TBM

NEWS FROM NIH

News from the NIH: leveraging big data in the behavioral sciences

Robert M. Kaplan, PhD,[1] William T. Riley, PhD,[2] Patricia L. Mabry, PhD[3]

[1]Agency for Healthcare Research and Quality, Office of the Director, 540 Gaither Road, Suite 2000, Rockville, MD 20850, USA

[2]Office of Behavioral and Social Sciences Research, National Institutes of Health, 31 Center Drive, Building 31, Room B1C19, Bethesda, MD 20892, USA

[3]Office of Disease Prevention, National Institutes of Health, 6100 Executive Blvd., Room 2B03, Bethesda, MD 20892, USA

Correspondence to: R Kaplan
Robert.Kaplan@ahrq.hhs.gov

Cite this as: *TBM* 2014;4:229–231
doi: 10.1007/s13142-014-0267-y

Discussions about "big data" have come to dominate visions of the future of behavioral and social sciences research. So what is big data, and how can it be of benefit to researchers in the behavioral and social sciences? In this article, we describe some of the challenges and opportunities big data may present to behavioral investigators. We also point readers to a variety of NIH resources on topics relevant to big data including training, funding opportunity announcements, Listserv, and websites.

WHAT IS "BIG DATA"?

"Big data" generally refers to "*data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time.*" [1] Another widely accepted definition [2, 3] refers to Big Data as complex data that is high on the dimensions of *volume* (see previous definition), *velocity* (i.e., intensive longitudinal data such as streaming data), *variety* (i.e., a range of data types and sources being integrated), and low on *veracity* (i.e., data which cannot easily be verified, such as data from social media).

For much of the big data that comes from secondary sources, signals about behavior and its putative mediators are distributed about a vast and noisy sea of digital traces. For example, data on the use of search engines (topics searched), social media sites (e.g., Twitter, Facebook, Instagram, Foursquare), cable television viewing, and cell/smart phone use (including who was contacted, duration of the contact, location, activity level, and app use) are now widely available. Moreover, commercial wearable sensor technologies capture big data in real-time and by geographic location, such as activity levels throughout the day, sleep/wake states, exposure to noise and air pollution, and a variety of psychophysiological data.

In comparison to traditional data collection methods, these emerging technologies produce larger data sets much more rapidly and easily. These data also create new opportunities to address research questions, including those in the behavioral science domain. Big data provide the opportunity to obtain archival data on huge swaths of the population and/or prospectively collect rich streams of information from selected individuals by applying sensor technologies.

As big data become increasingly available and inexpensive, behavioral science will move from a field that relies predominately on collection of small data to one focused on both big data collection–identifying and extracting relevant data from public sources and leveraging technology to capture big data from people in a free living context–and on the development of new analytic methods to make sense of it all. However, big data from secondary sources are not likely to wholly replace new data collection.

CHALLENGES IN TURNING BIG DATA INTO KNOWLEDGE

Big data presents significant opportunities for advancing knowledge relevant to improving health, and that the associated challenges of big data can be overcome. To encourage and facilitate the use of big data for biomedical (including behavioral) research, the NIH has developed the Big Data to Knowledge program (BD2K) http://bd2k.nih.gov. Over the coming years, the NIH is slated to devote many millions of dollars to this effort.

Some of the challenges to the use of big data relevant to behavioral science are as follows:

1. *Data Access.* Data are accumulating so fast and from so many sources that a formidable challenge is being able to identify and access big data sources. Accessibility may be hampered by privacy concerns, ownership (e.g., companies own transactional data, search data), and lack of data standards. NIH's BD2K program (workgroup 1) is addressing barriers to big data access by changing NIH data access policies and by efforts to make the data more usable. This program is supporting the development of data standards and the creation of indices that will identify and describe big data sets. See http://www.youtube.com/watch?v=za0xQbI79vo&feature=youtube_gdata

2. *Consent from research participants. A priori* consent to use public big data for research purposes is rarely provided. While anonymized, archival data are typically exempt from IRB review, privacy concerns extend beyond IRB purview, and the power to integrate data from various sources will often require some form of identification that requires consent. Researchers are encouraged to familiarize themselves with privacy issues around big data [4].
3. *Sampling bias.* All samples of convenience, including those obtained from public big data are biased, often in unknown ways. For example, social media data could provide interesting and unique insights into social networks, but it remains unclear to what degree those who use social media are representative of the population, and in many cases the biases are non-trivial [4]. Post hoc methods for estimating national population representativeness from non-probability samples exist [5] and when appropriate should be applied to open data. In other cases, the non-generalizability of the data may compel the researcher to: (a) collect additional data, (b) eschew the big data source in favor or primary data collection or other data source, (c) limit the generalizability of the results, (d) use the data for hypotheses generation only, and/or (e) take other remedial action.
4. *Estimating behavior constructs.* Behavioral constructs may or may not be represented in secondary big data sources, and care should be taken to assess what is measured versus the desired construct [4]. In some cases a construct may be appropriately triangulated from multiple signals. For example, psychometrically sound instruments have been designed to measure self-efficacy, but most public data sources do not include such measures. However, self-efficacy might be estimated via semantic analysis of social media content. Such approaches should be validated against traditional methods.
5. *Development of and training in new analytic methods.* Methods appropriate for big data need to be learned and applied, and in some cases new methods need to be developed. Analysis of big data may rely on analytic approaches drawn from engineering and computer science. These methods include, data mining (to extract a subset of data for examination), pattern recognition (to help classify data), and computational modeling approaches (e.g., social network analysis, agent-based modeling, system dynamics). Moreover, technological advances and their adoption (e.g., smartphones) bring the potential for frequent measurement within subjects, such as ecological momentary assessment [6] which requires methods capable of handling intensive longitudinal data [7]. The NIH BD2K program includes a workgroup focused on analysis methods and software, as well as one on training (see more in the next section). In February 2014, BD2K issued RFA-HG-14-020, *Development of Software and Analysis Methods for Biomedical Big Data in Targeted Areas of High Need (U01)* http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-14-020.html.
6. *Causal inferences.* Random assignment and null hypothesis testing have been the primary approaches for inferring causality in behavioral science. Manipulation of independent variables is still possible in big data sets (e.g., pushing out health promotion interventions to web sites and smartphones), but Bayesian approaches and computational modeling of observed behavioral changes over time will become increasingly important methods for inferring causality. Due caution about causal inferences from such analytic approaches is appropriate, and comparisons of traditional and big data findings will be important to perform. It is also important to acknowledge, however, that other areas of science are able to infer causality without a randomized trial using big data (e.g., the movement of tectonic plates causes earthquakes).

These are important challenges in the application of big data to behavioral science questions, but the value of addressing these challenges is considerable. When used to address appropriate research questions and in combination with traditional approaches, research can be done more efficiently, rapidly, and inexpensively. Big data uses sample sizes sufficiently large to detect potentially small but important effects of behavioral moderators and mediators. Key variables are measurable via direct, tech-based, observation of behavior, not from retrospective and subjective reports of behavior.

NIH RESPONSES TO BIG DATA TRAINING ISSUES

The issues in the analysis of big data/open data are much different than the concerns that investigators faced in the past; therefore, new training paradigms are needed. For example, fields like experimental psychology are built upon small experiments often involving less than 50 participants. Inferential statistical techniques, such as the analysis of variance or the *t* test, were crucial for making inferences about larger populations based on observations in small samples. These statistical tests are highly dependent on sample size. Larger experiments had a greater probability of finding a significant effect at the level $p<0.05$. Most PhD programs develop students' analytic expertise in statistical methods to the exclusion of other approaches. Using traditional statistical inference techniques when analyzing big data can lead to every test being correctly, but inappropriately, statistically significant.

The average PhD program now takes about 6 years complete. That means that students admitted in 2014 will get their degrees in 2020. Will they be ready for the world they will face when they have completed their studies? The methodology courses in many current PhD programs in the behavioral and social sciences are quite similar to those that have been in the curriculum for the last four decades. To prepare for the new world of big data, we need a serious reexamination of the core methods courses in our PhD programs. These courses might cover issues in data privacy and

storage, analysis of masses of data, techniques of pattern recognition, and the application of systems science methods.

The NIH has responded to the need for training in big data in several ways. The NIH BD2K program recently issued an RFA to encourage applicants to develop short courses to help meet the nation's biomedical, behavioral, and clinical research needs for utilizing biomedical big data, including the application of computational and statistical sciences in a biomedical context, NIH RFA-HG-14-009 *Courses for Skills Development in Biomedical Big Data Science (R25)* http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-14-008.html. BD2K also has FOAs for career development (K01) and for open educational resources (R25)–see the BD2K Funding Opportunities website for details: http://bd2k.nih.gov/funding_opportunities.html#sthash.lqXPzP2S.7WMx4KQH.dpbs. In the spring of 2013, the NIH Office of Behavioral and Social Sciences (OBSSR) issued RFA-OD-13-009, *Short Courses on Innovative Methodology in the Behavioral and Social Sciences (R25)* http://grants.nih.gov/grants/guide/rfa-files/RFA-OD-13-009.html. This funding announcement encourages short course development for many areas relevant to big data including (but not limited to) systems science methodologies, mobile health research, and working with spatial data. The applications are currently under review and the courses should become available sometime in 2015. OBSSR has also developed several training opportunities in systems science http://obssr.od.nih.gov/scientific_areas/methodology/systems_science/index.aspx and mHealth, which are described in detail elsewhere [8, 9].

While none of the above programs addresses the need to revamp doctoral training, this is an area of considerable interest to the NIH. For example, we believe that a better understanding of the behavioral and social science workforce dynamics can inform future policies and programs, including proposed changes to training, that affect the workforce. To this end, OBSSR joined NIGMS in issuing, RFA-GM-14-011, *Modeling the Scientific Workforce (U01)* http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-14-011.html to support the development of mathematical models of the BSSR workforce.

Big data are certain to have a profound effect on nearly all fields of scientific investigation. This revolution in science is expected to cause considerable turmoil. Proactive curriculum change may be an appropriate next step. To stay informed about future opportunities, including training, in big data, check the NIH BD2K website regularly. To be notified of training and activities in systems science, join the NIH BSSR-Systems Science Listserv; email Dr. Mabry (mabryp@od.nih.gov) to subscribe. To be notified about mobile health research training and activities join the NIH mHealth listserv http://obssr.od.nih.gov/scientific_areas/methodology/mhealth/index.aspx.

CONCLUSIONS

The nature of research investigation is changing. In the past, data collection was difficult and expensive. Now data are abundant and often easy to capture. Most of us were trained to apply inferential statistical methods to estimate population parameters from sample data. These methods may be of little value when sample sizes include millions of observations. To adapt to the new world of big data, we need to adjust our investigational and training programs. PhD programs may need to devote more curriculum time to the challenges of analyzing millions of observations often captured outside of the laboratory. The NIH has initiated a new program to translate big data into knowledge (BD2K). In addition, NIH will support short term training programs on new research methods.

1. Wikipedia, "Big Data". http://en.wikipedia.org/wiki/Big_data. accessed March 12, 2014
2. Laney D. *3D data management: Controlling data volume, velocity, and variety*. Application Delivery Strategies, META Group/Gartner http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. 2001.
3. Zikopoulos PC, Eaton C, deRoos D, Deutsch T, Lapis G. *Understanding big data. Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw-Hill; 2012.
4. Boyd D. *Privacy and Publicity in the Context of Big Data*. WWW. Raleigh, North Carolina, April 29. http://www.danah.org/papers/talks/2010/WWW2010.html. 2010: Accessed March 15, 2014.
5. Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile KJ, & Tourangeau R. Report of the AAPOR Task Force on Non-Probability Sampling. American Association for Public Opinion Research. http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=5963.2013: Accessed March 20, 2014.
6. Shiffman S, Stone A, Hufford M. Ecological momentary assessment. *Annu Rev Clin Psychol*. 2008; 4: 1-32.
7. Ginexi E, Riley WT, Atienza A, & Mabry P. (in press). *The promise of intensive longitudinal data capture for behavioral health research*. Nicotine and Tobacco Research, in a special issue entitled: *New methods for advancing research on tobacco dependence using ecological momentary assessments*.
8. Mabry PL, Kaplan RM. Systems science: a good investment for the public's health. *Health Educ Behav*. 2013; 40(1): suppl 9S-12S.
9. Nilsen W, Riley WT, Heetderks W. News from the NIH: using mobile and wireless technologies to improve health. *Transl Behav Med*. 2013; 3(3): 227-228.