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

Isaac Riley
Brigham Young University

Jacob Van Leeuwen
Brigham Young University

Abstract

A key challenge for research on many questions in the social sciences is that it is difficult to link Census 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 record links created by individual contributors to a large, public wiki-style family tree. This information is used to create a training data set that allows us to significantly improve upon previous machine learning methods. We describe the procedure we use and discuss our criteria for making decisions at each stage of the process. We illustrate the potential of our approach by implementing it on the 100% sample of the US Decennial Censuses from 1900, 1910, and 1920, and suggest several ways to verify the quality of links.

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 is the case in, for example, Sweden and Norway).

Recently, researchers studying the U.S. have solved this problem by acquiring restricted-use data with information on Social Security numbers that allow, for example, tax records to be linked across generations (Chetty and Hendren, 2018) or to education histories (Chetty et al., 2017). This innovative work 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 use name-matching methods to link individuals across censuses and other older data sets like military enlistment records (Abramitzky, Boustan, and Eriksson, 2014; Evans et al., 2016). 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. Another promising strategy is to combine historical records with genealogy information that is provided by users on online platforms. The University of Minnesota has formed a partnership with one such website, Ancestry.com, to make the transcribed records of 100% samples of the US census records available for academic research. Several different approaches are being used to link these records together and there is considerable discussion about the advantages and disadvantages of different approaches (Feigenbaum 2016; Abramitzky, Mill, and Perez 2018; Bailey, et. al. 2017).

In this paper, we propose a new approach that can be used in conjunction with other methods to link individuals across census records. We focus specifically on a novel way to create large training sets that can be used with supervised machine learning algorithms. Training data plays a key role in supervised machine learning algorithms and the lack of training data has been one the main barriers to using these methods to link historical records. Both Feigenbaum (2016) and Bailey et al. (2017) use a resource-intensive process that relies heavily on skilled human trainers to create their training data.

Our approach instead relies on the 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 can potentially be used to train the data to make new matches. Thus, these profiles provide a relatively low-cost way to create very large training sets with multiple sources of information attached. The training data are highly reliable, as the family members doing the linking typically have private information that can be used to identify the person of interest across multiple data sets. For example, family members are more likely to know maiden names, or to know which census record for someone with a common name belongs to their family member. Kaplanis et al. (2018) use data from one such website to "demonstrate that millions of genealogists can collaborate in order to produce high quality population-scale family trees."

The genealogy platform we use for our study is FamilySearch. FamilySearch is a large, public wiki-style family tree that includes a profile for over 1 billion deceased individuals with over 7 million people actively contributing information to those profiles. Individuals can upload

information and sources to 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, who then make a decision about 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 training set with 4.6 million 1900-1910 links, 4.9 million 1910-1920 links, and 2.9 million 1900-1920.

Our large training data set allows us to examine several important decisions that need to be made when using a machine learning approach to link historical records. These decisions include which features are used to identify matches (blocking), whether to pre-process the data (for example, by identifying common nicknames), how much training data to use, which machine learning algorithm to use, and how to evaluate the quality of the matches that are created. We use three key measures of success in record linking to guide our decision-making: the match rate (or recall), the false link rate (or precision) and representativeness (how the sample compares to the population of interest) (Bailey et al, 2017). There is generally an inherent trade-off among these different measures of success, as improving along one dimension often requires sacrificing along one or both of the remaining two.

We begin with an approach that gives more weight to maximizing recall and less weight to the representativeness of our matched sample. We then use an iterative approach that allows us to continue to match additional records to produce a sample that is more representative. For example, it is relatively easy to link people living in the same town or living with the same family members in adjacent censuses. However, both of these matching strategies create samples in which people who migrate or who do not live with family are under-represented. Our iterative approach involves removing from our sample all of the matches that we are able to do at each stage and then continue the matching process on the unmatched sample that is left over. As a result, the size of each of the

blocks that we use gets smaller and smaller, making it possible to identify more and more unique matches.

Once we have exhausted this iterative process, we provide the unmatched sample to human trainers who employ a variety of traditional family history tools to identify matches for these people. This provides a new set of training data that is based solely on linking the very hard-to-link observations. We find that our process failed to catch many of these links because their first or last name was misspelled by the original enumerator or by the person transcribing the data from the census record. Other people have a birth year that differs between the two records, which can occur with rounding or when people incorrectly assess or report their age. The machine learning approach will match people even if the information about them is reported incorrectly, but sometimes a combination of reporting errors causes the match to fall outside the parameters set by the model. These follow-up data checks provide us additional training data for hard-to-link records that we feed back into our iterative process of linking records.

We provide two illustrative examples of this entire process by focusing on two specific groups that we can approach as a self-contained linking project. Subdividing the data in this way allows us to take advantage of parallel processing and dramatically reduce the amount of time needed to link records. First, we attempt to link records for men with the surname Rockwood (or likely to be misspelled versions of Rockwood) across the 1900, 1910, and 1920 censuses. We then describe a process that could be used to group together other surnames based on the insights from our training data. Second, we link the 1910 and 1920 census records for everyone who was born in a subset of 19 states. Here, we achieve a match rate of 51 percent for our sample. We conclude by suggesting ways to measure the quality of matches generated by our approach.

I. Background

The 100 percent samples of the U.S. 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 economists to create large linked samples. These include creating pre-determined rules to identify unique matches (Ferrie 1996; Abramitzky, Boustan and Eriksson 2014; Collins and Wanamker 2014), employing a statistical algorithm such as expectation-maximization (Abramitsky, Mill, and Pérez 2018), using hand-linked data (Bailey et al. 2017; Costa et. al. 2018), or combining human-created training data with machine learning algorithms (Feigenbaum 2016; Goeken et. al. 2011). Each of these approaches have their advantages and disadvantages, and they are likely to complement each other in the common effort of linking as many individuals across historical records as possible. In this section, we describe in more detail those papers that have used supervised machine learning to link historical records, since those papers relate most to the 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) was one of the first to use a machine learning approach that is readily accessible for other researchers to use. He linked a sample of men in the 1915 Iowa Census with their record in the 1940 census and restricts the set of comparisons to pairs of records with the same birth state, born within 2 years of each other, and with similar first and last names. Specifically, names are identified as a match if they are within a Jaro-Winkler distance of 0.2, which is a measure of how similar two string variables are. 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 pair of records is labeled a match if the predicted score is above a threshold and if the second best match is below a different threshold.

Thus, a correct match is defined as one that is that has a high match score and has a significantly higher match score than any other possible matches.

Goekeven et. al. (2011) provides an earlier example of using machine learning to link census records; their approach was used to create the IPUMS Linked Representative Samples of the 1850-1930 US Censuses. They block on race, gender, and birthplace and 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 names and ages of the possible links. They also create some training data by comparing possible links that they identify with 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).

The other project that has created a large training data set is the Longitudinal Intergenerational Family Electronic Microdatabase (LIFEM) which is described in Bailey et al. (2016). They note that one of the strengths of deploying record linking algorithms on historical records (relative to modern administrative data) is that the data and code can be shared with other researchers as a way to be fully transparent about the samples and techniques. The LIFEM project has created a linked sample of individuals from birth records to census records. This training data is created through a clerical review process in which each candidate match is reviewed by two human data trainers to determine whether it is a true match or not. When there is a disagreement between the two trainers, the candidate match is re-reviewed by an additional three trainers. The training data created from this process includes 19,090 boys linked from Ohio birth records to the 1940 census and 25,352 boys linked from North Carolina birth records to the 1940 census. The final goal of the

LIFEM project is to use this training data to link together four generations across birth, marriage, and death records and the 1900 and 1940 US census.

II. 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 the father and mother. We can also identify 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 a training set of linked census records that were provided to us from FamilySearch. These matched pairs come from their online, wiki-style genealogy platform called the Family Tree. The Family Tree was created in 2001 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 7.3 million registered users who make contributions and includes over 1.2 billion individual profiles of deceased individuals.

The merge to Census records is done by FamilySearch users themselves, who find the records using publicly available sources and then attach them to each individual profile. FamilySearch provided us with a file with a personal identification number (PID) for the individual profile on the Family Tree and for the Census record that allows us to observe these matches. In addition to Census records, an individual profile could have vital statistics records, military records,

school records, city directories, places of birth and death, and date of death; we already have access to all of this information as well. 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.

This process produces a large, detailed, and highly representative training set. The data are highly reliable, 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. This type of data collection strategy has been validated in recent work by Kaplanis et al. (2018), who use data from 86 million profiles from Geni.com, a website that allows users to create individual profiles and upload family trees. The program merges individuals that are the same, pulling the individual family trees into larger combined trees, the largest of which include 13 million people. Kaplanis et al. (2018) were able to confirm the linkages in the family trees using DNA data and find that the links made by the website's users were highly reliable. They also compare the demographic patterns against traditional demographic datasets and find a strong amount of concordance.

Table 1 provides some information about the size of the training set that we will be using in this paper. 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 our training set are attached to all three census records from 1900 to 1920, while others are only linked to two of the censuses. The most common linkage type in our training set consists of those attached to just the 1910 and 1920 census, which accounts for 1.4 million women and 1.6 million men. 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. We also note that our training set is much larger than any previous training set used for linking records.

The accuracy of machine learning algorithms rely not only on the size of the training data set, but also on the quality of the data. We validate the quality of our training data in two ways. First, we compare training data created using data from the Family Tree with the links created by the human trainers working on the LIFEM project (Bailey et al, 2017). LIFEM 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 LIFEM 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 find the correct match (without having been shown the LIFEM or Family Tree match). They found that 75.4% of the time the link based on the Family Tree was correct, 26.1% of the time the LIFEM 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 LIFEM and Family Tree agree as correct reveals that the LIFEM links were correct 95.2% of the time and the links based on the Family Tree were correct 98.4% of the time. This suggests that the training data set based on pairs of links attached to individual profiles on the Family Tree provides 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 training data by having humans hand match a random sample of the records in the training set. 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 sample from the 1920. The 1910 census record that they labeled as a match for each 1920 census record agreed with the match in our training data 98% of the time. We replicated this with a random sample of 350 record links from our full training set. Of those 350 records, they were unable to find a link 6% of the time; when a link was found, they matched our training set with 99% accuracy.

The Family Tree is a public wiki-style resource so it is possible for outside researchers to use the FamilySearch API to obtain similar training data directly from the Family Tree. There are also other websites that have public trees for which data could potentially be gathered using automated approaches. The public member trees on Ancestry.com could also provide a large training set. On Ancestry, individuals curate their own family tree, without the wiki feature that the Family Tree has. The lack of a wiki is a disadvantage in that there is less information sharing, but it does mean that it is easier to identify the person who contributed the information. This could possibly be exploited to achieve a high level of accuracy in the training data, by focusing only on those public member trees that are created by professional genealogists. We are currently working with Ancestry.com to provide a training set that could be shared with academics through the same process at the University of Minnesota that provides access to the census data files.

III. Method

The process for linking census records has three steps. First, we pre-process the data to clean up obvious misspellings and abbreviations for names and places. Second, we create the features used for blocking and matching and decide on the specific blocking strategies to use. Third, we choose which machine learning algorithm to use and combine it with our training data to train the model to predict the matching records for each individual. In this section, we describe these steps in detail, and identify the important decisions that need to be made and some of the insights that we draw from our training data in making them.

Step 1: Pre-processing the data

There are three key features in our input data from the Census: 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 is 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. Thus when we link 1900 and 1910, we can use the information from 1900 to identify the age the person would be on April 15th in 1910 (with some error for those born in April).

Second, in the US census records, the most specific birthplace listed for those born in the US is the state they were born in. For those born outside of the US, 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 data in two ways. For those born in the US, we clean the spelling of each birth state to have a single standardized name. For example, in our data we find that the state

Connecticut has 97 different ways that it is spelled in the data. For those born outside of the US, we standardize the birth place to be the name of the country of birth, though certain abbreviations such as "ata," "o," and "aus," are difficult to classify.

Third, we do some cleaning to convert nicknames, abbreviations, and misspelled names to a standardized set of formal names. We start by using our training data to create a list of the most common nicknames and abbreviations. As examples, we provide a list of the 20 most common nicknames and abbreviations that we observe for both men and women in Table 2. We then report the Jaro-Winkler score for each of these pairs of names as a way to highlight the fact that employing a cutoff for the string similarity scores of 0.8 or 0.9 is likely to miss many of these matches. Our full list includes 1,704 nickname and abbreviations. We replace each of these nicknames and abbreviations with the formal name associated with it. For example, we have replaced every first name that appears as Wm in the data with William.

After cleaning the nicknames and abbreviations in our data, we observe 2.8 million unique first names. Over 99% appear less than 250 times in our data and are likely to be simply misspelled versions of other more common names. We split our set of names between those that appear more than 250 times and those that appear less than 250. We link each name in the less frequent set with a name from the more frequent set with which it has the highest level of similarity based on Jaro-Winkler scores. We require that the best match have a Jaro-Winkler score of at least 0.90 and that the next best match have a Jaro-Winkler scores less than 0.75. This is similar to the type of matching rules used in Feigenbaum when evaluating cases with multiple possible matches. This approach reduced the number of unique first names from 2.8 million in the original data to 2.65 million and reduces the number of unique last names from 6.5 million to 6.2 million. In contrast, if we were to use a phonetic coding system like Soundex to reduce the number of unique names, the final count of unique first names would be 6,550 and unique last names would be 6,628.

Step 2: 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 US during the war (Fouka 2018).

In Table 3, we use the training data for the 1900-1910 and 1910-1920 links for people living in Ohio to provide some information about the stability that we observe in different features of these individuals. The second column in this table provides the fraction of our training set for which each of the characteristics is the same for the individual between the 1900 and 1910 census and the third column does the same for 1910 to 1920. For example, only 75% of people have the exact same first name in both censuses but 94% have the same first initial of their first name. Four of the most stable characteristics of individuals are their race (99.8%), gender (99.7%), first initial of last name (98.9%), and birthplace (96.8%). Except for race, these characteristics have provided the blocking features 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 where the characteristics that are the most stable are also the least unique. The uniqueness of the characteristics

directly affects the size of the blocks. The level of consistency for specific features is also very similar across the two different pairs of years that we consider.

In Table 4, we examine combinations of characteristics that might be used as a block strategy. We focus on the blocking strategies used by Ferrie (1996), Feigenbaum (2016), and Abramitzky et al (2018). These blocking strategies are based on state of birth, gender, first initial of first name, first initial of last name, and birth year. The column labeled consistency indicates the fraction of linked pairs in our training data where the two linked records are included in the blocking strategy. 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 1, we see that the blocking strategy used by Ferrie (1996) would have included 73% of the true matches from our training set (consistency). This means that there would have been no way to link the other 27% because they would not have been included in the set of possible matches. The other two approaches perform better on this dimension, with consistency of 82.8% for Feigenbaum (2016) and 85.6% for Abramitzkey et al. (2018). However, the next two columns show the advantages 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, while the Abrimitzky et al. approach would require almost 200 times more comparisons. The number of potential matches has a direct effect on the computing time required to classify all of the possible matches. The third column indicates the number of unique matches that are identified with that set of potential matches; the Ferrie strategy produces the most unique matches. The results are similar in Panel B, which uses matches between the 1910 and 1920 censuses. The main point we want to illustrate is the block strategy chosen can have a dramatic effect on computing time and the results in Table 4 highlight this natural trade-off between consistency and computing time.

The main variables that we will use for matching census records will include gender, race, name, birth place and birth year. From these variables, we can construct multiple features based on each of these variables. For example, we include 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 include indicators for the similarity of birth year, birth place, race, and gender. We also do some matching on additional features including mother and father's birthplace, the names of family members, and place of residence.

Step 3: Choosing the Machine Learning Algorithm

We chose to create our model using XGBoost, 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 in addition to 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 not normal.

Table 5 provides three key metrics for five of the machine learning algorithms that we examined. We chose XGBoost because it outperformed other commonly used classification models on metrics such as precision, recall, and processing time. Since all of the metrics for precision or

recall are above 90%, it may seem that the classifiers all have similarly high performance. It is helpful also to consider the flip of each of these metrics. Subtracting the precision from 1 provides the proportion of false matches that are created; XGboost creates about half as many false positives at Random Forest (3.5% vs. 7.6%). Subtracting the recall from 1 provides the proportion of true matches that we miss and in this case; Gradient Boosting misses about 4 times as many possible matches as XGBoost (8.1% vs. 1.9%). Finally, since linking the census records together will require a few billion comparisons, processing time can become very important and in this case, XGBoost is more than 20 times faster than Gradient Boosting.

IV. Application

Our aim is to produce a linked dataset of individuals across the 1900, 1910 and 1920 US Censuses. We start by providing some analysis of the number of matches that should exist between these census records which provide a benchmark to use when evaluating match rates. Next, we provide an illustrative example using a subset of the data – people with a particular surname. Then we provide some statistics about our match rates for linking the 100% sample of the 1910 and 1920 census.

Calculating the number of potential matches

One way to visualize the matching process is as a giant matrix in which we are trying to combine rows that are duplicates of the same person. Figure 2 provides an illustrative example of how this works. Suppose that our sample includes three people who each appear in the 1900, 1910, and 1920 censuses. Suppose also that for one of those people (PID1), a contributor has already linked their 1900 and 1910 census records, and for another (PID2), a contributor has linked their 1910 and 1920 records. These two pairs of census links would provide the training data for our

machine learning algorithm in this simple example. In practice, we would use training data across multiple blocks and have a much larger training set to work with.

The other rows in this data are the possible census records for the three people for each of the three census years. Thus our starting dataset has 7 rows and 3 unique people. Each time we match a pair of censuses, we reduce one of the rows. For example, if the census records B and H are a match, then record H will move into the open slot for PID1 and the row labeled ID4 would be removed. Once all of the records have been matched, we will end up with our final dataset that has three rows and a record for each person in each census year as shown in the bottom part of Figure 2. Since each match removes a row, this indicates that in order to fully match this sample, we will need to identify 4 matched pairs. This simple example illustrates how we can calculate the number of possible matches by subtracting the number of unique people in our sample from the number of unique records.

We can use information from the censuses to figure out how many unique individuals are in our sample. Table 6 provides the number of records that appear in each census year for all of the censuses from 1850 to 1940. It also includes a column for the number of people that census that would not have been present in the previous censuses based on their birth year or immigration year. This provides a measure of the number of new additions to the census sample each census year. The numbers in Table 6 indicate that there are 76.2 million people in the 1900 census. An additional 27.5 million people joined the sample in the 1910 census and an additional 28.8 million people joined the sample in 1920. This provides a total sample of 132.5 million unique individuals that should appear in at least one of these three census years. The total number of records across these three census years is 274.9 million records. Thus, we can potentially create 142.4 matches from these three censuses. This is the overall number that we will use as a benchmark to determine our overall match rate.

We use a similar approach to determine the total number of possible links between each pair of censuses. Between 1900 and 1910, there are 168.4 million total records (76.2 million + 92.2 million) and 103.7 million unique people (76.2 million + 27.5 million). This means that there should be 64.7 million matched pairs between the 1900 and 1910 census. We can also subtract the number of matched pairs from the number of people in 1900 to get the number of people in our sample who died between 1900 and 1910, which is 11.5 million. For 1910 to 1920, there should be 77.7 matched pairs between 1910 and 1920 and 14.5 million people who died between 1910 and 1920. Using death rates provided by the National Vital Statistics System and population estimates from the US Census Bureau, we estimate there were approximately 13 million deaths between 1900 and 1910 and approximately 14 million deaths between 1910 and 1920. These are close to our estimates for the number of deaths in these decades, which provides some external validity to our approach.

Illustrative Example – Last Name

To demonstrate our process using a smaller sample, we begin with every male in the 1910 and 1920 censuses with the last name Rockwood. This sample includes 1639 individuals from 1910 and 1728 from 1920. Based on the birth and immigration years of each person, 1421 of the Rockwoods in 1920 should have a match in 1910. The training data that we began with from FamilySearch includes 11.7% of the possible 1910-1920 matches.

We use the XGboost model that was trained on our full training sample and then apply that trained model to our sample of Rockwoods. This approach identifies an additional 570 of the possible 1910-1920 matches bringing our total (including the training data) up to 636 matches or a 44.8% match rate. We then incorporate additional information about family members in the household and compare individuals within households that share an already-matched pair or have similar last names, keeping matches from households where at least 2 individuals matched on

approximate name and age, with an additional check for their relationship to each other. This family-based matched allows us to identify an additional 201 of the possible 1910-1920 matches bringing our final match rate up to 59%. In the next section, where we created predictions for the full sample of the US census, we don't include this step of family-based matching but it illustrates the degree to which multiple matching methods can be combined to get an even higher overall match rate.

After these steps, we end up with 484 Rockwoods in the 1920 census for which we could not find a match in the 1910 census. We provided the unmatched records to a set of research assistants trained in family history to identify the factors that did not allow our machine learning algorithm to identify the match. These hand-linked matches were generally very difficult to make and did not lend themselves to straightforward rule-based or pattern-based methods, and many required outside verification using records or public member trees on Ancestry.com.

Full Sample Match Results

In Table 7, we present the match rates that we were able to obtain for the full sample of the 1910 and 1920 census. We were able were able to identify 41.8 million matches between the 1910 and 1920 census, providing an overall match rate of 51.2%. We also split this data into mutually exclusive groups based on 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. Using the second year as the base sample for estimating match rates is important because it accounts for mortality since not everyone present in the 1910 census will still be alive in 1920. For this reason, we also exclude anyone in the 1920 census who list an age of 10 or younger or who immigrated to the United States in 1910 or later.

Table 7 also provides the match rate and number of matches based on the individual's gender and relationship to the household head in 1920. The groups with the highest match rates are sons (65.4%) and daughters (63.4%). 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. The group with the next highest match rate are male head of households (54.5%) followed by the spouse of the head of household (47.3%). The lowest match rates occur for other household members who are not part of the immediate nuclear family where the match rates fall to 31.4% for women and 26.2% for men.

V. Conclusion

Recent developments in data access and record linking methodology have created exciting opportunities for social science research using large populations. We contribute to this work by developing a unique way to create training data that has been largely untapped by the research community. This training data is created from the contributions of millions of individuals who are investigating their own family histories. This quality of the training data is very high, as these family members often have private information that allows them to make accurate matches that other methods cannot. This new method can generate training sets that are several orders of magnitude larger than previous training sets and require fewer resources (person-hours and computing time) to create. In addition, this approach can complement other record-linking approaches since we can use the links identified from other methods as part of the set of possible matches that we score with our supervised machine learning approach.

For just the three years of Census records that we use in this paper, we are able to create a training set with 12.6 million matched pairs. We use this new training data to provide several insights that will be helpful for advancing the state of art in machine-based records linking. First, we provide

a crosswalk of common nicknames and abbreviations that are used in US historical records. This crosswalk can be used to pre-process data to improve the accuracy of match features based on string distances. Second, we document the individual-level characteristics that are the most stable over time and highlight how the choice of characteristics that have to exactly match can influence the possible match rate. Third, we use this large training data to evaluate different machine learning algorithms and find that XGBoost, a variation of Random Forests, has the highest levels of precision and recall in this context and is also relatively fast in terms of computation time (more than 5 times faster than a standard Random Forest model).

As the approach that we describe in this paper is developed and refined, it will be important to measure the quality of the matches that are generated. We propose four ways for doing this. First, we could compare the links that the model predicts with other projects that have developed training data for comparing census records. Two notable examples include LIFEM and the Early Indicators Project; the LIFEM process has not yet constructed census-to-census links but we plan to compare the links from our approach with those once they become available. Second, we can draw a random sample of our matches and provide them to research assistants trained in family history and compare the matches they create using those time-intensive methods compared to what we find using machine learning. Third, 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. Fourth, we can share our predicted matches with the FamilySearch record hinting system and then observe what large groups of users decide about the predictions we make. FamilySearch has a system for emailing individuals using their platform about possible record hints for individuals that they are related to; we plan to work with FamilySearch to

create a specific email campaign in which we share predicted links that we have identified with experienced users, and then observe whether the user decides to attach the record or indicate that it was not a match.

One advantage of the source of training data that we use is that the same process could be applied to any two types of records that are available on various genealogical platforms. As such, training data could be created with the same approach for link vital records, military records, school records, and Census records. In addition, by collaborating with the family history research community we can identify additional ways to validate the quality of our linked samples. It is likely that this integration of family history and machine learning will be instrumental in efforts to create a fully linked sample across Census records of everyone that lived in the United States between 1850 and 1940.

References

Abramitzky, R., L. Platt Boustan and K. Eriksson (2014). "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122(3): 467-506.

Abramitzky, R., Mill, R., & Pérez, S. (2018). "Linking Individuals Across Historical Sources: a Fully Automated Approach." *National Bureau of Economic Research Working Paper*, 24324.

Bailey, M. J., S. Anderson, A. Karimova and C. G. Massey (2016). "Creating Life-M: The Longitudinal, Intergenerational Family Electronic Micro-Database."

Bailey, M., Cole, C., Henderson, M. & Massey, C. (2017). "How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth." Tech. Rep,National Bureau of Economic Research.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). Mobility Report Cards: The Role of Colleges in Intergenerational Mobility. *National Bureau of Economic Research Working Paper*, 23618.

Chetty, R. and Hendren, N. (2018). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *Quarterly Journal of Economics*, forthcoming.

Collins, William J. and Marianne H. Wanamaker. (2014). "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics* 6:220–252.

Costa, D. L., Kahn, M. E., Roudiez, C. & Wilson, S. (2018). Data set from the Union Army samples to study locational choice and social networks. *Data in Brief,* 17, 226–233.

Evans, M. F., Helland, E., Klick, J., & Patel, A. (2016). The Developmental Effect of State Alcohol Prohibitions at the Turn of the Twentieth Century. *Economic Inquiry*, *54*(2), 762-777.

Feigenbaum, J. J. (2016). Automated Census Record Linking: A Machine Learning Approach. *Working Paper*.

Feigenbaum, J. J. (2018). "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940." *Economic Journal*, forthcoming.

Ferrie, Joseph P. (1996). "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* 29:141–156.

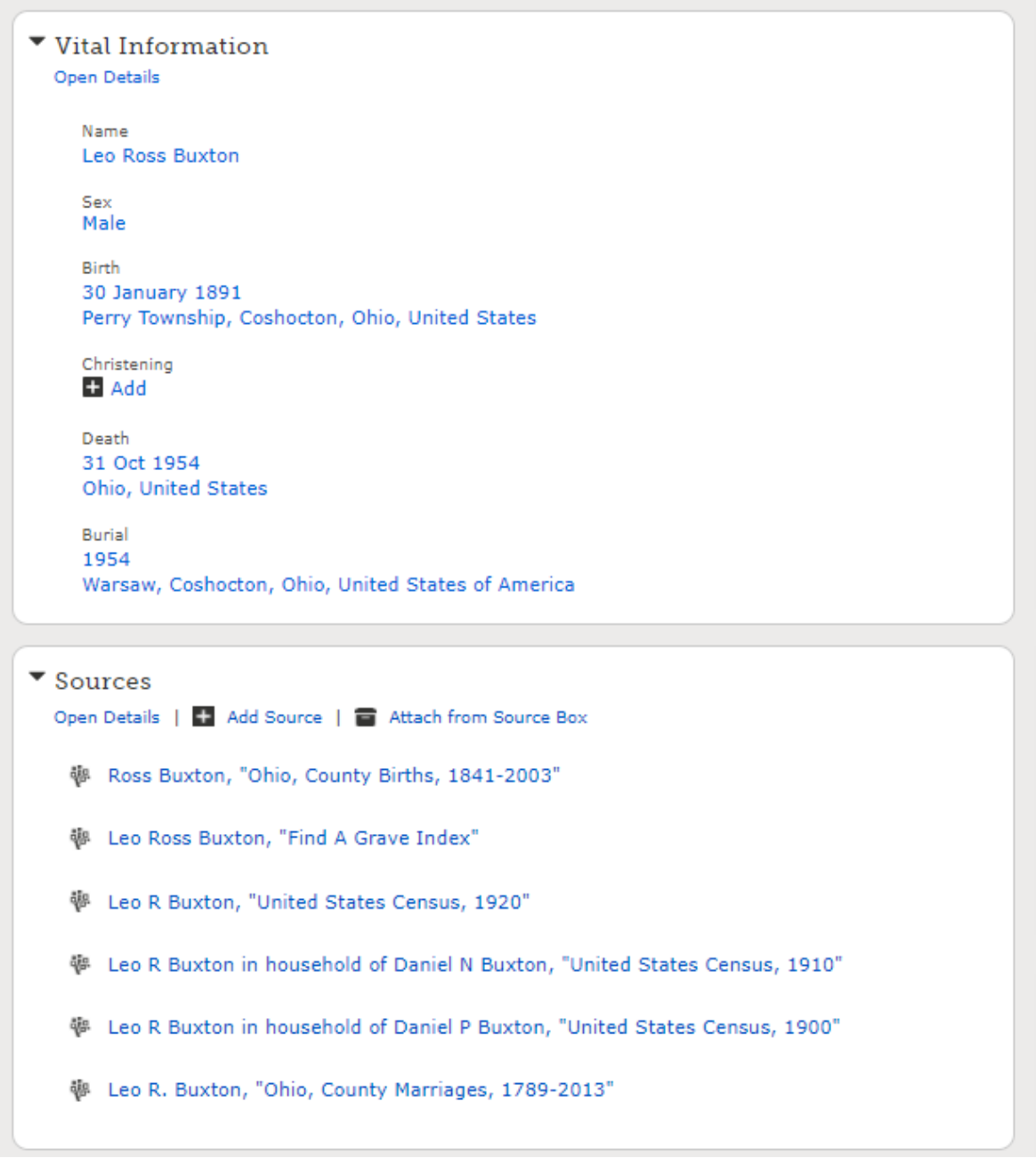
Goeken, Ron, Lap Huynh, Thomas Lenius, and Rebecca Vick. (2011). "New Methods of Census Record Linking." *Historical Methods* 44:7–14.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., and Bhatia, G. (2018). Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives. *Science*,

360(6385), 171-175.

Mill, Roy and Stein, Luke. (2016). "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America." *SSRN Working Paper* 2741797.

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



Notes: This is an example of what an individual profile page would look like on FamilySearch. There is a section that includes vital information about the person (name, birth, and death) and then a separate 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. Illustrative example of linking census records

A. Initial dataset with training set marked in blue

ID	Ark1900	Ark1910	Ark1920
PID1	A	B	
PID2		C	D
ID1	E		
ID2	F		
ID3		G	
ID4			H
ID5			I

B. Final dataset with all matches completed

ID	Ark1900	Ark1910	Ark1920
PID1	A	B	H
PID2	E	C	D
ID2	F	G	I

Notes: The example above is a case in which we have three unique individuals who each appear in the 1900, 1910, and 1920 census. The pairs marked in the blue boxes are individuals (PID1 and PID2) that already had linked census records on the Family Tree and would be the training data that we would use for the machine learning algorithm. The other rows are unmatched records. Each time a match is identified, a row is removed from the original table. The final product has a single row for each unique individual.

Table 1. Characteristics of the Training Set

	Female	Male
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: 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 our training set. For example, the first row provides the number of men and women in our training set that are matched to just the 1910 and 1920 census.

Table 2. Frequently Used Nicknames and their Associated Common Names.

Male Names			Female Names		
Full Name	Nickname	JW Score	Full Name	Nickname	JW Score
William	Wm	0.593	Anna	Annie	0.848
Charles	Chas	0.808	Elizabeth	Lizzie	0.704
Joseph	Joe	0.867	Margaret	Maggie	0.778
George	Geo	0.883	Lillian	Lilly	0.874
Frederick	Fred	0.889	Rosa	Rosie	0.848
Charles	Charlie	0.943	Lillian	Lillie	0.910
Samuel	Sam	0.883	Caroline	Carrie	0.874
William	Willie	0.910	Catherine	Kate	0.694
William	Will	0.914	Sarah	Sadie	0.680
James	Jim	0.720	Lillian	Lily	0.808
Thomas	Tom	0.850	Susan	Susie	0.813
Fredrick	Fred	0.900	Katherine	Kate	0.870
Henry	Harry	0.760	Jane	Jennie	0.675
Francis	Frank	0.874	Catherine	Katie	0.665
Edward	Ed	0.822	Harriet	Hattie	0.797
Benjamin	Ben	0.854	Martha	Mattie	0.733
Thomas	Thos	0.922	Jane	Jenny	0.670
Robert	Robt	0.922	Katherine	Katie	0.850
Alexander	Alex	0.889	Mary	Mollie	0.525
John	Jno	0.750	Sarah	Sallie	0.662

Notes: Each panel provides the 20 most commonly used nicknames or abbreviations for both men and women. The JW score is the Jaro-Winkler measure of the similarity of two string variables.

Table 3. Stability of Blocking Features between 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
Given Name	0.494	0.511	261,028	263,991	257,166

Notes: 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 our training data.

Table 4. Stability of Combinations of Features and Resulting Block Size

A. 1900 and 1910 Censuses

Blocking Strategy	Consistency	Potential Matches	Unique Match
Ferrie (1996)	0.730	7.4	516,952
Feigenbaum (2016)	0.828	10.5	373,695
Abramitzky et al. (2018)	0.856	1,369.3	4,559

B. 1910 and 1920 Censuses

Blocking Strategy	Consistency	Potential Matches	Unique Match
Ferrie (1996)	0.761	6.0	657,044
Feigenbaum (2016)	0.861	8.6	459,963
Abramitzky et al. (2018)	0.877	1,323.9	2,316

Notes: The first column indicates the fraction of true matches for which all of the characteristics used in each blocking strategy that 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 unique matches that occur within a block. he sample size for panel A is 2.63 million and the sample size for panel B is 2.85 million.

Table 5. Performance Measures of Classifiers

Model	Precision	Recall	Processing Time (seconds)
XGBoost	96.52%	98.15%	13.90
Random Forest	92.12%	97.99%	71.17
Gradient Boosting	96.32%	93.83%	311.89
Logit Regression	93.84%	95.92%	3.41
Neural Nets	93.63%	95.94%	9.38

Notes: The performance measures are based on creating a test set from our training set. Precision is the fraction of identified matches that are true matches. Recall 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 300,000 comparisons.

Table 6. US Population and New Additions in Each Census Year

	US Population	New Additions
1850	23.2	23.2
1860	31.4	10.9
1870	38.6	12.9
1880	50.2	16.4
1900	76.2	41.4
1910	92.2	27.5
1920	106.5	28.8
1930	123.1	29.3
1940	132.1	26.8
Total	673.5	217.2

Notes: The US population column indicates the number of individual-level records that are included in each of the decennial censuses from 1850 to 1940. The new additions column includes everyone in that year who was not present in the previous census based on their birth year or immigration year. All numbers are measured in millions.

Table 7. Full Sample Match Rates by Relationship Type

	Possible matches	Match Rate	Matches
Overall	81,685,065	51.2%	41,814,537
Sons	12,872,481	65.4%	8,424,717
Daughters	11,958,839	63.4%	7,586,881
Male Head of Household	21,249,895	54.5%	11,583,612
Spouse	19,158,935	47.3%	9,063,319
Female Head of Household	2,976,290	44.7%	1,329,547
Other Females	6,046,456	31.4%	1,899,808
Other Males	7,380,580	26.2%	1,940,091

Notes: The estimates above are based on linking the entire 1910 and 1920 censuses. Possible matches in the 1920 census for these states excludes anyone who was born in 1910 or later or immigrated to the US in 1910 or later.