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

instead. Take a recent paper by the economists Stelios Michalopoulos and Elias Papaioannou. Michalopoulos and Papaioannou (2016) were interested in the effects of the "Scramble for Africa" among European powers in the late nineteenth and early twentieth centuries. They argue that European powers divided up Africa with little regard for the shape or politics of the societies on which they imposed colonial borders. The design of their paper is very simple. First, they used a georeferenced version of George Peter Murdock's (1959) Ethnolinguistic Map, which depicts the territory of African ethnic groups at the time of European colonization.[1] Then they projected onto the Murdock map a map of the national borders that divided up the continent. This allowed them to divide ethnic groups into those that were split by a national border and those that were not. Finally, they added data on the locations of civil conflicts that took place in the late twentieth and early twenty-first centuries. They find that ethnic groups whose historical territory was partitioned by a colonial border experienced more, longer, and more devastating political conflicts.

Another example is Neil Fligstein's (1981) *Going North*. Fligstein's book is a sociological account of the causes of the Great Migration. To assess the claim that the boll weevil infestation of 1892-1922 inspired African Americans to leave the South, Fligstein hand-coded a map of the boll weevil's migration published in a United States Department of Agriculture (USDA) report. Looking decade by decade, he finds that infested counties generally had higher rates of black outmigration. Using georeferenced versions of the USDA boll-weevil maps and data on historical county borders (Minnesota Population Center 2016), Deirdre Bloome, James Feigenbaum, and I estimate the infestation's effects on other dimensions of the South's economy and demography (Bloome, Feigenbaum, and Muller Forthcoming).

Both of these examples show how historical sociologists can use maps to turn data-sparse historical periods into data-rich ones. In both cases, the most important source of data was neither numbers nor text, but the simple intersection of two geographical boundaries. Using maps in this way is not new, as the publication date on Fligstein's book shows. But creating georeferenced historical maps and combining them with the growing library of maps produced by other researchers makes it a lot faster and easier than it used to be.[2]

Sociologists have also been linking data for a very long time. Finding the same person in two records often allows you to learn more about them than you could by consulting one record alone. In the past, researchers did much of this work by hand, moving, person by person, from one list to another. Simple string-matching algorithms let you do this kind of work much more quickly.

Despite their fancy name, string-matching algorithms are straightforward. At their most basic, they compare two strings of characters and penalize the substitutions, deletions, transpositions, and so forth needed to turn one string into the other.[3] Different algorithms give weight to different things, like where in the strings a discrepancy appears or whether certain letters sound like others.

Knowing just a little bit about string matching can help you even if you never plan to do any quantitative analysis. In my dissertation research, I was interested in knowing about the circumstances under which people confined in Georgia's convict lease system in 1880 were apprehended. Luckily, the Georgia Archives had a collection of all of the pardon requests sent to the governor from 1858 to 1942. The only problem was that the information about the convicts who had sent requests was limited to a webpage that listed them alphabetically, with no additional information except for their county of commitment. In the past, I would

have had to compare my list of prisoners to this list of over 10,000, one by one. Instead, I wrangled the data on the webpage into a machine-readable format, used the sdists() package in R to construct a list of possible matches, then wrote to the Archives asking them to pull about 200 boxes so that they would be ready when I arrived. What would have taken me weeks took only two days in the archive.[4]

Looking for a prisoner in two lists is much easier if both lists only contain prisoners. What if instead we wanted to match everyone in the 1920 U.S. Census to everyone in the 1940 U.S. Census? We are still working on solving that problem, but we are getting better at it. The economist James Feigenbaum has a new paper in which he describes some ways to use machine learning to link census records. Feigenbaum (2016) hired a research assistant to link a small sample of people across two censuses. Then he taught an algorithm to replicate how the assistant created the links. He found that most of the gains from using the algorithm came after the assistant had linked only about 500 records. We are still a long way from creating a census-based genealogy of the entire U.S. population, but we are starting to make some impressive advances in our ability to link large numbers of people across multiple sources.

These are just two examples of ways historical researchers are creating new sources of data about the past. But why would we want to create datasets like this in the first place? One answer is that data of the kind I have described can help us to get a better handle on foundational questions not only in historical sociology but in the discipline more broadly.

Data linked over long stretches of time, for instance, can reopen research on the topic of historical persistence (Abbott 2005; Patterson 2004). The economist Nathan Nunn (2008) has shown that the countries in Africa that had the greatest concentration of people who were forcibly removed and enslaved have the lowest average income today. Nunn's paper provides compelling quantitative evidence about a question historians have long debated, and it helped to spawn a whole literature on the effects of historical institutions on economic development today. But one notable feature of this literature is that it tends to use geographical territories as its units of analysis. With linked multigenerational data, historical sociologists could begin studying how historical events and institutions affect not just the aggregate statistics of geographical territories, but also lineages of people, some of whom moved away from the territories of their ancestors. This kind of work could create a bridge between historical sociologists and demographers studying multigenerational mobility.

What would have taken me weeks took only two days in the archive.

With tools like the ones I have described, we can also uncover missing data that could change our understanding of contemporary problems. It is easy to find examples of errors social scientists have made by studying too short a time period. The first graph in Thomas Piketty's (2014) *Capital in the Twenty-First Century*, for instance, shows how Simon Kuznets built a theory about the relationship between economic growth and inequality based on a narrow slice of a time series. This wasn't entirely his fault: he was dealing with the data he had.[5] But it offers a cautionary tale for all of us: it is very easy to build grand and sweeping theories based on anomalous periods of history. We need to have more humility about this. The future can always confound us, as when Blumstein and Cohen (1973) developed a theory about the stability of imprisonment in

the early 1970s, right before the U.S. imprisonment rate exploded. But, with the increasing availability of historical data, we have a much weaker excuse for letting the past do the same.

Endnotes

1. Maps that are "georeferenced" are encoded with coordinates allowing them to be combined with other maps. Michalopoulos and Papaioannou (2016: 1812) discuss how imperfections in the Murdock map could affect their analysis.

2. Harvard University's World Map is one example of such a library.

3. For an introduction to the stringdist() package for string matching in R, see van der Loo (2014). Another useful function for string matching in R is sdists().

4. Of course, during other phases of historical research this kind of efficiency is not desirable. See, for instance, chapter 6 of Abbott (2014).

5. Taking a longer historical view, Milanovic (2016) introduces the idea of "Kuznets waves."

References

Abbott, Andrew D. 2005. "The Historicality of Individuals." *Social Science History* 29:1-13.

Abbott, Andrew D. 2014. *Digital Paper: A Manual for Research and Writing with Library and Internet Materials.* Chicago: University of Chicago Press.

Bloome, Deirdre, James J. Feigenbaum, and Christopher Muller. Forthcoming. "Tenancy, Marriage, and the Boll Weevil Infestation, 1892-1930." *Demography.*

Blumstein, Alfred and Jacqueline Cohen. 1973. "A Theory of the Stability of Punishment." *Journal of Criminal Law and Criminology* 64:198-207.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking." Working Paper, Princeton University.

Fligstein, Neil. 1981. *Going North: Migration of Blacks and Whites from the South, 1900-1950.* New York: Academic Press.

Fogel, Robert William. 1982. "'Scientific' History and Traditional History." *Studies in Logic and the Foundations of Mathematics* 104:15-61.

Michalopoulos, Stelios and Elias Papaioannou. 2016. "The Long-Run Effects of the Scramble for Africa." *American Economic Review* 106:1802-1848.

Milanovic, Branko. 2016. *Global Inequality: A New Approach for the Age of Globalization.* Cambridge, MA: Harvard University Press.

Minnesota Population Center. 2016. *National Historical Geographic Information System: Version 11.0* [Database]. Minneapolis: University of Minnesota.

Murdock, George Peter. 1959. *Africa: Its Peoples and their Culture History.* New York: McGraw-Hill.

Nunn, Nathan. 2008. "The Long-Term Effects of Africa's Slave Trades." *Quarterly Journal of Economics* 123:139-176.

Patterson, Orlando. 2004. "Culture and Continuity: Causal Structures in Sociocultural Persistence." Pp. 71-109 in *Matters of Culture: Cultural Sociology in Practice*, edited by John Mohr and Roger Friedland. Cambridge: Cambridge University Press.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century.* Cambridge, MA: Harvard University Press.

van der Loo, Mark P. J. 2014. "The stringdist Package for Approximate String Matching." *The R Journal* 6:111-122.

Computational Methods, Meaning, and Comparative Historical Sociology

Laura K. Nelson
Northeastern University and University of California, Berkeley

Larry Irving—widely credited with coining the phrase the "digital divide"—recently commented on the relationship between sports analytics and Black fans: "Sports is emotional. And analytics represent the absence of emotion, the antithesis. Nobody gets into sports to be dispassionate. And it just seems to me we are the feel it, smell it, touch it people."[1] In discussions I have had with sociologists, particularly qualitative sociologists, this comment rings true. Instead of emotion, many sociologists care about meaning and interpretation, and believe analytics represent the absence of meaning. And, of course, nobody gets into sociology to be dispassionate.

Comparative historical sociology has historically tackled the big questions—democracy versus totalitarianism, the causes of revolution, the formation of the state, the causes of inequality. These topics elicit a great