

# The Changing Evidence Base of Social Science Research\*

Gary King<sup>†</sup>

I believe the evidence base of political science and the related social sciences are beginning an underappreciated but historic change. As a result, our knowledge of and practical solutions for problems of government and politics will begin to grow at an enormous rate — if we are ready.

For the last half-century, we have learned about human populations primarily through sample surveys taken every few years, end-of-period government statistics, and in-depth studies of particular places, people, or events. These sources of information have served us well but, as is widely known, are limited: Survey research produces occasional snapshots of random selections of isolated individuals from unknown geographic locations, and the increases in cell phone use and growing levels of nonresponse are crumbling its scientific foundation. Aggregate government statistics are valuable, but in many countries are of dubious validity and are reported only with intentionally limited resolution or after obscuring valuable information. One-off in-depth studies are highly informative but for the most part do not scale, are not representative, and do not measure long-term change.

In the next half-century, these existing data collection mechanisms will surely continue to be used and improved — such as with inexpensive web surveys, if the problems with their representativeness can be addressed — but they will be supplemented by the profusion of massive data bases already becoming available in many areas. Some produce extensive or continuous time information on individual political behavior and its causes, such as based on text sources (via automated information extraction from blogs, emails, speeches, government reports, and other web sources), electoral activity (via ballot images, precinct-level results, and individual-level registration, primary participation, and campaign contribution data), commercial activity (through every credit card and real estate transaction and via product RFIDs), geographic location (by carrying cell phones or passing through toll booths with Fastlane or EZPass transponders), health information (through digital medical records, hospital admittances, and accelerometers and other devices being included in cell phones), and others. Parts of the biological sciences are now effectively becoming social sciences, as developments in genomics, proteomics, metabolomics, and brain imaging produce huge numbers of person-level variables. Satellite imagery is increasing in scope, resolution, and availability. The internet is spawning numerous ways for individuals to interact, such as through social networking sites, social bookmarking, comments on blogs, participating in product reviews, and entering virtual worlds, all of which are possibilities for observation and experimentation. (Ensuring privacy and protection of personal information during the analyses to be conducted with this information will require considerable effort, care, and new work in research ethics, but should not be markedly more difficult than the now routine medical research involving experiments on human subjects with drugs and surgical procedures of unknown safety and efficacy.)

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

---

\*To appear in Chapter 38, Pp. 91–93, Gary King, Kay Scholzman, and Norman Nie eds., *The Future of Political Science: 100 Perspectives*. New York: Routledge Press, 2009.

<sup>†</sup>Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138; <http://GKing.Harvard.Edu>, [King@Harvard.Edu](mailto:King@Harvard.Edu), (617) 495-2027.

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 clickthroughs. Partly as a result of new technology, governmental bureaucracies are improving their record keeping by moving from paper to electronic data bases, many of which are 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.

These data put numerous advances within our reach for the first time. 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, we can use new methods to mine the tens of millions of political opinions expressed daily in published blogs. Instead of studying the effects of context and interactions among people by asking respondents to recall their frequency and nature of 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. The number, extent, and variety of questions we can address are considerable and increasing fast.

If we can tackle the substantial privacy issues, build more powerful and more widely applicable theories with observable implications in these new forms of data, help create informatics techniques to ensure that the data are accessible and preserved, and develop new statistical methods adapted to the new types of data, political science can make more dramatic progress than ever before. The challenge before us as a profession, before each of us as researchers, and before the broader community of social scientists, is to prepare for the collection and analysis of these new data sources, to unlock the secrets they hold, and to use this new information to better understand and ameliorate the major problems that affect society and the well-being of human populations.