Preliminary Working Draft Only — Not for Dissemination – Do Not Cite

Privacy Protecting Research: Challenges and Opportunities

Daniel Goroff[*]
Jules Polonetsky[**]

"A secret isn't invalidated by its disclosure, it's defined by its disclosure. What makes a secret a secret is simply the operating instructions that accompany its movement from one person to the next."

– Malcolm Gladwell writing in *The New Yorker* (December 19, 2016) about the work of sociologist Beryl Bellman.

I. Stating the Problem

To call a policy "evidence-based" means that research on data played a role in its formulation. Often the data studied contain private or sensitive information about individuals. Supposing such data should not be released more widely than necessary, how should researchers gain access? Answering that question presents challenges and opportunities outlined in this essay.

To formulate the problem more precisely, it is useful to be explicit about certain distinctions and definitions. The word "policy," for example, refers here to a course of action with broad and significant consequences across a given population. It is not a matter of making a personal decision, or a decision about another given individual. It may be my personal policy not to shop online. I may even have reasons for reaching this conclusion based on observations. That is not the kind of policy issue of concern in this essay. Whether to tax online purchases could be.

In the case of online retail, tax decisions could be based on empirical research findings such as: "if you raise online taxes to x%, local purchases will go up by y%." What counts as convincing empirical evidence usually takes the form of a causal implication like this rather than, say, a simple correlation such as "states that tax internet transactions also have larger in-store sales." The statement "if you do x, expect y" has a direction, after all. But observing that "x covaries with y" is the same as observing that "y covaries with x." In other words, correlation is not causation, and causal inference is the kind of "evidence base" that policymaking ultimately needs.

[*] Vice President and Program Director, Alfred P. Sloan Foundation. Opinions or errors expressed here are his own rather than those of the Foundation or its grantees.

[**] CEO of the Future of Privacy Forum, a Washington, D.C.-based think tank that seeks to advance responsible data practices. FPF is supported by the chief privacy officers of more than 110 leading companies, several foundations, as well as by an advisory board comprised of leading academics and advocates.

Finding correlation among variables may be interesting and suggestive, or misleading and spurious. It is a fine and fun way of generating plausible hypotheses in any case, and thus falls under the category of "exploratory research." That is different from "confirmatory research" whose purpose is to test particular hypotheses about the relationship between, say, variables x and y. Confirming causal inferences can be particularly difficult. Such work demands careful attention to confounders, endogeneity, overfitting, or other potential sources of statistical bias. Researchers trying to generate an evidence base for a policymaking in this way therefore face different data requirements and restrictions than researchers seeking simple access to data for exploratory purposes.

And about that data, another distinction worth keeping in mind is between information whose collection was specifically designed to help answer the particular research question at hand and information collected for some other purpose. We refer to the latter as "administrative data," though others also refer to transactional, observational, or "found" data. Sources of administrative data typically include government, company, or other organizational records. Big data of this sort holds great promise for evidence-based policymaking. To realize that potential, however, requires more than just granting access to researchers. Using administrative data well also requires attention to data cleaning, metadata, reproducibility, data linking, transactions costs, and—perhaps even more than in experimental situations—attention to privacy and ethics, too.

The fact is that nearly any kind of research on data about individuals carries risk. In the case of policy research, the consequences can be especially significant. Granting access to data may end up damaging some individuals even if the intent is thoroughly beneficial. This goes for both natural persons as well as organizations concerned about the privacy of proprietary information, for example.

So it only makes sense to begin considering questions about access to sensitive data by clarifying how and why the research to be enabled would take place. Everyone might agree that "evidence-based policy making" represents an excellent goal that holds great promise. But no one wants to give up their data or their privacy for shoddy or unreliable science. It is all about trade-offs, and so the terms of access matter, and not just the granting of access.

How such trade-offs are being posed and how the terms of access are being decided are the subject of several examples presented in the next section. Based on these case studies, Section III posits three objectives for measuring the success of any plan to supply data for policy research. These provide criteria for answering how, what, and when questions. Section IV lays out technical alternatives for achieving our objectives, with special emphasis on solutions based on data enclaves, de-identification, and differential privacy in Section V. Then Sections VI takes up who and where considerations, including the credentialing of researchers and institutional options for organizing data access. The final section speculates about next steps, costs, organization, and execution.

II. Examining Examples

Researchers at academic institutions as well as private sector businesses tap a variety of sources of administrative data to pursue projects that promise great societal benefits. Companies collect massive amounts of administrative data through the Internet, mobile communicatins, and a vast infrastructure of devices and sensors embedded in healthcare facilities, retail outlets, public transportation, social networks, workplaces and homes. They use administrative data to test new products and services, improve existing offerings, conduct research and foster innovation. For example, in an article recently published in the *Journal of Oncology Practice*, a group of Microsoft scientists demonstrated that, by analyzing a large sample of search engine queries, researchers could in some cases identify Internet users who were suffering from pancreatic cancer even before those users were diagnosed with the disease.[1]

Increasingly, the collection and analysis of large-scale administrative data drives advances in the measurement and tracking of economic activities. Stanford economists Liran Einav and Jonathan Levin report, for example, that "Visa generates periodic reports that successfully predict survey- based outcomes ahead of time. Similarly, Automatic Data Processing (ADP) and Moody's Analytics release a monthly report on private- sector employment, based on data from the roughly 500,000 firms for which ADP provides payroll software."[2]

While compelling from a societal standpoint, such data-intensive research projects face formidable transaction costs, including real and perceived legal and ethical challenges. Consequently, data could remain locked in corporate coffers, weighed down by concerns about individuals' privacy, data security, and re-identification risk, as well as an incentive to protect trade secrets and intellectual property. These challenges affect all links of the data chain, including access to administrative data, its protection, analysis, sharing and linking. Moreover, the lack of a clear legal framework and ethical guidelines for use of administrative data jeopardizes the value of important research, either because the public perceives it as ethically tainted or because research outputs remain hidden from public view. Concerns over legal impediments and ethical restrictions threaten to diminish productive collaboration between researchers and private sector businesses, restricting funding opportunities and potentially locking administrative data and research projects behind corporate walls. Further complicating matters, companies struggle to define the line between scientific research projects and A/B testing for marketing or product improvement.

[1] John Paparrizos, Ryen W. White & Eric Horvitz, Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results, Journal of Oncology Practice, June 7, 2016, doi: 10.1200/JOP.2015.010504.

[2] Liran Einav & Jonathan Levin, The data revolution and economic analysis, in *Innovation Policy and the Economy,* J. Lerner, S. Stern, Eds. (Univ. of Chicago Press, Chicago, 2014), vol. 14, pp. 1–24. See also Liran Einav & Jonathan Levin, Economics in the Age of Big Data, 346 *Science* 715 (2014).

Similarly, data held by government agencies, or entities such as universities, school districts, or other quasi-governmental institutions, are also often inaccessible for important analysis by researchers. Benefits can include everything from improving official government statistics using corporate transactions records to rigorously testing causal hypotheses about policy improvements by using randomized controlled experiments. Despite important steps to make data available advanced by open data movements, by efforts to use data to encourage accountability in education systems, smart city efforts, and other programs, the ability of researchers to access significant government data sets are often limited by a range of concerns, in large part consisting of privacy and security objections.

<u>State Bar Data Case Study:</u>

In 2008, UCLA economist and law professor Richard Sander filed a lawsuit, originated from a request for access to information from the State Bar's admissions database. Sander requested "'individual-level' data concerning all applicants for the California Bar Examination from 1972 to 2008" that contains the following categories of information: race, law school, whether an applicant was a 'transfer student,' year of law school graduation, law school grade point average ('GPA'), Law School Admission Test ('LSAT') score, whether an applicant passed the bar, and raw and scaled scores for each component of each California bar exam taken."[3] According to Sander, the data was to be utilized in researching "the effect that attending particular law schools has upon students who have been admitted to those schools with the use of significant preferences."[4]

The State Bar opposed the release of the data, asserting it would violate promises it made to law students regarding privacy and limited use of the records.[5] In 2010 the San Francisco Superior Court concluded that no law required the State Bar to disclose the records.[6] However, in 2013, the Supreme Court reversed, holding that: "Under the common law right of public access, there is a sufficient public interest in the information contained in the admissions database such that the State Bar is required to provide access to it *if the information can be*

[3] https://assets.documentcloud.org/documents/3215383/Order-Re-Sander-v-State-Bar-CPF-08-508880.pdf *Sander v. State Bar of California, et al.*, CPF-08-508880 Order Denying Petition for Writ of Mandate, p3

[4] *Id. See also "Using data from the LSAC National Longitudinal Bar Passage Study, Sander argued that minorities who attended more elite law schools performed did not do as well as minorities who were admitted to an elite school, but chose to attend a non-elite school at which their LSAT scores and undergraduate GPAs more closely matched those of other students--the so-called 'mismatch hypothesis.' To test the mismatch hypothesis, Sander and others proposed to study candidates for the California Bar exam."* *http://www.thefacultylounge.org/2011/06/sander-v-state-bar-of-california.html*

[5] *Id.* at 4.

[6] *Id.*

provided in a form that protects the privacy of applicants and if no countervailing interest outweighs the public's interest in disclosure."[7]

In 2016, Senate Bill 387 both made the State Bar subject to the California Public Records Act, and mandated that any identifying information submitted by an applicant to the State Bar for admission be confidential, and not disclosed pursuant to any law (including the CPRA).[8] Given the legislation, the State Bar had requested an end to the case, but San Francisco Superior Court Judge Mary E. Wiss denied the motion and the case went to trial in mid-2016.[9]

At trial, the arguments largely centered around whether the protocols proposed by Sander and his experts truly anonymize the data. Expert witnesses for the Sander described four proposed data de-identification protocols and testified that each of the protocols would adequately protect the applicants' privacy.[10] An opposing expert witness for the State Bar described risks and adverse effects of re-identification or mis-identification associated with each of the four proposed protocols. The court also considered evidence regarding whether the process of de-identifying the records would require the creation of new documents under Freedom of Information laws, and how much of a burden the de-identification would put on the State Bar.

The court denied Sander's petition for the admission data on five grounds. First, the court found that the de-identification protocols, which would require the State Bar to recode its original data into new values, would amount to requiring the State Bar to create new records. One of the suggested de-identification protocols, for example, would truncate law school GPAs to two significant digits, while another would replace some applicants' actual LSAT scores with a calculated median.[11] As the CPRA (and freedom of information laws generally) only requires the disclosure of existing documents and does not require public agencies to create new records in order to respond to requests, the sophistication of the de-identification protocols was sufficient to deny the records request.

During this discussion, it should be noted that the court flatly rejected the idea of a "data enclave" providing the researchers with restricted access to the admissions database as a valid remedy under the CPRA.[12]

Next, the court analyzed the four proposed de-identification protocols under a provision of the CPRA barring disclosure of any data that "may" identify an individual bar applicant. Notably, this section of law does not contain a balancing analysis.[13] The court decided that three of the de-identification protocols could not guarantee zero risk of re-identification, thus leaving open

[7] *Id.*

[8] *Id.* at 4-5.

[9] *Id.*

[10] *Id.* at 6.

[11] *Id.* at 8-9.

[12] *Id.* at 9.

[13] *Id.* at 11.

the possibility that an individual "may" be re-identified and prohibiting the state from disclosing the records.[14] The court further decided that the final protocol would have rendered the data of minimal or no value, making disclosure unwarranted in any event.

Next, the court analyzed a state statute barring disclosure of records that would constitute an unwarranted invasion of privacy, which required the court to balance the public's interest in disclosure against the private interests in non-disclosure.[15] The court found that while there was a public interest in the activities of the State Bar in administering the bar exam and in the admissions process, the private interests in non-disclosure outweigh it.

Evidence to this point included testimony from a Professor at the University of North Carolina School of Law, who described several concerns: a fear of being re-identified from the data in view of the fact that she was one of just four black women to graduate from UCLA Law School in her class year; a fear that the information would affect her ability to obtain tenure whether or not she were correctly re-identified; and a fear of group stigmatization that might result if her and other applicants' data were used to draw broad conclusions about the abilities of black lawyers. She also testified that she provided the information with the understanding it would be kept confidential, and "probably would not have provided such information had she known that the information would be publicly disclosed."[16] Three additional witnesses provided similar testimony, including fears about group stigmatization. Accordingly, the court found that "applicants have a *strong interest* in preserving their expectation of privacy relating to information they provide to the State Bar. They also have a *strong interest* in avoiding any adverse consequences that may flow from disclosure of their data."[17]

Finally, the court analyzed a catch-all exemption in the CPRA justifying a public agency's withholding records where the public interest of *not* disclosing the record outweighs the public interest served by disclosing it. The court found that the public interest in *not* disclosing the records included:

a) protecting the general public from the adverse consequences of disclosure (e.g., the data could generate "unhealthy comparisons among lawyers, law students, and other professionals, and impede the goal of achieving greater diversity in the legal profession"),
b) protecting the State Bar's ability to collect data in the future
c) protecting the State Bar's ability to release data in the future (i.e., releasing these records would present a significant risk that any other data the State Bar releases in the future could be used to re-identify applicants)

[14] Given this finding, the court also found that the State Bar had sufficiently demonstrated that disclosure of the requested records was prohibited by state law, excusing it from disclosure under another state statute, *Government Code section 6254(k)*.
[15] *Id.* at 12.
[16] *Id.* at 13.
[17] *Id.* at 14 (emphasis added).

d) protecting the State Bar from the burden of having to implement the de-identification protocols.

The court was also concerned that permitting these records to be released would inspire others to make similar requests in the future, especially given that "the State Bar is not equipped to perform the expert analysis undertaken in this case" and "would have to hire data privacy experts."[18]

On the other hand, the court found that there is a public interest served by disclosing the records, primarily in understanding the activities of the State Bar in administering the exam, and whether different groups of applicants perform differently on the bar exam, and whether disparities in performance are the result of the admissions process or other factors (e.g., the State Bar used to publish statistical analyses of its bar passage rates, including race and gender breakdowns).[19]

Nevertheless, the court concluded that the public interest in *non*-disclosure should prevail. It describes the general public's interest in "avoiding the adverse consequences that could result from public disclosure of the records requested in this case. In particular, *the public has a strong interest in avoiding unhealthy and unwarranted comparisons among legal professionals .* . . [and] in maintaining and encouraging diversity in the legal professional, *and in avoiding the stigmatization of individuals or groups of individuals.* The release of the information requested here. . . is unprecedented and *presents significant risks of re-identification*."[20]

The court also weighed in on the technical merits of Sander's four proposed de-identification protocols. Protocols 1, 2, and 4 applied k-anonymity techniques to the admissions database, along with other supplemental anonymization techniques. K-anonymity ensures that for every record that appears in the dataset always corresponds to at least "k" indistinguishable copies.

Protocol 1 proposed to create a physical data enclave that would provide controlled access to a modified version of the admissions database. In that database, personally identifying information would be first deleted from the data, then an algorithm for k=5 would be applied. Members of the public wishing to access the data would then sign data user agreements, including contractual promises to not attempt to identify individuals or to disclose any inadvertent identifications and to take only limited items into and out of the enclave.

Protocol 2 proposed to create a public use file from the admissions database, and to apply an algorithm for k=11. The protocol would also eliminate or combine cells with only a few matches.

[18] *Id.* at 16.
[19] *Id.* at 17.
[20] *Id.* at 17.

Protocol 4 proposed to create a public data file similar to protocol 2, but would additionally: eliminate all data for a randomly selected 25% of applicants, round law school GPAs to two significant digits, and suppress unique law school GPAs in certain situations.

However, the State Bar's expert witness Dr. Latanya Sweeney calculated that after protocol 1, 47% of the data would be unique records across all fields, with 46% after protocol 2, and 31% under protocol 4. Her calculations also showed a higher likelihood of re-identification for members of minority groups. Mr. Luk Arbuckle, Sander's expert witness, however, argued that this calculation should only take into account data that is publicly knowable (e.g., law school attended, year of graduation, race/ethnicity, bar success status) – in which case, 0% of records would be unique in both protocols 2 and 4.

The court, however, decided that the protocols improperly ignored the possibility of that a person who had insider knowledge of an applicant could combine it with information from the State Bar data in order to re-identify the applicant. "Indeed," the court said, "**given the increasingly vast amount of personal data that can be found on the internet and or social media sites, as well as modern technological advancements that aid in data re-identification, this Court cannot ignore the risk that information that may not be known to the general public, such as law school GPAs, is available, and may become known and used to re-identify individual applicants from the data."**[21]

The court was also concerned that, even if individuals could not be linked to particular records, "inferences can be drawn about applicants simply by virtue of their belonging to a particular group," and noted that no evidence was presented about how protocols 1, 2, and 4 might protect individuals against attribute disclosures. The court found that "**the possibility of being linked to a negative attribute can be just as harmful as a verified, 1:1 identification."**[22]

Finally, the court also found protocol 3 inadequate. In Protocol 3, a public use file from the admissions database would be created, but with law school names eliminated from the data and with LSAT scores and law school GPAs standardized or recoded to reflect a computed average. The court determined that the utility of the data would be seriously degraded, offering "minimal or no value" to justify requiring complex anonymization procedures.

For the above reasons, Sander's petition for a writ of mandate to disclose the records was denied by the San Francisco Superior Court.

Student Unit Record Database Case Study

The Commission on Evidence-Based Policymaking was created by Congress in 2015 to, among other priorities, consider "whether a Federal clearinghouse should be created for government

[21] *Id.* at 20.
[22] *Id.* at 21.

survey and administrative data." [23] Many education groups saw this as an opportunity to revisit the need for a federal database that collects student-level data, something that was banned by Congress in 2008 due, in part, to undefined privacy concerns. These groups submitted several comments to the Commission and testified on how a student-level data system could help both students and policymakers make better decisions.

However, privacy pushback occurred almost immediately: more than half of the comments the Commission received by mid-December 2016 were vehemently opposed to a student record data base. Comments, most from parents, ranged from concerns about federal data security[24] to misunderstandings about the type of data that would be collected[25] to worries about how the data could be misused.[26] A letter from the ACLU, the Parent Coalition for Student privacy and other organizations summarized many of those concerns, stating that a federal student-level data system "could effectively create life-long dossiers on nearly every individual in the nation."[27]

The top concerns of the advocacy groups were potential breaches and unauthorized access to the student data; use of very sensitive student data for originally unintended purposes, such as law enforcement; and how a federal student-level data system could generally undermine privacy. Their letter listed numerous examples of security vulnerabilities and prior breaches in the federal government, such as the Office of Personnel Management breach in 2015 that exposed the personnel records of 22.1 million people and a November 2015 audit of the U.S. Department of Education that found "especially weak security standards."

[23] https://www.whitehouse.gov/omb/management/commission_evidence

[24] "In the past few years, data held by federal agencies has been hacked, including the personal information more than 22 million individuals, not only federal employees and contractors but also their families and friends, from the records of the Office of Personnel Management. The US Department of Education has especially weak security standards in its collection and storage of student data, and recently received a grade of "D" for its security protections." Comment ID: USBC-2016-0003-0221.

[25] "I am not comfortable with everything about a student and their family being loaded on a data base right down to astudents [sic] social media comments and social security number." Comment ID: USBC-2016-0003-0020. "We wish to register our opposition to standardized record-keeping of all children's behavior, health or educational progress. I understand that in the 'big picture' this information could be valuable to researchers. However, our constitution does not exist to make life easier for those conducting research or formulating government policies. Our constitution exists to protect the independence and liberty of the individual from governmental overreach." Comment ID: USBC-2016-0003-0058.

[26] "Collecting and sharing personal data/information on our children (though not exclusively to our children) always [sic] runs the risk of profiling them, which will result in great prejudices and injustices when it makes its way into the wrong hands; which it surely will." Comment ID: USBC-2016-0003-0137.

[27] ACLU/Parent Coalition Comment Letter. http://www.studentprivacymatters.org/wp-content/uploads/2016/11/letter-to-CEP-w-signers-final-11.14.16-pdf.pdf.

Next, the letter expressed concern that extremely sensitive information held in state longitudinal data systems, “such as student immigration status, disabilities, disciplinary incidents, and homelessness status,” would be shared with a federal student-level data system, combined with other sensitive data like “military service, tax returns, [and] criminal and health records,” and be used for originally unintended “purposes that should never be allowed.” While the letter acknowledged the current intention to just use a federal student-level database for research purposes, it noted that England’s National Pupil Database, also created just for research purposes, has been used by police for purposes including curbing “abuse of immigration control.”

Finally, the letter expressed concern about increased surveillance by the federal government in general, noting that, despite the potential ability of data to solve “complex problems,” the government holds a responsibility to “protect the privacy of their most personal information, especially that of vulnerable children.” The letter urged the Commission to avoid creating “life-long dossiers on nearly every individual” in the country, and instead only allow aggregate and de-identified information already maintained by states or districts for research and policy decisions.

These and other cases [INCLUDE MORE BUT SHORTER ONES?] illustrate some of the many risks associated with researcher access to private data.   These risks are not limited to re-identification, but also include: ethical concerns about the lack of effective risk-benefit analysis; expectations or rights concerning the performance of societally beneficial research; problems with consent, informed or otherwise; law enforcement and regulatory access to data; discriminatory uses of sensitive data; proprietary interests and fear of commercialism; lack of trusted intermediaries; lack of standard contracts and researcher certification; media and policymaker confusion; emerging regulations and standards in other countries, especially the EU; and security and data breach concerns.

III. Competing Objectives

On what basis should a data-access system for researchers be judged successful? Here are three criteria:

a. **Accuracy Objective**: The system should promote *research reliability*. That is, the results should be robust, unbiased, fully documented, and easily contestable or validated by other researchers. Judgement on this score rests with academic editors, referees, and other gatekeepers.

b. **Privacy Objective**: The system should promote *ethical integrity*. That is, the studies undertaken should be publicly defensible with respect to privacy protection for both natural persons and organizations (even though this is not always the only consideration that enters into legal or ethical cost/benefit calculations). Judgement on this score rests with the public generally, but especially with anyone whose data may be under study.

c. **Efficacy Objective**: The system should promote *practical sustainability*. That is, its demands should not be too burdensome in terms of the financial support, popular support, legal support, or bureaucratic support necessary to maintain smooth operations. Judgement on this score rests with system funders and managers as well as data donors and users.

As usual with multiple objectives, pursuing any one affects the others as well. There is a fundamental trade-off between accuracy and privacy, for example. Increasing privacy protection necessarily entails additional data obfuscation or reproducibility obstructions, as discussed further in the next section. This is not just an observation. Nor is it simply that no one has figured out how to reconcile privacy with research accuracy yet. These two goals are inevitably at odds. It is a theorem, not a temporary challenge to be overcome, that too much accuracy compromises privacy. Technological advances may change and improve the options available, but not the fact that some trade-offs will be necessary.

One consideration when deciding how to supply data for a given research project is, therefore, the appropriate balance to strike between the privacy and accuracy objectives. This will depend on the nature of the study and the nature of the data along with other factors. Another kind of consideration involves familiar constraints on time, funding, attention, etc. Because there are all sorts of monetary and non-monetary costs to supplying data, some priority setting among projects must take place. This is where project-specific expectations about policy applications and societal benefits would naturally come into play.

So how to handle competing objectives more specifically? When considering a request to access data in order to conduct research, there may be many ethical, practical, or other reasons to proceed or not. But to the extent privacy *per se* is the concern of interest, here are three steps to take:

Step 1: **Admissibility**. Compile a menu of all the protocols available that would govern how the data are released and used. Options might range from de-identification to nondisclosure agreements (see next section for an overview of alternatives). Rank each according to the privacy it affords, then according to the accuracy it affords. Call a possible protocol admissible if: a) there is no other protocol available that would provide at least as good privacy, at least as good accuracy, and do better in one respect or the other; and b) the privacy protection exceeds some minimum acceptable level and the accuracy exceeds some minimum acceptable level.

Step 2: **Appropriateness**. From among the admissible protocols, some might be more appropriate for certain kinds of data than others. For example, sensitive IRS information typically can only be studied under the strictest privacy protections. It is rarely even possible to check the accuracy of any such results independently. In contrast, scanner card data about individuals' grocery purchases are already sold and studied by corporations. Before making such datasets available for academic research, there may be little point in fiddling with the information at all other than replacing each name with a random number—especially since the names are not necessarily verified anyway.

Step 3: **Affordability**. The marginal cost of providing data to one more researcher may sometimes be small, but there are also fixed costs for infrastructure, legalities, bandwidth, documentation, etc. Besides using up dollars, each research study necessarily leaks some privacy and accuracy. So it is neither practical nor desirable to make a particular dataset containing sensitive information accessible to everyone who might ever be interested. Priority should instead be given to projects that stand to provide the most net benefit to society by advancing knowledge and understanding. This cannot be predicted with certainty ahead of time (if we knew what the answers would be in advance of conducting the study, we would not call it research), but academic panels, funders, and researchers do make judgments about such matters all the time.

What about the minimum levels mentioned in Step 1 above? These may vary from one time or place to another depending on societal norms. Here are two suggestions, though. With regard to privacy, any protocol that could be overcome in a few days by a skilled hacker should be ruled out as providing substandard protection. One can imagine "white hat" teams charged with randomly checking. It may not make sense, in any case, to set minimal privacy guarantees for research data that are much more stringent than those routinely accepted by customers as governing similar data gathered by corporations for commercial use.

As for minimal accuracy, there is an argument that exploratory research projects are so unreliable that they are never warrant a sacrifice of privacy. Why not perform such preliminary work on synthetic datasets rather than sensitive ones? Access to the actual data could be reserved for confirmatory research that tests a specified hypothesis using a pre-registered analysis plan. This prevents "p-hacking," that is, dredging through data until the researcher sooner or later finds coincidences that can masquerade as statistical significance.[28] The somewhat counterintuitive truth is that, the more statistical information of any kind that is released about a given dataset, the harder it becomes for anyone to discover results that are genuinely significant.[29] This has nothing to do with protecting privacy, but does suggest granting data access to researchers only sparingly and for well-designed projects.

Note that this whole way of looking at decisions to provide data is quite different from simple attempts at weighing risks against benefits. That is not easy, as the legal cases presented in Section II illustrate. It is never clear, for example, what common scale can be used to measure both risks and benefits other than dollars. But then translating everything into monetary terms often raises more questions than it answers.

In contrast, the approach here assumes that, for every research project in need of sensitive information, there is in principle a data-use protocol that would provide reasonable privacy protection. It may be inconvenient or expensive, but if academics can even study IRS returns under appropriate circumstances, then there should be ways of allowing access to nearly any kind of data while still protecting privacy. The more privacy protection required, however, the less accurate the results we can expect. As mentioned above, this is due to data obfuscation or reproducibility obstructions. Without the prospect of sufficiently accurate results, some research projects may not be worth pursuing. But that is not so much a privacy matter as one of efficacy and affordability. Those judgements always have to be made about research proposals. The use of private data does not change that.

Similarly, research proposals often raise ethical concerns, too, regardless of whether the data to be studied is private or not. Certainly, the release of certain information not otherwise available could be unethical—especially if confidentiality was promised to individuals described by the data. That is why research on sensitive data should not proceed without appropriate privacy protections in place. In that case, questions about how the results obtained are applied may still raise very serious ethical issues, but then those issues need not be viewed as questions about privacy *per se.* Again, there are rules and procedures for handling ethical concerns about proposed research, largely through Internal Review Boards (IRBs) as described in Section VI below.

Biohazard research is, in some ways, analogous. Work with certain dangerous but not very threatening material goes on routinely in universities and other laboratories. Precautions,

[28] Simonsohn, U., Nelson, L.D. and Simmons, J.P., 2014. P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), p.534.

[29] That is because statistical significance consists of finding a “p-value” of less than .05, where the p-value should be understood as the probability, conditioned on *all* the null hypotheses tested to date, of finding data supporting those null hypotheses at least as extreme as that observed.

procedures, special equipment, and training keep everyone safe. And there are spot checks as well as emergency plans just in case. For research that poses greater risks, there are more stringent levels of protections that range from Biosafety Level 1 to 4. Research on HIV or Salmonella, for example, requires the second level of protection, BSL-2. Ebola can only be handled at BSL-4 facilities. These must have special airlocks, positive pressure suits, decontamination processes, etc. There are only a few such facilities, and the work there is cumbersome and expensive. Attention to developing an Ebola vaccine may temporarily crowd out work on other scary viruses or, for that matter, more poorly understood pathogens, but the research is proceeding and promising. Technological and scientific advances may change classifications and capabilities over time, but the system itself is a trustworthy means of facilitating important research.

IV. Surveying Alternatives

Someone is collecting data about me. An independent researcher would like to study the dataset being compiled. It contains personal information about me and lots of other people, but the researcher claims to be interested only in aggregate statistics and not in sensitive details about any individual.

What are the data-handling protocols that can protect my privacy while also facilitating the collection, analysis, and release of information for scientific purposes? Under what restrictions should researchers be able to see which kinds of personal or proprietary information?

Designing protocols for research on confidential topics entails trade-offs between accuracy and privacy. Of course, any suggestion that would make empirical work less precise, open, representative, or replicable seems contrary to the needs and values of science. That is why a careful re-examination has begun of what "accuracy" or "privacy" should mean in this context, and what it therefore means to balance one against the other.

In particular, my attitude towards research on my personal data depends both on how well the investigator's protocol could generate valuable statistics and on how well it can also protect confidential details. There is always some risk of a leak, and so it makes little sense to participate in a study unless it is capable of producing valid results. Even assuming away human error or maliciousness, assurance would be welcome that the researcher could not use some other protocol to obtain findings with the same or better scientific reliability while also providing more reliable privacy protection.

Common research protocols sometimes purport to deliver accuracy or privacy that they do not. One useful way to classify proposed protocols is according to what stage in the research process data obfuscation occurs.[30] Three possibilities are: during input; during computations; or during output. Turning restrictions either on or off at each stage gives rise to eight possible categories (see table). Illustrations of each follow, starting with traditional methods whose strengths and shortcomings motivate more recent approaches.

[30] Goroff, D.L., 2015. Balancing privacy versus accuracy in research protocols. *Science*, *347*(6221), pp.479-480.

Table: At which of three research stages do protocols impose restrictions

	Input	Computation	Output	Protocol Example
1				Open Data
2			X	Data Enclave
3	X		X	Nondisclosure Agreement
4		X		Anonymization
5	X			Randomized Response
6	X	X		Multiparty Computation
7	X	X	X	Fully Homomorphic Encryption
8		X	X	Differential Privacy

1. No Restrictions: **Open Data**

Suppose a researcher wishes to study faculty wages. Some states publish names, salaries, and other information about university employees in downloadable formats. In that case, there are no restrictions on data collecting or sampling, linking or analyzing, or release and reuse.

This is the ideal supported by "open data" advocates.[31] It facilitates accuracy but not confidentiality. People who care about keeping their pay private can at least know about such disclosure policies before they decide to take a position.

2. Restricted Output: **Federal Data Enclaves**

Suppose a researcher wishes to study U.S. wage and employment trends more broadly. The most comprehensive datasets are compiled from state and federal administrative records under the Longitudinal Employment and Household Dynamics Program (LEHD). Academics can apply for access to personal information at one of several "Research Data Centers" run by the U.S. Census Bureau.[32] If approved, a researcher's "Special Sworn Status" makes him or her subject to prosecution for misuse of private information under the same terms as a government official. Computations typically take place on site, in a "data enclave" disconnected from the rest of the world. To protect against improper disclosures, papers must be approved by the Census Bureau before they can be released.

[31] See, for example, the Open Knowledge Foundation, the Center for Open Science, the Mozilla Science Lab, and the Berkeley Initiative for Transparency in the Social Sciences.

[32] https://www.census.gov/ces/rdcresearch/

Typically, the Census Bureau checks that any information reported is aggregated enough to obfuscate the identity of individuals. This is akin to pixelating faces in photographs to hide identities, which works well for unfamiliar people and less well for those who you have other information about. Such procedures have produced no known security breaches to date and are gradually becoming less cumbersome, but the replication of research results obtained at a data enclave remains problematic.

3. Restricted Input and Output: **Commercial Non-Disclosure Agreements**

Companies often draw inferences about users' salaries and other characteristics based on their online behavior. Inaccuracies occur at the input stage because of the indirect, obscure, and irremediable nature of this process, not to mention potential sample bias.   Researchers rarely gain access to such datasets without signing "non-disclosure agreements" that give the company control over what private or proprietary details may be released.

Again, this arrangement usually precludes replication of the results or reuse of the data. The *American Economic Review*, a premier academic journal whose authors are supposed to post the data they use, reported having to waive this requirement for nearly half the empirical papers published in 2014 because of such non-disclosure agreements.[33]

4. Restricted Computation: **Anonymization**

New York City recently released "anonymized" data about every taxi trip taken in 2013. Hackers quickly re-identified all this data by exploiting weak techniques used to encode the information and by linking with other publicly available datasets. Not only is it possible to track the earnings of each cabbie by name, you can also trace the times, fares, and tips of trips made by celebrities, or map all the precise GPS coordinates on the other end of trips to or from The Hustler Club, for example.[34]

This joins many other examples of datasets that were released with assurances that they had been scrubbed of any "personally identifiable information" (PII), but that nevertheless were easily linked with other public information to yield private confidences--including Governor William Weld's health records[35] as well as the movie rental histories of NetFlix customers.[36] It has been claimed that 87% of the U.S. population can be uniquely identified by just three pieces

[33] See L. Einav, J. Levin, *Science* 346 (2014)

[34] http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset

[35] L. Sweeney. *Journal of Law, Medicine, and Economics* 25:98-110.

[36] A. Narayanan, V. Shmatikov. Proc. IEEE Symposium on Security and Privacy, 2008, 111-125.

of information: gender, zip code, and date of birth.[37] Then anyone can look up or purchase the individual's name and social security number, along with all kinds of other intimate data.[38]

In most of these cases, it was correct that re-identifying individuals from the de-identified dataset alone would be extremely difficult if not impossible. The privacy violations came about instead by linking the anonymized data with other publicly available datasets. The possibility of such "linkage attacks" would seem to make safety guarantees for a given anonymization dependent on knowing all the current and future datasets that could ever conceivably be used to re-identify individuals. Clearly this is not a practical approach to privacy protection. That is one of the many reasons why—despite creative efforts to aggregate, average, edit, adjust, or otherwise impose obfuscation at the computational stage—experts are now reaching the conclusion that "sanitizing data doesn't" and "de-identified data isn't."[39]

5. Restricted Input: **Randomized Response**

Suppose a researcher wants to estimate what percentage of a group have incomes below the poverty line without compromising anyone's confidentiality about this. Give each person a coin to flip without anyone else seeing the outcome. If it lands heads, they should truthfully answer yes or no to the question of whether they are below the poverty line. If it lands heads, they should flip again. If the second toss is a head, they should again answer truthfully, but if the second toss is a tail, they should lie—that is, answer yes if they are above the poverty line and no if they are below. Then twice the number of yes responses minus a half is a good estimate of the fraction of the group who actually are below the poverty line.[40]

Even if you know who answered what, that does not tell you who is impoverished. The usefulness of this technique depends on having lots of participants who are all willing to follow instructions. There are also privacy-preserving variants that are more efficient estimators, but some accuracy is sacrificed in any case.

6. Restricted Input and Computation: **Multi-Party Computation**

Suppose a researcher would like to calculate the average salary of group, but without anyone ever communicating his or her own salary. Surprisingly, this can be done quite precisely as long

[37] http://aboutmyinfo.org/

[38] C. O'Neil. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing Group (NY), 2016.

[39] C. Dwork in *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (eds), Cambridge University Press, 2014.

[40] If *r* is the reported fraction of yes responses and *p* is the true fraction, then the expected value of *r* is $E(r)=\frac{p}{2}+\frac{p}{4}+\frac{(1-p)}{4}$.

as everyone cooperates. Say there are three people. Each person generates two random numbers and gives one to each of the other two participants. Next, each person adds the two random numbers she generated to her own salary, subtracts the two numbers she was given, and reports the result. Adding those results and dividing by three gives the average salary.[41] Similarly, special and more convoluted computations can secretly carry out operations beyond just taking averages.[42]

Although no individual's salary was ever communicated, this protocol does not necessarily protect the privacy of those who participate. If, for example, all but one of them collude by using the same method to compute their average salary, that new group could easily deduce what the salary was of their original colleague. Applications of blockchain technology have been proposed to make deviations from agreed protocols more easily detectable.[43]

7. Restricted Input, Computation, and Output: **Fully Homomorphic Encryption**

Banking and other sensitive information routinely travels over the internet without interception. Suppose that you could not only send your salary information to a researcher using similar or stronger encryption, but that the researcher could perform computations and return the encrypted results to you without ever being able to decrypt either your information or the results. Long thought impossible, such "fully homomorphic encryption" methods have now been devised.[44] Though practical applications are coming online, many algorithms are still too slow for practical applications. Proposed protocols would perform regressions and other analyses on encrypted data supplied by survey participants, for example, but then only allow

[41] In other words, let S_i denote the secret salary of person *i,* and let R_{ij} denote the random number generated by person *i* and given to person *j*. Then person *i* reports the result X_i where

$$\begin{aligned} X_1 &= S_1 + (R_{12} + R_{13}) - (R_{21} + R_{31}) \\ X_2 &= S_2 + (R_{21} + R_{23}) - (R_{12} + R_{32}) \\ X_3 &= S_3 + (R_{31} + R_{32}) - (R_{13} + R_{23}) \end{aligned}$$

When you add up these three equations, all the random numbers cancel so you get:

$$X_1 + X_2 + X_3 = S_1 + S_2 + S_3 .$$

This says that the sum of the numbers people report actually equals the sum of their salaries. So just divide by three, and voilà.

[42] M. Prabhakaran and A. Sahai. *Secure Multi-Party Computation*, IOS Press, 2013. For an application to collecting regulatory data, see E. Abbe, A. Khandani, A. Lo, *American Economic Review* 102:3 (2012).

[43] Zyskind, G., Nathan, O. and Pentland, A., 2015. Enigma: Decentralized computation platform with guaranteed privacy. arXiv preprint arXiv:1506.03471.

[44] C. Gentry. STOC. Vol. 9. 2009.

the statistics to be decrypted if those participants verify that the calculations have been done properly and to their satisfaction.[45]

Once practical, fully homomorphic encryption could have profound implications for the privacy of everything from cloud computing to search engines, and from tax preparation to "personal data lockers." But even effectively hiding the inputs to research does not necessarily preserve the privacy of participants, especially if the statistical findings are subject to "linkage" or "differencing" attacks. As an example of the latter, consider that asking two simple questions—how many employees of this company make more than $1 million in salary, and how many employees of this company who are not the CEO make more than $1 million—tells you whether or not the CEO makes over a million. It may seem straightforward to rule out lines of questioning like this. Provably, however, no algorithm can reliably determine whether a given set of questions that seem to ask only about statistical aggregates would nevertheless have answers that, taken together, reveal private information.[46]

8. Restricted Computation and Output: **Differential Privacy**

Consider a dataset *D* that contains my personal information in one "row," and another dataset *D'* that is missing that row but otherwise the same. Two datasets are said to be adjacent if they differ by one row like this. A research protocol would certainly count as privacy preserving if it could not distinguish between D and an adjacent D'. It also would not be very useful. But what if the protocol could barely make such a distinction? Specifically, consider the probabilities that it generates a given answer to a given question when applied to D as compared to D'. The ratio of those two probabilities should be as close to one as possible. In fact, the log of that ratio measures the loss of privacy incurred when the protocol answers the given question.[47] If the log of that probability ratio is always less than ε for any adjacent datasets, the protocol is said to provide ε-differential privacy.

Dwork et. al. not only formulated this definition and showed it captures basic intuitions about privacy loss, she also devised explicit research protocols that provide ε-differential privacy.[48] The data is held by a trusted curator who only accepts certain questions from the investigator. Calculations are performed behind a firewall, but the answers are returned only after adding on a small amount of carefully chosen noise. It suffices, for example, to draw that noise from a Laplace distribution with parameter 1/ε when responding to a counting query. Other aggregate statistics, including regression coefficients and contingency tables, can be handled similarly.

[45] A. Lopez-Alt, E. Tromer, V. Vaikuntanathan. STOC. 2012.

[46] C. Dwork, *op. cit.*

[47] For a protocol *M* that yields research result $\gamma = M(D)$ when applied to database *D,* Dwork defines the loss of privacy as $L = \ln \frac{\Pr[M(D) = \gamma]}{\Pr[M(D') = \gamma]}$ where *D'* is adjacent to *D.*

[48] C. Dwork, F. McSherry, K. Nissim, A. Smith in *Theory of Cryptography Conference* (TCC), Springer, 2006.

There is a limit on the number of such questions that can be allowed, however, since each one could deplete any given privacy budget by as much as ε.

Conceptually, choosing the parameter ε for a differentially private protocol determines how the research will trade accuracy against privacy. The smaller ε is, the less leakage of information--but at the cost of more noise. As a practical matter, how to provide differential privacy by designing, implementing, or combining various algorithms is the subject of intense research. One promising example is the Census Bureau's *OnTheMap* Project that provides probabilistic differential privacy.[49] A "synthetic database" has been constructed by carefully perturbing and aggregating actual payroll tax records in each state. By querying this dataset, members of the public can then receive approximate but quite accurate answers to a large class of counting and geographic questions, including ones of the form: how many people above a certain age who earn at least this much live in this town but work in that industry located in a specific nearby city. Similar differentially private protocols have also been implemented for studying cell phone records.[50]

The examples above emphasize salary data. As norms change, some people consider this information more sensitive than others.[51] Different kinds of private information collected by health providers, educators, or government officials are subject to specific laws such as HIPPA, FERPA, and the Privacy Act of 1974 respectively. No existing legislation specifically regulates or facilitates the use of personal information by researchers, however. Other countries, including Sweden, Austria, Germany, Denmark, and Norway, each have a well-funded and well-organized agency that enables sophisticated privacy preserving research on comprehensive population and tax records. Europe tends to regulate what companies can do with personal data more strictly than what governments can do, whereas the opposite is true in the United States.

Each society must decide how it will trade off privacy concerns against research potential. This differs from my choice between, say, taking an expensive vacation or buying a new car. That decision affects only me. The reliability levels of privacy protection and of scientific research produced by society are, in contrast, public goods like defense or lighthouses that are neither excludable nor rival. On the other hand, the trade-off is like any other between commodities in that no one should accept a research protocol if there is another that would deliver more accuracy without sacrificing privacy or more privacy without sacrificing accuracy.

How well society understands, facilitates, and regulates privacy-preserving research will, in turn, determine the extent to which "big data" allows everyone to benefit from advances in

[49] http://onthemap.ces.census.gov

[50] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J. and Varshavsky, A., 2011, June. Identifying important places in people's lives from cellular network data. In *International Conference on Pervasive Computing* (pp. 133-151). Springer Berlin.

[51] Acquisti, Alessandro, Curtis Taylor and Liad Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature*, 54(2): 442-92.

empirical behavioral and social science or only those private interests who hold enormous and growing stores of sensitive information about us all.

V. Risking Re-identification

In the overview of privacy protecting research protocols presented above, it is useful to distinguish between those that affect the collection of data as opposed to those that only deal with computations and output. The former, if built into a data project from the beginning, can help ease privacy concerns later. Researchers should therefore be aware of techniques like Secure Multi-Party Computation or Fully Homomorphic Encryption before setting out to collect data, and encouraged to use them when appropriate.

In many cases, however, researchers are eager to work with datasets that have already been compiled for other purposes. For so-called “administrative data” like this, privacy protecting protocols can only be applied at the computational and output stages of research. Because so much of the evidence base for policymaking derives from administrative data, it is worth focusing in more detail on the comparing and contrasting post-collection protocols. These are data enclaves, de-identification, and differential privacy.

For these post-collection scenarios, the issue is re-identification risk since nothing has been done to initially obscure or protect personal information when it was first recorded. Most controversial is the practice of de-identification. If done carefully, it is often considered acceptable in many applications. But there are also many actual and potential examples of re-identification. By contrast, research protocols that enforce differential privacy are effectively meant to rule out the identification of individuals altogether, not matter what post-processing or linkages might ever be attempted. Data enclaves are similarly considered quite safe. Re-identification is unheard of, but the theoretical possibility remains that the release of precise statistics—even if aggregated, averaged, or otherwise “sanitized”—could allow linkage, differencing, or other attacks to compromise privacy.

[ADD paragraphs here about data enclaves, including FSRDC’s as well as NORC, ICPSR, and ADRN-UK (especially its linking capabilities). Virtual enclaves, too. The point being that there is little if any re-identification risk (maybe more than anyone thought?), but accuracy cannot be checked easily if ever.]

Re-identification Risk

In recent years, a number of incidents have shown that data sets that were claimed to be de-identifiied were in fact vulnerable to re-identification attacks. One case involved the public release of de-identifiied hospitalization records of state employees including then-Massachusetts Governor Weld. Another was the posting of twenty million search queries of

650,000 AOL users to a site aimed at researchers and an additional; case focuses on more than 100 million ratings from over 480,000 Netflix customers on nearly 18,000 movie titles. All three incidents involved linkage attacks, in which an individual or entity trying to re-identifiy a data subject takes advantage of auxiliary or background information to link an individual to a record in the de-identifiied data set.

For many, these re-identification cases have called into question effectiveness of de-identification techniques.[52] Many leading experts now doubt the extent to which de-identification remains a credible method for using and deriving value from large data sets while protecting privacy. Some argue that it is currently impossible to eliminate privacy harms from publicly released data using de-identification due to the growing availability of background data, which allow attackers to identify data subjects by mounting linkage attacks.

A number of leading experts (El Imam, Barth Jones) that have long worked on the de-identification of health data vigorously dispute the claim that these de-identification attacks should be viewed as critiques of de-identification techniques. They note that in some of the cases (AOL, Netflix) the data sets were not anonymized in any credible manner or that the de-identification attacks were limited in scope, proving only the possibility of re-identifying public figures for who extensive data was available (Weld). Members of the disclosure control community counter that despite the theoretical and demonstrated ability to mount such attacks, the likelihood of re-identification for most data sets remains minimal.

Some commentators argue that the competing experts fall into distinct camps of "pragmatists" and formalists."[53] In general, pragmatists share an expertise in de-identification methods and value practical solutions for sharing useful data to advance the public good. Accordingly, they devote a great deal of effort to devising methods for measuring and managing the risk of re-identification for clinical and other specific disclosure scenarios.[54] In sharp contrasts, formalists are less concerned with finding practical solutions than with achieving mathematical rigor in defining privacy, modeling adversaries, and quantifying the probability of re-identification. They seek provable privacy guarantees using methods first developed in cryptography and more recently applied in theoretical research associated with differential privacy.[55]

[52] See Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. REV. 1701, 1717-23. (2010).

[53] See Ira Rubinstein & Woodrow Hartzog, Anonymization and Risk, 91 WASH. L. REV. 703, 714-17 (2016).

[54] See, e.g., KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (2013).

[55] Differential privacy has been described as "a set of techniques based on a mathematical definition of privacy and information leakage from operations on a data set by the introduction of non-deterministic noise. Differential privacy holds that the results of a data analysis should be roughly the same before and after the addition or removal of a single data record (which is usually taken to be the data from a single individual). In its basic form differential privacy is applied to online query systems, but differential privacy can also be used to produce machine-

Pragmatists take into account what they believe is the very low risk of the availability of information that can be leveraged for de-identification attacks and consequently give little weight to well-known re-identification attacks, noting the very limited number of people ever identified in all known attacks. Formalists object to such studies on the grounds that these efforts to quantify the efficacy of de-identification "are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do."[56] Unlike the pragmatists, they take very seriously proof-of-concept demonstrations of re-identification, while minimizing the importance of empirical studies showing low rates of re-identification in practice.

De-identification Solutions: Technology

Among the leading models for quantifying the privacy protection offered by de-identification are *k-anonymity* and *differential privacy*.

From **NIST Special Publication 800-188 (2nd DRAFT)**
http://csrc.nist.gov/publications/PubsDrafts.html#SP-800-188

De-Identifying Government Datasets
K-anonymity[57] *is a framework for quantifying the amount of manipulation required of the quasi-identifiers to achieve a given desired level of privacy. The technique is based on the concept of an equivalence class, the set of records that have the same values on the quasi-identifiers. (A quasi-identifier is a variable that can be used to identify an individual through association with other information.) A dataset is said to be k-anonymous if, for every specific combination of quasi-identifiers, there are no fewer than k matching records. For example, if a dataset that has the quasi-identifiers (birth year) and (state) has k=4 anonymity, then there must be at least four records for every combination of (birth year, state). Subsequent work has refined k-anonymity by adding requirements for diversity of the sensitive attributes within each equivalence class (known as l-diversity*[58]*) and requiring that the resulting data are statistically close to the original data (known as t-closeness).*[59]

learning statistical classifiers and synthetic data sets." SIMSON L. GARFINKEL, NAT'L INST. OF STANDARDS & TECH., DE-IDENTIFICATION OF PERSONAL INFORMATION (NISTIR 8053) 4 (2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

[56] ARVIND NARAYANAN & EDWARD W. FELTEN, NO SILVER BULLET: DE-IDENTIFICATION STILL DOESN'T WORK (2014), http://randomwalker.info/ publications/no-silver-bullet-de-identification.pdf.

[57] Latanya Sweeney. 2002. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557-570. DOI=10.1142/S0218488502001648 http://dx.doi.org/10.1142/S0218488502001648

[58] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[59] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy beyond k-anonymity and l-diversity". ICDE (Purdue University).

Differential privacy[60] *is a model based on a mathematical definition of privacy that considers the risk to an individual from the release of a query on a dataset containing their personal information. Differential privacy is also a set of mathematical techniques that can achieve the differential privacy definition of privacy. Differential privacy prevents both identity and attribute disclosure by adding non-deterministic noise (usually small random values) to the results of mathematical operations before the results are reported.*[61] *Unlike k-anonymity and other de-identification frameworks, differential privacy is based on information theory and makes no distinction between what is private data and what is not. Differential privacy does not require that values be classified as direct identifiers, quasi-identifiers, and non-identifying values. Instead, differential privacy assumes that all values in a record might be identifying and therefore all must potentially be manipulated.*

Differential privacy's mathematical definition holds that the result of an analysis of a dataset should be roughly the same before and after the addition or removal of the data from any individual. This works because the amount of noise added masks the contribution of any individual. The degree of sameness is defined by the parameter $\boldsymbol{\epsilon}$ *(epsilon). The smaller the parameter* $\boldsymbol{\epsilon}$*, the more noise is added, and the more difficult it is to distinguish the contribution of a single individual. The result is increased privacy for all individuals, both those in the sample and those in the population from which the sample is drawn who are not present in the dataset. The research literature describes differential privacy being used to solve a variety of tasks including statistical analysis, machine learning, and data sanitization.*[62] *Differential privacy can be implemented in an online query system or in a batch mode in which an entire dataset is de-identified at one time. In common usage, the phrase "differential privacy" is used to describe both the formal mathematical framework for evaluating privacy loss, and for algorithms that provably provide those privacy guarantees.*

Note that the use of differentially private algorithms does not guarantee that privacy will be preserved. Instead, the algorithms guarantee that the amount of privacy risk introduced by data processing or data release will reside within specific mathematical bounds.

[60] Cynthia Dwork. 2006. Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12. DOI Foundations of Differential Privacy, in Foundations and Trends in Theoretical Computer Science Vol. 9, Nos. 3–4 (2014) 211–407, https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf; http://dx.doi.org/10.1007/11787006_1

[61] Cynthia Dwork, Differential Privacy, in ICALP, Springer, 2006.

[62] Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy* (Foundations and Trends in Theoretical Computer Science). Now Publishers, August 11, 2014. http://www.cis.upenn.edu/~aaroth/privacybook.html

When data releases containing information about the same individual accumulate, then privacy loss accumulates. Organizations should keep this in mind and try to assess the overall accumulated risk, and differential privacy can be used to help them make this assessment.

Comparing traditional disclosure limitation, k-anonymity and differential privacy, the first two approaches start with a mechanism and attempt to reach the goal of privacy protection, whereas the third starts with a formal definition of privacy and has attempted to evolve mechanisms that produce useful (but privacy-preserving) results. These techniques are currently the subject of academic research, so it is reasonable to expect new techniques to be developed in the coming years that simultaneously increase privacy protection while providing for high quality of the resulting de-identified data.

One major concern for researchers using de-identified data is that they are often unlikely to know how inaccurate their statistical results are due to statistical distortions introduced by the de-identification process.

De-identification: Disclosure Control, Risk based Model

De-identification based on the removal of identifiers and transformation of quasi-identifiers is the approach for de-identification currently widely in use. It has long been used by federal statistical agencies and the healthcare industry, but it is not based on formal methods for assuring privacy protection. This process can take into account factors such as whether a data set will be made public, shared with trusted partners, used only internally and whether the data is sensitive or low risk.

Below is a sample process for de-identifying data by removing identifiers and transforming quasi-identifiers:[63]

Step 1. Determine the re-identification risk threshold. The organization determines acceptable risk for working with the dataset and possibly mitigating controls, based on strong precedents and standards

Step 2. Determine the information in the dataset that could be used to identify the data subjects. Identifying information can include:

a. **Direct identifiers**, such as names, phone numbers, and other information that unambiguously identifies an individual.

[63] This process is based on a process developed by Professors Khaled El Emam and Bradley Malin. See K. El Emam and B. Malin, "Appendix B: Concepts and Methods for De-Identifying Clinical Trial Data," in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015.

b. **Quasi-identifiers** that could be used in a linkage attack. Typically, quasi-identifiers identify multiple individuals and can be used to triangulate on a specific individual.

c. **High-dimensionality data**[64] that can be used to single out data records and thus constitute a unique pattern that could be identifying, if these values exist in a secondary source to link against.[65]

Step 3. Determine the direct identifiers in the dataset. An expert determines the elements in the dataset that serve only to identify the data subjects.

Step 4. Mask (transform) direct identifiers. The direct identifiers are either removed or replaced with pseudonyms. Options for performing this operation are discussed below. [NEED REFERENCE?]

Step 5. Perform threat modeling. The organization determines the additional information they might be able to use for re-identification, including both quasi-identifiers and non-identifying values that an adversary might use for re-identification.

Step 6. Determine the minimal acceptable data quality. In this step, the organization determines what uses can or will be made with the de-identified data.

Step 7. Determine the transformation process that will be used to manipulate the quasi-identifiers. Pay special attention to the data fields containing dates and geographical information, removing or recoding as necessary.

Step 8. Import (sample) data from the source dataset. Because the effort to acquire data from the source (identified) dataset may be substantial, some researchers recommend a test data import run to assist in planning.[66]

Step 9. Review the results of the trial de-identification. Correct any coding or algorithmic errors that are detected.

Step 10. Transform the quasi-identifiers for the entire dataset.

[64] Charu C. Aggarwal. 2005. On k-anonymity and the curse of dimensionality. In Proceedings of the 31st international conference on Very large data bases (VLDB '05). VLDB Endowment 901-909.

[65] For example, Narayanan and Shmatikov demonstrated how the set of movies that a person had watched could be used as an identifier, given the existence of a second dataset of movies that had been publicly rated. See Narayanan, Arvind and Shmatikov, Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125.

[66] Khaled El Emam and Bradley Malin, Concepts and Methods for De-Identifying Clinical Trial Data, Appendix B, in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, National Academies Press, 2015.

Step 11. Evaluate the actual re-identification risk. The actual identification risk is calculated. As part of this evaluation, every aspect of the released dataset should be considered in light of the question, “can *this* information be used to identify someone?”

Step 12. Compare the actual re-identification risk with the threshold specified by the policy makers.

Step 13. If the data do not pass the actual risk threshold, adjust the procedure and repeat Steps 11 and 12. For example, additional transformations may be required. Alternatively, it may be necessary to remove outliers.

The challenge of this model is that it is often used without serious application of all its requirements. For example, online ad companies may remove direct identifiers and claim their cookie linked profiles are anonymous. Contractual controls may not fully restrict data sharing. Identifiers may be transformed in ways that are reversible.

Expertise to implement this type of de-identification resided primarily in the health care sector, due to its experience in complying with HIPAA, and there is a very limited pool of experts who can provide expert disclosure control guidance or certification. Training a cadre of experts available for de-identification guidance and consulting would be an invaluable contribution for this sector. Additionally, formalizing the administrative contracts and controls necessary for non-public data sharing among entities in this sector would be useful.

VI. Building Institutions

To generate evidence for policymakers, there are ways that researchers can access sensitive data while protecting privacy. Solutions not only exist, many are already in use somewhere or another. But nearly everything is being done on an *ad hoc* basis, project by project, based on one-off Data Usage Agreements. Rather than learning much from experience, from best practices, or from other countries, the incentives for researchers are just to do what they can to get out their next paper. The incentives for companies, government agencies, or other data holders are not to risk much time or reputation dealing with researchers' requests. And the incentives for the individuals described by the data are to worry about other matters until some harm is done and it is too late. For everyone, the *transactions costs*—monetary or otherwise—are so high that more responsible behavior is thwarted.

Economic theory teaches that difficulties due to high transactions costs are often ameliorated by forming *institutions*. This view, often associated with Nobel Laureate Oliver Williamson, emphasizes the study of transactions, institutions, and their governance rather than goods, services, and their markets.[67] The way we tend to think or talk about a piece of data as a commodity is not particularly helpful, especially since the usual stories about pricing and profits do not really apply.

One kind of institution that already tries to deal with data privacy is the Institutional Review Board or IRB. But as this section first describes, they were set up to handle research ethics generally. The need for new and more specific institutional approaches as well follows, including a proposal to support a network of Administrative Data Research Facilities.

Institutional Review Boards

The ethical framework applying to human subject research in the biomedical and behavioral research fields dates back to the Belmont Report.[68] Drafted in 1976 and adopted by the United States government in 1991 as the Common Rule,[69] the Belmont principles were geared towards a paradigmatic controlled scientific experiment with a limited population of human subjects interacting directly with researchers and manifesting their informed consent. These days, researchers in academic institutions as well as private sector businesses not subject to the Common Rule, conduct analysis of a wide array of data sources, from massive commercial or government databases to individual tweets or Facebook postings publicly available online, with

[67] Williamson, O.E., 1996. *The mechanisms of governance*. Oxford University Press.

[68] *NATIONAL COMM'N FOR THE PROT. OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RESEARCH, BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH (1979), available at http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.*

[69] HHS, FEDERAL POLICY FOR THE PROTECTION OF HUMAN SUBJECTS ('COMMON RULE'), http://www.hhs.gov/ohrp/humansubjects/commonrule/.

little or no opportunity to directly engage human subjects to obtain their consent or even inform them of research activities. The challenge of fitting the round peg of data-focused research into the square hole of existing ethical and legal frameworks will determine whether society can reap the tremendous opportunities hidden in the data exhaust of governments and cities, health care institutions and schools, social networks and search engines, while at the same time protecting privacy, fairness, equality and the integrity of the scientific process. One commentator called this "the biggest civil rights issue of our time."[70]

These difficulties afflict the application of the Belmont Principles to even the academic research that is directly governed by the Common Rule. In many cases, the scoping definitions of the Common Rule are strained by new data-focused research paradigms. For starters, it is not clear whether research of large datasets collected from public or semi-public sources even constitutes human subject research. "Human subject" is defined in the Common Rule as "a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information."[71] Yet, data driven research often leaves little or no footprint on individual subjects ("intervention or interaction"), such as in the case of automated testing for security flaws.[72] As Michael Zimmer notes in his paper, "the perception of a human subject becomes diluted through increased technological mediation."[73] Arvind Narayanan and Bendet Zevenbergen explain that "the Internet is more properly understood as a sociotechnical system in which humans and technology interact."[74]

Not only the definitional contours of the Common Rule but also the Belmont principles themselves have required reexamination. The first principle, respect for persons, is focused on individual autonomy and its derivative application, informed consent. While obtaining individuals' informed consent may be feasible in a controlled research setting involving a well-defined group of individuals, such as a clinical trial, it is untenable for researchers experimenting on a database that contains the footprints of millions, or indeed billions, of data subjects. The second principle, beneficence, requires a delicate balance of risks and benefits to not only respect individuals' decisions and protect them from harm but also to secure their wellbeing. Difficult to deploy even in traditional research settings, such cost-benefit analysis is daunting in a data research environment where benefits could be probabilistic and incremental and the definition of harm subject to constant wrangling between minimalists who reduce

[70] Alistair Croll, Big data is our generation's civil rights issue, and we don't know it, O'Reilly Radar, Aug. 2, 2012, http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-wedont-know-it.html.

[71] 45 CFR 46.102(f).

[72] See, e.g., Arvind Narayanan & Bendert Zevenbergen*, No Encore for Encore? Ethical Questions for Web-Based Censorship Measurement.*

[73] Michael Zimmer, Research Ethics in the Big Data Era: Addressing Conceptual Gaps for Researchers and IRBs.

[74] Narayanan & Zevenbergen, supra note 5.

privacy to pecuniary terms and maximalists who view any collection of data as a dignitary infringement.[75]

In response to these developments, the Department of Homeland Security commissioned a series of workshops in 2011-2012, leading to the publication of the Menlo Report on Ethical Principles Guiding Information and Communication Technology Research.[76] That report remains anchored in the Belmont Principles, which it interprets to adapt them to the domain of computer science and network engineering, in addition to introducing a fourth principle, respect for law and public interest, to reflect the "expansive and evolving yet often varied and discordant, legal controls relevant for communication privacy and information assurance."[77] In addition, on September 8, 2015, the U.S. Department of Health and Human Services and 15 other federal agencies sought public comments to proposed revisions to the Common Rule.[78] The revisions, which address various changes in the ecosystem, include simplification of informed consent notices and exclusion of online surveys and research of publicly available information as long as individual human subjects cannot be identified or harmed.[79]

For federally funded human subject research, the responsibility for evaluating whether a research project comports with the ethical framework lies with Institutional Review Boards (IRBs). Yet, one of the defining features of the data economy is that research is increasingly taking place outside of universities and traditional academic settings. With information becoming the raw material for production of products and services, more organizations are exposed to and closely examining vast amounts of often personal data about citizens, consumers, patients and employees. This includes not only companies in industries ranging from technology and education to financial services and healthcare, but also non-profit entities, which seek to advance societal causes, and even political campaigns.[80] Whether the proposed revisions to the Common Rule address some of the new concerns or exacerbate them is hotly debated. But whatever the final scope of the rule, it seems clear that while raising challenging ethical questions, a broad swath of academic research will remain neither covered by the rules

[75] *Case C-362/14, Maximillian Schrems v. Data Protection Commissioner, 6 October 2015, http://curia.europa.eu/juris/document/document.jsf?docid=169195&doclang=EN; also see Ryan Calo, The Boundaries of Privacy Harm, 86 IND. L.J. 1131 (2011).*

[76] DAVID DITTRICH & ERIN KENNEALLY, THE MENLO REPORT: ETHICAL PRINCIPLES GUIDING INFORMATION AND COMMUNICATION TECHNOLOGY RESEARCH, U.S. Dept. of Homeland Sec., (Aug. 2012), available at https://www.predict.org/%5CPortals%5C0%5CDocuments%5CMenlo-Report.pdf.

[77] Ibid, at 5.

[78] *HHS, NPRM for Revisions to the Common Rule, Sept. 8 , 2015, http://www.hhs.gov/ohrp/humansubjects/regulations/nprmhome.html.*

[79] Also see Association of Internet Researchers, Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee (Version 2.0), 2012, http://aoir.org/reports/ethics2.pdf (original version from 2002: http://aoir.org/reports/ethics.pdf).

[80] Ira S. Rubinstein*, Voter Privacy in the Age of Big Data,* 2014 WISC. L. REV. 861*.*

nor subject to IRB review. Katie Shilton shows that academic researchers today have inconsistent views about how to handle these issues.[81] Currently, gatekeepers for ethical decisions range from private IRBs to journal publication standards, association guidelines and peer review. A key question for further debate is whether there is a need for new principles as well as new structures for review of academic research that is not covered by the current or expanded version of the Common Rule.[82]

In Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings, Tene and Polonetsky noted that even research initiatives that are not governed by the existing ethical framework should be subject to clear principles and guidelines. Whether or not a research project is federally funded seems an arbitrary trigger for ethical review. Urs Gasser et al note "[the] larger trend of big data research conducted outside of traditional oversight mechanisms due to the limited scope of research subject to existing regulations."[83] To be sure, privacy and data protection laws provide an underlying framework governing commercial uses of data with boundaries like consent and avoidance of harms. But in many cases where informed consent is not feasible and where data uses create both benefits and risks, legal boundaries are more ambiguous and rest on vague concepts such as "unfairness"[84]or the "legitimate interests of the controller."[85] This uncertain regulatory terrain could jeopardize the value of important research that could be perceived as ethically tainted or become hidden from the public domain to prevent scrutiny.[86] Concerns over data ethics could diminish collaboration between researchers and private sector entities, restrict funding opportunities, and lock research projects in corporate coffers contributing to the development of new products

[81] Katie Shilton*, Emerging Ethics Norms in Social Media Research.*

[82] National Research Council. (2014). *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences.* Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences. Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[83] Urs Gasser, Alexandra Wood, David R. O'Brien, Effy Vayena, and Micah Altman, *Towards a New Ethical and Regulatory Framework for Big Data Research.*

[84] *FTC Policy Statement on Unfairness, Appended to International Harvester Co*., 104 F.T.C. 949, 1070 (1984). See 15 U.S.C. § 45(n).

[85] *Article 29 Working Party, WP 217, Op. 06/2014 on the Notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, Apr. 9, 2014, http://ec.europa.eu/justice/dataprotection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf*.

[86] The Common Rule's definition of "research" is "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to *generalizable* knowledge." (Emphasis added).

without furthering generalizable knowledge.[87]

In a piece he wrote for a Stanford Law Review Online symposium,[88] Ryan Calo foresaw the establishment of "Consumer Subject Review Boards" to address ethical questions about corporate data research.[89] Calo suggested that organizations should "take a page from biomedical and behavioral science" and create small committees with diverse expertise that could operate according to predetermined principles for ethical use of data. The idea resonated in the White House legislative initiative, the Consumer Privacy Bill of Rights Act of 2015, which requires the establishment of a Privacy Review Board to vet non-contextual data uses.[90] In Europe, the European Data Protection Supervisor has recently announced the creation of an Advisory Group to explore the relationships between human rights, technology, markets and business models from an ethical perspective, with particular attention to the implications for the rights to privacy and data protection in the digital environment.[91]

Administrative Data Research Facilities

Because IRB's are mainly designed to examine research ethics before a project starts, they can recommend the whole range of pre-collection and post-collection protocols for data privacy. As noted above, though, more and more information is compiled by agencies, offices, or companies for purposes other than research. As above, we call such records "administrative data," though others also refer to transactional, observational, or "found" data. By-products of enterprise activity stand in contrast to the results of surveys, lab experiments, or field trials that are designed by academics and approved by their IRB's to test specific hypotheses.

If handled properly, administrative datasets can nevertheless be hugely valuable for research. The challenges were already evident in 1984 when Zvi Griliches wrote a handbook chapter that

[87] Jules Polonetsky, Omer Tene, & Joseph Jerome, Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings, 13 COLO. TECH. L. J. 333 (2015).

[88] Stan. L. Rev. Online Symposium Issue*, Privacy and Big Data: Making Ends Meet, September, 2013, http://www.stanfordlawreview.org/online/privacy-and-big-data; also see stage setting piece*, Jules Polonetsky & Omer Tene, Privacy and Big Data: Making Ends Meet, September 3, 2013 66 STAN. L. REV. ONLINE 25.

[89] Ryan Calo, Consumer Subject Review Boards: A Thought Experiment, 66 STAN. L. REV. ONLINE 97 (2013), available at http://www.stanfordlawreview.org/online/privacy-and-big-data/consumersubject-review-boards.

[90] *CONSUMER PRIVACY BILL OF RIGHTS §103(c) (Administration Discussion Draft 2015), available at https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf.*

[91] European Data Protection Supervisor*,* Ethics Advisory Group, Dec. 3, 2015, *https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/Ethics.* See also
Curtis Naser, The IRB Sledge-Hammer, Freedom and Big-Data.

presents "a review of the ambivalent relationship between data and econometricians, due largely to the second-hand nature of economic data and the consequences that flow from the distance between econometricians as users of data and its producers." [92]

Using data for secondary purposes, such as research or program evaluation, can entail high transaction costs. When researchers deal with administrative data, current practice fails in at least seven widely acknowledged ways:

A. *Obtaining* administrative data often relies on a personal contact who is less risk averse than most lawyers. Arrangements are typically ad hoc, ad hominem, time-consuming, expensive, and unstable. And data holders have limited patience for negotiating such deals one researcher at a time.

B. *Preparing* administrative data can be difficult and arbitrary but nevertheless consequential. In many cases, after the "cleaning" process is complete, there are few if any clues left about how data sifting protocols were carried out, much less what effect they may have on research findings.

C. *Protecting* administrative data against breaches of privacy or proprietary restrictions is sometimes haphazard, without rigorous standards for ensuring security, confidentiality, reuse, etc.

D. *Supplying* administrative data to other researchers is often needlessly or carelessly constrained in ways that hamper the checking, reliability, and broad use of findings.

E. *Sustaining* administrative data access and utility requires careful attention to metadata, archival standards, curation, and costs that researchers rarely take into account.

F. *Studying* administrative data poses special challenges when distinguishing causation from correlation, for example. Transaction records may not constitute a representative sample of an appropriate population. And information about control groups or exogenous variation may be non-existent or missing.

G. *Linking* administrative datasets can create exciting opportunities for answering research questions, but the process often gets bogged down by technical, legal, or administrative impediments.

Given a collective action problem with high transaction costs like this, economics teaches us to look for institutional solutions—especially ones that can engender trust. So imagine a network

[92] Zvi Griliches, "Data Problems in Econometrics" (NBER Technical Working Paper No. 39, July 1984), accessed at http:// http://www.nber.org/papers/t0039/.

of trustworthy data brokers called Administrative Data Research Facilities (ADRF's). Each has expertise solving problems like those cited above for data from a given sector of the economy. Today, for instance, researchers interested in supermarket scanner data would do well to consult the Kilts Center at Chicago Booth. Those interested in university data can turn to the Institute for Research on Innovation and Science (IRIS) at the University of Michigan. Or to study cities, there are experts, datasets, and systems for using them at the Center for Urban Science Progress (CUSP) at NYU.

Support for more data intermediaries like these would help, of course. Ideally there would be at least one covering each major data-producing sector the economy. Those sectors might include, for example: online retailers; traditional retail chains; credit card companies; financial services; automobile companies; payroll processors; not to mention all kinds of state, local, and federal agencies. An ADRF with a given specialization would negotiate a Data Use Agreement once with each member of such a sector on behalf of researchers generally. The ADRF would then be responsible for preparing each dataset for study by supplying documentation, metadata, basic cleaning and versioning, approximate summary statistics, citation information, archiving services, etc. For credentialed researchers with sound research plans, access to the data would be granted using admissible privacy protecting protocols as appropriate. The data may also be useful to federal agencies that compile official statistics, too.

While an ADRF like this could go a long way to lowering transaction costs for data suppliers and users, setting up lots of them is only Phase I of a more comprehensive plan. Phase II calls for organizing an Administrative Data Research Network (ADRN) whose members would be ADRF's committed to sharing best practices and high standards. This association of intermediaries would, for example, have working groups on topics like: Compliance and Legal Matters; Researcher Credentialing; ADRF Accreditation; Data Security; Systems and Operations; Private and Proprietary Data Protections; Government and Public Relations; Research and Reproducibility Standards; Corporate Relations and Contracting; Data Interfaces and Linking; etc. Any researcher or facility found to be jeopardizing the ADRN's reputation for trustworthiness would lose the status, privileges, and data access afforded to members in good standing.

While it is natural to think of an ADRF as hosted at a university, others could play that role, too, including NGO's, corporations, or government agencies—especially ones that, like the Census Bureau, are not only actively engaged in generating data but also in facilitating its use by independent researchers. Having received funding to establish a clearinghouse for government and other administrative data, for example, the Census Bureau can be an important collaborator. This could be especially significant since all network members would abide by explicit and streamlined procedures for sharing, linking, and protecting datasets.

The U.K. already has an organization called the Administrative Data Research Network (ADRN). Funded by the Economic and Social Research Council and run by the Administrative Data Service (see www.adrn.ac.uk). It concentrates on government datasets. Two of the sector specialists within the U.K. initiative are the Consumer Data Research Centre based at the

University College London and the Urban Big Data Centre based at the University of Glasgow. Laws and traditions vary considerably between one country and another, of course, but the U.K. experience provides inspiration to create similar institutions in the U.S. Existing U.K., German, Danish, or other international entities are also potential partners.

VII. Planning a System

To promote policymaking based on evidence derived from private or public data, at least four kinds of players need to cooperate: government agencies, private sector corporations, independent researchers, and philanthropic funders. Academics like Gary King of Harvard have had conversations about the need for a “grand bargain” among these groups concerning data. Lowering the transactions costs associated with studying data would be good for all. Led by the Census Bureau, government agencies at all levels are beginning to work better with academics. Elements of the private sector are beginning to tire of dealing with repeated data requests from one academic after another or one government agency after another. In some cases, they are also realizing that outside researchers can add enormous value by cleaning, compiling, archiving, and analyzing corporate data in ways they would or could not do themselves.

Researchers, in contrast, have been slow to organize facilities or networks to deal with access to administrative data. Those who, for one reason or another, have such access already are not eager to share and do not see the problem. Journal editors make noise about reproducibility requirements but offer waivers freely. As always, academics tend to focus only on completing their own next paper. This takes precedence over cooperating with colleagues, much less other parties, to make a more sensible system for granting data access or for generating reliable evidence.

Philanthropies can, at their best, provide excitement and incitement to solve collective action problems like this. Some already are. About $12 million annually for five years could substantially improve the conduct of empirical research and, in turn, the process of policymaking as well. There are, for example, Administrative Data Research facilities already in operation. A few are even beginning to bring in revenue other than foundation support. Besides launching more, next steps include setting up the board, offices, working groups, and membership criteria for an Administrative Data Research Network.

In the end, there is no one simple solution to the problem of research access to private data. But there are several technical and institutional solutions that could be quite serviceable, along with quite a few proposals or practices that are not. Enough planning, organization, and support could bring about the cooperation and standards development needed to reduce transactions costs that hinder progress. Achieving the potential for data science to improve policymaking while also protecting privacy is well within reach.