To: "Dr. Baruch Fischhoff - Chair, National Academy Committee on Improving Intelligence" <baruch@cmu.edu>, "Dr. Theda Skocpol - National Academy of Sciences and Past President, APSA" <ts@wjh.harvard.edu>, "Bill Nordhaus - National Academy of Sciences" <william.nordhaus@yale.edu>, "Dr. David Shaw - PCAST"<dshaw@blackpointgroup.com>, "Dr, Gene Rosa - Chair, AAAS Section K" <rosa@wsu.edu>, "Dr. Carole Pateman - President, APSA" <pateman@ucla.edu>, "Dr. Robert Keohane-National Academy of Sciences" <rkeohane@princeton.edu>, "Dr. Robert Axelrod - National Academy of Sciences" <axe@umich.edu>, "Dr. Jonathan Cole - CASBS" <jrc5@columbia.edu>, "Dr. Richard Atkinson - Chair - NRC/DBASSE" <rcatkinson@ucsd.edu>, "Dr. G. Bingham Powell, Jr. - APSA Vice President" <gb.powell@rochester.edu>, "Dr. Kenneth Prewitt - Chair, Committee on Social Science Evidence for Use" <kp2058@columbia.edu>, "Dr. Anne-Marie Slaughter - Director, Policy Planning Staff via Ms. Marisa S. McAuliffe" <mcauliffems@state.gov>, "Dr. Kwame Anthony Appiah - Chair, Exec. Committee, American Council of Learned Societies" <kappiah@Princeton.EDU>, Dean David Ellwood <david_ellwood@harvard.edu>, "Prof. Derek Bok" <derek_bok@harvard.edu>, "Dr. Mitchel B. Wallerstein - Dean" <mwallers@syr.edu>, "Dr. Nina Fedoroff - AAAS President" <nvf1@psu.edu>
From: Lloyd Etheredge <lloyd.etheredge@policyscience.net>

Subject: 246. Red Team: Recapitalizing social science and upgrading NSF Merit Review/Performance Measures by 3/31/2011

Dear Dr. Fischhoff, Dr, Prewitt, Dr. Atkinson, Dr. Skocpol, and Colleagues:

I discussed earlier (e.g., # 242 at www.policyscience.net at II.D) formal institutional mechanisms that can be used, under the Government Performance and Results Act, to upgrade and energize NSF support for the SBE sciences.

I enclose a copy of a recent letter to Dr. Suresh and the National Science Board about activating these mechanisms. The public comment period, which is the most appropriate time to identify new recommendations for the NSF Merit Review system, is open until the end of this month and there is an online feature that can be used. My online submission emphasized the benefits of specific stakeholder consultations with the DNI system to inform NSF's agenda for understanding the world beyond the water's edge, its allocation of funds, and the design of its systems for rapid national learning. Also: addressing the catastrophic failure of the NSF economics program.

Especially, I believe that an urgent need is for new government-funded "everything included" data systems [subject to privacy restrictions] online, that everyone can use. These include content analysis capabilities where the DNI system also can help by moving reference databases and analysis software, quickly, into the public domain. And new data systems for multi-disciplinary economics that can raise GDP/capita growth by 1% above the pre-crisis baseline in the US and other countries.

NSF and the Legacy of Political Pressures

General Clapper and his senior staff may have the incorrect impression that NSF operates independently of political influence. In fact, unless he and other government stakeholders ask for many lines of research that could be interpreted as critical of government (especially Republican) policies, these may not be funded. How many original, evidence-based criticisms of important government policy, funded by NSF, have you heard from academic social science in recent decades? The history of NSF's declining programs is discussed, from my perspective, in the letter to Dr. Suresh, which also includes additional detail about the secret accommodationist battles inside our national scientific Establishment.

LE

Dr. Lloyd S. Etheredge - Director, Government Learning Project
Policy Sciences Center

URL: www.policyscience.net
301-365-5241 (v); lloyd.etheredge@policyscience.net (email)

[The Policy Sciences Center, Inc. is a public foundation that develops and integrates knowledge and practice to advance human dignity. Its headquarters are 127 Wall St., Room 322 PO Box 208215 in New Haven, CT 06520-8215. It may be contacted at the office of its Chair, Michael Reisman (michael.reisman@yale.edu), 203-432-1993. Further information about the Policy Sciences Center and its projects, Society, and journal is available at www.policysciences.org.]

THE POLICY SCIENCES CENTER, INC.

Project Director: DR. LLOYD ETHEREDGE
7106 Bells Mill Rd.
Bethesda, MD 20817-1204
Tel: (301)-365-5241
E-mail: lloyd.etheredge@policyscience.net

March 6, 2011

Dr. Subra Suresh, Director
National Science Foundation
4201 Wilson Blvd., Room 1205 N
Arlington, VA 22230

Dear Dr. Suresh:

I replied to the National Science Board's invitation to recommend improvements for NSF's Merit Review process. However, this raises many serious issues of NSF's quiet political accommodations that damaged the social, behavioral, and economic (SBE) sciences during the era of Republican mindlessness. I write to bring these breakdowns to your personal attention because you must understand them to restore the integrity of the NSF system and restore health and rapid scientific progress in these areas of your responsibility.[1]

The fate of the SBE sciences is part of the background to President Obama's new directive to assure political independence and the rights of all applicants to receive just, honest, and competent evaluation based on scientific merit. Many years of cumulative breakdowns of integrity and (known) merit review problems ("folded lies" - Auden) have layered upon themselves and probably contribute to the catastrophic failure of the NSF economics program.

The Historical Context

Detailed reviews and filings with the Department of Justice will be available to you in NSF files and online.[2] In brief: NSF's problems began in the Reagan years when the late Donald Campbell ("Reforms as Experiments") was one of the most exciting and enrolling social scientists in the country. The vision to use scientific methods for rapid learning and evidence-based social, political, economic and international policy built upon a foundation by Lasswell and others and inspired a generation. (It shaped the undergraduate curriculum at MIT, where I taught for eight years and was a founding and core faculty member of the public policy concentration in course 17.) Then, Reagan's first OMB Director (Stockman), with *hubris* and juvenalia, launched a "defund the Left" pre-emptive strike and threatened to zero-out all behavioral science in the federal budget. Senior members of our national science Establishment, including NSF's leadership, disregarded their legal, professional, and ethical obligations to scientific integrity and merit-based evaluations.[3] Rather than fight Stockman - as an earlier generation fought the Right Wing attack of Senator McCarthy - they agreed to a politically neutralized future for the SBE sciences. The Luce Commission (NAS/NRC) was funded by NSF to designate new "leading edges." Although the new Republican policy ideas were an obvious (indeed, an ideal) scientific opportunity for rapid learning in the Campbell tradition, Frank Press (President of the National Academy of Sciences) and Luce gave Stockman everything that he wanted: Of 1700+ scientists cited in the index of the new National Academy/NSF "leading edges" plan, none was Donald Campbell. He became a "non-person." Stockman et al. altered the traditional civic role of our universities without the public battle that they would have lost.

Many people should have stood up to Republican zealots over the years, but they did not do so. You have inherited the problem.

The Later Battles

There were many further battles as other scientists began to realize what the NSF-funded project had quietly done.[4] I contacted Dr. David Hamburg who organized an off-the-record meeting of the Carnegie Commission on Science, Technology and Government (co-chaired by Joshua Lederberg) but we lost the earlier rounds. Republican ideologues and lobbyists especially demanded an unobstructed path for deregulation and a political victory for a "free market." The fate of the NSF program in economics alarmed me and I also was involved in several later initiatives to warn that econometric models and data systems had too many missing variables, too many self-limiting and untested ideological assumptions, and were (demonstrably) losing contact with reality in a changing world. [The enclosed supporting letter from Robert Reischauer, former head of the CBO and a member of the Executive Committee of Harvard's Board of Overseers, was seen by the Bush-era appointees at NSF and on the National Science Board - who ignored honest warnings by this route (and from others) several times before the catastrophic failure.]

Scientific professionals are granted autonomy for self-governance (and academic tenure) because they are trusted to behave with honor. Today, NSF has an obligation to Justice, to the country, and to individuals to make compensation for its breakdowns of integrity and the damage that it caused. Great damage has been done within the SBE disciplines, to Departments at leading universities, to evidence-based undergraduate education in the social sciences, to the careers of individual scientists who deserved to have their life work evaluated on the basis of its scientific merit and civic contribution, and - very painfully and visibly - to the growth of multi-disciplinary, reality-connected economics and the well-managed market economy that is part of NSF's stewardship portfolio. There is a large backlog of investments in new methods and data systems (e.g., hierarchical psychodrama models, computer-assisted content analysis of media databases (domestic and international), comparative political psychology/behavior, an Honest Broker project to reduce the range of ideological disagreement, Lotka-Volterra models of global finance and political economy).[5]

Righting Past Wrongs and Building the Future

It will be easier to establish the conceptual scientific case for compensation than to decide upon specific amounts that justice demands. However you may find that the legal framework for the equivalent of a class action settlement is informative because, as outlined for the Justice Department review, intentional fraud was involved. The Luce Commission did its work dishonestly, with a quiet agreement from Frank Press to suspend the traditional standards of ethics and integrity at the National Academy of Sciences. There was no independent review to audit how Luce's "merit review" was to be conducted and how the decision was made to designate the "leading edge" beneficiaries v. the research programs that were to be killed or deemed unworthy. Luce and his inner circle established a coalition of safe and politically innocuous research programs whose members merely mutually certified one another as "leading edge" winners. Frank Press, under his NSF contract, also allowed Luce et al. to use the study process to prepare detailed 10-year budgets for winners (including Luce and his friends) and these were transmitted to all government agencies and major foundations, with the imprimatur of the National Academy of Sciences, without telling the wider scientific community of the plan. Thus, Luce and his friends were freed from the established conflict of interest expectations of the scientific community governing scientific advice and merit review and they steered the process brilliantly to secure competitive advantages and personal financial benefit. [And, in the Luce v. Campbell battle for national scientific prominence, Luce was able to dispatch a rival. I infer from a telephone call from Campbell (a member of the National Academy of Sciences) that there were strong exchanges within the National Academy of Sciences (behind closed doors and via email) about these issues.] Next, the NSF Director stonewalled and did not ask that the Report's priorities be redone, nor warn other government agencies and foundations - or even NSF's professional staff - of the scientific, due process, and ethical corruptions and political bias. Thus, by American standards of justice, the honorable compensation appears to include at least triple (punitive) damages plus compounded interest.

Scientific Consultations

These issues may appear to involve only Washington-based institutions. However, your ability to understand what has happened is greater, with your MIT background, because the breakdowns also involved Cambridge-based processes and divisions within our nation's scientific Establishment.[6] Today, Prof. Gary King at Harvard has re-engaged the dream to apply science to social, economic, political and foreign policy questions and rekindle, into an illuminating blaze, what Carl Sagan called our "candle in the dark." In his recent article in Science (2/11/2011) King notes the "severe challenges . . . holding back progress" in data systems for the SBE sciences.[7] [For our scientific Establishment to allow - after fifteen years, as King discusses - an honest article in Science to mention "severe" SBE problems in print also is a breakthrough.][8] The problems discussed by King imply names of specific institutions and individuals that you should know before you make the required personnel and institutional changes to right these past wrongs and follow President Obama's instructions that such NSF breakdowns must never happen again.[9]

Yours truly,

Lloyd S. Etheredge

(Dr.) Lloyd S. Etheredge, Director
Government Learning Project

cc: Dr. John Holdren - Science Adviser; Dr. Ray Bowen - Chair, National Science Board; Dr. Kenneth Prewitt (Columbia) - President, COSSA

Attachments:

- Robert Reischauer (personal communication), December 23, 2002.

- - Gary King, "Ensuring the Data-Rich Future of the Social Sciences," Science, February 11, 2011, pp. 719-721.

Notes

1. NSF and the National Academy of Sciences have an abundance of clear rules requiring scientific integrity and merit review from their officials, employees, and advisers. Merely rewriting the rules will not solve NSF's problems nor bring our full national resources online to meet the challenges ahead.

2. The filing with the Department of Justice concerning the Luce Commission is at http://www.policyscience.net at II. A. (September, 2007). See also the filings with AAAS in this section; and *passim.*

3. There are scientists, among the accommodationists, who claim that trying to save the SBE disciplines "would not have made much difference." This is like Southern judges violating the law and imposing death penalties against Blacks with the excuse that a "lynch mob would have killed them anyway." The American system of government is designed with many checks and balances to stop abuses of political power by zealots and the first duty of NSF and scientists is for scientific integrity.

The Supreme Court, with honor and integrity, can publicly decline to decide a case because it lacks jurisdiction or believes that the issue is properly decided by the political process: However, this is not what NSF did.

4. Re further details: The National Research Council's staff knew of my NSF grant to study and improve government learning rates and they had invited me to submit ideas to the Luce Commission process. Later, other members of the professional staff recommended an oversight review of the Luce project and a Campbell/Lasswell tradition "leading edge" national project to evaluate ideological arguments quickly, developing new measures and using the model of the Michelson-Morley experiment in physics. [I was invited to submit a draft and was able, personally, to follow the history as the National Academy of Sciences President (Frank Press, from MIT) decided to kill this "leading edge" project without sending the idea to a social science advisory panel for written evaluations of scientific merits and support. The scientists were able (later) to get an "informal discussion," without written minutes. I spoke with two scientists who attended: Philip Converse told me of deep concerns (from the early days of his University of Michigan projects and centers) that ideological objections might wipe-out all government funding for behavioral science research concerning American citizens; and Sidney Verba told me that "if I was a younger man, I would jump on [these hierarchical psychodrama models]" that I had proposed as part of the strategy; but, he said, there was nothing that he could do. A member of an NSF advisory committee in the Bush era told me that the new paradigm "scared people" - but he gave no details.

5. A very partial list of the SBE innovations and lines of investigation killed, *de facto*, by the Bush-era NSF is at www.policyscience.net at II. D. for the Fischhoff advisory process to the Director of National Intelligence. See also the Recapitalization ideas at II. A. (2010) and the ideas for the last NSF Five-Year plan that (to judge from an FOIA request) did not receive a written merit review that NSF was willing to release, at II. A. (March, 2007).

6. I received my first NSF grant, at MIT, to develop the multi-disciplinary study of government learning and my project generated the new paradigm to render ideological truth claims quickly testable. Robert Solow sent me a note congratulating me on the approach to testing and commending the empathy-based account of Reagan's different way of thinking about economics as "exactly right." Frank Press came to Washington from MIT and his lack of political courage (and his astuteness) crafted this dark chapter that enrolled scientists themselves in dishonorable complicity. My friend and MIT's former Dean of Humanities, Bruce Mazlish, emerged on the other side, arguing that "the American people aren't ready" for evidence-based public policy. David Hamburg, formerly of Harvard, organized high-level Establishment opposition. A former President of MIT (as Chair of the Sloan Foundation, which had given support to the Luce project) also worked quietly to raise issues without a public confrontation. John Holdren at Harvard became involved in these breakdowns as AAAS President when a key issue was the long-standing decision by the last Editor-in-Chief of Science, Donald Kennedy (a Harvard doctorate and former President of Stanford), not to permit any news reporting in Science to alert the rank-and-file scientific community, SBE scientists, and/or AAAS members to the off-the-record meetings of the Carnegie Commission (of which he had personal knowledge), and many years of deadly and secretly contentious accommodations to Republicans.

7. See references in footnote 5, above. By now, political neutralization also has produced extraordinary institutional problems, especially with the rise of new communications systems and the wide awareness of uncorrected breakdowns of integrity in merit review.

8. See also the correspondence with AAAS at www.policyscience.net at II. A.

9. Relying upon NSF's Inspector General in these cases of top-level political derailing of Merit Review will not work. I have a degree of sympathy for the NSF Inspector General, to whose staff very few people were willing to speak candidly. The Inspector General assigned a woman with a background in computer science, and without training to understand the scientific issues and personalities. She did not have subpoena power, nor could she offer whistle blower protection to people who would put careers and professional relations at risk. [In my initial discussion with her, when I mentioned scientific advisory panels who had altered their recommendations to avoid political attack by Republicans, she asked, "Don't you think that is common in Washington?"] The secret decision to prejudice and kill entire classes of applications/lines of investigation, without merit review, already had been made and reviewed above her pay grade, by the top levels of NSF and the NSB and had been reviewed by the scientific Establishment via the Carnegie Commission. The knowledgeable NSF Directors and NSB President, who could have opened doors and allowed her to penetrate the intense social pressures supporting the code of silence at the National Academy of Sciences, did not do so - and the people she interviewed were aware that they had not received any telephone calls on her behalf. Later, her superior told me that they feared being "slapped down" if they ventured into "scientific" areas. The NSF Inspector General learned less than what participants in the off-the-record session already knew.

THE URBAN INSTITUTE 2100 M STREET, N.W. / WASHINGTON D.C. 20037

ROBERT D. REISCHAUER
President

Direct Dial: 202-261-5400
Fax: 202-223-1335
E-mail: RReischa@ui.urban.org

December 23, 2002

Dr. Lloyd S. Etheredge, Director
Government Learning Project
The Policy Sciences Center, Inc.
P. O. Box 208215
New Haven, CT 06520-8215

Dear Dr. Etheredge:

Thank you for your letter and thoughtful attachment. I am in complete agreement that the economic data we collect has significant deficiencies that limit our ability to understand the economy's problems and chart future policy.

We don't collect some information that is needed and gather much that we could do without. We collect other data in insufficient detail and almost always take too long to release the data for it to be useful in policy decisions.

As you know better than I, there are many reasons for this situation. What we collect and how we collect it reflects the forces at play in the first half of the last century and those forces do not want to give anything up. Congress has little interest in devoting more scarce budget resources to collect new and better information. Few economists who use the data appreciate its limitations. They have been raised on certain data sets and treat them as if they are part of the underlying environment, not subject to change. They put a premium on continuity and don't want discontinuity in the data sets they know and use.

I don't think I would be as critical as you are about CNSTAT/NCR. I don't think they would have much of an impact even if they had done the studies and made the recommendations you think warranted. Nor do I think universities (Yale or Harvard) or the Fed could make much of a dent in the problem. Rather, I think a presidential or congressional study commission is called for—one with a clear mandate and a promise that added resources will be devoted to strengthening the statistical system based on the commission's report. Unfortunately, the prospects for such an initiative rising to the top of policymakers' lists of things to do is very, very low.

Nevertheless, I wish you well in your efforts.

Sincerely,

3. The CMS Collaboration, *J. Instrumentation* **3**, S08004 (2008).
4. U.S. National Academy of Engineering and Royal Academy of Engineering, Frontiers of Engineering, EU-US Symposium, Cambridge, UK, 31 August to 3 September 2010; www.raeng.org.uk/international/activities/frontiers_engineering_symposium.htm.
5. E. J. Candès, J. Romberg, T. Tao, *IEEE Trans. Inf. Theory* **52**, 489 (2006).
6. D. L. Donoho, *IEEE Trans. Inf. Theory* **52**, 1289 (2006).
7. J. B. Tenenbaum, V. de Silva, J. C. Langford, *Science* **290**, 2319 (2000).
8. R. G. Baraniuk, M. B. Wakin, *Found. Comput. Math.* **9**, 51 (2009).
9. S. Muthukrishnan, *Found. Trends Theor. Comput. Sci.* **1** (issue 2), 117 (2005).
10. N. Snavely, S. M. Seitz, R. Szeliski, *ACM Trans. Graph.* **25**, 835 (2006).

10.1126/science.1197448

PERSPECTIVE

Ensuring the Data-Rich Future of the Social Sciences

Gary King

Massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress. I address these changes and challenges and suggest what can be done.

Fifteen years ago, *Science* published predictions from each of 60 scientists about the future of their fields (*1*). The physical and natural scientists wrote about a succession of breathtaking discoveries to be made, inventions to be constructed, problems to be solved, and policies and engineering changes that might become possible. In sharp contrast, the (smaller number of) social scientists did not mention a single problem they thought might be addressed, much less solved, or any inventions or discoveries on the horizon. Instead, they wrote about social science scholarship—how we once studied *this*, and in the future we're going to be studying *that.*

Fortunately, the editor's accompanying warning was more prescient: "history would suggest that scientists tend to underestimate the future" (*2*).

Indeed. What the social scientists did not foresee in 1995 was the onslaught of new social science data—enormously more informative than ever before—and what this information is now making possible. Today, huge quantities of digital information about people and their various groupings and connections are being produced by the revolution in computer technology, the analog-to-digital transformation of static records and devices into easy-to-access data sources, the competition among governments to share data and run randomized policy experiments, the new technology-enhanced ways that people interact, and the many commercial entities creating and monetizing new forms of data collection (*3*).

Analogous to what it must have been like when they first handed out microscopes to microbiologists, social scientists are getting to the point in many areas at which enough information exists to understand and address major previously intractable problems that affect human society. Want to study crime? Whereas researchers once relied heavily on victimization surveys, huge quantities of real-time geocoded incident reports are now available. What about the influence of citizen opinions? Adding to the venerable random survey of 1000 or so respondents, researchers can now harvest more than 100 million social media posts a day and use new automated text analysis methods to extract relevant information (*4*). At the same time, parts of the biological sciences are effectively becoming social sciences, as genomics, proteomics, metabolomics, and brain imaging produce large numbers of person-level variables, and researchers in these fields join in the hunt for measures of behavioral phenotypes. In parallel, computer scientists and physicists are delving into social science data with their new methods and data-collection schemes.

The potential of the new data is considerable, and the excitement in the field is palpable. The fundamental question is whether researchers can find ways of accessing, analyzing, citing, preserving, and protecting this information. Although information overload has always been an issue for scholars (*5*), today the infrastructural challenges in data sharing, data management, informatics, statistical methodology, and research ethics and policy risk being overwhelmed by the massive increases in informative data. Many social science data sets are so valuable and sensitive that when commercial entities collect them, external researchers are granted almost no access. Even when sensitive data are collected originally by researchers or acquired from corporations, privacy concerns sometimes lead to public policies that require the data be destroyed after the research is completed—a step that obviously makes scientific replication impossible (*6*) and that some think will increase fraudulent publications (*7*).

Indeed, we appear to be in the midst of a massive collision between unprecedented increases in data production and availability about individuals and the privacy rights of human beings worldwide, most of whom are also effectively research subjects (Fig. 1).

Consider how much more informative to researchers, and potentially intrusive to people, the new data can be. Researchers now have the possibility of continuous-time location information from cell phones, Fastlane or EZPass transponders, IP addresses, and video surveillance. We have information about political preferences from person-level voter registration, primary participation, individual campaign contributions, signature campaigns, and ballot images. Commercial information is available from credit card transactions, real estate purchases, wealth indicators, credit checks, product radio-frequency identification (RFIDs), online product searches and purchases, and device fingerprinting. Health information is being collected via electronic medical records, hospital admittances, and new devices for continuous monitoring, passive heart beat measurement, movement indicators, skin conductivity, and temperature. Extensive quantities of information in unstructured textual format are being produced in social media posts, e-mails, product reviews, speeches, government reports, and other Web sources. Satellite imagery is increasing in resolution and scholarly usefulness. Social everything—networking, bookmarking, highlighting, commenting, product reviewing, recommending, and annotating—has been sprouting up everywhere on the Web, often in research-accessible ways. Participation in online games and virtual worlds produces even more detailed data. Commercial entities are scrambling to generate data to improve their business operations through tracking employee behavior, Web site visitors, search patterns, advertising click-throughs, and every manner of cloud services that capture more and more information.

Efforts in the social sciences that make data, code, and information associated with individual published articles available to other scholars have been advancing through software, journal policies, and improved researcher practices for some time (*8*, *9*). However, this movement is at risk of

Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, MA 02138, USA. E-mail: king@harvard.edu

collapsing unless the improvements in methods for sharing sensitive, private, or proprietary data (*10*) are able to be modified fast enough to keep up with the changes in the types and quantities of data becoming available and unless public policy adapts to permit and encourage researchers to use them. The necessary technological innovations are more difficult than it may seem. For example, the venerable strategy of anonymizing data is not very useful when, for example, date of birth, gender, and ZIP code alone are enough to personally identify 87% of the U.S. population (*11*). And the cross-classification of 10 survey questions of 10 categories each contains more unique classifications than there are people on the planet. And now think of the challenges of sharing continuous-time cell phone–location information from a whole city, or biological information with hundreds of thousands of variables. The political situation is also complicated, with a media storm generated by each new revelation of how personal information is becoming publicly available, but at the same time citizens are voluntarily giving up more privacy than ever, such as via the rapid transition from private e-mail to public or semi-public social media posts.

If privacy can be protected in a way that still allows data sharing, considerable progress can be made for people everywhere without harm coming to any one research subject. This seems easier than, for example, the situation with most randomized medical experiments, in which if everything works as expected those in one treatment arm will be harmed relative to those in the other arms. Moreover, most concern about data sharing involves individuals, whereas social scientists usually seek to make generalizations about aggregates, and so spanning the divide is often possible with appropriate statistical methods.

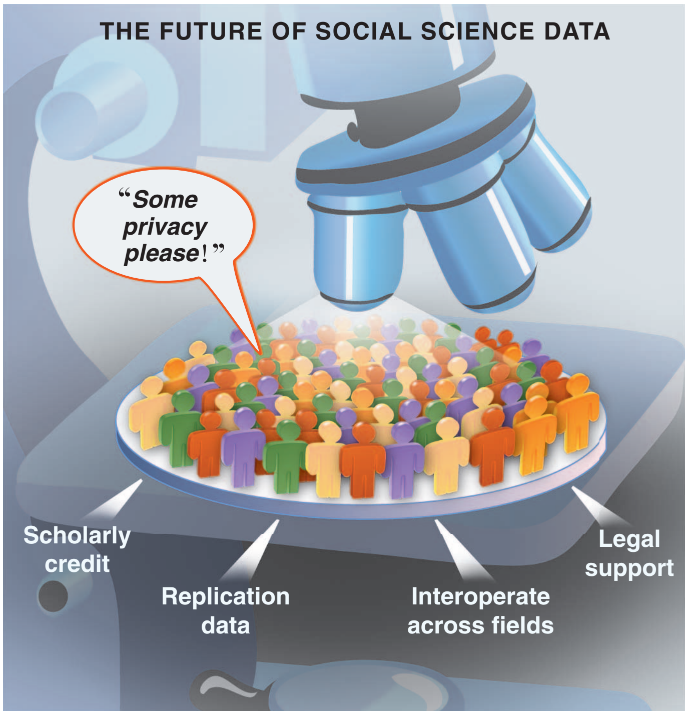

Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

What can we do to take advantage of the new data while facilitating data sharing and at the same time protecting privacy? First, before we try to convince other parts of society to give us some leeway, we social scientists need to get our own act together. At present, large data sets collected by social scientists in most fields are routinely shared, but the far more prevalent smaller data sets that are unique or derived from larger data sets are regularly lost, hidden, or unavailable—often making the related publications unreplicable. In most cases, many data sets associated with individual publications, and the related computer code and other information necessary to reproduce the published tables and figures from the input data, are not available unless you obtain permission of the original author, with no enforceable rules governing when access must be provided. This deserves serious reconsideration and action. We need to devolve Web visibility and scholarly credit for the data to the original author while ensuring that the data are professionally archived with access standards formalized in rules that do not require ad hoc decisions of or control by the original author (*12*, *13*).

Second, we need to nurture the growing replication movement (*14*, *15*). More individual scholars should see it as their responsibility to deposit data and replication information in public archives, such as those associated with the Data Preservation Alliance for the Social Sciences (*16*). More journals should encourage or require authors to make data available as a condition of publication, and granting agencies should continue to encourage data sharing norms. More importantly, when we teach we should explain that data sharing and replication is an integral part of the scientific process. Students need to understand that one of the biggest contributions they or anyone is likely to be able to make is through data sharing (*8*).

Third, we need to continue research into privacy-enhanced data sharing protocols (*10*) and to communicate better what is possible to government officials. Modern technology allows hundreds of millions of people to do electronic banking, commerce, and investing on the web; to view their personal medical records; to store their photographs, videos, and personal documents online; and to share with selected individuals their most private thoughts and secrets. So why, when analyzing these and other personally identifiable sensitive data for the public good, does policy regularly require researchers (through university Institutional Review Boards) to do their work in locked rooms without access to the Internet, other data sources, electronic communication with other researchers, or many of their usual software and hardware tools? Surely we can develop policies, protocols, legal standards, and computer security so that privacy can be maintained while data sharing and analysis proceeds in far more convenient, efficient, and productive ways. Progress in social science research would be greatly accelerated if policies merely allowed researchers more often—as they do corporations, governments, and private citizens—to analyze sensitive data using appropriate digital rather than physical security.

Fourth, even when privacy is not an issue, data sharing involves more than putting the data on a Web site. Scientists and editors of scholarly journals are not professional archivists, and many homegrown one-off solutions do not last long. Data formats have been changing so fast that archiving standards require special preservation formatting, using internationally agreed-upon metadata protocols and appropriate data citation standards. Social scientists need to continue to build a common, open-source, collaborative infrastructure that makes data analysis and sharing easy (*9*, *16*). However, unless we are content to let data sharing work only within disciplinary silos—which of course makes little sense in an era when social science research is more interdisciplinary than ever—we need to develop solutions that operate, or at least interoperate, across scholarly fields.

Last, social scientists could use additional help from the legal community (*17*). Standard intellectual property rules and data use agreements need to be developed so that every data set does not have its own essentially artisan legal work that merely increases transaction costs and reduces data sharing. The federal government should reconsider and relax the rules that prevent academic researchers from collecting, sharing, and publishing from data that those in other sectors of society do routinely.

Of course, social scientists have plenty to do even before we publish and share data. We must find ways of educating students about nonstandard data types, computational methods that scale, legal protocols, data sharing norms, and statistical tools that can take advantage of the new opportunities. Data are now arriving fast enough that the work life of many current social scientists is observably changing: Whereas they once sat in their offices working on their own, rates of co-authorship are increasing fast, and a collaborative laboratory-type work model is emerging in many subfields. These trends would be greatly facilitated by universities and funding agencies recognizing the need to build the infrastructure to support social science research.

For the first time in many areas of the social sciences, new forms and quantities of information may well make dramatic progress possible. Will we be ready?

References and Notes

1. H. Weintraub *et al.*, *Science* **267**, 1609 (1995).
2. D. E. Koshland, *Science* **267**, 1575 (1995).
3. G. King, K. Scholzman, N. Nie, Eds., *The Future of Political Science: 100 Perspectives* (Routledge, New York, 2009), pp. 91–93.
4. D. Hopkins, G. King, *Am. J. Pol. Sci.* **54**, 229 (2010).
5. A. M. Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (Yale Univ. Press, New Haven, 2010).
6. C. Mackie, N. Bradburn, Eds., *Improving Access to and Confidentiality of Research Data* (National Research Council, Washington, DC, 2000), p. 49.
7. R. F. White, *The Independent Review* **XI**, 547 (2007).
8. G. King, *PS Pol. Sci. Polit.* **39**, 119 (2006).
9. The Dataverse Network, http://TheData.org.
10. C. C. Aggarwal, P. S. Yu, Eds., *Privacy-Preserving Data Mining: Models and Algorithms* (Springer, New York, 2008).
11. L. Sweeney, *J. Law Med. Ethics* **25**, 98 (1997).
12. G. King, *Sociol. Methods Res.* **36**, 173 (2007).
13. M. Altman, G. King, *D-Lib* **13**, 10.1045/march2007-altman (2007).
14. G. King, *PS Pol. Sci. Polit.* **28**, 494 (1995).
15. R. G. Anderson, W. H. Green, B. D. McCullough, H. D. Vinod, *J. Econ. Methodol.* **15**, 99 (2008).
16. DATA-Pass, www.icpsr.umich.edu/icpsrweb/DATAPASS/.
17. V. Stodden, *Int. J. Comm. Law Pol.* **13**, 1 (2009).

18. My thanks to M. Altman and M. Crosas for helpful comments on an earlier version.

10.1126/science.1197872

PERSPECTIVE

Metaknowledge

James A. Evans* and Jacob G. Foster

The growth of electronic publication and informatics archives makes it possible to harvest vast quantities of knowledge about knowledge, or "metaknowledge." We review the expanding scope of metaknowledge research, which uncovers regularities in scientific claims and infers the beliefs, preferences, research tools, and strategies behind those regularities. Metaknowledge research also investigates the effect of knowledge context on content. Teams and collaboration networks, institutional prestige, and new technologies all shape the substance and direction of research. We argue that as metaknowledge grows in breadth and quality, it will enable researchers to reshape science—to identify areas in need of reexamination, reweight former certainties, and point out new paths that cut across revealed assumptions, heuristics, and disciplinary boundaries.

What knowledge is contained in a scientific article? The results, of course; a description of the methods; and references that locate its findings in a specific scientific discourse. As an artifact, however, the article contains much more. Figure 1 highlights many of the latent pieces of data we consider when we read a paper in a familiar field, such as the status and history of the authors and their institutions, the focus and audience of the journal, and idioms (in text, figures, and equations) that index a broader context of ideas, scientists, and disciplines. This context suggests how to read the paper and assess its importance. The scope of such knowledge about knowledge, or "metaknowledge," is illustrated by comparing the summary information a first-year graduate student might glean from reading a collection of scientific articles with the insight accessible to a leading scientist in the field. Now consider the perspective that could be gained by a computer trained to extract and systematically analyze information across millions of scientific articles (Fig. 1).

Metaknowledge results from the critical scrutiny of what is known, how, and by whom. It can now be obtained on large scales, enabled by a concurrent informatics revolution. Over the past 20 years, scientists in fields as diverse as molecular biology and astrophysics have drawn on the power of information technology to manage the growing deluge of published findings. Using informatics archives spanning the scientific process, from data and preprints to publications and citations, researchers can now track knowledge claims across topics, tools, outcomes, and institutions (*1*–*3*). Such investigations yield metaknowledge about the explicit content of science, but also expose implicit content—beliefs, preferences, and research strategies that shape the direction, pace, and substance of scientific discovery. Metaknowledge research further explores the interaction of knowledge content with knowledge context, from features of the scientific system such as multi-institutional collaboration (*4*) to global trends and forces such as the growth of the Internet (*5*).

The quantitative study of metaknowledge builds on a large and growing corpus of qualitative investigations into the conduct of science from history, anthropology, sociology, philosophy, psychology, and interdisciplinary studies of science. Such investigations reveal the existence of many intriguing processes in the production of scientific knowledge. Here, we review quantitative assessments of metaknowledge that trace the distribution of such processes at large scales. We argue that these distributional assessments, by characterizing the interaction and relative importance of competing processes, will not only provide new insight into the nature of science but will create novel opportunities to improve it.

Patterns of Scientific Content

The analysis of explicit knowledge content has a long history. Content analysis, or assessment of the frequency and co-appearance of words, phrases, and concepts throughout a text, has been pursued since the late 1600s, ranging from efforts in 18th-century Sweden to quantify the heretical content of a Moravian hymnal (*6*) to mid–20th-century studies of mass media content in totalitarian regimes. Contemporary approaches focus on the computational identification of "topics" in a corpus of texts. These can be tracked over time, as in a recent study of the news cycle (*7*). "Culturomics" projects now follow topics over hundreds of years, using texts digitized in the Google Books project (*3*). Topics can also be used to identify similarities between documents, as in topic modeling, which represents documents statistically as unstructured collections of "topics" or phrases (*8*).

With the rise of the Internet and computing power, statistical methods have also become central to natural language processing (NLP), including information extraction, information retrieval, automatic summarization, and machine reading. Advances in NLP have made it one of the most rapidly growing fields of artificial intelligence. Now that the vast majority of scientific publications are produced electronically (*5*), they are natural objects for topic modeling (*9*) and NLP. Some recent work, for example, uses computational parsing to extract relational claims about genes and proteins, and then compares these claims across hundreds of thousands of papers to reconcile contradictory results (*10*) and identify likely "missing" elements from molecular pathways (*11*). In such fields as biomedicine, electronic publications are further enriched with structured metadata (e.g., keywords) organized into hierarchical ontologies to enhance search (*12*). Citations have long been used in "scientometric" investigations to explore dependencies among

Department of Sociology, University of Chicago, Chicago, IL 60637, USA.

*To whom correspondence should be addressed. E-mail: jevans@uchicago.edu