Combining Family History and Machine Learning to Link Historical Records[†]

Joseph Price
Brigham Young University,
NBER, and IZA

Kasey Buckles
University of Notre Dame,
NBER, and IZA

Jacob Van Leeuwen
Brigham Young University

Isaac Riley
Brigham Young University

Abstract

A key challenge for research on many questions in the social sciences is that it is difficult to link historical records in a way that allows investigators to observe people at different points in their life or across generations. In this paper, we develop a new approach that relies on millions of record links created by individual contributors to a large, public, wiki-style family tree. First, we use these "true" links to inform the decisions one needs to make when using traditional linking methods. Second, we use the links to construct a training data set for use in supervised machine learning methods. We describe the procedure we use and illustrate the potential of our approach by linking individuals across the 100% samples of the US decennial censuses from 1900, 1910, and 1920. We obtain an overall match rate of about 70 percent, with a false positive rate of about 12 percent. This combination of high match rate and accuracy represents a point beyond the current frontier for record linking methods.

[†] This work has been supported in part by grant #G-1063 from the Russell Sage Foundation. Any opinions expressed are those of the principal investigators alone and should not be construed as representing the opinions of the Foundation. We are grateful for helpful comments and assistance from Ran Abramitzky, Martha Bailey, Katherine Eriksson, James Feigenbaum, Joe Ferrie, Ian Fillmore, Cathy Fitch, Brigham Frandsen, Katie Genadek, Jonas Helgertz, Bob Pollack, Steve Ruggles, and Anne Winkler. This project would not have been possible without the careful and thoughtful work of many research assistants at the Brigham Young University Record Linking Lab, including Ben Branchflower, Alison Doxey, Neil Duzzett, Nicholas Grasley, Amanda Marsden, and Joseph Young.

I. Introduction

For many of the most pressing questions in the social sciences, empirical analysis relies on access to data that allow the researcher to observe people at different points in their life or across generations. For example, to measure the intergenerational transmission of socio-economic status, we need to be able to link a parent to his or her adult child; to estimate the long-term impacts of childhood circumstances, we typically need to observe a person as both a child and as an adult. Unfortunately, this kind of data has been hard to come by in the United States, due to a lack of a consistent individual registration number that is recorded in census data and in many administrative data sets (as exists, for example, in Sweden and Norway).

Recently, researchers studying the U.S. have solved this problem by acquiring restricted-use data with information that permits this linking. This includes work that uses Social Security numbers to link tax records across generations (Chetty and Hendren, 2018), to education histories (Chetty et al., 2017), or to survey data (Mazumder and Davis, 2013). This innovative approach is limited by the fact that the data are only available for recent decades, and Social Security numbers are not recorded in many data sets where we would like to have them, such as censuses or vital statistics data. Another strategy has been to link individuals across censuses and other records by matching on characteristics like the person's name, birth year, and birth place (Ferrie, 1996; Abramitzky, Boustan, and Eriksson, 2014; Evans et al., 2016; Abramitzky, Mill, and Pérez, 2018). A drawback of this method is that it is known to produce non-representative samples and has typically omitted women, whose names often change between childhood and adulthood (Beach et al. 2016; Pérez 2019). A relatively new approach to record linking is to use supervised machine learning algorithms. Unlike the unsupervised matching methods described above, supervised methods require a training data set that includes examples of both correct and incorrect matches that the algorithms can "learn" from to make new matches. While machine learning methods show

promise, quality training data can be difficult and costly to obtain. Both Feigenbaum (2016) and Bailey (2018) use a process that relies on skilled human trainers to create their training data.

In this paper, we propose a new approach for linking individuals across historical records. At the center of our method is a data set created from decisions that are made by millions of people who are researching their own family histories. These researchers often gather source documents—including census records—to establish various life events and relationships for a family member, and then post their conclusions on genealogical websites like Ancestry, FamilySearch, FindMyPast, MyHeritage, Geni, and Wikitree. The key feature we exploit is that when the profile for a deceased individual on one of these websites has multiple sources attached, each pair of these sources establishes a "correct" match that can potentially be used to inform the decisions made when employing various linking strategies, and as training data for supervised machine learning methods.[1] The data are highly reliable, as the family members doing the linking have a personal interest in making a correct match. Furthermore, the family members typically have private information that can be used to identify the person of interest across multiple data sets, such as maiden names or the names of other household members.

The genealogy platform we use for our study is FamilySearch. FamilySearch has created a large, public, wiki-style family tree that includes a profile for over 1.2 billion deceased individuals with over 12.6 million registered users who can contribute information to those profiles. Individuals can upload information and sources to the profiles of their own ancestors and relatives and can make edits to the conclusions and sources attached by other contributors working on the same people. In addition, FamilySearch provides regular record hints as suggestions to these contributors,

[1] Of course, in most cases there is no way to determine conclusively whether a match is true. However, both Abramitzky et al. (2019) and Bailey et al. (2019) refer to genealogy data as the "gold standard" of hand linking, and use links from our FamilySearch database as a benchmark for evaluating the accuracy of their methods.

who then decide whether the source should be attached to that person. We use a sample of individuals from this family tree that are attached to at least two census records between 1900 and 1920. This provides a data set with 4.6 million 1900-1910 links, 4.9 million 1910-1920 links, and 2.9 million 1900-1920 links.

We describe a process that builds on these data to create millions more links among these three censuses. First, the FamilySearch data allow us to examine several important decisions that need to be made when using automated methods to link historical records. These decisions include how to pre-process the data, which features to use to identify potential matches (blocking), and which machine learning algorithm to use. We then use the FamilySearch links as training data for a supervised machine learning algorithm and combine the links we get from this machine learning approach with other methods to link records. Our final data set, which we call the "Census Tree" data, contains 67.8% of the potential matches between the 1900 and 1910 full-count US censuses, and 71.4% of the potential matches between the 1910 and 1920 full-count US censuses (or 42.7 and 54.8 million matches, respectively). We hand-check a random sample of our predicted matches and also use a transitivity test to show that the false positive rate among our predicted matches is about 12%. Abramitzky et al. (2019) document a "production possibilities frontier" showing the tradeoff between accuracy and the number of possible matches made for several widely-used record linking approaches and our method reaches a point well beyond this frontier.

Ultimately, our goal for this project is to create every possible link among the full-count US decennial censuses from 1850 to 1940.[2] But the method we describe in this paper could be applied to any pair of datasets for which there is a sufficient number of individuals with records in both

[2] The effort will extend past 1940 as later censuses become available under the Census Bureau's 72-year release policy.

collections linked by users of a genealogy platform. As a result, the potential of the method will grow even beyond this ambitious goal, as the use of genealogy websites like FamilySearch spreads around the globe and as more historical records are digitized.[3]

II. Background

The 100 percent samples of the US decennial censuses are made available to the public after 72 years, which opens up unique opportunities to link individuals over long periods of time. Several approaches have been used by social scientists to create large linked samples. These include creating pre-determined rules to identify unique matches (Ferrie 1996; Abramitzky, Boustan and Eriksson 2014; Collins and Wanamaker 2015; Beach et al. 2016; Alexander and Ward 2018), employing a statistical algorithm such as expectation-maximization (Abramitsky, Mill, and Pérez 2018; Pérez 2019), using hand-linked data (Bailey 2018; Costa et al. 2018), or combining human-created training data with machine learning algorithms (Feigenbaum 2016; Goeken et al. 2011; Bailey et al. 2019). Each of these approaches have their advantages and disadvantages, and they are likely to complement each other in a combined effort to link as many individuals across historical records as possible (see Bailey et al. (2019) and Abramitzky et al. (2019) for summaries of these approaches). In this section, we describe in more detail two papers that have used supervised machine learning to link historical records, as their approaches are most closely related to the main approach that we use in this paper.

Supervised machine learning requires training data with examples of both correct and incorrect matches. An algorithm then uses training data to determine which characteristics (or features) are best able to predict whether two records are a match. Feigenbaum (2016) proposes a

[3] In the case of FamilySearch, outside researchers can use the FamilySearch API to obtain links directly from the Family Tree. A similar strategy can be used for other websites.

machine learning procedure for linking a sample of men in the 1915 Iowa Census with their record in the 1940 census. He limits the set of potential matches to those with the same birth state, born within 2 years of each other, and with similar first and last names (based on Jaro-Winkler distance). He creates 17 features, all of which are based on name, birth year, and the number of possible matches, and uses a probit regression to estimate which of these features predict the likelihood that a particular pair of records is a correct match. Finally, a correct match is defined as one that has a high match score and has a significantly higher match score than any other possible matches. For his sample of 7,580 boys in the 1915 Iowa Census, he is able to find a match in the 1940 census for 57% of them, with an estimated false positive rate of around 13%.

Goeken et al. (2011) use a machine learning approach to create the IPUMS Linked Representative Samples of the 1850-1930 US censuses. Their blocking features are race, gender, and birthplace, and they require that matches have birth years within seven years of one another. They use a Support Vector Machine as their machine learning algorithm and combine this with two sources of training data. The first set of training data was created by data entry operators who coded a set of potential matches as true or false based on a visual examination of the names and ages of the possible links. For the second, they assessed links created by a company that produces record linkage software for genealogical research. Along with features based on name, birthplace, gender, and birth year, they also include features on parental birthplace, name commonality, and birth density (which is the fraction of the census born in particular states by race and gender). This approach yielded a data set with nearly 179,000 matches between the 1880 complete-count census and the 1% samples of the 1850-1930 censuses.

III. Data

We use two sources of data for this project. The first dataset is the 100% sample of the US decennial census for 1900, 1910, and 1920. These data provide the raw records that we aim to link together. These data include each person's name, birth year, birthplace, gender, race, and place of residence, and the birthplace of their father and mother. All of these variables were transcribed (or indexed) from the original digitized images by volunteers recruited by FamilySearch. The data also include other family members who are living in the same household which allows us to observe the same characteristics for the individual's parents, siblings, spouse, or children, depending on who they are living with in each census.

The second dataset is comprised of linked census records that were provided to us by FamilySearch. These matched pairs come from their online, wiki-style genealogy platform called the Family Tree. The Family Tree was created in 2012 and allows anyone to contribute once they have set up a free account. The website is structured to allow individuals to collaborate when they have a family member in common, and various relatives of the same individual on the tree can contribute information about vital events, family members and historical sources. This is an active crowdsourcing platform with over 450,000 site visits per day, 12.6 million registered users, and over 1.2 billion individual profiles of deceased people.

The link between census records and FamilySearch profiles is made by FamilySearch users themselves, who find records on the FamilySearch platform and attach them to an individual's profile. An individual profile could also be attached to vital records, military records, school records, and city directories. We include an example profile in Figure 1 to illustrate the potential of the data. For this person, we can observe the dates of birth, death, and marriage, and links to several public records. The record links include the 1900, 1910, and 1920 censuses, which allow us to create a panel with observations for this person at ages 9, 19, and 29. The first two observations are from

a time when he lived in his parents' household, while in the latter he was the head of his own household.

This process produces a large data set of highly reliable links among records, as the family members doing the linking identify the person of interest across multiple data sets more accurately than can be done by name matching methods. For example, family members are more likely to know maiden names, or to know which census record for a "John Williams" belongs to their family member.[4] This type of data collection strategy has been validated in recent work by Kaplanis et al. (2018), who used data from 86 million profiles from Geni.com, a similar website that allows users to create individual profiles and upload family trees. The authors attempted to confirm the links in the family trees using DNA data, and found very low rates of non-maternity and non-paternity that were consistent with rates of adoption, and that the lifespan and death information had a 98% concordance with historical distributions from the Human Mortality Database. The authors also compared the data to a population sample from Vermont, and found that the Geni.com sample was highly representative of the population. Kaplanis et al. (2018) conclude "that millions of genealogists can collaborate in order to produce high quality population-scale family trees" (p. 172).

We call this data set, which consists of matches made by FamilySearch users themselves, the "Family Tree" data set. Table 1 provides information about the size of the Family Tree data. We split the individuals in our training set into mutually exclusive groups based on their gender and which census records they are attached to. Some of the people in the Family Tree data are attached to all three census records from 1900 to 1920, while others are only linked to two of the censuses. Individuals linked to all three censuses provide three sets of matched pairs. Altogether, the Family

[4] One way family members do this is by using the names of the other people in the household. For example, if I know that the John Williams I am looking for had an older sister named Sarah and a younger brother named Joseph, that can drastically reduce the number of potential matches. The "Household Matching" strategy we describe below mimics this approach.

Tree data provide 12.3 million true matched pairs across the three censuses. There are more men than women, as men are easier for contributors to link across multiple records because their surnames rarely change. Nevertheless, the fact that women make up nearly half of our sample is a substantial advance in this literature; the poor performance of conventional name-matching methods for linking women's records has meant that women have been completely omitted from some important research.[5]

We validate the quality of the Family Tree data in two ways. First, we compare links from the Family Tree with the links created by the human trainers working on the LIFE-M project (Bailey 2018). LIFE-M provided us a set of 54,000 individuals that they had linked from an Ohio birth certificate to the 1940 census. We were able to find 12,000 people from their sample that were attached to both an Ohio birth certificate and the 1940 census on the Family Tree. Of this overlapping sample, we found that that the links on the Family Tree and those identified by LIFE-M agreed 93.4% of the time.

We then took the 991 cases where there was disagreement and asked hand research assistants to use traditional family tools to determine which match was correct. They found that 75.4% of the time the link based on the Family Tree was correct, 26.1% of the time the LIFE-M link was correct, 4.3% of the time they were both right (because the individual showed up twice in the 1940 census), and 4.3% of the time neither of the links were correct. Treating all of the links where LIFE-M and the Family Tree data agree as correct indicates that, for this overlap group, the LIFE-M links were correct 95.2% of the time and the links based on the Family Tree were correct 98.4% of

[5] While this paper describes a process for using the Family Tree data to create additional matches, the Family Tree links alone constitute an incredible data set, with millions of links among census records that include many matches between women before and after marriage.

the time. This suggests that the Family Tree links achieve a level of accuracy similar to or better than that created by skilled human trainers, at a much lower cost.

We also validate the quality of the Family Tree data by having humans hand-match a random sample of the records. Among the 500,000 matches for our Ohio sample between 1910 and 1920, we randomly sampled 100 records from the 1920 census and provided them to trained research assistants and asked them to use the search tools on Ancestry to identify the number of potential matches for that person in the 1910 census and which of those possible matches they determined was the correct one based on their inspection of the information from the two records. On average, they identified 12 individuals in the 1910 census that were a possible match for each person in the sample from the 1920 census. The 1910 census record that they labeled as a match for each 1920 census record agreed with the match in the Family Tree data 98% of the time. We replicated this with a random sample of 350 record links from our full data set. Of those 350 records, they were able to find a link 94% of the time and of these links that were found, they agreed with the link in the Family Tree data 99% of the time.

IV. Method

The process we use to link census records is depicted in Figure 2. We begin with the Family Tree data as described above. These user-made links not only serve as an input into our supervised machine learning algorithms, but also as a resource for making informed decisions about how to pre-process the data, which blocking and matching features to choose, and which machine learning algorithm to use. We now describe these different uses of the data in more detail.

A. Pre-processing the data

There are three features in the census that are especially important for our linking methods: birth year, birthplace, and name. We employ some pre-processing to each of these features to

improve the accuracy of our machine learning models. First, the birth year variable is imputed in our data in 1910 and 1920 based on the age that the person reported. In 1910, age was based on the age of the person on April 15th of that year, and in 1920 was based on the person's age on January 1st of that year. In 1900, the individual reported both their birth month and birth year, so no imputation is necessary.

Second, in the census records, the most specific birthplace listed for those born in the U.S. is the state in which they were born. For those born outside of the U.S., there are varying levels of specificity used; for example, birthplaces in the Netherlands were sometimes listed with their city of birth (e.g. "Amsterdam Netherlands") or province of birth ("Friesland Netherlands"). We pre-process the birthplace in two ways. For those born in the U.S., we clean the spelling of each birth state to have a single standardized name (the state of Connecticut is spelled 97 different ways in the data). For those born outside of the U.S., we standardize the birth place to be the name of the country of birth, though certain abbreviations such as "ata" and "o" are difficult to classify.

Third, we do some cleaning to convert nicknames and abbreviations to a standardized set of formal names. Each matched pair in the Family Tree data allows us to see two potential ways to spell an individual's first name. We use this information to create a network between every combination of uniquely spelled names and use the strength of the edges between nodes in this network to identify common nicknames and abbreviations. This provides us with a list of 1,704 nick names and abbreviations that we can convert into a formal name equivalent. We also create features that help address common misspellings by employing Soundex and Jaro-Winkler distances to handle transcription errors.[6]

[6] See Massey (2017) for an analysis of the tradeoff between linkage rates and accuracy when using Jaro-Winkler distances in matching.

B. Blocking and matching features

Blocking features are the characteristics of an individual for which you require an exact match in your matching algorithm. Blocking is required for nearly all matching to make it computationally possible to do the linking. Past studies have often required exact matching on birth state (Feigenbaum, 2018); birth year within a given number of years (Goeken et al., 2011; Feigenbaum, 2018); and the letters of the first and last names (Mill & Stein, 2016). These blocking strategies can be problematic when fields are indexed incorrectly, when information is not reported or recorded correctly on the census, or when people change aspects of their identity over time (such as race or last name). One notable case occurred after World War I when the number of people who reported being born in Germany or Austria dropped by roughly 40% between the 1910 and 1920 census (Charles et al. 2018). Many of these people likely changed their last name and birth place in response to the discrimination occurring in the U.S. during the war (Fouka 2018).

In Table 2, we use the 1900-1910 and 1910-1920 links from the Family Tree data to provide some information about the level of stability across adjacent census years in some of the potential blocking features. The first column in this table provides the fraction of the Family Tree data for which each of the characteristics is the same for the individual between the 1900 and 1910 census and the second column does the same for 1910 and 1920. For example, between the 1900 and 1910 censuses only 75% of people have the exact same first name, but 94% have the same first initial of their first name. Three of the most stable characteristics of individuals are their race (99.8%), gender (99.8%), and birthplace (96.6%). Values for the 1910 and 1920 links are similar. The stability of these features is a reason why they have frequently been used for blocking in previous studies. We also include columns that provide the number of unique values for each of the characteristics; this highlights the natural tradeoff for blocking strategies as the characteristics that are the most stable

are also the least unique. The uniqueness of the characteristics directly affects the size of the blocks, with less unique features producing larger blocks that make it more difficult to make a match.

The Family Tree data allow us to evaluate the performance of blocking strategies used in prior work. We focus on the blocking strategies used by Ferrie (1996), Abramitzky et al. (2014), Feigenbaum (2016), and Abramitzky et al. (2018), which are based on different combinations of state of birth, gender, first initial of first name, first initial of last name, and birth year. The results are in Table 3. The column labeled "consistency" indicates the fraction of the linked pairs in the Family Tree data for which all of the characteristics used in the blocking strategy are the same across the two records. This measure provides a proxy for the upper bound of the match rate that is possible using each of the blocking strategies.

Beginning with the first column in Panel A of Table 3, we see that the blocking strategy used by Ferrie (1996) would have included 73% of the true matches from the Family Tree data and ABE would include 49%. This means that there would have been no way to link the other 27% or 51% of the sample because they would not have been included in the set of possible matches. The other two approaches perform better on this dimension, with 83% for Feigenbaum (2016) and 86% for Abramitzky et al. (2018). However, the next two columns show the advantages of Ferrie's strategy. Each observation in the Ferrie approach would, on average, require 7.4 comparisons to be computed when linking the 1900 and 1910 census (and ABE would only have 2.1), while the Abramitzky et al. (2018) approach would require almost 200 times as many comparisons. The number of potential matches has a direct effect on the computing time required to create predicted scores for all of the possible matches. The third column indicates the number of unique matches that are identified with that set of potential matches; the ABE blocking strategy produces the most unique matches. The results are similar in Panel B, which uses matches between the 1910 and 1920 censuses. This exercise shows that the choice of blocking strategy can have a dramatic effect on

computing time, and that there is a natural trade-off between consistency and the number of unique matches a strategy is able to produce.

Informed by this exercise, we block on race, gender, birth year within 3 years, and the Soundex values for the first and last name. However, we can construct multiple features based on each of these variables to use in the matching process for the machine learning algorithm. For example, we create features for whether the first name matches exactly, whether the first initial of the first name matches, whether the Soundex value of the first name matches, and the Jaro-Winkler similarity score of the first name. We construct similar measures for the middle name and last name. We also create indicators for the similarity of birth year, birth place, race, and gender. Finally, the machine learning algorithm can include information on the similarity of the mother and father's birthplace, the names of family members, and place of residence.

C. Choosing the machine learning algorithm

There are many different algorithms that one can use to link records using a supervised machine learning method. Ideally, an algorithm performs well on three dimensions. First, it should be accurate—that is, most of the identified matches are true matches (often referred to as precision in the machine learning literature). Second, it should be efficient—many true matches are identified among those that are possible (often referred to as recall). Third, it should be computationally fast. Because we observe true matches in the Family Tree data, it can help inform this choice as well. In Table 4, we show how five different machine learning algorithms perform along these three dimensions when we ask them to make matches among the Family Tree links, using a subset of the links. We see that XGBoost performs best in terms of both accuracy and efficiency. We therefore

choose it as our classifier, though it is a bit slower than neural nets, random forests, and logit regression.[7]

D. Other steps in the process

After pre-processing the data and selecting the blocking strategy and machine learning algorithm, we implement the supervised machine learning method. In the early stages of this process, we asked trained research assistants to evaluate a random sample of the predicted matches and code them as either "true" or "false"; these links were then added to the training data set to further refine the matching algorithm.[8]

We also employ other linking strategies. First, we use the automated matching strategy developed by Abramitzky, Boustan, and Eriksson (2014) (ABE). Second, we use a dyad matching method that is described in more detail in Helgertz and Price (2019) and involves linking multiple people across households at the same time. This approach is helpful when the individual level

[7] XGBoost is a library that builds high-performing gradient boosting tree models. XGBoost has the benefits of a decision tree model with the added advantage of boosting through an ensemble learning method. XGBoost works by creating gradient boosted decision trees which split our data based on included features in order to predict an outcome. Gradient boosted decision trees are many-decision trees that are produced one after another where each sequential tree is specifically built using the residual errors of the previous model as target areas to improve upon and minimize loss and misclassification. XGBoost does this using a leaf-wise growth strategy meaning that the next tree splits at the leaf that reduces the greatest amount of loss. The benefits of using a tree-based model include scalability to large data sets, outlier robustness, and natural handling of missing data. This is important in our data as missing values are common and features have a variety of distributions, many of which are non-normal.

[8] While there are over 12.3 million links in the Family Tree data, we actually use only a subset of these when implementing the XGBoost algorithm. Using the full data set slows the processing speed considerably without increasing predictive power. This is consistent with the findings of Feigenbaum (2016) who shows that gains to accuracy and efficiency plateau at around 500 observations in the training data. The training data we use has 43,800 true matches and 794,607 false matches, where the false matches are other potential matches for the true matches. The advantage of the Family Tree data as training data is not its size, but rather the fact that it is high quality and can be obtained at a relatively low cost. The size is more important when using the data to inform the data cleaning and matching processes as described above.

information is sufficiently different that the other approaches are not sure if it is a match but when multiple close matches cluster in the same household, confidence in all of the matches increases.[9] Third, we use a household matching method which takes advantage of links created through the other methods. Once a person in a household in one census has been linked to a person in a household in another census, we employ a rules-based algorithm to identify other individuals in the two households that should be linked together. This works well for married couples, parent-child relationships, and siblings who are often in the same household together in adjacent censuses.[10] Fourth, we also include additional record hints that FamilySearch shared with us using their own machine learning algorithm. The files they shared with us provided us with 54 million links; of these, 5.7 million were links that we did not identify using one of the other methods.[11] Finally, we implemented the full process to link 1920 to 1900 and used a transitivity rule to create additional implied links for 1910 to 1900 and 1920 to 1910. That is, we take advantage of the fact that if we link a record between 1900 and 1920 and 1910 and 1920, we also have a link between 1900 and 1910. This last method will become even more valuable as we add other censuses to our data.

V. Results

We now apply this process to produce a linked dataset of individuals across the 1900, 1910 and 1920 US censuses. Table 5 reports the number of links we are able to make between adjacent censuses, using each of the strategies described in the previous section. We provide the total

[9] For example, if Andrew Buckles and Mary Buckles are siblings in two adjacent censuses, but both are excluded from our blocking strategy due to a mismatched feature (such as a difference in birth years of more than three years), the fact that the family relationship was the same across the two censuses gives us more confidence in linking the records for both Andrew and Mary.
[10] That is, if we have successfully linked Andrew using one of the other methods, and Andrew has a sister Mary in both censuses, we can then link the two records for Mary.
[11] We note that this is the only one of our linking strategies that is not readily available to other researchers.

number of matches that are obtained from each strategy as well as the number of new matches obtained when we apply the methods in sequence. Focusing on links between the 1910 and 1920 censuses, we see that the Family Tree data itself provides about 4.9 million links. The ABE automated linking method and the XGBoost machine learning method contribute an additional 19 million links each. Household matching and dyad matching add 2.6 and 4.2 million links, respectively.

Once we have employed the first four methods, we remove the linked records from our sample and then implement the ABE method a second time. This generates an additional 1.1 million links and highlights the potential for this approach to create additional matches as the number of unlinked individuals shrinks, in an interative process. FamilySearch hints and the implied links provide 3.9 million more links, so that the final Census Tree linked sample consists of 54.8 million links from 1910 to 1920. Given our estimate of the number of possible matches, we conclude that we have identified 71.4% of the possible links between these two censuses.[12] For 1900 to 1910, we identify 42.7 million links for a match rate of 67.8%.

In Table 6, we show the number of matches and match rates by the individual's relationship to the household head. We construct our base sample and use the household relationship code for the second of the two years for each pair of years, such that the 1910-1920 links are based on the 1920 data. The groups with the highest match rates are sons (74.0%) and daughters (72.7%). This is expected, as children living with their birth family in the 1920 census would likely have been living with their birth families in the 1910 census as well (here again, children who were born after 1910 are not included when calculating match rates). The group with the next highest match rates are

[12] See the Appendix for a description of how we estimate the number of potential matches.

male heads of household (64.0%), followed by the spouses of the head of household (55.1%).[13] The lowest match rates occur for other household members who are not part of the immediate nuclear family where the match rates fall to 44.6% for women and 39.5% for men.

We employ two methods to evaluate the false positive rate among the predicted matches that we obtain. First, we can examine the transitivity property between the predicted matches that we create. For example, our machine learning algorithm allows us to create predicted matches between the 1900 and 1910 census, the 1910 and 1920 census, and the 1900 and 1920 census. This triangle of links provides a number of transitivity tests that we can use to provide a measure of the quality of our matches—that is, if the model had a prediction for all three possible links, they should agree. Our final Census Tree linked sample includes 15.1 million of these transitivity tests and for this set, we find an agreement rate of 87.3% which implies a false positive rate of 12.7%.

Second, we drew a random sample of 1,000 records from the 1920 census and asked research assistants to use traditional genealogy tools to hand link these individuals to the 1910 census. They were able to find a link for all of those in the random sample and the link they identified agreed with our predicted link 88% of the time, again implying a false positive rate of

[13] The lower match rate for spouses of the household head reflects the difficulty in matching women before and after marriage, when their surnames usually changed. The subset of the Family Tree data that we use as training data include many of these links, as family members often know maiden names. However, the training data are unable to "teach" the algorithm to make these matches since the algorithm does not have this private information. We therefore still miss many of these links before and after marriage. Nevertheless, the across-marriage links from the Family Tree data set alone constitute a valuable resource for researchers.

about 12%.[14] In addition, we find that this false positive rate was about the same for men and women with a false positive rate of 12.8% for men and 11.1% for women.[15]

How do these measures compare to those achieved with other methods? Abramitzky et al. (2019) report the levels of accuracy and efficiency for various linking methods using a linked sample between the 1900 census and Union Army records.[16] They define accuracy as the number of correct matches divided by the total number of matches, and efficiency is calculated as the number of correct matches divided by the total possible number of matches. [17] Figure 1 of their paper shows the tradeoff between accuracy and efficiency, and can be thought of as a "production possibilities frontier" in the record linking literature. We have reproduced this figure as Figure 3, and added the two points achieved by the 1900-1910 and 1910-1920 Census Tree samples.[18] While we are using a different sample, the fact that we are trying to create accurate matches between two data sets with tens of millions of observations each makes linking *more* difficult—the sample Abramitzky et al. is attempting to match contains 1,647 possible links. As described above, about 12% of our matches

[14] While this may seem high, as we discuss below this false positive rate combined with our high match rate puts us well beyond the frontier for linking methods. Furthermore, it is possible that for many questions in the social sciences, the false positives that linking methods generate do not bias the results because the falsely linked people would often have had similar outcomes on average, given that they share so many other characteristics (name, birthplace, and birth year). This hypothesis is supported by the work of Olivetti and Paserman (2015), who show that first names contain important information about socioeconomic status; future work could formally test this hypothesis using the Census Tree data.

[15] We can also use the FamilySearch record-hinting system to continue to monitor the quality of our matches and use error analysis to improve the matching methods. FamilySearch has a system for emailing individuals using their platform about possible record hints for individuals that they are related to. These record hints also show up on the right side of the screen on the profile for each person on the Family Tree. We are sharing all of the predicted links that we identified through our pipeline with FamilySeach and will be able to observe in the future the decisions that individuals make with our predicted matches.

[16] They compare their links to those in the Union Army-Oldest Old sample, treating the latter as the "truth."

[17] Note that the efficiency rate is different from the match rate; the former has only correct matches in the numerator, while the latter has both correct and incorrect matches.

[18] We are grateful to the authors for providing the data and code to allow us to replicate this figure.

are false, for an accuracy rate of 88%, and our efficiency rate is 59.6% for the 1900-1910 Census Tree sample and 62.9% for 1910-1920. Figure 3 shows that this combination of accuracy and efficiency is well beyond the current frontier.

Finally, we aim to produce a sample that is representative of the entire US population for each census year. Table 7 compares those individuals that we are able to match to the previous census to those that we are not, for the 1910 and 1920 censuses. The samples are limited to those age eleven and over, to omit those who would not have been born in the previous census and therefore cannot possibly be matched. In the third and sixth columns, we also remove those who immigrated since the previous census. First, we see that while the unmatched samples are closer to 50% female, it is still the case that 47.1% (47.5%) of our matched sample is female in 1910 (1920). This level of representation of women constitutes a tremendous advance in the census linking literature. We do see that the matched sample is more likely to be white and less likely to be black, is slightly older, and is more likely to be a household head. The matched sample is also more likely to have been born in the U.S., which is unsurprising given that state of birth is one of the matching features.[19]

Despite these differences between the matched and unmatched samples, the fact that the final Census Tree data set is so large means that we still observe millions of links for groups that are under-represented in our data. For example, we have about 5.7 million links between the 1910 and 1920 censuses for people who were not born in the U.S. Thus, our data should be useful to researchers who are studying small or minority populations and to those who are interested in

[19] We also implement the balance test recommended by Bailey et al (2019) that employs a heteroscedasticity-corrected F-test for joint significance and find that we reject the null that our sample is balanced.

comparative analyses. The large sample sizes should also permit researchers to re-weight the data to construct a sample that is representative of the US population.

VI. Conclusion

Recent developments in data access and record linking methodology have created exciting opportunities for social science research using large populations (Gutmann, Merchant, and Roberts, 2018). We contribute to this work by developing novel ways to use data created from the contributions of millions of individuals who are investigating their own family histories on FamilySearch, a genealogy web platform. These researchers often gather records from censuses and other sources and link them together on a family member's profile. Effectively, the FamilySearch users do the work that trained research assistants would do to try to link records but at a much lower cost, and with a personal interest in identifying correct matches and private information that allows them to make accurate matches that other methods cannot. The result is a high-quality data set with links among censuses and other records for millions of people.

In this paper, we document the value of this new source of data. Taking the links created by the FamilySearch users alone, we observe 12.3 million links among the 1900, 1910, and 1920 censuses. These include links for women before and after marriage, which have typically been very difficult to make using other methods due to the change in surname. But we also show that these data provide several insights that will be helpful for advancing the state of art in machine-based records linking. For example, the data can be used to identify common nicknames and abbreviations for common names, to explore the stability and uniqueness of common blocking and matching features, and to test the performance of different machine learning algorithms. Finally, we show that the data can be used as a very large and reliable training data set for use in supervised machine learning approaches.

To demonstrate the potential of the data, we combine a supervised machine learning algorithm that uses data from the FamilySearch Family Tree as training data with other record linking methods to generate links among the 1900, 1910, and 1920 full count US censuses. We are able to create 68% of the potential matches between the 1900 and 1910 censuses, and 71% of the matches between 1910 and 1920. With a false positive rate of about 12%, our approach reaches a level of accuracy and efficiency that is beyond the frontier for record linking methods (Abramitzky et al. 2019).

The integration of family history research with automated record linking methods has the potential to dramatically improve the quality and quantity of data available to researchers in the social sciences, and to economic historians in particular. We are working to expand the current census-linking project to include all full-count censuses between 1850 and 1940. Beyond this effort, we note that training could be created with the same approach for any two types of records that are available on various genealogical platforms, including vital records, military records, and school records. Furthermore, as the use of genealogy web platforms expands around the world, researchers will be able to use our method to link records across and within other countries.

References

Abramitzky, Ran, Leah Boustan, and Katerine Eriksson. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122, no. 3 (2014): 467-506.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Boustan, Leah, Eriksson, Katherine, James Feigenbaum, and Santiago Pérez. "Automated Linking of Historical Data." NBER Working Paper No. 25825, Cambridge, MA, 2019.

Abramitzky, Ran, Roy Mill, and Santiago Pérez. "Linking Individuals Across Historical Sources: A Fully Automated Approach." NBER Working Paper No. 24324, Cambridge, MA, 2018.

Alexander, Rohan, and Zachary Ward. "Age at Arrival and Assimilation During the Age of Mass Migration." *The Journal of Economic History* 78, no. 3 (2018): 904-937.

Bailey, Martha J. "Creating LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database." Working Paper, University of Michigan, Ann Arbor, MI, 2018.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey. "How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth." *Journal of Economic Literature,* forthcoming, 2019.

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

Charles, Kerwin, Tanner Eastmond, Joseph Price, and Daniel Rees. "Long-Run Consequences of Prejudice." Working paper. 2018.

Chetty, Raj, John Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility." NBER Working Paper No. 23618, Cambridge, MA, 2017.

Chetty, Raj, and Nathaniel Hendren. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects." Q*uarterly Journal of Economics* 133, no. 3 (2018): 1107-1162.

Collins, William, and Marianne Wanamaker. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants." *The Journal of Economic History* 75, no. 4 (2015): 947-992.

Costa, Dora, Matthew Kahn, Christopher Roudiez, and Sven Wilson. "Data set from the Union Army samples to study locational choice and social networks." *Data in Brief,* 17, (2018): 226-233.

Evans, Mary, Eric Helland, Jonathan Klick, and Ashwin Patel. "The Developmental Effect of State Alcohol Prohibitions at the Turn of the Twentieth Century." *Economic Inquiry* 54, no. 2 (2016): 762-777.

Feigenbaum, James J. "Automated Census Record Linking: A Machine Learning Approach." W*orking Paper*, 2016.

---- "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940." *Economic Journal* 128, no. 612 (2018): F446-F481.

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." *Historical Methods* 29, (1996): 141-156.

Fouka, Vasiliki. "How Do Immigrants Respond to Discrimination? The Case of Germans in the US During World War I." *American Political Science Review* 113, no. 2 (2019): 405-422.

Goeken, Ron, Lap Huynh, Thomas Lenius, and Rebecca Vick. "New Methods of Census Record Linking." *Historical Methods* 44, (2011): 7-14.

Gutmann, Myron, Emily Merchant, and Evan Roberts. "'Big Data' in Economic History." *The Journal of Economic History* 78, no. 1 (2018): 268-299.

Hacker, J. David. "New Estimates of Census Coverage in the United States, 1850-1930." *Social Science History* 37, no. 1 (2013): 71-101.

Helgertz, Jonas, and Joseph Price. "Improving Precision in Linking Historical Census Data: Applying a Two-Stage Linking Procedure on the 1900-1910 US Census Data." *Working paper*, 2019.

Kaplanis, Joanna, Assaf Gordon, Tal Shor et al. "Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives." S*cience* 360, no. 6385 (2018): 171-175.

Massey, Catherine G. "Playing with Matches: An Assessment of Accuracy in Linked Historical Data." *Historical Methods* 50, no. 3 (2017): 129-43.

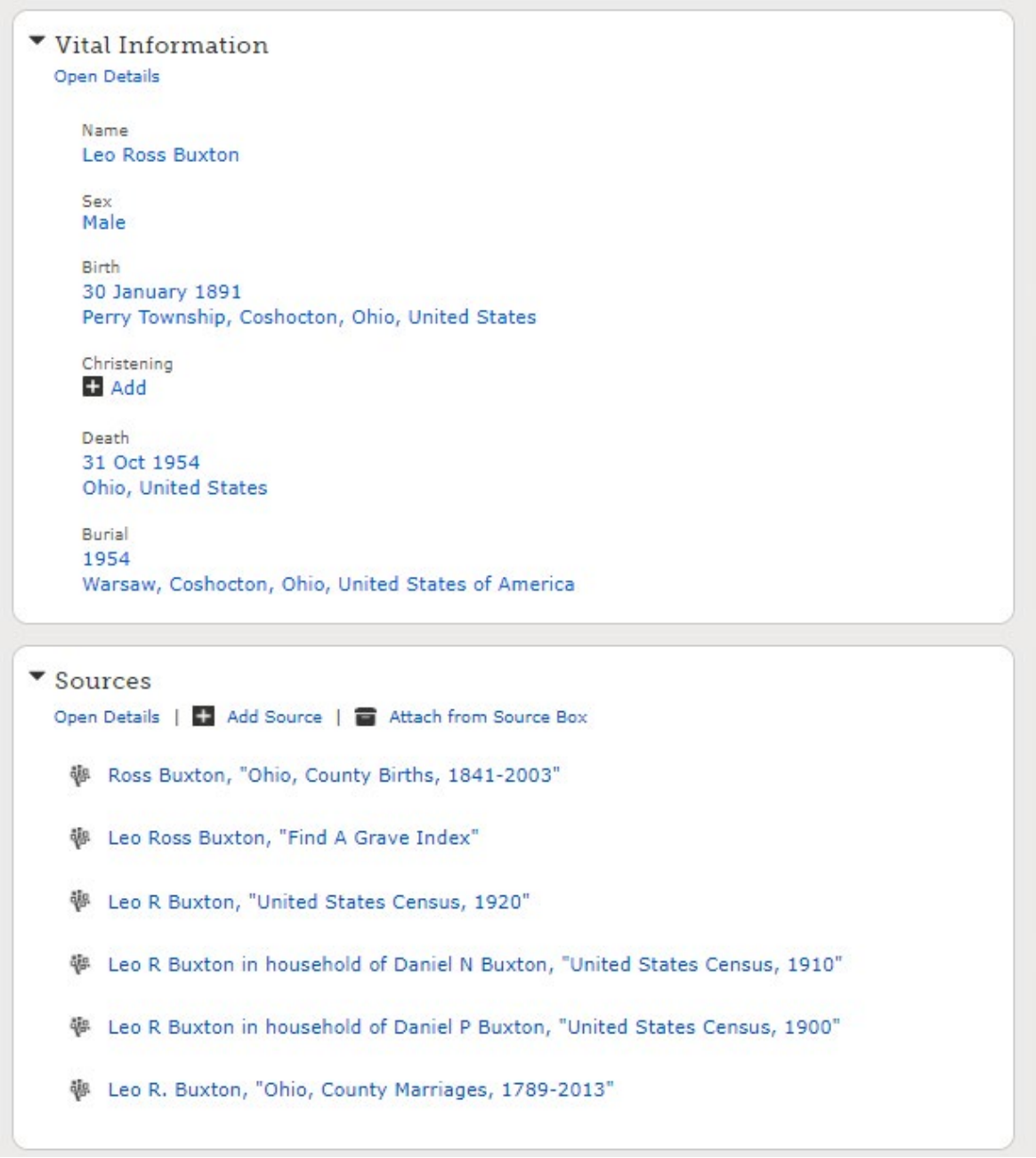
Mazumder, Bhashkar, and Jonathan M.V. Davis. "Parental Earnings and Children's Well-Being: An Analysis of the Survey of Income and Program Participation Matched to Social Security Administration Earnings Data." *Economic Inquiry* 51 no. 3 (2013): 1795-1808.

Mill, Roy, and Luke C. D. Stein. "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America." *SSRN*, Working Paper no. 2741797, 2016.

Olivetti, Claudia, and M. Daniele Paserman. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." A*merican Economic Review* 105, no. 8 (2015): 2695-2724.

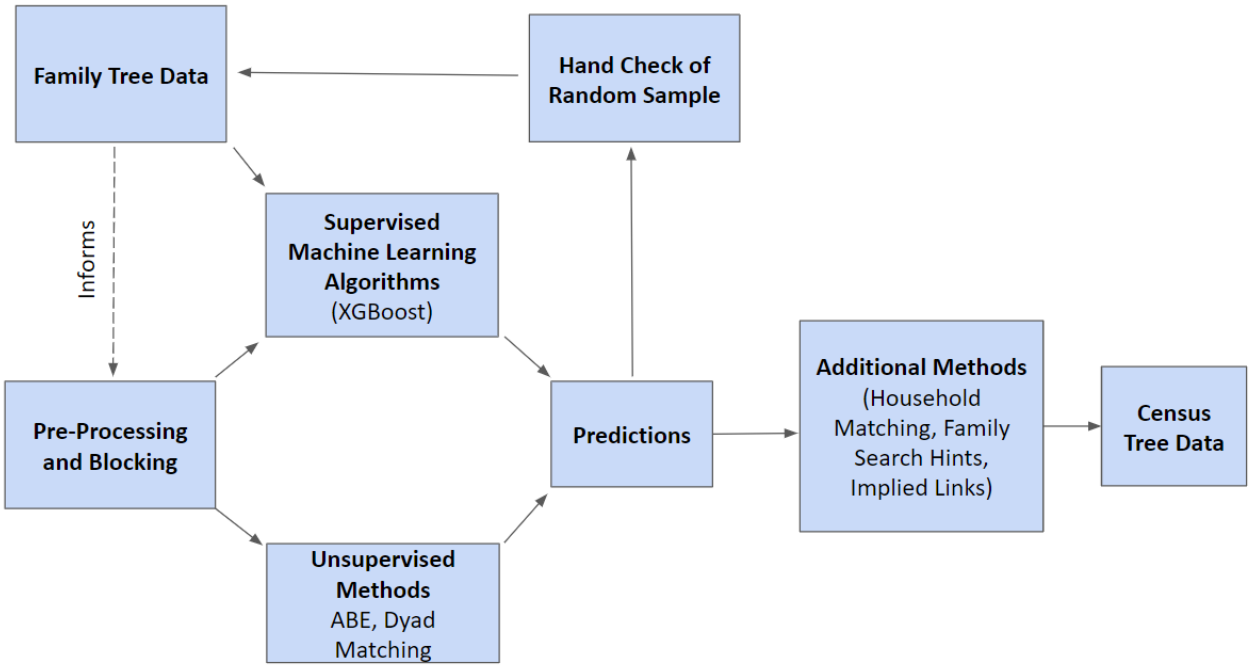
Pérez, Santiago. "Intergenerational Occupational Mobility across Three Continents." *The Journal of Economic History* 79, no. 2 (2019): 383-416.

Figure 1. Example of Person Profile with Sources on the Family Tree



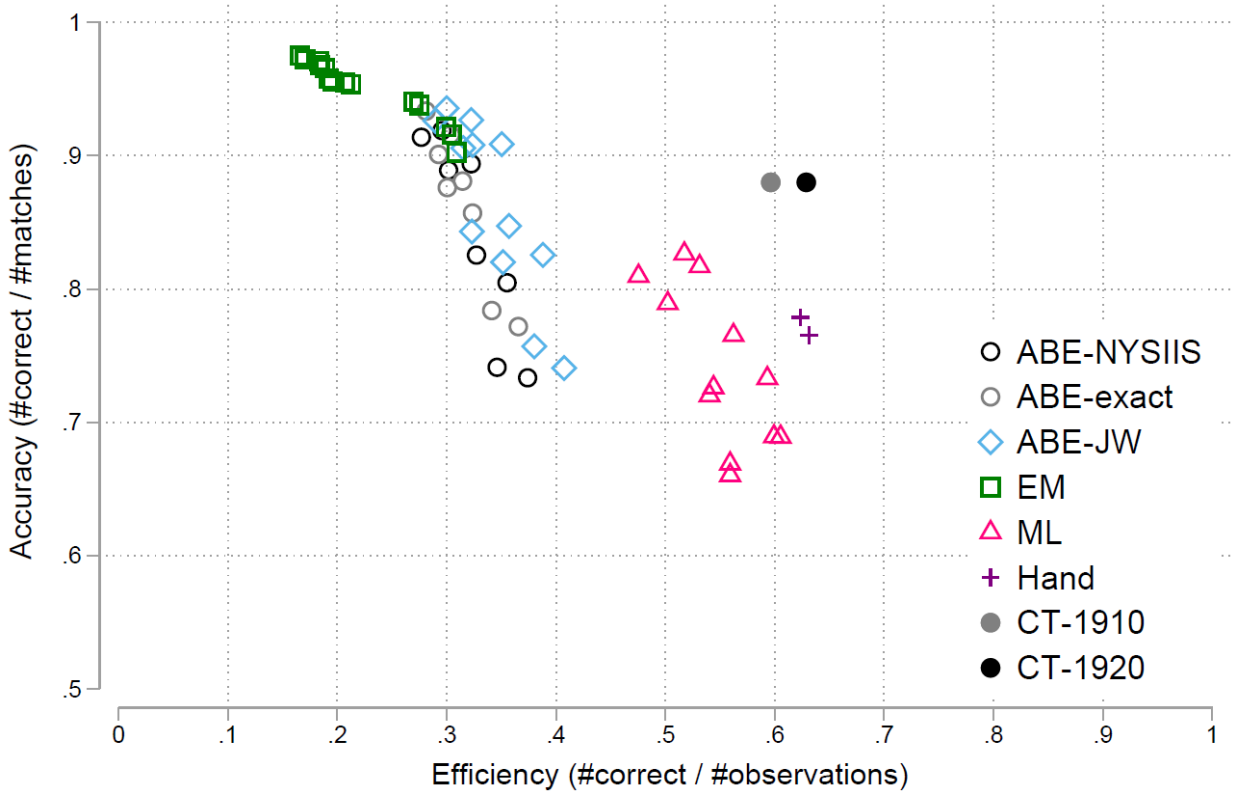
Notes: This is an example of an individual profile page on FamilySearch. There is a section that includes vital information about the person (name, birth, and death) and then a section with each of the sources attached to the person. Not shown is a separate section that provides names and links to each of the familial relations of the individual (parents, siblings, spouse, and children).

Figure 2. Process for Linking Census Records



Notes: Figure shows the process for going from the Family Tree data consisting of user-made links to the final Census Tree data set, which includes the Family Tree data as well as additional links made by supervised and unsupervised machine learning methods and additional linking strategies.

Figure 3. Accuracy vs. Efficiency for Various Linking Algorithms (Based on Abramitzky et al. 2019)



Notes: Figure is a reproduction, with permission, of Figure 1 in Abramtizky et al. (2019). For all but the Census Tree points (CT-1910 and CT-1920), data are from their exercise linking Union Army records to the 1900 census using different versions of the ABE automated matching strategy, an expectation maximization algorithm (EM), a machine learning method (ML), or a hand-linking strategy (Hand). CT-1910 documents the levels of accuracy and efficiency achieved in our Census Tree sample for 1900-1910; CT-1920 does the same for 1910-1920.

Table 1. Size of the Family Tree Data Set

	Women	Men
Only 1900 & 1910	1,275,583	1,356,810
Only 1910 & 1920	1,433,637	1,557,970
Only 1900 & 1920	442,814	536,256
1900 & 1910 & 1920	905,095	1,003,087

Notes: The table shows the number of links in the Family Tree data set, based on record matches made by FamilySearch users. Each of the cells in this table are mutually exclusive. The rows in the table indicate the censuses that are attached to each individual in the data. For example, the first row provides the number of women and men that are matched to just the 1910 and 1920 census in the Family Tree.

Table 2. Stability of Potential Blocking Features between the 1900, 1910, and 1920 Censuses

Feature	Stable 1900-1910	Stable 1910-1920	1900 Unique Values	1910 Unique Values	1920 Unique Values
Race/Ethnicity	0.998	0.998	14	17	15
Sex	0.998	0.999	2	2	2
Birth year within 3	0.978	0.984	-	-	-
Birthplace	0.966	0.984	2,927	648	1,937
Birth year within 2	0.964	0.973	-	-	-
Last initial	0.944	0.946	26	26	26
Last JW > 0.8	0.939	0.942	-	-	-
First initial	0.935	0.942	26	26	26
Last Soundex	0.908	0.915	5,209	5,209	5,235
Last JW > 0.9	0.908	0.915	-	-	-
Birth year within 1	0.885	0.923	-	-	-
First JW > 0.8	0.881	0.894	-	-	-
First Soundex	0.850	0.868	4,187	4,149	4,168
Mother's birthplace	0.840	0.852	3,862	1,052	3,230
Father's birthplace	0.837	0.848	4,158	1,125	3,522
First JW > 0.9	0.822	0.842	-	-	-
Last name	0.806	0.823	179,008	185,581	191,484
First name	0.743	0.769	79,837	85,326	85,170
County	0.741	0.746	1,628	1,724	1,818
Middle initial	0.656	0.654	26	26	26
Middle JW > 0.8	0.643	0.638	-	-	-
Middle JW > 0.9	0.635	0.638	-	-	-
Township	0.601	0.555	29,870	17,313	19,986

Notes: Data are from the full Family Tree data set (12.3 million observations). Features where the number of unique values are not reported are features where the values are binary (0 or 1). Data are from the 1900, 1910, and 1920 census records in the Family Tree data.

Table 3. Performance of Common Blocking Strategies Using the Family Tree Data

Panel A. 1900 and 1910 Censuses

Blocking Strategy	Consistency	Potential Matches	Unique Match
Ferrie (1996)	0.728	7.2	902,135
Abramitzky et al. (2014) (ABE)	0.485	2.1	1,005,944
Feigenbaum (2016)	0.829	10.4	652,387
Abramitzky et al. (2018)	0.842	1,367.6	7,159

Panel B. 1910 and 1920 Censuses

Blocking Strategy	Consistency	Potential Matches	Unique Match
Ferrie (1996)	0.756	5.9	1,069,929
Abramitzky et al. (2014) (ABE)	0.518	1.8	1,148,645
Feigenbaum (2016)	0.861	8.5	766,206
Abramitzky et al. (2018)	0.868	1,241.0	4,183

Notes: The analysis is performed on the full Family Tree data set. The first column indicates the fraction of true matches for which all of the characteristics used in each blocking strategy are the same across the two records (consistency). The second indicates the average number of potential matches across censuses for each individual. The third provides the number of pairs of records that are a unique match in one of the blocks. The sample size for Panel A is 4.52 million and the sample size for Panel B is 4.90 million.

Table 4. Performance Measures of Classifiers

Model	Accuracy (%)	Efficiency (%)	Processing Time (min)
XGBoost	91.95	89.95	16.72
Gradient boosting	91.74	89.54	28.45
Neural nets	90.01	86.86	4.06
Random forest	89.18	81.85	8.97
Logit regression	87.59	80.35	0.31

Notes: The analysis is performed on a subset of the Family Tree data. Accuracy is the fraction of identified matches that are true matches. Efficiency is the fraction of possible true matches that were identified. Processing time provides a relative measure of speed based on a representative run of the training data. In this case, the processing time represents the time it takes to train the classifier and run predictions on a dataset of 838,407 comparisons, with 43,800 true matches and 794,607 false matches. We use 80% of the data to train the model and 20% as a test set to get the measures of accuracy and efficiency.

Table 5. Contributions of the Different Methods Used to Link Records

	1900-1910		1910-1920	
Method	Matches	New Matches	Matches	New Matches
Family Tree data	4,567,392	4,567,392	4,912,838	4,912,838
ABE	17,093,182	14,665,647	21,811,611	19,064,068
XGBoost	26,855,325	14,845,508	29,063,701	19,028,103
Household matching	13,656,233	2,364,018	6,710,830	2,593,163
Dyad matching	12,846,431	548,382	25,085,386	4,214,342
ABE (second time)	961,544	656,070	3,424,294	1,111,118
FamilySearch hints	23,527,806	3,712,348	30,578,568	2,036,346
Implied from 1900-1920	3,087,749	1,335,657	2,579,389	1,847,714
Total		42,695,022		54,807,692
Match rate		67.8%		71.4%

Notes: Table summarizes the full Census Tree data set. "Matches" indicates the total number of matches created at each step of the iterative matching process. "New Matches" indicates the number of additional matches added to the cumulative total. The match rate is the total number of unique matches divided by the number of individuals in the later census that could have a match in the earlier census (see the Appendix for details).

Table 6. Match Rates by Relationship Type

	1900-1910		1910-1920	
	Matches	Match Rate (%)	Matches	Match Rate (%)
Male head	10,089,985	59.8	12,838,744	64.0
Female head	1,191,723	50.6	1,523,930	55.1
Spouse	7,862,816	51.8	10,670,849	57.7
Sons	7,524,531	67.2	9,594,710	74.0
Daughters	6,769,585	64.4	8,733,776	72.7
Other males	2,167,038	38.4	2,596,060	39.5
Other females	1,894,616	40.9	2,380,334	44.6

Notes: Table summarizes the full Census Tree data set. Each of the rows is mutually exclusive and based on the relationship to the household head that was reported in the later census record. The match rate is the total number of unique matches divided by the number of individuals in the later census that could have a match in the earlier census (see the Appendix for details).

Table 7. Summary Statistics for Matched and Unmatched Records in the 1910 and 1920 Censuses

	1910 Census			1920 Census		
	Matched to 1900	Unmatched	Unmatched (No Immigrants)	Matched to 1910	Unmatched	Unmatched (No Immigrants)
Female	0.471	0.498	0.520	0.475	0.509	0.515
	[0.499]	[0.500]	[0.500]	[0.499]	[0.500]	[0.500]
White	0.916	0.855	0.839	0.917	0.853	0.845
	[0.278]	[0.352]	[0.368]	[0.276]	[0.354]	[0.361]
Black	0.081	0.129	0.148	0.079	0.134	0.141
	[0.273]	[0.335]	[0.355]	[0.270]	[0.340]	[0.348]
Married	0.511	0.511	0.515	0.524	0.547	0.546
	[0.500]	[0.500]	[0.500]	[0.499]	[0.498]	[0.498]
Household head	0.312	0.264	0.272	0.318	0.282	0.285
	[0.463]	[0.441]	[0.445]	[0.466]	[0.450]	[0.451]
Age	34.55	32.32	32.98	35.43	33.19	33.40
	[17.38]	[15.44]	[16.01]	[17.70]	[15.49]	[15.72]
Born in US	0.874	0.718	0.828	0.879	0.722	0.769
	[0.331]	[0.450]	[0.377]	[0.326]	[0.448]	[0.421]
At least one parent born in US	0.712	0.587	0.677	0.721	0.580	0.618
	[0.453]	[0.492]	[0.468]	[0.449]	[0.494]	[0.486]
Lives in birth state	0.665	0.467	0.539	0.651	0.496	0.529
	[0.471]	[0.499]	[0.498]	[0.477]	[0.500]	[0.499]
N	39,805,878	30,000,265	26,010,807	46,778,504	30,944,708	29,024,181

Notes: Data are from the full Census Tree data set and from the 100% samples of the 1910 and 1920 decennial census. The terms "matched" and "unmatched" refers to whether we are able to link the individuals to a record in the previous census. The columns marked "No Immigrants" exclude individuals who immigrated after the previous census had taken place. Columns 1, 2, and 3 report summary statistics for the 1910 census. Columns 4, 5, and 6 report summary statistics for the 1920 census. All samples are restricted to individuals who are at least 11 years old in the latter census. Standard deviations reported in brackets.

Appendix: Calculating the Match Rate

Here we describe the procedure for calculating the number of matches we could possibly make between two censuses, for use in calculating match rates. For a pair of censuses, we begin with the number of people in the latter survey. We then subtract the number of people in the census whose information indicates that they either were born or immigrated since the previous census. We also account for the undercount in the previous census, as we will not be able to create a match between people in the latter census who were not enumerated in the former. Our numbers for the undercount are based on the estimates found in Hacker (2013). The numbers used to calculate the potential number of matches are shown in Appendix Table 8. We estimate that we could potentially find 63.0 million links between the 1900 and 1910 censuses (92.2 – 20.4 – 4.7 – 4.1), and 76.7 million links between 1910 and 1920 (106.5 – 22.3 – 2.2 – 5.3).

Appendix Table 8. Number of Possible Matches

	Number of Records in Census	Number Under-Enumerated	Number Born Since Prior Census	Number Immigrated Since Prior Census	Number of Possible Matches
1900	76.2	4.1	32.5	5.5	
1910	92.2	5.3	20.4	4.7	63.0
1920	106.5	6.9	22.3	2.2	76.7

Notes: All numbers are in millions. The number under-enumerated is calculated using the estimates in Hacker (2013).