Toward Next Generation Social Analytics: A Platform for Analysis of Quantitative, Qualitative, Geospatial, and Temporal Factors of Community Resilience

Dennis J. Folds, C.J. Hutto, Thomas A. McDermott

Georgia Institute of Technology, Atlanta, Georgia

Email: dennis.folds@gatech.edu; clayton.hutto@gtri.gatech.edu; tom.mcdermott@gtri.gatech.edu

***Abstract*— Social science stands on the brink of a revolution – or of failure. It needs powerful new tools, methods, and paradigms in order to succeed. These will include advances in computational capabilities, machine-based knowledge assimilation, quantitative analysis, and measurement. Human social analytics in the next generation will need to embrace more multifaceted representations of human behavior with more complex models. Such models will need to integrate data of disparate forms, using disparate units of measure, collected from disparate sources, at disparate scales. The development of a complex model of societal well-being (an inherently qualitative construct) forms the basis of research for a next-generation societal resilience model. The model combines traditionally separate socio-environmental and psychological constructs of resilience, a representation that requires large scale quantitative, geospatial, and temporally referenced data of disparate forms, units, sources, and scales. The research forms a framework for the development of data analytic experimentation platforms in the social sciences. The platform will be used to demonstrate tools and methods that facilitate the progression towards next generational social analytics at large scales. These concepts, tools, and methods are intended to empower social science in transformative ways.**

Keywords- Computational social science; human social analytics; human-centered data science; sociotechnical systems.

I. A VISION FOR THE NEXT GENERATION OF SOCIAL SCIENCE AND ANALYTICS

In a special session on Next Generation Social Analytics, held as part of the HUSO 2016 conference in Barcelona, Spain, a call to action and two papers were presented that discuss the challenges faced by, and payoffs expected from, the tools and methods that will facilitate future conduct of social science research [1][2][3]. The results of that session along with additional related research activities are discussed herein.

Social science is under intense scrutiny from politicians, funding agencies, leaders of scientific societies, and from within the constituent disciplines [4][5]. Well-publicized instances of fraud and misconduct may dominate popular headlines, but underlying problems with replicability, poor generalizability, poor methods, and biased interpretation of results are larger problems, and do not involve malfeasance. Into this domain, the internet's global interconnectivity and massive data availability provide an opportunity to transform social science methods. Next generation social analytics can take advantage of the internet's massive reach and information capacity to produce families of new methods and tools that address these underlying problems.

However, it is dangerous to just assume that today's connectivity and access to information will improve formal social science methods. New approaches to study human behavior in real time using social media analysis are relevant but must be placed in the context of larger behavioral theories with decades of longitudinal study. Adding to this, even with decades of studies of human behavior at the individual and group levels, comprehensive theories that adequately account for behavior in real world conditions remain illusive. Behavior is indeed complex, but at the root of social science is the conviction that behavior is lawful. Much of basic psychology (sensation and perception, for example) is well established. But studies of such constructs as beliefs, political action, and organized violence lack unifying theories that have any success in accounting for wide ranges of social phenomena. Methods that correctly use the internet to study and create new theories of social behavior must be linked.

Study after study is trumpeted in the popular media, as long as it fits the ideological preferences of the media gatekeepers, despite ongoing lack of replication and obvious failures in generalizability. Whether these shortcomings are largely the result of poor methods, poor interpretation, or simply the complexity of the phenomena studied, is not known.

Many other sciences, notably astronomy and biology, have come under intense scrutiny and criticism when results contradicted the received wisdom, often from religious authorities. Methods were questioned, interpretations of results were challenged, and scientists were attacked when science threatened to undermine religious and civil authority. In most circles, though, these sciences prevailed because the soundness of the methods, data, and interpretation withstood objective scrutiny. Social science has not yet achieved that status.

What accounts for the difficulty in achieving robustness in social science theory? Is it the very complexity of social behavior? Problems with the way data are collected, analyzed, and interpreted? Or is it that much of the subject matter of social science is at the heart of political and religious spheres of interest? Astronomy may have benefited from the fact that the power of political and religious authorities did not in fact reside in whether the earth was at the center of the universe. Biology faced a stronger challenge, but it, too, benefited when political and clerical leaders realized that the origins and evolution of species were not central to their sphere. Social science, though, must address topics that are at the heart of political and religious discourse. Many of these are also at the heart of today's debates on facts/alternative facts, fake news, etc. It is imperative that the social sciences harness the power of the internet to make the underlying science more robust and transparent.

Moreover, scientists are also human beings, behaving in social situations, as they practice science. A physicist who changes her position on, say, string theory may face social pressure for and against the change, but string theory does not lie at the heart of public policy debates. But social scientists who study violence in inner cities, for example, are studying issues that do affect elections. To proclaim theories that question the wisdom of public policies related to reduction of inner-city violence is to court opposition from supporters of those policies, and adulation from those who wish to change them. There is little reason to think that either side is particularly interested in science for science's sake.

The disciplines that seek to address social phenomena include experimental psychology, social psychology, sociology, cultural anthropology, cognitive science, medicine, evolutionary biology, and political science. The computational sciences are increasingly interested in addressing social phenomena, and environmental sciences are also quite relevant in this arena. It is daunting to imagine theories and methods that could satisfactorily span these disciplines.

A. Revolutionary Concepts Needed

With respect to reproducibility, repeatability, and generalization of experiments, social scientists must accept that there is a problem and mount efforts to address it [6]. As other sciences matured, repeatability of results became expected, and lack of repeatability besmirched both the scientist who reported the study and the theory it supported. Social science must reach this point of maturation.

To reach this point, social science must develop a culture of sharing data, and agreement on methods of measurement and analysis. The infrastructure is in place, the methods are not matured. Repeatability of results cannot be expected when constructs are not defined the same way and measured the same way across studies. Results from studies that lack internal validity cannot be expected to have external validity, that is, to be generalizable beyond the specific conditions under which those results were obtained. Thus, long-term success in social science must address construct definition, measurement methods, and theoretical frameworks that span multiple academic disciplines, not just sharing of results and underlying data.

A revolution in connectivity and computational resources available to support social science is underway. "Big data" gives rise to the need for big platforms that support collection, maintenance, and sharing of social science data. Advances in machine-based text processing will produce methods to automatically scour the world's literature for new findings, new methods, and new interpretation of vast repositories of social science data. Cognitive systems may well scrutinize published studies and identify the topic studied (even if called by different names in different studies), results and interpretation, and potential errors and biases in the study. This will allow for ongoing meta-analyses of prior studies and assimilation of multiple diverse data sets. For social scientists to understand these analyses, new visualization techniques are needed, and machine-generated interpretations expressed in natural language must accompany those visualizations.

We can envision, then, a future in which social science studies are routinely conducted in the context of massive, ongoing collection of data about human behavior around the globe. These data sets will include everything from casual social media utterances to economic and policy decisions made by corporate and government leaders. One source of enabling technology is what is being called the Internet of Things (IoT): data from cyber physical objects such as mobile phones, automobiles, and home appliances will provide data about the behavior of people using those things. These data sets can be continuously updated. New hypotheses can be generated by scientists and by software systems, and competing theories can be subjected to ongoing tests as new data arrive.

A comparable situation emerged in meteorology as the community converged on the attributes to measure, the measurement methods, data representation conventions, and protocols for sharing. Nowadays, a typical study in meteorology does not necessarily involve developing new measurement capability and collecting new data (although such studies do exist). Rather, a typical study might simply involve formulating a new hypothesis about causal mechanisms in weather patterns, and testing that hypothesis using massive data sets freely available across the community.

Perhaps social analytics will follow a similar pattern. Perhaps the globe will be instrumented with data collection capability for social phenomena the way that it is instrumented for local temperature, wind, and precipitation. These social data will be validated and loaded into accepted registries, and will immediately be used to update ongoing studies. New studies can be implemented in those registries, to test new hypotheses about causal mechanisms in human social behavior. These could be exciting times for social scientists.

Even more exciting is how these capabilities can positively impact the human condition – not just the advancement of science. These new capabilities can help us address social problems more effectively – not just measure them more reliably. Problems related to human health, standard of living, and subjective well-being (SWB) are intricately related to the phenomena studied by social scientists. In the developed democracies, re-election of incumbents is also affected by these phenomena. Politicians and business leaders alike will have a vested interest in the integrity of the social sciences and will therefore be more likely to keep them properly resourced.

B. A Timely Case Study: Well-being and Societal Resilience

The fitness and function of infrastructure in cities – with shelter, water, energy, transportation, and social interaction perhaps the most primal – is critically important for the development, survival, sustainability, resilience, and overall success of communities. Sustainability and resilience of critical infrastructure, and the related human concepts of livelihood and SWB, are becoming the subject of greater and greater scientific study. Human communities and their city infrastructure and institutions are strongly coupled interdependent systems, and the concept of community resilience cannot be evaluated predictively using simple indices or optimization of individual components. There is a need to model these systems using complex representations of

human and community development, participatory methods that address system complexity to engage communities and planners, and next generation social analytics tools to evaluate predicted, short-term, and long-term effects of resilience building. In other words, this is a perfect opportunity for revolutionary development of new methods and tools.

Our research on societal resilience is investigating four necessary features of future community resilience models that effectively address contextual factors and predicted emergence:

1. they accurately reflect the complexity of the problem of resilience building in the desired context, taking advantage of emerging computing methods to build complex social analytics;
2. they focus a set of core constructs and measurement models that scale effectively from local to regional to national level;
3. they can be used easily in decision analysis tools that exploit emerging large volume data analysis and machine learning algorithms; and
4. they provide predictive guidance to community planners on likely outcomes of community redevelopment projects including associated stress scenarios.

Each one of these features is a need driver in the opportunity space of next generation social analytics. However, the current state of research in this domain is not taking full advantage of emerging capabilities. Community development practice still recommends reduction into a few simple to understand (by stakeholders) measures. As a result, the complexity of the environment is lost and the effectiveness of the intervention becomes a debate. At the national level, planners still struggle to find measures that are meaningful at both local and national levels [7]. In today's era of big data analytics and social network analysis, much richer measures and deeper understanding of results are possible. Our community resilience case study includes a complex structural equation model (SEM) that relates over 130 human capital development measures to measures of critical infrastructure redevelopment [2]. This model is novel in that it captures a rich representation of the combined constructs of standard of living and SWB in the context of city infrastructure change. A challenge problem for next generation social analytics is to model the optimization of these disparate measures and predict likely outcomes of interventions in decision analysis tools used in participatory community design.

C. We Should Not Miss Out on the Future

The latest calls from researchers and city planners for simplified measures and independently defined interventions in resilience development continue [8]. This viewpoint represents continuation of business as usual in the social sciences. Studies will continue to be conducted with students or other convenient samples of small size, and results will continue to lack robustness. And city planners will still be searching for tools that help them predict the broader impact of their designs. A significant program that develops and tests the next generation of social science and social analytics methods and tools is sorely needed.

Funding for machine learning and data analytics is exploding, and true cross-disciplinary research is needed to meet the scale necessary for social science to succeed. Researchers in engineering, science, and computing disciplines are finding large and varied uses for these new technologies in socially related problem spaces. These communities are and will continue to study social phenomena themselves, and will attract funding and other forms of support in part due to the lack of acceptance of next generation technologies and methods of the established social sciences.

As a result, some of the core challenges of the human condition will continue to evolve without the benefits that rigorous science in the social domain could potentially provide. Community resilience is just one area. Throughout the developing world, and in many population segments in the developed world, such problems as infant mortality, vulnerability to crime, malnutrition, unemployment, financial insecurity, and mental illness remain rampant. Vulnerable populations continue to be at higher risk in terms of health outcomes, economic outcomes, and social outcomes because of these problems. Policy makers might well be willing to help alleviate those problems if only they could get guidance on the steps to take. Widespread adoption of vaccination for childhood diseases occurred once medical science was able to identify and understand the disease and to develop effective methods of prevention. Until then, policy makers were divided on approaches to address such problems as polio. Once an effective vaccine was developed, policy fell in line, and those problems were greatly reduced. Similar advances are needed in the social sciences for the social problems that plague humanity across the globe. Until then, social science will continue to have a diminished place in the public forum.

The connectivity and ability to access data in today's internet connected society, along with continually evolving solutions to make that access more broad and agile, creates a huge opportunity space. However, this must be addressed broadly across the social science community as an opportunity to transform methods and tools, not just enable individual studies. In this paper, we describe a conceptual approach and a computational platform that is intended to facilitate conduct of social science research in the age of "big data". Section II provides a general overview. In Sections III and IV we present a complex model of individual and societal well-being and the computational tools that support testing and extending that model. In Section V we briefly discuss the various issues and questions surrounding conduct of such research. Then, in Sections VI – IX we describe large-scale data sharing and analytic platforms that are emerging in other disciplines and discuss how they could be used in social science. Section X summarizes our conclusions.

II. NEXT GENERATION SOCIAL ANALYTICS

All sorts of new human social and behavioral data are now available, and on unprecedented scales. Of course, social scientists still rely heavily on traditional sources of social and behavioral data such as in-person, telephone, or computer assisted interviews, questionnaires and survey instruments, and

sources of "thick descriptions" [9] of human behavior compiled from ethnographic or anthropological observation research. However, new sources of human social behavior data are now available due to our increased use of mobile phone, GPS technology, and personal wearable technology (such as fitness trackers), as well as the digital traces of technology-mediated communications and online social interactions. These new data sources will allow researchers to conduct human social analytics for extraordinary levels of insights ranging from intra-individual scale investigations, through inter-personal and group level interactions, to organizational and even population scale research. Over the next 25 years (a generally accepted duration of a generation), social scientists and data analysts will need to modernize their ways of thinking about and interacting with human behavior data, else risk their research becoming obsolete and irrelevant.

The research goal is to address issues facing the next generation of social data scientists. In Section III, we present our case study, in which we progress beyond simple representations of human social behavior by constructing a complex model of individual and societal well-being. We describe the integration and analysis of data of varying forms, collected via diverse methods from a variety of sources by different groups, consisting of varied units of measure, spanning a temporal range of more than 40 years, and representing human behavioral data at disparate scales. In short, we present a case study of blending quantitative, geospatial, and temporally diverse data for the purpose of advancing human social analysis for an inherently qualitative construct using a more complex (and, we argue, more representative) model of human social behavior.

In Section IV, using the case study, we describe how new methods borrowed from the field of computer science can be leveraged to support next generation human social analysis of qualitative data. Computational natural language processing (NLP) and statistical machine learning (ML) techniques have the potential to be extremely useful for blending *thick data* (which is most commonly qualitative in form: e.g., descriptive text, audio, imagery, video, or similar multimedia) with the concepts of *big data* (typically more quantitative in nature). Here, we discuss three specific "tools" that embody NLP and ML techniques to support large-scale human social analysis on qualitative data. The first tool, called VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner), provides researchers the ability to quantify both the direction (positive or negative) and magnitude of affective expressions in textual documents ranging from word-level to tome-level scales, processing millions of sentences in a matter of seconds [10]. The second tool, CASTR (**C**ommon-ground **A**cquisition for **S**ocial **T**opic **R**ecognition), produces supporting text-based information needed to establish so called *common ground*, whereby sharing mutual facts and knowledge generally facilitates faster, better understanding [11][12]. The third tool, EAGLE-ID (**E**thnicity, **A**ge, **G**ender, **L**iteracy/**E**ducation **Id**entifier), automatically aids in characterizing demographic features of individuals based on social profile data. Finally, we discuss how digital crowdsourcing economies such as Amazon Mechanical Turk (a massive, distributed, anonymous crowd of individuals willing to perform human-intelligence micro-tasks for micro-payments) can be leveraged as a valuable resource for the next generation of social science research and practice [13].

In Section V, we discuss several open questions with regards to human social analytics, including those related to ethics, data ownership and use, and personal privacy concerns. We then look at the concept of federated data platforms to accelerate the social science community development and learning of these new constructs.

III. Increasing Representational Complexity of Data Models for Human Social Analytics

Traditional social scientific models of human behavior are often over-simplified representations of what in actuality are very complex aspects of the world. Human social analytics in the next generation will need to embrace more multifaceted representations of human behavior with more complex models. Such models will need to integrate data of disparate forms, using disparate units of measure, collected from disparate sources, at disparate scales. In this section, we contribute an example in which we develop a complex, system-of-systems representation of societal well-being.

A. From Simple to Complex Modeling of Well-being

Individual and societal constructs of well-being are well established in traditional social science and economic literature as a person's assessment of their own general *happiness* and overall *satisfaction* with their personal life [14][15]. Following from [16], we further posit that happiness and satisfaction are themselves complex social constructs, which holistically comprise four principal constituents:

1. Affective Experiences: the longer-term experiences of pleasant affect (as well as a lack of unpleasant affect) as indicated, for example, via their general perceived happiness in life, in their marriage, and with their cohabitation companion (e.g., partner or roommates).
2. Global Life Judgements: a person's overall belief regarding how interesting they find their own life in general (e.g., whether they consider life to be dull, routine, or exciting), as well as a judgement about the general nature of humanity (whether they believe most other people to be trustworthy, fair, and helpful).
3. Cognitive Appraisals: a person's subjective self-assessment of their own current socioeconomic state relative to their life goals, as well as broader social comparisons. Determinants include financial status self-appraisals, social status self-appraisals (e.g., social rank and social class), and self-appraisals regarding their health, the relative quality of their domicile, and aspects of the city in which they reside.
4. Domain Specific Satisfaction: the degree of fulfillment or contentment with important social elements such as satisfaction with their family life, friendships, hobbies and recreational interests, job/career, and their wages.

Traditional social analytics tend to focus on a narrowly scoped subset of the above constituents. While such studies do provide useful insights, they are limited precisely because they

are narrow; due to the inherent interconnectedness of these constituents, complex interactions abound. Nevertheless, they hold much greater analytical value when they are considered in conjunction with one another. The whole is greater than the sum of its parts, and aggregate-level insights may never emerge unless and until the underlying relationships are expressly represented.

To this end, we present an example in which we incorporate 130 different manifest indicators for- and correlates of- individual and societal well-being. This is represented in the "Community Population" oval of Figure 1. To do so, we blend qualitative, quantitative, geospatial, and temporal data from several sources. While detailed model specification is beyond the scope of this paper, we find the model useful as a reference for discussing next generation social analytics.

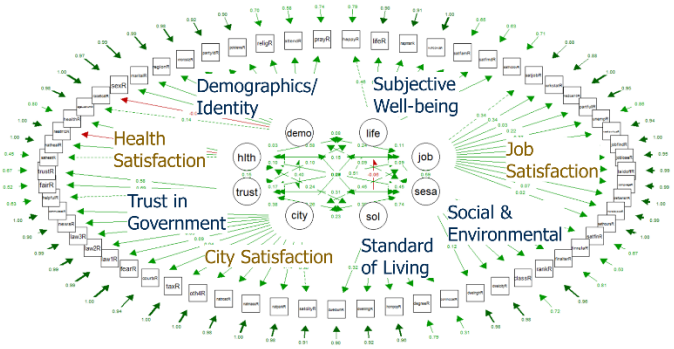

Figure 1. Complex Model of Well-being.

B. *Blending Qualitative, Quantitative, Geospatial, & Temporal Data*

The data for our complex model of well-being are drawn from several public data sets comprising records from 30 different collection activities spanning 42 years (from 1972 to 2014) across nine different divisions of the United States Census Bureau [17]. This data integrates 25 manifest indicators of societal well-being, organized into latent variable constructs representing the four principal constituents described in Section II-A. An additional 17 indicators provide data providing more objective measures of individual *quality of life and standard of living*, such as highest education level attained, number of people living in a household, type of dwelling (and whether owned or rented), various employment characteristics (part time, full time, student/homemaker, unemployed, retired, etc.), and constant (i.e., annual inflation adjusted) income in dollars. Also included are data capturing information about each respondent's *demographic* details, the *general political climate* (public opinion regarding amount of taxes paid, the efficacy of the courts, and national programs related to healthcare, transportation, and public transit), established local and regional *geographic boundary data*, annually recorded data regarding the *general economic climate* of the nation (such as inflation rates, consumer price indices, prime lending rates, and annual gross domestic product (GDP) per capital growth), and data characterizing the *general security climate* (e.g., individual and community exposure to crimes, perceptions of fear, etc.).

As one might imagine, the data are operationalized in multifaceted ways, taking multiple forms, units, and scales of measurement. In all, we integrate data from nearly 60,000 respondents spanning 42 years with regard to 130 different variables of interest, where each variable puts (on average) potentially 7 unique degrees of positive or negative pressure on individual and/or societal well-being. All told, this leverages approximately 55 million data points for our model, allowing for a very rich and complex representation of well-being – much more sophisticated than many other typical, prevailing social science models.

We argue that this representation, as opposed to a simpler model (for example, one based primarily on measures of *happiness*) is a more accurate reflection of true societal well-being. To illustrate this point, consider Figure 2, in which we visually depict how a simplistic representation of well-being (happiness scales) compare to a more complex representation of societal well-being for different geographic regions in the United States. Different insights emerge (especially in the southern regions) when a more complex construct capturing affective experiences, global life judgements, cognitive appraisals, domain specific satisfaction, objective socioeconomic quality of life and standard of living data, the general political climate, general economic climate, and the general security climate are incorporated when considering societal well-being.

We can also demonstrate how the model produces interesting insights in relation to political aspects of the national population, especially when considered in conjunction with temporal information. For instance, in Figure 3 the scatterplot dots indicate national-level averages for each year of data collection (1972-2014) for each self-identified political

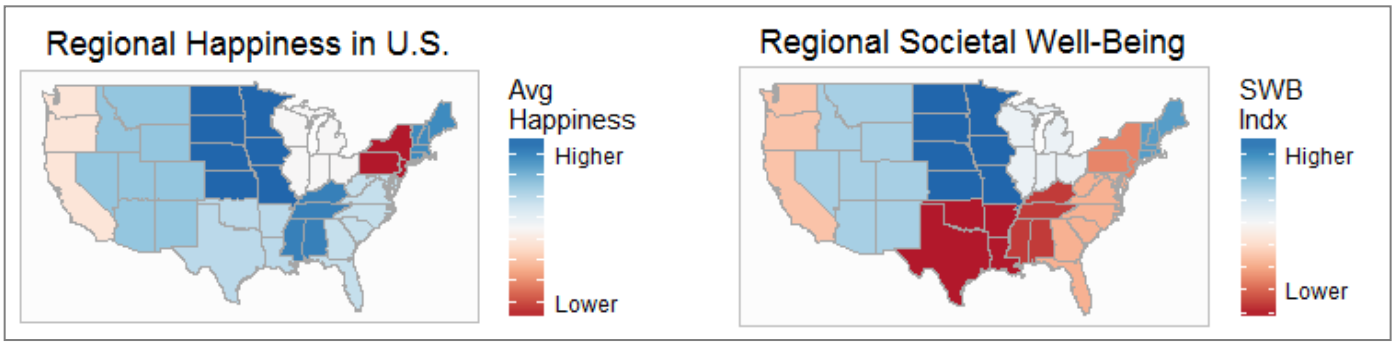

Figure 2. Comparing a simple representation of well-being (happiness scales, on left) to a more complex representation of societal well-being (on right) to derive different insights for different geographic regions in the United States.

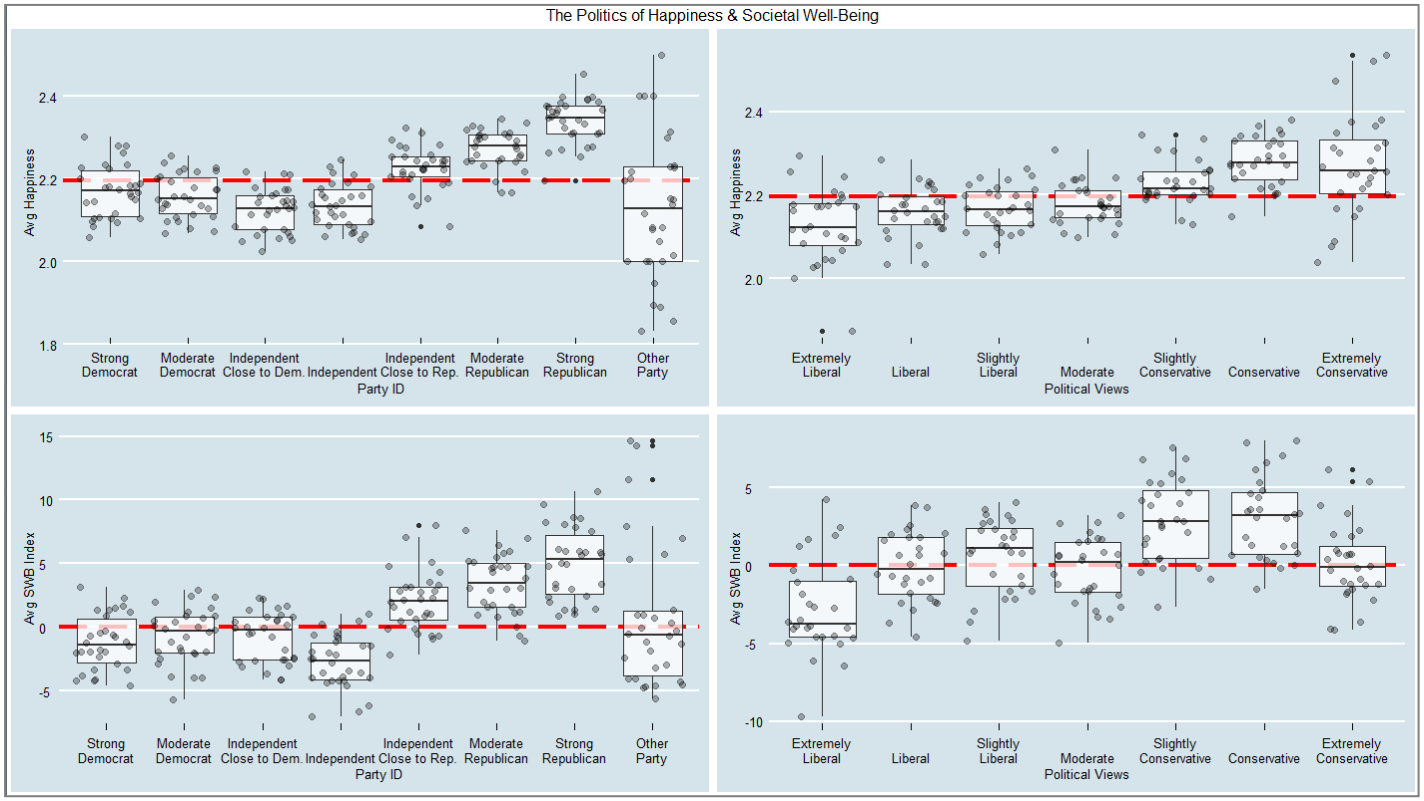

Figure 3. Aggregates of temporal data for political party and ideological views for a simplistic model of happiness versus a complex model of societal well-being.

community as measured by party affiliations in the left column plots (Party ID) or by ideological views in the right column plots (Political Views). The simple model of happiness (Avg Happiness) is plotted in the top row and the complex model of societal well-being (Avg SWB Index) is plotted in the bottom row. Boxes depict the middle fifty percent of the data (with mean lines) within each category, and whiskers show the range from minimum to maximum scores. The red dashed horizontal lines show overall means (across all categories). Especially interesting is how robust the results are for individual constructs; the general trends are qualitatively similar regardless of whether modeled with simplistic or complex representations of well-being.

C. *Monte Carlo Simulations and Predictions of Well-being*

The complex model, once derived as described in the previous section, may be used in Monte Carlo processes to explore the probability distributions associated with how potential changes in any subset of the input variables would impact societal well-being. The model can be extremely useful, for example, to government policy decision makers when the impacts of their decision alternatives could be vetted within a data-derived, model-driven trade space analysis tool. For example, Monte Carlo simulation modelers would be able to reliably quantify the effect that policy and funding decisions might have on societal well-being. Such considerations will enable next generation social analytics to generate better predictions, going beyond the prevailing social science policy of typically concluding a study upon reporting descriptive and inferential statistics.

IV. Methods, Techniques, and Tools for Next Generation Social Analytics of Qualitative Data

Next generation social scientists will also face issues related to developing methods and tools to help facilitate the collection, processing, analyzing, and visualizing of such multifaceted social data in near real-time. Our example model of individual and societal well-being is based on a static data set collected over many years. It is extremely valuable for generating structural equation models representing the interdependencies among the related input variables, and for paving the way for exploratory and predictive analyses.

Given the vast amount of qualitative data available in social media platforms such as Twitter, Facebook, and a host of blogging and microblogging technologies, it is possible to create “social sensors”, which monitor important indicators of societal well-being, on massive scales, in near real-time. Traditional social science methods rely on labor and time intensive qualitative data analysis techniques to transform qualitative data into quantitative representations of affect (e.g., manually reading and coding individual text entries to determine if a person is expressing positive or negative affect). In contrast to most typical quantitative methods, qualitative data analysis methods do not easily scale up. Datasets are too large (consider the entire internet of social media, SMS/text messages, emails, blogs, etc.), and they are produced at extreme velocities (e.g., 500 million tweets per day, or status updates from 1.8 billion active Facebook users per day [18]). It is impossible for human researchers to even look at all the data, much less analysis it in a timely manner.

Whereas previous generations of Computer Assisted Qualitative Data AnalysiS (CAQDAS) software supported the traditional toolkit of qualitative researchers, i.e., sorting, searching, and annotating, the newest generation of tools is adding features powered by computerized natural language processing (NLP) and statistical machine learning (ML) techniques to enable automated rapid, massively large scale assessment of digital text, audio, video, and other multimedia traces of people's affective experiences as portrayed in their social media posts. The norm for next generation social analytics will be to employ such computational tools to facilitate blending of social media *thick data* (rich, descriptive qualitative data) with *big data* (i.e., data that is characterized by massive volume (amount of data), velocity (speed of data in or out), and variety (range of data types and sources)).

A. VADER: Automated Analysis of Affect in Social Media

VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner) [10] is a computational tool for conducting automated large scale sentiment analysis [19][20]. Sentiment analysis is useful to a wide range of problems that are of interest to next generation social analysts, practitioners, and researchers from fields such as sociology, marketing and advertising, psychology, economics, and political science. The inherent nature of microblog content - such as those observed on Twitter and Facebook - poses serious challenges to practical applications of sentiment analysis. Some of these challenges stem from the sheer rate and volume of user generated social content, combined with the contextual sparseness resulting from shortness of the text and a tendency to use abbreviated language conventions to express sentiments. VADER is a simple rule-based algorithm and model for general sentiment analysis. In previous work [10], we compared VADER's effectiveness to eleven typical state-of-practice benchmarks for automated sentiment analysis, including LIWC [21][22], ANEW [23], the General Inquirer [24], SentiWordNet [25], and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms. We used a combination of qualitative and quantitative methods to produce, and then empirically validate, a *gold-standard* sentiment lexicon that is especially attuned to affective expressions in microblog-like contexts. VADER combines these lexical features with consideration for five generalizable rules that embody grammatical and syntactical conventions that humans use when expressing or emphasizing sentiment *intensity*. We found that incorporating these heuristics improves the accuracy of the sentiment analysis engine across several domain contexts (social media text, NY Times editorials, movie reviews, and product reviews). Notably, the VADER affective sentiment lexicon performs exceptionally well in the social media domain. The correlation coefficient shows that the VADER computational engine performs as well (r = 0.881) as individual *human* raters (r = 0.888) at matching ground truth (i.e., the aggregated group mean from 20 human raters for sentiment intensity of each text-based affective expression). Surprisingly, when we further inspect the classification accuracy, we see that VADER (*F1* = 0.96) actually even outperforms individual human raters (*F1* = 0.84) at correctly classifying the sentiment of tweets into positive, neutral, or negative classes.

B. CASTR: Aid to Automated Topic Models of Social Text

CASTR (**C**ommon-ground **A**cquisition for **S**ocial **T**opic **R**ecognition), produces the supporting text-based information needed to establish so called *common ground*, a well-known construct from psycholinguistics whereby individuals engaged in communication share mutual facts and knowledge in order to be better understood [11][12]. CASTR is intended to aid in *computational topic modeling* [26] by automatically acquiring this background knowledge.

Computational topic modeling techniques are used to uncover the hidden, or latent, concept-based semantic structures (i.e., topics) within text documents. Topic modeling is useful for a broad collection of activities, from automatically tagging newspaper articles with their appropriate newspaper sections (e.g., sports, finance, lifestyle, etc.) to automatically clustering like-minded social media users into groups based on the similarity of their expressed interests. Unfortunately, however, these automated approaches will sometimes infer topics that match poorly to – and are less semantically meaningful than – human inferred topics [27]. The issue is compounded when mining so-called *social text*, i.e., sparse text produced explicitly for informal social consumption (e.g., via social media, instant messages, SMS/texts, personal email, and so on where people rely on one another's common knowledge, rather than extended textual documentation, to understand intended meanings). In designing and developing CASTR's algorithms, we qualitatively assess the unique characteristics of social text, which present challenges to computational topic models, and which are not prevalent in other typical (non-social) text corpora like newspaper articles, scientific publications, or books. We find that a) constraints imposed by typical social media technologies, b) implicit social communication norms, and c) evolving conventions of use often confound typical computational topic modeling techniques for social text. For example, tweets are much terser than other kinds of text documents, and this sparsity is troublesome for computational topic modeling algorithms that perform posterior inference of the text. Also, tweets are often laden with a great deal of social communication "noise" (such as emoticons, emojis, hashtags, and URL links) that confuse computational models, and yet present very little trouble to humans.

CASTR leverages the concept of common ground to present a theoretically informed social and cognitive psychological framing of we refer to as the "human interpretability problem" as observed in computationally-produced topic models of text mined from social media. Additionally, CASTR employs a well-established theory from the field of Human-Centered Computing, namely Distributed Cognition (DCog) [28][29], as a basis for mitigating the issues of developing common ground for computational topic modeling efforts. DCog is a theoretical perspective that proposes knowledge and cognition are not confined to any single individual or referent resource; instead, they are distributed across individuals, objects, artefacts, and tools in the environment, and constructed in context.

As an example of how CASTR implements the DCog inspired mitigation strategies, consider a fictitious (but

representative) social media post that expresses a person's positive affective experience related to attending a musical concert at a popular venue near Atlanta, Georgia: "*Headed to Stone Mountain to see the Rolling Stones. Mick Rocks! www.rollingstones.com/band/ #StonesOnFire*". Although it is a relatively simple thing for humans to immediately understand the meaning of this social text (most Americans know who The Rolling Stones are, most people from Georgia know what Stone Mountain is, and most people understand what it means when "rock" is used as a verb in this context, even if they are not immediately sure who Mick refers to, and most people recognize the conventional use of hashtags, as well as URL links). However, the shared, socially constructed knowledge (common-ground) necessary to understand the intended meaning of the above example social text is often not readily available to computational topic models.

CASTR automatically retrieves the (previously missing) background distributed knowledge about key words, phrases, and named entities (proper nouns) within the terse text, and provides this information to the computational topic model processes. The result is a much more accurate representation of which topic(s) a particular short social media document should be belong. For example, the social text above would be appropriately grouped with music and entertainment related topics, rather than geological science related topics.

C. EAGLE-ID: Automated Demographic Profiling

EAGLE-ID (**E**thnicity, **A**ge, **G**ender, and **L**iteracy/**E**ducation **Id**entifier) automatically aids in characterizing important human social demographic features based on social media profile data. The EAGLE-ID system consists of software (currently in beta stage) that performs automatic classification of a person's ethnicity (given the person's surname), their likely age range and gender (based on their first name), and their literacy and education level based solely on information mined from the person's digital social media data (including user profile data as well as shared content). The majority of this is done via text-based computational linguistic processing (in conjunction with comparisons to data from the U.S. Census Bureau database, Social Security Administration records, and U.S. Dept. of Health and Human Services data), but it also uses computer vision for image processing on profile pictures to boost ethnicity/age/gender classification accuracy.

In addition to the obvious uses for user profiling and user modeling, the EAGLE-ID software could be useful for automatically collecting and associating demographic information with particular social media accounts. When used in conjunction with VADER and CASTR, EAGLE-ID facilitates rapid, large scale analysis of social data for use in real-time monitoring of individual and societal well-being with realistically representational complex models.

While the design and development of tools such as VADER, CASTR, and EAGLE-ID is not necessarily in the direct purview of social science, the employment and use of such tools will almost certainly be a significant part of next generation social analytics. It is already a major part of the new field of Computational Social Science. Eventually, the word "computational" will be dropped, and methods, tools, and techniques like the ones discussed in this section will be commonplace in social science research – integrated into social science education right alongside experimental study design, research ethics, and statistical analysis.

D. Crowdsourcing for Scaling-Up Qualitative Data Coding

An interesting interim step preceding fully automated artificial intelligent machine learning algorithms for conducting large scale qualitative data analyses are the emergence of digital crowdsourcing economies such as Amazon Mechanical Turk. These platforms are typically comprised of a massive, distributed, anonymous crowd of individuals willing to perform general human-intelligence micro-tasks for micro-payments, and they can be leveraged as a valuable resource for the next generation of social science research and practice. Indeed, in the past half-decade, Amazon Mechanical Turk has radically changed the way many social science scholars do research. The availability of a massive, distributed, anonymous crowd of individuals willing to perform general human-intelligence micro-tasks for micro-payments is a valuable resource for researchers and practitioners.

In other work [13], we addressed many of the challenges facing researchers using crowd-sourced platforms. Particularly, we reported on how to better ensure *high quality* qualitative data annotations for tasks of varying difficulty from a transient crowd of anonymous, non-experts. Crowdsourcing has already had a significant impact on social analytics, and we believe it will continue to play a substantial role in the next generation of social analytics.

V. From Traditional to Next Generation Social Analytics

The complex model of well-being described earlier differs from traditional social science in several meaningful ways:

1. *Representational complexity*: In next generation social analytics, model complexity will increase beyond what is typical for much of social science research today. Our example integrates more than 130 indicators for- and correlates of- individual and public well-being. These data are garnered from many sources, measured in numerous different units, stored using many data types at different scales representing individuals, communities, and entire societies. Just as other disciplines such as systems engineering, economics, and computer science have embraced the notion of incorporating "big data" into their typical data models, the next generation of social analytics will need to likewise expand their scope such that social analytics like the ones we illustrate are the norm, rather than the exception.
2. *Large-N and Multiple-T*: In order to achieve useful statistical power while incorporating the expanded scope resulting from increased representational complexity, and at the same time preserving broad generalization and application capacities, next generation social analysts will need to design and conduct studies with much larger sample sizes (i.e., "Large N" studies) collected over multiple instances in time (i.e., "Multiple T", or longitudinal studies). In our example, we integrate data

from nearly 60,000 respondents spanning 42 years with regard to 130 different variables of interest, where each variable puts (on average) potentially 7 unique degrees of positive or negative pressure on individual or societal well-being. All told, this leverages approximately 55 million data points for our model. Such study designs will eventually become more prevalent for social analytics.

3. *Extending exploratory and predictive analytics*: Our example model lays the foundations for predictive analysis (e.g., via Monte Carlo simulations), which would be extremely useful to government policy decision makers because the impacts of their decision alternatives could be vetted within a data-derived, model-driven trade space analysis tool. For example, we would be able to answer important questions such as: *in order to improve overall community/public well-being, should government decision makers invest tax dollars in a better public transportation system, economic development program, roads, schools, or security services*? Such considerations will enable next generation social analytics to generate better predictions, going beyond the prevailing social science policy of typically concluding a study upon reporting descriptive and inferential statistics.

Combining the increase in representational complexity with the methods, techniques and tools, a vision of how next generation social analytics will be conducted begins to emerge in which large-scale, individual and national-level, near real-time analysis of the following are common:

- social media data
- mobile and GPS technology data
- personal wearable technology data
- internet of things data

We outlined how new tools and techniques could be leveraged to marshal in the next generation of qualitative social analytics on heretofore unprecedented scales. VADER provides researchers the ability to automatically quantify both the direction (e.g., positive or negative) and magnitude of affective expressions in textual documents ranging from word-level to tome-level scales. In a matter of seconds, VADER is capable of automatically transforming millions of rich qualitative social media documents (e.g., tweets) into quantified measures of positive and negative affect for a given Twitter user. This capability alone allows us to produce a simple representation of well-being on a national scale in near-real time [10]. When we combine it with the ability to also understand the topic towards which the affective expressions apply, we can begin to incorporate other elements of the more complex representation of well-being previously discussed.

For example, consider when a Twitter user laments (or praises) aspects of her job, her health, her family or friends, her city/community, or her financial situation. Or consider how often she might express satisfaction (or dissatisfaction) for aspects of the general political, security, or economic climate of her community or nation. Now consider how prevalent such expressions are in aggregate for all Twitter users. Next think about how many other publically available forms of such data currently exist (other social networks like Facebook and Snapchat, place-based platform Foursquare, review platform Yelp, internet chat rooms, topical blogs, and discussion forums such as Reddit). Next generation social analytics should embrace such resources, as well as the tools needed for analyzing them at internet scale.

Typically, these social media data are time-stamped, so that temporal aspects can be incorporated (c.f., [30]). Slower changing data variables such as a person's demographic characteristics (e.g., ethnicity, age, gender, literacy and education level) can also be automatically extracted from a person's social media data. In many cases, these data can be combined with meta-information regarding the geolocated origins of the content producers, or otherwise merged with GPS, mobile, or other location-aware wearable technologies. Real-time assimilation of national, regional, or local unemployment rates, crime data, housing market data, inflation, consumer price index, prime rates, and gross domestic product round out the capability to produce timely, realistically complex models of societal well-being.

To achieve the vision of next generation social analytics, further research is needed in the following areas:

1) Model Complexity vs Model Interpretability

Increasing representational complexity in the way we discuss in Section II, while more characteristic of real-world human social behavior, is not devoid of its own issues; complex models are by their very nature more difficult to interpret. We offer a brief discussion of three avenues for mitigating the challenge of interpreting complex models. First, social science data analysts will need simple and intuitive interfaces for exploring the trade-space of the data. Such tools will increasing model transparency, and incorporating interactive data exploration will aid analysts in easily and quickly uncovering complex interrelationships within and among the variables of any complex model. Second, analysts need simple interfaces that allow them to rapidly build and assess Monte Carlo simulations regarding how potential changes in input variables impact selected response variables of interest. Third, advanced interactive data and information visualization tools will be critical for next generation social analytics to make sense of data at varying levels of aggregation and combination.

2) Ethical Considerations of Widespread Human Social Data Analytics

Ethical considerations related to privacy and confidentiality are often cited when human social analytics are discussed. Privacy (not collecting data that is not needed for the study) and confidentiality (protecting identifiable information from inappropriate dissemination) are fundamental principles of ethical research with human subjects. These principles must find new implementation when the context of research is large, shared data sets. By extension, as on-going studies continue (including longitudinal studies), mechanisms for individuals to monitor how their data are being used, and to have appropriate safeguards, must be developed.

Other issues include data ownership and potential for financial gain – both for individuals (about whom the data are

collected) and for institutions that otherwise possess the data of interest. Owning institutions must take care as data is updated over time that it does not become used or cited for purposes that are outside the agreed upon collection context, lest the whole dataset becomes discredited. Possible financial gain suggests possible financial loss, perhaps from liability that might arise from compromise of privacy or confidentiality, or perhaps from errors in algorithms or in other study methods. These issues, and others that arise from them, deserve careful attention, but are beyond the scope of the present paper.

3) Skill Sets and Education for NGSA

We must educate and train the next generation of social data analysts to be comfortable embracing representational complexity and incorporating methods, tools, and techniques like the ones discussed above. It will need to become standard parts of social science education, integrated into social science curricula right alongside research methods and experimental study design, research ethics, and statistical analysis.

4) Collaborative study and experimentation

We must build platforms where social scientists can come together and conduct joint experiments or related experiments in common contexts with next generation methods and tools. Such platforms are becoming a central component of biomedical research, and are expanding into other fields as diverse as international affairs, materials research, and system design. Digital network technologies supporting cloud computing, federated data architectures, knowledge graphs, data mining and machine learning, standardized web ontologies, digital annotation, experimental workflow sharing, computer visualization, crowdsourcing, and computer gaming are creating unprecedented capability for shared study of social behaviors. Emerging shared data experimentation platforms will provide a means to transform access to and sharing of social science research and social data analytics.

VI. NEXT-GENERATION RESEARCH FEDERATIONS

Although data sharing platforms like Harvard Dataverse are available to share the detailed results of scientific studies, in this section we discuss the idea of federated data models for experimentation – platforms that allow geographically dispersed cohorts of researchers to work together on scientific experiments around a common problem or area of study. To our knowledge such platforms have not yet entered use in the social sciences community. We discuss the challenges and opportunities associated with an experimentation platform concept, methodologies that can support development of such platforms, and an example case where a shared experimentation platform would be useful.

Unlike many other scientific areas of study, social situations represent complex adaptive systems that are characterized by independent agents who self-organize, adapt, and learn. In complex adaptive systems, broadly applicable models of behavior are difficult to generalize. The situation under study and the context of the situation must be studied together, and generalization across multiple contexts is not always wise or possible. Adaptation often makes generalized results short-lived. Intervention in social situations focuses heavily on causal relationships, but generalizing to purely linear causal relationships is often unsuccessful. Study of such systems must eventually account for *linear causal* relationships and also *circular causal* relationships, self-organization or *adaptive causal* relationships, and *reflexivity*, which acknowledges the act of studying the system can effect causal relationships [31]. Generalization of results using linear regressions is most common and appropriate, but can only be accomplished by applying assumptions with respect to the other three causal models that are often not captured with the data. These assumptions are often about which of a number of potential causes aggregate to larger populations, making explanations of causality difficult.

Because of such "shifts in causality," reduction to linear models makes the generalization of effects across multiple contexts difficult. They can also limit the reproducibility and replicability of social science study [6]. Issues related to reproducibility can be reduced by use of common datasets with access to original study data, models, and tools. Study generalization requires access to sampling methods as well as both positive and negative results, and more difficult, the original assumptions and abstractions used by the researcher to conceptualize the study. However, because many of these assumptions are related to selection of causal factors, effective conceptual models that capture context in the form of broader causal factors with hypotheses related to context-specific selections can help. The ability to do this has been until recently limited by the time and effort required to collect and analyze data, a condition which is changing rapidly.

Designing data analytic and computational models that accurately reflect performance measures at different layers of society, and the aggregation of measures from one layer to the next, is the primary conceptualization problem in social analysis and policy practice. Behavioral aspects of complex sociotechnical systems can be influenced at any layer of the system, but initiatives that try to analyze and improve factors at one level do not necessarily translate into positive influence at other layers. Moreover, the timeframes for measuring effects can vary greatly across different factors and societal layers [32] [33]. Lack of common methods and tools to define model abstraction and aggregation of data create further barriers to generalization, which tie back to the original conceptualization of the study and related selection of constructs and dependent variables.

Figure 4 places our complex model of societal well-being in the context of a city, where the built environment, institutions, and shared infrastructure provide the capital necessary for people's livelihoods. This model expands the total dataset required to evaluate well-being tremendously, and also introduces causal feedback into the model of well-being. This is a complex adaptive system that can be explored via complex models but will have no deterministic solution sets. Issues and concerns with use of data analytic methods in social experiments reflect the complex adaptive systems aspects of social phenomena like this. These include determining appropriate context, understanding both linear and non-linear causality, representing differing time scales, uncertainty about what constitutes entities that affect the system, and issues with agency or agent identification [34]. These can be overcome by

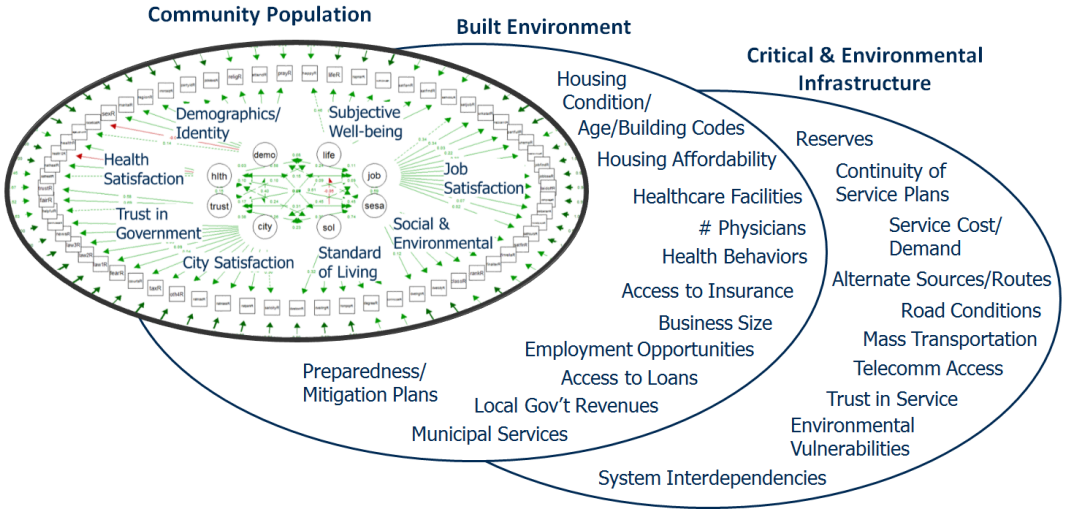

Figure 4. Complex Model of Human Capital in a Community Population situated in the Context of City Built Environment, Institutions, and Critical & Environmental Infrastructure.

viewing the social problem of interest as a system then conceptualizing both the problem system and response system as a set of conceptual and then dynamic models. Research related to enterprise systems of systems and sociotechnical systems analysis introduces a methodology to address these issues.

Shared experimentation implies agreement on paradigms that reflect the problem definition and contexts of interest, as well as the semantic descriptions of the sociotechnical system of interest, and the conceptual model of the current systems' behaviors and future states. The concept of an experimentation platform implies a set of methods and tools to define and address these agreements, which we discuss prior to descriptions of the tool framework.

In Section VII, we introduce the concept of an experimentation platform, using references from a United States Air Force concept as an appropriate framework for this application. We describe emerging computer platforms that make this concept a viable approach, and a methodology for building community-wide models in these platforms. In Section VIII, we describe the characteristics of a tool platform for experimentation, and the technological approaches that might be used to build it. We do not at this point describe a complete toolset, but a call for research to create these tools. In Section IX, we discuss early work in next generation social science study design tools necessary to complete the experimentation process.

VII. Experimentation Platform Concept

A shared data federation combined with a shared research and experimentation platform can serve to rapidly distribute knowledge and accelerate the development of new knowledge in scientific study. The concept of "System Level Experimentation" combined with next generation analytics is an approach that has not been explored yet in the social science domain, but is gaining prevalence in other areas of study. We discuss this first as a conceptual platform, then describe some of the emerging technology that can be used for implementation in the social science domain.

A. System Level Experimentation

Alberts et al. [35][36] captured a useful vision for information age transformation of social theories and related analytics in pursuit of a set of methods we refer to as System Level Experimentation. The authors define this as a "campaign of experimentation," or a "set of related activities that explore and mature knowledge about a concept of interest." Although developed as an approach for transforming military command and control, the general model of such a campaign provides a framework for joint experimentation in any social decision making domain. The framework is a scientific method for experimentation, which includes theory development, conceptualization or conceptual modeling, formulation of questions and hypotheses, collection of evidence, and analysis. The approach views system transformation as a campaign of multiple experiments that produces a body of knowledge that creates a foundation for future experiments. Such campaigns have leaders and goals, research cohorts who use and create knowledge aligned with the goals, and a shared knowledge capture framework that allows federated cohorts and experiments against a common knowledge model.

With respect to reproducibility, repeatability, and generalization of experiments, the idea of a campaign focuses the research process on aligned goals with deliberate urgency and resource allocation. Alberts and Hayes note, "*reuse here applies to ideas, information about investigations conducted, data collected, analyses performed, and tools developed and applied. In terms of experiments, it implies replication. Reuse, and hence progress, is maximized when attention is paid to the principles of science that prescribe how these activities should be conducted, how peer reviews should be executed, and when attention should be paid to the widespread dissemination of findings and conclusions*."

The authors stress the importance of a shared conceptual model as a key to generalization, reproducibility, and replicability. Although in many scientific studies there exists a shared paradigm of study and generally shared conceptualization, this is difficult to achieve in social situations where stakeholder perspectives, even those of research communities, are difficult to align. For example the community measurement paradigm for "standard of living" is moving from a Gross-Domestic Product (GDP)-based measure of production to more representative consumption-based representations. However, the GDP measure was conceptually simple, and consumption measures are conceptually complex. Although the community is accepting the paradigm shift, there do not exist common agreed upon conceptual models of standard of living that can drive shared and replicable experimentation. A debate over the conceptualization of our complex model of well-being would be counterproductive. We need a platform where the agreed upon factors can be organized and shared, research cohorts can experiment with models and empirical study in their contexts, and the common conceptualization in terms of factors, abstractions, and weightings can be updated over time via community experimentation. Thus an effective shared experimentation platform must address common conceptualization artifacts as well as data and potentially dynamic models. Such a platform will serve both cross-sectional and longitudinal studies. Longitudinal studies conducted in such a platform will have opportunity to use dynamically-computed weightings for different data collection epochs as new information is added to the platform.

B. Emerging Data Analytics Platforms

What we can do much more easily these days is collect the data. Public datasets that report social variables in both broad and localized contexts are becoming widespread. Shared community data warehouses and models for experimentation purposes are becoming more widely used in complex health and medical studies, leading one to believe that such approaches may also have use in social research and analysis. Notable examples of medical research platforms include the Global Alzheimer's Association Interactive Network (GAAIN) [37] and the Medical Informatics Platform (MIP) of the European Union's Human Brain Project [38]. Common features of these projects include a federated data model, shared schemas or data codings, community agreed upon ontologies and sematic tagging, machine learning tools for extraction and matching of data, and web-based interfaces to data, research cohorts, and visualizations. In all such projects, a shared database is created where an entity-relationship model defines the schema of the resultant "data warehouse," and agreed upon data codings provide a map between the larger sets of data and the phenomena of interest. We will further explore the possibility of designing similar projects for social data experimentation.

To reach this point, the community must develop not just common data, but also methods for agreement on research paradigms, related stakeholder perspectives of problem and solution spaces, associated viewpoints, and shared conceptualizations. Thus long-term success in social analytics must address the capture of both the data and conceptual relationship models that make the data meaningful. These conceptual relationships are often determined using soft systems approaches, which are appropriate, but existing methods and tools do not adequately connect the conceptual artifacts with the data-driven analytics. In the social analytics field, there is a need for research that connects the resulting collected data to its conceptual model artifacts. Without this problems with abstraction, generalization, reproducibility, and replicability cannot be resolved. Research from the systems engineering community centered on management of enterprise systems-of-systems provides a set of useful methods and tools.

C. Enterprise Systems of Systems Methodology

Sociotechnical systems analysis is a specific methodology that supports assessment of multiple factors across all layers of a complex enterprise or societal construct using sets of tools derived from system science and system modeling. The methods recognize that factors arise from the interaction of many and diverse enterprises that can be defined by their entities, relationships, established processes, pursued strategies, and emergent phenomena. The sociotechnical systems analysis attempts to capture the combined conceptual, data, and analytical modeling artifacts necessary to completely describe the problem [39][40].

With respect to social situations, the method produces a set of artifacts that describe the system context and boundaries, system entities and relationships, primary construct variables, potential causal variables, and phenomena of interest. The process is conducted such that insight can be fed into dynamic computer models. Hypotheses that intervene in lower level causal factors can then be viewed as they aggregate up into larger population behaviors. The sociotechnical systems analysis produces artifacts that communicate the abstractions and aggregation of behaviors across different scales, helping to explicitly document both the assumed and modeled variables.

At the core of a sociotechnical systems model are entities and their relationships, which can be organized into associated databases and warehouses. The entity-relationship model can be created, modified, and refined over periods of short and long term study. Standardized codings of the data entities then make relevant data elements accessible to researchers and analysts. One use of this is for data collection and analysis, but the sociotechnical systems analysis methods are focused on development of experimentation platforms. Experimentation requires that not only the data but also the underlying conceptual models context of study be updated over time.

The conceptual model representations produced by the sociotechnical systems analysis serve as a bridge between the soft systems aspects of the problem (systems thinking) and the quantitative analysis approach (design). This is an area that needs significant additional research as related to methods and tool design. However, recent advances in machine learning and semantic graphs can bring the semantic model and mathematical model artifacts into the same toolsets. The bridge between the two is a conceptual model that uses semantic models to specify the analytical models. We identify these as metamodels as they should describe broader conceptual models and data, while individual experiments

explore a subset of executable models and constructs related to central questions of interest. Figure 5 describes that bridge.

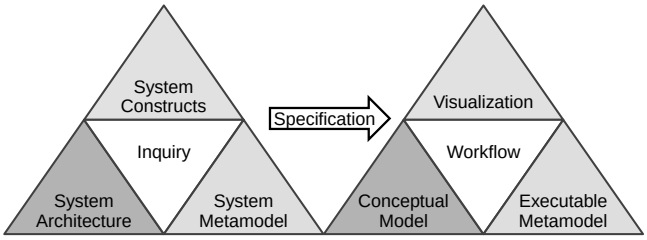

Figure 5. The bridge between soft systems analysis and social analytic model specification.

We define the soft systems aspects in Figure 5 as "*System Metamodeling*" using three fundamental abstraction approaches: system metamodels, system constructs, and system architecture models. These are determined in a participative, inquiry-based process. We describe hard system aspects as "*Executable Metamodeling*" determined by a specification and design workflow using conceptual models, executable metamodels, and data visualization. It is useful to think about this as a tool framework. The tools support structuring the systems metamodel, creating the conceptual models, creating the executable metamodels, analyzing and visualizing the decision space, and managing the contained knowledge over time [41].

The system metamodel is described as the set of constructs and rules used to define semantic relationships across information sets, associated data sets, and methodologies or processes [42]. The metamodel definition on the semantic side is an architectural description of the system using modeling views and stakeholder viewpoints. The executable metamodel is the dataset design and any associated computational models.

D. Metamodels and Federated Data Models

The emerging medical community models link together research cohorts by providing a common data model for integrating federated datasets. As experimentation platforms they provide a cohort discovery tool to link research communities, a federated data model integration architecture, and a common data visualization toolset that allows data exploration across multiple cohort data. The federated approach to data model integration allows individual cohorts to maintain their own working datasets while sharing and using data from other cohorts via a common data model representation. State of the art tools for data discovery, transformation, and integration automate most of the source data integration into the common data model. The common data model is implemented as a schema in a relational database using agreed upon codings for data tables and variables.

In a federated data model design, metadata or data descriptions are essential to data harmonization – integrating data from different sets and integrating experimental data back into the common data warehouse. Emerging data mining and machine learning tools can automate data harmonization assuming the metadata has a rich enough natural language description of the data elements to link multiple sets. Mapping variables between federated datasets and the common data model is accomplished by extracting and matching the data entities via descriptive data mapped from element descriptions in data dictionaries, a component of metadata. Adequate metadata provides a path to harmonizing the often cryptic tags placed on data elements in databases. Transformation tools are provided to map data between the common model representation and federated datasets [43].

The conceptualization of most existing common data model examples were developed initially from manual coding and integration of existing datasets [44][45]. In the social analytics area, a common conceptual definition of the data tables and entities would be a huge undertaking due to the tremendous differences in terminology, conceptual data relationships, and assumptions made around data generalizations across societal scales. Emerging approaches

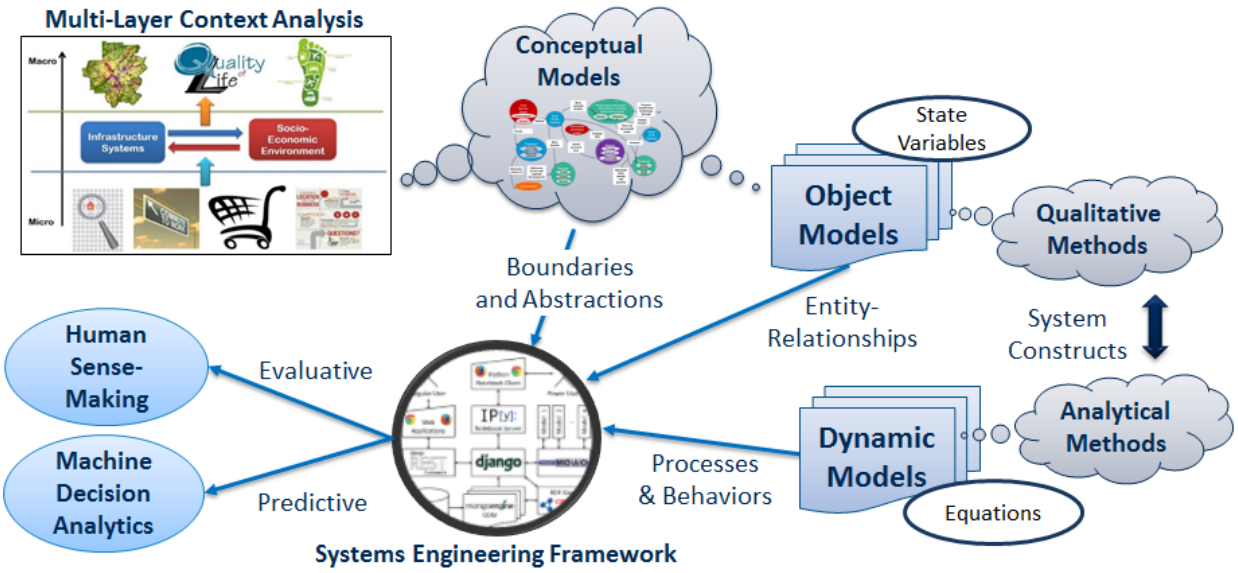

Figure 6. Conceptual Architecture Blending Qualitative and Quantitative Models into a Single Platform.

for graph representation of data entities and relationships should be explored in the social sciences arena as a tool for amassing large volumes of linked data and knowledge supporting both generalized and contextual research results.

VIII. SOCIAL EXPERIMENTATION TOOL FRAMEWORK

We present a generalized concept for social experimentation and analytics using both bottoms-up software environment and top-down conceptual architecture descriptions. The purpose of this discussion is not to present the design of an existing tool (none exist), but to describe the characteristics and architectural constructs of future frameworks for social experimentation and analysis. Figure 6 presents our high level system and process architecture.

Alberts et al. note that "*For purposes of building knowledge, the most important elements are (1) consistent language (clear and operational definitions and measures), (2) explicit use of metatags (meta-data) on data, and (3) clear and complete descriptions of assumptions. These are part and parcel of an explicit conceptual model*." [37]

A *consistent language* and *use of metatags* relate to the semantic model of the system of interest. This is often described as an ontology, but the term "System Metamodel" is more appropriate. The *description of assumptions* refers to appropriate documentation of construct variables and associated contextual assumptions of lower level abstractions.

The use of inconsistent language to name the data elements in the resulting database is the major limitation of a common data model, it can take years to agree on data element definitions and a static data schema can make the data model difficult to modify. Data element names are often useless to infer meaning. These issues can be abated by consistent mapping generated from data element descriptions in data dictionaries, a primary component of metadata. Data providers that create rich metadata and share this across the data federation will aid in effective model and data sharing. Metadata has additional benefit as it can hide the actual data if it is restricted, without impacting the federation [46]. Data value ranges and units must also be consistent or readable from the metadata.

Three general developments emerging from modern web standards aid in linking different data collections from different domains. The first is the Web Ontology Language (OWL) and widely used Resource Description Framework (RDF) stores such as Google's FreeBase. The standard subject-predicate-object or object-attribute-value framework and semantic linking ease in the standardization of semantic terms and relationships. Various domains are rapidly creating large RDF stores or web ontologies describing their domain. To date relatively little development and standardization of common web ontologies have been undertaken across the social sciences domain. However, as researchers opt to use existing ontologies and create domain specific ones, conditions will improve. A consistent language representation is the foundation of a good system metamodel.

A second development is extensive use of web linked data standards. Most database schemas remain defined in eXtensible Markup Language (XML) form but the web community is transitioning to JavaScript Object Notation (JSON) format for standard document annotation and linking of data to research. JSON is a computer language independent format for sharing objects and attribute-value relationships across different datasets, documents, etc. in addition, the use of annotated Hyper-Text Markup Language (HTML) documents to describe research experiments and link input data and results will aid in broader community sharing.

A third area of exploration is the evolution of linked graphs of semantic and mathematical information, an area that is rapidly developing due to Google's introduction of Knowledge Graph and similar entity-driven stores of large information sets. Graph structures support semantic integration and structuring of linked data by compiling text into linked nodes and then relating these to concepts that provide shared meaning to the text. In the graph structure the metadata of our data federation could be linked into a semantic network that can be grown over time with new data. This is an area of needed research; the ability to create large curated sets of community shared and agreed upon causal data and linked experimental results could transform social science research.

A significant hurdle in social science use of these tools is reconciling the linking of different actors' viewpoints to the standard object-attribute-value ontologies. Different actors assign different meaning to social entities and relationships, making contextual features of language by the actor an important variable. The specific meaning associated with the language used by different actors requires a different structuring of shared ontologies than used in most of these applications today. This is an area for further research.

The use of these new technologies does not inherently capture the conceptualizations that defined that data to be important in the first case, and it does not capture assumptions made about missing data elements in the graph. Discerning real causality from experimental measurement of a social construct often requires a qualitative analysis of the underlying causal variables that cannot be measured directly. This is an underlying conceptual model that is often not fully documented in the research results, particularly those potentially causal variables that were purposefully not assessed in the research. This is where context becomes critical – discussions of why these variables are assumed to be causal in this context versus different variables in another context – becomes a key component of the knowledge base. Existing computer-based data models and analytical models are not linked to their conceptual parent models, primarily because the available modeling tools have not been built. A related area of research is specific to this problem, which is how to formally link more freeform conceptual diagramming or facilitation artifacts with more constrained formal modeling and simulations tools.

The federation model recognizes the need to link in the dynamic aspects of predictive models with feedback and adaptation. Research cohorts should be able to extract the fundamental model from the central data model and conceptualization, apply their own dynamic or empirical results in their context, then provide updates back to the whole as new information and ideally new datasets. For example

system dynamics models can provide a larger systems context by connecting key social and human capital factors from the societal well-being model with system dynamics models of infrastructure, spatial communities, and social communities. Medina-Borja and Pasupathy [45] demonstrated the combination of complex structured equation models for uncovering causal relationships in data-rich scenarios and elucidating these to stakeholders through system dynamics models. These dynamic models are going to be context dependent, and should not be considered part of the data federation itself, although they will produce evidence that matures the conceptual model over time.

The "clear and operational definitions and measures" noted by Alberts et al. in the military context is a difficult hurdle in less well governed social situations [36]. Operational definitions and measures in social situations tend to be an area of great debate between different communities of interest. A GAAIN-like common data model is doomed to fail unless we can also define methods and tools to reach agreement on the conceptual models that drive entities, relationships, data definitions, and assumptions. Much of this disagreement involves data conceptualization, definition, and abstraction/aggregation at different scales (for example macroscale measures like "GDP per capita" versus microscale measures like "owning a dishwasher" – both used to describe standard of living). Emerging computer approaches to semantic integration offer hope for much richer microscale measurement sets, as long as the community can clearly see the need for research in this area.

IX. Extending to Next Generation Social Science

The explosive growth of computational tools and methods for analyzing social science data are not limited to use only during the analytics stages of the scientific process. Such tools, along with the massive increase in global digital connectivity, has opened new possibilities for both designing and conducting social science research in addition to data analytics. In this section, we briefly discuss this research.

A. Next Generation Social Science Study Design

Technology in the next generation will aid social science researchers with many of the typical tasks required for sound study design by providing automated aid in finding and vetting authoritative sources; automatically summarizing, categorizing, and organizing the concepts and ideas within scientific texts; cueing researchers to emergent concepts; and helping to identify potential novel hypotheses based on prior literature (using, for example, Microsoft's Academic Graph [46] as a data source). This technology, which we refer to as the Study Design Tool (SDT), will utilize scientometric analysis to automatically ingest and parse scientific publications using computational natural language processing.

Current research and development efforts are underway to build the SDT. These efforts include a collaboration with the Open Science Framework (OSF) [47] in which we are working to develop a social science study schema, which captures relevant study design information in a structured format. To inform iterative design of the schema and associated metrics, the research effort involves eliciting information from researchers regarding their personal design process during each study cycle. Existing (traditional) research design processes and capabilities will be enhanced through the development of new annotation, search, and machine learning-based classification functions in the SDT to allow researchers to rapidly explore and discover social science studies stored in OSF according to topics, keywords, methods used, dependent/independent variables studied, sampling techniques employed, research subject pool demographics, hypotheses tested, cross-references, and/or forward/backward citation context mapping. For example, by having researchers provide keywords relevant to their studies and references to foundational studies, the SDT will report metrics based on a co-citation analysis that indicate the degrees to which foundational research may be biased – such as when it only cites particular subsets of past work (i.e., cliqued or clustered scientific communities, in a graph analytical sense). These analyses may be run over either external publication databases (e.g., Scopus [48] or Web of Science [49]) or over all data stored in OSF. Another function will be to suggest relevant journal articles to researchers based on unbiased sampling over a clustered topic space that may suggest new avenues for inquiry. SDT will also capture insights from researcher-conducted literature reviews, allowing for a reduction in labor for future studies. This technology will aid in novel hypothesis generation and innovative experimental methods (e.g., by cueing researchers to interesting, but as yet untested combinations of dependent/independent variables, methods, domain contexts, and so on) to advance rigorous, reproducible social science studies at scales necessary to develop and validate causal models of human social behaviors.

B. Next Generation Social Science Study Deployment

Once limited by practical constraints to experiments involving just a few dozen participants - often university students or other easily available groups - or to correlational studies of large datasets without any opportunity for determining causation, scientists can now engage thousands of diverse volunteers online and explore an expanded range of important topics and questions. New tools and methods for harnessing virtual or alternate reality and massively distributed platforms will be developed and objectively validated, helping to mitigate many of the vexing challenges in social science. By developing and applying new methods and models to larger, more diverse, and more representative groups of individuals - such as through globally connected web-based platforms - we seek to validate new tools that may empower social science in the same way that sophisticated telescopes and microscopes have helped advance astronomy and biology.

X. Conclusions

This paper serves both as a general call for new social science methods and tools, and as a review of several efforts across a number of domains that address the call. It is exploratory but also representative of current technology. Community resilience, in the face of climate change, aging infrastructure, migration, and other looming grand challenges, represents a perfect opportunity to test these new concepts. Community resilience is but one of the many societal issues that is need of enlightenment from social science. Issues of

privacy and security must be acknowledged and addressed, but should not be insurmountable barriers to progress in the social sciences. We will be interested participants in and observers of the next generation in social analytics.

REFERENCES

[1] D.J. Folds, "Next Generation Social Analytics: Challenges and Payoffs," paper presented at HUSO 2016, The Second International Conference on Human and Social Analytics, 2016.

[2] C.J. Hutto, "Blending Quantitative, Qualitative, Geospatial, and Temporal Data: Progressing Towards the Next Generation of Human Social Analytics." Proceedings of HUSO 2016, The Second International Conference on Human and Social Analytics, 2016.

[3] T.A. McDermott, M. Nadolski, and D.J. Folds, "System-Level Experimentation: Social Computing and Analytics for Theory Building and Evaluation," Proceedings of HUSO 2016, The Second International Conference on Human and Social Analytics, 2016.

[4] Roberto Unger on what's wrong with social science today: http://www.socialsciencespace.com/2014/01/robertomangabeira-unger-what-is-wrong-with-the-social-sciencestoday/, accessed 31-May-2017.

[5] Open Science Collaboration, "Estimating the reproducibility of psychological science," Science. 349 (6251): aac4716. August 28, 2015

[6] K. Bollen, J Cacioppo, R.M. Kaplan, J.A. Krosnick, and J.L. Olds, Social, Behavioral, and Economic Science Perspectives on Robust and Reliable Science, Report of the Subcommittee on Replicability in Science, Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Science, May 2015.

[7] U. S. Federal Emergency Management Agency (FEMA), Draft Interagency Concept for Community Resilience Indicators and National-Level Measures, Published for Stakeholder Comment in June 2016.

[8] National Institute of Standards and Technology, NIST GCR 15-993, Community Resilience Workshop, February 18-19, 2015.

[9] C. Geertz, "Thick Description: Toward an Interpretive Theory of Culture," in The interpretation of cultures: selected essays, New York, NY: Basic Books, 1973, pp. 3–30.

[10] C.J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014, pp. 216–255.

[11] H.H. Clark, Using Language. Cambridge University Press, 1996.

[12] H.H. Clark and S.E. Brennan, "Grounding in communication," in Perspectives on socially shared cognition, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. Washington DC: APA Books, 1991.

[13] T. Mitra, C.J. Hutto, and E. Gilbert, "Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk," in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 1345–1354.

[14] E. Diener, "Assessing subjective well-being: Progress and opportunities," Soc. Indic. Res., vol. 31, no. 2, pp. 103–157, Feb. 1994.

[15] E. Diener, E.M. Suh, R.E. Lucas, and H.L. Smith, "Subjective well-being: Three decades of progress.," Psychol. Bull., vol. 125, no. 2, pp. 276–302, 1999.

[16] D.J. Folds and V.M. Thompson, "Engineering human capital: A system of systems modeling approach," in Proceedings of the 8th International IEEE Conference on Systems of Systems Engineering (SoSE-13), 2013, pp. 285–290.

[17] T.W. Smith, P.V. Marsden, M. Hout, and J. Kim, "General Social Surveys, 1972-2014 [machine-readable data file]." NORC at the University of Chicago [producer and distributor], 2014.

[18] InternetLiveStats.com, "Internet Live Stats," Internet Live Stats - Internet Usage and Social Media Statistics, 2016. [Online]. Available: http://www.internetlivestats.com/, accessed: 31-May-2017.

[19] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Found. Trends Inf. Retr., vol. 2, no. 1–2, pp. 1–135, 2008.

[20] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool, 2012.

[21] J.W. Pennebaker, M. Francis, and R. Booth, Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Erlbaum Publishers, 2001.

[22] J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth, The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net, 2007.

[23] M.M. Bradley and P.J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," NIMH Center for the Study of Emotion and Attention, Center for Research in Psychophysiology, University of Florida, Technical Report C-1, 1999.

[24] P.J. Stone, D.C. Dunphy, M.S. Smith, and D.M. Ogilvie, General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press, 1966.

[25] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in Proc. of LREC, 2010.

[26] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[27] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, 2009.

[28] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed Cognition: Toward a new foundation for human computer interaction research," ACM Trans. Comput.-Hum. Interact. TOCHI, vol. 7, no. 2, pp. 174–196, 2000.

[29] E. Hutchins, "Distributed Cognition," in International Encyclopedia of the Social & Behavioral Sciences, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 2068–2072.

[30] C.J. Hutto, S. Yardi, and E. Gilbert, "A Longitudinal Study of Follow Predictors on Twitter," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 2013, pp. 821–830.

[31] S.A. Umpleby, "Second-order science: logic, strategies, methods," Constructivist Foundations 2014, vol. 10, no. 1, pp. 16-23, 15 November 2014.

[32] J. Rotmans, R. Kemp, and M. van Asselt, "More evolution than revolution: transition management in public policy", Foresight, vol. 3, no. 1, pp. 15-31, February 2001. ISSN 1463-6689.

[33] F.W. Geels, "Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study." Research Policy, vol. 31, pp. 1257–1274, 2002.

[34] R. Wagner-Pacifici, J.W. Mohr, and R.L. Breiger, "Ontologies, methodologies, and new uses of Big Data in the social and

cultural sciences," Big Data & Society, vol. 2 iss. 2, pp. 1-11, December 2015. DOI: 10.1177/2053951715613810.

[35] D.S. Alberts, R.E. Hayes, D.K. Leedom, J.E. Kirzl, and D.T. Maxwell, Code of Best Practice for Experimentation, Washington DC: CCRP Publication Series, 2002.

[36] D.S. Alberts and R.E. Hayes, Code of Best Practice for Campaigns of Experimentation: Pathways to Innovation and Transformation, Washington DC: CCRP Publication Series, 2002.

[37] www.gaain.org/ , accessed: 31-May-2017.

[38] www.humanbrainproject.eu/ , accessed: 31-May-2017

[39] W.B. Rouse and D. Bodner, Multi-level modeling of complex socio-technical systems – phase 1, A013 - final technical report, SERC-2013-TR-020-2, Systems Engineering Research Center, 2013.

[40] W.B. Rouse and M. Pennock, Multi-level modeling of socio-technical systems a013 - final technical report, SERC-2013-TR-020-3, Systems Engineering Research Center, 2013.

[41] T.A. McDermott and D. Freeman, Systems thinking in the systems engineering process: new methods and tools, in Systems Thinking: Foundation, Uses and Challenges, Eds. Frank, Shaked, Kordova, Nova Publications, 2016.

[42] J. Ernst, "What is metamodeling, and what is it good for," http://infogrid.org/trac/wiki/Reference/WhatIsMetaModeling, retrieved: November 2015.

[43] N. Ashish and A.W. Toga, "Medical data transformation using rewriting," Frontiers in Neuroinformatics, vol. 9, no. 2, pp. 1-8, 20 February 2015. doi: 10.3389/fninf.2015.00001

[44] N. Ashish, P. Dewan, J.L. Ambite, and A.W. Toga, GEM: The GAAIN Entity Mapper, in Data Integration in the Life Sciences, 11th International Conference, DILS 2015, Eds. Ashish, N. and Ambite, J., Springer 2015.

[45] A. Medina-Borja, K.S. Pasupathy, and K. Triantis, "Large-scale data envelopment analysis (DEA) implementation: a strategic performance management approach," Journal of the Operational Research Society, 58(8), 1084-1098, 2007.

[46] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/, accessed 31-May-2017.

[47] https://osf.io/ , accessed: 31-May-2017

[48] https://www.scopus.com/, accessed: 31-May-2017

[49] http://webofknowledge.com/, accessed: 31-May-2017