ROUTLEDGE

Research Ethics in Applied Economics

A Practical Guide

Anna Josephson and Jeffrey D. Michler

5

DATA MANAGEMENT

5.1 Introduction

In the mid-1990s, the Group Insurance Commission (GIC) in Massachusetts released data on approximately 135,000 state employees and their families (Lasalandra, 1997). The GIC was responsible for purchasing health insurance for state employees, and the goal of making these data public was to allow researchers in academia and industry to design and recommend improvements to the state health care system. The release of individual medical data made many people nervous regarding their personal privacy. Dolores Mitchell, Executive Director of GIC, worked to reassure people, stating, "[M]y own records are in there and I care very deeply about security" (Lasalandra, 1997). Prior to the release of the data, GIC made sure to remove what it considered personally identifiable information (PII), such as names, exact addresses (though not ZIP Codes), and Social Security numbers.

Latanya Sweeney, who at the time was a graduate student in computer science at MIT, doubted GIC's assurances of data confidentiality.[1] She purchased a voter registration list for Cambridge, Massachusetts, for $20, which contained information such as individual names, addresses (including ZIP Codes), birth date, and gender (Sweeney, 2002). As the governor of Massachusetts at the time, William Weld had medical records in the GIC data. Weld's birth date, ZIP Code, and gender were public information. Using the voter registration data, Sweeney identified six people in Cambridge with Weld's birthday, three of whom were men, and he was the only one who lived in his ZIP Code. Knowing that Weld was the only man with his birthday in his ZIP Code allowed Sweeney (2007) to search the GIC data and identify Weld's medical records. While GIC had promised to anonymize the data before making it public, Sweeney was able to reidentify that data by combining two different data

DOI: 10.4324/9781003025061-8

sets. This process is known today as a reidentification attack. Once the news of Sweeney's reidentification attack broke, Joseph Heyman, President of the Massachusetts Medical Society, concluded that "there is no patient confidentiality. It's gone" (Lasalandra, 1997).

While the story of Sweeney's identification of Weld is sometimes identified as an extreme example, the opportunity for privacy violations in data is widespread – and the chance to identify people, as Sweeney identified Weld, are hardly uncommon. This is particularly true in the social sciences. As applied economists, we rely on microeconomic data that describe the choices, actions, and attributes of individuals, households, and firms. Ethical norms, and in many cases legal codes, require the de-identification of data when using or sharing it. Laws in many countries, and ethical review boards at universities, require that PII be kept confidential. But making data truly anonymous requires more than just removing names or addresses from a data set, as Sweeney demonstrated by identifying Weld by linking only two data sets.

Another similar reidentification attack was carried out in 2006 by Arvind Narayanan, then a graduate student, and Vitaly Shmatikov, a professor, at the University of Texas at Austin (Narayanan and Shmatikov, 2008).[2] In 2006, Netflix released data from its subscriber database on users' ratings of movies. In order to protect privacy, Netflix removed direct personal identifiers and released only a random sample of all subscribers and only a random sample of the ratings from a given subscriber. The data release was part of the Netflix prize, a $1 million prize for the individual or team that could improve upon Netflix's own algorithm for recommending new movies to subscribers based on a subscriber's past ratings.

The Netflix data contained only four pieces of information: an anonymous user identifier, the movie, the date of the grade, and the grade. Narayanan and Shmatikov combined these data with public data from the movie rating site IMDb.com, which contained a user's IMDb user name, the movies they have watched, their ratings, and the date of the rating. With this combined data set, Narayanan and Shmatikov were able to use an individual's public IMDb rating to identify them and reveal all their ratings in the supposedly private Netflix data. While a person's tastes in movies might not seem to be vital private information, one can imagine someone wanting to keep private their love (or hate) of movies involving topics such as queerness religion, politics, race, and the like.

The objective of this chapter is to understand the challenges of managing and sharing data in an ethically responsible way. We start by discussing the need for data sharing in order to achieve the goal of open science to make research transparent and reproducible. Many researchers are reluctant to share data and code and so the profession needs to consider ways in which to reward data sharing. We then turn to the different types of data that economists use and the unique threats each data type posses to privacy and confidentiality.

The remainder of the chapter shifts focus to the actual how of data sharing and, in particular, the measures that are necessary to ensure shared data preserves the anonymity and confidentiality that is promised to research participants on

informed consent forms. What the stories of Weld and the Netflix Prize illustrate is that "removing identifying information is not sufficient for anonymity" (Narayanan and Shmatikov, 2008). A data base administrator or a data set's author must take into account the current and future existence of other data that could be combined in a reidentification attack. This insight has lead to a variety of approaches to ensuring privacy protection in public use data sets. However, the anonymization of data comes with a cost: the more anonymous a data set (the more privacy is protected), the more statistical accuracy is lost. There is a trade-off between protecting privacy and sharing accurate data that social scientists are just now beginning to grapple with.

5.2 The Need for Data Sharing

In 2020, a group of researchers lead by Ariella Kristal attempted to replicate a now-retracted three-study paper (two lab, one field experiment) by Shu et al. (2012). The original study (ironically) was about people's honesty in reporting to a third party in a context where the third party is unlikely to verify honesty. Examples include tax documents and insurance policies, where individuals self-report information and sign their name. The retracted Shu et al. (2012) paper tests for differences in honesty, depending on whether an individual is asked to sign their name at the beginning or end of a document. They report simply having people sign at the beginning of the document increases honest reporting. The replication by Kristal et al. (2020) included all five of the original authors and attempted both direct and conceptual replications of the two lab experiments. None of the results from the original held up under the scrutiny of the replications.

The story could have ended there as an excellent example of ethical research. The original authors conducted several experiments and published their findings. These original authors then engaged with other authors to replicate the original results, and when those original findings did not hold up, they published their new results. The replicating authors write, "Given the policy applications of this result, it is important to update the scientific record regarding the veracity of these results" (Kristal et al., 2020).

However, the fact that there was no attempt, or reported attempt, to replicate the findings of the field experiment in the retracted study raised questions among some researchers.[3] The field experiment involved an auto insurance company that requested policyholders to report the current odometer reading on their insured vehicles. Individuals were randomized into a "sign at the top" or "sign at the bottom" reporting form. In the original 2013 publication, none of the data for the three studies were made available as part of the publication. But all the data were made available as part of the 2020 publication. This allowed a group of researchers unaffiliated with either of the 2013 or 2020 study to take a look at the data. As reported in Simonsohn et al. (2021), several anomalies emerged from just examining the raw data from the field experiment. There were implausible distributions of values, no rounding in self-reported values, and near-duplicate values. These anomalies lead the new group of

researchers to conclude that the original data were fabricated. Based on the evidence in Simonsohn et al. (2021), the editors of *Proceedings of the National Academy of Sciences* retracted Shu et al. (2012). It is still unclear who fabricated the data: whether it was the authors who ran the experiment, an individual in the authors' lab, or someone at the insurance company. What is clear is that without the data being shared as part of Kristal et al. (2020), the fraud would not have been discovered.

Communality is one of the scientific norms discussed in the opening chapter. As opposed to secrecy, communality calls for the open sharing of research methods, practices, and findings. It is the foundation of open science. Open science not only contributes to efficiency in producing new scientific research by avoiding duplication and promoting access, but it also makes science more robust by allowing for the identification of bad science and, in extreme cases, falsified results. Data sharing is an integral component of communality to facilitate improvement in the quality and practice of science.

For a data-driven science, like applied economics, there are two key components that must be shared to ensure transparency and reproducibility. The first is data, and the second is code. Code sharing has become relatively common due to journal policies and websites like GitHub. But the sharing of data, beyond the final, cleaned, regression-ready data, remains relatively uncommon for a variety of reasons that we discuss in this section.

5.2.1 Current Practices are Insufficient

As recently as 2010, one could find at least one book on research ethics in social science that encouraged the destruction of data after publication. The justification is that preserving data over time increases the probability that someone other than the original researcher(s) may gain access to the data and be less scrupulous in preserving subject privacy. Oliver (2010) writes,

> Generally speaking it is not necessary to store all of the raw data from a research study, once that study has been written up as a thesis or as a journal article. . . . One might argue that there could be the necessity for another researcher to reanalyse the data in order to confirm the results, and that this is a justification for data storage. However, this could be achieved shortly after the first analysis, thus removing the necessity to store the data. It is possible for another researcher to replicate the research design and to collect more data in a comparable context.
>
> *(p. 90)*

The American Economic Association (AEA) and the Econometric Society instituted data availability policies for their journals in the mid-2000s, so this view was clearly in decline by 2010. But for many years enforcement of data-sharing policies at economics journals was lax.

Perhaps the earliest attempt to replicate results from a set of studies published in a single economics journal was Dewald et al. (1986). The paper was part of a project,

funded by the National Science Foundation (NSF), to reproduce existing research that had been published in the *Journal of Money, Credit, and Banking* and improve data availability and transparency. Prior to the NSF grant, the journal's policy established a norm of data sharing by asking authors to make available data upon request. To facilitate the study, the journal's editors adopted a new policy requesting data and code at time of publication. Dewald et al. (1986) were only able to get data from 34 percent of papers published prior to the switch in policy, despite the allegedly established norm of data sharing. Even after the change in policy, Dewald et al. (1986) could only obtain data from 78 percent of papers – a marked increase – but far from full compliance with the explicit policy that data needed to be shared at time of publication.

Though Dewald et al. (1986) published their results in the *American Economic Review* (*AER*), not much changed in terms of data sharing in economics for more than two decades (Christensen et al., 2019). It was not until McCullough and Vinod (2003) attempted a similar replication exercise, this time using papers published in a single issue of the *AER*, that editors at other journals began to take notice. McCullough and Vinod (2003) were only able to obtain data from four of eight papers, despite the *AER* having a policy that data be made "readily available to any researcher for the purpose of replication" (Bernanke, 2004). As a result of authors failing to comply with the norms of data sharing, the editors at the *AER* updated the data-sharing policy to require, as a condition of publication, data and code sufficient to replicate results. With the change in policy at the *AER*, most other economics journals, especially those published by associations, followed suit.

However, as is often the case when it comes to ethical concerns, requirements diverge from practice. Stodden et al. (2018) examine 204 papers that were published in *Science*. They find only partial data availability and resistance to sharing when contracted by authors to request data. by authors, when contacted. This results in a somewhat amusing set of excerpts shared from emails with corresponding authors. The authors include 204 papers, which were published in *Science*. Of the authors contacted, 36 percent shared some material, 11 percent were unwilling to provide data or code without "information regarding intentions," 11 percent asked the authors to contact someone else who worked on the article, 7 percent refused, and 2 percent gave reasons they could not ethically or legally share the requested information. Stodden et al. (2018) is well worth reading, if only for the condescending and shockingly ill-informed responses from some of the authors.[4] Consider:

> I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

Or

> When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

Both of these are absurd responses, given the policy of *Science* to make available data and code sufficient for replication purposes upon request after publication. Ultimately, Stodden et al. (2018) conclude that fieldnorms are essential to incentivize data-sharing behavior, suggesting that journal policies may be insufficient in practice, as adherence to the policies remain imperfect.

Further studies support the findings in Stodden et al. (2018). Tedersoo et al. (2021) evaluate data availability in research articles in nine disciplines (biomaterials and biotechnology, ecology, forestry, humanities, materials for energy and catalysts, microbiology, optics and photonics, psychology, social sciences) in *Nature* and *Science*, both of which have requirements for data sharing. The authors find that there was only partial data availability for about 55 percent of papers of the 845 considered. Upon contacting authors, partial availability improved to nearly 70 percent. Tedersoo et al. (2021) report that a number of authors declined to share their data, citing issues of data sharing as well as "certain aspects of [the] request."

Within economics, data-sharing policies are still developing and, for the most part, are not yet mandated, as is the case in *Science* and *Nature*. As of 2022, the journals of the Agricultural and Applied Economics Association (AAEA) still framed/index-framed their data-sharing policy in normative terms: "Authors are expected to submit their datasets and associated documentation. . . . Authors are encouraged to comply with all of this policy, but the editors would prefer partial compliance over non-compliance" (AJAE, nd). In 2015, we along with a set of collaborators attempted to replicate studies in *Econometrica* and the *Journal of Development Economics*, both journals that we considered as having strong data-sharing policies at the time. Yet, we found the data were not available, despite the stated policies. Working with the editors, we were eventually able to get the data from the authors, but the lax enforcement signals that current practices, standards, and incentives for data sharing are insufficient in economic journals, just as in the general-interest science journals. The rules must be enforced before norms will change.

The AEA has recognized this as a problem and in 2018 hired a Data Editor for its journals and updated its Data and Code Availability Policy (AEA, 2020). The new policy states that papers will be published only after data and code have been submitted to the journal and the Data Editor has verified that the results do replicate with the materials provided. To our knowledge, the AEA policy is the most forceful policy among economic journals as it no longer relies on norms for data sharing but mandates data sharing as a condition of publication.

While we admire the AEA for its strong policies and we hope more journals adopt a similarly strong data-sharing policy, it is instructive to consider how other journals might implement such a policy. Successful implementation requires two things: desirability and enforcement. First, for effective implementation, journals must be desirable enough that authors are willing to provide data as a condition of publication. It is costly for authors to construct replication

packages for data sharing. If two journals, Journal *A* and Journal *B*, have similar impact factors but only *A* requires data sharing, authors might forgo publication in *A* for publication in *B*.[5] To put it simply, the AEA is able to implement such a strong data-sharing policy because there is extremely high demand to publish in one of the AEA's journals.

Second, for effective implementation, there must be someone verifying and enforcing compliance. The AEA has employed a Data Editor who has hired a team of assistants to run replication code on the shared data and verify the code produces the results in the paper. This might seem like a manageable task, but as the current AEA Data Editor, Lars Vilhuber, reports, the task in fact is enormous, time-consuming, and costly (Vilhuber, 2020). The median size of replication packages his team has received is 2MB, while the size at the 3rd quartile is 30MB and the max is 30GB. The number of files in each replication package is also substantial, with the median number of files at 13 and the 3rd quartile at 39. The median time that it takes an undergraduate research assistant to run the code for one paper and write up a report about the replicability is 15–20 hours (Vilhuber et al., 2022). Few professional associations or journals will have the resources to ensure compliance with forceful data-sharing policies. These implementation issues mean that we as a profession cannot rely on changes in journal policy alone to increase data sharing among researchers. We must also work to understand researchers' reluctance to share data so that we can better encourage data sharing within the profession.

PERSPECTIVE 5.1 DATA TRANSPARENCY BY LARS VILHUBER

The availability of data that underlies research in our papers is key for transparency. When we collect our own data via surveys or work hard to combine many different data sources into a cohesive new data set, the resulting object – let's call it a database – should be made broadly available, because it is under our control to do so. But not all data that we create or reuse can be archived in an open trusted repository. The data may be considered sensitive – now or in the future – and we may not have been given the rights to "post" the data. I will illustrate this with two examples, based on recent articles in economics journals.

In particular, I want to consider here the case when sharing might be considered problematic even when the original data used by the authors is, or was, public. In the U.S., much information surrounding the criminal justice system is public. For instance, when a person is arrested, the names of the officer, the arrested person, and the charge are all public records, even if the person is subsequently cleared of the charge or if the charge is dropped. This

information can be scraped from public websites or can be requested via Freedom of Information Act (FOIA) requests. The information is thus public, sometimes subtracting it from the purview of IRBs and leaving it up to the authors to decide what to do with it. However, that does not make the data nonsensitive. Arrest records might be expunged en masse – erased from the database – with a legal requirement for others who may have copied the data to do the same, including researchers. Thus, what was once public information now becomes illegal to retain. This is the issue faced by Ouss and Stevenson (Forthcoming), who therefore cannot make data available that was public when they first scraped it, since mass expungement laws have made large parts of the data they collected illegal to disseminate. Similar constraints may be imposed on data holders through various "rights to be forgotten" laws, such as the European General Data Protection Regulation or the California Consumer Privacy Act.

However, even when the data can be preserved, what researchers do with it might make it more sensitive than most are comfortable simply "posting." Consider Goncalves and Mello (2021a), who took data on police officers obtained via Freedom of Information Act, and "estimate the degree to which individual police officers practice racial discrimination." In other words, they estimate how racist a particular police officer is, based on demographics and observed behavior, where every police officer can be identified from the public records by name.

So what can authors do in these cases? Preferably, they can still preserve the data they had collected and make it available to other researchers on a more restricted basis. For instance, use may be restricted to verifying that the authors' calculations are correct and limited to academic researchers. While such data sharing no longer qualifies as "open" data sharing, it does enable transparency and the verification thereof. In the case of Goncalves and Mello (2021a), the authors deposited their data at openICPSR (Goncalves and Mello, 2021b), with the restriction that researchers wishing to use the restricted data be affiliated with academic institutions, have a data security plan, and demonstrate IRB approval. The authorization process is, for better or worse, out of their hands: neither can the authors withhold approval based on personal bias, nor can they speed up the process or influence its parameters.

A more prosaic workaround when even restricted archiving of data may not be possible is to keep robust personal archives. Journal policies, like the American Economic Association's Data and Code Availability Policy (AEA, 2019), require that authors maintain an archive of their materials when the data cannot be made public, for a minimum of five years. Authors might want to investigate to what extent their own (academic) institution provides mechanisms to preserve robustly such archives, through internal mechanisms or institutional subscriptions to cloud providers.

5.2.2 Why Researchers Do Not Share Data

In 2014, the academic publisher Wiley surveyed 2,250 researchers across disciplines to find out why they were reluctant to make their scientific data publicly available. Researchers responded with over a dozen reasons for why they were hesitant to share their data (Ferguson, 2014). In a well-timed article in *The Atlantic* titled "Scientists Have a Sharing Problem," journalist Maggie Puniewska distilled the many reasons given by researchers for not sharing data down to two: (1) competition and (2) disorganization (Puniewska, 2014). Some of the fears about sharing data are valid, some are self-serving, and some reflect common misconceptions. But we agree with Puniewska (2014) that almost all reasons for not sharing data are typically due to fear of competition from other researchers or disorganization by the researcher.

A common reason given by researchers for not sharing data is the fear of being scooped or outcompeted in making a new discovery (Ferguson, 2014). In economics, this fear amounts to being afraid to share data until the researcher has wrung all the useful information from it. Imagine a researcher collects household survey data containing a variety of information on education, livelihood, and health outcomes. They publish an initial paper, on education, and shares the data per journal policies. The researcher then starts to get to work on analyzing livelihood outcomes. But a second researcher, interested in livelihood outcomes, comes in, takes the shared data, analyzes it, and publishes a paper before the initial research can. The profession is indifferent to who wrote the paper (assuming both livelihood papers were of equal quality). But the initial researcher has expended money, time, and effort only to get scooped before they could fully use the data that they collected. Given that the initial research has not been able to fully internalize the benefits of publication (while fully internalizing the costs of data collection), a natural response would be to stop collecting data. This reaction, the suspension of new data collection efforts, is frequently mentioned in editorials and opinion pieces. For example, Longo and Drazen (2016) worry that science will be taken over by "research parasites" while Gibson (1995) warns of "data vultures" feeding on the data of others. While an individual researcher's decision to stop collecting new data because they cannot fully internalize the benefits of their work seems plausible, it is hard to know how much credence to give to predictions of general equilibrium outcomes in which no new data are collected, particularly in an environment in which so much data are collected with the express purpose of public dissemination for widespread use (e.g., the World Bank Living Standards Measurement Survey (LSMS) surveys or the USAID Demographic and Health Surveys (DHS)).

The second category under which most reasons for not sharing data fall is disorganization. We ourselves fell into this category early in our careers. When we were grad students or young faculty members mostly working on data on our own, our only incentive for organization was to ensure we were efficient in our work.

For the most part, no one was asking to see our data as we wrote our dissertations, and so we kept things only as organized as was necessary to know for ourselves what we were doing. This leads to a pretty clear causal chain in which a researcher keeps their files organized in a way that is only clear to them, and when someone does request the data, the disorganization (or the time required to create order for an outsider) is reason enough to deny the request.[6]

A related reason for not sharing data is a fear of the failure of privacy protection (Sardanelli et al., 2018). While this concern is certainly admirable and valid, it is also a symptom of disorganization because, as we discuss later, there are numerous ways to guarantee privacy – they just require the researcher be organized enough to implement them. Researchers will also claim, with good reason, that it is too costly to overcome the technical barriers, such as data conformity and documentation, to share data. Again, this reason is a symptom of disorganization in the research, or rather the profession's failure to incentivize the researcher to improve on organization.

There are two additional reasons, commonly given, for why researchers do not share data that do not fall into the competition and disorganization categories (Christensen et al., 2019). The first reason is because the researcher has committed some form of research misconduct and they do not want to be found out. Simonsohn (2013) reports using statistical methods to detect what he believed were three cases of fraudulent data. In two cases, the authors eventually made the data available, and Simonsohn was able to confirm fraud. In the third case, the author claimed to have lost the data, leaving Simonsohn unable to confirm if the fraud had in fact occurred. A second reason why a researcher might not share data is because they lack the ownership of the data or the authority to share the data. Economists frequently use administrative data from firms or population data from government agencies. In the former case, the firm may not want to allow the data to be released, and as the researcher does not own the data, there may be little they can do. For the latter, governments often require researchers to access restricted use data by physically traveling to a secure site to use the data. In these cases, often involving Census or Social Security data, the government's need to protect the privacy of its citizens overrides a researcher's desire to contribute to open science. In terms of these final two reasons for not sharing data, there is little one can do to encourage data sharing. But for the vast majority of reasons given by researchers to keep their data private, most can be addressed by changes in a researcher's habits and in the expectations of the profession.

5.2.3 Ways to Encourage Data Sharing

During the 2012 U.S. presidential election, Barack Obama delivered a speech on infrastructure in which he said, "Somebody invested in roads and bridges. If you've got a business, you didn't build that." Obama's opponent, Mitt Romney, used the ambiguity regarding what "that" referred to to claim Obama was telling

business owners that they had not built their own firms, inflaming the feelings of many entrepreneurs. Similarly, researchers often think in terms of "their" data, as if they built that data set on their own without any contributions from others. But in reality, building data sets relies on contributions in money, time, effort, and expertise from many people and institutions. Most data collection is funded with grants from the government or private foundations. Students or staff members frequently vet surveys before use, performing quality and logic tests. Enumerators deliver the survey and collect the actual data. Most important, hundreds to thousands of individuals and/or firms consent to having their data collected and used by a researcher. This is not to minimize the researcher's own contribution or to say that the researcher should not feel a sense of ownership of the data. Rather, it is to highlight the collaborative nature of data collection. If the public (i.e., funders, students, research participants) helped contribute to building the data set, then we as "the" researcher should feel some obligation to repay this contribution. One way is by sharing data.

Relying on a norm for data sharing based on a sense of obligation to those who helped contribute to building a data set is insufficient to overcome the reluctance of many researchers to share data (Tedersoo et al., 2021). Thus, we must carefully consider mechanisms to incentivize data collect that overcome the reluctance that exists on the part of the researcher. In the previous section we categorized the reasons for reluctance into competition and disorganization. We also discussed two reasons for reluctance that do not fall into our categories (i.e., fear of being caught in fraud, lack of control of data). For these latter two reasons, incentive schemes to encourage data sharing are unlikely to have much of an impact. But for the other reasons, there are practices that the profession can implement to address the reluctance and make researchers more comfortable with and more likely to share data.

In addressing the fear of competition from data sharing, journals are likely the best institution to provide appropriate incentives. This is because the competition that researchers fear is that other researchers will be able to exploit the data first, publish the results in a superior journal (due to primacy), and capture more citations. One way journals could address researchers' fear of being scooped is to require that authors only release data extracts. A data extract is the portion of the complete data set that is required to produce the analysis in the published paper. In fact, this is what most journals require in their data policy statements. The journal is concerned with ensuring the results published in their journal can be replicated, and thus, their focus, to date, has been on only requiring data sharing that is sufficient to alleviate their concern. Asking authors to submit a data extract, instead of the complete data and cleaning code, incentivizes data sharing because a competing research is unlikely to be able to publish a new paper using only the variables present in an existing paper. In our example of the researcher with household survey data on education, livelihoods, and health outcomes, the data extract accompanying the published paper on education would not need

to contain the variables on livelihood and health outcomes. This allows the researcher to then develop the second paper on livelihood outcomes without fear of being scooped. The trouble with journals only requiring data extracts is that it does not allow for the profession to determine if published results are due to *p*-hacking or specification search. If the data extract contains only the variables used in the final analysis, there is no way to determine what other specifications or what other variables were tried in the early analysis. If the goal of open science is to reduce the incentives to commit fraud by ensuring transparency and reproducibility, then the use of data extracts can only be considered half open science. This is one reason why the AEA's new Data and Code Availability Policy requires authors to submit the raw data as well as "the programs used to create any final and analysis data sets from raw data" (AEA, 2020).

An alternative to overcome the reluctance to share data due to competition is a data embargo (Christensen et al., 2019). Data embargoes are a period of time under which the original data is kept private. This provides the research with a time-limited "patent" to the data in which they have exclusive use of that data. The National Institutes of Health (NIH) in the U.S. has established a central repository for genomics data generated by the grants they award. This database, called the database of Genotypes and Phenotypes, has an embargo period, usually one year, in which researchers have exclusive use of the data generated from their grant. This incentivizes researchers to collect new data without fear that "data vultures" will extract all value from the data before the original researchers have a chance to exploit it. The embargo period also incentivizes researchers to get results out quickly instead of letting the data sit on a computer unused – a value to the granting agency. Embargoes are already in use by the AEA and the Open Science Foundation (OSF) for pre-analysis plans. On the AEA RCT Registry, some information about the trial is immediately made public while the researcher can choose to keep other information hidden until the trial is completed. This feature was built into the registry to calm fears that if all the information on a trial was made public at the time of registration, other researchers could come in and scoop the original researcher. A similar feature is available in OSF's registry. While we are unaware of any economics journals that currently implement a data embargo, one could imagine a simple policy whereby replication material is required at the time of publication and then links to the material appear at a set time after publication.

A third way that journals could encourage data sharing by researchers is to establish policies regarding data citation. If researchers fear that sharing data will lead to no new data collection, then one way to address this fear is for journals (and the profession) to reward new data collection. While data citation is still uncommon in economics, the field is changing. Part of the AEA Data and Code Availability Policy is that "[a]ll source data used in the paper shall be cited" (AEA, 2020).[7] One reason for the change of citation standards in economics is due to the introduction of the Transparency and Openness Promotion (TOP) Guidelines from the Center for Open Science (COS), the same group the runs the OSF. Nosek et al.

(2015) introduce the idea of TOP guidelines and establish three levels of journal compliance with these guidelines.[8] The guidelines include establishing citation standards for data, code, and research materials so that the intellectual contribution of those involved in their creation can be recognized and rewarded. TOP provides a sample data citation, as does the AEA Data and Code Availability Policy.

> Romer, Christina D., and David H. Romer. 2010. "Replication data for: The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." *American Economic Association (AEA)* [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E112357V1.

A data citation, like any citation, should include author, date, title of the data set, and publisher. In the case of data, the publisher is the location of the data, either in a permanent repository and with some data distributor.

While journals can provide the best mechanisms for overcoming fear of competition, funding organizations have the best institutional tools for overcoming the issue of disorganization. Disorganization arises as a justification for not sharing data because if no one requires sharing reproducible data and code then researchers are unlikely to be more organized than is absolutely necessary for them to complete their own work. Unlike AEA institution journals, most journals will not have the capacity to hire a data editor to verify replicability of data and code. That leaves granting agencies (both private and public) as the institutions that can provide the strongest incentives to encourage researchers to get organized. After publications and citations, generating external funding is typically the most important criteria for promotion and tenure in academia. Much like the NIH requires filing of pre-analysis plans for drug trials and the sharing of data, large governmental and non-governmental agencies could require, as a condition of funding, that recipients develop and execute a clear and organized data management and data availability plan. The agencies would not need to verify that all award recipients adhered to their plans but could engage in random audits to help reduce the incentive to "cheat" on their data management plans.

While funding agencies could provide a strong incentive to get organized, that only partially addresses the problem. There remains the learning of organizational skills. Here, funding agencies could allow researchers to build in some cost for training in methods or skills. Additionally, professional agencies and organizations, like the Berkeley Initiative for Transparency in Social Sciences (BITSS), can (and do) offer workshop and training in organizing data. The combination of pressure from funding sources to obtain training, the availability of training from professional organizations, and mainstreaming open science can help lower the hurdle that disorganization poses to sharing truly replicable and reproducible data and code. To facilitate movement toward this goal, the remainder of the chapter focuses on best practices regarding data management, privacy protection, and creating replicability packages.

5.3 Data Storage and Management

Regarding many practices in empirical science, economics tends to follow the leads and/or mandates of the health sciences. This is true for the adoption of randomized control trials, pre-analysis plans, and IRB. Thus, as economists look to improve standards around data management and data sharing, we can look to current practices in the health sciences. The trends in these fields can help us to understand where our own field is likely headed.

Much of the rules and regulations governing health science research in the U.S. comes from the NIH. The NIH (2020) defines scientific data as "the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications." This definition indicates that data must be properly managed and shared beyond what is just necessary for publication, that is, more than just a data extract. In economics, we use scientific data generated by a wide variety of sources, at various levels of aggregation, and with differing needs for confidentiality. Managing the various types of data can be complicated and a natural starting place, when feeling overwhelmed by the task of managing data, is to develop a data management plan (DMP). Such a plan need not be anything formal (unless required by a donor or IRB). At its most basic, a DMP simply provides concrete and actionable answers to the questions laid out in Box 5.3.

BOX 5.3 DATA MANAGEMENT PLAN

1. What type of data will be collected? What is the unit record or unit of observation? How many observations?
2. How will the data be collected? Where will it be stored during the collection process?
3. Where and how will the data be permanently stored? How will backups be provided?
4. How will the data be organized? What is the folder structure? In what file formats?
5. Who gets access to the data? When? How will access be managed?
6. What data will be archived after the analysis is complete? Where and for how long?
7. Under what conditions will the archived data be made available to others? When? Under what license?
8. Who owns the data? Who is responsible for the management of the data, particularly archived data?
9. Are there costs associated with the data management?

DMPs have long been a component of grants required to the U.S. government, including the NIH. In October 2020, the NIH took the requirements for data further, issuing their Policy for Data Management and Sharing. Three things are important about this document. First, it applies to all work done using NIH funding. Previously, the NIH policy had only applied to "large" grants and genomics data. Second, the plan defines not just a data-sharing policy but a data management policy. As discussed earlier in this chapter, a primary reason given by researchers for not sharing their data is the disorganization of the data. The new NIH policy directly addresses this by defining how data should be managed so that it can more easily be shared. Last, the policy provides a strong incentive for compliance. The NIH, as the granting agency, will determine a researcher's compliance with the stated policy, and failure to conform to the policy puts future NIH funding to the recipient institution in jeopardy. Similar to the U.S. Government's requirement for IRB, the failure of individual researchers to comply puts funding for the entire institution at risk. These latter incentives are the types that, as IRBs have shown, can create real, institutional change, at least with respect to stated compliance.

5.3.1 Individual, Personal, or Firm Data

The most common types of data used by economists are individual, personal, or firm data. If these data were generated by a third party as administrative data or as public use data, then the data publisher controls the process of data management and data anonymization. But applied economists frequently engage in their own data collection, generating their data sets through surveys. In this case, the researcher is responsible for data storage and management.

Assuming that the researcher is at a academic or research institution, the researcher will have obtained IRB approval prior to data collection. As part of the approval process, the IRB will typically require the researcher to define a data storage plan. The IRB's interest here is solely in data storage to ensure the privacy of research participants and the confidentiality of their data.[9] In accordance with these policies, two things must be considered:

1. Is the data "sensitive"? Sensitive data require more attention, for example, physical security and/or encryption.
2. How vulnerable are data storage resources? Are there sufficient protections and redundancies to ensure that the data are kept confidential?

Sensitive data are any data where its disclosure could have adverse consequences for the research participant, including not only financial and educational harm but also reputational harm or risk of criminal or civil liability. For economists, who (outside of the lab) collect data on personal or firm income, assets, and consumption, almost all data are considered sensitive.

Prior to the revolution in information and communication technology (ICT), most data were collected on paper and then transcribed to a computer. Security of data focused on the physical storage of surveys or records that might be locked in a filing cabinet or an office. Nowadays, almost all data are collected and stored electronically, either locally (on a physical drive) or remotely (on a server in the cloud). Each institution's IRB will typically have well-defined guidelines for how to ensure storage locations are not vulnerable. These guidelines will likely be strict – but will ensure confidentiality of data and privacy of participants' information.

The guidelines for ensuring secure yet easily accessible digital storage varies from project to project. Yet some practicalities and methods are consistent across data type and storage device. As a general guideline, all data collected on portable devices, such as tablets using computer-assisted personal interviewing (CAPI) software, should be transferred to an approved service as soon as possible after collection and deleted from the portable collection devices.

When we work with local enumerators to collect survey data, we have them upload the data to remote server storage every night and then, after verifying the data are on the server, have the enumerator delete the local copy. The choice of server is typically a function of the CAPI software being used, as the software on the tablets must be able to upload the encrypted data to the server. We also restrict access to the identifiable data on the server to just those researchers who have completed Collaborative Institutional Training Initiative (CITI) training as part of the IRB application. And, it should go without saying, but, with respect to access to these data: all storage devices must be password protected with a strong password.

PERSPECTIVE 5.2 DEVELOPING FORMAL PRIVACY MODELS BY IAN M. SCHMUTTE

Formal privacy models like *differential privacy* give us a tool for quantifying privacy loss associated with a data publication and the trade-off between privacy loss and data quality. The model in Abowd and Schmutte (2019) establishes what (I hope) is a useful framework for working toward the optimal trade-off. Our model assumes that the data provider has high-quality information on social preferences for data privacy and data accuracy. In practice, managing this trade-off requires making decisions and judgment calls, often in consultation with stakeholders. The involvement of economists and other social science researchers is crucial.

For example, the disclosure avoidance system for the 2020 U.S. Census was originally tuned to achieve data-quality targets on a set of tables needed for reapportionment and enforcement of the Voting Rights Act. But the final privacy parameters were determined by interacting with data users, who were able to experiment with "demonstration tables." By running their analyses against the demonstration tables, they could illustrate areas where the system was not

meeting their needs and argue for a relaxation of the privacy loss budget or for other changes to implementation.

Before tuning a disclosure avoidance system, it must be developed in the first place. Building practical formal privacy systems depends on the nature of the input data and the desired end uses. Again, the process requires interaction with stakeholders and a detailed understanding of what the data are needed for. On the privacy side, we must determine *what characteristics or properties of the underlying data need to be protected against disclosure.* On the data-quality side, we must determine *what are the most important applications for these data*. With these in hand, we should develop a system for data dissemination that measures the kind of privacy losses that are relevant and the corresponding loss of quality for important applications.

What does this look like in practice? Standard differential privacy can be understood as limiting inference about whether a particular unit appears in the data at all or whether it has a particular attribute. But for data on businesses, it can be the case that the appearance of a business in the data is not considered to be sensitive. Policymakers might also decide that the places and the sectors in which businesses operate are not secrets in need of protection. However, business revenue and operating costs are highly sensitive and should not be disclosed. With these requirements, a standard differential privacy requirement is probably excessive, adding more noise to the data than is strictly necessary.

Hopefully these examples make clear the importance of social scientists' involvement in statistical disclosure limitation and data privacy. Our domain expertise is critical for the successful development and implementation of these systems, and we have a strong interest in them working as effectively as possible.

Once data collection is complete, we transfer the data off of the server used for data collection and into a file backup system like Google Drive, Dropbox, and the like for permanent storage.[10] We use commercial or institutional servers or cloud storage to ensure that the data in those locations is sufficiently encrypted. Once the data are in a permanent storage location, the first thing we do is create a key: a unique identifier for each observation or unit record. We then anonymize the data by separating names, addresses, Global Positioning System (GPS) locations, and other PII, from the rest of the data. We then move the identifiable, private information to a separate password-protected and/or -encrypted folder stored in a different, secure location. Access to the identifiers should be extremely limited. In most cases, only one person needs to have access to them. Separating the identifiers from the rest of the data adds a level of redundancy to the security of that data by ensuring that even if someone obtains access to the data, that data by itself would not identify anyone. The key allows, if it is ever necessary, for the data to be reidentified.

Given that most data collection now occurs electronically, either using CAPI software, being provided administrative data in electronic form, or gathering data on the internet, the storage and management of that data are straightforward. First, one should restrict access to the data during the collection process. Then, after data collection, one again separates any personal identifiers from the data and stores them in a different location. As before, all data should be kept, both the identifiers and the deidentified data, in file backup systems that are encrypted and that require a password to access. Once this is done, one can allow access to the de-identified data to those on the research project with minimal risk that those on the project will accidentally violate the privacy of research participants or the confidentiality of data.

5.3.2 GPS and Remote-Sensing Data

In recent years, economists have begun to use remote-sensing data in a myriad of different analyses. Economists use weather data to help understand human capital formation (Garg et al., 2020), labor markets (Morten, 2019), conflict and institutions (Sarsons, 2015), agricultural production and economic growth (Yeh et al., 2020), intra-household bargaining power (Corno et al., 2020), technology adoption (Tesfaye et al., 2021), and extreme weather impacts (Michler et al., 2019). GPS information has also been used to improve measurement of crop area (Carletto et al., 2017) and, when combined with remote-sensing data, used to improve estimates of harvests (Lobell et al., 2020). Frankly, we are just beginning to explore the potential of remote-sensing data in economic analysis (Burke et al., 2021).

All these new data and new opportunities, however, raise the issue of storage and management of GPS information and remote-sensing data. Although often used by economists as synonyms, Geographic Information System (GIS) and GPS are different systems. GIS is a framework for gathering, managing, and analyzing data. It is rooted in the science of geography and integrates many types of data. One of these types of data comes from the GPS, a satellite-based radionavigation system. GPS is just one of the global navigation satellite systems that provides geo-location and time information to a receiver anywhere on or near Earth. By comparison, another type of data used in GIS is remote-sensing data, which are data obtained about objects or areas from a distance, including aircraft, drones, or satellites.

Generally, economists use GPS data to determine the location of a natural resource (e.g., crop plot, mine), individual, household, firm, or institution and combine this information with remote-sensing data. Most remote-sensing data that economists use is publicly available and is collected, processed, and disseminated by national or international weather or space agencies. As the data are public, no special actions to protect privacy need be taken when collecting or storing the data.

Where privacy concerns arise is when researchers collect GPS data for individuals or households. For example, the specific GPS location where someone lives is sensitive data and requires protection. In some contexts, this might also be true for

firm locations, particularly if the firm is engaged in illegal or unregistered activity, such as marijuana farmers or unregistered mines. But in developing countries, many firms operate in legal dark or gray zones in terms of taxation, employment, registration, and compliance. The specific geo-location of these firms may be viewed as sensitive information.

When recording GPS locations as part of a data collection effort, researchers should use similar methods to those described earlier. The GPS information should be included with other identifiers and separated from the rest of the data at the earliest possible stage of the analysis. Care must also be taken when using the GPS information to obtain remote-sensing data. As economists, we rarely download remote-sensing data ourselves. Rather, we rely on colleagues in geography or with GIS expertise to get the data for us. In these situations, we share only the GPS coordinates with the colleague so that there are no other identifiers or data with the GPS information. When we obtain remote-sensing data, we use the key to matching it to the rest of the data set, dropping the GIS coordinates in the process. This provides us with anonymized data for use in analysis. However, as with individual, personal, or firm data, things become more complicated when it comes time to share the data.

5.3.3 Aggregate Data

To this point, we have largely presented data management in cases in which the unit of observation for the data of interest are an individual, a household, or a firm. In these cases, the privacy violation of releasing the data with the PII is obvious. With names of individuals and firms, with addresses, or with GPS coordinates, a user could immediately identify who the data are related to and the confidentiality promised in the informed consent document is lost. Moreover, as we have seen with the case of Governor Weld and the Netflix Prize, releasing data without direct identifiers may not be sufficient to truly anonymize data.

But what if one were to aggregate data that individual-, household-, or firm-level data into an altogether new unit of observation? Would sharing data that reported village-level averages for income rather than individual income data achieve a sufficient level of anonymity for research participants? Could one report country-level averages of profitability for firms in a given industry and be confident that the participant firms whose data underlie the averages would remain anonymous?

Unfortunately, the answer to this query is actually relatively straightforward: aggregating data does not provide sufficient anonymization of data for sharing with other researchers. This comes as a surprise to most people: How could one possible reidentify an individual or firm if all one has is village or county level aggregate values and how would one even know if a given individual or firm is in the data?

One economist who has thought deeply about the issue of privacy in aggregate data is Raj Chetty. Chetty developed the Opportunity Atlas, a Census tract-level database that shows which neighborhoods in the U.S. offer children the best

opportunity to get out of poverty (Chetty et al., 2018). The Atlas is based on U.S. Census Bureau data on 20 million children and follows them into their adulthood. The Atlas allows one to see, by parental income, child race, and child gender what neighborhoods provide children with the best potential to achieve a set of outcomes, such as income, while accounting for other factors, including birth rate and incarceration rate. All of the individual Census data that goes into the Atlas is aggregated to the Census tract-level. And most outcome variables are not reported as simple means or medians of the data. Instead, most of the outcome variables are based on ordinary least squares (OLS) regression estimates of the relationship between childhood characteristics and outcomes in adulthood. But even so, as Chetty and Friedman (2019a) show, the use of anonymized individual data to create aggregate data are insufficient to ensure privacy.

Consider the statistical outcome which results from aggregating income for households by reporting only the mean of income (Chetty and Friedman, 2019b). The potential for identification of a specific household arises when the statistic (mean income) changes substantially based on whether a specific household is included or not included in the data. If the population that is being aggregated is relatively large or if the household's income is very close to the mean, the sensitivity of the statistic to inclusion or exclusion of the household will be small. But if a very wealthy household is included in a small group of households, the inclusion or exclusion of that specific observation makes it more likely that someone can infer (1) if the specific household is included or excluded and (2) what the income of that household is likely to be. In the contemporary discussion on privacy protection, privacy is protected if the statistic in question is the same regardless of whether or not specific individual, household, or firm data are included in the data.

Now, consider a concrete example which comes directly from the Opportunity Atlas (Chetty and Friedman, 2019a). The authors of the Atlas want to release estimates from OLS regressions estimated on small samples. Recall that in the Atlas, the unit of observation is the Census tract (effectively, a neighborhood), and results are reported based on parental income, child race, and child gender. The Atlas releases predicted values from an OLS regression of the income percentile ranks of children in adulthood on their parents' income ranks. Based on the economic, racial, and gender makeup of households in a Census tract, the OLS regression results may vary substantially when one includes or excludes a given household.

Figure 5.1 illustrates this idea using hypothetical data presented in Chetty and Friedman (2019a). The figure presents a scatterplot of children's income ranks in adulthood against their parents' income rank for a hypothetical Census tract. The dotted light gray line shows the fit of a simple univariate regression to the data. Reviewing this, we can then ask what the predicted income rank for a child will be if their parent's income rank was at the 25th percentile, which is $\beta_1 = 0.212$. To demonstrate sensitivity, we can add an extreme outlier to the data. When we fit the same regression to the data that includes the outlier the predicted income rank for

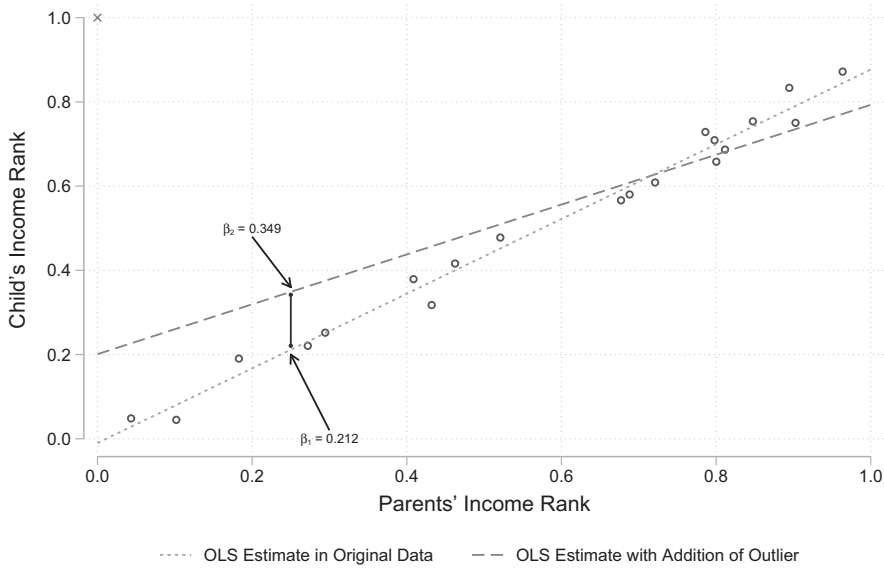

FIGURE 5.1 Sensitivity of Aggregate Statistics

Note: OLS = ordinary least squares. The figure shows how sensitive estimates can be to changes in the inclusion/exclusion of a single data point in a hypothetical Census tract. The figure presents a scatterplot of children's income ranks in adulthood against their parents' income rank. We then fit a simple univariate regression to the data and calculate the predicted value of child income rank at the 25th percentile of the parent income distribution (dotted light gray line). To demonstrate sensitivity, we add to the data an outlier point (0,1) marked by an x. We then fit the same regression to the data, including the outlier point, and calculate the predicted value of child income rank at the 25th percentile of the parent income distribution (dashed dark gray line). The figure is adapted from Chetty and Friedman (2019a) using data and code in Chetty and Friedman (2019c) under Modified BSD License and CC BY 4.0.

child with parental income at the 25th percentile is $\beta_2 = 0.349$. Although an extreme example, it illustrates the difference in potential loss of privacy that can occur if an individual, albeit an outlier, is included in a data set. Privacy is only protected to the extent that the aggregate statistics do not vary based on whether or not a specific person's or firm's data were included.

Removing what are traditionally considered PII (names, birth dates, social security numbers, telephone numbers) or spatial identifiers (addresses, GPS coordinates) does not sufficiently anonymize data. Nor does the aggregation of individual-, household-, or firm-level data sufficiently protect privacy. The reason simply is the existence of other data sets. As Latanya Sweeney demonstrated, it is possible to combine "anonymized" health records with voter records to identify Governor Weld. Like Arvind Narayanan and Vitaly Shmatikov showed, it is possible to link "anonymized" Netflix ratings with IMDb ratings to identify individuals. The danger is that by combining data, be it individual or aggregate, with other

existing data could result in a person reidentifying the private information present in the data.

5.4 Privacy Protection

The key tenet of modern data privacy is that a given data set should be private enough that one cannot reconstruct the missing information by combining that original data set with additional sources or data sets. We present three hypothetical cases to illustrate how and why the de-identification of a data set by removing explicit PII is insufficient for protecting the privacy of research participants.

Case 1: Consider a data set that contains information on farm households in the U.S. After collecting and analyzing the data, the researcher who collected the data posts it as part of a replication package for a paper. The researcher has removed names and addresses, but the data contain the county in which the farm is located, the number of adults and children in the household, and the farm and nonfarm income of the household. Given population density and county size in rural areas of the U.S., one can imagine an individual obtaining school yearbooks from the public library and determining which families in a given county had the same number of children as a household in the data set. The individual could then use telephone books at the same library to determine which family lived on which street. Finally, that individual could then drive to each address and determine if the family lived on a farm. With this information, the individual would then know the income of that particular family and could make that information public.

Case 2: Consider a nationally representative household survey data set from Zimbabwe that contains information on the health outcomes for individuals in the household. The researcher who collected the data captured GPS coordinates for the households and matched the data to remote-sensing data on rainfall and temperature. After analyzing the data and publishing a paper, the researcher archives the data in a public repository, having removed information on names as well as the GPS coordinates. The data still contain information on what diseases individuals in the household have contracted as well as the age of each household member and the matched time series weather data. Given that the data are nationally representative, rural households in the data set are fairly dispersed, and so an individual could match the time series whether data with the original remote-sensing data to determine in which grid cell a given household was located. Furthermore, the individual could access public voting records, which contain names, biological sex, addresses, and birth year of registered voters. With this information, an individual could identify the names of people in a household and the location of that household in the health data set.

The individual could then determine what diseases the voting members of that household had contracted and could make that information public.

Case 3: Consider a data set on sexual harassment complaints at firms in Canada, which includes the share of female employees at each firm. These data were collected by a graduate student. And, because of the sensitive nature of the data, the graduate student only makes a data extract available to go with published papers that use the data. A faculty member in the department requests to use the anonymized data in exchange for a data citation and the graduate student agrees but only shares a version of the data aggregated by Census subdivisions (municipalities), the International Classification System (ICS) of industries, and the quartile rank of the firm based on the share of female employees. The faculty member conducts their research and, believing the data have been sufficiently anonymized, publishes the aggregate data with the paper. An individual could use the public Directories of Canadian companies to determine the location and types of businesses in each Census subdivision. They could then use the employee directory on a firm's website to generate a rough breakdown of the share of female employees. Based on this information and the aggregate data (which likely contains very few observations in each aggregate "cell") the individual could identify a specific firm and determine the number of sexual harassment complaints at each firm and could make that information public.

Although hypothetical, the preceding three cases, and the real life cases discussed throughout this chapter, illustrate that de-identification of a data set by removing explicit PII is insufficient for protecting the privacy of research participants (Sweeney, 2007). Recall that when economists collect survey data, IRBs require two things. First, IRBs require that participants give informed consent, which includes consent to the use of their data. Most informed consent forms promise confidentiality of the information that a research participant provides. Second, IRBs require that the researcher complete a Protocol for Human Subjects, which includes describing how the researcher will protect the privacy of research participants and the confidentiality of the participants' data. Typically, IRBs only require that data be de-identified prior to analysis and data sharing. Yet, as we have demonstrated, de-identification is insufficient to ensure anonymity of research participants. There is a clear gap between what researchers promise participants (anonymity, privacy, confidentiality) and what we actually provide.

Fortunately, the risk of re-identification is well known to publishers of large, public use data sets like the public-use microdata samples (PUMSs) from the U.S. Census, the American Community Survey (ACS), Statistics of U.S. Businesses (SUSB), the Medical Expenditure Panel Survey (MEPS), the Living Standards Measurement Survey (LSMS), the Demographic and Health Surveys (DHS), the

Opportunity Atlas, and many others. Administrators of these and many other public use data sets are cognizant of the promises made to participants as well as the importance of protecting these sensitive data. As such, administrators for all of these public-use data sets undertake some version of statistical anonymization in order to provide privacy protection that goes beyond simply removing PII. Yet, while the use of these methods is ubiquitous, and sometimes required by law, in large public-use data sets, their use is very uncommon among individual researchers looking to share their data. What this means is that individual researchers are failing to sufficiently protect their research participants' privacy when they engage in data sharing. In the remainder of this chapter, we discuss the most common statistical de-identification methods used by large public-use data sets as well as how individual researchers can use these methods to ensure sufficient anonymization prior to sharing their data.

5.4.1 Statistical Disclosure Limitation

A set of commonly used methods to protect confidentiality in public use data fall under the umbrella of statistical disclosure limitation (SDL), sometimes called statistical disclosure control (SDC). SDL methods include noise infusion, aggregation, record swapping, and suppression. In the spatial dimension, SDL is often achieved through coordinate masking and noise infusion on derived spatial variables. The specific SDL method used is a function of the form of the statistical output, the sensitivity of that output to the inclusion or exclusion of a particular record, and the type of threat to confidentiality that exists (Skinner, 2009).

To understand the how and the what of specific SDL methods, it is important to have a clear understanding of the different types of threats to confidentiality. Computer scientists and cryptographers distinguish between identity disclosure and attribute disclosure (Abowd and Schmutte, 2015). Identity disclosure is when the identity, the PII, of a research participant is revealed in the data themselves or through combining data from multiple sources. This type of disclosure is what researchers are typically thinking about when they de-identify data prior to sharing. Separately, attribute disclosure is when it is possible to infer confidential attributes about a participant from the data or some combination of data. Examples include being able to infer the protected class (race, sexual orientation, religion, disability status, religion, etc.) of a participant. It is uncommon for researchers to consider protecting against attribute disclosure when they are preparing data for sharing.

In working to protect against both types of disclosure, SDL methods evaluate disclosure risk as a probabilistic event. This idea, known as inferential disclosure, is concerned not with the perfect identification of identity or attribute, but with one's ability to infer an identity or attribute with high probability. A person seeking to identify an individual identity or attribute has some prior belief about that identity or attribute. After data are published, the person updates their prior beliefs about the identity or attribute. If the prior and posterior beliefs are small, or the

differences are small, then the published data are confidential and preserve privacy. If, however, the difference in prior and posterior beliefs is large, then an inferential disclosure has occurred.

Abowd and Schmutte (2015) point out that addressing privacy issues through inferential disclosure provides two important insights. First, releasing any useful data (data that are not complete noise) results in a nonzero risk of disclosure. Second, certain SDL methods will require that the details of the methods remain confidential in order to be effective. Considering this from the perspective of the economic researcher, SDL inherently distorts the data, which can lead to bias in statistical analysis (Abowd et al., 2019). And, because data providers may be unable to publish SDL critical parameters, it may not be possible to determine the magnitude or direction of the bias (Abowd and Schmutte, 2015). With these insights and limitations in mind, we discuss specific SDL methods.

Suppression

Suppression eliminates entire records or specific attributes of a record from data (Abowd and Schmutte, 2015). Alternatively, specific values can be suppressed and then replaced with inputed values (Skinner, 2009). This is the most frequently used form of SDL and is commonly applied to tabular summaries published by the U.S. Census Bureau. A summary table for a specific Census tract might have missing values in cells for the number of single adults or the number of black or African American males in the tract. Because suppression is often used to deal with outliers or highly sensitive values, the exact suppression rule is generally not published because it could result in inferential disclosure. For users of the data, this means that it is often unclear how the SDL methods might bias the results of an analysis.

Data Swapping

Data swapping exchanges the value of a variable with the value from a different record (Abowd and Schmutte, 2015). The swaps can be done in such a way as to preserve certain characteristics of the distribution of the variable, such as the mean and covariance. However, in general, one cannot preserve the covariances between all variables (Skinner, 2009). Typically, a sensitive record will be identified and certain attributes of that record will be used to locate a "neighbor." Then the sensitive values of the original record will be swapped with the identified neighbor. Frequently, this involves swapping the geographic location of a record so that the sole very wealthy 20-something individual in a relatively poor Census tract will be swapped into a Census tract with many very wealthy individuals across various ages and someone of similar demographics (although either not as young or not as wealthy) will replace the original record. With data swapping, as with suppression, releasing information about the decision rule for when to swap, or where to swap, is typically unpublished. Publication of

this information could allow for the reconstruction of the original data set and the reidentification of the sensitive record. Without information on the swapping rules, researchers cannot determine how the SDL method might affect their analysis. Abowd and Schmutte (2015) consider suppression and data swapping "insidious" because neither method allows for data users to determine the magnitude or the direction of bias introduced by SDL.

Modification

Modification involves transforming a variable in such a way as to reduce the detail contained in a specific value (Skinner, 2009). This includes top coding and aggregation. Top coding is similar to winsorizing, which replaces outlier values with some maximum or minimum value. In the case of categorical variables, this top coding might be explicit, such as replacing any household size greater than 9 with the value 10+. Or, in the case of continuous variables, the process may be implicit, replacing all values of income above the 95 percentile with the income value at the 95 percentile. Similarly, aggregation involves coarsening the data (Abowd and Schmutte, 2015). In terms of continuous variables, such as revenue from individual firms, the data can be aggregated up to the ICS industry level. Alternatively, one could group firms by ICS category, calculate the median value for revenues within the group, and then replace each individual firm's revenue with the group median. Or, with individual income data, one could simply report the income quartile into which an individual falls. In terms of categorical variables, one could coarsen the categories. So, instead of reporting age or race and religious denomination, one could report age cohorts or broader religious categories (e.g., Judaism, instead of Orthodox or Reformed).

The degree of modification necessary is determined by the sensitivity of data to individual values. For postmodification, the usefulness or interoperability of the data is determined by the research question one would like to answer. Broad religious categories will be useful for answering some questions but will also limit one's ability to answer other questions. Similarly, top coding of income data can make the data useless in answering certain questions. Piketty and Saez (2003) demonstrate that income inequality in the U.S. looks very different if one analyzes income in publicly available data, such as using the top-coded Current Population Survey (CPS), than if one analyzes restricted data, such as the uncensored IRS income data.

Noise Infusion

Noise infusion adds stochastic perturbation to the values of a variable (Skinner, 2009). This addition of random noise is similar to the existence of measurement error for a variable. Methods for noise infusion are very sophisticated and can be designed to preserve certain characteristics of the original distribution, such as the

mean and the standard deviation of a variable, or the correlation between variables (Abowd and Schmutte, 2015). One example of this practice is adding noise to the population counts for each age cohort in each Census block in the Bureau's summary tables. Similarly, in the DHS and LSMS data, noise infusion is used to achieve a degree of spatial anonymity for unit records (Blankespoor et al., 2021). The DHS and LSMS unit record are typically at the household-level. In adding noise infusion in these data, the GPS data are first aggregated to a cluster, often referred to as an enumeration area (EA). The EA centerpoint for this cluster is then perturbed (displaced), such that centerpoints for urban clusters are within a 2km buffer of their true location and rural clusters are within a 5km buffer, with 1 percent of rural clusters displaced within a 10km buffer. The direction of displacement and the distance are both random variables (Perez-Haydrich et al., 2013).

If the variable that has been infused with the noise, and the parameters of the distribution are published, then a researcher could correct for any bias introduced by the mismeasurement. However, in general, the variances of any estimated parameters will remain inflated. But while noise infusion is less problematic for a researcher than suppression, it still limits the types of questions that can be asked. While using LSMS data, Michler et al. (2022) show that the current SDL methods to achieve spatial anonymization does not impact estimates of the impacts of rainfall and temperature on agricultural yields. Conversely, any amount of noise infusion would make a study like Lobell et al. (2020), where remote-sensing is used to measure plot-level agricultural yields, impossible.

Synthetic Data

Synthetic data are similar to noise infusion, but instead of targeting specific variables, all variables are perturbed (Skinner, 2009). Models to generate synthetic data are designed to allow analysis of the synthetic data to generate point estimates that are consistent with those generated from analyzing the confidential data. The synthetic data are drawn from the same data-generating process as the original data so as to preserve the same structure of the original data. The challenge is that a single synthetic data set can typically only preserve the relationships between a limited number of variables. So the original relationships between all variables cannot be preserved. That means that any individual synthetic data set can only be used to answer a limited set of research questions (Abowd and Schmutte, 2015). Alternatively, to address this limitation to some extent, a large number of synthetic data sets can be generated and multiple imputation methods can be used to analyze the original data and bind the bias introduced by the synthetic nature of the data (Skinner, 2009).

While all SDL methods distort data in order to preserve privacy, some SDL methods are more conducive to conducting SDL-aware analysis on. Suppression and data swap, by their very nature, require withholding key parameters of the process, leaving data users uncertain about the accuracy of their analysis.

In contrast, conditional on sensitivity and the exact research question, key parameters for modification, noise infusion, and synthetic data can be released to the public. The publication of these parameters allow for data users to adjust or bound their estimates, lending a degree of certainty to the accuracy of the analysis.

5.4.2 Differential Privacy

The use of SDL methods, like data swapping and imputation, were first used by the U.S. Census in 1990, allowing the Bureau to release for the first time Census block-level data. Since then, the list of SDL methods has grown as the ability to access and analyze data has simultaneously expanded. With the 2020 Census, the Bureau has moved beyond traditional SDL methods and now uses what is known as differential privacy (DP). The precipitating event for the shift from SDL to DP was the demonstration of the database reconstruction theorem (Abowd et al., 2019). The theorem establishes that publishing too many statistics too accurately from a confidential database exposes the entire database with near certainty (Dinur and Nissim, 2003). Abowd (2018) pessimistically calls the database reconstruction theorem the "death knell for traditional data publication systems."

DP is similar to SDL in that neither are a single technique to protect privacy but both are suites of tools. However, DP differs from SDL in that DP techniques allow for the precise measurement of disclosure risk, thereby avoiding excessive data manipulation while meeting anonymization objectives. DP provides a provable guarantee of privacy in exchange for a certain level of noise injected into the data within the constraints of a privacy-loss budget (Dwork et al., 2006). This formal approach to privacy rules is another advantage of DP over SDL in that DP makes explicit the social choice between more privacy protection or more data accuracy. Since the early 2010s, companies like Apple, Facebook, and Google have used DP techniques in preserving confidentiality of user data (Wood et al., 2018). Only recently, however, have economists begun to enter the conversation regarding the right mix of data privacy and data accuracy.[11]

To be differentially private, the published statistic would have to be similar when an individual's information was included in the data to when an individual's information was excluded from the data. The basic idea of DP is to inject enough noise into the data so that a user of the data cannot infer whether a given individual's information is in the data or not. How much noise is required depends on the statistics to be revealed and how sensitive those statistics are to the inclusion or exclusion of individuals. For a statistic such as the number of minority-owned farms in a Census block, DP would require that the published statistic not reveal if an individual was included or excluded from the data. DP gives the individual plausible deniability regarding their participation or lack of participation in a study or database (Christensen et al., 2019). This concept is what underlies the approach

taken by Chetty and Friedman (2019b) regarding aggregation of data discussed earlier in this chapter.

To further illustrate the idea of DP, we adapt a common example from the literature (Heffetz and Ligett, 2014). A researcher conducts a DP analysis on a health data set that contains information about patient smoking habits and the occurrence of different types of cancer. The analysis reveals a correlation between smoking and lung cancer. The researcher then publishes a paper stating this correlation. An individual smoker might feel that their privacy was violated, because third parties, like insurance companies, can now infer the probability that the smoker will develop lung cancer and will therefore charge them more for health insurance. However, because the outcome of the analysis would be no *different* whether or not the individual smoker's health data were included in the analysis, the *differential* privacy of the individual smoker has been preserved.

As with SDL, DP protects against inferential disclosure. Imagine two data sets that are exactly the same except that data set $\Theta_{\{A\}}$ contains a record for person A and data set $\Theta_{\{-A\}}$ is lacking that record. DP adds noise to any statistic or analysis so that the probability of getting any given value for that statistic or from that analysis is similar under $\Theta_{\{A\}}$ and $\Theta_{\{-A\}}$. Said differently, with a certain degree of probability, a researcher cannot infer if the statistic or the analysis comes from data set $\Theta_{\{A\}}$ or $\Theta_{\{-A\}}$. Some might object that inferring person A's identity or attribute is not the same as knowing with certainty that person A was included in or excluded from the data set. Or one might argue that privacy is really only violated if the researcher knows for certain that person A was included in the data set. And furthermore, one could claim that the only way to go from inference to certainty is if the researcher knew the method by which the statistic was produced. As a result, one might then conclude that as long as the method for protecting privacy was kept secret, such as the decision rule for record swapping in SDL, then person A's privacy is protected. However, a basic tenet of modern cryptography is that a system is not secure if its security depends on the internal algorithms or methods of the system being kept secret (Wood et al., 2018). Simply hiding the key is not a sure method to keeping something locked.

As DP treats disclosure risk probabilistically, one needs to define the degree of probability with which a researcher can infer if the statistic or the analysis comes from data set $\Theta_{\{A\}}$ or $\Theta_{\{-A\}}$. In DP, a privacy loss parameter, ε, measures the impact that each person's information has on the statistic or analysis, similar to the sensitivity parameter defined in Chetty and Friedman (2019b). The database manager, or the individual researcher looking to share their data, can choose a desired level of privacy protection, ε. The DP algorithm then injects noise into the statistic or analysis to ensure that the probability of disclosure is no greater than ε. In this way, DP can provide a mathematically provable guarantee of a certain level of privacy. Conversely, SDL methods are unable to provide the certainty of a particular level of privacy.

Another difference between DP and SDL methods is that DP is designed in such a way as to allow for the publication of the specific parameters used to infuse noise. This allows researchers to conduct DP-aware analysis, correcting in part for the noise introduced by DP. The major cost of DP is that it requires the specification of a privacy-loss budget. In order to continue to guarantee a certain level of privacy protection, DP must limit the total queries made on the data. Eventually, this privacy-loss budget may be exhausted, at which point data could no longer be used by the public without violating the original guaranteed level of privacy. This hard limit on how many times a data set can be used while maintaining a certain level of privacy is the result of the database reconstruction theorem (Dinur and Nissim, 2003). While detractors of DP view this limit on use as a detriment of DP, proponents point out that this trade-off is not unique to DP and among competing methods DP makes this trade-off explicit. It is then a social choice problem how best to balance privacy and accuracy and prioritize different uses of the data (Abowd and Schmutte, 2019).

As of 2022, DP has only just begun to be adopted by the statistical agencies and the managers of the databases most commonly used by economists. This includes the U.S. Census Bureau, which adopted DP for the 2020 Census, and plans to implement DP in several of the other data sets it publishes (Abowd et al., 2019). The Opportunity Atlas, which is published at the Census tract level, also protects privacy by methods that build on DP (Chetty and Friedman, 2019b). However, to date, public-use household survey data sets used in development economics still rely on SDL to protect participant privacy. And it is exceedingly rare for individual researchers to apply these methods to their data prior to sharing the data.

5.4.3 Implications for Data Sharing

As should be obvious by now, removing PII from data prior to sharing is not sufficient to meet the standards of confidentiality that we promise to our research participants during the process of obtaining informed consent. But as stated at the outset of this chapter, open data is a key part of open science. Applied economic researchers need to grapple with the trade-off between protecting privacy and sharing truly replicable data.

Privacy protection is not trivial, nor is it limited to large public use data sets or censuses of populations. Even a few simple data points can be used to perform a successful reidentify attack (Heffetz and Ligett, 2014). For most people in the U.S., all that is needed is birthday ($d = 365$) and year ($y = 100$), sex ($s = 2$), and a five digit ZIP code ($z = 32{,}000$). These pieces of data result in more than 2 billion possible combinations, but there are only 330 million people in the U.S. Most individuals are the only person of their sex born on their birth date living in their ZIP code. Given how unique just birth date, sex, and location are, consider of how identifiable an individual is within a data set that includes education, career, household size, farm size, income, race, or any of the other socioeconomic variables typically collected and released as part of a replication data set.

The burden is on us, as researchers who collect, manage, and analyze data to share data in such a way that truly protects the privacy of our research participants. Greely (2007) writes, "It is no solution to say that 'anonymity' means only 'not terribly easy to identify,' . . . or that 'informed consent' is satisfied by largely ignorant blanket permissions." However, for the field of economics it is unclear how best to manage the privacy–accuracy trade-off. Regardless, what is clear is that the current practices fail to satisfy the promise of informed consent regarding the privacy of participants and the confidentiality of data. Without taking what Heffetz and Ligett (2014) term the "naive" approach of refraining from all data sharing, the ethical researcher needs to invest the time and energy to learn how to share data in a truly private way. Fortunately, there are already a growing catalog of user-written R and Stata code for implementing SDL and DP methods (Heffetz and Ligett, 2014; Chetty and Friedman, 2019a). As more and more economists take seriously the need to protect privacy in open data, we hope and believe that the supply of new methods and codes will rise to meet the demand.

5.5 How to Share Data

Sharing data can be difficult and time consuming. It is a challenge to understand how to share data in a way that adequately preserves the privacy of research participants. Additionally, it can take a substantial amount of time to convert disorganized code and folder structures into a truly replicable package. While real, these hurdles are insufficient justification to refrain from sharing data and contributing to open science.

The best way to overcome these hurdles is to proactively develop a data management and sharing plan at the earliest stages of the research life cycle. Funding agencies increasingly require a detailed data management and sharing plan as part of the preliminary project proposal. At the initial idea stage, discussing a data management plan may seem premature, but having a plan in place can help guide the data collection process and streamline the data analysis process. Once a researcher has developed one or two data management and sharing plans, and learned what worked and what did not work, adapting the plans to new projects is fairly straightforward. In our Applied International Development Economics (AIDE) Lab, we have an OSF page (https://osf.io/3vtng/) that outlines standard procedures regarding coding style and data folder structure that are portable across different projects.

While all data management plans vary slightly, all contain the same basic elements. According to to the NIH, a comprehensive data management and sharing plan should:

- identify the data types and resources that will be generated, including the file types and software that will be used to clean and analyze the data.
- propose a timeline for sharing the data and resources, including any embargos.

- determine where the resources will be stored (public or institutional repository), preferably with a DOI address.
- describe how others can access the resources. This entails not just where the data are located but also where the code is located and how the code can be deployed to clean the data and reproduce results.

The plan should also govern the management and sharing of metadata, or data about data, which the NIH (2020) defines as "data that provide additional information intended to make scientific data interpretable and reusable."

5.5.1 Replication Packages

The dual, and simultaneously sought, objectives of open science and data sharing are to create efficiencies in the discovery process while allowing for the verification of past discoveries. Open science makes it easier for future researchers to build on past work without duplicating it. Open science additionally permits for an easier replication and/or reproduction of past work, strengthening the scientific record. *Replicability* and *reproducibility* are often used as synonyms but they have slight different meanings. We discuss these further in Chapter 6, in terms of data analysis but here discuss them with respect to data management.

Replicability is to obtain consistent results across studies that are working to answer the same question but using new data or different methods. This procedure is often called a conceptual replication, an extension, or a verification test. Reproducibility is the degree of similarity (or extent of differences) between the results of measurements with the same data, carried out using the same method. This procedure is often called direct replication or reanalysis (Christensen et al., 2019). To assist with achieving both these goals, applied economists need to create replication packages that include data, code, and implementation details in a README file such that others can replicate and reproduce their work.

In producing a replication package, it is important to distinguish between two different types of reproducibility (Stodden, 2014): computational and statistical. This differentiation is essential because how much data one chooses to share will affect how reproducible one's research is. First, computational reproducibility, what Clemens (2017) calls verification, is when the same procedures are used on the same data to determine if the exact results in a paper can be reproduced. To conduct a verification test requires the data and code, as well as detailed information about software, hardware, and implementation details. Ideally, a replication package that allows for computational reproduction would include the raw data, the cleaning code, and information about software version and hardware type. At minimum, however, all that is required for computational reproducibility is a data extract and the code to produce the tables and figures in the paper. This minimum version of a replication package for achieving computational reproducibility is what is required by most economics journals.

Next, distinct from computational reproducibility, is statistical reproducibility (Stodden, 2014). Statistical reproducibility, or what Clemens (2017) calls reanalysis, is using the same data, or a subsample from the same data, but tweaking the methods and procedures. The goal is to statistically assess the validity or robustness of a result under different permutations. Most empirical research papers check the statistical reproducibility of their findings within the paper itself. In economics, we generally refer to these as robustness checks.[12] To conduct a reanalysis test requires a bit more information than a verification test. One needs both the data and code, as well as detailed information about the choice of statistical tests, model parameters, threshold values, and more. Limiting the replication package to a data extract is not useful in verifying statistical reproducibility because one would like to verify the robustness of results on subsamples of the data and to different choices in data cleaning and variable definition. Thus, both more detailed information and more types of information are required to make a replication package that is statistically reproducible. However, at this time, including sufficient information in a replication package so as to ensure research is reproducible is not the norm in economics and so is infrequently done.

Following the guidelines put out by Jacoby and Luption (2016) for the *American Journal of Political Science* (*AJPS*), a replication package for statistical reproducibility should contain the following: (1) a README, (2) data sets, and (3) software commands.

README

Consider the task of assembling IKEA furniture – a relationship-challenging endurance test in its own right. Now, consider this same assembly task if you were not given directions. A dramatically more difficult ask, to be sure. The directions for assembly are akin to a README file in a replication package. A README file provides the directions for how to take the various components of a replication package and put them together to build the figures and tables in the final paper. README files should be produced as either a plain text (.txt) file or a portable document file (.pdf) file. The README should be readable by as large a set of software types as possible and should not rely on proprietary or pay-for software, like Microsoft Word. Nor should it rely on software that requires an internet connection, like Google Docs. README files should be designed to be widely accessible, as not everyone has the same software programs or shares the same level of internet access.

The README file should contain a list of all the components within the replication package along with a short description of each component. Ideally, one's data analysis workflow (see Chapter 6) will allow for the creation of a single "project" script file that calls and runs all subsidiary script file and produces all tables and graphs. If not, the README should describe which files produce which tables and figures, referencing the number or name used in the published paper. Figure 5.2 provides an example.

Readme Data File for "Money matters: The role of yields and profits in agricultural technology adoption"

Jeffrey D. Michler, Emilia Tjernström, Simone Verkaart, Kai Mausch

15 May 2018

This folder contains the data and Stata programs (version 14) required to replicate the tables and figures in Money Matters as well as the associated appendices. The programs generate all summary tables and figures. In order to construct the final regression tables, some manual editing (and copy-ing/pasting across files) is required, so we do not reproduce them here. However, all relevant numbers are in the regression output generated by the programs files listed below.

The programs are as follows:

- `mm_master_program.do`: Replicates the all tables and graphics in the paper. This program will run all subsidiary .do files.
- `mm_pub_tables.do`: Replicates the regression results in the paper and the appendix. Specifically Tables 3-5 and B1-D1. Plus Figures 2-4 which are generated using regression results.
- `mm_sum_stats.do`: Replicates the summary statistics tables plus MW-tests in the paper and the appendix. Specifically Tables 1-2, 6, and A1.
- `mm_graphs.do`: Replicates Figures 1, A1-A3.

The data are included in `mm.dta`. Several of the programs produce additional data files as inputs into regressions contained in the .do files.

Note that the code requires installing the Stata package `-randcoef-` as discussed in Barriga Cabanillas et al. (2018) and `-tuple-`.

References

Barriga Cabanillas, O., J. D. Michler, A. Michuda, and E. Tjernström (2018). Fitting and interpreting correlated random-coefficient models using Stata. *Stata Journal 18*(1), 159–73.

FIGURE 5.2 Example of a README file

Note: Reproduction of a README file for Michler et al. (2019).

the data file is read into a different software program, those labels may be stripped from the data. Fortunately, Stata and similar programs allow one to automatically generate a codebook. So, even if data files contain labels, its a good idea to include a codebook with the data file, in case there is a decoupling of labels from variables.

While including or linking to a data extract is the most common form of data sharing in economics at this time, an ideal replication package would provide the original data source and all information to reconstruct the analysis data set. In fact, Jacoby and Luption (2016) include "analysis data sets" and "information to reconstruct analysis data sets" as two separate components in a replication package. Following the advice of Vilhuber (2022) and speculating on the the trajectory to

Beyond a list and description of the components in the replication package, the README should detail any folder structure required for the data, what is known as the development environment, and the order the files need to be run in. Furthermore, the README should include information on what programs are necessary for replicating the analysis, including any user-written programs, the version of those programs, and whether any manual manipulation of tables and figures is required. One might also include information about the runtime and hardware requirements, particularly if the analysis relies are simulations or big data.

Data Sets

Any replication package should reference the data used in the analysis. For data sets of reasonable size, the data can be directly included in the replication package as a file. For larger data sets, or for those who use internet repositories or an online version control systems like GitHub for hosting replication packages, the data can be archived in a separate location. Data archives, such as Harvard's Dataverse or the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, can hold data or the entire replication package. Alternatively, for analysis that relies on public-use data sets, the replication package should reference where those data can be obtained and the edition or vintage of the data (i.e., the original download date of the data). Regardless of what type of data are necessary for the replication package, what is crucial is that if the data are not included in the replication package, then the replication package must include citations to the data. These citations should be as detailed as a bibliographic citation. Ideally, the citation will include a digital object identifier (DOI), which is a permanent web address for the location of the data. Placing data in an internet archive automatically generates a DOI for that specific archive entry.

The format a researcher should store their data in, either in the replication package or in an archive, tends to be driven by the standards of the research field. Jacoby and Luption (2016) suggest storing the data as .csv files, which can be read by almost every statistical software. However, if certain types of software dominate the field, then providing data in the software's own format is acceptable. As Figure 5.3 shows, Stata and Matlab dominate among the software programs used by economists.

Although storing data in .csv format makes data readable by many different types of software, the cost is that the format does not allow for any labeling of variables. Because of this, the *AJPS* requires data be accompanied by a codebook that details what each variable measures and what their value means (Jacoby and Luption, 2016). The variable y in economics typically signifies output, but if the variable y shows up in a data set, what does it measure? The output of what? And how is the value measured? In kilograms? Or in kilograms per hectare? One benefit of data file formats like that of Stata is that it allows for the attachment of both variable and value labels. This mitigates the need for a separate codebook. But, if

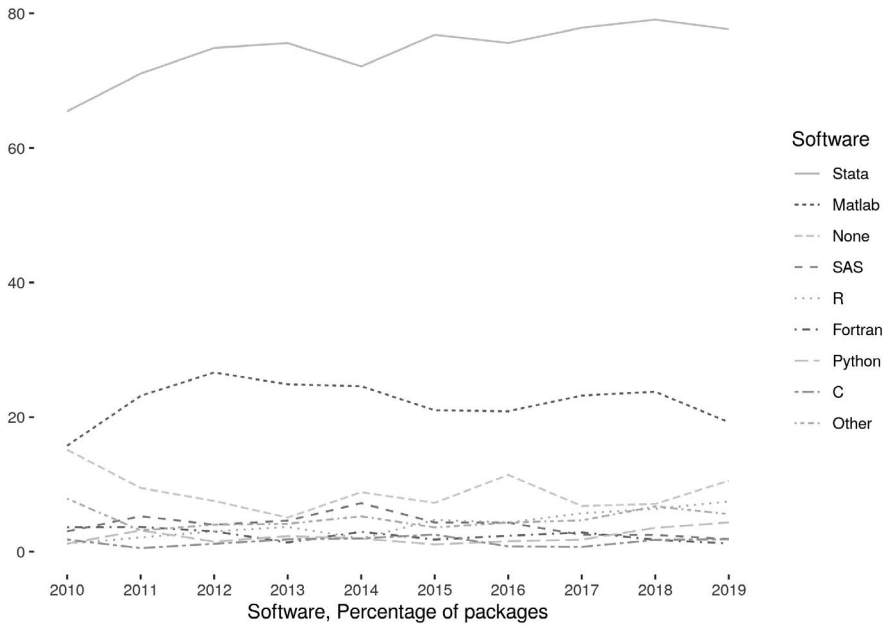

FIGURE 5.3 Software Used by Economists

Note: Figure presents line graphs of the number of papers submitted to the Data Editor of the AEA journals that used a specific software package. Reprinted from Vilhuber et al. (2020) under CC BY 4.0.

come in economics at large, we believe that there is no need to include the original version of the data and a final "analysis data set." Rather, economists should strive to build a replication package that is "push button reproducible." This means that there is a single project or master script file that runs all the code for both data cleaning and analysis.

Beyond format and location of data, a critical decision that researchers need to make is how to protect the privacy of research participants whose data are included in a data set. As should be clear by now, simply removing PII from the data is insufficient to ensure privacy and confidentiality. As of now, there is no standard for how economists should anonymize data prior to sharing them. And so, at present, it is up to the individual researcher to decide how much effort to put into making their data truly anonymous.

Software Commands

Obviously, to be actually reproducible, a replication package must include the software commands used to execute the analysis. As with the analysis data sets, the format of the code or script files is less important then their clarity and detail. We provide more details on organizing a workflow in Chapter 6. Here it is

sufficient simply to say that treating your code like a lab notebook is an excellent habit to develop. Script files should include not just what was done (the software commands) but also notes and commentary on why certain actions (commands) were used and the outcome of those actions. A lab notebook without this sort of editorial detail is useless to anyone other than the author, and even the author might have difficulty reproducing their work after a few months or years working on other projects.

As mentioned in the previous subsection, the software commands in a replication package should be organized so as to be push button reproducible. That means that a user needs only to place the data in the correct location and then open up a single script file and click run. That script file will then set up the development environment, with all necessary folders, run all the cleaning code, run all the analysis code, and output all tables and figures in a readable format. Ideally, this overarching project file would even query the archive where the original data are stored and download it to the right local folders. Vilhuber (2022) and the AEA Data Editor's GitHub page include many recommendations and tools to make one's replication package as automated as possible.

5.5.2 Rapidity of Change

In applied economics, the norms and standards around replication are changing rapidly. The reasons for the changes are twofold. First, researchers are realizing that the current norm of just providing a data abstract and analysis code to journals, through only self-enforcing mechanisms, is too weak. Even at journals with explicit policies, these policies are often not followed. Second, computing, network, and storage technologies are changing so rapidly, that what is possible to include in a replication package has greatly expanded in just a few years.

Regardless of where the norms and standards for sharing data and code go, we as individual researchers should be guided by one key principle: computational empathy (https://aeadataeditor.github.io/aea-de-guidance/preparingfor-data-deposit.html). Computational empathy means remembering the following two points:

- The replication package is meant to be run by others who have none of the setup, packages, and data, that the original author might have, on computers that may not run the same operating system.
- The replication package should be treated as one of the methods to convey the processes that lead to a manuscript's conclusions. Consider it a teaching tool, targeting graduate students or others who may not be in one's field.

To achieve computational empathy, posting a bunch of data and code online is not sufficient. The author of a paper is the most qualified person on the subject of that paper. This includes all the various components which went into the paper, including the data and code. By extension, this means that whoever is attempting to

replicate this work is less experienced and less qualified than the original author. To that end, the author should do all that they can to ensure that their work is replicable, reproducible, and transparent. As methods are changing fast, graduate students or early-career researchers need to be leaders on this front – both with co-authors and in lab or research groups. It is likely that students and early-career researchers will use tools that principal investigators and advisers have not used before. Just in the five to seven years since we completed our PhDs, how data and code can be shared is completely different. When we completed our PhDs the standard was uploading replication packages as .zip files to the author's website or the journal's website. Now data and code can and should be archived on Dataverse or ICPSR and given a DOI. Containers, like Docker, now provide a standardized way to package data, software, and code so that replication packages can be executed regardless of machine or operating system. With these changes in the past decade alone, what the decade to come will hold is sure to be tremendous with respect to advancing open-science practices.

There are also practical considerations for ensuring a comprehensive and usable replication package. Specifically, the pathway to publication will be eased with an appropriately replicable package. Figure 5.4 outlines the process flow of providing replication materials for papers accepted to AEA journals. While the specific process flow is likely to be different at different journals, the broad outlines are instructive. What should be obvious is that making research truly reproducible takes time. It cannot be done by tossing a bunch of data and code files into a folder, zipping it, and then posting it online. Early-stage decisions about data management and workflow need to be made with the end product (paper and replication package) in mind.

As for current resources on practicing open science and creating optimal replication packages, we recommend a few sources. First, the AEA Data Editor maintains a GitHub web page that contains guidance and resources for putting together a replication package that meets the current highest standards in the economics profession (https://aeadataeditor.github.io/). Additionally, BITSS maintains a web resource with advice and links to a variety of tools for building replication packages (www.bitss.org/resource-library/). While these specific resources and their contents may change as standards for replication packages increase, so will the available resources. Learning and using these tools is a long-term investment in terms of career success but also for making science better.

5.6 Conclusion

Open data are a cornerstone of open, ethical, and reproducible research. Yet, open data presents its own ethical challenges. Simply sharing data violates promises of anonymity made to the research participants whose confidential information is in the data. The applied economics profession is just beginning to grapple with the trade-offs between more accurate data and more privacy protection. New methods

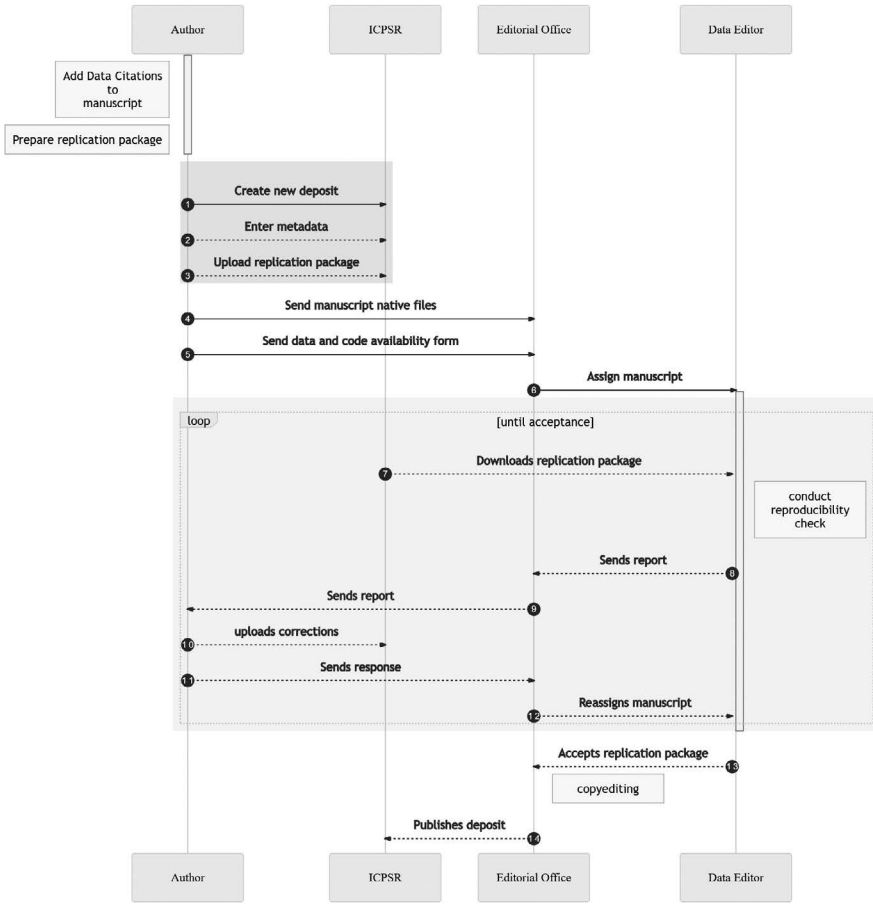

FIGURE 5.4 Replication Process Flow at American Economic Association Journals

Note: ICPSR = Inter-university Consortium for Political and Social Research. The figure presents the approximate flow process for providing replication materials for papers accepted to AEA journals. Reproduced from https://aeadataeditor.github.io/aeade-guidance under CC BY 4.0.

of privacy protection, like SDL and DP, offer opportunities to better protect the privacy and confidentiality of research participants. But these methods are not yet the norm in the profession.

De-identification, which is what IRBs typically require for data, does not provide the anonymity that informed consent forms promise. The language on informed consent forms needs to change so that it truthfully represent what sorts of privacy protection can be guarantee. But we echo Heffetz and Ligett (2014), in cautioning against an alternative of reverting to a norm of closed, private data sets. Currently the field is left with the unenviable choice between our commitment of anonymity to our research participants and our commitment of openness to the scientific community.

As a profession, we are still working to arrive at the optimal allocation. As individual researchers, we must keep learning and keep adapting new methods. We are learning to create better replication packages and practice computational empathy in their creation. All the complexities and complications of data management offer an opportunity to learn and practice better science – in line with available technology – rather than be a burden. As such, as researchers, we must take it upon ourselves to ensure that our work ensures the privacy and confidentiality of the participants in our research while simultaneously ensuring that our work is replicable and reproducible for others in the community.

Notes

1 Sweeney is now the Daniel Paul Professor of the Practice of Government and Technology at the Harvard Kennedy School and in the Harvard Faculty of Arts and Sciences.
2 Dr. Narayanan is now an associate professor of computer science at Princeton University, and Dr. Shmatikov is now a professor of computer science at Cornell University and Cornell Tech
3 Kristal et al. (2020) did not completely ignore the field experiment. They report in their paper that they found large differences in the baseline data and conclude that the initial randomization failed in some way. They did not pursue the reanalysis any further.
4 It is worth highlighting that the official data sharing policy of *Science* reads:

> All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including fees and original data obtained from other sources (Materials Transfer Agreements), must be disclosed to the editors upon submission.

5 Anecdotally, one of the authors can speak to a co-author suggesting exactly this due to the "onerous" requirement.
6 The policy also encourages citation of software packages, another relatively uncommon practice in economics.
7 We discuss TOP guidelines in detail in Chapter 7.
8 This is to say that IRBs are (generally speaking) unconcerned with data management practices.
9 Obviously, one could cut out this step if the server used for data collection was also the server for permanent data storage. The only reason we use different locations is that the server we use for data collection is not our university's preferred (and institutionally paid-for) cloud storage service.
10 See the symposium at the 2019 AEA Annual Meeting (Abowd et al., 2019; Abraham, 2019; Chetty and Friedman, 2019b; Ruggles et al., 2019).
11 See more about robustness checks in Section 6.4.