Big Data: Challenges and Opportunities for Comparative Historical Sociology

Bart Bonikowski
Harvard University

Large-scale digital data and related computational methods—often grouped under the catchall label of "big data"—have recently become a focus of active debate in our discipline. In academic conference panels, special journal issues, and graduate admissions and hiring committees, sociologists are weighing the pros and cons of these new approaches to empirical analysis. The exchanges are frequently enlightening, but just as often, they reveal tensions between impassioned views typical of early stages of innovation: proponents promise that big data will revolutionize social science and detractors warn that it will encourage method fetish and privilege theoretically uninteresting research questions.

Yet, the choice between embracing and rejecting these new developments is a false dichotomy: big data is already a reality in the social sciences and humanities, and its importance is only likely to grow. It is time for sociology, which has lagged behind other fields in adapting to this reality, to move past abstract programmatic debates and begin developing best practices for carrying out and evaluating big data research. Comparative historical sociology in particular stands to gain much from these developments, as massive volumes of digitized archival material become increasingly accessible. Moreover, this subfield's experience with non-random samples, data shaped by substantively important historical processes, and the difficulties associated with cross-case comparison and the temporal continuity of units of analysis—as well as a wealth of case knowledge—place comparative historical sociology in an advantageous position for interrogating big data research on conceptual and substantive grounds.

Even though comparative historical sociology is often identified with small-N qualitative research, scholars in this field have been actively making use of large-scale digital sources. This has enabled analyses of such wide-ranging topics as social network formation in early-modern overseas trade (Erikson 2014); the reshaping of mainstream discourse by radical movements following national crises (Bail 2014a); long-term shifts in the contours of national political discourse (Rule, Cointet, and Bearman 2015); the role of nation-state formation in promoting military conflicts (Wimmer and Min 2006); fluctuations in populism among U.S. presidential candidates (Bonikowski and Gidron 2016); and local variation in belief structures within the women's rights movement (Nelson 2015). While the analytical strategies employed in these studies vary, they all make use of the affordances of big data: the ability to compile large compendia of micro-level observations, link records across sources, aggregate the data to various desired levels of analysis, inductively detect empirical patterns otherwise hidden to the researcher, and carry out these complex operations more quickly and at a larger scale than would have been possible using traditional methods.

> ***It is time for sociology, which has lagged behind other fields in adapting to this reality, to move past abstract programmatic debates and begin developing best practices for carrying out and evaluating big data research.***

These approaches are not without limitations, of course. Whether the data take the form of political texts, as in studies of discourse, or of behavioral traces, as in studies of network interaction and institutional practices, powerful computational algorithms are able to produce results quickly and efficiently. This can create disincentives for delving deeply into case knowledge and verifying the validity of the automated output. It is precisely this kind of analytical reflection and painstaking validation that separates high-quality research from methodological ostentation. As Grimmer and Stewart (2013) repeatedly remind us in their excellent overview of computational text analysis, big data methods are not a replacement for scholarly elbow grease: without extensive validation, algorithms cannot be trusted (I would add that the data themselves should be viewed with equal doses of healthy suspicion). It is up to the peer review process to ensure that researchers do not take shortcuts when working with big data, particularly when exciting new methods can serve as a distraction from rigorous scholarship.

Big data raises two other challenges that are relevant to comparative historical research. The first concerns the question of what constitutes a unit of observation (Wagner-Pacifici, Mohr, and Breiger 2015). Big data scholars work with massive samples—and often entire populations—which can create the appearance of exhaustiveness and with it, empirical legitimacy (even as it creates problems for standard statistical inference methods). But it is worth remembering that these large data sets often originate in very specific organizational settings that may not generalize to other contexts. For instance, should we think of millions of tweets as representing millions of cases of public speech or as a high-resolution view into a single case, that of Twitter? Are 100,000 political speeches from the European Parliament, as in my ongoing research on populism in legislative discourse, representative of European politics or of the Parliament as a particular institution? The extensiveness and precision of large-scale digitized data is all-too-easily mistaken for an unmediated representation of entire domains of social reality, when in fact the data are likely to bear a strong imprint of their particular conditions of production (Adams and Brückner 2015; Bail 2014b).

Second, one of the advantages of big data is that it lends itself to inductive inquiry. But this raises an epistemological question: who (or what) is the better judge of the validity of the patterns inductively observed in data, the algorithm or the analyst? The call for extensive validation of automated methods (Grimmer and Stewart 2013) assumes that human coding is the gold standard that algorithms can at best approximate, because of human coders' sensitivity to nuance and context. And yet, human coding is prone to extensive biases that are less likely to affect algorithmic estimates, especially those generated by unsupervised models (DiMaggio 2015). This has led Lee and Martin (2015) to propose that formal methods—ideally those that reduce the complexity of the data without resorting to arbitrary coding schemes—be privileged over hermeneutic approaches to text analysis. These are provocative claims, but the cartographic approach advocated by Lee and Martin (2015) cannot do away with interpretation altogether (even if it does make it more transparent), and so the quandary remains: if there is no such thing as an analytically neutral position vis-à-vis data, how should we evaluate our models?

These and other critiques are often leveraged against computational approaches to large-scale digital data, but it is important to realize that similar doubts about the completeness of data, the conditions of their production, the choice of units of analysis, the challenges of induction, and the non-representativeness of samples can be raised against more traditional research designs, from surveys and archival work to interviews and ethnographies. That we often do not question these approaches is less a reflection of their robustness than their extensive institutionalization (but see Biernacki 2012; Jerolmack and Khan 2014). The fact that recent advances in data collection and analysis are challenging our taken-for-granted assumptions, therefore, should not be seen solely as a problem for big data research, but also as an opportunity to critically reflect on the shortcomings of all forms of empirical research. This, in fact, may be one of the most valuable unintended consequences of the big data revolution.

I started this essay by arguing that hand wringing about the potential pitfalls of big data should give way to the positive development of best practices for research. What might those look like? Scholars are still working out informal guidelines, but I would tentatively suggest six, three for data collection and three for analytical methods. First, given that most big data is found data, published work should explicitly consider the organizational and technological conditions of the data's production and how those conditions affect what the data are able to reveal about the social world. Second, because data curation (i.e., their collection, cleaning, and organization) is the most arduous aspect of big data research and one that involves considerable researcher discretion (Diesner 2015; DiMaggio 2015), every step in the curation process should be documented and presented to readers in methodical appendices. Third, caution should be taken with generalization based on populations of observations produced in highly specialized settings; such limitations should be clearly reflected in statements of studies' scope conditions. Fourth, while big data can reveal new empirical patterns, it should not serve as a source of methodological hubris; all methods obfuscate aspects of social reality and triangulation across multiple approaches remains the best remedy against analytical myopia. Fifth, scholars should be careful with and forthright about the interpretive steps involved in inductive inquiry (Lee and Martin 2015); openness to being surprised by data is a virtue, but getting duped by faulty algorithms is not. Careful and honest abduction—whereby surprising patterns observed in data are used to revise theory and then the updated theory is brought back to bear on the data—may offer one solution, however imperfect, to this problem (Goldberg 2015; Timmermans and Tavory 2012). Finally, there is no such thing as free lunch in empirical research: while computational methods may process millions of observations with unprecedented speed, they are no substitute for painstaking and time-consuming validation, the results of which should be carefully documented and reported (Grimmer and Stewart 2013).

For all its significant limitations, big data offers us unique opportunities to study old problems in new ways, to occasionally pose new, previously unanswerable questions, and to carry out research more efficiently than in the past. This suggests that one concern often expressed by detractors of these approaches is largely misplaced: big data does not inherently produce atheoretical research that tackles uninteresting questions. Bad research is not a function of types of data or methodological approaches; faulty and uninteresting studies can be found in every subfield and in every research tradition. While early attempts at big-data sociology may have been preoccupied with showcasing methods for their own sake, this research approach has matured and is increasingly producing sophisticated, interesting, and important studies. It is up to the scholarly community to encourage this trend by acknowledging the tremendous potential of big data, while holding its practitioners to the same exacting theoretical and empirical standards expected of other traditions. In the meantime, we would all do well to use the discussions surrounding big data to question the thoroughly institutionalized practices of other methodological approaches, both qualitative and quantitative. Comparative historical sociology and the discipline as a whole will be better off as a result.

References

Adams, Julia, and Hannah Brückner. 2015. "Wikipedia, Sociology, and the Promise and Pitfalls of Big Data." *Big Data & Society* 2(2):1–5.

Bail, Christopher A. 2014a. *Terrified: How Anti-Muslim Fringe Organizations Became Mainstream*. Princeton, NJ: Princeton University Press.

Bail, Christopher A. 2014b. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43:465–82.

Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables*. New York: Palgrave MacMillan.

Bonikowski, Bart, and Noam Gidron. 2016. "The Populist Style in American Politics: Presidential Campaign Rhetoric, 1952-1996." *Social Forces*. 94:1593-621.

Diesner, Jana. 2015. "Small Decisions with Big Impact on Data Analytics." *Big Data & Society* 2(2):1-6.

DiMaggio, Paul. 2015. "Adapting Computational Text Analysis to Social Science (and Vice Versa)." *Big Data & Society* 2(2):1-5.

Erikson, Emily. 2014. *Between Monopoly and Free Trade: The English East India Company.* Princeton, NJ: Princeton University Press.

Goldberg, Amir. 2015. "In Defense of Forensic Social Science." *Big Data & Society* 2(2):1-3.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21:267–97.

Jerolmack, Colin, and Shamus Khan. 2014. "Talk Is Cheap." *Sociological Methods & Research* 43:178–209.

Lee, Monica, and John L. Martin. 2014. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3(1):1–33.

Nelson, Laura K. 2015. "Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City." Working Paper. Department of Sociology and Anthropology, Northeastern University.

Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman. 2015. "Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790–2014." *Proceedings of the National Academy of Sciences* 112:10837–44.

Timmermans, Stefan, and Iddo Tavory. 2012. "Theory Construction in Qualitative Research From Grounded Theory to Abductive Analysis." *Sociological Theory* 30:167–86.

Wagner-Pacifici, Robin, John W. Mohr, and Ronald L. Breiger. 2015. "Ontologies, Methodologies, and New Uses of Big Data in the Social and Cultural Sciences." *Big Data & Society* 2(2):1–11.

Wimmer, Andreas, and Brian Min. 2006. "From Empire to Nation-State: Explaining Wars in the Modern World, 1816–2001." *American Sociological Review* 71:867–97.

Tools for Historical Sociologists

Christopher Muller
University of California, Berkeley

There is no substitute for the intuition for a historical period you can get by immersing yourself in primary and secondary source material. Robert Fogel (1982: 51) was right to acknowledge that "No amount of mathematical wizardry or computer magic can shortcut this process." But one of the distinct advantages of historical sociology relative to historical research in other social sciences is that it recognizes the importance of both primary-source qualitative work and historical data analysis of other kinds. Learning a little bit about some new tools for historical data collection and analysis can both speed up your archival research and allow you to supplement more traditional archival work with data that only recently would have been too laborious to collect or construct. In this essay, I will describe some ways historical researchers have extracted data from maps and linked people across multiple records. I'll then offer some brief thoughts on why I believe these tools and the approach to studying history that they facilitate are important.

One problem historical researchers often encounter is finding high-quality representative data about the past. Where data of the kind we are used to working with—censuses, surveys, events, and so forth—are unavailable, we can sometimes find helpful information in maps