Combining family history and machine learning to link historical records: The Census Tree Data Set[†]

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 records in a way that allows investigators to observe people at different points in their life or across generations. In this paper, we contribute to recent efforts to create these links with a new approach that relies on millions of record links created by individual contributors to a large, public, wiki-style family tree. We use these "true" links both to inform the decisions one needs to make when using automated methods to link records and as a training data set for use in a supervised machine learning approach. We describe our procedure and illustrate its potential by linking individuals across the 100% samples of the US censuses from 1900, 1910, and 1920. When linking adjacent censuses, we obtain an overall match rate of 62-65 percent (for over 88.9 million matches), with a false positive rate that is around 6-7 percent and with links that are similar to the population along observable characteristics. Thus, our method allows us to link records with a combination of a high match rate, precision, and representativeness that is beyond the current frontier. Finally, we demonstrate the potential of the data by estimating the degree of intergenerational transmission of literacy among father-son and mother-daughter pairs.

Keywords: record linking, genealogy data, machine learning, intergenerational transmission

JEL Codes: N01, N11, N12, C8

[†] 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 and at Notre Dame, including Ben Branchflower, Ben Busath, Alison Doxey, Neil Duzett, Laren Edwards, Brianna Felegi, Nicholas Grasley, Adrian Haws, Brandon Ly, Amanda Marsden, Jalen Morgan, Daniel Sabey, and Joseph Young.

* Corresponding author: kbuckles@nd.edu; 3060 Jenkins-Nanovic Halls, Notre Dame, IN 46556.

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 experiences, we typically need to observe a person as both a child and as an adult. Unfortunately, this type of data has been hard to come by in the United States, since census data and many administrative data sets lack a consistent individual registration number (as exists, for example, in Sweden and Norway).

Recently, researchers studying the U.S. have created large linked samples in three ways. One approach is to employ restricted-use data with unique individual identifiers that permit 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). Another strategy is to link individuals across records by matching on characteristics such as the person's name, birth year, and birthplace (Ferrie 1996; Abramitzky, Boustan, and Eriksson 2014; Evans et al. 2016; Abramitzky, Mill, and Pérez 2018). A third approach is to use supervised machine learning algorithms. Intuitively, machine learning methods use a training set with examples of both correct and incorrect matches to "learn" which features of the data best predict a correct match. This information can be used to create an algorithm to identify new matches (Christen 2012; Feigenbaum 2016; Folkman, Furner, and Pearson 2018).[2] 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.

[2] For an introduction to machine learning methods, see Müller and Guido (2016) and Mullainathan and Spiess (2017).

In this paper, we propose a new approach for linking individuals across historical records that builds on and incorporates many of these other methods. Our unique contribution arises from our use of a data set created from decisions made by millions of people doing research on their own family histories. These researchers often gather source documents—including census records—to establish various life events and relationships for a family member, 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 "true" match.[3] These matches can be used to inform the decisions made when employing various linking strategies and can also be used as training data for supervised machine learning methods. The data are highly reliable, as many of the links are created by family members who have a personal interest in making a correct match. Furthermore, these family members often 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 information 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 decide whether or not the source should be attached to that person. We use a sample of

[3] 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.

individuals from this family tree that are attached to at least two census records between 1900 and 1920. This provides a data set with 14.5 million 1900-1910 links, 16.3 million 1910-1920 links, and 9.8 million 1900-1920 links.

We describe a process that uses 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, and which machine learning algorithm to use. We show how key properties of the data—precision, recall, and representativeness—respond to these choices, and we demonstrate the potential for transfer learning with our algorithm. Second, we 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 61.6% of the potential matches between the 1900 and 1910 full-count US censuses, and 65.2% of the potential matches between the 1910 and 1920 full-count US censuses (or 38.8 and 50.1 million matches, respectively).

In Section V of the paper, we summarize the properties of the Census Tree data set and compare it to other linking methods and efforts. First, we show that people who are linked to a prior census in the Census Tree are similar to the full population in terms of gender, age, household characteristics, and occupation score, but that white Americans and those who were born in their birth state are over-represented. We also hand-check a random sample of our predicted matches and use a transitivity test to show that the false positive rate among our predicted matches is about 7%. Prior work has documented a "production possibilities frontier" that shows the tradeoff between recall (finding more matches) and precision (avoiding false positives) (Abramitzky et al. 2019); we show that when combining our methods with those of others, we achieve a combination of these two qualities that is beyond the frontier achieved by the groundbreaking Census Linking Project, or CLP (Abramitzky, Boustan, and Rashid 2020). Moreover, our method adds tens of millions of new links among these censuses, including many links among women (who are excluded entirely from the CLP) and minority groups.

As an application to demonstrate the potential of the data, we produce estimates of the intergenerational transmission of literacy between parents in the 1900 Census and their adult children in the 1920 Census. We are able to do this separately for Black and white Americans, *and* for both father-son and mother-daughter pairs. We show that our estimation samples are larger than those currently available from the CLP and are more representative of the full population than either the FamilySearch links alone or the CLP links. The estimates suggest that the greater precision of the Census Tree links works to mitigate the well-known problem of attenuation bias from measurement error due to incorrect links (Solon 1992).

Ultimately, our goal for this project is to create every possible link among the full-count US decennial censuses from 1850 to 1940, and to make these links available to other researchers.[4] But the method we describe in this paper could be applied to any pair of data sets for which there is a sufficient number of individuals with records in both collections linked by users on a genealogy platform. Moreover, the potential for transfer learning from the census links will likely aid efforts to create new links among these data sets. As a result, the potential of the methods introduced in this paper 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.[5]

[4] The effort will extend past 1940 as later censuses become available under the Census Bureau's 72-year release policy. We describe our plans for sharing the data and code in Section VII.

[5] In the case of FamilySearch, outside researchers can use the FamilySearch API to obtain links directly from the Family Tree. We describe the process for accessing the FamilySearch API in the FAQ document in Appendix D (also available at https://sites.google.com/view/family-tree-faq/home).

II. Background

The 100 percent samples of the US decennial censuses are made available to the public after 72 years, which creates the possibility for linking 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 (Abramitzky, Mill, and Pérez 2018; Pérez 2019), using hand-linked data (Costa et al. 2018), and combining human-created training data with machine learning algorithms (Feigenbaum 2016; Goeken et al. 2011; Bailey et al. 2019). Bailey et al. (2019) and Abramitzky et al. (2019) provide detailed summaries of these approaches and have posted helpful code for others to use.

Most recently, Abramitzky, Boutstan, and Rashid (2020) have published the links they have created among the 1850 to 1940 censuses on a website that allows users to download crosswalks for pairs of census years. Users then acquire the census data from IPUMS or other sources, which contain an identification code (a "histid") that can be used to identify the links. Crosswalks are available for four different variations on the linking method developed by Abramitzy, Boustan, and Eriksson (2014). This effort, known as the Census Linking Project (or CLP, available at censuslinkingproject.org), is a tremendous service to the social science community, and has the potential for use in a wide range of research projects. Below, we show how our data have the potential to add many more links to this effort, including links for women, who are currently excluded from the CLP and many similar projects. We also compare the precision, recall, and representativeness we are able to obtain when we use our machine learning method (alone or in combination with other approaches) to those obtained in the CLP data.

We now describe in more detail two papers that have used supervised machine learning to link historical records, as their approaches are most closely related to ours. 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 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 between the two censuses 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, which is a metric for measuring the similarity of two text strings based on the number of matching characters and the order in which they appear). He creates 17 features 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 with a match probability that is sufficiently higher 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. They identify potential matches as those that have the same race, gender, and birthplace, and birth years within seven years of one another. They use 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 98,330 matches for men between the 1880 complete-count census and the 1% samples of the 1850-1930 censuses, and 41,762 matches for women.

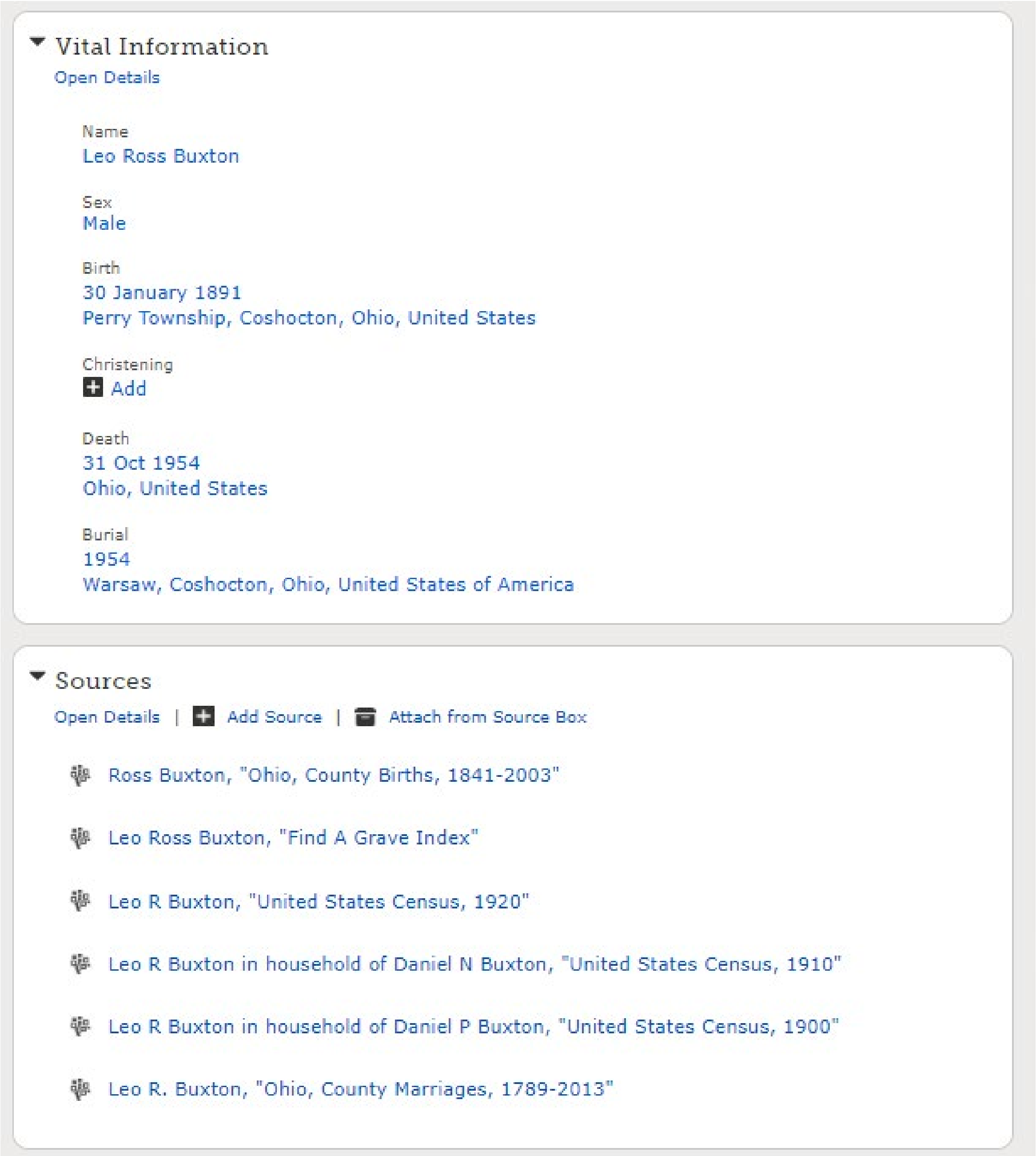
III. Data

We use two sources of data for this project. The first data set 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 and include characteristics such as the person's name, birth year, birthplace, gender, race, place of residence, and the birthplaces 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 are organized by household, allowing us to observe family relationships and characteristics of household members.

The second data set is comprised of linked census records that were provided to us by FamilySearch. These matched pairs come from FamilySearch's online, wiki-style genealogy platform called the Family Tree. The Family Tree 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 census records on the FamilySearch platform and attach them to an individual's profile. An individual profile could also include attached sources such as vital records, military records, school records, and city directories. We include an example profile in Figure 1 to

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



Vital Information

Open Details

Name
Leo Ross Buxton

Sex
Male

Birth
30 January 1891
Perry Township, Coshocton, Ohio, United States

Christening
⊞ Add

Death
31 Oct 1954
Ohio, United States

Burial
1954
Warsaw, Coshocton, Ohio, United States of America

Sources

Open Details | ⊞ Add Source | Attach from Source Box

Ross Buxton, "Ohio, County Births, 1841-2003"

Leo Ross Buxton, "Find A Grave Index"

Leo R Buxton, "United States Census, 1920"

Leo R Buxton in household of Daniel N Buxton, "United States Census, 1910"

Leo R Buxton in household of Daniel P Buxton, "United States Census, 1900"

Leo R. Buxton, "Ohio, County Marriages, 1789-2013"

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).

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.[6] 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 this data 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 Tree data provide 40.6 million true matched pairs across the three combinations of census years (1900-1910, 1910-1920, and 1900-1920).[7] 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

[6] 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.

[7] Modern administrative records provide additional ways to generate these type of large training sets to provide key insights about automated matching algorithms. Gross and Mueller-Smith (2020) is an excellent example that uses biometric identifiers to generate a massive training set for linking administrative records from the criminal justice system.

Table 1. Size of the Family Tree Data Set

	Women	Men	Total
Only 1900 & 1910	3,414,671	3,545,823	6,960,494
Only 1910 & 1920	4,249,552	4,538,200	8,787,752
Only 1900 & 1920	1,047,899	1,199,242	2,247,141
1900 & 1910 & 1920	3,636,123	3,905,109	7,541,232

Notes: The table shows the number of links in the Family Tree data set, which are record matches made by FamilySearch users. Each of the rows 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 the 1900 and 1910 census in the Family Tree, but not the 1920 census.

poor performance of conventional name-matching methods for linking women's records has meant that women have been completely omitted from some important research.[8]

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. LIFE-M provided us a set of 54,000 people that they had linked from an Ohio birth certificate to the 1940 census. We were able to find about 12,000 people from their sample that were attached to both an Ohio birth certificate and the 1940 census on the Family Tree. Of these, 1,060 links were identified by both LIFE-M and the Family Tree, and we found that that the links agreed 94% of the time. We then took the few cases where there was disagreement and asked research assistants to use traditional genealogy tools to determine which match was correct, without knowing the source of the match. They found that 75% of the time the link based on the Family Tree was correct, 26% of the time the LIFE-M link was correct, 4% of the time they were both right (because the individual

[8] While this paper focuses on the process for using the Family Tree data to create additional matches, we note that the Family Tree links *alone* constitute an incredible data set, with millions of links among census records that include hundreds of thousands of matches between women before and after marriage. We discuss this further in Sections V and VI below. One notable example is Feigenbaum and Gross (2020) who use Family Tree links to connect women across the 1920-1940 census records in order to study the impacts of an automation shock on telephone operators.

showed up twice in the 1940 census), and 4% of the time neither of the links were correct.[9] Adding the links the research assistants identified as correct to those where LIFE-M and the Family Tree agree, we conclude that the LIFE-M links were correct 95% of the time and the links based on the Family Tree were correct 98% of 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. These research assistants used the search tools on Ancestry to identify the potential matches for that person in the 1910 census and to choose the correct match. 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. [10]

[9] These numbers do not add up to 100% since they are not mutually exclusive groups. The 4% of the time when they were both right occurred when the person showed up in more than one census record in the same year. This occurs when a person moves while the census is being taken and shows up in two places or when a person appears as a family member in one household and a servant or boarder in another household. It also occurs when the exact same family is enumerated in the census twice in the same place.

[10] Kaplanis et al. (2018) validate web-based genealogy data from Geni.com, a genealogy website that is similar to FamilySearch. 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).

IV. Method

We begin by describing how we use the Family Tree data as training data for our supervised machine learning methods, and 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 then describe the full pipeline that we use to link census records, which includes both supervised and unsupervised methods.

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 it 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 birthplace to be the name of the country of birth, though certain abbreviations such as "ata" and "o" are difficult to classify.

Third, we clean the data by converting 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 nicknames and abbreviations that we use to convert into a formal name equivalent.[11] We also help address common misspellings and transcription errors by employing Jaro-Winkler distances and NYSIIS classifications (an algorithm that classifies text strings based on their pronunciation).[12]

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 first 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 is recorded incorrectly 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 birthplace in response to the discrimination occurring in the U.S. during the war (Fouka 2018).

[11] As an example, if a person's name is Joseph in one record and Joe in another, that will create a link between those two names. If we see that same combination of names across two records for another person the strength of the link will increase. We end up with very strong links between common name/nickname pairs like Joseph/Joe, since that occurs many times; unusual pairings that are likely family-specific would have weak links (e.g. Eldrick Woods was given the nickname "Tiger" by his family but the Eldrick/Tiger pairing is very rare). Our list consists of the strongest links.

[12] See Massey (2017) for an analysis of the tradeoff between linkage rates and accuracy when using Jaro-Winkler distances in matching. Among phonetic matching algorithms, NYSIIS has been shown to be more accurate than Soundex and several other name matching methods (Snae 2007).

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. Here, we take advantage of the fact that the users have used their private information to verify that the records are a match. 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 68% of the user-verified links have the exact

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

Feature	Stable 1900-1910	Stable 1910-1920	1900 # of Unique Values	1910 # of Unique Values	1920 # of Unique Values
Race	99.79	99.36	4	4	4
Sex	99.70	99.36	2	2	2
Birthplace	97.70	97.45	69	69	69
Birth year within 3	96.62	97.12	-	-	-
Birth year within 2	94.54	95.51	-	-	-
Last initial	92.29	92.46	26	26	26
Last JW > 0.8	92.25	92.11	-	-	-
First initial	88.03	92.32	26	26	26
Mother's birthplace	87.91	87.58	69	69	69
Last JW > 0.9	87.69	87.35	-	-	-
Father's birthplace	87.55	87.20	69	69	69
First JW > 0.8	87.09	88.32	-	-	-
Birth year within 1	85.96	89.94	-	-	-
Last NYSIIS	84.15	84.76	106,713	111,690	117,275
First JW>0.9	79.90	80.75	-	-	-
First NYSIIS	77.01	79.10	53,654	48,045	37,984
Last name	73.47	74.90	365,522	384,423	408,780
First name	67.80	69.98	249,336	250,983	243,868
County	67.47	75.51	1,600	2,909	2,899
Township	31.56	55.92	18,126	17,791	20,769

Notes: Data are from the full Family Tree data set (14.5 million observations for 1900-1910 and 16.3 million observations for 1910-1920). Features where the number of unique values is not reported are features where the values are binary (0 or 1). Stability is based on record pairs where the feature for each row in not missing in either record. Features are sorted on the stability of the 1900-1910 links.

same first name in both records, but 88% have the same first initial of their first name. Three of the most stable characteristics of individuals are their race (99.8%), sex (99.7%), and birthplace (97.7%). Values for the 1910-1920 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 look at the blocking strategies used by Ferrie (1996), Abramitzky et al. (2014), and Abramitzky, Mill, and Pérez (2020), 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 row 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.

For the 1900-10 matches, we see that the blocking strategies used by Ferrie (1996) and ABE (2014) would have included about 63% of the true matches from the Family Tree data. This means that there would have been no way to link the other 37% of the sample because they would not have been included in the set of possible matches. The Abramitzky et al. (2020) approach performs better on this dimension, with 84% of the true matches included in the blocking strategy. However, the next row shows the advantage of the first two approaches, which require far fewer comparisons to be computed when linking the 1900 and 1910 censuses. The number of potential matches has a

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

Blocking Strategy	Ferrie (1996)	Abramitzky et al. (2014) (ABE)	Abramitzky et al. (2020)
Key Features	1st 4 letters of first name NYSIIS of last name Birth year within 5 yrs.	NYSIIS of first name NYSIIS of last name Birth year within 3 yrs.	1st letter of first name 1st letter of last name Birth year within 5 yrs.
1900-1910			
Consistency	0.627	0.625	0.842
Potential Matches	7.59	2.56	1,440.51
1910-1920			
Consistency	0.651	0.655	0.846
Potential Matches	7.35	2.20	1,458.95

Notes: The analysis is performed on the full Family Tree data set. Each of the methods include race, birthplace, and gender as blocking features. Consistency indicates the fraction of true matches for which all of the characteristics used in each blocking strategy are the same across the two records. "Potentials matches" is the average number of potential matches across censuses for each individual.

direct effect on the computing time required to create predicted scores for all of the possible matches. The results are similar when making 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, birthplace, birth year within 3 years, and the NYSIIS values for the first and last name.[13] 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 NYSIIS value of the first name matches, and the Jaro-Winkler

[13] We use NYSIIS to improve computational efficiency, as the blocking strategies listed in Table 3 that do not use NYSIIS or Soundex (similar to NYSIIS) in blocking have much larger blocks and are much more computationally intensive. Ferrie and ABE both use NYSIIS.

similarity score of the first name. We construct similar measures for the last name. We also create indicators for the similarity of birth year, birthplace, race, and gender. Finally, the machine learning algorithms can include information on the similarity of the mother and father's birthplace, and place of residence. In total, we use 12 features in our machine learning algorithm.

C. Choosing the machine learning algorithm

There are many different supervised machine learning algorithms available for linking records. Ideally, an algorithm performs well on three dimensions. First, most of the identified matches should be true matches (often referred to as precision in the machine learning literature). Second, many true matches are identified among those that are possible (often referred to as recall). Third, it should be computationally fast to train. 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 as training data. We see that XGBoost

Table 4. Performance Measures of Classifiers

Model	Precision (%)	Recall (%)	Training Time (min)
Gradient boosting	88.22	85.10	1.95
XGBoost	88.55	84.70	0.13
Neural nets	88.58	83.03	3.60
Random forest	86.56	83.59	5.58
Logit regression	85.69	75.62	0.12

Notes: The analysis is performed on a subset of the Family Tree data. Precision is the fraction of identified matches that are true matches. Recall is the fraction of possible true matches that were identified. Training time represents the time required to train the classifier. We use a data set of 205,550 pairs between the years 1900 and 1910, of which 176,612 were false pairs and 28,938 were true pairs. Each reported value in the table is an average from 50 splits of this data set, where 98% of the data was used to train the model and 2% was used to compute precision and recall measures.

performs similar to gradient boosting in terms of both precision and recall but is much quicker to train. We therefore choose it as our classifier.[14]

One drawback of the XGBoost method is that it is not as transparent as (for example) the logit regression method, where the coefficients on the various linking features have clear interpretations that help the researcher understand which pieces of information are being used. We can, however, identify how important a feature is in the XGBoost algorithm by counting the number of times it is used across the different decision trees that are created. In Table 5, we show the ten most important features used by our XGBoost algorithm. The difference in birth years is the most important feature, and the second is the distance between the place of residence in the two records. Four of the other eight are different measures of the similarity and commonality of the first and last names, with last name Jaro-Winkler score being the most informative. Thus, the XGBoost algorithm is using the same types of information that are commonly relied upon by other matching methods—geography, names, and birth year—but the algorithm learned which specific features were most important from the training data and is relying on those features more frequently. One important thing to note is that most important features (gender, birthplace, race, etc.) are built into our blocking strategy and as such are not used as features in our model since all possible matches are required to match exactly on these characteristics.

There are some concerns about using place of residence in our machine learning model as it could result in a matched sample that over-represents individuals who are less geographically mobile. We provide some insight about this issue in Figure 2, in which we show the tradeoff between precision and recall for three variations of our XGBoost model using Family Tree links that were

[14] See Appendix A for additional information about the XGBoost classifier. All of the code that we use to implement the XGBoost model is available to other researchers on Open-ICPSR (Price and Buckles 2021).

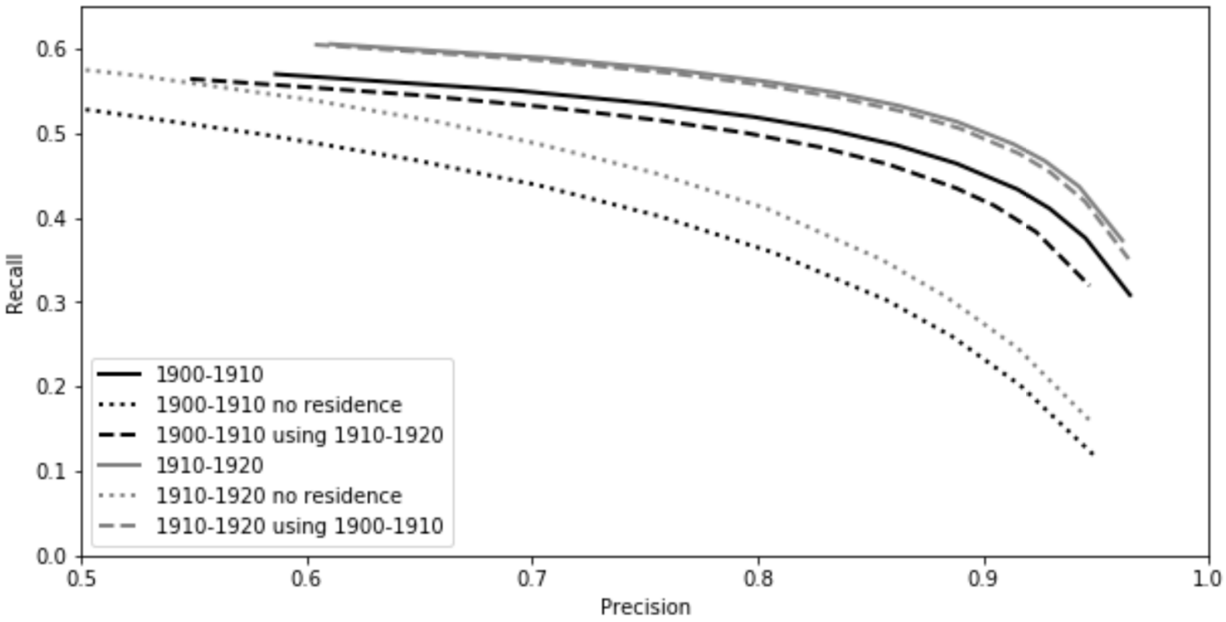
Table 5. Ten Most Important Features in XGBoost Model

Feature	Description	Feature Importance
Birth Year	The absolute difference in birth years.	0.237
Residential Geographical Distance	The distance between the coordinates (using latitude and longitude) of the residences.	0.226
Last Name Jaro-Winkler Score	The string similarity score of the last names as calculated by the Jaro-Winkler string similarity algorithm.	0.127
Mother's Birthplace	An indicator for whether the recorded Mother's Birthplace is the same.	0.094
First Name Jaro-Winkler Score	The string similarity score of the first names as calculated by the Jaro-Winkler string similarity algorithm.	0.074
Relation to Head of House	An indicator for whether the recorded Relation to Head of Household is the same	0.070
Father's Birthplace	An indicator for whether the recorded Mother's Birthplace is the same.	0.056
Commonality of First Name	The difference between the normalized commonality of the first names.	0.040
Commonality of Last Name	The difference between the normalized commonality of the last names.	0.033
Marriage Status	An indicator for whether the recorded marriage status is the same between censuses.	0.023

Notes: Features describe the characteristics of the records in a potential link. Feature weight is defined as the number of times a feature is used to make a decision in a decision tree used by the XGBoost algorithm. Feature importance is defined as the proportion of decisions that use the given feature. The model used has n=2,500 trees and N=17,303 decisions, and 12 possible features were available.

not part of the set of links we used to train the model. To create each curve, we incrementally adjust the cut-off that we use to label a pair of records as a match. When we lower that cut-off, we are able to identify more of the correct matches but do so at the cost of letting in more matches that are

Figure 2. Precision and Recall Curves for our XGBoost Predictions



Notes: Each curve is for a different machine learning model and the points along the curve provide the trade-off between precision and recall within each model based on the cut-off values that we use for a predicted match. We include our preferred specification along with the same model but without using place of residence as a feature and also our preferred specification but training the model with data from a different census year.

incorrect. Thus, the curves demonstrate the recognized trade-off between precision and recall within each model.[15]

To assess the value of including place of residence as a feature, we show the precision-recall curves for models that do and do not include it for each census pair. We see that we can still achieve high levels of precision and recall when not using the information on residence, but there is a cost to doing so. As an example, if we wanted to achieve a match rate of 40% for the 1900-1910 census, we

[15] The tradeoffs that we display in Figure 2 are all evaluated based on the links that we observe from the Family Tree. The links on the Family Tree are likely to be the easier links for humans to find and so this approach might overestimate the overall precision of these methods and may differ from the results for a random sample of the potentially linkable data. However, it provides a good measure of how well each of the methods can approximate what is possible through the time-consuming process of people using family history tools to find links by hand.

increase our false positive rate from 7% to 24% when we omit place of residence. In some cases—especially those where the inclusion of place of residence (or other time-varying characteristics) does not change the composition of the sample much—researchers may decide to include it as a way to reduce false positives. We assess the impact of using this information on the representativeness of our sample in Appendix Table 11, where we provide summary statistics for the 1920 samples that are linked to 1910 using either the full model or the model that excludes residence as a feature. As expected, both models over-represent those who live in their birth state, and the full model does so to a greater degree. However, on every other dimension, the samples produced by the two models are extremely similar. Thus, we find no evidence that using residence as a feature introduces greater selection bias in this setting.[16]

We are able to do one additional exercise with our data, to explore the possibility for "transfer learning," or the extent to which what we learn by making links between a pair of censuses can be used to make links between other data sets. If our machine learning algorithm is successful in making these links, this greatly expands the potential of the approach—an algorithm could be trained on the census data but used to make links between other data sets where training data are not available (i.e., among census data, vital records, school records, military records, etc.). While a full investigation of the transfer-learning performance of our algorithm is beyond the scope of this paper, we do have an informative experiment available to us—we can explore whether we can make successful matches when using data from the "wrong" pair of census years to train our machine learning models. The results of this exercise are shown using the dashed lines in Figure 2; we see that our precision and recall curves are only slightly worse when we use the 1910-1920 data and algorithm to find matches between 1900 and 1910, and *vice versa*. These results show that there is a

[16] See Antoine et al. (2020) and Helgertz et al. (2020) for recent discussions on the costs and benefits of including time-varying characteristics in matching models.

high level of transfer learning between census years, suggesting that our models are likely to have some success in creating links between other pairs of records.

D. Other steps in the process

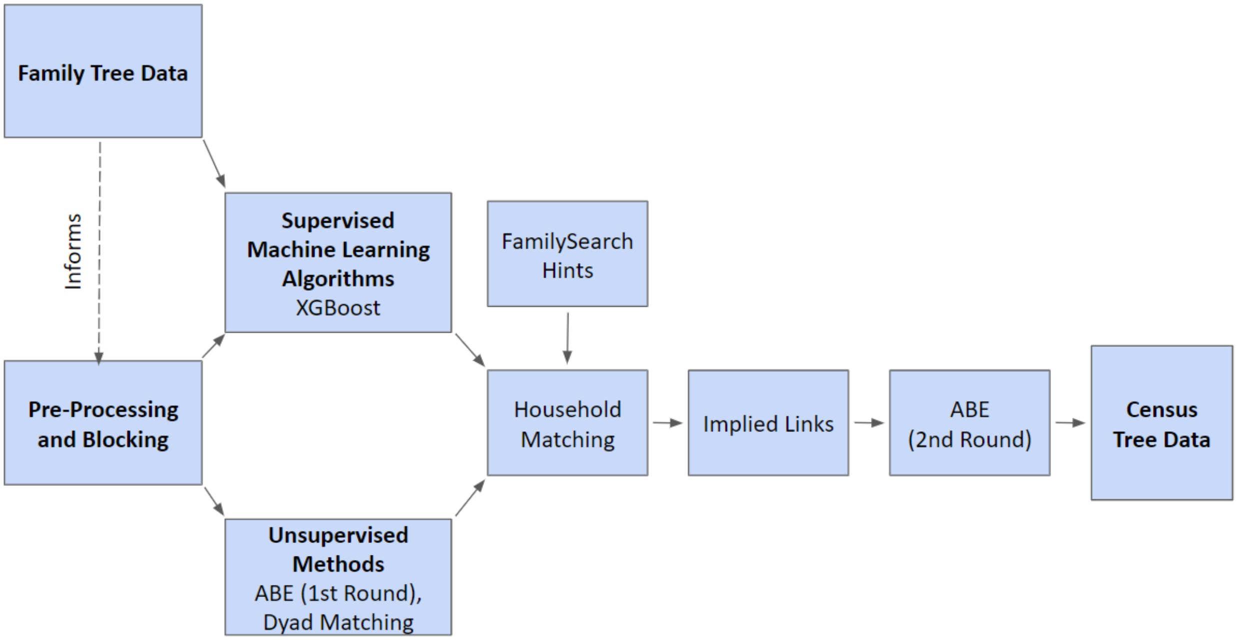
After pre-processing the data and selecting the blocking strategy and machine learning algorithm, we implement the supervised machine learning method.[17] In the early stages of this process, we asked trained research assistants to evaluate a random sample of the predicted matches and code them as either "true" or "false"; these links were then added to the training data set to further refine the matching algorithm.

In our full pipeline for linking census records to create the "Census Tree" data set, we also employ other linking strategies. This process is depicted in Figure 3. We use the same pre-processing and blocking strategies described above to inform our use of the automated matching strategy developed by Abramitzky, Boustan, and Eriksson (2014) (ABE). We also use a dyad matching method that involves linking multiple people across households at the same time. This approach is helpful when the individual-level 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.[18]

[17] While there are over 40.6 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 precision and recall plateau at around 500 observations in the training data. The training data we use has 42,165 true matches and 257,835 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.

[18] 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.

Figure 3. Process for Linking Census Records



Notes: Figure 3 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.

The methods to this point are performed separately and independently, and they may make conflicting predictions. To resolve these conflicts, we use a process we call "sheet-checking," which is based on the number of links between census sheets. Each census sheet has up to 50 individuals, and when we combine the links together, we can observe how many links occur between any pair of sheets. Because there are over a trillion combinations of sheet-to-sheet pairs, most incorrect matches are the only match between those two sheets, but correct matches will often have other pairs from family members or neighbors. We hand-tested this approach for 300 cases and found that the individual link with the highest sheet-to-sheet count was nearly always the correct match (only one was incorrect). Thus, when two sets of predictions conflict with each other (for example, one 1900 record is matched to two different 1910 records), we count how many matches there are between each pair of census sheets. Whenever there is a disagreement between two links, we choose the link that is part of a sheet-to-sheet pair that has the most additional links. In cases where the sheet-to-sheet count is the same for the two links, we do not use either link (which is similar to the ABE approach of excluding non-unique matches).[19]

Next, 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.[20] At this stage, we also include additional record hints that FamilySearch shared with us using their own machine learning algorithm. The files they shared provided us with 55 million links; of these, 6.5 million were links that we did not identify using one of the other methods.[21] Next, once we have implemented the full process to link 1920 to 1900, we can use 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 from 1920 to a record in both 1900 and 1910, we also now have a link between 1900 and 1910.

Finally, we run the ABE model once more, on the set of all still-unmatched records. Because the set of unmatched records is much smaller, this method is able to find additional new matches in

[19] In a previous version of this paper (Price et al., 2019), we did not implement sheet matching. As a result, the match rate was higher but precision was lower.

[20] 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.

[21] Potential links identified by FamilySearch and suggested to users are never automatically added to a person's profile; the user must make the decision to attach. This type of clerical review by the users is an important aspect of the platform. FamilySearch sets the threshold for precision in their record hints very high, at 95%. While this high threshold means lower recall, it ensures that the vast majority of the records that are attached are created by users conducting their own searches. We note that this is the only one of our linking strategies that is not readily available to other researchers.

cases where previously multiple matches were possible, but now some of those have been removed because they were linked to another record. To produce the final data set, we sheet-check one last time to get rid of any conflicting matches.

V. Results

We now apply this process to produce a linked data set of individuals across the 1900, 1910 and 1920 US censuses. Table 6 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 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 our XGBoost data, after sheet checking, provides about 28.0 million links. The ABE automated linking method and the household matching method contribute an additional 3.8 and 3.5

Table 6. Contributions of the Different Methods Used to Link Records

	1900-1910		1910-1920	
Method	Matches	New Matches	Matches	New Matches
XGBoost	20,271,149	20,271,149	28,033,131	28,033,131
ABE	17,093,182	3,435,666	21,811,611	3,790,831
Household Matching	13,656,233	3,602,120	6,710,830	3,494,771
Dyad Matching	12,846,431	851,977	25,085,386	4,566,309
Family Tree Data	15,079,238	4,520,685	17,234,083	4,249,998
FamilySearch Hints	29,850,308	3,212,511	24,705,392	3,319,395
Implied from 1900-1920	30,021,427	1,836,498	26,846,493	1,522,155
ABE round 2	1,327,768	1,055,220	1,291,530	1,110,166
Total		38,785,826		50,086,756
Match rate		61.6%		65.2%

Notes: Table summarizes the full Census Tree data set. The matches column provides the total number of matches created before conflicting matches were dropped at each step of the iterative matching process. The new matches column provides 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 B for details).

million links each. Dyad matching provides 4.6 million links, the original Family Tree data adds 4.2 million links, and the FamilySearch hints provide an additional 3.3 million links. The marginal contribution of each of these methods depends on the order that we list them since the various methods find many of the same matches as each other. For example, by subtracting the new matches from the matches for ABE, we find that XGBoost and ABE share 18 million links in common.

The results in Table 6 also demonstrate the value of implementing the ABE method a second time. Because the number of unlinked individuals is now much smaller, so that some matches that were non-unique the first time it was implemented now are, we are able to create an additional 1.1 million links. Altogether, the final Census Tree linked sample for 1910 to 1920 consists of 50.1 million links. Given our estimate of the number of possible matches, we conclude that we have identified 65.2% of the possible links between these two censuses.[22] For 1900 to 1910, we identify 42.7 million links for a match rate of 61.6%.

Before we present an in-depth analysis of precision and recall, we first want to turn our attention to the representativeness of the Census Tree sample. It is critical to think about whether the samples are representative of the population; if they are not, empirical analysis that uses these samples is vulnerable to sample selection bias. This is a particular concern given that our methods are informed by links made by FamilySearch users, who might be selected along characteristics including religion, race and ethnicity, education, or access to the internet. Furthermore, they might be more likely to attach records to profiles for relatives who are more successful, or easier to find.

[22] These calculations exclude children under the age of eleven, as we would not expect to be able to find them in the previous decennial census. See Appendix B for a detailed description of how we estimate the number of possible matches.

While we do not have information about the users themselves, we can compare the characteristics of samples produced by our various methods, to assess how well they compare to the US population.

These results are in Table 7. The samples are limited to those ages eleven and over, to omit those who would not have been born in the previous census and therefore cannot possibly be matched. For brevity we show the results for the 1910-20 links; results for the 1900-10 links are available in Appendix C. We also include summary statistics for links from the CLP.

First, comparing the people in 1920 that we are able to link to 1910 using our full linking pipeline (the Census Tree) to the 1920 population, we see that the Census Tree is very comparable to the population in terms of gender, age, household characteristics, and occupation score. The most important differences are that the Census Tree over-represents white Americans and those who were born in their birth state—though on the former, the Census Tree does include a higher percent of Black Americans than any of the other methods. We discuss the representation of Black Americans in our data in more detail below. The matches made by the XGBoost algorithm alone are slightly less representative in most cases, suggesting that the other steps in the pipeline are helping to add observations from groups where the algorithm does not perform as well. The fourth column shows the subset of census records that are attached to the Family Tree; this gives us a sense of the types of records that FamilySearch users are more likely to search for and find. The individuals with records attached to the Family Tree are more likely to be white, married, and have larger families.

The fifth column of Table 7 allows us to compare the characteristics of matches in the Census Tree data set to those available from the Census Linking Project. Along many dimensions, the Census Tree and the CLP data sets are very similar, but there are two big differences. First, looking at the number of observations, we see that the complete Census Tree data set, which includes links taken directly from the Family Tree, includes 3.7 times as many links as

Table 7. Summary Statistics for Subsamples of the 1920 Census

	Full 1920 Census	Matched in Census Tree	Matched by XGBoost	On the FamilyTree	Census Linking Project	Full 1920 Census (Men Only)
Female	0.488	0.478	0.456	0.450	0	0
	(0.502)	(0.500)	(0.498)	(0.498)	-	-
White	0.894	0.928	0.942	0.986	0.931	0.903
	(0.308)	(0.259)	(0.234)	(0.116)	(0.254)	(0.296)
Black	0.099	0.069	0.055	0.012	0.056	0.097
	(0.298)	(0.254)	(0.229)	(0.108)	(0.230)	(0.296)
Married	0.533	0.524	0.486	0.716	0.501	0.527
	(0.499)	(0.499)	(0.500)	(0.451)	(0.500)	(0.499)
HH head	0.303	0.312	0.313	0.301	0.497	0.419
	(0.460)	(0.463)	(0.464)	(0.459)	(0.500)	(0.493)
Age	34.570	35.071	34.864	33.810	33.386	34.871
	(16.895)	(17.584)	(17.946)	(15.797)	(17.720)	(16.841)
Lives in	0.593	0.657	0.678	0.691	0.651	0.577
birth state	(0.491)	(0.475)	(0.467)	(0.462)	(0.477)	(0.494)
HH size	7.718	7.657	7.555	8.223	4.740	7.953
	(6.821)	(6.030)	(5.800)	(5.686)	(2.553)	(7.307)
Speaks	0.944	0.961	0.962	0.964	0.952	0.949
English	(0.229)	(0.194)	(0.191)	(0.186)	(0.213)	(0.221)
Literate	0.935	0.958	0.965	0.972	0.951	0.936
	(0.246)	(0.200)	(0.183)	(0.164)	(0.216)	(0.245)
Occupation	8.771	8.638	8.946	6.976	13.368	14.188
score	(12.441)	(12.559)	(12.766)	(11.630)	(13.862)	(13.644)
N	81,492,832	48,191,662	26,962,493	4,119,447	15,362,014	41,241,991

Notes: Column 1 shows the summary statistics for all people in the 1920 census who are at least 11 years old. Column 2 includes only those who are also matched to the 1910 census in our Census Tree. Column 3 includes only matches generated by the XGBoost algorithm. Column 4 is restricted to census records that are attached to a profile on the Family Tree. Column 5 includes links from the Census Linking Project (using the standard method and exact name matching). Column 6 includes only the men from the Column 1 sample. Standard deviations reported in parentheses.

the CLP data set. Second, the CLP excludes women entirely, while nearly half of the Census Tree sample is female.[23]

We want to be clear about the kinds of matches we are able to make for women—the majority of these links are for cases where both records are either before or after marriage. Matching women *across* a marriage, when their surnames usually change, remains a challenge with our approach. The subset of the Family Tree data that we use as training data include many of these across-marriage 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. Nevertheless, the across-marriage links from the Family Tree data set alone constitute a valuable resource for researchers—we have 241,466 user-generated across-marriage links between the 1900 and 1910 censuses, and 248,805 between 1910 and 1920. These large samples have the potential to be very useful, as previous matching methods have not been able to produce across-marriage samples anywhere near this size for women. More generally, the large sample sizes in the Family Tree should also permit researchers to re-weight the data to construct a sample that is more representative of the US population along any desired dimension.

In Table 8, 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 (70.2%/74.0%) and daughters (67.7%/72.6). 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

[23] The fact that the CLP data set omits women also explains why the occupation score, percent that are household head, and household size from that sample are not as representative of the population as they are in our data. In the last column of Table 7 we restrict the full census sample to include only men, and the differences in these characteristics are much smaller.

Table 8. Match Rates by Relationship Type

	1900-1910		1910-1920	
	Matches	Match Rate (%)	Matches	Match Rate (%)
Male head	10,240,366	59.84	13,446,515	63.21
Female head	1,118,276	46.57	1,565,690	52.71
Spouse	8,544,223	55.57	11,045,153	57.58
Sons	7,948,475	70.15	9,918,259	73.96
Daughters	7,208,082	67.74	9,020,592	72.58
Other males	2,007,312	32.56	2,702,90	38.1
Other females	1,732,953	35.17	2,394,388	41.33

Notes: Table 8 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 Appendix B for details).

were born after 1910 are not included when calculating match rates). The group with the next highest match rates is male heads of household (59.8%/63.2%), followed by the spouses of the head of household (55.6%/57.6%). The lowest match rates occur for other household members who are not part of the immediate nuclear family, where the match rates fall to 35.2% and 41.3% for women and 32.6% and 38.1% for men.

Having described the characteristics of those who are likely to be linked in the Census Tree, we now return to our discussion of the number and quality of the matches. To begin, we employ two methods to evaluate the false positive rate among the predicted matches that we obtain. First, we 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 (excluding the implied links) includes 21.2 million of these transitivity tests and for this set, we find an agreement rate of 94.1% which implies a false positive rate of 5.9%.

Second, we drew a random sample of 1,000 records from the 1920 census that had been linked by our method to the 1910 census. We asked research assistants to use traditional genealogy tools to hand link these individuals to the 1910 census without seeing the link that we had identified with our automated approach. They were able to find a 1910 census record for 97% of the people from the random sample, where many of the ones for whom they could not find a link were lodgers or boarders in the 1920 census. The link they identified agreed with our predicted link 93% of the time, again implying a false positive rate of about 7%.[24] In addition, we find that this false positive rate was about the same for men and women with a false positive rate of 6.7% for men and 7.5% for women.[25]

How do these measures compare to those achieved with other methods? To investigate this, we again use all of the links that are currently available at the CLP for 1900-1910 and for 1910-1920. As our "truth" set, we use the links that we obtained from the Family Tree (14.5 million for 1900-1910 and 16.3 million for 1910-1920).[26] Thus, recall is defined as the fraction of the links in the Family Tree that are identified by each method, while precision is the fraction of the identified links

[24] This is a very low false positive rate, particularly when compared to other methods that achieve a match rate that is as high as ours. Nevertheless, researchers may be concerned about the remaining false positives. We note that for some questions in the social sciences, the false positives that our linking process generates might 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.

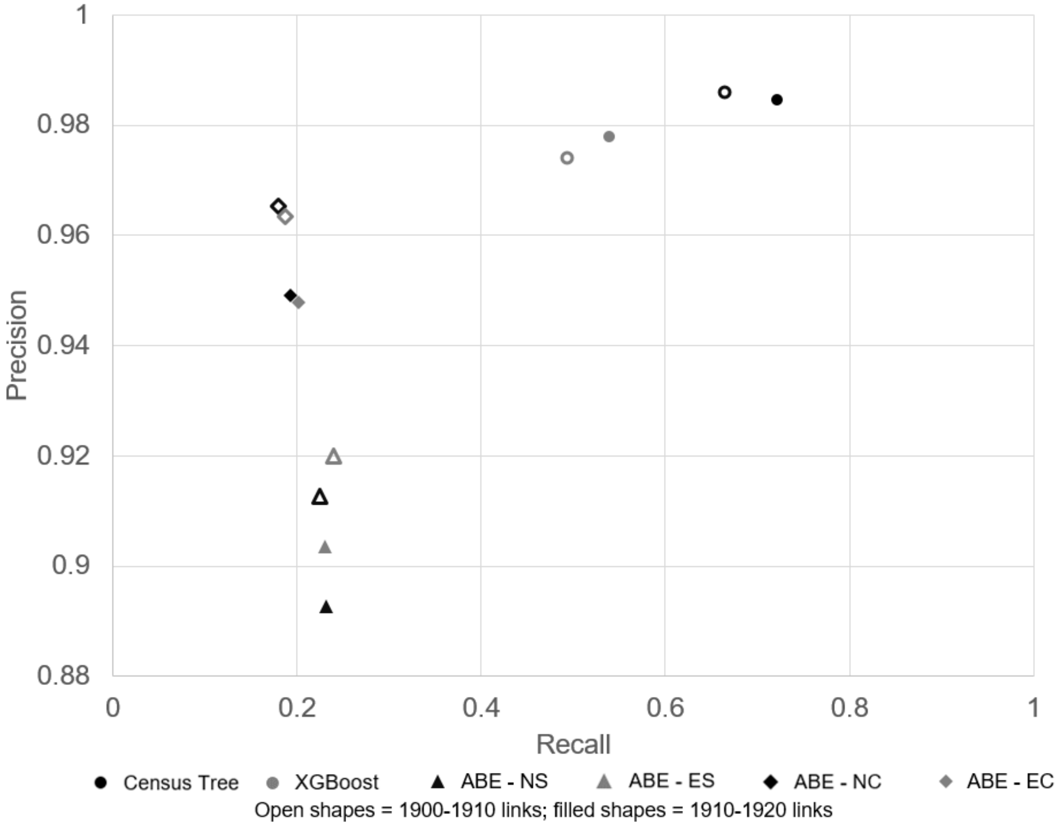
[25] We can also use the FamilySearch record-hinting system to continue to monitor the quality of our matches and use error analysis to improve the matching methods. FamilySearch has a system for emailing individuals using their platform about possible record hints for individuals that they are related 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.

[26] This is similar to Abramitzky et al. (2019), who use the links from the Union Army data set (uadata.org) as a "truth" set to compare accuracy and efficiency across linking methods.

that agree with the Family Tree. As in Figure 2, the Family Tree links are likely easier links for humans to find thus causing us to overstate the precision of these methods, but it approximates what is possible through time-consuming efforts to hand link records.

Figure 4 provides these estimates for four types of links that are currently available on the CLP website. All four sets of links are variations on the ABE approach that we use in our own linking process. The precision and recall of each of the four combinations of ABE for 1900-1910 and 1910-1920 are clustered on the left side of Figure 4. The more conservative methods are able to achieve a precision of 95-97%, which is comparable to the rate of precision that can be achieved by humans doing genealogy research. However, only around 20% of possible matches are made. For comparison, the figure also shows the precision and recall for the links created using only our machine learning model (XGBoost), and for the complete Census Tree data set. For the complete Census Tree, we removed any matches that were identified only by the Family Tree data which includes 4.5 million links for 1900-1910 and 4.2 million for 1910-1920. The XGBoost method alone achieves a level of precision and recall that is beyond the ABE frontier. Implementing our full linking approach (which includes ABE matching) allows us to attain precision of over 98% with recall of 67% and 72% for 1900-10 and 1910-20, respectively. In other words, we make three times as many matches as are currently available from the CLP, while improving on precision, including women, and obtaining a sample that is at least as representative of the full population.

Figure 4. Precision and Recall for Various Linking Algorithms



Notes: This figure shows the levels of precision and recall in data sets achieved by various linking methods, where the validation set is the 1900-1910 and 1910-1920 census links from the Family Tree excluding any links that were included in the training set. XGBoost refers to the matches created using only the machine learning algorithm described in this paper. "Census Tree" refers to matches in the full Census Tree data set, excluding matches that were made only by the Family Tree data. The remaining points show results for variations of the Abramitzky, Boustan, and Eriksson (2014) automated matching strategy. "N" indicates that NYSIIS was used to match names, while "E" indicates that an exact match was required. "S" indicates that their standard matching procedure was used, while "C" indicates the conservative method.

VI. Application: The Intergenerational Transmission of Literacy

As a final demonstration of the potential of our data for research, we calculate estimates of the intergenerational transmission of literacy using three different samples—the links taken directly from the Family Tree, links from our full Census Tree (which include the Family Tree links), and links from the CLP. To do this, we estimate OLS regressions where the dependent variable is a dummy variable indicating that the child is literate, and the independent variable is a dummy variable for the parent's literacy. The parent's literacy is observed in the 1900 census, when the child was age 5-15, and the child's literacy is observed in the linked 1920 census, when the child was age 25-35. For men, we regress the son's literacy on the father's; for women, we regress the daughter's literacy on the mother's. We estimate the intergenerational transmission of literacy separately for white and Black men and women.[27] Finally, we report coefficients from unweighted regressions, and from regressions which are weighted using inverse propensity score weights, following the procedure described in Bailey et al. (2019). While it is impossible to know the true value of the parameter of interest, we believe the weighted estimates from the Family Tree data set best approximate the "truth," as they are based on "ground truth" links made by users that are then weighted to account for sample selection bias.

The results from this exercise are in Table 9. Focusing first on the results for white men, we see that the Family Tree sample *alone* is able to provide 1.34 million links between fathers in 1900 and their sons in 1920. The sample size increases to 2.68 million with the implementation of our full Census Tree pipeline. This is about one million more links than are available in the sample from the CLP. The estimates from the unweighted regressions are similar, suggesting that having a literate father increases the son's probability of being literate by 0.065 to 0.077. When we weight the observations to better match the full population in 1900, the estimates are even more similar, and the Census Tree estimate is identical to the "true" value produced by the Family Tree sample. The smaller estimate using data from the Census Linking Project is consistent with lower precision in linking (a form of measurement error) attenuating coefficients towards zero—an issue that has long

[27] Summary statistics for all regression samples are provided in Appendix Table 13; we discuss these in more detail below.

Table 9. Estimates of the Intergenerational Transmission of Literacy

	Unweighted			Weighted		
			Census			Census
	Family	Census	Linking	Family	Census	Linking
	Tree	Tree	Project	Tree	Tree	Project
White Men	0.077	0.069	0.065	0.068	0.068	0.063
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
	1,335,514	2,684,550	1,595,654	1,330,115	2,669,788	1,587,659
Black Men	0.138	0.089	0.074	0.152	0.086	0.072
	(0.007)	(0.002)	(0.002)	(0.012)	(0.002)	(0.002)
	9,092	122,295	102,893	9,059	121,164	101,937
White Women	0.060	0.058	-	0.052	0.056	-
	(0.000)	(0.000)	-	(0.001)	(0.001)	-
	1,285,042	2,039,825		1,279,842	2,029,969	
Black Women	0.104	0.061	-	0.113	0.059	-
	(0.007)	(0.002)	-	(0.009)	(0.002)	-
	8,036	99,228		8,003	98,415	

Notes: Coefficients are from a regression of an indicator for the child's literacy on an indicator for that of the same-sex parent (fathers for boys and mothers for girls). Standard errors are in parentheses, and the number under the standard error is the number of observations in the regression. In the last three columns, the regressions are weighted using inverse propensity score weights, following the procedure described in Bailey et al. (2019). The Family Tree sample includes the links taken directly from the FamilySearch platform. The Census Tree sample includes the observations from our entire linking pipeline, including the Family Tree links. The Census Linking Project links are from their exact match/standard linking approach; these data do not include links for women.

been identified in the literature on intergenerational transmission estimates (Solon 1992) and documented in modern administrative data by Gross and Mueller-Smith (2020).[28]

[28] Gross and Muller-Smith (2020) use modern administrative data from the criminal justice system that allows them to link people based on fingerprints. They do a series of simulations in which they increasingly corrupt the record linking process, thus lowering the precision. This lower precision attenuates coefficients towards zero. They also found that lower recall can bias coefficients, but this depends on the issue of representativeness of the linked sample; weighting the data to be more representative of the population as we do in Table 9 can reduce this bias.

Consistent with the results from Table 7, the number of links available on the Family Tree for Black men is much smaller. Still, we have over 9,000 of these links, and we are able to increase that number to 122,295 with our full linking process. Looking to the weighted estimates, we see that the Census Tree estimate is closer to the "true" estimate from the Family Tree than the CLP estimate, though here we have less confidence that the Family Tree estimate is approximating the truth because it is based on a much smaller sample size.

Turning now to the results for women, we see that the Family Tree contains links between childhood and adulthood for 1.29 million white women. Our full Census Tree adds an additional 754,783 links—largely for women whose names do not change, or who can be linked using household or dyad matching. For both the weighted and unweighted estimates, the Family Tree and the Census Tree samples produce very similar results, which suggests that our full Census Tree process does not introduce significant selection bias. As we have emphasized above, the fact that we are able to provide a set of links for women that is nearly the size of the set for men is a substantial contribution that opens up new possibilities for research.

For Black women, we have many fewer observations, though we are able to identify nearly 100,000 links with our full process. Mirroring the results for Black men, we find that estimates of the intergenerational transmission coefficients for Black women vary across the samples, and we are less confident that we are able to approximate the "truth" with any of these estimates.

To summarize, this exercise highlights four advantages of the Census Tree data set for empirical applications. First, the Census Tree sample includes more parent-child links for this period than any other data set that currently exists. This will increase the precision of empirical estimates, which will likely be most valuable for researchers wanting to produce estimates using a small subsample (for example, people in a particular profession or from a specific location). Second, the fact that the Census Tree data set has higher precision (a lower rate of false positives) than other samples means that there is less measurement error in the links; this should reduce attenuation bias in estimates of intergenerational transmission. Third, as the results in Appendix Table 13 indicate, the Census Tree sample is more representative of the full population than either the Family Tree alone or the CLP. As one example, Black men who are literate are over-represented in all three of the linked data sets, but the problem is less acute in the Census Tree; as another, men on the Census Tree have very similar occupation scores to those in the full census. Fourth, we are able to produce estimates of the intergenerational transmission of literacy for women—something that is not possible using the CLP or any other large-scale linked data set that currently exists.

The results in Tables 7 and 9 also demonstrate that our methods have advanced efforts to link records for Black Americans. Between the 1910 and 1920 censuses, we have nearly 6 million Black Americans who can be linked to the prior census. We are aware of no other matching effort that has linked records for Black Americans at this scale. Nevertheless, our data fall short of being fully representative for racial minorities, and a top priority in our ongoing work is the continual improvement of our methods for linking records for Black Americans. As an example, we are currently working with the African Diaspora Experience Team at FamilySearch to increase further the coverage of African Americans on the Family Tree. This effort will grow the number of census-to-census links available on the Tree for this population, which in turn will improve our methods for making new matches with machine learning methods. To date, we have created profiles on FamilySearch for over 600,000 Black families and are working with volunteers from FamilySearch to attach additional census records to these profiles.

VII. Accessing the Data and Code for Research

We are committed to making the data and methods that we have described in this paper available to other researchers. To link the restricted version of the census to a subset of our training data from the Family Tree, researchers will first need to obtain access to the restricted versions of the complete count censuses for the relevant years. They can then use the data and code that we provide in our Open ICPSR repository to create the links based on our machine learning model (Price and Buckles 2021). The repository also contains the code needed to create the tables and figures in this paper, including the code needed to generate the features and to implement our blocking strategy. We have also arranged with the Census Linking Project to publicly disseminate all of the links that we create using our machine learning model (XGBoost) through their website and have shared those with them already. We will be expanding the approach that we use in this paper to include links for all census pairs between 1850-1940 and when that work is complete, those links will be available through the Census Linking Project website as well, and through a new ICPSR repository that we will create for this purpose.

While the data and code that we provide in the repository and on the Census Linking Project site will aid those who want to reproduce our Census Tree data or create a similar version of their own, our hope is that researchers will apply our methods to a wide range of linking projects. Anyone can access the Family Tree on familysearch.org once they have set up a free account. While FamilySearch does not allow individuals to download the entire Family Tree (which includes 1.2 billion profiles), there are multiple ways to query the Tree for the information that can be used to create a new training data set like the ones used in this paper. We describe these approaches in more detail in the "FamilySearch FAQ" that are included in Appendix D and available at https://sites.google.com/view/family-tree-faq/home.

VIII. 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 40.6 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. 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 properties of links made using alternative 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 approach, we combine the links made by our 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 62% of the potential matches between the 1900 and 1910 censuses, and 65% of the matches between 1910 and 1920, with a false positive rate of about 7%. Our approach therefore yields tens of millions of reliable links that can be used in academic research. As an application, we produce estimates of the intergenerational transmission of literacy for father-son and mother-daughter pairs. Our estimation samples are larger and more representative of the full population than are currently available from other state-of-the-art linked data sets, and the greater precision of our links appears to reduce attenuation bias from measurement error. Overall, we show that our links are highly representative of the population for white Americans, but that Black Americans continue to be under-represented. This is an important area for future work.

Ultimately, 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 Tree to include all full-count censuses between 1850 and 1940. Beyond this effort, we note that training data 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, and that transfer learning could potentially allow the Family Tree links to be used as training data for making links among these other sources. 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. Information, sample training data, and code are available in our ICPSR repository (Price and Buckles 2021) and on the Record Linking Lab website (https://rll.byu.edu) as an ongoing resource to access the innovations that we describe in this paper.

References

Abramitzky, Ran, Leah Boustan, and Katherine 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, James Feigenbaum, and Santiago Pérez. "Automated Linking of Historical Data." *Journal of Economic Literature,* forthcoming, 2019.

Abramitzky, Ran, Leah Boustan, and Myera Rashid. *Census Linking Project: Version 1.0 [data set].* 2020. https://censuslinkingproject.org

Abramitzky, Ran, Roy Mill, and Santiago Pérez. "Linking Individuals Across Historical Sources: A Fully Automated Approach." Historical Methods, 53, no. 2 (2020): 94-111.

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

Antoine, Luiza, Kris Inwood, Chris Minns, and Fraser Summerfield. "Selection Bias Encountered in the Systematic Linking of Historical Census Records." *Social Science History* 44, no. 3 (2020): 555-570.

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." *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.

Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. 2016.

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.

Christen, Peter. *Data Matching.* Springer, Berlin (2012).

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

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.

Feigenbaum, James J. "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.

Folkman, Tyler, Rey Furner, and Drew Pearson. "GenERes: A Genealogical Entity Resolution System." In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 495-501. IEEE, 2018.

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.

Gross, Matthew, and Michael Mueller-Smith. "Modernizing Person-Level Entity Resolution with Biometrically Linked Records." Working paper, 2020.

Feigenbaum, James, and Daniel Gross. "Automation and the Fate of Young Workers: Evidence from Telephone Operators in the Early 20th Century." National Bureau of Economic Research, Working Paper No. w28061, 2020.

Gutmann, Myron, Emily Merchant, and Evan Roberts. "'Big Data' in Economic History." *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, Joseph Price, Kelly Thompson, and Jacob Wellington. "A New Strategy for Linking Census Data: A Case Study Linking the 1900 and 1910 Full-Count US Censuses." *Working paper*, 2020.

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.

Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31, no. 2 (2017): 87-106.

Müller, Andreas C., and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. " O'Reilly Media, Inc.", 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." *Journal of Economic History* 79, no. 2 (2019): 383-416.

Price, Joseph, and Kasey Buckles. "Data and Code from Price, Buckles, Van Leeuwen, & Riley (Explorations in Economic History)." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. (2021). https://doi.org/10.3886/E130961V4.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. "Combining Family History and Machine learning to Link Historical Records." National Bureau of Economic Research, Working Paper No. w26227, 2019.

Solon, Gary. "Intergenerational income mobility in the United States." *The American Economic Review* 82, no. 3 (1992): 393-408.

"XGBoost Documentation." Accessed January 28 2020. https://xgboost.readthedocs.io/en/latest/.

Appendix A: The XGBoost Classifier

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.

In training our XGBoost model, we use a grid search method to select hyperparameters, including the number of trees, the depth of the trees, and the learning rate. We use 2500 trees and set the maximum depth of the trees to 3. Maximum depth indicates the number of nodes from the root node to the furthest leaf node in a tree. We also set the learning rate to 0.01, which indicates how the feature weights are updated in the gradient descent algorithm that XGBoost uses.

Additional resources:

- The XGBoost documentation is available at https://xgboost.readthedocs.io/en/latest/.
- See Chen and Guestrin (2016) for an introduction to XGBoost.
- Ancestry.com uses XGBoost to identify cases where multiple versions of the same person exist in their database; see Folkman, Furner, and Pearson (2018) for a description of their model and their reasons for choosing the XGBoost method.

- The code that we use to implement the XGBoost model in our setting is available on Open ICPSR (Price and Buckles 2021).

Appendix B: 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 10. 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 10. 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).

Appendix C: Additional Tables

Appendix Table 11: Summary statistics for samples in Figure 2

	Full 1920 Census	Full Model	Model Without Residence
Female	0.488	0.459	0.465
	(0.502)	(0.498)	(0.499)
White	0.894	0.945	0.950
	(0.308)	(0.227)	(0.219)
Black	0.099	0.055	0.050
	(0.298)	(0.227)	(0.219)
Married	0.533	0.464	0.457
	(0.499)	(0.499)	(0.498)
HH head	0.303	0.243	0.226
	(0.460)	(0.429)	(0.418)
Age	34.570	33.837	32.878
	(16.895)	(18.312)	(17.968)
Lives in birth state	0.593	0.712	0.674
	(0.491)	(0.453)	(0.469)
HH size	7.718	7.648	7.744
	(6.821)	(5.748)	(5.932)
Speaks English	0.944	0.962	0.960
	(0.229)	(0.190)	(0.196)
Literate	0.935	0.967	0.967
	(0.246)	(0.178)	(0.179)
Occupation score	8.771	8.580	8.342
	(12.441)	(12.625)	(12.515)
N	81,492,832	25,940,819	22,249,691

Notes: Column 1 shows the summary statistics for all people in the 1920 census who are at least 11 years old. Columns 2 and 3 correspond to the samples from the full 1910-20 model and the 1910-20 model without residence, respectively, used in Figure 2. Standard deviations in parentheses.

Appendix Table 12: Replication of Table 7, using 1910 census instead of 1920

	Full 1910 Census	Matched in Census Tree	Matched by XGBoost	On the FamilyTree	Census Linking Project	Full 1910 Census (Men Only)
Female	0.481	0.478	0.455	0.453	0	0
	(0.502)	(0.499)	(0.498)	(0.498)	-	-
White	0.889	0.926	0.941	0.983	0.923	0.901
	(0.315)	(0.261)	(0.236)	(0.131)	(0.267)	(0.298)
Black	0.110	0.071	0.057	0.015	0.058	0.099
	(0.303)	(0.256)	(0.231)	(0.122)	(0.234)	(0.298)
Married	0.510	0.512	0.463	0.684	0.479	0.499
	(0.500)	(0.499)	(0.499)	(0.465)	(0.500)	(0.500)
HH head	0.291	0.303	0.301	0.297	0.478	0.435
	(0.454)	(0..459)	(0.458)	(0.457)	(0.500)	(0.496)
Age	33.590	34.180	34.075	33.451	32.924	33.790
	(16.611)	(17.390)	(18.023)	(15.741)	(17.648)	(16.498)
Lives in	0.579	0.663	0.687	0.687	0.654	0.557
birth state	(0.494)	(0.472)	(0.464)	(0.464)	(0.476)	(0.497)
HH size	7.286	7.100	6.957	7.225	4.798	7.599
	(6.655)	(5.507)	(5.247)	(4.904)	(2.652)	(7.347)
Speaks	0.955	0.985	0.985	0.993	0.980	0.955
English	(0.208)	(0.119)	(0.123)	(0.083)	(0.141)	(0.207)
Literate	0.918	0.947	0.956	0.960	0.940	0.921
	(0.274)	(0.223)	(0.205)	(0.195)	(0.238)	(0.270)
Occupation	9.352	9.121	9.468	7.431	14.169	15.117
score	(12.663)	(12.810)	(13.040)	(11.931)	(13.826)	(13.493)
N	70,231,997	37,512,921	19,603,678	4,392,387	12,210,620	35,935,375

Notes: Column 1 shows the summary statistics for all people in the 1910 census who are at least 11 years old. Column 2 includes only those who are also matched to the 1900 census in our Census Tree. Column 3 includes only matches generated by the XGBoost algorithm. Column 4 is restricted to census records that are attached to a profile on the Family Tree. Column 5 includes links from the Census Linking Project (using the standard method and exact name matching). Column 6 includes only the men from the Column 1 sample. Standard deviations reported in parentheses.

Appendix Table 13: Summary Statistics for Samples Used in Table 9

Panel A: White Men

	Full 1920 Census	Matched in Census Tree	On the FamilyTree	Census Linking Project
Married	0.54	0.59	0.70	0.64
	(0.50)	(0.49)	(0.46)	(0.48)
HH head	0.39	0.43	0.52	0.48
	(0.49)	(0.50)	(0.50)	(0.50)
Age	27.43	29.47	29.77	29.73
	(4.58)	(3.13)	(3.15)	(3.14)
Lives in birth state	0.58	0.76	0.76	0.71
	(0.49)	(0.43)	(0.43)	(0.45)
HH size	7.83	7.20	7.16	7.13
	(7.47)	(6.06)	(5.29)	(6.38)
Speaks English	0.95	0.97	0.97	0.97
	(0.22)	(0.17)	(0.16)	(0.17)
Literate	0.96	0.98	0.99	0.98
	(0.20)	(0.13)	(0.12)	(0.13)
Occupation score	17.41	17.55	17.24	18.09
	(13.31)	(13.67)	(13.22)	(13.77)
# Observations	12,310,842	2,684,550	1,335,514	1,595,654

Panel B: Black Men

	Full 1920 Census	Matched in Census Tree	On the FamilyTree	Census Linking Project
Married	0.61	0.66	0.71	0.71
	(0.49)	(0.47)	(0.46)	(0.45)
HH head	0.44	0.46	0.50	0.52
	(0.50)	(0.50)	(0.50)	(0.50)
Age	26.94	29.19	29.59	29.58
	(4.62)	(3.11)	(3.17)	(3.15)
Lives in birth state	0.67	0.77	0.85	0.70
	(0.47)	(0.42)	(0.36)	(0.46)
HH size	7.70	7.76	8.41	7.53
	(7.82)	(6.63)	(5.86)	(7.02)
Speaks English	0.96	0.96	0.97	0.96
	(0.20)	(0.19)	(0.17)	(0.19)
Literate	0.79	0.84	0.86	0.85
	(0.41)	(0.36)	(0.35)	(0.36)
Occupation score	13.95	14.07	13.46	15.09
	(8.94)	(9.71)	(9.14)	(10.01)
# Observations	1,283,064	122,295	9,092	102,893

Panel C: White Women

	Full 1920 Census	Matched in Census Tree	On the FamilyTree	Census Linking Project
Married	0.68	0.58	0.77	-
	(0.47)	(0.49)	(0.42)	
HH head	0.02	0.02	0.01	-
	(0.13)	(0.13)	(0.08)	
Age	27.20	29.52	29.75	-
	(4.55)	(3.16)	(3.14)	
Lives in birth state	0.61	0.77	0.76	-
	(0.49)	(0.42)	(0.43)	
HH size	7.27	7.38	7.30	-
	(6.17)	(5.58)	(4.98)	
Speaks English	0.94	0.97	0.97	-
	(0.24)	(0.18)	(0.17)	
Literate	0.96	0.99	0.99	-
	(0.19)	(0.12)	(0.10)	
Occupation score	4.58	5.69	3.11	-
	(9.35)	(10.28)	(8.13)	
# Observations	12,226,733	2,039,825	1,285,042	0

Panel D: Black Women

	Full 1920 Census	Matched in Census Tree	On the FamilyTree	Census Linking Project
Married	0.71	0.63	0.74	-
	(0.45)	(0.48)	(0.44)	
HH head	0.07	0.05	0.02	-
	(0.25)	(0.23)	(0.13)	
Age	26.74	29.24	29.59	-
	(4.58)	(3.15)	(3.17)	
Lives in birth state	0.71	0.79	0.85	-
	(0.45)	(0.41)	(0.36)	
HH size	7.06	8.01	8.88	-
	(5.66)	(6.00)	(5.48)	
Speaks English	0.96	0.96	0.96	-
	(0.20)	(0.20)	(0.19)	
Literate	0.83	0.87	0.90	-
	(0.38)	(0.33)	(0.30)	
Occupation score	3.69	4.04	2.86	-
	(6.20)	(7.33)	(6.33)	
# Observations	1,476,066	99,228	8,036	0

Notes: Samples correspond to those used in the unweighted regressions in Table 9. For the full 1920 Census, sample is limited to those with non-missing literacy measures. Standard deviations are in parentheses.

Appendix D: Family Search FAQ

This is an adapted version of the FAQ that appear online at https://sites.google.com/view/family-tree-faq/home.

This document will provide some basic information about the Family Tree at FamilySearch.org. The goal is to help researchers understand how this data platform works, how the research team at the Brigham Young University Record Linking Lab has used the data to create new data sources and research, and how unaffiliated researchers might be able to use it in their own work.

What is FamilySearch and the Family Tree?

FamilySearch is a "non-profit family history organization dedicated to connecting families across generations" that is offered as a service of the Church of Jesus Christ of Latter-day Saints (https://www.familysearch.org/en/home/about). FamilySearch started in 1894 as the Utah Genealogical Society. It has a website (familysearch.org) that provides access to historical record collections and a wiki-style platform, called the Family Tree, through which individuals can gather information about their ancestors. In 2020 there were over 1.2 billion individual profiles on the Family Tree and over 12 million registered users.

How does FamilySearch work for its registered users?

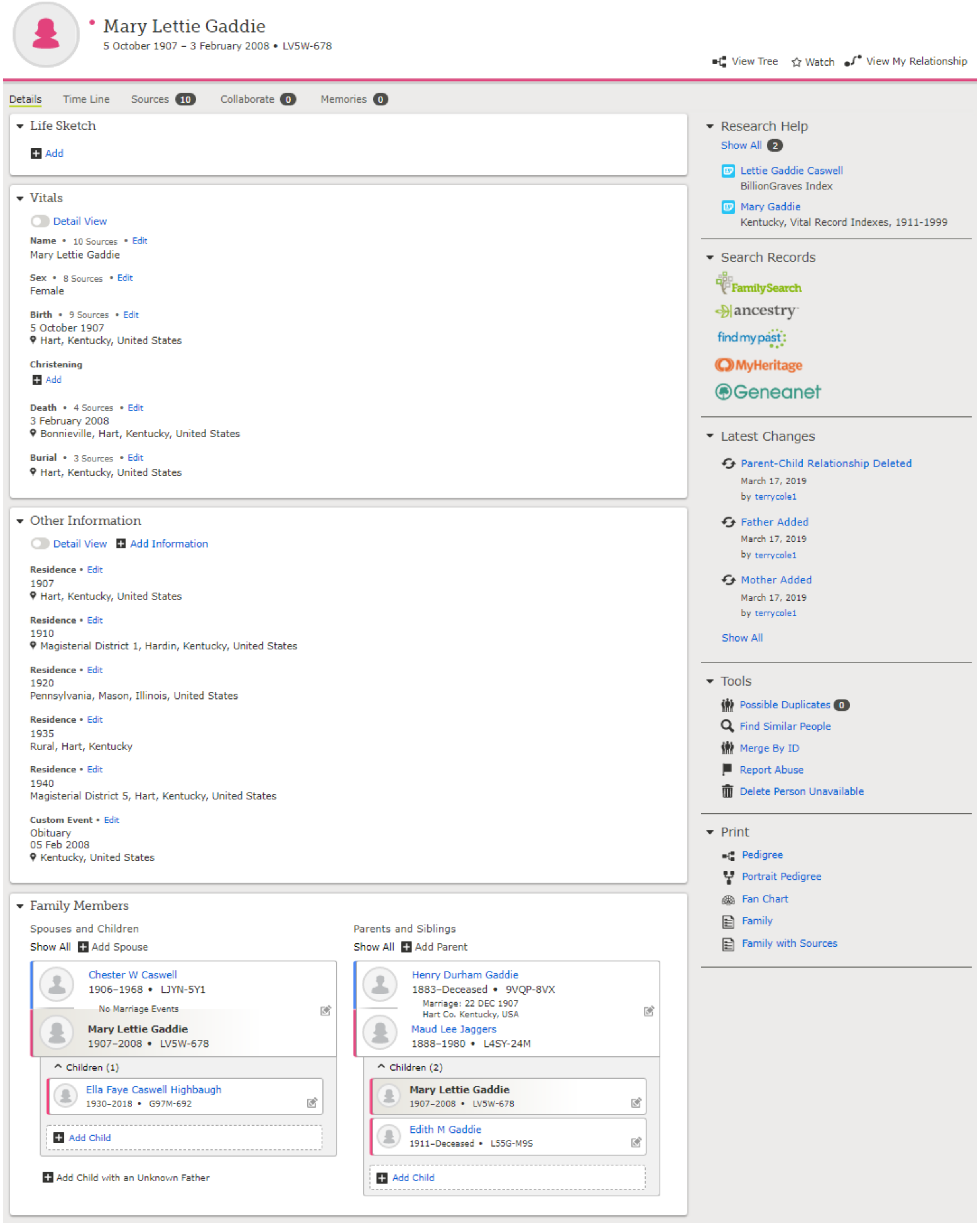
When someone first registers on FamilySearch.org, they enter information that they know about their parents, grandparents, and other relatives. If a deceased relative appears to be similar to a profile that is already on the Tree (e.g. in terms of name, dates of birth or death, birth location) then the site will suggest that the existing profile be linked to the user's family tree. In this way, the user's individual tree becomes connected to the large, wiki-style Family Tree—the largest of which connects over 400 million profiles. It is very common for users with ancestors in the United States—and increasingly with ancestors elsewhere—to quickly find a relative with an existing profile that allows them to link into the Family Tree. The Family Tree is a "wiki" in the sense that it is a public, shared platform, and when individuals have ancestors in common, any one of them can add and edit information and anyone else will see those edits when they visit the profile of the individuals.

Users may also start to attach records to relatives' FamilySearch profiles, including census records, birth and death certificates, images from yearbooks or newspapers, church records, and military records. FamilySearch has a growing collection of over 4 billion digital images for these records, and it partners with other record sources including Ancestry.com, Findmypast.com, and others. Users can search the digitized record collection themselves using the site's search forms. The FamilySearch website can also suggest possible record matches to users when the digitized information is similar to that provided by the user on the profile.

It is important to remember that while individual users can include living persons in their own personal family tree, only information about deceased persons appears on the wiki-style Family Tree. As a result, researchers only have access to records for deceased persons.

What is included in an individual profile on the Family Tree?

The image below includes an example of an individual profile on the Family Tree. Each profile includes a section about vital information, family members, sources, notes, and memories. On the right of the profile there are also help features that provide links to record hints for the person, flags for errors about their profile (e.g. being born after a parent dies), search tools, an edit history, and other features.



Mary Lettie Gaddie
5 October 1907 – 3 February 2008 • LV5W-678

View Tree · Watch · View My Relationship

Details · Time Line · Sources 10 · Collaborate 0 · Memories 0

▾ Life Sketch

➕ Add

▾ Vitals

Detail View

Name • 10 Sources • Edit
Mary Lettie Gaddie

Sex • 8 Sources • Edit
Female

Birth • 9 Sources • Edit
5 October 1907
Hart, Kentucky, United States

Christening
➕ Add

Death • 4 Sources • Edit
3 February 2008
Bonnieville, Hart, Kentucky, United States

Burial • 3 Sources • Edit
Hart, Kentucky, United States

▾ Other Information

Detail View ➕ Add Information

Residence • Edit
1907
Hart, Kentucky, United States

Residence • Edit
1910
Magisterial District 1, Hardin, Kentucky, United States

Residence • Edit
1920
Pennsylvania, Mason, Illinois, United States

Residence • Edit
1935
Rural, Hart, Kentucky

Residence • Edit
1940
Magisterial District 5, Hart, Kentucky, United States

Custom Event • Edit
Obituary
05 Feb 2008
Kentucky, United States

▾ Family Members

Spouses and Children
Show All ➕ Add Spouse

Chester W Caswell
1906–1968 • LJYN-5Y1
No Marriage Events
Mary Lettie Gaddie
1907–2008 • LV5W-678

^ Children (1)
Ella Faye Caswell Highbaugh
1930–2018 • G97M-692
➕ Add Child

➕ Add Child with an Unknown Father

Parents and Siblings
Show All ➕ Add Parent

Henry Durham Gaddie
1883–Deceased • 9VQP-8VX
Marriage: 22 DEC 1907
Hart Co. Kentucky, USA
Maud Lee Jaggers
1888–1980 • L4SY-24M

^ Children (2)
Mary Lettie Gaddie
1907–2008 • LV5W-678
Edith M Gaddie
1911–Deceased • L55G-M9S
➕ Add Child

▾ Research Help

Show All 2

Lettie Gaddie Caswell
BillionGraves Index

Mary Gaddie
Kentucky, Vital Record Indexes, 1911-1999

▾ Search Records

FamilySearch
ancestry
findmypast
MyHeritage
Geneanet

▾ Latest Changes

Parent-Child Relationship Deleted
March 17, 2019
by terrycole1

Father Added
March 17, 2019
by terrycole1

Mother Added
March 17, 2019
by terrycole1

Show All

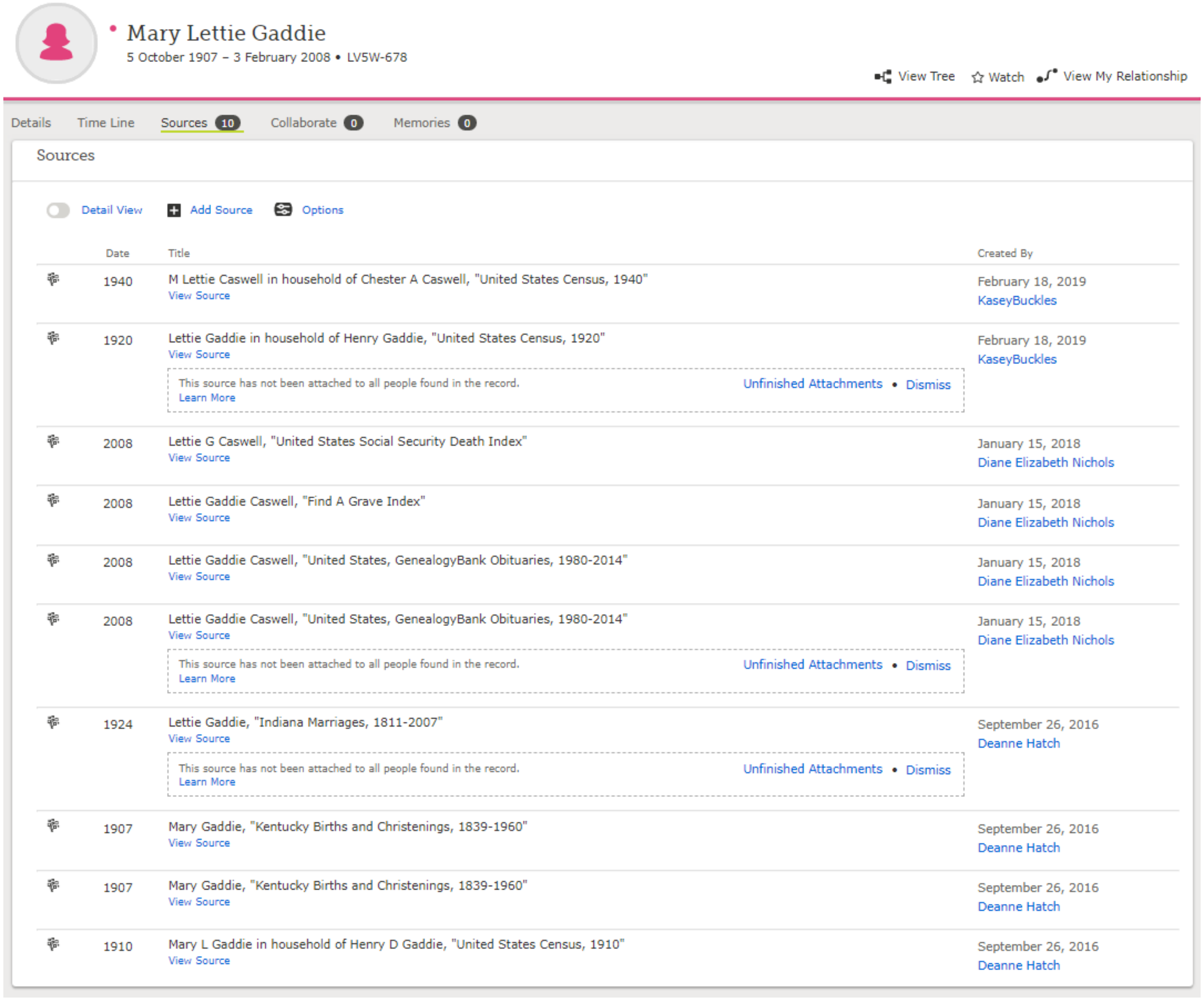
▾ Tools

Possible Duplicates 0
Find Similar People
Merge By ID
Report Abuse
Delete Person Unavailable

▾ Print

Pedigree
Portrait Pedigree
Fan Chart
Family
Family with Sources

The next image shows the source page for this profile. The source page identifies the user that attached the record, includes links for viewing the source, and provides hints for additional record linking based on the established records (described in detail below).



Mary Lettie Gaddie
5 October 1907 – 3 February 2008 • LV5W-678

Sources

Date	Title	Created By
1940	M Lettie Caswell in household of Chester A Caswell, "United States Census, 1940" View Source	February 18, 2019 KaseyBuckles
1920	Lettie Gaddie in household of Henry Gaddie, "United States Census, 1920" View Source This source has not been attached to all people found in the record. Learn More — Unfinished Attachments • Dismiss	February 18, 2019 KaseyBuckles
2008	Lettie G Caswell, "United States Social Security Death Index" View Source	January 15, 2018 Diane Elizabeth Nichols
2008	Lettie Gaddie Caswell, "Find A Grave Index" View Source	January 15, 2018 Diane Elizabeth Nichols
2008	Lettie Gaddie Caswell, "United States, GenealogyBank Obituaries, 1980-2014" View Source	January 15, 2018 Diane Elizabeth Nichols
2008	Lettie Gaddie Caswell, "United States, GenealogyBank Obituaries, 1980-2014" View Source This source has not been attached to all people found in the record. Learn More — Unfinished Attachments • Dismiss	January 15, 2018 Diane Elizabeth Nichols
1924	Lettie Gaddie, "Indiana Marriages, 1811-2007" View Source This source has not been attached to all people found in the record. Learn More — Unfinished Attachments • Dismiss	September 26, 2016 Deanne Hatch
1907	Mary Gaddie, "Kentucky Births and Christenings, 1839-1960" View Source	September 26, 2016 Deanne Hatch
1907	Mary Gaddie, "Kentucky Births and Christenings, 1839-1960" View Source	September 26, 2016 Deanne Hatch
1910	Mary L Gaddie in household of Henry D Gaddie, "United States Census, 1910" View Source	September 26, 2016 Deanne Hatch

How do users find records to attach to the Family Tree?

There are three ways that sources become attached to a profile on the Family Tree: record hints, search, and private information.

First, the most common way that a source becomes attached to a profile is through a record hint. These appear in the upper right hand of the screen when on an individual's profile page or are sometimes sent to users through an email campaign. These hints are generated using the matching algorithm that FamilySearch has developed which is based on a neural net using training data generated by genealogists.

Second, users will use the search features on FamilySearch to look for a person in various records. This approach allows the user to specify which features to use to search for the person including their name, birthplace, birth year, family member names, residence place, and race. They can employ narrow or wide ranges around dates of life events and include wild cards that allow a search query like: J* Pri?e which would search for anyone with a first name that that starts with J and has a surname that has the letters Pri_e where the blank could be any character. These advanced search features allow users to find records that get missed by the record hint algorithms.

Third, users find records in some other way, either through a manual search through all of the records for a particular town or using research compiled in other books or sources. Also, all of the other major genealogical websites (Ancestry, FindMyPast, MyHeritage) provide record hints and search tools and their users us this information that they find on other websites and attach the same sources to profiles on FamilySearch.

If I want to use the Family Tree to create a training set for machine learning, how will it compare to other methods of creating training data?

Social science researchers can use the FamilySearch platform to gather training data for machine learning algorithms. To see how this works, consider that one common way of creating training data is to train research assistants to search for possible matches (or provide them with potential matches found using a blocking strategy) and then ask them to label those possible matches as either true or false. The FamilySearch website effectively "crowdsources" this process, so that the site's millions of users are making the true matches.

There are two ways to think about the value of the training data that can be gathered from the Family Tree. The first way is in terms of the number of matches that it generates. Each day, an average of 720,000 sources are attached to the profiles on the Family Tree. Each of these attached records create a new pair of true links for each of the sources that were already attached to each of those individual's profiles. As a result, the amount of training data that can be gathered from the Family Tree will continue to grow and encompass links between a greater variety of data collections and countries. In fact, one of the most valuable aspects of this data will be the ability to create training sets for data sets between different countries as a way to create larger samples of migrants.

The second way this training data is valuable is in terms of the quality and variety of links that can be included. The traditional way of labeling training data requires the researcher to specify the blocking strategy that will be used to generate the possible comparisons. The true matches from the Family Tree provide important insights into what those blocking strategies should be and the trade-offs inherent in that decision between the size of the block and the probability that the true match is in the block. The research that people do on their own family members draws on a broad set of evidence and involves much more time considering the evidence across multiple sources and possible conflicts between those sources. This detailed detective work helps identify the matches in cases where the first and last name were switched on the census, where a 70 was erroneously transcribed as a 10, or where the human that was doing the transcription accidentally looked at the wrong row on the image. These are all mistakes that can be identified when users on FamilySearch have a strong incentive to discover the truth about their family. This type of truth discovery is nearly impossible and much too expensive for many of the current methods of creating training data.

There are two important caveats when using the Family Tree data as training data. First, under the traditional method of creating training data, the researcher knows each of the record pairs that the research assistants considered and which they labeled as true matches. With the FamilySearch data, the researcher only sees the conclusions that the users made and not the possible comparisons that they considered. For any given record that is attached to an individual profile on the Family Tree, it is not possible to know what process was used to make the decision. However, the researcher can use traditional blocking methods to identify the likely matches that would have been considered for any true match, and then label any non-chosen links as false.

Second, the level of skill and attention provided by different users can vary at lot. Many of the people on FamilySearch are professional genealogists and others are brand new. In theory, we could potentially gather information on the characteristics of the person that attached the source, but this is not something we have access to at this point. Instead, what we do is trust the decisions made by the humans, rely on the wiki feature of the website in which mistakes are frequently corrected, and acknowledge the fact that our training data will includes some mistakes. We provide some evidence on the quality of the data in response to a specific question in this FAQ.

How do I go about creating a training data set from the Family Tree?

Anyone can access the Family Tree on familysearch.org once they have set up a free account. While FamilySearch does not allow individuals to download the entire Family Tree (which includes 1.2 billion people), there are two ways to query the Tree for the information you need to create a training data set like the one described in this paper.

First, researchers can use the FamilySearch API to query for people on the Tree that meet certain criteria or who are attached to a particular record collection. For example, if we are trying to create a training set to identify links from the 1920 census to the 1910 census, then we can use the API to find people in the Tree who are already attached to the 1920 census, and then see which of them are already attached to the 1910 census. Alternatively, we can query the Tree for a particular type of person and use the API to gather all of the sources attached to the relevant profiles, and then use pairs of sources from those profiles to generate training data. Using the API requires an API key from FamilySearch; we provide information on obtaining acquiring an API key elsewhere in this FAQ.

Second, for smaller, customized linking projects, researchers have the option to interact directly with the profiles of individuals on the Family Tree. For example, we have already helped other researchers with a project linking Mothers' Pension records to vital records and census records. In this case, the researchers had names and other information from the Pension records, which they used to find the person on the Tree and link to any records contained within the profile. A researcher could also use the Family Tree as a platform for identifying correct links using the built-in record hints and search features on the website (including features that broaden the scope of possible comparisons). The links they create on the Family Tree could then be recorded and used as a training set. We are currently using this approach to help a researcher link records on politicians to census records, and also military records to census records.

Under either approach, the researcher still needs to have obtained their own access to the relevant record collection through (for example) the University of Minnesota/IPUMS, NBER, Ancestry, or FamilySearch and have a legal data agreement to use the actual data, as the Family Tree points to record sources but is separate from the record collections.

Does FamilySearch verify the links?

FamilySearch does not have an automated process to check whether records are attached to the correct person. Many of the sources that are attached are based on record hints provided by FamilySearch. Their match algorithm sets their precision threshold at 95% so that the links are highly likely to be a match, and the user plays a key role in providing a final validation before linking the source.

There are also features within the FamilySearch platform that help users catch and fix mistakes. For example, if someone attempts to attach a profile on the Family Tree to a source that is already attached to another profile, they will see the profile to which the record is already attached. At this point, it is easy for the user to compare the profiles, and to detach the source from the original profile if that is deemed to be an incorrect link. They are also able to put in a reason statement to help explain their decision. Furthermore, when users are doing research on a particular individual, they will often look over the sources attached to that person to look for additional information that should be included on the profile or identify family members that might have been missed. In the process of comparing information across multiple sources and reconciling conflicting information they will often discover that one of sources was incorrectly attached to that person and they can easily detach the source from the person. This wiki-style aspect of the Tree is one of the most important features to ensure high quality in the long-run (though it has the ability to lower quality in the short-run).

Is there any evidence on the quality of the links?

Here we summarize two exercises to examine the quality of the data; see the main text for more detail.

In the first exercise, we compare links from the Family Tree with the links created by the human trainers working on the LIFE-M project. 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 about 12,000 people from their sample that were attached to both an Ohio birth certificate and the 1940 census on the Family Tree. Of these, 1,060 links were identified by both LIFE-M and the Family Tree, and we found that that the links agreed 94% of the time. For the few cases where there was disagreement, we asked hand research assistants to use traditional family tools to determine which match was correct. Adding the links that the research assistants identified as correct to those where LIFE-M and the Family Tree data agree, we conclude that the links based on the Family Tree were correct 98% of the time.

In the second exercise, we began with 500,000 matches for our Ohio sample between the 1910 and 1920 censuses and randomly sampled 100 records from the 1920 census. We gave these 100 records 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 correct 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.

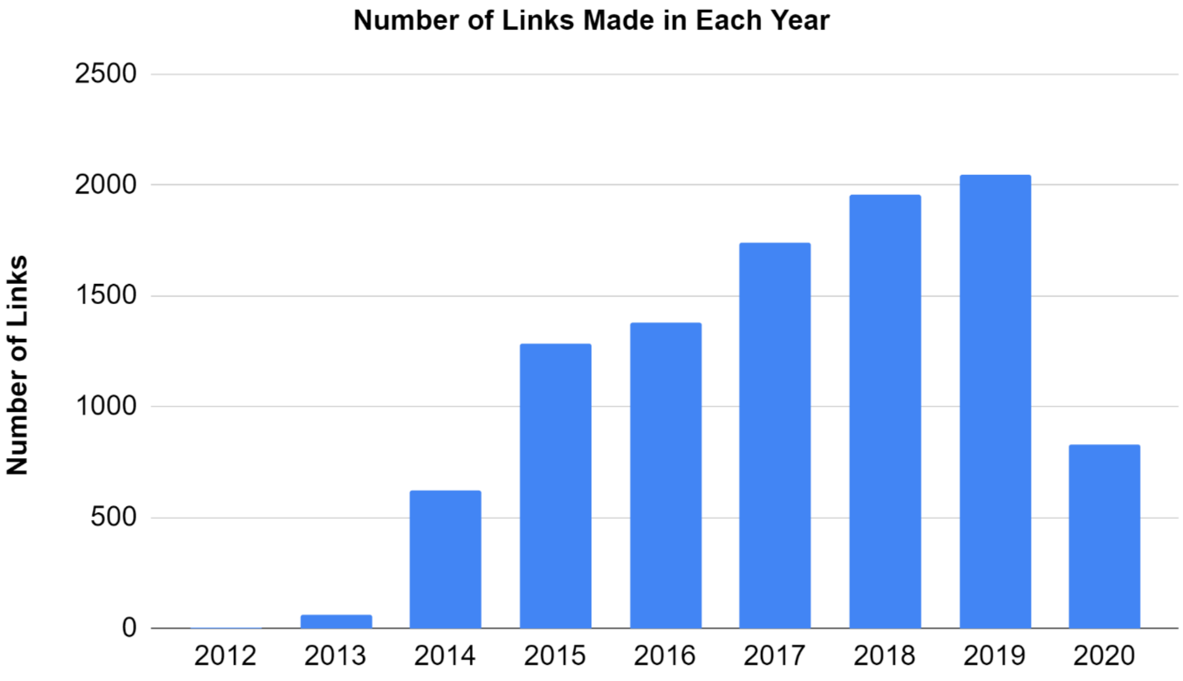
Both of these exercises suggest that the Family Tree links achieve a level of accuracy similar to or better than that created by skilled human trainers, at a much lower cost.

How much does the Family Tree change over time?

The wiki-style structure of the Family Tree means that the profiles on the Family Tree are being continually updated and edited. As such, the training data that we gather from the Family Tree at one point in time might differ from the training data we would obtain at a different point in time.

Each profile on the Family Tree includes an edit history that records every single change that has been made to the profile and the date it occurred and the user who made the change. In order to test how much the training data we use in this paper changes over time, we examined the edit history for a random sample of 10,000 linked pairs in our training data. For each linked pair, we gathered data from the edit history to see when each census source in the pair was attached to the profile. We find that 69% of the time, the two census records were linked to the individual's profile on the same day and 84% of the time they were linked in the same year. In cases where the year differed, we used the later of the two dates. The table below provides the year that each of our linked pairs were added to the Tree.

Note that almost half of our training pairs were created in 2018 or later. This confirms that the Tree continues to grow rapidly; those interested in using the Tree for training data will therefore want to use the most recent data available. Nevertheless, we have found that in practice, updating the training data with newly made links does not meaningfully change the number or composition of the links made by the machine learning algorithm.



This random sample of our training data also provides some interesting insight about the contributors to the training data. There were 9,141 unique registered users who attached one of the

20,000 sources that were used for our random sample of 10,000 linked pairs (two attached sources for each pair). Of these attached sources, 17% were attached by users who only attached one source, 41% were attached by users who attached two sources, 20% were attached by users who attached 3-4 sources, and the other 22% of sources were attached by users who attached five or more of the sources within this random sample of our training. In addition, of the matches in this random sample, 71% were attached by the same registered user.

Who are the users of FamilySearch, and how representative is the Family Tree of the general population?

There are two key questions to consider when thinking about how FamilySearch users might produce a Family Tree that is not representative of a population. First, what are the characteristics of the users themselves? Here, specific concerns include the facts that FamilySearch users have access to computers/smart phones, internet, and time, or that members of the Church of Jesus Christ of Latter-day Saints may be more prevalent among FamilySearch users than they are in the general population. Second, how might the behavior of users affect who ends up on the Tree and who does not? For example, are users more likely to look for or find information about successful relatives, which would lead to their over-representation on the Tree?

Unfortunately, we do not have demographic information that allows us to provide summary statistics for the 12 million+ FamilySearch users. However, we can compare the characteristics from records on the Family Tree to other population records, to help us assess the representativeness of the Tree. We summarize the key findings here; see Table 7 in the paper for the full results.

When comparing the census profiles that are on the Family Tree to the full population, we see that those on the tree are similar in terms of gender, age, household size, and the probability of being the household head. However, those on the Tree are more likely to be white, married, literate, and are more likely to be living in their birth state. Interestingly, we find that those on the Tree have a *lower* occupation score, which suggests that users are not more likely to look for or find information on more successful relatives.

While these results suggest that there is selection into the Tree along some characteristics, we note that when using the Family Tree data (or any samples produced using it as training data), it is possible to re-weight the sample to be representative of the desired population by following the procedure outlined in Bailey et al. (2019). The fact that the Family Tree includes over 1.2 billion profiles means that even for under-represented groups, there will likely be sufficient support in the data for this approach.

What about survivor bias?

Since people using a genealogical platform often seek out their own ancestors first, the Family Tree might under-represent individuals who never had descendants. One thing that allays this concern is that while users on FamilySearch tend to start by focus on finding information about their direct ancestors, they generally then turn their attention to doing descendancy research. Descendancy research is gathering information on the children, grandchildren, and great grandchildren of each of your ancestors. This approach to family history allows everyone the chance to be gathered into the Family Tree, including those with no living descendants.

To test for the extent of survivorship bias on the Family Tree, we took a random sample of women age 35 from the 1910 census and compared the coverage rate on the Family Tree of women based on whether or not they had ever had children (that census year includes a question for women about how many children they have ever had). We created a random sample of 5,000 women who had had children and 5,000 women who had never had children. We found that 46% of women with children in the 1910 census had a profile on the Tree, compared to 18% of women who had never had children. These results suggest that individuals who do not have children are less likely to currently have a profile on the Family Tree, but they do still have a significant presence and are included in our training data.

Are there other uses for the Family Tree data, beyond its use as training data?

Absolutely. As one example, the Family Tree itself provides a rich set of links that traditional linking methods and even machine learning are unlikely to be able to provide. To see this, consider the case of "maiden" names. Women's records from before and after marriage have historically been difficult to link because her surname usually changes. Neither the research assistant nor the machine learning algorithm has the information needed to link "Mary Gaddie" as a child to "Mary Caswell" as a married adult. However, family members often have this private information and can successfully create this link. Family members also have private information about occupations, geographical moves, and the names of family members. Given that the Family Tree contains tens of millions of record links, this is a valuable source of difficult-to-link records.

Additionally, because the Family Tree data contain millions of links that are considered "ground truth," they can be used to check the validity of other methods. For example, Abramitzky et al. (2019) and Bailey et al. (2019) have used the Family Tree data in this way.

How can I access the FamilySearch API?

Data from the Family Tree can be accessed using the FamilySearch API. Documentation for the API is provided on their API resources page (https://www.familysearch.org/developers/docs/api/resources). FamilySearch also provides a helpful getting started page (https://www.familysearch.org/developers/docs/guides/getting-started) . The FamilySearch API is designed primarily for App developers who use data from the Family Tree to create family history experiences; however, there is increasing interest to partner with academics as a way to increase the quality and growth of the Family Tree.

An App Key is required to use the FamilySearch API and instructions for how to obtain one are provided on the FamilySearch API getting started page. The BYU Record Linking Lab is willing to share the API that they have created to access data on the Family Tree; requests can be made by emailing rll@byu.edu. The BYU Record Linking Lab can also help identify ways that academic projects can facilitate the growth of data in the Family Tree or add value back to the Family Tree, as these are two of the primary considerations that FamilySearch uses when determining whether to grant access to an App Key. Data acquisition and record linking efforts in economics and other fields have tremendous potential to add value back to the Family Tree.

What is the Census Tree?

The Census Tree project will link the 217 million people that lived in the United States between 1850 and 1940 across each of the census records for these years. It will also connect each person to all of their one-hop relatives (parents, siblings, spouses, and children). The Census Tree will provide the largest longitudinal data set ever created in the United States and will open up many opportunities for research in economics, demography, sociology, and public health. To do this, we are combining the original Family Tree data with machine learning methods that use it as training data and with other linking methods. See Figure 3 and the text of this paper for more detail. The current version of the Census Tree data includes 38.8 million links between the 1900 and 1910 censuses, and 50.1 million links between the 1910 and 1920 censuses.

Where can I find the publicly available code and data that has been used in the Record Linking Lab's Family Tree and Census Tree projects?

All of the code used to produce the data, tables, and figures in this paper is available to other academic researchers and can be accessed on Open ICPSR (Price and Buckles 2021). The code is a combination of Python, SQL, and Stata. Detailed documentation is included to provide instructions on how to run each part of the process. Much of the code is related to getting the data into the correct format for the machine learning. For researchers with their data already prepared, they can use our trained model to do automated record linking and that step in the process is just a simple python file and we can provide a pre-trained model (based on census-to-census training data). The code is also easy to adapt to linking other combinations of records and we can provide example code for those types of linking as well.

One of the contributions of this paper is to highlight a previously untapped source for large, low cost, and high-quality training data. We are continuing to improve the size and quality of the training data and expanding our current code to including other data sets and other countries. We are open to helping others use these resources and work to improve the quality and coverage of the Family Tree.

As we continue to expand the Census Tree data set to other years, we plan to make the *histid* crosswalks for linking censuses available on multiple platforms, including Open ICPSR, IPUMS, and the Census Linking Project website.

Any other advice?

Yes! The best way to understand the FamilySearch platform and the Family Tree is to create an account and build your family tree. Doing so will help you gain a deeper understanding of the user experience, and how that experience is reflected in the data you want to use.