NATIONAL CANCER INSTITUTE

SCIENCE OF RESEARCH AND TECHNOLOGY BRANCH

BIG D.A.T.A. (DATA AND THEORY ADVANCEMENT) WORKSHOP

EXECUTIVE SUMMARY

OVERVIEW

The National Cancer Institute (NCI) Science of Research and Technology Branch convened experts in big data analytics, systems science, and theory development and testing at the September 2013 workshop, **Big D.A.T.A. (Data And Theory Advancement).** Workshop participants gathered at the National Institutes of Health (NIH) campus to address how to leverage big data and dynamical systems models to advance health behavior theory in the context of modifying cancer and other disease risk factors. The era of big data presents opportunities to substantively test, refine, and improve health behavior theories. The goal the Big D.A.T.A workshop was to stimulate new directions in theory development, testing, and integration with the use of big data, dynamic systems modeling, and measurement advances.

Advances in health behavior theory development, integration, and testing are clearly needed, as current theories have shown to be insufficiently explanatory, unnecessarily fragmented, and inadequately tested. Big D.A.T.A workshop participants envisioned a shift in focus in theory research to advance the study of health risk behaviors that remain the primary causes of morbidity and mortality.

Recently, researchers have begun to apply dynamical systems models to behavior theory. These innovative modeling approaches create more robust and testable models of human behavior. While continued modeling efforts will provide needed simulation data, actual data obtained across multiple perspectives and levels (e.g., individual, environment) are needed to fully test and refine these dynamical models of human behavior. Therefore, these modeling efforts should take advantage of the increasing availability of big data, which can estimate some theoretical constructs. The Big D.A.T.A. workshop objectives and anticipated outcomes are consistent with the NIH Big Data to Knowledge (BD2K) initiative (http://bd2k.nih.gov). Behavioral theory and related sources of data were largely underrepresented in the BD2K effort. In contrast, there is a greater focus on the use of large data sets and analytics specifically to advance behavior theory in the Big D.A.T.A. workshop initiative. In fact, the Big D.A.T.A. workshop was designed to help connect the BD2K and theory research community, while incorporating systems and big data analytic approaches.

WORKSHOP STRUCTURE AND PROCESS

The two day Big D.A.T.A. workshop included participants from various extramural scientific communities involved in health behavior theory, systems science, methodology, data analytics, and behavioral interventions. Participants also included scientists from various governmental entities including the Department of Health and Human Services (HHS) and the National Science Foundation. Representatives of the NIH BD2K initiative also participated in the workshop.

Welcome and opening statements were given on Day 1 by Robert Croyle, Director of the Division of Cancer Control and Population Sciences, and on Day 2 by William Klein, Director of the Behavioral Research Program at the National Cancer Institute. Workshop discussions were organized around the opportunities and challenges within five thematic topic areas:

1. Health Behavior Theory
2. Systems Modeling
3. Social Network Data Analysis
4. Big Data Mash-ups and Statistical Modeling
5. Dynamic Interventions

For each of these topics, two workshop participants provided their perspectives on the challenges and opportunities for advancing theory, followed by a discussant who provided initial responses to these perspectives to stimulate and facilitate discussion.

In addition to the thematic topic discussions, two workshop participants, Genevieve Dunton and Daniel Rivera, were asked to provide a demonstration on application of systems modeling approaches (Rivera) to theory-based ecological momentary assessment (EMA) data (Dunton). Following the topic discussions, breakout groups were formed to discuss the most promising future opportunities to advance behavioral theory using a) systems dynamics, b) social network analysis, c) data integration, and d) dynamic interventions. A Federal panel responded to the breakout reports and discussed interests in big data and theory advancement. The workshop concluded with a summary discussion among participants and suggestions for next steps.

PRESENTER LIST AND AGENDA FORMAT

Day 1: September 19, 2013

Welcome
Robert Croyle, National Cancer Institute

Workshop Orientation:
Audie Atienza, National Cancer Institute

Health Behavior Theory: Opportunities and Challenges
Alex Rothman, University of Minnesota
Jasmin Tiro, University of Texas Southwestern
Bob Evans, Google

Systems Modeling: Opportunities and Challenges
Ross Hammond, The Brookings Institute
Daniel Rivera, Arizona State University
Stephen Intille, Northeastern University

Applying Dynamic Systems Modeling to Time-Intensive Data
Daniel Rivera, Arizona State University
Genevieve Dunton, University of Southern California

Social Network Data Analyses: Opportunities and Challenges
Nosh Contractor, Northwestern University
Nathan Cobb, MeYouHealth
Holly Jimison, Northeastern University

Big Data Mash-ups & Statistical Modeling: Opportunities and Challenges
Patrick Curran, University of North Carolina at Chapel Hill
Donna Coffmann, Pennsylvania State University
Eric Hekler, Arizona State University

Dynamic Interventions: Opportunities and Challenges
Linda Collins, Penn State University
Bonnie Spring, Northwestern University
Genevieve Dunton, University of Southern California

Breakout groups
A. Systems Modeling and Behavioral Theory
B. Network Analyses and Behavioral Theory
C. Data Mash-ups, Stats Modeling, and Behavioral Theory
D. Dynamic Interventions and Behavioral Theory

Lightning Round Presentations (all participants in each breakout group):
(1) Introduce yourself & your work related to the Big Data, Health IT, and/or Behavioral Theory

(2) In your opinion, what are the most promising future opportunities for leveraging Big Data to Advance Behavioral Theory and how can different disciplines work together toward this goal?

(3) From your perspective (given your background and expertise), what do you see as the most significant gaps and/or barriers in the field of Big Data and Behavioral Theory?

Day 2, September 20, 2013

Welcome Day 2 & Report of Breakout Groups:
Bill Klein, National Cancer Institute

Federal Panel Discussants
Wendy Nilsen, National Institutes of Health
Misha Pavel, National Science Foundation
Lynda Hardy, National Institutes of Nursing Research & BD2K
Damon Davis, U.S. Dept. of Health and Human Services

Synopsis of "International Workshop on New Computationally-Enabled Theoretical Models to Support Health Behavior Change and Maintenance"
Donna Spruijt-Metz, University of Southern California

Closing Remarks
Bill Riley, National Cancer Institute

Thematic Topic Areas	Observational Notes	Challenges & Needs	Advantages/Opportunities
Health Behavior Theory	• Theories are a set of concepts and their interrelationships that describe the expected relationship between phenomena • Underlying assumption is that theory is useful both for explaining and changing health behaviors • Theory development and testing is stagnant • Measurement of latent variables, and variations of measurement equivalence across modes of administration and across languages and cultural contexts • Machine learning comes from a Bayesian tradition, distinct from traditional hypothesis testing	• Insufficient guidance on action aspects of theory (i.e. how to change the theoretical construct) than on the conceptual aspects of theory (how the construct affects behavior) • Inadequate competitive theory testing • Big Data as the “end of theory” • Constrained by what is measured and theory guides what is measured • Need to consider when outliers and missing data provide information relevant to theoretical questions • Need discrete and precise measurement, and sufficient variance in the data to inform the model • Can the language of behavior become executable (machine learning) functions • Sensor data is proxy, messy, biased, and needs its own models • Need an overarching model to inform data collection methods • Need to move theory from “what is interesting” to “what is important”	• Link survey data that measure cognitions and perceptions unknowable from any other source with other data sources that measure behaviors in real-world contexts • Specify when a theory does not apply - for whom, for what, for when (boundary conditions) • Elucidate how behavior unfolds over time, especially different factors operating during initiation vs. maintenance • Model both linear and curvilinear relationships • Stimulate new and more complex theoretical premises • Forces us to challenge and make explicit our theoretical assumptions • Potential to use decision trees as a way of describing theory • Set up clear mathematical deductions that are falsifiable • Overlap of constructs with similar conceptual and/or operational definitions across theories

Thematic Topic Areas	Observational Notes	Challenges & Needs	Advantages/Opportunities
Systems Modeling	• Systems science is composed of many different modeling techniques • Rapidly growing area intersecting across public health	• Difficult to map actual data elements onto theoretical constructs • Need sufficiently dense data sets which requires sustain participant engagement • Need to develop a coherent base of knowledge from which to test theory • Current theories are not precise or detailed enough • Need a common set of constructs and common set of theories • Need large enough data sets to insure the models work • Health data are generally large but sparse • Need consensus methodologies for intensive longitudinal data, including handling missing data, timing and spacing of data, measurement theory, experimental designs, etc.	• Captures social and environmental influences and co-evolution of social influences with individual behavior (multi-level) • Allows a better understanding of change and effect in behavioral systems (e.g. the speed, shape, and magnitude of response) • Enables a more efficient use of intensive longitudinal data • Dynamic modeling allows for adaptation, including adaptive responses to an intervention over time Big data in social science models • System science models can inform big data collection • Common and clear language around what we mean by “theory,” “model,” and other terms that have different meanings among theoreticians and modelers to facilitate interdisciplinary team science

Thematic Topic Areas	Observational Notes	Challenges & Needs	Advantages/Opportunities
Social Network Data Analysis	• At a confluence of theory, data, and methods – and the computational infrastructure to test theories at scale (e.g. Lazar – computational social science) • By "Big Data," we really mean "Broad Data"	• In social network analyses, data has outraced utility • Ethical issues regarding data gathering of "others" • Considerable commercial or "dark data" that cannot be obtained for theory research purposes • Mapping social connections does not specify the value or processes involved in those connections (e.g. social influence vs. homophily) • Work needed on developing predictive models, determining the appropriate experimental design, and defining outcomes • Need more team science approaches that produce a "creative friction," and more integrated approaches across teams • Funding needs to be faster and incentive working together	• Invites consideration of new combinations of variables – drawing from data that will expand theories • Highlights the importance of boundary conditions • Movement toward citizen science • Ability to test interventions at scale via web companies • Ability to computationally model new theories • Assess if theories scale up to large population-based behaviors • Allows exploration of social context, socialization, norms, and influences • Provides alternative of targeting social network catalyst

Thematic Topic Areas	Observational Notes	Challenges & Needs	Advantages/Opportunities
Big Data Mash-ups and Statistical Modeling	• Integrative Data Analysis (IDA) is the fitting of statistical models to raw data pooled across samples • Tie data together via cross-calibration of different measures, or via probabilistic mapping • Health behavior theories are "non-bayesian" in that they never update the prior based on evidence	• Extensive data preparation, especially for EHRs • Correlation is not causation • Low signal to noise ratio • Need to insure that statistical thinking remains the bedrock of data science • Continued questions about what constitutes "big data." Broad data? Time-intensive data? Broad data that links everything related to context? Five V's –Volume, Variety, Velocity, Value, Veracity. • Need to consider how to integrate experimental design in big data to shift from correlative models • Need bridging studies to integrate disparate data sets • Address issue of measuring what we can easily measure, not what is most important • How do we measure success of these big data efforts?	• Efficient use of resources • Improved power • Adequate sampling of rare phenomena • Greater sample heterogeneity • Study replication • Potential value on prediction • Use of data mining or machine learning for propensity scores • Big data analytics focus on variance, not just central tendency which loses information

Thematic Topic Areas	Observational Notes	Challenges & Needs	Advantages/Opportunities
Dynamic Interventions	• Tailoring variables on each individual measured periodically to adjust treatment • Adaptive Treatment Designs – MOST, SMART, control system approaches	• Drowning in information – how do we capture, visualize and make data actionable • Lack dynamic theories to guide intervention development • Lack predictive models that guide preemptive vs. reactive interventions • Need to optimize interventions based on theoretical constructs vs. atheoretical strategies • Focus not only on adding but also subtracting from interventions over time during maintenance	• Just in time adaptive intervention - immediacy of feedback – and use machine learning to learn decision rules for each individual • Access to continuous data that speeds learning about treatment • Can help us explore theoretical mediators over time

MAIN POINTS FROM BREAKOUT GROUPS

SYSTEMS MODELING AND BEHAVIORAL THEORY

A large, Trans-NIH effort to develop a single broad dataset to model behavior and interventions at the population level would refine techniques for anonymization, encryption, data integration, and data analytics, etc. This effort may also include collaborations with private companies and their data sets; and provide a platform for building and testing competing models, simulating the effects of various interventions. Privacy and data integration issues will need to be addressed, as well as collaborations of large and diverse teams of researchers.

NETWORK ANALYSES AND BEHAVIORAL THEORY

Areas to consider:

1) How to encourage novel “nodes” for social network analysis and research that treats concepts, places, etc. as nodes, not just individuals. (i.e.,Can social network analysis be used to map and connect theoretical constructs?)

2) How to capitalize on existing social network, data but augment with more behavioral theory data.

3) How to increase access to private social network data sets using an “honest broker” between private and the public health researchers.

DATA MASH-UPS, STATS MODELING, AND BEHAVIORAL THEORY

Less burdensome assessment approaches (e.g. passive sensing, low-burden CAT) will allow for more intensive data collection at multiple levels. This will provide the basis for building models of interaction between person and environment, studying variability and trends over time, and better model the shape of effect curves. This requires openness to single case designs and better alignments of incentives for team science approaches

DYNAMIC INTERVENTIONS AND BEHAVIORAL THEORY

Opportunities to consider include:

1) Team science approaches facilitate pragmatic innovation.

2) Context-contingent behavior theories delineate the boundary conditions and facilitate common theory ontologies, and

3) Extensive training is warrented in these approaches, including workshops on dynamic intervention methods, training institutes on comprehensive adaptive interventions, and demonstration projects on just-in-time adaptive interventions (perhaps with smoking cessation as a first model). Development and refinement of methods for the creation of just-in-time adaptive interventions would be a worthwhile investment.

POTENTIAL PARTNERS

- NIH Office of Behavioral and Social Science Research
- National Science Foundation
- NIH BD2K
- HHS Big Data Initiative (in conjunction with the yearly *Health Datapalooza*)

SYNOPSIS OF "INTERNATIONAL WORKSHOP ON NEW COMPUTATIONALLY-ENABLED THEORETICAL MODELS TO SUPPORT HEALTH BEHAVIOR CHANGE AND MAINTENANCE"

Donna Spruijt-Metz presented a summary of the *2012 European Union* meeting in Brussels on "New Computationally-enabled Theoretical Models to Support Health Behavior Change and Maintenance." The overarching question of the 2012 meeting was "what are the major discoveries and innovations that would enable use of existing and merging technology to its full potential to understand human behavior?" The three main action areas from the meeting were: 1) Generation of new computational models of behavior; 2) development of data sources that are rich, temporally dense, and longitudinal for a large number of individuals; and 3) creation of a shared behavior change vocabulary and ontology. The workshop also identified a number of challenges including the need for new research methods to analyze longitudinal real-world data, greater awareness and education of science gatekeepers (e.g. Deans, Editors), and the need for an inventory of current tools for complex data and modeling, as well as more partnerships between behavioral science, computer science, and engineering.

CONCLUSION

The Big D.A.T.A. workshop concluded with a general discussion of major themes and next steps. Participants noted that the use of big data and system modeling for theory advancement is a major and potentially paradigmatic shift from the current science of theory development and testing. This research area needs to move quickly but there are academic and funder barriers to disruptive and rapid research changes. In addition, there is an urgent need for more training and infrastructure. We need access and influence in large data sets, better public-private partnerships, and a substantial training effort.

Training is a critical need buy we need to be cognizant of the time costs. One possible model is the Penn State lunch seminar series, *Taste of Methodology.* Seminars give researches an overview of the new approach to consider its applicability to their work and determine if more intensive training is worth the time. Other training efforts include targeted immersion, virtual training, and rapid consulting efforts. Matchmaking of behavioral theorists, researchers with intensive longitudinal data expertise, and computational modelers would be beneficial. The Big D.A.T.A. workshop demonstration of pairing a modeler (Rivera) with an EMA researcher (Dunton) showed how this process can quickly achieve new synergies and interesting results. Clearly, with more time and resources, this sort of pairing can yield new findings. The field can benefit from "bonding agents" – research teams with the technical expertise – that can show others how this work is done and facilitate new connections.

There is also an urgent need for the research infrastructure to facilitate modeling of big data for theory advancement. There is a lack of computer programming and modeling manpower at an affordable level under traditional funding. We need to consider how to support an organization of programmers for behavioral research. There is also a clear need for more data repositories and common data elements that facilitate the merging and integrated use of these datasets. Libraries of the future will be data repositories, and tapping the infrastructures of industry could facilitate this effort.