Published in final edited form as:

Ann Am Acad Pol Soc Sci. 2015 May 1; 659(1): 16–32. doi:10.1177/0002716215570007.

From Big Data to Knowledge in the Social Sciences

Bradford W. Hesse, **Richard P. Moser**, and **William T. Riley**

Abstract

One of the challenges associated with high-volume, diverse datasets is whether synthesis of open data streams can translate into actionable knowledge. Recognizing that challenge and other issues related to these types of data, the National Institutes of Health developed the *Big Data to Knowledge* or *BD2K* initiative. The concept of translating "big data to knowledge" is important to the social and behavioral sciences in several respects. First, a general shift to data-intensive science will exert an influence on all scientific disciplines, but particularly on the behavioral and social sciences given the wealth of behavior and related constructs captured by big data sources. Second, science is itself a social enterprise; by applying principles from the social sciences to the conduct of research, it should be possible to ameliorate some of the systemic problems that plague the scientific enterprise in the age of big data. We explore the feasibility of recalibrating the basic mechanisms of the scientific enterprise so that they are more transparent and cumulative; more integrative and cohesive; and more rapid, relevant, and responsive.

Keywords

big data; data visualization; integrative data analysis; informatics

In 2003, the U.S.-based National Science Foundation (NSF) received a report from a Blue Ribbon Advisory Panel on the importance of strengthening collaborative, electronic networks for science (Atkins et al. 2003). The advisory report was followed by an action plan from the NSF in 2007 announcing a "Cyberinfrastructure Vision for 21st Century Discovery" (National Science Foundation 2007). Both reports captured a vision for a collaborative, data-intensive research environment that—according to the NSF director at the time, Arden Bement—would transform every aspect of science from theory building within high-energy physics; to mapping new frontiers in molecular medicine; to supporting new interdisciplinary views of human endeavor within the social, behavioral, and economic sciences. The reports heralded the need to bring big data into the full pantheon of scientific endeavor, including medicine and the social sciences.

Big data in the life sciences and medicine

Two years after the release of the cyberinfrastructure report, the newly appointed director of the National Institutes of Health (NIH), Francis Collins, stressed that the development of high-throughput computing technologies would be an integral part of the NIH mission. Admittedly, Dr. Collins's emphasis on big data technologies in the life sciences stemmed

from his observation that the gene sequencing technologies responsible for documenting the 3 billion+ base pairs of the human genome had increased exponentially in computing speed while decreasing in cost (F. S. Collins 2010). The NIH was becoming aware that in the not-so-distant future the availability of high-throughput computing technologies would create a clinical environment conducive to an entirely new type of biomedical enterprise. Thought leaders have referred to this as “4P Medicine,” or medicine that would use the power of data to become more predictive, preemptive, personalized (or some would say more precise), and participative than its industrial-age counterpart (Hood and Flores 2012; Shaikh et al. 2014). Data flowing back into the research enterprise from the natural laboratories of clinical practice could also inform a new era of basic discovery; one in which biomedical science and health systems research could be informed by real world application in a virtuous cycle of continuous improvement within a “rapid learning healthcare system” (Etheredge 2007).

At the same time, new technologies have been emerging in the consumer sphere that have been catalyzing a need to accommodate the influx of voluminous, and high-velocity, data streams in health and medicine. For example, the market from real-time “wearable devices” such as the Apple Watch® and the Fitbit® Daily Activity Tracker is expected to exceed $6 billion by 2016 in the United States (Mearian 2012). The Food and Drug Administration has been actively engaged in discussions related to its role in monitoring the efficacy and reliability of these devices in medicine, while keeping a close eye on the privacy, security, and confidentiality issues associated with the collection of these data for research and marketing. Similarly, the number of office-based physicians reporting adoption of electronic health records (EHRs) rose precipitously from 17 percent in 2002 to 78 percent in 2013 due in large part to market incentives from the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009. The sociotechnical challenges of creating new clinical workflows that are enhanced by—and not impeded by—this data capacity will remain at the heart of research and engineering activities in the years to come (Robert Wood Johnson Foundation 2014).

With all of these innovations in data collection, what remains as a bottleneck is the analytic capacity to turn raw data into actionable information; or put more succinctly, to turn big data into knowledge. It was for this reason that the NIH launched the BD2K initiative (Ohno-Machado 2014) in 2013. BD2K includes efforts not only to develop new methods and analytics but also to train biomedical and behavioral researchers in existing big data approaches, which are not represented in traditional research training programs.

Relevance of BD2K to the social sciences

The opportunities associated with moving big data to knowledge are relevant to the social sciences in two ways. First, the computational techniques needed to integrate and analyze voluminous amounts of data will be just as important in building a foundation for new knowledge in the social sciences as they are becoming in the physical sciences. The big data movement will require all disciplines to revisit their basic methodological and epistemological foundations in an era of data-intensive, networked science, and the social sciences will be no exception. In fact, the social and behavioral sciences have been criticized by the popular press recently for being unable to replicate some of the foundational studies

in their constituent disciplines (Winerman 2013). Other critics have bemoaned the slow pace of translation from laboratory to practice in the social sciences, and have noted that changes in the fabric of the underlying disciplines should be enacted to accelerate progress on important social problems (Baker, McFall, and Shoham 2009). Social science communities are joining other scientific disciplines in revisiting the roles of theory, publication, data sharing, intellectual property rights, knowledge accumulation, and public accountability as they reinvent themselves in the current era.

Second, the translation of data into knowledge is itself a social enterprise. Social science research is needed to understand how big data assets can be created, accessed, shared, and utilized for advances in knowledge across disciplines. The President's Council of Advisors on Science and Technology explained it this way in their 2010 report on "Designing a Digital Future" in the United States and abroad: "One of the most striking features of the revolution enabled by the Internet and the World Wide Web over the last two decades is the extent to which it has been fueled by the contributions of millions of users, the vast majority of whom have little or no technological or programming prowess. … This is just the beginning of the new field of collective intelligence, in which modern technology yields new understanding of collective human behavior and new methods for problem-solving in complex systems and networks."

In this article, we explore the implications of moving from big data to knowledge in the social sciences. We begin with the challenges and opportunities associated with refining the basic structures underlying the creation of a truly cumulative science through revised policy and new, participative platforms. We then examine the necessary structural and analytic changes needed to extract value from converging and parallel data streams. In this context, we introduce the concept of integrative data analysis, and we illustrate how new techniques have begun to improve awareness of the social and behavioral processes that play out across multiple levels of analysis within health-oriented communities. We complete our treatment of BD2K in the social sciences by examining ways in which these new data structures can be engineered to facilitate a faster, and more efficient and relevant course of science in the future.

Refining the Structures of Cumulative Science

In 1993, the first author of this article reported data from an online survey conducted within a community of physical oceanographers in what was arguably one of the first evaluations of the sociological and psychological effects associated with moving collaborative science into a networked, computer-mediated communication environment (Hesse et al. 1993). Findings from the study were enlightening. In general, electronic connectivity was positively associated with increased scientific productivity across respondents as measured by papers published, presentations made, honors received, and scientists known (equivalent to a main effect). Interestingly, associations were differentially strong for oceanographers that were geographically isolated from the resources amassed at the more prestigious coastal institutions (an interaction effect). Taken together, this early research highlighted the potential of electronic connectivity to support distributed work in ways that would be productive within the "invisible college" (Crane 1972) of science. Research in other contexts

revealed similar effects (Sproull and Kiesler 1991). Recognizing the potential of electronic networks to accelerate knowledge discovery, the NSF increased its investments in electronic connectivity to improve data collection, analysis, and reporting in science.

A movement toward open access

With investments under way to support science through electronic networks, there would soon be a sharpening focus among policy-makers to ensure that the benefits of publicly funded research would be distributed transparently and equally to all possible beneficiaries. In what is commonly known as the “Bethesda statement on open access publishing” issued in June 2003, the NIH led the pack by announcing its intention to require that any scholarly publications resulting from publicly funded biomedical research be made available freely to the public within 12 months of formal publication. Other international bodies offered similar resolutions to take advantage of the benefits associated with the global information infrastructure. Several bills were introduced through the U.S. legislature to follow suit, suggesting that publicly funded research be made freely available to the public to spur innovation. Although the majority of those bills never came up for a vote, the intent of the legislation was codified during the budget negotiations of 2013 and 2014 in the United States.

The Consolidated Appropriations Act of 2014—the negotiated funding bill to keep the U.S. government open after the political impasse of 2013—included specific provisions for public access to government-funded research findings. Affected agencies included the Departments of Health and Human Services, Education, and Labor with annual expenditures over $100 million. The legislation directed the agencies to make published research articles funded by taxpayer dollars available freely to the public in electronic format no later than 12 months following publication. The expanded coverage of the bill, it has been estimated, would now make $31 billion of the total $61 billion annual research budget for the U.S. government open to the public. As might be imagined, the legislation has led to ongoing debates among the scientific publishing houses over how to comply with the requirements of the legislation while meeting publication costs. Social science publishers typically rely on subscription fees, whereas medical publishers often use author fees to cover costs. Debates over new business models are under way.

Moving toward open access to data

In a similar vein, the NIH and the White House have taken interest in assuring that the building blocks of research, specifically the data upon which publications are based, also be made available to the broader scientific enterprise. In 2003, the NIH initiated a data sharing policy for all grants receiving more than $500,000 per year in direct costs, and the NIH recently published a genomic data sharing (GDS) policy that requires all NIH-supported genomic data, regardless of funding level, to be made available in an appropriate NIH data repository for secondary analyses.[1] This policy addresses many of the challenges of open data efforts including informed consent for future secondary data use, appropriate de-identification and additional privacy protections of sensitive information, protection of

[1]See http://gds.nih.gov.

intellectual property, and the considerations for controlled versus unrestricted access to these data repositories. The NIH continues to support the database of Genotypes and Phenotypes (dbGaP)[2] along with a number of other data repositories to facilitate data access and integration, but behavioral and social sciences data are not included among the data repositories that the NIH supports.

At the executive level, the first presidentially appointed Federal Chief Information Officer (a position made possible by passage of the E-Government Act of 2002) announced plans to make federally funded datasets available to the public in machine-readable format through the establishment of the "data.gov" website and the more recent "healthdata.gov" site. On February 22, 2013, the Office of Science and Technology Policy at the White House reinforced that policy by issuing a memorandum to the heads of executive departments and agencies instructing them to make access to digital datasets resulting from federally funded research a priority. The memo highlighted the provision of weather data to power the forecasting industry and the provision of genomic sequencing data to power biomedical innovation as compelling use cases.

Improving rigor and reproducibility

Another justification for moving into a paradigm of open science is to improve the rigor and reproducibility of research as a fully transparent, collaborative endeavor. Proponents of open science have often pointed out that an historical reliance on paper-based, limited publication venues has created a set of unanticipated obstacles that stand in the way of cumulative knowledge building (Nosek and Bar-Anan 2012). Page limits in print journals, for example, may often lead to an exclusion of nonsignificant findings based on the somewhat arbitrary heuristic of a $p < .05$ threshold for statistical significance testing (Cumming 2014). When tenure, funding, and professional recognition are all predicated on publication rates, there is a not-so-subtle pressure on investigators to strain the assumptions of a priori hypothesis testing to explore ways of reaching a .05 level of significance for at least some findings in their dataset (Ioannidis et al. 2014), a custom referred to by some as "p hacking" (Simonsohn, Nelson, and Simmons 2014). When evaluated for expected frequencies of positive and negative findings, and for evidence of successful replication of core findings, much of the literature in the life sciences did not appear to measure up to a priori expectations (Ioannidis, Nosek, and Iorns 2012).

In response to these concerns, professional societies and funding agencies have initiated efforts to identify the systems-level constraints on cumulative science, and to experiment with potential remedies. In 2012, the Association for Psychological Science in conjunction with the NIH Office of Behavioral and Social Science Research published a special issue of the journal *Perspectives on Psychological Science* on the topic of reproducibility of research findings in the psychological sciences. At around the same, the American Psychological Association launched an experimental open-access journal called the *Archives of Scientific Psychology* as a foray into data archiving and open-access publishing. In February 2014, the Social, Behavioral, and Economics Science Directorate at the NSF convened a panel of

[2]See www.ncbi.nlm.nih.gov/gap.

invited experts to discuss obstacles to rigor and reliability in the social sciences and, in an era of big data, propose promising solutions for further exploration.

Integrating Data Streams

Another thrust of the big data initiatives as facilitated through cyberinfrastructure support for science is the ability to turn isolated data streams into an integrative picture of converging patterns to facilitate situational awareness among social scientists, policy-makers, practicing professionals, and the general public (Thacker, Qualters, and Lee 2012). One example of this capacity can be described using the implications of the work being done by physical oceanographers as described earlier in this article. One of the reasons physical oceanographers were early adopters of distributed network technologies is that they were reliant on these technologies to integrate signals from remote buoys, satellite telemetry and sensing, oceangoing vessels, airborne weather balloons, and other sources of high-volume, high-velocity data inputs covering large geographic areas. Government agencies such as the NSF, the National Aeronautics and Space Administration, and the National Oceanic and Atmospheric Administration have all contributed joint funding to ensure that inputs from these sensors conform to high standards of fidelity and reliability.

What the funding agencies realized is that they could return value to the public by allowing third-party vendors to build applications based on these data. Commercial meteorologists translate daily readings of these inputs into daily weather and ocean condition reports for reporting through news outlets and more recently through mobile device weather apps. Geographic position system (GPS) device developers have created an entirely new sector of the economy built on nautical, aeronautic, and automobile navigational systems. Information technology powerhouses such as Google, Apple, Android, and others have been able to augment these systems with complementary data streams from traffic sensors, open geographic information system architecture, and commercial vendors to enable consumer-facing map software and have even begun experimenting with autonomous driving technologies.

Constructing a prototype for situational awareness

Using these integrative data activities from the physical sciences as a backdrop, we sought in 2007 to illustrate how complementary data streams related to knowledge, attitudes, and behaviors in health might add value to communities striving to meet Healthy People 2010—and subsequently Healthy People 2020—goals.[3] Through the National Cancer Institute's Behavioral Research Program, we commissioned a set of developmental studies designed to explore the feasibility of using integrative data analytic techniques to present users with an interactive map of trends from converging data sources over time. Figure 1 illustrates the technical blueprint that we followed in constructing a Population Science Grid (i.e., the PopSci Grid) prototype application. The bottom layer of the schematic depicts how common data elements, along with other semantically interoperable ontologies, could be used to link publicly available surveillance data to an interoperable Grid architecture using the Open

[3]The Healthy People initiative is a public engagement program designed to provide science-based, 10-year objectives for improving health in America (see www.healthypeople.gov).

Access Globus toolkit and extensible markup language (XML/XMi) for metadata. Sourced datasets could, in theory, be drawn from a large number of publicly available surveillance datasets maintained across the federal government. The idea was that by making these datasets discoverable on the Grid, it would be possible to construct an interactive public layer for bringing relevant components together to inform public policy planners, journalists, researchers, and the public.

Figure 2 illustrates how one such interface was constructed, using elements derived from the Gapminder animated statistics conceptualized by Hans Rosling (Rosling and Zhang 2011). The prototype offered dynamic access to juxtapositions of data on state-based cigarette tax policy, self-reported smoking rates, and knowledge and attitudes related to the deleterious effects of smoking, while a slider allowed for explorations of trends across a temporal dimension. The “Science of Network Computing in Communities” (SONIC) group at Northwestern University conducted the initial design and program, with experimental enhancements offered by Deborah McGuinness and her team at Rensselaer Polytechnic Institute (McGuinness et al. 2011). Hua Min and her colleagues at the Fox Chase Cancer Center offered additional development from a medical and public health informatics perspective (Min et al. 2013). These collaborations illustrate how multidisciplinary teams can be brought together to create new architectures for public health participation, with the intention of moving big data assets into a framework for collective intelligence (Hesse et al. 2011).

This team effort, however, did have its own challenges, mostly related to the data themselves. Since data from several independent sources were merged together, it was critical to identify measures that were common to all sources; and in some cases, the data had to be reformatted to be put on a common scale (e.g., creating common education levels across variables) or were left out of the analysis if this was not possible. Inherent limitations in the data also precluded the types of information that could be shown (and types of analyses that could be done). For example, though cigarette tax information was available at the state level, smoking behavior and other demographic information was only available at the censusregion level and so this became the smallest geographic level that could be shown on the map; the same applied to the years shown as not all data were collected annually. Another challenge, not related to the data per se, were the types of statistical analyses that were supported through the site. We wanted to avoid overinterpretation of the results, especially given the cross-sectional nature of the data, so we limited the analysis to descriptive statistics and bivariate correlations as we did not want users to interpret causal effects. There were few challenges, however, in working within our multidisciplinary team. Given that the team included a wide range of disciplines, that the members had a history of working together, and that members were not competing for resources and had a common goal, a sense of trust had been established to guide project team members toward collective goals. Teams without that background would likely have struggled.

Refining methods for integrating data

As new data streams become available through investments in NIH BD2K activities, there will be a concomitant need to expand our repertoire of data management and analytic

methods to deal with the variety of new types of data. One area of analytic development that is garnering attention is in the sphere of integrative data analysis (IDA). IDA refers to a set of strategies in which two or more independent datasets are pooled or combined into one and are then statistically analyzed (Curran and Hussong 2009). Relevant data for IDA can be both quantitative and qualitative (Castro et al. 2010). IDA approaches differ from and offer advantages over other methodological techniques that also strive to build cumulative knowledge bases, such as meta-analysis (Cooper, Hedges, and Valentine 2009). In meta-analysis, summary statistics across multiple studies are pooled together (Cooper, Hedges, and Valentine 2009; Glass 1976). Because IDA techniques pool original raw data, there is no loss of individual information as found within meta-analytic approaches, allowing researchers to answer not only what works, but for whom, and in what context (Cooper and Patall 2009). Combining data across studies can provide sufficient power to detect moderator and mediator effects that individual clinical studies are seldom sufficiently powered to detect, or to facilitate unique cross-study comparisons. In addition, use of IDA affords expanded inquiry within many areas of social science research. For example, IDA can be used to augment surveillance datasets focused on biomedical tracking with variables related to social context, understanding, beliefs, or behaviors.

Though a powerful analytic technique, IDA does have its own limitations related to the data being integrated. To successfully combine common data elements, at a minimum the constructs being assessed must be the same. Once this is established, challenges arise, especially when working with self-report survey data, when the constructs are being assessed with measures (or items) that may have different wording, response options, or groups being assessed (usually due to different skip patterns across survey iterations). Differences in regard to time (when the data were collected), sampling methods, geography, and other sources of heterogeneity can also be problematic when integrating data though they can also be used for cross-study comparisons and thus be a strength (Curran and Hussong 2009). In many ways, measurement issues present the biggest challenge though there are psychometric and statistical methods—both traditional and recently developed—that allow for direct integration and comparison of measures that are assessed differently (Bauer and Hussong 2009; Choi et al. 2014).

The second author on this article has taken the lead in illustrating how IDA techniques can be used to extend analytic value for the NCI's Health Information National Trends Survey (HINTS), addressing some of the challenges with integrating survey data. The HINTS program is a nationally representative, cross-sectional survey that was fielded for the first time in fall 2002–2003. Since then there have been seven successive iterations of the survey over an 11-year period. Data from the surveys have been mounted in downloadable format in the spirit of open access, as have the instruments, methodology reports, and research and technical documentation. Since 2003, a robust user community has been active in analyzing the open data and in integrating items as common data elements in their own studies, enabling comparisons of local results with national data. Over the past decade there have been adaptations of HINTS items administered to Korean speakers, Spanish speakers (within the continental United States as well as in the territory of Puerto Rico), and Mandarin speakers in the People's Republic of China. Items have been used in several state surveys including a concentration among Appalachian states, and in a special adaptation of

respondent-driven sampling in the territory of Guam. The modes in which questions have been administered have also varied, with a telephone-based random digit dial frame utilized in 2003 and 2005; a mixed telephone/postal frame in 2007–2008; postal only frame used from 2012 to 2015; and a proposal to translate items into American Sign Language (Finney Rutten et al. 2012, 2010; Moser et al. 2011; Nelson et al. 2004; Tortolero-Luna et al. 2010).

The challenge, then, has been to create an analytic bridge between the successive waves of HINTS data collection over time, especially when items are not assessed consistently or are not assessed at all; across changes in modality that can create qualitative differences in the results; across national and special emphases populations; and across levels of analysis for users moving from national to regional and state decision-making. To meet these challenges, the HINTS analytic team compiled a publicly available technical report that could serve as a pragmatic guide for researchers located in schools of communication, psychology, public health, medicine, and political science (Moser et al. 2013). The report demonstrated through step-by-step examples (and related statistical software syntax) how to use IDA principles to investigate and control for biases, and methods to allow for stable estimates at more “local” levels (i.e., smaller geographic units) and trending across iterations while expanding the breadth of research questions.

While noting the successful application of these methods in the report, it is also important to understand that the authors of two of the studies had access to restricted data that provided state-level estimates. These geographic data are not available to the public, though they can be accessed if users sign an agreement to use the data in an ethical manner and report any confidentiality breaches.

Enabling Rapid, Responsive, and Relevant Research

The consumer-facing, and often provocative, gene sequencing company 23andMe caught the attention of biomedical scientists when it demonstrated how it was possible to replicate the findings of a large NIH-funded trial (Neumann et al. 2009) in less than one-sixth of the time and a fraction of the cost for the original study. The NIH-funded trial followed a very customary, six-year trajectory to move from hypothesis generation, to proposal development, funding, data gathering, data submission, analysis, writing, and acceptance for publication. Its methods yielded an evidentiary conclusion suggesting that genetic mutations in the GBA gene were five times more likely to develop Parkinson’s than those without the anomalous gene. The 23andMe trial, on the other hand, took only 12 months to conceptualize, execute, and conclude. It did this by leveraging the willingness of its customers to donate data in an exercise of citizen science, while leveraging the capacity of a massively distributed electronic network to upload data from thousands of customers simultaneously. The end result was a significant reduction in the time it took to collect, analyze, and confirm the GBA-Parkinson link (U.S. Institute of Medicine 2012).

Consumer- or patient-initiated registries and repositories such as 23andMe and Quantified Self[4] provide unique infrastructures for rapid study, but these data sources also suffer from

[4]See http://quantifiedself.com.

being highly self-selective, nonprobability samples, which limits the generalizability of the findings. Participants in these repositories, however, intend for their data to be used for research. In contrast, consumers who leave digital traces on search engines, mobile applications, and social media sites do not intend for their data to be used for research. These data are extensively mined for quality improvement and marketing purposes, and these data are often considered archival and exempt from informed consent. Moreover, these digital services frequently engage in controlled trials, termed "A–B testing" to select optimal features, functions, and interfaces for their products, but as evident from the recent criticisms of a study that manipulated mood content on Facebook (Kramer, Guillory, and Hancock 2014) experimental manipulations of consumer behavior for research purposes require greater research participant protections than those required for the experimental manipulation of an interface feature or function.

Reinventing discovery

What these new capacities may deliver is an opportunity to "reinvent discovery" in an era of networked science, according to some (Nielsen 2012). Along those lines, the third author on this article has been engaged in efforts as a contributor to the NIH BD2K steering committee to accelerate the pace and relevance of social science research processes given the enhanced capacities of the information revolution. His goal, as articulated in a 2013 paper, has been to engineer a new research environment that is more rapid in catalyzing discovery; is more responsive to real community needs; and is more relevant to the task of translating scientific knowledge into replicable behavioral interventions (Riley et al. 2013). This articulation is timely given that the Institute of Medicine is moving toward using interoperable data flows made possible through electronic health record systems to move efficacious treatment recommendations more expeditiously from "bench to bedside" and then "back to bench" for further refinement in a learning healthcare system (Abernethy et al. 2010; Etheredge 2007).

Part of the focus on creating a more rapid environment for moving data into knowledge lies in revisiting the basic assumptions that underlie much of what we do in social and biomedical research. The new mobile sensing technologies that are becoming ubiquitous as part of the "wearable device" revolution can provide the capability to collect rapidly recorded behavioral data, often unobtrusively, within an "n-of-1" paradigm (Ginexi et al. 2014; Riley et al. 2011). Scientists who are building theoretical foundations based on between-subject designs arrayed over a sparingly collected set of sparse data points will likely not be up to the task of accelerating their discoveries beyond the slow, cumbersome pace of time-consuming trials (Riley et al. 2011). Research must become more rapid if it is to be responsive and relevant to those making treatment and policy decisions now, not 7 to 14 years from now; and more rapid research reduces the risk of producing findings on techniques and procedures that could be dated or obsolete by the time the findings are made available.

Fortunately, research can be made more rapid and responsive without compromising rigor due to the emerging pallet of new designs and analytic methods—many borrowed from complementary disciplines—that can be applied to the challenge of driving knowledge more expeditiously from accumulating data sources. For example, the fractional factorial

methodologies used by engineers to make decisions quickly about the critical features of an engineered system, after which the design is modified and retested through rapid iteration, has been modified by methodologists to create sequential multiple assignment research trials (SMART) models to test interventions in behavioral medicine (L. M. Collins, Murphy, and Strecher 2007). Likewise, the assumptions of statistical process control underlying many "six sigma" quality improvement efforts in manufacturing, particularly computational system dynamics, can be appropriated to improve the contextual fit between an intervention and behavior at both the individual (e.g., Timms et al. 2013) and systems levels (e.g., Gaglio, Shoup, and Glasgow 2013).

These rapid research methods and approaches are not intended to replace the traditional randomized controlled trial (RCT), large nationally represented epidemiologic studies, or other "slow" research methods. However, slow is not synonymous with rigorous. RCTs may be an optimal method for testing the efficacy of a new intervention, but questions such as the effectiveness of the intervention among real patients in real settings, the safety and side effects of the intervention, and the determination of for whom the intervention may be most effective are questions that are better addressed by leveraging health system EHRs and other large data sources. New data technologies are also emerging from other spheres. From the biomedical sciences, genomic researchers are seeking to explore the potential interactions between underlying genetic mutations and influences from the physical and social environments, a sphere of influence referred to by some as the "exposome" (Wild 2005), to predict risk for disease and to modulate treatment. One of the emerging study designs in this area is the "genome-wide association study," a technique that compares DNA extracted from individuals with a particular disease against DNA from a comparison group without the disease. Literally millions of genetic variants are read using Single Nucleotide Polymorphism (SNP) microarrays and then associations are explored over individuals between polymorphisms and presence of the disease. What results is a graphical scatter plot, referred to as a Manhattan plot (because the resulting spikes look like skyscrapers), in which the strength of statistical associations are listed on the Y axis and a list of identified alleles (genomic coordinates) are listed on the X axis. The purpose of the plot is to help researchers look for alleles that stand out as exhibiting strong associations with the disease when computed over many subjects. These techniques could be explored in other types of association studies, including the influence of exposome characteristics on disease pathogenesis (Topol 2012).

Participating broadly in science

The 23andMe example given at the beginning of this section highlights another trend to explore in social science research, and that is the trend referred to by some as "citizen science." Researchers funded by the NSF hypothesized that under the right circumstances ordinary citizens could find themselves motivated and capable of contributing data to researchers as partners in the scientific enterprise. In one such study, characterized as an experiment in "participatory sensing," individuals suffering from asthma volunteered to use specially designed apps on their mobile phones to monitor air quality in Los Angeles over the course of their days. The citizen-donated data could then be compiled into data-rich atmospheric maps, indicating where air quality was bad and where the air quality was

relatively cleaner. The approach had the advantage of being rapid (these were real-time sensors), responsive, relevant to participants' needs, and surprisingly robust (Chen et al. 2012).

Other manifestations of citizen science have been arising with varying degrees of success. In the biomedical arena, the crowdsourcing platform "Foldit" uses gamification techniques (i.e., interface ideas borrowed from the video game industry to engage attention and promote interaction) in prompting a general audience to "solve puzzles for science." The puzzle in question has to do with folding proteins, a task that is extremely difficult to deduce through automated routines but can be tackled by the lay public interested in helping biochemists solve real world problems (Parslow 2013). Another manifestation is the "Patients Like Me" web presence, which gives patients diagnosed with any number of diseases a chance to interact with others suffering from the same ailment and then, together, to offer up data about their conditions and treatment in an act of "data altruism" for the good of others. Finally, the Food and Drug Administration's "Mini-Sentinel" pilot program also represents an effort to use electronically distributed data collection techniques to aggregate postmarket surveillance data on the safety of marketed medical products including pharmaceuticals, devices, and biologics (Platt et al. 2012).

Conclusion

In this article we examined the implications of a rapidly evolving, electronically distributed work environment for scientists collaborating across a host of issues from physical oceanography and biomedicine on one hand, to the social and behavioral sciences on the other. In doing so, we have attempted to explore the feasibility of recalibrating the basic mechanisms of the scientific enterprise to be more transparent and cumulative; to be more integrative and cohesive; and to be more rapid, relevant, and responsive than they ever have been before. We recognize, as did the President's Council of Advisors on Science and Technology (2010), that these opportunities are being enabled by the hard work of information scientists endeavoring to realize the benefits of a "digital future" across all sectors of the economy. We also recognize, as did the President's Council, that this is a sociotechnical endeavor; that at its core, it is about a new era of "social computing." The social sciences will not only benefit from the endeavor, they must also be part of it.

Biographies

Bradford (Brad) W. Hesse is a psychologist by training and chief of the Health Communication and Informatics Research Branch within the Behavioral Research Program at the National Cancer Institute. His work emphasizes the use of computer-mediated collaborative environments to improve outcomes in medicine and science.

Richard P. Moser is acting chief of the Science of Research and Technology Branch within the Behavioral Research Program at the National Cancer Institute. He has interests in integrative data analysis and measure standardization efforts.

William (Bill) T. Riley is acting director of the National Institute of Health Office of Behavioral and Social Sciences Research and is the chief of the Science of Research and Technology Branch, Behavioral Research Program in the Division of Cancer Control and Population Sciences at the National Cancer Institute.

References

Abernethy, Amy P.; Etheredge, Lynn M.; Ganz, Patricia A.; Wallace, Paul; German, Robert R.; Neti, Chalapathy; Bach, Peter B.; Murphy, Sharon B. Rapid-learning system for cancer care. Journal of Clinical Oncology. 2010; 28(27):4268–4274. [PubMed: 20585094]

Atkins, Daniel E.; Droegemeier, Kelvin K.; Feldman, Stuart I.; Garcia-Molina, Hector; Klein, Michael; Messerschmitt, David G.; Messina, Paul; Ostriker, Jeremiah P.; Wright, Margaret H. Revolutionizing science and engineering through engineering. Arlington, VA: National Science Foundation; 2003. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.

Baker, Timothy B.; McFall, Richard M.; Shoham, Varda. Current status and future prospects of clinical psychology: Toward a scientifically principled approach to mental and behavioral health care. Psychological Science in the Public Interest. 2009; 9(2):67–103. [PubMed: 20865146]

Bauer, Daniel J.; Hussong, Andrea M. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychological Methods. 2009; 14(2):101–125. [PubMed: 19485624]

Castro, Felipe G.; Kellison, Joshua G.; Boyd, Stephen J.; Kopak, Albert. A methodology for conducting integrative mixed methods research and data analyses. Journal of Mixed Methods Research. 2010; 4(4):342–360. [PubMed: 22167325]

Chen, Connie; Haddad, David; Selsky, Joshua; Hoffman, Julia E.; Kravitz, Richard L.; Estrin, Deborah E.; Sim, Ida. Making sense of mobile health data: An open architecture to improve individual-and population-level health. Journal of Medical Internet Research. 2012; 14(4):e112. [PubMed: 22875563]

Choi, Seung W.; Schalet, Benjamin; Cook, Karon F.; Cella, David. Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychological Assessment. 2014; 26(2):513–527. [PubMed: 24548149]

Collins, Francis S. Research agenda. Opportunities for research and NIH. Science. 2010; 327(5961): 36–37. [PubMed: 20044560]

Collins, Linda M.; Murphy, Susan A.; Strecher, Victor. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent eHealth interventions. American Journal of Preventive Medicine. 2007; 32(5 Suppl.):S112–S118. [PubMed: 17466815]

Cooper, Harris M.; Hedges, Larry V.; Valentine, Jeff C. The handbook of research synthesis and meta-analysis. 2nd ed.. New York, NY: Russell Sage Foundation; 2009.

Cooper, Harris M.; Patall, EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychological Methods. 2009; 14(2):165–176. [PubMed: 19485627]

Crane, Diana. Invisible colleges: Diffusion of knowledge in scientific communities. Chicago, IL: University of Chicago Press; 1972.

Cumming, Geoff. There's life beyond .05: Embracing the new statistics. APS Observer. 2014; 27(3): 19–21.

Curran, Patrick J.; Hussong, Andrea M. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14(2):81–100. [PubMed: 19485623]

Etheredge, Lynn M. A rapid-learning health system. Health Affairs (Millwood). 2007; 26(2):w107–w118.

Finney Rutten, Lila J.; Davis, Terisa; Beckjord, Ellen B.; Blake, Kelly D.; Moser, Richard P.; Hesse, Bradford W. Picking up the pace: Changes in method and frame for the Health Information

National Trends Survey (2011–2014). Journal of Health Communication. 2012; 17(7):979–989. [PubMed: 23020763]

Finney Rutten, Lila J.; Hesse, Bradford W.; Moser, Richard P.; Kreps, Gary L. Building the evidence base in cancer communication. Cresskill, NJ: Hampton Press; 2010.

Gaglio, Bridget; Shoup, Jo Ann; Glasgow, Russell E. The RE-AIM framework: A systematic review of use over time. American Journal of Public Health. 2013; 103(6):e38–e46. [PubMed: 23597377]

Ginexi, Elizabeth. M; Riley, William; Atienza, Audie A.; Mabry, Patricia L. The promise of intensive longitudinal data capture for behavioral health research. Nicotine Tobacco Research. 2014; 16(Suppl. 2):S73–S75. [PubMed: 24711629]

Glass, Gene V. Primary, secondary, and meta-analysis. Educational Researcher. 1976; 5:3–8.

Hesse, Bradford W.; O'Connell, Mary; Augustson, Erik M.; Chou, Wen-Ying; Shaikh, Abdul R.; Finney Rutten, Lila J. Realizing the promise of web 2.0: Engaging community intelligence. Journal of Health Communication. 2011; 16(Suppl. 1):10–31. [PubMed: 21843093]

Hesse, Bradford W.; Sproull, Lee; Kiesler, Sara B.; Walsh, John P. Returns to science: Computer networks in oceanography. Communication of the ACM. 1993; 36(8):90–101.

Hood, Leroy; Flores, Mauricio. A personal view on systems medicine and the emergence of proactive P4 medicine: Predictive, preventive, personalized and participatory. New Biotechnology. 2012; 29(6):613–624. [PubMed: 22450380]

Ioannidis, John PA.; Munafo, Marcus R.; Fusar-Poli, Paulo; Nosek, Brian A.; David, Sean P. Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. Trends in Cognitive Science. 2014; 18(5):235–241.

Ioannidis, John PA.; Nosek, Brian A.; Iorns, Elizabeth. Reproducibility concerns. Nature Medicine. 2012; 18(12):1736–1737.

Kramer, Adam D.; Guillory, Jamie E.; Hancock, Jeffrey T. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences. 2014; 111(24):8788–8790.

McGuinness, Deborah L.; Shaikh, Abdul R.; Moser, Richard P.; Hesse, Bradford W.; Morgan, Glen D.; Jacobs, Mike; Hunt, Yvonne; Tatalovich, Zaria; Willis, Gordon; Blake, Kelly, et al. A semantically-enabled community health portal for cancer prevention and control. Paper presented at the Third International Web Science Conference; Koblenz, Germany. 2011.

Mearian, Lucas. Wearable technology market to exceed $6B by 2016. Computerworld. 2012

Min, Hua; Ohira, Riki; Collins, Mindy A.; Bondy, Jessica; Avis, Nancy E.; Tchuvatkina, Olga; Courtney, Paul K.; Moser, Richard P.; Shaikh, Abdul R.; Hesse, Bradford W., et al. Sharing behavioral data through a grid infrastructure using data standards. Journal of American Medical Information Assocation. 2013; 21(4):642–649.

Moser, Richard P.; Hesse, Bradford W.; Shaikh, Abdul R.; Courtney, Paul; Morgan, Glen; Augustson, Erik; Kobrin, Sarah; Levin, Kerry Y.; Helba, Cynthia; Garner, David, et al. Grid-enabled measures: using Science 2.0 to standardize measures and share data. American Journal of Preventive Medicine. 2011; 40 Suppl. 2(5):S134–S143. [PubMed: 21521586]

Moser, Richard P.; Naveed, Sana; Cantor, David; Blake, Kelly D.; Finney Rutten, Lila J.; Ramirez, Susana G.; Liu, Benmei; Yu, Mandi. Integrative analytic methods using population-level cross-sectional data. Bethesda, MD: National Institutes of Health; 2013.

National Science Foundation. Cyberinfrastrure vision for 21 st century discovery. Arlington, VA: National Science Foundation; 2007.

Nelson, David E.; Kreps, Gary L.; Hesse, Bradford W.; Croyle, Robert T.; Willis, Gordon B.; Arora, Neeraj K.; Rimer, Barbara K.; Viswanath, Kasisomayajula; Weinstein, Neil; Alden, Sarah. The Health Information National Trends Survey (HINTS): Development, design, and dissemination. Journal of Health Communication. 2004; 9(5):443–460. [PubMed: 15513791]

Neumann, Juliane; Bras, Jose; Deas, Emma; O'Sullivan, Sean S.; Parkkinen, Laura; Lachmann, Robin H.; Li, Abi; Holton, Janice; Guerreiro, Rita; Paudel, Reema, et al. Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. Brain. 2009; 132(Pt. 7):1783–1794. [PubMed: 19286695]

Nielsen, Michael A. Reinventing discovery: The new era of networked science. Princeton, NJ: Princeton University Press; 2012.

Nosek, Brian A.; Bar-Anan, Yoav. Scientific utopia: I. Opening scientific communication. Psychological Inquiry: An International Journal for the Advancement of Psychological Theory. 2012; 23(3):217–243.

Ohno-Machado, Lucila. NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. Journal of American Medical Information Association. 2014; 21(2)

Parslow, Graham R. Commentary: Crowdsourcing, Foldit, and scientific discovery games. Biochemistry and Molecular Biology Education. 2013; 41(2):116–117. [PubMed: 23483655]

Platt, Richard; Carnahan, Ryan M.; Brown, Jeffrey S.; Chrischilles, Elizabeth; Curtis, Lesley H.; Hennessy, Sean; Nelson, Jennifer C.; Racoosin, Judith A.; Robb, Melissa; Schneeweiss, Sebastian, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: Status and direction. Pharmacoepidemioly and Drug Safety. 2012; 21(Suppl. 1):1–8.

President's Council of Advisors on Science and Technology. Designing a digital future: Federally funded research and development in networking and information technology. Washington, DC: Executive Office of the President of the United States; 2010.

Riley, William T.; Glasgow, Russell E.; Etheredge, Lynn M.; Abernethy, Amy P. Rapid, responsive, relevant (R3) research: A call for a rapid learning health research enterprise. Clinical and Translational Medicine. 2013; 2(1)

Riley, William T.; Rivera, Daniel E.; Atienza, Audie A.; Nilsen, Wendy; Allison, Susannah M.; Mermelstein, Robin. Health behavior models in the age of mobile interventions: Are our theories up to the task? Translational Behavioral Medicine. 2011; 1(1):53–71. [PubMed: 21796270]

Robert Wood Johnson Foundation. Health information technology in the United States: Progress and challenges ahead, 2014. Princeton, NJ: Robert Wood Johnson Foundation; 2014.

Rosling, Hans; Zhang, Zhongxing. Health advocacy with Gapminder animated statistics. Journal of Epidemiology and Global Health. 2011; 1(1):11–14. [PubMed: 23856371]

Shaikh, Abdul R.; Butte, Atul J.; Schully, Sheri D.; Dalton, William S.; Khoury, Muin J.; Hesse, Bradford W. Collaborative biomedicine in the age of big data: The case of cancer. Journal of Medical Internet Research. 2014; 16(4)

Simonsohn, Uri; Nelson, Leif D.; Simmons, Joseph P. P-curve: A key to the file-drawer. Journal of Experimental Psychology: General. 2014; 143(2):534–547. [PubMed: 23855496]

Sproull, Lee; Kiesler, Sara. Connections: New ways of working in the networked organization. Cambridge, MA: MIT Press; 1991.

Thacker, Stephen B.; Qualters, Judith R.; Lee, Lisa M. Public health surveillance in the United States: Evolution and challenges. Morbidity and Mortality Weekly Report. 2012; 61(3):3–9. [PubMed: 22695457]

Timms, Kevin P.; Rivera, Daniel E.; Collins, Linda M.; Piper, Megan E. Control systems engineering for understanding and optimizing smoking cessation interventions. Proceedings of the American Control Conference. 2013; 2013:1964–1969. [PubMed: 24362946]

Topol, Eric J. The creative destruction of medicine: How the digital revolution will create better health care. New York, NY: Basic Books; 2012.

Tortolero-Luna, Guillermo; Finney Rutten, Lila J.; Hesse, Bradford W.; Davis, Terisa; Kornfeld, Julie; Sancheza, Marta; Moser, Richard P.; Ortiza, Ana Patricia; Serrano-Rodríguez, Ruby A.; Davis, Kia L. Health and cancer information seeking practices and preferences in Puerto Rico: Creating an evidence base for cancer communication efforts. Journal of Health Communication. 2010; 15(Suppl. 3):30–45. [PubMed: 21154082]

U.S. Institute of Medicine. Public engagement and clinical trials: New models and disruptive technologies: Workshop summary. Washington, DC: National Academies Press; 2012.

Wild, Christopher P. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiology Biomarkers and Prevention. 2005; 14(8):1847–1850.

Winerman, Lea. Interesting results: Can they be replicated? Monitor on Psychology. 2013; 44:38–41.

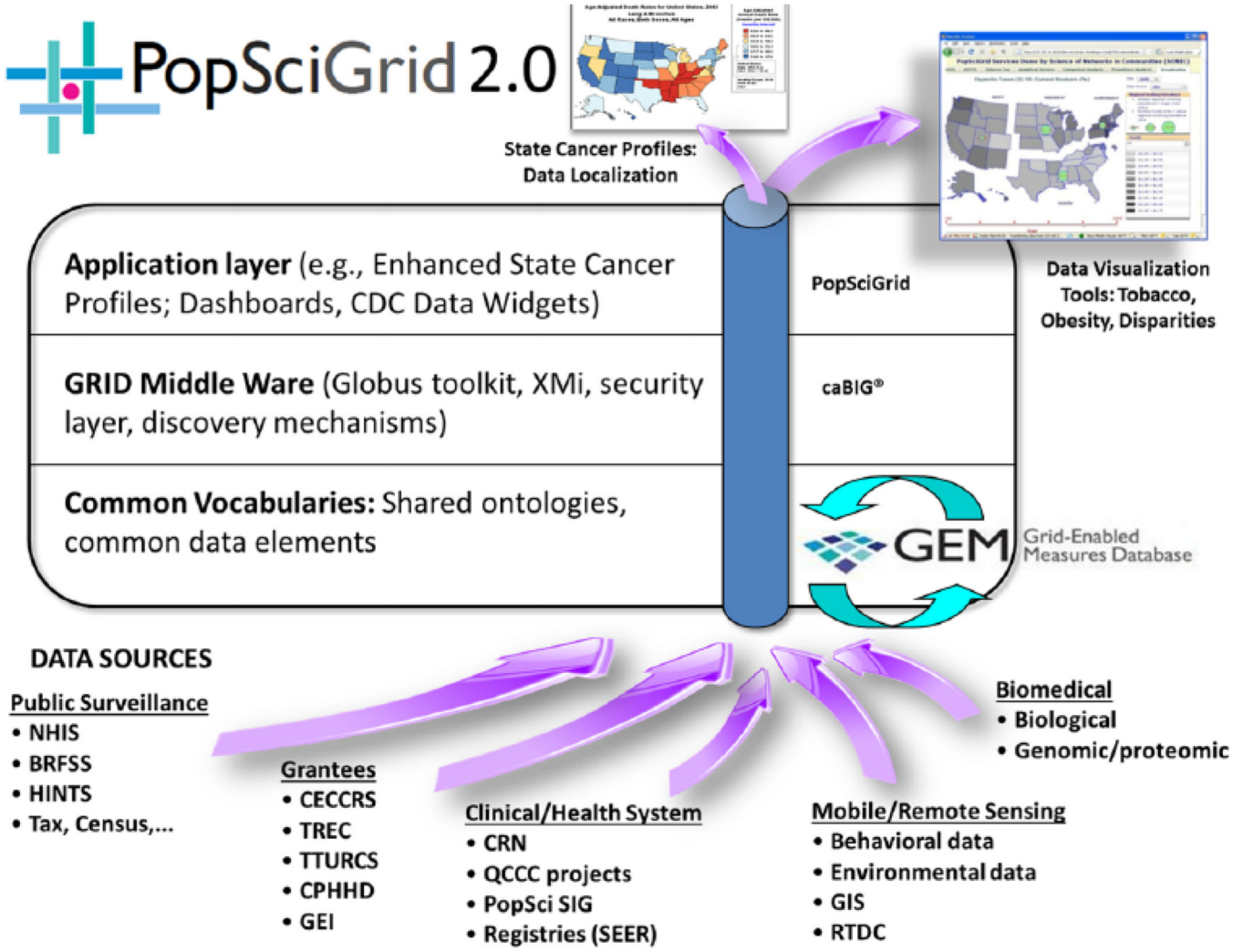

Figure 1.

"Population Science Grid 2.0"

NOTE: From the schematic, multiple data sources are integrated in real time using common vocabularies and an extensible Web Service middleware layer to present users with an integrative view of health behaviors, social determinants, and public health outcomes. The Grid Enabled Measures (GEM) tool utilized a web 2.0, crowdsourcing approach to data harmonization. Population Science Grid 2.0 was developed between 2007 and 2009.

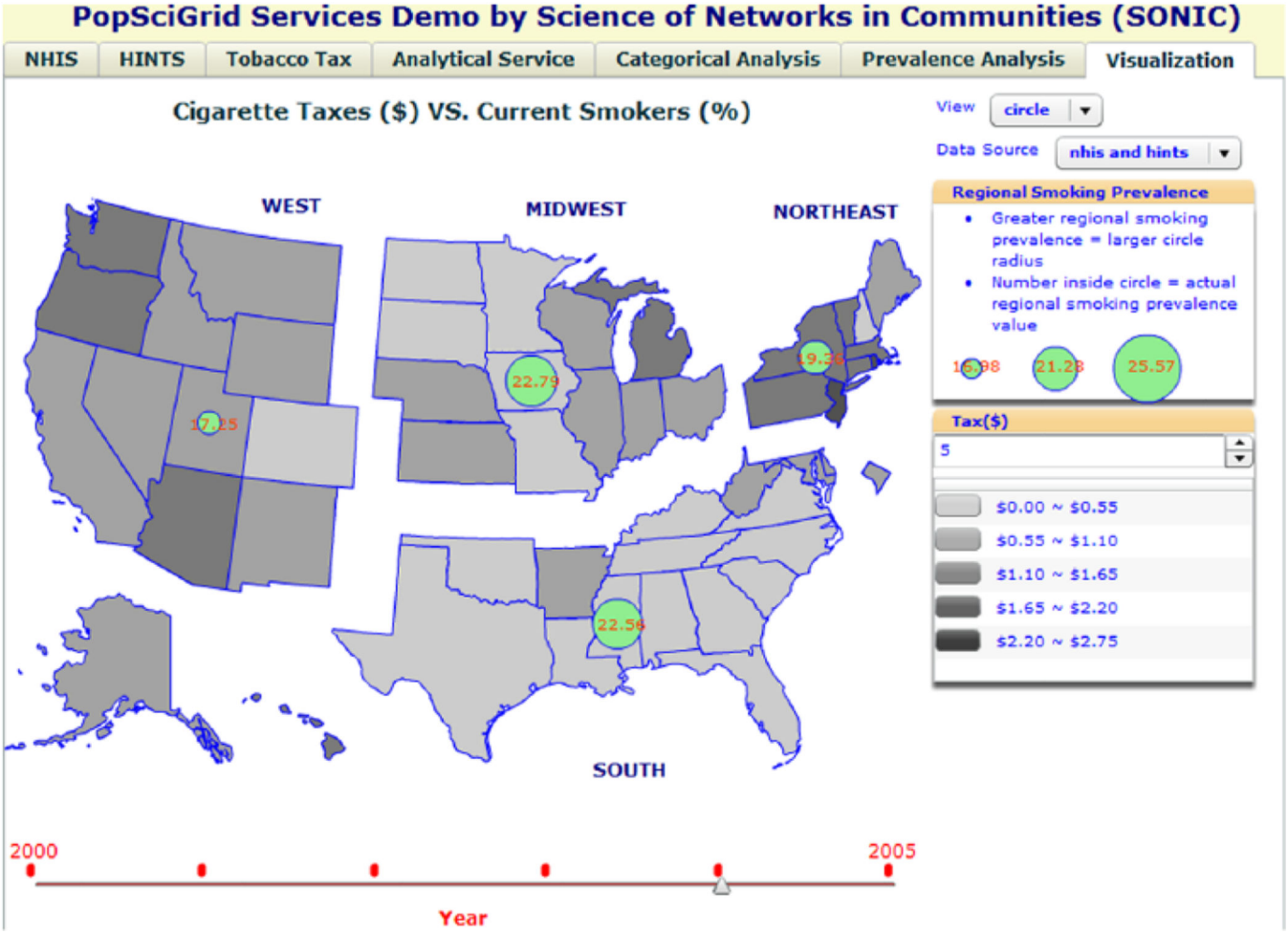

Figure 2.

How Multiple Data Streams Could be Integrated through a Common Interface to Steer Community-Level Situational Awareness and Action.