Privacy, Big Data, and the Public Good

Frameworks for Engagement

Massive amounts of data on human beings can now be analyzed. Pragmatic purposes abound, including selling goods and services, winning political campaigns, and identifying possible terrorists. Yet "big data" can also be harnessed to serve the public good: scientists can use big data to do research that improves the lives of human beings, improves government services, and reduces taxpayer costs. In order to achieve this goal, researchers must have access to this data – raising important privacy questions. What are the ethical and legal requirements? What are the rules of engagement? What are the best ways to provide access while also protecting confidentiality? Are there reasonable mechanisms to compensate citizens for privacy loss?

The goal of this book is to answer some of these questions. The book's authors paint an intellectual landscape that includes legal, economic, and statistical frameworks. The authors also identify new practical approaches that simultaneously maximize the utility of data access while minimizing information risk.

Julia Lane is Senior Managing Economist for the American Institutes for Research in Washington, DC. She holds honorary positions as professor of economics at the BETA University of Strasbourg CNRS, chercheur associée at Observatoire des Sciences et des Techniques, Paris, and professor at the University of Melbourne's Institute of Applied Economics and Social Research.

Victoria Stodden is Assistant Professor of Statistics at Columbia University and is affiliated with the Columbia University Institute for Data Sciences and Engineering.

Stefan Bender is head of the Research Data Center (RDC) at the German Federal Employment Agency in the Institute for Employment Research (IAB).

Helen Nissenbaum is Professor of Media, Culture and Communication, and Computer Science at New York University, where she is also director of the Information Law Institute.

Privacy, Big Data, and the Public Good

Frameworks for Engagement

Edited by

Julia Lane
American Institutes for Research, Washington DC

Victoria Stodden
Columbia University

Stefan Bender
Institute for Employment Research of the German Federal Employment Agency

Helen Nissenbaum
New York University

CAMBRIDGE
UNIVERSITY PRESS

CONTENTS

Contributors *page* vii

Editors' Introduction ix

Part I. Conceptual Framework 1

1 **Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context** 5
Katherine J. Strandburg

2 **Big Data's End Run around Anonymity and Consent** 44
Solon Barocas and Helen Nissenbaum

3 **The Economics and Behavioral Economics of Privacy** 76
Alessandro Acquisti

4 **Changing the Rules: General Principles for Data Use and Analysis** 96
Paul Ohm

5 **Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency** 112
Victoria Stodden

Part II. Practical Framework 133

6 **The Value of Big Data for Urban Science** 137
Steven E. Koonin and Michael J. Holland

7 **Data for the Public Good: Challenges and Barriers in the Context of Cities** 153
Robert M. Goerge

8 **A European Perspective on Research and Big Data Analysis** 173
Peter Elias

9 **The New Deal on Data: A Framework for Institutional Controls** 192
Daniel Greenwood, Arkadiusz Stopczynski, Brian Sweatt, Thomas Hardjono, and Alex Pentland

10 **Engineered Controls for Dealing with Big Data** 211
Carl Landwehr

11 **Portable Approaches to Informed Consent and Open Data** 234
John Wilbanks

Part III. Statistical Framework 253

12 **Extracting Information from Big Data: A Privacy and Confidentiality Perspective** 257
Frauke Kreuter and Roger Peng

13 **Using Statistics to Protect Privacy** 276
Alan F. Karr and Jerome P. Reiter

14 **Differential Privacy: A Cryptographic Approach to Private Data Analysis** 296
Cynthia Dwork

5 Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency

Victoria Stodden

Introduction

The 21st century will be known as the century of *data*. Our society is making massive investments in data collection and storage, from sensors mounted on satellites down to detailed records of our most mundane supermarket purchases. Just as importantly, our reasoning about these data is recorded in software, in the scripts and code that analyze this digitally recorded world. The result is a deep digitization of scientific discovery and knowledge, and with the parallel development of the Internet as a pervasive digital communication mechanism we have powerful new ways of accessing and sharing this knowledge. The term *data* even has a new meaning. Gone are the days when scientific experiments were carefully planned prior to data collection. Now the abundance of readily available data creates an observational world in itself suggesting hypotheses and experiments to be carried out after collection, curation, and storage of the data has already occurred. We have departed from our old paradigm of data collection to resolve research questions – nowadays, we collect data simply *because we can*.

In this chapter I outline what this digitization means for the independent verification of scientific findings from these data, and how the current legal and regulatory structure helps and hinders the creation and communication of reliable scientific knowledge.[1] Federal mandates and laws regarding data disclosure, privacy, confidentiality, and ownership all influence the ability of researchers to produce openly available and reproducible research. Two guiding principles are suggested to accelerate research in the era of big data and bring the regulatory infrastructure in line with scientific norms: the Principle of Scientific Licensing and the Principle of Scientific Data and

Code Sharing. These principles are then applied to show how intellectual property and privacy tort laws could better enable the generation of verifiable knowledge, facilitate research collaboration with industry and other proprietary interests through standardized research dissemination agreements, and give rise to dual licensing structures that distinguish between software patenting and licensing for industry use and open availability for open research. Two examples are presented to give a flavor of how access to data and code might be managed in the context of such constraints, including the establishment of 'walled gardens' for the validation of results derived from confidential data, and early research agreements that could reconcile scientific and proprietary concerns in a research collaboration with industry partners.

Technological advances have complicated the data privacy discussion in at least two ways. First, when datasets are linked together, a richer set of information about a subject can result but so can an increased risk of a privacy violation. Linked data presents a challenging case for open scientific research, in that it may permit privacy violations from otherwise non-violating datasets. In this case privacy tort law is suggested as a viable remedy for privacy violations that arise from linking datasets.

Second, the subjects of studies are becoming more knowledgeable about privacy issues, and may wish to opt for a greater level of access to their contributed data than that established by traditional research infrastructures, such as Institutional Review Boards. For data collection and release that happens today, research subjects have very little say over the future openness of their data. A suggestion is made to permit individuals to share their own data with provisions regarding informed consent. For example, an enrollee in a clinical trial for a new Crohn's disease treatment may wish to permit other Crohn's researchers access to the data arising from her participation, perhaps in an effort to help research advance in an area about which she cares deeply. At the moment, this is not only nonstandard, but downstream data use is difficult for the participant to direct.

Ownership itself can be difficult to construe since many resources typically go into creating a useful dataset, from research scientists who design the experiment, to data collectors, to participants, to curators, to industry collaborators, to institutes and funding agencies that support the research, further complicating the discussion of data and code access. Data access becomes increasingly complex, underscoring the need for a broad understanding of the value of maximizing open access to research data and code.

Trust and Verify: Reliable Scientific Conclusions in the Era of Big Data

Scientific research is predicated on an understanding of scientific knowledge as a public good – this is the rationale underlying today's multibillion-dollar subsidies of scientific research through various federal and state agencies. The scientific view is not one of adding nuggets of truth to our collective understanding, but instead one of weighing evidence and assigning likelihoods to a finding's probability of being true. This creates a normative structure of skepticism among scientists: the burden is on the discovering scientist to convince others that what he or she has found is more likely to be correct than our previous understanding. The scientific method's central motivation is the *ubiquity of error* – the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error. As a result, standards of scientific communication evolved to incorporate full disclosure of the methods and reasoning used to arrive at the proffered result.

The case for openness in science stems from Robert Boyle's exhortations in the 1660s for standards in scientific communication. He argued that enough information should be provided to allow others in the field to independently reproduce the finding, creating both the greatest chance of the accurate transmission of the new discoveries and also maximizing the likelihood that errors in the reasoning would be identified. Today, communication is changing because of the pervasive use of digital technology in research. Digital scholarly objects such as data and code have become essential for the effective communication of computational findings. Computations are frequently of such a complexity that an explanation sufficiently detailed to enable others to replicate the results is not possible in a typical scientific publication. A solution to this problem is to accompany the publication with the code and data that generated the results and communicate a *research compendium*.[2] However, the scientific community has not yet reached a stage where the communication of research compendia is standard.[3] A number of delicate regulatory and policy changes are essential to catalyze both scientific advancement and the development of applications and discoveries outside academia by making the data and code associated with scientific discoveries broadly available.

Challenge 1: Intellectual Property Law and Access to Digital Scholarly Objects

The Intellectual Property Clause of the United States Constitution has been interpreted to confer two distinct powers, the first providing the basis for copyright law: Securing for a limited time a creator's exclusive right to their original work;[4] and the second giving the basis for patent law: Endowing an inventor with a limited-term exclusive right to use their discoveries in exchange for disclosure of the invention. In this section the barrier copyright creates to open reproducible research will be discussed first, then the role of patent law in potentially obfuscating computational science.

Creators do not have to apply for copyright protection, as it adheres automatically when the original expression of the idea is rendered in fixed form. Many standard scientific activities, such as writing a computer script to filter a dataset or fit a statistical model, will produce copyrighted output, in this case the code written to implement these tasks. Building a new dataset through the original selection and arrangement of data will generate ownership rights through copyright for the dataset creator, to give another example.[5] The default nature of copyright creates an *intellectual property* framework for scientific ideas at odds with longstanding scientific norms in two key ways.[6] First, by preventing copying of the research work it can create a barrier to the legal reproduction and verification of results.[7] Second, copyright also establishes rights for the author over the creation of derivative works. Such a derivative work might be something as scientifically productive as, say, the application of a software script for data filtering to a new dataset, or the adaptation of existing simulation codes to a new area of research.

As computation becomes central to scientific investigation, copyright on code and data become barriers to the advancement of science. There is a copyright exception, titled *fair use,* which applies to "teaching (including multiple copies for classroom use), scholarship, or research"[8] but this does not extend to the full research compendium including data, code, and research manuscript. In principle, a relatively straightforward solution to the barrier copyright imposes would be to broaden the fair use exception to include scientific research that takes place in research institutions such as universities or via federal research grants; however, this is extremely challenging in practice.[9] Distinguishing legal fair use is not a clear exercise in any event, and an extension to research more broadly may still not sufficiently clarify rights. A more practical mechanism for realigning intellectual property rights with scientific norms is the Reproducible Research Standard (RRS), applying appropriate open licenses to remove restrictions on copying and reuse of the scientific work, as well as possibly adding an attribution requirement to elements of the research compendium. Components of the research compendium have different features that necessitate different licensing approaches and a principle for licensing scientific digital objects can guide choices:

Principle of Scientific Licensing *Legal encumbrances to the dissemination, sharing, use, and reuse of scientific research compendia should be minimized, and require a strong and compelling rationale before their application.*[10]

For media components of scientific work, the Reproducible Research Standard suggests the Creative Commons attribution license (CC BY), which frees the work for replication and re-use without prior author approval, with the condition that attribution must accompany any downstream use of the work.

Many licenses exist that allow authors to set conditions of use for their code. In scientific research code can consist of scripts that are essentially stylized text files (such as python or R scripts) or the code can have both a compiled binary form and a source representation (such as code written in C). Use of the CC BY license for code is discouraged by Creative Commons.[11] The Reproducible Research Standard suggests the Modified Berkeley Software Distribution (BSD) license, the MIT license, or the Apache 2.0 license, which permit the downstream use, copying, and distribution of either unmodified or modified source code, as long as the license accompanies any distributed code and the previous authors' names are not used to promote modified downstream code.[12] The Modified BSD and MIT licenses differ in that the MIT license does not include a clause forbidding endorsement.[13] The Apache 2.0 license differs in that it permits the exercise of patent rights that would otherwise extend only to the original licensor, meaning that a patent license is granted for those patents needed for use of the code.[14] The license further stipulates that the right to use the work without patent infringement will be lost if the downstream user of the code sues the licensor for patent infringement.

Collecting, cleaning, and preparing data for analysis can be a significant component of empirical scientific research. Copyright law in the United States forbids the copyrighting of 'raw facts' but original products derived from those facts can be copyrightable. In *Feist Publications,* *Inc. v. Rural Telephone Service*, the Court held that the *original* "selection and arrangement" of databases is copyrightable:[15] the component falling under copyright must be original in that "copyright protection extends only to those components of the work that are original to the author, not to the facts themselves."[16] Attaching an attribution license to the original "selection and arrangement" of a database may encourage scientists to release the datasets they have created by providing a legal framework for attribution and reuse of the original selection and arrangement aspect of their work.[17] Since the raw facts themselves are not copyrightable, such a license cannot be applied to the data themselves. The selection and arrangement may be implemented in code or described in a text file accompanying the dataset, either of which can be appropriately licensed. Data can however be released to the public domain by marking with the Creative Commons CC0 standard.[18]

This licensing structure that makes the total of the media, code, data components – the research compendium – available for reuse, in the public domain or with attribution, is labeled the *Reproducible Research Standard*.

Patent law is the second component of intellectual property law that affects the disclosure of scientific scholarly objects. In 1980 Congress enacted two laws, the Stevenson-Wydler Act and the Bayh-Dole Act, both intended to promote the commercial development of technologies arising from federally funded research. This was to be facilitated through licensing agreements between research entities, such as universities, and for-profit companies. The Bayh-Dole Act explicitly gave federal agency grantees and contractors, most notably universities and research institutions, title to government-funded inventions and charged them with using the patent system to disclose and commercialize inventions arising in their institution. In 2009 this author carried out a survey of computational scientists, in order to understand why they either shared or withheld the code and data associated with their published papers. In the survey one senior professor explained that he was not revealing his software because he was currently seeking a patent on the code.[19] In fact, 40% of respondents cited patent seeking or other intellectual property constraints as a reason they were not sharing the code associated with published scientific results.[20] Rates of software patenting by academic institutions have been increasing over the last decade, posing a potentially serious problem for scientific transparency and reproducibility.[21] Instead of ready access to the code that generated published results, a researcher may be required to license access to the software through a university's technology transfer office, likely being prohibitively expensive for an academic scientist in both time and money. In December of 1999, the National Institutes of Health stated that

> the use of patents and exclusive licenses is not the only, nor in some cases the most appropriate, means of implementing the [Bayh-Dole] Act. Where the subject invention is useful primarily as a research tool, inappropriate licensing practices are likely to thwart rather than promote utilization, commercialization, and public availability.[22]

The federal funding agencies are without authority to issue regulations regarding patentable inventions, and the NIH viewpoint above does not appear to have been adopted by technology transfer offices at the university and institutional research level. A typical interpretation is that of Columbia University, where this author is employed, which follows: "The University claims, as it may fairly and rightfully do, the commercial rights in conceptions that result primarily from the use of its facilities or from the activity of members of its faculty while engaged in its service."[23] Not all universities make such an a priori claim to determine the patenting and licensing fate of research inventions. For example, Stanford University's *Research Policy Handbook* says that as a researcher, "I am free to place my inventions in the public domain as long as in so doing neither I nor Stanford violates the terms of any agreements that governed the work done."[24] The Bayh-Dole Act also grants agencies 'march-in' rights to obtain intellectual property (presumably to grant nonexclusive licenses, but not necessarily), but the process is long with multiple appeal opportunities. In July of 2013, however, in a letter to Francis Collins, head of the NIH, Senator Leahy recommended the use of march-in rights on patented breast cancer genetic research "to ensure greater access to genetic testing for breast and ovarian cancer."[25]

Challenge 2: Scale, Confidentiality, and Proprietary Interests

Even without intellectual property law encumbrances to the dissemination of digital scholarly objects, other barriers can create obstacles to access. For example, the sheer size of many datasets may require specialized computational infrastructure to permit access, or scale itself can even prohibit access. For example, the July 2013 release of the Sloan Digital Sky Survey (SDSS) is 71.2 terabytes in size, making a conventional download of data to a personal laptop impossible.[26] The approach of the SDSS is to create different websites for different data types, and provide a variety of tools for access including SkyServer SQL search, CasJobs, and Schema Browser, each with a different purpose in mind.[27] This infrastructure permits search and user-directed access to significantly smaller subsets of the entire database.

In some fields however even 70 terabytes would not seem large. CERN director general Rolf Heuer said in 2008 that, "[t]en or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data."[28] In March of 2013, the CERN data center passed a storage milestone by exceeding 100 petabytes of data.[29] It is not clear how this can be made open data in the sense discussed in this chapter, as Director Heuer suggests. The traditional approaches to making data and code available seem intractable for such datasets at the present time. I use these examples to introduce a Principle of Scientific Data and Code Sharing:

Principle of Scientific Data and Code Sharing *Access to the data and methods associated with published scientific results should be maximized, only subject to clearly articulated restrictions such as: privacy or confidentiality concerns, legal barriers including intellectual property or HIPAA regulations, or technological or cost constraints.*

This principle can also be phrased as 'Default to Open', meaning that it takes compelling and convincing reasons, articulated in detail (i.e. the precise section of HIPAA that is restricting disclosure, or the part of intellectual property law that is creating a barrier) to close data and code from public access.[30,31] A careful defense of any deviation from full openness will have the effect of maximizing the availability of data and code. A corollary effect is an uncovering of the reasons for not sharing data and presumably a greater understanding of the precise nature of legal barriers to disclosure and their appropriateness given the nature of the underlying data and code.[32] Sequestering a dataset due to 'confidentiality', with no further justification, should no longer be acceptable practice.

The second corollary from the Principle of Scientific Data and Code Sharing is that it implies *levels* of access. Whether due to privacy concerns, technological barriers, or other sources, restrictions on data and code availability do not necessarily imply absolute barriers. In the case of CERN, internal research design mechanisms exist to make up for some of the shortcomings in openness of research data and the inability of independent groups to verify findings obtained from empirical data. Specifically, either independent research groups within CERN access the data from the collider and carry out the research in isolation from each other, or the same group will verify analyses using independent toolsets.[33] Of crucial importance, these internal groups have access to the infrastructure and technologies needed to understand and analyze the data. In this case, there has been some openness of the data and the use of independent parallel research increases the chances of catching errors, all improvements over the more commonly seen research context where the data are accessed only by the original researchers and analyzed without any reported validation or verification cross-checks.

A second illustrative example originates from the Wandell Lab in the Psychology Department at Stanford University. Brian Wandell, the Isaac and Madeline Stein Family Professor of Psychology, has an MRI machine for his lab research. For the lifetime of the machine, each image has been carefully stored in a database with metadata including equipment settings, time, date, resolution, and other pertinent details of the experimental setup. The output image data are, however, subject to HIPAA regulations in that each image is a scan of a subject's brain and therefore privacy restrictions prevent these from being made publicly available. The Wandell Lab belongs to a consortium with several other research groups at different universities in California. In order to permit some potential verification of results based upon these images, there is no legal barrier to giving researchers within these authorized groups access to the database, and thereby creating the possibility for independent cross-checking of findings inside this 'walled garden'. While this does not achieve the same potential for finding errors as open release would (more eyes making more bugs increasingly shallow), it satisfies the Principle of Scientific Data and Code Sharing by maximizing access subject to the inherent legal constraints with which the data are endowed. Although the implementation details may differ for different data, understanding and developing infrastructure to facilitate these middle-ground data access platforms or walled gardens, will be essential for the reliability of results derived from confidential data.[34] One could also cast the CERN approach as a type of walled garden since it is characterized by independent research on the same question on closed data, carried out by different internal groups.

Another potential barrier to data and code release derives from collaboration with partners who may be unwilling to release the data and software that arise from the project, and may not be academic researchers bound by the same notions of scientific transparency. For example, industry research partners do not necessarily have the goal of contributing their research findings to the public good, but are frequent research collaborators with academics. A conflict can ensue, for example, at the point of publication when the academic partner wishes to publish the work in a journal or a conference proceedings that requires data and code disclosure, or when the researcher simply wishes to practice really reproducible research and make the data and code openly available.[35] One possible solution is to offer template agreements for data and code disclosure at the beginning of the collaboration, possibly through the institution's technology transfer office or through funding agency access policy.[36] Unfortunately the issue of data and code access is often ignored until the point at which one party would like to make them available after the research has been completed.[37]

When a patent is being sought on the software associated with the research, broader access can be achieved by implementing patent licensing terms that distinguish between commercial and research applications, in order to permit reuse and verification by researchers, while maintaining the incentives for commercialization and technology transfer provided by the Bayh–Dole Act. The Stanford Natural Language Processing Group for example uses such a dual licensing strategy. Their code is available for download by researchers under an open license and groups that intend commercial reuse must pay licensing fees.[38]

Challenge 3: Linked Data and Privacy Tort Law

Access to datasets necessarily means data with common fields can and will be linked. This is very important for scientific discovery as it enriches subject-level knowledge and opens new fields of inquiry, but it comes with risks such as revealing private information about individuals that the datasets in their isolated, unlinked form would not reveal. As has been widely reported, data release is now mandated for many government agencies through Data.gov. In 2009 Vivek Kundra, then–federal chief information officer,[39] was explicit – saying that, "the dream here is that you have a grad student, sifting through these datasets at three in the morning, who finds, at the intersection of multiple datasets, insight that we may not have seen, or developed a solution that we may not have thought of." On February 22, 2013, the Office of Science and Technology Policy directed federal agencies with significant research budgets to remit plans to make data arising from this research openly available.[40] This includes academic research funded by the National Science Foundation and the National Institutes for Health, for example.

An instructive example about the privacy risks from data linking that Kundra describes comes from the release of genomic information. An individual's genomic information could be uncovered by linking their relatives' genomic information together, when this individual has not shared any of his or her genetic information directly. Recall, we carry 50% of the DNA from each of our parents and children, and an average of 50% from each of our siblings. Privacy risks could include, for example, an insurance company linking the genetic signature information to medical records data, possibly through a genetic diagnostic test that was performed, and then to other insurance claims, for individuals whose relatives had made their DNA available though they themselves did not.[41] A number of cities are releasing data, for example public school performance data, social service agency visits, crime reports, and other municipal data, and there has been controversy over appropriate privacy protection for some of these data.[42] Research that links these datasets may have laudable aims – better understanding the factors that help students succeed in their education – but the risks to linking datasets can include privacy violations for individuals.

Much of the policy literature around privacy in digitally networked environments refers to corporate or government collected data used for commercial or policy ends.[43] Insufficient attention has been paid to the compelling need for access to data for the purposes of verification of data-driven research findings. This chapter does not advocate that the need for reproducibility should trump privacy rights, but instead that scientific integrity should be part of the discussion of access to big data, including middle ground solutions such as those as discussed earlier in this chapter.

Traditional scientists in an academic setting are not the only ones making inferences from big data and linked data, as Chapters 6 and 7 in this volume show. The goal of better decision making is behind much of the current excitement surrounding big data, and supports the emergence of 'evidence-based' policy, medicine, practice, and management. For conclusions that enter the public sphere, it is not unreasonable to expect that the steps that generated the knowledge be disclosed to the maximal extent possible, including making the data they are based on available for inspection, and making the computer programs that carried out the data analysis available.

We cannot know how data released today, even data that all would agree carry no immediate privacy violations, could help bring about privacy violations when linked to other datasets in the future. These other datasets may not be released, or even imagined, today. It is impossible to guard completely against the risk of these types of future privacy violations. For this reason a tort-based approach to big data access and privacy is an important alternative to creating definitive guidelines to protect privacy in linked data. Perhaps not surprisingly, however, privacy tort law developed in the pre-digital age and is not a perfect fit with today's notions of privacy protection and big data access.

Much of the current scholarly literature frames the online privacy violation question as protection again defamation or the release of private information by others, and does not explicitly consider the case of linked data. For example, privacy torts are often seen as redress for information made available online, without considering the case of harm from new information derived from combination of non-private sources. This can happen in the case of data linking, as described above, but differs in that a privacy violation can be wholly inadvertent and unforeseen, and may not be individually felt but can affect an entire class of people (those in the dataset). This, along with persistence of privacy-violating information on the web, changes the traditional notion of an individual right to privacy.[44] In current privacy tort law one must establish that the offender intended to commit a privacy invasion,[45] that the conduct was "highly offensive to the reasonable person," and that the information revealed was sufficiently private.[46]

Current privacy tort law protects against emotional, reputational, and proprietary injuries caused by any of: a public disclosure of private facts; an intrusion on seclusion; a depiction of another in a false light, or an appropriation of another's image for commercial enrichment.[47] Articulating privacy rights in big data and linked data founders on accountability since it is unlike securing private (real) property or a landlord ensuring his or her building is secured.[48] Potential privacy violations deriving from linked data cannot always be foreseen at the time of data release. The Principle of Scientific Data and Code Sharing frames a possible way forward: research data that does not carry any immediate privacy violations should be released (and otherwise released in a way that makes the data maximally available for independent reproducibility purposes that safeguards privacy); linked datasets should either be released or the methods to link the datasets should be released with caveats to check for new privacy violations; and if privacy violations still arise, redress could be sought through the tort system. If tort law responds in a way that matches our normative expectations regarding privacy in data, this will permit a body of law to grow around big data that protects privacy. In order for this to be effective, a broadening of tort law beyond the four types of privacy-violating behaviors needs to occur. Harms arising from the release of private information derived from data, and from linked data, could be included in the taxonomy of privacy torts. These may not be intentioned or foreseeable harms, and may potentially be mass torts as datasets with confidentiality violations are likely to contain records on a large number of people. The issue of liability and responsibility for privacy violations becomes more complex than in the past, and there may be chilling effects on the part of institutions and funding agencies with regard to open data. Finally, making code and data available is not costless as databases and access software can cost a considerable amount of money, and innovative middle-ground solutions that may be project specific can add to that cost.[49]

Research data poses yet another unique challenge to privacy law. Many research collaborations exist across international boundaries, and it is common for some members of a research team to be more heavily involved with the associated data than other members. Access to data on the Internet is not generally restricted by country and enforcing privacy violations across international borders poses a considerable challenge for scientific research. Data and code must be made available to maximally permit verification, subject to privacy and other barriers, and these data may be accessible from anywhere in the world through the Internet. Privacy violations from linked data can thus occur in countries with more stringent privacy standards though the release of the data may occur in a country that does not have a mechanism for legal redress of privacy violations.

Challenge 4: Changing Notions of Data Ownership and Agency

The notion of a data owner is a rapidly changing concept as many entities contribute to dataset creation, increasing the complexity of the data-sharing issue. Data is collected both by people and by automated systems such as sensor arrays, and goes through myriad processing in the course of information extraction. Different entities may carry out data cleaning and filtering, data curation and warehousing, facilitation of data access, recombination of datasets to create novel databases, or preservation and provenance through repositories and institutions – each possibly creating intellectual property and ownership rights in the data. There is a similar story for research code, as it evolves through different applications and extensions by different people and becomes an amalgam of many contributions. The open release of data and code means untangling ownership and tracking contributions. Versions of code and data are vitally important for reproducibility – as code is modified, even as bugs are fixed, or data are extended, corrected, or combined, it is important to track which instantiation produced which scientific findings.

There is a new source of potential ownership as well. Subjects in a study can feel a sense of ownership over information about themselves, including medical descriptions or choices they have made. It is becoming increasingly the case that study participants wish to direct the level of access to data about themselves and traditional notions of privacy protection may not match their desires. Some data owners would prefer that data about themselves, that might traditionally be considered worthy of privacy protection such as medical data or data resulting from clinical trials participation, should be made more fully available.[50] As noted in a World Economic Forum Report, "[o]ne of the missing elements of the dialogue around personal data has been how to effectively engage the individual and give them a voice and tools to express choice and control over how data about them are used."[51,52] Traditional mechanisms, such as the Institutional Review Board or national laboratory policy, may be overprotecting individuals at the expense of research progress if they are not taking individual agency into account.

These changing notions of ownership can impede sharing, if permission from multiple parties is required to grant open access, or to relinquish data, or even to simply participate in the development of infrastructure to support access. A careful assessment of ownership and contributions to dataset development will inform liability, in the case of breaches of privacy. While some of this assessment and tracking is done today for some datasets, for the majority of datasets there is very little provenance available and little clarity regarding ownership rights.

Conclusion

The goal of this chapter is to bring the consideration of scientific research needs to the discussion around data disclosure and big data. These needs comprise a variety of issues, but a primary one is the need for independent verification of results, for reproducibility from the original data using the original code. This chapter asserts two principles to guide policy thinking in this area: the **Principle of Scientific Licensing**, that legal encumbrances to the dissemination, sharing, use, and re-use of scientific research compendia should be minimized, and require a strong and compelling rationale before their application; and the **Principle of Scientific Data and Code Sharing**, that access to the data and methods associated with published scientific results should be maximized, only subject to clearly articulated restrictions interpreted in the most minimally restrictive way, including intellectual property or HIPAA restrictions, or technological or cost constraints.

The chapter outlines intellectual property barriers to the open release of research data and code, and proposes open licensing solutions. Templated sharing agreements are suggested to guide data and code release at the beginning of collaboration with industry partners who may have a different agenda to the open sharing of data and code that arise from the research. The chapter also argues for dual licensing of patented research code: license fees for commercial reuse, and open availability for academic research purposes. To address privacy and confidentiality in sharing there must be a move to maximize openness in the face of these concerns. Sharing within a group of authorized researchers in the field, or with scientists who have sought permission, can create a 'walled garden' that, while inferior to open sharing, can still obtain some of the properties and benefits of independent verification that is possible from public access. 'Middle-ground' platforms such as walled gardens are possible solutions to maximize the reliability of scientific findings in the face of privacy and confidentiality concerns.

The linking of open data sets is framed as an open-ended threat to privacy. Individuals may be identified through the linking of otherwise non-identifiable data. Since these linkages cannot, by definition, be foreseen and are of enormous benefit to research and innovation, the use of privacy tort law is suggested both to remedy harm caused by such privacy violations and to craft a body of case law that follows norms around digital data sharing.

Finally, privacy can be an overly restrictive concept, both legally and as a guiding principle for policy. Data ownership can be difficult to construe since many resources can create a useful dataset, and individuals may prefer to release what might be considered private information by some. In the structure of data collection and release today, such individuals have very little say over the future openness of their data. A sense of agency should be actively restored to permit individuals to share data.

Some of the concern about open data stems from the potential promulgation of misinformation as well as perceived privacy risks. In previous work I have labeled that concern 'Taleb's Criticism'.[53] In a 2008 essay, Taleb worries about the dangers that can result from people using statistical methodology without having a clear understanding of the techniques.[54]

An example of Taleb's Criticism appeared on UCSF's EVA website, a repository of programs for automatic protein structure prediction.[55] The UCSF researchers refuse to release their code publicly because, as they state on their website, "[w]e are seriously concerned about the 'negative' aspect of the freedom of the Web being that any newcomer can spend a day and hack out a program that predicts 3D structure, put it on the web, and it will be used." However, an analogy can be made to early free speech discussions that encouraged open dialog. In a well-known quote Justice Brandeis elucidated this point in *Whitney v. California* (1927), writing that "If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence." In the open data discussion this principle can be interpreted to favor a deepening of the dialog surrounding research, which is in keeping with scientific norms of skepticism and the identification of errors. In the case of the protein structure software, the code remains closed and a black box in the process of generating research results.[56]

Increasing the proportion of verifiable published computational science will stem from changes in four areas: funding agency policy, journal publication policies, institutional research policies, and the attitudes of scientific societies and researchers themselves. Although there have been significant recent advances from each of these four stakeholder groups, changing established scientific dissemination practices is a collective action problem. Data and code sharing places additional burdens on all these groups, from curation and preparation through to hosting and maintenance, which go largely unrewarded in scientific careers and advancement. These burdens can be substantial for all stakeholders in terms of cost, time, and resources. However, the stakes are high. Reliability of the results of our investments in scientific research, the acceleration of scientific progress, and the increased availability of scientific knowledge are some of the gains as we begin to recognize the importance of data and code access to computational science.

Acknowledgement I would like to thank two anonymous and extraordinarily helpful reviewers. This research was supported by Alfred P. Sloan Foundation award number PG004545 "Facilitating Transparency in Scientific Publishing" and NSF award number 1153384 "EAGER: Policy Design for Reproducibility and Data Sharing in Computational Science."

NOTES

1. Because of the wide scope of data considered in this article, the term *computational science* is used in a very broad sense, as any computational analysis of data.

See V. Stodden, “Resolving Irreproducibility in Empirical and Computational Research,” *IMS Bulletin*, November 2013, for different interpretations of reproducibility for different types of scientific research.

2. R. Gentleman and D. Temple Lang, “Statistical Analyses and Reproducible Research,” Bioconductor Working Series, 2004. Available at http://biostats.bepress.com/bioconductor/paper2/.
3. D. Donoho, A. Maleki, I. Ur Rahman, M. Shahram, and V. Stodden, “Reproducible Research in Computational Harmonic Analysis,” *Computing in Science and Engineering* 11, no. 1 (2009): 8–18. Available at http://www.computer.org/csdl/mags/cs/2009/01/mcs2009010008-abs.html.
4. For a discussion of the Copyright Act of 1976 see e.g. Pam Samuelson, “Preliminary Thoughts on Copyright Reform Project,” *Utah Law Review* 2007 (3): 551–571. Available at http://people.ischool.berkeley.edu/~pam/papers.html.
5. See Feist Publications Inc. v. Rural Tel. Service Co., 499 U.S. 340 (1991) at 363–364.
6. For a detailed discussion of copyright law and its impact on scientific innovation see V. Stodden, “Enabling Reproducible Research: Licensing for Scientific Innovation,” *International Journal for Communications Law and Policy*, no. 13 (Winter 2008–09). Available at http://www.ijclp.net/issue_13.html.
7. See V. Stodden “The Legal Framework for Reproducible Scientific Research: Licensing and Copyright,” *Computing in Science and Engineering* 11, no. 1 (2009): 35–40.
8. U.S. 17 Sec. 107.
9. This idea was suggested in P. David, “The Economic Logic of ‘Open Science’ and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer.” Available at http://ideas.repec.org/p/wpa/wuwpdc/0502006.html. For an analysis of the difficulty of an expansion of the fair use exception to include digital scholarly objects such as data see J. H. Reichman and R. L. Okediji, “When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale,” *Minnesota Law Review* 96 (2012): 1362–1480. Available at http://scholarship.law.duke.edu/faculty_scholarship/267.
10. A research *compendium* refers to the triple of the research article, and the code and data that underlies its results. See Gentleman and Temple Lang, “Statistical Analyses and Reproducible Research.”
11. See “Can I Apply a Creative Commons License to Software?” http://wiki.creativecommons.org/FAQ.
12. http://opensource.org/licenses/bsd-license.php.
13. http://opensource.org/licenses/mit-license.php.
14. http://www.apache.org/licenses/LICENSE-2.0.
15. Miriam Bitton, “A New Outlook on the Economic Dimension of the Database Protection Debate,” *IDEA: The Journal of Law and Technology* 47, no. 2 (2006): 93–169. Available at http://ssrn.com/abstract=1802770.
16. Feist v. Rural, 340. The full quote is “Although a compilation of facts may possess the requisite originality because the author typically chooses which facts to include, in what order to place them, and how to arrange the data so that readers

may use them effectively, copyright protection extends only to those components of the work that are original to the author, not to the facts themselves. . . . As a constitutional matter, copyright protects only those elements of a work that possess more than de minimis quantum of creativity. Rural's white pages, limited to basic subscriber information and arranged alphabetically, fall short of the mark. As a statutory matter, 17 U.S.C. Sec. 101 does not afford protection from copying to a collection of facts that are selected, coordinated, and arranged in a way that utterly lacks originality. Given that some works must fail, we cannot imagine a more likely candidate. Indeed, were we to hold that Rural's white pages pass muster, it is hard to believe that any collection of facts could fail."

17. See A. Kamperman Sanders, "Limits to Database Protection: Fair Use and Scientific Research Exemptions," *Research Policy* 35 (July 2006): 859, for a discussion of the international and WIPO statements of the legal status of databases.
18. For details on the CC0 protocol see http://creativecommons.org/press-releases/entry/7919.
19. V. Stodden, "The Scientific Method in Practice: Reproducibility in the Computational Sciences," MIT Sloan School Working Paper 4773-10, 2010. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193.
20. Ibid.
21. V. Stodden and I. Reich, "Software Patents as a Barrier to Scientific Transparency: An Unexpected Consequence of Bayh-Dole," Conference on Empirical Legal Studies, 2012. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2149717.
22. See National Institutes of Health, "Principles for Recipients of NIH Research Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources: Request for Comments." Available at http://www.ott.nih.gov/policy/rt_guide.html.
23. See Columbia University, *Faculty Handbook*, Appendix D: "Statement of Policy on Proprietary Rights in the Intellectual Products of Faculty Activity." Available at http://www.columbia.edu/cu/vpaa/handbook/appendixd.html (accessed August 21, 2013).
24. See Stanford University, *Research Policy Handbook*. Available at http://doresearch.stanford.edu/sites/default/files/documents/RPH%208.1_SU18_Patent%20and%20Copyright%20Agreement%20for%20Personnel%20at%20Stanford.pdf (accessed August 21, 2013).
25. "Leahy Urges Action to Ensure Access to Affordable Life-Saving Diagnostic Tests for Breast and Ovarian Cancer" (press release). See http://www.leahy.senate.gov/press/leahy-urges-action-to-ensure-access-to-affordable-life-saving-diagnostic-tests-for-breast-and-ovarian-cancer (accessed August 21, 2013).
26. See http://www.sdss3.org/dr10/.
27. See http://www.sdss3.org/dr10/data_access/, including http://skyserver.sdss3.org/dr10/en/help/docs/sql_help.aspx, http://skyserver.sdss3.org/CasJobs/, and http://skyserver.sdss3.org/dr10/en/help/browser/browser.aspx (accessed August 23, 2013).
28. "In Search of the Big Bang," *Computer Weekly*, August 2008. Available at http://www.computerweekly.com/feature/In-search-of-the-Big-Bang (accessed August 23, 2013).

29. “CERN Data Centre Passes 100 Petabytes,” *CERN Courier*, March 28, 2013. Available at http://cerncourier.com/cws/article/cern/52730 (accessed August 23, 2013). 100 petabytes is about 100 million gigabytes or 100,000 terabytes of data. This is equivalent to approximately 1500 copies of the Sloan Digital Sky Survey.
30. See D. H. Bailey, J. Borwein, and V. Stodden, “Set the Default to ‘Open’,” *Notices of the American Mathematical Society*, June/July 2013, available at http://www.ams.org/notices/201306/rnoti-p679.pdf, and V. Stodden, J. Borwein, and D. H. Bailey, “‘Setting the Default to Reproducible’ in Computational Science Research,” *SIAM News*, June 3, 2013, available at http://www.siam.org/news/news.php?id=2078.
31. For a complete discussion of HIPAA, see Chapters 1 (Strandburg) and 4 (Ohm) in this volume.
32. Some of these barriers were elucidated through a survey of the machine learning community in 2009. See Stodden, “The Scientific Method in Practice.”
33. E.g. “All results quoted in this paper are validated by using two independent sets of software tools. . . . In addition, many cross checks were done between the independent combination tools of CMS and ATLAS in terms of reproducibility for a large set of test scenarios” (from http://cds.cern.ch/record/1376643/files/HIG-11-022-pas.pdf).
34. Reichman and Uhlir proposed that contractual rules governing data sharing, for example providing licensing terms or compensating creators, create a knowledge “semi-commons.” A ‘semi-commons’ can exist through data pooling and thus sharing the burden of warehousing and supporting access infrastructure and tools, in exchange for increased access to the data. However, the concept of the ‘walled garden’ is slightly different in this example in that authorized independent researchers are given full access to the resources for verification and/or reuse purposes thereby mimicking open data as fully as possible under the privacy constraints inherent in the data. J. H. Reichman and Paul F. Uhlir, “A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment,” *Law and Contemporary Problems* 66 (Winter 2003): 315–462. Available at http://scholarship.law.duke.edu/lcp/vol66/iss1/12.
35. For an assessment of the reach of data and code disclosure requirements by journals, see V. Stodden, P. Guo, and Z. Ma, “Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals.” *PLoS ONE* 8, no. 6 (2013). Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0067111.
36. For a further discussion of such template agreements, see V. Stodden, “Innovation and Growth through Open Access to Scientific Research: Three Ideas for High-Impact Rule Changes,” in *Rules for Growth: Promoting Innovation and Growth through Legal Reform* (Kansas City, MO: Kauffman Foundation, 2011). Available at http://www.kauffman.org/~/media/kauffman_org/research%20reports%20and%20covers/2011/02/rulesforgrowth.pdf.
37. Of course, researchers in private sector for-profit firms are not the only potential collaborators who may have a different set of intentions regarding data and code availability. Academic researchers themselves may wish to build a start-up around

the scholarly objects deriving from their research, for example. In a survey conducted by the author in 2009, one senior academic wrote he would not share his code because he intended to start a company around it. See Stodden, "The Scientific Method in Practice."

38. See http://nlp.stanford.edu/software/.
39. See http://www.whitehouse.gov/the_press_office/President-Obama-Names-Vivek-Kundra-Chief-Information-Officer/ (accessed September 1, 2013).
40. See "Expanding Public Access to the Results of Federally Funded Research," http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research (accessed September 1, 2013).
41. For other examples see e.g. J. E. Wiley and G. Mineau, "Biomedical Databases: Protecting Privacy and Promoting Research," *Trends in Biotechnology* 21, no. 3 (March 2003): 113–116. Available at http://www.sciencedirect.com/science/article/pii/S0167779902000392.
42. See e.g. https://data.cityofchicago.org/ and http://schoolcuts.org, http://nycopendata.socrata.com/ and https://data.ny.gov/ (state level), and https://data.sfgov.org/. The Family Educational Rights and Privacy Act (FERPA) attempts to address this with a notion of student privacy; see http://www.ed.gov/policy/gen/guid/fpco/ferpa/index.html. The State of Oklahoma recently passed a bill, the Student DATA Act, to protect student school performance data; see http://newsok.com/oklahoma-gov.-mary-fallin-signs-student-privacy-bill/article/3851642. In New York State a case was filed in 2013 to prevent a third party from accessing student data without parental consent; see http://online.wsj.com/article/AP0d716701df9f4c129986a28a15165b4d.html.
43. See e.g. the report from the World Economic Forum, "Unlocking the Value of Personal Data: From Collection to Usage," (Geneva, 2013). Available at http://www.weforum.org/reports/unlocking-value-personal-data-collection-usage.
44. See e.g. D. Citron, "Mainstreaming Privacy Torts," *California Law Review* 98 (2010): 1805–1852.
45. See e.g. McCormick v. Haley, 307 N.E.2d 34, 38 (Ohio Ct. App. 1973).
46. Restatement (Second) of Torts § 652B (1977).
47. See Citron, 1809.
48. Citron; e.g. Kline v. 1500 Massachusetts Ave. Apartment Corp., 439 F.2d 477, 480–81 (D.C. Cir. 1970), holding the landlord liable for a poorly secured building when tenants were physically beaten by criminals.
49. See e.g. F. Berman and V. Cerf, "Who Will Pay for Public Access to Research Data?" *Science* 341, no. 6146 (2013): 616–617. Available at http://www.sciencemag.org/content/341/6146/616.summary.
50. Individuals may direct their data to be used for research purposes only, or to be placed in the public domain for broad reuse, for example. See e.g. Consent to Research, http://weconsent.us, which supports data owner agency and informed consent for data sharing beyond traditional privacy protection.
51. World Economic Forum, 12.
52. Some restrictions on subject agency exist; see e.g. Moore v. Regents of University of California 51 Cal.3d 120 (Supreme Court of California July 9, 1990). This case dealt with ownership over physical human tissue, and not digital data, but

the tissue could be interpreted as providing data for scientific experiments and research, in a role similar to that of data. See also the National Institutes of Health efforts to continue research access to the Henrietta Lacks cell line, taking into account Lacks family privacy concerns. E. Callaway, "Deal Done over HeLa Cell Line," *Nature News*, August 7, 2013. Available at http://www.nature.com/news/deal-done-over-hela-cell-line-1.13511.

53. V. Stodden, "Optimal Information Disclosure Levels: Data.gov and 'Taleb's Criticism'," http://blog.stodden.net/2009/09/27/optimal-information-disclosure-levels-datagov-and-talebs-criticism/.
54. N. Taleb, "The Fourth Quadrant: A Map of the Limits of Statistics," http://www.edge.org/3rd_culture/taleb08/taleb08_index.html (accessed September 1, 2013).
55. See http://eva.compbio.ucsf.edu/~eva/doc/concept.html (accessed September 1, 2013).
56. See e.g. A. Morin, J. Urban, P. D. Adams, I. Foster, A. Sali, D. Baker, and P. Sliz, "Shining Light into Black Boxes," *Science* 336, no. 6078 (2012): 159–160. Available at http://www.sciencemag.org/content/336/6078/159.summary.