The National Institutes of Health

BD2K Behavioral and Social Sciences and Big Data Workshop

March 19-20, 2018

Revised July 5, 2018

RLA
ROSE LI & ASSOCIATES

This meeting summary was prepared by David Clarke, Rose Li and Associates, Inc., under contract to the National Institutes of Health (HHSN271201400038C) in collaboration with Regina Bures. The views expressed in this document reflect both individual and collective opinions of the meeting participants and not necessarily those of the National Institutes of Health or the U.S. Department of Health and Human Services. Review of earlier versions of this meeting summary by the following individuals is gratefully acknowledged: Stan Ahalt, Elisabeth Bruch, Regina Bures, Cheryl Eavey, Barbara Entwisle, Carlos Gallo, Elizabeth Ginexi, Jeffrey Groen, John Haaga, Della Hann, Vasant Honavar, Chandra Keller-Allen, Jonathan W. King, Bruce Y. Lee, Margaret Levenstein, Shari Ling, Tyler McCormick, Wendy Nilsen, John W. R. Phillips, Michael Rendall, Amy Rose, Susan Tenney, and Nancy Tuvesson.

Table of Contents

Executive Summary

Introduction

On March 19-20, 2018, behavioral, social science, computational big data, and informatics researchers participated in a workshop convened by the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) Common Fund Program. The workshop goals were to encourage researchers from these fields to engage in cross-disciplinary discussion and collaboration through a series of individual and panel presentations. Six invited experts presented papers on critical big data issues and concepts as well as case studies of big data use in behavioral and social sciences research. Panels of invited speakers addressed topics including federal big data and health research, big data training for behavioral and social sciences researchers, opportunities for incorporating behavioral and social sciences in big data research, and central data resources.

State of the Science

In the NIH context, big data are generally associated with biomedical research fields such as genomics, where petabyte sized datasets are common and field-specific data sharing policies have been established. Big data computational models have historically had more traction with systems biology and neuroscience researchers than behavioral and social sciences researchers. More recently, behavioral and social scientists have begun applying systems approaches to maximize insights from big data through applications to public health problems such as addressing health disparities.

There is a growing acknowledgment that the behavioral and social sciences can benefit from and contribute to the analysis of data characterized by the 4V's: high volume, velocity, variety, and veracity. These big data are generated by digital sensors, mobile devices, geographic information systems, and other means in amounts vastly larger than traditional self-report surveys, questionnaires, direct observation, or randomized controlled trials (RCTs), and include the voluminous administrative datasets generated by federal agencies.

Although rapidly becoming incorporated into many types of scientific research, big data are not specifically designed for research purposes and may lack population representativeness. So-called found data have an inherent susceptibility to bias because people self-select into populations from which data are generated. However, this problem is not limited to big data; questions about representativeness also arise in RCTs, whose participants are volunteers. Bias should not be feared but managed, for example, by pooling data to increase analytic power or applying tools developed to understand survey measurement error to big data.

Key Challenges

Linking Big Datasets

While big data presents new opportunities for researchers, linking disparate datasets can create complex technical and ethical challenges related to a loss of representativeness, a lack of common identifiers, and issues with respondent consent and data privacy. Advances in matching technologies that increase the likelihood of correct linkages, as well as methods that balance the cost of false positives (linking records from two distinct individuals) and false negatives (discarding records when a successful linkage is not made), have increased the feasibility of linking datasets. As cross-linkages create new privacy concerns, researchers need to better understand the expectations of study participants whose data may be used in ways not originally stipulated. Meeting speakers agreed that cross-linkages of federal datasets should be pursued to further health-related research, but only after development of standard protocols for doing so. Federal Statistical Research Data Centers, where linkages can be made, are a positive step in this direction. Linked data are expected to be key resources in the future: linked data can enable researchers to address new research questions and reduce research costs. NIH researchers should make greater use of linkages between study data and available administrative datasets (e.g., from the Centers for Medicare & Medicaid Services and the National Death Index).

Training for Big Data Science

Current behavioral and social sciences graduate training is inadequate to meet emerging data analytics needs. Multidisciplinary collaboration will be an essential element of training for the next generation of biomedical behavioral and social science researchers. Universities should encourage programs featuring collaboration and integration between traditional behavioral and social sciences programs and programs in data science fields such as computer science, applied math, and informatics. Effective training in the analysis of big data requires an application domain. Behavioral and social scientists need knowledge about data science, and data scientists need knowledge about behavioral and social sciences. To move forward collaboratively, training programs will need to commit to developing a dialogue about common terminologies, ethics, reconciliation of privacy levels for different data platforms, and proactively addressing other issues as they arise.

Incorporating Behavioral and Social Sciences into Computational Health Science

Although time consuming, communication is essential to the success of the team science required by big data. Cross-disciplinary communication about big data is complicated by the fact that big data research models are often quantitative computer programs written in non-standard languages, in proprietary code, with high processor demands, run on remote servers. Nonetheless, the gap between the language of the behavioral and social sciences and that of big data must be bridged. The field of computational biology has evolved over the past 30 years because computer science specialists and biologists have built working relationships. The next generation of scientists skilled in working at the nexus of big data and behavioral and social sciences can be created through a similar process. A model exists in funding agencies, like NSF, that have supported multidisciplinary groups of investigators working on emerging problems and receiving, to some degree, cross-training.

Key Conclusions

In general, the integration of big data sciences with behavioral and social sciences presents key opportunities to:

- Enhance the value of single datasets through linkages to other datasets that enable new kinds of analyses to answer different research questions.
- Incorporate computer-based methods such as computational linguistics and machine learning to measure the fidelity with which behavioral interventions are implemented.
- Link timely administrative and commercial data with survey data to advance social science research at a time of declining survey participation and rising data collection costs.
- Train behavioral and social scientists in data science (e.g., approaches for text, data, and image content organization, storage, and mapping) and data scientists in behavioral and social sciences (e.g., sources, methods, and formats of data acquisition).

The growth of data resources presents new opportunities for collaboration between behavioral and social scientists and data scientists. Realizing this potential will require innovative multidisciplinary training and educational approaches as well as ongoing communication. Contrary to pronouncements that big data may render theory obsolete, researchers increasingly need behavioral and social science models and theory to help guide the curation and analysis of big data. NIH is well positioned to lead the integration of behavioral and social sciences into big data analytics and computational health science.

Meeting Summary

Introduction

On March 19-20, 2018, behavioral, social science, computational big data, and informatics researchers participated in a workshop convened by the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) Common Fund Program. The workshop goals were to encourage researchers from these fields to engage in cross-disciplinary discussion and collaboration through a series of individual and panel presentations. Six invited experts presented papers on critical big data issues and concepts as well as case studies of big data use in behavioral and social science research. Panels of invited speakers addressed topics such as federal big data and health research, big data needs for behavioral and social sciences researchers, opportunities for incorporating behavioral and social sciences in big data research, and central data resources.

The meeting was organized by co-chairs Dr. Regina Bures, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), and Dr. Jonathan W. King, National Institute on Aging (NIA). This meeting summary provides an overview of the presentations and a thematic summary of discussions. The meeting agenda and participant list are provided in the appendices.

Welcome Remarks

Dr. Della Hann, Director of the NICHD Division of Extramural Research, explained that the Common Fund BD2K effort fosters an interesting set of relationships by including behavioral and social scientists in the NIH initiative for big data, which has been primarily focused on biomedical partners. The workshop panels were structured to elicit broad brainstorming on federal big data, training, integration, and collaboration so that the next generation of researchers finds behavioral sciences already part of the big data system. A major question concerns data resources. Collaboration and data sharing are major aims among NICHD and other Institutes and Centers, which collect vast and diverse amounts of data. For example, NICHD supports the Data and Specimen Hub (DASH), a centralized resource for researchers to store and access data from more than 40 datasets to use for secondary research. NICHD also supports the Data Sharing for Demographic Research archive, which hosts approximately 3,000 datasets from 70 different studies used by about 45,000 researchers annually. Dr. Bures' branch has been a forerunner of such efforts, spearheading a small R03 grant program that assists researchers in de-identifying their data and developing standards for pre-planning datasets for future sharing. The goal is to bridge different science traditions to benefit all researchers with a richer dataset while limiting burden.

Dr. William Riley, NIH Associate Director for Behavioral and Social Sciences Research and Director of the Office of Behavioral and Social Sciences Research (OBSSR), noted that the biomedical world has enviable big data use cases involving analyses of genetics, genomics, and health records. For example, the database of Genotypes and Phenotypes (dbGaP) includes data from various genetic data repositories and other sources and users must adhere to a data-sharing policy. Common data elements, taxonomies, and other tools support data linkages. Systems biology and neuroscience system scientists have greater traction around big data computational models than do behavioral and social scientists, who have traditionally used surveys, questionnaires, performance measures, and direct observation. However, behavioral and social scientists are experiencing an explosion of big data, including "digital detritus" and administrative datasets. The big data challenge is to parse out, clean, and organize the massive incoming data streams. Behavioral and social sciences need compelling use cases, originating from the top-down and the bottom-up, similar to the biomedical world.

Research Presentation: Creating the CenHRS

Margaret C. Levenstein, PhD, University of Michigan

The CenHRS project links the NIA-funded Health and Retirement Study (HRS)[1] data on more than 20,000 Americans over age 50 to the U.S. Census Bureau's Business Register (BR) and other survey and administrative data on the characteristics of employers and co-workers. The goal is to produce an enhanced HRS data resource that enables analyses of work-context and co-worker impacts on HRS respondents' health and well-being.

The CenHRS project illuminates important issues that arise in the context of linking and classifying big data and traditional survey data resources for social, behavioral, and health-related research. The process of linking datasets introduces a new source of uncertainty. It can generate false positives and false negatives, even with purportedly unique, direct identifiers. Careful measurement of this uncertainty allows us to draw more accurate inferences from analyses of linked data. It also allows us to retain and use information about possible matches, rather than discarding classes of records with a high probability of impossible matches. Levenstein and colleagues developed a linkage process using blocked pairs of candidate matches between BR and HRS data. Potential matches are blocked on 3-digit zip codes or 10-digit telephone numbers and classified by Jaro-Winkler string comparator scores for employer names and addresses. Humans review pairs of employers to create a training dataset. A machine learning model is estimated on the training data and then used to estimate the probability of a match between any HRS employer and any BR employer within the block. These probabilities are then used to impute a match. The imputation process is repeated 10 times so that researchers can measure the uncertainty in the imputed match (e.g., if the process imputes a linkage between the same two establishments 10 times, we have much greater confidence in the match than if it matched five times to one establishment and five times to another establishment). The machine learning model performs well and enables the retention of pairs that are more difficult to match. The CenHRS creates a new data resource that supports a wide range of research on the impacts of employment context on health, well-being, and the degree of employee attachment to the work.

Panel Discussion: Federal Big Data and Health Research

Panelists included representatives of federal producers and users of big data. They described the types of data their programs collect and use for various purposes, as well as the types of dataset linkages and innovative studies that have been accomplished with diverse federal data.

John W.R. Phillips, Social Security Administration (SSA) Office of Research, Evaluation, and Statistics (ORES), explained that ORES is the official resource for foundational SSA program data covering nearly 500 million individuals. ORES produces extracts of program data on beneficiaries and applicants to Social Security retirement and disability programs (including the Supplemental Security Income program). Researchers can request access to program data for research purposes through a centralized SSA data exchange office, with ORES serving as the business partner for the applicant. Examples of big data efforts include: (1) linking HRS data to beneficiary benefits and earnings data, creating an extraordinary source of data to examine the influence of socioeconomic factors on health. More than 200 SSA

[1] See http://hrsonline.isr.umich.edu/

extramural studies using HRS data have generated new findings; (2) regular Census Bureau updates of Survey of Income and Program Participation (SIPP) data enhance the ORES Modeling Income in the Near-Term microsimulation model. Model results are used by the Office of Management and Budget (OMB) and Congress; and (3) linked SIPP data make possible the Financial Eligibility Model exploring how changes in program eligibility factors influence social welfare. Where appropriate and within its mission, the SSA is committed to expanding access to these and other essential SSA program data.

Shari Ling, Centers for Medicare & Medicaid Services (CMS), Department of Health and Human Services, noted that CMS, which covers 1 in 3 Americans through Medicare and Medicaid, has amassed enormous amounts of data, especially administrative claims data, reflecting how people interact with the U.S. health care system. Beneficiaries are defined by age, disability status, and end-stage renal disease status, and the data are used to track beneficiaries' quality of care, health outcomes, and other information. The data may be used to inform public reporting of quality, quality improvement, coverage, and payment. CMS data can be linked to other data, as exemplified by the Surveillance, Epidemiology, and End Results (SEER) program. NIH has a memorandum of understanding with CMS that allows bi-directional data sharing, with virtual access through the Virtual Research Data Center.[2] CMS also publishes various public use files and welcomes feedback from data users about their experiences accessing these data. Ultimately, CMS intends for this data usage to contribute to the national effort to improve health outcomes.

Jeffrey Groen, U.S. Bureau of Labor Statistics (BLS), described the National Longitudinal Surveys (NLS) Program's data linkage efforts. The NLS follows three cohorts of approximately 10,000 individuals each, collecting information on many aspects of their lives, including labor market experience. Cohorts are grouped as NLSY79 (birth years 1957-1964), NLSY79 Child and Young Adult, and NLSY97 (birth years 1980-1984). In one project, the NLS program matched employers of NLSY97 respondents to BLS Quarterly Census of Employment and Wages (QCEW) microdata at the establishment level. The program is also pursuing matching NLSY data to the National Death Index (NDI) at the individual level, to Decennial Census and American Community Survey data, and to environmental pollution variables. Challenges to matching efforts relate to respondent consent, funding, data security, lack of common identifiers, and dissemination of matched data. Other matching projects are designed to compare employment and wage estimates for the nonprofit and for-profit sectors and to combine occupational microdata with injury and illness microdata at BLS.

Shelly Martinez, Office of Management and Budget (OMB) and Executive Director of the Commission on Evidence-Based Policymaking (CEP), noted that the United States has the most decentralized federal statistical system in the world. The CEP's work fits into the statistical system's goals of modernizing the traditional survey-based data tracking system and removing obstacles to matching datasets from multiple agencies. Because OMB is primarily concerned with data availability and quality, staff focus on such issues as the quality of blended or alternative data, including private-sector data created through unknown processes. The Federal Committee on Statistical Methodology and Interagency Council on Statistical Policy experts have started work on statistical standards for future data products. Martinez encouraged workshop participants to engage with the standards effort because it deals with blended datasets and characterization of individual datasets. The CEP's 2017 consensus report envisioned making the routine, efficient collection of statistical data a core part of government operations and addressed the issue of creating secure, confidential data with an understanding that microdata might later be used to link two or more datasets. Perhaps most importantly, the CEP concluded that a

[2] Available at https://www.resdac.org/cms-data/request/cms-virtual-research-data-center

functional national secure data system best practices agency is needed to help the other statistical ecosystem actors store and link data and to help states build evidence.

Charles Rothwell, National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention, noted the importance of and challenges to linking data from large surveys and datasets. Strong statistical methods enable data linkage and analyses of issues. For more than 40 years, health statistics has focused on specific diseases, but health results from the broader social milieu that is influenced by community, family, and employment. For example, NCHS linked health survey data with U.S. Department of Housing and Urban Development (HUD) data and found that blood-lead levels were higher in HUD-supported housing—an example of how one federal agency's responsibilities impact areas outside its missions. Data curation is vital and includes explaining the data collected, survey terminology, data quality and possible uses, and the linkage process, as well as enabling the replication of linkages. Different results can occur if the same datasets are linked differently. Standard protocols are needed for the data linkage process. Federal Statistical Research Data Centers are a small step forward. Also needed is a third party to conduct secure, trusted linkages for researchers. Linkages make financial and analytic sense and are predicted to be key resources for researchers in the future.

Discussion and Q&A

Participants made several points about data linkage and use:

1. NIH-funded studies rarely link to CMS or NDI data, which could provide insight into NIH study participants' later outcomes. NIH should make much greater use of linkages.
2. Ethics discussions should explore study participant expectations regarding cross-linkage of data by federal agencies, such as linking 1099R data and medical records to better understand relationships between health and pensions. The public may want government to do more with their data to improve lives; in one case, 18,000 people have signed consent forms allowing all of their medical information to be publicly shared, including their genome. However, the government must assure that sufficient public safeguards are in place. The CEP recommended radical transparency regarding potentially sensitive issues of data collection and use.
3. Systems currently exist to enable finer-grain research than is possible with public use files; a similar system may be needed for linked files, in part for educational purposes. At present, most analysts specialize in single data files and do not know how to use linked files. Because characterizing massive data from different agencies is extremely challenging and requires understanding of program context, the CEP recommendation for a national secure data system to handle multiple-agency datasets may be infeasible to implement.
4. OMB is soliciting help from big data scientists in areas such as researcher access and process improvement.

Panel Discussion: Big Data Training Needs for Behavioral and Social Sciences Researchers

Panelists represented federal and nonfederal entities with programs or activities focused on building the capacity of behavioral and social sciences researchers by training them to use and integrate big data.

Elizabeth Ginexi, OBSSR, explained that traditional data for social research have expanded exponentially to now include the Internet and many other sources of personal data that increasingly are used to study health. Big data opportunities are proliferating, but university graduate training may be inadequate to meet emerging data analytics needs. The required paradigm shift is moving slowly, as evidenced by studies in 1990,[3] 2008,[4] and 2014[5] as well as a National Academy of Sciences (NAS) conference, which demonstrate the unchanging nature of graduate education. Future Ph.D. curricula must prepare candidates for team science and multidisciplinary work and should include elements such as big data analysis, data linkages, pattern recognition, and computational modeling. At OBSSR, and NIH more generally, data sharing traditionally has been done upon request and has depended on personal relationships; now, however, NIH encourages the sharing of government-funded data where possible, some journals require data sharing, platforms that facilitate and promote data sharing are proliferating,[6] and knowledge is regarded as a community enterprise. No single data-sharing model exists, which complicates training of students to access data in diverse formats. Reconciling terminology, ethics, privacy levels, and other issues will require dialogue. The NAS highlighted an array of new data mining tools that are not part of many graduate school programs but whose use could be encouraged by innovative funding agency incentives.

Valerie Florance, National Library of Medicine (NLM), noted that two data science training models, the university-based NLM and the BD2K predoctoral training programs, have unique prerequisites and goals. NLM focuses on research careers, while BD2K focuses on more traditional NIH training. The content possibilities for core curricula in behavioral and social sciences are diverse, as revealed by an examination of different existing programs. They include, for example, core concept training, such as biomedical and health informatics and human-computer interaction; electives, such as systematic review procedures and meta-analysis; and team science, such as work in clinical care settings. Content for BD2K core curricula would have unique core concept, elective, and team science elements. Behavioral and social scientists need knowledge about data science (e.g., approaches for text, data and image content organization, storage and mapping), and data scientists need knowledge about social sciences (e.g., sources, methods, formats of data acquisition). Both sciences need additional skills in key areas (e.g., basics of knowledge representation and management). Free resources are available to train social scientists, such as Massive Open Online Courses (MOOCs) and open educational resources.

Vasant Honavar, Pennsylvania State University, emphasized that an application domain is needed when training data scientists. Data science tools and education, which typically focus on such areas as data collection, organization analysis, and sharing, should be expanded to include algorithmic thinking to create abstractions of scientific discourse domains and reproducible data science practices. Communication with domain scientists is necessary, and the gap between the language of specific scientific domains and machine learning methods must be closed. In the era of big data, data scientists must understand the sources of data and their limitations (e.g., sampling bias) to be able to draw valid conclusions. Machine learning methods must extend beyond black box models to offer explainable models. Integrating diverse data types is difficult enough, but data scientists will also need to integrate models across different levels of abstraction and to link them to different types of data at different temporal and spatial scales. Data scientists need to be trained in data ethics, data privacy, algorithmic bias, fairness, and related issues. Some have argued that petabytes enable analysis of data for correlations and that science can advance without hypotheses, models, or theories, but nothing could

[3] Aiken, L. S., West, S. G., Sechrest, L, & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist*, 63(1), 32-50.

[4] Aiken, L. S., West, S. G., et al. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, V63(1), 32-50.

[5] Kaplan, R. M., Riley, W. T., & Mabry, P. L. (2014). News from the NIH: leveraging big data in the behavioral sciences. *Translational Behavioral Medicine*, 4(3), 229-231.

[6] See, for example, https://osf.io/

be further from truth. Data science training must offer more than a facility to find correlations. Training at the interface of data sciences and social sciences should produce a new generation of scientists who are well-versed in both state-of-the-art data science methods and tools (including their relative strengths and limitations) as well as grounded in social and behavioral sciences to be able to harness the power of data to advance social and behavioral sciences. Here, joint mentoring of students by faculty with complementary expertise in data sciences and social and behavioral sciences—experiences that help develop the ability to effectively communicate across disciplinary boundaries—can be extremely beneficial. The training efforts can benefit from availability and sharing of data, software, and course materials.

Cheryl Eavey, National Science Foundation (NSF), reported on NSF's activities related to big data training for behavioral and social sciences researchers and the large future potential in this area. As examples, NSF has funded a big data project on an integrative education and research program in social data analytics and two workshops on methods and infrastructure for scalable computing in neuroscience. For the future, NSF has identified "Harnessing the Data Revolution" as a big 10 long-term research and process idea for future investment at the frontiers of science and engineering. Education, cyber-infrastructure, algorithms, and tools are encompassed by this big-10 idea, though how education will manifest itself as a priority is still to be determined.[7]

Discussion and Q&A

The panelists' remarks revealed several salient themes: (1) the importance of interdisciplinary team science; (2) the need to infuse data science into domain-specific scientific inquiries; and (3) the need to not only focus on exposure, skills, and training as big data's importance to the field of behavioral and social sciences grows, but also to avoid repeating the past response to shape the field as data science evolves. One participant noted that computational biology evolved over 30 years through dialog between computer science specialists and biologists. A similar development could occur if specialists in one field left their comfort zones, eventually creating the next generation of scientists skilled in working at the nexus of big data and social sciences. Another participant stated that a model exists in funding agencies that have supported multidisciplinary groups of investigators working on interesting problems and being, to some degree, required to cross train. An example is the NSF decision-making under uncertainty projects. Although these projects were not about data science, their large scale and diversity of disciplines led to participants learning about science outside their specialization. The challenge is to determine the right placement for scientists with cross-disciplinary skills.

A participant asked about on-the-job training, both for data scientists and people who work with them. Previously, small funding amounts were made available as supplements to add information scientists to ongoing projects, a possible model because many researchers need advanced data management skills. NIH has allowed researchers to add supplementary data scientists, but too often the data experts were marginalized, so the integration process must be refined. NIH is specifically interested in use cases from domain areas. Behavioral and social sciences researchers could define important problems to solve, without necessarily knowing how, and draw big data and computer scientists' interest through a discussion of the challenges. Data sciences often focus on prediction and computation, which are not necessarily core to social sciences; therefore, the answer is not simply to send social scientists to computer classes. However, conversations about the potential for data sciences to transform research are ongoing, although social scientists struggle with when and how to use machine learning to deal with

[7] New opportunities are posted on the NSF website at https://www.nsf.gov/news/news_summ.jsp?cntn_id=244678&org=CI

correlation versus causality. Even the machine learning community is moving toward explainable models. In prevention research, the focus is less on explanation than on determination of who will respond well to an intervention. There is great promise for behavioral and social sciences to benefit from big data computational capabilities to personalize research in new experimental domains that can define for whom and why treatments work, moving beyond explanations to prediction and possibly optimization. Communication is an essential and time-consuming factor when merging two different fields.

Panel Discussion: Opportunities for Incorporating Behavioral and Social Sciences in Big Data Research

Panelists provided various perspectives on the multiple opportunities for applying and using big data and big data methods in behavioral and social sciences research.

Brian Athey, University of Michigan and co-director of the Michigan Institute for Data Science (MIDAS), noted that major opportunities to incorporate behavioral and social sciences in big data research exist. However, a 2009 article in *Science* called for computational social science that has not yet occurred. When MIDAS issued a call for participants in computational social science teams, responses were scant, and Athey has struggled to define what the field needs. The 2009 article noted that ubiquitous digital data are created as users log on to social media and other platforms. With bioinformatics, large amounts of data were consolidated as if created at one time, but today the temporal and longitudinal aspects of network data can be parsed, and numerous analytical methods applied. Knowledge about the mathematics of networks and natural language and speed processing could be applied to create computational social science, which is a goal that should not be abandoned. Regarding data science education, statisticians can become computer scientists, which is at the core of data science. Important opportunities exist to create new knowledge, for example with research into the epidemic of “despair syndrome” in the United States involving co-morbid addiction, liver disease, and depression. U.S. incomes have doubled since 1972, but happiness indices have either stagnated or declined, with data showing correlations between obesity, opioids, and depression. New computational social sciences methods could increase understanding of such phenomena using cellphone and other data.

Wendy Nilsen, National Science Foundation, remarked on the unimaginably large amount of data generated through YouTube and other platforms, which creates unprecedented opportunities to understand behavior. She manages many projects that deal with image, sound, and sensor data, which are all major computational areas. Data scientists deal extensively with the behavior of interest to behavioral and social scientists, but they lack understanding of basic social science concepts and the associated literature. One NSF-funded big data study collects real-time data on Family Eating Dynamics using in-home sensors and other technologies. Another studies dementia using body-worn and in-home sensors to model the relationship between agitation and the environmental factors. Smart communities present opportunities to effectively use big data and behavioral sciences for planning. Using a large amount of observational data to make causal inference presents significant opportunities as well as challenges, in areas such as data quality, data validity, data acquisition, scalable performance metrics, and privacy and security. Collaboration is a key to obtaining the best research and must be part of training.

John Eltinge, U.S. Census Bureau, presented an idealized regression model analysis as a basis for discussing inferential nuances, microdata, and bias, which should not be feared but managed. All of the workshop topics could be regarded as a variation on the simple regression model structure involving an outcome, predictors, and coefficients. In some cases, an outcome variable may be a distillation of very complex sources of information, such as temporal features or information extracted from images. Historically, design data collection, such as sample surveys, prevailed, focusing attention on sources of variability, which remains necessary when working with today's non-design data. A threshold question centers on how non-design data should be handled once obtained and depends on a researcher's inferential goals. Inferential nuances have been an issue for decades, such as in John Tukey's studies in which either a rigorous inference or a predictive outcome was the goal, with careful uncertainty analysis. An alternative approach that is implicit in big data discussions does not seek rigorous proof of causality but explores indications of what might explain a phenomenon and therefore warrant further study. That distinction has important implications for training or for deeper education. In the evolving discussion, behavioral and social sciences can contribute to cases that entail an understanding of the underlying phenomena, such as despair syndrome, and that highlight the need to assess the quality of measures and to define the targeted predictors. Issues of transparency, reproducibility, and replicability—and their alignment with specific cases of data variability—must be considered, such as, for example, when dealing with proxy variables. Microdata quality also must be considered. Over 30 years, a wealth of studies has centered on cognitive aspects of survey methodology (e.g., individual reactions to data collection instrument designs), and the same issues can be extended into the big data arena to understand the social processes affecting big data observations.

Paul Beatty, U.S. Census Bureau, although excited by big data's game-changing possibilities for enhancing analyses, powering studies, and reducing costs, highlighted methodological caveats that can be easily overlooked. Declines in survey participation, rising costs, and other issues make using big data potentially attractive. For example, records of immunizations are more reliable than self-reports of immunizations. However, administrative records also pose problems, such as the need for respondents' permission to access records, missing or incorrect information, and insufficient volume for some purposes. An appeal of big data is its potential to transcend the limitations of typical survey analyses, offering greater linkages, filling survey gaps, and perhaps replacing some surveys altogether. However, measurement error remains an issue, because the extent, causes, and statistical consequences are less well known. Generally, the issues relate to validity, missing data, and representation, and various types of error can be found in individual data systems, such as inaccuracies during data curation. Additional obstacles can arise when datasets with different data sources, production processes, and statistical properties are combined. These and other issues regarding the use of non-design data and multiple data systems deserve consideration. To avoid being misled by uncritical acceptance of big data, we can apply the tools used to understand survey measurement error to big data.

Carlos Gallo, Northwestern University, as a computational social scientist, is focused on how best to implement proven evidence-based programs (EBPs) using big data that target outcomes on preventing HIV infection, drug abuse, and other health-related outcomes. To sustain efficacy, effectiveness, and effect sizes, the fidelity of EBP implementation must be assessed and modifications to as-designed interventions must be tracked. In the early stages of intervention development, independent observers can assess fidelity. However, this approach is not feasible as EBPs are scaled up for use by local agencies. Currently, approximately 10 percent of fidelity assessments are conducted by human raters, a procedure that involves expensive micro-coding. Also, with older assessment tools, feedback can occur weeks later, and personnel to deploy these tools are limited. The challenge is to design a system that can monitor 100 percent of sessions so that feedback can be provided to funding agencies, supervisors, and individuals who deliver interventions. Automated systems based on computational linguistics and machine learning, such as essay graders, can provide instant feedback. With valid computer-based measuring methods, 100 percent of sessions can be checked, with thousands of ratings for each providing a “dashboard” of insights into a session’s high- or low-fidelity, including problematic minutes within an otherwise high-quality session. The goal is not to replace humans but to enhance their capability to implement various interventions using computational linguistics and other tools.

Discussion and Q&A

A participant asked about the difference between the predictions that Google seeks to make with big data and the knowledge that behavioral and social scientists seek to gain with big data. Several participants responded that there is likely extensive overlap: Google and behavioral scientists seek to change behaviors, although of differing kinds. Google runs thousands of randomized controlled trials (RCTs) every day. Both Google and behavioral scientists seek to know what works and to assess the effectiveness of a given tool for solving a defined problem.

Research Presentation: Behavioral and Social Science Insights for Big Data Research

Barbara Entwisle, PhD, University of North Carolina at Chapel Hill

Big data presents both opportunities and challenges for innovative behavioral and social science. Big data provides incredible precision in time and space when studying, for example, how the opportunities and constraints on behavior in a neighborhood affect health. Most studies operationalize neighborhoods as census tracts. In contrast, focusing on a neighborhood as an activity space over a day or week, as well as the timing of the activity, is closer to reality because most people spend time outside of their census tract. In conceptualizing neighborhoods as activity spaces, researchers can align a dynamic map created using GPS-enabled smart phones with a detailed accounting of activities over a given period. Besides drilling down into detailed neighborhood studies, researchers can expand their studies to a national and historical scope using big data and simulation software to understand trends emerging over long time periods. However, a significant issue has emerged as data collection has evolved from collection using designed surveys to collection involving “organic” big data developed for purposes other than research. The mechanism used to generate data determines who or what is potentially captured in the data and, critically, the mechanism is often a behavior that constrains the generality of the results. While it is well known that convenience samples affect external validity, it is less recognized that the data-generating mechanism can bias results if it inappropriately conditions on common effects of the outcome of interest *and* one or more key predictors. This is known as “collider variable bias.” It is also well known that social variables, such as education and income, affect how data are generated, but health outcomes can also have an effect, as shown by studies on volunteering. Collider variable bias is thus a potential concern. Greater understanding of data-generating mechanisms is needed and should be incorporated into the analysis of socially produced big data to advance big data science.

A brief question and answer follow-up session confirmed that “collider variable bias” is also called “endogenous selection bias.” Participants agreed that so-called found data, or big data, have an inherent susceptibility to bias as people self-select into populations from which data are generated. Behavioral and social scientists can inform understanding of the issue of causal inference, which is not limited to big data, because there are questions regarding who volunteers for RCTs. Bias poses a significant problem for trials, especially behavioral therapy trials, which depend on the same characteristics that affect whether people are in the trial group at all.

Panel Discussion: Central Data Resources

Panelists represented organizations in the process of building, maintaining, and sharing big data resources for research.

Stan Ahalt, Renaissance Computing Institute (RENCI), described a large effort funded by the NIH Data Commons that seeks to provide scientists with a platform to conduct research as if they had Google's resources, but he underscored a workshop theme that caution is necessary when combining massive amounts of data and inferring meaning. The project goal is to understand how knowledge sources can be used, typically to map associations between genes and diseases or a cell and a pathway. The National Center for Advancing Translational Sciences (NCATS), for example, seeks to build the infrastructure to enable researchers to someday answer very difficult questions about drug interactions and other matters. The team names (e.g., Team Helium, Team Argon) emphasize the fact that the researchers come from diverse institutions but are equals conducting team science. The teams' efforts will be interlinked to build data tools and an environment for discovery and to scale research across very large datasets. Guidelines ensure compliance with Findable, Accessible, Interoperable, and Reusable (FAIR) and other standards, especially the requirement for a cloud-agnostic architecture. Complex interacting systems, entirely virtualized, will support scientists to work faster, with ethics, privacy, and security woven throughout to avoid bias. Scientists will be to share the "apps" they develop. Machine data translation and reasoning are essential to this process.

Amy Rose, Oak Ridge National Laboratory (ORNL), discussed the opportunities presented by spatial big data, noting that such data are often not considered in big data discussions. As a starting point, spatial big data can be used to understand what exists in places, and where, as a step toward understanding the populations in areas. Satellite data, which are unstructured as to social phenomena, are an example. ORNL combines satellite images and other diverse data sources to create global population estimates. The Laboratory is trying to understand new development, which is not well defined as the world changes minute by minute. ORNL receives increasingly high-resolution imagery requiring massive computer processing and storage capacity to identify where people live. Stakeholders use this information to make vaccination deliveries more effective. In some regions, massive population migrations are occurring, which has implications for a global census. Neighborhood mapping is a critical piece of information that can help to identify patterns in spatial areas that can serve as indicators of socio-economic strength, energy use, and other variables important to understanding the people who live in areas and their activities. Machines, which might classify spatial imagery differently than humans, can highlight information that otherwise could be overlooked—for example, differences in formal and informal settlements. Thus, spatial big data can produce basic understanding of population location issues.

Susan Tenney, Booz Allen Hamilton, described the NICHD-funded Data and Specimen Hub (DASH), a centralized resource for NICHD-supported researchers to store and access de-identified data with the goal of facilitating data sharing and accelerating scientific findings to improve human health. Launched in 2015, DASH has more than 12,000 users from 102 countries and houses 65 completed studies spanning 27 diverse study topics. There have been 73 data requests and three publications to-date, with six to eight manuscripts in progress. DASH's design is based on FAIR principles is completely meta-data driven. The system features a study overview page so that users can quickly glimpse the study characteristics and decide whether to dig deeper. DASH has overcome various data-sharing challenges. Tenney stated that to ask the right questions, the right data are needed, and Barbara Entwisle's paper ("Sample Generation Mechanisms and the Potential for Bias in 'Big Data'") explained that bias can impact all areas of public health, especially precision medicine. DASH and other centralized resources can help to minimize bias in the areas of study design, data quality, and data analysis. For example, to enhance scientific rigor, best practices relating to study methodologies could be captured over time by centralized data repositories. Although bias cannot be wholly eliminated, it can be managed through powering studies with pooled data. Despite the challenges, it is worth considering a transition from big data to open data to facilitate robust data analysis.

Alastair Thomson, National Heart, Lung, and Blood Institute and the NIH Data Commons, discussed what the Data Commons could mean to workshop participants and provided an overview of the datasets. NIH is creating massive amounts of data, especially genomic data but also other data such as imaging data. The massive data amounts are a barrier to use, making the cloud, with its attendant costs, the only viable approach. In recognition of this issue, NIH funded pilot projects to build a unified data and computational platform in the cloud. DASH and other initiatives are limited because they are silos of excellence aimed at solving a particular community's problems. Combining datasets, including spatial data, requires interoperability, which is a core FAIR principle. The Data Commons seeks to rapidly expand the data, including behavioral data, and make it available in one place. But true interoperability in a cloud environment is a complex IT project that poses significant technical challenges. The government, however, is not good at large, complex IT projects, agility, and real innovation—all capabilities required by the Data Commons. Traditional grants and contracts will not suffice, so a new mechanism for funding—the Other Transaction Research Opportunity Announcement—was adopted for the initiative, allowing real decision-making and agile experimentation and producing greater collaboration across a global consortium. The investment is critical to both data science and all of the kinds of science that NIH funds; increasingly, science is data science. The initiative explores the value of NIH data, bringing in commercial groups to assess what they might leverage from the data, thereby introducing a profit motive that could support the data's cost sustainability. The commercial aspects raise ethical issues regarding returning health-related data to individuals if risks are found and permissions for using data in studies. Machine learning, with open-ended rather than hypothesis-driven data queries, also raises complex issues of consent and possible bias. It is hoped that behavioral and social scientists join the consortium as it grows.

Discussion and Q&A

Participants inquired about how data of diverse temporal scales can be combined to obtain insights into the speed at which people's environments change. The challenges are difficult because few tools are available for involved space-time analytical work, which is an issue that extends beyond big data to big methods. Typically, spatial- and temporal-change data are handled separately, then combined to understand changes on the ground. A participant asked what behavioral and social scientists can contribute to solving the big data management issues. One concern is communication with faculty to spur collaboration, especially when crossing domains. At ORNL, for example, collaborations with social scientists across big engineering projects is beneficial, and workshops such as this one expose individuals to ideas, datasets, and projects that they typically do not encounter in their circles. The hope is that universities will soon be able to encourage faculty to engage with the Data Commons initiative in its early stage because real scientific cases will be needed to make the system work. Social scientists, including economists, have already contributed to big data–related efforts, and those contributions should be recognized and understood. For example, confidentiality rules have been developed. Centralized data systems will be helpful to train junior investigators, who could explore hypotheses without having to collect data. Over the past 20 years, federal statistical agencies have developed a federal statistical center network in partnership with academic colleagues. Some dovetailing of federal datasets and university researchers may already be occurring.

Research Presentation: Big Data Health Research in Vulnerable Populations

Jay Bhattacharya, MD, PhD, Stanford University

A key distinction in defining big data exists between inductive versus deductive reasoning that, if ignored, leads to confusion about the standards of big data analysis. Bhattacharya applied this distinction to two examples: p-hacking, or data fishing, and identifying vulnerable populations using big data–like methods. To alleviate this confusion, it is important to consider why statistical goodness of fit for modeling is disparaged in econometrics but deemed essential for model selection in machine learning and decision-making. Big data analysis consists of two types of activities: (1) machine learning and statistical decision-making and (2) testing for causal relationships in the data, as with econometrics. In science, deductive reasoning is mainly concerned with hypothesis testing, while inductive reasoning is mainly concerned with theory formation. Bhattacharya offered three propositions. First, big data is primarily inductive, striving to create new models and theories that make accurate predictions; by implication, then, goodness of fit is a key criterion to judge whether a big data model is performing well. Second, big data involves treating the data as if they were "the population" rather than a sample, making statistical inference secondary; as a result, big data methods avoid over-fitting models. Third, following from the first two propositions, big data methods and causal statistical analysis are complements. Analysts lacking clarity about whether they are engaging in the inductive or deductive analytical mode will mistakenly apply one mode's rules to the other. In one of his applied examples, Bhattacharya used big data methods to differentiate new and existing Supplemental Nutrition Assistance Program (SNAP) recipients. The analysis found that new SNAP recipients generally share more demographic similarities with existing SNAP recipients than with never-SNAP recipients but, comparatively, are thinner and more food secure. Overall, he concluded that big data methods formalize inductive thinking, can conditionally co-exist with econometric methods, and are ideal for identifying vulnerable populations.

Research Presentation: Big Data, Big Models, Uncertainty, and Bias—Data Collection in Development

Tyler McCormick, PhD, University of Washington

In most lower- and middle-income countries, no comprehensive registration system of births and deaths exists. Limitations of big data—including that it is not intended for research and that researchers do not control the design mechanism—raise statistical issues regardless of data size. Data scientists are developing exciting new methods and software, and their contributions to enabling science should be acknowledged. Individuals must be trained to ask appropriate questions as they use the tools in the scientific process. Themes in the context of collecting data in developing countries include the related issues of data combination and pooling, targeted sampling, and opportunistic data collection. Ethics and preserving privacy are important open challenges for all methods used in big data demography. Another issue is the lack of calibration data needed to correct for known sources of bias. Big data models pose extrapolation challenges pertaining to statistical frameworks and decision-making. Education, training, and transparency must be addressed. Such themes apply to specific studies, such as the world population project that combined data to understand mobility, poverty, and other measures, and a study that used cellphone data to infer wealth and make extrapolations. Spatially and temporally localized estimates are necessary for policy and resource allocation, but available data do not typically support that level of resolution. Spatial and temporal smoothing, while tempting, create a fundamentally different kind of generalization uncertainty than the typical sampling uncertainty taught in most statistics or biostatistics programs. Communicating and incorporating such uncertainty into decisions must be discussed. Underscoring the point, a study with gold standard data on 7,841 adult deaths from six cities did not provide a case with very high extrapolation error, which serves as a cautionary tale regarding data inputs used to extrapolate from one context to another.

Discussion and Q&A

Discussant Michael Rendall, University of Maryland, commented that social statisticians have been anticipating the arrival of big data for some time. The two main challenges—population representativeness and data not designed for research—can be addressed through methods that combine traditional and big data sources, including statistical calibration methods. As for the move from the source of estimation uncertainty from sampling error to model assumptions, that issue has been a statistical method topic that has received considerable attention and should not be underestimated as a challenge. Dr. Bhattacharya’s study using “big” administrative SNAP data with “small” but representative survey SNAP data (in this case, from the Panel Study of Income Dynamics) to calibrate the administrative data to target populations is an exemplar of this challenge. The sampling error in the “small” survey data will contribute to estimation error as a potentially substantial source of overall uncertainty. Big data’s arrival presents wonderful opportunities for social scientists to improve the accuracy and temporal relevance of their work, greatly reducing the limitations created by survey sampling error and processing times. Survey samples are already outdated by the time they are completed; big data, which leverages survey sample results, can provide more up-to-date perspectives. Bayesian methods are critically important for bringing together social and statistical science with big data because, unlike frequentist methods, they allow for multiple uncertainty sources. Big data challenges have been present for longer than recognized and have been managed with data-combining methods, as exemplified by graphs of Hispanic total fertility rates between 1991-2001 that reveal the imperfection of population data deemed perfect. That example and others demonstrate that social scientists have calibrated “big” administrative data by using various kinds of population-representative survey data and statistical combining methods.

The discussion focused on the population sample provided by big data, which Google collects for trends and Amazon collects for its users. The key issues are whether the populations sampled are the ones that researchers care about and how to understand the relationship between the populations that Google and Amazon care about and those that researchers care about. Most analyses presume possession of data on their target population, but sampling error standard protocols do not apply in these cases.

Research Presentation: Systems Science and Data Science

Elizabeth Bruch, PhD, University of Michigan

Systems science needs data science to make strong inferences and vice versa. Over the past decade, two mostly disparate developments have occurred in public health: (1) the rise of data science (i.e., big data, machine learning) and (2) the rise of systems science (i.e., a framework for considering health challenges, mechanisms to explicitly capture feedback and interdependent behavior). Faced with complex problems that are difficult to solve with targeted policy interventions, public health officials’ interest in systems science has grown. Diverse examples illustrate that feedback effects make health problems difficult, such as, for example, low-fat foods leading people to eat larger portions and thereby contributing to obesity. Several methods have become prominent in public health for studying complex problems, such as agent-based modeling that simulates how individuals interact with their environments and each other. However, using systems science in policymaking is challenging; for example, “deep parameters” of behavior and structure are needed to enable extrapolation. Data science techniques and big data can help overcome the challenges. Regarding the challenge of model complexity, data science might provide hybrid models that combine mechanistic formal models with machine learning's predictive power or ensemble models that help analysts to rigorously incorporate model uncertainty and overcome knowledge gaps. Big data can assist development of higher quality model representations of social systems. As for how systems science can benefit data science, data science needs formal modeling to make inferences about how to impact nonstationary social systems.

Research Presentation: Obesity and Big Data

Bruce Y. Lee, MD, MBA, Johns Hopkins University

Obesity provides an example of how the combination of big data and systems science can help inform understanding of and address a public health problem. Big data is not simply large volumes of data. It has unique features that can either generate new opportunities and insights or lead to misinformation and misunderstandings, depending on how it is handled, analyzed, interpreted, and used. Obesity is a complex system issue that includes the interactions of diet, physical activity, metabolism, and chronic disease with many different complex biological, behavioral, social, cultural, economic, and other systems and subsystems. Many public health problems are symptoms of broken systems (e.g., lack of access to healthful food), and failure to treat them as such often leads to band aids rather than sustainable solutions, missing secondary and tertiary effects, unintended consequences, and wasted time, effort, and resources. The key is whether big data can be properly harnessed to inform understanding of complex systems rather than simply offering more correlations and associations that may be misleading.

The proper use of big data includes understanding the relative strengths and weaknesses of different sources (which can range from machine logs to sensors to social media to archives of text documents) and different methods of transmitting, storing, managing, and analyzing big data. In general, big data analytic methods can be divided into "top down" (analogous to hovering over the data from above and searching for patterns that give some insight as to structure and function of the system) versus "bottom up" (attempts to build mechanistic representations of the systems of interest and then use big data to further develop this representation) approaches. While "top down" approaches start with the data and then use that data to discover patterns, associations, and possible relationships, "bottom-up" approaches (otherwise known as systems approaches) instead start with a representation of the actual components, structure, and mechanisms that comprise the system or systems of interest. With the latter, big data serves a different supporting purpose rather than being the origin of the insight. Big data can help refine and further develop a structure and mechanisms that have already been established.

Systems mapping and systems modeling are examples of "bottom up" or systems approaches. An example are the Virtual Population Obesity Prevention simulation models of various cities, developed by the Global Obesity Prevention Center (GOPC). These agent-based models include geospatially explicit representations of all the people, food sources, households, schools, parks, and other locations in a city such as Baltimore. The GOPC has used various traditional data and big data sources to populate these models and to answer questions such as "what is the impact of increasing children's physical activity, reducing crime, or implementing sugar-sweetened beverage warning labels on obesity and various health outcomes and costs?"

Discussion and Q&A

Discussant Ross Hammond, Brookings Institution, noted the high potential for gains resulting from the intersection of data science and systems science, especially for addressing health disparities. Many developments are already occurring; however, when the intersection is examined seriously, issues arise that must be addressed, such as sample bias and collider variables Even if these issues can be resolved, there are others for which answers remain elusive. Three questions pertain to agent-based modeling that use longitudinal and personal big data:

1. Exactly how granular must the data be for any given model? Potentially data could be collected on the time scale of seconds or, conversely, at an intergenerational scale. Work with in-person physiology models in the context of obesity demonstrates that small, good datasets, such as cohort studies, are extremely sensitive to the granularity frame chosen for the data. Ecosystem modelers confront the same issues and have some promising answers.
2. How extensive must a model be? That is, while data are available on people's locations, what else needs to be known? While people's behavior in a location would likely be of interest, is it necessary to know about physiologically relevant variables, psychology, or the social structure context? In what cases might such information be needed or not needed?
3. Regarding the tradeoff between data accessibility and privacy issues, often the data exist in a carefully controlled computational ecosystem; the models, however, are often quantitative computer programs written in non-standard languages, in proprietary code, with high processor demands, and therefore cannot be run on a server where the data are stored. Enabling data to be used on models while preserving privacy is a significant issue. Additionally, in work by policymakers and stakeholders, often at a local level, the demand for location-specific visualization raises even greater privacy challenges.

Many of these issues will not be solved one investigator at a time; rather, they demand centralized solutions, which federal funding agencies are positioned to provide. In addition, mechanism-focused models that communicate the data's origins and capture the dynamic processes of interest are greatly needed. Simply collecting data without defining the need and purpose is not optimal. Google's pronouncement that theory is dead is mistaken. Clearly defined questions and goals are necessary when working with the vast data pools created using big data mechanisms, and different standards are needed for policy as opposed to etiology modeling.

Participant discussion focused on the use of models to optimizing policies and interventions. Optimization is a difficult issue because there is not necessarily one optimal solution; rather, models can be used to explore many different scenarios to better understand different relationships and identify the key drivers, which can serve as the targets for data collection and intervention. The goal is not to always select the optimal policy or intervention for a specific situation but instead to identify policies and interventions that would be robust across a spectrum of situations. Caveats notwithstanding about models not serving as crystal balls, policy analysts nevertheless need deep confidence in the credibility of their inferences, a goal that can be supported by ensemble modeling approaches.

The relationship between data mining and data modeling is also of interest. Data mining, which is an inductive approach, seeks to extract actionable insights from data; revealed patterns can later be explored for mechanistic explanations. Systems models, in which sensitivity analyses play a large part, can help guide data mining by delineating the most relevant parameters driving various phenomena. If the purpose is to intervene in a system, understanding the mechanisms at play is important, as in disease and health research where understanding mechanisms of adaptation is vital. Social scientists emphasize conceptual models that inform the computational models that they run. A participant expressed interest in whether big data and computational tools will now accelerate the speed at which answers to research questions are provided and how speculative those answers will be. Analyses using different random samples of datasets to avoid overfitting demonstrated that each sample highlighted different variables as being important. Simple data mining is unlikely to accurately extract the significant mechanistic factors. Faster answers are unlikely in today's environment because the speed is limited by researchers' ability to think, but new tools will reveal answers to previously unanswerable questions. Models and data collection will evolve through a cyclical iterative process leading to more targeted data collection and additional mechanistic insights, with both data collection and modeling improving together.

Closing Remarks

Meeting co-chair Regina Bures emphasized the key workshop themes, which included (1) data linkage processes and issues, exemplified by federal big data providers; (2) big data training needs, especially multidisciplinary training with a substantive domain focus to integrate data science into more scientific disciplines; (3) the language and communication challenges between big data and social and behavioral scientists, which are not insurmountable if specialists make the extra effort; (4) the potential unique sources of bias with big data, including demographic, population data of interest to social scientists; (5) the underlying interoperability challenges arising from central data resources, such as efforts to create more interoperability between medical science big data and behavioral and social sciences resources; and (6) the issue of modeling and the continuing importance of mechanistic theory in using big data. Workshop participants demonstrated their willingness to look beyond their disciplinary boundaries and to engage with each other on the immense opportunities and interesting challenges presented by big data and the interlinking of data science and behavioral and social sciences.

Appendix 1: Agenda

Revised March 15, 2018

Monday, March 19, 2018

HANDS-ON TRAINING WORKSHOP: R and RSTUDIO FOR REPRODUCIBLE SCIENTIFIC ANALYSIS

8:30 a.m.	Introduction to RStudio and R Data Types	Doug Joubert
9:45	**BREAK**	
10:00	Data Wrangling with R	Christopher Belter
12:00 p.m.	**BREAK**	

BEHAVIORAL AND SOCIAL SCIENCES AND BIG DATA WORKSHOP

1:15	**Welcome, Introductions, and Purpose**	Della Hann, NICHD William Riley, OBSSR
1:45	**Research Presentation**: Creating the CenHRS: A New Method for Probabilistic Linkage of Employers in Survey and Administrative Data	Margaret Levenstein
2:15	**Panel Discussion**: Federal Big Data and Health Research *Panel Members* John W. R. Phillips, Social Security Administration Shari Ling, Centers for Medicare & Medicaid Services Jeffrey Groen, U.S. Bureau of Labor Statistics Shelly Martinez, Office of Management and Budget Charles Rothwell, Centers for Disease Control and Prevention	Moderator: John Haaga, NIA
3:30	Discussion and Q&A	
3:45	**BREAK**	
4:00	**Panel Discussion**: Big Data Training Needs for BSS Researchers *Panel Members* Elizabeth Ginexi, Office of Behavioral and Social Sciences Research Valerie Florance, National Library of Medicine Vasant Honavar, Penn State University Cheryl Eavey, National Science Foundation	Moderator: Christine Hunter, OBSSR
5:00	**ADJOURN**	

Tuesday, March 20, 2018

9:00 a.m.	**Panel Discussion**: Opportunities for Incorporating Behavioral and Social Sciences in Big Data Research	Moderator Robert Carter, NIAMS
	Panel Members Brian Athey, University of Michigan Wendy Nilsen, National Science Foundation John Eltinge, U.S. Census Bureau Paul Beatty, U.S. Census Bureau Carlos Gallo, Northwestern University	
10:15	**Research Presentation**: Behavioral and Social Science Insights for Big Data Research	Barbara Entwisle
10:45	**Panel Discussion**: Central Data Resources	Moderator: Elaine Collier, NCATS
	Panel Members Stan Ahalt, Renaissance Computing Institute Amy Rose, Oak Ridge National Laboratory Susan Tenney, Booz Allen Hamilton Alastair Thomson, NIH Data Commons	
11:45	Discussion and Q&A	
12 p.m.	**LUNCH**	
12:45	**Research Presentation**: Big Data and Health Research in Vulnerable Populations	Jay Bhattacharya
1:15	**Research Presentation**: Big Data, Big Models, Uncertainty, and Bias: Data Collection in Development	Tyler McCormick
1:45	**Discussant**: Michael Rendall, University of Michigan	
2:00	Discussion and Q&A	
2:15	**BREAK**	
2:30	**Research Presentation**: Systems Science and Data Science	Elizabeth Bruch
3:00	**Research Presentation**: Obesity and Big Data	Bruce Y. Lee
3:30	**Discussant**: Ross Hammond, Brookings Institution	
3:45	Discussion and Q&A	
4:00	**Closing Remarks**	Regina Bures, NICHD
4:15	**ADJOURN**	

Appendix 2: Participant List

Revised March 20, 2018

Invited Speakers

Stan Ahalt, University of North Carolina at Chapel Hill
Brian Athey, University of Michigan
Paul Beatty, U.S. Census Bureau
Christopher Belter, Office of Research Services, NIH
Jay Bhattacharya, Stanford University
Elizabeth Bruch, University of Michigan
Robert Carter, National Institute of Arthritis and Musculoskeletal and Skin Diseases
Elaine Collier, National Center for Advancing Translational Sciences
Cheryl Eavey, National Science Foundation
John Eltinge, U.S. Census Bureau
Barbara Entwisle, University of North Carolina at Chapel Hill
Valerie Florance, National Library of Medicine
Carlos Gallo, Northwestern University
Elizabeth Ginexi, Office of Behavioral and Social Sciences Research, NIH
Jeffrey Groen, U.S. Bureau of Labor Statistics
John Haaga, National Institute on Aging
Ross Hammond, Brookings Institution
Della Hann, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Vasant Honavar, Pennsylvania State University
Christine Hunter, Office of Behavioral and Social Sciences Research, NIH
Doug Joubert, Office of Research Services, NIH
Bruce Y. Lee, Johns Hopkins University
Margaret Levenstein, University of Michigan
Shari Ling, Centers for Medicare & Medicaid Services
Shelly Martinez, Office of Management and Budget
Tyler McCormick, University of Washington
Wendy Nilsen, National Science Foundation
Candace Norton, Office of Research Services, NIH
Michael Rendall, University of Maryland, College Park
William Riley, Office of Behavioral and Social Sciences Research, NIH
Amy Rose, Oak Ridge National Laboratory
Charles Rothwell, National Center for Health Statistics, Centers for Disease Control and Prevention
Susan Tenney, Booz Allen Hamilton
Alastair Thomson, National Heart, Lung, and Blood Institute

Participants

Farheen Akbar, Office of Behavioral and Social Sciences Research, NIH
Ruben Alvarez, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Julia Berzhanskaya, National Institute on Drug Abuse
Partha Bhattacharyya, National Institute on Aging
Gregory Bloss, National Institute on Alcohol Abuse and Alcoholism
Regina Bures, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Minki Chatterji, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development

Juanita Chinn, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Rebecca Clark, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Elan Cohen, National Institute of Mental Health
Richard Conroy, Office of Strategic Coordination, NIH
Leslie Derr, Office of Strategic Coordination, NIH
Rashida Dorsey, Office of the Assistant Secretary for Planning and Evaluation, HHS
Elena Fazio, National Institute on Aging
Anna Fernandez, Booz Allen Hamilton
Shana Gillette, U.S. Department of Agriculture
Paige Green, National Cancer Institute
Patricia Jones, National Center for Advancing Translational Sciences
L. Kurt Kreuger, University of Michigan
Karen Lee, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Stephen Marcus, National Institute of General Medical Sciences
Venkata Mavuri, National Heart, Lung, and Blood Institute
Brett Miller, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Nancy Miller, National Cancer Institute
Kathryn Morris, Office of Behavioral and Social Sciences Research, NIH
Lis Nielsen, National Institute on Aging
Emmanuel Peprah, National Heart, Lung, and Blood Institute
Ronna Popkin, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development
Rajni Samavedam, Booz Allen Hamilton
Elad Sharon, National Cancer Institute
Susan Spillane, National Cancer Institute
Michael Spittel, Office of Behavioral and Social Sciences Research, NIH
Erica Spotts, Office of Behavioral and Social Sciences Research, NIH
Luke Stoeckel, National Institute of Diabetes and Digestive and Kidney Diseases
Ken Wilkins, National Institute of Diabetes and Digestive and Kidney Diseases
Carolyn Williams, National Institute of Allergy and Infectious Diseases
Joshua Williams, Office of the Assistant Secretary for Planning and Evaluation, HHS
Sung Sug Yoon, National Institute of Nursing Research

Contractor Staff

David Clarke, Rose Li and Associates, Inc.
Chandra Keller-Allen, Rose Li and Associates, Inc.
Valery Leng, Rose Li and Associates, Inc.