SCIENTIFIC INTEGRITY AND TRANSPARENCY

HEARING

BEFORE THE

SUBCOMMITTEE ON RESEARCH

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HOUSE OF REPRESENTATIVES

ONE HUNDRED THIRTEENTH CONGRESS

FIRST SESSION

TUESDAY, MARCH 5, 2013

Serial No. 113–10

Printed for the use of the Committee on Science, Space, and Technology

Available via the World Wide Web: http://science.house.gov

U.S. GOVERNMENT PRINTING OFFICE

79–929PDF WASHINGTON : 2013

For sale by the Superintendent of Documents, U.S. Government Printing Office
Internet: bookstore.gpo.gov Phone: toll free (866) 512–1800; DC area (202) 512–1800
Fax: (202) 512–2104 Mail: Stop IDCC, Washington, DC 20402–0001

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HON. LAMAR S. SMITH, Texas, *Chair*

DANA ROHRABACHER, California
RALPH M. HALL, Texas
F. JAMES SENSENBRENNER, JR., Wisconsin
FRANK D. LUCAS, Oklahoma
RANDY NEUGEBAUER, Texas
MICHAEL T. McCAUL, Texas
PAUL C. BROUN, Georgia
STEVEN M. PALAZZO, Mississippi
MO BROOKS, Alabama
RANDY HULTGREN, Illinois
LARRY BUCSHON, Indiana
STEVE STOCKMAN, Texas
BILL POSEY, Florida
CYNTHIA LUMMIS, Wyoming
DAVID SCHWEIKERT, Arizona
THOMAS MASSIE, Kentucky
KEVIN CRAMER, North Dakota
JIM BRIDENSTINE, Oklahoma
RANDY WEBER, Texas
CHRIS STEWART, Utah
VACANCY

EDDIE BERNICE JOHNSON, Texas
ZOE LOFGREN, California
DANIEL LIPINSKI, Illinois
DONNA F. EDWARDS, Maryland
FREDERICA S. WILSON, Florida
SUZANNE BONAMICI, Oregon
ERIC SWALWELL, California
DAN MAFFEI, New York
ALAN GRAYSON, Florida
JOSEPH KENNEDY III, Massachusetts
SCOTT PETERS, California
DEREK KILMER, Washington
AMI BERA, California
ELIZABETH ESTY, Connecticut
MARC VEASEY, Texas
JULIA BROWNLEY, California
MARK TAKANO, California
VACANCY

SUBCOMMITTEE ON RESEARCH

HON. LARRY BUCSHON, Indiana, *Chair*

STEVEN M. PALAZZO, Mississippi
MO BROOKS, Alabama
STEVE STOCKMAN, Texas
CYNTHIA LUMMIS, Wyoming
JIM BRIDENSTINE, Oklahoma
LAMAR S. SMITH, Texas

DANIEL LIPINSKI, Illinois
ZOE LOFGREN, California
AMI BERA, California
ELIZABETH ESTY, Connecticut
EDDIE BERNICE JOHNSON, Texas

CONTENTS

Tuesday, March 5, 2013

SCIENTIFIC INTEGRITY AND TRANSPARENCY

TUESDAY, MARCH 5, 2013

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON RESEARCH
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY,
Washington, D.C.

The Subcommittee met, pursuant to call, at 10:01 a.m., in Room 2318 of the Rayburn House Office Building, Hon. Larry Bucshon [Chairman of the Subcommittee] presiding.

LAMAR S. SMITH, Texas
CHAIRMAN

EDDIE BERNICE JOHNSON, Texas
RANKING MEMBER

Congress of the United States

House of Representatives

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

2321 RAYBURN HOUSE OFFICE BUILDING

WASHINGTON, DC 20515-6301

(202) 225-6371
www.science.house.gov

Subcommittee on Research

Scientific Integrity and Transparency

Tuesday, March 5, 2013
10:00 a.m. to 12:00 p.m.
2318 Rayburn House Office Building

Witnesses

Prof. Bruce Alberts, Professor of Biochemistry, University of California San Francisco

Prof. Victoria Stodden, Assistant Professor of Statistics, Columbia University

Dr. Stanley Young, Assistant Director for Bioinformatics, National Institute of Statistical Sciences

Mr. Sayeed Choudhury, Associate Dean for Research Data Management at Johns Hopkins University and Hodson Director of the Digital Research and Curation Center

U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
SUBCOMMITTEE ON RESEARCH

HEARING CHARTER

Scientific Integrity and Transparency

TUESDAY, MARCH 5, 2013
10:00 A.M. TO 12:00 P.M.
2318 RAYBURN HOUSE OFFICE BUILDING

Purpose

At 10 AM on Tuesday, March 5, 2013, the Subcommittee on Research will hold a hearing titled Scientific Integrity and Transparency. This hearing will provide Members an opportunity to understand the problem of access to underlying data from published research funded by the federal government, and why access to this underlying data is vital to scientific integrity and transparency for peer reviewed research. On March 29th, 2012 the Investigation and Oversight Subcommittee held a hearing entitled, "Federally Funded Research: Examining Public Access and Scholarly Publication Interests."[1] The focus of this past hearing was on open access to publications, whereas the focus of this hearing is on open access to data used in federal research.

Witnesses

- Prof. Bruce Alberts, Professor of Biochemistry, University of California San Francisco
- Prof. Victoria Stodden, Assistant Professor of Statistics, Columbia University
- Dr. Stanley Young, Assistant Director for Bioinformatics, National Institute of Statistical Sciences
- Mr. Sayeed Choudhury, Associate Dean for Research Data Management at Johns Hopkins University and Hodson Director of the Digital Research and Curation Center

Overview

The bedrock of the scientific process is the ability to replicate the experimental claims made by researchers. These claims include both the generation of data and the analysis of data by computer software and code. Scientists rarely reproduce the work of others since they neither have the time nor the resources to reliably replicate the work of their colleagues; instead, they often trust these claims and rely on the peer review process and their colleagues to share their data and analysis methods when needed. This exchange allows for scientists and companies to exploit the latest insights to develop new directions in their research, and allows them to maximize the impact of federal research investment. Thus, scientific progress cannot occur unless there is a strong culture of integrity and transparency.

Unfortunately, the current system has demonstrated several flaws. The current incentive system rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure. The process of peer review, which the scientific community views as its primary means to check scientific integrity in journal publications, oftentimes does not try to replicate the results of submitted papers. Fellow researchers conducting the peer review for publication rarely ask for the original data of the submitted paper they are reviewing, and focus instead on whether the claims made in the paper are plausible. They simply assume the underlying data is valid. In a recent study by Young and Karr, up-

[1] http://science.house.gov/hearing/subcommittee-investigations-and-oversight-hearing-examining-public-access-and-scholarly

wards of 90% of clinical trial claims for new medicines cannot be replicated.[2] The inability to replicate published results is not unique to clinical trials and occurs across scientific disciplines.[3]

This hearing will attempt to understand the scope of the problem with scientific integrity, especially how thorough researchers deal with underlying data. This issue of scientific integrity should be differentiated from cases of scientists knowingly and intentionally committing scientific fraud, fabricating data, or plagiarism though these might be inter-related depending on individual circumstances. This hearing will focus primarily on how data is collected, shared, and analyzed by the scientific community and policies for what, how, and when federally funded research data should be shared, as well as the cost of making this data available to the scientific community and public. Current federal laws governing the sharing of data include the Data Access Act (DAA) of 1999 and the Information Quality Act (IQA) of 2001.[4] Introduced by Senator Richard Shelby, the DAA (sometimes known as "the Shelby Amendment" within the science community) requires that data from federally funded research be made available under the Freedom of Information Act procedures. The IQA requires the OMB to issue regulations for ensuring the quality and integrity of all information disseminated by federal agencies. However, the Government Accountability Office reported in September 2007 that federal agencies rarely monitor whether researchers make data available.[5]

In response to these aforementioned issues, the Office of Science and Technology Policy (OSTP) released guidance to federal agencies on February 22nd about increasing access to the results of federally funded scientific research which includes a discussion about access to non-classified digital data. In this memo, OSTP outlines the following principles for federal funding agencies to follow when a issuing a data access plan[6]:

- Maximize access to scientific data created with federal funds;
- Ensure that researchers develop data management plans, and allow inclusion for costs in proposals along with proper evaluations of these proposals;
- Include mechanisms to ensure compliance with data management plans and policies;
- Promote the deposit of data in publicly accessibly databases;
- Encourage cooperation with the private sector to improve data access and compatibility;
- Develop approaches for identifying/providing appropriate attribution to data sets;
- Support the training, education and workforce development related to data management; and
- Provide assessment of long-term needs for the preservation of scientific data.

This hearing will address how such principles might best be implemented by federal research agencies and members of the scientific community conducting such research.

---

[2] http://science.house.gov/sites/republicans.science.house.gov/files/documents/hearings/HHRG-112-SY20-WState-SYoung-20120203.pdf

[3] "Again, and again, and again." p1225 Science Vol 334 2 December 2011

[4] National Research Council, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age (Washington, DC: National Academy Press), 2009.

[5] http://www.gao.gov/products/GAO-07-1172

[6] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_*public_access_memo_2013.pdf*

Chairman BUCSHON. The Subcommittee on Research will now come to order.

Good morning. Welcome to today's hearing entitled "Scientific Integrity and Transparency." In front of you are packets containing the written testimonies, biographies and Truth-in-Testimony disclosures for today's witness panel. I recognize myself for five minutes for an opening statement.

I want to welcome everyone to today's Research Subcommittee hearing on the issue of scientific integrity and transparency.

An editorial in the March 29, 2012, edition of Nature magazine entitled: "Must try harder: too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data and at themselves." I found this editorial particularly interesting because of my background as a cardiothoracic surgeon and my professional interest in medicine. The editorial goes on to cite a recent study contained in this specific issue by Glenn Begley and Lee Ellis, which analyzes the low number of cancer research studies that have been converted into clinical success, and concludes that a major factor is the overall poor quality of published preclinical data. This is one of the many similar studies that I have read.

The growing lack of scientific integrity and transparency has many causes but one thing is very clear: without open access to data, there can be neither integrity nor transparency from the conclusions reached by the scientific community. Furthermore, when there is no reliable access to data, the progress of science is impeded and leads to inefficiencies in the scientific discovery process. Important results cannot be verified, and confidence in scientific claims dwindles.

The Federal Government is the main sponsor of basic scientific research, with over $140 billion spent in fiscal year 2013. The American scientific community has made enormous contributions in many scientific fields from federally sponsored research. I believe our Nation's scientists will continue to develop the breakthrough discoveries and innovations of tomorrow. However, scientists receiving federal funding need to be accountable and responsible stewards of taxpayers' resources. Hardworking Americans trust our scientists to be genuine and authentic in the way they conduct and share federally funded research.

The focus of this hearing will be on scientific research data funded by the Federal Government. There are key issues that data-sharing policies should address including what is data, how it should be shared, when it should be shared, and what potential costs might result in making this data available to the research community. We want to maximize access to data while protecting personal privacy, avoid any negative impact on intellectual property rights and innovation, and preserve data without ridiculous cost or administrative burdens.

In an attempt to begin addressing this issue, the Office of Science and Technology Policy released guidelines on February 22nd of this year that recognized the problem of data access. These guidelines, intended for federal science agencies, are to be followed when determining a policy for public access to scientific data in dig-

ital formats. As part of this hearing, I look forward to hearing the witnesses' opinions on these federal guidelines.

Our witnesses today offer input from a variety of scientific fields, as this problem is not exclusive to one scientific field, community or discipline. I would like to thank them for coming and taking the time to offer their expertise. I would also like to thank Ranking Member Lipinski and everyone else participating in today's hearing.

[The prepared statement of Mr. Bucshon follows:]

PREPARED STATEMENT OF CHAIRMAN LARRY BUCSHON

I want to welcome everyone to today's Research subcommittee hearing on the issue of scientific integrity and transparency.

An editorial in the March 29, 2012 edition of Nature magazine was entitled: "Must try harder: too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data—and at themselves." I found this editorial particularly interesting because of my background as a cardiothoracic surgeon and my professional interest in medicine. The editorial goes on to cite a recent study (contained in this specific issue) by Glenn Begley and Lee Ellis which analyzes the low number of cancer-research studies that have been converted into clinical success, and concludes that "a major factor is the overall poor quality of published preclinical data." This is one of many similar studies that I have read.

The growing lack of scientific integrity and transparency has many causes but one thing is very clear: without open access to data, there can be neither integrity nor transparency from the conclusions reached by the scientific community. Furthermore, when there is no reliable access to data, the process of science is impeded and leads to inefficiencies in the scientific discovery process. Important results cannot be verified, and confidence in scientific claims dwindles.

The federal government is the main sponsor of basic science research, with over $140 billion spent in fiscal year 2013. The American scientific community has made enormous contributions in many scientific fields from federally sponsored research. I believe our nation's scientists will continue to develop the breakthrough discoveries and innovations of tomorrow. However, scientists receiving federal funding need to be accountable and responsible stewards of tax-payer resources. Hard-working Americans trust our scientists to be genuine and authentic in the way they conduct and share federally funded research.

The focus of this hearing will be on scientific research data funded by the federal government. There are key issues that data-sharing policies should address including: what is data, how it should be shared, when it should be shared, and what potential costs might result in making this data available to the research community. We want to maximize access to data while protecting personal privacy, avoid any negative impact on intellectual property rights and innovation, and preserve data without ridiculous cost or administrative burdens. In an attempt to begin addressing this issue, the Office of Science and Technology Policy released guidelines on February 22nd of this year that recognized the problem of data access. These guidelines, intended for federal science agencies, are to be followed when determining a policy for public access to scientific data in digital formats. As part of this hearing, I look forward to hearing the witness's opinions on these federal guidelines.

Our witnesses today offer input from a variety of scientific fields, as this problem is not exclusive to one scientific field, community, or discipline. I'd like to thank them for coming and taking time to offer their expertise. I'd also like to thank Ranking Member Lipinski and everyone else participating in today's hearing.

Chairman BUCSHON. With that, I now recognize the Ranking Member, the gentleman from Illinois, Mr. Lipinski, for an opening statement.

Mr. LIPINSKI. Thank you, Chairman Bucshon. I think this is our third hearing in three weeks, and we have another one next week that I will now label you the hardest-working Chairman in Washington, D.C. So it is good to be at work here and I want to thank all the witnesses for being here.

The United States has for decades represented the world's gold standard for scientific integrity. But no one should mistake this observation as an argument for complacency. In the COMPETES Act of 2007, which we worked on in this Subcommittee, then-Subcommittee Chairman Brian Baird included a provision on Responsible Conduct of Research that required every institution receiving NSF grant funding to provide training on the ethical conduct of science to all students and postdocs covered under those grants. Today, all U.S. research universities have implemented research ethics training for their STEM students and trainees, which we all can agree is a good thing.

The bigger challenge to the progress of science is not misconduct, but rather poor methodology and bad statistical analysis that take a long time to uncover. Or for that matter, discoveries in one field that have broad multidisciplinary relevance but take time to be known in other fields. To that end, the open sharing of scientific data is good for science and it is good for society. We must, of course, respect issues of privacy and intellectual property. But the more data are open, the faster we will validate new theories and overturn old ones, and the more efficiently we will transform new discoveries into innovations that will create jobs and make us healthier and more prosperous. The movement toward open data is not primarily about scientific integrity; it is mostly about speeding up the process of scientific discovery and innovation.

However, there are some big challenges to the widespread implementation of open data. Someone must define what exactly data sharing is going to mean and how it is going to be done, beginning with a standard. The February 22nd OSTP memo, which the Chairman mentioned, on increasing access to the results of federally funded scientific research, which by the way was also a direct response to requirements in the COMPETES Act, takes on many of these issues in detail. But specifically, here are some questions that we have to consider, and some of these questions were questions raised by the Chairman. First, what does it entail and how much does it cost for researchers to develop a data management plan and to prepare their own data for sharing? Do they have adequate assistance from professional information managers? Are funding agencies sufficiently aware of the costs and skills required for good data management plans, and how should they evaluate and budget for data management proposals? What are the IT infrastructure needs for data sharing, including technical standards, and what, if any, scientific or technical barriers exist to developing that infrastructure? What are the most important factors to consider in the economics of digital data access and preservation? What should be the respective roles of science agencies, universities, and the private sector in supporting and preserving public databases? How can these groups work together to minimize costs and maximize benefit to the scientific community? And finally, are there any policy or legal barriers for sustainable digital access and preservation?

In light of the majority's suggestion of a possible legislative outcome for this hearing, I hope that today's dialogue will include a thoughtful discussion of some of these practical issues of implementation. I know that all four expert witnesses before us have a lot

to contribute to this discussion and I look forward to learning from them because this is certainly something that is important for us to pursue but we need to make sure that we are covering all our bases here and do this in the right manner.

With that, I yield back.

[The prepared statement of Mr. Lipinski follows:]

PREPARED STATEMENT OF RANKING MINORITY MEMBER DANIEL LIPINSKI

Thank you Chairman Bucshon and thanks to all of the witnesses for being here.

The U.S. has for decades represented the world's gold standard for scientific integrity. But no one should mistake this observation as an argument for complacency. In the COMPETES Act of 2007, which we worked on in this subcommittee, then Subcommittee Chairman Brian Baird included a provision on Responsible Conduct of Research that required every institution receiving NSF grant funding to provide training on the ethical conduct of science to all students and postdocs covered under those grants. Today, all U.S. research universities have implemented research ethics training for their STEM students and trainees.

The bigger challenge to the progress of science is not misconduct, but rather poor methodology and bad statistical analysis that take a long time to uncover. Or for that matter, discoveries in one field that have broad multidisciplinary relevance but take time to be known in other fields. To that end, the open sharing of scientific data is good for science and it's good for society. We must, of course, respect issues of privacy and intellectual property. But the more data are open, the faster we will validate new theories and overturn old ones, and the more efficiently we will transform new discoveries into innovations that will create jobs and make us healthier and more prosperous. The movement toward open data is not primarily about scientific integrity, it's mostly about speeding up the process of scientific discovery and innovation.

However, there are some big challenges to the widespread implementation of open data. Someone must define what exactly data sharing is going to mean and how it is going to be done, beginning with a standard. The February 22nd OSTP memo on increasing access to the results of federally funded scientific research, which by the way was also a direct response to a requirement in COMPETES, takes on many of these issues in detail.

Specifically, we must consider such questions as:

- What does it entail and how much does it cost for researchers to develop a data management plan and to prepare their own data for sharing? Do they have adequate assistance from professional information managers?
- Are funding agencies sufficiently aware of the costs and skills required for good data management plans, and how should they evaluate and budget for data management proposals?
- What are the IT infrastructure needs for data-sharing, including technical standards, and what, if any, scientific or technical barriers exist to developing that infrastructure?
- What are the most important factors to consider in the economics of digital data access and preservation?
- What should be the respective roles of science agencies, universities, and the private sector in supporting and preserving public databases? How can these groups work together to minimize costs and maximize benefit to the scientific community?
- And finally, are there any policy or legal barriers for sustainable digital access and preservation?

In light of the Majority's suggestion of a possible legislative outcome for this hearing, I hope that today's dialogue will include a thoughtful discussion of some of these practical issues of implementation. I know that all four expert witnesses before us have a lot to contribute to this discussion and I look forward to learning from them.

With that I yield back.

Chairman BUCSHON. Thank you, Mr. Lipinski.

If there are Members who wish to submit additional opening statements, your statements will be added to the record at this point.

At this time I would like to introduce our witnesses. Our first witness is Dr. Bruce Alberts, Editor-in-Chief of Science Magazine and Professor Emeritus of Biochemistry and Biophysics at the University of California-San Francisco. Welcome. Our next witness is Dr. Victoria Stodden, Assistant Professor of Statistics at Columbia University. Our third witness is Dr. Stanley Young, the Assistant Director of Bioinformatics at the National Institutes of Statistical Sciences. That was hard to say. Our fourth and final witness today is Mr. Sayeed Choudhury, Associate Dean for Research Data Management at Johns Hopkins University and Hodson Director of the digital Research and Curation Center.

As our witnesses should know, spoken testimony is limited to five minutes each after which Members of the Committee will have five minutes each to ask questions.

I now recognize Dr. Alberts to present his oral testimony.

TESTIMONY OF DR. BRUCE ALBERTS, EDITOR-IN-CHIEF, SCIENCE MAGAZINE AND PROFESSOR EMERITUS OF BIOCHEMISTRY AND BIOPHYSICS, UNIVERSITY OF CALIFORNIA – SAN FRANCISCO

Dr. ALBERTS. It is a pleasure to be here today. I would just like to start by emphasizing something that Science Magazine covers repeatedly, which is the fact that our strength in science and technology in the United States underlies both our economic success and our military dominance in the world. As you all know, many other nations are increasingly making investments in this area, and I find it distressing that although this Committee has long supported fundamental, long-term scientific research, the investment in the United States has been stagnant for many years. The investment in this kind of research was 1.25 percent of GDP in 1985, has dropped to .87 percent of GDP in 2013, a big drop, and of course, the current sequester will now make our situation even worse. I believe that this is dangerous for America's future, for my grandchildren's future.

But this hearing, of course, is to focus on the quality and not the quantity of U.S. research. I would like to address first the data availability issue, which of course is crucial for science. Science builds by one scientist testing and building on and maybe refuting the data of other scientists, very much a community endeavor. And the privilege of publishing in a journal like ours demands data sharing. Otherwise science doesn't work.

So our journal has been working on this. This is a special issue we published, 14 long articles about all these issues, February 2011, and we are publishing more and more about this. It is accompanied by a survey, a useful survey of scientists, how they use data and whether they have enough access. And we have stressed over and over again that our policy is "that all data necessary to understand, assess and extend the conclusions of the manuscript must be available to any reader of science." In this issue, we announced a new policy. This includes computer codes involved in the creation

or analysis of data, and I am pleased to say that we are getting good compliance with those policies.

Of course, there are problems that remain. You will hear about them from the rest of the group here. But one I would like to emphasize is guaranteeing funding for the public databases, the critical ones, funding long term so that the community and journals like ours can rely on them. This is really a major issue. In my field, the protein database, for example, is absolutely crucial. It has got 100,000 different protein coordinates in it. You know, if funding lapses, then we lose all this, and these places play major roles in setting standards as well.

And secondly, I would like to emphasize that we need tools for interacting with the largest data sets that are now increasingly provided as supplemental online information and journal publications like ours, so when we demand the data, we put the data not in the written paper but most of it in a big electronic supplement, and other journals are doing that as well, but we need ways to help people analyze that data who are not the original authors. And of course, every journal needs to stress clear and complete presentation of all the materials and methods that were used in the research.

So the other issue is data reproducibility. Mr. Chairman, you quoted from that paper. My conclusion, and talking to people at Genentech who would agree with that paper that you cited from Bayer Health Care is that the scientific standards are lower in some fields of science and others that we need to work on setting higher standards.

In addition, human cells are incredibly complex and it is easy to get a result that looks right when it is really wrong, and one can easily be fooled. Every scientist must be trained to be highly suspicious about his or her own results, and this again is a major issue. And finally, I believe we are overemphasizing research directly aimed at finding drugs at the expense of the high-quality discovery-driven basic research that is urgently needed to improve the search for disease treatments. We are just mostly stabbing in the dark.

So my suggestions for improving this situation would demand a community effort from scientific journals like ours. We have new policies in the last three years that every senior author for each part of the results being published must confirm that he or she has personally reviewed the original data generated by that unit, specifying where exactly those results appear in the paper. It used to be that we wanted one author to take responsibility. That is totally unreasonable now. Half of our papers have authors in different countries. We would have to have a set of senior authors. We are developing checklists in various fields of science to help journals and scientists. There is a biosketch issue. People should not be listing huge lists of publications to impress other people who are giving them grant funds. They need to focus on their five or ten most important contributions, and quality is critical, not quantity. And funding agencies have a role to play here as well.

I just want to emphasize my own role at universities. I am still teaching. I am going to be teaching a 2-week minicourse on ethics

and research standards this May, so I am very much involved in these issues. Thank you.

[The prepared statement of Dr. Alberts follows:]

1

Written Testimony of
Dr. Bruce M. Alberts
Editor-in-Chief, *Science* magazine;
Professor Emeritus, UCSF
Before the Subcommittee on Research,
Committee on Science, Space, and Technology
U.S. House of Representatives
Hearing on
"Scientific Integrity and Transparency"
March 5, 2013

Mr. Chairman, Ranking Member Lipinski, and Members of the Subcommittee, my name is Bruce Alberts and I currently serve as the Editor-in-Chief of *Science* magazine. I thank you for the opportunity to speak to you today on this important topic for the future of science and the United States.

Science magazine is a leading weekly science journal (100,000 subscriptions) published by the American Association for the Advancement of Science (AAAS). I am a biochemist and cell biologist whose major research contributions have concerned the mechanism of DNA replication, which is the process that duplicates chromosomes before a cell divides. A Professor Emeritus in the School of Medicine at the University of California, San Francisco (UCSF), I have recently served as one of the first three U. S. Science Envoys, appointed by Secretary of State Hillary Clinton. My previous positions include: full-time president of the National Academy of Sciences (1993-2005), president of the American Society for Cell Biology, and chairman of the Department of Biochemistry and Biophysics at UCSF. I am a member of the National Academy of Sciences, and a foreign member of the Royal Society (UK), the Indian National Science Academy, and the academies of several other nations.

As I have written in many editorials for *Science*, the strength of US science and technology (S & T) has been, and will long be, critical for our position as the leading nation of the world. It underlies both our economic success and our military dominance. In recent years, nations like China have focused intensely on strengthening their own S & T as they increasingly challenge our leadership position. Critical to maintaining the position of the US in the world will be both the amount and the quality of our long-term fundamental research in science, engineering, and medicine. The National Academies outlines the value of basic science in a series of twenty pamphlets on such research that has led in the past to breakthroughs with great human and economic benefit. Three of the 20 highlighted examples were the global positioning system, modern communications, and the antiviral therapy for AIDS. (See www.beyonddiscovery.org.) Exactly how future advances in our fundamental understanding of the universe will lead to such benefits can never be predicted in advance. Nevertheless, based on past experience, we can

confidently expect striking breakthroughs to emerge from such research that are completely unimaginable now.

Although the subject of basic science funding is not a focus of today's hearing, the House Science, Space, and Technology Committee has long emphasized the critical importance of our investments in America's future through governmental support of fundamental, long-term research. This is an investment that has remained stagnant in the U.S., while other nations are increasing their research intensity at an alarming pace. According to the AAAS, this type of investment will have decreased in the U.S. as a percent of GDP from 1.25 percent in 1985 to 0.87 percent of GDP in 2013. And in a ranking of total R&D spending as a share of GDP, America came in tenth in 2011, whereas we were sixth in 2001. The sequester will now make the situation considerably worse.

But this Hearing is entirely focused on the quality of U.S. scientific research and how we might improve it. I shall now proceed to address the specific questions posed.

Why is the integrity of scientific results and data sharing so important for both the scientific community and the general public?

Science is a remarkable community endeavor, in which a reliable body of knowledge about how the world works, called Science (with an upper case S), is built up over time from the many small bits of science (with a lower case s) that is carried out by large numbers of individual scientists. The rules established for individual scientists that make it all work demand that – in return for being given the privilege of publishing any particular research finding – each scientist must provide access to the methods that he or she has used, as well as to the data, so that any one else in the world can try to repeat the work to either confirm or deny what the first scientist has claimed. Once thereby, confirmed, new knowledge is developed by building on this knowledge in novel ways, through the work of many other scientists. Integrity and data sharing are crucial, because scientists are constantly relying on the discoveries of others as they carry on their own research. Without both the integrity of scientific results and data sharing, Science cannot develop from science.

Why does the public have such a strong interest in this issue? It is because of the enormous benefits that the public derives from Science, as explained in each of the 20 case studies that I described earlier entitled "Beyond Discovery: the Path from Research to Human Benefit" (www.beyonddiscovery.org). Such benefits are precisely why governments invest so heavily in supporting scientific research for the public good. Thus the scientific community places great emphasis on promoting the highest scientific ethics, using aids such as the freely available publication from the National Academies "On Being a Scientist: A Guide to Responsible Conduct in Research" to help imbue the needed scientific values in the next generation of scientists.[1]

What factors have contributed toward a scientific culture where unreliable results are being published and data sharing is difficult?

I begin with the data sharing issue, which is the easier half of this question to answer. The others testifying today will address ways to make the scientific data that is produced by one scientist more widely accessible and reusable for other scientists. This is an important issue for all fields of science, and *Science* magazine has long strongly supported such efforts. In early 2011, we published a large special issue of the magazine entitled "Dealing with Data" that contained 14 articles on its different aspects, in fields from astronomy to genomics. In our editorial for that issue, entitled "Making Data Maximally Available," we stressed that "*Science's* policy for some time has been that "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*" (see www.sciencemag.org/site/feature/contribinfo/)." And we announced a new policy that extended this requirement "to include computer codes involved in the creation or analysis of data." [2]

In general, we feel that data availability has increased dramatically in recent decades and that more data is available now than ever before. Standards are firmer and the community norms have improved. Problems of course remain, in part due to the massive amounts of data that can now be rapidly collected. Many of these were highlighted in our "Dealing with Data" special issue.

The main current challenges that we see with regard to data sharing are:

1) Developing standards on what data to keep, inasmuch as some scientists are collecting terabytes (TB) of data daily.

2) Developing community standards for how to organize and describe the data that is kept and where exactly to deposit it.

3) Guaranteeing funding for public databases long term, so that the community and journals like ours can rely on them.

4) Developing standards for how to deal with huge datasets that have to be housed locally, and providing protocols to access the data.

5) Developing tools for interacting with the large datasets that are now increasingly provided as supplemental online information in journal publications like ours.

The other half of the Committee's question is more difficult to address. At least in large part, I believe that the concern about unreliable results being published reflects reports stating that many of the results in the field of "translational medical research" cannot be reproduced by other scientists.[3] Much of the research that cannot be reproduced aims at identifying the specific protein targets that could be useful to pharmaceutical and biotechnology companies seeking to develop new drugs.

Even though I have been a faculty member in the School of Medicine at UCSF since 1976, I have never carried out this type of research myself. Instead, like many of my colleagues, I have pursued basic mechanistic studies aimed at understanding biological

systems at the molecular level. In preparation for this testimony, I have therefore spoken to top scientists at Genentech, a very successful biotechnology company that frequently makes use of results published in the scientific literature for their own research into potential drug targets. In general, they agree with the conclusions concerning drug target reproducibility published by the Bayer HeathCare scientists in reference 3. One of the groups of Genentech scientists whom I consulted was Dr. Frederic de Sauvage, who pointed me to a paper that he published in 2008 that refuted the results of 7 earlier publications in very prestigious journals (references 1 to 7 in his paper). [4]

From this and many other discussions that I have had on this issue, I have reached a few tentative conclusions.

1) The first is that the scientific standards are lower in some subfields of science than others. For example, I am told that many published papers in medically related fields have not been officially retracted by either the journal or the authors, even though the authors have agreed with those unable to reproduce their results that their original publication is wrong. We need to develop a value system where simply "moving on from one's mistakes without publicly acknowledging them" severely damages, rather than protects, a scientific reputation.

2) Human cells are incredibly complex. Because their behavior is determined by huge networks of interlocked signaling pathways, an off-target effect -- one that is due to affecting a protein other than the intended one -- will often mimic the expected effect for a hit on the desired drug target. Every scientist should be trained to be highly suspicious about his or her own results. But a scientist whose career advancement requires finding a drug target may fail to carry out all of the many controls needed to avoid reaching a false conclusion. And the pressures on and incentives for a young researcher whose focus is finding a potential drug target can make it difficult to avoid inadvertent data selection. .

3) We are currently overemphasizing research directly aimed at finding drugs at the expense of the high quality discovery-driven basic research that is urgently needed to improve the search for disease treatments. As elegantly pointed out in a recent editorial in *Science* by Dr. Huda Zoghbi, a leading researcher in translational medicine and a member of the U.S. National Academy of Sciences: [5]

> "Science, like most human endeavors, is susceptible to fads and fashions driven by money and status; and today many highly qualified basic scientists feel compelled to jump on the "translational medicine" bandwagon. For quite some time, it has been apparent that biomedical research in the United States is more likely to get funded if it is tied to a practical outcome, such as a step toward a cure for some disorder. There is no doubt that such targeted and in-depth disease-oriented research is sorely needed. But it is at least as important to support investigators dedicated to discovery-driven basic research." She then goes on to observe that the "task of translational research is not unlike the act of translating a book from one language into another. Fluency in both languages is a given; beyond that, there must be a talent, a feel, for those concepts unique to one language or culture that cannot be directly translated

but must somehow still be conveyed. The challenge in translational medicine is that scientists are trying to translate a text with the sophistication and depth of Shakespeare using a first-grader's vocabulary and experience, because our knowledge about the functions of most pathways in various cell types, during different developmental stages, and under normal physiological conditions, is still rudimentary and piecemeal."

What issues must be considered when promoting the publication and responsible sharing of data? From your experience, what are some models that have worked?

Scientific journals like ours have an important role to play in enforcing the responsible sharing of data. As stated previously, when a scientist publishes research with us, he or she must agree that "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*." There have only been rare times when we have had to reinforce this provision with an author. In the early 1990s, several journals, including *Science*, *Nature*, and *Cell* joined together to require X-ray crystallographic data to be made publicly available in a shared database immediately upon publication (some of the scientists involved instead wanted a 1-year moratorium on this release). All genomic data must meet the same type of standard before being published. In addition, in the late 1990's, many journals started to publish data supplements and require electronic data deposition with the journal. A continuing problem is presented by huge datasets in fields where there is no public database for deposition. Here *Science* has had an archival agreement with authors. (They have to house large datasets for 5 years and we get an escrow copy). But such data storage must be paid for by the journal and the cost is a perpetual one; thus, it is not clear how long this type of service can be maintained.

A different critical need that all journals should enforce is the clear and complete presentation in each publication of all of the materials and methods that were used in the research. This goal has become much easier to attain due to the Internet, because the limited space in the printed journal is now routinely supplemented by online supplementary material that is made readily available electronically.

In December 2011, *Science* published a special issue entitled "Data Replication and Reproducibility." This is a topic that we shall return to again in the future. As have other journals, *Science* has on occasion been fooled into publishing articles that contain data that was fabricated by one or more of the authors. As soon as possible after either an honest error or a fraud is detected, the retracted papers are specifically highlighted as incorrect, so that anyone accessing the paper on our website will know that it is wrong. Although ideally a paper will be publicly retracted by its authors, the Editor-in-Chief has retracted incorrect papers in the absence of such consent.

To help protect against both data selection (scientists fooling themselves) as well as against the much rarer intentional fabrication of data, *Science* has initiated a policy to help senior scientists enforce standards in their own laboratories. As I announced in an editorial on January 1, 2010 entitled "Promoting Scientific Standards":[6]

> "*Science* will require that the senior author for each laboratory or group confirm that he or she has personally reviewed the original data generated by that unit, ascertaining that the data selected for publication in specific figures and tables have been appropriately presented. Thus, for example, a researcher who prepares a digitally processed figure displaying an assortment of electrophoretic gel separations will need to present all of the original gel data to a specified senior author, who must certify that this has been done when the manuscript is returned for revision. In this way, *Science* aims to identify a few senior authors who collectively take responsibility for all of the data presented in each published paper. Traditionally, a single individual has been asked to accept this responsibility. But the former requirement has become increasingly unrealistic, considering that a large fraction of publications now contain contributions from groups with very different expertise—and that half of the papers published in 2009 by *Science* had authors from more than one nation."

I believe that there is more that can and should be done to enforce scientific standards by the community. For example, I strongly favor the proposal that the biosketches routinely used to help evaluate an individual researcher for research support, appointments, and promotions be limited to a small number of publications, for each of which both the significance and the contribution of the individual must be carefully described. It is time to stop allowing long lists of publications to be presented, many of which (in some fields) may have contained major errors, despite having been cited extensively in the literature.

I also believe that new experiments are in order, aimed at creating a much lower barrier for reporting any serious effort to reproduce results that has failed, and insuring that such information becomes attached directly to the original publication in a way that cannot easily be missed.

To summarize, improving the quality of scientific publications will require an on-going effort by many different players in the scientific community. Scientific journals like ours will need to play leadership roles in enforcing standards. Checklists are beginning to be developed by the community to help both scientists and journals guard against the most common errors in research in selected fields like drug target development. Funding agencies can help by facilitating and rewarding the publication of failures to replicate important published results, as well as by changing the way that the biosketches in grant submissions are presented and evaluated. And research institutes and universities should place more emphasis on short courses that teach research methodology, ethics, and important technical skills such as how to avoid statistical errors to all of their research trainees. In fact, I myself will be co-teaching such an intensive two-week "minicourse" to PhD students at UCSF this coming May.

REFERENCES

1. Committee on Science, Engineering, and Public Policy, *On Being a Scientist: A Guide to Responsible Conduct in Research.* The National Academies Press, 2009
2. B. Hanson, A. Sugden, and B. Alberts, Making Data Maximally Available. *Science* 331: 649 (2011).
3. F. Prinz, T. Schlange, and K. Asadullah, Believe it or not: How much can we rely on published data on potential drug targets? *Nature Rev. Drug Discov.* 10: 712 (2011).
4. R. L. Yauch, et al., A paracrine requirement for hedgehog signalling in cancer. *Nature* 455:406 (2008).
5. H. Zoghbi, The Basics of Translation. *Science* 339:250 (2013).
6. B. Alberts, Promoting Scientific Standards. *Science* 327:12 (2010).

Witness Biography

BRUCE ALBERTS EDITOR-IN-CHIEF, SCIENCE MAGAZINE; PROFESSOR EMERITUS, UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Bruce Alberts, a prominent biochemist with a strong commitment to the improvement of science and mathematics education, serves as Editor-in-Chief of *Science* and served as one of President Obama's first three Science Envoys. Alberts is also Professor Emeritus in the Department of Biochemistry and Biophysics at the University of California, San Francisco, to which he returned after serving two six-year terms as the president of the National Academy of Sciences (NAS).

During his tenure at the NAS, Alberts was instrumental in developing the landmark National Science Education standards that have been implemented in school systems nationwide. The type of "science as inquiry" teaching we need, says Alberts, emphasizes "logical, hands-on problem solving, and it insists on having evidence for claims that can be confirmed by others. It requires work in cooperative groups, where those with different types of talents can discover them – developing self confidence and an ability to communicate effectively with others."

Alberts is also noted as one of the original authors of *The Molecular Biology of the Cell*, a preeminent textbook in the field now in its fifth edition. For the period 2000 to 2009, he served as the co-chair of the InterAcademy Council, an organization in Amsterdam governed by the presidents of 15 national academies of sciences and that was established to provide scientific advice to the world.

Committed in his international work to the promotion of the "creativity, openness and tolerance that are inherent to science," Alberts believes that "scientists all around the world must now band together to help create more rational, scientifically-based societies that find dogmatism intolerable."

Widely recognized for his work in the fields of biochemistry and molecular biology, Alberts has earned many honors and awards, including 16 honorary degrees. He currently serves on the advisory boards of more than 20 non-profit institutions, including the Gordon and Betty Moore Foundation.

Chairman BUCSHON. Thank you.

I now recognize Dr. Stodden for five minutes to present her testimony.

TESTIMONY OF DR. VICTORIA STODDEN,
ASSISTANT PROFESSOR OF STATISTICS,
COLUMBIA UNIVERSITY

Dr. STODDEN. Thank you for the privilege of addressing you, and thank you for your very lucid comments. I agree with just about everything both of you have said, and I also agree on how important this issue is. So I would like to spend my remaining time on two aspects. One is, I would like to scope the problem for you, and the second is, I would like to scope the action I think that is available to you here.

So the first thing I want to say is that there is not a crisis of integrity in terms of scientists and scientists' behavior. What has happened in science is that like all sectors of the economy, and all across America, we are taking advantage of technological revolutions. What we are doing is using far more computers, far more data-oriented and data-driven research, far more high-powered investigation in all the research all across the sciences. This isn't just in the life sciences. This is in engineering, this is in English departments who are doing word counts in Shakespeare. This is something that is really pervasive in the scientific enterprise as a whole, and this is something that is having ramifications in the way that we disseminate and communicate science. It is not a question of personal integrity.

So what this means is, to scope the issue, I think that we need to think about this issue in terms of reproducibility, so as Dr. Alberts outlined, open data itself is a very broad notion. I think this needs to be scoped to data and software required to reproduce published results, and what that means to a scientist is clear. There are details, of course, but that is something that a scientist can understand. This is something that institutions in the scientific enterprise can understand. And I reiterate that it is not just about data, it must include the codes and the software that take that data to the published results so that those results can be validated and verified.

You mentioned in your opening remarks about statistical errors, about other issues. I would like to scope the problem to this computational issue, which I believe is reflected in the language around this, digital data, and the reason for that is clarity. I agree with you that as a statistician, there are lots of statistical errors that are in the literature that are being worked out. This is in part because doing computational work is new to many fields, and I believe the core issue is sharing data, sharing code and things like sort of biological materials or the mathematics and the statistics, those will work out as corollary issues. Right now the issue needs to be scoped on data and code that allow those results to be understood, validated and reproduced by other members in the community.

So secondly, I would like to talk about the scope of action that I think is available and important for you to think about. The first thing is, as Dr. Alberts outlined, scientists are very interested in

these issues of reproducibility. As we know, it is a cornerstone. We don't accept scientific findings until there is replication, until there is validation by other people—at least that is the theory. And in my testimony, I included two articles that are in some sense manifestos from computational scientists calling for greater reproducibility. The reason computational scientists are banding together and creating these manifestos is because there is a collective action problem. It does take time to make your data available and to make your software available. It is easier to hack things up on your machine and produce a paper and never really look at the code or the data in the sense of sharing it. That does take extra time. So what this means is that scientists who want to do reproducible research and sharing the code and data that replicates their results are at a disadvantage because they don't receive credit for this right now. They generally receive credit for the publications. So steps like what Science Magazine has taken with data-sharing requirements and code-sharing requirements are extraordinary and laudable and very important. This is Science, though, our highest-impact journal, and it is much harder for lower-impact journals to demand that of the authors. But this is where the federal funding agencies come in as another lever that exerts pressure on scientists and what they are required to do.

So in these manifestos that I included in my testimony, you will see computational scientist after computational scientist calling for help in a broad sense because people who stick their nose out get it cut off and we need the federal funding agencies to work in an integrated way to help overcome this collective action problem.

Now, how does this happen? This happens through the creation of and financial support for repositories that can house code and can house data, and this is something that can't just happen, I don't believe, from added money on grants, on NIH grants and so on, that are supposed to fund these things in an ethereal way. I think this is more serious and this is something that needs to be directly confronted, more similar to a mandate when you take federal funds for your research.

Now, standards, as Dr. Alberts mentioned, the protein data bank and these institutional repositories, other institutional repositories are very important for setting standards. They come from the community level. I don't believe they come from the federal level down. But this needs to be addressed and recognized. There is no point in saying we need to have reproducibility, we need to share data, we need to share code when they don't know where to put it and there aren't ways for people to share it and access it and curate it.

So I will move to questions here.

[The prepared statement of Dr. Stodden follows:]

Testimony of
Dr. Victoria Stodden
Columbia University

Before the House Committee on Science, Space and Technology
Subcommittee on Research

Hearing On
Scientific Integrity & Transparency
March 5, 2013

Thank you Chairman Bucshon, Ranking Member Lipinski, and other members of the Subcommittee for the opportunity to speak with you today.

I am Victoria Stodden, assistant professor of statistics at Columbia University. My research is on reproducibility of results in computational science. Reproducibility is a new challenge, brought about by advances in scientific research capability due to immense changes in technology over the last two decades. It is widely recognized as a defining hallmark of science and directly impacts the transparency and reliability of findings, and is taken very seriously by the scientific community.

Federally Funded Digital Archives are Necessary for Scientific Integrity and Accelerate Scientific Discovery

Massive computation has begun a transformation of the scientific enterprise that will finish with computation absolutely central to the scientific method. From the ability to capture data, methods, create simulations, and provide dissemination mechanisms, science has gone digital. We need federally funded archives for the scientific data and software associated with research publications. Convenient access to data and software is a necessary step in enabling reproducibility in computational science, and preservation ensures reproducibility persists. Because of their broad impact, the federal agencies that fund scientific research play a key role in facilitating the dissemination and archiving of the data and software associated with scientific findings that scientists or universities cannot play on their own. Data archives that are discipline specific and directly funded are necessary for the validation of published results, and permits others to use these resources to accelerate economic and scientific competitiveness. Openly available data and methods will maximize the downstream discoveries that could be made the information contain in the data and the know-how of methods contained in the

code. This availability means curious STEM students, for example, can try their hand at replicating published results from the data and software, and learn about the science (and perhaps contribute further discoveries!).

For example other countries, such as Sweden, the U.K., the Netherlands, and Germany, are steps ahead in creating a long-term data archive for the social sciences with a standing similar to that of a national archive. This is a solution to the public good problem of access to scientific data and code. I believe separate funding is required to establish such archives in America, since using research grant funds is unpredictable and unreliable. Funding agencies need to treat this as a mandate and plan to protect data and code availability for 25 years. Archived data and code should be linked with all publications that use either of them, in order for reproducibility to be effective.

Background on the Reproducibility Issue

First, I will provide some background on the reproducibility issue. Recent technological advances have accelerated scientific research in three principal ways: increasing in our ability to collect and store vast amounts of data; increasing the computer power needed to analyze these data and perform computationally intensive science; and providing a mechanism for the rapid transmission of digital scholarly objects (such as research articles, data, or computer software) via the Internet. These three changes have revolutionized scientific research and computation is emerging as absolutely central to the scientific enterprise. In keeping with longstanding scientific norms, the scientific community has responded to these technological changes by calling for modifications of the standards of scientific communication: making available the computational aspects of the research – the code and data – that generated published scientific findings at the time of publication.[1] This is commonly called the "Reproducible Research Movement."

The communication of scientific research was established around the goal of reproducibility – providing sufficient information to other researchers so that they are able to verify the new results. This is still the overarching goal of scientific publishing, but these technological changes are requiring us to update our standards of communication. Computational steps are typically too complex and numerous to be described in the traditional scientific publication. Researchers will need to provide both the data and the code with the computational steps as a routine part of scientific publishing. In computational science today, the published research article is rarely sufficient for the findings to be validated.

[1] The *Reproducible Research* movement: see e.g. "Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science," with Yale Roundtable Participants, Computing in Science and Engineering, 12(5) 8-13, Sep./Oct. 2010 (attached); and D. Donoho et al. "Reproducible Research in Computational Harmonic Analysis," Computing in Science and Engineering, 11(1) 8-18, Jan 2009.

This is not to say published results are necessarily wrong, or that there is a lack of integrity on the part of scientists. What is happening is that access to the data and software is needed in order to validate and understand new scientific claims. In short, scientific communication needs to catch up with the recent technological changes in scientific research and this is not something any single researcher can do on their own. The scientific community is responding with piecemeal independent efforts however, including sessions on reproducibility at major scientific conferences and meetings, dedicated workshops and journal issues (see appendices), standards on data and code access requirements by journals, subject specific repositories with data deposit requirements, independently releasing data and code, and the development of software research tools to help with data and code sharing. These efforts, while immensely laudable since they do not result in direct career advancement or reward for the scientists responsible, are minuscule and largely token compared to the scale of change that needs to happen. Science is a peer-guided endeavor and these are the main options scientists have for creating change. A larger effort is needed and this is where the federal funding agencies come in.

The scientific community has been rocked by episodes like the case at Duke University where published results about a new statistical medical assessment test could not be verified prior to the start of clinical trials. In an embarrassing scandal, the trials were eventually cancelled after the underlying research was found contain errors. Many in the scientific community feel that these errors would have been caught much earlier, well before clinical trials had started, if the associated data and software were made routinely available when computational results are published.

Some scientists feel strongly enough about the importance of reproducible research they have self archived their data and code. For example, David Donoho's Wavelab package (http://www-stat.stanford.edu/~wavelab), my Sparselab package (http://sparselab.stanford.edu), and the papers contained in http://www.RunMyCode.org. The event at Duke University prompted the Institute of Medicine to produce a report requiring data and software submission, for validation and reproducibility purposes, to be submitted to the FDA prior to clinical trial approval. Their report, "Evolution of Translational Omics: Lessons Learned and the Path Forward," was released on March 23, 2012 at http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx . These and other efforts, while laudable, cannot come close to enabling reproducibility for all computational findings that are published today and going forwards. A funding agency level solution is needed.

Open Data and Software Accelerate Scientific Discovery, Innovation, and Economic Growth

For an individual researcher, making data and software available takes time. It takes time for professionals to archive and curate these objects, and to ensure they are properly linked to the published results. I believe that these efforts are both essential to the integrity of the scholarly record and vastly more efficient over the long run than the current method of publication (omitting the associated research data and code) since it is then much easier to ensure the accuracy of published scientific findings.

Making research data and software conveniently available also has valuable corollary effects beyond validating the original associated published results. Other researchers can use them for new research, linking datasets and augmenting results in other areas, or applying the software and methods to new research applications. These powerful benefits will accelerate scientific discovery. Benefits can also accrue to private industry. Again, data and software availability permit business to apply these methods to their own research problems, link with their own datasets, and accelerate innovation and economic growth.

Scientific research is not intended to produce viable market-ready products. It produces scientific knowledge about our world. When the data and code are made conveniently available this opens entirely new possibilities for others to commercialize and ready these discoveries for market. The discoveries and technologies are made openly available as part of publication. Raw facts are not subject to copyright (499 US 340 (1991)) and data can be readily open to catalyze innovation across scientific disciplines and across industry.

American competitiveness can only be increased as we increase the integrity of our scholarly record, and as we make access to scientific innovations, data, and their implementation broadly available to other researchers and to industry.

The Federal Agencies are Vital to Ensuring Reproducible Computational Science

Since January 18 of 2011 the National Science Foundation has required a two page "Data Management Plan" be submitted with every grant application. The Plan requested that the applicant explain how "the proposal will conform to NSF policy on the dissemination and sharing of research results." The NSF policy referred to follows, "NSF ... 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 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." The National Institutes for Health has a similar policy, "We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers" (NIH grants greater than $500,000 must include a data sharing plan).

If enforced, these guidelines would help shift computational research toward reproducibility. These guidelines are generally not enforced however, and I believe this is for two reasons. One, the guidelines are not well defined in terms of what constitutes data and how it should be shared. Two, sharing is costly and the funding agency should provide mechanisms and appropriate repositories for data and code deposit. At the moment, these guidelines seem like an unfunded mandate and this should change. Federal agencies should be provided with the funds to support the open availability of data and code. This should take the form of repositories maintained by the funding agencies that can curate, preserve, and make these digital scholarly objects openly available.

The OSTP Executive Memorandum Reinforces Efforts Towards Reproducible Computational Research

On February 22, 2013, the Office of Science and Technology Policy in the Whitehouse released an executive memorandum giving federal agencies with more than $100 million in research funding six months to devise plans to facilitate open access to scientific publications and scientific data. Data is defined as "digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications." Software is equally as important as data in validating computational science and I hope the committee understands "data" as referred to in the Executive Memorandum as including both data and software.

Standards for data sharing will vary by discipline and by research problem. Some areas, especially those that tend to make big capital investments in data collection devices such as telescopes or genome sequencing machines, are relatively advanced in terms of data sharing. Others have almost no experience with the issue. Since software is typically generated by the researchers themselves, organized methods for software sharing have not come into prominence. The different types of research funded by federal agencies may require different sharing requirements. What is clear is that they

will need funds and resources to support the need for open data and software. The costs of data sharing increase markedly with the amount of curation desired, for example, meta-data, variable labeling, versioning, citation and unique identifiers, archiving, repository creation, data standards, release requirements and sequestration, among others.

Barriers to Open Data and Software: A Collective Action Problem Faces Scientists

As the scientific community reaches toward reproducibility of computational science through open data and software, there are a number of barriers. The first, mentioned earlier, is that a change in the standards of research dissemination is a classic collective action problem. One person may change their behavior but unless others follow, he or she is penalized for deviating from the norm, in this case spending time on data and code release rather than more publications (publications are recognized and rewarded in scientific careers, unlike data and code production). Federal agency action is required to break this gridlock and shift the community toward data and software sharing together. Federal agency action is also required to ensure that scientists receive credit, through citation and attribution, for their data and software contributions to science. One important step was taken by the National Science Foundation in October of 2012 when it permitted grant applicants to list research products, such as citable data and software, in biographical sketches, rather than restricting the list of contributions to publications only. More steps like this should be taken, including providing citation recommendations for data and software re-use, and expectations that use of data or software be cited or claimed as original to the author.

Not all datasets or software are worthy of the same levels of curation. Curation can be costly in terms of human hours and it stands to reason that widely used datasets and software with potentially broad applicability should receive the majority of the curation resources. Provisions can be made that curation levels be increased for data or software that is used more than expected.

There must be ways for data and software users to provide feedback on difficulties they found in re-use, and ways for these corrections and improvements to be acted upon. Such a mechanism can help establish standards for data and code curation and release within a community.

The kind of sharing infrastructure associated with data and associated with software are very different. Data is typically shared as an annotated file in repository, whereas software is much more interaction, typically shared through a version control system,

perhaps with an overlay for web access such as GitHub.com (for open source software, not scientific software specifically). Reproducibility demands that we consider both the data and code associated with published computational results, and each of these have very different infrastructure needs for open accessibility. What version is used, how to manage updates and changes to data or software, what meta-data is needed, what standards apply and what documentation is expected, and how to link to the associated publication differ for data and for software.

Intellectual Property Issues in Open Scientific Data and Software

Intellectual Property Law comes to bear on both scientific data and software. Longstanding scientific norms encourage reproducible research, and scientists find it natural to openly share their findings, data, software, such that results may be understood and validated by others. Copyright adheres by default to both the scientific manuscript and software, and adhere to the original "selection and arrangement" of the data, although not to the raw facts themselves. This has resulted in efforts to apply open licensing to scholarly digital objects such that they may be shared as is natural in the scientific community: use my work, validate it, build on it, but make sure I am given appropriate citation for my contributions.[2] A broad fair use exception for scientific research that includes data and software would align Intellectual Property Law with scientific norms and needs for reproducibility, and maximize future discoveries and use of the data and code, both within the research community and within industry.

With software established as an indispensible tool in scientific discovery, a computational researcher can be faced with an unexpected conflict: conform to scientific norms of reproducibility and reveal the software that generated the results, or seek a software patent and license access to the code. Traditionally the research methodology was contained in the published report, but in the computational sciences methodology is encapsulated within a potentially patentable medium, software. It is important that the needs of science, especially those of reproducibility, remain paramount to patenting in order to promote high integrity scientific research. Making data and code available in an archives the goals for transparency and technology transfer embodied in the Bayh-Dole Act, and can be done in a way that is coordinated and harmonious between the relevant funding agencies.

[2] For discussions of open licensing for computational scientific research see e.g. V. Stodden, "The Legal Framework for Reproducible Scientific Research: Licensing and Copyright," Computing in Science and Engineering, 11(1), 2009; and see also V. Stodden, "Enabling Reproducible Research: Open Licensing for Scientific Innovation," International Journal of Communications Law and Policy, Issue 13, 2009. Available at http://ijclp.net/old_website/article.php?doc=1&issue=13_2009

Conclusion

The issue of reproducibility in computational science cuts across all manner of disciplines and research areas, from the liberal arts to engineering. The solutions are not obvious, but it is clear they can emerge with experience and action. It is imperative that data and code are made conveniently available with published research findings. Data and software availability do not, by themselves, ensure reproducibility of published computational findings, but they are an essential step toward the solution.

Thank you for the opportunity to testify this morning. I look forward to answering any questions you may have.

Setting the Default to Reproducible

Reproducibility in Computational and Experimental Mathematics

Developed collaboratively by the ICERM workshop participants[1]

Compiled and edited by the Organizers

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein

Abstract

Science is built upon foundations of theory and experiment validated and improved through open, transparent communication. With the increasingly central role of computation in scientific discovery this means communicating all details of the computations needed for others to replicate the experiment, i.e. making available to others the associated data and code. The "reproducible research" movement recognizes that traditional scientific research and publication practices now fall short of this ideal, and encourages all those involved in the production of computational science – scientists who use computational methods and the institutions that employ them, journals and dissemination mechanisms, and funding agencies – to facilitate and practice really reproducible research.

This report summarizes discussions that took place during the ICERM Workshop on Reproducibility in Computational and Experimental Mathematics, held December 10-14, 2012. The main recommendations that emerged from the workshop discussions are:

1. It is important to promote a culture change that will integrate computational reproducibility into the research process.
2. Journals, funding agencies, and employers should support this culture change.
3. Reproducible research practices and the use of appropriate tools should be taught as standard operating procedure in relation to computational aspects of research.

The workshop discussions included presentations of a number of the diverse and rapidly growing set of software tools available to aid in this effort. We call for a broad implementation of these three recommendations across the computational sciences.

Introduction

The emergence of powerful computational hardware, combined with a vast array of computational software, presents unprecedented opportunities for researchers in mathematics and science. Computing is rapidly becoming the backbone of both theory and experiment, and essential in data analysis, interpretation, and inference.

Unfortunately the scientific culture surrounding computational work has evolved in ways that often make it difficult to verify findings, efficiently build on past research, or even to apply the basic tenets of the scientific method to computational procedures. Bench scientists are taught to keep careful lab notebooks documenting all aspects of the materials and

[1]For a list of participants see Appendix H or the workshop webpage `http://icerm.brown.edu/tw12-5-rcem`.

Version of February 16, 2013.

methods they use including their negative as well as positive results, but computational work is often done in a much less careful, transparent, or well-documented manner. Often there is no record of the workflow process or the code actually used to obtain the published results, let alone a record of the false starts. This ultimately has a detrimental effect on researchers' own productivity, their ability to build on past results or participate in community efforts, and the credibility of the research among other scientists and the public [6].

There is increasing concern with the current state of affairs in computational science and mathematics, and growing interest in the idea that doing things differently can have a host of positive benefits that will more than make up for the effort required to learn new work habits. This research paradigm is often summarized in the computational community by the phrase "reproducible research." Recent interest and improvements in computational power have led to a host of new tools developed to assist in this process. At the same time there is growing recognition among funding agencies, policy makers, and the editorial boards of scientific journals of the need to support and encourage this movement. [2]. A number of workshops have recently been held on related topics, including a Roundtable at Yale Law School [14] a workshop as part of the Applied Mathematics Perspectives 2011 conference [2, 7], and several minisymposia at other conferences, including SIAM Conferences on Computational Science and Engineering 2011 and ICIAM 2011.

The ICERM Workshop on Reproducibility in Computational and Experimental Mathematics, held December 10-14, 2012, provided the first opportunity for a broad cross section of computational scientists and mathematicians, including pure mathematicians who focus on experimental mathematics or computer-assisted proofs, to discuss these issues and brainstorm ways to improve on current practices. The first two days of the workshop focused on introducing the themes of the meeting and discussing policy and cultural issues. In addition to introductory talks and open discussion periods, there were panels on funding agency policies and on journal and publication policies. The final three days featured many talks on software tools that help achieve reproducibility and other more technical topics in the mornings. Afternoons were devoted to breakout groups discussing specific topics in more depth, which resulted in recommendations and other outcomes. Breakout group topics included: reproducibility tools, funding policies, publication policies, numerical reproducibility, taxonomy of terms, reward structure and cultural issues, and teaching reproducible research techniques. [3] We also held a tutorial on version control the day before the official start of the workshop. [4]

Both in the workshop and in this report the terms "reproducible research" and "reproducibility" most often refer to the ability to recreate computational results from the data and code used by the original researcher [11]. This is related to but distinct from both the notions of "numerical reproducibility" of computational results, referring to when the same program may give different results due to hardware or compiler issues, particular in the context of parallel computing, and "repeatability," when an experiment is conducted independently from first principles. A taxonomy of reproducibility concepts is developed in Appendix A and a discussion of numerical reproducibility appears in Appendix B.

[2] See National Science Foundation Data Management Plan `http://`, ACM Publications Policy `http://`

[3] See the workshop program at `http://icerm.brown.edu/tw12-5-rcem`.

[4] `http://icerm.brown.edu/tw12-5-rcem-tutorial`.

About this document

This document reports on the three main recommendations emerging from the workshop discussions:

1. It is important to promote a culture change that will integrate computational reproducibility into the research process.
2. Journals, funding agencies, and employers should support this culture change.
3. Reproducible research practices and the use of appropriate tools should be taught as standard operating procedure in relation to computational aspects of research.

The recommendations are each discussed in turn in the three sections of this document, and we include five appendices that develop important topics in further detail. Besides the appendices mentioned above on taxonomy and numerical reproducibility, there are appendices on best practices for publishing research, the state of reproducibility in experimental mathematics, and tools to aid in reproducible research. An initial draft of this document was presented to participants and discussed on the final day of the workshop, and participants were able to give input on the final draft before submission.

In addition to this document, a number of other products emerged from the workshop. Video of the talks is available at `http://icerm.brown.edu/video_archive`, and numerous topical references were collected on the workshop wiki [5]. The workshop webpage and the wiki also contain participant thought pieces, slides from the talks, and breakout group reports. Readers are invited to contribute to the wiki. A snapshot of the wiki is appended at the end of the report as Figure 1.

1. Changing the Culture and Reward Structure

For reproducibility to be fostered and maintained, workshop participants agreed that cultural changes need to take place within the field of computationally based research that instill the open and transparent communication of results as a default. Such a mode will increase productivity — less time wasted in trying to recover output that was lost or misplaced, less time wasted trying to double-check results in the manuscript with computational output, and less time wasted trying to determine whether other published results (or even their own) are truly reliable. Open access to any data used in the research and to both primary and auxiliary source code also provides the basis for research to be communicated transparently creating the opportunity to build upon previous work, in a similar spirit as open software provided the basis for Linux. Code and data should be made available under open licensing terms as discussed in Appendix F. [9] This practice enables researchers both to benefit more deeply from the creative energies of the global community and to participate more fully in it. Most great science is built upon the discoveries of preceding generations and open access to the data and code associated with published computational science allows this tradition to continue. Researchers should be encouraged to recognize the potential benefits of openness and reproducibility.

[5] Available at `http://is.gd/RRlinks`, see Figure 1.

It is also important to recognize that there are costs and barriers to shifting to a practice of reproducible research, particularly when the culture does not recognize the value of developing this new paradigm or the effort that can be required to develop or learn to use suitable tools. This is of particular concern to young people who need to earn tenure or secure a permanent position. To encourage more movement towards openness and reproducibility, it is crucial that such work be acknowledged and rewarded. The current system, which places a great deal of emphasis on the number of journal publications and virtually none on reproducibility (and often too little on related computational issues such as verification and validation), penalizes authors who spend extra time on a publication rather than doing the minimum required to meet current community standards. Appropriate credit should given for code and data contributions including an expectation of citation. Another suggestion is to instantiate yearly award from journals and/or professional societies, to be awarded to investigators for excellent reproducible practice. Such awards are highly motivating to young researchers in particular, and potentially could result in a sea change in attitudes. These awards could also be cross-conference and journal awards; the collected list of award recipients would both increase the visibility of researchers following good practices and provide examples for others.

More generally, it is unfortunate that software development and data curation are often discounted in the scientific community, and programming is treated as something to spend as little time on as possible. Serious scientists are not expected to carefully test code, let alone document it, in the same way they are trained to properly use other tools or document their experiments. It has been said in some quarters that writing a large piece of software is akin to building infrastructure such as a telescope rather than a creditable scientific contribution, and not worthy of tenure or comparable status at a research laboratory. This attitude must change if we are to encourage young researchers to specialize in computing skills that are essential for the future of mathematical and scientific research. We believe the more proper analog to a large scale scientific instrument is a supercomputer, whereas software reflects the intellectual engine that makes the supercomputers useful, and has scientific value beyond the hardware itself. Important computational results, accompanied by verification, validation, and reproducibility, should be accorded with honors similar to a strong publication record [7].

Several tools were presented at the workshop that enable users to write and publish documents that integrate the text and figures seen in reports with code and data used to generate both text and graphical results, such as IPython, Sage notebooks, Lepton, knitr, and Vistrails. Slides for these talks are available on the wiki [1] and Appendix E discusses these and other tools in detail.

The following two sections and the appendices outline ideas from the workshop on ways in which journals, funding agencies, and employers can support reproducibility.

2. Funding Agencies, Journals, Employers Should Support This Change

Incentives in scholarly research are influenced by three main sources, the funding agency, dissemination processes such as journals, and employers such as those on tenure committees and lab managers. The workshop discussions mentioned the role of each of them in shifting to a culture of reproducible computational research.

The Role of Funding Agency Policy

Workshop participants suggested that funding sources, both government agencies and private foundations, consider establishing some reasonable standards for proposals in the arena of mathematical and computational science. If such standards become common among related agencies this would significantly simplify the tasks involved in both preparing and reviewing proposals, as well as supporting a culture change toward reproducible computational research.

For example, workshop participants recommend that software and data be "open by default" unless it conflicts with other considerations. Proposals involving computational work might be required to provide details such as:

- Extent of computational work to be performed.
- Platforms and software to be utilized.
- Reasonable standards for dataset and software documentation, including reuse (some agencies already have such requirements [8]).
- Reasonable standards for persistence of resulting software and dataset preservation and archiving.
- Reasonable standards for sharing resulting software among reviewers and other researchers.

In addition, we suggest that funding agencies might add "reproducible research" to the list of specific examples that proposals could include in their requirements such as "Broader Impact" statements. Software and dataset curation should be explicitly included in grant proposals and recognized as a scientific contribution by funding agencies. Templates for data management plans could be made available that include making software open and available, perhaps by funding agencies, or by institutional archiving and library centers. [6]

Participants also suggested that statements from societies and others on the importance of reproducibility could advance the culture change. In addition, funding agencies could provide support for training workshops on reproducibility, and cyberinfrastructure for reproducibility at scale, for both large projects and long-tail research efforts. Funding agencies are key to the promotion of a culture that embraces reproducible research, due to their central importance in the research process. We turn to journals next, and then employers.

The Role of Journal Policy

There is a need to produce a set of "best practices" for publication of computational results i.e. any scientific results in which computation plays a role, for example in empirical research, statistical analysis, or image processing. We recommend that a group representing several professional societies in the mathematical sciences be formed to develop a set of best practices for publication of research results. Such guidelines would be useful

[6] For examples see `http://scholcomm.columbia.edu/data-management/data-management-plan-templates/`, `http://www2.lib.virginia.edu/brown/data/NSFDMP.html`

to the editorial boards of many journals, as well as to authors, editors, and referees who are concerned about promoting reproducibility. Best practices may be tailored to different communities, but one central concern, for which there was almost unanimous agreement by the workshop participants, is the need for full disclosure of salient details regarding software and data use. This should include specification of the dataset used (including URL and version), details of the algorithms employed (or references), the hardware and software environment, the testing performed, etc., and would ideally include availability of the relevant computer code with a reasonable level of documentation and instructions for repeating the computations performed to obtain the results in the paper. [7].

There is also a need for better standards on how to include citations for software and data in the references of a paper, instead of inline or as footnotes. Proper citation is essential both for improving reproducibility and in order to provide credit for work done developing software and producing data, which is a key component in encouraging the desired culture change [7].

Workshop participants agreed that it is important that a set of standards for reviewing papers in the computational arena be established. Such a set of standards might include many or all of the items from a "best practices" list, together with a rational procedure for allowing exceptions or exclusions. Additionally, provisions are needed to permit referees to obtain access to auxiliary information such as computer codes or data, and the ability to run computational tests of the results in submitted papers, if desired.

Different journals may well adopt somewhat different standards of code and data disclosure and review [12], but it is important that certain minimal standards of reproducibility and rigor be maintained in all refereed journal publications. Along these lines, it may be desirable for the computational claims of a manuscript to be verifiable at another site such as RunMyCode.org, or on another computer system with a similar configuration.

Some related issues in this arena include: (a) anonymous versus public review, (b) persistence (longevity) of code and data that is made publicly available, and (c) how code and data can be "watermarked," so that instances of uncited usage (plagiarism) can be detected and provenance better established (d) how to adjudicate disagreements that inevitably will arise.

Very rigorous verification and validity testing, along with a full disclosure of computational details, should be required of papers making important assertions, such as the computer-assisted proof of a long-standing mathematical result, new scientific breakthroughs, or studies that will be the basis for critical policy decisions [13].

Proper consideration of openness constraints can enable a larger community to participate in the goals of reproducible research. This can include issues such as copyright, patent, medical privacy, personal privacy, security, and export issues. This is discussed further in Appendix F.

It was recognized that including such details in submitted manuscripts (or, at the least, in supplementary materials hosted by the journal) is a significant departure from established practice, where few such details are typically presented. But these changes will be required if the integrity of the computational literature is to be maintained. Computational approaches have become central to science and cannot be completely documented and

[7] See for example `http://software.ac.uk/so-exactly-what-software-did-you-use`

transparent without the full disclosure of computational details. Appendix D contains the full list of workshop suggestions.

The Role of Employers and Research Managers

The third source of influence on the research process stems from employers – tenure and promotion committees and research managers at research labs. Software and dataset contributions, as described in the previous two subsections, should be rewarded as part of expected research practices. Data and code citation practices should be recognized and expected in computational research. Prizes for reproducible research should also be recognized in tenure and promotion decisions.

Institutional libraries can also play a role in supporting a culture change toward reproducible research. As mentioned above, they can and do provide template data management plans, but they are also highly experienced in archiving, stewardship and dissemination of scholarly objects. Greater coordination between departments and the institute's library system could help provide the support and resources necessary to manage and maintain digital scholarly output, including datasets and code [4].

3. Teaching Reproducibility Skills

Proficiency in the skills required to carry out reproducible research in the computational sciences should be taught as part of the scientific methodology, along with teaching modern programming and software engineering techniques. This should be a standard part of any computational research curriculum, just as experimental or observational scientists are taught to keep a laboratory notebook and follow the scientific method. Adopting appropriate tools (see Appendix E) should be encouraged, if not formally taught, during the training and mentoring of students and postdoctoral fellows. Without a change in culture and expectations at this stage, reproducibility will likely never enter the mainstream of mathematical and scientific computing.

We see at least five separate ways in which these skills can be taught: full academic courses, incorporation into existing courses, workshops and summer schools, online courses or self-study materials, and last but certainly not least, teaching-by-example on the part of mentors.

Although a few full-scale courses on reproducibility have been attempted (see the wiki for links), we recognize that adding a new course to the curriculum or the students' schedules is generally not feasible. It seems more effective as well as more feasible to incorporate teaching the tools and culture of reproducibility into existing courses on various subjects, concentrating on the tools most appropriate for the domain of application. For example, several workshop participants have taught classes in which version control is briefly introduced and then students are required to submit homework by pushing to a version control repository as a means of encouraging this habit.

A list of potential curriculum topics on reproducibility are listed in Appendix G. Ideally, courseware produced at one institution should be shared with others under an appropriate open license.

Conclusion

The goal of computational reproducibility is to provide a solid foundation to computational science, much like a rigorous proof is the foundation of mathematics. Such a foundation permits the transfer of knowledge that can be understood, implemented, evaluated, and used by others. This reports discusses the efforts of participants and organizers of the ICERM workshop on "Reproducibility in Computational and Experimental Mathematics" to formulate steps toward the ideal of reproducible computational research. We identified three key recommendations emerging from workshop discussions, calling for a culture change toward reproducible research, mapping roles for funding agencies, journals, and employers to support this change, and emphasizing that methods and best practices for reproducible research must be taught. We also include detailed appendices on related issues that arose in the workshop discussions, including a taxonomy of terms, numerical reproducibility, best practices for publishing reproducible research, a summary of the state of experimental mathematics, and tools to aid in reproducible research. To capture the phenomenal level of engagement by workshop participants, we collate further information, including their talk slides, thought pieces, and further references on the workshop wiki.

References

[1] Applied mathematics perspectives 2011 workshop on reproducible research: Tools and strategies for scientific computing. `http://wiki.stodden.net/AMP2011/`, 2012.

[2] Applied mathematics perspectives 2011 workshop on reproducible research: Tools and strategies for scientific computing. `http://stodden.net/AMP2011/`, 2009.

[3] David H. Bailey, Roberto Barrio, and Jonathan M. Borwein. High precision computation: Mathematical physics and dynamics. *Applied Mathematics and Computation*, 218:10106–10121, 2012.

[4] Christine Borgman. Research data, reproducibility, and curation. In *Digital Social Research: A Forum for Policy and Practice, Oxford Internet Institute Invitational Symposium*, 2012.

[5] Jonathan M. Borwein and David H. Bailey. *Mathematics by Experiment: Plausible Reasoning in the 21st Century*. A K Peters (Taylor and Francis), Natick, MA, 2008.

[6] David Donoho, Arian Maleki, Morteza Shahram, Victoria Stodden, and Inam Ur Rahman. Reproducible research in computational harmonic analysis. *Computing in Science and Engineering*, 11, January 2009.

[7] Randall LeVeque, Ian Mitchell, and Victoria Stodden. Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science and Engineering*, pages 13–17, July 2012.

[8] Brian Matthews, Brian McIlwrath, David Giaretta, and Ester Conway. The significant properties of software: A study. `http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf`, 2008.

[9] Victoria Stodden. Enabling reproducible research: Licensing for scientific innovation. *International Journal of Communications Law and Policy*, pages 1–25, 2009.

[10] Victoria Stodden. The legal framework for reproducible scientific research. *Computing in Science and Engineering*, 11, January 2009.

[11] Victoria Stodden. Trust your science? open your data and code. *Amstat News*, 2011.

[12] Victoria Stodden, Peixuan Guo, and Zhaokun Ma. How journals are adopting open data and code policies. In *The First Global Thematic IASC Conference on the Knowledge Commons: Governing Pooled Knowledge Resources*, 2012.

[13] Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Steven H. D. Haddock Richard T. Guy, Katy Huff, Ian M. Mitchell, Mark Plumbley, Ben Waugh, Ethan P. White, and Paul Wilson. Best practices for scientific computing.

[14] Yale law school, data and code sharing roundtable. `http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/`, 2009.

Appendices

These appendices contain some additional material arising from workshop discussions. We have avoided including long lists of references in these appendices. Instead, many links have been collected and categorized on the workshop wiki, which can be referred to for more examples, additional tools, articles, editorials, etc.[8]

A Terminology and Taxonomy

The terms "reproducible research" and "reproducibility" are used in many different ways to encompass diverse aspects of the desire to make research based on computation more credible and extensible. Lively discussion over the course of the workshop has led to some suggestions for terminology, listed below. We encourage authors who use such terms in their work to clarify what they mean in order to avoid confusion.

There are several possible levels of reproducibility, and it seems valuable to distinguish between the following:

- *Reviewable Research.* The descriptions of the research methods can be independently assessed and the results judged credible. (This includes both traditional peer review and community review, and does not necessarily imply reproducibility.)
- *Replicable Research.* Tools are made available that would allow one to duplicate the results of the research, for example by running the authors' code to produce the plots shown in the publication. (Here tools might be limited in scope, e.g., only essential data or executables, and might only be made available to referees or only upon request.)
- *Confirmable Research.* The main conclusions of the research can be attained independently without the use of software provided by the author. (But using the complete description of algorithms and methodology provided in the publication and any supplementary materials.)
- *Auditable Research.* Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.
- *Open or Reproducible Research.* Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.

Other terms that often arise in discussing reproducibility have specific meanings in computational science. In particular the widely-used acronym V&V (verification & valida-

[8] For the wiki see `http://icerm.brown.edu/tw12-5-rcem-wiki.php` or `http://is.gd/RRlinks`

tion) makes it difficult to use "verify" or "validate" more generally. These terms are often defined as follows:

- *Verification.* Checking that the computer code correctly solves the mathematical problem it claims to solve. (Does it solve the equation right?)
- *Validation.* Checking that the results of a computer simulation agree with experiments or observations of the phenomenon being studied. (Does it solve the right equation?)

The term "Uncertainty Quantification (UQ)" is also commonly used in computational science to refer to various approaches to assessing the effects of all of then uncertainties in data, models, and methods on the final result, which is often then viewed as a probability distribution on a space of possible results rather than a single answer. This is an important aspect of reproducibility in situations where exact duplication of results cannot be expected for various reasons.

The *provenance* of a computational result is a term borrowed from the art world, and refers to a complete record of the source of any raw data used, the computer programs or software packages employed, etc. The concept of provenance generally includes a record of changes that the dataset or software has undergone.

B Numerical Reproducibility

Numerical round-off error and numerical differences are greatly magnified as computational simulations are scaled up to run on highly parallel systems. As a result, it is increasingly difficult to determine whether a code has been correctly ported to a new system, because computational results quickly diverge from standard benchmark cases. And it is doubly difficult for other researchers, using independently written codes and distinct computer systems, to reproduce published results.

One solution is to utilize some form of higher precision arithmetic, such as Kahan's summation or "double-double" arithmetic. In many cases, such higher precision arithmetic need only be used in global summations or other particularly sensitive operations, so that the overall runtime is not greatly increased. Such measures have dramatically increased reproducibility in various codes, ranging from climate simulations to computational physics applications [3].

But it is clear that this solution will not work for all applications. Other approaches include interval arithmetic (which potentially can provide provable bounds on final results), affine arithmetic (which works better than interval arithmetic in some cases), and also some proposed new tools, currently under development at U.C. Berkeley, that can pin down numerically sensitive sections of code and take corrective actions. in any event, additional study and research is in order. Certainly the available tools for high-precision computation need to be significantly refined so as to be usable and highly reliable for a wide range of users.

It is clear that these issues must be addressed with greater diligence by authors of all manuscripts presenting results of numeric computations. They must be more careful

to state exactly what levels of numeric precision (32-bit, 64-bit or higher precision) have been used, and to present evidence that their selected precision is adequate to achieve a reasonable level of reproducibility in their results.

One of the foundations of reproducibility is how to deal with (and set standards for) difficulties such as numerical round-off error and numerical differences when a code is run on different systems or different numbers of processors. Such difficulties are magnified as problems are scaled up to run on very large, highly parallel systems.

Computations on a parallel computer system present particularly acute difficulties for reproducibility since, in typical parallel usage, the number of processors may vary from run to run. Even if the same number of processors is used, computations may be split differently between them or combined in a different order. Since computer arithmetic is not commutative, associative, or distributive, achieving the same results twice can be a matter of luck. Similar challenges arise when porting a code from one hardware or software platform to another.

The IEEE Standards for computer arithmetic resulted in significant improvements in numerical reproducibility on single processors when they were introduced in the 1970s. Some work is underway on extending similar reproducibility to parallel computations, for example in the Intel Mathematics Kernel Library (MKL), which can use used to provide parallel reproducibility for mathematical computations.

Additional issues in this general arena include: (a) floating-point standards and whether they being adhered to on the platform in question, (b) changes that result from different levels of optimization, (c) changes that result from employing library software, (d) verification of results, and (e) fundamental limits of numerical reproducibility, what are reasonable expectations and what are not.

The foundation of numerical reproducibility is also grounded in the computing hardware and in the software stack. Studies on silent data corruption (SDC) have documented SDC in field testing, as discussed in some of the references on the wiki.

Field data on supercomputer DRAM memory failures have shown that advanced error correcting codes (ECC) are required and technology roadmaps suggest this problem will only get worse in the coming years. Designing software that can do some or all of identification, protection, and correction will become increasingly important. Still, there is much work being done to quantify the problem on current and next generation hardware and approaches to addressing it. Several United States and international governmental reports have been produced on the need for, outlining ongoing research in, and proscribing roadmaps.

These foundational components set a limit to the achievable reproducibility and make us aware that we must continually assess just how reproducible our methods really are.

C The State of Experimental Mathematics

Automatic theorem proving has now achieved some truly impressive results such as fully formalized proofs of the Four color theorem and the Prime number theorem. While such tools currently require great effort, one can anticipate a time in the distant future when all

truly consequential results are so validated.

The emerging discipline of experimental mathematics, namely the application of high-performance computing technology to explore research questions in pure and applied mathematics, raises numerous issues of computational reproducibility [5]. Experimental mathematics research often press the state-of-the-art in very high precision computation (often hundreds or thousands of digits), symbolic computation, graphics and parallel computation. There is a need to carefully document algorithms, implementation techniques, computer environments, experiments and results, much as with other forms of computation-based research. Even more emphasis needs to be placed on aspects of such research that are unique to this discipline: (a) Are numeric precision levels (often hundreds or even thousands of digits) adequate for the task at hand? (b) What independent consistency checks have been employed to validate the results? (c) If symbolic manipulation software was employed (e.g., Mathematica or Maple), exactly which version was used?[9] (c) Have numeric spot-checks been performed to check derived identities and other results? (d) Have symbolic manipulations been validated, say by using two different symbolic manipulation packages?

Such checks are often required, because even the best symbolic and numeric computation packages have bugs and limitations, which bugs are often only exhibited when doing state-of-the-art research computations. Workshop participants identified numerous instances of such errors in their work, underscoring the fact that one cannot place unquestioned trust in such results.

D Best Practices for Publishing Research

Publishing can take many forms – traditional journal publication is one avenue but other electronic options are increasingly being used. Traditional publications are also frequently complemented by "supplementary materials" posted on a journal's website or in other archival-quality data or code repositories.

A number of suggestions were made regarding best practices for publications of research results. To aid in reproducibility, the available materials should ideally contain:

- A precise statement of assertions to be made in the paper.
- A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.
- Complete statements of, or references to, every algorithm employed.
- Salient details of auxiliary software (both research and commercial software) used in the computation.
- Salient details of the test environment, including hardware, system software and the number of processors utilized.

[9] Indeed, one needs to know which precise functions were called, with what parameter values and environmental settings?

- Salient details of data reduction and statistical analysis methods.
- Discussion of the adequacy of parameters such as precision level and grid resolution.
- Full statement (or at least a valid summary) of experimental results.
- Verification and validation tests performed by the author(s).
- Availability of computer code, input data and output data, with some reasonable level of documentation.
- Curation: where are code and data available? With what expected persistence and longevity? Is there a site for site for future updates, e.g. a version control repository of the code base?
- Instructions for repeating computational experiments described in the paper.
- Terms of use and licensing. Ideally code and data "default to open", i.e. a permissive re-use license, if nothing opposes it.
- Avenues of exploration examined throughout development, including information about negative findings.
- Proper citation of all code and data used, including that generated by the authors.

Several publications have adopted some requirements for reproducibility (e.g., Biostatistics, TOMS, IPOL, or conferences such as SIGMOD). In addition to those discussed in the main article, some other recommendations arose in discussions and break-out groups to change the culture in relation to reproducibility in publications. Journals or other publications could offer certifications of reproducibility that would kite-mark a paper satisfying certain requirements, as done by the journal Biostatistics, for example. Certification could also come from an independent entity such as RunMyCode.org. Journals could also create reproducible overlay issues for journals that collect together reproducible papers. Linking publications to sites where code and data are hosted will help shift toward reproducible research. For example, the *SIAM Journal on Imaging Science* provides cross-referencing with the peer-reviewed journal *Image Processing On Line (IPOL)* and encourage authors to submit software to IPOL. Other sites such as RunMyCode.org or Wakari might be used in a similar way. Finally, all code and data should be labeled with author information.

E Tools to aid in reproducible research

A substantial portion of the workshop focused on tools to aid in replicating past computational results (by the same researcher and/or by others) and to assist in tracking the provenance of results and the workflow used to produce figures or tables, along with discussion of the policy issues that arise in connection with this process.

Some tools are aimed at easing literate programming and publishing of computer code, either as commented code or in notebook environments. Other tools help capture the provenance of a computational result and/or the complete software software environment

used to run a code. Version control systems have been around for decades, but new tools facilitate the use of version control both for collaborative work and for archiving projects along with the complete history. Collaborative high performance computational tools, while still infrequently used, now allow researchers at multiple locations to explore climate or ocean flow models in real time. Less sophisticated but instructive applets generated in geometry or computer algebra packages can easily be shared and run over the internet. We gives an overview of tools in these various categories. A list of links to these tools and many others can also be found on the wiki.

Literate programming, authoring, and publishing tools. These tools enable users to write and publish documents that integrate the text and figures seen in reports with code and data used to generate both text and graphical results. In contrast to notebook-based tools discussed below, this process is typically not interactive, and requires a separate compilation step. Tools that enable literate programming include both programming-language-specific tools such as WEB, Sweave, and knitr, as well as programming-language-independent tools such as Dexy, Lepton, and noweb. Other authoring environments include SHARE, Doxygen, Sphinx, CWEB, and the Collage Authoring Environment.

Tools that define and execute structured computation and track provenance. Provenance refers to the tracking of chronology and origin of research objects, such as data, source code, figures, and results. Tools that record provenance of computations include VisTrails, Kepler, Taverna, Sumatra, Pegasus, Galaxy, Workflow4ever, and Madagascar.

Integrated tools for version control and collaboration. Tools that track and manage work as it evolves facilitate reproducibility among a group of collaborators. With the advent of version control systems (e.g., Git, Mercurial, SVN, CVS), it has become easier to track the investigation of new ideas, and collaborative version control sites like Github, Google Code, BitBucket, and Sourceforge enable such ideas to be more easily shared. Furthermore, these web-based systems ease tasks like code review and feature integration, and encourage collaboration.

Tools that express computations as notebooks. These tools represent sequences of commands and calculations as an interactive worksheet with pretty printing and integrated displays, decoupling content (the data, calculations) from representation (PDF, HTML, shell console), so that the same research content can be presented in multiple ways. Examples include both closed-source tools such as MATLAB (through the publish and app features), Maple, and Mathematica, as well as open-source tools such as IPython, Sage, RStudio (with knitr), and TeXmacs.

Tools that capture and preserve a software environment. A major challenge in reproducing computations is installing the prerequisite software environment. New tools make it possible to exactly capture the computational environment and pass it on to someone who wishes to reproduce a computation. For instance, VirtualBox, VMWare, or Vagrant can be used to construct a virtual machine image containing the environment. These images are typically large binary files, but a small yet complete text description (a recipe to create the virtual machine) can be stored in their place using tools like Puppet, Chef, Fabric, or shell scripts. Blueprint analyzes the configuration of a machine and outputs its text description. ReproZip captures all the dependencies, files and binaries of the experiment, and also creates a workflow specification for the VisTrails system in order to make

the execution and exploration process easier. Application virtualization tools, such as CDE (Code, Data, and Environment), attach themselves to the computational process in order to find and capture software dependencies.

Computational environments can also be constructed and made available in the cloud, using Amazon EC2, Wakari, RunMyCode.org and other tools. VCR, or Verifiable Computational Research, creates unique identifiers for results that permits their reproduction in the cloud.

Another group are those tools that create an integrated software environment for research that includes workflow tracking, as well as data access and version control. Examples include Synapse/clearScience and HUBzero including nanoHUB.

Interactive theorem proving systems for verifying mathematics and computation. "Interactive theorem proving", a method of formal verification, uses computational proof assistants to construct formal axiomatic proofs of mathematical claims. Examples include coq, Mizar, HOL4, HOL Light, ProofPowerHOL, Isabelle, ACL2, Nuprl, Veritas, and PVS. Notable theorems such as the Four Color Theorem have been verified in this way, and Thomas Hales's Flyspeck project, using HOL Light and Isabelle, aims to obtain a formal proof of the Kepler conjecture. Each one of these projects produces machine-readable and exchangeable code that can be integrated in to other programs. For instance, each formula in the web version of NIST's authoritative Digital Library of Mathematical Functions may be downloaded in TeX or MathML (or indeed as a PNG image) and the fragment directly embedded in an article or other code. This dramatically reduces chances of transcription error and other infelicities being introduced.

While we have organized these tools into broad categories, it is important to note that users often require a collection of tools depending on their domain and the scope of reproducibility desired. For example, capturing source code is often enough to document algorithms, but to replicate results on high-performance computing resources, for example, the build environment or hardware configuration are also important ingredients. Such concerns have been categorized in terms of the depth, portability, and coverage of reproducibility desired.

The development of software tools enabling reproducible research is a new and rapidly growing area of research. We think that the difficulty of working reproducibly will be significantly reduced as these and other tools continue to be adopted and improved. The scientific, mathematical, and engineering communities should encourage the development of such tools by valuing them as significant contributions to scientific progress.

F Copyright and licensing

The copyright issue is pervasive in software and can affect data, but solutions have been created through open licensing and public domain dedication. Copyright adhere to all software and scripts as an author types, and care must be taken when sharing these codes that permission is given for others to copy, reproduce, execute, modify and otherwise use the code. For reproducibility of scientific findings an attribution-only license is recommended, such as the Apache, MIT, or Modified BSD license [10]. Copyright does not adhere to raw facts, and so the raw numbers in a dataset do not fall under copyright. But datasets can

have copyright barriers to reuse if they contain "original selection and arrangement" of the data, and for this reason dedication to the public domain is suggested using the Creative Commons CC0 license for example [9]. In addition, dataset authors can provide a citation suggestion for others who use the dataset. These steps will permit shared code and data to be copied, run on other systems, modified, and results replicated, and help encourage a system of citation for both code and data.

Limits to disclosure of data also include issues such as release of individual data for medical records, census data, and for example Google search data is not publicly shareable except in the aggregate. Of course "the aggregate" is defined differently in each domain. We also recognize that legal standards in different jurisdictions (e.g. European Union, United States, Japan) can vary and that each individual needs to apprise themselves of the most substantial differences.

The algorithms embodied in software can be patentable and the author or institution may choose to seek a patent. Patents create a barrier to access and it is recommended to license the software to commercial entities through a traditional patent, and permit open access for research purposes. If patents restrict access to code this can inhibit reproducibility, access to methods, and scientific progress. Within the commercial sphere, there is a need for avenues to allow audit such as non-disclosure agreements (NDA) and independent agents for auditing similar to financial audits. Public disclosure of algorithms and code can prevent patenting by others, and ensure that such scholarly objects remain in the public domain.

G The teaching and training of reproducibility skills

The breakout group on Teaching identified the following topics as ones that instructors might consider including in a course on scientific computing with an emphasis on reproducibility. Some subset of these might be appropriate for inclusion in many other courses.

- version control and use of online repositories,
- modern programming practice including unit testing and regression testing,
- maintaining "notebooks" or "research compendia",
- recording the provenance of final results relative to code and/or data,
- numerical / floating point reproducibility and nondeterminism,
- reproducibility on parallel systems,
- dealing with large datasets,
- dealing with complicated software stacks and use of virtual machines,
- documentation and literate programming,
- IP and licensing issues, proper citation and attribution.

The fundamentals/principles of reproducibility can and should taught already at the undergraduate level. However, care must be taken to not overload the students with technicalities whose need is not clear from the tasks assigned to them. Collaborative projects/assignments can be a good motivation.

H Workshop Participants

Aron Ahmadia, Dhavide Aruliah, Jeremy Avigad, David Bailey, Lorena Barba, Blake Barker, Sara Billey, Ron Boisvert, Jon Borwein, Brian Bot, Andre Brodtkorb, Neil Calkin, Vincent Carey, Ryan Chamberlain, Neil Chue Hong, Timothy Clem, Noah Clemons, Constantine Dafermos, Andrew Davison, Nathan DeBardeleben, Andrew Dienstfrey, David Donoho, Katherine Evans, Sergey Fomel, Juliana Freire, James Glimm, Sigal Gottlieb, Josh Greenberg, Tom Hales, Nicolas Hengartner, David Ketcheson, Matt Knepley, David Koop, Randall LeVeque, Nicolas Limare, Elizabeth Loew, Ursula Martin, Bruce McHenry, Chris Mentzel, Sarah Michalak, Ian Mitchell, Victor Moll, Hatef Monajemi, Akil Narayan, Peter Norvig, Travis Oliphant, Peter Olver, Geoffrey Oxberry, Fernando Perez, Konrad Polthier, Bill Rider, Robert Robey, Todd Rosenquist, Michael Rubinstein, Thomas Russell, Fernando Seabra Chirigati, Li-Thiao-Te Sebastien, Benjamin Seibold, Loren Shure, Philip Stark, William Stein, Victoria Stodden, Benjamin Stubbs, Andrew Sutherland, Matthias Troyer, Jan Verschelde, Stephen Watt, Greg Wilson, Carol Woodward, Yihui Xie.

Victoria Stodden

Assistant Professor
Department of Statistics, Columbia University.

Victoria is an assistant professor of Statistics at Columbia University, and affiliated with the Columbia University Institute for Data Sciences and Engineering. She completed her PhD in statistics and her law degree at Stanford University. Her research centers on the multifaceted problem of enabling reproducibility in computational science. This includes studying adequacy and robustness in replicated results, designing and implementing validation systems, developing standards of openness for data and code sharing, and resolving legal and policy barriers to disseminating reproducible research.

She is the developer of the award winning "Reproducible Research Standard," a suite of open licensing recommendations for the dissemination of computational results.

She is a co-founder of http://www.RunMyCode.org, an open platform for disseminating the code and data associated with published results, and enabling independent and public cloud-based verification of methods and findings.

She is the creator and curator of SparseLab, a collaborative platform for reproducible computational research in underdetermined systems.

She was awarded the NSF EAGER grant "Policy Design for Reproducibility and Data Sharing in Computational Science."

She serves on the National Academies of Science committee on "Responsible Science: Ensuring the Integrity of the Research Process" and the American Statistical Association's "Committee on Privacy and Confidentiality" (2013). She also serves as a member of the National Science Foundation's Advisory Committee on Cyberinfrastructure (ACCI), the Mathematics and Physical Sciences Directorate Subcommittee on "Support for the Statistical Sciences at NSF," and Columbia University's Senate Information Technologies Committee.

She co-chaired a working group on Virtual Organizations for the NSF's Office of Cyberinfrastructure Task Force on Grand Challenge Communities in 2010. She is a nominated member of the Sigma Xi scientific research society, and serves on several advisory boards including hackNY.org, Galaxy, and the Science Exchange.

Her Erdös Number is 3.

Chairman BUCSHON. Thank you very much.

I recognize Dr. Young for five minutes to present his testimony.

TESTIMONY OF DR. STANLEY YOUNG,
ASSISTANT DIRECTOR FOR BIOINFORMATICS,
NATIONAL INSTITUTES OF STATISTICAL SCIENCES

Dr. YOUNG. Thank you for the opportunity of testifying.

As an abstract principle, the sharing of research data is a noble goal and meets with little opposition. However, when data sharing is attempted in a particular circumstance, the conflicting interests of the parties can thwart the exchange. So said Joe Cecil of the Justice Department in 1985.

What is the current status of science in general and data availability in particular? First, where are we with science claims? In 2005, John Ioannidis published two papers of interest. In one, he asserted that 90 percent of the claims made in science papers are wrong in the sense that they are not expected to replicate. In another, he noted that five out of six papers based on observational studies failed to replicate. I published a paper in 2011 and showed that of 52 hypotheses suggested from observational studies, none replicated in the expected direction and five were statistically significant, but in the opposite direction. Begley and Ellis reported that 47 out of 53 claims made in major science journals failed to usefully replicate.

Where are we on data sharing? John Ioannidis selected 10 papers from each of 50 of the highest-impact journals—New England Journal of Medicine, Nature, Science, et cetera—and asked, is the data used in these papers publicly available? Overall, only 47 of 500 papers deposited full primary raw data online. None of the 149 papers not subjected to data availability policies made their full primary data publicly available.

I report on two personal experiences. Dr. Beate Ritz of UCLA made a claim in Environmental Health Perspectives that air pollution in L.A. county leads to low birth weights. Dr. Frederica Perera of Columbia University asserted in the journal Pediatrics that air pollution decreased IQ in children. NIEHS provided funding for both studies. In both cases, I asked for the data sets from the authors. I also asked for help from NIEHS. I resorted to FOI. I received neither data set. Recently, I was informed that NIEHS does not have the legal authority to compel and an author to proved data that was funded by them. Operationally, NIH funding, the Shelby amendment, etc. mean very little with respect to data availability. Mostly, authors do not provide data sets used in their publications. It is technically easy to share data used in publications. Others will discuss reproducible Research, so I will leave that aside.

Just why are we in this situation, where most claims do not replicate and authors will not make data sets available? In a long and illustrious career, Edwards Deming made the point that if a system is failing, it is not the workers' fault—that is the scientist—it is the fault with management, in this case funding agencies and journal editors. For over 30 years, workers have been admonished to do their work better and to make their data sets available. It was re-

ported in Science in 1988 that there were serious problems with observational studies. Nothing has changed in 25 years.

Congress, funding agencies and journal editors need to step up and manage the scientific process. They should require authors to deposit data sets on publication of their papers. Funding of data set construction and analysis should be separate. They should require data analysis strategies that demonstrate reproducibility. For example, any claim should be replicated in a separate data set before publication. Remember, the reliability of current scientific claims is only 10 to 20 percent. John Holdren's thing on the Office of Science and Technology Policy I think is a welcomed thing in this area.

It is not enough to agree with sharing data. It is almost 30 years since Joe Cecil stated the problem. Management should make the depositing of data sets on publication mandatory. This is a management problem; it is not a science worker problem.

Thank you very much.

[The prepared statement of Dr. Young follows:]

01 Testimony

Words : 628

Title: Make data used in federally funded research publicly available

S. Stanley Young, PhD, FASA, FAAAS
National Institute of Statistical Sciences

"As an abstract principle, the sharing of research data is a noble goal and meets with little opposition. However, when data sharing is attempted in a particular circumstance, the conflicting interests of the parties can thwart the exchange." so said Joe Cecil of the Justice Department in 1985. What is the current status of science in general and data availability in particular? First, where are we with science claims? In 2005, John Ioannidis published two papers of interest. In one he asserted that 90% of the claims made in science papers are wrong in the sense that they are not expected to replicate. In another he noted that 5/6 papers based on observational data failed to replicate. I published a paper in 2011 and showed that of 52 hypotheses suggested from observational studies none replicated in the expected direction and five were statistically significant, but in the opposite direction. Begley and Ellis (2012) reported that 47/53 claims made in major science journals failed to usefully replicate.

Where are we with data sharing? Ioannidis (2011) selected 10 papers each from the 50 highest impact journals, NEJM, Nature, Science, etc. and asked, Is the data used in these papers publicly available? "Overall, only 47 of 500 papers (9%) deposited full primary raw data online. None of the 149 papers not subject to data availability policies made their full primary data publicly available." I report on two personal experiences. Dr. Beate Ritz, UCLA, made a claim in Environmental Health Perspectives (2012) that air pollution in LA county leads to low birth weights. Dr. Frederica Perera (2009) of Columbia University asserted in the journal Pediatrics that air pollution decrease IQ in children. NIEHS provided funding for both studies. In both cases I asked for the data sets from the authors and also asked for help from NIEHS and resorted to FOI. I received neither data set. Recently, I was informed that NIH/NIEHS does not have the legal authority to compel and an author to proved data that was funded by them. Operationally NIH funding, the Shelby amendment, etc. mean very little with respect to data availability. Mostly authors do not provide data sets used in their publications.

It is technically very easy to share data used in publications. Others will discuss "Reproducible Research," provide study protocol, statistical analysis code and an electronic copy of data sets use in the paper. There are technical methods for dealing with de-identifying people.

Just why are we in this situation, where most claims do not replicate and authors will not make data sets available? In a long and illustrious career, W. Edwards Deming made the point that if a system is failing it is not the workers' fault. The fault is with management, in this case funding agencies and journal editors. For over 30 years, workers have been admonished to do their work in better ways and to make their data sets available. It was

reported in Science in 1988 that there were serious problems with observational studies. Nothing has changed in 25 years.

Congress, funding agencies and journal editors need to step up and manage the scientific process. They should require authors to deposit study protocol, statistical analysis code and data sets on publication of their paper. Funding of data set construction and analysis should be separate. They should require data analysis strategies that demonstrate reproducibility. For example, any claim should be replicated in a separate data set before publication. Remember current scientific claims only replicated only 10 to 20% of the time.

John Holdren on 22Feb2012 of the Office of Science and Technology Policy issued a memorandum, "Expanding Public Access to the Results of Federally Funded Research." This memorandum should be supported legislatively by requiring data availability for papers cited in support of rule-making.

Appendix I: Proposed laws

1. Use of Science Transparency Act

Any federal agency proposing rule-making or legislation shall specifically name each document used to support the proposed rule-making or legislation and provide all data used in said document for viewing by the public.

2. Federal Study Transparency Act

If federal funds are provided for a study, all data relating to the reporting of results of said study must be provided for scrutiny by the public at the time of publication.

NISS
National Institute of Statistical Sciences

Dr. Stanley Young

Dr. S. Stanley Young is the Assistant Director for Bioinformatics at the National Institute of Statistical Sciences (NISS) in Research Triangle Park, North Carolina. NISS' mission is to identify, catalyze and foster high-impact, cross- disciplinary research involving the statistical sciences. He is also the CEO of Omicsoft Corporation.

Dr. Young graduated from North Carolina State University, BS, MES and a PhD in Statistics and Genetics.

He worked in the pharmaceutical industry on all phases of pre-clinical research, first at Eli Lilly and then at GlaxoSmithKline. He has authored or co-authored over 50 papers including six "best paper" awards, and a highly cited book, *Resampling-Based Multiple Testing*. He has two issued patents. He is interested in all aspects of applied statistics, with special interest in chemical and biological informatics. He conducts research in the area of data mining.

Dr. Young is a Fellow of the American Statistical Association and the American Association for the Advancement of Science. He is an adjunct professor of statistics at North Carolina State University, the University of Waterloo and the University of British Columbia where he co-directs thesis work.

Chairman BUCSHON. Thank you.

I now recognize Mr. Choudhury to present his testimony, five minutes.

TESTIMONY OF MR. SAYEED CHOUDHURY, ASSOCIATE DEAN FOR RESEARCH DATA MANAGEMENT AT JOHNS HOPKINS UNIVERSITY AND HODSON DIRECTOR OF THE DIGITAL RESEARCH AND CURATION CENTER

Mr. CHOUDHURY. Chairman Bucshon, Ranking Member Lipinski, Members of the Subcommittee, thank you for the opportunity to be here today.

I have been asked to address questions related to data sharing, access and preservation. I would like to do so from the perspective of infrastructure development. The other witnesses have already addressed the importance of persistent scientific data archives for reproducibility. I believe that strategic investments in data infrastructure also have important implications for our overall competitiveness.

There are important lessons from our historical infrastructure development that are relevant as we consider data sharing, access and preservation. The development of railroads initially led to systems that served regional networks but eventually merged into a national network through a standard track gauge. With the development of automobiles, we adapted from early mistakes to adjust drivers' behavior through education, driving rules and seat belts. The development of the Internet reflects a layered approach of different technologies connected through a key component in the form of two protocols known as TCP and IP.

Broadly speaking, successful infrastructure development has relied on a flexible balance of community and national approaches, social aspects relating to human behavior, and key components. In each case, as infrastructure evolved through community efforts, we reached the point where national coordination moved us to a more cohesive situation. In previous cases, the more cohesive infrastructure led to greater societal benefits from both the private and public sector. I believe we have reached a similar point with certain aspects of data infrastructure.

From a policy perspective, the recent Executive Memorandum from the Office of Science and Technology Policy provides a useful framework for federal policies that would maximize data sharing, access and preservation. The memorandum acknowledges the need for flexibility by federal agencies for the communities they support balanced with the need for uniform guidelines when appropriate. There is one specific example that I will mention in my oral remarks. The memorandum outlines the need for appropriate data attribution and citation. The method for meeting this need is the persistent identifier, which is a long-lasting reference to data. You can think of persistent identifiers as an improved version of Web site addresses such as Congress.gov. It is a rough analogy, but the persistent identifier may be compared to having the same role as track gauge in the development of railroads.

From an economics perspective, there is a greater need for understanding of costs. For example, some cost studies focus only on storage, ignoring related costs such as data center operations or

longer-term costs related to preservation. Preservation of data ensures that we can extract value for the long term, noting that with data, preservation issues can arise in as little as five years. The development of data preservation infrastructure represents a case where effective partnerships could be formed between the public sector, private sector and university sector, in which I include libraries and national laboratories. It is possible that the private sector will not focus on data preservation because there are unresolved research problems, it is unlikely to be profitable, and it benefits from large-scale coordination. Federal agencies could provide the funding for research, prototypes and initial deployment of data preservation infrastructure. The university sector could then set up production systems that the scientific community and private sector could exploit for discovery and profit.

From a technology perspective, it is important to remember that there are different types of data and different stages of scientific projects. Consequently, there is a need for a layered approach to diverse systems spanning individual researchers to large-scale national projects. Even with this in mind, it is possible to identify gaps that are common across this landscape. For example, today's storage systems work well for many purposes but they do not currently meet some preservation requirements. It is worth mentioning that some storage companies view this situation as an opportunity for code development with the university sector.

From a non-technical perspective, scientists do their best to manage their data but they do not always have a full understanding. Raising awareness and reinforcing the importance of data sharing, access and preservation will be important. This type of awareness building and education is similar to the adjustment of automobile drivers' behaviors over time.

In conclusion, I believe that we have an important opportunity to advance our data networks into more cohesive, large-scale infrastructure that will advance the scientific process and generate benefits for the public sector, industry and the scientific community.

I thank you again for the opportunity to be here, and I look forward to answering your questions.

[The prepared statement of Mr. Choudhury follows:]

Written Testimony of G. Sayeed Choudhury
Associate Dean for Research Data Management
Hodson Director of the Digital Research and Curation Center
Sheridan Libraries
The Johns Hopkins University

Given before the Testimony to the Committee on Science, Space, and Technology
Subcommittee on Research
House of Representatives
Hearing on -
Scientific Integrity & Transparency
March 5, 2013

Mr. Chairman and Members of the Subcommittee, thank you for inviting me to address the following questions on data sharing, access, and preservation. I will address these questions from the perspective of infrastructure development. With prior infrastructure development (e.g., railroads, roads), there was a natural stage at which point national coordination and strategic planning moved regional systems into a cohesive national infrastructure. I believe we have reached this point with certain aspects of public access to data. The existing networks of research systems and processes at universities, scientific societies, publishers, etc. ("ecosystem") that relate to data sharing can be complemented with common, wide-scale infrastructure. The opinions expressed herein are my own and do not necessarily reflect the views of The Johns Hopkins University.

I have spent over a decade dealing with scientific data management beginning with early work associated with the Sloan Digital Sky Survey (SDSS) and continuing today through my leadership of Data Conservancy, one of the awards through the National Science Foundation's DataNet program. In addition to my experience with scientific data management, I have also had long-term experience with humanities data management, most notably through a digital manuscripts program. These diverse experiences have given me a keen appreciation for varying disciplinary needs, practices and cultures regarding data sharing but also an understanding of common infrastructure requirements that span a wide range of diverse domains and contexts. My two roles at Johns Hopkins – one related to research and development and one related to administration – allow me to focus on migration or translation of research results into operational environments.

I have led projects with funding from diverse sources including federal agencies, private foundations, corporations and a venture capital group. In addition to my experiences within the United States, I have been fortunate to work closely with colleagues and collaborators in the United Kingdom, European Union, Australia and New Zealand.

I believe that these diverse experiences, funding sources and interactions have given me a comprehensive opportunity to identify useful conditions for wide-scale implementation of data infrastructure.

Before addressing the questions directly, it is useful to consider lessons learned from historical infrastructure development. With the development of railroads within the United States, there was a period of regional railroads that served portions of the country. The recognition that a national railroad network would confer greater benefits for the transport of people and goods prompted the development of a national railroad gage that resulted in interoperability and efficiency. The evolution of automobiles reflected a process of learning and adapting from early mistakes. Eventually, we produced safer automobiles and built new roads, regional highways and eventually interstate highways. The Internet was designed and modeled with a stack model that delineated different functions and protocols, with the TCP/IP protocols being the most important.

Each of these historical infrastructure developments offers insights that are relevant when considering data infrastructure for sharing, access, and preservation, particularly relating to the balance between local versus global frameworks. The United States' investment in these earlier forms of infrastructure resulted in benefits for a range of private and public stakeholders. I believe similar investments in data infrastructure will result in benefits for scientists, the public and the private sector.

With these insights in mind, I will address the specific questions sent in advance for this hearing:

1. *What are the issues that we need to consider for wide-scale implementation of data sharing? Specifically, what are the IT infrastructure needs, including hardware, software, and technical standards, and what, if any, scientific or technical barriers to developing that infrastructure? Are there policy or non-technical barriers for sustainable digital access and preservation?*

One of the overarching issues to consider for wide-scale implementation of data sharing relates to an "ecosystem" viewpoint for infrastructure. Related to this point is the reality that all data are not alike. Scientific data comes in various levels that range from the raw, unprocessed signals generated directly by instruments (e.g., telescope, genome sequencer) to more calibrated data to highly refined, processed data cited within publications. These different levels of data possess different requirements for IT infrastructure. Additionally, the type of instrument, presence or absence of standards, community practices and other factors can result in different IT infrastructure needs (and costs as mentioned later).

Consequently, there is a need for a layered approach for data sharing, access, and preservation that includes a diversity of systems for active use of data during projects (most often directly managed by researchers); staging areas that house data for less active use (such as repositories managed by libraries; universities or data centers and cloud-based storage offered by commercial providers); data archives that preserve data but retain access and sharing provisions (nascent infrastructure that is evolving); and "dark" archives that preserve content for long-term periods without direct access. It is important to stress that these various layers of an overall infrastructure must be designed for data, which are fundamentally different than documents. Attempting to use

or re-engineer existing document management systems will result in inadequate functionality and possibly additional costs, particularly in the long-term.

From a hardware perspective, there is a need to consider enhancements to existing storage systems particularly from a data preservation perspective. Over the last three years, my colleagues at Johns Hopkins have learned firsthand regarding the issues of storage hardware and software as we have managed the data from the Sloan Digital Sky Survey. Examples include storage system block size being too large compared to smallest unit of data, inadequate methods for generating fixity (machine generated code to verify data integrity), and performance issues related to throughput (volume of data processed in a particular unit of time). Our current engagements with storage companies indicate that they view development of new capabilities as a business opportunity between the private sector and universities.

From a standards perspective, it is important to note that many scientific communities have existing standards for data sharing and access. Even in these cases, developing infrastructure and mechanisms (e.g., semantic Web) for sharing across disciplines or communities remains a challenge. It may be possible to span across two disciplines or communities through bilateral agreements. However, this approach does not scale for multiple disciplines or communities. While it is possible to develop common denominator standards for discovery of data, there remain fundamental research problems to address interdisciplinary or cross-disciplinary data sharing and access. Federal agency funding to support this type of research with the goal of developing working systems or infrastructure would be helpful.

One of the most important non-technical barriers for sustainable digital access and preservation relates to a lack of awareness regarding comprehensive data management. Terms such as storage, archiving, preservation and curation are often used interchangeably and inappropriately. My colleagues and I from the Data Conservancy have developed a data management layer stack model that conceptualizes the concepts of storage, archiving, preservation and curation. This model is not intended to be definitive, but rather reflective of our lessons learned. For this model, storage describes bits on disk, tape or in the cloud with backup and restore services. Archiving focuses on persistent identification and data protection through actions such as generating and verifying fixity and maintaining or tracking multiple copies. The term "preservation" is perhaps most often mentioned loosely or vaguely. For our model, preservation involves providing enough representation, context, metadata, fixity, and provenance information such that someone -- or some machine -- other than the original data producer can use and interpret the data. Provenance can be defined simply as whom or what machine handled the data and what did they do with the data. Finally, curation refers to adding value to foster discovery, access and re-use of data.

Researchers do not always realize the full extent of sustainable digital access and preservation. Educating researchers and changing their data management practices and behavior represents an important social component of infrastructure development. This type of behavioral change is not unlike the process that automobile drivers went through in the United States. Drivers have changed their behavior over time as we have gained greater understanding regarding safe driving and greater willingness to introduce safety through seat belts, speed limits, laws, etc. This type of social or cultural change represents an important aspect of the social-technical dimension of infrastructure development.

2. *What are the most important factors to consider in the economics of digital data access and preservation? What funding models have proven effective and how scalable are they? What should be the role of federal science agencies in supporting and preserving accessible databases? What should be the role of the private sector and of universities? How can all three work together to minimize costs and maximize benefit to the scientific community?*

The economics of digital data access and preservation require greater examination of both costs and benefits. There has been relevant work for cost models in the UK and even recent application of those models for scientific data. The Australian National Data Services has developed a business plan.

Within the US, there is a need to conduct more analyses in the full accounting sense of costs including hardware, software, human labor, utilities, etc. Furthermore, cost estimates must consider the long-term implications. For example, referring to the previous discussion about the data management layer stack model (storage, archiving, preservation, curation), some cost models account for storage only. As mentioned previously, not all data are alike so there is a need to consider cost issues according to data levels, types, presence of standards, etc. For example, a terabyte of data produced from a single instrument according to well defined standards and a single processing pipeline will probably require less cost for access and preservation than a terabyte of data produced by a single investigator using multiple instruments and within a discipline without well defined community standards.

One of the most important costs that are often unconsidered relates to data center operating costs. The power and cooling requirements for these data centers can be significant. Technologies that use less power and space will reduce these escalating costs.

On the benefits side, there is a greater need for understanding the demand for accessing, re-using and preserving data. There are potential organizations from both the private sector and university environment that would provide highly useful information case studies for costs and benefits. Examples from my own experience include the National Snow and Ice Data Center and Inter-university Consortium for Political and Social Research, both of which have successful, long-term track records with providing access to and preserving scientific data. These case studies could lead to the development of business models and eventually economic models that could be applied in a scalable manner.

In this context, it is worth mentioning that archival principles such as appraisal and intrinsic value are important, particularly as they relate to unanticipated use. There are cases where re-use of the data or secondary uses by individuals other than the original data producer generates unforeseen benefits. There is evidence that some astronomers use data archives even more often than new telescopes and that some use of high-performance computing facilities relates to re-use of existing data.

The development of wide-scale IT infrastructure for data sharing, access, and preservation is multi-faceted in that there are reinforcing roles for the federal agencies, the private sector, and

universities or national laboratories. The case for preservation highlights the possible delineation and coordination of roles. Preservation of data ensures persistent use and re-use for scientific and commercial reasons. However, it is likely that preservation for public access by itself is not a profitable activity and therefore possible that the private sector would not develop relevant capacity and service. Universities, libraries and national laboratories – which have established relationships with researchers, the public, and the private sector – have developed nascent infrastructure for data preservation but require additional resources for further development. Federal agencies could develop contracts with universities, libraries and national laboratories to further develop data preservation infrastructure that supports a range of scientific and commercial uses.

3. *What federal policies are necessary to maximize data sharing and access? Do you have any recommendation with respect to current science agency data management policies at NSF or at other agencies?*

The recent memorandum for the Heads of Executive Departments and Agencies from the Office of Science and Technology Policy's Director John Holdren offers a useful framework for considering federal policies to maximize data sharing and access, including potential extensions to existing data management policies. It reinforces the benefits of public access to data for "the public, industry, and the scientific community."

The memorandum acknowledges that federal agencies need flexibility in developing and implementing plans for data sharing, access, and preservation given the diverse set of disciplines, missions and approaches. However, the memorandum also identifies some uniform guidelines. To the extent possible, federal agencies should coordinate their responses to this memorandum and their associated plans to minimize burden and costs associated with compliance.

It is encouraging to note that each federal agency's response and plan must "ensure appropriate evaluation of the merits of submitted data management plans." In order to meet this condition, reviewers will need guidelines for effective evaluation of data management plans. My colleagues at Johns Hopkins have developed such guidelines based on our experience to date with data management plans and reviewers' responses to those plans. Many other universities and libraries can collect such information to develop community-based guidelines that federal agencies might use to inform their proposal reviewers.

The memorandum also asks federal agencies to "develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan." In this regard, it is worth examining the recent workshop report from US CODATA and the Board on Research Data and Information (BRDI) "**For Attribution—Developing Data Attribution and Citation Practices and Standards.**"

This report discusses and outlines examples of persistent identifiers—a long-lasting reference to a digital object consisting of a single file or a set of files. As an analogy, often when a webpage is not found, one encounters a "404" error and little other information to resolve the problem. Persistent identifiers mitigate this problem by assigning a permanent reference that tracks the movement of the associated digital object.

The persistent identifier is a key piece of infrastructure that demonstrates the value of using systematic approaches for data citation or identification that can be used for sharing, access and preservation. A balanced approach between local and global dimensions would include a requirement that researchers use persistent identifiers for data without prescribing the specific choice of identifier. Even though different communities will probably choose different identifier schemes initially, doing so represents progress analogous (in a rough sense) to regional railroads. As communities choose and adopt persistent identifiers, the opportunity to consider cross-community or global approaches becomes possible similar to the equivalent of TCP/IP in the Internet model.

I hope that my testimony has provided background, context and recommendations that can advance the development of data infrastructure within the United States. Such infrastructure, developed through partnership of the public and private sectors, would result in benefits for science, industry and the public. While there remain important research and social issues to consider, there are practical steps we can take now to advance our scientific enterprise especially in light of the recent OSTP memorandum related to public access to data.

Thank you again Mr. Chairman and Members of this Subcommittee for the opportunity to address these questions.

Sayeed Choudhury

Associate Dean for Research Data Management

The Johns Hopkins University

G. Sayeed Choudhury is the Associate Dean for Research Data Management and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of Johns Hopkins University. He is also the Director of Operations for the Institute of Data Intensive Engineering and Science (IDIES) based at Johns Hopkins. He is a member of the National Academies Board on Research Data and Information, the ICPSR Council, DuraSpace Board, and a Senior Presidential Fellow with the Council on Library and Information Resources. Previously, he was a member of the Digital Library Federation advisory committee, Library of Congress' National Digital Stewardship Alliance Coordinating Committee and Federation of Earth Scientists Information Partnership (ESIP) Executive Committee. He has been a Lecturer in the Department of Computer Science at Johns Hopkins and a Research Fellow at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. He is the recipient of the 2012 OCLC/LITA Kilgour Award.

Choudhury has served as principal investigator for projects funded through the National Science Foundation, Institute of Museum and Library Services, the Andrew W. Mellon Foundation and Microsoft Research. He is the Principal Investigator for the Data Conservancy, one of the awards through NSF's DataNet program. He has oversight for data curation research and development at the Sheridan Libraries at Johns Hopkins University. Choudhury has published articles in journals such as the International Journal of Digital Curation, D-Lib, the Journal of Digital Information, First Monday, and Library Trends. He has served on committees for the Digital Curation Conference, Open Repositories, Joint Conference on Digital Libraries, and Web-Wise. He has presented at various conferences including Educause, CNI, DLF, ALA, ACRL, and international venues including IFLA, the Kanazawa Information Technology Roundtable, eResearch Australasia and the North America-China Conference.

Chairman BUCSHON. Thank you very much, and I thank all the witnesses for their testimony, reminding Members that the Committee rules limit questioning to five minutes. The Chair will at this point open the round of questions. The Chair recognizes himself for five minutes.

As a cardiothoracic surgeon, I am very interested in this issue because I have to translate what is written into clinical practice, and so this type of issue really does affect real people. I can tell you the difficulty that people like me have in figuring out when to change your clinical practice, when you are doing something that turns out wasn't the right thing to do, it is a very difficult process that is ongoing, so I am very interested in this particular subject.

I will start with Dr. Young. Could you give me some examples of where State and federal regulations were made without public release of data used to make those regulations?

Dr. YOUNG. Yes. I have taken an interest pro bono in air pollution questions, and an expert in the area worked with me and we developed 100 papers that are key papers in that area. Then being a statistician, I selected 50 of those papers at random and asked the authors for the data sets. I received no data sets at all. Many of these data sets were funded by the Federal Government and there are many regulations that are based on these data sets. They are key data sets. For the most part, these data sets are not available.

Chairman BUCSHON. Just so you know, I had the same problem getting the data out of the Federal Government. It can be an issue.

Mr. Choudhury, could you give me what specific infrastructure technology requirements are required for the storage of scientific data research?

Mr. CHOUDHURY. There are several layers that are necessary to actually preserve scientific data. It begins with storage, which is basically just the bits residing on a hard disc or a tape or even in the cloud, but eventually we also need to do things to ensure data protection. We also need to have to then do things to ensure that we can migrate the data over time, so as we start to use new storage systems or if we have new file formats, we have to be able to move those data into those new environments. As Dr. Stodden mentioned, we also need to have access to the software or the tools that process the data because in many cases, it is not sufficient just to get access to the data alone. So the actual preservation of the data is this complex set of layers that go beyond storage. Storage is necessary but it isn't sufficient. So we have to do all these other things to understand the context and the reusability of the data as well.

Chairman BUCSHON. Do you think currently that university libraries or national laboratories are equipped for this type of infrastructure?

Mr. CHOUDHURY. At Johns Hopkins, we have taken an approach of looking at two stages. The first is prior to investigators submitting proposals—they need some sort of consultation and support to develop their data management plans. In this respect, I do believe that the university sector, and particularly university libraries, have stepped up very well. I think most research university libraries are providing that kind of consultation to their investigators.

The second stage is that once an award is made, then we actually have to handle the data and we actually have to start preserving it for the long term. In this respect, there is a subset of that library community that has come forward to help provide that kind of support, and then there is the long-term preservation need, and even there, it is a smaller subset again. It is in the preservation of the data where I think there remains some research questions which ultimately when they are addressed they can migrate the support into the university library sector.

Chairman BUCSHON. Great. Dr. Alberts, on February 11, 2011, in a Science magazine editorial, you write, "We will ask authors to provide a specific statement regarding the availability and curation of data as part of their acknowledgments requesting that reviewers consider this as a responsibility of the authors." Do you think this self-policing policy works in practice?

Dr. ALBERTS. We find that it has been working for Science magazine. Our senior author, deputy editor, Brooks Hanson, has been deeply involved in this. On rare occasions we have had to make authors do things that they should have done themselves but I guess we are fortunate we have the threat, which is, we are not going to publish any more papers from you, and they want to publish in Science magazine, and as Victoria said, not every journal can make that threat. So I think this is a very important issue to emphasize. We haven't talked about the fact—I am a biochemist, and I had lots of data from my laboratory when I was an active scientist. Not all of it should be preserved. I mean, if I tried to preserve everything, I couldn't find anything. So we also need different fields to decide what it is that we really need to preserve and make available. There is so much material being collected now that it is really important to get standards for different fields of what needs to be preserved and what needs to be put in your publication.

Chairman BUCSHON. Great. Thank you all. I now yield to Mr. Lipinski from Illinois.

Mr. LIPINSKI. Thank you. I wanted to start out by saying I am sort of going back to my days as a social scientist and thinking about not just the research I did and the data that I had but also thinking about behavior, and it is—there are not rewards generally for having—someone had mentioned, I think Dr. Stodden, that you are rewarded for a result in a publication but you are not rewarded—the rewards aren't there to spend the time and the effort to have the data in a format even that is accessible to others, and if you are talking about going further than that, how exactly you went through and you analyzed the data. I can't tell you how much paper I had printed out of different ways, all these different models that I ran and trying to keep track of all that. So it is not simple to do and there has to be incentives. So somehow the culture has to be changed. And the question is, how do we change that culture? Now, the National Science Foundation requires that you have a data management plan when you are applying for a grant, so the NSF puts that in there.

My question is, in a short period of time if you can do it, how do we change this, and should this be a situation where it is data available upon request or should it all be available? Should it be put out there published somewhere or put on a site that everyone

can access? And how far do we go with the data? Is it, okay, this is how I analyze it, this is the statistical package I used, this is how exactly I did it. So let me start with Dr. Stodden. I mean, what is your quick sort of suggestion on it for your 30,000 foot? What would you do if you could?

Dr. STODDEN. So I think the efforts that have been taken so far are really this on request and so on, and there are a number of experiments and studies, and Dr. Young mentioned a couple, where that doesn't seem to work as well. You don't simply get the response. So I think it is time to move forward to this being a standard. Now, having said, as Dr. Alberts said, there are data sets and problems of different importance, and you can imagine investing a lot more time curating a data set that has broad use and applicability and might underlie 50 or 100 studies and so on versus one one-off. But the changes really something that I believe scientists are willing to do and are working on standards. For example, in economics this is a very forward-thinking community and many of the journals have standards and they do engage in data sharing and code sharing but not even as much as they would like. And so I think the complexity of the problem means that it really is not a one-size-fits-all solution. As you mentioned, it is something that comes from the field.

But I would suggest that this is a standard that it should be understood that this code and the data go open for reproducibility and changing the culture is something scientists are talking about. There is a special issue I can point you to in Computing and Science in Engineering that is called Changing the Culture, and it is about giving these rewards. So as Dr. Choudhury mentioned, having these persistent identifiers allows citation for data and for code NSF steps towards allowing scholarly objects like data and code listed on the biosketch and not just publication is a real step in this direction, and I think the scientific community will sort out how it values data contribution and code contribution and publication contribution. They may not be all valued equally but we have a long history of doing this. Not all publications are valued equally. But I think that bringing this through citation and having citation standards is a way to really change the culture and reward people.

And I will add one last point, which is there is a generational difference here because these changes in technology, young people and young scientists and people who want to go into research, it is very natural for them to share data and to share code, and it is discouraging for them to enter a situation where suddenly this is not the norm. So this is something where I think there is also this opportunity that the culture is changing naturally on its own just with time as younger people come in and have these expectations for sharing what they are doing digitally. And so that is also something to capitalize on. And again, I go back to the testimony in that there is this collective action problem because, as you mentioned, it takes time, and so something particularly from federal agencies that can help push through that is really very important.

Mr. LIPINSKI. I thank you. My time is up. I yield back.

Chairman BUCSHON. I now yield to Mr. Stockman for five minutes.

Mr. STOCKMAN. I have a question for Dr. Alberts. My wife is a NASA privacy officer, and I want to follow up on something the Chairman related. In February in your editorial, you wrote, "We recognize that exceptions may be needed to these general requirements for sharing data, for example, preserve the privacy of individuals or in some cases when data materials are obtained from third parties and for security reasons but we accept those rare exceptions." Is this your view today?

Dr. ALBERTS. For example, we had an experience with a Department of Energy lab where they weren't allowed to give us the code because presumably it had some security implications. So we do encounter those one-off occasions. But they have been rare. So we have to live with the law, and we try our best to do what we can.

Mr. STOCKMAN. Do you see other exceptions?

Dr. ALBERTS. Not that—I don't know of any exceptions since that policy was made.

Mr. STOCKMAN. Okay. The other question I have is for all the witnesses. Many of you today also practice science. You are also members of the United States scientific community. You have been a world leader in producing first-class research. How do you envision the mechanism of enforcing the sharing of data without hindering the process of scientific discovery and simultaneously minimizing the administrative burden of a scientist? Because I know a lot of professors and everything a lot of time fill our more paperwork than they do research. If you could each just go quickly through the——

Dr. ALBERTS. Well, I think Victoria said it right. We need to mobilize our communities. I mean, I am a cell biologist and the American Society of Cell Biology used to help us. What does it mean for our community, and we have to take responsibility for it, and it is going to be different for statisticians. Different people will have different requirements and it has to make sense, and I agree with you that it has gone way overboard now at universities. Every time I want to do anything, I have to fill out a form. So I think we should try to avoid legislating more flat requirements. You know, if I want to interview students, graduate students at UCSF about their career options, I have to fill out a 50-page human youth form. It drives me nuts. So this Committee might work on pushing back on some of the meaningless paper and get some requirements that are more meaningful.

Dr. STODDEN. That is a great question, and I think it goes back to these issues of reproducibility. If you are publishing a paper where you claim that data and code are out there and available for it to be reproducible, then that is in a sense the starting point of standards in a community. Now, as Dr. Alberts mentioned, this will change for different communities and different research problems and they can be quite different, but there needs to be this expectation that the results, the computational results will be reproducible and then when you go and get your hands dirty and you try and do the reproducibility, then if it doesn't work or it does work, then that is value too in the community, and I think that scaffolding and that framework is really there. It is a question of moving towards this default of openness rather than the default of being closed and then you request and so on, and as I was men-

tioning to Ranking Member Lipinski, the default needs to be open, and then as you mentioned, we have exceptions for confidentiality and so on but those are the exceptions, and then the standard is really about reproducibility.

Dr. YOUNG. The first thing to keep in mind is that many estimates say that 80 to 90 percent of the claims that appear in scientific papers are wrong in the sense that they will not replicate. So I would focus on cost per valid result. Additional costs can be put into reproducible research and things like that. The total number of claims that are checked will go down but the number of valid claims can easily go up if we do our research better. Thank you.

Mr. CHOUDHURY. I think one thing that is becoming clear is data management is a complex and demanding set of activities on its own. It may not be reasonable to expect scientists to conduct their own data management but rather work with a set of professionals who sit somewhere between the domain sciences, say, library information science. So I think there is a workforce development issue here. We don't expect scientists to be experts in IT systems or other kinds of systems. We provide support for them, and I think data management may be in that category.

Mr. STOCKMAN. Thank you. I yield back.

Chairman BUCSHON. I now recognize Mr. Bera from California.

Mr. BERA. Thank you, Mr. Chairman.

Now, to start off with, I would want to make sure we don't give the impression that our scientific community and our research institutions are producing faulty data. We maintain a competitive advantage. As a scientist myself, as someone who spent countless hours in the lab as a medical student and has spent time as a faculty member and associate dean at the University of California-Davis, working with our medical students and our resident physicians, we maintain a competitive superiority in our research institutions, and I think Dr. Alberts touched on the importance of the federal investment in our research institutions. We also need to recognize our journals and particularly our leading peer review journals. There is a rigorous process having again submitted articles and worked with countless students that you go through as you are submitting articles. Replicability is an important component but also putting the information out there so others can look at it and provide feedback is very important. So we want to be conscious of that as well.

As we set up our research institutions, we often are doing it and our trials are in a very transparent way, you know, funding multicenter trials. When we look at major projects like the Human Genome Project, as we talk about brain mapping, we will set that up in as transparent a way as possible using multiple of our institutions. And it isn't always just about replicability. It is about sharing that data and working together, but at the same time—and my question is this—as we move into this era of wanting to share data, we also have to maintain our competitive advantage. We do have competitor nations that every day are trying to get to our data and get to the research institutions. We talk about cybersecurity on this Committee. We need to be very conscious of what we are putting out there as well.

I would direct a question to Dr. Alberts. You talked about the importance of research funding as well as the threats to research funding in our academic institutions. Why don't you touch on that, and then if the rest of the panel wants to talk about how we move forward in kind of an open, transparent way but maintaining our competitive advantage and protecting those discoveries that we are making.

Dr. ALBERTS. As I wrote in my written testimony, I referred to this major project from the National Academy of Sciences when I was president to explain to Congress and the public how fundamental knowledge produces breakthroughs. The first pamphlet we produced was on the global positioning system. Somewhere started with the fact that physicists invented atomic clocks. They won a Nobel Prize but everybody thought it was useless because it enabled us to keep time to a billionth of a second, and why should we want to do that. Well, you follow this progression, and I recommend that whole series. It is still up on the Web. That combined with many other findings of knowledge about the world enabled us to put up these 24 satellites that produce this wonderful device that we all use and the military uses, and we did that over and over.

And what has been true in the United States, remarkably, and I don't think people recognize this, we have been a magnet for the most talented people from all around the world coming here, and you just look at Silicon Valley and places like that. So if we don't keep our leading position as scientific research, a place to come to, our universities, then those people won't come here and they won't subsequently contribute their genius to the American economy and the American strength of our Nation. So I am quite worried right now because many other countries, China, for one, they see this very clearly. This is where we have our competitive advantage and they are trying to gain it, and if we don't pay attention to that, I think we are going to lose this game. We are taking it for granted that all these great people are going to come to this country but they are not going to do that anymore if we are not the best place to do research.

Dr. STODDEN. So I couldn't agree with your comments more, and also with Dr. Alberts that American science is absolutely superb, and as evidence of this, I believe our discussion today actually reflects the high integrity and the honesty of that community in trying to grapple with these problems. I mean, these manifestos and so on I put in the testimony here, these are scientists who are concerned about the quality of the science and trying to fix it. This is not anything other than the highest-integrity profession.

I also want to make one quick comment about corollary benefits of open data, going back to your earlier point, which is, you probably gathered by now that I think reproducibility is important but there are also issues in terms of access to the technology. So if you have the ability, the software tools and the data to replicate those results and those findings, not only can you therefore build on them more easily as well as validating them but it also opens them to industry and to others who can then capitalize on this for commercial use. I mean, whatever they see as appropriate. So it opens

all of these avenues towards economic growth that can't be overlooked that are extremely important.

And to your point about, well, what if open data helps our competitors, I think that there is a long history in the United States of being able to capitalize on this and move ahead, and I don't think that maintaining a closure around our scientific enterprise does anything but restrict American enterprise and competitiveness internationally and also threaten the integrity of our results. I mean, science moves forward, as Dr. Alberts mentioned, through skepticism and through questioning and through transparency and openness, and being able to share those methods and giving others the tools to replicate and also build on, commercialize, capitalize on all of this, I think is an avenue towards economic growth and an avenue towards STEM understanding too. When it is open, you can imagine smart high school kids getting their hands on this stuff and figuring things out and playing with it, and that is very real.

Chairman BUCSHON. Thank you. I now yield to Ms. Lummis five minutes.

Ms. LUMMIS. Thank you, Mr. Chairman.

Now, my first question is for any of you who cares to answer. It is about OSTP guidance. My question is, do you think that the guidances provides appropriate flexibility to agencies in developing plans to improve access to federally funded research?

Dr. YOUNG. Stan Young. I read the guidelines very carefully. I think they are a major advance forward. The history is that if scientists are not compelled to make their data sets available, they generally don't make it available. The American Psychological Association, for example, just started a huge effort on reproducibility. Their journals, there are 50 of them, have the author sign a paper saying I will make my data set available. Studies have shown that two-thirds of the authors that have signed those statements do not make their data sets available, so I think there is—some scientists are great. In general, there is no data sharing.

Mr. CHOUDHURY. I do think the memorandum provides a good deal of flexibility for federal agencies and the communities they support. I do think it is also important to think about those opportunities where something may be uniform across different agencies. Another example that I would give is the memo talks very clearly about enforcing data management plans. Well, most reviewers in these early days don't even know what constitutes a good data management plan, so I think providing guidelines to reviewers about what constitutes a rigorous data management plan would be a very important thing that any federal agency could do, and it would, of course, be customized to their communities.

Ms. LUMMIS. Well, I had an experience like you have mentioned with the greater Yellowstone interagency brucellosis committee where we trying to get data on elk and the transmission of brucellosis from elk to bison, bison to domestic livestock, and it was tremendously important because we finally have that disease pretty well isolated to the greater Yellowstone area after trying for, what, almost 100 years now to isolate it because it does—it used to be prevalent in milk cows, but after years of destroying entire herds of dairy cattle, we finally have that disease isolated to the greater Yellowstone area. But it is raising havoc, and there was a woman

who was an employee of Yellowstone National Park who gave her entire career paid by the taxpayers to studying elk and she would not share her data with us. I mean, she was taxpayer funded. So I have had personal experience with your frustrations here.

Another question. Could you comment on the difference between what has been written in statute versus what is happening in practice regarding obtaining data in federally funded research, you know, any of you in your experience?

Dr. YOUNG. I have a lot of experience asking for data sets, and I will call out the country of Finland. Every time I ask a scientist in Finland to send me a data set, I get it in return email. Given the electronic age that we are in, it is reasonably easy to pass data sets around. My experience in the United States is not nearly so good. I mentioned requests for 50 data sets in the area of air pollution, and I got none. The psychologists know very well that data sharing, even though it is compelled by their journals, it is not done there. There is a huge difference between what beautiful-thinking people say about sharing data, and then Joe Cecil is right. In practice, quite often it is to the advantage of the person that holds the data not to share it, and so there is a real problem and a difference. NIEHS or NIH, for example, has a wonderful data-sharing policy. However, they have no legal authority to compel anyone to share data, and so many times I have gone all the way up through very high levels of the NIH asking for data sets and have not gotten them. So the practice is very different from the publicity.

Dr. STODDEN. I would like to just reiterate Stan's point there. Both NIH and NSF grant guidelines require data sharing, and even encourage software sharing, and these have been around for at least a decade, and it seems to be unenforceable. And so when the Executive Memorandum talked about mechanisms for enforceability, I found that very exciting because, like Stan says, things can be on paper and then without that enforcement, then things don't proceed, and that, I think, is a real bridge to breaking the collective action problem and providing those incentives for sharing and rewarding scientists to do this.

Ms. LUMMIS. Thank you, panel. My time is up, so I will yield back to the Chairman.

Chairman BUCSHON. Thank you. I now yield to Mr. Palazzo for five minutes.

Mr. PALAZZO. Thank you, Mr. Chairman.

Dr. Stodden, allowing open access to federally funded scientific data may also create new business opportunities. What are your thoughts on this issue?

Dr. STODDEN. I think the evidence is clear, and one of the reasons that scientific research is funded by the Federal Government is because we can discover scientific facts and inventions and so on that then can, among other things, undergird economic growth through these creations of opportunity for industry. So something like economic open data and open methods that allow reproduction of these discoveries, I don't think it can help but fuel economic growth in the sense that you can take these discoveries—scientists don't develop things for market. They don't do commercialization or full development, particularly not of software and so on. And then

it is perfectly plausible that these can be taken out and developed into products and taken to market if that is viable, and I think that that is something that is a very compelling reason behind open data and open code.

Mr. PALAZZO. Do you have any examples of products and services that companies may be able to offer?

Dr. STODDEN. So, for example, some of my background is in image processing and working on standards like the JPEG 2000 standard. So this came out of academic research on how to do image compression and then that is released openly with open code, and that is something that can be implemented and become standard in the Web for faster loading of Facebook or whatever it is or Flickr or whatnot, and it is these types of things that are done in the scientific labs and then sometimes, as Dr. Alberts said, you don't even see the end application. You are making these discoveries and then it takes ingenuity and industry to then turn it into different other applications, but this happens absolutely all the time.

Mr. PALAZZO. And I think you mentioned this in your testimony, that it is definitely a potential economic growth area for our country?

Dr. STODDEN. Absolutely.

Mr. PALAZZO. Now, on the flip side, allowing open access to federally funded scientific research and the impact, or what would be the impact on the intellectual property rights, which innovation and U.S. competitiveness and things of that nature?

Dr. STODDEN. That is a great question, and it has, unfortunately, a complex answer that I tried to touch on in my testimony. The intellectual property structure that affects scientists was not designed for science, and there is two principal ways that it touches scientific output, and one is copyright and the other is patents, and copyright is something that works against—in the scientific context that works against openness in the sense that a scientist who produces code or produces other copyrighted outputs like a paper, I actually would need to give you explicit permission to do this. The default is not openness. So this is something I mentioned in my testimony, that maybe this is something that we need to rethink how the intellectual property system interacts with scientists who have completely different normative structure to say, for example, a poet or someone creating a movie or something like this, it is a very different model.

The other way that it interacts is through patents, and this is largely around inventions, not touching so much the computational work that we have been discussing today but software is patentable, and I can imagine—and this is actually increasing now, that patentable code is something that is coming out of the academic institution. So I think this is something that we need to think about very carefully. If you think back to 1980 and Bayh-Dole, this was something that was put into place to encourage transparency, the idea being that giving these intellectual property rights to institutions would then allow them to patent and give them this incentive, a financial incentive, to be open. Now if we have standards of reproducibility where code is open and data is open, it doesn't make sense to have that same incentive to patent because it actu-

ally becomes more of a barrier because in 1980, no one imagined you would just go to a repository or get hub or whatnot and click and get the code. It had to be this whole thing through a tech transfer and so on, which is completely different and now that is the barrier. So I think there is some careful thinking that needs to happen in terms of IP and also around how we collaborate with industry too. Industry has very fruitful collaborations with academia, and those need to be worked out in terms of what intellectual property remains over the scientific output so that industry has—essentially they can sort of get some return on their investment.

Mr. PALAZZO. I yield back, Mr. Chairman.

Chairman BUCSHON. Thank you very much. I would like to thank all the witnesses for their valuable very interesting testimony and the Members for their questions. The Members of the Committee may have additional questions for you, and they we will ask you to respond to those in writing. The record will remain open for two weeks for additional comments and written questions from Members.

The witnesses are excused and the hearing is adjourned. Thank you, everyone.

[Whereupon, at 11:06 a.m., the Subcommittee was adjourned.]

Appendix I

Answers to Post-Hearing Questions

ANSWERS TO POST-HEARING QUESTIONS

Responses by Dr. Bruce Alberts

QUESTIONS FOR THE RECORD
THE HONORABLE LARRY BUCSHON (R-IN)
U.S. House Committee on Science, Space, and Technology

Scientific Integrity and Transparency

Tuesday, March 5, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

1. One of the witnesses on the panel, Dr. Stan Young, wrote in his written testimony that "funding of data set construction and analysis should be separate." What do you think of this suggestion, and could this be executed in a manner that is practical for the scientific community?

   For the vast majority of science, this is neither necessary nor practical.

2. In your written testimony, you write: "Funding agencies can help by facilitating and rewarding the publication of failures to replicate important published results." Is this recommendation practical? How would you recommend that this be implemented?

   One way would be to competitively provide financial support to a set of scientific journals that agree to publish an open-access (immediately free) subsection that publishes brief "failure to reproduce" articles that pass a minimal screen for quality. One might start by focusing on the large amount of such data that is generated by biotech and pharmaceutical companies, inasmuch as these entities will always attempt to reproduce results before they use them to develop drug-development programs. Presently, almost none of this information reaches the public. What would it take for them to share their negative results in this way? One could start with a few journals to work out the mechanisms, as one would need to work out ways to insure a proper screening, so that the inevitable cranks cannot publish such reports based on no data. Also needed is a way to prevent these papers, which are expected to be poorly cited, from lowering the pernicious "impact factors" calculated for those journals that volunteer to try to help in this way. (*Science Translational Medicine* could possibly be one).

   In addition, repositories like PubMed should agree to link such failure to reproduce articles prominently to the original publication, and the government should insist that this information be included in all grant applications (that is, the information needs to be available when reviewers are considering the previous work of each applicant).

See also my answers to #3 and #4 below.

3. On February 22nd 2013, the Office of Science and Technology Policy (OSTP) released guidelines, which outlined objectives for public access to scientific data in digital formats. The guidelines defined data as: "*the digital recorded factual material commonly accepted in the scientific community as necessary to validate research finding, including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects such as laboratory specimens.*" Do you agree with this definition? Does it reasonably describe what data should be shared?

   Yes, but the critical statement is "*the digital recorded factual material commonly accepted in the scientific community as necessary to validate research finding{s}* ". This will be different for different types of science, and a select set of scientific societies could be commissioned (and funded) to help guide what is needed (see my answer to #2 above).

4. The guidelines by OSTP outlines ten points that agencies must consider regarding the public access to scientific data in digital formats. Do you have any concerns with anything on this list? Is there any policy recommendation that you would like to see changed?

   I am not sure how to interpret: "Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, ……". Critical to me is the "as appropriate" statement, because not every researcher should be burdened with developing such a formal plan. The burden of required (and often unnecessary) paperwork for researchers in the US is ever increasing, and we need to be sure that such a new requirement is not universally applied. Again, an appropriate set of scientific societies can be critical in specifying exactly where a formal data management plan does, and does not, make sense. (See answer to #3 above).

5. Are there some situations or scientific fields where it would be cost-prohibitive to store and share data? Please explain. How should data be shared in these cases?

   Yes. As pointed out in the special Data issue of Science magazine (11 Feb. 2011), in fields like nuclear physics and genomics, it is not possible to store all of the data, and much must be discarded. The particular community in each case must set thoughtful standards of what is to be saved and how it can be shared appropriately.

6. One of the reasons for not releasing data in experiments is that it may contain personal identifying information. Is this a legitimate reason on the part of researchers not to share

data? Please explain. How can we promote the sharing of such data while also assuring that confidentiality will be maintained?

I am not an expert here, but others are. For the advance of medical science, there need to be systems set up that allow patient data to be effectively shared. These should minimize the risk of loss of privacy. But it seems likely that insisting on systems with zero risk will prevent ANY meaningful data aggregation, and thus a balance must be sought.

7. As an editor, have you encountered any situations where the research was funded both by the federal government and a non-profit or for-profit third party? Would these cases, or any other similar circumstances, merit any special consideration?

   Yes, this is frequently the case. In the view of *Science* magazine, even having a private sector company involved does not eliminate the need to share data as part of the "cost" of publication. Otherwise the self-correcting nature of science becomes impossible.

8. Would companies (such as pharmaceutical, oil, or technology companies) be less willing to publish their results with federally funded scientists, and would data-sharing policies stifle any potential research collaborations between the two?

   This is certainly possible, but the downsides of data-sharing policies are clearly outweighed by their many positive aspects. Note that many journals, including *Science*, have had data sharing requirements for some time, and this has not stifled publications from company scientists. Indeed, it would be harmful (e.g., including to shareholders) for companies to be able to make claims in the scientific literature while holding back data necessary to evaluate a claim. The data-sharing requirement is quite different from company decisions on whether to announce findings to competitors.

9. Would a move towards open-access of published data cause additional administrative costs for Universities and other Institutions that receive federal funding for scientific research? How can we minimize administrative burdens while simultaneously maximize access to data?

   To minimize such ever-increasing burdens, I would not place the onus on the institution hosting the scientist. I suggest that the screening for a data management plan, where needed, be instead carried out by the journals before publication (using expert peer review), and that each funding agency be responsible for producing the incentives required for those who do not behave responsibly. Those who hold the purse strings can best insure compliance. And again, for each field of science, I recommend that selected scientific societies be funded to set the standards and make them both reasonable and widely accepted by the scientists in that field.

10. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

    I agree with the need for caution to minimize the burden on scientists, and this is the reason for many of my answers above. Any policy needs to be very sensitive to each type of science involved, which is why I have suggested that select scientific societies be centrally involved. This is despite the fact that it can be hard to predict which data will or will not be needed for some future research. Many new fields are emerging that are taking advantage of individual data sets in ways that were not considered when those data were collected.

11. There are differing types and sizes of data used by various science disciplines. Is a "one-size-fits-all" policy appropriate?

    No, as I repeatedly emphasize above.

12. We now live in an age where virtually every academic and non-profit research institution has a webpage, and where all researchers (with some rare exceptions) have the means to maintain their own websites. Authors of publications could be required to archive data on an institutional website. However, in a Science Magazine Editorial, you write that such attempts are "only a stopgap solution." Could you explain what you meant by this statement?

    What we need are reliable repositories for the most critical kinds of data. Institutional websites cannot be expected to be permanent, and for huge datasets, the storage costs are prohibitive. Of equal importance is the setting of standards that make the data accessible, which is best done by a central depository. I suggest that you request testimony from successful examples such as the "Protein Data Bank", in order to get a good feel for exactly what is involved and why we need better support mechanisms for such entities. The good news is that the expansion of science around the world should make it possible for the US to shoulder less of the costs (and countries like China more), since these types of resources are essential for science to progress all around the world.

13. What specific support could the federal government contribute towards a permanent community-maintained archive for storing research data, that non-federal organizations could not provide?

We would never want a massive archive for storage of all types of research data, because each type of science that requires such an archive has different needs and requires the development of its own standards for storage. Making each such database effective for data aggregation and access takes real expertise and a close association with the particular scientific community involved. Thus for example, invaluable, separate databanks currently exist for different model organisms -- such as yeast, the model plant *Arabidopsis*, the fruit fly *Drosophila*, and the worm *C. elegans*, mice, etc. Each database is operated and maintained by experts for that organism. As increasing amounts of biological data accumulates, one can foresee a need for new such databases. Meanwhile, the older ones will need to be maintained. Thus increasing resources will be required to support such databanks to keep them freely accessible on the Web. Since these databases are needed by a huge range of scientists, and supporting them by charging user fees would severely limit access (and thus both waste resources and retard the progress of science), the governments of nations around the world would seem to be the only feasible source of long-term, reliable funding.

14. A 2007 GAO Report entitled "Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research" states: "*The scientific community generally rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure*." What are your suggestions to change this structure?

    Incentives for scientists will change if journals like *Science* insist on their data availability policies, if grant funds are specifically added for the "preparation of data for others' use" to those competitively awarded grants that need them, and if shared databases like those I describe in my answer to question 13 exist that make each scientist's data deposition a straightforward, readily certified process. As a condition for publication, each journal can then require proof of data availability in the form of a registration number that the database provides upon data deposition. As one example, this is what is currently done for three-dimensional protein coordinates via the Protein Data Bank (PDB), as I describe in my written testimony.

    An additional problem, not included in the above question, is that we need to develop more respect and support in academia for what might be called "data scientists" – outstanding specialists who make discoveries by clever analyses of the data collected by other scientists. Otherwise, much of the vast amounts of data that can be productively mined will go unused. The Gordon and Betty Moore Foundation, where I am a Trustee, has for this reason recently begun to fund a major initiative with this aim (information can be found on the Moore Foundation website, under its "Science" program).

15. On July 29, 2010 Dr. David Lipman testified before the House Subcommittee on Information Policy, Census and National Archives. While most of his testimony

centered around open-access issues, he noted that the National Center for Biotechnology Information (NCBI) produces more than 40 databases, including GenBank and dbGaP. He also mentioned other data intensive activities that his center is currently handling. Based on his testimony, and other publically available information about the activities at NIH Pubmed Central and NCBI, do you think that they have the technical capability and infrastructure to store, archive, and handle large amounts of data (i.e. achieve the purposes of open-data)? Please explain. If there was a movement towards a national repository for scientific data, would it not be better to build off of existing infrastructure at NIH and NCBI? What are other issues that should be taken into consideration when going towards a single repository model? Finally, based on your experience, do you see any potential cross-agency issues (for example between NIH and NSF) that might make a single federal repository inefficient or not worthy of pursuing?

My answer to question 13 is relevant here. I do not believe that we should put all of our eggs in one basket, despite the excellence of NCBI. Instead, I believe in competitions to select the best solutions for each case. In addition, our mechanisms should make it easy for other governments to pay their fair share – which share should increase dramatically as science expands around the world in the years ahead.

However, your question is an intriguing one that needs a detailed study by experts. This is the type of question that can best be answered through a major study by the National Academy of Sciences, and I would recommend that the US government seriously consider financing the production of such a formal investigation soon.

Responses by Dr. Victoria Stodden

QUESTIONS FOR THE RECORD
THE HONORABLE LARRY BUCSHON (R-IN)
U.S. House Committee on Science, Space, and Technology

Scientific Integrity and Transparency

Tuesday, March 5, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

1. One of the witnesses on the panel, Dr. Stan Young, wrote in his written testimony that "funding of data set construction and analysis should be separate." What do you think of this suggestion, and could this be executed in a manner that is practical for the scientific community?

I don't think there is a clear cut answer that applies in all cases. I can imagine a psychologist, for example, who carries out a survey to answer questions about religious impact; and I can imagine a data arising from a capital investment, such as a telescope, sequencer, or collider, where the data collection is a separate operation from the analysis already. It doesn't seem necessary to impose that structure on the psychology experiment.

An important aspect of this distinction is the need to reward data contributions to scientific research. In the case of the psychologist she should render a usable dataset along with the final paper (and the source code she applied to filter the data and test her hypotheses), and she should garner citation for both subsequent dataset use as well as for her research article (same goes for her source code contribution). Rendering the data and code at the time of publication is a new step for many researchers.

2. On February 22nd 2013, the Office of Science and Technology Policy (OSTP) released guidelines, which outlined objectives for public access to scientific data in digital formats. The guidelines defined data as: *"the digital recorded factual material commonly accepted in the scientific community as necessary to validate research finding, including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects such as laboratory specimens."* Do you agree with this definition? Does it reasonably describe what data should be shared?

This is largely a legal definition to sit comfortably with Bayh-Dole and Feist (499 U.S. 340 (1991)). I think the emphasis on the need to validate results is key in defining what data should be shared at publication of the first research article that uses it (otherwise the definition is overbroad to the point of being unclear). I believe there is a bright line with regard to sharing in science – it occurs at the time of publication. Once the scientist decides to publish her work, she is then subject to validation and community skepticism, and needs to disclose enough for other to understand and replicate her work if they choose (this is not new, scientific communication is currently this way, except the need for data and code disclosure is new). Before publication, I believe the scientist may work privately is she chooses, and not all do, since this is fundamentally a creative effort. However, during this private phase, she must keep adequate records of her work so that when the work is published she can conform with standards of reproducibility (data and code along with the writeup of the experiment). I

that sense I agree that drafts, lab notebooks, etc do not need to be disclosed at the time of publication, unless needed for validation – but this information should be in the paper, data, code without need for the other objects. (If the scientist is accused of misconduct, the university or funding agency may demand objects such as lab notebooks, but this is an entirely different case than typical publication.)

3. The guidelines by OSTP outlines ten points that agencies must consider regarding the public access to scientific data in digital formats. Do you have any concerns with anything on this list? Is there any policy recommendation that you would like to see changed?

I feel a) ii) must be interpreted carefully and broadly, otherwise just about every dataset or code could be construed to fall under here. It must be interpreted subject to the constraint that access be maximized.

b) munges who is responsible for long term access somewhat. I believe that falls to archivists and repositories, not the researcher himself. He does need to know which archive he'll use though.

e) I am very happy to see enforcement and compliance included. As you may know, both NSF and the NIH have data sharing requirements in their grant guidelines, and have for more than a decade, but does not enforce them.

For g), again, this must be construed in the public interest. Already Elsevier is planning to charge for gateway access to scientific data it plans to host. I think the publishers are too used to a situation of exploiting scientists, rather than working with scientific norms, to be involved in data sharing and code sharing.

h) needs to reference software.

4. Are there some situations or scientific fields where it would be cost-prohibitive to store and share data? Please explain. How should data be shared in these cases?

Yes. CERN for example has data too large to share, at least with current technology. Even if the petabytes could be made available online, it is not useful to potential users since it cannot be downloaded as you would a typical dataset. Some datasets can be used for research while existing in the cloud, so not needing to be downloaded, but that requires additional infrastructure (cloud computation, online interface to the data). This is a good solution for large data, although probably CERN's data is still too big, and for some confidential data, since access can be controlled. I believe there are still some exceptions to data sharing due to size and scale, ie. CERN, but these are very rare and can be dealt with on a case by case basis, without disruption to default of openness.

5. One of the reasons for not releasing data in experiments is that it may contain personal identifying information. Is this a legitimate reason on the part of researchers not to share data? Please explain. How can we promote the sharing of such data while also assuring that confidentiality will be maintained?

Solutions will need to be developed for these cases. I can imagine sharing confidential data with an authorized subset of users, say other independent researchers on the topic, in a "walled garden" for example. This is not as useful for scientific integrity as open data (the principle of "many eyes make all

bugs shallow"), but it is much more useful that not sharing, since at least some researchers will have the opportunity to independently validate results.

6. Would companies (such as pharmaceutical, oil, or technology companies) be less willing to publish their results with federally funded scientists, and would data-sharing policies stifle any potential research collaborations between the two?

That's possible, although I believe it is unlikely and here's why. Pharma is under pressure to make the results of its clinical trial available (see e.g. http://www.alltrials.net and http://www.nytimes.com/2012/10/11/business/glaxo-opens-door-to-data-on-its-research.html) Industry is changing as well in regard to openness and arguably it is *appealing* to work with academics who routinely share data. Not only because of perceptions and public pressure, but also because it simply will produce more reliable results.

7. What is the economic benefit for the U.S. if we go to an open-access policy of federally funded scientific data?

If someone asked, at the down of the Internet, "what's the benefit?" would we have been able to list all the ways in which we use it today? No, of course not. If this had been made a criterion for its funding we would not have an Internet today. This is a case where data sharing is the right thing to do, and there will be uses of the data we haven't imagined.

Having said that I think there is clear economic benefit from greater credibility in scientific findings and this will cause fields to progress more quickly. There is great potential for acceleration of discoveries in climate science and computational biology with routine data and code sharing, for example. Drug trials will be more reliable if results can be checked.

There are also direct economic benefits, such as faster translation of scientific discoveries into commercial products (for example, image compression standards like mpeg and jpeg and algorithms improved MRI functioning, data driven ventures for shopping or health (especially the "quantified self" movement), information theory results that gave rise to our current cell phone network, and many yet to come).

8. There are other countries, such as China, that do not have a strong history of respecting copyright or intellectual property rights. What are your thoughts on other countries having access to our federally funded research data?

Assuming the data does not contain national security information or private information about Americans, this is great. It means there are more voices in the scientific conversation, which are better able to find mistake in the science and find improvement and new discoveries. As long as we have standards for data and code openness so that results can be validated, this is nothing but a win for us since it is a win for scientific progress. Better this than a closed system when data driven results can be published without their data and code and are unverifiable.

With regard to economic growth resulting from open data and code, yes, other countries would be able to capitalize on this as well. But in the economic competitiveness arena, I would bet on America every time to succeed and dominate in capitalizing on business opportunities.

9. What are the potential business platforms and areas that you envision would develop from open access to data? Could you give some possible examples?

For example data driven ventures for shopping or health (especially the "quantified self" movement), use of geo-location data. There are existing way these data can be useful for businesses (banks and airlines for example having more information to inform their businesses and provide services and even to place relevant ads (which is Google's multibillion dollar business model)). There are also entrepreneurship ventures that can arise from data. The "startup scene" in New York is centered around capitalizing on data driven opportunities. Startups like Foursquare, Tumblr, Hunch, FogCreek, Etsy, 10gen, AOL Ventures, Betaworks, Union Square Ventures, all in NYC, focus on data as their primary driver.

10. Would a move towards open-access of published data cause additional administrative costs for Universities and other Institutions that receive federal funding for scientific research? How can we minimize administrative burdens while simultaneously maximize access to data?

It might. In fact it might be good to involve university research offices in compliance to facilitate the transmission of data and code to repositories. I don't foresee this potential expense as anything other than marginal, but it may be worth considering this as a line item on grants or an explicit part of indirect cost deductions.

11. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

This is possible, but like all science it is not possible to tell what the most useful results will be in advance. This implies that standards of reporting should be as consistent as possible for all research. Certainly larger datasets and more obviously useful code can receive greater investments in usability and preservation, but I can't imagine excepting great swaths of publications because someone thinks they won't be useful in the future. If this is the case they shouldn't be published at all, so if we publish the efforts for code and data sharing should be in place.

12. There are differing types and sizes of data used by various science disciplines. Is a "one-size-fits-all" policy appropriate?

No, there must be community involvement and community decisions on appropriate sharing, subject to the criterion that whatever is sharing must be sufficient to validate the published computational results.

13. What specific support could the federal government contribute towards a permanent community-maintained archive for storing research data, that non-federal organizations could not provide?

This is important. A federal effort would be trusted and reliable, unlike a private effort that may in future because closed so the private entity can generate revenues (we have seen this happen with the publishers of journal articles). The concern is that any federal effort be effective. For example, NSF's FASTLANE is considered ineffective, awkward to use, and annoying. We do not need a FASTLANE-style federal data repository, we need a tightly considered and well constructed useful repository. Luckily, there are many good repositories in existence (The DataVerse Network, IPCSR, and UK efforts such as the UK Data Archive) that have pioneers ways to do this effectively.

14. A 2007 GAO Report entitled "Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research" states: *"The scientific community generally rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure."*
What are your suggestions to change this structure?

We need to demand and expect citation, even if it is nonstandard citation to begin with, for data and code. Uncited data and code use should be considered plagiarism. (Note however, that using someone else's code in an unmodified form is very desirable – this is not plagiarism).

15. What specific technical standards need to be considered when storing data for open access?

Access by the public is not the scientist's responsibility. The scientist is responsible for making his or her data and code usable by other researchers in the field. These are the data that the public get – if others (not the researchers) wish to build upon the data to provide further explanations for usability features, that is great, but it is not for the researchers to do that.

16. What federal agency and/or other entities would be appropriately suited to determine standards for storing data?

Researchers need to determine this for their data, and the relevant research community will give feedback to develop appropriate sharing standards for their data and code. There is no clear answer to who should determine relevant additional layers for public use. I would suggest leaving this question open – perhaps the public and the larger community will annotate data, or repositories can start to develop this during the curation process.

17. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

See previous answer to this question.

18. On July 29,2010 Dr. David Lipman testified before the House Subcommittee on Information Policy, Census and National Archives. While most of his testimony centered around open-access issues, he noted that the National Center for Biotechnology Information (NCBI)

produces more than 40 databases, including GenBank and dbGaP. He also mentioned other data intensive activities that his center is currently handling. Based on his testimony, and other publically available information about the activities at NIH Pubmed Central and NCBI, do you think that they have the technical capability and infrastructure to store, archive, and handle large amounts of data (i.e. achieve the purposes of open-data)? Please explain. If there was a movement towards a national repository for scientific data, would it not be better to build off of existing infrastructure at NIH and NCBI? What are other issues that should be taken into consideration when going towards a single repository model? Finally, based on your experience, do you see any potential cross-agency issues (for example between NIH and NSF) that might make a single federal repository inefficient or not worthy of pursuing?

Yes, I think they do since they have considerable experience in these area for complex biological data. Most of their experience is in genomic and –omic data, which is only one type of data. I believe expanding the existing infrastructure at NIH to involve other agencies is a wise move, just as I believe PubMed Central should be expanded to other agencies to become PubCentral (or similar). It may be worth commission a study for both of these extension to the NIH core infrastructure.

Responses by Dr. Stanley Young

QUESTIONS FOR THE RECORD
THE HONORABLE LARRY BUCSHON (R-IN)
U.S. House Committee on Science, Space, and Technology

Scientific Integrity and Transparency

Tuesday, March 5, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

1. On February 22nd 2013, the Office of Science and Technology Policy (OSTP) released guidelines, which outlined objectives for public access to scientific data in digital formats. The guidelines defined data as: "*the digital recorded factual material commonly accepted in the scientific community as necessary to validate research finding, including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects such as laboratory specimens.*" Do you agree with this definition? Does it reasonably describe what data should be shared?

The definition of data is fine, but necessary material is not complete.

To reproduce the finding in a paper three things are necessary:
1. An electronic copy of the data sets used to produce the tables, figures, and statistical analysis presented in the paper.
2. The protocol of the study.
3. The statistical analysis code used.

2. The guidelines by OSTP outlines ten points that agencies must consider regarding the public access to scientific data in digital formats. Do you have any concerns with anything on this list? Is there any policy recommendation that you would like to see changed?

Point 1 Policy principles.
1. NIH does not have the legal authority to require grantees to make data available. It makes sense to give them the legal authority.
2. Peer review does not provide careful checking of analysis or re-analysis with alternative methods. All of the claims that fail to replicate appeared in peer reviewed studies, Feinstein, Science, 1988.

Point 2 Agency Public Access Plan
1. Where possible protocol, analysis code and e data should be placed in a public repository. Design of de-identification should be part of the research plan.

2. De-identification will often come up. There is considerable literature and practice in this area. De-identification is almost always possible so this should not be an excuse to not provide data.

Point 3 Objectives for Public Access to Scientific Publications

1. Item f)ii) Meta data should also include study protocol and statistical analysis code as well as e data.

Point 4 Objectives for Public Access to Scientific Data in Digital Formats

1. "resource constraints" To produce tables, figures and statistical analysis raw data has to be formatted for analysis. Once this is done making data available should not present a resource constraint. See also, cost example given in Question 10.
2. Note that research by OMOP indicates that decisions made in the processing or raw data into the file used for analysis can dramatically influence the claims made in an analysis. Access to the data used in a paper not perfect, but it is a very good start. People could reasonably ask for raw data as well.
3. Item a) i) De-identification is most often technically possible. The research plan should address how this is to be done.
4. Item a) ii) Intellectual Property. The key point in time is publication or making a report to a government agency. IP can be protected by publishing AFTER filling for patents. When a paper or report is public, data should also be available.

Point 5 Implementation of Public Access Plans

1. Implementation date. Key papers need to be identified that are used to support regulations. The papers and the data used in these key papers needs to be made public. If it is not possible to provide data for these papers, the paper can not be used to support agency regulations. See suggested laws, Point 7.

Point 6 General Provisions

1. In general the OSTP requirements are very good. How are the OSTP provisions to be enforced?

3. Are there some situations or scientific fields where it would be cost-prohibitive to store and share data? Please explain. How should data be shared in these cases?

One size does not fit all, clearly. A good step is to require that the data sets used to produce the tables, figures and statistical analysis in a paper/report be placed in a public repository at the time of publication of the paper/report. Usually these files are much smaller than the raw data files. Also note there are costs associated with false claims. See response to Question 10.

4. One of the reasons for not releasing data in experiments is that it may contain personal identifying information. Is this a legitimate reason on the part of researchers not to share data? Please explain. How can we promote the sharing of such data while also assuring that confidentiality will be maintained?

There is good technology for de-identifying personal data. How this is to be done should be planned from the beginning of the study.

5. Could you comment on the difference between what has been written in statute versus what is happening in practice regarding the obtaining of data in federally funded research?

The NIH, for example, has a wonderful policy which is given on their web site. There is the Shelby amendment. There is FOI. Etc. In practice, there is essentially very little data sharing. I requested a data set that was used in a journal that required authors to sign a statement that they would make their data available. They refused. The editors strongly intervened and six months later I was given the data set. The claim made by the authors was not supported by re-analysis of the data set. Several times I have asked for data funded by NIEHS. I made my request to the university in addition to the author. I made a FOI request to NIEHS. I received no data. I requested that NIH become involved. They said they had no legal authority to make the author provide data. Federal agencies should be given the legal authority to make authors post their data.

In applying for a NIH grant the scientist has to say what will be done for data sharing. However, what they say is not used in the evaluation of the grant request. The word appears to be out. Authors understand that they do not have to share data.

6. How effective is the Information Quality Act (IQA) or the Shelby Amendment in obtaining federally funded data? Are these current federal guidelines adequate for reproducing the scientific claims of scientists whose research is sponsored by the federal government?

In my experience both of these laws are completely ineffective. Authors thumb their noses. Universities do not support those making a request. NIH appears powerless. Essentially authors do what they like. They may provide data to friends. They withhold data at their pleasure. One solution would be to fund data set collection and building separately from data set analysis. Once a data set is build, it is publicly posted. In that case it is in the interest of the data set builder to post the data (to collect final payment for their work).

Much data is beyond the reach of IQA, Shelby, etc. Often the agency contracts the work and never takes possession of the data. If the agency is requested to provide the data, they can say in truth, We don't have it. This flaw was pointed out by Cecil and Griffen, The Role of legal policies in data sharing, 1985, page 171-172:

> "...However, the private or public status of a data set can be difficult to determine when research data sets are developed through public funding of private researchers. This is a common circumstance."

> "Recent interpretations of the term "agency records" have been rather restrictive and not likely to aid researchers who seek access to data sets maintained by private researchers but developed with public funds through either contracts or grants."

It appears that a law is needed to reach these data sets.

7. Given the current requirements posed by existing law, are new laws necessary, or should existing requirements be enforced more strictly? What is wrong with current enforcement mechanisms?

There is currently no effective enforcement. The rule/law should be "when a paper is written or a report turned over to a federal agency the data used in that paper/report should be placed in a public repository along with study protocol and statistical analysis code. The major problem of the current system is that it is often in the best interest of the author to keep the data for themselves; hence the suggestion to separate data set building from data set analysis.

Useful laws:
Use of Science Transparency Act
Any federal agency proposing rule-making or legislation shall specifically name each document used to support the proposed rule-making or legislation and provide all data used in said document for viewing by the public.

Federal Study Transparency Act
If federal funds are provided for a study, all data relating to the reporting of results of said study must be provided for scrutiny by the public at the time of publication.

8. Would a move towards open-access of published data cause additional administrative costs for Universities and other Institutions that receive federal funding for scientific research? How can we minimize administrative burdens while simultaneously maximize access to data?

This question tacitly assumes that the vast majority of claims made in paper are valid and that we are interested in the few critical claims that may not be valid. If that were true, then depositing of data sets would be burdensome. First let's disabuse everyone that most claims made in science papers are true. Ioannidis, PLoS (2005) reports that 90% of claims are expected to fail to replicate. Young and Karr (2011) give evidence that over 90% of claims from observational studies fail to replicate. Begley and Ellis Nature (2012) report that 47/53 claims made in experimental biology papers fail to replicate. Fang et al. PNAS (2012) report that 40% of retracted papers are retracted due to fraud. Also note that if making data available is a requirement from the start of the research, then good planning will reduce the cost. Oversight is the issue. Currently there is no oversight and it is estimated that 90% of claims fail to replicate. The most cost effective way to have oversight is to have interested scientists take the time to carefully look at the paper, the

data, the analysis code, and the protocol. There can be no effective oversight without access to the data (code and protocol).

9. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

If a scientist institutes a study and publishes the work there is a presumption that the question was a good one and the results worth knowing. If scientists' claims were highly reproducible, then yes, there is extra work. To move society and science forward, claims should be reproducible. See comment on Question 8.

10. What specific support could the federal government contribute towards a permanent community-maintained archive for storing research data, that non-federal organizations could not provide?

Public or private archives should work.

One thought is that funding fewer, higher quality studies might actually be cost effective. False claims cost money as scientists attempt to replicate the finding or base work on false claims. The public can react to false claims, e.g. coffee causes pancreatic cancer, so far as we know, a false claim. Suppose that Ioannidis is correct that 90% of claims are false. So for 100 studies, you expect 10 correct claims and 90 false claims. Suppose that you fund 50 studies and that 20% of those claims are reproducible. You would still expect 10 valid claims, but you would only have 40 false claims. So by improving the quality you can have the same number of valid claims and you can dramatically reduce false claims. Even with depositing of data sets, running 50 studies are expected to be less expensive that 100 studies. The fact that the scientist knows that there can be oversight should inspire higher quality studies.

11. A 2007 GAO Report entitled "Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research" states: "*The scientific community generally rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure.*" What are your suggestions to change this structure?

The cleanest way to change the reward system is to fund data creating separately from data analysis. There is now a journal that is aimed at describing and making access to

data their goal. So if data construction is funded separately, there is a publication place for the effort. There would be a strong incentive for researchers to build good data sets and make them public.

There are many data sources that could be joined to address important societal questions. For example, health data joined with air pollution data. The building of joined data sets is a different skill set from the analysis of a joined data set. Fund the building and analysis of data sets separately is very attractive.

12. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

Is depositing of data worth it? See Question 9.

Responses by Mr. Sayeed Choudhury

U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

Questions for the Record

Scientific Integrity and Transparency

Tuesday, March 5, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

---

1. On February 22nd 2013, the Office of Science and Technology Policy (OSTP) released guidelines, which outlined objectives for public access to scientific data in digital formats. The guidelines defined data as: "*the digital recorded factual material commonly accepted in the scientific community as necessary to validate research finding, including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects such as laboratory specimens*." Do you agree with this definition? Does it reasonably describe what data should be shared?

   The definition of data from the OSTP memo represents a useful starting point for public access to scientific data in digital formats. It is important to balance the needs for public access with the costs and burden that would be placed on researchers, universities, publishers, etc. The validation of research findings, particularly those from published articles, represents an important criterion from which to identify relevant data. It is also important to note that federal funding agencies generally do not provide funding to digitize print or physical materials though there are some exceptions (e.g., Institute of Museum and Library Services, National Endowment for the Humanities). Note that even if there are valid reasons for not offering public access to data, there may be still be valid reasons for preserving the data.

   There are some cases where documentation (in addition to data) is necessary to validate research findings. For example, notes from a laboratory notebook might be necessary to fully understand the processing of data. Even in such cases, the goal of validating research findings remains relevant rather than an overarching policy that could raise costs or burdens unnecessarily. Finally, even if physical data items are not available through public access, it is nonetheless important that researchers describe within their data management plans the means through which they (or their institutions) maintain, provide physical access to and preserve these objects.

2. The guidelines by OSTP outlines ten points that agencies must consider regarding the public access to scientific data in digital formats. Do you have any concerns with anything on this list? Is there any policy recommendation that you would like to see changed?

   The ten points from the OSTP memo describe a useful set of recommendations. There are a few additions or suggestions that I would recommend for the list:

   d) Ensure appropriate evaluation of the merits of submitted data management plans;

   - In order to properly evaluate merits of submitted data management plans, federal agencies should consider instructing their reviewers to comment on the plans specifically. For effective review, agencies should provide general guidelines noting that communities of practice vary by discipline or community.

   e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;

   - Such mechanisms should be oriented toward enforcement of plans and actions as stated within the researchers' data management plans. To this end, federal agencies could support and highlight the development of machine-based mechanisms for compliance, audit, provable possession of data, etc. Additionally, data repositories that have undergo external certification and audit through mechanisms such as the Data Seal of Approval may provide a systematic means for addressing compliance.

   f) Promote the deposit of data in publicly accessible databases, where appropriate and available;

   - I am unsure what is meant by "databases" in this context but it would seem that publicly accessible "repositories" or "archives" would be a better choice of terms. Regardless of the type of system or technology, I believe that deposit of data should ensure the assignment of a unique persistent identifier.

   j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.

   - It would be challenging for federal agencies to develop methods for assessment of long-term needs. Beginning with short-term assessments is more likely, particularly as it relates to metrics for assessing value of data. At this point, one could assert that the only metric is citation within a publication. However, as data repositories evolve and proliferate, there will be value with using, discovering, analyzing, etc. data independent of publications. The development of these metrics could represent

another opportunity for partnership between libraries, publishers, scholarly societies and the private sector.

3. Are there some situations or scientific fields where it would be cost-prohibitive to store and share data? Please explain. How should data be shared in these cases?

   There are nascent or planned scientific projects (e.g., Pan-STARRS and LSST in astronomy) that generate so much data with each individual survey of the night sky that there is not even sufficient hard disk to capture all of the data from the entire project. In such cases it is clearly cost-prohibitive to provide public access. It is my understanding that researchers have developed techniques for analyzing or sifting through such data in real-time. For these types of projects, it is perhaps most useful to document the procedures, processes, etc. that are used to analyze the data and the decisions regarding data acquisition, retention, deaccession, etc. in case there is a need to conduct additional surveys in the future.

   On a smaller scale, it is worth noting that in some situations costs could be lowered if researchers relied on economies of scale offered through community-based data repositories and archives. That is, there should be some third party or community based assertion of prohibitive costs, rather than an individual researcher who may not be using the most efficient options or means for data management.

4. One of the reasons for not releasing data in experiments is that it may contain personal identifying information. Is this a legitimate reason on the part of researchers not to share data? Please explain. How can we promote the sharing of such data while also assuring that confidentiality will be maintained?

   Please note that this response includes input from the Inter-university Consortium of Political and Social Research (ICPSR), which has extensive experience with data possessing personal identifying information. Disclosure: I am a member of the ICPSR Council (or Advisory Board).

   In certain domains such as social and behavioral sciences, it is not uncommon to collect personal identifying information in the course of doing research. The success of the social science research enterprise relies on the willingness of research participants to take part in experiments and surveys, and researchers are very aware of their obligation to protect such information. Procedures have been developed to protect confidential information during the research process and to assure that subjects cannot be identified in research publications. Disclosure risk is a term that is often used for the possibility that data from a research study might be linked to a specific person thereby revealing personal information that otherwise could not be known or known with as much certainty.

Concerns about disclosure risk have grown as more datasets have become available online and it has become easier to link research datasets with publicly available external databases.

Safeguards can be applied that allow access to data while at the same time ensuring confidentiality. Archive and repository data managers have developed skills in assessing and mediating disclosure risk and now can apply several approaches and technologies to ensure confidentiality throughout the data lifecycle. Working with these professionals, especially in the data collection planning phases, can allay concerns regarding disclosure risk. These approaches include creating public-use files by modifying the data (e.g., removing identifying numbers such as social security numbers), "coarsening" data (e.g., mentioning time intervals rather than specific dates), suppressing highly unique cases, sub-sampling and adding "noise" to the data.

In cases where data cannot be modified to protect confidentiality without significantly compromising the research potential of the data, access to the data must be restricted and stringent confidentiality safeguards imposed.

In these situations, archives require an application, review, and vetting process. Applicants are required to provide a research plan, Institutional Review Board approval, and a data protection plan. Approved users sign a Data Use Agreement, which establishes the rules for acquiring and using the data, a security pledge, and institutional approval and signatures. The agreement is particularly important because it specifies the guidelines that researchers must follow in the release of statistics derived from a dataset. Violations of the agreement are treated as research misconduct and violations of policies governing scientific integrity. Severe consequences are possible, including suspending research grants and legal liability. After an agreement is processed and approved, data are sent securely on CD, made available for secure download, or provided in a virtual data enclave (VDE), whereby the user must access and analyze the data on secure servers of the data provider. Results of data accessed via a VDE are vetted for disclosure risk prior to being sent to the user.

For data that present especially high disclosure risk, access can be provided in a data enclave where researchers must enter a secure facility to access the data. Investigators must undergo an application and approval process, as previously described, and archive staff reviews their notes and analytic output.

5. Would a move towards open-access of published data cause additional administrative costs for Universities and other Institutions that receive federal funding for scientific research? How can we minimize administrative burdens while simultaneously maximize access to data?

   A movement toward open-access of published data would almost certainly cause additional, administrative costs for universities and institutions that receive federal funding for scientific research. There is a challenging and delicate balance that needs to be struck between the benefits of open-access to data and new, additional costs. On the national scale, we may need to consider this balance in terms of how much new science we wish to support as compared to how much value we wish to extract from existing data.

   There is also a time dimension to consider. As noted earlier, the OSTP memo emphasizes data to validate research findings. This tangible goal represents a useful goal with which to make decisions regarding selection criteria for data. Additionally, systematic approaches to data management will almost certainly require lower costs than relying upon individual researchers' to manage their own data. As data infrastructure evolves, economies of scale arise and marginal costs reduce, it may become possible to consider other, tangible goals or classes of data for open access.

6. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

   It is difficult to know the community reaction to open-access data. While it seems likely that scientists will *initially* want data from a small fraction of paper, the availability of such data might encourage greater discovery, re-use, etc. Focusing on specific goals such as verification of results and citation provide a useful, initial set of objectives for identifying data which should be deposited into repositories or archives. It is important to remember that federal funding is supposed to result in reproducible, citable science. As scalable, more efficient data infrastructure becomes available, both costs and time related to data management should diminish. With more data available, the prospects of unanticipated uses may increase over time. One of my Data Conservancy colleagues once said: "one scientist's noise is another scientist's signal" referring to the conventional wisdom of "one person's garbage is another person's treasure."

It is also worth noting the public's potential interest in scientific data. The experience of PubMed Central has demonstrated that the public does indeed refer to scientific literature for various reasons. The experience with the Sloan Digital Sky Survey (SDSS) provides evidence that similar trends may apply with data. There are approximately 10,000 professional astronomers but there are nearly 1 million registered users of the SkyServer that provides access to SDSS data.

Finally, greater availability of data could inspire the development of tools and services by a host of stakeholders such as scientists, publishers, professional societies and even the general public.

7. What specific infrastructure-technology requirements are required for the storage of scientific research data? Are University libraries or National Laboratories currently equipped with this type of infrastructure technology? Would an entirely new infrastructure need to be developed for the massive storage of data?

   I can only speak to the experience that my colleagues and I have gained through our process of dealing with Sloan Digital Sky Survey (SDSS) data for over a decade. Through our evaluation of storage systems, we have identified that current systems have limitations in terms of data preservation. For example, current storage systems do not possess formal auditing that is necessary for full-fledged preservation. Through personal interactions, I have heard similar concerns from other large-scale storage users such as the Internet Archive and the Science and Technology Council of the Academy of Motion Picture Arts and Sciences. I believe that development of new storage hardware and software based on these data infrastructure requirements represents an ideal opportunity for private-public partnerships that respond to federal funding programs. These funding programs should require working systems in operational environments as an outcome.

8. What are the potential cost-drivers for storing data? What are other costs that need to be considered?

   It is important to note that storing data is necessary, but not sufficient for sustained data sharing, access and preservation. In addition to storing data, archiving (e.g., protection such as checksums or computer generated codes to check integrity of data), preserving (e.g., format migration), and curating (e.g., adding value for re-use) are required.

   Regarding costs of storage, there is an unfortunate perception that storage is cheap so therefore we can store data easily. Not only does this perception ignore archiving, preservation, and curation, it also ignores the reality that storage *management* is not cheap. For example, the costs (in the form of computing cycles) for generating checksums can be significant or for migrating from one format to another (e.g., jpeg to tiff), depending on the amount of data.

There have been systematic attempts to measure costs associated with managing digital assets though the emphasis on data is more recent. For example, the LIFE project (http://www.life.ac.uk/) in the UK has "developed a methodology to model the digital lifecycle and calculate the costs of preserving digital information for the next 5, 10 or 20 years." The Australian National Data Service (ANDS; http://www.ands.org.au/) has developed a business plan. More recently, the OpenAIRE project and the European Commission has announced a tender seeking input for a Sustainability Model and Business Plan for digital infrastructure.

9. Are there any countries that have successfully implemented open-access data-sharing? Could the models used in those countries be used here in the US? Why or why not?

It is fair to assert that, in many ways, Europe and Australia are both better organized than the US with respect to open-access data sharing. In the UK, some funding agencies require deposit of data into publicly accessible repositories. In Australia, the Australian National Data Service (ANDS) provides a national discovery service for open data deposited throughout their country. Arguably, these countries have also implemented data systems at the institutional, community and national levels, understanding that diverse "ecosystem" of approaches and systems are necessary for different functions related to open-access data.

It would be difficult to imagine adopting these models verbatim within the US. There is a difference in scale and diversity of funding sources with the US. That is, there are fewer researchers, universities, etc. that generate data and fewer funding agencies that provide funding in Europe and Australia, many of which share common data management plan requirements. There is much the US can learn from our colleagues in Europe and Australia. We may possibly adopt elements of their approach.

Having noted this, one could make a reasonable argument that while other countries are more advanced in the deposit, discovery and access realms, they are not more advanced in the data preservation realm and, in some cases, US-based data centers such as the Inter-university Consortium for Political and Social Research (ICPSR) and the National Snow and Ice Data Center (NSIDC) have long-term track records with data preservation (at least for certain types of data). Additionally, some new US-led data infrastructure development efforts such as the one I lead at Johns Hopkins (the Data Conservancy) have focused specifically on data preservation. Given this situation, there is comparative advantage to working with our colleagues in Europe and Australia.

NSF (and perhaps other federal funding agencies) often seeks international collaboration as part of solicitations but do not allow use of funds to support international participants. Understandably, this reality makes it challenging to secure international partnerships. There have been joint NSF/JISC and NSF/EU funding programs but these programs can

lead to greater administrative burdens in terms of reporting, oversight, etc. Streamlined programs that foster international partnerships would be worthwhile. The Research Data Alliance (rd-alliance.org) has been launched with a goal of fostering collaboration on a global scale toward data sharing and interoperability. At this point, NSF and NIST are the only two federal agencies directly supporting the Research Data Alliance (RDA). Disclosure: I am involved in RDA, particularly as the leader for the task force planning the 2nd meeting of RDA in Washington, DC from September 16-18, 2013.

10. What specific support could the federal government contribute towards a permanent community-maintained archive for storing research data, that non-federal organizations could not provide?

    The federal government can and should provide funding toward the development of community-maintained data archives. There is value to building infrastructure at scale (i.e., beyond individual universities). While the private sector has an important role to play, certain functions such as preservation – while essential – are unlikely to be profitable. It is worth considering the role of the federal government with other types of existing infrastructure that rely upon a combination of federal, state, university and private funding and resources. If one considers other forms of infrastructure to support data-intensive science such as high-performance computing, there is a diversity of options ranging from university-based or company-based services. Some of these options such as supercomputing centers receive federal funding support. However, even in cases of federal support, there should be a real sustainability plan that does not rely upon additional rounds of federal investment.

11. A 2007 GAO Report entitled "Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research" states: "*The scientific community generally rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure.*" What are your suggestions to change this structure?

    This matter relates to the reward and recognition structure that is part of universities' academic policies and practices. There is a tremendous diversity and complexity to this framework that the federal government cannot address. Having said this, there are existing mechanisms within the federal funding environment that can be leveraged effectively. For example, NSF recently changed its guidelines such that instead of mentioning "five most relevant publications" within the NSF-compliant two-page bios, one can know list "five most relevant products" ostensibly to include other output of research such as data. Similar mechanisms should be leveraged as well. If data are included in this manner (e.g., NSF two-page bio), then they should be cited using a persistent identifier to ensure reliable, sustained ability to discover and review such data.

12. What specific technical standards need to be considered when storing data for open access?

    There are many existing standards. Consider the growing list that the Digital Curation Centre in the UK maintains at http://www.dcc.ac.uk/resources/metadata-standards/list. Each scientific community has its own set of metadata standards. There are attempts to map between these standards but it is perhaps more important to focus on data types. The aforementioned Research Data Alliance (RDA) has two working groups focused on persistent identifier types and data type registries. These groups are considering the various *types* of data (e.g., images, videos), the salient or representative properties of these types, and the role of persistent identifiers with these data types. This type of foundational work focused on data types and identifiers is necessary before considering a universal set of metadata standards that may be applied across a variety of domains and contexts.

13. What federal agency and/or other entities would be appropriately suited to determine standards for storing data?

    As mentioned, the Research Data Alliance has undertaken global community-driven and guided work in this regard. The National Institute of Standards and Technology (NIST) would seem to be an appropriate agency in this context. Various federal funding agencies have natural connections to various scientific communities (e.g., NASA with space sciences and earth sciences) in a manner that facilities development of community-based standards.

14. On July 29, 2010 Dr. David Lipman testified before the House Subcommittee on Information Policy, Census and National Archives. While most of his testimony centered around open-access issues, he noted that the National Center for Biotechnology Information (NCBI) produces more than 40 databases, including GenBank and dbGaP. He also mentioned other data intensive activities that his center is currently handling. Based on his testimony, and other publically available information about the activities at NIH Pubmed Central and NCBI, do you think that they have the technical capability and infrastructure to store, archive, and handle large amounts of data (i.e. achieve the purposes of open-data)? Please explain. If there was a movement towards a national repository for scientific data, would it not be better to build off of existing infrastructure at NIH and NCBI? What are other issues that should be taken into consideration when going towards a single repository model? Finally, based on your experience, do you see any potential cross-agency issues (for example between NIH and NSF) that might make a single federal repository inefficient or not worthy of pursuing?

    I do not know enough about the technical capability and infrastructure of NCBI to comment in detail. I was the Principal Investigator of an NSF-funded evaluation of pros

and cons for a potential open-access repository of publications resulting from NSF funding. Based on this evaluation, I can offer the following observations or comments.

One of the main reasons that NIH can provide infrastructure for publications and data is the existence of the National Library of Medicine (NLM), which is itself a type of infrastructure. Noting that other funding agencies such as NSF do not have an equivalent resource, it is worth considering whether NIH or NLM could provide relevant infrastructure or services. Having said this, while the approaches and processes that NIH or NLM have undertaken might be useful, it is not clear that the specific choices and workflows would apply effectively to other scientific domains or communities.

As with other infrastructure development, there needs to be a balance between national or centralized approaches and community or decentralized approaches. A national repository could offer significant economies of scale (e.g., for storage) but might result in too rigid a framework to effectively describe or share data across a diverse set of domains or communities.

It may be more effective for the federal government to identify cross cutting, common components of data infrastructure that could be applied across different funding agencies. For example, referring to the aforementioned discussion of data types and identifiers, the federal government could require funding agencies to mandate the use of persistent identifiers but not prescribe the specific choices. This type of approach represents a balance between an overarching national approach that recognizes the need for flexibility within scientific communities.