THE PLAYBOOK

Data Sharing for Research

DECEMBER 2022

FUTURE OF PRIVACY FORUM

AUTHORED BY

Dr. Sara Jordan

PwC, Senior Manager, AI Governance;
Future of Privacy Forum, former Senior Researcher,
Artificial Intelligence and Ethics

CONTRIBUTORS

Elizabeth Arledge

FPF Consultant and PhD Candidate

Amie Stepanovich

Future of Privacy Forum, Vice President of US Policy

Shea Swauger

Future of Privacy Forum, Senior Researcher for Data Sharing and Ethics

Marjory Blumenthal

Future of Privacy Forum, Senior Fellow

Roy Auh

Future of Privacy Forum, Policy Intern

FUTURE OF
PRIVACY
FORUM

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

EXECUTIVE SUMMARY

The act of data sharing for research appears to be simple: a company or organization holds large volumes of important data about people or the world, and a researcher, who wants to uncover important new knowledge, seeks access to that data in their pursuit. The simplest solution is to share the data.

However, the real dynamic of data sharing is one rife with legal, reputational, or financial risks. These risks apply both to the researcher and the organization from which data is being sought. Some glaring examples are inadequate privacy and cybersecurity protections while the data is in transit or in storage, contractual or internal prohibitions against general transfer of data to third parties, and potential misuse of data by researchers, to name a few. Technical and legal mechanisms, such as formal data sharing agreements, institutional review boards, and access controls, may mitigate some of the more significant risks, but they cannot solve every challenge we are just beginning to understand. The sparse regulatory guidance in the United States provides little assistance for practicing or supporting data sharing for research. Even when considering the risks, data sharing has the potential to improve every sector of society and the benefits are compelling enough to justify thoughtful and informed data sharing programs.

This Playbook, after more deeply analyzing the current landscape of data sharing and its risks, offers recommendations to organizations, research institutions, and researchers for a more effective, manageable, and policy-compliant data sharing program.

FOR ORGANIZATIONS

First, we recommend that businesses increase awareness of how data sharing fits with the organization's values and strategy, the risks and benefits of data sharing, its capacity to absorb workload and costs associated with data sharing, the legal requirements regarding data sharing, and its capability and technology stack for secure processing and transmission of data. Then, implement the current best practices for data sharing that promote privacy and cybersecurity, such as notifying the subjects of personal data, creating data sharing agreements with the key company stakeholders, and employing oversight and accountability mechanisms for ethical use of the shared data. Finally, consider additional practices that streamline data usage, such as refining data fitness and installing personnel, privacy, and cybersecurity controls that are appropriate for the type, value, and sensitivity of the data.

FOR RESEARCH INSTITUTIONS

First, we recommend research institutions build open lines of communications with the data-providing organizations to manage expectations, responsibilities, and necessary changes to data sharing procedures to address data protection concerns. Then, develop a specific process for reviewing and approving proposals involving data sharing and assess whether and how institutional review boards and ethics committees will be involved. Finally, ensure personnel, privacy, and cybersecurity controls are in place that are appropriate for the type, value, and sensitivity of the data.

FOR RESEARCHERS

First, we recommend researchers evaluate whether their current internal policies, resources, and capacity are sufficient to manage an external data sharing partnership. Then, establish a process for communication and collaboration with legal, privacy, and cybersecurity personnel at both their institution and the corporation to enhance privacy, compliance, and cybersecurity measures according to the needs of the partnership. Finally, in that endeavor, consider adopting a zero-trust approach to data access and use.

TABLE OF CONTENTS

PROLOGUE

Why is the Future of Privacy Forum (FPF) writing this Playbook? FPF has addressed issues of research data access and research data sharing for over half of its twelve years. Starting in 2012 with organizing panels and publications addressing Big Data Research[1] and Designing Ethical Review Processes for Big Data Research[2], FPF led thinking about how researchers' use of data generates insights that improve lives while also maintaining high standards of data protection. In 2015, FPF gathered the support of the US National Science Foundation, the Alfred P. Sloan Foundation, and the Washington & Lee School of Law to produce "Beyond IRBs," a conference and proceedings that sparked significant interest and discussion of ethical standards for research data sharing.[3] While some of these same concerns from 2015 were echoed in FPF's latest work on data sharing for research, other thought leadership on protecting privacy in administrative data access[4] and protecting children's data privacy changed[5] practices such that the risk of research uses of data changed. Thought leadership on Corporate Data Sharing with Researchers[6] led to the establishment of the FPF Award for Research Data Stewardship, now in its third year.[7] FPF is proud of its long-standing collaboration with stakeholders to improve researchers' access to data and is excited to publish this playbook as the latest effort in this area.[8]

This playbook started with an idea as old as "The Symposium" or Daoist disputations: host a series of dinners wherein individuals on all sides of an issue can come together over a shared meal to discuss their compatible and competing interests.[9] Between the months of November and December 2021, we (the authors) convened four virtual salon dinners designed to bring together many stakeholders around the themes we (initially) supposed would organize common and competing interests in data sharing for research. The participants invited included both junior and senior academic researchers, legal scholars, chief data officers of companies both massive and small, congressional staff, heads and key personnel of scientific agencies, research archive leads, data protection experts in non-profit organizations, organization leaders for tribal and minority communities, leads of IRBs and clinical trials organizations, and editors of high-profile research journals.[10]

We set the scene in our first convening, explored ethical considerations in our second meeting, discussed the role of corporate actors in our third gathering, and in our fourth and final conclave, discussed the legislative future of data sharing.[11] Owing to the mutations in the corona virus that kept the meetings virtual, the more than 100 people that joined this conversation were never able to physically break bread together or enjoy the warm conviviality of a glass of wine. However, over 100 pages of transcribed notes were captured by appointed note takers and hundreds of messages were exchanged in the chat function of our virtual conference utility. Because we conducted these dinners in the tradition of the Chatham House rule, none of the participants' direct words are used here and no one attending the dinners is directly quoted. Each participant, however, whether they attended or contributed ideas in emails and pre-meeting calls, inspired at least a piece of this text. As with all philosophical dialogues, there are many lessons to be learned from our dinner series. And, that there are many lessons is itself the first lesson: there is not one story of data sharing for research to be told, but very many stories.

We set out to draft a playbook on data sharing for research and discovered that the tangled web of relationships around these activities amounted to more than a contest on a sporting field — it is an epic opera. The story told far and wide about researchers' thwarted access to corporate data is an adversarial page-turner peppered with uncomplicated characters. Tales of bad, profit driven, corporate actors obstructing access to data by beneficent and well-intentioned independent researchers who promise tireless toiling to expand public knowledge to new heights have driven the story all the way to the steps of Congress.[12] Along the way, the pitched battles of "corporate evil" and "research for good" have generated fruitful social and conventional media engagement.[13] But, this simple view of research data sharing reduces the complexity and nuance of the story in ways that limit productive conversation between and about the main actors and their supporting cast and communities.

On the crowded stage of the opera of data sharing for research, the sordid tale of social media platform malfeasance out-shadows the small stories of data sharing success such as the Stanford University Medical School and Empatica partnership in researching physiological signals of COVID-19 infection.[14] In fact, the story is largely one of unseen actors and unseen successes. Corporations large and small today are sharing data with researchers. They are sharing data for single projects or as part of a larger, coordinated, effort seeking to build trusted partnerships with research organizations, government agencies, research data archives, and one another.[15] Many more actors beside the biggest social media platforms and Ivy League universities are part of the full cast of characters that facilitate these interactions. Players on, under, or around the stage include directors of massive research data archives,[16] data editors for high profile research journals,[17] leaders of patient communication platforms,[18] research privacy officers at large state universities,[19] leaders of data integration platforms at health data companies,[20] lawyers in large and small firms,[21] data scientists for agricultural information platforms,[22] researchers from disciplines as varied as computational biology and political science, and officers of international charities,[23] and legislators and regulators from agencies large and small.

A key play for any company seeking to stand up or improve their research data sharing practice is to listen carefully to the real story of data sharing for research that goes beyond the headlines and tweets. Data sharing for research is tangled in competing needs, complex contracts, stern conversations about cybersecurity and data ethics, and more than a little hand-wringing and sleepless nights as all sides strive to manage reputational, financial, and asset-based risks.

THE CAST

The opera of data sharing is performed on a stage crowded with players and props and complex relationships and plot lines. To make sense of these stories and to learn from them, an important first step is to explore the interests, expectations, and norms of the many characters. The characters in this opera may be grouped by organization type (e.g., research institution, company, or platform), size (e.g., large technology companies or small- and medium-sized enterprises), or influence (e.g., "Big Tech" or national science). Researchers seeking data, research institutions that support and encourage them, and companies holding massive amounts of potentially useful data are the main characters. These main characters, however, are supported by a cast of many others who often have reciprocal relationships of their own, and whose interests shape the nuances in the story told. Key among those supporting players are those that ingest, process, cultivate, and curate data, such as (for example) data architects, chief data officers, and data librarians. Just as important are the leaders of research organizations and companies whose influence can spur or spurn a data sharing program.

The players that take the stage to support researchers' use of company data assets enter at different phases of the research process. Legislators, governors, and even presidents set the tone for pressing researchers to pursue lines of inquiry that serve state, national, and global goals.[24] Through their words and actions, these leaders shape the priorities of funding agencies, press universities, and research labs to take on grand challenges, and spur agencies and private business to more aggressively support (or de-fund) specific research paradigms. In a hypothetical future environment of mandated data sharing for research, government agencies with regulatory authority to enforce the terms of the relationships between these players would join the cast of government leaders shaping the story.

Research involving data about Europeans is subject to an extensive legal and regulatory framework centered on the General Data Protection Regulation (GDPR). Discussion in this Playbook has been framed from a U.S. perspective. Legislators are part of a longer list of actors who provide legal authorities, oversight, and accountability for research and data sharing for research.[25] Oversight actors also include peer-reviewers, journal editors, and readers. Accountability actions taken by these players include multiple forms of review and feedback that occur throughout the data sharing story.

Leaders at research institutions, from university presidents and faculty senators to deans and department heads, shape the priorities for many researchers to pursue and incentives to use external data resources. Researchers also are influenced by their peers and disciplinary visionaries whose theories and hypotheses beg for testing with novel data at scale.[26] Funding organizations both shape and respond to these pressures by creating new funding mechanisms, changing the terms of fulfillment for existing mechanisms, and encouraging novel uses of data or use of new publication outlets.[27]

Researchers build data assets from shared resources and, depending on the terms of the data sharing agreements negotiated between legal counsels working for companies and research institutions, curate those into datasets retained by archives or repositories and data libraries.[28] Some of those datasets become assets for further use by other researchers who find them attached to published journal articles, preprint articles, or in the digital appendices to books.[29] The data assets published with journals are checked (in some but not all cases) by data editors, peer reviewers, and research integrity sleuths.[30] Students and other consumers of research, such as journalists, may use summaries and extracts of the shared data assets to further the knowledge that becomes common sense or startling conclusions of extensive investigative reporting. Each of these players has a role to play in generating the benefits that the public and the greater research community might reap from data sharing for research.

COMPANIES AND ORGANIZATIONS

Data sharing is one of the core business data processing functions that private companies do as part of the normal conduct of business.[31] For example, companies share data with one another to contribute to product development, sometimes for mutual benefit such as through sharing of data for market research. Separately, companies may share data with government agencies, including to build important economic indicator values such as the consumer price index or measures of inflation. Companies also share data to build collaborative relationships with partners that provide complementary lines of service. Data sharing activities such as those in these examples are not only the acts of large firms or those with significant market value or reputation. Instead, newly incubating businesses, start-ups, small or medium sized enterprises, and massive multinational firms all share data. And, of course, these companies also sell data to one another for business purposes and, in some cases, sell data to researchers.

The idea of data sharing for research is not new. Many companies around the world, in fact, already share data with researchers. Some of these companies joined our dinner series on this topic, including Mercy Corps AgriFin,[32] Lioness,[33] and Datavant[34]. These companies sharing data are often aided by research data repositories, like the ICPSR[35], CESSDA[36], and FigShare[37], who also joined our dinner series. Google was presented with one of the Future of Privacy Forum's Awards for Research Data Stewardship for sharing COVID-19 Mobility Reports with several universities, as was digital biomarker Empatica, which shared data from smartwatches and related wearables with Stanford Medical researchers to determine if COVID-19 could be detected early.[38]

RESEARCHERS

Researchers, as used in this playbook, are individuals with specialized training in theories, research questions, research methods, and the history of a field that equips them with knowledge appropriate to answer questions that are both theoretical and applied. Appropriateness of theoretical and methodological training can be measured by referring to norms of the fields, as well as by examining the standards issued by organizations that accredit, certify, or publish on behalf of a field. What constitutes a research field can also be estimated by reviewing lists of fields recognized as fundable by national or international research funding bodies.[39] Examples of such organizations include the professional associations of specific research fields, such as the American Medical Association (AMA), American Statistical Association (ASA), Accreditation Board for Engineering and Technology (ABET), and many others.

The professional organization of a field sets the vocabulary and definitive methods that can help differentiate lay or citizen research from professional research. While citizen scientists can and do perform critical scientific tasks — collecting data, interpreting research results — their most common role is to fulfill tasks within the profession of research (although they sometimes form an alternative community of researchers). As an example, the U.S. Environmental Protection Agency (EPA) uses citizen scientists in water monitoring programs to spot dangerous bacterial growths and evaluate the ecological health of wetlands.[40] Its citizen science programs put the tools of research data gathering in the hands of citizens to broaden the quantity of data acquisition efforts.

Journalists, non-profit advocacy organizations, and government agencies also ask private sector companies for access to their data assets to perform investigations and studies. But, unlike the research data needs that journalists, specific public interest bodies, and government agencies have, researchers use data to build a common pool resource known as *generalizable knowledge*.[41] While excellence in investigative journalism and non-profit reporting answers fascinating questions about specific events or persons that provide an evidence base for testing or inducing new theories and frameworks, these studies do not themselves produce new, generally applicable, theories. Independent, scientifically minded, research, however, does produce generalizable knowledge.

Two characteristics make research lead to generalizable knowledge:

- When the research speaks to theory development and higher-level abstraction as a goal.
- When it provides a common-pool good — scientific knowledge — that all future researchers and even lay persons could potentially use. Common pool goods like scientific knowledge can be enjoyed by all without the good itself being diminished by any other actors' enjoyments.

When considered alongside other organizations' use of corporate data, researchers are also considered special because of their scientific training and posture of scientific neutrality and independence.[42] While journalists, think-tanks, and other nonprofits might be financially independent of their data sources, they are dependent on their advertisers, editorial teams, boards, and even stakeholder contributors. Researchers, on the other hand, are argued to be accountable to public democratic bodies, such as state legislatures and national research funding bodies, but most importantly they are also accountable to science, scientific norms, and to the accumulation of knowledge in their field. Research institutions, as discussed below, feature mechanisms to foster accountability.

RESEARCH INSTITUTIONS

While research institutions share the concerns about independence, reputation, and integrity that researchers themselves do, it is important to recognize that research institutions, such as universities, are subject to a range of economic incentives.[43] Regardless of their status as non-profit corporations, universities, think-tanks, and research institutes compete for a limited supply of funding, talent, data, and attention. Their espoused mission and vision statements may extol their public mindedness and focus on independent research, but research institutions are not free of agendas, endowments, incentives, and motivations that are specific to a narrow slice of the public.[44]

Research institutions provide support, backing, and resources for researchers. In addition to the mechanisms put in place by units like IRBs, Human Resources, Faculty Affairs, and the Faculty Senate, universities also support researchers with resources needed to engage with legal and technical barriers to using shared data. For example, while researchers may be a signatory on a data sharing agreement or memorandum of understanding or agreement (MOU or MOA), their university is also likely to be a signatory and may even be the "responsible official" to whom blame falls in the event of data loss or data breach.[45] Further, a university can incentivize its data librarians or data repository to commit resources to facilitating the appropriate handling mechanisms for data shared by an organization subject to use restrictions. Where a university perceives corporate data to be a strongly valuable part of the data used, they can commit technical resources, like software and servers, to create the access control mechanisms through which their researchers can more safely interface with companies or other universities.

PREPARING FOR ACTION

Recommended Actions for Researchers:
Collaborate with institutional cyber and physical security specialists to build a data and software securitization and management plan for the shared data aNd relevant analytical software used.

LEGISLATORS AND OVERSIGHT ACTORS

Oversight and accountability for data sharing for research is not only provided by agreements between companies and research institutions but also through legislation, including legislative mandates.[46] Corporate leaders, data officers, and legal staff all work to ensure both technical and operational safeguards are in place to meet contractual obligations and protect consumer data. Research institutions and researchers will have contractual obligations as well as commitments to university governance systems and other key stakeholders to ensure the integrity of the institution. Each of these players will need to ensure appropriate oversight and accountability mechanisms to meet these requirements.

In addition, legislative bodies at both the federal and state level are considering legislation to protect citizen data during data sharing and/or mandate corporate data sharing with researchers. Legislation may establish oversight mechanisms and accountability requirements, which may include assigning a government agency as a central regulatory authority. Even in the simplest of agreements or legislation, additional players will become involved, and mechanisms will be put into place.

OTHER ACTORS

Individuals

While not a direct actor in the data sharing relationship, individuals or groups of individuals are often the subject and the object of the research being conducted. Predicting or analyzing individual behavior may be one purpose for research that requires data access at an individual level. To study why individuals share news articles with sensational or "clickbait" headlines, researchers will need to know something about the individuals.[47] Although all individual data shared should be de-identified and protected with the highest applicable levels of privacy engineering techniques, the individual is an inescapably important part of research.

Unfortunately, individuals — the lay public — often have an incomplete understanding of how their data is collected, shared, or used. On one end of the spectrum, there is considerable fear that corporations' use of data allows them to "know everything about us" while at the other end, there is a belief that corporate knowledge of customers' true values or desires is frighteningly low.[48] Individuals' perceptions of the reasons that companies use data are also similarly confusing: people believe that companies use their data solely for the benefit of the company but also for their benefit as product users. As recent research on attitudes towards their data uses show, "...when data is used to improve a product or service, users generally feel the enhancement itself is a fair trade for their data. But they expect more value in return for data used to target marketing, and the most value for data that will be sold to third parties. In other words, the value people place on their data rises as its sensitivity and breadth increase from basic information that is voluntarily shared to detailed information about the person that the firm derives through analytics, and as its uses go from principally benefiting the individual (in the form of product improvements) to principally benefiting the firm (in the form of revenues from selling data)."[49]

Although the average person may be confused or concerned by companies' uses of their data, they are often also bewildered by researchers' use of data, particularly data shared to them as third parties by other organizations, such as hospitals or universities.[50] Perceptions of the uses and benefits of personal data by researchers also spans a wide range. At one end, there is a belief that science and scientists advance the public interest and that analysis of research data will lead to findings that will save even the most catastrophically ill patients or solve the most difficult environmental problems.[51] At the other end is a public disdain for the pace, nuance, and stilted communication of scientific research and a belief in the impractical, "ivory tower" and otherworldly nature of research.[52] Spanning the spectrum is a low level of statistical acumen and scientific illiteracy that prevents most laypersons from competently reading and understanding an entry level (social) scientific textbook or journal article.[53] Furthermore, individuals may have concerns about who will profit from research using their data.

Data Intermediaries

Recent changes in the law in the European Union (EU) established a new type of data organization — the data intermediary — which is a "a catch-all term for those who help broker the flow of data from data source to data user who otherwise could be described as middlemen, data aggregators, data brokers, etcetera".[54] These new organizations could, as described by the Center for Data Ethics and Innovation, "provide technical infrastructure and expertise to support interoperability between datasets, or act as a mediator negotiating sharing arrangements between parties looking to share, access, or pool data. They can also provide rights-preserving services — for example, by acting as a data custodian allowing remote analysis through privacy-enhancing technologies (PETs) or providing independent analytical services in a siloed environment. Data intermediaries could assume the roles and obligations of a data controller and/or processor, depending on the circumstances". One instance under development is the European Digital Media Observatory (EDMO), which is developing multiple communal resources for researchers (including platform-data access with privacy and security protections) with independent governance.[55]

The true effect that data intermediaries will have on facilitating data sharing for research is unknown, but data trusts, data collaboratives, and data archives have shown that second and third sector initiatives have spurred access to data in innovative ways in the past. Learning lessons from these existing data intermediaries and keeping an ear to the ground for new movements will be key to leveraging these organizations in the future. For example, EDMO has proposed a Code of Conduct leveraging EU data-protection mechanisms to guide research using platform data.[56] Data intermediaries might also develop shareable insights into possible avenues for reconciling researcher requests to see microdata and the privacy protection available through PETs.

Vendors

For some companies, data sharing is a component of the business. For others, facilitating data sharing is the business. One way to measure the importance of data sharing as a key business function is to examine the growth in the industries providing security of data transfer or analysis, such as through trusted execution environments, zero-knowledge and zero-trust environments, and secure multi-party computation. Some of the biggest technology companies, such as IBM, Microsoft, and Google, have software and/or hardware-based methods for secure data sharing or confidential cloud computing that allow multiple parties to use the same data for analysis without sharing it.[57]

Trusted execution environments (TEE) and other utilities that allow organizations to share access to data can facilitate research with shared data. Improving access while reducing risk through transfer and off-site storage is a promising avenue for more secure data sharing. Allowing access without explicitly transferring data will also reduce overall data storage costs for data users. However, these utilities may be cost prohibitive for some researchers to access. Establishing cost sharing arrangements or public-interest research pricing structures may forge a path forward toward innovation in data sharing for research.

Publishers

For any given data sharing collaboration, the relationship changes once the data is analyzed and is out for publication; the researcher is no longer the only entity who has a claim to the shared data. The research publication process involves multiple individuals and organizations who can lay claim to a data resource. This may include research repositories at the institutional level, or the repository of data at a journal, or even claims from a data journal itself. As research goes from analysis and discovery to publication, journal editors and peer reviewers take on a special role with respect to a data resource. Journal editors can request raw data files for replication analysis. Peer reviewers can also request data files to validate claims made in text. When companies share research data with academics, they should work closely with their academic partners to identify publication venues of choice to determine whether or not a data resource must be shared with those publication outlets. Once the data is available for review for a publication, the question becomes less about the relationship between companies and researchers, and more between companies and the research data enterprise itself. Once material is submitted to a journal, it becomes the province of research as a profession and general knowledge building as such.

Funders

Research funders play a powerful role in terms of shaping the career trajectory of researchers and the reputation and capacity of research institutions. They not only evaluate but encourage quality research, and they can encourage or discourage use of shared data resources. Organizations that fund research support capacity building for using any form of shared data through the usual mechanisms of supporting institutional on-costs. Without the administrative costs associated with large-scale grants, university offices, such as research administration offices and legal counsel, would not be able to operate robustly. Without support from grants universities may not have the capacity to house data and to build the necessary infrastructure around that data. Research funders can also incentivize the creation of shared data assets by explicitly funding data development. Just as data journals will provide specific and tailored opportunities for faculty to publish data as a published product, funding data creation can also create opportunities for faculty to achieve the goals of funded research.

Research funders priorities can help to foster a culture of open and efficacious research data sharing. Research funders' requirements set the terms for performance of research related tasks, from training requirements (e.g., RCR training), publication of abstracts or findings in specific venues, dissemination of data to required platforms (e.g., clinicaltrials.gov), to management of data throughout the lifecycle of research (e.g., NIH data sharing and management plans). Research funders also play a role in studying the barriers and benefits to research data sharing. The process of working through the implementation of mid-2022 White House guidance to federal researchers, discussed below, will help clarify paths forward.

ACT I: SETTING THE SCENE

Data sharing is commanding the attention of companies, researchers, and legislators now, but it is not new. The promises and pitfalls of data sharing hinge on the degree to which data sharing and data receiving organizations feel that the other is fulfilling the terms of their relationship. Recent stories of failed data sharing relationships describe situations where breakdowns in communication between parties, exploration of novel methods to satisfy data acquisition under loose contracts, or changes in the perceived risk of data sharing by one party fractured the sharing relationship.[58]

There are as many denotations to "data sharing" as there are connotations of the phrase. For database administrators, data sharing (a.k.a. data conferencing) is "[t]he ability to share the same data resource with multiple applications or users."[59] For clinical researchers and funders of clinical research, data sharing involves many actions, ranging from publication of summary data in a clinical study report, publication of data in a research journal, and/or uploading of data to a repository like clinicaltrials.gov.[60] Since there is no definition of data sharing available that covers all of its dimensions in technical fields or specific research areas, stipulating such a definition is an important task of this playbook and for companies or research institutions striving to create the infrastructure to support research data use.

In the broadest strokes, data sharing relevant to research is the transfer of data assets to third parties, with or without intervention of an intermediary such as a cloud provider or data trust. Data sharing connects distinct organizations, where one party transfers data to another so that both parties can learn more about the contents of that data, answer common questions, or exploit shared market opportunities. Finally, data sharing can occur between individuals and organizations when individuals grant unique levels of access to their data in return for access to knowledge or promises of future services. What differentiates data sharing from other forms of data transfer is that data sharing is not predicated on an exchange of data for compensation but presumes establishment of a relationship that will lead to benefits for data sharers and data receivers.[61]

Figure 1: Terms and Definitions

Terms	Definition
Data Sharing	The non-compensated exchange of data for services, such as access, analysis, or insight
Data Sharing for research	Used when data is produced (collected, curated, stored) by organizations that do not have research as their primary economic or social function and given to other organizations who do have a research function

Source: Created by Future of Privacy Forum, 2022

The concept of "data sharing for research" spotlights data that is produced (collected, curated, stored) by organizations that do not have research as their primary economic or social function who then give that data to other organizations who do have a research function. For example, when a health insurance company shares transaction data with health economics experts performing health economics research as part of their research roles at a university, the phrase "data sharing for research" applies.

Defining and Differentiating Research

Research is a term with many definitions and interpretations. In the conversations that led to this playbook, speakers often pointed out that the lack of a clear definition of research that differentiates research from other forms of data use hinders development of a coordinated approach to data sharing. The extensive discussion of research in the massive collection of texts from philosophy of science, research ethics, research methods and individual research disciplines, is beyond what could be synthesized here.[62] Companies striving to define the research they could support with their data assets should examine the discussion of research in the scientific and social scientific disciplines that build knowledge essential for their products and processes. Other accessible places to find a general definition of research include legal and regulatory sources, for example the definition of "research" in the Common Rule (45 CFR 46),[63] the definition of "basic research" from the U.S. Department of Defense (DoD 7000-R),[64] or the definitions of "basic research," "applied research," and "experimentation" as defined in the "Frascati Manual" published by the Organization for Economic Cooperation and Development (OECD).[65]

Research is "systematic investigations, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge and which conforms to methods of investigation of a recognized discipline or subdiscipline. Research designed to contribute to generalizable knowledge supports development or modification of theories and general abstractions, such as models of social or natural phenomena, that can be used by other researchers or non-researchers."[66]

The definition of basic research reflects similar tenets: "Basic research is systematic study directed toward greater knowledge or understanding of the fundamental aspects of phenomena and of observable facts without specific applications towards processes or products in mind. It includes all scientific study and experimentation directed toward increasing fundamental knowledge and understanding in those fields of the physical, engineering, environmental, and life sciences related to long-term national security needs. It is farsighted high payoff research that provides the basis for technological progress."[67]

Contributions to generalizable knowledge is not all there is to research. Applied or practical research uses the techniques, tools, and methods of generalizable research to address more specific problems faced by specific audiences. Applied research has similarities with "development" in descriptions of "Research & Development (R&D)."

Breakout 1: Frascati Manual Definition of Basic and Applied Research

The term R&D covers three types of activity: basic research, applied research, and experimental development. Basic research is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use in view. Applied research is original investigation undertaken in order to acquire new knowledge. It is, however, directed primarily towards a specific, practical aim or objective. Experimental development is systematic work, drawing on knowledge gained from research and practical experience and producing additional knowledge, which is directed to producing new products or processes or to improving existing products or processes."[68]

Applied research consists of methodologically sound inquiries designed to answer and confirm specific questions for a specific purpose or community of interest. Applied research is often explicitly divorced from the research task of theory building or creating high level abstractions for use by any other disciplines or field of study. There are few regulatory definitions of applied research, but the U.S. Department of Defense defines it as "Applied Research is the systematic study to understand the means to meet a recognized and specific need. It is a systematic expansion and application of knowledge to develop useful materials, devices, and systems or methods. It may be oriented, ultimately, toward the design, development, and improvement of prototypes and new processes to meet general mission area requirements."[69]

To be called research, whether basic or applied, the investigation must use methods and have a basis in the theories that define a field. As has been revealed throughout the coronavirus pandemic, with individuals crowing about "doing their own research", many lay persons have come to believe that "doing research" means a-systematically watching videos or listening to talks from experts (both self-professed or peer-vetted), or triangulating between social media and conventional media (news) sources looking for "evidence" to support favored views.[70] By contrast, as described above, professional researchers share a professional orientation toward technical and ethical norms surrounding the processes of research in their area and create an argument for the validity and contribution of their research by referring to theories and methods recognized by fellow researchers as suitable to answer questions in that field.

Research Data

Research data is information and material that is used by researchers for research purposes. The effort to expand researchers' access to company data assumes that all data could, in theory at least, be relevant to answer research questions. Research data could be data designed for research use or data that is wrangled and reconstructed and then used to test hypotheses that contribute to the knowledge base of a research field. Research data can include data about people, including personally identifiable information. Where researchers need data that describes personal behaviors, they are supposed to take many ethical and statistical steps to ensure that the results are presented at a sufficiently high level of abstraction that the readers do not learn about a specific person but learn about general traits of people.[71] But, in many research fields, the data needed to carry out important investigations that shape our virtual and physical lives have no personally identifying information in them. For example, significant research on cybersecurity, such as how data breaches are perpetrated, often does not require personally identifiable information to have an effect on our virtual lives.[72]

ACT II: BUILDING A NARRATIVE

Epic stories rarely have clear-cut "good" and "bad" characters. Similarly, it is hard to find any quantifiable metric to qualify "good" or "beneficial" research, though it is inarguable that there have been benefits to the undertaking of research across time, both to society at large as well as to smaller groups and communities. Despite this, it is impossible to separate any future research from the risks that it may create, both in terms of privacy risks to individuals and communities as well as risks related to undermining widely understood ethical norms and standards. While there may exist some strategies to mitigate those risks, those strategies may not also be appropriate or conducive to the full scope of the risk presented. It is necessary to understand both the propensity of research to create benefits as well as the scope and scale of underlying risk to fully prepare to create programs to provide or receive data for research.

Benefits of Data Sharing for Research

A hypothesis driving the story told about research data sharing is that a modern company's vast data holdings are an untapped resource for answering the complex problems of science at a scale traditionally inaccessible to researchers.[73] Research questions argued to be answerable using the troves of corporate data include those in myriad areas of social and economic life, engineering, health, education, finance, and security.[74]

But what is the potential for research outputs to return benefits? The academic research literature is replete with problems.[75] Some of those problems are economic, such as paywalls for journals and for journals and for books, and some of those challenges are endemic to the research writing profession, such as dense, jargonistic, theoretically rich, but practically inconclusive, expositions and on narrow topics. How are cost-conscious corporations to pay for the opportunity for creating research outputs that do not explicitly and assuredly improve their lines of business? While benefits from research may be attenuated, conversations revealed examples of research benefiting actors across the ecosystem.

Benefits to Society

The story of data sharing for research is a multi-act opera and like other operas it does come with a moral. The moral of the data sharing story is that the public — those who contribute the data, pay for research projects (directly or indirectly), and read the products of researcher's analyses — is better off, and believes they are better off, when research is advanced because data is shared.[76] The public includes all the people, past and present, whose choices generate the data used by both corporations and researchers. Public betterment comes from companies sharing data with researchers who leverage their skills and position to build knowledge from which everyone could learn. The benefits supposed to accrue to the public are new and better knowledge. While advocates often couch their appeals for researchers' access to company data in the language of building new knowledge, there may be a stronger case to be made for using company data to improve the evidence base for existing knowledge.

A company's vast data holdings are, in principle, an untapped resource for answering the complex problems of science at a scale inaccessible to researchers.[77] Research questions argued to be answerable using the troves of corporate data include those in myriad areas of social and economic life, engineering, health, education, finance, and security.[78] The answers to those questions are supposed to provide tangible benefits to the public, whether in segments like disease sufferers and their families,[79] or as a whole like all users of cybersecurity protections or all members of the human genetic species.[80] For example, research using data from tech companies allowed multiple researchers around the globe and in various government and journalistic organizations to study public travel behavior during the early days of the COVID-19 pandemic.[81] Such data contributed to (at the time of writing) over 400 academic research publications indexed in the US National Library of Medicine PubMed Central Database.[82] Researchers are also using data shared from private companies to examine if there may be early biometric indicators of disease, including infections of COVID-19.[83]

Within the previous 40 years, researchers with expanded data access, whether from use of corporate data or through multi-institutional and even global collaborations, have been able to reexamine previous research findings and learned that results may vary when those results are re-tested with much larger amounts of data, much finer resolution data, or data that is more representative of a population of people.[84] Areas where replication of studies using company data has exposed previous research shortfalls include psychology,[85] neuroscience,[86] health behavior,[87] and political behavior research (to name a few).[88]

Refinement of research conclusions allows for groups to be treated with more culturally appropriate and efficacious interventions and for students to be met "where they are" rather than where standardized metrics suggest they should be. With specific respect to the uses of corporate data to facilitate more tractable research results, corporate data access allows for research to move toward truly generalizable conclusions because hypotheses are tested and inferences drawn from a more diverse population, measured with greater frequency, and with less possibility of researchers' interventional biases, than could be done through conventional research channels.[89]

Independent, scientifically minded, research using large datasets can be used to address some of society's most complex problems both in generating new knowledge as well as in refining previous conclusions.

Benefits to Research Institutions and Researchers

For researchers, access to corporate data, when done properly, can lower costs and barriers to engaging in many types of research and open new opportunities for testing hypotheses on a large scale. Researchers' work product and professional reputations are directly burnished when granted access to company data. In addition, universities and colleges, research funders, and students stand to reap direct and secondary benefits.[90] Research institutions, whether they are for-profit or non-profit organizations, are large-scale businesses that compete with one another for reputational benefits, financial benefits, and talent.[91] When research institutions partner with companies it increases perception of their technological sophistication, business acumen, networking and political strength, and ability to forge connections with the future state of the world. As research institutions are increasingly asked by legislative bodies to engage in applied research that affects the communities in which they are situated, they must seek new channels for data and real-world evidence that allow their research products to outcompete that of other research institutions and even outcompete the findings of policy analysts, journalists, and other civil society organizations.

Benefits to Corporations and Organizations

Research institutions' use of corporate data also benefits the companies whose data they use.[92] This can occur by design. Corporations bear part of the costs of data sharing for research and may reasonably expect that researchers provide them something in return. Spending to stand up a data sharing program is a form of corporate philanthropy, but it is also a form of corporate strategy.[93] While paying to share data for research is, in one sense, paying for an uncertain positive return, research is also a boon for new product ideation, innovation, and identification of talent or markets.

Despite their size and the influence of their software and related products on our lives, corporations are neither omniscient nor omnipotent. Small companies and large corporations must gather ideas and insights from outside persons, such as researchers. Companies have a strong interest in sharing research data and opportunities with researchers to generate new ideas, pull novel insights from untapped data, and to identify new talent. This allows them to more keenly solve the applied problems that customers will ultimately pay for and to reap reputational benefits, such as positive public perception, competitive advantage with their stakeholders and shareholders, and credibility as purveyors of next-generation products. Companies also may wish to share research data to push ideas into the research world for testing by thinkers without having to overcome readers' worries about corporate influence.

Benefits to Other Actors

While research institutions and data sharing businesses stand to benefit from research data sharing, legislators, journalists, civil society, and interest groups also benefit from researchers' access to company data.

Legislators benefit because researchers can conduct research that compliments legislators' oversight functions. Research data sharing is enjoying a renewed moment of resurgence because legislators and leaders do not trust answers from companies about how much of the world their software "eats."[94] Since legislators and community leaders cannot answer this question themselves they must rely on insights and guidance from academic researchers (and sometimes lay-researchers, like journalists), to build a more reliable picture of how much companies consume our attention, social behavior, and environmental resources. Companies seeking to share data should understand that their data sharing activities may inform projects that extend either form of benefit described above.

With specific reference to social media data used by researchers, legislators may be able to use the results of that research to more readily answer the questions such as, "why did I win?" or "why did I lose an election?" by understanding the dynamics of political advertisement and consequences of political speech at scale.[95] Likewise, legislators can more adequately understand why issues are arising amongst their constituents at a scale and pace that exceeds that of ordinary political operations.[96] Journalists have an interest in researchers' access to company data. First, journalists have an interest in researchers' access to data because research findings may make for interesting journalistic content. Second, journalists may also be interested in researchers' access to data so that they may also be able to access data beyond the reach of mechanisms like Freedom of Information Act (FOIA) requests. Third, journalists also have an interest in researcher pursuit of company data to better understand their advertisers and supporters or detractors. Civil Society groups, such as nonprofit organizations, have an interest in research access to data because the applied research conducted may inform their positions or help them educate their constituents. Likewise, Civil Society organizations such as think tanks have an interest in research access to data to understand issues and produce competitive content. Lastly, sharing data with researchers can help in part to identify, measure, and provide interventions for pressing social issues such as foreign influence on elections, the negative impacts of social media on adolescents, or the prevalence of hate speech on platforms.

Risks

As with all other compelling stories, the story of data sharing for research includes intrigue, confusion, disappointment, fear, and risk. Just as the benefits that each player gains from a data sharing relationship can be specific to their role, so are the risks each must mitigate. However, some risks may appear as themes across different roles. For example, as discussed above, research institutions share many of the same drivers as private corporations, and, as such, many of the concerns that private corporations have about managing real costs and reputational risks for data sharing programs have analogs in the setting of research institutions.

Figure 3: Best Practices for Reputational Risk Management for Universities[97]

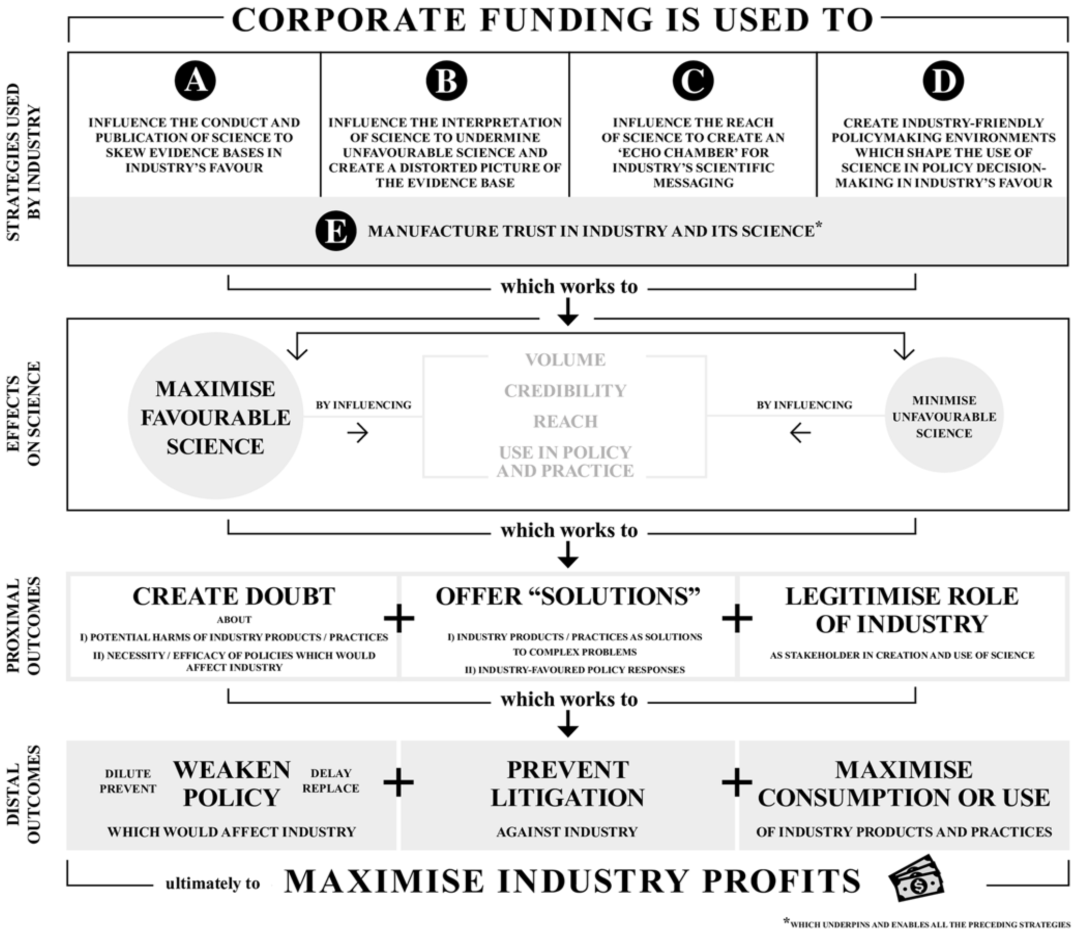
BEST PRACTICES FOR MANAGING RISK
Understand the institution's current reputation through social media mentions and rankings in guidebooks.
Assess the culture of the institution and make clear the mission and values of the institution.
Assign ownership of the institution's reputational risks and create specific lines of communication between leaders.
Consider the institution's programs, people, or areas that are highly esteemed and perceived to be above the rules.
Obtain a collection of all risks and understand how non-reputational risks may affect reputational risks.
Communicate the risk portfolio and mitigation plans with the Board on a regular basis, conveying the resilience of the institution should a risk event occur.
Create and maintain a risk monitoring system to proactively identify potential risk events.

Reputation

Researchers are not unalloyed beneficiaries of corporate data sharing programs. In a profession where independence and rejection of a status quo are prized, researchers bear risks from relationships with corporations such as perceived (or actual) conflict of interest, loss of opportunities from other funding sources, and broken trust if shared data is revealed to be incomplete or inaccurate.[98] For some research fields, there is a strong perception that association with a corporate entity — even if only for the purposes of data use — is "selling out" or "shilling" or even foregoing research norms for the sake of corporate money.[99] And, in other fields, uses of data gathered by others, whether that comes from research repositories or from corporate holdings, is described as being "parasitical" or even "cheating".[100]

Companies must accept risks to their reputation when sharing data as well. If consumers, legislators, partners such as vendors, and competitor firms disagree with data sharing or the conclusions of the research, companies risk reputational and financial damage.[101] Companies who share research data that turns out to have statistical errors or is poorly documented also risk losing the trust of future research partners.[102]

Figure 4: The Science for Profit Model — Corporate influence on science and the use of science in policy and practice[103]


CORPORATE FUNDING IS USED TO
STRATEGIES USED BY INDUSTRY
A
INFLUENCE THE CONDUCT AND PUBLICATION OF SCIENCE TO SKEW EVIDENCE BASES IN INDUSTRY'S FAVOUR
B
INFLUENCE THE INTERPRETATION OF SCIENCE TO UNDERMINE UNFAVOURABLE SCIENCE AND CREATE A DISTORTED PICTURE OF THE EVIDENCE BASE
C
INFLUENCE THE REACH OF SCIENCE TO CREATE AN 'ECHO CHAMBER' FOR INDUSTRY'S SCIENTIFIC MESSAGING
D
CREATE INDUSTRY-FRIENDLY POLICYMAKING ENVIRONMENTS WHICH SHAPE THE USE OF SCIENCE IN POLICY DECISION-MAKING IN INDUSTRY'S FAVOUR
E
MANUFACTURE TRUST IN INDUSTRY AND ITS SCIENCE*
which works to
EFFECTS ON SCIENCE
MAXIMISE FAVOURABLE SCIENCE
BY INFLUENCING
VOLUME
CREDIBILITY
REACH
USE IN POLICY AND PRACTICE
BY INFLUENCING
MINIMISE UNFAVOURABLE SCIENCE
which works to
PROXIMAL OUTCOMES
CREATE DOUBT
ABOUT
I) POTENTIAL HARMS OF INDUSTRY PRODUCTS / PRACTICES
II) NECESSITY / EFFICACY OF POLICIES WHICH WOULD AFFECT INDUSTRY
+
OFFER "SOLUTIONS"
I) INDUSTRY PRODUCTS / PRACTICES AS SOLUTIONS TO COMPLEX PROBLEMS
II) INDUSTRY-FAVOURED POLICY RESPONSES
+
LEGITIMISE ROLE OF INDUSTRY
AS STAKEHOLDER IN CREATION AND USE OF SCIENCE
which works to
DISTAL OUTCOMES
DILUTE
PREVENT
WEAKEN POLICY
DELAY
REPLACE
WHICH WOULD AFFECT INDUSTRY
+
PREVENT LITIGATION
AGAINST INDUSTRY
+
MAXIMISE CONSUMPTION OR USE
OF INDUSTRY PRODUCTS AND PRACTICES
ultimately to MAXIMISE INDUSTRY PROFITS
*WHICH UNDERPINS AND ENABLES ALL THE PRECEDING STRATEGIES

User Expectations

An undeniable challenge to sharing corporate data with researchers is doing so within the limits of the expectations and appetite for data risks internalized by the people to whom the data pertains. What are those expectations? What are some of the risks that individuals perceive? These are difficult questions to answer without probing some prior, related, questions, such as the state of general knowledge about corporate data use, or of uses of data for research?

Research on individual knowledge of data breaches, cybersecurity, and privacy reveals low levels of technical knowledge but high perceived risk.[104] There are also low levels of perceived self-efficacy to do anything to change their level of knowledge or to change personal habits to effectuate better personal data protection or cybersecurity behaviors. In this respect, researchers are not markedly better than average people: studies of researchers' knowledge of data security and cybersecurity practices reveal that they are often no better than average individuals in terms of data protection and data security.

Compounding the problem of a flummoxed public are cases where "researchers'" use of corporate data turned out to be inapposite to the above definition of research, and anything but public-interested, generalizable, theory-building, or seeking knowledge. Further compounding this problem is that organizations claim that they are "sharing data" that may provide long-term benefits to individuals, such as helping build machine learning applications, but are also reaping near-term financial benefits from that sharing relationship.[105] The degree of sensitivity of the data being shared, or the context in which it is collected, may further complicate this dynamic: more sensitive data is likely to be perceived as less acceptable to share in order to financially benefit the organization.[106]

Contractual Limitations and Legal Liability

Researchers generally obtain access to corporate data in the context of a legal agreement. They face legal risks and liability when using shared data.[107] A first issue might be nondisclosure — what kind of publication, for example, is consistent with protection of proprietary information and other kinds of confidentiality that are codified? The contracts such as data sharing agreements or memoranda of understanding governing the relationships, can be Byzantine in their complexity and contain provisions that complicate the relationship that researchers have with their students and research staff.[108] For example, prohibitions of onward transfer may seem like prudent inclusions to limit sharing of a corporations' proprietary information with others, but, as the student-researcher moves forward with their career, it can limit the uses of a doctoral dissertation or post-doctoral research paper through limits of the data they were built on. Violating the terms of a data sharing agreement, even unintentionally, may put researchers at odds with the expensive legal experts that enforce corporate terms. Finding the funds to fight legal battles over data use is beyond the means of grant-funded research support or university salaries.

PREPARING FOR ACTION

Recommended Actions for Researchers:

Assess available legal, financial, technical, and personnel resources to support a corporate partnership.

Companies also face considerable legal risks when sharing data or even using the shared data from other companies and researchers. Among others, those risks include costs to build contracts that can govern data sharing programs. Costs borne to keep contracts fresh for data sharing arrangements include not only time for legal research and expertise, but also regulatory attention and interpretation. Those contracts may have language that places limits on later innovations such as requirements for data to be shared in a specific hard format (e.g., by CD or DVD) that are inapt for the present computing environment.[109] Likewise, there are changing circumstances for liability for company data uses and sharing that come from changes to case law and interpretation.[110]

Finally, there are international dimensions to the legal considerations. To begin with, again, companies and researchers alike that share data about Europeans will work in a detailed legal and regulatory framework that establishes liability and provides for oversight. Australia, India, and other countries also have restrictions or conditions that might need to be addressed.

Costs

The cost of sharing data for research is non-trivial.[111] Data sharing costs can be grouped into two buckets of spending: relationship management and accountability structures.[112] Costs of data sharing include spending on internal and external personnel costs to build relationships with research entities as well as costs to implement changes to search, file architecture, metadata, and file types or file sizes so researchers can access and manage used data assets.[113] Accountability costs accumulate from the moment data could be extracted: there are costs to building a data source that can be used to power more than ordinary business activities. For data sharing programs specifically, there are personnel and cloud data costs to extract-transform and curate data for sharing, costs for cybersecurity experts and infrastructure to create and supervise the appropriate securitization of data assets, and legal fees to negotiate tractable data sharing agreements.[114]

Data sharing for research is also a source of potential negative returns, such as reputational costs. For instance, what is often missed in the retelling of the Cambridge Analytica scandal is that the individuals responsible for transfer of social media data to political intelligence operations were "researchers". Creating or maintaining a data sharing for research program amid uncertainty concerning research results or the intentions of researchers themselves is a possible source of risk for companies.

Breakout 2: Cambridge Analytica[115]

In 2018, *The Guardian* and *The New York Times* ran exposés on the use of Facebook user data in psychological research performed by Cambridge Analytica. Cambridge Analytica was a political consulting firm. In 2013, a Cambridge University academic created an app that asked users to answer questions for a psychological profile, but the app also was able to access information on the individuals' Facebook friends. The academic and his company, Global Science Research (GSR), contracted with Cambridge Analytica to disclose the data he collected in his research, which, reportedly, was subsequently analyzed and used in Cambridge Analytica's work on major campaigns for politicians in the U.S. and around the world.

Responses

Managing risks and benefits is a complicated dance between many partners, moving on a floor of uncertainty. The maneuvers that these players take to accomplish their goals of data sharing include numerous forms of risk estimation and management techniques, such as ethical review boards and privacy impact analysis.[116] While no technique is foolproof — not all risks can be eliminated under all circumstances — there are ways to identify risks and mitigate foreseen harms early.

Education

The remedy proposed for most situations where people suffer from a heightened sense of risk and a low level of trust, knowledge, and self-efficacy is "education."[117] Educating consumers about the complexity of the data ecosystem surrounding them is no mean feat.

Education-oriented programs can provide individuals with a more sophisticated basis of knowledge that can work to overcome gaps in user expectations. However, developing and targeting useful education tools to important audiences can be a challenge.

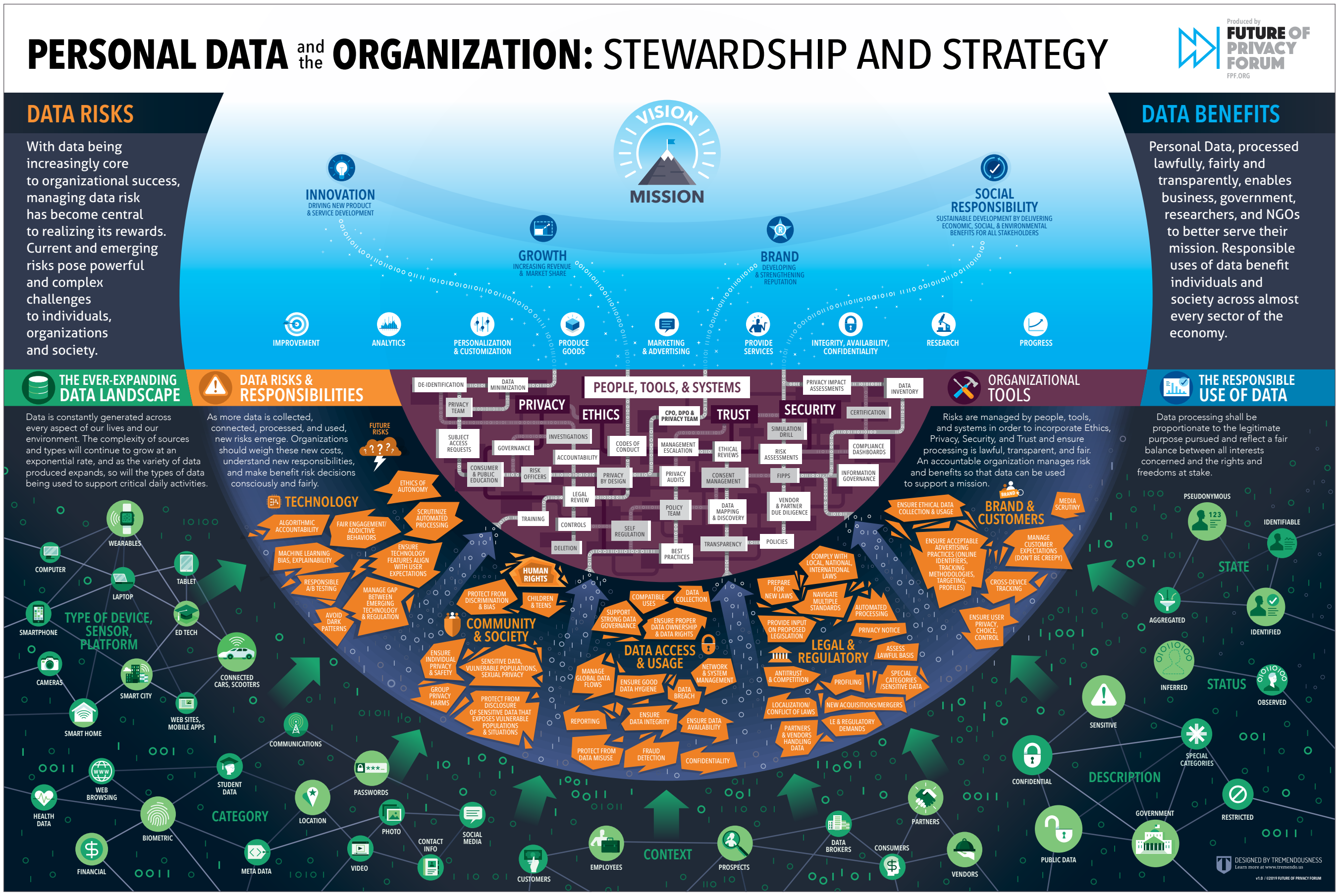
Infographic resources from the Future of Privacy Forum have provided some insights to technically inclined individuals into how organizations use data, including geolocation data and student data. Converting that education to a tractable perception of risk requires more careful and individualized attention. Consumer level interventions to reduce the probability that one's data is included in massive breaches or used in papers later determined to be fraudulent are few, far-between, and have not been sufficiently studied for long term efficacy.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Ensure individuals are informed of how their data will be shared with researchers.

Figure 5: Personal Data and the Organization*



PERSONAL DATA and the ORGANIZATION: STEWARDSHIP AND STRATEGY

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

DATA RISKS

With data being increasingly core to organizational success, managing data risk has become central to realizing its rewards. Current and emerging risks pose powerful and complex challenges to individuals, organizations and society.

DATA BENEFITS

Personal Data, processed lawfully, fairly and transparently, enables business, government, researchers, and NGOs to better serve their mission. Responsible uses of data benefit individuals and society across almost every sector of the economy.

THE EVER-EXPANDING DATA LANDSCAPE

Data is constantly generated across every aspect of our lives and our environment. The complexity of sources and types will continue to grow at an exponential rate, and as the variety of data produced expands, so will the types of data being used to support critical daily activities.

DATA RISKS & RESPONSIBILITIES

As more data is collected, connected, processed, and used, new risks emerge. Organizations should weigh these new costs, understand new responsibilities, and make benefit risk decisions consciously and fairly.

ORGANIZATIONAL TOOLS

Risks are managed by people, tools, and systems in order to incorporate Ethics, Privacy, Security, and Trust and ensure processing is lawful, transparent, and fair. An accountable organization manages risk and benefits so that data can be used to support a mission.

THE RESPONSIBLE USE OF DATA

Data processing shall be proportionate to the legitimate purpose pursued and reflect a fair balance between all interests concerned and the rights and freedoms at stake.

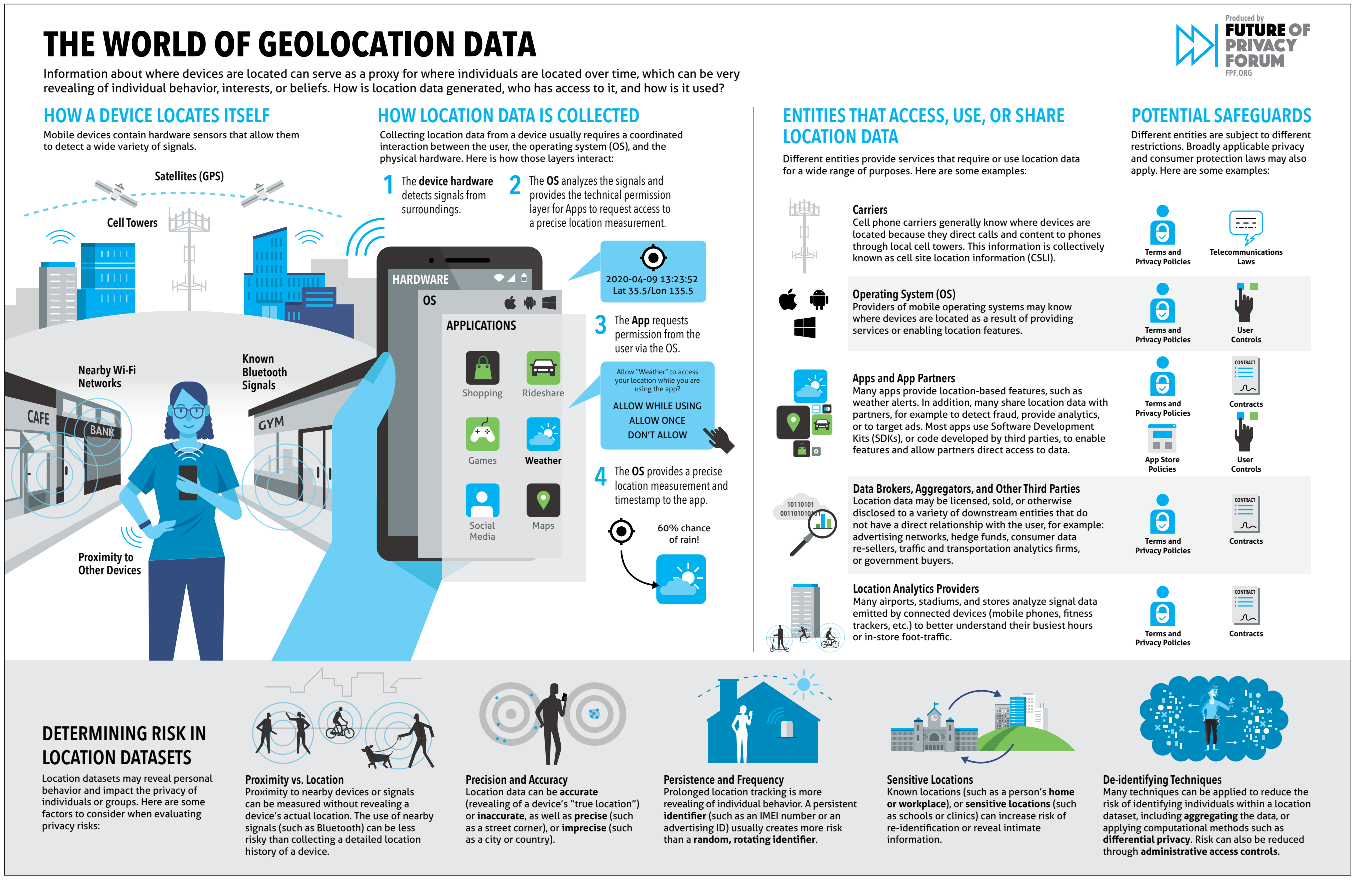
FPF's **Personal Data and the Organization: Stewardship and Strategy** infographic illustrates:

- The complexities of how organizations collect and use data
- The risks involved
- How principled data stewardship supports the goals of innovation, growth, brand development, and social responsibility

Download the infographic at: https://fpf.org/blog/personal-data-and-the-organization-stewardship-and-strategy/[118]

** Organizations can use these resources in the creation of helping individuals understand how their data is used.*

Figure 6: Understanding the World of Geolocation Data*



THE WORLD OF GEOLOCATION DATA

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

Information about where devices are located can serve as a proxy for where individuals are located over time, which can be very revealing of individual behavior, interests, or beliefs. How is location data generated, who has access to it, and how is it used?

HOW A DEVICE LOCATES ITSELF

Mobile devices contain hardware sensors that allow them to detect a wide variety of signals.

HOW LOCATION DATA IS COLLECTED

Collecting location data from a device usually requires a coordinated interaction between the user, the operating system (OS), and the physical hardware. Here is how those layers interact:

1. The **device hardware** detects signals from surroundings.
2. The **OS** analyzes the signals and provides the technical permission layer for Apps to request access to a precise location measurement.
3. The **App** requests permission from the user via the OS.
4. The **OS** provides a precise location measurement and timestamp to the app.

ENTITIES THAT ACCESS, USE, OR SHARE LOCATION DATA

Different entities provide services that require or use location data for a wide range of purposes. Here are some examples:

POTENTIAL SAFEGUARDS

Different entities are subject to different restrictions. Broadly applicable privacy and consumer protection laws may also apply. Here are some examples:

Carriers
Cell phone carriers generally know where devices are located because they direct calls and content to phones through local cell towers. This information is collectively known as cell site location information (CSLI).

Terms and Privacy Policies
Telecommunications Laws

Operating System (OS)
Providers of mobile operating systems may know where devices are located as a result of providing services or enabling location features.

Terms and Privacy Policies
User Controls

Apps and App Partners
Many apps provide location-based features, such as weather alerts. In addition, many share location data with partners, for example to detect fraud, provide analytics, or to target ads. Most apps use Software Development Kits (SDKs), or code developed by third parties, to enable features and allow partners direct access to data.

Terms and Privacy Policies
Contracts
App Store Policies
User Controls

Data Brokers, Aggregators, and Other Third Parties
Location data may be licensed, sold, or otherwise disclosed to a variety of downstream entities that do not have a direct relationship with the user, for example: advertising networks, hedge funds, consumer data re-sellers, traffic and transportation analytics firms, or government buyers.

Terms and Privacy Policies
Contracts

Location Analytics Providers
Many airports, stadiums, and stores analyze signal data emitted by connected devices (mobile phones, fitness trackers, etc.) to better understand their busiest hours or in-store foot-traffic.

Terms and Privacy Policies
Contracts

DETERMINING RISK IN LOCATION DATASETS

Location datasets may reveal personal behavior and impact the privacy of individuals or groups. Here are some factors to consider when evaluating privacy risks:

Proximity vs. Location
Proximity to nearby devices or signals can be measured without revealing a device's actual location. The use of nearby signals (such as Bluetooth) can be less risky than collecting a detailed location history of a device.

Precision and Accuracy
Location data can be **accurate** (revealing of a device's "true location") or **inaccurate**, as well as **precise** (such as a street corner), or **imprecise** (such as a city or country).

Persistence and Frequency
Prolonged location tracking is more revealing of individual behavior. A persistent **identifier** (such as an IMEI number or an advertising ID) usually creates more risk than a **random, rotating identifier**.

Sensitive Locations
Known locations (such as a person's **home or workplace**), or **sensitive locations** (such as schools or clinics) can increase risk of re-identification or reveal intimate information.

De-identifying Techniques
Many techniques can be applied to reduce the risk of identifying individuals within a location dataset, including **aggregating** the data, or applying computational methods such as **differential privacy**. Risk can also be reduced through **administrative access controls**.

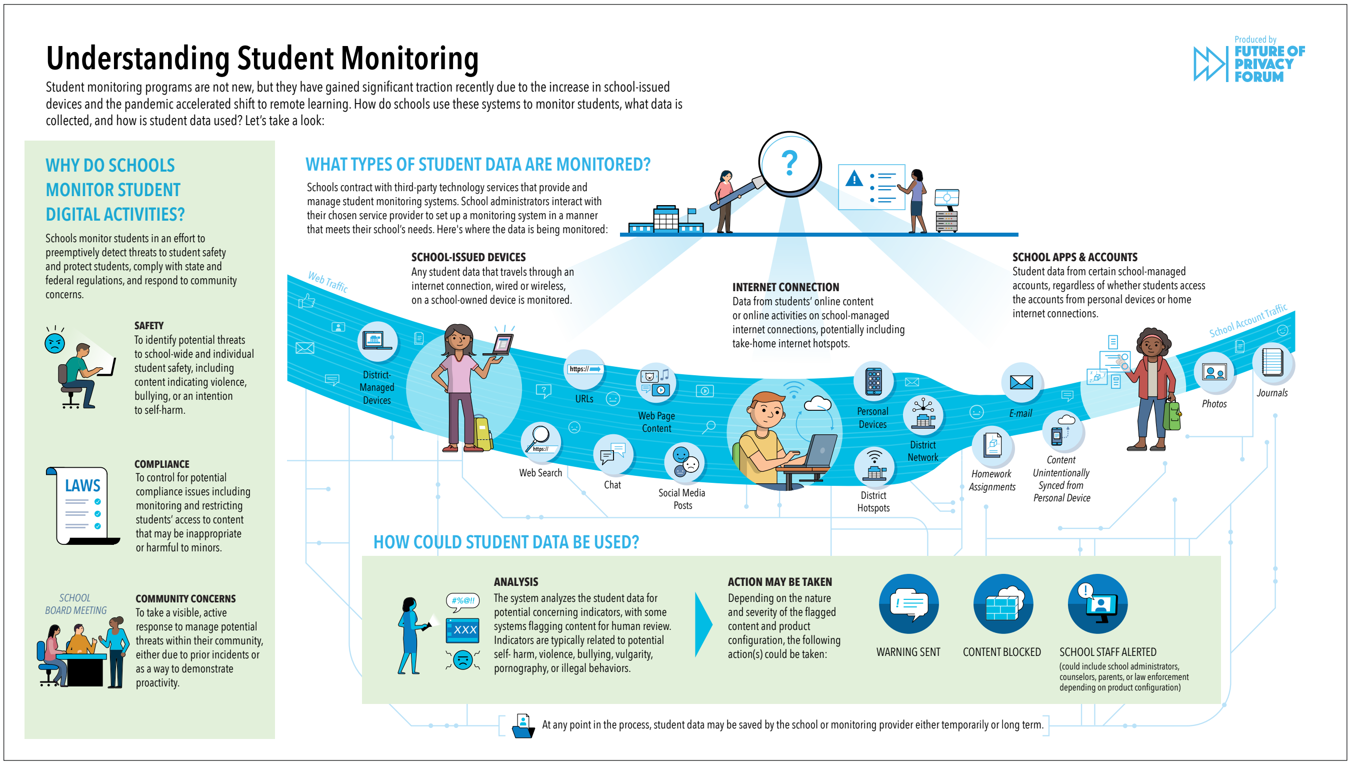
FPF's **The World of Geolocation Data** infographic illustrates:

- The practical basics of how mobile operating systems work
- How apps request access to information
- How location datasets can be more or less risky or revealing for individuals or groups

Download the infographic at: https://fpf.org/blog/understanding-the-world-of-geolocation-data/[119]

** Organizations can use these resources in the creation of helping individuals understand how their data is used.*

Figure 7: Understanding Student Monitoring*



Understanding Student Monitoring

Student monitoring programs are not new, but they have gained significant traction recently due to the increase in school-issued devices and the pandemic accelerated shift to remote learning. How do schools use these systems to monitor students, what data is collected, and how is student data used? Let's take a look:

WHY DO SCHOOLS MONITOR STUDENT DIGITAL ACTIVITIES?

Schools monitor students in an effort to preemptively detect threats to student safety and protect students, comply with state and federal regulations, and respond to community concerns.

SAFETY
To identify potential threats to school-wide and individual student safety, including content indicating violence, bullying, or an intention to self-harm.

COMPLIANCE
To control for potential compliance issues including monitoring and restricting students' access to content that may be inappropriate or harmful to minors.

COMMUNITY CONCERNS
To take a visible, active response to manage potential threats within their community, either due to prior incidents or as a way to demonstrate proactivity.

WHAT TYPES OF STUDENT DATA ARE MONITORED?

Schools contract with third-party technology services that provide and manage student monitoring systems. School administrators interact with their chosen service provider to set up a monitoring system in a manner that meets their school's needs. Here's where the data is being monitored:

SCHOOL-ISSUED DEVICES
Any student data that travels through an internet connection, wired or wireless, on a school-owned device is monitored.

INTERNET CONNECTION
Data from students' online content or online activities on school-managed internet connections, potentially including take-home internet hotspots.

SCHOOL APPS & ACCOUNTS
Student data from certain school-managed accounts, regardless of whether students access the accounts from personal devices or home internet connections.

HOW COULD STUDENT DATA BE USED?

ANALYSIS
The system analyzes the student data for potential concerning indicators, with some systems flagging content for human review. Indicators are typically related to potential self- harm, violence, bullying, vulgarity, pornography, or illegal behaviors.

ACTION MAY BE TAKEN
Depending on the nature and severity of the flagged content and product configuration, the following action(s) could be taken:

WARNING SENT

CONTENT BLOCKED

SCHOOL STAFF ALERTED
(could include school administrators, counselors, parents, or law enforcement depending on product configuration)

At any point in the process, student data may be saved by the school or monitoring provider either temporarily or long term.

FPF's *Understanding Student Monitoring* infographic illustrates:

- Why schools may choose to adopt a monitoring system
- What student information and activities a monitoring service can access
- How a school and monitoring provider may process and use student information collected through a system
- What actions may be taken as a result of a monitoring system flagging a student's activity or information

The following sections further detail how monitoring works as presented in our new infographic, which can be downloaded at: https://studentprivacycompass.org/resource/understanding-student-monitoring/[120]

* *Organizations can use these resources in the creation of helping individuals understand how their data is used.*

Ethics Review Boards

Universities, colleges, and hospitals that conduct research with human participants often are the home to an entity frequently discussed in the context of shared research data — research ethics boards. Research ethics boards, also known as research ethics committees, Institutional Review Boards (IRB), Human Subjects Protections Committees, and other titles,[121] have a regulatory remit to ensure that research performed by qualified researchers using information gathered from living humans participating through intervention and interaction abide by specific requirements.[122] These regulatory requirements outline basic actions, such as requiring informed consent be obtained from participants prior to conduct of most (but not always all) research.[123] Actions like informed consent are required to instantiate ethical principles, such as respect for a person's autonomy.

PREPARING FOR ACTION

Recommended Actions for Research Institutions:

Determine the role institutional review boards and ethics committees will play, if any, in review of corporate data sharing for research.

IRBs (and similar boards) serve as gatekeepers for research to move from ideation to implementation. But importantly, once researchers obtain these permissions, they are able to advance their research out of the institution and into the hands of funding bodies and publication venues. Where funding bodies require a letter or other documentation suggesting the university ethics board has approved the research project proposed to the funder, they place a considerable amount of power in the hands of the board and the university. In the context of conversations involving data sharing for research, IRBs are held up as protectors of research participants and/or as hindrances to researchers' use of shared data.[124] The truth is somewhere in the middle. In the minds of many outside of the IRB system, research ethics committees are envisioned as a genial graduate philosophy seminar. This is incorrect: research ethics committees are highly professionalized regulatory and administrative organizations with considerable expertise in research regulations, grants management, research methodology, data collection, sampling strategies, applied ethics, and statistical analysis.[125] The administrative side of IRBs support the review of projects of all levels of research risk, not only supporting review of that limited number of proposals requiring "full board review" or "expedited review." The limitation that the faculty board members only see a limited number of research protocols means that the IRB administrators themselves review the vast majority of "exempt from review" and draft protocols.[126] In large institutions, there may be multiple specific boards, each with their own research administration professionals, that review hundreds and sometimes thousands of protocols per year.[127]

Of importance to the data sharing discussion is that IRBs do not have a clear regulatory remit to review secondary data uses outside of use of biospecimen data.[128] Calls for IRBs to review any form of corporate data sharing for research puts these institutions into a gray area of "mission creep" and asks these already overtaxed institutions to perform review of complex protocols with no specific supporting expertise on the budget of an unfunded mandate.[129] Such recommendations are likely to lead to repeated cries that IRBs are inconsistent and unreliable arbiters of scientific access and permissions.[130] Universities keen to reduce any risks to their reputation or to avoid incurring costs from data sharing for research will shy away from pushing their research ethics infrastructure into the path of supervising corporate data sharing for research until a clear regulatory remit to do so is established.[131]

Data Sharing Agreements

Ensuring oversight of data sharing for research requires establishing the terms of data sharing programs. The terms of these programs stipulate the parameters against which requests are measured and serve as a passive oversight mechanism to the process. For example, by linking a data catalog or clearly specifying proposal requirements, describing the target public to be benefitted, or the types of personnel or institutions to which data could be shared, organizations create gates to moderate the flow of proposed projects.

PREPARING FOR ACTION

Recommended Actions for Research Institutions:

Develop a specific process for reviewing and approving proposals involving organization data sharing for research, including data management plans and partnership agreements.

There is, however, confusion in the way in which some data philanthropy programs are described. For example, one of the names that data sharing for research is known by in some cases is "data for good." Yet, it is not always clear that researchers are an intended user community for the data holdings under programs with a "data for good" label. "Data for good" programs, such as those described in the box below, often target applied researchers, including journalists, civil society organizations, or non-profits engaged in active problem solving. For some companies, the distinction between applied and theoretical research may be a distinction without significant difference and thus the difference between data sharing for research and "data for good" may be nothing much. Clarifying at the outset which of the user communities. are the targets for an intended data sharing program is one way of creating passive oversight of such programs.[132] [133] [134]

The language companies use to describe their data sharing programs varies widely.

Breakout 3: Describing Data for Good

"SAS is proud to be part of the Data for Good movement, which encourages using data in meaningful ways to solve humanitarian issues around poverty, health, human rights, education, and the environment. From preventing life-threatening illnesses to protecting endangered species to rebuilding after natural disasters, organizations across the globe are harnessing data to make a difference. Applying data for social good has led to new and creative ways to address global issues…" (Data for Good, SAS).

"We empower partners with privacy-preserving data that strengthens communities and advances social issues" (Data for Good, Meta).

"We use the tagline "Data for Good" to capture succinctly the who, what, when, why, and how of data science at Columbia" (Data for Good, Data Science Institute at Columbia University).

"FarmStack is an open-source protocol to power the secure transfer of data across the agricultural sector. It helps users share data directly and enforce usage policy restricting misuse of data" (FarmStack.co).

Breakout 4: Describing Research Data Sharing

"Make real change through discovery and accelerate your research. We help researchers store, host, and analyze their data with easy-to-use solutions. We also provide cloud credits, hands-on technology consultations, introductions to peers, opportunities, press and media support, and more" (Oracle for Research).

"Facilitating efficient and quality research, ensuring data integrity, and fostering a culture of data sharing" (Duke Research Data Initiative).

"Academic Research access: Advance your research objectives with public data on nearly any topic. Enhance your academic research with global, real-time and historical data. Get more precise, complete, and unbiased data from the public conversation for free" (Twitter API for Academic Research).

"Connect patient data at scale to power observational studies" (Academic Researchers & Nonprofits, Datavant).

Where corporations want to narrow the scope of research, they may accomplish this by specifying certain requirements of researchers as conditions to the sharing. This could be data, early presentations of findings, or ownership of copyright on papers. This can also include an obligation of researchers to deliver useful summaries and syntheses to corporate data benefactors. In practice, this means creating lay persons and expert level summaries of papers. This also includes participation in conferences and panels where corporate actors go for insights. For example, academics might present at corporate focused events in addition to professional associations meetings.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Ensure key company stakeholders (including technical, legal, and data personnel) are involved in the agreement process.

Other ways in which data sharing for research can be useful for organizations is to train students in the data wrangling and data analysis techniques most relevant for those corporate actors' business environment. While training students on shared corporate data raises some potential risks as students' understanding of appropriate and necessary data protections safeguards is evolving as they learn, training students to approach data as a part of work in a corporate environment speaks to the broad goal of meaningful collaboration.

Access Controls

Discussions about improving conditions for sharing particularly sensitive types of data with researchers are sprinkled with references to two types of technical environments for facilitating secure and private research data sharing: "data clean rooms" and "secure data facilities".[135] Data clean rooms, which might also be described as confidential virtual machines or secure virtual machines, are well known components of the cloud data sharing environment that corporations already use to share data between themselves.[136] Secure data facilities are physical spaces that house secure hardware and software for use by researchers and others whose security and privacy needs are high.

Breakout 5: Data Clean Room

A “data clean room” is not a physical location, it is a software environment where a user, such as a researcher, can bring together data from one or more sensitive or private sources to perform specific analytical tasks without durable exchange of data between users.

Breakout 6: Secure data facilities

“Secure data facilities’’ are part of the Federal Statistical Data Research center networks, which provide “secured research environments”. According to the Census Bureau, a major source of valuable research data, “Federal Statistical Research Data Centers (RDCs) are Census Bureau facilities, housed in partner institutions, that meet all physical and information security requirements for access to restricted — use micro data of the agencies whose data are accessed there. RDC researchers have access to computing capacity to handle large datasets and complex calculations. Standard statistical, econometric, and programming software, including Stata, SAS®, R, MATLAB and Anaconda python are available in a Linux environment. RDC researchers can collaborate with other RDC researchers across the U.S. through the secure RDC computing environment.”[137]

Data clean rooms might also be known to cybersecurity researchers, computer science researchers, and artificial intelligence researchers, as a sandbox environment.[138] Sandboxes are similar to data clean rooms in that the cordoned off environment of a sandbox allows researchers to test research hypotheses in live or production environments without interrupting the services provided.

Challenges and Opportunities

We have reviewed some of the known responses used to coordinate the many players interested in sharing data for research. There are other techniques that are less clear because they are nascent, such as evaluation of data fitness or because they overlap other areas, such as security. Evaluating each of these challenges is a growth area and opportunity for data sharing partners that are eager to work together on as many fronts as needed to move collaboratively towards a future where researchers use company data to provide insights for research projects that serve both generalizable knowledge and company interests.

Data Fitness

Data fitness is a term of art used to describe the degree to which a data asset — whether that is a single data set or a full enterprise data fabric — is well governed from end to end.[139] As data governance experts define it, “To understand data fitness, you need to first have a good understanding of data quality. A helpful and well-adopted definition of data quality throughout the data quality industry is the fitness to the purpose of use. In other words, the way that you use certain measures, analytics or reports defines its quality or integrity. When it comes to evaluating your health care organization’s data fitness, you need to think about how fit each data element is for its ultimate purpose.”[140]

In the sense of fitness for research uses, well-governed data is company data whose acquisition is documented from start to finish, but also whose metadata, transformations, extractions, connections, and architecture is traceable by external partners like researchers. A data set has a high degree of fitness when it has been managed well from planning to dissemination, allowing for outside personnel, such as researchers, to read the full story of the dataset. This story includes knowing how the data has been transformed by the tools used in the corporate environment.

Making data fit for use includes use of multiple software tools and data management techniques. Unfortunately, each tool or technique leaves a mark on data that researchers may need to know about. Documenting the names, versions, and dependencies of software tools used is already an essential component of good data governance and model-risk-management practices.[141] Researchers need to know the consequences to the data provenance that arise from companies' choices of software tools (aka the tech stack). This is important as up-to-date versions of vendor specific software and even up-to-the-moment versions of programming libraries may be incompatible with out-of-date versions of the same.

Also, within a corporate environment there are bespoke tools, such as desktop automations for data entry and cleaning, that complicate the data story as need to be retold for research use. Revealing the full history of all systems that interacted with a data asset may seem extreme but, in the process of publishing research findings, researchers are increasingly asked to reveal the full provenance of data from byte to table in order to fulfill requirements of peer review and replication analysis.

Ensuring such a high level of data fitness helps companies to share data with researchers but will also help them to share data resources internally. Maintaining data fitness also helps to enable reuse of data for novel applications like machine learning; a research ready level of data fitness is commensurate with the level needed to reduce time spent cleaning data for analytics and machine learning applications. Finally, it is important to acknowledge the dynamic nature of data: data pipelines and their governance change over time, affecting fitness. As a result, researcher and company needs for a given set of data might diverge, which could affect the research, the costs of data sharing, or both.

Security and Privacy

Basic cybersecurity is the floor for all data sharing, but protection of private, proprietary, and other kinds of sensitive data requires more intensive protection. Creating data is not the same as transmitting data, and the tooling available for secure data access may either help or hinder research data sharing. There are many ways in which data can be secured for transit. Many of the techniques to ensure secure transfer of data are already used by sophisticated companies and universities striving to keep their data secure in a distributed work environment. Using first line tools, such as strong encryption and secure cloud environments, to secure data assets for sharing across organizational types is appropriate for many, if not most, research data sharing exercises. There are also more robust and secure tools for data sharing, such as "data clean rooms" (discussed above), that can secure highly sensitive or high-value data assets for transfer or use.[142]

Both researchers and companies may wish to approach the data sharing relationship by acknowledging that it is an undeniably risky affair for both. Following the ethical and technical norms of good data governance is an essential component of well-considered data sharing for research programs. From the perspective of companies, insights into researchers' academic training are important to evaluate their research potential. But companies may rightfully be curious about the other forms of data integrity and regulatory compliance training that academic researchers might have. Companies should consult with the institutional compliance officials who employ the researchers they will work with.

Many companies may be heartened to know that researchers work in an increasingly regulated environment whose ethics and compliance training suite may look similar to their own.[143] For example, companies should know that researchers are trained to handle sensitive data as part of their professional obligations to their institutions.[144] This training is in addition to the research data management and responsible conduct of research (RCR) virtually all research faculty must have.[145] Researchers also work in a data security environment that is similar to that of corporate data security environments: some of the same tooling for data security, cybersecurity, and data management that corporations use, universities use.[146] Alas, just as corporate professionals who are outside of IT or security do not know the details of their security tech stack, research professionals themselves are not likely to know the details of their data security environment. Prior to considering sharing sensitive data, companies should partner their data security professionals with those in the research institution to ensure that there is corresponding or otherwise appropriate technical data security expertise and infrastructure available.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Assess the organization's capability and tech stack for secure transmission of data.

Regardless how well controls are envisioned, there will invariably arise situations where researchers make a credible request for company data that is, ultimately, quite sensitive. Companies may be tempted to take a strongly risk-averse posture to sharing any data that could be considered sensitive. However, a truly open posture to data sharing should include at least considerations for how a company might share sensitive customer information. It is important to note that the sensitivity of data shared does not imply sensitivity of the research performed. Just because researchers are accessing highly refined personal data, does not mean that they must report on their findings at that level of analysis. Assessing the risk of sharing sensitive data for research should always be paired with collaborative discussions about the statistics and metrics planned for presentation by the researcher. Protocols, such as keycoding, pseudonymization, and identifier stripping, should be discussed in the research data sharing agreement and should also be a key component of research data sharing proposals and data management plans.

PREPARING FOR ACTION

Recommended Actions or Organizations and Research Institutions:

Ensure personnel, privacy, and cybersecurity controls are in place that are appropriate to the level of sensitivity and value of the data.

One of the biggest challenges to data sharing is the belief that all data that is shared is subject to a risk of exposure through re-identification attacks. Re-identification through a combination of public and private data sources is known to introduce risk. It is good practice, as is the case of many of the contracts and law related to data sharing, whether for research or for market research purposes, to stipulate against re-identification.

How likely is it that a person or persons could be re-identified from a shared data resource? How likely is it that the same person or persons could be identified given application of the many combinations of security and privacy controls that are already in use? The summaries of the re-identification research already done are compelling, but these do not measure the risk of re-identification in a way that gives a measure of the per customer amount of time, effort, or energy an attacker would need to spend to achieve re-identification.[147] In addition to taking affirmative steps toward preventing specific re-identification attacks, companies and researchers should collaborate on research that measures the overall utility of privacy enhancing technology (PETs), including use of synthetic data, as a tractable and scalable way to prevent re-identification attacks.[148]

Synthetic data is often held up as a panacea for concerns about re-identification. Synthetic data and creation of synthetic data sets can provide researchers the opportunity to conduct this research without raising the specter of specific persons re-identification. The issue of phantom data re-identification, or the belief that someone has been re-identified in a synthetic or systematically augmented data set is also something that needs to be measured through rigorous research.[149]

Creating carefully drawn contracts, strong collaborative relationships, and engaging in careful sharing of research resources including personnel know-how for re-identification research must be done if all parties to data sharing are going to move forward in a realistic, research-informed, and consumer-protective manner. Companies, researchers, consumers, and research institutions should work closely together to more carefully research and assess the true breadth and depth of privacy risks that arise when data is shared between organizations. Ultimately, for companies struggling to share data with researchers, the challenge is to pair sharing tools with the level of sensitivity or security required by the data.[150]

The Regulatory Environment

U.S. and international law and policy may provide specific requirements or limitations on data sharing. Just as corporations learning about research data standards might find themselves bewildered by the number of available standards, researchers striving to learn about corporate data standards may struggle to understand the complexity of the overlapping regulations that companies must abide by when dealing with consumer data.[151] Companies spend millions, if not billions, of dollars per year to comply with data security and data protections standards.[152] How can researchers catch up to the knowledge that companies have so that they can operate as good stewards of shared resources? A good place for researchers to start to understand the complexity of this landscape is by looking closely at a single regulatory scheme, such as that under the Health Information Portability and Accountability Act, or HIPAA, as well as the accompanying regulations, known as the HIPAA Privacy Rule and HIPAA Security Rule.[153]

The many detailed terms for compliance with HIPAA may be beyond the scope that researchers can understand and implement fully but understanding the basics of the HIPAA rules for Privacy and Security is a useful place to start. Fortunately, most researchers working for universities or other institutions like hospitals will have access to both HIPAA experts and related training resources. The HIPAA Privacy Rule, which occupies an outsized and somewhat confused place in the public (and professionals) views of consumer data protection, sets the protection standard for personal health information (PHI).[154] The HIPAA Security Rule sets the terms for data breach. This archetypal personal information protection standard sets the tone for requirements for data privacy, storage, transfer, and retention. The eighteen identifiers that must be removed under the Safe Harbor standard set the tone for data privacy. Data must be stored in areas with physical access limits in place. But, locking up paper files, office doors, and screens is not enough. HIPAA requires technical and administrative safeguards, such as limiting access to data to the minimum number of persons needed to accomplish a legitimate goal and ensuring that data is encrypted to NIST standards whether the data is at rest or in transit. Likewise, HIPAA sets the terms for data transfer by designating what can be transferred between covered entities (CEs), such as other health providers or health plans, and business associates (BAs), such as lawyers, accountants, and data storage or data encryption providers.

HIPAA is only one of many rules that provide a large patchwork of standards with which companies that gather data, transfer data, or use data must comply. Other data standards that constrain data availability for third party uses include data protection and data security for payment cards (PCI DSS), for financial services data (Gramm-Leach-Bliley Act and Fair Credit Reporting Act), for government data (FISMA), and for education data (FERPA).[155] For researchers working in areas for which corporate actors will need to comply with these requirements, learning more can promote end-to-end data protection for consumers but also promote collaboration and even reduce barriers to acceptance of requests to share data for research.

Trade Secrets

What about the use of software or data gathering techniques that are a trade secret to the company? Where restricted company property (e.g., specific software accelerators) must be discussed as part of the story of a research dataset, the appetite a company has for risk of disclosure must be re-examined. If unintentional exposure of the use of a specific tool through the research review process is something the company can tolerate, then moving forward with statements describing the use of proprietary data may be acceptable. But, where any exposure of company products and systems through the processes involved in fulfilling the terms of a data sharing for research relationship, the relationship may need to be reexamined. One alternative is to “in-board” a researcher to conduct their research projects on premise, and/or in a trusted execution environment, and accept lesser degrees of publicity about research products. Those tighter controls on research outputs can be captured in the contractual mechanisms that govern the relationship, such as a data sharing agreement.

Publication Requirements

A clear norm of research is that research is not complete until it is published. Experts in research ethics go so far as to suggest that performing research using human-generated data, such as from clinical trials or educational surveys, without publishing findings contravenes the principle of respect for persons.[156] If publication is an ethical requirement for researchers overall, then publication of findings from studies using shared data is also ethically required of researchers. Researchers also confront publication incentives within their organizations and professional communities. However, in some cases, the outlets for publication — journals, preprint archives, conference proceedings — can create barriers or incentives for publication.

One way that publication venues create challenges for data sharing is by requiring researchers to publish their data or to share data for the purpose of peer review. Obligations of researchers to publish raw data may introduce complexity into the data sharing arrangements made between companies, research institutions, and researchers. The obligations to publish data extend to journals across multiple disciplines as described in Appendix II: Chart of Publication Requirements for Data.

Encouraging journals and conferences and book publishers to be conscientious and flexible in terms of their data publication requirements can be one way for research outlets to encourage data sharing. Editors of journals and book series should identify how their terms and conditions for publication do not unduly contribute complexity for situations where researchers used shared data assets. Researchers and companies sharing data should carefully collaborate with research publication outlets when they do share data.

Research Misconduct

In the process of peer review or replication analysis, a dark side of research may be revealed — research misconduct. According to the Office of Research Integrity,

> "Research misconduct means fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results. (a) Fabrication is making up data or results and recording or reporting them. (b) Falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record. (c) Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit. (d) Research misconduct does not include honest error or differences of opinion."[157]

Not only does misconduct in a research environment create substantial questions about the process and results from that research, but it may raise ethical or reputational questions about the researcher, the research institution, or even the organization originating the data. Should misconduct be alleged, organizations should consider in advance if they are willing to share data that could be subject to a misconduct investigation within research institutions or even at federal agencies and that decision should be commemorated in any agreements between the relevant entities.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Ensure oversight and accountability mechanisms are included in data sharing agreements.

Within the context of data sharing agreement language, the willingness of a company to share information about the provenance of shared data to institutional research integrity and federal research integrity offices should be considered as this may be required for thorough investigation. Finally, in a well-developed data sharing program, companies may need to be prepared to share data in support of misconduct investigations or careful replication research even where they were not the originators of the data used.[158]

ACT III: MANAGING THE STAGE

The story of data sharing for research is a multi-part opera, not a one act play. When companies, researchers, and their many supporting actors enter into a data sharing relationship, they must plan to participate in a long-term, multi-partner engagement. While that arrangement will be exciting and beneficial, the relationship will also cause feelings of anxiety and heightened sensitivity to uncertain risks. The players in the opera can properly prepare for the engagement by considering the risks and challenges laid out above and taking proper steps to prepare for a data sharing relationship. Tasks to prepare for a data sharing relationship include program development, such as gauging available resources and ensuring value alignment on the purpose of the program.

Building the Program

Companies considering building a data sharing program have some essential pre-work to complete prior to building a data sharing program. These tasks can include assessing readiness for data sharing, including, for example, thinking through the appetite for the risks, benefits, and spending such programs will entail. Of course, assessing readiness will also entail amassing knowledge of data that could be shared and determining the levels of both personnel and system capacity for building and maintaining the sharing relationships. These components cannot be assessed in isolation from each other but are part of an iterative process all actors should take stock of when starting, measuring, modifying, or even ending, their data sharing programs.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Ensure clarity of expected benefits and likely challenges involved in a potential data sharing for research partnership.

Assess Readiness

The first task is to revisit your organizational values. Sharing data requires a posture of openness to the joys and challenges of collaboration under conditions of uncertainty. The extent to which your organization values openness and external collaboration will set the tone for your ability to accept the risks and costs of a data sharing relationship. Organizations contemplating a data sharing for research program need to evaluate whether their values align with a philanthropic approach that includes "data philanthropy" as a non-traditional form of philanthropy.[159] Just as traditional philanthropic giving carries risk, data philanthropy, which covers data sharing for research purposes, carries risk that must be balanced with expected benefits and values.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Determine if and how data sharing for research fits the organization's values and strategy.

Openness and collaboration are terms loaded with positive connotations; suggesting that your organization may not be able to accept a maximally open approach to data sharing for research may seem like a negative statement. It needn't be. In fact, limited approaches to data sharing for research can still have a powerful impact. These approaches may include limiting the types of research supported, the number of projects supported, the scale and the scope of data shared, and the mechanisms used for continuous monitoring and accountability.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Ensure open lines of communication between all parties.

Preparing for even a limited relationship will require measuring organizational readiness to engage on each of these indicators.

Breakout 7: Are You Ready to Share Your Data?

Essential questions to ask as part of your value assessment and alignment might include:

- Does your organization have a model research project that you would be proud to sponsor?
- Can you identify important KPIs from that research project that you could translate into performance indicators for future projects?
- Does a model research project align with your business needs, including R&D innovation or talent identification needs?
- Can your organization realistically measure a model research project against others to determine how well the sponsored project meets valuable KPIs?
- Does your organization have a data architecture that allows for extraction and transformation for non-business purposes?
- Do your data workers (including database administrators and software engineers) have the time and resources necessary to perform additional work? Can they perform this work more than once as there may be modifications to data requests to be filled?
- Can your organization spare the capacity of managers and project managers to oversee the relationship with the research team?
- Do your employees support the endeavor?
- Are you confident that your firms' reputation will weather any adverse findings by the research team?

Understanding Essential Components of Data and Partnerships

If the first task is value alignment with fulfilling this non-traditional philanthropic mission, the second task is to determine the fitness of the data and take stock of your data tooling. Completing the second task includes more than checking if your data holdings include data you can lawfully share. It includes evaluating whether you are willing to allow others access to that data.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Know current legal requirements regarding data sharing.

For companies not accustomed to treating data as a specific product to be shared repeatedly outside of the data-owning teams, it can be daunting to decide to share large volumes of structured data or even small volumes of unstructured data. There are two strategies to reduce the sense that building shareable research data is overwhelming or too costly. The first is to actively develop collaborative techniques for data documentation that will meet researchers' needs. Doing so might have benefits to the company, such as surfacing sources of data that have not been analyzed internally. The second is to crosswalk data documentation for research data sharing programs with data documentation for activities like data operations, feature engineering, or [machine learning] model risk management. Both strategies represent part of a credible corporate data quality management process and collaboration with researchers on this can be valuable. For example, recent studies suggest that, "When pulled together, the tally [of data management costs] can be jarring. A midsize institution with $5 billion of operating costs, for example, spends more than $250 million on data across third-party data sourcing, architecture, governance, and consumption."[160] To the extent that collaboration with researchers helps to reduce this cost while also building collateral with research institutions, companies may do more than recoup unseen costs.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Assess capacity to absorb workload and costs associated with data sharing for research partnerships.

PREPARING FOR ACTION

Recommended Actions for Research Institutions:

Assess cyber and data management capacity for the additional workload and costs associated with secure data sharing.

There are certainly overlaps between a companies' well-governed data and a researchers' well-managed data. Since a corporate data set that is fit for research use, as described above, will have some of the characteristics that a fit research data has, organizations like the Inter-university Consortium for Political and Social Research (ICPSR) or the Consortium of European Social Science Data Archives (CESSDA) may provide a resource for organizations looking to create fit research data. Likewise, performing a crosswalk between a data governance plan and a data management plan required for researchers in associated disciplines is a good starting point for improvement of overall data governance beyond data sharing.

Figure 8: Steps in Data Management

Steps in Data Management
Store
Organize and Document
Process
Store
Protect
Archive & Publish
Discovery

Source: CESSDA DMEG: Data Management Expert Guide, January 2020 (*Data Management Expert Guide*, 2017- 2020)[161]

Another strategy that could create mutual benefits for an organization striving to manage their data well and researchers hoping to access useful and credible data is to start with the ideal normative and technical guidance for research data sharing. (Remember that data sharing for research is transfer of data between organizations with a research purpose). The current normative guidance that researchers should follow to ensure that their data fulfills research ethics norms is that data should be Findable, Accessible, Interoperable, and Recoverable ("FAIR").

Figure 9: FAIR

Term	Meaning
Findable	Easy to find for humans and computers
Accessible	Ability to access the data
Interoperable	Standardzed terms, use with other applications, workflows or processing
Reusable	Described in such a way that humans and computers can understand with a clear data usage license

Source: FAIR Principles[162] (*FAIR Principles*)

Building research data use into corporate data activities includes collaborating with researchers and/or the artifacts of the research profession. But direct researcher collaboration or re-imagining a corporate data governance program to also fit research norms will require intense periods of collaboration, interpretation, strategy setting, and careful implementation.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Determine the fitness of company data for data sharing for research.

A less collaborative, but no less useful, approach to building (or retrofitting) corporate data holdings to meet known data standards is to explore and crosswalk the elements of a quality research data management program, including use of explicit data standards, metadata standards, and careful attention to versioning data assets. Data standards, including metadata standards, controlled terminology tables, data taxonomies, and data ontologies, represent a valuable, but burgeoning and bewilderingly complex, set of resources for translation into company practices. How might companies approaching data sharing identify whether a particular data standard is right given data holdings and research data sharing intentions? At a glance, there are over 200 metadata standards that are specific to particular research disciplines.[163] For large companies whose data holdings could be used by numerous academic researchers from various fields, it can be extraordinarily challenging to determine which of these standards they should aim at.

Breakout 8: The Right Standard for the Right Field

Different fields of research may follow different data standards, and therefore data fit for research within that field should be aware of the appropriate standards. For example:

- For companies that employ social scientists and anticipate they will collaborate with social sciences, a general standard might be the Data Documentation Initiative or DDI.[164]
- For medical and healthcare-oriented businesses that collaborate with health researchers, the many resources from Clinical Data Interchange Standards Consortium or CDISC is a good place to start.[165]
- For companies with data pertinent to environmental management, the US Environmental Protections Agency or EPA has many standards available to describe its data and data for related fields.[166]

Striving to meet some general-purpose data standards, such as The Dublin Core Metadata Initiative, or the National Information Exchange Model, is a good place to start.[167] Where general standards do not seem appropriate, a good rule is to look inward, toward your research and development personnel. Leveraging employees and existing partners' research expertise is one way to design data that could be useful for training future researchers, and future employees, for your company.

Executing the Program

Building the data, hardening transmission channels, and managing known sensitivities are all essential parts of a research data sharing program. The next task to make data ready for research sharing is communicating that data is available, soliciting proposals for data use, and evaluating those proposals.

Simply identifying research data sharing opportunities by navigating program descriptions or even platforms for data sharing access are a first and low gate that researchers must cross to receive permission to use company data. This gate can be circumvented by researchers in some cases through personal relationships or by simply emailing a request. Whether proposals come through established gates or are submitted in paths around them, the real work begins when the recipient of a data use proposal begins to check the quality of research proposed, the tenability of proposed technical mechanisms to protect privacy, or the fitness of their data holdings to fulfill the request.

There are few reliable metrics for good research that work for all research disciplines or interdisciplinary fields. Mining the abstracts of funded projects such as, are available on foundation and other funder websites, gives some insights into what others thought was fundable. But these repositories do not clarify why the vast unseen troves of what was submitted do not make the grade. Were the proposals only a little bit off or were they way off? Did they speak to some funder priorities and not others? Were they so novel as to be path breaking or so conventional as to be droll? Recent research suggests that the conventional ways of evaluating research proposals may not perform any better than random chance when important outcome measures, such as number and impact of papers published, are evaluated.[168] Consequently, taking a novel approach to reviewing proposals should not be seen as introducing any additional risks to companies sharing data.

PREPARING FOR ACTION

Recommended Actions for Organizations:

Develop a process to communicate data sharing opportunities, review research proposals, contract with research institutions, and share data.

Evaluation Criteria

One metric of good research should be obvious enough to include by now: good research proposals include plans to manage research data. Other metrics include the salience of the proposed research to answering questions of burning interest to the company sharing the data, the relevance of the research to challenges the company faces with respect to customers or employee's health, safety, and welfare, and the probability that the research produces innovative ideas that could be translated into organizational products and processes.

There are other salient measures of research proposals that go beyond the research itself. With respect to organization values, a key question is to determine if company efforts to ensure diversity, equity, and inclusion (DEI) extend to those adjacent to the company. Some researchers argue that data sharing for research programs are presently like a "17th century medieval castle", accessible only by a chosen, hereditary, few. To the extent that DEI is a company-wide goal, companies will need to attend to communications with smaller, low-profile, institutions. Companies may also need to evaluate whether their risk appetite in the data and research space includes a willingness to explain why data was shared with Stanford University and not Stamford University.

The most important question to ask concerning a company data use project is whether the research project has potential to benefit the individuals whose data is shared. But asking whether the proposed research also benefits the company or the research institutions are also important questions. These questions may need to be answered before some of the more narrow and conventional questions are asked: is the research plan well-reasoned, well-organized, and based upon a sound grasp of the field? Does the research plan incorporate a mechanism to assess success of the data analysis? How well qualified is the individual team or organization to conduct the proposed activities? Are there adequate resources available to the principal investigator either at the home organization or through collaborations to carry out the proposed activities? Answering these questions is not easy as the answers bring together statements of values and preferences for the terms of important relationships.

Evaluating research proposals is made more difficult by the lack of open, general, guidance for how to evaluate such proposals. Through the multi-stakeholder listening sessions conducted as part of the background research for this playbook, we identified five actionable terms for evaluating research proposals: ethical soundness, privacy preservation, scientific validity, compellingness, and methodological rigor.

Ethical review

During the FPF salon dinners, speakers frequently raised the point that data should only be shared for research that meets high ethical standards. Achieving consensus about ethical standards is challenging. Questions about which principles serve as adequate standards, how to measure the achievement of those principles, and how to evaluate the degree to which these principles are achieved are weighty questions that motivated over 100 years of research ethics research and continue to provoke considerable discussion today.[169] Focusing on a specific context, such as research using platform data in Europe, has facilitated the developing of targeted codes of ethics.[170]

Throughout our discussions, we returned to existing research ethics review organizations, such as Institutional Review Boards (IRBs) that focus on the ethics of human subject research as defined by the US Common Rule. IRBs were held up as the example of organizations that *should not* be the sole mechanism for conduct of ethical reviews for projects employing shared data. The reasons that IRBs should not be the place for review of research data sharing include lack of regulatory remit to do so, lack of resources (expertise, time), and pressures to protect their own institution and researchers interests over a broad public interest. However, the seventy years of history and extensive professionalization of ethical review by IRBs provides a treasure trove of guidance and insights for companies to use if they wish to evaluate the ethical profile of research projects. These learnings are a resource that companies and other organizations, such as FPF, can use to build and provide third party options for ethical review of research (see EDUC box below).

PREPARING FOR ACTION

Recommended Actions for Organizations:

Consider establishing or partnering with an ethical data use committee to assess data sharing.

Privacy Review

When shared data includes personally identifiable information (PII) or includes data that could foreseeably be recombined with other data to reveal information that is personally identifiable, that data should be evaluated for its impact on individuals' privacy. A Privacy Impact Assessment should be a routine component of any data sharing activity, even if that data is shared between disparate units of the same corporate entity who use the data for different reasons and recombine it with different data sources. FPF issued best practices and contract guidance for protecting privacy in the context of research data sharing.[171] The ten points emphasized: 1. Data sharing agreements, 2. Due diligence and oversight, 3. Data minimization and de-identification, 4. Data security and integrity, 5. Vendor management, 6. Data retention and deletion, 7. Ethical data use, 8. Independent review, 9. Publication expectations, and 10. Training and education. These points are distilled into contract-level guidance for use as a basis against which contracts governing proposed could be compared.

Scientific Merit

Scientific merit is a well-known standard for review of research. But what is meritorious science? An analogous case to the determination of scientific merit of proposed research can be found in the legal field in the "Daubert Standard".

"Under the *Daubert* standard, the factors that may be considered in determining whether the [expert testimony] methodology is valid are: (1) whether the theory or technique in question can be and has been tested; (2) whether it has been subjected to peer review and publication; (3) its known or potential error rate; (4) the existence and maintenance of standards controlling its operation; and (5) whether it has attracted widespread acceptance within a relevant scientific community.

A research project that meets the norm of scientific merit is one that answers a question in more than a yes or no fashion by testing a deductive hypothesis or inductively reasoning to a new conclusion to fill a gap in a theoretical framework recognized by a scientific field. The professional associations responsible for overseeing, accrediting, or supporting a field of science have mechanisms for evaluating the merit of research for presentation at professional conferences or recognition of best papers. The merit of interdisciplinary or multidisciplinary studies can be evaluated by scientists from those interdisciplinary approaches or the unity of approaches represented in the interdisciplinary proposal. There is a small risk here, however, that the scientific merit of studies will be judged conservatively and that truly novel approaches may take time to be rewarded with approval for shared data use. Companies eager to share data may wish to assess their appetite for reputational risk that may arise from being associated with as-yet-untested or truly novel research methods not yet clearly supported by the conventions of a research field.

Compellingness

That something is scientifically meritorious does not mean that the research is compelling. Compelling research is science that a reasonably well-informed reviewer could judge would be useful when making a case for action by public bodies or even serve to produce factual evidence that would be permissible in a court of law (see scientifically meritorious). Compellingness can be a meatier standard, for example that such research would be used to make decisions that would reduce existential or otherwise grievous harms to known groups.

That research is compelling to companies supporting it with their data does not mean that the research will be widely accepted. In fact, it may be deeply critiqued or scrutinized. But a compelling research proposal might be sufficient to alter a research paradigm and advance the field in a new and unforeseen direction. Additionally, while companies may engage in compelling research, compellingness needs to be defined by the scientific community rather than the interests of the company.

Methodological Soundness

An important dimension of intellectual merit is methodological soundness. In clinical research, for example, randomized controlled trials — properly executed — are an exemplar. Methodological rigor is relevant to any kind of research, however, and should be addressed in the proposal and/or description of the work.

Importance (or not) of Identification

Should companies base their grants of access to research data primarily on the characteristics of the research project or on the reputation and background of the researcher? Deciding to share valuable assets such as data requires considerable trust in people and institutions. It can be easier to build trust in tangibles, such as persons or institutions, than in intangible processes. But, in the research data sharing environment, trust in the processes and professional practices that define research, such as are reflected in proposals that exhibit soundness of methods and ethical reasoning, present a more durable and defensible basis for decisions to support research. Further, companies can avoid the appearance of discrimination by reviewing research proposals strictly on their merits, such as through a double-blind peer review process.

The professional background of researchers gives companies insights into the capabilities of that person to follow through on producing credible research. Likewise, the background of a research institution gives companies a window into the reputational and other resources at the disposal of any project team. If a research data asset needs to be shared with stringent security conditions or if the size of a research data asset will require extraordinary storage or use capacity (e.g., high-performance computing facilities), the infrastructure

Breakout 9: The Ethical Data Use Committee (EDUC) at the Future of Privacy Forum

The Future of Privacy Forum designed an Ethical Data Use Committee (EDUC) as an ethics review committee with the specific purpose to review research projects using shared data. We established the committee on the model of a human subject research ethics committee, such as an Institutional Review Board (IRB). The EDUC is led by a standing committee of experts from fields essential to understanding data protection and privacy from both a legal and a technical perspective; expertise in areas of the human sciences (e.g., social sciences and clinical research); and expertise in physical sciences (e.g., agronomy and physics). Following the models of existing review committees, we also established the terms for participation of ex-officio experts who can advise on review of research projects in specific areas of expertise, such as computational social science or artificial intelligence.

The membership of a research review committee is an important part of its constitution, but the principles against which it reviews proposed research may be the most important to the research data subjects that the committee protects. Using the existing rich language concerning data ethics and research ethics, we identified nine ethical principles essential for thoroughgoing review of research data sharing projects.

We did not stop at stipulating principles and brief definitions for review. We adopted the spirit according to which the original drafting of the Belmont Report (1975) was done, arguing for a pragmatic approach to ethics wherein each ethical principle should sit on a spectrum of achievement of practical tasks that fulfill a particular ethical principle. For example, instead of requiring a review board to determine that proposed research fits a definition of accountability, we ask reviewers to determine which accountability tasks are described as accomplished in a proposal and to grade the proposal against a lower limit for accountability, a middle ground of accountability or what should be reasonably expected of research data sharing projects, and what a superlative level of accountability related efforts would require. Where researchers' proposed and planned activities fall below the mean standard of accountability, they are encouraged by the review board to improve in terms of accomplishment of essential accountability tasks. Where they have exceeded the mean of this then they are rewarded with recognition for superlative ethics achievement. The choice of an accomplishments-oriented method for review of data sharing reflects the belief that data sharing is an interactive process of actors relating to one another over time.[172]

of a research institution will play an important role in decision making. These capabilities should be evaluated directly, rather than by making assumptions based on the name or reputation of an institution. A lingering challenge to reviewing research proposals based on a researcher's or a research institution's characteristics is determining whether there might be a financial conflict of interest or a conflict of commitment. The proposal process can call for disclosures, research administration operations at an institution can assist in vetting their researchers, and government entities such as the Office of Research Integrity (ORI) at the Department of Health and Human Services can provide additional guidance.

Special Considerations for Research Institutions and Researchers

Stories of data sharing for research are retold with a subtext of "David vs Goliath" or Charles Dickens' Oliver Twist: the sole or small researcher seeking something from a gargantuan, wealthy, and unfeeling adversary. This morality tale neglects the role of another large and significant player in the data sharing space: research institutions, like the University of California system schools, and other research institutions, such as Woods Hole Oceanographic Institute.[173] Researchers are backed by these culturally valuable, resourceful, businesses in their own right, who can perform the critical functions that push data through from one set of hands to the next.

Just as organizations and individual researchers must be ready to participate in data sharing relationships, research institutions must also be ready. The dimensions of readiness for research institutions are largely similar to those of company readiness, discussed above, and will not be elaborated on as extensively except where their specific role vis-a-vis evaluation and support of research and researchers might either help or hinder.

PREPARING FOR ACTION

Recommended Actions for Research Institutions:

Establish open lines of communication with corporate partners' legal, development, and data governance personnel to assess readiness to engage in a corporate partnership.

An important distinction to be made up front in the discussion of the roles of research institutions in data sharing is that of publicly funded and privately funded universities. Publicly funded universities, or those whose budgets come from the coffers managed by state legislators, are under pressure from those states to push forward (or push under) specific types of research. For example, many state legislatures have pushed forward bills establishing centers of excellence or research centers for big data, artificial intelligence, or data analytics. Other states encourage "academic-industry connections", "knowledge transfer", or "town and gown partnerships" that come close to or even explicitly tie funds to acquisition of corporate support, including data sharing.[174] Targeted state initiatives can sometimes provide incentives and resources to pursue specific research agendas.

Breakout 10: Legislative language requiring academic industry connections, including data sharing

Example Language from Legislation Addressing Data Sharing

HR3912: "Financial Data Sharing Choice Act. This bill requires financial institutions to obtain consent from a consumer before sharing that consumer's nonpublic personal information to a nonaffiliated third party. Currently, a consumer must opt out of such information sharing." From HR 3426 "Sec 6, (a) (c) (3) the adoption of shared data privacy, data sharing, and data archiving standards among the United States and partner countries and relevant economic and political unions, including harmonized data protection regulations;"

HR 2225: "award grants to support research and development activities to encourage greater collaboration and coordination between institutions of higher education and industry to enhance education, foster hands-on learning experiences, and improve alignment with workforce needs;"

S1397: "Among other activities, the CDC must (1) develop guidance for state and local health agencies to improve birth and death record data for American Indians and Alaska Natives; (2) enter into cooperative agreements with tribes, tribal organizations, urban Indian organizations, and tribal epidemiology centers to analyze and address certain inaccuracies related to records for American Indians and Alaska Natives; (3) adopt uniform standards for the collection of health data on race and ethnicity; and (4) encourage states to enter into data sharing agreements with tribes and tribal epidemiology centers to improve the quality and accuracy of health data."

Where legislatures create funding incentives or disincentives for universities to seek corporate support, universities' interests may lie in creating smoother pathways, including technical and legal support, for data sharing. Legislative initiatives may press universities to partner with industry, but those pressures invariably flow down into the requirements that deans and department heads must put into employment contracts for faculty. Often missing in the simplified story of data sharing for research is the salient point that faculty researchers are employees of universities. While tenure protections create substantial barriers to removal of (the dwindling numbers of) tenured faculty,[175] even the highest-ranking endowed chair professors still have employment contracts and evaluation of performance. To the extent that universities include creation of data as a component in evaluation of tenure, they support creation of research data. Likewise, to the extent that universities include "knowledge transfer" or "academic industry connections" as part of the evaluation of research faculty members, they also incentivize researchers to seek relationships with companies.[176] Of course, universities can also create disincentives to partner with companies or to use shared data, just as companies can create disincentives to partner with researchers or to share data. For example, universities can press faculty to engage in more and more teaching or relationship cultivation with other types of organizations (e.g., partnerships with local governments and civic leaders) that would cut into time needed to liaise with companies or build data.[177] Employment contracts are the tip of the incentive iceberg: universities often create smaller pools of incentive funding to pursue novel projects with new or specific partners or create incentives to collaborate with company R&D teams. Universities can also release researchers on sabbatical or leave to pursue a project with a company.[178]

Universities and research institutions pave the way for data sharing relationships between companies and researchers. But, ultimately, production of the creative and publicly beneficial uses of company data fall to researchers. How can researchers prepare to be good stewards of shared data?

PREPARING FOR ACTION

Recommended Actions for Researchers:

Ensure open lines of communication between the data sharing company and your institution to manage expectations, responsibilities, and necessary changes to address data protection concerns.

PREPARING FOR ACTION

Recommended Actions for Research Institutions:

Ensure open lines of communication between company, university, and researcher including expectations, responsibilities, and notification of and responses to data protection concerns.

Many of the documents addressing data sharing for academic research focus on what companies must do to be stewards of the public interest and to serve this role by making consumers' data available to researchers. Few focus on what researchers should do to be good stewards of data. As some scholars who study researchers' data practices have found, what often stands in the way of researchers either sharing data with others or using shared data is concern about the integrity and privacy of that data.[179]

> "We argue that a prerequisite to data sharing is to have a data management and sharing policy as well as associated processes, tools and governance mechanisms in place. We acknowledge that data sharing is indeed occurring, albeit without the existence of institutional policies and with gaps such as inequity in data access and reuse. In addition, much data sharing occurs without the implementation of basic data management standards, e.g., the sending of datasets via non-secure channels such as email. A policy would help an institution, department or research group generate high quality data, maximize the use of its data, and gain better control over its data assets".[180]

With respect to use of shared data, researchers should endeavor to learn more about the legal and technical data environment in which their corporate partners operate to be good stewards of consumers' shared data. In a previous section of this playbook, we discussed how corporations can learn from researchers' data standards, which they must follow to make their data useful and FAIR (findable, accessible, interoperable, and reliable).[181] But, researchers should also improve their knowledge of the standards that corporations must abide by to make their data compliant with legal requirements, such as for financial model risk management, health data protection, and educational data protection (to name a few). Where researchers understand the requirements of corporate data governance, they can more carefully craft research data requests; better understand how corporate data protection and data security places limitations on what types, sizes, formats, and under which transfer protocols corporate data might be made available; and match their data use practices with that of corporate best practices. They will be supported by research administration and other support units at their institutions.

PREPARING FOR ACTION

Recommended Actions for Researchers:

Evaluate your internal policies for accountability, oversight, auditing, and system controls for use and analysis of the shared data.

Privacy and Security by Design

Organizations can follow best practices for secure data sharing by maintaining a clear line of sight towards the purpose of protecting the shared data asset.

PREPARING FOR ACTION

Recommended Actions for Researchers:

Collaborate with privacy specialists, including privacy engineers, to implement the highest levels of privacy protection for the data.

Data should be protected to safeguard individual privacy and ensure confidentiality, data integrity, restrict prohibited reuse, and ensure that data is not deleted.[182]

Breakout 11: Data Privacy Considerations

- Assessing the contents of the research data landscape to understand where you have data, what the risks are from specific research data, and the responsibilities towards a corporate data holder, its customers, and relevant regulatory authorities.
- Protecting research data using a zero trust/ least-privilege access approach to apply appropriate data privacy protection
- Responding swiftly to data subjects, including corporate customer requests, and compliance requirements by paying careful attention to and then swiftly addressing and notifying data privacy breaches; carrying out periodic data use audit reports that include reviewing subjects’ rights to withdrawal from a study; identifying suspicious behavior on behalf of any research data analysts or any data traffic through software systems; and handling any subject or regulatory requests promptly.

PREPARING FOR ACTION

Recommended Actions for Researchers:

Consider adopting a zero-trust approach to data access and use.

Breakout 12: Network Security Controls

- Access controls reduce the amount of risk by restricting permissions to only those persons who need access to the data.
- Anti-malware software is an essential component of a data protection strategy. Increasingly, research data is a source for ransomware and other malicious attacks. Preventing such intrusions is essential to mitigate against research data breach.
- Anomaly detection, particularly through use of automated, “Network anomaly detection engines (ADE)”, allow for hands-off monitoring of access or use by persons without authorizations. Establishing methods for monitoring anomalous behavior for research data access is a wise choice to researchers using sensitive shared data.
- Application security is an important component of network security, particularly when researchers hold data in cloud servers with application-based interfaces on their own devices (e.g., Google Docs or Google Sheets on mobile). This is increasingly important in a “BYOD” or bring your own device environment.
- Data loss prevention (DLP) training can help to keep data from being lost through incidents or accidents (e.g., loss of storage hardware like USBs or theft of laptops). Relatedly, research leads should ensure that they ensure effective management of personnel curiosity to prevent “honest but curious” data breach incidents.
- Researchers rely on email to perform their roles as both researchers and as instructional faculty. Email security is an essential part of good data practices and shoring up email security, such as against phishing attacks and using automated technology to identify dangerous emails can prevent the unintended re-sharing of hard-won data with malicious actors.
- Firewalls and Intrusion prevention systems (also called intrusion detection) can prevent entry into systems but can also analyze network traffic, track known attack methods, and recognize threats and respond immediately.
- Network segmentation strategies, such as establishment of trusted execution environments and analytics sandboxes, allows those with the appropriate level of access rights to perform their essential functions out of reach of others and while restricting traffic from unauthorized sources.
- Finally, and perhaps of course, ordinary web and wireless security techniques are necessary measures that researchers, like businesses, must take to ensure safe web use. This helps prevent use of browsers and wireless networks as access points to peer into data files on a network.

The National Institute of Standards and Technology (NIST) has drawn together crosswalks to specific data standards to provide more refined guidance. This includes crosswalks for cybersecurity and for privacy protection.

Figure 10: Framework Core[183]

Function	Category	ID
Identity	Asset Management	ID.AM
	Business Environment	ID.BE
	Governance	ID.GV
	Risk Assessment	ID.RA
	Risk Management Strategy	ID.RM
	Supply Chain Risk Management	ID.SC
Protect	Identity Management and Access Control	PR.AC
	Awareness and Training	PR.AT
	Data Security	PR.DS
	Information Protection and Procedures	PR.IP
	Maintenance	PR.MA
	Protective Technology	PR.PT
Detect	Anomalies and Events	DE.AE
	Security Continuous Monitoring	DE.CM
	Detection Processes	DE.DP
Respond	Response Planning	RS.RP
	Communications	RS.CO
	Analysis	RS,RP
	Mitigation	RS.MI
	Improvements	RS.IM
Recover	Recovery Planning	RC.RP
	Improvements	RC.IM
	Communications	RC.CO

Figure 11: NIST Privacy Framework[184]

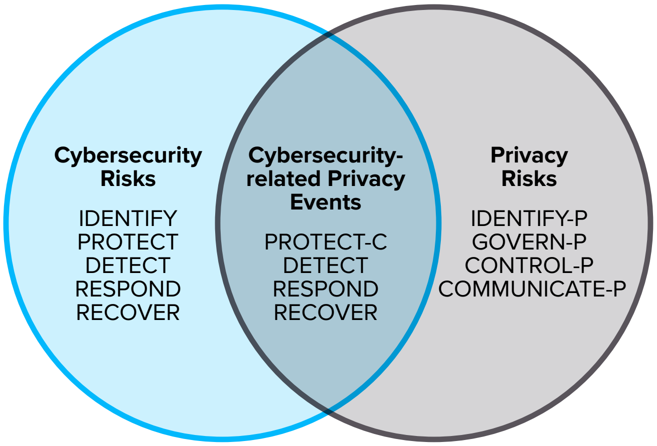


Figure 12: NIST Privacy Framework [185]

NIST Privacy Framework	
Identify-P	Develop the organizational understanding to manage privacy risk for individuals arising from data processing
Govern-P	Develop and implement the organizational governance structure to enable an ongoing understanding of the organizations risk management priorities that are informed by privacy risk
Control-P	Develop and implement appropriate activities to enable organizations or individuals to manage data with sufficient granularity to manage privacy risks
Protect-P	Develop and implement appropriate activities to enable organizations or individuals to have reliable understanding and engage in a dialogue about how data are processed and associated privacy risks
Recover	Develop and implement appropriate data processing safeguards

As these resources can help organizations properly secure and manage their own data, gaining a working understanding of the NIST data security and data privacy standards can also help researchers to understand the magnitude of their data requests and to grasp why some requests might be rejected based on application of data privacy and network security controls that have an influence. Learning more about the standards of corporate data governance can help researchers to propose data requests in the idiom of company data protections.

For researchers to increase their level of awareness of corporate data practices can be just as expensive as corporations work to structure data according to research data standards. Beyond learning the legal landscape of corporate data protection, researchers must also navigate the vocabulary (e.g., data lakes, data fabric) and platforms (e.g., AWS, Azure) of the multiple data management vendors used by corporate actors. Alas, access to some of the data management platforms or even training on those platforms may be beyond the cost point researchers and research institutions can accept. Building in costs to access or learn such systems may fall on research institutions or research funders. However, streamlining the training on tools and methods for corporate data management is one way that researchers, universities, and grant funders can create the basis of shared vocabulary essential for a collaborative data sharing environment.

ACT IV: TO BE CONTINUED

The Role of Law and Policy in Determining the Future Direction of Data Sharing for Research

The plot that will unfold in the sequel to this opera of company data sharing for research depends in part on how relevant policy evolves. In a growing number of instances, for example, the federal (and sometimes state and local) government is opening up data for researchers. At the time of this writing, data sharing by companies remains voluntary, although incentives promoting it could emerge from public policy.

Legislators and other policy makers have three general options to change the story told thus far: they could allow the status quo to continue, they could impose a mandate for data sharing, or they could create incentives and disincentives that would re-shape the current voluntary approach. When deciding which approach to take, the policy makers tackling research data sharing programs will have to creatively balance three competing priorities: protecting consumers' data from breaches or misuse by either researchers or the organizations to which legislators delegate the power to share these resources; spurring research innovation for the benefit of multiple overlapping and non-overlapping publics; and pushing companies' data practices in new directions that challenge practices of intellectual property protection, data governance, and corporate competition. Policymakers will also have to decide whether the attention currently being paid to this topic generates sufficient interest among relevant constituents to make additional legislation valuable to them during subsequent elections.

A new bill or new bills are only part of the future of high-level data sharing oversight. Because legislators are likely to determine both that no oversight of data sharing is unacceptable and that they are unable to tackle this issue wholly through legislation, they are also likely to look to federal agencies to craft regulation to govern data sharing. In the US, multiple federal agencies could become involved either as oversight agencies or as agencies crafting supporting guidance — roles some agencies already play in specific instances, as touched on above. For example, legislators could ask that anyone or all the 20 Common Rule agencies investigate how to modify the exemptions for uses of secondary data for research enumerated in the 2018 revisions to the common rule.[186]

Proposed Mandated Data Sharing

One proposal tabled during the FPF salon dinner conversations, which is intended to amplify the social benefits of data sharing, is to make sharing of data for research something that is mandatory for some or all companies. By contrast, the concepts discussed in this Playbook generally presume voluntary sharing and mutual interest among researchers and companies in the sharing of data for research. Historically, academic researchers and life sciences/biomedical companies have had a long history of voluntary and mutually beneficial data sharing for research. Most recently, the rise of governmental and other concerns about the societal effects of social media have led to broad consideration of and demand for research access to related data, demand that extends to the potential for mandatory sharing.[187]

What could mandating data sharing entail? As described throughout this playbook, a mandatory requirement that companies share data with researchers gathers many players into the data sharing story. Researchers, organizations, research institutions, research funders, publishers, repositories, research administrators, ethicists, statisticians and research methodologists, teachers, corporate CEOs, data engineers, data managers, legislators, regulators and many more personnel would engage to fulfill this mandate, which implies extensive coordination of values, purposes, platforms, policies, and vocabulary.

Based upon the lessons learned from the experts consulted for this project, we anticipate that the following challenges could arise if a mandatory data sharing program were implemented. The observations below echo the findings and recommendations presented earlier in this Playbook:

- Funding data sharing includes provisioning funds for all stages of data creation and management. This includes data transformation, encryption, transfer, security, storage, and archiving.
- Soliciting research proposals for use of shared data will include creating program descriptions that clearly describe the types of research supported, the data already available for use, data that could be made available for use, and requirements for research proposals. The solicitations would likely need to be on a web-based platform which would itself need to be built, securitized, and maintained.
- Reviewing research proposals for mandatory data sharing programs would require synthesis of the research needs that each of the relevant stakeholder groups have into a review mechanism. This will include research ethics review, research methods review, and an evaluation of the compellingness that the research has with respect to benefitting both the narrow public of organizations mandating the sharing and the wider public of constituents to each of those organizations. This will also include reviewing the qualifications of people and institutions who can perform the proposed research.
- Performing research under a mandatory data sharing program will include coordination between research institutions, researchers, and the holder of the research data. Coordination will involve ensuring that contracts and memos acknowledge the terms each institution will need to abide by, management of privacy and security assessments including checks of necessary controls, and management of legal requirements that influence each actor within their jurisdiction.
- A mandatory research data sharing will require coordination of the legal and regulatory landscape of data protection. The many laws, policies, and standards that form the patchwork of data protection in the US will need to be harmonized to the purpose of expediting shared data to researchers. As data is an international asset, international laws will also need to be accounted

for. Centralized oversight, whether in existing agencies (e.g., NSF or FTC) or in new agencies, will require funding, time, expertise, and context to meaningfully review the benefits and risks of local and hyper-local research. Decentralized oversight options, such as creating a federal statistical data center in each state or supporting decentralized review through greater funding and support to build expertise in university systems will build on existing institutions who already require considerable funding.

- Mandatory research data sharing programs will also need to be supported by actors that support research publication, such as research journals, preprint archives, and research conferences. There will need to be coordination of the many disparate requirements for data preparation, data editing, data publication, preparation of research papers, review of research data and research papers, and publication data and papers. This will require coordination of researchers' professional organizations, international publication houses, and commercial and non-profit publication venues. Researchers have already noted potential regressive impacts of publication charges associated with open publication today.[188]
- Safe Harbor protections will need to be considered for companies complying with mandatory data sharing programs as they potentially risk liability if researchers misuse their data.

The coordination of all the relevant actors to build the architecture of a mandatory research data sharing program at a national level will require coordination on specific values as well. For example, ensuring that small, medium, and massive companies can participate will ensure that researchers' needs in various disciplines are met. Research proposals from myriad institutions — universities, institutes, think-tanks — will need to be reviewed on equal footing regardless of the existing reputational or resource strength of the institution. Researchers will need to be reviewed on the merit of their ideas and not based upon brute-luck characteristics, such as gender identity or ethnic origin.[189]

Some insight might come from European consideration of mandating research access to platform data.[190] Finally, as this Playbook was being completed, the White House issued guidance calling for open publication of research funded by federal agencies, a process that—subject to further analysis and limitations—would involve publication of data. Progress on its implementation will be instructive for the broader arena of data sharing for research.[191]

CONCLUSION

Data sharing for research makes for a complicated opera, but it doesn't have to be a tragedy. Applause will come if all the players essential to make data sharing for research work well are, first and foremost, acknowledged, and second, accounted for in the orchestration of mechanisms and means that facilitate transfer of data and conduct of research.

There are careful steps that each player in this opera must take to see that data sharing moves to a next act. Some of these actions are comfortable steps taken on well-worn paths — this includes use of data governance and data standards tools already known so well to each side. Other moves are like a Verdi operatic due — maneuvers that will take years of training to perform and whose execution requires support by many coordinated experts—such as creating research specific trusted execution environments. We have compiled some of these steps into the actions listed below in Appendix One.

Whether the next act for data sharing for research is a triumph or a tragedy turns on whether the actions surfaced here are addressed. The actions we identified in this playbook and the story told here are only a start to the story. The next steps in this story are largely in the hands of the principal players in this opera — the researchers, the companies, the policymakers — whose actions or inactions shaped the past. As new players, including non-profits and civil society organizations, step onto the stage, the plot will twist, and a new act will begin.

APPENDICES

Appendix I: Preparing for Action: Summary of Recommended Actions

Recommended Actions for Organizations

1. Determine if and how data sharing for research fits the organization's values and strategy.
2. Ensure clarity of expected benefits and likely challenges involved in a potential data sharing for research partnership.
3. Determine the fitness of company data for data sharing for research.
4. Assess capacity to absorb workload and costs associated with data sharing for research partnerships.
5. Know current legal requirements regarding data sharing.
6. Ensure individuals are informed of how their data will be shared with researchers.
7. Assess the organization's capability and tech stack for secure processing and transmission of data.
8. Develop a process to communicate data sharing opportunities, review research proposals, contract with research institutions, and share data.
9. Ensure key company stakeholders (including technical, legal, and data personnel) are involved in the agreement process.
10. Ensure personnel, privacy, and cybersecurity controls are in place that are appropriate to the level of sensitivity and value of the data.
11. Consider establishing or partnering with an ethical data use committee to assess data sharing.
12. Ensure oversight and accountability mechanisms are included in data sharing agreements.
13. Ensure open lines of communication between all parties.

Recommended Actions for Research Institutions

1. Establish open lines of communication with corporate partners' legal, development, and data governance personnel to assess readiness to engage in a corporate partnership.
2. Assess cyber and data management capacity for the additional workload and costs associated with secure data sharing.
3. Develop a specific process for reviewing and approving proposals involving organization data sharing for research, including data management plans and partnership agreements.
4. Determine the role institutional review boards and ethics committees will play, if any, in review of corporate data sharing for research.

5. Ensure personnel, privacy, and cybersecurity controls are in place that are appropriate to the level of sensitivity and value of the data.
6. Ensure open lines of communication between the data sharing company and your research teams to manage expectations, responsibilities, and necessary changes to address data protection concerns.

Recommended Actions for Researchers

1. Assess available legal, financial, technical, and personnel resources to support a corporate partnership.
2. Evaluate your internal policies for accountability, oversight, auditing, and system controls for use and analysis of the shared data.
3. Collaborate with institutional cyber and physical security specialists to build a data and software securitization and management plan for the shared data and relevant analytical software used.
4. Consider adopting a zero trust approach to data access and use.
5. Collaborate with privacy specialists, including privacy engineers to implement the highest levels of privacy protection for the data.
6. Ensure open lines of communication between the data sharing company and your institution to manage expectations, responsibilities, and necessary changes to address data protection concerns.

Appendix II: Chart of Publication Requirements for Data

Journal Title and Publisher	Data Publication Statement
AI and Ethics, Springer[192]	"This journal operates a type 1 research data policy. The journal encourages authors, where possible and applicable, to deposit data that support the findings of their research in a public repository. Authors and editors who do not have a preferred repository should consult Springer Nature's list of repositories and research data policy."
Journal of Labor Economics, University of Chicago Press[193]	"It is the policy of the Journal of Labor Economics to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the Journal, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the JOLE Web site. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met."
The Lancet, Lancet Journals[194]	"Data sharing From September 21, 2020, all submitted research Articles must contain a data sharing statement, to be included at the end of the manuscript. Data sharing statements must include: • Whether data collected for the study, including individual participant data and a data dictionary defining each field in the set, will be made available to others ("undecided" is not an acceptable answer); • What data will be made available (deidentified participant data, participant data with identifiers, data dictionary, or other specified data set); • Whether additional, related documents will be available (e.g., study protocol, statistical analysis plan, informed consent form); • When these data will be available (beginning and end date, or "with publication", as applicable); • Where the data will be made available (including complete URLs or email addresses if relevant); • By what access criteria data will be shared (including with whom, for what types of analyses, by what mechanism — e.g., with or without investigator support, after approval of a proposal, with a signed data access agreement — or any additional restrictions)."
American Journal of Political Science, Wiley[195]	"The corresponding author of a manuscript that is accepted for publication in the American Journal of Political Science must provide materials that are sufficient to enable interested researchers to verify all of the analytic results that are reported in the text and supporting materials. The document titled "American Journal of Political Science Guidelines for Preparing Replication Files"* provides useful information about what information is needed and how it should be organized. All verification files must be stored in a Dataset within the AJPS Dataverse, on the Harvard Dataverse Network. Note that authors also can make their verification files available elsewhere (e.g., their personal website, other data repositories, etc.) as long as all of the necessary files are included in the Dataset on the AJPS Dataverse."
Astronomy and Astrophysics Review, Springer[196]	"All authors are requested to make sure that all data and materials as well as software application or custom code support their published claims and comply with field standards. Please note that journals may have individual policies on (sharing) research data in concordance with disciplinary norms and expectations."

ENDNOTES

[1] https://fpf.org/blog/big-data-research/

[2] https://fpf.org/blog/beyond-irbs-designing-ethical-review-processes-for-big-data-research/

[3] https://fpf.org/blog/conference-proceedings-beyond-irbs-designing-ethical-review-processes-big-data-research/

[4] https://fpf.org/blog/privacy-protective-research-facilitating-ethically-responsible-access-administrative-data/

[5] https://fpf.org/blog/dqc-report-effective-data-use-and-research-partnerships-between-seas-and-education-researchers/

[6] https://fpf.org/blog/understanding-corporate-data-sharing-decisions-practices-challenges-and-opportunities-for-sharing-corporate-data-with-researchers/; https://fpf.org/blog/fpf-companies-academics-developing-best-practices-on-data-sharing/

[7] https://fpf.org/blog/event-recap-using-corporate-data-for-research-lessons-from-an-award-winning-project/

[8] https://fpf.org/blog/fpf-publishes-report-supporting-stakeholder-engagement-and-communications-for-researchers-and-practitioners-working-to-advance-administrative-data-research/. Funding for the salon discussion series was provided by FPF's NSF Research Coordination Network, FPF's general fund and a donation made by Meta.

[9] Ronnie Littlejohn & Qingjun Li (2020) The concept of dialogue in Chinese philosophy, Educational Philosophy and Theory, DOI: 10.1080/00131857.2020.1799945; Plato, Seth Benardete, and Allan Bloom. Plato's Symposium. Chicago: University of Chicago Press, 2001. Print.

[10] A list of participating organizations is available on request.

[11] Sara Jordan, *Overcoming Hurdles to Effective Data Sharing for Researchers*, Future of Privacy Forum, (January 13, 2022), Accessed on April 4, 2022), https://fpf.org/blog/overcoming-hurdles-to-effective-data-sharing-for-researchers/

[12] C-SPAN. Senate Commerce Subcommittee on Consumer Protection hearing [Video] (Oct. 5, 2021), *available at* https://www.c-span.org/video/?515042-1/whistleblower-frances-haugen-calls-congress-regulate-facebook.

[13] *See* Alex Abdo, Ramya Krishnan, Stephanie Krent, Evan Welber Falcón, & Andrew Keane Woods, *A Safe Harbor for Platform Research*, Knight First Amendment Institute (Jan. 19, 2022), https://knightcolumbia.org/content/a-safe-harbor-for-platform-research; Gilad Edelman, *Facebook's Reason for Banning Researchers Doesn't Hold Up*, Wired (Aug. 4, 2021), https://www.wired.com/story/facebooks-reason-banning-researchers-doesnt-hold-up/; Barbara Ortutay, *Facebook Shuts Out NYU Academics' Research on Political Ads*, APNews (Aug. 4, 2021), https://apnews.com/article/technology-business-5d3021ed9f193bf249c3af158b128d18; Shirin Ghaffary, *"People do not trust that Facebook is a healthy ecosystem"; Leading Social Media Researcher Laura Edelson Explains Her Misinformation Fight with Facebook*, Recode (Aug. 6, 2021), https://www.vox.com/recode/22612151/laura-edelson-facebook-nyu-ad-observatory-social-media-researcher.

[14] Future of Privacy Forum, *FPF Issues Award for Research Data Stewardship to Stanford Medicine & Empatica, Google* (Jun. 28, 2021), https://fpf.org/press-releases/fpf-issues-2021-award-for-research-data-stewardship/.

[15] Robert C. Weber, *A Letter to Our Clients About Government Access to Data*, IBM (Mar. 14, 2014), https://www.ibm.com/blogs/think/2014/03/open-letter-data/; Mercy Corps AgriFin, *Optimizing Digital Data Sharing in Agriculture* (Dec. 9, 2020), https://www.mercycorpsagrifin.org/2020/12/09/optimizing-digital-data-sharing-in-agriculture/.

[16] Margaret C. Levenstein, *Brief Biography*, Inter-university Consortium for Political and Social Research, Accessed on Mar. 29, 2022, https://www.icpsr.umich.edu/web/pages/about/staff-profile.html?node=1719.

[17] Lars Vilhuber, *People at ILR*, ILR School, Accessed on Mar. 29, 2022, https://www.ilr.cornell.edu/people/lars-vilhuber.

[18] Richard Tsai, *About Richard Tsai*, Inspire, Accessed on Mar. 29, 2022, https://corp.inspire.com/author/richard/.

[19] Mary Potter*, Mary Potter*, Research and Innovation, Virginia Tech, Accessed on Mar. 29, 2022, https://www.research.vt.edu/sirc/prdp/contacts/mary-potter.html.

[20] Claire Cravero, *DATAVANT Team*, Datavant, Accessed on Mar. 29, 2022, https://datavant.com/about/team/#.

[21] David Peloquin, *David Peloquin*, Ropes & Gray LLP, Accessed on Mar. 29, 2022, https://www.ropesgray.com/en/biographies/p/david-peloquin.

[22] Brian King, *Brian King*, LinkedIn, Accessed on Mar. 29, 2022, https://www.linkedin.com/in/brian-king-1b86434/?originalSubdomain=co.

[23] Casey Greene, *Casey Greene*, The Greene Lab, Accessed on Mar. 29, 2022, https://greenelab.com/members/casey-greene.html; Libby Hemphill, *Libby Hemphill*, School of Information University of Michigan, Accessed on Mar. 29, 2022, https://www.si.umich.edu/people/libby-hemphill; Emmanuel Makau, *The Team*, Mercy Corps AgriFin, Accessed on Mar. 29, 2022, https://www.mercycorpsagrifin.org/emmanuel-makau/.

[24] The White House, *Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking,* (Jan. 27, 2021), https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/memorandum-on-restoring-trust-in-government-through-scientific-integrity-and-evidence-based-policymaking/; U.S. Department of State, *PEPFAR Data Governance* (Aug. 9, 2017), https://learn.pepfar.net/assets/courseware/v1/e09bc70ae756515f14e096aa033628f2/asset-v1:learn-pepfar-net+PMDATVW_1+2018_9+type@asset+block/20170809_PEPFAR_Data_Governance.pdf; PEPFAR, *PEPFAR Panorama Spotlight,* Accessed on Mar. 29, 2022, https://data.pepfar.gov/.

[25] European Commission (25 November 2020). Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) — COM/2020/767 final. Brussels, Belgium: European Commission. Retrieved 2021-07-01. Document 52020PC0767

[26] Janet Box-Steffensmeier, Jean Burgess, Maurizio Corbetta, Kate Crawford, Esther Duflo, Laurel Fogarty, Alison Gopnik, Sari Hanafi, Mario Herrero, Ying-yi Hong, Yasuko Kameyama, Tatia M. C. Lee, Gabriel M. Leung, Daniel S. Nagin, Anna C. Nobre, Merete Nordentoft, Aysu Okbay, Andrew Perfors, Laura M. Rival, Cassidy R. Sugimoto, Bertil Tungodden, and Claudia Wagner, *The Future of Human Behaviour Research*, Nature Human Behaviour (2022), 6 (1): 15-24.

[27] Alfred P. Sloan Foundation, *Data and Computational Research*, Accessed on March 29, 2022, *https://sloan.org/programs/digital-technology/data-and-computational-research*

[28] Re3data, *Registry of Research Data Repositories*, Accessed on March 29, 2022, https://www.re3data.org/; Kelsey Finch, *FPF Best Practices and Contract Guidelines Help Companies Share Data with Academic Researchers* (Oct. 28, 2020), https://fpf.org/blog/fpf-best-practices-and-contract-guidelines-help-companies-share-data-with-academic-researchers.

[29] Michigan State University, *How to Find Data & Statistics: Finding Data*, Accessed on March 29, 2022, https://libguides.lib.msu.edu/c.php?g=96631&p=626754.

[30] *Office of the AEA Data Editor*, American Economic Association, https://aeadataeditor.github.io/; Adam Marcus, *Which Takes Longer to Produce: An Infant Who Can Sit on His Own, or a Retraction?*, Retraction Watch (Feb. 23, 2022), https://retractionwatch.com/2022/02/23/which-takes-longer-to-produce-an-infant-who-can-sit-on-his-own-or-a-retraction/; Natalia Mesa, *Q&A: A Randomized Approach to Awarding Grants*, The Scientist (February 25, 2022), https://www.the-scientist.com/news-opinion/q-a-a-randomized-approach-to-awarding-grants-69741.

[31] European Commission, *Data Act: Commission Proposes Measures for a Fair and Innovative Data Economy* (Feb. 23, 2022), https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113.

[32] Mercy Corps AgriFin, Optimizing Digital Data Sharing in Agriculture, (December 9, 2020), Accessed on March 29, 2022, https://www.mercycorpsagrifin.org/2020/12/09/optimizing-digital-data-sharing-in-agriculture/.

[33] Lioness, *Sex Research Platform*, Accessed on Apr. 1, 2022, https://lioness.io/pages/lioness-research-platform.

[34] Datavant, *Research*, Accessed on April 4, 2022, https://datavant.com/resources/research/

[35] Inter-university Consortium for Political and Social Research, *ICPSR*, University of Michigan, Accessed on March 29, 2022, https://www.icpsr.umich.edu/web/pages/.

[36] Consortium of European Social Science Data Archives, *About*, Accessed on April 4, 2022, https://www.cessda.eu/About.

[37] Figshare, *Store, Share, Discover Research*, Accessed on April 4, 2022, https://figshare.com/.

[38] Future of Privacy Forum, FPF Issues Award for Research Data Stewardship to Stanford Medicine & Empatica, Google (Jun. 28, 2021), https://fpf.org/press-releases/fpf-issues-2021-award-for-research-data-stewardship/.

[39] Consortium of European Social Science Data Archives, *CESSDA Controlled Vocabulary for CESSDA Topic Classification* (Feb. 2, 2022), https://vocabularies.cessda.eu/vocabulary/TopicClassification.

[40] Kasantha Moodley and George Wyeth, *Citizen Science Programs at Environmental Agencies: Case Studies*, Environmental Law Institute, (Oct. 2020), Accessed on Mar. 29, 2022, https://www.eli.org/research-report/citizen-science-programs-environmental-agencies-case-studies.

[41] Organized Crime and Corruption Reporting Project, *Catalogue of Research Databases*, Accessed on Mar. 29, 2022, https://id.occrp.org/databases/; Lawrence Leung, *Validity, Reliability, and Generalizability in Qualitative Research,* Journal of Family Medicine and Primary Care (2015), 4 (3) 324-327, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4535087/; U.S. Department of Health & Human Services, *Definitions*, Accessed on Mar. 29, 2022, https://ori.hhs.gov/content/chapter-3-The-Protection-of-Human-Subjects-Definitions.

[42] David Mandel & Philip Tetlock, Debunking the Myth of Value-Neutral Virginity: Toward Truth in Scientific Advertising, Frontiers in Psychology (Mar. 30, 2016), https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00451/full; 500 Women Scientists Leadership, *Silence is Never Neutral; Neither is Science*, Scientific American (Jun. 6, 2020), https://blogs.scientificamerican.com/voices/silence-is-never-neutral-neither-is-science/; Nick Howe, *Stick to the Science: When Science Gets Political*, Nature Podcast, (Nov. 3, 2020), *available at* https://www.nature.com/articles/d41586-020-03067-w; Steven Rose and Hilary Rose, *Can Science Be Neutral*, Perspectives in Biology and Medicine (1973), 16 (4): 605-624.

[43] Danielle Douglas-Gabriel, *College Endowments Aren't Piggy Banks. But Some Experts Say Wealthy Schools Could Spend More*, Washington Post (Feb. 19, 2022), https://www.washingtonpost.com/education/2022/02/19/wealthy-university-endowments/; Emma Whitford, *College Endowments Boomed in Fiscal 2021*, Inside Higher Ed (Feb. 18, 2022), https://www.insidehighered.com/news/2022/02/18/college-endowments-boomed-fiscal-year-2021-study-shows; Emma Whitford, *Divestment Gap Emerges,* Inside Higher Ed (Apr. 28, 2021), https://www.insidehighered.com/news/2021/04/28/divestment-gains-some-colleges-can-it-spread-where-oil-rules.

[44] Jesse Saffron & Stephanie Keaveney, *The Higher Education Establishment's Self-Interest Goes Unchecked—Again*, James G. Martin Center for Academic Renewal (Dec. 12, 2016), https://www.jamesgmartin.center/2016/12/higher-education-establishments-self-interest-goes-unchecked/; ERM Initiative Faculty & Erika Baker, *Are Universities & Colleges Doing Enough to Manage Reputational Risk?,* North Carolina State University (Feb. 21, 2019), https://erm.ncsu.edu/library/article/are-universities-colleges-doing-enough-to-manage-reputational-risk.

[45] U.S. Department of Health & Human Services, *Federalwide Assurance Instructions* (Jul. 31, 2017), https://www.hhs.gov/ohrp/register-irbs-and-obtain-fwas/forms/fwa-instructions/index.html; *see also* Sarah Coble, *Data Breach at University of Kentucky,* InfoSecurity Magazine (Aug. 6, 2021), https://www.infosecurity-magazine.com/news/data-breach-at-university-of/.

[46] U.S. Department of Health and Human Services. "Code of Federal Regulations - Title 45 Public Welfare CFR 46". https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

[47] Kevin Munger, Mario Luca, Jonathan Nagler, & Joshua Tucker, *The (Null) Effects of Clickbait Headlines on Polarization, Trust, and Learning*, Public Opinion Quarterly (2020), 84 (1): 49-73; Canyu Zhang, and Paul D. Clough, *Investigating Clickbait in Chinese Social Media: A Study of WeChat*, Online Social Networks and Media (2020) 19: 100095.

[48] Lee Rainie, *Americans' Complicated Feelings About Social Media in an Era of Privacy Concerns,* Pew Research Center (Mar. 27, 2018), https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/; Sarah, Gilbert, Jessica Vitak, and Katie Shilton, *Measuring Americans' Comfort with Research Uses of Their Social Media Data*, Social Media + Society (2021), 7, (3): 1-13.

[49] Timothy Morey, Theodore Forbath, and Allison Schoop, *Customer Data: Designing for Transparency and Trust*, Harvard Business Review (2015), 93(5): 96-105.

[50] *See* Juntae DeLane, *Consumer Opinion On Data Collection And Personalization Is Mixed: What Should Marketers Do?*, Digital Branding Institute, Accessed on Mar. 29, 2022, https://digitalbrandinginstitute.com/consumers-data-collection/.

[51] Carey Funk, *Key Findings About Americans' Confidence in Science and Their Views on Scientists' Role in Society*, Pew Research Center (Feb. 12, 2020), https://www.pewresearch.org/fact-tank/2020/02/12/key-findings-about-americans-confidence-in-science-and-their-views-on-scientists-role-in-society/; Paul Appelbaum, Milena Anatchkova, Karen Albert, Laura Dunn, and Charles Lidz, *Therapeutic Misconception in Research Subjects: Development and Validation of a Measure, Clinical Trials* (2012), 9(6): 748-761.

[52] Adrian Rauchfleisch, Mike Schäfer, and Dario Siegen, *Beyond the Ivory Tower: Measuring and Explaining Academic Engagement with Journalists, Politicians and Industry Representatives Among Swiss Professors*, PLoS ONE (2021), 16(5): e0251051; Stefan de Jong, Elena Ketting, and Leonie van Drooge, *Highly Esteemed Science: An Analysis of Attitudes towards and Perceived Attributes of Science in Letters to the Editor in Two Dutch Newspapers*, Public Understanding of Science (2020), 29 (1): 37—52.

[53] Mirjam Annina Jenny, Niklas Keller, and Gerd Gigerenzer, *Assessing Minimal Medical Statistical Literacy Using the Quick Risk Test: A Prospective Observational Study in Germany,* BMJ Open (2018), 8 (8): e020847; Rainer Bromme and Eva Thomm, *Knowing Who Knows: Laypersons' Capabilities to Judge Experts' Pertinence for Science Topics*, Cognitive Science (2016), 40: 241-252.

[54] The role of data intermediary was codified in the recent EU Data Governance Act (DGA), which will also affect research data sharing strategies. While the GDPR is well known as a standard with de facto global force, the DGA is a new act. "The DGA proposal covers three key areas: (1) access to data held by public sector bodies; (2) regulation of data sharing services through "data intermediaries"; and (3) encouraging "data altruism," which means donating data for the common good, such as health care research." Sandeep Sangwan, *How To Know You Are a 'Data Intermediary' Under the Data Governance Act*, International Association of Privacy Professionals, (April 27, 2021), Accessed on April 4, 2022, https://iapp.org/news/a/how-to-know-you-are-a-data-intermediary-under-the-data-governance-act

[55] See: https://digital-strategy.ec.europa.eu/en/policies/european-digital-media-observatory

[56] European Digital Media Observatory and George Washington University Institute for Data, Democracy, and Politics, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher-Data Access, 31 May 2022. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

[57] *See, e.g.,* Microsoft, *Confidential Computing*, Accessed on Mar. 30, 2022, https://www.microsoft.com/en-us/research/theme/confidential-computing/; Google, *Confidential Computing*, Accessed on Mar. 30, 2022, https://cloud.google.com/confidential-computing; IBM, *Confidential Computing on IBM Cloud*, Accessed on Mar. 30, 2022, https://www.ibm.com/cloud/confidential-computing.

[58] Issie Lapowsky, *Platforms vs. PhDs: How Tech Giants Court and Crush the People Who Study Them*, Protocol, (Mar. 19, 2021), https://www.protocol.com/nyu-facebook-researchers-scraping; Shannon Bond, *NYU Researchers Were Studying Disinformation On Facebook. The Company Cut Them Off,* NPR (Aug. 4, 2021), https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f.

[59] PC, *Data Sharing*, Accessed on Mar. 30, 2022, https://www.pcmag.com/encyclopedia/term/data-sharing.

[60] Institute of Medicine, *Data Sharing Elements and Activities. Discussion Framework for Clinical Trial Data Sharing: Guiding Principles, Elements, and Activities*, (2014), Washington (DC): National Academies Press.

[61] Sara Jordan, *Data Sharing…By Any Other Name*, Future of Privacy Forum (Nov. 4, 2021), https://fpf.org/blog/data-sharing-by-any-other-name/.

[62] Harris Cooper, Larry Hedges, and Jeffrey Valentine, *The Handbook of Research Synthesis and Meta-Analysis*, (2019), Russell Sage Foundation; Sally Aboelela, Elaine Larson, Suzanne Bakken, Olveen Carrasquillo, Allan Formicola, Sherry Glied, Janet Haas, and Kristine Gebbie, *Defining Interdisciplinary Research: Conclusions from a Critical Review of the Literature*, Health Services Research, (2007), 42(1p1): 329-346; Delbert Miller, and Neil Salkind, *Handbook of Research Design and Social Measurement*, Sage, (2002); Ian Hacking, *Imre Lakatos's Philosophy of Science*, The British Journal for the Philosophy of Science, (1979), 30(4): 381-402; Jacob Tebes, *Community Science, Philosophy of Science, and the Practice of Research*, American Journal of Community Psychology, (2005), 35(3-4): 213-230; Jan Remme, Adam Taghreed, Francisco Becerra-Posada, Catherine D'Arcangues, Michael Devlin, Charles Gardner, Abdul Ghaffar, Joachim Hombach, Jane Kengeya, Anthony Mbewu, Michael Mbizvo, Zafar Mirza, Tikki Pang, Robert Ridley, Fabio Zicker, and Robert Terry, *Defining Research to Improve Health Systems*, PLoS Medicine, (2010), 7(11): e1001000.

[63] Office of the Federal Register, National Archives and Records Administration, *Protection of Human Subjects,* govinfo, (September 30, 2000), Accessed on Mar. 31, 2022, https://www.govinfo.gov/app/details/CFR-2000-title45-vol1/CFR-2000-title45-vol1-part46.

[64] U. S. Department of Defense, *Financial Management Regulation*, DoD 7000-R, (June 2004), Vol 2B, Ch 5: 5-2, Accessed on Mar. 31, 2022, https://comptroller.defense.gov/Portals/45/documents/fmr/archive/02barch/02b_05old.pdf

[65] National Science Foundation, *Definitions of Research and Development: An Annotated Compilation of Official Sources*, (Mar. 22, 2018), Accessed on Mar. 31, 2022, https://www.nsf.gov/statistics/randdef/.

[66] 45 CFR 46, *available at* Office of the Federal Register, National Archives and Records Administration, *Protection of Human Subjects,* govinfo, (Sept. 30, 2000), Accessed on Mar. 31, 2022, https://www.govinfo.gov/app/details/CFR-2000-title45-vol1/CFR-2000-title45-vol1-part46.

[67] 32 CFR 272.3, *available at* Office of the Federal Register, National Archives and Records Administration, *Definition of Basic Research*, govinfo, (Jul. 1, 2012), Accessed on Mar. 31, 2022, https://www.govinfo.gov/app/details/CFR-2012-title32-vol2/CFR-2012-title32-vol2-sec272-3.

[68] National Science Foundation, *Definitions of Research and Development: An Annotated Compilation of Official Sources* (Mar. 22, 2018), https://www.nsf.gov/statistics/randdef/.

[69] 32 CFR 37.1220, *available at* U. S. Department of Defense, *Financial Management Regulation*, DoD 7000-R, (Jun. 2004), Vol 2B, Ch 5: 5-2, Accessed on Mar. 31, 2022, https://comptroller.defense.gov/Portals/45/documents/fmr/archive/02barch/02b_05old.pdf

[70] Helen Xun, *Waverly He, Jonlin Che*n, Scott Sylvester, Sheera Lerman, and Julie Caffrey, *Characterization and Comparison of the Utilization of Facebook Groups Between Public Medical Professionals and Technical Communities to Facilitate Idea Sharing and Crowdsourcing During the COVID-19 Pandemic: Cross-sectional Observational Study*, JMIR Formative Research (2021), 5(4): e22983.

[71] Neelima Bhatnagar and Michael Pry, *Student Attitudes, Awareness, and Perceptions of Personal Privacy and Cybersecurity in the Use of Social Media: An Initial Study*, Information Systems Education Journal, (2020) 18(1): 48-58. https://files.eric.ed.gov/fulltext/EJ1246231.pdf.

[72] Kenneth Olmstead & Aaron Smith, *Americans and Cybersecurity*, Pew Research Center (Jan. 26, 2017), https://www.pewresearch.org/internet/2017/01/26/americans-and-cybersecurity/; Berkeley Center for Law & Technology, UC Berkeley School of Information & the International Computer Science Institute, Report of a Workshop, Cybersecurity Research: Addressing the Legal Barriers and Disincentives (Sept. 28, 2015), https://www.ischool.berkeley.edu/sites/default/files/cybersec-research-nsf-workshop.pdf.

[73] Institute of Medicine, *Sharing Clinical Research Data: Workshop Summary* (2013), Washington DC: The National Academies Press, *available at* https://nap.nationalacademies.org/download/18267.

[74] Janet Box-Steffensmeier, Jean Burgess, Maurizio Corbetta, Kate Crawford, Esther Duflo, Laurel Fogarty, Alison Gopnik, Sari Hanafi, Mario Herrero, Ying-yi Hong, Yasuko Kameyama, Tatia M. C. Lee, Gabriel M. Leung, Daniel S. Nagin, Anna C. Nobre, Merete Nordentoft, Aysu Okbay, Andrew Perfors, Laura M. Rival, Cassidy R. Sugimoto, Bertil Tungodden, and Claudia Wagner, *The Future of Human Behaviour Research*, Nature Human Behaviour (2022), 6 (1): 15-24.

[75] Ivan Oransky, Weekend Reads: '*Published Crap;' Randomized Grant Awards; 'Problems in Science Publishing'*, Retraction Watch, (Feb. 26, 2022), Accessed on Mar. 31, 2022, https://retractionwatch.com/2022/02/26/weekend-reads-published-crap-randomized-grant-awards-problems-in-science-publishing/#more-124323.

[76] Carey Funk, *What the Public Really Thinks About Scientists*, American Scientist, (2021), Accessed on Mar. 29, 2022, https://www.americanscientist.org/article/what-the-public-really-thinks-about-scientists.

[77] Institute of Medicine, *Sharing Clinical Research Data: Workshop Summary*, (2013), Washington DC: The National Academies Press. https://nap.nationalacademies.org/download/18267.

[78] Janet Box-Steffensmeier, Jean Burgess, Maurizio Corbetta, Kate Crawford, Esther Duflo, Laurel Fogarty, Alison Gopnik, Sari Hanafi, Mario Herrero, Ying-yi Hong, Yasuko Kameyama, Tatia M. C. Lee, Gabriel M. Leung, Daniel S. Nagin, Anna C. Nobre, Merete Nordentoft, Aysu Okbay, Andrew Perfors, Laura M. Rival, Cassidy R. Sugimoto, Bertil Tungodden, and Claudia Wagner, *The Future of Human Behaviour Research*, Nature Human Behaviour (2022), 6 (1): 15-24.

[79] Stephan Dombroski, Matthew McDonald, Marjon Van Der Pol, Mark Grindle, Alison Avenell, Paula Carroll, Eileen Calveley, Andrews Elders, Nicola Glennie, Cindy Gray, Fiona Harris, Adrian Hapca, Claire Jones, Frank Kee, Michelle McKinley, Rebecca Skinner, Martin Tod, and Pat Hoddinott, *Games of Stones: Feasibility Randomised Controlled Trial of How to Engage Men with Obesity in Text Message and Incentive Interventions for Weight Loss*, BMJ Open (2020), 10:e032653, *available at* https://bmjopen.bmj.com/content/10/2/e032653.

[80] *See, e.g.,* Cybersecurity and Infrastructure Security Agency, *CYBER INFORMATION SHARING AND COLLABORATION PROGRAM (CISCP)*, Accessed Mar. 31, 2022, https://www.cisa.gov/ciscp; Kathryn Maxson Jones, Rachel Ankeny, and Robert Cook-Deegan, *The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project,* Journal of the History of Biology (2018), 51: 693-805, *available at* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7307446/.

[81] Google, *COVID-19 Community Mobility Reports*, Accessed on Mar. 31, 2022, https://www.google.com/covid19/mobility/; *see also* Future of Privacy Forum, *FPF Issues Award for Research Data Stewardship to Stanford Medicine & Empatica, Google*, (Jun. 28, 2021), Accessed on Mar. 21, 2022, https://fpf.org/press-releases/fpf-issues-2021-award-for-research-data-stewardship/.

[82] U. S. National Library of Medicine, National Institutes of Health, *PMC Full-Text Search Results*, Accessed on Mar. 31, 2022, https://www.ncbi.nlm.nih.gov/pmc/?term=Google+%22Community+Mobility+Reports%22

[83] The Future of Privacy Forum, with the support of the Alfred P. Sloan Foundation, established an Award for Research Data Stewardship to recognize corporate-researcher partnerships for their privacy commitments and best practices to corporate data sharing for research purposes. In 2021, the award recognized two corporate-researcher teams and their efforts during the COVID-19 global pandemic. *See* https://fpf.org/press-releases/fpf-issues-2021-award-for-research-data-stewardship/.

[84] Sabyasachi Dash, Sushil Kumar, Mohit Sharma and Sandeep Kaushik, *Big Data in Healthcare: Management, Analysis and Future Prospects*, Journal of Big Data (2019), 6(54), *available at* https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0.

[85] Lisa Harlow and Frederick Oswald, *Big Data in Psychology: Introduction to the Special Issue*, Psychological Methods (2016), 21(4): 447-457.

[86] Terrence Sejnowski, Patricia Churchland, and J. Anthony Movshon, *Putting Big Data to Good Use in Neuroscience*, Nature Neuroscience (2014), 17:1440-1441, *available at* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224030/.

[87] Sunny Jung Kim, Lisa Marsch, Jeffrey Hancock and Amarendra Das, *Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data,* Journal of Medical Internet Research, (2017), 19(10): e353, *available at* https://pubmed.ncbi.nlm.nih.gov/29089287/.

[88] European Consortium for Political Research, *Political Sciences and the Big Data Challenge. From Big Data in Politics to the Politics of Big Data*, (2017), Accessed Mar. 31, 2022, https://ecpr.eu/Events/Event/SectionDetails/640.

[89] Heidi Ledford, *How Facebook, Twitter and Other Data Troves are Revolutionizing Social Science*, Nature, (Jun. 17, 2020), Accessed on Mar. 31, 2022, *available at* https://www.nature.com/articles/d41586-020-01747-1.

[90] Abby Benson, Jodi Hubble, and Kristen Freaney, *The Benefits of Building University Corporate Partnerships*, Academic Futures White Paper, Accessed Mar. 31, 2022, https://www.colorado.edu/academicfutures/sites/default/files/attached-files/benson_et_al.pdf.

[91] Christine Musselin, *New Forms of Competition in Higher Education*, Socio-Economic Review, (2018), 16(3): 657-683, https://academic.oup.com/ser/article/16/3/657/5067568.

[92] Lars Frølund, Fiona Murray, and Max Riedel, *Developing Successful Strategic Partnerships with Universities*, MITSloan Management Review, (2017), Accessed on Mar. 31, 2022, https://sloanreview.mit.edu/article/developing-successful-strategic-partnerships-with-universities/.

[93] Igor Tulchinsky and Robert Kirkpatrick, *The Power of Data Philanthropy*, Milken Institute (May 10, 2019), https://milkeninstitute.org/article/power-data-philanthropy.

[94] Marc Andreessen, *Why Software is Eating the World*, The Wall Street Journal (Aug. 20, 2011), https://www.wsj.com/articles/SB10001424053111903480904576512250915629460.

[95] Sean Illing, *A Political Scientist Explains How Big Data is Transforming Politics*, Vox (Mar. 16, 2017), https://www.vox.com/conversations/2017/3/16/14935336/big-data-politics-donald-trump-2016-elections-polarization.

[96] Kevin Rands, *How Big Data Has Changed Politics*, CIO, (Jun. 28, 2018), Accessed Mar. 31, 2022, https://www.cio.com/article/221882/how-big-data-has-changed-politics.html.

[97] ERM Initiative Faculty and Erika Baker, *Are Universities & Colleges Doing Enough to Manage Reputational Risk?,* North Carolina State University, (Feb. 21, 2019), Accessed on Mar. 29, 2022, https://erm.ncsu.edu/library/article/are-universities-colleges-doing-enough-to-manage-reputational-risk.

[98] Amy Westervelt, *Revealed: Leading Climate Research Publisher Helps Fuel Oil and Gas Drilling*, The Guardian (Feb. 24, 2022), https://www.theguardian.com/environment/2022/feb/24/elsevier-publishing-climate-science-fossil-fuels.

[99] Alice Fabbri, Alexandra Lai, Quinn Grundy, and Lisa Anne Bero, *The Influence of Industry Sponsorship on the Research Agenda: A Scopy Review,* American Journal of Public Health (2018), 108(11): e9-e16, *available at* https://pubmed.ncbi.nlm.nih.gov/30252531/; Tess Legg, Jenny Hatchard, and Anna Gilmore, *The Science for Profit Model-How and Why Corporations Influence Science and the Use of Science in Policy and Practice*, PLoS ONE (2021), 16(6): e0253272, *available at* https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0253272; Joel Lexchin, Lisa Bero, Courtney Davis, and Marc-Andre Gagnon, *Achieving Greater Independence from Commercial Influence in Research*, BMJ, (2021), 372: n370, *available at* https://pubmed.ncbi.nlm.nih.gov/33687982/.

100 Casey Greene, Lana Garmire, Jack Gilbert, Marylyn Ritchie, and Lawrence Hunter, *Celebrating Parasites*, Nature Genetics, (2017), 49(4): 483-484, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5710834/; Dan Longo and Jeffrey Drazen, *Data Sharing*, The New England Journal of Medicine, (2016), 374: 276-277, *available at* https://www.nejm.org/doi/full/10.1056/nejme1516564.

101 Tiffany Hsu and Gillian Friedman, *CVS, Dunkin', Lego: The Brands Pulling Ads From Facebook Over Hate Speech*, The New York Times (Jul. 7, 2020), https://www.nytimes.com/2020/06/26/business/media/Facebook-advertising-boycott.html; Kim Lyons, *Coca-Cola, Microsoft, Starbucks, Target, Unilever, Verizon: All the Companies Pulling Ads from Facebook*, The Verge (Jul. 2, 2020), https://www.theverge.com/21307454/unilever-verizon-coca-cola-starbucks-microsoft-ads-facebook.

102 TechRepublic Staff, *Facebook Data Privacy Scandal: A Cheat Sheet,* TechRepublic, (Jul. 30, 2020), Accessed on Mar. 31, 2022, https://www.techrepublic.com/article/facebook-data-privacy-scandal-a-cheat-sheet/; Altexsoft, *Preparing Your Dataset for Machine Learning: 10 Basic Techniques That Make Your Data Better* (Mar. 19, 2021), https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/.

103 Legg, T., Hatchard, J., & Gilmore, A. B. (2021). The science for profit Model—How and why corporations influence science and the use of science in policy and practice. PloS One, 16(6), e0253272-e0253272. https://doi.org/10.1371/journal.pone.0253272

104 Kenneth Olmstead and Aaron Smith, Americans and Cybersecurity, Pew Research Center (Jan. 26, 2017), https://www.pewresearch.org/internet/2017/01/26/americans-and-cybersecurity/.

105 Alexandra Levine, *Suicide Hotline Shares Data with for-Profit Spinoff, Raising Ethical Questions*, Politico (Jan. 28, 2022), https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617.

106 Alexandra Levine, *Suicide Hotline Shares Data with for-Profit Spinoff, Raising Ethical Questions*, Politico (Jan. 28, 2022), https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617.

107 Aaron Carroll, *Why a Lot of Important Research Is Not Being Done*, The New York Times (Dec. 4, 2017), https://www.nytimes.com/2017/12/04/upshot/health-research-lawsuits-chilling-effect.html.

108 The University of Chicago, *Data-Sharing Agreements* (Apr. 1, 2011), https://ura.uchicago.edu/page/data-sharing-agreements.

109 Anne Diekema, Andrew Wesolek, and Cheryl Walters, *The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories*, The Journal of Academic Librarianship, (2014), 40(3-4): 322-331.

110 Netanel Weinstock, *Making Sense of Fair Use*, Lewis & Clark Law Review (2011) 15:715-771. https://escholarship.org/content/qt5mh7w8hc/qt5mh7w8hc.pdf; Lee Bygrave, *Data Protection Law, Approaching its Rationale, Logic and Limits, (Vol 10)*, (2002), Information Law Series), The Hague: Kluwer Law International.

111 European Chemicals Agency, *Typical Cost Elements in Data Sharing*, (May 2017), Accessed on Mar. 31, 2022, https://www.echa.europa.eu/documents/10162/17223/factsheet_costs_datasharing_en.pdf/c4595798-0634-4f3b-a247-d518b999ba1f; Sara Shaw, Van-Kim Lin, and Kelly Maxwell, *Guidelines for Developing Data Sharing Agreements to Use State Administrative Data for Early Care and Education Research*, (2018), OPRE Research Brief #2018-67, U.S. Department of Health and Human Services, Washington D.C., https://www.childtrends.org/wp-content/uploads/2018/09/data-sharing-agreements_Child-Trends_June-2018.pdf.

112 Paige Backlund Jarquin, *Data Sharing: Creating Agreements*, Colorado Clinical and Translational Sciences Institute & Rocky Mountain Prevention Research Center, Accessed on Mar. 31, 2022, http://trailhead.institute/wp-content/uploads/2017/04/tips_for_creating_data_sharing_agreements_for_partnerships.pdf.

113 Andrew Joss, *Value of an Enterprise Intelligent Data Governance Framework*, Informatica (Nov. 28, 2017), https://www.informatica.com/blogs/the-value-of-enterprise-intelligent-data-governance-framework.html.

114 *See* Cybersecurity and Infrastructure Security Agency, *TRAFFIC LIGHT PROTOCOL, (TLP) DEFINITIONS AND USAGE*, Accessed Mar. 31, 2022, https://www.cisa.gov/tlp.

115 *See* Srishti Deoras, *Cambridge Analytica Controversy: A Timeline of Events,* Analytics India Mag Careers, (Mar. 26, 2018), https://analyticsindiamag.com/cambridge-analytica-controversy-a-timeline-of-events/; Sam Meredith, *Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal*, CNBC, (Apr. 10, 2018), https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html; Sam Meredith, *Here's Everything You Need to Know About the Cambridge Analytica Scandal*, CNBC, (Mar. 21, 2018), https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html; Harry Davies, *Ted Cruz Using Firm That Harvested Data on Millions of Unwitting Facebook Users*, The Guardian, (Dec. 11, 2015), https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data; Nicholas Confessore, *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*, The New York Times, (Apr. 4, 2018), https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

116 Sydney Johnson, *Chief Privacy Officers: A Small But Growing Fleet in Higher Education*, EdSurge, (Mar. 25, 2019), Accessed on Mar. 31, 2022, https://www.edsurge.com/news/2019-03-25-chief-privacy-officers-a-small-but-growing-fleet-in-higher-education.

117 *See* Sue Poremba, *COVID-19 Leads to Greater Consumer Awareness of Data Security*, Security Boulevard (Sept. 21, 2020), https://securityboulevard.com/2020/09/covid-19-leads-to-greater-consumer-awareness-of-data-security/; Federal Trade Commission, *Data Breach Response: A Guide for Business*, Accessed on Apr. 1, 2022, https://www.ftc.gov/business-guidance/resources/data-breach-response-guide-business; Jian Ming Colin Wee, Masooda Bashir, and Nasir Memon, *Self-Efficacy in Cybersecurity Tasks and Its Relationship with Cybersecurity Competition and Work-Related Outcomes* [Conference Presentation], USENIX Workshop on Advances in Security Education 2016 (Aug. 9, 2016), https://www.usenix.org/conference/ase16/workshop-program/presentation/wee.

118 Editor. 2019. Personal Data and the Organization: Stewardship and Strategy, The Future of Privacy Forum. https://fpf.org/blog/personal-data-and-the-organization-stewardship-and-strategy/

119 Gray, S. 2020. Understanding the World Geolocation Data, The Future of Privacy Forum. https://fpf.org/blog/understanding-the-world-of-geolocation-data/

120 Sharifi, Siegl & Vance. 2021. Understanding Student Monitoring. Student Privacy Compass, The Future Privacy Forum. https://studentprivacycompass.org/resource/understanding-student-monitoring/

121 Herein, we will use the acronym "IRBs" to collectively refer to all similar boards regardless of their specific title.

122 Pankaja Desai, Priyanka Nasa, Jackie Soo, Cunhui Jia, Michael Berbaum, James Fischer, and Timothy Johnson, *Effects of Regulatory Support Services on Institutional Review Board Turnaround Times*, Journal of Empirical Research on Human Research Ethics: JERHRE, 12(3): 131-139, *available at* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5546085/.

123 U. S. Department of Health and Human Services, *FWAs*, Accessed Mar. 31, 2022, https://www.hhs.gov/ohrp/register-irbs-and-obtain-fwas/fwas/index.html.

124 Kelsey Finch, *Conference Proceedings - Beyond IRBs Designing Ethical Review Processes for Big Data Research*, Future of Privacy Forum (Jan. 5, 2017), https://fpf.org/blog/conference-proceedings-beyond-irbs-designing-ethical-review-processes-big-data-research/.

125 Public Responsibility in Medicine and Research, *CIP Eligibility*, Accessed on Mar. 31, 2022, https://primr.org/cip/eligibility; *see also* U. S. Department of Health and Human Services, *Lesson 3: What are IRBs?*, Accessed Mar. 31, 2022, https://www.hhs.gov/ohrp/education-and-outreach/online-education/human-research-protection-training/lesson-3-what-are-irbs/index.html.

126 University of Michigan, *IRB Review Process*, Accessed on Mar. 31, 2022, https://research-compliance.umich.edu/human-subjects/irb-health-sciences-and-behavioral-sciences-hsbs/irb-review-process.

127 Patrick Varley, Ulrike Feske, Shasha Gao, Roslyn Stone, Sijian Zhang, Robert Monte, Robert Arnold, and Daniel Hall, *Time Required to Review Research Protocols at 10 VA Institutional Review Boards*, The Journal of Surgical Research, (2016), 204(2): 481-489, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7224356/.

128 U. S. Department of Health and Human Services, *Coded Private Information or Biospecimens Used in Research* (Jan. 19, 2018), https://www.hhs.gov/ohrp/coded-private-information-or-biospecimens-used-research.html; https://biospecimens.cancer.gov/bestpractices/elp/ic.asp; National Institute of Health, *Informed Consent* (Mar. 29, 2016), https://biospecimens.cancer.gov/bestpractices/elp/ic.asp.

129 Christine Grady, *Institutional Review Boards: Purpose and Challenges*, Chest (2015), 148(5): 1148-1155, *available at* https://pubmed.ncbi.nlm.nih.gov/26042632/.

130 Marjolein Timmers, Jeroen van Dijck, roel van Wijk, Valerie Legrand, Ernest van Veen, Andrew Maas, David Menon, Giuseppe Citerio, Nino Stocchetti, and Erwin Kompanje, *How Do 66 European Institutional Review Boards Approve One Protocol for an International Prospective Observational Study on Traumatic Brain Injury? Experiences from the CENTER-TBI Study*, BMC Medical Ethics, (May 12, 2020), 21(1): 36, *available at* http://europepmc.org/article/MED/32398066.

[131] Again, there are more mechanisms and rules associated with research in Europe and research about Europeans. For research using platform data in particular, a new code of ethics has been proposed that extends the existing legal framework. See: European Digital Media Observatory and George Washington University Institute for Data, Democracy, and Politics, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher-Data Access, 31 May 2022. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

[132] Accenture, *Data for Good*, (Sept. 23, 2020), Accessed on Mar. 31, 2022, https://www.accenture.com/us-en/insights/technology/data-good.

[133] *See,* SAS, *Data for Good. Analytics Helping Humanity*, Accessed on Apr. 1, 2022, https://www.sas.com/en_us/data-for-good.html; Columbia University, Data Science Institute, *Data for Good*, Accessed on Apr. 1, 2022, https://datascience.columbia.edu/about-us/data-for-good/; Meta, *Data for Good*, Accessed on Apr. 1, 2022, https://dataforgood.facebook.com/; FarmStack, *Free, Open-Source Software for Trusted Data Exchange*, Accessed on Apr. 1, 2022, https://farmstack.co/.

[134] Oracle, *Oracle for Research*, Accessed on Apr. 1, 2022, https://www.oracle.com/research/?source=:ad:pas:go:awr:a_nas:71700000088709647-58700007499775371-p67490766328:RC_DEVT211021P00001:research%20data%20sharing; Duke Clinical Research Institute, *Data Sharing*, Accessed on Apr. 1, 2022, https://dcri.org/our-work/analytics-and-data-science/data-sharing/; Twitter, Developer Platform, *Academic Research Access*, Accessed on Apr. 1, 2022, https://developer.twitter.com/en/products/twitter-api/academic-research.

[135] Snowflake, *Distributed Data Clean Rooms Powered by Snowflake* (Jan. 27, 2020), https://www.snowflake.com/blog/distributed-data-clean-rooms-powered-by-snowflake/.

[136] Microsoft, *Azure Guidance for Secure Isolation* (Dec. 1, 2021), https://docs.microsoft.com/en-us/azure/azure-government/azure-secure-isolation-guidance.

[137] U. S. Census Bureau, *Secure Research Environment*, Accessed on Apr 1, 2022, https://www.census.gov/about/adrm/fsrdc/about/secure_rdc.html.

[138] Chris Sanders and Jason Smith, *Chapter 14 - Friendly and Threat Intelligence*, Editor(s): Chris Sanders, Jason Smith, Applied Network Security Monitoring, (2014), Syngress, Pages 385-420, *available at* https://www.sciencedirect.com/science/article/pii/B9780124172081000143.

[139] Vicky Mahn-DiNicola, *Six Dimensions of Data Fitness*, Medisolv (Jan. 25, 2019), https://blog.medisolv.com/articles/six-dimensions-of-data-fitness.

[140] Vicky Mahn-DiNicola, *Six Dimensions of Data Fitness*, Medisolv (Jan. 25, 2019), https://blog.medisolv.com/articles/six-dimensions-of-data-fitness.

[141] U. S. Department of the Treasury, *Comptroller's Handbook* (Aug. 2021), https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf.

[142] Habu, *The Future of Data Collaboration in a Privacy-First, Industry-Driven World*, Accessed on Apr. 1, 2022, https://assets.habu.com/img/Habu_Data_Collaboration_Playbook.pdf.

[143] *See, e.g.*, the compliance areas at University of Connecticut: https://compliance.uconn.edu/compliance-areas/ or those areas required by University of North Carolina— Pembroke: https://www.uncp.edu/facultystaff/braveskickoff/annual-compliance-training.

[144] *See* the Regulated Data training requirements chart outlined by Cornell University, *available at* https://it.cornell.edu/regulated-data-chart-0.

[145] National Science Foundation, *RCR Frequently Asked Questions (FAQs)*, (Aug. 16, 2010), Accessed on Apr. 1, 2022, https://www.nsf.gov/pubs/policydocs/rcr/rcrfaqs.jsp#3.

[146] *See, e.g.,* University of Texas, https://security.utexas.edu/content/contractors-required-training.

[147] Luc Rocher, Julien Hendrickx and Yves-Alexandre de Montjoye, *Estimating the Success of Re-Identification in Incomplete, Datasets Using Generative Models,* Nature Communications, (2019), 10: 3069, *available at* https://www.nature.com/articles/s41467-019-10933-3; Mats Hansson, Hanns Lochmüller, Olaf Riess, Franz Schaefer, Michael Orth, Yaffa Rubinstein, Caron Molster, Hugh Dawkins, Domenica Taruscio, Manuel Posada, and Simon Woods, *The Risk of Re-Identification Versus the Need to Identify Individuals in Rare Disease Research*, European Journal of Human Genetics, (2016), 24: 1553-1558, *available at* https://www.nature.com/articles/ejhg201652.

[148] Sara Jordan, Clara Fontaine, and Rachele Hendricks-Sturrup, *Selecting Privacy-Enhancing Technologies for Managing Health Data Use*, Frontiers in Public Health, (Mar. 16, 2022), 10:814163, *available at* https://www.frontiersin.org/articles/10.3389/fpubh.2022.814163/full.

[149] Lindsay Walker, Michael Curry, Amritha Nayak, Nicholas Lange, Carlo Pierpaoli, and the Brain Development Group, *A Framework for the Analysis of Phantom Data in Multicenter Diffusion Tensor Imaging Studies*, Human Brain Mapping, (2013), 34(10):2439-2454, *available at* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3458186/.

[150] Uses of FTP or SFTP file transfer protocols employing usernames and passwords add a layer of security to data transfer. Secure https creates more secure gateways for uses of cloud services when transferring data. Network based techniques include use of connections managers that restrict or filter through source, destination, or host names. Likewise, valid node checking, database-enforced network access (virtual private databases), secure-multiparty computation, secure federated data architectures, and uses of encryption algorithms such as RDA or even Triple Data Encryption (3DES) can add multiple layers of security that go further than commonplace techniques for security research data assets. Finally, less obvious solutions, such as peer to peer communication and near-field communications can also be used to safely encrypt and transfer files between parties with similar software and intentions to directly share with one another.

[151] *See, e.g.*, https://www.medicaleconomics.com/view/hipaa-what-cost.

[152] Mary Pratt, *Cybersecurity Spending Trends for 2022: Investing in the Future*, CSO (Dec. 20, 2021), https://www.csoonline.com/article/3645091/cybersecurity-spending-trends-for-2022-investing-in-the-future.html.

[153] U. S. Department of Health and Human Services, *The HIPAA Privacy Rule*, Accessed on Apr. 1, 2022, https://www.hhs.gov/hipaa/for-professionals/privacy/index.html; U. S. Department of Health and Human Services, *Summary of the HIPAA Security Rule*, Accessed on Apr. 1, 2022, https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

[154] Person Centered Tech, *HIPAA Security Reminders*, Accessed on Apr. 1, 2022, https://personcenteredtech.com/memes/; Blog HIPAA, *HIPAA Funnies*, Accessed on Apr. 1, 2022, https://bloghipaa.com/category/hipaa/funny/#.YkdhQzfMIqt.

[155] *See, e.g.,* https://www.bu.edu/cfo/comptroller/departments/cashier/resources/pci-data-security-standards; https://www.ftc.gov/business-guidance/resources/how-comply-privacy-consumer-financial-information-rule-gramm-leach-bliley-act; https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/2349; https://digitalguardian.com/blog/what-fisma-compliance-fisma-definition-requirements-penalties-and-more; https://officeofresearch.ucsc.edu/compliance/services/irb40_educational_records.html.

[156] Gavin Yamey, *Scientists Who Do Not Publish Trial Results are "Unethical"*. British Medical Journal (1999), 319(7215): 939. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1116795/.

[157] U. S. Department of Health and Human Services, *Definition of Research Misconduct*, Accessed on Apr. 1, 2022, https://ori.hhs.gov/definition-research-misconduct.

[158] Friederike Hendiks, Dorothe Kienhues, Rainer Bromme, *Replication Crisis = Trust Crisis? The Effect of Successful vs Failed Replications on Laypeople's Trust in Researchers and Research*, Public Understanding of Science, (2020), 29(3): 270-288, *available at* https://pubmed.ncbi.nlm.nih.gov/32036741/.

[159] Brice McKeever, Solomon Greene, Graham MacDonald, Peter Tatian, Deondré Jones, *Data Philanthropy: Unlocking the Power of Private Data for Public Good* (Jul. 24, 2018), https://www.urban.org/research/publication/data-philanthropy-unlocking-power-private-data-public-good.

[160] David Grande, Jorge Machado, Bryan Petzold, and Marcus Roth, *Reducing Data Costs Without Jeopardizing Growth*, McKinsey (Jul. 31, 2020), https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/reducing-data-costs-without-jeopardizing-growth.

[161] Consortium of European Social Science Data Archives, *Data Management Expert Guide*, Accessed on Apr. 1, 2022, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide.

[162] GO FAIR, *FAIR Principles*, Accessed on Apr. 1, 2022, https://www.go-fair.org/fair-principles/.

[163] Kelsey Finch, FPF Best Practices and Contract Guidelines Help Companies Share Data with Academic Researchers, Future of Privacy Forum (Oct. 28, 2020), https://fpf.org/blog/fpf-best-practices-and-contract-guidelines-help-companies-share-data-with-academic-researchers/

[164] Data Documentation Initiative Alliance, *Document, Discover and Interoperate,* Accessed on Apr. 1, 2022, https://ddialliance.org/.

[165] Clinical Data Interchange Standards Consortium, Accessed on Apr. 1, 2022, https://www.cdisc.org/.

[166] U. S. Environmental Protection Agency, *Data Standards*, Accessed on Apr. 1, 2022, https://www.epa.gov/data-standards.

[167] Dublin Core, *Dublin Core™ Metadata Element Set, Version 1.1: Reference Description,* (Jun. 14, 2012), Accessed on Apr. 1, 2022, https://www.dublincore.org/specifications/dublin-core/dces/; National Information Exchange Model, Accessed on Apr. 1, 2022, https://www.niem.gov/.

[168] Natalia Mesa, *Q&A: A Randomized Approach to Awarding Grants*, The Scientist, (Feb. 25, 2022), Accessed on Mar. 29, 2022, https://www.the-scientist.com/news-opinion/q-a-a-randomized-approach-to-awarding-grants-69741.

[169] David Resnik, *Research Ethics Timeline*, National Institute of Environmental Health Sciences, Accessed on Apr. 1, 2022, https://www.niehs.nih.gov/research/resources/bioethics/timeline/index.cfm.

[170] European Digital Media Observatory and George Washington University Institute for Data, Democracy, and Politics, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher-Data Access, 31 May 2022. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

[171] Marianne Varkiani, *Call for Public Comments: Resources for Companies Sharing Personal Data with Academic Researchers*, Future of Privacy Forum (Feb. 18, 2020), https://fpf.org/blog/call-for-public-comments-resources-for-companies-sharing-personal-data-with-academic-researchers/.

[172] Sara Jordan, FPF Ethical Data Use Committee will Support Research Relying on Private Sector Data, Future of Privacy Forum (May 5, 2021), https://fpf.org/blog/fpf-ethical-data-use-committee-will-support-research-relying-on-private-sector-data/. The Ethical Data Use Committee was designed and developed with the support of Schmidt Futures building on previous FPF work funded by the Alfred P. Sloan Foundation and the National Science Foundation.

[173] Woods Hole Oceanographic Institution, Accessed on Apr. 1, 2022, https://www.whoi.edu/.

[174] *See, e.g.,* General Assembly of the State of South Carolina, A319, R343, H4840, 116th Session, 2005-2006, (2006). https://www.scstatehouse.gov/sess116_2005-2006/bills/4840.htm; Connecticut General Assembly, sSB 1258, Public Act No. 05-198, (2005), https://www.cga.ct.gov/2005/ACT/PA/2005PA-00198-R00SB-01258-PA.htm; The Florida Senate, SB 52: Postsecondary Education, (2021), https://www.flsenate.gov/Session/Bill/2021/52/?Tab=BillText.

[175] The Editorial Board, *Academic Tenure is in Desperate Need of Reform*, (May 9, 2021), Accessed on Apr. 1, 2022, https://www.bostonglobe.com/2021/05/09/opinion/academic-tenure-is-desperate-need-reform/.

[176] Paul Sanberg, Morteza Gharib, Patrick Harker, Eric Kaler, Richard Marchase, Timothy Sands, Nasser Arshadi, and Sudeep Sarkar, *Changing the Academic Culture: Valuing Patents and Commercialization Toward Tenure and Career Advancement*, Proceedings of the National Academy of Sciences of the United States of America, (2014), 111(18): 6542-6547, *available at* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4020064/.

[177] Kerry Ann O'Meara & Audrey Jaeger, *Preparing Future Faculty for Community Engagement: Barriers, Facilitators, Models, and Recommendations*, Journal of Higher Education Outreach and Engagement, (2006), 11(4): 3-26, *available at* https://openjournals.libs.uga.edu/jheoe/article/view/537.

[178] Ishani Chettri, *Clinical Research Office Will Promote Project Collaboration, Medical Faculty Say*, the GW Hatchet (Jan. 18, 2022), https://www.gwhatchet.com/2022/01/18/clinical-research-office-will-promote-project-collaboration-medical-faculty-say.

[179] Liz Ferguson, *How and Why Researchers Share Data (and Why They Don't)*, Wiley (Nov. 3, 2014), https://www.wiley.com/network/researchers/licensing-and-open-access/how-and-why-researchers-share-data-and-why-they-dont; Naomi Waithira, Brian Mutinda, & Phaik Yeong Cheah, *Data Management and Sharing Policy: The First Step Towards Promoting Data Sharing*, BMC Medicine, (2019), 17:80, *available at* https://link.springer.com/article/10.1186/s12916-019-1315-8.

[180] Naomi Waithira, Brian Mutinda, and Phaik Yeong Cheah, *Data Management and Sharing Policy: The First Step Towards Promoting Data Sharing*, BMC Medicine, (2019), 17:80, *available at* https://link.springer.com/article/10.1186/s12916-019-1315-8.

[181] GO FAIR, *FAIR Principles*, Accessed on Apr. 1, 2022, https://www.go-fair.org/fair-principles/.

[182] These last two points— data integrity checking— should be a robust part of the data sharing arrangement. Use of MD5Checksums, for example, helps ensure that shared data is not tampered with in the researchers' environment or corrupted through the interaction of company systems with that data asset.

[183] U. S. Department of Commerce, National Institute of Standards and Technology, *An Introduction to the Components of the Framework*, Accessed on Apr. 1, 2022, https://www.nist.gov/cyberframework/online-learning/components-framework.

[184] U. S. Department of Commerce, National Institute of Standards and Technology, *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0* (Jan. 16, 2020), https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

[185] Matt Dumiak, *NIST Privacy 101: An Intro to the NIST Privacy Framework*, JDSupra (Feb. 9, 2021),https://www.jdsupra.com/legalnews/nist-privacy-101-an-intro-to-the-nist-6416366/.

[186] U. S. Department of Health and Human Services, *Federal Policy for the Protection of Human Subjects ('Common Rule')*, Accessed on Apr. 1, 2022, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html; *see also* U. S. Department of Health and Human Services, *Annotated Comparison of the Pre-2018 Common Rule with the Revised Common Rule*, Accessed on Apr. 1, 2022, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/annotated-2018-requirements/index.html.

[187] The Center for Democracy and Technology recently made the case for such sharing by drawing from three regulated contexts: clinical trials (life sciences/biomedical), smart meters for electricity, and environmental impact statements. By contrast, this report addresses the potential for any kind of company to share data for research, since data has become more important to all companies. See: Gabriel Nicholas and Dhanuraj Thakur, Learning to Share: Lessons from Beyond Social Media, Center for Democracy and Technology, September 2022. https://cdt.org/wp-content/uploads/2022/09/20220907-learningtoshare-final.pdf

[188] These concerns were expressed in an open letter responding to a new White House call for open publication, including for research data. See: https://ostp-letter.github.io

[189] Martin Sandbu, *On Dworkin's Brute-Luck-Option-Luck Distinction and the Consistency of Brute-Luck Egalitarianism*, Politics, Philosophy and Economics, (2004), 3(3): 283-312.

[190] This potential mandate would be consistent with the EU Digital Services Act as well as GDPR. See: European Digital Media Observatory and George Washington University Institute for Data, Democracy, and Politics, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher-Data Access, 31 May 2022. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

[191] White House, "OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay," August 25, 2022. https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/

[192] Springer, *Research Data Policy*, Accessed on Apr. 1, 2022, https://www.springer.com/journal/43681/submission-guidelines#Instructions%20for%20Authors_Research%20Data%20Policy.

[193] The University of Chicago Press Journals, Journal of Labor Economics, *Data Policy,* Accessed on Apr. 1, 2022, https://www.journals.uchicago.edu/journals/jole/data-policy.

[194] The Lancet, *Information for Authors*, Accessed on Apr. 1, 2022, https://www.thelancet.com/pb-assets/Lancet/authors/tl-info-for-authors.pdf.

[195] American Journal of Political Science, *AJPS Verification Policy*, Accessed on Apr. 1, 2022, https://ajps.org/ajps-verification-policy/.

[196] The Astronomy and Astrophysics Review, *Submission Guidelines*, Accessed on Apr. 1, 2022, https://www.springer.com/journal/159/submission-guidelines.

Notes

Notes

FUTURE OF PRIVACY FORUM

1350 EYE STREET NW | SUITE 350 | WASHINGTON, DC 20005 FPF.ORG | info@fpf.org