

## **Providing the Web of Social Science Knowledge for the Future: A Network of Social Science Data Collaboratories**

Karen S. Cook, Gary King, and David Laitin

### Abstract

The evidence base of the social sciences is changing rapidly as we enter a historically unprecedented phase in the production and availability of data and other information about the social, political, and economic world. These new and abundant sources of information hold the promise of enabling social scientists to address the most significant issues facing us as a society -- from governance, health and welfare, to the environment and the economy-- if the information is harnessed appropriately. If we can gear up fast and build the research infrastructure necessary to manage effectively and make accessible the immense infusion of data, successfully provide training to a new generation of scholars who will work with these data, and tackle the substantial privacy and security issues, social science can make more dramatic progress than ever before imagined. No single university or research group is likely to be able to manage all of these tasks, so it is proposed that NSF create a major national resource -- a collaboratory of networked institutions to support a wide-range of activities that would make these tasks manageable, creating a shared resource of unparalleled value to the world of social, behavioral and economic science.

## A New Social Science Infrastructure: Networked Collaboratories

### The Premise, the Promise and the Problem:

This White Paper is based on a premise that for the first time in history, the social, behavioral and economic sciences have the possibility of providing the types of information necessary for the solution of some of the most complex problems facing us as a nation. These include problems in health and health care delivery, in environmental degradation, in the use and misuse of technologies, in revitalization of various types of industry, in educational institutions and their effectiveness, in countering insurgencies and terrorism, and in the revival of our institutions of democracy. Instead of picking one of these consequential issues for the focus of an SBE initiative at the NSF, we propose the development and the nurturing of an infrastructure that will enable creative social, behavioral and economic scientists to make inroads on problems with data heretofore unimaginable. In light of this historic opportunity to address social problems with immense infusion of high quality data, we will in this White Paper propose the funding of a “collaboratory” with several sites, each of which would manage and disseminate data to the broad scientific community.

Our premise is evident. The evidence base of the social sciences is changing rapidly as we enter a historically unprecedented phase of digital data collection and storage. The population of humans worldwide is leaving digital data traces in ways that can inform us about the most significant issues facing us as a society from governance, health and welfare, to the environment and the economy. How we make this information useful to the scientific community and those who make policy and determine our future represents one of the vital challenges of our time. Over the past fifty years or so social scientists have collected data to understand public opinion, primarily using survey instruments, sample surveys taken every few years, or using government statistics, often flawed. Anthropologists, sociologists, and others have used these methods, but have also studied specific places, events or groups using more intensive qualitative methods. These sources of information and standard methods have been relatively informative, but they are also limited in significant ways. For example, increases in cell phone use and the growing levels of non-response are crumbling the scientific foundations of random surveys of isolated individuals. And, while aggregated government statistics are valuable, in many countries they are of dubious validity for various reasons. In-depth case studies are informative, but do not scale, are not representative and cannot measure long-term change.

While the existing data collection mechanisms will surely continue to be used and improved, such as with the increased use of web surveys, the great promise of transforming the social, behavioral and economic sciences lies in creating the scientific infrastructure to capture and make accessible the terabytes of data just now becoming available to the research community.

Examples include:

Extensive and continuous time information can be collected now on individual political and social behavior. Potential sources include unstructured text (via automated information extraction from social media posts, emails, speeches, government reports, social networking sites, and other web sources.) Electoral activity may be coded based on ballot images, precinct-level results, individual-level registration, primary participation, and campaign contribution data. Commercial activity can be recorded based on credit card and real estate transactions, among other types of economic exchange. Geographic location of people and events can be detected using cell phone, GPS systems, and tracking through tollbooths via Fastlane or EZPass transponders.

Health information will become ever more detailed and available through the use of digital medical records, linked data systems, hospital admittance information records, and specialized devices (e.g. accelerometers) being included in cell phones. Many aspects of the biological sciences are now effectively becoming social sciences, as developments in genomics, proteomics, and brain imaging produce huge numbers of person-level variables being recorded for research and eventually treatment purposes. Satellite imagery is increasing in scope, resolution, and availability.

## A New Social Science Infrastructure: Networked Collaboratories

The Internet is spawning numerous ways for individuals to interact such as through social networking sites, social bookmarking, comments on blogs or tweets, participating in product and service delivery reviews, and entering virtual worlds, all of which create possibilities for observation and experimentation.

The analogue-to-digital transformation of numerous devices people own makes them work better, faster, and less expensively, but also enables each one to produce data in domains not previously accessible via systematic analysis. This includes everything from real-time changes in the web of contacts among people in society (the Bluetooth in your cell phone knows whether other people are nearby!) to records kept of individuals' web clicking, searches, and advertising click-throughs. Similarly, the paper-to-web-based transformation of government agencies' record keeping is making valuable data increasingly available to researchers. Some governmental policies are furthering these changes by requiring more data collection, such as the "No Child Left Behind Act" in education and via the proliferation of randomized policy experiments. All these changes are being supplemented by the replication movement in academia that encourages or requires social scientists to share data we have created with other researchers to allow for a more open science and more rigorous tests of our theories and policy prescriptions.

These data, appearing with increasing rapidity, put numerous advances within our reach for the first time in history. Instead of trying to extract information from a few thousand activists' opinions about politics every two years, in the necessarily artificial conversation initiated by a survey interview, or the coding of uneven newspaper reports by undergraduates, we can use new methods to mine the tens of millions of political opinions expressed daily in published social media posts. Instead of studying the effects of context and interactions among people by asking respondents to recall the frequencies and nature of their social contacts, we now have the ability to obtain a continuous record of all phone calls, emails, text messages, and in-person contacts among a much larger group. In place of dubious or nonexistent governmental statistics to study economic development or population spread in Africa, we can use satellite pictures of human-generated light at night or networks of roads and other infrastructure measured from space during the day. And, getting biomarker data on individuals is becoming standard in medical research with important implications for behavioral research as well as health policy.

The implications of these data for big problems hardly need justification. One example should suffice. While we know that an increase in generalized trust in society is associated with stable democracy, the provision of high quality public goods and economic growth, our tools to measure changes in trust over time in individuals are paltry; meanwhile massive data on the emergence of trust are available in the databases such as eBay and StubHub. Instead of surveys or experiments on trust, we can now observe microscopically its emergence in natural settings.

If we can tackle the substantial privacy and security issues, build more powerful and more widely applicable theories, help create informatics techniques to ensure that the data are accessible and preserved, and develop new statistical methods adapted to these new types of data, the social, behavioral and economic sciences can make more dramatic progress than ever before imagined.

What we are missing is the infrastructure to make these many forms of data widely available through proper human subjects protections, appropriate modes of storage and accessibility, access to high performance computing solutions, and new methodological research designed to take advantage of the promise these data offer. Providing this infrastructure would make possible massive new leaps of knowledge in the behavioral and social sciences. In other fields of science, and different eras, America understood the potential of large science projects such as the need for a "supercollider." Our challenge now is to conceive of creating a supercollider for the social, behavioral and economic sciences.

Methods of analysis are being developed that aid in the taming of massive amounts of data from millions of lines of text to millions of snippets of conversation, from pages and pages of congressional testimony to minute-by-minute records of chat, information sharing, networked data,

## A New Social Science Infrastructure: Networked Collaboratories

and many types of continuous time, geo-located data trails. Computer scientists have jumped into the fray aggressively competing to produce software and hardware that facilitate the taming of such information and even its visualization. Collaborations in various corners of the United States and abroad are springing up to support these proto-endeavors. But there is no clearinghouse for the collection, storage, protection, and facilitation of the analysis of such data, and thus our proposal for a collaboratory.

The key fact is that no single university or research group is likely to be able to manage all of these tasks, so we propose that NSF create a major national resource, a collaboratory of networked institutions to support a wide-range of activities that would make these tasks manageable, creating a shared resource of unparalleled value in the world of behavioral and social science. Most of the benefits offered by the collaboratory would also directly benefit the social science community at large through shared digital and computational resources.

### A Collaboratory:

In order to create the infrastructure for this new Web of Social, Behavioral and Economic Science Data we suggest the formation of a national advisory committee (NAC) to help connect each of the individual sites and important relevant initiatives at the foundation.

The first task of the NAC would be to administer RFPs to compete for one of several social, behavioral and economic science institutes (or collaboratories) that would provide both the leadership and the organizational framework for development of this SBE-Web. These proposals would outline the data that would be mined, and how those data would be managed and disseminated. Many of the features these collaboratories will provide will be online and completely networked, and so their services will be available to the all social scientists. The field is well past the time when resources, which are necessarily awarded to one institution, only benefit that institution. However, they do need a physical location, and experienced leadership and staff. Many institutions capable of performing a full collaboratory role exist now. Some of the institutions moving in this direction include IQSS at Harvard, IRiSS at Stanford, and ISR and ICPSR at Michigan. All the major social science archives now collaborate in a unified structure through DATA-Pass, including also the Odum Institute at UNC, the Roper Center at the University of Connecticut, and the U.S. National Archives and Records Administration. And, there are many other potential relevant institutional participants, who have collaborated in the past on different types of efforts (including the ANES) and have experience with training in cutting edge methods of analysis in the social sciences.

Each chosen location would require large-scale NSF funding for servers and data storage devices along with the relevant personnel to manage the facility and create protocols for national access. Obviously expertise in managing digital archives (as well as coping with the difficulties of storage given rapidly changing environments for both hardware and software) would be advantageous. We would clearly need to involve a range of technicians with the relevant expertise as we design each SBE-Web based collaboratory. Moreover, a staff able to make the data accessible to the social science community broadly, with careful attention to issues of human subjects review, confidentiality, and data security, would be essential.

Eventually implementation will require new personnel to administer proposals for new protocols and enhanced data management, proposals that would be administered through peer review processes and judged through NSF panels. Staff with the proper technical expertise and a few postdoctoral students in the social, behavioral, economic and computational sciences (statistics, bioinformatics and computer science) will be necessary to create a "test-bed" for research into mining such information by and for social scientists as well as for those in the relevant policy world with a need for access to such information in a condensed and digestible form. In sum, we envision the creation of a broad community of scholars working to solve the inevitable challenges in making this SBE-Web work to fulfill its promise.

## A New Social Science Infrastructure: Networked Collaboratories

### Footnotes:

\* This essay incorporates portions of Gary King "The Changing Evidence Base of Social Science Research," Chapter 38, Pp. 91–93 in Gary King, Kay Scholzman, and Norman Nie eds., *The Future of Political Science: 100 Perspectives*. New York: Routledge Press, 2009, copy at <http://gking.harvard.edu/files/abs/evbase-abs.shtml>.

† Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138;

^The IGERT Proposal to Establish a Computational Social Science Program at Stanford University (PIs: Dan McFarland and Dan Jurafsky)

\*\*David Lazer; Alex Pentland; Lada Adamic; Sinan Aral; Albert-László Barabási; Devon Brewer; Nicholas Christakis; Noshir Contractor; James Fowler; Myron Gutmann; Tony Jebara; Gary King; Michael Macy; Deb Roy; Marshall Van Alstyne "SOCIAL SCIENCE: Computational Social Science," *Science*, 323, 5915 (6 February 2009): 721--723, copy at <http://www.sciencemag.org/cgi/content/full/323/5915/721?maxtoshow=&hits=10&RESULTFORMAT=&fulltext=lazer&searchid=1&FIRSTINDEX=0&resourcetype=HWCIT>

### License:

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.