17: Innovation and Growth through Open Access to Scientific Research: Three Ideas for High-Impact Rule Changes

Victoria Stodden*

Our stock of scientific knowledge is now accumulating in digital form. Our DNA is now encoded as genome sequence data, scans of brain activity exist in functional magnetic resonance image datasets, and records of our climate are stored in myriad time series datasets—to name a few examples. Equally as important, our reasoning about these data is recorded in software, in the scripts and code that analyze the digitally recorded world. The result is a deep digitization of scientific knowledge, spreading across fields and generating new ways of understanding our surroundings. With the parallel development of the Internet as a pervasive communication mechanism for digital data, an unprecedented opportunity for access to society's scientific understanding is at hand.

At present, the notion of unmitigated access to scientific knowledge largely remains an unrealized opportunity. This paper pro-

* Victoria Stodden is Assistant Professor of Statistics at Columbia University, completing both her PhD in statistics in 2006 and her law degree in 2007 at Stanford University. Her current research focuses on how pervasive and large-scale computation is changing our practice of the scientific method; reproducibility of computational results; understanding factors underlying code and data sharing among researchers; and the role of legal framing for scientific advancement.

poses three changes to our current regulatory system designed to take into account the new reality of scientific innovation in a digital world and thereby promote innovation and economic growth. Our current intellectual property framework developed with a view to protecting original expressions of ideas in established media, such as literature, film, and sound recordings. Scientific innovations, such as code written to implement a new algorithm or an image produced for academic journal publication, now fall within a copyright structure that was developed for an entirely different normative environment, and the result is the creation of barriers to scientific innovation.

Part I of this essay explores the mismatch of intellectual property laws with scientific norms regarding the treatment of ideas, and proposes an alternative structure designed to facilitate deep sharing of scientific innovation through open code and data, thereby realigning the legal environment with long-standing scientific norms.

Part II addresses the role of federal agency policy in funding scientific research. Federal funding agencies create incentives for openness through their grant guidelines and enforcement. Changes to both accommodate the impact of computation on reproducibility, and therefore openness, are suggested.

Part III proposes ideas for the facilitation of code and data sharing through a process of disentanglement of ownership rights and the establishment of sharing protocols. Scientific research is becoming increasingly collaborative, particularly with industry researchers, and without a clear understanding of ownership rights in data and code, open sharing is hampered, if not obstructed completely. Methods for streamlining this process at the university level to permit the disclosure of the underlying code and data at the time of publication are presented.

The case for openness in science is not a new one. Scientific research is predicated on an understanding of scientific knowledge as a public good—this is the rationale underlying today's multibillion-dollar subsidies of scientific research through various federal and state agencies. The scientific view is not one of adding nuggets of truth to our collective understanding, but instead one of weighing evidence and assigning likelihoods to a finding's probability of being true. This creates a normative structure of skepticism among scientists: the burden is on the discovering scientist to convince others that what he or she has found is more likely to be correct than our previous understanding. The scientific method's central motivation is therefore the *ubiquity of error*—the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error. As a result, standards of scientific communication evolved to incorporate full disclosure of the methods and reasoning used to arrive at the result. Since the 1660s, the gold standard for scientific communication has been reproducibility, to create both the greatest chance of the accurate transmission of the new discoveries and also to maximize the likelihood that any errors in the reasoning would be identified.

Today, massive computation is transforming science, as researchers from numerous fields, even historically nontechnical ones, launch ambitious projects involving large-scale computations. A rapid transition is under way—visible particularly over the past two decades—that will finish with computation as absolutely central to scientific enterprise. From the newcomer's struggle to make even the simplest computer program run, to the seasoned professional's frustration when a server crashes in the middle of a large job, all is struggle against error. The understanding necessary for reproducibility is typically not transmitted, as computational results are frequently of a complexity that makes the effective explanation of the methodology all but impossible in a typical scientific publication today. To affect reproducibility, and the transfer of the knowledge embodied in the scientific finding, the code and data on which the result is derived must be communicated such that the result can be independently replicated and verified.

As a contribution to society's stock of knowledge, a scientific finding has the potential to be both developed and extended into commercial settings and to become the foundation for further scientific discoveries. Acceleration of innovation is facilitated by the incorporation of the open release of code and data in today's computational science practice. A number of changes are essential to catalyze both scientific advancement and the development of applications and discoveries outside academia.

Part I. The Legal Framework for Scientific Innovation and Dissemination: Copyright is a Barrier

The Intellectual Property Clause of the Constitution has been interpreted to confer two distinct powers: the first power provides the basis for copyright law—securing for a limited time a creator's exclusive right to their original work;[1] the second power provides the basis for patent law—giving inventors a limited-term exclusive right to their discoveries in exchange for disclosure of the invention. Authors do not have to apply for copyright protection, as it adheres automatically when the original expression of the idea is rendered in fixed form. Many perfectly standard scientific activities, such as writing a script to filter a dataset or fit a statistical model, will produce a copyrighted output, in this case the code written to implement these tasks. Building a new dataset through the original selection and arrangement of data will generate ownership rights through copyright for the dataset creator, to give another example.[2]

The default nature of copyright confers an intellectual property framework for scientific ideas at odds with long-standing scientific norms in two key ways.[3] First, by preventing copying of

[1] For a discussion of the Copyright Act of 1976 see e.g. Pam Samuelson, "Preliminary Thoughts on Copyright Reform Project," *Utah Law Review* 3 (2007): 551, accessed March 7, 2009, http://people.ischool.berkeley.edu/~pam/papers.html.

[2] See *Feist Publications Inc. v. Rural Telephone Services* Co., 499 U.S. 340 (1991), 363-364.

the research work, it creates a barrier to the possibility of legally reproducing and verifying another scientist's results without the need to obtain prior permission from the authoring scientist.[4] Second, copyright also establishes rights for the owner over the creation of derivative works. Scientific norms guide scientists to build on previous discoveries—using copyrighted work in derivative research typically requires obtaining the permission of the copyright holder, thus creating a block to the generation of new scientific discoveries. Particularly as computation becomes increasingly central to the scientific method, copyright on code and the potential for copyright in data are barriers to the advancement of science and economic growth. When scientists share their research on the Web, for example, the original expression of their ideas automatically falls under copyright.

Copyright law is often understood as a trade-off between providing incentives for the production of creative works by granting the author certain limited-term exclusive rights over their work, and the public's desire to access the work. By blocking the ability of others to copy and reuse research, copyright law acts counter to the prevailing scientific norms that encourage scientists to openly release their work to the community in exchange for citation.

An exception is made in our federal copyright code under fair use for "teaching (including multiple copies for classroom use), scholarship, or research,"[5] but this does not extend to the full research project. A relatively straightforward solution to the barrier copyright imposes would be to broaden the fair use exception to include scientific research that takes place in research institutions such as universities or via federal research grants.[6]

[3] For a detailed discussion of copyright law and its impact on scientific innovation, see Victoria Stodden, "Enabling Reproducible Research: Licensing for Scientific Innovation," *International Journal for Communications Law and Policy* 13 (Winter 2008-9), http://www.ijclp.net/issue_13.html.

[4] See Victoria Stodden, "The Legal Framework for Reproducible Scientific Research: Licensing and Copyright," *Computing in Science and Engineering* 11, no. 1 (January/February 2009): 35.

[5] U.S. 17 § 107.

Distinguishing legal fair use is not a clear exercise, and an extension to research more broadly may still not sufficiently clarify rights. A preferable step would be to include academic research, identified perhaps by federal funding, directly in the fair use exception.

Another mechanism for realigning intellectual property rights with scientific norms is the Reproducible Research Standard (RRS). The first component of this standard is the application of an appropriate license to remove restrictions on copying and reusing the scientific work, as well as adding an attribution requirement to elements of the research compendium.

Components of the research compendium have different features that necessitate different licensing approaches. Licensing is given strength through rights created by the underlying copyright law: if these licenses are found invalid by a court, the work will still be considered under copyright. Effectively, this means that even if a license fails to be recognized as a valid contract by a court, use of the work will remain subject to injunction and other remedies associated with copyright violation.[7]

With myriad options for licensing copyright-protected work, a principle for scientific licensing can guide choices:

***Principle of Scientific Licensing**: Legal encumbrances to the dissemination, sharing, use, and re-use of scientific research compendia should be minimized, and require a strong and compelling rationale before application.*[8]

The goal of an intellectual property legal framework for scientific research must be to increase what Benkler terms "that most precious of all public domains—our knowledge of the world that

[6] This idea was suggested in Paul A. David, "The Economic Logic of 'Open Science' and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer," Accessed January 12, 2009, http://ideas.repec.org/p/wpa/wuwpdc/0502006.html.

[7] This recourse to copyright for enforcement may not be necessary: a recent case (*Jacobsen v. Katzer*, 535 F.3d 1373 (Fed. Cir. 2008)) found a software license to be enforceable like a copyright condition for which courts can apply the remedy of injunction.

surrounds us."[9] This effort involves an alignment of the private incentives faced by a scientific researcher and the societal benefit of increasing our stock of public knowledge. Scientific norms have arisen to align these interests in practice, and an associated intellectual property structure should reflect these norms to allow scientific research to flourish.[10]

The Paper, Figures, and Other Media Files

For media components of scientific work, alignment with scientific norms is most readily and simply achievable through use of the Creative Commons attribution license (CC BY), which frees the work for replication and re-use, with the condition that attribution must accompany any downstream use of the work.

The Code

A plethora of licenses exist that allow authors to set conditions of use for their code. In scientific research, code can consist of scripts that are essentially stylized text files (such as MATLAB or R scripts) or the code can have both a compiled binary form and a source representation (such as code written in C). Use of the CC BY license for code is actively discouraged by Creative Commons.[11]

[8] A research *compendium* refers to the triple of research paper, and the code and data that underlies its results. See Robert Gentleman and Duncan Temple Lang, "Statistical Analyses and Reproducible Research" *Journal of Computational and Graphical Statistics* 16, no. 1 (2007): 1-23, http://www.bepress.com/bioconductor/paper2/.

[9] Yochai Benkler, "Constitutional Bounds of Database Protection: The Role of Judicial Review in the Creation and Definition of Private Rights in Information," *Berkeley Technology Law Journal* 15 (Fall 1999): 3, http://ssrn.com/abstract=214973.

[10] See Robert K. Merton, *The Sociology of Science* (Chicago: University of Chicago Press, 1973), for a description of the four scientific norms. Of particular interest to us is the "Communitarian" norm: that scientists relinquish ownership rights over their work in exchange for acknowledgement through citation or perhaps the naming of discoveries. This, in conjunction with the norm of "Skepticism" that establishes the close inspection and review of research work by the community, implies open access to scientific research, satisfying the interests of the larger community in the openness and availability of scientific research work. Paul David has made this observation in "The Economic Logic of 'Open Science'," 5.

The (Modified) Berkeley Software Distribution (BSD) license permits the downstream use, copying, and distribution of either unmodified or modified source code, as long as the license accompanies any distributed code and the previous authors' names are not used to promote modified downstream code.[12] The Modified BSD license is very similar to the MIT license, with the exception that the MIT license does not include a clause forbidding endorsement.[13]

The Apache 2.0 license is another common method for developers to specify terms of use of their work.[14] Like the Modified BSD and MIT licenses, the Apache license requires attribution. It differs from the previously discussed licenses in that it permits the exercise of patent rights that otherwise would extend only to the original licensor, meaning that a patent license is granted for those patents needed for use of the code. The license further stipulates that the right to use the work without patent infringement will be lost if the downstream user of the code sues the licensor for patent infringement. Attribution under Apache 2.0 requires that derivative works carry a copy of the license, with notice of any files modified. All copyright, trademark, and patent notices that pertain to the work must be included. Attribution can also be done in such a notice file.

Scientific Data

Collecting, cleaning, and otherwise preparing data for analysis is often a significant component of scientific research. Copyright law in the United States does not permit the copyrighting of "raw facts," but original products derived from those facts are copy-

[11] "[W]e do not recommend that you apply a Creative Commons license to software code," "FAQ." accessed January 5, 2009, http://wiki.creativecommons.org/FAQ.

[12] "Open Source Initiative OSI—The BDS License," accessed January 2, 2009, http://www.opensource.org/licenses/bsd-license.php.

[13] "Open Source Initiative OSI—The MIT License," accessed March 5, 2009, http://www.opensource.org/licenses/mit-license.php.

[14] "Apache License, Version 2.0," accessed January 1, 2009, http://www.apache.org/licenses/LICENSE-2.0.

rightable. In *Feist Publications, Inc. v. Rural Telephone Service*, the Supreme Court found that the white pages from telephone directories are not themselves directly copyrightable, since copyrightable works must have creative originality:[15]

> ...the copyright in a factual compilation is thin. Notwithstanding a valid copyright, a subsequent compiler remains free to use the facts contained in another's publication to aid in preparing a competing work, so long as the competing work does not feature the same selection and arrangement.[16]

Currently, the Court holds *original* "selection and arrangement" of databases protectable:[17] the component falling under copyright must be original in that "copyright protection extends only to those components of the work that are original to the author, not to the facts themselves...."[18] The extraction of facts from a database does not violate copyright. Attaching an attribution license to the original "selection and arrangement" of a database can encourage scientists to release the datasets they have created by providing a legal framework for attribution and re-use of the original selection and arrangement aspect of their work.[19] Since

[15] *Feist Publications Inc. v. Rural Telephone Service Co.*, 363-364.

[16] Ibid., 349. See also Miriam Bitton, "A New Outlook on the Economic Dimension of the Database Protection Debate" and Hongwei Zhu and Stuart E. Madnick, "One Size does not Fit All: Legal Protection for Non-Copyrightable Data" (working paper CISL# 2007-04), accessed January 4, 2009, http://web.mit.edu/smadnick/www/wp/2007-04.pdf.

[17] Miriam Bitton, "A New Outlook," 4.

[18] *Feist Publications, Inc. v. Rural Telephone Services Co.*, 340. The full quote reads "Although a compilation of facts may possess the requisite originality because the author typically chooses which facts to include, in what order to place them, and how to arrange the data so that readers may use them effectively, copyright protection extends only to those components of the work that are original to the author, not to the facts themselves... As a constitutional matter, copyright protects only those elements of a work that possess more than *de minimis* quantum of creativity. Rural's white pages, limited to basic subscriber information and arranged alphabetically, fall short of the mark. As a statutory matter, 17 U.S.C. sec. 101 does not afford protection from copying to a collection of facts that are selected, coordinated, and arranged in a way that utterly lacks originality. Given that some works must fail, we cannot imagine a more likely candidate. Indeed, were we to hold that Rural's white pages pass muster, it is hard to believe that any collection of facts could fail." For a discussion of the Constitutional limits on Congress's ability to create property rights in facts see Yochai Benkler, "Constitutional Bounds of Database Protection."

the raw facts themselves are not copyrightable, it does not make sense to apply such a license to the data themselves. The selection and arrangement may be implemented in code or described in a text file accompanying the dataset, either of which can be appropriately licensed.

Since the components of research compendia are varied, licenses should be applied as appropriate to each component in accordance with the Principle of Scientific Licensing. Using CC BY on the media components of the research, such as text and figures, permits other scientists to freely use and reuse this work provided the original author is attributed. The same result is obtained by using a software license that provides an attribution component for the code components, such as the Apache License 2.0, the Modified BSD License,[20] or the MIT License. The original selection and arrangement of data can be similarly licensed depending on whether it takes a code or text format. Since an attribution license cannot be attached to raw facts, data can be released to the public domain by marking with the CC0 standard.[21] A licensing structure that makes media, code, data, and data arrangements—the research compendium—available for re-use, in the public domain or with attribution, is termed the Reproducible Research Standard.

Part II. Government Funding Agency Policy Should Require Openness

Government funding agencies such as the National Institutes of Health (NIH), the National Science Foundation (NSF), and the

[19] See Anselm Kamperman Sanders, "Limits to database protection: Fair use and scientific research exemptions," *Research Policy* 35, no. 6 (July 2006): 854-874 (859 for a discussion of the international and WIPO statements of the legal status of databases).

[20] Creative Commons provides the BSD as a CC license, accessed March 5, 2009, See http://creativecommons.org/licenses/BSD/.

[21] For details on the CC0 protocol, see Creative Commons, "Creative Commons Launches CC0 and CC+ Programs," news release, December 17, 2007, http://creativecommons.org/press-releases/entry/7919 (accessed February 12, 2009).

Department of Energy (DOE) support an overwhelmingly large percentage of academic research in the United States. They often have policies that recommend and even require open release of funded research, including data and code, yet there is very little implementation or enforcement of these policies.[22] Washington is currently considering the extension of the open access implementation for manuscripts policies enacted by the NIH to other agencies,[23] but two things need to occur. Data and code must be included in the discussion of open access, and these policies of open access must be extended to agencies beyond the NIH. Each agency addresses very different bodies of research and thus implementation of open research may vary by agency, permitting each to face issues such as privacy, confidentiality, scientific norms including versioning and citation, and legal issues such as appropriate licensing of manuscripts, code, and data, as appropriate to the research communities involved. To aid in this effort, a number of research projects could be selected as pilots for the implementation of reproducible research thus providing an experiment in the full release of the code and data. Such carefully chosen pilot projects could help map out needs for open research, and then support could be given for these projects to facilitate their production of really reproducible research. This would create a scenario where it would be possible to learn what

[22] On May 10, 2010, the NSF announced that it would require the submission of a two-page data management plan along with grant applications for funding, beginning in October 2010. The impetus for the requirement is the need for open shared data: "Science is becoming data-intensive and collaborative," noted Ed Seidel, acting assistant director for NSF's Mathematical and Physical Sciences directorate. "Researchers from numerous disciplines need to work together to attack complex problems; openly sharing data will pave the way for researchers to communicate and collaborate more effectively." "This is the first step in what will be a more comprehensive approach to data policy," added Cora Marrett, NSF acting deputy director. "It will address the need for data from publicly funded research to be made public." See National Science Foundation, "Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans," news release, May 10, 2010, http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF.

[23] On April 15, 2010, Rep. Doyle reintroduced the Federal Research Public Access Act (H.R. 5037), seeking to make published papers from federally funded research publicly available over the Internet. See http://thomas.loc.gov/cgi-bin/bdquery/z?d111:HR05037:@@@P for the full text of the bill. A similar bill was introduced in the Senate on June 25, 2009 (S. 1373), by Senators Lieberman and Cornyn.

support, in terms of repositories, funding, or infrastructure, is needed and at what expense.[24]

To give an example, the National Science Foundation (NSF), through its role as a funding agency, makes a key contribution to the research incentives faced by many computational scientists and is in a unique position to address issues regarding verification of results, both through its research funding activities and policy leadership. There are five interlocking barriers to code and data release within funding agency purview: crafting appropriate release guidelines; collaborative tool development; intellectual property issues; facilitating access to research compendia;[25] and provision of "best practices" statements.

Issue 1: Enforcement of Existing Grant Guidelines

The NSF, for example, requires data and other supporting materials for any research it funds to be made available to other researchers at no more than incremental cost (with a provision for safeguards the right of individuals and subjects). The following passage is from the January 2009 NSF Grant General Conditions:

> **38. Sharing of Findings, Data, and Other Research Products**
>
> a. NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other support-

[24] On January 18, 2011 the National Science Foundation will require the submission of a two-page *Data Management Plan* with every grant, thereby creating a complementary experiment that also will reveal research needs and the costs associated with supporting reproducible research. See http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

[25] The term research "compendium" was coined by Robert Gentleman and Duncan Temple Lang to refer to "both a container for the different elements that make up the document and its computations (i.e. text, code, data...), and as a means for distributing, managing and updating the collection." See Gentleman and Lang, "Statistical Analyses and Reproducible Research."

> ing materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.[26]

This passage requires the release of data collected through NSF-funded activities, and recommends the release of accompanying software.

Recommendation 1.1: NSF Policy Expression

An important step would be to open the discussion of rewording the General Conditions to include the release of software, just as data are required to be released. Section 38 could be modified in the spirit of the following:

> **38. Sharing of Findings, Data, and Other Research Products**
>
> a. NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials, *such as software and inventions, created or* gathered in the course of the work. *Data and software should be made available in such a way that they are easily reusable by someone knowledgeable in the field.* (emphasis added)

Often the steps taken to generate computational results are embodied in software scripts or code. Computational research can includes a large number of small decisions—from data collation and filters, to software invocation sequences and parameter settings used in algorithms—that are impossible to capture com-

[26] National Science Foundation, *Grant General Conditions (GC-1)*, January 5, 2009, http://www.nsf.gov/pubs/gc1/jan09.pdf (accessed Sept. 26, 2009).

pletely in the final published paper, simply due to their multiplicity. A potentially fruitful way of communicating research methodology in these cases is to release the underlying code for inspection. Release of the accompanying data is the second necessary step for reproducibility of published computational findings.

Recommendation 1.2: Grantee-developed Release Plans

A blanket requirement of code and data release indicates funding agency intent but is not sufficient to create a regulatory environment in which researchers share easily reusable code and data, due to the difficulty in preparing code and data for release and widespread use. The use of computational tools is appearing in an increasing number of aspects of modern scientific research, making the myriad research settings in which these tools are used very complex, highly differentiated, and granular. One size does not fit all research problems, and a heavy-handed release requirement could result in *de jure* compliance—release of code and data—without the extra effort necessary to create usable code and data that facilitates the verification of the results. A solution partially under way (see footnote 28) would be to require grant applicants to formulate plans for release of the code and data generated through their research proposal, if funded. This creates a natural experiment where grantees, who know their research environments best, contribute complete strategies for release. This experiment would allow the funding agency to gather data on needs for release (repositories, further support); understand which research problem characteristics engender what particular solutions; identify what solutions are most appropriate in what settings; and uncover as yet unrecognized problems particular researchers may encounter. These findings would permit the funding agency to craft code and data release requirements that are more sensitive to barriers researchers face and the demands of their particular research problems, and implements strategies for enforcement of these requirements. This approach also permits researchers to address confidentiality and privacy issues associated with their research. This would not be the first implementation of this approach to policy crafting. The Wellcome Trust in the

United Kingdom began requiring grant applicants to submit comprehensive data release plans more than two years ago, and they are on the cusp of enforcing and observing these plans in action as grantees are now beginning to generate datasets from their funded research.[27]

Issue 2: Tools for Collaboration and Work Sharing

In the world of computing in general, not just scientific computing, the ubiquity of error has led to many responses, including special programming languages, error-tracking systems, disciplined programming efforts, and organized program-testing schemes. These efforts are key in developing a system of code and data release that does not create an overwhelming burden on the part of the computational scientist.

Recommendation 2.1: Funding of Software and Tool Development

Researchers use computational resources in very different ways. Examples range from short MATLAB scripts to the millions of lines of code, perhaps spanning several languages, that can underlie a complex simulation. The underlying software was typically not designed with scientific needs in mind and is generally a dialog with a single user, who would like to implement an algorithm or other innovation. It is up to the user to take extra steps to save the coding efforts and decisions taken, to record program invocation sequences and parameter settings, and otherwise track provenance of their research. These are exactly the steps it is important to share for verifiability, yet they are often not recorded as a natural part of the computational research process. In the heat of a computational project, researchers store many things in short-term memory that are needed at that moment to use the code productively. Facilitating the burden of code and data release means avoiding reliance on this soft, transient knowledge

[27] Nicole Perrin, Senior Policy Advisor, Wellcome Trust Limited, UK, "Data Matters: A Research Funder's Perspective" (keynote speech at COMMUNIA, Torino, June 29, 2009). See also "Wellcome Trust, policy on data management and sharing," last modified August 2010, http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm.

and, more specifically, codifying that knowledge objectively and reproducibly. An analogy could be drawn to the lab notebook kept by experimentalists. Its purpose is to record experimental methodology precisely, and is standard practice for all experimental sciences.

Tools for provenance are emerging but need to be developed at a much faster rate and for a much wider number of research problems, at a wider range of scales.[28] Even the solo researcher running software on his or her laptop can benefit from a system designed with the understanding that elements of the work that produce published results will be shared. Version control systems for code exist but are not routinely used by all computational scientists. Provenance tools must be easy to use since many researchers who use computational methods are not computer specialists. Aspects such as unit testing and standardized test beds (as is typical in open source code development) should be emphasized and even required for scientific code.

The scope of this problem is broad enough to warrant a discussion of targeted funding from the NSF and other agencies that fund computational work, particularly as research begins to move into the cloud and increasingly takes place in shared virtual spaces. Many computational scientists will require retraining in the use of software that tracks provenance and allows for such workflow sharing.

In a recent survey of computational scientists, the incremental amount of work involved in preparing code and data for release was the primary barrier to open code and data sharing.[29] Routinely used software tools typically lack a system of incorpo-

[28] See http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge; the UK-funded Taverna software package, http://www.mygrid.org.uk/; the Sumatra package for reproducible simulations, http://neuralensemble.org/trac/sumatra; the Pegasus system developed at the University of Southern California, http://pegasus.isi.edu/; and Galaxy software developed at Penn State University, http://galaxy.psu.edu/. See Microsoft's Trident Workbench for an oceanography example, http://research.microsoft.com/en-us/collaboration/tools/trident.aspx.

[29] See Victoria Stodden, "The Scientific Method and Computation: Reproducibility in the Computational Sciences" (forthcoming).

rating the preparation of code as data for release, as the research is progressing. Recreating steps previously taken is difficult for any scientist when working in a programming environment designed for running code, but not for sharing or working collaboratively. Such tools will not only facilitate the release of coherent and reusable code and data, but ease research collaboration and facilitate communication of work in progress between coauthors. A challenge for scientific research is developing software environments to enable collaborative research, and facilitating reproducibility of computational results is a key step in this process.

Recommendation 2.2: Funding of Statistical Methods for Simulation-Based Modeling

An increasingly pervasive methodological tool is the use of massive simulations of a physical system's complete evolution, repeated numerous times while varying simulation parameters systematically. Such models in climate research provide the foundation for some of our most crucial public policy decisions and are beginning to represent scientific research in the public dialogue. Statistical machinery analogous to such long-standing tools in conventional modeling as error bounds on prediction, parameter estimation, and overall model fit must be developed in the case of computer simulation. An important step is the development of workflow-tracking software environments to facilitate tracing of error sources as mentioned previously, but further research is needed to understand how to evaluate the output of simulations. Since this is a new area of research, framing of the uncertainty quantification problem should be carefully undertaken as a preliminary step to a broader research agenda.

Issue 3: The Intellectual Property Framework for Code and Data Release

Even though scientists produce public goods, their work is not immune to intellectual property strictures, as elaborated in Part I.

Recommendation 3.1: Adopting the Reproducible Research Standard

As discussed in the previous section, the Reproducible Research Standard (RRS) realigns the intellectual property framework faced by computational researchers with longstanding scientific norms.[30] The RRS suggests a licensing structure for research compendia, including code and data, which permits others to use and reuse code and data without having to obtain prior permission or assume a fair use exception to copyright, so long as attribution is given.[31] Using the RRS on all components of computational scholarship will encourage reproducible scientific investigation, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.

Issue 4: Access to Published Research Papers and Supporting Materials

Asking computational scientists to embrace reproducibility poses questions with regard to location of research compendia on the Internet and access to published results.

Recommendation 4.1: Funding Agency Public Access Policy

Reproducibility requires not only access to underlying code and data, but access to the original published article. Funding agencies such as the NSF and DOE could create a digital archive, analogous to the National Institutes of Health's PubMed Central, and require the deposit of their funded final manuscripts. The NIH

[30] For a full discussion of the Reproducible Research Standard, see Victoria Stodden, "Enabling Reproducible Research."

[31] Fair use is how the U.S. copyright law provides for the use of copyrighted works without the need to obtain the copyright holder's permission, in order to provide flexibility in balancing the interests of copyright holders and the public's desire to make use of copyrighted works. The copyright statute states that "...the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright." (17 U.S.C. § 107). Whether or not use of copyrighted material can be deemed fair use is fact specific and subject to a four-factor test. How far the scholarship exception extends is unclear, and scientists may not feel comfortable relying on it when building on another scientist's research through, for example, reusing code.

requires that papers that arise from NIH funds comply with their public access to policy: final peer-reviewed journal manuscripts must be submitted to PubMed Central upon acceptance for publication, and become accessible to the public no longer than twelve months after publication.[32]

The NIH further requires that copyright be addressed. If publicly funded research falls under the Reproducible Research Standard as described in Section 3, articles will be licensed using the Creative Commons attribution license, therefore removing copyright barriers from the paper. Many journals, however, require authors to assign copyright to the journal as a condition of publication, but will allow an earlier version to be posted publicly. The NIH has made publication in journals that permit the article, or a version thereof, to be posted in PubMed Central a requirement of funding—this strategy is an option for other funding agencies as well.

The final requirement the NIH makes of grant recipients is to use the PubMed Central identifier at the end of citations. Encouraging the use of unique identifiers of papers, as well as code and data, can encourage the release and hence citation of all forms of computational research.[33] Such a unique identifier would indicate compliance with funder agency open-access policies.

It is important that these requirements be tied to grant funding and a mechanism established that allows compliance to be reflected in future grant determinations. Strategies for release of code and data arising from a particular grant should be subject to peer review in the grant evaluation process.

[32] The twelve-month post-publication grace period could be applied to code and data release, upon researcher request. This strategy was advocated for genome data in “Prepublication data sharing,” *Nature* 461, (10 September 2009): 168-170, http://www.nature.com/nature/journal/v461/n7261/full/461168a.html.

[33] See e.g. Altman and King’s Uniform Numerical Identifier proposal for data citation, http://thedata.org/citation/standard. This also can ensure that an unmodified version of a dataset is used in different research studies, when confidentiality or other concerns prohibit open release of the dataset.

Recommendation 4.2: Funding Agencies Support Digital Archiving for Data and Code

For papers whose results can be replicated from short scripts and small datasets, many computational scientists who do engage in reproducible research are able to host their research compendia on their institutional web pages or using hosting resources their institution is willing to provide.[34] Not all computation research involves small amounts of supplemental code and data; hosting very large datasets or complex bodies of code may be necessary and home institutional support may not be available to the researcher. A funding agency could create code and data repositories as for papers (perhaps even jointly among agencies), or seek to increase support of the growing set of data repositories emerging at institutions.[35] Data is necessary for reproducibility of computational research, but an equal amount of concern should be directed at code sharing. As yet, code sharing repositories are not established to the extent that data repositories are.

Tagging of research compendia is an important issue for communicating work, facilitating topical web searches, and aggregating a researcher's contributions, including their code and dataset building activities. Development of a standard RDFa vocabulary for HTML tags for publicly funded research would enable searches for code, data, and research as well as facilitating the transmission of licensing information, authorship, and sources. That such a standard would enable searches by author would allow a more granular understanding of a scientist's research contributions, beyond citations. This would provide an incentive to release code and data, and give groups, such as funders, award committees, and university hiring and promotion committees, access to a more accurate representation of the researcher's work. Such a tagging vocabulary could include unique identifiers for code and data, ideally the same as those required for repository deposit as

[34] See e.g. http://sparselab.stanford.edu and http://www-stat.stanford.edu/~wavelab.

[35] See e.g. The Stanford Microarray Database, http://smd.stanford.edu/.

discussed in the previous section, and thus facilitate and encourage their citation.

Issue 5: Reproducible Research "Best Practice" Recommendations

Computational scientists may be unaware of the need to work reproducibly, researchers may be unaware of what it means to do so, and funding agencies and journals may find it useful to have a clear explanation of the issue and its implementation at the funding agency.

Recommendation 5.1: Release of Funding Agency "Best Practice" Recommendations

Such a document would be publicly available at a stable URL, updated with versions, and intended to provide clarity on all relevant issues. It would be framed to suggest ideal recommendations, rather than list a series of requirements. Some points that such a list may wish to touch on follow below.

Reproducibility is a goal of computational science, and practicing reproducible research means:

- Uploading the final peer-reviewed journal manuscripts that arise from agency-funded research to a digital archive upon acceptance for publication;
- Making the code and data required to reproduce results in agency-funded works publicly available online within twelve months of publication (or less);
- Utilizing appropriate licensing structures for agency-funded research, such as the Reproducible Research Standard; and
- Utilizing tagging structures for agency-funded compendia release, as part of inclusion in repositories or posting on institutional repositories.

The Necessity of a Multifaceted Approach

This discussion is intended to frame issues that arise with the implementation of reproducibility in computational science.

These recommendations reflect a set of interlocking issues, and progress from one recommendation will be facilitated by implementation of other recommendations.

PART III. UNTANGLING OWNERSHIP ISSUES FOR SCIENTIFIC COLLABORATION AND OPEN DISSEMINATION: A NEW VISION FOR UNIVERSITY LEADERSHIP

With the advent of large-scale data and the pervasiveness of computing in scientific research, ownership issues for code and data have yet to be fully addressed. Data often are generated in collaboration with co-researchers, who may be in academia, government, or the private sector, and funding sources can be equally as varied. Copyright endows authors of code with exclusive rights, contracts with universities often give home institutions a claim, and evidence suggests that journals are turning their publishing models for articles toward hosting and releasing the associated code and data. To make matters more complex, repositories for both code and data are coming online with their own ownership and licensing schemes for scientific products.

When data and code are widely shared, such ownership issues come sharply to the fore. What is missing today is clarity regarding ownership rights, which can vary by case, and one ownership model may not transfer from one research setting to the next. To accelerate the wide dissemination of newly discovered scientific knowledge, an ombudsman position needs to be created at the university level, perhaps within the Copyright Office or Provost's Office, to streamline the process of rights ascertainment and negotiate agreements for sharing of collaboratively created code and data. This position would be regarded as temporary, perhaps lasting a decade, during which a set of typical sharing arrangements would emerge. In the longer term, negotiation over ownership and sharing rights would be shifted to the beginning of the project, when collaborators could typically adopt one of the small number of established emergent ownership models.

There are a limited number of possible claims in data and code ownership. The scientists themselves, their university, or funding bodies (public and private) are principal stakeholders. Ownership rights vested in the scientist present the least complex case, in that norms of openness in methods and in reproducible research exist, even if they are not always carefully implemented. In my recent survey of computational scientists, there emerged a clear tension between open science and code patenting.[36] Some respondents noted that a reason not to share their code even after publication was the possibility of patents (and the possibility of forming a company around the patented technology). A perhaps unexpected result of the Bayh-Dole Act of 1980, passed on the eve of the computer revolution in scientific research, was the creation of incentives for universities and academic researchers to lock scientific knowledge in patents. With the intention of providing an impetus for universities to transfer innovations outside academia and thus facilitate commercial and industrial development, the result has been the creation of a barrier to both scientific integrity and openness in the communication of scientific discovery, insofar as innovations are not shared openly.[37] For scientific findings to be reproducible, code and data must be open and verifiable, and to accelerate scientific innovation the code and data must be modifiable, reusable, and able to be applied to novel research problems. Patented code inserts the university's Office of Technology Licensing into this process, disrupting the open flow of downstream scientific research. A second recommendation is an automatic exception from patent use restrictions on code used for academic research purposes, still permitting commercial development of new technologies. This could be achieved through an open licensing structure that distinguishes between commercial and noncommercial downstream use of the scientific output.

[36] Victoria Stodden, "The Scientific Method in Practice: Reproducibility in the Computational Sciences" (MIT Sloan Research Paper no. 4773-10), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193.

[37] For a further description, see Joseph Stiglitz and John Sulston, "The Manchester Manifesto" (November 2009), Available at http://www.isei.manchester.ac.uk/TheManchesterManifesto.pdf.

Open code is emerging as a requirement for publication and is a clear component of reproducible computational research. A clash is emerging between the requirements for scientific integrity in the computer age—open code and data—and the incentives of the university to extract licensing fees from patented code written by university researchers. Without the creation of an exemption for code re-use in the academic setting, including the verification of published results and the application to new research problems, scientific integrity will suffer, deepening the current credibility crisis in computational science.[38] The university is uniquely positioned to play a key leadership role in establishing standard protocols and sharing agreements among scientific collaborators that facilitate the wide dissemination of discoveries and knowledge, thereby accelerating innovation and growth.

[38] An analogous proposal has been made by Paul David, when he suggested expanding the Fair Use provision in copyright law to encompass all academic research output. See Paul David, "The Economic Logic of 'Open Science.'"