Title: Better Evidence-Based Policy by Infusing Big Data into Social Sciences Research

Point of Contact: Maureen A. Pirog
School of Public and Environmental Affairs, SPEA 410H, IUB
812-855-0747 (direct)
812-360-0232 (cell)

One Sentence Summary: The Grand Challenge (GC) for Indiana University (IU) is to inform policy making in Indiana and at all levels of government in ways not previously possible by harnessing its' technical expertise in informatics and high performance computing and make those resources accessible to world-class social science researchers and doctoral students at IU who have no training in the types of data collection, storage, analysis and visualization needed to realistically incorporate big data into their research.

Abstract: Governments at all levels, including the State of Indiana, want data driven policy but social scientists are not well poised to take advantage of big data to help. While IU has outstanding computing resources, extraordinary talent in data analytics, and gifted social scientists, the ability to use big data for social science research has been hampered by lack of training and experience. This GC is to disseminate knowledge regarding big data and data analytics to the community of IU social scientists, very broadly defined. The following tasks, taken together, will place IU’s social sciences at the leading edge of the big data revolution, raise the profile of IU’s scholarship and provide practical insights to important social problems. (1) Fund 10-15 rotating social science research projects that actually use big data for 1-2 years each. The research projects must have important practical and policy implications for diverse audiences. (2) Host Data Carpentry workshops regarding the use of big data. (3) Establish a speaker series on using big data for social scientists. (4) Expand the capacity of the Center for Survey Research (CRS) to include big data. CRS has been utilized for decades by social scientists, making it an ideal vehicle for this task. (5) Expand educational opportunities in the School of Informatics and Computing (SOIC) that target social scientists. This project includes collaborators from IUB and IUPUI, from six schools/colleges at IU, the CRS, and the Social Science Research Commons (SSRC) as well as external collaborators.
THE GRAND CHALLENGE (GC): INFUSING BIG DATA INTO SOCIAL SCIENCE RESEARCH

State of Indiana’s leaders and policy makers everywhere care about healthcare costs, public health, homeland security, the state’s position in the global economy, economic development, new mechanisms for improving the educational outcomes of children, energy, and how we can better evaluate environmental risks. This GC tackles all of these issues utilizing big data not previously accessible to IU’s social science scholars. A keyword search of the top ISI-ranked journals found hundreds of articles with big data terminology but only 10 actually used big data.¹ Big data are not being used by social scientists who currently lack the expertise to collect and analyze new types of very different and challenging data. The GC for IU is to inform policy making in Indiana and at all levels of government in ways not previously possible. To do this requires harnessing IU’s technical expertise in informatics and high performance computing and make those resources accessible to world-class social science researchers and doctoral students at IU who have never been trained in the types of data collection, storage, analysis and visualization needed to realistically incorporate big data into their research. Federal agencies are hosting big data conferences, the National Science Foundation is keenly interested in big data – IU should be leading the way, not following in the wake of other institutions.

GOALS: Succinctly stated, there are three goals of this GC.

1. Make big data accessible to the existing social science research community at IU. There are five components of this goal.
   a. Locate technical data sampling and gathering for “big data” in the IU Center for Survey Research and fully subsidize the use of this expertise for at least five years to encourage social scientists to conduct studies that use big data.
   b. Host Data Carpentry workshops for IU faculty and doctoral students (our social science PhD students will benefit tremendously) in the “how-to” of collecting, storing and analyzing big data throughout the academic year.
   c. Create a university-wide speakers series to bring in experts from Google, Internet Explorer, Mozilla, Telecommunications, business and government agencies to discuss uses and access to big data. The recent visit to IU by Dr. Tom Cook who spoke on the topic of “big data” should be viewed as a first-step in this direction.
   d. Create and host an Annual Summer Workshop on Using Big Data. Workshop participants will be selected competitively based on short proposals for specific research projects. Workshops open to the academic community will serve to highlight the strategic, competitive advantage of IU in this arena.
   e. Expand curricular offerings targeted to the social sciences in SOIC.

¹ The journals searched were the Annual Review of Anthropology (rank=4, Impact Factor (IF)= 2.717, the Quarterly Journal of Economics (rank=1, IF=6.654), Political Analysis (rank=1, IF=4.655), American Political Science Review (rank=2, IF=3.688), Journal of Applied Psychology (rank=2, IF=4.799), Review of Education Research (rank=1, IF=3.897), Journal of Public Administration Research and Theory (rank=1, IF=2.833), Journal of Policy Analysis and Management (rank=2, IF=2.576), Demography (rank=1, IF=2.616), Global Environmental Change-Human and Policy Dimensions (rank=1, IF=5.089), Harvard Law Review (rank=1, IF=4.979), Journal of Criminal Justice (rank=1, IF=3.154), Criminology (rank=2, IF=3.098), Nature Climate Change (rank=1, IF=14.547), Academy of Management Review (rank=1, IF=7.475). The Journal of Policy Analysis and Management is the top ranked policy analysis journal but is included by ISI in the Public Administration category. Two journals were included from criminology and political science because the content of the top journals were known to be quite different in each other.
(2) Fund 10-15 social science research projects each year that will use big data on topics of social and scientific merit. Nine proposed projects are discussed in this pre-proposal. A broader Call for Proposals will be issued from all campuses of IU should this pre-proposal move forward. Social science researchers will be connected with technical support from SOIC, SLIS, COAS-CL and/or the CRS. In each case, we have and will continue to consider the potential of these projects to secure external funding.

(3) Further expand capacity for using big data in the social sciences by strategically hiring new faculty and/or staff with big data expertise. Having experts conveniently located proximately will encourage and facilitate the use of these types of data. It is suggested that specific department and schools for new hiring be based in part on the receptivity of faculty to engage with big data (across schools and departments) as well as bolster the excellence of high-performing departments.

PROPOSED RESEARCH AND ITS IMPACT: This project brings the full array of data into the social science portfolio with the express purpose of providing rigorous, scientific evidence to support government decision making in areas where this is not possible without incorporating big data. An ancillary benefit is that this project provides important information about the best contexts in which to use big data as well as the reliability of big data (as several projects will compare results which also include more traditional sources of social science data). Another benefit is that the research will increase the visibility of IU as the social science papers appear to garner roughly 10 times the citations as articles in the most highly cited social science journal.

To demonstrate the feasibility of our approach at IU, social scientists from IU and IUPUI, and six schools or colleges have proposed nine research projects to be initiated in year-one of the project. They are:


With the legalization of marijuana in Colorado, sales have skyrocketed and roughly one third of the sales are of marijuana edibles: candies (e.g., gummies, chocolate bars), brownies, cookies, and sodas. Most of the edibles are difficult to distinguish from similar edibles that are not laced with THC, the active psychoactive constituent in cannabis. News media have provided unsubstantiated and anecdotal reports of THC poisoning in children who mistake these edibles for regular treats. Nonetheless, in light of the revenue raised by Colorado, and the reduction in prison crowding, many states are looking to Colorado and Washington’s experiences with legalization of marijuana. Standard emergency room medical coding does not yet include special coding for marijuana poisoning.

Because medical records do not specifically code THC poisoning or marijuana poisoning as a reason for admission to emergency rooms, this project proposes to use geographically based Internet search and Twitter to examine the frequency of searches for terms closely related to THC poisoning in children (e.g., marijuana poisoning in children, overdoses of THC in children) before and after the legalization of marijuana in Colorado. The state of Washington had a different timeline for the legalization of marijuana will also be included in this study as well as geographically proximate control states (or border counties). Comparison states will be selected as well. To compare and contrast the advantages of standard hospital admissions and big data, we also will incorporate standard emergency room admission
data on poisoning in children (as this should increase after the easy availability of marijuana edibles) in this study. Our project looks at the medical outcomes for an important and typically protected subpopulation, children. It is a topic that will be of keen interest to policy makers.

Ashley Clark (Center for Survey Research), Comparing Traditional and Big Data Methods for Gathering and Analyzing Public Opinion

Researchers across a wide range of disciplines who measure attitudes, knowledge, and behaviors have relied on focus groups, field research methods, interviews and large-scale sample survey methods for decades as important methods for capturing data that address important societal issues. The explosion in big data presents a new world of possibilities for researchers. We are also in a time when nonresponse and rising costs threaten the viability of traditional methods for social, behavioral and health research. Under what circumstances should big data replace or supplement traditional methods is an important question.

CRS is planning a series of studies to compare traditional survey methods with big data. The first is a study of the US public’s support for the use of fracking to extract natural gas. Information on the level of support for fracking and the possible factors influencing such support is crucial for policymakers, industry and environmental organizations as well as academic researchers. We will compare public opinion data from an online survey using a panel of respondents who were recruited using random sampling methods with results from a media content analysis and an analysis of Twitter data. We will assess to what extent these sources produce different or similar results in gauging the amount of pro or anti-fracking support and to what extent they can explain the determinants of support. Additionally, we will assess the cost-effectiveness and timeliness of these sources, as these are also crucial considerations in obtaining and using public opinion data.

William Kindred Winecoff (COAS- Political Science), Mapping the Global Structure of Economic Corporations.

Mapping the global structure of economic corporations is critical for understanding the production and investment decisions that collectively determine the development of the world’s economy. Taken together, these economic actors comprise a network of collaboration, competition, co-ownership, and hierarchical organization. In recent years several datasets have been developed that are comprised of millions of observations of firms, executives, intra-conglomerate divisions, and inter-firm interdependencies. These data are cross-sectional and inter-temporal. But methodologies to analyze these data in a principled way with links to existing theory in business, economics, and political economy are much less well developed. This result is that academics, policymakers, and private actors are unable to make use of all of the information that is available.

This project would develop such a methodology that can be used on existing data but is flexible enough for use on data that will become available in the future. A significant component will involve developing best practices for use on network data which are currently too large to be accommodated computationally. These data are not only large, but also complex (meaning that they involve endogenous processes and have a non-trivial topology). Substantively it would principally be concerned with two issues. First, it would analyze the structure of global ownership and corporate management as a network, in order to understand how the performance of national economies is dependent upon the decisions taken at the core of this inter-linked system. Second, it would look at the social structure of
firms and their managers themselves by examining shared board memberships and inter-firm transactions.

Cindy Hmelo-Silver (School of Education), *Big Data from Small Groups: Using Learning Analytics to Support Collaborative Learning.*

Technology has enabled new forms of collaboration that allow learners to benefit from a broader range of sources and allow teachers to tap into different methodologies than ever before. Many studies on the intersection between technology and collaboration have already been conducted but leave room for more work to be done.

Our next step is to delve deeper and extract actionable indicators from research on collaboration that could be used to develop new technologies that would support developing models of collaboration that lend themselves well to learning analytics and to better adaptive support for collaborative learning.

A major goal of new research on collaborative learning is to produce new detailed models of collaboration that could lead to the development of adaptive scaffolding. Studies to examine this would be set in the context of large-technology rich classrooms in STEM subjects and professional education to better understand how adaptive support can be used to allow innovative pedagogies such as problem-based and other inquiry pedagogies to be used on a broader scale. First steps include understanding what the indicators are and studying how experts use this kind of data to adaptively scaffold collaboration. This can be used to create instructor dashboards with visualizations and later to determine how adaptive support for collaboration can be distributed across the instructor and automated support tools.

Jeffrey Gruenwald (SPEA-IUPUI-Criminal Justice), *Temporal and Geospatial Patterns of New Media Use Prior to and Following Terrorist Attacks*

The goal of this proposal is to harness the power of big data for identifying temporal and geospatial patterns of new media use prior to and following terrorist attacks and foiled plots. Several analyses utilizing big data are planned that will map how specific uses of new media by extremist groups, advocacy organizations, and others are distributed over time and place in relation to several outcomes of interest, such as the successfulness of terrorist plots and target selection. Other planned studies will develop statistical models to predict how patterned behaviors affect these same outcomes.

The proposed research will extend the work of Dr. Gruenewald already has some in creating large, open-source databases on terrorism and other forms of extremist crime. This work will also open the door for multi-disciplinary collaborations between SPEA and SOIC faculty who have expertise in utilizing big data and homeland security across IUPUI and IUB campuses. Several stakeholders stand to benefit from the planned analyses. In particular, findings will inform the policies and practices of law enforcement, intelligence analysts, and private security organizations responsible for the prevention of terrorism. This research holds significant potential for future external funding, as the DOJ, DOD, and DHS continue to fund research programs that promote the use of alternative data sources and the development of innovative methodologies for furthering homeland security and national security interests.

The goal is to reveal the production and ownership structure of the world economy, which in turn should help us better understand key issues related to development, inequality, regulatory politics, and state-firm interactions.
Haizhen Lin (Kelley School of Business). *What is Valued More in Consumer Choice: Government or Social Media Ratings?*

With online ratings playing an increasing role in consumer decisions, little is known about their impact, especially relative to official government ratings. The impacts of government and social media ratings will be disentangled from each other in the context of hospitals: involuntary ratings based for Centers for Medicare and Medicaid Services (CMS) will be compared with voluntary ratings from consumer reviews on social media such as Facebook.

More specifically, we propose to answer the following research questions. First, does social media rating tell consumers anything beyond what they already know? Do social media reviews substitute for or complement government ratings? Second, are these two types of ratings valued differentially by consumers? Do consumers trust government ratings or social media ratings more? Third, to what extent do these ratings affect consumers’ decisions over their choices of hospitals? Last, how do these ratings affect hospitals’ behaviors and operations, for example, the implementation of quality improvement initiatives? Four data sources are going to be used: the hospital compare database from CMS, consumer ratings on social media including Facebook, the American Hospital Association annual hospital survey, and hospital cost reports. This project will shed light on how online resources shape the healthcare industry.

Dan Sacks (Kelley School of Business), *Health Insurance, Health Care, and Health*

The 20th and 21st centuries have seen an enormous growth in both health care spending and health, especially life expectancy. Surprisingly, however, we about the connectedness of these two trends. Increases in life expectancy derive from low-cost public health interventions like closed sewers or anti-smoking campaigns. Much of the rise in health care spending comes from the use of costly medical technologies with uncertain health value, such as robot-assisted surgery. A critical component of evaluating the rise in health care spending is addressing whether new medical technologies bring about sufficient health gains to justify their cost.

This question is hard to answer because historically it has been impossible to assemble the necessary data. However, new data sets are coming together which provide an opportunity to overcome previous data limitations. Health insurance companies are creating databases which link insurance claims with medical charts, making it possible to see, in tremendous detail, patients’ health as they experience different treatments. Linking in additional information on insurance plan choices and options creates exciting research opportunities, for it would let research trace out the links from health insurance coverage to health care choices to health outcomes, and provide evidence on the value and costs of recent health care innovations.

Diane Henshel (SPEA) and Sandra Kuebler (COAS-Computational Linguistics), Geoff Scott – University of South Carolina, and Pat Fair, National Oceanic and Atmospheric Administration, *Minimizing Social, Economic, Environmental and Health Risks from the Charleston Harbor, SC, Dredging Project*

The Charleston Harbor dredging project is a federally funded project that will deepen Charleston Harbor from 45 to 52 feet to enable larger container ships to use the harbor and its ports. The dredging project will make Charleston Harbor the deepest harbor on the East Coast, enhancing the economic efficiency of East Coast commercial navigation.
This proposed research project, being planned with collaborators in South Carolina (and with a proposed advisory board including top risk and decision analysts from across the country) will expand landscape scale risk assessment modeling to integrate metrics for environmental and human health (typical risk assessment metrics) across multiple scales of biological organization with metrics for ecosystem services (benefits to humans as well as inherent ecosystem benefits), socioeconomic and economic development costs and benefits, and disparate socioeconomic impacts using both classic sources of information (federal and state environmental, census, health and economic databases and data resources) as well as social media as a means to track acute social and economic impacts during the project. Social media (Facebook, Twitter, Google searches) can be analyzed both longitudinally, but also acutely to assess impacts on such factors as quality of life, jobs, and personal perspectives on air quality and water quality long before such factors would be quantified and available in any aggregated monitoring or tracking metrics. The acute and on-going assessments will also be used to provide feedback to regulatory and implementing agencies to help modify policies guiding the dredging project so that the project has minimal adverse impact on ongoing life within the Charleston Harbor region.

Seth Freedman (IUB-SPEA) and Ezra Golberstein, School of Public Health, University of Minnesota.  

Effects of Time Pressures in Primary Care

The project explores how increased time pressure in primary care affects physician behavior. This research sits at the intersection of two key areas for health services and policy. One area is the concern about shortages of primary care providers (PCPs) and the related concerns that PCPs routinely lament the limited time they have with their patients. The second area is the current emphasis on shifting the health care system toward delivering “high value” care. Many health policy experts believe that an enhanced foundation of primary care is crucial to increasing accessibility while avoiding the use of unnecessary or low-value services. We will examine whether primary care encounters with more or less time pressure lead to different patterns of prescribing medications, including any prescriptions, brand versus generic prescription, and potentially-inappropriate prescriptions (e.g., antibiotics for acute respiratory infections); orders for diagnostic tests, including potentially inappropriate orders (e.g., imaging for low-back pain); and referrals to specialty care.

This project will utilize rich, encounter-level data from the Fairview Electronic Health Records system, including daily clinic-level scheduling data, patient demographics, reason for visits, diagnoses and procedures within the visit, de-identified provider identifiers, and outcomes of the visit (tests ordered, referrals, prescriptions written). We will obtain data on clinics served by Fairview from hundreds of thousands of primary care encounters. In health care research, it is often difficult to get a full picture of a patient’s contact with the healthcare system. While high quality data exist on hospital visits, and Medicare data can be used to observe all encounters for elderly patients, there are few datasets that allow researchers to track detailed information on primary care visits over time from the non-elderly population. Electronic health records from large provider organizations provide new opportunities to observe detailed information on primary care encounters and to link these encounters to downstream health care utilization, such as specialist office visits and hospitalizations.

RESOURCES: Informatics Human and Technical Infrastructure

IU is well positioned to aid the transition of social science research at IU towards Big Data driven research because of its data and computing infrastructure (Scholarly Data Archive, Karst, Jetstream, Big Red II), its novel classroom technology (Mosaic) and existing capability in network science (IU Network
Institute, IUNI). It capitalizes on the talent of gifted social scientists as well as key collaborators who are world-experts in informatics, data analytics, and data visualization.

This proposal builds off this solid foundation to build new infrastructure to span three intended needs: education to *enable*; expertise to *support*; and a shared data commons to *suggest*. The infrastructure *enables* through short-term measures and long-term plans for education of researchers in the social sciences in Big Data skills. The infrastructure *supports* the ongoing needs of Big Data-integrated social science research through embedded expertise through all stages of the data lifecycle. Finally, by bringing together methods, models, datasets into a social science data commons with attention to the lifecycle of each, the infrastructure becomes an ecosystem that *suggest* new research questions into the future. The infrastructure piece of this proposal draws expertise from multiple faculty in the SoIC (SoIC) including but not limited to B. Plale, K. Borner, F. Menczer, G. Fox, and D. Wild, and Pervasive Technology Institute (PTI) including but not limited to B. Plale and C. Stewart.

(1) Education to *enable*: education to enable is a strategy to both quickly orient researchers around Big Data so as to have maximal impact early, and impact longer term the research careers and research direction of IU. There has been much success nationally with Data Carpentry sessions ([www.datacarpentry.org](http://www.datacarpentry.org)). Data Carpentry events would become yearly events at IU, sponsored by this GC. Longer term, IU is uniquely positioned with its new data science program. While currently germinating and growing from within the SoIC, the Data Science Program is structured with the intention of easily extending into discipline-specific areas of interest. The Data Science program is highly relevant to the needs of those wanting to learn to manipulate Big Data in the context of discipline-specific research questions. The Data Science program would be extended under this GC to include courses of relevance to the social science research being proposed in the GC which could take the form of a PhD minor, certificate, or degree. The new faculty hires proposed as part of this GC would teach courses of relevance to the Data Science program and be invited to affiliate with the program.

(2) Expertise to *support*: for the GC to be a success, the SoIC colleagues in the GC will have to engage in the research questions early in their study so as to share methodology, and remove barriers to access to the datasets of interest. Staff and students will participate in the research GC hosted discussion forums and thus learn as well. In out years (beyond year 2 or 3) the skill set is transferred and can be used to strengthen IU Center for Survey Research. SoIC has faculty and research scientists who themselves have social science backgrounds. The forum of the GC will foster their engagement as well.

(3) Shared data commons to *suggest*: Informatics and Computing researchers in Big Data (and more broadly) look at research problems as a source of generalization. Generalization, in fact, is a methodology built into computer science. As such, to the extent possible by policy and law, the SoIC researchers will accumulate the data sets, models, results, and software into a shared data commons. F. Menczer has already agreed to share his massive Twitter dataset. The shared data commons will draw on research in secure, virtual compute environments pioneered by PTI. The commons will be able to leverage efforts such as the HathiTrust Research Center ([www.hathitrust.org/htrc](http://www.hathitrust.org/htrc)) for the analysis tools it provisions for analysis of digitized books from research libraries, and the Cyberinfrastructure for Network Science Center for tools it provides for
analysis of network data. The former is available through one of the technical leads (Plale) and the latter is available through IUNI of which one of the technical leads (Menczer) is a founding member. Large-scale data management can leverage systems and tools such as Karst, Jetstream, and Big Red II in UITS through affiliation of C. Stewart and FutureSystems in SoIC through G. Fox.

The coupling of excellent and motivated social science research with human and technical informatics infrastructure, leveraging the existing expertise in IUNI and resources in UITS creates the underlying infrastructure (education to enable, expertise to support, shared data commons to suggest) that will propel social science at IU into a leading position in Big Data-driven research outcomes.

New resources requested will be funding for the research projects, speaker series, Data Carpentry and summer workshops, and strategic hires. The rationale for these activities has already been described.

TEAM: This project will be organized by faculty from SPEA and SOIC. SPEA faculty coordinating the project include Drs. Maureen Pirog, Kosali Simon and Diane Henshel. Drs. Beth Plale and Fil Menczer from the SOIC will also be instrumental in organizing the project. Each of these individuals and their respective roles are described below. Additionally, the Drs. Craig Stewart, Geoffrey Fox, Katy Borner, and Sandra Kuebler are part of the technical team that will support this project. The CSR and the Social Science Research Commons both endorse and support this GC fully. Individual faculty from IUPUI and IUB representing SPEA, SLIS, SOIC, the COAS, School of Education, Kelley School of Business, CRS and SSRC want to participate as researchers.

**Dr. Diane Henshel.** Dr. Henshel is an Associate Professor in SPEA, working at the edge between science and policy. An internationally known expert in Risk Assessment and Toxicology, Dr.Henshel has carried out multiple landscape scale analyses incorporating what was, at the time, the cutting edge of “big data” (a phrase for which the cutting edge changes with new technological developments). Dr. Henshel is currently a co-PI and Risk Lead on a DoD funded project developing a new approach to Risk Assessment for CyberSecurity, a project which generates its own “Big Data” associated with modeling network flows and changing risks in each risk model. Diane will conduct a research project in year one and additionally will take responsibility for GC 1 (parts b, c and d).

**Dr. Maureen Pirog.** Dr. Pirog is the Rudy Professor of Policy Analysis. Dr. Pirog has authored several papers on the potential for big data and how data drives innovation in research arguing that data and methods are co-dependent – hence, the necessity to facilitate learning new methodological approaches. Dr. Pirog spent ten years as the editor-in-chief of the *Journal of Policy Analysis and Management*, the top-ranked policy analysis in the world. The journal is interdisciplinary giving her some credibility outside her specific areas of research. She will take responsibility for coordinating GC (1a) along with the director of the Center for Survey Research, Dr. Ashley Clark. Dr. Pirog will also gather data and facilitate discussion around GC #3.

**Dr. Kosali Simon.** Dr. Simon is a Professor in SPEA, and a Research Associate of the National Bureau of Economic Research. She is a health economist with considerable experience conducting policy relevant research involving large scale data sets, and with developing capacity for computing advances in social sciences. At IU, she has spearheaded the development of infrastructure for sensitive data computing for social science. She is also active in boards of professional organizations and editorial positions of
journals in health economics and policy. Dr. Simon will take the lead in the solicitation and selection of projects for GC #2.

Dr. Beth Plale. Dr. Beth Plale is a Professor in the School of Informatics and Computing, an affiliated faculty member of the Ostrom Workshop on Political Theory and Policy Analysis, and Science Director of the Pervasive Technology Institute. She is founding director of the HathiTrust Research Center, a member of the IU Data Science Program, and founding director of the Data To Insight Center at IU. Alson with Fil Menczer, she plays a pivotal role in coordinating the technical side of this proposal with the social sciences.

Dr. Fil Menczer. Dr. Fil Menczer is a Professor in the School of Informatics and Computing at Indiana University, Director of the Center for Complex Networks and Systems Research, and member of the Cognitive Science program. Dr. Menczer is an ACM Distinguished Scientist, a Fellow of the ISI Foundation, and a Senior Research Fellow of the Kinsey Institute for Research in Sex, Gender, and Reproduction. His research is supported by the NSF, DARPA, and the McDonnell Foundation. It focuses on Web science, social networks, social media, social computation, Web mining, distributed and intelligent Web applications, and modeling of complex information networks.

SUSTAINABILITY: As new sources of data become increasingly available, researchers re-tool and adapt to the richer resource environment. New approaches to the collection and analysis of big data will become part of the tools-kit portfolio of social scientists and integrated into their curriculum once faculty and our social science doctoral students gain facility with these new sources of data. Researchers who enter this domain early will have competitive advantages in high-profile publishing, grant seeking and advising the conduct of business, government and the third sector. Our social science faculty are resourceful, brilliant and motivated. This GC is intended to facilitate this transformation and place IU faculty at the head of this shift – to be leaders, not followers, in exploiting these data. Once the “know-how” or proverbial cat is out of the bag, it will be incorporated into the daily conduct of business until new data generating technologies are developed and disseminated on a broad scale. We believe that there should be a ten year commitment to this challenge on the part of the university with resources loaded at the front-end (years 1-3) of the project.

PARTNERS: This GC tackles important social problems (e.g., public health, healthcare affordability, energy, homeland security, assessing environmental risks, new technologies for student learning). Each research team will interact with public officials in Indiana and at the national level who are most closely connected to each individual project. Because the link between these research projects is their reliance on big and it is not driven by a single social problem, our external partners will be more diffuse than those found in other GC pre-proposals.

This broader portfolio of social problems addressed by this pre-proposal suggests that this project would benefit from the assistance of an Advisory Board to help guide the implementation of this GC. Should this pre-proposal move to the next stage, we will seek the advice and assistance of the Office of the Provost for Research and the President’s Office for assistance in identifying members for this advisory board. We will focus on key stakeholders in Indiana, national agencies with interests in this research, as well as data gate-keepers, and applied researchers who have bridged this gap. The entire team of faculty who are organizing this project will work together with university officials to comprise this board.
METRICS: The goal of this project is to facilitate the use of big data by IU’s social science faculty and PhD students. As such, the metrics used will include the standard ones at the university. Were research projects competed? Was new external funding solicited that capitalized on the new skills learned? To what extent were the expanded capabilities of the university used by project participants and others? Were new courses on using big data developed in the social sciences? Did doctoral dissertations incorporate big data because of the activities supported by this challenge? Were workshops, talks and summer seminars attended? Well received? To what extent did the research teams funded by this challenge interact with key stake-holders? Influence their operations?
GRAND CHALLENGE

BIOS/CVS OF PARTICIPANTS

INTEGRATING BIG DATA INTO SOCIAL SCIENCE RESEARCH

Katy Börner (SLIS and School of Informatics and Computing)
Geoffrey C. Fox (School of Informatics and Computing)
Jeff Gruenewald (SPEA-IUPUI-Criminologist)
Sandra Kuebler (COAS-Computational Linguistics)
Cindy E. Hmelo-Silver (School of Education)
Haizhen Lin (Kelley School of Business, Department of Business Economics and Public Policy)
Filippo Menczer (School of Informatics and Computing)
Jill Nicholson-Crotty (missing)
Jill’s co-author from the Kinsey Institute (missing)
Victoria Perez (Health Policy-SPEA)
Maureen Anne Pirog (SPEA-Social Policy)
Beth Plale (SoIC)
Sean Reardon (SPEA)
Dan Sachs (Kelley School of Business)
Kosali Simon (SPEA-Health Policy)
David Wild (School of Informatics)
William Kindred Winecoff (Department of Political Science)

Biographical Sketch