"Big Data" in Economic History

Myron P. Gutmann, Emily Klancher Merchant, and Evan Roberts

Big data is an exciting prospect for the field of economic history, which has long depended on the acquisition, keying, and cleaning of scarce numerical information about the past. This article examines two areas in which economic historians are already using big data – population and environment – discussing ways in which increased frequency of observation, denser samples, and smaller geographic units allow us to analyze the past with greater precision and often to track individuals, places, and phenomena across time. We also explore promising new sources of big data: organically created economic data, high resolution images, and textual corpora.

Talk of "big data" has become nearly ubiquitous in popular and academic circles in the last decade. In economic history, an inherently empirical field of scholarship, the promise of big data is particularly tantalizing, suggesting an end to long hours spent acquiring, keying, and cleaning scarce and scattered numerical traces of the past. In this review article we examine what big data means to economic history, survey how economic historians are using big data now, assess available and forthcoming sources of big data, and discuss the possibilities for big data that could transform economic history in the next decade or two.

We focus on big numeric data in the study of populations, the environment, and prices, and include a brief discussion of the use of digital images and textual big data in economic history. At present, the most readily available sources of numeric big data describe historical populations and environments. Recent price and transaction figures also constitute big data, but may not be regarded as historical quite yet. In both the population and environment domains there are large, well organized quantities of analog

The Journal of Economic History, Vol. 78, No. 1 (March 2018). © The Economic History Association. All rights reserved. doi: 10.1017/S0022050718000177

Myron P. Gutmann is Professor, Department of History and Director, Institute of Behavioral Science, University of Colorado Boulder, 483 UCB, Boulder, CO 80309. E-mail: Myron.Gutmann@colorado.edu. Emily Klancher Merchant is Assistant Professor, Science and Technology Studies, University of California, Davis, One Shields Avenue, Davis, CA 95616. E-mail: ekmerchant@ucdavis.edu. Evan Roberts is Assistant Professor, Department of Sociology and Minnesota Population Center, University of Minnesota, 909 Social Sciences, 267 19th Ave S, Minneapolis, MN 55455. E-mail: eroberts@umn.edu.

We thank the editors of the Journal for the opportunity to undertake this review, and for their comments and feedback throughout the writing of the paper. Jeremy Mikecz provided assistance in the research on agriculture and environment data. Evan Roberts gratefully acknowledges support from the Minnesota Population Center (Project 5R24HD041023), funded through grants from the Eunice Kennedy Shriver National Institute for Child Health and Human Development. Myron Gutmann acknowledges support from the University of Colorado Population Center funded through a grant from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (Project 2P2CHD066613-06) for research and administrative support.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

data—text or images—that can be converted to a machine-readable format for statistical analysis at relatively modest cost, or that have already been converted by genealogists or environmental scientists. Similarly, there has been wide scholarly and commercial interest in converting textual corpora to electronic format, creating new possibilities for the analysis of language and its relationship to economic activity.

The current review is motivated by recent scholarly interest in changing volumes and forms of data collection and distribution. Specifically, the development of internet and digital imaging technologies has allowed personal, commercial, and scholarly datasets to grow in the various ways outlined later. Surveys of the growth and potential of big data have been published in cognate social sciences, including demography (Ruggles 2014), economics (Einav and Levin 2014; Varian 2014), epidemiology and health research (Bates, Saria, Ohno-Machado, et al. 2014; Khoury 2015; Wyber, Vaillancourt, Perry, et al. 2015), geography (Graham and Shelton 2013), political science (Monroe 2013), and sociology (Bearman 2015; Burrows and Savage 2014; Tinati, Halford, Carr, et al. 2014). General surveys of what big data means for social science attempt to synthesize the perspectives in field-specific surveys (King 2011; Shah, Cappella, and Neuman 2015). Our comparative advantage as economic historians is coming to the discussion at this particular juncture rather than earlier.

BIG DATA: DEFINITIONS AND FORMS

Economic historians accustomed to the problem of comparing prices and quantities over long periods of time will recognize that the "big" in "big data" needs a clear metric. A key characteristic of modern "big data" is that the volume of stored data exceeds human analytic capacity and pushes against the boundaries of currently-available computing power. For that reason, the magnitude of "big" is continually growing. The National Academies of Sciences, Engineering, and Medicine and the National Institutes of Health have described the recent growth of scientific data as a "deluge" (Anderson 1997; National Research Council 2013), though the quantity of printed data produced by governments in the nineteenth century has also been termed an "avalanche" compared with what was available previously (Hacking 1982). The problem of analyzing an ever-growing trove of data is not new in kind, though its scale has changed. To put the issue in perspective, in 1997 the National Academies published a report on *Massive Datasets* in which the social sciences were represented by a chapter on the computational difficulties of tabulating frequencies and running ordinary least square (OLS) regressions, with equipment available at the time, on a 5 percent sample of the U.S. census from 1990, a dataset of 12.5 million records (Anderson 1997). By contrast, in a recent article in this JOURNAL, Brian Beach, Joseph Ferrie, Martin Saavedra, et al. (2016) report analyses based on the linked records of more than 8 million individuals from complete databases of the 1900 and 1940 U.S. censuses, a considerably more computationally intensive task. It is now becoming common for scholars to work with datasets of 40 million or more records, whether from one source or pooled in a common format from many (Aaronson, Dehejia, Jordan, et al. 2017; Gutmann, Brown, Cunningham, et al. 2016). As with "top incomes," what constitutes "big data" varies over time.

Economic history has traditionally relied on the survival of original sources in manuscript form. Source survival has been both deliberate and accidental. The way in which

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

past actors and societies chose to preserve some records and destroy others in itself provides useful information about their values (Ashplant and Wilson 1988). Storing and preserving records is costly; large-scale record collections that have been preserved reflect what those in the past considered to have continuing value for their own activities or for posterity. These collections tended to be those of powerful individuals or institutions, such as states or churches, both of which have historically collected information about people and their property, though that information originally served the purposes of taxation, governance, and salvation, rather than scientific analysis (Hall, McCaa, and Thorvaldsen 2000). Economic historians in Britain, Canada, and the United States have made ample use of the population censuses in their research, which is highlighted in this article. Yet in Australia, Ireland, and New Zealand, which shared much in common with these countries (Lloyd, Metzer, and Sutch 2013), census authorities sometimes deliberately destroyed the manuscripts after the material had been tabulated. Even in countries where records were routinely preserved, some were lost to war, fire, or neglect after surviving for decades (Blake 1996; Dorman 2008).

Historically, the quantity of data available for scholarly research has increased as the costs of storing information have decreased. The falling cost of paper in the nineteenth century facilitated the storage of increasing numbers of records by governments and businesses, requiring in turn a large number of clerical workers to organize, analyze, and physically move the records. In the twentieth century, the cost of reproducing paper records declined, leading to a further increase in the volume of material that could be stored.

Contemporary interest in the scholarly potential of "big data" analysis derives from large reductions in the costs of producing, sharing, and storing electronic information since the late 1990s. As more data are "born digital"—that is, created and stored electronically from the outset—the cost of production, storage, and sharing are further reduced. However, the increasing availability of electronic data engenders some of the same core challenges faced by scholars and archivists in earlier eras when societies chose to store increasing quantities of analog information: organization and preservation in a format that allows useful analysis. Discussing the costs and benefits of processing into the archives 1,500 boxes of records from a former Illinois governor, one archivist lamented in the 1980s that "if it cost nothing to access and preserve records, we could save everything," but on further examination concluded that "much of the data resemble more the noise and distortions of a badly tuned television set than useful information" (Ham 1984). When information storage is costly, the task of distinguishing signal from noise falls to archivists and other data preservation experts. When storage is cheap and more information can be preserved, scholars become increasingly responsible for distinguishing signal from noise, opening opportunities for the discovery of unexpected signals in what might have otherwise been discarded as noise.

So far we have defined "big data" only in terms of logical size and only in relation to computing resources (Gandomi and Haider 2015; Ward and Barker 2013). Every author who surveys the landscape of big data agrees that *volume* of data is a critical part of the definition, but few have discussed the specific ways in which volume of data has increased. Overall, the volume of data available to scholars has increased because more people and other entities of interest are undertaking activities that generate stored data, meaning that more information is being collected for more people and about more activities. It is important to note, however, that some big datasets may be highly self-selected

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

when individuals themselves decide whether or not to participate in the data generating process, particularly in processes that generate born-digital data (Hargittai 2015), as will be discussed later at greater length.

Volume of data increases not only because more datasets are being created and stored, but also because those datasets include more observations and more characteristics of each observation. More observations can produce an effect akin to increased sample density, or can even produce datasets that contain the entire population of interest (Chetty, Stepner, Abraham, et al. 2016; Ruggles 2014). More observations can also mean more frequent data collection, which is often referred to as *velocity* in discussions of big data. The same tools that allow more people to participate in data generating process, and that allow for higher-density and more frequent samples, such as internet-based surveys and mobile devices that track activity, also make it possible to collect data on more characteristics of the activity or entity of interest, often referred to as *variety*. We can think of variety as the collection of more variables and velocity as the collection of more observations in the familiar format of a dataset of rows and columns of text and numbers. However, *variety* also refers to the increasing collection of different data formats, for example, video or images. Still images have become particularly important to the research process in many fields, including economic history, as they allow researchers to work with primary materials away from the data collection site, an aspect of the research process we discuss in the final section. High resolution images require a significant amount of disk space, thus returning us to *volume* as a defining characteristic of big data.

USES OF BIG DATA IN ECONOMIC HISTORY

Population and the environment are two areas where economic historians are already making use of big data. The collection and analysis of big data are not new activities. Governments have taken complete censuses since the eighteenth century, tabulating and printing the results for the entire population after each one. Researchers have extensively employed these aggregate data particularly for studies of the nineteenth century. In some countries the census tabulated quantities of interest at subnational levels, permitting analysis of geographically grouped data (Easterlin 1976; Fleisig 1976; Humphries 1987). Aggregate data on entire populations are typically structured as counts or averages over geographic areas such as counties or municipalities, so their size depends on the administrative geography of a given country. To put their size in perspective, they might be on the order of several hundred (e.g., 659 municipalities in Norway in 1910) to several thousand observations (e.g., 13,565 parishes in England and Wales in 1881). These data are big in the sense of describing large populations, though they reduce individual variation to a relatively small number of observations describing geographical aggregates or averages.

Economic historians have made ample use of individual-level data from high-density samples of complete population data, such as the IPUMS 1 percent and 5 percent samples of the United States census (Ruggles, Genadek, Goeken, et al. 2015).[1] Such samples have allowed for more precise measures of human behavior, for example, with

[1] Integrated Public Use Microdata Series [hereafter IPUMS].

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

individual family size replacing aggregate child-woman ratios as a measure of fertility. Since the microdata revolution of the 1970s, more individual-level samples of censuses, social surveys, and civil registration records have become available for many countries, with increasing sample density. The computational resources for analyzing them have also improved. These datasets were large for the time in which they were created; as noted earlier, just 20 years ago, analyzing a 5 percent sample of one census with 12.5 million records stretched the computing capacity of the modal scholar.

Population aggregates and samples of individual-level census data have provided vital sources in economic history, for example, in studies of slavery, Reconstruction, and civil rights in the United States. In the 1950s, Alfred Conrad and John Meyer (1958) used census aggregates for their research; in the 1960s, scholars developed samples of census data to study slavery (Fogel and Engerman 1974) and Reconstruction (Goldin 1977). Since the 1990s, the IPUMS has been a critical data source to explain the slow convergence of African American and white economic status (Boustan 2009; González, Marshall, and Naidu 2017; Johnson 2004; Sundstrom 2007).

Today, a prominent form of big population data is the complete universe of individual-level census records. Between 2003 and 2013 the North Atlantic Population Project (NAPP) made available full-count microdata for censuses of Canada (1881); Denmark (1787, 1801); Great Britain (1881, 1911); Norway (1801, 1865, 1900, 1910); Sweden (1880, 1890, 1900); the United States (1880); and Iceland (1703, 1729, 1801, 1901, 1910) (Minnesota Population Center 2015a; Ruggles, Roberts, Sarkar, et al. 2011), totaling just over 100 million individual records. Since 2013 the availability of complete count censuses has expanded dramatically, including every United States census from 1790 to 1940 (except 1890), every British census from 1851 to 1911, and the 1901 and 1911 Irish censuses. Just under one billion historical census records from North America and Europe are now available for research through IPUMS and the NAPP.[2] By 2020 all records will have standardized codes for economically relevant variables such as occupation and industry, nativity, educational attainment, and place of residence. In addition to these historical records, which come with reduced privacy restrictions allowing researchers to see the names of individuals, the complete short- and long-form versions of the United States census since 1960 are now available to researchers in the Census Bureau's Federal Statistical Research Data Centers.[3] As sample density reaches its logical limit of complete individual records, the quantity of census data an economic historian can analyze currently pushes the limits of computational power for more complex procedures such as record linkage or duration modeling (Ruggles 2014).

Given that analysis of samples of the census have been so fruitful, what are the scholarly benefits to working with bigger data and specifically complete-count census datasets? Two important advantages of big data for economic historians are *precision* and the potential to *link* observations.

At the limit with population data, we can ignore sampling error, which can make it impossible to say anything definitive about less-common phenomena when working with smaller samples, and focus on the substantive interpretation of statistical results.

[2] These resources are located at ipums.org and nappdata.org.

[3] Similar expansions of access to complete post-1970 census records have occurred in many countries.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Complete population data make it possible to analyze small subgroups either by themselves or in comparison to other populations. For example, while the 1.9 million Irish and 1.9 million German-born migrants in the United States in 1880 can be adequately studied with a 1 percent sample, important but smaller groups, such as the 125,000 French-born migrants in the same year, are harder to study using even a 5 percent sample. Marginal totals and main effects can be derived, but interactions would evaporate in a "welter of empty cells" (Ruggles and Menard 1995). Recent scholarship taking advantage of this increased precision has examined assortative mating among immigrants from specific countries in U.S. cities (Logan and Shin 2012) and interracial marriage in the United States (Gullickson 2006).

Additionally, complete population data often identify people's location quite precisely, allowing researchers to examine social and economic processes at small geographic scales. Geographic precision in census data derives from the original and fundamental purpose of censuses: to provide data to governments for administration and apportionment of populations to political boundaries. Scholarship that utilizes geographical precision includes work on residential segregation in the United States (Logan and Zhang 2012; Logan and Parman 2017) and the influence of kin in neighboring households on fertility (Hacker and Roberts 2017).

The availability of full-count census data increases the prospects of linking individuals across time to better address questions about behavior and causality. Linkage is possible between complete count censuses (Beach, Ferrie, Saavedra, et al. 2016), between a complete count and a sample (Long and Ferrie 2013), and between a complete census and some other source (Bleakley and Ferrie 2016; Roberts and Warren 2017). Linking individuals across time to examine inter- and intra-generational economic mobility, for example, allows scholars to revisit classic questions in American economic history about the degree to which new institutions and environment in the new world allowed individuals to fare better than their parents had done in terms of economic status. The hypothesis that the United States (and similar settler societies such as Canada, Australia, and New Zealand) had a high degree of intergenerational mobility dates to contemporary arguments in the nineteenth century (Turner 1893), and generated research in the 1960s and 1970s by economic and social historians who developed longitudinal data on individual economic progress in specific cities and towns (Katz 1975; Knights 1971; Pearson 1980; Thernstrom 1973, 1964). In the 1990s, Ferrie created more representative samples using indices (Jackson 1992) to the American census to search over a wider geographic space (Ferrie 1994, 1999) for migrants who had moved out of their community of origin. Subsequently, Jason Long and Ferrie (2013) have shown using parallel samples of American and British men that the level of intergenerational mobility among these men in the United States was high in the nineteenth century, a contrast to the more pessimistic conclusions of earlier scholars. However, in the twentieth century the relationship reversed, with more mobility in Britain. Thus, big data have helped refine our understanding of the changing relationship between place, migration, and occupational mobility.

Ideally, record linkage would be based on complete population listings from each time point. As an intermediate step to this goal, the IPUMS project developed a set of Linked Representative Samples, in which individuals in the 1 percent samples of censuses for 1850–1870 and 1900–1930 are linked to the 1880 full-count census.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Linked data are particularly useful for the study of geographic mobility (Ferrie 2005), occupational mobility (Abramitzky, Boustan, and Eriksson 2014; Long and Ferrie 2013, 2007), socioeconomic mobility (Long 2005), and even mobility between census racial categories (Saperstein and Gullickson 2013), a phenomenon that would be difficult to track with even a relatively high-density sample of the population. Studies of mobility are not limited to the United States. Taking advantage of the availability of full-count census data for Norway in 1865 and 1900 and a dataset for the entire Norwegian-born population of the United States in 1900, Ran Abramitzky, Leah Boustan, and Katherine Eriksson examine the relationship between inheritance and the decision to migrate from Norway to the United States (2013), and estimate the economic returns on migration for those who made that decision (2012).

Censuses, by definition, record information for full universes of population, usually at the national level, but they are not the only source of universal data. Administrative records can also produce data about complete populations, though usually at a smaller scale. Administrative data often represent continuous rather than periodic collection, allowing scholars to answer questions about demographic processes at a very fine level of detail. For example, using German social security system data for 61.4 million individuals to calculate the daily number of births between 1920 and 1989, Thomas Bauer, Stefan Bender, Jörg Heining, et al. (2013) examine the relationship between births, lunar cycles, and sunspots, finding that the lunar cycle does not affect the number of births but that births and the number of sunspots are positively correlated. Oriana Bandiera, Imran Rasul, and Martina Viarengo (2013) use administrative data on the full universe of 24 million immigrants who entered the United States through Ellis Island between 1892 and 1924 to re-estimate migration flows in and out of the United States. They find that immigration was considerably more prevalent between 1900 and 1920 than the official statistics suggest. By comparing these administrative records to the number of immigrants found in the censuses of 1900–1920 and accounting for expected mortality, they estimate that out-migration during this period was more than twice as frequent as recorded in official estimates.

Population registers are another valuable and heavily used source of universal administrative data. They were often a product of church and state co-operation, collected to monitor, more or less continuously, the residential location of populations (Bengtsson, Campbell, and Lee 2004). Population registers provide today's scholars with a continuous record of the stock and flow of administratively defined historical populations, facilitating the exploration of questions that require large numbers of records or intergenerational linkages. For example, the China Multi-Generational Panel Dataset, publicly available at the Inter-University Consortium for Political and Social Research (ICPSR), includes the records of 370,000 people between 1749 and 1913 (Lee and Campbell 2016; Lee, Chen, Campbell, et al. 2017).[4] Sweden is, perhaps, the country with the longest run of population register data. Swedish data have been used to examine the relationship between childbearing and longevity (Barclay, Keenan, Grundy, et al. 2016), the effect of birth order on mortality (Barclay and Kolk 2015),

[4] China Multi-Generational Panel Datasets Series. Lee and Campbell (2016), available at www.icpsr.umich.edu/icpsrweb/DSDR/series/00265.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

and the effects on fertility of size of family of origin (Kolk 2014). Digitized population registers have allowed for similar studies in Norway (Grundy and Kravdal 2014, 2010). As these registers continue in computational forms and are automatically updated, they become truly big data (Dribe and Helgertz 2016).

In North America there is no history of continuous population registration.[5] Genealogical projects have produced multigenerational datasets for the populations of Quebec in Canada (Dillon, Amorevieta-Gentil, Caron, et al. 2017) and Utah in the United States. The BALSAC database, managed by the Université Laval, McGill University, and Université de Montréal, was created by linking marriage, birth, and death certificates in Quebec from the seventeenth century to the present (Université du Québec à Chicoutimi 2017). It currently includes five million individuals over four centuries. The Utah Population Database (UPDB) includes over 7.7 million descendants of those who experienced a vital event on the Mormon Trail and is linked to a host of other medical and administrative records (Smith and Huntsman Cancer Institute 2017). It is currently maintained by the University of Utah and is updated annually. In common with population registers, big genealogical databases such as BALSAC and the UPDB are particularly suitable for answering historical questions about intergenerational social mobility or transmission of fertility behavior (Gagnon, Tremblay, Vézina, et al. 2011; Jennings, Sullivan, and Hacker 2012; Maloney, Hanson, and Smith 2014). These genealogies also provide an essential source of data for the study of inheritance of health and disease (Broeckel, Hengstenberg, Mayer, et al. 2007; Kerber, O'Brien, Smith, et al. 2001). Martha Bailey's LIFE-M project is now developing geographically broader population linkages of vital records and census data for the late nineteenth and twentieth century United States (Bailey 2017).

Certainly, economic history has long featured analysis of individual-level data and longitudinal data. The practice of creating a dataset with observations from different time points, or from a variety of data sources, has merely changed in degree, rather than in kind. The Cambridge *Population History of England*, for example, created linked datasets from parish records, working in conjunction with volunteers in parishes around England to abstract the data from the original manuscripts (Wrigley and Schofield 1981). The continued use of the resulting data attests to their value (Boberg-Fazlic, Sharp, and Weisdorf 2011).

In addition to linking individuals across time, economic historians have created big population datasets by pooling data from multiple censuses, genealogies, or population registers. For censuses, this process was facilitated by the NAPP, described earlier, and the IPUMS-International project, which provides harmonized individual-level census data since 1960 for 82 countries, including a total of 614 million person records. Pooled microdata permit analysis of both within-country and between-country variation, as well as cross-sectional change over time. Recent studies using such data have examined the relationship between socioeconomic status and fertility at the turn of the twentieth century (Dribe, Hacker, and Scalone 2014) and international variation and

[5] The absence of population registration in Canada and the United States reflects a shared British influence on statistical structures. Similarly Britain, Ireland, South Africa, Australia, and New Zealand have no state or church population registration.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

changes over time in the living arrangements of the elderly (Ruggles 2009). Combined data from genealogical databases have been used to examine the relationship between fertility, aging, and mortality across populations (Eijkemans, Van Poppel, Habbema, et al. 2014; Gagnon, Smith, Tremblay, et al. 2009). A landmark project using population register data from Belgium, Sweden, China, Japan, and Italy—the Eurasia Project—has explored several demographic processes, including mortality, fertility, and nuptiality, teasing apart biological universals and cultural differences (Bengtsson, Campbell, and Lee 2004; Lundh and Kurosu 2014; Tsuya, Wang, Alter, et al. 2010).

"Big population data" most often refers to large collections of individual microdata, but aggregate-level population data can be big as well if they cover a long period of time and aggregate over many relatively small spatial units. Ian Gregory, Jordi Marti-Henneberg, and Francisco J. Tapiador (2010) have created a database that points in this direction, mapping regional-level population data for European censuses from 1870 to 2000, facilitating the long-term analysis of population change throughout Europe at the sub-national level. This geographic information systems (GIS) database includes decadal population data from 1870 to the present, interpolated to the boundaries of 562 intermediate-level administrative units currently used by the statistical office of the European Community. While this database does not qualify as "big" in terms of number of observations or computational requirements, it suggests that it is possible to go one step further and to divide aggregate population data into small gridded units, using sophisticated interpolation techniques. One example is the Fourth Version of the Gridded Population of the World, which distributes population over a 30 arc-second grid, representing approximately 1 km square at the equator (http://sedac.ciesin.columbia.edu/data/collection/gpw-v4). These data describe population at five-year intervals from 2000 to 2020; the high resolution of the grid dramatically increases the granularity of spatially-oriented population data of use to researchers. The potential for linkage of small-scale population data to other factors, especially the environment, has also been brought to the forefront by the Terra Populus project which, when completed, will make available yet more high resolution historical population data that are and can be linked to environmental and other data (Minnesota Population Center 2015b).[6]

Several recent articles in this JOURNAL demonstrate how big population data are changing the practice of economic history, and will continue to do so. William Collins and Marianne Wanamaker (2015) address a perennial question in American economic history: How the Great Migration improved economic outcomes for African Americans. They begin with a 1 percent sample of the 1910 United States census, from which they link forward 26,829 out of an initial sample of 111,524 southern males aged 0–40. Although a recently published article, some of Collins and Wanamaker's research methods have already aged slightly, showing how quickly big data are changing economic history. Collins and Wanamaker searched digital genealogical indices of the 1930 census to link with the 1910 sample; today linkage between publicly available files of the complete 1910 and 1930 censuses could create a much larger sample. They find a low degree of selection, suggesting wide participation in migration out

[6] Minnesota Population Center, Terra Populus Project. 2016. terrapop.org.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

of the South. However, they also found that white and black movers of comparable skill level and background made significantly different destination choices. As Collins and Wanamaker note in conclusion, longitudinal data are fundamental to understanding migration decisions. Cross-sectional data collected on migrants at their destination cannot show migrants' status and conditions before they departed, which are necessary to understand why people chose to migrate. This article is representative of the frontier of economic history research on late nineteenth and early twentieth century migration, with scholars pursuing similar research strategies in studies of British (Long 2005) and Norwegian (Abramitzky, Boustan, and Eriksson 2012) migrants, both domestic and international.

While Collins and Wanamaker show how big data can be used to *link* people across time, another recent article in the JOURNAL demonstrates how big data can be used to measure concepts with greater *precision*. Trevon Logan and John Parman (2017) return to another important topic in American economic history: residential segregation of black and white households. Logan and Parman use the complete individual-level returns of the 1880 and 1940 censuses to construct a new measure of segregation that is based on the race of a household's next-door neighbors. Previous studies of historical segregation had used racial composition at no smaller than the ward level, because wards were the smallest unit for which population by race was consistently reported (Cutler, Glaeser, and Vigdor 1999). Logan and Parman's measure shows that segregation increased substantially between 1880 and 1940. The chance of a black household having a white neighbor declined 25 percent in 60 years, with similar changes across the entire United States. While previous studies with aggregate data (Cutler, Glaeser, and Vigdor 1999) had documented the overall rise in segregation, Logan and Parman are able to show precisely where segregation was more prevalent, and where it grew more rapidly. As Logan and Parman acknowledge, their measure of segregation is simple, relying only on the immediate two neighbors of a household. More complex measures of segregation could use a wider window of households/neighbors or include measures that incorporate indicators of socioeconomic status. As more complete count datasets from population censuses become available, we expect economic historians will become increasingly creative in their construction of new measures of social and economic behavior.

Increased granularity in population data facilitates linkages to environmental data, which have also become available at finer levels of geographic and temporal detail. Data about the environment are well poised to become "big data" because of their potential for large-scale coverage, high frequency through time, and high levels of geographic granularity. Put another way, data about the environment, recorded frequently and divided into relatively small spatial units, are increasingly available for large swaths of territory. Good examples are records of weather through time and across space, which have been recorded systematically for many parts of the world since the nineteenth century, and land cover, which has been systematically documented in the United States by aerial photographs since the 1930s and globally by satellites since the 1970s. Agricultural land use data, which overlap conceptually with land cover change data, also fall into this category, with data collected at the county level for the United States since the mid-nineteenth century and similar data collected elsewhere over various periods of time.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Other large collections of environmental data document the distribution of soils and the elevation and slope of terrain, all useful for understanding the environmental context in which economic activity takes place.

As is the case with other types of data discussed here, most historical environment and agriculture data were originally collected and available in analog (paper or film) formats, with an increasing fraction of those converted to digital. An early example is the Parker-Gallman study, based on a sample of data digitized from manuscript records of the 1860 U.S. Census schedules for population, slaves, and agriculture for 405 counties in what would become the Confederate states (Parker and Gallman 1991; Parker 1970). At state and county levels, Michael Haines (2010) has digitized published volumes of the U.S. Census of Agriculture as well as the Census of Population, providing a valuable data source about land cover, crop yields, and livestock rearing. These datasets, though they continue to be well-utilized (Olmstead and Rhode 2015), are not what we would term "big." William Parker and Robert Gallman digitized only a sample of the 1860 Census manuscripts for the counties in question, which was no trivial task given the technology of the period. While the Haines dataset includes all U.S. counties in all censuses, population censuses have been taken only at 10-year intervals (the agricultural censuses since 1920 have been more frequent, generally at five-year intervals), and the United States today contains only 3,144 counties and county equivalents. Neither dataset included all variables in the analog sources from which they were digitized. Historical weather data have also traditionally been digitized from the paper sources in which they were originally published, but here too both print publication of the original sources and the process of digitization have imposed limits on the volume of data that could be collected and processed.

Increasingly, environmental and agricultural data are becoming available at finer levels of granularity with respect to both space and time. The introduction of digital methods of data collection and preservation have facilitated the availability of more detailed environmental and agricultural data, resulting in data that are "born digital" and are limited neither by the constraints of publication nor by those of digitization. The U.S. agricultural censuses are an excellent example. Like other census-type data, they were published as county-level aggregates in books through the 1970s, succeeded by digital versions on CD, and have been available for internet download since the 1980s. Similarly, weather data for the United States are now available in born-digital form, in some cases with frequencies as great as hourly for individual stations (National Centers for Environmental Information 2016).

By definition, virtually all environmental data are spatial; each data cell in a table represents the attributes of some piece of the earth, water, or atmosphere, located in three-dimensional space. Because of those characteristics, and with the assistance of modern GIS technologies, it is possible to manipulate and eventually to subdivide the data. Data that begin, for example, as the attributes of weather at a given moment at a given group of weather stations can be interpolated into the attributes of weather for grid cells of almost any size, often as small as a single square kilometer or a single degree (or less) of latitude and longitude on a side. Those millions or billions of data cells representing, for example, the temperature or precipitation in a one-kilometer grid cell for every minute or hour of an extended period of time, constitute genuinely "big data" (Daly, Gibson, Taylor, et al. 2002; Daly, Halbleib, Smith, et al. 2008). They can be used

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

as they are or re-aggregated in other ways to capture the weather characteristics of a city, county, or some other spatial unit. These processes are analogous to the creation of the gridded population data described earlier.

Spatially defined data also make it possible to integrate various types of information into a single framework, based either on common geospatial definitions (such as counties in the United States) or on a gridded approach. Valuable integrated data sources have recently become available, including the Great Plains Population and Environment Project's data at ICPSR (Gutmann 2005, 2007; Parton, Gutmann, Hartman, et al. 2012), and, even more ambitiously, the global integration of population and environment data in the Terra Populus project at the University of Minnesota (Minnesota Population Center 2015b).

Historians have made a good start on using these sources, as researchers learn about the availability of data and their possibilities. One area where substantial progress has already been made is in the study of land cover change in the context of social, economic, and policy change. Work done by Ken Sylvester and colleagues (Maxwell and Sylvester 2012; Sylvester, Brown, Deane, et al. 2013; Sylvester, Gutmann, and Brown 2016; Sylvester and Rupley 2012) on transitions in land cover in the Great Plains using digitized aerial photographs and satellite remote sensing data is especially notable in this regard. It uses large datasets derived at the pixel or small-scale grid level and demonstrates how the impact of economic and policy change can be directly viewed on the ground. Other notable efforts to understand large-scale land use and land cover change in historical perspective include work by Kees Klein Goldewijk (2001); Jed Kaplan, Kristen Krumhardt, and Niklaus Zimmerman (2009); and Mingliang Liu and Hanqin Tian (2010). These environmental datasets are likely to become important resources for economic historians studying agricultural productivity under changing environmental conditions (Olmstead and Rhode 2011).

Weather and climate data also have the potential to play an important role in economic history research. Analysis of high-resolution weather and climate data is already underway in understanding the history of natural disasters and their relationship with demographic and economic outcomes. Studies of drought and its impact, for example, increasingly make use of large-scale data sources. These include work by Glenn Deane and Myron Gutmann (2003) on the drought and dust storms of the United States in the 1930s, Qiang Chen (2015) and Ruixue Jia (2014) on China, Klas Rönnbäck (2014) on Africa, and Greg Bankoff (2007) on more general experiences. Celine Herweijer, Richard Seager, Edward Cook, et al. (2007) show the potential for extremely long-term studies by examining a millennium of droughts in North America.

High-resolution data also have a strong potential to make a major contribution to economic history research through the detailed simulation of conditions at the intersection of climate, agriculture, and population. Biogeochemical modeling to estimate the historical production of greenhouse gases from agriculture is one way that these approaches have been used, producing high resolution data for future analysis (Parton, Gutmann, Hartman, et al. 2012) and promising valuable results (Hartman, Merchant, Parton, et al. 2011; Parton, Gutmann, Hartman, et al. 2013; Parton, Gutmann, Merchant, et al. 2015). Historical agent-based modeling can also take advantage of high-resolution data to examine the ways people have made economic decisions in particular environmental conditions (Sylvester et al. 2015).

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

FUTURE OF BIG DATA IN ECONOMIC HISTORY

The datasets available to economic historians studying the 1980s and beyond are increasingly large. While the data creation process varies across domains, an important factor in the production of larger datasets is that the marginal cost of creating and storing a single data point has declined. Researchers continue to make their own datasets from analog sources. Such collections have also grown larger as costs have fallen. Increasingly, economic historians are working with three new types of data: numeric data that have been created and stored organically (Groves 2011) in the course of routine economic actions, whether private market transactions or interactions between people and government social and fiscal programs; high-resolution digital images; and digitized texts.

Organically Created Economic Data

Many forms of records familiar to economic historians are now created electronically and stored at relatively low cost, compared to paper records. Some is personal—such as email messages, social media entries, images, logs of physical activity—and some proprietary—business emails, transaction records, logs of computer or machine actions (Kay and Harmelen 2012).

Another category of "big data" organically created comes from people's use of the internet to communicate, manage, and transact. These data are created by internet users without conscious intent to create an archive. Economists have begun using these forms of data to study familiar topics: Dolan Antenucci et al. (2014) used Twitter posts to create a leading indicator of unemployment. Contemporary social scientists are using the social networking site, Facebook, to measure human behavior (Kramer, Guillory, and Hancock 2014). With a significant amount of behavior being recorded on sites such as these, it will soon be possible to examine changing behavior over time: the task of economic historians.

Much of the data traditionally used by economic historians does not come up to the standard of "big data" (even several centuries of monthly data on the prices of several dozen commodities in multiple places do not make large datasets), but researchers who study the period from the 1980s onwards *will* encounter truly large datasets. Whereas historical data in these areas are often summaries of market prices, scholars of current prices, wages, and financial markets work with large volumes of transaction-level data.

The introduction of electronic scanners to conduct transactions in retail stores in the 1970s and 1980s (Levin, Levin, and Meisel 1992) has led to more detailed price indices (Hausman 2003; Melser 2006; Silver and Heravi 2001), with high frequency price data and a finer classification of the commodity being transacted (Ivancic, Diewert, and Fox 2011). At the household level, a panel of more than 10,000 Japanese households has been tracked by a market research firm (Kohara and Kamiya 2016), while more than 25,000 British households participate in a panel that records between 600,000 and 1 million weekly transactions (Griffith and O'Connell 2009; Leicester and Oldfield 2009; Lusk and Brooks 2011). To give a sense of the size of the datasets created, a recent working paper by Greg Kaplan and Sam Schulhofer-Wohl (2016) analyzed dispersion

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

in household-level inflation rates using data on 500 million transactions from 50,000 households over a decade.

The migration of commerce to the internet, where prices are posted in both public and machine-readable form, has brought additional research opportunities, along with challenges. The Billion Prices Project at MIT has been "scraping" data on prices since 2008 to create alternatives to official price indices (Cavallo and Rigobon 2016). Of course both scanner data and online prices have important selection issues. Households that participate in consumer panels differ from a random sample of the population. Moreover, goods and services are not equally likely to have bar-codes and be recorded easily, or to be sold online where their prices can be scraped. This issue is not new. Researchers have long recognized that households participating in "family budget" or consumer expenditure surveys do not resemble the population as a whole (Index Committee 1948).

In the past two decades, the volume of transactions on world financial markets has grown dramatically (Miller and Shorter 2016). In a process similar to that which has taken place in other fields, the operations of financial markets are increasingly undertaken by computers with little direct human input, and have sharply increased in volume. Data on the characteristics of each trade are archived as part of the transaction process. The size of these datasets can be substantial. For example, Cheng Gao and Bruce Mizrach analyze 30 million stock quotes over a 20-year period from 1993 to 2013 (Gao and Mizrach 2016), an amount that would be dwarfed by the data generated today. Central banks and financial regulators, who relied on summary statements of financial positions in the past, are now beginning to analyze the characteristics of individual financial transactions and loans, requesting more granular data from market participants (Bholat 2015; Fitzgerald 2016).

Individual-level data about income reported and taxes paid constitute another increasingly detailed and voluminous data source. While available to researchers since the 1960s after approval by the appropriate government agency (Clotfelter 1983), their scale has increased significantly in recent decades, reaching millions of records. For example, Naomi Feldman, Peter Katuščák, and Laura Kawano (2016) are able to focus on the narrow universe of taxpayers with a child turning 17 to analyze how households respond to losing a child-tax credit. Raj Chetty, Nathaniel Hendren, Patrick Kline, et al. (2014) study inter-generational mobility for more than 50 million U.S. children born between 1980 and 1993. Studies using the entire universe of tax returns are not limited to the United States (Atkinson, Piketty, and Saez 2011; Claus, Creedy, and Teng 2012). Because taxation data often include some demographic information, several countries have developed linked employee-employer datasets that combine individual earnings data from personal tax returns, information on the employer, and in some countries demographic information from census records or national health care systems (Abowd, Stephens, Vilhuber, et al. 2009; Bagger and Seltzer 2014; Lazear and Shaw 2009).[7]

[7] Matched employee-employer datasets are available in countries including Denmark, Sweden, Norway, Finland, Germany, the Netherlands, Belgium, France, Austria, Italy, Portugal, Slovenia, Slovakia, Japan, New Zealand, Brazil, Colombia, and the United States. See the references for further information. Data typically extends from the mid-1990s forward.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Government records describing health, education, and labor market activities are increasingly accessible electronically. Since the 1950 census round, population censuses in the United States have been processed electronically, and internationally censuses have been processed electronically from the 1960s onward. Where censuses have survived, they are increasingly available for research (Hall, McCaa, and Thorvaldsen 2000). Records from national health care systems (Lynge, Sandegaard, and Rebolj 2011), taxation records (Chetty, Stepner, Abraham, et al. 2016; Feldman, Katuščák, and Kawano 2016), population registers (Devereux, Black, and Salvanes 2007), and birth and death registers (Ferrie and Rolf 2011) are being used by economists and other social scientists in many countries. Access to data describing living populations is often restricted for privacy reasons, and ease of access varies across countries. Generally, access conditions are relaxed for cohorts that are deceased. Although much of the research with these datasets does not yet frame questions historically (Black, Devereux, and Salvanes 2013), as the chronological span of the data increases, it will be feasible to ask how economic behavior changed over time.

High-Resolution Images

Sensing, monitoring, and recording devices used in many areas of science and social science can now capture a large volume of images or other data types with little intervention. An exemplary example of the transformation comes in wildlife ecology, the social science of animal behavior. As recently as the late 1990s ecologists were limited to film cameras tripped by motion or pressure sensors, or capturing images at set intervals. Standard film cameras were constrained, physically, to capturing fewer than 40 images before needing to be re-loaded. Higher capacity was only available at significant cost, or by compromising image quality. Thus, camera trap research was limited to researchers who could afford the necessary equipment, and often conducted on datasets of several hundred images that could be analyzed by the lead researchers. In the past 15 years, significant advances have been made in digital camera technology, so that a standard digital camera today can capture and store nearly 100,000 high-resolution images. Accordingly, there has been a dramatic increase in the use of camera trap technology, bringing the new challenge of cleaning, classifying, and analyzing datasets that have increased in size by several orders of magnitude (O'Connell, Nichols, and Karanth 2011; Swanson, Kosmala, Lintott, et al. 2015). Comparable increases in the size of raw image datasets have occurred in other environmental sciences (Porter, Hanson, and Lin 2011), biology and medicine (Candido dos Reis, Lynn, Ali, et al. 2015), and astronomy (Willett, Lintott, Bamford, et al. 2013).

The social sciences have made less use of high-resolution images as a form of data collection, but recent work suggests they are beginning to do so. For example, Edward L. Glaeser, Scott Duke Kominers, Michael Luca, et al. (2018) demonstrate how social scientists can use similarly large, and growing, collections of street-level images of the urban environment to study economic behavior. From a different perspective, satellite data have become widely used by economists to study the extent and form of economic activity (Henderson, Storeygard, and Weil 2012), even if databases of images and other outputs from monitoring devices are not yet a common form of data in the field.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Automatic monitoring technologies have the potential to reduce a variety of selection and compliance issues in data generation. Instead of burdening respondents to complete a time diary, for example, their activities can just be recorded. Thus, social scientists in a wide range of areas will increasingly produce data from these technologies. As these datasets develop a time dimension, they will eventually become big data for economic historians of the twenty-first century.

Textual Datasets

In the last decade or so, economic historians have begun to analyze a new type of big data: textual corpora. These datasets, usually unstructured collections of words and documents rather than structured rows and columns of numbers, invite analysts to combine more familiar econometric methods with tools borrowed from the new field of digital humanities and computational linguistics. The traditional sources used by economic historians well describe populations, prices and wages, and agricultural inputs and outputs, and provide estimates of how economic behavior has responded to measurable changes in circumstances. Yet even a complete universe of these traditional quantitative sources provides relatively little insight into tastes, values, and motivations. Textual data in various forms can provide insight into what past economic actors thought about the decisions they were making. Textual corpora provide economic historians with a new quantitative approach to questions sometimes addressed in a more narrative style.

For example, there has long been an interest in whether differences in religious and political affiliation affect institutions, human capital accumulation, and economic performance (Becker and Woessmann 2009; Cantoni 2015). Jeremiah Dittmar and Skipper Seabold (2015) use statistical models for high-dimensional data to identify characteristically Protestant and Catholic language in the titles of books published in German between 1454 and 1600 to examine the relationship between media competition, religious content, and institutional change during the Protestant Reformation. Using much more recent textual data, Matthew Gentzkow and Jesse Shapiro (2010) identify sets of phrases from the 2005 *Congressional Record* used more frequently by one party than the other, and compare the occurrence of these phrases in 2005 U.S. newspapers to identify each paper's "slant" between left and right. They then compare newspapers' ideological orientation to that of their potential markets and examine to what extent the "slant" of each paper has been calculated to maximize profit by mirroring reader ideology. Jacob Jensen, Suresh Naidu, Ethan Kaplan, et al. (2012) apply similar methods to studying more than a century of Congressional debate to identify trends in partisan polarization.

These studies classify texts according to a predetermined set of ideological language. Another approach to textual data is the classification of texts through unsupervised machine learning algorithms. David Newman and Sharon Block (2006) apply these methods to the *Pennsylvania Gazette* from 1728 to 1800 (a corpus of 80,000 articles and advertisements) to examine what topics the paper covered and how they changed over time. Fitting a probabilistic latent semantic analysis model with 40 topics, they find that the largest topics related to economics and politics, with the most prevalent reflecting ads for escaped slaves and indentured servants. They demonstrate time trends

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

in the topics, showing a dramatic increase in discussion of government from the 1760s to the 1790s. The prevalence in discussions of cloth, for example, tracks a rise and fall in imports over the period. These unsupervised models offer new approaches to the analysis of large quantities of text.

Scholars have also begun to use computational textual analysis to generate structured databases out of unstructured texts. In analysis of newspaper accounts of lynching in the United States between 1875 and 1930, Roberto Franzosi, Gianluca De Fazio, and Stefania Vicari (2012) classify parts of speech to identify the way the media attribute agency in cases of violence. They use quantitative narrative analysis to identify semantic triplets (subject, verb, object) that include one person or corporate actor exerting violence on another, creating a database of lynching that can be analyzed in terms of directed networks or the spatial distribution or chronology of lynching.

The Trading Consequences project, a joint effort of Canadian environmental historians and computational linguists and computer scientists in the U.K., has mined massive volumes of nineteenth-century papers and trading records to create a database of commodities in geographical space and across time, including nearly 2,000 commodities that were regularly traded in the nineteenth century.[8] Efforts to turn unstructured texts into structured databases produce new forms of "big data" that are tractable to a variety of quantitative and spatial methods.

Economists have turned to the analysis of textual data to chart the history of and trends in their own discipline. Efforts to identify trends in the field over time are not new, but the ability to analyze large textual corpora allows for the analysis of trends at the level of the journal article, rather than the title or abstract. Lea-Rachel Kosnik (2015) uses computational linguistics to identify time trends in 20,321 articles in seven top economics research journals from 1960 to 2010. She identifies a set of keywords and key-phrases representative of well-defined economic fields, and uses their frequencies in the corpus to track the prevalence of the subfields over time. Zubin Jelveh, Bruce Kogut, and Suresh Naidu (2015) use natural language processing to predict the individual political behavior of economists on the basis of their scholarly writings from 1973 to 2011. These studies have found that attention to the various subfields of economics has remained relatively constant over time, with the exception of macroeconomics, which has decreased in prevalence, and that the political ideology of economists influences both the topics and results of their research. More recently, a working paper by Lino Wehrheim (2017) applies topic modeling to 2,675 articles published in this JOURNAL between 1941 and 2016. He finds that topic modeling produces results very similar to those obtained through human classification methods, and specifically identifies the "cliometric revolution" of the 1960s, when economic historians increasingly engaged with economic theory and used quantitative or econometric methods. Such examples suggest the broad range of possibilities for the analysis of large corpora of economic texts.

Analysis of large-scale textual corpora is still in its infancy but suggests new forms data archiving and analysis might take in the future. For example, as organizational documents and correspondence shift increasingly to digital formats, preserving and analyzing the data they contain becomes simpler and less costly. Although it may still

[8] This database is publicly available at tradingconsequences.blogs.edina.ac.uk.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

be some time before born-digital material becomes a standard component of government archives and individual manuscript collections, digital business records are already available for research. David Kirsch (2009) describes several digital collections that promise a wealth of textual data to business historians. As more textual sources are digitized and more born-digital texts archived and made accessible to scholars, the development of methods to analyze these texts will open up new possibilities for economic historians to study economic opinions and beliefs in the past, for which quantitative data in rectangular form are often lacking. Making these data usable, however, presents its own challenges; some have been addressed by crowdsourcing, or the involvement of non-scientists in the production of scientific data.

Creating Large Data: Citizen Science

The involvement of lay people in scholarship has a lengthy history, dating at least to the founding of the Royal Society of London and continuing in various ways to the present. The relationship between professional scholars in the academy and amateur and citizen scholars ranges from citizen challenges to scientific paradigms to collaborative research (Epstein 1995; Irwin 1995; Wynne 1992). In the past decade, the label "citizen science" has been applied to efforts by scholars to enlist the public in the labor of scientific classification and processing of research sources. The origins of this trend are the increasing ease and decreasing cost of digital imaging that has made it possible to collect quantities of raw data that are several orders of magnitude larger than what could be collected as recently as the early 2000s. Paradigmatic examples come from the physical and biological sciences: digital images of wildlife and galaxies (Swanson, Kosmala, Lintott, et al. 2016; Willett, Lintott, Bamford, et al. 2013). The cost of creating images of historical manuscripts has also declined significantly, such that an individual researcher can easily photograph several thousand pages of manuscript material in a day. Digital images allow for data entry outside the archive, so that it can be undertaken at lower cost by undergraduate research assistants or data entry professionals in lower-income countries, a development heralded by Collins (2015) and Kris Mitchener (2015) separately in the 75th anniversary issue of the *Journal of Economic History*. But the challenge for researchers is that the costs of digital imaging have decreased significantly, while the time it takes to classify an image as containing a particular object (galaxy or animal) or to transcribe the text has not decreased nearly as much.

In some instances, it is still possible to obtain grants to digitize and transcribe data. Yet many raw datasets require more classifications or transcriptions than government or private funding agencies will support. Biological and physical scientists have turned to "citizen science" to cover this gap. Classification or transcription tasks are reduced to the simplest possible element that can be performed by a volunteer working at his or her own computer, without any interaction with the researcher. The use of volunteer labor in data collection and classification is not new (Star and Griesemer 1989) but now occurs on a much larger scale. Today classification or transcription is often done multiple times, ranging from three independent iterations in transcription projects to 10–20 in galaxy classification. Researchers are then responsible for developing a consensus value for each data field, or quantifying the degree of uncertainty about the value.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

The largest citizen science organization, Zooniverse (www.zooniverse.org) grew out of astronomy classification projects and now encompasses more than 60 different projects. Physical and environmental sciences predominate, but the organization also supports more than ten historical transcription projects, working with sources as diverse as historical weather logs (Blaser 2014), New Zealand soldiers' enlistment records, and artists' letters and notes.[9] As readers of this JOURNAL will appreciate, transcribing old text is hard work and has, perhaps, less intrinsic appeal to the general population than identifying wild animals or stars (Eveleigh, Jennett, Blandford, et al. 2014); for that reason, projects in the physical and biological sciences have achieved their goals more quickly than have transcription projects. Nevertheless, the quality of transcription obtained from these projects is comparable to that done by supervised undergraduate or graduate research assistants, and data from crowd-sourced transcriptions are beginning to be used in research (Grayson 2016).

The key software products underlying the Zooniverse projects—Panoptes and Scribe—have recently been opened up so that individual scholars can build their own citizen science projects (Bowyer, Lintott, Hines, et al. 2015).[10] Economic historians will recognize that citizen science is essentially trying to reduce the labor costs of producing raw data by relying on volunteer labor. There are still fixed costs to the researcher of organizing the material before transcription, cleaning and coding after; and variable costs of motivating an unpaid labor force (Jennett, Kloetzer, Schneider, et al. 2016). It is important to recognize that the largest datasets currently used by economic historians—the transcriptions of census and other population data by FamilySearch.org—were created by crowd-sourcing organized by the Church of Jesus Christ of Latter-day Saints, where the labor was motivated by incentives internal to the church to participate in the effort. Similarly, the Cambridge Group for the History of Population and Social Structure collected data by working with volunteer transcribers in parishes across England, predating internet citizen science by several decades (Wrigley and Schofield 1981).

The open nature of citizen science software will allow economic historians more control over what material is transcribed by volunteers for scientific research. Projects created through crowd-sourcing may be small in conventional terms, limited by the number of observations extant in the archives (rows) and the limited variables (columns) in the archival material itself. However, the datasets created through citizen science are often more complex than a rectangular dataset. Each individual field is transcribed multiple times, expanding the dataset linearly by the number of iterations. Moreover, each field comes with a significant amount of *paradata*—data describing and auditing the process by which that single field was created, such as who created it, when it was created, and the Cartesian co-ordinates of the field on the image. A "consensus" algorithm must then be chosen to transform multiple transcriptions of each field into a more familiar cross-sectional or longitudinal dataset.

[9] All Zooniverse projects can be accessed at zooniverse.org. Old Weather is using citizen science to transcribe tabular data in a format that is similar to many economic records of the same era: oldweather.org/. Measuring the ANZACs (Australian and New Zealand Army Corps) uses citizen science to generate data for anthropometric history: measuringtheanzacs.org/.

[10] Researchers can begin building projects at zooniverse.org by following the "Build a Project" link.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Future Prospects and Issues in Big Data

Working with samples or data aggregated into geographic units, economic historians have often been concerned to estimate average values for quantities of interest and how they have changed over time, or to estimate the difference in a quantity of interest between two groups and how it has changed. Extending this framework to a multivariate situation, coefficients in regression can be interpreted as the average effect of a change in the independent variable on the dependent variable. We expect that the form of analysis scholars will undertake with big data will be slightly different. With data on a complete population, sampling error is no longer a consideration, allowing us to estimate precisely other moments of the distribution. Moreover, in a population or otherwise very large dataset it is possible to estimate parameters of interest for different groups, which may be used to infer differences in behavior among those groups. However, big datasets may have substantial non-sampling error and selection issues that demand attention. Data derived from social media sites are an excellent example of this issue (Grimmer 2015; Hargittai 2015).

Big data allow us to focus on the structure of the variance in a population. Standard deviations, not standard errors, are likely to become a measure of greater interest in a big data world. Similarly, big datasets are more likely to include a mix of data at different levels, such as a complete population census or tax register with household and geographic identifiers, and characteristics of the geographic units in which people live (or work). Thus, in a big data world we can identify relationships between units. The assumption we make in sample data that the units are statistically independent of each other is no longer valid or necessary. The characteristics of neighboring geographic units are unlikely to be independent, and their dependence can be measured when data are available for all units. In short, big data will require that economic historians take greater account of correlation between units in their dataset, whether defined in terms of spatial proximity or a relationship within some organization such as a household, school, firm, or government unit. Variance and correlation are computationally more intensive to estimate than averages. If closed-form solutions are not available, estimates of higher moments may need to be bootstrapped.

Record linkage is another computationally intensive task likely to grow in importance in a big data environment. As with the estimation of variance and correlation, the computational demands of record linkage increase faster than the linear increase in the number of observations. Economic and demographic historians have been at the forefront of record linkage in historical population data (Abramitzky, Boustan, and Eriksson 2012; Atack, Bateman, and Gregson 1992; Feigenbaum 2016; Ferrie 1999; Goeken, Huynh, Lynch, et al. 2011; Mill and Stein 2016). Where individual entities in a dataset can be identified unambiguously and assigned a unique identifier that is common across datasets, combining datasets is a trivial one-to-one matching problem. But historical sources are rarely so generous as to provide these identifiers, and modern sources provide them with error. Shared data resources, such as the Union Army datasets (Fogel, Costa, Haines, et al. 2000), the University of Lund's Scanian Economic Demographic Database (Bengtsson, Dribe, Quaranta, et al. 2014), and the China Multigenerational Panel dataset (Lee and Campbell 2016), can be used by many scholars to answer a variety of questions, reducing the need for people to undertake bespoke record linkage.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Beyond linking individual person records between datasets, big historical data facilitate analyses that utilize information from multiple domains: population, climate, and price, for example. When data become highly granular with respect to individual, location, and time, the possibilities to create custom measures combining or linking these grows dramatically. Economic history has a strong tradition of identifying new sources as they become available and linking them to old sources (or vice versa), so we expect that record linkage—across both datasets and domains—will be an important part of economic history's future. Combining sources in this way demands careful attention to detail, an awareness of historical change, and critical examination of the primary data sources, traits that have long characterized research in economic history. Big data in economic history will build on the strengths of the existing research tradition in the discipline.

REFERENCES

Aaronson, Daniel, Rajeev Dehejia, Andrew Jordan, et al. "The Effect of Fertility on Mothers' Labor Supply over the Last Two Centuries." NBER Working Paper No. 23717, Cambridge, MA, 2017.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, et al. "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators." In *Producer Dynamics: New Evidence from Micro Data,* 149–230. Chicago: University of Chicago Press.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102, no. 5 (2012): 1832–56.

———. "Have the Poor Always Been Less Likely to Migrate? Evidence from Inheritance Practices During the Age of Mass Migration." *Journal of Development Economics* 102 (2013): 2–14.

———. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122, no. 3 (2014): 467–506.

Anderson, Albert F. "Statistics and Massive Data Sets: One View from the Social Sciences." In *Massive Datasets: Proceedings of a Workshop,* Committee on Applied and Theoretical Statistics, National Research Council, 33–38. Washington, DC: National Academies Press, 1997.

Antenucci, Dolan, Michael J. Cafarella, Margaret C. Levenstein, et al. "Using Social Media to Measure Labor Market Flows." NBER Working Paper No. 20010, Cambridge, MA, 2014.

Ashplant, T.G., and Adrian Wilson. "Present-Centred History and the Problem of Historical Knowledge." *Historical Journal* 31, no. 2 (1988): 253–74.

Atack, Jeremy, Fred Bateman, and Mary Eschelbach Gregson. "'Matchmaker, Matchmaker, Make Me a Match' a General Personal Computer-Based Matching Program for Historical Research." *Historical Methods* 25, no. 2 (1992): 53–65.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49, no. 1 (2011): 3–71.

Bagger, Jesper, and Andrew Seltzer. "Administrative and Survey Data in Personnel Economics." *Australian Economic Review* 47, no. 1 (2014): 137–46.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Bailey, Martha. "Longitudinal, Intergenerational Family Electronic Micro-Database Project." Ann Arbor: University of Michigan, 2017. Available at https://sites.lsa.umich.edu/life-m/.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo. "The Making of Modern America: Migratory Flows in the Age of Mass Migration." *Journal of Development Economics* 102 (2013): 23–47.

Bankoff, Greg. "Comparing Vulnerabilities: Toward Charting an Historical Trajectory of Disasters." *Historical Social Research/Historische Sozialforschung* (2007): 103–14.

Barclay, Kieron, Katherine Keenan, Emily Grundy, et al. "Reproductive History and Post-Reproductive Mortality: A Sibling Comparison Analysis Using Swedish Register Data." *Social Science & Medicine* 155 (2016): 82–92.

Barclay, Kieron, and Martin Kolk. "Birth Order and Mortality: A Population-Based Cohort Study." *Demography* 52, no. 2 (2015): 613–39.

Bates, David W., Suchi Saria, Lucila Ohno-Machado, et al. "Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients." *Health Affairs* 33, no. 7 (2014): 1123–31.

Bauer, Thomas K., Stefan Bender, Jörg Heining, et al. "The Lunar Cycle, Sunspots and the Frequency of Births in Germany, 1920–1989." *Economics & Human Biology* 11, no. 4 (2013): 545–50.

Beach, Brian, Joseph Ferrie, Martin Saavedra, et al. "Typhoid Fever, Water Quality, and Human Capital Formation." *Journal of Economic History* 76, no. 1 (2016): 41–75.

Bearman, Peter. "Big Data and Historical Social Science." *Big Data & Society* 2, no. 2 (2015).

Becker, Sascha O., and Ludger Woessmann. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124, no. 2 (2009): 531–96.

Bengtsson, Tommy, Cameron Campbell, James Z. Lee, et al. *Life Under Pressure: Mortality and Living Standards in Europe and Asia, 1700–1900.* Cambridge: MIT Press 2004.

Bengtsson, Tommy, Martin Dribe, Luciana Quaranta, et al. "The Scanian Economic Demographic Database, Version 4.0 (Machine-Readable Database)." Lund: University of Lund, 2014. https://www.ed.lu.se/databases/sedd.

Bholat, David. "Big Data and Central Banks." *Big Data & Society* 2, no. 1 (2015).

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. "Under Pressure? The Effect of Peers on Outcomes of Young Adults." *Journal of Labor Economics* 31, no. 1 (2013): 119–53.

Blake, Kellee. " 'First in the Path of the Firemen': The Fate of the 1890 Population Census." *Prologue Magazine* 28, no. 1 (1996): 64–81.

Blaser, Lucinda. "Old Weather: Approaching Collections from a Different Angle." In *Crowdsourcing Our Cultural Heritage,* edited by Mia Ridge, 45–56. London: Routledge, 2014.

Bleakley, Hoyt, and Joseph Ferrie. "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital across Generations." *Quarterly Journal of Economics* 131, no. 3 (2016): 1455–95.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Boberg-Fazlic, Nina, Paul Sharp, and Jacob Weisdorf. "Survival of the Richest? Social Status, Fertility and Social Mobility in England 1541–1824." *European Review of Economic History* 15, no. 3 (2011): 365–92.

Boustan, Leah Platt. "Competition in the Promised Land: Black Migration and Racial Wage Convergence in the North, 1940–1970." *Journal of Economic History* 69, no. 3 (2009): 755–82.

Bowyer, Alex, Chris Lintott, Greg Hines, et al. "Panoptes, a Project Building Tool for Citizen Science." *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Human Computation and Crowdsourcing (HCOMP 15).* San Diego: AAAI Press, 2015.

Broeckel, Ulrich, Christian Hengstenberg, Bjoern Mayer, et al. "A Locus on Chromosome 10 Influences C-Reactive Protein Levels in Two Independent Populations." *Human Genetics* 122, no. 1 (2007): 95–102.

Burrows, Roger, and Mike Savage. "After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology." *Big Data & Society* 1, no. 1 (2014). doi: 10.1177/2053951714540280.

Candido dos Reis, Francisco J., Stuart Lynn, H. Raza Ali, et al. "Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer." *EBioMedicine* 2, no. 7 (2015): 681–89.

Cantoni, Davide. "The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands." *Journal of the European Economic Association* 13, no. 4 (2015): 561–98.

Cavallo, Alberto, and Roberto Rigobon. "The Billion Prices Project: Using Online Prices for Measurement and Research." *Journal of Economic Perspectives* 30, no. 2 (2016): 151–78.

Chen, Qiang. "Climate Shocks, Dynastic Cycles and Nomadic Conquests: Evidence from Historical China." *Oxford Economic Papers* 67, no. 2 (2015): 185–204.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, et al. "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility." *American Economic Review* 104, no. 5 (2014): 141–47.

Chetty, Raj, Michael Stepner, Sarah Abraham, et al. "The Association between Income and Life Expectancy in the United States, 2001–2014." *JAMA* 315, no. 16 (2016): 1750–66.

Claus, Iris, John Creedy, and Josh Teng. "The Elasticity of Taxable Income in New Zealand." *Fiscal Studies* 33, no. 3 (2012): 287–303.

Clotfelter, Charles T. "Tax Evasion and Tax Rates: An Analysis of Individual Returns." *Review of Economics and Statistics* 65, no. 3 (1983): 363–73.

Collins, William J. "Looking Forward: Positive and Normative Views of Economic History's Future." *Journal of Economic History* 75, no. 4 (2015): 1228–33.

Collins, William J., and Marianne H. Wanamaker. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants." *Journal of Economic History* 75, no. 4 (2015): 947–92.

Conrad, Alfred J., and John R. Meyer. "The Economics of Slavery in the Antebellum South." *Journal of Political Economy* 66, no. 2 (1958): 95–130.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. "The Rise and Decline of the American Ghetto." *Journal of Political Economy* 107, no. 3 (1999): 455–506.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Daly, Christopher, Wayne P. Gibson, George H. Taylor, et al. "A Knowledge-Based Approach to the Statistical Mapping of Climate." *Climate Research* 22, no. 2 (2002): 99–113.

Daly, Christopher, Michael Halbleib, Joseph I. Smith, et al. "Physiographically Sensitive Mapping of Climatological Temperature and Precipitation across the Conterminous United States." *International Journal of Climatology* 28, no. 15 (2008): 2031–64.

Deane, Glenn, and Myron P. Gutmann. "Blowin'down the Road: Investigating Bilateral Causality between Dust Storms and Population in the Great Plains." *Population Research and Policy Review* 22, no. 4 (2003): 297–331.

Devereux, Paul J., Sandra E. Black, and Kjell G Salvanes. "From the Cradle to the Labor Market? The Effect of Birth Weight." *Quarterly Journal of Economics* 122, no. 1 (2007): 409–39.

Dillon, Lisa, Marilyn Amorevieta-Gentil, Marianne Caron, et al. "The Programme De Recherche En Démographie Historique: Past, Present and Future Developments in Family Reconstitution." *History of the Family* (2017): 1–34.

Dittmar, Jeremiah, and Skipper Seabold. "Media, Markets and Institutional Change: The Protestant Reformation." Center for Economic Performance Discussion Paper, London, UK, 2015.

Dorman, Robert L. "The Creation and Destruction of the 1890 Federal Census." *American Archivist* 71 (2008): 350–83.

Dribe, Martin, J. David Hacker, and Francesco Scalone. "The Impact of Socio-Economic Status on Net Fertility During the Historical Fertility Decline: A Comparative Analysis of Canada, Iceland, Sweden, Norway, and the USA." *Population Studies* 68, no. 2 (2014): 135–49.

Dribe, Martin, and Jonas Helgertz. "The Lasting Impact of Grandfathers: Class, Occupational Status, and Earnings over Three Generations in Sweden 1815–2011." *Journal of Economic History* 76, no. 4 (2016): 969–1000.

Easterlin, Richard A. "Population Change and Farm Settlement in the Northern United States." *Journal of Economic History* 36, no. 1 (1976): 45–75.

Eijkemans, Marinus J.C., Frans Van Poppel, Dik F. Habbema, et al. "Too Old to Have Children? Lessons from Natural Fertility Populations." *Human Reproduction* (2014): 1–9.

Einav, Liran, and Jonathan Levin. "Economics in the Age of Big Data." *Science* 346, no. 6210 (2014): 1243089.

Epstein, Steven. "The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials." *Science, Technology & Human Values* 20, no. 4 (1995): 408–37.

Eveleigh, Alexandra, Charlene Jennett, Ann Blandford, et al. "Designing for Dabblers and Deterring Drop-Outs in Citizen Science." In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems,* 2985–94. Toronto, Canada: Association for Computing Machinery, 2014.

Feigenbaum, James J. "A Machine Learning Approach to Census Record Linking." Cambridge: Harvard University, 2016. Available at http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf.

Feldman, Naomi E.; Peter Katuščák, and Laura Kawano. "Taxpayer Confusion: Evidence from the Child Tax Credit." *American Economic Review* 106, no. 3 (2016): 807–35.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Ferrie, Joseph P. "The Wealth Accumulation of Antebellum European Immigrants to the U.S., 1840–60." *Journal of Economic History* 54, no. 1 (1994): 1–33.

———. *Yankeys Now: Immigrants in the Antebellum United States*. New York: Oxford University Press, 1999.

———. "History Lessons: The End of American Exceptionalism? Mobility in the United States since 1850." *Journal of Economic Perspectives* 19, no. 3 (2005): 199–215.

Ferrie, Joseph P., and Karen Rolf. "Socioeconomic Status in Childhood and Health after Age 70: A New Longitudinal Analysis for the U.S., 1895–2005." *Explorations in Economic History* 48, no. 4 (2011): 445–60.

Fitzgerald, Michael. "Better Data Brings a Renewal at the Bank of England." *MIT Sloan Management Review* (May 2016): 3–13.

Fleisig, Heywood W. "Slavery, the Supply of Agricultural Labor, and the Industrialization of the South." *Journal of Economic History* 36, no. 3 (1976): 572–97.

Fogel, Robert W., Dora L. Costa, Michael Haines, et al. "Aging of Veterans of the Union Army: Version M-5." Center for Population Economics University of Chicago Graduate School of Business, Department of Economics Brigham Young University, and The National Bureau of Economic Research, 2000. http://uadata.org.

Fogel, Robert William, and Stanley L. Engerman. *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little, Brown and Company, 1974.

Franzosi, Roberto, Gianluca De Fazio, and Stefania Vicari. "Ways of Measuring Agency: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875–1930)." *Sociological Methodology* 42, no. 1 (2012): 1–42.

Gagnon, Alain, Ken R. Smith, Marc Tremblay, et al. "Is There a Trade-Off between Fertility and Longevity? A Comparative Study of Women from Three Large Historical Databases Accounting for Mortality Selection." *American Journal of Human Biology* 21, no. 4 (2009): 533–40.

Gagnon, Alain, Marc Tremblay, Hélène Vézina, et al. "Once Were Farmers: Occupation, Social Mobility, and Mortality During Industrialization in Saguenay-Lac-Saint-Jean, Quebec 1840–1971." *Explorations in Economic History* 48, no. 3 (2011): 429–40.

Gandomi, Amir, and Murtaza Haider. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35, no. 2 (2015): 137–44.

Gao, Cheng, and Bruce Mizrach. "Market Quality Breakdowns in Equities." *Journal of Financial Markets* 28 (2016): 1–23.

Gentzkow, Matthew, and Jesse M Shapiro. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* 78, no. 1 (2010): 35–71.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, et al. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." *Economic Inquiry* 56, no. 1 (2018): 114–37.

Goeken, Ron, Lap Huynh, T. A. Lynch, et al. "New Methods of Census Record Linking." *Historical Methods* 44, no. 1 (2011): 7–14.

Goldewijk, Kees Klein. "Estimating Global Land Use Change over the Past 300 Years: The Hyde Database." *Global Biogeochemical Cycles* 15, no. 2 (2001): 417–33.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Goldin, Claudia. "Female Labor Force Participation: The Origin of Black and White Differences, 1870 and 1880." *Journal of Economic History* 37, no. 1 (1977): 87–108.

González, Felipe, Guillermo Marshall, and Suresh Naidu. "Start-up Nation? Slave Wealth and Entrepreneurship in Civil War Maryland." *Journal of Economic History* 77, no. 2 (2017): 373–405.

Graham, Mark, and Taylor Shelton. "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography* 3, no. 3 (2013): 255–61.

Grayson, Richard. "A Life in the Trenches? The Use of Operation War Diary and Crowdsourcing Methods to Provide an Understanding of the British Army's Day-to-Day Life on the Western Front." *British Journal for Military History* 2, no. 2 (2016).

Gregory, Ian N., Jordi Marti-Henneberg, and Francisco J. Tapiador. "Modelling Long-Term Pan-European Population Change from 1870 to 2000 by Using Geographical Information Systems." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173, no. 1 (2010): 31–50.

Griffith, Rachel, and Martin O'Connell. "The Use of Scanner Data for Research into Nutrition." *Fiscal Studies* 30, no. 3–4 (2009): 339–65.

Grimmer, Justin. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48, no. 1 (2015): 80–83.

Groves, Robert M. "Three Eras of Survey Research." *Public Opinion Quarterly* 75, no. 5 (2011): 861–71.

Grundy, Emily, and Øystein Kravdal. "Fertility History and Cause-Specific Mortality: A Register-Based Analysis of Complete Cohorts of Norwegian Women and Men." *Social Science & Medicine* 70, no. 11 (2010): 1847–57.

———. "Do Short Birth Intervals Have Long-Term Implications for Parental Health? Results from Analyses of Complete Cohort Norwegian Register Data." *Journal of Epidemiology and Community Health* 68, no. 10 (2014): 958–64.

Gullickson, Aaron. "Black/White Interracial Marriage Trends, 1850–2000." *Journal of Family History* 31, no 3 (2006): 289–312.

Gutmann, Myron P. "Great Plains Population and Environment Data: Agricultural Data, 1870–1997 [United States]." Ann Arbor: Inter-university Consortium for Political and Social Research, 2005.

———. "Great Plains Population and Environment Data: Social and Demographic Data, 1870–2000 [United States]." Ann Arbor: Inter-university Consortium for Political and Social Research, 2007.

Gutmann, Myron P., Daniel Brown, Angela R. Cunningham, et al. "Migration in the 1930s: Beyond the Dust Bowl." *Social Science History* 40, no. 4 (2016): 707–40.

Hacker, J. David, and Evan Roberts. "The Impact of Kin Availability, Parental Religiosity, and Nativity on Fertility Differentials in the Late Nineteenth-Century United States." *Demographic Research* 37, no. 34 (2017): 1049–80.

Hacking, Ian. "Biopower and the Avalanche of Printed Numbers." *Humanities in Society* 5 (1982): 279–95.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Haines, Michael R. "Historical, Demographic, Economic, and Social Data: The United States, 1790–2002." Ann Arbor: Inter-university Consortium for Political and Social Research, 2010.

Hall, Patricia Kelly, Robert McCaa, and Gunnar Thorvaldsen, eds. *A Handbook of International Historical Microdata for Population Research*. Minneapolis: Minnesota Population Center, 2000.

Ham, F. Gerald. "Archival Choices: Managing the Historical Record in an Age of Abundance." *Armerican Archivist* 47, no. 1 (1984): 11–22.

Hargittai, Eszter. "Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites." *Annals of the American Academy of Political and Social Science* 659, no. 1 (2015): 63–76.

Hartman, Melannie D., Emily R. Merchant, William J. Parton, et al. "Impact of Historical Land-Use Changes on Greenhouse Gas Exchange in the US Great Plains, 1883–2003." *Ecological Applications* 21, no. 4 (2011): 1105–19.

Hausman, Jerry. "Sources of Bias and Solutions to Bias in the Consumer Price Index." *Journal of Economic Perspectives* 17, no. 1 (2003): 23–44.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. "Measuring Economic Growth from Outer Space." *American Economic Review* 102, no. 2 (2012): 994–1028.

Herweijer, Celine, Richard Seager, Edward R. Cook, et al. "North American Droughts of the Last Millennium from a Gridded Network of Tree-Ring Data." *Journal of Climate* 20, no. 7 (2007): 1353–76.

Humphries, Jane. " 'The Most Free from Objection ...' The Sexual Division of Labor and Women's Work in Nineteenth-Century England." *Journal of Economic History* 47, no. 4 (1987): 929–49.

Index Committee. "Report of Index Committee." Wellington, NZ: Appendices to the Journals of the House of Representatives, 1948.

Irwin, Alan. *Citizen Science: A Study of People, Expertise and Sustainable Development*. New York: Routledge, 1995.

Ivancic, Lorraine, W. Erwin Diewert, and Kevin J. Fox. "Scanner Data, Time Aggregation and the Construction of Price Indexes." *Journal of Econometrics* 161, no. 1 (2011): 24–35.

Jackson, R. V. *Index to the Eighth Census of the United States*. Salt Lake City: Accelerated Indexing Systems International, 1992.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu. "Political Language in Economics." Research Paper No. 14-57, Columbia Business School, New York, NY, 2015.

Jennett, Charlene, Laure Kloetzer, Daniel Schneider, et al. "Motivations, Learning and Creativity in Online Citizen Science." *Journal of Science Communication* 15, no. 3 (2016): 1–23.

Jennings, Julia A., Allison R. Sullivan, and J. David Hacker. "Intergenerational Transmission of Reproductive Behavior During the Demographic Transition." *Journal of Interdisciplinary History* 42, no. 4 (2012): 543–69.

Jensen, Jacob, Suresh Naidu, Ethan Kaplan, et al. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech [with Comments and Discussion]." *Brookings Papers on Economic Activity* 2 (2012): 1–81.

Jia, Ruixue. "Weather Shocks, Sweet Potatoes and Peasant Revolts in Historical China." *Economic Journal* 124, no. 575 (2014): 92–118.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Johnson, Ryan S. "The Economic Progress of American Black Workers in a Period of Crisis and Change, 1916–1950." *Journal of Economic History* 64, no. 2 (2004): 552–58.

Kaplan, Greg, and Sam Schulhofer-Wohl. "Inflation at the Household Level." NBER Working Paper No. 22331, Cambridge, MA, 2016.

Kaplan, Jed O., Kristen M. Krumhardt, and Niklaus Zimmermann. "The Prehistoric and Preindustrial Deforestation of Europe." *Quaternary Science Reviews* 28, no. 27 (2009): 3016–34.

Katz, Michael B. *The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth-Century City*. Cambridge: Harvard University Press, 1975.

Kay, David, and Mark van Harmelen. *Activity Data - Delivering Benefits from the Data Deluge*. London: Joint Information Systems Committee, 2012.

Kerber, Richard A., Elizabeth O'Brien, Ken R. Smith, et al. "Familial Excess Longevity in Utah Genealogies." *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 56, no. 3 (2001): B130–B39.

Khoury, Muin J. "Planning for the Future of Epidemiology in the Era of Big Data and Precision Medicine." *American Journal of Epidemiology* 182, no. 12 (2015): 977–79.

King, Gary. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331, no. 6018 (2011): 719–21.

Kirsch, David A. "The Record of Business and the Future of Business History: Establishing a Public Interest in Private Business Records." *Library Trends* 57, no. 3 (2009): 352–70.

Knights, Peter R. *The Plain People of Boston, 1830–1860: A Study in City Growth*. New York: Oxford University Press, 1971.

Kohara, Miki, and Yusuke Kamiya. "Maternal Employment and Food Produced at Home: Evidence from Japanese Data." *Review of Economics of the Household* 14, no. 2 (2016): 417–42.

Kolk, Martin. "Multigenerational Transmission of Family Size in Contemporary Sweden." *Population Studies* 68, no. 1 (2014): 111–29.

Kosnik, Lea-Rachel D. "What Have Economists Been Doing for the Last 50 Years? A Text Analysis of Published Academic Research from 1960–2010." *Economics* 9 (2015): 1–38.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111, no. 24 (2014): 8788–90.

Lazear, Edward P., and Kathryn L Shaw. *The Structure of Wages: An International Comparison*. Chicago: University of Chicago Press, 2009.

Lee, James Z., and Cameron D. Campbell. "China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN), 1749–1909." Ann Arbor: Inter-university Consortium for Political and Social Research, 2016. http://www.icpsr.umich.edu/icpsrweb/DSDR/series/00265.

Lee, James Z., Shuang Chen, Cameron D. Campbell, et al. "China Multi-Generational Panel Dataset, Shuangcheng (CMGPD-SC), 1866–1913." Ann Arbor: Inter-university Consortium for Political and Social Research, 2017. http://www.icpsr.umich.edu/icpsrweb/DSDR/series/00265.

Leicester, Andrew, and Zoe Oldfield. "Using Scanner Technology to Collect Expenditure Data." *Fiscal Studies* 30, no. 3–4 (2009): 309–37.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Levin, Sharon G., Stanford L. Levin, and John B. Meisel. "Market Structure, Uncertainty, and Intrafirm Diffusion: The Case of Optical Scanners in Grocery Stores." *Review of Economics and Statistics* 74, no. 2 (1992): 345–50.

Liu, Mingliang, and Hanqin Tian. "China's Land Cover and Land Use Change from 1700 to 2005: Estimations from High-Resolution Satellite Data and Historical Archives." *Global Biogeochemical Cycles* 24, no. 3 (2010).

Lloyd, Christopher, Jacob Metzer, and Richard Sutch. *Settler Economies in World History*. Leiden: Brill, 2013.

Logan, John R., and Hyoung-jin Shin. "Assimilation By the Third Generation? Marital Choices of White Ethnics at the Dawn of the Twentieth Century." *Social Science Research* 41, no. 5 (2012): 1116–1125.

Logan, John R., and Weiwei Zhang. "White Ethnic Residential Segregation in Historical Perspective: US Cities in 1880." *Social Science Research* 41, no. 5 (2012): 1292–306.

Logan, Trevon, and John Parman. "The National Rise in Residential Segregation." *Journal of Economic History* 77, no. 1 (2017): 127–70.

Long, Jason. "Rural-Urban Migration and Socioeconomic Mobility in Victorian Britain." *Journal of Economic History* 65, no. 1 (2005): 1–35.

Long, Jason, and Joseph P. Ferrie. "The Path to Convergence: Intergenerational Occupational Mobility in Britain and the US in Three Eras." *Economic Journal* 117, no. 519 (2007): C61–C71.

———. "Intergenerational Occupational Mobility in Britain and the U.S. Since 1850." *American Economic Review* 103, no. 4 (2013): 1109–37.

Lundh, Christer, and Satomi Kurosu, et al. *Similarity in Difference: Marriage in Europe and Asia, 1700–1900*. Cambridge: MIT Press, 2014.

Lusk, Jayson L., and Kathleen Brooks. "Who Participates in Household Scanning Panels?" *American Journal of Agricultural Economics* 93, no. 1 (2011): 226–40.

Lynge, Elsebeth, Jakob Lynge Sandegaard, and Matejka Rebolj. "The Danish National Patient Register." *Scandinavian Journal of Public Health* 39, no. 7 suppl. (2011): 30–33.

Maloney, Thomas N., Heidi Hanson, and Ken Smith. "Occupation and Fertility on the Frontier: Evidence from the State of Utah." *Demographic Research* 30 (2014): 853–86.

Maxwell, Susan K., and Kenneth M. Sylvester. "Identification of 'Ever-Cropped' Land (1984–2010) Using Landsat Annual Maximum NDVI Image Composites: Southwestern Kansas Case Study." *Remote Sensing of Environment* 121 (2012): 186–95.

Melser, Daniel. "Accounting for the Effects of New and Disappearing Goods Using Scanner Data." *Review of Income and Wealth* 52, no. 4 (2006): 547–68.

Mill, Roy, and Luke C.D. Stein. "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America." Tucson: Arizona State University, 2016. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2741797.

Miller, Rena S., and Gary Shorter. "High Frequency Trading: Overview of Recent Developments." Washington, DC: Congressional Research Service, 2016.

Minnesota Population Center. "North Atlantic Population Project: Complete Count Microdata. Version 2.2" [machine readable database]. Minneapolis, MN: Minnesota Population Center [distributor], 2015a.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

———. 2015b. *Terra Populus [Dataset]*. Minneapolis: Minnesota Population Center.
Mitchener, Kris James. "The 4D Future of Economic History: Digitally-Driven Data Design." *Journal of Economic History* 75, no. 4 (2015): 1234–39.
Monroe, Burt. "The Five Vs of Big Data Political Science: Introduction to the Virtual Issue on Big Data in Political Science." *Political Analysis*, Virtual Issue 4 (2013): 1–9.
National Centers for Environmental Information. "Climate Data Online." Silver Spring, MD: National Oceanic and Atmospheric Administration, 2016.
Newman, David J., and Sharon Block. "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *Journal of the American Society for Information Science and Technology* 57, no. 6 (2006): 753–67.
O'Connell, Allan F., James D. Nichols, and K. Ullas Karanth. *Camera Traps in Animal Ecology: Methods and Analyses*. Dordrecht: Springer, 2011.
Olmstead, Alan L., and Paul W. Rhode. "Adapting North American Wheat Production to Climatic Challenges, 1839–2009." *Proceedings of the National Academy of Sciences* 108, no. 2 (2011): 480–85.
———. "Were Antebellum Cotton Plantations Factories in the Field?" In *Enterprising America: Businesses, Banks, and Credit Markets in Historical Perspective,* edited by William J. Collins and Robert A. Margo, 245–76. Chicago: University of Chicago Press, 2015.
Parker, William N. *The Structure of the Cotton Economy of the Antebellum South*. Washington, DC: Agricultural History Society, 1970.
Parker, William N., and Robert E. Gallman. "Southern Farms Study, 1860." Ann Arbor: Inter-university Consortium for Political and Social Research, 1991.
Parton, William J., Myron P. Gutmann, Melannie D. Hartman, et al. "Great Plains Population and Environment Data: Biogeochemical Modeling Data, 1860–2003." Ann Arbor: Inter-university Consortium for Political and Social Research, 2012.
Parton, William J., Myron P. Gutmann, Melannie D. Hartman, et al. "Simulated Biogeochemical Impacts of Historical Land-Use Changes in the U.S. Great Plains from 1870 to 2003." In *Land Use and the Carbon Cycle: Science and Applications in Coupled Natural-Human Systems*, edited by Daniel G. Brown, Derek T. Robinson, Nancy H. F. French, et al., 287–304. New York: Cambridge University Press, 2013.
Parton, William J., Myron P. Gutmann, Emily R. Merchant, et al. "Measuring and Mitigating Agricultural Greenhouse Gas Production in the US Great Plains, 1870–2000." *Proceedings of the National Academy of Sciences* 112, no. 34 (2015): E4681–E88.
Pearson, David. *Johnsonville, Continuity and Change in a New Zealand Township*. Sydney: George Allen & Unwin, 1980.
Porter, John H., Paul C. Hanson, and Chau-Chin Lin. "Staying Afloat in the Sensor Data Deluge." *Trends in Ecology & Evolution* 27, no. 2 (2011): 121–29.
Roberts, Evan, and John Robert Warren. "Family Structure and Childhood Anthropometry in Saint Paul, Minnesota in 1918." *History of the Family* 22, no. 2–3 (2017): 258–90.
Rönnbäck, Klas. "Climate, Conflicts, and Variations in Prices on Pre-Colonial West African Markets for Staple Crops." *Economic History Review* 67, no. 4 (2014): 1065–88.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Ruggles, Steven. "Reconsidering the Northwest European Family System: Living Arrangements of the Aged in Comparative Historical Perspective." *Population and Development Review* 35, no. 2 (2009): 249–73.

———. "Big Microdata for Population Research." *Demography* 51, no. 1 (2014): 287–97.

Ruggles, Steven, Katie Genadek, Ronald Goeken, et al. "Integrated Public Use Microdata Series: Version 6.0" [machine-readable database]. Minneapolis, MN: University of Minnesota, 2015. www.ipums.org.

Ruggles, Steven, and Russell Menard. "The Minnesota Historical Census Projects." *Historical Methods* 28, no. 1 (1995): 6–10.

Ruggles, Steven, Evan Roberts, Sula Sarkar, et al. "The North Atlantic Population Project: Progress and Prospects." *Historical Methods* 44, no. 1 (2011): 1–6.

Saperstein, Aliya, and Aaron Gullickson. "A 'Mulatto Escape Hatch' in the United States? Examining Evidence of Racial and Social Mobility During the Jim Crow Era." *Demography* 50, no. 5 (2013): 1921–42.

Shah, Dhavan V., Joseph N. Cappella, and W. Russell Neuman. "Big Data, Digital Media, and Computational Social Science: Possibilities and Perils." *Annals of the American Academy of Political and Social Science* 659, no. 1 (2015): 6–13.

Silver, Mick, and Saeed Heravi. "Scanner Data and the Measurement of Inflation." *Economic Journal* 111, no. 472 (2001): F383–F404.

Smith, Ken R. and Huntsman Cancer Institute. "Utah Population Database." Salt Lake City: University of Utah, 2017. https://healthcare.utah.edu/huntsmancancerinstitute/research/updb/.

Star, Susan Leigh, and James R. Griesemer. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39." *Social Studies of Science* 19, no. 3 (1989): 387–420.

Sundstrom, William A. "The Geography of Wage Discrimination in the Pre–Civil Rights South." *Journal of Economic History* 67, no. 2 (2007): 410–44.

Swanson, Alexandra, Margaret Kosmala, Chris Lintott, et al. "Snapshot Serengeti, High-Frequency Annotated Camera Trap Images of 40 Mammalian Species in an African Savanna." *Scientific Data* 2 (2015): 150026.

Swanson, Alexandra, Margaret Kosmala, Chris Lintott, et al. "A Generalized Approach for Producing, Quantifying, and Validating Citizen Science Data from Wildlife Images." *Conservation Biology* 30, no. 3 (2016): 520–31.

Sylvester, Kenneth M., Daniel G. Brown, Glenn D. Deane, et al. "Land Transitions in the American Plains: Multilevel Modeling of Drivers of Grassland Conversion (1956–2006)." *Agriculture, Ecosystems & Environment* 168 (2013): 7–15.

Sylvester, Kenneth M., Daniel G. Brown, Susan H. Leonard, et al. "Exploring Agent-Level Calculations of Risk and Returns in Relation to Observed Land-Use Changes in the U.S. Great Plains, 1870–1940." *Regional Environmental Change* 15, no. 2 (2015): 301–15.

Sylvester, Kenneth M., Myron P. Gutmann, and Daniel G. Brown. "At the Margins: Agriculture, Subsidies and the Shifting Fate of North America's Native Grassland." *Population and Environment* 37, no. 3 (2016): 362–90.

Sylvester, Kenneth M., and Eric S. A. Rupley. "Revising the Dust Bowl: High above the Kansas Grasslands." *Environmental History* 17, no. 3 (2012): 603–33.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press

Thernstrom, Stephan. *Poverty and Progress; Social Mobility in a Nineteenth Century City*. Cambridge: Harvard University Press, 1964.

———. *The Other Bostonians; Poverty and Progress in the American Metropolis, 1880–1970*. Cambridge: Harvard University Press, 1973.

Tinati, Ramine, Susan Halford, Leslie Carr, et al. "Big Data: Methodological Challenges and Approaches for Sociological Analysis." *Sociology* 48, no. 4 (2014): 663–81.

Tsuya, Noriko O., Feng Wang, George Alter, et al. *Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700–1900*. Cambridge: MIT Press, 2010.

Turner, Frederick Jackson. "The Significance of the Frontier in American History." *Proceedings of the State Historical Society of Wisconsin* 41 (1893): 79–112.

Université du Québec à Chicoutimi; Université Laval; McGill University and Université de Montréal. "BALSAC Population Database." Chicoutimi: Université du Québec à Chicoutimi, 2017. http://balsac.uqac.ca/.

Varian, Hal R. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28, no. 2 (2014): 3–27.

Ward, Jonathan Stuart, and Adam Barker. "Undefined by Data: A Survey of Big Data Definitions." arXiv preprint arXiv:1309.5821, 2013.

Wehrheim, Lino. "Economic History Goes Digital: Topic Modeling the *Journal of Economic History*." BGPE Discussion Paper No. 177 (November 2017): https://www.researchgate.net/profile/Lino_Wehrheim/publication/321213391_Economic_History_Goes_Digital_Topic_Modeling_the_Journal_of_Economic_History/links/5a15485b458515005213298e/Economic-History-Goes-Digital-Topic-Modeling-the-Journal-of-Economic-History.pdf.

Willett, Kyle W., Chris J. Lintott, Steven P. Bamford, et al. "Galaxy Zoo 2: Detailed Morphological Classifications for 304 122 Galaxies from the Sloan Digital Sky Survey." *Monthly Notices of the Royal Astronomical Society* 435, no. 4 (2013): 2835–60.

Wrigley, Edward A., and Roger Schofield. *The Population History of England 1541–1871*. Cambridge: Cambridge University Press, 1981.

Wyber, Rosemary, Samuel Vaillancourt, William Perry, et al. "Big Data in Global Health: Improving Health in Low-and Middle-Income Countries." *Bulletin of the World Health Organization* 93, no. 3 (2015): 203–208.

Wynne, Brian. "Misunderstood Misunderstanding: Social Identities and Public Uptake of Science." *Public Understanding of Science* 1, no. 3 (1992): 281–304.

https://doi.org/10.1017/S0022050718000177 Published online by Cambridge University Press