

AFC ASSOCIATION FRANÇAISE
DE CLIMÉTRIE

WORKING PAPERS

Nr. 8, 2019

Measuring Success:
Clio and the Value of Database Creation

Claude Diebolt, Michael Hauptert

Measuring Success: Clio and the Value of Database Creation

By

Claude DIEBOLT

CNRS, University of Strasbourg: cdiebolt@unistra.fr

and

Michael J. HAUPERT

University of Wisconsin-La Crosse: mhaupt@uwlax.edu

Draft (July 15, 2019)

Abstract: In a recent article Stefano Fenoaltea (2018) bemoaned the loss of respect and focus on the importance of creating databases, or “measurement” as he referred to it. Cliometrics has made and continues to make valuable contributions not just to the field of economic history, but economics in general. In particular, we focus on the contribution of cliometrics to the creation of datasets. We highlight several important cases in both the past and present, of recognized important contributions of new datasets to the economics discipline. We argue that Clio has continually focused on, and valued, the creation of new data sets and the clever and novel ways they have been exploited to further the frontiers of knowledge, and that these efforts are both appreciated and recognized.

Keywords: Cliometrics, Databases, Economic history, Measurement.

JEL Codes: A10, A12, B41, C81, C82, N00, N01.

Introduction

The oft unheralded but important role of data set creation is one of the pillars of the cliometric movement. It is less sexy than its brethren, economic theory and econometric techniques, but it has grown in prominence and value as a result of the computer revolution, as has the ability to conduct cutting edge econometric tests on said data. Simply put, without the data, there would be no theory to empirically test.

Unheralded, however, means neither unappreciated nor unrecognized. And with the growing impact of computing power has grown the creation, recognition, availability, and appreciation for the cliometric contribution of data set creation to the discipline of economics.

Data set creation ranges from the heroic, labor intensive, gathering of data from dusty archives, to the downloading of already digitized data. In either case, useful data must be organized into a format that is easily accessible and useable by other scholars, and it must have been cleansed of errors so that future users can be confident of its veracity and validity. But to be truly useful, it really needs to be widely available in a format useful to not just a single researcher, but to a wide range of scholars, who can use the data to test a variety of hypotheses, with a variety of tools, from a wide spectrum of vantage points. And of course, to test and confirm prior results. New additions to databases can provide particularly insightful results that may change our understanding of the past in substantial ways.

The first difference between dataset creation today and the early days of the cliometric movement is the method of gathering it. A half century ago, when cliometrics was the “new” economic history, there was no internet, there were no digital cameras or laptop computers, and hard copies of books, articles, and monographs were the only way to share data. It was more expensive to gather data from archives, a process that required first traveling to the archives, where data had to be transcribed by hand, then collated and organized, all at great expense. If one had access to a mainframe computer, punch cards needed to be coded for the final step of computerized data analysis. The data gathering alone could take weeks, and the time lag between the gathering process and final publication of the dataset could be years. The process was long and labor intensive, but necessary. And the efforts were applauded, because the results could be surprising and spectacular, not to mention pathbreaking and controversial. It is no wonder these datasets (on slavery and railroads, for example) are so famous.

Consider the database created by Robert Fogel and Stanley Engerman that they used as the basis of their body of work on slavery. They originally built upon the Parker-Gallman sample of slave farms, and extended that database by collecting and coding data from 20,000 northern farms.¹ The growth of their database expanded the issues they considered, and spawned an entire literature on the economics of the American slavery institution.²

Assembling the Parker-Gallman sample, and subsequent additions to it, were extraordinary feats of perseverance in the pre-digital, pre-laptop computer days, and as a result, tended to be glorified and held up as examples of outstanding effort and service to the discipline. Today, with digitization, electronic cameras, and computing technology that can read, sort, and analyze data on a scale unimaginable a generation ago, the creation of databases does not draw the attention that it once did. The task is less labor intensive, takes less time, and is done more frequently. The decreasing cost of gathering data has led to an increase in the production of datasets. The task is, however, no less important than it ever was.

Another obvious evolution of datasets is the ability to share them. No longer is a hard copy of data necessary, nor in most cases even desired. This alone makes a dataset more valuable, because once created, downstream users no longer have to spend time or effort to enter it into their own database. Digital databases can easily and cheaply be assimilated into future research. This only increases the value of previously created datasets, and lowers the cost of further research on the topic and verification of the original research findings.

The lower cost of gathering data has meant that economic historians can examine issues in new ways by creating novel datasets. This can be especially important when more traditional data are unavailable. For example, Spoerer, et al (2019) compiled a database of 16,000 bells from the German bell classification scheme of WWII. Bells from churches in Germany, Austria, and parts of Poland, Czechoslovakia, and Belgium were classified by their casting date and artistic or musical importance as a way of determining which should be preserved and which could be melted down for the war effort. The authors use the data to suggest that the geographical distribution of church bell casting is a useful proxy for comparative regional economic growth from the 13th to the 18th century, a period of time for which GDP data are scarce, unreliable, or nonexistent.³

Another example of unique datasets being put to the test can be found in Giddings and Hauptert (2019) and Hauptert and Murray (2012), who compiled a database of 10,000 salary observations of professional baseball players. These were matched to individual output metrics to determine the cause of salary variations across time, gender, and race. Vincent Geloso (2019) used a novel dataset of prices and wages collected from the account books of religious congregations in Quebec to measure colonial-era living standards. And Bailey, et al (2018) created a comprehensive dataset on private, non-corporate banks in antebellum Michigan to examine the business of private banking and exchange brokering.

Gust and Baten (2019) constructed a data set of regicides in South Asia beginning in 900 CE. They use regicide as a proxy for homicide in both Europe and South Asia in an effort to explain divergent long run growth patterns. Keywood and Baten (2017) use the relative number of kings for whom a birth year is known as a proxy for human capital in medieval Europe. They find that “ruler numeracy” is positively correlated with other estimates of education levels for the medieval and modern periods.

Big picture questions can also be approached in new ways. Research on the Industrial Revolution, which has been going on for over a century, can still offer us new insights gleaned from new data. Kevin O’Rourke and Jeff Williamson (2005) exploited recently collected data on relative factor price trends over the very long run to argue that the opening of the European economy to international trade was instrumental in growth during the Industrial Revolution. Mills (2008) used new price and interest rate data and modern time series econometrics to reassess the relationship between interest rates, prices and inflation in Britain over two and a half centuries, from 1750 to 2006. And Nuvolari et al (2011) examined the diffusion of steam technology across eighteenth century Britain using new estimates of the pace and extent of steam engine implementation to bring new insights to the British Industrial Revolution.

Allen (2016) built on the work of Lindert and Williamson (1982, 1983), using newly found information to update English social tables. They used new wage and occupation data to improve the tables, increasing their reliability for their own and future work of others on the Industrial Revolution, social structure, and income inequality.

The use of newly created datasets is common in economic history research. A survey of recent articles published in four leading economic history journals shows the reliance on measurement or exploitation of newly discovered or created databases for new research. Looking at the 2018 volume of the *Journal of Economic History*, *Explorations in Economic History*, *Cliometrica*, and the *Economic History Review*, 58% of articles relied on newly created or discovered data for their analysis.

Computing power has opened other venues for new data. Optical character recognition and topic models, which are statistical algorithms that automatically infer themes from large collections of texts, have made possible the creation of new qualitative databases.⁴ The ongoing effort of archives and libraries to digitize books, articles, and archival material makes these new tools ever more powerful. It also preserves more data and makes it accessible to a wider audience of scholars.

1. Cliometrics

New Economic History, or Cliometrics, represents a move from the historical, descriptive approach of *describing* a historical event, toward the use of economic theory to *analyze* an event. The first practitioners of the art of cliometrics “proposed that economic historians use the techniques and insights of modern economic theory to frame the questions asked of history, to influence the hypotheses advanced about the past, and to suggest the nature and type of data to be collected from the archives.”⁵ Further, they advocated for the rigorous testing of any hypotheses advanced against the alternatives, particularly those found in the “old” economic history. This required the collection of data and its analysis using econometric techniques.

The seeds of the cliometric movement already existed after WWII. With the American economy booming, economists gained cachet. Economics, with its rigorous models tested from an abundance of numerical data by use of advanced, mathematically expressed formulae came to be regarded as the paradigm of the social sciences. At the same time as this increasingly technical focus, economists were increasingly interested in the determinants of economic growth and what they saw as the widening gap between developed and underdeveloped regions of the world. They saw the study of economic history as a source of insight into the issues of economic growth and economic development, and the new quantitative methods as the ideal tools for analysis.

Arguing against those who cliometricians would later label “old” economic historians, Simon Kuznets claimed that little would be gained from a study of the past unless it was systematic and quantitative. In his view, this was the only way to weigh the relative effects of factors and events. One reason for the scarcity of quantitative work in economic history was due to the extraordinary effort necessary before the computer to sift and classify quantitative information, and the relatively recent development of statistical theory and techniques capable of handling these problems.

The reception of the “new” economic history was chilly by some due to its perceived threat to traditional historical methods, but warmly welcomed by others for the possibilities it promised. Jonathan Hughes and Stan Reiter (1958) compared the computational effort it took them to analyze their steamship data to that of Newmarch (1857), who compiled more than 13,000 individual pieces of information and then performed a mere three arithmetical calculations, but all by hand. His efforts represented a lifetime of work, while the steamship

paper was but one of many “big data” (not to be confused with “Big Data,” which counts observations in the millions) projects cliometricians could explore with the power of new techniques and technology.⁶ The steamship study had a total of nearly twice as many data points as the Newmarch data set, but after coding punch cards, the computer then did all of the computational work.

The “new” economic history can be dated to the 1957 joint meeting of the EHA and the Conference on Research in Income and Wealth (under the guidance of the NBER). In particular, two joint papers by Alfred Conrad and John Meyer (1957 and 1958) constituted the manifesto for the new era. The first paper, on methodology, explained what scientific method was really all about and how it applied to economic historians. The second paper is one of the most influential in the evolution of economic history. It added enormous force to the methodological prescription by claiming to follow it in an analysis of the profitability of slavery on the eve of the Civil War. The analytical method, the data, the economic and accounting framework, and the choice of slavery as a subject, were to have vast consequences for the next generation of economic historians.

Kuznets may have inspired the cliometric movement, but it was Robert Fogel who reunified economics and history. He used the latest techniques of modern economics and gathered reams of historical data to reinterpret American economic growth in sectors as diverse as railroads, slavery, and nutrition. Rather than conjecture about the causes of growth, he carefully measured them. He pioneered the use of large-scale cross-sectional and longitudinal data sets harvested from original sources to examine policy issues.

Fogel (1964b) highlighted the changes in economic history that justified its being “new.” It was not a change in subject, economic historians still remained interested in the description and explanation of economic growth. It was the approach to measurement and theory that was new. Economic history always had a quantitative dimension. But much of the past work had been limited to the simple organization of data contained in government and business records. While continuing this pursuit, the new economic history placed its primary emphasis on reconstructing measurements and organizing primary data in a manner allowing them to obtain measurements that were never before possible. It thus followed that the most critical issue in the work of the new economic historians was the logical and empirical validity of the theories on which their measurements were based.

Cliometricians have contributed to the development of both economics and history by combining theory with quantitative methods, constructing and revising databases, and adding the variable of time to traditional economic theories. This has made it possible to question and reassess earlier findings, expanding the frontier of our knowledge of the past and its ability to portend the future. The use of history as a crucible to examine economic theory has deepened our knowledge of how, why and when economic growth and development occurs.

The main achievements of cliometricians have been to slowly but surely establish a solid set of economic analyses of historical evolution by means of measurement and theory. Nothing can now replace rigorous statistical and econometric analysis based on systematically ordered data.

Cliometrics has had a profound impact on economic history. The recent articles published in *Cliometrica* serve as an excellent sample of what is new in economic history research. They exploit new data and databases, with newly developed methods and with newly developed

hypotheses, models and theories in economics, history and statistics. Cliometrics is responsible for transforming the discipline from a primarily narrative to a mathematical approach. This transformation has combined theory with quantitative methods, new and revised databases, and innovative techniques, expanding our knowledge of the process of economic growth.

2. The major contributions of cliometrics

Cliometricians make use of the whole gamut of economic theory and statistical models, and the measurements they have obtained have yielded considerably more precise information than previously available. Perhaps the most famous example of this is Fogel's railroad studies (1962, 1964a), about which we will have more to say. Previously we have argued that the contributions of cliometrics can be placed into four categories: new techniques, new approaches, revisions of previously held beliefs, and new data sets.⁷ Because data is at the heart of cliometrics, it is a critical factor in each of these categories.

2.1. New techniques

Technique is what likely first comes to mind when one hears the term cliometrics. Certainly, the advancement of econometric theory and computing power have contributed greatly to the techniques used by cliometricians. Here we focus on the importance of technique in advancing the importance of data.

Technique goes beyond the latest advances of mathematical sophistication. One of the earliest techniques used by cliometricians was the counterfactual, made famous (but not created) by Robert Fogel's work on the railroads. The counterfactual is still a useful tool. Vasta, *et al* (2017) provide a recent implementation of it. They combine it with a large data set of more than 300,000 directors of Italian joint-stock companies. Their counterfactual models what would have happened to the Italian corporate network in the two decades before WWII had there been no German-type universal banks.

Among the newer techniques popularized by cliometricians are age heaping models and the use of church book registries. Both of these apply new techniques to new databases. The former was necessary in order to best exploit the latter, providing an example of how technique and dataset construction can work hand in hand. Church book registries have been used to study a wide range of demographic issues, none bigger than the question of why some countries are rich and others are poor.⁸

Age heaping can be applied to approximate the basic numerical skills and hence basic education of a population, and its impact on a variety of variables, including the impact of numeracy on long-run growth (Acemoglu, *et al* 2001, 2002), the role of religion in human capital formation (Becker and Woessmann 2009), gender inequalities (De Moor and Van Zanden 2010, Manzel and Baten 2009), and labor market outcomes (Charette and Meng 1998). Tollnek and Baten (2016) provide an exhaustive overview of age-heaping models and their applications. Age heaping is a useful proxy for human capital when traditional measures like literacy rates and level of education are not available. The technique cannot be used without sufficient data to analyze, but it is the technique that often garners more attention than the dataset itself.

2.2. New approaches

Cliometrics has spawned entirely new approaches to the study of economics. The most prominent were those pioneered by its Nobel laureates. Institutional economics, promoted by Douglass North, grew throughout the 1980s, spreading across disciplines from economics to anthropology, law, management, political science, psychology, sociology, and cognitive science. Anthropometrics, which counts Robert Fogel among its earliest practitioners, is another example.

Anthropometrics is the study of patterns in human body size over time. The field has its roots in the natural sciences but came into vogue as a field of study in the social sciences in the 1970s. Cliometricians originally used it as a means of measuring changes in the standard of living, using human heights as a measure of net nutrition, which in turn proxies standard of living. They have also used anthropometrics to contribute to research in mortality trends (Fogel 1986, Floud and Harris 1997), slavery (Engerman 1976, Steckel 1979, Margo and Steckel 1982), and the outcomes of industrialization and economic development (Floud and Wachter 1982, Steckel and Floud, 1997, Haines 2004). The genesis of much of this research in the United States was an NBER study on American and European mortality trends coordinated by Robert Fogel in the 1980s. Since then the scope of the field has grown to include countries around the world. These studies were based on datasets created from plantation records, slave manifests, probate records, military records, the 19th century national growth study, epidemiological studies, newspaper ads, and skeletal remains, among other sources.⁹

Demography has also drawn the interest of numerous cliometricians, in large part due to the ability to create and analyze large databases. Federal and state censuses have long been available as sources of big data, but only relatively recently has technology made them accessible for serious research. Joe Ferrie has long been a leader in this field. One of his earliest contributions was a sample of males linked from Federal censuses of 1850 to 1870 (Ferrie 1996). This has created longitudinal datasets allowing scholars to track the economic and geographic mobility of individuals and families over time. When combined with 20th century data compiled from the National Longitudinal Surveys (NLS) and the Panel Study of Income Dynamics (PSID), Ferrie's data set provides a historical benchmark, and the linked samples provide information on occupation, wealth, family structure, and location for individuals across time.

The construction of longitudinal population databases is not confined to the United States. Pfister and Fertig (2010) created an aggregative reconstruction of the population of Germany from the sixteenth to the mid-eighteenth century. Their estimates of population size and an annual series of crude birth, marriage and death rates were built on partial censuses, parish registers, and the protostatistical material on population size and vital events that states began to collect in the mid-18th century. Without modern computing power, it would have taken an army of scholars a lifetime just to compile the data, let alone make use of the results. Without cliometrics, the compiled data would lay fallow.

2.3. Revisions

The revision of misunderstandings in history is both important and necessary. Developing a clear understanding of the causes of economic growth is among the most important tasks of economic historians. Cliometrics has overturned some accepted wisdoms and in the process created its fair share of controversy. However, they have also pushed forward the frontier of our understanding of economic growth and development.

Among the notable “revisions” made by the first generation of cliometricians were the findings of Conrad and Meyer (1958), Yasuba (1961) and Sutch (1965), who used capital theory models to determine that slavery was indeed a profitable investment. Fogel (1964a) showed that the railroad was not the determinant of American economic development that it was believed to have been, while Fishlow (1965) overturned the notion that the railroads were built ahead of demand, and Temin (1969) showed that President Jackson did not cause the financial panics of the 1830s. Revisions occur when new data are discovered, or new techniques are applied to existing data, that otherwise could not be properly analyzed. The aforementioned use of census data is an example.

Demographic data have been used in place of the traditional measures of income and output to gain a fresh perspective on the Industrial Revolution. Greg Clark (2014, 2015) explores an alternative to the standard institution and market based stories by focusing on surnames, in particular the idea that the economically successful in a society will likely be the demographically successful. Voitländer and Voth (2013) argue that the Black Death gave rise to a European marriage pattern that in turn set in motion a process that led to the Industrial Revolution, a bold claim that leads to a dramatic revision of the economic history of Western Europe.¹⁰

And then there is Douglass North. In his 1968 ocean shipping article, he used data gathered from sailor wages, shipbuilding costs, freight rates, ship tonnage, and ship records to determine that institutions, not technology, were responsible for the increase in the productivity of ocean shipping from the 17th to the 19th century. The decrease in piracy and quicker turnaround times in port contributed more to productivity gains than did the previously credited technological advances.

2.4. Compilation of data sets

While cliometrics is often associated with technique, it is data that is at its center. Without good data, the best technique is limited in its ability to illuminate the truth. The application of sound theory to solid, unbiased data, is central to cliometrics. The most significant break from narrative history is the application of theory to data. The collection, collation, and careful vetting of the data to clean it of errors remains central to the task that cliometricians set out to perform.

It is the lack of relevant data more than the lack of relevant theory that is often the greater problem in historical research. In this way, cliometricians have made some of the greatest contributions to the fields of economics and history by discovering and compiling new data sets that have been, and will continue to be, used by future researchers to better understand the evolution and growth of economies over time.

The building of databases has a long and storied history among cliometricians. As previously mentioned, the slavery database built by Parker-Gallman and expanded upon by Fogel and Engerman is an example of early cliometric emphasis on building data sets. Robert Fogel, in his railroad work, provides another example of clio’s early focus on data.

Fogel’s breakthrough work was *Railroads and American Economic Growth* (1964a). At the time of its publication, economists believed they had established that modern economic growth was due to certain important industries having played a vital role in development. Fogel set out to measure this impact, which he did with extraordinary precision. He famously found that the railroad was not necessary to explain economic development and that its effect on the

growth of GNP was minimal. Herein was the difference between the “old” economic history and the “new:” the use of newly created data series and cutting edge techniques - made more useful, applicable, powerful and easy to replicate and reconsider, with the growth of computing power, to bring a laser focus to a problem.

While cliometrics promoted quantitative analysis of the highest quality data available, cliometricians did not inaugurate the concept. During his service to the U.S. government in the First World War, Edwin Gay¹¹ became convinced of the need for better economic statistics. He and Wesley Mitchell headed the *Central Bureau of Planning and Statistics*, responsible for the gathering and reporting of statistical data. Together they helped found the NBER to stimulate the collection and interpretation of historical statistics. The NBER ultimately served as a catalyst for the change in emphasis from narrative to quantitative studies in economic history.

The accumulation of the data is in itself monumental in many respects, but its usefulness has been expanded by the rapid growth of computing power. The ability to handle “Big Data” is not a cliometric issue by itself, but the construction of significant, important historical data sets, which can then be analyzed using cutting edge econometric techniques utilizing the latest software, is very much a contribution of cliometrics.

The marriage of cliometrics and “Big Data” is a natural one, and has been exploited by economic historians in new and creative ways. The work of James Feigenbaum (2015) is one recent example. He uses new automated linking methods to manage mammoth volumes of census data. In less obvious ways, large-scale qualitative databases are now being used to analyze text (Gentzkow, *et al* 2014, Wehrheim 2019), and GIS mapping allows geographic data to be quantified (Atack 2019). On a broader level is the Integrated Public Use Microdata Series (IPUMS), which provides census and survey data from around the globe in easy to use formats for a broad range of research on economic, social, and health research topics. IPUMS USA collects, preserves, and harmonizes U.S. census microdata and provides easy and free access to the data, which includes all available census data and 21st century American Community Surveys.

Economic historians have always been known for their work in building databases by collecting data from archival sources and gathering it into useful forms for analysis. Before the advent of the computer and digitization this was done by hand, painstakingly and slowly. With the evolution of advances in technology including the computer, scanning, and digital photography, the task was not eliminated, but sped up, allowing for the building of larger and richer data sets.

Demography has benefitted from technology and techniques that have allowed for the explosion of individual level microdata. The digitization of marriage, birth, and death records, state censuses, immigration records, and the federal census have made it possible to gather large amounts of data on individuals by linking these databases in order to follow them across generations in an effort to better understand behavior, causality, and intergenerational mobility.

Aizer et al (2016) collected individual-level administrative records of applicants to the Mothers’ Pension program, an early U.S. welfare program, and matched them to death records and the census in order to study the long-run impact of cash payments to poor families. Their efforts revealed the positive health and income outcomes for children whose mothers received the benefits.

Bailey et al (2019), Ferrie (1996), and Abramitzky et al (2019) are at the forefront of a burst of research using long existing census data that technology has made more useful. Their efforts at using machine-learning techniques to improve the size and accuracy of linked data sets using the federal census over time and linking it across other data sets, such as state censuses and birth and death records, have increased our ability to study intergenerational effects of economic events such as migration, education, health, and occupation.

Advances in data linkage techniques have led to an increase in the quality and applicability of the data. “New large-scale linked data are revolutionizing empirical social science . . . Examples abound across subfields in economics, including health economics and medicine, industrial organization, development economics, criminal justice, political economy, macroeconomics, and economic history.”¹² New projects are underway to link national surveys, administrative data, and research samples to recently digitized U.S. census records.¹³ These “Big Data” have the potential to break new ground on old questions and open entirely novel areas of inquiry.

What we have is an example of changing technology (computer and software) increasing the usefulness of preexisting data. The census data is not new, but its usefulness has been increased by the technology allowing us to link people across time, creating new panel data sets. “New large-scale linked data hold the potential to shift the frontier of knowledge.”¹⁴

2.5. Sharing Data

As mentioned earlier, technology has improved our ability to share data. Files can easily and instantly be sent across the ether, or even better, permanently stored there for the world to use. The collection of data has been cataloged at sites such as EH.net, MeasuringWorth.com, the Global Price and Income History Group, and the Cambridge Group for the History of Population and Social Structure. In France, the ClioData database of the French Cliometric Association completes the Carolus database, which compiles data connected with the economics of education. Carolus has actually contributed numerous cliometric and econometric works. In Germany, the Histat database is a “must”.¹⁵

Eh.net is host to more than two dozen datasets uploaded by scholars. The available data include financial series, trade statistics, labor statistics, commodity prices, manufacturing censuses, and public debt. One of the larger projects housed on the site is the Historical Labor Statistics Project Series. The project was established in 1990 for the purpose of collecting detailed data on U.S. labor markets gathered by state Bureaus of Labor Statistics from the late 19th to the early 20th century. Numerous cross-sectional surveys of firms and workers, which include information on working conditions, living standards, and family demographics, are currently available.

The Global Price and Income History Group has gathered vast quantities of data on prices and incomes for the period prior to 1950 from around the world. MeasuringWorth.com includes series for real and nominal GDP for the US (since 1790), UK (since 1300), Japan (since 1879), China (since 1952), wages, price indices, daily closing values of the Dow Jones since 1885, interest rates, and exchange rates. And this is only a partial list.

The Cambridge Group for the History of Population and Social Structure, founded in 1964, has compiled massive quantities of demographic, economic, and political data on the United Kingdom dating back to the medieval period. The data have been organized and aggregated at a variety of levels into numerous datasets, all of which are publicly available.

The Longitudinal, Intergenerational Family Electronic Microdata (LIFE-M) is an ongoing project that exemplifies new data sets, new techniques, and new approaches. LIFE-M is a large-scale public database that extends from the late 19th into the 21st century. It uses vital records as a basis for linking with census data from 1880 to 1940, providing birth to death coverage of individuals identified in the census. When completed, the combination of birth, death, and marriage records with data across censuses will produce a four generation database, including for the first time substantial numbers of women and minorities.

Sam Williamson created MeasuringWorth, which he founded in 2006 and has continued to maintain with no institutional support. The mission of MeasuringWorth is to make the highest quality and most reliable historical data on economic aggregates available to the public. The database includes nominal and real price measures for the United States, the United Kingdom and Australia. In addition, the site provides carefully designed comparators using these data that explain the many issues involved in making value comparisons over time and across countries. The data and accompanying text are designed to be useful to both professional researchers and the general public alike. The site is continually updated and expanded upon, and includes a variety of tutorials, user guides, and essays on the construction and use of the data. In a 2018 survey of its membership, the Economic History Association reported that over 40% of their members accessed the site for either research or teaching purposes on a regular basis.

In 2018 the Economic History Association inaugurated two prizes to recognize significant contributions to public databases. The Engerman-Goldin Prize, named after Stanley Engerman and Claudia Goldin, whose research famously involved creation of significant data sets covering education, labor, and slavery, would be awarded in even numbered years to honor individuals for creating, compiling, and sharing data and information with scholars in the recent past. The Gallman-Parker Prize, whose namesakes Robert Gallman and William Parker were also noted for their creation of data sets that transformed our understanding of the growth of the American economy, is awarded in odd-numbered years for lifetime construction, maintenance, and dissemination of databases that have been publicly available.

Michael Haines was the first recipient of the Gallman-Parker Prize for his work on historical demographics. Among other publicly available datasets emanating from his work is county level U.S. data, which includes population, agricultural and other economic and social statistics dating from the late 18th century. Jeremy Atack was the inaugural Engerman-Goldin Prize winner for his contributions to largescale datasets for public use. His historical transportation files, which consist of multiple files in GIS readable format of river, canal, and railroad transportation networks in the United States from the late 18th through the early 20th century, are freely available on his personal website. His joint work with Fred Bateman and Tom Weiss contributed to the manufacturing census database found on eh.net and the ICPSR at the University of Michigan.

The Cliometric Society began inducting Fellows of the Society in 2010. Among those inducted, several were cited for their contributions to database creation, including the aforementioned Robert Fogel and Douglass North. Sam Williamson was also cited for his efforts at creating and maintain the MeasuringWorth database, as well as Jeremy Atack and Fred Bateman, for their work on the transportation and manufacturing censuses. Other Fellows noted for contributions to databases included Alan Olmstead, Richard Sutch, and Michael Haines, who were part of the editorial team that put together the *Historical Abstract of the United States Millennial Edition*, considered the standard source for quantitative indicators of America's history, with more than 37,000 data series.

Matt Jaremski (2019) makes a case for the value of database creation in the education of graduate students. Graduate school is the most likely environment when young scholars will have support and time to collect and build a new dataset that has the potential to advance the frontiers of knowledge. "By necessity, economic history classes must cover how data are collected and what to do when data are not available . . . most research in economic history is built on new data and archival work rather than unknown historical events . . . Therefore, to understand that research, one must understand how authors gathered or created the data and why." ¹⁶ He argues that young scholars tend to gravitate towards the same publically available data, and because of its heavy use, they are forced to make more marginal contributions rather than opening new areas of study.

What is on the horizon for economists in terms of publicly available data? Gutman, et al (2018) highlight what they see as the future of "Big Data" for economic historians. Increasingly, they will be working with three new types of data: generically created numeric data, high-resolution digital images, and digitized texts. As economic historians begin to research more modern data, from the 1980s forward, they will encounter "Big Data" in many generically generated forms, including prices (online shopping), income (income and tax data), and government records on health, education and labor markets (much of which has been available since the 1950s). High-resolution image data is not frequently used by economists, but Glaeser et al (2018) have discussed the ability to use images to study economic behavior, and satellite data have already been used to study geographic impacts on economic behavior. As mentioned earlier, text analysis is a technique that will only grow in importance as more texts, especially those intimately related to the interests of economic historians, such as government documents, become more widely available in digital format.

Conclusion

Measurement requires that the data used be reliable, valid, and meet specific criteria. They must be standardised so as not to compare information that cannot be compared statistically. In addition, specificities must be taken into account in order to understand what apparently similar data can mean in very different contexts across time and countries, and to avoid meaningless analyses making daring comparisons of figures representing different realities. This will be our greatest data challenge going forward as more data becomes more easily available in ready-to-use digital format.

It isn't that we don't value the construction of datasets anymore, it is that they have become commonplace. Digitization and technology have made the creation of new datasets for research more common, because they are more easily available in digital form. This reduces the cost in terms of time and money to travel to archives, transcribe data, codify it, and clean it. It is easier to share, manipulate, and analyse with modern technology. Hence, new data sets more frequently appear and provide fodder for new research.

The ability to create a new data set for research not only increases the likelihood of novel work, it can create the potential for future work using the same data set, and when shared, can spawn an industry on the use of the data set. The creation of a new data set has become part of the research project for many economic historians. It is built into the research project, not a separate project itself. The creation of a data set in and of itself has never been regarded as a final product, but when the cost of assembling the dataset was greater, the desire to exploit it for more projects may have been greater. The decreasing cost of data has increased the number of datasets, thus giving the appearance of less value being placed on their creation. It isn't that data are valued less, it is that we now live in the luxury of cheap and abundant data.

Theory, technique, and data are the three pillars on which Cliometrics rests. No two of these pillars by themselves can uphold the discipline without the other. While it may seem that data has lost its importance, it most assuredly has not. As we have shown, the importance of data has not diminished over time, and new datasets are still being created, used, and shared to advance research. The importance of data is recognized in the discipline.

What has changed is the perception of the importance of data. Data can now be shared in easy to access and use formats. It is made ever more useful with advancing technology, which has made it easier to gather, codify, share, and use data. Technology has also allowed for the proliferation and exploitation of "Big Data," providing us with a new universe of research possibilities. This embarrassment of riches has likely contributed to the impression that the creation of new datasets is no longer valued in the discipline. To the contrary, data whether "Big" or small, is more important now than it ever has been.

Sources

- Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James J. Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," NBER Working Paper No. 25825, (May 2019)
- Acemoglu, Daron, Simon Johnson, and James A. Robinson, "The colonial origins of comparative development: An empirical investigation," *American Economic Review* 91, no. 5, (Dec 2001), 1369-1401
- Acemoglu, Daron, Simon Johnson, and James A. Robinson, "Reversal of fortune: Geography and institutions in the making of the modern world income distribution," *Quarterly Journal of Economics* 117, no. 4, (Nov 2002), 1231-1294
- Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long Term Impact of Cash Transfers to Poor Families," *American Economic Review* 106, no. 4 (2016): 935-71
- Allen, Robert C., "Revising England's Social Tables Once Again," *University of Oxford Discussion Papers in Economic and Social History*, no. 146, (July 2016)

- Atack, Jeremy, "Railroads," in Diebolt, Claude, and Hauptert Michael, eds., *Handbook of Cliometrics second edition*, Berlin: Springer-Verlag, forthcoming 2019
- Bailey, Christopher, Tarique Hossain, and Gary Pecquet, "Private banks in early Michigan, 1837-1884," *Cliometrica* 12, no. 1 (January 2018): 153-80
- Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well do Automated Linking Methods Perform? Lessons from U.S. Historical Data," *working paper University of Michigan*, (February 20, 2019)
- Becker, Sascha O., and Woessmann, Ludger, "Was Weber wrong? A human capital theory of Protestant economic history," *Quarterly Journal of Economics* 124, no 2, (May 2009), 531-596
- Charette, Michael F., and Ronald Meng, "The determinants of literacy and numeracy, and the effect of literacy and numeracy on labour market outcomes," *Canadian Journal of Economics* 31, no 3, (Aug 1998), 495-517
- Clark, Gregory, *The Son Also Rises: Surnames and the History of Social Mobility*, Princeton, NJ: Princeton University Press, 2014
- Clark, Gregory, "Markets before economic growth: the grain market of medieval England", *Cliometrica*, 9, 3, 2015: 265-287
- Conrad, Alfred H. and John R. Meyer, "The Economics of Slavery in the Antebellum South," *Journal of Political Economy* 66, April 1958, 75-92
- Davis, Lance E., "Sources of Industrial Finance: The American Textile Industry, A Case Study," *Explorations in Entrepreneurial History*, IX (April 1957), pp 189-203
- Davis, Lance E. Davis, "Stock Ownership in the Early New England Textile Industry," *The Business History Review*, XXXII (Summer 1958), 204-22 2
- Davis, Lance E., "The New England Textile Mills and the Capital Markets: A Study of Industrial Borrowing, 1840-1860," *The Journal of Economic History* XX (March 1960), 1-30
- Davis, Lance E., and Jonathan R. T. Hughes, "A Dollar Sterling Exchange 1803-1895," *Economic History Review* (August 1960)
- De Moor, Tine, and Jan Luiten Van Zanden, "'Every woman counts': A gender-analysis of numeracy in the Low Countries during the early modern period," *Journal of Interdisciplinary History* 41, no 2, (Autumn 2010), 179-208
- Diebolt, Claude and Michael Hauptert, "Clio's Contributions to Economics and History," *Revue d'Economie Politique* 126, no. 5, (2016), pp 971-89
- Diebolt, Claude, Gabrielle Franzmann, Ralph Hippe, and Jürgen Sensch, "The Power of Big Data: Historical Time Series on German Education," *Journal of Demographic Economics*, 83, no 3, 2017, pp 329-376
- Engerman, Stanley, "The Height of U.S. Slaves," *Local Population Studies* 16, (1976), 45-50
- Engerman, Stanley L., John R. T. Hughes, Donald N. McCloskey, Richard C. Sutch, and Samuel H. Williamson, *Two Pioneers of Cliometrics: Robert W. Fogel and Douglass C. North, Nobel Laureates of 1993*, Miami, OH: The Cliometric Society, 1994
- Feigenbaum, James, "Automated Census Record Linking: A Machine Learning Approach," *unpublished manuscript*, 2015

- Fenoaltea, Stefano, "Spleen: The Failures of the Cliometric School," *working paper University of Turin*, (November 2018)
- Ferrie, Joseph, "A New Sample of Americans Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules," *Historical Methods* 29, (Fall 1996), 141-156
- Fishlow, Albert, *American Railroads and the Transformation of the Ante-bellum Economy*, Cambridge, MA: Harvard University Press, 1965
- Floud, Roderick, and Bernard Harris, "Health, Height, and Welfare: Britain 1700-1980," 91-126. in Richard H. Steckel and Roderick Floud, eds., *Health and Welfare during Industrialization*, Chicago: University of Chicago Press, 1997
- Floud, Roderick and Kenneth Wachter, "Poverty and Physical Stature, Evidence on the Standard of Living of London Boys 1770-1870," *Social Science History* 6, (1982), 422-52
- Fogel, Robert, "A Quantitative Approach to the Study of Railroads in American Economic Growth: A Report of Some Preliminary Findings." *Journal of Economic History* 22, no. 2 (June 1962): 163-197
- Fogel, Robert, *Railroads and American Economic Growth: Essays in Econometric History*. Johns Hopkins, Baltimore, 1964a
- Fogel, Robert, "Discussion," *American Economic Review* 54, no.3 (May 1964b): 377-389.
- Fogel, Robert, "Physical Growth as a Measure of the Economic Well-being of Populations: The Eighteenth and Nineteenth Centuries," 263-281, in F. Falkner and J.M. Tanner, eds., *Human Growth: A Comprehensive Treatise*, second edition, volume 3, New York: Plenum, 1986.
- Geloso, Vincent, "Distinct within North America: living standards in French Canada, 1688-1775," *Cliometrica* 13, no. 2 (May 2019): 277-321
- Gentzkow, Matthew, Jesse Shapiro, and Michael Sinkinson, "Competition and Ideological Diversity: Historical Evidence from US Newspapers," *American Economic Review* 104, no. 10, (2014), 3073-3114
- Giddings, Lisa, and Michael Hauptert, "Earning Like a Woman: the gender gap in professional baseball 1944-1954," *Journal of Sports Economics* 20, no. 2 (February 2019), 198-217
- Glaeser, Edward L. Scott Duke Kominers, Michael Luca, and Nikhil Naik, "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life," *Economic Inquiry* 56, no. 1 (2018): 114-37
- Gust, Sarah, and Joerg Baten, "Interpersonal violence in South Asia, 900-1900," *working paper Universität Tübingen*, (March 2019)
- Gutman, Myron P., Emily Klancher Merchant, and Evan Roberts, "'Big Data' in Economic History," *Journal of Economic History* 78, no. 1 (March 2018): 268-299
- Haines, Michael R. "Growing Incomes, Shrinking People – Can Economic Development Be Hazardous to Your Health? Historical Evidence for the United States, England, and the Netherlands in the Nineteenth Century." *Social Science History* 28, (2004), 249-70
- Hauptert, Michael, "A Brief History of Cliometrics and the Evolving View of the Industrial Revolution," *European Journal of the History of Economic Thought*, forthcoming 2019
- Hauptert, Michael, and James Murray, "Regime Switching and Wages in Major League Baseball Under the Reserve Clause," *Cliometrica* 6, no. 2, (June 2012), 143-162

- Hughes, Jonathan R. T. and Stanley Reiter, "The First 1,945 British Steamships," *Journal of the American Statistical Association*, LIII (June 1958), pp 360-81
- Jaremski, Matt, "Today's Economic History and Tomorrow's Scholars," *Cliometrica* online first, accessed June 2019
- Keywood, Thomas., and Joerg Baten, "Elite violence and elite numeracy in the Middle East from 500 CE to 1900 CE," presented at the *Mobility and Migration in Historical Perspective* conference, Regensburg, Germany, (March 20-22, 2019)
- Lindert, Peter H., and Jeffrey G. Williamson, "Revising England's Social Tables, 1688-1812," *Explorations in Economic History* 19, no. 4 (October 1982): 385-408
- Lindert, Peter H., and Jeffrey G. Williamson, "Reinterpreting Britain's Social Tables, 1688-1913," *Explorations in Economic History* 20, no. 1 (January 1983): 94-109
- Manzel K, and Joerg Baten, "Gender equality and inequality in numeracy: The case of Latin America and the Caribbean 1880–1949," *Review of Economic History* 27, (2009), 37–74
- Margo, Robert and Richard Steckel. "The Height of American Slaves: New Evidence on Slave Nutrition and Health," *Social Science History* 6, (1982), 516-38
- Meyer, John R., and Conrad, Alfred H., "Economic Theory, Statistical Inference, and Economic History," *Journal of Economic History*, vol 17, 4, 1957, December, 524-544
- Mills, Terence C., "Exploring historical economic relationships: two and a half centuries of British interest rates and inflation," *Cliometrica* 2, no. 3 (October 2008): 213-28
- Mitch, David, "The Contributions of Robert Fogel to Cliometrics," *The Handbook of Cliometrics*, Springer Reference Live, https://link.springer.com/referenceworkentry/10.1007/978-3-642-40458-0_49-1, accessed June 2019
- Newmarch, William, in collaboration with Thomas Tooke, A History of Prices, and of the State of the Circulation during the nine years, 1848–56, forming the fifth and sixth volumes of the History of Prices from 1792 to the present time,' London, 8vo, 1857
- North, Douglass, "Sources of Productivity Change in Ocean Shipping, 1600-1850," *The Journal of Political Economy* 76, (September/October 1968), 953-70
- Nuvolari, Alessandro, Bart Verspagen, and Nick von Tunzelmann, "The early diffusion of the steam engine in Britain, 1700–1800: a reappraisal," *Cliometrica* 5, no. 3 (October 2011): 291-321
- O'Rourke, Kevin H. and Jeffrey G. Williamson, "From Malthus to Ohlin: trade, growth and distribution since 1500". *Journal of Economic Growth* 10, no. 1 (2005):5–34
- Pfister U and Fertig G. 2010 The Population History of Germany: Research Strategy and Preliminary Results, *Max Planck Institute for Demographic Research working paper no 35*
- Ruggles, Steven, "Linked Historical Censuses: A New Approach," *History and Computing* 14 (2006): 213-224
- Ruggles, Steven, Katie Genadek, Josiah Grover, and Matthew Sobeck, Integrated Public Microdat Series (Version 6.0) [machine-Readable database], edited by University of Minnesota, Minneapolis: University of Minnesota, 2015

- Spoerer, Mark, Thomas Brenner, Alexander Gebhardt, and Patrick Schwabl, "Is the geographical distribution of church bell casting dates a useful proxy for comparative regional economic growth? Preliminary results from the Upper Palatinate, 13th to 18th centuries," *working paper Universität Regensburg* (March 21, 2019)
- Steckel, Richard. "Slave Height Profiles from Coastwise Manifests." *Explorations in Economic History* 16, (1979), 363-80
- Steckel, Richard, "Anthropometrics," in Diebolt, Claude, and Hauptert Michael, eds., *Handbook of Cliometrics second edition*, Berlin: Springer-Verlag, forthcoming 2019
- Steckel, Richard H., and Roderick Floud. *Health and Welfare during Industrialization*, Chicago: University of Chicago Press, 1997
- Sutch, Richard, "The Profitability of Ante-Bellum Slavery: Revisited," *Southern Economic Journal* 31, no 4, (April 1965), 365-77
- Temin, Peter, *The Jacksonian Economy*, New York: W. W. Norton and Company, 1969
- Tollnek, Franziska, and Joerg Baten, "Age-Heaping-Based Human Capital Estimates," in Diebolt, Claude, and Hauptert Michael, eds., *Handbook of Cliometrics*, Berlin: Springer-Verlag, 2016
- Vasta, Michelangelo, Carlo Drago, Roberto Ricciuti, and Alberto Rinaldi, (2017) "Reassessing the bank-industry relationship in Italy, 1913-1936: a counterfactual analysis," *Cliometrica*, vol 11 no 2, (May 2017), pp 183-216
- Voitländer, Nico, and Voth, Hans-Joachim, "How the west 'invented' fertility restriction," *American Economic Review* 103, no 6, (2013), 2227-2264
- Wehrheim, Lino, "Economic history goes digital: topic modeling the Journal of Economic History," *Cliometrica* 13, no. 1 (January 2019): 83-125
- Weisdorf, Jacob, "Church Book Registry: A Cliometric View," in Diebolt, Claude, and Hauptert Michael, eds., *Handbook of Cliometrics*, Berlin: Springer-Verlag, 2016
- Yasuba, Yasukici, "The Profitability and Viability of Plantation Slavery in the United States," *Economic Studies Quarterly*, (September 1961)

¹ James Faust and Fred Bateman provided substantial help in the collection and coding process.

² For an overview of this literature see Mitch, 2019

³ Their preliminary results are for the Upper Palatinate region of Germany.

⁴ See Wehrheim 2017

⁵ Engerman *et al*, 1994, p 71

⁶ For example, they cited four additional data-processing studies in economic history carried out at Purdue in the late 1950s that had developed entirely new statistical series, and could not have been conducted without the latest technology or mathematical models: Lance Davis's textile studies (1957, 1958 and 1960), and the Davis and Hughes exchange rate study (1960)

⁷ Diebolt and Hauptert 2016

⁸ See Weisdorf 2016

⁹ See Steckel (2019) for a thorough coverage of the history of anthropometrics and its uses by cliometricians.

¹⁰ See Hauptert (2019) for a survey of the contributions cliometricians have made to our understanding of the Industrial Revolution.

¹¹ Gay became the first president of the Economic History Association in 1940.

¹² Bailey et al 2019, p 1

¹³ See Ruggles, et al 2015, Ruggles 2006

¹⁴ Bailey et al 2019, p 41

¹⁵ Diebolt et al 2017

¹⁶ Jaremsky online first, accessed June 2019