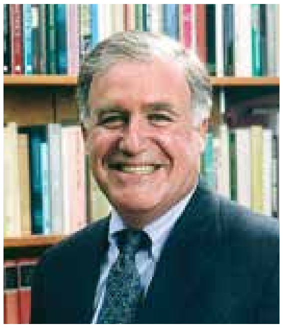
President's Letter

Dr. Paul L. Joskow



September 14, 2015

Research Integrity and Reproducibility

Science Is Not Broken, But It Can Be Better[1]

The October 19, 2013 issue of *The Economist* contains two articles that report and discuss the large fraction of important scientific papers in medicine, computer science, and other fields where scientists have been unable to reproduce the original authors' results ("How science goes wrong" 2013, "Trouble at the lab" 2013). A study by Stanford researcher Daniele Fanelli found that published papers in empirical economics report results that confirm the tested hypothesis at a rate five times higher than published papers in space science (Fanelli 2010).[2] A recent paper by Robert M. Kaplan and Veronica L. Irvin (Kaplan and Irvin 2015) examined 55 large randomized controlled trials studying the effects (positive, negative, or null) of drugs or dietary supplements used to treat cardiovascular disease. They found that studies that did not pre-register the hypotheses under investigation were significantly less likely to report null results than the studies that did pre-register. Science watchdog sites like Retraction Watch[3] have sprung up to catalog retractions, withdrawals, and significant post-publication amendments of scientific papers. Cases of scientific retraction now regularly turn into major media events. These findings and others like them have led some to conclude that something is rotten at the core of twenty-first century research. "Science," they say, "is broken."

In June 2015, a group of prominent researchers took to the pages of *Science* magazine to rebut this charge. (Alberts, et al. 2015). Recent cases of high profile retractions, they argue, far from being evidence that science is broken, are in fact evidence that science is *working*. When scientific papers are discovered to be unreproducible or fraudulent, after all, the people who have done the discovering are invariably other scientists. Paper withdrawals and retractions are

[1] Thanks are due to my colleagues Daniel Goroff, Josh Greenberg, and Nate Williams for very useful comments on an earlier draft of this letter.

[2] This "confirmation bias"—the tendency to disproportionately publish only those studies that confirm the hypothesis under investigation—appears to be growing in many other scientific fields as well.

[3] http://www.retractionwatch.com

instances not of the failure of the scientific process but of its proper functioning. They are the key mechanisms through which the scientific community polices itself.

I agree that the recent flurry of studies that raise questions about the credibility of results reported in important scientific papers should not lead to the conclusion that "science is broken." However, this is not to say that there is no room for improvement.[4] Confirmation bias in academic journals is genuinely worrying. The publication of false results, however infrequent, is worrying. The publication of research that cannot be replicated is worrying. The integrity of scholarly research is essential if it is to achieve its full potential. Research must be perceived to be highly reliable by those who use it, by funding agencies, and by the public. "Reproducibility is essential for validating empirical research." (Ioannidis and Doucouliagos 2013). The credibility of science is seriously undermined when other scientists find that they cannot reproduce research results that appear in the literature. Why are there seemingly so many published research papers whose empirical results cannot be readily reproduced by other scientists? Why do there seem to be so many high profile retractions of published research papers? How can we change scientific practice, institutions, incentives, or norms in ways that lead to more reproducible, more reliable research?

The Alfred P. Sloan Foundation has a keen interest in these questions. A large fraction of the Sloan Foundation's grants support basic and applied research in science, technology, and economics. Because the Foundation does not itself publish any papers or reports drawn from the research we support, we expect our grantees to disseminate the results of their research in working papers, journal articles, books, and a variety of online media. The integrity and reliability of the academic literature is thus a topic of immense importance to us. Since 2008, the Foundation has committed nearly $30 million to projects designed to help scientists conduct their work in ways conducive to greater reproducibility and transparency. In what follows I will discuss how the structure of scientific careers and the incentives facing scientists and journals create barriers to conducting fully reproducible research. I will then discuss the opportunities grantmakers have to reduce or remove those barriers and what the Sloan Foundation is doing to make the published scientific literature more reproducible, reliable, and credible.

What Is Reproducibility?

The concerns I wish to address fall under a wide array of rubrics. Some authors talk about *reproducibility*. Others talk about *replicability*. Still others talk about the *integrity* of research or its *reliability*, its *trustworthiness* or *dependability* or *robustness*. Still others talk about the need for research *transparency*. More recent entrants to the discussion have introduced new phrases influenced by Internet culture. They call for *open science*, or *open data*, or *open code* (and usually all three). Sometimes we talk about the reproducibility of *experiments*. Other times we talk about the reproducibility of *findings*.

The National Science Foundation's Subcommittee on Replicability in Science has recently produced a report (Social, Behavioral and Economic Sciences Perspectives on Robust and Reliable Science 2015) that provides a helpful framework for discussing these issues, a framework I shall adopt in what follows. The report draws a distinction between three ways in which science may aspire to be robust, which it calls reproducibility, replicability, and generalizability. An experimental[5] finding is *reproducible* according to this framework if a researcher is able to duplicate the results of a prior study using the same methods, procedures, code, and data as the original author of the research. An experimental finding is *replicable* if a researcher is able to duplicate the results of a prior study by applying the same procedures and methods of the original experiment to newly collected data. Finally, an experimental finding is *generalizable* if a researcher is able to duplicate the results of a prior

[4] Alberts et al.(2015) concede as much and make many useful suggestions for how science might be improved in ways conducive to better reliability of published findings.

[5] I use the term "experiment" broadly in this letter to include laboratory experiments in controlled environments; randomized controlled trials (RCTs) applied to individuals, groups, or organizations; and hypothesis testing that relies on data drawn from non-randomized controlled trials, but where statistical modelling and other analytical methods are used to appropriately control for causal variables other than the variable of interest. Including the latter case as a type of "experiment" deviates somewhat from common usage in the physical sciences, but recent work in economics and econometrics has, in my opinion, made a persuasive case that in some situations a methodologically rigorous analysis of an appropriately large but non-random sample can yield insights comparable in breadth and power to those produced by controlled environments and RCTs. See Angrist (2009) and Angrist (2014).

study using an entirely different experimental design and associated data. If a given experimental finding is reproducible, replicable, and generalizable according to these definitions, we will call the finding *robust*. Ideally, we want scientists to produce robust research. Too often, they do not. The question I turn to now is why.

Failures of Reproducibility, Replicability, and Generalizability

We should begin by cutting researchers some slack, if only a little. It is unreasonable to expect all empirical research to be robust. The conclusions of empirical science—that a particular treatment or causal variable is "significant," whether its sign is positive or negative, the magnitude of the discovered effect, etc.— are probabilistic statements. For even the most rigorous, well-designed experiments, a replication study using the same experimental design, the same computational methods, but new data will yield exactly the same results only with some probability less than 100 percent. Accordingly, specific findings cannot be expected to be duplicated 100 percent of the time. Yet failures to reproduce experimental findings occur more frequently than would be predicted merely by the probabilistic nature of scientific findings. We need to understand why reproducibility failures arise more frequently than would be implied by the uncertainties inherent in empirical research.

In general, experimental findings fail to be reproducible (according to the definition above) for one of two reasons:

Inadequate description of experimental design, experimental methods, and computational methods. To perform a reproducibility study a researcher must have access to relevant information about the original experiment's design, how it was implemented, and the computational and statistical methods used to analyze the data and draw conclusions. Absent this information, the reproducibility study may actually fail to be a *reproduction* at all, instead running a different experiment altogether or using different experimental or computational methods. Moreover, without certain methodological information it becomes impossible to interpret whether a failure to duplicate the original study's result is due to flaws in the original experimental design, its implementation, or the computational methods used. In theory, this information is included in the traditional methodology section of a published paper. In practice, however, methodology sections of papers are increasingly unhelpful to the would-be-reproducer. Space constraints in scholarly journals rarely allow for adequate methodological detail for a subsequent researcher to accurately reproduce the original experiment. Yet even if more space were allotted to such discussions inside papers themselves, the complexity of modern research makes it difficult and time consuming to describe one's methodology in a way that allows research results to be reproduced. Methodological information needed to reproduce a typical study involving the analysis of survey data (say) would include the sampling procedure, the survey instrument, the procedure used to field the survey, the type and version of the software used to collect responses, the methods used to code respondents' answers, the techniques used to "scrub" respondent data, the information collected on respondents, the model used to analyze the data and the assumptions powering that model, the software platform and platforms (and their versions) used to perform analyses, the particular algorithms used, etc. Much of this information often does not appear in the paper itself but must be obtained in other ways. Yet, there is no standard practice governing how to store this information and make it available. Typically, one must go to the authors and request this information, with notoriously inconsistent responses to these requests.

Unavailability of data. Even the most complete information about methods is of little use in reproduction unless one has the data used in the original experiment. But like detailed information on methods, the data analyzed in a scholarly paper are not routinely made available to researchers. In some cases, the analyzed data *cannot* be shared, legally, because they belong to some private or corporate entity or because sharing the data would violate the privacy of the individuals whose information was used in the original research. Indeed, one of the paradoxes of the current era of science is that it increasingly utilizes datasets with information about individual attributes and behavior of a size, complexity, and usefulness dwarfing those available to researchers in the past. Yet the use of these datasets is often restricted to protect individual privacy or because the data are proprietary.[6]

These two factors—inadequate description of methods and the unavailability of data—are further

[6] For more on the challenges posed to science by the Big Data revolution and the corresponding opportunities for grantmakers, see my letter to the 2013 Alfred P. Sloan Foundation Annual Report.

complicated by the fact that reproducibility studies are often conducted years after the original experiment. The significance of any given finding is often not fully appreciated immediately upon publication and subsequent work either confirming or contradicting the finding may spur interest in reproducing the original study. In the intervening years, however, the original research team has often moved on to other pursuits. Researchers will have moved institutions or careers. Lab personnel will have changed. Memories will have faded. Data may be trapped in old, outdated formats or systems. To be useful, data, metadata, and methods must be stored in *durable, permanent* archives. Where those archives are to be hosted, how to make them easily discoverable, how to ensure their permanence, and how to pay for them are questions that have yet to be answered.

The two obstacles described above are barriers to reproducibility, to duplicating an experiment exactly as it was originally performed. There can also be failures of replicability and generalizability, failures to duplicate the *findings* of the original experiment either by running the original experiment on new data or in duplicating old findings through new experimental methods. Researchers are sometimes unable to duplicate the findings of previous research because the original finding is incorrect. The research reports a significant causal relationship between variables that is not, in fact, there. How do such results find their way into the published, peer-reviewed, academic literature? Any adequate analysis of the failures of replicability and generalizability in research will have to give some account of the sources of scientific error.

Data errors and data ”cleaning.” In a typical modern experimental study, data must be collected; transferred; aggregated; “cleaned” for obvious errors, missing entries, and outliers; converted into one or more formats suitable for analysis; and, finally, analyzed. Errors may creep in at any step in this process. The original data may have errors; there may be errors in transcription between the original source and the datasets used for analyses; and efforts to “clean” the data obtained from the original source may throw away relevant data, adjust data in questionable ways, or mischaracterize the data in some fashion. In addition, a replication study may make similar errors or fail to reproduce the original results due to errors of its own.

Modelling and Statistical Deficiencies: Most empirical research involves questions about causality (Angrist 2014) and the sign and magnitudes of the causal impacts of one “independent” variable on another “dependent” variable. Do variations in some variable x affect variations in another variable y? What is the sign of the impact of variations in x on variations in y? How large is the impact of variations in x on variations in y? In an uncontrolled environment or a poorly designed controlled environment, however, there may be one or more other variables z that also affect variations in y or x or both . The failure to take the relationships between the variable z and the variables x and y into account can lead to incorrect conclusions about causality and biased estimates of key parameter values. The simplest case is one in which there is no causal relationship between x and y at all but a third variable z has a causal impact on both x and y. That is, z is correlated with x and y and, as a result, x and y happen to be correlated as well. A popular real life example of this scenario is a series of Dutch statistics showing a positive correlation between the number of storks nesting in springs around the city and the number of human babies born at that time. Of course, there was no causal connection, but a third variable, the weather nine months before the observations, caused both variables to comove. Absent a sound theory of why x and y should be causally related, we run the danger of drawing conclusions about causality when we are just measuring spurious correlations. Another very common situation is one in which there actually is a causal relationship between x and y, but there is a third variable z that is causally correlated with both x and y. In this case, using ordinary least squares regression methods to estimate the relationship between x and y while leaving the variable z out of the regression will lead to a biased estimate of the coefficient of x. This occurs because, by leaving the variable z out of the regression, the estimate of the coefficient of x now captures both the “direct effect” of x on y plus the “indirect effect” of z on y through x, making the identification of the isolated causal effect of x on y, our primary objective, impossible. In such cases, the production of reliable findings requires resorting to more sophisticated solutions, for example, a more complete modelling of the full system of relationships between causal variables that demonstrates that the parameter of interest can be identified, or the application of appropriate statistical methods that take account of the endogeneity of one or more independent variables. Other common statistical issues that often arise include sample selection bias, measurement error, and inadequate sample sizes required to correctly reject (in a binary hypothesis test) the null hypothesis when the alternative hypothesis is true with a sufficiently high probability (the power of the statistical test).

Data analysis practices can also lead to results that are not replicable or generalizable. Repeated empirical analysis of the same dataset using variations on the causal model and then reporting only the results that the researcher "likes" (overfitting or data mining) makes standard statistical tests meaningless, especially in relatively small samples where out-of-sample verification of the results is not feasible. Other bad practices include the failure to identify and exclude outliers (e.g. one observation drives the results) or, at the other extreme, the unmotivated exclusion of some observations whose inclusion would lead to results that are inconsistent with the hypothesis being tested.

Fraud: People will disagree about how exactly to define fraud in research. Whatever its proper definition, there will be borderline cases, those where reasonable people will disagree about whether the experimental design, implementation, or analysis of some study are so obviously flawed as to constitute willful scientific malpractice. Perhaps it is one of those cases where "you know it when you see it." But not every case is near the border. If data are fabricated or consciously doctored to yield particular conclusions, if the experiments described in a paper have not actually been conducted, or if the computational and statistical analyses described were never performed, we have a clear case of research fraud. Research that is fraudulent will fail replication tests[7], though, of course, research that cannot be replicated does not imply fraud and may be due to one or more of the factors I have already described. While cases of real or alleged research fraud garner much media attention and can, therefore, seem to be common, in my more than forty years of experience doing empirical research in economics, I must conclude that such cases of obvious fraud are rare. Reading through the reports on Retraction Watch, however, it is clear that fraudulent research does get published. How often such research is eventually identified, corrected, or retracted is unknown, but it would be surprising if all cases of fraud have been or will be identified. While the number of paper retractions appear to be growing, it is not at all clear whether this is due to an increasing incidence of flawed or fraudulent research or merely an artifact of the increased attention being paid to these issues. In any case, adoption of standards that foster reproduction and replication will deter the publication of fraudulent research since it increases the likelihood of detection.

Incentives[8]

Research findings fail to be reproducible or replicable either because the original methods and data are unavailable or because flaws in the experimental design, implementation, or analysis led a researcher to reach a conclusion that was not supported by the evidence (including cases of fraudulent evidence). Minimizing the publication of such research requires not just understanding how it happens, but also understanding the underlying incentives faced by researchers and academic publishers which facilitate the barriers to reproducibility that I have discussed here.

Researcher Incentives. The academic incentive environment is partially responsible for the barriers to reproducibility and replicability. There are few incentives for researchers to perform replication studies.[9] Funding for replication is hard to find—funders tend to want to fund original, not duplicative, research—replication studies tend to be difficult to publish in the most highly respected journals and garner lower citation counts[10] than

[7] Unless, that is, the fraudster got *very* lucky and happened to have manufactured evidence for a replicable hypothesis.

[8] Come on, I'm an economist. You think there wasn't going to be a section on incentives?

[9] As a professor, I often gave students replication assignments to help them to learn how to do research. And they often had difficulty replicating published research, often due to their own mistakes rather than mistakes in the original research. (I frequently asked students to try to replicate the results of a very famous study published in a respected economics journal where the methods and data sources were especially clear. They could rarely replicate the econometric results exactly, but also never found a significant change in the most important result in the paper. Years later an experienced economist, armed with the actual data used in the original study, also tried to replicate it and found a very significant mistake in the data (and it was clearly simply a careless mistake). Correcting the mistake changed the magnitude of the coefficient of interest quite significantly and in turn changed the interpretation of the paper's most important result.

[10] In my experience, the primary factor that affects tenure and promotion decisions at top research universities is the faculty member's research. And, in my experience, the evaluation of the quality and impact of the scholar's research is evaluated by colleagues and external experts who read the candidate's papers, and not by simply counting the *number* of publications

papers reporting original research. As one scholar said to me "nobody ever got tenure based on the replication studies she performed."

The pressure to publish and the ticking of the tenure clock make time a very scarce resource. Labor-intensive activities are very costly in such an environment. Carefully documenting one's experimental design and how it was implemented and then assembling, annotating, and archiving the data, code, and statistical methods used in an experiment takes a lot of time[11], time that is poorly rewarded, if at all. Little credit is given to those who create well-documented and easily-accessible datasets or who follow best practice by carefully documenting every component of the research that underlies their published work.

Publisher Incentives. In theory, the editors and referees at scholarly journals are the "guardians at the gate" who safeguard research quality by deciding whether a paper should be published and identifying changes that are necessary to make a paper publishable. In practice, however, there is only so much that can be expected from editors and referees. Academic journal editors typically take those positions as a service to their profession with little in the way of financial or scholarly reward and with continuing academic responsibilities. At high quality journals where the submission rate is high and the acceptance rate low, the task of triaging papers, selecting referees, badgering referees to get their reports in, producing "revise and resubmit letters," and ultimately deciding what gets published is very burdensome. Like editors, referees typically accept assignments as a service to their profession. Honoraria, when they exist at all, are modest, and referees are expected to continue to meet their teaching, research, and professional service obligations. Because they do not receive detailed methodological information or data, referees rarely have the resources to turn the refereeing process into a replication process. Nor would they have the time to do so even if proper documentation were available. The best they can do is determine whether the research question makes sense, whether the experiment appears to be designed in a way that yields defensible results, and whether the data look sensible based on their own experience.

Academic journals are subject to many of the same pressures that plague individual researchers. The pressure to be a "high impact journal," one that publishes papers that go on to be influential and highly cited, is immense. Maximizing the chances that one's published pieces will be influential means maximizing, as far as is economically feasible, the number of articles published. Because print journals are space constrained—there's an upper bound on how many pages a print journal can reasonably contain—there's pressure to shorten average paper length, publishing more papers in the same number of pages[12]. This leads to pressures to make published papers shorter and shorter and to include fewer and fewer details about the experiments, data, and code in the paper itself. This "extra" (though in fact, crucial) information is then included in separate documents "available from the authors" or posted on their web pages. This increasingly common practice, however, is a barrier to reproducibility and replicability. Authors may not post information on their data and methods, data that does get posted is not standardized or controlled for quality, journals do not verify that information has been made available, web pages where material is posted often change, and data can be removed at any time by the author without consequence. While the journals could remedy these problems by creating standards and hosting an archive of this "residual" material themselves, this would require valuable time and resources, with little benefit accruing to the journal itself. The bottom line is that reproducibility and replicability are public goods that yield few if any private returns to the journals, the editors, or the scholars that produce them. (See, for example, Nosek, et al. 2015)

or citations. However, some departments and universities apparently do count publications and citations, either because they lack the internal expertise in the scholar's research area and/or because they cannot get external reviewers to accept the tedious task of reading a scholar's papers and writing a letter about their quality and impact. The threat that evaluation letters may become public, further diminishes incentives to provide frank appraisals. Unfortunately, some universities have bureaucratic procedures for promotion and tenure that do include crude counts of publications and citations. In these situations, powerful incentives are created to get as many publications out as quickly as possible, with predictably deleterious consequences for the quality and reproducibility of this research.

[11] In the short term, at least. Some argue that keeping a tidy scientific house, like keeping a tidy regular house, is a time saver in the long run.

[12] Though digital journals are not similarly space-constrained, print norms have tended to carry over into the online publishing environment.

Journals, it is true, could require researchers to "pre-register" their hypotheses, the experimental design and methods, the expected outcomes, the models to be used and relationships to be tested, etc. This would counteract many of the practices that make replication difficult. But journals are poorly positioned to make such demands. To be effective, pre-registration must occur *before* an experiment begins, which is also before an author knows which journal she will submit her findings to or even whether she will seek to publish them at all. Since journals have the most leverage over researchers at the time of publication, pre-registration occurs too early in the research cycle for individual journal policies to hold much sway.

Journals are understandingly hesitant to place burdensome transparency requirements on their authors for fear that talented scientists will simply publish elsewhere. Why submit yourself to pain-in-the-neck transparency requirements from Journal X when Journal Y will publish your paper without the hassle? Without altering the incentives to both journals and researchers alike, we find ourselves in a bad equilibrium. The current system is non-ideal, but no individual actor in the system has reason to make things better.

Opportunities for Grantmakers

Grantmakers and other funders of research can lead the way to a better equilibrium. Here I will discuss what the Sloan Foundation is doing, recognizing that some other private foundations and government funders of research are also making efforts to respond to the issues that I have discussed. The Foundation's Digital Information Technology Program, led by program director Josh Greenberg, and the Economic Institutions, Behavior & Performance program, led by Vice President Daniel Goroff, have invested significant resources in changing the ways we conduct and publish research. What follows are some areas in which funders might "move the needle" in ways that change incentives and help move academic publishing in a direction more conducive to reproducibility and replicability.

1. Insist on high methodological standards

Funders have the most influence before a project gets funded, when researchers, eager to acquire funding, are willing to make commitments they might not otherwise make. Funders can increase transparency and replicability by insisting on the highest standards before agreeing to fund research. The Alfred P. Sloan Foundation, expanding on federal data management plan requirements, has adopted grant proposal guidelines[13] that require potential grantees to specify clearly and completely, in a separate Information Products Appendix, the anticipated products of their research, including working papers, publications, data, and code, and to specify whether and how those products will be made available to other researchers and to the public. We strongly encourage grantees to adhere to the principle that "making research products freely and openly accessible can increase the reach and value of what we fund." The requirement that prospective grantees produce an Information Products Appendix as part of the grant application process begins an often-fruitful dialogue, making it possible for Foundation staff and external reviewers to work with potential grantees to increase the accessibility of their work product. Funders who wish to include similar requirements in their application process should proceed with caution however. First, as I've already noted, properly making the results of one's work accessible is not costless; funders must be willing to increase grant budgets accordingly. Second, requirements should be sensitive to the fact that norms, standards, and research practices differ widely between fields and institutions.[14] Third, research products themselves differ widely and may require different treatments. Data and patents, for example, are governed by widely varying sets of intellectual property laws, institutional regulations, and professional norms. Fourth, funders should be mindful that requiring open access to data or methods may limit the journals authors may publish in or the data sources they may use.

In our case, the Sloan Foundation has opted for a very flexible Information Products policy. We ask prospective grantees to provide some plan laying out how research products will be made available, and to be mindful of certain principles when constructing it, but do not otherwise mandate specific forms of access. This gives us the flexibility

[13] See www.sloan.org/fileadmin/media/files/application_documents/proposal_guidelines_research_trustee_grants.pdf

[14] The Foundation has supported efforts by UCLA professor Christine Borgman to compile detailed scientific ethnographies focusing on how data is impacting scientific practice across different disciplines. Her recent book, *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, 2015), reports some of this research. The Foundation has supported similar research, focused on scholarly work with big data, by Eric Meyer at the Oxford Internet Institute.

to work with grantees on a case-by-case basis and craft a policy that makes sense for each particular researcher and project.

Insisting on the highest standards for reproducibility and replicability means not merely asking grantees to commit to making their data and methods available, but to ensuring that their research findings can be replicated if other scientists seek to try to replicate them. Funders of scientific research can make an impact by insisting that prospective grantees make explicit their empirical methodology and by subjecting that methodology to independent expert scrutiny. All empirical research grant applications to the Sloan Foundation must include an Empirical Research Methods appendix that specifies relevant theoretical and empirical models, experimental designs, data sources and attributes, sampling methods, identification of key parameters, estimation techniques, power calculations, and robustness tests. This information helps our staff and outside reviewers determine whether the proposed research will employ appropriate empirical methods, whether conclusions reached are likely to be robust given the data collected and analyses deployed, and how to improve the experimental design.

Lastly, funders can support the increased generalizability of scientific research by funding multiple experimental approaches to the same scientific question. For instance, a recent grant by the Foundation to the Environmental Defense Fund (EDF) supported the study of the magnitude of methane gas leaks associated with the extraction of shale gas. Rather than fund a single measurement approach, EDF commissioned 14 studies, conducted by independent research teams using different measurement and estimation methods.[15] This diversity of approaches permits a confidence in the studies' findings that a single study could not match.

2. Support the development of new tools that lower the costs of making data and methods available

Software packages, computing platforms, and digital archiving infrastructure have the potential to significantly lower the costs to researchers of making their data and methods available. These new technologies also have beneficial knock-on effects. If documenting, archiving, and sharing one's data and methods ceases to be burdensome, journals should be more willing to require such activities as a condition of publication.

The Foundation has made several major grants towards developing these new technologies. A $1 million grant to Gary King at Harvard University is supporting the expansion of the Dataverse repository[16]. King and his team have linked the Dataverse with the popular Open Journal System and the Open Monograph Press, two digital workflow platforms used by numerous academic journals. The new linkage allows authors to upload their data to the Dataverse repository as part of the standard article submission process. This, in turn, allows editors and referees to view the data a paper is based on when making publication decisions. The Foundation is also supporting the development of the Jupyter notebook[17], an exciting new computing platform designed to bring the traditional lab notebook into the digital age. At present, much scientific analysis requires using multiple computing package and programming languages: one to clean data, another for analysis, yet another to turn data into charts, graphs, and other visualizations. Developed by a team led by scholars Brian Granger and Fernando Perez, the Jupyter Notebook allows researchers to document their computational work *in situ*, combining narrative text, computational formulas from multiple languages, visualizations, and data into a single, useful research log that can be shared and collaboratively edited by others.

Other Foundation grants include support for the development of standardized modules for accessing commonly used scientific databases using the R programming language; and several grants to open source software developer Max Ogden for development of the DAT data versioning software, which helps researchers properly *version* datasets, allowing them to cite which version of a frequently updated dataset they used in their research. The Foundation has also supported the expansion of the University of California's Data Management Plan Tool, an online platform that helps researchers craft data management plans that comply with the requirements of funding agencies. A Sloan-supported collaboration between Sayeed Choudhury of The Johns Hopkins University and the Institute of Electrical and Electronics Engineers is crafting a software system that will link publications with

[15] See (Johnson 2015)

[16] https://dataverse.harvard.edu/

[17] https://jupyter.org/

the data on which they are based. In addition, the Foundation was an early supporter of the Open Science Framework, a new online platform being developed by the Center for Open Science that aspires to provide end-to-end research support in a single online interface, allowing scholars to use one system to document every part of their scientific process from data collection, to analysis, to archiving. In these cases and others, we seek opportunities to fund the development and dissemination of platforms that offer immediate, tangible value to researchers while nudging them towards more reproducible and replicable research practices.

3. Fund replication projects

Scientific funders are under many of the same pressures that fall on academic journals. Federal agencies are under pressure to maximize value to the taxpayer. Private foundations have an obligation to be prudent stewards of the funds left in their trust. Funding replication studies may thus seem to be a waste of scarce time and resources. Yet if we think reproducibility and replicability are important, then we have some obligation to fund it. The Sloan Foundation is currently supporting an innovative replication study by Colin Camerer at the California Institute of Technology, who has initiated a project to replicate some of the most famous findings in experimental economics with the cooperation of the original authors. The study is innovative because Camerer is surveying a group of economists to collect their predictions about what the replication studies will find, and confirming the experimental design and implementation with them. The project is the best of both worlds – a first class replication study of important experiments in economics and original research that will teach us something new. A complementary replication project by the University of Chicago's Devin Pope is also receiving Sloan support.

4. Foster the development and adoption of norms and institutions devoted to good replication practices

Scientists' work is supported by a host of interrelated institutions and norms. If scientific practice is to include efforts to increase its reproducibility, there must be norms and institutional infrastructure supporting those efforts. The Foundation is supporting a number of initiatives in this area. We have partnered with the National Academy of Sciences on a project to spearhead the creation of new norms for data citation. Such standards streamline the process whereby researchers can acknowledge which datasets underlie their work and reward the creators and curators of useful scientific data. Other foundation grants support research on data archives and storage infrastructure. With Sloan support, Kristen Eschenfelder at the University of Wisconsin, Madison is compiling a set of case studies on the sustainability strategies, successful or not, of data archives, helping us understand which roads are promising avenues to archival permanence. The Foundation is also supporting a project by Phoenix Bioinformatics to experiment with a flexible paywall service for TAIR, a popular repository of molecular and biological data. The pilot, if successful, could serve as a potential model for other data archives.

Other Foundation efforts in this area include a focus on fostering institutions and norms that encourage the pre-registration of research. Pre-registration of hypotheses to be tested and methods to be deployed in experiments have been shown to be an effective deterrent against hypothesis fishing and data mining, practices where researchers hunt through data in search of statistically significant connections between variables. Since roughly one in twenty statistically significant correlations will be an artifact of the data, data mining is widely seen to contribute toward the confirmation bias that has been observed in the scientific literature. With Sloan support the American Economic Association launched a public registry of randomized controlled trials in economics, asking researchers who are using RCTs to publicly register their methods and the hypotheses they will be testing. Since 2014, more than 400 randomized controlled trials have been registered on the site. Success could serve as a model for other disciplines in the social sciences. Sloan is also supporting research on new mathematical techniques for designing and analyzing randomized controlled trials that can achieve a given level of statistical power with smaller sample sizes and at less cost.

5. Train scientists directly

The Internet Age is changing the skills scientists need to do their jobs well. Datasets are often too large to be manually searched, cleaned, manipulated, or analyzed. Working with data often means working with computers capable of handling millions of records and that means programming them to do what you need them to do. In addition, the increased power of computers has opened up new channels for scientific analysis, allowing researchers to probe large datasets using complex statistical methods that would have been impractical only a generation ago. The twenty-first century scientist increasingly needs to understand statistics; computational methods; and data organization, manipulation, and curation techniques. But these skills are not yet a standard part of scientific training in some disciplines. Better science means training a new generation of working scientists in the best practices of software development and statistical methods. The Foundation is funding a number of training initiatives aimed at helping scientists master the increasingly complex skills required by modern science. Sloan grants to the Mozilla Foundation supported the growth of Software Carpentry, an organization of science-minded coders who provide software development boot camps for scientists, helping train them in the best practices of iterative software development and versioning. The Foundation is also supporting a project at Haverford College to experiment with early interventions with advanced undergraduates, incorporating best practices in experimental integrity, transparency, and reproducibility into the basic scientific curriculum. The Berkeley Initiative for Transparent Social Science (BITSS) actively promotes reproducible research, too, with funding from the Sloan, Templeton, and Arnold Foundations.

6. Support data science professionals

Taking scientific reproducibility and replicability seriously means taking data seriously, and that means cultivating a cadre of professionals whose job it is to work with data, to collaborate with disciplinary-scientists on data-intensive projects, to adapt existing data management tools for scientific use, and to develop new tools to aid scientists in data-driven discovery. We need institutions that value those researchers who do the important work of documenting, curating, and archiving data and we need to develop compelling and secure career paths for them to pursue. The Foundation is currently supporting several initiatives to support data professionals. Partnering with the Andrew W. Mellon Foundation, the Foundation is supporting an innovative fellowship program created by the Council on Library and Information Resources. Hosted at university libraries, supported fellows work to make data accessible and durable by advising scholars on how best to handle their data-intensive projects at every stage of the research process. The fellowships have been strikingly popular and many of the host institutions have committed to supporting the positions after external funding lapses. The Foundation is also working with the Research Data Alliance on creating fellowships for doctoral students who want to work on projects in effective data management and access in connection with scientific research.

Last, but most importantly, the Foundation has launched a five-year, $37.8 million initiative with the Gordon and Betty Moore Foundation to help empower data scientists and accelerate data-intensive, replicable scientific research. Partnering with New York University, the University of Washington, and the University of California, Berkeley, Sloan is helping create new data science centers focused, in part, on building durable, fulfilling career paths for data scientists, and in fostering their interaction with disciplinary scientists across the university. In addition to their other activities, the centers have convened a cross-university working group on Reproducibility and Open Science, with leadership provided by computer scientist Juliana Freire (NYU), mathematician Randy LeVeque (UW) and statistician Philip Stark (Berkeley). The result will be data-intensive research projects that are better conceived, better managed, more open, and more accessible to the scholarly world and the public.

7. Explore ways to facilitate the repeated scientific analysis of private and proprietary data

In an increasingly large number of cases, studies are not reproducible because the data the original study is based upon cannot be shared due to privacy restrictions or because the data are proprietary. Interesting opportunities abound for funders to facilitate the scientific use of such data. Recent Sloan grants support efforts by mathematicians and computer scientists to develop ways to query sensitive datasets that are both mathematically rigorous and protect the anonymity of the data they are analyzing. "Differentially private" techniques, for example, allow researchers to make aggregate statistical queries about a dataset while provably protecting the privacy of individuals' information contained in that dataset. Grants in this area have supported pioneers like Cynthia Dwork as well as implementers based at Harvard and at the MIT Libraries. Other Foundation work funds further development of the mathematical theory behind fully homomorphic encryption, a method of reliably analyzing data without having to decrypt them. Sloan supported work also includes a project led by George Alter at the University of Michigan's Interuniversity Consortium for Political and Social Research (ICPSR). In addition to exploring "secure multiparty computation," he is collecting samples of non-disclosure agreements (NDAs) signed by scientists in exchange for access to the company's proprietary data. Standardizing such agreements, so that businesses could safely rely on one of a few standard NDA templates, would make getting access to proprietary data much easier and would facilitate replication and reproduction. Cornell economist John Abowd is also being funded to examine the economics of privacy, cataloging the most popular privacy-preserving algorithms and evaluating their tradeoffs between accuracy and privacy. Abowd and others are also developing novel new ways to measure how much people value privacy, a topic that will be of significant interest should it turn out that some scientifically important datasets contain "private" information that no one actually wants to keep private. With Sloan support, computer scientists Adam Smith and Aaron Roth are investigating how techniques originally designed to protect privacy can also prevent false discovery and enhance reproducibility regardless of whether the data under study contains confidential information.

Conclusions

The Alfred P. Sloan Foundation is fully committed to supporting efforts to facilitate access to all components of the process that characterizes modern scholarly empirical research and its dissemination. We believe that these efforts can improve the quality of research in the long run, lower the costs of adopting better research practices, reduce mistakes that find their way into the published literature, facilitate reproducibility, deter both honest mistakes and fraud, and ultimately enhance the integrity of scientific research.

References

Alberts, Bruce, Ralph J. Cicerone, Stephen E. Fienberg, Alexander Kamb, Marcia McNutt, Robert M. Nerem, Randy Schekman, et al. 2015. "Self-correction in science at work." *Science* 348 (6242): 1420-1422.

Angrist, Joshua D. 2014. *Mastering Metrics: The Path from Cause to Effect.* Princeton: Princeton University Press.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Bollen, Kenneth, John T. Cacioppo, Robert M. Kaplan, Jon A. Korsnick, and James L. Olds. 2015. "Social, Behavioral and Economic Sciences Perspectives on Robust and Reliable Science." Subcommittee on Replicability in Science, National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Fanelli, Daniele. 2010. ""Positive" Results Increase Down the Hierarchy of the Sciences." *PLOS One* 5. doi:10.1371/journal.pone.0010068.

Ioannidis, John, and Chris Doucouliagos. 2013. "What's to Know About the Credibility of Empirical Economics." *Journal of Economic Surveys* 27 (5): 997-1004.

Johnson, Scott K. 2015. *Measuring the heck out of shale gas leakage in Texas.* July 24. http://arstechnica.com/science/2015/07/measuring-the-heck-out-of-shale-gas-leakage-in-texas/.

Kaplan, Robert M., and Veronica L. Irvin. 2015. "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time." *PLoS ONE* 10. doi:10.1371/journal.pone.0132382.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. "Promoting an open research culture." *Science* 348 (6242): 1422-1425. doi:10.1126/science.aab2374.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11). doi:10.1177/0956797611417632.

The Economist. 2013. "How science goes wrong." *The Economist*, October 19.

—. 2013. "Trouble at the lab." *The Economist*, October 19.