Ensuring the Data-Rich Future of the Social Sciences

Gary King

Science **331**, 719 (2011);
DOI: 10.1126/science.1197872

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

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.

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.

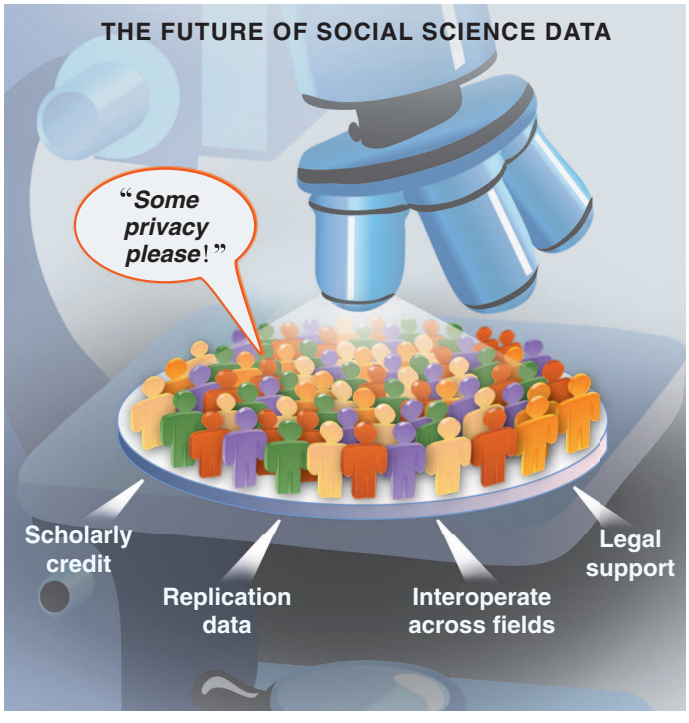

Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

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