Commentary

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–5

DOI: 10.1177/2053951715612497
bds.sagepub.com
SAGE

Big Data and historical social science

Peter Bearman

Abstract

"Big Data" can revolutionize historical social science if it arises from substantively important contexts and is oriented towards answering substantively important questions. Such data may be especially important for answering previously largely intractable questions about the timing and sequencing of events, and of event boundaries. That said, "Big Data" makes no difference for social scientists and historians whose accounts rest on narrative sentences. Since such accounts are the norm, the effects of Big Data on the practice of historical social science may be more limited than one might wish.

Keywords

Big Data, historical social science

Pretty much at the same time as network scientists have been discovering power laws and other connectivity-based epiphenomenal features of social networks in the digital exhaust of on-line searches, social media friendship choices, purchase recommendations, product reviews and the like, historians have been quietly building massive archival data structures from the extant records of crucially important institutions and contexts and making those data structures available to the public. These include—and this is only an idiosyncratic sample drawn mainly from Britain—the complete text record of the Old Bailey, the extant records of the Atlantic slave trade, and the British East India Company. With the simplest of search strategies answers to descriptive questions that the lone historian poring through a dusty archive could barely imagine possible are now achievable in a matter of days, perhaps hours, or even minutes. *How will these Big Data archives and many others change our understanding of the past, and hence our present?*

As these archives come on-line, with even less fanfare than the newest study showing that the day length has an influence on people's emotions indexed by the proportion of happy over sad tweets, vast textual corpora spanning long periods of time are now becoming widely available in digital form. And not surprisingly, these textual corpora (once again providing only an idiosyncratic sample from the United States—State of the Union (SoU) speeches, the Congressional Record, transcripts of Supreme Court decisions) also arise from important institutions whose form is largely continuous over often very long periods of time.[1] *Can we answer old questions in new ways from these old texts?*

Similarly, cultural objects such as photographs, menus from New York City restaurants, seating charts for the New York Philharmonic, also previously largely inaccessible, are increasingly made available in on-line repositories (Accominetti and Khan, 2015). Most of these data are precisely time-stamped and much of it can be precisely geo-referenced. *Do these new cultural repositories make possible changed understandings of crucially important dynamics such as elite formation, ethnic group integration, disease diffusion and even long-term climate change?*

And we are just at the beginning of the archival revolution. Historical archives of enormous significance are being digitized rapidly. Equally amazing, new archives of unknown importance but on a massive scale—for example, the two billion annual emails to and from the department of state under Clinton—dwarf the number of unique documents that ever existed from

Columbia University, New York, USA

Corresponding author:
Peter Bearman, Columbia University, 606 W. 122nd Street, New York, NY 10027, USA.
Email: psb17@columbia.edu

the entire Lincoln administration.[2] *Does any of this matter for what historians and historical social scientists do?*

Not surprisingly, the answer is no and yes.

Golder and Macy report (2011), as quoted in this volume by Breiger, that "the web sees everything and forgets nothing". As a fantasy this is not so different from the ideal chronicle that records everything "just as it happened, as it happened" imagined in 1968 by Arthur Danto (1968) who wondered whether such a chronicle would change anything of significance that historians do. Danto acknowledges that after doing whatever historians do in the archives that the existence of such a chronicle would allow them to instantly check their facts, but this is at best a trivial element of their practice. What historians do uniquely is write narrative sentences—sentences such as "on Christmas Eve 1642, the father of modern physics was born"—and a chronicle that recorded everything that happened just as it happened would not be of any use for such an activity, for only after modern physics was born could Newton be its father. The web or the chronicle that saw everything and forgot nothing would record absolutely faithfully that on Christmas Eve 1642 Isaac Newton was born. But it could never record at the moment that some hundreds of years later he would father modern physics. So for this central activity of the historian's craft—the writing of narrative sentences—our new data structures will be of little importance.

But for a social science history not concerned with writing such sentences, these data structures may induce across a wide array of substantive problems of note a radical re-imagination of the past, and consequently a radical reorganization of our understanding of the present because they invite and provide supports for a new kind of history, one focused less on dynamics, pattern recognition, and the identification of the mechanisms by which the actions of actors cumulate into macro-level outcomes. Think of our deepest intellectual problems—for example, something like the emergence of "modern society". What could Big Data contribute to our understanding?

Jose Atria—a PhD student at Columbia—is working on using the Old Bailey archive to answer just that question (2015). The starting point for this project is his observation that historians and social scientists agree that a whole set of process widely believed to have started in the 16th century resulted in the emergence of modern society. There is no question about this. Atria (2015) writes:

> These include changes like the emergence of ways of life that made violence and the body seem repugnant and shameful (Elias, 2000), the development of centralized state authority and the emergence of national communities beyond direct bonds of kinship (Bearman, 1993; Anderson, 2006; Tilly, 1975), the development of market society (Polanyi, 2001), the dissolution of traditional social relationships based on status in favor of a social structure based on social classes (Thompson, 1963), and the emergence of disciplinary methods of control and power (Foucault, 1977).

Others could be added, of course, but this is a sufficient set for our purposes and the key issue is not that there may be others, but that even though there is agreement about the fact that these processes are central, there is very little agreement about the structure of the causal connections between them.[3] And answering that question—which can be decomposed into a series of other critical questions, such as: How did the various semi-autonomous processes involved in the emergence of modern society interact with one another? Which changes came first? Which followed either in timing of onset or speed of change?—is where "Big Data" (in this case, the Old Bailey archive) comes into play.

Importantly, the Old Bailey archive arises directly from an institution that is absolutely central to all of the critical processes that constitute modernity's emergence. Other institutions play equally critical roles. An often observed character of historical data and also true for big historical data that promises to change what we know about causal dynamics in historical context is that it arises from institutional contexts strongly implicated in those processes. Things that were preserved more often than not are things that people then believed worth preserving. The archive, which covers the period from 1674 to 1913 (this is, after all, a period bounding the emergence of modernity) contains precise time-stamped data on millions of unique persons, thousands of places, and hundreds of thousands of interactions, both criminal and quotidian. Indexed in these records are precise relational data structures that enable one to focus on changing patterns of social relations, the social geography of violence, the displacement of kinship, the breakdown of localisms of all sorts, and the emergence of new strategies and discourses about governance. And from this, one can identify the pace of institutional change, the causal dynamics underlying the complex cluster of interlocked and tied processes that constitute modernity. Or at the least, one can anticipate that such an analysis—otherwise impossible—is within reach.

Or consider a related problem: how should we case historical event sequences; that is, how should we induce periods? Newly available textual corpora spanning long periods of time provide a powerful setting for identifying turning points. As with the previous discussion, American historians agree that at some time

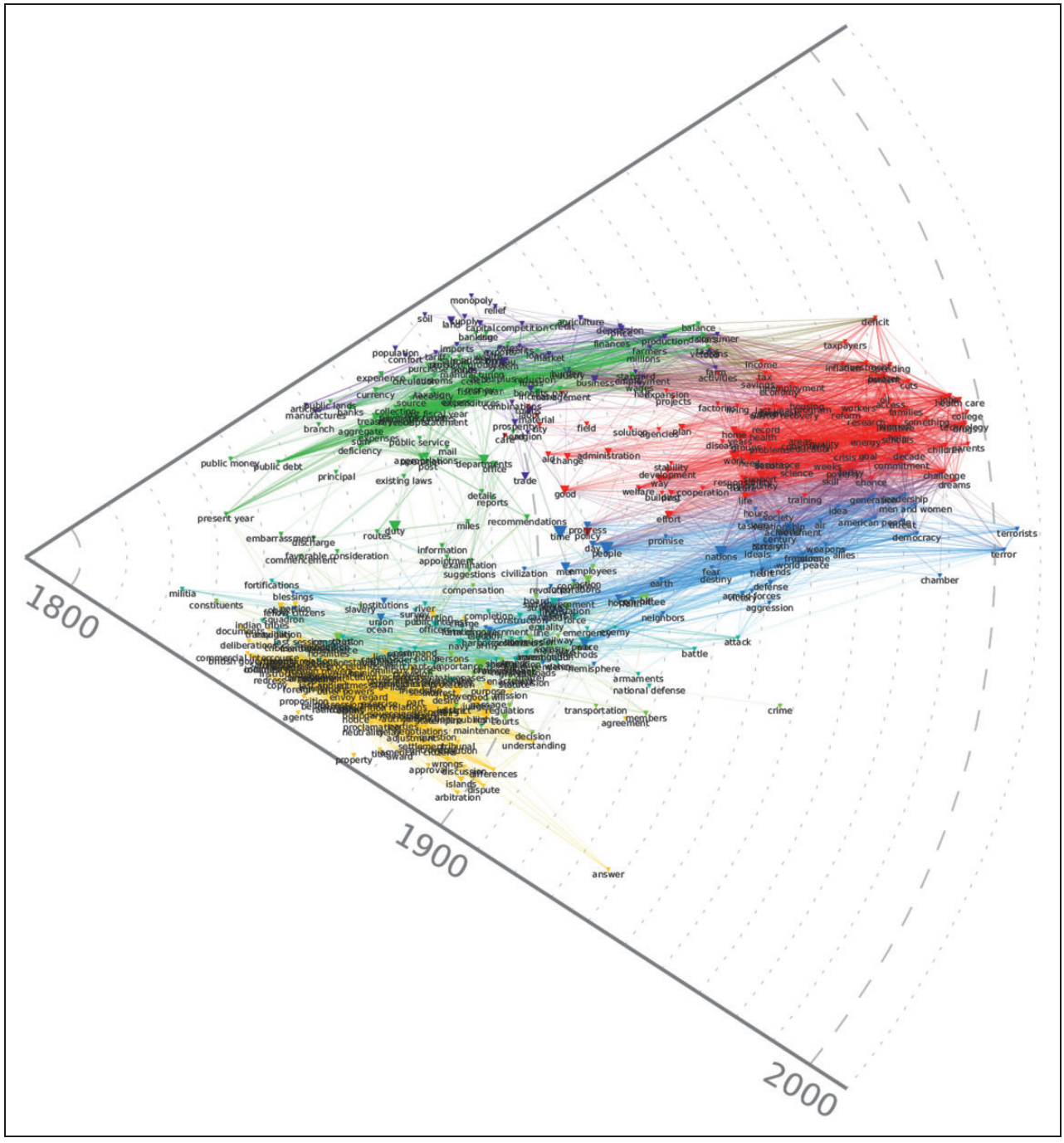

Figure 1. Radar Plot of the State of the Union Speeches: Words are located temporally when 50% of their occurences have taken place.

American social and political discourse over the course of the more than two centuries from 1790 to the present changed. They also agree about the nature of this change—that one can with confidence distinguish modern discourse from earlier forms. Yet when this change happened and how we can best identify it are matters of significant debate. The canonical explanations center on 1865, 1877, 1896, and 1913, but just about any year has a proponent! This debate necessarily implicates scholarly strategies for casing history, that is, strategies for identifying partitions in the necessarily uninterrupted flow of the past sufficient to meaningfully induce discrete temporal clusters such as 'the gilded age'', ''the progressive period'', ''the age of the frontier'', and so on. The partitions that one selects depend, of course, on the standpoint from which one

starts. The history of America as a country of immigrants induces a different partition from the history of America as a loose confederacy of states. Given this, the challenge is generating a meaningful partition between modern and pre-modern social and political discourse where the *contents* of that discourse are rhetorical devices for periodization! To do this, one needs to identify a *form* that is continuous, and of continuous significance. Enter Big Data.

Drawing from the archive of Presidential SoU addresses, Rule et al. (2015) show that the annual SoU addresses of Presidents provides, despite changes in mode of delivery, one such form, and thus provides an Archimedean point from which they can view the content and structure of American social and political discourse unfolding over time. Using new strategies for analyzing historical texts in which terms, concepts, language use and words' meanings change over time, Rule et al. show that modern social and political discourse over the nature of governance emerges as a distinct object *after* 1917, though elements of such discourse are identifiable earlier. This provides new insights into our understanding of American history and transforms in its wake our understanding of what modern and pre-modern discourse looks like and, equally critically, the elements of continuity that link them. Along the way, they demonstrate a new strategy for identifying meaningful categories in textual corpora that span long periods of time. Specifically, their approach is able to account for the fluidity of discursive categories over time, and to analyze their continuity by identifying the conversational *stream* as the object of interest. That it gives rise to visualizations that allow for the identification of lexical change over the *longue duree* (as in Figure 1) is simply an additional benefit.

A whole class of new problems that could not be precisely articulated may be accessible with Big Data. Using a new data structure that precisely geo-references season ticket holders of the New York Philharmonic with respect to seat location (!) and residence, Accominetti and Khan (2015), for example, can show how the NYC elite used patronage of cultural institutions to recognize themselves, and how this capacity for joint recognition led to a subsequent reorganization of residential patterns in the City during the Gilded Age. One insight here is that the elite class came to see itself through joint observation across multiple settings, neighborhoods, concerts, and so on, while at the same time creating avenues for non-elites to circulate on the borders of the elite sub-world.

Likewise, Hoffman et al. (2015), using data from the Atlantic slave trade archive, bolstered with detailed occupational data gleaned from newly available geo-referenced street directories for Liverpool, are able to show why the Liverpool slave trade ownership network, was able to ride out failure caused by privateering and "defeat" London and Bristol for control of the Triangle Trade after 1750. In the same vein, Muller and his colleagues are currently digitizing prison incarceration data for Lawrence Mass, during the famed Bread and Roses strike, which, coupled with other administrative (census, mortality, hospital) data, will make it possible to model the diffusion of resistance in one of the canonical events in US Labor history (2015). All of these studies were theoretically possible before "Big Data", but none could be accomplished within the lifetime of a single scholar. The promise of such studies is both their contribution to the specific history with which they engage and the identification of mechanisms that may be transposable across context. Using Big Data to identify transposable mechanisms is, I believe, central. Otherwise, we are left with mere description, a useful project, but not a particularly useful sociological project.

These few examples barely touch the surface of what the Big Data revolution is making possible in historical studies. Whole arenas of work reliant on administrative data structures (like the census); institutional records (for example, state-level records on de-institutionalization of mental patients in the 1970–1985 period); data that link standard social science instruments with large-scale historical archives (like the GSS with the mortality schedules for the US, and so on) are not described herein. In so far as they share the characteristics of the projects described above: oriented towards answering causally important questions through the identification of transposable mechanisms, as versus writing narrative sentences, focused on identifying the pace and structure of change arising from key institutional locales known to undergird the large historical processes of interest, these Big Data structures can revolutionize historical social science. In so far as they capture merely epiphenomenal materials arising from non-central institutional locales, Big Data seems less promising. After all, what goes in comes out.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For the congressional record, cf. http://thomas.loc.gov/home/LegislativeData.php?n=Record; for the SoU, cf: http://www.presidency.ucsb.edu/sou.php; for Court Records cf; http://www.loc.gov
2. I am indebted to Matt Connelly for this observation.

3. By causal I mean the broadly Weberian sense of the term, where for example, the *Protestant Ethic and the Spirit of Capitalism* deals with one side of the causal chain connecting the spirit of capitalism with ascetic Protestantism, not the much narrower version of the term prevalent in the discipline today.

References

Accominetti F and Khan S (2015) The NY Philharmonic Project, Mellon Foundation. Available at: http://incite.columbia.edu/subscribers-to-the-ny-philharm (accessed 10 January 2015).

Anderson B (2006) *Imagined Communities*. London: Verso.

Atria J (2015) *Dissertation Proposal, Columbia Sociology*. Unpublished Manuscript.

Bearman P (1993) *Relations into Rhetorics: Elite Social Structure in Norfolk England, 1540–1640*. New Brunswick, NJ: Rutgers University Press.

Danto AC (1985) *Narration and Knowledge: Including the Integral Text of Analytical Philosophy of History*. New York, NY: Columbia University Press.

Elias N (2000) *The Civilizing Process: Sociogenetic and Psychogenetic Investigations*. Revised edition. Oxford: Blackwell Publishing.

Eltis D. A brief overview of the trans-atlantic slave trade voyages: The trans-atlantic slave trade database. Available at: http://www.slavevoyages.org (accessed 10 January 2015).

Foucault M (1977) *Discipline & Punish*. New York: Random House Digital, Inc.

Golder SA and Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051): 1878–1881.

Hitchcock T, Shoemaker R, Emsley C, et al. *The Old Bailey Proceedings Online, 1674—1913*. version 7.0. Available at: www.oldbaileyonline.org (accessed 24 March 2012).

Hoffman M, Makovi K and Bearman P (2015) *The structure of the atlantic slave trade ownership network*. Unpublished Manuscript.

Klingenstein S, Hitchcock T and DeDeo S (2014) The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences of the United States of America* 111(26): 9419–9424.

Muller, Srivastava CS, Makovi K, et al. (2015) *The Lawrence Bread and Roses strike, RWJ-HSS proposal*. Unpublished Manuscript.

Polanyi K (2001) *The Great Transformation: The Political and Economic Origins of Our Time*, 2nd ed. Boston, MA: Beacon Press.

Rule AR, Cointet JP and Bearman PS (forthcoming) Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences of the United States of America*, PNAS Early Edittion. www.pnas.org/cgi/doi/10.1073/pnas.1512221112 (accessed 30 June 2015).

Thompson EP (1963) *The Making of the English Working Class*. London: Victor Gollancz.

Tilly C (1975) *Coercion, Capital, and European States, AD 990–1992. 1975 Revised. Studies in Social Discontinuity*. Cambridge, MA: Blackwell.

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.