MIT Sloan School of Management

MIT Sloan School Working Paper 4773-10

The Scientific Method in Practice: Reproducibility in the Computational Sciences

Victoria Stodden

THE SCIENTIFIC METHOD IN PRACTICE: REPRODUCIBILITY IN THE COMPUTATIONAL SCIENCES

Victoria Stodden*

February 2010

* MIT Sloan School of Management
50 Memorial Drive, Cambridge, MA 02141, USA;
Postdoctoral Associate in Law and Kauffman Fellow in Law and Innovation
Yale Law School
127 Wall St., New Haven, CT 06511, USA
vcs@stanford.edu

THE SCIENTIFIC METHOD IN PRACTICE: REPRODUCIBILITY IN THE COMPUTATIONAL SCIENCES

ABSTRACT

Since the 1660's the scientific method has included reproducibility as a mainstay in its effort to root error from scientific discovery. With the explosive growth of digitization in scientific research and communication, it is easier than ever to satisfy this requirement. In computational research experimental details and methods can be recorded in code and scripts, data is digital, papers are frequently online, and the result is the potential for "really reproducible research."[1] *Imagine the ability to routinely inspect code and data and recreate others' results: Every step taken to achieve the findings can potentially be transparent. Now imagine anyone with an Internet connection and the capability of running the code being able to do this.*

This paper investigates the obstacles blocking the sharing of code and data to understand conditions under which computational scientists reveal their full research compendium. A survey of registrants at a top machine learning conference (NIPS) was used to discover the strength of underlying factors that affect the decision to reveal code, data, and ideas. Sharing of code and data is becoming more common as about a third of respondents post some on their websites, and about 85% self report to have some code or data publicly available on the web. Contrary to theoretical expectations, the decision to share work is grounded in communitarian norms, although when work remains hidden private incentives dominate the decision. We find that code, data, and ideas are each regarded differently in terms of how they are revealed and that guidance from scientific norms varies with pervasiveness of computation in the field. The largest barriers to sharing are time involved in preparation of work and the legal Intellectual Property framework scientists face.

This paper does two things. It provides evidence in the debate about whether scientists' research revealing behavior is wholly governed by considerations of personal impact or whether the reasoning behind the revealing decision involves larger scientific ideals, and secondly, this research describes the actual sharing behavior in the Machine Learning community.

[1] See J. Claerbout, http://sepwww.stanford.edu/sep/jon/reproducible.html

INTRODUCTION

As scientific computing becomes increasingly central to the modern scientific enterprise, standards on openness and verifiability have not kept pace with standards for reproducibility of results. This paper surveys computational scientists to understand the factors that underlie their decisions to share or not share code and data on the Internet. The investigation is built on three pillars of intellectual work, but the underlying interest is expressed with this comment from a survey respondent, "`my opinion: if it's not open and verifiable by others, it's not science, or engineering, or whatever it is you call what we do.`"

Sociological research predicts that scientists will act in self-interested ways with regard to exposing their work. This survey provides evidence that scientists' decisions to share are most typically motivated by higher ideals such as concern for scientific advancement and the encouragement of sharing in the research community generally. The survey confirms conclusions from the literature regarding the free revealing of information in industry: that even proprietary knowledge is shared among competitors in the spirit of generalized reciprocity. We also found that legal barriers deriving from Intellectual Property law place highly as factors that prevent sharing.

The paper is constructed from a sociological point of view: the digitization of science combined with the Internet create a new transparency in scientific knowledge, potentially moving scientific progress from building with black boxes, to one where the boxes themselves remain wholly transparent. Science of course has not always been open. After mathematician Ramanujan's death his collected works were published in1927 and "the floodgates opened" with a rush of subsequent publications, building on Ramanujan's newly released work. (Kanigel 1991, p337). Before this publication his work was kept secret and meted out through filters such as co-authorship with western mathematical luminaries such as G. H. Hardy, who had access to his work.[2] Scientific communication is at a new juncture of transparency and reproducibility.

[2] Atle Selberg, one of the world's most famous number theoreticians credits seeing Ramanujan's *Collected Works* with giving him the impetus for his own mathematical work. There is no way to tell, but because Ramanujan's work was revealed after his untimely death, perhaps *more* research was engendered than if he had actually lived, and his work remained relatively cloistered for a longer period of time. (See Kanigel, p338)

TRANSPARENCY IN SCIENTIFIC REASONING AND REPRODUCIBILITY

Openness and the sharing of scientific reasoning are long established norms within the scientific community, and a vital part of the process of evolving a finding to an accepted fact. Two threads in the literature have addressed incentives scientists face in deciding to reveal aspects of their work. The first is grounded in the study of the sociology of science and incentives regarding replication, and the second in the nature of revealing of proprietary information between firms. The scientific setting differs from the industrial setting, but understanding the influences acting on decisions to reveal proprietary information by competing firms gives structure to a fruitful inquiry into the sharing of research in the scientific context.

Merton (1942) was one of the first to study the mechanisms by which society's stock of scientific knowledge is created, defining a scientist as one who follows 'the ethos of science,' comprised of four norms, repeated below for convenience:

1. *Communism*. Property rights only extend to the naming of scientific discoveries (Arrow's Impossibility Theorem, for example). All other intellectual property rights are given up in exchange for recognition and esteem;
2. *Universalism*. Scientific discoveries must conform with previously confirmed knowledge and are evaluated in terms of universal or impersonal criteria, not on a subjective basis such as class, gender, religion, race, nationality, or affiliation;
3. *Disinterestedness*. Scientists must appear to act in ways that are selfless in their pursuit of claims to truth, thereby ensuring integrity in the research process and reducing fraud;
4. *Organized Skepticism*. Scientists are critical: All ideas must be tested and are subject to rigorous structured community scrutiny.

In Merton's model, voluntary sharing of methods, results, and scholarship is implied: "Communism" suggests scientists not create property rights and instead give their work over to the community; the selflessness inherent in "Disinterestedness" implies a social aspect to scientific investigation evident in both sharing and transparency; and "Organized Skepticism" suggests sharing by requiring results and ideas to be subject to verification by the community. These objects of scholarship are not seen by Merton to belong to the individual scientist but to the larger community. David (2003) observed that the full disclosure of findings and methods actually prescribes the Organized Skepticism norm, since it creates an expectation that claims to truth contributing to the stock of knowledge will have been subject to "trials of verification." (p3). Polanyi (1962) supported Merton's characterization of scientists as a community of independent men and women freely cooperating, calling it "a highly simplified example of a free society." Kuhn (1962) proposes an alternative to Merton's norm-based scientific organization by suggesting that scientists follow a pre-established paradigm dictating what problems are to be investigated, which creates a guide to appropriate future research. Scientists' current practices may or may not conform to the established rules of the scientific method, in that previously established results are taken for granted as building blocks for their own discovery work, such as the solving of concrete problems.

Bourdieu (2002) argues change comes from innovators confronting the defenders of the paradigm, where these innovators may draw intellectual strength from societal currents outside the scientific community. (p15). According to Bourdieu disputes are settled though strength of capital: scientific, symbolic, and social capital, used to exert pressure to enforce conformity with this particular researcher's views.[3] This is how others are convinced of the truth and, according to Bourdieu, how truth is established. For Bourdieu, scientific behavior is defined by those trying to maximize the amount of those forms of capital in their possession. Merton's norms of cooperation and disclosure, coupled with the verification of results create an "incentive compatibility" according to David (2003), in that these two norms self-reinforce and imply each other. This behavior is also encouraged by science's "reputation-based reward system grounded upon validated claims to priority in discovery or invention" which requires voluntary openness and sharing of scientific results, including sufficient information that they can be validated by the scientific community (p4).

Merton (1949) introduced the concept of 'obliteration by incorporation' to describe sharing in the sciences: A concept becomes so popularized that its inventor is forgotten, no longer cited, and the idea is considered common knowledge. Merton himself contributed a number of these when he coined phrases in our regular lexicon such as "self-fulfilling prophecy" (1968) and "unintended consequences" (1936). Facts are artificial in the sense that they are manufactured: "a fact is nothing but a statement with . . . no trace of authorship." (Latour and Woolgar 1979: 82). The steps taken to establish this fact are forgotten: the traces of research, the disputes and negotiations between research groups. Latour (1987) takes this idea further when he describes scientific progress as the construction of "black boxes," meaning that successful scientific discoveries are those that are identified by a shorthand name and survive without citation or authorship attribution.[4] Latour proposes the mechanism of the "black box" as both an explanation for the progress of scientific research, and as a method of resolving disputes. A "black box" is a term used to describe a complex set of commands, machinery, or a methodology underlying a result, that is too intricate to be represented by its full description, and for which only the inputs and the outputs need to be known.

Latour's black boxes are created when agreement on new knowledge is established and the uncertainty inherent in the process of discovery is removed to an outside observer. In his theory this is an essential mechanism for scientific progress, since it allows for the modularization of scientific knowledge and enables new ideas to be built upon the old. A black box is costly and difficult to open since this implies questioning results and understanding how they came to be. Thus this system of black box creation discourages dissent and questioning of results. According to Latour, science proceeds as an effort to close black boxes which, in turn, have been built upon previous black boxes.

---

[3] Scientific capital is that which augments a scientist's ability to make scientific achievements, specifically the ability to get his or her arguments noticed, such as standing in his or her field, prestige, or resources. Social capital refers to the strength of relationships the scientists has managed to develop within his or her field through which he or she can mobilize power and resources. Symbolic capital refers to that which a scientist can use to advance his or her interests, and maximize his or her symbolic profit, in particular his or her scientific authority.

[4] This is slightly different to Merton's formulation where, in *Communism*, he asserts that scientists give up property rights over discoveries with the exception of names. Theorems and factual discoveries are permitted to be named after people without violation of the Communism norm.

Barnes & Bloor (1982) suggest that scientific theories can never be determined or confirmed by data, given that several theories can point to the same data. Thus, consensus is fragile and controversies are often not resolved by the evidence but come to an end in any event, by other means. Collins and Pinch (1998) have found that when scientists replicate others' work, they tend not to do so exactly, but subvert the previous procedure for their own ends/programs.[5] Bourdieu postulates that this may be due to the opacity of the methodology in published papers (p20). He also suggests that scientific papers conform to specific norms for written presentation of results rather than describe what actually happened, and scientists can repeatedly get good results without explaining how. This is a manifestation of Polanyi's concept of "Tacit Knowledge:" Some elements of what we do aren't reducible to a series of instructions and thus are impossible to share, even though they may be vital to understanding the scientific experiment and results. As an example he asks readers try explaining how to ride a bike through instruction alone (p.). "Scientific research – in short – is an art" (1951; 57).

In the words of Bourdieu, Collins and Pinch directly contradict Merton's Universalism norm by suggesting both negotiations regarding fact creation and acceptance of results revolve around "judgments about questions of personal honesty, technical competence, institutional affiliation, style of presentation and nationality." And, "[i]n short, Popperian falsification gives an idealized image of the solutions provided by the 'core set' of scientists in the course of their disputes." (p20).[6] Polanyi (1951) gives an instructive analogy for understanding the organization of scientists. He imagines a group of workers trying to assemble a large jigsaw puzzle (p35). Each worker alone may have a few pieces, but cannot see how they fit into the overall puzzle. Polanyi suggests scientific research is an efficient solution to this problem in that it provides a way for each worker to keep track of what the other workers are doing in their jigsaw puzzle assembling work. Due to the openness inherent in scientific research, when someone fits in a piece of the puzzle, the others can watch and see the next steps that become possible. Each worker acts on his or her own initiative, and according to private interests, but helps to further the entire group. Polanyi suggests that when new research avenues are created, scientists become aware of them via perfectly open communication, and decide to work on areas that maximize their intellectual and emotional reward. Polanyi (1951) suggests that the freedom to pursue truth, acquire knowledge for it's own sake, and react to claims made by peers is a prerequisite for contributions to society's stock of knowledge (p40, 69). He envisions scientists as community members bound by a commitment to truth, what he termed "The Republic of Science," and exhibiting behavior resembling that of a free market for truth, rather than profit. Hagstrom (1965) states that scientists give each other new information that he or she has discovered, to receive recognition for those discoveries in return (p16-22).

""Replicability... is the Supreme Court of the scientific system. In the scientific value system replicability symolizes the indiference of science to race, creed, class,

---

[5] Collins also develops the concept of "experimenter's regress," that each replication, or attempt at confirmation of results, requires a further experiment to confirm it and so on ad infinitum. Collins, Changing Order, University of Chicago Press, 1992, p2.

[6] Popperian falsification itself is questioned by the Duhem-Quine critique: that empirical falsification is ambiguous. Ie. one cannot be sure that the theory has been falsified, rather than another aspect of the experimental setup.

colour and so forth. It corresponds to what the sociologist Robert Merton (1945) called the 'norm of universality". Anybody, irrespective of who or what they are, in principle ought to be able to check for themselves through their own experiments that a scientific claim is valid." (Collins 1985). Collins discusses several case studies in replication of scientific results, building a case that this aspect of scientific practice is highly culturally influenced, rather than a straightforward implementation of Merton's norms. "Repeatability...is the touchstone of common sense philosophy of science. .. It is crucial to separate the simple idea of repeatability from the complexities of its practical accomplishment." (p18-19). He chooses examples that represent Kuhn's three phases of science, the revolutionary, extraordinary, and normal phases (1962), showing how replication is nearly impossible in practice, in non-computational science. The rules of the game have changed with the increased use of computational tools in science, that afford, in theory, precise replication of published results by independent parties. This focus of this paper is on error checking and the transmission of the scientific knowledge underlying computational results, such that they may be verified, not with the extension of our body of scientific knowledge through the testing of previous results that therefore "must be neither exactly the same nor too different." (Collins p 34)

In modeling information sharing in scientific research Fröhlich (1998) postulates that scientists strategically hold back information from peers in their field, group, or laboratory, in their publications, and at conferences. He suggests that this behavior is not the exception, but a prevalent practice. Fröhlich describes three principles of scientific communication, countering Merton's four norms of scientific behavior (p541):

1. Communicate informally just as much as absolutely necessary to keep scientific groups functional;
2. Publish the minimum needed to preserve one's claim for priority on the findings;
3. Publish as little information of practical use as possible to prevent giving competitors a competitive advantages.

Information that could be withheld could be experimental details, particularly specialized local knowledge about the methodology. Fröhlich also suggests that scientific jargon is an attempt to frustrate efforts of other scientists seeking to gain knowledge and thus advantage from the work (p540). David (2003) describes the output of scientific research as a public good, and notes that public goods create an incentive to "free-ride" on the co-operative actions of others (p3). David notes that even with this temptation, the fact that the system has survived in tact indicates that scientists must still find a greater reward in sharing, despite the potential for free-riding.

This discussion suggests that individual scientists are rational actors and will only freely reveal if it is in their best interest, as they see it. But "best interest" can be evaluated in several ways. If a scientist is rewarded by an increase in capital in any form, such as public recognition (Bourdieu) or claims on priority (Fröhlich), then there is an incentive to freely reveal. If free revealing hastens the closure of a black box, then again there is a reason for the scientist to freely reveal (Latour).[7] If sharing allows a scientist to

[7] This can be interpreted as a restatement of an accrual of capital if scientists are publicly recognized for black box closure, and it can also be interpreted in Fröhlich's framework if the creation of a black box is

advance his or her work, by participating in a community that shares, then there is an incentive to reveal (Polyani). The work of Polanyi gives a second rationale for black box closure and the withholding of experimental or methodological details: That it is impossible to transmit details of the work that are "tacit."

In Latour's model the creation of black boxes is an epistemological strategy for the construction of new knowledge. Latour insists this cannot be done by one person in a vacuum, but that a new fact is a collective object and others must be marshaled in support, bringing to mind Bourdieu's description of science as the accumulation of social, scientific and symbolic capital. The black box has typically been difficult to reopen: involving a re-examination of the methodology that led to the results and a possible re-opening of any controversies that were involved with the original experiment, or the generation of new controversies. Within each black box are, necessarily, many further black boxes that need to be unpacked, questioned, and understood. Indeed, in Latour's world scientists have an incentive to make the black box as difficult to open as possible since on a macro level it is no longer a useful building block in further research if opened, and on an individual level one's own results being re-questioned removes the reputational reward for research. With the sharing mechanism of the Internet, scientists now have the option of not sealing the black box, even keeping it transparent.

Borgman (p196; 2007), focusing on the role of data sharing in scientific scholarship, directly suggests four reasons scientists may fail to reveal their data: 1) insufficient perceived reward, such as promotion or subsequent citation, 2) effort in documenting, 3) concerns for priority, including control of results and sources, and 4) intellectual property issues. Borgman suggests a sense of ownership over data can lead to the creation of bargaining power for scientists against those who may want to see the data. Freely revealing data would erode this advantage over other scientists. What is unclear, Borgman continues, is the true nature of ownership of data when a diverse group of funders, scientists, and collaborators have contributed to its creation. This latter point was reflected in at least one comment by a survey respondent, "Intellectual property is usually a major consideration. It is unusual for me to have complete and sole ownership of a data set. Good data is so difficult to collect that there are inevitably other people and organizations involved." Concerning this point, Stodden (2009) addresses IP issues scientists are subject to in the U.S. as well as typical funding requirements that data be made available to the public. Supporting Borgman's description of data ownership, another respondent commented that "the data (e.g., from animals) is hard to collect, and even harder to fully document and standardize. Thus data becomes a bargaining token. For theoreticians, it means dibs on new data and a say in new experimental design; for experimentalists it means much more visibility for their data/experimental paradigm, and an opportunity to more thoroughly investigate the problem area. Plus an opportunity to make sure that their interpretation/worldview gets solid representation."

Borgman's four factors can be applied to the code component of the research as well as to the data component, with the exception that code and data each fall under a different Intellectual Property structure. Code is typically copyrightable by default, where

---

either an assurance of the scientist's claims on priority or a method of obscuring information to prevent advantaging other scientists.

the raw facts encased in a dataset are not. There may not be sufficient rewards for revealing code, for taking the time to document it for release, and sharing may engender a loss of control over priority in future work, just as Borgman articulates in the case of data. Effort in documentation was broken down by a survey respondent into three parts, "The time to find a place to put the code, or write the web page -- I guess it comes under documentation, but there are 3 time factors: (1) setting up or finding a platform, (2) explaining how to use or run the code and (3) internal documentation within the code."

Allen (1983), first reporting free revealing in industry and labeling it "collective invention," found three factors he used to explain why this was successful in blast furnace design in nineteenth century England. First, he showed that a solid technical understanding of the blast furnace was lacking, and thus performance uncertain and previous experience valuable; second, reputation was augmented for those who shared useful information; and thirdly, despite revealing, blast furnace owners still made economic gains from their innovations, although they were diffuse in that they benefited other firms as well since the value of iron ore was increased through the technological improvements. Aside from directly realized profit, an increase in firm reputation motivated the sharing of proprietary information both informally and formally in professional societies and through publication in technical journals by employees of competing firms. Scientific research, especially when analogized as a process of puzzle completion by Polanyi, bears a striking resemblance to Allen's setting: scientific scholarship centers on problems with no known solutions; reputation is the main mode of reward for research and discovery; and a discovery in one area may "spill over" and have a positive impact beyond the immediate application for which it was derived. This is evident when Polanyi describes the information gained by scientists as they keep track of progress on the entire puzzle (their field), even though from their perspective they only contribute in an isolated region. A scientist's reputation may also be increased through free revealing by factors beyond citation: high quality elements of the work might come to light on the web, which might not have otherwise. Clever code or well-organized data, for example, can have the effect of increasing trust and thus reputation (through the mechanism of Bourdieu's scientific capital), in addition to facilitating citation.

A similar free revealing phenomenon to Allen's furnace example was found by Nuvolari (2004) when he studied nineteenth century Cornish mining. Mine managers from competing firms founded a journal, *Engine Reporter*, to share information on engine use for water removal from mines. The two motivations here appear to be different: in Allen's case standing was increased, and thus presumably profitability, through the sharing of designs, whereas in Nuvolari's case sharing seems to have been a rebellious reaction to stringent patent enforcement by Watt, inventor of the steam engine. Nevertheless, the *Engine Reporter* permitted attribution and reputational reward for innovators publishing in the journal. As in Allen's case the mine pumping problem was not yet fully understood and there were spillover effects in that iron ore became more valuable overall as any one mining company improved their technology. Lerner and Tirole (2002) and Raymond (1999) studied the conditions for free revealing in the open source software context and found that programmers who freely reveal code are similarly rewarded by gains in reputation, and possible increases in their value on the job market. Raymond also postulates that only a very low level of damage could accrue to the free revealing programmer. In a comprehensive survey of coders who participate on the open

source website SourceForge.com, Lakhani and Wolf (2005) found that many programmers shared their code simply because they enjoyed learning from the coding experience. von Hippel and von Krogh (2006) found that being part of a community of developers and learning through feedback from peers motivated coders to reveal their work.

In scientific research there is a norm allowing a commensurate increase in the time data can remain sequestered with the difficulty level inherent in its collection. Unique data, more likely when the difficulty level is high, is more likely to remain closed because of its higher value induced by its monopoly-like status. For example, data collected from the Large Hadron Collider requires a series of rules to enforce openness.[8] As Rolf-Dieter Heuer, director general of CERN, explains, "Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data.…"[9] Non-revealers in science, as opposed to the open source world, may be acting shrewdly if the disclosure of methodology through making code and data public is not required. Scientists who do make their entire research compendium available risk having their work inspected more closely and perhaps increase the chance of errors being found.

Being the first to reveal an innovation in industry may have a payoff that outweighs any potential loss from revealing freely. For example, being first with a new idea or result increases the chances of wider adoption or the establishment of a standard around the revealed product. (Harhoff and von Hippel (1985)). This occurs in scientific research: Donoho (2002), who was first to freely reveal wavelet research including code and data, laments the repeated use of his testbeds for signal processing applications beyond those for which they were intended.[10] His test cases have become a standard benchmark by which to test certain types of algorithms.

Ekeh (1974: 48) states that generalized norms of reciprocity refer to occasions when "an individual feels obligated to reciprocate another's action, not by directly rewarding his benefactor, but by benefiting another actor implicated in a social exchange situation with his benefactor and himself." There is an analogy among scientists, if one benefits from publicly shared data or code, there is an expectation that modifications are subsequently shared. Bouty (2000) reported informal information trading among scientists working in the R&D units of firms. She obtained detailed reports of 128 accomplished and attempted "resource exchanges" by firm scientists with external colleagues. In these exchanges, scientists would informally provide information on their research and other research-related services to colleagues in competing firms who request it. She found that these incidents involved expectations of reciprocity understood both by people who give the information and those who receive it.

Lerner and Tirole (2001) identify several key research questions they felt required answering in any study of the Open Source software community (Nuvolari 2005). Two

---

[8] See http://www.w3.org/History/1989/proposal.html

[9] Computer Weekly, August 6, 2008. Available at http://www.computerweekly.com/Articles/2008/08/06/231762/in-search-of-the-big-bang.htm (last accessed July 16, 2009).

[10] See http://www-stat.stanford.edu/~wavelab/

can be adapted for our context:

1. As a spontaneously provided "pure" public good, scientific research should be prone to the free–rider problem; how can research projects encourage the active participation of talented investigators?
2. Research that is fully revealed on the web is subject to default copyright restrictions that work against the scientific norms of reproducibility and building upon others; work. Can an adjustment of this intellectual property scheme increase the rate of discovery in the sciences?

The second question is taken up in (Stodden 2009), which addresses the appropriate copyright structure for scientific works released on the web. The first allows us to structure our information on industrial sharing to the scientific context. This evidence supports our main hypothesis: that scientists share when it pays personally, that is, the private gains from sharing behavior outweigh the costs. But we can formulate a more precise analysis. The free rider problem is not so clear cut in science as it is in the open software community or in industry. In science the only way to free ride is to use another scientist's work without citation or to preclude a publication he or she was poised to make.[11] Unless publications are precluded, use of another scientist's materials, properly attributed, boosts the reputation of the revealing scientist, rather than siphoning away return. The loss to a revealer to free riders is an opportunity cost: the loss of future publications or recognition, and the loss of time spent preparing work for release that could have been spent otherwise. Scientific research, in terms of code and data, is not costly to hide, except in terms of forgone citation from scientists who may have built on the work. Data and code usually reside on private machines and of course no effort is required to leave them there.

Lakhani and von Hippel (2003) point out that there may also be pressure for programmers to reveal: if they do not contribute their code, someone else could contribute code with the same functionality, thereby blocking their contribution. There are many examples of scientists racing to publish a result that they fear others might report first.[12] If priority can be established through code and data releases, establishing priority could be a compelling reason to reveal publicly, although typically in scientific research priority is gained through publication.

These two threads in the literature reveal two hypotheses:

Hypothesis 1: Sharing occurs because scientists perceive personal gain from doing so, through either reputation gains, claims on priority, black box closure, acceleration of the research process, encouragement of others to work on the problem, or advantage through the establishment of standards, despite the possibility of free-riding. Sharing does not

[11] I am assuming that the revealed research is not stolen and is used within the scientific norm of giving attribution. As mathematician G. H. Hardy said in 1908, "Surely it is obvious that, if I were to make any illegitimate use of your results, nothing would be easier than for you to expose me." (Kanigel, p181). Free revealing on the Internet has the convenient property of creating a time stamp of when the revealing took place, so subsequent uses of ideas or material released on the web can be easily exposed.

[12] See, for example, Ingrid Daubechies, Ten Lectures on Wavelets, 1992.

occur when scientists perceive a personal loss from doing so, such as time taken to prepare code and data for release, insufficient rewards, or loss of control over priority on derivative works.

This hypothesis describes scientists as concerned only with their immediate self-interest, in both their decisions to share and not share their research. From the preceding discussion a second hypothesis emerges: scientists view themselves as belonging to a community and seek community membership and feedback through sharing their work. Seeking feedback can be viewed as satisfying private incentives, in accordance with Hypothesis 1, but seeking acceptance in a community can be seen as behavior induced from larger communitarian ideals. Being a good community member implies an acceptance of and support for the community structure.

Hypothesis 2: The willingness to reveal work reflects a scientist's desire to belong to a community and to gain feedback on his or her work.

Both of the reasons given in Hypothesis 2 address why a scientist may want to share, not why he or she may choose to keep work private. In this sense Hypothesis 2 is a subset of Hypothesis 1 which makes an assertion about the nature of reasons for both sharing and not sharing. These two hypotheses suggest a survey design: measuring the influence of factors on the decision to reveal or not reveal code, data, and ideas.

EMPIRICAL STUDY OF SHARING BEHAVIOR

In surveying "computational scientists" the first step was to define the term. A computational scientist is *an academic whose research has both code and data components*. I chose to focus on scientists using computing heavily in their work and interacting primarily with other computational scientists, in particular those working in the field of machine learning. Examining a representative cross-section of this population, from those new to the field to experienced icons, from highly ranked schools and smaller departments, and across a variety of fields, would shed the most light on the factors underlying sharing. I chose to question researchers registered for one of the largest and most prestigious machine learning conferences spanning many different fields, the Neural Information Processing Systems (NIPS) conference,[13] held every January in Whistler, British Columbia, Canada.[14]

A. The Survey Sample

NIPS and ICML (International Conference on Machine Learning) are the two most prestigious and biggest conferences in the Machine Learning community. An important benefit to examining this conference is that part of its focus is on biological systems and bioinformatics, giving a wide variety of academic backgrounds in the study. Of the program committee for NIPS '08 about 40% of the 27 members are involved in

[13] http://nips.cc/

[14] I was a speaker at NIPS in 2003. See David L. Donoho and Victoria Stodden, "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?" NIPS Proceedings, 2003.

work that has a biological dimension: genetics, neuroscience, or medical imaging for example. Registrants come from a wide variety of departments: mathematics, statistics, psychology, neurology, computer science, electrical engineering, linguistics, robotics, medicine are all represented for example.

Studying one group of conference attendees has compelling advantages. The group is coherent in the sense of sharing similarly situated computational research. The registrants have common research interests in machine learning and its application in computational and biological problems and span a broad range of fields. The NIPS registrants vary from master's student through emeritus professor, and span a large number of schools that are geographically diverse and vary widely in terms of size, wealth, and prestige. Their common feature is their interest in and success at publishing in computational machine learning.

Up to and including 2008, there have been 1,758 NIPS registrants. I eliminated from the study those without an email address ending in .edu to restrict my research to scientists affiliated with American universities and thus under the same Intellectual Property legal framework. Doing this also allowed me to study a group with reasonably consistent institutional standards regarding expectations and valuation of academic work. Restricting the target sample in this way dropped from the survey group both registrants who were affiliated with a company and those who provided no email address.[15] After reducing to .edu email addresses I was left with 1,008 researchers.

B. Survey Design

The survey was designed to address questions that came to light in both interviews and in the two threads in the literature. I conducted pre-pilot interviews to solicit input from computational scientists, and ran a pilot survey. The story drawn from the sociology and industrial free revealing literature isn't a complete description of code and data sharing for computational scientists. From the pilot and interviews, additional reasons underlying the sharing decision included normalizing understanding in a field, the time to document code and data, the potential to find collaborators, protection from commercial application, publicity, the desire to track use of your work; contentiousness of the topic; the number of requests received for the code or data, the amount of effort exerted in created the code or data, the impact on future grant funding, and privacy concerns with data are all factors that were cited as important to sharing decision making. "Table 1: Factors Reflecting Private Incentive" lays out the factors the final survey asked participants to rate for influence. The patent system provides a disincentive to reveal code since software can be patented and posting code publicly could establish prior art and make obtaining the patent more difficult, if not impossible. Accordingly, the survey includes factors to measure the influence of the Intellectual Property framework scientists are subject to. "Table 1: Factors " and "Table 2: Factors Reflecting Communitarian Ideals" list each of the factors included in the final survey, with its source as discussed earlier in the previous section. The factors are divided into those reflecting private incentives and those reflecting communitarian ideals, and those that do neither.

[15] Registrants who gave no email address are unlikely to have had a paper accepted at that conference and so helps reduce the survey group to those making publishable contributions to computational science.

Table 1: Factors Reflecting Private Incentive

	Source
Increase in publicity	Hagstrom, Interviews
The potential to set a standard for the field	Harhoff & von Hippel
Opportunity to get feedback on your work	Polanyi, von Krogh, Interviews
The number of requests you receive for the code	Schrader
Being first to release code in this area	Borgman, Interviews
Impact on future grant applications	Pilot Survey
Whether you put in a large amount of work building the code or data	Interviews
Requirement for publication	Interviews
Required by research funders	Pilot Survey
Dealing with questions from users about the code or data	Schrader
The time it takes to clean up and document for release	Borgman, Interviews
The time it takes to verify privacy or other administrative data concerns (data only)	Borgman, Interviews
The potential loss of future publications using this code	Interviews
The possibility that your code may be used without citation	Interviews
Competitors may get an advantage	Von Hippel, Latour, Frohlich, Borgman

Table 2: Factors Reflecting Communitarian Ideals

	Source
Encouraging scientific advancement	Pilot Survey
Encouraging sharing and having others share with you	Ekeh, Bouty
Being a good community member	Polanyi, von Krogh
Improvement in the caliber of research	Pilot Survey
Normalizing understanding in a field	Pilot Survey
Conforming with the requirements of the scientific method	Merton, Pilot Survey

Table 3: Factors Reflecting both Private Incentive and Communitarianism

	Source
The potential to set a standard for the field	Harhoff & von Hippel
Potential to encourage others to work on the problem	Allen, Interviews
The topic is receiving a lot of attention	Pilot Survey
Conforming with the usual practices of the research community	Polanyi
Potential to close a line of inquiry and move the research to the next step	Latour, Baldwin
Whether there is intense competition in the topic	Schrader, Allen
Availability of other code or data that might substitute for your own	Schrader

The code might be used in commercial applications (code only)	Pilot Survey
The possibility of patents, or other IP constraints (code only)	Stodden
Legal barriers, such as copyright	Stodden
Technical limitations, ie. webspace platform space constraints (pilot)	Pilot Survey

Respondents were asked to indicate the level of influence of the factors on a scale 7-point scale ranging from "Very Strong Influence to Share," through "No Influence" to "Very Strong Influence NOT to Share." As indicated in the tables above, the respondents were asked slightly different factors for code and data, because of the differing IP structures for each. The order the factors were presented to each participant varied randomly, to eliminate any possible biasing effect due to factor ordering. Each respondent was asked for any further influencing factors in the event there were other reasons they decided to share or not share their most recent NIPS paper. Additional questions asked them which components of their research they were comfortable revealing on the web, how many potential papers they feel they have had scooped, and to estimate the proportion of code or data they have posted on the web. I also asked them for further thoughts and whether they are interested in receiving a copy of this paper. (See appendix for a copy of the final survey.)

C. Survey Results

The final survey was sent in two waves to the remaining 638 scientists in the sample, with 37 bounces and 5 away from their mail.[16] Three further researchers were dropped from the study when their response indicated they were at a firm or in a foreign university. The final response rate was 134 of 593 or 23%. I obtained demographic and other information for each respondent from his or her website. The distribution of characteristics is displayed in "Table 4: Descriptive Characteristics (from Respondent's Website)". Respondents' university affiliation was grouped into 5 groups using the U.S. News and World Reports ranking for computer science departments for 2008. This is not a ideal proxy for prestige since 34% of respondents were not from computer science departments. Respondents tend to be early in their career, male, from a computer science department.

Table 4: Descriptive Characteristics (from Respondent's Website)

Characteristic	Count	Proportion
Position		
Master's Student	1	0.01
PhD Student	56	0.42
Postdoc or Fellow	20	0.15
Assistant Professor	26	0.19
Associate Professor	14	0.10
Professor	15	0.11

[16] See http://www.stanford.edu/~vcs/Survey2009/SharingSurvey.html for an example of the webform used in the survey.

Gender		
Male	120	0.90
Female	13	0.10
Department		
Electrical Engineering or Computer Science	88	0.66
Neuroscience, Neurobiology, Biomedical engineering	10	0.07
Statistics, Biostatistics, Medical informatics	9	0.07
Medicine, Psychiatry	6	0.04
Mathematics	6	0.04
Physics	3	0.02
Biology	2	0.01
Music	2	0.01
Psychology	2	0.01
Finance, Operations Research	2	0.01
Philosophy	1	0.01
University (by US News CS Department Ranking 2008)		
MIT, Stanford, Berkeley; CMU; UIUC	35	0.26
Cornell, Princeton, U Washington; GATech, UTAustin	18	0.13
CalTech, UW-Madison; UCLA, U Maryland, U Michigan	11	0.08
Columbia, Harvard, UCSD; Purdue; Brown	15	0.11
All others	55	0.41

Respondents were asked about their comfort level in sharing their final publication, pre and post publication code and data, and ideas pre-publication. "Table 5: Research Component Sharing Rates (Survey Results)" reveals unsurprisingly that essentially all respondents were willing to post their final paper on the web, but surprisingly a strong majority were willing to share their post publication code and data, and more seemed willing to share pre-publication code than pre-publication data.[17] Respondents self-reported an average of 32% of their code available on the web, and 48% of their data, with 81% claiming to reveal some code and 84% claiming to reveal some data.

Table 5: Research Component Sharing Rates (Survey Results)

Proportion Comfortable Sharing on the Web…	
Final Paper	99%
Draft Paper	26%
Pre-publication Data	13%
Post-publication Data	67%
Pre-publication Code	21%
Post-publication Code	74%

[17] A difference in proportions test applied to the difference in willingness to share prepublication code and data was insignificant, with p-value = 0.1929, indicating that there is no difference in willingness to share pre-publication code and data (21% and 13% respectively, as shows in Table 5: Research Component Sharing Rates (Survey Results)).

Unpublished research ideas, ie. through a blog	21%

From inspecting respondent websites, 30% of respondents shared some code and 20% shared some data on their own websites. The discrepancy between the self-reporting sharing rates and those I compiled is probably due to the fact that much data is available in public corpora and may not be directly linked to on the scientist's homepage. Similarly, code may be developed and housed in public repositories which may not be linked from the respondent's webpage. In one respondent's words, "All papers that I have published use corpora that can be licensed from other sources."[18] Of the shared code, half had no provision for licensing, and of the licensed half, the majority chose the GPL license[19] and about a quarter wrote their own license. The remaining respondents chose an established license other than the GPL.

Table 6: Sharing and Licensing (from Respondent's Website)

Characteristic	Count	Proportion
Code Sharing	40	0.30
No License	20	0.15
GPL	12	0.09
Own license	5	0.04
Other license	3	0.02
Data Sharing	27	0.20

If Hypothesis 1 holds, we would expect reasons to share and reasons not to share to reflect predominantly private incentives, with little regard for larger ideals. This is precisely not the case. The factors that influence computational scientists to share their code and data are those established as communitarian, but when scientists chose not to share their code and data, the decision is ruled by private incentives. "Table 8: Comparing Reasons or Sharing Code to Reasons Not to Share" and "Table 7: Comparing Reasons for Sharing Data to Reasons Not to Share" show the proportion of respondents that cited each factor as a reason to share or a reason not to share. For both code and data, reasons not to share a predominantly driven by private incentives (in italics) and reasons to share by communitarian norms (in bold). The ordering of the communitarian and non-communitarian factors is significantly different from random, at the 10% level, for sharing both code and data.[20]

[18] In this respondent's case, I did find a link from the webpage to data used in publication. In the biosciences, a general requirement is that all relevant data be made available at a publicly accessible website (usually an internationally coordinated database) at the time of a paper's publication (Cech T.R. 2003 *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*; available at www.nap.edu/books/0309088593/html).

[19] See http://www.gnu.org/copyleft/gpl.html (last accessed June 19, 2009).

[20] This significance level was obtained using the bootstrap to generate a null-distribution over all factor orderings.

Table 7: Comparing Reasons for Sharing Data to Reasons Not to Share[21]

	Share Data	Not Share Data
Encouraging scientific advancement	81.95%	0.75%
Being a good community member	80.45%	0.00%
Potential to encourage others to work on the problem	79.70%	2.26%
Encouraging sharing and having others share with you	78.95%	0.00%
The potential to set a standard for the field	77.44%	0.00%
Improvement in the caliber of research	74 44%	0.00%
Increase in publicity	73.68%	0.00%
Opportunity to get feedback on your work	73.68%	0.75%
Potential for finding collaborators	72.18%	0.00%
Normalizing understanding in a field	69.92%	0.00%
The topic is receiving a lot of attention	67.67%	3.76%
The number of requests you receive for the data	58.65%	3.01%
Conforming with requirements of the scientific method	58.65%	3.01%
Being first to release data in this area	56.39%	3.01%
Conforming with the usual practices of the research community	50.38%	6.77%
Potential to close a line of inquiry and move research to next step	48.87%	3.01%
Impact on future grant applications	48.12%	3.01%
Required by research funders	45.45%	1.52%
Requirement for publication	43.94%	3.03%
Whether you put in a large amount of work building the dataset	34.33%	24.63%
Availability of other data that might substitute for your own	24.63%	14.93%
Dealing with questions from users about the data	20.30%	33.83%
Whether there is intense competition in the topic	19.55%	24.81%
Legal barriers, such as copyright	10.47%	40.70%
The time it takes to clean up and document for release	8.27%	55.64%
Technical limitations, ie. webspace platform space constraints	6.02%	27.82%
Time to verify privacy or other administrative data concerns	6.02%	39.10%
The potential loss of future publications using this data	5.22%	35.82%
The possibility that your data may be used without citation	3.01%	43.61%
Competitors may get an advantage	2.24%	34.33%

[21] As in the case for code, the proportion of respondents citing each factors as an influential reason to share was significantly different from that citing it as a reason not to share, at the 5% level, with the exception of "Whether you put in a large amount of work building the dataset," "Whether there is intense competition in the topic" and "Availability of other data that might substitute for your own" ($p = 0.1079$).

Table 8: Comparing Reasons or Sharing Code to Reasons Not to Share[22]

	Share Code	Not Share Code
Encouraging scientific advancement	91.11%	0.00%
Encouraging sharing and having others share with you	89.63%	0.00%
Being a good community member	86.67%	0.00%
Increase in publicity	85.19%	0.74%
Improvement in the caliber of research	84.44%	0.00%
The potential to set a standard for the field	82.22%	1.48%
Potential to encourage others to work on the problem	81.48%	1.48%
Opportunity to get feedback on your work	77.78%	0.74%
Potential for finding collaborators	71.85%	1.48%
The topic is receiving a lot of attention	71.11%	0.74%
The number of requests you receive for the code	67.41%	4.44%
Normalizing understanding in a field	66.67%	0.00%
Being first to release code in this area	64.44%	0.00%
Conforming with requirements of the scientific method	61.48%	1.48%
Conforming with the usual practices of the research community	58.52%	1.48%
Impact on future grant applications	51.11%	4.44%
Potential to close a line of inquiry and move research to next step	46.67%	0.74%
Whether you put in a large amount of work building the code	44.44%	20.00%
Requirement for publication	44.03%	2.24%
Required by research funders	43.28%	1.49%
Whether there is intense competition in the topic	28.89%	17.04%
Availability of other code that might substitute for your own	25.37%	21.64%
Dealing with questions from users about the code	18.52%	51.85%
The code might be used in commercial applications	16.30%	28.15%
The time it takes to clean up and document for release	8.89%	77.78%
The possibility of patents or other IP constraints	8.15%	40.00%
Legal barriers, such as copyright	5.81%	33.72%
The potential loss of future publications using this code	5.19%	31.11%
The possibility that your code may be used without citation	3.73%	44.78%
Technical limitations, ie. webspace platform space constraints	3.70%	20.00%
Competitors may get an advantage	2.22%	31.85%

This provides evidence that Hypothesis 1 is violated: computational scientists' reasons not to share their code and data do seem to arise from private concerns, but their reasons to share are grounded in a communitarian sense. "Dealing with questions from users about the code" is the second most cited reason for now sharing, after the time it takes to

[22] For each factor, the proportion citing it as an influence to share was significantly different from the proportion citing it as an influence not to share at the 5% level, as measured by a difference in proportions test., with the exception of "Availability of other code that might substitute for your own."

clean up and document.[23] It seems appropriate to label this the "Newton Effect" for its parallel in science history: After Newton published his first scientific paper in *Philosophical Transactions*[24] he was so inundated with questions from members of the Royal Society, it also became his last journal article.[25] (Willinsky 2005, chapter 13, p198).

Table 9: Top Reasons Not to Share Data

	Not Share
The time it takes to clean up and document for release	55.64%
The possibility that your data may be used without citation	43.61%
Legal barriers, such as copyright	40.70%
Time to verify privacy or other administrative data concerns	39.10%
The potential loss of future publications using this data	35.82%
Competitors may get an advantage	34.33%
Dealing with questions from users about the data	33.83%
Technical limitations, ie. webspace platform space constraints	27.82%
Whether there is intense competition in the topic	24.81%
Whether you put in a large amount of work building the dataset	24.63%
Availability of other data that might substitute for your own	14.93%

Table 10: Top Reasons Not to Share Code

	Not Share
The time it takes to clean up and document for release	77.78%
Dealing with questions from users about the code	51.85%
The possibility that your code may be used without citation	44.78%
The possibility of patents or other IP constraints	40.00%
Legal barriers, such as copyright	33.72%
Competitors may get an advantage	31.85%
The potential loss of future publications using this code	31.11%
The code might be used in commercial applications	28.15%
Availability of other code that might substitute for your own	21.64%
Whether you put in a large amount of work building the code	20.00%
Technical limitations, ie. webspace platform space constraints	20.00%

[23] "Dealing with questions from users about the data" is the seventh most highly cited reason not to share data.

[24] This was also the first substantive scientific paper published in the *Transactions*, Thomas Kuhn (1978, 27) cited in Willinsky p 200.

[25] "the publication of this letter proved to be a more open and immediate forum for his work than Newton was willing to bear, and he did not again use the journal to publish his experimental pursuits but relied exclusively on the unhurried book, most notably with the *Principia*, published fifteen years later in 1687." Willinsky p200. This firestorm of exchange seems to have lead to the creation of the blind review process.

Whether there is intense competition in the topic	17.04%

"Table 9: Top Reasons Not to Share Data" and "Table 10: Top Reasons Not to Share Code" show the top reasons why respondents hold back on releasing their code and data. For both, the far and away biggest reason is the time it takes to prepare each for public release. As one respondent comments, "Not wanting to clean up, document and support the code is an incredibly strong influence not to reveal. It eclipses the other pros and cons so much that you can consider my answer on that line to be on a log scale." This is interesting because it speaks to an incentive misalignment in the reward structure for scientific research. There are many aspects of research which are tedious and time consuming, yet they get done when the expectations and reward structures are in place. This suggests a strong need to account for code and data release directly in the research review process. It seems change could occur on many levels, from policy set by University presidents, grant requirements, to validation in peer review and reward in promotion and award selection committees. Fear of use without proper citation ranked second for data and third for code. The "Newton effect" ranked second for code and seventh for data, implying data that is released tends to be perceived as reusable by others, more so than code. This implies an interesting phenomena, that perhaps researchers largely write their code in house and are more likely to use others' data. As one respondent noted,

> Unlike data, people seem eager to write code. What amazes me is how many people write code that already exists. I have written nearly a hundred software tools in my field. All of the software is professionally written, carefully engineered for portability, with copious internal documentation and superb complete user documentation. Yet the tools are *underutilized*. Instead, people re-write the same functions over and over with bad documentation, bugs, etc. I conclude that writing software is more fun than collecting data, and students would much rather write software from scratch rather than spend a little time learning to use existing research software. Ironically, people seem to hate collecting data. There is an asymmetry here: people seem more eager to share software than data, and they enjoy writing software more than collecting data.

It is also interesting to note the block to data and code sharing that the intellectual property structure instills is very real: ranking as the third most influential reason not to share data and the fourth and fifth most influential reasons not to share code. In fact, the use of the phrase copyright is something of a misnomer in describing the IP structure for raw facts, that are not copyrightable.[26] Unfortunately this factor is ambiguous. Scientists may find copyright a real block to sharing data, which is a misunderstanding of copyright law, or they may simply be stating that there are real legal blocks to data sharing other than copyright. Verifying administrative restrictions on the data, such as privacy, ranks

[26] See Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991).

under legal restrictions. For both code and data, concern over the loss of future publications and giving competitors an advantage influenced approximately the same proportion of respondents – about 35% for data and 31% for code. What is interesting to note is that although concern over loss of a publication stream to others is an influential reason not to share, it pales when compared to the top reason, time for preparation for release. Researchers are simply much more concerned about spending their time preparing the auxiliary components of their work than they are about being scooped. Intense competition and the amount of work put into building the data were also important factors. That computational researchers recognize a "sweat of the brow" property right in data was also evident in the interviews. The availability of similar code stopped scientists from releasing their own, as it did with data, but to a lesser extent.

Actual code and data sharing behavior was obtained from direct inspection of each respondent's website. If any code or data was posted on the respondent's website, he or she was classified as "open," and if no code or data was posted the respondent was classified as "closed." This gave 44 open respondents and 90 closed, or 33% open and 67% closed. A full 81% and 84% of respondents self-reported revealing a nonzero amount of code and data on the web, respectively. The discrepancy can perhaps be explained in that I only examined the respondent's webpage for code and data, and if he or she was revealing through a co-author's page, a central repository, or a lab page for example, these components of the research would not be included in my count. Using a central repository or lab webpage to share data and code is not uncommon, especially in biological sciences.

Several respondents went to great lengths to share code and data on their websites. Of these top sharers, typically code packages are introduced on a separate page within their website and tend to be more general than an association with a single paper. Links to data files may be with the link to the published paper, or on a separate page within the site. These top sharers generally reported, similarly to the entire sample, that time to prepare the code and data for release was a negative factor in their decision to share. The very top sharers tended to have webpages that included information on their scientific philosophy, possibly even about the sharing of code and data and reproducibility in computational science. One top sharer used http://www.github.com to track code changes and share code publicly.

Since respondents can be divided into open and closed groups, analysis on the importance of factors underlying each group's sharing decision can be carried out. Examining the reasons both sharers and nonsharers choose not to reveal their work is instructive: respondents who do not reveal their code and data cite concerns about attribution and tracking subsequent data use. In this case, it is not unexpected that respondents cite primarily private incentives are influencing them, since the data in the tables is for the decision not to share.

Between the Open and Closed groups, significant differences in factor influence were found in concerns about dataset citation and use tracking. Those who do not reveal data or code on their website appear to have a different perception of the risks of not being given appropriate credit than those who do.[27]

[27] See http://www.stanford.edu/~vcs/Survey2009/CodeAppendix.txt for code used in generating these p-values.

Table 11: Most Influential Reasons Not to Share Data, by Non-sharer and Sharer

	Closed	Open	p-value for diff
The time it takes to document for release	57.95%	52.38%	0.6818
The possibility that your dataset may be used without citation	50.00%	28.57%	**0.0342**
Legal barriers, such as copyright	42.37%	40.00%	1.0000
The potential loss of future publications using these data	39.33%	30.95%	0.4629
Dealing with questions from users about the data	38.64%	26.19%	0.2310
The time it takes to verify privacy or other admin data concerns	38.20%	41.46%	0.8724
Competitors may get an advantage	37.08%	30.95%	0.6245
The web doesn't allow me to track others use of the data	30.68%	14.28%	**0.0729**
Technical limitations, ie. webspace platform space constraints	29.54%	26.19%	0.8504
Whether there is intense competition in the topic	29.55%	16.67%	0.1731
Whether you put in a large amount of work building the dataset	24.72%	26.19%	1.0000
Availability of other data that might substitute for your own	12.36%	19.05%	0.4540
	<10%	<10%	

Table 12: Most Influential Reasons Not to Share Code, by Non-Sharer and Sharer

	Closed	Open	p-value for diff
The time it takes to clean up and document for release	82.22%	71.43%	0.2363
Dealing with questions from users about the code	54.44%	47.62%	0.5863
The possibility that your code may be used without citation	47.19%	37.71%	0.2964
The possibility of patents, or other IP constraints	38.89%	40.48%	1.0000
Competitors may get an advantage	34.44%	23.81%	0.3040
The potential loss of future publications using this code	31.11%	28.57%	0.9264
The code might be used in commercial applications	28.89%	23.81%	0.6888
Legal barriers, such as copyright	28.81%	44.00%	0.2727
The web doesn't allow me to track others use of the code	26.67%	14.29%	0.1745
Technical limitations, ie. webspace platform space constraints	23.33%	14.29%	0.3327
Availability of other code that might substitute for your own	22.22%	17.07%	0.6580
Whether you put in a large amount of work building the code	22.22%	14.29%	0.4049
Whether there is intense competition in the topic	15.56%	21.43%	0.5604
	<10%	<10%	

One question on the survey asked respondents to estimate the number of papers, if any, they feel they have had scooped. "Table 13: Idea Scooping (reported in the Survey)" gives the results: a majority of scientists who answered the survey felt they had had at least one paper scooped.

Table 13: Idea Scooping (reported in the Survey)

Idea Scooping	Count	Proportion
At least one publication scooped	53	0.51
2 or more scooped	31	0.30

No ideas stolen	50	0.49

This points to a seemingly pervasive problem in sharing of code and data: the perceived risk of losing subsequent publications and inadequate citation.

Hypothesis 2 can be evaluated by measuring the strength of influence of both the desire for community membership and the desire for feedback. Three factors address this: "Opportunity to get feedback on your work," "Being a good community member," and "Encouraging sharing and having others share with you."

Table 14: Hypothesis 2

	Share Data	Not Share Data	Share Code	Not Share Code
Being a good community member	80.45%	0.00%	86.67%	0.00%
Encouraging sharing and having others share with you	78.95%	0.00%	89.63%	0.00%
Opportunity to get feedback on your work	73.68%	0.75%	77.78%	0.74%

These three factors are significantly in their effect on the decision to share compared to the decision not to share, for both code and data, and the influence of all three is significantly different from zero. The real question seems to be whether these are the primary reasons for sharing compared to the other factors. Once again, I can use the bootstrap to find the likelihood of observing this particular ranking of factors if we expect all factors to be equally distributed. The high ranking we do observe is significant at the 12% level for both code and data, providing support that these three factors are unusually highly ranked as observed in the survey response. It does appear that "Opportunity to get feedback on your work," "Being a good community member," and "Encouraging sharing and having others share with you" are predominantly important factors underlying the decision to share code and data, supporting Hypothesis 2.

Using the previous classification respondents as open or closed (whether any code or data appears on the respondent's website) I used a regression model to understand the predictive ability of the covariates. Using University Ranking, Department, whether the respondent is in the Life Science, Gender, and Position a logistic model was fit. University Ranking is on a 0 to 4 scale, based on the U.S. News and World Reports 2008 Ranking as described earlier, binary variables for gender (1 for female, 0 for male) and life sciences (1 for life science, 0 otherwise), and Position is on a 0 to 5 scale from Master's Student to Full Professor. Department indicates whether the respondent is in a more computational department (CS, EE, physics, statistics, computational biology, medical imaging, bioinformatics, psychology, organizational research) or a less computational department (music, philosophy, medicine, biology, mathematics). "Department" is 1 for less computational and 0 for the more computational.

"Table 15: Logistic Regression Results" gives the parameter estimates for the fitted logistic model. Position is a significant predictor of sharing of code and data, with a higher position indicating a higher likelihood the respondent will be open. Whether the respondent is in a more computational department indicates he or she is more likely to be open. Gender, working in the life sciences, and university ranking do not correlate with propensity to be open.

Table 15: Logistic Regression Results

	Estimate	P-value
(Intercept)	-1.21607	0.0099
University	-0.04654	0.7018
Department	-1.33245	**0.1089**
LifeSciences	-0.07585	0.8700
Position	0.29609	**0.0373**
Gender	0.29625	0.6435

The significance at the 10% level of the level of computational of respondent's department may simply be an artifact of those in more computational departments having more data and code to share. The significance of the Position variable supports Polanyi and von Krogh's postulation of scientists as a community of revealers. If their assertion is so, we would expect the more established within the community to be those more likely to reveal. These results show the propensity to reveal, among the machine learning community, increases significantly with seniority of status. Predicting only whether or not some code was shared increased the significance level of Position to a p-value of 0.0258 and a parameter estimate of 0.3294, and the department's level of computation increased significance to a p-value of 0.0608 with a parameter estimate of -2.0463. These results strengthen the conclusions of the overall regression in predicting openness generally, since the signs of the coefficient estimates are the same and the magnitude has increased.

D. Design Shortcomings

Statistical analysis of the survey responses shows that, contrary to previous literature, scientists' reasons for sharing their code and data are largely communitarian in nature, whereas their concerns when not sharing reflect more immediate personal consequences. But how are we to know that this isn't simply scientists saying what they feel is the appropriate thing to say, in line with the well-known Mertonian norms, rather than what they really think? It is well known in survey analysis that asking people about behavior that is seen as against norm, can induce the respondent to be reluctant to admit such behavior. There is no way to know definitively that these survey responses reflect the true considerations computational scientists take into account when making a decision but several safeguards point to the veracity of respondents. First, the survey asked about a particular decision they had taken, the sharing made with respect to their last NIPS publication, and if that paper was not computational, then their last published computational paper. Giving a specific reference point, consistent across all respondents, maximized not only comparability in respondents' answers but the most likelihood to penetrate to the rationale behind the actual sharing behavior. (Tversky and Kahneman, 1981). Second, there is evidence that suggests scientists are not giving their behavior a Mertonian gloss: the revealers are claiming communitarian norms, rather than Mertonian norms, as their basis for revealing. Third, the survey itself made every effort not to communicate a bias toward Mertonian norms. Over 80% of respondents indicated they did share code or data, indicating, if true, their behavior is largely in line with Mertonian norms before being questioned. When the analysis was reduced to two groups, those whom I found evidence of sharing or not sharing on their website, the same results held.

As Christine Borgman states in her recent book, the very neatness of Mertonian norms tends "to oversimplify the mechanisms of scholarship. Much of the subsequent social studies of science literature takes a more constructionist perspective, finding scientists may state norms as a convenient shorthand to explain themselves, but in fact, their practices are local and vary widely." (2007: p37).

That fully 81% and 84% of respondents indicated they shared code or data raises concern about respondent bias. A number of respondents prefaced their response with a statement indicating their interest in this topic, and frequently request a copy of the paper that results from this work.[28] This type of indicates that researchers sympathetic to openness and reproducibility, a priori, are more likely to reply. Of final survey respondents who answered the question, 89% requested I send them a copy of the final paper resulting from this research. Having presented at NIPS myself in 2003, I received a small number of replies to the pilots from researchers who remembered my talk or remembered meeting me. This may improve response rates in that researchers might be more likely to respond to someone they recognize as in the Machine Learning community, and so long as their mental association is not with reproducibility (my NIPS presentation was unrelated (Stodden 2003)), this should not introduce a bias.

DISCUSSION

This paper does two things. It provides evidence in the debate about whether scientists' research revealing behavior is wholly governed by considerations of personal impact or whether the reasoning behind the revealing decision involves larger scientific ideals, and secondly, this research describes the actual sharing behavior in the Machine Learning community. As measured by my analysis of code and data posting on respondents' webpages, about 30% of computational scientists in Machine Learning post some code or some data on their own site.[29] Only a handful of respondents made comprehensive posting of code and data used in their work readily and easily available through their website. David's aforementioned "trials of verification" are not systematically taking place. (2003: p3).

In the survey scientists indicate that the closing of black boxes as a method of resolving disputes is not a compelling reason not to share as only 3% or respondents cited it as a reason not to share data, and 1% for code. If anything, it appears that transparency in research is actually viewed as way to settle scientific disputes, with 49% of respondents saying the "Potential to close a line of inquiry and move research to next step" influenced them to share their data, and 47% indicating the same for code, although

[28] The two pilots were communicated via email with no link to a webform, so sending an email back to me was the only way to answer the survey questions. When I gave recipients the option of using a webform on the final survey, without exception every respondent opted to use the webform, rather than email me their response, and the personalized reply dropped to a couple.

[29] As mentioned previously, this percentage doesn't include data or code deposited in repositories not linked to from the respondent's website. Nor does it include publicly available data corpora or repositories associated with journals that may be well referenced in the text of the published paper. Because of this, the 30% figure must be considered a lower bound for the general level of revealing of code and data, although it is an appropriate measure for final survey respondents' revealing of code and data publicly on their website.

this factor did not rank particularly highly (16th most influential for data of 34, and 17th of 35 factors for code) in influence on the decision to reveal.

Generalized reciprocity is highly cited as a reason to share code and data (2nd and 4th ranked for code and data respectively). As a respondent stated, "what goes around comes around." This finding lends support to the generalization of Ekeh and Bouty's work in free revealing in industry to the scientific context. Another respondent notes, "Much of my research would have not been possible if other people had not released their data or code. This is one of the main reasons for which I also want to contribute by releasing my own data and code" and another respondent cites knowing that the effort to release can save time for others: "Once you've spent time solving a problem, it softens the pain a little to think that others can benefit from it."

The potential to set a standard in a research area ranked 5th for data and 6th for code. This suggests the generalization of similar results found by von Hippel and Harhoff in the setting with free revealing of proprietary information by firms. von Hippel (1986) models the sharing and trading of proprietary information by rival (and non-rival) firms as a prisoner's dilemma. He finds underlying conditions that determine when participating in this information trading game makes economic sense. Although von Hippel describes a situation in which engineers within firms trade information between each other rather than making the information publicly available in published form, and I am studying for the case in computational scientific research revealed publicly on the web, my results are supportive of his in two important ways.[30] von Hippel (1986:11) observes situations when firms in direct competition nonetheless engage in information sharing. Two conditions exist for know-how trading by engineers: that the traded information is not vital to the firm, and that is could be independently developed by other firms if necessarily. These factors imply the highest competitive advantage information is not traded. This is supported in the research on sharing between computational scientists, in that scientists do not seem particularly worried about the intensity of competition when they chose not to share their code and data. The factor on the survey "Whether there is intense competition in the topic" ranked as the 8th highest reason not to share data and the 12th for code (25% of respondents for data and 17% for code). Maintaining a competitive advantage is cited at the sixth highest reason for not sharing code and data (32% and 34% of respondents cited it as a reason for not sharing code or data, respectively). von Hippel bases his model of personal proprietary information trading on the notion of reciprocity. As pointed out in von Hippel's paper, "Collins (1982) has shown that scientists employed by non-profit laboratories (university and governmental) selectively revealed data to colleagues interested in know-how related to the "TEA laser"." Collins notes that individuals and laboratories make conscious and careful discriminations as to what know-how would be revealed to what recipient, and noted also that "'"[n]early every laboratory expressed a preference for giving information only to those who had something to return."(Collins 1982:59)" (von Hippel 1986:27). The

[30] Von Hippel's paper also investigates whether the know-how is in fact valuable. I am assuming throughout this study that code and data are valuable to other researchers. It seems clear this is the case for data and at least some code (e.g. WaveLab and SparseLab), but as a survey respondent pointed out researcher seem to like to recode anew rather than adapt preexisting code. Investigating this assumption is an area for future research: to what extent is science remixed? Regardless, scientific principle of reproducibility and verification demand openness in data and methods.

conditions von Hippel gives for a prisoner's dilemma exist in the scientific context as well: there is a temptation to "defect" each round (ie. not share code and data upon publication) and the gains are maximized if each researcher shares his or her code with every publication.

A number of factors were included in the survey to ascertain the constraints inhibiting sharing that scientists perceive due to the Intellectual Property framework. When citing reasons not to share data and code, the top of the list of factors preventing the sharing of data was time to prepare for release (cited by 56% percent of respondents) and concern over lack of citation in downstream use (44%), then legal barriers such as copyright (41%). The story is similar for code: the top reason for not sharing is time to prepare (78%), then dealing with questions from users (52%), concerns over use without citation (45%), and third the possibility of patents or other IP constraints (40%) and legal barriers such as copyright (32%). This indicated the IP framework is having a deleterious effect on reproducibility in computational science.

CONCLUSION

The traditional mechanism for the communication of new scientific discoveries has been through publishing a research paper in a journal. But the Internet has changed this: not only are papers available to researchers much more quickly, but scientists are not limited to sharing only the aspects of their work that lend themselves to a journal's format. Now scientists routinely share preprints, published papers, and other forms of traditional scientific knowledge transmission mechanisms, but they also share entirely new forms such as datasets, code, high resolution images, software designed to entail the manipulation of results by others, links and lists of related works. This facility can make the black boxes significantly easier to open.

Scholars can concerned about the creation of derivative works from their writing, specifically by someone who uses large chunks of their work to recontextualize and distort their ideas. I don't think this is so much of a problem in an objective field where code verifies the work, but it might be worth asking.

Keeping a secret can be costly. Innovators may choose not to protect their product, through either patenting, trade secrets, or trademarking, because the cost of doing so outweighs the potential profit (Harhoff et al. 2003). In the sciences it is not typical to reveal work before publication implying a need for a window of protection while ideas develop in preparation for public exposure. Patent law provides a window of protection to inventors in exchange for disclosure, and so is structured to create incentives both for knowledge to be revealed and the gains reaped by inventors, thus internalizing the benefits accruing to research and development. A patent grants a limited term exclusive right to the invention to the inventor, in exchange for openly revealing the knowledge. Landes and Posner (2003) give an incentives explanation for the rationale behind patent law.: As costs of inventing around and defending an infringement lawsuit increase, the patent becomes more valuable and thus the inventor becomes more likely to seek a patent.

There is an analogy to the production of scientific research: just as inventors are thought to require certain limited monopoly rights to encourage their inventing behavior, so scientists may find a window of proprietary rights over their work encourages further

scientific discoveries. It is plausible that with the Internet's facilitation of rendering Latour's black boxes transparent, scientists may seek other ways to keep them closed and thereby maintain their lead over other scientists. This analogy ends with the incentives to invent around and push the limits of noninfringement. In scientific research the idea must be attributed but there is little need on the part of the discoverer for exclusive use of the idea once it has been attributed to him or her, in fact this is counter to the Mertonian norms described previously. Note that there is no need for a limited term, as in the case of patents, because academic scientific research is intended to be open and usable by its very nature, at least by other researchers. Scientific research is quality filtered not by a central office, as in the case of patents, but by a diffuse network of journals and reputation mechanisms. Of course fundamental, usually scientific, ideas cannot be patented but even if the analogy to patents is not apt, researchers can reveal their work in such a way that it is difficult for fellow researchers to follow. Such obfuscation does occur in patent applications: the invention is revealed but there is an incentive for the inventor to reveal as little as possible, making it harder for his or her work to be used, thus collecting more return on the investment and incurring less risk of defending the patent. Because a similar incentive mechanism is operating – scientists may wish to conceal the details of their work from competitors, yet still appear superficially to be revealing – policy makers must use a delicate hand in mandating revealing in scientific research.

There is a vibrant culture around the sharing of artistic works, especially digital ones such as images, music and film. The Free Culture movement seeks to facilitate this sharing and mobilize artistic creation. This view of cultural creation, based at core on notions of remix, can be extended to include the process of scientific discovery and its addition to culture, thereby fleshing out the appropriate of the Free Culture arguments for scientific enterprise.

This work could also be verified in a second direction. Both Polanyi and von Krogh envision science as a community of revealers seeking feedback from peers. Thus we would expect free revealing scientists to be highly engaged members of their respective subfield. It would be interesting to test this hypothesis – preliminary investigation in this study reveals that more senior and engaged members do tend to reveal their work more, but controls are needed to understand whether this is due to years of experience, number of research assistants or other factors, if the effect is even real at all.

Computational science is a special case of scientific research: the work is easily shared via the Internet since the paper, code, and data are digital and those three aspects are all that is required to reproduce the results, given sufficient computation tools. It may be the case that the theory of scientist's sharing incentives applies differently in non-computational areas where tools and materials are less easily shared. A natural next step would be to examine the sharing behavior of other materials used in scientific research, for example wet lab materials, animals, and specimens (Murray 2009).

BIBLIOGRAPHY

Allen, R., "Collective Invention," Journal of Economic Behavior and Organization, 4(1), 1-24, 1983.

Antelman, K., "Do Open-Access Articles Have a Greater Research Impact?" College and Research Libraries 65: 372-382, 2004.

Barnes, B. and D. Bloor, "Relativism, Rationalism and the Sociology of Knowledge", in M. Hollis and S. Lukes (eds.), Rationality and Relativism, Oxford: Blackwell, 21-47, 1982.

Birney, et al., "Prepublication Data Sharing", with the Toronto International Data Release Workshop Authors, Nature, Vol 461, Issue 10, September 2009, p. 168-70, 2009.

Bouty, I., Interpersonal and Interaction Influences on Informal Resource Exchanges Between R&D Researchers Across Organizational Boundaries, Academy of Management Journal, Vol 43, No.1, pp 50-65, 2000.

Bourdieu, P., Science of Science and Reflexivity, University of Chicago Press, 2004.

Collins, H., Changing Order, University of Chicago Press, 1992.

Collins, H., "Tacit Knowledge and Scientific Networks," in Science in Context, edited by Barry Barnes and David Edge, Cambridge, MA: MIT Press, 1982.

Collins, H. and Pinch, T., The Golem: What Everyone Should Know About Science. Cambridge University Press, 1998.

Dasgupta, P. and David, P., Towards a New Economics of Science. Research Policy, 23, 5, 487–521, 1994.

David, P., "Knowledge, Property, and the System Dynamics of Technological Change," Proceedings of the World Bank Annual Conference on Development Economics, 215–247, 1992.

David, P., "Knowledge spillovers, technology transfers, and the economic rationale for public support of exploratory research in science," Background Paper for the European Committee for Future Accelerators, 1998.

David P., "The Economic Logic of 'Open Science' and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer" SIEPR Discussion Paper No. 02-30, 2003.

Donoho, D., "How to be a Highly Cited Author in the Mathematical Sciences," In-cites, March 2002. http://www.in-cites.com/scientists/DrDavidDonoho.html

Donoho, D. and Stodden, V., "Breakdown Point of Model Selection when the Number of Variables Exceeds the Number of Observations," Proceedings of the International Joint Conference on Neural Networks, 2006.

Donoho, D. et al., "15 Years of Reproducible Research in Computational Harmonic Analysis," IEEE Computing in Science and Engineering, 11(1), p.8-18, January 2009.

Ekeh, P., Social Exchange Theory: The Two Traditions, Harvard University Press, 1974.

Good, I., "How much science can you have at your fingertips?" IBM J. Res. Develop. 2, 282–288, 1958.

Hagstrom W., The Scientific Community. New York: Basic Books, 1965.

Harhoff, D., J. Henkel, E. von Hippel, "Profiting from voluntary information spillovers: How users benefit by freely revealing their innovations," Research Policy, 32(10), 1753-1769, 2003.

Lakhani, K., Wolf, R., "Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects" in Perspectives on Free and Open Source Software, edited by J. Feller, B. Fitzgerald, S. Hissam, and K. R. Lakhani (MIT Press), 2005.

Lakhani, K. and von Hippel, E. "How Open Source Works: 'Free' User-to-User Assistance," Research Policy, 32(6), p923-943, 2003.

Landes, W., and Posner, R., The Economic Structure of Intellectual Property Law, Belknap Press, 2003.

Latour, B., Science in Action, Harvard University Press, 1987.

Latour , B., and Woolgar, S., Laboratory Life: The Social Construction of Scientific Facts, Sage Publications, London, 1979.

Kanigal, R., The Man Who Knew Infinity: A Life of the Genius Ramanujan, Washington Square Press, 1991.

Kuhn, T., The Structure of Scientific Revolutions, University of Chicago Press, 1962.

Merton, R. K. "The Unanticipated Consequences of Purposive Social Action," American Sociological Review, 1(6), 894-904, 1936.

Merton, R.K. "The Normative Structure of Science". In: Merton, R. K., The Sociology of Science: Theoretical and Empirical Investigations, Chicago, IL: University Of Chicago Press, 1942.

Merton, R. K., Social Theory and Social Structure, New York: Free Press, 1968

Merton, R. K., The Sociology of Science: Theoretical and Empirical Investigations, Chicago, University of Chicago Press, 1973.

Murray, F., "The OncoMouse that Roared: Hybrid Exchange Strategies as a Source of Productive Tension at the Boundary of Overlapping Institutions." American Journal of Sociology, forthcoming.

Nuvolari, A., "Collective invention during the British Industrial Revolution: The case of the Cornish Pumping Engine," Cambridge Journal of Economics, 28(3), 347–363, 2004.

Nuvolari, A., "Open Source Software Development: Some Historical Perspectives," First Monday, 10(10), 3 October 2005.

Polanyi, M., Science, Faith, and Society, University of Chicago Press, 1946.

Polanyi, M., The Tacit Dimension, Doubleday & Co., 1966.

Polanyi, M., The Republic of Science, Roosevelt University, 1962.

Polanyi, M., The Logic of Liberty, The University of Chicago Press, 1951.

Stodden, V., "Enabling Reproducible Research: Open Licensing for Scientific Innovation," International Journal of Communication Law and Policy, Issue 3, Winter 2009.

Stodden, V. "Model Selection When the Number of Variables Exceeds the Number of Observations," Doctoral Dissertation, Department of Statistics, Stanford University, 2006.

Sulston, J. and Ferry G., The Common Thread, Corgi Books, 2003.

Tversky, A. and Kahneman, D., "The Framing of Decisions and the Psychology of Choice," Science, New Series, Vol. 211, No. 4481. (Jan. 30, 1981), pp. 453-458.

von Hippel, E., "Cooperation Between Rivals: Informal Know-How Trading," Research Policy, 16: 291-302, 1987.

von Hippel, E., and G. von Krogh, "Free Revealing and the Private-Collective Model for Innovation Incentives," R&D Management, 36(3), 291-302, 2006.

von Hippel, E., Democratizing Innovation, MIT Press, 2005.

Willinsky, John. The Access Principle: The Case for Open Access to Research and Scholarship, MIT Press 2005.