The IT of Demography

Emily Klancher Merchant, *University of California, Davis, CA, 95616, USA*
Myron P. Gutmann, *University of Colorado Boulder, CO, 80309, USA*

The 1787 drafting of the U.S. Constitution triggered what Ian Hacking [28] has termed an "avalanche of printed numbers." Democratic self-government required an accounting of the demos that was to govern itself through elected representatives. As democratic revolutions swept across Europe and Latin America, countries in those regions also began to enumerate their citizens in new ways. In addition to counting, governments collected information about their citizens that facilitated state power over how people lived, and over life itself [21].

In the United States, decennial censuses increased in size and complexity across the nineteenth century, and the information they collected became the focal points of a series of policy debates, beginning with questions about the abolition of slavery [2]. In 1850, after a data-collection debacle in the 1840 census that appeared to provide ammunition to supporters of slavery [17], the individual replaced the household as the unit of enumeration, swelling the quantity of data collected every ten years. The task of converting manuscript census returns into volumes of printed numbers inspired new methods of data storage and manipulation: first data sheets and tally marks, and then in 1890 punched cards and tabulating machines [64]. Since then, the histories of computing technology and demographic research and analysis have been closely intertwined.

The electromechanical tabulator, in use until the 1950 census, was developed by Herman Hollerith, who had previously worked as a statistician on the 1880 Census.[1] Hollerith's Tabulating Machine Company eventually became the International Business Machines Corporation (IBM), which supplied machines to tabulate the 1950 Census, the last to use punched cards for tabulation. In 1960, the Census Bureau began to use electronic computers to capture and analyze data. This story is now well known [2], [64]. What is less well known are the cycles of innovation that have taken place in the 20th and 21st centuries that link the transformation of computing technology with an increasing capacity to understand and analyze human populations. The articles in this issue engage with that history, exploring the impact of rapid technological change on demographic understanding and the parallel ways that the need to analyze population dynamics spurred the development of new technologies.

Our contributors tell stories of how governments, social scientists, and private industry have perceived, utilized, and preserved demographic data using a variety of information technologies, and how they promoted the development of information technologies to suit their analytic needs. Their work provides a series of examples, in which innovations in computing technology are embedded in the analysis of population dynamics. These stories go beyond the process of counting the number of inhabitants of the United States, and reveal the variety of ways that our understanding of human population has been studied since the time of the U.S. Constitution. This introduction places those stories into a broader narrative about the coproduction of demography and information technology, examining the mutual transformation of computational technology and the collection and analysis of demographic data.

DATA TABULATION BY GOVERNMENT STATISTICAL AGENCIES

Hollerith's tabulating machine was the starting point for twentieth-century developments that increased the speed and efficiency with which U.S. Census data were captured and processed. Mechanical tabulation relied on a new technology of data capture and storage: a card onto which information could be punched, which would then be sorted and counted by the electromechanical tabulator. Over the first half of the 20th century, the Census Bureau, in competition with IBM and Remington Rand (another company with Census Bureau roots), improved the card punching, sorting, and tabulation process as much as possible [64].

[1]https://www.census.gov/history/www/census_then_now/notable_alumni/herman_hollerith.html

1058-6180 © 2022 IEEE
Digital Object Identifier 10.1109/MAHC.2022.3214634
Date of current version 29 November 2022.

In the 1940s, the Bureau contracted for the first commercial computer with the newly-founded Eckert-Mauchly Computer Corporation, which was soon purchased by Remington Rand. The resulting machines, known as the Universal Automatic Computer or UNIVAC, became available after the 1950 Census and aided in its tabulation. The UNIVAC stored data in a new way, on magnetic tape, but information still had to be punched onto cards in order to be transferred to the tape, and this process had become a reverse salient, holding back the efficiency of the entire system [64].[2]

Even before the UNIVAC was delivered, the Census Bureau had partnered with the U.S. National Bureau of Standards to develop a new data capture technology that could rapidly scan the images on specially designed pages that had been microfilmed and encode them on magnetic tape, bypassing punched cards altogether. The Census Bureau first used the Film Optical Sensing Device for Input to Computers (FOSDIC) for tabulation of the 1960 Census. In that year, enumerators went door-to-door, filling in bubble sheets with the information for each person at each residence. Beginning in 1970, households would fill in the bubble sheets themselves and mail them back to the Census Bureau [64].

THE PRESERVATION OF MANUSCRIPT CENSUS RETURNS HAS ALSO MADE IT POSSIBLE TO EXTRACT ADDITIONAL DATA THAT WERE NEVER BEFORE CAPTURED DIGITALLY.

FOSDIC remained in use through the 1990 Census. It only read in "data" items, however, bubbled-in answers to multiple-choice questions. Handwritten answers to open-ended questions, such as individual names and places of birth, still needed to be keyed in by hand if they were to be captured digitally. In the 1990s, the Census Bureau privatized data capture, working with contractors who introduced optical character recognition (OCR) to encode hand-written responses, beginning with the 2000 Census [64]. The 2020 Census was the first to allow for an online response, in which household members directly keyed their information.

What is perhaps most remarkable about this entire history is that the Census Bureau preserved the original manuscript census returns, even after the information in them had been transferred to punched cards and magnetic tape. With the exception of the 1890 Census, which was inadvertently destroyed in a fire, manuscript returns for every U.S. Census are still kept in the National Archives, either on microfilm or, more recently, as scanned images. These records have been a boon to researchers, particularly in the field of historical demography, which emerged in the decades following World War II. In the article "The Present of the Past: A Sociotechnical Framework for Understanding the Availability of Research Materials," Rebecca Emigh and Johanna Hernández-Pérez consider the factors that lead to the preservation of historical documents and traces of the past and facilitate the use of preserved documents and traces in historical demography. Among the important arguments made by Emigh and Hernández-Pérez is that the preservation of documents that record information about the human population allows analysis and understanding that the original creators of the documents never foresaw. This is particularly true in historical demography, where the documents and traces used by today's researchers were almost invariably collected for other purposes, such as democratic self-governance (census records) or eternal salvation (ecclesiastical records). The use by researchers of what they call both "documents" and "traces" has permitted the emergence and development of a rich field of retrospective research, especially over the past 75 years.

The preservation of manuscript census returns has also made it possible to extract additional data that were never before captured digitally. In "The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Census Infrastructure," Katie Genadek and J. Trent Alexander describe a project aimed at using OCR technology to capture information from the U.S. Censuses of 1950 to 1990 that was not previously keyed in or captured by FOSDIC. Genadek and Alexander report on a research program that aims to link the records of individuals who were enumerated in multiple censuses, tracking them from one decade to another, a project that will facilitate sophisticated analysis of social and geographical mobility, family change, and more. Such a program (which we discuss further later), needs to have the record of every person named in the census fully digitized, something not remotely considered prior to the 2000 census. As Genadek and Alexander explain, the availability of original census documents and the rapid transformation of information technology has begun to make this goal achievable.

COMPLEX DESCRIPTIONS AND ANALYSES

Prior to the introduction of UNIVAC, the Census Bureau distributed data through published reports, which

[2]For the concept of the reverse salient in science and technology studies, see MacKenzie and Wajcman [43].

spanned more and more volumes with each census. These reports presented tables of data aggregated by administrative units: cities, counties, states, and the country as a whole. They provided numbers rather than analyses; users who wanted to analyze the data did so primarily by hand or with calculators.

When the Census Bureau began using magnetic tapes for the tabulation of data from the 1960 Census, data users outside the Census Bureau requested access to them. The Census Bureau granted many of these requests, but the data tapes were not particularly user-friendly, nor did they come with sufficient documentation [75]. With the 1970 Census, however, the Census Bureau aimed to produce a data product that could be broadly distributed. Since then, it has made tabulated data widely available in machine-readable form, using new media to do so as they have become available. Today users can download tabulated census data at a variety of scales from the Census Bureau's website, or access them using an application programming interface.

Initial users of Census data tapes—those for whom the tapes were first produced—were primarily large public agencies and social scientists working in universities. The anecdote by Barbara Anderson, "The Effects of Increases in Computing Power on Demographic Analysis Over the Last 50 Years," describes the excitement and frustrations experienced by social scientists using computers to analyze demographic data on mainframes during the 1970s. Her descriptions of work in the computing environments of the mainframe and early personal computer era give life to the opportunities and challenges of work at that time. Her frank characterization of her work and that of colleagues who sometimes misunderstood their data demonstrates the issues faced by early adopters of new technology, prior to the development of user-friendly software and standard documentation, something described in this issue and later in our introduction.

Once the Census Bureau made summary data tapes available to wider audiences, new types of users emerged. The business community made particular use of these data, engendering the rise of demographics as a tool for identifying, constructing, and advertising to narrowly-defined markets [67]. Magnetic tapes facilitated the release of data at ever-smaller geographic units, such as zip codes and census tracts.[3] "When the New Magic Was New: The Claritas Corporation and the Clustering of America" by Fenwick McKelvey describes the way in which an innovative marketing company, the Claritas Corporation, applied an old statistical method—factor analysis—to these new small-area census data in order to identify every zip code in the United States with one of forty markets, facilitating direct mail advertisement. The products sold by Claritas combined the tabulated data produced by the Census Bureau with analytic approaches made possible by the introduction of a new generation of digital computers, new media for digital storage, and new programming languages that could examine, process, and cluster large quantities of data, building on the UNIVAC's innovations.

WHEN THE CENSUS BUREAU BEGAN USING MAGNETIC TAPES FOR THE TABULATION OF DATA FROM THE 1960 CENSUS, DATA USERS OUTSIDE THE CENSUS BUREAU REQUESTED ACCESS TO THEM.

Even before the U.S. Government's approach to the collection and analysis of population data expanded to the business realm, it traveled overseas. As the United States began to extend its reach into the Atlantic and Pacific Oceans around the turn of the twentieth century, the federal government implemented censuses in the new territories of Puerto Rico, Cuba, and the Philippines [19]. During the Vietnam War, the U.S. government adapted demographic data collection and analysis—facilitated by the use of electronic computers and magnetic tape—to the needs of counterinsurgency. In "Computing Counterinsurgency: The Hamlet Evaluation System (HES) and Databasing During the Vietnam War," Moritz Feichtinger examines "databasing" as a governmental activity and a means of waging warfare. Building on a combination of census-taking and the sort of analytics used by Claritas, the government attempted—not always successfully—to bring data-based analytics to the management of a war.

The increase in computing speed and the availability of new input and output devices opened up other doors as well. Feichtinger discusses the mapping of data collected in Vietnam, a precursor to the development in the 1970s and 1980s of sophisticated Geographic Information Systems capable of organizing data spatially and allowing their analysis and display. These systems offered innovative visualization tools

[3]The Census Bureau began to publish tract-level data in the 1940s, but it was not until 2000 that census tracts were defined for the entire country (https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf).

that made the display of spatial information quicker and more intuitive, but they also did more. By managing information that encoded the substance of the data point (attributes of a person, business, or place, for example), along with its location, new forms of statistical analysis became possible that went beyond visualization and considered the role of space and proximity in demographic processes [23].

ANALYZING INDIVIDUAL-LEVEL DATA

The increasing availability of high-capacity computing hardware and software in the late 1960s and 1970s opened a number of doors for demographic researchers. With those resources, it became possible to move from the analysis of already-tabulated data (as practiced by Claritas) to analysis of the underlying individual-level data, and the more effective analysis of data that originated from individually-conducted surveys.

Social scientists began using surveys as a research method in the 1930s. Survey methodology developed rapidly in the work of the U.S. Department of Agriculture during World War II, and the use of surveys in the social sciences expanded dramatically in the decades following the war [14], [35]. Demographers were early adopters of survey methods, using survey research to predict and influence childbearing decisions, first in the United States and then worldwide [45]. Analyzing survey or census data at the individual level allowed social scientists to identify statistical relationships that were not subject to the ecological fallacy, the attribution of group-level correlations to individuals [57], [72].

The statistical methods leveraged in these analyses—correlation, regression, and chi-squared tests, among others—were not new; they had been developed around the turn of the twentieth century [53], [70]. Doing them by hand, however, was cumbersome, all the more so as the number of observations increased. Aggregating data limited the number of observations, facilitating statistical analysis. With the availability of high-capacity computers, however, it became feasible to perform complex statistical analyses on individual-level data, even with hundreds or thousands of observations. Private companies also participated in the computer-assisted analysis of individual-level survey data, with the Simulmatics Corporation selling its analyses of voting behavior to the Kennedy campaign in preparation for the 1960 election [40].

Early adopters of these new technologies became masters at programming using FORTRAN or other languages, but not everyone had the time or skill needed for that work, and there was a fundamental replicability problem: it was not always clear that the statistics generated by those individual efforts were accurate or comparable. To meet this need for scientific, biomedical, engineering, and business researchers, universities in the mid-1960s began to develop "packaged" programs designed for statistical analysis. UCLA (BMDP), Princeton (P-Stat), and the University of Michigan (Osiris) all made significant investments. Private corporations soon joined the scene, offering more sophisticated and commercially viable products, first SPSS (1971) and SAS (1971), later S (1976), STATA (1985), and most recently R (1995), an open-source and noncommercial implementation of S. These systems offered more user-friendly environments than FORTRAN, simple methods for common statistical analyses, and the replicability needed by academic reviewers. If an author reported having generated a specific statistic from one of these packages, reviewers understood how it was computed and could compare it to other published statistics. Barbara Anderson describes how the introduction of these statistics packages increased the autonomy of demographers, reducing their reliance on programmers and statisticians, though it also increased their conformity to the types of analysis facilitated by the packages. More recently, the development of machine-readable and machine-actionable documentation, such as that produced by the Data Documentation Initiative, has built on technological developments that further improved the capacity of researchers to use large quantities of data.[4]

Even with these technologies in place, however, there was still a barrier to analyzing U.S. Census data at the individual level: census returns are protected by a 72-year privacy wall (for example, manuscript returns for the 1950 U.S. Census of Population became publicly available in April, 2022). To meet the growing demand of researchers for individual-level data without violating privacy restrictions, the Census Bureau released a public-use microdata sample of the 1960 Census in 1963. This 0.1% sample excluded any potentially identifying information for the individuals included [66]. As the production and dissemination of public use microdata samples continued through the 1990 census, and as computing capabilities improved, historically-oriented researchers began to expand the time frame for which data were available. Teams at the Universities of Wisconsin, Washington, Pennsylvania, and Minnesota created samples of individual

[4]https://ddialliance.org

records, first for the 1940 and 1950 censuses (Wisconsin), then for the 1900 (Washington) and 1910 (Pennsylvania) censuses, and later for all earlier censuses back to 1850 (Minnesota). While these samples were initially small (later expanded to complete populations), they came to constitute a unique resource for the study of the development of the U.S. population, a valuable tool for demographers researching the present as well as the past.

THE DEVELOPMENT OF TECHNOLOGY FOR LINKING INDIVIDUAL CENSUS RECORDS ACROSS TIME AND TO OTHER ADMINISTRATIVE RECORDS IS PERHAPS ONE OF THE CLEAREST EXAMPLES OF THE EXIGENCIES OF HISTORICAL RESEARCH DRIVING THE APPLICATION OF TECHNOLOGICAL INNOVATION.

With this long series of census microdata increasingly available, historians at the University of Minnesota realized that longitudinal research would be greatly facilitated by a harmonized coding scheme that overcame the differences from decade to decade in how the census was conducted, recorded, and coded. The anecdote "Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience" by Diana Magnuson and Steven Ruggles describes the inception and development of the Integrated Public Use Microdata Series (IPUMS) project to meet this need. The IPUMS project provides a window into the ways that interest in and access to large-scale individual-level data resources benefited from the rapid transformation of computing and communication technology that took place in the 1980s and 1990s. Magnuson and Ruggles describe three important changes: a rapid improvement in the processing power of small computers; increased communication speed and capacity that extended high-speed computer communication throughout the research community and to individual homes and businesses; and the creation of new software and communication protocols (especially the World Wide Web) that facilitated computer communication on the new networks. IPUMS and the projects facilitated by it in the early 1990s were greatly enriched by the capacity to deliver data through an interactive system that operated over the Web, enabled by the capacity to manage very large quantities of data and share them over fast data networks. By the early 2020s harmonized U.S. data for censuses going back to 1850 (with much of it for the full population up to 1950) were available, along with an ever-increasing body of also-integrated international census data [65].

The capacity to study long-term demographic trends, using IPUMS data, has greatly expanded our understanding of historical U.S. population dynamics. Topics covered in recent publications include fertility [24], [26], [27], family structure [46], [59], [61], migration [1], [22], [30], racial segregation [15], [42], and census quality [25].

While the idea of harmonizing data across census years has born significant fruit, a second technology-enabled insight promises still more. This is the subject of Genadek and Alexander's contribution ("The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Census Infrastructure"), which considers the opportunity to go beyond data harmonization to create a dataset linking individuals across censuses. Recent work in this area shows the extraordinary promise of these efforts [3], [55], [60], [68].

Over the past 75 years a large volume of research, mostly in social history and historical demography, has shown the value of studies that rely on data about individuals linked across multiple documents, including censuses and other sources. Until recently, this research was limited by the work required to acquire (and digitize) the data, and then to manually link records together. For two generations of researchers, studies focused either on very small areas (a village or small town, or possibly a rural county) or on very small samples, often based on names that were rare or began with only a few letters of the alphabet. The best-known research, such as studies of social mobility [12], [38], [71], urban social change in Philadelphia [29], [34], the relationship between religion and fertility among the Mormon pioneers in Utah [6], [7], and the demographic history of France [8] and the Belgian city of Antwerp [54], were limited by the effort required.

Accomplishing the goal outlined by Genadek and Alexander requires innovation on two fronts. First, the names need to be digitized. Second, another process needs to figure out which John Smith in 1980 is the same John Smith in 1990. Advances in OCR have facilitated the first. For the second, historians and demographers have developed a variety of techniques to link records automatically or semi-automatically over the past fifty years. The best available methods currently use machine learning but are still far from perfect [63]. The development of technology for linking individual census records across time and to other administrative

records is perhaps one of the clearest examples of the exigencies of historical research driving the application of technological innovation.

Individuals can be linked across censuses because censuses (in theory at least) capture every individual living in the United States in years ending with zero. Demographers researching contemporary population dynamics use panel surveys to follow individuals over time, interviewing the same sample of people at regular (or irregular) intervals [32], [33], [36], [69]. These studies have taken advantage of and pushed developments in computer-assisted personal interviewing and computer-assisted telephone interviewing, both of which are also used in government surveys and market research. These techniques combine interviewing with data capture, eliminating the step of keying in data from a paper-based questionnaire or using OCR. When these technologies were first introduced in the 1980s, respondents worried that having their information captured by a computer made it less confidential [4]. Today, however, surveys are often designed on the assumption that respondents, now familiar with using computers for personal purposes in their own homes, feel more comfortable entering sensitive information on a screen than discussing it with an interviewer [58]. In the past decade, the near ubiquity of internet access has made it possible to reinterview the same people at much more frequent intervals, and demographers have experimented with novel modes of internet-based survey administration [5]. Demographic panel studies have also been at the forefront of data sharing, pioneering a variety of methods for archiving data and making them publicly available for secondary analysis.

Secondary analysis of individual data, particularly datasets that link information about individuals from multiple sources, raises important questions about the protection of the privacy and confidentiality of individuals and groups who are research subjects in demographic studies and large-scale data collections. On the one hand, census data and data collected using public funds should be a public resource. On the other hand, democracy and research ethics and integrity require the preservation of respondent confidentiality. Improvements in technology and the expanding business applications of data mining have dramatically increased the amount of data available about individuals and groups, the kinds of information about them that are known, and the ways that those information items can be linked. It is not difficult to imagine a data record about an individual emerging from the topics described in this issue that includes linked references to multiple census records over time, the spatial location of the individual involved, and their biological and genetic characteristics, which we discuss more below. This concern is not new and has led to multiple government-sponsored studies by the National Academies [48], [49]. Nor is it one that is ignored by Congress, which has legislated tight controls over Census and other federally-collected data, or by federal agencies, which have implemented those controls. The protection of the anonymity of Census data is a topic discussed by several of our authors, especially Magnuson and Ruggles, Genadek and Alexander, and Anderson. The creation of a network of Federal Statistical Research Data Centers, as mentioned by Anderson, is one mechanism that has protected data while facilitating research.[5] That is where the linked data that Genadek and Alexander describe will reside when the process is complete.

Maintaining confidentiality protection in the face of ever-increasing computational capacity and ever-larger volumes of data that might be linked together to reveal information that might put respondents at risk is a subject of continuing debate among experts in the fields of demographic research, computer science, and related fields. It is already customary for locational information to be perturbed when individual-level survey research is made publicly available in order to avoid respondent identification. For the 2020 census, even aggregate data will have noise injected in order to deter the identification of the individuals who constitute small-scale aggregates [62]. Such approaches promise to be a source of debate and research innovation in the future.

FUTURE DIRECTIONS: NEW TECHNOLOGIES AND THE RISE OF BIOLOGICAL AND GENETIC DEMOGRAPHY

Questions about privacy have become even more pressing with the emergence of biological and genetic demography, made possible by new computational innovations in the analysis of DNA and other biomarkers. Beginning in the 1990s, contemporary demographic studies began collecting biomarkers—indicators from blood and urine samples, anthropometric measures, and field tests such as cognition and grip strength—along with social information, giving rise to biodemography. The initial focus of biodemography was aging, a topic that received considerable research funding through the National Institute on Aging [51]. Biodemography's rise was accelerated by the fact that fundamentally biological characteristics could

[5]https://www.census.gov/about/adrm/fsrdc.html

be represented within existing datasets as a result of rapid improvements in the speed of computation and the development of new algorithms. The addition of biomarkers to such panel surveys as the Health and Retirement Study (HRS) and the National Longitudinal Study of Adolescent to Adult Health (Add Health) expanded the range of demographic research into biomedical territory [74]. It also generated new questions about how to store biological specimens and the information derived from them [20]. As was the case with manuscript census records, where preservation of paper forms on microfilm facilitated the later extraction of additional data, the storage of biological samples allows their reanalysis as technologies develop to extract different types of information.

The possibility of reanalyzing stored samples has been a particular boon to genetic demography, as the technology to identify more of an individual genome at lower costs has developed rapidly over the past fifteen years [73]. Today, single nucleotide polymorphism (SNP) arrays can genotype multiple individuals at hundreds of thousands or even millions of points along the genome. For each locus, a probe binds to specific segments of the DNA, producing signals that are converted algorithmically into inferences about each individual's genotype (for example AA or AT or TT) at each locus [39]. A genome-wide association study (GWAS) then assesses, for each locus identified by a SNP array, whether an individual's genotype is correlated with the outcome in question, which could be height, a disease state, or a socioeconomic measure like educational attainment or income [10]. The regression coefficients can then be multiplied across a person's genome and summed to produce a polygenic score (PGS), which is often interpreted as an individual's genomic propensity to experience a particular outcome [41]. Long-running demographic panel studies began to collect DNA samples in the early 2000s, and have re-analyzed these samples as SNP array technology has improved and as the price has fallen. While SNP arrays provide researchers with potentially hundreds of thousands or millions of new variables, GWAS offers a kind of dimensionality reduction, making it possible to summarize all of those SNPs in a single PGS (for a specific outcome) that social scientists can include in their models just like any other variable [31]. Today, HRS, Add Health, the Wisconsin Longitudinal Study, and the Panel Study of Income Dynamics all make a variety of PGS available to researchers, along with all of the social and biometric data they collect.

Demographers have primarily used PGS to control for genetic variation in order to better understand the social world [13]. This type of research has shown (for example) that people living in unsafe neighborhoods are more likely to develop type 2 diabetes, even when controlling for their PGS for the disease [56]. Similarly, childhood socioeconomic status remains an important predictor of educational outcomes even when controlling for the PGS for educational attainment [52]. Over the past five years, however, it has become clear that PGS are not an accurate assessment of individual risk for medical or social outcomes. GWAS are conducted primarily on individuals with exclusively European genetic ancestry, and PGS explain very little of the variance in outcomes among people with non-European genetic ancestry. Their use in clinical and research settings (both biomedical and social) therefore threatens to perpetuate existing outcome disparities between differently-racialized groups [44]. Even among individuals with exclusively European ancestry, however, PGS capture only one source of genetic heterogeneity (SNPs), and they are irremediably confounded by environmental and social factors [9]. This is particularly true of PGS for social outcomes, such as educational attainment, or socially motivated behaviors, such as smoking [18]. Their use in social scientific research therefore threatens to mask rather than reveal the operation of the social world.

The problems with GWAS and PGS point to two unique and related challenges associated with using genomic data in demography, neither of which can readily be resolved by further technological innovation. Any potential genetic effects on social outcomes are tiny, and therefore require enormous samples to identify [11]. Samples that were considered large in social demography, such as HRS and Add Health, each with around 20,000 respondents, are not nearly large enough for a GWAS. GWAS for social outcomes therefore require pooling data, not just from multiple cohorts, but from multiple countries [50]. While some of these data sources, such as HRS and Add Health, may have been representative of some population (Americans over age 50 in the case of HRS and Americans who were in high school in the 1990s in the case of Add Health), the largest sources of genomic data, such as the U.K. Biobank, are not representative of any population [37]. Nor is the pooled sample. It is therefore unclear who is represented in GWAS and for whom PGS are most predictive [47]. This is the first challenge. The second challenge is that creating a nationally-representative sample of genomic data in the United States would require large-scale buy in, and possibly even the same kind of legal compulsion involved in census-making.[6] Yet genomic data are highly identifiable, and if not very well protected could make research participants

[6]For the concept of "census-making," see Curtis [16].

and their family members vulnerable to criminal prosecution and various new forms of discrimination, along with all of the existing forms.

These advances will continue to increase opportunities for research that will improve our knowledge of demographic processes and the human condition more generally, but we will also need to use existing expertise in the social sciences and humanities to ensure that emerging technologies do not harm research subjects, particularly members of groups who have historically sacrificed the most and benefited the least from research in the human sciences. Our contributors point the way to understanding both the history of these interconnected fields and ways that they will help us in the future.

ACKNOWLEDGMENTS

We are grateful to George Alter and David Hemmendinger for advice about this introduction. We received generous support for the workshop that led up to this special issue from the University of Colorado Population Center and the Institute of Behavioral Science at the University of Colorado Boulder. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award P2CHD066613. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

BIBLIOGRAPHY

[1] R. Abramitzky and L. Boustan, *Streets of Gold: America's Untold Story of Immigrant Success*. New York, NY, USA: Public Affairs, 2022.

[2] M. Anderson, *The American Census: A Social History*, 2nd ed. New Haven, CT, USA: Yale Univ. Press, 2015.

[3] M. Bailey, C. Cole, and C. Massey, "Simple strategies for improving inference with linked data: A case study of the 1850-1930 IPUMS linked representative historical samples," *Historical Methods, J. Quant. Interdiscipl. Hist.*, vol. 44, no. 2, pp. 80–93, 2020.

[4] R. Baker, "New technology in survey research: Computer-assisted personal interviewing," *Social Sci. Comput. Rev.*, vol. 10, no. 2, pp. 145–157, 1992.

[5] J. Barber, Y. Kusunoki, H. Gatny, and P. Schulz, "Participation in an intensive longitudinal study with weekly web surveys over 2.5 years," *J. Med. Internet Res.*, vol. 18, no. 6, 2016, Art. no. e105.

[6] L. L. Bean, G. Mineau, and D. L. Anderton, "Residence and religious effects on declining family size: An historical analysis of the Utah population," *Rev. Religious Res.*, vol. 25, no. 2, pp. 91–101, 1983.

[7] L. L. Bean, G. P. Mineau, and D. L. Anderton, *Fertility Change on the American Frontier: Adaptation and Innovation*. Berkeley, CA, USA: Univ. California Press, 1990.

[8] J. Bourdieu, L. Kesztenbaum, and G. Postel-Vinay, "L'enquête TRA, une matrice d'histoire," *Population*, vol. 69, no. 2, pp. 217–248, 2014.

[9] C. Burt, "Challenging the utility of polygenic scores for social science: Environmental confounding, downward causation, and unknown biology," *Behav. Brain Sci.*, vol. 13, pp. 1–36, 2022, doi: 10.1017/S0140525X22001145.

[10] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS Comput. Biol.*, vol. 8, no. 12, 2012, Art. no. e1002822.

[11] C. F. Chabris et al., "Most reported genetic associations with general intelligence are probably false positives," *Psychol. Sci.*, vol. 23, no. 11, pp. 1314–1323, 2012.

[12] H. P. Chudacoff, *Mobile Americans: Residential and Social Mobility in Omaha 1880-1920*. New York, NY, USA: Oxford Univ. Press, 1972.

[13] D. Conley and J. Fletcher, *The Genome Factor: What the Social Genomics Revolution Reveals About Ourselves, Our History and the Future*. Princeton, NJ, USA: Princeton Univ. Press, 2017.

[14] J. M. Converse, *Survey Research in the United States: Roots and Emergence, 1890-1960*. New York, NY, USA: Routledge, 2009.

[15] L. D. Cook, T. D. Logan, and J. M. Parman, "Racial segregation and southern lynching," *Social Sci. Hist.*, vol. 42, no. 4, pp. 635–675, 2018.

[16] B. Curtis, *The Politics of Population: State Formation, Statistics, and the Census Ofinin Canada, 1840-1875*. Toronto, ON, Canada: Univ. Toronto Press, 2022.

[17] A. Deutsch, "The first U.S. census of the insane (1840) and its use as pro-slavery propaganda," *Bull. Hist. Med.*, vol. 15, no. 5, pp. 469–482, 1944.

[18] B. W. Domingue, D. Conley, J. Fletcher, and J. D. Boardman, "Cohort effects in the genetic influence on smoking," *Behav. Genet.*, vol. 46, pp. 31–42, 2016.

[19] V. R. Dominguez, "When the enemy is unclear: US censuses and photographs of Cuba, Puerto Rico, and the Philippines from the beginning of the 20th century," *Comp. Amer. Stud. Int. J.*, vol. 5, no. 2, pp. 173–203, 2007.

[20] C. E. Finch, J. W. Vaupel, and K. Kinsella, Eds., *Cells and Surveys: Should Biological Measures Be Included in Social Science Research?* WA, DC, USA: Nat. Acad. Press, 2001.

[21] M. Foucault, *The History of Sexuality. vol. 1: An Introduction*. New York, NY, USA: Vintage, 1990 (Transl.: in Robert Hurley, Ed.).

[22] B. Gratton and E. Merchant, "Immigration, repatriation, deportation: The Mexican-origin population in the United States, 1920-1950," *Int. Migration Rev.*, vol. 47, no. 4, pp. 944–975, 2013.
[23] M. P. Gutmann, G. D. Deane, E. Merchant, and K. M. Sylvester, Eds., *Navigating Time and Space in Population Studies*. New York, NY, USA: Springer, 2011.
[24] J. D. Hacker, "Rethinking the 'early' decline of marital fertility in the United States," *Demography*, vol. 40, no. 4, pp. 605–620, 2003.
[25] J. D. Hacker, "New estimates of census coverage in the United States, 1850-1930," *Social Sci. Hist.*, vol. 37, no. 1, pp. 71–101, 2013.
[26] J. D. Hacker, "Ready, willing, and able? Impediments to the onset of marital fertility decline in the United States," *Demography*, vol. 53, no. 6, pp. 1657–1692, 2016.
[27] J. D. Hacker and E. Roberts, "Fertility decline in the United States, 1850-1930: New evidence from complete-count datasets," *Annales de Démographie Historique*, vol. 2, no. 138, pp. 143–177, 2019.
[28] I. Hacking, "Biopolitics and the avalanche of printed numbers," *Humanities Soc.*, vol. 5, no. 3/4, pp. 279–295, 1982.
[29] M. R. Haines, "Fertility and marriage in a nineteenth-century industrial city: Philadelphia, 1850-1880," *J. Econ. Hist.*, vol. 40, no. 1, pp. 151–158, 1980.
[30] P. K. Hall and S. Ruggles, "'Restless in the midst of their prosperity': New evidence on the internal migration of Americans, 1850-2000," *J. Amer. Hist.*, vol. 91, no. 3, pp. 829–846, 2004.
[31] K. P. Harden, "'Reports of my death were greatly exaggerated': Behavior genetics in the postgenomic era," *Annu. Rev. Psychol.*, vol. 72, pp. 37–60, 2021.
[32] K. M. Harris et al., "Cohort profile: The national longitudinal study of adolescent to adult health (Add health)," *Int. J. Epidemiol.*, vol. 48, no. 5, pp. 1415–1415k, 2019.
[33] P. Herd, D. Carr, and C. Roan, "Cohort profile: Wisconsin longitudinal study (WLS)," *Int. J. Epidemiol.*, vol. 43, no. 1, pp. 34–41, 2014.
[34] T. Hershberg, "The Philadelphia social history project: An introduction," *Historical Methods Newslett.*, vol. 9, no. 2/3, pp. 43–58, 1976.
[35] S. E. Igo, *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Cambridge, MA, USA: Harvard Univ. Press, 2008.
[36] D. S. Johnson, K. A. McGonagle, V. A. Freedman, and N. Sastry, "Fifty years of the panel study of income dynamics: Past, present, and future," *Ann. Amer. Acad. Political Social Sci.*, vol. 680, no. 1, pp. 9–28, 2018.
[37] K. M. Keyes and D. Westreich, "U.K. biobank, big data, and the consequences of non-representativeness," *Lancet*, vol. 393, no. 10178, 2019, Art. no. P1297.
[38] P. R. Knights, *The Plain People of Boston, 1830-1860: A Study in City Growth*. New York, NY, USA: Oxford Univ. Press, 1971.
[39] T. LaFramboise, "Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances," *Nucleic Acid Res.*, vol. 37, no. 13, pp. 4181–4193, 2009.
[40] J. Lepore, *If Then: How the Simulmatics Corporation Invented the Future*. New York, NY, USA: Liveright, 2020.
[41] C. M. Lewis and E. Vassos, "Polygenic risk scores: From research tools to clinical instruments," *Genome Med.*, vol. 12, 2020, Art. no. 44.
[42] T. D. Logan and J. M. Parman, "The national rise in residential segregation," *J. Econ. Hist.*, vol. 77, no. 1, pp. 127–170, 2017.
[43] D. MacKenzie and J. Wajcman, "Introductory essay: The social shaping of technology," in *The Social Shaping of Technology*, D. MacKenzie and J. Wajcman, Eds., 2nd ed. Buckingham, U.K.: Open Univ. Press, 1999.
[44] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, "Clinical use of current polygenic risk scores may exacerbate health disparities," *Nature Genet.*, vol. 51, no. 4, pp. 584–591, 2019.
[45] E. K. Merchant, *Building the Population Bomb*. New York, NY, USA: Oxford Univ. Press, 2021.
[46] E. Merchant, B. Gratton, and M. P. Gutmann, "A sudden transition: Household changes for middle-aged U.S. women in the twentieth century," *Popul. Res. Policy Rev.*, vol. 31, no. 5, pp. 703–726, 2012.
[47] H. Mostafavi, A. Harpak, I. Agarwal, D. Conley, J. K. Pritchard, and M. Przeworski, "Variable prediction accuracy of polygenic scores within an ancestry group," *eLife*, vol. 9, 2020, Art. no. e48376.
[48] National Research Council, *Putting People on the Map: Protecting Confidentiality With Linked Social-Spatial Data*, M. P. Gutmann and P. Stern, Eds., WA, DC, USA: Nat. Acad. Press, 2007.
[49] National Research Council, *Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens*, R. M. Biodata, M. Hauser, R. P. Weinstein, and B. Cohen, Eds., WA, DC, USA: Nat. Acad. Press, 2010.
[50] A. Okbay et al., "Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals," *Nature Genet.*, vol. 54, pp. 437–449, 2022.

[51] J. Olshansky, "On the biodemography of aging: A review essay," *Popul. Develop. Rev.*, vol. 24, no. 2, pp. 381–393, 1998.
[52] N. W. Papageorge and K. Thom, "Genes, education, and labor market outcomes: Evidence from the health and retirement study," *J. Eur. Econ. Assoc.*, vol. 18, no. 3, pp. 1351–1399, 2020.
[53] T. Porter, *The Rise of Statistical Thinking, 1820-1900*. Princeton, NJ, USA: Princeton Univ. Press, 1986.
[54] P. Puschmann, H. Matsuo, and K. Matthijs, "Historical life courses and family reconstitutions. The scientific impact of the Antwerp COR*-Database," *Historical Life Course Stud.*, vol. 12, pp. 260–278, 2022.
[55] E. Rauscher, "Does educational equality increase mobility? Exploiting nineteenth-century U.S. compulsory schooling laws," *Amer. J. Sociol.*, vol. 121, no. 6, pp. 1697–1761, 2016.
[56] J. W. Robinette, J. D. Boardman, and E. M. Crimmins, "Differential vulnerability to neighbourhood disorder: A gene x environment interaction study," *J. Epidemiol. Community Health*, vol. 73, pp. 388–392, 2019.
[57] W. S. Robinson, "Ecological correlations and the behavior of individuals," *Amer. Sociol. Rev.*, vol. 15, pp. 351–357, 1950.
[58] S. S. Rostosky, M. D. Regnerus, and M. L. C. Wright, "Coital debut: The role of religiosity and sex attitudes in the add health survey," *J. Sex Res.*, vol. 40, no. 4, pp. 358–367, 2003.
[59] S. Ruggles, "The decline of intergenerational coresidence in the United States, 1850 to 2000," *Amer. Sociol. Rev.*, vol. 72, no. 6, pp. 964–989, 2007.
[60] S. Ruggles, "Intergenerational coresidence and family transitions in the United States, 1850-1880," *J. Marriage Fam.*, vol. 73, no. 1, pp. 136–148, 2011.
[61] S. Ruggles, "Patriarchy, power, and pay: The transformation of American families, 1800-2015," *Demography*, vol. 52, no. 6, pp. 1797–1823, 2015.
[62] S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder, "Differential privacy and census data: Implications for social and economic research," *AEA Papers Proc.*, vol. 109, pp. 403–408, 2019.
[63] S. Ruggles, C. Fitch, and E. Roberts, "Historical census record linkage," *Annu. Rev. Sociol.*, vol. 44, pp. 19–37, 2018.
[64] S. Ruggles and D. L. Magnuson, "Census technology, politics, and institutional change, 1790-2020," *J. Amer. Hist.*, vol. 107, no. 1, pp. 19–51, 2020.
[65] S. Ruggles, R. McCaa, M. Sobek, and L. Cleveland, "The IPUMS collaboration: Integrating and disseminating the world's population microdata," *J. Demographic Econ.*, vol. 81, pp. 203–216, 2015.
[66] S. Ruggles and R. R. Menard, "The Minnesota historical census projects," *Historical Methods, J. Quant. Interdiscipl. Hist.*, vol. 28, no. 1, pp. 6–10, 1995.
[67] C. Russell, "The business of demographics," *Popul. Bull.*, vol. 39, no. 3, pp. 1–40, 1984.
[68] A. Saperstein and A. Gullickson, "A 'mulatto escape hatch' in the United States? Examining evidence of racial and social mobility during the Jim Crow era," *Demography*, vol. 50, no. 5, pp. 1921–1942, 2013.
[69] A. Sonnega, J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. R. Phillips, and D. R. Weir, "Cohort profile: The health and retirement study (HRS)," *Int. J. Epidemiol.*, vol. 43, no. 2, pp. 576–585, 2014.
[70] S. Stigler, *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA, USA: Harvard Univ. Press, 1999.
[71] S. Thernstrom, *The Other Bostonians: Poverty and Progress in the Americaninin Metropolis, 1880-1970*. Cambridge, MA, USA: Harvard Univ. Press, 1976.
[72] E. L. Thorndike, "On the fallacy of imputing the correlation found for groups to the individuals or smaller groups containing them," *Amer. J. Psychol.*, vol. 51, pp. 122–124, 1939.
[73] E. B. Ware and J. D. Faul, "Genomic data measures and methods: A primer for social scientists," in *Handbook of Aging and the Social Sciences*, K. F. Ferraro and D. Carr, Eds., 9th ed. Cambridge, MA, USA: Academic, pp. 49–62, 2021.
[74] M. Weinstein, J. W. Vaupel, and K. W. Wachter, Eds., *Biosocial Surveys*. WA, DC, USA: Nat. Acad. Press, 2008.
[75] P. T. Zeisset, "Making decennial census data available," *Govern. Inf. Quart.*, vol. 2, no. 4, pp. 419–431, 1985.