they hold and encouraging uses that rely less on the viewing of personal information of other users.[56]

Large technology companies have also begun instituting internal ethics review processes for their big data activities, though such mechanisms do not fully mirror the level of review and protection provided by a formal, independent IRB process.[57] One prominent example is Facebook, which has established an internal research ethics review process influenced by the ethical principles outlined in the Belmont Report and reflected in the Common Rule.[58] It involves training for employees on privacy and research ethics, review by senior managers with substantive expertise in the area of proposed research, and, where necessary, extended reviews by a committee of substantive area experts and experts in law, ethics, communications, and policy.[59] The value of the research to Facebook, the Facebook community, and society at large, its contributions to general knowledge, and other 'positive externalities and implications for society' are considered.[60] Against these benefits, the committee weighs potential adverse consequences from the research, especially with respect to vulnerable populations or sensitive topics, and 'whether every effort has been taken to minimize them'.[61] Other criteria include '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.[62]

A real-world example illustrates the lack of systematic evaluation of and protection against privacy risks in the commercial context

Another notable commercial example is Acxiom, which holds what is by some measures the largest commercial database on consumers in the world.[63] The company collects, combines, analyses, and sells sensitive personal data from a number of sources including public records, surveys and questionnaires, retail purchases, web browsing cookies, and social media postings. To protect the sensitive data it holds, Acxiom complies with a number of regulatory and industry standards, including those found in HIPAA, HITRUST, NIST, and PCI frameworks. The company also purports to engage in 'very rigorous' privacy impact assessments with a focus on

52 See Molly Jackman and Lauri Kanerva, 'Evolving the IRB: Building Robust Review for Industry Research' (2016) 72 Wash & Lee L Rev Online 442.

53 See Natasha Singer, 'Acxiom Lets Consumers See Data It Collects' (The New York Times, New York 2013) B6.

54 See Úlfar Erlingsson, Vasyl Pihur and Aleksandra Korolova, 'RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response' Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (2014) <https://research.google.com/pubs/archive/42852.pdf> accessed 25 January 2018; Andy Greenberg, 'Apple's 'Differential Privacy' Is About Collecting Your Data–But Not Your Data' (*Wired*, 13 June 2016) <https://www.wired.com/2016/06/apples-differential-privacy-collecting-data> accessed 25 January 2018.

55 See, eg Benjamin Edelman, Michael Luca and Dan Svirsky, 'Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment' (2017) 9 American Economic Journal: Applied Economics 1.

56 See Katie Benner, 'Airbnb Has Enlisted Hosts; Now It Must Fight Their Bias' (The New York Times, New York, 9 September 2016) A1.

57 See, eg Zoltan Boka, 'Facebook's Research Ethics Board Needs to Stay Far Away from Facebook' (*Wired*, 23 June 2016) <https://www.wired.com/2016/06/facebooks-research-ethics-board-needs-stay-far-away-facebook> accessed 25 January 2018.

58 See Jackman and Kanerva (n 52) 442.

59 Ibid at 451–53.

60 Ibid at 452–53.

61 Ibid at 455.

62 See ibid at 455.

63 See Natasha Singer, 'Mapping, and Sharing, the Consumer Genome' *The New York Times* (16 June 2012) <http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html> accessed 25 January 2018.

ethical use of data, involving 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'.[64]

While these descriptions imply a robust, careful, and systematic ethical review process, there are indications it is applied in an ad hoc fashion in practice. 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'.[65] 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.[66] This example illustrates how, despite commitments to rigorous processes, in practice decisions may be based on ad hoc, gut judgments by a small number of decision-makers and their opinions about unspecified 'social norms'.[67] 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.[68] While this is just one example, it is emblematic of the lack of systematic data privacy reviews across a large number of commercial and government settings.

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

Long-term data activities affect privacy risk in different ways and threaten to further erode the effectiveness of traditional approaches to privacy. Scholars and practitioners are now exploring new technical, procedural, and legal interventions for managing data privacy that can complement traditional approaches and better address emerging challenges.

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. Many concepts are embedded in a notion of privacy risk, and decomposing the relevant dimensions and analysing them separately can inform the selection among interventions that address different risk drivers. Examining privacy risk as a function of three separate dimensions—identifiability, threats, and vulnerabilities—where threats and vulnerabilities are often bundled together in discussions of information sensitivity, is instructive.[69] Privacy risk is not simply an additive function of these components. Identifiability and sensitivity may be better modelled as multiplicative factors, and their effects are not evenly distributed in the population, as some individuals may be more vulnerable to particular threats.

A review of various literatures guiding IRB practice, describing methodologies for data management in longitudinal research, and presenting findings from the scientific study of privacy, taken together, indicate at least three characteristics related to time as components that influence data privacy risk: age, period, and frequency. Table 2 summarizes the complex

64 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)' (25 May 2016) <http://www.ncvhs.hhs.gov/transcripts-minutes/transcript-of-the-may-25-2016-ncvhs-subcommittee-on-privacy-confidentiality-security-hearing> accessed 25 January 2018.

65 See ibid.

66 See ibid.

67 See ibid.

68 See Testimony of Pam Dixon, 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?' (18 December 2013).

69 See Altman and others (n 42).

Table 2. Key risk drivers for big data over time and their effects on privacy risk components

	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

relationship between these temporal characteristics and the components of privacy risk (identifiability, threats, and vulnerabilities), as elaborated below.

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

Older information is often argued to reduce the risk of identifiability, as individuals' observable characteristics generally change, and the availability and accuracy of data have historically decreased, over time. For instance, an individual who currently has red hair may not possess this attribute 30 years later, making this attribute less identifying with time. Arguably, this is a weak reduction, as some characteristics such as DNA do not appreciably change over time. The availability of external sources of information that could be used to infer sensitive information about individuals in a dataset is growing. Real estate and criminal records created decades ago are being digitized and made publicly available online, lowering the barrier to access and enabling uses far removed from the contexts likely envisioned at the time the data were created.[70] These factors challenge traditional approaches to privacy such as de-identification, the efficacy of which depends in large part on limited sources of external information that can be linked to information in a dataset.

Data are stored over extended timeframes, increasing the risk of data breach. Industry standards often require the encryption of data in storage, and applicable laws may require encryption where it can be reasonably implemented. As the time between data collection and use increases, the potential for applying the data to uses that could not be anticipated at the time of collection grows, thereby increasing threats from data use. New analytical methods, such as machine learning and social network analyses, can unlock new uses of information originally collected for different purposes. Social media postings are being used to track the spread of illnesses, measure behavioural risk factors, and infer individuals' personality traits.[71]

The more time that elapses between collection and use, the greater the likelihood that circumstances will change, creating challenges related to obtaining consent, complying with privacy laws and regulations, and disclosing risks to data subjects. Search engines evaluate removal requests in accordance with the right to be forgotten in the European Union, based on criteria such as accuracy, adequacy, relevance, and proportionality of the content, and the age of the data is one of the factors considered when weighing relevance.[72] Guidelines on evaluating this criterion reflect considerations such as an individual's status as a public figure changing over time, or the significance of a conviction for a minor crime diminishing after many years have passed.[73] In other situations, the age of the data is less relevant, such as when an individual has a conviction for a violent or fraud-related crime and is applying for an employment position involving interaction with children or financial responsibility.[74]

Older data may also leave some individuals in the data less vulnerable to privacy-related harms. Information in the distant past may be commonly viewed as less relevant and less likely to cause harm. Consider, for instance, the magnitude of potential harm from a release of a high school student's grades at a time when the subject is a high school student or a recent graduate, versus such a disclosure 30 years later. This rationale is reflected in the Census Bureau's procedures for protecting confidentiality, which consider potential distortions in information due to the passage of time as one factor in the risk analysis.[75] When the age of the data is great enough, the subjects will be deceased and unaffected by many of the consequences of personal

70 See generally Federal Trade Commission, Data Brokers: A Call for Transparency and Accountability (May 2014).

71 See, eg Paul and Dredze (n 13).

72 See Art 29 Data Protection Working Party, Guidelines on the Implementation of the Court of Justice of the European Union Judgment on 'Google Spain and Inc. vs. Agencia Española de Protección de Datos (AEPD) and Mario Costeja Gonzalez' C-131/12, 14/EN WP 225 (2014).

73 See Advisory Council to Google on the Right to be Forgotten, <https://archive.google.com/advisorycouncil/advisement/advisory-report.pdf> (2015) accessed 25 January 2018.

74 See ibid.

75 Susan M Miskura, 'Disclosure Avoidance in Publication of Race and Hispanic Origin Data' Census 2000 Informational Memorandum No 54 (8 May 2000).

information disclosure, though one must consider its impact on the subject's children or on groups to which he or she belonged. Individually identifying information from decennial census records is restricted for 72 years from the date of collection, after which the risk is considered to be low enough to permit its public release.[76] Similarly, laws such as the Common Rule apply only to data about living individuals,[77] and research studies are sometimes designed to release data after a substantial period of time has passed.

Long periods of data collection, ie data that describe trends, create additional privacy risks

Data collected over an extended period may result in increased threats, as they enable analysis of trends over time revealing sensitive characteristics related to health, behaviour, and interests. In the Framingham Heart Study, investigators have been continuously collecting data from participants and their descendants since 1948, and the data 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.[78] Another example is the Panel Study of Income Dynamics, which has relied on the continuous collection of data since 1968 and covers topics such as employment, income, wealth, expenditures, health, marriage, and childbearing, among others, and has enabled research to understand individuals' and families' socio-economic, health, educational, marital, and consumption trajectories and the factors influencing them.[79] Extensive controls on access to and use of detailed data from this study, including secure data enclaves, have been implemented, due in large part to concerns about risks from multiple measurements about individuals over a long period.[80]

There are additional, weaker risks related to interactions between period and other dimensions. 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. Because human behaviour exhibits patterns at multiple temporal scales, the interaction of extended period of collection and high frequency may enable increased detection of trends, further increasing threats and enabling stronger behavioural fingerprinting, thereby increasing identifiability.[81]

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

Data collected at frequent intervals can also reveal identifiable or sensitive details about individuals. Mobile health apps and devices use sensors to continuously monitor and record features related to an individual's health and behaviour. For example, a research study in 2015 monitored the activity levels of multiple sclerosis patients in their daily lives, by continuously measuring the number of steps and distance walked by patients wearing activity tracking devices.[82] High-frequency data dramatically increases identifiability, with as few as four data points on an individual's spatiotemporal location or retail purchases being sufficient to uniquely identify her records.[83] In many cases, commercial and government big data collection leads to much more frequent observations than those collected in the research setting. For example, microphones, cameras, accelerometers, GPS receivers, 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.

Increases in frequency in data analysis and release—defined broadly to include internal uses of personal information within an organization, publications of data and statistical summaries of data, inadvertent leakages of data, and analyses of released data—also create heightened challenges for privacy protection. More frequent analysis of data about the same individuals inherently increases the risk of learning information specific to the individuals in the data.[84] High-frequency data collection, while not a privacy threat by itself, is a

76 44 USC s 2108(b).

77 See 45 CFR s 46.102.

78 See History of the Framingham Heart Study, <https://www.framinghamheartstudy.org/about-fhs/history.php> accessed 25 January 2018.

79 See Katherine A McGonagle and others, 'The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research' (2012) 3 Longitudinal and Life Course Studies 268.

80 See Panel Study of Income Dynamics, Process and requirements for obtaining restricted data, <https://simba.isr.umich.edu/restricted/ProcessReq.aspx> accessed 14 August 2017; Katherine McGonagle and Robert Schoeni, 'The Panel Study of Income Dynamics: Overview and Summary of Scientific Contributions After Nearly 40 Years', Technical Series Paper #06-01 (2006), <https://psidonline.isr.umich.edu/publications/Papers/tsp/2006-01_PSID_Overview_and_summary_40_years.pdf> accessed 25 January 2018.

81 See, eg Nathan Eagle and Alex (Sandy) Pentland, 'Reality Mining: Sensing Complex Social Systems' (2006) 10 J Personal and Ubiquitous Computing 255; Nathan Eagle and Alex (Sandy) Pentland, 'Eigenbehaviors: Identifying Structure in Routine' (2009) 63 Behav Ecol & Sociobiol 1057.

82 See James McIninch and others, 'Remote Tracking of Walking Activity in MS Patients in a Real-World Setting', (2015) 84 Neurol Supp P3.209.

83 See Yves-Alexandre de Montjoye and others, 'Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata' (2015) 347 Science 536 Yves Alexandre de Montjoye and others, 'Unique in the Crowd: The Privacy Bounds of Human Mobility', (2013) 3 Nature Sci Rep 1376.

84 This observation has been referred to as the fundamental law of information recovery, which states, informally, that '"overly accurate" estimates of "too many" statistics completely destroys privacy'. See, eg Cynthia

prerequisite for release, and is associated with threats from analysis and release, including data breaches. Many approaches to privacy in common use, such as de-identification, are insufficient to protect against such risks across multiple releases and analyses.

Although the capacity to collect and analyse data at greater frequency carries clear benefits for research and innovation, it can also 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 other temporal dimensions. High frequency data collection may interact with long periods of data collection, increasing the threats from data release due to the uniqueness of individual records in a database. It can also interact with the age of the data, which may be associated with a greater availability of external sources of information that could be linked to individual records.

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 current practice and the state of the art for privacy protection are independent of the extended timescale of data activities. Table 3 summarizes the relationship between components of privacy risk and non-temporal characteristics of big data, including the size and diversity of the population being studied, the size of the sample, the dimensionality of the data, and the broader analytic uses enabled by the data.

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

Dimensionality, or the number of independent attributes for each data subject, is a factor correlated with privacy risk. While from a mathematical standpoint high-dimensional data is broadly defined and encompasses high-frequency data, this article's discussion of dimensionality focuses on the number of *independent* attributes measured for each data subject in order to analyse the risks associated with these factors separately. This makes it possible to analyse how high-frequency behavioural data carry heightened risks due to the availability of auxiliary information about the same attributes, the potential to create durable behavioural fingerprints for individuals, and the likelihood of repeated measures of the same characteristic to reveal attributes of behaviour that are expressed over time.

The composition of even seemingly benign measures can create unique patterns, or 'fingerprints' of behaviour that can be used to link a named individual with a distinct record in the data. For instance, although names were not provided in a 2006 release of AOL Inc. search query records, sets of search queries alone were found to be revealing of an individual's identity and sensitive personal attributes.[85] A release of research data extracted from Facebook profiles also enabled identification based on the large number of attributes provided by each profile.[86] A dataset of movie ratings by Netflix users was vulnerable to re-identification attacks due to the number of ratings from each user making many of the records in the dataset unique and identifiable by cross-reference with other information.[87] 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.[88] Examples of algorithmic discrimination raise questions about the use of personal information to classify people in ways that may harm individuals and groups.[89]

There are few constraints on linking commercial data, compared to those imposed in the research context. Companies often have an incentive to combine as many data points as possible, and draw linkages at the individual level, in order to compile more accurate profiles about individuals in their databases. Data brokers accumulate and link data about the same individuals from many different sources, including administrative records from multiple government agencies and commercial data from other sources. The profiles compiled are sold to other businesses, including banks, automotive companies, and department stores,[90] and linked to even more data sources. As of 2012, Acxiom purportedly held data on approximately 500 million individuals, including

Dwork and Guy N.Rothblum, 'Concentrated Differential Privacy', Working Paper (2016), <https://arxiv.org/abs/1603.01887v2> accessed 25 January 2018.

85 See Michael Barbaro and Tom Zeller, Jr, 'A Face Is Exposed for AOL Searcher No. 4417749' (The New York Times, New York, 9 August 2006) A1.

86 Michael Zimmer, 'But the data is already public': On the Ethics of Research in Facebook' (2010) 12 Ethics and Infor Technol 313

87 See Narayanan and Shmatikov (n 28).

88 See Carter Jernigan and Behram FT Mistree, 'Gaydar: Facebook Friendships Expose Sexual Orientation' 14 First Monday 10 (2009).

89 See, eg Julia Angwin and others, 'Machine Bias' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 25 January 2018; Alessandro Acquisti, Laura Brandimarte and George Loewenstein, 'Privacy and Human Behavior in the Age of Information' (2015) 347 Science 509.

90 Natasha Singer, 'Mapping, and Sharing, the Consumer Genome' (*New York Times*, 16 June 2012) <http://www.nytimes.com/2012/06/17/technology/acxiom-the- quiet-giant-of-consumer-database-marketing.html> accessed 25 January 2018.

Table 3. Non-temporal characteristics of big data that drive privacy risks and their effects on components of privacy risk

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

about 1500 pieces of information about each person,[91] and these figures are likely much higher today. It also offers tools to advertiser clients for linking data across databases, even where there are variations between records due to a change of name or address.[92]

The dimensionality of big data may be greater than it first appears, as richer data types, such as network data, unstructured text, audio, and video, are subject to multiple independent measurements. Informational risks from social network analyses are a function not only of the nodes but also the structure of the network connections.[93] Pieces of text may be associated with metadata (eg Twitter posts may have embedded geolocation codes), may embed direct identifiers such as names (eg 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, and heart rate can be detected using smartphone video cameras. Research has shown it may even be possible to extract conversations from video recorded images of vibrations on surrounding materials, or to determine the occupancy of a room based on Wi-Fi signal strength. Data may also be highly precise, enabling linkages that are only possible at fine levels of precision; high-precision geolocation data, for example, can reveal the exact residence or business an individual visited.

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

Another driver of privacy risk in big data is the potential for broader analytic uses, including uses that could be described as algorithmic discrimination. New big data sources and analytical techniques, including a wide variety of datamining algorithms, are enabling classification of individuals in areas such as online behavioural advertising to credit decisions to crime forecasting by law enforcement to predictive healthcare analytics, in ways that reflect or even amplify inherent bias.[95] Examples of algorithmic discrimination fall along a spectrum based on the extent to which the discrimination is enabled by learning about individual characteristics. An example from one end of the spectrum is differential pricing, whereby firms aim to generate individualized predictions or interventions based on personal 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. An example at the other end of the spectrum is fraud detection, in which the goal is inherently to make inferences about the predicted behaviour of a specific individual. Such predictions and interventions expand the set of threats that must be considered beyond those that arise from population estimates, through differential pricing, redlining, recidivism scores, or micro-targeted advertising.

Traditional approaches to privacy such as individual control, consent, and transparency fail to adequately address problems of discrimination. De-identification techniques, as well as techniques based on formal privacy models such as differential privacy, are designed to enable accurate estimations of population parameters and do not protect against learning facts about populations that could be used to discriminate. These techniques are also not suited where the goal is explicitly to make inferences about or intervene with respect to individuals.

91 See ibid.

92 Jim Edwards, 'Facebook's Big Data Partner Knows Who You Are Even When You Use a Different Name on the Web' (*Business Insider*, 26 September 2013) <http://www.businessinsider.com/facebook-and-acxioms-big-data-partnership-2013-9> accessed 25 January 2018.

93 See, eg Lars Backstrom, Cynthia Dwork and Jon Kleinberg, 'Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography' Proceedings of WWW 2007 (2007).

94 See, eg Ahmed Al Faresi, Ahmed Alazzawe and Anis Alazzawe, 'Privacy Leakage in Health Social Networks' (2014) 30 Computational Intelligence 514.

95 See, eg Megan Smith, DJ Patil and Cecilia Muñoz, 'Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights' Report from the White House Executive Office of the President (2016).

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

Increases in sample size are associated with greater identifiability risks. 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. As samples grow to be very large, they are quite likely to include members of vulnerable populations, who may be entitled to special protection. Reflecting ethical principles, the Common Rule establishes greater protections for vulnerable populations, such as pregnant women, children, and prisoners.[96] Commercial entities not bound by these regulations routinely collect and analyse personal information about such groups.[97]

When the size of the sample is large enough with respect to the overall population, it can even support uses such as surveillance, generating more than moderate risks of harm for society and democratic institutions. 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 for a dataset containing only US individuals, though it could for a dataset of a broader population that includes individuals living under non-democratic regimes.

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

In order to ensure robust protection of privacy, similar risks should be addressed similarly. This requires applying principles for balancing privacy and utility in data releases more systematically. Risk-benefit assessments 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.

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 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 conceptual purposes, 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. Real-world data management should be designed to utilize appropriate tools from the full selection of interventions available and incorporate them at each stage of the information lifecycle, from collection, to transformation, retention, release, and post-release.

As illustrated in Figure 1, rather than relying on a single intervention such as de-identification or consent, corporate and government actors may consider the suitability of combinations of a wide range of interventions. There is a growing recognition that de-identification alone is not sufficient as a general standard. New procedural, legal, and technical tools for evaluating and mitigating risk, balancing privacy and utility, and providing enhanced transparency, review, and accountability are being explored, and some are beginning to be deployed as part of comprehensive data management plans.

Practical data sharing models can also combine various legal and technical approaches. For instance, a data release may 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 some direct or indirect identifiers, could be instructed to submit an application to a review board. 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, and the uses desired by different users.

As illustrated in Figures 2(a)–(d), the characteristics of long-term data activities, such as the increasing frequency of collection, shift the recommended sets of controls. Changes in one of the big data features, all else equal, correspond to shifts in identifiability and harm, as shown in the following stylized examples involving a smartphone weather app.

In Figure 2(a), the developer of a smartphone weather app decides to collect coarse geolocation data on a daily basis from mobile devices as the app runs in the background, rather than relying on users to self-report their location upon installing the app. This represents a shift from one-time data collection to higher frequency data collection. It is likely to substantially increase the identifiability of the data collected, as

96 See 45 CFR pt 46, sub-pts B, C, and D.

97 See, eg Kashmir Hill, 'How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did' (*Forbes*, 16 February 2012) <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did>.

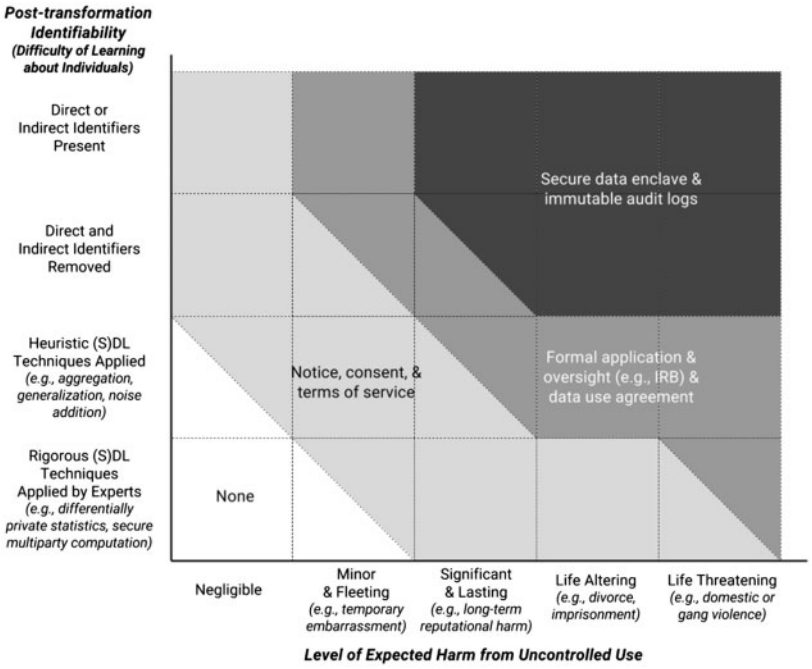

Figure 1. How identifiability and sensitivity guide the recommendations of sets of privacy and security controls. This diagram is reprinted from Micah Altman et al., Towards a Modern Approach to Privacy-Aware Government Data Releases (2015) 30 Berkeley Tech LJ 1967, 2046.

geolocation data collected on a daily basis will likely reveal places of residence or employment that can be used to determine the identity of an individual. A shift from weekly to daily collection is not expected to reveal significantly more sensitive attributes about individuals in most circumstances, resulting in a smaller effect on the level of expected harm. The large increase in identifiability and small increase in the level of expected harm points to the need for stronger controls.

As frequency increases further, eg from daily to hourly samples, identifiability of the data increases. Data collection on an hourly basis is likely to reveal an individual's residence, place of employment, children's school locations, and friends' and family members' homes. If frequency increases to minute-by-minute samples (while keeping the collection period constant), it enables analysis of behaviour that is expressed at finer timescales. Such high-frequency data may enable inferences about subjects' walking and visiting patterns, enabling limited inferences about their fitness, exercise habits, or shopping habits. These factors create new privacy threats and increase the level of expected harm.

Figure 2(b) illustrates a scenario in which the developer begins analysing stored records of users' old geolocation information from their devices. Analysis of older data, absent any increase in period, is associated with a small decrease in identifiability and a moderate increase in sensitivity. Attempting to match a record to an individual, when the record pertains to a geographic location data point from years prior, would in many cases be more difficult than it would be using more recent data. Use of older data is associated with greater risks of harm because the information is more likely to be put to uses that were unanticipated at the time of collection.

Data from the distant past may be less relevant and less likely to cause harm to the data subject, leading to a moderate decrease in identifiability and only a moderate increase in the level of expected harm, as illustrated by the dotted line that reverses direction. As with an increase in the frequency of data collection, an increase in the age of the data being analysed points to the need for stronger controls, unless the data are from the distant past.

In Figure 2(c), the developer makes a decision to collect geolocation data over a longer period of time and to customize its suggestions based on the user's full location history, rather than using only the data collected during the user's current session. A search query history over a period of weeks may contain details related to an individual's location, employment, and interests. This increases the likelihood that a unique pattern of behavior can be found in the data and used to identify an individual. Patterns of behaviour revealed by the data can also reflect sensitive attributes, such as an individual's sexual preferences, medical ailments, or substance

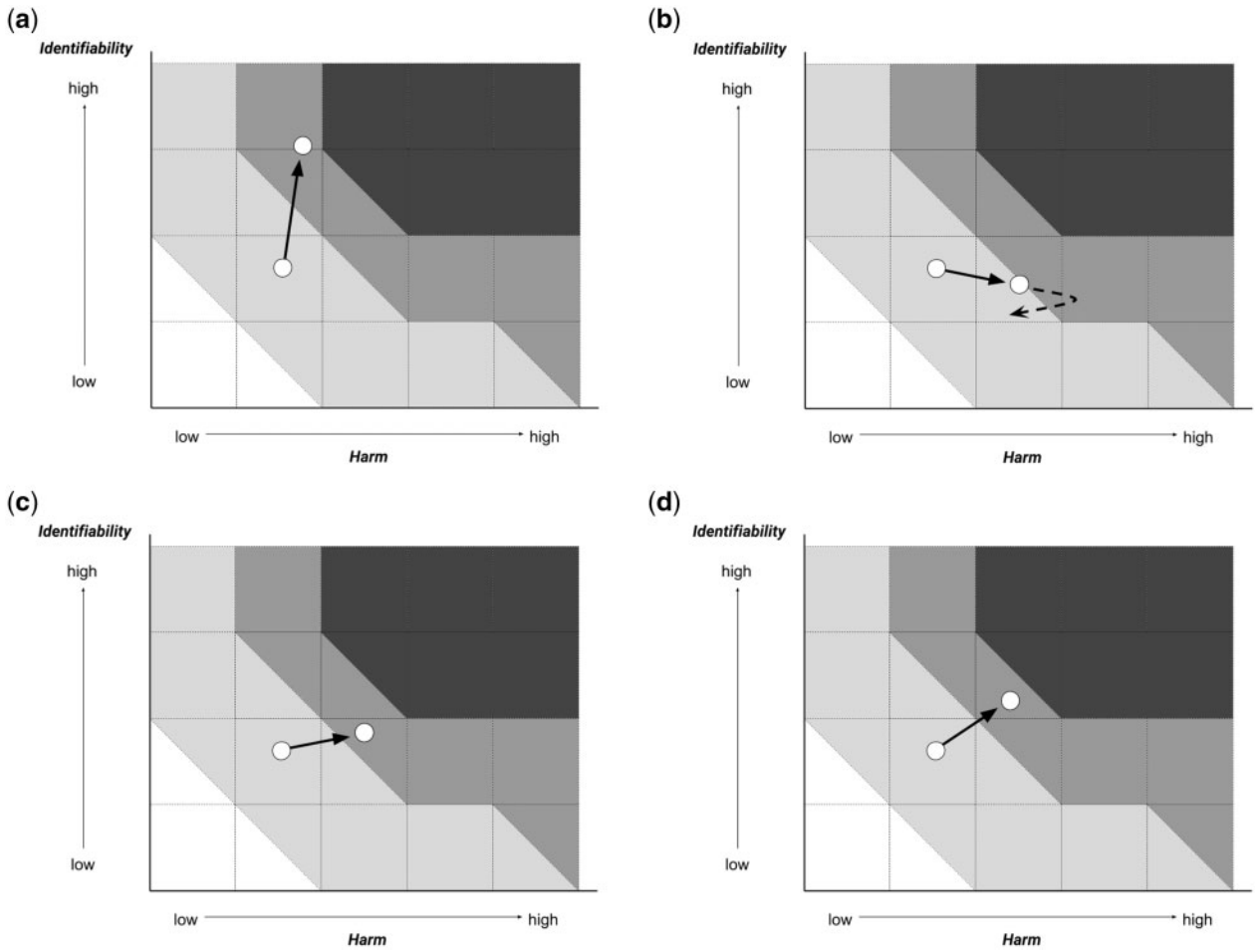

Figure 2. Figures 2(a)–(d). How big data characteristics shift recommendations of privacy and security controls. (a) A smartphone weather app developer makes a decision to collect geolocation data from the mobile device daily, rather than relying on the user's self-reported location. (b) The developer decides to offer users access to stored records of their old geolocation information. (c) The developer decides to collect geolocation data over a longer period of time and to base its suggestions on the user's location history over this period. (d) The developer makes a decision to require users to log in using their Facebook accounts, which grants the app access to the user's Facebook data.

abuse. Therefore, this shift to a longer period of data collection is likely to be associated with a small increase in identifiability and a moderate increase in sensitivity.

Figure 2(d) illustrates the effect of a developer's decision to require users to log in using their Facebook accounts, thereby granting the app access to the user's Facebook data. This shift from low-dimensional data collection to high-dimensional data collection is likely to be associated with moderate increases in both identifiability and sensitivity. Allowing the combination of the information stored by Facebook with the information previously held by the service increases the number of data points that can be used to learn sensitive information about that individual with respect to a number of topics such as political affiliation, sexual preference, and substance use behaviours.

In Figures 2(a)–(d), an increase in expected harm implies that stronger privacy controls should be implemented. As illustrated in Figure 2(a), if notice, consent, terms of service would be considered sufficient for sharing a dataset containing one-time geolocation data, then a dataset containing higher frequency location data would likely require stronger controls, such as formal application and oversight and a data use agreement prior to granting access to the dataset.

Upon determining that a particular risk factor increases, one is naturally tempted to mitigate the factor at play, but this may not be the optimal solution. For example, attempting to decrease the frequency of collection by GPS sensors would reduce the detail and 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, by collecting high-frequency data, but storing only low-frequency samples. However, this can still have significant effects on utility; 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,[98] is computationally costly, and substantially reduces the utility of the data.

Different characteristics of data may combine to create privacy risks, and the effects are often cumulative and superlinear. The risks of combining an increase in the period of data collection and an increase in the dimensionality of the data are likely to be cumulative, both broadening the range of inferences possible from the data and creating new threats. It is also possible that the range of inferences and threats grows rapidly as multiple factors change. When data are collected over longer terms and at higher frequency, it allows detection of behaviours that manifest at many scales of time than would otherwise be possible. Consider, for example, how the privacy risks associated with Figures 2(a) and (c) combine, as illustrated in Figure 3 below. Increasing frequency alone would increase identifiability, and increasing the period would allow for different longer-term behaviours to be inferred. Increasing both enables identification and measures of a range of behaviours that involve multiple time-scales and is thus superadditive.

Beyond frequency and period, other factors may also combine superadditively. For example, if the weather application increases the period of data collection, while the sample size increases to capture a large fraction of the population, it may enable inferences not only about individuals but, through colocation, about group activities. From this information, one could derive colocation of individuals in order to infer, with reasonable reliability, connections between people, and even group action, creating threats from the potential for broad surveillance activities.

Combinations of controls can be used to manage risks in big data over time

Prior research has documented a wide range of available procedural, economic, educational, and legal controls for protecting individual privacy while enabling beneficial uses of data.[99] In broad strokes, such controls can be thought of as affecting computation, inference, and use.

- Controls on *computation* aim to limit the direct operations that can be meaningfully performed on data. Commonly used examples are file-level encryption and interactive analysis systems or model servers. Emerging approaches include secure multiparty computation, functional encryption, homomorphic encryption, and secure public ledgers, eg blockchain technologies.
- Controls on *inference* aim to limit how much can be learned from computations about the constituent components of the database, eg records, individuals, or groups. Examples in common use include redaction and traditional statistical disclosure limitation methods. Increasingly, differentially private mechanisms are being used to provide strong limits on inferences specific to individuals in the data.
- Controls on *use* aim to limit the domain of human activity in which computations and inferences are used. Controls on use are commonly implemented through regulation, contract (eg data use agreements and terms of service), and oversight. Personal data stores and the executable policies they incorporate are emerging technical approaches that aims to control use.

Variants of each type of controls may be applied at different times throughout a multi-stage information lifecycle.[100] Addressing the risk factors associated with long-term data management 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. 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, including a combination of controls, to address new risks and new analytic methodologies.

Examples of clusters of controls for addressing identifiability and sensitivity that can be effectively implemented are summarized below in Table 4 and in the subsequent discussion.

98 See, eg Benjamin CM Fung and others, 'Privacy-Preserving Data Publishing: A Survey of Recent Developments' (2010) 42 ACM Computing Surveys 14.

99 Altman and others (n 42).

100 Ibid.

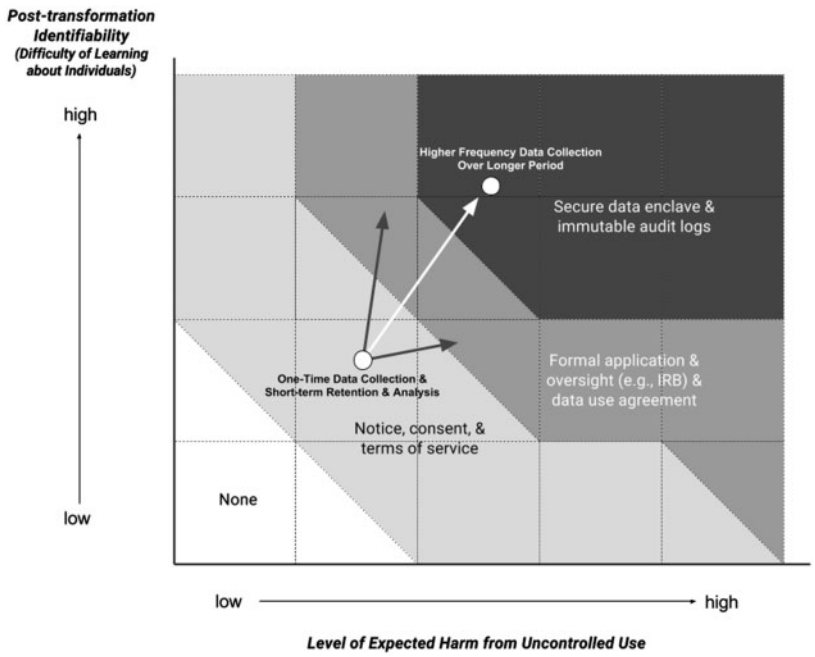

Figure 3. Combining Figures 2(a) and (c), the developer begins collecting geolocation data continuously and over a longer period of time. In combination, there are large increases in both identifiability and sensitivity.

With long-lived large-scale data collections, threats and vulnerabilities continue to evolve, as do the privacy risks posed to individuals and the controls that emerge. There is a need for periodic review and evaluation of these activities, similar to the continuing review process that is required in the human subjects research context.[101] Continual review and adjustment based on newly discovered risks and intended uses has the potential to bring substantial benefits for both privacy and utility.

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

Regardless of the relevant risk drivers, limiting identifiability where it can be accomplished without drastically reducing utility reduces overall risk. For example, although an increase in age does not increase identifiability, the risks that greater 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. A number of controls are available for addressing identifiability risks, including simple redaction approaches, heuristic statistical disclosure limitation techniques, and robust disclosure limitation techniques and tools that provide formal privacy guarantees like differential privacy. While actors most often limit identifiability of data at the stage of publication, identifiability can be controlled at other stages of the information lifecycle,[102] through information minimization or by interactive mechanisms that dynamically add noise to query results.

Note, however, that de-identification using common approaches such as simple redaction of identifiers often does not prevent many types of learning about individuals in a set of data.[103] Some big data risk factors make efficient identifiability limitation especially 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. Also, individuals may be readily identified from their longitudinal spatial trace.[104] Due to the high-dimensionality and sparseness of large-scale datasets containing Netflix users' ratings for each of the films they have watched, traditional approaches to de-identification were shown to fail to provide meaningful privacy protection and to destroy data utility.[105] With respect to high-dimensional data

101 See 45 CFR s 46.1099(e).

102 See Altman and others (n 42).

103 See Arvind Narayanan and Edward W Felten, 'No Silver Bullet: De-identification Still Doesn't Work' (2014), <http://www.randomwalker.info/publications/no-silver-bullet-de-identification.pdf>.

104 See Fung and others, (n 98).

105 Narayanan and Shmatikov (n 28).

Table 4. Examples of feasible privacy and security controls based on the risk drivers and intended mode of analysis identified in a big data use case

	Lower risk: Age, period, sample size, or population diversity	Moderate risk: High dimensional *or* High frequency	High risk: Combined high dimensional and high frequency
Statistical analysis	Notice, consent, terms of service; formal oversight	Differential privacy; formal oversight	Secure data enclaves/model servers; restricted access; formal oversight
Individual analytics	Notice, consent, terms of service; formal oversight	Personal data stores; secure public ledgers; secure multiparty computation; formal oversight	Secure data enclaves/model servers; restricted access; formal oversight

that are less-structured, such as text or video, 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, any names mentioned in it, geolocation information attached to it, and even the unique writing style embedded in the content.[106]

To address identifiability in high-dimensional data, advanced tools such as synthetic data and differential privacy are beginning to be implemented. The Census Bureau has experimented with releasing data using synthetic data models,[107] some of which have been shown to meet a variant of differential privacy.[108] There are several other practical implementations of differential privacy, and off-the-shelf tools that can be applied without specific expertise are beginning to emerge.[109] Tiered access models incorporating a combination of different types of legal, computational, and procedural controls tailored to the risks and intended uses involved can offer more effective risk reduction from learning about individuals. 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.

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

The aforementioned methods for limiting identifiability from simple redaction to differential privacy, have a similar goal: to reduce the ability of a data user to make inferences about individuals, or small groups of individuals. Limiting identifiability is not the only way to limit risk, however. Traditional controls on use include restrictions on who is allowed to make computations on the data. New computational methods, such as secure multiparty computation, secure public ledgers, and executable policies, enable one to limit the computations that can be successfully performed, what can be learned from those computations, and the uses to which the results of those computations can be put, respectively.

The primary way in which computations are currently limited is through the use of secure enclaves with embedded auditing procedures. Federal statistical research data centres operate across the country, and some large research universities run secure enclaves as well, such as the NORC Data Enclave at the University of Chicago. These systems employ strong data security measures, such as those required by FISMA,[110] maintain

106 See Mudit Bhargava, Pulkit Mehndiratta and Krishna Asawa, 'Stylometric Analysis for Authorship Attribution on Twitter' Proceedings of the 2nd International Conference on Big Data Analytics (2013).

107 See Satkartar K Kinney and others, 'Towards Unrestricted Public use Business Microdata: The Synthetic Longitudinal Business Database' (2011) 79 International Statistical Review 362.

108 See ibid.

109 See, eg 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); Prashanth Mohan and others, 'GUPT: Privacy-Preserving Data Analysis Made Easy' Proceedings of 2012 ACM SIGMOD International Conference on Management of Data (2012); Marco Gaboardi and others, 'PSI: A Private data Sharing Interface' Working Paper (2016) <https://arxiv.org/abs/1609.04340> accessed 25 January 2018.

110 See, eg NIST, FIPS Pub 199, Standards for Security Categorization of Federal Information and Information Systems (2004).

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 analytics make it difficult to individually vet results before publication. For longitudinal data analyses, period and age often drive utility, and data destruction would have a high utility cost. Furthermore, although many data management plans rely on data destruction as a technique for protecting privacy, this approach alone should not be considered sufficient for eliminating risk, as deleting data does not mitigate all risks if the data have previously been used or shared.

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 pose 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.[111] Emerging approaches 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 policies do not restrict inference but may be used to restrict domains of use, or classes of authorized users, and enable further auditability.[112] 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.[113] Secure public ledgers,[114] including blockchain technologies, implement tamperproof records of transactions, enabling robust review and auditing procedures. Approaches such as secure multiparty computation techniques, personal data stores, secure public ledger tools,[115] and privacy icons,[116] 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 IRB, in combination with a data use agreement restricting future uses and re-disclosures of the data, as well as data privacy and security requirements, 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,[117] participant-led review boards,[118] and personal data cooperatives,[119] 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.

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 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

111 See, eg Aleecia M McDonald and Lorrie Faith Cranor, 'The Cost of Reading Privacy Policies' (2008) 4 I/S: A Journal of Law and Policy for the Information Society 543.

112 See, eg Lalana Kagal and Joe Pato, 'Preserving Privacy Based on Semantic Policy Tools' (2010) 8 IEEE Security & Privacy 25.

113 See, eg Yves-Alexandre de Montjoye and others, 'On the Trusted Use of Large-Scale Personal Data' (2013) 35 IEEE Data Eng Bull 5.

114 See Jing Chen and Silvio Micali, 'Algorand' Working Paper (2017), <https://arxiv.org/abs/1607.01341> accessed 25 January 2018.

115 See, eg Guy Zyskind, Oz Nathan and Alex "Sandy" Pentland, 'Enigma: Decentralized Computation Platform with Guaranteed Privacy' (2015) <https://www.enigma.co/enigma_full.pdf> accessed 25 January 2018.

116 See, eg Patrick Gage Kelley and others, 'A "Nutrition Label" for Privacy' (2009) 5 Symp on Usable Privacy & Security, art 4.

117 See M Ryan Calo, 'Consumer Subject Review Boards: A Thought Experiment' 66 (2013) Stan L Rev Online 97, 101–02.

118 See Effy Vayena and John Tasioulas, 'Adapting Standards: Ethical Oversight of Participant-Led Health Research' (2013) 10 PLoS Med. e1001402.

119 See Ernst Hafen, Donald Kossmann and Angela Brand, 'Health Data Cooperatives—Citizen Empowerment' (2014) 53 Methods Info Med 82, 84; Effy Vayena and Urs Gasser, 'Between Openness and Privacy in Genomics' (2016) 13 PLoS Med e1001937.

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 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 the privacy literature can offer robust privacy protection for individuals.

Current frameworks and practices for privacy protection in the human subjects research setting have shortcomings as well. However, there are opportunities for commercial actors to leap ahead of current privacy practice in the research setting. 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, the US Census Bureau, Google, Apple, and Uber have begun deploying implementations of differential privacy within tools for collecting or releasing statistics while protecting privacy. 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 for long-term, large-scale data management.

doi:10.1093/idpl/ipx027