ISSUE PAPER

Leveraging Big Data for Science & Innovation Indicators: Pitfalls and Promises

Jeffrey M. Alexander, Ph.D.
Senior Manager, Innovation Policy
RTI International
701 13th Street, NW, Suite 750
Washington, DC 20005
jmalexander@rti.org

Abstract

As seen in other domains of social science and policy, the field of science and innovation policy is focusing recently on the perceived potential benefits to be gained from the capabilities of "Big Data." In this context, Big Data encompasses the integration of machine learning, "organic" data (such as administrative records), and knowledge of the dynamics of innovation. In 2015, the Industrial Research Institute launched a research effort to study the potential impact of "Big Data" on the management of R&D in large organizations. This and other activities have revealed both the promises and the pitfalls of these new modes of research when applied to science and innovation measurement. Direct experience in these techniques and approaches highlight some of those cautionary considerations, including:

- The dangers of exploiting datasets that have possible defects and biases not recognized by the researchers;
- The difficulties in evaluating Big Data techniques and analysis, especially by conventional criteria such as falsifiability; and
- The complexities involved in explaining these techniques and their value as evidence in policy evaluation to decision-makers and to the public.

In this paper, we present some observations on ways in which Big Data can provide a constructive complement to more traditional forms of social science research. We conclude by presenting a few propositions on how Big Data might be leveraged more effectively in the development of next-generation indicators for science and innovation.

Acknowledgments

This paper benefited tremendously from numerous discussions during the author's work at SRI International with a number of sponsoring organizations. My understanding of the technical nature of machine learning and "Big Data" was grounded in discussions with Dr. John J. Byrnes of SRI Advanced Analytics, along with Prof. Padhraic Smyth of UC Irvine, Prof. David Newman (formerly at UC Irvine, now at Google), and Prof. Eytan Adar of the University of Michigan. My thoughts were honed through many interactions with my colleagues at SRI's Center for Innovation Strategy & Policy, particular John Chase and Dr. Christina Freyman. I also owe much to the insights from Dr. Dewey Murdick of the U.S. Department of Homeland Security, Patrick Lambe of Straits Knowledge, Dr. Julia Lane of New York University, and Jeri Mulrow (formerly at the National Science Foundation, now at the U.S. Bureau of Justice Statistics). Any errors or misinterpretations are the sole responsibility of the author.

Prepared for the OECD Blue Sky Forum III: Towards the Next Generation of Science and Engineering Indicators, 19-21 September 2016, Ghent, Belgium

1 Introduction

Traditionally, R&D activity is measured in the statistical community via surveys, per the recommendations of the OECD Frascati Manual (OECD, 2015).[1] Recently, prominent members of the community have raised concerns about trends affecting the reliability, validity, and feasibility of surveys for collecting detailed economics and social data. Notably, Dr. Robert Groves (formerly director of the U.S. Bureau of the Census, and now at Georgetown University) wrote:

> …at this moment in survey research, uncertainty reigns. Participation rates in household surveys are declining throughout the developed world. Surveys seeking high response rates are experiencing crippling cost inflation. Traditional sampling frames that have been serviceable for decades are fraying at the edges. (Groves, 2011)

In that essay, he goes further to discuss the potential to leverage what he calls "organic data" as a supplement to surveys. Organic data refers to data produced to track activities for purposes other than statistical measurement, but which can be processed and analyzed to generate statistics. Generally, organic data has been numerical in nature. An example of this is the Billion Prices Project at MIT, which seeks to replicate the Consumer Price Index by "scraping" the prices of common products off of e-commerce sites like Amazon, rather than deploying in-store or point-of-sale collectors (human and automated) to observe prices in "brick-and-mortar" stores.

The advent of natural language processing and statistical analysis of digitized text now presents new opportunities for the measurement of human activity. By analyzing descriptions of activities found in organic data repositories, we may be able to model with great accuracy relationships between entities, activities, and outcomes. The intersection of three phenomena could enable new advances the measurement of science and innovation in society:

[1] See, in particular, "Design of data collection methodology" at p. 187.

- Administrative records and other repositories of documents provide a new source of raw data for analysis
- Machine learning, taken as the combined advances in statistics and computation, can process textual information to aid in the summarization and interpretation of large sets of documents
- Science and innovation studies (referring to the body of work on innovation management, sociology of science, and related domains) offer a set of theories and extant findings that provide new propositions about how scientific and technical activities occur, and how innovation manifests itself in the economy and society.

The work initiated by the U.S. federal government under the STAR METRICS (Science & Technology for America's Recovery: Measuring the Effect of Research on Innovation, Competitiveness and Science) effort exemplifies the types of data infrastructure and capabilities enabled by these developments. (Lane, Owen-Smith, Rosen, & Weinberg, 2015) One result is the data system maintained by the Institute for Research on Innovation in Science (IRIS), linking records of federal government research grants to human resource and financial management records from universities receiving those funds. Advanced analytics enable IRIS researchers to match grants to the individuals whose research is funded by those grants, and then to the articles produced by those researchers from the funded projects. Topic modeling and other techniques in text analysis can then group together projects based on the similarity of their subject matter, which in turn helps to understand the different ways that federal funds can impact specific research fields.

Based on several different efforts to investigate the use of digitized documents in exploring the dynamics of science and innovation, we can make some summary observations about both the problematic aspects of this convergence and more promising approaches to using "Big Data" techniques in developing next-generation indicators.

2 Understanding Text Records as an Alternative Data Source

Digitized text repositories offer a number of advantages over surveys as a source of new data on science and innovation. The primary advantage is that this type of "organic data" is readily available in data systems. Data collection becomes a task of identifying, locating, and extracting record files, rather than preparing a survey instrument, testing it, fielding it, and then working to maximize the response rate. Even compared to Web-based surveys, administrative records collections appear to be an efficient method of gathering data on phenomena of interest rapidly and inexpensively.

Experience with these repositories belies that appearance. Administrative records, like other types of data, are often filled with errors, as they generally rely at some point on fallible human data entry. In one project, this researcher attempted to use data from the Federal Procurement Data System, a regularly-updated database of U.S. federal contracting transactions, to calculate R&D investments by geographic region. (U.S. National Research Council, 2010) The U.S. General Services Administration collected these data from multiple agencies, offering the promise of a government-wide assessment of this type of extramural R&D spending.

Close inspection of the records revealed a number of systematically-flawed entries. Due to the very rigid data structure in FPDS, some agencies found that their operational needs were not met by the design of the data system. In response, data entry operators would "hijack" data fields—for example, in place of a project summary, the agency would list the name of the office that issued the contract. Therefore, summarizing the contracts by topic or description left many gaps. Also, a large number of

contracts that were classified as R&D contracts using the FPDS taxonomy were clearly *not* R&D funding. Those had to be reclassified so that they were excluded from the final dataset.[2]

Using administrative records will often involve extensive efforts to inspect, correct, organize, parse, and cleanse those data. This process, known colloquially as "data munging," can be extremely time-consuming and complex. Even the use of automated software, such as "Extract-Transform-Load" (ETL) packages, may require time and effort to set up properly. One researcher in e-Science estimated that in "big data" research projects, as much as 90 percent of a data scientist's time would be spent on data management rather than data analysis. (Howe, 2015)

There also may be systematic biases in large sets of administrative records that researchers gloss over, or ignore outright. The size of the dataset may give it the appearance of a census, rather than a sample—if one is analyzing the entire population, doesn't bias become irrelevant? In this sense, the *provenance* of the record set is critical. For example, a number of firms and researchers are developing systems for "real-time labor-market information"—analyzing online job posting in a region (from, for example, Monster.com), comparing them to online resume databases (such as Indeed.com), and using those data to build supply/demand models of labor-market dynamics. These systems assume that all employers recruiting for certain occupations are posting those job descriptions online, and that all job-seekers are posting their resumes. One issue encountered by researchers is that employers are posting announcement for positions that don't actually exist, simply to take a measure of the market for candidates. Employers also post different announcements for the same position, using different language in each posting to attract different types of candidates. Therefore, the "demand" side of the

[2] See U.S. National Research Council (2010) at pp. 94-95.

analysis is being systematically skewed because it is not an accurate representation of the population of interest.

Understanding the nature of the administrative records also reveals potential pitfalls in analysis. For example, researchers seeking to understand technical progress often apply citation analysis to patents in a manner similar to its application in journal bibliometrics. In theory, citing a work in a patent provides an acknowledgement that the work was somehow influential in the invention presented in the patent claims. More recently, researchers have investigated the role of citations in patent approval, using data from the U.S. Patent and Trademark Office denoting which citations are added by patent examiners versus those added by the inventor or patent agent. An initial study found that in for some technological fields and especially some patent applicants, citations were commonly added to an application by the examiner, not the applicant. Therefore, using citation analysis to model "influence" among patents in the same way as among scholarly articles could have misleading results (Alcacer, Gittelman, & Sampat, 2009).

In one project, researchers looked at differences in the syntax contained in scholarly article abstracts and patent text. This investigation revealed differences in how articles and patents are written. For a given set of articles and patents concerning the same technology (for example, DNA microarrays), a topic model run across both sets of documents showed very little overlap in the topics found in each type of document. Topics discovered in the article set were generally not present in the patent set, and vice versa, even if both sets were generated by searching for the same set of keywords. This reflects in part the fact that article authors and patent applicants have very different motives for why they write what they do. Academic researchers want to "advertise" the work that they have performed and its value to the field. Thus, they will be explicit about their research findings and contributions to knowledge. By contrast, patent authors (who often are not the actual inventors) have an interest in obfuscating the exact nature of the invention. This approach has strategic value in both the process of

patent prosecution, and in preventing potential competitors from learning how to engineer around the patent. Therefore, deploying text analytics to measure concepts in these two types of documents must be done differently to account for their divergence in language and intent.[3]

These three problems—dataset quality, sample bias, and behavioral variance—have long-plagued conventional statistical analysis. The danger is that proponents of Big Data may argue mistakenly that these problems are mitigated, or even eliminated, by using larger and larger datasets. (McFarland & McFarland, 2015) Administrative record sets may offer greater coverage of a population of interest, but they still have defects, and those defects may just be more difficult to isolate and minimize.

3 Machine Learning, "Black Boxes," and Problems of Interpretability

As we move into a mode where "big data" feeds into the domain now called "computational social science," we should be careful to establish some points of reference. The term "big data" itself sparks some controversy, as different data users and analysts apply different parameters to measure what we mean by "big." The Industrial Research Institute, the professional association of R&D executives of large firms and other organizations, launched in 2015 a study group looking at the phenomenon of Big Data and how it might change R&D in the future. As co-chair of that group, I presided over early definitional discussions. One proposal that gained traction across many participants was to reframe Big Data as "uncomfortable data"—datasets that are too large to be handled by "conventional" tools and techniques. (Alexander, Blackburn, & Legan, 2015)

This view has three advantages. It takes into account that different organizations, disciplines, and domains will have differing views on the size of a big dataset. As one participant noted, in some

[3] Based on comments by Profs. Padhraic Smyth and David Newman of UC Irvine at SRI International, June 20, 2012

organizations, data becomes "big" when it exceeds the processing capabilities of Microsoft Excel or similar off-the-shelf spreadsheets. A second advantage is that it provides a moving target. Technical and computational advances will always expand our capacity to store and analyze data. Big Data should be viewed as the type of dataset that is just beyond our standard operating environment. Finally, our view includes the fact that the complexity of the dataset, rather than its absolute volume, may make it "Big Data." Digital text, for example, can contain very rich information that can be extracted and interpreted only with careful attention to both content and context. Processing even a few thousand detailed documents can take on the properties of Big Data, if the researcher is not already very familiar with that repository. In particular, unstructured text (where the documents and their content are not described by well-documented, standardized metadata) can be very challenging to analyze with computational techniques, as the disparities between documents confound routine processing.

In the computational analysis of text and similar datasets, computer scientists build models that assist in the interpretation of a set of documents by parsing and summarizing their content. One popular technique is "probabilistic topic modeling." (Blei, 2012) In this approach, a set of documents is treated as a set of terms that appear with some pattern in colocation and co-occurrence. These patterns derive the "hidden" structure of documents—the topics discussed in the documents—from the observed terms. Topic modeling is attractive for analyzing large documents sets because it utilizes unsupervised learning. The topics are derived by calculating the relationships between terms, and those that appear most closely related are grouped together as a "topic." Topics are simultaneously associations between terms and associations between documents.

This mode of analysis is described as "probabilistic" because it uses statistical algorithms to identify groups of terms most salient to each topic, and to assign documents to those topics. The most popular method of model generation is Latent Dirichlet Allocation, or LDA. In this process, the topic modeling software will randomly begin calculating associations between terms and documents, and produce

topics "generatively" through iterative calculations. Once the process has completed calculating all of these relationships, it can identify a given number of topics (generally 10 to 100 topics) from the set of given documents. While a very powerful method for extracting key themes, this reveals some challenges in using machine learning for text analysis:

- The "topics" generated by a topic model are strings of terms that are difficult to interpret without further human review. (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) While probabilistic modeling does not require prior "labeling" of text by experts, which can be time-consuming, it may require follow-on inspection and labeling of topics by those experts. Technical and scientific document sets can require significant post-processing to generate human-interpretable labels.
- As the topic modeling process begins by sampling terms in random order, successive topic models run on the same set of documents can yield different results. Therefore, probabilistic topic models may not be replicable.[4] This is especially true when generating large numbers of topics on smaller sets of documents. In such cases, the relationships calculated by the process are path-dependent; the associations between terms calculated using LDA depend on the random starting point of the process. Depending on the application, the variance in the topics generated from the same set of documents may cause some uncertainty in how a given model's results should be analyzed.

These limitations reveal some key points that should be considered when deploying machine learning algorithms for the analysis of Big Data:

[4] See discussion at http://datascience.stackexchange.com/questions/8193/replicability-reproducibility-in-topic-modeling-lda accessed 20 July 2016.

- Researchers and analysts should have some basic understanding of the mechanisms by which machine learning algorithms process data and build models. Using these algorithms naively may generate findings that not fully reliable or replicable.
- There is a close relationship between the nature of the documents being analyzed, and the nature of the topics produced by topic modeling. A strong understanding of the particular features of the document set will help in interpreting the results of machine-generated models.

The second point is related to a more general issue. Researchers who understand the fundamental nature of these datasets are likely to be those steeped in the particular domain of the phenomenon being studied. A further danger of the "Big Data" movement is the belief by some computational researchers that this kind of study can be "theory-free." Rather than developing detailed causal models based on the particular dynamics of the domain, these researchers run machine learning analyses with extremely large, multivariate datasets to isolate particular correlations. As noted in Section 2, if one assumes that a Big Data approach measures and models every variable of significance, it is possible that a purely computational analysis will reveal some trend or influence of interest. In practice, Big Data does not make domain expertise obsolete, and may make it even more critical to successful research. A scholar in the field can contribute useful insight into what is already established knowledge in the field, where new phenomena and variables of interest are likely to be found, and how a Big Data approach might be deployed to complement and extend more traditional investigations, rather than attempting to supplant existing scientific knowledge. While Big Data claims to provide more powerful tools of investigation, that power is increased when deployed in a thoughtful way by teams that include domain experts. This is discussed further in the following section.

4 Towards a Computational Science of Science & Innovation Policy

The community of scholarship in the Science of Science and Innovation Policy (SciSIP), such as it exists, is embracing more computational methods for measuring, describing, and analyzing innovation. One issue holding back the field is its balkanization. Within the science and innovation policy fields, these studies were typically the realm of bibliometricians, or later scientometricians, who utilized early databases of bibliographic metadata about scientific articles. Researchers in computational linguistics also addressed issues related to the nature of progress in science and innovation, often as way of developing new methods in discourse analysis or natural language processing. Another stream of research flowed from the library sciences, particularly in information retrieval. We are now seeing a convergence among these communities, as both funding opportunities and venues create the motivation for collaboration between innovation researchers, computational linguistics, and the 'iSchool' (information school) movement. There is an increase in journal articles in the noted publication *Research Policy* that rely on computational methods of analysis (see, for example, works by Small et al. and Breitzman & Thomas). (Small, Boyack, & Klavans, 2014) (Breitzman & Thomas, 2015) The 2013 iConference featured a workshop on "computational scientometrics," and papers from a 2013 workshop on "Combining Bibliometrics and Information Retrieval" were published in a 2015 issue of the journal *Scientometrics.* (Mayr & Scharnhorst, 2015)

As the science of science and innovation policy is brought into the domain of computational social science, what are the implications for the community of researchers and practitioners involved in developing and disseminating science and innovation (S&I) indicators? I offer a few observations about views to keep in mind.

First, we need to keep sight of the "why" of S&I indicators, without becoming too enamored with the "how." Yes, Big Data and advanced analytics will give us new tools with which we can devise an even

greater array of indicators. But at the end, indicators are intended to assess and support policy decision-making. Increasing the complexity of the processing pipeline for indicators can make them more accurate and granular, but at a cost of making them explicable and transparent. Our role is to deliver the right data and indicators appropriate for the types of decisions that need to be made. This still requires more judgment and understanding about the decision-making process, rather than additional computation.

This does have a strong implication on approaching the "how" of computational S&I indicators. One notable trend is that as we making indicators more computationally intensive, we become part of a broader debate between what University of Michigan professor Eytan Adar calls the "two cultures" of big data—the culture that pursue explanatory models, and the culture that pursues predictive models. (Adar, 2015) One of the possible temptations of having access to huge swaths of data is the conceit that we can capture virtually every variable of consequence and process them simultaneously in a single model. Exotic techniques with labels like "random forest" and "Hidden Markov Models" can churn through variables and find those most predictive of particular outcomes.

Prediction relies on a few fundamentals, however. One is that we have an idea of the actual "outcome" that matters. This is not always obvious. In a series of investigations on identifying and predicting the trajectories of "emerging technologies," we discovered that a serious limitation came from a lack of formalized definitions. There is no clear agreement, for example, on what makes a technology "emergent." As one example, it seems somewhat ridiculous to continue studying nanotechnology as an "emerging" technology, when it has been in development for decades. How will we know when it is no longer "emerging" but has, in fact, emerged? There is no clear threshold marking that boundary. Similarly, what makes nanotechnology a "technology?" At what level of analysis are we aiming? The absence of a rigorous, formalized framework renders data analysis fairly useless.

Even if we know what to predict, how does prediction help us make decisions? In a debate with a computer scientist, I was presented with the argument that an accurate prediction is more important than an explicable one. A key feature of certain machine learning techniques, like LDA, is that there is an element of randomness in their processing. Therefore, in running complex prediction models, a researcher may not be able to decompose the final prediction into a logical sequence tied back to specific variables. In other words, we arrive at a prediction with little idea on how that prediction was generated.

While accuracy is commendable, the shortfall here is that even the most sophisticated model relies on underlying data structures, and data is collected inevitably with some biases and assumptions. We gather data because we believe that they describe the world in a certain way. In computing a prediction, we are not able to expose those assumptions clearly. As a result, a predictive model may be very *brittle*. If its predictive power relies on a specific implicit assumption, and then reality changes somehow so that the key assumption no longer holds true, the model will fail catastrophically and without warning. In academic research, this may be an inconvenient failure. In policymaking, the decisions made based on that model could have very significant consequences. The "flash crashes" observed in stock market trading may be one manifestation of computing models' inability to foresee the perverse effects of implicit assumptions on model behavior.

Therefore, explanatory modeling and predictive modeling should be viewed as a single enterprise. This highlights the key role of theory and testing in computational indicator development. We still need an underlying framework of understanding how innovation happens, so that data collection and analysis are aligned with our understanding of the phenomena at work. This is especially important when we observe that successful innovation is believed to be a very rare event (although again, we lack formal definitions for identifying discrete examples of unsuccessful innovations—are they simply innovations that haven't succeeded yet?). Predicting a metaphorical "black swan" event without some kind of causal

model is very difficult. It becomes more treacherous in the world of Big Data, where we can collect so many variables that even the most spurious correlations appear to be statistically significant. (McFarland & McFarland, 2015) Thus, we may attribute causation to features and signals that are, in fact, mostly irrelevant to the phenomenon of innovation. A strong, robust theoretical framework for analyzing the forces affecting innovation will help us to constrain the forces in the Big Data community that lead us to generate data that is all-encompassing but not necessarily salient.

This, in turn, will highlight even more the value of qualitative, domain-specific investigation to inform the quantitative, computational aspect of S&I indicators. The "grounded theory" approach could, as an example, take primarily qualitative investigations (such as the NSF's historic TRACES study) and isolate candidate indicators that could be inputs to a Big Data research effort. (Glaser & Strauss, 1967) (IIT Research Institute, 1968) Computational researchers should ensure that their study designs and data collection efforts are informed by the input of domain experts in the study of science and innovation.

Finally, this all argues for a continuing focus on parsimony when we pursue computational research on S&I indicators. We must realize that decision-makers have limited capacity to ingest complex findings, due to time constraints and competing priorities. A better understanding of the decision-making environment might help to define the parameters for selecting and designing a useful indicator set. (Alexander, Hart, & Hill, 2016) We also want to limit the cross-correlation and reduce the degrees of freedom to make models useful and defensible. In this sense, we should strive to ensure that when we venture into the realm of "uncomfortable" data, we still maintain a sufficient comfort level that contributes to credible and reasonable decisions.

5 Works Cited

Adar, E. (2015). The two cultures and Big Data research. *I/S: A Journal of Law and Policy for the Information Society, 10*(3), 765-781.

Alcacer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy, 38*, 413-427.

Alexander, J. M., Blackburn, M., & Legan, D. (2015, Nov-Dec). Digitalization in R&D management: Big Data. *Research-Technology Management, 58*(6), 45-49.

Alexander, J., Hart, D. M., & Hill, C. T. (2016). *Enhancing the Usefulness of Science of Science and Innovation Policy Research: An Agenda-Setting Workshop.* Prepared for the National Science Foundation by George Mason University & SRI International. Arlington: George Mason Univerity.

Blei, D. M. (2012, April). Probabilistic Topic Models. *Communications of the ACM, 55*(4), 77-84.

Breitzman, A., & Thomas, P. (2015). The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy, 44*(1), 195-205.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861-871.

Howe, W. (2015, April 28). A confluence of big data skills in academic and industry R&D. *Annual Meeting of the Industrial Research Institute*. Seattle, WA.

IIT Research Institute. (1968). *Technology in Retrospect And Critical Events in Science.* Illinois Institute of Technology. Washington: National Science Foundation.

Lane, J. I., Owen-Smith, J., Rosen, R. F., & Weinberg, B. A. (2015). New linked data on research investments: Scientific workforce, productivity, and public value. *Research Policy, 44*, 1659-1671.

Mayr, P., & Scharnhorst, A. (2015, March). Combining bibliometrics and information retrieval: preface. *Scientometrics, 102*(3), 2191-2192.

McFarland, D. A., & McFarland, H. R. (2015, July-Dec). Big Data and the danger of being precisely inaccurate. *Big Data & Society, 2*(2), 1-4.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262-272). Edinburgh, Scotland: Association for Computational Linguistics.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467.

U.S. National Research Council. (2010). *Data on Federal Research and Development Investments: A Pathway to Modernization.* (Panel on Modernizing the Infrastructure of the NSF Federal Funds Survey, Committee on National Statistics ed.). Washington, DC: National Academy Press.