The Five Vs of Big Data Political Science

Introduction to the Virtual Issue on Big Data in Political Science *Political Analysis*

Burt L. Monroe
Department of Political Science
Pennsylvania State University
University Park, PA 16803
email: burtmonroe@psu.edu

Introduction

In the last three years, the concept of "Big Data" has emerged from technical obscurity to fully fledged, overblown, memetic phenomenon complete with Rorschach-test perceptions and backlash. For political scientists, the phrase seems to evoke a cognitive mapping onto our old continuum of "small-N" vs. "large-N," implying perhaps "*really* large N." This in turn tends to evoke the social science instinct to sample: "You don't have to eat the whole cow to know that it's tough." (Firebaugh 2008) But this misunderstands the defining characteristics of "big," the qualitative differences from what has come before, and the challenges and opportunities that big data present to political science.

Over a decade ago, IT analyst Doug Laney laid out a framework for understanding the challenges that were then emerging in data management for business, now ubiquitously referred to as "The Three Vs of Big Data": Volume, Velocity, and Variety (Laney 2001). Simply put, data were beginning to be created at a scale, speed, and diversity of forms sufficient to overwhelm relational databases and other conventional modes of enterprise information management. This framework proved both catchy and useful for the "data science" and "analytics" communities, where technological innovations and new data-driven companies, products, and services are often characterized as addressing the challenges posed by, or extracting new value from, one of the Vs.

Broadly conceived, this framework captures much of the challenge and opportunity that big data presents to political science, and social science more generally. As is *de rigueur* I extend the Vs,[1] and those that arise at the fuzzy borders among political science, computer science, and data science. Included in this virtual issue are ten articles from the last decade of *Political Analysis*, which are individually excellent and worthy of your attention, and which are collectively useful for illustrating these five Vs of big data political science: volume, velocity, variety, vinculation, and validity:

- Stephen Ansolabehere and Eitan Hersh, "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* (2012), 20: 437-59.
- Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* (2012), 20: 351-368.
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn, "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* (2008), 16: 372-403.
- Pierre F. Landry and Mingming Shen, "Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China." *Political Analysis* (2005), 13: 1-22.
- Justin Grimmer, "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* (2011), 19: 32-47.

[1] There are as many "fourth Vs of big data" as there are "fifth Beatles."

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

- Skyler J. Cranmer and Bruce A. Desmarais, "Inferential Network Analysis with Exponential Graph Models." *Political Analysis* (2011), 19: 66-86.
- Marc T. Ratkovic and Kevin H. Eng, "Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes." *Political Analysis* (2010), 18: 57-77.
- Jacob M. Montgomery, Florian M. Hollenbach and Michael D. Ward, "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* (2011), 20: 271-91.
- Ryan Bakker and Keith T. Poole, "Bayesian Metric Multidimensional Scaling." *Political Analysis* (2013), 21: 125-40.
- Tom S. Clark and Benjamin E. Lauderdale, "The Genealogy of Law." *Political Analysis* (2012), 20: 329-50.

Volume

For most, the "Big" in Big Data means *volume*. Data have volume because mobile and distributed technology has created more observations (e.g., tweets), sensor technology has created observations that are individually massive (e.g., digital images), storage technology has enabled us to keep more, and networking technology has enabled us to share and combine more. It is the simplest measure of volume, file size, for which there are obvious upper limits to standard techniques – the data cannot be stored in one place, cannot be easily queried, subsetted, or manipulated, or cannot be held in memory for computation – beyond which something different must be done.[2]

In a few circumstances, data are created in volume with their end purpose as data in mind. Generally, this requires resources on a commercial or governmental scale. Examples include purchasing card data collected by grocery stores, or the individual Decennial U.S. Census. More typically, data obtain volume passively, created as the digital traces of human behavior and interaction (Webb 1966 at web-scale), a phenomenon referred to as *data exhaust*. Further, data can reach volume through *linkage*, the merger of data sets through shared fields. All of these are at work in the background, but linkage is what gives the first paper in the issue, **Ansolabehere and Hersh 2012**, its volume and its unambiguous claim to the Big Data label. Ansolabehere and Hersh partnered with Catalist, a commercial firm built on the creation of massive individual voter data, in order to investigate the validity of survey measures of turnout and related voting behaviors at an unprecedented granularity. The volume in play here is not immediately obvious – Catalist's raw data are opaque – but certainly meets any reasonable criterion for Big. The voting age population for recent elections in the U.S. is roughly 200 million, and a single record in the underlying voter files represents one individual on a list in a particular jurisdiction at a particular point in time. The effort to de-duplicate, validate, etc., would have to involve processing at least tens of billions of individual records.

Note that the target of this analysis is not a quantity that can be obtained by sampling from those billions of records. The target quantity is a comparison of survey response and truth, which can only be estimated for those specific individuals surveyed by the CCES or NES. This requires an exact match of a few thousand specific individual-jurisdiction-time records out of those billions. The big data technology that does most of the work here is linkage – also called *merging*, *entity disambiguation*, *mash-up*, and in this paper *matching* (an unfortunately confusing choice for the *Political Analysis* audience) – which comes into play in two places. The proximate one is the linkage of CCES and NES data with the Catalist database. The less obvious one is the linkage Catalist is doing between voter files and commercial databases (like credit card information) and public data (like Post Office change of address data). This is the core of many data aggregation businesses, and the source of Catalist's value as a provider of microtargeting information to campaigns.

Linkage is a challenge, but opens up many new scientific opportunities, and individuals are not the only entities on which data sets can be linked. Event data projects for example seek to disambiguate entities in news reports like collective political actors (like "Hezbollah") and actions (e.g., verb phrases like "attacked"), from which data about political events and sequences of events can be generated (Schrodt and Yonamine 2012). Geo-referencing of

[2]This is unfortunately an operational definition of Big Data that many in the data science community carry in their heads. I defy you to attend a conference session on Big Data, at any conference in any setting, and not witness a panelist or audience member wag his finger and say, in effect, "If you can store it on your laptop's hard drive, it's not Big Data."

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

conflicts by PRIO Conflict Site for the UCDP/PRIO Armed Conflict Dataset allows linkage across data sets by location (Hallberg 2012). Linkage also carries with it the difficult statistical challenge of how to characterize uncertain links, and how to propagate that uncertainty through analyses when alternative discrete mergers would create alternative datasets.

Big Data technology also has the potential to help political science scale up conventional approaches to data collection. There are many modes of political science research that use humans directly for the purpose of generating data: surveys, experiments, content analysis and other "coding." Each piece of such data is expensive – perhaps in more ways than monetarily – and this limits scale. One approach to scaling up is to use statistical learning techniques to replicate human judgment and behaviors, turning a small amount of human-generated data into a lot (e.g., Quinn et al. 2010; D'Orazio et al. 2011; Cantú and Saiegh 2011). Another approach is to design the data generation task so that people provide data at much lower cost, even for free, through *crowdsourcing* and related approaches to human-in-the-loop computation. The familiar reCAPTCHA system, for example, is a technology that functions both as an internet security system and as a massive source of free labor generating character recognition data for Google Books (von Ahn et al. 2008).

The best-known crowdsourcing system is Amazon's Mechanical Turk (MTurk), a market exchange through which "turkers" – anonymous individuals from around the globe – accept micro-payments for micro-tasks. For social scientists this has tempting potential as a pool of coders, survey respondents, and experimental subjects who are cheaper and more responsive than, say, undergraduate students at their universities. But under what circumstances is this pool of subjects worse or better, by how much, and how would we know? In the second paper in the issue, **Berinsky, Huber, and Lenz (2012)** address the validity of turkers as experimental subjects, an exercise so useful and carefully done that I doubt I go out on much of a limb to project this as the most citeable piece in the issue. Berinsky et al. compare MTurk samples (unfavorably, but not as wildly so as one might expect) to those from representative population samples, and favorably to convenience and student samples, while also noting the relative strengths of MTurk for studies focused on internal validity. They then replicate three different social science experiments, finding roughly similar results even in a complex experiment with weak effects.

Finally, data can have volume by having massive numbers of covariates or features, a situation referred to as *high dimensional* – not big N but big k or p, depending on your notational preferences. A standard example is gene expression microarrays, which may have tens of thousands of genes (features) from a few hundred or even a few dozen people. Other examples include textual data (with many more words than documents), functional MRI data (with many more voxels than subjects), and intensive longitudinal data from heart monitors (with many more heart beats than subjects). Moreover, nearly any data set can be made high-dimensional with higher order interaction terms. Such data challenge standard notions of identification and bump against the *curse of dimensionality*. In the third paper in the issue, **Monroe, Colaresi, and Quinn (2008)** address the issues of feature selection and feature evaluation in the context of text data. Specifically, they seek to identify the words that are most informative about the differences between two or more sets of documents, with the primary running example the Senate speeches of the two US parties on a topic and leveraging the context of language to make each point visually. Demonstrating first the face invalidity of commonly used metrics, they develop a (computationally trivial) Bayesian metric for the task. Finally, they demonstrate applications of the metric for the study of representation, realignment, polarization, and dimensionality.

Variety

Variety is one of the most challenging for social scientists to grapple with, as much Big Data comes in forms and data structures that do not map easily onto the rectangular spreadsheets of familiar statistics packages. Tweets are a useful example. They are delivered in structured text formats, either XML or JSON, that impose a nested tree structure on each tweet. The levels of this tree depend on whether the tweeter references a previous tweet. The number of fields varies depending on what the user provides and can include time stamps, geo-references, text, and images. Taken together, a set of tweets has a complex network structure that moves in time and space.

Despite the unfamiliarity, new types of data open up possibilities for new types of analysis and new research designs. In the fourth paper in the issue, **Landry and Shen (2005)** demonstrate a new opportunity arising from the availability of cheap and universal GPS devices. Landry and Shen construct a method of spatial sampling, leveraging

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

GPS devices to improve coverage in sample surveys of poorly documented populations. Their specific application involves an attempt to obtain a representative sample of urban Chinese residents, in an environment where the official registration lists are missing up to a quarter of the population, and leading to biased results. Theoretically, an unbiased estimate can be achieved if there exists a well defined geographic partition of the survey area into units, regardless of the population distribution, size, and shape of the units. A unit is chosen at random, and then *every* household in the area is enumerated, in the manner of traditional block listing, thus ensuring each household overall was chosen with equal probability. But this is extremely expensive for even moderately large units, as well as easier said than done in settings, like Beijing, where natural blocks or other partitions are difficult to define. GPS devices allow the construction of arbitrary sampling units – they construct a grid of "square seconds" – that interviewers can both find and enumerate in reasonable time. For comparison, Landry and Shen implement their procedure alongside a traditionally drawn sample used by the Beijing Area Study, and find overwhelming evidence of dramatic bias in the traditional procedure.

There is also an explicitly social interpretation of variety in Big Data, which is particularly well illustrated by the Landry and Shen paper. The social sciences are inherently interested in variation and diversity in human behavior and experience. The most interesting populations and phenomena, and those that will mislead us if ignored – Landry and Shen's undocumented migrants – are small and inherently more difficult to observe. Technological improvements in scale, scope, and representativeness of data improve our capacity to understand smaller populations. At the extreme, as seen in the Ansolabehere and Hersh paper, the collection of billions of observations enables insight into the behavior of specific individuals.

Velocity

Velocity refers to the speed at which data can or must be created and analyzed, an obvious challenge with Big Data. There are extreme examples in industry. Contemplate, for example, what must happen for Facebook to continuously update the model generating billions of news feeds based on millions of asynchronous posts, likes, and comments. We have few examples of political science models that attempt to ingest data and update in real-time, although there are efforts underway in the arena of event forecasting (Brandt et al. 2011).

At more typical social science speeds, velocity is a constraint instead in cases where our models cannot be estimated or evaluated in reasonable time with conventional means, and new algorithmic technologies can loosen that constraint. Common approaches involve breaking the computation into component pieces that can be solved in parallel, or finding ways to approximate solutions with simpler computations. For example, Markov Chain Monte Carlo (MCMC) approaches to the estimation of Bayesian models have created increasingly common velocity problems in political science. On the one hand, Bayesian estimation became practical only when MCMC became computationally feasible. On the other hand, MCMC scales poorly with model complexity and data set size. In the fifth paper in the issue, **Grimmer (2011)** describes a technique developed in computer science – *variational inference* (Jordan et al. 1999) – for the efficient, but approximate, estimation of Bayesian posteriors in complex models with large data. He notes that for some models MCMC methods may not just be too slow for the impatient, but too slow to be true, never converging or, worse, converging to the wrong posterior. In a particular compelling example, he demonstrates how variational inference allows for the "fast approximation to the true posterior" of a nonparametric Bayesian topic model that would defy MCMC estimation.

Vinculation

I offer *vinculation* – to "vinculate" is to bind together, to attach in a relationship – as the fourth V, in order to emphasize the fundamentally interdependent nature of social data.[3] Social data are data about interactions – conflict, transaction, communication – and the questions asked in social science disciplines like political science, sociology, and economics are often focused directly on understanding not the individuals but their interactions. Moreover, much of political methodology, and of statistics, is premised on the need to correctly account for interdependencies like serial correlation, spatial correlation, clustering, hierarchy, confounding, SUTVA violations, and so on.

[3] As Stephen Colbert would say: "Trademark."

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

Social networks are a particularly clear example, as the focus on relationships and interdependencies is baked in from the beginning, and it takes almost no effort to bump against challenges of volume, velocity, and variety with social network data. In the sixth paper in the issue, **Cranmer and Desmarais (2011)** describe, critique, and extend the Exponential Random Graph Model (ERGM) for statistical inference on network data. While network data are increasingly common, and big social data very frequently have a relational structure built in, most exercises in social network analysis are descriptive. A typical example might involve the calculation of network centrality measures, perhaps then used as a variable in a subsequent analysis of influence. ERGMs, and extensions described by Cranmer and Desmarais, represent an effort to take both network structure and statistical inference seriously, demonstrating both the potential in the approach and its remaining pitfalls. They provide two examples – one of bill cosponsorship in Congress and another of international conflict. The latter example, in particular, makes a compelling case that modeling the data as a network, rather than as a set of conditionally independent dyads, leads to different conclusions.

Validity

Issues of *validity* are at the heart of the scientific enterprise. Is it true? How true? How sure are you? How do you know? Political scientists, computer scientists, and now data scientists are already defensive from the endless snark that "anything that calls itself 'science' probably isn't" (Searle 1986), and statisticians have to deal with that "lies, damned lies" business. We approach the process of learning from data so differently that sideways glances of condescension and confusion are par for the course. But this is the fundamental issue. How do we use data to learn true things about the political and social world from Big Data, and how do we know?

Political science has hardly settled this issue itself. Many readers of the journal will skew too young to recall the ferocity with which theorists and empiricists tore at each other throughout political science in the 1990s, reaching an uneasy (and at best partially-accepted) detente under the NSF-enabled rubric of "EITM": Empirical Implications of Theoretical Models (Granato and Scioli 2004). In essence, this represented a truce loosely based on a deductive positivist approach to science, under which theorists would concede that their models should almost always be testable and tested, and empiricists would concede that theory should almost always come first.

The data science community has reached no such agreement. One clue to variations in the philosophical bent of practitioners is the variety of labels given the collection of methods – support vector machines, random forests, etc. – discussed in Hastie et al. (2009). The label *statistical learning*, common among statisticians, is most likely to indicate common ground with political methodologists on notions of inference. The label *machine learning*, common among computer scientists, indicates more interest in the efficiency and accuracy of algorithms, and less concern for interpretation of causality or measures of uncertainty. The label *data mining*, unfortunately most common of all, sends shivers up social science spines. This is partly semantic coincidence, but isn't entirely inapt. Many data science exercises are in fact almost entirely inductive, and proudly so. One data scientist, Kaggle president Jeremy Howard recently asserted that prediction challenges in data science had demonstrated that 'Specialist knowledge is useless and unhelpful' (Aldhous 2012), a direct challenge to the primacy of theory.

The seventh paper in the virtual issue, **Ratkovic and Eng 2010**, is illustrative of a familiar enough setting, from time series, where induction is clearly appropriate. Ratkovic and Eng consider the problem of finding discontinuities – "jumps," or "critical events" – in time series that are otherwise smooth. Within time series, these are typically referred to as "structural breaks" (Caldeira and Zorn 1998) or "change points" (Spirling 2007). Absent a means of detecting a structural break, a smoother like loess dismisses jumps as noise and under-reacts, flattening the curve. Most of these smoothers ignore the unidirectionality of time as well, allowing time series to react prior to the break. Ratkovic and Eng start with their "9/11 problem," in which the smoother becomes a Truther, detecting a rise in presidential approval prior to the attack. Ratkovic and Eng propose an alternative sequential segmentation method that performs well relative to alternatives in simulations, but there are limited theoretical bases – the premise here is that a process is smooth except when it's not – on which to judge results estimated from sufficiently noisy real data.

A common approach to validation in data science is out-of-sample prediction, particularly compelling if these are genuine forecasts of future events. In the eighth paper of the virtual issue, **Montgomery, Hollenbach, and Ward (2012)** introduce, extend, and apply the ensemble Bayesian model averaging (EWBA) approach of Raftery et al. (2005) to out-of-sample forecasting examples from international relations and American politics. As Montgomery et al. note, genuine forecasting is relatively rare in political science. The primary examples are in the event data community in

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

international relations (Brandt et al. 2011; Gleditsch and Ward 2013), and election forecasters. In the 2012 presidential election, several political scientists did exactly what Montgomery et al. cite as "nearly nonexistent" elsewhere: create forecasts built from combinations of multiple forecasts. The forecasts of Drew Linzer, posted on `votamatic.org`, and based on a model described in Linzer (Forthcoming), and of Simon Jackman, posted on *HuffPost Pollster*, were among the few who called every state correctly. Conversely, prediction is the bread and butter of Big Data. The canonical data science exercise is a "collaborative filtering" or "recommender" system, the engines behind the "you might also like" recommendations from Amazon, Netflix, and similar services. This is an exercise in prediction: based on what you've liked in the past, what might you like in the future? Popularized by the $1 million Netflix Prize, challenging all comers to improve its movie rating prediction model, such prediction contests are now common (and often managed by `Kaggle.com`, which is how Howard comes by his specialist knowledge on the subject). While it is arguable how scientifically useful these exercises are – even Netflix didn't use the solution it paid $1 million for – the discipline imposed by true out-of-sample validation is noteworthy and this is a challenge to be taken seriously.

It is also worth noting that these contests are nearly always won by ensemble methods. Broadly, the insight of ensembling is to combine the forecasts of multiple models – ideally, very different models – in a weighted average, producing a single forecast more accurate than that of any of the component models. The logic is similar to that which drives arguments about the merits of diverse inputs in decision-making (Page 2011). Methods vary in how the weights are calculated. EBMA assumes an ensemble of ongoing forecast models, so there is a set of past predictions for each. EBMA calculates weights based on this past performance, giving heavier weight to forecasts that have been more correct, and particularly heavy weight to forecasts that have been uniquely correct. The component models need not even be stochastic. Montgomery et al. offer three compelling examples of EBMA applications in political science with very different technical and substantive features.

Conversely, social science approaches to validity are relevant to Big Data as well. Colleagues and I have argued elsewhere (e.g., Quinn et al. 2010) that many problems in data science – including topic modeling and collaborative filtering – are measurement exercises that could and should be subjected to the same standards of measurement validation we would apply in social science. One overarching standard is construct validity: the degree to which the technique measures what it claims to be measuring. One of the most compelling social science approaches to construct validity is the estimation of measurement models that closely adhere to an underlying theoretical mechanism. Examples include the tight match between ideal point estimation models based on an underlying spatial theory of voting, or between unfolding models based on an underlying spatial model of perception. In the ninth paper in the issue, **Bakker and Poole (2012)**, describe a Bayesian approach to metric multidimensional scaling (MDS) problems, including scaling of similarities / dissimilarities data and the unfolding problem. Conventional approaches to such problems are unable to produce estimates of uncertainty, while uncareful Bayesian MCMC approaches are flummoxed by the identification problems inherent in the MDS problems. In particular, the posterior has mirror image reflections in each dimension, rendering the posterior mean of every spatial coordinate at zero. Bakker and Poole provide a computationally straightforward solutions to these problems, computing scalings recovered for US Senators using rollcall similarity scores, and unfoldings for survey respondents who provided feeling thermometer ratings of presidential candidates.

The unfolding problem is almost identical in its starting point to the collaborative filtering / recommender system problem. In the Bakker and Poole feeling thermometer example, the input data is a matrix with one row for each of 1392 survey respondents, one column for each of 12 presidential candidates, and each non-missing cell a rating by the respondent of the candidate on a 100-point feeling thermometer. The Netflix Prize (training) data are a (sparse) matrix of 480,189 customers by 17770 movies, with each non-missing cell a rating on a 5-point scale.[4] Despite the substantial similarities, the questions that are asked of these and other similar datasets are quite different, as is the evidence for validity, suggesting considerable room for interdisciplinary arbitrage.

Conclusion

I conclude with a discussion of the final paper in the issue, **Clark and Lauderdale (2012)**, which touches on almost all of these challenges and opportunities facing Big Data political science.[5] Clark and Lauderdale posit a theoretical

[4]Each rating also has a time stamp, which is also relevant for that problem.

[5]Indeed, the Clark and Lauderdale paper covers so much ground, with so much originality and technical facility, that it may be altogether unciteable.

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

structure by which legal doctrine develops, which they term the "genealogy of law." The analogy to a family tree is apt in that it captures the structural reality of each case as a "child" of a chronologically prior "parent," and inapt in that a case may have only one parent and multiple children. The statistical model builds on the bag-of-words modeling tradition seen elsewhere in this issue with text data (Monroe et al.; Grimmer), modified to capture this structural constraint.

Clark and Lauderdale deal with volume, in at least three ways. The source data for the study are the texts of 18713 Supreme Court opinions, which had to be downloaded and parsed for citation information via Perl script. A decade ago, this step passed for magic and required extensive explanation, but here merits only a few sentences. The resulting numerical data that feeds the statistical model is the 18713×18713 matrix of citations, a data set of over 35 million counts (17 million if the impossibility of citation forward in time is enforced). Clark and Lauderdale reduce this by operating on substantively meaningful subsets of the data, also taking advantage of sparsity (most cases don't cite most other cases) in the data. But then we reach the third element of volume: the model space. The number of possible trees on any one data set is $(n-1)!$. So, for example, the relatively small subset of abortion cases, $n = 42$, could be generated by one of 3.3×10^{49} possible citation trees.

Clark and Lauderdale thus have to address velocity. It's one thing to write their model down, but estimating it in finite time requires some computational innovation. The space is much too large for conventional MLE algorithms and too multimodal for conventional MCMC. They develop an alternative independence Metropolis approach which allows for reasonable tradeoffs. Among other features, the algorithm preserves work done on subtrees when it chooses to explore a solution in which that subtree is effectively grafted onto an alternative precedent. This is clever, and has potential application in other partially discrete problems, including the linkage problem discussed above. As with Grimmer, Montgomery et al., and Bakker and Poole, the world of Big Data political science requires concern for computation and the scaling of algorithms, a subject of study rarely touched upon in political science training.

Clark and Lauderdale worked with variety. The final analysis shown, in Table 1, is an OLS regression with 950 observations and six covariates. This is data that fits neatly in a small rectangular spreadsheet. But consider all the data types and structures traveled through to reach that point – text, time, networks, sparse matrices, dependency trees – that do not fall into the conventional rectangular spreadsheet framework. As with Landry and Shen, Berinsky et al., and Cranmer and Desmarais, we see the potential and challenges associated analysis of novel forms of data. As with Ansolabehere and Hersh and Monroe et al., we also see again that one payoff of Big Data for political science is not overkill statistical power from billion-scale N, but finer grain insight obtained from large-enough N over a variety of contexts.

This is an excellent example of the importance and challenges of vinculation. The basic unit of data is an interconnection – a citation – between two cases. This is network data, but the statistical and theoretical models are premised on a particular class of structural dependence among the cases. An ERGM, for example, could not recover this structure. The primary theoretical object of interest is a characterization of those interdependencies and how they cohere as legal doctrine.

Finally, the paper features multiple efforts to demonstrate validity. First, as with Bakker and Poole, we see a strong focus on construct validity, in which the measurement model is directly built from a substantive theory of the data-generating process. Second, we see a demonstration of face validity, that obvious expected structure is recovered and recovered structure is not nonsensical. Many seem to be dismissive of face validity, as "it looks reasonable" is certainly an insufficient criterion for validity. The point is rather that this should be a *necessary* criterion. If we look at the results and they are clearly implausible – we see examples of this in Ratkovic and Eng Figure 1; Monroe et al., Figures 1-3, Bakker and Poole Figure 4 – then we know a measure lacks validity. This is a challenge with Big Data, often requiring visual forms of communication. Third, we see two tests of model fit, reinforcing the validity of the structural model. Fourth, as with Bakker and Poole and Ratkovic and Eng, we see several discussions of discriminant validity, demonstrating that their measure is different from what has come before in the places where it should be different. Finally, and most important, we see a demonstration of what is variously called hypothesis or criterion validity, that the measure can be used to learn something new about a political phenomenon of interest. Similar efforts conclude all of the papers here, offering the "So What?" that not only validates the exercise, but establishes that the exercise has a (political) scientific point.

Taken together, I hope these papers demonstrate that there is tremendous potential for political science to benefit from Big Data, and vice versa.

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

References

Aldhous, Peter. 2012. "Specialist knowledge is useless and unhelpful." *New Scientist* December 7.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. 2011. "Real time, time series forecasting of inter- and intra-state political conflict." *Conflict Management and Peace Science* 28(1):48-64.

Caldeira, Gregory A. and Christopher J. W. Zorn. 1998. "Of time and consensual norms in the Supreme Court." *American Journal of Political Science* 42:874-902.

Cantú, Francisco, and Sebastián Saiegh. 2011. "Fraudulent democracy? An analysis of Argeninta's infamous decade using supervised machine learning." *Political Analysis* 19(4):409-433.

D'Orazio, Vito, Stephen T. Landis, Glenn Palmer, and Philip A. Schrodt. 2011. "Separating the wheat from the chaff: Application of two-step support vector machines to MID4 text classification." Paper presented to the Midwest Political Science Association, Chicago, April.

Firebaugh, Glenn 2008. *Seven rules for social research*, Princeton, NJ: Princeton University Press.

Gleditsch, Kristian Skrede, and Michael D. Ward. 2013. "Forecasting is difficult, especially about the future." *Journal of Peace Research* 50(1):17-31.

Granato, Jim, and Frank Scioli. 2004. "Puzzles, proverbs, and Omega matrices: The scientific and social significance of empirical implications of theoretical models (EITM)." *Perspectives on Politics* 2(2):313-23.

Hallberg, Johan Dittrich. 2012. " PRIO Conflict Site 1989-2008: A geo-referenced dataset on armed conflict." *Conflict Management and Peace Science* 29(2):219-32.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*, second edition. New York: Springer.

Jordan, Michael, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. 1999. "An introduction to variational methods for graphical models." *Machine Learning* 37:183-233.

Laney, Doug. 2001. "3D management: Controlling data volume, velocity, and variety." *Application Delivery Strategies*, META Group, Inc., Feb 6.

Linzer, Drew. Forthcoming. "Dynamic Bayesian forecasting of presidential elections in the states." *Journal of the American Statistical Association*.

Page, Scott. 2011. *Diversity and complexity* Princeton: Princeton University Press.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209-28.

Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. 2005. "Using Bayesian model averaging to calibrate forecast ensembles." *Monthly Weather Review* 133:1155-74.

Schrodt, Philip and Jay Yonamine. 2012. "Automated coding of very large scale political event data." Presentation to the New Directions in Text as Data Workshop, Harvard University, October.

Searle, John R. *Minds, brains and science (1984 Reith lectures)* Cambridge: Harvard University Press.

Spirling, Arthur. 2007. "Bayesian approaches for limited dependent variable change point problems." *Political Analysis* 15:387-405.

von Ahn, Luis, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. "reCAPTCHA: Human-based character recognition via web security measures." *Science* 321(5895):1465-8.

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press

Webb, Eugene J, Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. 1966. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.

About the Author: Burt L. Monroe is an Associate Professor of Political Science, and Director of the Quantitative Social Science Initiative and the Big Data Social Science Program at Pennsylvania State University. This work was supported in part by National Science Foundation IGERT Grant DGE-114860, "Big Data Social Science."

https://doi.org/10.1017/S1047198700014315 Published online by Cambridge University Press