Published in final edited form as:

Am J Prev Med. 2011 May ; 40(5 Suppl 2): S159–S161. doi:10.1016/j.amepre.2011.02.001.

Making Sense of the Data Explosion:

The Promise of Systems Science

Patricia L. Mabry, PhD

Office of Behavioral and Social Sciences Research, National Institutes of Health, Bethesda, Maryland

The Data Explosion

The past decade has given birth to a data explosion; data are now captured and shared at unprecedented levels, and many of the data are relevant to human health. A good portion of the data have emanated from personal devices (mobile phones, personal computers), but other data abound. Transactional data are generated by purchases at the point of sale and can be associated with individuals via their credit card and/or store club card. Surveillance cameras capture behavior surreptitiously in vending areas, at banks, on street corners/highways, and in sports arenas. Across these modalities, data exist relevant to a wide range of health behaviors (e.g., commuting patterns, food and tobacco consumption, recreational behavior), along with contextual data (e.g., network membership, movement across geographic space) and biomedical data (e.g., genetic data, physiologic indicators, clinical measures) as well as other health data (e.g., access to care, insurance coverage).

The data explosion has been aided and abetted by rapid advances in technology that have greatly increased the range and capacity of the types of data that are collected. “Phones” are no longer used solely for real-time voice communication; smart phones can capture text, audio, video, and GPS data virtually anywhere, anytime. The portability, affordability, vast data capacity, and versatility of mobile devices along with their ubiquity have made the data-collection possibilities innumerable. The iPhone alone boasts over 200,000 apps. The recent 2010 mHealth Summit (www.mhealthsummit.org) cosponsored by the NIH Foundation and a host of private companies drew over 2400 attendees from over 50 countries and is a testament to the growing interest in this area.

In the meantime, connectivity of the social space has increased dramatically with the advent of sites like MySpace, Facebook, YouTube, PatientsLikeMe, the blogosphere, and other “participatory” sites where content is user driven. Sharing has not only become easier by virtue of the cyberinfrastructure, but the social cyberspace also provides an *incentive*, a social reward, for doing so. Thus, the data that are being captured are being shared like never before.

Address correspondence to Patricia L. Mabry, PhD, Office of Behavioral and Social Sciences Research, Office of the Director, National Institutes of Health, 31 Center Drive, Bldg 31, B1-C19, MSC 2027 Bethesda MD 20892-2027. mabryp@od.nih.gov.

Disclaimer: The article presented here is the sole work of the author and does not imply or reflect official positions of the National Institutes of Health or the Office of Behavioral and Social Sciences research.

No financial disclosures were reported by the authors of this paper.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Facing a data deluge, it is easy to become overwhelmed. The task facing the research community is to find ways to assemble shards from the mountains of data into comprehensive pictures that can help us understand and act on problems in the population/public health domain. A straightforward and illustrative example of this is the use of online search data on influenza symptoms to predict regional outbreaks; Google data showed that online searches of influenza symptoms peaked just before a CDC-documented outbreak in 2008.¹ A tremendous potential exists to leverage data, not just from single sources, but also by combining data from multiple sources (aka “data fusion”) to understand and make inferences about online and offline behavior. The remainder of this commentary describes opportunities for employing systems science methodologies in public health research within the context of the data explosion. Human subjects protections and privacy concerns are important, but are not within the scope of this paper.

Systems Sciences Approaches for Complex Problems

In the past 5 years, the behavioral and social science research community has begun to appreciate the complexity inherent in most public health problems.^{2,3} A growing number in this community now recognize that methodologies designed to address complexity exist; a few have availed themselves of training in systems science, and a handful are receiving NIH funding to apply systems science methodologies in their research. The recognition that public health problems are complex—multifaceted with interacting components, time delays, and nonlinear and bidirectional relationships—is not revolutionary; however, the adoption of systems science methods in the population/public health community most certainly is. Traditional methods taught in behavioral and social science graduate programs and applied in NIH-funded projects in this area have been predominantly correlation-based (e.g., regression), reductionist, and designed to factor out complexity rather than study it directly. Systems science methods allow a wider range of aspects of the problem to be explored without negating any findings made available via traditional methods. As stated elsewhere,³ systems science methods augment, and do not replace, traditional methodologies.

Making Sense of the Data Explosion

The recognition that public health problems are complex and that there are methodologies suitable for addressing these complexities is taking place against the backdrop of the data explosion. The data tsunami compels researchers to use systems science methodologies, which are designed to get at complex phenomena,^{2,3} to make sense of the data, while simultaneously providing the data needed to feed the models generated under these approaches.

To make sense of this large and complex volume of health-related data, cyberinfrastructure is needed. That is, the scaffolding for depositing, coordinating, and organizing data, and facilitating their analysis represented by efforts including the Office of the National Coordinator for Health Information Technology’s Nationwide Health Information Network (NHIN), and the National Cancer Institute’s (NCI) cancer Biomedical Informatics Grid (caBIG®). The NCI’s PopSciGrid conceptual framework is one approach for understanding the application of cyberinfrastructure in the public health arena (cancercontrol.cancer.gov/hcirb/cyberinfrastructure/). The PopSciGrid framework is composed of three layers: (1) a data layer, which will facilitate the use of a variety of data sources (i.e., federal, state and local, clinical, and academic) by stipulating common data elements, data standards, and common vocabularies; (2) a GRID layer, which includes ontology development, communications protocol, and metadata registry; and (3) an application layer, consisting of tools that can easily interact with the data for a variety of

aims: discovery, visualization, decision support, fusion, and policy planning.⁴ To address public health challenges effectively, systems science approaches can and should play a central role in the application layer of PopSciGrid as well as in other similar initiatives that are currently under development, for example, the DHHS's Community Health Data Initiative, and *Splash!* by IBM.⁵

By housing data, modeling tools, and models themselves, cyberinfrastructure can provide an efficient means for leveraging the data explosion. Specifically, they can greatly facilitate linking data and models so that they can be used in an iterative fashion; models and simulation can synthesize existing data, expose data gaps, and inform health policy decisions.³ Data can fuel the models and be used to check the accuracy of previous forecasts, informing further adjustments to the models.

In this vein, a variety of modeling efforts have already begun influencing policy decisions for population health. Perhaps two of the best known are CISNET and MIDAS, NIH modeling efforts aimed at cancer and infectious disease, respectively. The Interagency Modeling and Analysis Group (IMAG) is a consortium of federally sponsored groups and emphasizes modeling across multiple scales of biomedical systems (e.g., atomic, genetic, cellular, tissue, organ, whole body, and population). The CDC has sponsored the development of Health Bound, a modeling and simulation tool for health system planning. A very recent effort, Envision, is a network of modeling teams, each using a different method and model of childhood obesity policy; Envision is part of the National Collaborative on Childhood Obesity Research (NCCOR) collaborative. At the state level, Karen Minyard and colleagues at the Georgia Health Policy Center have been trailblazers in engaging the state legislature in using systems science methodologies to help aid their decision making.⁶

To date, system dynamics modeling, agent-based modeling, and network analysis have been the methodologies featured prominently in NIH funding opportunities and training initiatives for population and public health. Other systems science methodologies that have potential to help make sense of the data explosion include control engineering, fuzzy logic, and neural networks.

Control engineering methods are showing great promise for aiding in behavioral intervention design by guiding decisions about the optimal ordering and strength of intervention components (e.g., a weight-loss intervention⁷). Engineering approaches have also been used to inform the design of adaptive, time-varying interventions for drug, alcohol, and other relapsing behavioral problems. In these designs, the dosage and/or type of treatment varies between individuals and within individuals across time according to the patient's changing needs (e.g., a preventive intervention aimed at children at risk⁸).

Fuzzy logic allows degrees of truth to be built into models as a way of handling uncertainty. Membership in a set or acceptance of a truth is expressed in degrees of belongingness/truthfulness rather than as a binary outcome. For example, a glass can be some degree of full, in contrast to a binary approach in which the glass is full or empty. Models built on fuzzy logic can more closely mimic human decision making and are robust under highly variable, volatile or unpredictable conditions. Fuzzy logic conveys the concept of *how much* a variable belongs to a set whereas probability measures express *how likely* it is for a member to belong to a set.

Artificial neural networks are mathematical or computational models that have been designed to mimic important aspects of the structure and function of networks of biological neurons. An artificial neural network:

consists of an interconnected group of artificial neurons...[and in most cases] changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools [that are] used to model complex relationships between inputs and outputs or to find patterns in data.⁹

Some public/population health problems may be amenable to being characterized as neural networks.

Hand in hand with cyberinfrastructure, systems science methodologies are a powerful and necessary way to make sense of the data for public health/population health impact. Readers interested in learning more are encouraged to consider the International Conference on Social Computing, Behavioral Modeling, and Prediction; and the Workshop on Dynamic Modeling for Health Policy and may contact the author for additional information on systems science.

Acknowledgments

The sole source of funding for this work was provided by the National Institutes of Health.

References

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*. 2009; 457(7232):1012–1014. [PubMed: 19020500]
2. Mabry PL, Olster DH, Morgan GD, Abrams D. Interdisciplinarity and Systems Science to Improve Population Health: A View from the NIH Office of Behavioral and Social Sciences Research. *Am J Prev Med*. 2008; 35(2 Suppl):S211–S224. [PubMed: 18619402]
3. Mabry PL, Marcus SE, Clark PI, Leischow SJ, Mendez D. Systems Science: A Revolution in Public Health Policy Research. *Am J Pub Health*. 2010; 100(7):1161–1163. [PubMed: 20530757]
4. Shaikh, AR.; Contractor, N.; Moser, R., et al. PopSciGrid: Using cyberinfrastructure to enable data harmonization, collaboration, and advanced computation of nationally representative behavioral, demographic, and economic data. Proceedings of the 137th APHA Annual Meeting; 2009. apha.confex.com/apha/137am/webprogram/Paper196543.html
5. Maglio, P.; Cefkin, M.; Hass, P.; Selinger, P. Social Factors in Creating an Integrated Capability for Health System Modeling and Simulation. In: Salerno, J.; Chai, S-K.; Mabry, PL., editors. Proceedings of the 2010 International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP10); Bethesda, MD. March 29-April 1, 2010; Springer; p. 44-51. Lecture Notes in Computer Science 6007/2010
6. Minyard, K. Using Systems Thinking and Collaborative Modeling to Improve State Policymaking. Presentation at the 2010 Institute on Systems Science and Health (ISSH). isssh.aed.org/Files/Minyard_COModel_ISSH_0610FINAL7132010.pdf
7. Navarro-Barrieitos, J-E.; Rivera, DE.; Collins, LM. A Dynamical Systems Model for Understanding Behavioral Interventions for Weight Loss. In: Salerno, J.; Chai, S-K.; Mabry, PL., editors. Proceedings of the 2010 International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP10); Bethesda, MD. March 29-April 1, 2010; Springer; p. 170-179. Lecture Notes in Computer Science 6007/2010
8. Zafra-Cabeza A, Rivera DE, Collins LM, Ridao MA, Camacho EF. A Risk-Based Model Predictive Control Approach to Adaptive Interventions in Behavioral Health. *IEEE Transaction on Control Systems Technology*. 2010; (99):1–11. PP.
9. Wikipedia. Entry: Artificial Neural Network. en.wikipedia.org/wiki/Artificial_neural_network